

## Entrega 2 Proyecto DSA

### Problema a abordar y contexto

En el núcleo de nuestra misión de transformar vidas, se encuentra Proyección Infantil SOS (PI), una ONG dedicada a brindar un hogar con futuro a niños y jóvenes separados de sus familias. Los tiempos cambian, y esta organización enfrenta desafíos cruciales: cambios en el comportamiento de los donantes, la búsqueda de la autosostenibilidad y la necesidad de alinear estrategias. En este proyecto, exploraremos cómo la tecnología y el aprendizaje automático pueden ayudar a Proyección Infantil a definir los nuevos perfiles de donantes y personalizar estrategias para retener a aquellos que hacen posible su noble labor.

### Pregunta de negocio y alcance del proyecto

PI busca responder la siguiente pregunta clave: "¿Cuáles son los nuevos perfiles de donantes teniendo en cuenta las variables externas e internas que afectan su comportamiento a partir de técnicas de Machine Learning?". Para resolver esta pregunta, se buscará desarrollar un modelo de segmentación que ayude a alinear correctamente las estrategias diseñadas dentro de la organización para tener un mejor aprovechamiento de los recursos y llegar a ser costo-eficientes en todos los esfuerzos que se ejecutan. Para desarrollar el modelo de segmentación, primero se debe calcular el "Customer Lifetime Value" o "Donor Lifetime Value" (DLTV) para lograr tener una segmentación basada en el retorno o valor que genera cada uno de los donantes. Posteriormente, según su bajo, medio o alto valor, lo que se quiere es identificar las características principales que representan cada segmento además de su valor con el fin de proponerle a la organización un perfil de donante en el que se deberían enfocar en buscar. Finalmente, se procede a cruzar cada uno de estos segmentos identificados contra el comportamiento de variables como la TRM, o el desempleo para identificar si hay algún patrón o relación entre el promedio de donación y estas variables macroeconómicas para que se puedan anticipar con sus estrategias y minimizar el impacto si es que el comportamiento de los donantes es disminuir su aporte. Todo lo anterior se mostrará en un tablero de control que permita la fácil identificación de estos segmentos y de las estrategias que debo aplicar a cada grupo de donantes.

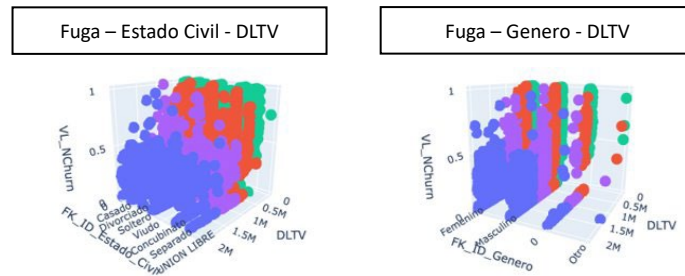
### Descripción de conjuntos de datos a emplear

Dentro de los datos suministrados por Proyección Infantil SOS, se cuenta con dos tablas principales, Base de Donantes y Transacciones Donantes Individuales. La primera contiene información sobre todos los donantes, mientras que la segunda contiene información sobre las transacciones. La base de datos de donantes cuenta con 99.594 filas y 19 variables. La base de datos de transacciones contiene por su parte 1.348.521 observaciones y 12 variables. Para la limpieza de esta información, se eliminaron las observaciones con error en fecha de aniversario de pago, se eliminaron menores de edad y mayores de 90 años, se ajustaron todas las fechas en su formato correspondiente y se corrigieron las probabilidades de Churn y Lapsed, dividiendo entre 100 en los casos en los que se asignaron valores mayores a 1. La siguiente tabla muestra algunas descriptivas para la base de datos depurada. Adicionalmente, se describe el género y las edades de los donantes.

