



PROCESO GESTIÓN DEL TALENTO HUMANO
FORMATO INFORME MENSUAL EJECUCIÓN CONTRACTUAL

Medellín, 30 de agosto del 2022

Señor (a)

John Jairo Castro Maldonado

SUPERVISOR(A) CONTRATO No. CO1.PCCNTR.3367636

Coordinador académico

Centro de Servicios y Gestión Empresarial

Medellín

Asunto: Informe mensual de ejecución contractual
de agosto del año 2022.

Referencia: No 3367636 del año 2022

Jesus Arbey Benavides Guevara, identificado con la cédula de ciudadanía No. 1.085.260.713 de Pasto, en mi calidad de Contratista del SENA, en Centro de Servicios y Gestión Empresarial, en cumplimiento del Contrato de Prestación de Servicios de la referencia, a continuación presento el Informe de actividades realizadas en el mes objeto de cobro.

Valor y forma de Pago: Se fija como valor total para el contrato la suma de TREINTA Y CUATRO MILLONES TRESCIENTOS MIL PESOS MCTE (\$ 34.300.000) Incluido IVA. Esta suma será pagada por el SENA al contratista de la siguiente manera: Valor por mes: CUATRO MILLONES NOVECIENTOS MIL PESOS MCTE (\$ 4.900.000), y el equivalente por fracción de mes.

Plazo: Será hasta el treinta (30) de agosto (08) de 2022.

OBJETO: Prestar servicios profesionales para el desarrollo de las actividades de levantamiento de indicadores, bases de datos y el modelo matemático con base en analítica de datos y Big Data en el prototipo de software para modelo de Innovación en el desarrollo del proyecto de I+D+i SGPS 9748

Obligaciones Específicas:

No	Obligaciones	Acciones realizadas	Evidencias
1	Apoyar el desarrollo de aplicaciones o algoritmos usados en la actualidad para solucionar de forma eficiente y eficaz problemas o retos identificados	Acompañamiento de implementación de mapas geoespaciales en la Interfaz de programación de aplicaciones (API), de los resultados de conglomerados generados por Kmeans y PCA para	ANEXOS 6. Actas de reuniones. ANEXO 10. Reporte de códigos para compatibilidad de API con resultado analítico.



	en Big Data y analítica de datos con aplicación en gestión de la innovación	los centros de formación de SENA Antioquia, con sus clasificaciones de alta, media y baja productividad y gestión de manera respectiva.	
2	Revisar las herramientas tecnológicas o digitales de Big Data y analítica de datos necesarias para dar cumplimiento a los objetivos propuestos dentro del proyecto.	Gestión realizada	<p>ANEXO 7. Participación en reuniones con equipo para el desarrollo del prototipo de software.</p> <p>ANEXO 9. Reporte de códigos para selección de modelos de analítica de datos.</p> <p>ANEXO 10. Reporte de códigos para compatibilidad de API con resultado analítico.</p>
3	Revisar y actualizar indicadores para la gestión de la innovación según objetivos del proyecto	Gestión realizada	<p>ANEXO 7. Participación en reuniones con equipo para el desarrollo del prototipo de software.</p> <p>ANEXO 9. Reporte de códigos para selección de modelos de analítica de datos.</p>
4	Realizar pruebas relacionadas con la compatibilidad de la tecnología y la aplicación de los sistemas de Big Data y analítica de datos para la evaluación y valoración de la Gestión de CTel para los Centros SENA Regional Antioquia	Gestión realizada	<p>ANEXO 7. Participación en reuniones con equipo para el desarrollo del prototipo de software.</p> <p>ANEXO 10. Reporte de códigos para compatibilidad de API con resultado analítico.</p>



5	Apoyar la elaboración del manual del sistema y usuario y las actividades de transferencia	Gestión realizada	ANEXO 10. Reporte de códigos para compatibilidad de API con resultado analítico.
6	Presentar los informes de seguimiento del proyecto de acuerdo con las solicitudes del líder del área o supervisor	Se elaboró el informe de pago mensual GTH-F-062 V09 Se realizó la redacción de informe de investigación resultados del proyecto.	Informe en formato pdf. ANEXOS 2. Evidencia de informe de investigación
7	Divulgar en eventos académicos/empresariales los resultados del proyecto.	Se recibió certificación de participación en Workshop on Modeling and Simulation for Science and Engineering, en modalidad de Artículo. Se realizó acompañamiento de modificación de autor de correspondencia y modificaciones a ser realizadas en artículo publicado en Workshop on Modeling and Simulation for Science and Engineering.	ANEXO 1. Medios de divulgación: Notificación de cambio de autor de correspondencia en Workshop on Modeling and Simulation for Science and Engineering ANEXO 8. Correos asociados a Proceso editorial artículos [V WMSSE-JPCS]
8	Redactar el informe de investigación y de resultados del proyecto.	Se elaboró el informe de pago mensual GTH-F-062 V09 Se realizó la redacción de informe de investigación resultados del proyecto.	Informe en formato pdf. ANEXOS 2. Evidencia de informe de investigación
9	Redactar y someter a publicación artículo de investigación, libro	Gestión realizada	ANEXO 1. Medios de divulgación: Notificación de cambio de autor de correspondencia en Workshop on



	de investigación o capítulo de libro de investigación relacionado con el proyecto, en medio reconocido por Minciencias.		Modeling and Simulation for Science and Engineering ANEXO 8. Correos asociados a Proceso editorial artículos [V WMSSE-JPCS]
10	Apoyar actividades de Grupo de Investigación, Semilleros de investigación y SENNOVA que sean propias de su área de conocimiento y en relación con el objeto contractual	Se realizó analítica de datos de 5 encuestas en apoyo a la instructora Luz Carime Chia del Centro de Comercio y Servicios- SENA Cauca, de la siguiente manera: <ol style="list-style-type: none"> 1. Analítica de datos CUESTIONARIO 1- Aprendices. 2. Analítica de datos CUESTIONARIO 2 – Instructores 3. Analítica de datos _ENCUESTA_EQ UIPO PEDAGÓGICO 4. Analítica de datos _ENCUESTA_EQ UIPO_EJECUTOR 	ANEXO 4. Analítica de datos – Encuestas. Soporte de correos, entrega de resultados en código. ANEXOS 5. Códigos de apoyo en analítica de datos - Centro de Comercio y Servicios- SENA Cauca Link de códigos y datos: https://bit.ly/3K7IKFS
11	Apoyar la evaluación técnica, administrativa y presupuestal de los proyectos de investigación aplicada, desarrollo tecnológico e innovación que se formulan por el SENA para el desarrollo de actividades misionales, acorde con las	Se realizó taller de apoyo a aprendices del “Curso de estadística aplicada a inteligencia artificial” para la apropiación de conceptos de estadística descriptiva.	ANEXOS 3. Apoyo a la formación - Curso de estadística aplicada a inteligencia artificial.



	indicaciones dadas por la Dirección General.		
12	Desarrollar sus actuaciones en el marco de la normatividad vigente y presentar los informes mensuales de acuerdo con el objeto del presente contrato.	Se elaboró un (1) informe de pago mensual GTH-F-062 V09.	Informe en formato pdf.

A continuación, relaciono los desplazamientos que realicé previo a la presentación de este informe. Una vez finalizado cada desplazamiento presenté al ordenador del gasto el informe en el Formato Informe Legalización Desplazamiento Contratista GTH-F-087, en el que se describieron las actividades desarrolladas y los resultados de cada desplazamiento. Cada informe cuenta con el visto bueno del Supervisor.

Se lista a continuación el soporte de la legalización de los desplazamientos realizados, los cuales forman parte integral del presente informe de ejecución contractual.

ITEM	No DE LA ORDEN DE VIAJE	LUGAR DE DESPLAZAMIENTO	FECHAS DE DESPLAZAMIENTOS	
			INICIO	FINALIZACIÓN
1.		No he realizado ningún desplazamiento		

Nota 1: Por cada desplazamiento que haya realizado el contratista, adjuntará el respectivo informe que la soporte. En caso de haber realizado el desplazamiento en fecha posterior a la presentación del informe de ejecución contractual, deberá reportarlo en el siguiente informe de ejecución contractual.

Para el trámite de la cuenta me permito adjuntar: Documentos electrónicos enunciados como evidencias del cumplimiento de las obligaciones contractuales y los desplazamientos realizados y el No. 60626908 de la planilla, miplanilla correspondiente al mes de julio del año 2022. (Decreto Ley 2106 de 2019 – “Decreto Ley Antitrámites”)

Declaración bajo la gravedad de juramento.

Me permito declarar **bajo la gravedad de juramento** que durante el periodo objeto de cobro SI___NO_X_ he suscrito otro u otros contratos con entidad(es) del Estado el (los) cual(es) se



encuentra(n) en ejecución; la anterior declaración de acuerdo con lo establecido en el numeral 5º de la Directiva Presidencial No. 001 de 2022, a fin de identificar potenciales conflictos de interés y hacer la gestión del riesgos para la toma de decisiones dentro del principio de la debida diligencia frente a la ejecución de los contratos de prestación de servicios profesionales o de apoyo a la gestión.

Nota: En el evento de señalar **SI**, el análisis efectuado por el supervisor y ordenador de pago deberá quedar documentado y publicado en el SECOP II en los documentos relacionados en el archivo Gestión Contractual GC_ del periodo objeto de cobro.

Evidencias en folios:

- ANEXOS.
- https://drive.google.com/drive/folders/1yEDbPoaNjp_LMjKuMufA0nyHAufxe9iC?usp=sharing

Cordialmente,

JESUS ARBEY BENAVIDES GUEVARA
Contratista
C.C. No. 1.085.260.713 DE PASTO (NARIÑO)

Recibí a satisfacción:

JOHN JAIRO CASTRO MALDONADO
SUPERVISOR(A) CONTRATO No. CO1.PCCNTR.3367636 DE 2022
COORDINADOR ACADÉMICO
C.C. No. 88.265.552



ANEXOS



ANEXO 1. Medios de divulgación: Notificación de cambio de autor de correspondencia en *Workshop on Modeling and Simulation for Science and Engineering*

Solicitud para notificaciones de publicación Abstract Id: 20

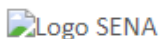
Editor
Ely Dannier V. Niño

Le escribo para solicitarle que las notificaciones como autor correspondiente sean enviadas al correo paola.buitrago.betancur@gmail.com, toda vez que mi cuenta de correo pabuitragoc@sena.edu.co expira esta semana ya que he terminado contrato como investigadora de la institución.

Envío copia del correo al grupo de investigadores co-autores del artículo con Abstract Id: 20.

Saludos cordiales,

--



Paola Andrea Buitrago Cadavid

Investigador Junior (IJ)

https://scienti.minciencias.gov.co/cvlac/visualizador/generarCurriculoCv.do?cod_rh=0001487730

Ingeniera Electrónica

Universidad de Antioquia

Magister en Modelación y Ciencia Computacional

Universidad de Medellín

Investigadora Sennova - SENA

Centro De Servicios y Gestión Empresarial - Complejo Central - SENA

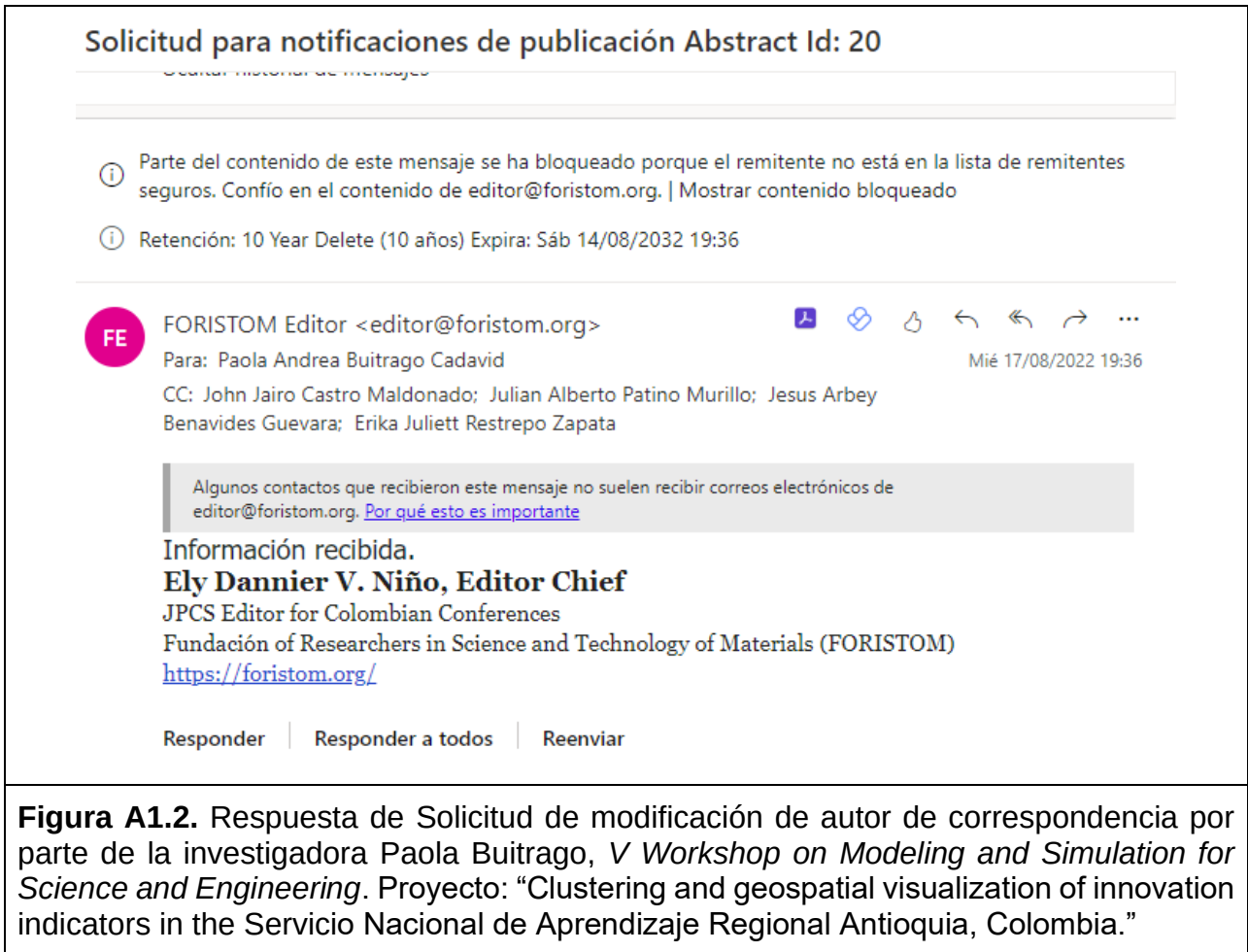
pabuitragoc@sena.edu.co

paola.buitrago.betancur@gmail.com

+573113671040

Calle 51 Nro 57-70 - Complejo Central

Figura A1.1. Solicitud de modificación de autor de correspondencia por parte de la investigadora Paola Buitrago, V *Workshop on Modeling and Simulation for Science and Engineering*. Proyecto: “Clustering and geospatial visualization of innovation indicators in the Servicio Nacional de Aprendizaje Regional Antioquia, Colombia.”



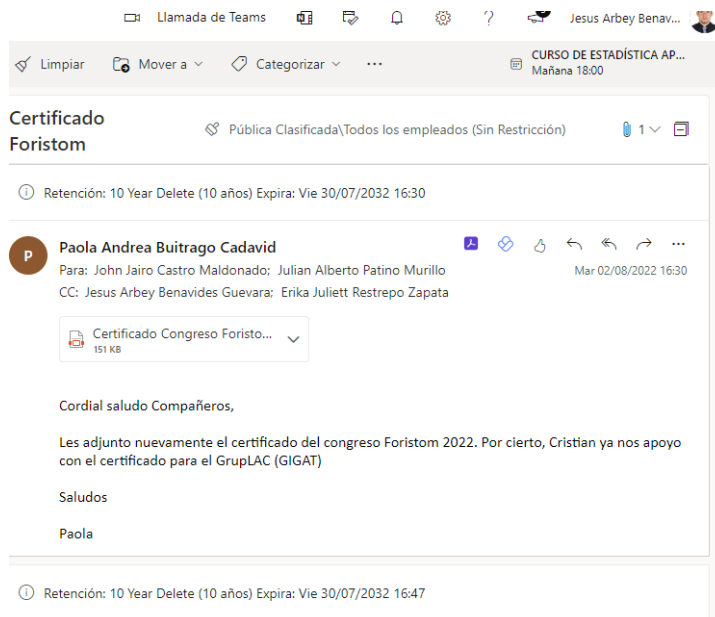


Figura A1.3. Recepción de certificación de participación *V Workshop on Modeling and Simulation for Science and Engineering*. Proyecto: “Clustering and geospatial visualization of innovation indicators in the Servicio Nacional de Aprendizaje Regional Antioquia, Colombia.”

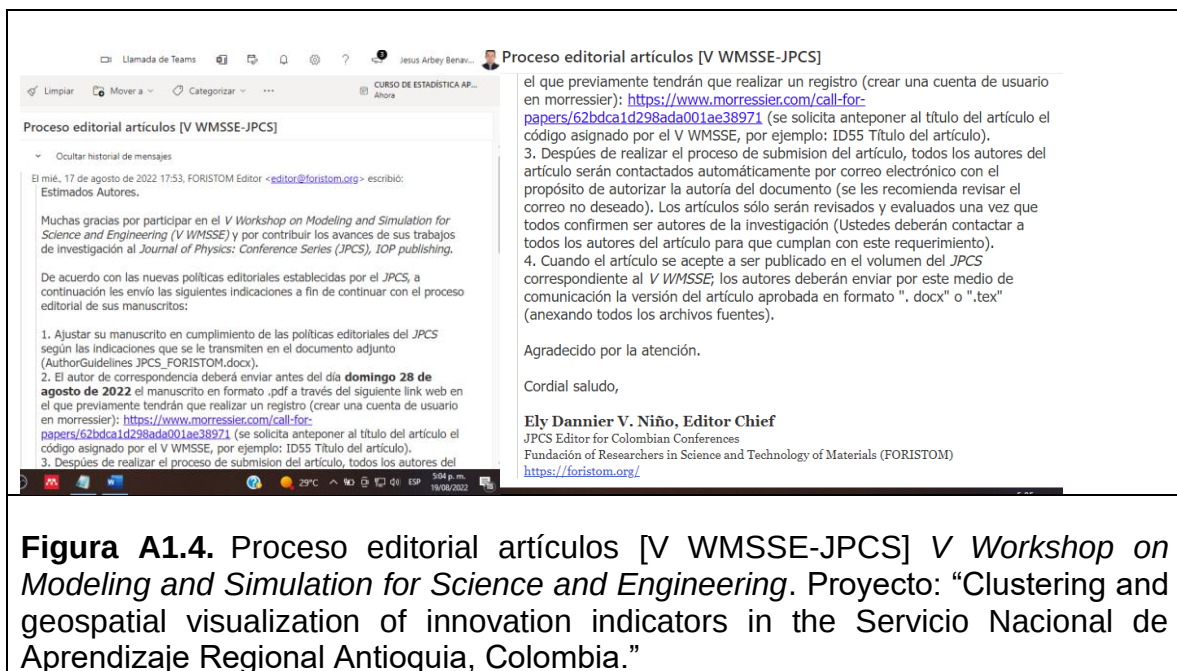


Figura A1.4. Proceso editorial artículos [V WMSSE-JPCS] *V Workshop on Modeling and Simulation for Science and Engineering*. Proyecto: “Clustering and geospatial visualization of innovation indicators in the Servicio Nacional de Aprendizaje Regional Antioquia, Colombia.”

ANEXOS 2. Evidencia de informe de investigación

**Modelo de evaluación y valoración de la Gestión de CTel para los Centros SENA Regional
Antioquia aplicando analítica de datos - Proyecto SGPS 9748 - 2021**

Informe técnico de investigación

Presentado por

Grupo de Investigación en Gerencia y Aplicación de la Ciencia y Tecnología - GIGAT

Paola Buitrago Cadavid

Jesus Arbey Benavides Guevara

John Jairo Castro Maldonado

Julián Alberto Patiño Murillo

Erika Juliett Restrepo Zapata

Jhon Alexis Ferla Pérez



Centro de Servicios y Gestión

Empresarial

Sena, regional Antioquia

Medellín

2022



1 ¡Error! Marcador no definido.2 ¡Error! Marcador no definido.3 ¡Error! Marcador no
definido.4 ¡Error! Marcador no definido.5 ¡Error! Marcador no definido.6 ¡Error! Marcador
no definido.7 ¡Error! Marcador no definido.7.1 ¡Error! Marcador no definido.7.2
¡Error! Marcador no definido.8 ¡Error! Marcador no definido.9 ¡Error! Marcador no
definido.9.1 ¡Error! Marcador no definido.9.2 ¡Error! Marcador no definido.10
¡Error! Marcador no definido.10.1 ¡Error! Marcador no definido.10.2 ¡Error!
Marcador no definido.10.2.1 3010.3 3110.3.1 ¡Error! Marcador no definido.10.3.2
¡Error! Marcador no definido.10.4 ¡Error! Marcador no definido.10.5 ¡Error!
Marcador no definido.10.5.1 ¡Error! Marcador no definido.10.6 ¡Error! Marcador no
definido.10.6.1¡Error! Marcador no definido.10.6.2 ¡Error! Marcador no definido.10.6.3
¡Error! Marcador no definido.10.7 ¡Error! Marcador no definido.10.7.1 ¡Error!
Marcador no definido.10.7.2 ¡Error! Marcador no definido.10.7.3 ¡Error! Marcador no
definido.11 ¡Error! Marcador no definido.12 ¡Error! Marcador no definido.12.1
¡Error! Marcador no definido.12.1.1 ¡Error! Marcador no definido.12.2 ¡Error!
Marcador no definido.12.3 ¡Error! Marcador no definido.13 ¡Error! Marcador no
definido.13.1 ¡Error! Marcador no definido.13.1.1 ¡Error! Marcador no definido.13.2
¡Error! Marcador no definido.13.3 ¡Error! Marcador no definido.13.3.1 ¡Error!
Marcador no definido.13.3.2 ¡Error! Marcador no definido.13.3.3 ¡Error! Marcador no
definido.13.3.4¡Error! Marcador no definido.13.3.5 ¡Error! Marcador no definido.13.3.6
¡Error! Marcador no definido.13.3.7 ¡Error! Marcador no definido.13.4 ¡Error!
Marcador no definido.13.4.1 ¡Error! Marcador no definido.13.5 ¡Error! Marcador no
definido.13.5.1¡Error! Marcador no definido.13.5.2 ¡Error! Marcador no definido.13.5.3
¡Error! Marcador no definido.13.5.4 ¡Error! Marcador no definido.13.5.4.1 ¡Error!
Marcador no definido.13.5.4.2 ¡Error! Marcador no definido.13.5.4.3 ¡Error! Marcador no
definido.13.5.4.4 ¡Error! Marcador no definido.13.5.4.5 ¡Error! Marcador no
definido.13.5.4.6 ¡Error! Marcador no definido.13.5.5 ¡Error! Marcador no
definido.13.5.5.1 ¡Error! Marcador no definido.14 ¡Error! Marcador no
definido.14.1 ¡Error! Marcador no definido.14.2 ¡Error! Marcador no definido.15
¡Error! Marcador no definido.15.1 ¡Error! Marcador no definido.15.2 ¡Error!
Marcador no definido.15.3 ¡Error! Marcador no definido.16 ¡Error! Marcador no
definido.17 ¡Error! Marcador no definido.18 ¡Error! Marcador no definido.18.1.1
¡Error! Marcador no definido.18.1.1.1 ¡Error! Marcador no definido.



1 DATOS DEL PROYECTO

Nombre del Proyecto:

Modelo de evaluación y valoración de la Gestión de CTel para los Centros SENA Regional Antioquia aplicando analítica de datos - Proyecto SGPS 9748 - 2021

Autores

Paola Buitrago Cadavid, Ingeniería Electrónica, Magister en Modelación y Ciencia Computacional, Investigadora SENNOVA. Email: pabuitragoc@sena.edu.co

Jesus Arbey Benavides Guevara, Físico, Magíster en Física. Doctorado(candidato) en Física, Especialista en ciencias de datos(Estudiente). Investigador SENNOVA. Email: jabenavides@sena.edu.co

Erika Juliett Restrepo Zapata, Administradora de Empresas, Tecnóloga en Formulación de Proyectos. Investigador SENNOVA, Email: erestrepoz@sena.edu.co

Julián Alberto Patiño, Ingeniero de Control, Magíster en Ingeniería Automatización Industrial, Doctor en Ingeniería Automática, Email: japatino@sena.edu.co

John Jairo Castro Maldonado, Ingeniero Electromecánico, Especialización en Práctica Pedagógica Universitaria, Especialización en Seguridad Informática, Maestría en Ciencia y Tecnología de Materiales, Maestría en Controles Industriales, Email: jcastrom@sena.edu.co

Nombre de semilleros Involucrados

Semillero de Investigación en Proyectos e Ingeniería (SIPI)

Semillero de Investigación en Móviles,
y Control (MERLIN)



Educación, Robótica, Automatización

Línea de Investigación

Innovación

Aliados

Unidad de Servicios Tecnológicos del Centro de Servicios y Gestión Empresarial

Tecnoparque

Nombre del grupo de investigación

GIGAT (Grupo de Investigación en Gerencia y Aplicación de la Ciencia y la Tecnología)

Fuente de financiación

Convocatoria SENNOVA 2021 – Resolución 0012020

Duración del proyecto

7 meses

2 LISTA DE ANEXOS



Anexo 1. Implementación de código en R statistics para generación de base de datos	90
Anexo 2. Implementación de código en R statistics del método de clustering	91
Anexo 3. Implementación de código en R statistics del método K-means	92
Anexo 4. Implementación de código en R statistics del método PCA	94
Anexo 5. Implementación de código en R statistics del método Random Forest	96
Anexo 6. Implementación de código en R statistics del método K-NN	98
Anexo 7. Implementación de código en Python del método de clustering	90
Anexo 8. Implementación de código en Python del método K-means	91
Anexo 9. Implementación de código en Python del método PCA	92
Anexo 10. Implementación de código en Python de una red neuronal tipo perceptrón	94
Anexo 11. Implementación de código en Python del método de regresión logística	96
Anexo 12. Implementación de código en Python del método SVM	98

3 LISTA DE TABLAS

Tabla 1. Variables identificadas que intervienen en la medición de indicadores	19
--	----

**Tabla 2. Variables identificadas que
procesos**

19



intervienen en el módulo de

Tabla 3. Variables de salida

20

Tabla 4. Identificación de tablas y campos para la base de datos

25

Tabla 5. Identificación de campos para la tabla de datos Procesos

27

Tabla 6. Identificación de campos para la tabla de datos Productividad

27

4 LISTADO DE FIGURAS

Figura 1. Diagrama representativo de clúster

28

Figura 2. Iteración 2. Elaboración con base en los resultados de VOSviewer

22

Figura 3. Clasificación Actividades de Ciencia y Tecnología OCDE.

24

Figura 4. Tipos de análisis. Fuente:(Riahi &  Riahi, 2018)	40
Figura 5. Etapas del proceso de investigación	49
Figura 6. Etapas del proyecto	51
Figura 7. Ciclo de vida del proyecto. Elaboración propia	54
Figura 8. Modelo Conceptual de la evaluación de la gestión de ciencia, tecnología e innovación	58
Figura 9. Componentes del Sistema de Información de Proyectos de CTel	66
Figura 10. Dependencia de las variables de los proyectos de CTel	68
Figura 11. Oportunidades de captura de la información del Sistema de Proyectos de CTel	69
Figura 12. Flujo de proceso de la metodología CRISP® Tomado de (García Vélez, 2018)	70



- Acervo de recursos humanos en ciencia y tecnología, ARHCyT (Manual de Canberra, OECD) (C. Manual, 1995) Personas que cumplen con una de las siguientes condiciones: a) haber completado exitosamente educación de tercer nivel en un campo de la ciencia y la tecnología, b) personas no tan formalmente calificadas como lo anterior, pero que están empleadas en una ocupación de ciencia y tecnología que generalmente requiere de habilidades y conocimientos de tercer nivel.
- Activos (Miller y Meiners) (Miller et al., 1990) Cualquier bien que proporciona un flujo de utilidad y/o ingreso en el futuro. Los activos incluyen acciones, bonos, automóviles, estéreos, un grado universitario, etc.
- Actividades de innovación tecnológica (Manual de Frascati, OECD) (F. Manual, 2002) Conjunto de etapas científicas, tecnológicas, organizativas, financieras y comerciales, incluyendo las inversiones en nuevos conocimientos, que llevan o que intentan llevar a la implementación de productos y de procesos nuevos o mejorados. La I+D no es más que una de estas actividades y puede ser llevada a cabo en diferentes fases del proceso de innovación, siendo utilizada no sólo como la fuente de ideas creadoras sino también para resolver los problemas que pueden surgir en cualquier fase hasta su culminación.
- Ciencia (Real Academia Española) (GONZALEZ et al., 2013) Conjunto de conocimientos obtenidos mediante la observación y el razonamiento, sistemáticamente estructurados y de los que se deducen principios y leyes generales. En otra acepción, ciencia es el conjunto de conocimientos relativos a las ciencias exactas, fisicoquímicas y naturales.
- Costes salariales del personal de I+D (Manual de Frascati, OECD) (F. Manual, 2002) Estos costes comprenden los salarios y remuneraciones anuales y todos los gastos complementarios de personal o remuneraciones diversas, tales como primas, vacaciones pagadas, contribuciones a fondos de pensiones y otros pagos a la Seguridad Social, impuestos salariales, etc. Los costes salariales de las personas que prestan servicios indirectos y que no se tienen en cuenta en los datos de personal (tales como el personal de seguridad y de mantenimiento o el personal de bibliotecas centrales, de servicios informáticos y de las oficinas de dirección) deben excluirse y contabilizarse en el apartado de otros gastos corrientes.



- Desarrollo experimental (Manual de Frascati, OECD) (F. Manual, 2002)
El desarrollo experimental consiste en trabajos sistemáticos fundamentados en los conocimientos existentes obtenidos por la investigación o la experiencia práctica, que se dirigen a la fabricación de nuevos materiales, productos o dispositivos, a establecer nuevos procedimientos, sistemas y servicios, o a mejorar considerablemente los que ya existen. En las ciencias sociales, el desarrollo experimental puede definirse como el proceso que permite convertir los conocimientos adquiridos a través de la investigación en programas operativos, incluidos los proyectos de demostración que se llevan a cabo con fines de ensayo y evaluación. Esta categoría tiene escasa o nula significación en el caso de las humanidades.
- Investigación y desarrollo experimental, I+D (Manual de Frascati, OECD) (F. Manual, 2002) La investigación y el desarrollo experimental comprenden el trabajo creativo llevado a cabo de forma sistemática para incrementar el volumen de conocimientos, incluido el conocimiento del hombre, la cultura y la sociedad, y el uso de esos conocimientos para crear nuevas aplicaciones. Comprende investigación básica, aplicada y el desarrollo experimental.
- Investigadores (Manual de Frascati, OECD) (F. Manual, 2002) Profesionales que se dedican a la concepción o creación de nuevos conocimientos, productos, procesos, métodos y sistemas y también a la gestión de los proyectos respectivos. Los investigadores se encuentran incluidos en el Grupo Principal 2 de la clasificación ISCO-88, Profesionales. Por convención, también se incluyen en estas categorías los miembros de las fuerzas armadas con formación equivalente y que realicen actividades de I+D. Igualmente, están incluidos en esta categoría los gestores y administradores que desarrollan actividades de planificación y gestión de los aspectos científicos y técnicos del trabajo de los investigadores. Normalmente, tienen una categoría igual o superior a la de las personas empleadas directamente como investigadores, tratándose a menudo de investigadores veteranos o a tiempo parcial. Los títulos profesionales pueden variar de una institución a otra, de un sector a otro y de un país a otro. Los estudiantes de postgrado a nivel de doctorado que participan en tareas de I+D deben considerarse como investigadores. Habitualmente suelen poseer titulaciones universitarias básicas (CINE/ISCED nivel 5A) y realizan trabajos de investigación a la vez que preparan su doctorado (CINE/ISCED nivel 6).



- Personal de I+D (Manual de Frascati, OECD) (F. Manual, 2002)
Todo el personal empleado directamente en I+D, así como las personas que proporcionan servicios directamente relacionados con actividades de I+D, como los directores, administradores y personal de oficina. Las personas que proporcionan servicios indirectos, como el personal de los comedores y de seguridad, deben ser excluidas, aunque sus sueldos y salarios se contabilicen como gastos generales para la medida del gasto de I+D.

6 RESUMEN DEL PROYECTO

Las actividades de Ciencia, Tecnología e Innovación están relacionadas con la ejecución de actuaciones que impliquen investigación, desarrollo experimental, apoyo a la educación y formación y prestación de servicios científicos y tecnológicos, actividades de innovación, servicios científicos y tecnológicos, administración y otras actividades de gestión, que son realizadas por entidades públicas y entidades privadas para contribuir a la generación y apropiación del conocimiento.

En este sentido, debido al incremento exponencial que ha tenido la generación de nuevos productos de conocimiento a nivel educativo y con el fin de responder a una necesidad evidente de contribuir al desarrollo del sector productivo en Colombia, a través del Sistema de Investigación SENNOVA del SENA Se potencian tanto los recursos humanos como los tecnológicos con los que cuenta la entidad para contribuir al crecimiento económico y social de la región. Teniendo en cuenta lo anterior, en la región de Antioquia del SENA se trabaja en el desarrollo de la visualización de un modelo de analítica de datos para la medición de indicadores de innovación, donde se identifican gráficamente las diferentes categorías de actividades de Ciencia, Tecnología e Innovación que vienen trabajando en los últimos años en la entidad, con el propósito de facilitar la medición y análisis de los resultados obtenidos desde la perspectiva tanto de la estadística descriptiva como de la ciencia de datos. Entre los indicadores utilizados se encuentran los procesos de gestión y productividad de los proyectos ejecutados en la entidad en un periodo de tiempo, a los cuales se les aplicaron técnicas como K-means grouping y el Análisis de Componentes Principales (PCA), que permitieron agrupar los indicadores seleccionados. Se utilizó Python Folium como una herramienta de visualización que permite generar comparaciones de manera gráfica entre los diferentes centros del SENA de la Regional de Antioquia en un período de tiempo determinado. Posteriormente, se desarrolló una plataforma API que permite visualizar la

estadística descriptiva de los datos  sintéticos, así como los resultados analíticos representados en mapas y valores predictivos usando KNN como modelo predictivo en dimensiones de productividad y procesos. Se observó que, a través de las técnicas utilizadas, existe una mayor facilidad de interpretación, comparación y análisis de resultados estadísticos y análisis de datos en la medición de indicadores de actividades de Ciencia, Tecnología e Innovación en comparación con las herramientas de visualización de información utilizadas convencionalmente.

Palabras clave: Clustering, Folium, KPIs, K-means, PCA, Random forest, K-NN, ACTI, SENA

7 OBJETIVOS

7.1 Objetivo general:

Proponer un Modelo de Gestión de Indicadores de Ciencia, Tecnología e Innovación para los Centros SENA Regional Antioquia a partir de técnicas de analítica de datos.

7.2 Objetivos específicos:

1. Identificar metodologías de analítica de datos asociada al establecimiento de indicadores y modelos de CTel que puedan aplicarse a los centros de formación del SENA.
2. Determinar un modelo de evaluación y valoración de las actividades de CTel innovación que más se ajuste al comportamiento de estas sinergias en los centros de formación de la regional Antioquia.
3. Validar el modelo por medio de datos sintéticos que permitan generar una metodología para la analítica de datos.
4. Crear modelos de analítica de datos a partir de aprendizaje de máquina supervisado y no supervisado para la determinación de indicadores de CTel.



5. Generar mapas geoespaciales de alertas de alto, medio y bajo de las dimensiones de indicadores establecidos.

6. Implementar modelos predictivos a partir de aprendizaje supervisado para la determinación de estados de alerta de alto medio y bajo de las dimensiones de indicadores, a lo largo del tiempo.

8 INTRODUCCIÓN

Comenzamos definiendo las actividades de Ciencia, Tecnología e innovación (ACTI) como aquellas que se relacionan con producir, promover, difundir y aplicar conocimientos científicos en los campos que involucran la ciencia y la tecnología [1]. En los últimos años ha surgido la necesidad de proporcionar de forma sistemática información sobre la recolección, análisis, publicación y normalización de los datos asociados a estas actividades [2], de allí que se empiecen a diseñar indicadores que aporten a la toma de decisiones estratégicas sobre su gestión.

A nivel internacional la Organización para la Cooperación y el Desarrollo Económico (OCDE) ha elaborado diferentes manuales con lineamientos que pueden ser utilizados para la definición de indicadores que permitan la medición de ACTI [3], como el Manual de Frascati y el Manual de Oslo, que compuestos por una serie de propuestas para la recolección e interpretación de datos sobre innovación tecnológica, por otro lado, el Manual de Bogotá que contiene pautas para construir indicadores de innovación tecnológica en Latinoamérica y el Caribe, entre otros [4].

Esta investigación tiene como objetivo proponer un método de medición de Indicadores de Ciencia, Tecnología e Innovación mediante analítica de datos, la fundamentación teórica de la propuesta se basó en técnicas estadísticas no paramétricas supervisadas, entre las que encontramos métodos como clustering basado en algoritmos k-means y el uso de análisis de componentes principales (PCA). Para ello se simuló una base de datos sintéticos de 500 proyectos distintos con el propósito de realizar un análisis exploratorio que permitieron caracterizar los 16 centros del SENA de la regional Antioquia, de los cuales posteriormente se crean mapas utilizando como herramienta de base Folium [5].

Uno de los indicadores de mayor  preponderancia es el de gasto y/o inversión en actividades de investigación y desarrollo, I+D, ya que permite hacer seguimiento a las actividades de ciencia y tecnología – ACTI mediante el apalancamiento de proyectos de I+D. La mejor muestra de ello es la Organización para la Cooperación y el Desarrollo Económico, OCDE, uno de los organismos más reconocidos por el trabajo de estandarización conceptual para la medición de la Ciencia y la Tecnología. Según el Observatorio Colombiano de Ciencia y Tecnología, una de las variables más importantes a tener en cuenta es la cuantificación de los recursos humanos, toda vez que estos son utilizados para la identificación de las actividades de investigación y desarrollo I+D, así como la clasificación e identificación del gasto, [6].

Mediante el manejo y gestión de los recursos, y el seguimiento de sus actividades científicas; es posible medir los indicadores de gestión y productividad al interior de una organización y de esta forma detectar sus fortalezas y debilidades, promoviendo incentivos que permitan mejorar sectores específicos que se consideren estratégicos.

Por ello, se han llevado a cabo diferentes esfuerzos en el SENA para la medición de indicadores de ciencia, tecnología e innovación. Lo que ha permitido generar un modelo de analítica de datos basado en clustering mediante k-means y análisis de componentes principales, y posteriormente la implementación de dos modelos de predicción basados en random forest y K-NN.



9 ANTECEDENTES

Muchos autores han identificado diferentes indicadores que ajustan de acuerdo con el comportamiento productivo de la organización, no obstante, algunos de estos son comunes entre Sí y que son importantes tenerlos en cuenta debido a su impacto y resultados en los procesos de CTel de las organizaciones, por ejemplo Mario Albornoz en su documento de indicadores y política científica y tecnológica expone un serie de preguntas que nos invita a determinar a partir de ellas los indicadores relevantes en la identificación de capacidades de CTel como son: ¿Cómo aprovechar los conocimientos disponibles? ¿Cómo formar e incorporar los recursos humanos necesarios?, ¿Qué instituciones e instrumentos son los más indicados?, entre otras preguntas que formula (Albornoz, 2012).

Por otro lado, Jorge Robledo Velásquez en su libro Introducción a la gestión de la tecnología y la innovación empresarial, relaciona con gran pertinencia las áreas temáticas desde las cuales se deben construir los indicadores de innovación y que se han establecido en los últimos años desde los manuales de Oslo y Frascati hasta la políticas de cada uno de los países, además, se pueden generar algunos indicadores globales pertinentes a países con características sociodemográficas similares como la calidad del entorno de la organización, las capacidades de adquisición de conocimiento y los resultados logrados orientados a patentes y contenido tecnológico para exportar (Robledo Velásquez, 2020).

En este sentido, en México se procedió a realizar un estudio donde se pretendía identificar las condiciones de las capacidades CTel de los clústeres y con este análisis verificar cuál de ellos estaban todavía activos y cuales ya no estaban en funcionamiento, para este proyecto se tomó como base los indicadores establecidos en la CEPAL (2007) tomando información de diferentes fuentes de dicho país (Pérez Hernández et al., 2017).

Específicamente en el contexto educativo también se han tenido desarrollos importantes en el establecimiento de indicadores de CTel, por ejemplo Jackeline Serrano en su artículo variables para la medición de las capacidades de innovación tecnológica en instituciones universitarias, desde un concepto de organización académica, identifica el contexto de acción de la universidad en un estadio de triple hélice, el cual, le demanda algunas condiciones para el establecimiento de estos indicadores, en este sentido, plantea inicialmente un modelo de comportamiento organizacional de la universidad alineado a las dimensiones organizacionales de (Nadler & Tushman, 1980) para de esta forma identificar

las variables definitivas, dentro de las  cuales se pueden mencionar las capacidades en Investigación y desarrollo (I+D), Capacidad de producción, Capacidad de planeación, Capacidad de relacionamiento organizacional, capacidad de aprendizaje organizacional, capacidad de gestión de recursos y capacidad de mercadeo (Serrano García et al., 2017).

Igualmente, dentro del SENA ya se han realizado algunos acercamientos para la identificación y valoración de indicadores de CTel como el trabajo titulado procesos de I+D+i en el Centro de Servicios y Gestión Empresarial del SENA 2015 – 2017 de autoría de John Jairo Castro, el cual, muestra los avances en materia de actividades con resultados que se pueden categorizar en productos de CTel y a su vez, cómo se realizaron diferentes actividades para poder proporcionar espacios para su consolidación y desarrollo eficaz tanto con aprendices como con los instructores de la institución, asimismo, muestra el comportamiento de apropiación de una cultura investigativa dentro del entorno académico (Jairo et al., 2018).

Para concluir, el establecimiento adecuado de los indicadores de evaluación y medición de indicadores de CTel ha estado siempre presente en los estudios tanto a nivel administrativo como a nivel de ingeniería, con el fin, de definir la adecuada ejecución de los recursos de las organizaciones en estos procesos, es por esto que la implementación de modelos que permitan inferir el comportamiento de estos indicadores también ha sido menester continuo de los investigadores en las áreas de la gestión de la CTel debido a la importancia que tiene encontrar puntos de equilibrio donde se pueda establecer la asignación certera de recursos eficientes y eficaces para estas actividades sin sacrificar la calidad de los productos ni exceder asignaciones presupuestales y más cuando los recursos son del estado como es el caso del SENA, por tanto, es necesario implementar modelos de medición de indicadores de CTel usando tecnologías emergentes como la analítica de datos, con el fin, de dar herramientas objetivas a la institución al momentos evaluar y valorar las actividades de CTel desarrolladas en los diferentes centros de formación en todo el departamento de Antioquia.

Nuestra investigación se basa en implementar un modelo de analítica de datos que permita medir los indicadores de gestión y de productividad en los 16 centros de la regional Antioquia del SENA, mediante una base de datos generada pseudoaleatoriamente en R statistics. Posteriormente se generó una tabla denominada procesos, teniendo en cuenta las tablas Principal_Proyectos y Principal_Productos, donde se implementó un código en R statistics para obtener los siguientes campos: Centro_Año (representado por 208 etiquetas de 16 centros en 13 años), Q_Proyectos (que representa la cantidad de proyectos por cada etiqueta Centro_Año), QNA_5007 (que representa la cantidad de profesionales por Centro_Año), QNA_5009 (que representa la cantidad de magister por Centro_Año), QNA_5010 (que representa la cantidad de doctores por Centro_Año) y TV_Proyectos (que

representa el total del valor de cada  proyecto por Centro_Año). Por último se generó una tabla denominada Productos, teniendo en cuenta las tablas de datos Principal_Proyectos, Principal_Productos y Principal_Talento_Humano, Donde se implementó un código en R statistics para obtener los siguientes campos: Centro_Año (representado por 208 etiquetas de 16 centros en 13 años), Cantidad_Proyectos (que representa la cantidad de proyectos por Centro_Año), Fecha_inicio_Proyecto (que representa la fecha en la que inicia cada proyecto por Centro_Año), Cantidad_Productos (representada por la cantidad total de productos de CTel por Centro_Año), y las cantidad para cada una de las seis categorías de productos de CTel por Centro_Año. Se aplicó el método de clustering, y K-means mediante el uso de análisis de componentes principales a las tablas de datos obtenidas, para poder medir los indicadores de gestión y procesos, identificándose en tres niveles, alta gestión, media gestión y baja gestión para la tabla de procesos, y alta productividad, mediana productividad y baja productividad para la tabla de productividad. Y por último se aplicó el método supervisado K-NN y Random forest, donde se identificó si las tablas de datos procesos y productividad estaban bien clasificadas mediante la evaluación de las siguientes métricas: sensibilidad, F1 (precisión y sensibilidad), exactitud y especificidad.



10.1 VARIABLES IDENTIFICADAS

Como variables de entrada que intervienen en actividades de ciencia, tecnología e innovación se identificaron las siguientes:

Tabla 1. Variables identificadas que intervienen en la medición de indicadores

DIMENSIÓN	CATEGORIA	INDICADOR
Insumos para actividades de Ciencia, Tecnología e Innovación (ACTI) (Entradas)	Recursos económicos	Gasto/inversión en ACTI
	Recursos físicos y tecnológicos	Inversión de equipos de para ACTI
	Talento humano	Número de personas dedicadas a ACTI
	Recursos estructurales y relacionales	Número de unidades organizativas (grupos de investigación - semilleros)
		Número de programas impactados de educación superior ofertados (programas del SENA por ACTI)
		Número de convenios de colaboración para ACTI

Dentro de nuestro modelo conceptual implementado se tienen en cuenta todas las variables que permitieron obtener las variables de salida para medir indicadores de gestión y de productividad.

Tabla 2. Variables identificadas que intervienen en el módulo de procesos

DIMENSIÓN	CATEGORIA	INDICADOR
Proceso	Proyectos	Número de proyectos



	Talento humano	proporción de gasto en ACTI por personas dedicadas a ACTI
	Recursos económicos	porcentaje de gasto en ACTI en relación con el presupuesto total
		porcentaje de gasto en investigación aplicada en relación con el gasto total por ACTI
	Relaciones con el entorno	Eventos de movilidad para ACTI
	Formación científica y tecnológica	Personal de talento humano calificado que interviene en ACTI

Tabla 3. Variables de salida

DIMENSIÓN	CATEGORÍA	INDICADOR
Resultados (Salidas)	Relaciones con el medio ambiente	Número de eventos de movilidad para ACTI
	Productos resultados de las actividades de CTI	Número de productos de I+D (productos de I+D por tipología Colciencias)
	Transferencia de conocimiento	Número de asistentes a eventos de CTI organizados por la universidad (por campo de conocimiento según la clasificación de la OCDE)
	Formación científica y tecnológica	Personal de talento humano calificado que interviene en ACTI

10.2 ESTRATEGIA SENNOVA

En el Servicio Nacional de Aprendizaje – SENA, además de las actividades administrativas y de formación para el trabajo, se realizan ACTI como parte de la Formación Profesional

Integral que se ofrece en la entidad, Investigación y Desarrollo Tecnológico

(Ramírez Ariza, 2014).



algunas de ellas están enfocadas en la como la conformación de nuevos grupos y semilleros de investigación, además de consolidar y fortalecer el trabajo al interior de ellos vinculando a sus actividades a los aprendices con el objetivo de que desarrollen sus habilidades y conocimientos investigativos mediante la ejecución de proyectos formativos que cuenten con el procedimiento y rigurosidad requeridos en la investigación aplicada

En este sentido, en el ACUERDO 16 DE 2012 se regula el programa de Investigación, Desarrollo Tecnológico e Innovación, además de justificar que con dicho programa se busca contribuir al desarrollo tecnológico, la competitividad de las empresas del país, así como al fomento de la gestión en investigación, innovación y el conocimiento, ejecutando acciones para incentivar el desarrollo de proyectos y/o estrategias que generen resultados cuantificables y verificables en el sector productivo, la formación profesional integral y la formación para el trabajo, involucrando el sector productivo e instituciones generadoras de conocimiento y al SENA.

10.2.1 Listado de las actividades de Ciencia Tecnología e Innovación

En la política de investigación del SENA (Ramírez Ariza, 2014) se menciona una serie de ACTI que buscan fortalecer este proceso entre las que se encuentran:

- Fomentar la conformación de nuevos grupos y semilleros de investigación en el SENA.
- Consolidar y fortalecer el trabajo de los grupos y semilleros de investigación registrados en los centros de formación.
- Facilitar el registro y validación de los grupos de investigación dentro del SENA ante MINCIENCIAS.
- Fomentar el desarrollo tecnológico a partir de la investigación aplicada realizada.
- Aprovechar el conocimiento generado por los grupos de investigación para retroalimentación y rediseño curricular de los programas de formación.
- Fomentar la participación de aprendices en los grupos y semilleros de investigación para incrementar sus capacidades investigativas y asegurar su formación integral.

- Fomentar las alianzas entre  investigadores del Sena y universidades o centros de investigación.
- Asegurar el acceso a la información necesaria para el desarrollo de proyectos de investigación y vigilancia tecnológica en todos los centros de formación del Sena.
- Propiciar la vinculación de los productos y procesos desarrollados por investigadores o grupos de investigación en el Sena con el sector productivo.
- Apoyar el proceso de acreditación de calidad de la Formación Profesional a través del apoyo a los grupos de investigación aplicada de los Centros de Formación Profesional.
- Propiciar la difusión de productos académicos desarrollados en procesos de investigación para el fortalecimiento de los programas de formación.
- Estandarizar los criterios de reporte y publicación de los grupos de investigación.
- Fortalecer las políticas de propiedad intelectual de los productos y procesos desarrollados por investigadores y grupos de investigación asociados al Sena.
- Fomentar la comunicación y colaboración entre grupos de investigación en el Sena.
- Reforzar la cultura escrita en los centros de investigación para que los saberes empíricos lleven a nuevos conocimientos y no desaparezcan por falta de documentación adecuada.
- Difundir ampliamente los resultados de las investigaciones desarrolladas en el Sena para comunicar a empresarios, gobierno y comunidad científica en general para acelerar la innovación y evitar la repetición de trabajos.

10.3 GESTIÓN

La gestión es un término integral en función genérica que está relacionado con la planeación, la organización, la dirección y el control de los recursos en una organización, con el fin de lograr el mayor beneficio para la misma (Aktouf, 1989; Henao Castrillón & Zapata Giraldo, 2018). Por otro lado, empleando las palabras de Henao & Zapata (2018) cuando mencionan Schumpeter, esta juega un papel importante en las organizaciones al coordinar e integrar sus procesos, recursos y capacidades en beneficio de su desempeño, la creación de nuevos productos y el desarrollo de competencias adicionales.

10.3.1 INDICADOR DE GESTIÓN



Un indicador de gestión es la expresión cuantitativa del comportamiento y desempeño de un proceso, cuya magnitud, al ser comparada con algún nivel de referencia, puede estar señalando una desviación sobre la cual se toman acciones correctivas o preventivas según el caso.

Para trabajar con los indicadores debe establecerse todo un sistema que vaya desde la correcta comprensión del hecho o de las características hasta la de toma de decisiones acertadas para mantener, mejorar e innovar el proceso del cual dan cuenta.

Los indicadores de gestión deben cumplir con unos requisitos y elementos para poder apoyar la gestión y conseguir el objetivo propuesto. Estas características son:

Simplicidad. Puede definirse como la capacidad para definir el evento que se pretende medir, de manera poco costosa en tiempo y recurso.

Adecuación. Entendida como la facilidad de la medida para describir por completo el fenómeno o efecto. Debe reflejar la magnitud del hecho analizado y mostrar la desviación real del nivel deseado.

Validez en el tiempo. Puede definirse como la propiedad de ser permanente por un periodo deseado.

Participación de los usuarios. Es la habilidad para estar involucrados desde el diseño, y debe proporcionárseles los recursos y formación necesarios para su ejecución. Este es quizás el ingrediente fundamental para que el personal se motive en torno al cumplimiento de los indicadores.

Utilidad. Es la posibilidad del indicador para estar siempre orientado a buscar las causas que han llevado a que alcance un valor particular y mejorarlas.

Oportunidad. Entendida como la capacidad para que los datos sean recolectados a tiempo. Igualmente requiere que la información sea analizada oportunamente para poder actuar.

10.4 PRODUCTIVIDAD

La productividad es la relación cuantitativa entre lo que se produce y lo que se gasta para producir. La productividad no es más que una reducción en el desperdicio de recursos como los hombres, el material, la máquina, el tiempo, el espacio, el capital, etc. Denota la relación

entre el producto y uno o todos los



insumos asociados. En palabras de Peter Drucker, productividad significa un equilibrio entre todos los factores de producción que proporcionan el máximo rendimiento con el menor esfuerzo, es, por lo tanto, la cantidad de bienes o servicios (producto) que producen. La tierra y los materiales de construcción, máquinas, mano de obra (mano de obra), tecnología, etc. son los recursos a disposición de la organización. Por lo tanto, una mayor (mejor) productividad significa que se produce más con el mismo gasto de recursos, es decir, al mismo costo en términos de tierra, materiales, máquina, tiempo o mano de obra, o bien, se produce la misma cantidad a un menor costo en términos de tierra, materiales, tiempo de la máquina o mano de obra que se utiliza.

¿Cuáles son los factores que afectan la productividad de los procesos?

Hay varios factores que afectan la productividad de los procesos y, por lo tanto, el rendimiento de la organización. Algunos factores tienen un efecto positivo en la productividad de los procesos, mientras que otros factores tienen una influencia negativa en la productividad. Algunos de estos factores se pueden controlar mientras que los otros no se pueden controlar. Los factores importantes que afectan la productividad se enumeran a continuación.

Inversión de capital: las inversiones de capital en la tecnología, los equipos y las instalaciones influyen en la productividad del proceso. Cualquier ahorro en estos elementos de proceso puede tener un gran efecto negativo en la productividad del proceso.

Materiales de entrada: la calidad del material de entrada tiene una gran influencia en la productividad del proceso. Algunos materiales de entrada pueden tener un menor costo por unidad de peso, pero tienen un gran impacto adverso en la productividad del proceso debido a que su calidad es inferior a la calidad necesaria para el proceso.

Capacidad de proceso: la capacidad del proceso influye en su productividad. Con una gran capacidad, la productividad del proceso mejora debido a las economías de escala.

Tecnología del proceso: la tecnología seleccionada para el proceso juega un papel muy importante en la determinación de la productividad del proceso. Para varios procesos, hay tecnologías alternativas disponibles para producir el mismo producto. Estas diferentes tecnologías normalmente tienen diferentes niveles de productividad. En algunos casos, la tecnología seleccionada puede ser la misma, pero la velocidad del proceso influye en la productividad. Como ejemplo, un tren de laminación con menor velocidad máxima de laminación no puede tener la misma productividad que el tren de laminación que tiene una velocidad máxima de laminación más alta.

Mano de obra: el conocimiento, la  experiencia y el comportamiento de la fuerza de trabajo son los factores principales que ayudan a mantener la productividad del proceso. Para maximizar la productividad del proceso, tanto el conocimiento como la experiencia de la fuerza laboral son necesarios ya que estas cualidades mejoran las habilidades de la fuerza de trabajo y, por lo tanto, tienen una influencia vital en la productividad del proceso. Las habilidades de la fuerza laboral se pueden mejorar mediante capacitación y educación. La capacitación y la educación también ayudan a moldear el comportamiento de la fuerza de trabajo.

10.4.1 INDICADOR DE PRODUCTIVIDAD

Los indicadores de productividad son un valor medible (también KPI) que demuestra la eficacia con la que una empresa está logrando objetivos comerciales clave. Las organizaciones usan indicadores clave de productividad en múltiples niveles para evaluar su éxito en alcanzar objetivos.

Los indicadores de alto nivel pueden centrarse en el rendimiento general de una organización, mientras que los indicadores de bajo nivel pueden centrarse en los procesos en departamentos como ventas, marketing, recursos humanos, soporte y otros.

Tipos de indicadores: Existen principalmente dos tipos de indicadores de productividad que son:

Indicadores Cuantitativos

Son aquellos que se basan en cantidades o tiempo. Ya que se calculan con ayuda del cociente entre cantidad producida y el tiempo necesario para su realización. Igualmente, se utilizan en la relación entre inversión y número de productos consumados. Se utilizan para medir la productividad de la empresa, un equipo de trabajo o un producto o servicio.

Para nuestro proyecto se identificaron seis tipos de variables de CTel que permitieron medir el nivel de productividad en el SENA, estas se describen a continuación: Apropiación_social, Artículo_Sometido_en_revista_Indexada, Divulgación_Pública_de_la_CTel_Publicaciones_editoriales_no_especializadas_producciones_de_contenido_digital, Informe_de_investigación y Libros_resultados_de_investigación.

Indicadores Cualitativos

Miden la calidad del producto final o servicio prestado. Estos indicadores revelan las cualidades positivas o valor de lo realizado, sin limitarse a su cantidad. En nuestro caso se

establecieron niveles de productividad en CTel como alta productividad, mediana productividad y baja productividad.



11 MARCO REFERENCIAL

Para la medición de indicadores de gestión y de productividad, se implementaron los siguientes métodos no supervisados: Clustering, k-means y análisis de componentes principales, y como métodos supervisados random forest y K-NN. El proceso buscaba primero encontrar la variable de clasificación por clustering para 16 centros de la regional Antioquia durante 13 años y de esta forma identificar el nivel de gestión y productividad en tres niveles a saber: alta gestión, mediana gestión y baja gestión y alta productividad, mediana productividad, y baja productividad, por último, se implementaron dos modelos de predicción para evaluar el modelo de clasificación. A continuación, se explican los pasos que se llevaron a cabo:

I. Identificación de las variables de la base de datos a generar

La base de datos tiene en total 21 tablas, cada una de ellas representada por los siguientes campos.

Tabla 4. Identificación de tablas y campos para la base de datos

Regionales	Grupo_investigacion	Semilleros
_Codigo_Regional	_Codigo_Grupo_Investigacion	_Codigo_Semillero
_Nombre_Regional	_Nombre_Grupo_Investigacion	_Nombre_Semillero
_Codigo_Departamento	_Categoria_Grupo	_Codigo_Grupo_Investigacion
_Codigo_Municipio	_Codigo_Centro	_Codigo_Centro
	_Codigo_Regional	_Codigo_Regional



Rubro _Codigo_Rubro _Nombre_Rubro	Rol_Sennova _Codigo_Rol _Nombre_Rol	Tipo_Contratacion _Codigo_Tipo_Contrato _Nombre_Tipo_Contrato
Principal_Productos _Codigo_Producto _Nombre_Producto _Descripcion_producto _Codigo_Regional _Codigo_Centro _Codigo_Rubro _Codigo_Tipologia _Codigo_Proyecto _Nombre_Proyecto _Codigo_Grupo_Investigacion _Codigo_Semillero _Codigo_persona _Codigo_area_OCDE _Fecha_registro	Principal_Proyectos _Codigo_Proyecto _Nombre_Proyecto _Valor_Proyecto _Codigo_Rubro _Fecha_inicio_Proyecto _Fecha_Cierre_Proyecto _Codigo_Regional _Codigo_Centro _Codigo_Estado_Proyecto _Codigo_area_OCDE _Serial_equipo _Valor_Servicios_Personales _Valor_Compra_Equipos _Valor_Software _Valor_Otros_Gastos	Principal_Talento_Humano _Codigo_Persona _Codigo_Nivel_Academico _Codigo_Tipo_Contrato _Objeto_Contrato _Codigo_Rubro _Codigo_Area_OCDE _Codigo_Estado_Contratacion _Codigo_Regional _Codigo_Centro _Codigo_Proyecto _Codigo_Producto
Area_conocimiento_OCDE _Codigo_area_OCDE _Nombre_area_OCDE	Estado_Proyecto _Codigo_Estado_Proyecto _Nombre_Estado_Proyecto	Nivel_Academico _Codigo_Nivel_Academico _Nombre_Nivel_Academico
Municipio _Codigo_Municipio _Nombre_Municipio	Centros_Formacion _Codigo_Centro _Nombre_Centro	Principal_Equipos _Serial_equipo _Nombre_del_equipo



_Codigo_Departamento	_Codigo_Regional _Codigo_Municipio _Codigo_Departamento _latitud _longitud	_Valor_equipo _Fecha_de_compra _Codigo_Proyecto _Codigo_Regional _Codigo_Centro _Codigo_Rubro _Codigo_Tipo_Equipo _Codigo_area_OCDE _Codigo_persona
Tipologia_Productos _Codigo_Tipologia _Nombre_Tipologia	Tipo_Contratación _Codigo_Tipo_Contrato _Nombre_Tipo_Contrato	Tipo_equipo _Codigo_Tipo_Equipo _Nombre_Tipo_Equipo
Rol_Proyectos _Codigo_Rol_Proyecto _Nombre_Rol	Estado_Contratación _Codigo_Estado_Contratacion _Nombre_Estado_Contratacion	Departamento _Codigo_Departamento _Nombre_Departamento

II. Identificación de la tabla de procesos y productividad, con las respectivas variables.

Tabla 11.1 . Identificación de campos para la tabla de datos Procesos

Centro_Año Es una etiqueta que posee 16 centros en 13 años, para un total de 208	Q_Proyectos Representa la cantidad de proyectos por Centro_Año	QNA_5007 Representa la cantidad de profesionales por Centro_Año
--	--	---



<p>QNA_5009</p> <p>Representa la cantidad de magister por Centro_Año</p>	<p>QNA_510</p> <p>Representa la cantidad de doctores por Centro_Año</p>	<p>TV_Proyectos</p> <p>Representa el total del valor de cada proyecto por Centro_Año</p>
---	--	---

Tabla 11.2. Identificación de campos para la tabla de datos Productividad

<p>Centro_Año</p> <p>Es una etiqueta que posee 16 centros en 13 años, para un total de 208</p>	<p>Cantidad_Proyectos</p> <p>Representa la cantidad de proyectos para cada Centro_Año</p>	<p>Fecha_inicio_Proyecto</p> <p>Representa la fecha en la que inicia cada proyecto</p>
<p>Cantidad_Productos</p> <p>Representa la cantidad de productos de Ciencia, Tecnología e Innovación por Centro_Año</p>	<p>P_8020</p> <p>Representa la cantidad de productos de Apropiación_social</p>	<p>P_8025</p> <p>Representa la cantidad de productos de Artículo_Sometido_en_revista_Indexada</p>
<p>P_8019</p> <p>Representa la cantidad de productos de Divulgación_Pública_de_la_CTel_Publicaciones_editoriales_no_especializadas_producciones_de_contenido_digital</p>	<p>P_8039</p> <p>Representa la cantidad de productos de Informe_de_investigación</p>	<p>P_8016</p> <p>Representa la cantidad de productos Libros_resultados_de_investigación</p>
<p>P_8031</p> <p>Representa la cantidad de productos de</p>		



III. Implementación del método de clustering basado en K-means y análisis de componentes principales.

Definición de Clustering

Considere un conjunto de datos que contiene n instancias/observaciones cada una descrita por m atributos. El problema de agrupamiento de k -medias es dividir el n instancias en k clústeres, con los clústeres particionando las instancias $(x_1, x_2, \dots, x_n)$ en los subconjuntos $Q_{1..k}$. Los subconjuntos se pueden resumir como puntos $(C_{1..k})$ en el espacio m dimensional, comúnmente conocido como **centroides** o centros de conglomerados, cuyas coordenadas son el promedio de todos los puntos pertenecientes al subconjunto. Nos referiremos a esta colección de centroides obtenidos como un modelo de aplicación del algoritmo de agrupamiento. El agrupamiento de K-means también se puede considerar como una cuantificación vectorial con el objetivo de minimizar el error de cuantificación del vector (también conocido como distorsión) que se muestra en la ecuación (1). La solución matemática trivial es tener un clúster para cada instancia, pero normalmente $k \ll n$.

$$VQ = \frac{1}{2} \sum_{i=1}^n D(x_i, C_{f(x_i)}) \quad (1)$$

Donde D es una función distancia y $f(x)$ devuelve el índice del clúster en la instancia i

Calculamos el k óptimo por **Método de Elbow**: El método del codo se usa con el error cuadrático (sse) o la suma cuadrática del grupo (wcsc). Para nuestro caso, usaré wcsc para encontrar la cantidad óptima de clústeres. El siguiente código es una manera fácil de obtener el valor de wcsc para diferentes números de clústeres:

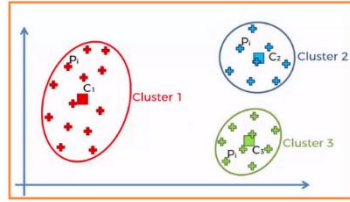


Figura 11.1. Diagrama representativo de clúster

Matemáticamente

$$WCSS = \sum_{P_i \text{ in Cluster 1}} distancia (P_i, C_1)^2 + \sum_{P_i \text{ in Cluster 2}} distancia (P_i, C_2)^2 + \sum_{P_i \text{ in Cluster 3}} distancia (P_i, C_3)^2$$

Método de Clustering

Esta técnica se basa en la clusterización y consiste en particionar un conjunto de m observaciones, en un conjunto de datos con características similares, se parte de una muestra E de m individuos, $X_1, \dots, X_m$ de variables aleatorias, cada una de los cuales está representada por un vector n – dimensional, $X_j = (x_{j1}, x_{j2}, \dots, x_{jn})'$, $j = 1, \dots, m$, [14] debemos encontrar una partición de la muestra en regiones $\omega_1, \dots, \omega_c$ de forma que:

$$\bigcup_{i=1}^c \omega_i = E \quad (1)$$

$$\omega_i \cap \omega_j = \emptyset ; i \neq j \quad (2)$$

Método K-means



Se desea particionar un conjunto de datos determinado en varios grupos: cada observación del conjunto de datos en el mismo grupo comparte más similitudes. El algoritmo busca minimizar la varianza total del sistema. Por lo tanto, para un conjunto de datos $\{x_i\}_{i=1}^n \in R^p$, tal que, se busca encontrar k centroides que minimicen la distancia total entre estos puntos de datos y sus centroides correspondientes, donde $\{c_j\}_{j=1}^k$ son los centroides y $\{\Psi_j\}_{j=1}^k$ son sus particiones correspondientes. Aquí $\{\Psi_j\}_{j=1}^k$ son mutuamente disyuntos y $\cup_{j=1}^k \Psi_j = [n]$. La suma total de la distancia es [15]:

$$f(c_j, \Psi_j) := \sum_{j=1}^k \sum_{i \in \Psi_j} d(x_i, c_j) \quad (3)$$

Se denomina función k-means, que depende de los centroides $\{c_j\}_{j=1}^k$ y $\{\Psi_j\}_{j=1}^k$. Para este caso $d(x_i, c_j)$ se denota como la distancia, que mide la similitud entre x y y .

Método de componentes principales

Es una técnica estadística multivariada cuyo objeto es representar un conjunto de variables observadas en un grupo de individuos por un número pequeño de variables nuevas construidas mediante combinaciones lineales de las variables originales. A continuación, se definen los componentes principales en términos poblacionales [16]:

Sea $X = (X_1, \dots, X_p)'$ un vector aleatorio p – dimensional y sean μ y Σ su esperanza y covarianza

$$E[X] = \mu = 0 \quad Cov[X] = \Sigma = (\sigma_{ij}) \quad (4)$$

Sea $\alpha = (\alpha_1, \dots, \alpha_p)' \in R^p$ tal que $\alpha' \alpha = 1$. Se consideran combinaciones lineales

$$\alpha' X = (\alpha_1, \dots, \alpha_p) \begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix} = \alpha_1 X_1 + \dots + \alpha_p X_p = \sum_{i=1}^p \alpha_i X_i \quad (5)$$

Cuya varianza vendrá dada por



$$Var[\alpha'X] = E[(\alpha'X)^2] = E[\alpha'XX'\alpha] = \alpha'\Sigma\alpha \quad (6)$$

Cálculo de las componentes principales

Sea una matriz X con $n \times p$ datos

$$X = (x_{11} \cdots x_{1p} \vdots \vdots x_{n1} \cdots x_{np}) \quad (7)$$

Definida en una población con matriz de covarianza Σ o correlación para las p variables involucradas.

Se denotan las variables de este conjunto como elementos del vector aleatorio:

$$\underline{x} = (x_1, x_2, \dots, x_p) \quad (8)$$

Con el PCA intentaremos construir las siguientes combinaciones lineales:

$$\begin{aligned} y_1 &= \underline{a}'_1 \underline{x} = a_{11}x_1 + a_{12}x_2 + \cdots + a_{1p}x_p, \\ y_2 &= \underline{a}'_2 \underline{x} = a_{21}x_1 + a_{22}x_2 + \cdots + a_{2p}x_p, \\ &\vdots \\ y_p &= \underline{a}'_p \underline{x} = a_{p1}x_1 + a_{p2}x_2 + \cdots + a_{pp}x_p \end{aligned} \quad (9)$$

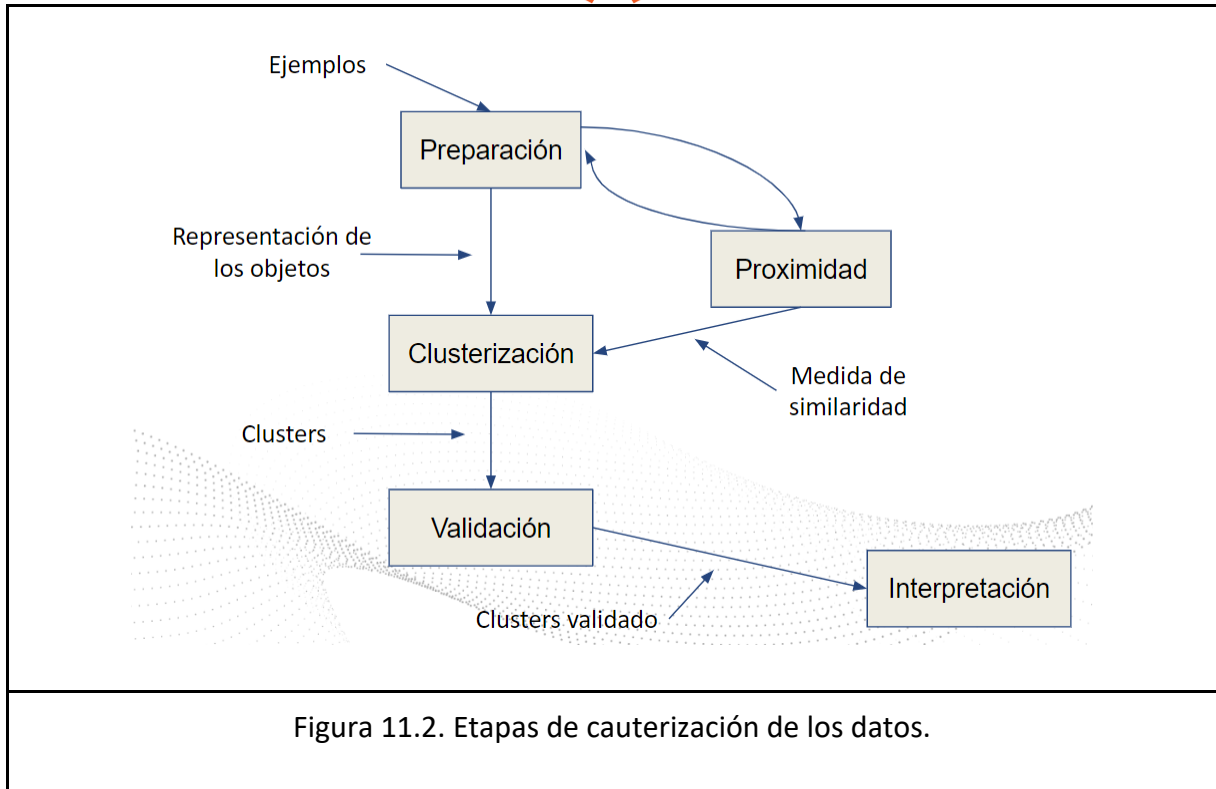
Donde $\underline{a}'_1, \underline{a}'_2, \dots, \underline{a}'_p$ son desconocidos y  los $p \times 1$ son los vectores que contienen los pesos de las combinaciones lineales. Por lo tanto, la primera componente principal y_1 explica la mayor varianza posible, lo que resultará de las combinaciones lineales de las variables iniciales en $\underline{x}$. Se impone una restricción para la magnitud de los elementos de $\underline{a}_1$ ya que, si no se hace, la $var(y_1)$ aumentará desmedidamente, lo cual impide encontrar el vector $\underline{a}_1$. Por lo general asumimos que la longitud es de 1, esto es $\|\underline{a}_1\| = 1$, una vez que se encuentre la solución y el vector, óptimo de pesos $\underline{a}_1$ la combinación lineal $\underline{a}'_1 \underline{x}$ es llamada la primera componente principal. El siguiente paso es encontrar un segundo vector $\underline{a}_2$ tal que la segunda componente $\underline{a}'_2 \underline{x}$ tenga las siguientes propiedades [16]:

1. Debe de explicar la mayor varianza posible de los datos restantes (considerando de nuevo la restricción).
2. No debe estar correlacionado con la primera componente principal.

Ambas componentes no deben estar correlacionadas, se busca hallar el número de combinaciones lineales que expliquen la porción más grande posible de las varianzas de las variables iniciales. Una vez encontrado este conjunto de valores $\underline{a}_2$, las combinaciones lineales $\underline{a}'_2 \underline{x}$ se denominará la segunda componente principal. Podemos continuar con este comportamiento hasta que encontremos p vectores $\underline{a}_1, \underline{a}_2, \dots, \underline{a}_p$ que aparecen en el lado derecho de la ecuación (9), por definición las sucesivas componentes principales no están correlacionadas y formalmente hay tantas componentes principales como variables estudiadas [16].

Etapas del proceso de clusterización

Los procesos de clusterización necesitan de unas etapas que validen la estrategia de conglomeración de los datos.



- **Preparación:** normalizaciones, conversiones de tipos, reducción del número de atributos
- **Proximidad:** definición (elección) de una medida de similitud apropiada al dominio de la aplicación, por ejemplo. distancia.
- **Clustering:** aplicación de algoritmos de clustering
- **Validación:** evaluación del resultado de la agrupación. Debe determinar si los conglomerados son significativos.
- **Interpretación:** etiquetado del clúster, definiendo su naturaleza, a través del análisis de sus objetos.

IV. Implementación del método de predicción basado en random forest y K-NN para evaluación del método de clasificación a través del uso de métricas.

¿Qué es un algoritmo de aprendizaje supervisado?

En aprendizaje supervisado, un  algoritmo recibe un conjunto de datos que están etiquetados con los valores de salida correspondientes sobre los que puede entrenarse y definir un modelo de predicción. Este algoritmo podrá luego ser utilizado sobre datos nuevos para predecir sus valores de salida correspondientes.

El algoritmo de las K vecinas más cercanas o K-nearest neighbors (kNN) es un algoritmo de Machine Learning que pertenece a los algoritmos de aprendizaje supervisado simples y fáciles de aplicar que pueden ser utilizados para resolver problemas de clasificación y de regresión. En este artículo veremos la definición de este algoritmo, su funcionamiento, así como también una aplicación directa en programación.

Método K-NN

La lógica detrás del algoritmo de las K vecinas más cercanas es una de las más sencillas de todos los algoritmos de Machine Learning supervisados:

- Etapa 1: Seleccionar el número de K vecinas
- Etapa 2: Calcular la distancia
- Etapa 3: tomar las K vecinas más cercanas según la distancia calculada
- Etapa 4: entre las K vecinas, contar el número de puntos en cada categoría.
- Etapa 5: atribuir un nuevo punto a la categoría más presente entre las K vecinas
- Etapa 6: El modelo está listo

Método Random Forest

- Se considera como la “panacea” en todos los problemas de ciencia de datos.
- Útil para regresión y clasificación.
- Un grupo de modelos “débiles”, se combinan en un modelo robusto.
- Sirve como una técnica para reducción de la dimensionalidad.
- Se generan múltiples árboles (a diferencia de CART: Comparación entre árboles de regresión).
- Cada árbol da una clasificación (vota por una clase). Y el resultado es la clase con mayor número de votos en todo el **random forest**.

- Para regresión, se toma el  promedio de las salidas (predicciones) de todos los árboles.

¿Cómo se construye un modelo random forest?

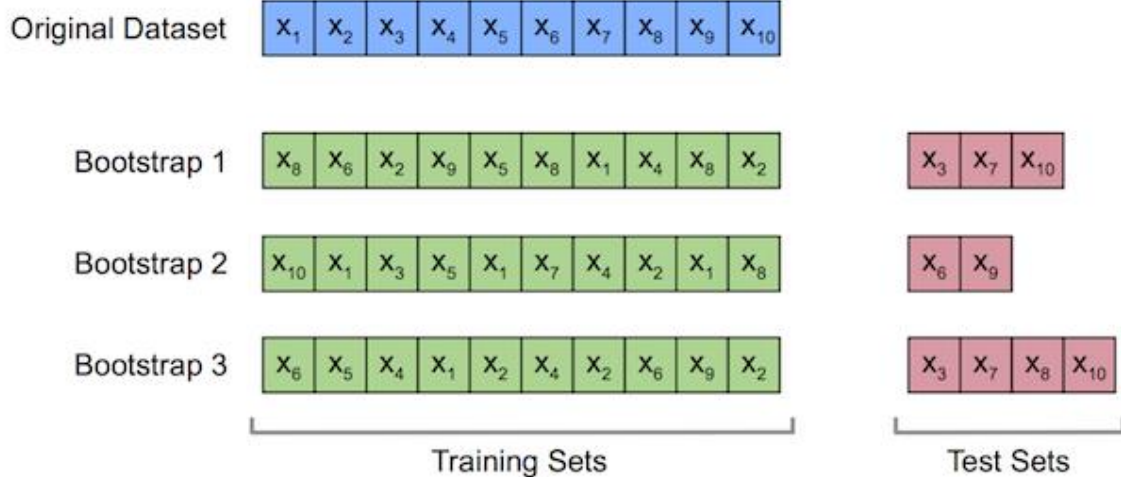
Cada árbol se construye así:

- Dado que el número de casos en el conjunto de entrenamiento es N . Una muestra de esos N casos se toma aleatoriamente, pero **CON REEMPLAZO**. Esta muestra será el conjunto de entrenamiento para construir el árbol i .
- Si existen M variables de entrada, un número $m < M$ se especifica tal que para cada nodo, m variables se seleccionan aleatoriamente de M . La mejor división de estos m atributos es usado para ramificar el árbol. El valor m se mantiene constante durante la generación de todo el bosque.
- Cada árbol crece hasta su máxima extensión posible y **NO hay proceso de poda**.
- Nuevas instancias se predicen a partir de la agregación de las predicciones de los x árboles (i.e., mayoría de votos para clasificación, promedio para regresión)

Funcionamiento del algoritmo Random Forest

Out of bag samples y out of bag error (Muestras fuera de bolsa y error fuera de bolsa)

- El proceso de muestreo de los datos con reemplazo se denomina **bootstrap**.
- Un tercio de los datos no se usan para el entrenamiento y pueden ser usados para test.
- Este conjunto se denomina out of bag (OOB) samples.



This work by Sebastian Raschka is licensed under a Creative Commons Attribution 4.0 International License.

Figura 11.3 Iteraciones en el método random forest

- El error estimado en estos out of bag samples se conoce como out of bag error (OOB error)
- Usar este conjunto de test (OOB) es tan preciso como si se usara un conjunto de test del mismo tamaño que el de entrenamiento.
- Sería posible no usar un conjunto de test adicional.

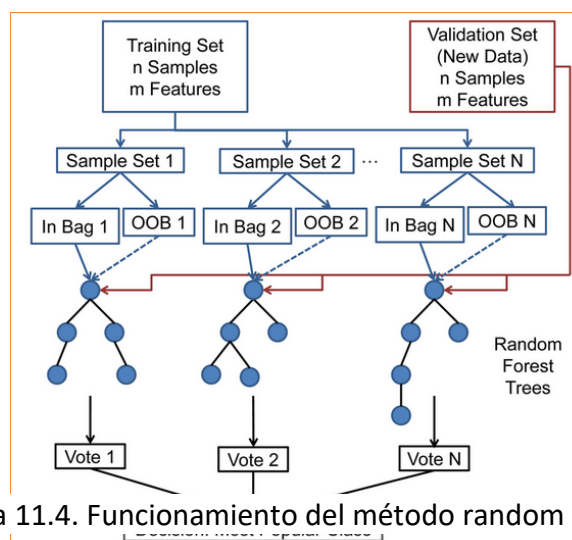


Figura 11.4. Funcionamiento del método random forest

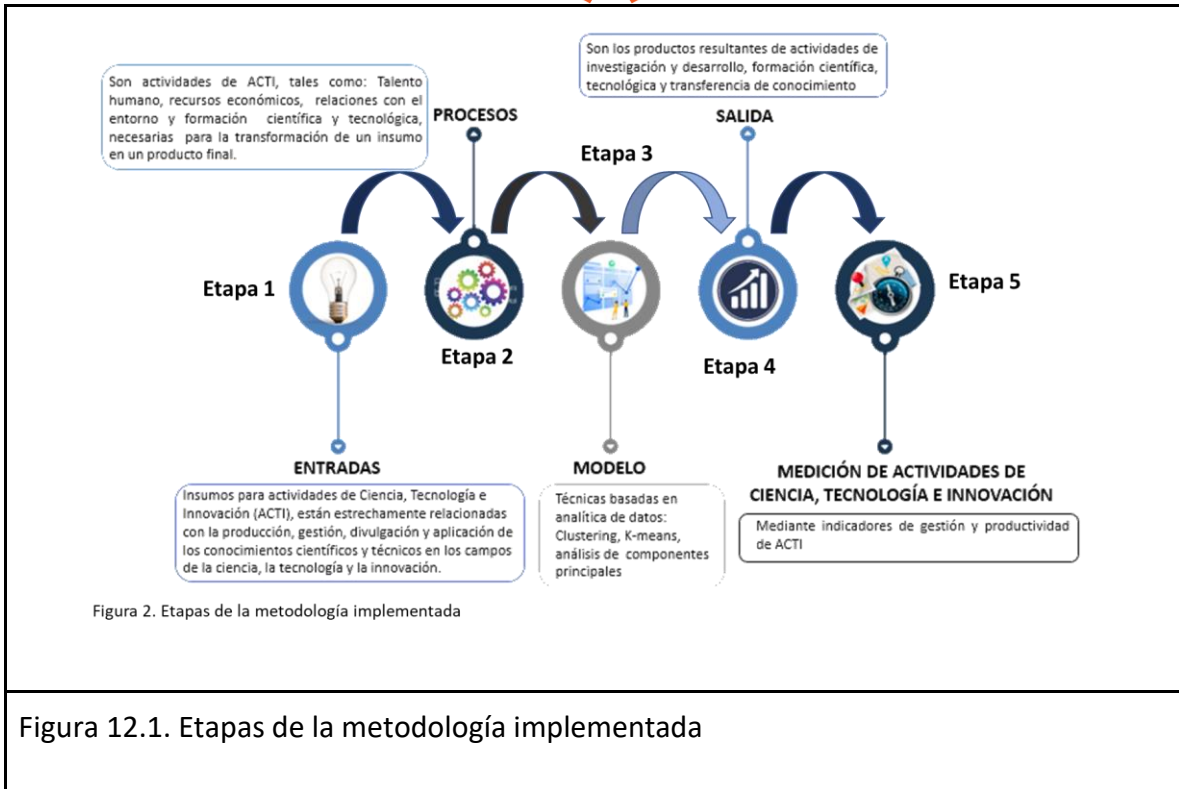
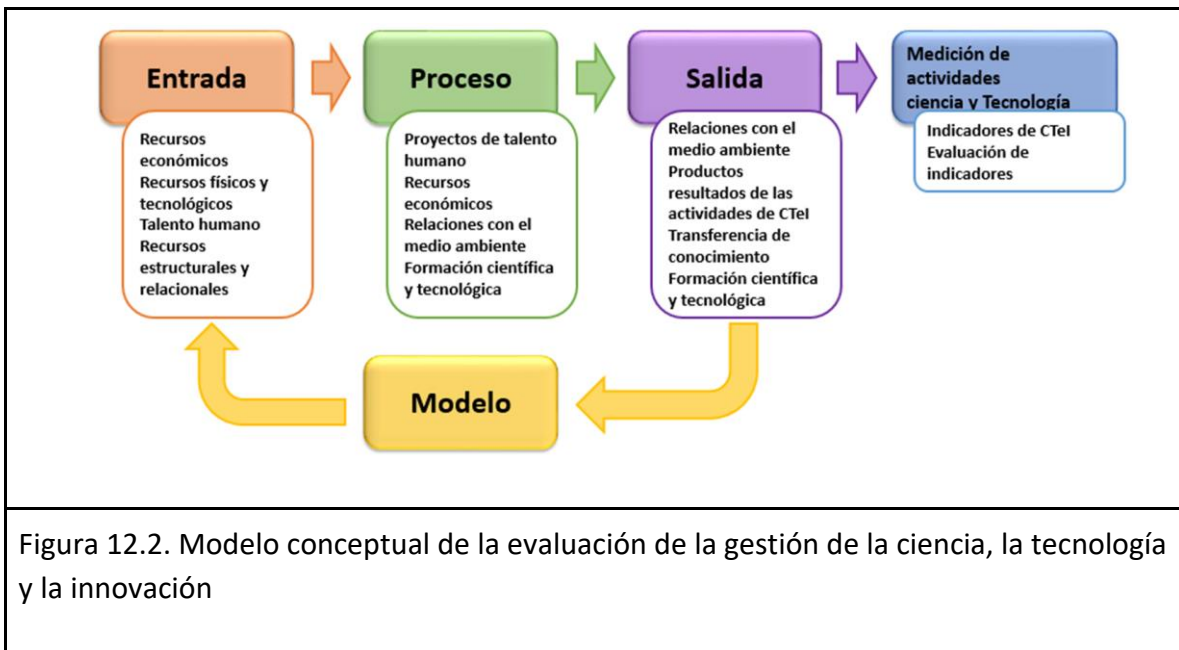


12 METODOLOGÍA

12.1 Metodología de la investigación

El enfoque desarrollado en nuestra investigación es de tipo mixto este implica la recolección y el análisis de datos cuantitativos y cualitativos, permitiendo integrarlos de manera conjunta, para realizar inferencias de la información realizada y lograr una mayor comprensión del objeto de estudio propuesto [6], ya que el tema se abordó por primera vez en este contexto, el método de estudio es fenomenológico ya que intenta interpretar y comprender el proceso de desarrollo, ejecución y validación del modelo de evaluación y valoración de capacidades de (CTel) en los centros de formación de la regional Antioquia del SENA, asimismo, el proceso de investigación es inductivo - deductivo. Los fundamentos expuestos en el marco del sistema de I + D [7] pretenden medir las capacidades de ciencia, tecnología e innovación desde cuatro dimensiones, que son entrada, proceso, salida y entorno.

Sin embargo, la propuesta se realizó teniendo en cuenta sólo las dimensiones de entrada y salida, dado que; en primer lugar, ignoramos la variable entorno debido a que las políticas de (CTel) son cambiantes de acuerdo con la industria y la región. En segundo lugar, el proceso juega un papel clave en la formación de capacidades, sin embargo, está excluido por dos razones: la primera referida a la dificultad en cuantificar para algunas organizaciones los procesos y actividades de ciencia, tecnología e innovación y la segunda en cuanto a la intensidad de los insumos de (ACTI) y los resultados de producción que podrían representar el proceso de (CTel) en conjunto. Se propuso un modelo de medición de indicadores para que las distintas organizaciones evalúen el impacto de las (ACTI), en la figura 1 se realizó a través de indicadores de insumos y productos [8,9,10]. Siendo los insumos de (CTel) todos aquellos recursos que respaldan las actividades de (CTel). Por otro lado, los indicadores de salida (CTel), son los resultados producto de las ACTI [11].



12.2 Metodología CRISP-DM



La metodología CRISP - DM de sus siglas en inglés de *Cross-Industry Standard Process for Data Mining*, es uno de los métodos más reconocidos que permiten orientar trabajos de minería de datos. Esta permite la extracción del conocimiento de un grupo de datos en diferentes etapas

12.2.1 Comprensión de los datos: El objetivo de esta fase es familiarizarse con los datos disponibles.

12.2.2 Preparación de los datos: Se realiza una etapa de preprocesamiento y exploración de los datos, tales como limpieza, normalización y preparación de la información que lleven a alcanzar los objetivos que persigue el proyecto.

12.2.3 Modelado: Se determina la técnica matemática a implementar y se buscan los algoritmos que permitan resolver el problema objetivo con los parámetros establecidos.

12.2.4 Evaluación: En esta fase se realiza la etapa de validación de modelos, en el que permite dar veracidad de los resultados obtenidos a partir de métricas que permite dar certeza de esto.

12.2.5 Implementación: En esta fase se busca lograr que el modelo propuesto cumpla con los objetivos del proyecto, en base a los requisitos comerciales de la organización en este caso la medición de los indicadores de ciencia, tecnología e innovación.

12.2.6 Modelo de analítica de datos: Integración de técnicas que permitan la toma de decisiones en base a los datos propuestos, todo esto orientado al propósito de medir indicadores de Ciencia Tecnología e Innovación.

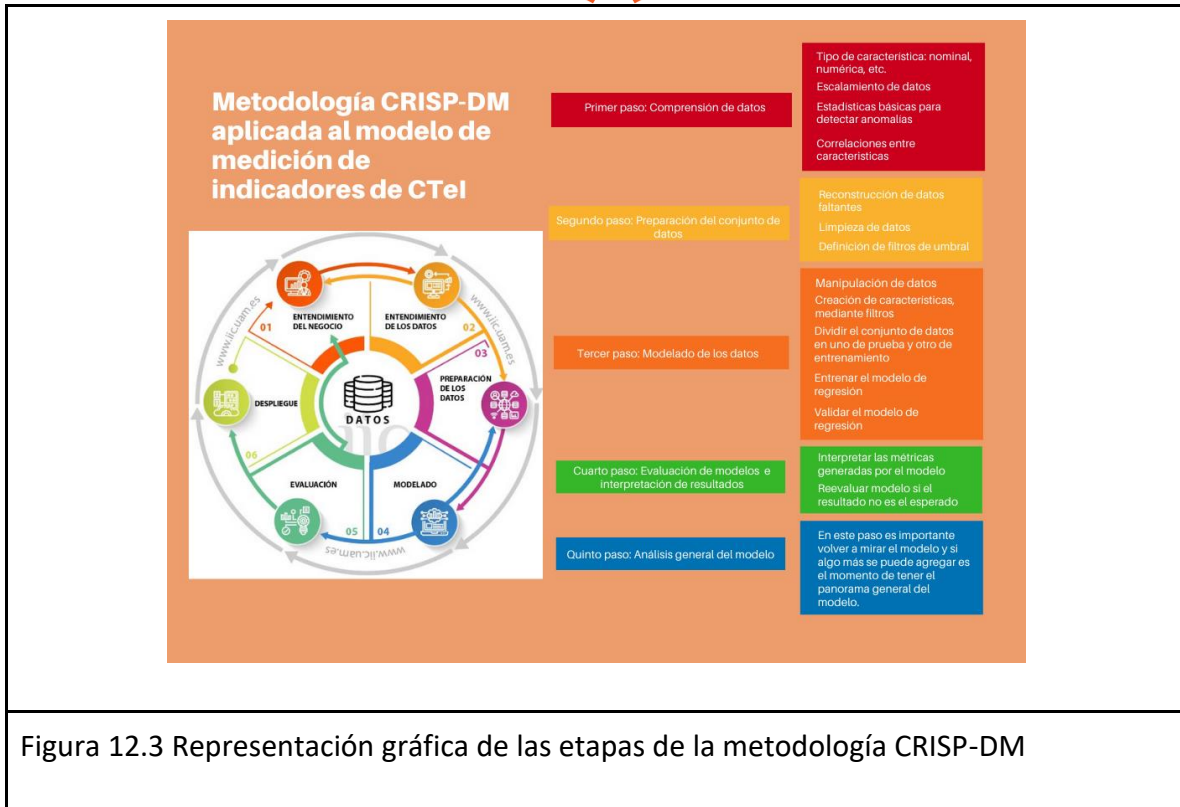


Figura 12.3 Representación gráfica de las etapas de la metodología CRISP-DM

12.3 Descripción de los datasets

Esta sección describe todos los conjuntos de datos utilizados para el desarrollo de la solución propuesta. Se creó un conjunto de datos sintéticos que permitieron generalizar los resultados de los modelos de la analítica de datos.

12.3.1 Resumen de los datasets y modelo relacional. En la siguiente tabla se resumen las principales características de los conjuntos de datos que tendremos en cuenta a lo largo del desarrollo de nuestro proyecto.



Regionales	Grupo_investigacion	Semilleros
_Codigo_Regional	_Codigo_Grupo_Investigacion	_Codigo_Semillero
_Nombre_Regional	_Nombre_Grupo_Investigacion	_Nombre_Semillero
_Codigo_Departamento	_Categoria_Grupo	_Codigo_Grupo_Investigacion
_Codigo_Municipio	_Codigo_Centro	_Codigo_Centro
	_Codigo_Regional	_Codigo_Regional
Rubro	Rol_Sennova	Tipo_Contratacion
_Codigo_Rubro	_Codigo_Rol	_Codigo_Tipo_Contrato
_Nombre_Rubro	_Nombre_Rol	_Nombre_Tipo_Contrato
Principal_Productos	Principal_Proyectos	Principal_Talento_Humano
_Codigo_Producto	_Codigo_Proyecto	_Codigo_Persona
_Nombre_Producto	_Nombre_Proyecto	_Codigo_Nivel_Academico
_Descripcion_producto	_Valor_Proyecto	_Codigo_Tipo_Contrato
_Codigo_Regional	_Codigo_Rubro	_Objeto_Contrato
_Codigo_Centro	_Fecha_inicio_Proyecto	_Codigo_Rubro
_Codigo_Rubro	_Fecha_Cierre_Proyecto	_Codigo_Area_OCDE
_Codigo_Tipologia	_Codigo_Regional	_Codigo_Estado_Contratacion
_Codigo_Proyecto	_Codigo_Centro	_Codigo_Regional
_Nombre_Proyecto	_Codigo_Estado_Proyecto	_Codigo_Centro
_Codigo_Grupo_Investigacion	_Codigo_area_OCDE	_Codigo_Proyecto
_Codigo_Semillero	_Serial_equipo	_Codigo_Producto
_Codigo_persona	_Valor_Servicios_Personales	
_Codigo_area_OCDE	_Valor_Compra_Equipos	
_Fecha_registro	_Valor_Software	
	_Valor_Otros_Gastos	
Area_conocimiento_OCDE	Estado_Proyecto	Nivel_Academico
_Codigo_area_OCDE	_Codigo_Estado_Proyecto	_Codigo_Nivel_Academico
_Nombre_area_OCDE	_Nombre_Estado_Proyecto	_Nombre_Nivel_Academico
Municipio	Centros_Formacion	Principal_Equipos
_Codigo_Municipio	_Codigo_Centro	_Serial_equipo
_Nombre_Municipio	_Nombre_Centro	_Nombre_del_equipo
_Codigo_Departamento	_Codigo_Regional	_Valor_equipo
	_Codigo_Municipio	_Fecha_de_compra
	_Codigo_Departamento	_Codigo_Proyecto
	_latitud	_Codigo_Regional
	_longitud	_Codigo_Centro
		_Codigo_Rubro
		_Codigo_Tipo_Equipo
		_Codigo_area_OCDE
		_Codigo_persona
Tipologia_Productos	Tipo_Contratacion	Tipo_equipo
_Codigo_Tipologia	_Codigo_Tipo_Contrato	_Codigo_Tipo_Equipo
_Nombre_Tipologia	_Nombre_Tipo_Contrato	_Nombre_Tipo_Equipo
Rol_Proyectos	Estado_Contratacion	Departamento
_Codigo_Rol_Proyecto	_Codigo_Estado_Contratacion	_Codigo_Departamento
_Nombre_Rol	_Nombre_Estado_Contratacion	_Nombre_Departamento

Tabla 12.4. Base de datos de tablas y variables a analizar

De este conjunto de datos, se  seleccionaron las tablas que se relacionan con la dimensión de procesos de gestión y productividad, como se muestra en el siguiente diagrama de entidad-relación de la Tabla 7.

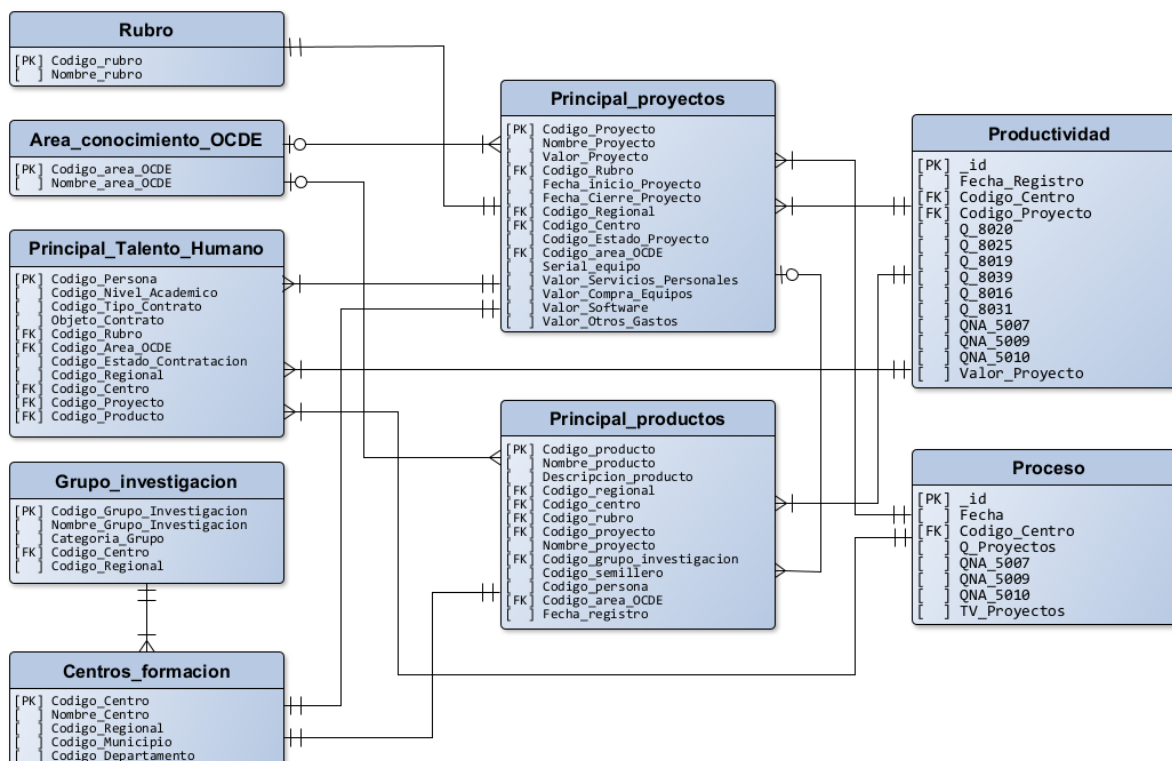


Figura 12.5. Modelo entidad-relación para dimensiones de productividad y procesos de los centros de formación SENA simulados con datos sintéticos.

12.3.2 Generación de datos pseudoaleatorios. Como mencionado anteriormente, para la generación de datos sintéticos pseudoaleatorios para cada una de las variables correspondientes al modelo entidad-relación, se crearon utilizando R.



RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

docs Tipología_Productos Tipo_Contratacion Regionales Principales_Productos Codigo.R Codigo.R

	Codigo_Proyecto	Nombre_Proyecto	Valor_Proyecto	Codigo_Rubro	Fecha_Inicio_Proyecto	Fecha_Cierre_Proyecto	Codigo_Regional	Codigo_Centro	Codigo_Estado_Proyecto	Codigo_area_OCDE
1	10651	Corporación de tecnología Carahsoft	147801416	76	2019-11-25	2019-05-22	5	9274	3	
2	8327	Tecnología SilkRoad	100264266	61	2014-01-24	2015-01-07	5	9164	3	
3	5549	Corporación de tecnología Carahsoft	235580787	30	2017-06-13	2019-05-17	5	9479	4	
4	7624	Tecnología Trek Wheel	50434232	47	2021-11-25	2014-04-04	5	9419	4	
5	2369	Grupo Techcess	21930280	73	2019-11-10	2012-09-27	5	9483	3	
6	5478	Aspect Technology Inc	412231127	27	2018-10-10	2015-02-15	5	9329	3	
7	8939	Tecnología Trek Wheel	191597410	80	2014-09-13	2011-01-13	5	9297	3	
8	1028	Infosys	45734696	35	2012-01-17	2020-02-09	5	9461	3	
9	1732	Micron Technology Inc	497082564	81	2015-04-24	2019-09-16	5	9445	3	
10	3479	Falcon Tech	308855993	72	2017-11-05	2019-12-21	5	9142	4	
11	11033	Señales de alta tecnología	99623441	81	2020-07-21	2017-05-15	5	9487	4	
12	11490	DuPontmakers Technology Inc	166624849	47	2019-04-12	2016-07-22	5	9312	4	
13	7520	PTC	438635920	79	2020-09-27	2022-07-11	5	9417	3	
14	445	Kingston Technology Co Inc	459763526	82	2021-12-29	2019-04-10	5	9120	4	
15	7669	Infosys	284382483	59	2011-09-11	2022-02-02	5	9532	3	
16	7349	NetStandard Inc	230668625	36	2014-05-23	2020-06-16	5	9463	4	
17	3436	Corporación de tecnología Carahsoft	431271879	70	2019-07-20	2017-05-29	5	9140	4	
18	10569	Aspect Technology Inc	199256297	70	2016-05-22	2014-07-13	5	9429	4	

Showing 1 to 18 of 500 entries, 15 total columns

Figura 12.6. Organización de base de base de datos y creación de datos sintéticos



13 RESULTADOS

13.1 Análisis exploratorio de los datos



El principal objetivo de este análisis exploratorio de datos (AED) de esta investigación es analizar y comprender a partir de la generación de diferentes gráficos estadísticos y cálculo de determinadas métricas, patrones de comportamiento y características conjuntas de las actividades de ACTI de los datos pseudoaleatorios generados para distintas métricas de indicadores. Estos son algunos de estos.

Como una etapa inicial de análisis de los datos, se realizaron extracciones de informaciones que nos permitiera obtener datos estadísticos, así como de variables correspondientes al modelo ER. Para Para nuestros análisis, dentro del modelo relacional cuatro (4) bases de datos de estudio, principal_proyecto, principal_talento_humano, principal_productos y principal_equipos que se describen en la Tabla 4, esto a raíz de su alta relacionamiento con tablas adicionales, y así permitir la generación de métricas en modelos de analítica de datos.

13.1.1. base de datos principal_proyecto

En la tabla a continuación se documenta las diferentes características de los principales proyectos a nivel de SENA - Regional Antioquia, tal como se muestra en la Tabla 13.1.

Codigo_Proyecto	Nombre_Proyecto	Valor_Proyecto	Codigo_Rubro	Fecha_inicio_Proyecto	Fecha_Cierre_Proyecto	Codigo_Regional	Codigo_Centro	Codigo_Estado_Pi
10651	Corporación de tecnología Carahsoft	147801416	65	2010-04-30	2017-11-06	5	9501	
8327	Tecnología SilkRoad	100264266	66	2012-02-13	2022-05-04	5	9503	
5549	Corporación de tecnología Carahsoft	235580787	82	2017-11-25	2021-07-04	5	9202	
7624	Tecnología Trek Wheel	50434232	23	2022-03-29	2011-11-30	5	9206	
2369	Grupo Techcess	21930280	66	2015-02-25	2011-07-31	5	9127	

Tabla 13.1 Información de tabla con sus respectivos códigos de variable.

En la Tabla 13.2, que sigue a  continuación, se describe la cantidad de variables que en este caso son 15, con 500 filas de datos sintéticos.

#	Column	Non-Null Count	Dtype
0	Codigo_Proyecto	500 non-null	int64
1	Nombre_Proyecto	500 non-null	object
2	Valor_Proyecto	500 non-null	int64
3	Codigo_Rubro	500 non-null	int64
4	Fecha_inicio_Proyecto	500 non-null	datetime64[ns]
5	Fecha_Cierre_Proyecto	500 non-null	datetime64[ns]
6	Codigo_Regional	500 non-null	int64
7	Codigo_Centro	500 non-null	int64
8	Codigo_Estado_Proyecto	500 non-null	int64
9	Codigo_area_OCDE	500 non-null	int64
10	Serial_equipo	500 non-null	int64
11	Valor_Servicios_Personales	500 non-null	int64
12	Valor_Compra_Equipos	500 non-null	int64
13	Valor_Software	500 non-null	int64
14	Valor_Otros_Gastos	500 non-null	int64

dtypes: datetime64[ns](2), int64(12), object(1)
memory usage: 58.7+ KB

Tabla 13.2. Características de la información registrada para principal_proyectos con datos sintéticos

En la Figura 13.1. se puede observa en gráficas de series temporales la evolución de gestión de proyectos de inicio (Figura 13.1.a) así como el fin del proyecto (Figura 13.1.b).

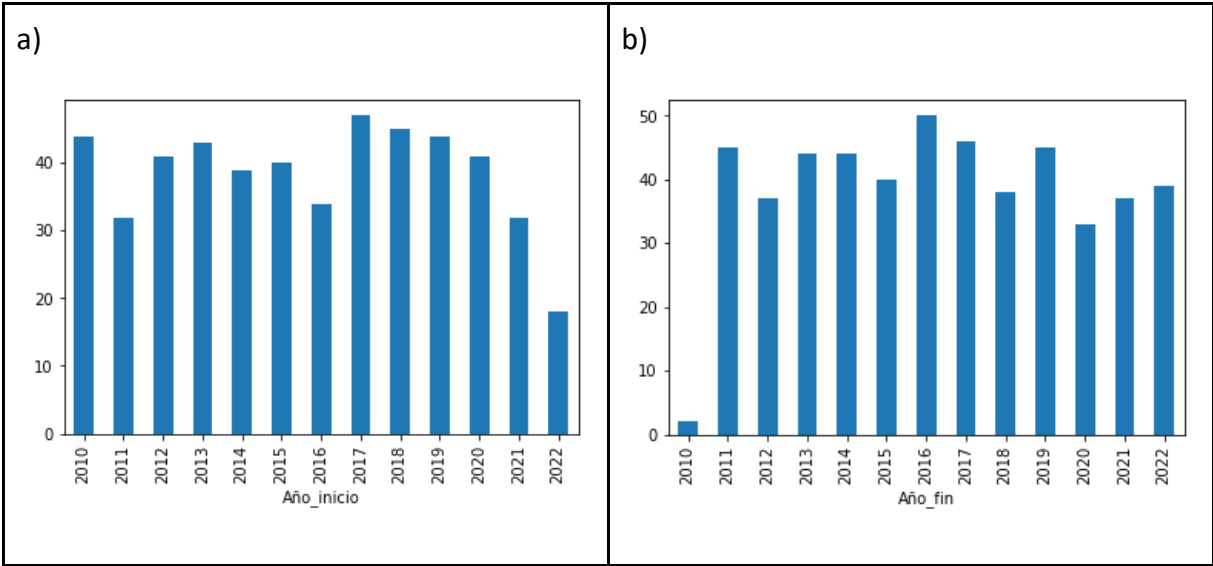


Figura 13.1. Serie temporal de inicio y fin de proyectos para los años 2010 al 2022 a) Inicio de proyecto; b) Fin de proyecto.



13.1.2. base de datos principal_talento_humano

En la tabla a continuación se documenta las diferentes características de talento humano a nivel de SENA - Regional Antioquia, tal como se muestra en la Tabla 13.3.

	Codigo_Persona	Codigo_Nivel_Academico	Codigo_Tipo_Contrato	Objeto_Contrato	Codigo_Rubro	Codigo_Area_OCDE	Codigo_Estado_Contratacion	Codigo_Regional	Cov
0	3029	5007	2003	NaN	23	4	1	5	
1	3019	5006	2005	NaN	23	4	1	5	
2	3016	5007	2002	NaN	23	2	1	5	
3	3020	5010	2004	NaN	23	4	1	5	
4	3013	5007	2002	NaN	23	6	1	5	

Tabla 13.3 Contenido de tabla con sus respectivos códigos de variable.

En la Tabla 13.4a, que sigue a continuación, se describe la cantidad de variables que en este caso son 11, con 500 filas. Cabe destacar que, la Tabla 13.4.b a pesar de no representar lo contenido a talento humano, si representa la estructura de niveles educativos al que es clasificada dentro la institución, y es registrada por una llave primaria que posteriormente será relacionada con la tabla de talento humano.

a) <class 'pandas.core.frame.DataFrame'> RangeIndex: 500 entries, 0 to 499 Data columns (total 11 columns): # Column Non-Null Count Dtype --- - 0 Codigo_Persona 500 non-null int64 1 Codigo_Nivel_Academico 500 non-null int64 2 Codigo_Tipo_Contrato 500 non-null int64 3 Objeto_Contrato 0 non-null float64 4 Codigo_Rubro 500 non-null int64 5 Codigo_Area_OCDE 500 non-null int64 6 Codigo_Estado_Contratacion 500 non-null int64 7 Codigo_Regional 500 non-null int64 8 Codigo_Centro 500 non-null int64 9 Codigo_Proyecto 500 non-null int64 10 Codigo_Producto 500 non-null int64 dtypes: float64(1), int64(10) memory usage: 43.1 KB	b) <table><thead><tr><th>Codigo_Nivel_Academico</th><th>Nombre_Nivel_Academico</th></tr></thead><tbody><tr><td>5001</td><td>Bachiller</td></tr><tr><td>5002</td><td>Técnico_certificado</td></tr><tr><td>5003</td><td>Técnico_profesional</td></tr><tr><td>5004</td><td>Técnico_profesional_con_especialización</td></tr><tr><td>5005</td><td>Tecnólogo</td></tr><tr><td>5006</td><td>Tecnólogo_con_especialización</td></tr><tr><td>5007</td><td>Profesional</td></tr><tr><td>5008</td><td>Profesional_con_especialización</td></tr><tr><td>5009</td><td>Profesional_con_maestría</td></tr><tr><td>5010</td><td>Profesional_con_doctorado</td></tr></tbody></table>	Codigo_Nivel_Academico	Nombre_Nivel_Academico	5001	Bachiller	5002	Técnico_certificado	5003	Técnico_profesional	5004	Técnico_profesional_con_especialización	5005	Tecnólogo	5006	Tecnólogo_con_especialización	5007	Profesional	5008	Profesional_con_especialización	5009	Profesional_con_maestría	5010	Profesional_con_doctorado
Codigo_Nivel_Academico	Nombre_Nivel_Academico																						
5001	Bachiller																						
5002	Técnico_certificado																						
5003	Técnico_profesional																						
5004	Técnico_profesional_con_especialización																						
5005	Tecnólogo																						
5006	Tecnólogo_con_especialización																						
5007	Profesional																						
5008	Profesional_con_especialización																						
5009	Profesional_con_maestría																						
5010	Profesional_con_doctorado																						



Tabla 13.4. a) Cantidad de variables, filas y su tipología; b) tabla de datos alterna a principal_talento_humano, que describe el código respectivo en este caso a Nivel Académico.

13.1.3. base de datos principal_productos

En esta tabla se documenta las diferentes características de productos de ciencia y tecnología generados a nivel de SENA - Regional Antioquia, de manera sintética, tal como se muestra en la Tabla 13.5

	Codigo_Producto	Nombre_Producto	Descripcion_producto	Codigo_Regional	Codigo_Centro	Codigo_Rubro	Codigo_Proyecto	Nombre_Proy
0	8039	Articulo_Sometido_en_revista_Indexada	NaN	5	9549	82	8602	
1	8029	Productos_tecnológicos_certificados_o_validados	NaN	5	9202	65	6465	
2	8019	Articulo_Sometido_en_revista_Indexada	NaN	5	9504	82	11205	
3	8016	Libros_resultados_de_investigación	NaN	5	9504	65	10481	
4	8020	Apropiación_Social	NaN	5	9204	65	1770	
5	8013	Apropiación_Social	NaN	5	9203	66	4372	
6	8031	Productos_tecnológicos_certificados_o_validados	NaN	5	9501	23	12075	
7	8016	Libros_resultados_de_investigación	NaN	5	9206	66	1012	
8	8012	Divulgación_Pública_de_la_CTeI_Publicaciones_e...	NaN	5	9203	66	3647	

Tabla 13.5. Contenido de tabla principal_productos con sus respectivos códigos de variable.



En la Tabla 13.6, que sigue a continuación, se describe la cantidad de variables que en este caso son 11, con 500 filas de datos sintéticos.

