PROCESO GESTIÓN DEL TALENTO HUMANO

FORMATO INFORME MENSUAL EJECUCIÓN CONTRACTUAL

Medellín, 30 de abril 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 abril 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	Se realizaron avances en la etapa de compresión del problema así como en el entendimiento de los datos y preparación de	ANEXO 2. Preprocesamiento de datos en python y modelos de Machine Learning Supervisado - Collaboratory ANEXOS 4. Presentación de avances técnicos abril 2022.

	para solucionar de forma eficiente y eficaz problemas o retos identificados en Big Data y analítica de datos con aplicación en gestión de la innovación	estos siguiendo la metodología CRISP-DM.	ANEXOS 5. Actas de avance abril 2022
		Se organizaron datasets, variables y modelo relacional, para determinación de variables críticas.	ANEXO 2. Preprocesamiento de datos en python y modelos de Machine Learning Supervisado - Collaboratory ANEXOS 4. Presentación de avances técnicos abril 2022. ANEXOS 5. Actas de avance abril 2022
		Se iniciaron las etapas de preprocesamiento de datos desde python -colab.	ANEXO 2. Preprocesamiento de datos en python y modelos de Machine Learning Supervisado - Collaboratory ANEXOS 4. Presentación de avances técnicos abril 2022. ANEXOS 5. Actas de avance abril 2022
		Se exploraron modelos de machine learning supervisado de datasets. ● Regresión Logística ● Regresión múltiple ● Máquina de Soporte Vectorial (SVM)	ANEXO 2. Preprocesamiento de datos en python y modelos de Machine Learning Supervisado - Collaboratory ANEXOS 4. Presentación de avances técnicos abril 2022. ANEXOS 5. Actas de avance abril 2022
2	Revisar las herramientas tecnológicas o digitales de Big Data y analítica de datos necesarias para dar cumplimiento a los	Se implementó modelos de machine learning supervisado desde python y R en la plataforma Colaboratory google y ver su rendimiento ante posibles grandes volúmenes de datos.	ANEXO 2. Preprocesamiento de datos en python y modelos de Machine Learning Supervisado - Collaboratory ANEXOS 4. Presentación de avances técnicos abril 2022. ANEXOS 5. Actas de avance abril 2022

	objetivos propuestos dentro del proyecto.		
3	Revisar y actualizar indicadores para la gestión de la innovación según objetivos del proyecto	Se determinaron 4 dataset a trabajar: Principal_Productos Principal_Proyectos Principal_Talento_Humano Principal_Equipos Se prepararon estos datos en la búsqueda de cuantificar indicadores para la gestión de la innovación, proyecto SGPS 9748 prototipo innovación.	ANEXO 2. Preprocesamiento de datos en python y modelos de Machine Learning Supervisado - Collaboratory ANEXOS 4. Presentación de avances técnicos abril 2022. ANEXOS 5. Actas de avance abril 2022
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 CTeI para los Centros SENA Regional Antioquia	Se implementaron modelos de machine learning supervisado desde python y R en la plataforma Colaboratory google , evaluando rendimiento ante posibles grandes volúmenes de datos.	ANEXO 2. Preprocesamiento de datos en python y modelos de Machine Learning Supervisado - Collaboratory ANEXOS 4. Presentación de avances técnicos abril 2022. ANEXOS 5. Actas de avance abril 2022
5	Apoyar la elaboración del manual del sistema y usuario y las actividades de transferencia	Gestión por realizar	Gestión por realizar
6	Presentar los informes de seguimiento del proyecto de	Se elaboró el informe de pago mensual GTH-F-062 V08.	Informe en formato pdf.

	acuerdo con las solicitudes del líder del área o supervisor	Se elaboraron diapositivas en presentación de informes técnicos para el supervisor y coordinador del contrato del mes en curso.	ANEXOS 4. Presentación de avances técnicos abril 2022.
7	Divulgar en eventos académicos/empresariales los resultados del proyecto.	Se aplicó a un (1) evento de divulgación de ciencia, tecnología e innovación para aprendices de semilleros, evento Redcolsi XXI ENCUENTRO DEPARTAMENTAL DE SEMILLEROS DE INVESTIGACIÓN. Nodo Antioquia.	ANEXO 1. Medios de divulgación: Redcolsi
8	Redactar el informe de investigación y de resultados del proyecto.	Se elaboró un (1) informe de pago mensual GTH-F-062 V08. Se elaboraron diapositivas en presentación de informes técnicos para el supervisor y coordinador del contrato del mes en curso.	Informe en formato pdf. ANEXOS 4. Presentación de avances técnicos abril 2022.
9	Redactar y someter a publicación artículo de investigación, libro de investigación o capítulo de libro de investigación relacionado con el proyecto, en medio reconocido por Minciencias.	Se realizó apoyo en la redacción de FORMATOS DE INSCRIPCIÓN DE PROYECTOS PARA PARTICIPACIÓN EN EVENTOS RedCOLSI: proyecto SGPS 9748 para el 2021. Título del proyecto: *Desarrollo de un prototipo para la medición de la Gestión de CTeI para los Centros SENA Regional Antioquia aplicando analítica de datos.*	ANEXO 1. Medios de divulgación: Redcolsi

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 participó en el evento virtual: "Lanzamiento Índice de Innovación para Colombia" Organizado por el Departamento Nacional de Planeación - Gobierno de Colombia.	ANEXOS 3. Participación en evento virtual: Índice de Innovación para Colombia.
11	Apoyar la evaluación técnica, administrativa y presupuestal de los proyectos de investigación aplicada, desarrollo tecnológico e innovación que se formulen por el SENA para el desarrollo de actividades misionales, acorde con las indicaciones dadas por la Dirección General.	Gestión por realizar	Gestión por realizar
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 V08.	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.57925719 de la planilla, miplanilla correspondiente al mes de marzo 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 entidade(s) 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.

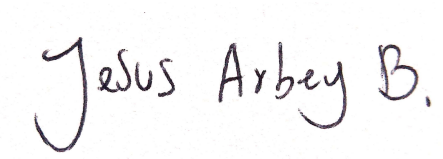
Nota: En el evento de señalar **SI,** el análisis efectuado por el supervisor y ordenador de pago deberá quedar documentado y publicado en el SECOP II en los documentos relacionados en el archivo Gestión Contractual GC_ del periodo objeto de cobro.

Evidencias en 125 folios:

- ANEXOS.

- https://drive.google.com/drive/folders/1j61SX1uD8L8vP3UX7BeYt0NUtED3ZxwB?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: Redcolsi

XXI ENCUENTRO DEPARTAMENTAL DE SEMILLEROS DE INVESTIGACIÓN.
Fecha de inicio: 2022-05-11
Fecha de finalización: 2022-05-13
Medellín, Colombia
Organiza: Fundación Redcolsi
http://fundacionredcolsi.org/eventos/index.php?evento_id=330

FORMATOS DE INSCRIPCIÓN DE PROYECTOS PARA PARTICIPACIÓN EN EVENTOS RedCOLSI

CONSENTIMIENTO EXPRESO Con el envío y de acuerdo con la Ley Estatutaria 1581 de 2012 de Protección de Datos y con el Decreto 1377 de 2013, los responsables del proyecto, autorizamos como Titulares de los datos que plasmamos, que éstos sean incorporados en una base de datos de responsabilidad de la Red Colombiana de semilleros de Investigación RedCOLSI, siendo tratados con la finalidad de gestión administrativa, formativa, evaluación y de información institucional, de semilleristas y de proyectos entre otros, de conformidad con el aviso de privacidad publicado en www.fundacionredcolsi.org. De igual modo, los autores declaran haber sido informados que pueden ejercitar los derechos de acceso, corrección, supresión, revocación o reclamo por infracción sobre datos, mediante escrito dirigido a la dirección de correo electrónico nodoantioquia@fundacionredcolsi.org, indicando en el asunto el derecho que desea ejercitar, o mediante correo ordinario remitido a carrera 12 Número 53 – 42 Edificio Barcelona la Castellana Montería.

Con base en la ley 23 de 1982 de derechos de autor desea que los datos del proyecto puedan ser parte de las memorias de los eventos. Marque con una X Si: ____ No: ____

FORMATO UNICO DE INSCRIPCIÓN - PROYECTOS DE INVESTIGACIÓN

País	Colombia						
Nodo	Antioquia						
Universidad	SENA						
Nombre del Semillero	Semillero de Investigación en Móviles, Educación, Robótica, Automatización y Control (MERLIN) Semillero de Investigación en Proyectos e Ingeniería (SIPI)						
Nivel de Formación (Indique Grado o Semestre)	Tecnólogo						
Programa Académico	Gestión de proyectos económicos y sociales						
Título del Proyecto	Desarrollo de un prototipo para la medición de la Gestión de CTeI para los Centros SENA Regional Antioquia aplicando analítica de datos						
Ponente(s)	Karen Yulieth Estrada Montoya						
Autor(es) (luego del nombre completo agregue n° identificación	Karen Yulieth Estrada Montoya CC 1001760579 Julián Alberto Patiño Murillo CC 8102363 John Jairo Castro Maldonado CC 88265552 Jesús Arbey Benavides Guevara CC 1085260713 Paola Andrea Buitrago CC 43641579 Erika Juliett Restrepo Zapata CC 43973982						
E-mail de Contacto	erestrepoz@sena.edu.co; karenyuliethestradamontoya@gmail.com; japatino@sena.edu.co; jcastrom@sena.edu.co; jabenavides@sena.edu.co; pabuitragoc@sena.edu.co;						
Teléfonos de Contacto	3005017253; 3145117659; 3013004624; 3204185552; 3105595765; 3113671040;						
CATEGORIA (seleccionar una)	Propuesta de Investigación		Investigación en Curso	X	Investigación Terminada		
Área de la investigación (seleccionar una)	Ingeniería				Subárea: Ingeniería		

1. TITULO: Desarrollo de un prototipo de software para la medición de la Gestión de CTeI en los Centros SENA Regional Antioquia aplicando analítica de datos.

2. INTRODUCCIÓN: Descripción breve del tema de investigación, dirigido a orientar al lector sobre la condición a investigar.

Las actividades de Ciencia, Tecnología e innovación (ACTI) están relacionadas con la producción, promoción, difusión y aplicación de conocimientos científicos en los campos que involucran la ciencia y la tecnología (Fuentes Cuadrado & Angulo Cuentas, 2016). Hace algunos años surgió la necesidad de suministrar información que evidencien la recolección, análisis, publicación y normalización de los datos asociados a estas actividades con el objetivo de analizarla a través de indicadores que contribuyan a la toma de decisiones estratégicas (Aguilar, 2017).

Para facilitar lo anterior, desde diferentes áreas cada día es más frecuente utilizar herramientas de análisis de datos ya que permiten analizar rápidamente la información para utilizarla como base para la toma de decisiones, una de esas áreas es la Ciencia, Tecnología e Innovación. En los últimos años en la Regional Antioquia del SENA, los centros de formación han aumentado el desarrollo de ACTI, destacándose por el aumento exponencial de productos y resultados de investigación e innovación generados desde los semilleros y

grupos de investigación existentes, aportando nuevas tecnologías y conocimientos al sector productivo de la región.

Igualmente, se ha evidenciado la continua articulación de las empresas con las diferentes líneas programáticas de SENNOVA en cada centro de formación, consolidando sinergias para el aprovechamiento de los recursos tanto humanos como tecnológicos con los que cuenta la entidad, que aportan al desarrollo económico y social del departamento, a través de los Tecnoparques, los servicios tecnológicos de los centros, las Tecnoacademias, semilleros y grupos de investigación, aportando a la consolidación de la regional Antioquia como una de las más activas a nivel país, no obstante, algunos de los productos no son incluidos dentro de las evidencias que soportan los procesos de Ciencia, Tecnología e Innovación – CTeI en dichos centros de formación, ya que se encuentran debilidades en la definición de indicadores claros que sirvan de guía para orientar las actividades de los entes investigativos y de esta forma ajustarse a lo que realmente acepta el ecosistema de CTeI.

El presente trabajo surgió de la necesidad de desarrollar un prototipo de software como apoyo para la medición de la Gestión de las Actividades de Ciencia, Tecnología e Innovación (ACTI) que facilite el análisis de la información y el diseño de estrategias teniendo en cuenta la información generada y los indicadores definidos. Para ello, se tuvieron en cuenta cuatro de las nueve líneas programáticas de SENNOVA que son Apropiación de la Cultura (Divulgación), Fomento de la Innovación en Centros y Empresas, Investigación y Modernización, con el fin de ser analizadas en la analítica de datos bajo el modelo CRISP – DM (Cantú y Aldeco, 2019).

3. PLANTEAMIENTO DEL PROBLEMA Y JUSTIFICACIÓN:

El índice global de innovación (GII) que diseña en conjunto la organización mundial de la propiedad intelectual (OMPI), la universidad de Cornell y la escuela de negocios INSEAD para realizar la evaluación de las capacidades y resultados de innovación de las economías a nivel mundial, ubican a Colombia con un puntaje bajo en relación a las estadísticas de los demás países analizados, ocupando el puesto 68 de 131 países con una eficiencia de innovación del 40% en contraste con el promedio a nivel mundial que es del 82% (Córdoba, 2020).

En los últimos años se ha logrado el incremento de las actividades de CTeI en los centros de formación SENA de Antioquia, gracias a la continua gestión a los actores del Sistema de investigación, Desarrollo Tecnológico e Innovación del SENA – SENNOVA y la comunidad educativa, sin embargo, se han identificado falencias en la consolidación de la información para la cuantificación de indicadores que permitan articularlos con los estándares que regula los procesos de CTeI en pro de aportar al incremento de los indicadores globales en Colombia.

Teniendo en cuenta lo anterior, algunos de los productos que se desarrollan dentro de las estrategias de Investigación, Desarrollo Tecnológico Investigación – I+D+i de los centros de formación no cuentan con el valor agregado suficiente para generar impacto en las regiones y además no permite desarrollar modelos que demuestren la capacidad de gestión en CTeI de cada una de estas estrategias, lo que dificulta las actividades de seguimiento y control indispensables en los procesos de innovación, por tanto, el desarrollo de indicadores de CTeI es base para el fomento, consolidación de modelos y resurgimiento de nuevos estudios prospectivos, con el fin, de identificar las fortalezas y debilidades de los centros de la regional Antioquia.

En este sentido, en el SENA como institución de carácter nacional se identifica la necesidad de implementar la cuantificación de indicadores y la formulación de modelos de CTeI coherentes con las políticas de CTeI emanadas desde el Ministerio de Ciencia, Tecnología e Innovación de Colombia (MINCIENCIAS) y basados en los estudios que promueven la identificación de características de innovación alineados a políticas y regulaciones de nivel internacional. Por lo expuesto anteriormente se plantea la siguiente pregunta: ¿Cómo aplicar una metodología desde la analítica de datos para medir las actividades de Ciencia, Tecnología e Innovación de los centros de formación en la regional Antioquia del SENA?

4. OBJETIVOS:

Desarrollar un prototipo de software para la medición de la Gestión de CTeI en los Centros SENA Regional Antioquia aplicando analítica de datos.

- Identificar los indicadores que permiten la cuantificación de las ACTI de las líneas SENNOVA de los Centros SENA de la Regional Antioquia.
- Reconocer los requerimientos de la metodología para la aplicación de la analítica de datos.
- Desarrollar el prototipo teniendo en cuenta los requerimientos obtenidos.
- Validar el funcionamiento del prototipo de software con usuarios de los centros de la Regional Antioquia.

5. **REFERENTE TEORICO:**

Cada día es más frecuente utilizar herramientas que permitan analizar de manera rápida y confiable grandes volúmenes de información, esta situación capta cada vez más la atención de diferentes tipos de organizaciones, ya que permiten la visualización de datos que resultan de sus actividades productivas, permitiendo evidenciar su situación actual y facilitando el análisis de la información a través de diferentes modelos, la cual sirve de base para la toma de decisiones (Tejada, Murrieta, Villao & Garzón, 2018).

Teniendo en cuenta la relevancia que tiene para las instituciones educativas para evaluación de la gestión en CTeI, en el trabajo de Serrano & Robledo (2013) se presentan las variables para medir la gestión de innovación tecnológica en instituciones universitarias para medir las capacidades de innovación - CIT, en él construyeron un diagrama de afinidad teniendo en cuenta las dimensiones organizacionales como apoyo a la formulación de lineamientos, políticas y estrategias de gestión organizacional para contribuir a fortalecer el sistema regional y nacional de innovación. Igualmente, en la Universidad del Magdalena se realizó un estudio para proponer un modelo de evaluación de la gestión en CTeI como base para actualizar el plan de CTeI de la institución, utilizando

modelos que se han propuesto en los países como Colombia, Indonesia, Vietnam, Egipto, Turquía y Sudáfrica (Fuentes & Ángulo, 2016).

Igualmente, el trabajo de Zapata (2022) fue realizado con el objetivo de definir un modelo de analítica de datos para monitorear el desempeño académico del colegio Francisco de Miranda, como resultado obtuvo una serie de indicadores que permitió medir las actividades que tienen influencia en el desempeño de los estudiantes, además le permitió generar un cuadro de mando integral para la visualización y comprensión de la situación del colegio en el aspecto académico.

6. METODOLOGIA:

La investigación a desarrollar en el presente proyecto será de enfoque mixto ya que se utilizarán tanto variables de tipo cuantitativo como cualitativo, el tipo de investigación será exploratoria ya que es un tema que será abordado por primera vez en este contexto, el método de estudio será fenomenológico ya que intentará interpretar y comprender el proceso de desarrollo, ejecución y validación del modelo de evaluación y valoración de capacidades de CTeI en los centros de formación, asimismo, el proceso de investigación será inductivo - deductivo. Para la ejecución del proyecto se definen las siguientes fases teniendo en cuenta lo sugerido en la metodología CRISP DM: Comprensión del negocio; Comprensión de los datos; Preparación del conjunto de datos; Modelado de los datos; Evaluación de modelos e interpretación de resultados; Análisis general del modelo.

Para la implementación del desarrollo de software, se determinaron una lista de programas que permiten realizar la analítica de datos, así como el desarrollo del prototipo de software para el almacenamiento de información: SQL-Server; SQL Management; Studio; Node JS; Java Runtime; Git; Anaconda; Rstudio; R y Python.

7. RESULTADOS:

Dentro de los resultados que se tienen hasta el momento se puede destacar que fueron identificadas 4 dimensiones que permiten clasificar los diferentes tipos de indicadores, entre ellos son:

a) Insumos para actividades de Ciencia, Tecnología e Innovación (ACTI)

esta dimensión permite cuantificar indicadores relacionados con: gasto/inversión en ACTI, número de laboratorios para ACTI, inversión de equipos de para ACTI, número de licencias, número de personas dedicadas a ACTI, 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.

b) Proceso

Esta dimensión permite cuantificar indicadores relacionados con: número de proyectos, proporción de gasto en ACTI por personas dedicas a ACTI, 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, eventos de movilidad para ACTI, número de aprendices en formación por semilleros de investigación.

c) Resultado

esta dimensión permite cuantificar indicadores relacionados con: número de convenios con otras instituciones, número de productos de I+D (productos de I+D por tipología Minciencias), número de eventos de CTI organizados por la universidad (número de eventos de CTI por clasificación ACTI), número de asistentes a eventos de CTI organizados por la universidad (por campo de conocimiento según la clasificación de la OCDE), numero de aprendices con formación científica.

d) Productividad

Esta dimensión permite cuantificar indicadores relacionados con: proporción de productos de CTeI por persona dedicadas a ACTI, en relación con el total de personal, proporción de productos de CTeI por unidad organizativa, proporción de eventos de movilidad para ACTI por personas dedicadas a ACTI.

Adicionalmente, se cuenta con un prototipo funcional que permite el almacenamiento de información y al que posteriormente se le adicionará el modelo de analítica de datos que sea elegido de acuerdo a las necesidades y características de la institución, para lo cual se están siguiendo los pasos establecidos en la metodología CRISP – DM, en este momento la investigación ha superado las fases de comprensión del negocio, comprensión de los datos a los que se les está realizando un preprocesamiento para poder continuar con los siguientes pasos.

Se espera en próximas actividades completar las etapas faltantes de la metodología de analítica de datos más apropiada para las necesidades de la institución y luego de ser desarrollada proceder la validación, inicialmente en los centros de formación que integran la Regional Antioquia del SENA.

8. CONCLUSIONES:

La medición de la gestión de las actividades de ciencia, tecnología e innovación se realiza teniendo en cuenta las necesidades y características de cada organización, incluyendo diferentes tipos de medidas. Sin embargo, lo principal es medir las actividades relacionadas con la capacidad de innovación o algunos de sus elementos.

Por su parte las herramientas de análisis de datos permiten adquirir, administrar, compartir y proporcionar datos de primer nivel. Se espera que con la integración del prototipo desarrollado con el modelo de analítica datos se espera el beneficio de un trabajo más simple y de mayor calidad en todas las áreas involucradas en el proceso de toma de decisiones de los centros de la Regional de Antioquia del SENA y esta solución en un momento dado pueda ser utilizado a nivel nacional.

9. BIBLIOGRAFIA:

Tejada-Escobar, F., Murrieta-Marcillo, R., Villao-Santos, F., & Garzón-Balcázar, J. (2018). Big Data en la Educación: Beneficios e Impacto de la Analítica de Datos. *Revista Científica Y Tecnológica UPSE*, *5*(2), 80-88. https://doi.org/10.26423/rctu.v5i2.424

Fuentes Cuadrado, C., & Angulo Cuentas, G. (2016). *Diagnóstico de las capacidades institucionales de ciencia, tecnología e innovación (CTI) de la universidad del magdalena 2014-2015* (Número June). https://doi.org/10.13140/RG.2.2.21646.97605

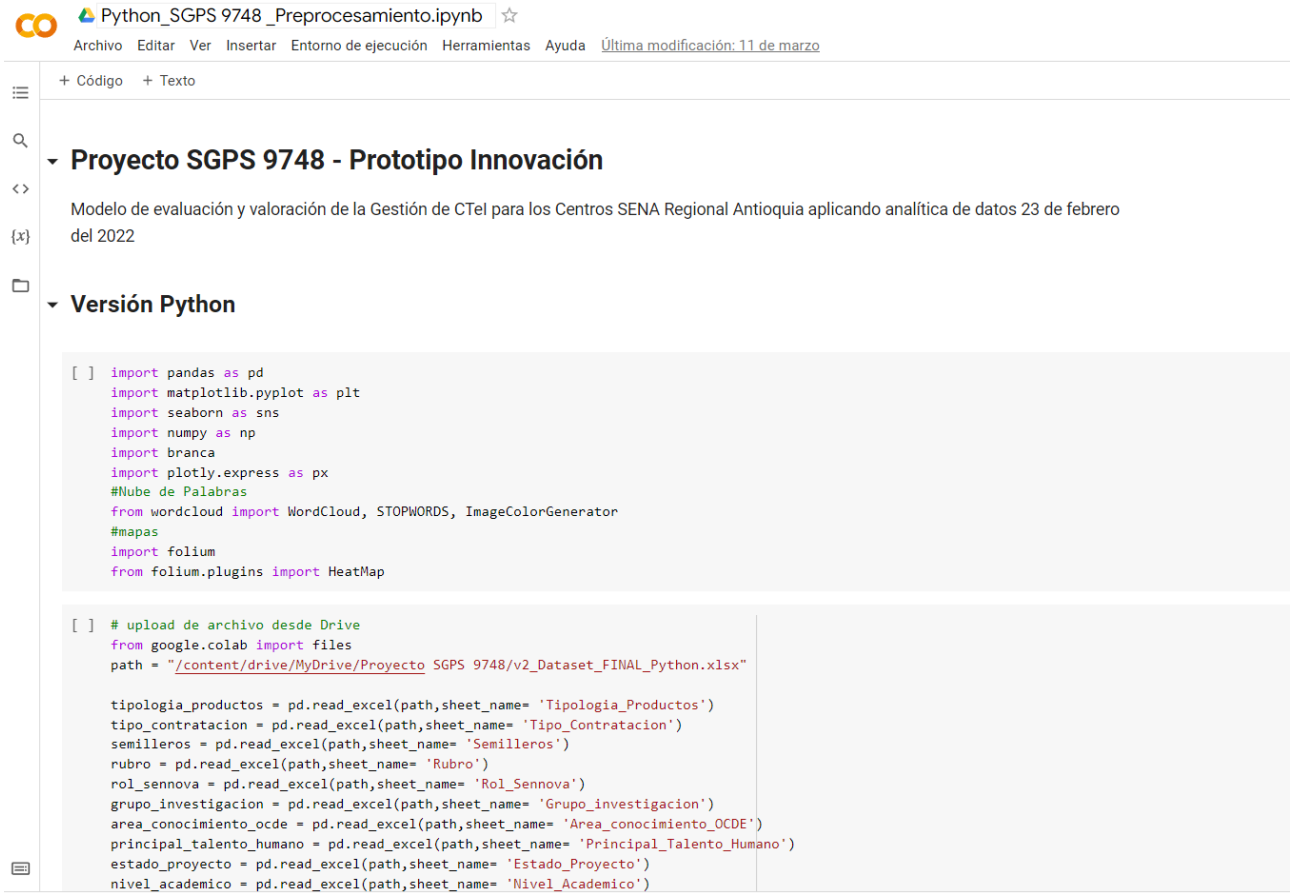
Aguilar, M. (2017). Indicadores de ciencia, tecnología e innovación en Venezuela y su impacto en el desarrollo de políticas públicas. *Telos: revista de Estudios Interdisciplinarios en Ciencias Sociales*, *19*(1), 119–146.

Cantú, Francisco, y Rocío Aldeco. 2019. *Conocimiento y Razonamiento Computacional*. México: Academia Mexicana de Computación, A, C.http://148.247.102.227/wp-content/uploads/2021/01/Libro-CyR.pdf.

Córdoba, R. (2020). Innovación, prioridad para el desarrollo. Consejo Privado de Competitividad. https://compite.com.co/innovacion-prioridad-para-el-desarrollo/

Zapata-Prada, Laura Daniela. 2022. Analítica de datos para monitorear el desempeño de los estudiantes del colegio Francisco De Miranda. Universidad Católica de Colombia. https://repository.ucatolica.edu.co/handle/10983/27216.

ANEXO 2. Preprocesamiento de datos en python y modelos de Machine Learning Supervisado - Collaboratory



Python_SGPS 9748 _Preprocesamiento.ipynb

Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda Última modificación: 11 de marzo

+ Código + Texto

Proyecto SGPS 9748 - Prototipo Innovación

Modelo de evaluación y valoración de la Gestión de CTeI para los Centros SENA Regional Antioquia aplicando analítica de datos 23 de febrero del 2022

Versión Python

```
[ ] import pandas as pd
    import matplotlib.pyplot as plt
    import seaborn as sns
    import numpy as np
    import branca
    import plotly.express as px
    #Nube de Palabras
    from wordcloud import WordCloud, STOPWORDS, ImageColorGenerator
    #mapas
    import folium
    from folium.plugins import HeatMap
```

```
[ ] # upload de archivo desde Drive
    from google.colab import files
    path = "/content/drive/MyDrive/Proyecto SGPS 9748/v2_Dataset_FINAL_Python.xlsx"

    tipologia_productos = pd.read_excel(path,sheet_name= 'Tipologia_Productos')
    tipo_contratacion = pd.read_excel(path,sheet_name= 'Tipo_Contratacion')
    semilleros = pd.read_excel(path,sheet_name= 'Semilleros')
    rubro = pd.read_excel(path,sheet_name= 'Rubro')
    rol_sennova = pd.read_excel(path,sheet_name= 'Rol_Sennova')
    grupo_investigacion = pd.read_excel(path,sheet_name= 'Grupo_investigacion')
    area_conocimiento_ocde = pd.read_excel(path,sheet_name= 'Area_conocimiento_OCDE')
    principal_talento_humano = pd.read_excel(path,sheet_name= 'Principal_Talento_Humano')
    estado_proyecto = pd.read_excel(path,sheet_name= 'Estado_Proyecto')
    nivel_academico = pd.read_excel(path,sheet_name= 'Nivel_Academico')
```

Python_SGPS 9748 _Preprocesamiento.ipynb
Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda Última modificación: 11 de marzo
+ Código + Texto
[ ] principal_equipos = pd.read_excel(path,sheet_name= 'Principal_Equipos')
[ ] 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).
Descripción
tipologia_productos
Describe la tipológía de los productos clasificados por Sena
[ ] tipologia_productos.head()
_Codigo_Tipologia _Nombre_Tipologia
0 7500.0 2.2.1.1 Artículos de investigación A1, A2, B y...
1 7501.0 2.2.1.1 Artículos de investigación D.
2 7502.0 2.2.1.2 Notas científicas.
3 7503.0 2.2.1.3 Libros resultados de investigación.
4 7504.0 2.2.1.4 Capítulos en libro res...
[ ] tipologia_productos.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 30 entries, 0 to 29
Data columns (total 2 columns):
Column Non-Null Count Dtype
0 _Codigo_Tipologia 30 non-null float64
1 _Nombre_Tipologia 30 non-null object

Python_SGPS 9748 _Preprocesamiento.ipynb

Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda Última modificación: 11 de marzo

+ Código + Texto

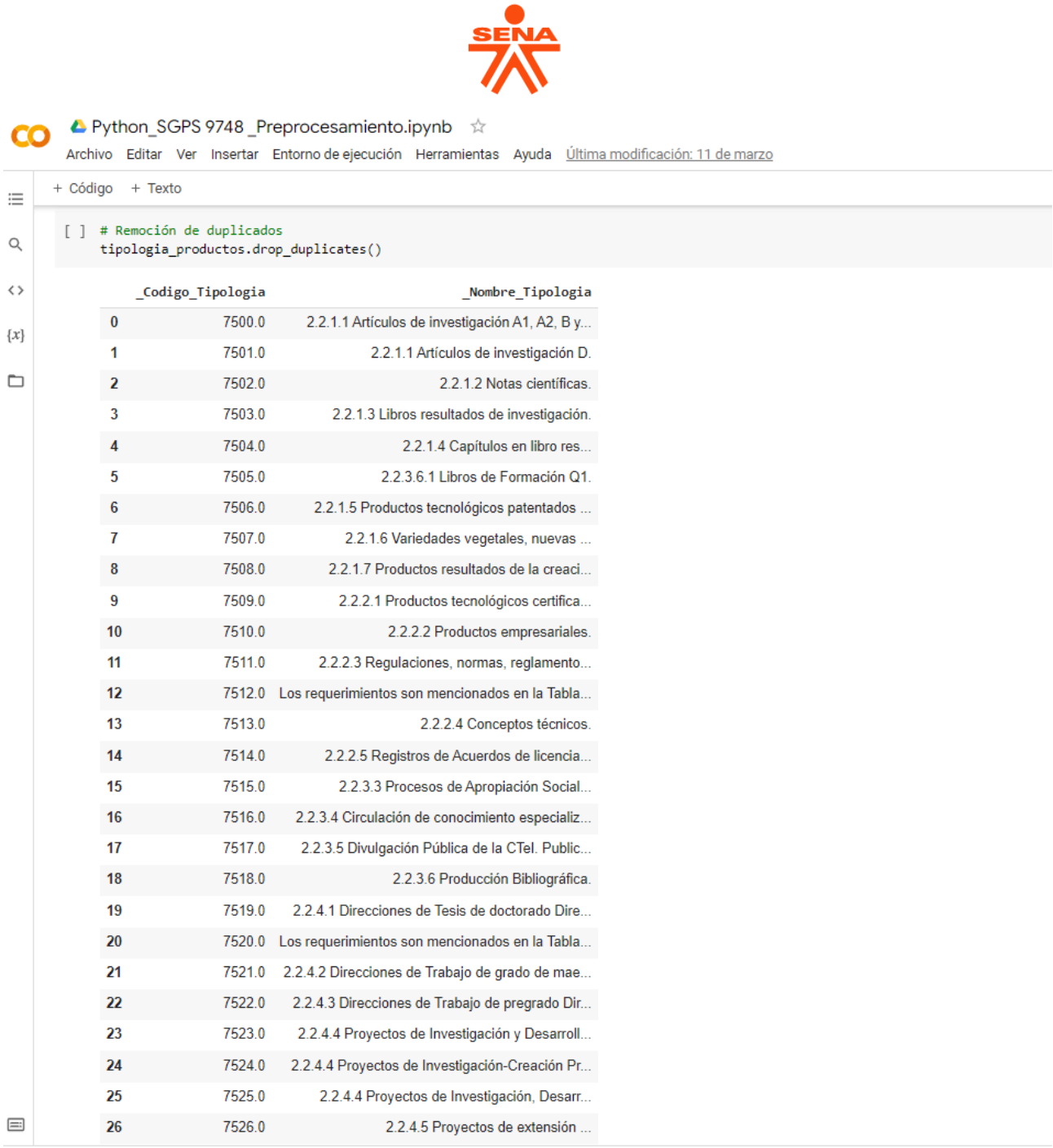
```
[ ] <class 'pandas.core.frame.DataFrame'>
    RangeIndex: 30 entries, 0 to 29
    Data columns (total 2 columns):
     #   Column             Non-Null Count  Dtype
    ---  ------             --------------  -----
     0   _Codigo_Tipologia  30 non-null     float64
     1   _Nombre_Tipologia  30 non-null     object
    dtypes: float64(1), object(1)
    memory usage: 608.0+ bytes
```

```
[ ] tipologia_productos.describe()
```

	_Codigo_Tipologia
count	30.000000
mean	7514.500000
std	8.803408
min	7500.000000
25%	7507.250000
50%	7514.500000
75%	7521.750000
max	7529.000000

```
[ ] # Verficacion de valores null
    f, ax = plt.subplots(figsize=(17, 15))
    ax = sns.heatmap(tipologia_productos.notna())
```

```
[ ]  # Verficacion de valores null
     f, ax = plt.subplots(figsize=(17, 15))
     ax = sns.heatmap(tipologia_productos.notna())
```



	_Codigo_Tipologia	_Nombre_Tipologia
0	7500.0	2.2.1.1 Artículos de investigación A1, A2, B y...
1	7501.0	2.2.1.1 Artículos de investigación D.
2	7502.0	2.2.1.2 Notas científicas.
3	7503.0	2.2.1.3 Libros resultados de investigación.
4	7504.0	2.2.1.4 Capítulos en libro res...
5	7505.0	2.2.3.6.1 Libros de Formación Q1.
6	7506.0	2.2.1.5 Productos tecnológicos patentados ...
7	7507.0	2.2.1.6 Variedades vegetales, nuevas ...
8	7508.0	2.2.1.7 Productos resultados de la creaci...
9	7509.0	2.2.2.1 Productos tecnológicos certifica...
10	7510.0	2.2.2.2 Productos empresariales.
11	7511.0	2.2.2.3 Regulaciones, normas, reglamento...
12	7512.0	Los requerimientos son mencionados en la Tabla...
13	7513.0	2.2.2.4 Conceptos técnicos.
14	7514.0	2.2.2.5 Registros de Acuerdos de licencia...
15	7515.0	2.2.3.3 Procesos de Apropiación Social...
16	7516.0	2.2.3.4 Circulación de conocimiento especializ...
17	7517.0	2.2.3.5 Divulgación Pública de la CTeI. Public...
18	7518.0	2.2.3.6 Producción Bibliográfica.
19	7519.0	2.2.4.1 Direcciones de Tesis de doctorado Dire...
20	7520.0	Los requerimientos son mencionados en la Tabla...
21	7521.0	2.2.4.2 Direcciones de Trabajo de grado de mae...
22	7522.0	2.2.4.3 Direcciones de Trabajo de pregrado Dir...
23	7523.0	2.2.4.4 Proyectos de Investigación y Desarroll...
24	7524.0	2.2.4.4 Proyectos de Investigación-Creación Pr...
25	7525.0	2.2.4.4 Proyectos de Investigación, Desarr...
26	7526.0	2.2.4.5 Proyectos de extensión ...

Python_SGPS 9748 _Preprocesamiento.ipynb ☆

Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda Última modificación: 11 de marzo

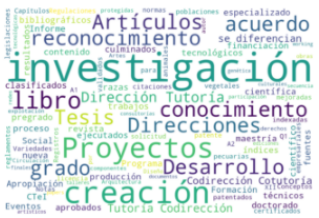
+ Código + Texto

[ ]

22	7522.0	2.2.4.3 Direcciones de Trabajo de pregrado Dir...
23	7523.0	2.2.4.4 Proyectos de Investigación y Desarroll...
24	7524.0	2.2.4.4 Proyectos de Investigación-Creación Pr...
25	7525.0	2.2.4.4 Proyectos de Investigación, Desarr...
26	7526.0	2.2.4.5 Proyectos de extensión ...
27	7527.0	2.2.4.6 Apoyos a la creación de programas y cu...
28	7528.0	2.2.4.7 Acompañamientos y asesorías de lí...
29	7529.0	Informe de investigación

[ ]

```
#Palabras más mecionadas en tipología de productos
#Stopword list:
stopwords = set(STOPWORDS)
stopwords.update(["de", "en", 'la','son','Los','del','los','el','una','un','uno','las','con','Tabla',
                  'trabajo','productos','anexo','resultado','requerimientos', 'mencionados'])

# Word cloud image

text = ' '.join(tipologia_productos['_Nombre_Tipologia'].astype('str'))
wordcloud = WordCloud(width = 1500, height = 1000, stopwords=stopwords, max_font_size=400, max_words=100, background_color="white").generate(text)
plt.figure()
plt.imshow(wordcloud, interpolation="bilinear")
plt.axis("off")
plt.show()
```

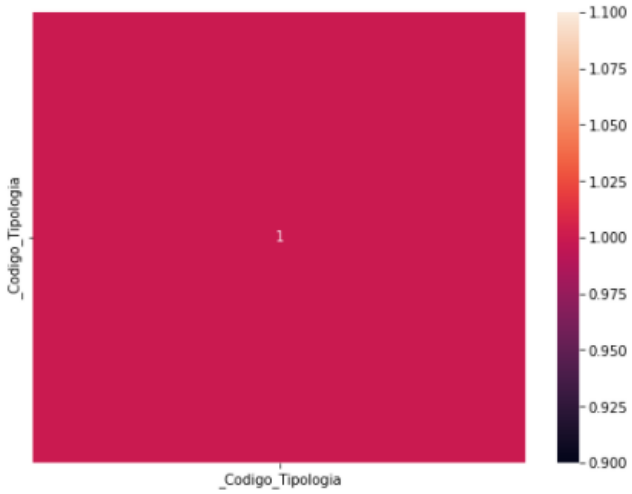


Python_SGPS 9748 _Preprocesamiento.ipynb ☆

Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda Última modificación: 11 de marzo

+ Código + Texto

```
[ ] # Correlación de Pearson
    corr_df = tipologia_productos.corr(method='pearson')
    plt.figure(figsize=(8, 6))
    sns.heatmap(corr_df, annot=True)
    plt.show()
```



tipo_contratacion

```
[ ] tipo_contratacion.head()
```

	_Codigo_Tipo_Contrato	_Nombre_Tipo_Contrato
0	2001.0	Aprendiz
1	2002.0	Contratista
2	2003.0	Planta
3	2004.0	Temporal
4	2005.0	Provisional

Python_SGPS 9748 _Preprocesamiento.ipynb ☆

Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda Última modificación: 11 de marzo

+ Código + Texto

tipo_contratacion

```
[ ] tipo_contratacion.head()
```

	_Codigo_Tipo_Contrato	_Nombre_Tipo_Contrato
0	2001.0	Aprendiz
1	2002.0	Contratista
2	2003.0	Planta
3	2004.0	Temporal
4	2005.0	Provisional

```
[ ] tipo_contratacion.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5 entries, 0 to 4
Data columns (total 2 columns):
 #   Column                Non-Null Count  Dtype  
---  ------                --------------  -----  
 0   _Codigo_Tipo_Contrato  5 non-null      float64
 1   _Nombre_Tipo_Contrato  5 non-null      object 
dtypes: float64(1), object(1)
memory usage: 208.0+ bytes
```

```
[ ] tipo_contratacion.describe()
```

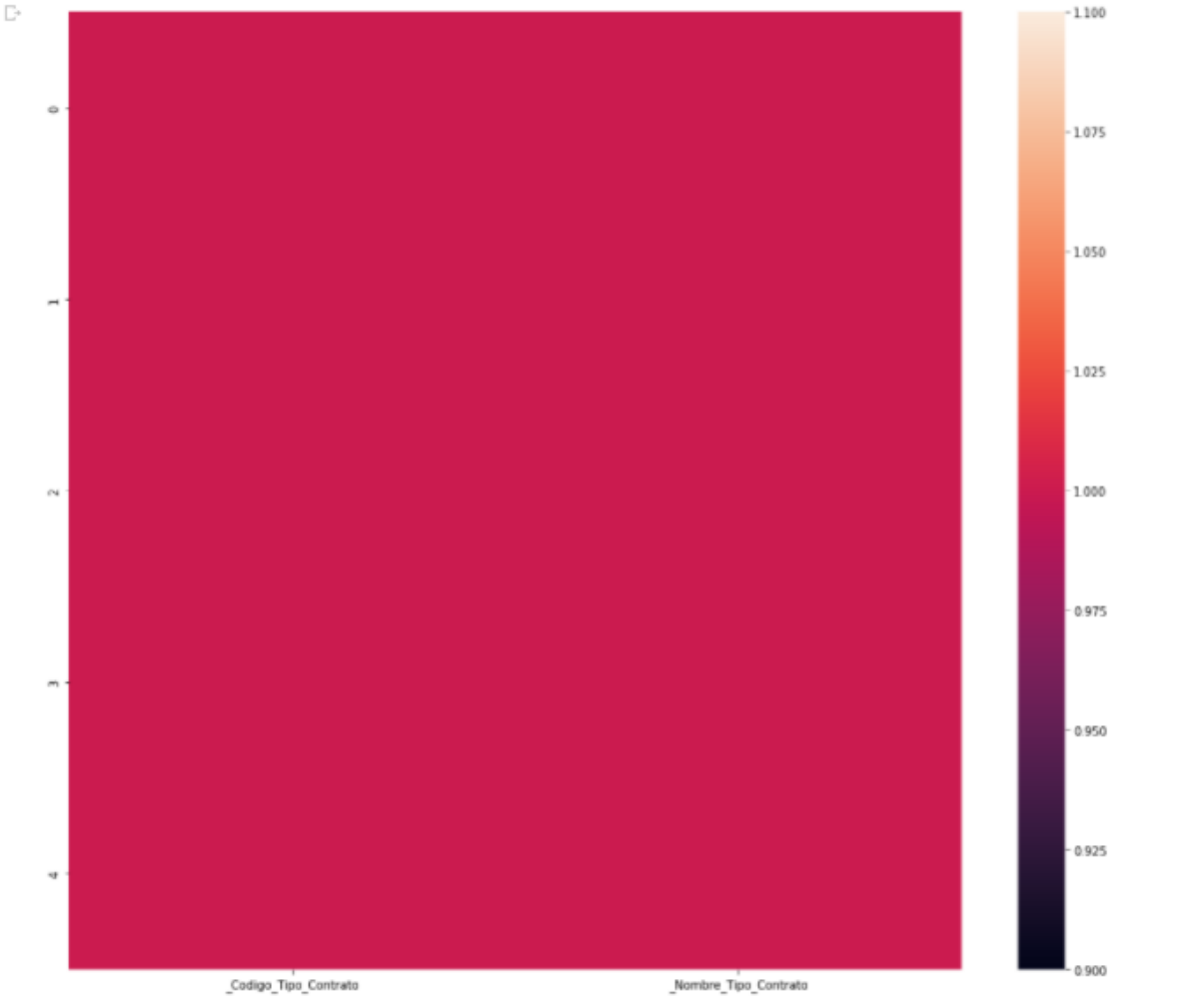
	_Codigo_Tipo_Contrato
count	5.000000
mean	2003.000000
std	1.581139
min	2001.000000
25%	2002.000000
50%	2003.000000
75%	2004.000000
max	2005.000000