## Modelos desarrollados

### 1. Segmentación con k-prototypes

En este modelo se intentaron varias combinaciones de variables categóricas como la edad, estado civil y género, pero los resultados arrojaron que ninguna de estas variables es representativa para caracterizar los segmentos, pues solo el valor del donante (DLTV) es el que predominaba en la segmentación como podemos ver en las siguientes imágenes.



Instancia lanzada:


The screenshot shows the AWS Management Console for an EC2 instance. The instance is named 'i-0b8156cf4f4b836ab (Trabajo\_DSA\_S5)' and is in the 'En ejecución' (Running) state. The console displays various details about the instance, including its ID, public IP address (3.84.27.224), private IP address (172.31.85.237), DNS name (ec2-3-84-27-224.compute-1.amazonaws.com), and the VPC (vpc-00e6c4b68c8b72). The instance is running on a t2.medium instance type with a subnet of subnet-06296b6b46d0cb4c.

Experimentos en MLFlow:

The screenshot shows the MLFlow Experiments interface. The experiment is named 'Experimento\_k\_prototype' and has a description of 'metrics.rmse < 1 and params.model = "tree"'. The interface displays a table of runs, with two runs listed: 'inductive-midge-269' and 'bemused-Res-499'. Both runs have a duration of 3.1min and are associated with the source file 'Experimento\_k\_prototype2.py' and the model 'sklearn'.

Run Name	Created	Dataset	Duration	Source	Models
inductive-midge-269	21 minutes ago	-	3.1min	Experimento_k_prototype2.py	sklearn
bemused-Res-499	35 minutes ago	-	3.1min	Experimento_k_prototype2.py	sklearn

Archivos cargados a Github:

**Proyecto\_DSA** Public

forked from [ggomez1803/Proyecto\\_DSA](#)

main

1 branch

0 tags

Go to file


Add file










Code

This branch is up to date with ggomez1803/Proyecto\_DSA:main.