#	Column	Non-Null Count	Dtype
0	Codigo_Producto	500 non-null	int64
1	Nombre_Producto	500 non-null	object
2	Descripcion_producto	0 non-null	float64
3	Codigo_Regional	500 non-null	int64
4	Codigo_Centro	500 non-null	int64
5	Codigo_Rubro	500 non-null	int64
6	Codigo_Proyecto	500 non-null	int64
7	Nombre_Proyecto	500 non-null	int64
8	Codigo_Grupo_Investigacion	500 non-null	object
9	Codigo_Semillero	500 non-null	int64
10	Codigo_persona	500 non-null	int64
11	Codigo_area_OCDE	500 non-null	int64
12	Fecha_registro	500 non-null	datetime64[ns]

dtypes: datetime64[ns](1), float64(1), int64(9), object(2)
memory usage: 50.9+ KB

Tabla 13.6. Características de la información registrada para principal_productos con datos sintéticos

En la FIGURA 13.2. se puede observar la cantidad de productos científicos categorizados por SENA y el Ministerio de Ciencia Tecnología e Innovación. En esta se observa una distribución casi constante en la producción de productos, exceptuando a artículo sometido en revista científica.



cantidad de productos científicos categorizados por SENA y el Ministerio de Ciencia Tecnología e Innovación.

En esta se observa una distribución casi constante en la producción de productos, exceptuando a artículo sometido en revista científica.

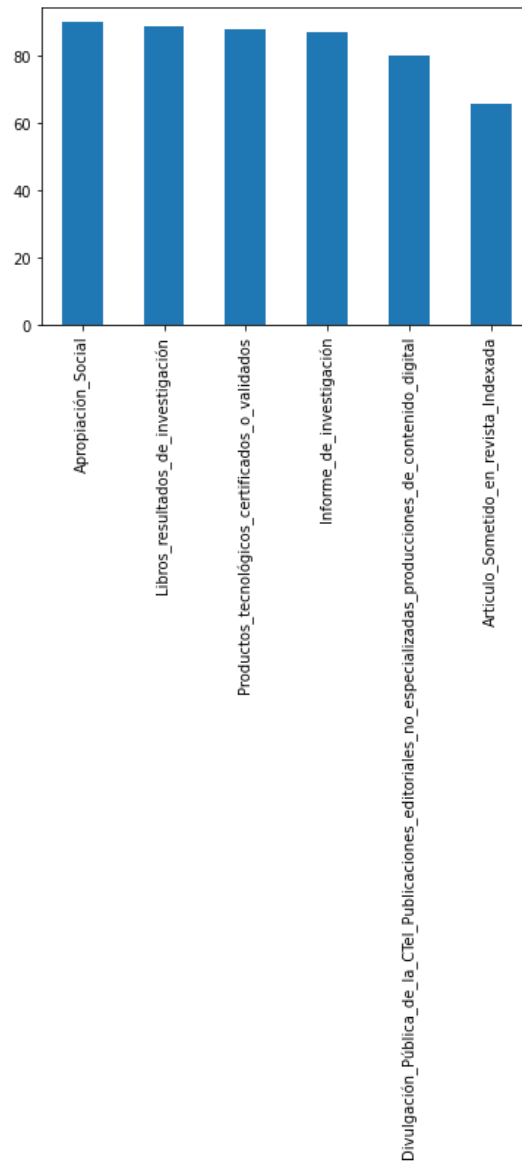


Figura 13.2. Frecuencia de cantidad de códigos por principal_productos.



13.1.4. Determinación geoespacial de Centros de formación SENA - Regional Antioquia.

Para la generación de mapas geoespaciales, se creó una base de datos alterna que dieran la posición de latitud y longitud para los dieciséis (16) diferentes centros de formación del SENA regional Antioquia. Para ello se hizo uso de la librería Folium de python, que permitió dar esta visualización de manera más adecuada, como se observa en la Figura 13.3. En este mapa, se puede observar la densidad de centros de formación a lo largo del departamento de Antioquia, donde claramente se destaca la alta densidad de centros en torno a Medellín así como de municipios aledaños.

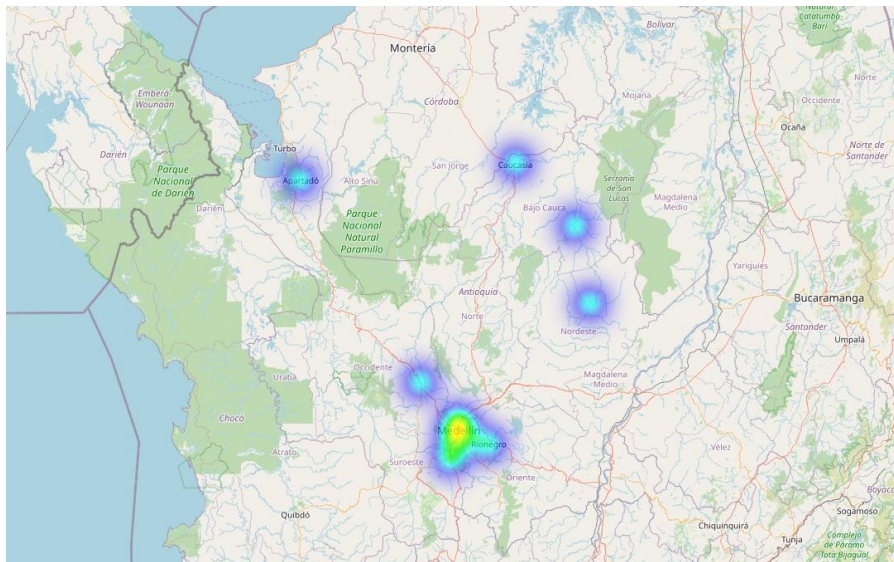


Figura 13.3. Mapa de posición geoespacial de centros de formación



13.2. Generación de mapas

Geoespaciales

Para la generación de indicadores de productividad y procesos de gestión en centros de formación, se procedió a realizar una estrategia de aprendizaje profundo no supervisada, tales como K-means y PCA, que permitieron generar clasificación de los centros en términos de productividad académica y gestión de procesos en alta, media y baja respectivamente. Para ello, se hizo uso de los datos preprocesados en R de *Principal_Proyectos* y *Principal_Productos*, como se observa en la Tabla 11.1 y Tabla 11.2. Para las etapas de clusterización se tuvieron en cuenta las etapas que se ilustran en la Figura 11.2. así como de los índices de validación internos y externos de los mismos.

Posteriormente a este proceso de clasificación, se procede a una etapa de visualización geospacial de alertas de alto, medio y bajo sea esto de *principal_proyectos* y *Principal_productos*.

13.2.1. Mapas geoespaciales para productividad de centros formación

Como mencionamos anteriormente, la generación de estos mapas se realiza en dos etapas: una etapa de clasificación a partir de herramientas de aprendizaje no supervisado, como la clusterización y PCA, y posteriormente se realiza una etapa de visualización de estos mapas.

13.2.1.1 Kmeans y PCA.

Para la selección de la cantidad de clusters adecuados, se aplicó el método del codo, que se visualiza en la Figura 13.4. Debido a nuestro interés de determinar un clasificador de tres alertas de alto, medio y bajo, se opta a seleccionar $k = 3$, que es también valido desde el punto de vista del método del codo.

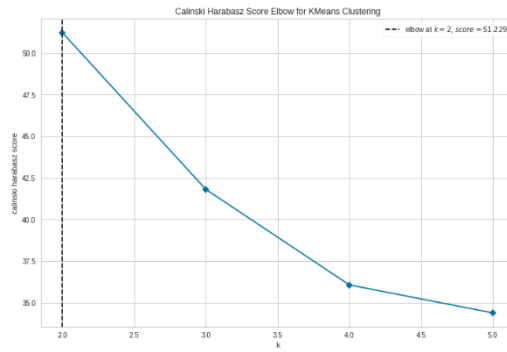
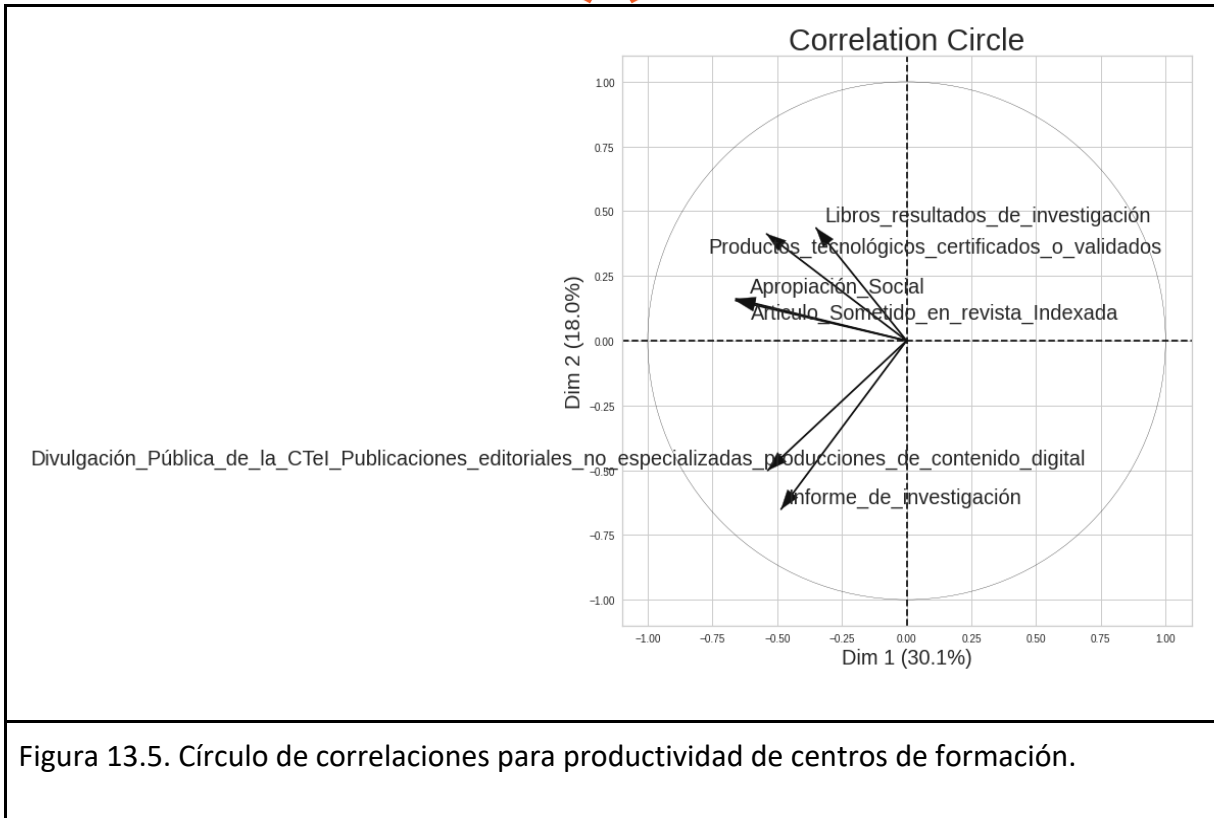


Figura 13.4. Método del codo para la determinación de cantidad de clusters.

Externos						Internos	
Indicador	Homogeneidad	Compleitud	Índice Rand	Medida V	Silueta	Calinski - Harabasz	Davies - Bouldin
Productividad	0,197	1,00	0	0,330	0,171	41,810	1,891

Tabla 13.7. Métricas de cluster.

Se determinan las métricas de cluster tanto internas como externas que se muestran en la Tabla 13.7 En se muestra la tendencia aleatoria de estos clusters, como consecuencia de la generación aleatoria de los datos. Se anexan códigos de proceso de clusterización a partir de K-Means (Ver Anexos).



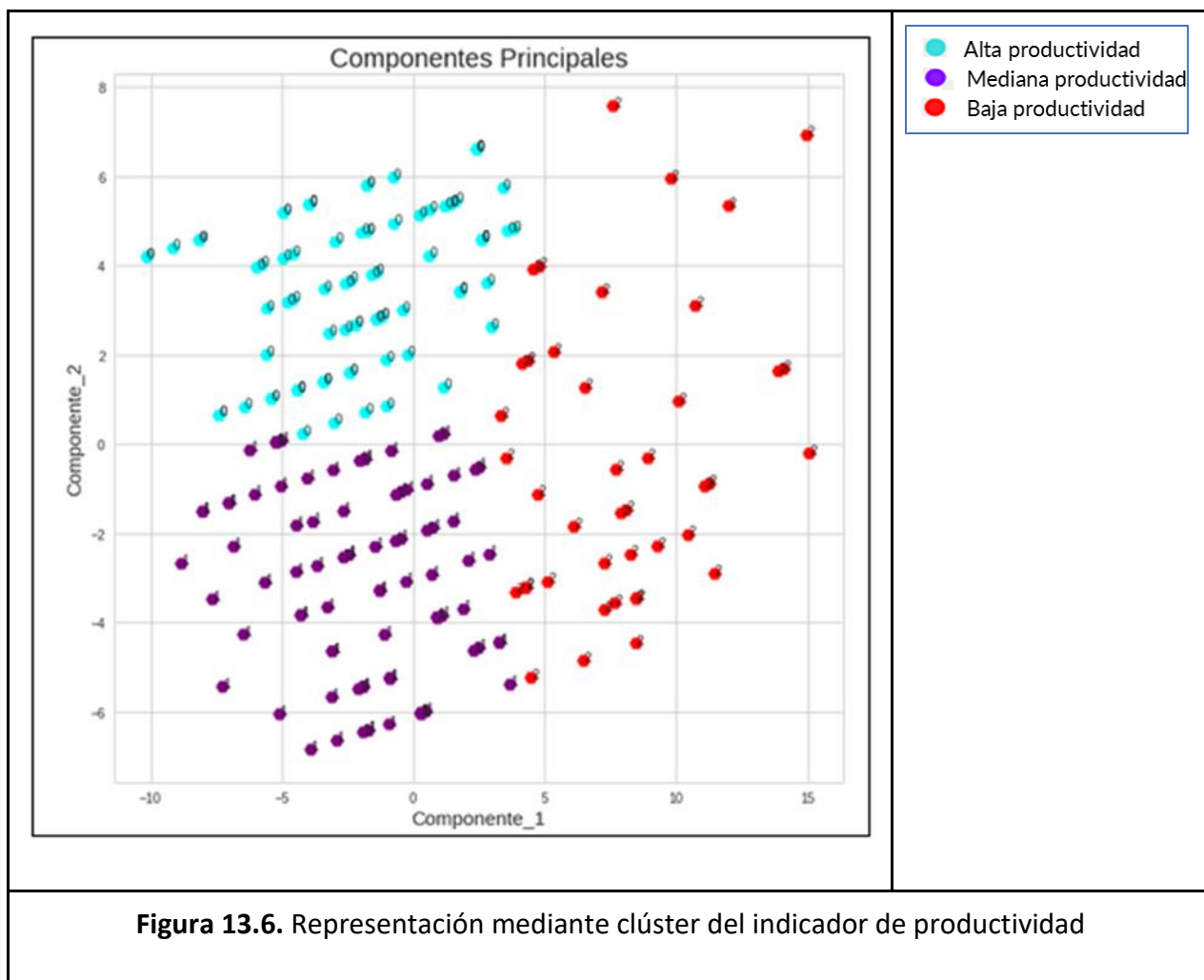
Posteriormente, a estos procesos de clasificación, se procede a realizar PCA para la reducción de dimensionalidad, y así proceder, de acuerdo a ciertas características del gráfico de correlación, encontrar un rango de categorías entre alta, media y baja productividad. La Figura 13.5 muestra el diagrama de correlación para el caso de productividad. En esta se puede observar que en la separación de cuadrantes de cada variable respecto a los productos generados por cada centro de formación. Existe mayor cantidad de productos en el segundo cuadrante del diagrama de correlaciones, lo que permite asociar que este cuadrante corresponde a un índice alto de producción científica, y que para el tercer cuadrante están únicamente dos tipos de productos. Esta hipótesis generada, es relativa, ya que depende netamente del interés institucional que permite evaluar a sus centros de formación entre alto, medio o bajo de acuerdo criterios generados por el ministerio de las ciencias, pero que para esta etapa hemos dado igualdad de peso.

En la Figura 13.6. se puede observar las gráficas de componentes principales de cada uno de los clusters generados por K-means. Si hacemos una comparación entre el círculo de correlaciones de la Figura 13.5 y Figura 13.6 podemos generar tres tipos de clusters, que se pueden clasificar entre alta, media y baja productividad, esto basado en la cantidad de

vectores del círculo de correlación. De este modo hemos generado una serie de alertas de productividad científica para los centros de formación de estudio.



este modo hemos generado una serie de alertas de productividad científica para los centros de formación de estudio.

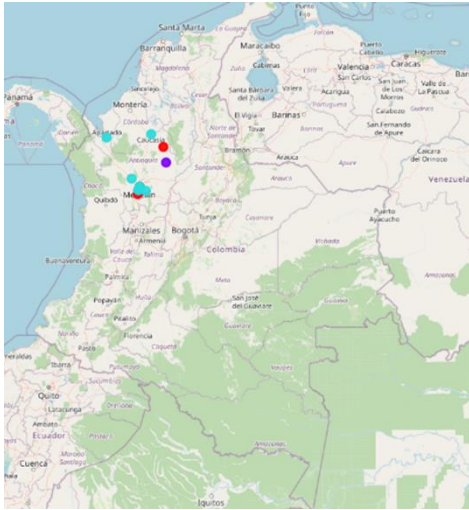
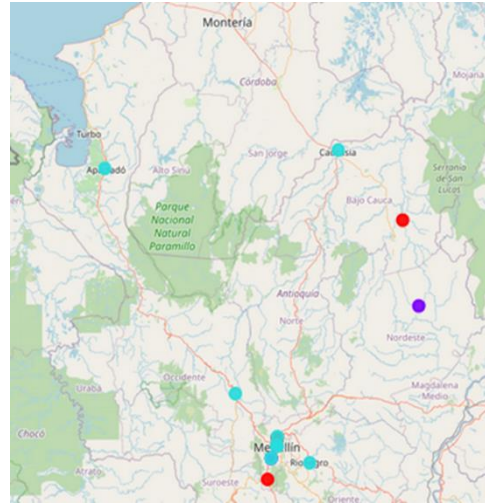


13.2.1.2 Mapas Geospaciales.

Generado la clasificación de centros de formación, se procede a generar los mapas georreferenciados que permiten clasificar cada uno de los centros de formación en medio, alto y bajo, tal como se muestra en la Figura 13.7.

a)

b)



c)

d)

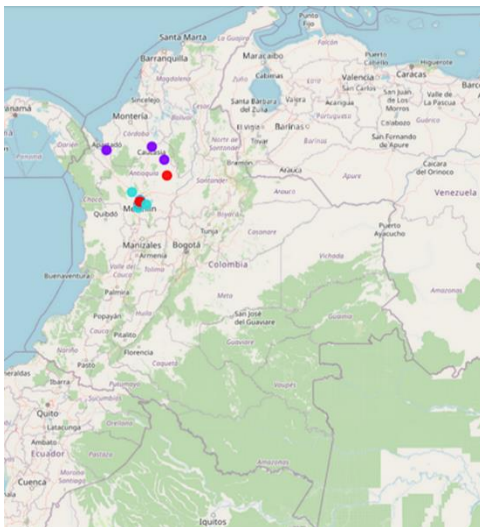
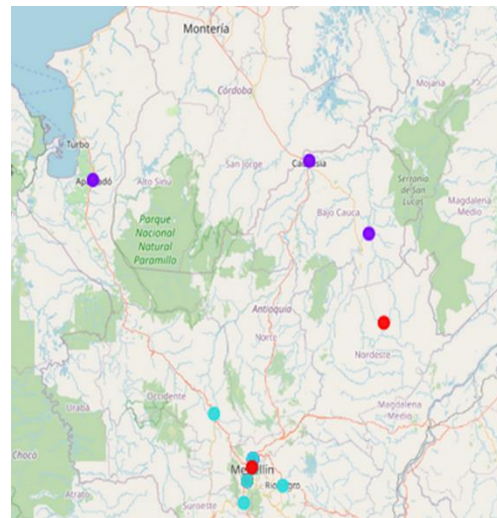


Figura 13.7. Representación geoespacial clúster del indicador productividad a) y b) 2018; y c) y d) 2021.



13.2.2. Mapas geospaciales para gestión de procesos de centros formación

Como mencionamos anteriormente, la generación de estos mapas se realiza en dos etapas: una etapa de clasificación a partir de herramientas de aprendizaje no supervisado, como la clusterización y PCA, y posteriormente se realiza una etapa de visualización de estos mapas.

13.2.2.1 Kmeans y PCA.

Para la selección de la cantidad de clusters adecuados, se aplicó el método del codo, que se visualiza en la Figura 13.8. Debido a nuestro interés de determinar un clasificador de tres alertas de alto, medio y bajo, se opta a seleccionar $k = 3$, que es también valido desde el punto de vista del método del codo.

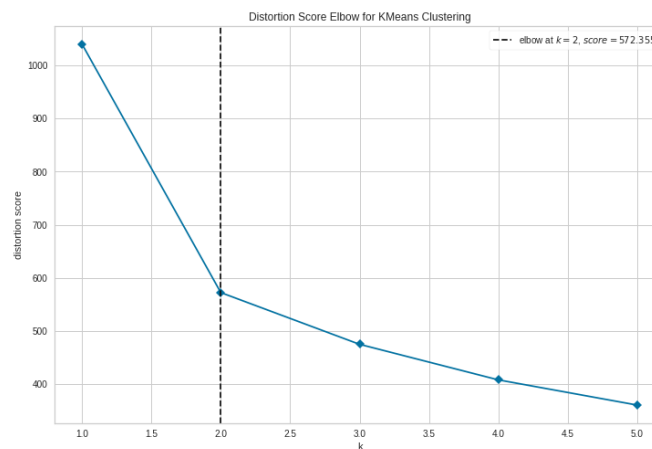


Figura 13.8. Método del codo para la determinación de cantidad de clusters en procesos de gestión.

Externos						Internos	
Indicador	Homogeneidad	Compleitud	Índice Rand	Medida V	Silueta	Calinski - Harabasz	Davies - Bouldin



Procesos	0,183	1,00	0	0,309	0,301	121,774	1,328
----------	-------	------	---	-------	-------	---------	-------

Tabla 13.8. Métricas de cluster.

Se determinan las métricas de cluster tanto internas como externas. Que se muestran en la Tabla 13.8 En se muestra la tendencia aleatoria de estos clusters, como consecuencia de la generación aleatoria de los datos. Se anexan códigos de proceso de clusterización a partir de K-Means (Ver Anexos).

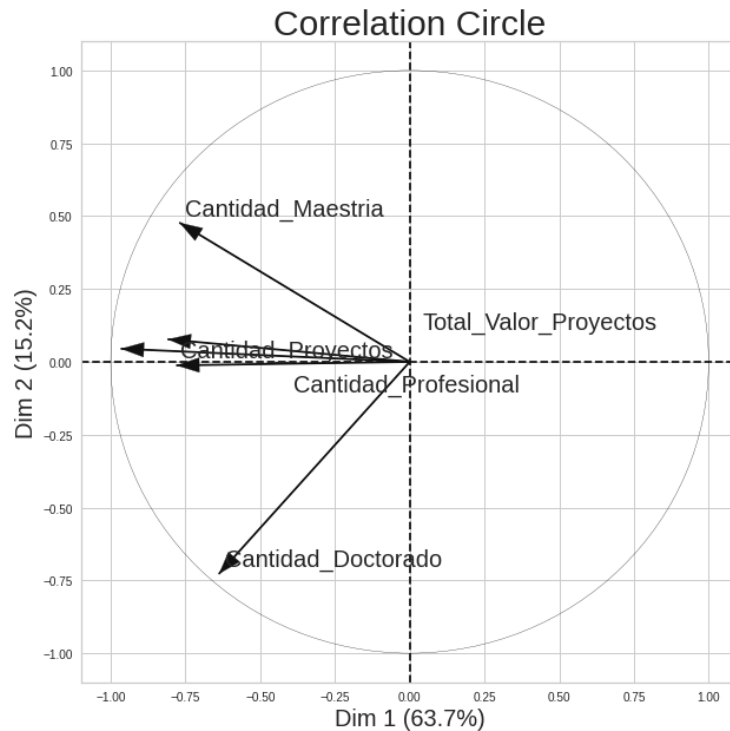
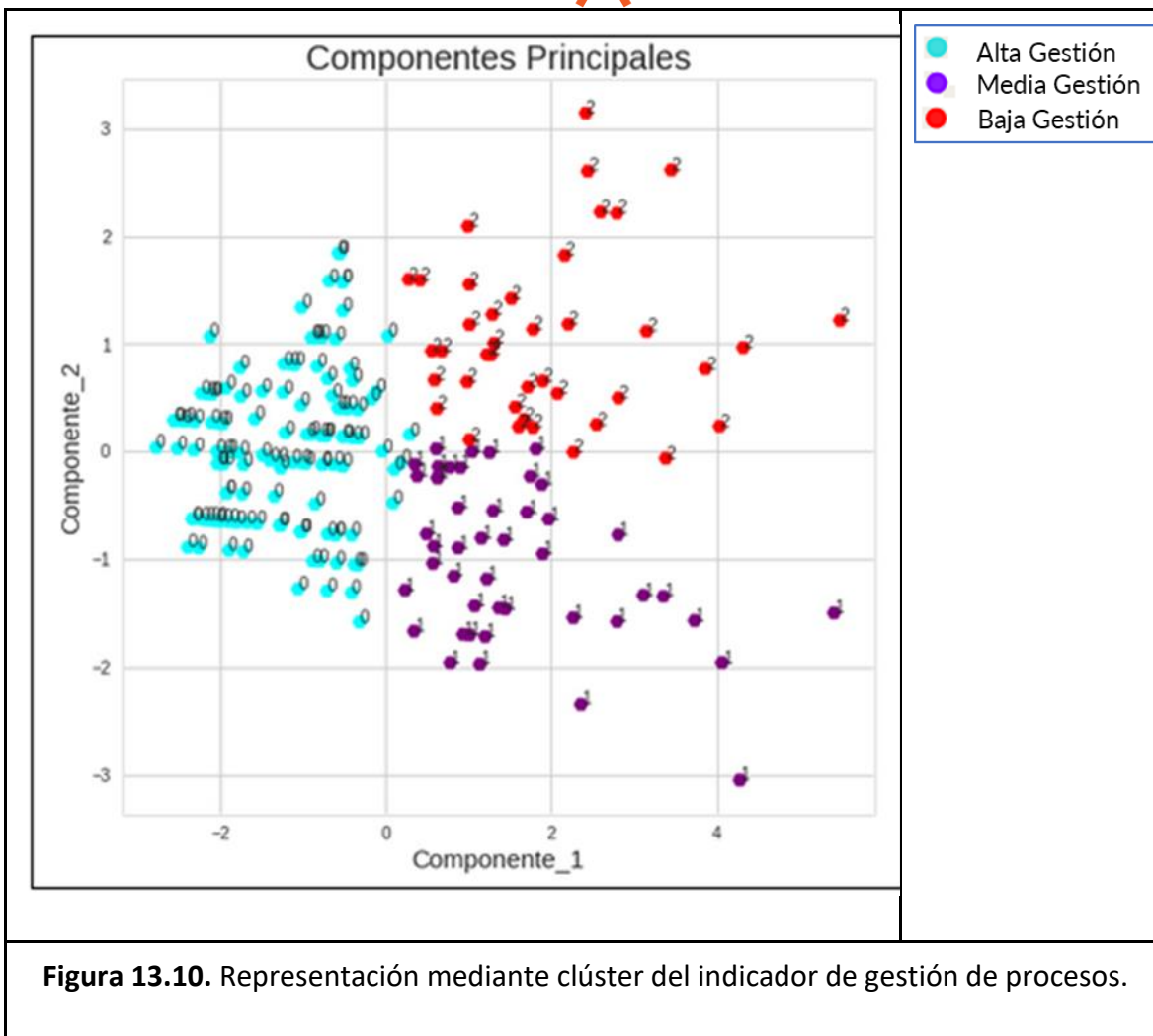


Figura 13.9. Círculo de correlaciones para procesos de centros de formación.



Posteriormente, a estos procesos de clasificación, se procede a realizar PCA para la reducción de dimensionalidad, y así proceder, de acuerdo a ciertas características del gráfico de correlación, encontrar un rango de categorías entre alta, media y baja gestión. La figura 13.9 muestra el diagrama de correlación para el caso de gestión de procesos. En esta se puede observar que, en la separación de cuadrantes de cada variable respecto a la cantidad de talento humano formado en los diferentes niveles educativo, así como con la cantidad de proyectos por cada centro de formación. Existe mayor cantidad de vectores en el segundo cuadrante del círculo de correlaciones, lo que permite asociar que este cuadrante corresponde a un índice de alta gestión, y que para el tercer cuadrante están únicamente un tipo de vector que es la variable cantidad de doctores que por cantidad correspondería a media gestión. Esta hipótesis generada, es relativa, ya que depende netamente del interés institucional que permite evaluar a sus centros de formación entre alto, medio o bajo de acuerdo criterios generados a nivel institucional y gubernamental, pero que para esta etapa hemos dado igualdad de peso.

En la Figura 13.10. se puede observar las gráficas de componentes principales de cada uno de los clusters generados por K-means. Si hacemos una comparación entre el círculo de correlaciones de la Figura 13.9 y Figura 10 podemos generar tres tipos de clusters, que se pueden clasificar entre alta, media y baja gestión, esto basado en la cantidad de vectores del círculo de correlación. De este modo hemos generado una serie de alertas de gestión de procesos para los centros de formación en estudio.



13.2.2.2 Mapas Geospaciales.

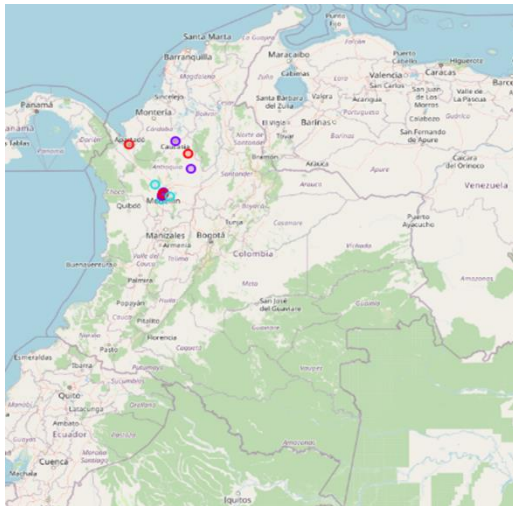
Generado la clasificación de centros de formación, se procede a generar los mapas georreferenciados que permiten clasificar cada uno de los centros de formación en alta, media y baja gestión, tal como se muestra en la Figura 13.11.

a)

b)



c)



d)

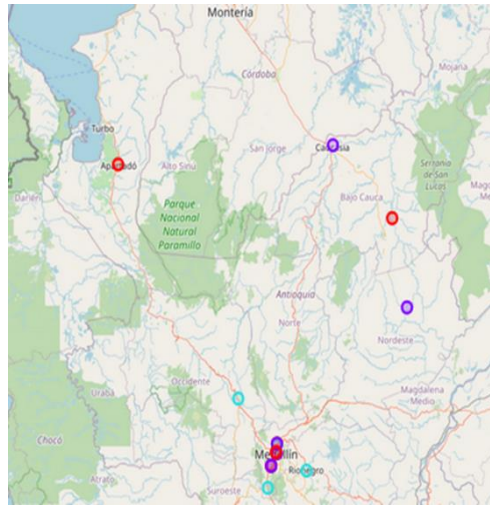


Figura 13.11. Representación geoespacial clúster del indicador gestión de procesos a) y b) 2018; y c) y d) 2021.



13.3. Modelos predictivos utilizando analítica de datos para medición de gestión y productividad por centros – Algoritmo KNN

Otros de los objetivos del proyecto fue generar un modelo predictivo que permita predecir, bajo las variables descritas en las TablaXX y la TablaXX, si el centro de formación entra en una categoría de alta, media o baja gestión o productividad respectivamente. Para ellos se hace implementación del algoritmo K-NN en R statistics

13.3.1. Modelos predictivos para medición de gestión de procesos en centros de formación – Algoritmo KNN

Se realiza el escalamiento de los datos

```
#=====
#Escalamiento de datos
#=====
datos_TPProceso <- scale(data_procesos[,c(2:6)])
View(datos_TPProceso)
#Se adiciona nuevamente la columna de etiqueta Centro-Año, Cluster y Gestion
datos_scale_TPProceso <- as.data.frame(cbind(data_procesos$Centro_Año, datos_TPProceso, data_procesos$cluster,
data_procesos$Gestion))
View(datos_scale_TPProceso)
str(datos_scale_TPProceso)
dim(datos_scale_TPProceso)
datos_scale_TPProceso$Q_Proyectos <- as.numeric(datos_scale_TPProceso$Q_Proyectos)
datos_scale_TPProceso$QNA_5007 <- as.numeric(datos_scale_TPProceso$QNA_5007)
datos_scale_TPProceso$QNA_5009 <- as.numeric(datos_scale_TPProceso$QNA_5009)
datos_scale_TPProceso$QNA_5010 <- as.numeric(datos_scale_TPProceso$QNA_5010)
datos_scale_TPProceso$TV_Proyectos <- as.numeric(datos_scale_TPProceso$TV_Proyectos)
str(datos_scale_TPProceso)
names(datos_scale_TPProceso) <- names(data_procesos)
View(datos_scale_TPProceso)
row.names(datos_scale_TPProceso) <- datos_scale_TPProceso$Centro_Año
View(datos_scale_TPProceso)
proceso <- datos_scale_TPProceso[, c(-1,-7)]
View(proceso)
#=====
```

Figura 13.12. Escalamiento de datos.

Se transforma la variable Gestión en una variable factor de tres niveles

```
#=====
# La variable factor está indexada en la 6ª posición
#=====

proceso$Gestion <- as.factor(proceso$Gestion)
View(proceso)
str(proceso)
#=====
```



```
> str(proceso)
'data.frame': 208 obs. of 6 variables:
 $ Q_Proyectos : num -1.257 -0.362 -1.257 0.534 -0.362 ...
 $ QNA_5007 : num -1.103 -0.689 0.137 0.55 0.137 ...
 $ QNA_5009 : num -0.8238 -0.0502 -1.2106 0.3366 -0.437 ...
 $ QNA_5010 : num -1.108 -0.225 -0.225 0.658 -0.225 ...
 $ TV_Proyectos: num -0.126 -0.733 -0.206 -0.607 -1.573 ...
 $ Gestion : Factor w/ 3 levels "Alta Gestion",...: 2 2 2 1 2 3 2 2 3 2 ...
```

Figura 13.13 Transformación de variable Gestión en un factor de tres niveles.

Se obtiene una tabla, donde se evidencia la cantidad de observaciones por nivel de gestión

```
#=====
table(proceso$Gestion)
# Alta Gestion Baja Gestion Media Gestion
# 47 122 39
#=====
```

```
> table(proceso$Gestion)

Alta Gestion Baja Gestion Media Gestion
47 122 39
```

Figura 13.14 Cantidad de observaciones por nivel de gestión

```
#=====
# Porcentaje de centros
#=====
round(prop.table(table(proceso$Gestion)) * 100,digits = 1)
# Alta Gestion Baja Gestion Media Gestion
# 22.6 58.7 18.8
#=====
```

Figura 13.15. Porcentaje de centros por nivel de gestión



Dadas estas etapas, se genera el modelo de analítica de datos utilizando KNN. Con esto se obtienen la siguiente matriz de confusión.

```
> Validacion_modelo_procesos <- CrossTable(x = data_procesos_test_labels, y = data_procesos_test_predi, prop.chisq=TRUE)
```

Cell Contents				

Chi-square contribution	N			
	N / Row Total			
	N / Col Total			
	N / Table Total			

Total Observations in Table: 62

	data_procesos_test_predi			
data_procesos_test_labels	Alta Gestion	Baja Gestion	Media Gestion	Row Total

Alta Gestion	16	0	1	17
	24.805	9.323	1.107	
	0.941	0.000	0.059	0.274
	0.889	0.000	0.100	
	0.258	0.000	0.016	

Baja Gestion	1	34	0	35
	8.260	11.422	5.645	
	0.029	0.971	0.000	0.565
	0.056	1.000	0.000	
	0.016	0.548	0.000	

Media Gestion	1	0	9	10
	1.248	5.484	33.833	
	0.100	0.000	0.900	0.161
	0.056	0.000	0.900	
	0.016	0.000	0.145	

Column Total	18	34	10	62
	0.290	0.548	0.161	

Figura 13.16. Matriz de confusión para el algoritmo K-NN en R, con k=3



```
> con=confusionMatrix(testData[,ncol(testData)], data_procesos_test_pred1, mode = "everything")
> con
Confusion Matrix and Statistics

          Reference
Prediction  Alta Gestion Baja Gestion Media Gestion
Alta Gestion      16          0          1
Baja Gestion       1         34          0
Media Gestion      1          0          9

Overall Statistics

          Accuracy : 0.9516
          95% CI   : (0.865, 0.9899)
    No Information Rate : 0.5484
    P-value [Acc > NIR] : 1.495e-12

          Kappa : 0.9173

  Mcnemar's Test P-Value : NA

Statistics by Class:

               Class: Alta Gestion Class: Baja Gestion Class: Media Gestion
Sensitivity                0.8889                1.0000                0.9000
Specificity                0.9773                0.9643                0.9808
Pos Pred Value              0.9412                0.9714                0.9000
Neg Pred Value              0.9556                1.0000                0.9808
Precision                   0.9412                0.9714                0.9000
Recall                      0.8889                1.0000                0.9000
F1                          0.9143                0.9855                0.9000
Prevalence                  0.2903                0.5484                0.1613
Detection Rate              0.2581                0.5484                0.1452
Detection Prevalence        0.2742                0.5645                0.1613
Balanced Accuracy           0.9331                0.9821                0.9404
```

Figura 13.17. Matriz de confusión para el algoritmo K-NN en R, con k=3, y sus respectivas métricas

Para nuestro caso la tabla de contingencia arrojada por la matriz de confusión consta de $N=62$ observaciones clasificadas en forma cruzada en las variables de fila y columna, que normalmente tendrían los mismos niveles, sin embargo, el estadístico de McNemar se usa para probar la hipótesis nula de simetría, es decir, que la probabilidad de que una observación se clasifique en celda $[i,j]$ es igual a la probabilidad de ser clasificado en la celda $[j,i]$, (mcnemar.test)

Para nuestro caso al obtener un P-Value: NA, es un indicio de que la clasificación está fallando, por lo tanto, debemos calcular cual es el óptimo k-vécin cercano en nuestra base de datos, como se describe a continuación.

Hasta el momento solo hemos utilizado un solo valor para k, 3 vecinos más cercanos. Sin embargo, podríamos ver la exactitud o *accuracy* para diferentes valores de k. Como

tenemos 146 valores en nuestra data de
tomando máximo 146 vecinos más



entrenamiento veremos los aciertos
cercanos.

```
#=====
k <- 1:146
resultado <- data.frame(k, precision = 0)

for(n in k){
  proceso_test_output_knn <- knn(train = trainData1,
                                cl = data_procesos_train_labels,
                                test = testData1,
                                k = n)

  resultado$precision[n] <- mean(proceso_test_output_knn == testData[,6])
}
resultado
library(extrafont)
#resultado %>%
ggplot(resultado,aes(x=k, y=precision))+
  labs(title = 'Evolución de la precision segun valor de k')+
  labs(subtitle = 'Precision asociados a valores distintos de k')+
  geom_line(color="red")+ geom_point()+
  scale_x_continuous(breaks=pretty(resultado$k, n = 10), name="k")+
  theme(legend.position="top", legend.text = element_text(colour="blue", size=6)) +
  geom_text(aes(label=k),hjust = 0, nudge_x = 0.05, color="blue",size=3)+
  theme(text=element_text(size=12, family="Calibri"))+
  theme_bw()
#=====
```

Figura 13.18. Fragmento de código modelo KNN para procesos de gestión en centros de formación.

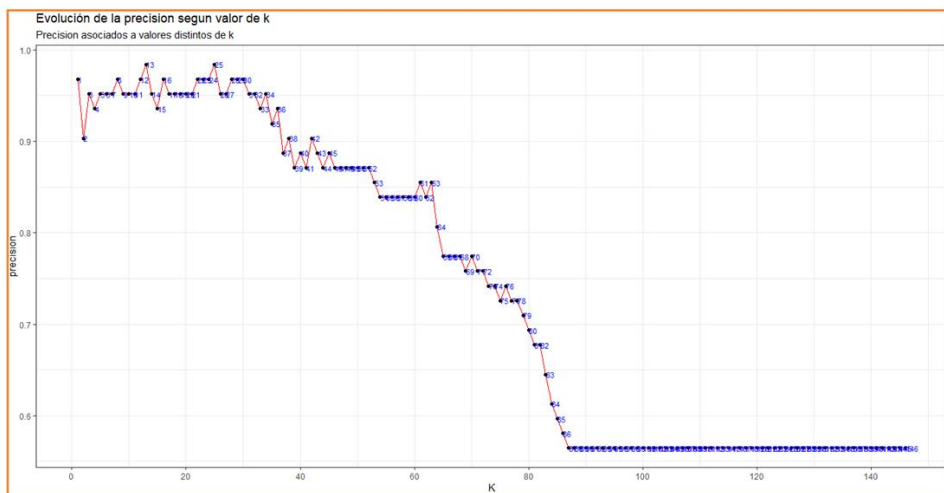


Figura 13.19. Precisión del modelo para diferentes valores de k



Como vemos, para este caso, a partir de cierto número de vecinos más cercanos el nivel de aciertos de nuestro algoritmo empieza a reducirse. Dependerá de cada caso el escoger el mejor “k” para nuestro modelo, para el caso se elige $k=2$

```
> con
Confusion Matrix and Statistics

Prediction      Reference
Alta Gestion    Alta Gestion Baja Gestion Media Gestion
Baja Gestion    15          0          2
Baja Gestion    1          33         1
Media Gestion    2          0          8

Overall Statistics

Accuracy : 0.9032
95% CI : (0.8012, 0.9637)
No Information Rate : 0.5323
P-Value [Acc > NIR] : 3.354e-10

Kappa : 0.8363

McNemar's Test P-value : 0.5724

Statistics by Class:

Class: Alta Gestion Class: Baja Gestion Class: Media Gestion
Sensitivity          0.8333          1.0000          0.7273
Specificity          0.9545          0.9310          0.9608
Pos Pred Value       0.8824          0.9429          0.8000
Neg Pred Value       0.9333          1.0000          0.9423
Precision            0.8824          0.9429          0.8000
Recall               0.8333          1.0000          0.7273
F1                   0.8571          0.9706          0.7619
Prevalence           0.2903          0.5323          0.1774
Detection Rate       0.2419          0.5323          0.1290
Detection Prevalence 0.2742          0.5645          0.1613
Balanced Accuracy     0.8939          0.9655          0.8440
```

Figura 13.20. Matriz de confusión para el algoritmo K-NN en R, con $k=2$, y sus respectivas métricas

Evaluación del rendimiento del modelo: métricas asociadas al clasificador K-NN, con $k=2$

```
> clasificador_metricas
Sensitivity Specificity Precision Recall F1 Accuracy
Metricas_del_Clasificador 0.9103128 0.9422794 0.9032258 0.9103128 0.9058236 0.9032258
```

Hemos clasificado 15 centros de Alta Gestion como Alta Gestion, lo que indica que está bien clasificado y 2 centros de Alta Gestion como de Media Gestion, lo que indica una incorrecta clasificación. A su vez analizando los centros en Baja Gestion la predicción en la clasificación arroja que 33 centros de Baja Gestion quedaron clasificados como de Baja Gestion lo cual indica que quedaron bien clasificados. Por otro lado, 1 centro siendo de Baja Gestion quedo clasificado como que fuera de Alta Gestion, asimismo otro centro siendo de Baja Gestion

quedo clasificado en Media Gestion. Por último 8 centros de Media Gestion fueron clasificados como de Media Gestion, lo cual indica una correcta clasificación y 2 centros de pertenecientes a Media Gestion quedaron clasificados como si fueran de Alta Gestion, lo cual indica que esta incorrectamente clasificado



13.3.2. Modelos predictivos para medición de productividad en centros de formación – Algoritmo KNN

Se realiza el escalamiento de los datos

```
#=====
# La variable factor está indexada en la 7ª posición
#=====

productividad$Nivel_Productividad <- as.factor(productividad$Nivel_Productividad)
view(productividad)
str(productividad)

#=====

> str(productividad)
'data.frame': 208 obs. of 7 variables:
 $ P_8020      : num -0.559 -0.559 -1.187 -1.187 2.582 ...
 $ P_8025      : num 0.855 -1.271 -1.271 -0.562 0.147 ...
 $ P_8019      : num -0.559 -0.559 -1.187 -1.187 2.582 ...
 $ P_8039      : num 0.855 -1.271 -1.271 -0.562 0.147 ...
 $ P_8016      : num -0.559 -0.559 -1.187 -1.187 2.582 ...
 $ P_8031      : num 0.855 -1.271 -1.271 -0.562 0.147 ...
 $ Nivel_Productividad: Factor w/ 3 levels "Alta Productividad",...: 1 2 2 2 1 2 2 3 2 2 ...
```

Figura 13.21. Escalamiento de datos.

Se transforma la variable Gestión en una variable factor de tres niveles

```
#=====
# La variable factor está indexada en la 7ª posición
#=====

productividad$Nivel_Productividad <- as.factor(productividad$Nivel_Productividad)
view(productividad)
str(productividad)

#=====
```



```
> str(productividad)
'data.frame': 208 obs. of 7 variables:
 $ P_8020      : num -0.559 -0.559 -1.187 -1.187 2.582 ...
 $ P_8025      : num 0.855 -1.271 -1.271 -0.562 0.147 ...
 $ P_8019      : num -0.559 -0.559 -1.187 -1.187 2.582 ...
 $ P_8039      : num 0.855 -1.271 -1.271 -0.562 0.147 ...
 $ P_8016      : num -0.559 -0.559 -1.187 -1.187 2.582 ...
 $ P_8031      : num 0.855 -1.271 -1.271 -0.562 0.147 ...
 $ Nivel_Productividad: Factor w/ 3 levels "Alta Productividad",...: 1 2 2 2 1 2 2 3 2 2 ...
```

Figura 13.22. Transformación de variable Nivel_Productividad en un factor de tres niveles.

Se obtiene una tabla, donde se evidencia la cantidad de observaciones por nivel de productividad.

```
#=====
table(productividad$Nivel_Productividad)
# Alta Productividad  Baja Productividad Media Productividad
#           39           112           57
#=====

> round(prop.table(table(productividad$Nivel_Productividad)) * 100,digits = 1)

Alta Productividad  Baja Productividad Mediana Productividad
           18.8           53.8           27.4
```

Figura 13.23. Cantidad de observaciones por nivel de gestión

```
> round(prop.table(table(productividad$Nivel_Productividad)) * 100,digits = 1)

Alta Productividad  Baja Productividad Mediana Productividad
           18.8           53.8           27.4
```

Figura 13.24. Porcentaje de centros por nivel de productividad



Dadas estas etapas, se genera el modelo de analítica de datos utilizando KNN. Con esto se obtienen la siguiente matriz de confusión.

```
> library(gmodels)
> validacion_modelo_productividad <- CrossTable(x = data_productividad_test_labels, y = data_productividad_test_pred1, prop.chisq=TRUE)
```

cell contents				
	N			
Chi-square contribution				
N / Row Total				
N / Col Total				
N / Table Total				

Total Observations in Table: 74

data_productividad_test_labels	data_productividad_test_pred1			Row Total
	Alta Productividad	Baja Productividad	Mediana Productividad	
Alta Productividad	2 2.211 0.111 0.667 0.027	8 1.157 0.444 0.167 0.108	8 1.034 0.444 0.348 0.108	18 0.243
Baja Productividad	1 0.167 0.027 0.333 0.014	35 5.042 0.946 0.729 0.473	1 9.587 0.027 0.043 0.014	37 0.500
Mediana Productividad	0 0.770 0.000 0.000 0.000	5 4.353 0.263 0.104 0.068	14 11.095 0.737 0.609 0.189	19 0.257
Column Total	3 0.041	48 0.649	23 0.311	74

Figura 13.25. Matriz de confusión para el algoritmo K-NN en R, con k=3



```
> con=confusionMatrix(testData[,ncol(testData)], data_productividad_test_pred1, mode = "everything")
> con
Confusion Matrix and Statistics

              Reference
Prediction    Alta Productividad Baja Productividad Mediana Productividad
Alta Productividad      2              8              8
Baja Productividad      1             35              1
Mediana Productividad    0              5             14

Overall Statistics

          Accuracy : 0.6892
          95% CI   : (0.571, 0.7917)
    No Information Rate : 0.6486
    P-value [Acc > NIR] : 0.274045

          Kappa : 0.4696

  McNemar's Test P-value : 0.001076

Statistics by Class:

               Class: Alta Productividad Class: Baja Productividad Class: Mediana Productividad
Sensitivity                0.66667                0.7292                0.6087
Specificity                0.77465                0.9231                0.9020
Pos Pred value             0.11111                0.9459                0.7368
Neg Pred value             0.98214                0.6486                0.8364
Precision                  0.11111                0.9459                0.7368
Recall                     0.66667                0.7292                0.6087
F1                         0.19048                0.8235                0.6667
Prevalence                 0.04054                0.6486                0.3108
Detection Rate             0.02703                0.4730                0.1892
Detection Prevalence       0.24324                0.5000                0.2568
Balanced Accuracy          0.72066                0.8261                0.7553
```

Figura 13.26. Matriz de confusión para el algoritmo K-NN en R, con k=3, y sus respectivas métricas

Hasta el momento solo hemos utilizado un solo valor para k, 3 vecinos más cercanos. Sin embargo, podríamos ver la exactitud o *accuracy* para diferentes valores de k. Como tenemos 134 valores en nuestra data de entrenamiento veremos los aciertos tomando máximo 134 vecinos más cercanos.



```
#=====
k <- 1:134
resultado <- data.frame(k, precision = 0)

for(n in k){
  productividad_test_output_knn <- knn(train = trainData1,
                                       cl = data_productividad_train_labels,
                                       test = testData1,
                                       k = n)

  resultado$precision[n] <- mean(productividad_test_output_knn == testData[,7])
}
resultado
library(extrafont)
#resultado %>%
ggplot(resultado,aes(x=k, y=precision))+
  labs(title = 'Evolución de la precision segun valor de k')+
  labs(subtitle = 'Precision asociados a valores distintos de k')+
  geom_line(color="red")+ geom_point()+
  scale_x_continuous(breaks=pretty(resultado$k, n = 10), name="K")+
  theme(legend.position="top", legend.text = element_text(colour="blue", size=6)) +
  geom_text(aes(label=k),hjust = 0, nudge_x = 0.05, color="blue",size=3)+
  theme(text=element_text(size=12, family="Calibri"))+
  theme_bw()
#=====
```

Figura 13.27. Fragmento de código modelo KNN para productividad en centros de formación.

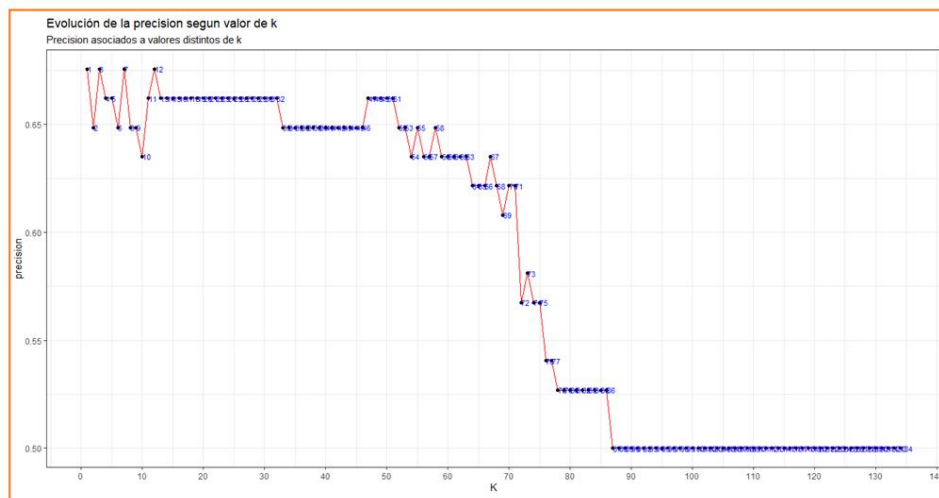


Figura 13.28. Precisión del modelo para diferentes valores de k

Como vemos, para este caso, a partir de cierto número de vecinos más cercanos el nivel de aciertos de nuestro algoritmo empieza a reducirse. Dependerá de cada caso el escoger el

mejor “k” para nuestro modelo, para  nuestro caso k=3, ofrece la mejor precisión para nuestro modelo.

```
> con
Confusion Matrix and Statistics

Prediction      Reference
Alta Productividad 1      8      9
Baja Productividad 0     36     1
Mediana Productividad 0      7     12

Overall Statistics

Accuracy : 0.6622
95% CI : (0.5428, 0.7681)
No Information Rate : 0.6892
P-value [Acc > NIR] : 0.738

Kappa : 0.4133

McNemar's Test P-Value : 8.288e-05

Statistics by Class:

Class: Alta Productividad Class: Baja Productividad Class: Mediana Productividad
Sensitivity      1.00000      0.7059      0.5455
Specificity      0.76712      0.9565      0.8654
Pos Pred Value   0.05556      0.9730      0.6316
Neg Pred Value   1.00000      0.5946      0.8182
Precision        0.05556      0.9730      0.6316
Recall          1.00000      0.7059      0.5455
F1              0.10526      0.8182      0.5854
Prevalence      0.01351      0.6892      0.2973
Detection Rate   0.01351      0.4865      0.1622
Detection Prevalence 0.24324      0.5000      0.2568
Balanced Accuracy 0.88356      0.8312      0.7054
```

Figura 13.29. Matriz de confusión para el algoritmo K-NN en R, con k=7, y sus respectivas métricas

Evaluación del rendimiento del modelo: métricas asociadas al clasificador K-NN, con k=3

```
> clasificador_mtricas
Sensitivity Specificity Precision Recall F1 Accuracy
Metricas_del_Clasificador 0.9103128 0.9422794 0.9032258 0.9103128 0.9058236 0.9032258
```

Hemos clasificado 2 centros de Alta Productividad como Alta Productividad, lo que indica que está bien clasificado, 8 centros que siendo de Alta Productividad como de Baja Productividad y 8 centros que siendo de Alta productividad como de Mediana Productividad, lo que indica una incorrecta clasificación para estos. A su vez analizando los centros de Baja Productividad tenemos que la predicción en la clasificación arroja que 35 centros de Baja Productividad quedaron clasificados como de Baja Productividad lo cual indica que quedaron bien clasificados. Por otro lado, 1 centro siendo de Baja Productividad quedo clasificado como que fuera de Alta Productividad, asimismo otro centro siendo de Baja Productividad quedo clasificado en Mediana Productividad, lo que indica una

incorrecta clasificación para estos. Por último 14 centros de Mediana Productividad fueron clasificados como de Mediana Productividad, lo cual indica una correcta clasificación y 5 centros pertenecientes a Mediana Productividad quedaron clasificados como si fueran de Baja Productividad, lo cual indica que esta incorrectamente clasificado.



14. LECCIONES APRENDIDAS

Se considera relevante el levantamiento de datos

15. PARTICIPACIÓN EN EVENTOS

16. CONCLUSIONES

Sensibilidad de acuerdo al valor obtenido del 0.6830322, esto me indica **¿qué porcentaje de los centros de acuerdo a su nivel de productividad somos capaces de identificar?** Para lo cuál diríamos que el 68,30% de los centros se identifican acertadamente de acuerdo al nivel de productividad.

El valor **F1** se utiliza para combinar las medidas de **precision** y **recall (sensibilidad)** en un sólo valor. Asume que nos importa de igual forma la precisión y la exhaustividad. Esto no tiene que ser así en todos los problemas. Ya que necesitamos predecir si el nivel de productividad del centro se detecto adecuadamente, por lo tanto podemos valorar más la sensibilidad a riesgo de tener una peor precisión. Sin embargo para nuestro caso el F1 es igual 0,6292679 equivalente al 62,93% que no es muy aproximado al 68,30% de sensibilidad del modelo y a la precisión del modelo equivalente a 68,92%. Sin embargo el F1 permite comparar modelos donde si un modelo arroja un F1 alto representa el modelo que mejor puede clasificar las observaciones en clases.

Para nuestro caso un F1 alto indicaría que el modelo sería capaz de clasificar las observaciones de la clase nivel de productividad de manera más precisa.

Exactitud (Accuracy): La exactitud (accuracy) mide el porcentaje de casos que el modelo ha acertado. Es decir, el modelo acierta el 68,92% de las veces. Como se observa, la exactitud es una métrica muy engañosa. De hecho, si tuviésemos un modelo que siempre predijera que el centro siempre va a estar en un nivel de productividad determinado, su accuracy

podría ser del 80% o más y esto  normalmente se observa en problemas con clases desbalanceadas por lo tanto es mucho mejor usar precision, recall y F1. Estas métricas dan una mejor idea de la calidad del modelo. IArtificial.net. (2020).

Especificidad: el modelo tiene una alta capacidad para identificar los centros que no fueron clasificados correctamente según su nivel de productividad entre todos los centros cuyos niveles de productividad fueron clasificados correctamente. Para nuestro caso la especificidad del clasificador es de 88,16%.

De lo anterior tenemos, el modelo es mucho más específico que sensible en la clasificación esto se traduce en que si un falso positivo significa que el modelo tiene una muy alta capacidad para identificar los centros que no fueron clasificados correctamente según su nivel de productividad evitando los falsos positivos.

17. REFERENCIAS

[1] mcnemar.test. McNemar's Chi-Square Test for Count Data.

https://docs.tibco.com/pub/enterprise-runtime-for-R/3.1.1/doc/html/Language_Reference/stats/mcnemar.test.html

[2] Fleiss, J. L. (1981). Statistical Methods for Rates and Proportions, 2nd ed. New York: Wiley.

[3] Zar, J. H. (1984). Biostatistical Analysis, 2nd ed. Englewood Cliffs: Prentice-Hall.

Alan Agresti (1990). Categorical data analysis. New York: Wiley. Pages 350--354.

IArtificial.net. (2020). <https://www.iartificial.net/precision-recall-f1-accuracy-en-clasificacion/>



ANEXOS 3. Apoyo a la formación - Curso de estadística aplicada a inteligencia artificial.

☰ Análisis estadístico de datos

Datasets para ejercitar JOHN JAIRO CASTRO MALDONADO · 22 abr (Última modificación: 19 ago)

Datos_Usa. Referencia: <https://www.kaggle.com/code/carlolepelaars/statistics-tutorial/data>

	flores.xlsx Excel		CeramicasPorosasAF.xlsx Excel
	Datos_USA.xlsx Excel		pecesr.txt Texto

Comentarios de la clase

Añade un comentario de clase...

Figura A3.1. Soporte de envío de apoyo a la formación: actividad de estadística básica: Curtosis y coeficiente de asimetría.

meet.google.com/wjr-eunu-oxp?authuser=1

Participants: JOHN JAIRO CASTRO MALDONADO, felipe gutierrez, RICARDO URIBE GOMEZ, nora del carmen arango tabares, Doris Elena Monsalve Sossa, Nicolas Emilio Velez Perra, Oscar David Luna Julio, 18 más

Mensajes de la llamada:

- Permitir a todo el mundo enviar mensajes ☒
- Los mensajes solo se muestran a los participantes de la llamada y se eliminan cuando termina
- Juan Guillermo Ruiz Zapata 17:54: seleccionar el ultimo que se llama estadística
- fabian Jaramillo Coronado 17:54: Al inicio del curso hablaban de un grupo de whatsapp quien tiene la administración del grupo, me pueden incluir?
- Claudia Bustamante 17:55: gracias
- nora del carmen arango tabares 17:55: Fabian me puedes escribir al 3004544540 para incluirte

Envía un mensaje a todos

Figura A3.2. Soporte de envío de apoyo a la formación. Realización de clases por medio de la plataforma Meet de Google.



ANEXO 4. Analítica de datos – Encuestas. Soporte de correos, entrega de resultados en código.



Figura A4.1. Recepción de datos para analítica de información.



Figura A4.2. Reunión técnica de avances de desarrollo de código en analítica de información. Desarrollada en 2022-08-08. **Parte 1.**

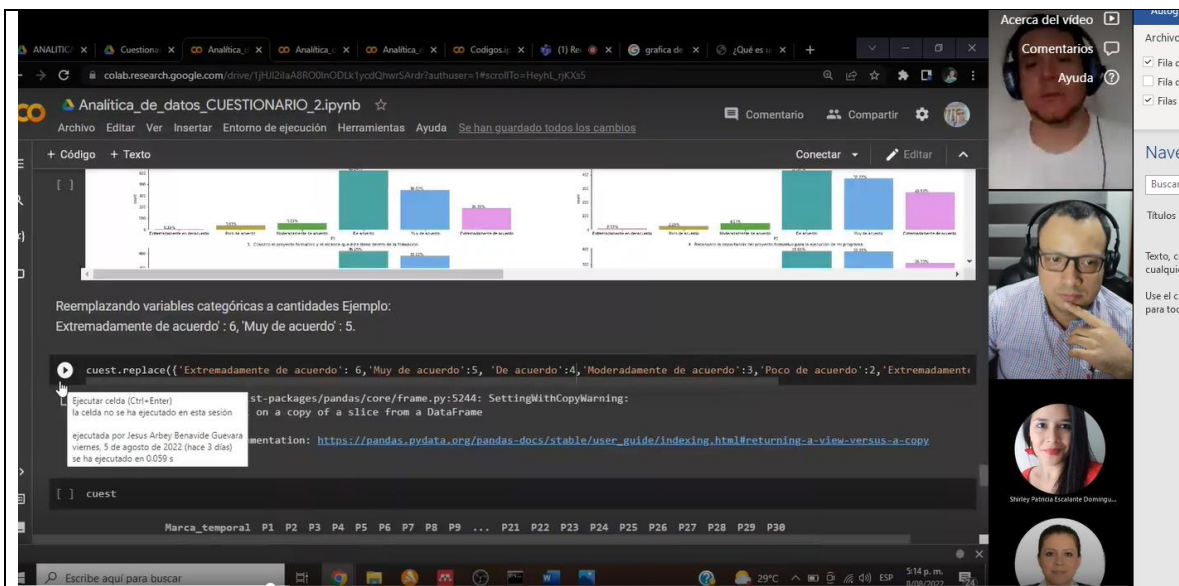


Figura A4.3. Reunión técnica de avances de desarrollo de código en analítica de información. Desarrollada en 2022-08-08. **Parte 2.**



Llamada de Teams

Categorizar Deshacer

CURSO DE ESTADÍSTICA Ahora

Avances Analítica de datos - Encuestas

Pública Clasificada\Todos los empleados (Sin Restricción)

Jesus Arbey Benavides Guevara

Para: Luz Carime Chia Cepeda

CC: John Jairo Castro Maldonado

3 MB 5 MB

4 archivos adjuntos (11 MB) Guardar todo en OneDrive - Servicio Nacional de Aprendizaje Descargar todo

Estimada

Luz Carime Chia,

Instructora - Centro de Comercio y Servicios

Reciba un cordial saludo. Comparto por este medio los avances de la analítica de las encuestas realizadas por su equipo. Sigue en anexo:

1. Link de códigos y datos: <https://bit.ly/3K7IKFS>
2. Pdf de códigos de los resultados avanzados hasta el momento.

Por favor, cualquier inquietud no dude en entrar en contacto y quedo atento a sus indicaciones.

Cordialmente,

Jesus Arbey Benavides Guevara
Investigador SENNOVA - Centro de Servicios y Gestión Empresarial,
jabenavides@sena.edu.co
3105595765
Calle 51 # 59 -70 Torre Norte, Piso 5 -Medellín

Figura A4.4. Soporte de envío de resultados de analítica de datos vía e-mail, para supervisor Ing. John Jairo Castro e instructora Luz Carime Chia Sena- Regional Cauca.



ANEXOS 5. Códigos de apoyo en analítica de datos - Centro de Comercio y Servicios- SENA Cauca

▼ Analítica de datos CUESTIONARIO 1 - Instructores

```
pip install factor_analyzer
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public
Requirement already satisfied: factor_analyzer in /usr/local/lib/python3.7/dist-packages
Requirement already satisfied: numpy in /usr/local/lib/python3.7/dist-packages (from factor_analyzer)
Requirement already satisfied: pandas in /usr/local/lib/python3.7/dist-packages (from factor_analyzer)
Requirement already satisfied: scipy in /usr/local/lib/python3.7/dist-packages (from factor_analyzer)
Requirement already satisfied: scikit-learn in /usr/local/lib/python3.7/dist-packages (from factor_analyzer)
Requirement already satisfied: python-dateutil>=2.7.3 in /usr/local/lib/python3.7/dist-packages (from pandas)
Requirement already satisfied: pytz>=2017.3 in /usr/local/lib/python3.7/dist-packages (from pandas)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.7/dist-packages (from pandas)
Requirement already satisfied: threadpoolctl>=2.0.0 in /usr/local/lib/python3.7/dist-packages (from scikit-learn)
Requirement already satisfied: joblib>=0.11 in /usr/local/lib/python3.7/dist-packages (from scikit-learn)
```

```
import pandas as pd
import numpy as np
#Gráfica
import matplotlib.pyplot as plt
import seaborn as sns
import plotly.express as px
from plotly.subplots import make_subplots
import plotly.figure_factory as ff
import plotly.offline as offline
import plotly.graph_objs as go
```

```
from google.colab import drive
drive.mount('/content/drive')
```

```
Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount(
```

```
from google.colab import files
path = "/content/drive/MyDrive/Proyecto SGPS 9748/Analítica Luz Karime_ Regional Cauca/Analítica"
#Dataframes
cuest= pd.read_excel(path)

cuest.head()
```


		1. Existe un documento formal por parte de la Coordinación Académica para la designación del equipo ejecutor de la ficha asociada al proyecto formativo.	2. El proyecto formativo se formula en conjunto por los miembros del equipo ejecutor asociado a la ficha del programa de formación.	3. La formulación del proyecto formativo se realiza teniendo en cuenta el contexto y necesidades del sector productivo.	4. El proyecto formativo es evaluado previa revisión del Equipo Pedagógico de Centro y posteriormente publicado por Coordinación Académica.	1. El proyecto formativo estimula acciones de investigación aplicada de acuerdo con el Modelo Pedagógico y lineamientos de la estrategia SENNOVA.	
0	2021-07-19 11:08:01.165	En desacuerdo	En desacuerdo	Neutral	De acuerdo	En desacuerdo	,
1	2021-07-19 19:49:56.365	En desacuerdo	En desacuerdo	En desacuerdo	Neutral	En desacuerdo	
2	2021-07-26 17:37:55.557	En desacuerdo	En desacuerdo	Totalmente de acuerdo	Neutral	De acuerdo	
3	2021-07-26 23:04:29.893	En desacuerdo	En desacuerdo	En desacuerdo	Neutral	Totalmente en desacuerdo	
4	2022-05-23 10:25:25.554	De acuerdo	De acuerdo	Muy de acuerdo	De acuerdo	Muy de acuerdo	

▾ EDA(exploratory data analysis)

renombramos columnas

```
columns_descriptions = list(cuest.columns)
columns_descriptions
```

```
['Marca temporal',
 '1. Existe un documento formal por parte de la Coordinación Académica para la designación del equipo ejecutor de la ficha asociada al proyecto formativo.',
 '2. El proyecto formativo se formula en conjunto por los miembros del equipo ejecutor asociado a la ficha del programa de formación.',
 '3. La formulación del proyecto formativo se realiza teniendo en cuenta el contexto y necesidades del sector productivo.',
 '4. El proyecto formativo es evaluado previa revisión del Equipo Pedagógico de Centro y posteriormente publicado por Coordinación Académica.',
 '1. El proyecto formativo estimula acciones de investigación aplicada de acuerdo con el Modelo Pedagógico y lineamientos de la estrategia SENNOVA. ',
```

'2. Todas las actividades de aprendizaje se diseñan teniendo en cuenta las actividades del proyecto formativo.',

'3. Existen actividades de aprendizaje con productos específicos que permiten resolver el problema de investigación propuesto en el proyecto formativo.',

'4. El dinamizador SENNOVA participa en la formulación de los proyectos formativos del Centro de formación.',

'5. El Aprendizaje Basado en Problemas es una estrategia /didáctica utilizada para la promoción y desarrollo de acciones investigativas en el Centro de Formación.',

'6. Los productos del proyecto formativo corresponden con las evidencias solicitadas a los aprendices a lo largo del proceso formativo.',

'7. Se socializan los resultados de los productos de investigación del proyecto formativo como estrategia para el fortalecimiento de las habilidades investigativas de los aprendices del centro.',

'8. Reconozco que el Aprendizaje Basado en Problemas (ABP), es una de las didácticas que se pueden emplear para el desarrollo del proyecto formativo.',

'9. Los proyectos formativos que hacen uso del ABP guardan una mayor relación (o relación fuerte) con los procesos de investigación.',

'1. Para hacer parte de la estrategia SENNOVA es necesario estar vinculado a un Semillero de Investigación o Grupo de Investigación de mi Centro de Formación.',

'2. Conozco el Grupo de Investigación y los Semilleros de Investigación que hacen parte de mi Centro de Formación.',

'3. En los Semilleros de Investigación los actores más importantes son los instructores-investigadores y los aprendices.',

'4. La estrategia de vinculación de los aprendices a los Grupos(s) de Investigación y Semillero(s) de Investigación es una responsabilidad exclusiva de los instructores-investigadores.',

'5. Como instructor he tenido la oportunidad de realizar trabajo colaborativo en SENNOVA en donde participan aprendices e instructores de diferentes programas.',

'6. Puedo presentar cualquier tipo de proyecto de investigación a SENNOVA.',

'7. Los proyectos que se presentan a la convocatoria SENNOVA tienen en cuenta el Plan Tecnológico y la oferta formativa del centro.',

'8. Los productos de los proyectos de investigación financiados en las convocatorias SENNOVA deben registrarse periódicamente en el CvLAC del investigador y el GrupLAC.',

'9. Los Grupos de Investigación y Semilleros de Investigación involucran tareas adicionales a la ejecución de los proyectos en el marco de las convocatorias SENNOVA.',

'1. Los programas de formación, especialmente los que son objeto de registro calificado, deben guardar una relación directa con los Semilleros de Investigación y las actividades SENNOVA.',

'2. El tiempo de dedicación a las actividades de investigación las define el Subdirector(a)/Coordinador(a) Académico(a) de forma autónoma en cada Centro de formación.',

'3. Para diseñar e incluir las competencias de investigación en los proyectos formativos se vinculan a los instructores-investigadores en el equipo ejecutor.',

'4. Para mantener el Registro Calificado es necesario que los instructores se registren en las diferentes bases de datos de MINCIENCIAS.',

'5. Las horas de investigación asignadas a los instructores se deben reportar en el

```
columns_shorten = ['Marca_temporal', 'P1', 'P2', 'P3', 'P4', 'P5', 'P6', 'P7', 'P8', 'P9', 'P10', 'P11', 'P24', 'P25', 'P26', 'P27', 'P28', 'P29', 'P30']
```

```
col_renames = {}
col_description = {}
for desc, short in zip(columns_descriptions, columns_shorten):
```

```
print('\033[1m'+short,':\033[0;0m')
print('\tDescription: ', desc)
col_renames[desc] = short
col_description[short] = desc
```

Marca_temporal :

```
Description: Marca temporal

P1 :
Description: 1. Existe un documento formal por parte de la Coordinación Acad

P2 :
Description: 2. El proyecto formativo se formula en conjunto por los miembro

P3 :
Description: 3. La formulación del proyecto formativo se realiza teniendo en

P4 :
Description: 4. El proyecto formativo es evaluado previa revisión del Equipo

P5 :
Description: 1. El proyecto formativo estimula acciones de investigación apl

P6 :
Description: 2. Todas las actividades de aprendizaje se diseñan teniendo en

P7 :
Description: 3. Existen actividades de aprendizaje con productos específicos

P8 :
Description: 4. El dinamizador SENNOVA participa en la formulación de los pro

P9 :
Description: 5. El Aprendizaje Basado en Problemas es una estrategia /didáct

P10 :
Description: 6. Los productos del proyecto formativo corresponden con las ev

P11 :
Description: 7. Se socializan los resultados de los productos de investigaci

P12 :
Description: 8. Reconozco que el Aprendizaje Basado en Problemas (ABP), es u

P13 :
Description: 9. Los proyectos formativos que hacen uso del ABP guardan una m

P14 :
Description: 1. Para hacer parte de la estrategia SENNOVA es necesario esta

P15 :
Description: 2. Conozco el Grupo de Investigación y los Semilleros de Invest

P16 :
Description: 3. En los Semilleros de Investigación los actores más important

P17 :
Description: 4. La estrategia de vinculación de los aprendices a los Grupos(

P18 :
Description: 5. Como instructor he tenido la oportunidad de realizar trabajo

P19 :
Description: 6. Puedo presentar cualquier tipo de proyecto de investigación

P20 :
Description: 7. Los proyectos que se presentan a la convocatoria SENNOVA tie

P21 :
Description: 8. Los productos de los proyectos de investigación financiados

P22 :
Description: 9. Los Grupos de Investigación y Semilleros de Investigación in

P23 :
Description: 1. Los programas de formación, especialmente los que son objeto

P24 :
Description: 2. El tiempo de dedicación a las actividades de investigación l

P25 :
```

Description: 3. Para diseñar e incluir las competencias de investigación en

P26 :

Description: 4. Para mantener el Registro Calificado es necesario que los in

P27 :

Description: 5. Las horas de investigación asignadas a los instructores se d

P28 :

```
cuest.rename( columns=col_renames, inplace=True)
```

```
cuest
```

	Marca_temporal		P1	P2	P3	P4	P5
0	2021-07-19 11:08:01.165	En desacuerdo	En desacuerdo	Neutral	De acuerdo	En desacuerdo	
1	2021-07-19 19:49:56.365	En desacuerdo	En desacuerdo	En desacuerdo	Neutral	En desacuerdo	
2	2021-07-26 17:37:55.557	En desacuerdo	En desacuerdo	Totalmente de acuerdo	Neutral	De acuerdo	
3	2021-07-26 23:04:29.893	En desacuerdo	En desacuerdo	En desacuerdo	Neutral	Totalmente en desacuerdo	
4	2022-05-23 10:25:25.554	De acuerdo	De acuerdo	Muy de acuerdo	De acuerdo	Muy de acuerdo	
...	...	...	...	...	...	...	
326	2022-06-26 09:53:50.169	Moderadamente de acuerdo	De acuerdo	Opción 3	Poco de acuerdo	De acuerdo	
327	2022-06-28 11:53:29.262	Poco de acuerdo	Muy de acuerdo	Extremadamente de acuerdo	Muy de acuerdo	Muy de acuerdo	
328	2022-06-28 12:00:50.734	Extremadamente en desacuerdo	De acuerdo	Muy de acuerdo	De acuerdo	De acuerdo	
329	2022-06-28 22:32:04.694	Muy de acuerdo	Muy de acuerdo	Muy de acuerdo	De acuerdo	De acuerdo	
330	2022-07-20 14:49:11.347	Extremadamente en desacuerdo	De acuerdo	Extremadamente de acuerdo	De acuerdo	Moderadamente de acuerdo	

331 rows × 31 columns

```
cuest.info()
```

```

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 331 entries, 0 to 330
Data columns (total 31 columns):
#   Column                Non-Null Count  Dtype
---  -
0   Marca_temporal        331 non-null    datetime64[ns]
1   P1                     331 non-null    object
2   P2                     331 non-null    object
3   P3                     331 non-null    object
4   P4                     331 non-null    object
5   P5                     331 non-null    object
6   P6                     331 non-null    object
7   P7                     331 non-null    object
8   P8                     331 non-null    object
9   P9                     331 non-null    object
10  P10                    331 non-null    object
11  P11                    331 non-null    object
12  P12                    331 non-null    object
13  P13                    331 non-null    object
14  P14                    331 non-null    object
15  P15                    331 non-null    object
16  P16                    331 non-null    object
17  P17                    331 non-null    object
18  P18                    331 non-null    object
19  P19                    331 non-null    object
20  P20                    331 non-null    object
21  P21                    331 non-null    object
22  P22                    331 non-null    object
23  P23                    331 non-null    object
24  P24                    331 non-null    object
25  P25                    331 non-null    object
26  P26                    331 non-null    object
27  P27                    331 non-null    object
28  P28                    331 non-null    object
29  P29                    331 non-null    object
30  P30                    331 non-null    object
dtypes: datetime64[ns](1), object(30)
memory usage: 80.3+ KB

```

```

#Verificación de contenido en variables
cuest['P11'].value_counts()

```

```

De acuerdo                101
Muy de acuerdo            80
Moderadamente de acuerdo  48
Poco de acuerdo           45
Extremadamente de acuerdo 42
Extremadamente en desacuerdo 12
Totalmente en desacuerdo  1
Neutral                   1
En desacuerdo             1
Name: P11, dtype: int64

```

```
#Reemplazando Totalmente en desacuerdo, Neutral y En desacuerdo por NaN
```

```
cuest.replace({'Totalmente en desacuerdo':np.nan,'Neutral': np.nan,'En desacuerdo': np.nan, 'En desacuerdo': np.nan})
```

```
#Eliminando NaN
```

```
cuest = cuest.dropna()
```

```
cuest
```

	Marca_temporal	P1	P2	P3	P4	
4	2022-05-23 10:25:25.554	De acuerdo	De acuerdo	Muy de acuerdo	De acuerdo	Mu
6	2022-05-23 10:59:31.935	De acuerdo	De acuerdo	Moderadamente de acuerdo	De acuerdo	Poc
7	2022-05-23 11:12:12.866	De acuerdo	De acuerdo	Muy de acuerdo	De acuerdo	
9	2022-05-23 11:38:58.366	Extremadamente de acuerdo	Muy de acuerdo	Extremadamente de acuerdo	Extremadamente de acuerdo	Extr
10	2022-05-23 11:41:26.864	De acuerdo	Moderadamente de acuerdo	Muy de acuerdo	De acuerdo	
...	...	...	...	...	...	
325	2022-06-24 23:35:50.200	Extremadamente de acuerdo	Poco de acuerdo	Extremadamente de acuerdo	Extremadamente de acuerdo	
327	2022-06-28 11:53:29.262	Poco de acuerdo	Muy de acuerdo	Extremadamente de acuerdo	Muy de acuerdo	Mu
328	2022-06-28 12:00:50.734	Extremadamente en desacuerdo	De acuerdo	Muy de acuerdo	De acuerdo	
329	2022-06-28 22:32:04.694	Muy de acuerdo	Muy de acuerdo	Muy de acuerdo	De acuerdo	
330	2022-07-20 14:49:11.347	Extremadamente en desacuerdo	De acuerdo	Extremadamente de acuerdo	De acuerdo	Mo

312 rows × 31 columns

```
#Muestra de variables con cuestionario corregido
```

```
cuest['P14'].value_counts()
```

```
De acuerdo          100
Muy de acuerdo      80
Extremadamente de acuerdo  58
```

```

Moderadamente de acuerdo    46
Poco de acuerdo             24
Extremadamente en desacuerdo  4
Name: P14, dtype: int64

```

cuest

	Marca_temporal	P1	P2	P3	P4	
4	2022-05-23 10:25:25.554	De acuerdo	De acuerdo	Muy de acuerdo	De acuerdo	Mu
6	2022-05-23 10:59:31.935	De acuerdo	De acuerdo	Moderadamente de acuerdo	De acuerdo	Poc
7	2022-05-23 11:12:12.866	De acuerdo	De acuerdo	Muy de acuerdo	De acuerdo	
9	2022-05-23 11:38:58.366	Extremadamente de acuerdo	Muy de acuerdo	Extremadamente de acuerdo	Extremadamente de acuerdo	Extr
10	2022-05-23 11:41:26.864	De acuerdo	Moderadamente de acuerdo	Muy de acuerdo	De acuerdo	
...	...	...	...	...	...	
325	2022-06-24 23:35:50.200	Extremadamente de acuerdo	Poco de acuerdo	Extremadamente de acuerdo	Extremadamente de acuerdo	
327	2022-06-28 11:53:29.262	Poco de acuerdo	Muy de acuerdo	Extremadamente de acuerdo	Muy de acuerdo	Mu
328	2022-06-28 12:00:50.734	Extremadamente en desacuerdo	De acuerdo	Muy de acuerdo	De acuerdo	
329	2022-06-28 22:32:04.694	Muy de acuerdo	Muy de acuerdo	Muy de acuerdo	De acuerdo	
330	2022-07-20 14:49:11.347	Extremadamente en desacuerdo	De acuerdo	Extremadamente de acuerdo	De acuerdo	Mor

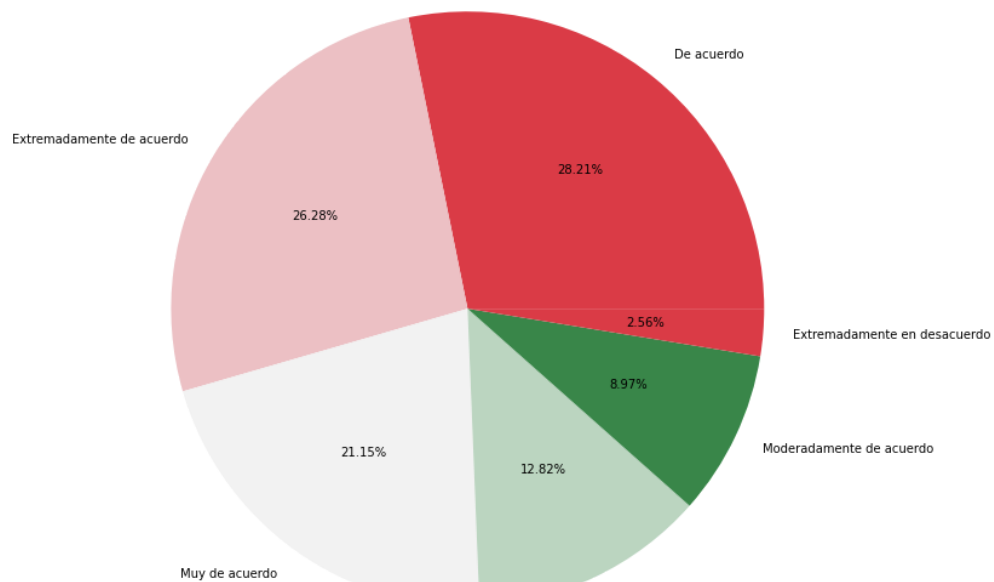
312 rows × 31 columns

```

cuest['P1'].value_counts().plot(kind='pie', autopct='%.2f%%', figsize=(20,10))
plt.axis('equal');

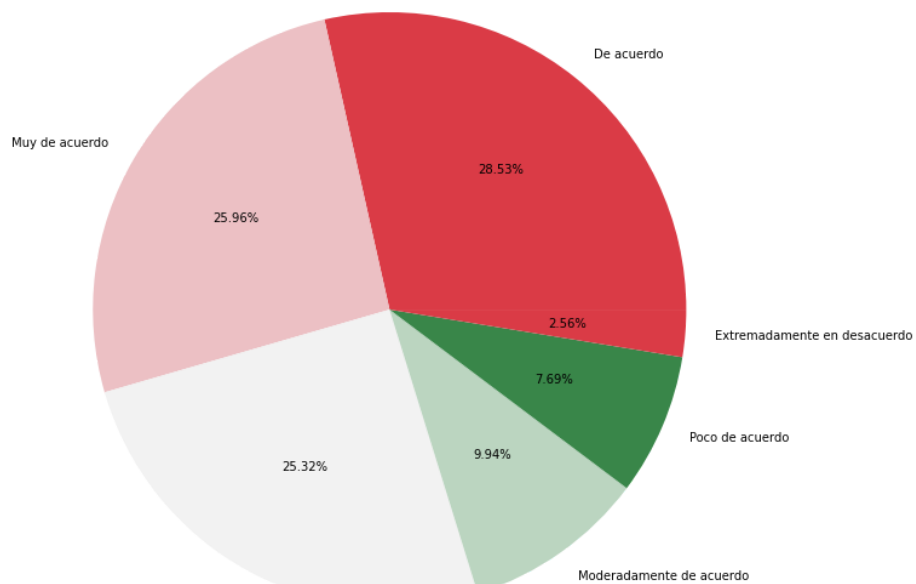
```

P1

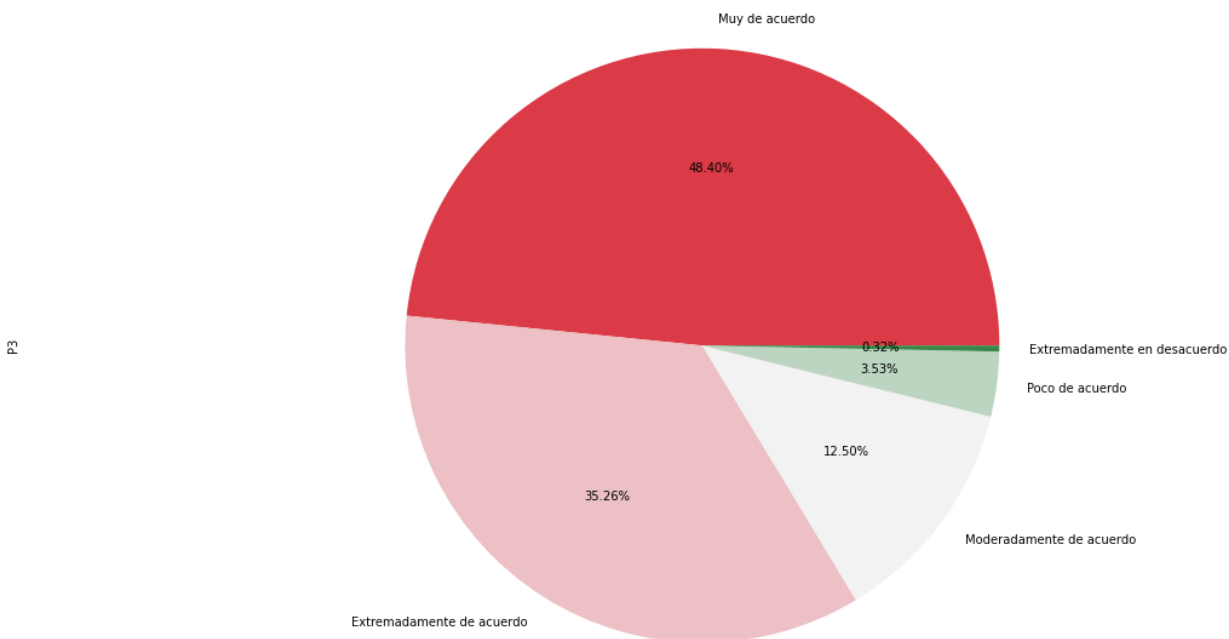


```
cuest['P2'].value_counts().plot(kind='pie', autopct='%0.2f%%', figsize=(20,10))
plt.axis('equal');
```

P2




```
cuest['P3'].value_counts().plot(kind='pie', autopct='%0.2f%%',figsize=(20,10))
plt.axis('equal');
```



Gráficas de barras resumidas de preguntas

```
answering_columns = list (cuest.columns[1:29])
```

```
rows = round( len(answering_columns)//2)
```

```
fig, axarr = plt.subplots(rows, 2, figsize=(40, 4*rows))
```

```
c = 0
```

```
r = 0
```

```
answers = {'Extremadamente de acuerdo': 6, 'Muy de acuerdo':5, 'De acuerdo':4, 'Moderadamente d
```

```
#sns.set_palette(sns.cubehelix_palette(5, start=.5, rot=-.75))
sns.set_palette(sns.diverging_palette(10, 133, sep=80, n=5))

size = len(cuest)
for count,col in enumerate(answering_columns):

    #sns.countplot(base[col].sort_values(), ax=axarr[r][c])

    ax = sns.countplot( cuest.iloc[cuest[col].map(answers).argsort()][col], ax=axarr[r][c] )
    axarr[r][c].set_title( col_description[col], fontsize=11)

    #ax = (base[col].value_counts()/len(base)*100).sort_index().plot(kind="bar", rot=0)
    #ax.set_yticks(np.arange(0, 900, 10))

    #ax2 = ax.twinx()
    #ax2.set_yticks(np.arange(0, 110, 10)*len(base)/100)

    for p in ax.patches:
        ax.annotate('{:.2f}%'.format(p.get_height()/len(cuest) * 100), (p.get_x()+0.15, p.get

    if c >= 1:
        c = 0
        r = r + 1
    else:
        c = c + 1
plt.subplots_adjust(hspace=.3)
sns.despine()
```

```
/usr/local/lib/python3.7/dist-packages/seaborn/_decorators.py:43: FutureWarning:
```

```
Pass the following variable as a keyword arg: x. From version 0.12, the only valid posit
```

```
/usr/local/lib/python3.7/dist-packages/seaborn/_decorators.py:43: FutureWarning:
```

```
Pass the following variable as a keyword arg: x. From version 0.12, the only valid posit
```

```
/usr/local/lib/python3.7/dist-packages/seaborn/_decorators.py:43: FutureWarning:
```

```
Pass the following variable as a keyword arg: x. From version 0.12, the only valid posit
```

```
/usr/local/lib/python3.7/dist-packages/seaborn/_decorators.py:43: FutureWarning:
```

```
Pass the following variable as a keyword arg: x. From version 0.12, the only valid posit
```

```
/usr/local/lib/python3.7/dist-packages/seaborn/_decorators.py:43: FutureWarning:
```

```
Pass the following variable as a keyword arg: x. From version 0.12, the only valid posit
```

```
/usr/local/lib/python3.7/dist-packages/seaborn/_decorators.py:43: FutureWarning:
```

```
Pass the following variable as a keyword arg: x. From version 0.12, the only valid posit
```

```
/usr/local/lib/python3.7/dist-packages/seaborn/_decorators.py:43: FutureWarning:
```

```
Pass the following variable as a keyword arg: x. From version 0.12, the only valid posit
```

```
/usr/local/lib/python3.7/dist-packages/seaborn/_decorators.py:43: FutureWarning:
```

```
Pass the following variable as a keyword arg: x. From version 0.12, the only valid posit
```

```
/usr/local/lib/python3.7/dist-packages/seaborn/_decorators.py:43: FutureWarning:
```

```
Pass the following variable as a keyword arg: x. From version 0.12, the only valid posit
```

Reemplazando variables categóricas a cantidades Ejemplo: Extremadamente de acuerdo' : 6, 'Muy de acuerdo' : 5.

```
cuest.replace({'Extremadamente de acuerdo': 6, 'Muy de acuerdo':5, 'De acuerdo':4, 'Moderadamen
```

```
Pass the following variable as a keyword arg: x. From version 0.12, the only valid posit
```

```
cuest
```

21/8/22, 22:51

Analítica_de_datos_CUESTIONARIO_1.ipynb - Colaboratory

	Marca_temporal	P1	P2	P3	P4	P5	P6	P7	P8	P9	...	P21	P22	P23	P24	P25	P
4	2022-05-23 10:25:25.554	4	4	5	4	5	5	4	4	4	...	5	4	4	4	4	
6	2022-05-23 10:59:31.935	4	4	3	4	2	5	4	3	4	...	4	5	6	5	2	
7	2022-05-23 11:12:12.866	4	4	5	4	4	5	5	4	4	...	5	5	4	2	5	
9	2022-05-23 11:38:58.366	6	5	6	6	6	6	6	6	4	...	6	5	5	3	5	
	2022-05-23																

▼ Estadística descriptiva de los datos.

2022-06-24

```
cuest.describe()
```

	P1	P2	P3	P4	P5	P6	P7
count	312.000000	312.000000	312.000000	312.000000	312.000000	312.000000	312.000000
mean	4.314103	4.435897	4.983974	4.717949	4.035256	4.849359	4.477564
std	1.415597	1.313690	1.106717	1.177467	1.354535	1.045462	1.189238
min	1.000000	1.000000	1.000000	1.000000	1.000000	2.000000	1.000000
25%	4.000000	4.000000	5.000000	4.000000	3.000000	4.000000	4.000000
50%	4.000000	5.000000	5.000000	5.000000	4.000000	5.000000	5.000000
75%	6.000000	6.000000	6.000000	6.000000	5.000000	6.000000	5.000000
max	6.000000	6.000000	6.000000	6.000000	6.000000	6.000000	6.000000

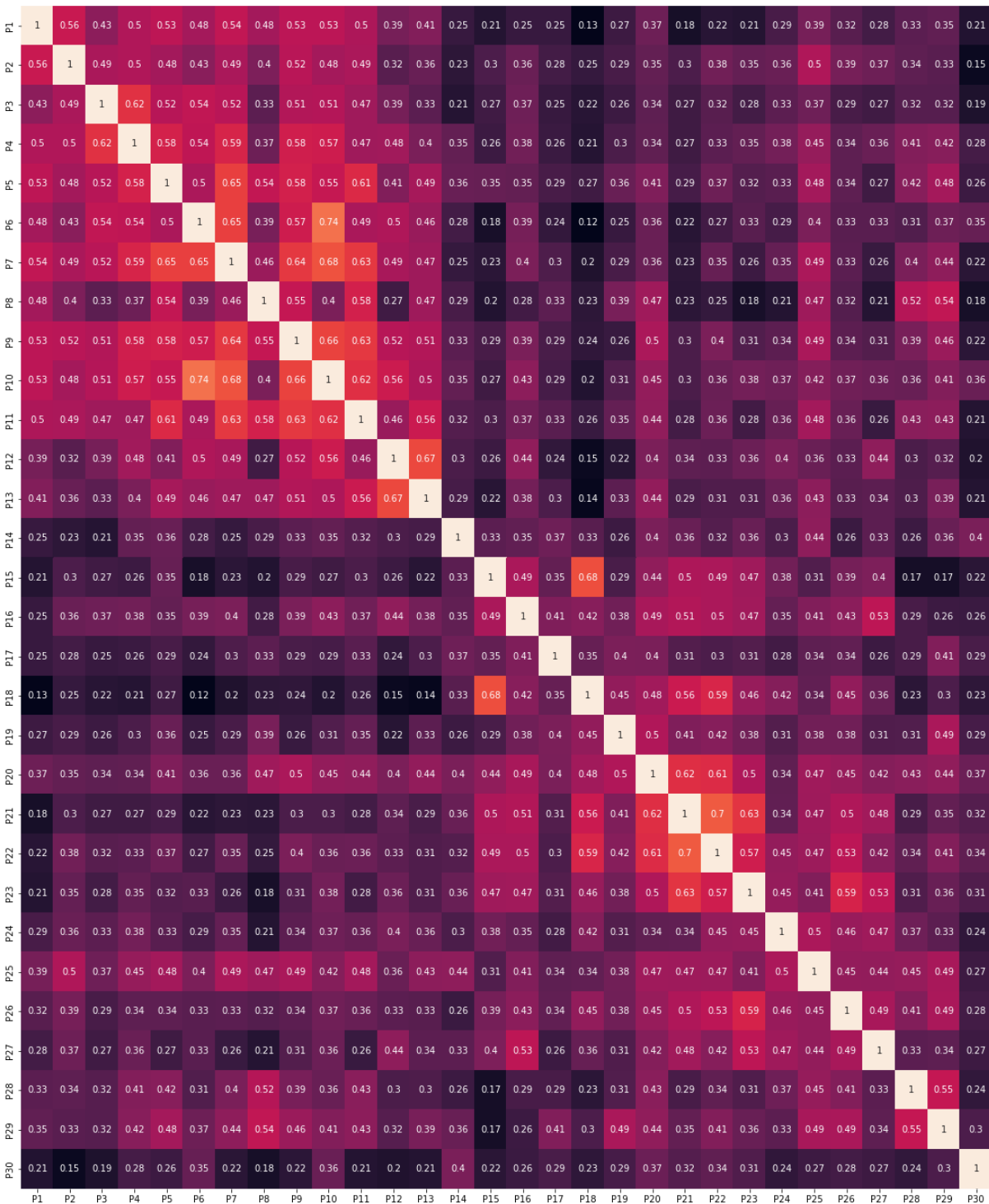
8 rows × 30 columns

/usr/local/lib/python3.7/dist-packages/seaborn/_decorators.py:43: FutureWarning:

Correlación de Pearson.

```
corr_df = cuest.drop(columns = ['Marca_temporal'],axis = 1).corr(method='pearson')

plt.figure(figsize=(24, 24))
sns.heatmap(corr_df, annot=True)
plt.show()
```



Coeficiente de variación

```
#define function to calculate cv
cv = lambda x: np.std(x, ddof=1) / np.mean(x) * 100

df = cuest.drop(columns = ['Marca_temporal'],axis = 1)

#calculate CV for each column in data frame
df1 = df.apply(cv)

columns_shorten.remove('Marca_temporal')

#sns.scatterplot(x=list(results.keys() ),y= list(results.values()))
plt.figure(figsize=(15,5))
ax = sns.barplot(x=columns_shorten,y= df1 )
plt.title(title)
#ax = sns.lineplot(x = list(range(0,len(df))),y= df[target].mean(),label='mean ({:0.2f})'.for
for p in ax.patches:
    _x = p.get_x() + p.get_width() / 2
    _y = p.get_y() + p.get_height()+1
    value = str( int(p.get_height() ) )
    ax.text(_x, _y, value, ha="center")

    #ax.text(_x, _y, value, ha="left")

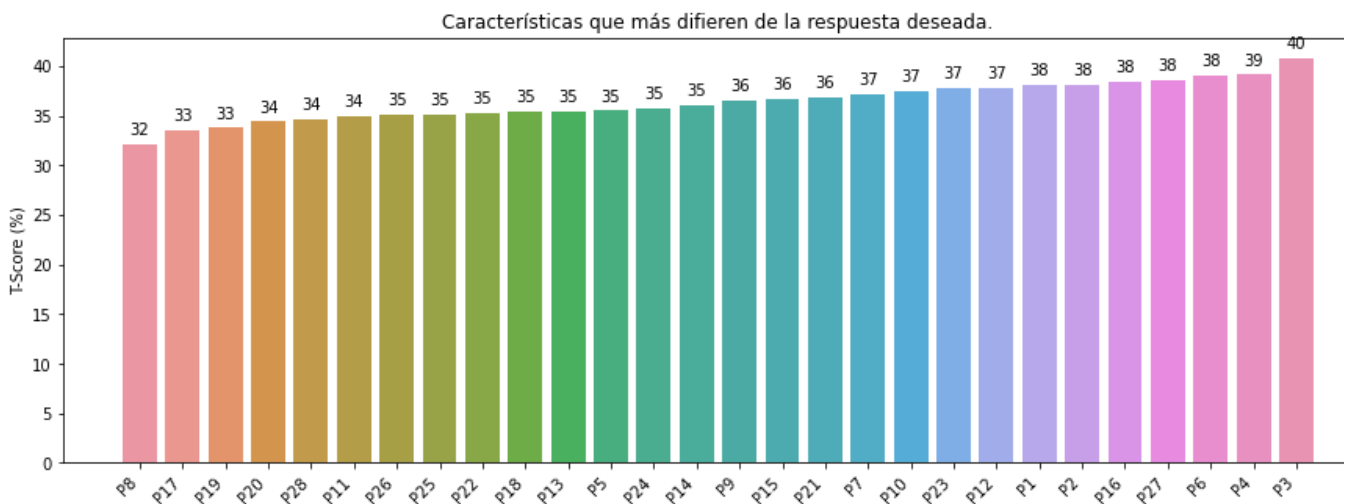
ax.set_xticklabels(ax.get_xticklabels(), rotation=45, horizontalalignment='right')
plt.ylabel('Coeficiente de variación')
plt.title('Coeficiente de variación - Cuestionario Instructores')
plt.plot()
```



```

P1 4.314102564102564
P2 4.435897435897436
P3 4.983974358974359
P4 4.717948717948718
P5 4.035256410256411
P6 4.8493589743589745
P7 4.477564102564102
P8 3.2371794871794872
P9 4.403846153846154
P10 4.7243589743589745
P11 3.967948717948718
P12 4.875
P13 4.458333333333333
P14 4.288461538461538
P15 4.019230769230769
P16 4.685897435897436
P17 3.7275641025641026
P18 3.4551282051282053
P19 3.730769230769231
P20 4.221153846153846
P21 4.487179487179487
P22 4.342948717948718
P23 4.6506410256410255
P24 4.246794871794871
P25 4.141025641025641
P26 4.07051282051282
P27 4.7756410256410255
P28 3.858974358974359
[32.05, 33.58, 33.76, 34.39, 34.66, 34.99, 35.08, 35.16, 35.21, 35.43, 35.46, 35.5, 35.6
[ ]

```



▼ Análisis Factorial

```
from factor_analyzer import FactorAnalyzer
```

Prueba de esfericidad de Bartlett

comprueba si las variables observadas se interrelacionan o no usando la matriz de correlación observada contra la matriz de identidad. Si la prueba resultó estadísticamente insignificante, no debe emplear un análisis factorial.

```
from factor_analyzer.factor_analyzer import calculate_bartlett_sphericity
chi_square_value,p_value=calculate_bartlett_sphericity(df)
chi_square_value, p_value

(5466.290354148843, 0.0)
```

En esta prueba de Bartlett, el valor p es 0. la prueba fue estadísticamente significativa, indicando que la matriz de correlación observada no es una matriz de identidad.

prueba Kaiser-Meyer-Olkin (KMO)

Mide la idoneidad de los datos para el análisis factorial. Determina la adecuación para cada variable observada y para el modelo completo. KMO estima la proporción de varianza entre todas las variables observadas. Identificación de proporción más baja más adecuada para el análisis factorial. Los valores de KMO oscilan entre 0 y 1. Un valor de KMO inferior a 0,6 se considera inadecuado.

```
from factor_analyzer.factor_analyzer import calculate_kmo

kmo_all,kmo_model=calculate_kmo(df)

kmo_model

0.9410779212845063
```

El KMO general para nuestros datos es 0,94 , lo cual es excelente. Este valor indica que puede continuar con su análisis factorial planificado.