Python_SGPS 9748 _Preprocesamiento.ipynb ☆

Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda Última modificación: 11 de marzo

+ Código + Texto

```
# Verficacion de valores null
f, ax = plt.subplots(figsize=(17, 15))
ax = sns.heatmap(tipo_contratacion.notna())
```




```
[ ] semilleros.head()
```

	_Codigo_Semillero	_Nombre_Semillero	_Codigo_Grupo_Investigacion	_Codigo_Centro	_Codigo_Regional
0	8888.0	REDTIS (Semillero de Investigación en Redes, t...	9999.0	9402.0	5.0
1	8889.0	SIALTA (Semillero de Investigación en Logístic...	9999.0	9402.0	5.0
2	8890.0	SIDIN (Semillero de Investigación en Desarroll...	9999.0	9402.0	5.0
3	8891.0	SIG-SENA (Semillero de investigación en sistem...	9999.0	9402.0	5.0
4	8892.0	SIPAEM (Semillero de Investigación del Program...	9999.0	9402.0	5.0

Python_SGPS 9748 _Preprocesamiento.ipynb ☆

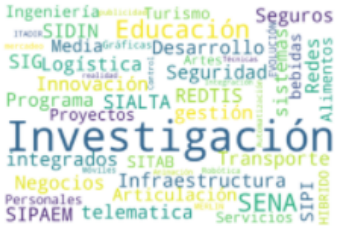
Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda Última modificación: 11 de marzo

+ Código + Texto

```
[ ] semilleros.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 11 entries, 0 to 10
Data columns (total 5 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   _Codigo_Semillero            11 non-null     float64
 1   _Nombre_Semillero            11 non-null     object
 2   _Codigo_Grupo_Investigacion  11 non-null     float64
 3   _Codigo_Centro               11 non-null     float64
 4   _Codigo_Regional             11 non-null     float64
dtypes: float64(4), object(1)
memory usage: 568.0+ bytes
```

```
[ ] #Palabras más mecionadas en tipología de productos
    #Stopword list:
    stopwords = set(STOPWORDS)
    stopwords.update(["de", "en", 'la','son','Los','del','los','el','una','un','uno','las','con', "Semillero"])

    # Word cloud image

    text = ' '.join(semilleros['_Nombre_Semillero'].astype('str'))
    wordcloud = WordCloud(width = 1500, height = 1000, stopwords=stopwords, max_font_size=600, max_words=200, background_color="white").generate(text)
    plt.figure()
    plt.imshow(wordcloud, interpolation="bilinear")
    plt.axis("off")
    plt.show()
```




```
[ ] rubro.head()
```

	_Codigo_Rubro	_Nombre_Rubro
0	23.0	23 - Modernización
1	65.0	65 - Divulgación
2	66.0	66 - Investigación aplicada
3	82.0	82 - Innovación

```
[ ] rubro.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4 entries, 0 to 3
Data columns (total 2 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   _Codigo_Rubro   4 non-null      float64
 1   _Nombre_Rubro   4 non-null      object
dtypes: float64(1), object(1)
memory usage: 192.0+ bytes
```

Python_SGPS 9748 _Preprocesamiento.ipynb ☆

Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda Última modificación: 11 de marzo

+ Código + Texto

```
rubro.head()
```

	_Codigo_Rubro	_Nombre_Rubro
0	23.0	23 - Modernización
1	65.0	65 - Divulgación
2	66.0	66 - Investigación aplicada
3	82.0	82 - Innovación

```
[ ] rubro.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4 entries, 0 to 3
Data columns (total 2 columns):
 #   Column         Non-Null Count  Dtype  
---  ------         --------------  -----  
 0   _Codigo_Rubro  4 non-null      float64
 1   _Nombre_Rubro  4 non-null      object 
dtypes: float64(1), object(1)
memory usage: 192.0+ bytes
```

```
[ ] rol_sennova.head()
```

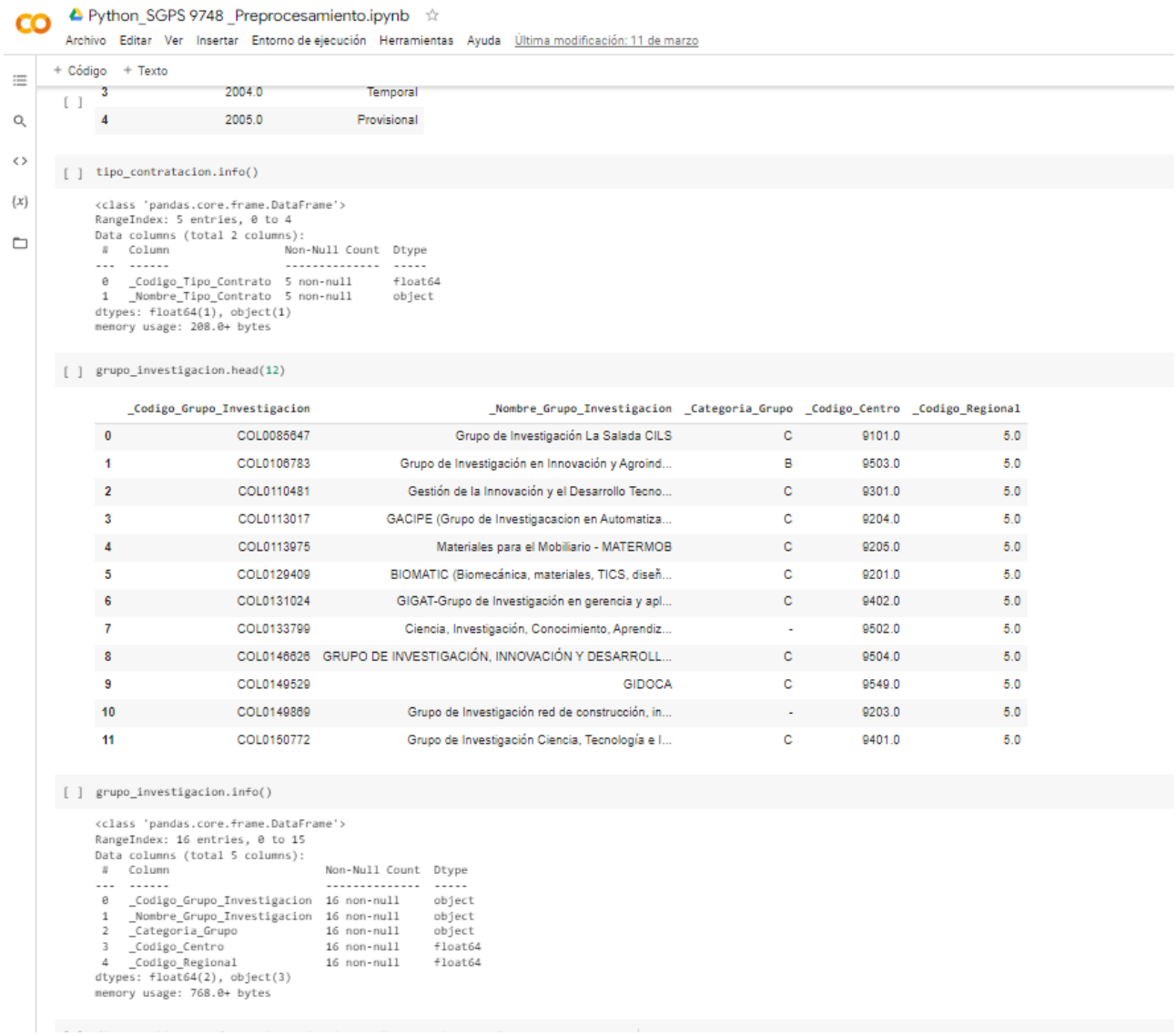
	_Codigo_Rol	_Nombre_Rol
0	1.0	Aprendiz
1	2.0	Semillero
2	3.0	Aprendiz Sennova con Contrato de aprendizaje
3	4.0	Monitor
4	5.0	Dinamizador del conocimiento

```
[ ] rol_sennova.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 13 entries, 0 to 12
Data columns (total 2 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   _Codigo_Rol  13 non-null     float64
 1   _Nombre_Rol  13 non-null     object 
dtypes: float64(1), object(1)
memory usage: 336.0+ bytes
```

```
[ ] tipo_contratacion.head()
```

	_Codigo_Tipo_Contrato	_Nombre_Tipo_Contrato
0	2001.0	Aprendiz
1	2002.0	Contratista
2	2003.0	Planta
3	2004.0	Temporal
4	2005.0	Provisional



Python_SGPS 9748 _Preprocesamiento.ipynb

Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda Última modificación: 11 de marzo

+ Código + Texto

3	2004.0	Temporal
4	2005.0	Provisional

```
[ ] tipo_contratacion.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5 entries, 0 to 4
Data columns (total 2 columns):
 #   Column                Non-Null Count  Dtype  
---  ------                --------------  -----  
 0   _Codigo_Tipo_Contrato  5 non-null      float64
 1   _Nombre_Tipo_Contrato  5 non-null      object 
dtypes: float64(1), object(1)
memory usage: 208.0+ bytes
```

```
[ ] grupo_investigacion.head(12)
```

	_Codigo_Grupo_Investigacion	_Nombre_Grupo_Investigacion	_Categoria_Grupo	_Codigo_Centro	_Codigo_Regional
0	COL0085647	Grupo de Investigación La Salada CILS	C	9101.0	5.0
1	COL0106783	Grupo de Investigación en Innovación y Agroind...	B	9503.0	5.0
2	COL0110481	Gestión de la Innovación y el Desarrollo Tecno...	C	9301.0	5.0
3	COL0113017	GACIPE (Grupo de Investigacacion en Automatiza...	C	9204.0	5.0
4	COL0113975	Materiales para el Mobiliario - MATERMOB	C	9205.0	5.0
5	COL0129409	BIOMATIC (Biomecánica, materiales, TICS, diseñ...	C	9201.0	5.0
6	COL0131024	GIGAT-Grupo de Investigación en gerencia y apl...	C	9402.0	5.0
7	COL0133799	Ciencia, Investigación, Conocimiento, Aprendiz...	-	9502.0	5.0
8	COL0146626	GRUPO DE INVESTIGACIÓN, INNOVACIÓN Y DESARROLL...	C	9504.0	5.0
9	COL0149529	GIDOCA	C	9549.0	5.0
10	COL0149869	Grupo de Investigación red de construcción, in...	-	9203.0	5.0
11	COL0150772	Grupo de Investigación Ciencia, Tecnología e I...	C	9401.0	5.0

```
[ ] grupo_investigacion.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 16 entries, 0 to 15
Data columns (total 5 columns):
 #   Column                      Non-Null Count  Dtype  
---  ------                      --------------  -----  
 0   _Codigo_Grupo_Investigacion  16 non-null     object 
 1   _Nombre_Grupo_Investigacion  16 non-null     object 
 2   _Categoria_Grupo             16 non-null     object 
 3   _Codigo_Centro               16 non-null     float64
 4   _Codigo_Regional             16 non-null     float64
dtypes: float64(2), object(3)
memory usage: 768.0+ bytes
```

Python_SGPS 9748 _Preprocesamiento.ipynb

Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda Última modificación: 11 de marzo

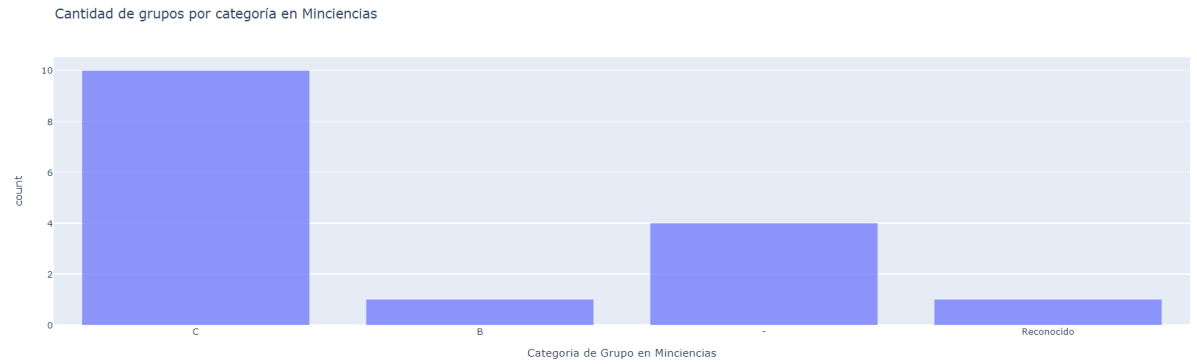
Comentario Compartir

+ Código + Texto

Conectar Edit

```
memory usage: 768.0+ bytes
```

```
[ ] fig = px.histogram(grupo_investigacion, x="_Categoria_Grupo",
                       title="Cantidad de grupos por categoría en Minciencias",
                       labels={'_Categoria_Grupo':' Categoria de Grupo en Minciencias'},
                       opacity=0.7)
    fig.show()
```



Python_SGPS 9748 _Preprocesamiento.ipynb

Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda Última modificación: 11 de marzo

+ Código + Texto

[ ]

Categoria de Grupo en Minciencias

```
[ ] area_conocimiento_ocde.head()
```

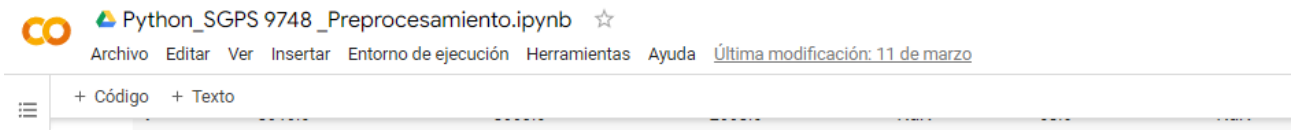
	_Codigo_area_OCDE	_Nombre_area_OCDE
0	1.0	CIENCIAS NATURALES Y EXACTAS
1	2.0	INGENIERÍAS Y TECNOLOGÍAS
2	3.0	CIENCIAS MÉDICAS Y DE LA SALUD
3	4.0	CIENCIAS AGRÍCOLAS
4	5.0	CIENCIAS SOCIALES

```
[ ] area_conocimiento_ocde.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6 entries, 0 to 5
Data columns (total 2 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   _Codigo_area_OCDE  6 non-null      float64
 1   _Nombre_area_OCDE  6 non-null      object
dtypes: float64(1), object(1)
memory usage: 224.0+ bytes
```

```
[ ] principal_talento_humano.head()
```

	_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
0	3000.0	5009.0	2004.0	NaN	66.0	NaN	1.0	5.0	9101.0	3412.0	8000.0
1	3001.0	5005.0	2002.0	NaN	66.0	NaN	1.0	5.0	9127.0	3412.0	8000.0
2	3002.0	5008.0	2002.0	NaN	66.0	NaN	1.0	5.0	9201.0	3412.0	8000.0
3	3008.0	5005.0	2002.0	NaN	65.0	NaN	1.0	5.0	9202.0	3515.0	8012.0
4	3010.0	5009.0	2003.0	NaN	65.0	NaN	1.0	5.0	9203.0	3515.0	8012.0




```
[ ] principal_talento_humano.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 534 entries, 0 to 533
Data columns (total 11 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   _Codigo_Persona              534 non-null    float64
 1   _Codigo_Nivel_Academico      534 non-null    float64
 2   _Codigo_Tipo_Contrato        534 non-null    float64
 3   _Objeto_Contrato             0 non-null      float64
 4   _Codigo_Rubro                534 non-null    float64
 5   _Codigo_Area_OCDE            0 non-null      float64
 6   _Codigo_Estado_Contratacion  534 non-null    float64
 7   _Codigo_Regional             534 non-null    float64
 8   _Codigo_Centro               534 non-null    float64
 9   _Codigo_Proyecto             534 non-null    float64
 10  _Codigo_Producto             534 non-null    float64
dtypes: float64(11)
memory usage: 46.0 KB
```

```
[ ] estado_proyecto.head()
```

	_Codigo_Estado_Proyecto	_Nombre_Estado_Proyecto
0	1.0	Sin iniciar
1	2.0	Ejecución
2	3.0	Terminado
3	4.0	Suspedido

```
[ ] estado_proyecto.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4 entries, 0 to 3
Data columns (total 2 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   _Codigo_Estado_Proyecto  4 non-null      float64
 1   _Nombre_Estado_Proyecto  4 non-null      object
dtypes: float64(1), object(1)
memory usage: 192.0+ bytes
```

2	3.0	Terminado
3	4.0	Suspedido

```
[ ] estado_proyecto.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4 entries, 0 to 3
Data columns (total 2 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   _Codigo_Estado_Proyecto  4 non-null      float64
 1   _Nombre_Estado_Proyecto  4 non-null      object
dtypes: float64(1), object(1)
memory usage: 192.0+ bytes
```

```
[ ] nivel_academico.head(10)
```

	_Codigo_Nivel_Academico	_Nombre_Nivel_Academico
0	5001.0	Bachiller
1	5002.0	Técnico certificado
2	5003.0	Técnico profesional
3	5004.0	Técnico profesional con especialización
4	5005.0	Tecnólogo
5	5006.0	Tecnólogo con especialización
6	5007.0	Profesional
7	5008.0	Profesional con especialización
8	5009.0	Profesional con maestría
9	5010.0	Profesional con doctorado

```
[ ] nivel_academico.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10 entries, 0 to 9
Data columns (total 2 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   _Codigo_Nivel_Academico  10 non-null     float64
 1   _Nombre_Nivel_Academico  10 non-null     object
dtypes: float64(1), object(1)
memory usage: 288.0+ bytes
```

```
dtypes: float64(1), object(1)
memory usage: 288.0+ bytes
```

```
[ ] municipio.head()
```

	_Codigo_Municipio	_Nombre_Municipio	_Codigo_Departamento
0	1.0	MEDELLIN	232.0
1	2.0	ABEJORRAL	232.0
2	4.0	ABRIAQUI	232.0
3	21.0	ALEJANDRIA	232.0
4	30.0	AMAGA	232.0

```
[ ] municipio.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 125 entries, 0 to 124
Data columns (total 3 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   _Codigo_Municipio    125 non-null    float64
 1   _Nombre_Municipio    125 non-null    object
 2   _Codigo_Departamento  125 non-null    float64
dtypes: float64(2), object(1)
memory usage: 3.1+ KB
```

```
[ ] centros_formacion.head(12)
```

	_Codigo_Centro	_Nombre_Centro	_Codigo_Regional	_Codigo_Municipio	_Codigo_Departamento	latitud	longitud
0	9101.0	Centro de los Recursos Naturales Renovables -L...	1.0	129.0	5.0	6.053157	-75.628486
1	9127.0	Centro de Formación Minero Ambiental	1.0	250.0	5.0	7.584361	-74.800660
2	9201.0	Centro del Diseño y Manufactura del Cuero	1.0	360.0	5.0	6.180794	-75.605884
3	9202.0	Centro de Formación en Diseño, Confección y Moda	1.0	360.0	5.0	6.181273	-75.605935
4	9203.0	Centro para el Desarrollo del Hábitat y la Con...	1.0	1.0	5.0	6.303487	-75.567591
5	9204.0	Centro de Tecnología de la Manufactura Avanzada	1.0	1.0	5.0	6.300949	-75.568664
6	9205.0	Centro Tecnológico del Mobiliario	1.0	360.0	5.0	6.180750	-75.606430
7	9206.0	Centro Textil y de Gestión Industrial	1.0	1.0	5.0	6.302387	-75.567365
8	9301.0	Centro de Comercio	1.0	1.0	5.0	6.254794	-75.575466
9	9401.0	Centro de Servicios de Salud	1.0	1.0	5.0	6.251389	-75.571571

Python_SGPS 9748 _Preprocesamiento.ipynb ☆

Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda Última modificación: 11 de marzo

+ Código + Texto

```
[ ] centros_formacion.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 16 entries, 0 to 15
Data columns (total 7 columns):
 #   Column                Non-Null Count  Dtype  
---  ------                --------------  -----  
 0   _Codigo_Centro        16 non-null     float64
 1   _Nombre_Centro        16 non-null     object 
 2   _Codigo_Regional      16 non-null     float64
 3   _Codigo_Municipio     16 non-null     float64
 4   _Codigo_Departamento  16 non-null     float64
 5   latitud               16 non-null     float64
 6   longitud              16 non-null     float64
dtypes: float64(6), object(1)
memory usage: 1.0+ KB
```

```
[ ] #Dados de coordenadas geoespaciais
    my_zip = zip(centros_formacion['latitud'], centros_formacion['longitud'])
    list_of_my_zip = list(my_zip)
```

```
[ ] hm_layer = HeatMap(list_of_my_zip,
                       # These are parameters that we tweak manually to adjust color
                       # See folium docs for more information
                       min_opacity=3,
                       radius= 20,
                       blur= 30,
                      )
```

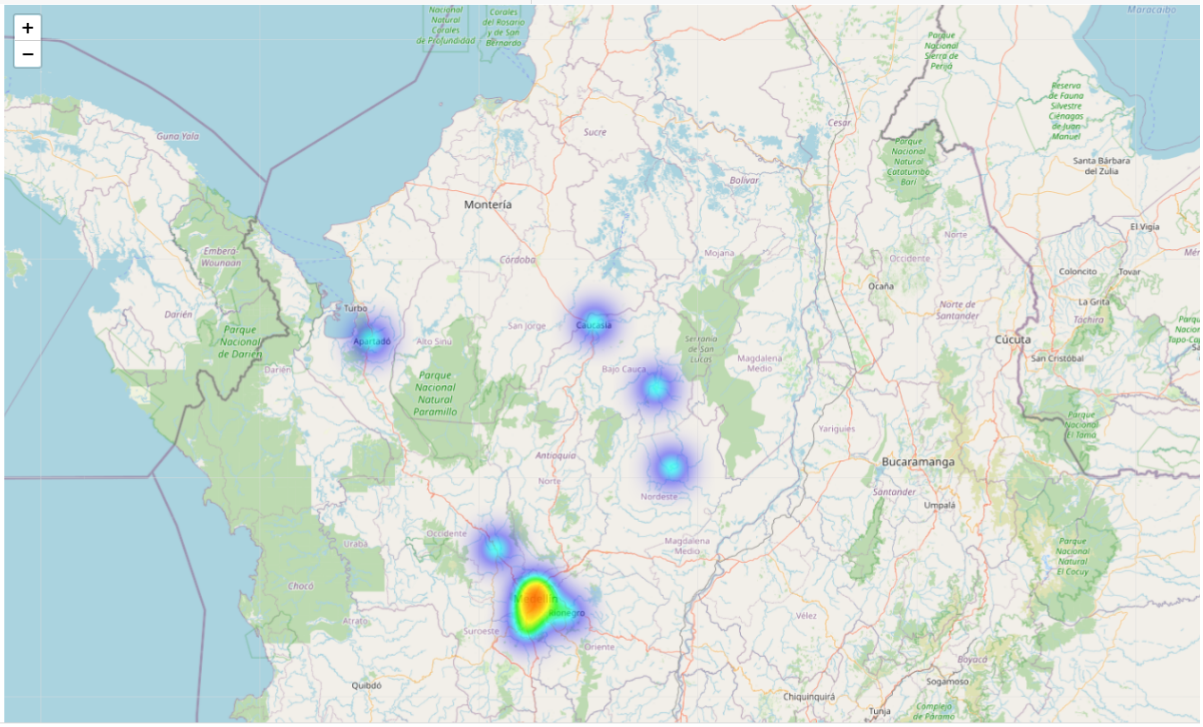
```
[ ] folium_map = folium.Map()
    ny_coords = [6.55, -75.817 ] # lat, long
    folium_map = folium.Map(location=ny_coords,zoom_start=8) #tiles="Stamen Terrain")
    folium_map.add_child(hm_layer)
```

Python_SGPS 9748 _Preprocesamiento.ipynb

Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda Última modificación: 11 de marzo

+ Código + Texto

```
[ ]  folium_map = folium.Map()
     ny_coords = [6.55, -75.817 ] # lat, long
     folium_map = folium.Map(location=ny_coords,zoom_start=8) #tiles="Stamen Terrain")
     folium_map.add_child(hm_layer)
```

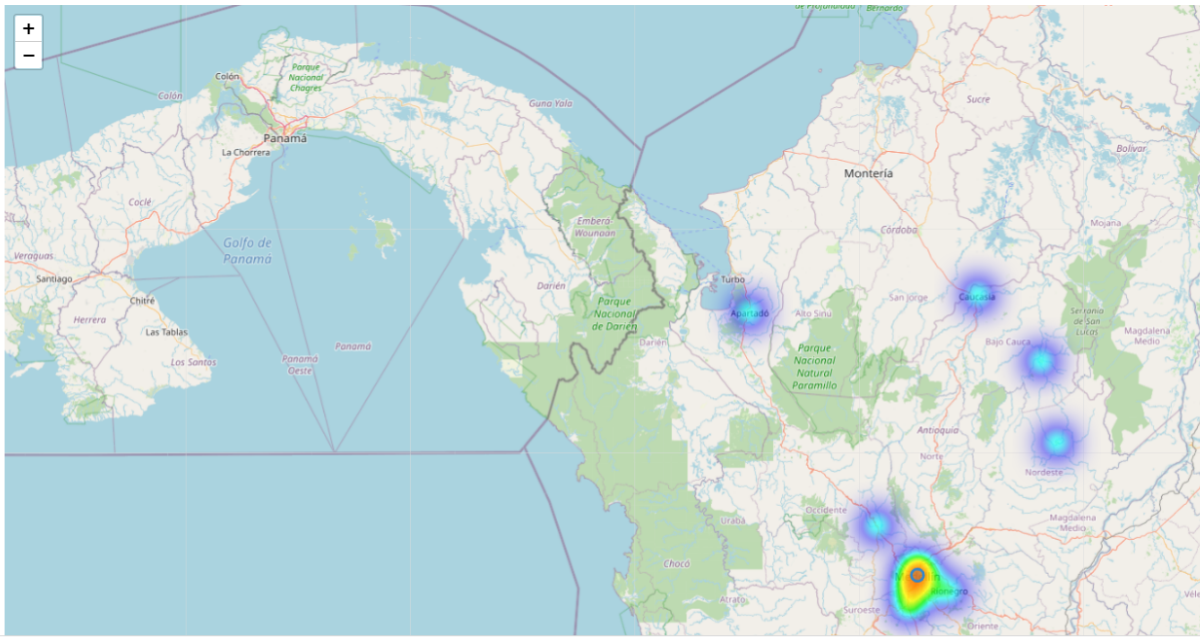


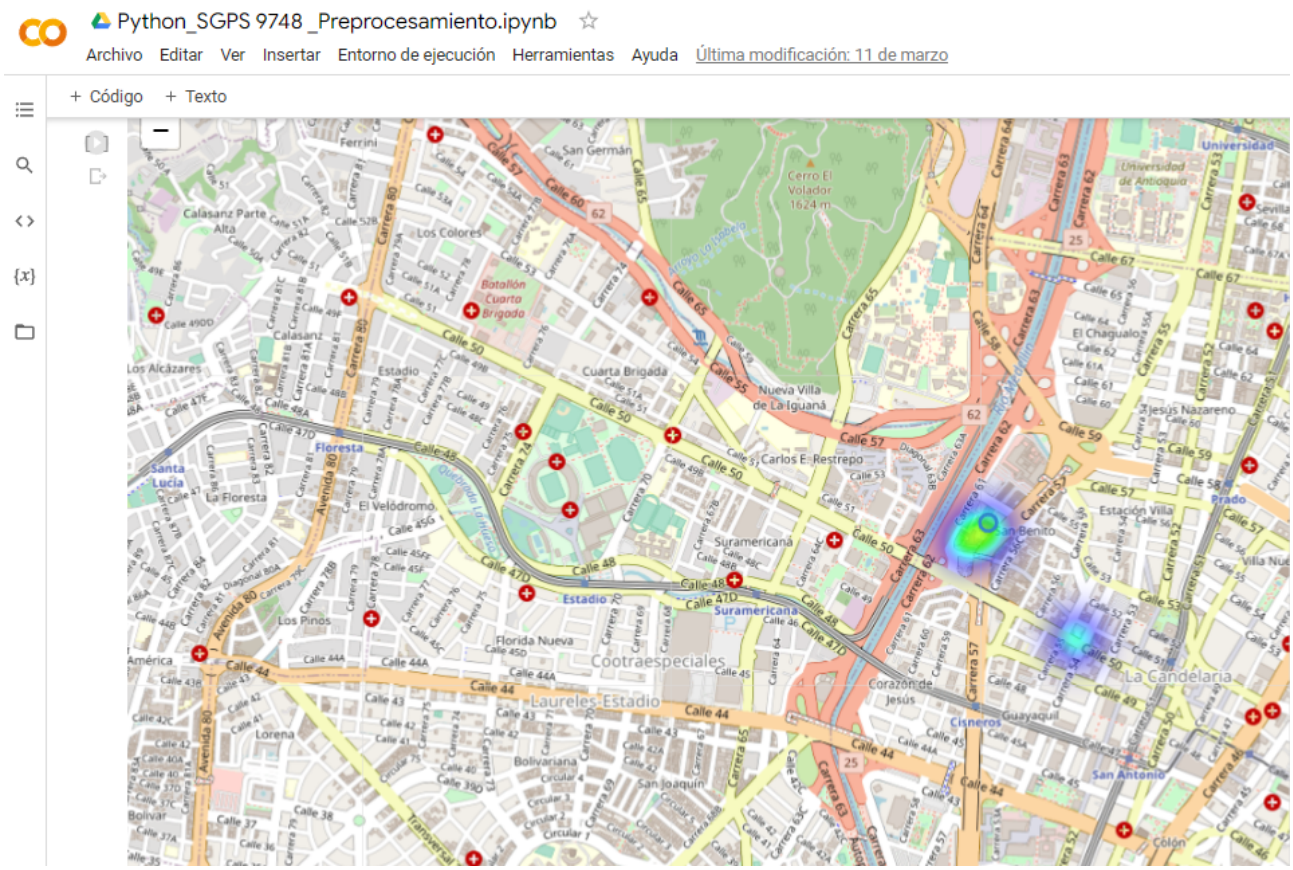
Python_SGPS 9748 _Preprocesamiento.ipynb

Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda Última modificación: 11 de marzo

+ Código + Texto

```
[ ] #Ubicación centro de servicios y gestion empresarial
    folium.CircleMarker(
        location=[6.255546889481773, -75.57468130270018],
        radius=7,
        popup="Laurelhurst Park",
        color="#3186cc",
        fill=True,
        fill_color="#3186cc",
    ).add_to(folium_map)
    folium_map
```




Python_SGPS 9748 _Preprocesamiento.ipynb
Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda Última modificación: 11 de marzo
+ Código + Texto
Ferrini
San Germán
Cerro El Volador 1624 m
Universidad de Antioquia
Calasanz Parte Alta
Los Colores
Batallón Cuarta Brigada
Calasanz
Los Alcázares
Estadio
Cuarta Brigada
Nueva Villa de La Iguaná
Carlos E. Restrepo
Floresta
Santa Lucía
La Floresta
El Velódromo
Suramericana
Estadio
Los Pinos
Florida Nueva
Cootraespeciales
América
Laureles Estadio
Lorena
Bolivariana
San Joaquín
Boliva
Corazón de Jesús
La Candelaria
Cisneros
San Antonio
Jesús Nazareno
El Chagualo
Estación Villa
Prado
Universidad

Python_SGPS 9748 _Preprocesamiento.ipynb ☆

Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda Última modificación: 11 de marzo

+ Código + Texto

[ ]


```
[ ] #Descarga en  un archivo .html
    folium_map.save("Centros de Formación Antioquia.html")
```

```
[ ] departamento.head()
```

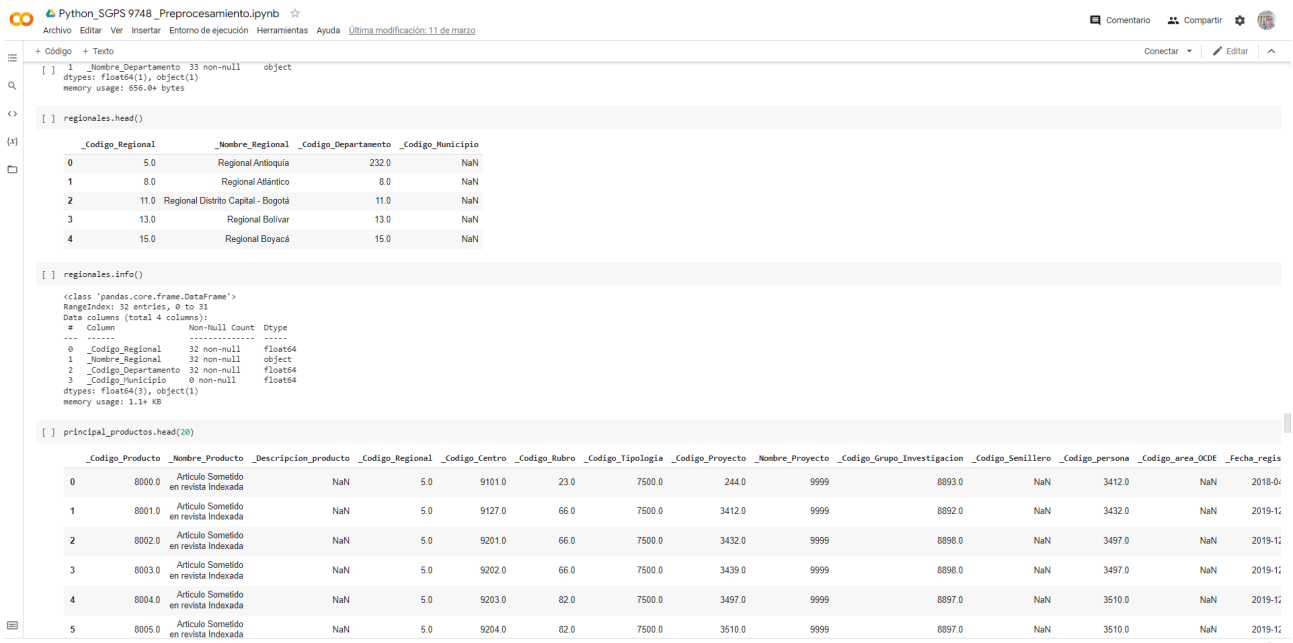
	_Codigo_Departamento	_Nombre_Departamento
0	8.0	ATLANTICO
1	11.0	BOGOTA
2	13.0	BOLIVAR
3	15.0	BOYACA
4	17.0	CALDAS

```
[ ] departamento.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 33 entries, 0 to 32
Data columns (total 2 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   _Codigo_Departamento  33 non-null     float64
 1   _Nombre_Departamento  33 non-null     object
dtypes: float64(1), object(1)
memory usage: 656.0+ bytes
```

```
[ ] regionales.head()
```

	_Codigo_Regional	_Nombre_Regional	_Codigo_Departamento	_Codigo_Municipio
0	5.0	Regional Antioquía	232.0	NaN
1	8.0	Regional Atlántico	8.0	NaN
2	11.0	Regional Distrito Capital - Bogotá	11.0	NaN
3	13.0	Regional Bolívar	13.0	NaN
4	15.0	Regional Boyacá	15.0	NaN



Python_SGPS 9748 _Preprocesamiento.ipynb

Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda Última modificación: 11 de marzo

Comentario Compartir

+ Código + Texto

Conectar Editar

```
 1   _Nombre_Departamento  33 non-null     object
dtypes: float64(1), object(1)
memory usage: 656.0+ bytes
```

```
regionales.head()
```

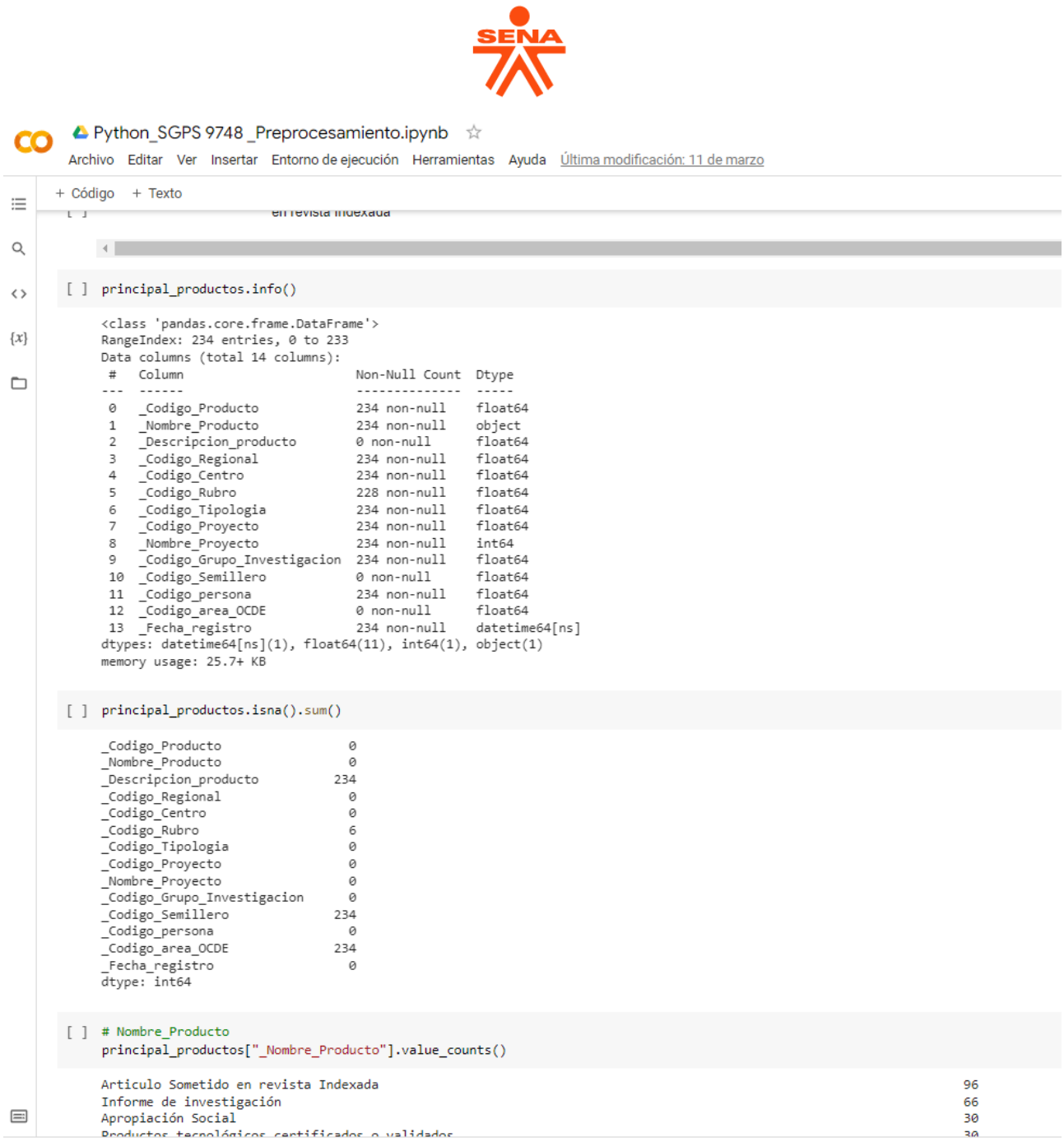
	_Codigo_Regional	_Nombre_Regional	_Codigo_Departamento	_Codigo_Municipio
0	5.0	Regional Antioquia	232.0	NaN
1	8.0	Regional Atlántico	8.0	NaN
2	11.0	Regional Distrito Capital - Bogotá	11.0	NaN
3	13.0	Regional Bolívar	13.0	NaN
4	15.0	Regional Boyacá	15.0	NaN

```
regionales.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 32 entries, 0 to 31
Data columns (total 4 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   _Codigo_Regional      32 non-null     float64
 1   _Nombre_Regional      32 non-null     object
 2   _Codigo_Departamento  32 non-null     float64
 3   _Codigo_Municipio     0 non-null      float64
dtypes: float64(3), object(1)
memory usage: 1.1+ KB
```

```
principal_productos.head(20)
```

	_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_regis
0	8000.0	Articulo Sometido en revista Indexada	NaN	5.0	9101.0	23.0	7500.0	244.0	9999	8893.0	NaN	3412.0	NaN	2018-04
1	8001.0	Articulo Sometido en revista Indexada	NaN	5.0	9127.0	66.0	7500.0	3412.0	9999	8892.0	NaN	3432.0	NaN	2019-12
2	8002.0	Articulo Sometido en revista Indexada	NaN	5.0	9201.0	66.0	7500.0	3432.0	9999	8898.0	NaN	3497.0	NaN	2019-12
3	8003.0	Articulo Sometido en revista Indexada	NaN	5.0	9202.0	66.0	7500.0	3439.0	9999	8898.0	NaN	3497.0	NaN	2019-12
4	8004.0	Articulo Sometido en revista Indexada	NaN	5.0	9203.0	82.0	7500.0	3497.0	9999	8897.0	NaN	3510.0	NaN	2019-12
5	8005.0	Articulo Sometido en revista Indexada	NaN	5.0	9204.0	82.0	7500.0	3510.0	9999	8897.0	NaN	3510.0	NaN	2019-12


Python_SGPS 9748 _Preprocesamiento.ipynb
Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda Última modificación: 11 de marzo
+ Código + Texto
en revista indexada
[ ] principal_productos.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 234 entries, 0 to 233
Data columns (total 14 columns):
Column Non-Null Count Dtype
--- ------ -------------- -----
0 _Codigo_Producto 234 non-null float64
1 _Nombre_Producto 234 non-null object
2 _Descripcion_producto 0 non-null float64
3 _Codigo_Regional 234 non-null float64
4 _Codigo_Centro 234 non-null float64
5 _Codigo_Rubro 228 non-null float64
6 _Codigo_Tipologia 234 non-null float64
7 _Codigo_Proyecto 234 non-null float64
8 _Nombre_Proyecto 234 non-null int64
9 _Codigo_Grupo_Investigacion 234 non-null float64
10 _Codigo_Semillero 0 non-null float64
11 _Codigo_persona 234 non-null float64
12 _Codigo_area_OCDE 0 non-null float64
13 _Fecha_registro 234 non-null datetime64[ns]
dtypes: datetime64[ns](1), float64(11), int64(1), object(1)
memory usage: 25.7+ KB
[ ] principal_productos.isna().sum()
_Codigo_Producto 0
_Nombre_Producto 0
_Descripcion_producto 234
_Codigo_Regional 0
_Codigo_Centro 0
_Codigo_Rubro 6
_Codigo_Tipologia 0
_Codigo_Proyecto 0
_Nombre_Proyecto 0
_Codigo_Grupo_Investigacion 0
_Codigo_Semillero 234
_Codigo_persona 0
_Codigo_area_OCDE 234
_Fecha_registro 0
dtype: int64
[ ] # Nombre_Producto
principal_productos["_Nombre_Producto"].value_counts()
Articulo Sometido en revista Indexada 96
Informe de investigación 66
Apropiación Social 30
Productos tecnológicos certificados o validados 30

Python_SGPS 9748 _Preprocesamiento.ipynb

Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda Última modificación: 11 de marzo