Contribute

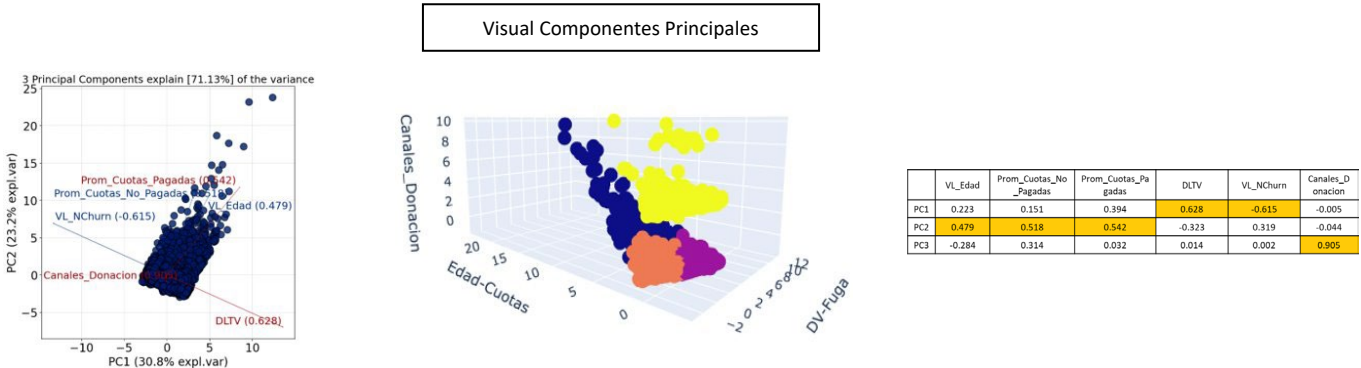
Sync fork

 **ggomez1803** Add files via upload 8105e07 4 minutes ago 47 commits

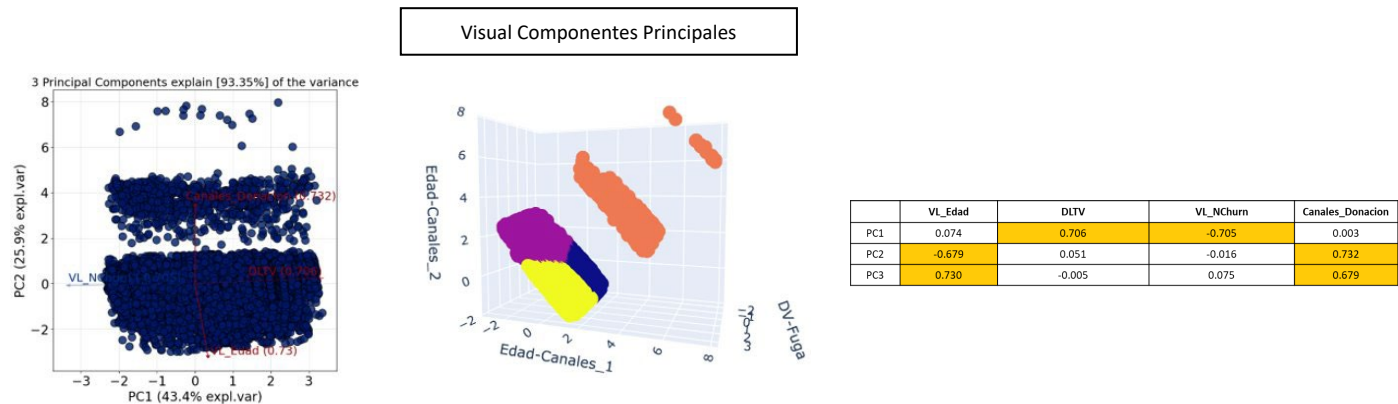
	.gitignore	Flujo preprocesamiento de datos, rutas y gitignore	2 weeks ago
	DLTV.py	Add files via upload	yesterday
	Experimentacion (1).ipynb	experimentacion PCA	50 minutes ago
	Experimentacion PCA.py	experimentacion PCA	50 minutes ago
	Experimento_cluster_jerarquico.py	Add files via upload	4 hours ago
	Experimento_kmeans.py	Add files via upload	4 minutes ago
	Exploracion_de_Datos.ipynb	Add files via upload	2 weeks ago
	Funciones_DLTV.py	Add files via upload	yesterday
	Funciones_preprocesamiento.py	funciones para corregir ciudades y transacciones	2 weeks ago

## 2. PCA para reducción de dimensionalidad y K-means

En el análisis de componentes principales, inicialmente se usaron 6 variables para reducirlas a tres componentes principales, con el fin de ayudar a mejorar la interpretabilidad de los datos, estos tres componentes explicaron el 71% de la varianza. Sin embargo, dentro de las variables utilizadas se encontraban el promedio de cuotas pagadas y no pagadas anuales donde se evidenciaron datos sin sentido en análisis previos.