Análisis factorial

```
df = df.dropna()

fa = FactorAnalyzer(n_factors= 5,rotation = 'varimax')

fa.fit(df)
```

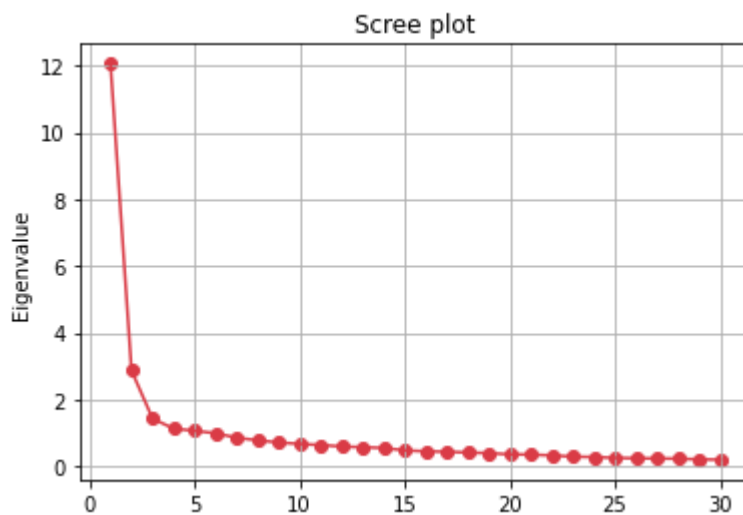
```
FactorAnalyzer(n_factors=5, rotation='varimax', rotation_kwangs={})
```

```
#Get the loadings
loadings = fa.loadings_
```

```
# Eigen vectores
ev, v = fa.get_eigenvalues()
```

```
xvals = range(1,df.shape[1]+1)
```

```
plt.scatter(xvals,ev)
plt.plot(xvals,ev)
plt.title('Scree plot')
plt.ylabel('Eigenvalue')
plt.grid()
plt.show()
```



```
# análisis factorial
df_pca = pd.DataFrame.from_records(loadings)
df_pca['Pregunta'] = columns_shorten
first_column = df_pca.pop('Pregunta')
df_pca.insert(0, 'Pregunta', first_column)
df_pca.columns = ['Pregunta', 'Factor 1', 'Factor 2', 'Factor 3', 'Factor 4', 'Factor 5']
display(df_pca)
```

	Pregunta	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
0	P1	0.626583	0.099170	0.253447	0.105868	0.006173
1	P2	0.600908	0.280697	0.181221	-0.019444	0.076727
2	P3	0.671203	0.225136	0.075387	0.017010	0.061477
3	P4	0.699081	0.223817	0.158453	0.067945	0.186800
4	P5	0.652987	0.223972	0.337308	0.087694	-0.095997
5	P6	0.673909	0.140008	0.118781	0.287237	0.138656
6	P7	0.755994	0.139709	0.226604	0.166860	-0.001848
7	P8	0.414063	0.096109	0.664833	0.117840	-0.199211
8	P9	0.674862	0.208371	0.273877	0.242186	-0.066530
9	P10	0.700090	0.227165	0.143362	0.327337	0.084342
10	P11	0.613682	0.197573	0.362275	0.233248	-0.167915
11	P12	0.444029	0.255382	0.044720	0.633039	0.127621
12	P13	0.394426	0.185062	0.296373	0.591491	-0.038423
13	P14	0.210507	0.378073	0.253238	0.162143	0.029217
14	P15	0.207384	0.731413	-0.029283	-0.007815	-0.235042
15	P16	0.305363	0.582400	0.068901	0.252600	0.018860
16	P17	0.184176	0.377987	0.333789	0.101011	-0.054028
17	P18	0.078156	0.783762	0.149579	-0.140144	-0.197303
18	P19	0.155750	0.429605	0.432513	0.067576	-0.023213
19	P20	0.229023	0.573136	0.391467	0.254098	-0.073691
20	P21	0.072777	0.751787	0.180708	0.151642	0.075266
21	P22	0.202149	0.714226	0.204784	0.070757	0.061197
22	P23	0.170985	0.680781	0.118048	0.149245	0.281867
23	P24	0.304828	0.465226	0.151060	0.097140	0.224371
24	P25	0.401793	0.399819	0.392979	0.102893	0.146620
25	P26	0.228233	0.547238	0.302360	0.063257	0.274683

```
# Get variance of each factors
```

```
df_var =pd.DataFrame.from_records(fa.get_factor_variance())
```

```
df_var['Parámetros'] =['SS Loadings','Proportion Var','Cumulative Var']
```

```
first_column = df_var.pop('Parámetros')
```

```
df_var.insert(0, 'Parámetros', first_column)
```

```
df_var.columns = ['Parámetros', 'Factor 1', 'Factor 2', 'Factor 3', 'Factor 4', 'Factor 5']
display(df_var)
```

	Parámetros	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
0	SS Loadings	5.927927	5.533386	2.748929	1.455524	0.698704
1	Proportion Var	0.197598	0.184446	0.091631	0.048517	0.023290
2	Cumulative Var	0.197598	0.382044	0.473675	0.522192	0.545482

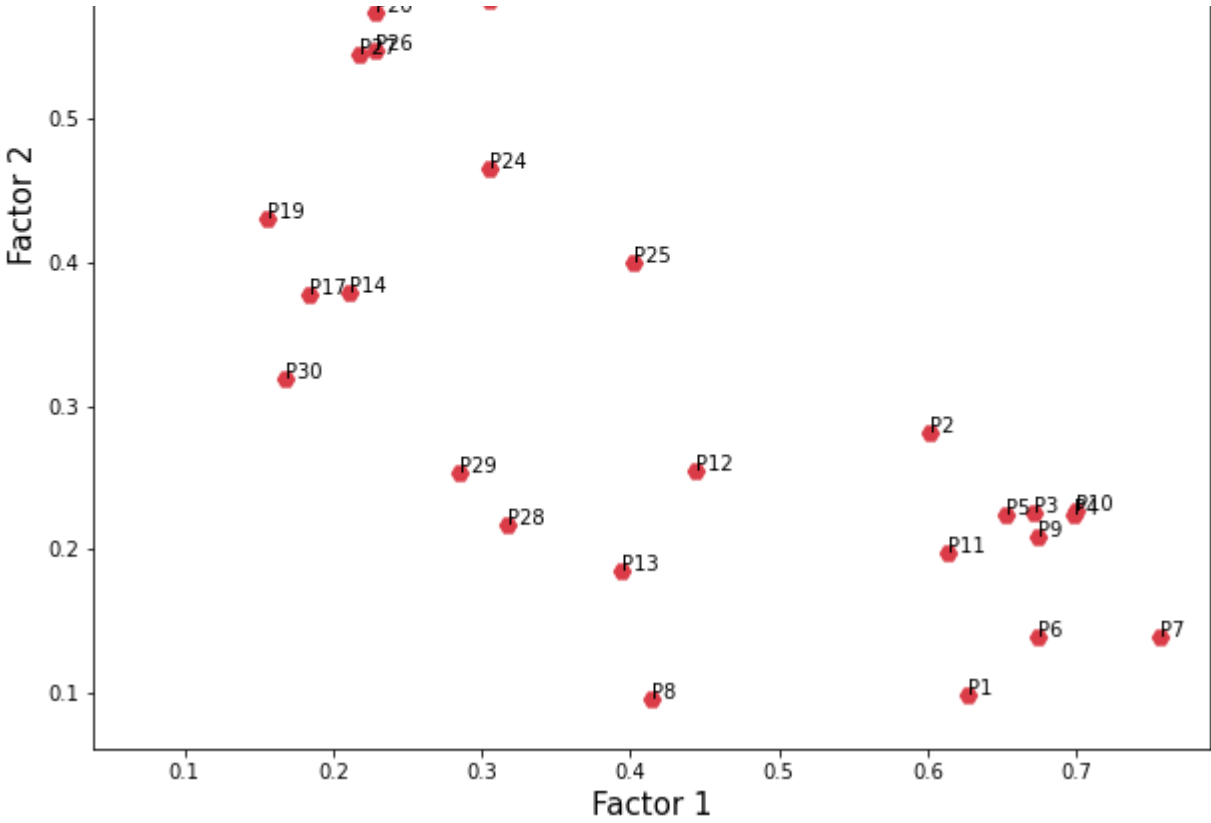
```
#gráfico de análisis factorial
fig=plt.figure(figsize=(10,10))
ax=fig.add_subplot(1,1,1)
annotations = df_pca['Pregunta']
ax.set_xlabel("Factor 1",fontsize=15)
ax.set_ylabel("Factor 2", fontsize=15)
ax.set_title("Análisis Factorial", fontsize=20)
color_theme=np.array(["CYAN",'GREENYELLOW','mediumpurple'])
ax.scatter(x=df_pca['Factor 1'], y=df_pca['Factor 2'],marker = "H" ,s=70)
for i, label in enumerate(annotations):
    plt.annotate(label, (df_pca['Factor 1'][i], df_pca['Factor 2'][i]))
plt.show()
```

Análisis Factorial



Referencias:

- <https://bit.ly/3PCfZBb>
- <https://bit.ly/3QY5iK7>



▼ Analítica de datos CUESTIONARIO 2 - Aprendizices

```
pip install factor_analyzer
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public
Requirement already satisfied: factor_analyzer in /usr/local/lib/python3.7/dist-packages (1.0.0)
Requirement already satisfied: scikit-learn in /usr/local/lib/python3.7/dist-packages (1.0.2)
Requirement already satisfied: scipy in /usr/local/lib/python3.7/dist-packages (from factor_analyzer) (1.7.3)
Requirement already satisfied: pandas in /usr/local/lib/python3.7/dist-packages (from factor_analyzer) (1.3.4)
Requirement already satisfied: numpy in /usr/local/lib/python3.7/dist-packages (from factor_analyzer) (1.21.6)
Requirement already satisfied: pytz>=2017.3 in /usr/local/lib/python3.7/dist-packages (from factor_analyzer) (2022.1)
Requirement already satisfied: python-dateutil>=2.7.3 in /usr/local/lib/python3.7/dist-packages (from factor_analyzer) (2.8.2)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.7/dist-packages (from factor_analyzer) (1.16.0)
Requirement already satisfied: threadpoolctl>=2.0.0 in /usr/local/lib/python3.7/dist-packages (from factor_analyzer) (3.1.0)
Requirement already satisfied: joblib>=0.11 in /usr/local/lib/python3.7/dist-packages (from factor_analyzer) (1.2.0)
```

```
import pandas as pd
import numpy as np
#Gráfica
import matplotlib.pyplot as plt
import seaborn as sns
import plotly.express as px
from plotly.subplots import make_subplots
import plotly.figure_factory as ff
import plotly.offline as offline
import plotly.graph_objs as go
```

```
from google.colab import drive
drive.mount('/content/drive')
```

```
Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount('/content/drive', force_remount=True)
```

```
from google.colab import files
path = "/content/drive/MyDrive/Proyecto SGPS 9748/Analítica Luz Karime_Regional Cauca/Analítica de datos"
#Dataframes
cuest= pd.read_excel(path)

cuest
```

	Marca temporal	1. Reconozco el perfil de salida establecido en el programa de formación.	2. Dentro del ambiente de formación se ha socializado el proyecto formativo de mi programa.	3. Conozco el proyecto formativo y el alcance que éste tiene dentro de la formación.	4. Reconozco la importancia del proyecto formativo para la ejecución de mi programa.	aci api desai pi +
0	2021-07-18 11:36:35.227	Totalmente de acuerdo	De acuerdo	Neutral	De acuerdo	Tota
1	2021-07-18 11:38:21.228	De acuerdo	De acuerdo	De acuerdo	De acuerdo	[
2	2021-07-18 12:53:16.714	De acuerdo	De acuerdo	Neutral	De acuerdo	[
3	2021-07-18 15:01:10.977	Totalmente de acuerdo	De acuerdo	De acuerdo	De acuerdo	[
4	2021-07-18 22:07:04.492	De acuerdo	De acuerdo	Neutral	De acuerdo	
...	...	...	...	...	...	
1178	2022-07-10 12:29:07.771	Extremadamente de acuerdo	Extremadamente de acuerdo	Extremadamente de acuerdo	Extremadamente de acuerdo	Extren c
1179	2022-07-11 18:03:01.593	Muy de acuerdo	Muy de acuerdo	De acuerdo	Extremadamente de acuerdo	Muy c
1180	2022-07-14 16:50:19.157	De acuerdo	De acuerdo	De acuerdo	De acuerdo	[
1181	2022-07-19 13:09:34.327	Moderadamente de acuerdo	Muy de acuerdo	Muy de acuerdo	Muy de acuerdo	Muy c
1182	2022-07-20 03:59:37.368	Poco de acuerdo	Poco de acuerdo	Poco de acuerdo	De acuerdo	Muy c

1183 rows × 31 columns

▼ EDA(exploratory data analysis)

renombramos columnas

```
columns_descriptions = list(cuest.columns)
```

```
columns_descriptions
```

```
[ 'Marca temporal',
  '1. Reconozco el perfil de salida establecido en el programa de formación.',
  '2. Dentro del ambiente de formación se ha socializado el proyecto formativo de mi programa.',
  '3. Conozco el proyecto formativo y el alcance que éste tiene dentro de la formación.',
  '4. Reconozco la importancia del proyecto formativo para la ejecución de mi programa.',
  '5. Las actividades de aprendizaje se desarrollan a partir del proyecto formativo',
  '6. Las actividades del proyecto están relacionadas con los resultados de aprendizaje de mi programa de formación.',
  '7. Las actividades planeadas en las acciones de formación corresponden a los productos del proyecto formativo del programa que se adelanta.',
  '8. El proyecto formativo da respuesta al problema formulado teniendo en cuenta una necesidad u oportunidad del sector productivo.',
  '1. El aprendizaje que recibo en el SENA es igual al que recibí en mi colegio y/o Institución Educativa - IE',
  '2. La formación que imparte el Servicio Nacional de aprendizaje - SENA me permite el desarrollo de habilidades para desempeñarme en el sector productivo.',
  '3. En los ambientes de formación me enseñan a resolver problemas del mundo laboral.',
  '4. Soy capaz de resolver problemas de mi entorno por medio de la estrategia de Aprendizaje Basado en Proyectos.',
  '5. En mi proceso de formación he aprendido a identificar y formular un problema.',
  '6. Cuando realizo trabajo en equipo se valoran las diferentes opiniones y posiciones argumentadas.',
  '7. La búsqueda de información en libros, revistas, bases de datos, entre otros, es complicada y poco útil.',
  '8. Como aprendiz puedo identificar un problema y puedo plantear unos objetivos para resolverlo.',
  '9. Dentro de las actividades planteadas en las sesiones formativas, el instructor implementa estrategias que estimulan ejercicios de investigación (estudio de caso, método experimental, formulación de problemas, delimitación del problema).',
  '10. En los procesos de formación se promueve la divulgación de los resultados de investigación.',
  '11. Me siento con las capacidades y competencias para presentar los resultados de los proyectos de investigación desarrollados durante mi proceso formativo.',
  '12. En mi Centro de formación se promueven eventos donde se presentan los resultados de las investigaciones realizadas en mi proceso formativo.',
  '1. Como aprendiz, he recibido sensibilización sobre la estrategia SENNOVA',
  '2. En los Semilleros de Investigación los actores más importantes son los aprendices y los instructores-investigadores.',
  '3. Los aprendices pueden participar en cualquier momento de las actividades que realizan los Grupo(s) de Investigación y Semillero(s) de Investigación.',
  '4. El trabajo colaborativo en SENNOVA promueve la participación de aprendices e instructores que hacen parte de los diferentes programas de formación.',
  '5. Para hacer parte de la estrategia SENNOVA debo vincularme a un Grupo de Investigación o a un Semillero de Investigación que pertenezca a mi Centro de formación.',
  '6. Conozco el Grupo de Investigación y los Semilleros de Investigación que hacen parte de mi Centro de formación.',
  '7. Mi programa de formación tiene relación con uno de los Semilleros de
```



```

Investigación de mi Centro.',
'8. A través de los proyectos SENNOVA puedo culminar mi etapa práctica en
modalidades de contrato de aprendizaje, monitorias y pasantías.',
'9. En mi Centro de formación se realizan actividades periódicas de divulgación de
la estrategia SENNOVA, los Grupo de Investigación, Semilleros de Investigación,
proyectos y resultados de investigación '

```

```

columns_shorten = ['Marca_temporal', 'P1', 'P2', 'P3', 'P4', 'P5', 'P6', 'P7', 'P8', 'P9', 'P10', 'P11', 'P12', 'P13', 'P14', 'P15', 'P16', 'P17', 'P18', 'P19', 'P20', 'P21', 'P22', 'P23', 'P24', 'P25', 'P26', 'P27', 'P28', 'P29', 'P30']

```

```

col_renames = {}
col_description = {}
for desc, short in zip(columns_descriptions, columns_shorten):
    print('\033[1m'+short,':\033[0;0m')
    print('\tDescription: ', desc)
    col_renames[desc] = short
    col_description[short] = desc

```

```

Marca_temporal :
    Description:  Marca temporal
P1 :
    Description:  1.  Reconozco el perfil de salida establecido en el programa de
P2 :
    Description:  2.  Dentro del ambiente de formación se ha socializado el proye
P3 :
    Description:  3.  Conozco el proyecto formativo y el alcance que éste tiene d
P4 :
    Description:  4.  Reconozco la importancia del proyecto formativo para la eje
P5 :
    Description:  5.  Las actividades de aprendizaje se desarrollan a partir del p
P6 :
    Description:  6.  Las actividades del proyecto están relacionadas con los resu
P7 :
    Description:  7.  Las actividades planeadas en las acciones de formación corr
P8 :
    Description:  8.  El proyecto formativo da respuesta al problema formulado te
P9 :
    Description:  1.  El aprendizaje que recibo en el SENA es igual al que recibí
P10 :
    Description:  2.  La formación que imparte el Servicio Nacional de aprendizaje
P11 :
    Description:  3.  En los ambientes de formación me enseñan a resolver problema
P12 :
    Description:  4.  Soy capaz de resolver problemas de mi entorno por medio de l
P13 :
    Description:  5.  En mi proceso de formación he aprendido a identificar y form
P14 :
    Description:  6.  Cuando realizo trabajo en equipo se valoran las diferentes o
P15 :
    Description:  7.  La búsqueda de información en libros, revistas, bases de dato
P16 :
    Description:  8.  Como aprendiz puedo identificar un problema y puedo plantear
P17 :
    Description:  9.  Dentro de las actividades planteadas en las sesiones formati

```

P18 : Description: 10. En los procesos de formación se promueve la divulgación de

P19 : Description: 11. Me siento con las capacidades y competencias para presenta

P20 : Description: 12. En mi Centro de formación se promueven eventos donde se pr

P21 : Description: 1. Como aprendiz, he recibido sensibilización sobre la estrateg

P22 : Description: 2. En los Semilleros de Investigación los actores más important

P23 : Description: 3. Los aprendices pueden participar en cualquier momento de las

P24 : Description: 4. El trabajo colaborativo en SENNOVA promueve la participació

P25 : Description: 5. Para hacer parte de la estrategia SENNOVA debo vincularme a

P26 : Description: 6. Conozco el Grupo de Investigación y los Semilleros de Inves

P27 : Description: 7. Mi programa de formación tiene relación con uno de los Sem

P28 :

```
cuest.rename( columns=col_renames, inplace=True)
```

```
cuest
```

	Marca_temporal	P1	P2	P3	P4
0	2021-07-18 11:36:35.227	Totalmente de acuerdo	De acuerdo	Neutral	De acuerdo
1	2021-07-18 11:38:21.228	De acuerdo	De acuerdo	De acuerdo	De acuerdo
2	2021-07-18 11:38:21.228	De acuerdo	De acuerdo	Neutral	De acuerdo

cuest.info()

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1183 entries, 0 to 1182
Data columns (total 31 columns):
#   Column                Non-Null Count  Dtype
---  -
0   Marca_temporal        1183 non-null   datetime64[ns]
1   P1                     1183 non-null   object
2   P2                     1183 non-null   object
3   P3                     1183 non-null   object
4   P4                     1183 non-null   object
5   P5                     1183 non-null   object
6   P6                     1183 non-null   object
7   P7                     1183 non-null   object
8   P8                     1183 non-null   object
9   P9                     1183 non-null   object
10  P10                    1183 non-null   object
11  P11                    1183 non-null   object
12  P12                    1183 non-null   object
13  P13                    1182 non-null   object
14  P14                    1182 non-null   object
15  P15                    1183 non-null   object
16  P16                    1183 non-null   object
17  P17                    1183 non-null   object
18  P18                    1183 non-null   object
19  P19                    1183 non-null   object
20  P20                    1156 non-null   object
21  P21                    1183 non-null   object
22  P22                    1183 non-null   object
23  P23                    1183 non-null   object
24  P24                    1182 non-null   object
25  P25                    1183 non-null   object
26  P26                    1183 non-null   object
27  P27                    1183 non-null   object
28  P28                    1183 non-null   object
29  P29                    1183 non-null   object
30  P30                    1183 non-null   object
dtypes: datetime64[ns](1), object(30)
memory usage: 286.6+ KB
```

```
#Verificación de contenido en variables
cuest['P11'].value_counts()
```

Muy de acuerdo

409

De acuerdo	363
Extremadamente de acuerdo	311
Moderadamente de acuerdo	55
Poco de acuerdo	22
Totalmente de acuerdo	14
Extremadamente en desacuerdo	7
Neutral	1
Totalmente en desacuerdo	1

Name: P11, dtype: int64

```
#Reemplazando Totalmente en desacuerdo, Neutral y En desacuerdo por NaN
```

```
cuest.replace({'Totalmente en desacuerdo':np.nan,'Neutral': np.nan,'En desacuerdo': np.nan, '
```

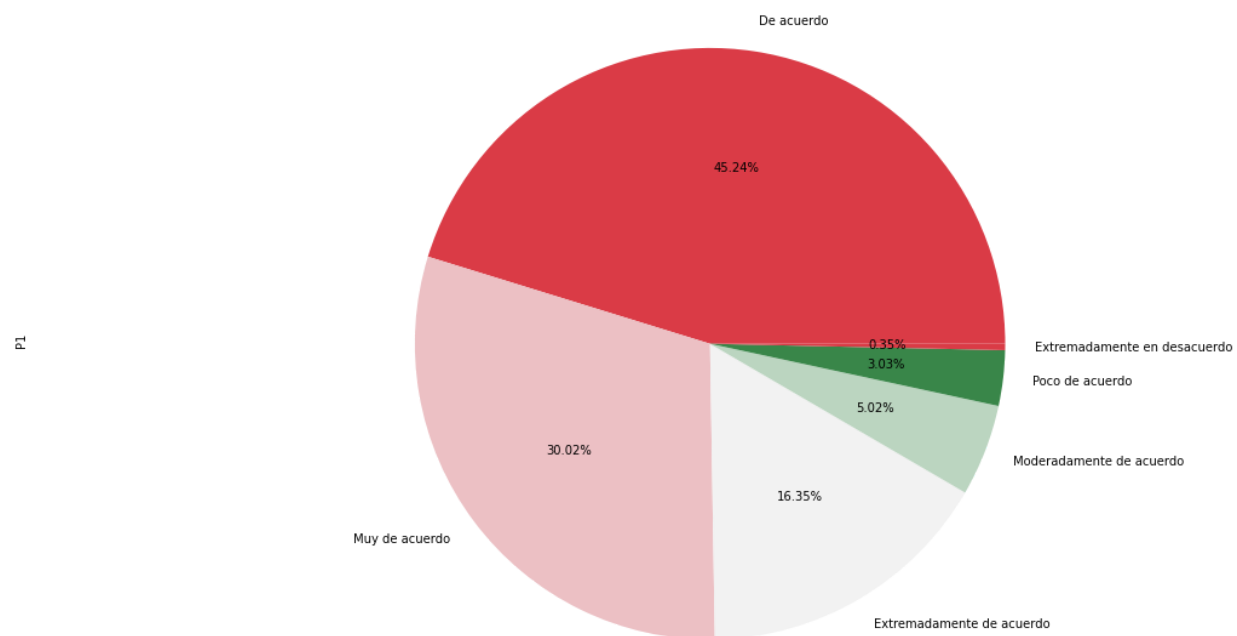
```
#Eliminando NaN
```

```
cuest = cuest.dropna()
```

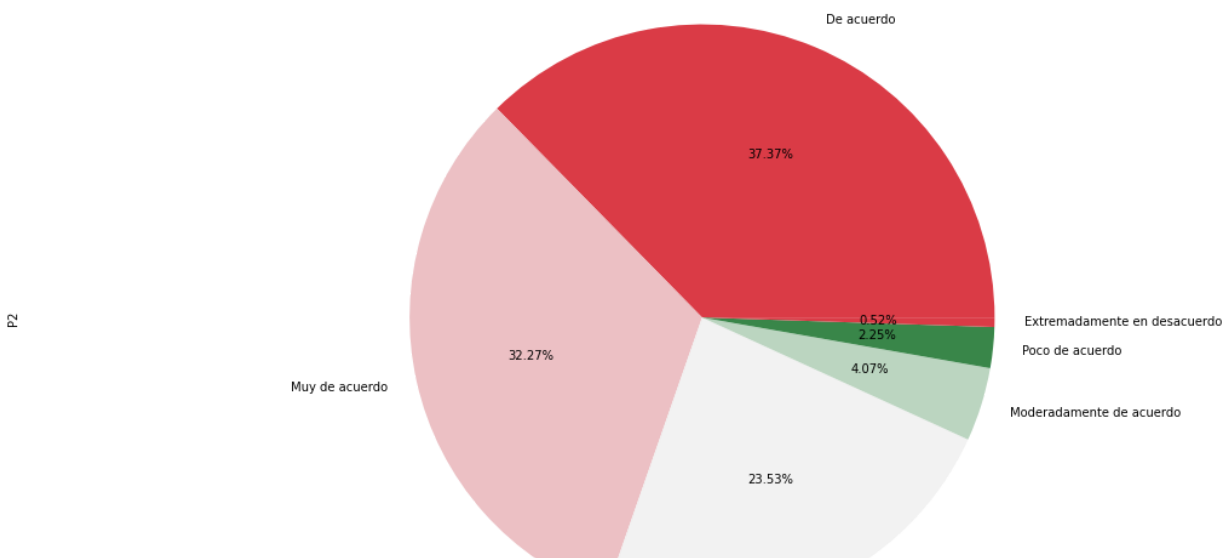
```
cuest
```

	Marca_temporal	P1	P2	P3	P4
27	2022-05-23	Muy de acuerdo	Muy de acuerdo	Muy de acuerdo	Muy de acuerdo
#Muestra de variables con cuestionario corregido					
cuest['P14'].value_counts()					
	Muy de acuerdo	417			
	De acuerdo	349			
	Extremadamente de acuerdo	336			
	Moderadamente de acuerdo	33			
	Poco de acuerdo	14			
	Extremadamente en desacuerdo	7			
Name: P14, dtype: int64					
31	10:44:00.452	De acuerdo	Muy de acuerdo	de acuerdo	Muy de acuerdo
cuest					
	Marca_temporal	P1	P2	P3	P4
27	2022-05-23 09:45:11.003	Muy de acuerdo	Muy de acuerdo	Muy de acuerdo	Muy de acuerdo
28	2022-05-23 10:11:17.308	De acuerdo	De acuerdo	Moderadamente de acuerdo	Muy de acuerdo
29	2022-05-23 10:24:05.790	De acuerdo	Muy de acuerdo	De acuerdo	De acuerdo
30	2022-05-23 10:40:58.829	De acuerdo	De acuerdo	De acuerdo	De acuerdo
31	2022-05-23 10:44:00.452	De acuerdo	Muy de acuerdo	Extremadamente de acuerdo	Muy de acuerdo
...	...	...	...	...	...
1178	2022-07-10 12:29:07.771	Extremadamente de acuerdo	Extremadamente de acuerdo	Extremadamente de acuerdo	Extremadamente de acuerdo
1179	2022-07-11 18:03:01.593	Muy de acuerdo	Muy de acuerdo	De acuerdo	Extremadamente de acuerdo
1180	2022-07-14 16:50:19.157	De acuerdo	De acuerdo	De acuerdo	De acuerdo
1181	2022-07-19 13:09:34.327	Moderadamente de acuerdo	Muy de acuerdo	Muy de acuerdo	Muy de acuerdo
1182	2022-07-20 03:59:37.368	Poco de acuerdo	Poco de acuerdo	Poco de acuerdo	De acuerdo
1156 rows × 31 columns					

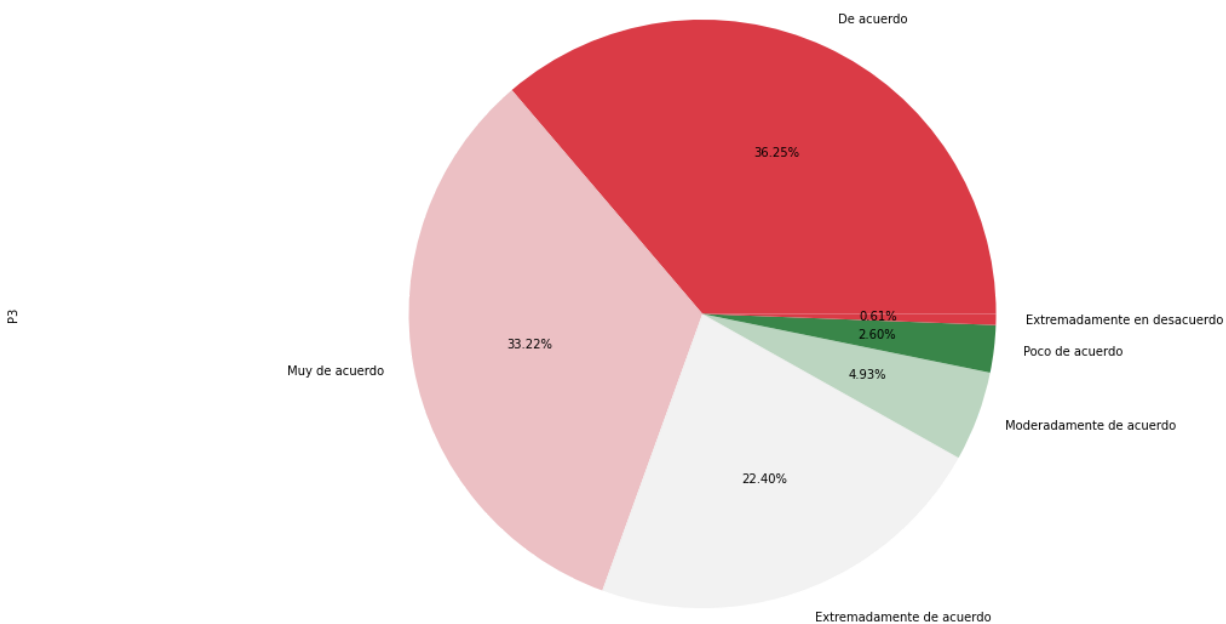
```
cuest['P1'].value_counts().plot(kind='pie', autopct='%0.2f%%',figsize=(20,10))  
plt.axis('equal');
```



```
cuest['P2'].value_counts().plot(kind='pie', autopct='%0.2f%%',figsize=(20,10))  
plt.axis('equal');
```

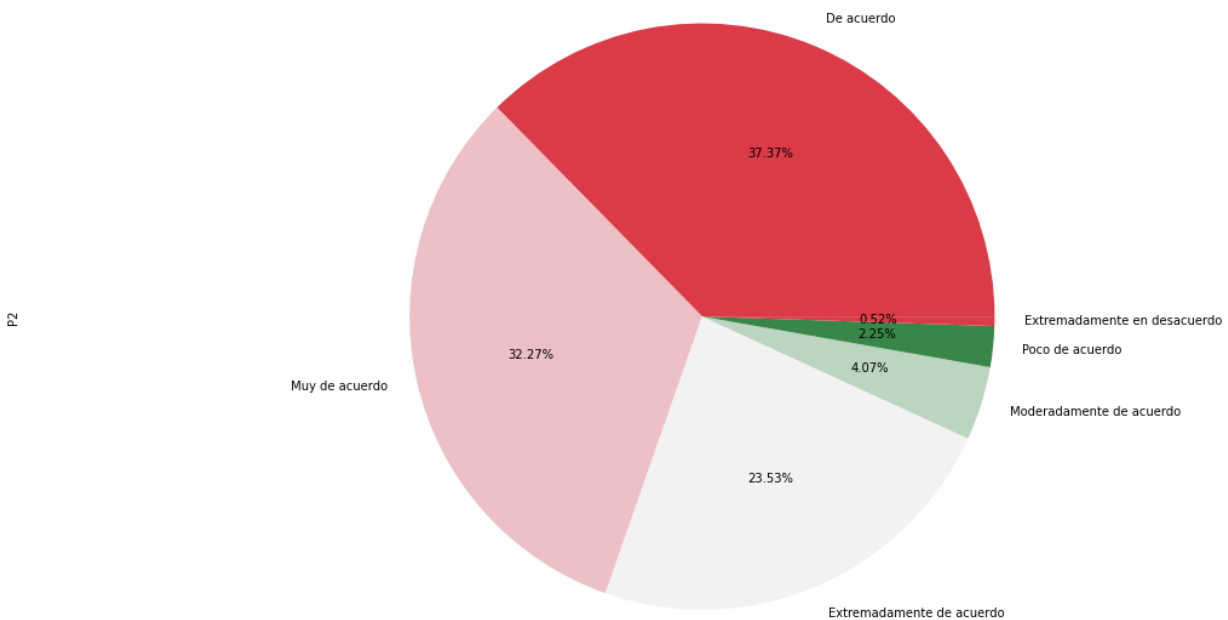


```
cuest['P3'].value_counts().plot(kind='pie', autopct='%0.2f%%',figsize=(20,10))
plt.axis('equal');
```

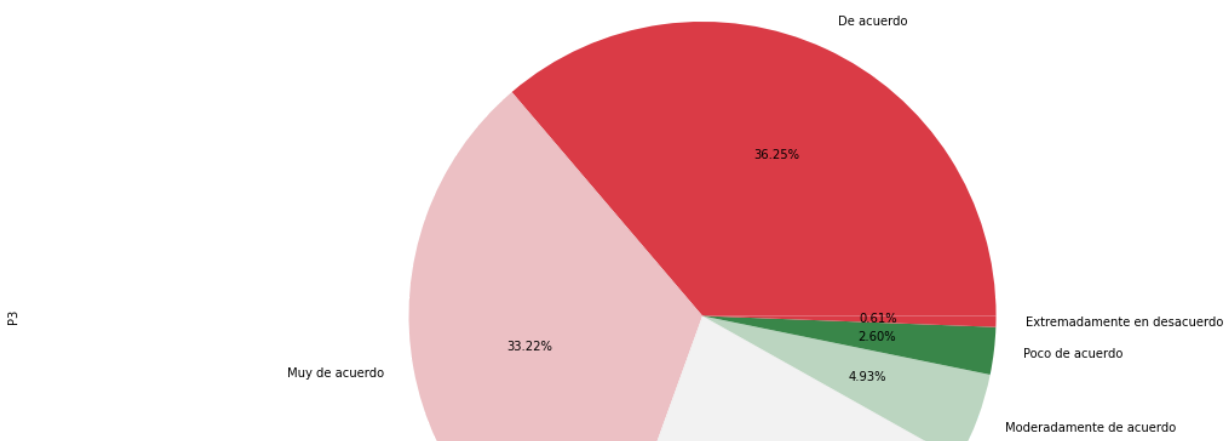


```
cuest['P2'].value_counts().plot(kind='pie', autopct='%0.2f%%',figsize=(20,10))
```

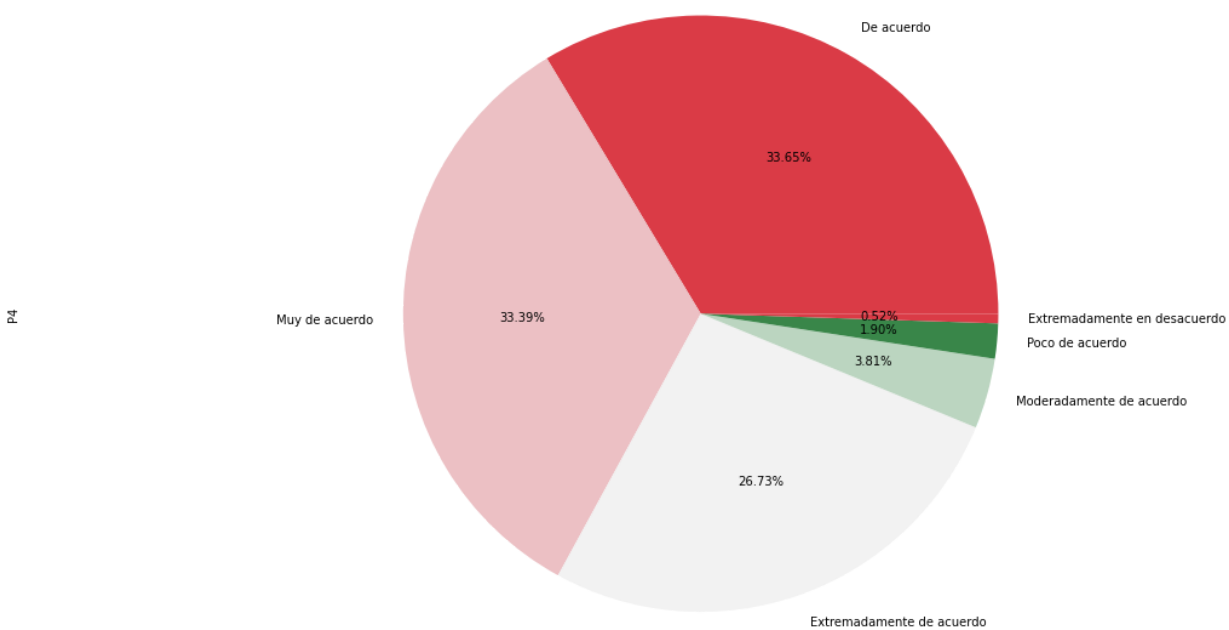
```
plt.axis('equal');
```



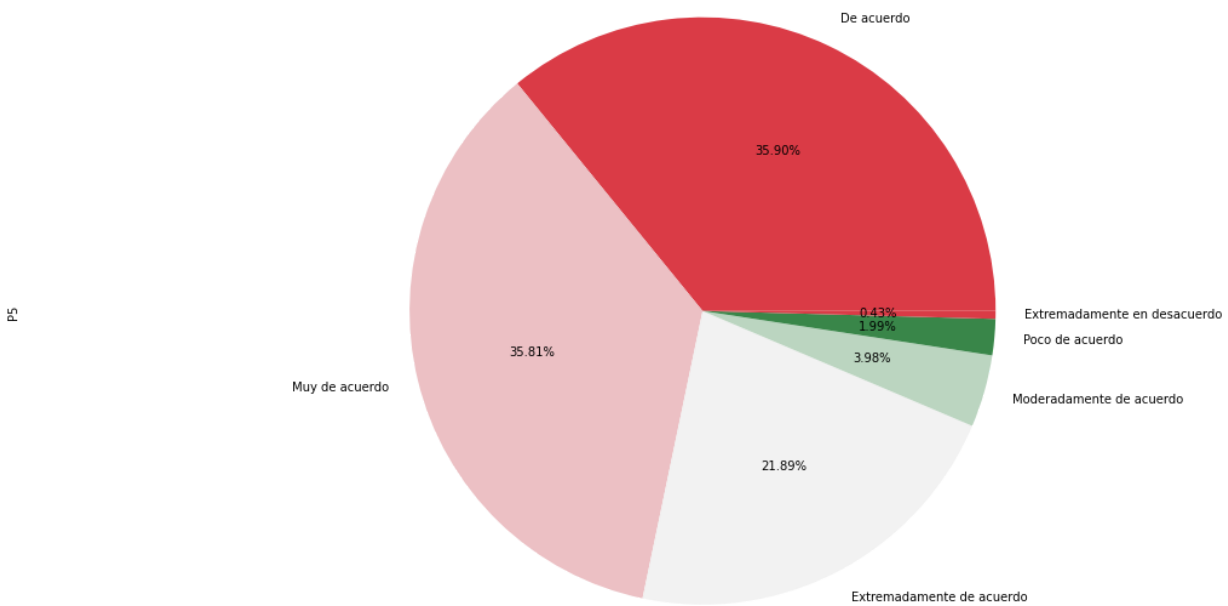
```
cuest['P3'].value_counts().plot(kind='pie', autopct='%.2f%%',figsize=(20,10))  
plt.axis('equal');
```

```
cuest['P4'].value_counts().plot(kind='pie', autopct='%0.2f%%',figsize=(20,10))
plt.axis('equal');
```

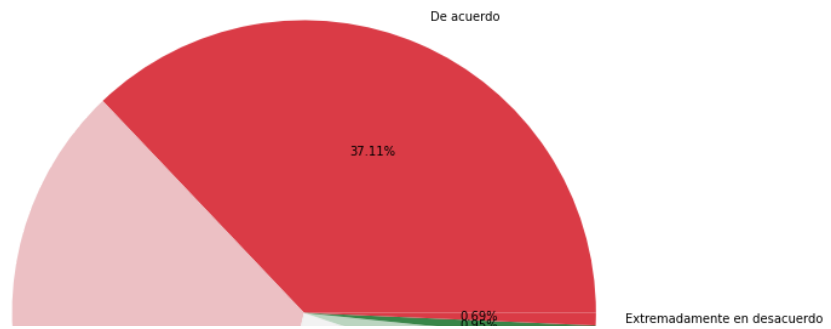


```
cuest['P5'].value_counts().plot(kind='pie', autopct='%0.2f%%',figsize=(20,10))
plt.axis('equal');
```



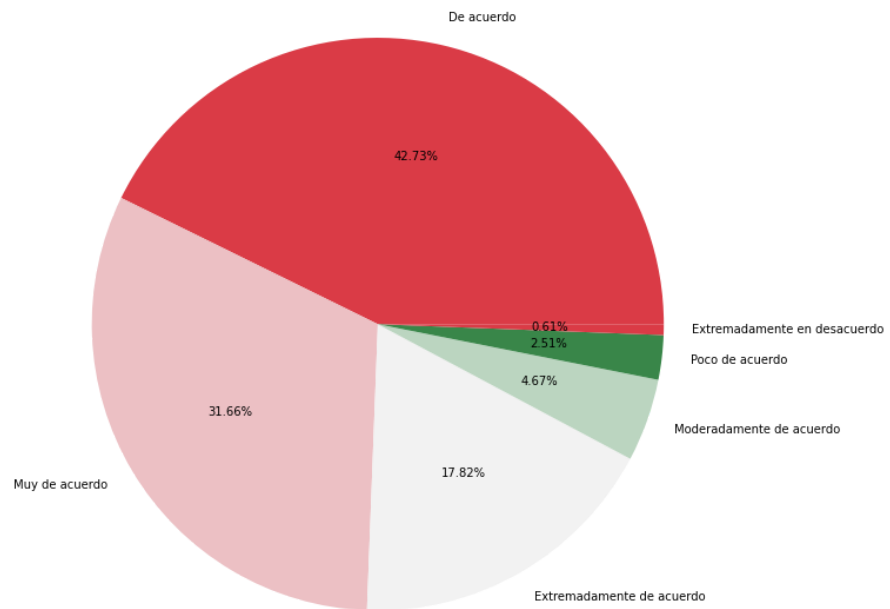
```
cuest['P6'].value_counts().plot(kind='pie', autopct='%0.2f%%',figsize=(20,10))  
plt.axis('equal');
```

P6

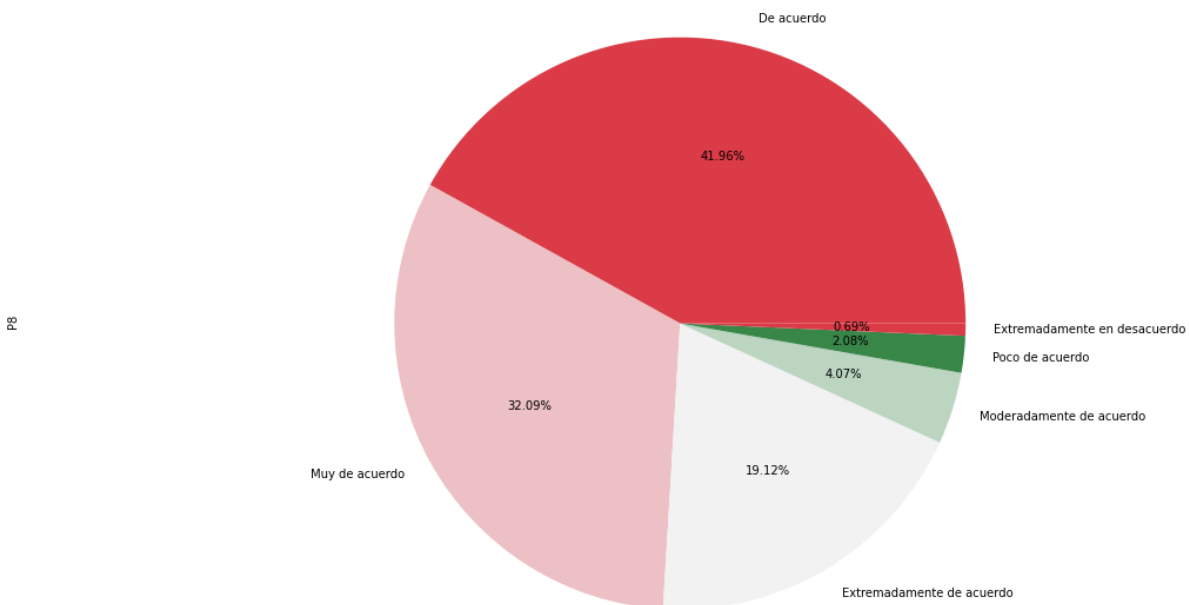


```
cuest['P7'].value_counts().plot(kind='pie', autopct='%0.2f%%', figsize=(20,10))
plt.axis('equal');
```

P7



```
cuest['P8'].value_counts().plot(kind='pie', autopct='%0.2f%%', figsize=(20,10))
plt.axis('equal');
```



```
cuest['P9'].value_counts().plot(kind='pie', autopct='%.2f%%', figsize=(20,10))  
plt.axis('equal');
```



```
cuest.head()
```

	Marca_temporal	P1	P2	P3	P4	P5	P6	
27	2022-05-23 09:45:11.003	Muy de acuerdo	Muy de acuerdo	Muy de acuerdo	Muy de acuerdo	De acuerdo	De acuerdo	M ac
28	2022-05-23 10:11:17.308	De acuerdo	De acuerdo	Moderadamente de acuerdo	Muy de acuerdo	Muy de acuerdo	De acuerdo	ac
29	2022-05-23 10:24:05.790	De acuerdo	Muy de acuerdo	De acuerdo	De acuerdo	De acuerdo	Moderadamente de acuerdo	M ac
30	2022-05-23 10:40:58.829	De acuerdo	De acuerdo	De acuerdo	De acuerdo	De acuerdo	De acuerdo	ac
31	2022-05-23 10:44:00.452	De acuerdo	Muy de acuerdo	Extremadamente de acuerdo	Muy de acuerdo	Muy de acuerdo	Muy de acuerdo	ac

5 rows × 31 columns

Gráficas de barras resumidas de preguntas

```
answering_columns = list (cuest.columns[1:29])

rows = round( len(answering_columns)//2)
fig, axarr = plt.subplots(rows, 2, figsize=(40, 4*rows))

c = 0
r = 0
answers = {'Extremadamente de acuerdo': 6, 'Muy de acuerdo':5, 'De acuerdo':4, 'Moderadamente d

#sns.set_palette(sns.cubehelix_palette(5, start=.5, rot=-.75))
sns.set_palette(sns.diverging_palette(10, 133, sep=80, n=5))

size = len(cuest)
for count,col in enumerate(answering_columns):

    #sns.countplot(base[col].sort_values(), ax=axarr[r][c])
```

```
ax = sns.countplot( cuest.iloc[cuest[col].map(answers).argsort()][col], ax=axarr[r][c] )
axarr[r][c].set_title( col_description[col], fontsize=11)

#ax = (base[col].value_counts()/len(base)*100).sort_index().plot(kind="bar", rot=0)
#ax.set_yticks(np.arange(0, 900, 10))

#ax2 = ax.twinx()
#ax2.set_yticks(np.arange(0, 110, 10)*len(base)/100)

for p in ax.patches:
    ax.annotate('{:.2f}%'.format(p.get_height()/len(cuest) * 100), (p.get_x()+0.15, p.get_

if c >= 1:
    c = 0
    r = r + 1
else:
    c = c + 1
plt.subplots_adjust(hspace=.3)
sns.despine()
```

```

/usr/local/lib/python3.7/dist-packages/seaborn/_decorators.py:43: FutureWarning:
Pass the following variable as a keyword arg: x. From version 0.12, the only valid po
/usr/local/lib/python3.7/dist-packages/seaborn/_decorators.py:43: FutureWarning:
Pass the following variable as a keyword arg: x. From version 0.12, the only valid po
/usr/local/lib/python3.7/dist-packages/seaborn/_decorators.py:43: FutureWarning:
Pass the following variable as a keyword arg: x. From version 0.12, the only valid po
/usr/local/lib/python3.7/dist-packages/seaborn/_decorators.py:43: FutureWarning:
Pass the following variable as a keyword arg: x. From version 0.12, the only valid po
/usr/local/lib/python3.7/dist-packages/seaborn/_decorators.py:43: FutureWarning:
Pass the following variable as a keyword arg: x. From version 0.12, the only valid po
/usr/local/lib/python3.7/dist-packages/seaborn/_decorators.py:43: FutureWarning:

```

Reemplazando variables categóricas a cantidades Ejemplo:

Extremadamente de acuerdo' : 6, 'Muy de acuerdo' : 5.

```
cuest.replace({'Extremadamente de acuerdo': 6, 'Muy de acuerdo':5, 'De acuerdo':4, 'Moderadamen
```

```

/usr/local/lib/python3.7/dist-packages/seaborn/_decorators.py:43: FutureWarning:
cuest

```

	Marca_temporal	P1	P2	P3	P4	P5	P6	P7	P8	P9	...	P21	P22	P23	P24	P25
27	2022-05-23 00:15:44.000	5	5	5	5	4	4	5	5	3	...	5	4	5	5	5

▼ Estadística descriptiva de los datos.

```
2022-05-23
cuest.describe()
```

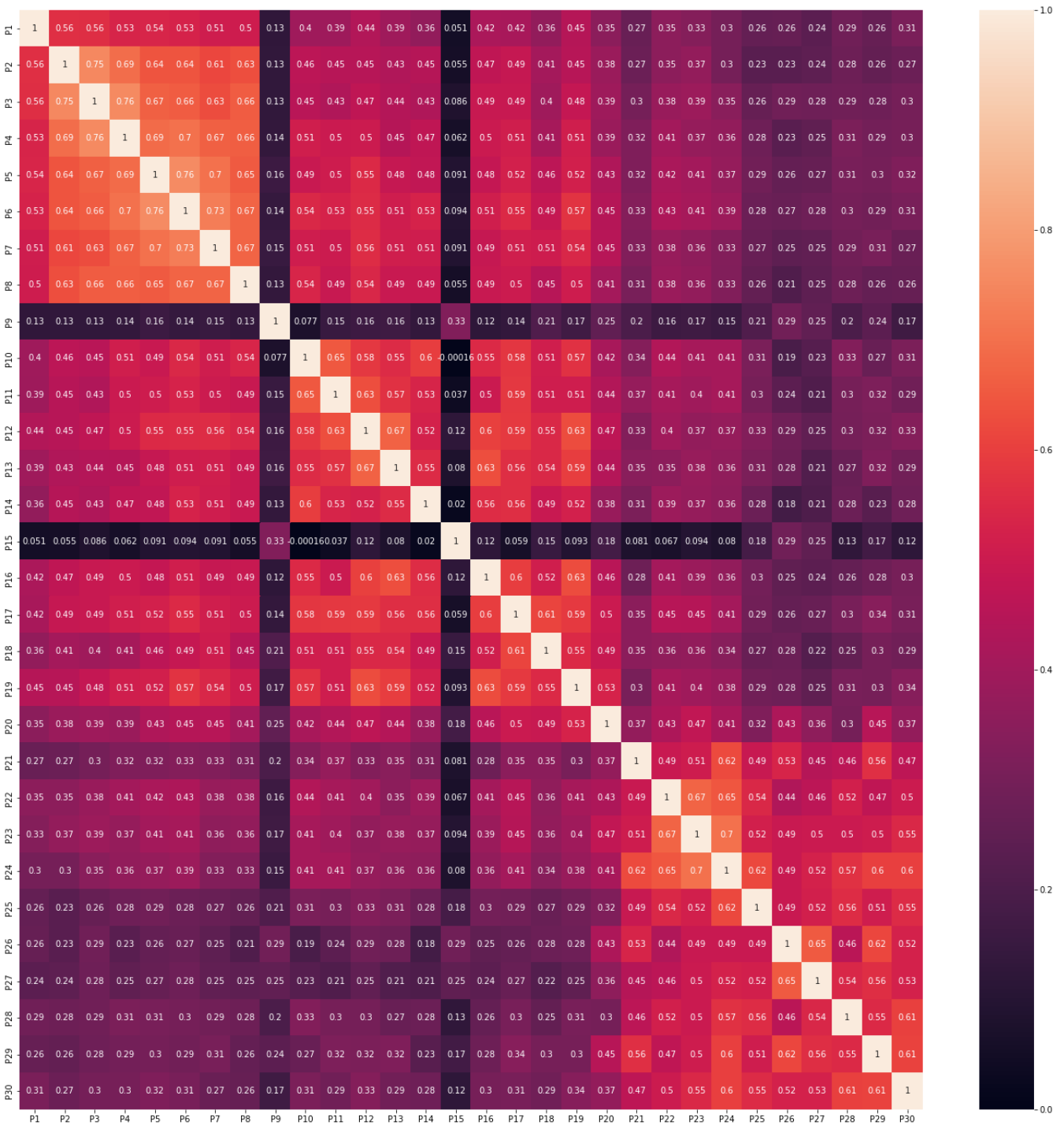
	P1	P2	P3	P4	P5	P6	
count	1156.000000	1156.000000	1156.000000	1156.000000	1156.000000	1156.000000	1156.0
mean	4.506055	4.692042	4.660900	4.776817	4.703287	4.737024	4.5
std	0.949348	0.981371	1.000031	0.980511	0.947646	0.936076	0.9
min	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.0
25%	4.000000	4.000000	4.000000	4.000000	4.000000	4.000000	4.0
50%	4.000000	5.000000	5.000000	5.000000	5.000000	5.000000	4.0
75%	5.000000	5.000000	5.000000	6.000000	5.000000	5.000000	5.0
max	6.000000	6.000000	6.000000	6.000000	6.000000	6.000000	6.0

8 rows × 30 columns

Correlación de Pearson.

```
corr_df = cuest.drop(columns = ['Marca_temporal'],axis = 1).corr(method='pearson')

plt.figure(figsize=(24, 24))
sns.heatmap(corr_df, annot=True)
plt.show()
```

Coeficiente de variación

```
#define function to calculate cv
cv = lambda x: np.std(x, ddof=1) / np.mean(x) * 100

df = cuest.drop(columns = ['Marca_temporal'],axis = 1)
```

```
#calculate CV for each column in data frame
```

```
df1 = df.apply(cv)
```

```
columns_shorten.remove('Marca_temporal')
```

```
#sns.scatterplot(x=list(results.keys() ),y= list(results.values())) )
```

```
plt.figure(figsize=(15,5))
```

```
ax = sns.barplot(x=columns_shorten,y= df1 )
```

```
#plt.title(title)
```

```
#ax = sns.lineplot(x = list(range(0,len(df))),y= df[target].mean(),label='mean ({:0.2f})'.for
```

```
for p in ax.patches:
```

```
    _x = p.get_x() + p.get_width() / 2
```

```
    _y = p.get_y() + p.get_height()+1
```

```
    value = str( int(p.get_height() ) )
```

```
    ax.text(_x, _y, value, ha="center")
```

```
#ax.text(_x, _y, value, ha="left")
```

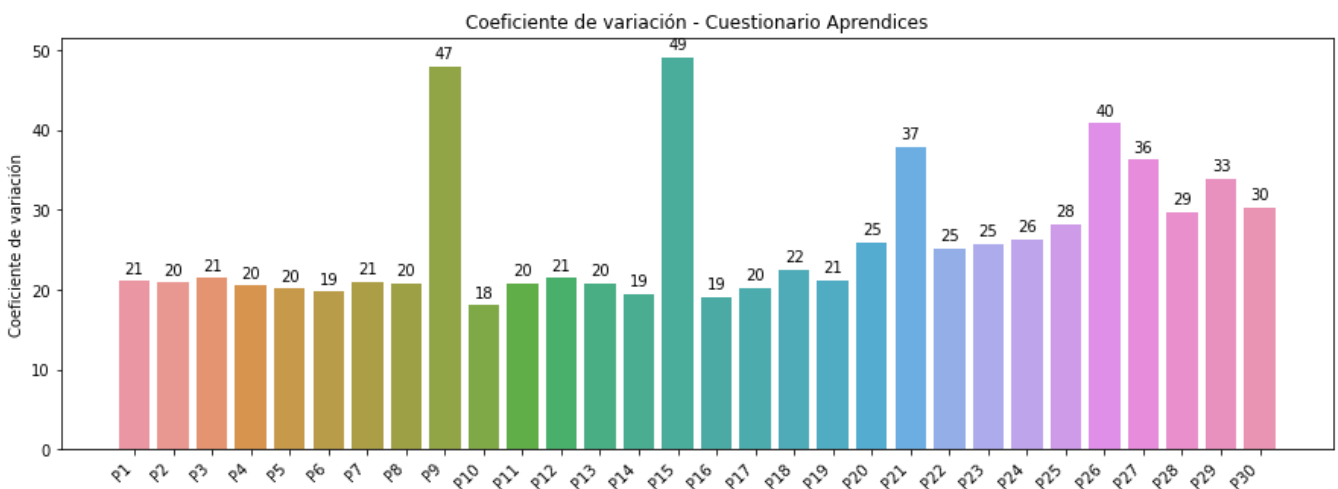
```
ax.set_xticklabels(ax.get_xticklabels(), rotation=45, horizontalalignment='right')
```

```
plt.ylabel('Coeficiente de variación')
```

```
plt.title('Coeficiente de variación - Cuestionario Aprendices')
```

```
plt.plot()
```

```
[]
```



▼ Características que más difieren de la Respuesta Deseada

T-Score: número de desviaciones estándar con respecto al valor medio de la población.

```

#T-Score = (Puntuación bruta - Puntuación promedio) / Desviación estándar * 10 + 50
results = {}

answering_columns_ = answering_columns[:]

for ans in answering_columns_:
    t_s = (cuest[ans].mean() - cuest[ans].max()) / cuest[ans].std() * 10 + 50
    results[ans] = t_s
    print(ans,cuest[ans].mean())

values_sorted = sorted(results, key=results.get)
values_ = []
for r in values_sorted:
    values_.append( float( '{:0.2f}'.format(results[r]) ) ) )

print(values_)

#sns.scatterplot(x=list(results.keys() ),y= list(results.values()))
plt.figure(figsize=(15,5))
ax = sns.barplot(x=values_sorted,y= values_ )
#plt.title(title)
#ax = sns.lineplot(x = list(range(0,len(df))),y= df[target].mean(),label='mean ({:0.2f})'.for
for p in ax.patches:
    _x = p.get_x() + p.get_width() / 2
    _y = p.get_y() + p.get_height()+1
    value = str( int(p.get_height() ) )
    ax.text(_x, _y, value, ha="center")

    #ax.text(_x, _y, value, ha="left")

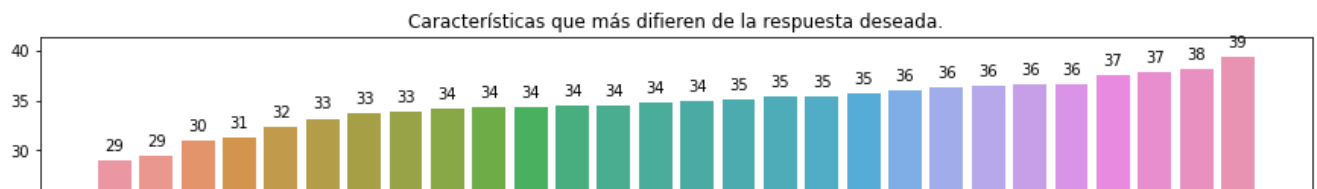
ax.set_xticklabels(ax.get_xticklabels(), rotation=45, horizontalalignment='right')
plt.ylabel('T-Score (%)')
plt.title('Características que más difieren de la respuesta deseada.')
plt.plot()

```

```

P1 4.506055363321799
P2 4.692041522491349
P3 4.660899653979238
P4 4.77681660899654
P5 4.703287197231834
P6 4.737024221453288
P7 4.557958477508651
P8 4.600346020761246
P9 2.9948096885813147
P10 5.038062283737024
P11 4.788062283737024
P12 4.567474048442906
P13 4.6219723183391
P14 4.8711072664359865
P15 2.982698961937716
P16 4.629757785467128
P17 4.678200692041522
P18 4.446366782006921
P19 4.538062283737024
P20 4.250865051903114
P21 3.6660899653979238
P22 4.3088235294117645
P23 4.331314878892734
P24 4.215397923875432
P25 3.9982698961937717
P26 3.374567474048443
P27 3.5717993079584773
P28 4.045847750865052
[29.05, 29.42, 30.95, 31.25, 32.28, 33.19, 33.76, 33.92, 34.16, 34.26, 34.37, 34.47, 34
[]

```



▼ Análisis Factorial

```

from factor_analyzer import FactorAnalyzer

```

Prueba de esfericidad de Bartlett

comprueba si las variables observadas se interrelacionan o no usando la matriz de correlación observada contra la matriz de identidad. Si la prueba resultó estadísticamente insignificante, no debe emplear un análisis factorial.

```

from factor_analyzer.factor_analyzer import calculate_bartlett_sphericity
chi_square_value,p_value=calculate_bartlett_sphericity(df)
chi_square_value, p_value

```

(22590.229248296386, 0.0)

En esta prueba de Bartlett, el valor p es 0. la prueba fue estadísticamente significativa, indicando que la matriz de correlación observada no es una matriz de identidad.

prueba Kaiser-Meyer-Olkin (KMO)

Mide la idoneidad de los datos para el análisis factorial. Determina la adecuación para cada variable observada y para el modelo completo. KMO estima la proporción de varianza entre todas las variables observadas. Identificación de proporción más baja más adecuada para el análisis factorial. Los valores de KMO oscilan entre 0 y 1. Un valor de KMO inferior a 0,6 se considera inadecuado.

```
from factor_analyzer.factor_analyzer import calculate_kmo
```

```
kmo_all,kmo_model=calculate_kmo(df)
```

```
kmo_model
```

```
0.9655206550591892
```

El KMO general para nuestros datos es 0,97 , lo cual es excelente. Este valor indica que puede continuar con su análisis factorial planificado.

Análisis factorial

```
df = df.dropna()
```

```
fa = FactorAnalyzer(n_factors= 5,rotation = 'varimax')
```

```
fa.fit(df)
```

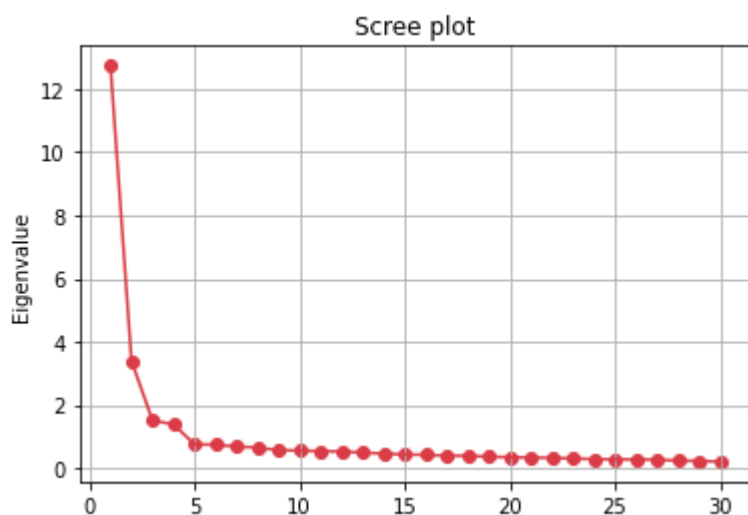
```
FactorAnalyzer(n_factors=5, rotation='varimax', rotation_kwargs={})
```

```
#Get the loadings  
loadings = fa.loadings_
```

```
# Eigen vectores  
ev, v = fa.get_eigenvalues()
```

```
xvals = range(1,df.shape[1]+1)
```

```
plt.scatter(xvals,ev)  
plt.plot(xvals,ev)  
plt.title('Scree plot')  
plt.ylabel('Eigenvalue')  
plt.grid()  
plt.show()
```



```
# análisis factorial  
df_pca = pd.DataFrame.from_records(loadings)  
df_pca['Pregunta'] = columns_shorten  
first_column = df_pca.pop('Pregunta')  
df_pca.insert(0, 'Pregunta', first_column)  
df_pca.columns = ['Pregunta', 'Factor 1', 'Factor 2', 'Factor 3', 'Factor 4', 'Factor 5']  
display(df_pca)
```

	Pregunta	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
0	P1	0.201783	0.282968	0.550510	0.071643	-0.013091
1	P2	0.150767	0.270910	0.754280	0.051353	0.037134
2	P3	0.185252	0.241282	0.793759	0.083644	0.044370
3	P4	0.195128	0.309862	0.769931	0.014768	-0.008701
4	P5	0.194621	0.360972	0.712214	0.081873	0.042575
5	P6	0.184890	0.412644	0.704663	0.072633	0.062847
6	P7	0.149210	0.418649	0.666133	0.114383	0.008632
7	P8	0.151523	0.394283	0.675057	0.034380	0.000924
8	P9	0.167308	0.099945	0.065920	0.439202	0.013048
9	P10	0.244719	0.662112	0.330741	-0.119253	-0.016261
10	P11	0.235828	0.641099	0.309984	-0.013284	-0.000381
11	P12	0.195353	0.701056	0.325743	0.133950	-0.100881
12	P13	0.181358	0.704398	0.269051	0.118678	-0.059631
13	P14	0.187915	0.612599	0.325441	-0.040074	0.008591
14	P15	0.101371	0.022863	0.021633	0.516746	-0.017988
15	P16	0.174471	0.646835	0.325010	0.101352	0.033586
16	P17	0.223851	0.651308	0.334689	0.057513	0.160803
17	P18	0.164769	0.622408	0.256227	0.207824	0.125993
18	P19	0.193059	0.642418	0.345723	0.137533	0.051116
19	P20	0.313628	0.445323	0.248906	0.313262	0.240718

```
# Get variance of each factors
```

```
df_var = pd.DataFrame.from_records(fa.get_factor_variance())
```

```
df_var['Parámetros'] = ['SS Loadings', 'Proportion Var', 'Cumulative Var']
```

```
first_column = df_var.pop('Parámetros')
```

```
df_var.insert(0, 'Parámetros', first_column)
```

```
df_var.columns = ['Parámetros', 'Factor 1', 'Factor 2', 'Factor 3', 'Factor 4', 'Factor 5']
```

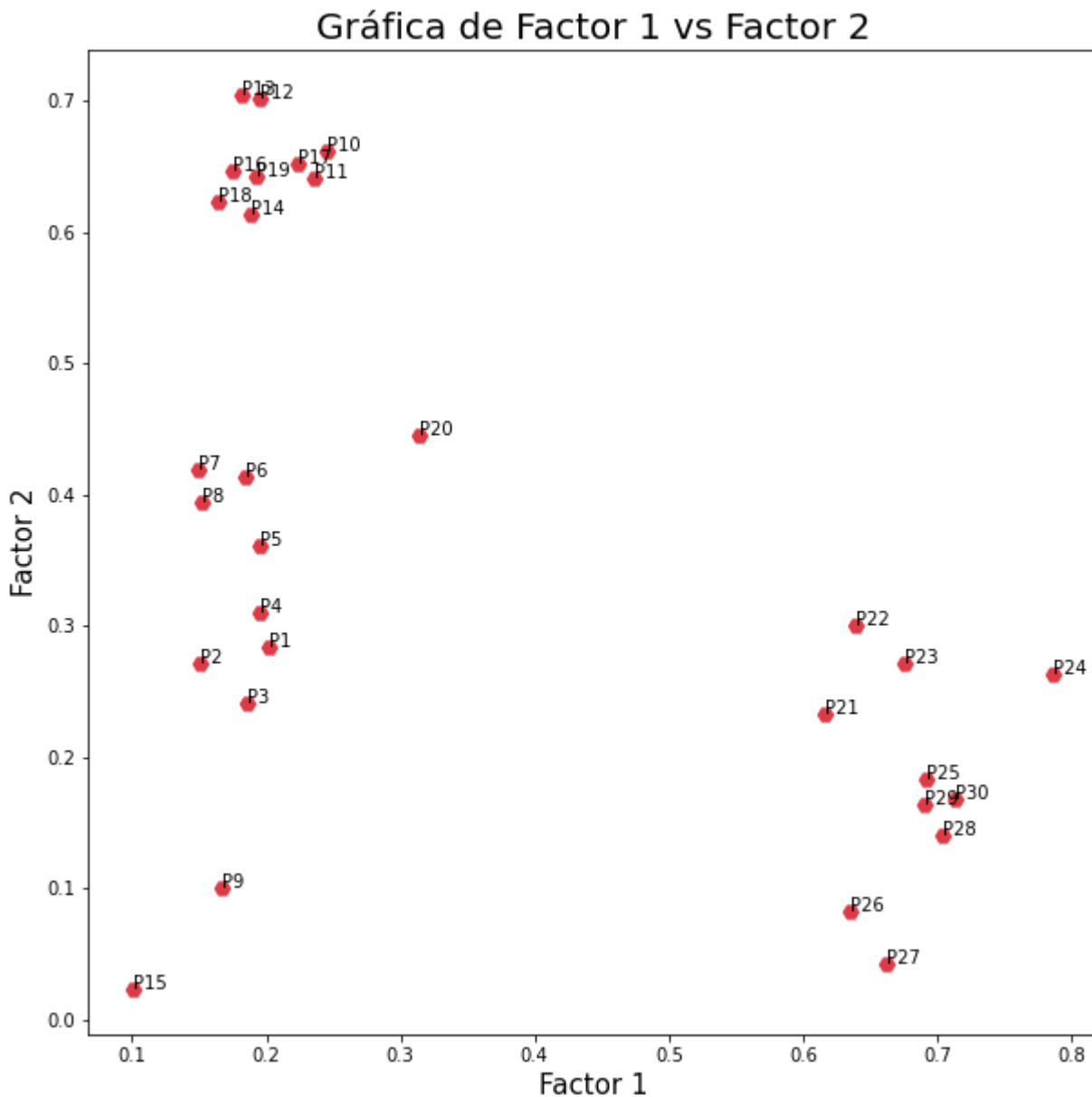
```
display(df_var)
```

	Parámetros	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
0	SS Loadings	5.42191	5.407259	5.228928	1.175595	0.362110
1	Proportion Var	0.18073	0.180242	0.174298	0.039187	0.012070
2	Cumulative Var	0.18073	0.360972	0.535270	0.574456	0.586527

```

#gráfico de factores
fig=plt.figure(figsize=(10,10))
ax=fig.add_subplot(1,1,1)
annotations = df_pca['Pregunta']
ax.set_xlabel("Factor 1",fontsize=15)
ax.set_ylabel("Factor 2", fontsize=15)
ax.set_title("Gráfica de Factor 1 vs Factor 2", fontsize=20)
color_theme=np.array(["CYAN",'GREENYELLOW','mediumpurple'])
ax.scatter(x=df_pca['Factor 1'], y=df_pca['Factor 2'],marker = "H" ,s=70)
for i, label in enumerate(annotations):
    plt.annotate(label, (df_pca['Factor 1'][i], df_pca['Factor 2'][i]))
plt.show()

```



Referencias:

- <https://bit.ly/3PCfZBb>

- <https://bit.ly/3QY5iK7>



▼ Analítica de datos ENCUESTA EQUIPO PEDAGÓGICO

```
pip install pyLDAvis
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Requirement already satisfied: pyLDAvis in /usr/local/lib/python3.7/dist-packages (3.3.1)
Requirement already satisfied: gensim in /usr/local/lib/python3.7/dist-packages (from pyLDAvis) (4.0.0)
Requirement already satisfied: sklearn in /usr/local/lib/python3.7/dist-packages (from pyLDAvis) (0.24.2)
Requirement already satisfied: joblib in /usr/local/lib/python3.7/dist-packages (from pyLDAvis) (1.1.0)
Requirement already satisfied: pandas>=1.2.0 in /usr/local/lib/python3.7/dist-packages (from pyLDAvis) (1.3.4)
Requirement already satisfied: scikit-learn in /usr/local/lib/python3.7/dist-packages (from pyLDAvis) (0.24.2)
Requirement already satisfied: scipy in /usr/local/lib/python3.7/dist-packages (from pyLDAvis) (1.7.3)
Requirement already satisfied: jinja2 in /usr/local/lib/python3.7/dist-packages (from pyLDAvis) (3.0.3)
Requirement already satisfied: numexpr in /usr/local/lib/python3.7/dist-packages (from pyLDAvis) (2.8.1)
Requirement already satisfied: numpy>=1.20.0 in /usr/local/lib/python3.7/dist-packages (from pyLDAvis) (1.21.6)
Requirement already satisfied: fancy in /usr/local/lib/python3.7/dist-packages (from pyLDAvis) (0.3.1)
Requirement already satisfied: future in /usr/local/lib/python3.7/dist-packages (from pyLDAvis) (0.18.2)
Requirement already satisfied: setuptools in /usr/local/lib/python3.7/dist-packages (from pyLDAvis) (57.5.0)
Requirement already satisfied: python-dateutil>=2.7.3 in /usr/local/lib/python3.7/dist-packages (from pyLDAvis) (2.8.2)
Requirement already satisfied: pytz>=2017.3 in /usr/local/lib/python3.7/dist-packages (from pyLDAvis) (2022.1)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.7/dist-packages (from pyLDAvis) (1.16.0)
Requirement already satisfied: smart-open>=1.2.1 in /usr/local/lib/python3.7/dist-packages (from pyLDAvis) (5.2.1)
Requirement already satisfied: MarkupSafe>=0.23 in /usr/local/lib/python3.7/dist-packages (from pyLDAvis) (2.1.1)
Requirement already satisfied: packaging in /usr/local/lib/python3.7/dist-packages (from pyLDAvis) (21.3)
Requirement already satisfied: pyparsing!=3.0.5,>=2.0.2 in /usr/local/lib/python3.7/dist-packages (from pyLDAvis) (3.0.9)
Requirement already satisfied: threadpoolctl>=2.0.0 in /usr/local/lib/python3.7/dist-packages (from pyLDAvis) (3.1.0)
```

```
pip install factor_analyzer
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Requirement already satisfied: factor_analyzer in /usr/local/lib/python3.7/dist-packages (0.11.0)
Requirement already satisfied: pandas in /usr/local/lib/python3.7/dist-packages (from factor_analyzer) (1.3.4)
Requirement already satisfied: scikit-learn in /usr/local/lib/python3.7/dist-packages (from factor_analyzer) (0.24.2)
Requirement already satisfied: scipy in /usr/local/lib/python3.7/dist-packages (from factor_analyzer) (1.7.3)
Requirement already satisfied: numpy in /usr/local/lib/python3.7/dist-packages (from factor_analyzer) (1.21.6)
Requirement already satisfied: python-dateutil>=2.7.3 in /usr/local/lib/python3.7/dist-packages (from factor_analyzer) (2.8.2)
Requirement already satisfied: pytz>=2017.3 in /usr/local/lib/python3.7/dist-packages (from factor_analyzer) (2022.1)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.7/dist-packages (from factor_analyzer) (1.16.0)
Requirement already satisfied: joblib>=0.11 in /usr/local/lib/python3.7/dist-packages (from factor_analyzer) (1.1.0)
Requirement already satisfied: threadpoolctl>=2.0.0 in /usr/local/lib/python3.7/dist-packages (from factor_analyzer) (3.1.0)
```

```
#Importanto bibliotecas
import pandas as pd
import numpy as np
import math
import re
#Nubes de Palabras y editores de texto
```

```

from wordcloud import WordCloud, ImageColorGenerator
from nltk.corpus import stopwords # para stopwords en español
import nltk
from nltk import tokenize
from nltk.tokenize import word_tokenize
import spacy
# Importacion de spacy
import spacy.cli
#Gráfica
import matplotlib.pyplot as plt
import seaborn as sns
# Importando el CountVectorizer
from sklearn.feature_extraction.text import CountVectorizer
import pyLDAvis.sklearn
nltk.download('punkt') # necesario para descargar el tokenizador
nltk.download('stopwords')
spacy.cli.download("es_core_news_sm")
import es_core_news_sm
spc_es = es_core_news_sm.load()

```

```

[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
✓ Download and installation successful
You can now load the package via spacy.load('es_core_news_sm')

```

```
stopwords_es = stopwords.words("spanish")
```

```

from google.colab import drive
drive.mount('/content/drive')

```

Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.moun



```

from google.colab import files
path = "/content/drive/MyDrive/Proyecto SGPS 9748/Analítica Luz Karime_ Regional Cauca/Analít
#Dataframes
equ_pdg= pd.read_excel(path)

```

```
equ_pdg
```

			1. ¿Se generan espacios de intercambio de experiencias investigativas - formativas que puedan convertirse en proyectos de investigación SENNOVA?	Por favor, justifique la respuesta anterior:	2. ¿El equipo pedagógico tiene conocimiento de los proyectos de investigación que se desarrollan a través de SENNOVA en su Centro de formación?	Por favor, justifique la respuesta anterior: .1	3. ¿El equipo SENNOVA socializa con el equipo pedagógico los hallazgos de la revisión o vigilancia tecnológica del contexto, para incorporar los resultados en el proyecto formativo?
	Marca temporal						
0	2022-05-04 12:40:51.987	Si	Sugiere que los proyectos formativos, tenga un...		Si	Los proyectos SENNOVA son presentado al equipo...	Si
1	2022-05-23 10:35:31.738	Si	Como estrategia por parte del EPC de centro, s...		No	Sennova maneja diversas líneas y proyectos que...	No
2	2022-05-23 11:30:31.118	No	pocas veces		Si	En la ultima reunion del EPC se explicaron los...	No
	2022-05-23		Desde el ecosistema			Durante las jornadas de	

▼ EDA(exploratory data analysis)

Muchos de

El equipo de

renombramos columnas

revisa...

reuniones ...

```
columns_descriptions = list(equ_pdg.columns)
```

Sen Técnica

Hay personas

```
columns_shorten = ['Marca_temporal', 'P1', 'P2', 'P3', 'P4', 'P5', 'P6', 'P7', 'P8', 'P9', 'P10', 'P11',
                   'P24', 'P25', 'P26', 'P27', 'P28']
```

```
col_renames = {}
col_description = {}
for desc, short in zip(columns_descriptions, columns_shorten):
    print('\033[1m'+short,':\033[0;0m')
    print('\tDescription: ', desc)
    col_renames[desc] = short
    col_description[short] = desc
```

Marca_temporal :

Description: Marca temporal

P1 :

Description: 1. ¿Se generan espacios de intercambio de experiencias investig

P2 :

Description: Por favor, justifique la respuesta anterior:

P3 :

Description: 2. ¿El equipo pedagógico tiene conocimiento de los proyectos de

P4 :

Description: Por favor, justifique la respuesta anterior: .1

P5 :

Description: 3. ¿El equipo SENNOVA socializa con el equipo pedagógico los ha

P6 :

Description: Por favor justifique la respuesta anterior:

P7 :

Description: 4. Con qué frecuencia el equipo pedagógico implementa las sigui

P8 :

Description: 4. Con qué frecuencia el equipo pedagógico implementa las sigui

P9 :

Description: 1.Determine el nivel de frecuencia de las siguientes estrategia

P10 :

Description: 1.Determine el nivel de frecuencia de las siguientes estrategia

P11 :

Description: 2. Con el fin de implementar los lineamientos de la Dirección d

P12 :

Description: 2. Con el fin de implementar los lineamientos de la Dirección d

P13 :

Description: 2. Con el fin de implementar los lineamientos de la Dirección d

P14 :

Description: 2. Con el fin de implementar los lineamientos de la Dirección d

P15 :

Description: 3. ¿Cuáles son los parámetros que utiliza en su Centro de Form

P16 :

Description: 3. ¿Cuáles son los parámetros que utiliza en su Centro de Form

P17 :

Description: 3. ¿Cuáles son los parámetros que utiliza en su Centro de Form

P18 :

Description: Cuales son los parámetros que utiliza en su Centro de Formació

```

P19 : Description: Cuales son los parámetros que utiliza en su Centro de Formació
P20 : Description: Cuales son los parámetros que utiliza en su Centro de Formació
P21 : Description: 5. En su rol como miembro del Equipo Pedagógico: ¿con qué frecu
P22 : Description: 5. En su rol como miembro del Equipo Pedagógico: ¿con qué frecu
P23 : Description: 5. En su rol como miembro del Equipo Pedagógico: ¿con qué frecu
P24 : Description: 4. Con qué frecuencia el equipo pedagógico implementa las sigui
P25 : Description: 1.Determine el nivel de frecuencia de las siguientes estrategia
P26 : Description: 3. ¿Cuáles son los parámetros que utiliza en su Centro de Form
P27 : Description: Cuales son los parámetros que utiliza en su Centro de Formació
P28 :

```

```

equ_pdg.columns = ['Marca_temporal', 'P1', 'P2', 'P3', 'P4', 'P5', 'P6', 'P7', 'P8', 'P9', 'P10', 'P11',
                  'P24', 'P25', 'P26', 'P27', 'P28']

```

```
equ_pdg.info()
```

```

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 99 entries, 0 to 98
Data columns (total 29 columns):
#   Column                Non-Null Count  Dtype
---  -
0   Marca_temporal        99 non-null    datetime64[ns]
1   P1                     99 non-null    object
2   P2                     99 non-null    object
3   P3                     99 non-null    object
4   P4                     95 non-null    object
5   P5                     99 non-null    object
6   P6                     91 non-null    object
7   P7                     99 non-null    object
8   P8                     99 non-null    object
9   P9                     99 non-null    object
10  P10                    99 non-null    object
11  P11                    99 non-null    object
12  P12                    99 non-null    object
13  P13                    99 non-null    object
14  P14                    99 non-null    object
15  P15                    99 non-null    object
16  P16                    99 non-null    object
17  P17                    99 non-null    object
18  P18                    99 non-null    object
19  P19                    99 non-null    object
20  P20                    99 non-null    object
21  P21                    99 non-null    object
22  P22                    99 non-null    object
23  P23                    99 non-null    object

```

```

24 P24          99 non-null    object
25 P25          99 non-null    object
26 P26          99 non-null    object
27 P27          99 non-null    object
28 P28          99 non-null    object
dtypes: datetime64[ns](1), object(28)
memory usage: 22.6+ KB

```

```
equ_pdg['P4']
```

```

0    Los proyectos SENNOVA son presentado al equipo...
1    Sennova maneja diversas líneas y proyectos que...
2    En la ultima reunion del EPC se explicaron los...
3    Durante las jornada de inducción y reinducción...
4    El equipo de sennova informa en las reuniones ...
...
94   Hay personas que están gestionando estos procesos
95                                     NaN
96   Se socializan los proyectos constantemente
97   se agenda cuando se tienen proyectos por socli...
98   Cada año el equipo de SENNOVA envía la informa...
Name: P4, Length: 99, dtype: object

```

```
equ_pdg['P8']
```

```

0    Rara vez
1    Nunca
2    Rara vez
3    Rara vez
4    Usualmente
...
94   Siempre
95   Usualmente
96   Usualmente
97   Rara vez
98   Rara vez
Name: P8, Length: 99, dtype: object

```

```
equ_pdg['P11'].value_counts()
```

```

Usualmente    44
Rara vez      30
Siempre       21
Nunca         4
Name: P11, dtype: int64

```

```
#-----Organización por tipo de variables-----
```

```
#variables textuales
```

```
equ_pdg_text_1 = equ_pdg[['P2','P4','P6']]
```

```
#Variables textuales si/no
```

```
equ_pdg_text_2 = equ_pdg[['P1','P3','P5']]
```

```
#Variables con escala likert
```

```
equ_pdg_likert = equ_pdg[['P7','P8','P9','P10','P11','P12','P13','P14','P15','P16','P17','P18',
                          'P21','P22','P23','P24','P25','P26','P27','P28']]
```

▾ variables equ_pdg_likert

```
equ_pdg_likert.replace({'Siempre, Usualmente': np.nan, 'Rara vez, Nunca':np.nan,'Usualmente, Siempre, Nunca':np.nan},inplace=True)
```

/usr/local/lib/python3.7/dist-packages/pandas/core/frame.py:5244: SettingWithCopyWarning: A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_method=method,



```
equ_pdg_likert = equ_pdg_likert.dropna()
```

```
equ_pdg_likert.info()
```

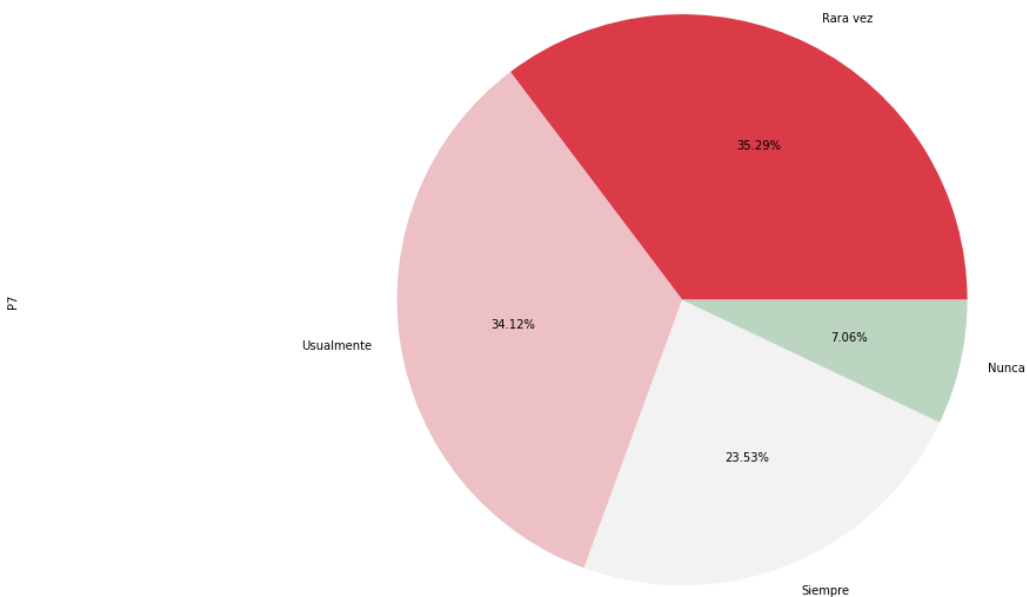
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 85 entries, 0 to 98
Data columns (total 22 columns):
#   Column  Non-Null Count  Dtype
---  -
0    P7      85 non-null      object
1    P8      85 non-null      object
2    P9      85 non-null      object
3    P10     85 non-null      object
4    P11     85 non-null      object
5    P12     85 non-null      object
6    P13     85 non-null      object
7    P14     85 non-null      object
8    P15     85 non-null      object
9    P16     85 non-null      object
10   P17     85 non-null      object
11   P18     85 non-null      object
12   P19     85 non-null      object
13   P20     85 non-null      object
14   P21     85 non-null      object
15   P22     85 non-null      object
16   P23     85 non-null      object
17   P24     85 non-null      object
18   P25     85 non-null      object
19   P26     85 non-null      object
20   P27     85 non-null      object
21   P28     85 non-null      object
```



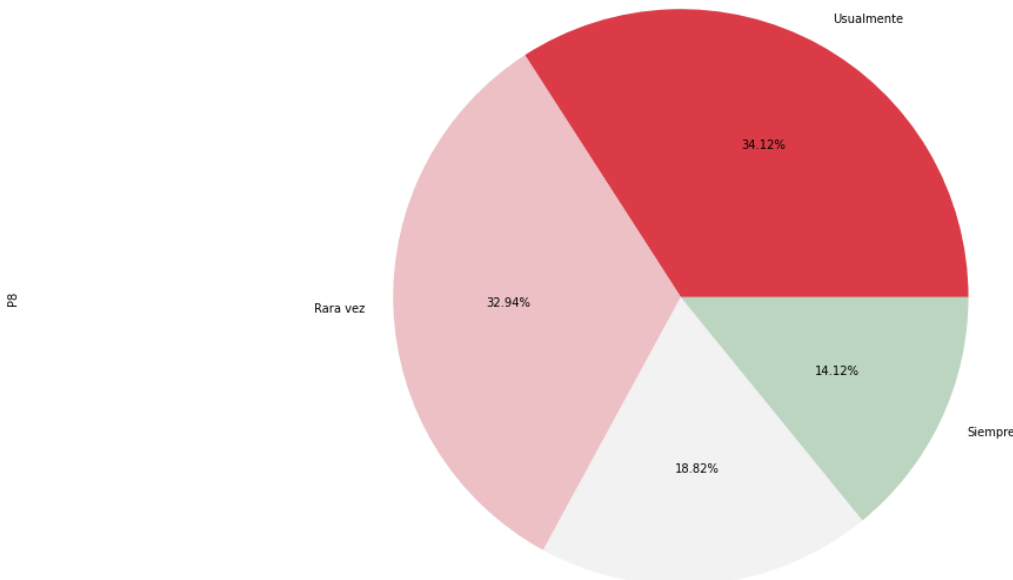
```
dtypes: object(22)
```

```
memory usage: 15.2+ KB
```

```
equ_pdg_likert['P7'].value_counts().plot(kind='pie', autopct='%0.2f%%',figsize=(20,10))  
plt.axis('equal');
```



```
equ_pdg_likert['P8'].value_counts().plot(kind='pie', autopct='%0.2f%%',figsize=(20,10))  
plt.axis('equal');
```



Gráficas de barras resumidas de preguntas

```

answering_columns = list (equ_pdg_likert.columns[1:21])

rows = round( len(answering_columns)//2)
fig, axarr = plt.subplots(rows, 2, figsize=(40, 4*rows))

c = 0
r = 0
answers = {'Siempre': 4, 'Usualmente': 3, 'Rara vez': 2, 'Nunca':1}

#sns.set_palette(sns.cubehelix_palette(5, start=.5, rot=-.75))
sns.set_palette(sns.diverging_palette(10, 133, sep=80, n=5))

size = len(equ_pdg_likert)
for count,col in enumerate(answering_columns):

    #sns.countplot(base[col].sort_values(), ax=axarr[r][c])

    ax = sns.countplot( equ_pdg_likert.iloc[equ_pdg_likert[col].map(answers).argsort()][col],
    axarr[r][c].set_title( col_description[col], fontsize=11)

    #ax = (base[col].value_counts()/len(base)*100).sort_index().plot(kind="bar", rot=0)
    #ax.set_yticks(np.arange(0, 900, 10))

    #ax2 = ax.twinx()
    #ax2.set_yticks(np.arange(0, 110, 10)*len(base)/100)

    for p in ax.patches:
        ax.annotate('{:.2f}%'.format(p.get_height()/len(equ_pdg_likert) * 100), (p.get_x()+0.

```

```
if c >= 1:
    c = 0
    r = r + 1
else:
    c = c + 1
plt.subplots_adjust(hspace=.3)
sns.despine()
```

[illegible]

Reemplazando variables categóricas a cantidades 'Siempre': 4,'Usualmente': 3,'Rara vez': 2,'Nunca':1

Futurewarning

```
equ_pdg_likert.replace({'Siempre': 4, 'Usualmente': 3, 'Rara vez': 2, 'Nunca': 1}, inplace=True)
```

```
/usr/local/lib/python3.7/dist-packages/seaborn/_decorators.py:43: FutureWarning: Pass the following variables as keyword arguments: {'x': 'x', 'y': 'y'}. This warning will disappear after six months without action.
```

equ_pdg_likert

	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16	...	P19	P20	P21	P22	P23	P24	P2
0	4	2	3	3	3	3	3	3	3	2	...	3	3	3	3	3	3	
1	2	1	3	3	3	2	2	2	2	2	...	3	2	2	3	2	1	
2	2	2	2	3	2	2	3	3	3	2	...	2	2	2	2	2	2	
3	4	2	2	2	3	2	2	3	3	3	...	2	2	3	2	4	2	
4	3	3	4	3	3	3	3	3	4	3	...	3	3	3	3	3	3	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
93	4	3	2	3	3	3	3	4	3	3	...	3	3	4	3	4	4	
94	4	4	4	4	4	4	4	4	4	4	...	4	4	4	4	4	4	
95	2	3	4	4	2	2	2	2	3	3	...	3	3	3	3	3	2	
97	3	2	4	3	4	3	2	3	4	3	...	3	3	2	2	3	2	
98	2	2	3	3	3	2	2	2	3	3	...	2	2	3	3	3	3	

85 rows × 22 columns



Estadística descriptiva de los datos.



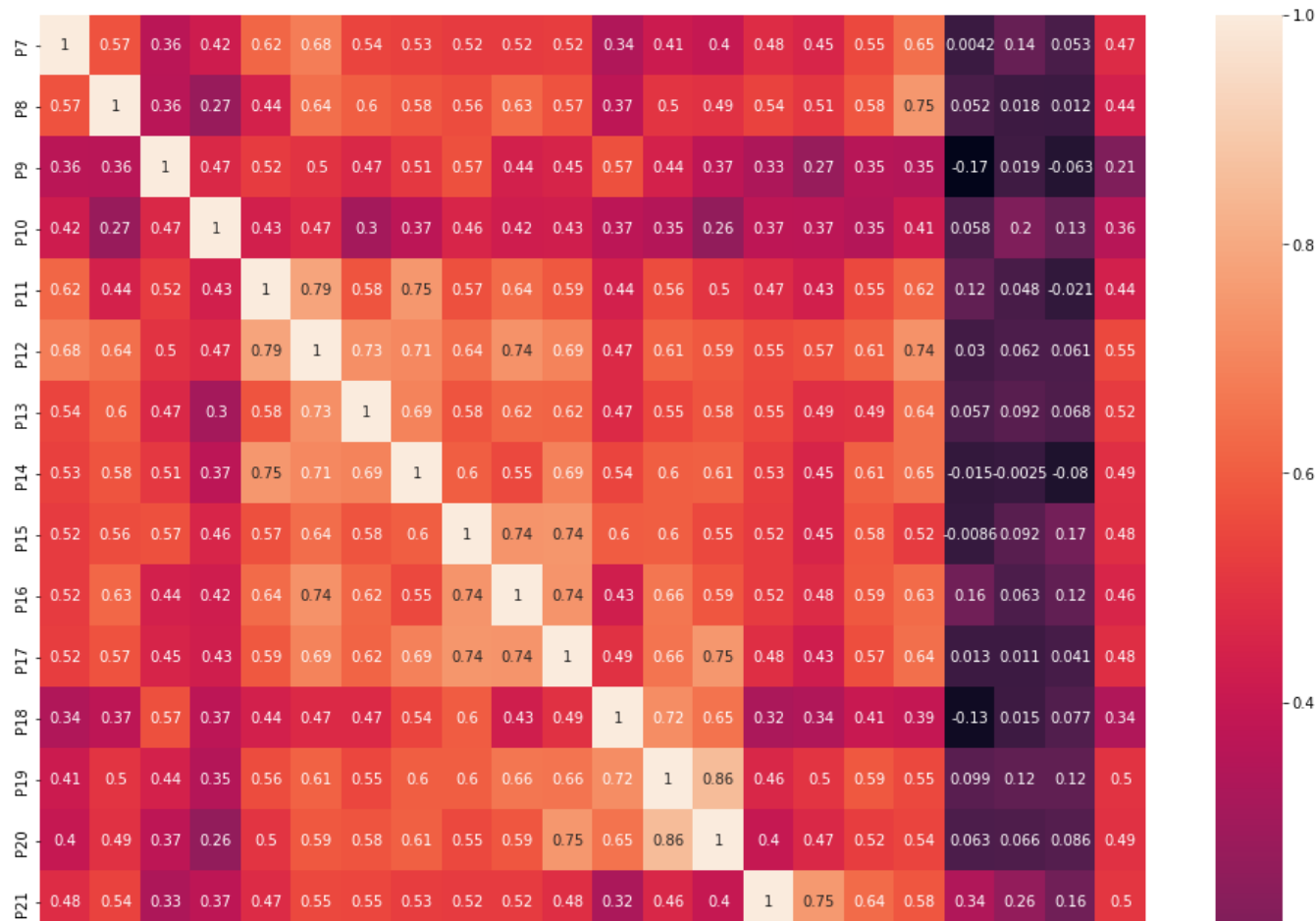
```
equ_pdg_likert.describe()
```

	P7	P8	P9	P10	P11	P12	P13	
count	85.000000	85.000000	85.000000	85.000000	85.000000	85.000000	85.000000	85.000000
mean	2.741176	2.435294	3.082353	2.952941	2.835294	2.623529	2.400000	2.741176
std	0.901757	0.956769	0.789993	0.722197	0.814264	0.925517	0.847967	0.847967
min	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
25%	2.000000	2.000000	3.000000	3.000000	2.000000	2.000000	2.000000	2.000000
50%	3.000000	2.000000	3.000000	3.000000	3.000000	3.000000	2.000000	3.000000
75%	3.000000	3.000000	4.000000	3.000000	3.000000	3.000000	3.000000	3.000000
max	4.000000	4.000000	4.000000	4.000000	4.000000	4.000000	4.000000	4.000000

Correlación de Pearson.

```
corr_df = equ_pdg_likert.corr(method='pearson')
```

```
plt.figure(figsize=(17, 17))
sns.heatmap(corr_df, annot=True)
plt.show()
```



Coeficiente de variación



equ_pdg_likert

```

P7 P8 P9 P10 P11 P12 P13 P14 P15 P16 ... P19 P20 P21 P22 P23 P24 P2
#define function to calculate cv
cv = lambda x: np.std(x, ddof=1) / np.mean(x) * 100

#calculate CV for each column in data frame
df = equ_pdg_likert.apply(cv)

3 4 2 2 2 3 2 2 3 3 3 ... 2 2 3 2 4 2

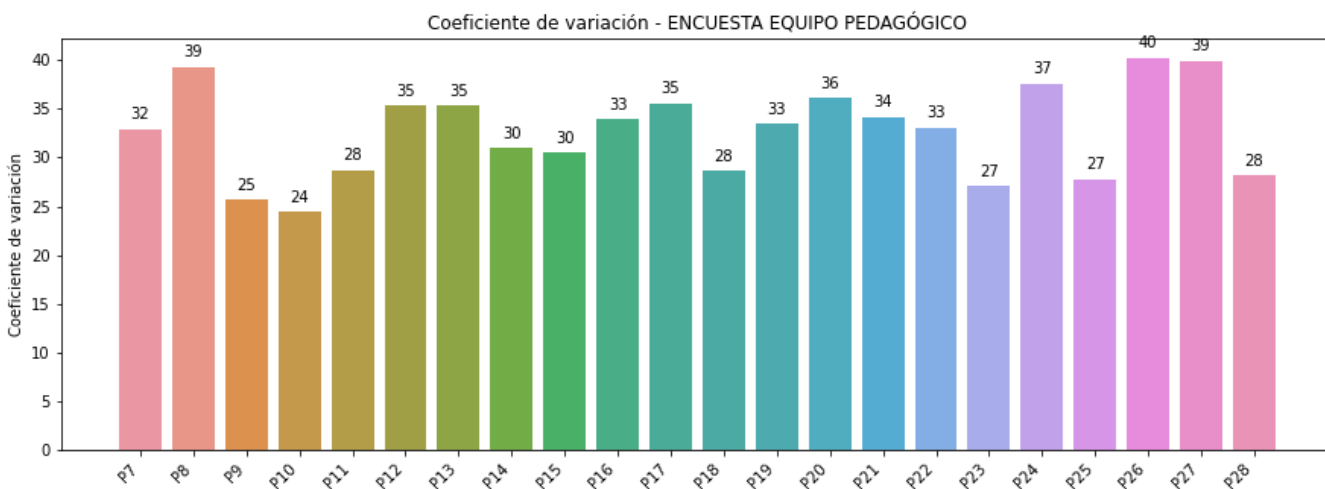
#sns.scatterplot(x=list(results.keys() ),y= list(results.values()))
plt.figure(figsize=(15,5))
ax = sns.barplot(x=equ_pdg_likert.columns,y= df )
#plt.title(title)
#ax = sns.lineplot(x = list(range(0,len(df))),y= df[target].mean(),label='mean ({:0.2f})'.for
for p in ax.patches:
    _x = p.get_x() + p.get_width() / 2
    _y = p.get_y() + p.get_height()+1
    value = str( int(p.get_height() ) )
    ax.text(_x, _y, value, ha="center")

#ax.text(_x, _y, value, ha="left")

ax.set_xticklabels(ax.get_xticklabels(), rotation=45, horizontalalignment='right')
plt.ylabel('Coeficiente de variación')
plt.title('Coeficiente de variación - ENCUESTA EQUIPO PEDAGÓGICO')
plt.plot()

```

[[



▼ Características que más difieren de la Respuesta Deseada

T-Score: número de desviaciones estándar con respecto al valor medio de la población.

```

#T-Score = (Puntuación bruta - Puntuación promedio) / Desviación estándar * 10 + 50
results = {}

answering_columns_ = answering_columns[:]

for ans in answering_columns_:
    t_s = (equ_pdg_likert[ans].mean() - equ_pdg_likert[ans].max()) / equ_pdg_likert[ans].std(
    results[ans] = t_s
    print(ans,equ_pdg_likert[ans].mean())

values_sorted = sorted(results, key=results.get)
values_ = []
for r in values_sorted:
    values_.append( float( '{:0.2f}'.format(results[r]) ) ) )

print(values_)

#sns.scatterplot(x=list(results.keys() ),y= list(results.values()))
plt.figure(figsize=(15,6))
ax = sns.barplot(x=values_sorted,y= values_ )
#plt.title(title)
#ax = sns.lineplot(x = list(range(0,len(df))),y= df[target].mean(),label='mean ({:0.2f})'.for
for p in ax.patches:
    _x = p.get_x() + p.get_width() / 2
    _y = p.get_y() + p.get_height()+1
    value = str( int(p.get_height() ) )
    ax.text(_x, _y, value, ha="center")

    #ax.text(_x, _y, value, ha="left")

ax.set_xticklabels(ax.get_xticklabels(), rotation=45, horizontalalignment='right')
plt.ylabel('T-Score (%)')
plt.title('Características que más difieren de la respuesta deseada.')
plt.plot()

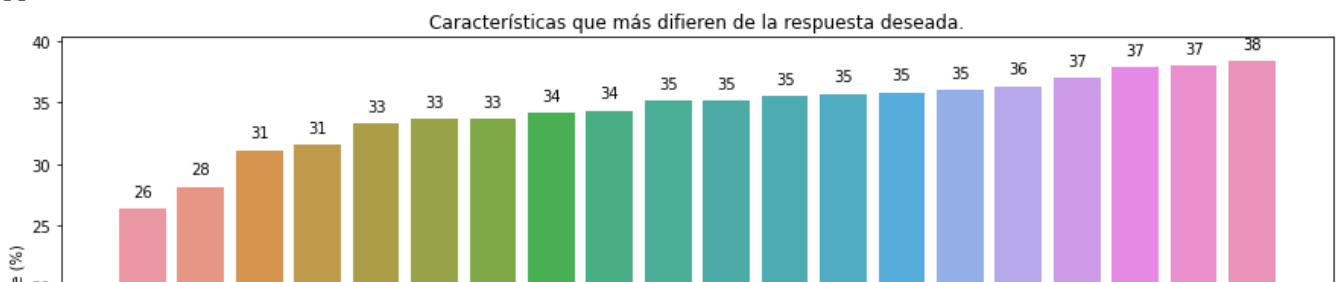
```



```

P8 2.4352941176470586
P9 3.0823529411764707
P10 2.9529411764705884
P11 2.835294117647059
P12 2.623529411764706
P13 2.4
P14 2.7411764705882353
P15 2.9176470588235293
P16 2.552941176470588
P17 2.6588235294117646
P18 2.9176470588235293
P19 2.611764705882353
P20 2.5176470588235293
P21 2.7058823529411766
P22 2.635294117647059
P23 2.9176470588235293
P24 2.7529411764705882
P25 2.6470588235294117
P26 2.1294117647058823
P27 2.0588235294117645
[26.37, 28.15, 31.13, 31.6, 33.29, 33.65, 33.67, 34.12, 34.33, 35.13, 35.14, 35.5, 35.7,
[]

```



▼ Análisis Factorial

```

from factor_analyzer import FactorAnalyzer

```

Prueba de esfericidad de Bartlett

comprueba si las variables observadas se interrelacionan o no usando la matriz de correlación observada contra la matriz de identidad. Si la prueba resultó estadísticamente insignificante, no debe emplear un análisis factorial.

```

from factor_analyzer.factor_analyzer import calculate_bartlett_sphericity
chi_square_value, p_value = calculate_bartlett_sphericity(equ_pdg_likert)
chi_square_value, p_value

```

```
(1454.749954181742, 2.6021288753826213e-176)
```

En esta prueba de Bartlett, el valor p es cercano a 0. la prueba fue estadísticamente significativa, indicando que la matriz de correlación observada no es una matriz de identidad.

prueba Kaiser-Meyer-Olkin (KMO)

Mide la idoneidad de los datos para el análisis factorial. Determina la adecuación para cada variable observada y para el modelo completo. KMO estima la proporción de varianza entre todas las variables observadas. Identificación de proporción más baja más adecuada para el análisis factorial. Los valores de KMO oscilan entre 0 y 1. Un valor de KMO inferior a 0,6 se considera inadecuado.

```
from factor_analyzer.factor_analyzer import calculate_kmo
```

En esta prueba de Bartlett, el valor p es 0. la prueba fue estadísticamente significativa, indicando que la matriz de correlación observada no es una matriz de identidad.

```
kmo_all,kmo_model=calculate_kmo(equ_pdg_likert)
```

```
/usr/local/lib/python3.7/dist-packages/factor_analyzer/utils.py:249: UserWarning: The inverse of the variance-covariance matrix is not positive definite. The matrix is singular and cannot be inverted.
warnings.warn('The inverse of the variance-covariance matrix is not positive definite. The matrix is singular and cannot be inverted.'
```

```
kmo_model
```

```
0.877778922567115
```

El KMO general para nuestros datos es 0,88 , lo cual es excelente. Este valor indica que puede continuar con su análisis factorial planificado.

Análisis factorial

```
equ_pdg_likert = equ_pdg_likert.dropna()
```

```
fa = FactorAnalyzer(n_factors= 5,rotation = 'varimax')
```

```
fa.fit(equ_pdg_likert)
```

```
/usr/local/lib/python3.7/dist-packages/factor_analyzer/factor_analyzer.py:356: DeprecationWarning: sp.sum(residual) is deprecated. Use np.sum(residual) instead.
error = sp.sum(residual)
/usr/local/lib/python3.7/dist-packages/factor_analyzer/factor_analyzer.py:356: DeprecationWarning: sp.sum(residual) is deprecated. Use np.sum(residual) instead.
error = sp.sum(residual)
/usr/local/lib/python3.7/dist-packages/factor_analyzer/factor_analyzer.py:356: DeprecationWarning: sp.sum(residual) is deprecated. Use np.sum(residual) instead.
error = sp.sum(residual)
/usr/local/lib/python3.7/dist-packages/factor_analyzer/factor_analyzer.py:356: DeprecationWarning: sp.sum(residual) is deprecated. Use np.sum(residual) instead.
error = sp.sum(residual)
```

```

/usr/local/lib/python3.7/dist-packages/factor_analyzer/factor_analyzer.py:356: Deprec
error = sp.sum(residual)
/usr/local/lib/python3.7/dist-packages/factor_analyzer/factor_analyzer.py:356: Deprec
error = sp.sum(residual)
/usr/local/lib/python3.7/dist-packages/factor_analyzer/factor_analyzer.py:356: Deprec
error = sp.sum(residual)
/usr/local/lib/python3.7/dist-packages/factor_analyzer/factor_analyzer.py:356: Deprec
error = sp.sum(residual)
/usr/local/lib/python3.7/dist-packages/factor_analyzer/factor_analyzer.py:356: Deprec
error = sp.sum(residual)
/usr/local/lib/python3.7/dist-packages/factor_analyzer/factor_analyzer.py:356: Deprec
error = sp.sum(residual)
/usr/local/lib/python3.7/dist-packages/factor_analyzer/factor_analyzer.py:356: Deprec
error = sp.sum(residual)
/usr/local/lib/python3.7/dist-packages/factor_analyzer/factor_analyzer.py:356: Deprec
error = sp.sum(residual)
/usr/local/lib/python3.7/dist-packages/factor_analyzer/factor_analyzer.py:356: Deprec
error = sp.sum(residual)
/usr/local/lib/python3.7/dist-packages/factor_analyzer/factor_analyzer.py:356: Deprec
error = sp.sum(residual)
/usr/local/lib/python3.7/dist-packages/factor_analyzer/factor_analyzer.py:356: Deprec
error = sp.sum(residual)
/usr/local/lib/python3.7/dist-packages/factor_analyzer/factor_analyzer.py:356: Deprec
error = sp.sum(residual)
/usr/local/lib/python3.7/dist-packages/factor_analyzer/factor_analyzer.py:356: Deprec
error = sp.sum(residual)
/usr/local/lib/python3.7/dist-packages/factor_analyzer/factor_analyzer.py:356: Deprec
error = sp.sum(residual)
/usr/local/lib/python3.7/dist-packages/factor_analyzer/factor_analyzer.py:356: Deprec
error = sp.sum(residual)
/usr/local/lib/python3.7/dist-packages/factor_analyzer/factor_analyzer.py:356: Deprec
error = sp.sum(residual)
/usr/local/lib/python3.7/dist-packages/factor_analyzer/factor_analyzer.py:356: Deprec
error = sp.sum(residual)
/usr/local/lib/python3.7/dist-packages/factor_analyzer/factor_analyzer.py:356: Deprec
error = sp.sum(residual)
/usr/local/lib/python3.7/dist-packages/factor_analyzer/factor_analyzer.py:356: Deprec
error = sp.sum(residual)
/usr/local/lib/python3.7/dist-packages/factor_analyzer/factor_analyzer.py:356: Deprec
error = sp.sum(residual)
/usr/local/lib/python3.7/dist-packages/factor_analyzer/factor_analyzer.py:356: Deprec
error = sp.sum(residual)
/usr/local/lib/python3.7/dist-packages/factor_analyzer/factor_analyzer.py:356: Deprec
error = sp.sum(residual)
/usr/local/lib/python3.7/dist-packages/factor_analyzer/factor_analyzer.py:356: Deprec
error = sp.sum(residual)

```

```

#Get the loadings
loadings = fa.loadings_

```

```

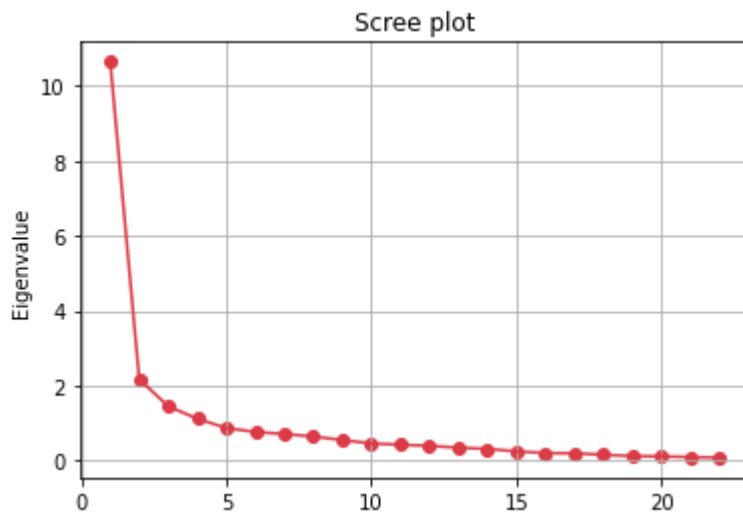
# Eigen vectores

```

```
ev, v = fa.get_eigenvalues()
```

```
xvals = range(1, equ_pdg_likert.shape[1]+1)
```

```
plt.scatter(xvals, ev)  
plt.plot(xvals, ev)  
plt.title('Scree plot')  
plt.ylabel('Eigenvalue')  
plt.grid()  
plt.show()
```



```
# componente principales  
df_pca = pd.DataFrame.from_records(loadings)  
df_pca['Pregunta'] = equ_pdg_likert.columns  
first_column = df_pca.pop('Pregunta')  
df_pca.insert(0, 'Pregunta', first_column)  
df_pca.columns = ['Pregunta', 'Factor 1', 'Factor 2', 'Factor 3', 'Factor 4', 'Factor 5']  
display(df_pca)
```

	Pregunta	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
0	P7	0.715986	0.060622	0.126862	0.208290	0.066052
1	P8	0.694092	0.221349	0.198859	0.083338	-0.035180
2	P9	0.331681	0.236754	-0.081730	0.755018	-0.042783
3	P10	0.379868	0.094012	0.120168	0.424849	0.157536
4	P11	0.674077	0.211518	0.118783	0.336339	-0.027717
5	P12	0.828738	0.236344	0.109351	0.255582	0.035195
6	P13	0.664437	0.304548	0.146619	0.205842	0.029582
7	P14	0.660412	0.351504	0.121222	0.307017	-0.115409
8	P15	0.594335	0.395083	0.049457	0.375390	0.128259
9	P16	0.694201	0.357890	0.123542	0.166044	0.088315
10	P17	0.682001	0.482308	0.022896	0.163110	0.015906
11	P18	0.249460	0.644266	-0.009080	0.465164	0.017401
12	P19	0.407066	0.780180	0.192697	0.193039	0.060512

```
# Get variance of each factors
```

```
df_var =pd.DataFrame.from_records(fa.get_factor_variance())
```

```
df_var['Parámetros'] =['SS Loadings','Proportion Var','Cumulative Var']
```

```
first_column = df_var.pop('Parámetros')
```

```
df_var.insert(0, 'Parámetros', first_column)
```

```
df_var.columns = ['Parámetros','Factor 1','Factor 2','Factor 3','Factor 4','Factor 5']
```

```
display(df_var)
```

	Parámetros	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
0	SS Loadings	6.652923	2.929958	1.736342	1.724923	1.641506
1	Proportion Var	0.302406	0.133180	0.078925	0.078406	0.074614
2	Cumulative Var	0.302406	0.435585	0.514510	0.592916	0.667530

Se observa que quien contiene mayores propociones en variance, son los componentes 1 y componentes 2.

```
#gráfico de análisis factorial
```

```
fig=plt.figure(figsize=(10,10))
```

```
ax=fig.add_subplot(1,1,1)
```

```
annotations = df_pca['Pregunta']
```

```
ax.set_xlabel("Factor 1",fontsize=15)
```

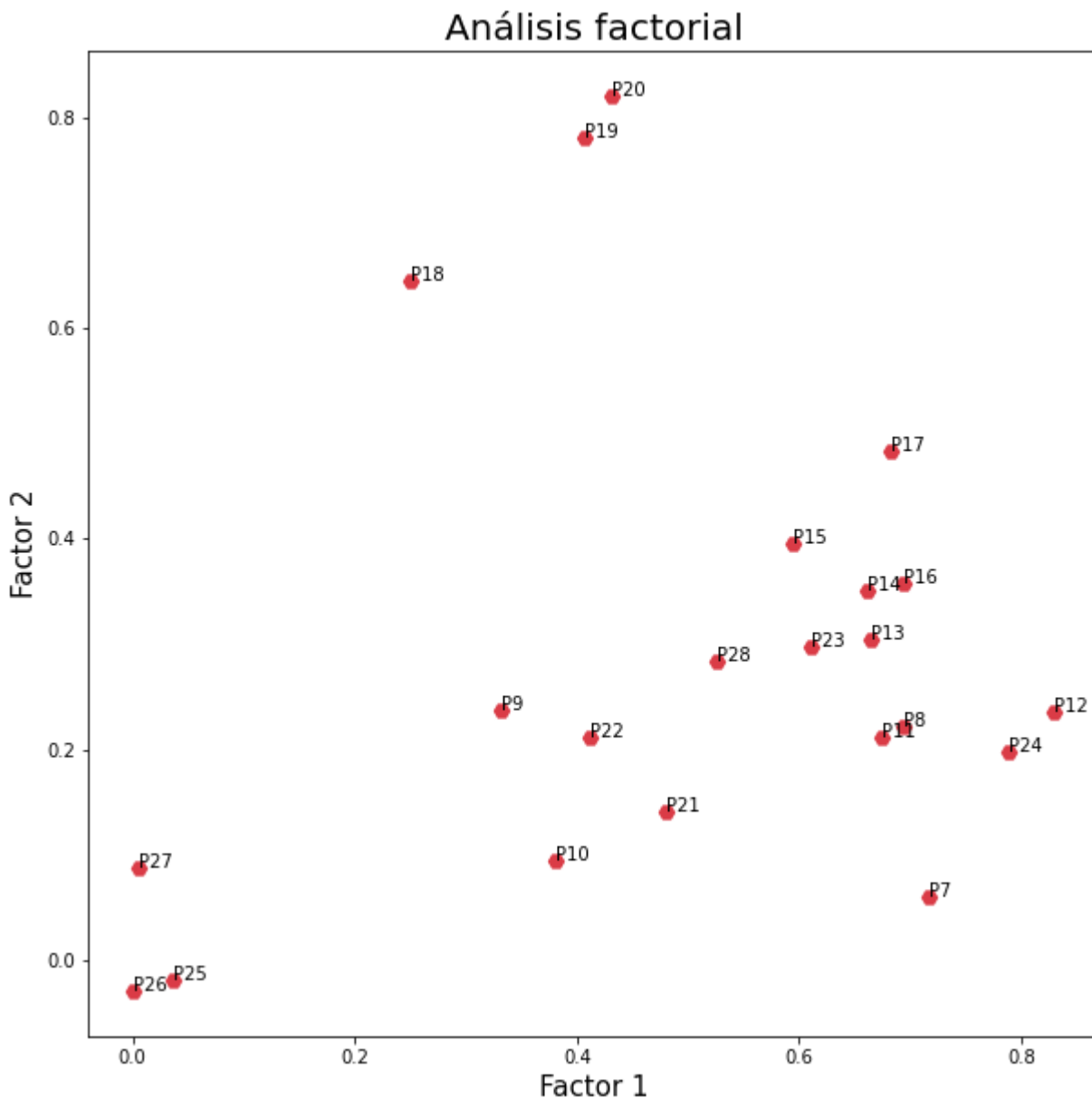
```
ax.set_ylabel("Factor 2", fontsize=15)
```

```
ax.set_title("Análisis factorial", fontsize=20)
```

```

color_theme=np.array(["CYAN", 'GREENYELLOW', 'mediumpurple'])
ax.scatter(x=df_pca['Factor 1'], y=df_pca['Factor 2'], marker = "H" ,s=70)
for i, label in enumerate(annotations):
    plt.annotate(label, (df_pca['Factor 1'][i], df_pca['Factor 2'][i]))
plt.show()

```

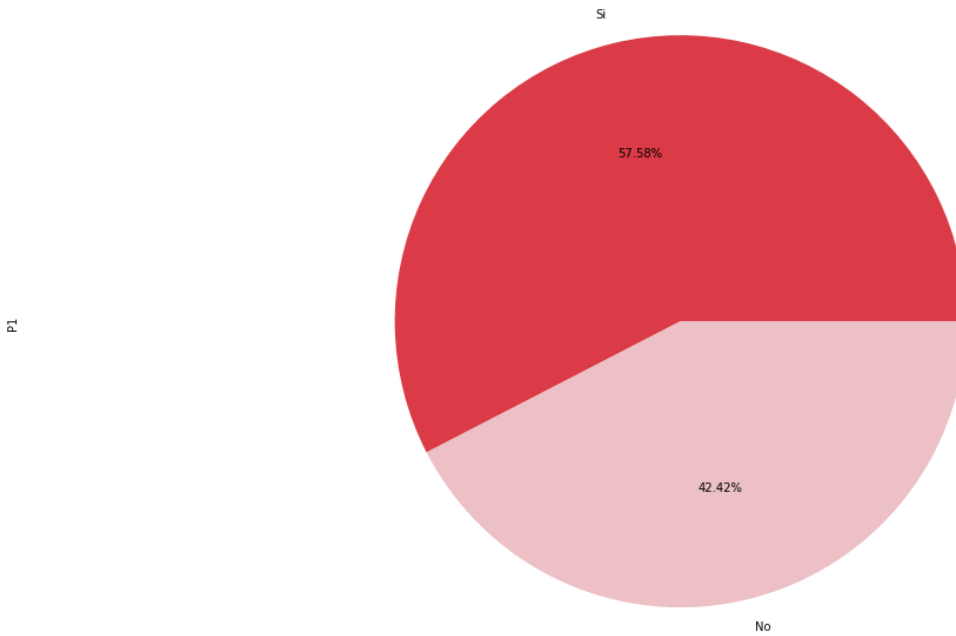


Referencias:

- <https://bit.ly/3PCfZBb>
- <https://bit.ly/3QY5iK7>

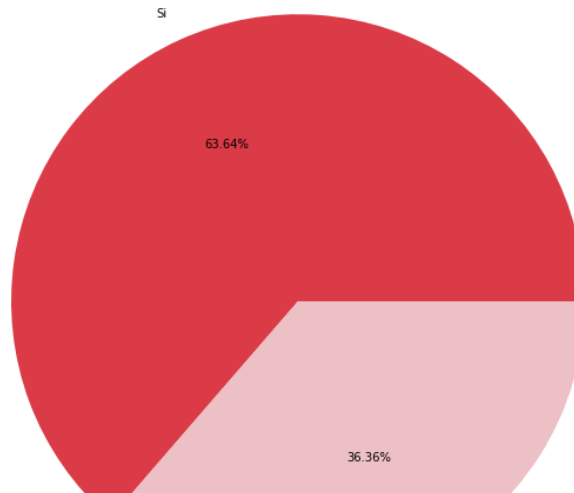
▾ Variables equ_pdg_text_2 tipo Si y No

```
equ_pdg_text_2['P1'].value_counts().plot(kind='pie', autopct='%0.2f%%',figsize=(20,10))  
plt.axis('equal');
```



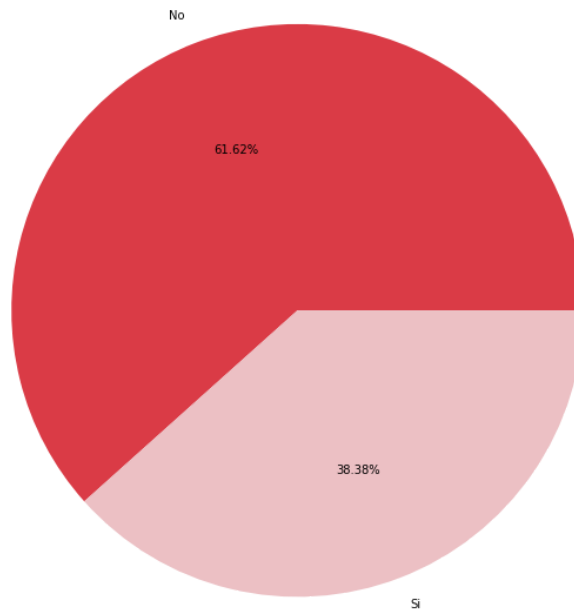
```
equ_pdg_text_2['P3'].value_counts().plot(kind='pie', autopct='%0.2f%%',figsize=(20,10))  
plt.axis('equal');
```

P3



```
equ_pdg_text_2['P5'].value_counts().plot(kind='pie', autopct='%2f%%',figsize=(20,10))  
plt.axis('equal');
```

P5



Reemplazando variables categóricas a cantidades 'No': 1, 'Si':2


```
equ_pdg_text_2.replace({'No': 1, 'Si':2},inplace=True)
```

/usr/local/lib/python3.7/dist-packages/pandas/core/frame.py:5244: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_method=method,



equ_pdg_text_2

	P1	P3	P5
0	2	2	2
1	2	1	1
2	1	2	1
3	2	2	2
4	2	2	2
...	...	...	...
94	1	2	2
95	1	1	1
96	2	2	2
97	2	2	2
98	2	2	1

99 rows × 3 columns

Estadística descriptiva de los datos.