+ Código + Texto

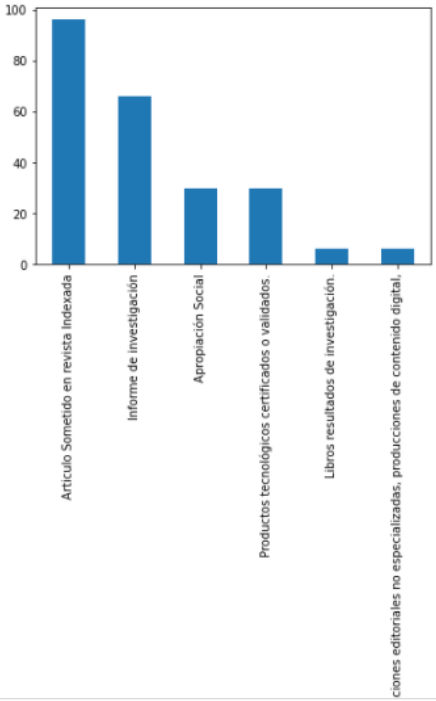
```
dtype: int64
```

```
[ ] # Nombre_Producto
    principal_productos["_Nombre_Producto"].value_counts()
```

```
Articulo Sometido en revista Indexada                                                    96
Informe de investigación                                                                 66
Apropiación Social                                                                       30
Productos tecnológicos certificados o validados.                                         30
Libros resultados de investigación.                                                       6
Divulgación Pública de la CTeI. Publicaciones editoriales no especializadas, producciones de contenido digital,     6
Name: _Nombre_Producto, dtype: int64
```

```
[ ] # Nombre_Producto
    plot_principal_productos = principal_productos["_Nombre_Producto"].value_counts()
    plot_principal_productos.plot.bar()
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f89d98bd310>
```



Python_SGPS 9748 _Preprocesamiento.ipynb ☆

Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda Última modificación: 11 de marzo

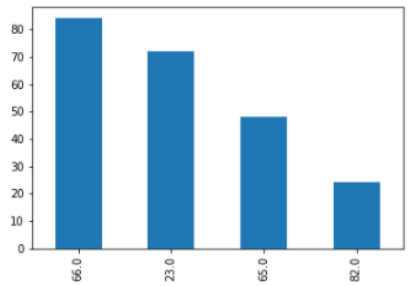
+ Código + Texto

```
[ ] # Rubro
    principal_productos["_Codigo_Rubro"].value_counts()
```

```
66.0    84
23.0    72
65.0    48
82.0    24
Name: _Codigo_Rubro, dtype: int64
```

```
[ ] plot__Codigo_Rubro = principal_productos["_Codigo_Rubro"].value_counts()
    plot__Codigo_Rubro.plot.bar()
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f89d92d5dd0>
```




```
[ ] #Cantidad de Grupos de investigación
    principal_productos["_Codigo_Grupo_Investigacion"].value_counts()
```

```
8897.0    54
8893.0    44
8892.0    41
8898.0    41
8894.0    22
8888.0    18
8889.0    14
Name: _Codigo_Grupo_Investigacion, dtype: int64
```

Python_SGPS 9748 _Preprocesamiento.ipynb

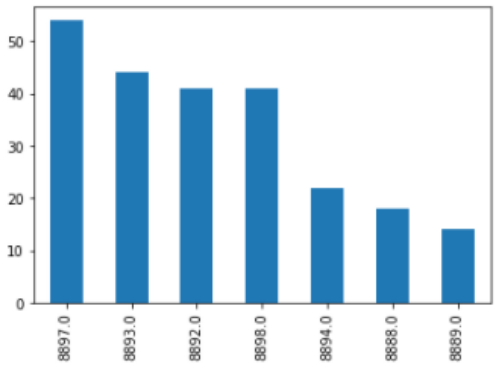
Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda Última modificación: 11 de marzo

+ Código + Texto

```
Name: _Codigo_Grupo_Investigacion, dtype: int64
```

```
[ ] plot_Codigo_Grupo_Investigacion = principal_productos["_Codigo_Grupo_Investigacion"].value_counts()
    plot_Codigo_Grupo_Investigacion.plot.bar()
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f89d995e390>
```




```
[ ] #Cantidad de productos por persona
    principal_productos["_Codigo_persona"].value_counts()
```

```
3515.0     30
3510.0     24
3439.0     24
6263.0     24
244.0      24
3432.0     18
3497.0     18
6299.0     18
6234.0     18
3961.0     18
3412.0     12
12336.0     6
Name: _Codigo_persona, dtype: int64
```

```
[ ] plot_Codigo_persona = principal_productos["_Codigo_persona"].value_counts()
    plot_Codigo_persona.plot.bar()
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f89d986c910>
```

Python_SGPS 9748 _Preprocesamiento.ipynb

Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda Última modificación: 11 de marzo

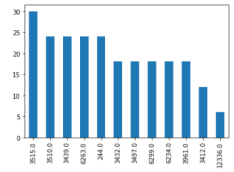
Comentario Compartir

+ Código + Texto

Conectar Editar

```
[ ]
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f89d986c910>
```




```
[ ] #Agrupación para canditidad de producción por grupo
    #g = principal_productos.groupby(by= principal_productos['_Codigo_Centro'])
    #principal_productos[['_Codigo_Centro']].value_counts()
    principal_productos_centros = principal_productos.groupby("_Codigo_Centro").count()
    principal_productos_centros
```

_Codigo_Centro	_Codigo_Producto	_Nombre_Producto	_Descripcion_producto	_Codigo_Regional	_Codigo_Rubro	_Codigo_Tipologia	_Codigo_Proyecto	_Nombre_Proyecto	_Codigo_Grupo_Investigacion	_Codigo_Semillero	_Codigo_persona	_Codigo_area_OCDE	_Fecha_registro
9101.0	7	7	0	7	7	7	7	7	7	0	7	0	7
9127.0	7	7	0	7	7	7	7	7	7	0	7	0	7
9201.0	4	4	0	4	4	4	4	4	4	0	4	0	4
9202.0	7	7	0	7	7	7	7	7	7	0	7	0	7
9203.0	7	7	0	7	7	7	7	7	7	0	7	0	7
9204.0	10	10	0	10	10	10	10	10	10	0	10	0	10
9205.0	16	16	0	16	15	16	16	16	16	0	16	0	16
9206.0	13	13	0	13	13	13	13	13	13	0	13	0	13
9301.0	13	13	0	13	13	13	13	13	13	0	13	0	13
9401.0	7	7	0	7	6	7	7	7	7	0	7	0	7
9402.0	103	103	0	103	100	103	103	103	103	0	103	0	103
9501.0	4	4	0	4	4	4	4	4	4	0	4	0	4
9502.0	4	4	0	4	4	4	4	4	4	0	4	0	4

Python_SGPS 9748 _Preprocesamiento.ipynb

Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda Última modificación: 11 de marzo

Comentario Compartir

+ Código + Texto

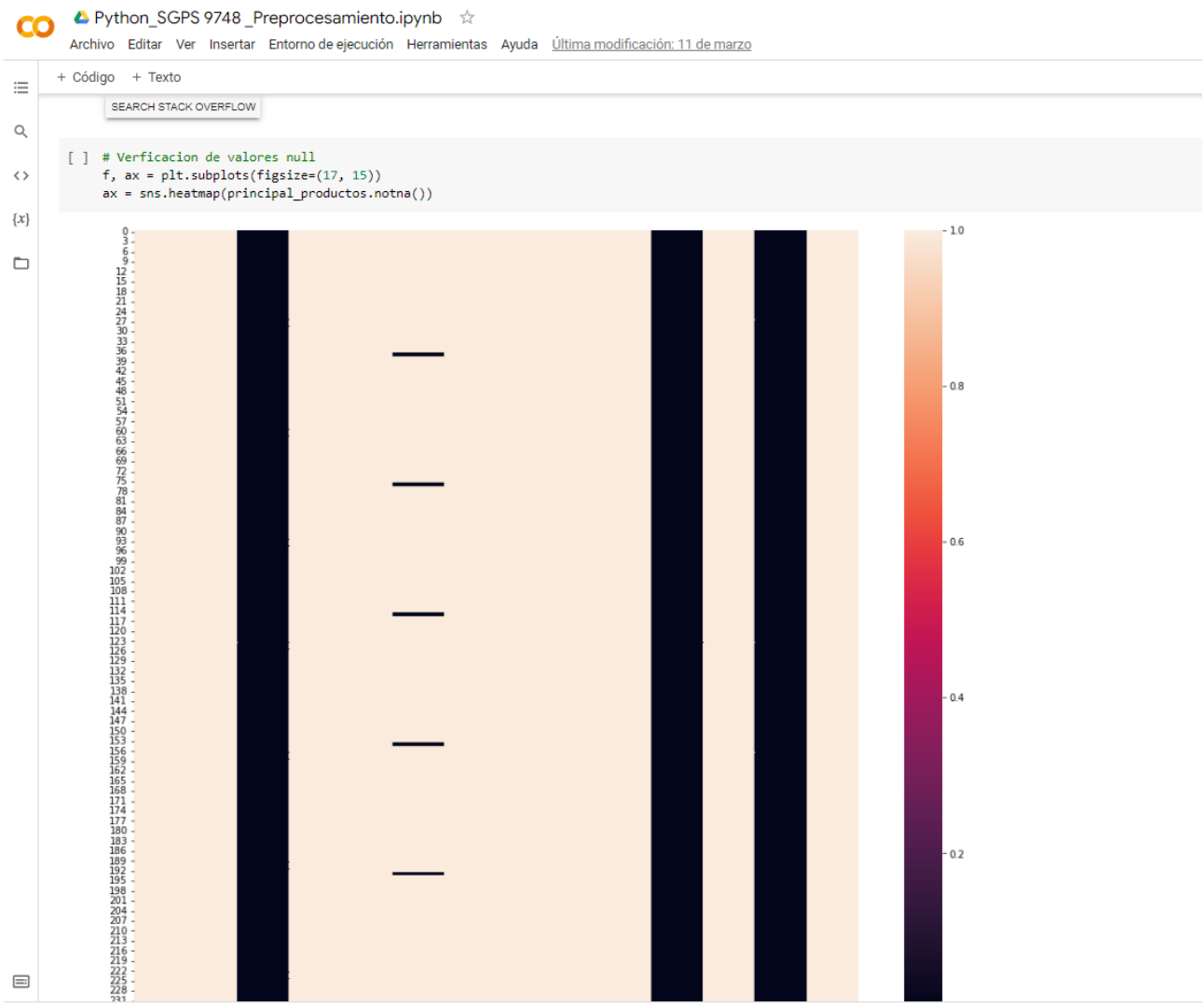
Conectar Editar

9549.0	18	18	0	18	17	18	18	18	18	0	18	0	18

```
[ ] def
      if  centros_formacion['_Codigo_Centro'] = principal_productos_centros["_Codigo_Centro"]
    return


    principal_productos_centros['latitud'],principal_productos_centros['longitud'] =  centros_formacion['latitud'], centros_formacion['longitud']
```

```
[ ] principal_productos_centros
```

_Codigo_Centro	_Codigo_Producto	_Nombre_Producto	_Descripcion_producto	_Codigo_Regional	_Codigo_Rubro	_Codigo_Tipologia	_Codigo_Proyecto	_Nombre_Proyecto	_Codigo_Grupo_Investigacion	_Codigo_Semillero	_Codigo_persona	_Codigo_area_OCDE	_Fecha_registro
9101.0	7	7	0	7	7	7	7	7	7	0	7	0	7
9127.0	7	7	0	7	7	7	7	7	7	0	7	0	7
9201.0	4	4	0	4	4	4	4	4	4	0	4	0	4
9202.0	7	7	0	7	7	7	7	7	7	0	7	0	7
9203.0	7	7	0	7	7	7	7	7	7	0	7	0	7
9204.0	10	10	0	10	10	10	10	10	10	0	10	0	10
9205.0	16	16	0	16	15	16	16	16	16	0	16	0	16
9206.0	13	13	0	13	13	13	13	13	13	0	13	0	13
9301.0	13	13	0	13	13	13	13	13	13	0	13	0	13
9401.0	7	7	0	7	6	7	7	7	7	0	7	0	7
9402.0	103	103	0	103	100	103	103	103	103	0	103	0	103
9501.0	4	4	0	4	4	4	4	4	4	0	4	0	4
9502.0	4	4	0	4	4	4	4	4	4	0	4	0	4
9503.0	4	4	0	4	4	4	4	4	4	0	4	0	4
9504.0	10	10	0	10	10	10	10	10	10	0	10	0	10
9549.0	18	18	0	18	17	18	18	18	18	0	18	0	18

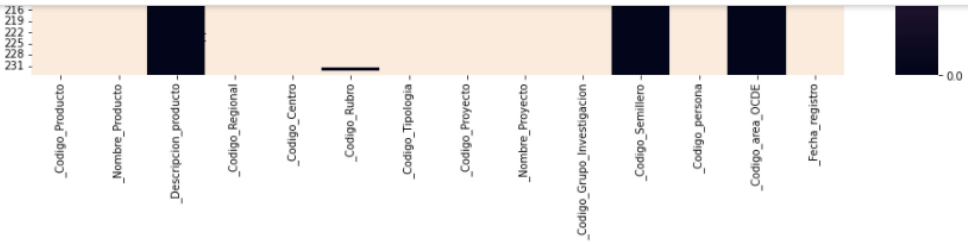

Python_SGPS 9748 _Preprocesamiento.ipynb
Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda Última modificación: 11 de marzo
+ Código + Texto
SEARCH STACK OVERFLOW
[ ] # Verficacion de valores null
f, ax = plt.subplots(figsize=(17, 15))
ax = sns.heatmap(principal_productos.notna())
1.0
0.8
0.6
0.4
0.2

Python_SGPS 9748 _Preprocesamiento.ipynb ☆

Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda Última modificación: 11 de marzo

+ Código + Texto

[ ]



[ ] `tipo_equipo.head()`

	_Codigo_Tipo_Equipo	_Nombre_Tipo_Equipo

[ ] `tipo_equipo.info()`

```
<class 'pandas.core.frame.DataFrame'>
Index: 0 entries
Data columns (total 2 columns):
 #   Column               Non-Null Count  Dtype 
---  ------               --------------  ----- 
 0   _Codigo_Tipo_Equipo  0 non-null      object
 1   _Nombre_Tipo_Equipo  0 non-null      object
dtypes: object(2)
memory usage: 0.0+ bytes
```

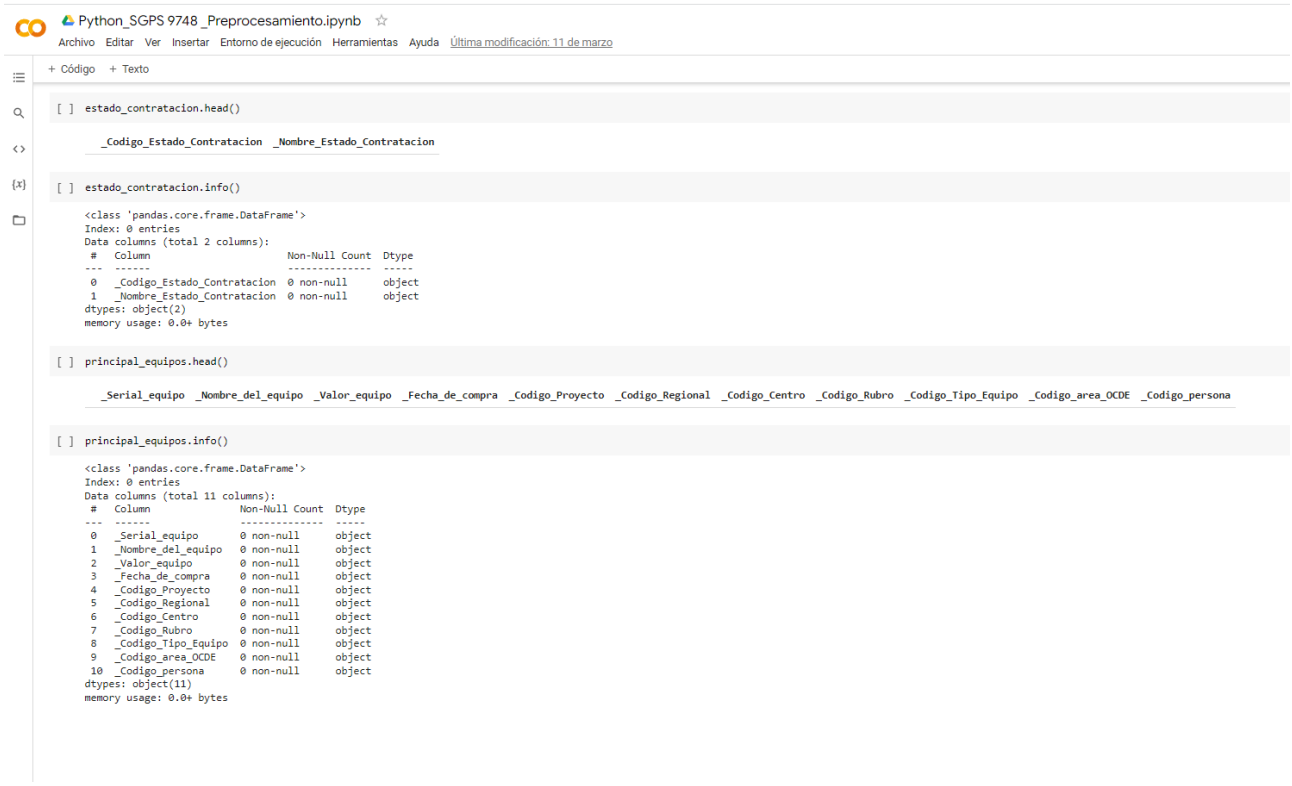
[ ] `rol_proyectos.head()`

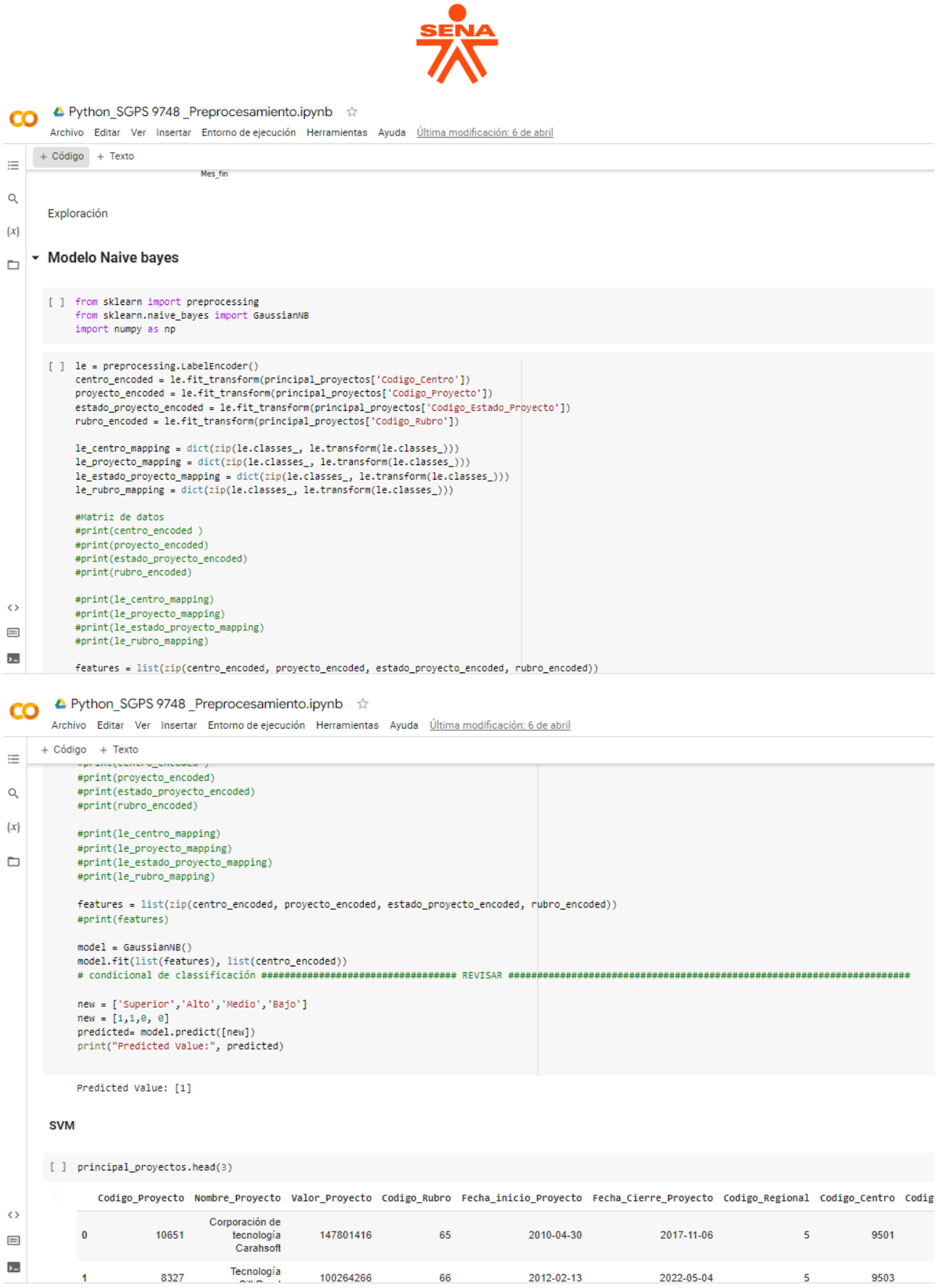
	_Codigo_Rol_Proyecto	_Nombre_Rol

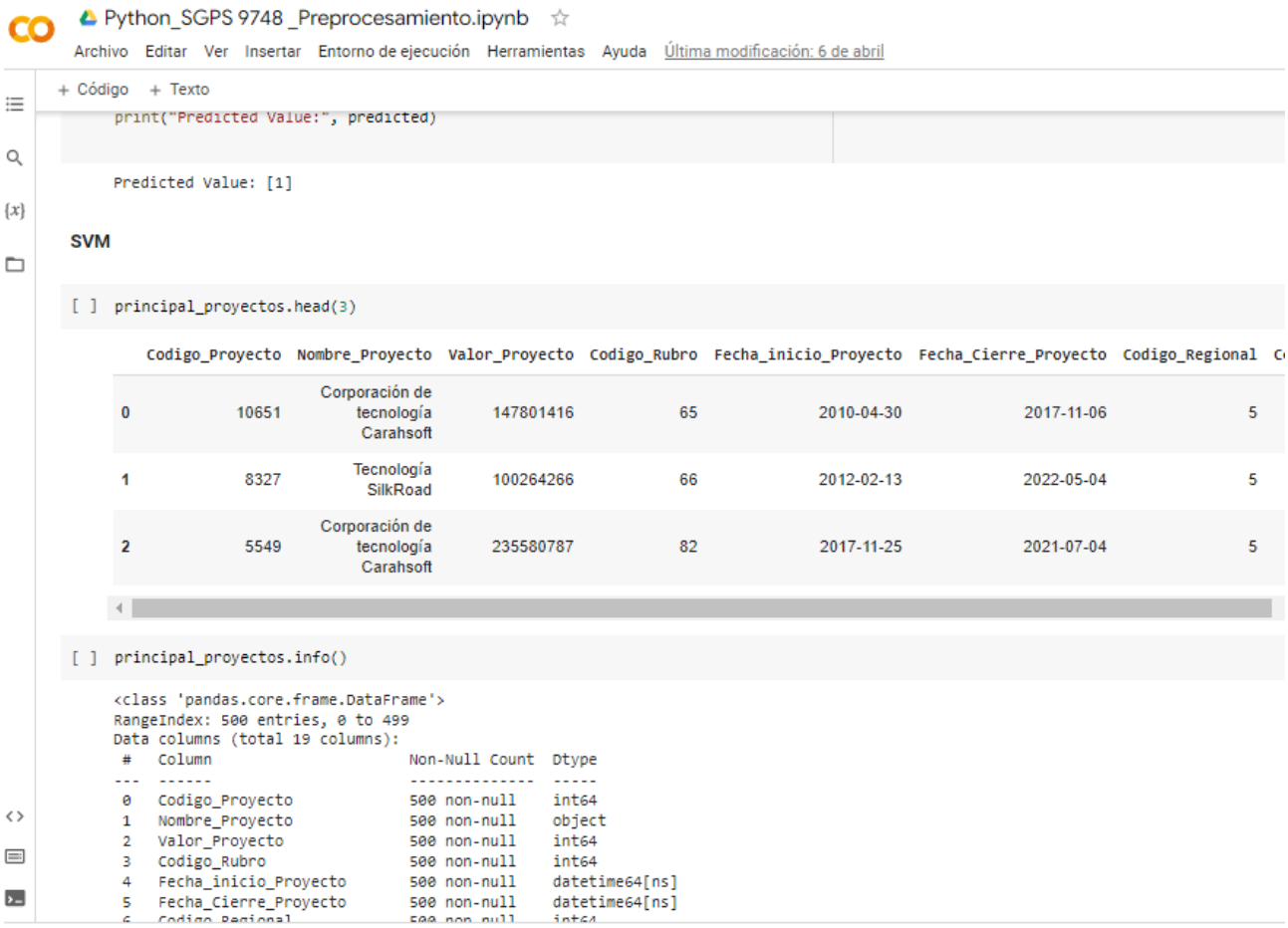
[ ] `rol_proyectos.info()`

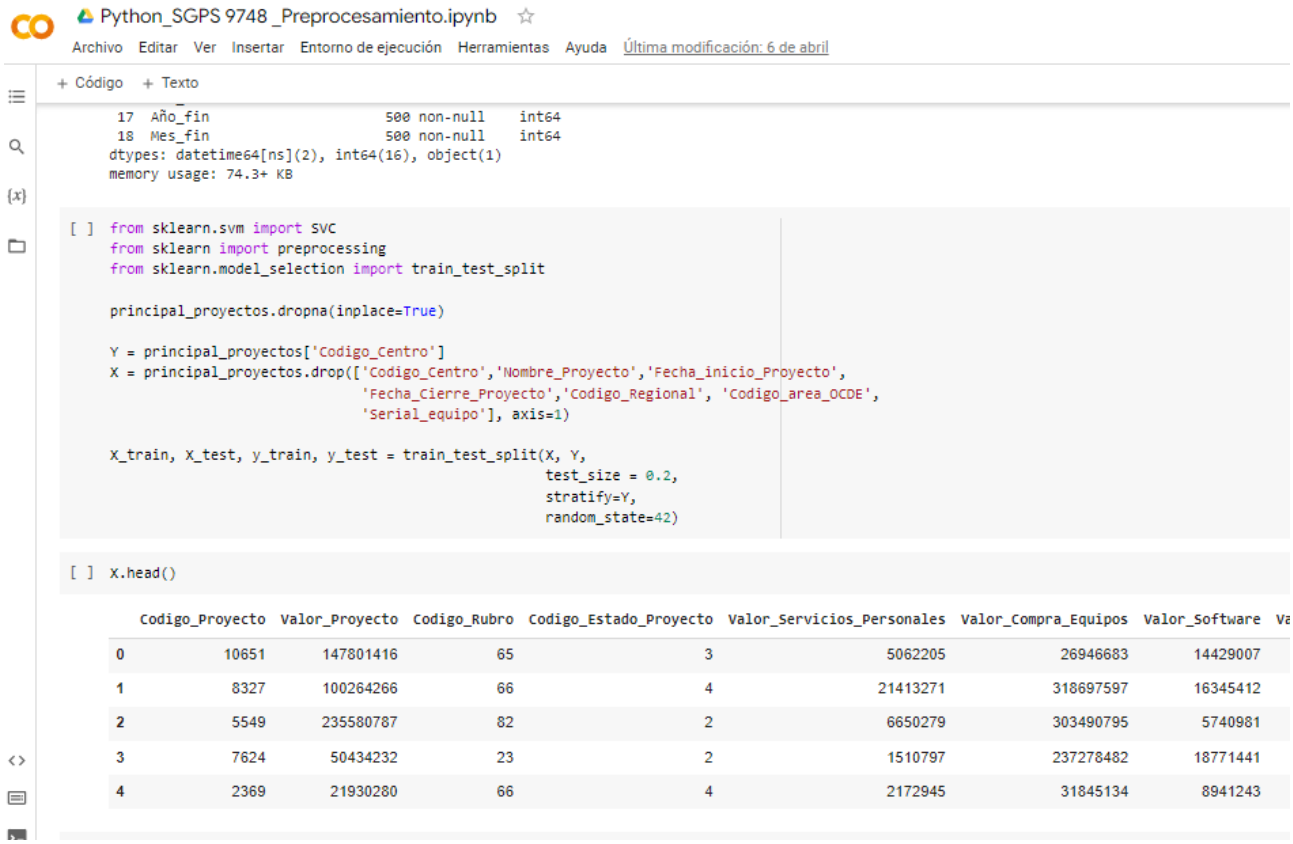
```
<class 'pandas.core.frame.DataFrame'>
Index: 0 entries
Data columns (total 2 columns):
 #   Column               Non-Null Count  Dtype 
---  ------               --------------  ----- 
 0   _Codigo_Rol_Proyecto  0 non-null      object
 1   _Nombre_Rol          0 non-null      object
dtypes: object(2)
memory usage: 0.0+ bytes
```

[ ] `estado_contratacion.head()`

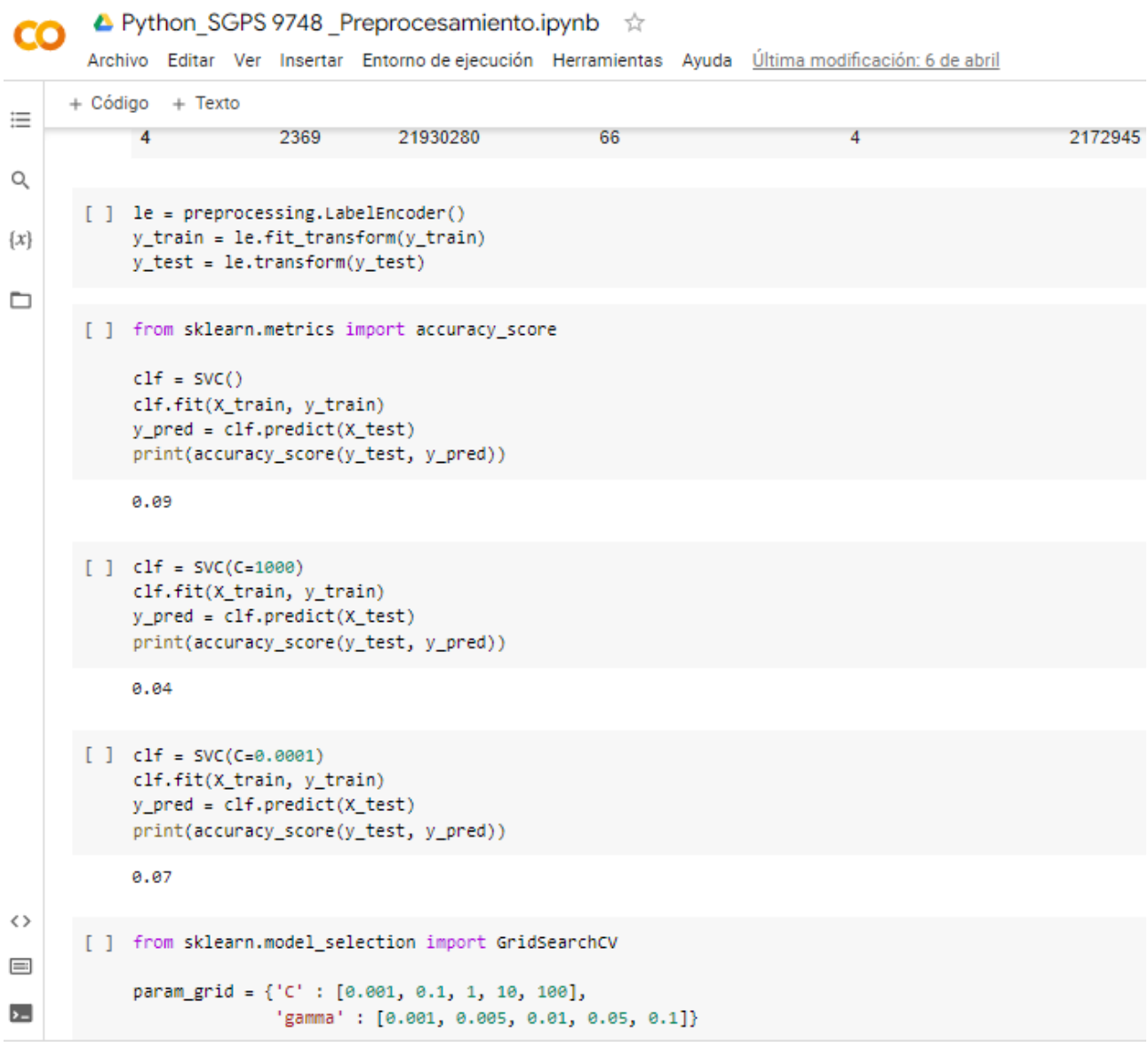

Python_SGPS 9748 _Preprocesamiento.ipynb
Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda Última modificación: 11 de marzo
+ Código + Texto
[ ] estado_contratacion.head()
_Codigo_Estado_Contratacion _Nombre_Estado_Contratacion
[ ] estado_contratacion.info()
<class 'pandas.core.frame.DataFrame'>
Index: 0 entries
Data columns (total 2 columns):
Column Non-Null Count Dtype
0 _Codigo_Estado_Contratacion 0 non-null object
1 _Nombre_Estado_Contratacion 0 non-null object
dtypes: object(2)
memory usage: 0.0+ bytes
[ ] principal_equipos.head()
_Serial_equipo _Nombre_del_equipo _Valor_equipo _Fecha_de_compra _Codigo_Proyecto _Codigo_Regional _Codigo_Centro _Codigo_Rubro _Codigo_Tipo_Equipo _Codigo_area_OCDE _Codigo_persona
[ ] principal_equipos.info()
<class 'pandas.core.frame.DataFrame'>
Index: 0 entries
Data columns (total 11 columns):
Column Non-Null Count Dtype
0 _Serial_equipo 0 non-null object
1 _Nombre_del_equipo 0 non-null object
2 _Valor_equipo 0 non-null object
3 _Fecha_de_compra 0 non-null object
4 _Codigo_Proyecto 0 non-null object
5 _Codigo_Regional 0 non-null object
6 _Codigo_Centro 0 non-null object
7 _Codigo_Rubro 0 non-null object
8 _Codigo_Tipo_Equipo 0 non-null object
9 _Codigo_area_OCDE 0 non-null object
10 _Codigo_persona 0 non-null object
dtypes: object(11)
memory usage: 0.0+ bytes

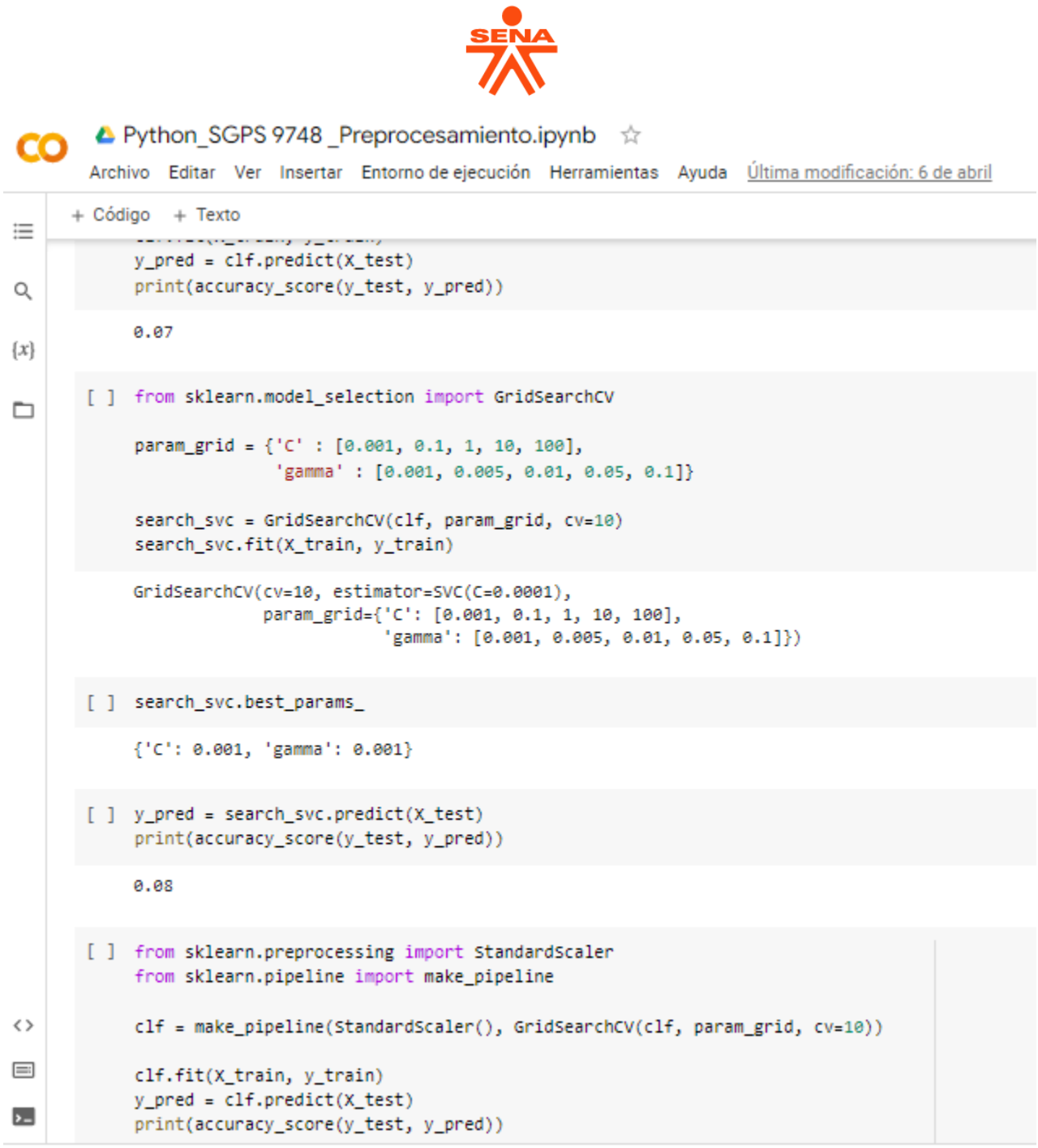

Python_SGPS 9748 _Preprocesamiento.ipynb
Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda Última modificación: 6 de abril
+ Código + Texto
Mes_fin
Exploración
Modelo Naive bayes
[ ] from sklearn import preprocessing
from sklearn.naive_bayes import GaussianNB
import numpy as np
[ ] le = preprocessing.LabelEncoder()
centro_encoded = le.fit_transform(principal_proyectos['Codigo_Centro'])
proyecto_encoded = le.fit_transform(principal_proyectos['Codigo_Proyecto'])
estado_proyecto_encoded = le.fit_transform(principal_proyectos['Codigo_Estado_Proyecto'])
rubro_encoded = le.fit_transform(principal_proyectos['Codigo_Rubro'])
le_centro_mapping = dict(zip(le.classes_, le.transform(le.classes_)))
le_proyecto_mapping = dict(zip(le.classes_, le.transform(le.classes_)))
le_estado_proyecto_mapping = dict(zip(le.classes_, le.transform(le.classes_)))
le_rubro_mapping = dict(zip(le.classes_, le.transform(le.classes_)))
#Matriz de datos
#print(centro_encoded)
#print(proyecto_encoded)
#print(estado_proyecto_encoded)
#print(rubro_encoded)
#print(le_centro_mapping)
#print(le_proyecto_mapping)
#print(le_estado_proyecto_mapping)
#print(le_rubro_mapping)
features = list(zip(centro_encoded, proyecto_encoded, estado_proyecto_encoded, rubro_encoded))
Python_SGPS 9748 _Preprocesamiento.ipynb
Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda Última modificación: 6 de abril
+ Código + Texto
#print(proyecto_encoded)
#print(estado_proyecto_encoded)
#print(rubro_encoded)
#print(le_centro_mapping)
#print(le_proyecto_mapping)
#print(le_estado_proyecto_mapping)
#print(le_rubro_mapping)
features = list(zip(centro_encoded, proyecto_encoded, estado_proyecto_encoded, rubro_encoded))
#print(features)
model = GaussianNB()
model.fit(list(features), list(centro_encoded))
# condicional de classificación ################################## REVISAR ##
new = ['Superior','Alto','Medio','Bajo']
new = [1,1,0, 0]
predicted= model.predict([new])
print("Predicted Value:", predicted)
Predicted Value: [1]
SVM
[ ] principal_proyectos.head(3)
Codigo_Proyecto Nombre_Proyecto Valor_Proyecto Codigo_Rubro Fecha_inicio_Proyecto Fecha_Cierre_Proyecto Codigo_Regional Codigo_Centro Codig
0 10651 Corporación de tecnología Carahsoft 147801416 65 2010-04-30 2017-11-06 5 9501
1 8327 Tecnología 100264266 66 2012-02-13 2022-05-04 5 9503

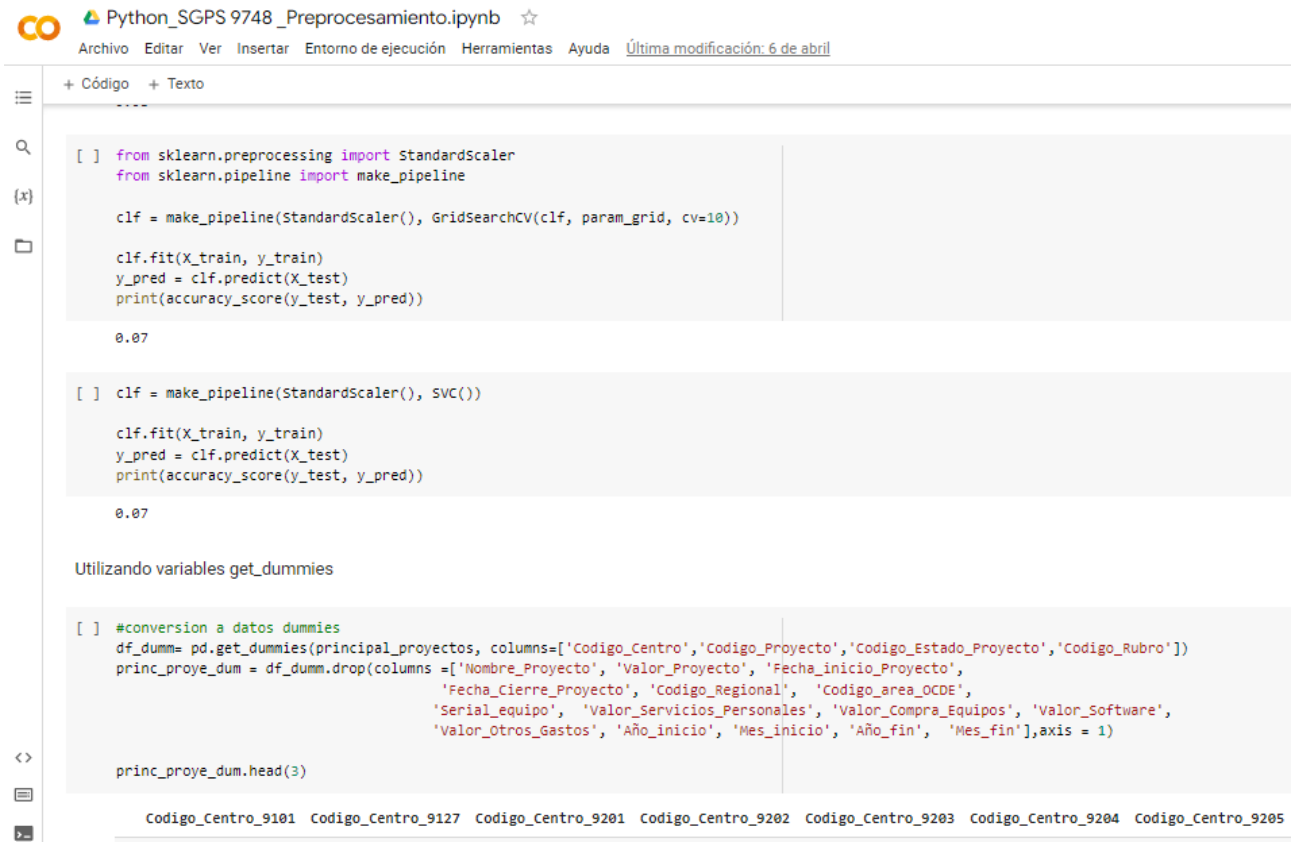

Python_SGPS 9748 _Preprocesamiento.ipynb
Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda Última modificación: 6 de abril
+ Código + Texto
print("Predicted Value:", predicted)
Predicted Value: [1]
SVM
[ ] principal_proyectos.head(3)
Codigo_Proyecto Nombre_Proyecto Valor_Proyecto Codigo_Rubro Fecha_inicio_Proyecto Fecha_Cierre_Proyecto Codigo_Regional
0 10651 Corporación de tecnología Carahsoft 147801416 65 2010-04-30 2017-11-06 5
1 8327 Tecnología SilkRoad 100264266 66 2012-02-13 2022-05-04 5
2 5549 Corporación de tecnología Carahsoft 235580787 82 2017-11-25 2021-07-04 5
[ ] principal_proyectos.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 500 entries, 0 to 499
Data columns (total 19 columns):
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]