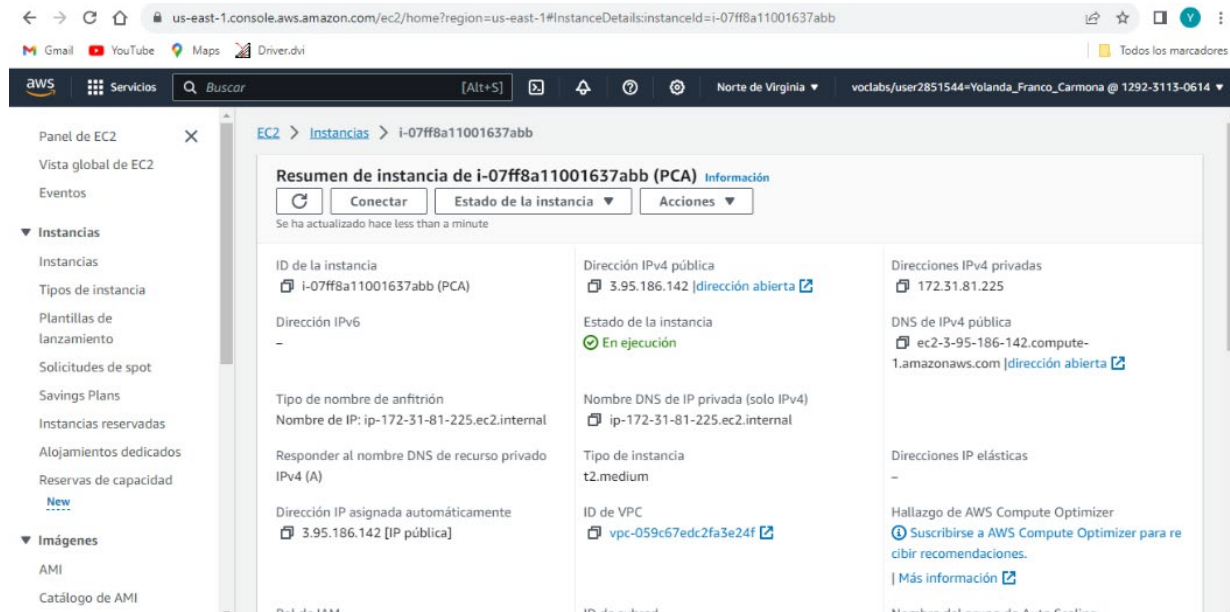
Se hizo un segundo intento con cuatro variables reduciéndolas a tres componentes y explicando el 93% de la varianza, pero los componentes dos y tres eran una combinación lineal de las mismas variables, por lo que se procede a reducirlo a dos componentes principales, explicando el 69% de la varianza. Dentro de este último modelo la variable de canales de donación no parece representar o ser parte de algún perfil específico dentro de los segmentos, por lo que se decidió quitarla y usar solamente k-means.



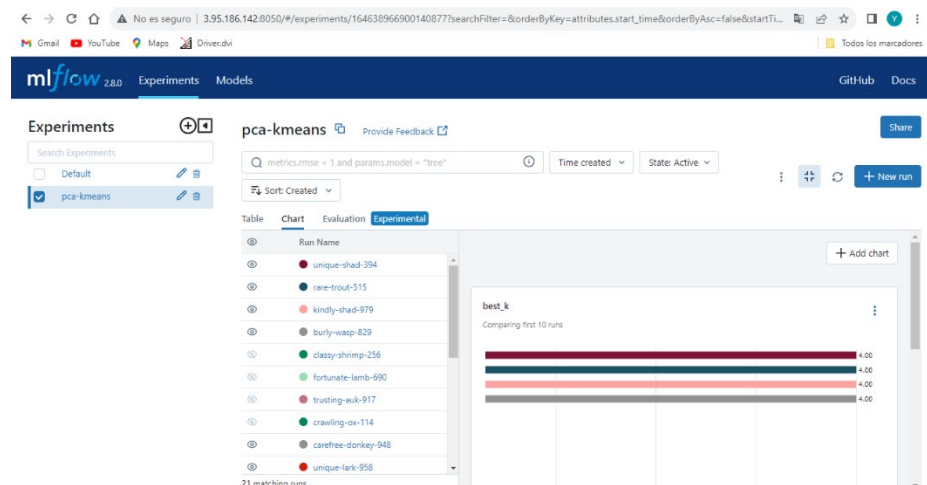
3. PCA para reducción de dimensionalidad y K-means con MIFlow

Se utilizó el control de versiones para verificar el número óptimo para el parámetro k y adicionalmente se calculó la métrica de silhouette\_score para establecer el grado de disimilaridad de los clusters creados. Al modificar el parámetro de número de componentes principales se obtuvo que en todos los casos el número óptimo de componentes era 4, por lo que el índice de silhouette dio igual a 0.48 en los modelos evaluados. Los resultados de los diferentes experimentos realizados se pueden ver en MIFlow