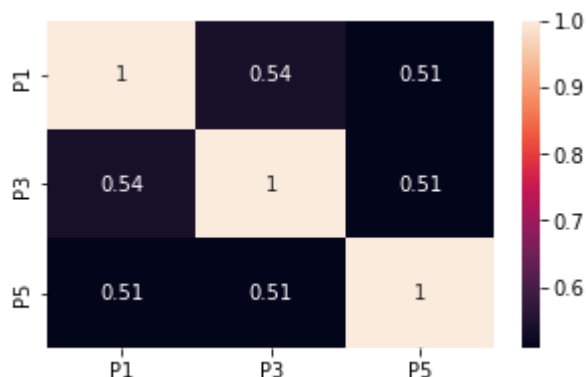
```
equ_pdg_text_2.describe()
```

	P1	P3	P5
P1	00.000000	00.000000	00.000000
P3	00.000000	00.000000	00.000000
P5	00.000000	00.000000	00.000000

Correlación de Pearson.

```
corr_df = equ_pdg_text_2.corr(method='pearson')
```

```
plt.figure(figsize=(5, 3))
sns.heatmap(corr_df, annot=True)
plt.show()
```



Coeficiente de variación

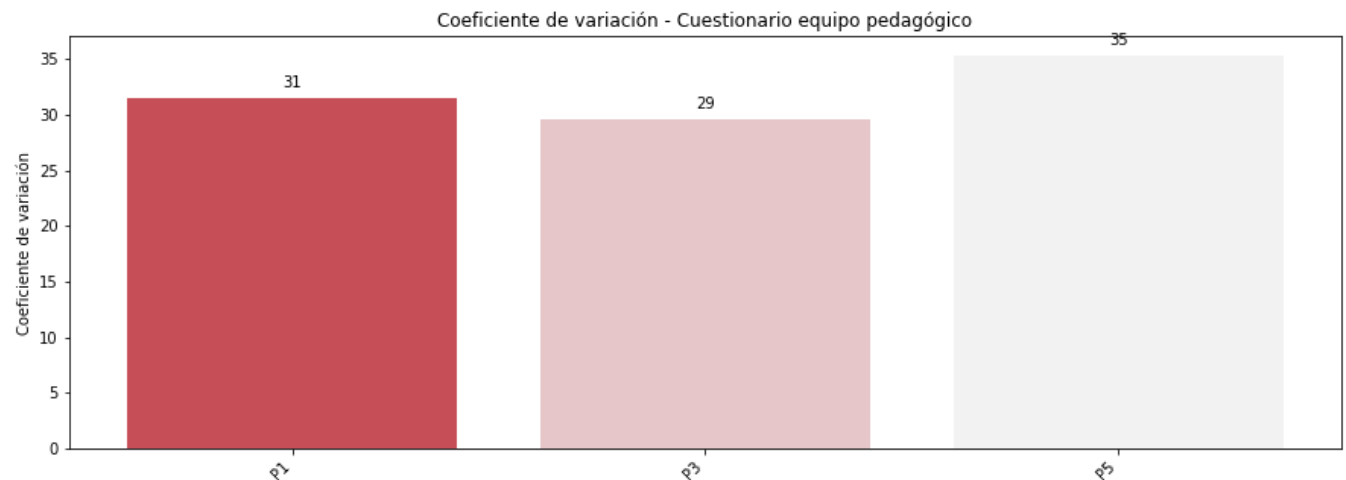
```
cv = lambda x: np.std(x, ddof=1) / np.mean(x) * 100
#calculate CV for each column in data frame
df1 = equ_pdg_text_2.apply(cv)
```

```
#sns.scatterplot(x=list(results.keys() ),y= list(results.values()))
plt.figure(figsize=(15,5))
ax = sns.barplot(x=equ_pdg_text_2.columns,y= df1 )
plt.title(title)
#ax = sns.lineplot(x = list(range(0,len(df))),y= df[target].mean(),label='mean ({:0.2f})'.for
for p in ax.patches:
    _x = p.get_x() + p.get_width() / 2
    _y = p.get_y() + p.get_height()+1
    value = str( int(p.get_height() ) )
    ax.text(_x, _y, value, ha="center")

    #ax.text(_x, _y, value, ha="left")

ax.set_xticklabels(ax.get_xticklabels(), rotation=45, horizontalalignment='right')
plt.ylabel('Coeficiente de variación')
plt.title('Coeficiente de variación - Cuestionario equipo pedagógico')
plt.plot()
```

[]



Características que más difieren de la Respuesta Deseada

T-Score: número de desviaciones estándar con respecto al valor medio de la población.

▼ Análisis Factorial

```
from factor_analyzer import FactorAnalyzer
```

Prueba de esfericidad de Bartlett

comprueba si las variables observadas se interrelacionan o no usando la matriz de correlación observada contra la matriz de identidad. Si la prueba resultó estadísticamente insignificante, no debe emplear un análisis factorial.

```
from factor_analyzer.factor_analyzer import calculate_bartlett_sphericity
chi_square_value, p_value = calculate_bartlett_sphericity(equ_pdg_text_2)
chi_square_value, p_value
```

```
(75.7741639679984, 2.472703577960159e-16)
```

En esta prueba de Bartlett, el valor p es cercano a 0. la prueba fue estadísticamente significativa, indicando que la matriz de correlación observada no es una matriz de identidad.

prueba Kaiser-Meyer-Olkin (KMO)

Mide la idoneidad de los datos para el análisis factorial. Determina la adecuación para cada variable observada y para el modelo completo. KMO estima la proporción de varianza entre todas las variables observadas. Identificación de proporción más baja más adecuada para el análisis factorial. Los valores de KMO oscilan entre 0 y 1. Un valor de KMO inferior a 0,6 se considera inadecuado.

```
from factor_analyzer.factor_analyzer import calculate_kmo
```

```
kmo_all,kmo_model=calculate_kmo(equ_pdg_text_2)
```

```
kmo_model
```

```
0.6925450818923018
```

El KMO general para nuestros datos es 0,7 , lo cual está bien porque no es inferior a 0.6. Este valor indica que puede continuar con su análisis factorial planificado.

Análisis factorial

```
equ_pdg_text_2 = equ_pdg_text_2.dropna()
```

```
fa = FactorAnalyzer(n_factors= 5,rotation = 'varimax')
```

```
fa.fit(equ_pdg_text_2)
```

```
/usr/local/lib/python3.7/dist-packages/factor_analyzer/factor_analyzer.py:356: Deprecati
error = sp.sum(residual)
/usr/local/lib/python3.7/dist-packages/factor_analyzer/factor_analyzer.py:356: Deprecati
error = sp.sum(residual)
/usr/local/lib/python3.7/dist-packages/factor_analyzer/factor_analyzer.py:356: Deprecati
error = sp.sum(residual)
/usr/local/lib/python3.7/dist-packages/factor_analyzer/factor_analyzer.py:356: Deprecati
error = sp.sum(residual)
/usr/local/lib/python3.7/dist-packages/factor_analyzer/factor_analyzer.py:356: Deprecati
error = sp.sum(residual)
/usr/local/lib/python3.7/dist-packages/factor_analyzer/factor_analyzer.py:356: Deprecati
error = sp.sum(residual)
/usr/local/lib/python3.7/dist-packages/factor_analyzer/factor_analyzer.py:356: Deprecati
error = sp.sum(residual)
/usr/local/lib/python3.7/dist-packages/factor_analyzer/factor_analyzer.py:356: Deprecati
error = sp.sum(residual)
FactorAnalyzer(n_factors=5, rotation='varimax', rotation_kwargs={})
```

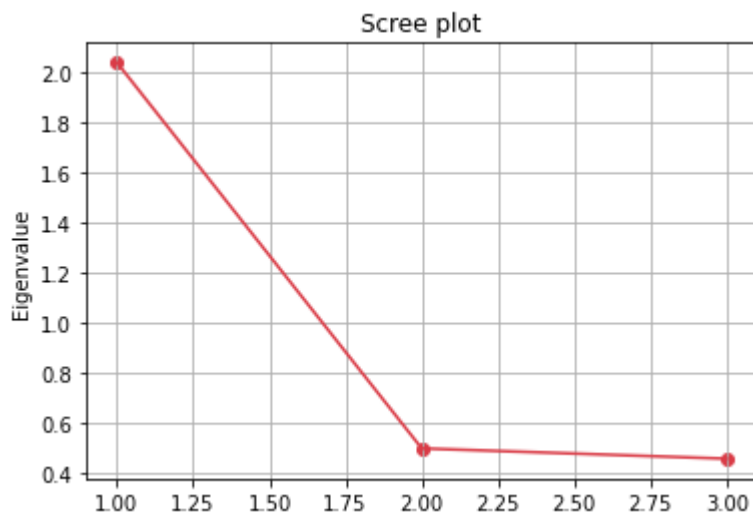
```
#Get the loadings
```

```
loadings = fa.loadings_
```

```
# Eigen vectores
ev, v = fa.get_eigenvalues()
```

```
xvals = range(1, equ_pdg_text_2.shape[1]+1)
```

```
plt.scatter(xvals, ev)
plt.plot(xvals, ev)
plt.title('Scree plot')
plt.ylabel('Eigenvalue')
plt.grid()
plt.show()
```



```
# componente principales
df_pca = pd.DataFrame.from_records(loadings)
df_pca['Pregunta'] = equ_pdg_text_2.columns
first_column = df_pca.pop('Pregunta')
df_pca.insert(0, 'Pregunta', first_column)
df_pca.columns = ['Pregunta', 'Factor 1', 'Factor 2', 'Factor 3']
display(df_pca)
```

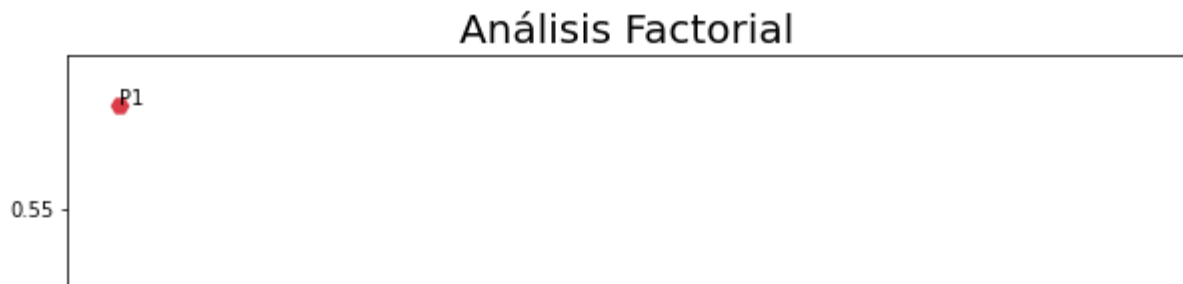
	Pregunta	Factor 1	Factor 2	Factor 3
0	P1	0.359041	0.576952	0.355995
1	P3	0.360216	0.357442	0.576354
2	P5	0.582743	0.321894	0.321516

```
# Get variance of each factors
df_var = pd.DataFrame.from_records(fa.get_factor_variance())
df_var['Parámetros'] = ['SS Loadings', 'Proportion Var', 'Cumulative Var']
first_column = df_var.pop('Parámetros')
```

```
df_var.insert(0, 'Parámetros', first_column)
df_var.columns = ['Parámetros', 'Factor 1', 'Factor 2', 'Factor 3']
display(df_var)
```

	Parámetros	Factor 1	Factor 2	Factor 3
0	SS Loadings	0.598255	0.564254	0.562289
1	Proportion Var	0.199418	0.188085	0.187430
2	Cumulative Var	0.199418	0.387503	0.574933

```
#gráfico de componente principales- plano factorial
fig=plt.figure(figsize=(10,10))
ax=fig.add_subplot(1,1,1)
annotations = df_pca['Pregunta']
ax.set_xlabel("Factor 1",fontsize=15)
ax.set_ylabel("Factor 2", fontsize=15)
ax.set_title("Análisis Factorial", fontsize=20)
color_theme=np.array(["CYAN",'GREENYELLOW','mediumpurple'])
ax.scatter(x=df_pca['Factor 1'], y=df_pca['Factor 2'],marker = "H" ,s=70)
for i, label in enumerate(annotations):
    plt.annotate(label, (df_pca['Factor 1'][i], df_pca['Factor 2'][i]))
plt.show()
```



Se muestra la independencia de cada pregunta

▼ NUBE DE PALABRAS

Argumento pregunta P1: ¿Se generan espacios de intercambio de experiencias investigativas - formativas que puedan convertirse en proyectos de investigación SENNOVA?

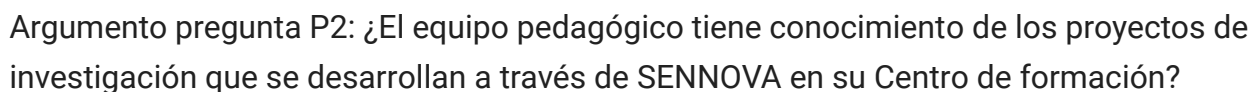
```
stopwords = set(stopwords.words('spanish')) # stopwords de nltk en español
```

```
#Lista de Stopword:
```

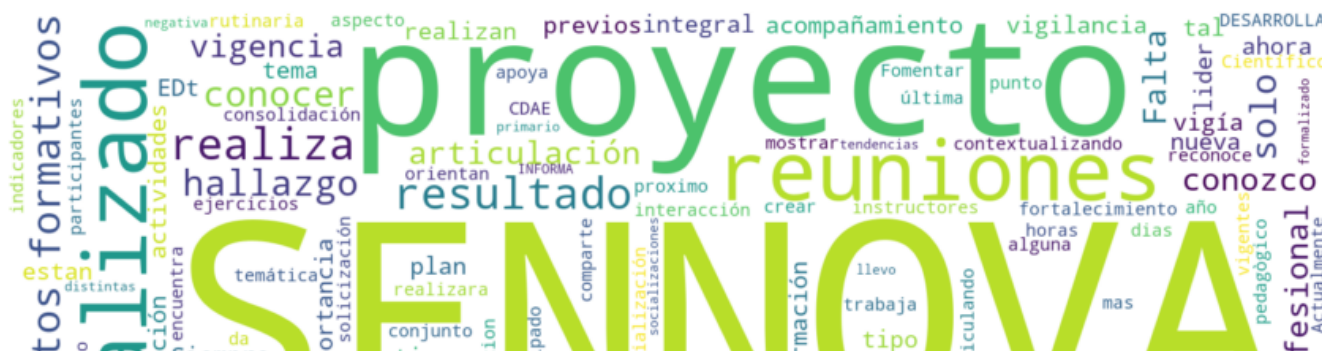
```
stopwords.update(["|", "-", "NaN"]) # adicionales que no están en stopwords
```

```
# Imagen de nube de palabras
```

```
text = ' '.join(equ_pdg_text_1['P2'].astype('str'))
wordcloud = WordCloud(width = 1500, height = 1000, stopwords=stopwords, max_font_size=300, ma
plt.figure(figsize = (15,15))
plt.imshow(wordcloud, interpolation="bilinear")
plt.axis("off")
plt.show()
```



```
text = ' '.join(equ_pdg_text_1['P4'].astype('str'))
wordcloud = WordCloud(width = 1500, height = 1000, stopwords=stopwords, max_font_size=300, ma
plt.figure(figsize = (15,15))
plt.imshow(wordcloud, interpolation="bilinear")
plt.axis("off")
plt.show()
```

Preprocesamiento de Texto Preprocesamiento del texto:

Tokenizar, Limpieza, Stopwords, Lematizar

```
equ_pdg_text_1=equ_pdg_text_1.astype(str)
equ_pdg_text_1.replace({'nan':np.nan,'Na':np.nan},inplace=True)
equ_pdg_text_1 = equ_pdg_text_1.dropna()
equ_pdg_text_1
```

	P2	P4	P6
0	Sugiere que los proyectos formativos, tenga un...	Los proyectos SENNOVA son presentado al equipo...	Se espera que en la solicialización oriente a ...
1	Como estrategia por parte del EPC de centro, s...	Sennova maneja diversas líneas y proyectos que...	Falta mas interacción con el EPC para establec...
2	pocas veces	En la ultima reunion del EPC se explicaron los...	La vigilancia tecnologica estan contenidas en ...
3	Desde el ecosistema SENNOVA se establecen espa...	Durante las jornada de inducción y reinducción...	El grupo SENNOVA socializa los ejercicios de v...
6	En el centro el estilo de trabajo en la defini...	Todos aportan a la definición de los proyectos...	se realizan EDT de forma rutinaria para mostra...
...	...	...	...
93	Se identifican oportunidades de mejora en l ej...	En diferentes reuniones el líder sennova Juliá...	En reuniones de seguimiento a las actividades ...
94	Son Técnico y el tiempo no da para realizar in...	Hay personas que están gestionando estos procesos	Siempre hay acompañamiento del equipo pedagógico

stopwords es

```
['de',
 'la',
 'que',
 'el',
 'en',
```

```
'y',  
'a',  
'los',  
'del',  
'se',  
'las',  
'por',  
'un',  
'para',  
'con',  
'no',  
'una',  
'su',  
'al',  
'lo',  
'como',  
'más',  
'pero',  
'sus',  
'le',  
'ya',  
'o',  
'este',  
'sí',  
'porque',  
'esta',  
'entre',  
'cuando',  
'muy',  
'sin',  
'sobre',  
'también',  
'me',  
'hasta',  
'hay',  
'donde',  
'quien',  
'desde',  
'todo',  
'nos',  
'durante',  
'todos',  
'uno',  
'les',  
'ni',  
'contra',  
'otros',  
'ese',  
'eso',  
'ante',  
'ellos',  
'e',  
'esto'.
```

Esta función toma una cadena, reduce todo, filtra solo letras, elimina palabras vacías, lematiza y

```
def limpia_texto(texto):
    '''(str) -> str
    Esta función toma una cadena, reduce todo, filtra solo letras, elimina palabras vacías, lem
    ...
    texto = texto.lower()

    texto = re.sub(r"[\W\d_]+", " ", texto)

    texto = [pal for pal in texto.split() if pal not in stopwords_es]

    spc_texto = spc_es(" ".join(texto))
    tokens = [word.lemma_ if word.lemma_ != "-PRON-" else word.lower_ for word in spc_texto]

    return " ".join(tokens)

#preprocesamiento de texto filtra solo letras, elimina palabras vacías, lematiza y devuelve l
equ_pdg_text_1['P2'] = equ_pdg_text_1['P2'].apply(limpia_texto)
equ_pdg_text_1['P4'] = equ_pdg_text_1['P4'].apply(limpia_texto)
equ_pdg_text_1['P6'] = equ_pdg_text_1['P6'].apply(limpia_texto)
#Ejemplo con P6
equ_pdg_text_1.dropna()
```

	P2	P4	P6
0	sugerir proyecto formativo componente investig...	proyecto sennova presentado equipo pedagógico ...	esperar solicialización oriente proyección pro...
1	estrategia parte epc centro invitar cada reuni...	sennova manejar diverso línea proyecto importa...	faltar mas interacción epc establecer propuest...
2	poco vez	ultimo reunion epc explicar proyecto investiga...	vigilancia tecnologico estar contenido documen...
3	ecosistema sennova establecer espacio apropiac...	jornada inducción reinducción así espacio apro...	grupo sennova socializar ejercicio vigilancia ...
6	centro estilo trabajo definición acción colegiado	aportar definición proyecto vigencia anterior	realizar edt forma rutinario mostrar avance re...
...	...	...	...
93	identificar oportunidad mejor l ejecución form...	diferente reunión líder sennova julián cely co...	reunión seguimientir actividad programado epc
94	técnico tiempo dar realizar investigación dent...	persona gestionar proceso	siempre acompañamiento equipo pedagógico

▼ Matriz termino-documento - Document-Term Matrix

Notaremos como "Com" los comentarios por cada usuario

1. ¿Se generan espacios de intercambio de experiencias

▼ investigativas - formativas que puedan convertirse en proyectos de investigación SENNOVA?

```
# Count Vectorizer
vect = CountVectorizer()
vects = vect.fit_transform(equ_pdg_text_1['P2'])
# Seleccione las primeras cinco filas del conjunto de datos
td = pd.DataFrame(vects.todense()).iloc[:5]
td.columns = vect.get_feature_names()
term_document_matrix = td.T
term_document_matrix.columns = ['Com '+str(i) for i in range(1, 6)]
term_document_matrix['total_count'] = term_document_matrix.sum(axis=1)

# Top 25 palabras
term_document_matrix = term_document_matrix.sort_values(by='total_count', ascending=False)[:1]

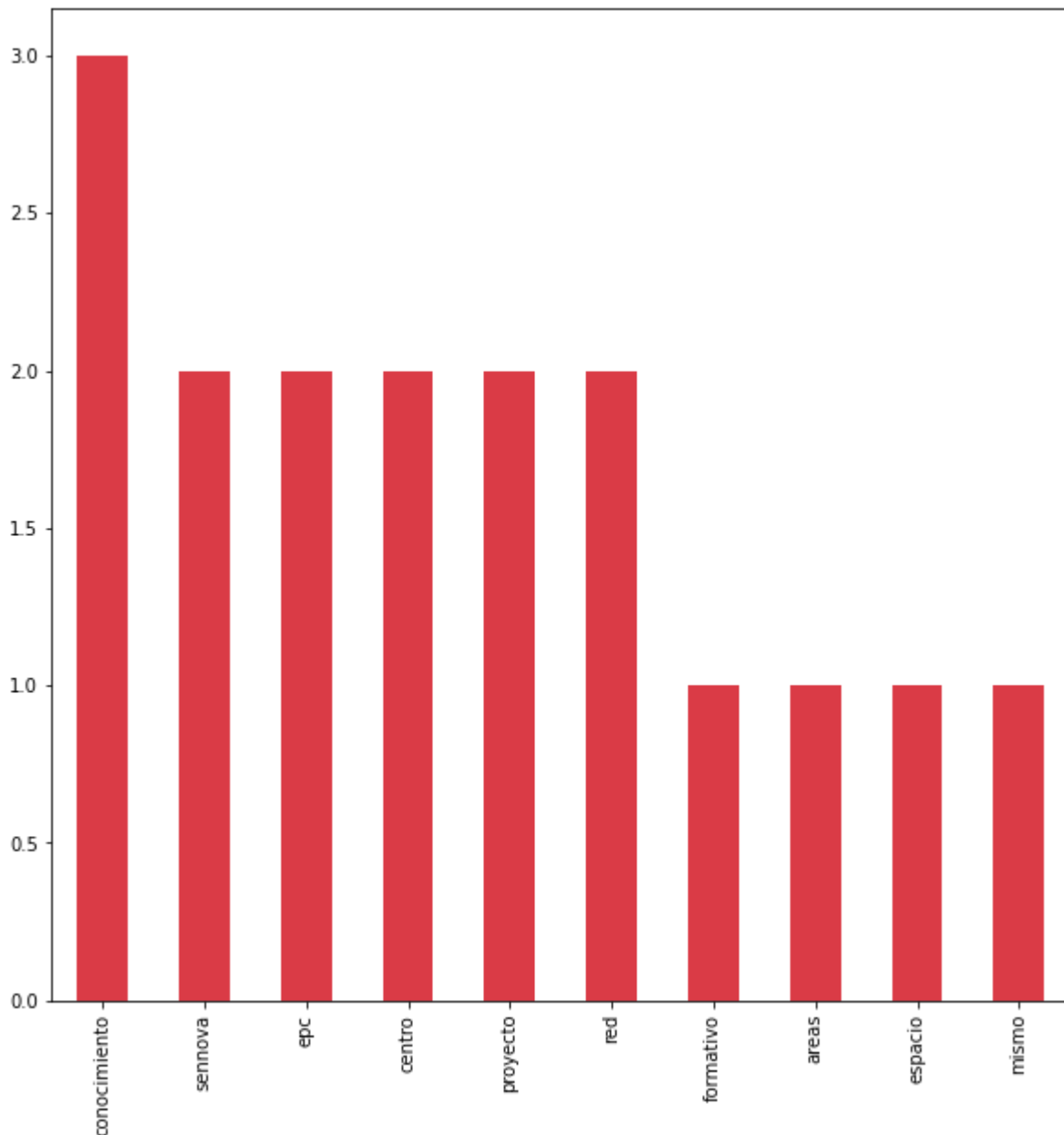
# Imprime las primeras 10 filas
print(term_document_matrix.drop(columns=['total_count']).head(20))
```

	Com 1	Com 2	Com 3	Com 4	Com 5
conocimiento	0	1	0	2	0
sennova	0	1	0	1	0
epc	0	2	0	0	0
centro	0	1	0	0	1
proyecto	1	0	0	1	0
red	0	1	0	1	0
formativo	1	0	0	0	0
areas	0	1	0	0	0
espacio	0	0	0	1	0
mismo	0	1	0	0	0

```
/usr/local/lib/python3.7/dist-packages/sklearn/utils/deprecation.py:87: FutureWarning: f
warnings.warn(msg, category=FutureWarning)
```

▼ Histograma de palabras

```
plt.figure(figsize=(10,10))
term_document_matrix['total_count'].plot.bar()
plt.show()
```



▼ Semejanza de coseno

```
# Define la función de similitud del coseno.
def cosine_similarity(a,b):
    "calcula la similitud del coseno de v1 a v2: (a punto b)/{||a||*||b||}"
    sumxx, sumxy, sumyy = 0, 0, 0
    for i in range(len(a)):
        x = a[i]; y = b[i]
        sumxx += x*x
        sumyy += y*y
        sumxy += x*y
    return sumxy/math.sqrt(sumxx*sumyy)
```

▼ Latent Dirichlet Allocation - Modelo no supervisado de Tópicos.

```
from sklearn.decomposition import LatentDirichletAllocation
```

```
vectorizer_ = CountVectorizer(min_df=5, max_df=0.9, stop_words= stopwords_es, lowercase=True,
com_vectorized = vectorizer_.fit_transform(equ_pdg_text_1['P2'])
```

```
<>:1: DeprecationWarning: invalid escape sequence \-
<>:1: DeprecationWarning: invalid escape sequence \-
<>:1: DeprecationWarning: invalid escape sequence \-
<ipython-input-286-69e88de69145>:1: DeprecationWarning: invalid escape sequence \-
    vectorizer_ = CountVectorizer(min_df=5, max_df=0.9, stop_words= stopwords_es, lowercase=
/usr/local/lib/python3.7/dist-packages/sklearn/feature_extraction/text.py:401: UserWarning:
    % sorted(inconsistent)
```

```
lda_ = LatentDirichletAllocation(n_components=10, max_iter=5, learning_method='online', verbose=1)
tweets_lda = lda_.fit_transform(com_vectorized)
```

```
iteration: 1 of max_iter: 5
iteration: 2 of max_iter: 5
iteration: 3 of max_iter: 5
iteration: 4 of max_iter: 5
iteration: 5 of max_iter: 5
```

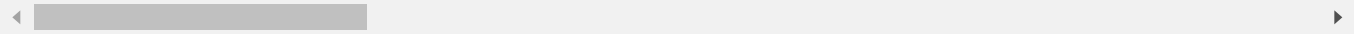
▼ keywords

```
def selected_topics(model, vectorizer, top_n=10):
    for idx, topic in enumerate(model.components_):
        print("Topic %d:" % (idx))
        print([(vectorizer.get_feature_names()[i], topic[i])
                for i in topic.argsort()[: -top_n - 1: -1]])
```

```
print("Modelo de comentarios LDA:")
selected_topics(lda_, vectorizer_)
```

```
Modelo de comentarios LDA:
Topic 0:
[('proyecto', 4.281542046507682), ('investigaci', 4.176824773262229), ('socializar', 2.811111111111111), ('formaci', 2.7777777777777777), ('realizar', 2.7777777777777777), ('intercambio', 2.7777777777777777)]
Topic 1:
[('intercambio', 1.6907491113671007), ('realizar', 1.078373927578417), ('formativo', 0.7407407407407407), ('sennova', 0.7407407407407407), ('equipo', 0.7407407407407407)]
Topic 2:
[('sennova', 5.234995463982507), ('equipo', 4.391664107101011), ('experiencia', 3.3216875000000002), ('proyecto', 3.3216875000000002), ('realizar', 3.3216875000000002)]
Topic 3:
[('espacio', 11.182147772810369), ('proyecto', 7.8817644923785855), ('investigaci', 6.911111111111111), ('formaci', 6.911111111111111), ('realizar', 6.911111111111111)]
Topic 4:
[('realizar', 5.598468878278119), ('formaci', 3.8652555322385735), ('investigaci', 3.5911785860954826), ('sennova', 3.5911785860954826), ('proyecto', 3.5911785860954826)]
Topic 5:
[('proyecto', 7.842641160260371), ('formativo', 3.5911785860954826), ('sennova', 3.5137500000000002), ('equipo', 3.5137500000000002), ('realizar', 3.5137500000000002)]
Topic 6:
```

```
[('gico', 0.690233674222807), ('pedag', 0.6329275105269168), ('socializar', 0.5909752223  
Topic 7:  
[('sennova', 3.513224179029562), ('actividad', 1.6859049028516333), ('poder', 1.68405614  
Topic 8:  
[('generar', 7.393730756846865), ('espacio', 4.758994075295806), ('experiencia', 2.25319  
Topic 9:  
[('gico', 6.631406078189716), ('pedag', 6.0696198941961335), ('equipo', 4.22360092134788  
/usr/local/lib/python3.7/dist-packages/sklearn/utils/deprecation.py:87: FutureWarning: F  
warnings.warn(msg, category=FutureWarning)
```



```
pyLDAvis.enable_notebook()  
dash = pyLDAvis.sklearn.prepare(lda_, com_vectorized, vectorizer_, mds='tsne')  
dash
```



```

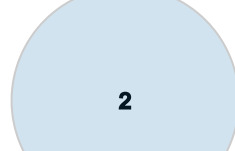
/usr/local/lib/python3.7/dist-packages/pyLDAvis/_prepare.py:247: FutureWarning: In a fut
by='saliency', ascending=False).head(R).drop('saliency', 1)
/usr/local/lib/python3.7/dist-packages/sklearn/manifold/_t_sne.py:783: FutureWarning: Th
FutureWarning,
/usr/local/lib/python3.7/dist-packages/sklearn/manifold/_t_sne.py:793: FutureWarning: Th
FutureWarning,
/usr/local/lib/python3.7/dist-packages/sklearn/manifold/_t_sne.py:827: FutureWarning: 's
FutureWarning,

```

Selected Topic:

Slid

Intertopic Distance Map (via multidimensional scaling)



0

generar

espacio

2. ¿El equipo pedagógico tiene conocimiento de los proyectos de investigación que se desarrollan a través de SENNOVA en su Centro de formación?



socializar

```

# Count Vectorizer
vect = CountVectorizer()
vects = vect.fit_transform(equ_pdg_text_1['P4'])
# Seleccione las primeras cinco filas del conjunto de datos
td = pd.DataFrame(vects.todense()).iloc[:5]
td.columns = vect.get_feature_names()
term_document_matrix = td.T
term_document_matrix.columns = ['Com ' + str(i) for i in range(1, 6)]
term_document_matrix['total_count'] = term_document_matrix.sum(axis=1)

# Top 25 palabras
term_document_matrix = term_document_matrix.sort_values(by = 'total_count', ascending=False)[:1]

# Imprime las primeras 10 filas
print(term_document_matrix.drop(columns=['total_count']).head(20))

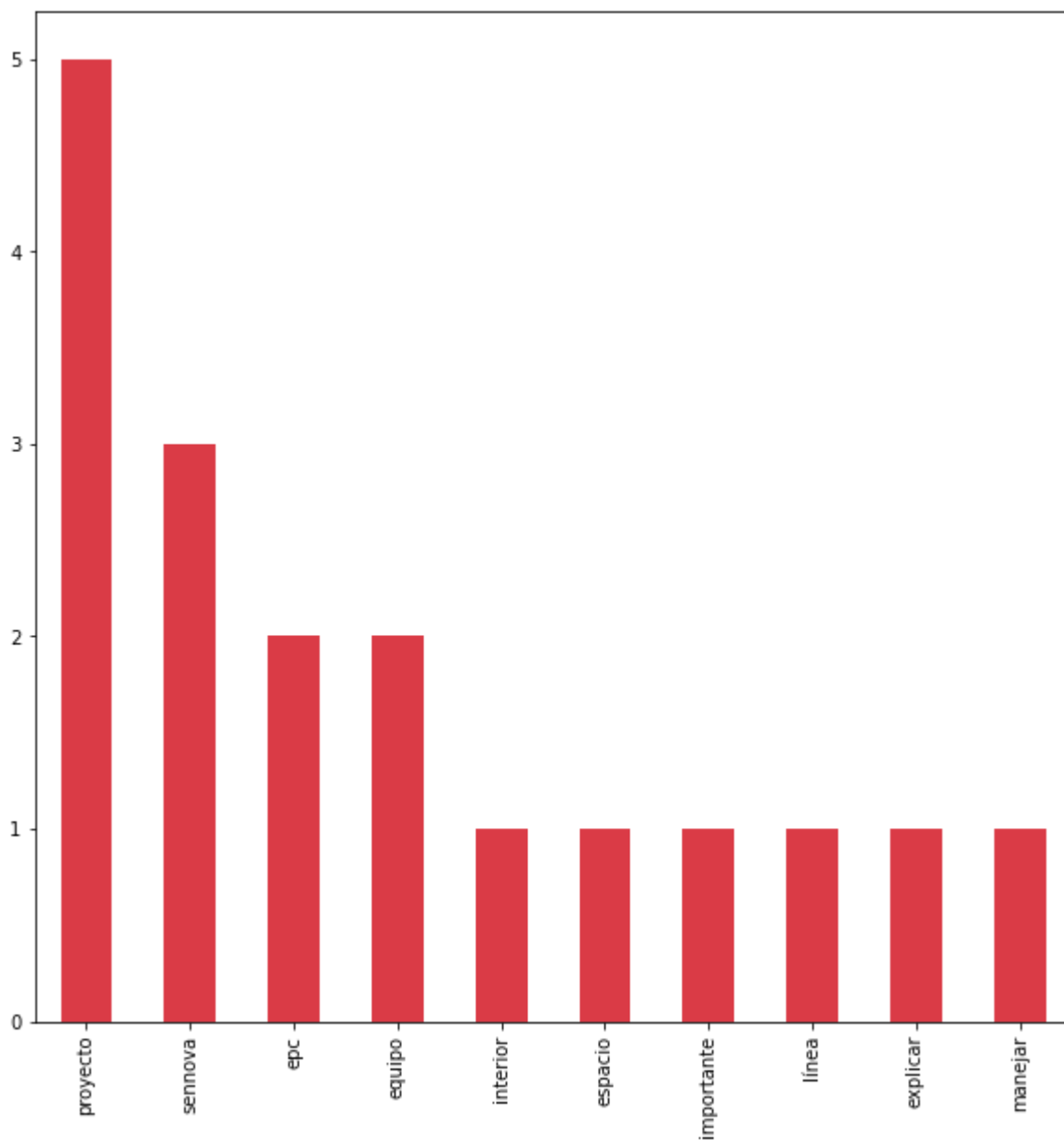
```

	Com 1	Com 2	Com 3	Com 4	Com 5
proyecto	1	1	1	1	1
sennova	2	1	0	0	0
epc	0	1	1	0	0
equipo	2	0	0	0	0
interior	0	1	0	0	0
espacio	0	0	0	1	0
importante	0	1	0	0	0
línea	0	1	0	0	0
explicar	0	0	1	0	0
manejar	0	1	0	0	0

```
/usr/local/lib/python3.7/dist-packages/sklearn/utils/deprecation.py:87: FutureWarning: F  
warnings.warn(msg, category=FutureWarning)
```

▼ Histograma de palabras

```
plt.figure(figsize=(10,10))  
term_document_matrix['total_count'].plot.bar()  
plt.show()
```



▼ Semejanza de coseno

```

# Define la función de similitud del coseno.
def cosine_similarity(a,b):
    "calcula la similitud del coseno de v1 a v2: (a punto b)/{||a||*||b||}"
    sumxx, sumxy, sumyy = 0, 0, 0
    for i in range(len(a)):
        x = a[i]; y = b[i]
        sumxx += x*x
        sumyy += y*y
        sumxy += x*y
    return sumxy/math.sqrt(sumxx*sumyy)

# Guardar el índice de similitud entre los comentarios
def pair(s):
    for i, v1 in enumerate(s):
        for j in range(i+1, len(s)):
            yield [v1, s[j]]

dic={}
for (a,b) in list(pair(['Com 1', 'Com 2', 'Com 3', 'Com 4', 'Com 5'])):
    dic[(a,b)] = cosine_similarity(term_document_matrix[a].tolist(), term_document_matrix[b].to

{('Com 1', 'Com 2'): 0.3779644730092272, ('Com 1', 'Com 3'): 0.19245008972987526, ('Com

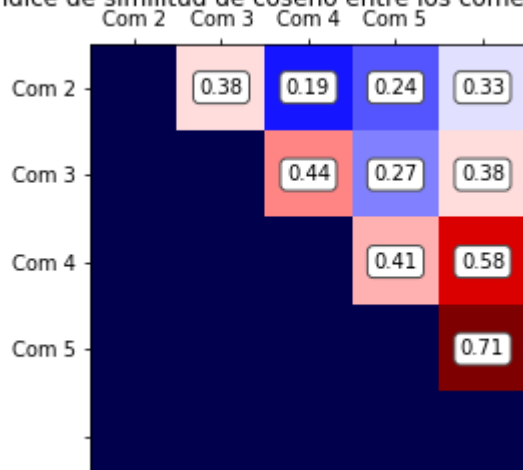
documents= ['Com 1', 'Com 2', 'Com 3', 'Com 4', 'Com 5']
final_df = pd.DataFrame(np.asarray([[dic[(x,y)] if (x,y) in dic else 0) for y in documents]
final_df.columns = documents
final_df.index = documents

fig, ax = plt.subplots()
ax.set_xticks(np.arange(len(documents)))
ax.set_yticks(np.arange(len(documents)))
ax.set_xticklabels(documents)
ax.set_yticklabels(documents)
ax.matshow(final_df, cmap='seismic')
for (i, j), z in np.ndenumerate(final_df):
    if z != 0 :
        ax.text(j, i, '{:0.2f}'.format(z), ha='center', va='center',
                bbox=dict(boxstyle='round', facecolor='white', edgecolor='0.3'))
    else:
        None

fig.suptitle('Índice de similitud de coseno entre los comentarios')
plt.show()

```

Índice de similitud de coseno entre los comentarios



▼ Latent Dirichlet Allocation - Modelo no supervisado de Tópicos.

```
from sklearn.decomposition import LatentDirichletAllocation
```

```
vectorizer_ = CountVectorizer(min_df=5, max_df=0.9, stop_words= stopwords_es, lowercase=True,
com_vectorized = vectorizer_.fit_transform(equ_pdg_text_1['P4'])
```

```
<>:1: DeprecationWarning: invalid escape sequence \-
<>:1: DeprecationWarning: invalid escape sequence \-
<>:1: DeprecationWarning: invalid escape sequence \-
<ipython-input-297-edf9d7f36983>:1: DeprecationWarning: invalid escape sequence \-
vectorizer_ = CountVectorizer(min_df=5, max_df=0.9, stop_words= stopwords_es, lowercas
/usr/local/lib/python3.7/dist-packages/sklearn/feature_extraction/text.py:401: UserWarni
% sorted(inconsistent)
```

```
lda_ = LatentDirichletAllocation(n_components=10, max_iter=5, learning_method='online', verbos
tweets_lda = lda_.fit_transform(com_vectorized)
```

```
iteration: 1 of max_iter: 5
iteration: 2 of max_iter: 5
iteration: 3 of max_iter: 5
iteration: 4 of max_iter: 5
iteration: 5 of max_iter: 5
```

▼ keywords

```
def selected_topics(model, vectorizer, top_n=10):
    for idx, topic in enumerate(model.components_):
        print("Topic %d:" % (idx))
        print([(vectorizer.get_feature_names()[i], topic[i])
                for i in topic.argsort()[: -top_n - 1: -1]])
```

```

print("Modelo de comentarios LDA:")
selected_topics(lda_, vectorizer_)

Modelo de comentarios LDA:
Topic 0:
[('sennova', 12.884433532042241), ('equipo', 12.235745234944316), ('proyecto', 11.053488
Topic 1:
[('presentar', 1.4258470871057054), ('proyecto', 1.33332143437624), ('der', 1.251938238
Topic 2:
[('proyecto', 5.398781075237268), ('investigaci', 4.796713148397315), ('epc', 2.7974412
Topic 3:
[('espacio', 2.333218564944151), ('der', 1.1747753304893107), ('reuni', 1.0929819156509
Topic 4:
[('equipo', 4.440479600679263), ('pedag', 3.454238434565307), ('gico', 3.43975835051001
Topic 5:
[('epc', 7.7900519810301585), ('proyecto', 4.166795920623417), ('sennova', 3.4644754112
Topic 6:
[('sennova', 2.0898918335557672), ('proyecto', 2.047454106686373), ('socializaci', 1.86
Topic 7:
[('conocimiento', 2.7761447072519707), ('reuni', 1.5919509752116736), ('pedag', 0.64462
Topic 8:
[('conocer', 1.8945764475869575), ('epc', 1.327394708441441), ('reuni', 0.9964358783945
Topic 9:
[('proyecto', 9.114136492152214), ('sennova', 3.964318825950709), ('conocer', 3.5153423
/usr/local/lib/python3.7/dist-packages/sklearn/utils/deprecation.py:87: FutureWarning: F
    warnings.warn(msg, category=FutureWarning)

```

```

pyLDAvis.enable_notebook()
dash = pyLDAvis.sklearn.prepare(lda_, com_vectorized, vectorizer_, mds='tsne')
dash

```

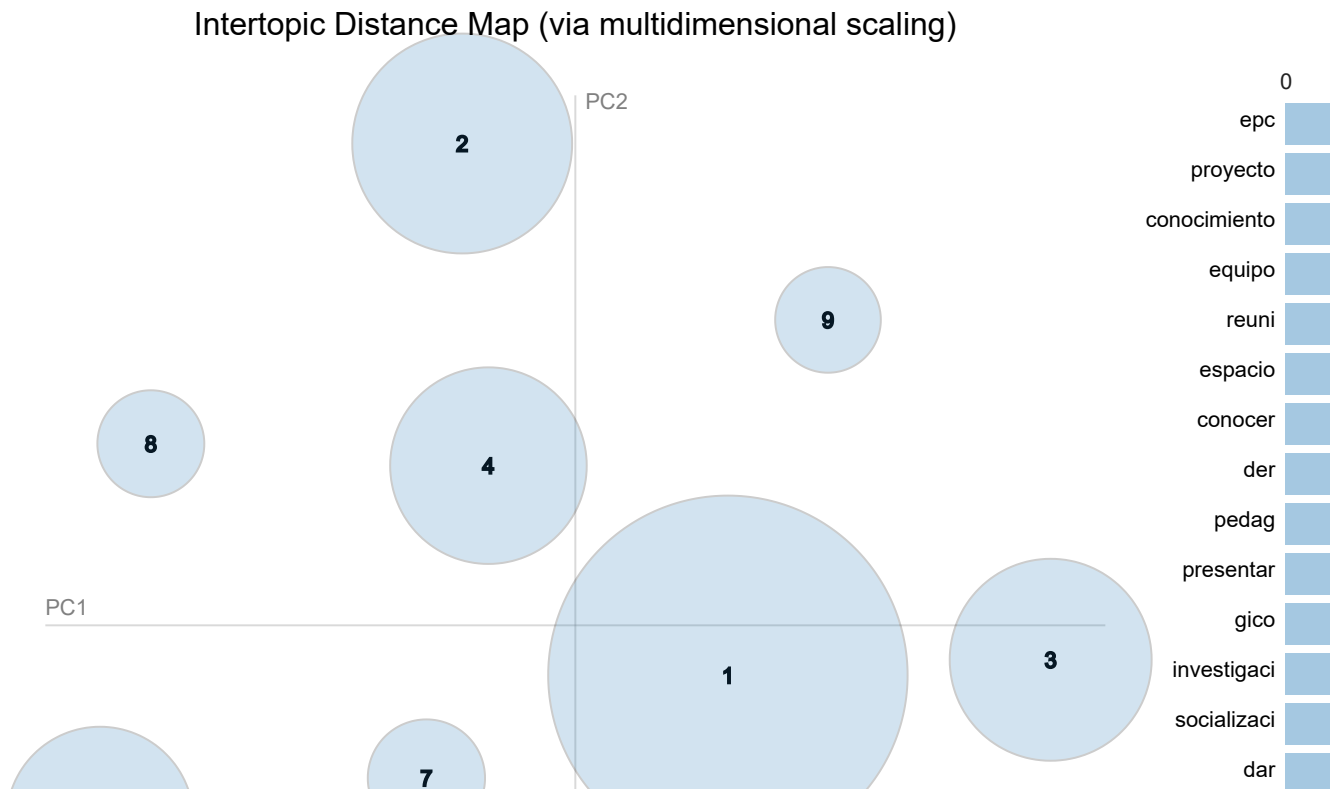
```

/usr/local/lib/python3.7/dist-packages/pyLDAvis/_prepare.py:247: FutureWarning: In a fut
by='saliency', ascending=False).head(R).drop('saliency', 1)
/usr/local/lib/python3.7/dist-packages/sklearn/manifold/_t_sne.py:783: FutureWarning: Th
FutureWarning,
/usr/local/lib/python3.7/dist-packages/sklearn/manifold/_t_sne.py:793: FutureWarning: Th
FutureWarning,
/usr/local/lib/python3.7/dist-packages/sklearn/manifold/_t_sne.py:827: FutureWarning: 's
FutureWarning,

```

Selected Topic:

Slid



3. ¿El equipo SENNOVA socializa con el equipo pedagógico los

- ▼ hallazgos de la revisión o vigilancia tecnológica del contexto, para incorporar los resultados en el proyecto formativo?



socializar

```

# Count Vectorizer
vect = CountVectorizer()
vects = vect.fit_transform(equ_pdg_text_1['P6'])
# Seleccione las primeras cinco filas del conjunto de datos
td = pd.DataFrame(vects.todense()).iloc[:5]
td.columns = vect.get_feature_names()
term_document_matrix = td.T
term_document_matrix.columns = ['Com '+str(i) for i in range(1, 6)]
term_document_matrix['total_count'] = term_document_matrix.sum(axis=1)

```

```
# Top 25 palabras
```

```
term_document_matrix = term_document_matrix.sort_values(by = 'total_count', ascending=False)[:1
```

```
# Imprime las primeras 10 filas
```

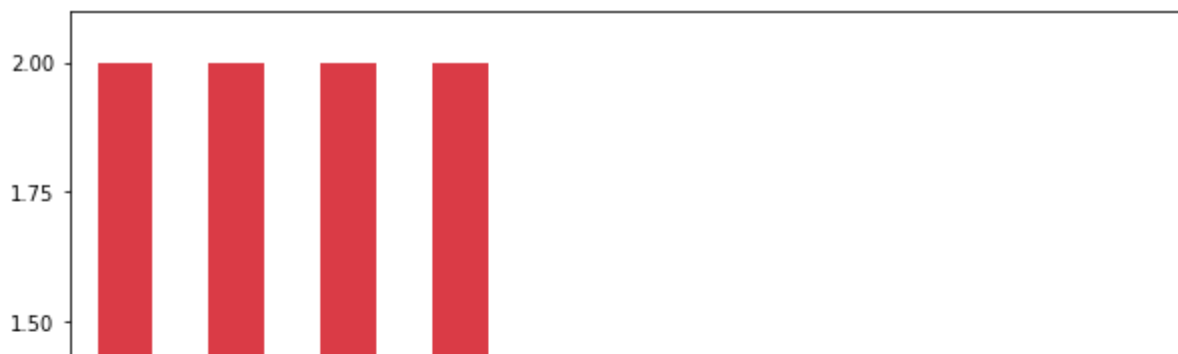
```
print(term_document_matrix.drop(columns=[ 'total_count' ]).head(20))
```

	Com 1	Com 2	Com 3	Com 4	Com 5
tecnológico	0	0	0	2	0
socializar	0	0	1	1	0
vigilancia	0	0	1	1	0
documento	0	0	2	0	0
rutinario	0	0	0	0	1
contenido	0	0	1	0	0
sennova	0	0	0	1	0
formación	0	1	0	0	0
redser	1	0	0	0	0
tecnologico	0	0	1	0	0

```
/usr/local/lib/python3.7/dist-packages/sklearn/utils/deprecation.py:87: FutureWarning: F
warnings.warn(msg, category=FutureWarning)
```

▼ Histograma de palabras

```
plt.figure(figsize=(10,10))
term_document_matrix['total_count'].plot.bar()
plt.show()
```



▼ Semejanza de coseno

Define la función de similitud del coseno.

```
def cosine_similarity(a,b):
    "calcula la similitud del coseno de v1 a v2: (a punto b)/{||a||*||b||}"
    sumxx, sumxy, sumyy = 0, 0, 0
    for i in range(len(a)):
        x = a[i]; y = b[i]
        sumxx += x*x
        sumyy += y*y
        sumxy += x*y
    return sumxy/math.sqrt(sumxx*sumyy)
```

Guardar el índice de similitud entre los comentarios

```
def pair(s):
    for i, v1 in enumerate(s):
        for j in range(i+1, len(s)):
            yield [v1, s[j]]

dic={}
for (a,b) in list(pair(['Com 1', 'Com 2', 'Com 3', 'Com 4', 'Com 5'])):
    dic[(a,b)] = cosine_similarity(term_document_matrix[a].tolist(), term_document_matrix[b].to
```

Print the cosine similarity index

```
print(dic)
```

```
{('Com 1', 'Com 2'): 0.0, ('Com 1', 'Com 3'): 0.0, ('Com 1', 'Com 4'): 0.0, ('Com 1', 'Com 5'): 0.0, ('Com 2', 'Com 3'): 0.0, ('Com 2', 'Com 4'): 0.0, ('Com 2', 'Com 5'): 0.0, ('Com 3', 'Com 4'): 0.0, ('Com 3', 'Com 5'): 0.0, ('Com 4', 'Com 5'): 0.0}
```

```
documents= ['Com 1', 'Com 2', 'Com 3', 'Com 4', 'Com 5']
final_df = pd.DataFrame(np.asarray([[dic[(x,y)] if (x,y) in dic else 0] for y in documents]
final_df.columns = documents
final_df.index = documents
```

```
fig, ax = plt.subplots()
ax.set_xticks(np.arange(len(documents)))
ax.set_yticks(np.arange(len(documents)))
```

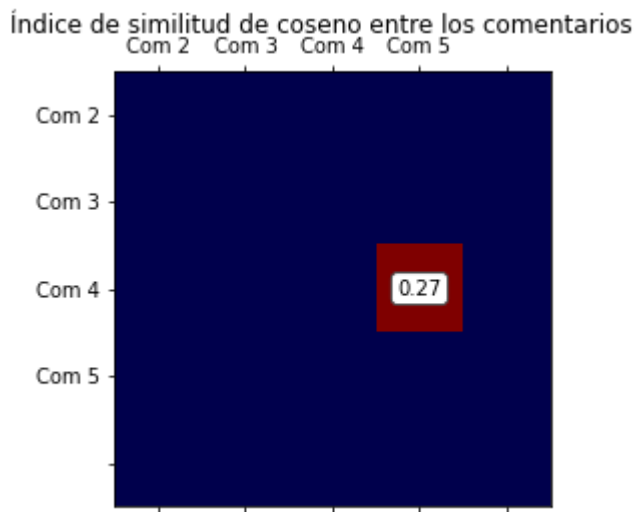


```

ax.set_xticklabels(documents)
ax.set_yticklabels(documents)
ax.matshow(final_df, cmap='seismic')
for (i, j), z in np.ndenumerate(final_df):
    if z != 0 :
        ax.text(j, i, '{:0.2f}'.format(z), ha='center', va='center',
                bbox=dict(boxstyle='round', facecolor='white', edgecolor='0.3'))
    else:
        None

fig.suptitle('Índice de similitud de coseno entre los comentarios')
plt.show()

```



▼ Latent Dirichlet Allocation - Modelo no supervisado de Tópicos.

```

vectorizer_ = CountVectorizer(min_df=5, max_df=0.9, stop_words= stopwords_es, lowercase=True,
com_vectorized = vectorizer_.fit_transform(equ_pdg_text_1['P6'])

```

```

<>:1: DeprecationWarning: invalid escape sequence \-
<>:1: DeprecationWarning: invalid escape sequence \-
<>:1: DeprecationWarning: invalid escape sequence \-
<ipython-input-307-548dd9c02b7d>:1: DeprecationWarning: invalid escape sequence \-
vectorizer_ = CountVectorizer(min_df=5, max_df=0.9, stop_words= stopwords_es, lowercase=
/usr/local/lib/python3.7/dist-packages/sklearn/feature_extraction/text.py:401: UserWarni
% sorted(inconsistent)

```

```

lda_ = LatentDirichletAllocation(n_components=10, max_iter=5, learning_method='online', verbose=1)
tweets_lda = lda_.fit_transform(com_vectorized)

```

```

iteration: 1 of max_iter: 5
iteration: 2 of max_iter: 5
iteration: 3 of max_iter: 5
iteration: 4 of max_iter: 5
iteration: 5 of max_iter: 5

```

▼ keywords

```
def selected_topics(model, vectorizer, top_n=10):
    for idx, topic in enumerate(model.components_):
        print("Topic %d:" % (idx))
        print([(vectorizer.get_feature_names()[i], topic[i])
                for i in topic.argsort()[::-top_n - 1:-1]])

print("Modelo de comentarios LDA:")
selected_topics(lda_, vectorizer_)

Modelo de comentarios LDA:
Topic 0:
[('espacio', 2.338969703519724), ('tecnol', 0.823526715530892), ('gico', 0.745145618534)
Topic 1:
[('socializar', 6.432557053168243), ('epc', 2.7122148544601368), ('proyecto', 2.4843335)
Topic 2:
[('reuni', 1.0840289972540054), ('actividad', 1.066762841504616), ('epc', 1.03777876291)
Topic 3:
[('proyecto', 3.856329159091621), ('conocimiento', 3.6382078230217814), ('sennova', 2.8)
Topic 4:
[('actividad', 1.1474860862483596), ('realizar', 1.1201320993192398), ('espacio', 1.069)
Topic 5:
[('conocer', 4.46738816358936), ('proyecto', 2.854177345305704), ('realizado', 2.733424)
Topic 6:
[('sennova', 0.7082484719303066), ('proyecto', 0.6851988375961179), ('equipo', 0.616038)
Topic 7:
[('socializaci', 8.195035028418207), ('realizar', 3.906421752161277), ('reuni', 3.41185)
Topic 8:
[('gico', 11.70593146740003), ('equipo', 9.635080720432931), ('pedag', 6.35361438062144)
Topic 9:
[('sennova', 0.6918640897340296), ('socializar', 0.6573764128505081), ('proyecto', 0.59)
/usr/local/lib/python3.7/dist-packages/sklearn/utils/deprecation.py:87: FutureWarning: F
    warnings.warn(msg, category=FutureWarning)
```

```
pyLDAvis.enable_notebook()
dash = pyLDAvis.sklearn.prepare(lda_, com_vectorized, vectorizer_, mds='tsne')
dash
```

```
/usr/local/lib/python3.7/dist-packages/pyLDAvis/_prepare.py:247: FutureWarning: In a fut
by='saliency', ascending=False).head(R).drop('saliency', 1)
/usr/local/lib/python3.7/dist-packages/sklearn/manifold/_t_sne.py:783: FutureWarning: Th
FutureWarning,
/usr/local/lib/python3.7/dist-packages/sklearn/manifold/_t_sne.py:793: FutureWarning: Th
FutureWarning,
/usr/local/lib/python3.7/dist-packages/sklearn/manifold/_t_sne.py:827: FutureWarning: 's
FutureWarning,
```

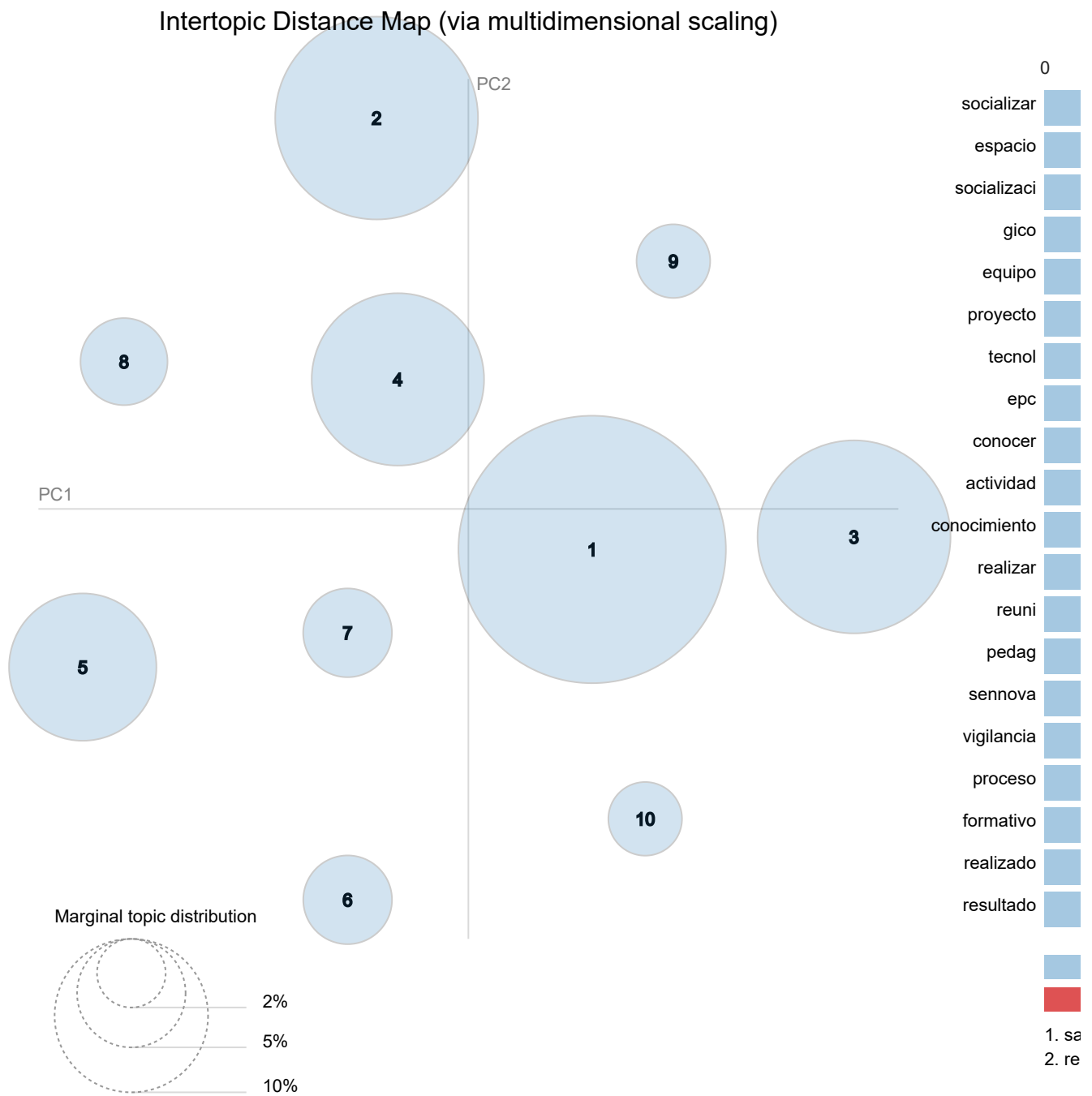
Selected Topic:

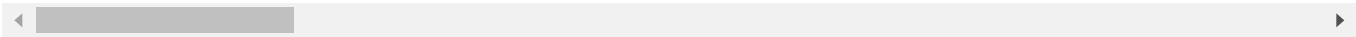
Previous Topic

Next Topic

Clear Topic

Slid





▼ Analítica de datos ENCUESTA EQUIPO EJECUTOR

```
pip install pyLDAvis
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Requirement already satisfied: pyLDAvis in /usr/local/lib/python3.7/dist-packages (3.3.1)
Requirement already satisfied: numpy>=1.20.0 in /usr/local/lib/python3.7/dist-packages (from pyLDAvis)
Requirement already satisfied: pandas>=1.2.0 in /usr/local/lib/python3.7/dist-packages (from pyLDAvis)
Requirement already satisfied: gensim in /usr/local/lib/python3.7/dist-packages (from pyLDAvis)
Requirement already satisfied: numexpr in /usr/local/lib/python3.7/dist-packages (from pyLDAvis)
Requirement already satisfied: scipy in /usr/local/lib/python3.7/dist-packages (from pyLDAvis)
Requirement already satisfied: joblib in /usr/local/lib/python3.7/dist-packages (from pyLDAvis)
Requirement already satisfied: funcy in /usr/local/lib/python3.7/dist-packages (from pyLDAvis)
Requirement already satisfied: sklearn in /usr/local/lib/python3.7/dist-packages (from pyLDAvis)
Requirement already satisfied: future in /usr/local/lib/python3.7/dist-packages (from pyLDAvis)
Requirement already satisfied: scikit-learn in /usr/local/lib/python3.7/dist-packages (from pyLDAvis)
Requirement already satisfied: Jinja2 in /usr/local/lib/python3.7/dist-packages (from pyLDAvis)
Requirement already satisfied: setuptools in /usr/local/lib/python3.7/dist-packages (from pyLDAvis)
Requirement already satisfied: pytz>=2017.3 in /usr/local/lib/python3.7/dist-packages (from pyLDAvis)
Requirement already satisfied: python-dateutil>=2.7.3 in /usr/local/lib/python3.7/dist-packages (from pyLDAvis)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.7/dist-packages (from pyLDAvis)
Requirement already satisfied: smart-open>=1.2.1 in /usr/local/lib/python3.7/dist-packages (from pyLDAvis)
Requirement already satisfied: MarkupSafe>=0.23 in /usr/local/lib/python3.7/dist-packages (from pyLDAvis)
Requirement already satisfied: packaging in /usr/local/lib/python3.7/dist-packages (from pyLDAvis)
Requirement already satisfied: pyparsing!=3.0.5,>=2.0.2 in /usr/local/lib/python3.7/dist-packages (from pyLDAvis)
Requirement already satisfied: threadpoolctl>=2.0.0 in /usr/local/lib/python3.7/dist-packages (from pyLDAvis)
```

```
pip install factor_analyzer
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Requirement already satisfied: factor_analyzer in /usr/local/lib/python3.7/dist-packages (0.1.1)
Requirement already satisfied: pandas in /usr/local/lib/python3.7/dist-packages (from factor_analyzer)
Requirement already satisfied: numpy in /usr/local/lib/python3.7/dist-packages (from factor_analyzer)
Requirement already satisfied: scikit-learn in /usr/local/lib/python3.7/dist-packages (from factor_analyzer)
Requirement already satisfied: scipy in /usr/local/lib/python3.7/dist-packages (from factor_analyzer)
Requirement already satisfied: python-dateutil>=2.7.3 in /usr/local/lib/python3.7/dist-packages (from factor_analyzer)
Requirement already satisfied: pytz>=2017.3 in /usr/local/lib/python3.7/dist-packages (from factor_analyzer)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.7/dist-packages (from factor_analyzer)
Requirement already satisfied: joblib>=0.11 in /usr/local/lib/python3.7/dist-packages (from factor_analyzer)
Requirement already satisfied: threadpoolctl>=2.0.0 in /usr/local/lib/python3.7/dist-packages (from factor_analyzer)
```

```
#Importanto bibliotecas
import pandas as pd
import numpy as np
import math
import re
#Nubes de Palabras y editores de texto
```

```

from wordcloud import WordCloud, ImageColorGenerator
from nltk.corpus import stopwords # para stopwords en español
import nltk
from nltk import tokenize
from nltk.tokenize import word_tokenize
import spacy
# Importacion de spacy
import spacy.cli
#Gráfica
import matplotlib.pyplot as plt
import seaborn as sns
# Importando el CountVectorizer
from sklearn.feature_extraction.text import CountVectorizer
import pyLDAvis.sklearn
nltk.download('punkt') # necesario para descargar el tokenizador
nltk.download('stopwords')
spacy.cli.download("es_core_news_sm")
import es_core_news_sm
spc_es = es_core_news_sm.load()

```

```

[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data] Package punkt is already up-to-date!
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data] Package stopwords is already up-to-date!
✓ Download and installation successful
You can now load the package via spacy.load('es_core_news_sm')

```

```
stopwords_es = stopwords.words("spanish")
```

```

from google.colab import drive
drive.mount('/content/drive')

```

Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.moun



```

from google.colab import files
path = "/content/drive/MyDrive/Proyecto SGPS 9748/Analítica Luz Karime_ Regional Cauca/Encues
#Dataframes
equ_ejr= pd.read_excel(path)

equ_ejr.head()

```

			1. ¿La Coordinación Académica asigna tiempos y recursos a los Equipos Ejecutores, para realizar el proceso de análisis de pertinencia de los proyectos formativos a los programas de formación?	Por favor, argumente la respuesta anterior:	2. ¿La Coordinación Académica es un actor activo en la formulación de los proyectos formativos?	Por favor, argumente la respuesta anterior: .1	3. ¿Se realiza concertación con el Coordinador Académico sobre los requerimientos para la formación (Talento Humano, materiales de formación, ambientes de formación, equipos y herramientas, etc.) en la formulación del proyecto formativo?	P
0	2021-07-21 11:09:51.106		No	NaN	No	NaN	No	
1	2021-07-21 14:35:51.739		Si	La Coordinación Académica permite espacios par...	Si	La Coordinación Académica realiza verificación...	Si	C ha
2	2021-07-27 00:20:13.586		Si	Pueda que se asignen esos tiempos para realiza...	No	En la mayoría de las ocasiones y en un porcent...	Si	n rec
3	2021-08-04 14:03:34.914		Si	A través de la semana de alistamiento y los in...	No	es realizado por el equipo de instructores	No	c a e

▼ EDA(exploratory data analysis)

renombramos columnas

```
columns_descriptions = list(equ_ejr.columns)
```

```
columns_shorten=['Marca_temporal','P1','P2','P3','P4','P5','P6','P7','P8','P9','P10','P11','P12','P13','P14','P15','P16','P17','P18','P19','P20','P21','P22','P23','P24','P25','P26','P27','P28','P29','P30','P31','P32','P33']
```

```
col_renames = {}
col_description = {}
for desc,short in zip(columns_descriptions,columns_shorten):
    print('\033[1m'+short,':\033[0;0m')
    print('\tDescription: ', desc)
    col_renames[desc] = short
    col_description[short] = desc
```

Marca_temporal :

Description: Marca temporal

P1 :

Description: 1. ¿La Coordinación Académica asigna tiempos y recursos a los E

P2 :

Description: Por favor, argumente la respuesta anterior:

P3 :

Description: 2. ¿La Coordinación Académica es un actor activo en la formulac

P4 :

Description: Por favor, argumente la respuesta anterior: .1

P5 :

Description: 3. ¿Se realiza concertación con el Coordinador Académico sobre

P6 :

Description: Por favor, argumente la respuesta anterior: .2

P7 :

Description: 4. ¿Para la formulación del proyecto formativo, el Coordinador ,

P8 :

Description: Por favor, argumente su respuesta anterior:

P9 :

Description: 5. ¿La Coordinación Académica, socializa los resultados de las I

P10 :

Description: Por favor, argumente la respuesta anterior:

P11 :

Description: 6. La designación del líder del Programa de formación se realiz

P12 :

Description: 6. La designación del líder del Programa de formación se realiz

P13 :

Description: 6. La designación del líder del Programa de formación se realiz

P14 :

Description: 6. La designación del líder del Programa de formación se realiz

P15 :

Description: 1. ¿El Equipo Ejecutor recibe acompañamiento en la formulación

P16 :

Description: Por favor, justifique la respuesta anterior:

P17 :

Description: 2. ¿El Equipo Ejecutor recibe apoyo en la construcción de las G

P18 :

Description: Por favor, argumente la respuesta anterior:.1


```

P19 : Description: 3. ¿El Equipo Pedagógico hace recomendaciones en la inducción a
P20 : Description: Por favor, argumente la respuesta anterior: .3
P21 : Description: 4. ¿El Equipo Pedagógico incentiva la incorporación de estrateg
P22 : Description: Por favor, argumente la respuesta anterior: .4
P23 : Description: 5. El Equipo Ejecutor reconoce, que los proyectos pedagógicos a
P24 : Description: 5. El Equipo Ejecutor reconoce, que los proyectos pedagógicos a
P25 : Description: 5. El Equipo Ejecutor reconoce, que los proyectos pedagógicos a
P26 : Description: 1. ¿El Equipo SENNOVA realiza charlas, sensibilizaciones, invit
P27 : Description: Por favor, argumente la respuesta anterior:
P28 :

```

```

equ_ejr.columns =['Marca_temporal','P1','P2','P3','P4','P5','P6','P7','P8','P9','P10','P11','
                  'P24','P25','P26','P27','P28','P29','P30','P31','P32','P33']

```

```
equ_ejr.info()
```

```

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 32 entries, 0 to 31
Data columns (total 34 columns):
#   Column                Non-Null Count  Dtype
---  -
0   Marca_temporal        32 non-null    datetime64[ns]
1   P1                     32 non-null    object
2   P2                     31 non-null    object
3   P3                     32 non-null    object
4   P4                     30 non-null    object
5   P5                     32 non-null    object
6   P6                     30 non-null    object
7   P7                     32 non-null    object
8   P8                     31 non-null    object
9   P9                     32 non-null    object
10  P10                    30 non-null    object
11  P11                    32 non-null    object
12  P12                    28 non-null    object
13  P13                    28 non-null    object
14  P14                    28 non-null    object
15  P15                    32 non-null    object
16  P16                    31 non-null    object
17  P17                    32 non-null    object
18  P18                    30 non-null    object
19  P19                    32 non-null    object
20  P20                    30 non-null    object
21  P21                    32 non-null    object
22  P22                    31 non-null    object
23  P23                    32 non-null    object

```

```

24 P24          32 non-null    object
25 P25          32 non-null    object
26 P26          32 non-null    object
27 P27          30 non-null    object
28 P28          32 non-null    object
29 P29          31 non-null    object
30 P30          32 non-null    object
31 P31          30 non-null    object
32 P32          32 non-null    object
33 P33          30 non-null    object
dtypes: datetime64[ns](1), object(33)
memory usage: 8.6+ KB

```

```
equ_ejr['P8']
```

```

0          NaN
1  El Coordinador Académico ofrece orientaciones ...
2  Quizás se haga la recomendación de dar a conoc...
3  a través de la muestra de proyectos formativos...
4          lo hacen los instructores
5  se parte de una anlysis de la información de l...
6          De eso no se ha dialogado
7  Si. Porque se sabe que de esta forma se genera...
8          SIEMPRE SE TIENE EN CUENTA LAS EMPRESAS
9  A través del equipo pedagógico se trata de ten...
10         al momento no se ha dado esta situacion
11  Depende del programa para establecer la necesi...
12  Como lo mencione antes, la coordinación académ...
13  Los proyectos formativos en su planteamiento d...
14          Por norma y defecto
15  Las necesidades se tienen en cuenta por las c...
16  Siempre hace énfasis en la importancia de cubr...
17  Por su puesto en todo momento se nos solicita ...
18  Partiendo de las solicitudes, requerimientos d...
19  Cada diseño curricular interviene con sus prop...
20  No solo desde el coordinador académico sino ta...
21          Se siguen los procedimientos de diseño
22  Sii tal vez sin. Las recomendaciones que se da...
23          Se promueve el análisis del entorno.
24  Propone acciones y estrategias de articulación...
25  El coordinador no propone ese tipo de activida...
26  Se propone pero no es usual validar el cumplim...
27          Este 2022 ha sido permanente esto
28  Se realiza una revisión y diagnóstico del sector
29          Lo propone en las jornadas de alistamiento
30          Desconozco
31          Algunas veces
Name: P8, dtype: object

```

```
equ_ejr['P12'].value_counts()
```

```

Siempre      13
Usualmente   7
Nunca        5

```

Rara vez 3
 Name: P12, dtype: int64

equ_ejr.info()

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 32 entries, 0 to 31
Data columns (total 34 columns):
#   Column                Non-Null Count  Dtype
---  -
0   Marca_temporal        32 non-null     datetime64[ns]
1   P1                     32 non-null     object
2   P2                     31 non-null     object
3   P3                     32 non-null     object
4   P4                     30 non-null     object
5   P5                     32 non-null     object
6   P6                     30 non-null     object
7   P7                     32 non-null     object
8   P8                     31 non-null     object
9   P9                     32 non-null     object
10  P10                    30 non-null     object
11  P11                    32 non-null     object
12  P12                    28 non-null     object
13  P13                    28 non-null     object
14  P14                    28 non-null     object
15  P15                    32 non-null     object
16  P16                    31 non-null     object
17  P17                    32 non-null     object
18  P18                    30 non-null     object
19  P19                    32 non-null     object
20  P20                    30 non-null     object
21  P21                    32 non-null     object
22  P22                    31 non-null     object
23  P23                    32 non-null     object
24  P24                    32 non-null     object
25  P25                    32 non-null     object
26  P26                    32 non-null     object
27  P27                    30 non-null     object
28  P28                    32 non-null     object
29  P29                    31 non-null     object
30  P30                    32 non-null     object
31  P31                    30 non-null     object
32  P32                    32 non-null     object
33  P33                    30 non-null     object
dtypes: datetime64[ns](1), object(33)
memory usage: 8.6+ KB
```

#-----Organización por tipo de variables-----

#variables textuales

equ_ejr_text_1 = equ_ejr[['P2','P4','P6','P8','P10','P16','P18','P20','P22','P27','P29','P31']

#Variables textuales si/no

equ_ejr_text_2 = equ_ejr[['P1','P3','P5','P7','P9','P15','P17','P19','P21','P26','P28','P30'],

```
#Variables con escala likert
```

```
equ_ejr_likert = equ_ejr[['P11','P12','P13','P14','P23','P24','P25']]
```

▼ variables equ_ejr_likert

```
equ_ejr_likert.replace({'Siempre, Rara vez': np.nan, 'Usualmente, Rara vez':np.nan},inplace=T
```

```
/usr/local/lib/python3.7/dist-packages/pandas/core/frame.py:5244: SettingWithCopyWarning
A value is trying to be set on a copy of a slice from a DataFrame
```

```
See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user\_
```

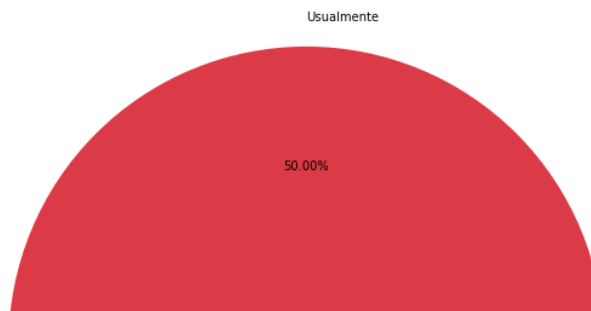


```
equ_ejr_likert = equ_ejr_likert.dropna()
```

```
equ_ejr_likert['P11'].value_counts()
```

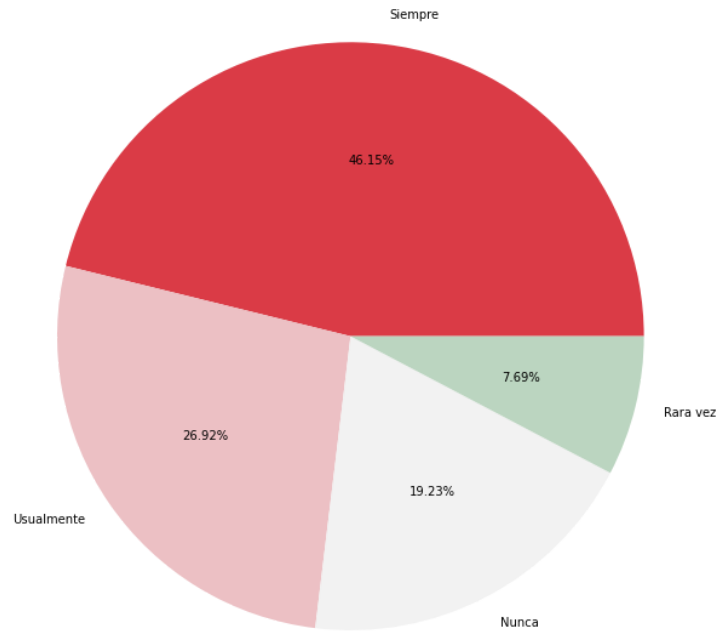
```
Usualmente    13
Siempre       7
Nunca         4
Rara vez      2
Name: P11, dtype: int64
```

```
equ_ejr_likert['P11'].value_counts().plot(kind='pie', autopct='%0.2f%%',figsize=(20,10))
plt.axis('equal');
```



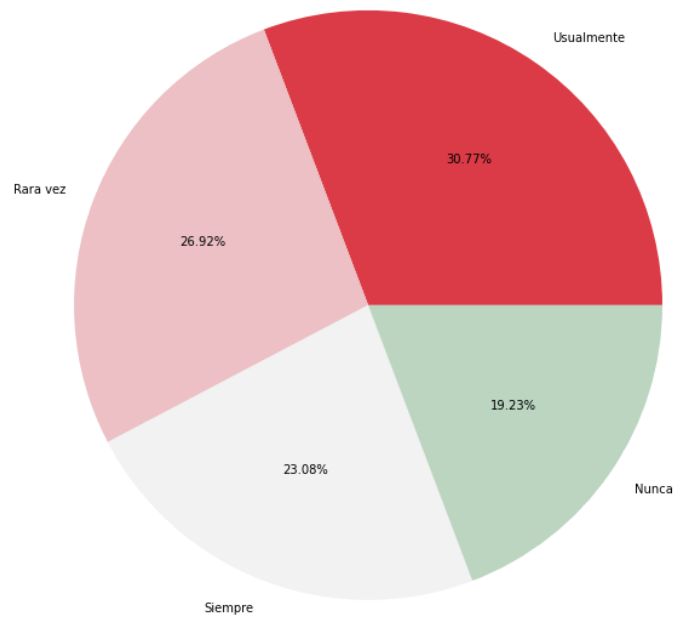
```
equ_ejr_likert['P12'].value_counts().plot(kind='pie', autopct='%.2f%%',figsize=(20,10))  
plt.axis('equal');
```

P12



```
equ_ejr_likert['P13'].value_counts().plot(kind='pie', autopct='%.2f%%',figsize=(20,10))  
plt.axis('equal');
```

p.13



Gráficas de barras resumidas de preguntas

```

answering_columns = list (equ_ejr_likert.columns[1:8])

rows = round( len(answering_columns)//2)
fig, axarr = plt.subplots(rows, 2, figsize=(40, 4*rows))

c = 0
r = 0
answers = {'Siempre': 4, 'Usualmente': 3, 'Rara vez': 2, 'Nunca':1}

#sns.set_palette(sns.cubehelix_palette(5, start=.5, rot=-.75))
sns.set_palette(sns.diverging_palette(10, 133, sep=80, n=5))

size = len(equ_ejr_likert)
for count,col in enumerate(answering_columns):

    #sns.countplot(base[col].sort_values(), ax=axarr[r][c])

    ax = sns.countplot( equ_ejr_likert.iloc[equ_ejr_likert[col].map(answers).argsort()][col],
    axarr[r][c].set_title( col_description[col], fontsize=11)

    #ax = (base[col].value_counts()/len(base)*100).sort_index().plot(kind="bar", rot=0)
    #ax.set_yticks(np.arange(0, 900, 10))
  
```

```

#ax2 = ax.twinx()
#ax2.set_yticks(np.arange(0, 110, 10)*len(base)/100)

for p in ax.patches:
    ax.annotate('{:.2f}%'.format(p.get_height()/len(equ_ejr_likert) * 100), (p.get_x()+0.

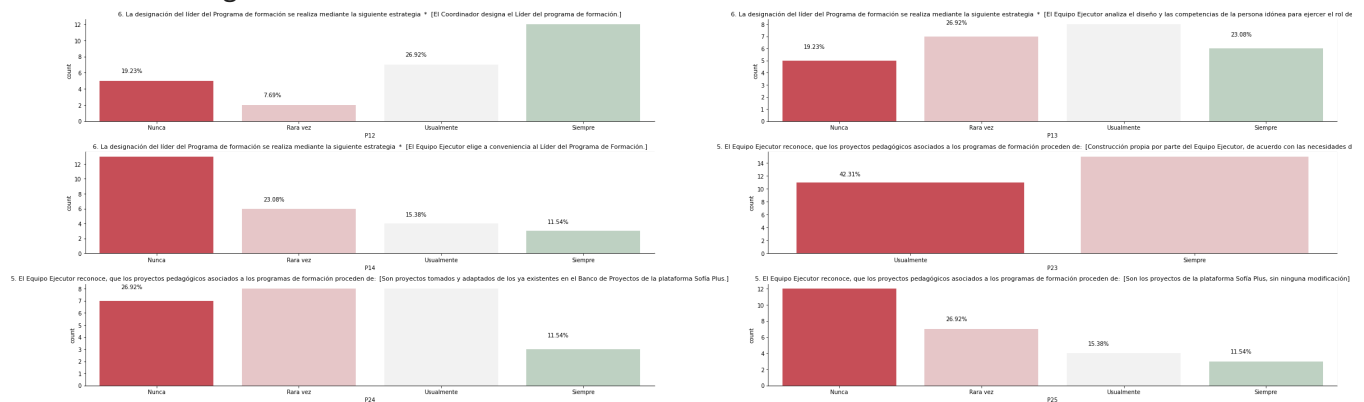
if c >= 1:
    c = 0
    r = r + 1
else:
    c = c + 1
plt.subplots_adjust(hspace=.3)
sns.despine()

```

```

/usr/local/lib/python3.7/dist-packages/seaborn/_decorators.py:43: FutureWarning: Pass the
FutureWarning
/usr/local/lib/python3.7/dist-packages/seaborn/_decorators.py:43: FutureWarning: Pass the
FutureWarning
/usr/local/lib/python3.7/dist-packages/seaborn/_decorators.py:43: FutureWarning: Pass the
FutureWarning
/usr/local/lib/python3.7/dist-packages/seaborn/_decorators.py:43: FutureWarning: Pass the
FutureWarning
/usr/local/lib/python3.7/dist-packages/seaborn/_decorators.py:43: FutureWarning: Pass the
FutureWarning
/usr/local/lib/python3.7/dist-packages/seaborn/_decorators.py:43: FutureWarning: Pass the
FutureWarning

```



Reemplazando variables categóricas a cantidades 'Siempre': 4,'Usualmente': 3,'Rara vez': 2,'Nunca':1

```
equ_ejr_likert.replace({'Siempre': 4, 'Usualmente': 3, 'Rara vez': 2, 'Nunca': 1}, inplace=True)
```

```
equ_ejr_likert
```

	P11	P12	P13	P14	P23	P24	P25
4	3	4	4	2	3	2	1
5	4	4	4	3	4	2	2
6	4	4	3	1	3	3	2
7	3	4	3	4	4	4	3
8	1	1	1	1	3	1	3
9	3	3	2	1	3	3	2
10	3	2	3	1	3	3	1
11	2	4	1	1	4	4	4
12	3	4	2	1	4	1	1
13	2	4	2	2	4	2	1
14	1	1	1	4	4	3	2
15	3	4	3	2	3	2	1
16	4	4	2	2	4	1	1
17	4	4	4	4	4	4	4
20	4	3	4	3	4	2	2
21	1	4	1	1	3	3	3
22	3	1	3	1	4	1	1
23	3	3	2	2	3	3	4
24	4	2	4	1	4	1	1
25	3	3	2	1	4	1	1
26	3	3	3	2	3	3	2
27	3	1	3	1	4	2	1
28	4	1	4	1	4	1	1
29	3	3	3	3	3	2	1
30	1	4	1	1	4	2	2
31	3	3	2	3	3	3	3

Estadística descriptiva de los datos.

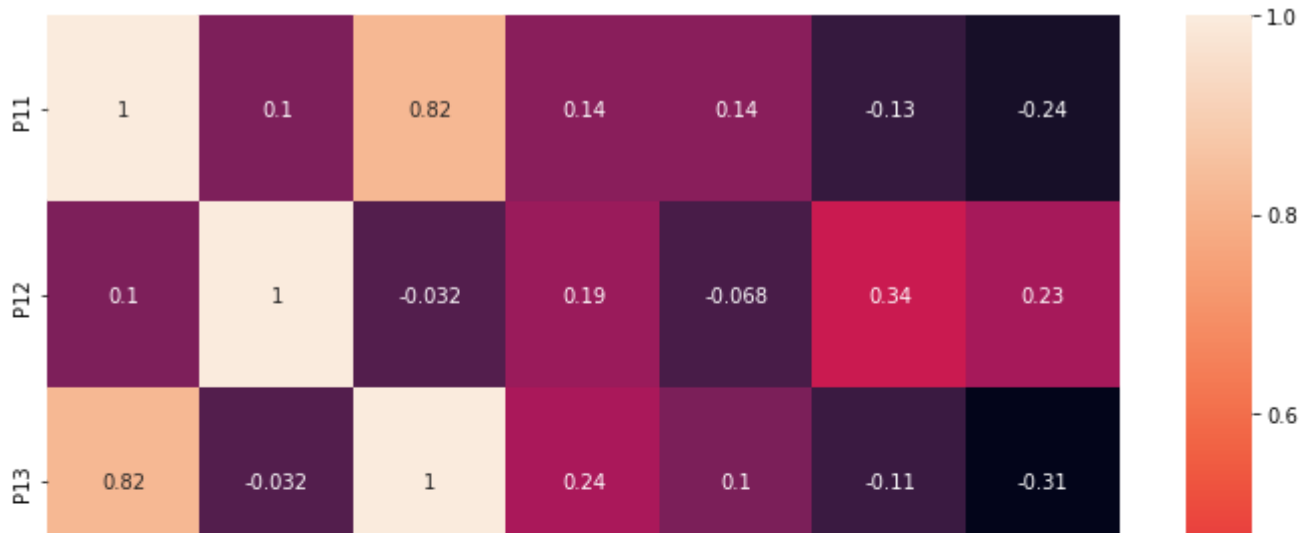

```
equ_ejr_likert.describe()
```

	P11	P12	P13	P14	P23	P24	P25
count	26.000000	26.000000	26.000000	26.000000	26.000000	26.000000	26.000000
mean	2.884615	3.000000	2.576923	1.884615	3.576923	2.269231	1.923077
std	0.993053	1.16619	1.064822	1.070586	0.503831	1.002305	1.055389
min	1.000000	1.000000	1.000000	1.000000	3.000000	1.000000	1.000000
25%	3.000000	2.250000	2.000000	1.000000	3.000000	1.250000	1.000000
50%	3.000000	3.000000	3.000000	1.500000	4.000000	2.000000	2.000000
75%	3.750000	4.000000	3.000000	2.750000	4.000000	3.000000	2.750000
max	4.000000	4.000000	4.000000	4.000000	4.000000	4.000000	4.000000

Correlación de Pearson.

```
corr_df = equ_ejr_likert.corr(method='pearson')
```

```
plt.figure(figsize=(12, 12))
sns.heatmap(corr_df, annot=True)
plt.show()
```



Coeficiente de variación

```
#define function to calculate cv
cv = lambda x: np.std(x, ddof=1) / np.mean(x) * 100

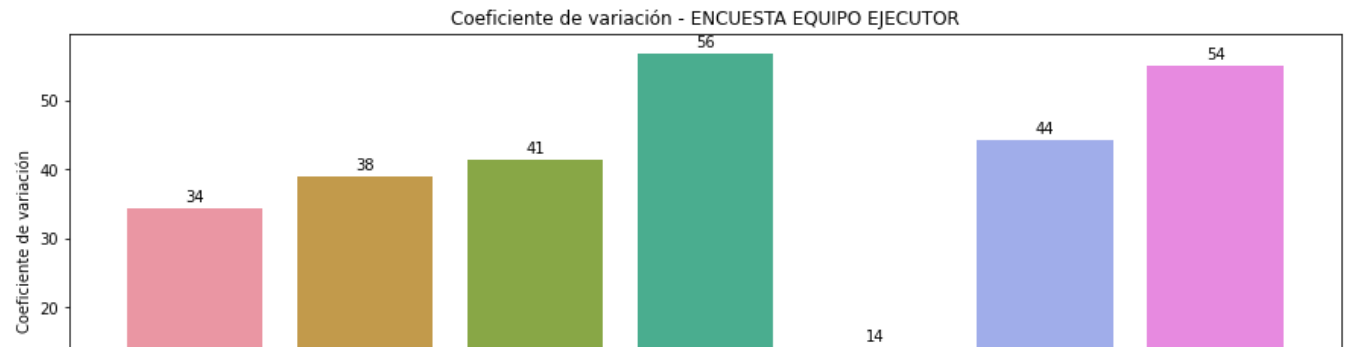
#calculate CV for each column in data frame
df = equ_ejr_likert.apply(cv)

#sns.scatterplot(x=list(results.keys() ),y= list(results.values()))
plt.figure(figsize=(15,5))
ax = sns.barplot(x=equ_ejr_likert.columns,y= df )
#plt.title(title)
#ax = sns.lineplot(x = list(range(0,len(df))),y= df[target].mean(),label='mean ({:0.2f})'.for
for p in ax.patches:
    _x = p.get_x() + p.get_width() / 2
    _y = p.get_y() + p.get_height()+1
    value = str( int(p.get_height() ) )
    ax.text(_x, _y, value, ha="center")

#ax.text(_x, _y, value, ha="left")

ax.set_xticklabels(ax.get_xticklabels(), rotation=45, horizontalalignment='right')
plt.ylabel('Coeficiente de variación')
plt.title('Coeficiente de variación - ENCUESTA EQUIPO EJECUTOR')
plt.plot()
```

[]



▼ Características que más difieren de la Respuesta Deseada

T-Score: número de desviaciones estándar con respecto al valor medio de la población.

```
#T-Score = (Puntuación bruta - Puntuación promedio) / Desviación estándar * 10 + 50
results = {}
```

```
answering_columns_ = answering_columns[:]
```

```
for ans in answering_columns_:
    t_s = (equ_ejr_likert[ans].mean() - equ_ejr_likert[ans].max()) / equ_ejr_likert[ans].std()
    results[ans] = t_s
    print(ans, equ_ejr_likert[ans].mean())
```

```
values_sorted = sorted(results, key=results.get)
values_ = []
for r in values_sorted:
    values_.append( float( '{:0.2f}'.format(results[r]) ) ) )
```

```
print(values_)
```

```
#sns.scatterplot(x=list(results.keys() ),y= list(results.values()))
plt.figure(figsize=(15,6))
ax = sns.barplot(x=values_sorted,y= values_ )
#plt.title(title)
#ax = sns.lineplot(x = list(range(0,len(df))),y= df[target].mean(),label='mean ({:0.2f})'.for
for p in ax.patches:
    _x = p.get_x() + p.get_width() / 2
    _y = p.get_y() + p.get_height()+1
    value = str( int(p.get_height() ) )
    ax.text(_x, _y, value, ha="center")

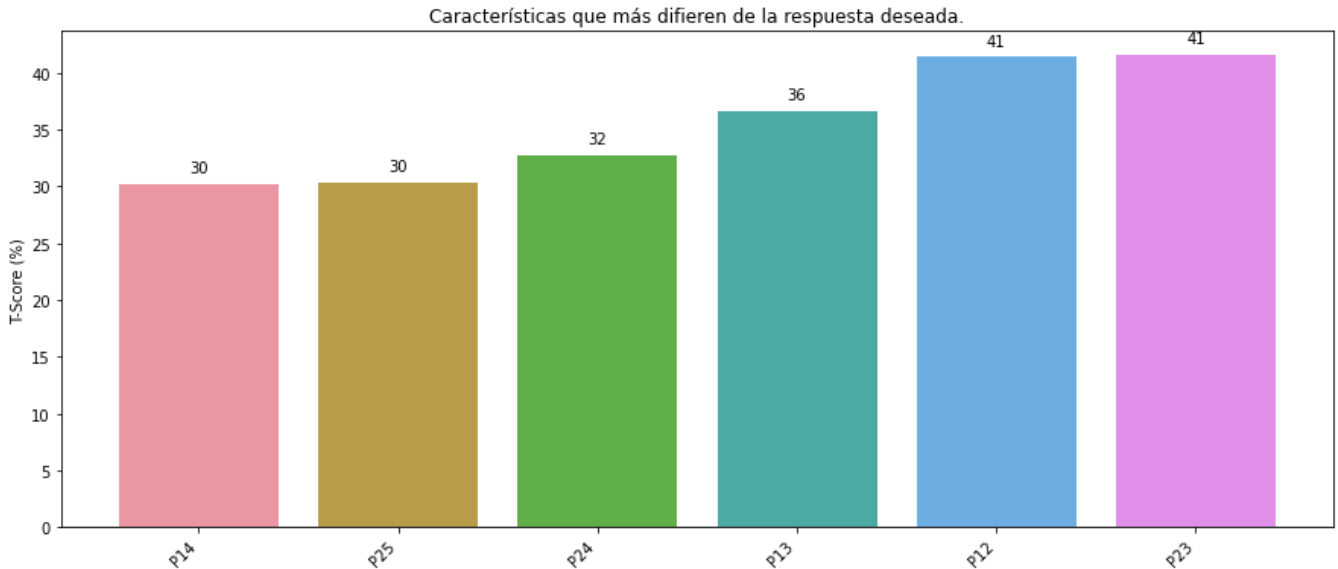
    #ax.text(_x, _y, value, ha="left")
```

```
ax.set_xticklabels(ax.get_xticklabels(), rotation=45, horizontalalignment='right')
plt.ylabel('T-Score (%)')
plt.title('Características que más difieren de la respuesta deseada.')
plt.plot()
```

```

P12 3.0
P13 2.576923076923077
P14 1.8846153846153846
P23 3.576923076923077
P24 2.269230769230769
P25 1.9230769230769231
[30.24, 30.32, 32.73, 36.64, 41.43, 41.6]
[]

```



▼ Análisis Factorial

```
from factor_analyzer import FactorAnalyzer
```

Prueba de esfericidad de Bartlett

comprueba si las variables observadas se interrelacionan o no usando la matriz de correlación observada contra la matriz de identidad. Si la prueba resultó estadísticamente insignificante, no debe emplear un análisis factorial.

```

from factor_analyzer.factor_analyzer import calculate_bartlett_sphericity
chi_square_value,p_value=calculate_bartlett_sphericity(equ_ejr_likert)
chi_square_value, p_value

```

```
(63.9535664485149, 3.163019575858237e-06)
```

En esta prueba de Bartlett, el valor p es cercano a 0. la prueba fue estadísticamente significativa, indicando que la matriz de correlación observada no es una matriz de identidad.

prueba Kaiser-Meyer-Olkin (KMO)

Mide la idoneidad de los datos para el análisis factorial. Determina la adecuación para cada variable observada y para el modelo completo. KMO estima la proporción de varianza entre todas las variables observadas. Identificación de proporción más baja más adecuada para el análisis factorial. Los valores de KMO oscilan entre 0 y 1. Un valor de KMO inferior a 0,6 se considera inadecuado.

```
from factor_analyzer.factor_analyzer import calculate_kmo
```

En esta prueba de Bartlett, el valor p es 0. la prueba fue estadísticamente significativa, indicando que la matriz de correlación observada no es una matriz de identidad.

```
kmo_all,kmo_model=calculate_kmo(equ_ejr_likert)
```

```
kmo_model
```

```
0.517651173834809
```

El KMO general para nuestros datos es 0,51 , lo cual indica que no es apto para continuar con su análisis factorial.

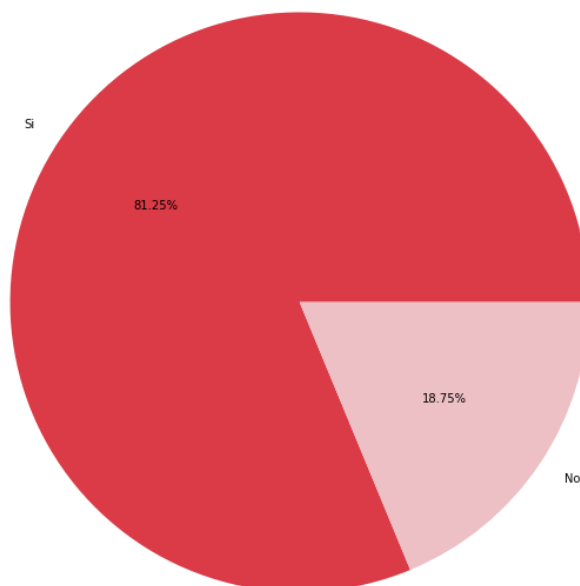
▾ Variables equ_ejr_text_2 tipo Si y No

```
equ_ejr_text_2.head(2)
```

	P1	P3	P5	P7	P9	P15	P17	P19	P21	P26	P28	P30	P32
0	No	No	No	Si	No	No	No	No	No	No	No	Si	Si
1	Si	Si	Si	Si	No	Si	No	No	Si	No	Si	Si	Si

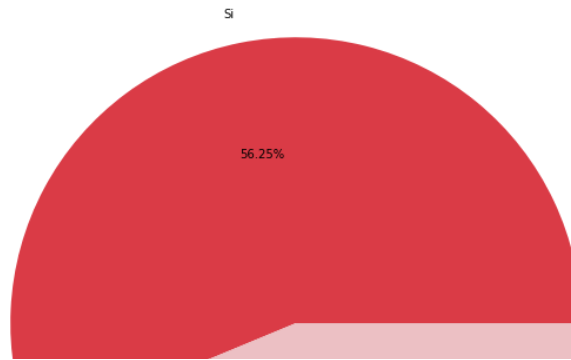
```
equ_ejr_text_2['P1'].value_counts().plot(kind='pie', autopct='%.2f%%',figsize=(20,10))
plt.axis('equal');
```

p1



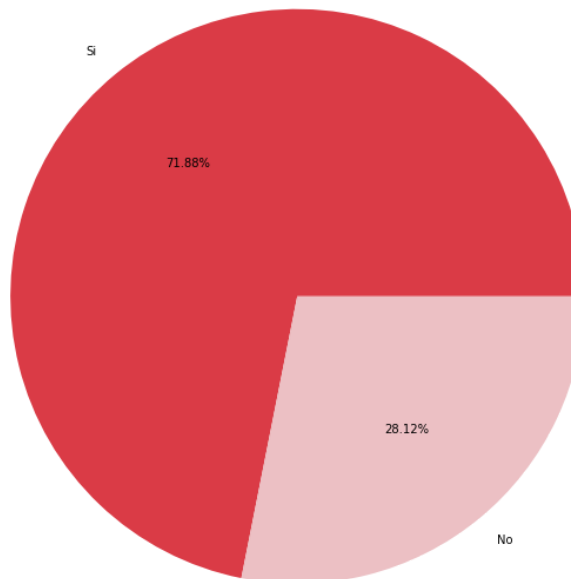
```
equ_ejr_text_2['P3'].value_counts().plot(kind='pie', autopct='%.2f%%',figsize=(20,10))  
plt.axis('equal');
```

P3



```
equ_ejr_text_2['P5'].value_counts().plot(kind='pie', autopct='%2f%%',figsize=(20,10))
plt.axis('equal');
```

P5



Reemplazando variables categóricas a cantidades 'No': 1, 'Si':2

```
equ_ejr_text_2.replace({'No': 1, 'Si':2},inplace=True)
```

```
/usr/local/lib/python3.7/dist-packages/pandas/core/frame.py:5244: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame
```

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_method=method,



equ_ejr_text_2

	P1	P3	P5	P7	P9	P15	P17	P19	P21	P26	P28	P30	P32
0	1	1	1	2	1	1	1	1	1	1	1	2	2
1	2	2	2	2	1	2	1	1	2	1	2	2	2
2	2	1	2	2	1	1	2	2	2	1	2	1	1
3	2	1	1	2	1	1	2	1	2	2	2	2	2

Estadística descriptiva de los datos.

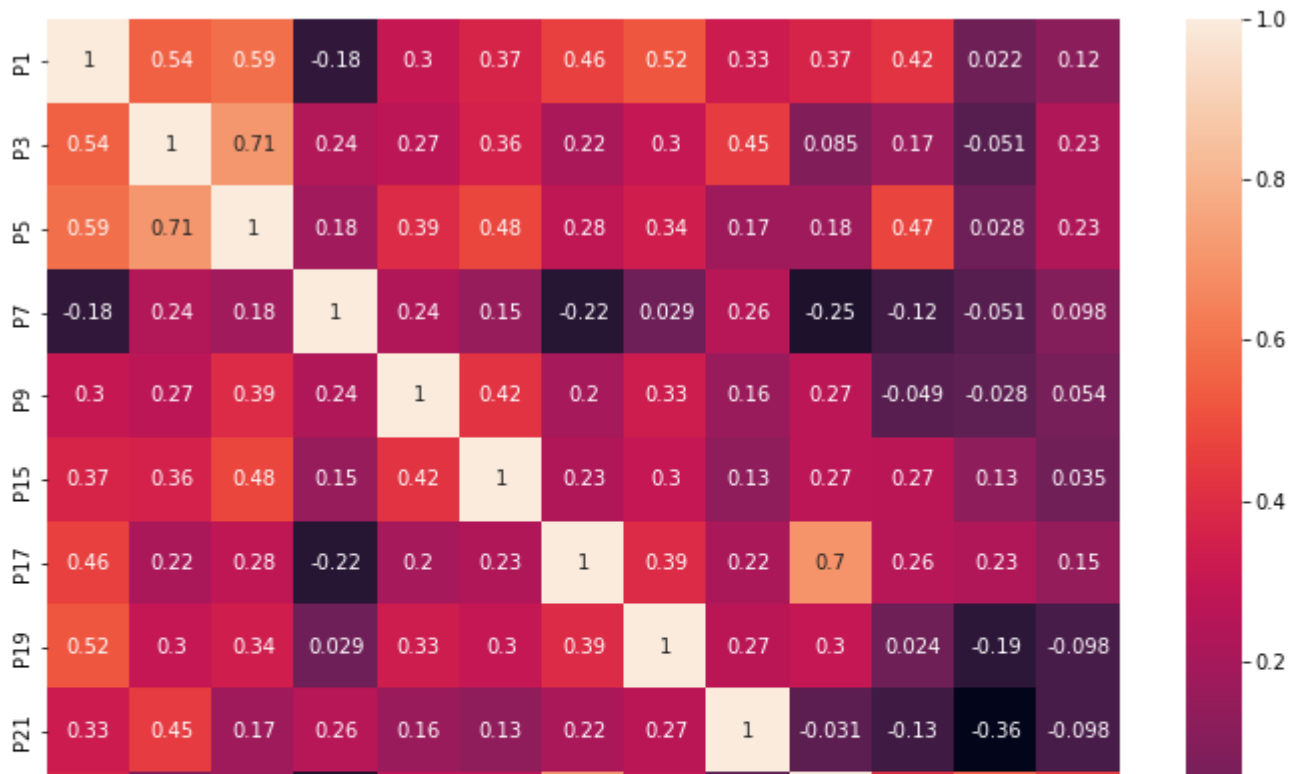
```
equ_ejr_text_2.describe()
```

	P1	P3	P5	P7	P9	P15	P17	P19
count	32.000000	32.000000	32.000000	32.000000	32.000000	32.000000	32.000000	32.000000
mean	1.812500	1.562500	1.718750	1.875000	1.281250	1.687500	1.750000	1.781250
std	0.396558	0.504016	0.456803	0.336011	0.456803	0.470929	0.439941	0.420000
min	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
25%	2.000000	1.000000	1.000000	2.000000	1.000000	1.000000	1.750000	2.000000
50%	2.000000	2.000000	2.000000	2.000000	1.000000	2.000000	2.000000	2.000000
75%	2.000000	2.000000	2.000000	2.000000	2.000000	2.000000	2.000000	2.000000
max	2.000000	2.000000	2.000000	2.000000	2.000000	2.000000	2.000000	2.000000

Correlación de Pearson.

```
corr_df = equ_ejr_text_2.corr(method='pearson')

plt.figure(figsize=(12, 10))
sns.heatmap(corr_df, annot=True)
plt.show()
```



Coeficiente de variación

```
cv = lambda x: np.std(x, ddof=1) / np.mean(x) * 100
```

```
#calculate CV for each column in data frame
```

```
df1 = equ_ejr_text_2.apply(cv)
```

```
#sns.scatterplot(x=list(results.keys() ),y= list(results.values() ) )
```

```
plt.figure(figsize=(15,5))
```

```
ax = sns.barplot(x=equ_ejr_text_2.columns,y= df1 )
```

```
#plt.title(title)
```

```
#ax = sns.lineplot(x = list(range(0,len(df))),y= df[target].mean(),label='mean ({:0.2f})'.for
```

```
for p in ax.patches:
```

```
    _x = p.get_x() + p.get_width() / 2
```

```
    _y = p.get_y() + p.get_height()+1
```

```
    value = str( int(p.get_height() ) )
```

```
    ax.text(_x, _y, value, ha="center")
```

```
    #ax.text(_x, _y, value, ha="left")
```

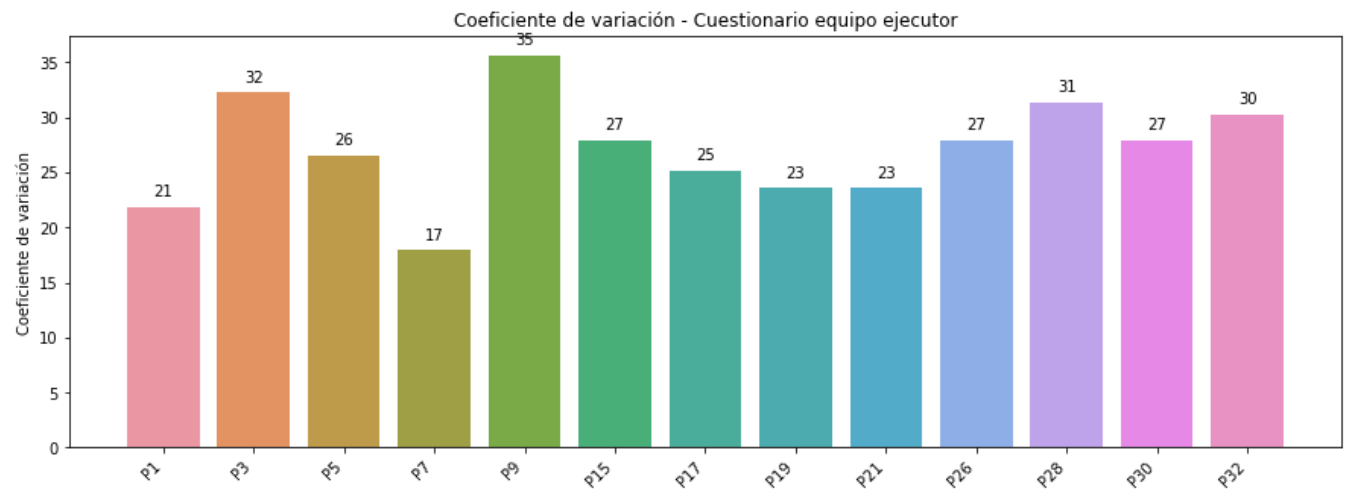
```
ax.set_xticklabels(ax.get_xticklabels(), rotation=45, horizontalalignment='right')
```

```
plt.ylabel('Coeficiente de variación')
```

```
plt.title('Coeficiente de variación - Cuestionario equipo ejecutor')
```

```
plt.plot()
```

[]



Características que más difieren de la Respuesta Deseada

T-Score: número de desviaciones estándar con respecto al valor medio de la población.

▼ Análisis Factorial

Prueba de esfericidad de Bartlett

comprueba si las variables observadas se interrelacionan o no usando la matriz de correlación observada contra la matriz de identidad. Si la prueba resultó estadísticamente insignificante, no debe emplear un análisis factorial.

```
chi_square_value, p_value = calculate_bartlett_sphericity(equ_ejr_text_2)
chi_square_value, p_value
```

```
(192.3847753687999, 1.1986147810484083e-11)
```

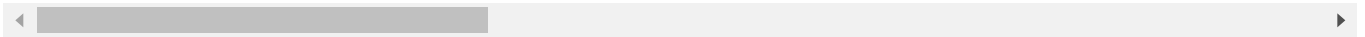
En esta prueba de Bartlett, el valor p es cercano a 0. la prueba fue estadísticamente significativa, indicando que la matriz de correlación observada no es una matriz de identidad.

prueba Kaiser-Meyer-Olkin (KMO)

Mide la idoneidad de los datos para el análisis factorial. Determina la adecuación para cada variable observada y para el modelo completo. KMO estima la proporción de varianza entre todas las variables observadas. Identificación de proporción más baja más adecuada para el análisis factorial. Los valores de KMO oscilan entre 0 y 1. Un valor de KMO inferior a 0,6 se considera inadecuado.

```
kmo_all,kmo_model=calculate_kmo(equ_ejr_text_2)
```

```
/usr/local/lib/python3.7/dist-packages/factor_analyzer/utils.py:249: UserWarning: The ir
warnings.warn('The inverse of the variance-covariance matrix '
```



```
kmo_model
```

```
0.6145615719269294
```

El KMO general para nuestros datos es 0,6 , lo cual indica que no es apto para continuar con su análisis factorial.

▼ NUBE DE PALABRAS

Argumento pregunta P1: ¿La Coordinación Académica asigna tiempos y recursos a los Equipos Ejecutores, para realizar el proceso de análisis de pertinencia de los proyectos formativos a los programas de formación?

```
stopwords = set(stopwords.words('spanish')) # stopwords de nltk en español
```

```
#Lista de Stopword:
```

```
stopwords.update(["|", "-", "NaN"]) # adicionales que no están en stopwords
```

```
# Imagen de nube de palabras
```

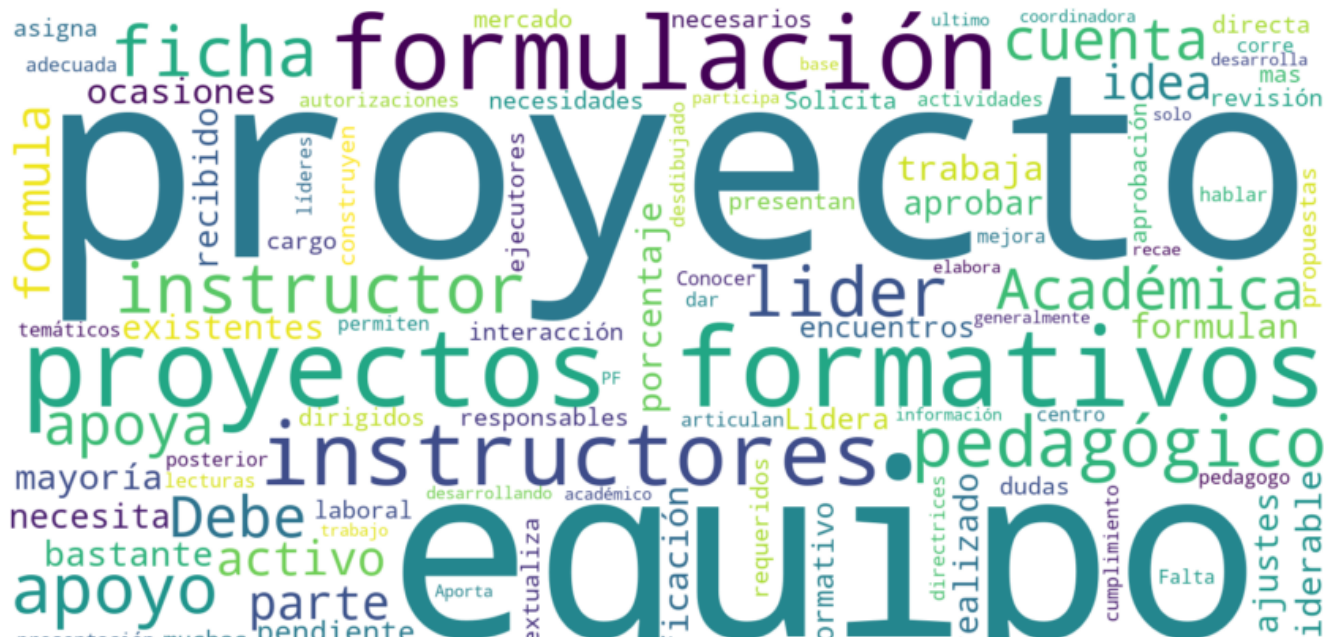
```
text = ' '.join(equ_ejr_text_1['P2'].astype('str'))
wordcloud = WordCloud(width = 1500, height = 1000, stopwords=stopwords, max_font_size=300, ma
plt.figure(figsize = (15,15))
plt.imshow(wordcloud, interpolation="bilinear")
plt.axis("off")
plt.show()
```



#Lista de Stopword:

Imagen de nube de palabras

```
text = ' '.join(equ_ejr_text_1['P4'].astype('str'))
wordcloud = WordCloud(width = 1500, height = 1000, stopwords=stopwords, max_font_size=300, ma
plt.figure(figsize = (15,15))
plt.imshow(wordcloud, interpolation="bilinear")
plt.axis("off")
plt.show()
```



Argumento pregunta P5: ¿Se realiza concertación con el Coordinador Académico sobre los requerimientos para la formación (Talento Humano, materiales de formación, ambientes de formación, equipos y herramientas, etc.) en la formulación del proyecto formativo?

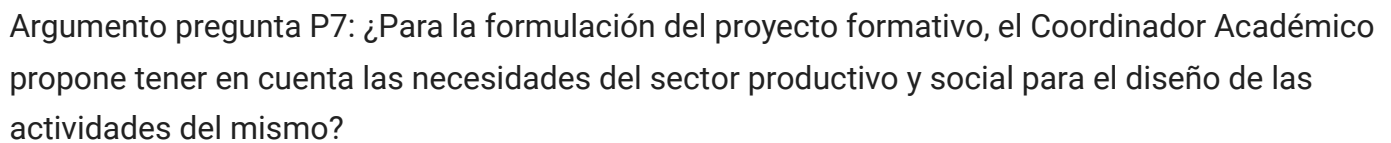


#Lista de Stopword:

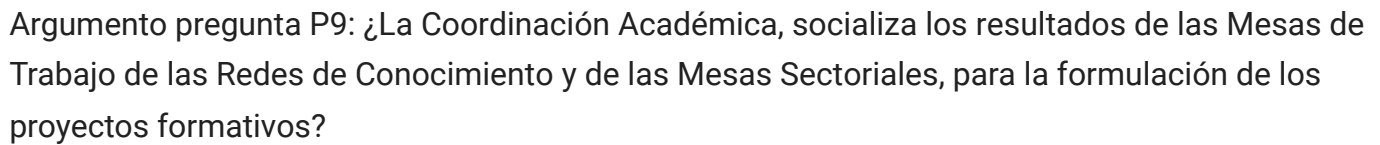
```
stopwords.update(["|", "-", "NaN"]) # adicionales que no están en stopwords
```

Imagen de nube de palabras

```
text = ' '.join(equ_ejr_text_1['P6'].astype('str'))
wordcloud = WordCloud(width = 1500, height = 1000, stopwords=stopwords, max_font_size=300, ma
plt.figure(figsize = (15,15))
plt.imshow(wordcloud, interpolation="bilinear")
plt.axis("off")
plt.show()
```



```
text = ' '.join(equ_ejr_text_1['P8'].astype('str'))
wordcloud = WordCloud(width = 1500, height = 1000, stopwords=stopwords, max_font_size=300, ma
plt.figure(figsize = (15,15))
plt.imshow(wordcloud, interpolation="bilinear")
plt.axis("off")
plt.show()
```



28/51



Argumento pregunta P15: ¿El Equipo Ejecutor recibe acompañamiento en la formulación técnica, metodológica y financiera, de los proyectos formativos?