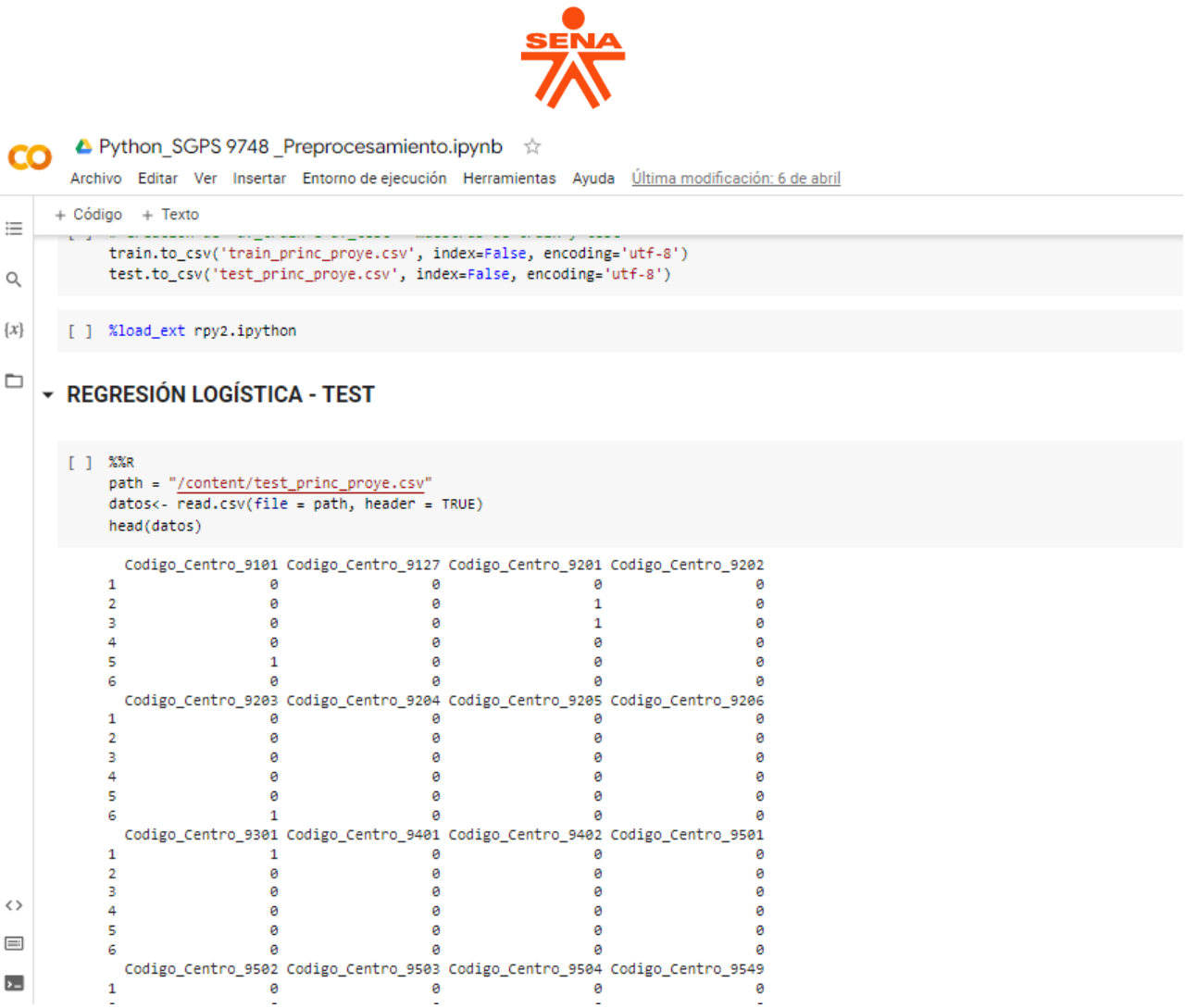


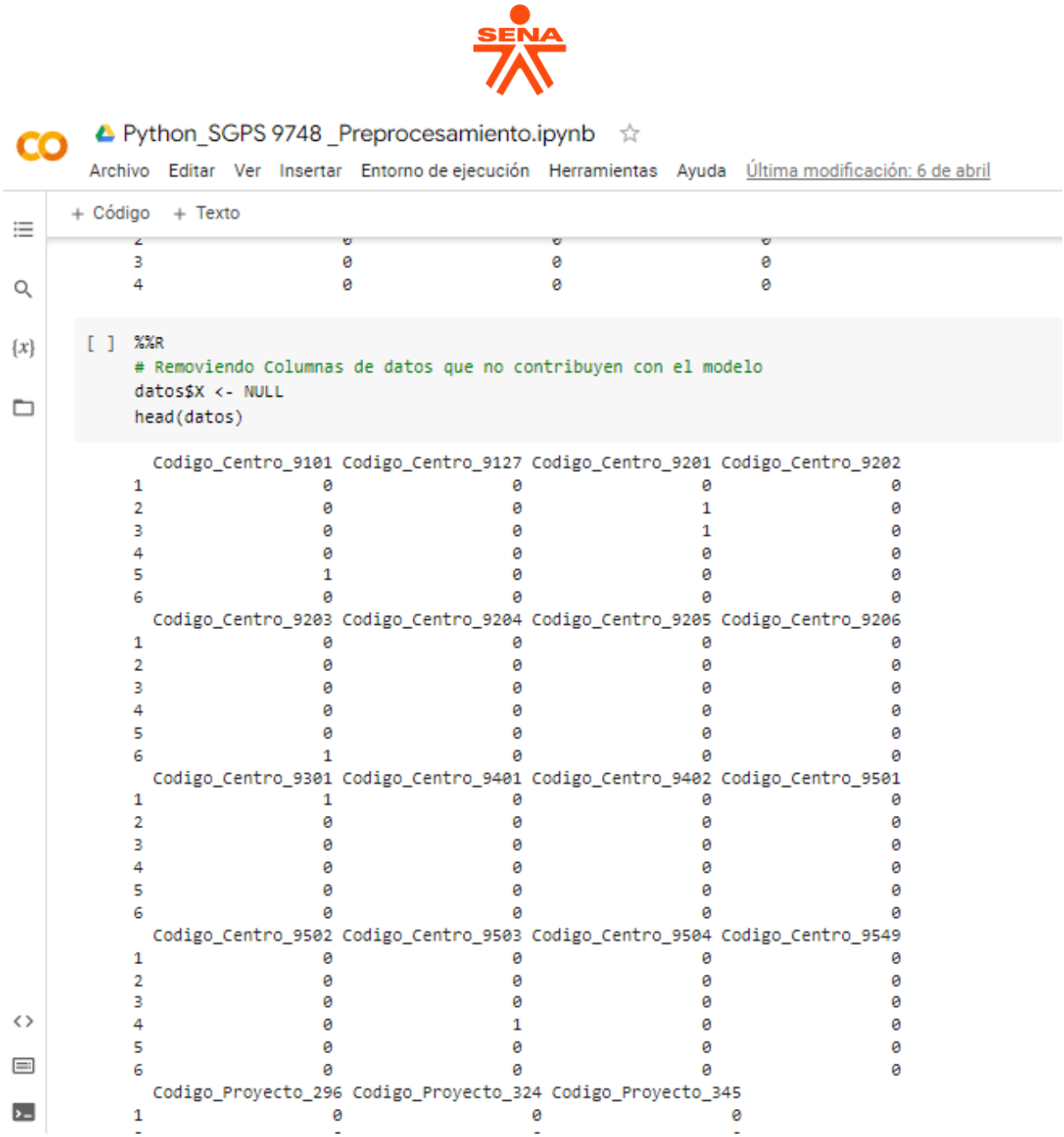
	Codigo_Proyecto	Valor_Proyecto	Codigo_Rubro	Codigo_Estado_Proyecto	Valor_Servicios_Personales	Valor_Compra_Equipos	Valor_Software	Va
0	10651	147801416	65	3	5062205	26946683	14429007	
1	8327	100264266	66	4	21413271	318697597	16345412	
2	5549	235580787	82	2	6650279	303490795	5740981	
3	7624	50434232	23	2	1510797	237278482	18771441	
4	2369	21930280	66	4	2172945	31845134	8941243	

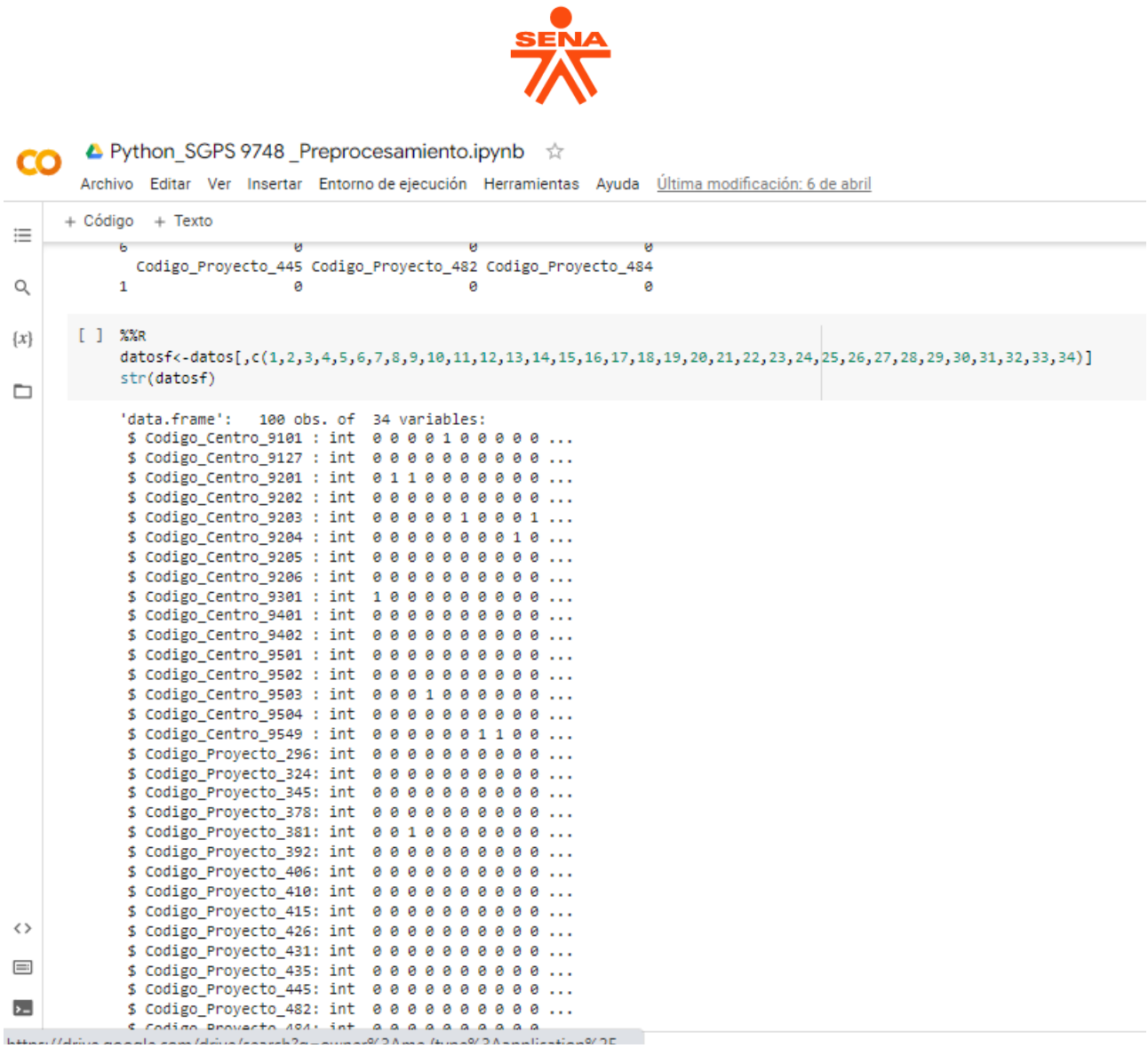

Python_SGPS 9748 _Preprocesamiento.ipynb
Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda Última modificación: 6 de abril
+ Código + Texto
4 2369 21930280 66 4 2172945
[ ] le = preprocessing.LabelEncoder()
y_train = le.fit_transform(y_train)
y_test = le.transform(y_test)
[ ] from sklearn.metrics import accuracy_score
clf = SVC()
clf.fit(X_train, y_train)
y_pred = clf.predict(X_test)
print(accuracy_score(y_test, y_pred))
0.09
[ ] clf = SVC(C=1000)
clf.fit(X_train, y_train)
y_pred = clf.predict(X_test)
print(accuracy_score(y_test, y_pred))
0.04
[ ] clf = SVC(C=0.0001)
clf.fit(X_train, y_train)
y_pred = clf.predict(X_test)
print(accuracy_score(y_test, y_pred))
0.07
[ ] from sklearn.model_selection import GridSearchCV
param_grid = {'C' : [0.001, 0.1, 1, 10, 100],
'gamma' : [0.001, 0.005, 0.01, 0.05, 0.1]}

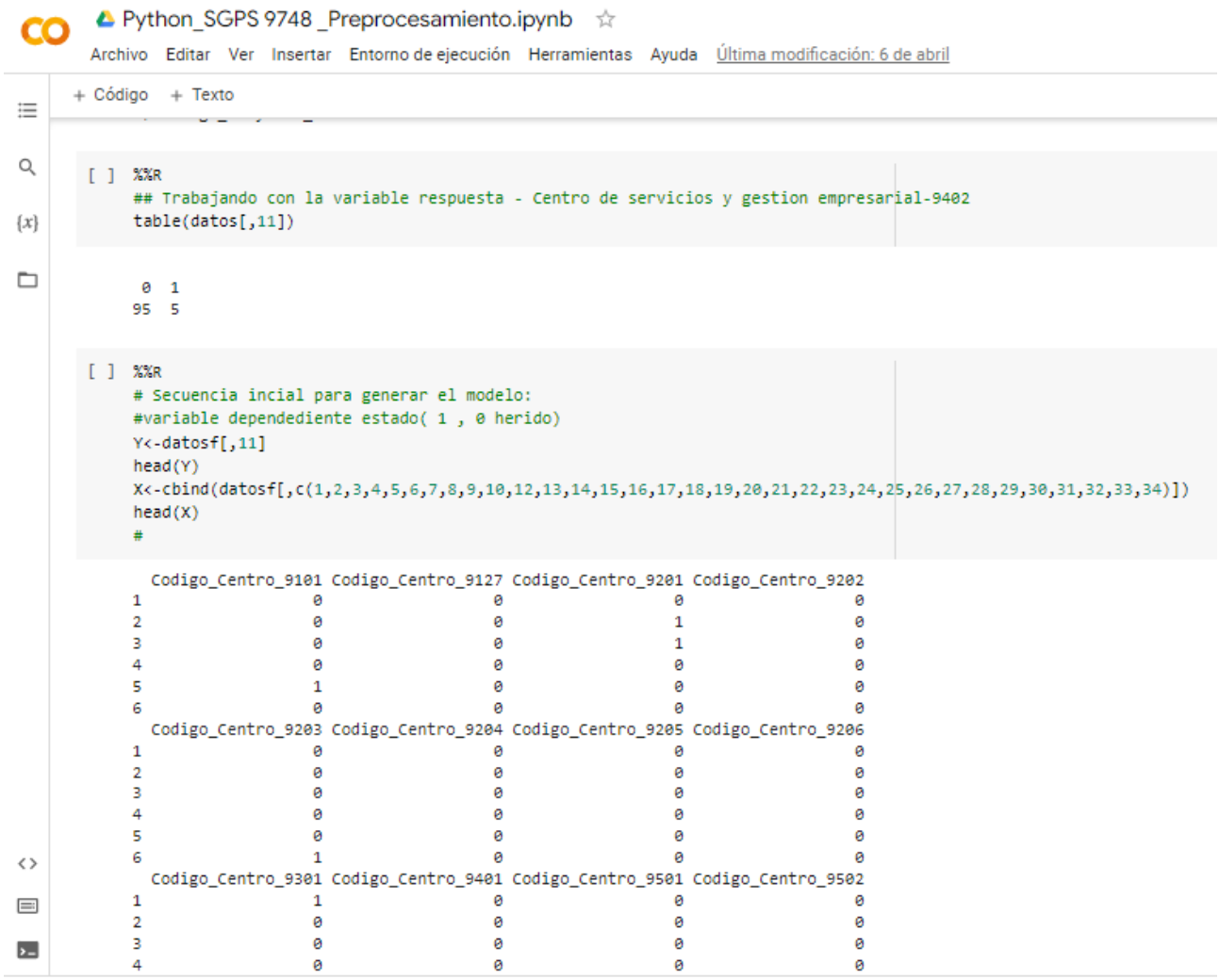

Python_SGPS 9748 _Preprocesamiento.ipynb
Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda Última modificación: 6 de abril
+ Código + Texto
y_pred = clf.predict(X_test)
print(accuracy_score(y_test, y_pred))
0.07
[ ] from sklearn.model_selection import GridSearchCV
param_grid = {'C' : [0.001, 0.1, 1, 10, 100],
'gamma' : [0.001, 0.005, 0.01, 0.05, 0.1]}
search_svc = GridSearchCV(clf, param_grid, cv=10)
search_svc.fit(X_train, y_train)
GridSearchCV(cv=10, estimator=SVC(C=0.0001),
param_grid={'C': [0.001, 0.1, 1, 10, 100],
'gamma': [0.001, 0.005, 0.01, 0.05, 0.1]})
[ ] search_svc.best_params_
{'C': 0.001, 'gamma': 0.001}
[ ] y_pred = search_svc.predict(X_test)
print(accuracy_score(y_test, y_pred))
0.08
[ ] from sklearn.preprocessing import StandardScaler
from sklearn.pipeline import make_pipeline
clf = make_pipeline(StandardScaler(), GridSearchCV(clf, param_grid, cv=10))
clf.fit(X_train, y_train)
y_pred = clf.predict(X_test)
print(accuracy_score(y_test, y_pred))

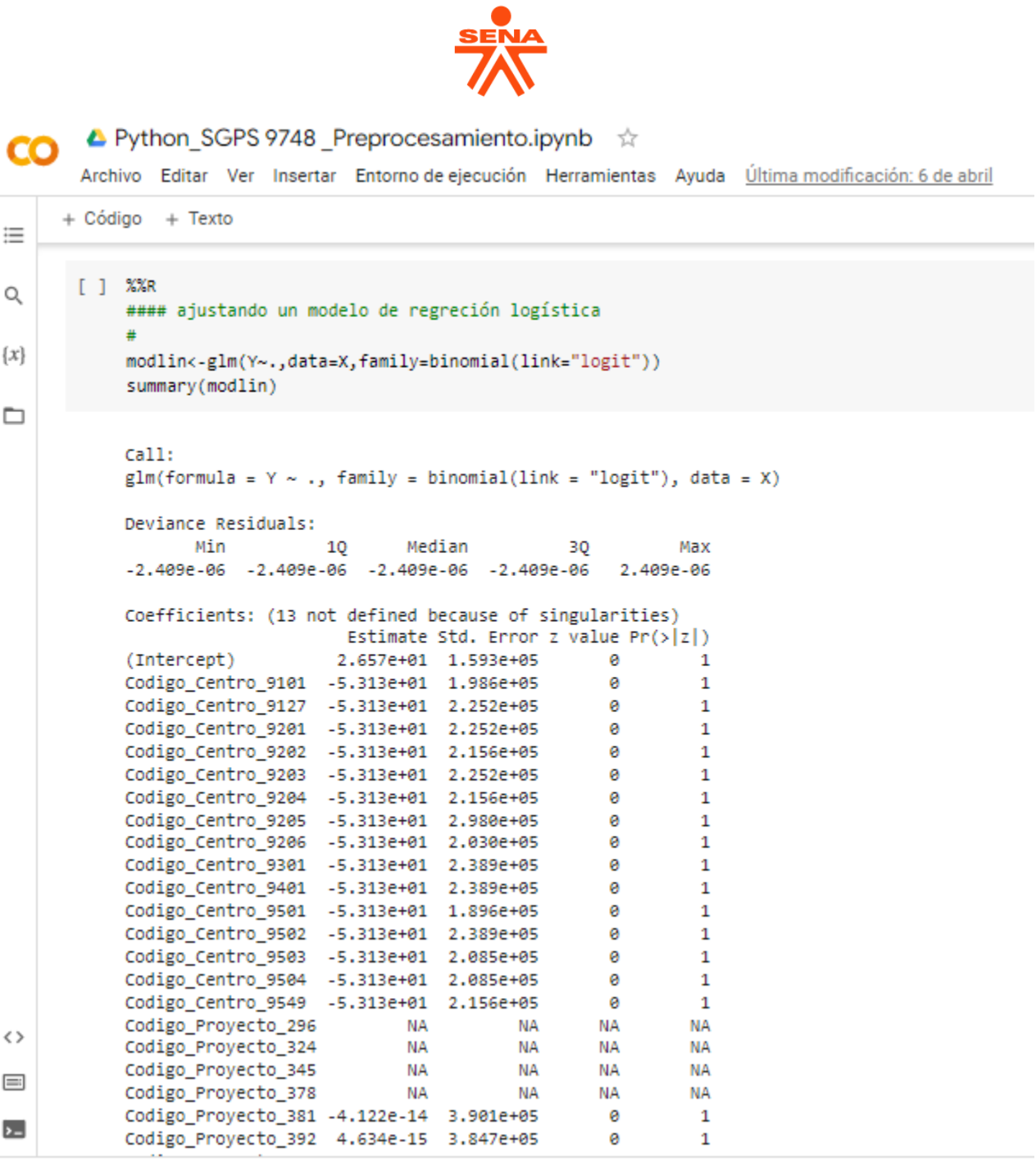

Python_SGPS 9748 _Preprocesamiento.ipynb
Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda Última modificación: 6 de abril
+ Código + Texto
[ ] from sklearn.preprocessing import StandardScaler
from sklearn.pipeline import make_pipeline
clf = make_pipeline(StandardScaler(), GridSearchCV(clf, param_grid, cv=10))
clf.fit(X_train, y_train)
y_pred = clf.predict(X_test)
print(accuracy_score(y_test, y_pred))
0.07
[ ] clf = make_pipeline(StandardScaler(), SVC())
clf.fit(X_train, y_train)
y_pred = clf.predict(X_test)
print(accuracy_score(y_test, y_pred))
0.07
Utilizando variables get_dummies
[ ] #conversion a datos dummies
df_dumm= pd.get_dummies(principal_proyectos, columns=['Codigo_Centro','Codigo_Proyecto','Codigo_Estado_Proyecto','Codigo_Rubro'])
princ_proye_dum = df_dumm.drop(columns =['Nombre_Proyecto', 'Valor_Proyecto', 'Fecha_inicio_Proyecto',
'Fecha_Cierre_Proyecto', 'Codigo_Regional', 'Codigo_area_OCDE',
'Serial_equipo', 'Valor_Servicios_Personales', 'Valor_Compra_Equipos', 'Valor_Software',
'Valor_Otros_Gastos', 'Año_inicio', 'Mes_inicio', 'Año_fin', 'Mes_fin'],axis = 1)
princ_proye_dum.head(3)
Codigo_Centro_9101 Codigo_Centro_9127 Codigo_Centro_9201 Codigo_Centro_9202 Codigo_Centro_9203 Codigo_Centro_9204 Codigo_Centro_9205


Python_SGPS 9748 _Preprocesamiento.ipynb
Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda Última modificación: 6 de abril
+ Código + Texto
train.to_csv('train_princ_proye.csv', index=False, encoding='utf-8')
test.to_csv('test_princ_proye.csv', index=False, encoding='utf-8')
[ ] %load_ext rpy2.ipython
REGRESIÓN LOGÍSTICA - TEST
[ ] %%R
path = "/content/test_princ_proye.csv"
datos<- read.csv(file = path, header = TRUE)
head(datos)
Codigo_Centro_9101 Codigo_Centro_9127 Codigo_Centro_9201 Codigo_Centro_9202
1 0 0 0 0
2 0 0 1 0
3 0 0 1 0
4 0 0 0 0
5 1 0 0 0
6 0 0 0 0
Codigo_Centro_9203 Codigo_Centro_9204 Codigo_Centro_9205 Codigo_Centro_9206
1 0 0 0 0
2 0 0 0 0
3 0 0 0 0
4 0 0 0 0
5 0 0 0 0
6 1 0 0 0
Codigo_Centro_9301 Codigo_Centro_9401 Codigo_Centro_9402 Codigo_Centro_9501
1 1 0 0 0
2 0 0 0 0
3 0 0 0 0
4 0 0 0 0
5 0 0 0 0
6 0 0 0 0
Codigo_Centro_9502 Codigo_Centro_9503 Codigo_Centro_9504 Codigo_Centro_9549
1 0 0 0 0


Python_SGPS 9748 _Preprocesamiento.ipynb
Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda Última modificación: 6 de abril
+ Código + Texto
3 0 0 0
4 0 0 0
[ ] %%R
Removiendo Columnas de datos que no contribuyen con el modelo
datos$X <- NULL
head(datos)
Codigo_Centro_9101 Codigo_Centro_9127 Codigo_Centro_9201 Codigo_Centro_9202
1 0 0 0 0
2 0 0 1 0
3 0 0 1 0
4 0 0 0 0
5 1 0 0 0
6 0 0 0 0
Codigo_Centro_9203 Codigo_Centro_9204 Codigo_Centro_9205 Codigo_Centro_9206
1 0 0 0 0
2 0 0 0 0
3 0 0 0 0
4 0 0 0 0
5 0 0 0 0
6 1 0 0 0
Codigo_Centro_9301 Codigo_Centro_9401 Codigo_Centro_9402 Codigo_Centro_9501
1 1 0 0 0
2 0 0 0 0
3 0 0 0 0
4 0 0 0 0
5 0 0 0 0
6 0 0 0 0
Codigo_Centro_9502 Codigo_Centro_9503 Codigo_Centro_9504 Codigo_Centro_9549
1 0 0 0 0
2 0 0 0 0
3 0 0 0 0
4 0 1 0 0
5 0 0 0 0
6 0 0 0 0
Codigo_Proyecto_296 Codigo_Proyecto_324 Codigo_Proyecto_345
1 0 0 0


Python_SGPS 9748 _Preprocesamiento.ipynb
Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda Última modificación: 6 de abril
+ Código + Texto
6 0 0 0
Codigo_Proyecto_445 Codigo_Proyecto_482 Codigo_Proyecto_484
1 0 0 0
[ ] %%R
datosf<-datos[,c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32,33,34)]
str(datosf)
'data.frame': 100 obs. of 34 variables:
$ Codigo_Centro_9101 : int 0 0 0 0 1 0 0 0 0 0 ...
$ Codigo_Centro_9127 : int 0 0 0 0 0 0 0 0 0 0 ...
$ Codigo_Centro_9201 : int 0 1 1 0 0 0 0 0 0 0 ...
$ Codigo_Centro_9202 : int 0 0 0 0 0 0 0 0 0 0 ...
$ Codigo_Centro_9203 : int 0 0 0 0 0 1 0 0 0 1 ...
$ Codigo_Centro_9204 : int 0 0 0 0 0 0 0 0 1 0 ...
$ Codigo_Centro_9205 : int 0 0 0 0 0 0 0 0 0 0 ...
$ Codigo_Centro_9206 : int 0 0 0 0 0 0 0 0 0 0 ...
$ Codigo_Centro_9301 : int 1 0 0 0 0 0 0 0 0 0 ...
$ Codigo_Centro_9401 : int 0 0 0 0 0 0 0 0 0 0 ...
$ Codigo_Centro_9402 : int 0 0 0 0 0 0 0 0 0 0 ...
$ Codigo_Centro_9501 : int 0 0 0 0 0 0 0 0 0 0 ...
$ Codigo_Centro_9502 : int 0 0 0 0 0 0 0 0 0 0 ...
$ Codigo_Centro_9503 : int 0 0 0 1 0 0 0 0 0 0 ...
$ Codigo_Centro_9504 : int 0 0 0 0 0 0 0 0 0 0 ...
$ Codigo_Centro_9549 : int 0 0 0 0 0 0 1 1 0 0 ...
$ Codigo_Proyecto_296: int 0 0 0 0 0 0 0 0 0 0 ...
$ Codigo_Proyecto_324: int 0 0 0 0 0 0 0 0 0 0 ...
$ Codigo_Proyecto_345: int 0 0 0 0 0 0 0 0 0 0 ...
$ Codigo_Proyecto_378: int 0 0 0 0 0 0 0 0 0 0 ...
$ Codigo_Proyecto_381: int 0 0 1 0 0 0 0 0 0 0 ...
$ Codigo_Proyecto_392: int 0 0 0 0 0 0 0 0 0 0 ...
$ Codigo_Proyecto_406: int 0 0 0 0 0 0 0 0 0 0 ...
$ Codigo_Proyecto_410: int 0 0 0 0 0 0 0 0 0 0 ...
$ Codigo_Proyecto_415: int 0 0 0 0 0 0 0 0 0 0 ...
$ Codigo_Proyecto_426: int 0 0 0 0 0 0 0 0 0 0 ...
$ Codigo_Proyecto_431: int 0 0 0 0 0 0 0 0 0 0 ...
$ Codigo_Proyecto_435: int 0 0 0 0 0 0 0 0 0 0 ...
$ Codigo_Proyecto_445: int 0 0 0 0 0 0 0 0 0 0 ...
$ Codigo_Proyecto_482: int 0 0 0 0 0 0 0 0 0 0 ...


Python_SGPS 9748 _Preprocesamiento.ipynb
Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda Última modificación: 6 de abril
+ Código + Texto
[ ] %%R
Trabajando con la variable respuesta - Centro de servicios y gestion empresarial-9402
table(datos[,11])
0 1
95 5
[ ] %%R
Secuencia incial para generar el modelo:
#variable dependediente estado(1 , 0 herido)
Y<-datosf[,11]
head(Y)
X<-cbind(datosf[,c(1,2,3,4,5,6,7,8,9,10,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32,33,34)])
head(X)
#
Codigo_Centro_9101 Codigo_Centro_9127 Codigo_Centro_9201 Codigo_Centro_9202
1 0 0 0 0
2 0 0 1 0
3 0 0 1 0
4 0 0 0 0
5 1 0 0 0
6 0 0 0 0
Codigo_Centro_9203 Codigo_Centro_9204 Codigo_Centro_9205 Codigo_Centro_9206
1 0 0 0 0
2 0 0 0 0
3 0 0 0 0
4 0 0 0 0
5 0 0 0 0
6 1 0 0 0
Codigo_Centro_9301 Codigo_Centro_9401 Codigo_Centro_9501 Codigo_Centro_9502
1 1 0 0 0
2 0 0 0 0
3 0 0 0 0
4 0 0 0 0


Python_SGPS 9748 _Preprocesamiento.ipynb
Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda Última modificación: 6 de abril
+ Código + Texto
[ ] %%R
ajustando un modelo de regreción logística
#
modlin<-glm(Y~.,data=X,family=binomial(link="logit"))
summary(modlin)
Call:
glm(formula = Y ~ ., family = binomial(link = "logit"), data = X)
Deviance Residuals:
Min 1Q Median 3Q Max
-2.409e-06 -2.409e-06 -2.409e-06 -2.409e-06 2.409e-06
Coefficients: (13 not defined because of singularities)
Estimate Std. Error z value Pr(>|z|)
(Intercept) 2.657e+01 1.593e+05 0 1
Codigo_Centro_9101 -5.313e+01 1.986e+05 0 1
Codigo_Centro_9127 -5.313e+01 2.252e+05 0 1
Codigo_Centro_9201 -5.313e+01 2.252e+05 0 1
Codigo_Centro_9202 -5.313e+01 2.156e+05 0 1
Codigo_Centro_9203 -5.313e+01 2.252e+05 0 1
Codigo_Centro_9204 -5.313e+01 2.156e+05 0 1
Codigo_Centro_9205 -5.313e+01 2.980e+05 0 1
Codigo_Centro_9206 -5.313e+01 2.030e+05 0 1
Codigo_Centro_9301 -5.313e+01 2.389e+05 0 1
Codigo_Centro_9401 -5.313e+01 2.389e+05 0 1
Codigo_Centro_9501 -5.313e+01 1.896e+05 0 1
Codigo_Centro_9502 -5.313e+01 2.389e+05 0 1
Codigo_Centro_9503 -5.313e+01 2.085e+05 0 1
Codigo_Centro_9504 -5.313e+01 2.085e+05 0 1
Codigo_Centro_9549 -5.313e+01 2.156e+05 0 1
Codigo_Proyecto_296 NA NA NA NA
Codigo_Proyecto_324 NA NA NA NA
Codigo_Proyecto_345 NA NA NA NA
Codigo_Proyecto_378 NA NA NA NA
Codigo_Proyecto_381 -4.122e-14 3.901e+05 0 1
Codigo_Proyecto_392 4.634e-15 3.847e+05 0 1

```
Codigo_Proyecto_345          NA          NA      NA       NA
Codigo_Proyecto_378          NA          NA      NA       NA
Codigo_Proyecto_381 -4.122e-14  3.901e+05       0        1
Codigo_Proyecto_392  4.634e-15  3.847e+05       0        1
Codigo_Proyecto_406          NA          NA      NA       NA
Codigo_Proyecto_410          NA          NA      NA       NA
Codigo_Proyecto_415          NA          NA      NA       NA
Codigo_Proyecto_426          NA          NA      NA       NA
Codigo_Proyecto_431  3.493e-14  3.754e+05       0        1
Codigo_Proyecto_435          NA          NA      NA       NA
Codigo_Proyecto_445          NA          NA      NA       NA
Codigo_Proyecto_482          NA          NA      NA       NA
Codigo_Proyecto_484 -1.666e-13  3.777e+05       0        1
Codigo_Proyecto_511          NA          NA      NA       NA
Codigo_Proyecto_514          NA          NA      NA       NA
Codigo_Proyecto_539 -6.081e-30  3.982e+05       0        1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3.9703e+01  on 99  degrees of freedom
Residual deviance: 5.8016e-10  on 79  degrees of freedom
AIC: 42

Number of Fisher Scoring iterations: 25
```

```
[ ] %%R
    summary(modlin)
    plot(modlin$fitted.values)
    plot(residuals(modlin,type='response')) # resíduos ordinários
```

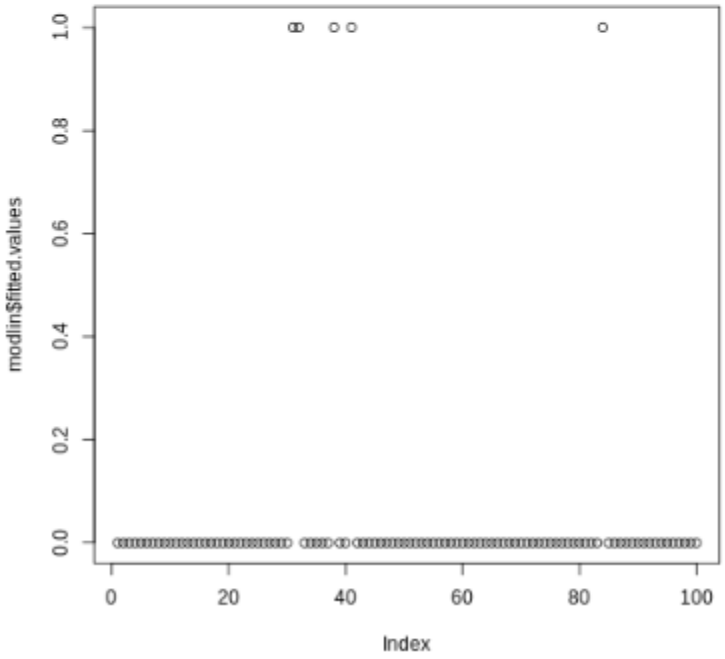
Python_SGPS 9748 _Preprocesamiento.ipynb ☆

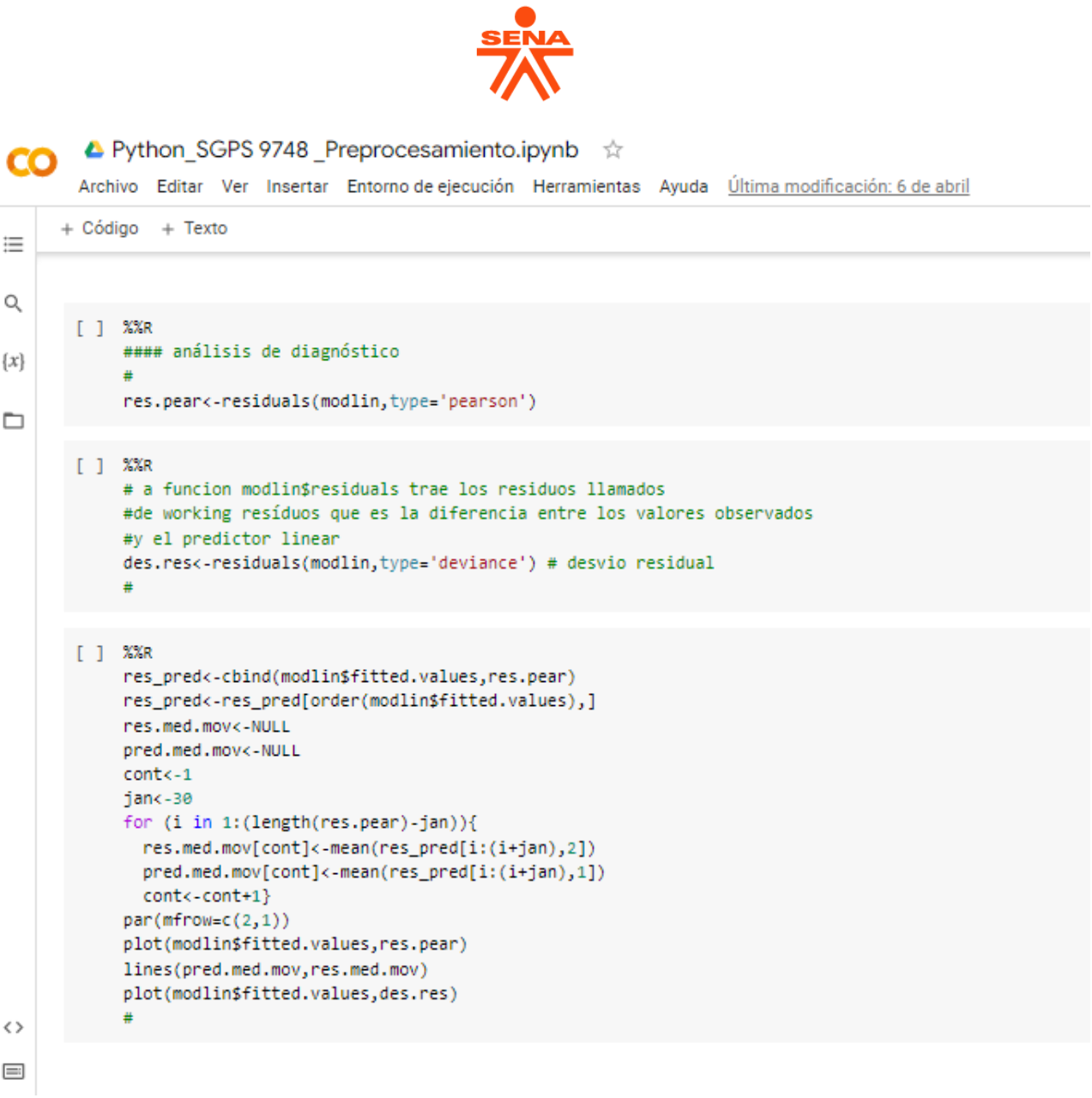
Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda Última modificación: 6 de abril

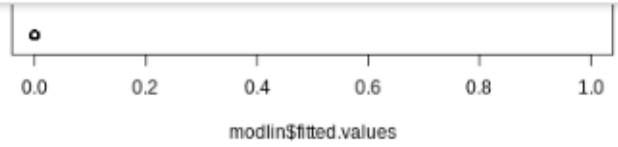
+ Código + Texto

```
Number of Fisher Scoring iterations: 25
```

```
[ ]  %%R
     summary(modlin)
     plot(modlin$fitted.values)
     plot(residuals(modlin,type='response')) # resíduos ordinários
```

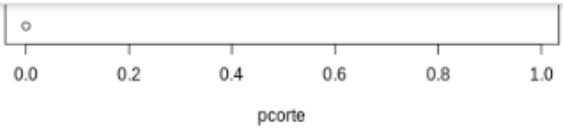



Python_SGPS 9748 _Preprocesamiento.ipynb
Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda Última modificación: 6 de abril
+ Código + Texto
[ ] %%R
análisis de diagnóstico
#
res.pear<-residuals(modlin,type='pearson')
[ ] %%R
a funcion modlin$residuals trae los residuos llamados
#de working resíduos que es la diferencia entre los valores observados
#y el predictor linear
des.res<-residuals(modlin,type='deviance') # desvio residual
#
[ ] %%R
res_pred<-cbind(modlin$fitted.values,res.pear)
res_pred<-res_pred[order(modlin$fitted.values),]
res.med.mov<-NULL
pred.med.mov<-NULL
cont<-1
jan<-30
for (i in 1:(length(res.pear)-jan)){
res.med.mov[cont]<-mean(res_pred[i:(i+jan),2])
pred.med.mov[cont]<-mean(res_pred[i:(i+jan),1])
cont<-cont+1}
par(mfrow=c(2,1))
plot(modlin$fitted.values,res.pear)
lines(pred.med.mov,res.med.mov)
plot(modlin$fitted.values,des.res)
#



