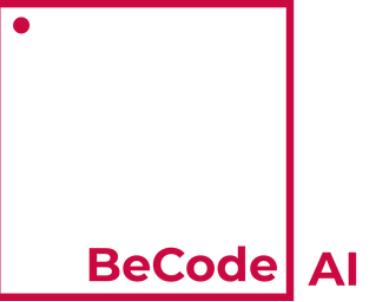




Proyecto Churn Rate

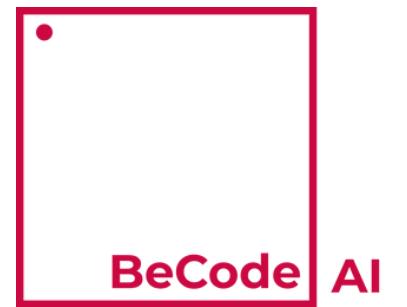
Victor Hugo Portilla Ortiz - German Wong del Toro - Misael Chavez Ramos-
Hector Miranda Garcia

Contexto del Proyecto



Este reto consiste en utilizar información de los clientes para su debido tratamiento y creación de clusters para categorizar a dichos clientes, además de un modelo predictivo de churn con base en su historial de comportamiento dentro del servicio.

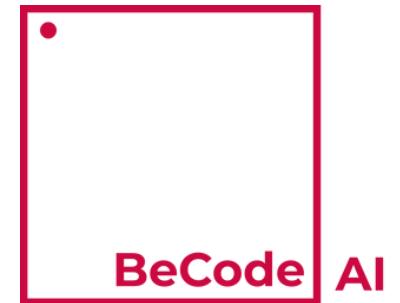
Solución pensada



Además de entrenar un modelo para la identificación de los usuarios que estén a punto de salirse, mostramos diversos datos con respecto a cada cluster para identificar múltiples causas por las cuales estos estén abandonando y por medio de una plataforma web poder enviar marketing dirigido para evitar la pérdida de clientes.

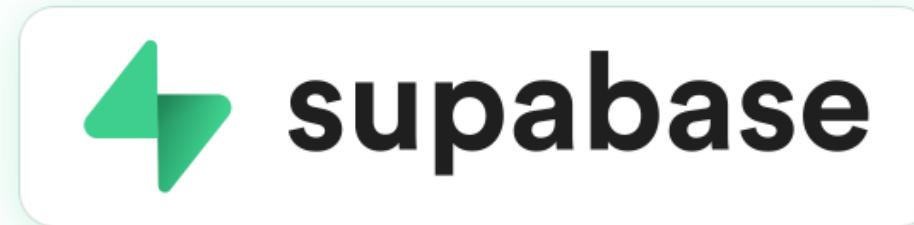
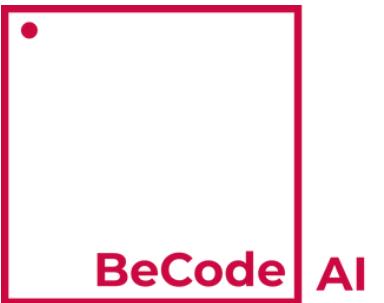
Herramientas utilizadas

Entrenamiento modelos



Herramientas utilizadas

Página Web



NodeMailer



Pre-estudio de los datos

BeCode AI

Utilizamos el módulo de ADS para obtener algunas recomendaciones por parte de este sobre el tratado de nuestro dataset

Summary

Name: User Provided DataFrame

Type: BinaryClassificationDataset

371,589 Rows, 42 Columns

Column Types:

- ordinal: 26 features
- categorical: 9 features
- continuous: 5 features
- datetime: 1 features
- address: 1 features

Note: Visualizations use a sampled subset of the dataset, this is to improve plotting performance. The sample size is calculated to be statistically significant within the confidence level: 95 and confidence interval: 1.0. The sampled data has 10,000 rows

- The confidence level refers to the long-term success rate of the method, that is, how often this type of interval will capture the parameter of interest.
- A specific confidence interval gives a range of plausible values for the parameter of interest

Features (42)

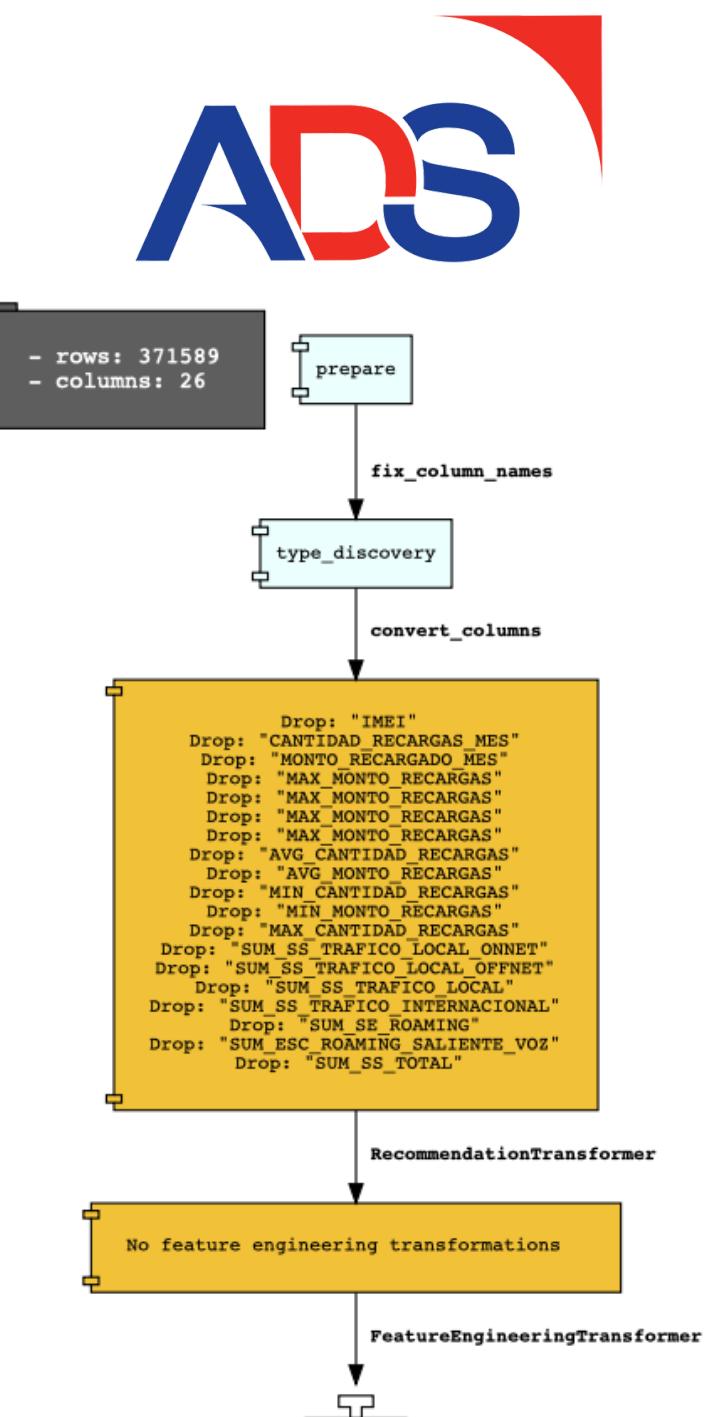
Correlations

Warnings (44)