```
#Lista de Stopword:
```

```
stopwords.update(["|", "-", "NaN"]) # adicionales que no están en stopwords
```

```
# Imagen de nube de palabras
```

```
text = ' '.join(equ_ejr_text_1['P16'].astype('str'))
```

```
wordcloud = WordCloud(width = 1500, height = 1000, stopwords=stopwords, max_font_size=300, ma
```

```
plt.figure(figsize = (15,15))
```

```
plt.imshow(wordcloud, interpolation="bilinear")
```

```
plt.axis("off")
```

```
plt.show()
```



Argumento pregunta P17: ¿El Equipo Ejecutor recibe apoyo en la construcción de las Guías de Aprendizaje e Instrumentos de Evaluación, para obtener los resultados propuestos desde el proyecto de formación?



#Lista de Stopword:

```
stopwords.update(["|", "-", "NaN"]) # adicionales que no están en stopwords
```

Imagen de nube de palabras

```
text = ' '.join(equ_ejr_text_1['P18'].astype('str'))
wordcloud = WordCloud(width = 1500, height = 1000, stopwords=stopwords, max_font_size=300, ma
plt.figure(figsize = (15,15))
plt.imshow(wordcloud, interpolation="bilinear")
plt.axis("off")
plt.show()
```



Argumento pregunta P19: ¿El Equipo Pedagógico hace recomendaciones en la inducción a instructores, para la formulación del proyecto formativo?



#Lista de Stopword:

```
stopwords.update(["|", "-", "NaN"]) # adicionales que no están en stopwords
```

Imagen de nube de palabras

```
text = ' '.join(equ_ejr_text_1['P20'].astype('str'))
wordcloud = WordCloud(width = 1500, height = 1000, stopwords=stopwords, max_font_size=300, ma
plt.figure(figsize = (15,15))
plt.imshow(wordcloud, interpolation="bilinear")
plt.axis("off")
plt.show()
```

--- 1: --- explicación proceso formativo capacitaciones

Argumento pregunta P21:¿El Equipo Pedagógico incentiva la incorporación de estrategias pedagógicas que orienten el proceso de formación hacia la investigación?

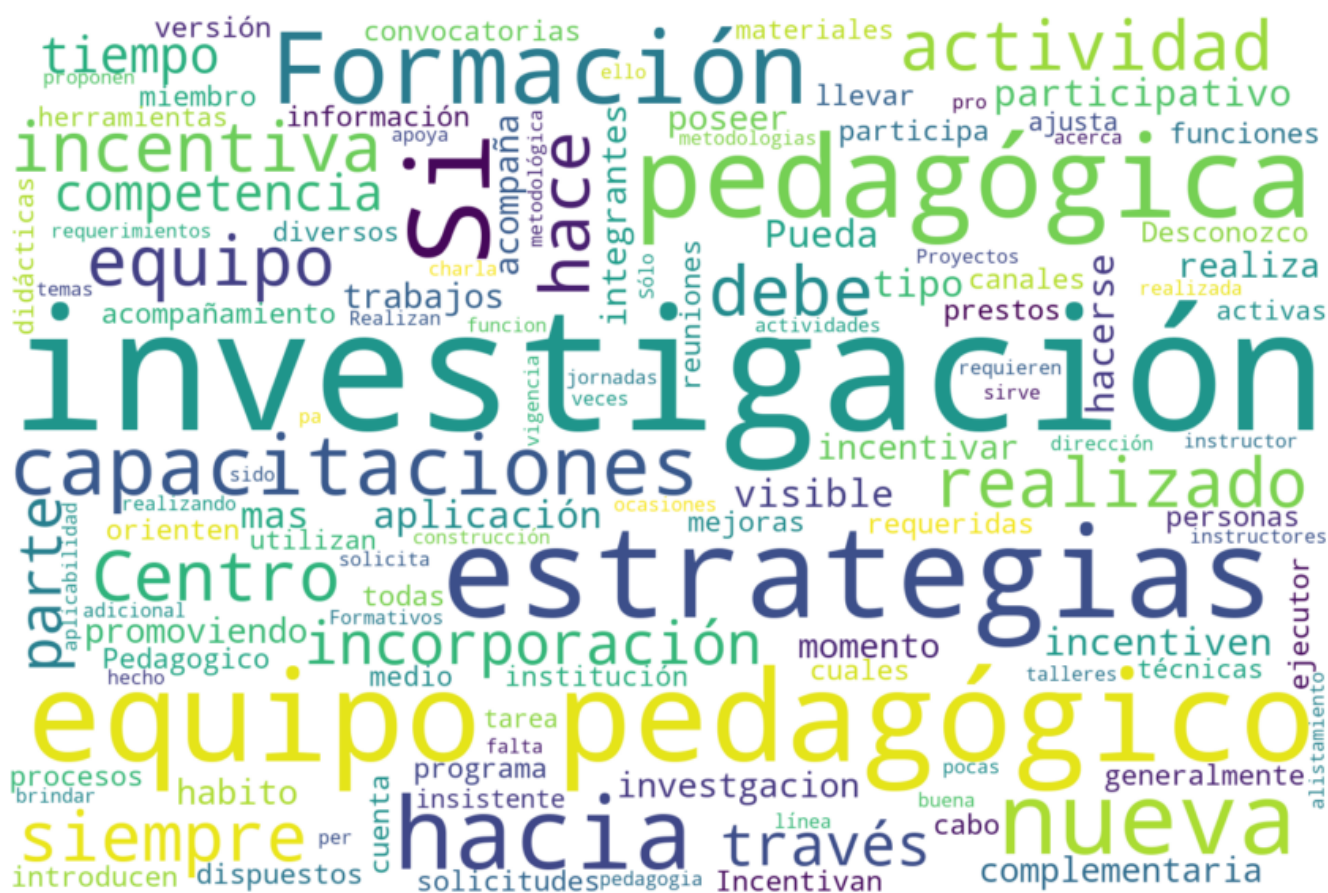
importante acción programa

#Lista de Stopword:

```
stopwords.update(["|", "-", "NaN"]) # adicionales que no están en stopwords
```

Imagen de nube de palabras

```
text = ' '.join(equ_ejr_text_1['P22'].astype('str'))
wordcloud = WordCloud(width = 1500, height = 1000, stopwords=stopwords, max_font_size=300, ma
plt.figure(figsize = (15,15))
plt.imshow(wordcloud, interpolation="bilinear")
plt.axis("off")
plt.show()
```



```
#Lista de Stopword:
stopwords.update(["|", "-", "NaN"]) # adicionales que no están en stopwords

# Imagen de nube de palabras

text = ' '.join(equ_ejr_text_1['P27'].astype('str'))
wordcloud = WordCloud(width = 1500, height = 1000, stopwords=stopwords, max_font_size=300, ma
plt.figure(figsize = (15,15))
plt.imshow(wordcloud, interpolation="bilinear")
plt.axis("off")
plt.show()
```



33/51

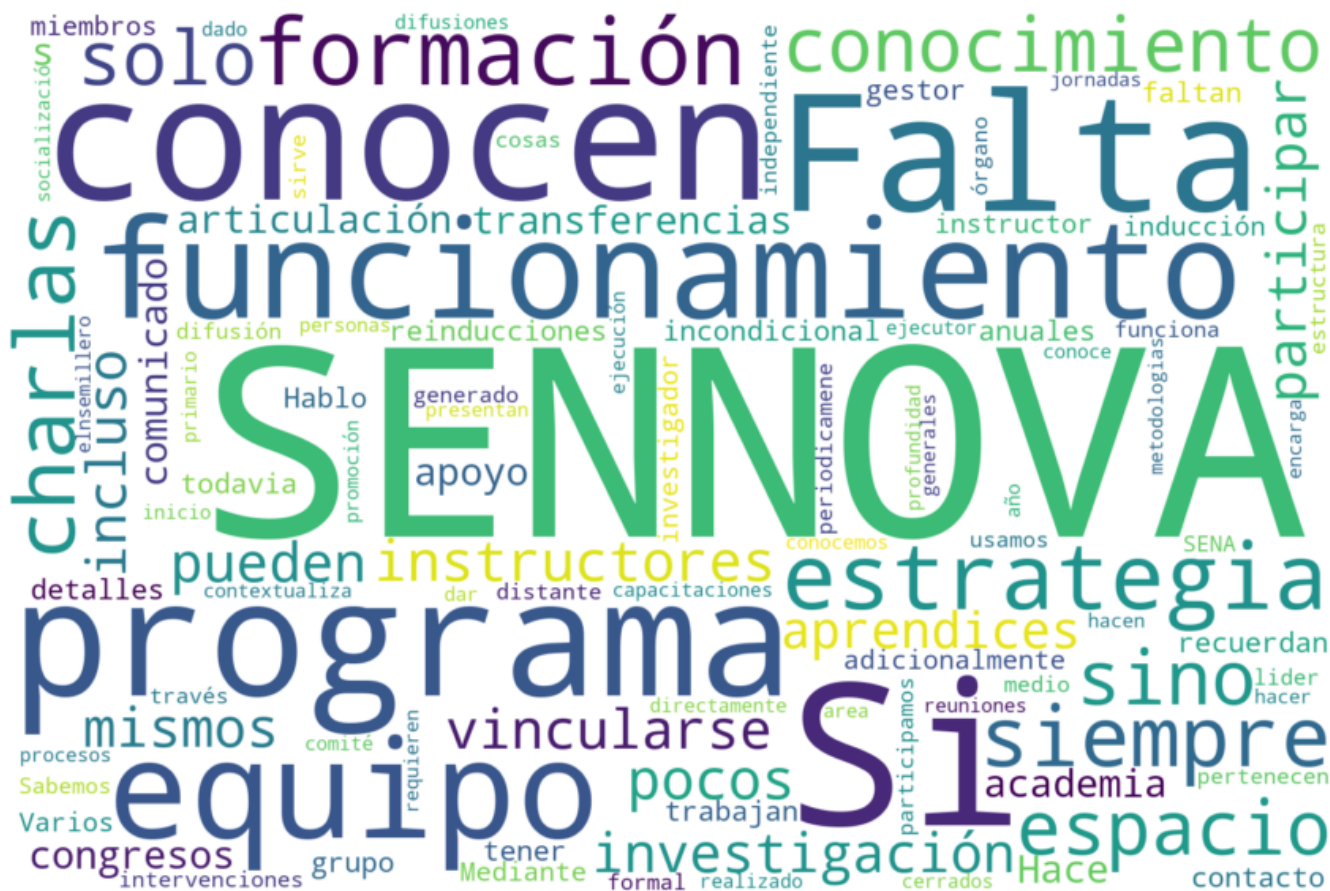
[illegible]

```
#Lista de Stopword:
stopwords.update(["|", "-", "NaN"]) # adicionales que no están en stopwords

# Imagen de nube de palabras
```



```
text = ' '.join(equ_ejr_text_1['P31'].astype('str'))
wordcloud = WordCloud(width = 1500, height = 1000, stopwords=stopwords, max_font_size=300, ma
plt.figure(figsize = (15,15))
plt.imshow(wordcloud, interpolation="bilinear")
plt.axis("off")
plt.show()
```



Argumento pregunta P32: ¿El Equipo Ejecutor conoce el funcionamiento de la estrategia de investigación SENNOVA?

#Lista de Stopword:

```
stopwords.update(["|", "-", "NaN"]) # adicionales que no están en stopwords
```

Imagen de nube de palabras

```
text = ' '.join(equ_ejr_text_1['P33'].astype('str'))
```

```
wordcloud = WordCloud(width = 1500, height = 1000, stopwords=stopwords, max font size=300, ma
```


	P2	P4	P6	P8	P10	
2	Pueda que se asignen esos tiempos para realiza...	En la mayoría de las ocasiones y en un porcent...	Si se concerta la necesidad y requerimiento pa...	Quizás se haga la recomendación de dar a conoc...	En la mayoría de las ocasiones no se cuenta co...	Prec por realiz
3	A través de la semana de alistamiento y los in...	es realizado por el equipo de instructores	casi siempre el alistamiento es posterior cuan...	a traves de la muestra de proyectos formativos...	No tengo información	lo ha en c
4	Tiempo	se necesita para aprobar	En lo que es necesario	lo hacen los instructores	No lo hace	no l
5	se realiza un trabajo en equipos para saber cu...	si	eso queda estipulado en el proyecto formativo	se parte de una analisis de la información de l...	no conozco dicha labor	el c
6	Se asignan tiempos antes de iniciar la formaci...	Se ha recibido apoyo cuando se ha informado de...	Para talento humano es difícil debido a que de...	De eso no se ha dialogado	No he tenido esa información	La durante recup
7	Con un proyecto que se inicio hace un mes me a...	Claro que si son activos están pendiente del p...	Si en todo lo que se puede siempre esta dispue...	Si. Porque se sabe que de esta forma se genera...	Respondí que no porque desconozco si la Coordi...	te pedag
9	Se dispone uno de los integrantes del equipo p...	Lidera los encuentros entre equipo ejecutor y...	Se dispone y alerta sobre la necesidad de real...	A través del equipo pedagógico se trata de ten...	No tengo conocimiento.	pedi: través c
10	Se incluye dentro de la agenda	Debe integrarse mas con los que construyen el ...	Se le da apertura a este proceso antes de abri...	al momento no se ha dado esta situacion	Por el momento no se ha realizado en ADSI	Hemo: acompa d
11	Se realizan reuniones mensuales de los equipos...	Desde coordinación se cuenta con el apoyo del ...	Se realiza reunión para establecer las necesid...	Depende del programa para establecer la necesi...	No se ha socializado	acompa por
12	La coordinación académica nos permite un día	La coordinación académica nos avova	En las reuniones mensuales del equipo	Como lo mencione antes, la coordinación	Con la nueva coordinación no hemos generado	pedagi ac

stopwords_es

```
['de',  
'la',  
'que',  
'el',  
'en',  
'y',  
'a',  
'los',  
'del',  
'se',  
'las',  
'por',  
'un',  
'para',  
'con',  
'no',  
'una',  
'su',  
'al',  
'lo',  
'como',  
'más',  
'pero',  
'sus',  
'le',  
'ya',  
'o',  
'este',  
'sí',  
'porque',  
'esta',  
'entre',  
'cuando',  
'muy',  
'sin',  
'sobre',  
'también',  
'me',  
'hasta',  
'hay',  
'donde',  
'quien',  
'desde',  
'todo',  
'nos',  
'durante',  
'todos',  
'uno',  
'les',  
'ni',  
'contra',  
'otros',  
'ese',  
'eso',  
'ante',
```

```
'ellos',
'e',
'esto'
```

Esta función toma una cadena, reduce todo, filtra solo letras, elimina palabras vacías, lematiza y devuelve la cadena resultante.

decisio...

```
def limpia_texto(texto):
    '''(str) -> str
    Esta función toma una cadena, reduce todo, filtra solo letras, elimina palabras vacías, lem
    '''
    texto = texto.lower()

    texto = re.sub(r"[\W\d_]+", " ", texto)

    texto = [pal for pal in texto.split() if pal not in stopwords_es]

    spc_texto = spc_es(" ".join(texto))
    tokens = [word.lemma_ if word.lemma_ != "-PRON-" else word.lower_ for word in spc_texto]

    return " ".join(tokens)
```

Este ultimo Cada que se

equ_ejr_text_1.info()

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 29 entries, 2 to 31
Data columns (total 13 columns):
 #   Column  Non-Null Count  Dtype
---  -
 0    P2      29 non-null      object
 1    P4      29 non-null      object
 2    P6      29 non-null      object
 3    P8      29 non-null      object
 4   P10     29 non-null      object
 5   P16     29 non-null      object
 6   P18     29 non-null      object
 7   P20     29 non-null      object
 8   P22     29 non-null      object
 9   P27     29 non-null      object
10   P29     29 non-null      object
11  P31     29 non-null      object
12  P33     29 non-null      object
dtypes: object(13)
memory usage: 3.2+ KB
```

31

durante la

idea de

lanplaneacion

Algunas veces

aislamiento

Alour

```
#preprocesamiento de texto filtra solo letras, elimina palabras vacías, lematiza y devuelve l
equ_ejr_text_1['P2'] = equ_ejr_text_1['P2'].apply(limpia_texto)
equ_ejr_text_1['P4'] = equ_ejr_text_1['P4'].apply(limpia_texto)
equ_ejr_text_1['P6'] = equ_ejr_text_1['P6'].apply(limpia_texto)
equ_ejr_text_1['P8'] = equ_ejr_text_1['P8'].apply(limpia_texto)
equ_ejr_text_1['P10'] = equ_ejr_text_1['P10'].apply(limpia_texto)
```

```
equ_ejr_text_1['P16'] = equ_ejr_text_1['P16'].apply(limpia_texto)
equ_ejr_text_1['P18'] = equ_ejr_text_1['P18'].apply(limpia_texto)
equ_ejr_text_1['P20'] = equ_ejr_text_1['P20'].apply(limpia_texto)
equ_ejr_text_1['P22'] = equ_ejr_text_1['P22'].apply(limpia_texto)
equ_ejr_text_1['P27'] = equ_ejr_text_1['P27'].apply(limpia_texto)
equ_ejr_text_1['P29'] = equ_ejr_text_1['P29'].apply(limpia_texto)
equ_ejr_text_1['P31'] = equ_ejr_text_1['P31'].apply(limpia_texto)
equ_ejr_text_1['P33'] = equ_ejr_text_1['P33'].apply(limpia_texto)
#Ejemplo con P6
equ_ejr_text_1.dropna()
```

	P2	P4	P6	P8	P10	
2	poder asignir tiempo realizar dicho actividad ...	mayoría ocasión porcentaje bastante considerab...	si concertar necesidad requerimiento formación...	quizás hacer recomendación dar conocer product...	mayoría ocasión contar retroalimentacion	precisar realizar for proye
3	través semana alistamiento inicio trimestre	realizado equipo instructor	casi siempre alistamiento posterior oferta apr...	trav mostrar proyecto formativo realizar trime...	información	hace ocasió

▼ Matriz termino-documento - Document-Term Matrix

realizar trabajo quedar parte analisis

col_description

```
{'Marca_temporal': 'Marca temporal',
 'P1': '1. ¿La Coordinación Académica asigna tiempos y recursos a los Equipos Ejecutores, para realizar el proceso de análisis de pertinencia de los proyectos formativos a los programas de formación? ',
 'P2': 'Por favor, argumente la respuesta anterior: ',
 'P3': '2. ¿La Coordinación Académica es un actor activo en la formulación de los proyectos formativos? ',
 'P4': 'Por favor, argumente la respuesta anterior: .1',
 'P5': '3. ¿Se realiza concertación con el Coordinador Académico sobre los requerimientos para la formación (Talento Humano, materiales de formación, ambientes de formación, equipos y herramientas, etc.) en la formulación del proyecto formativo? ',
 'P6': 'Por favor, argumente la respuesta anterior: .2',
 'P7': '4. ¿Para la formulación del proyecto formativo, el Coordinador Académico propone tener en cuenta las necesidades del sector productivo y social para el diseño de las actividades del mismo? ',
 'P8': 'Por favor, argumente su respuesta anterior: ',
 'P9': '5. ¿La Coordinación Académica, socializa los resultados de las Mesas de Trabajo de las Redes de Conocimiento y de las Mesas Sectoriales, para la formulación de los proyectos formativos? ',
 'P10': 'Por favor, argumente la respuesta anterior:',
 'P11': '6. La designación del líder del Programa de formación se realiza mediante la siguiente estrategia \u202f* [Es consensuada entre el Coordinador Académico y el Equipo ejecutor.]',
 'P12': '6. La designación del líder del Programa de formación se realiza mediante la siguiente estrategia \u202f* [El Coordinador designa el Líder del programa de formación.]',
 'P13': '6. La designación del líder del Programa de formación se realiza mediante la siguiente estrategia \u202f* [El Equipo Ejecutor analiza el diseño y las competencias de la persona idónea para ejercer el rol de Líder.]',
 'P14': '6. La designación del líder del Programa de formación se realiza mediante la siguiente estrategia \u202f* [El Equipo Ejecutor elige a conveniencia al Líder del Programa de Formación.]',
 'P15': '1. ¿El Equipo Ejecutor recibe acompañamiento en la formulación técnica, metodológica y financiera, de los proyectos formativos?\u202f* ',
```



```

realizar      1      0      0      1      0
través        1      1      0      0      0
lado          2      0      0      0      0
saber         0      0      0      1      0
asignir       1      0      0      0      0
nuevo         0      0      0      0      1
inicio        0      1      0      0      0
iniciar       0      0      0      0      1

```

```

/usr/local/lib/python3.7/dist-packages/sklearn/utils/deprecation.py:87: FutureWarning: F
warnings.warn(msg, category=FutureWarning)

```

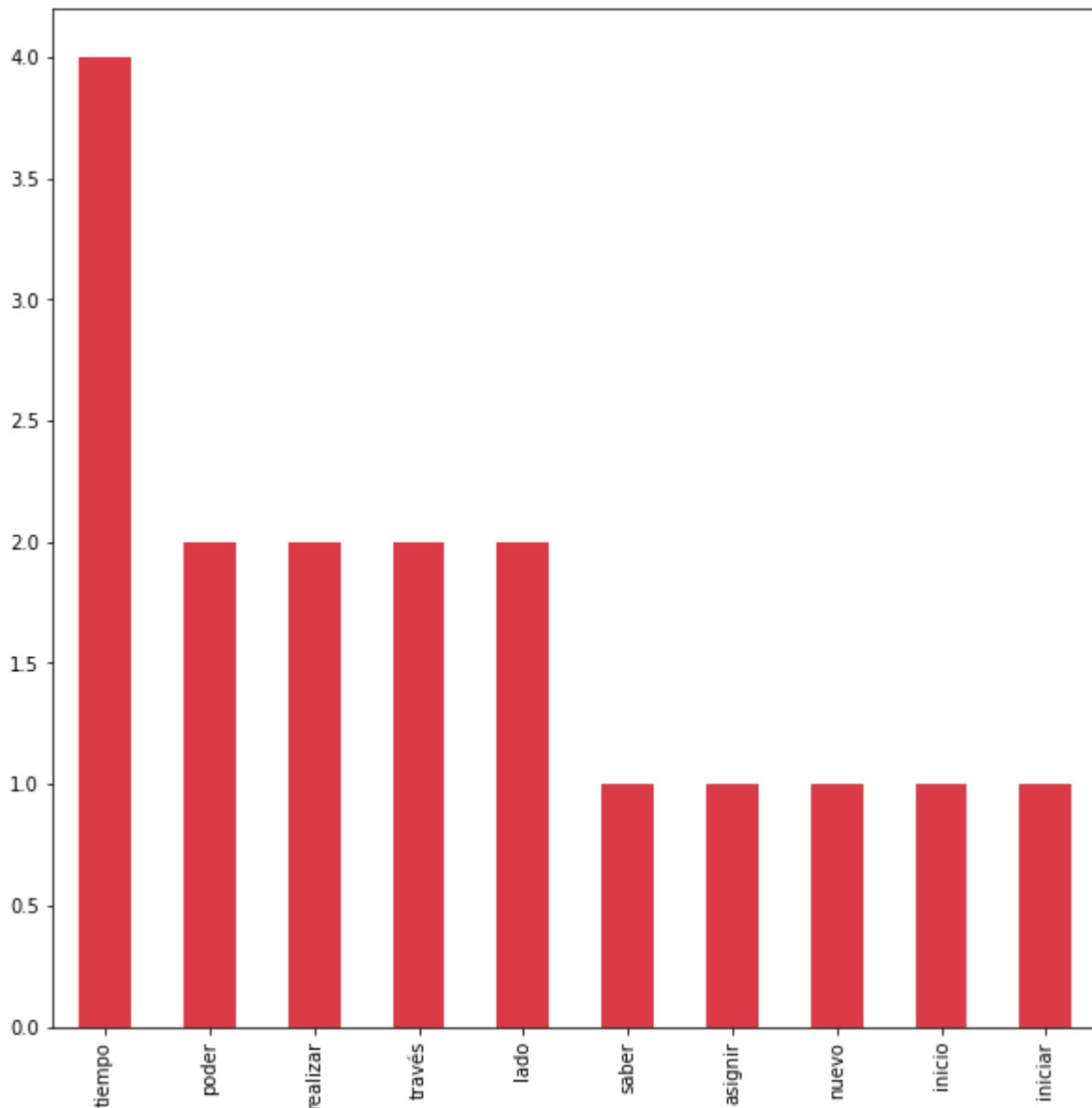
← aistamiento reenvir correos →

▼ Histograma de palabras

```

falta      nunto      ...      ...
plt.figure(figsize=(10,10))
term_document_matrix['total_count'].plot.bar()
plt.show()

```



▼ Semejanza de coseno

```
# Define la función de similitud del coseno.
def cosine_similarity(a,b):
    "calcula la similitud del coseno de v1 a v2: (a punto b)/{||a||*||b||}"
    sumxx, sumxy, sumyy = 0, 0, 0
    for i in range(len(a)):
        x = a[i]; y = b[i]
        sumxx += x*x
        sumyy += y*y
        sumxy += x*y
    return sumxy/math.sqrt(sumxx*sumyy)
```

▼ Latent Dirichlet Allocation - Modelo no supervisado de Tópicos.

```
from sklearn.decomposition import LatentDirichletAllocation
```

```
vectorizer_ = CountVectorizer(min_df=5, max_df=0.9, stop_words= stopwords_es, lowercase=True,
com_vectorized = vectorizer_.fit_transform(equ_ejr_text_1['P2'])
```

```
<>:1: DeprecationWarning: invalid escape sequence \-
<>:1: DeprecationWarning: invalid escape sequence \-
<>:1: DeprecationWarning: invalid escape sequence \-
<ipython-input-279-f932c6d82633>:1: DeprecationWarning: invalid escape sequence \-
    vectorizer_ = CountVectorizer(min_df=5, max_df=0.9, stop_words= stopwords_es, lowercas
/usr/local/lib/python3.7/dist-packages/sklearn/feature_extraction/text.py:401: UserWarni
% sorted(inconsistent)
```

```
lda_ = LatentDirichletAllocation(n_components=10, max_iter=5, learning_method='online', verbos
com_lda = lda_.fit_transform(com_vectorized)
```

```
iteration: 1 of max_iter: 5
iteration: 2 of max_iter: 5
iteration: 3 of max_iter: 5
iteration: 4 of max_iter: 5
iteration: 5 of max_iter: 5
```

▼ keywords

```
def selected_topics(model, vectorizer, top_n=10):
    for idx, topic in enumerate(model.components_):
        print("Topic %d:" % (idx))
```



```

print([(vectorizer.get_feature_names()[i], topic[i])
       for i in topic.argsort()[::-top_n - 1:-1]])

print("Modelo de comentarios LDA:")
selected_topics(lda_, vectorizer_)

Modelo de comentarios LDA:
Topic 0:
[('tiempo', 8.567635667570645), ('equipo', 2.2669085278681353), ('desarrollo', 2.2480000000000002)]
Topic 1:
[('tiempo', 0.5544435181899667), ('actividad', 0.5186274914474076), ('desarrollo', 0.48000000000000004)]
Topic 2:
[('equipo', 0.6801457918045917), ('realizar', 0.6661286219470014), ('curricular', 0.5050000000000001)]
Topic 3:
[('equipo', 2.9832343240731487), ('actividad', 1.552785627120182), ('realizar', 1.4944444444444444)]
Topic 4:
[('actividad', 0.6149586016467175), ('asignar', 0.503022533953781), ('curricular', 0.49000000000000005)]
Topic 5:
[('tiempo', 0.6392804197940477), ('curricular', 0.5981242134224559), ('desarrollo', 0.5000000000000001)]
Topic 6:
[('asignar', 0.6263015787302342), ('realizar', 0.6100811280494413), ('equipo', 0.5611800000000001)]
Topic 7:
[('asignar', 3.331479866066715), ('tiempo', 2.8072515936833753), ('actividad', 0.9808890000000001)]
Topic 8:
[('desarrollo', 2.703674478060763), ('curricular', 2.6787329770055432), ('realizar', 1.4000000000000001)]
Topic 9:
[('curricular', 0.5053465507965926), ('actividad', 0.49155864865401916), ('tiempo', 0.40000000000000006)]
/usr/local/lib/python3.7/dist-packages/sklearn/utils/deprecation.py:87: FutureWarning: F
  warnings.warn(msg, category=FutureWarning)

```

```

pyLDAvis.enable_notebook()
dash = pyLDAvis.sklearn.prepare(lda_, com_vectorized, vectorizer_, mds='tsne')
dash

```

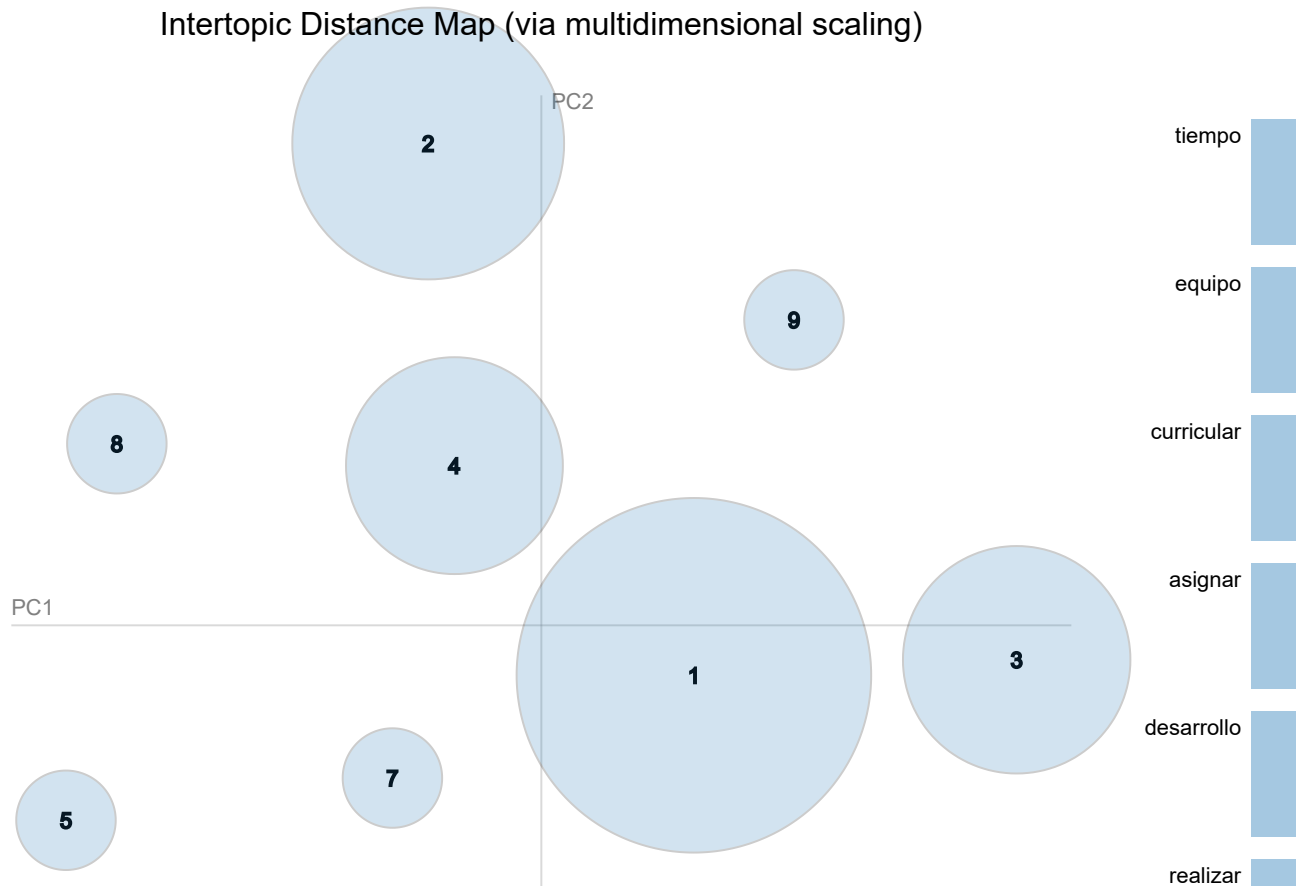
```

/usr/local/lib/python3.7/dist-packages/pyLDAvis/_prepare.py:247: FutureWarning: In a fut
by='saliency', ascending=False).head(R).drop('saliency', 1)
/usr/local/lib/python3.7/dist-packages/sklearn/manifold/_t_sne.py:783: FutureWarning: Th
FutureWarning,
/usr/local/lib/python3.7/dist-packages/sklearn/manifold/_t_sne.py:793: FutureWarning: Th
FutureWarning,
/usr/local/lib/python3.7/dist-packages/sklearn/manifold/_t_sne.py:827: FutureWarning: 's
FutureWarning,

```

Selected Topic:

Slid



¿La Coordinación Académica es un actor activo en la formulación de los proyectos formativos?

Marginal topic distribution

```

# Count Vectorizer
vect = CountVectorizer()
vects = vect.fit_transform(equ_ejr_text_1['P4'])
# Seleccione las primeras cinco filas del conjunto de datos
td = pd.DataFrame(vects.todense()).iloc[:5]
td.columns = vect.get_feature_names()
term_document_matrix = td.T
term_document_matrix.columns = ['Com '+str(i) for i in range(1, 6)]
term_document_matrix['total_count'] = term_document_matrix.sum(axis=1)

```

```
# Top 25 palabras
term_document_matrix = term_document_matrix.sort_values(by = 'total_count', ascending=False)[:1]

# Imprime las primeras 10 filas
print(term_document_matrix.drop(columns=['total_count']).head(20))
```

```
/usr/local/lib/python3.7/dist-packages/sklearn/utils/deprecation.py:87: FutureWarning: F
warnings.warn(msg, category=FutureWarning)
```

	Com 1	Com 2	Com 3	Com 4	Com 5
instructor	0	1	0	0	0
ocasión	1	0	0	0	0
considerable	1	0	0	0	0
realizado	0	1	0	0	0
formativo	1	0	0	0	0
proyecto	1	0	0	0	0
existente	1	0	0	0	0
formular	1	0	0	0	0
programa	0	0	0	0	1
bastante	1	0	0	0	0

▼ Histograma de palabras

```
plt.figure(figsize=(10,10))
term_document_matrix['total_count'].plot.bar()
plt.show()
```



▼ Latent Dirichlet Allocation - Modelo no supervisado de Tópicos.

| ■ ■ ■ ■ ■ ■ ■ ■ ■ ■ |

```
vectorizer_ = CountVectorizer(min_df=5, max_df=0.9, stop_words= stopwords_es, lowercase=True,
com_vectorized = vectorizer_.fit_transform(equ_ejr_text_1['P4'])
```

```
<>:1: DeprecationWarning: invalid escape sequence \-
<>:1: DeprecationWarning: invalid escape sequence \-
<>:1: DeprecationWarning: invalid escape sequence \-
<ipython-input-286-b846385e3eec>:1: DeprecationWarning: invalid escape sequence \-
vectorizer_ = CountVectorizer(min_df=5, max_df=0.9, stop_words= stopwords_es, lowercas
/usr/local/lib/python3.7/dist-packages/sklearn/feature_extraction/text.py:401: UserWarni
% sorted(inconsistent)
```

◀ | ■ ■ ■ ■ ■ ■ ■ ■ ■ ■ ▶

```
lda_ = LatentDirichletAllocation(n_components=10, max_iter=5, learning_method='online',verbo
com_lda = lda_.fit_transform(com_vectorized)
```

```
iteration: 1 of max_iter: 5
iteration: 2 of max_iter: 5
iteration: 3 of max_iter: 5
iteration: 4 of max_iter: 5
iteration: 5 of max_iter: 5
```

▼ keywords

```
def selected_topics(model, vectorizer, top_n=10):
    for idx, topic in enumerate(model.components_):
        print("Topic %d:" % (idx))
        print([(vectorizer.get_feature_names()[i], topic[i])
                for i in topic.argsort()[::-top_n - 1:-1]])

print("Modelo de comentarios LDA:")
selected_topics(lda_, vectorizer_)
```

Modelo de comentarios LDA:

```

Topic 0:
[('formulaci', 1.3718065213621262), ('instructor', 1.0712883042709322), ('equipo', 0.471
Topic 1:
[('equipo', 4.935961511363144), ('proyecto', 2.9195584385711353), ('instructor', 2.43234
Topic 2:
[('instructor', 1.0178238866944556), ('equipo', 1.0167695618816273), ('formativo', 0.601
Topic 3:
[('proyecto', 5.346051282881118), ('formativo', 4.273331012384828), ('coordinaci', 2.866
Topic 4:
[('formulaci', 0.8528950211235456), ('formativo', 0.46912708603407716), ('proyecto', 0.4
Topic 5:
[('proyecto', 0.5559513890881733), ('instructor', 0.5020745064349568), ('coordinaci', 0
Topic 6:
[('formativo', 0.5749917129577613), ('proyecto', 0.57288916180737), ('coordinaci', 0.555
Topic 7:
[('instructor', 0.4654144698504289), ('coordinaci', 0.4539348545485961), ('formativo', 0
Topic 8:
[('coordinaci', 1.61988565637658), ('formulaci', 1.0178180465617541), ('equipo', 0.46163
Topic 9:
[('proyecto', 3.6825197321041507), ('instructor', 1.6833309905751062), ('formativo', 0.5
/usr/local/lib/python3.7/dist-packages/sklearn/utils/deprecation.py:87: FutureWarning: F
warnings.warn(msg, category=FutureWarning)

```

```

pyLDAvis.enable_notebook()
dash = pyLDAvis.sklearn.prepare(lda_, com_vectorized, vectorizer_, mds='tsne')
dash

```

```
/usr/local/lib/python3.7/dist-packages/pyLDAvis/_prepare.py:247: FutureWarning: In a fut
by='saliency', ascending=False).head(R).drop('saliency', 1)
/usr/local/lib/python3.7/dist-packages/sklearn/manifold/_t_sne.py:783: FutureWarning: Th
FutureWarning,
/usr/local/lib/python3.7/dist-packages/sklearn/manifold/_t_sne.py:793: FutureWarning: Th
FutureWarning,
/usr/local/lib/python3.7/dist-packages/sklearn/manifold/_t_sne.py:827: FutureWarning: 's
FutureWarning,
```

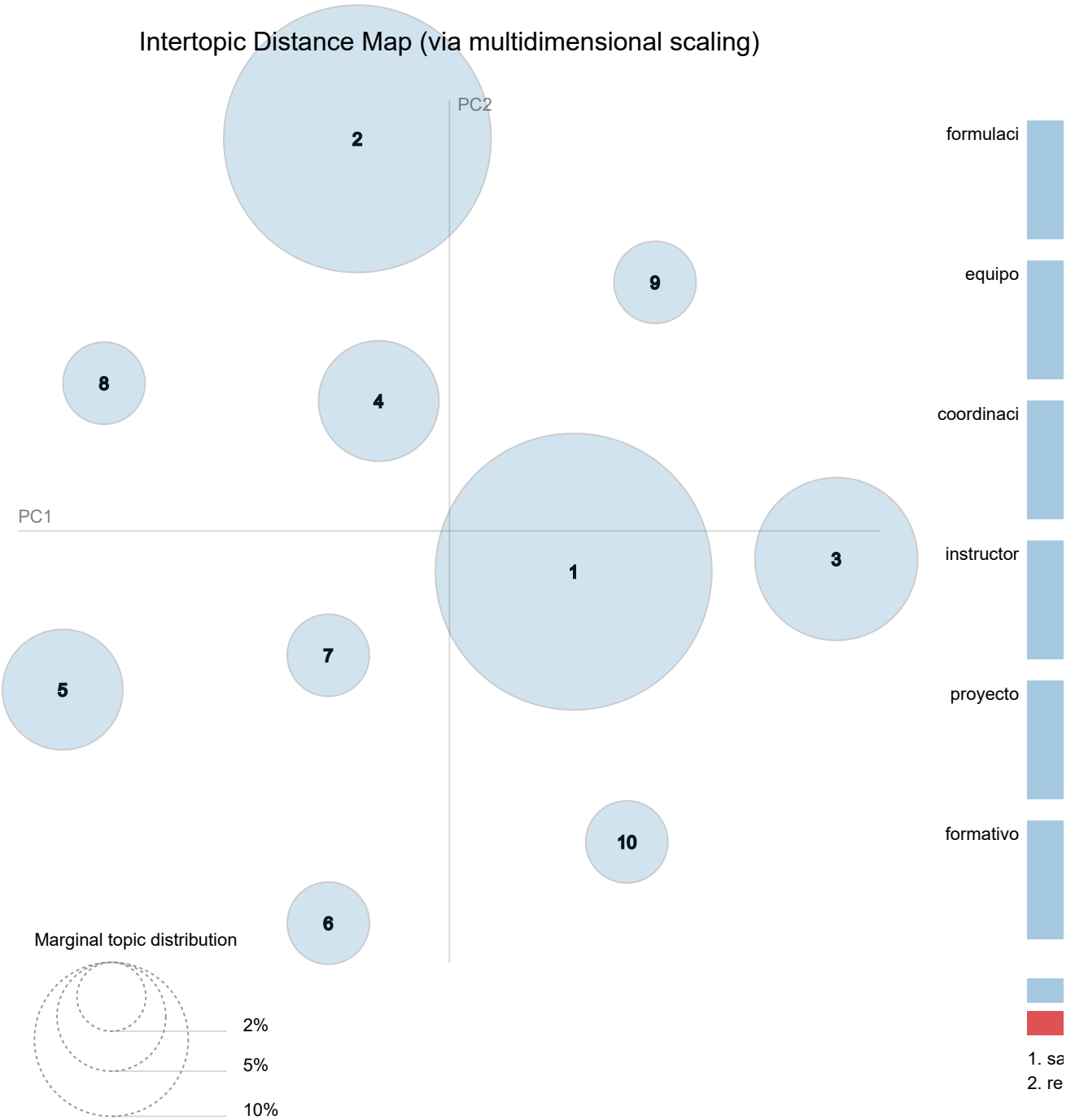
Selected Topic:

Previous Topic

Next Topic

Clear Topic

Slid





ANEXOS 6. Actas de reuniones.



ACTA No. 145		
NOMBRE DEL COMITÉ O DE LA REUNIÓN: Seguimiento Proyecto SGPS 9748 - Prototipo Innovación		
CIUDAD Y FECHA: Medellín, 02 de agosto de 2022	HORA INICIO: 10:40 a.m.	HORA FIN: 12:04 p.m.
LUGAR Y/O ENLACE: Laboratorio de Inteligencia Artificial Piso 5 https://bit.ly/3oC4ChB	DIRECCIÓN GENERAL / REGIONAL / CENTRO Centro de Servicios y Gestión Empresarial	
AGENDA O PUNTOS PARA DESARROLLAR: <ol style="list-style-type: none">1. Verificación del avance de los productos pendientes2. Sometimiento de artículos		
OBJETIVO(S) DE LA REUNIÓN: Realizar seguimiento a la ejecución de las actividades del Proyecto SGPS 9748 Prototipo Innovación.		
DESARROLLO DE LA REUNIÓN <ol style="list-style-type: none">1. Verificación del avance de los productos pendientes Se inicia la reunión comentando que el desarrollador John Ferla no trabajará esta semana en el prototipo debido a que en el prototipo AMMAS le entregaron el informe de los ajustes a realizar, luego de ser probado por Andrés Álzate, encargado del área ambiental del centro, en este sentido estará trabajando en esos ajustes para ya dedicarse de lleno a culminar con el desarrollo e integraciones pendientes en este prototipo. Por otro lado, la investigadora Paola Buitrago comentó que está trabajando en el informe de los códigos que se han trabajado durante la vigencia para los modelos de analítica del prototipo y cada código lo está documentando. Se hizo énfasis en que la mayor evidencia de la gestión es la entrega de los códigos y demás productos planteados en el proyecto. En cuanto al informe de investigación está en un 50% se entregará un borrador la próxima reunión con los espacios para la información del trabajo realizado por los compañeros John Ferla y Jesús Benavides, se aclaró que el 15 de agosto se debe el informe en su primer borrador.		



Se comentó que al investigador Jesús Arbey Benavides se le encargó la tarea de realizar el análisis de datos de la información de un trabajo de investigación que fue enviado desde la Regional Cauca y es un trabajo al que el investigador le dará prioridad ya que debe realizar una depuración y tratamiento de los datos para facilitar el posterior análisis, al tener el código generado se presentará el avance y se solicitará espacio para presentar el informe de analítica. Luego de generar el código el investigador Jesús Benavides retomará sus aportes al informe de investigación. También la investigadora Paola Buitrago propuso agregar un glosario de términos técnicos para facilitar la comprensión del informe de investigación.

En cuanto al manual de usuario y técnico a las investigadoras Paola Buitrago y Erika Restrepo se les encomendó la tarea de revisarlo y apoyar con los ajustes de redacción.

2. Sometimiento de artículos

Se realizó el sometimiento del artículo de la Revista Retos, se está a la espera de la respuesta.

La investigadora Erika Restrepo estuvo realizando el ajuste del formato del artículo que se envió al V Workshop on Modeling and Simulation for Science and Engineering (V WMSSE), en el que realizó ajustes en las especificaciones de forma, algunos detalles de redacción y en las referencias según el estilo solicitado por el evento.

También se envió el sometimiento de la propuesta del trabajo realizado en la vigencia anterior, solicitando ser tenidos en cuenta para la publicación en capítulo de libro en Expotecnológica 2022 del Pascual Bravo, donde se tendrá ponencia y posible publicación.

Se sugiere que los investigadores del prototipo Erika Restrepo, Paola Buitrago y Jesús Benavides, realicen sesiones para establecer una forma de realizar la transferencia y pulir el discurso, además Erika contará con el apoyo de Jesús Benavides en la parte técnica, así continúe como instructor se le dará en tiempo para apoyar en esta tarea de transferencia del prototipo.



Además, se socializó que en Redcolsi Regional se obtuvo una calificación de 91 puntos lo que da paso a participar en el encuentro Nacional, se gestionará un aprendiz en contrato de aprendizaje que pueda representar el prototipo en el evento.

Aunque no alcanzó el tiempo de los investigadores para explorar más técnicas de clusterización, para la predicción también hizo falta tiempo para poder realizar comparaciones entre métodos y realizar otro artículo. Esto es insumo para la convocatoria 2023 de SENNOVA. Sin embargo, se dejan muy buenos insumos y productos para continuar con el prototipo, se considera que hay tiempo suficiente para terminar la documentación del prototipo.

Para finalizar se mencionó que el equipo de trabajo alcanzó una muy buena sinergia y cohesión para el trabajo esto se ve reflejado en el cumplimiento de las actividades planteadas en los tiempos establecidos, se esperaba contar con presupuesto para continuar con las actividades, aunque no fue posible el resultado obtenido se considera satisfactorio.

Al enviar el certificado al líder del grupo, no fue aprobado inicialmente por quien hace las veces de líder del grupo; la investigadora Paola Buitrago realizará las gestiones que sean necesarias para que la participación en el evento sea aceptada y agregada a los productos del grupo de investigación.

CONCLUSIONES

Se está a la espera de que el desarrollador John Ferla continúe avanzando con la integración del aplicativo de recolección de información y los modelos de analítica de datos.

Se enviaron tres sometimientos a diferentes eventos, donde se espera tener una respuesta positiva frente a la publicación de los trabajos que se propusieron.

De acuerdo con La Ley 1581 de 2012, Protección de Datos Personales se debe garantizar la seguridad y protección de los datos personales que se encuentran almacenados en este documento. El Servicio Nacional de Aprendizaje SENA solicita la siguiente clasificación de la información:

La información de este documento se debe clasificar como:

PÚBLICA ☐ **PRIVADO** ☐ **SEMIPRIVADO** ☐ **SENSIBLE** ☐



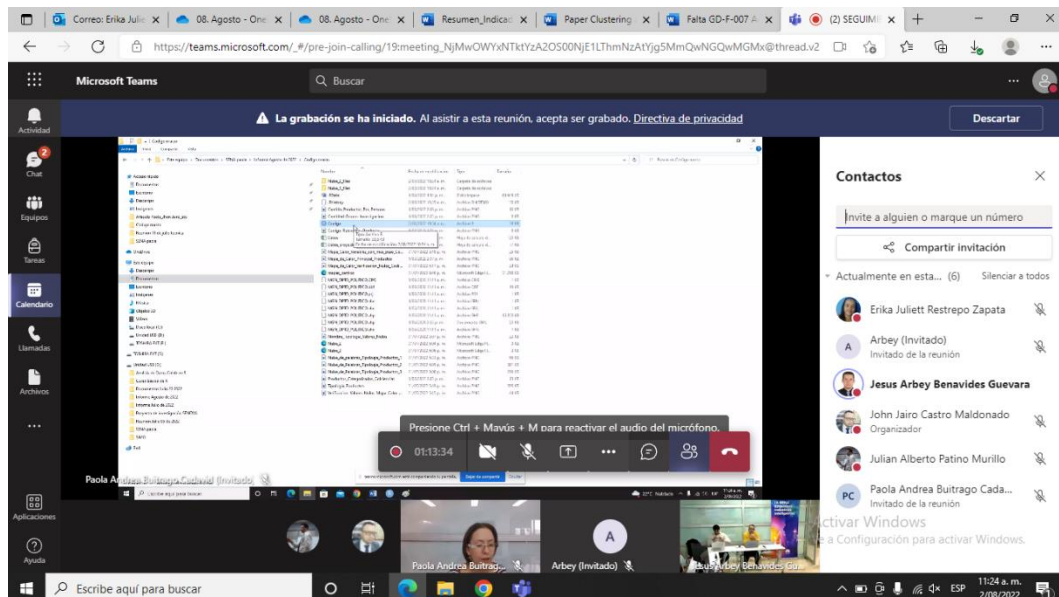
Nota: Antes de contestar esta informacional por favor remitirse al instructivo

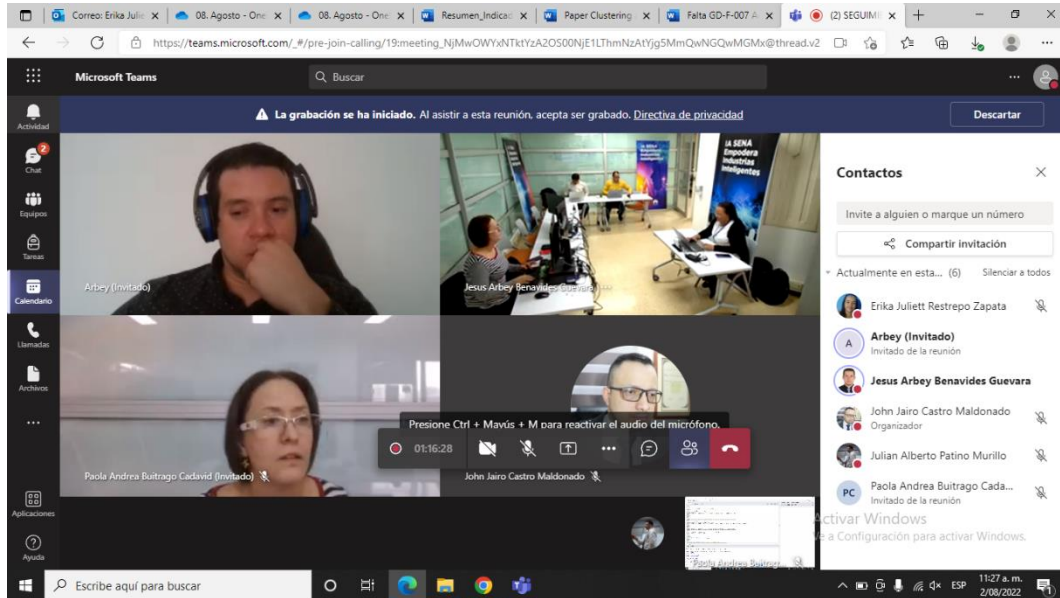
COMPROMISOS

ACTIVIDAD	RESPONSABLE	FECHA
Continuar apoyando el desarrollo y la elaboración de la documentación del prototipo.	Jesús Arbey Benavides Paola Andrea Buitrago	09 de agosto de 2022
Continuar integrando el modelo de analítica en el aplicativo.	John Ferla	09 de agosto de 2022
Continuar con las actividades de divulgación del prototipo.	Erika Restrepo	09 de agosto de 2022
Verificar con el líder del grupo de investigación el certificado de participación en el V Workshop on Modeling and Simulation for Science and Engineering (V WMSSE)	Paola Buitrago	09 de agosto de 2022

ASISTENTES: (Incorporar registro de asistencia)

Se anexa evidencia fotográfica de la reunión:





Nota: Se conectó de manera virtual Jesús Arbey Benavides

Enlaces de grabación:

[SEGUIMIENTO DE ACTIVIDADES DE PROYECTO DE INDICADORES DE CTEI CON ANALÍTICA DE DATOS-20220802_105047-Grabación de la reunión.mp4](#)



PROCESO GESTIÓN DEL TALENTO HUMANO
FORMATO INFORME MENSUAL EJECUCIÓN CONTRACTUAL

Medellín, 30 de agosto del 2022

Señor (a)
John Jairo Castro Maldonado
SUPERVISOR(A) CONTRATO No. CO1.PCCNTR.3367636
Instructor Grado 16
Centro de Servicios y Gestión Empresarial
Medellín

Asunto: Informe mensual de ejecución contractual
de agosto del año 2022.

Referencia: No 3367636 del año 2022

Jesus Arbey Benavides Guevara, identificado con la cédula de ciudadanía No. 1.085.260.713 de Pasto, en mi calidad de Contratista del SENA, en Centro de Servicios y Gestión Empresarial, en cumplimiento del Contrato de Prestación de Servicios de la referencia, a continuación presento el Informe de actividades realizadas en el mes objeto de cobro.

Valor y forma de Pago: Se fija como valor total para el contrato la suma de TREINTA Y CUATRO MILLONES TRESCIENTOS MIL PESOS MCTE (\$ 34.300.000) Incluido IVA. Esta suma será pagada por el SENA al contratista de la siguiente manera: Valor por mes: CUATRO MILLONES NOVECIENTOS MIL PESOS MCTE (\$ 4.900.000), y el equivalente por fracción de mes.

Plazo: Será hasta el treinta (30) de agosto (08) de 2022.

OBJETO: Prestar servicios profesionales para el desarrollo de las actividades de levantamiento de indicadores, bases de datos y el modelo matemático con base en analítica de datos y Big Data en el prototipo de software para modelo de Innovación en el desarrollo del proyecto de I+D+i SGPS 9748

Obligaciones Específicas:

No	Obligaciones	Acciones realizadas	Evidencias
1	Apoyar el desarrollo de aplicaciones o algoritmos usados en la actualidad para solucionar de forma eficiente y eficaz problemas o retos identificados	Acompañamiento de implementación de mapas geoespaciales en la Interfaz de programación de aplicaciones (API), de los resultados de conglomerados generados por Kmeans y PCA para	ANEXOS 6. Actas de reuniones. ANEXO 10. Reporte de códigos para compatibilidad de API con resultado analítico.



	en Big Data y analítica de datos con aplicación en gestión de la innovación	los centros de formación de SENA Antioquia, con sus clasificaciones de alta, media y baja productividad y gestión de manera respectiva.	
2	Revisar las herramientas tecnológicas o digitales de Big Data y analítica de datos necesarias para dar cumplimiento a los objetivos propuestos dentro del proyecto.	Gestión realizada	<p>ANEXO 7. Participación en reuniones con equipo para el desarrollo del prototipo de software.</p> <p>ANEXO 9. Reporte de códigos para selección de modelos de analítica de datos.</p> <p>ANEXO 10. Reporte de códigos para compatibilidad de API con resultado analítico.</p>
3	Revisar y actualizar indicadores para la gestión de la innovación según objetivos del proyecto	Gestión realizada	<p>ANEXO 7. Participación en reuniones con equipo para el desarrollo del prototipo de software.</p> <p>ANEXO 9. Reporte de códigos para selección de modelos de analítica de datos.</p>
4	Realizar pruebas relacionadas con la compatibilidad de la tecnología y la aplicación de los sistemas de Big Data y analítica de datos para la evaluación y valoración de la Gestión de CTel para los Centros SENA Regional Antioquia	Gestión realizada	<p>ANEXO 7. Participación en reuniones con equipo para el desarrollo del prototipo de software.</p> <p>ANEXO 10. Reporte de códigos para compatibilidad de API con resultado analítico.</p>



5	Apoyar la elaboración del manual del sistema y usuario y las actividades de transferencia	Gestión realizada	ANEXO 10. Reporte de códigos para compatibilidad de API con resultado analítico.
6	Presentar los informes de seguimiento del proyecto de acuerdo con las solicitudes del líder del área o supervisor	Se elaboró el informe de pago mensual GTH-F-062 V09 Se realizó la redacción de informe de investigación resultados del proyecto.	Informe en formato pdf. ANEXOS 2. Evidencia de informe de investigación
7	Divulgar en eventos académicos/empresariales los resultados del proyecto.	Se recibió certificación de participación en Workshop on Modeling and Simulation for Science and Engineering, en modalidad de Artículo. Se realizó acompañamiento de modificación de autor de correspondencia y modificaciones a ser realizadas en artículo publicado en Workshop on Modeling and Simulation for Science and Engineering.	ANEXO 1. Medios de divulgación: Notificación de cambio de autor de correspondencia en Workshop on Modeling and Simulation for Science and Engineering ANEXO 8. Correos asociados a Proceso editorial artículos [V WMSSE-JPCS]
8	Redactar el informe de investigación y de resultados del proyecto.	Se elaboró el informe de pago mensual GTH-F-062 V09 Se realizó la redacción de informe de investigación resultados del proyecto.	Informe en formato pdf. ANEXOS 2. Evidencia de informe de investigación
9	Redactar y someter a publicación artículo de investigación, libro	Gestión realizada	ANEXO 1. Medios de divulgación: Notificación de cambio de autor de correspondencia en Workshop on



	de investigación o capítulo de libro de investigación relacionado con el proyecto, en medio reconocido por Minciencias.		Modeling and Simulation for Science and Engineering ANEXO 8. Correos asociados a Proceso editorial artículos [V WMSSE-JPCS]
10	Apoyar actividades de Grupo de Investigación, Semilleros de investigación y SENNOVA que sean propias de su área de conocimiento y en relación con el objeto contractual	Se realizó analítica de datos de 5 encuestas en apoyo a la instructora Luz Carime Chia del Centro de Comercio y Servicios- SENA Cauca, de la siguiente manera: 1. Analítica de datos CUESTIONARIO 1- Aprendices. 2. Analítica de datos CUESTIONARIO 2 – Instructores 3. Analítica de datos _ENCUESTA_EQ UIPO PEDAGÓGICO 4. Analítica de datos _ENCUESTA_EQ UIPO_EJECUTOR	ANEXO 4. Analítica de datos – Encuestas. Soporte de correos, entrega de resultados en código. ANEXOS 5. Códigos de apoyo en analítica de datos - Centro de Comercio y Servicios- SENA Cauca Link de códigos y datos: https://bit.ly/3K7IKFS
11	Apoyar la evaluación técnica, administrativa y presupuestal de los proyectos de investigación aplicada, desarrollo tecnológico e innovación que se formulan por el SENA para el desarrollo de actividades misionales, acorde con las	Se realizó taller de apoyo a aprendices del “Curso de estadística aplicada a inteligencia artificial” para la apropiación de conceptos de estadística descriptiva.	ANEXOS 3. Apoyo a la formación - Curso de estadística aplicada a inteligencia artificial.



	indicaciones dadas por la Dirección General.		
12	Desarrollar sus actuaciones en el marco de la normatividad vigente y presentar los informes mensuales de acuerdo con el objeto del presente contrato.	Se elaboró un (1) informe de pago mensual GTH-F-062 V09.	Informe en formato pdf.

A continuación, relaciono los desplazamientos que realicé previo a la presentación de este informe. Una vez finalizado cada desplazamiento presenté al ordenador del gasto el informe en el Formato Informe Legalización Desplazamiento Contratista GTH-F-087, en el que se describieron las actividades desarrolladas y los resultados de cada desplazamiento. Cada informe cuenta con el visto bueno del Supervisor.

Se lista a continuación el soporte de la legalización de los desplazamientos realizados, los cuales forman parte integral del presente informe de ejecución contractual.

ITEM	No DE LA ORDEN DE VIAJE	LUGAR DE DESPLAZAMIENTO	FECHAS DE DESPLAZAMIENTOS	
			INICIO	FINALIZACIÓN
1.		No he realizado ningún desplazamiento		

Nota 1: Por cada desplazamiento que haya realizado el contratista, adjuntará el respectivo informe que la soporte. En caso de haber realizado el desplazamiento en fecha posterior a la presentación del informe de ejecución contractual, deberá reportarlo en el siguiente informe de ejecución contractual.

Para el trámite de la cuenta me permito adjuntar: Documentos electrónicos enunciados como evidencias del cumplimiento de las obligaciones contractuales y los desplazamientos realizados y el No. 60626908 de la planilla, miplanilla correspondiente al mes de julio del año 2022. (Decreto Ley 2106 de 2019 – “Decreto Ley Antitrámites”)

Declaración bajo la gravedad de juramento.

Me permito declarar **bajo la gravedad de juramento** que durante el periodo objeto de cobro SI___NO_X_ he suscrito otro u otros contratos con entidad(es) del Estado el (los) cual(es) se



encuentra(n) en ejecución; la anterior declaración de acuerdo con lo establecido en el numeral 5º de la Directiva Presidencial No. 001 de 2022, a fin de identificar potenciales conflictos de interés y hacer la gestión del riesgos para la toma de decisiones dentro del principio de la debida diligencia frente a la ejecución de los contratos de prestación de servicios profesionales o de apoyo a la gestión.

Nota: En el evento de señalar **SI**, el análisis efectuado por el supervisor y ordenador de pago deberá quedar documentado y publicado en el SECOP II en los documentos relacionados en el archivo Gestión Contractual GC_ del periodo objeto de cobro.

Evidencias en folios:

- ANEXOS.
- https://drive.google.com/drive/folders/1yEDbPoaNjp_LMjKuMufA0nyHAufxe9iC?usp=sharing

Cordialmente,

JESUS ARBEY BENAVIDES GUEVARA
Contratista
C.C. No. 1.085.260.713 DE PASTO (NARIÑO)

Recibí a satisfacción:

JOHN JAIRO CASTRO MALDONADO
SUPERVISOR(A) CONTRATO No. CO1.PCCNTR.3367636 DE 2022
INSTRUCTOR GRADO 16
C.C. No. 88.265.552



ANEXOS



ANEXO 1. Medios de divulgación: Notificación de cambio de autor de correspondencia en *Workshop on Modeling and Simulation for Science and Engineering*

Solicitud para notificaciones de publicación Abstract Id: 20

Editor

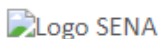
Ely Dannier V. Niño

Le escribo para solicitarle que las notificaciones como autor correspondiente sean enviadas al correo paola.buitrago.betancur@gmail.com, toda vez que mi cuenta de correo pabuitragoc@sena.edu.co expira esta semana ya que he terminado contrato como investigadora de la institución.

Envío copia del correo al grupo de investigadores co-autores del artículo con Abstract Id: 20.

Saludos cordiales,

--



Paola Andrea Buitrago Cadavid

Investigador Junior (IJ)

[https://scienti.minciencias.gov.co/cvlac/visualizador/generarCurriculoCv.do?](https://scienti.minciencias.gov.co/cvlac/visualizador/generarCurriculoCv.do?cod_rh=0001487730)

[cod_rh=0001487730](https://scienti.minciencias.gov.co/cvlac/visualizador/generarCurriculoCv.do?cod_rh=0001487730)

Ingeniera Electrónica

Universidad de Antioquia

Magister en Modelación y Ciencia Computacional

Universidad de Medellín

Investigadora Sennova - SENA

Centro De Servicios y Gestión Empresarial - Complejo Central - SENA

pabuitragoc@sena.edu.co

paola.buitrago.betancur@gmail.com

+573113671040

Calle 51 Nro 57-70 - Complejo Central

Figura A1.1. Solicitud de modificación de autor de correspondencia por parte de la investigadora Paola Buitrago, V *Workshop on Modeling and Simulation for Science and Engineering*. Proyecto: “Clustering and geospatial visualization of innovation indicators in the Servicio Nacional de Aprendizaje Regional Antioquia, Colombia.”

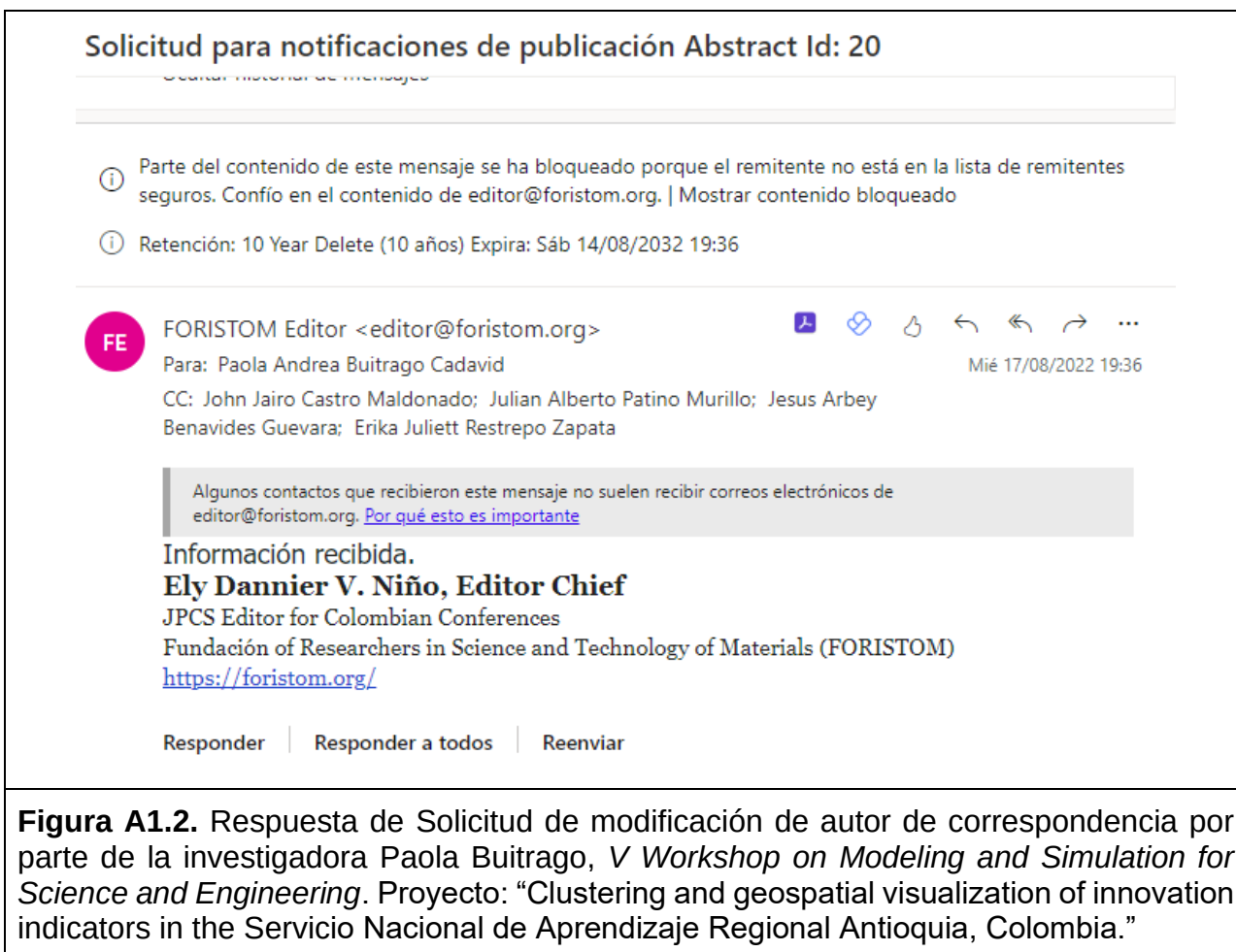


Figura A1.2. Respuesta de Solicitud de modificación de autor de correspondencia por parte de la investigadora Paola Buitrago, *V Workshop on Modeling and Simulation for Science and Engineering*. Proyecto: “Clustering and geospatial visualization of innovation indicators in the Servicio Nacional de Aprendizaje Regional Antioquia, Colombia.”

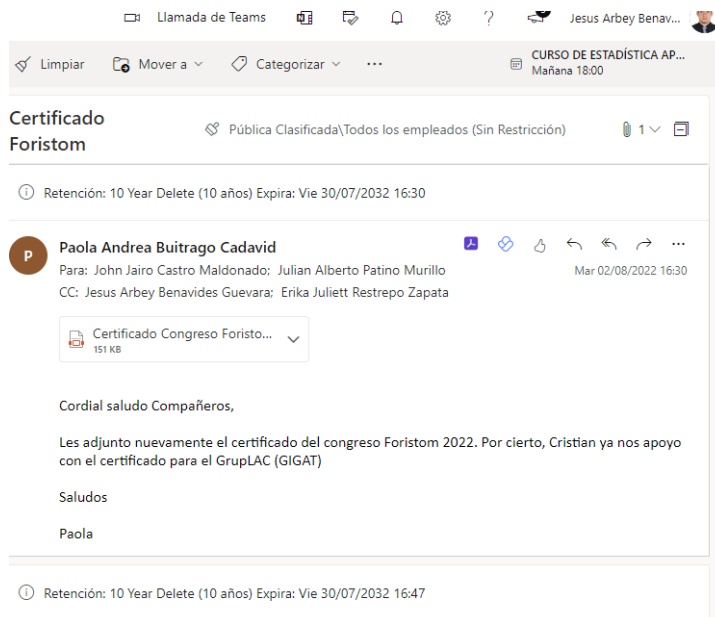


Figura A1.3. Recepción de certificación de participación *V Workshop on Modeling and Simulation for Science and Engineering*. Proyecto: “Clustering and geospatial visualization of innovation indicators in the Servicio Nacional de Aprendizaje Regional Antioquia, Colombia.”

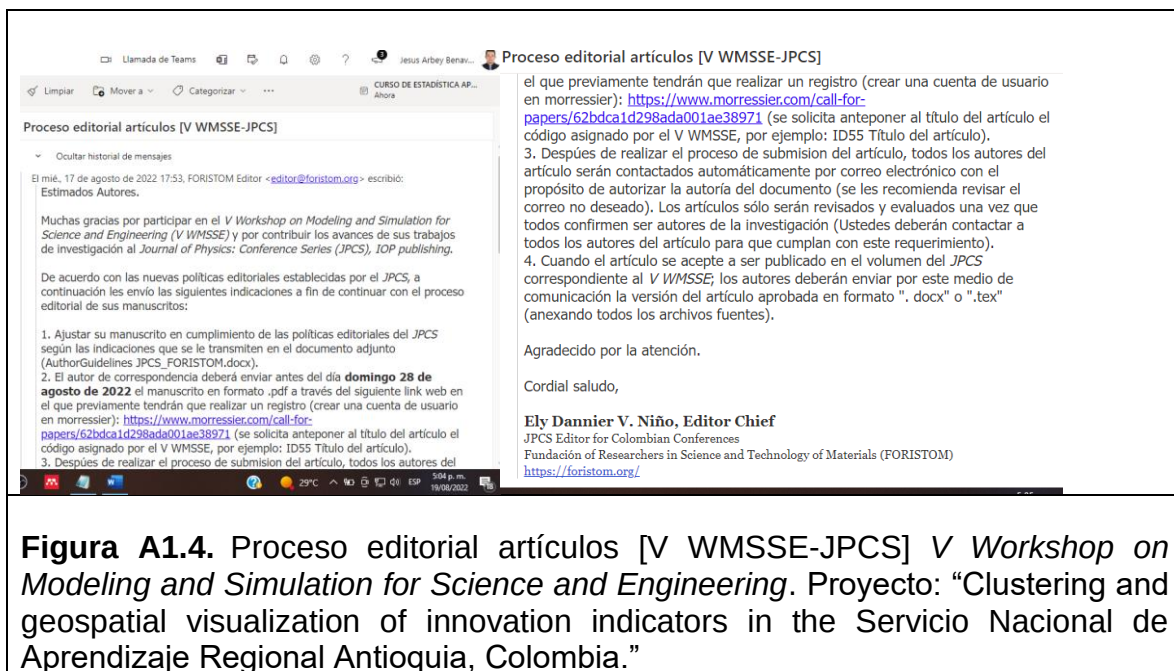


Figura A1.4. Proceso editorial artículos [V WMSSE-JPCS] *V Workshop on Modeling and Simulation for Science and Engineering*. Proyecto: “Clustering and geospatial visualization of innovation indicators in the Servicio Nacional de Aprendizaje Regional Antioquia, Colombia.”

ANEXOS 2. Evidencia de informe de investigación

**Modelo de evaluación y valoración de la Gestión de CTel para los Centros SENA Regional
Antioquia aplicando analítica de datos - Proyecto SGPS 9748 - 2021**

Informe técnico de investigación

Presentado por

Grupo de Investigación en Gerencia y Aplicación de la Ciencia y Tecnología - GIGAT

Paola Buitrago Cadavid

Jesus Arbey Benavides Guevara

John Jairo Castro Maldonado

Julián Alberto Patiño Murillo

Erika Juliett Restrepo Zapata

Jhon Alexis Ferla Pérez



Centro de Servicios y Gestión

Empresarial

Sena, regional Antioquia

Medellín

2022



1 ¡Error! Marcador no definido.2 ¡Error! Marcador no definido.3 ¡Error! Marcador no
definido.4 ¡Error! Marcador no definido.5 ¡Error! Marcador no definido.6 ¡Error! Marcador
no definido.7 ¡Error! Marcador no definido.7.1 ¡Error! Marcador no definido.7.2
¡Error! Marcador no definido.8 ¡Error! Marcador no definido.9 ¡Error! Marcador no
definido.9.1 ¡Error! Marcador no definido.9.2 ¡Error! Marcador no definido.10
¡Error! Marcador no definido.10.1 ¡Error! Marcador no definido.10.2 ¡Error!
Marcador no definido.10.2.1 3010.3 3110.3.1 ¡Error! Marcador no definido.10.3.2
¡Error! Marcador no definido.10.4 ¡Error! Marcador no definido.10.5 ¡Error!
Marcador no definido.10.5.1 ¡Error! Marcador no definido.10.6 ¡Error! Marcador no
definido.10.6.1¡Error! Marcador no definido.10.6.2 ¡Error! Marcador no definido.10.6.3
¡Error! Marcador no definido.10.7 ¡Error! Marcador no definido.10.7.1 ¡Error!
Marcador no definido.10.7.2 ¡Error! Marcador no definido.10.7.3 ¡Error! Marcador no
definido.11 ¡Error! Marcador no definido.12 ¡Error! Marcador no definido.12.1
¡Error! Marcador no definido.12.1.1 ¡Error! Marcador no definido.12.2 ¡Error!
Marcador no definido.12.3 ¡Error! Marcador no definido.13 ¡Error! Marcador no
definido.13.1 ¡Error! Marcador no definido.13.1.1 ¡Error! Marcador no definido.13.2
¡Error! Marcador no definido.13.3 ¡Error! Marcador no definido.13.3.1 ¡Error!
Marcador no definido.13.3.2 ¡Error! Marcador no definido.13.3.3 ¡Error! Marcador no
definido.13.3.4¡Error! Marcador no definido.13.3.5 ¡Error! Marcador no definido.13.3.6
¡Error! Marcador no definido.13.3.7 ¡Error! Marcador no definido.13.4 ¡Error!
Marcador no definido.13.4.1 ¡Error! Marcador no definido.13.5 ¡Error! Marcador no
definido.13.5.1¡Error! Marcador no definido.13.5.2 ¡Error! Marcador no definido.13.5.3
¡Error! Marcador no definido.13.5.4 ¡Error! Marcador no definido.13.5.4.1 ¡Error!
Marcador no definido.13.5.4.2 ¡Error! Marcador no definido.13.5.4.3 ¡Error! Marcador no
definido.13.5.4.4 ¡Error! Marcador no definido.13.5.4.5 ¡Error! Marcador no
definido.13.5.4.6 ¡Error! Marcador no definido.13.5.5 ¡Error! Marcador no
definido.13.5.5.1 ¡Error! Marcador no definido.14 ¡Error! Marcador no
definido.14.1 ¡Error! Marcador no definido.14.2 ¡Error! Marcador no definido.15
¡Error! Marcador no definido.15.1 ¡Error! Marcador no definido.15.2 ¡Error!
Marcador no definido.15.3 ¡Error! Marcador no definido.16 ¡Error! Marcador no
definido.17 ¡Error! Marcador no definido.18 ¡Error! Marcador no definido.18.1.1
¡Error! Marcador no definido.18.1.1.1 ¡Error! Marcador no definido.



1 DATOS DEL PROYECTO

Nombre del Proyecto:

Modelo de evaluación y valoración de la Gestión de CTel para los Centros SENA Regional Antioquia aplicando analítica de datos - Proyecto SGPS 9748 - 2021

Autores

Paola Buitrago Cadavid, Ingeniería Electrónica, Magister en Modelación y Ciencia Computacional, Investigadora SENNOVA. Email: pabuitragoc@sena.edu.co

Jesus Arbey Benavides Guevara, Físico, Magíster en Física. Doctorado(candidato) en Física, Especialista en ciencias de datos(Estudiente). Investigador SENNOVA. Email: jabenavides@sena.edu.co

Erika Juliett Restrepo Zapata, Administradora de Empresas, Tecnóloga en Formulación de Proyectos. Investigador SENNOVA, Email: erestrepoz@sena.edu.co

Julián Alberto Patiño, Ingeniero de Control, Magíster en Ingeniería Automatización Industrial, Doctor en Ingeniería Automática, Email: japatino@sena.edu.co

John Jairo Castro Maldonado, Ingeniero Electromecánico, Especialización en Práctica Pedagógica Universitaria, Especialización en Seguridad Informática, Maestría en Ciencia y Tecnología de Materiales, Maestría en Controles Industriales, Email: jcastrom@sena.edu.co

Nombre de semilleros Involucrados

Semillero de Investigación en Proyectos e Ingeniería (SIPI)

Semillero de Investigación en Móviles,
y Control (MERLIN)



Educación, Robótica, Automatización

Línea de Investigación

Innovación

Aliados

Unidad de Servicios Tecnológicos del Centro de Servicios y Gestión Empresarial

Tecnoparque

Nombre del grupo de investigación

GIGAT (Grupo de Investigación en Gerencia y Aplicación de la Ciencia y la Tecnología)

Fuente de financiación

Convocatoria SENNOVA 2021 – Resolución 0012020

Duración del proyecto

7 meses

2 LISTA DE ANEXOS



Anexo 1. Implementación de código en R statistics para generación de base de datos	90
Anexo 2. Implementación de código en R statistics del método de clustering	91
Anexo 3. Implementación de código en R statistics del método K-means	92
Anexo 4. Implementación de código en R statistics del método PCA	94
Anexo 5. Implementación de código en R statistics del método Random Forest	96
Anexo 6. Implementación de código en R statistics del método K-NN	98
Anexo 7. Implementación de código en Python del método de clustering	90
Anexo 8. Implementación de código en Python del método K-means	91
Anexo 9. Implementación de código en Python del método PCA	92
Anexo 10. Implementación de código en Python de una red neuronal tipo perceptrón	94
Anexo 11. Implementación de código en Python del método de regresión logística	96
Anexo 12. Implementación de código en Python del método SVM	98

3 LISTA DE TABLAS

Tabla 1. Variables identificadas que intervienen en la medición de indicadores	19
---	-----------

**Tabla 2. Variables identificadas que
procesos**

19



intervienen en el módulo de

Tabla 3. Variables de salida

20

Tabla 4. Identificación de tablas y campos para la base de datos

25

Tabla 5. Identificación de campos para la tabla de datos Procesos

27

Tabla 6. Identificación de campos para la tabla de datos Productividad

27

4 LISTADO DE FIGURAS

Figura 1. Diagrama representativo de clúster

28

Figura 2. Iteración 2. Elaboración con base en los resultados de VOSviewer

22

Figura 3. Clasificación Actividades de Ciencia y Tecnología OCDE.

24

Figura 4. Tipos de análisis. Fuente:(Riahi &  Riahi, 2018)	40
Figura 5. Etapas del proceso de investigación	49
Figura 6. Etapas del proyecto	51
Figura 7. Ciclo de vida del proyecto. Elaboración propia	54
Figura 8. Modelo Conceptual de la evaluación de la gestión de ciencia, tecnología e innovación	58
Figura 9. Componentes del Sistema de Información de Proyectos de CTel	66
Figura 10. Dependencia de las variables de los proyectos de CTel	68
Figura 11. Oportunidades de captura de la información del Sistema de Proyectos de CTel	69
Figura 12. Flujo de proceso de la metodología CRISP® Tomado de (García Vélez, 2018)	70



- Acervo de recursos humanos en ciencia y tecnología, ARHCyT (Manual de Canberra, OECD) (C. Manual, 1995) Personas que cumplen con una de las siguientes condiciones: a) haber completado exitosamente educación de tercer nivel en un campo de la ciencia y la tecnología, b) personas no tan formalmente calificadas como lo anterior, pero que están empleadas en una ocupación de ciencia y tecnología que generalmente requiere de habilidades y conocimientos de tercer nivel.
- Activos (Miller y Meiners) (Miller et al., 1990) Cualquier bien que proporciona un flujo de utilidad y/o ingreso en el futuro. Los activos incluyen acciones, bonos, automóviles, estéreos, un grado universitario, etc.
- Actividades de innovación tecnológica (Manual de Frascati, OECD) (F. Manual, 2002) Conjunto de etapas científicas, tecnológicas, organizativas, financieras y comerciales, incluyendo las inversiones en nuevos conocimientos, que llevan o que intentan llevar a la implementación de productos y de procesos nuevos o mejorados. La I+D no es más que una de estas actividades y puede ser llevada a cabo en diferentes fases del proceso de innovación, siendo utilizada no sólo como la fuente de ideas creadoras sino también para resolver los problemas que pueden surgir en cualquier fase hasta su culminación.
- Ciencia (Real Academia Española) (GONZALEZ et al., 2013) Conjunto de conocimientos obtenidos mediante la observación y el razonamiento, sistemáticamente estructurados y de los que se deducen principios y leyes generales. En otra acepción, ciencia es el conjunto de conocimientos relativos a las ciencias exactas, fisicoquímicas y naturales.
- Costes salariales del personal de I+D (Manual de Frascati, OECD) (F. Manual, 2002) Estos costes comprenden los salarios y remuneraciones anuales y todos los gastos complementarios de personal o remuneraciones diversas, tales como primas, vacaciones pagadas, contribuciones a fondos de pensiones y otros pagos a la Seguridad Social, impuestos salariales, etc. Los costes salariales de las personas que prestan servicios indirectos y que no se tienen en cuenta en los datos de personal (tales como el personal de seguridad y de mantenimiento o el personal de bibliotecas centrales, de servicios informáticos y de las oficinas de dirección) deben excluirse y contabilizarse en el apartado de otros gastos corrientes.



- Desarrollo experimental (Manual de Frascati, OECD) (F. Manual, 2002)
El desarrollo experimental consiste en trabajos sistemáticos fundamentados en los conocimientos existentes obtenidos por la investigación o la experiencia práctica, que se dirigen a la fabricación de nuevos materiales, productos o dispositivos, a establecer nuevos procedimientos, sistemas y servicios, o a mejorar considerablemente los que ya existen. En las ciencias sociales, el desarrollo experimental puede definirse como el proceso que permite convertir los conocimientos adquiridos a través de la investigación en programas operativos, incluidos los proyectos de demostración que se llevan a cabo con fines de ensayo y evaluación. Esta categoría tiene escasa o nula significación en el caso de las humanidades.
- Investigación y desarrollo experimental, I+D (Manual de Frascati, OECD) (F. Manual, 2002) La investigación y el desarrollo experimental comprenden el trabajo creativo llevado a cabo de forma sistemática para incrementar el volumen de conocimientos, incluido el conocimiento del hombre, la cultura y la sociedad, y el uso de esos conocimientos para crear nuevas aplicaciones. Comprende investigación básica, aplicada y el desarrollo experimental.
- Investigadores (Manual de Frascati, OECD) (F. Manual, 2002) Profesionales que se dedican a la concepción o creación de nuevos conocimientos, productos, procesos, métodos y sistemas y también a la gestión de los proyectos respectivos. Los investigadores se encuentran incluidos en el Grupo Principal 2 de la clasificación ISCO-88, Profesionales. Por convención, también se incluyen en estas categorías los miembros de las fuerzas armadas con formación equivalente y que realicen actividades de I+D. Igualmente, están incluidos en esta categoría los gestores y administradores que desarrollan actividades de planificación y gestión de los aspectos científicos y técnicos del trabajo de los investigadores. Normalmente, tienen una categoría igual o superior a la de las personas empleadas directamente como investigadores, tratándose a menudo de investigadores veteranos o a tiempo parcial. Los títulos profesionales pueden variar de una institución a otra, de un sector a otro y de un país a otro. Los estudiantes de postgrado a nivel de doctorado que participan en tareas de I+D deben considerarse como investigadores. Habitualmente suelen poseer titulaciones universitarias básicas (CINE/ISCED nivel 5A) y realizan trabajos de investigación a la vez que preparan su doctorado (CINE/ISCED nivel 6).



- Personal de I+D (Manual de Frascati, OECD) (F. Manual, 2002)
Todo el personal empleado directamente en I+D, así como las personas que proporcionan servicios directamente relacionados con actividades de I+D, como los directores, administradores y personal de oficina. Las personas que proporcionan servicios indirectos, como el personal de los comedores y de seguridad, deben ser excluidas, aunque sus sueldos y salarios se contabilicen como gastos generales para la medida del gasto de I+D.

6 RESUMEN DEL PROYECTO

Las actividades de Ciencia, Tecnología e Innovación están relacionadas con la ejecución de actuaciones que impliquen investigación, desarrollo experimental, apoyo a la educación y formación y prestación de servicios científicos y tecnológicos, actividades de innovación, servicios científicos y tecnológicos, administración y otras actividades de gestión, que son realizadas por entidades públicas y entidades privadas para contribuir a la generación y apropiación del conocimiento.

En este sentido, debido al incremento exponencial que ha tenido la generación de nuevos productos de conocimiento a nivel educativo y con el fin de responder a una necesidad evidente de contribuir al desarrollo del sector productivo en Colombia, a través del Sistema de Investigación SENNOVA del SENA Se potencian tanto los recursos humanos como los tecnológicos con los que cuenta la entidad para contribuir al crecimiento económico y social de la región. Teniendo en cuenta lo anterior, en la región de Antioquia del SENA se trabaja en el desarrollo de la visualización de un modelo de analítica de datos para la medición de indicadores de innovación, donde se identifican gráficamente las diferentes categorías de actividades de Ciencia, Tecnología e Innovación que vienen trabajando en los últimos años en la entidad, con el propósito de facilitar la medición y análisis de los resultados obtenidos desde la perspectiva tanto de la estadística descriptiva como de la ciencia de datos. Entre los indicadores utilizados se encuentran los procesos de gestión y productividad de los proyectos ejecutados en la entidad en un periodo de tiempo, a los cuales se les aplicaron técnicas como K-means grouping y el Análisis de Componentes Principales (PCA), que permitieron agrupar los indicadores seleccionados. Se utilizó Python Folium como una herramienta de visualización que permite generar comparaciones de manera gráfica entre los diferentes centros del SENA de la Regional de Antioquia en un período de tiempo determinado. Posteriormente, se desarrolló una plataforma API que permite visualizar la

estadística descriptiva de los datos  sintéticos, así como los resultados analíticos representados en mapas y valores predictivos usando KNN como modelo predictivo en dimensiones de productividad y procesos. Se observó que, a través de las técnicas utilizadas, existe una mayor facilidad de interpretación, comparación y análisis de resultados estadísticos y análisis de datos en la medición de indicadores de actividades de Ciencia, Tecnología e Innovación en comparación con las herramientas de visualización de información utilizadas convencionalmente.

Palabras clave: Clustering, Folium, KPIs, K-means, PCA, Random forest, K-NN, ACTI, SENA

7 OBJETIVOS

7.1 Objetivo general:

Proponer un Modelo de Gestión de Indicadores de Ciencia, Tecnología e Innovación para los Centros SENA Regional Antioquia a partir de técnicas de analítica de datos.

7.2 Objetivos específicos:

1. Identificar metodologías de analítica de datos asociada al establecimiento de indicadores y modelos de CTel que puedan aplicarse a los centros de formación del SENA.
2. Determinar un modelo de evaluación y valoración de las actividades de CTel innovación que más se ajuste al comportamiento de estas sinergias en los centros de formación de la regional Antioquia.
3. Validar el modelo por medio de datos sintéticos que permitan generar una metodología para la analítica de datos.
4. Crear modelos de analítica de datos a partir de aprendizaje de máquina supervisado y no supervisado para la determinación de indicadores de CTel.



5. Generar mapas geoespaciales de alertas de alto, medio y bajo de las dimensiones de indicadores establecidos.

6. Implementar modelos predictivos a partir de aprendizaje supervisado para la determinación de estados de alerta de alto medio y bajo de las dimensiones de indicadores, a lo largo del tiempo.

8 INTRODUCCIÓN

Comenzamos definiendo las actividades de Ciencia, Tecnología e innovación (ACTI) como aquellas que se relacionan con producir, promover, difundir y aplicar conocimientos científicos en los campos que involucran la ciencia y la tecnología [1]. En los últimos años ha surgido la necesidad de proporcionar de forma sistemática información sobre la recolección, análisis, publicación y normalización de los datos asociados a estas actividades [2], de allí que se empiecen a diseñar indicadores que aporten a la toma de decisiones estratégicas sobre su gestión.

A nivel internacional la Organización para la Cooperación y el Desarrollo Económico (OCDE) ha elaborado diferentes manuales con lineamientos que pueden ser utilizados para la definición de indicadores que permitan la medición de ACTI [3], como el Manual de Frascati y el Manual de Oslo, que compuestos por una serie de propuestas para la recolección e interpretación de datos sobre innovación tecnológica, por otro lado, el Manual de Bogotá que contiene pautas para construir indicadores de innovación tecnológica en Latinoamérica y el Caribe, entre otros [4].

Esta investigación tiene como objetivo proponer un método de medición de Indicadores de Ciencia, Tecnología e Innovación mediante analítica de datos, la fundamentación teórica de la propuesta se basó en técnicas estadísticas no paramétricas supervisadas, entre las que encontramos métodos como clustering basado en algoritmos k-means y el uso de análisis de componentes principales (PCA). Para ello se simuló una base de datos sintéticos de 500 proyectos distintos con el propósito de realizar un análisis exploratorio que permitieron caracterizar los 16 centros del SENA de la regional Antioquia, de los cuales posteriormente se crean mapas utilizando como herramienta de base Folium [5].

Uno de los indicadores de mayor  preponderancia es el de gasto y/o inversión en actividades de investigación y desarrollo, I+D, ya que permite hacer seguimiento a las actividades de ciencia y tecnología – ACTI mediante el apalancamiento de proyectos de I+D. La mejor muestra de ello es la Organización para la Cooperación y el Desarrollo Económico, OCDE, uno de los organismos más reconocidos por el trabajo de estandarización conceptual para la medición de la Ciencia y la Tecnología. Según el Observatorio Colombiano de Ciencia y Tecnología, una de las variables más importantes a tener en cuenta es la cuantificación de los recursos humanos, toda vez que estos son utilizados para la identificación de las actividades de investigación y desarrollo I+D, así como la clasificación e identificación del gasto, [6].

Mediante el manejo y gestión de los recursos, y el seguimiento de sus actividades científicas; es posible medir los indicadores de gestión y productividad al interior de una organización y de esta forma detectar sus fortalezas y debilidades, promoviendo incentivos que permitan mejorar sectores específicos que se consideren estratégicos.

Por ello, se han llevado a cabo diferentes esfuerzos en el SENA para la medición de indicadores de ciencia, tecnología e innovación. Lo que ha permitido generar un modelo de analítica de datos basado en clustering mediante k-means y análisis de componentes principales, y posteriormente la implementación de dos modelos de predicción basados en random forest y K-NN.



9 ANTECEDENTES

Muchos autores han identificado diferentes indicadores que ajustan de acuerdo con el comportamiento productivo de la organización, no obstante, algunos de estos son comunes entre Sí y que son importantes tenerlos en cuenta debido a su impacto y resultados en los procesos de CTel de las organizaciones, por ejemplo Mario Albornoz en su documento de indicadores y política científica y tecnológica expone un serie de preguntas que nos invita a determinar a partir de ellas los indicadores relevantes en la identificación de capacidades de CTel como son: ¿Cómo aprovechar los conocimientos disponibles? ¿Cómo formar e incorporar los recursos humanos necesarios?, ¿Qué instituciones e instrumentos son los más indicados?, entre otras preguntas que formula (Albornoz, 2012).

Por otro lado, Jorge Robledo Velásquez en su libro Introducción a la gestión de la tecnología y la innovación empresarial, relaciona con gran pertinencia las áreas temáticas desde las cuales se deben construir los indicadores de innovación y que se han establecido en los últimos años desde los manuales de Oslo y Frascati hasta la políticas de cada uno de los países, además, se pueden generar algunos indicadores globales pertinentes a países con características sociodemográficas similares como la calidad del entorno de la organización, las capacidades de adquisición de conocimiento y los resultados logrados orientados a patentes y contenido tecnológico para exportar (Robledo Velásquez, 2020).

En este sentido, en México se procedió a realizar un estudio donde se pretendía identificar las condiciones de las capacidades CTel de los clústeres y con este análisis verificar cuál de ellos estaban todavía activos y cuales ya no estaban en funcionamiento, para este proyecto se tomó como base los indicadores establecidos en la CEPAL (2007) tomando información de diferentes fuentes de dicho país (Pérez Hernández et al., 2017).

Específicamente en el contexto educativo también se han tenido desarrollos importantes en el establecimiento de indicadores de CTel, por ejemplo Jackeline Serrano en su artículo variables para la medición de las capacidades de innovación tecnológica en instituciones universitarias, desde un concepto de organización académica, identifica el contexto de acción de la universidad en un estadio de triple hélice, el cual, le demanda algunas condiciones para el establecimiento de estos indicadores, en este sentido, plantea inicialmente un modelo de comportamiento organizacional de la universidad alineado a las dimensiones organizacionales de (Nadler & Tushman, 1980) para de esta forma identificar

las variables definitivas, dentro de las  cuales se pueden mencionar las capacidades en Investigación y desarrollo (I+D), Capacidad de producción, Capacidad de planeación, Capacidad de relacionamiento organizacional, capacidad de aprendizaje organizacional, capacidad de gestión de recursos y capacidad de mercadeo (Serrano García et al., 2017).

Igualmente, dentro del SENA ya se han realizado algunos acercamientos para la identificación y valoración de indicadores de CTel como el trabajo titulado procesos de I+D+i en el Centro de Servicios y Gestión Empresarial del SENA 2015 – 2017 de autoría de John Jairo Castro, el cual, muestra los avances en materia de actividades con resultados que se pueden categorizar en productos de CTel y a su vez, cómo se realizaron diferentes actividades para poder proporcionar espacios para su consolidación y desarrollo eficaz tanto con aprendices como con los instructores de la institución, asimismo, muestra el comportamiento de apropiación de una cultura investigativa dentro del entorno académico (Jairo et al., 2018).

Para concluir, el establecimiento adecuado de los indicadores de evaluación y medición de indicadores de CTel ha estado siempre presente en los estudios tanto a nivel administrativo como a nivel de ingeniería, con el fin, de definir la adecuada ejecución de los recursos de las organizaciones en estos procesos, es por esto que la implementación de modelos que permitan inferir el comportamiento de estos indicadores también ha sido menester continuo de los investigadores en las áreas de la gestión de la CTel debido a la importancia que tiene encontrar puntos de equilibrio donde se pueda establecer la asignación certera de recursos eficientes y eficaces para estas actividades sin sacrificar la calidad de los productos ni exceder asignaciones presupuestales y más cuando los recursos son del estado como es el caso del SENA, por tanto, es necesario implementar modelos de medición de indicadores de CTel usando tecnologías emergentes como la analítica de datos, con el fin, de dar herramientas objetivas a la institución al momentos evaluar y valorar las actividades de CTel desarrolladas en los diferentes centros de formación en todo el departamento de Antioquia.

Nuestra investigación se basa en implementar un modelo de analítica de datos que permita medir los indicadores de gestión y de productividad en los 16 centros de la regional Antioquia del SENA, mediante una base de datos generada pseudoaleatoriamente en R statistics. Posteriormente se generó una tabla denominada procesos, teniendo en cuenta las tablas Principal_Proyectos y Principal_Productos, donde se implementó un código en R statistics para obtener los siguientes campos: Centro_Año (representado por 208 etiquetas de 16 centros en 13 años), Q_Proyectos (que representa la cantidad de proyectos por cada etiqueta Centro_Año), QNA_5007 (que representa la cantidad de profesionales por Centro_Año), QNA_5009 (que representa la cantidad de magister por Centro_Año), QNA_5010 (que representa la cantidad de doctores por Centro_Año) y TV_Proyectos (que

representa el total del valor de cada  proyecto por Centro_Año). Por último se generó una tabla denominada Productos, teniendo en cuenta las tablas de datos Principal_Proyectos, Principal_Productos y Principal_Talento_Humano, Donde se implementó un código en R statistics para obtener los siguientes campos: Centro_Año (representado por 208 etiquetas de 16 centros en 13 años), Cantidad_Proyectos (que representa la cantidad de proyectos por Centro_Año), Fecha_inicio_Proyecto (que representa la fecha en la que inicia cada proyecto por Centro_Año), Cantidad_Productos (representada por la cantidad total de productos de CTel por Centro_Año), y las cantidad para cada una de las seis categorías de productos de CTel por Centro_Año. Se aplicó el método de clustering, y K-means mediante el uso de análisis de componentes principales a las tablas de datos obtenidas, para poder medir los indicadores de gestión y procesos, identificándose en tres niveles, alta gestión, media gestión y baja gestión para la tabla de procesos, y alta productividad, mediana productividad y baja productividad para la tabla de productividad. Y por último se aplicó el método supervisado K-NN y Random forest, donde se identificó si las tablas de datos procesos y productividad estaban bien clasificadas mediante la evaluación de las siguientes métricas: sensibilidad, F1 (precisión y sensibilidad), exactitud y especificidad.



10.1 VARIABLES IDENTIFICADAS

Como variables de entrada que intervienen en actividades de ciencia, tecnología e innovación se identificaron las siguientes:

Tabla 1. Variables identificadas que intervienen en la medición de indicadores

DIMENSIÓN	CATEGORIA	INDICADOR
Insumos para actividades de Ciencia, Tecnología e Innovación (ACTI) (Entradas)	Recursos económicos	Gasto/inversión en ACTI
	Recursos físicos y tecnológicos	Inversión de equipos de para ACTI
	Talento humano	Número de personas dedicadas a ACTI
	Recursos estructurales y relacionales	Número de unidades organizativas (grupos de investigación - semilleros)
		Número de programas impactados de educación superior ofertados (programas del SENA por ACTI)
		Número de convenios de colaboración para ACTI

Dentro de nuestro modelo conceptual implementado se tienen en cuenta todas las variables que permitieron obtener las variables de salida para medir indicadores de gestión y de productividad.

Tabla 2. Variables identificadas que intervienen en el módulo de procesos

DIMENSIÓN	CATEGORIA	INDICADOR
Proceso	Proyectos	Número de proyectos



	Talento humano	proporción de gasto en ACTI por personas dedicadas a ACTI
	Recursos económicos	porcentaje de gasto en ACTI en relación con el presupuesto total
		porcentaje de gasto en investigación aplicada en relación con el gasto total por ACTI
	Relaciones con el entorno	Eventos de movilidad para ACTI
	Formación científica y tecnológica	Personal de talento humano calificado que interviene en ACTI

Tabla 3. Variables de salida

DIMENSIÓN	CATEGORÍA	INDICADOR
Resultados (Salidas)	Relaciones con el medio ambiente	Número de eventos de movilidad para ACTI
	Productos resultados de las actividades de CTI	Número de productos de I+D (productos de I+D por tipología Colciencias)
	Transferencia de conocimiento	Número de asistentes a eventos de CTI organizados por la universidad (por campo de conocimiento según la clasificación de la OCDE)
	Formación científica y tecnológica	Personal de talento humano calificado que interviene en ACTI

10.2 ESTRATEGIA SENNOVA

En el Servicio Nacional de Aprendizaje – SENA, además de las actividades administrativas y de formación para el trabajo, se realizan ACTI como parte de la Formación Profesional

Integral que se ofrece en la entidad, Investigación y Desarrollo Tecnológico de ellos vinculando a sus actividades a los aprendices con el objetivo de que desarrollen sus habilidades y conocimientos investigativos mediante la ejecución de proyectos formativos que cuenten con el procedimiento y rigurosidad requeridos en la investigación aplicada (Ramírez Ariza, 2014).



algunas de ellas están enfocadas en la como la conformación de nuevos grupos y semilleros de investigación, además de consolidar y fortalecer el trabajo al interior de ellos vinculando a sus actividades a los aprendices con el objetivo de que desarrollen sus habilidades y conocimientos investigativos mediante la ejecución de proyectos formativos que cuenten con el procedimiento y rigurosidad requeridos en la investigación aplicada

En este sentido, en el ACUERDO 16 DE 2012 se regula el programa de Investigación, Desarrollo Tecnológico e Innovación, además de justificar que con dicho programa se busca contribuir al desarrollo tecnológico, la competitividad de las empresas del país, así como al fomento de la gestión en investigación, innovación y el conocimiento, ejecutando acciones para incentivar el desarrollo de proyectos y/o estrategias que generen resultados cuantificables y verificables en el sector productivo, la formación profesional integral y la formación para el trabajo, involucrando el sector productivo e instituciones generadoras de conocimiento y al SENA.

10.2.1 Listado de las actividades de Ciencia Tecnología e Innovación

En la política de investigación del SENA (Ramírez Ariza, 2014) se menciona una serie de ACTI que buscan fortalecer este proceso entre las que se encuentran:

- Fomentar la conformación de nuevos grupos y semilleros de investigación en el SENA.
- Consolidar y fortalecer el trabajo de los grupos y semilleros de investigación registrados en los centros de formación.
- Facilitar el registro y validación de los grupos de investigación dentro del SENA ante MINCIENCIAS.
- Fomentar el desarrollo tecnológico a partir de la investigación aplicada realizada.
- Aprovechar el conocimiento generado por los grupos de investigación para retroalimentación y rediseño curricular de los programas de formación.
- Fomentar la participación de aprendices en los grupos y semilleros de investigación para incrementar sus capacidades investigativas y asegurar su formación integral.

- Fomentar las alianzas entre  investigadores del Sena y universidades o centros de investigación.
- Asegurar el acceso a la información necesaria para el desarrollo de proyectos de investigación y vigilancia tecnológica en todos los centros de formación del Sena.
- Propiciar la vinculación de los productos y procesos desarrollados por investigadores o grupos de investigación en el Sena con el sector productivo.
- Apoyar el proceso de acreditación de calidad de la Formación Profesional a través del apoyo a los grupos de investigación aplicada de los Centros de Formación Profesional.
- Propiciar la difusión de productos académicos desarrollados en procesos de investigación para el fortalecimiento de los programas de formación.
- Estandarizar los criterios de reporte y publicación de los grupos de investigación.
- Fortalecer las políticas de propiedad intelectual de los productos y procesos desarrollados por investigadores y grupos de investigación asociados al Sena.
- Fomentar la comunicación y colaboración entre grupos de investigación en el Sena.
- Reforzar la cultura escrita en los centros de investigación para que los saberes empíricos lleven a nuevos conocimientos y no desaparezcan por falta de documentación adecuada.
- Difundir ampliamente los resultados de las investigaciones desarrolladas en el Sena para comunicar a empresarios, gobierno y comunidad científica en general para acelerar la innovación y evitar la repetición de trabajos.

10.3 GESTIÓN

La gestión es un término integral en función genérica que está relacionado con la planeación, la organización, la dirección y el control de los recursos en una organización, con el fin de lograr el mayor beneficio para la misma (Aktouf, 1989; Henao Castrillón & Zapata Giraldo, 2018). Por otro lado, empleando las palabras de Henao & Zapata (2018) cuando mencionan Schumpeter, esta juega un papel importante en las organizaciones al coordinar e integrar sus procesos, recursos y capacidades en beneficio de su desempeño, la creación de nuevos productos y el desarrollo de competencias adicionales.

10.3.1 INDICADOR DE GESTIÓN



Un indicador de gestión es la expresión cuantitativa del comportamiento y desempeño de un proceso, cuya magnitud, al ser comparada con algún nivel de referencia, puede estar señalando una desviación sobre la cual se toman acciones correctivas o preventivas según el caso.

Para trabajar con los indicadores debe establecerse todo un sistema que vaya desde la correcta comprensión del hecho o de las características hasta la de toma de decisiones acertadas para mantener, mejorar e innovar el proceso del cual dan cuenta.

Los indicadores de gestión deben cumplir con unos requisitos y elementos para poder apoyar la gestión y conseguir el objetivo propuesto. Estas características son:

Simplicidad. Puede definirse como la capacidad para definir el evento que se pretende medir, de manera poco costosa en tiempo y recurso.

Adecuación. Entendida como la facilidad de la medida para describir por completo el fenómeno o efecto. Debe reflejar la magnitud del hecho analizado y mostrar la desviación real del nivel deseado.

Validez en el tiempo. Puede definirse como la propiedad de ser permanente por un periodo deseado.

Participación de los usuarios. Es la habilidad para estar involucrados desde el diseño, y debe proporcionárseles los recursos y formación necesarios para su ejecución. Este es quizás el ingrediente fundamental para que el personal se motive en torno al cumplimiento de los indicadores.

Utilidad. Es la posibilidad del indicador para estar siempre orientado a buscar las causas que han llevado a que alcance un valor particular y mejorarlas.

Oportunidad. Entendida como la capacidad para que los datos sean recolectados a tiempo. Igualmente requiere que la información sea analizada oportunamente para poder actuar.

10.4 PRODUCTIVIDAD

La productividad es la relación cuantitativa entre lo que se produce y lo que se gasta para producir. La productividad no es más que una reducción en el desperdicio de recursos como los hombres, el material, la máquina, el tiempo, el espacio, el capital, etc. Denota la relación

entre el producto y uno o todos los



insumos asociados. En palabras de Peter Drucker, productividad significa un equilibrio entre todos los factores de producción que proporcionan el máximo rendimiento con el menor esfuerzo, es, por lo tanto, la cantidad de bienes o servicios (producto) que producen. La tierra y los materiales de construcción, máquinas, mano de obra (mano de obra), tecnología, etc. son los recursos a disposición de la organización. Por lo tanto, una mayor (mejor) productividad significa que se produce más con el mismo gasto de recursos, es decir, al mismo costo en términos de tierra, materiales, máquina, tiempo o mano de obra, o bien, se produce la misma cantidad a un menor costo en términos de tierra, materiales, tiempo de la máquina o mano de obra que se utiliza.

¿Cuáles son los factores que afectan la productividad de los procesos?

Hay varios factores que afectan la productividad de los procesos y, por lo tanto, el rendimiento de la organización. Algunos factores tienen un efecto positivo en la productividad de los procesos, mientras que otros factores tienen una influencia negativa en la productividad. Algunos de estos factores se pueden controlar mientras que los otros no se pueden controlar. Los factores importantes que afectan la productividad se enumeran a continuación.

Inversión de capital: las inversiones de capital en la tecnología, los equipos y las instalaciones influyen en la productividad del proceso. Cualquier ahorro en estos elementos de proceso puede tener un gran efecto negativo en la productividad del proceso.

Materiales de entrada: la calidad del material de entrada tiene una gran influencia en la productividad del proceso. Algunos materiales de entrada pueden tener un menor costo por unidad de peso, pero tienen un gran impacto adverso en la productividad del proceso debido a que su calidad es inferior a la calidad necesaria para el proceso.

Capacidad de proceso: la capacidad del proceso influye en su productividad. Con una gran capacidad, la productividad del proceso mejora debido a las economías de escala.

Tecnología del proceso: la tecnología seleccionada para el proceso juega un papel muy importante en la determinación de la productividad del proceso. Para varios procesos, hay tecnologías alternativas disponibles para producir el mismo producto. Estas diferentes tecnologías normalmente tienen diferentes niveles de productividad. En algunos casos, la tecnología seleccionada puede ser la misma, pero la velocidad del proceso influye en la productividad. Como ejemplo, un tren de laminación con menor velocidad máxima de laminación no puede tener la misma productividad que el tren de laminación que tiene una velocidad máxima de laminación más alta.

Mano de obra: el conocimiento, la  experiencia y el comportamiento de la fuerza de trabajo son los factores principales que ayudan a mantener la productividad del proceso. Para maximizar la productividad del proceso, tanto el conocimiento como la experiencia de la fuerza laboral son necesarios ya que estas cualidades mejoran las habilidades de la fuerza de trabajo y, por lo tanto, tienen una influencia vital en la productividad del proceso. Las habilidades de la fuerza laboral se pueden mejorar mediante capacitación y educación. La capacitación y la educación también ayudan a moldear el comportamiento de la fuerza de trabajo.

10.4.1 INDICADOR DE PRODUCTIVIDAD

Los indicadores de productividad son un valor medible (también KPI) que demuestra la eficacia con la que una empresa está logrando objetivos comerciales clave. Las organizaciones usan indicadores clave de productividad en múltiples niveles para evaluar su éxito en alcanzar objetivos.

Los indicadores de alto nivel pueden centrarse en el rendimiento general de una organización, mientras que los indicadores de bajo nivel pueden centrarse en los procesos en departamentos como ventas, marketing, recursos humanos, soporte y otros.

Tipos de indicadores: Existen principalmente dos tipos de indicadores de productividad que son:

Indicadores Cuantitativos

Son aquellos que se basan en cantidades o tiempo. Ya que se calculan con ayuda del cociente entre cantidad producida y el tiempo necesario para su realización. Igualmente, se utilizan en la relación entre inversión y número de productos consumados. Se utilizan para medir la productividad de la empresa, un equipo de trabajo o un producto o servicio.

Para nuestro proyecto se identificaron seis tipos de variables de CTel que permitieron medir el nivel de productividad en el SENA, estas se describen a continuación: Apropiación_social, Artículo_Sometido_en_revista_Indexada, Divulgación_Pública_de_la_CTel_Publicaciones_editoriales_no_especializadas_producciones_de_contenido_digital, Informe_de_investigación y Libros_resultados_de_investigación.

Indicadores Cualitativos

Miden la calidad del producto final o servicio prestado. Estos indicadores revelan las cualidades positivas o valor de lo realizado, sin limitarse a su cantidad. En nuestro caso se

establecieron niveles de productividad en CTel como alta productividad, mediana productividad y baja productividad.



11 MARCO REFERENCIAL

Para la medición de indicadores de gestión y de productividad, se implementaron los siguientes métodos no supervisados: Clustering, k-means y análisis de componentes principales, y como métodos supervisados random forest y K-NN. El proceso buscaba primero encontrar la variable de clasificación por clustering para 16 centros de la regional Antioquia durante 13 años y de esta forma identificar el nivel de gestión y productividad en tres niveles a saber: alta gestión, mediana gestión y baja gestión y alta productividad, mediana productividad, y baja productividad, por último, se implementaron dos modelos de predicción para evaluar el modelo de clasificación. A continuación, se explican los pasos que se llevaron a cabo:

I. Identificación de las variables de la base de datos a generar

La base de datos tiene en total 21 tablas, cada una de ellas representada por los siguientes campos.