Instancia lanzada:



# Experimentos en MIFlow



## Archivos cargados a GitHub:

YolandaFrancoC Merge branch 'main' of https://github.com/ggomez1803/Proyecto_DSA	9865aa4 3 minutes ago	44 commits	Proyecto para Despliegue de Soluciones Analíticas
.gitignore	Flujo preprocesamiento de datos, rutas y gitignore	2 weeks ago	Readme
DLTV.py	Add files via upload	yesterday	Activity
Experimentacion (1).ipynb	experimentacion PCA	3 minutes ago	0 stars
Experimentacion PCA.py	experimentacion PCA	3 minutes ago	1 watching
Experimento_cluster_jerarquico.py	Add files via upload	3 hours ago	2 forks
Exploracion_de_Datos.ipynb	Add files via upload	2 weeks ago	Report repository
Funciones_DLTV.py	Add files via upload	yesterday	Releases
Funciones_preprocesamiento.py	funciones para corregir ciudades y transacciones	2 weeks ago	No releases published <a href="#">Create a new release</a>
Preprocesamiento_datos.py	Flujo preprocesamiento de datos, rutas y gitignore	2 weeks ago	Packages
README.md	readme2	2 weeks ago	No packages published <a href="#">Publish your first package</a>
Rutas.py	Flujo preprocesamiento de datos, rutas y gitignore	2 weeks ago	

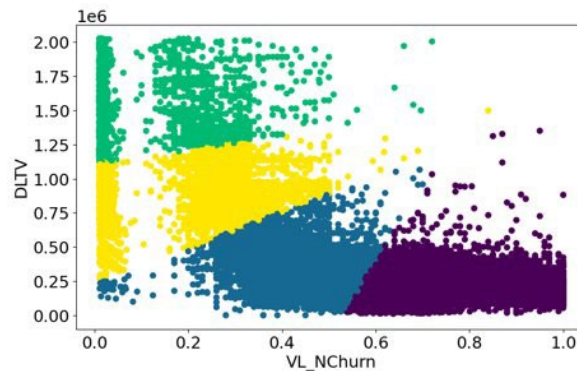
## K-means con diferentes grupos de variables

En este modelo se experimentó con diferentes variables y diferente número de segmentos. Inicialmente con variables como el DLTV, la edad y la probabilidad de fuga, sin embargo, durante varias corridas se evidenciaron que no existe una diferencia muy significativa entre clusters para la variable de edad, esto quiere decir que no es un eje de segmentación muy relevante y por esta razón se elimina de los ejes de segmentación para quedar finalmente con el valor del donante y su probabilidad de fuga obteniendo un resultado que hace sentido estratégico para alinear las estrategias actuales de la organización.

Los cuatro segmentos obtenidos son:

1. Segmento en fuga: donantes con bajo valor y alta probabilidad de fuga (Morado)
2. Segmento potencial: donantes con medio-bajo valor y medio-bajo probabilidad de fuga (Azul)
3. Segmento en crecimiento: donantes de medio valor y baja probabilidad de fuga (Amarillo)
4. Segmento estrella: donantes de alto valor y baja probabilidad de fuga (Verde)

Aprovechando la interpretabilidad de esta segmentación que se puede ver en la imagen a continuación, se pueden realizar recomendaciones coherentes sobre qué estrategias enfocar en cada grupo de donantes de la siguiente manera:



- Segmento estrella (Verde): Para estos donantes se pueden dirigir las estrategias más costosas con tal de seguir fidelizándolos tales como el programa de puntos y el club de beneficios
- Segmento en crecimiento (Amarillo): Se puede crear comunicaciones donde se muestre la labor realizada con las donaciones y exaltar su importancia para generar un estado aspiracional y filantrópico para que estos donantes lleguen a hacer parte del segmento de mayor valor
- Segmento potencial (Azul): Se recomienda promover las donaciones ya sea llamando a buscar un incremento de donación o crear nuevos canales para que donen
- Segmento en fuga (Morado): Se debe realizar estrategias de concientización y crear sentido de urgencia en la problemática para evitar que dejen de donar.