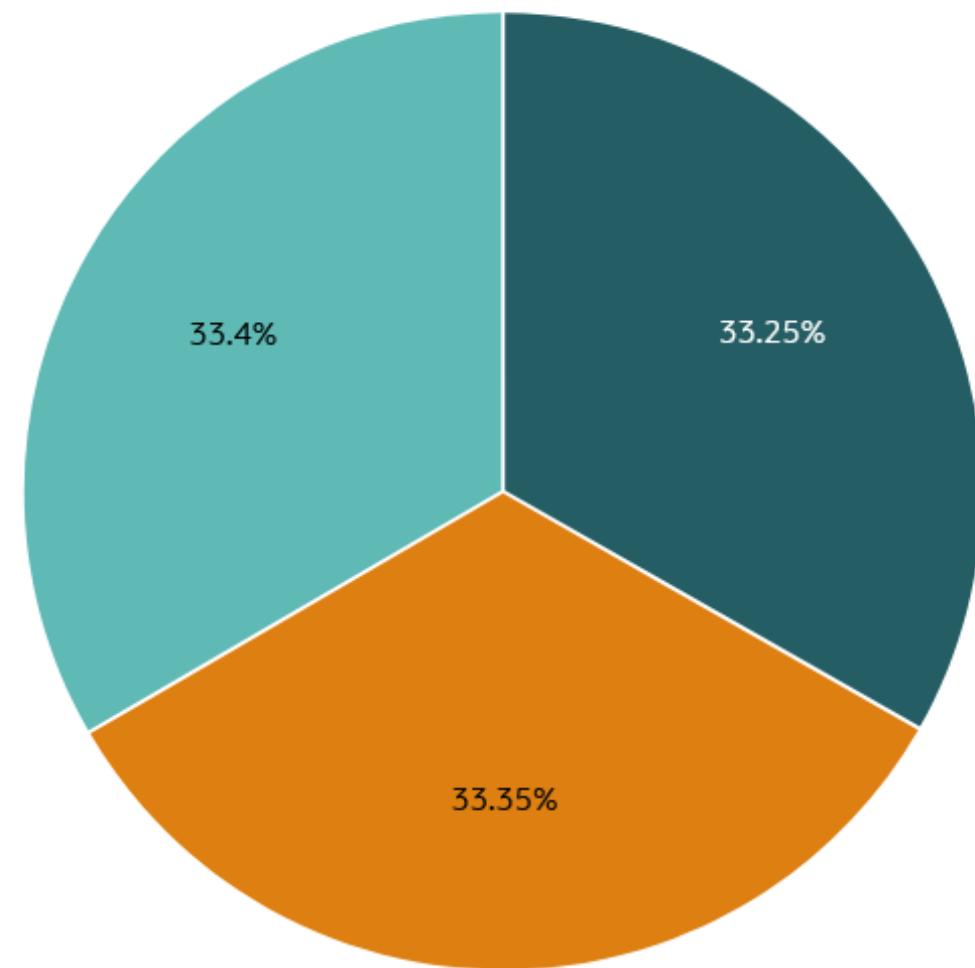
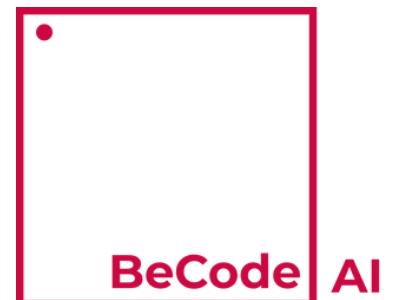
Potential Primary Key Columns

IMEI(type: int64) Contains mostly unique values(100.00%) Drop

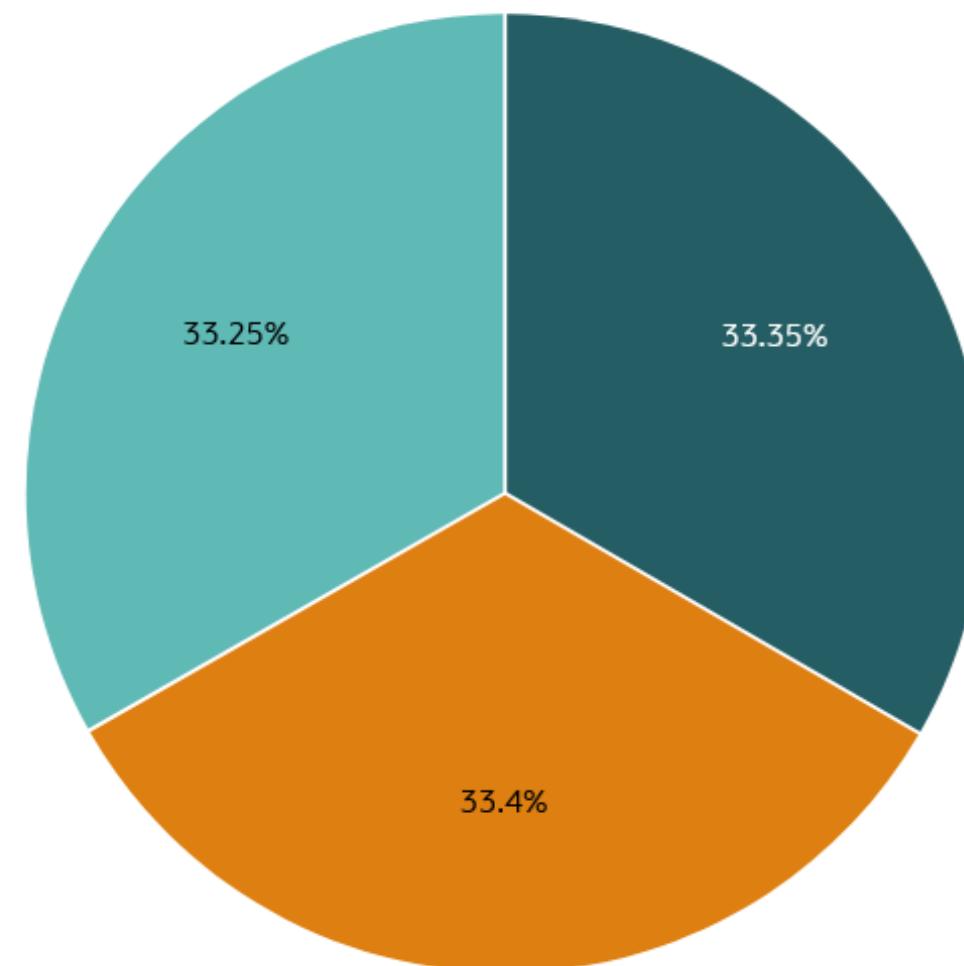
Next Reset All



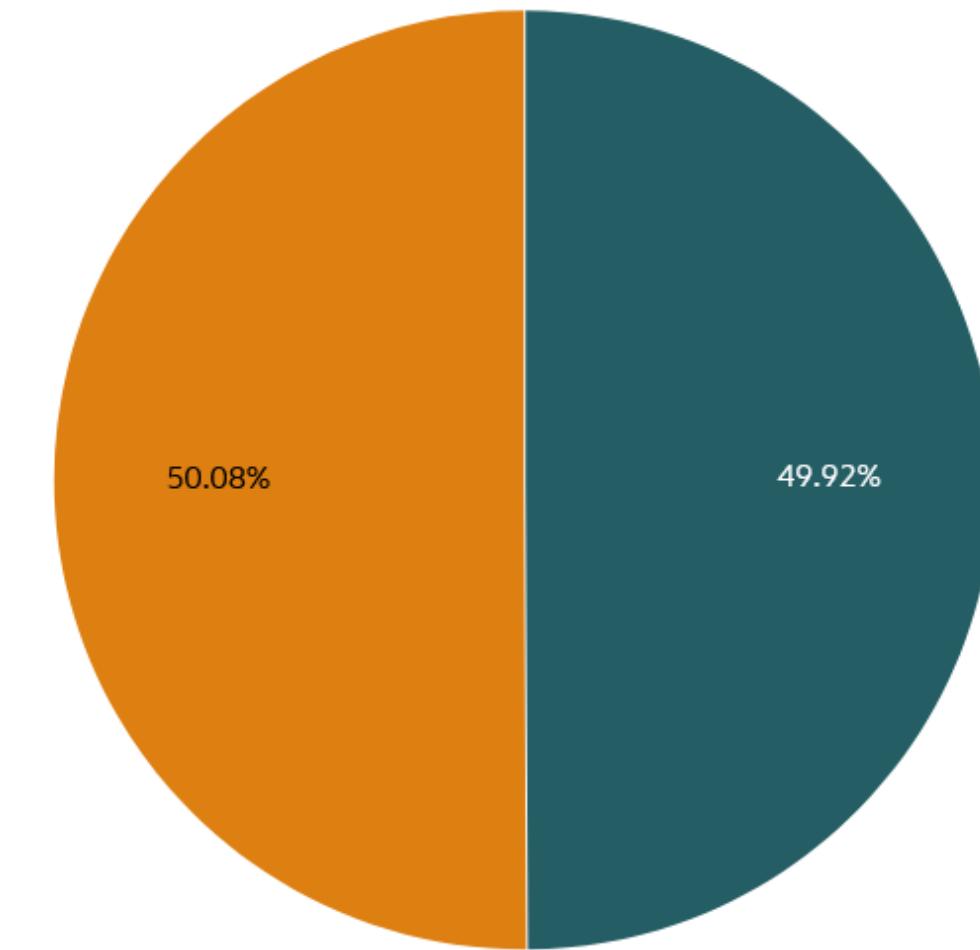
Distribución y análisis de los datos



MARITAL_STATUS ■ Married ■ Single ■ Widower



EMPLOYEE_STATUS ■ Unemployment ■ employee ■ self-employed



GENDER ■ FEMALE ■ MALE

Distribución y análisis de los datos

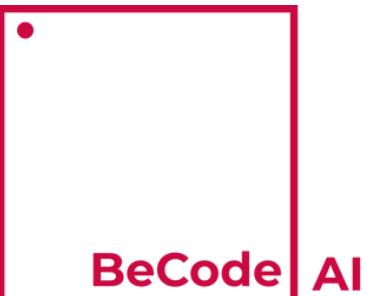
BeCode AI

Matriz de correlación

Con la matriz de correlación revela las relaciones entre cada una de las atributos en el conjunto, como se observa en la imagen existe mucha correlación positiva entre los atributos que contienen algún tipo de regarga.