Tabla 4. Identificación de tablas y campos para la base de datos

Regionales	Grupo_investigacion	Semilleros
_Codigo_Regional	_Codigo_Grupo_Investigacion	_Codigo_Semillero
_Nombre_Regional	_Nombre_Grupo_Investigacion	_Nombre_Semillero
_Codigo_Departamento	_Categoria_Grupo	_Codigo_Grupo_Investigacion
_Codigo_Municipio	_Codigo_Centro	_Codigo_Centro
	_Codigo_Regional	_Codigo_Regional



Rubro _Codigo_Rubro _Nombre_Rubro	Rol_Sennova _Codigo_Rol _Nombre_Rol	Tipo_Contratacion _Codigo_Tipo_Contrato _Nombre_Tipo_Contrato
Principal_Productos _Codigo_Producto _Nombre_Producto _Descripcion_producto _Codigo_Regional _Codigo_Centro _Codigo_Rubro _Codigo_Tipologia _Codigo_Proyecto _Nombre_Proyecto _Codigo_Grupo_Investigacion _Codigo_Semillero _Codigo_persona _Codigo_area_OCDE _Fecha_registro	Principal_Proyectos _Codigo_Proyecto _Nombre_Proyecto _Valor_Proyecto _Codigo_Rubro _Fecha_inicio_Proyecto _Fecha_Cierre_Proyecto _Codigo_Regional _Codigo_Centro _Codigo_Estado_Proyecto _Codigo_area_OCDE _Serial_equipo _Valor_Servicios_Personales _Valor_Compra_Equipos _Valor_Software _Valor_Otros_Gastos	Principal_Talento_Humano _Codigo_Persona _Codigo_Nivel_Academico _Codigo_Tipo_Contrato _Objeto_Contrato _Codigo_Rubro _Codigo_Area_OCDE _Codigo_Estado_Contratacion _Codigo_Regional _Codigo_Centro _Codigo_Proyecto _Codigo_Producto
Area_conocimiento_OCDE _Codigo_area_OCDE _Nombre_area_OCDE	Estado_Proyecto _Codigo_Estado_Proyecto _Nombre_Estado_Proyecto	Nivel_Academico _Codigo_Nivel_Academico _Nombre_Nivel_Academico
Municipio _Codigo_Municipio _Nombre_Municipio	Centros_Formacion _Codigo_Centro _Nombre_Centro	Principal_Equipos _Serial_equipo _Nombre_del_equipo



_Codigo_Departamento	_Codigo_Regional _Codigo_Municipio _Codigo_Departamento _latitud _longitud	_Valor_equipo _Fecha_de_compra _Codigo_Proyecto _Codigo_Regional _Codigo_Centro _Codigo_Rubro _Codigo_Tipo_Equipo _Codigo_area_OCDE _Codigo_persona
Tipologia_Productos _Codigo_Tipologia _Nombre_Tipologia	Tipo_Contratación _Codigo_Tipo_Contrato _Nombre_Tipo_Contrato	Tipo_equipo _Codigo_Tipo_Equipo _Nombre_Tipo_Equipo
Rol_Proyectos _Codigo_Rol_Proyecto _Nombre_Rol	Estado_Contratación _Codigo_Estado_Contratacion _Nombre_Estado_Contratacion	Departamento _Codigo_Departamento _Nombre_Departamento

II. Identificación de la tabla de procesos y productividad, con las respectivas variables.

Tabla 11.1 . Identificación de campos para la tabla de datos Procesos

Centro_Año Es una etiqueta que posee 16 centros en 13 años, para un total de 208	Q_Proyectos Representa la cantidad de proyectos por Centro_Año	QNA_5007 Representa la cantidad de profesionales por Centro_Año
--	--	---



<p>QNA_5009</p> <p>Representa la cantidad de magister por Centro_Año</p>	<p>QNA_510</p> <p>Representa la cantidad de doctores por Centro_Año</p>	<p>TV_Proyectos</p> <p>Representa el total del valor de cada proyecto por Centro_Año</p>
---	--	---

Tabla 11.2. Identificación de campos para la tabla de datos Productividad

<p>Centro_Año</p> <p>Es una etiqueta que posee 16 centros en 13 años, para un total de 208</p>	<p>Cantidad_Proyectos</p> <p>Representa la cantidad de proyectos para cada Centro_Año</p>	<p>Fecha_inicio_Proyecto</p> <p>Representa la fecha en la que inicia cada proyecto</p>
<p>Cantidad_Productos</p> <p>Representa la cantidad de productos de Ciencia, Tecnología e Innovación por Centro_Año</p>	<p>P_8020</p> <p>Representa la cantidad de productos de Apropiación_social</p>	<p>P_8025</p> <p>Representa la cantidad de productos de Artículo_Sometido_en_revista_Indexada</p>
<p>P_8019</p> <p>Representa la cantidad de productos de Divulgación_Pública_de_la_CTel_Publicaciones_editoriales_no_especializadas_producciones_de_contenido_digital</p>	<p>P_8039</p> <p>Representa la cantidad de productos de Informe_de_investigación</p>	<p>P_8016</p> <p>Representa la cantidad de productos Libros_resultados_de_investigación</p>
<p>P_8031</p> <p>Representa la cantidad de productos de</p>		



III. Implementación del método de clustering basado en K-means y análisis de componentes principales.

Definición de Clustering

Considere un conjunto de datos que contiene n instancias/observaciones cada una descrita por m atributos. El problema de agrupamiento de k -medias es dividir el n instancias en k clústeres, con los clústeres particionando las instancias $(x_1, x_2, \dots, x_n)$ en los subconjuntos $Q_{1..k}$. Los subconjuntos se pueden resumir como puntos $(C_{1..k})$ en el espacio m dimensional, comúnmente conocido como **centroides** o centros de conglomerados, cuyas coordenadas son el promedio de todos los puntos pertenecientes al subconjunto. Nos referiremos a esta colección de centroides obtenidos como un modelo de aplicación del algoritmo de agrupamiento. El agrupamiento de K-means también se puede considerar como una cuantificación vectorial con el objetivo de minimizar el error de cuantificación del vector (también conocido como distorsión) que se muestra en la ecuación (1). La solución matemática trivial es tener un clúster para cada instancia, pero normalmente $k \ll n$.

$$VQ = \frac{1}{2} \sum_{i=1}^n D(x_i, C_{f(x_i)}) \quad (1)$$

Donde D es una función distancia y $f(x)$ devuelve el índice del clúster en la instancia i

Calculamos el k óptimo por **Método de Elbow**: El método del codo se usa con el error cuadrático (sse) o la suma cuadrática del grupo (wcsc). Para nuestro caso, usaré wcsc para encontrar la cantidad óptima de clústeres. El siguiente código es una manera fácil de obtener el valor de wcsc para diferentes números de clústeres:

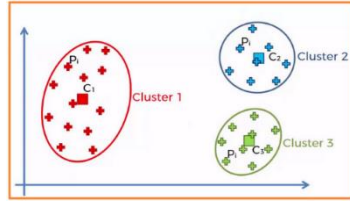


Figura 11.1. Diagrama representativo de clúster

Matemáticamente

$$WCSS = \sum_{P_i \text{ in Cluster 1}} distancia (P_i, C_1)^2 + \sum_{P_i \text{ in Cluster 2}} distancia (P_i, C_2)^2 + \sum_{P_i \text{ in Cluster 3}} distancia (P_i, C_3)^2$$

Método de Clustering

Esta técnica se basa en la clusterización y consiste en particionar un conjunto de m observaciones, en un conjunto de datos con características similares, se parte de una muestra E de m individuos, $X_1, \dots, X_m$ de variables aleatorias, cada una de los cuales está representada por un vector n – dimensional, $X_j = (x_{j1}, x_{j2}, \dots, x_{jn})'$, $j = 1, \dots, m$, [14] debemos encontrar una partición de la muestra en regiones $\omega_1, \dots, \omega_c$ de forma que:

$$\bigcup_{i=1}^c \omega_i = E \quad (1)$$

$$\omega_i \cap \omega_j = \emptyset ; i \neq j \quad (2)$$

Método K-means



Se desea particionar un conjunto de datos determinado en varios grupos: cada observación del conjunto de datos en el mismo grupo comparte más similitudes. El algoritmo busca minimizar la varianza total del sistema. Por lo tanto, para un conjunto de datos $\{x_i\}_{i=1}^n \in R^p$, tal que, se busca encontrar k centroides que minimicen la distancia total entre estos puntos de datos y sus centroides correspondientes, donde $\{c_j\}_{j=1}^k$ son los centroides y $\{\Psi_j\}_{j=1}^k$ son sus particiones correspondientes. Aquí $\{\Psi_j\}_{j=1}^k$ son mutuamente disyuntos y $\cup_{j=1}^k \Psi_j = [n]$. La suma total de la distancia es [15]:

$$f(c_j, \Psi_j) := \sum_{j=1}^k \sum_{i \in \Psi_j} d(x_i, c_j) \quad (3)$$

Se denomina función k-means, que depende de los centroides $\{c_j\}_{j=1}^k$ y $\{\Psi_j\}_{j=1}^k$. Para este caso $d(x_i, c_j)$ se denota como la distancia, que mide la similitud entre x y y .

Método de componentes principales

Es una técnica estadística multivariada cuyo objeto es representar un conjunto de variables observadas en un grupo de individuos por un número pequeño de variables nuevas construidas mediante combinaciones lineales de las variables originales. A continuación, se definen los componentes principales en términos poblacionales [16]:

Sea $X = (X_1, \dots, X_p)'$ un vector aleatorio p – dimensional y sean μ y Σ su esperanza y covarianza

$$E[X] = \mu = 0 \quad Cov[X] = \Sigma = (\sigma_{ij}) \quad (4)$$

Sea $\alpha = (\alpha_1, \dots, \alpha_p)' \in R^p$ tal que $\alpha' \alpha = 1$. Se consideran combinaciones lineales

$$\alpha' X = (\alpha_1, \dots, \alpha_p) \begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix} = \alpha_1 X_1 + \dots + \alpha_p X_p = \sum_{i=1}^p \alpha_i X_i \quad (5)$$

Cuya varianza vendrá dada por



$$Var[\alpha'X] = E[(\alpha'X)^2] = E[\alpha'XX'\alpha] = \alpha'\Sigma\alpha \quad (6)$$

Cálculo de las componentes principales

Sea una matriz X con $n \times p$ datos

$$X = (x_{11} \cdots x_{1p} \vdots \vdots x_{n1} \cdots x_{np}) \quad (7)$$

Definida en una población con matriz de covarianza Σ o correlación para las p variables involucradas.

Se denotan las variables de este conjunto como elementos del vector aleatorio:

$$\underline{x} = (x_1, x_2, \dots, x_p) \quad (8)$$

Con el PCA intentaremos construir las siguientes combinaciones lineales:

$$\begin{aligned} y_1 &= \underline{a}'_1 \underline{x} = a_{11}x_1 + a_{12}x_2 + \cdots + a_{1p}x_p, \\ y_2 &= \underline{a}'_2 \underline{x} = a_{21}x_1 + a_{22}x_2 + \cdots + a_{2p}x_p, \\ &\vdots \\ y_p &= \underline{a}'_p \underline{x} = a_{p1}x_1 + a_{p2}x_2 + \cdots + a_{pp}x_p \end{aligned} \quad (9)$$

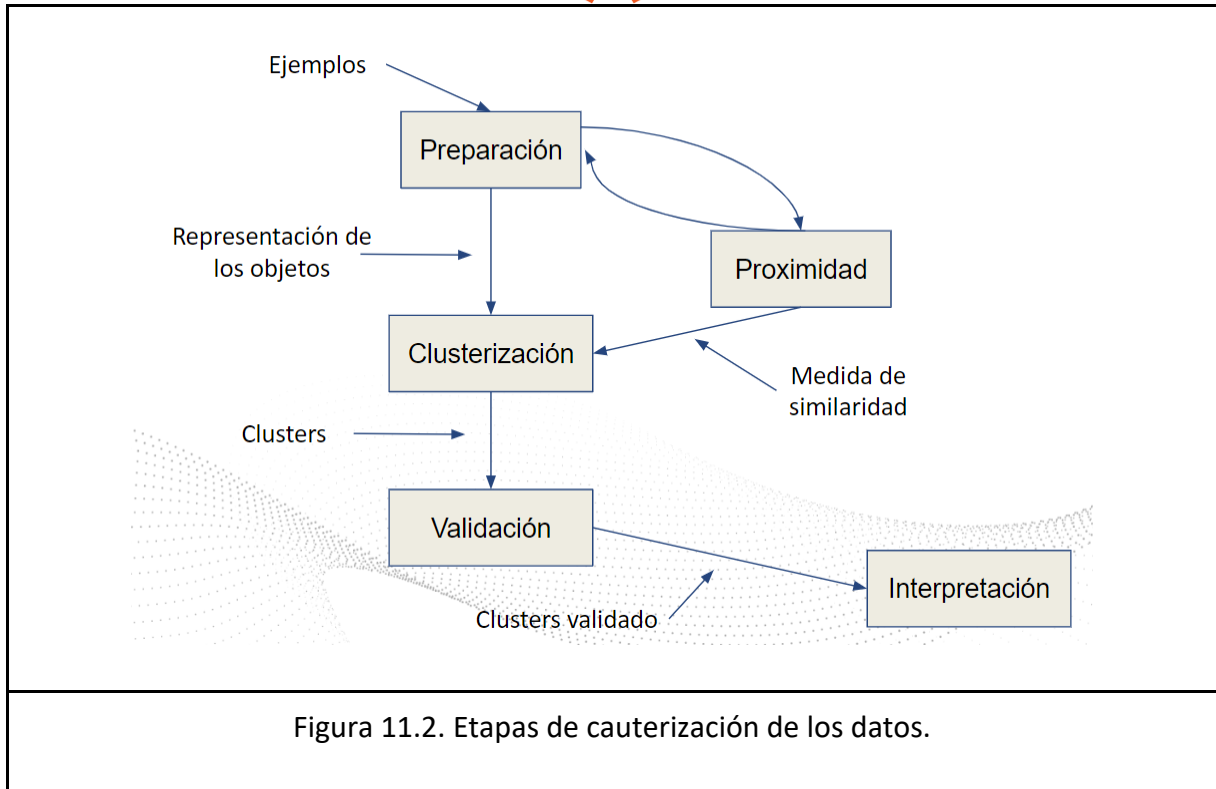
Donde $\underline{a}'_1, \underline{a}'_2, \dots, \underline{a}'_p$ son desconocidos y  los $p \times 1$ son los vectores que contienen los pesos de las combinaciones lineales. Por lo tanto, la primera componente principal y_1 explica la mayor varianza posible, lo que resultará de las combinaciones lineales de las variables iniciales en $\underline{x}$. Se impone una restricción para la magnitud de los elementos de $\underline{a}_1$ ya que, si no se hace, la $var(y_1)$ aumentará desmedidamente, lo cual impide encontrar el vector $\underline{a}_1$. Por lo general asumimos que la longitud es de 1, esto es $\|\underline{a}_1\| = 1$, una vez que se encuentre la solución y el vector, óptimo de pesos $\underline{a}_1$ la combinación lineal $\underline{a}'_1 \underline{x}$ es llamada la primera componente principal. El siguiente paso es encontrar un segundo vector $\underline{a}_2$ tal que la segunda componente $\underline{a}'_2 \underline{x}$ tenga las siguientes propiedades [16]:

1. Debe de explicar la mayor varianza posible de los datos restantes (considerando de nuevo la restricción).
2. No debe estar correlacionado con la primera componente principal.

Ambas componentes no deben estar correlacionadas, se busca hallar el número de combinaciones lineales que expliquen la porción más grande posible de las varianzas de las variables iniciales. Una vez encontrado este conjunto de valores $\underline{a}_2$, las combinaciones lineales $\underline{a}'_2 \underline{x}$ se denominará la segunda componente principal. Podemos continuar con este comportamiento hasta que encontremos p vectores $\underline{a}_1, \underline{a}_2, \dots, \underline{a}_p$ que aparecen en el lado derecho de la ecuación (9), por definición las sucesivas componentes principales no están correlacionadas y formalmente hay tantas componentes principales como variables estudiadas [16].

Etapas del proceso de clusterización

Los procesos de clusterización necesitan de unas etapas que validen la estrategia de conglomeración de los datos.



- **Preparación:** normalizaciones, conversiones de tipos, reducción del número de atributos
- **Proximidad:** definición (elección) de una medida de similitud apropiada al dominio de la aplicación, por ejemplo. distancia.
- **Clustering:** aplicación de algoritmos de clustering
- **Validación:** evaluación del resultado de la agrupación. Debe determinar si los conglomerados son significativos.
- **Interpretación:** etiquetado del clúster, definiendo su naturaleza, a través del análisis de sus objetos.

IV. Implementación del método de predicción basado en random forest y K-NN para evaluación del método de clasificación a través del uso de métricas.

¿Qué es un algoritmo de aprendizaje supervisado?

En aprendizaje supervisado, un  algoritmo recibe un conjunto de datos que están etiquetados con los valores de salida correspondientes sobre los que puede entrenarse y definir un modelo de predicción. Este algoritmo podrá luego ser utilizado sobre datos nuevos para predecir sus valores de salida correspondientes.

El algoritmo de las K vecinas más cercanas o K-nearest neighbors (kNN) es un algoritmo de Machine Learning que pertenece a los algoritmos de aprendizaje supervisado simples y fáciles de aplicar que pueden ser utilizados para resolver problemas de clasificación y de regresión. En este artículo veremos la definición de este algoritmo, su funcionamiento, así como también una aplicación directa en programación.

Método K-NN

La lógica detrás del algoritmo de las K vecinas más cercanas es una de las más sencillas de todos los algoritmos de Machine Learning supervisados:

- Etapa 1: Seleccionar el número de K vecinas
- Etapa 2: Calcular la distancia
- Etapa 3: tomar las K vecinas más cercanas según la distancia calculada
- Etapa 4: entre las K vecinas, contar el número de puntos en cada categoría.
- Etapa 5: atribuir un nuevo punto a la categoría más presente entre las K vecinas
- Etapa 6: El modelo está listo

Método Random Forest

- Se considera como la “panacea” en todos los problemas de ciencia de datos.
- Útil para regresión y clasificación.
- Un grupo de modelos “débiles”, se combinan en un modelo robusto.
- Sirve como una técnica para reducción de la dimensionalidad.
- Se generan múltiples árboles (a diferencia de CART: Comparación entre árboles de regresión).
- Cada árbol da una clasificación (vota por una clase). Y el resultado es la clase con mayor número de votos en todo el **random forest**.

- Para regresión, se toma el  promedio de las salidas (predicciones) de todos los árboles.

¿Cómo se construye un modelo random forest?

Cada árbol se construye así:

- Dado que el número de casos en el conjunto de entrenamiento es N . Una muestra de esos N casos se toma aleatoriamente, pero **CON REEMPLAZO**. Esta muestra será el conjunto de entrenamiento para construir el árbol i .
- Si existen M variables de entrada, un número $m < M$ se especifica tal que para cada nodo, m variables se seleccionan aleatoriamente de M . La mejor división de estos m atributos es usado para ramificar el árbol. El valor m se mantiene constante durante la generación de todo el bosque.
- Cada árbol crece hasta su máxima extensión posible y **NO hay proceso de poda**.
- Nuevas instancias se predicen a partir de la agregación de las predicciones de los x árboles (i.e., mayoría de votos para clasificación, promedio para regresión)

Funcionamiento del algoritmo Random Forest

Out of bag samples y out of bag error (Muestras fuera de bolsa y error fuera de bolsa)

- El proceso de muestreo de los datos con reemplazo se denomina **bootstrap**.
- Un tercio de los datos no se usan para el entrenamiento y pueden ser usados para test.
- Este conjunto se denomina out of bag (OOB) samples.



- El error estimado en estos out of bag samples se conoce como out of bag error (OOB error)
- Usar este conjunto de test (OOB) es tan preciso como si se usara un conjunto de test del mismo tamaño que el de entrenamiento.
- Sería posible no usar un conjunto de test adicional.

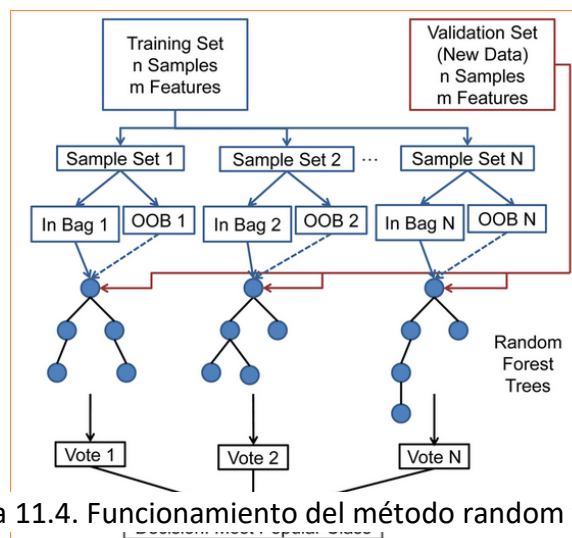


Figura 11.4. Funcionamiento del método random forest



12 METODOLOGÍA

12.1 Metodología de la investigación

El enfoque desarrollado en nuestra investigación es de tipo mixto este implica la recolección y el análisis de datos cuantitativos y cualitativos, permitiendo integrarlos de manera conjunta, para realizar inferencias de la información realizada y lograr una mayor comprensión del objeto de estudio propuesto [6], ya que el tema se abordó por primera vez en este contexto, el método de estudio es fenomenológico ya que intenta interpretar y comprender el proceso de desarrollo, ejecución y validación del modelo de evaluación y valoración de capacidades de (CTel) en los centros de formación de la regional Antioquia del SENA, asimismo, el proceso de investigación es inductivo - deductivo. Los fundamentos expuestos en el marco del sistema de I + D [7] pretenden medir las capacidades de ciencia, tecnología e innovación desde cuatro dimensiones, que son entrada, proceso, salida y entorno.

Sin embargo, la propuesta se realizó teniendo en cuenta sólo las dimensiones de entrada y salida, dado que; en primer lugar, ignoramos la variable entorno debido a que las políticas de (CTel) son cambiantes de acuerdo con la industria y la región. En segundo lugar, el proceso juega un papel clave en la formación de capacidades, sin embargo, está excluido por dos razones: la primera referida a la dificultad en cuantificar para algunas organizaciones los procesos y actividades de ciencia, tecnología e innovación y la segunda en cuanto a la intensidad de los insumos de (ACTI) y los resultados de producción que podrían representar el proceso de (CTel) en conjunto. Se propuso un modelo de medición de indicadores para que las distintas organizaciones evalúen el impacto de las (ACTI), en la figura 1 se realizó a través de indicadores de insumos y productos [8,9,10]. Siendo los insumos de (CTel) todos aquellos recursos que respaldan las actividades de (CTel). Por otro lado, los indicadores de salida (CTel), son los resultados producto de las ACTI [11].

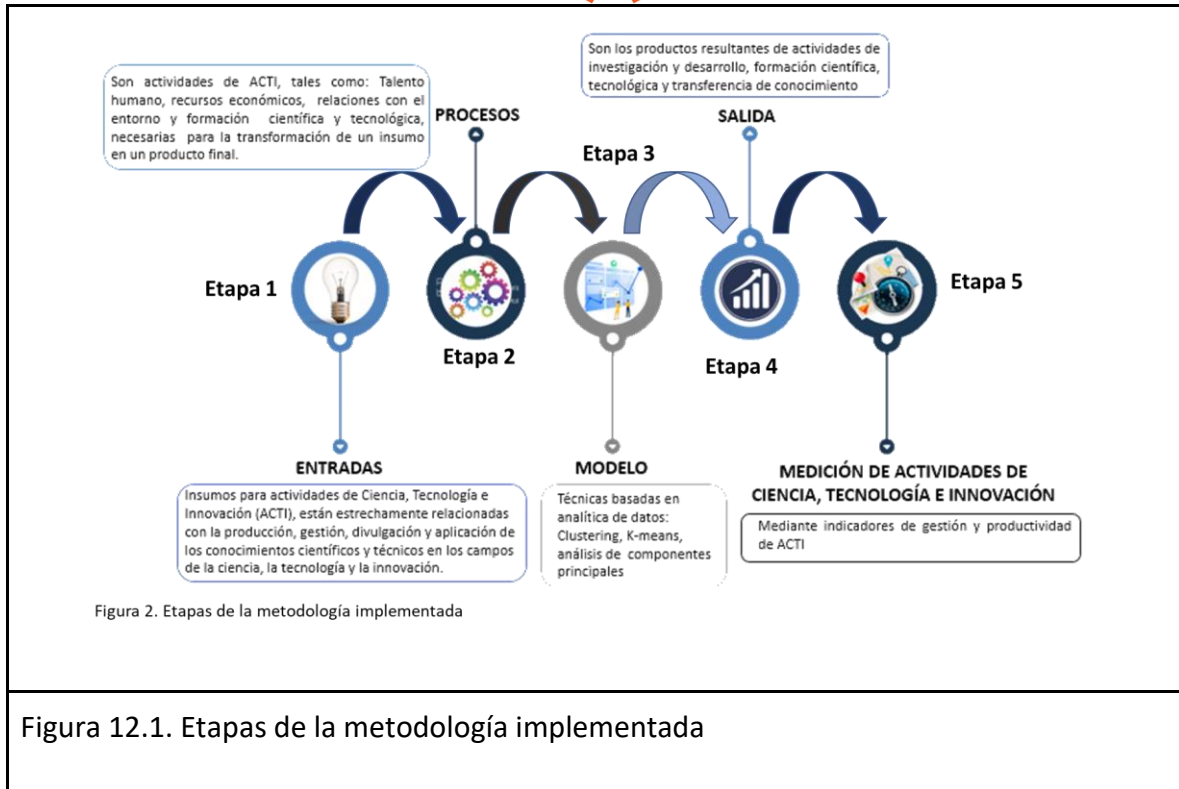


Figura 12.1. Etapas de la metodología implementada

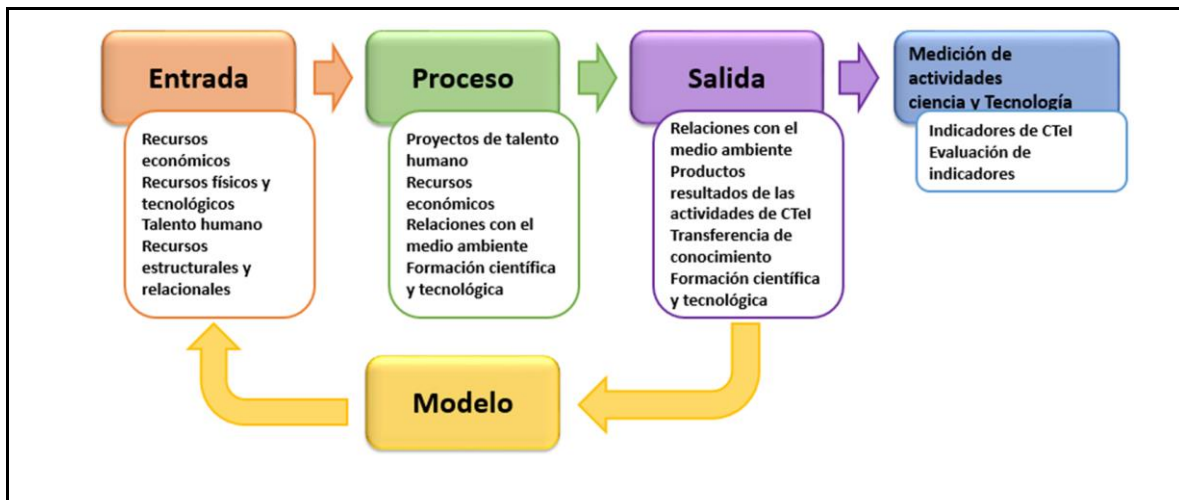


Figura 12.2. Modelo conceptual de la evaluación de la gestión de la ciencia, la tecnología y la innovación

12.2 Metodología CRISP-DM



La metodología CRISP - DM de sus siglas en inglés de *Cross-Industry Standard Process for Data Mining*, es uno de los métodos más reconocidos que permiten orientar trabajos de minería de datos. Esta permite la extracción del conocimiento de un grupo de datos en diferentes etapas

12.2.1 Comprensión de los datos: El objetivo de esta fase es familiarizarse con los datos disponibles.

12.2.2 Preparación de los datos: Se realiza una etapa de preprocesamiento y exploración de los datos, tales como limpieza, normalización y preparación de la información que lleven a alcanzar los objetivos que persigue el proyecto.

12.2.3 Modelado: Se determina la técnica matemática a implementar y se buscan los algoritmos que permitan resolver el problema objetivo con los parámetros establecidos.

12.2.4 Evaluación: En esta fase se realiza la etapa de validación de modelos, en el que permite dar veracidad de los resultados obtenidos a partir de métricas que permite dar certeza de esto.

12.2.5 Implementación: En esta fase se busca lograr que el modelo propuesto cumpla con los objetivos del proyecto, en base a los requisitos comerciales de la organización en este caso la medición de los indicadores de ciencia, tecnología e innovación.

12.2.6 Modelo de analítica de datos: Integración de técnicas que permitan la toma de decisiones en base a los datos propuestos, todo esto orientado al propósito de medir indicadores de Ciencia Tecnología e Innovación.

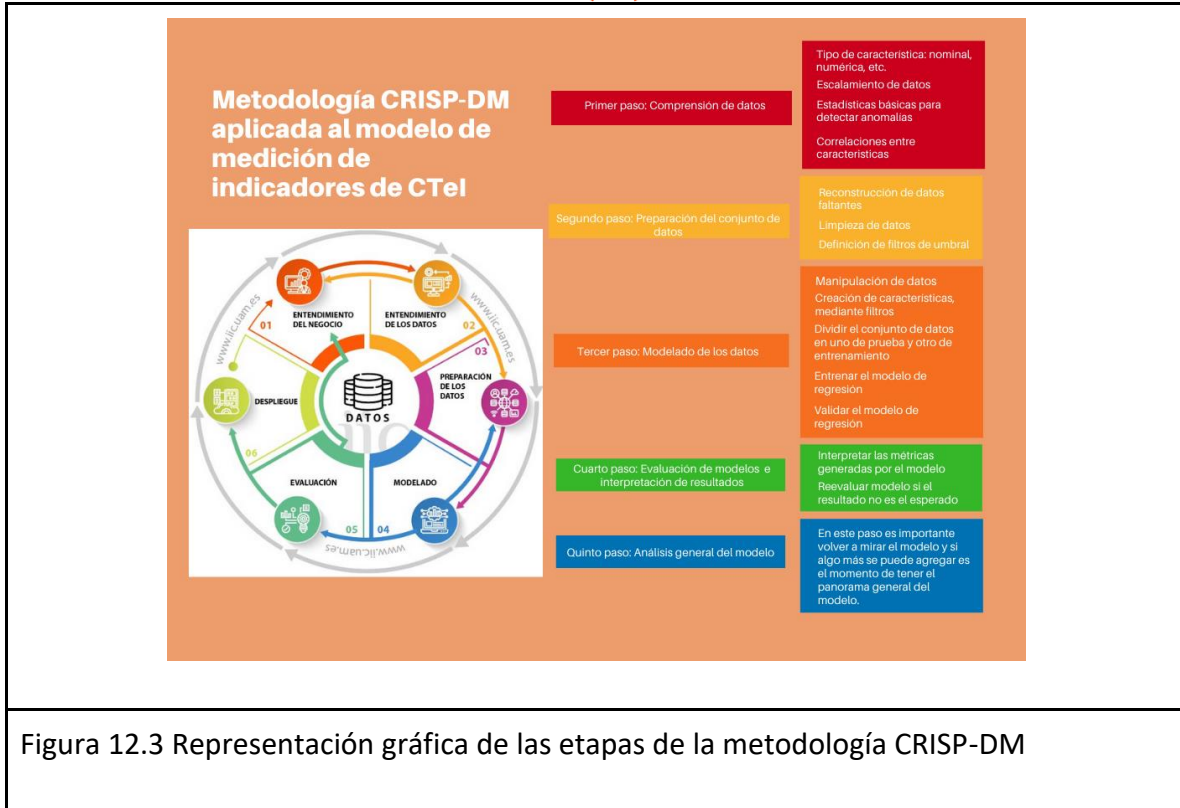


Figura 12.3 Representación gráfica de las etapas de la metodología CRISP-DM

12.3 Descripción de los datasets

Esta sección describe todos los conjuntos de datos utilizados para el desarrollo de la solución propuesta. Se creó un conjunto de datos sintéticos que permitieron generalizar los resultados de los modelos de la analítica de datos.

12.3.1 Resumen de los datasets y modelo relacional. En la siguiente tabla se resumen las principales características de los conjuntos de datos que tendremos en cuenta a lo largo del desarrollo de nuestro proyecto.



Regionales	Grupo_investigacion	Semilleros
_Codigo_Regional	_Codigo_Grupo_Investigacion	_Codigo_Semillero
_Nombre_Regional	_Nombre_Grupo_Investigacion	_Nombre_Semillero
_Codigo_Departamento	_Categoria_Grupo	_Codigo_Grupo_Investigacion
_Codigo_Municipio	_Codigo_Centro	_Codigo_Centro
	_Codigo_Regional	_Codigo_Regional
Rubro	Rol_Sennova	Tipo_Contratacion
_Codigo_Rubro	_Codigo_Rol	_Codigo_Tipo_Contrato
_Nombre_Rubro	_Nombre_Rol	_Nombre_Tipo_Contrato
Principal_Productos	Principal_Proyectos	Principal_Talento_Humano
_Codigo_Producto	_Codigo_Proyecto	_Codigo_Persona
_Nombre_Producto	_Nombre_Proyecto	_Codigo_Nivel_Academico
_Descripcion_producto	_Valor_Proyecto	_Codigo_Tipo_Contrato
_Codigo_Regional	_Codigo_Rubro	_Objeto_Contrato
_Codigo_Centro	_Fecha_inicio_Proyecto	_Codigo_Rubro
_Codigo_Rubro	_Fecha_Cierre_Proyecto	_Codigo_Area_OCDE
_Codigo_Tipologia	_Codigo_Regional	_Codigo_Estado_Contratacion
_Codigo_Proyecto	_Codigo_Centro	_Codigo_Regional
_Nombre_Proyecto	_Codigo_Estado_Proyecto	_Codigo_Centro
_Codigo_Grupo_Investigacion	_Codigo_area_OCDE	_Codigo_Proyecto
_Codigo_Semillero	_Serial_equipo	_Codigo_Producto
_Codigo_persona	_Valor_Servicios_Personales	
_Codigo_area_OCDE	_Valor_Compra_Equipos	
_Fecha_registro	_Valor_Software	
	_Valor_Otros_Gastos	
Area_conocimiento_OCDE	Estado_Proyecto	Nivel_Academico
_Codigo_area_OCDE	_Codigo_Estado_Proyecto	_Codigo_Nivel_Academico
_Nombre_area_OCDE	_Nombre_Estado_Proyecto	_Nombre_Nivel_Academico
Municipio	Centros_Formacion	Principal_Equipos
_Codigo_Municipio	_Codigo_Centro	_Serial_equipo
_Nombre_Municipio	_Nombre_Centro	_Nombre_del_equipo
_Codigo_Departamento	_Codigo_Regional	_Valor_equipo
	_Codigo_Municipio	_Fecha_de_compra
	_Codigo_Departamento	_Codigo_Proyecto
	_latitud	_Codigo_Regional
	_longitud	_Codigo_Centro
		_Codigo_Rubro
		_Codigo_Tipo_Equipo
		_Codigo_area_OCDE
		_Codigo_persona
Tipologia_Productos	Tipo_Contratacion	Tipo_equipo
_Codigo_Tipologia	_Codigo_Tipo_Contrato	_Codigo_Tipo_Equipo
_Nombre_Tipologia	_Nombre_Tipo_Contrato	_Nombre_Tipo_Equipo
Rol_Proyectos	Estado_Contratacion	Departamento
_Codigo_Rol_Proyecto	_Codigo_Estado_Contratacion	_Codigo_Departamento
_Nombre_Rol	_Nombre_Estado_Contratacion	_Nombre_Departamento

Tabla 12.4. Base de datos de tablas y variables a analizar

De este conjunto de datos, se  seleccionaron las tablas que se relacionan con la dimensión de procesos de gestión y productividad, como se muestra en el siguiente diagrama de entidad-relación de la Tabla 7.

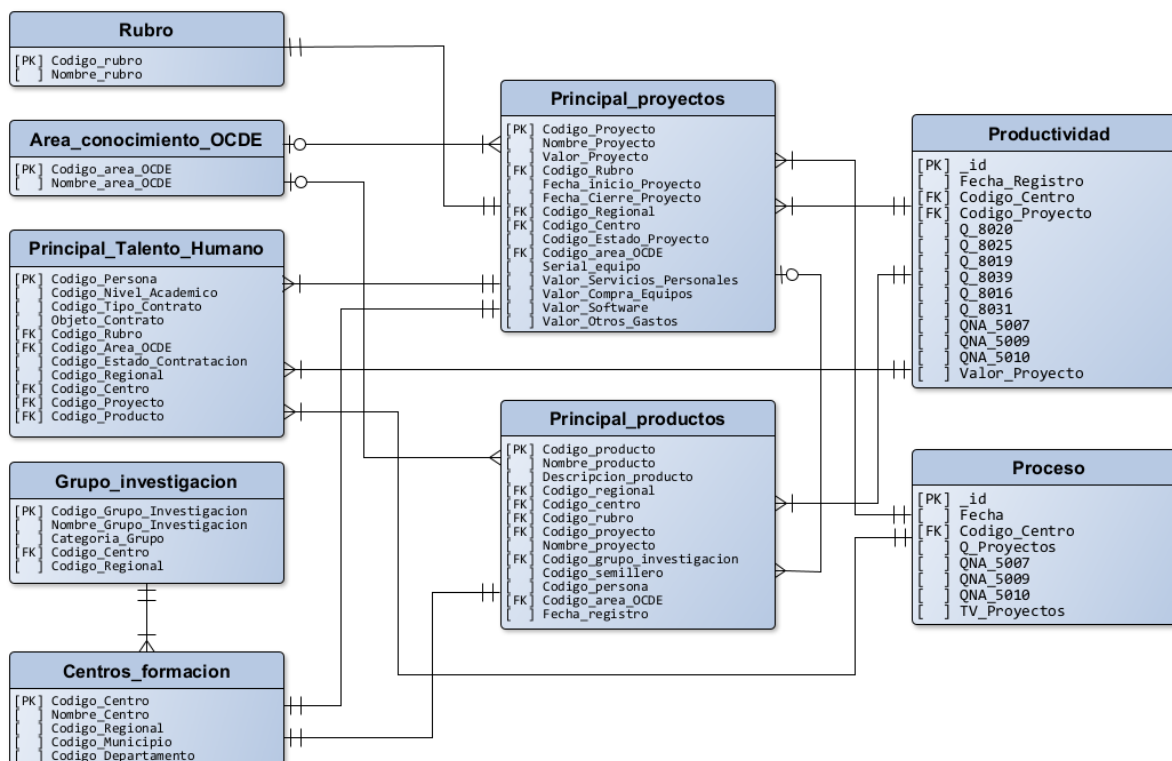


Figura 12.5. Modelo entidad-relación para dimensiones de productividad y procesos de los centros de formación SENA simulados con datos sintéticos.

12.3.2 Generación de datos pseudoaleatorios. Como mencionado anteriormente, para la generación de datos sintéticos pseudoaleatorios para cada una de las variables correspondientes al modelo entidad-relación, se crearon utilizando R.



RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

docs Tipología_Productos Tipo_Contratacion Regionales Principales_Productos Codigo.R Codigo.R

	Codigo_Proyecto	Nombre_Proyecto	Valor_Proyecto	Codigo_Rubro	Fecha_Inicio_Proyecto	Fecha_Cierre_Proyecto	Codigo_Regional	Codigo_Centro	Codigo_Estado_Proyecto	Codigo_area_OCDE
1	10651	Corporación de tecnología Carahsoft	147801416	76	2019-11-25	2019-05-22	5	9274	3	
2	8327	Tecnología SilkRoad	100264266	61	2014-01-24	2015-01-07	5	9164	3	
3	5549	Corporación de tecnología Carahsoft	235580787	30	2017-06-13	2019-05-17	5	9479	4	
4	7624	Tecnología Trek Wheel	50434232	47	2021-11-25	2014-04-04	5	9419	4	
5	2369	Grupo Techcess	21930280	73	2019-11-10	2012-09-27	5	9483	3	
6	5478	Aspect Technology Inc	412231127	27	2018-10-10	2015-02-15	5	9329	3	
7	8939	Tecnología Trek Wheel	191597410	80	2014-09-13	2011-01-13	5	9297	3	
8	1028	Infosys	45734696	35	2012-01-17	2020-02-09	5	9461	3	
9	1732	Micron Technology Inc	497082564	81	2015-04-24	2019-09-16	5	9445	3	
10	3479	Falcon Tech	308855993	72	2017-11-05	2019-12-21	5	9142	4	
11	11033	Señales de alta tecnología	99623441	81	2020-07-21	2017-05-15	5	9487	4	
12	11490	DuPontmakers Technology Inc	166624849	47	2019-04-12	2016-07-22	5	9312	4	
13	7520	PTC	438635920	79	2020-09-27	2022-07-11	5	9417	3	
14	445	Kingston Technology Co Inc	459763526	82	2021-12-29	2019-04-10	5	9120	4	
15	7669	Infosys	284382483	59	2011-09-11	2022-02-02	5	9532	3	
16	7349	NetStandard Inc	230668625	36	2014-05-23	2020-06-16	5	9463	4	
17	3436	Corporación de tecnología Carahsoft	431271879	70	2019-07-20	2017-05-29	5	9140	4	
18	10569	Aspect Technology Inc	199256297	70	2016-05-22	2014-07-13	5	9429	4	

Showing 1 to 18 of 500 entries, 15 total columns

Figura 12.6. Organización de base de base de datos y creación de datos sintéticos



13 RESULTADOS

13.1 Análisis exploratorio de los datos



El principal objetivo de este análisis exploratorio de datos (AED) de esta investigación es analizar y comprender a partir de la generación de diferentes gráficos estadísticos y cálculo de determinadas métricas, patrones de comportamiento y características conjuntas de las actividades de ACTI de los datos pseudoaleatorios generados para distintas métricas de indicadores. Estos son algunos de estos.

Como una etapa inicial de análisis de los datos, se realizaron extracciones de informaciones que nos permitiera obtener datos estadísticos, así como de variables correspondientes al modelo ER. Para Para nuestros análisis, dentro del modelo relacional cuatro (4) bases de datos de estudio, principal_proyecto, principal_talento_humano, principal_productos y principal_equipos que se describen en la Tabla 4, esto a raíz de su alta relacionamiento con tablas adicionales, y así permitir la generación de métricas en modelos de analítica de datos.

13.1.1. base de datos principal_proyecto

En la tabla a continuación se documenta las diferentes características de los principales proyectos a nivel de SENA - Regional Antioquia, tal como se muestra en la Tabla 13.1.

Codigo_Proyecto	Nombre_Proyecto	Valor_Proyecto	Codigo_Rubro	Fecha_inicio_Proyecto	Fecha_Cierre_Proyecto	Codigo_Regional	Codigo_Centro	Codigo_Estado_Pi
10651	Corporación de tecnología Carahsoft	147801416	65	2010-04-30	2017-11-06	5	9501	
8327	Tecnología SilkRoad	100264266	66	2012-02-13	2022-05-04	5	9503	
5549	Corporación de tecnología Carahsoft	235580787	82	2017-11-25	2021-07-04	5	9202	
7624	Tecnología Trek Wheel	50434232	23	2022-03-29	2011-11-30	5	9206	
2369	Grupo Techcess	21930280	66	2015-02-25	2011-07-31	5	9127	

Tabla 13.1 Información de tabla con sus respectivos códigos de variable.

En la Tabla 13.2, que sigue a  continuación, se describe la cantidad de variables que en este caso son 15, con 500 filas de datos sintéticos.

#	Column	Non-Null Count	Dtype
0	Codigo_Proyecto	500 non-null	int64
1	Nombre_Proyecto	500 non-null	object
2	Valor_Proyecto	500 non-null	int64
3	Codigo_Rubro	500 non-null	int64
4	Fecha_inicio_Proyecto	500 non-null	datetime64[ns]
5	Fecha_Cierre_Proyecto	500 non-null	datetime64[ns]
6	Codigo_Regional	500 non-null	int64
7	Codigo_Centro	500 non-null	int64
8	Codigo_Estado_Proyecto	500 non-null	int64
9	Codigo_area_OCDE	500 non-null	int64
10	Serial_equipo	500 non-null	int64
11	Valor_Servicios_Personales	500 non-null	int64
12	Valor_Compra_Equipos	500 non-null	int64
13	Valor_Software	500 non-null	int64
14	Valor_Otros_Gastos	500 non-null	int64

dtypes: datetime64[ns](2), int64(12), object(1)
memory usage: 58.7+ KB

Tabla 13.2. Características de la información registrada para principal_proyectos con datos sintéticos

En la Figura 13.1. se puede observa en gráficas de series temporales la evolución de gestión de proyectos de inicio (Figura 13.1.a) así como el fin del proyecto (Figura 13.1.b).

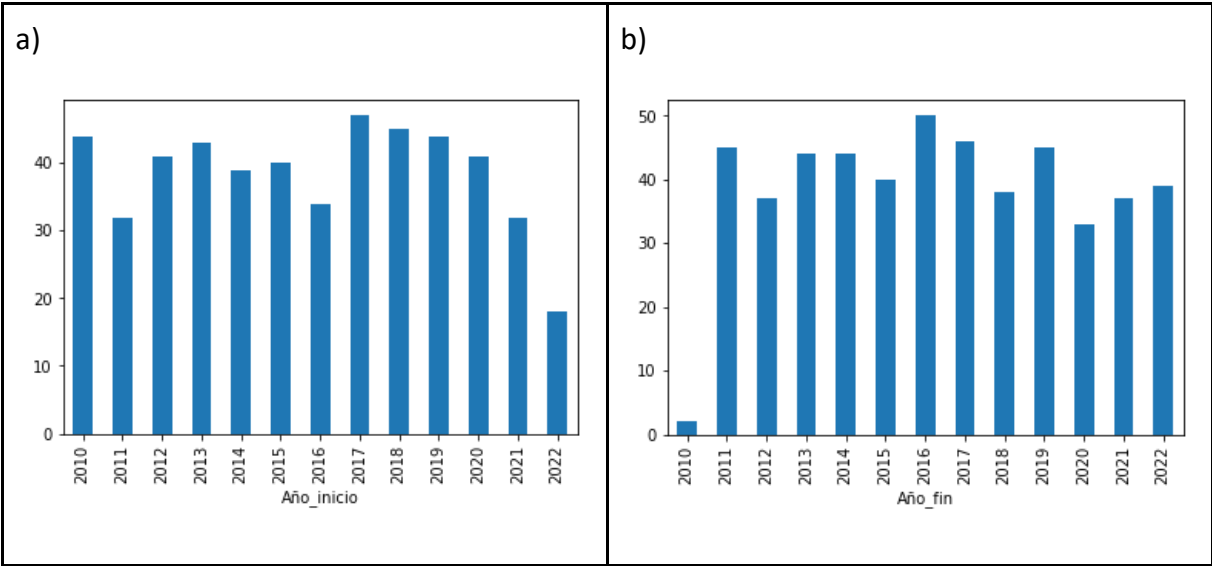


Figura 13.1. Serie temporal de inicio y fin de proyectos para los años 2010 al 2022 a) Inicio de proyecto; b) Fin de proyecto.



Tabla 13.4. a) Cantidad de variables, filas y su tipología; b) tabla de datos alterna a principal_talento_humano, que describe el código respectivo en este caso a Nivel Académico.

13.1.3. base de datos principal_productos

En esta tabla se documenta las diferentes características de productos de ciencia y tecnología generados a nivel de SENA - Regional Antioquia, de manera sintética, tal como se muestra en la Tabla 13.5

	Codigo_Producto	Nombre_Producto	Descripcion_producto	Codigo_Regional	Codigo_Centro	Codigo_Rubro	Codigo_Proyecto	Nombre_Proy
0	8039	Articulo_Sometido_en_revista_Indexada	NaN	5	9549	82	8602	
1	8029	Productos_tecnológicos_certificados_o_validados	NaN	5	9202	65	6465	
2	8019	Articulo_Sometido_en_revista_Indexada	NaN	5	9504	82	11205	
3	8016	Libros_resultados_de_investigación	NaN	5	9504	65	10481	
4	8020	Apropiación_Social	NaN	5	9204	65	1770	
5	8013	Apropiación_Social	NaN	5	9203	66	4372	
6	8031	Productos_tecnológicos_certificados_o_validados	NaN	5	9501	23	12075	
7	8016	Libros_resultados_de_investigación	NaN	5	9206	66	1012	
8	8012	Divulgación_Pública_de_la_CTeI_Publicaciones_e...	NaN	5	9203	66	3647	

Tabla 13.5. Contenido de tabla principal_productos con sus respectivos códigos de variable.



En la Tabla 13.6, que sigue a continuación, se describe la cantidad de variables que en este caso son 11, con 500 filas de datos sintéticos.

#	Column	Non-Null Count	Dtype
0	Codigo_Producto	500 non-null	int64
1	Nombre_Producto	500 non-null	object
2	Descripcion_producto	0 non-null	float64
3	Codigo_Regional	500 non-null	int64
4	Codigo_Centro	500 non-null	int64
5	Codigo_Rubro	500 non-null	int64
6	Codigo_Proyecto	500 non-null	int64
7	Nombre_Proyecto	500 non-null	int64
8	Codigo_Grupo_Investigacion	500 non-null	object
9	Codigo_Semillero	500 non-null	int64
10	Codigo_persona	500 non-null	int64
11	Codigo_area_OCDE	500 non-null	int64
12	Fecha_registro	500 non-null	datetime64[ns]

dtypes: datetime64[ns](1), float64(1), int64(9), object(2)
memory usage: 50.9+ KB

Tabla 13.6. Características de la información registrada para principal_productos con datos sintéticos

En la FIGURA 13.2. se puede observar la cantidad de productos científicos categorizados por SENA y el Ministerio de Ciencia Tecnología e Innovación. En esta se observa una distribución casi constante en la producción de productos, exceptuando a artículo sometido en revista científica.

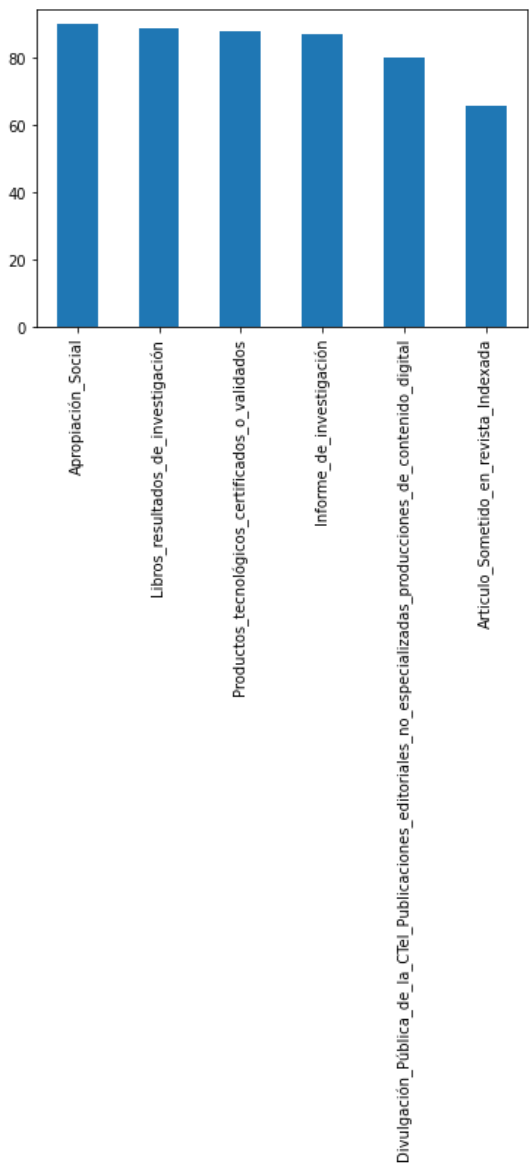


Figura 13.2. Frecuencia de cantidad de códigos por principal_productos.



13.1.4. Determinación geoespacial de Centros de formación SENA - Regional Antioquia.

Para la generación de mapas geoespaciales, se creó una base de datos alterna que dieran la posición de latitud y longitud para los dieciséis (16) diferentes centros de formación del SENA regional Antioquia. Para ello se hizo uso de la librería Folium de python, que permitió dar esta visualización de manera más adecuada, como se observa en la Figura 13.3. En este mapa, se puede observar la densidad de centros de formación a lo largo del departamento de Antioquia, donde claramente se destaca la alta densidad de centros en torno a Medellín así como de municipios aledaños.

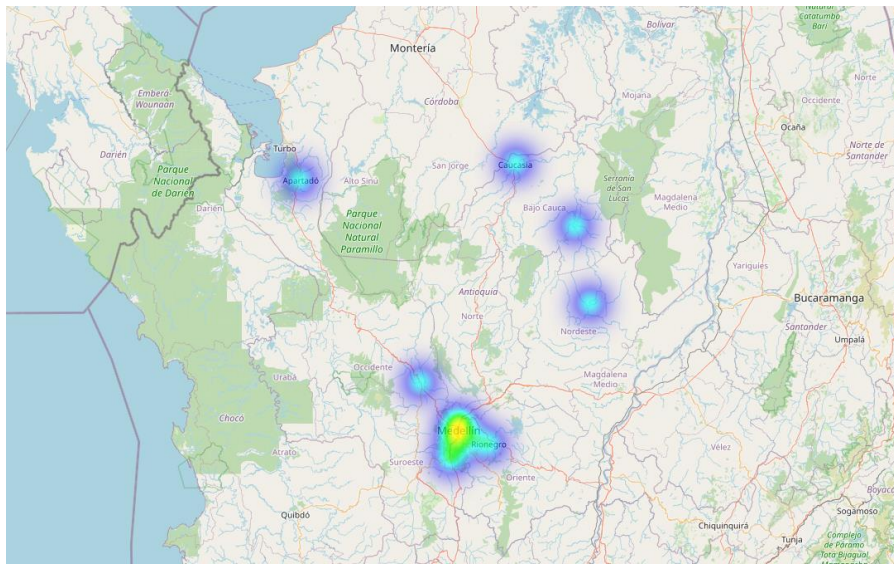


Figura 13.3. Mapa de posición geoespacial de centros de formación



13.2. Generación de mapas

Geoespaciales

Para la generación de indicadores de productividad y procesos de gestión en centros de formación, se procedió a realizar una estrategia de aprendizaje profundo no supervisada, tales como K-means y PCA, que permitieron generar clasificación de los centros en términos de productividad académica y gestión de procesos en alta, media y baja respectivamente. Para ello, se hizo uso de los datos preprocesados en R de *Principal_Proyectos* y *Principal_Productos*, como se observa en la Tabla 11.1 y Tabla 11.2. Para las etapas de clusterización se tuvieron en cuenta las etapas que se ilustran en la Figura 11.2. así como de los índices de validación internos y externos de los mismos.

Posteriormente a este proceso de clasificación, se procede a una etapa de visualización geospacial de alertas de alto, medio y bajo sea esto de *principal_proyectos* y *Principal_productos*.

13.2.1. Mapas geoespaciales para productividad de centros formación

Como mencionamos anteriormente, la generación de estos mapas se realiza en dos etapas: una etapa de clasificación a partir de herramientas de aprendizaje no supervisado, como la clusterización y PCA, y posteriormente se realiza una etapa de visualización de estos mapas.

13.2.1.1 Kmeans y PCA.

Para la selección de la cantidad de clusters adecuados, se aplicó el método del codo, que se visualiza en la Figura 13.4. Debido a nuestro interés de determinar un clasificador de tres alertas de alto, medio y bajo, se opta a seleccionar $k = 3$, que es también valido desde el punto de vista del método del codo.

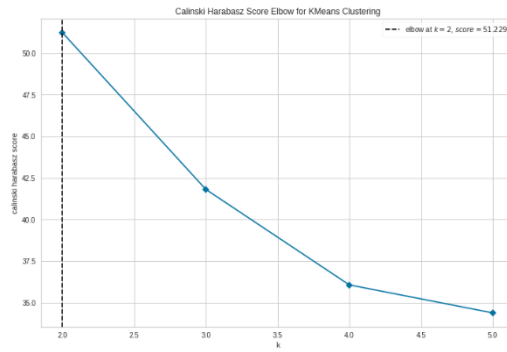


Figura 13.4. Método del codo para la determinación de cantidad de clusters.

Externos						Internos	
Indicador	Homogeneidad	Compleitud	Índice Rand	Medida V	Silueta	Calinski - Harabasz	Davies - Bouldin
Productividad	0,197	1,00	0	0,330	0,171	41,810	1,891

Tabla 13.7. Métricas de cluster.

Se determinan las métricas de cluster tanto internas como externas que se muestran en la Tabla 13.7 En se muestra la tendencia aleatoria de estos clusters, como consecuencia de la generación aleatoria de los datos. Se anexan códigos de proceso de clusterización a partir de K-Means (Ver Anexos).

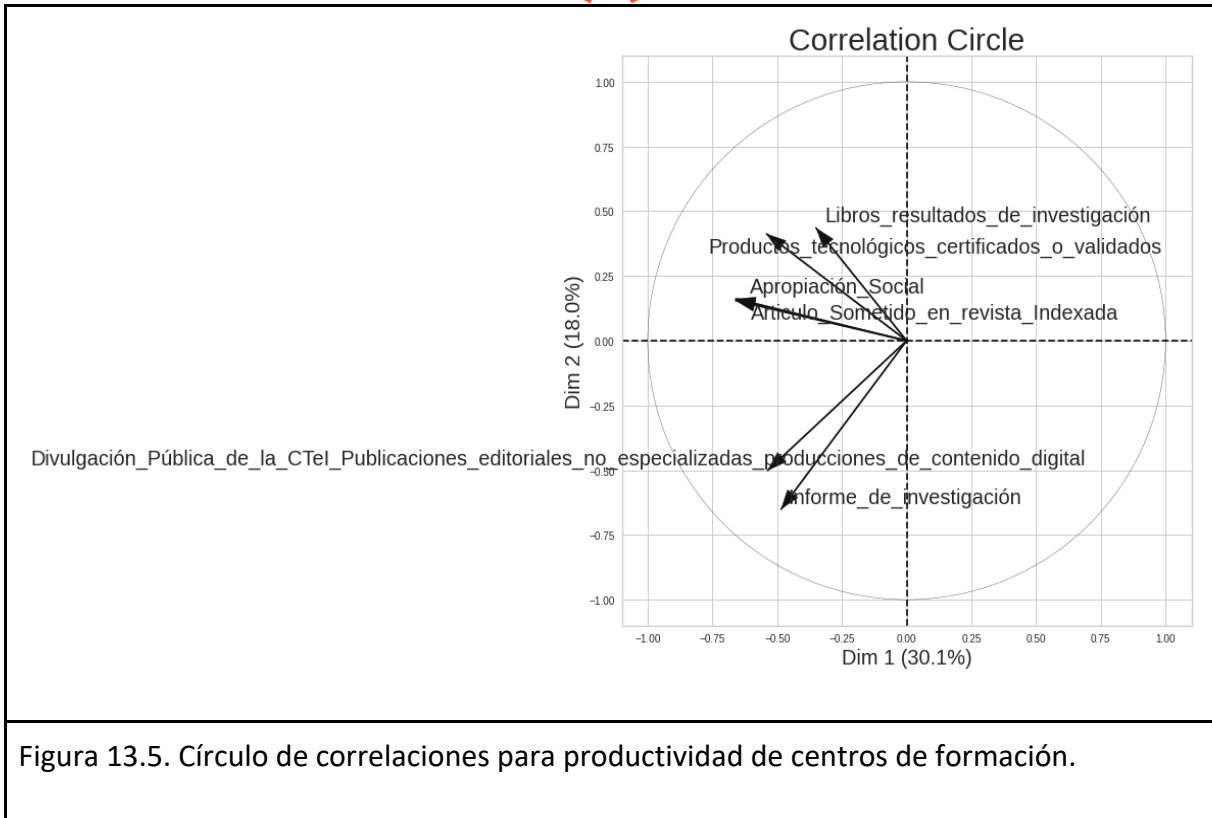


Figura 13.5. Círculo de correlaciones para productividad de centros de formación.

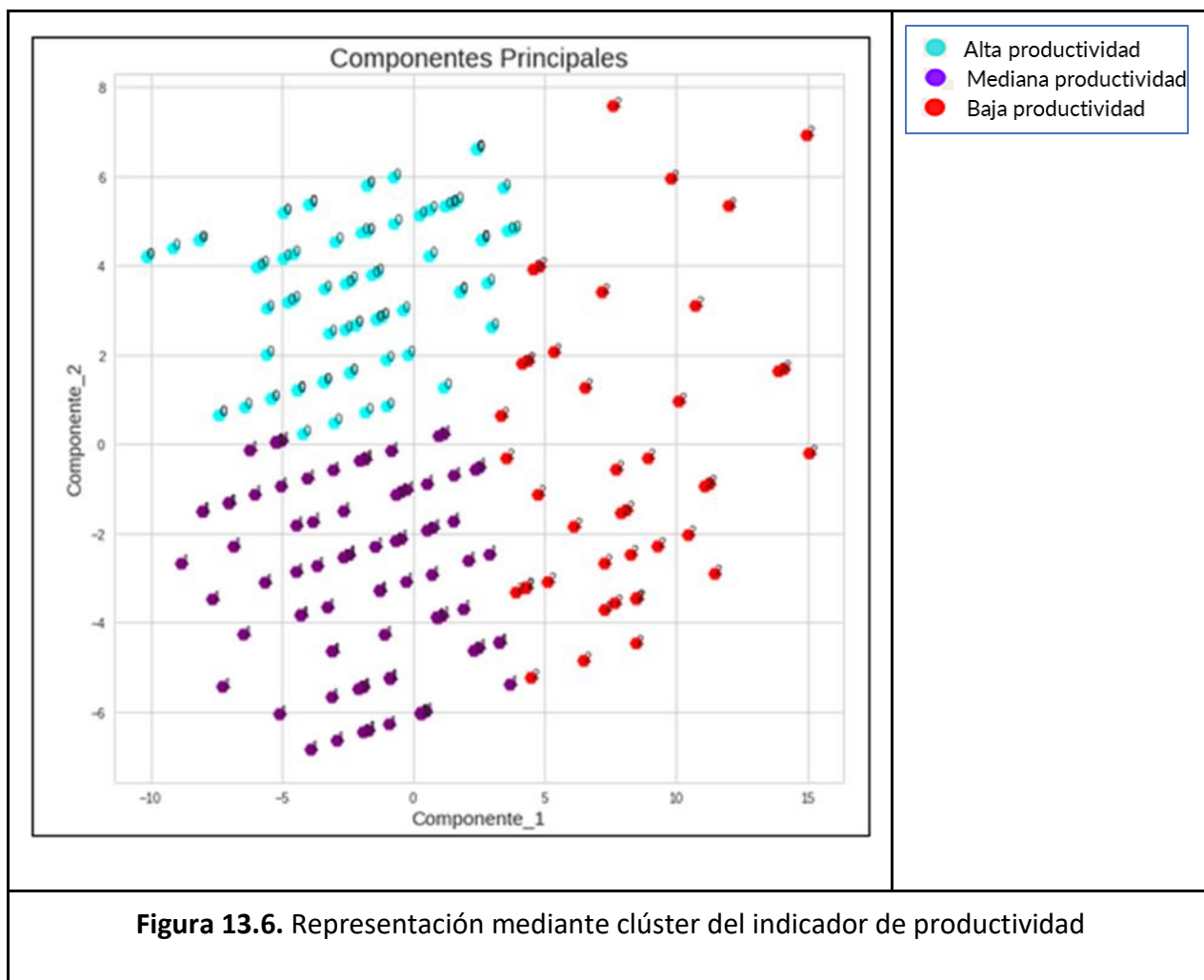
Posteriormente, a estos procesos de clasificación, se procede a realizar PCA para la reducción de dimensionalidad, y así proceder, de acuerdo a ciertas características del gráfico de correlación, encontrar un rango de categorías entre alta, media y baja productividad. La Figura 13.5 muestra el diagrama de correlación para el caso de productividad. En esta se puede observar que en la separación de cuadrantes de cada variable respecto a los productos generados por cada centro de formación. Existe mayor cantidad de productos en el segundo cuadrante del diagrama de correlaciones, lo que permite asociar que este cuadrante corresponde a un índice alto de producción científica, y que para el tercer cuadrante están únicamente dos tipos de productos. Esta hipótesis generada, es relativa, ya que depende netamente del interés institucional que permite evaluar a sus centros de formación entre alto, medio o bajo de acuerdo criterios generados por el ministerio de las ciencias, pero que para esta etapa hemos dado igualdad de peso.

En la Figura 13.6. se puede observar las gráficas de componentes principales de cada uno de los clusters generados por K-means. Si hacemos una comparación entre el círculo de correlaciones de la Figura 13.5 y Figura 13.6 podemos generar tres tipos de clusters, que se pueden clasificar entre alta, media y baja productividad, esto basado en la cantidad de

vectores del círculo de correlación. De este modo hemos generado una serie de alertas de productividad científica para los centros de formación de estudio.



este modo hemos generado una serie de alertas de productividad científica para los centros de formación de estudio.

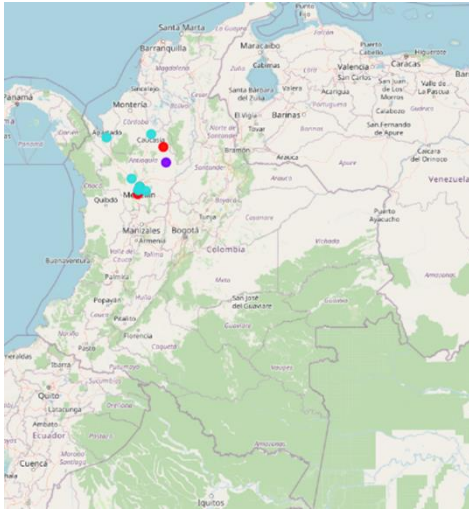
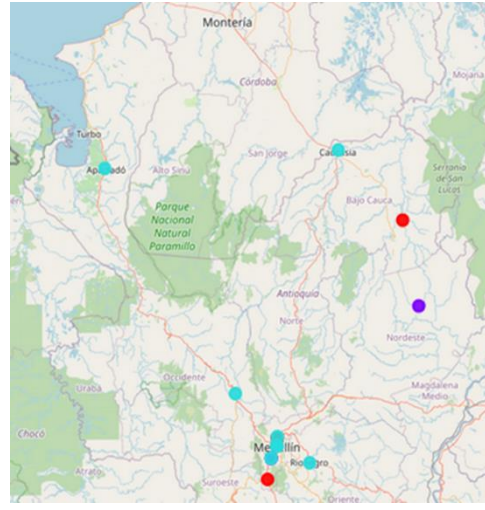


13.2.1.2 Mapas Geospaciales.

Generado la clasificación de centros de formación, se procede a generar los mapas georreferenciados que permiten clasificar cada uno de los centros de formación en medio, alto y bajo, tal como se muestra en la Figura 13.7.

a)

b)



c)

d)

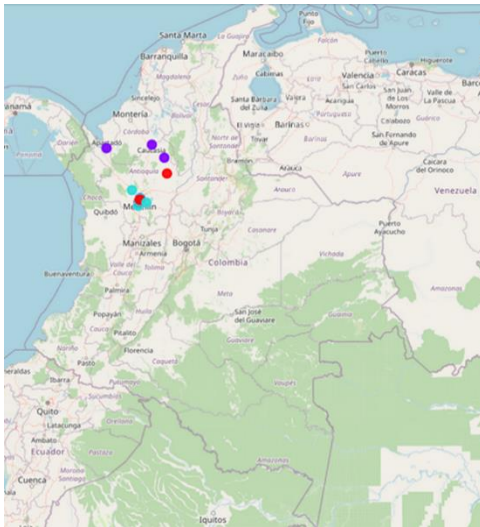
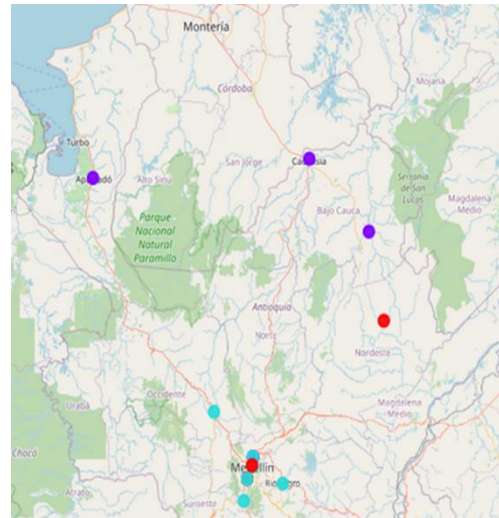


Figura 13.7. Representación geoespacial clúster del indicador productividad a) y b) 2018; y c) y d) 2021.

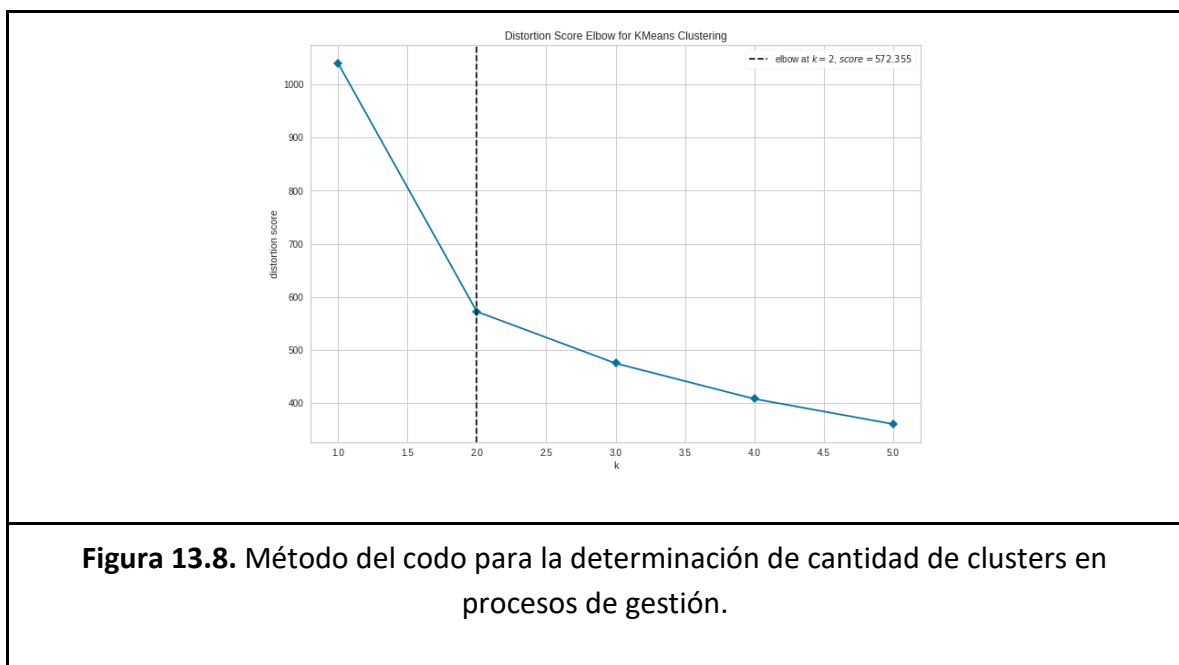


13.2.2. Mapas geospaciales para gestión de procesos de centros formación

Como mencionamos anteriormente, la generación de estos mapas se realiza en dos etapas: una etapa de clasificación a partir de herramientas de aprendizaje no supervisado, como la clusterización y PCA, y posteriormente se realiza una etapa de visualización de estos mapas.

13.2.2.1 Kmeans y PCA.

Para la selección de la cantidad de clusters adecuados, se aplicó el método del codo, que se visualiza en la Figura 13.8. Debido a nuestro interés de determinar un clasificador de tres alertas de alto, medio y bajo, se opta a seleccionar $k = 3$, que es también valido desde el punto de vista del método del codo.



Externos						Internos	
Indicador	Homogeneidad	Compleitud	Índice Rand	Medida V	Silueta	Calinski - Harabasz	Davies - Bouldin



Procesos	0,183	1,00	0	0,309	0,301	121,774	1,328
----------	-------	------	---	-------	-------	---------	-------

Tabla 13.8. Métricas de cluster.

Se determinan las métricas de cluster tanto internas como externas. Que se muestran en la Tabla 13.8 En se muestra la tendencia aleatoria de estos clusters, como consecuencia de la generación aleatoria de los datos. Se anexan códigos de proceso de clusterización a partir de K-Means (Ver Anexos).

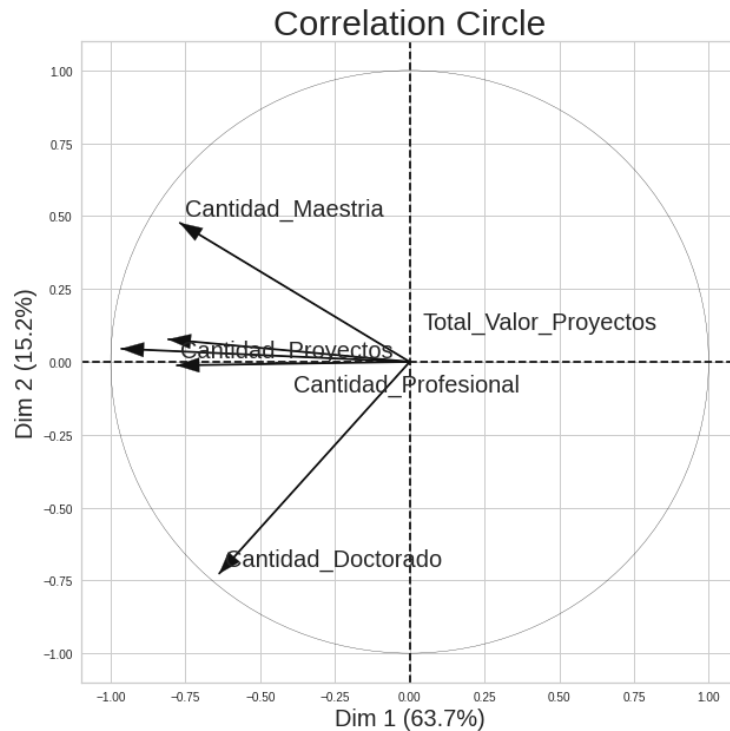
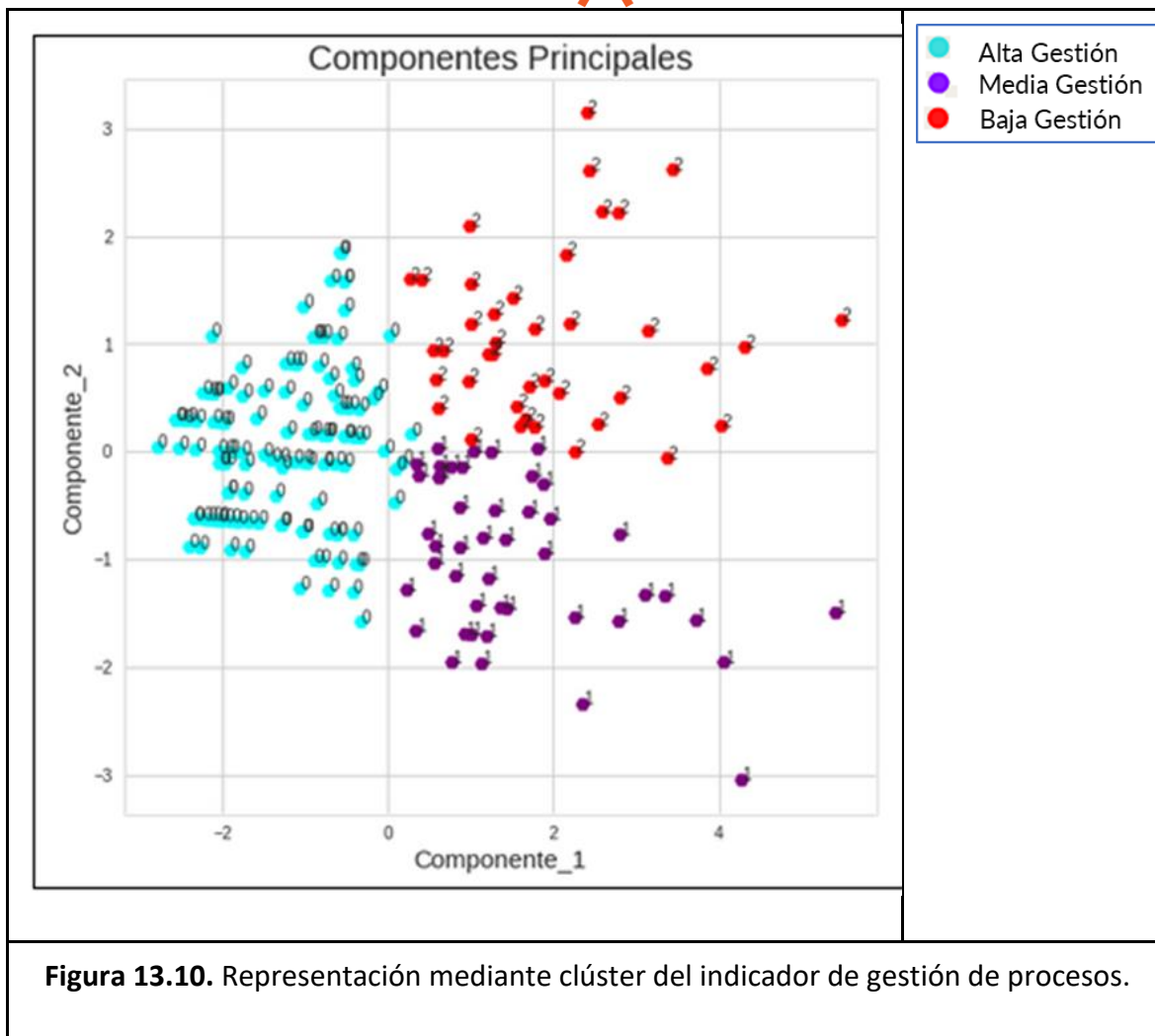


Figura 13.9. Círculo de correlaciones para procesos de centros de formación.



Posteriormente, a estos procesos de clasificación, se procede a realizar PCA para la reducción de dimensionalidad, y así proceder, de acuerdo a ciertas características del gráfico de correlación, encontrar un rango de categorías entre alta, media y baja gestión. La figura 13.9 muestra el diagrama de correlación para el caso de gestión de procesos. En esta se puede observar que, en la separación de cuadrantes de cada variable respecto a la cantidad de talento humano formado en los diferentes niveles educativo, así como con la cantidad de proyectos por cada centro de formación. Existe mayor cantidad de vectores en el segundo cuadrante del círculo de correlaciones, lo que permite asociar que este cuadrante corresponde a un índice de alta gestión, y que para el tercer cuadrante están únicamente un tipo de vector que es la variable cantidad de doctores que por cantidad correspondería a media gestión. Esta hipótesis generada, es relativa, ya que depende netamente del interés institucional que permite evaluar a sus centros de formación entre alto, medio o bajo de acuerdo criterios generados a nivel institucional y gubernamental, pero que para esta etapa hemos dado igualdad de peso.

En la Figura 13.10. se puede observar las gráficas de componentes principales de cada uno de los clusters generados por K-means. Si hacemos una comparación entre el círculo de correlaciones de la Figura 13.9 y Figura 10 podemos generar tres tipos de clusters, que se pueden clasificar entre alta, media y baja gestión, esto basado en la cantidad de vectores del círculo de correlación. De este modo hemos generado una serie de alertas de gestión de procesos para los centros de formación en estudio.



13.2.2.2 Mapas Geospaciales.

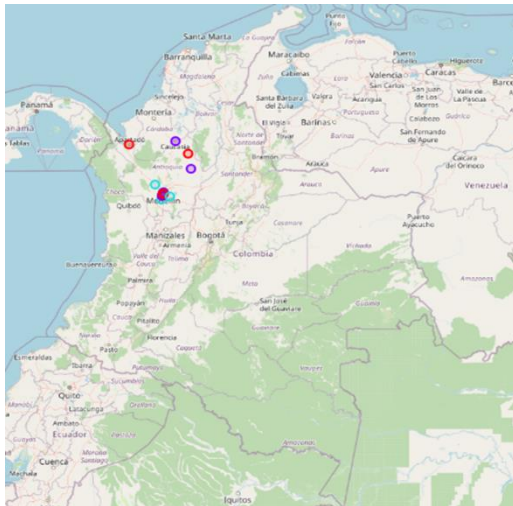
Generado la clasificación de centros de formación, se procede a generar los mapas georreferenciados que permiten clasificar cada uno de los centros de formación en alta, media y baja gestión, tal como se muestra en la Figura 13.11.

a)

b)



c)



d)

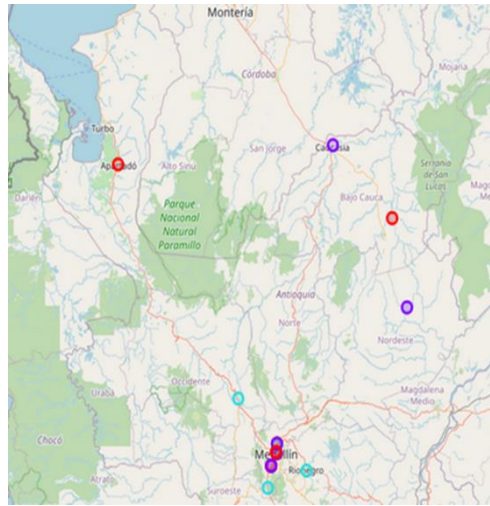


Figura 13.11. Representación geoespacial clúster del indicador gestión de procesos a) y b) 2018; y c) y d) 2021.



13.3. Modelos predictivos utilizando analítica de datos para medición de gestión y productividad por centros – Algoritmo KNN

Otros de los objetivos del proyecto fue generar un modelo predictivo que permita predecir, bajo las variables descritas en las TablaXX y la TablaXX, si el centro de formación entra en una categoría de alta, media o baja gestión o productividad respectivamente. Para ellos se hace implementación del algoritmo K-NN en R statistics

13.3.1. Modelos predictivos para medición de gestión de procesos en centros de formación – Algoritmo KNN

Se realiza el escalamiento de los datos

```
#=====
#Escalamiento de datos
#=====
datos_TPProceso <- scale(data_procesos[,c(2:6)])
View(datos_TPProceso)
#Se adiciona nuevamente la columna de etiqueta Centro-Año, Cluster y Gestion
datos_scale_TPProceso <- as.data.frame(cbind(data_procesos$Centro_Año, datos_TPProceso, data_procesos$cluster,
data_procesos$Gestion))
View(datos_scale_TPProceso)
str(datos_scale_TPProceso)
dim(datos_scale_TPProceso)
datos_scale_TPProceso$Q_Proyectos <- as.numeric(datos_scale_TPProceso$Q_Proyectos)
datos_scale_TPProceso$QNA_5007 <- as.numeric(datos_scale_TPProceso$QNA_5007)
datos_scale_TPProceso$QNA_5009 <- as.numeric(datos_scale_TPProceso$QNA_5009)
datos_scale_TPProceso$QNA_5010 <- as.numeric(datos_scale_TPProceso$QNA_5010)
datos_scale_TPProceso$TV_Proyectos <- as.numeric(datos_scale_TPProceso$TV_Proyectos)
str(datos_scale_TPProceso)
names(datos_scale_TPProceso) <- names(data_procesos)
View(datos_scale_TPProceso)
row.names(datos_scale_TPProceso) <- datos_scale_TPProceso$Centro_Año
View(datos_scale_TPProceso)
proceso <- datos_scale_TPProceso[, c(-1,-7)]
View(proceso)
#=====
```

Figura 13.12. Escalamiento de datos.

Se transforma la variable Gestión en una variable factor de tres niveles

```
#=====
# La variable factor está indexada en la 6ª posición
#=====

proceso$Gestion <- as.factor(proceso$Gestion)
View(proceso)
str(proceso)
#=====
```



```
> str(proceso)
'data.frame': 208 obs. of 6 variables:
 $ Q_Proyectos : num -1.257 -0.362 -1.257 0.534 -0.362 ...
 $ QNA_5007 : num -1.103 -0.689 0.137 0.55 0.137 ...
 $ QNA_5009 : num -0.8238 -0.0502 -1.2106 0.3366 -0.437 ...
 $ QNA_5010 : num -1.108 -0.225 -0.225 0.658 -0.225 ...
 $ TV_Proyectos: num -0.126 -0.733 -0.206 -0.607 -1.573 ...
 $ Gestion : Factor w/ 3 levels "Alta Gestion",...: 2 2 2 1 2 3 2 2 3 2 ...
```

Figura 13.13 Transformación de variable Gestión en un factor de tres niveles.

Se obtiene una tabla, donde se evidencia la cantidad de observaciones por nivel de gestión

```
#=====
table(proceso$Gestion)
# Alta Gestion Baja Gestion Media Gestion
# 47 122 39
#=====
```

```
> table(proceso$Gestion)

Alta Gestion Baja Gestion Media Gestion
47 122 39
```

Figura 13.14 Cantidad de observaciones por nivel de gestión

```
#=====
# Porcentaje de centros
#=====
round(prop.table(table(proceso$Gestion)) * 100,digits = 1)
# Alta Gestion Baja Gestion Media Gestion
# 22.6 58.7 18.8
#=====
```

Figura 13.15. Porcentaje de centros por nivel de gestión



Dadas estas etapas, se genera el modelo de analítica de datos utilizando KNN. Con esto se obtienen la siguiente matriz de confusión.

```
> Validacion_modelo_procesos <- CrossTable(x = data_procesos_test_labels, y = data_procesos_test_predi, prop.chisq=TRUE)
```

Cell Contents				

Chi-square contribution				
N				
N / Row Total				
N / Col Total				
N / Table Total				

Total observations in Table: 62				

data_procesos_test_labels				
data_procesos_test_predi				
Alta Gestion				
Baja Gestion				
Media Gestion				
Row Total				

Alta Gestion				
Baja Gestion				
Media Gestion				
Row Total				

Baja Gestion				
Media Gestion				
Row Total				

Media Gestion				
Row Total				

Column Total				

Figura 13.16. Matriz de confusión para el algoritmo K-NN en R, con k=3



```
> con=confusionMatrix(testData[,ncol(testData)], data_procesos_test_pred1, mode = "everything")
> con
Confusion Matrix and Statistics

          Reference
Prediction  Alta Gestion Baja Gestion Media Gestion
Alta Gestion      16          0          1
Baja Gestion       1         34          0
Media Gestion      1          0          9

Overall Statistics

          Accuracy : 0.9516
          95% CI   : (0.865, 0.9899)
    No Information Rate : 0.5484
    P-value [Acc > NIR] : 1.495e-12

          Kappa : 0.9173

  Mcnemar's Test P-Value : NA

Statistics by Class:

               Class: Alta Gestion Class: Baja Gestion Class: Media Gestion
Sensitivity           0.8889           1.0000           0.9000
Specificity           0.9773           0.9643           0.9808
Pos Pred Value        0.9412           0.9714           0.9000
Neg Pred Value        0.9556           1.0000           0.9808
Precision             0.9412           0.9714           0.9000
Recall                0.8889           1.0000           0.9000
F1                    0.9143           0.9855           0.9000
Prevalence            0.2903           0.5484           0.1613
Detection Rate        0.2581           0.5484           0.1452
Detection Prevalence  0.2742           0.5645           0.1613
Balanced Accuracy     0.9331           0.9821           0.9404
```

Figura 13.17. Matriz de confusión para el algoritmo K-NN en R, con k=3, y sus respectivas métricas

Para nuestro caso la tabla de contingencia arrojada por la matriz de confusión consta de $N=62$ observaciones clasificadas en forma cruzada en las variables de fila y columna, que normalmente tendrían los mismos niveles, sin embargo, el estadístico de McNemar se usa para probar la hipótesis nula de simetría, es decir, que la probabilidad de que una observación se clasifique en celda $[i,j]$ es igual a la probabilidad de ser clasificado en la celda $[j,i]$, (mcnemar.test)

Para nuestro caso al obtener un P-Value: NA, es un indicio de que la clasificación está fallando, por lo tanto, debemos calcular cual es el óptimo k-vécin cercano en nuestra base de datos, como se describe a continuación.

Hasta el momento solo hemos utilizado un solo valor para k, 3 vecinos más cercanos. Sin embargo, podríamos ver la exactitud o *accuracy* para diferentes valores de k. Como

tenemos 146 valores en nuestra data de
tomando máximo 146 vecinos más



entrenamiento veremos los aciertos
cercanos.

```
#=====
k <- 1:146
resultado <- data.frame(k, precision = 0)

for(n in k){
  proceso_test_output_knn <- knn(train = trainData1,
                                cl = data_procesos_train_labels,
                                test = testData1,
                                k = n)

  resultado$precision[n] <- mean(proceso_test_output_knn == testData[,6])
}
resultado
library(extrafont)
#resultado %>%
ggplot(resultado,aes(x=k, y=precision))+
  labs(title = 'Evolución de la precision segun valor de k')+
  labs(subtitle = 'Precision asociados a valores distintos de k')+
  geom_line(color="red")+ geom_point()+
  scale_x_continuous(breaks=pretty(resultado$k, n = 10), name="k")+
  theme(legend.position="top", legend.text = element_text(colour="blue", size=6)) +
  geom_text(aes(label=k),hjust = 0, nudge_x = 0.05, color="blue",size=3)+
  theme(text=element_text(size=12, family="Calibri"))+
  theme_bw()
#=====
```

Figura 13.18. Fragmento de código modelo KNN para procesos de gestión en centros de formación.

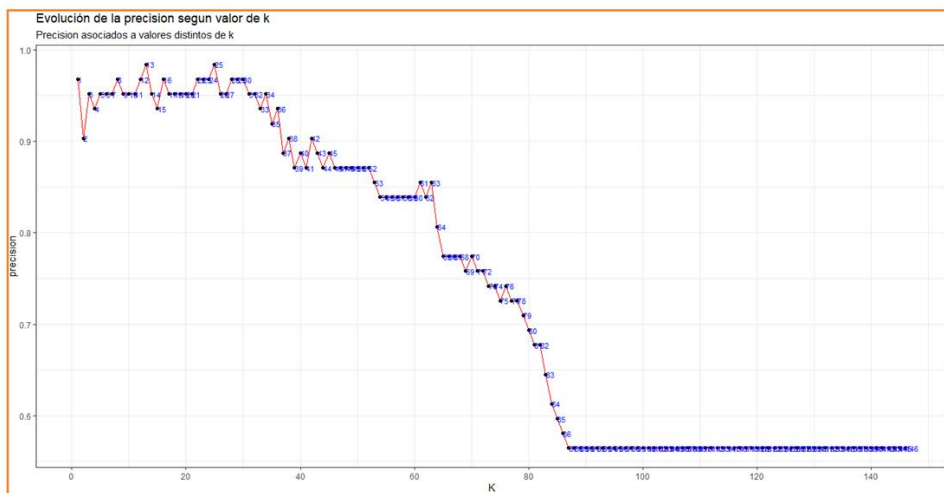


Figura 13.19. Precisión del modelo para diferentes valores de k



Como vemos, para este caso, a partir de cierto número de vecinos más cercanos el nivel de aciertos de nuestro algoritmo empieza a reducirse. Dependerá de cada caso el escoger el mejor “k” para nuestro modelo, para el caso se elige $k=2$

```
> con
Confusion Matrix and Statistics

Prediction      Reference
Alta Gestion    Alta Gestion Baja Gestion Media Gestion
Baja Gestion    15          0          2
Baja Gestion    1          33         1
Media Gestion    2          0          8

Overall Statistics

Accuracy : 0.9032
95% CI : (0.8012, 0.9637)
No Information Rate : 0.5323
P-Value [Acc > NIR] : 3.354e-10

Kappa : 0.8363

McNemar's Test P-value : 0.5724

Statistics by Class:

Class: Alta Gestion Class: Baja Gestion Class: Media Gestion
Sensitivity          0.8333          1.0000          0.7273
Specificity          0.9545          0.9310          0.9608
Pos Pred Value       0.8824          0.9429          0.8000
Neg Pred Value       0.9333          1.0000          0.9423
Precision            0.8824          0.9429          0.8000
Recall               0.8333          1.0000          0.7273
F1                   0.8571          0.9706          0.7619
Prevalence           0.2903          0.5323          0.1774
Detection Rate       0.2419          0.5323          0.1290
Detection Prevalence 0.2742          0.5645          0.1613
Balanced Accuracy     0.8939          0.9655          0.8440
```

Figura 13.20. Matriz de confusión para el algoritmo K-NN en R, con $k=2$, y sus respectivas métricas

Evaluación del rendimiento del modelo: métricas asociadas al clasificador K-NN, con $k=2$

```
> clasificador_metricas
Sensitivity Specificity Precision Recall F1 Accuracy
Metricas_del_Clasificador 0.9103128 0.9422794 0.9032258 0.9103128 0.9058236 0.9032258
```

Hemos clasificado 15 centros de Alta Gestion como Alta Gestion, lo que indica que está bien clasificado y 2 centros de Alta Gestion como de Media Gestion, lo que indica una incorrecta clasificación. A su vez analizando los centros en Baja Gestion la predicción en la clasificación arroja que 33 centros de Baja Gestion quedaron clasificados como de Baja Gestion lo cual indica que quedaron bien clasificados. Por otro lado, 1 centro siendo de Baja Gestion quedo clasificado como que fuera de Alta Gestion, asimismo otro centro siendo de Baja Gestion

quedo clasificado en Media Gestion. Por último 8 centros de Media Gestion fueron clasificados como de Media Gestion, lo cual indica una correcta clasificación y 2 centros de pertenecientes a Media Gestion quedaron clasificados como si fueran de Alta Gestion, lo cual indica que esta incorrectamente clasificado



13.3.2. Modelos predictivos para medición de productividad en centros de formación – Algoritmo KNN

Se realiza el escalamiento de los datos

```
#=====
# La variable factor está indexada en la 7ª posición
#=====

productividad$Nivel_Productividad <- as.factor(productividad$Nivel_Productividad)
view(productividad)
str(productividad)

#=====

> str(productividad)
'data.frame': 208 obs. of 7 variables:
 $ P_8020      : num -0.559 -0.559 -1.187 -1.187 2.582 ...
 $ P_8025      : num 0.855 -1.271 -1.271 -0.562 0.147 ...
 $ P_8019      : num -0.559 -0.559 -1.187 -1.187 2.582 ...
 $ P_8039      : num 0.855 -1.271 -1.271 -0.562 0.147 ...
 $ P_8016      : num -0.559 -0.559 -1.187 -1.187 2.582 ...
 $ P_8031      : num 0.855 -1.271 -1.271 -0.562 0.147 ...
 $ Nivel_Productividad: Factor w/ 3 levels "Alta Productividad",...: 1 2 2 2 1 2 2 3 2 2 ...
```

Figura 13.21. Escalamiento de datos.

Se transforma la variable Gestión en una variable factor de tres niveles

```
#=====
# La variable factor está indexada en la 7ª posición
#=====

productividad$Nivel_Productividad <- as.factor(productividad$Nivel_Productividad)
view(productividad)
str(productividad)

#=====
```



```
> str(productividad)
'data.frame': 208 obs. of 7 variables:
 $ P_8020      : num -0.559 -0.559 -1.187 -1.187 2.582 ...
 $ P_8025      : num 0.855 -1.271 -1.271 -0.562 0.147 ...
 $ P_8019      : num -0.559 -0.559 -1.187 -1.187 2.582 ...
 $ P_8039      : num 0.855 -1.271 -1.271 -0.562 0.147 ...
 $ P_8016      : num -0.559 -0.559 -1.187 -1.187 2.582 ...
 $ P_8031      : num 0.855 -1.271 -1.271 -0.562 0.147 ...
 $ Nivel_Productividad: Factor w/ 3 levels "Alta Productividad",...: 1 2 2 2 1 2 2 3 2 2 ...
```

Figura 13.22. Transformación de variable Nivel_Productividad en un factor de tres niveles.

Se obtiene una tabla, donde se evidencia la cantidad de observaciones por nivel de productividad.

```
#=====
table(productividad$Nivel_Productividad)
# Alta Productividad Baja Productividad Media Productividad
#          39          112          57
#=====

> round(prop.table(table(productividad$Nivel_Productividad)) * 100,digits = 1)

Alta Productividad Baja Productividad Mediana Productividad
          18.8          53.8          27.4
```

Figura 13.23. Cantidad de observaciones por nivel de gestión

```
> round(prop.table(table(productividad$Nivel_Productividad)) * 100,digits = 1)

Alta Productividad Baja Productividad Mediana Productividad
          18.8          53.8          27.4
```

Figura 13.24. Porcentaje de centros por nivel de productividad



Dadas estas etapas, se genera el modelo de analítica de datos utilizando KNN. Con esto se obtienen la siguiente matriz de confusión.

```
> library(gmodels)
> validacion_modelo_productividad <- CrossTable(x = data_productividad_test_labels, y = data_productividad_test_pred1, prop.chisq=TRUE)
```

cell contents				
	N			
Chi-square contribution				
N / Row Total				
N / Col Total				
N / Table Total				

Total observations in Table: 74

data_productividad_test_labels	data_productividad_test_pred1			Row Total
	Alta Productividad	Baja Productividad	Mediana Productividad	
Alta Productividad	2	8	8	18
	2.211	1.157	1.034	
	0.111	0.444	0.444	0.243
	0.667	0.167	0.348	
	0.027	0.108	0.108	
Baja Productividad	1	35	1	37
	0.167	5.042	9.587	
	0.027	0.946	0.027	0.500
	0.333	0.729	0.043	
	0.014	0.473	0.014	
Mediana Productividad	0	5	14	19
	0.770	4.353	11.095	
	0.000	0.263	0.737	0.257
	0.000	0.104	0.609	
	0.000	0.068	0.189	
Column Total	3	48	23	74
	0.041	0.649	0.311	

Figura 13.25. Matriz de confusión para el algoritmo K-NN en R, con k=3



```
> con=confusionMatrix(testData[,ncol(testData)], data_productividad_test_pred1, mode = "everything")
> con
Confusion Matrix and Statistics

              Reference
Prediction    Alta Productividad Baja Productividad Mediana Productividad
Alta Productividad      2              8              8
Baja Productividad      1             35              1
Mediana Productividad    0              5             14

Overall Statistics

          Accuracy : 0.6892
          95% CI   : (0.571, 0.7917)
    No Information Rate : 0.6486
    P-value [Acc > NIR] : 0.274045

          Kappa : 0.4696

  McNemar's Test P-value : 0.001076

Statistics by Class:

               Class: Alta Productividad Class: Baja Productividad Class: Mediana Productividad
Sensitivity                0.66667                0.7292                0.6087
Specificity                0.77465                0.9231                0.9020
Pos Pred value             0.11111                0.9459                0.7368
Neg Pred value             0.98214                0.6486                0.8364
Precision                  0.11111                0.9459                0.7368
Recall                    0.66667                0.7292                0.6087
F1                        0.19048                0.8235                0.6667
Prevalence                 0.04054                0.6486                0.3108
Detection Rate             0.02703                0.4730                0.1892
Detection Prevalence       0.24324                0.5000                0.2568
Balanced Accuracy          0.72066                0.8261                0.7553
```

Figura 13.26. Matriz de confusión para el algoritmo K-NN en R, con k=3, y sus respectivas métricas

Hasta el momento solo hemos utilizado un solo valor para k, 3 vecinos más cercanos. Sin embargo, podríamos ver la exactitud o *accuracy* para diferentes valores de k. Como tenemos 134 valores en nuestra data de entrenamiento veremos los aciertos tomando máximo 134 vecinos más cercanos.



```
#=====
k <- 1:134
resultado <- data.frame(k, precision = 0)

for(n in k){
  productividad_test_output_knn <- knn(train = trainData1,
                                       cl = data_productividad_train_labels,
                                       test = testData1,
                                       k = n)

  resultado$precision[n] <- mean(productividad_test_output_knn == testData[,7])
}
resultado
library(extrafont)
#resultado %>%
ggplot(resultado,aes(x=k, y=precision))+
  labs(title = 'Evolución de la precision segun valor de k')+
  labs(subtitle = 'Precision asociados a valores distintos de k')+
  geom_line(color="red")+ geom_point()+
  scale_x_continuous(breaks=pretty(resultado$k, n = 10), name="K")+
  theme(legend.position="top", legend.text = element_text(colour="blue", size=6)) +
  geom_text(aes(label=k),hjust = 0, nudge_x = 0.05, color="blue",size=3)+
  theme(text=element_text(size=12, family="Calibri"))+
  theme_bw()
#=====
```

Figura 13.27. Fragmento de código modelo KNN para productividad en centros de formación.

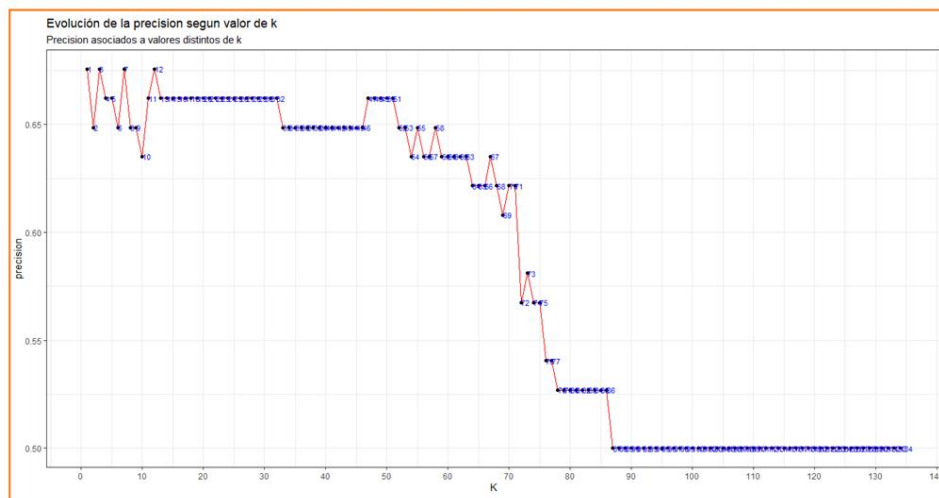


Figura 13.28. Precisión del modelo para diferentes valores de k

Como vemos, para este caso, a partir de cierto número de vecinos más cercanos el nivel de aciertos de nuestro algoritmo empieza a reducirse. Dependerá de cada caso el escoger el

mejor “k” para nuestro modelo, para  nuestro caso k=3, ofrece la mejor precisión para nuestro modelo.

```
> con
Confusion Matrix and Statistics

              Reference
Prediction    Alta Productividad Baja Productividad Mediana Productividad
Alta Productividad      1           8           9
Baja Productividad      0          36           1
Mediana Productividad   0           7          12

Overall Statistics

          Accuracy : 0.6622
          95% CI   : (0.5428, 0.7681)
    No Information Rate : 0.6892
    P-value [Acc > NIR] : 0.738

          Kappa : 0.4133

  McNemar's Test P-Value : 8.288e-05

Statistics by Class:

               Class: Alta Productividad Class: Baja Productividad Class: Mediana Productividad
Sensitivity              1.00000             0.7059             0.5455
Specificity              0.76712             0.9565             0.8654
Pos Pred Value            0.05556             0.9730             0.6316
Neg Pred Value            1.00000             0.5946             0.8182
Precision                 0.05556             0.9730             0.6316
Recall                   1.00000             0.7059             0.5455
F1                       0.10526             0.8182             0.5854
Prevalence                0.01351             0.6892             0.2973
Detection Rate            0.01351             0.4865             0.1622
Detection Prevalence      0.24324             0.5000             0.2568
Balanced Accuracy         0.88356             0.8312             0.7054
```

Figura 13.29. Matriz de confusión para el algoritmo K-NN en R, con k=7, y sus respectivas métricas

Evaluación del rendimiento del modelo: métricas asociadas al clasificador K-NN, con k=3

```
> clasificador_metricas
               Sensitivity Specificity Precision  Recall      F1 Accuracy
Metricas_del_Clasificador  0.9103128  0.9422794 0.9032258 0.9103128 0.9058236 0.9032258
```

Hemos clasificado 2 centros de Alta Productividad como Alta Productividad, lo que indica que está bien clasificado, 8 centros que siendo de Alta Productividad como de Baja Productividad y 8 centros que siendo de Alta productividad como de Mediana Productividad, lo que indica una incorrecta clasificación para estos. A su vez analizando los centros de Baja Productividad tenemos que la predicción en la clasificación arroja que 35 centros de Baja Productividad quedaron clasificados como de Baja Productividad lo cual indica que quedaron bien clasificados. Por otro lado, 1 centro siendo de Baja Productividad quedo clasificado como que fuera de Alta Productividad, asimismo otro centro siendo de Baja Productividad quedo clasificado en Mediana Productividad, lo que indica una

incorrecta clasificación para estos. Por último 14 centros de Mediana Productividad fueron clasificados como de Mediana Productividad, lo cual indica una correcta clasificación y 5 centros pertenecientes a Mediana Productividad quedaron clasificados como si fueran de Baja Productividad, lo cual indica que esta incorrectamente clasificado.



14. LECCIONES APRENDIDAS

Se considera relevante el levantamiento de datos

15. PARTICIPACIÓN EN EVENTOS

16. CONCLUSIONES

Sensibilidad de acuerdo al valor obtenido del 0.6830322, esto me indica **¿qué porcentaje de los centros de acuerdo a su nivel de productividad somos capaces de identificar?** Para lo cuál diríamos que el 68,30% de los centros se identifican acertadamente de acuerdo al nivel de productividad.

El valor **F1** se utiliza para combinar las medidas de **precision** y **recall (sensibilidad)** en un sólo valor. Asume que nos importa de igual forma la precisión y la exhaustividad. Esto no tiene que ser así en todos los problemas. Ya que necesitamos predecir si el nivel de productividad del centro se detecto adecuadamente, por lo tanto podemos valorar más la sensibilidad a riesgo de tener una peor precisión. Sin embargo para nuestro caso el F1 es igual 0,6292679 equivalente al 62,93% que no es muy aproximado al 68,30% de sensibilidad del modelo y a la precisión del modelo equivalente a 68,92%. Sin embargo el F1 permite comparar modelos donde si un modelo arroja un F1 alto representa el modelo que mejor puede clasificar las observaciones en clases.

Para nuestro caso un F1 alto indicaría que el modelo sería capaz de clasificar las observaciones de la clase nivel de productividad de manera más precisa.

Exactitud (Accuracy): La exactitud (accuracy) mide el porcentaje de casos que el modelo ha acertado. Es decir, el modelo acierta el 68,92% de las veces. Como se observa, la exactitud es una métrica muy engañosa. De hecho, si tuviésemos un modelo que siempre predijera que el centro siempre va a estar en un nivel de productividad determinado, su accuracy

podría ser del 80% o más y esto  normalmente se observa en problemas con clases desbalanceadas por lo tanto es mucho mejor usar precision, recall y F1. Estas métricas dan una mejor idea de la calidad del modelo. IArtificial.net. (2020).

Especificidad: el modelo tiene una alta capacidad para identificar los centros que no fueron clasificados correctamente según su nivel de productividad entre todos los centros cuyos niveles de productividad fueron clasificados correctamente. Para nuestro caso la especificidad del clasificador es de 88,16%.

De lo anterior tenemos, el modelo es mucho más específico que sensible en la clasificación esto se traduce en que si un falso positivo significa que el modelo tiene una muy alta capacidad para identificar los centros que no fueron clasificados correctamente según su nivel de productividad evitando los falsos positivos.

17. REFERENCIAS

[1] mcnemar.test. McNemar's Chi-Square Test for Count Data.

https://docs.tibco.com/pub/enterprise-runtime-for-R/3.1.1/doc/html/Language_Reference/stats/mcnemar.test.html

[2] Fleiss, J. L. (1981). Statistical Methods for Rates and Proportions, 2nd ed. New York: Wiley.

[3] Zar, J. H. (1984). Biostatistical Analysis, 2nd ed. Englewood Cliffs: Prentice-Hall.

Alan Agresti (1990). Categorical data analysis. New York: Wiley. Pages 350--354.

IArtificial.net. (2020). <https://www.iartificial.net/precision-recall-f1-accuracy-en-clasificacion/>



ANEXOS 3. Apoyo a la formación - Curso de estadística aplicada a inteligencia artificial.

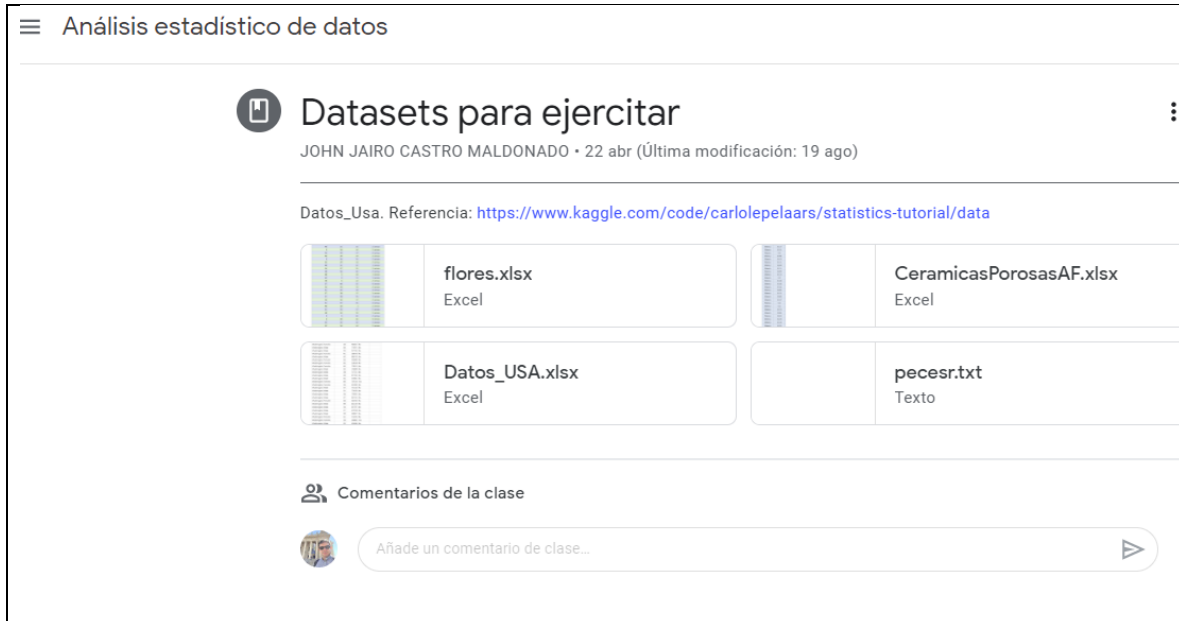


Figura A3.1. Soporte de envío de apoyo a la formación: actividad de estadística básica: Curtosis y coeficiente de asimetría.

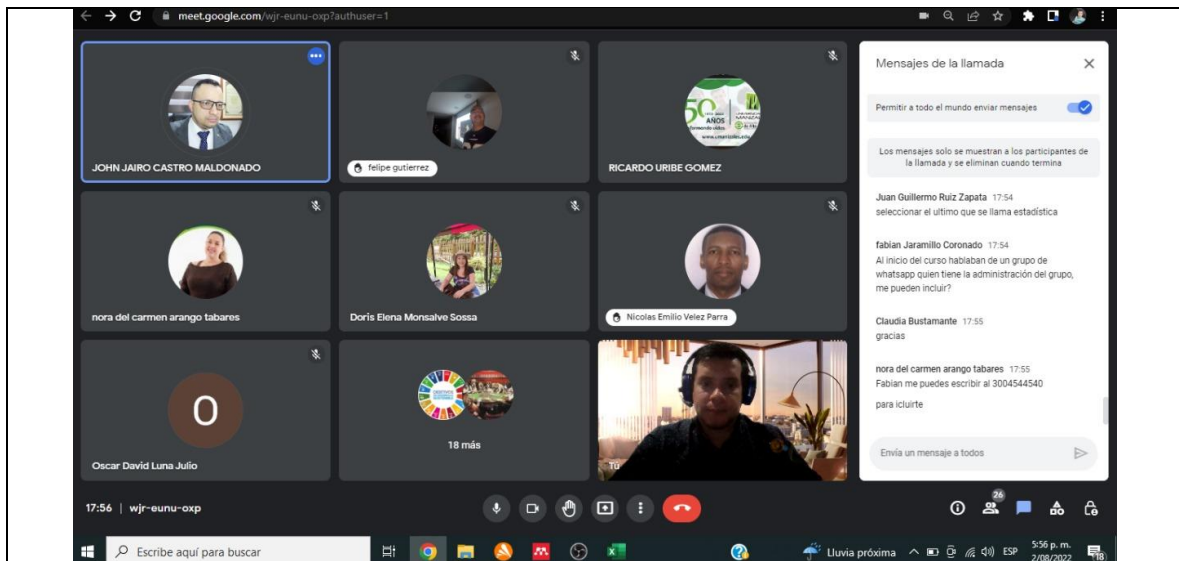


Figura A3.2. Soporte de envío de apoyo a la formación. Realización de clases por medio de la plataforma Meet de Google.



ANEXO 4. Analítica de datos – Encuestas. Soporte de correos, entrega de resultados en código.



Figura A4.1. Recepción de datos para analítica de información.