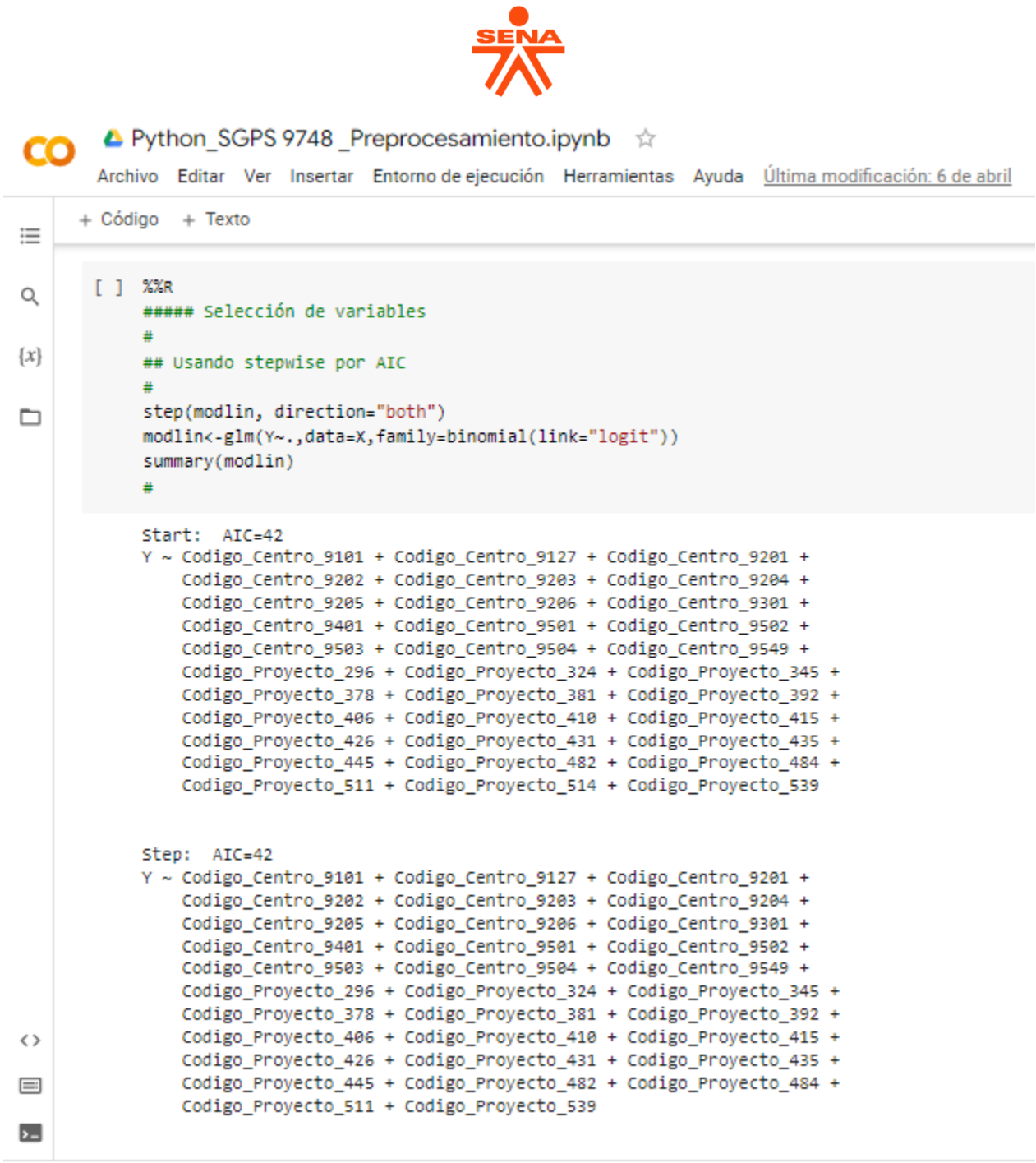

```
[ ]  %%R
     cor_clas<-NULL
     pcorte<-seq(from=min(modlin$fitted.values),to=max(modlin$fitted.values),by=0.005)
     for (i in 1:length(pcorte)){
       classif<-ifelse(modlin$fitted.values>pcorte[i],1,0)
       cor_clas[i]<-(table(Y,classif)[1,1]+table(Y,classif)[2,2])/nrow(X)*100}
     plot(pcorte,cor_clas)
     pcorte[cor_clas==max(cor_clas)]
     #
```

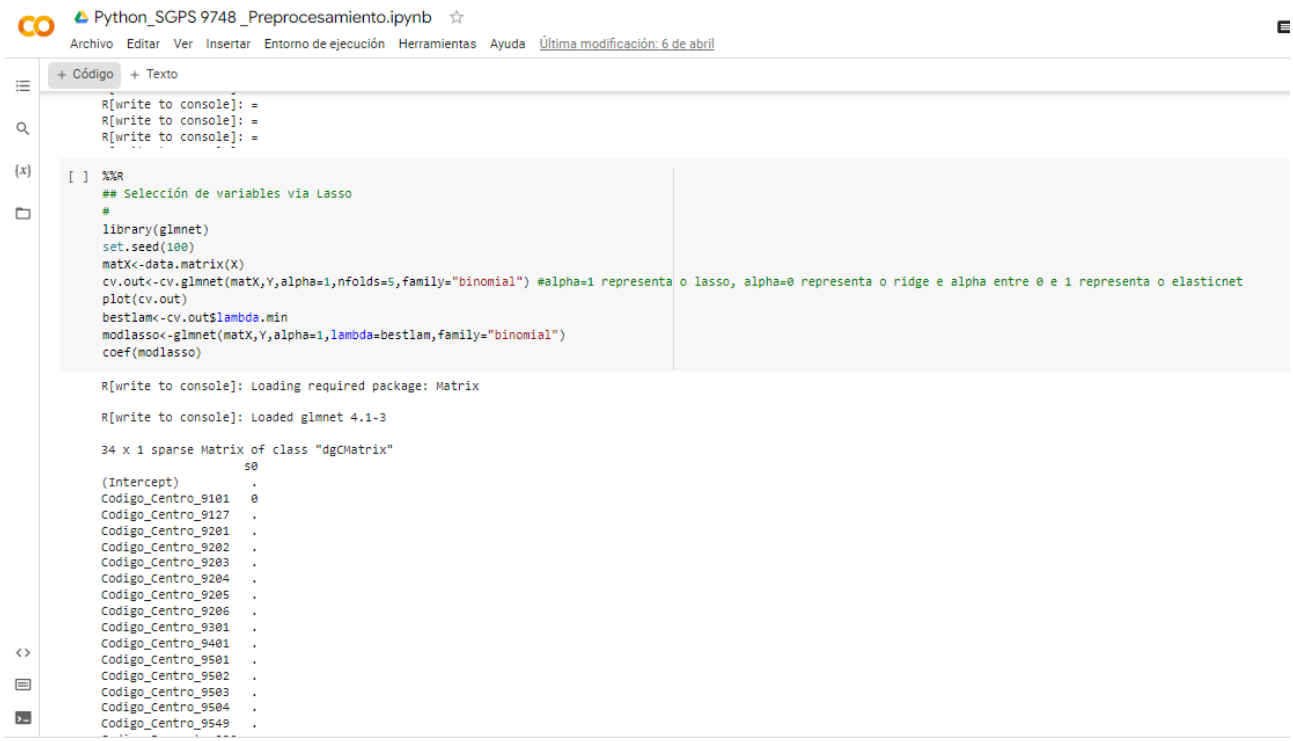
```
  [1] 0.005 0.010 0.015 0.020 0.025 0.030 0.035 0.040 0.045 0.050 0.055 0.060
 [13] 0.065 0.070 0.075 0.080 0.085 0.090 0.095 0.100 0.105 0.110 0.115 0.120
 [25] 0.125 0.130 0.135 0.140 0.145 0.150 0.155 0.160 0.165 0.170 0.175 0.180
 [37] 0.185 0.190 0.195 0.200 0.205 0.210 0.215 0.220 0.225 0.230 0.235 0.240
 [49] 0.245 0.250 0.255 0.260 0.265 0.270 0.275 0.280 0.285 0.290 0.295 0.300
 [61] 0.305 0.310 0.315 0.320 0.325 0.330 0.335 0.340 0.345 0.350 0.355 0.360
 [73] 0.365 0.370 0.375 0.380 0.385 0.390 0.395 0.400 0.405 0.410 0.415 0.420
 [85] 0.425 0.430 0.435 0.440 0.445 0.450 0.455 0.460 0.465 0.470 0.475 0.480
 [97] 0.485 0.490 0.495 0.500 0.505 0.510 0.515 0.520 0.525 0.530 0.535 0.540
[109] 0.545 0.550 0.555 0.560 0.565 0.570 0.575 0.580 0.585 0.590 0.595 0.600
[121] 0.605 0.610 0.615 0.620 0.625 0.630 0.635 0.640 0.645 0.650 0.655 0.660
[133] 0.665 0.670 0.675 0.680 0.685 0.690 0.695 0.700 0.705 0.710 0.715 0.720
[145] 0.725 0.730 0.735 0.740 0.745 0.750 0.755 0.760 0.765 0.770 0.775 0.780
[157] 0.785 0.790 0.795 0.800 0.805 0.810 0.815 0.820 0.825 0.830 0.835 0.840
[169] 0.845 0.850 0.855 0.860 0.865 0.870 0.875 0.880 0.885 0.890 0.895 0.900
[181] 0.905 0.910 0.915 0.920 0.925 0.930 0.935 0.940 0.945 0.950 0.955 0.960
[193] 0.965 0.970 0.975 0.980 0.985 0.990 0.995
```




```
[ ] %%R
    # dando costo 5 para clasificaciones equivocadas
    err_clas<-NULL
    pcorte<-seq(from=min(modlin$fitted.values),to=max(modlin$fitted.values),by=0.005)
    for (i in 1:length(pcorte)){
      classif<-ifelse(modlin$fitted.values>pcorte[i],1,0)
      err_clas[i]<-(table(Y,classif)[1,2]+(1/5)*table(Y,classif)[2,1])/((1/5)*sum(Y==0)+sum(Y==1))*100}
    plot(pcorte,err_clas)
    pcorte[err_clas==min(err_clas)]
    #
```

```
  [1] 0.005 0.010 0.015 0.020 0.025 0.030 0.035 0.040 0.045 0.050 0.055 0.060
 [13] 0.065 0.070 0.075 0.080 0.085 0.090 0.095 0.100 0.105 0.110 0.115 0.120
 [25] 0.125 0.130 0.135 0.140 0.145 0.150 0.155 0.160 0.165 0.170 0.175 0.180
 [37] 0.185 0.190 0.195 0.200 0.205 0.210 0.215 0.220 0.225 0.230 0.235 0.240
 [49] 0.245 0.250 0.255 0.260 0.265 0.270 0.275 0.280 0.285 0.290 0.295 0.300
 [61] 0.305 0.310 0.315 0.320 0.325 0.330 0.335 0.340 0.345 0.350 0.355 0.360
 [73] 0.365 0.370 0.375 0.380 0.385 0.390 0.395 0.400 0.405 0.410 0.415 0.420
 [85] 0.425 0.430 0.435 0.440 0.445 0.450 0.455 0.460 0.465 0.470 0.475 0.480
 [97] 0.485 0.490 0.495 0.500 0.505 0.510 0.515 0.520 0.525 0.530 0.535 0.540
[109] 0.545 0.550 0.555 0.560 0.565 0.570 0.575 0.580 0.585 0.590 0.595 0.600
[121] 0.605 0.610 0.615 0.620 0.625 0.630 0.635 0.640 0.645 0.650 0.655 0.660
[133] 0.665 0.670 0.675 0.680 0.685 0.690 0.695 0.700 0.705 0.710 0.715 0.720
[145] 0.725 0.730 0.735 0.740 0.745 0.750 0.755 0.760 0.765 0.770 0.775 0.780
[157] 0.785 0.790 0.795 0.800 0.805 0.810 0.815 0.820 0.825 0.830 0.835 0.840
[169] 0.845 0.850 0.855 0.860 0.865 0.870 0.875 0.880 0.885 0.890 0.895 0.900
[181] 0.905 0.910 0.915 0.920 0.925 0.930 0.935 0.940 0.945 0.950 0.955 0.960
[193] 0.965 0.970 0.975 0.980 0.985 0.990 0.995
```


Python_SGPS 9748 _Preprocesamiento.ipynb
Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda Última modificación: 6 de abril
+ Código + Texto
[ ] %%R
Selección de variables
#
Usando stepwise por AIC
#
step(modlin, direction="both")
modlin<-glm(Y~.,data=X,family=binomial(link="logit"))
summary(modlin)
#
Start: AIC=42
Y ~ Codigo_Centro_9101 + Codigo_Centro_9127 + Codigo_Centro_9201 +
Codigo_Centro_9202 + Codigo_Centro_9203 + Codigo_Centro_9204 +
Codigo_Centro_9205 + Codigo_Centro_9206 + Codigo_Centro_9301 +
Codigo_Centro_9401 + Codigo_Centro_9501 + Codigo_Centro_9502 +
Codigo_Centro_9503 + Codigo_Centro_9504 + Codigo_Centro_9549 +
Codigo_Proyecto_296 + Codigo_Proyecto_324 + Codigo_Proyecto_345 +
Codigo_Proyecto_378 + Codigo_Proyecto_381 + Codigo_Proyecto_392 +
Codigo_Proyecto_406 + Codigo_Proyecto_410 + Codigo_Proyecto_415 +
Codigo_Proyecto_426 + Codigo_Proyecto_431 + Codigo_Proyecto_435 +
Codigo_Proyecto_445 + Codigo_Proyecto_482 + Codigo_Proyecto_484 +
Codigo_Proyecto_511 + Codigo_Proyecto_514 + Codigo_Proyecto_539
Step: AIC=42
Y ~ Codigo_Centro_9101 + Codigo_Centro_9127 + Codigo_Centro_9201 +
Codigo_Centro_9202 + Codigo_Centro_9203 + Codigo_Centro_9204 +
Codigo_Centro_9205 + Codigo_Centro_9206 + Codigo_Centro_9301 +
Codigo_Centro_9401 + Codigo_Centro_9501 + Codigo_Centro_9502 +
Codigo_Centro_9503 + Codigo_Centro_9504 + Codigo_Centro_9549 +
Codigo_Proyecto_296 + Codigo_Proyecto_324 + Codigo_Proyecto_345 +
Codigo_Proyecto_378 + Codigo_Proyecto_381 + Codigo_Proyecto_392 +
Codigo_Proyecto_406 + Codigo_Proyecto_410 + Codigo_Proyecto_415 +
Codigo_Proyecto_426 + Codigo_Proyecto_431 + Codigo_Proyecto_435 +
Codigo_Proyecto_445 + Codigo_Proyecto_482 + Codigo_Proyecto_484 +
Codigo_Proyecto_511 + Codigo_Proyecto_539

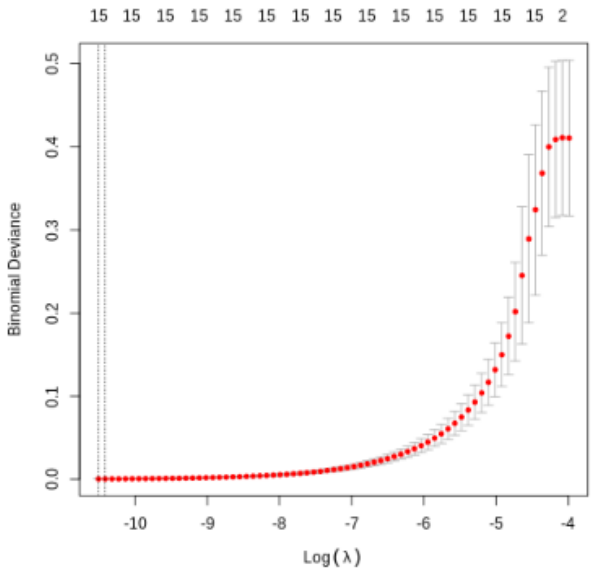

Python_SGPS 9748 _Preprocesamiento.ipynb
Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda Última modificación: 6 de abril
+ Código + Texto
R[write to console]: =
R[write to console]: =
R[write to console]: =
[ ] %%R
Selección de variables via Lasso
#
library(glmnet)
set.seed(100)
matX<-data.matrix(X)
cv.out<-cv.glmnet(matX,Y,alpha=1,nfolds=5,family="binomial") #alpha=1 representa o lasso, alpha=0 representa o ridge e alpha entre 0 e 1 representa o elasticnet
plot(cv.out)
bestlam<-cv.out$lambda.min
modlasso<-glmnet(matX,Y,alpha=1,lambda=bestlam,family="binomial")
coef(modlasso)
R[write to console]: Loading required package: Matrix
R[write to console]: Loaded glmnet 4.1-3
34 x 1 sparse Matrix of class "dgCMatrix"
s0
(Intercept) .
Codigo_Centro_9101 0
Codigo_Centro_9127 .
Codigo_Centro_9201 .
Codigo_Centro_9202 .
Codigo_Centro_9203 .
Codigo_Centro_9204 .
Codigo_Centro_9205 .
Codigo_Centro_9206 .
Codigo_Centro_9301 .
Codigo_Centro_9401 .
Codigo_Centro_9501 .
Codigo_Centro_9502 .
Codigo_Centro_9503 .
Codigo_Centro_9504 .
Codigo_Centro_9549 .

Python_SGPS 9748 _Preprocesamiento.ipynb ☆

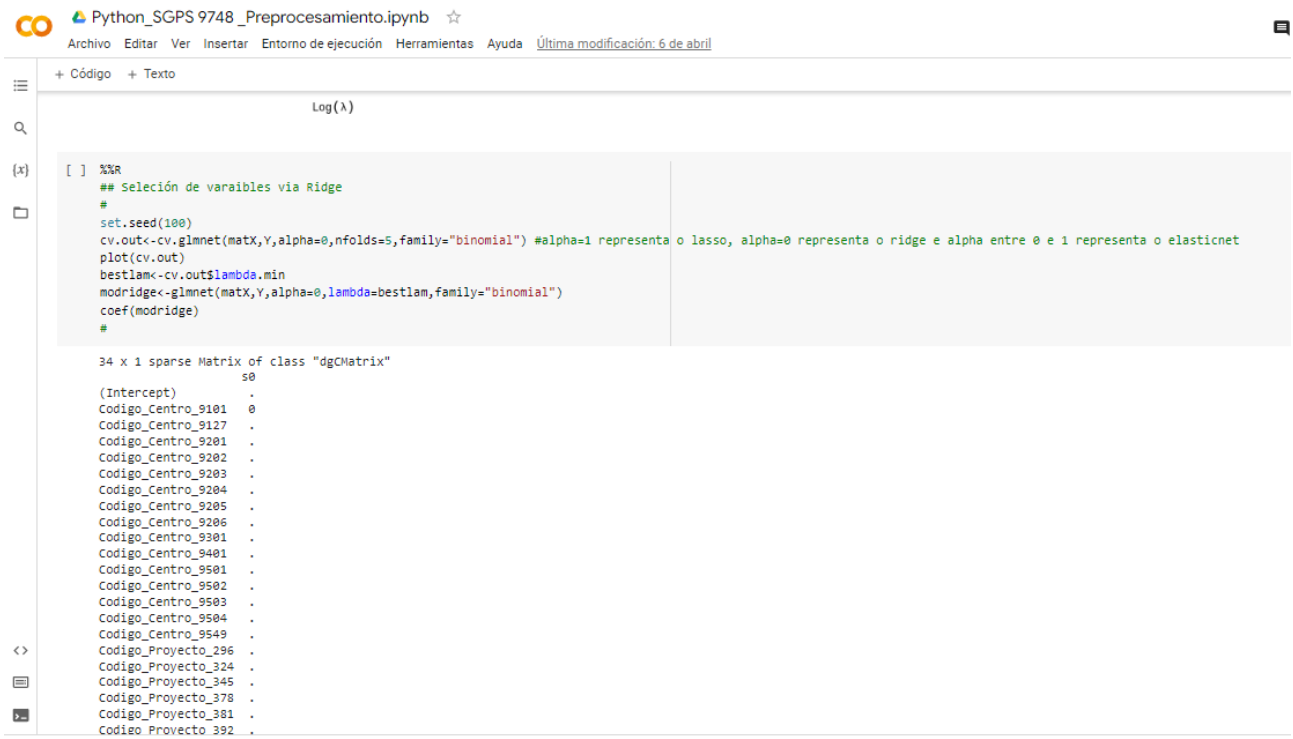
Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda Última modificación: 6 de abril

+ Código + Texto

```
Codigo_Proyecto_484  .
Codigo_Proyecto_511  .
Codigo_Proyecto_514  .
Codigo_Proyecto_539  .
```




```
[ ] %%R
    ## Seleción de varaibles via Ridge
    #
    set.seed(100)
    cv.out<-cv.glmnet(matX,Y,alpha=0,nfolds=5,family="binomial") #alpha=1 representa o lasso, alpha=0 representa o ridg
    plot(cv.out)
```

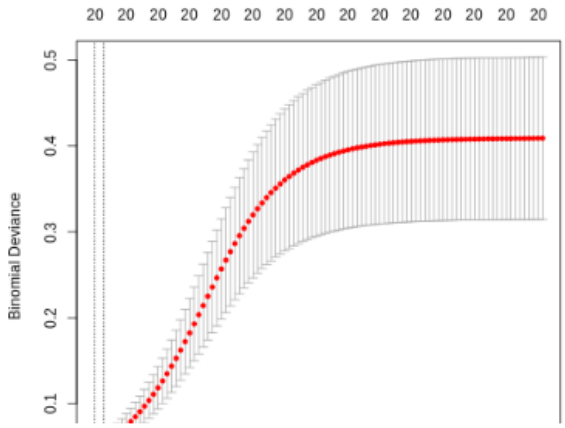

Python_SGPS 9748 _Preprocesamiento.ipynb
Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda Última modificación: 6 de abril
+ Código + Texto
Log(λ)
[ ] %%R
Selección de varaibles via Ridge
#
set.seed(100)
cv.out<-cv.glmnet(matX,Y,alpha=0,nfolds=5,family="binomial") #alpha=1 representa o lasso, alpha=0 representa o ridge e alpha entre 0 e 1 representa o elasticnet
plot(cv.out)
bestlam<-cv.out$lambda.min
modridge<-glmnet(matX,Y,alpha=0,lambda=bestlam,family="binomial")
coef(modridge)
#
34 x 1 sparse Matrix of class "dgCMatrix"
s0
(Intercept) .
Codigo_Centro_9101 0
Codigo_Centro_9127 .
Codigo_Centro_9201 .
Codigo_Centro_9202 .
Codigo_Centro_9203 .
Codigo_Centro_9204 .
Codigo_Centro_9205 .
Codigo_Centro_9206 .
Codigo_Centro_9301 .
Codigo_Centro_9401 .
Codigo_Centro_9501 .
Codigo_Centro_9502 .
Codigo_Centro_9503 .
Codigo_Centro_9504 .
Codigo_Centro_9549 .
Codigo_Proyecto_296 .
Codigo_Proyecto_324 .
Codigo_Proyecto_345 .
Codigo_Proyecto_378 .
Codigo_Proyecto_381 .
Codigo_Proyecto_392 .

Python_SGPS 9748 _Preprocesamiento.ipynb

Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda Última modificación: 6 de abril

+ Código + Texto

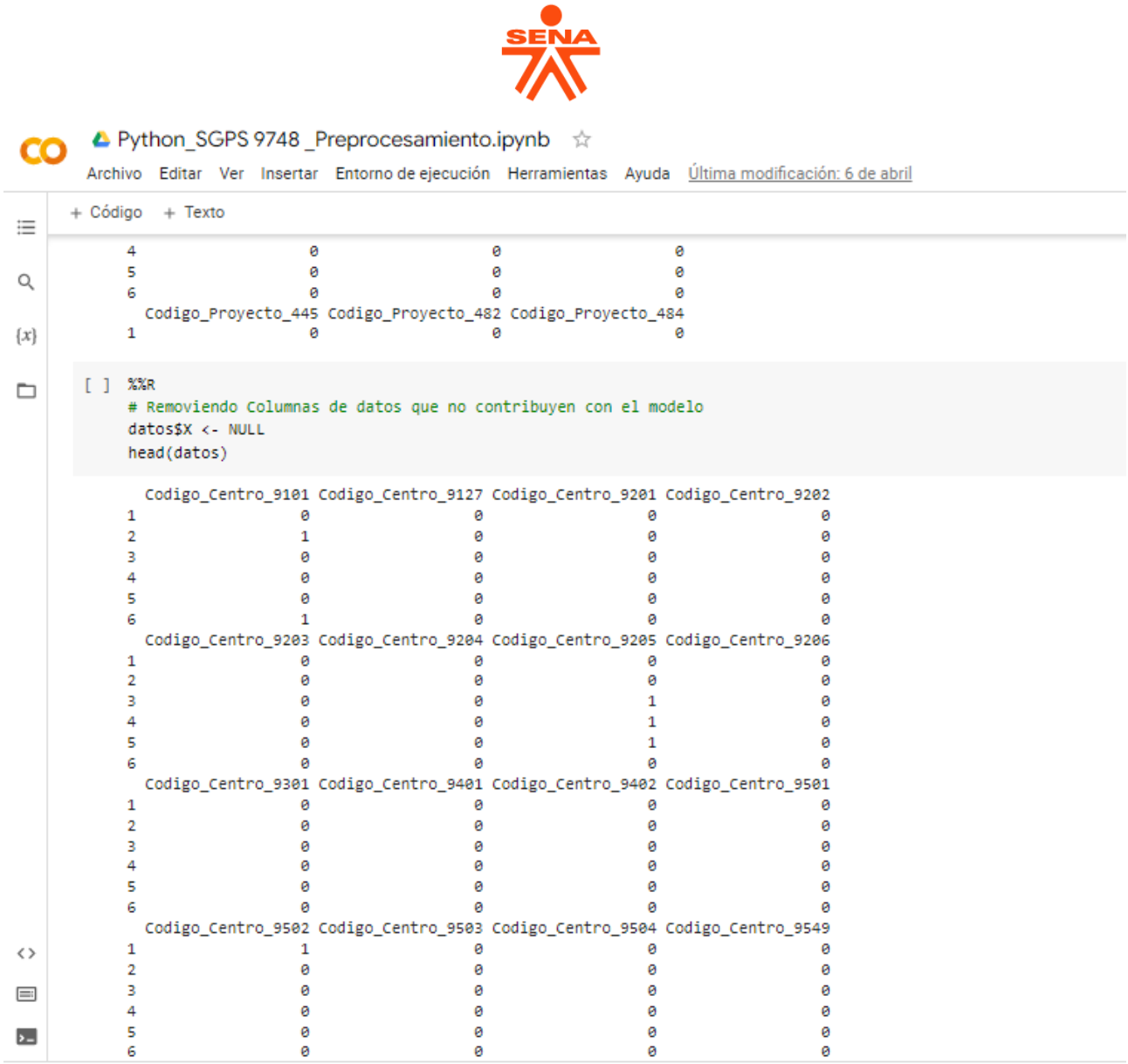
```
Codigo_Proyecto_514  .
Codigo_Proyecto_539  .
```

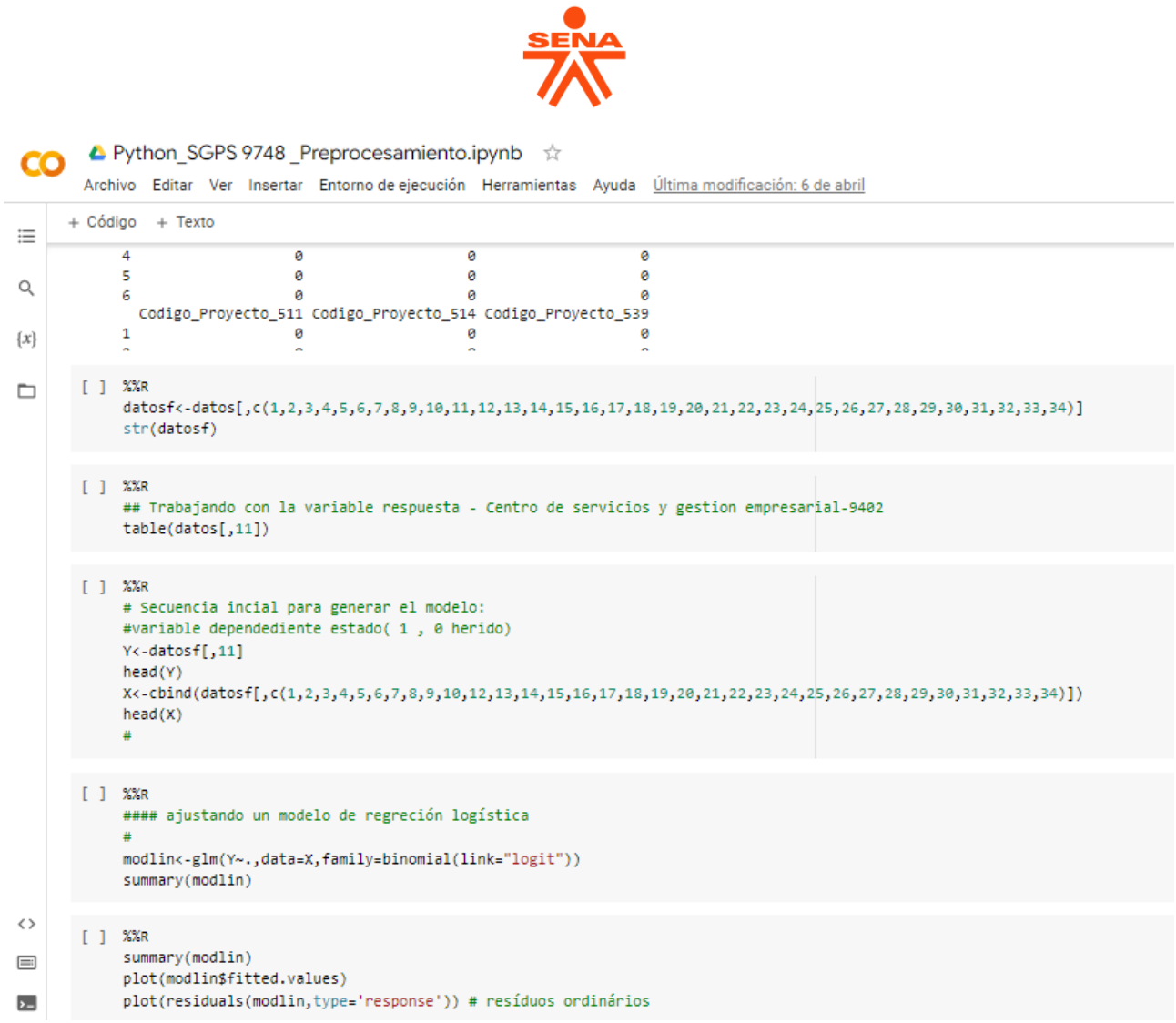


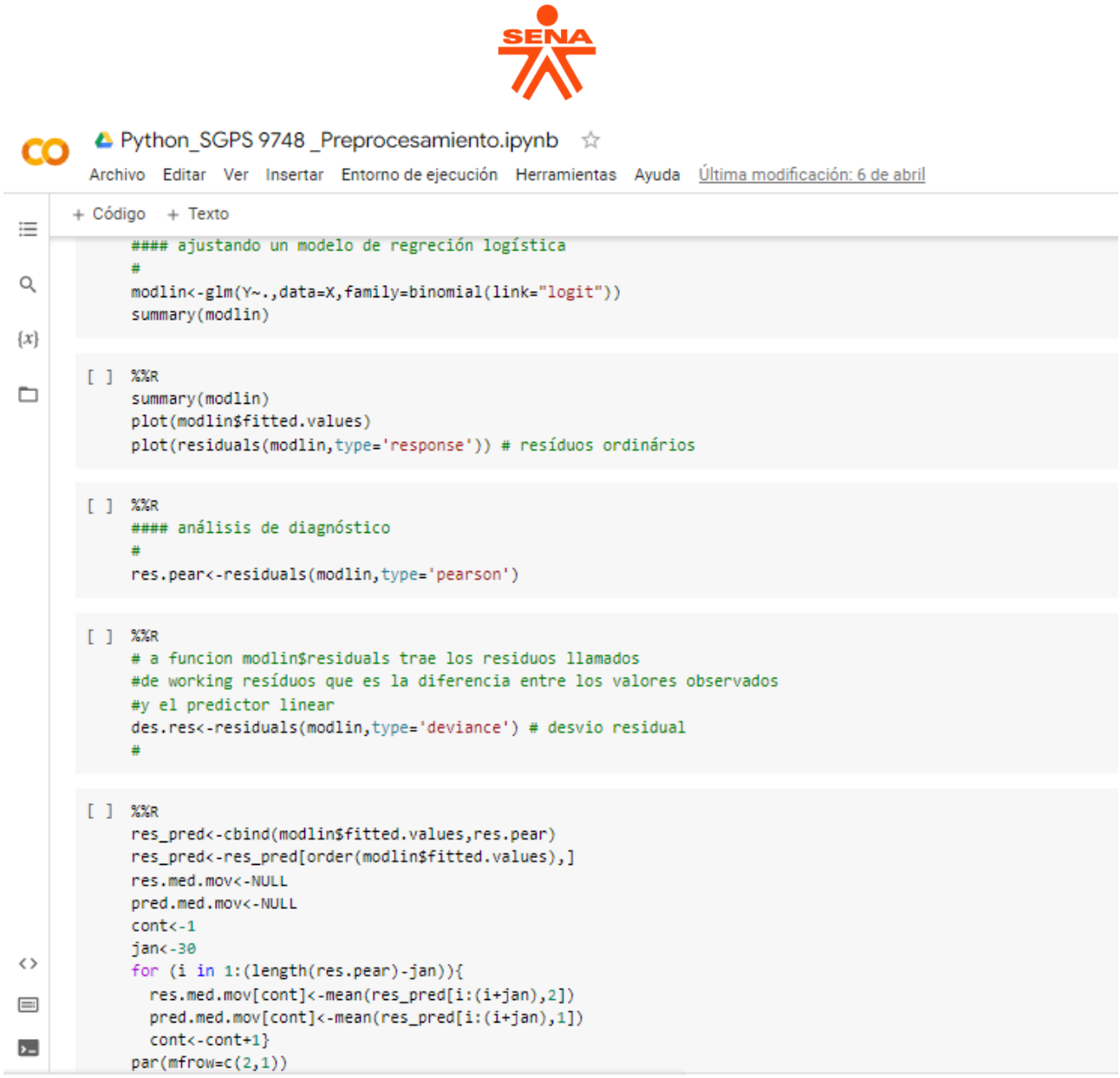
Train

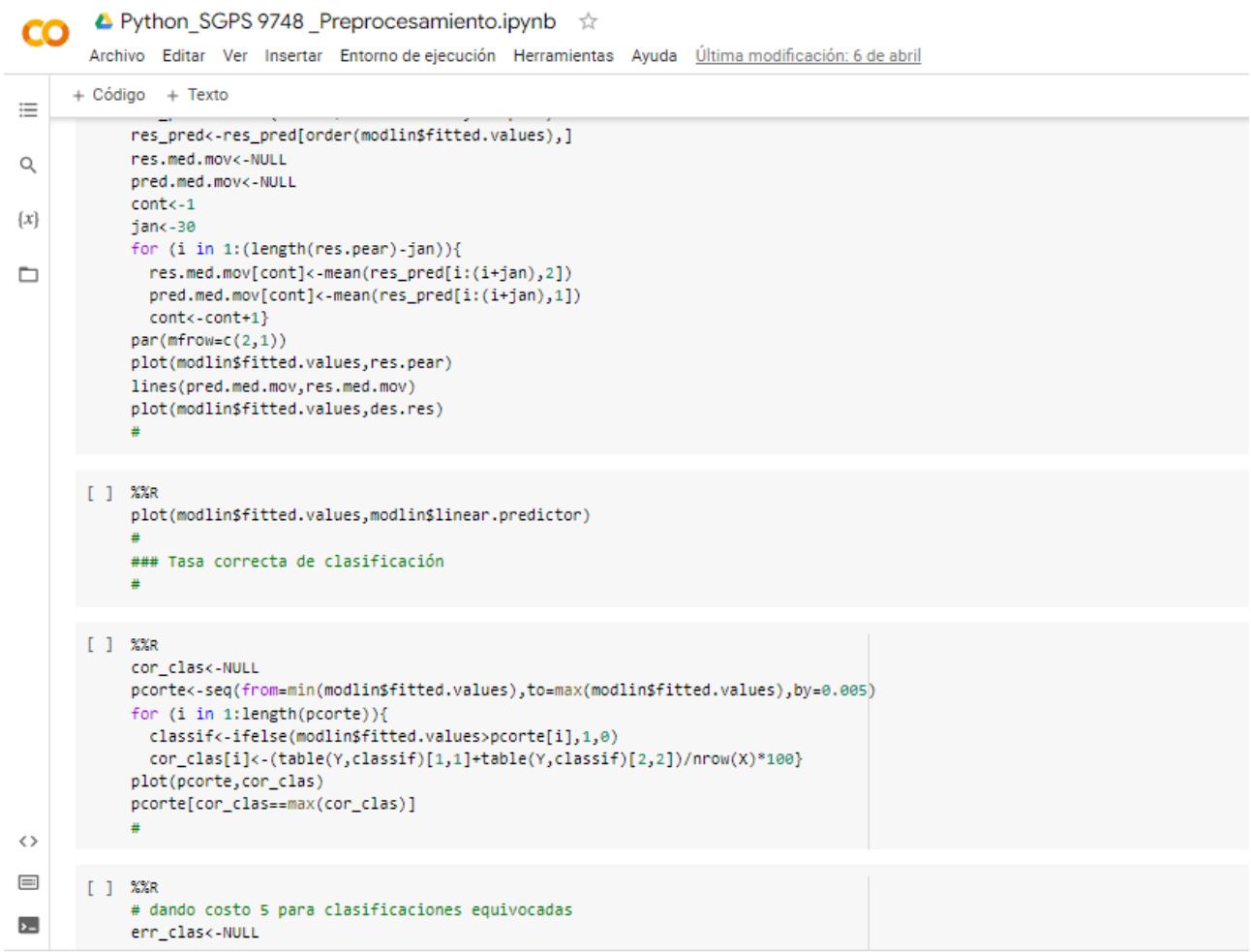
```
[ ] %%R
    path = "/content/train_princ_proye.csv"
    datos<- read.csv(file = path, header = TRUE)
    head(datos)
```

```
  Codigo_Centro_9101 Codigo_Centro_9127 Codigo_Centro_9201 Codigo_Centro_9202
1                  0                  0                  0                  0
2                  1                  0                  0                  0
3                  0                  0                  0                  0
4                  0                  0                  0                  0
5                  0                  0                  0                  0
```


Python_SGPS 9748 _Preprocesamiento.ipynb
Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda Última modificación: 6 de abril
+ Código + Texto
4 0 0 0
5 0 0 0
6 0 0 0
Codigo_Proyecto_445 Codigo_Proyecto_482 Codigo_Proyecto_484
1 0 0 0
[ ] %%R
Removiendo Columnas de datos que no contribuyen con el modelo
datos$X <- NULL
head(datos)
Codigo_Centro_9101 Codigo_Centro_9127 Codigo_Centro_9201 Codigo_Centro_9202
1 0 0 0 0
2 1 0 0 0
3 0 0 0 0
4 0 0 0 0
5 0 0 0 0
6 1 0 0 0
Codigo_Centro_9203 Codigo_Centro_9204 Codigo_Centro_9205 Codigo_Centro_9206
1 0 0 0 0
2 0 0 0 0
3 0 0 1 0
4 0 0 1 0
5 0 0 1 0
6 0 0 0 0
Codigo_Centro_9301 Codigo_Centro_9401 Codigo_Centro_9402 Codigo_Centro_9501
1 0 0 0 0
2 0 0 0 0
3 0 0 0 0
4 0 0 0 0
5 0 0 0 0
6 0 0 0 0
Codigo_Centro_9502 Codigo_Centro_9503 Codigo_Centro_9504 Codigo_Centro_9549
1 1 0 0 0
2 0 0 0 0
3 0 0 0 0
4 0 0 0 0
5 0 0 0 0
6 0 0 0 0


Python_SGPS 9748 _Preprocesamiento.ipynb
Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda Última modificación: 6 de abril
+ Código + Texto
4 0 0 0
5 0 0 0
6 0 0 0
Codigo_Proyecto_511 Codigo_Proyecto_514 Codigo_Proyecto_539
1 0 0 0
[ ] %%R
datosf<-datos[,c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32,33,34)]
str(datosf)
[ ] %%R
Trabajando con la variable respuesta - Centro de servicios y gestion empresarial-9402
table(datos[,11])
[ ] %%R
Secuencia incial para generar el modelo:
#variable dependediente estado(1 , 0 herido)
Y<-datosf[,11]
head(Y)
X<-cbind(datosf[,c(1,2,3,4,5,6,7,8,9,10,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32,33,34)])
head(X)
#
[ ] %%R
ajustando un modelo de regreción logística
#
modlin<-glm(Y~.,data=X,family=binomial(link="logit"))
summary(modlin)
[ ] %%R
summary(modlin)
plot(modlin$fitted.values)
plot(residuals(modlin,type='response')) # resíduos ordinários


Python_SGPS 9748 _Preprocesamiento.ipynb
Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda Última modificación: 6 de abril
+ Código + Texto
ajustando un modelo de regreción logística
#
modlin<-glm(Y~.,data=X,family=binomial(link="logit"))
summary(modlin)
[ ] %%R
summary(modlin)
plot(modlin$fitted.values)
plot(residuals(modlin,type='response')) # resíduos ordinários
[ ] %%R
análisis de diagnóstico
#
res.pear<-residuals(modlin,type='pearson')
[ ] %%R
a funcion modlin$residuals trae los residuos llamados
#de working resíduos que es la diferencia entre los valores observados
#y el predictor linear
des.res<-residuals(modlin,type='deviance') # desvio residual
#
[ ] %%R
res_pred<-cbind(modlin$fitted.values,res.pear)
res_pred<-res_pred[order(modlin$fitted.values),]
res.med.mov<-NULL
pred.med.mov<-NULL
cont<-1
jan<-30
for (i in 1:(length(res.pear)-jan)){
res.med.mov[cont]<-mean(res_pred[i:(i+jan),2])
pred.med.mov[cont]<-mean(res_pred[i:(i+jan),1])
cont<-cont+1}
par(mfrow=c(2,1))


Python_SGPS 9748 _Preprocesamiento.ipynb
Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda Última modificación: 6 de abril
+ Código + Texto
res_pred<-res_pred[order(modlin$fitted.values),]
res.med.mov<-NULL
pred.med.mov<-NULL
cont<-1
jan<-30
for (i in 1:(length(res.pear)-jan)){
res.med.mov[cont]<-mean(res_pred[i:(i+jan),2])
pred.med.mov[cont]<-mean(res_pred[i:(i+jan),1])
cont<-cont+1}
par(mfrow=c(2,1))
plot(modlin$fitted.values,res.pear)
lines(pred.med.mov,res.med.mov)
plot(modlin$fitted.values,des.res)
#
[ ] %%R
plot(modlin$fitted.values,modlin$linear.predictor)
#
Tasa correcta de clasificación
#
[ ] %%R
cor_clas<-NULL
pcorte<-seq(from=min(modlin$fitted.values),to=max(modlin$fitted.values),by=0.005)
for (i in 1:length(pcorte)){
classif<-ifelse(modlin$fitted.values>pcorte[i],1,0)
cor_clas[i]<-(table(Y,classif)[1,1]+table(Y,classif)[2,2])/nrow(X)*100}
plot(pcorte,cor_clas)
pcorte[cor_clas==max(cor_clas)]
#
[ ] %%R
# dando costo 5 para clasificaciones equivocadas
err_clas<-NULL

Python_SGPS 9748 _Preprocesamiento.ipynb

Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda Última modificación: 6 de abril

+ Código + Texto

```
[ ] %%R
    # dando costo 5 para clasificaciones equivocadas
    err_clas<-NULL
    pcorte<-seq(from=min(modlin$fitted.values),to=max(modlin$fitted.values),by=0.005)
    for (i in 1:length(pcorte)){
      classif<-ifelse(modlin$fitted.values>pcorte[i],1,0)
      err_clas[i]<-(table(Y,classif)[1,2]+(1/5)*table(Y,classif)[2,1])/((1/5)*sum(Y==0)+sum(Y==1))*100}
    plot(pcorte,err_clas)
    pcorte[err_clas==min(err_clas)]
    #
```

```
[ ] %%R
    ##### Selección de variables
    #
    ## Usando stepwise por AIC
    #
    step(modlin, direction="both")
    modlin<-glm(Y~.,data=X,family=binomial(link="logit"))
    summary(modlin)
    #
```

```
[ ] %%R
    ## Selección de variables via Lasso
    #
    library(glmnet)
    set.seed(100)
    matX<-data.matrix(X)
    cv.out<-cv.glmnet(matX,Y,alpha=1,nfolds=5,family="binomial") #alpha=1 representa o lasso, alpha=0 representa o ridge e alpha entre 0 e 1 representa o elasticnet
    plot(cv.out)
    bestlam<-cv.out$lambda.min
    modlasso<-glmnet(matX,Y,alpha=1,lambda=bestlam,family="binomial")
    coef(modlasso)
```

```
[ ] %%R
```