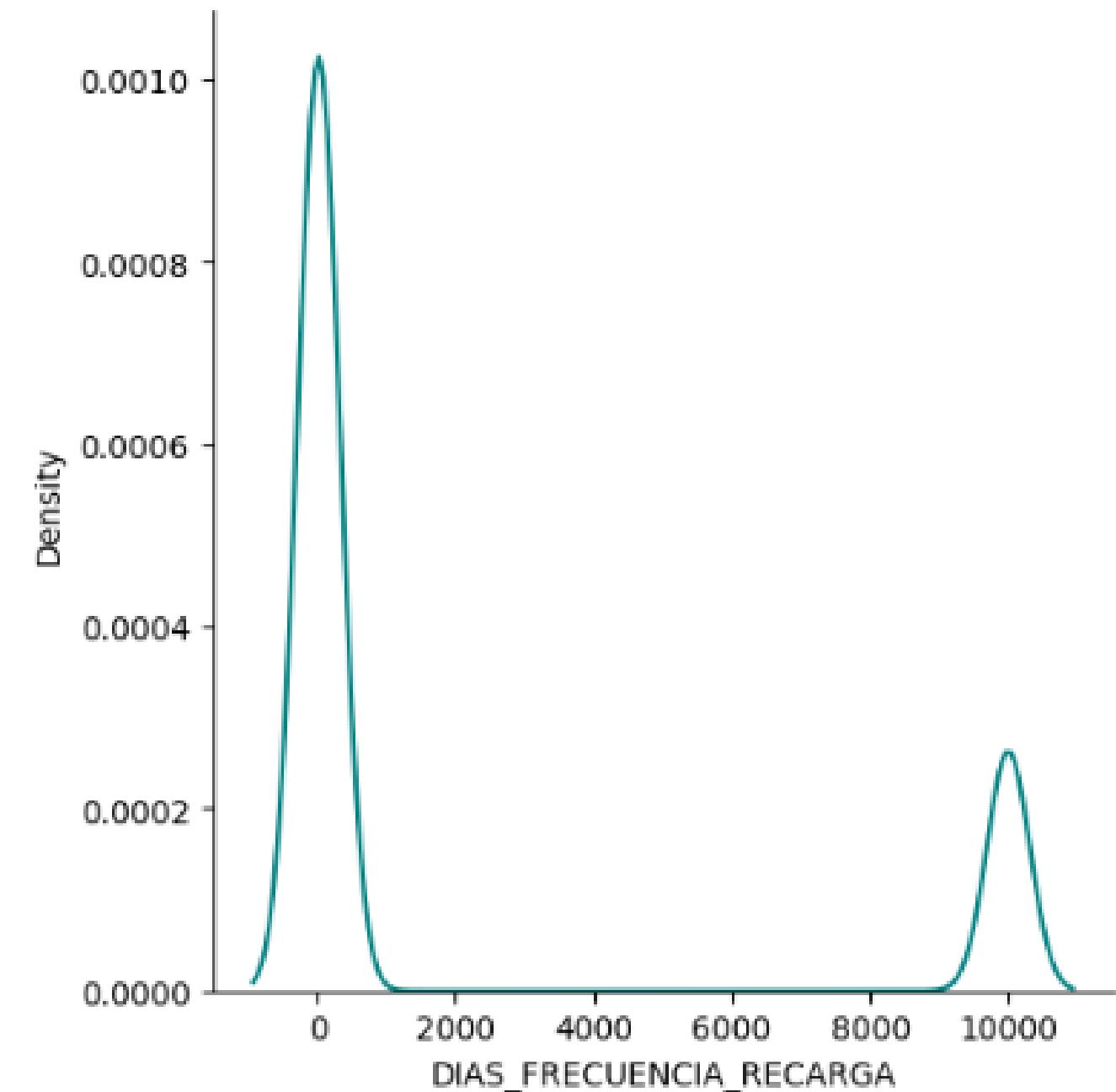


Distribución y análisis de los datos

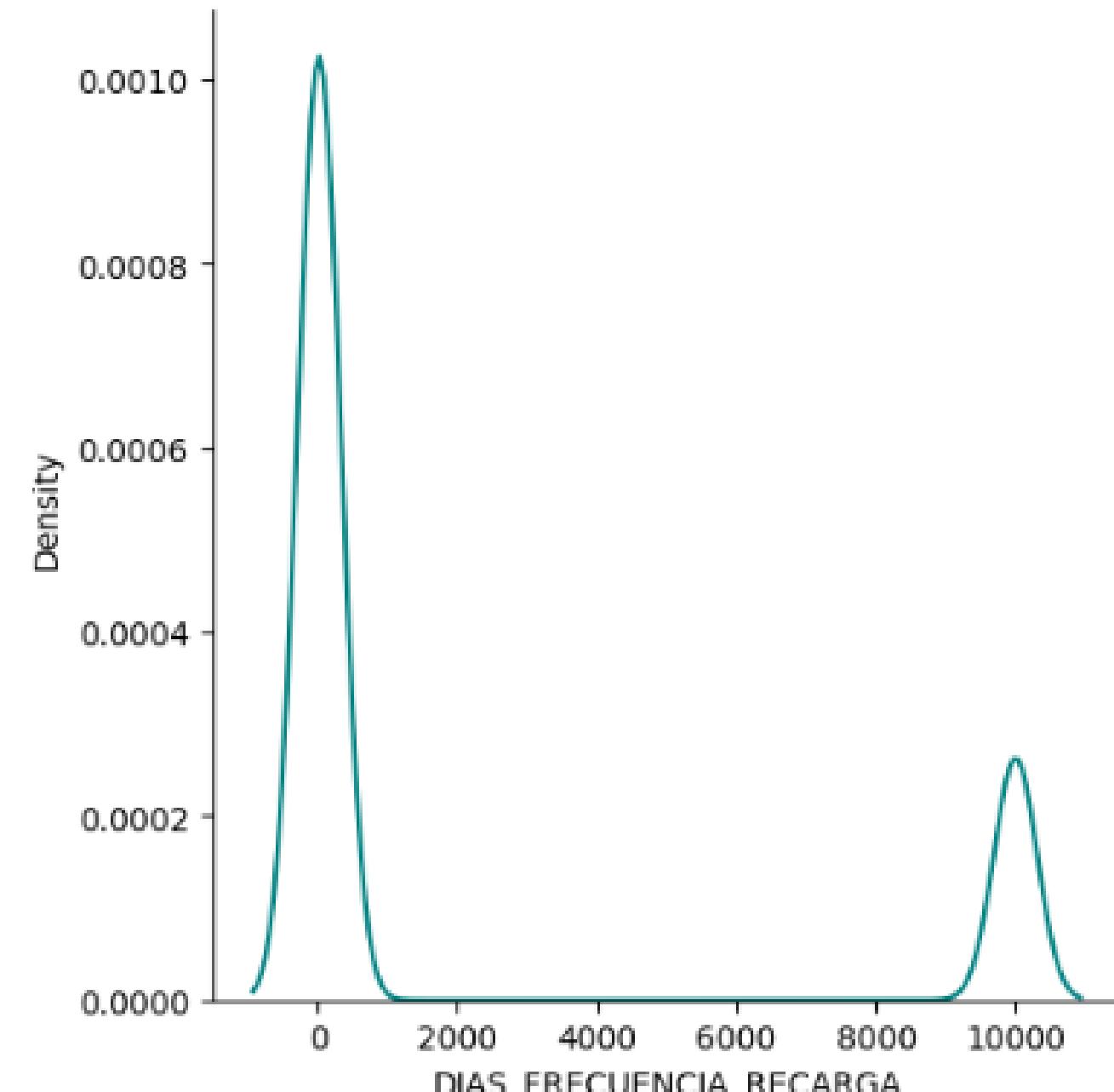
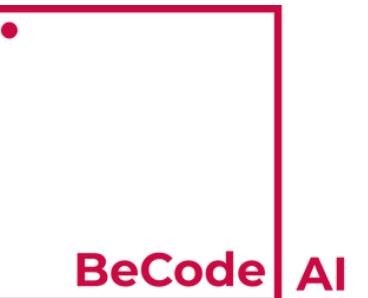