Figura A4.2. Reunión técnica de avances de desarrollo de código en analítica de información. Desarrollada en 2022-08-08. **Parte 1.**

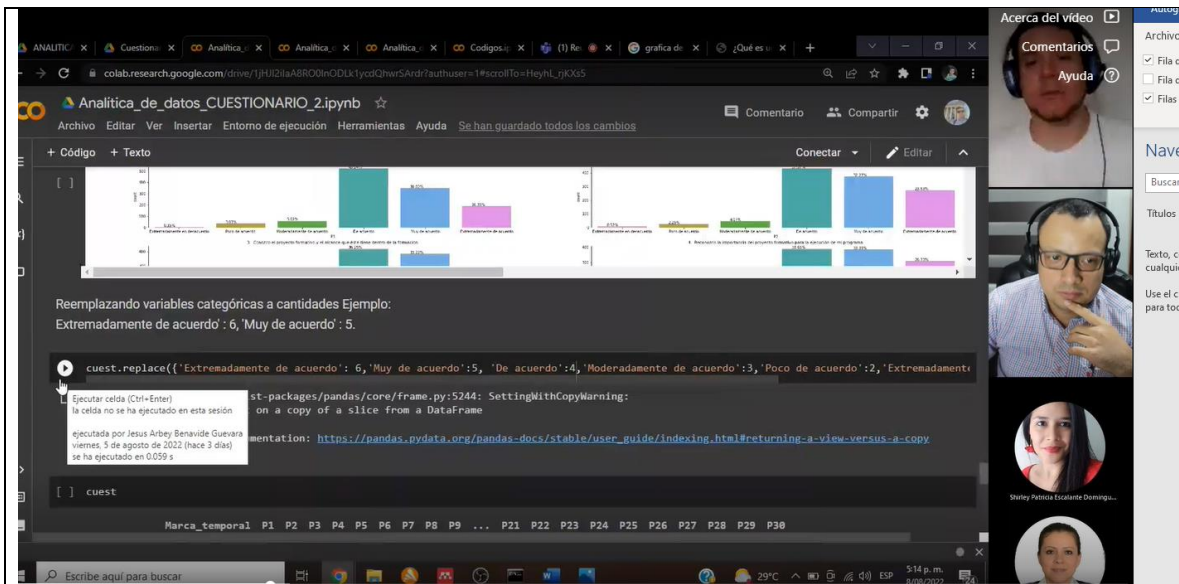


Figura A4.3. Reunión técnica de avances de desarrollo de código en analítica de información. Desarrollada en 2022-08-08. **Parte 2.**



Llamada de Teams

Categorizar Deshacer

CURSO DE ESTADÍSTICA Ahora

Avances Analítica de datos - Encuestas

Pública Clasificada\Todos los empleados (Sin Restricción)

Jesus Arbey Benavides Guevara

Para: Luz Carime Chia Cepeda

CC: John Jairo Castro Maldonado

3 MB 5 MB

4 archivos adjuntos (11 MB) Guardar todo en OneDrive - Servicio Nacional de Aprendizaje Descargar todo

Estimada

Luz Carime Chia,

Instructora - Centro de Comercio y Servicios

Reciba un cordial saludo. Comparto por este medio los avances de la analítica de las encuestas realizadas por su equipo. Sigue en anexo:

1. Link de códigos y datos: <https://bit.ly/3K7IKFS>
2. Pdf de códigos de los resultados avanzados hasta el momento.

Por favor, cualquier inquietud no dude en entrar en contacto y quedo atento a sus indicaciones.

Cordialmente,

Jesus Arbey Benavides Guevara
Investigador SENNOVA - Centro de Servicios y Gestión Empresarial,
jabenavides@sena.edu.co
3105595765
Calle 51 # 59 -70 Torre Norte, Piso 5 -Medellín

Figura A4.4. Soporte de envío de resultados de analítica de datos vía e-mail, para supervisor Ing. John Jairo Castro e instructora Luz Carime Chia Sena- Regional Cauca.



ANEXOS 5. Códigos de apoyo en analítica de datos - Centro de Comercio y Servicios- SENA Cauca



ANEXOS 6. Actas de reuniones.



ACTA No. 145		
NOMBRE DEL COMITÉ O DE LA REUNIÓN: Seguimiento Proyecto SGPS 9748 - Prototipo Innovación		
CIUDAD Y FECHA: Medellín, 02 de agosto de 2022	HORA INICIO: 10:40 a.m.	HORA FIN: 12:04 p.m.
LUGAR Y/O ENLACE: Laboratorio de Inteligencia Artificial Piso 5 https://bit.ly/3oC4ChB	DIRECCIÓN GENERAL / REGIONAL / CENTRO Centro de Servicios y Gestión Empresarial	
AGENDA O PUNTOS PARA DESARROLLAR: <ol style="list-style-type: none">1. Verificación del avance de los productos pendientes2. Sometimiento de artículos		
OBJETIVO(S) DE LA REUNIÓN: Realizar seguimiento a la ejecución de las actividades del Proyecto SGPS 9748 Prototipo Innovación.		
DESARROLLO DE LA REUNIÓN <ol style="list-style-type: none">1. Verificación del avance de los productos pendientes Se inicia la reunión comentando que el desarrollador John Ferla no trabajará esta semana en el prototipo debido a que en el prototipo AMMAS le entregaron el informe de los ajustes a realizar, luego de ser probado por Andrés Álzate, encargado del área ambiental del centro, en este sentido estará trabajando en esos ajustes para ya dedicarse de lleno a culminar con el desarrollo e integraciones pendientes en este prototipo. Por otro lado, la investigadora Paola Buitrago comentó que está trabajando en el informe de los códigos que se han trabajado durante la vigencia para los modelos de analítica del prototipo y cada código lo está documentando. Se hizo énfasis en que la mayor evidencia de la gestión es la entrega de los códigos y demás productos planteados en el proyecto. En cuanto al informe de investigación está en un 50% se entregará un borrador la próxima reunión con los espacios para la información del trabajo realizado por los compañeros John Ferla y Jesús Benavides, se aclaró que el 15 de agosto se debe el informe en su primer borrador.		



Se comentó que al investigador Jesús Arbey Benavides se le encargó la tarea de realizar el análisis de datos de la información de un trabajo de investigación que fue enviado desde la Regional Cauca y es un trabajo al que el investigador le dará prioridad ya que debe realizar una depuración y tratamiento de los datos para facilitar el posterior análisis, al tener el código generado se presentará el avance y se solicitará espacio para presentar el informe de analítica. Luego de generar el código el investigador Jesús Benavides retomará sus aportes al informe de investigación. También la investigadora Paola Buitrago propuso agregar un glosario de términos técnicos para facilitar la comprensión del informe de investigación.

En cuanto al manual de usuario y técnico a las investigadoras Paola Buitrago y Erika Restrepo se les encomendó la tarea de revisarlo y apoyar con los ajustes de redacción.

2. Sometimiento de artículos

Se realizó el sometimiento del artículo de la Revista Retos, se está a la espera de la respuesta.

La investigadora Erika Restrepo estuvo realizando el ajuste del formato del artículo que se envió al V Workshop on Modeling and Simulation for Science and Engineering (V WMSSE), en el que realizó ajustes en las especificaciones de forma, algunos detalles de redacción y en las referencias según el estilo solicitado por el evento.

También se envió el sometimiento de la propuesta del trabajo realizado en la vigencia anterior, solicitando ser tenidos en cuenta para la publicación en capítulo de libro en Expotecnológica 2022 del Pascual Bravo, donde se tendrá ponencia y posible publicación.

Se sugiere que los investigadores del prototipo Erika Restrepo, Paola Buitrago y Jesús Benavides, realicen sesiones para establecer una forma de realizar la transferencia y pulir el discurso, además Erika contará con el apoyo de Jesús Benavides en la parte técnica, así continúe como instructor se le dará en tiempo para apoyar en esta tarea de transferencia del prototipo.



Además, se socializó que en Redcolsi Regional se obtuvo una calificación de 91 puntos lo que da paso a participar en el encuentro Nacional, se gestionará un aprendiz en contrato de aprendizaje que pueda representar el prototipo en el evento.

Aunque no alcanzó el tiempo de los investigadores para explorar más técnicas de clusterización, para la predicción también hizo falta tiempo para poder realizar comparaciones entre métodos y realizar otro artículo. Esto es insumo para la convocatoria 2023 de SENNOVA. Sin embargo, se dejan muy buenos insumos y productos para continuar con el prototipo, se considera que hay tiempo suficiente para terminar la documentación del prototipo.

Para finalizar se mencionó que el equipo de trabajo alcanzó una muy buena sinergia y cohesión para el trabajo esto se ve reflejado en el cumplimiento de las actividades planteadas en los tiempos establecidos, se esperaba contar con presupuesto para continuar con las actividades, aunque no fue posible el resultado obtenido se considera satisfactorio.

Al enviar el certificado al líder del grupo, no fue aprobado inicialmente por quien hace las veces de líder del grupo; la investigadora Paola Buitrago realizará las gestiones que sean necesarias para que la participación en el evento sea aceptada y agregada a los productos del grupo de investigación.

CONCLUSIONES

Se está a la espera de que el desarrollador John Ferla continúe avanzando con la integración del aplicativo de recolección de información y los modelos de analítica de datos.

Se enviaron tres sometimientos a diferentes eventos, donde se espera tener una respuesta positiva frente a la publicación de los trabajos que se propusieron.

De acuerdo con La Ley 1581 de 2012, Protección de Datos Personales se debe garantizar la seguridad y protección de los datos personales que se encuentran almacenados en este documento. El Servicio Nacional de Aprendizaje SENA solicita la siguiente clasificación de la información:

La información de este documento se debe clasificar como:

PÚBLICA ☐ **PRIVADO** ☐ **SEMIPRIVADO** ☐ **SENSIBLE** ☐



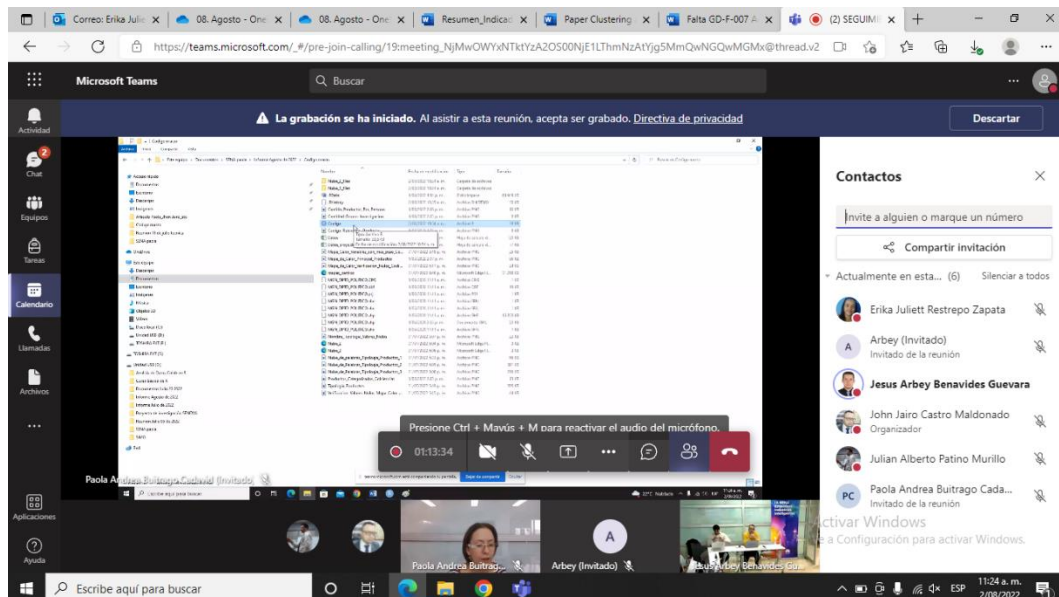
Nota: Antes de contestar esta informacional por favor remitirse al instructivo

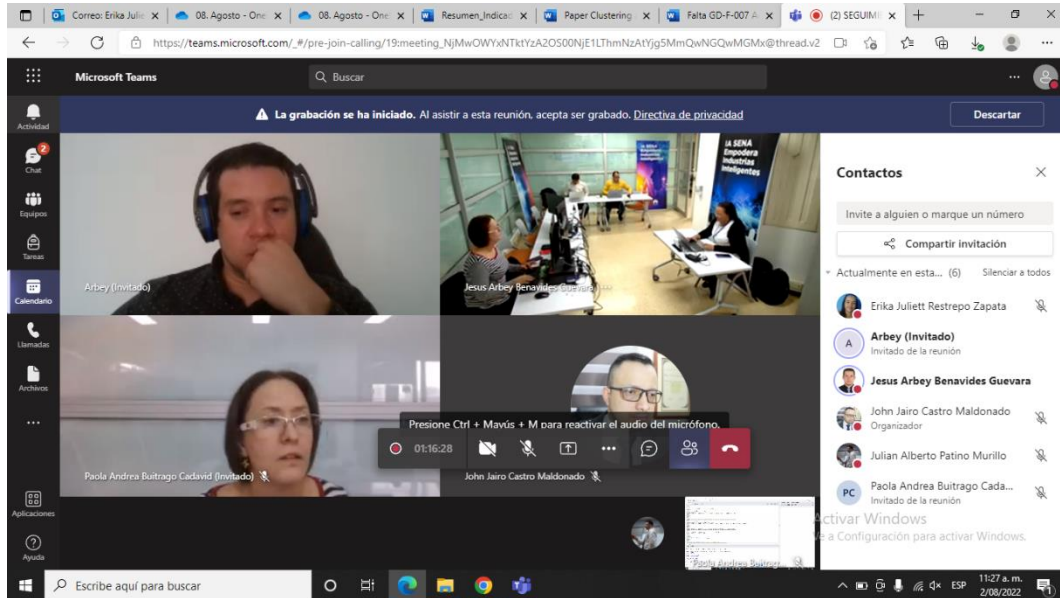
COMPROMISOS

ACTIVIDAD	RESPONSABLE	FECHA
Continuar apoyando el desarrollo y la elaboración de la documentación del prototipo.	Jesús Arbey Benavides Paola Andrea Buitrago	09 de agosto de 2022
Continuar integrando el modelo de analítica en el aplicativo.	John Ferla	09 de agosto de 2022
Continuar con las actividades de divulgación del prototipo.	Erika Restrepo	09 de agosto de 2022
Verificar con el líder del grupo de investigación el certificado de participación en el V Workshop on Modeling and Simulation for Science and Engineering (V WMSSE)	Paola Buitrago	09 de agosto de 2022

ASISTENTES: (Incorporar registro de asistencia)

Se anexa evidencia fotográfica de la reunión:





Nota: Se conectó de manera virtual Jesús Arbey Benavides

Enlaces de grabación:

[SEGUIMIENTO DE ACTIVIDADES DE PROYECTO DE INDICADORES DE CTEI CON ANALÍTICA DE DATOS-20220802_105047-Grabación de la reunión.mp4](#)



ANEXO 7. Participación en reuniones con equipo para el desarrollo del prototipo de software.

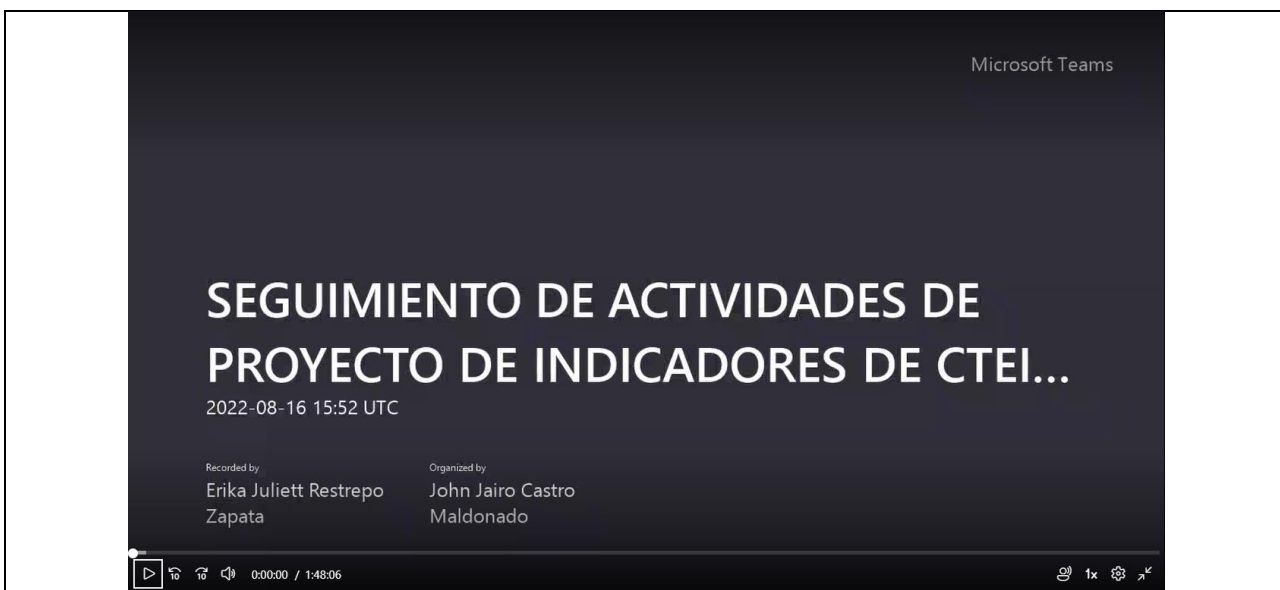


Figura A7.1. Reunión de seguimiento 2022-08-16

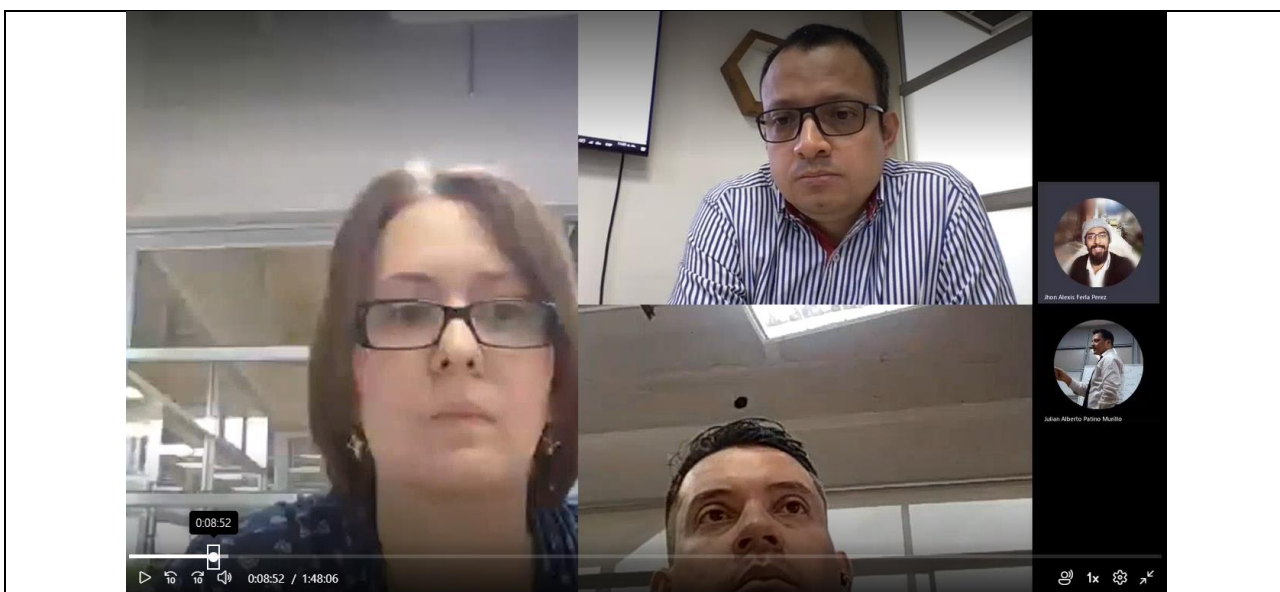


Figura A7.2. Reunión de seguimiento 2022-08-16. Reporte final y entrega de proyecto investigadora Paola Buitrago.

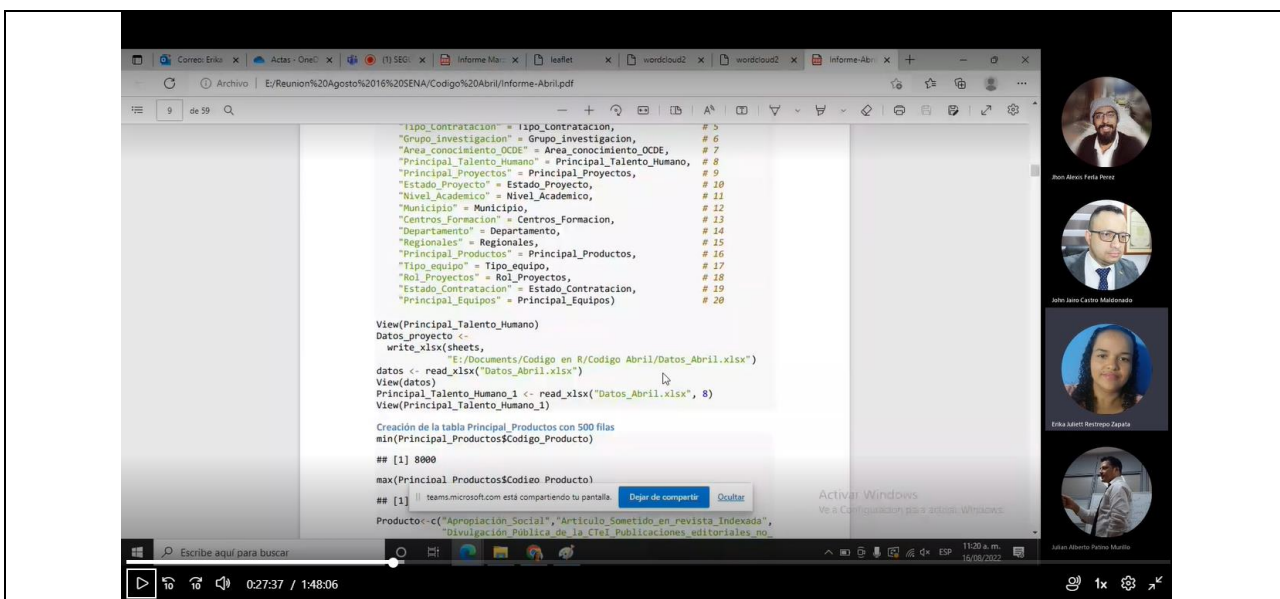


Figura A7.3. Reunión de seguimiento 2022-08-16. Reporte final y entrega de proyecto investigadora Paola Buitrago. Reporte de códigos en R



Figura A7.4. Reunión de seguimiento 2022-08-26. Modelo de predictor en plataforma API.

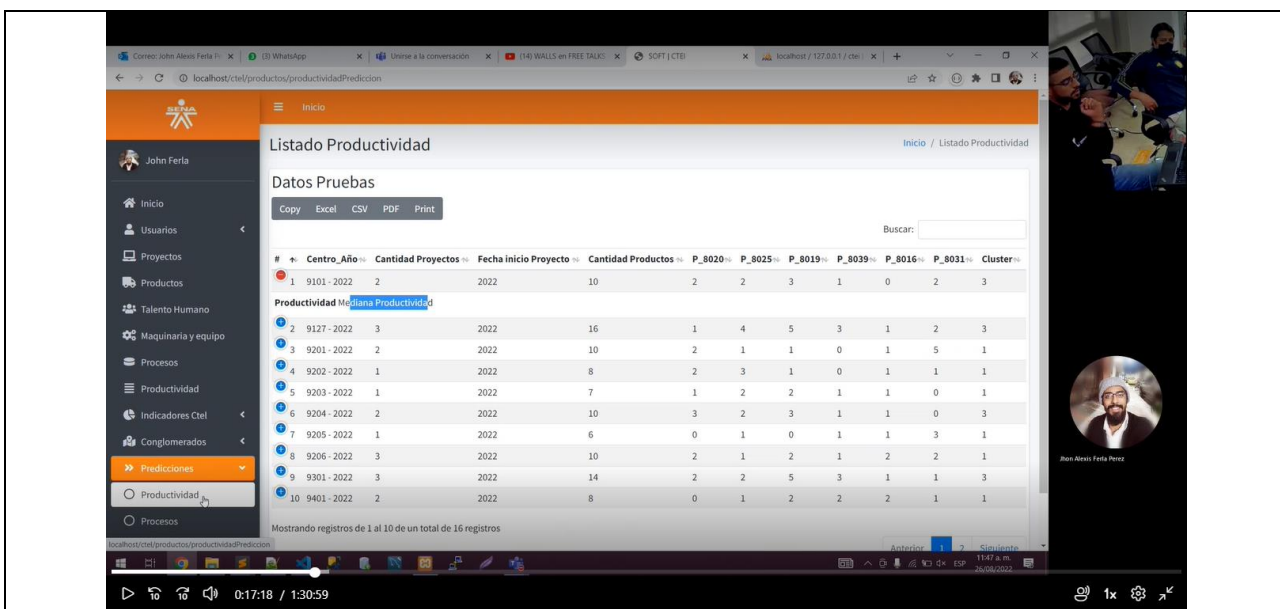


Figura A7.5. Reunión de seguimiento 2022-08-26. Modelo de predictor en plataforma API. Tabla de datos data sintética y plataforma API.

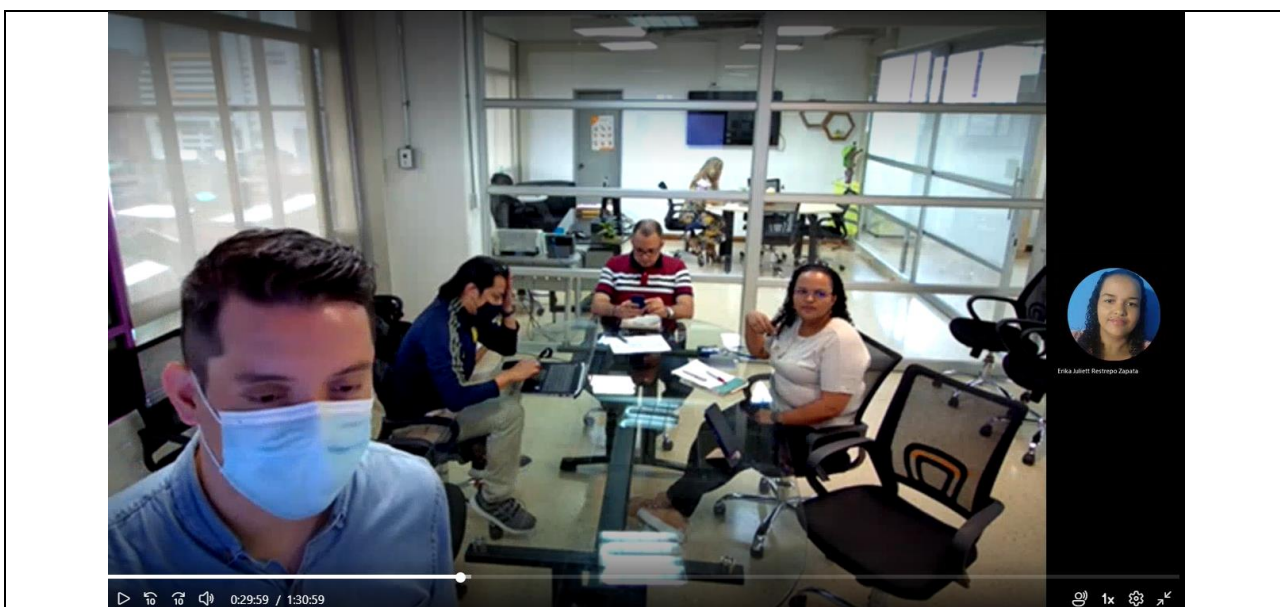


Figura A7.6. Reunión de seguimiento 2022-08-26. Reporte de agosto.



ANEXO 8. Correos asociados a Proceso editorial artículos [V WMSSE-JPCS].



Figura A8.1. Correos de acompañamiento asociados a Proceso editorial artículos [V WMSSE-JPCS].

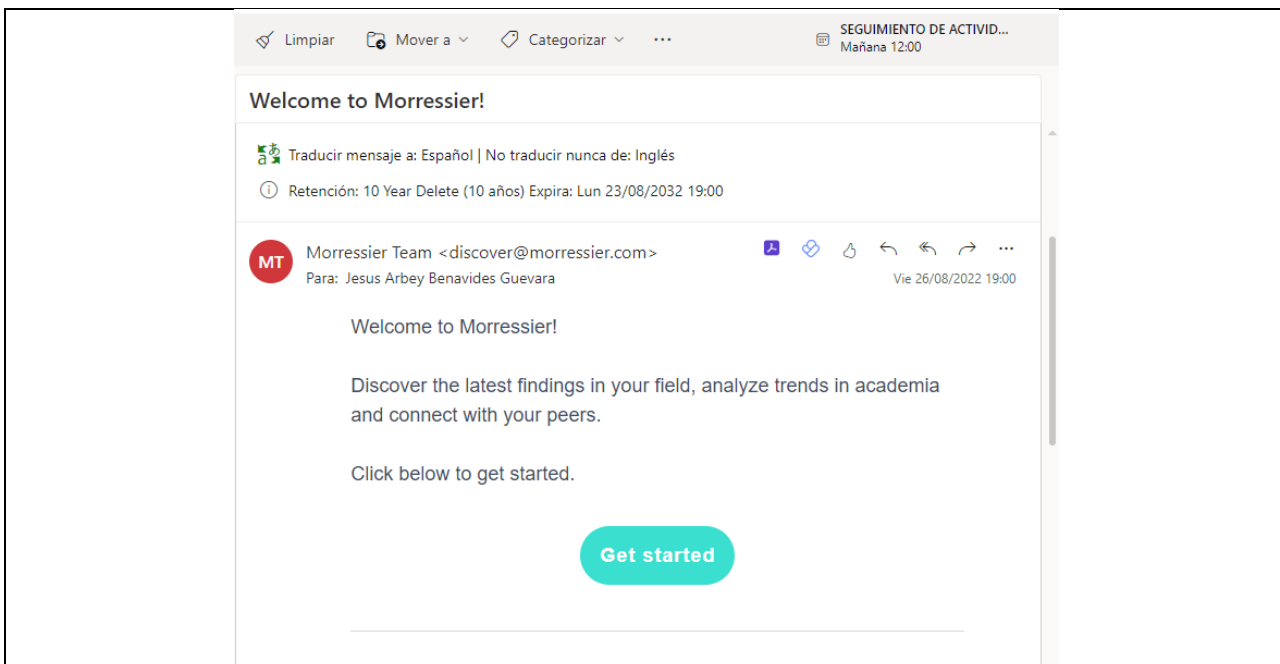


Figura A8.2. Correos de acompañamiento asociados a Proceso editorial artículos [V WMSSE-JPCS]. Vinculo de autoría en plataforma Morressier.

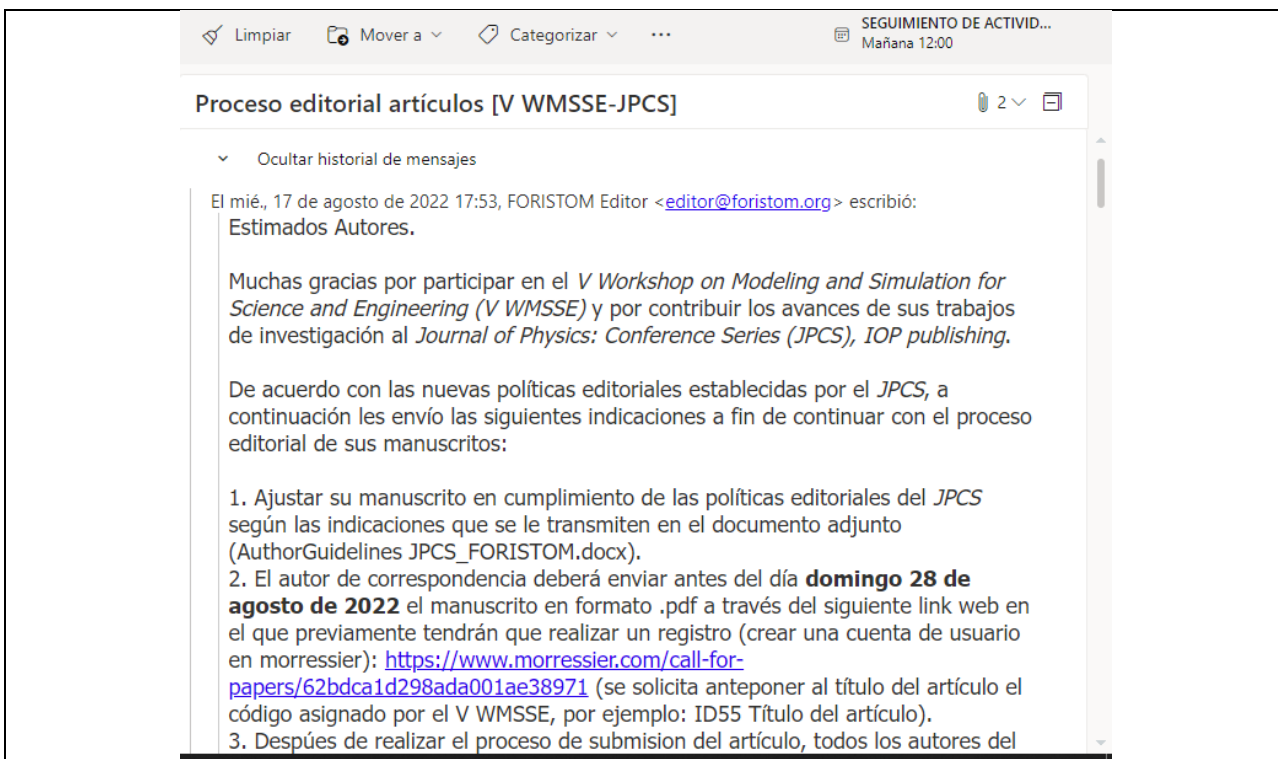


Figura A8.3. Correos de acompañamiento asociados a Proceso editorial artículos [V WMSSE-JPCS]. Recomendaciones de editorial.



Figura A8.4. Correos de acompañamiento asociados a Proceso editorial artículos [V WMSSE-JPCS]. Envío de respuesta para adecuación investigador Paola Buitrago.

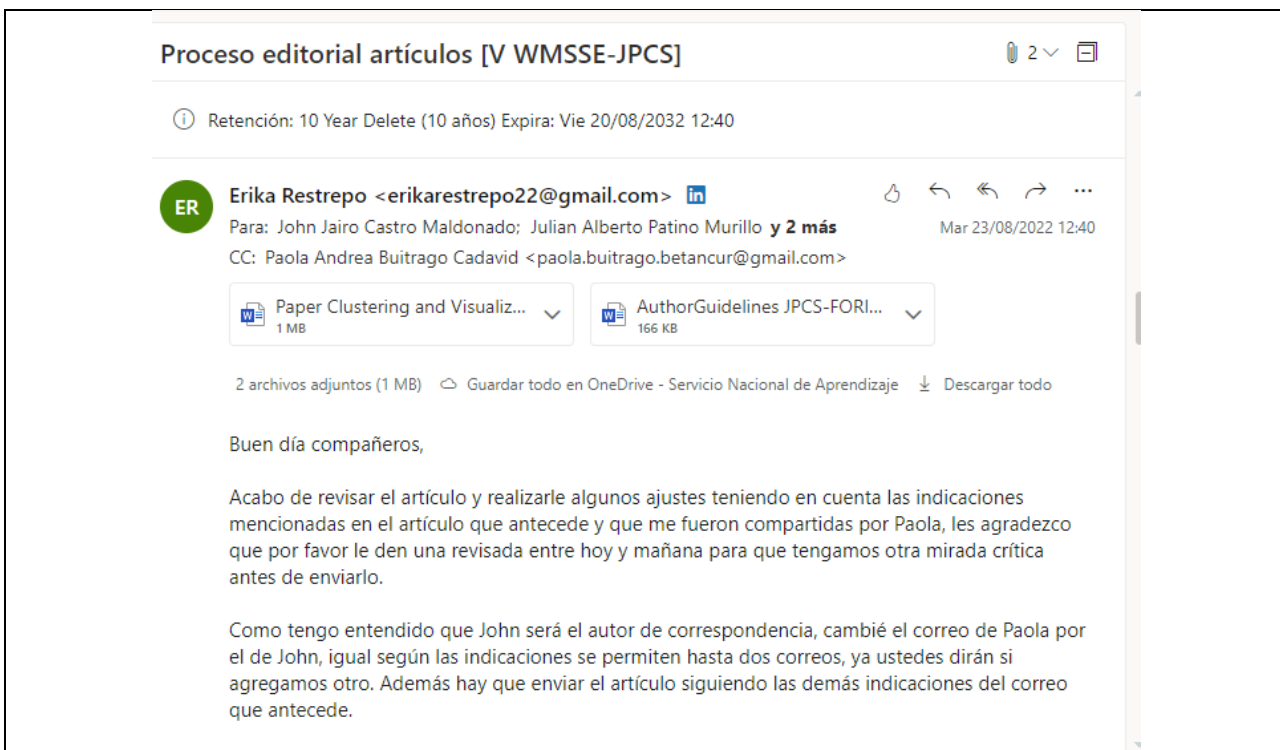


Figura A8.5. Correos de acompañamiento asociados a Proceso editorial artículos [V WMSSE-JPCS]. Envío de respuesta para adecuación investigador Paola Buitrago.

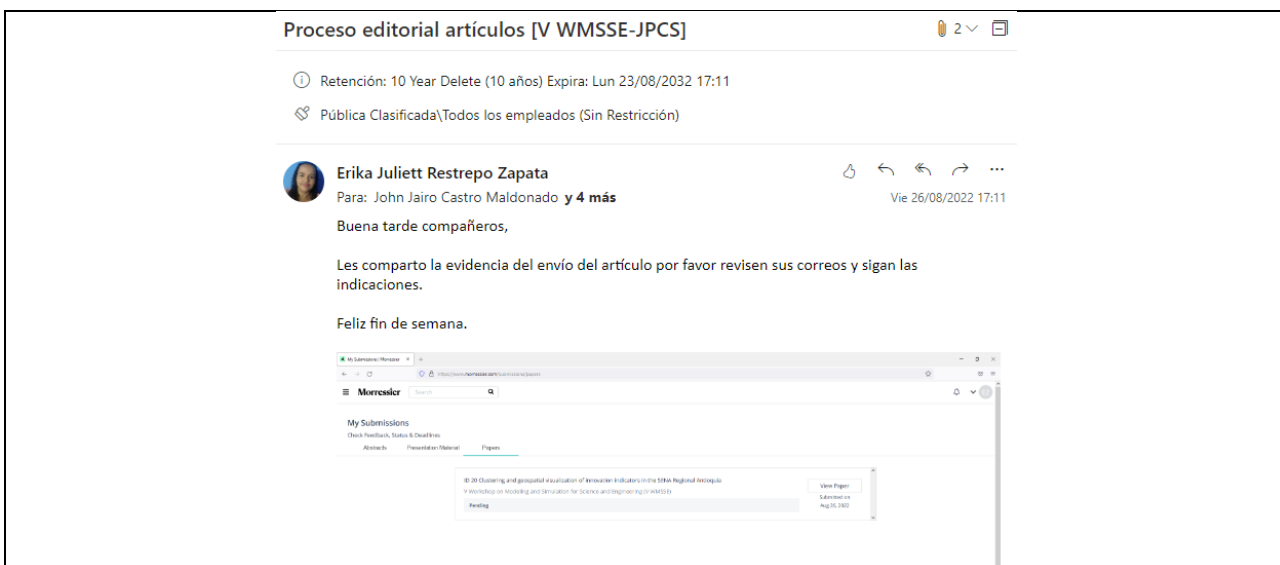


Figura A8.6. Correos de acompañamiento asociados a Proceso editorial artículos [V WMSSE-JPCS]. Evidencia de sometimiento de artículo Erika Restrepo.



ANEXO 9. Reporte de códigos para selección de modelos de analítica de datos.

```

import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)

import folium # plotting library
from folium import plugins
from folium.plugins import MarkerCluster

from sklearn.cluster import KMeans
from sklearn.decomposition import PCA
import sklearn.metrics as sm
from sklearn.decomposition import PCA
from sklearn.cluster import AgglomerativeClustering
from sklearn.cluster import DBSCAN
from io import IncrementalNewlineDecoder
import scipy.cluster.hierarchy as sch

#Cluster
from yellowbrick.cluster import SilhouetteVisualizer
from yellowbrick.cluster import KElbowVisualizer
from sklearn.datasets import make_blobs

#Graficas
import matplotlib.pyplot as plt
from matplotlib import pyplot as plot
import matplotlib.cm as cm
import matplotlib.colors as colors
from plotly import graph_objects as go
from plotly.subplots import make_subplots
import plotly.express as px
import seaborn as sns
import branca
from sklearn.metrics import plot_confusion_matrix

from google.colab import drive
drive.mount('/content/drive')

```

Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.moun



```

from google.colab import files
path = "/content/drive/MyDrive/Proyecto SGPS 9748/Datos.xlsx"
path2 = '/content/drive/MyDrive/Proyecto SGPS 9748/PCPA.xlsx'

#Dataframes
principal_proyectos= pd.read_excel(path,sheet_name= 'Principal_Proyectos')
principal_talento_humano = pd.read_excel(path,sheet_name= 'Principal_Talento_Humano')
principal_productos = pd.read_excel(path,sheet_name= 'Principal_Productos')

```

```
df_product_orig = pd.read_excel(path2)
df_centros = pd.read_excel(path, sheet_name= 'Centros_Formacion')
```

```
df_product_orig
```

	Centro_Año	P_8020	P_8025	P_8019	P_8039	P_8016	P_8031
0	9101-2010	1.0	3.0	1.0	6.0	0.0	3.0
1	9127-2010	1.0	0.0	0.0	2.0	3.0	2.0
2	9201-2010	0.0	0.0	0.0	0.0	0.0	2.0
3	9202-2010	0.0	1.0	0.0	2.0	0.0	1.0
4	9203-2010	6.0	2.0	2.0	4.0	2.0	3.0
...	...	...	...	...	...	...	...
203	9501-2022	3.0	6.0	6.0	1.0	1.0	1.0
204	9502-2022	2.0	3.0	2.0	1.0	1.0	1.0
205	9503-2022	2.0	0.0	2.0	1.0	2.0	1.0
206	9504-2022	4.0	2.0	0.0	1.0	1.0	1.0
207	9549-2022	1.0	0.0	3.0	3.0	2.0	4.0

208 rows × 7 columns

```
df_product_orig.columns
```

```
Index(['Centro_Año', 'P_8020', 'P_8025', 'P_8019', 'P_8039', 'P_8016',
      'P_8031'],
      dtype='object')
```

```
df_product_orig.columns = ['Centro_Año', 'Apropiación_Social', 'Articulo_Sometido_en_revista_
      'Divulgación_Pública_de_la_CTeI_Publicaciones_editoriales_no_especializadas_p
      'Informe_de_investigación', 'Libros_resultados_de_investigación', 'Productos_
df= df_product_orig
df
```

	Centro_Año	Apropiación_Social	Articulo_Sometido_en_revista_Indexada	Divulgación
0	9101-2010	1.0		3.0
1	9127-2010	1.0		0.0
2	9201-2010	0.0		0.0
3	9202-2010	0.0		1.0
4	9203-2010	6.0		2.0
...	...	...		...
203	9501-2022	3.0		6.0
204	9502-2022	2.0		3.0
205	9503-2022	2.0		0.0

df.columns

```
Index(['Centro_Año', 'Apropiación_Social',
      'Articulo_Sometido_en_revista_Indexada',
      'Divulgación_Pública_de_la_CTeI_Publicaciones_editoriales_no_especializadas_producciones',
      'Informe_de_investigación', 'Libros_resultados_de_investigación',
      'Productos_tecnológicos_certificados_o_validados'],
      dtype='object')
```

df.describe()

	Apropiación_Social	Articulo_Sometido_en_revista_Indexada	Divulgación_Pública_d
count	208.000000	208.000000	
mean	1.889423	1.793269	
std	1.591731	1.410983	
min	0.000000	0.000000	
25%	1.000000	1.000000	
50%	2.000000	2.000000	
75%	3.000000	3.000000	
max	8.000000	7.000000	

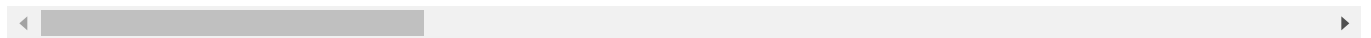
```

from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
df[['Apropiación_Social',
    'Articulo_Sometido_en_revista_Indexada',
    'Divulgación_Pública_de_la_CTeI_Publicaciones_editoriales_no_especializadas_produccion
    'Informe_de_investigación', 'Libros_resultados_de_investigación',
    'Productos_tecnológicos_certificados_o_validados']] = scaler.fit_transform(df[['Apropi
    'Articulo_Sometido_en_revista_Indexada',
    'Divulgación_Pública_de_la_CTeI_Publicaciones_editoriales_no_especializadas_produccion
    'Informe_de_investigación', 'Libros_resultados_de_investigación',
    'Productos_tecnológicos_certificados_o_validados']])
df

```

	Centro_Año	Apropiación_Social	Articulo_Sometido_en_revista_Indexada	Divulgación
0	9101-2010	-0.560125		0.857305
1	9127-2010	-0.560125		-1.274003
2	9201-2010	-1.189888		-1.274003
3	9202-2010	-1.189888		-0.563567
4	9203-2010	2.588687		0.146869
...	...	...		...
203	9501-2022	0.699400		2.988612
204	9502-2022	0.069637		0.857305
205	9503-2022	0.069637		-1.274003
206	9504-2022	1.329162		0.146869
207	9549-2022	-0.560125		-1.274003

208 rows × 7 columns



```

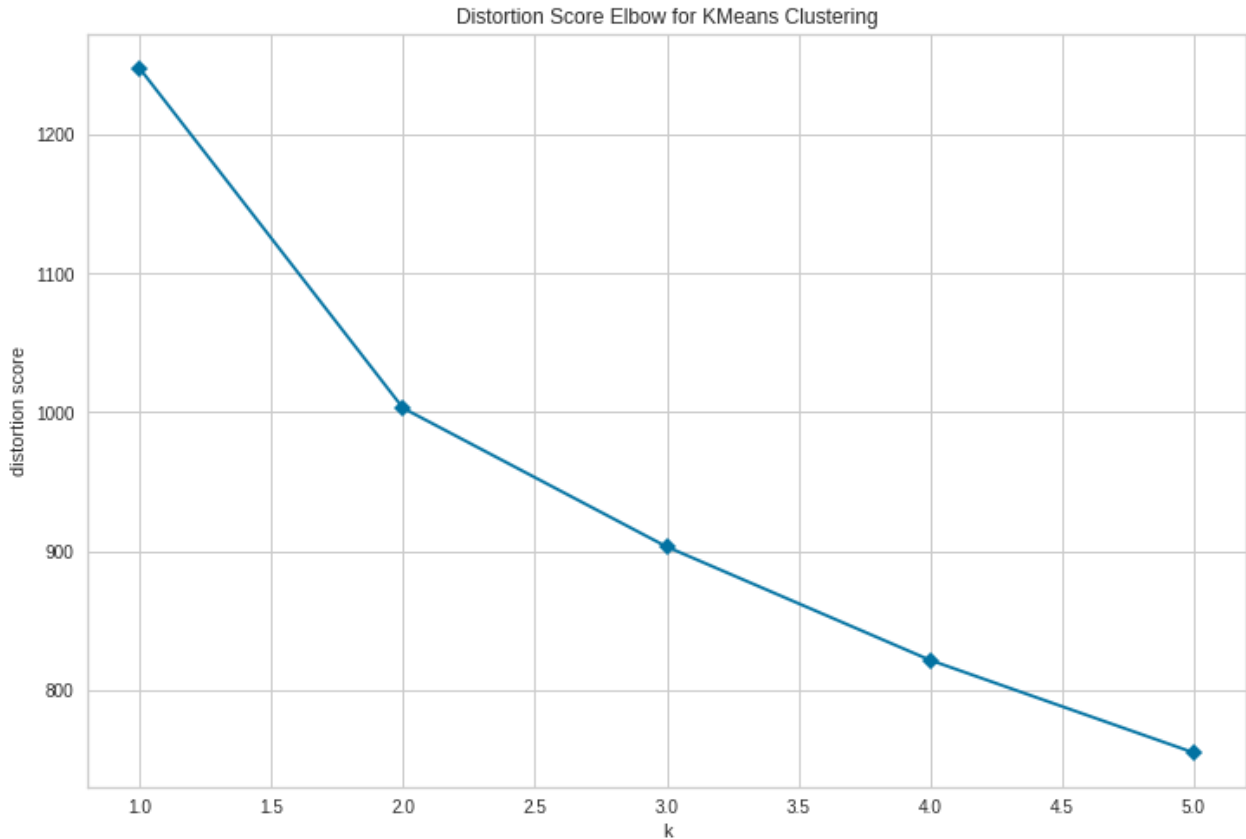
#Búsqueda de la cantidad óptima de clusters
##Calculando qué tan similares son los individuos dentro de los Custers
y = df[['Centro_Año']]
x = df.drop(['Centro_Año'], axis= 1)
wcss = []
for i in range (2,6):
    kmeans = KMeans(n_clusters=i, max_iter=300, random_state=0)
    kmeans.fit(x) #Aplico K-means a la base de datos
    wcss.append(kmeans.inertia_)

```

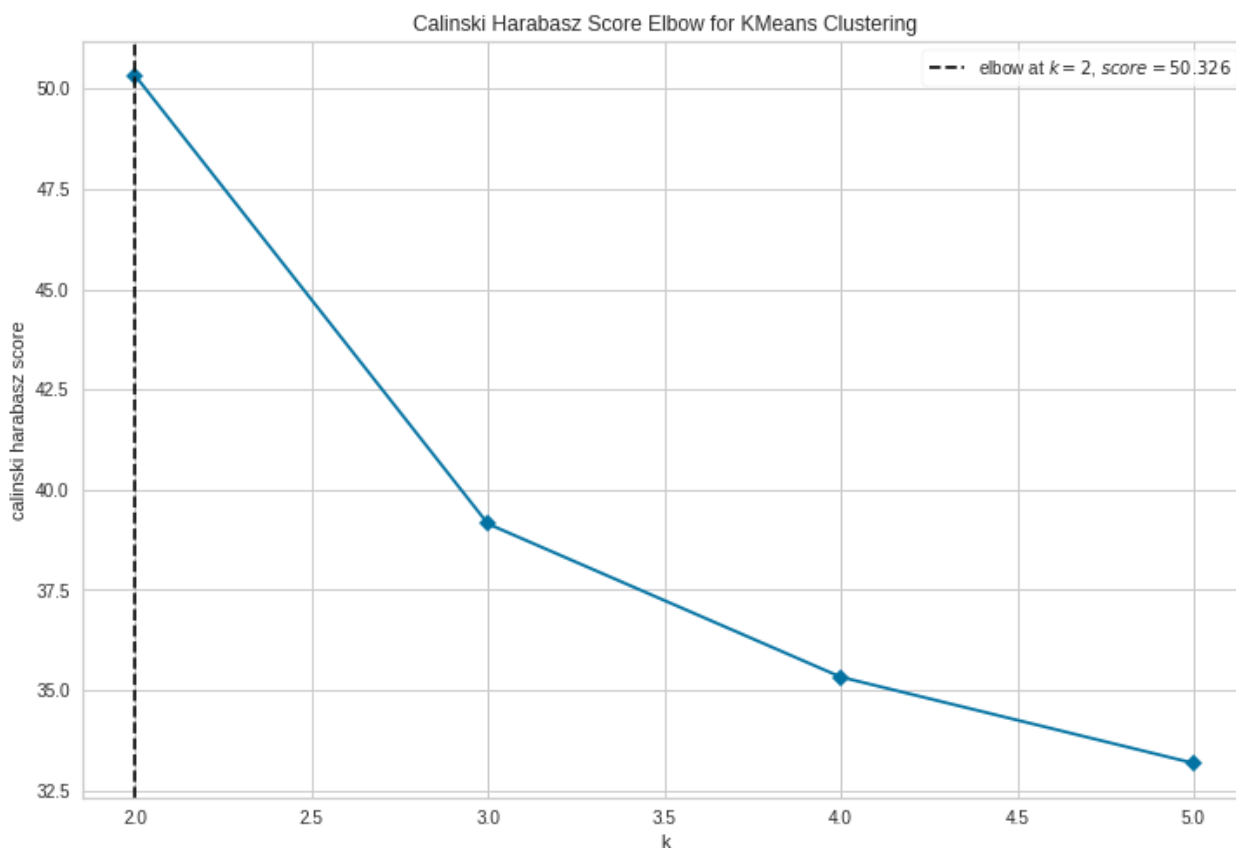
distortion: suma media de las distancias al cuadrado de los centroides (Error medio cuadrático)

```
#km_transformed = mms.transform(km)
plt.figure(figsize=(12,8))
model = KMeans(random_state=0)
visualizer = KElbowVisualizer(model, k=(1,6), timings=False)
visualizer.fit(x)
visualizer.show();
```

```
/usr/local/lib/python3.7/dist-packages/yellowbrick/utils/kneed.py:155: Yellowbrick
warnings.warn(warning_message, YellowbrickWarning)
/usr/local/lib/python3.7/dist-packages/yellowbrick/cluster/elbow.py:361: Yellowbri
warnings.warn(warning_message, YellowbrickWarning)
```



```
plt.figure(figsize=(12,8))
visualizer = KElbowVisualizer( model, k=(2,6), metric='calinski_harabasz', timings=False)
visualizer.fit(x)          # Fit the data to the visualizer
visualizer.show()          # Finalize and render the figure
```



```
plt.figure(figsize=(12,8))
visualizer = KElbowVisualizer( model, k=(2,6), metric='silhouette', timings=False)
visualizer.fit(x)          # Fit the data to the visualizer
visualizer.show()          # Finalize and render the figure
```



```
#Aplicando el método de k-means a la base de datos
clustering=KMeans(n_clusters=3, max_iter=300,random_state=0) #Crea el modelo
clustering.fit(x) #Aplica el modelo a la base de datos
clustering.cluster_centers_
```

```
array([[ 0.59187933,  0.28549053,  0.7278235 ,  1.07189163,  0.96450039,
         1.00649341],
       [ 0.1193553 ,  0.78252193,  0.2124389 , -0.22308574, -0.20644944,
        -0.43755308],
       [-0.36635226, -0.7821624 , -0.50534198, -0.29662682, -0.26213581,
        -0.0880461 ]])
```

```
#Agregando la clasificación en el archivo original
df["KMeans_Clusters"]=clustering.labels_
```

► Validación de clusters

```
#Calcular agrupación de índices de validación externa con KMeans
labels_true = df["Centro_Año"]
rotulos_km = clustering.labels_
h = sm.homogeneity_score(labels_true, rotulos_km)
print("Homogeneidad: %0.3f" % h)
c = sm.completeness_score(labels_true, rotulos_km)
print("Compleitud: %0.3f" % c)
v = sm.v_measure_score(labels_true, rotulos_km)
print("Medida V: %0.3f" % v )
ari = sm.adjusted_rand_score(labels_true, rotulos_km)
print("Índice Rand corregido: %0.3f" % ari)
ami = sm.adjusted_mutual_info_score(labels_true, rotulos_km, average_method='arithmetic')
print("Adjusted Mutual Information: %0.3f" % ami )
fm = sm.fowlkes_mallows_score(labels_true, rotulos_km)
print("Índice Fowlkes-Mallows: %0.3f" % fm )
```

```
Homogeneidad: 0.197
Compleitud: 1.000
Medida V: 0.329
Índice Rand corregido: 0.000
Adjusted Mutual Information: 0.000
Índice Fowlkes-Mallows: 0.000
```

```
from sklearn.metrics import silhouette_score
#silhouette_avg = silhouette_score(x, clustering.labels_)
s = sm.silhouette_score(x, clustering.labels_, metric='euclidean')
```



```
print("Coeficiente de Silhueta: %0.3f" % s)
ch = sm.calinski_harabasz_score(x, clustering.labels_)
print("Calinski_harabasz: %0.3f" % ch)
dbs = sm.davies_bouldin_score(x, clustering.labels_)
print("Davies Bouldin: %0.3f" % dbs)
```

```
Coeficiente de Silhueta: 0.148
Calinski_harabasz: 39.148
Davies Bouldin: 1.945
```

▼ DBSCAN CLUSTERING

```
x.head()
```

	Apropiación_Social	Articulo_Sometido_en_revista_Indexada	Divulgación_Pública_de_la
0	-0.560125		0.857305
1	-0.560125		-1.274003
2	-1.189888		-1.274003
3	-1.189888		-0.563567
4	2.588687		0.146869

```
db = DBSCAN(eps=0.3, min_samples=7, metric='euclidean').fit(x)
db.labels_
```

```
array([-1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1,
       -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1,
       -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1,
       -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1,
       -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1,
       -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1,
       -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1,
       -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1,
       -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1,
       -1, -1, -1, -1])
```

```
#Contar número de clusters
#labels = db.labels_
```

```
#n_clusters_ = len(set(labels)) - (1 if -1 in labels else 0)
#n_noise_ = list(labels).count(-1)
#print('Número estimado de clusters: %d' % n_clusters_)
#print('Número estimado de datos ruidosos: %d' % n_noise_)
#print("Adjusted Rand Index: %0.3f"
#      % metrics.adjusted_rand_score(labels_true, labels))
#print("Adjusted Mutual Information: %0.3f"
#      % metrics.adjusted_mutual_info_score(labels_true, labels))
#print("Silhouette Coefficient: %0.3f"
#      % metrics.silhouette_score(X, labels))
```

▼ PCA

```
#Visualización de los Clusters que se crearon
#Aplicaremos Análisis de componentes principales (PCA) para darnos una idea de cómo se crearon
pca=PCA(n_components=2)
pca_productividad=pca.fit_transform(x)
pca_productividad_2=pd.DataFrame(data=pca_productividad,columns=["Componente_1", "Componente_2"])
pca_productividad_3=pd.concat([pca_productividad_2,df[["KMeans_Clusters"]]],axis=1)
pca_productividad_3=pd.concat([pca_productividad_3,df[["Centro_Año"]]],axis=1)
pca_productividad_3
```

	Componente_1	Componente_2	KMeans_Clusters	Centro_Año
0	0.816244	-0.662001	0	9101-2010
1	-0.847087	-1.759406	2	9127-2010
2	-2.365945	-0.431087	2	9201-2010
3	-1.898267	-0.256295	2	9202-2010
4	2.185298	-0.601068	0	9203-2010
...	...	...	...	...
203	2.164333	3.604048	1	9501-2022
204	-0.179709	1.214482	1	9502-2022
205	-0.765314	-0.367135	2	9503-2022
206	-0.552922	0.376368	1	9504-2022
207	0.674241	-1.491958	0	9549-2022

208 rows × 4 columns

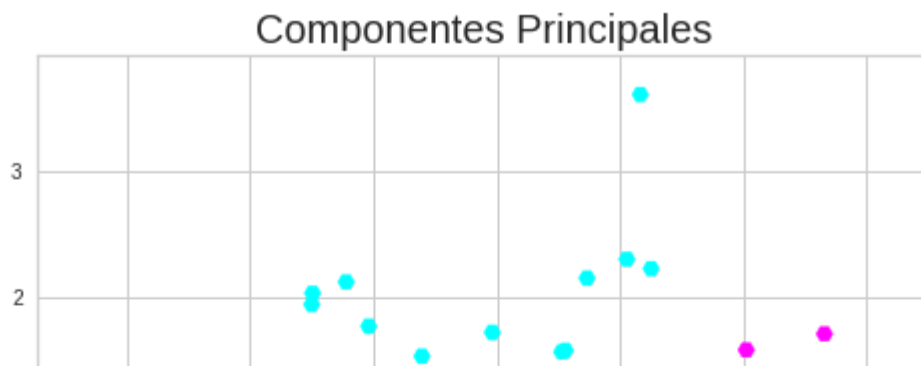
```
pca_productividad_3 = pca_productividad_3.dropna()
pca_productividad_3['KMeans_Clusters'] = pca_productividad_3['KMeans_Clusters'].astype(int)
pca_productividad_3
```

	Componente_1	Componente_2	KMeans_Clusters	Centro_Año
0	0.816244	-0.662001	0	9101-2010
1	-0.847087	-1.759406	2	9127-2010
2	-2.365945	-0.431087	2	9201-2010
3	-1.898267	-0.256295	2	9202-2010
4	2.185298	-0.601068	0	9203-2010
...	...	...	...	...
203	2.164333	3.604048	1	9501-2022
204	-0.179709	1.214482	1	9502-2022
205	-0.765314	-0.367135	2	9503-2022
206	-0.552922	0.376368	1	9504-2022
207	0.674241	-1.491958	0	9549-2022

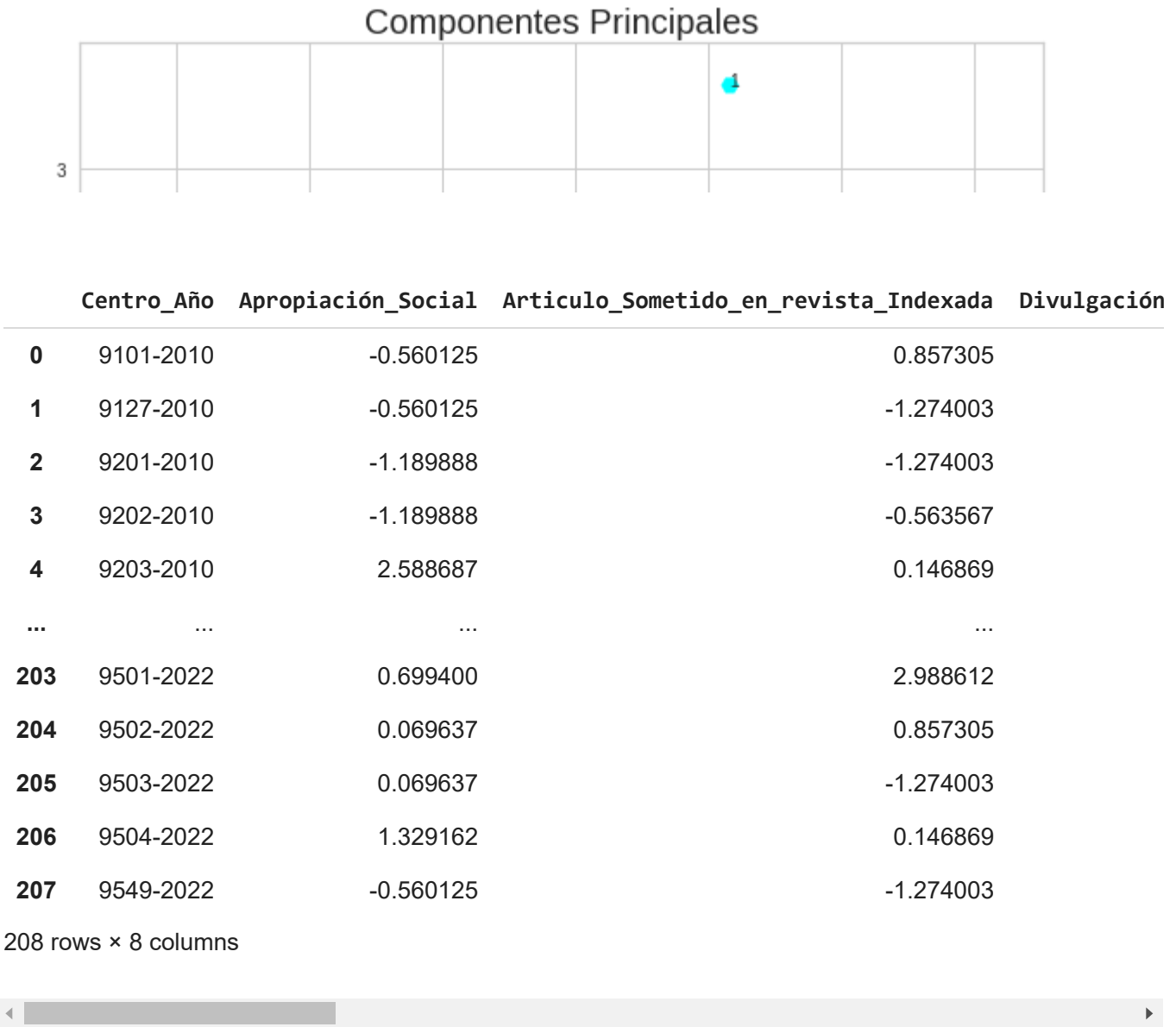
208 rows × 4 columns

```
fig=plt.figure(figsize=(8,8))
ax=fig.add_subplot(1,1,1)
ax.set_xlabel("Componente_1",fontsize=15)
ax.set_ylabel("Componente_2", fontsize=15)
ax.set_title("Componentes Principales", fontsize=20)
color_theme=np.array(["MAGENTA","CYAN",'GREENYELLOW','mediumpurple'])
ax.scatter(x=pca_productividad_3['Componente_1'], y=pca_productividad_3['Componente_2'],c=col
#for i, label in enumerate(annotations):
#    plt.annotate(label, (pca_productividad_3['KMeans_Clusters'][i], pca_productividad_3['KMe
#plt.show()
```

<matplotlib.collections.PathCollection at 0x7f90d421b6d0>



```
#revisar tamaño de etiqueta de punto
fig=plt.figure(figsize=(10,10))
ax=fig.add_subplot(1,1,1)
annotations = pca_productividad_3['KMeans_Clusters']
ax.set_xlabel("Componente_1",fontsize=15)
ax.set_ylabel("Componente_2", fontsize=15)
ax.set_title("Componentes Principales", fontsize=20)
color_theme=np.array(["MAGENTA","CYAN",'GREENYELLOW','mediumpurple'])
ax.scatter(x=pca_productividad_3['Componente_1'], y=pca_productividad_3['Componente_2'],c=col
for i, label in enumerate(annotations):
    plt.annotate(label, (pca_productividad_3['Componente_1'][i], pca_productividad_3['Compone
plt.show()
```



```
pip install mlxtend
```

Looking in indexes: <https://pypi.org/simple>, <https://us-python.pkg.dev/colab-wheels/public/python-wheels/>

Requirement already satisfied: mlxtend in /usr/local/lib/python3.7/dist-packages (0.20.0)

Requirement already satisfied: pandas>=0.24.2 in /usr/local/lib/python3.7/dist-packages (1.1.5)

Requirement already satisfied: numpy>=1.16.2 in /usr/local/lib/python3.7/dist-packages (1.21.0)

Requirement already satisfied: setuptools in /usr/local/lib/python3.7/dist-packages (57.0.0)

Requirement already satisfied: scikit-learn>=1.0.2 in /usr/local/lib/python3.7/dist-packages (0.24.2)

Requirement already satisfied: matplotlib>=3.0.0 in /usr/local/lib/python3.7/dist-packages (3.5.3)

Requirement already satisfied: joblib>=0.13.2 in /usr/local/lib/python3.7/dist-packages (1.1.0)

Requirement already satisfied: scipy>=1.2.1 in /usr/local/lib/python3.7/dist-packages (1.7.3)

Requirement already satisfied: pyparsing!=2.0.4,!=2.1.2,!=2.1.6,>=2.0.1 in /usr/local/lib/python3.7/dist-packages (3.0.9)

Requirement already satisfied: cyclert>=0.10 in /usr/local/lib/python3.7/dist-packages (0.10.0)

Requirement already satisfied: kiwisolver>=1.0.1 in /usr/local/lib/python3.7/dist-packages (1.4.4)

Requirement already satisfied: python-dateutil>=2.1 in /usr/local/lib/python3.7/dist-packages (2.8.2)

Requirement already satisfied: typing-extensions in /usr/local/lib/python3.7/dist-packages (4.1.1)

Requirement already satisfied: pytz>=2017.3 in /usr/local/lib/python3.7/dist-packages (2022.7.1)

```
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.7/dist-packages (from
Requirement already satisfied: threadpoolctl>=2.0.0 in /usr/local/lib/python3.7/dist-pac
```

◀ ▶

```
pip install mlxtend --upgrade --no-deps
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public
Requirement already satisfied: mlxtend in /usr/local/lib/python3.7/dist-packages (0.20.6)
```

◀ ▶

```
from mlxtend.plotting import plot_pca_correlation_graph
```

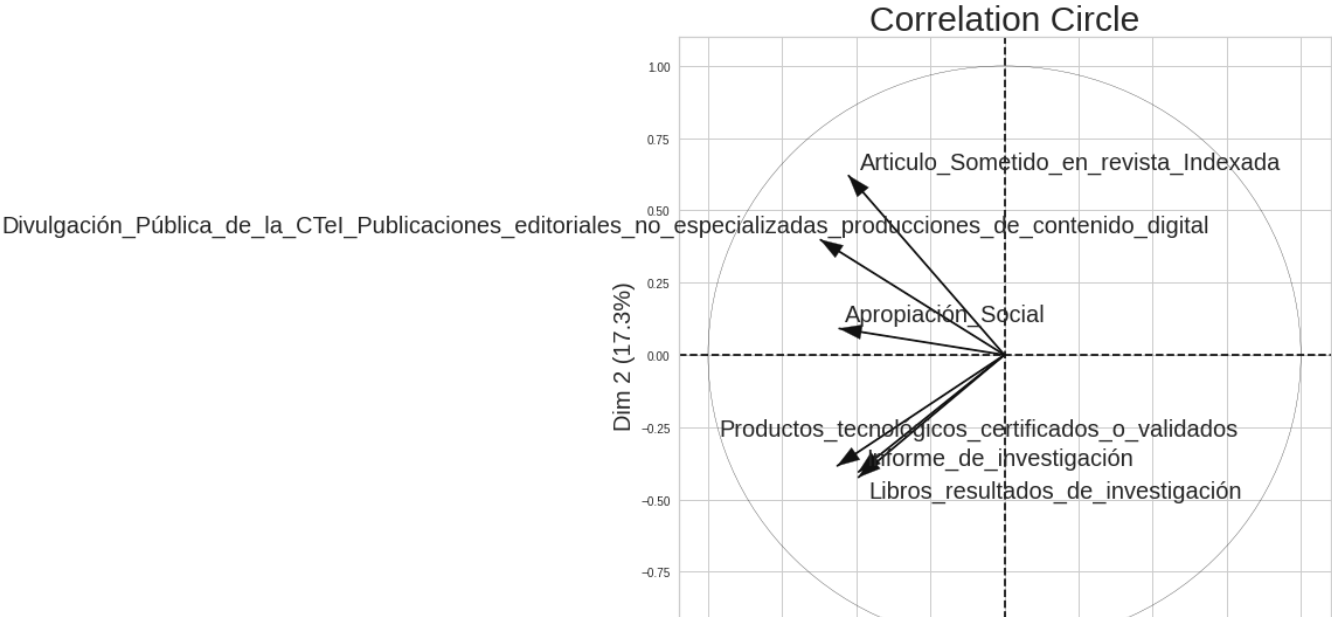
```
df.columns
```

```
Index(['Centro_Año', 'Apropiación_Social',
       'Articulo_Sometido_en_revista_Indexada',
       'Divulgación_Pública_de_la_CTeI_Publicaciones_editoriales_no_especializadas_producciones',
       'Informe_de_investigación', 'Libros_resultados_de_investigación',
       'Productos_tecnológicos_certificados_o_validados', 'KMeans_Clusters'],
      dtype='object')
```

◀ ▶

```
x1= df.drop(['Centro_Año','KMeans_Clusters'],axis = 1)
features = ['Apropiación_Social',
            'Articulo_Sometido_en_revista_Indexada',
            'Divulgación_Pública_de_la_CTeI_Publicaciones_editoriales_no_especializadas_produccion',
            'Informe_de_investigación', 'Libros_resultados_de_investigación',
            'Productos_tecnológicos_certificados_o_validados']
```

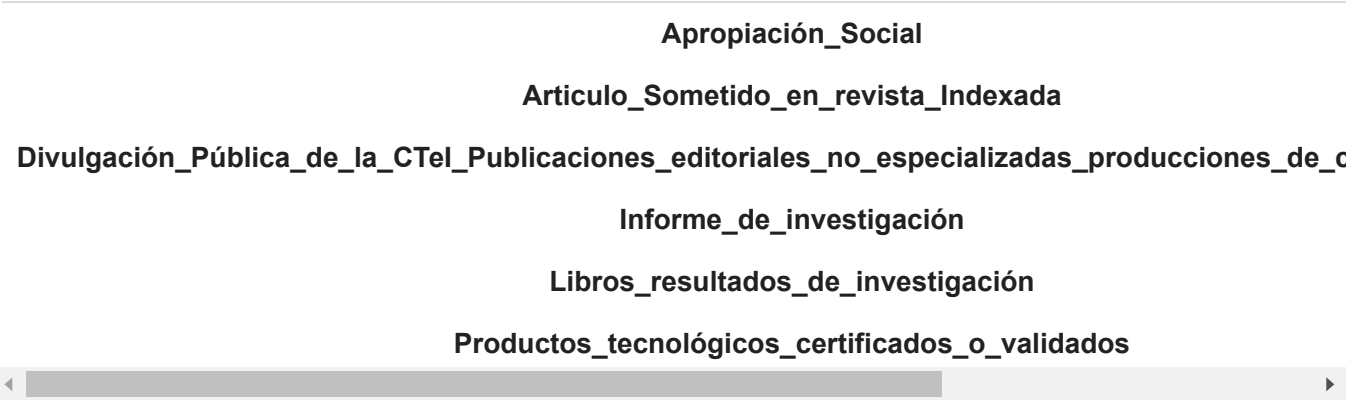
```
figure, correlation_matrix = plot_pca_correlation_graph(x1, features, dimensions=(1, 2), figu
```



Por tanto:

- **Cluster 2: Alta productividad**
- **Cluster 0: Media productividad**
- **Cluster 1: Baja Productividad**

correlation_matrix



x1

	Apropiación_Social	Articulo_Sometido_en_revista_Indexada	Divulgación_Pública_de_
0	-0.560125		0.857305
1	-0.560125		-1.274003
2	-1.189888		-1.274003
3	-1.189888		-0.563567
4	2.588687		0.146869
...	...		...
203	0.699400		2.988612
204	0.069637		0.857305
205	0.069637		-1.274003

```

gestion= df['KMeans_Clusters'].replace({0:0,2:1, 1:2})
gestion= df['KMeans_Clusters'].replace({0: 'Baja Productividad',2: 'Alta Productividad', 1: '
df['Productividad'] = pd.concat([gestion], axis=1)
df_k_clas = df
df_k_clas

```

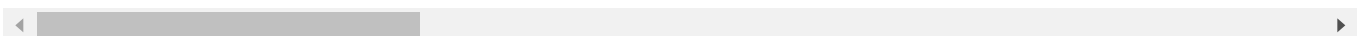

	Centro_Año	Apropiación_Social	Articulo_Sometido_en_revista_Indexada	Divulgación
0	9101-2010	-0.560125		0.857305
1	9127-2010	-0.560125		-1.274003
df_k_clas.to_excel('Productividad_Cluster_Kmeans.xlsx',index =False)				
2	9201-2010	-1.189888		-1.274003

▼ DENDOGRAMA

```
df = df.drop(['KMeans_Clusters','Productividad'], axis=1)
df
```

	Centro_Año	Apropiación_Social	Articulo_Sometido_en_revista_Indexada	Divulgación
0	9101-2010	-0.560125		0.857305
1	9127-2010	-0.560125		-1.274003
2	9201-2010	-1.189888		-1.274003
3	9202-2010	-1.189888		-0.563567
4	9203-2010	2.588687		0.146869
...	...	...		...
203	9501-2022	0.699400		2.988612
204	9502-2022	0.069637		0.857305
205	9503-2022	0.069637		-1.274003
206	9504-2022	1.329162		0.146869
207	9549-2022	-0.560125		-1.274003

208 rows × 7 columns



```
from scipy.cluster.hierarchy import linkage,dendrogram
df_el = df.drop(['Centro_Año'],axis =1)
labels = df['Centro_Año'].to_numpy().tolist()
```

```
x.columns
```

```
Index(['Apropiación_Social', 'Articulo_Sometido_en_revista_Indexada',
```

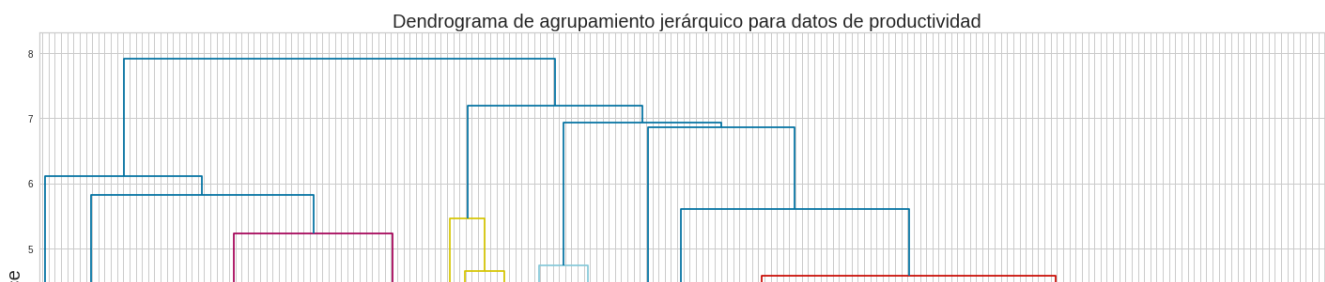
```
'Divulgación_Pública_de_la_CTeI_Publicaciones_editoriales_no_especializadas_producciones'

    'Informe_de_investigación', 'Libros_resultados_de_investigación',
    'Productos_tecnológicos_certificados_o_validados'],
dtype='object')
```



```
features = ['Apropiación_Social', 'Articulo_Sometido_en_revista_Indexada',
            'Divulgación_Pública_de_la_CTeI_Publicaciones_editoriales_no_especializadas_produccion',
            'Informe_de_investigación', 'Libros_resultados_de_investigación',
            'Productos_tecnológicos_certificados_o_validados']
X = x[features]
```

```
plt.figure(figsize=(20,10))
plt.title('Dendrograma de agrupamiento jerárquico para datos de productividad',fontsize=20)
plt.xlabel('sample index', fontsize=20)
plt.ylabel('distance',fontsize=20)
Z = linkage(X, 'complete','euclidean')
dn1 = sch.dendrogram(
    Z,
    labels=labels,
    leaf_rotation=90., # rotates the x axis labels
    leaf_font_size=15., # font size for the x axis labels
)
plt.tight_layout()
plt.show()
```



```
plt.figure(figsize=(20,10))
plt.title('Dendrograma de agrupamiento jerárquico para datos de productividad',fontsize=20)
plt.xlabel('sample index', fontsize=20)
plt.ylabel('distance',fontsize=20)
Z = linkage(X, 'ward','euclidean')
dn2 = sch.dendrogram(
    Z,
    labels=labels,
    leaf_rotation=90., # rotates the x axis labels
    leaf_font_size=15., # font size for the x axis labels
)
plt.tight_layout()
plt.show()
```

Dendrograma de agrupamiento jerárquico para datos de productividad

```

hac = AgglomerativeClustering(n_clusters=3,
                              affinity="euclidean",
                              linkage="ward")

```

```

hac.fit(x1)

```

```

AgglomerativeClustering(n_clusters=3)

```

```

hac.fit(X)

```

```

categorias = hac.labels_

```

```

categorias

```

```

array([0, 1, 1, 1, 0, 1, 1, 2, 1, 1, 1, 0, 1, 1, 0, 2, 1, 2, 1, 0, 1, 2,
       2, 0, 2, 0, 1, 0, 0, 1, 0, 1, 2, 1, 0, 0, 1, 0, 2, 1, 2, 1, 0, 0,
       2, 0, 1, 2, 0, 1, 0, 1, 0, 1, 1, 2, 2, 1, 1, 2, 1, 1, 1, 0, 1, 0,
       1, 0, 1, 1, 0, 1, 0, 0, 0, 1, 1, 1, 2, 0, 1, 0, 0, 1, 2, 1, 1, 1,
       1, 1, 2, 2, 1, 1, 1, 2, 1, 0, 1, 2, 1, 0, 1, 1, 1, 2, 0, 1, 1, 1,
       1, 1, 0, 1, 2, 0, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 2, 0, 0, 2, 1, 1,
       1, 2, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 0, 0, 2, 2, 0, 1, 2, 2, 1, 0,
       0, 1, 1, 0, 1, 1, 0, 2, 2, 0, 1, 0, 1, 2, 2, 0, 0, 1, 0, 0, 1, 0,
       1, 1, 2, 2, 2, 0, 1, 2, 0, 0, 2, 1, 2, 1, 0, 1, 1, 1, 2, 0, 1, 1,
       1, 1, 1, 1, 1, 2, 1, 1, 0, 2])

```

```

#Agregando la clasificación en el archivo original

```

```

df["Hac_Clusters"]=categorias

```

```

df

```

	Centro_Año	Apropiación_Social	Articulo_Sometido_en_revista_Indexada	Divulgación
0	9101-2010	-0.560125		0.857305

▼ Validación

índices de validación externa

```
#Calcular agrupación de índices de validación externa con KMeans
labels_true = df["Centro_Año"]
rotulos_km = hac.labels_
h = sm.homogeneity_score(labels_true, rotulos_km)
print("Homogeneidad: %0.3f" % h)
c = sm.completeness_score(labels_true, rotulos_km)
print("Compleitud: %0.3f" % c)
v = sm.v_measure_score(labels_true, rotulos_km)
print("Medida V: %0.3f" % v )
ari = sm.adjusted_rand_score(labels_true, rotulos_km)
print("Índice Rand corregido: %0.3f" % ari)
ami = sm.adjusted_mutual_info_score(labels_true, rotulos_km, average_method='arithmetic')
print("Adjusted Mutual Information: %0.3f" % ami )
fm = sm.fowlkes_mallows_score(labels_true, rotulos_km)
print("Índice Fowlkes-Mallows: %0.3f" % fm )
```

```
Homogeneidad: 0.192
Compleitud: 1.000
Medida V: 0.322
Índice Rand corregido: 0.000
Adjusted Mutual Information: 0.000
Índice Fowlkes-Mallows: 0.000
```

índices de validación interna

```
from sklearn.metrics import silhouette_score
#silhouette_avg = silhouette_score(x, clustering.labels_)
rotulos_hac = hac.labels_
s= sm.silhouette_score(x, rotulos_hac, metric='euclidean')
print("Coeficiente de Silhueta: %0.3f" % s)
ch = sm.calinski_harabasz_score(x,rotulos_hac)
print("Calinski_harabasz: %0.3f" % ch)
dbs = sm.davies_bouldin_score(x, rotulos_hac)
print("Davies Bouldin: %0.3f" % dbs)
```

```
Coeficiente de Silhueta: 0.162
Calinski_harabasz: 31.817
Davies Bouldin: 2.157
```

```
x.head()
```

	Apropiación_Social	Articulo_Sometido_en_revista_Indexada	Divulgación_Pública_de_la
0	-0.560125		0.857305
1	-0.560125		-1.274003
2	-1.189888		-1.274003
3	-1.189888		-0.563567
4	2.588687		0.146869

PCA

```
#Visualización de los Clusters que se crearon
#Aplicaremos Análisis de componetes principales (PCA) para darnos una idea de cómo se crearon
pca=PCA(n_components=2)
pca_productividad=pca.fit_transform(x)
pca_productividad_2=pd.DataFrame(data=pca_productividad,columns=["Componente_1", "Componente_2"])
pca_productividad_3=pd.concat([pca_productividad_2,df[["Hac_Clusters"]]],axis=1)
pca_productividad_3=pd.concat([pca_productividad_3,df[["Centro_Año"]]],axis=1)
pca_productividad_3
```

	Componente_1	Componente_2	Hac_Clusters	Centro_Año
0	0.816244	-0.662001	0	9101-2010
1	-0.847087	-1.759406	1	9127-2010
2	-2.365945	-0.431087	1	9201-2010
3	-1.898267	-0.256295	1	9202-2010
4	2.185298	-0.601068	0	9203-2010
...	...	...	...	...
203	2.164333	3.604048	2	9501-2022
204	-0.179709	1.214482	1	9502-2022
205	-0.765314	-0.367135	1	9503-2022
206	-0.552922	0.376368	0	9504-2022
207	0.674241	-1.491958	2	9549-2022

208 rows × 4 columns

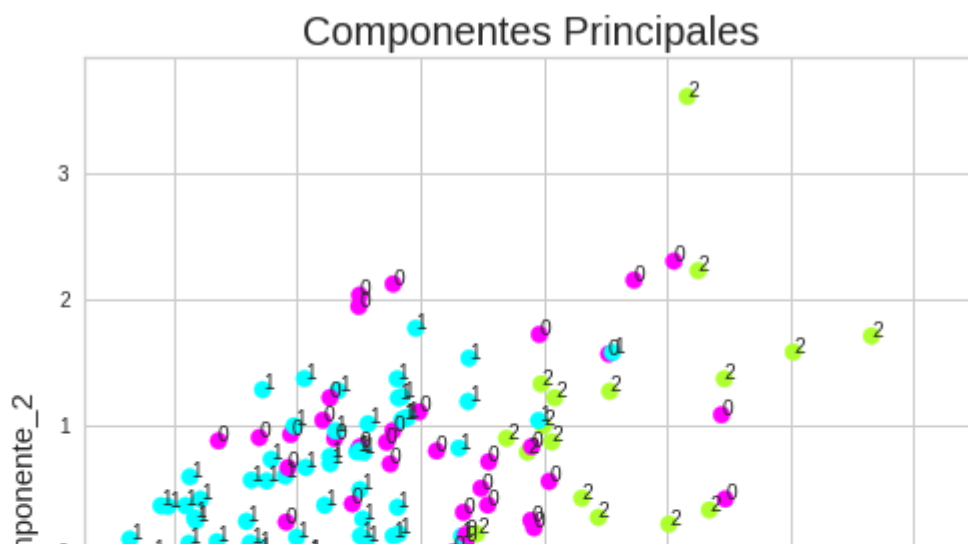
```
pca_productividad_3 = pca_productividad_3.dropna()
```

```
pca_productividad_3['Hac_Clusters'] = pca_productividad_3['Hac_Clusters'].astype(int)
pca_productividad_3
```

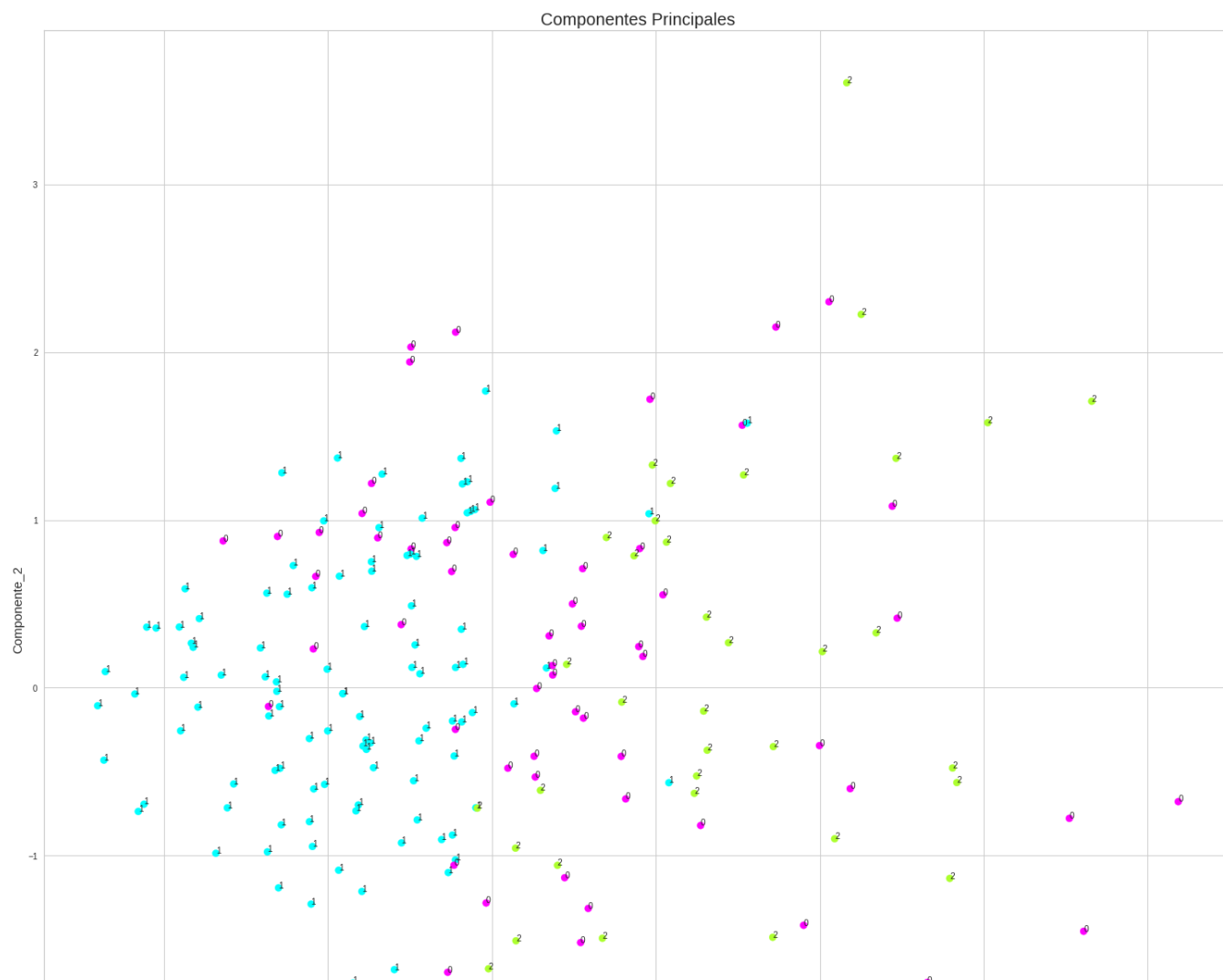
	Componente_1	Componente_2	Hac_Clusters	Centro_Año
0	0.816244	-0.662001	0	9101-2010
1	-0.847087	-1.759406	1	9127-2010
2	-2.365945	-0.431087	1	9201-2010
3	-1.898267	-0.256295	1	9202-2010
4	2.185298	-0.601068	0	9203-2010
...	...	...	...	...
203	2.164333	3.604048	2	9501-2022
204	-0.179709	1.214482	1	9502-2022
205	-0.765314	-0.367135	1	9503-2022
206	-0.552922	0.376368	0	9504-2022
207	0.674241	-1.491958	2	9549-2022

208 rows × 4 columns

```
fig=plt.figure(figsize=(8,8))
ax=fig.add_subplot(1,1,1)
annotations = df['Hac_Clusters']
ax.set_xlabel("Componente_1",fontsize=15)
ax.set_ylabel("Componente_2", fontsize=15)
ax.set_title("Componentes Principales", fontsize=20)
color_theme=np.array(["MAGENTA","CYAN",'GREENYELLOW','mediumpurple'])
ax.scatter(x=pca_productividad_3['Componente_1'], y=pca_productividad_3['Componente_2'],c=col
for i, label in enumerate(annotations):
    plt.annotate(label, (pca_productividad_3['Componente_1'][i], pca_productividad_3['Compone
plt.show()
```



```
fig=plt.figure(figsize=(25,25))
ax=fig.add_subplot(1,1,1)
annotations = pca_productividad_3['Hac_Clusters']
ax.set_xlabel("Componente_1",fontsize=15)
ax.set_ylabel("Componente_2", fontsize=15)
ax.set_title("Componentes Principales", fontsize=20)
color_theme=np.array(["MAGENTA","CYAN",'GREENYELLOW','mediumpurple'])
ax.scatter(x=pca_productividad_3['Componente_1'], y=pca_productividad_3['Componente_2'],c=col
for i, label in enumerate(annotations):
    plt.annotate(label, (pca_productividad_3['Componente_1'][i], pca_productividad_3['Compone
plt.show()
```

x1.columns

```
Index(['Apropiación_Social', 'Articulo_Sometido_en_revista_Indexada',
      'Divulgación_Pública_de_la_CTeI_Publicaciones_editoriales_no_especializadas_producciones',
      'Informe_de_investigación', 'Libros_resultados_de_investigación',
      'Productos_tecnológicos_certificados_o_validados'],
      dtype='object')
```

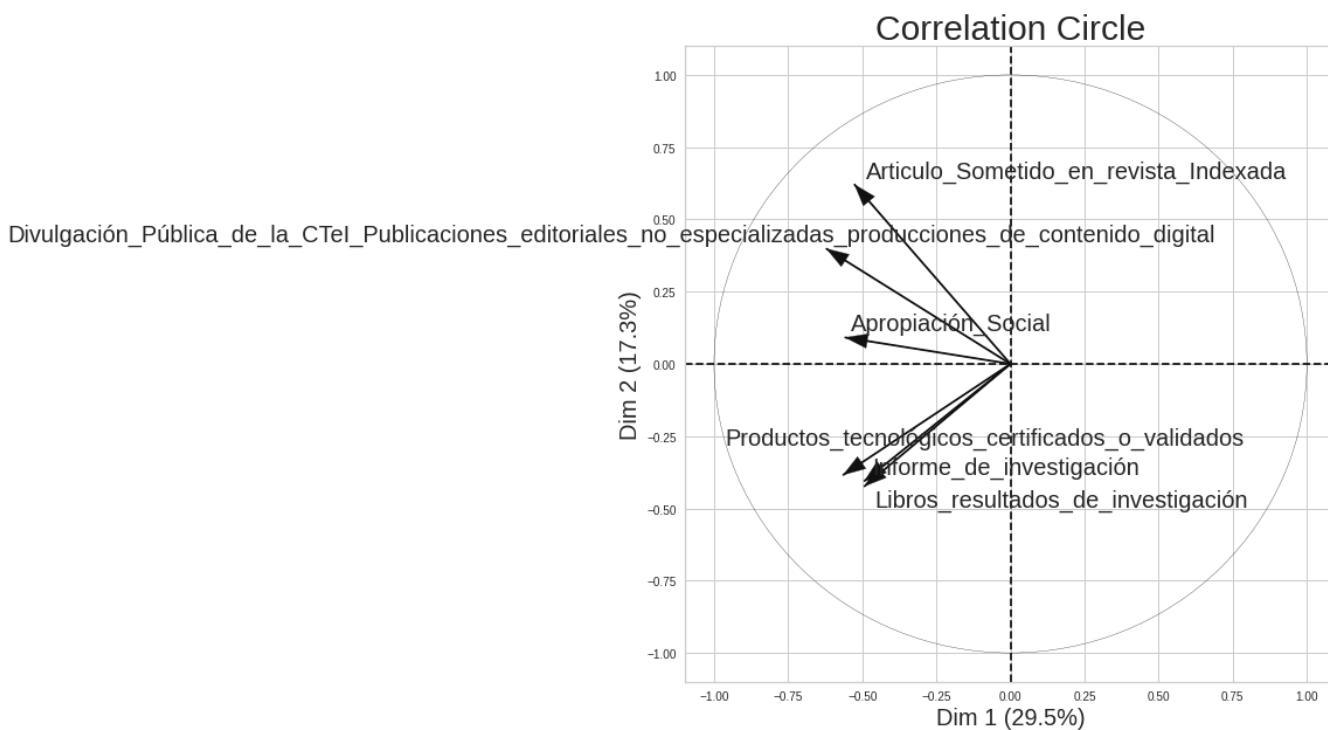
df.columns

```
Index(['Centro_Año', 'Apropiación_Social',
      'Articulo_Sometido_en_revista_Indexada',
      'Divulgación_Pública_de_la_CTeI_Publicaciones_editoriales_no_especializadas_producciones',
      'Informe_de_investigación', 'Libros_resultados_de_investigación',
      'Productos_tecnológicos_certificados_o_validados', 'Hac_Clusters'],
      dtype='object')
```

```

x1= df.drop(['Centro_Año', 'Hac_Clusters'],axis = 1)
features = ['Apropiación_Social', 'Articulo_Sometido_en_revista_Indexada',
            'Divulgación_Pública_de_la_CTeI_Publicaciones_editoriales_no_especializadas_produccion
            'Informe_de_investigación', 'Libros_resultados_de_investigación',
            'Productos_tecnológicos_certificados_o_validados']
figure, correlation_matrix = plot_pca_correlation_graph(x1, features,
                                                         dimensions=(1, 2),
                                                         figure_axis_size=10)

```



correlation_matrix

Apropiación_Social**Articulo_Sometido_en_revista_Indexada****Divulgación_Pública_de_la_CTel_Publicaciones_editoriales_no_especializadas_producciones_de_c**

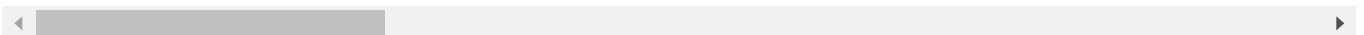
```

productividad= df['Hac_Clusters'].replace({0: 'Baja Productividad',2: 'Media Productividad',
df['Productividad'] = pd.concat([productividad], axis=1)
df_hac_clas = df
df_hac_clas

```

	Centro_Año	Apropiación_Social	Articulo_Sometido_en_revista_Indexada	Divulgación
0	9101-2010	-0.560125		0.857305
1	9127-2010	-0.560125		-1.274003
2	9201-2010	-1.189888		-1.274003
3	9202-2010	-1.189888		-0.563567
4	9203-2010	2.588687		0.146869
...	...	...		...
203	9501-2022	0.699400		2.988612
204	9502-2022	0.069637		0.857305
205	9503-2022	0.069637		-1.274003
206	9504-2022	1.329162		0.146869
207	9549-2022	-0.560125		-1.274003

208 rows × 9 columns



```
df_hac_clas.to_excel('Productividad_Cluster_Dendograma.xlsx', index =False)
```

```
df = df_k_clas
centros_años = df["Centro_Año"].str.split("-", expand=True)
centros_años.columns = ['Codigo_Centro', 'Año']
df = pd.concat([df, centros_años], axis=1)
df.head()
```

	Centro_Año	Apropiación_Social	Articulo_Sometido_en_revista_Indexada	Divulgación_P
0	9101-2010	-0.560125		0.857305
1	9127-2010	-0.560125		-1.274003
2	9201-2010	-1.189888		-1.274003
3	9202-2010	-1.189888		-0.563567
4	9203-2010	2.588687		0.146869

```
df.columns
```

```
Index(['Centro_Año', 'Apropiación_Social',
      'Articulo_Sometido_en_revista_Indexada',
      'Divulgación_Pública_de_la_CTeI_Publicaciones_editoriales_no_especializadas_producciones',
      'Informe_de_investigación', 'Libros_resultados_de_investigación',
      'Productos_tecnológicos_certificados_o_validados', 'KMeans_Clusters',
      'Productividad', 'Codigo_Centro', 'Año'],
      dtype='object')
```

```
df_map = df[['Codigo_Centro', 'Año', 'KMeans_Clusters', 'Productividad']]
df_map
```

	Codigo_Centro	Año	KMeans_Clusters	Productividad
0	9101	2010	0	Baja Productividad
1	9127	2010	2	Alta Productividad
2	9201	2010	2	Alta Productividad
3	9202	2010	2	Alta Productividad
4	9203	2010	0	Baja Productividad
...	...	...	...	...

```
df_map['Codigo_Centro'] = df_map['Codigo_Centro'].astype(int)
```

```
df_map['Año'] = df_map['Año'].astype(int)
```

```
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead
```

See the caveats in the documentation: <https://pandas.pydata.org/pandas-docs/stable/user>

```
"""Entry point for launching an IPython kernel.
```

```
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead
```

See the caveats in the documentation: <https://pandas.pydata.org/pandas-docs/stable/user>



```
df_map = df_map.merge(df_centros, on='Codigo_Centro', how='inner')
```

```
df_map.head()
```

	Codigo_Centro	Año	KMeans_Clusters	Productividad	
0	9101	2010	0	Baja Productividad	Centro_de_los_Recursos_Naturales_f
1	9101	2011	2	Alta Productividad	Centro_de_los_Recursos_Naturales_f
2	9101	2012	0	Baja Productividad	Centro_de_los_Recursos_Naturales_f
3	9101	2013	1	Media Productividad	Centro_de_los_Recursos_Naturales_f
4	9101	2014	2	Alta Productividad	Centro_de_los_Recursos_Naturales_f



```
df_map = df_map.drop(['Codigo_Regional','Codigo_Municipio','Codigo_Departamento'],axis= 1)
```

```
df_map
```

	Codigo_Centro	Año	KMeans_Clusters	Productividad	
0	9101	2010	0	Baja Productividad	Centro_de_los_Recurso
1	9101	2011	2	Alta Productividad	Centro_de_los_Recurso
2	9101	2012	0	Baja Productividad	Centro_de_los_Recurso
3	9101	2013	1	Media Productividad	Centro_de_los_Recurso
4	9101	2014	2	Alta Productividad	Centro_de_los_Recurso
...	...	...	...	...	...
203	9549	2018	2	Alta Productividad	Complejo_Tecnológico_Turístico
204	9549	2019	2	Alta Productividad	Complejo_Tecnológico_Turístico

▼ Mapa para 2010

```
df_map_2010 = df_map[df_map['Año']<2011]
df_map_2010 = df_map_2010.drop(['Año'],axis = 1)
```

```
latitude = 6.51033052
longitude = -75.48477721
```

```
map_clusters = folium.Map(location=[latitude, longitude], zoom_start=4)
```

```
# set color scheme for the clusters
x = np.arange(2)
ys = [i + x + (i*x)**2 for i in range(4)]
colors_array = cm.rainbow(np.linspace(0, 1, len(ys)))
rainbow = [colors.rgb2hex(i) for i in colors_array]
```

```
# add markers to the map
markers_colors = []
for lat, lng, cluster, cent, gest in zip(df_map_2010['latitud'],
                                         df_map_2010['longitud'],
```

```
df_map_2010['KMeans_Clusters'],  
df_map_2010['Nombre_Centro'],  
df_map_2010['Productividad']) :
```

```
label = folium.Popup(str(cent) + ' - Cluster ' + str(cluster), parse_html=True)  
folium.vector_layers.CircleMarker(  
    [lat, lng],  
    radius=6,  
    #popup=label,  
    tooltip = str(cent) + ' , ' + str(gest),  
    color=rainbow[cluster-1],  
    fill=True,  
    fill_color=rainbow[cluster-1],  
    fill_opacity=0.9).add_to(map_clusters)
```

```
map_clusters
```

Make this Notebook Trusted to load map: File -> Trust Notebook

▼ Mapa para 2014

```
df_map_2014 = df_map[df_map['Año']<2015]
df_map_2014 = df_map_2014.drop(['Año'],axis = 1)

latitude = 6.51033052
longitude = -75.48477721

map_clusters = folium.Map(location=[latitude, longitude], zoom_start=4)

# set color scheme for the clusters
x = np.arange(2)
ys = [i + x + (i*x)**2 for i in range(4)]
colors_array = cm.rainbow(np.linspace(0, 1, len(ys)))
rainbow = [colors.rgb2hex(i) for i in colors_array]

# add markers to the map
markers_colors = []
for lat, lng, cluster, cent, gest in zip(df_map_2014['latitud'],
                                         df_map_2014['longitud'],
                                         df_map_2014['KMeans_Clusters'],
                                         df_map_2014['Nombre_Centro'],
                                         df_map_2014['Productividad']) :

    label = folium.Popup(str(cent) + ' - Cluster ' + str(cluster), parse_html=True)
    folium.vector_layers.CircleMarker(
        [lat, lng],
        radius=6,
        #popup=label,
        tooltip = str(cent) + ' , ' + str(gest),
        color=rainbow[cluster-1],
        fill=True,
        fill_color=rainbow[cluster-1],
        fill_opacity=0.9).add_to(map_clusters)

map_clusters
```


Make this Notebook Trusted to load map: File -> Trust Notebook

▼ Mapa para 2018

```
df_map_2018 = df_map[df_map['Año']<2019]
df_map_2018 = df_map_2018.drop(['Año'],axis = 1)
```

```
latitude = 6.51033052
longitude = -75.48477721
```

```
map_clusters = folium.Map(location=[latitude, longitude], zoom_start=4)
```

```
# set color scheme for the clusters
x = np.arange(2)
ys = [i + x + (i*x)**2 for i in range(4)]
colors_array = cm.rainbow(np.linspace(0, 1, len(ys)))
rainbow = [colors.rgb2hex(i) for i in colors_array]
```

```
# add markers to the map
markers_colors = []
for lat, lng, cluster, cent, gest in zip(df_map_2018['latitud'],
                                         df_map_2018['longitud'],
                                         df_map_2018['KMeans_Clusters'],
                                         df_map_2018['Nombre_Centro'],
                                         df_map_2018['Productividad']) :

    label = folium.Popup(str(cent) + '- Cluster ' + str(cluster), parse_html=True)
    folium.vector_layers.CircleMarker(
        [lat, lng],
        radius=6,
        #popup=label,
        tooltip = str(cent) + ' , ' + str(gest),
        color=rainbow[cluster-1],
        fill=True,
        fill_color=rainbow[cluster-1],
        fill_opacity=0.9).add_to(map_clusters)

map_clusters
```

Make this Notebook Trusted to load map: File -> Trust Notebook

```
df_map_2022 = df_map[df_map['Año']<2023]
df_map_2022 = df_map_2022.drop(['Año'],axis = 1)

latitude = 6.51033052
longitude = -75.48477721

map_clusters = folium.Map(location=[latitude, longitude], zoom_start=4)

# set color scheme for the clusters
x = np.arange(2)
ys = [i + x + (i*x)**2 for i in range(4)]
colors_array = cm.rainbow(np.linspace(0, 1, len(ys)))
rainbow = [colors.rgb2hex(i) for i in colors_array]

# add markers to the map
markers_colors = []
for lat, lng, cluster, cent, gest in zip(df_map_2022['latitud'],
                                         df_map_2022['longitud'],
                                         df_map_2022['KMeans_Clusters'],
                                         df_map_2022['Nombre_Centro'],
                                         df_map_2022['Productividad']) :

    label = folium.Popup(str(cent) + ' - Cluster ' + str(cluster), parse_html=True)
    folium.vector_layers.CircleMarker(
        [lat, lng],
        radius=6,
        #popup=label,
        tooltip = str(cent) + ' , ' + str(gest),
        color=rainbow[cluster-1],
        fill=True,
        fill_color=rainbow[cluster-1],
        fill_opacity=0.9).add_to(map_clusters)

map_clusters
```

Make this Notebook Trusted to load map: File -> Trust Notebook

▼ RED NEURONAL: PERCEPTRON

```
df.columns
```

```
Index(['Centro_Año', 'Apropiación_Social',  
      'Articulo_Sometido_en_revista_Indexada',  
      'Divulgación_Pública_de_la_CTeI_Publicaciones_editoriales_no_especializadas_producciones',  
      'Informe_de_investigación', 'Libros_resultados_de_investigación',  
      'Productos_tecnológicos_certificados_o_validados', 'KMeans_Clusters',  
      'Productividad', 'Codigo_Centro', 'Año'],  
      dtype='object')
```



```
df_product_orig = pd.read_excel(path2)
```

```
df_product_orig.columns
df_prd = df[['Codigo_Centro', 'Año', 'Productividad']]

df_per = pd.concat([df_prd, df_product_orig], axis=1).drop(['Centro_Año'], axis=1)
df_per
```

	Codigo_Centro	Año	Productividad	P_8020	P_8025	P_8019	P_8039	P_8016	P_8031
0	9101	2010	Baja Productividad	1.0	3.0	1.0	6.0	0.0	3.0
1	9127	2010	Alta Productividad	1.0	0.0	0.0	2.0	3.0	2.0
2	9201	2010	Alta Productividad	0.0	0.0	0.0	0.0	0.0	2.0
3	9202	2010	Alta Productividad	0.0	1.0	0.0	2.0	0.0	1.0
4	9203	2010	Baja Productividad	6.0	2.0	2.0	4.0	2.0	3.0
...	...	...	...	...	...	...	...	...	...
203	9501	2022	Media Productividad	3.0	6.0	6.0	1.0	1.0	1.0
204	9502	2022	Media Productividad	2.0	3.0	2.0	1.0	1.0	1.0

```
Productividad = df_per['Productividad'].replace({'Baja Productividad': 1, 'Media Productividad': 2})
df_per['Productividad'] = pd.concat([Productividad], axis=1)
df_per
```

	Codigo_Centro	Año	Productividad	P_8020	P_8025	P_8019	P_8039	P_8016	P_8031
0	9101	2010	1	1.0	3.0	1.0	6.0	0.0	3.0

```
#Clases
```

```
y = (df_per['Productividad']).astype(np.int8)
print(y)
```

```
0      1
1      3
2      3
3      3
4      1
..
203    2
204    2
205    3
206    2
207    1
```

```
Name: Productividad, Length: 208, dtype: int8
```

```
208 rows x 0 columns
```

Dividimos los datos en un conjunto de aprendizaje y de prueba:

```
from sklearn.model_selection import train_test_split
datasets = train_test_split(df_per,
                             y,
                             test_size=0.3)
```

```
X_train, X_test, y_train, y_test = datasets
```

Ahora, creamos una instancia de Perceptron y ajustamos los datos de entrenamiento:

```
from sklearn.linear_model import Perceptron
p = Perceptron(random_state=42,
                max_iter=100,
                tol=0.001)
p.fit(X_train, y_train)

Perceptron(max_iter=100, random_state=42)
```

Recordando un poco sobre la matriz de confusión

		Actual Values	
		Yes	No
Predicted Values	Yes	True Positive	False Positive
	No	False Negative	True Negative

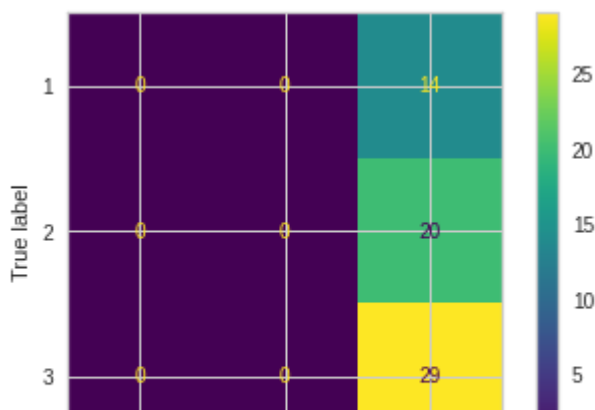
```
from sklearn.metrics import classification_report
y_perc_pred = p.predict(X_test)
print(classification_report(y_test, y_perc_pred))
```

	precision	recall	f1-score	support
1	0.00	0.00	0.00	14
2	0.00	0.00	0.00	20
3	0.46	1.00	0.63	29
accuracy			0.46	63
macro avg	0.15	0.33	0.21	63
weighted avg	0.21	0.46	0.29	63

```
/usr/local/lib/python3.7/dist-packages/sklearn/metrics/_classification.py:1318: Undefined
_warn_prf(average, modifier, msg_start, len(result))
/usr/local/lib/python3.7/dist-packages/sklearn/metrics/_classification.py:1318: Undefined
_warn_prf(average, modifier, msg_start, len(result))
/usr/local/lib/python3.7/dist-packages/sklearn/metrics/_classification.py:1318: Undefined
_warn_prf(average, modifier, msg_start, len(result))
```

```
plot_confusion_matrix(p, X_test, y_test)
plt.show()
```

```
/usr/local/lib/python3.7/dist-packages/sklearn/utils/deprecation.py:87: FutureWarning: F
warnings.warn(msg, category=FutureWarning)
```



Test con valores de Prueba

ACKNOWLEDGMENTS

```
path3 = '/content/drive/MyDrive/Proyecto Sennova borrador/sample.xlsx'
df_sample = pd.read_excel(path3)
df_sample.drop(['Productividad'], axis = 1)
```

	Codigo_Centro	Año	P_8020	P_8025	P_8019	P_8039	P_8016	P_8031
0	9101.0	2010.0	10.0	5.0	5.0	5.0	5.0	5.0
1	9127.0	2010.0	5.0	5.0	5.0	5.0	5.0	5.0
2	9201.0	2010.0	0.0	0.0	0.0	0.0	0.0	1.0
3	9202.0	2010.0	0.0	0.0	0.0	0.0	0.0	0.0
4	9203.0	2010.0	0.0	1.0	0.0	1.0	2.0	0.0
5	9504.0	2022.0	12.0	12.0	12.0	12.0	12.0	12.0

```
y_percep_perd = p.predict(df_sample)
y_percep_perd_s= pd.Series(y_percep_perd)
df_percep = df_sample[['Codigo_Centro', 'Año', 'Productividad']]
df_percep['Productividad'] = y_percep_perd_s.replace({1: 'Baja Productividad', 2: 'Media Produ
df_percep
```



```
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:4: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
```

```
Try using .loc[row_indexer,col_indexer] = value instead
```

See the caveats in the documentation: <https://pandas.pydata.org/pandas-docs/stable/user>
after removing the cwd from sys.path.

```
#muestra_metricas(df_sample['Productividad'], p.predict(df_sample))
```

▼ Regresión Logística

```
3          9202.0  2010.0  Alta Productividad
```

```
from sklearn.linear_model import LogisticRegression
```

```
from sklearn.model_selection import train_test_split
datasets = train_test_split(df_per,
                            y,
                            test_size=0.3)
```

```
X_train, X_test, y_train, y_test = datasets
```

```
# Instanciando a reg. logistica
reglog = LogisticRegression()
```

```
# Aplicando el modelo
reglog.fit(X_train, y_train)
```

```
/usr/local/lib/python3.7/dist-packages/sklearn/linear_model/_logistic.py:818: ConvergenceWarning:
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.
```

Increase the number of iterations (max_iter) or scale the data as shown in:

<https://scikit-learn.org/stable/modules/preprocessing.html>

Please also refer to the documentation for alternative solver options:

https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression

```
extra_warning_msg=_LOGISTIC_SOLVER_CONVERGENCE_MSG,
LogisticRegression()
```

```
LogisticRegression()
```

```
LogisticRegression()
```

```
# Predicción
y_reglog_pred = reglog.predict(X_test)
```

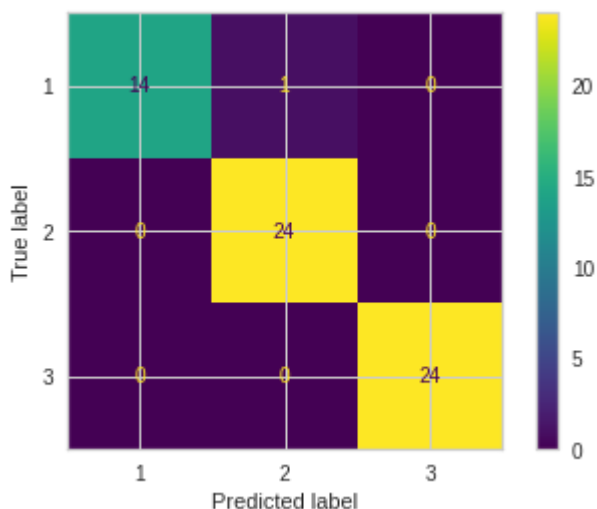
```
print(classification_report(y_test,y_reglog_pred))
```

	precision	recall	f1-score	support
1	1.00	0.93	0.97	15
2	0.96	1.00	0.98	24
3	1.00	1.00	1.00	24
accuracy			0.98	63
macro avg	0.99	0.98	0.98	63
weighted avg	0.98	0.98	0.98	63

Matriz de confusión

```
plot_confusion_matrix(reglog, X_test, y_test)
plt.show()
```

/usr/local/lib/python3.7/dist-packages/sklearn/utils/deprecation.py:87: FutureWarning: F
warnings.warn(msg, category=FutureWarning)



Test con valores de Prueba

```
df_sample.drop(['Productividad'], axis =1)
```

	Codigo_Centro	Año	P_8020	P_8025	P_8019	P_8039	P_8016	P_8031
0	9101.0	2010.0	10.0	5.0	5.0	5.0	5.0	5.0
1	9127.0	2010.0	5.0	5.0	5.0	5.0	5.0	5.0
2	9201.0	2010.0	0.0	0.0	0.0	0.0	0.0	1.0
3	9202.0	2010.0	0.0	0.0	0.0	0.0	0.0	0.0
4	9203.0	2010.0	0.0	1.0	0.0	1.0	2.0	0.0
5	9504.0	2022.0	12.0	12.0	12.0	12.0	12.0	12.0

```
y_reglog_pred = reglog.predict(df_sample)
y_reglog_pred = pd.Series(y_reglog_pred)
```

```
df_logist = df_sample[['Codigo_Centro', 'Año', 'Productividad']]
df_logist['Productividad'] = y_reglog_pred.replace({1: 'Baja Productividad', 2: 'Media Productividad'})
df_logist
```

/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:5: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/10min.html#setting-with-copy-warning

	Codigo_Centro	Año	Productividad
0	9101.0	2010.0	Baja Productividad
1	9127.0	2010.0	Baja Productividad
2	9201.0	2010.0	Alta Productividad
3	9202.0	2010.0	Alta Productividad
4	9203.0	2010.0	Media Productividad
5	9504.0	2022.0	Baja Productividad

▼ Random Forest

```
from sklearn.ensemble import RandomForestClassifier
```

```
rf = RandomForestClassifier()
```

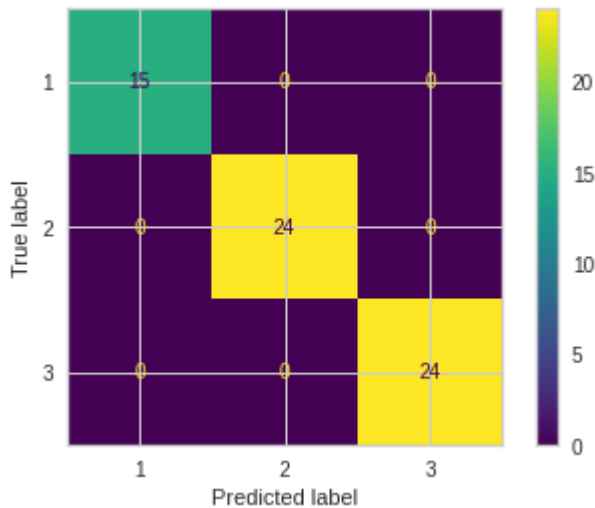
```
rf.fit(X_train, y_train)
```

```
y_dt_pred = rf.predict(X_test)
print(classification_report(y_test, y_dt_pred))
```

	precision	recall	f1-score	support
1	1.00	1.00	1.00	15
2	1.00	1.00	1.00	24
3	1.00	1.00	1.00	24
accuracy			1.00	63
macro avg	1.00	1.00	1.00	63
weighted avg	1.00	1.00	1.00	63

```
# Matriz de confusión
plot_confusion_matrix(rf, X_test, y_test)
plt.show()
```

/usr/local/lib/python3.7/dist-packages/sklearn/utils/deprecation.py:87: FutureWarning: f
warnings.warn(msg, category=FutureWarning)



Test con valores de Prueba

```
df_sample.drop(['Productividad'], axis =1)
```

	Codigo_Centro	Año	P_8020	P_8025	P_8019	P_8039	P_8016	P_8031
0	9101.0	2010.0	10.0	5.0	5.0	5.0	5.0	5.0
1	9127.0	2010.0	5.0	5.0	5.0	5.0	5.0	5.0
2	9201.0	2010.0	0.0	0.0	0.0	0.0	0.0	1.0
3	9202.0	2010.0	0.0	0.0	0.0	0.0	0.0	0.0
4	9203.0	2010.0	0.0	1.0	0.0	1.0	2.0	0.0
5	9504.0	2022.0	12.0	12.0	12.0	12.0	12.0	12.0

```
y_dt_pred_ram = rf.predict(df_sample)
y_dt_pred_ram = pd.Series(y_dt_pred_ram)
```

```
df_rd_forest = df_sample[['Codigo_Centro', 'Año', 'Productividad']]
df_rd_forest['Productividad'] = y_dt_pred_ram.replace({1: 'Baja Productividad', 2: 'Media Prod
df_rd_forest
```

/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:5: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.