Python_SGPS 9748 _Preprocesamiento.ipynb

Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda Última modificación: 6 de abril

+ Código + Texto

```
    coef(modlasso)
```

```
[ ] %%R
    ## Seleción de varaibles via Ridge
    #
    set.seed(100)
    cv.out<-cv.glmnet(matX,Y,alpha=0,nfolds=5,family="binomial") #alpha=1 representa o lasso, alpha=0 representa o ridge e alpha entre 0 e 1 representa o elasticnet
    plot(cv.out)
    bestlam<-cv.out$lambda.min
    modridge<-glmnet(matX,Y,alpha=0,lambda=bestlam,family="binomial")
    coef(modridge)
    #
```

Regresión lineal múltiple con Python principal_proyectos

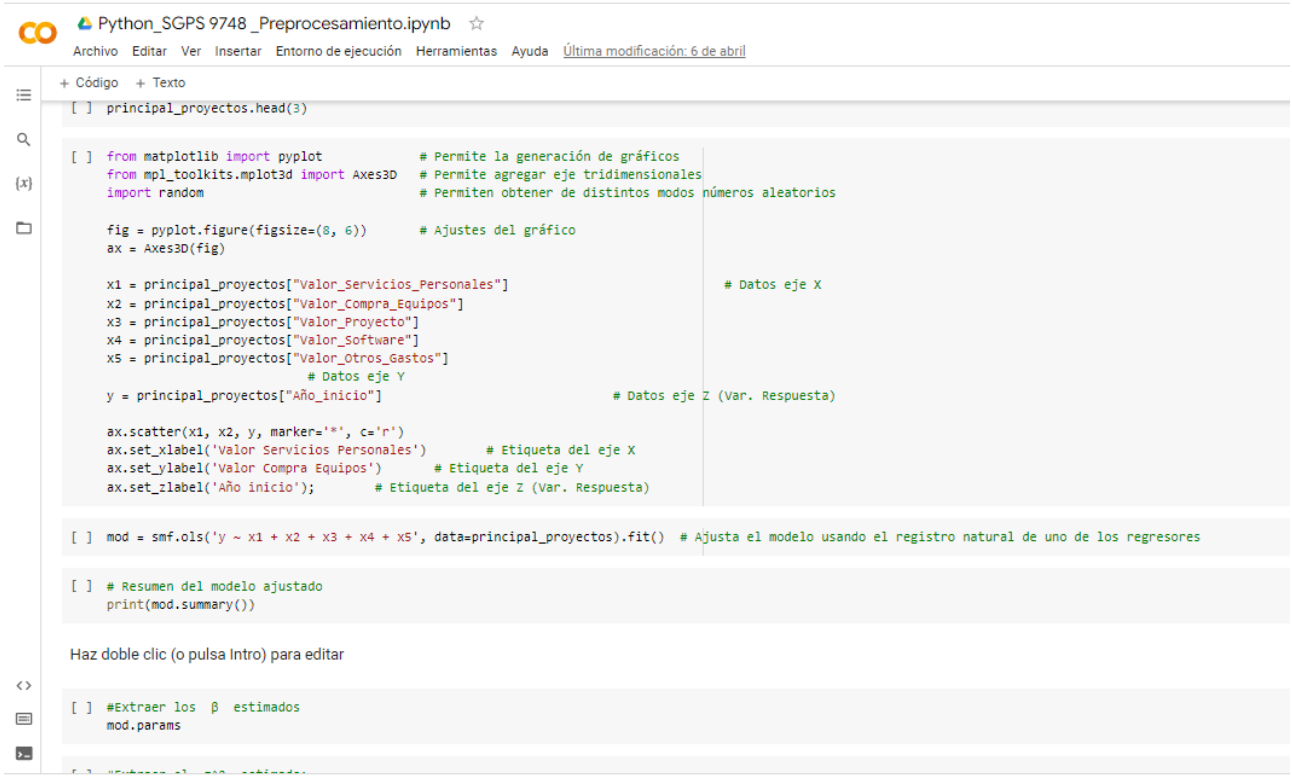
material de apoyo: https://yuasaavedraco.github.io/Docs/Regresi%C3%B3n_Lineal_M%C3%BAltiple_con_Python.html

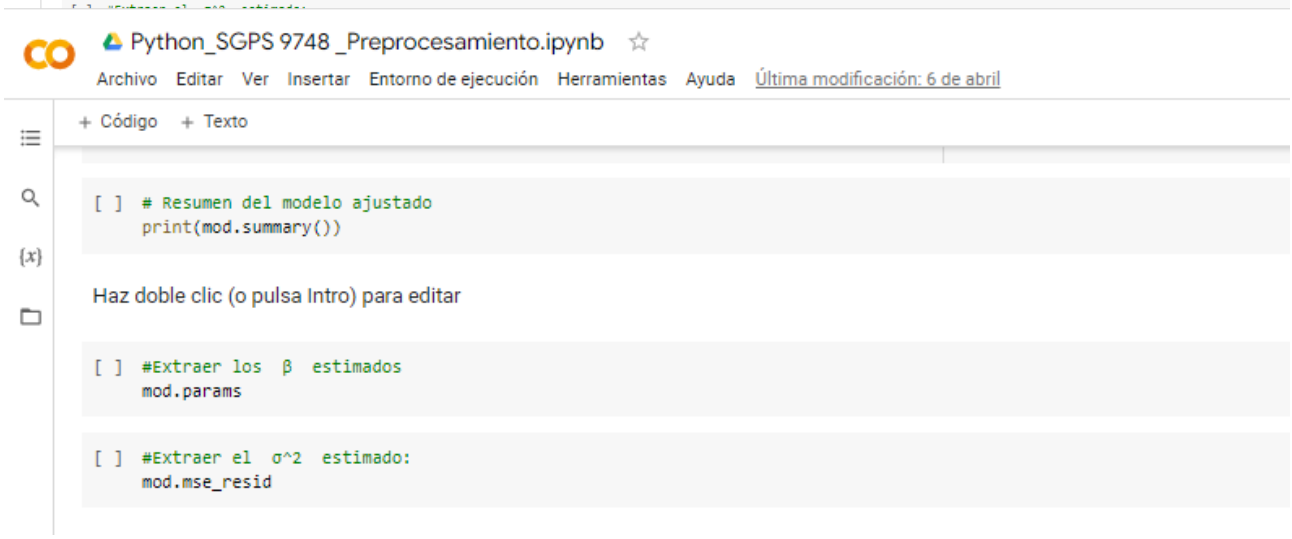
```
[ ] import statsmodels.api as sm          ## Este proporciona funciones para la estimación de muchos modelos estadísticos
    import statsmodels.formula.api as smf  ## Permite ajustar modelos estadísticos utilizando fórmulas de estilo R
```

Diagrama de dispersión

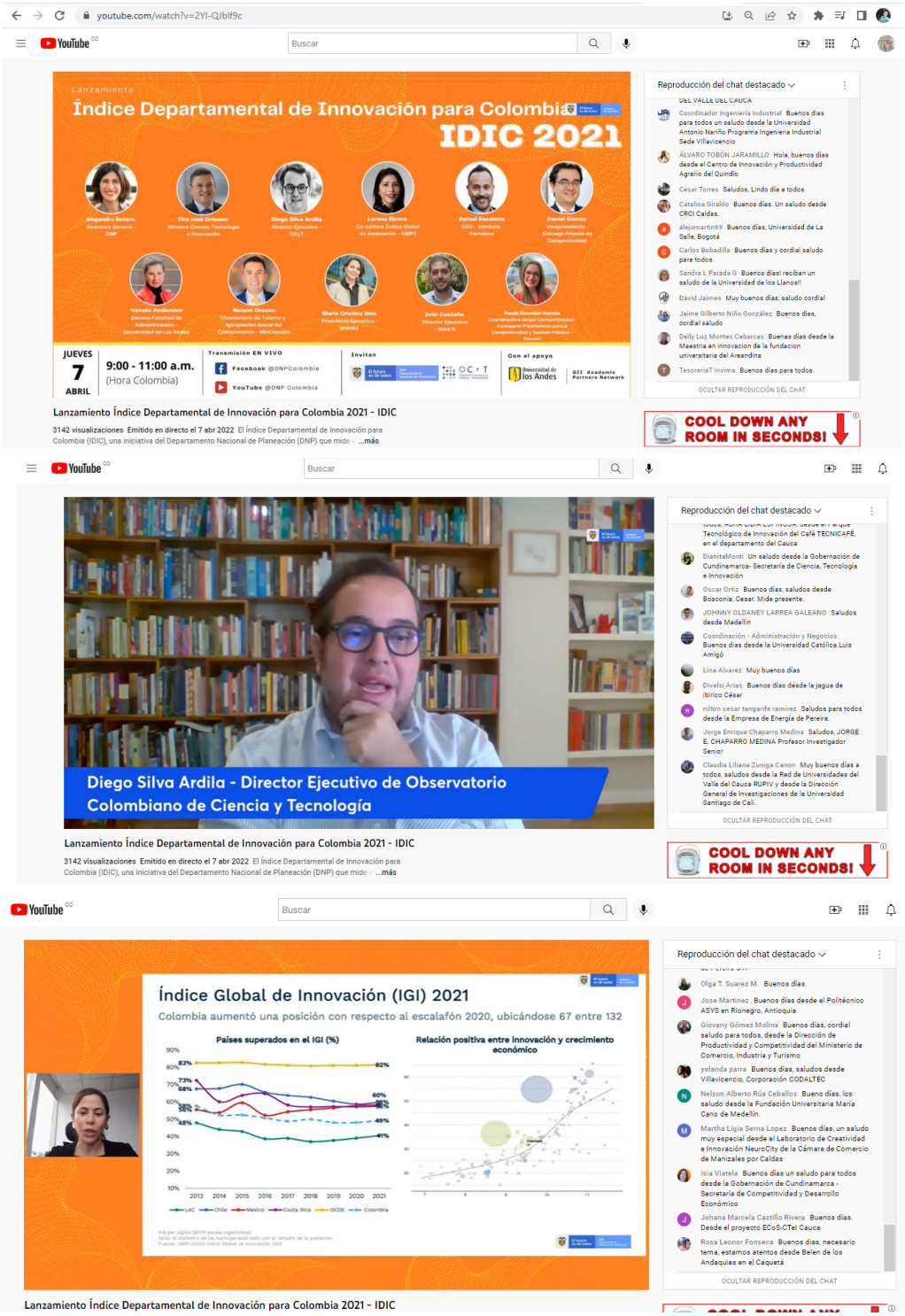
```
[ ] principal_proyectos.head(3)
```

```
[ ] from matplotlib import pyplot            # Permite la generación de gráficos
    from mpl_toolkits.mplot3d import Axes3D  # Permite agregar eje tridimensionales
    import random                            # Permiten obtener de distintos modos números aleatorios

    fig = pyplot.figure(figsize=(8, 6))      # Ajustes del gráfico
    ax = Axes3D(fig)

    x1 = principal_proyectos["Valor Servicios Personales"]                    # Datos eje X
```


Python_SGPS 9748 _Preprocesamiento.ipynb
Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda Última modificación: 6 de abril
+ Código + Texto
[ ] principal_proyectos.head(3)
[ ] from matplotlib import pyplot # Permite la generación de gráficos
from mpl_toolkits.mplot3d import Axes3D # Permite agregar eje tridimensionales
import random # Permiten obtener de distintos modos números aleatorios
fig = pyplot.figure(figsize=(8, 6)) # Ajustes del gráfico
ax = Axes3D(fig)
x1 = principal_proyectos["Valor_Servicios_Personales"] # Datos eje X
x2 = principal_proyectos["Valor_Compra_Equipos"]
x3 = principal_proyectos["Valor_Proyecto"]
x4 = principal_proyectos["Valor_Software"]
x5 = principal_proyectos["Valor_Otros_Gastos"]
# Datos eje Y
y = principal_proyectos["Año_inicio"] # Datos eje Z (Var. Respuesta)
ax.scatter(x1, x2, y, marker='*', c='r')
ax.set_xlabel('Valor Servicios Personales') # Etiqueta del eje X
ax.set_ylabel('Valor Compra Equipos') # Etiqueta del eje Y
ax.set_zlabel('Año inicio'); # Etiqueta del eje Z (Var. Respuesta)
[ ] mod = smf.ols('y ~ x1 + x2 + x3 + x4 + x5', data=principal_proyectos).fit() # Ajusta el modelo usando el registro natural de uno de los regresores
[ ] # Resumen del modelo ajustado
print(mod.summary())
Haz doble clic (o pulsa Intro) para editar
[ ] #Extraer los β estimados
mod.params


Python_SGPS 9748 _Preprocesamiento.ipynb
Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda Última modificación: 6 de abril
+ Código + Texto
[ ] # Resumen del modelo ajustado
print(mod.summary())
Haz doble clic (o pulsa Intro) para editar
[ ] #Extraer los β estimados
mod.params
[ ] #Extraer el σ^2 estimado:
mod.mse_resid

ANEXOS 3. Participación en evento virtual: Índice de Innovación para Colombia.



GTH-F-062 V09

ANEXOS 4. Presentación de avances técnicos abril 2022.

El empleo es de todos | Mintrabajo

Proyecto SGPS 9748

Prototipo Innovación

Presentación de avances 06

Modelo de evaluación y valoración de la Gestión de CTeI para los Centros SENA Regional Antioquia aplicando analítica de datos

6 de abril del 2022

SENA

Contenido

1. Etapas Modelo CRISP - DM - Proyecto SGPS-9748
2. Cronograma 2022 - Proyecto SGPS-9748
3. Organización de base de datos
4. Modelos de analítica de datos
   - 4.1. Machine learning no supervisado
   - 4.2. Machine learning supervisado
5. Actividades de divulgación

1. Etapas modelo CRISP - DM - Proyecto SGPS-9748

1. ~~Comprensión del negocio~~
2. ~~Comprensión de los datos~~
3. Preparación de los datos
4. Modelado
5. Evaluación
6. Implementación
7. Modelo de analítica de datos
8. Validación del modelo

Informe - V4 - Modelo de evaluación y valoración de la Gestión de CTeI para los Centros SENA Regional Antioquia aplicando analítica de datos

2. Cronograma 2022 - Proyecto SGPS-9748

	Fechas programadas	Febrero	Marzo	Abril	Mayo	Junio	Julio	Agosto
Fase 1 : Avance de la metodología del Big Data	3 marzo	100%						
Fase 2 : Identificación y selección del Dataset análogo al trabajo en particular	15 abril		100 %					
Fase 3: Caracterizar y seleccionar modelos analítica	19 mayo			30%				
Fase 4: Aplicación de API's para vincular con el desarrollo web	20 junio							
Fase 5 : Puesta en marcha del desarrollo web con los modelos de segmentación y clusterización	21 julio							
Entrega final	15 agosto							

GC-F-004v.05

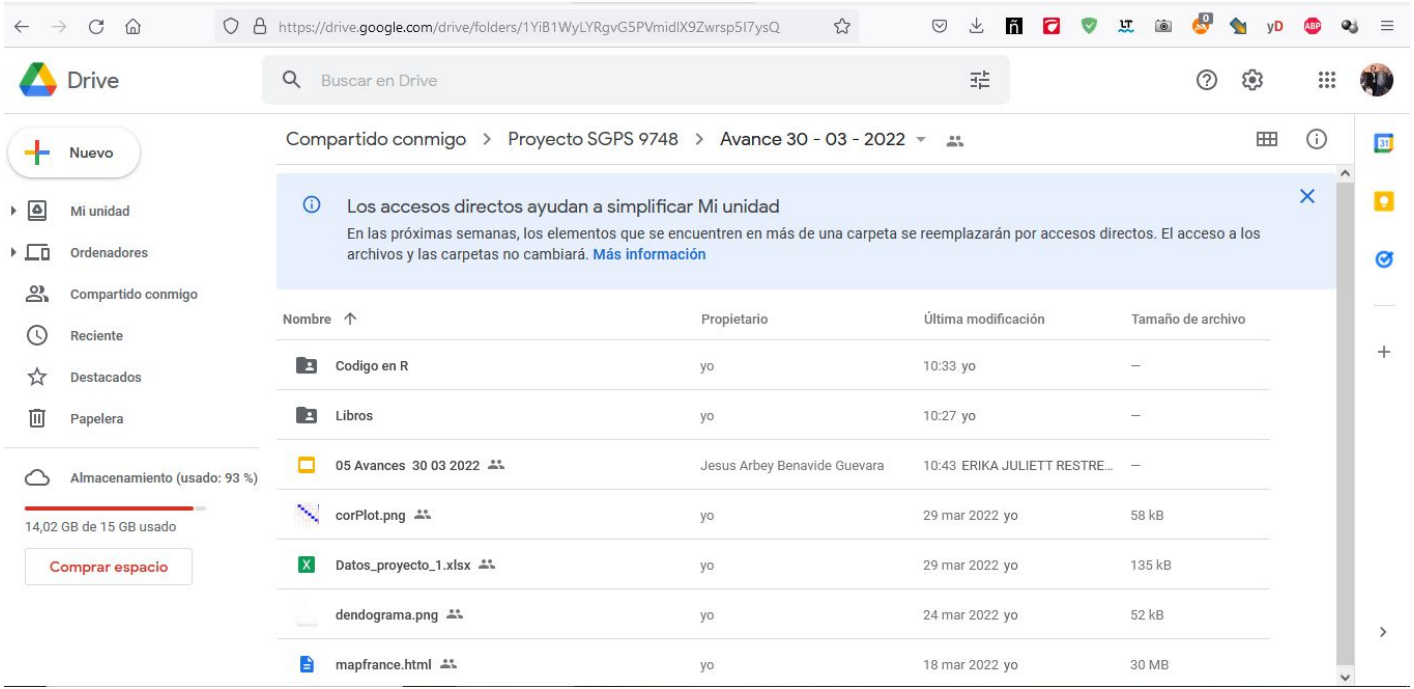
3. Organización de base de datos

Indicadores y correspondencia de variables de entrada y salida del modelo

Nro	DIMENSIÓN (Proceso de entradas y salidas)	CATEGORIA	INDICADOR (Variables)	INDICADOR FINAL
1	Insumos para actividades de Ciencia, Tecnología e Innovación (ACTI) (Entradas en el modelo)	RECURSOS ECONÓMICOS	GASTO/INVERSIÓN en ACTI	Suma de los diferentes rubros ejecutados.
2		RECURSOS FÍSICOS Y TECNOLÓGICOS	NÚMERO DE LABORATORIOS PARA ACTI	AMBIENTES DE APRENDIZAJE (Laboratorios dotados con recursos de SENNOVA)
3			INVERSIÓN DE EQUIPOS DE PARA ACTI	Suma del rubro de equipos y maquinaria,
4			NÚMERO DE LICENCIAS (Campo de conocimiento segun la OCDE) (Por programas de educación)	-Software de aplicaciones especializadas -Software de programación
5		TALENTO HUMANO	NUMERO DE PERSONAS DEDICADAS A ACTI	
6		RECURSOS ESTRUCTURALES Y RELACIONALES	NÚMERO DE UNIDADES ORGANIZATIVAS (GRUPOS DE INVESTIGACIÓN - SEMILLEROS)	
7			NÚMERO DE PROGRAMAS IMPACTADOS DE EDUCACIÓN SUPERIOR OFERTADOS (PROGRAMAS DEL SENA POR ACTI)	
8			NÚMERO DE CONVENIOS DE COLABORACIÓN PARA ACTI	
9	Proceso (Planta del modelo)	PROYECTOS	NÚMERO DE PROYECTOS	
10		TALENTO HUMANO	PROPORCIÓN DE GASTO EN ACTI POR PERSONAS DEDICAS A ACTI	
11		RECURSOS ECONÓMICOS	PORCENTAJE DE GASTO EN ACTI en relación con el PRESUPUESTO TOTAL	
12			Porcentaje de gasto en INVESTIGACIÓN APLICADA en relación con el gasto total por ACTI	
13		RELACIONES CON EL ENTORNO	EVENTOS DE MOVILIDAD PARA ACTI	
14		FORMACIÓN CIENTÍFICA Y TECNOLÓGICA	Formacion de aprendices en semilleros de investigacion	
15	Resultado (Salidas del modelo)	RELACIONES CON EL ENTORNO	Convenios con otras instituciones	
16		PRODUCTOS RESULTADOS DE LAS ACTIVIDADES DE CTI	NÚMERO DE PRODUCTOS DE I+D (Productos de I+D por tipología Colciencias)	
17		TRANFERENCIA DE CONOCIMIENTO	NÚMERO DE EVENTOS DE CTI ORGANIZADOS POR LA UNIVERSIDAD (Número de eventos de CTI por clasificación ACTI)	
19		FORMACIÓN CIENTÍFICA Y TECNOLÓGICA	Numero de aprendices con formación científica	
20	Productividad (Rendimiento del modelo)	TALENTO HUMANO	PROPORCIÓN DE PRODUCTOS DE CTI POR PERSONA DEDICADAS A ACTI, EN RELACIÓN CON EL TOTAL DE PERSONAL	
21			PROPORCIÓN DE PRODUCTOS DE CTI POR UNIDAD ORGANIZATIVA	
22		RELACIONES CON EL ENTORNO	PROPORCIÓN DE EVENTOS DE MOVILIDAD PARA ACTI POR PERSONAS DEDICADAS A ACTI	

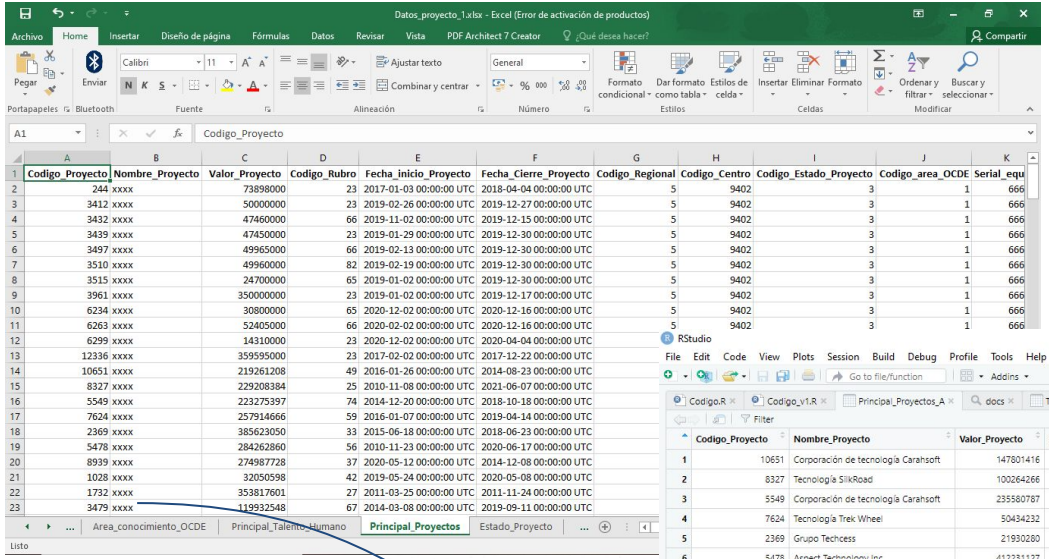
Identificación de indicadores: Indicadores y correspondencia de variables de entrada y salida del modelo
https://docs.google.com/spreadsheets/d/1vL_HVo1sFDdqqCgYSwjY8axuVdo8fJenDftpjcF_WN0/edit?usp=sharing

Organización de base de datos



Referencia: https://drive.google.com/drive/folders/1YiB1WyLYRgvG5PVmidlX9Zwrsp5I7ysQ

Dataframes



Referencia: https://drive.google.com/drive/folders/1YiB1WyLYRgvG5PVmidlX9Zwrsp5l7ysQ

GC-F-004v.05

4. Modelos de analítica de datos

Dataframes identificados en el modelo relacional.

Principal_Productos	Principal_Proyectos	Principal_Talento_Humano	Principal_Equipos
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	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	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	Serial_equipo Nombre_del_equipo Valor_equipo Fecha_de_compra Codigo_Proyecto Codigo_Regional Codigo_Centro Codigo_Rubro Codigo_Tipo_Equipo Codigo_area_OCDE Codigo_persona

dataframes y métodos de machine learning

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

Modelo de ML implementados

No supervisados :

- Análisis de componentes principales (ACP)
- Clustering - Kmeans
- Dendrograma

Supervisados:

- Regresión logística
- Support Vector Machine(SVM)

Identificación de *dataframes: Características globales*

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

Modelo de ML implementados

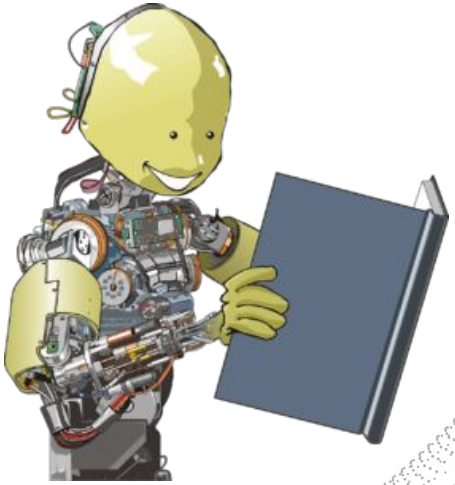
Supervisados :

- Regresión múltiple

Técnicas implementadas hasta el momento:

NO SUPERVISADO

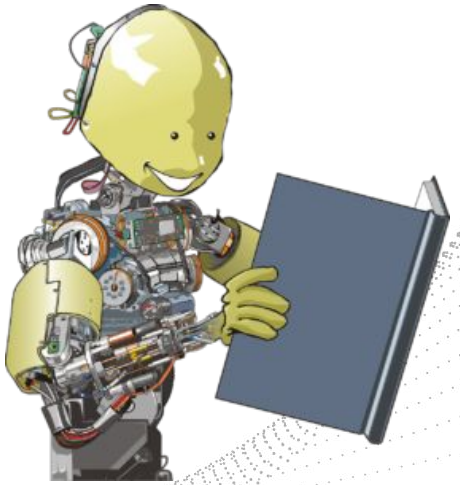
- Análisis de componentes principales (ACP)
- Clustering - Kmeans
- Dendrograma



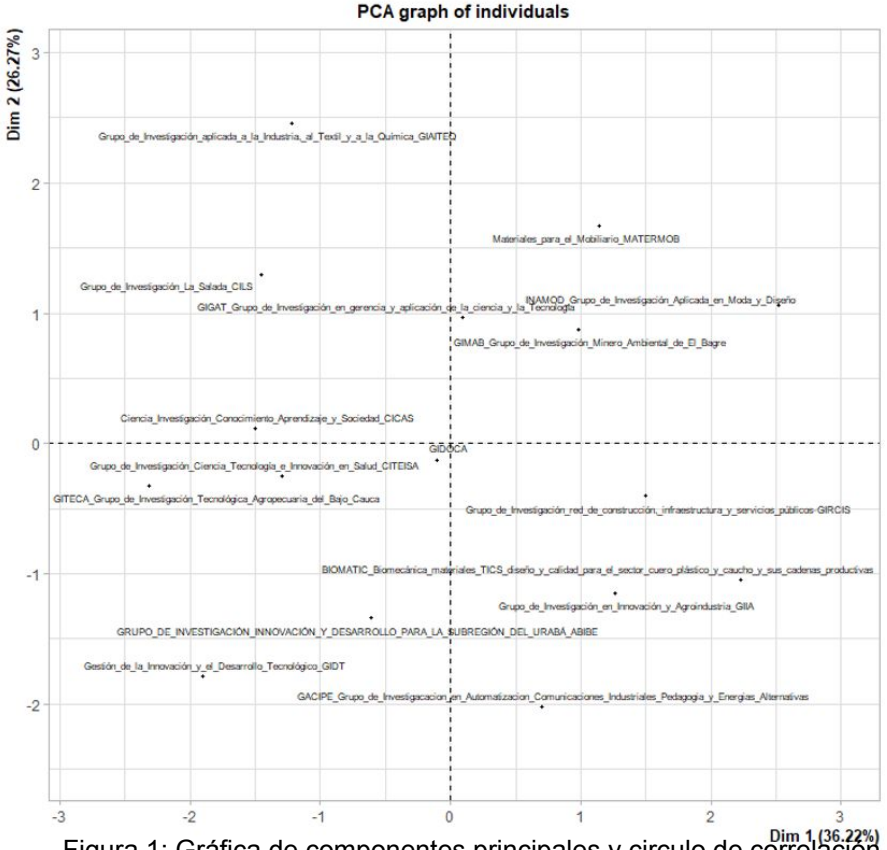
SUPERVISADO:

- Regresión Logística
- Regresión múltiple
- Máquina de Soporte Vectorial (SVM) - *Support vector machine*

4.1. MACHINE LEARNING NO SUPERVISADO



Análisis de componentes principales y Circulo de correlación



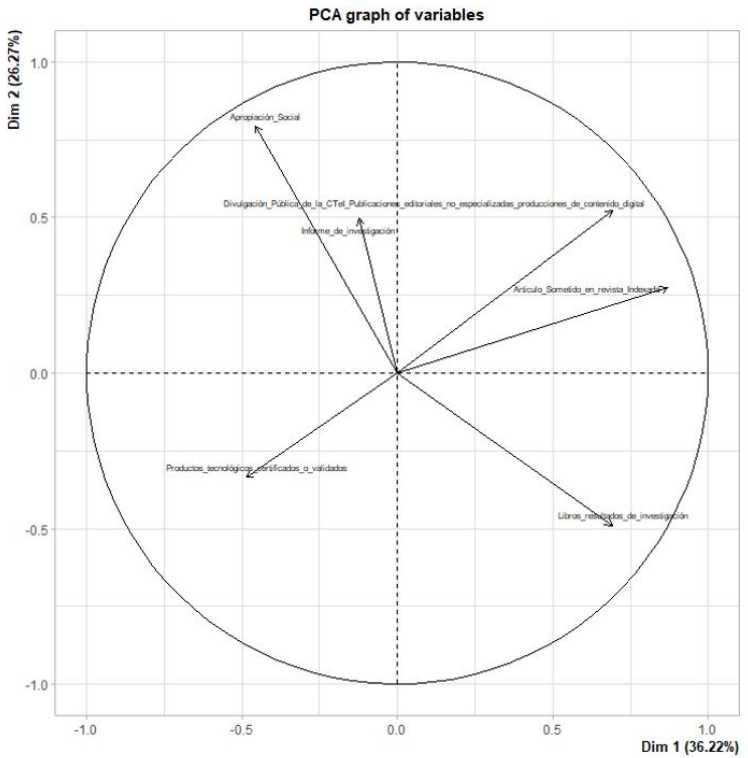


Figura 1: Gráfica de componentes principales y circulo de correlación de variables por productos de CTeI

Referencia: https://drive.google.com/drive/folders/1YiB1WyLYRgvG5PVmidlX9Zwrsp5l7ysQ

Análisis de componentes principales y Círculo de correlación

	Apropiación_Social	Articulo_Sometido_en_revista_Indexada	Divulgación_Pública_de_la_CTeI_Publicaciones_editoriales_no_especializadas_producciones_de_contenido_digital	Informe_de_investigación	Libros_resultados_de_investigación	Productos_tecnológicos_certificados_o_validados
Grupo_de_Investigación_La_Salada_CILS	9	3	4	12	5	7
Grupo_de_Investigación_en_Innovación_y_Agroindustria_GIIA	2	7	5	8	9	9
Gestión_de_la_Innovación_y_el_Desarrollo_Tecnológico_GIDT	3	1	1	6	5	7
GACIPE_Grupo_de_Investigacacion_en_Automatizacion_Comunicaciones_Industriales_Pedagogia_y_Energias_Alternativ	2	3	4	3	9	5
Materiales_para_el_Mobiliario_MATERMOB	7	7	7	4	3	2
BIOMATIC_Biomecánica_materiales_TICS_diseño_y_calidad_para_el_sector_cuero_plástico_y_caucho_y_sus_cadenas_	2	7	6	2	8	4
GIGAT_Grupo_de_Investigación_en_gerencia_y_aplicación_de_la_ciencia_y_la_Tecnología	6	3	6	8	5	3
Ciencia_Investigación_Conocimiento_Aprendizaje_y_Sociedad_CICAS	8	4	3	4	3	7
GRUPO_DE_INVESTIGACIÓN_INNOVACIÓN_Y_DESARROLLO_PARA_LA_SUBREGIÓN_DEL_URABÁ_ABIBE_GIIDS	5	3	3	2	6	6
GIDOCA	8	3	5	3	7	5
Grupo_de_Investigación_red_de_construcción_infraestructura_y_servicios_públicos-GIRCIS	4	5	8	4	8	7
Grupo_de_Investigación_Ciencia_Tecnología_e_Innovación_en_Salud_CITEISA	5	2	3	3	2	3
Grupo_de_Investigación_aplicada_a_la_Industria_al_Textil_y_a_la_Química_GIAITEQ	13	4	6	7	3	6
INAMOD_Grupo_de_Investigación_Aplicada_en_Moda_y_Diseño	5	8	7	8	8	2
GITECA_Grupo_de_Investigación_Tecnológica_Agropecuaria_del_Bajo_Cauca	6	0	5	6	3	10
GIMAB_Grupo_de_Investigación_Minero_Ambiental_de_El_Bagre	5	6	7	7	5	5

Cluster 1
Cluster 2
Cluster 3

Tabla 1: Caracterización de grupos de investigación por productos de CTeI

Referencia: https://drive.google.com/drive/folders/1YiB1WyLYRgvG5PVmidlX9Zwrsp5I7ysQ

Análisis de clustering

Los tres clusters se agrupan según los grupos establecidos en los cuatro cuadrantes de la PCA. Y se distribuyen por productos según los patrones de productos de CTeI descritos en los ítems 1,2,3 y 4.

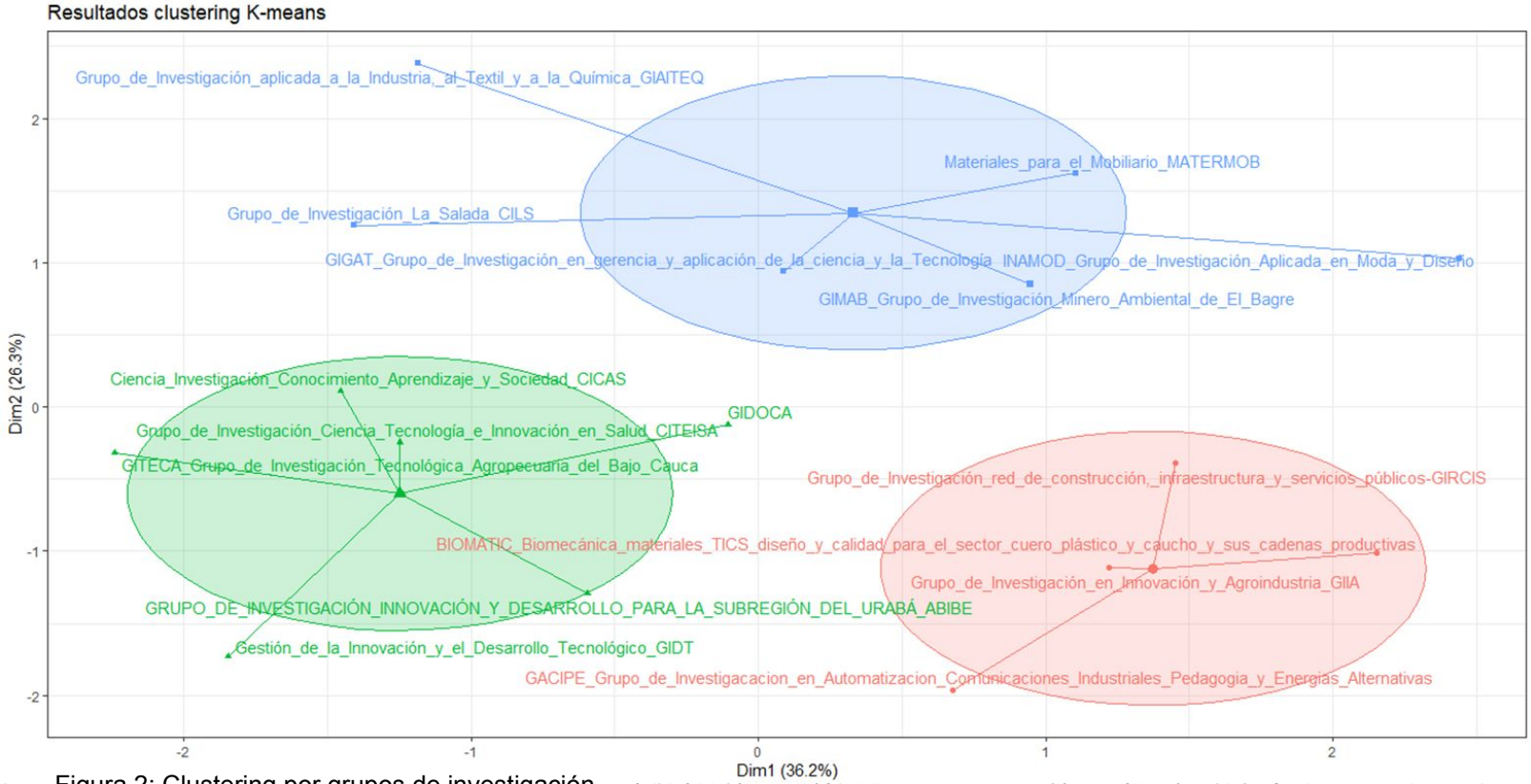


Figura 2: Clustering por grupos de investigación

Referencia: https://drive.google.com/drive/folders/1YiB1WyLYRgvG5PVmidlX9Zwrsp5I7ysQ

Dendograma

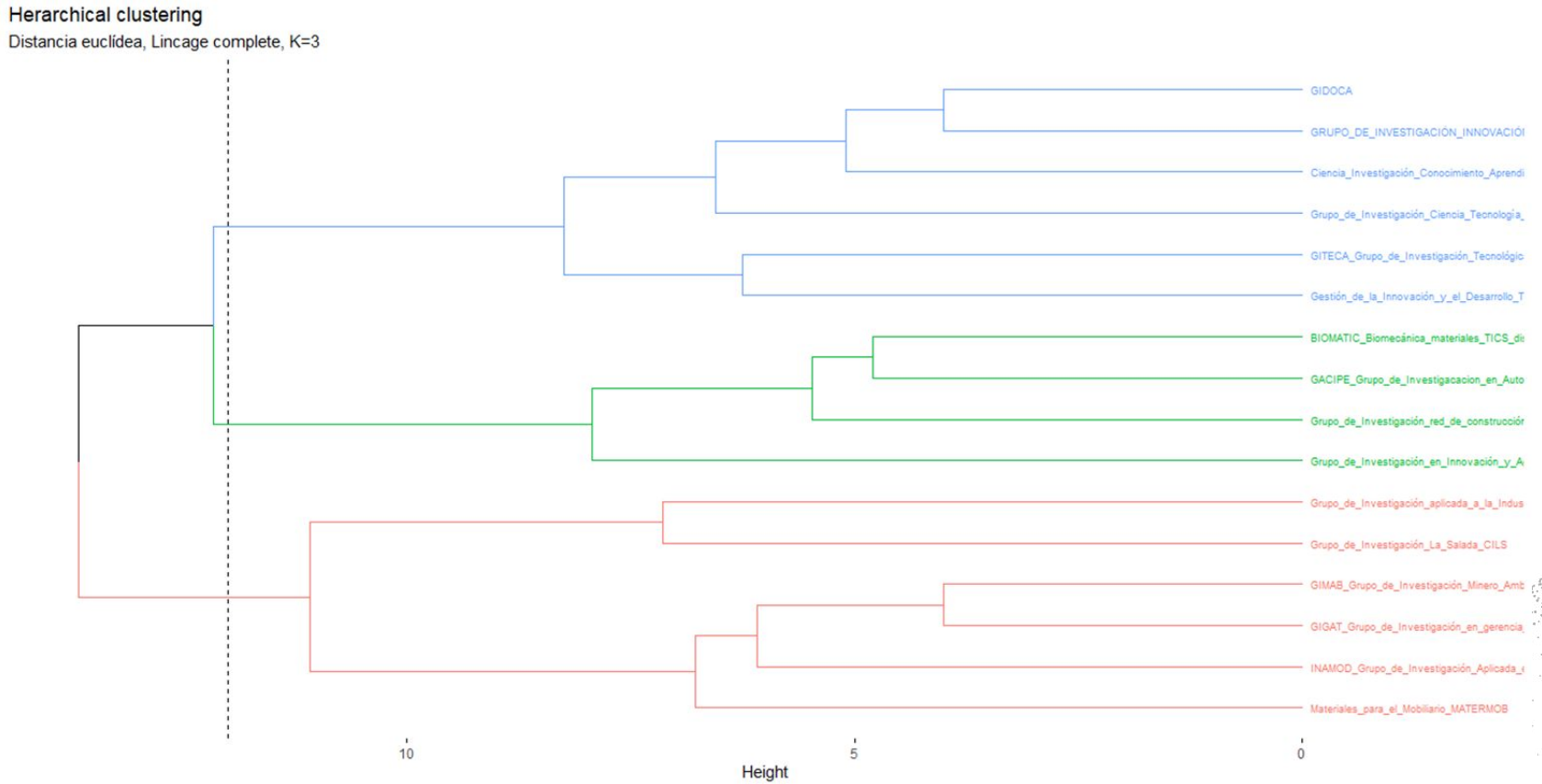


Figura 3: Clustering Jerárquico por grupos de investigación

Referencia: https://drive.google.com/drive/folders/1YiB1WyLYRgvG5PVmidlX9Zwrsp5I7ysQ

Mapa de calor de productos de CTeI por cluster (Datos sin escalar)

```
> center_df
  cluster Apropiación_Social Articulo_Sometido_en_revista_Indexada
1       1           2.500000                                5.500000
2       2           5.833333                                2.166667
3       3           7.500000                                5.166667
  Divulgación_Pública_de_la_CTeI_Publicaciones_editoriales_no_especializadas_producciones_de_contenido_digital
1                                                                                                    5.750000
2                                                                                                    3.333333
3                                                                                                    6.166667
  Informe_de_investigación Libros_resultados_de_investigación Productos_tecnológicos_certificados_o_validados
1                 4.250000                            8.500000                                       6.250000
2                 4.000000                            4.333333                                       6.333333
3                 7.666667                            4.833333                                       4.166667
```

Tabla 3: Productos de CTeI por clustering (Datos sin escalar)

```
> km_clusters.df$size
[1] 4 6 6
>
```

Tabla 4: Tamaño de observaciones por clustering

Referencia: https://drive.google.com/drive/folders/1YiB1WyLYRgvG5PVmidlX9Zwrsp5I7ysQ

Mapa de calor de productos de CTeI por cluster (Datos sin escalar)