## Instancia lanzada

[Opción+S]

Norte de Virginia

voclabs/user2847452=Gabriel\_Gomez\_Monta\_o @ 7879-4040-3014

Instancias (1/1) Información

Conectar

Estado de la instancia

Acciones

Lanzar instancias

Buscar Instance por atributo o etiqueta (case-sensitive)

<input checked="" type="checkbox"/>	Name	ID de la instancia	Estado de la instancia	Tipo de inst...	Comprobación ...	Estado de la ...	Zona de dispon...
<input checked="" type="checkbox"/>	Proyecto_DSA	i-0b70941f292b443f3	En ejecución	t2.medium	Inicializando	Sin alarmas	us-east-1a

Instancia: i-0b70941f292b443f3 (Proyecto\_DSA)

Detalles

Seguridad

Redes

Almacenamiento

Comprobaciones de estado

Monitoreo

Etiquetas

Resumen de instancia Información

ID de la instancia

i-0b70941f292b443f3 (Proyecto\_DSA)

Dirección IPv6

-

Dirección IPv4 pública

3.83.143.68 [dirección abierta](#)

Estado de la instancia

En ejecución

Direcciones IPv4 privadas

172.31.90.253

DNS de IPv4 pública

ec2-3-83-143-68.compute-1.amazonaws.com [dirección abierta](#)

# Experimentos en MLFlow

mlflow 2.8.0

Experiments

Models

GitHub Docs

Experiments

Search Experiments

☐ k\_means

☒ exp\_kmeans

exp\_kmeans

Provide Feedback

Share

metrics.rmse < 1 and params.model = "tree"

Time created

State: Active

Sort: Created

Columns

Table

Chart

Evaluation

Experimental

		Metrics								
<input type="checkbox"/>		Run Name	Created	Duration	cluster_0_chur	cluster_0_cuo	cluster_0_cuo	cluster_0_dltv	cluster_0_edac	
<input type="checkbox"/>		enthused-ram-898	7 minutes ago	3.0s	0.0516518...	1.4410887...	5.5414700...	899672.59...	38.607129...	
<input type="checkbox"/>		calm-newt-161	8 minutes ago	3.0s	0.2088117...	1.5434240...	9.6836448...	861574.46...	43.869886...	
<input type="checkbox"/>		awesome-shoot-802	8 minutes ago	2.8s	0.3280594...	1.7560683...	10.319811...	574645.46...	43.796044...	
<input type="checkbox"/>		polite-grub-974	9 minutes ago	3.2s	0.6110185...	1.5449423...	7.9270957...	294988.48...	38.535557...	
<input type="checkbox"/>		judicious-ant-680	10 minutes ago	3.0s	0.3486118...	1.6971586...	9.949223...	549396.94...	39.699253...	
<input type="checkbox"/>		youthful-loon-332	17 minutes ago	3.2s	0.3539691...	1.6759565...	9.9519490...	539795.05...	39.430532...	

# Repositorio de GitHub

github.com/ggomez1803/Proyecto\_DSA

Inicio de sesión -...