Distribucion de atributo DIAS_FRECUENCIA_RECARGA

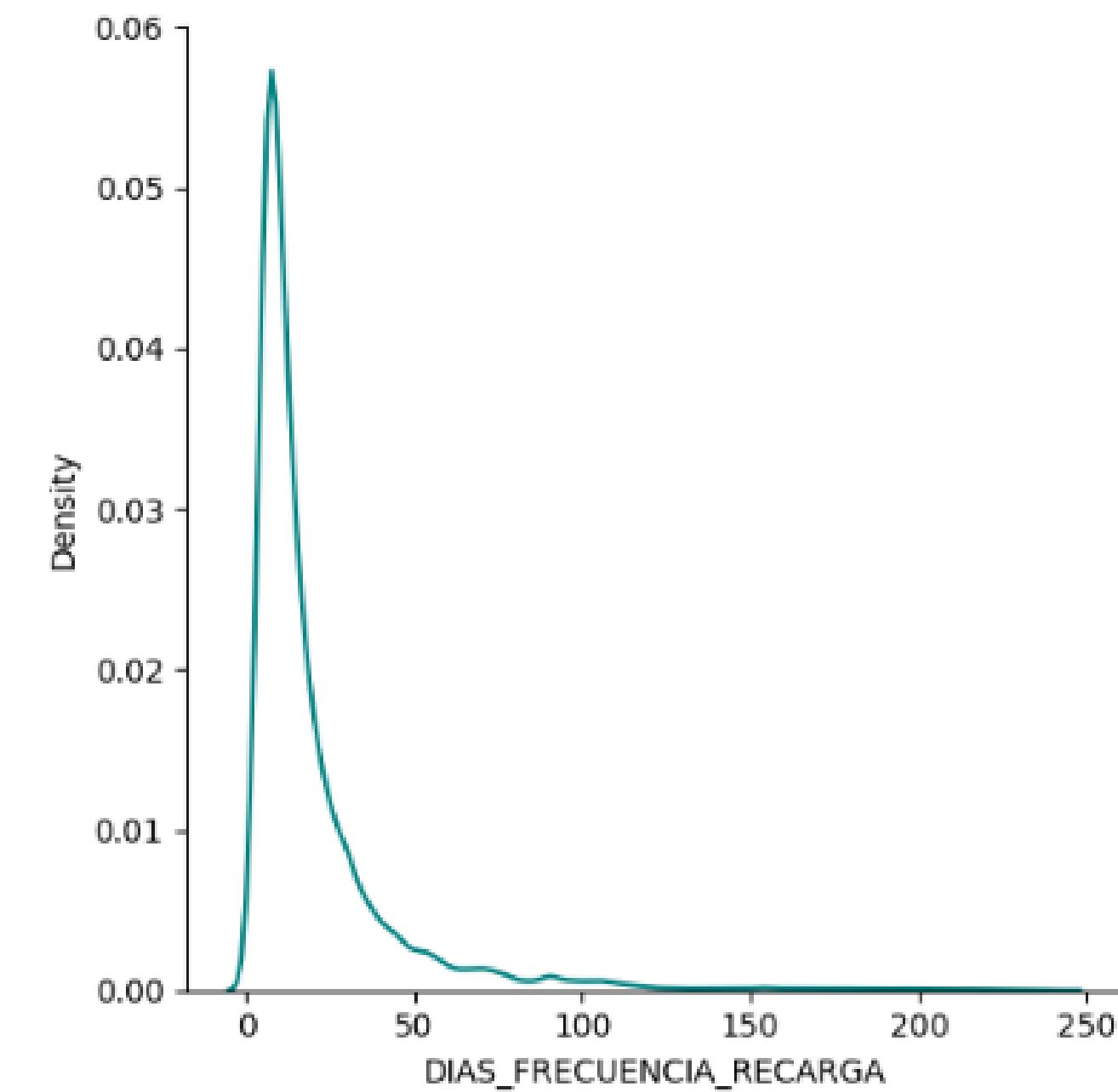
Encontramos un comportamiento extraño en el atributo DIAS_FRECUENCIA_RECARGA, al comparar sus diferentes metricas encontramos que el segmento de datos compuesto mayormente por valores de 9999 afectan tanto a nuestro modelo como a nuestra clasificación



Distribución y análisis de los datos



Antes



Después

Exploración y procesamiento de los datos



01 Verificación de datos faltantes

Al aplicar técnicas para la búsqueda de datos faltantes o erroneos, encontramos que había una pequeña cantidad de valores duplicados. Decidimos eliminar estos datos. De igual manera, encontramos un error en la dirección de una tienda.

02 Imputación de datos

La columna de la dirección de la tienda en Hawái presentaba un error de codificación que eliminaba registros. Se corrigió el nombre incompleto y se imputaron las coordenadas correctas para esta tienda.

03 Manejo de outliers

Utilizamos el método de Rango Intercuartílico (IQR) durante el entrenamiento del modelo no supervisado para el tratamiento de outliers. De igual manera, hicimos uso de Isolation Forest para encontrar outliers en el entrenamiento del modelo de clasificación.

Exploración y procesamiento de los datos

BeCode AI

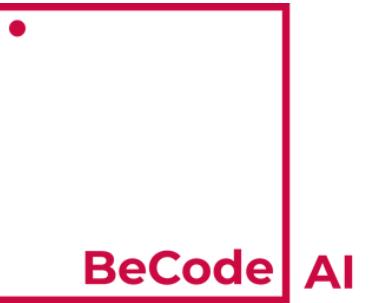
04 Datos categóricos

Los datos categóricos no fueron utilizados para el modelo de clustering, debido a los patrones irrelevantes que generaban.

05 Escalamiento de los datos

Escalamos los datos con sus valores mínimos y máximos para tratar el modelo de cluster y un escalamiento robusto (que escala en base a los rangos intercuartiles) para nuestro modelo de predicción de churn.