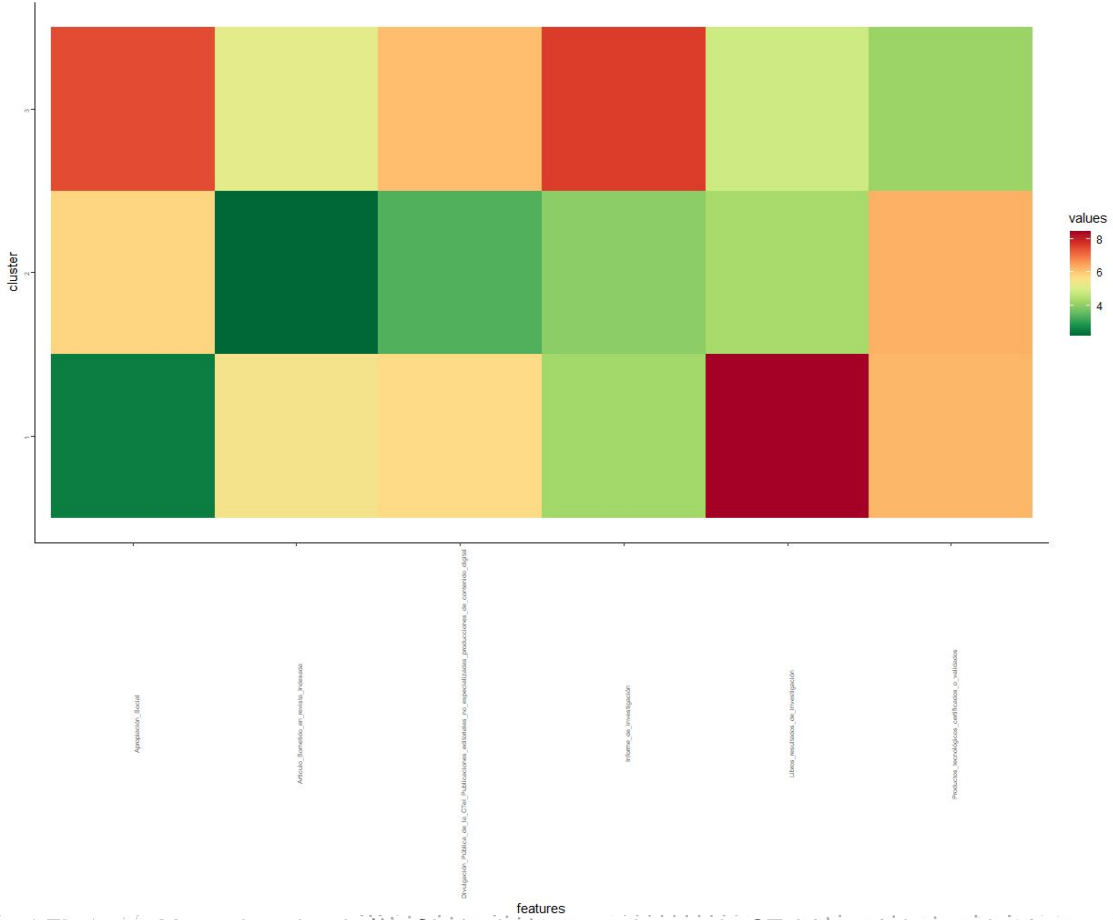


Figura 5: Mapa de calor de los Clustering por productos de CTeI, datos sin escalar

Referencia: https://drive.google.com/drive/folders/1YiB1WyLYRgvG5PVmidlX9Zwrsp5I7ysQ

Mapa de calor de productos de CTeI por cluster (Datos escalados)

	cluster	features	values
1	1	Apropiación_Social_scal	1.26087116
2	2	Apropiación_Social_scal	-0.40679792
3	3	Apropiación_Social_scal	-0.15583801
4	1	Articulo_Sometido_en_revista_Indexada_scal	-0.76746568
5	2	Articulo_Sometido_en_revista_Indexada_scal	0.86435503
6	3	Articulo_Sometido_en_revista_Indexada_scal	-0.62468137
7	1	Divulgación_Pública_de_la_CTeI_Publicaciones_editoriales_no_especializadas_producciones_de_contenido_digital_scal	0.00000000
8	2	Divulgación_Pública_de_la_CTeI_Publicaciones_editoriales_no_especializadas_producciones_de_contenido_digital_scal	0.82821558
9	3	Divulgación_Pública_de_la_CTeI_Publicaciones_editoriales_no_especializadas_producciones_de_contenido_digital_scal	-0.96625151
10	1	Productos_tecnológicos_certificados_o_validados_scal	0.92668217
11	2	Productos_tecnológicos_certificados_o_validados_scal	-0.39714950
12	3	Productos_tecnológicos_certificados_o_validados_scal	0.00000000
13	1	Informe_de_investigación_scal	1.04161500
14	2	Informe_de_investigación_scal	0.15094318
15	3	Informe_de_investigación_scal	-0.69690788
16	1	Libros_resultados_de_investigación_scal	-0.80143401
17	2	Libros_resultados_de_investigación_scal	0.42650884
18	3	Libros_resultados_de_investigación_scal	-0.09687664

Tabla 5: Productos de CTeI por clustering (Datos escalados)

```
> km_clusters$size
[1] 7 3 6
>
```

Tabla 6: Tamaño de observaciones por clustering

Referencia: https://drive.google.com/drive/folders/1YiB1WyLYRgvG5PVmidlX9Zwrsp5I7ysQ

Mapa de calor de productos de CTeI por cluster (Datos escalados)

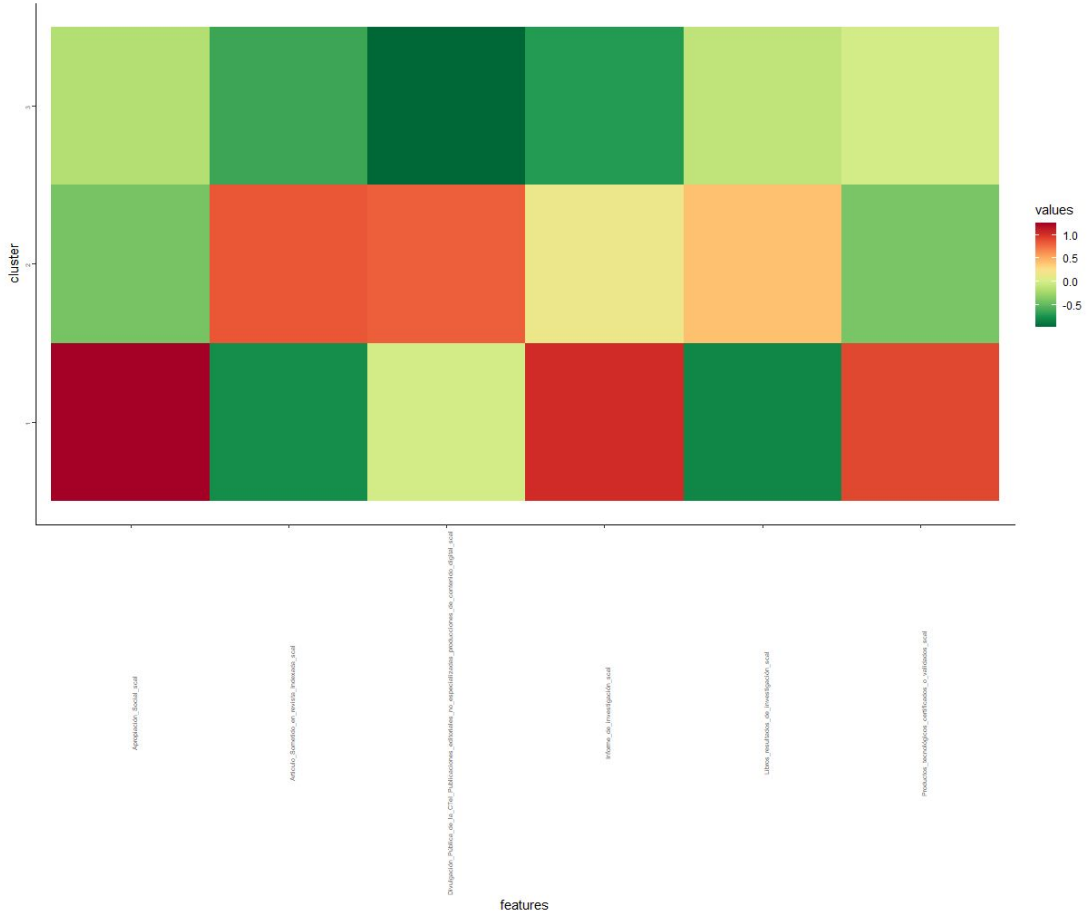
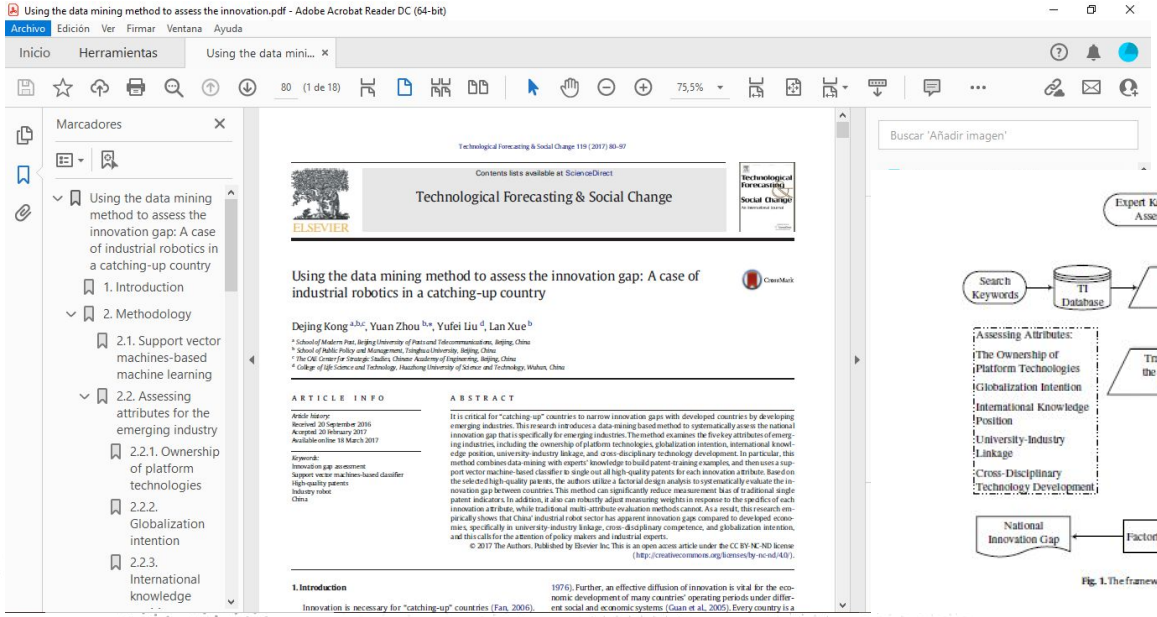


Figura 4: Mapa de calor de los Clustering por productos de CTeI, datos escalados.

Referencia: https://drive.google.com/drive/folders/1YiB1WyLYRgvG5PVmidlX9Zwrsp5I7ysQ

Análisis Bibliográfico de técnicas a implementar



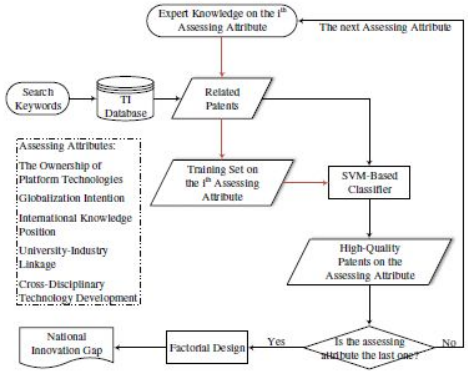
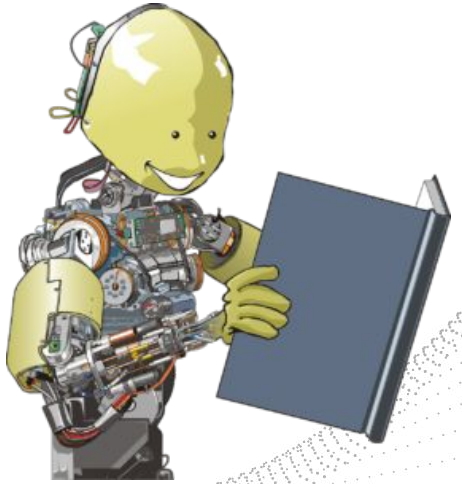


Fig. 1. The framework for innovation gap assessment.

Referencia: https://drive.google.com/drive/folders/1YiB1WyLYRgvG5PVmidIX9Zwrsp5l7ysQ

4.2. MACHINE LEARNING SUPERVISADO



Regresión logística : Variables ficticias (dummy)

Principal_Productos

Codigo_Centro	'Codigo_Proyecto'	Codigo_Estado_Proyecto	Codigo_Rubro
Codigo_Centro_9101	Codigo_Proyecto_12287	Codigo_Estado_Proyecto_1	Codigo_Rubro_65
Codigo_Centro_9127	Codigo_Proyecto_12325	Codigo_Estado_Proyecto_2	Codigo_Rubro_65
...	...	...	...

Tabla 3: Generación de variables ficticias para generación de modelos predictivos

Se generan 514 variables. Para un primer modelo, determinamos analizar para las primeras **32 variables** de manera arbitraria entre código centro y código proyecto.

Regresión logística : Base de test

```
Coefficients: (14 not defined because of singularities)
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)           2.657e+01  1.781e+05       0        1
Codigo_Centro_9101   -5.313e+01  2.299e+05       0        1
Codigo_Centro_9127   -5.313e+01  2.518e+05       0        1
Codigo Centro 9201   -5.313e+01  2.299e+05       0        1
Codigo_Centro_9202   -5.313e+01  2.181e+05       0        1
Codigo Centro 9203   -5.313e+01  2.232e+05       0        1
Codigo_Centro_9204   -5.313e+01  2.036e+05       0        1
Codigo_Centro_9205   -5.313e+01  2.181e+05       0        1
Codigo Centro 9206   -5.313e+01  3.982e+05       0        1
Codigo_Centro_9301   -5.313e+01  3.084e+05       0        1
Codigo Centro 9401   -5.313e+01  3.084e+05       0        1
Codigo_Centro_9501   -5.313e+01  2.056e+05       0        1
Codigo_Centro_9502   -5.313e+01  2.518e+05       0        1
Codigo Centro 9503   -5.313e+01  2.107e+05       0        1
Codigo_Centro_9504   -5.313e+01  2.299e+05       0        1
Codigo Centro 9549   -5.313e+01  2.720e+05       0        1
Codigo_Proyecto_296   1.373e-13  3.847e+05       0        1
Codigo_Proyecto_378  -2.582e-30  4.362e+05       0        1
Codigo Proyecto 392  -4.313e-14  3.696e+05       0        1
Codigo_Proyecto_514   2.005e-14  3.847e+05       0        1
```

Características Generales

- Se define la variable a predecir cómo `Codigo_Centro_9402`
- De las 32 variables, el método determina que 19 de estas son relevantes
- Se observa una alto valor de desviación estándar
- Aplicación aproximación por medio de Logit

Tabla 4: Coeficientes de modelo lineal de regresión logística en modelo predictivo para el centro de servicios y gestion empresarial, base de test

Regresión logística : Base train

```
Coefficients: (4 not defined because of singularities)
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)           2.657e+01  7.269e+04   0.000        1
Codigo_Centro_9101   -5.313e+01  9.554e+04  -0.001        1
Codigo_Centro_9127   -5.313e+01  1.094e+05   0.000        1
Codigo_Centro_9201   -5.313e+01  1.149e+05   0.000        1
Codigo_Centro_9202   -5.313e+01  1.028e+05  -0.001        1
Codigo_Centro_9203   -5.313e+01  1.094e+05   0.000        1
Codigo_Centro_9204   -5.313e+01  1.051e+05  -0.001        1
Codigo_Centro_9205   -5.313e+01  1.078e+05   0.000        1
Codigo_Centro_9206   -5.313e+01  9.438e+04  -0.001        1
Codigo_Centro_9301   -5.313e+01  9.494e+04  -0.001        1
Codigo_Centro_9401   -5.313e+01  1.028e+05  -0.001        1
Codigo_Centro_9501   -5.313e+01  9.827e+04  -0.001        1
Codigo_Centro_9502   -5.313e+01  1.172e+05   0.000        1
Codigo_Centro_9503   -5.313e+01  1.008e+05  -0.001        1
Codigo_Centro_9504   -5.313e+01  1.078e+05   0.000        1
Codigo_Centro_9549   -5.313e+01  1.008e+05  -0.001        1
Codigo_Proyecto_324   7.606e-30  3.649e+05   0.000        1
Codigo_Proyecto_345   3.757e-14  3.629e+05   0.000        1
Codigo_Proyecto_381  -2.332e-13  3.671e+05   0.000        1
Codigo_Proyecto_406   4.067e-15  3.629e+05   0.000        1
Codigo_Proyecto_410   5.697e-29  3.635e+05   0.000        1
Codigo_Proyecto_415   5.257e-29  3.635e+05   0.000        1
Codigo_Proyecto_426   1.010e-15  3.615e+05   0.000        1
Codigo_Proyecto_431   1.010e-15  3.615e+05   0.000        1
Codigo_Proyecto_435   1.062e-13  3.622e+05   0.000        1
Codigo_Proyecto_445   1.081e-15  3.613e+05   0.000        1
Codigo_Proyecto_482  -5.565e-31  3.678e+05   0.000        1
Codigo_Proyecto_484  -2.205e-15  3.612e+05   0.000        1
Codigo_Proyecto_511   1.010e-15  3.615e+05   0.000        1
Codigo_Proyecto_539  -8.854e-30  3.678e+05   0.000        1
```

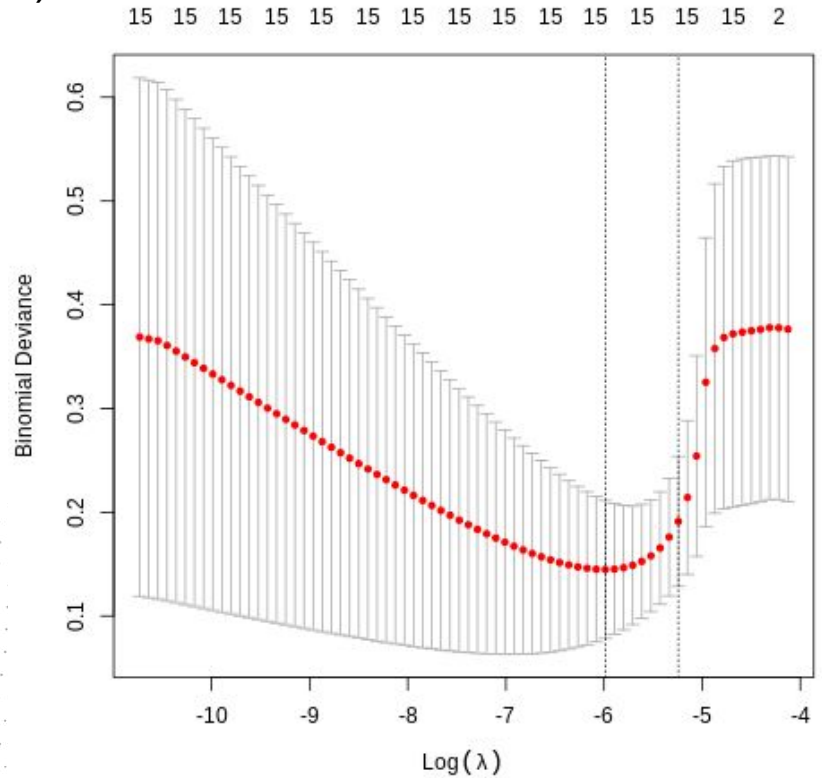
Características Generales

- Se define la variable a predecir cómo `Codigo_Centro_9402`
- De las 32 variables, el método determina que 30 de estas son relevantes
- Se observa una alto valor de desviación estándar
- Aplicación aproximación por medio de Logit

Tabla 5: Coeficientes de modelo lineal de regresión logística en modelo predictivo para el centro de servicios y gestion empresarial, base de entrenamiento

Regresión logística : Método Lasso de selección de variables

a) TEST



b) TRAIN

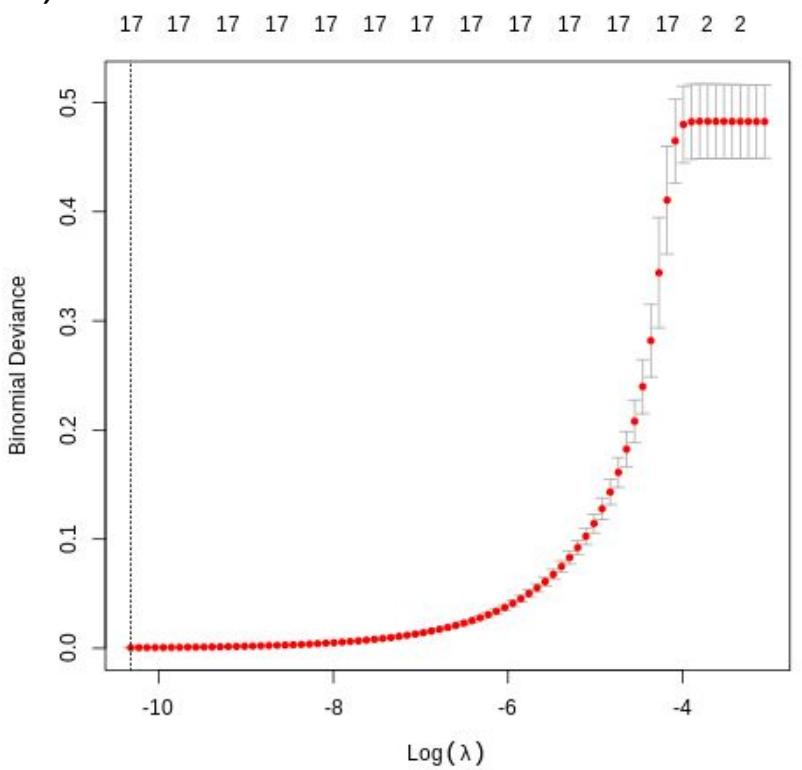
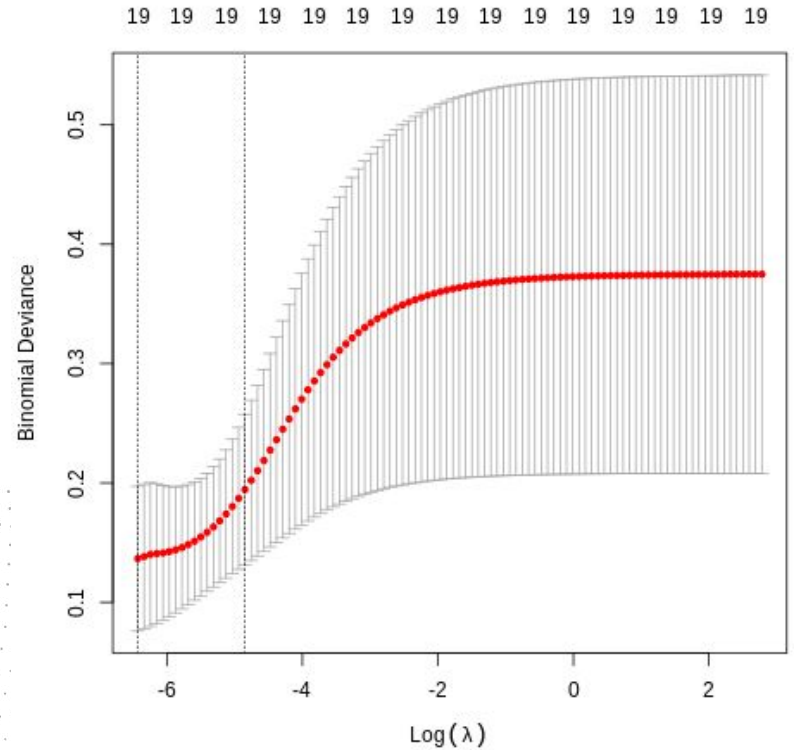


Figura 5: Gráfica de selección de variables modelo logístico método Lasso. a) datos de test y b) datos de entrenamiento.

Regresión logística : Método Ridge de selección de variables

a) TEST



b) TRAIN

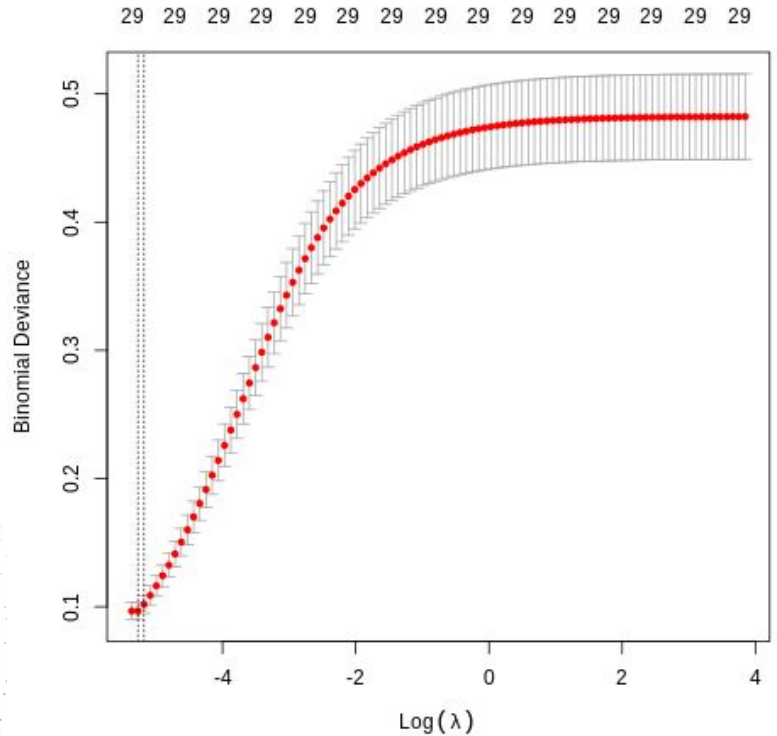


Figura 6: Gráfica de selección de variables modelo logístico método Ridge. a) datos de test y b) datos de entrenamiento.

Support Vector Machine (SVM)

Dataframe Principal_Productos

Y
Codigo_Centro

X
Codigo_Proyecto
Valor_Proyecto
Codigo_Rubro
Codigo_Estado_Proyecto
Valor_Servicios_Personales
Valor_Compra_Equipos
Valor_Software
Valor_Otros_Gastos
Año_inicio
Mes_inicio
Año_fin
Mes_fin

Se trabajan 13 variables de las 19 de Principal_Productos. Se determina como variable a predecir Codigo_Centro.

```
accuracy_score(y_test, y_pred) = 0.08
```

Regresión Múltiple

Dataframe Principal_Proyectos

Y
Año_inicio

X
Valor_Servicios_Personales
Valor_Compra_Equipos
Valor_Proyecto
Valor_Software
Valor_Otros_Gastos

Se trabajan 6 variables de las 19 de Principal_Proyectos. Se determina como variable dependiente `Año_inicio`.

Regresión Múltiple

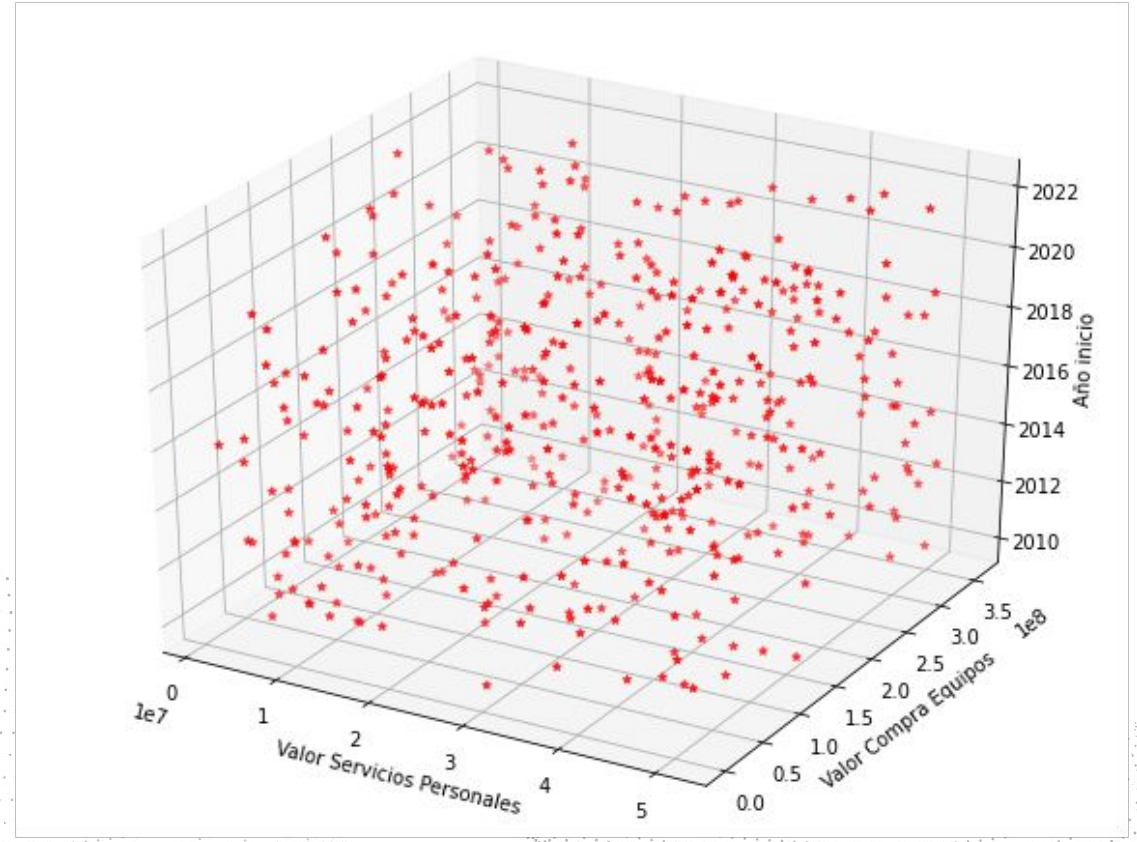


Figura 7. Muestra de la aleatoriedad de los datos entre Año de inicio; Valor Servicios Personales; Valor de Compra Equipos.

Regresión Múltiple

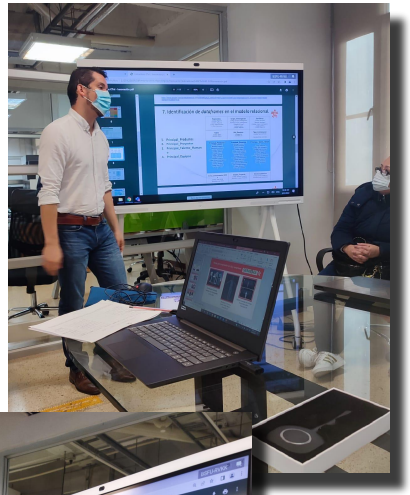
	coef	std err
Intercept	2014.7406	0.634
x1	1.287e-08	1.08e-08
x2	1.372e-09	1.5e-09
x3	2.709e-10	1.12e-09
x4	9.438e-09	2.54e-08
x5	1.444e-08	1.87e-08

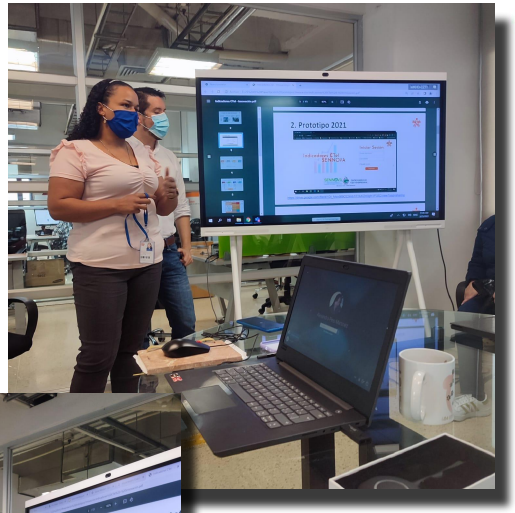
Tabla 6: Coeficientes de modelo lineal de regresión múltiple con desviación estándar

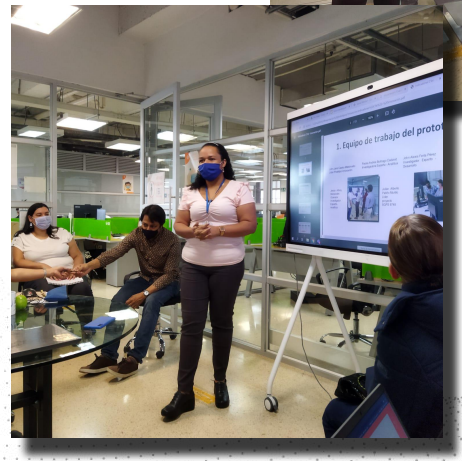
Características Generales

- Es posible encontrar una relación de gastos como modelos predictivos en años futuros, sin embargo al estar presente el componente aleatorio no se observa una tendencia definido como se observa en los valore de coeficientes lineales así
- Otra de las formas de identificar la aleatoriedad es la relación de los coeficientes y su desviación estándar que son equivalentes.

5. Actividades de divulgación







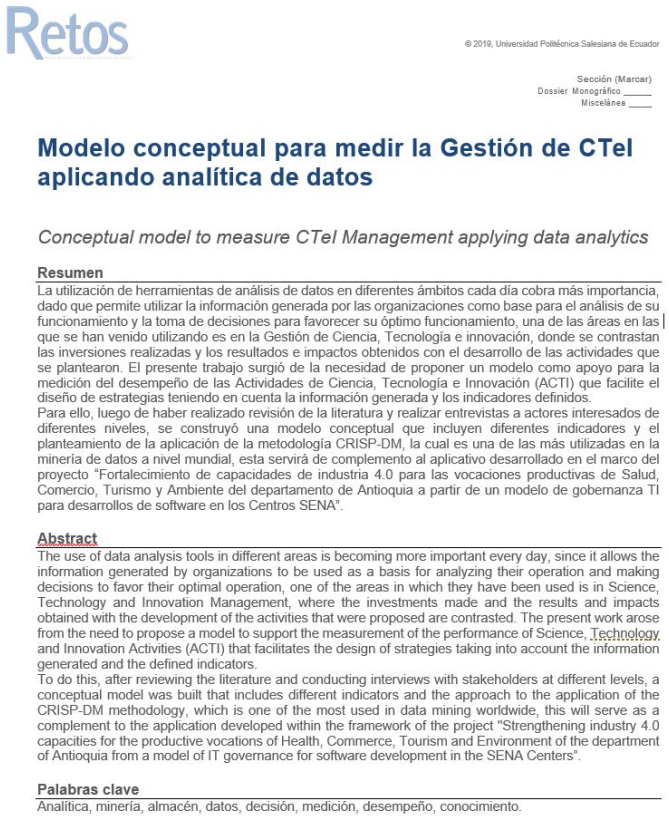
Presentación de avances - Dinamizadores SENNOVA

Actividades de Divulgación de resultados del prototipo

Universidad Politécnica
Salesiana de Ecuador

Envío de artículos hasta: 31/07/2022
Fecha de publicación: 01/10/2022

https://bit.ly/37mojoa



Retos

© 2019, Universidad Politécnica Salesiana de Ecuador

Sección (Marcar)
Dossier Monográfico ____
Miscelánea ____

Modelo conceptual para medir la Gestión de CTeI aplicando analítica de datos

Conceptual model to measure CTeI Management applying data analytics

Resumen

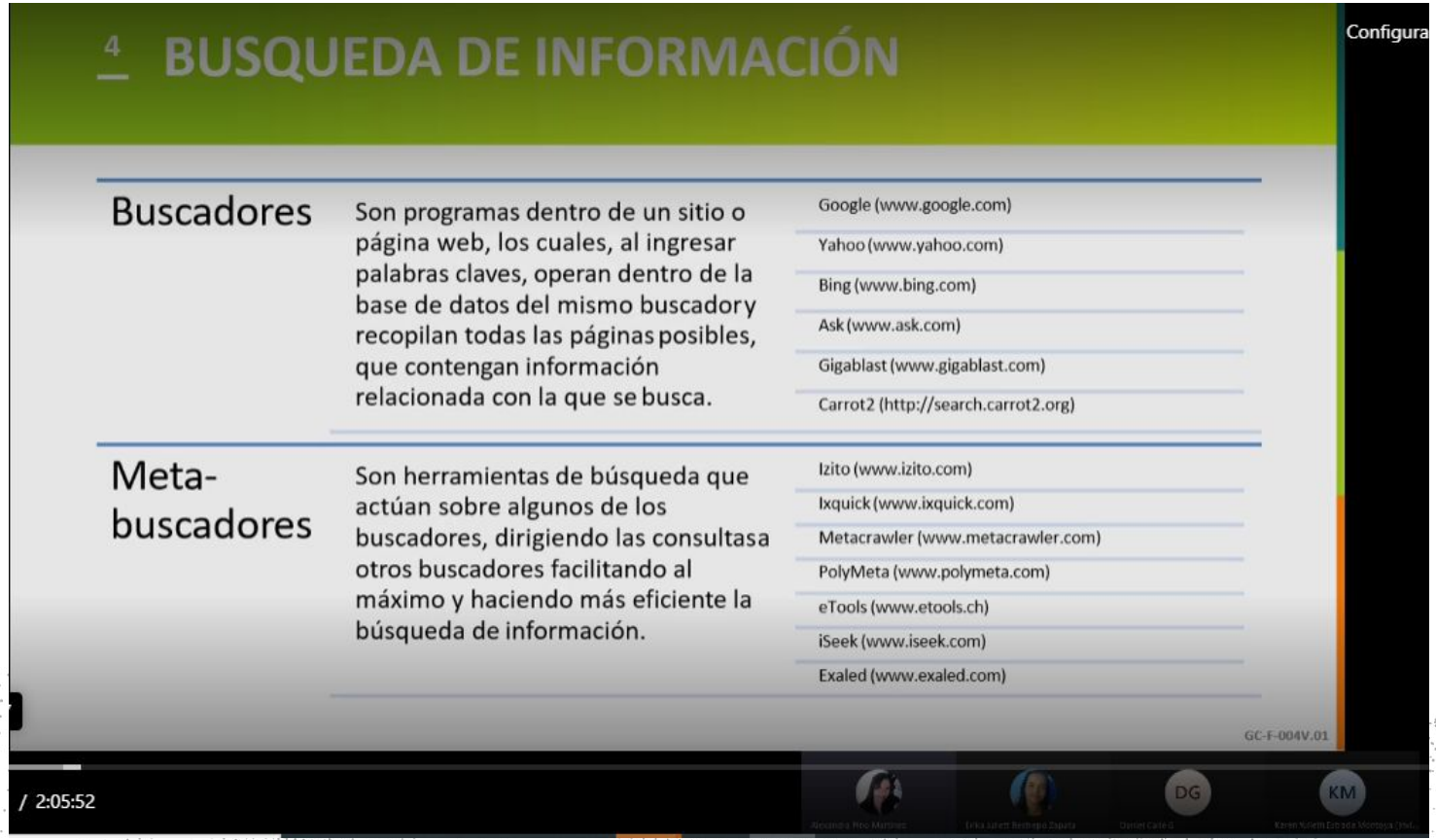
La utilización de herramientas de análisis de datos en diferentes ámbitos cada día cobra más importancia, dado que permite utilizar la información generada por las organizaciones como base para el análisis de su funcionamiento y la toma de decisiones para favorecer su óptimo funcionamiento, una de las áreas en las que se han venido utilizando es en la Gestión de Ciencia, Tecnología e innovación, donde se contrastan las inversiones realizadas y los resultados e impactos obtenidos con el desarrollo de las actividades que se plantearon. El presente trabajo surgió de la necesidad de proponer un modelo como apoyo para la medición del desempeño de las Actividades de Ciencia, Tecnología e Innovación (ACTI) que facilite el diseño de estrategias teniendo en cuenta la información generada y los indicadores definidos.
Para ello, luego de haber realizado revisión de la literatura y realizar entrevistas a actores interesados de diferentes niveles, se construyó una modelo conceptual que incluyen diferentes indicadores y el planteamiento de la aplicación de la metodología CRISP-DM, la cual es una de las más utilizadas en la minería de datos a nivel mundial, esta servirá de complemento al aplicativo desarrollado en el marco del proyecto "Fortalecimiento de capacidades de industria 4.0 para las vocaciones productivas de Salud, Comercio, Turismo y Ambiente del departamento de Antioquia a partir de un modelo de gobernanza TI para desarrollos de software en los Centros SENA".

Abstract

The use of data analysis tools in different areas is becoming more important every day, since it allows the information generated by organizations to be used as a basis for analyzing their operation and making decisions to favor their optimal operation, one of the areas in which they have been used is in Science, Technology and Innovation Management, where the investments made and the results and impacts obtained with the development of the activities that were proposed are contrasted. The present work arose from the need to propose a model to support the measurement of the performance of Science, Technology and Innovation Activities (ACTI) that facilitates the design of strategies taking into account the information generated and the defined indicators.
To do this, after reviewing the literature and conducting interviews with stakeholders at different levels, a conceptual model was built that includes different indicators and the approach to the application of the CRISP-DM methodology, which is one of the most used in data mining worldwide, this will serve as a complement to the application developed within the framework of the project "Strengthening industry 4.0 capacities for the productive vocations of Health, Commerce, Tourism and Environment of the department of Antioquia from a model of IT governance for software development in the SENA Centers".

Palabras clave

Analítica, minería, almacén, datos, decisión, medición, desempeño, conocimiento.



Transferencia de conocimientos a aprendices - monitores de excelencia SENNOVA

G R A C I A S

Línea de atención al ciudadano: 018000 910270
Línea de atención al empresario: 018000 910682

@SENAcomunica

www.sena.edu.co

ANEXOS 5. Actas de avance abril 2022

ACTA No. 22		
NOMBRE DEL COMITÉ O DE LA REUNIÓN: Seguimiento Proyecto SGPS 9748 - Prototipo Innovación		
CIUDAD Y FECHA: Medellín, 23 de marzo de 2022	**HORA INICIO:** 10:30 a.m.	**HORA FIN:** 1:15 p.m.
LUGAR Y/O ENLACE: Sala de juntas Piso 9	**DIRECCIÓN GENERAL / REGIONAL / CENTRO** Centro de Servicios y Gestión Empresarial	

AGENDA O PUNTOS PARA DESARROLLAR:

1. Revisión presentación para reunión de avances con los Dinamizadores SENNOVA
2. Avances de las actividades del prototipo

OBJETIVO(S) DE LA REUNIÓN:

Realizar seguimiento Proyecto SGPS 9748 Prototipo Innovación - Empalme Plataforma Software.

DESARROLLO DE LA REUNIÓN

1. Revisión presentación para reunión de avances con los Dinamizadores SENNOVA

Se socializó la presentación elaborada para la reunión con los dinamizadores y demás integrantes del proyecto 9748, la idea es realizar una contextualización del trabaja realizado el año pasado y compartir con los asistentes los avances de este año, para ello se utilizará la presentación y el video que se diseño finalizando el 2021 con el funcionamiento del prototipo. Se definió la distribución de la presentación y la información que socializará casa integrante del equipo.

2. Avances de las actividades del prototipo

Se presentó el algoritmo con el que se va a generar la información complementaria para la base de datos con la que se van a aplicar los modelos de analítica de datos. Se sugiere aprovechar este algoritmo para generar al menos 500 datos por cada variable.

Se presentaron los avances de la programación que se está realizando en Phyton y los resultados que se han obtenido hasta el momento con la regresión múltiple y se socializaron otros métodos que podrían ser utilizados.

Se sugirió tener en cuenta varios modelos tanto para modelos supervisados como no supervisados y establecer la forma en la que se pueden utilizar en el prototipo

CONCLUSIONES

Se continúa avanzando con la aplicación de la metodología CRISP – DM para diseñar el modelo de analítica de datos y la identificación de los modelos de analítica de datos a utilizar en el prototipo.

Se acordó la forma de presentar los avances en la reunión con los dinamizadores SENNOVA el próximo 25 de marzo.