ggomez1803 / Proyecto\_DSA

Type to search

<> Code

Issues

Pull requests

Actions

Projects

Wiki

Security

Insights

Settings

Proyecto\_DSA

Public

Pin

Unwatch 1

main

5 branches

0 tags

Go to file

Add file

<> Code

ggomez1803

Add files via upload

2687d25

30 minutes ago

51 commits

.gitignore

Flujo preprocesamiento de datos, rutas y gitignore

2 weeks ago

DLTV.py

Add files via upload

yesterday

Experimentacion (1).ipynb

experimentacion PCA

1 hour ago

Experimentacion PCA.py

experimentacion PCA

1 hour ago

Experimento\_cluster\_jerarquico.py

Add files via upload

4 hours ago

Experimento\_kmeans.py

correccion de metricas

25 minutes ago

Exploracion\_de\_Datos\_.ipynb

Add files via upload

2 weeks ago

Funciones\_DLTV.py

Add files via upload

yesterday

Funciones\_preprocesamiento.py

funciones para corregir ciudades y transacciones

2 weeks ago

Preprocesamiento\_datos.py

Flujo preprocesamiento de datos, rutas y gitignore

2 weeks ago

README.md

readme2

2 weeks ago

Rutas.py

Flujo preprocesamiento de datos, rutas y gitignore

2 weeks ago



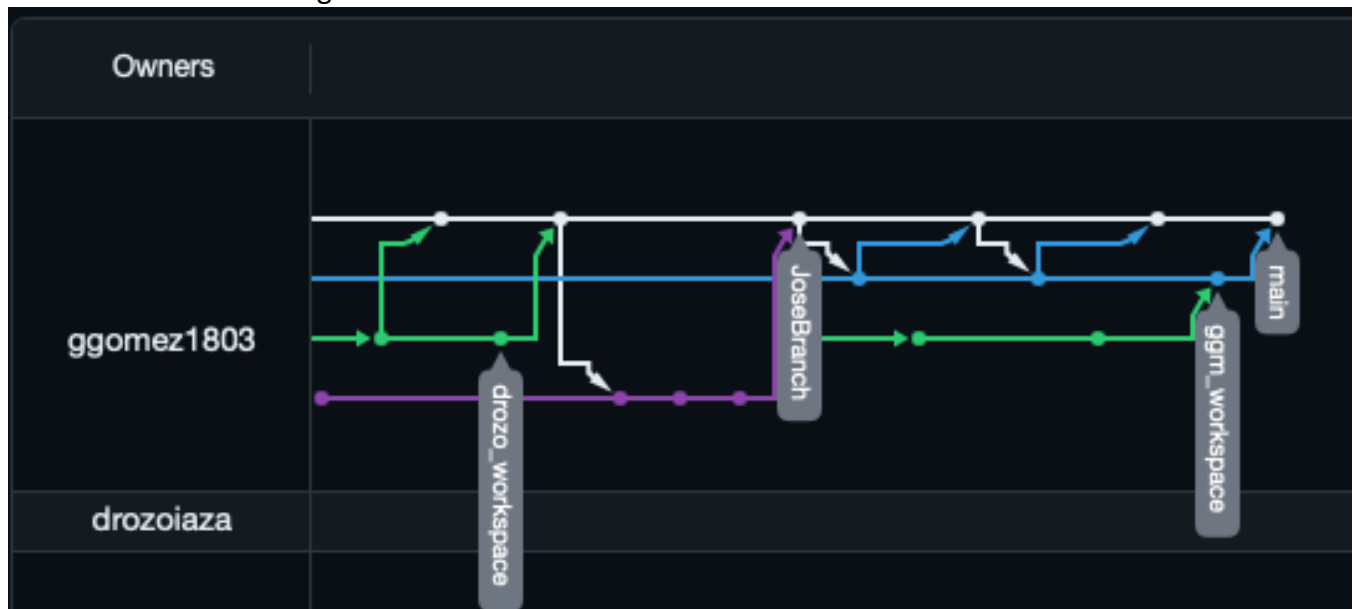
Dashboard desarrollado según la maqueta anteriormente presentada en la entrega 1.