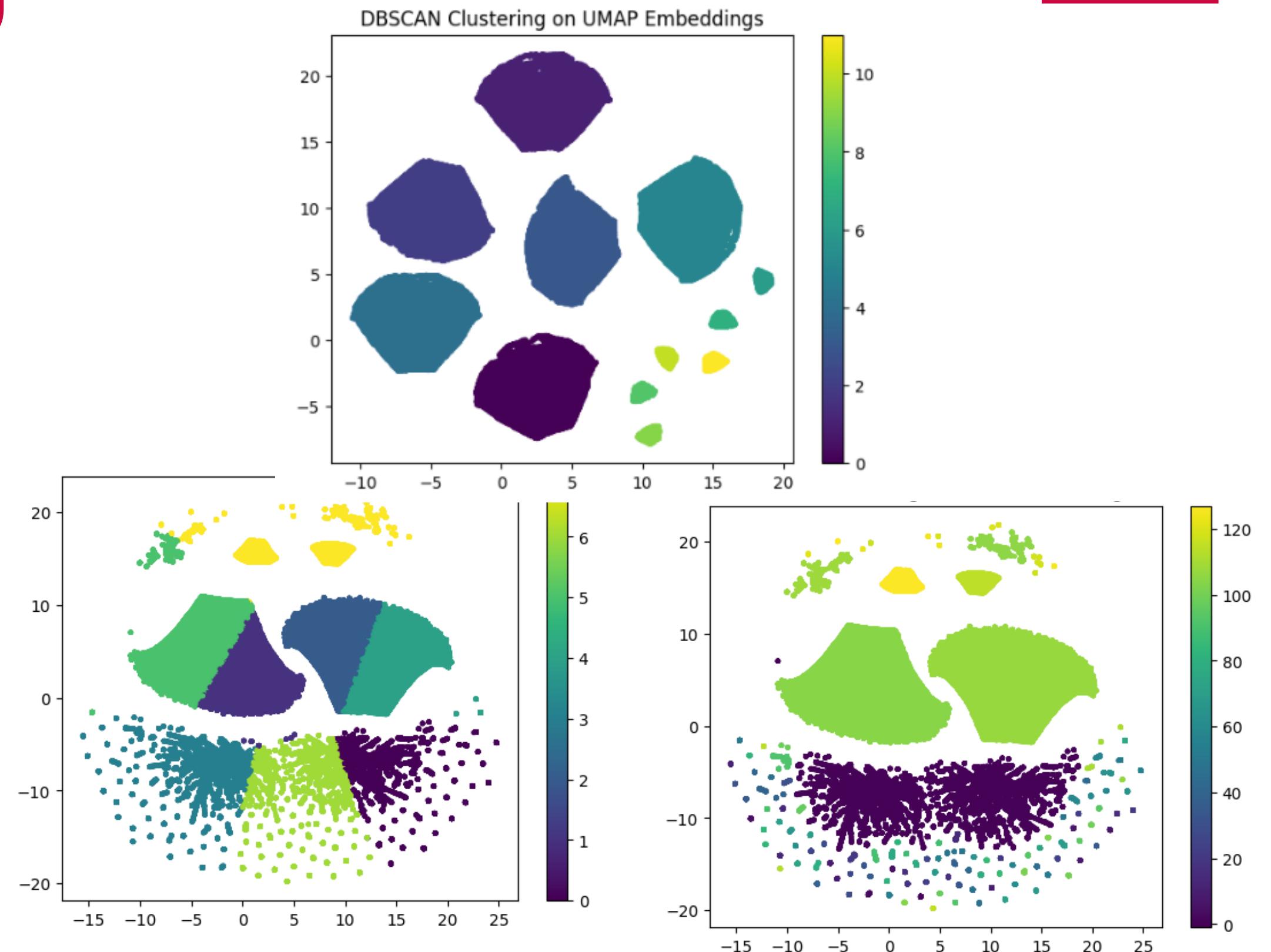
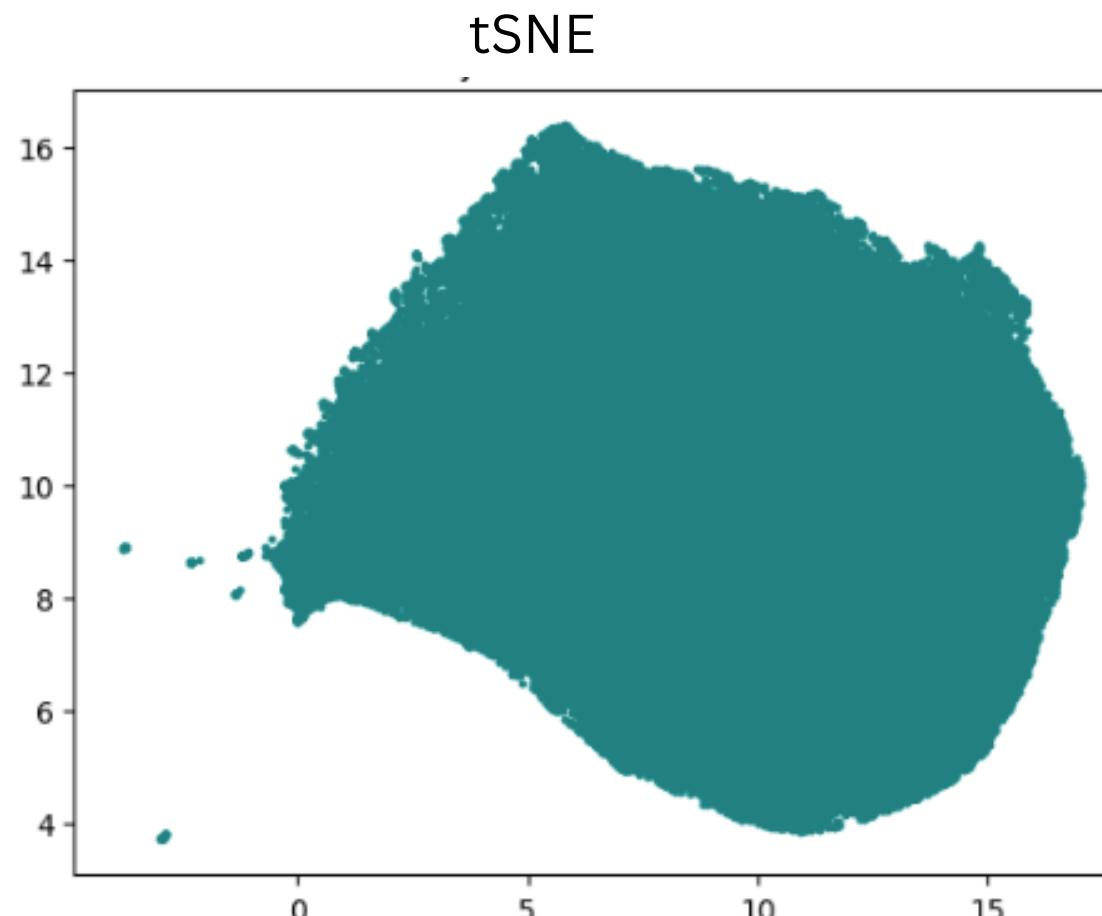
Outliers



- Nos enfrentamos a la dificultad de interpretar distribuciones debido a outliers que trivializaban el proceso.
- Implementamos un método robusto basado en el rango intercuartílico (IQR) para eliminar estas trivializaciones.
- Descartamos muestras fuera de los límites establecidos por el IQR en cada columna del conjunto de datos, reduciendo significativamente las muestras.

Clustering

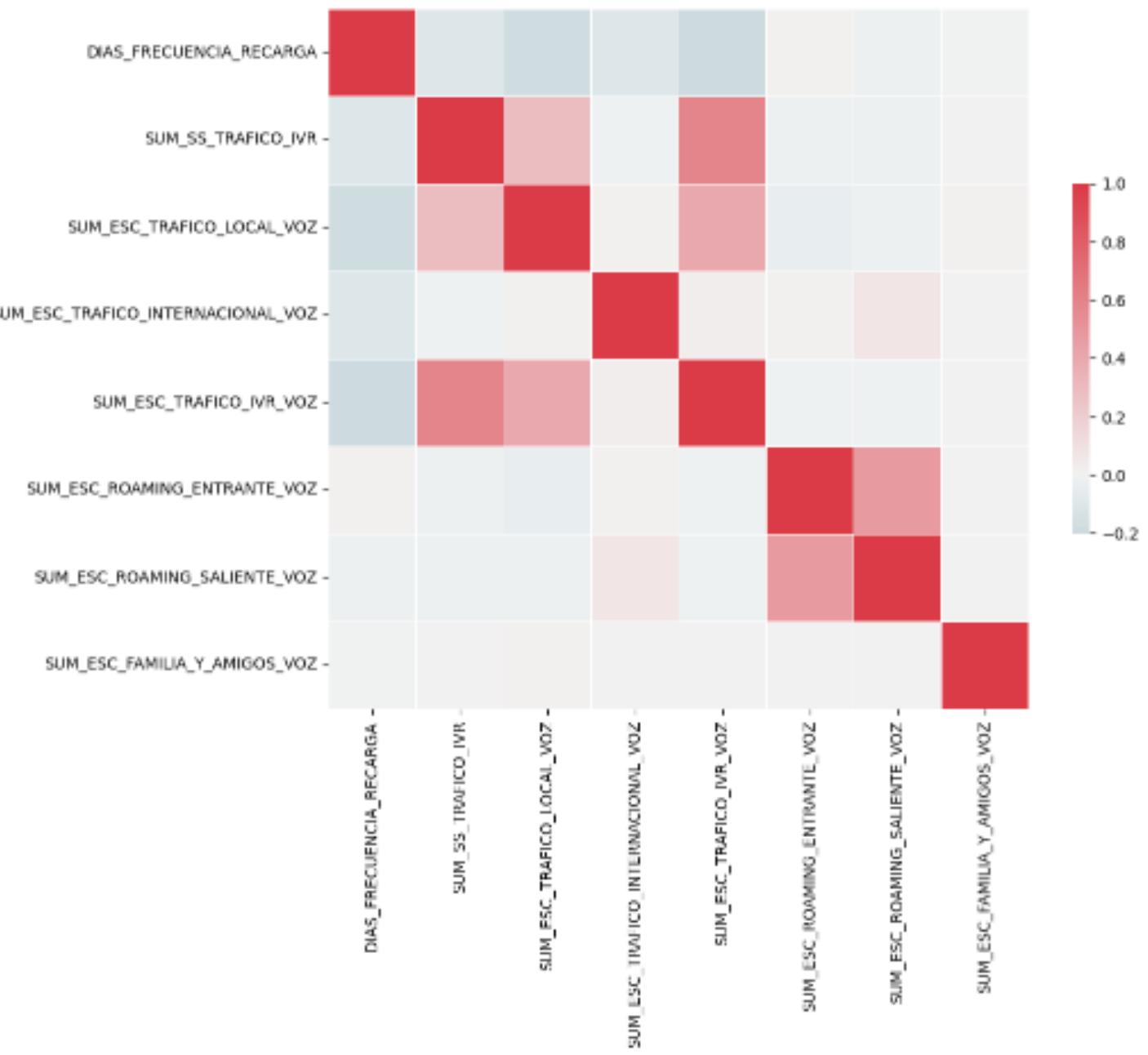
UMAP & tSNE



Matriz correlación final Clusters

BeCode AI

Para realizar el análisis de los componentes, primero utilizamos el dataset sin datos categóricos además de eliminar las columnas de SUM_CANTIDAD_RECARGAS y SUM_MONTO_RECARGAS, esto debido al nivel de importancia que tenían a la hora de hacer el clusterizado, obteniendo la siguiente matriz de correlación con los atributos restantes.