Try using `.loc[row_indexer,col_indexer] = value` instead

See the caveats in the documentation: <https://pandas.pydata.org/pandas-docs/stable/user>

```

"""

```

	Codigo_Centro	Año	Productividad
0	9101.0	2010.0	Baja Productividad
1	9127.0	2010.0	Baja Productividad
2	9201.0	2010.0	Baja Productividad
3	9202.0	2010.0	Baja Productividad
.			

▼ SVM

```

from sklearn.svm import SVC
y_svm = SVC(C = 100, kernel = 'linear', random_state=123)
y_svm.fit(X_train, y_train)

```

```
SVC(C=100, kernel='linear', random_state=123)
```

```

y_svm_pred = y_svm.predict(X_test)
print(classification_report(y_test,y_svm_pred))

```

	precision	recall	f1-score	support
1	0.81	0.87	0.84	15
2	0.92	0.96	0.94	24
3	1.00	0.92	0.96	24
accuracy			0.92	63
macro avg	0.91	0.91	0.91	63
weighted avg	0.92	0.92	0.92	63

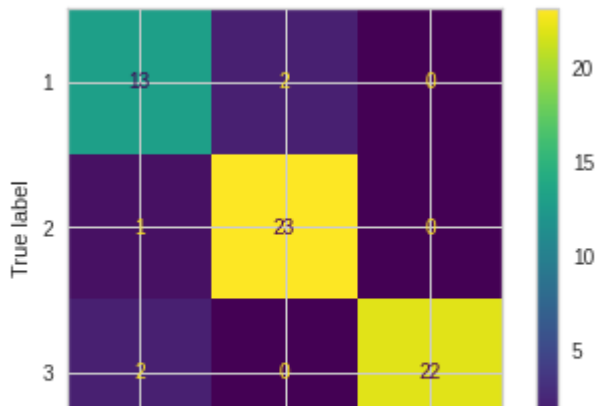
Test con valores de muestra

```

# Matriz de confusión
plot_confusion_matrix(y_svm, X_test, y_test, include_values =True)
plt.show()

```

```
/usr/local/lib/python3.7/dist-packages/sklearn/utils/deprecation.py:87: FutureWarning: F
warnings.warn(msg, category=FutureWarning)
```



```
df_sample.drop(['Productividad'], axis =1)
```

	Codigo_Centro	Año	P_8020	P_8025	P_8019	P_8039	P_8016	P_8031
0	9101.0	2010.0	10.0	5.0	5.0	5.0	5.0	5.0
1	9127.0	2010.0	5.0	5.0	5.0	5.0	5.0	5.0
2	9201.0	2010.0	0.0	0.0	0.0	0.0	0.0	1.0
3	9202.0	2010.0	0.0	0.0	0.0	0.0	0.0	0.0
4	9203.0	2010.0	0.0	1.0	0.0	1.0	2.0	0.0
5	9504.0	2022.0	12.0	12.0	12.0	12.0	12.0	12.0

```
y_dt_pred_svm = pd.Series(y_svm.predict(df_sample))
df_SVM = df_sample[['Codigo_Centro','Año','Productividad']]
df_SVM['Productividad'] = y_dt_pred_svm.replace({1: 'Baja Productividad',2: 'Media Productivi
df_SVM
```

```
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:3: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead
```

See the caveats in the documentation: <https://pandas.pydata.org/pandas-docs/stable/user>
This is separate from the ipykernel package so we can avoid doing imports until

	Codigo_Centro	Año	Productividad
0	9101.0	2010.0	Baja Productividad
1	9127.0	2010.0	Baja Productividad
2	9201.0	2010.0	Media Productividad
3	9202.0	2010.0	Media Productividad
4	9203.0	2010.0	Media Productividad
5	9504.0	2022.0	Baja Productividad

▼ Reporte de parametros de validación de modelos

```
from sklearn.metrics import accuracy_score, confusion_matrix, f1_score, precision_score, recall_score

p.predict(X_test)
reglog.predict(X_test)
rf.predict(X_test)
y_svm.predict(X_test)
y_test
```

Reporte de parámetros de validación

```
# perceptrón
acc_perc = accuracy_score(y_test, p.predict(X_test))
f1_perc = f1_score(y_test, p.predict(X_test), average='weighted')
#auc_perc = roc_auc_score(y_test, p.predict(X_test))
#recall_perc = recall_score(y_test, p.predict(X_test))

# Regresión logística
acc_log = accuracy_score(y_test, reglog.predict(X_test))
f1_log = f1_score(y_test, reglog.predict(X_test), average='weighted')

#Arboles de decisión
acc_rf = accuracy_score(y_test, rf.predict(X_test))
f1_rf = f1_score(y_test, rf.predict(X_test), average='weighted')

#Máquinas de soporte vectorial
acc_svm = accuracy_score(y_test, y_svm.predict(X_test))
f1_svm = f1_score(y_test, y_svm.predict(X_test), average='weighted')

Parametro = ['Accuracy', 'F1-Score']
Percep= [acc_perc, f1_perc]
r_log = [acc_log, f1_log]
rfor = [acc_rf, f1_rf]
svm_t = [acc_svm, f1_svm]

df_valid = pd.DataFrame()
df_valid['Parámetro'] = Parametro
df_valid['Perceptrón'] = Percep
df_valid['Regresión logística'] = r_log
df_valid['Random Forest'] = rfor
df_valid['Máquina de soporte vectorial'] = svm_t
```

df_valid

	Parámetro	Perceptrón	Regresión logística	Random Forest	Máquina de soporte vectorial
0	Accuracy	0.380952	0.984127	1.0	0.920635
1	F1-Score	0.210181	0.984015	1.0	0.921711

df_valid.to_excel('Validacion_Productividad.xlsx',index =False)


```

import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)

import folium # plotting library
from folium import plugins
from folium.plugins import MarkerCluster

from sklearn.cluster import KMeans
from sklearn.decomposition import PCA
import sklearn.metrics as sm
from sklearn.decomposition import PCA
from sklearn.cluster import AgglomerativeClustering
from sklearn.cluster import DBSCAN
from io import IncrementalNewlineDecoder
import scipy.cluster.hierarchy as sch

#Cluster
from yellowbrick.cluster import SilhouetteVisualizer
from yellowbrick.cluster import KElbowVisualizer
from sklearn.cluster import KMeans
from sklearn.datasets import make_blobs

#Graficas
import matplotlib.pyplot as plt
from matplotlib import pyplot as plot
import matplotlib.cm as cm
import matplotlib.colors as colors
from plotly import graph_objects as go
from plotly.subplots import make_subplots
import plotly.express as px
import seaborn as sns
import branca
from sklearn.metrics import plot_confusion_matrix

from google.colab import drive
drive.mount('/content/drive')

```

Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mour



```

from google.colab import files
path1 = "/content/drive/MyDrive/Proyecto SGPS 9748/TPProceso.xlsx"
path2 = "/content/drive/MyDrive/Proyecto SGPS 9748/Datos.xlsx"
df = pd.read_excel(path1)
df_centros = pd.read_excel(path2, sheet_name= 'Centros_Formacion')
df

```

	Centro_Año	Q_Proyectos	QNA_5007	QNA_5009	QNA_5010	TV_Proyectos
0	9101-2010	1	1	2	0	4.652924e+08
1	9127-2010	2	2	4	1	3.920052e+08
2	9201-2010	1	4	1	1	4.556627e+08
3	9202-2010	3	5	5	2	4.071852e+08
4	9203-2010	2	4	3	1	2.906175e+08
...	...	...	...	...	...	...
203	9501-2022	2	3	7	0	3.754077e+08
204	9502-2022	1	2	1	1	3.977569e+08
205	9503-2022	1	2	1	0	5.068536e+08
206	9504-2022	4	9	6	2	7.614273e+08
207	9549-2022	2	3	6	1	3.294535e+08

208 rows × 6 columns

df.columns

```
Index(['Centro_Año', 'Q_Proyectos', 'QNA_5007', 'QNA_5009', 'QNA_5010',
      'TV_Proyectos'],
      dtype='object')
```

```
df.columns = [ 'Centro_Año', 'Cantidad_Proyectos', 'Cantidad_Profesional', 'Cantidad_Maestria',
               'Cantidad_Doctorado', 'Total_Valores_Proyectos']
#df_2010_orig = df[df['Fecha']<2011]
#df_2010_orig = df_2010_orig.drop(['Fecha'],axis = 1)
df
```

	Centro_Año	Cantidad_Proyectos	Cantidad_Profesional	Cantidad_Maestria	Cant
0	9101-2010	1	1	2	
1	9127-2010	2	2	4	
2	9201-2010	1	4	1	

```
df.describe()
```

	Cantidad_Proyectos	Cantidad_Profesional	Cantidad_Maestria	Cantidad_Doctorado
count	208.000000	208.000000	208.000000	208.000000
mean	2.403846	3.668269	4.129808	1.250000
std	1.116579	2.419836	2.585257	1.116579
min	1.000000	0.000000	0.000000	0.000000
25%	2.000000	2.000000	2.000000	0.000000
50%	2.000000	3.000000	4.000000	1.000000
75%	3.000000	5.000000	6.000000	2.000000
max	6.000000	14.000000	13.000000	5.000000

```
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
df[['Cantidad_Proyectos', 'Cantidad_Profesional', 'Cantidad_Maestria', 'Cantidad_Doctorado', 'Total']] = scaler.fit_transform(df[['Cantidad_Proyectos', 'Cantidad_Profesional', 'Cantidad_Maestria', 'Cantidad_Doctorado', 'Total']])
```

Centro_Año	Cantidad_Proyectos	Cantidad_Profesional	Cantidad_Maestria	Cant
------------	--------------------	----------------------	-------------------	------

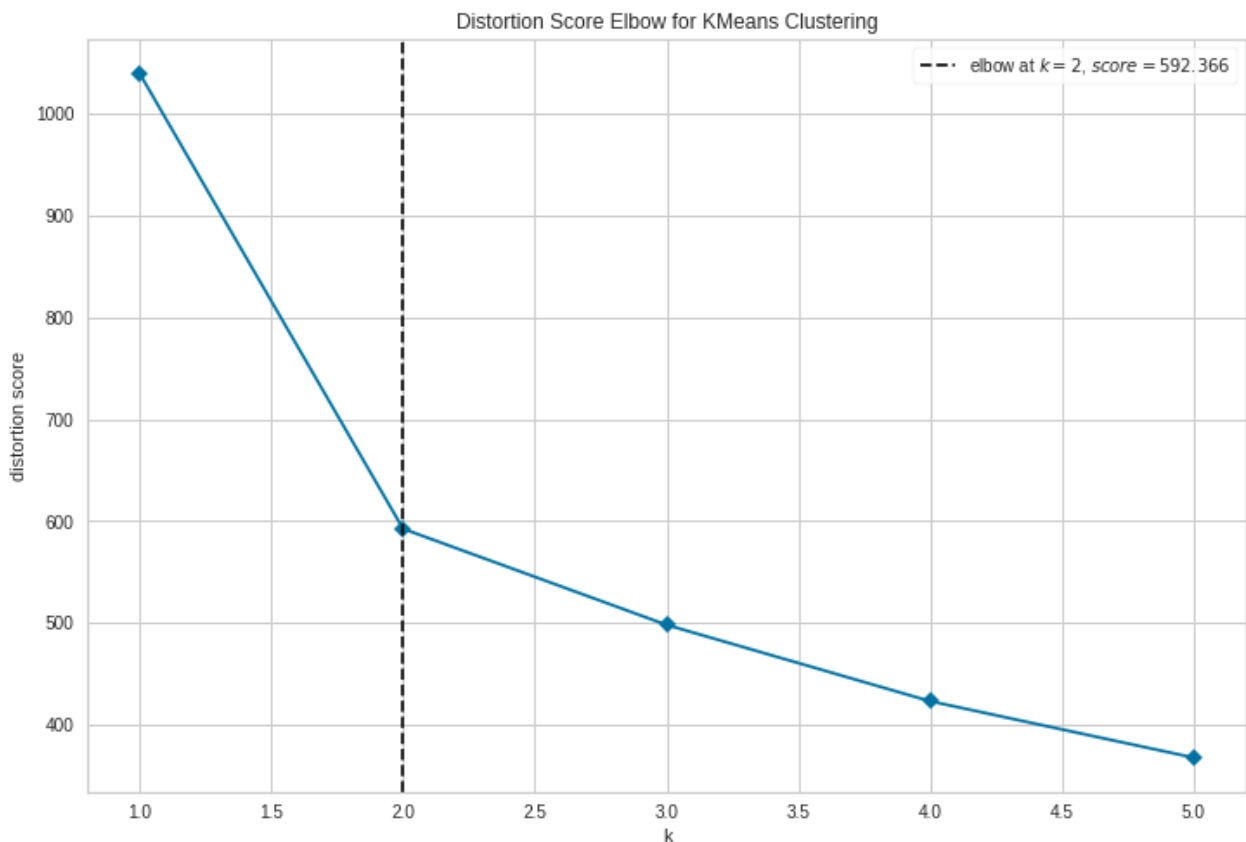
```
#Búsqueda de la cantidad óptima de clusters
##Calculando qué tan similares son los individuos dentro de los Custers
y = df[['Centro_Año']]
x = df.drop(['Centro_Año'], axis= 1)
wcss = []
for i in range (1,6):
    kmeans = KMeans(n_clusters=i, max_iter=300, random_state=0)
    kmeans.fit(x) #Aplico K-means a la base de datos
    wcss.append(kmeans.inertia_)
```

203	9501-2022	-0.362554	-0.276829	1.112894
-----	-----------	-----------	-----------	----------

distortion: suma media de las distancias al cuadrado de los centroides (Error medio cuadrático)

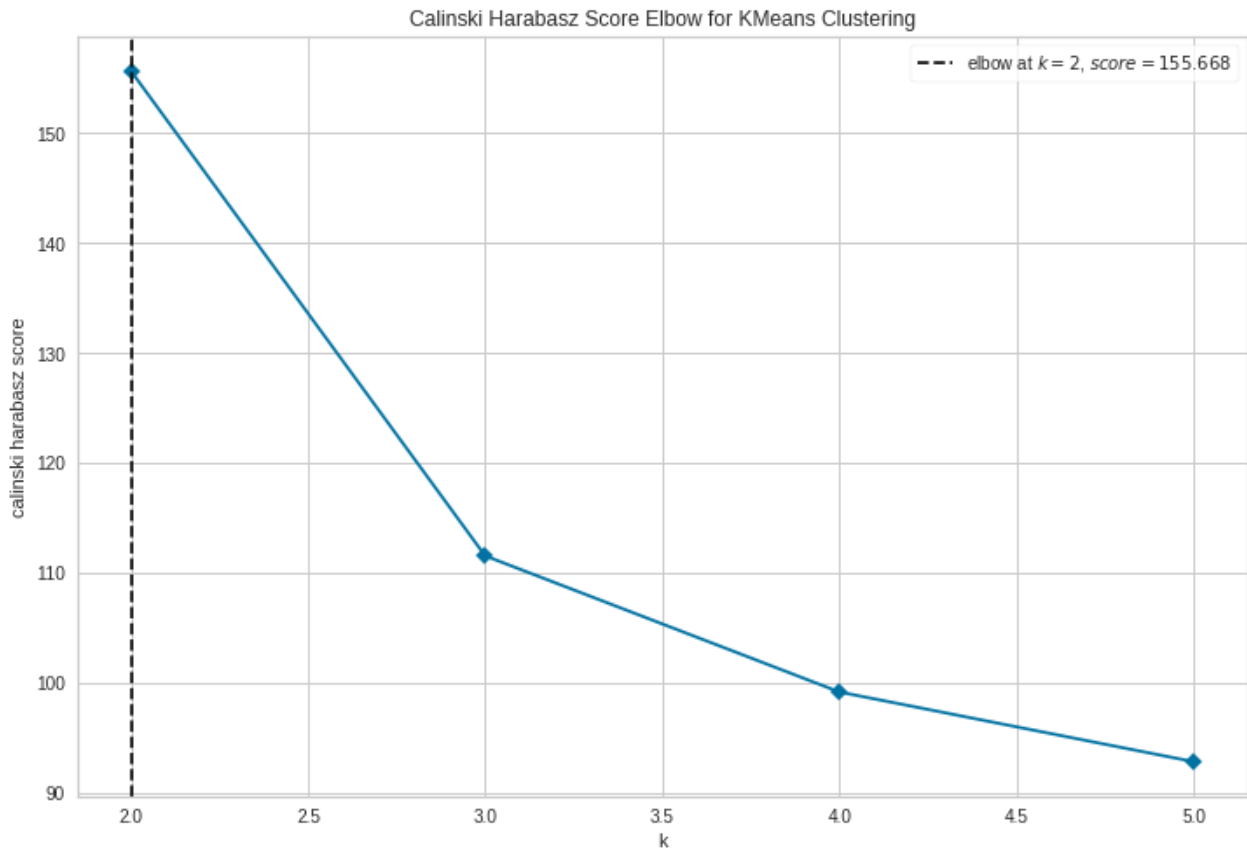
205	9502-2022	-1.260308	-0.601077	-1.213558
-----	-----------	-----------	-----------	-----------

```
#km_transformed = mms.transform(km)
plt.figure(figsize=(12,8))
model = KMeans(random_state=0)
visualizer = KElbowVisualizer(model, k=(1,6), timings=False)
visualizer.fit(x)
visualizer.show();
```



calinski_harabasz: relación de dentro entre a la dispersión del grupo

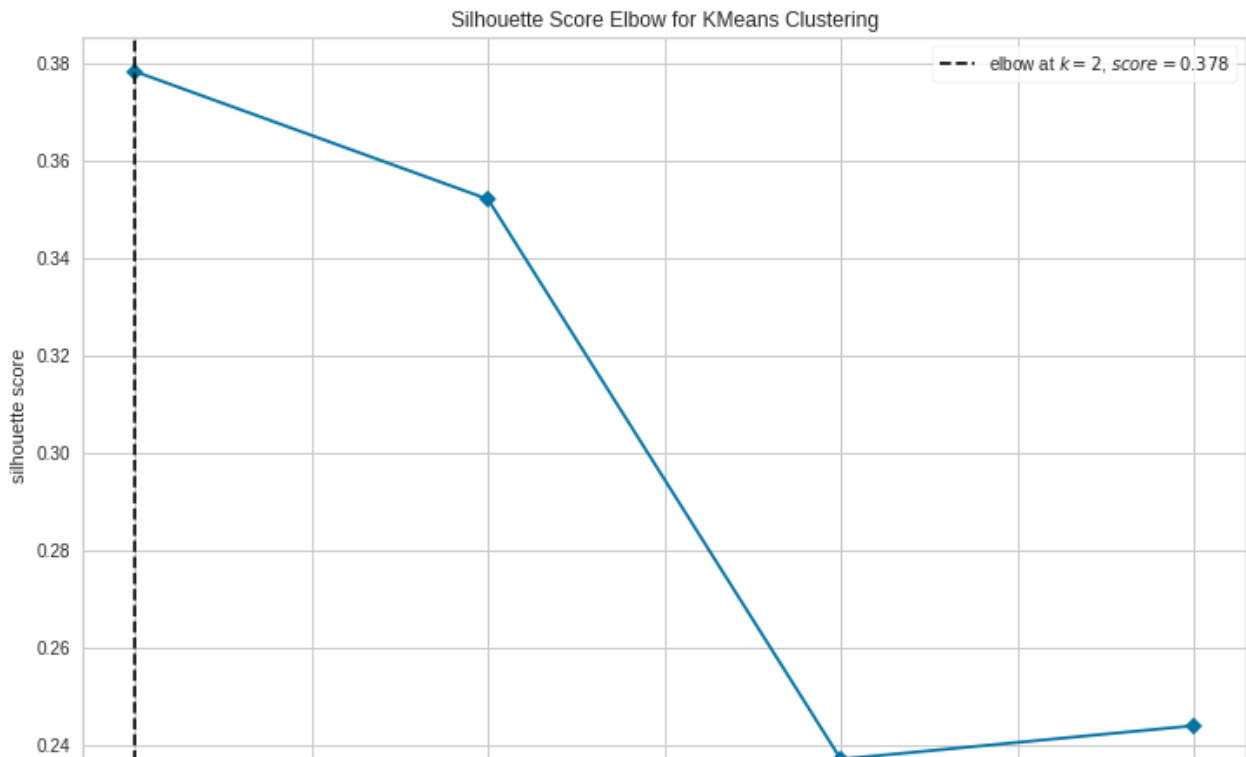
```
plt.figure(figsize=(12,8))
visualizer = KElbowVisualizer( model, k=(2,6), metric='calinski_harabasz', timings=False)
visualizer.fit(x)          # Fit the data to the visualizer
visualizer.show()         # Finalize and render the figure
```



<matplotlib.axes._subplots.AxesSubplot at 0x7f5ee4f17d10>

silhouette : relación media de la distancia intra-cluster y del cluster más cercano

```
plt.figure(figsize=(12,8))
visualizer = KElbowVisualizer( model, k=(2,6), metric='silhouette', timings=False)
visualizer.fit(x)          # Fit the data to the visualizer
visualizer.show()         # Finalize and render the figure
```



```
#Aplicando el método de k-means a la base de datos
clustering=KMeans(n_clusters=3, max_iter=300,random_state=0) #Crea el modelo
clustering.fit(x) #Aplica el modelo a la base de datos
clustering.cluster_centers_

array([[ -0.69369284, -0.55525834, -0.44760853, -0.44318905, -0.57404786],
       [ 0.86697762,  0.6597317 , -0.03347355,  1.31380732,  0.70543525],
       [ 1.11873891,  0.93484648,  1.40370059, -0.15915182,  0.93959545]])
```

Se obtienen dos centroides con coordenadas en 5 dimensiones(cantidad de variables), por lo que no se genera visualización. Se debe reducir dimensionalidad como PCA.

```
#Agregando la clasificación en el archivo original
df["KMeans_Clusters"]=clustering.labels_
df
```

	Centro_Año	Cantidad_Proyectos	Cantidad_Profesional	Cantidad_Maestria	Cant
0	9101-2010	-1.260308	-1.105326	-0.825816	
1	9127-2010	-0.362554	-0.691077	-0.050332	
2	9201-2010	-1.260308	0.137419	-1.213558	
3	9202-2010	0.535199	0.551667	0.337410	
4	9203-2010	-0.362554	0.137419	-0.438074	

```
df.to_excel('TPProceso_Cluster_Kmeans.xlsx', index = False)
```

▼ Validación de clusters

205	9503-2022	-1.260308	-0.691077	-1.213558
-----	-----------	-----------	-----------	-----------

Indices externos

207	9510-2022	-0.362554	0.376820	0.705150
-----	-----------	-----------	----------	----------

```
#Calcular agrupación de índices de validación externa con KMeans
labels_true = df["Centro_Año"]
rotulos_km = clustering.labels_
h = sm.homogeneity_score(labels_true, rotulos_km)
print("Homogeneidad: %0.3f" % h)
c = sm.completeness_score(labels_true, rotulos_km)
print("Compleitud: %0.3f" % c)
v = sm.v_measure_score(labels_true, rotulos_km)
print("Medida V: %0.3f" % v)
ari = sm.adjusted_rand_score(labels_true, rotulos_km)
print("Índice Rand corregido: %0.3f" % ari)
ami = sm.adjusted_mutual_info_score(labels_true, rotulos_km, average_method='arithmetic')
print("Adjusted Mutual Information: %0.3f" % ami)
fm = sm.fowlkes_mallows_score(labels_true, rotulos_km)
print("Índice Fowlkes-Mallows: %0.3f" % fm)
```

```
Homogeneidad: 0.181
Compleitud: 1.000
Medida V: 0.306
Índice Rand corregido: 0.000
Adjusted Mutual Information: 0.000
Índice Fowlkes-Mallows: 0.000
```

índices internos

```
from sklearn.metrics import silhouette_score
#silhouette_avg = silhouette_score(x, clustering.labels_)
s = sm.silhouette_score(x, clustering.labels_, metric='euclidean')
print("Coeficiente de Silhueta: %0.3f" % s)
ch = sm.calinski_harabasz_score(x, clustering.labels_)
```

```
print("Calinski_harabasz: %0.3f" % ch)
dbs = sm.davies_bouldin_score(x, clustering.labels_)
print("Davies Bouldin: %0.3f" % dbs)
```

Coeficiente de Silhueta: 0.352
Calinski_harabasz: 111.570
Davies Bouldin: 1.393

▼ DBSCAN CLUSTERING

```
x.head()
```

	Cantidad_Proyectos	Cantidad_Profesional	Cantidad_Maestria	Cantidad_Doctorado
0	-1.260308	-1.105326	-0.825816	-1.110658
1	-0.362554	-0.691077	-0.050332	-0.225536
2	-1.260308	0.137419	-1.213558	-0.225536
3	0.535199	0.551667	0.337410	0.659586
4	-0.362554	0.137419	-0.438074	-0.225536

```
db = DBSCAN(eps=0.3, min_samples=7, metric='euclidean').fit(x)
db.labels_
```

[illegible]

```
labels = db.labels_  
labels
```

```
array([-1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1,  
       -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1,  
        -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1,  
        -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1,  
        -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1])
```



```
-1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1,
-1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1,
-1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1,
-1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1,
-1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1,
-1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1,
-1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1,
-1, -1, -1, -1]]
```

```
#Contar número de clusters
#labels = db.labels_
#n_clusters_ = len(set(labels)) - (1 if -1 in labels else 0)
#n_noise_ = list(labels).count(-1)
#print('Número estimado de clusters: %d' % n_clusters_)
#print('Número estimado de datos ruidosos: %d' % n_noise_)
#print("Adjusted Rand Index: %0.3f"
#      % metrics.adjusted_rand_score(labels_true, labels))
#print("Adjusted Mutual Information: %0.3f"
#      % metrics.adjusted_mutual_info_score(labels_true, labels))
#print("Silhouette Coefficient: %0.3f"
#      % metrics.silhouette_score(X, labels))
```

▼ PCA

```
#Visualización de los Clusters que se crearon
#Aplicaremos Análisis de componentes principales (PCA) para darnos una idea de cómo se crearon
pca=PCA(n_components=2)
pca_productividad=pca.fit_transform(x)
pca_productividad_2=pd.DataFrame(data=pca_productividad,columns=["Componente_1", "Componente_2"])
pca_productividad_3=pd.concat([pca_productividad_2,df[["KMeans_Clusters"]]],axis=1)
pca_productividad_3=pd.concat([pca_productividad_3,df[["Centro_Año"]]],axis=1)
pca_productividad_3
```

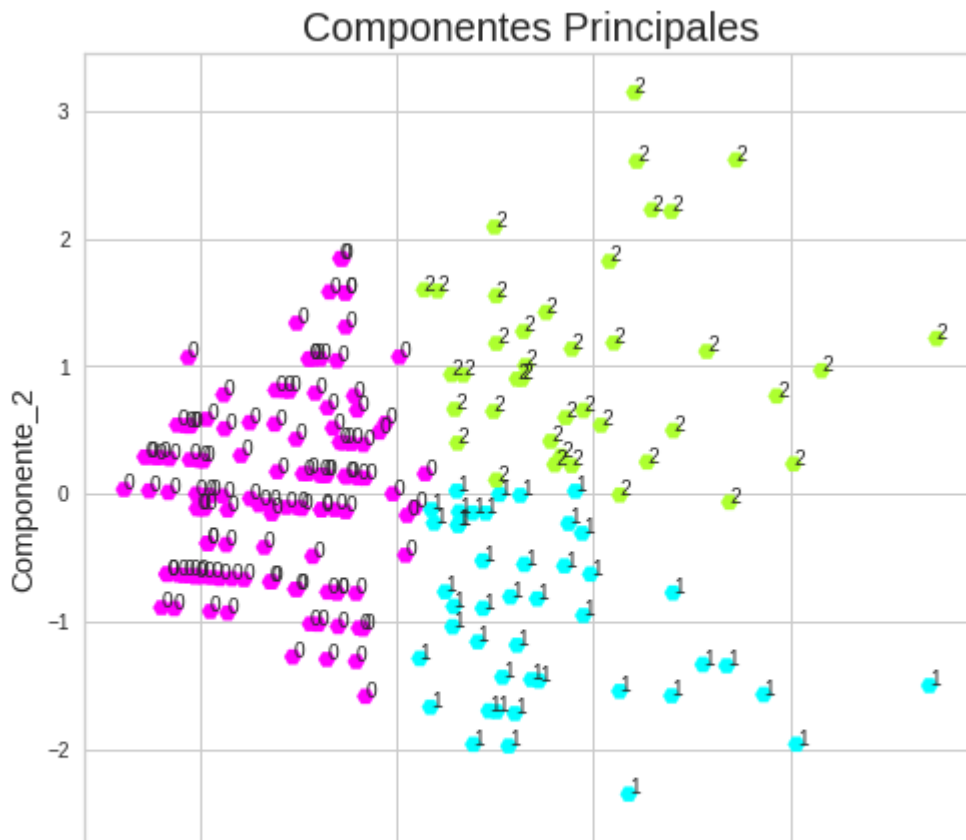
	Componente_1	Componente_2	KMeans_Clusters	Centro_Año
0	-1.963914	0.259801	0	9101-2010
1	-0.951451	0.156047	0	9127-2010
2	-1.285903	-0.690315	0	9201-2010

```
pca_productividad_3 = pca_productividad_3.dropna()
pca_productividad_3['KMeans_Clusters'] = pca_productividad_3['KMeans_Clusters'].astype(int)
pca_productividad_3
```

	Componente_1	Componente_2	KMeans_Clusters	Centro_Año
0	-1.963914	0.259801	0	9101-2010
1	-0.951451	0.156047	0	9127-2010
2	-1.285903	-0.690315	0	9201-2010
3	0.622891	-0.246133	1	9202-2010
4	-1.123852	-0.106162	0	9203-2010
...	...	...	...	...
203	-0.682772	1.580615	0	9501-2022
204	-1.881752	-0.655856	0	9502-2022
205	-1.769395	-0.017971	0	9503-2022
206	3.384948	-0.066004	2	9504-2022
207	-0.703090	0.672548	0	9549-2022

208 rows × 4 columns

```
fig=plt.figure(figsize=(8,8))
ax=fig.add_subplot(1,1,1)
annotations = df['KMeans_Clusters']
ax.set_xlabel("Componente_1",fontsize=15)
ax.set_ylabel("Componente_2", fontsize=15)
ax.set_title("Componentes Principales", fontsize=20)
color_theme=np.array(["MAGENTA","CYAN",'GREENYELLOW','mediumpurple'])
ax.scatter(x=pca_productividad_3['Componente_1'], y=pca_productividad_3['Componente_2'],c=col
for i, label in enumerate(annotations):
    plt.annotate(label, (pca_productividad_3['Componente_1'][i], pca_productividad_3['Compone
plt.show()
```



```
#revisar tamaño de etiqueta de punto
fig=plt.figure(figsize=(25,25))
ax=fig.add_subplot(1,1,1)
annotations = df['Centro_Año']
ax.set_xlabel("Componente_1",fontsize=15)
ax.set_ylabel("Componente_2", fontsize=15)
ax.set_title("Componentes Principales", fontsize=20)
color_theme=np.array(["MAGENTA","CYAN",'GREENYELLOW','mediumpurple'])
ax.scatter(x=pca_productividad_3['Componente_1'], y=pca_productividad_3['Componente_2'],c=col
for i, label in enumerate(annotations):
    plt.annotate(label, (pca_productividad_3['Componente_1'][i], pca_productividad_3['Compone
plt.show()
```



	Centro_Año	Cantidad_Proyectos	Cantidad_Profesional	Cantidad_Maestria	Cantidad_Doctorado
0	9101-2010	-1.260308	-1.105326	-0.825816	
1	9127-2010	-0.362554	-0.691077	-0.050332	
2	9201-2010	-1.260308	0.137419	-1.213558	
3	9202-2010	0.535199	0.551667	0.337410	
4	9203-2010	-0.362554	0.137419	-0.438074	

```
#df_2010 = df_2010.merge(df_centros,on='Codigo_Centro',how='inner')
#df_2010 = df_2010[['Codigo_Centro','Nombre_Centro','Cantidad_Proyectos','Cantidad_Profesional',
                    'Cantidad_Maestria','Cantidad_Doctorado','Total_Valor_Proyectos']]
204  9502-2022      -1.260308      -0.691077      -1.213558
#df_2010.head()
```

```
pip install mlxtend
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple
Requirement already satisfied: mlxtend in /usr/local/lib/python3.7/dist-packages (0.20.0)
Requirement already satisfied: setuptools in /usr/local/lib/python3.7/dist-packages (from mlxtend) (57.5.0)
Requirement already satisfied: pandas>=0.24.2 in /usr/local/lib/python3.7/dist-packages (from mlxtend) (1.1.5)
Requirement already satisfied: scikit-learn>=1.0.2 in /usr/local/lib/python3.7/dist-packages (from mlxtend) (0.24.2)
Requirement already satisfied: matplotlib>=3.0.0 in /usr/local/lib/python3.7/dist-packages (from mlxtend) (3.3.0)
Requirement already satisfied: joblib>=0.13.2 in /usr/local/lib/python3.7/dist-packages (from mlxtend) (1.1.0)
Requirement already satisfied: numpy>=1.16.2 in /usr/local/lib/python3.7/dist-packages (from mlxtend) (1.21.0)
Requirement already satisfied: scipy>=1.2.1 in /usr/local/lib/python3.7/dist-packages (from mlxtend) (1.7.3)
Requirement already satisfied: python-dateutil>=2.1 in /usr/local/lib/python3.7/dist-packages (from mlxtend) (2.8.2)
Requirement already satisfied: cycloper>=0.10 in /usr/local/lib/python3.7/dist-packages (from mlxtend) (0.10.0)
Requirement already satisfied: kiwisolver>=1.0.1 in /usr/local/lib/python3.7/dist-packages (from mlxtend) (1.3.2)
Requirement already satisfied: pyparsing!=2.0.4,!=2.1.2,!=2.1.6,>=2.0.1 in /usr/local/lib/python3.7/dist-packages (from mlxtend) (3.0.9)
Requirement already satisfied: typing-extensions in /usr/local/lib/python3.7/dist-packages (from mlxtend) (4.1.1)
Requirement already satisfied: pytz>=2017.3 in /usr/local/lib/python3.7/dist-packages (from mlxtend) (2022.1)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.7/dist-packages (from mlxtend) (1.16.0)
Requirement already satisfied: threadpoolctl>=2.0.0 in /usr/local/lib/python3.7/dist-packages (from mlxtend) (2.2.0)
```

```
pip install mlxtend --upgrade --no-deps
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple
Requirement already satisfied: mlxtend in /usr/local/lib/python3.7/dist-packages (0.20.0)
```

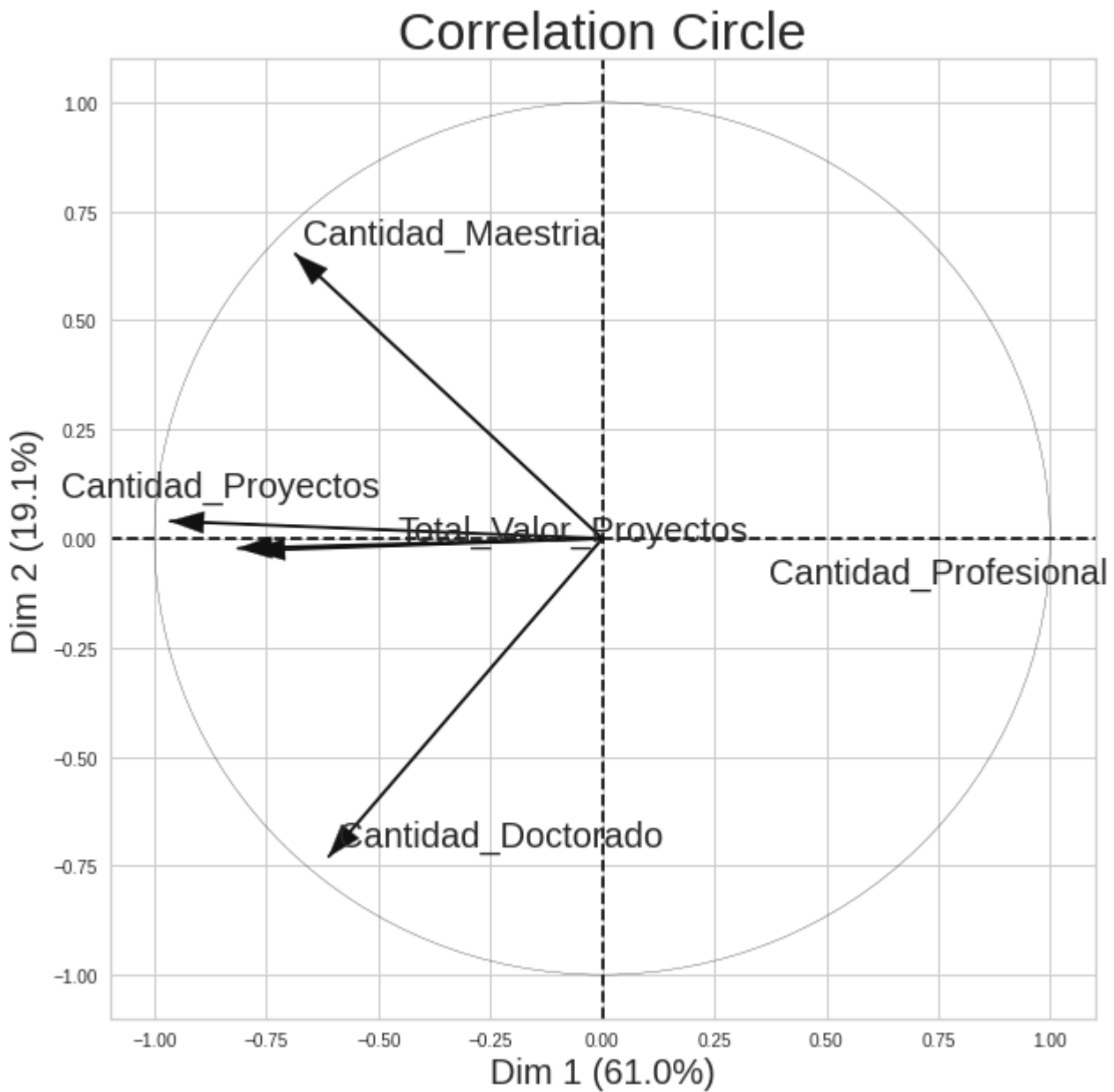
```
from mlxtend.plotting import plot_pca_correlation_graph
```

```
df.columns
```

```
Index(['Centro_Año', 'Cantidad_Proyectos', 'Cantidad_Profesional',
       'Cantidad_Maestria', 'Cantidad_Doctorado', 'Total_Valor_Proyectos',
```

```
'KMeans_Clusters'],  
dtype='object')
```

```
x1= df.drop(['Centro_Año','KMeans_Clusters'],axis = 1)  
features = ['Cantidad_Proyectos',  
            'Cantidad_Profesional', 'Cantidad_Maestria', 'Cantidad_Doctorado',  
            'Total_valor_Proyectos']  
figure, correlation_matrix = plot_pca_correlation_graph(x1, features,  
                                                         dimensions=(1, 2),  
                                                         figure_axis_size=10)
```



correlation_matrix

	Dim 1	Dim 2
Cantidad_Proyectos	-0.965193	0.039709
Cantidad_Profesional	-0.784122	-0.027681
Cantidad_Maestria	-0.685551	0.652113
Cantidad_Doctorado	-0.610983	-0.728949

x1

	Cantidad_Proyectos	Cantidad_Profesional	Cantidad_Maestria	Cantidad_Doctorado	T
0	-1.260308	-1.105326	-0.825816	-1.110658	
1	-0.362554	-0.691077	-0.050332	-0.225536	
2	-1.260308	0.137419	-1.213558	-0.225536	
3	0.535199	0.551667	0.337410	0.659586	
4	-0.362554	0.137419	-0.438074	-0.225536	
...	...	...	...	...	
203	-0.362554	-0.276829	1.112894	-1.110658	
204	-1.260308	-0.691077	-1.213558	-0.225536	
205	-1.260308	-0.691077	-1.213558	-1.110658	
206	1.432953	2.208659	0.725152	0.659586	
207	-0.362554	-0.276829	0.725152	-0.225536	

208 rows × 5 columns



```

gestion= df['KMeans_Clusters'].replace({1:2,0: 1, 2:0})
gestion= df['KMeans_Clusters'].replace({2: 'Alta Gestión',1: 'Media Gestión', 0: 'Baja Gestión'})
df['Gestión'] = pd.concat([gestion], axis=1)
df_k_clas = df
df_k_clas

```

	Centro_Año	Cantidad_Proyectos	Cantidad_Profesional	Cantidad_Maestria	Cantidad_
0	9101-2010	-1.260308	-1.105326	-0.825816	
1	9127-2010	-0.362554	-0.691077	-0.050332	
2	9201-2010	-1.260308	0.137419	-1.213558	
3	9202-2010	0.535199	0.551667	0.337410	

```
df_k_clas.to_excel('TPProceso_Cluster_Kmeans.xlsx',index =False)
```

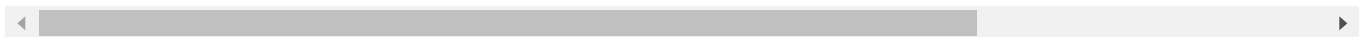
```
...
```

▼ DENDOGRAMA

```
df = df.drop(['Gestión', 'KMeans_Clusters'], axis=1)
df
```

	Centro_Año	Cantidad_Proyectos	Cantidad_Profesional	Cantidad_Maestria	Cantidad_
0	9101-2010	-1.260308	-1.105326	-0.825816	
1	9127-2010	-0.362554	-0.691077	-0.050332	
2	9201-2010	-1.260308	0.137419	-1.213558	
3	9202-2010	0.535199	0.551667	0.337410	
4	9203-2010	-0.362554	0.137419	-0.438074	
...	...	...	...	...	
203	9501-2022	-0.362554	-0.276829	1.112894	
204	9502-2022	-1.260308	-0.691077	-1.213558	
205	9503-2022	-1.260308	-0.691077	-1.213558	
206	9504-2022	1.432953	2.208659	0.725152	
207	9549-2022	-0.362554	-0.276829	0.725152	

208 rows × 6 columns



```
from scipy.cluster.hierarchy import linkage,dendrogram
df_el = df.drop(['Centro_Año'],axis =1)
labels = df['Centro_Año'].to_numpy().tolist()
```



```
x.columns
```

```
Index(['Cantidad_Proyectos', 'Cantidad_Profesional', 'Cantidad_Maestria',  
      'Cantidad_Doctorado', 'Total_Valor_Proyectos'],  
      dtype='object')
```

```
features = ['Cantidad_Proyectos', 'Cantidad_Profesional', 'Cantidad_Maestria',  
           'Cantidad_Doctorado', 'Total_Valor_Proyectos']  
X = x[features]
```

```
plt.figure(figsize=(20,10))  
plt.title('Dendrograma de agrupamiento jerárquico para datos de procesos',fontsize=20)  
plt.xlabel('sample index', fontsize=20)  
plt.ylabel('distance',fontsize=20)  
Z = linkage(X, 'complete','euclidean')  
dn1 = sch.dendrogram(  
    Z,  
    labels=labels,  
    leaf_rotation=90., # rotates the x axis labels  
    leaf_font_size=15., # font size for the x axis labels  
)  
plt.tight_layout()  
plt.show()
```



```
plt.figure(figsize=(20,10))
plt.title('Dendrograma de agrupamiento jerárquico para datos de procesos',fontsize=20)
plt.xlabel('sample index', fontsize=20)
plt.ylabel('distance',fontsize=20)
Z = linkage(X, 'ward','euclidean')
dn2 = sch.dendrogram(
    Z,
    labels=labels,
    leaf_rotation=90., # rotates the x axis labels
    leaf_font_size=15., # font size for the x axis labels
)
plt.tight_layout()
plt.show()
```

```

hac = AgglomerativeClustering(n_clusters=3,
                              affinity="euclidean",
                              linkage="ward")

```

```

hac.fit(x1)

```

```

AgglomerativeClustering(n_clusters=3)

```

```

^|||||

```

```

hac.fit(X)

```

```

categorias = hac.labels_

```

```

categorias

```

```

array([1, 1, 1, 0, 1, 0, 1, 1, 0, 1, 0, 1, 0, 1, 1, 1, 0, 0, 1, 1, 0, 1,
       1, 1, 1, 0, 2, 1, 1, 1, 2, 2, 0, 1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 1,
       1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 1, 2, 0, 2, 1, 1, 0, 0, 1, 1, 1, 1,
       1, 0, 1, 1, 1, 0, 1, 1, 2, 0, 0, 1, 1, 0, 1, 1, 2, 1, 0, 1, 0, 0,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0, 0, 1, 1, 1, 0, 0, 0, 1,
       0, 1, 0, 0, 0, 2, 1, 1, 2, 2, 1, 0, 2, 1, 0, 2, 1, 0, 0, 1, 0, 1,
       1, 0, 1, 0, 1, 2, 0, 1, 2, 0, 1, 0, 1, 0, 0, 1, 0, 0, 1, 2, 0, 0,
       1, 1, 0, 0, 0, 0, 2, 1, 0, 0, 0, 0, 0, 2, 1, 1, 2, 0, 1, 1, 1, 1,
       0, 1, 0, 1, 0, 0, 0, 2, 0, 0, 1, 0, 0, 0, 1, 0, 1, 1, 0, 1, 1, 1,
       1, 1, 1, 2, 1, 1, 1, 1, 2, 1])

```

```

#Agregando la clasificación en el archivo original

```

```

df["Hac_Clusters"]=categorias

```

```

df

```

	Centro_Año	Cantidad_Proyectos	Cantidad_Profesional	Cantidad_Maestria	Cantidad_
0	9101-2010	-1.260308	-1.105326	-0.825816	
1	9127-2010	-0.362554	-0.691077	-0.050332	
2	9201-2010	-1.260308	0.137419	-1.213558	
3	9202-2010	0.535199	0.551667	0.337410	
4	9203-2010	-0.362554	0.137419	-0.438074	
...	...	...	...	...	
203	9501-2022	-0.362554	-0.276829	1.112894	
204	9502-2022	-1.260308	-0.691077	-1.213558	
205	9503-2022	-1.260308	-0.691077	-1.213558	
206	9504-2022	1.432953	2.208659	0.725152	
207	9549-2022	-0.362554	-0.276829	0.725152	

208 rows × 7 columns



▼ Validación

índices de validación externa

```
#Calcular agrupación de índices de validación externa con KMeans
labels_true = df["Centro_Año"]
rotulos_km = hac.labels_
h = sm.homogeneity_score(labels_true, rotulos_km)
print("Homogeneidad: %0.3f" % h)
c = sm.completeness_score(labels_true, rotulos_km)
print("Compleitud: %0.3f" % c)
v = sm.v_measure_score(labels_true, rotulos_km)
print("Medida V: %0.3f" % v )
ari = sm.adjusted_rand_score(labels_true, rotulos_km)
print("Índice Rand corregido: %0.3f" % ari)
ami = sm.adjusted_mutual_info_score(labels_true, rotulos_km, average_method='arithmetic')
print("Adjusted Mutual Information: %0.3f" % ami )
fm = sm.fowlkes_mallows_score(labels_true, rotulos_km)
print("Índice Fowlkes-Mallows: %0.3f" % fm )
```

```
Homogeneidad: 0.174
Compleitud: 1.000
Medida V: 0.296
Índice Rand corregido: 0.000
Adjusted Mutual Information: -0.000
Índice Fowlkes-Mallows: 0.000
```

índices de validación interna

```
from sklearn.metrics import silhouette_score
#silhouette_avg = silhouette_score(x, clustering.labels_)
rotulos_hac = hac.labels_
s = sm.silhouette_score(x, rotulos_hac, metric='euclidean')
print("Coeficiente de Silhueta: %0.3f" % s)
ch = sm.calinski_harabasz_score(x, rotulos_hac)
print("Calinski_harabasz: %0.3f" % ch)
dbs = sm.davies_bouldin_score(x, rotulos_hac)
print("Davies Bouldin: %0.3f" % dbs)
```

```
Coeficiente de Silhueta: 0.317
Calinski_harabasz: 102.104
Davies Bouldin: 1.436
```

```
x.head()
```

	Cantidad_Proyectos	Cantidad_Profesional	Cantidad_Maestria	Cantidad_Doctorado	Tot
0	-1.260308	-1.105326	-0.825816	-1.110658	
1	-0.362554	-0.691077	-0.050332	-0.225536	
2	-1.260308	0.137419	-1.213558	-0.225536	

#Visualización de los Clusters que se crearon

#Aplicaremos Análisis de componetes principales (PCA) para darnos una idea de cómo se crearon

```
pca=PCA(n_components=2)
```

```
pca_productividad=pca.fit_transform(x)
```

```
pca_productividad_2=pd.DataFrame(data=pca_productividad,columns=["Componente_1", "Componente_2"])
```

```
pca_productividad_3=pd.concat([pca_productividad_2,df[["Hac_Clusters"]]],axis=1)
```

```
pca_productividad_3=pd.concat([pca_productividad_3,df[["Centro_Año"]]],axis=1)
```

```
pca_productividad_3
```

	Componente_1	Componente_2	Hac_Clusters	Centro_Año
0	-1.963914	0.259801	1	9101-2010
1	-0.951451	0.156047	1	9127-2010
2	-1.285903	-0.690315	1	9201-2010
3	0.622891	-0.246133	0	9202-2010
4	-1.123852	-0.106162	1	9203-2010
...	...	...	...	...
203	-0.682772	1.580615	1	9501-2022
204	-1.881752	-0.655856	1	9502-2022
205	-1.769395	-0.017971	1	9503-2022
206	3.384948	-0.066004	2	9504-2022
207	-0.703090	0.672548	1	9549-2022

208 rows × 4 columns

```
pca_productividad_3 = pca_productividad_3.dropna()
```

```
pca_productividad_3['Hac_Clusters'] = pca_productividad_3['Hac_Clusters'].astype(int)
```

```
pca_productividad_3
```

	Componente_1	Componente_2	Hac_Clusters	Centro_Año
0	-1.963914	0.259801	1	9101-2010
1	-0.951451	0.156047	1	9127-2010
2	-1.285903	-0.690315	1	9201-2010
3	0.622891	-0.246133	0	9202-2010
4	-1.123852	-0.106162	1	9203-2010
...	...	...	...	...
203	-0.682772	1.580615	1	9501-2022
204	-1.881752	-0.655856	1	9502-2022

```

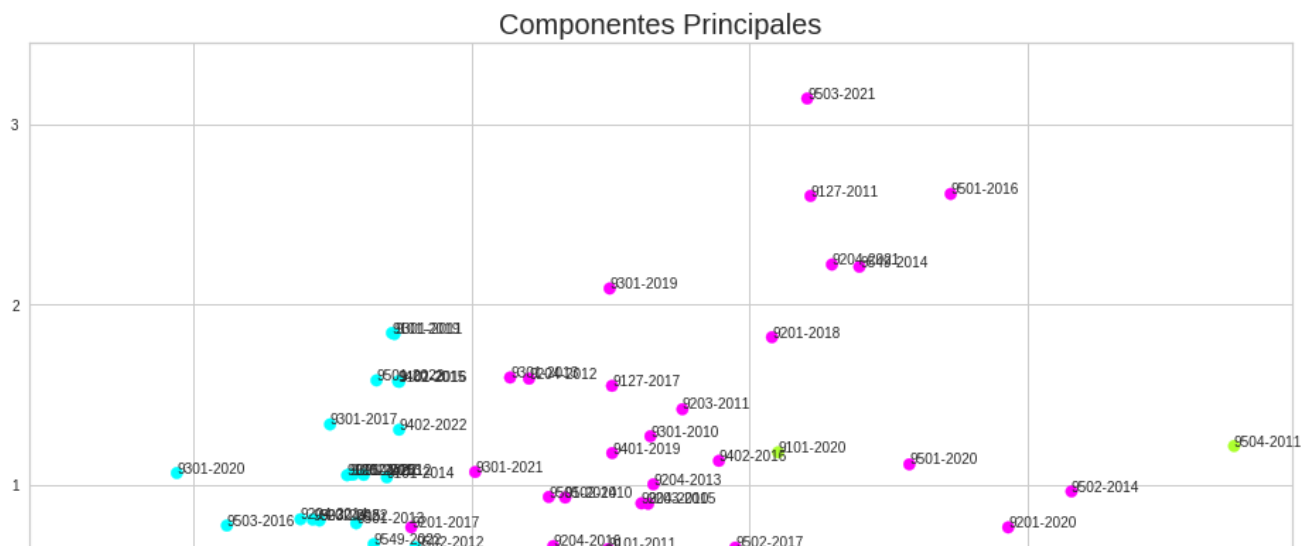
fig=plt.figure(figsize=(8,8))
ax=fig.add_subplot(1,1,1)
annotations = df['Hac_Clusters']
ax.set_xlabel("Componente_1",fontsize=15)
ax.set_ylabel("Componente_2", fontsize=15)
ax.set_title("Componentes Principales", fontsize=20)
color_theme=np.array(["MAGENTA","CYAN",'GREENYELLOW','mediumpurple'])
ax.scatter(x=pca_productividad_3['Componente_1'], y=pca_productividad_3['Componente_2'],c=col
for i, label in enumerate(annotations):
    plt.annotate(label, (pca_productividad_3['Componente_1'][i], pca_productividad_3['Compone
plt.show()

```

Componentes Principales



```
fig=plt.figure(figsize=(16,16))
ax=fig.add_subplot(1,1,1)
annotations = pca_productividad_3['Centro_Año']
ax.set_xlabel("Componente_1",fontsize=15)
ax.set_ylabel("Componente_2", fontsize=15)
ax.set_title("Componentes Principales", fontsize=20)
color_theme=np.array(["MAGENTA","CYAN",'GREENYELLOW','mediumpurple'])
ax.scatter(x=pca_productividad_3['Componente_1'], y=pca_productividad_3['Componente_2'],c=col
for i, label in enumerate(annotations):
    plt.annotate(label, (pca_productividad_3['Componente_1'][i], pca_productividad_3['Compone
plt.show()
```



```
x1= df.drop(['Centro_Año','Hac_Clusters'],axis = 1)
features = ['Cantidad_Proyectos',
            'Cantidad_Profesional', 'Cantidad_Maestria', 'Cantidad_Doctorado',
            'Total_valor_Proyectos']
figure, correlation_matrix = plot_pca_correlation_graph(x1, features,
                                                         dimensions=(1, 2),
                                                         figure_axis_size=10)
```


	Dim 1	Dim 2
Cantidad_Proyectos	-0.965193	0.039709
Cantidad_Profesional	-0.784122	-0.027681
Cantidad_Maestria	-0.685551	0.652113
Cantidad_Doctorado	-0.610983	-0.728949
Total Valor Proyectos	-0.813485	-0.022499

	Centro_Año	Cantidad_Proyectos	Cantidad_Profesional	Cantidad_Maestria	Cantidad_
0	9101-2010	-1.260308	-1.105326	-0.825816	
1	9127-2010	-0.362554	-0.691077	-0.050332	
2	9201-2010	-1.260308	0.137419	-1.213558	
3	9202-2010	0.535199	0.551667	0.337410	
4	9203-2010	-0.362554	0.137419	-0.438074	
...	...	...	...	...	
203	9501-2022	-0.362554	-0.276829	1.112894	
204	9502-2022	-1.260308	-0.691077	-1.213558	

25/50

```
df = df_k_clas
centros_años = df["Centro_Año"].str.split("-", expand=True)
centros_años.columns = ['Codigo_Centro', 'Año']
df = pd.concat([df, centros_años], axis=1)
df.head()
```

	Centro_Año	Cantidad_Proyectos	Cantidad_Profesional	Cantidad_Maestria	Cantidad_Do
0	9101-2010	-1.260308	-1.105326	-0.825816	-1
1	9127-2010	-0.362554	-0.691077	-0.050332	-0
2	9201-2010	-1.260308	0.137419	-1.213558	-0
3	9202-2010	0.535199	0.551667	0.337410	0
4	9203-2010	-0.362554	0.137419	-0.438074	-0

```
df.columns
```

```
Index(['Centro_Año', 'Cantidad_Proyectos', 'Cantidad_Profesional',
      'Cantidad_Maestria', 'Cantidad_Doctorado', 'Total_valor_Proyectos',
      'KMeans_Clusters', 'Gestión', 'Codigo_Centro', 'Año'],
      dtype='object')
```

```
df_map = df[['Codigo_Centro', 'Año', 'KMeans_Clusters', 'Gestión']]
df_map
```

	Codigo_Centro	Año	KMeans_Clusters	Gestión
0	9101	2010	0	Baja Gestión
1	9127	2010	0	Baja Gestión
2	9201	2010	0	Baja Gestión
3	9202	2010	1	Media Gestión

```
df_map['Codigo_Centro'] = df_map['Codigo_Centro'].astype(int)
df_map['Año'] = df_map['Año'].astype(int)
```

/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: <https://pandas.pydata.org/pandas-docs/stable/user>

"""Entry point for launching an IPython kernel.

/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: <https://pandas.pydata.org/pandas-docs/stable/user>

```
df_map = df_map.merge(df_centros, on='Codigo_Centro', how='inner')
df_map.head()
```

	Codigo_Centro	Año	KMeans_Clusters	Gestión	Nombre
0	9101	2010	0	Baja Gestión	Centro_de_los_Recursos_Naturales_Renova
1	9101	2011	2	Alta Gestión	Centro_de_los_Recursos_Naturales_Renova
2	9101	2012	1	Media Gestión	Centro_de_los_Recursos_Naturales_Renova
3	9101	2013	0	Baja Gestión	Centro_de_los_Recursos_Naturales_Renova
4	9101	2014	0	Baja Gestión	Centro_de_los_Recursos_Naturales_Renova

```
df_map = df_map.drop(['Codigo_Regional', 'Codigo_Municipio', 'Codigo_Departamento'], axis= 1)
```

```
df_map
```

	Codigo_Centro	Año	KMeans_Clusters	Gestión	Noml
0	9101	2010	0	Baja Gestión	Centro_de_los_Recursos_Naturales_Reno
1	9101	2011	2	Alta Gestión	Centro_de_los_Recursos_Naturales_Reno
2	9101	2012	1	Media Gestión	Centro_de_los_Recursos_Naturales_Reno
3	9101	2013	0	Baja Gestión	Centro_de_los_Recursos_Naturales_Reno
4	9101	2014	0	Baja Gestión	Centro_de_los_Recursos_Naturales_Reno
...	...	...	...	...	...
203	9549	2018	1	Media Gestión	Complejo_Tecnológico_Turístico_y_Agi
204	9549	2019	1	Media Gestión	Complejo_Tecnológico_Turístico_y_Agi

df_map.dtypes

```

Codigo_Centro    int64
Año              int64
KMeans_Clusters  int32
Gestión          object
Nombre_Centro    object
latitud          float64
longitud         float64
dtype: object

```

▼ Mapa para 2010

```

df_map_2010 = df_map[df_map['Año']<2011]
df_map_2010 = df_map_2010.drop(['Año'],axis = 1)

```

```

latitude = 6.51033052
longitude = -75.48477721

```

```
map_clusters = folium.Map(location=[latitude, longitude], zoom_start=4)
```

```

# set color scheme for the clusters
x = np.arange(2)
ys = [i + x + (i*x)**2 for i in range(4)]

```

```
colors_array = cm.rainbow(np.linspace(0, 1, len(ys)))
rainbow = [colors.rgb2hex(i) for i in colors_array]

# add markers to the map
markers_colors = []
for lat, lng, cluster, cent, gest in zip(df_map_2010['latitud'],
                                         df_map_2010['longitud'],
                                         df_map_2010['KMeans_Clusters'],
                                         df_map_2010['Nombre_Centro'],
                                         df_map_2010['Gestión']) :

    label = folium.Popup(str(cent) + '- Cluster ' + str(cluster), parse_html=True)
    folium.vector_layers.CircleMarker(
        [lat, lng],
        radius=6,
        #popup=label,
        tooltip = str(cent) + ' , ' + str(gest),
        color=rainbow[cluster-1],
        fill=True,
        fill_color=rainbow[cluster-1],
        fill_opacity=0.9).add_to(map_clusters)

map_clusters
```



▼ Mapa para 2014

```
df_map_2014 = df_map[df_map['Año']<2015]
df_map_2014 = df_map_2014.drop(['Año'],axis = 1)
```

Leaflet (<http://leafletjs.com/>)

```
latitude = 6.51033052
longitude = -75.48477721
```

```
map_clusters = folium.Map(location=[latitude, longitude], zoom_start=4)
```

```
# set color scheme for the clusters
x = np.arange(2)
ys = [i + x + (i*x)**2 for i in range(4)]
colors_array = cm.rainbow(np.linspace(0, 1, len(ys)))
rainbow = [colors.rgb2hex(i) for i in colors_array]
```

```
# add markers to the map
markers_colors = []
for lat, lng, cluster, cent, gest in zip(df_map_2014['latitud'],
                                         df_map_2014['longitud'],
                                         df_map_2014['KMeans_Clusters'],
                                         df_map_2014['Nombre_Centro'],
                                         df_map_2014['Gestión']) :
```

```
label = folium.Popup(str(cent) + '- Cluster ' + str(cluster), parse_html=True)
folium.vector_layers.CircleMarker(
    [lat, lng],
    radius=6,
    #popup=label,
    tooltip = str(cent) + ' , ' + str(gest),
    color=rainbow[cluster-1],
    fill=True,
    fill_color=rainbow[cluster-1],
```

```
fill_opacity=0.9).add_to(map_clusters)
```

```
map_clusters
```



▼ Mapa para 2018

```
df_map_2018 = df_map[df_map['Año']<2019]
```

```
df_map_2018 = df_map_2018.drop(['Año'],axis = 1)

latitude = 6.51033052
longitude = -75.48477721

map_clusters = folium.Map(location=[latitude, longitude], zoom_start=4)

# set color scheme for the clusters
x = np.arange(2)
ys = [i + x + (i*x)**2 for i in range(4)]
colors_array = cm.rainbow(np.linspace(0, 1, len(ys)))
rainbow = [colors.rgb2hex(i) for i in colors_array]

# add markers to the map
markers_colors = []
for lat, lng, cluster, cent, gest in zip(df_map_2018['latitud'],
                                         df_map_2018['longitud'],
                                         df_map_2018['KMeans_Clusters'],
                                         df_map_2018['Nombre_Centro'],
                                         df_map_2018['Gestión']) :

    label = folium.Popup(str(cent) + ' - Cluster ' + str(cluster), parse_html=True)
    folium.vector_layers.CircleMarker(
        [lat, lng],
        radius=6,
        #popup=label,
        tooltip = str(cent) + ' , ' + str(gest),
        color=rainbow[cluster-1],
        fill=True,
        fill_color=rainbow[cluster-1],
        fill_opacity=0.9).add_to(map_clusters)

map_clusters
```




▼ Mapa para 2022

```
df_map_2022 = df_map[df_map['Año']<2023]
df_map_2022 = df_map_2022.drop(['Año'],axis = 1)

latitude = 6.51033052
longitude = -75.48477721

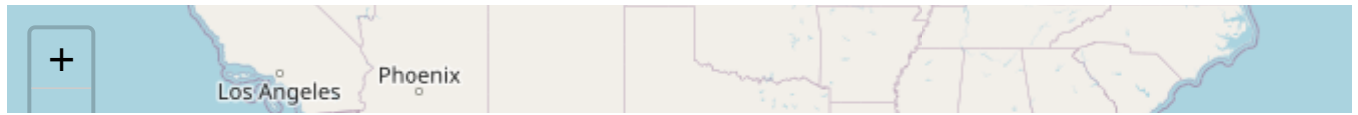
map_clusters = folium.Map(location=[latitude, longitude], zoom_start=4)

# set color scheme for the clusters
x = np.arange(2)
ys = [i + x + (i*x)**2 for i in range(4)]
colors_array = cm.rainbow(np.linspace(0, 1, len(ys)))
rainbow = [colors.rgb2hex(i) for i in colors_array]

# add markers to the map
markers_colors = []
for lat, lng, cluster, cent, gest in zip(df_map_2022['latitud'],
                                         df_map_2022['longitud'],
                                         df_map_2022['KMeans_Clusters'],
                                         df_map_2022['Nombre_Centro'],
                                         df_map_2022['Gestión']):
```

```
label = folium.Popup(str(cent) + '- Cluster ' + str(cluster), parse_html=True)
folium.vector_layers.CircleMarker(
    [lat, lng],
    radius=6,
    #popup=label,
    tooltip = str(cent) + ' , ' + str(gest),
    color=rainbow[cluster-1],
    fill=True,
    fill_color=rainbow[cluster-1],
    fill_opacity=0.9).add_to(map_clusters)
```

```
map_clusters
```



```
#from folium import plugins

#map_hooray = folium.Map(location=[51.5074, 0.1278],
#                          #                          zoom_start = 13)

# Ensure you're handing it floats
#df_acc['Latitude'] = df_acc['Latitude'].astype(float)
#df_acc['Longitude'] = df_acc['Longitude'].astype(float)

# Filter the DF for rows, then columns, then remove NaNs
#heat_df = df_acc[df_acc['Speed_limit']=='40'] # Reducing data size so it runs faster
#heat_df = df_acc[df_acc['Year']=='2007'] # Reducing data size so it runs faster
#heat_df = heat_df[['Latitude', 'Longitude']]

# Create weight column, using date
#heat_df['Weight'] = df_acc['Date'].str[3:5]
#heat_df['Weight'] = heat_df['Weight'].astype(float)
#heat_df = heat_df.dropna(axis=0, subset=['Latitude','Longitude', 'Weight'])

# List comprehension to make out list of lists
#heat_data = [[[row['Latitude'],row['Longitude']] for index, row in heat_df[heat_df['Weight']]

# Plot it on the map
#hm = plugins.HeatMapWithTime(heat_data,auto_play=True,max_opacity=0.8)
#hm.add_to(map_hooray)
# Display the map
#map_hooray

#import folium
#import folium.plugins as plugins
#import pandas as pd
#import json
#import requests
#import time
#from folium.plugins import HeatMapWithTime
#import datetime

#cumulative = pd.DataFrame()

#mins = input('How many mins do you want to see? ')
#number_of_times = (int(mins)*60)/30

#startTime = time.time()
#for i in range(int(number_of_times)):
#    url = ("https://api.data.gov.sg/v1/transport/taxi-availability")
```

```

# response = requests.get(url)
# data = response.json()
#df = pd.io.json.json_normalize(data['features'])
#coordinateslist = df['geometry.coordinates'].tolist()
#df1 = pd.DataFrame(coordinateslist)
#result = df1.transpose()
#result.columns = ['coordinates']
#result['Timestamp'] = (df['properties.timestamp'][0])
#cumulative = cumulative.append(result)
#time.sleep(30)
#endTime = time.time()
#elapsedTime = endTime - startTime
#print("Elapsed Time = %s" % elapsedTime, 'seconds')

#cumulative['coordinates'] = cumulative['coordinates'].astype(str)
#Latitude = []
#Longitude = []
#coordinates = []
#for i in cumulative['coordinates']:
#    i = i.split(", ")
#    lat = i[1][:1]
#    long = i[0][1:]
#    Latitude.append(lat)
#    Longitude.append(long)
#cumulative['Latitude'] = Latitude
#cumulative['Longitude'] = Longitude

#lat_long_list = []
#for i in cumulative['Timestamp'].unique():
#    temp=[]
#    for index, instance in cumulative[cumulative['Timestamp'] == i].iterrows():
#        temp.append([instance['Latitude'],instance['Longitude']])
#    lat_long_list.append(temp)

#converting it to datetime format
#cumulative['Timestamp']= pd.to_datetime(cumulative['Timestamp'])
#creating a time index
#time_index = []
#for i in cumulative['Timestamp'].unique():
#    time_index.append(i)
#formatting the index
#date_strings = [d.strftime('%d/%m/%Y, %H:%M:%S') for d in time_index]

#Choosing the map type
#m = folium.Map(location=[1.352083,103.819839],zoom_start = 11, tiles="https://tiles.stadiama
#Plot it on the map
#HeatMapWithTime(lat_long_list,radius=5,auto_play=True,position='bottomright',name="cluster",

```

```
# Display the map
#m
```

▼ RED NEURONAL: PERCEPTRON

```
df.columns
```

```
Index(['Centro_Año', 'Cantidad_Proyectos', 'Cantidad_Profesional',
      'Cantidad_Maestria', 'Cantidad_Doctorado', 'Total_valor_Proyectos',
      'KMeans_Clusters', 'Gestión', 'Codigo_Centro', 'Año'],
      dtype='object')
```

```
df_proc_orig = pd.read_excel(path1)
df_proc_orig
df_prc = df[['Codigo_Centro', 'Año', 'Gestión']]
```

```
df_per = pd.concat([df_prc, df_proc_orig], axis=1).drop(['Centro_Año'], axis=1)
df_per
```

	Codigo_Centro	Año	Gestión	Q_Proyectos	QNA_5007	QNA_5009	QNA_5010	TV_Proyec
0	9101	2010	Baja Gestión	1	1	2	0	4.652924e-
1	9127	2010	Baja Gestión	2	2	4	1	3.920052e-
2	9201	2010	Baja Gestión	1	4	1	1	4.556627e-
3	9202	2010	Media Gestión	3	5	5	2	4.071852e-
4	9203	2010	Baja Gestión	2	4	3	1	2.906175e-
...	...	...	...	...	...	...	...	...
203	9501	2022	Baja Gestión	2	3	7	0	3.754077e-
204	9502	2022	Baja Gestión	1	2	1	1	3.977569e-

```
proceso = df_per['Gestión'].replace({'Baja Gestión': 1, 'Media Gestión': 2, 'Alta Gestión': 3})
df_per['Gestión'] = pd.concat([proceso], axis=1)
df_per
```

	Codigo_Centro	Año	Gestión	Q_Proyectos	QNA_5007	QNA_5009	QNA_5010	TV_Proyec
0	9101	2010	1	1	1	2	0	4.652924e-
1	9127	2010	1	2	2	4	1	3.920052e-
2	9201	2010	1	1	4	1	1	4.556627e-
3	9202	2010	2	3	5	5	2	4.071852e-
4	9203	2010	1	2	4	3	1	2.906175e-
...	...	...	...	...	...	...	...	
203	9501	2022	1	2	3	7	0	3.754077e-
204	9502	2022	1	1	2	1	1	3.977569e-
205	9503	2022	1	1	2	1	0	5.068536e-
206	9504	2022	3	4	9	6	2	7.614273e-
207	9549	2022	1	2	3	6	1	3.294535e-

```
#Clases
```

```
y = (df_per['Gestión']).astype(np.int8)
```

```
print(y)
```

```
0      1
1      1
2      1
3      2
4      1
..
203    1
204    1
205    1
206    3
207    1
```

```
Name: Gestión, Length: 208, dtype: int8
```

Dividimos los datos en un conjunto de aprendizaje y de prueba:

```
from sklearn.model_selection import train_test_split
datasets = train_test_split(df_per,
                             y,
                             test_size=0.3)
```

```
X_train, X_test, y_train, y_test = datasets
```

Ahora, creamos una instancia de Perceptron y ajustamos los datos de entrenamiento:

```
from sklearn.linear_model import Perceptron
```

```
p = Perceptron(random_state=42,
               max_iter=100,
               tol=0.001)
p.fit(X_train, y_train)

Perceptron(max_iter=100, random_state=42)
```

Recordando un poco sobre la matriz de confusión

		Actual Values	
		Yes	No
Predicted Values	Yes	True Positive	False Positive
	No	False Negative	True Negative

```
from sklearn.metrics import classification_report
y_perc_pred = p.predict(X_test)
print(classification_report(y_test, y_perc_pred))
```

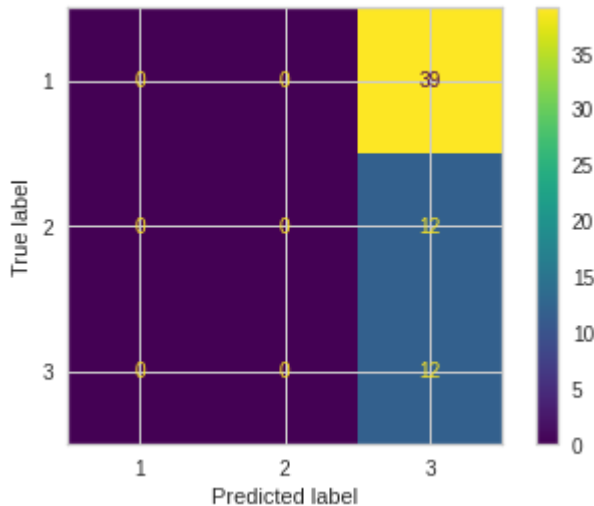
	precision	recall	f1-score	support
1	0.00	0.00	0.00	39
2	0.00	0.00	0.00	12
3	0.19	1.00	0.32	12
accuracy			0.19	63
macro avg	0.06	0.33	0.11	63
weighted avg	0.04	0.19	0.06	63

```
/usr/local/lib/python3.7/dist-packages/sklearn/metrics/_classification.py:1318: UndefinedWarning:
  _warn_prf(average, modifier, msg_start, len(result))
/usr/local/lib/python3.7/dist-packages/sklearn/metrics/_classification.py:1318: UndefinedWarning:
  _warn_prf(average, modifier, msg_start, len(result))
```

```
/usr/local/lib/python3.7/dist-packages/sklearn/metrics/_classification.py:1318: UndefinedWarning:
warn_prf(average, modifier, msg_start, len(result))
```

```
plot_confusion_matrix(p, X_test, y_test)
plt.show()
```

```
/usr/local/lib/python3.7/dist-packages/sklearn/utils/deprecation.py:87: FutureWarning:
warnings.warn(msg, category=FutureWarning)
```



```
path3 = '/content/drive/MyDrive/Proyecto Sennova borrador/sample_procesos.xlsx'
df_sample = pd.read_excel(path3)
df_sample
```

	Codigo_Centro	Año	Gestión	Q_Proyectos	QNA_5007	QNA_5009	QNA_5010	TV_Proyect
0	9101.0	2010.0	1.0	1.0	1.0	2.0	0.0	4.652924e+
1	9127.0	2010.0	1.0	2.0	2.0	4.0	1.0	3.920052e+
2	9201.0	2010.0	1.0	1.0	3.0	3.0	0.0	4.556627e+
3	9202.0	2010.0	1.0	3.0	8.0	5.0	0.0	4.071852e+
4	9203.0	2010.0	1.0	2.0	6.0	4.0	1.0	2.906175e+
5	9501.0	2022.0	1.0	2.0	4.0	4.0	2.0	3.754077e+
6	9502.0	2022.0	1.0	1.0	3.0	2.0	0.0	3.977569e+
7	9503.0	2022.0	1.0	1.0	0.0	0.0	1.0	5.068536e+
8	9504.0	2022.0	1.0	4.0	7.0	7.0	1.0	7.614273e+
9	9549.0	2022.0	1.0	2.0	5.0	2.0	1.0	3.294535e+

```
y_percep_perd = p.predict(df_sample)
y_percep_perd_s= pd.Series(y_percep_perd)
```



```
df_percep = df_sample[['Codigo_Centro', 'Año', 'Gestión']]
df_percep['Gestión'] = y_percep_perd_s.replace({1: 'Baja Gestión', 2: 'Media Gestión', 3: 'Alta Gestión'})
df_percep
```

/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:4: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/10min.html
after removing the cwd from sys.path.

	Codigo_Centro	Año	Gestión
0	9101.0	2010.0	Alta Gestión
1	9127.0	2010.0	Alta Gestión
2	9201.0	2010.0	Alta Gestión
3	9202.0	2010.0	Alta Gestión
4	9203.0	2010.0	Alta Gestión
5	9501.0	2022.0	Alta Gestión
6	9502.0	2022.0	Alta Gestión
7	9503.0	2022.0	Alta Gestión
8	9504.0	2022.0	Alta Gestión
9	9549.0	2022.0	Alta Gestión

▼ Regresión Logística

```
from sklearn.linear_model import LogisticRegression
```

```
from sklearn.model_selection import train_test_split
datasets = train_test_split(df_per,
                             y,
                             test_size=0.3)
```

```
X_train, X_test, y_train, y_test = datasets
```

```
# Instanciando a reg. logistica
reglog = LogisticRegression()
```

```
# Aplicando el modelo
reglog.fit(X_train, y_train)
```

```
LogisticRegression()
```

```
LogisticRegression()
```

```
LogisticRegression()
```

```
# Predicción
```

```
y_reglog_pred = reglog.predict(X_test)
```

```
print(classification_report(y_test,y_reglog_pred))
```

	precision	recall	f1-score	support
1	0.57	1.00	0.73	36
2	0.00	0.00	0.00	11
3	0.00	0.00	0.00	16
accuracy			0.57	63
macro avg	0.19	0.33	0.24	63
weighted avg	0.33	0.57	0.42	63

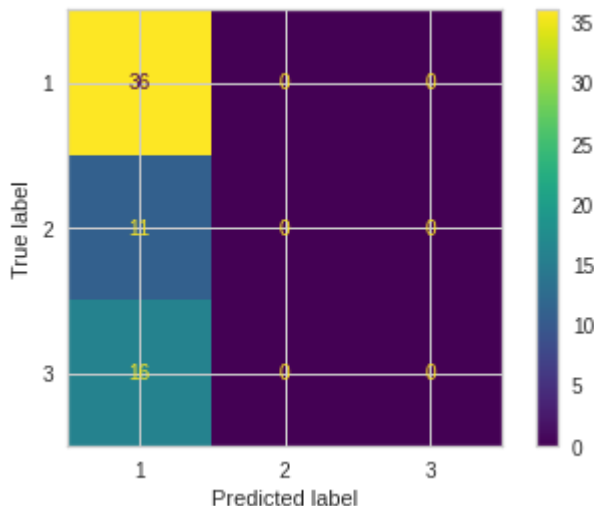
```
/usr/local/lib/python3.7/dist-packages/sklearn/metrics/_classification.py:1318: UndefinedWarning:
  _warn_prf(average, modifier, msg_start, len(result))
/usr/local/lib/python3.7/dist-packages/sklearn/metrics/_classification.py:1318: UndefinedWarning:
  _warn_prf(average, modifier, msg_start, len(result))
/usr/local/lib/python3.7/dist-packages/sklearn/metrics/_classification.py:1318: UndefinedWarning:
  _warn_prf(average, modifier, msg_start, len(result))
```

```
# Matriz de confusión
```

```
plot_confusion_matrix(reglog, X_test, y_test)
```

```
plt.show()
```

```
/usr/local/lib/python3.7/dist-packages/sklearn/utils/deprecation.py:87: FutureWarning:
  warnings.warn(msg, category=FutureWarning)
```



Test con valores de muestra

```
df_sample.drop(['Gestión'], axis = 1)
```

	Codigo_Centro	Año	Q_Proyectos	QNA_5007	QNA_5009	QNA_5010	TV_Proyectos
0	9101.0	2010.0	1.0	1.0	2.0	0.0	4.652924e+14
1	9127.0	2010.0	2.0	2.0	4.0	1.0	3.920052e+14
2	9201.0	2010.0	1.0	3.0	3.0	0.0	4.556627e+14
3	9202.0	2010.0	3.0	8.0	5.0	0.0	4.071852e+14
4	9203.0	2010.0	2.0	6.0	4.0	1.0	2.906175e+14
5	9501.0	2022.0	2.0	4.0	4.0	2.0	3.754077e+14
6	9502.0	2022.0	1.0	3.0	2.0	0.0	3.977569e+14
7	9503.0	2022.0	1.0	0.0	0.0	1.0	5.068536e+14
8	9504.0	2022.0	4.0	7.0	7.0	1.0	7.614273e+14
9	9549.0	2022.0	2.0	5.0	2.0	1.0	3.294535e+14

```
y_reglog_pred = reglog.predict(df_sample)
```

```
y_reglog_pred = pd.Series(y_reglog_pred)
```

```
df_logist = df_sample[['Codigo_Centro', 'Año', 'Gestión']]
```

```
df_logist['Productividad'] = y_reglog_pred.replace({1: 'Baja Gestión', 2: 'Media Gestión', 3:
```

```
df_logist
```

/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:5: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: <https://pandas.pydata.org/pandas-docs/stable/user>
"""

▼ Random Forest

```
1          9127 0  2010 0          1 0      Raia Gestión
from sklearn.ensemble import RandomForestClassifier
```

```
rf = RandomForestClassifier()
```

```
rf.fit(X_train, y_train)
```

```
y_dt_pred = rf.predict(X_test)
print(classification_report(y_test,y_dt_pred))
```

	precision	recall	f1-score	support
1	1.00	1.00	1.00	36
2	1.00	1.00	1.00	11
3	1.00	1.00	1.00	16
accuracy			1.00	63
macro avg	1.00	1.00	1.00	63
weighted avg	1.00	1.00	1.00	63

```
# Matriz de confusión
plot_confusion_matrix(rf, X_test, y_test)
plt.show()
```

/usr/local/lib/python3.7/dist-packages/sklearn/utils/deprecation.py:87: FutureWarning: F

Test con valores de muestra



df_sample

	Codigo_Centro	Año	Gestión	Q_Proyectos	QNA_5007	QNA_5009	QNA_5010	TV_Proyect
0	9101.0	2010.0	1.0	1.0	1.0	2.0	0.0	4.652924e+
1	9127.0	2010.0	1.0	2.0	2.0	4.0	1.0	3.920052e+
2	9201.0	2010.0	1.0	1.0	3.0	3.0	0.0	4.556627e+
3	9202.0	2010.0	1.0	3.0	8.0	5.0	0.0	4.071852e+
4	9203.0	2010.0	1.0	2.0	6.0	4.0	1.0	2.906175e+
5	9501.0	2022.0	1.0	2.0	4.0	4.0	2.0	3.754077e+
6	9502.0	2022.0	1.0	1.0	3.0	2.0	0.0	3.977569e+
7	9503.0	2022.0	1.0	1.0	0.0	0.0	1.0	5.068536e+
8	9504.0	2022.0	1.0	4.0	7.0	7.0	1.0	7.614273e+
9	9549.0	2022.0	1.0	2.0	5.0	2.0	1.0	3.294535e+

```
y_dt_pred_ram = rf.predict(df_sample)
y_dt_pred_ram = pd.Series(y_dt_pred_ram)

df_rd_forest = df_sample[['Codigo_Centro','Año','Gestión']]
df_rd_forest['Productividad'] = y_dt_pred_ram.replace({1: 'Baja Gestión',2: 'Media Gestión',
df_rd_forest
```

/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:5: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.

Try using `.loc[row_indexer,col_indexer] = value` instead

See the caveats in the documentation: <https://pandas.pydata.org/pandas-docs/stable/user>

"""

Codigo_Centro Año Gestión Productividad

▼ SVM

2 9204 0 2010 0 1 0 Baja Gestión

#Escalaamos el valor de costo total de proyectos

from sklearn.svm import SVC

scaler = StandardScaler()

X_train[['TV_Proyectos']] = scaler.fit_transform(X_train[['TV_Proyectos']])

X_train

	Codigo_Centro	Año	Gestión	Q_Proyectos	QNA_5007	QNA_5009	QNA_5010	TV_Proyec
37	9204	2012	3	3	2	7	0	0.833
131	9202	2018	1	1	3	2	1	-0.275
95	9549	2015	1	2	2	3	1	0.282
35	9202	2012	1	2	3	3	1	-0.945
55	9206	2013	2	5	14	2	5	0.512
...	...	...	...	...	...	...	...	
18	9201	2011	1	2	1	1	2	-0.257
65	9127	2014	1	2	3	2	2	-0.395
97	9127	2016	3	4	8	6	2	0.339
88	9301	2015	1	1	1	2	0	-1.410
205	9503	2022	1	1	2	1	0	0.219

145 rows × 8 columns

y_svm = SVC(C = 100, kernel = 'linear', random_state=123)

y_svm.fit(X_train, y_train)

SVC(C=100, kernel='linear', random_state=123)

y_svm_pred = y_svm.predict(X_test)

print(classification_report(y_test,y_svm_pred))

precision recall f1-score support

1	0.00	0.00	0.00	36
2	0.00	0.00	0.00	11
3	0.25	1.00	0.41	16
accuracy			0.25	63
macro avg	0.08	0.33	0.14	63
weighted avg	0.06	0.25	0.10	63

```

/usr/local/lib/python3.7/dist-packages/sklearn/metrics/_classification.py:1318: Undefined
_warn_prf(average, modifier, msg_start, len(result))
/usr/local/lib/python3.7/dist-packages/sklearn/metrics/_classification.py:1318: Undefined
_warn_prf(average, modifier, msg_start, len(result))
/usr/local/lib/python3.7/dist-packages/sklearn/metrics/_classification.py:1318: Undefined
_warn_prf(average, modifier, msg_start, len(result))

```

Test con valores de muestra

Matriz de confusión

```

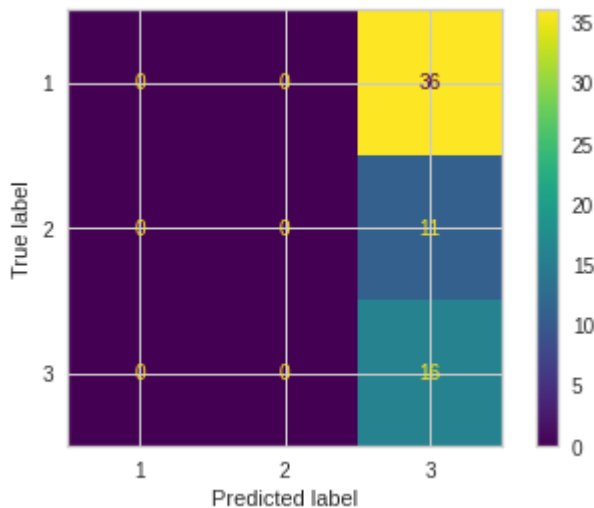
plot_confusion_matrix(y_svm, X_test, y_test, include_values =True)
plt.show()

```

```

/usr/local/lib/python3.7/dist-packages/sklearn/utils/deprecation.py:87: FutureWarning: F
warnings.warn(msg, category=FutureWarning)

```



```

y_dt_pred_svm = pd.Series(y_svm.predict(df_sample))
df_SVM = df_sample[['Codigo_Centro', 'Año', 'Gestión']]
df_SVM['Productividad'] = y_dt_pred_svm.replace({1: 'Baja Gestión', 2: 'Media Gestión', 3: 'Alta Gestión'})
df_SVM

```

/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:3: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.

Try using `.loc[row_indexer,col_indexer] = value` instead

See the caveats in the documentation: <https://pandas.pydata.org/pandas-docs/stable/user>

This is separate from the ipykernel package so we can avoid doing imports until

	Codigo_Centro	Año	Gestión	Productividad
0	9101.0	2010.0	1.0	Alta Gestión
1	9127.0	2010.0	1.0	Alta Gestión
2	9201.0	2010.0	1.0	Alta Gestión
3	9202.0	2010.0	1.0	Alta Gestión
4	9203.0	2010.0	1.0	Alta Gestión
5	9501.0	2022.0	1.0	Alta Gestión
6	9502.0	2022.0	1.0	Alta Gestión

➤ Reporte de parametros de validación de modelos

9 9549.0 2022.0 1.0 Alta Gestión

```
from sklearn.metrics import accuracy_score, confusion_matrix, f1_score, precision_score, recall_score
```

```
p.predict(X_test)
```

```
reglog.predict(X_test)
```

```
rf.predict(X_test)
```

```
y_svm.predict(X_test)
```

```
y_test
```

Reporte de parámetros de validación

```
# perceptrón
```

```
acc_perc = accuracy_score(y_test, p.predict(X_test))
```

```
f1_perc = f1_score(y_test, p.predict(X_test), average='weighted')
```

```
#auc_perc = roc_auc_score(y_test, p.predict(X_test))
```

```
#recall_perc = recall_score(y_test, p.predict(X_test))
```

```
# Regresión logística
```

```
acc_log = accuracy_score(y_test, reglog.predict(X_test))
```

```
f1_log = f1_score(y_test, reglog.predict(X_test), average='weighted')
```

```
#Arboles de decisión
```

```
acc_rf = accuracy_score(y_test, rf.predict(X_test))
```



```

f1_rf = f1_score(y_test, rf.predict(X_test), average='weighted')

#Máquinas de soporte vectorial
acc_svm = accuracy_score(y_test, y_svm.predict(X_test))
f1_svm = f1_score(y_test, y_svm.predict(X_test), average='weighted')

Parametro = ['Accuracy', 'F1-Score']
Percep= [acc_perc,f1_perc]
r_log = [acc_log,f1_log]
rfor = [acc_rf,f1_rf]
svm_t = [acc_svm,f1_svm]

df_valid = pd.DataFrame()
df_valid['Parámetro'] = Parametro
df_valid['Perceptrón'] = Percep
df_valid['Regresión logística'] = r_log
df_valid['Random Forest'] = rfor
df_valid['Máquina de soporte vectorial'] = svm_t

```

```
df_valid
```

	Parámetro	Perceptrón	Regresión logística	Random Forest	Máquina de soporte vectorial
0	Accuracy	0.253968	0.571429	1.0	0.253968
1	F1-Score	0.102873	0.415584	1.0	0.102873

```
df_valid.to_excel('Validacion_Procesos.xlsx',index =False)
```

[Productos de pago de Colab](#) - [Cancelar contratos](#)





ANEXO 10. Reporte de códigos para compatibilidad de API con resultado analítico.

```

#librerías
import numpy as np
import pandas as pd

import folium
from folium import plugins

from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler

import matplotlib.pyplot as plt
import matplotlib.cm as cm
import matplotlib.colors as colors

#Cargar archivos desde drive
from google.colab import drive
drive.mount('/content/drive')

    Mounted at /content/drive

#Path en Drive
from google.colab import files
path = "/content/drive/MyDrive/Proyecto SGPS 9748/Datos.xlsx"
path2 = '/content/drive/MyDrive/Proyecto SGPS 9748/PCPA.xlsx'

#Dataframes
principal_proyectos= pd.read_excel(path,sheet_name= 'Principal_Proyectos')
principal_talento_humano = pd.read_excel(path,sheet_name= 'Principal_Talento_Humano')
principal_productos = pd.read_excel(path,sheet_name= 'Principal_Productos')
df = pd.read_excel(path2)
df_centros = pd.read_excel(path,sheet_name= 'Centros_Formacion')

#Escalar datos: desviación estandar
scaler = StandardScaler()
df[['P_8020', 'P_8025', 'P_8019', 'P_8039', 'P_8016','P_8031']] = scaler.fit_transform(df[['

# variables
y = df[['Centro_Año']]
x = df.drop(['Centro_Año'], axis= 1)
#K-means a la base de datos
clustering=KMeans(n_clusters=3, max_iter=300,random_state=0) #Crea el modelo
clustering.fit(x) #Aplica el modelo a la base de datos

    KMeans(n_clusters=3, random_state=0)

#creando columna de clusters
df["KMeans_Clusters"]=clustering.labels_

```

df

	Centro_Año	P_8020	P_8025	P_8019	P_8039	P_8016	P_8031	KMeans_Cl
0	9101-2010	-0.560125	0.857305	-0.505342	2.814094	-1.143520	0.780675	
1	9127-2010	-0.560125	-1.274003	-1.170602	0.168528	0.930772	0.068480	
2	9201-2010	-1.189888	-1.274003	-1.170602	-1.154255	-1.143520	0.068480	
3	9202-2010	-1.189888	-0.563567	-1.170602	0.168528	-1.143520	-0.643715	
4	9203-2010	2.588687	0.146869	0.159918	1.491311	0.239341	0.780675	
...	...	...	...	...	...	...	...	
203	9501-2022	0.699400	2.988612	2.820960	-0.492864	-0.452089	-0.643715	
204	9502-2022	0.069637	0.857305	0.159918	-0.492864	-0.452089	-0.643715	
205	9503-2022	0.069637	-1.274003	0.159918	-0.492864	0.239341	-0.643715	
206	9504-2022	1.329162	0.146869	-1.170602	-0.492864	-0.452089	-0.643715	
207	9549-2022	-0.560125	-1.274003	0.825179	0.829919	0.239341	1.492871	

208 rows × 8 columns



```
# la asignación de etiqueta de cluster cambia al ser compilado, por lo que no debe ser consid
# para ser corregido, debe realizarse a partir del PCA y el mapa de correlaciones.
# Es posible corregirse, posiblemente con algun condicional en los signos y magnitudes del ve
produc= df['KMeans_Clusters'].replace({0:0,2:1, 1:2})
produc= df['KMeans_Clusters'].replace({0: 'Baja Productividad', 1: 'Media Productividad',2: '
df['Productividad'] = pd.concat([produc], axis=1)
centros_años = df["Centro_Año"].str.split("-",expand=True)
centros_años.columns = ['Codigo_Centro', 'Año']
df = pd.concat([df, centros_años], axis=1)
```

```
df_map = df[['Codigo_Centro', 'Año', 'KMeans_Clusters', 'Productividad']]
df_map['Codigo_Centro'] = df_map['Codigo_Centro'].astype(int)
df_map['Año'] = df_map['Año'].astype(int)
df_map = df_map.merge(df_centros, on='Codigo_Centro', how='inner')
df_map = df_map.drop(['Codigo_Regional', 'Codigo_Municipio', 'Codigo_Departamento'], axis= 1)
```

```
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead
```

See the caveats in the documentation: <https://pandas.pydata.org/pandas-docs/stable/user>

```
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:3: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead
```

See the caveats in the documentation: <https://pandas.pydata.org/pandas-docs/stable/user>
 This is separate from the ipykernel package so we can avoid doing imports until

```
# Separación de centros por años
df_map_2010 = df_map[(df_map['Año']<2011)&(df_map['Año']>=2010)]
df_map_2010 = df_map_2010.drop(['Año'],axis = 1)

latitude = 6.51033052
longitude = -75.48477721

map_clusters = folium.Map(location=[latitude, longitude], zoom_start=4)

# set color scheme for the clusters
x = np.arange(2)
ys = [i + x + (i*x)**2 for i in range(4)]
colors_array = cm.rainbow(np.linspace(0, 1, len(ys)))
rainbow = [colors.rgb2hex(i) for i in colors_array]

# add markers to the map
markers_colors = []
for lat, lng, cluster, cent, gest in zip(df_map_2010['latitud'],
                                         df_map_2010['longitud'],
                                         df_map_2010['KMeans_Clusters'],
                                         df_map_2010['Nombre_Centro'],
                                         df_map_2010['Productividad']) :

    label = folium.Popup(str(cent) + '- Cluster ' + str(cluster), parse_html=True)
    folium.vector_layers.CircleMarker(
        [lat, lng],
        radius=6,
        #popup=label,
        tooltip = str(cent) + ' , ' + str(gest),
        color=rainbow[cluster-1],
        fill=True,
        fill_color=rainbow[cluster-1],
        fill_opacity=0.9).add_to(map_clusters)

map_clusters
```

Make this Notebook Trusted to load map: File -> Trust Notebook



```
# Separación de centros por años
df_map_2014 = df_map[(df_map['Año']<2015)&(df_map['Año']>=2014)]
df_map_2014 = df_map_2014.drop(['Año'],axis = 1)

latitude = 6.51033052
longitude = -75.48477721

map_clusters = folium.Map(location=[latitude, longitude], zoom_start=4)

# set color scheme for the clusters
x = np.arange(2)
ys = [i + x + (i*x)**2 for i in range(4)]
colors_array = cm.rainbow(np.linspace(0, 1, len(ys)))
rainbow = [colors.rgb2hex(i) for i in colors_array]

# add markers to the map
markers_colors = []
for lat, lng, cluster, cent, gest in zip(df_map_2014['latitud'],
                                         df_map_2014['longitud'],
                                         df_map_2014['KMeans_Clusters'],
                                         df_map_2014['Nombre_Centro'],
                                         df_map_2014['Productividad']) :

    label = folium.Popup(str(cent) + ' - Cluster ' + str(cluster), parse_html=True)
    folium.vector_layers.CircleMarker(
        [lat, lng],
```

```
radius=6,  
#popup=label,  
tooltip = str(cent) + ' , ' + str(gest),  
color=rainbow[cluster-1],  
fill=True,  
fill_color=rainbow[cluster-1],  
fill_opacity=0.9).add_to(map_clusters)
```

map_clusters

Make this Notebook Trusted to load map: File -> Trust Notebook



Leaflet (<http://leafletjs.com>)

```
# Separación de centros por años
```

```
df_map_2018 = df_map[(df_map['Año']<2019)&(df_map['Año']>=2018)]
```

```
df_map_2018 = df_map_2018.drop(['Año'],axis = 1)
```

```
latitude = 6.51033052
```

```
longitude = -75.48477721
```

```
map_clusters = folium.Map(location=[latitude, longitude], zoom_start=4)
```



```
# set color scheme for the clusters
x = np.arange(2)
ys = [i + x + (i*x)**2 for i in range(4)]
colors_array = cm.rainbow(np.linspace(0, 1, len(ys)))
rainbow = [colors.rgb2hex(i) for i in colors_array]

# add markers to the map
markers_colors = []
for lat, lng, cluster, cent, gest in zip(df_map_2018['latitud'],
                                         df_map_2018['longitud'],
                                         df_map_2018['KMeans_Clusters'],
                                         df_map_2018['Nombre_Centro'],
                                         df_map_2018['Productividad']) :

    label = folium.Popup(str(cent) + '- Cluster ' + str(cluster), parse_html=True)
    folium.vector_layers.CircleMarker(
        [lat, lng],
        radius=6,
        #popup=label,
        tooltip = str(cent) + ' , ' + str(gest),
        color=rainbow[cluster-1],
        fill=True,
        fill_color=rainbow[cluster-1],
        fill_opacity=0.9).add_to(map_clusters)

map_clusters
```

Make this Notebook Trusted to load map: File -> Trust Notebook



```
# Separación de centros por años
df_map_2021 = df_map[(df_map['Año']<2022)&(df_map['Año']>=2021)]
df_map_2021 = df_map_2021.drop(['Año'],axis = 1)

latitude = 6.51033052
longitude = -75.48477721

map_clusters = folium.Map(location=[latitude, longitude], zoom_start=4)

# set color scheme for the clusters
x = np.arange(2)
ys = [i + x + (i*x)**2 for i in range(4)]
colors_array = cm.rainbow(np.linspace(0, 1, len(ys)))
rainbow = [colors.rgb2hex(i) for i in colors_array]

# add markers to the map
markers_colors = []
for lat, lng, cluster, cent, gest in zip(df_map_2021['latitud'],
                                         df_map_2021['longitud'],
                                         df_map_2021['KMeans_Clusters'],
                                         df_map_2021['Nombre_Centro'],
                                         df_map_2021['Productividad']) :

    label = folium.Popup(str(cent) + '- Cluster ' + str(cluster), parse_html=True)
    folium.vector_layers.CircleMarker(
        [lat, lng],
        radius=6,
        #popup=label,
        tooltip = str(cent) + ' , ' + str(gest),
        color=rainbow[cluster-1],
        fill=True,
        fill_color=rainbow[cluster-1],
        fill_opacity=0.9).add_to(map_clusters)

map_clusters
```

Make this Notebook Trusted to load map: File -> Trust Notebook



Leaflet (<http://leafletjs.com>)

```
#librerías
import numpy as np
import pandas as pd

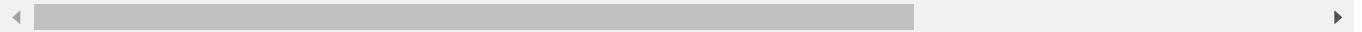
import folium
from folium import plugins

from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler

import matplotlib.pyplot as plt
import matplotlib.cm as cm
import matplotlib.colors as colors

#Cargar archivos desde drive
from google.colab import drive
drive.mount('/content/drive')
```

Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.moun



```
from google.colab import files
path1 = "/content/drive/MyDrive/Proyecto SGPS 9748/TPProceso.xlsx"
path2 = "/content/drive/MyDrive/Proyecto SGPS 9748/Datos.xlsx"
df = pd.read_excel(path1)
df_centros = pd.read_excel(path2, sheet_name= 'Centros_Formacion')
df
```



Centro Año Q Proyectos QNA 5007 QNA 5009 QNA 5010 TV Proyectos

```
df.columns
```

```
Index(['Centro_Año', 'Q_Proyectos', 'QNA_5007', 'QNA_5009', 'QNA_5010',
      'TV_Proyectos'],
      dtype='object')
```

```
2      0.000000000000000000      2      5      5      2      1.071852e+08
```

```
scaler = StandardScaler()
```

```
df[['Q_Proyectos', 'QNA_5007', 'QNA_5009', 'QNA_5010',
    'TV_Proyectos']] = scaler.fit_transform(df[['Q_Proyectos', 'QNA_5007', 'QNA_5009', 'QNA_5010',
    'TV_Proyectos']])
```

```
# variables
```

```
y = df[['Centro_Año']]
```

```
x = df.drop(['Centro_Año'], axis= 1)
```

```
#K-means a la base de datos
```

```
clustering=KMeans(n_clusters=3, max_iter=300,random_state=0) #Crea el modelo
```

```
clustering.fit(x) #Aplica el modelo a la base de datos
```

```
KMeans(n_clusters=3, random_state=0)
```

```
#creando columna de clusters
```

```
df["KMeans_Clusters"]=clustering.labels_
```

```
# la asignación de etiqueta de cluster cambia al ser compilado, por lo que no debe ser consid
# para ser corregido, debe realizarse a partir del PCA y el mapa de correlaciones.
```

```
# Es posible corregirse, posiblemente con algun condicional en los signos y magnitudes del ve
produc= df['KMeans_Clusters'].replace({1:2,0: 1, 2:0})
```

```
produc= df['KMeans_Clusters'].replace({2: 'Alta Gestión',1: 'Media Gestión', 0: 'Baja Gestión
```

```
df['Gestión'] = pd.concat([produc], axis=1)
```

```
centros_años = df["Centro_Año"].str.split("-",expand=True)
```

```
centros_años.columns = ['Codigo_Centro', 'Año']
```

```
df = pd.concat([df, centros_años], axis=1)
```

```
df_map = df[['Codigo_Centro', 'Año', 'KMeans_Clusters', 'Gestión']]
```

```
df_map['Codigo_Centro'] = df_map['Codigo_Centro'].astype(int)
```

```
df_map['Año'] = df_map['Año'].astype(int)
```

```
df_map =df_map.merge(df_centros,on='Codigo_Centro',how='inner')
```

```
df_map = df_map.drop(['Codigo_Regional', 'Codigo_Municipio', 'Codigo_Departamento'],axis= 1)
```

```
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead
```

```
See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user
```

```
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:3: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
```

Try using `.loc[row_indexer,col_indexer] = value` instead

See the caveats in the documentation: <https://pandas.pydata.org/pandas-docs/stable/user>
This is separate from the ipykernel package so we can avoid doing imports until

```
# Separación de centros por años
df_map_2010 = df_map[(df_map['Año']<2011)&(df_map['Año']>=2010)]
df_map_2010 = df_map_2010.drop(['Año'],axis = 1)

latitude = 6.51033052
longitude = -75.48477721

map_clusters = folium.Map(location=[latitude, longitude], zoom_start=4)

# set color scheme for the clusters
x = np.arange(2)
ys = [i + x + (i*x)**2 for i in range(4)]
colors_array = cm.rainbow(np.linspace(0, 1, len(ys)))
rainbow = [colors.rgb2hex(i) for i in colors_array]

# add markers to the map
markers_colors = []
for lat, lng, cluster, cent, gest in zip(df_map_2010['latitud'],
                                         df_map_2010['longitud'],
                                         df_map_2010['KMeans_Clusters'],
                                         df_map_2010['Nombre_Centro'],
                                         df_map_2010['Gestión']) :

    label = folium.Popup(str(cent) + '- Cluster ' + str(cluster), parse_html=True)
    folium.vector_layers.CircleMarker(
        [lat, lng],
        radius=6,
        #popup=label,
        tooltip = str(cent) + ' , ' + str(gest),
        color=rainbow[cluster-1],
        fill=True,
        fill_color=rainbow[cluster-1],
        fill_opacity=0.9).add_to(map_clusters)

map_clusters
```



```
# Separación de centros por años
df_map_2014 = df_map[(df_map['Año']<2015)&(df_map['Año']>=2014)]
df_map_2014 = df_map_2014.drop(['Año'],axis = 1)

latitude = 6.51033052
longitude = -75.48477721

map_clusters = folium.Map(location=[latitude, longitude], zoom_start=4)

# set color scheme for the clusters
x = np.arange(2)
ys = [i + x + (i*x)**2 for i in range(4)]
colors_array = cm.rainbow(np.linspace(0, 1, len(ys)))
rainbow = [colors.rgb2hex(i) for i in colors_array]

# add markers to the map
markers_colors = []
for lat, lng, cluster, cent, gest in zip(df_map_2014['latitud'],
                                         df_map_2014['longitud'],
                                         df_map_2014['KMeans_Clusters'],
                                         df_map_2014['Nombre_Centro'],
                                         df_map_2014['Gestión']) :

    label = folium.Popup(str(cent) + ' - Cluster ' + str(cluster), parse_html=True)
    folium.vector_layers.CircleMarker(
        [lat, lng],
```

```

radius=6,
#popup=label,
tooltip = str(cent) + ' , ' + str(gest),
color=rainbow[cluster-1],
fill=True,
fill_color=rainbow[cluster-1],
fill_opacity=0.9).add_to(map_clusters)

```

map_clusters



```

# Separación de centros por años
df_map_2018 = df_map[(df_map['Año']<2019)&(df_map['Año']>=2018)]
df_map_2018 = df_map_2018.drop(['Año'],axis = 1)

```

```

latitude = 6.51033052
longitude = -75.48477721

```

```

map_clusters = folium.Map(location=[latitude, longitude], zoom_start=4)

```



```

# set color scheme for the clusters
x = np.arange(2)
ys = [i + x + (i*x)**2 for i in range(4)]
colors_array = cm.rainbow(np.linspace(0, 1, len(ys)))
rainbow = [colors.rgb2hex(i) for i in colors_array]

# add markers to the map
markers_colors = []
for lat, lng, cluster, cent, gest in zip(df_map_2018['latitud'],
                                         df_map_2018['longitud'],
                                         df_map_2018['KMeans_Clusters'],
                                         df_map_2018['Nombre_Centro'],
                                         df_map_2018['Gestión']) :

    label = folium.Popup(str(cent) + ' - Cluster ' + str(cluster), parse_html=True)
    folium.vector_layers.CircleMarker(
        [lat, lng],
        radius=6,
        #popup=label,
        tooltip = str(cent) + ' , ' + str(gest),
        color=rainbow[cluster-1],
        fill=True,
        fill_color=rainbow[cluster-1],
        fill_opacity=0.9).add_to(map_clusters)

map_clusters

```



```
# Separación de centros por años
```

```
df_map_2021 = df_map[(df_map['Año']<2022)&(df_map['Año']>=2021)]
```

```
df_map_2021 = df_map_2021.drop(['Año'],axis = 1)
```



```
latitude = 6.51033052
```

```
longitude = -75.48477721
```

```
map_clusters = folium.Map(location=[latitude, longitude], zoom_start=4)
```

```
# set color scheme for the clusters
```

```
x = np.arange(2)
```

```
ys = [i + x + (i*x)**2 for i in range(4)]
```

```
colors_array = cm.rainbow(np.linspace(0, 1, len(ys)))
```

```
rainbow = [colors.rgb2hex(i) for i in colors_array]
```

```
# add markers to the map
```

```
markers_colors = []
```

```
for lat, lng, cluster, cent, gest in zip(df_map_2021['latitud'],
                                         df_map_2021['longitud'],
                                         df_map_2021['KMeans_Clusters'],
                                         df_map_2021['Nombre_Centro'],
                                         df_map_2021['Gestión']) :
```

```
    label = folium.Popup(str(cent) + '- Cluster ' + str(cluster), parse_html=True)
```

```
    folium.vector_layers.CircleMarker(
        [lat, lng],
        radius=6,
        #popup=label,
        tooltip = str(cent) + ' , ' + str(gest),
        color=rainbow[cluster-1],
        fill=True,
        fill_color=rainbow[cluster-1],
        fill_opacity=0.9).add_to(map_clusters)
```

```
map_clusters
```



● ×