Repositorio con todos los códigos

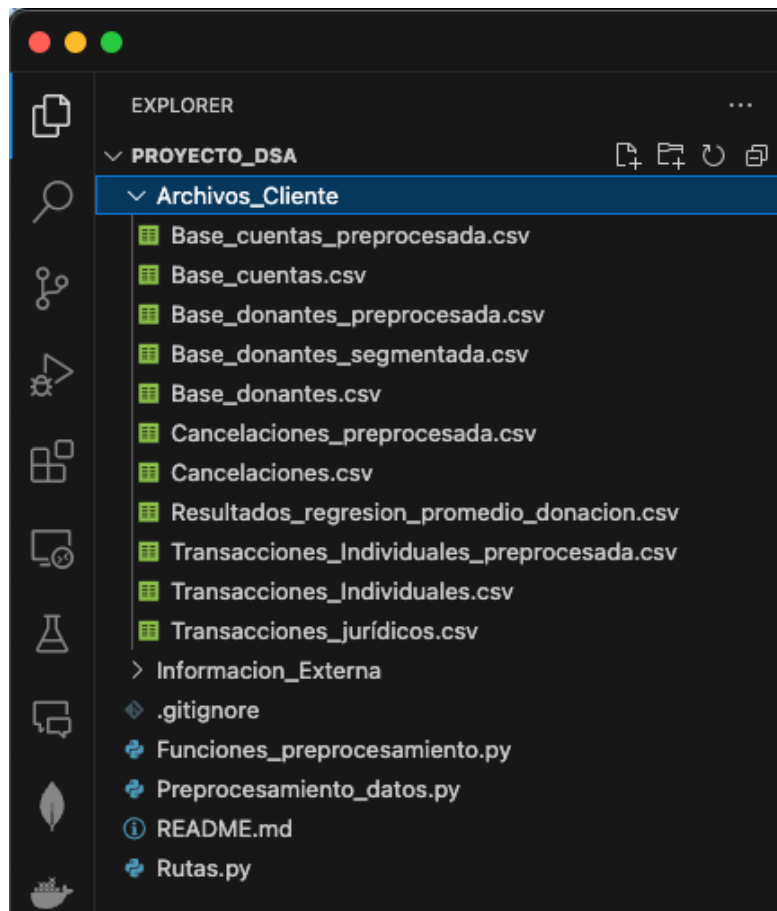
Link del repositorio: [https://github.com/ggomez1803/Proyecto\\_DSA.git](https://github.com/ggomez1803/Proyecto_DSA.git)

Contribución de los integrantes:





Fuentes de datos a emplear para los modelos (no se pueden enviar por un acuerdo de confidencialidad)



## Reporte de trabajo en equipo:

Yolanda Franco – Project Manager	<ul style="list-style-type: none"><li>- Organización y delegación de funciones del grupo.</li><li>- Exploración de datos de los donantes.</li><li>- Proyecto, pregunta de negocio y alcance del proyecto informe.</li></ul>
Daniel Rozo – Visualization Leader	<ul style="list-style-type: none"><li>- Reporte de trabajo en equipo.</li><li>- Diseño dashboard preliminar acorde a maqueta</li><li>- Colaboración en el procesamiento de experimentos</li><li>- Visualizaciones de donantes informe.</li></ul>
Gabriel Gomez – Analytics Leader	<ul style="list-style-type: none"><li>- Organización del repositorio en Github.</li><li>- Elaboración Experimentos PCA 1</li><li>- Elaboración experimentos Kmeans</li><li>- Código Dashboard preliminar</li></ul>
Jose Hoyos – Commercial Executive	<ul style="list-style-type: none"><li>- Elaboración Experimentos</li><li>- Informe descripción conjunto de datos a explorar.</li></ul>

## Reporte de trabajo en el repositorio de Github:

Todos los miembros del equipo colaboraron en parte del procesamiento de datos de los datos de donantes de nuestro proyecto. El siguiente es el reporte descargado de Github:

[ggomez1803/Proyecto\\_DSA: Proyecto para Despliegue de Soluciones Analíticas \(github.com\)](https://github.com/ggomez1803/Proyecto_DSA:Proyecto para Despliegue de Soluciones Analíticas)