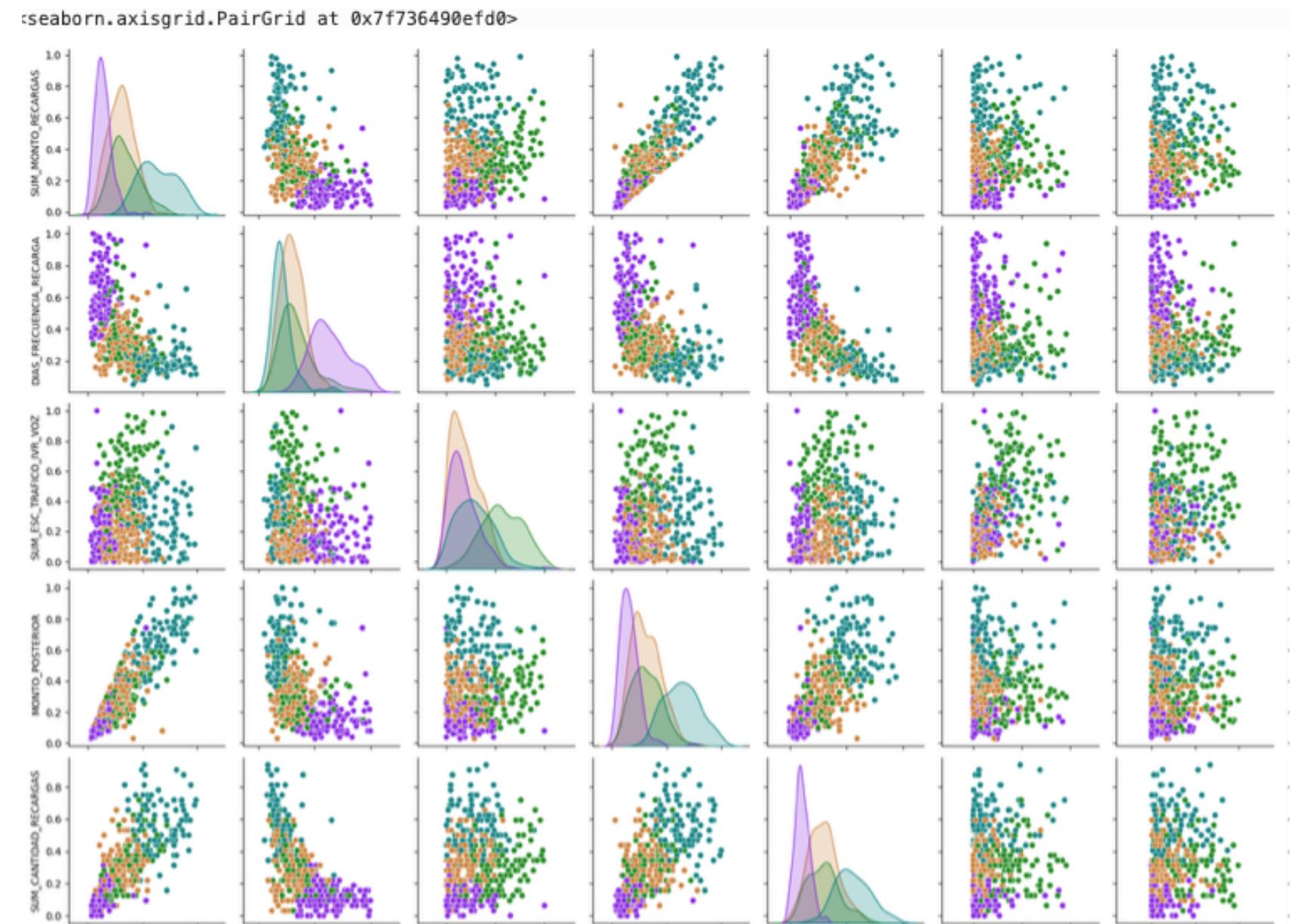


K-Means

BeCode AI

Tras probar con diferentes modelos de clustering, decidimos usar K-Means.

Este es modelo es fácil de usar, fácil de exportar y se acomoda a las distribuciones que obtuvimos al procesar nuestros outliers.

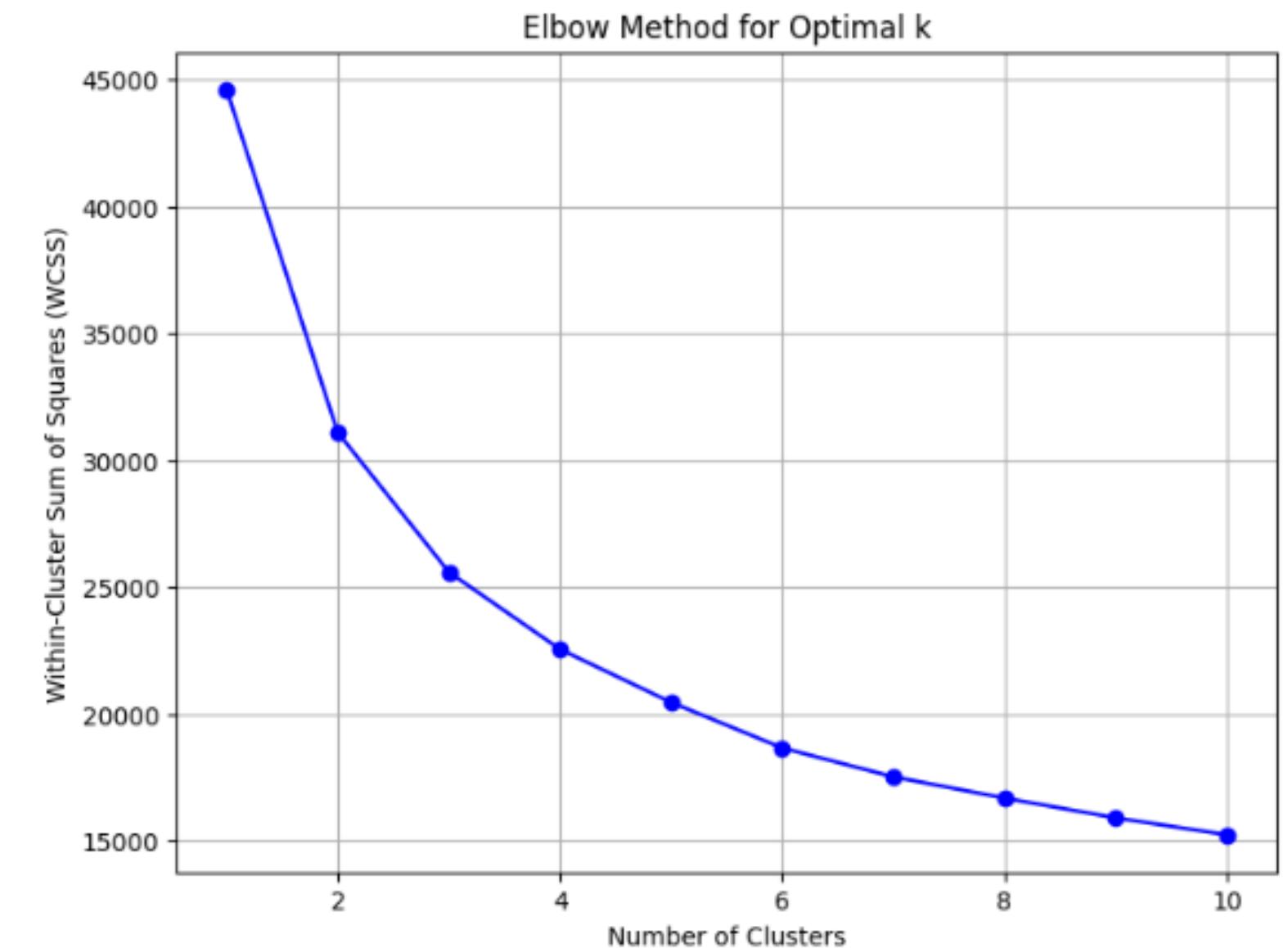


Método del codo



Este dataset fue utilizado para llevar acabo el método del codo con el objetivo de obtener un número adecuado de clusters.

Con esta gráfica nos percatamos que un numero ideal para clusters seria de 4.