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:

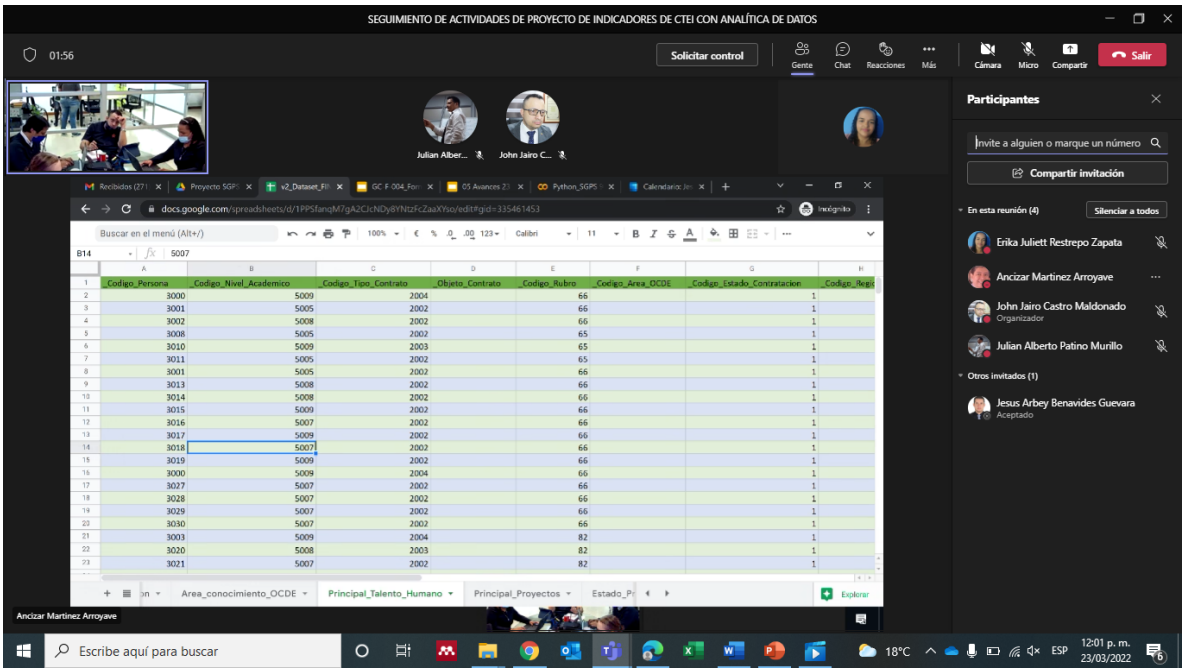
PÚBLICA ☐ **PRIVADO** **SEMIPRIVADO** **SENSIBLE**

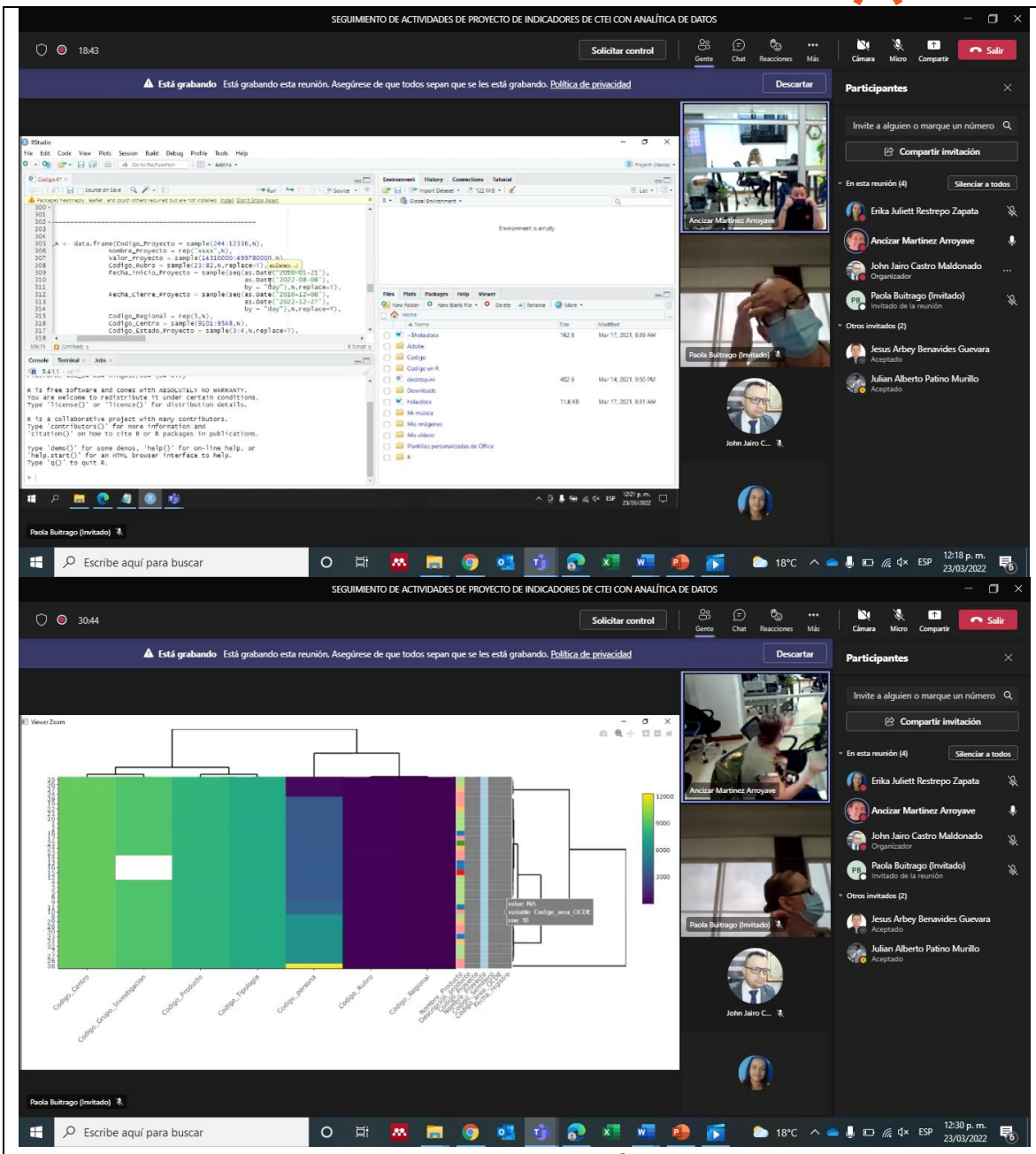
Nota: Antes de contestar esta informacional por favor remitirse al instructivo

COMPROMISOS

ACTIVIDAD	RESPONSABLE	FECHA
Identificar los modelos de analítica a utilizar en el prototipo	Jesús Arbey Benavides Paola Betancur	30 de marzo de 2022

ASISTENTES: (Incorporar registro de asistencia)





Enlace grabación:

https://sena4-my.sharepoint.com/personal/jcastrom_sena_edu_co/_layouts/15/onedrive.aspx?id=%2Fpersonal%2Fjcastrom%5Fsena%5Fedu%5Fco%2FDocuments%2FRecordings%2FSEGUIMIENTO%20DE%20ACTIVIDADES%20DE%20PROYECTO%20DE%20INDICADORES%20DE%20CTEI%20CON%20ANAL%C3%8DTICA%20DE%20DATOS%2D20220309%5F192131%2DMeeting%20Recording%2Emp4&parent=%2Fpersonal%2Fjcastrom%5Fsena%5Fedu%5Fco%2FDocuments%2FRecordings

REGISTRO DE ASISTENCIA Y APROBACIÓN DEL ACTA No- DEL DÍA 23 DEL MES DE marzo DEL AÑO 2022

OBJETIVO (S): Realizar seguimiento a la ejecución de las actividades proyecto SGPS9748-Prototipo Innovación

No.	NOMBRES Y APELLIDOS	No. DOCUMENTO	PLANTA	CONTRATISTA	OTRO ¿CUAL?	DEPENDENCIA/ EMPRESA	CORREO ELECTRÓNICO	TELÉFONO/EXT SENA	AUTORIZA GRABACIÓN	FIRMA O PARTICIPACIÓN VIRTUAL
1	John Jairo Castro H.	88265512	X			SENA CESCE	jjcastroh@sena.edu.co	42850	SC	
2	Jesus Arley Bonilla	1085260713		X		SENA	jabonillas@sena.edu.co	3105595765	Si	
3	Erika Juliett Restrepo	43973982		X		SENA	erestrepoz@sena.edu	3005017252	Si	
4	Julian Peñuela	8.102.363		X		Asesor/ Consejo	jpenuela@sena.edu.co	3013004624	Si	
5	Paola Andrea Burbano	43641579		X		SENA	pabuitrago@sena.edu.co	3113671040	Si	

De acuerdo con La Ley 1581 de 2012, Protección de Datos Personales, el Servicio Nacional de Aprendizaje SENA, se compromete a garantizar la seguridad y protección de los datos personales que se encuentran almacenados en este documento, y les dará el tratamiento correspondiente en cumplimiento de lo establecido legalmente.

ACTA No. 45		
NOMBRE DEL COMITÉ O DE LA REUNIÓN: Seguimiento Proyecto SGPS 9748 - Prototipo Innovación		
CIUDAD Y FECHA: Medellín, 06 de abril de 2022	**HORA INICIO:** 11:00 a.m.	**HORA FIN:** 1:30 p.m.
LUGAR Y/O ENLACE: Laboratorio de Inteligencia Artificial Piso 5 Virtual TEAMS https://bit.ly/35L9Hyb	**DIRECCIÓN GENERAL / REGIONAL / CENTRO** Centro de Servicios y Gestión Empresarial	

AGENDA O PUNTOS PARA DESARROLLAR:

1. Propuesta del desarrollador John Ferla
2. Avances de las actividades

OBJETIVO(S) DE LA REUNIÓN:

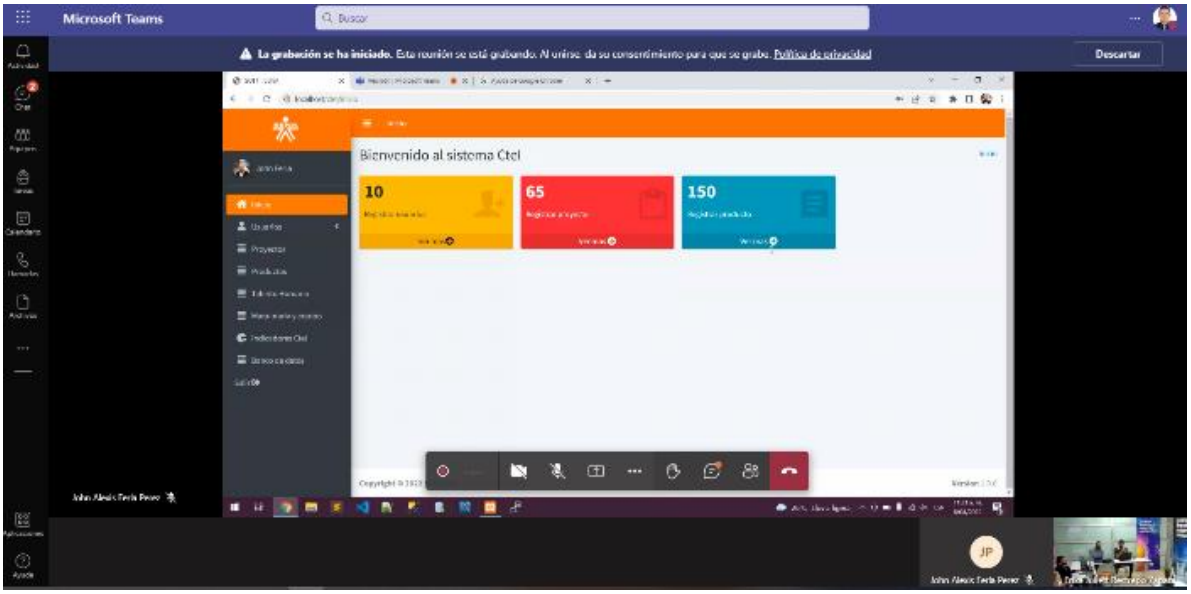
Realizar seguimiento Proyecto SGPS 9748 Prototipo Innovación - Empalme Plataforma Software.

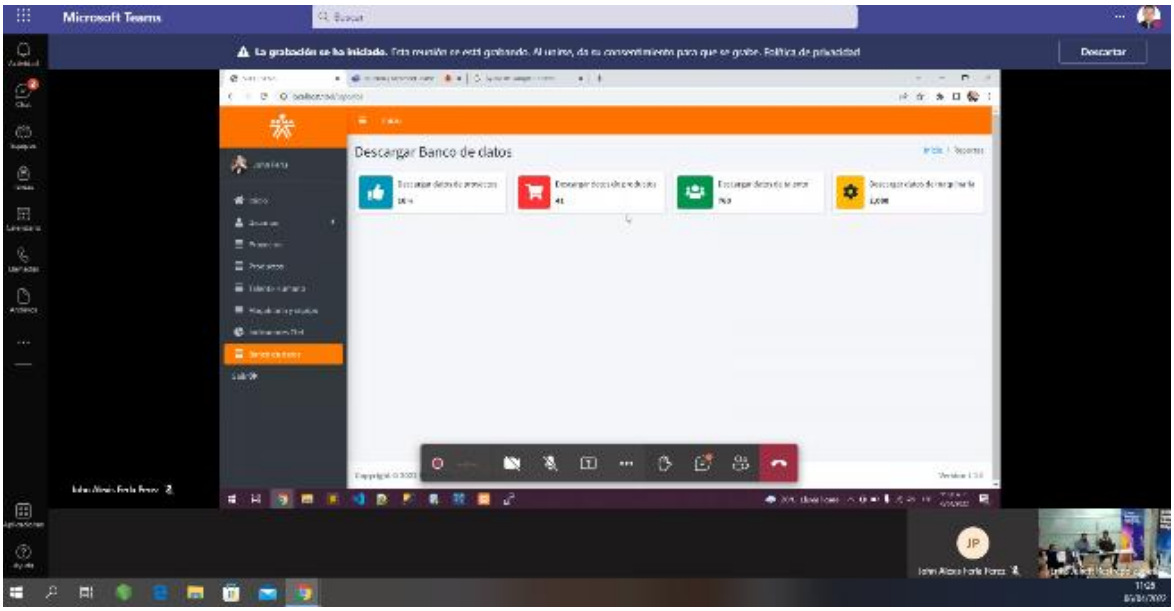
DESARROLLO DE LA REUNIÓN

1. Propuesta del desarrollador John Ferla

Para iniciar la reunión el desarrollador Jon Ferla socializó su propuesta para el aplicativo, donde considera que no es necesario utilizar Power Bi para realizar las gráficas, ya que según su experiencia se pueden realizar desde la base de datos directamente.

Adicionalmente, presentó algunos ajustes de visualización que le realizó al aplicativo, los cuales fueron aceptados por el equipo de trabajo, como que el administrador pueda crear o eliminar usuarios para que no dependa del desarrollador, cambios en la visualización y los gráficos. Se sugiere adicionar pestaña de visualizaciones y otras de gráficas para las visualizaciones de Phyton.





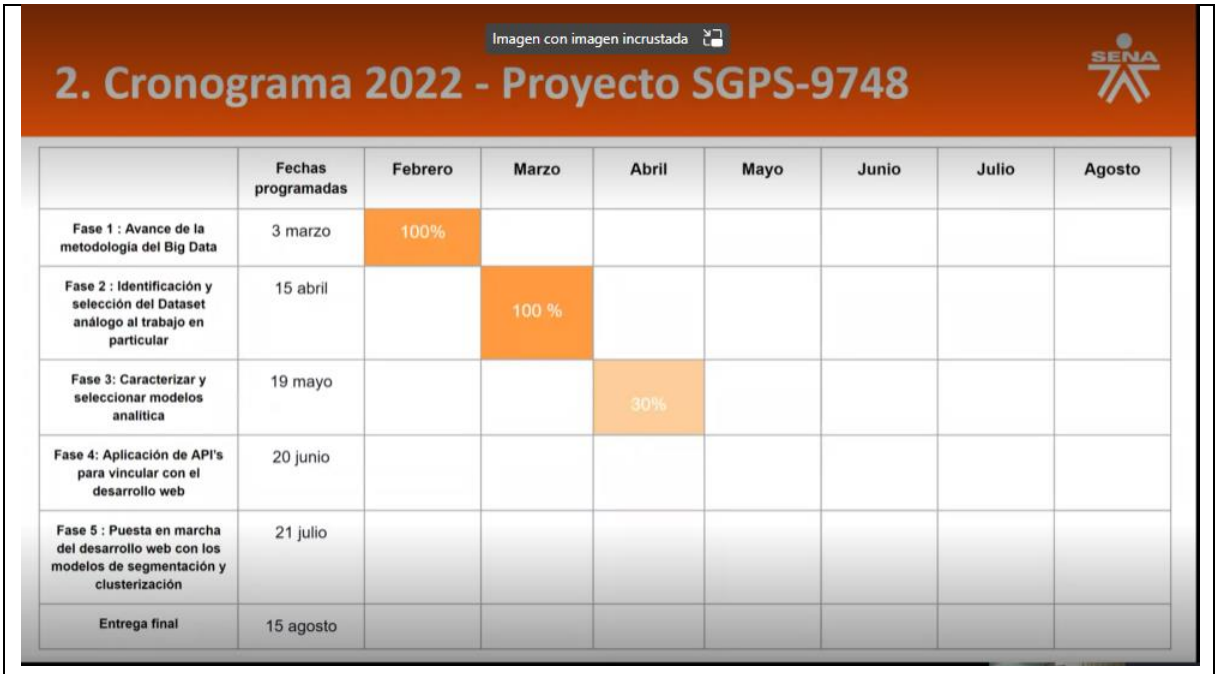
El desarrollador va a verificar hasta que nivel de detalle se puede llegar con los gráficos, teniendo en cuenta que el fuerte del desarrollo es la analítica predictiva y realizará la actualización del modelo relacional.

Los investigadores Jesús Arbey Benavides y Paola Buitrago le enviarán al Desarrollador información del código, el dataframe y la información de los gráficos que se requieren incluir en el prototipo.

2. Avances del modelo de analítica

En la aplicación de la metodología CRISP-DM se está en la etapa de preparación de los datos y modelado. Según el cronograma establecido al inicio de la ejecución del prototipo se está cumpliendo de manera adecuada, incluso un poco adelantados.

Se organizaron los indicadores para que se pueda actualizar el modelo teórico y el relacional del prototipo. Adicionalmente, se generó un algoritmo para obtener 500 reglones de información aleatorios en la base de datos.

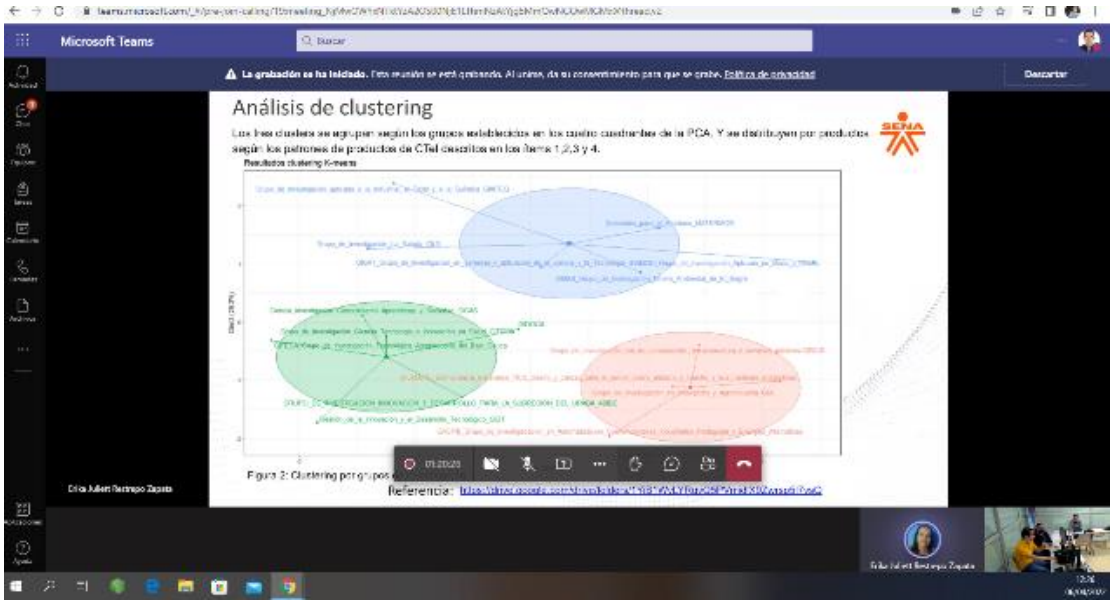


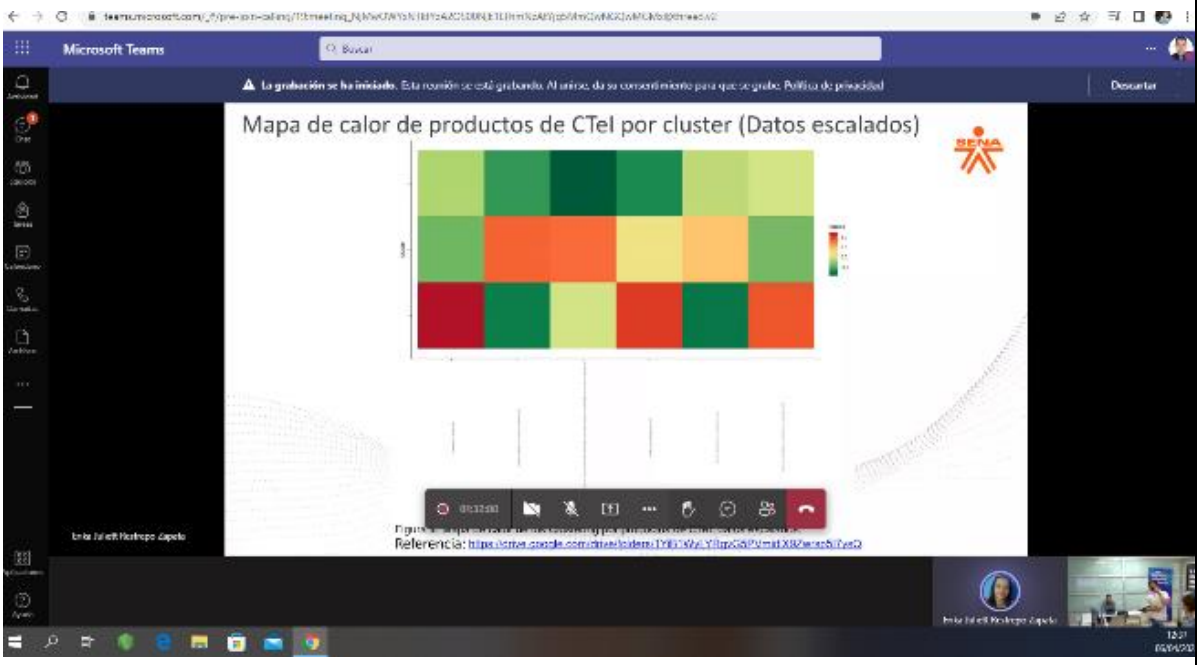
	Fechas programadas	Febrero	Marzo	Abril	Mayo	Junio	Julio	Agosto
Fase 1 : Avance de la metodología del Big Data	3 marzo	100%						
Fase 2 : Identificación y selección del Dataset análogo al trabajo en particular	15 abril		100 %					
Fase 3: Caracterizar y seleccionar modelos analítica	19 mayo			30%				
Fase 4: Aplicación de API's para vincular con el desarrollo web	20 junio							
Fase 5 : Puesta en marcha del desarrollo web con los modelos de segmentación y clusterización	21 julio							
Entrega final	15 agosto							

Además se socializaron los dataframes sobre los cuales se priorizó realizar la analítica de datos

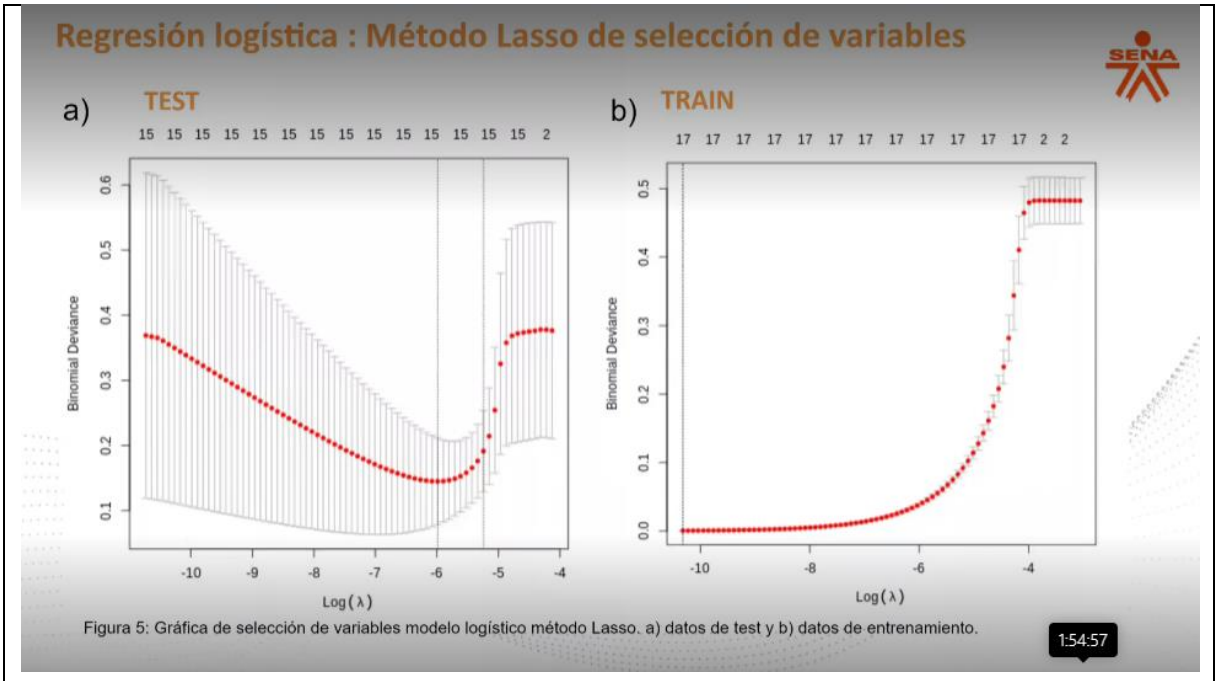
Los modelos de analítica a utilizar se dividen en dos los no supervisados, de los cuales se van a trabajar Análisis de Componentes Principales, Clustering – Kmeans, se sugiere incluir KNN y Árboles de Decisión, dentro de los supervisados se van a trabajar Regresión Logística y múltiple, Support Vector Machine, Árboles de Decisión se sugiere ampliar con otros modelos.

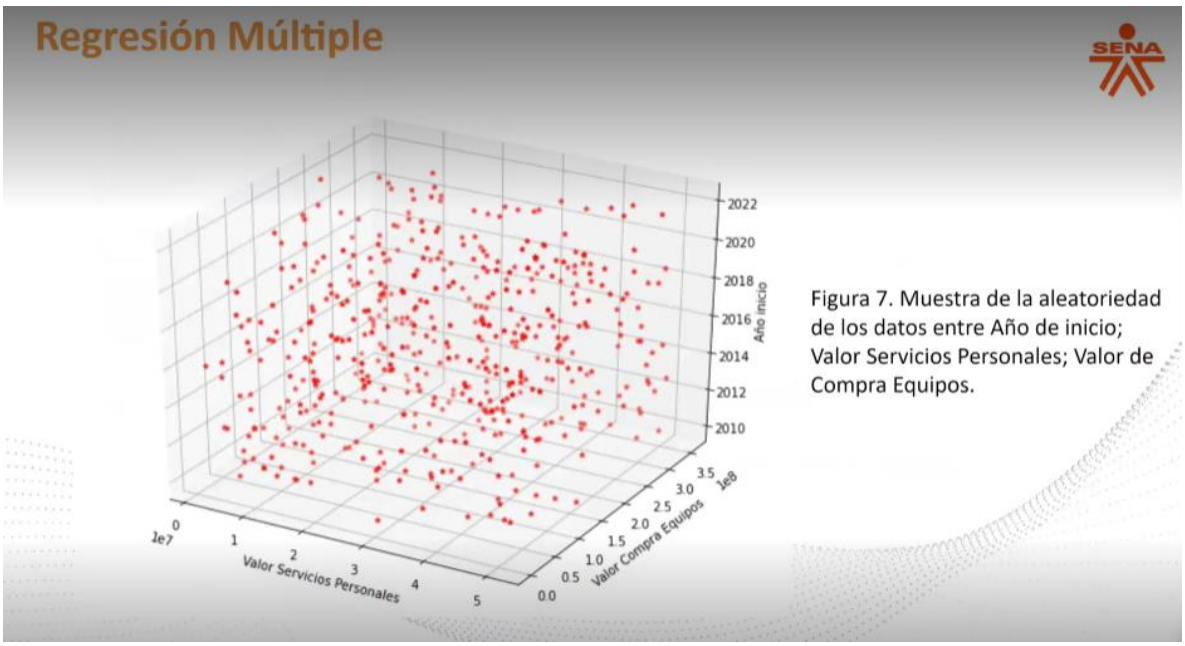
Se presentaron gráficos de componentes principales y círculo de correlación, se explicaron y analizaron los resultados obtenidos. Igualmente se presentó un análisis de Clustering y el dendograma obtenido con la información y se analizó la información presentada en ellos. También se analizó un mapa de calor obtenido analizando los grupos de investigación, el cual fue considerado interesante para la comparación de los grupos de investigación y establecer rutas de trabajo para fortalecerlos.

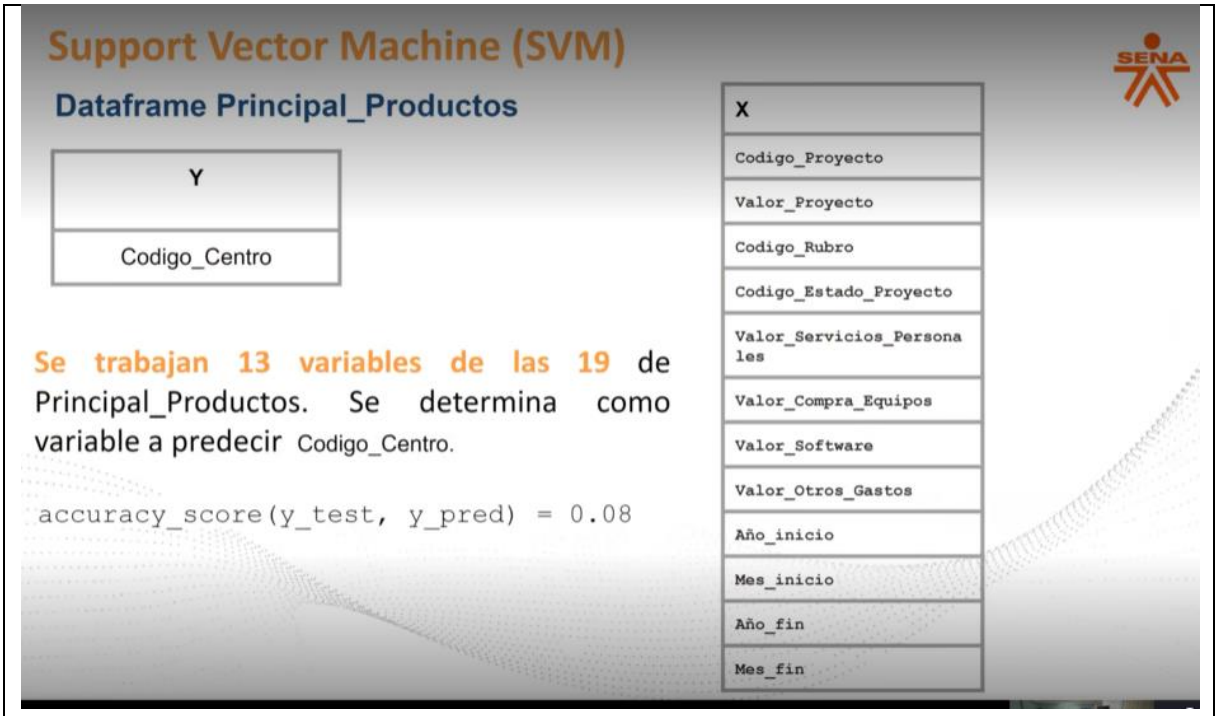




Además de lo anterior, se realizaron regresiones logísticas y múltiples teniendo en cuenta la información de los productos, también se presentó el trabajo realizado con la Support Vector Machine.


Regresión logística : Método Lasso de selección de variables
a)
TEST
b)
TRAIN
Binomial Deviance
Log(λ)
Figura 5: Gráfica de selección de variables modelo logístico método Lasso. a) datos de test y b) datos de entrenamiento.
1:54:57


Regresión Múltiple
Año inicio
Valor Servicios Personales
Valor Compra Equipos
Figura 7. Muestra de la aleatoriedad de los datos entre Año de inicio; Valor Servicios Personales; Valor de Compra Equipos.



Se realizó presentación ante los dinamizadores SENNOVA de la regional Antioquia y demás compañeros del proyecto 9748 para presentar los avances del prototipo, se obtuvieron muy buenos comentarios y de esa reunión hasta el momento se evidencian avances muy significativos.

Se trabajó en el artículo de investigación que se va a enviar a más tardar este mes, se enviará a revisión para aportes inicialmente de Julián Patiño, se elaboraron los documentos solicitados.

Se realizó una trasferencia de conocimiento a los monitores asignados a SENNOVA, se está trabajando con la Karen Estrada, monitora de excelencia para su preparación de la participación en Redcolsi.

Por último, se les solicitó a los investigadores Paola Buitrago y Jesús Arbey Benavides que vayan planteando el enfoque del artículo que aportarán como producto de la investigación, se mencionó que se cuenta con recursos para los posibles gastos en los que se puedan incurrir para la publicación.

CONCLUSIONES

Se continúa avanzando con la aplicación de la metodología CRISP – DM para diseñar el modelo de analítica de datos.

Con la colección de las herramientas utilizadas se acordó que es importante divulgar lo que se hizo, las lecciones aprendidas y las limitaciones que se presentaron.

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:

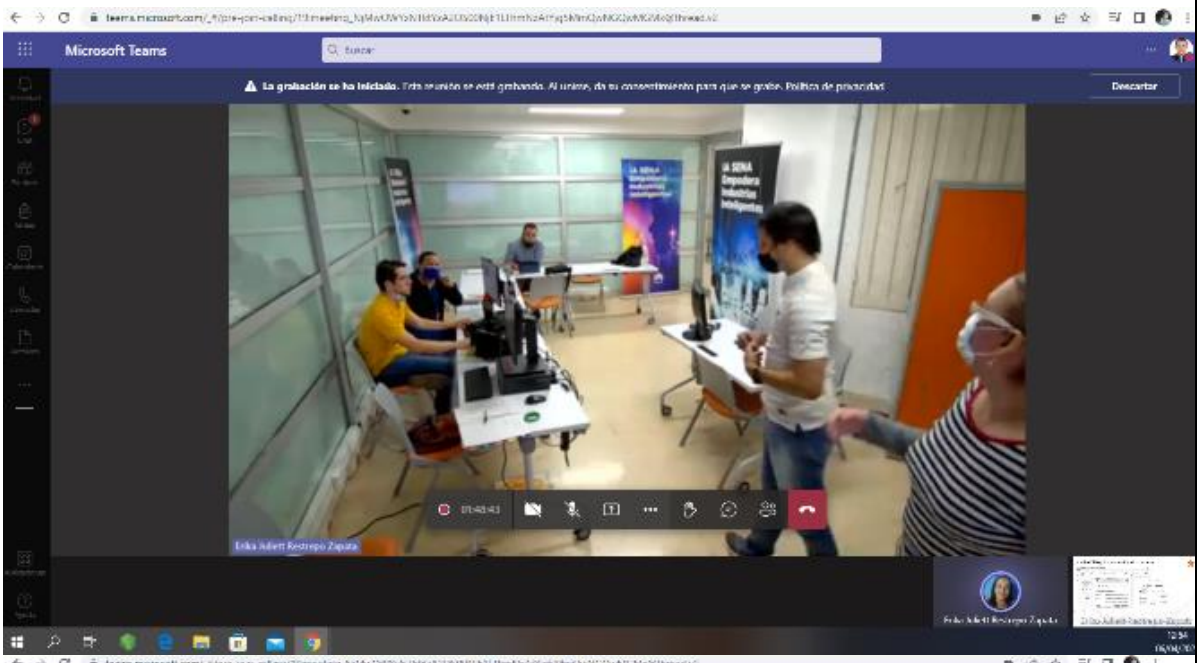
La información de este documento se debe clasificar como:

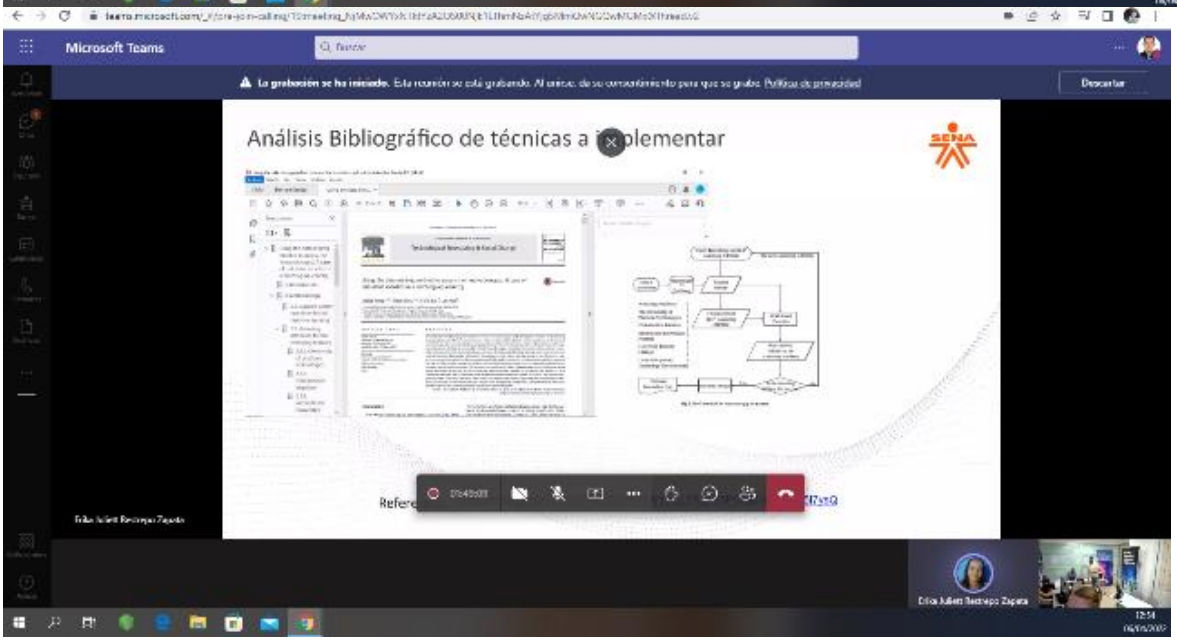
PÚBLICA ☐ **PRIVADO** ☐ **SEMIPRIVADO** ☐ **SENSIBLE** ☐

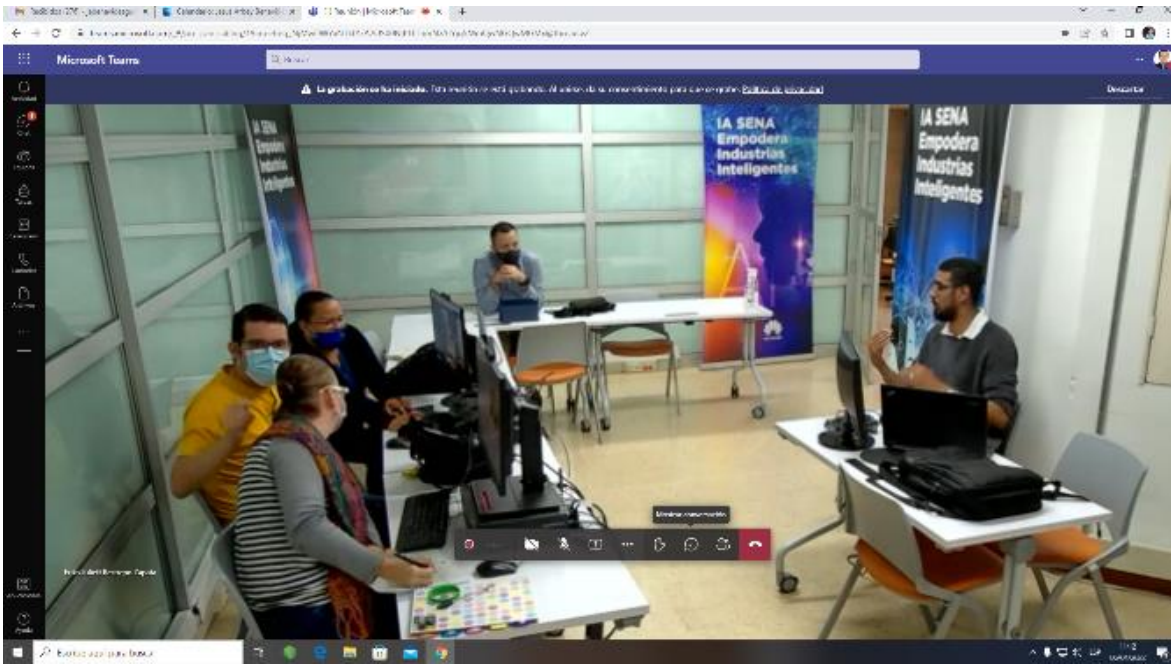
Nota: Antes de contestar esta informacional por favor remitirse al instructivo

COMPROMISOS		
ACTIVIDAD	**RESPONSABLE**	**FECHA**
Enviar información solicitada al desarrollador John Ferla	Jesús Arbey Benavides Paola Buitrago	18 de abril de 2022

ASISTENTES: (Incorporar registro de asistencia)






Microsoft Teams
IA SENA
Empodera
Industrias
Inteligentes

Enlace grabación:
https://sena4-my.sharepoint.com/personal/erestrepoz_sena_edu_co/_layouts/15/onedrive.aspx?id=%2Fpersonal%2Ferestrepoz%5Fsena%5Fedu%5Fco%2FDocuments%2FRecordings%2FSEGUIMIENTO%20DE%20ACTIVIDADES%20DE%20PROYECTO%20DE%20INDICADORES%20DE%20CTEI%20CON%20ANAL%C3%8DTICA%20DE%20DATOS%2D20220406%5F110704%2DGrabaci%C3%B3n%20de%20la%20reuni%C3%B3n%2Emp4&parent=%2Fpersonal%2Ferestrepoz%5Fsena%5Fedu%5Fco%2FDocuments%2FRecordings&ga=1

SENA

REGISTRO DE ASISTENCIA Y APROBACIÓN DEL ACTA No- DEL DÍA 06 DEL MES DE abril DEL AÑO 2022.

OBJETIVO (S): Reunión seguimiento proyecto 9748 - Prototipo innovación.

No.	NOMBRES Y APELLIDOS	No. DOCUMENTO	PLANTA	CONTRATISTA	OTRO ¿CUAL?	DEPENDENCIA/ EMPRESA	CORREO ELECTRÓNICO	TELÉFONO/EXT.SENA	AUTORIZA GRABACIÓN	FIRMA O PARTICIPACIÓN VIRTUAL
1	Jesus Albey Benavides Guevara	1085260713		X		SENA	jubenavidesg sena.edu.co	3105545765	Si	Jesus Albey B.
2	Paola Andrea Buitrago	43641579		X		SENA	pbuitragoc@sena.edu.co	3113671040	Si	Paola Buitrago
3	John Ferb	1036621674		X		SENA	Jaferb@sena-edu.co	3146664277	Si	
4	John Jairo Castro H.	88255552	X			SENA CESGE	jcastrom@sena.edu.co	42850	Si	
5	Erika Juliett Restrepo Z.	43973982		X		SENA CESGE	erestrepoz@sena.edu.co	3005017253	Si	
6	Julián Patiño	8102363		X		Sena/ Cesgt	jpatino@sena.edu.co	3013004624	Si	

De acuerdo con La Ley 1581 de 2012, Protección de Datos Personales, el Servicio Nacional de Aprendizaje SENA, se compromete a garantizar la seguridad y protección de los datos personales que se encuentran almacenados en este documento, y les dará el tratamiento correspondiente en cumplimiento de lo establecido legalmente.