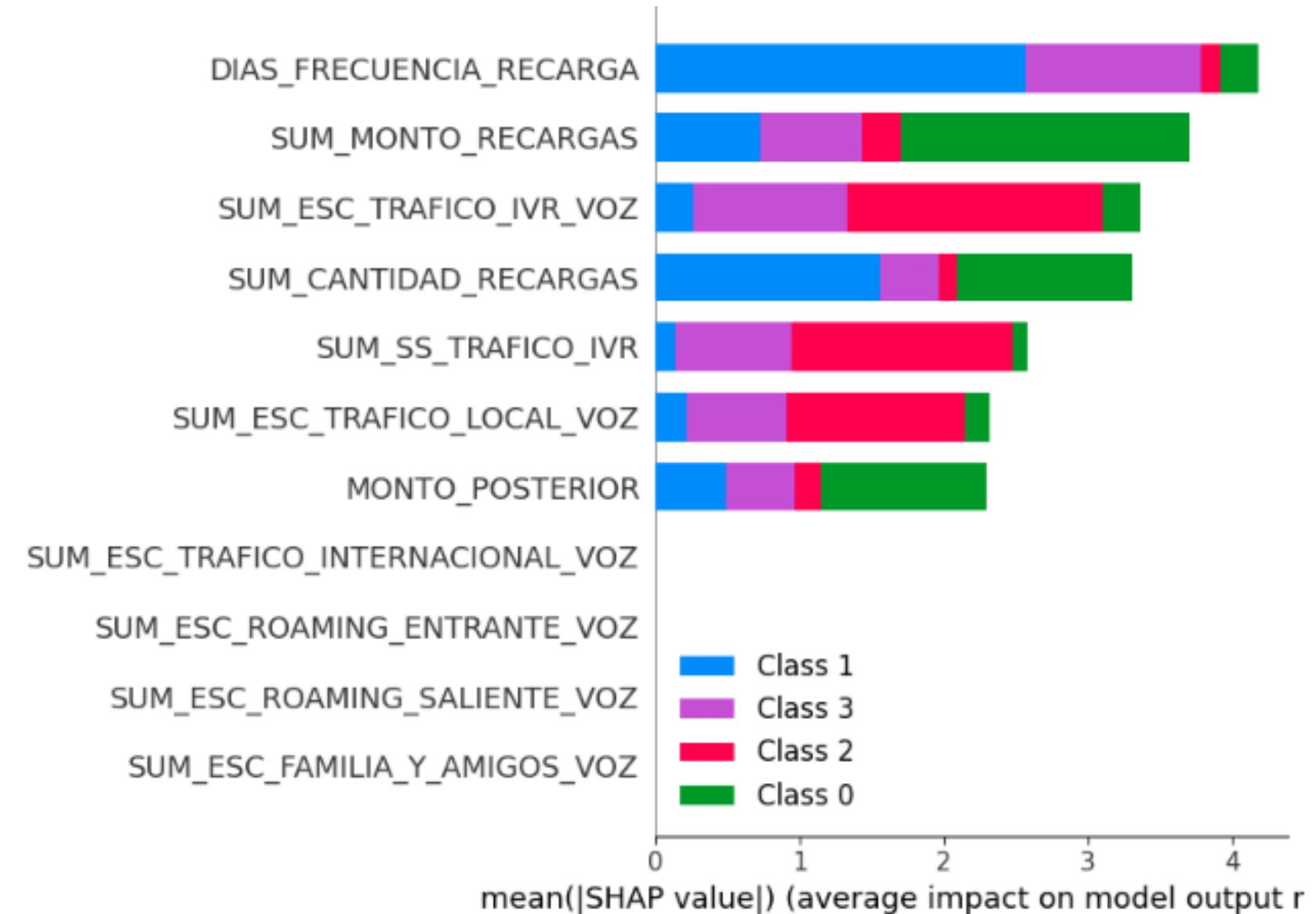


Métodos de Clustering

BeCode AI

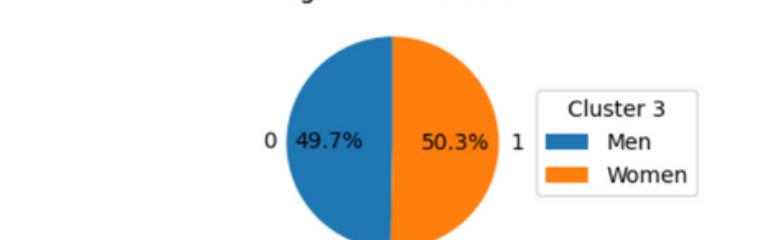
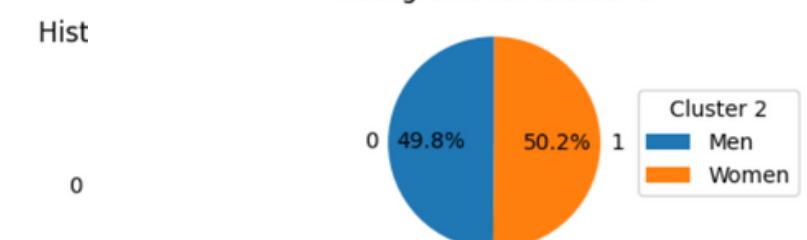
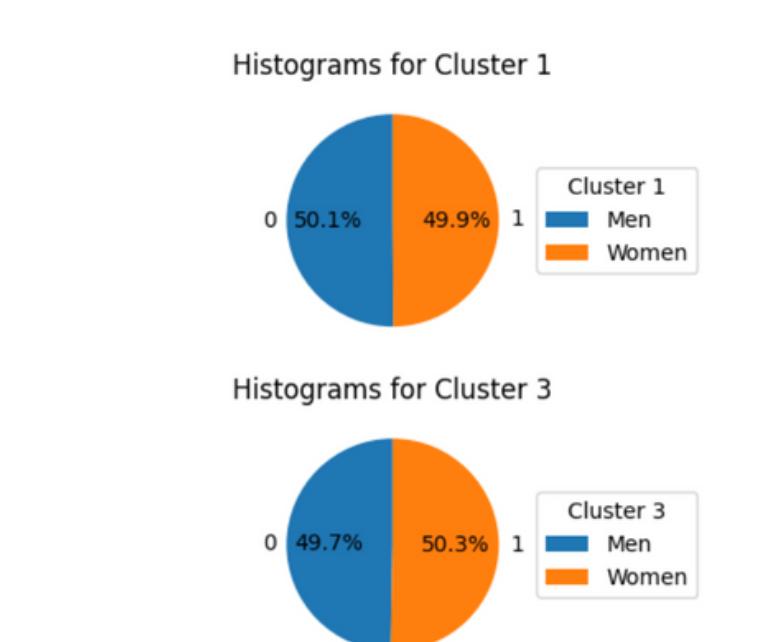
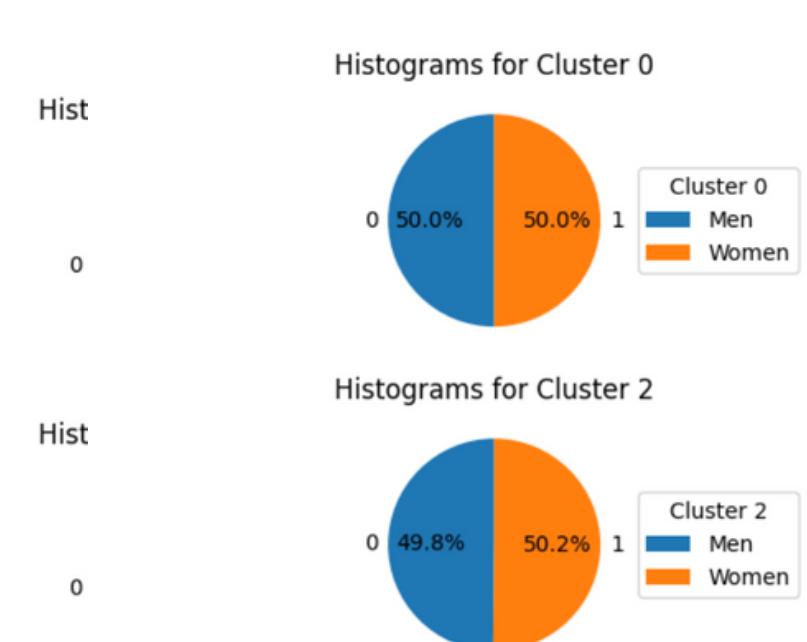
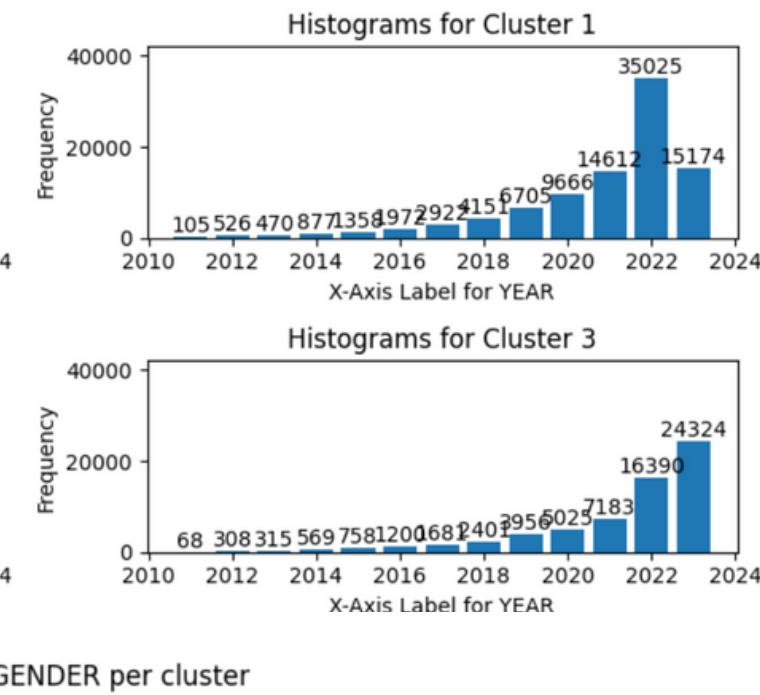
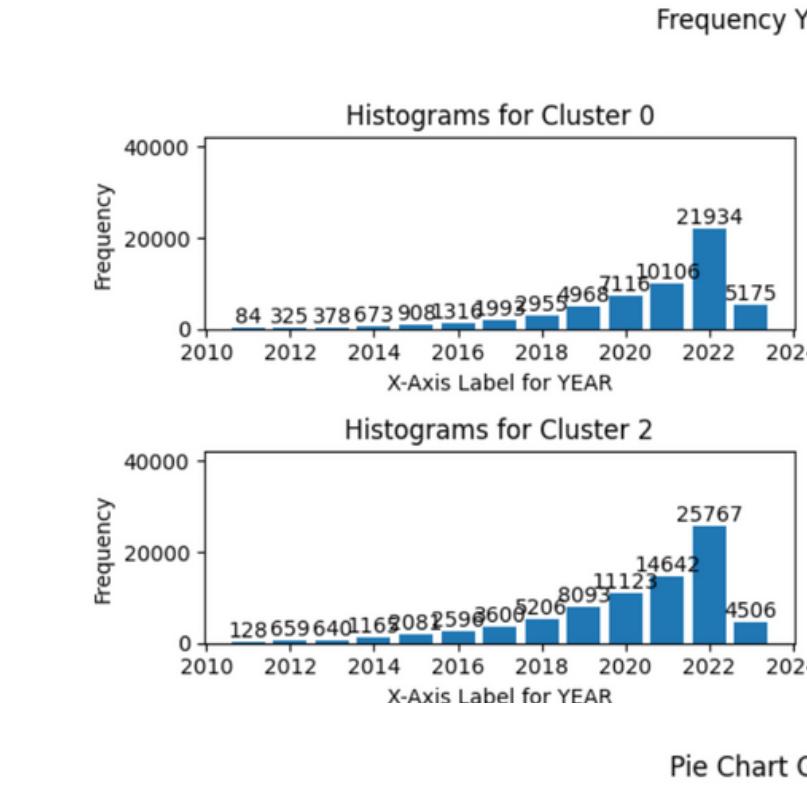
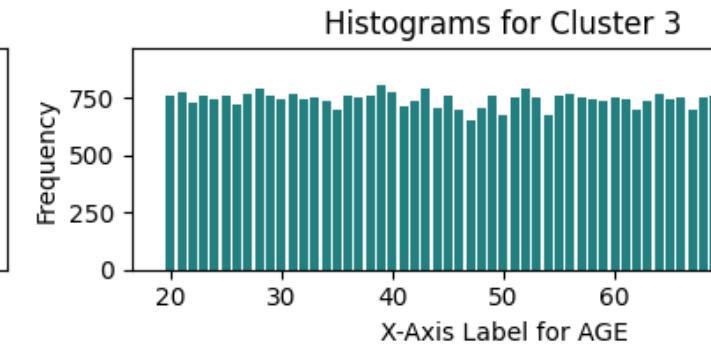
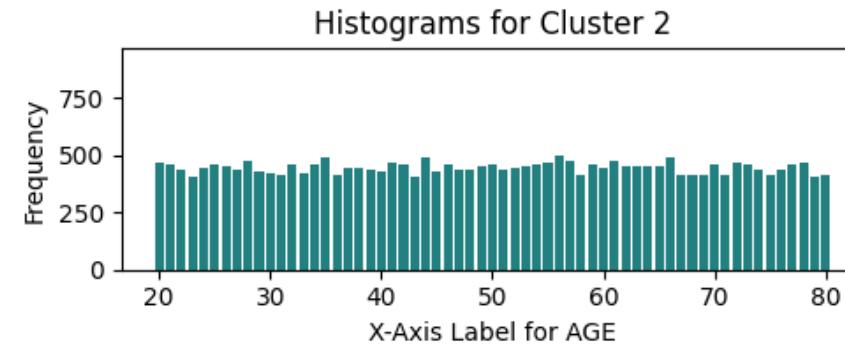
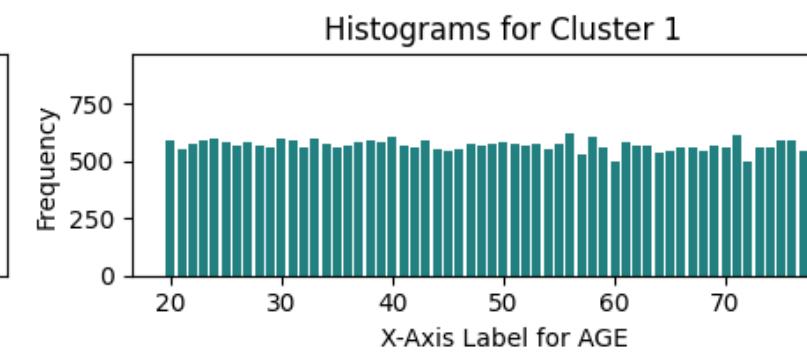
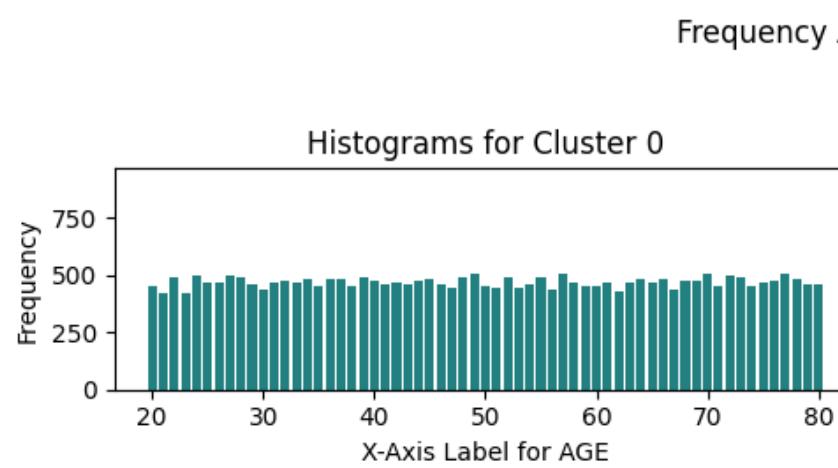
Hicimos uso de los valores SHAP (SHapley Additive exPlanations) para ver la importancia que le da a cada atributo un modelo de predicción.

De esta manera, entrenando a un modelo como LGBM usando los clusters como target, podemos ver los atributos que importan a cada cluster.



Distribución entre clusters

BeCode AI



Clustering

```
[104]: agent.run("What can you interpret from the LABEL_CLUSTER column and the main differences between clusters?")
```

> Entering new AgentExecutor chain...

Thought: The values in the LABEL_CLUSTER column seem to be categorical, and they may represent different customer segments. I can look at the unique values in this column to get a better understanding of the different clusters.

Action: python_repl_ast

Action Input: df['LABEL_CLUSTER'].unique()

Observation: [5 0 3 12 4 7 9 14 1 8 6 13 2 11 10 15]

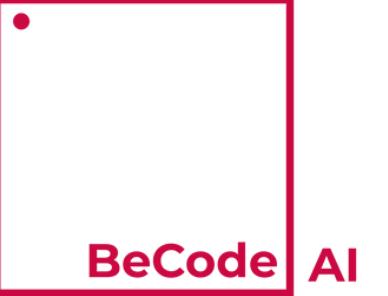
Thought: There are 16 unique values in the LABEL_CLUSTER column, which suggests that there are 16 different customer segments.

Final Answer: The LABEL_CLUSTER column contains 16 unique values, suggesting that there are 16 different customer segments.

> Finished chain.

```
[104]: 'The LABEL_CLUSTER column contains 16 unique values, suggesting that there are 16 different customer segments.'
```

Tipos de usuarios

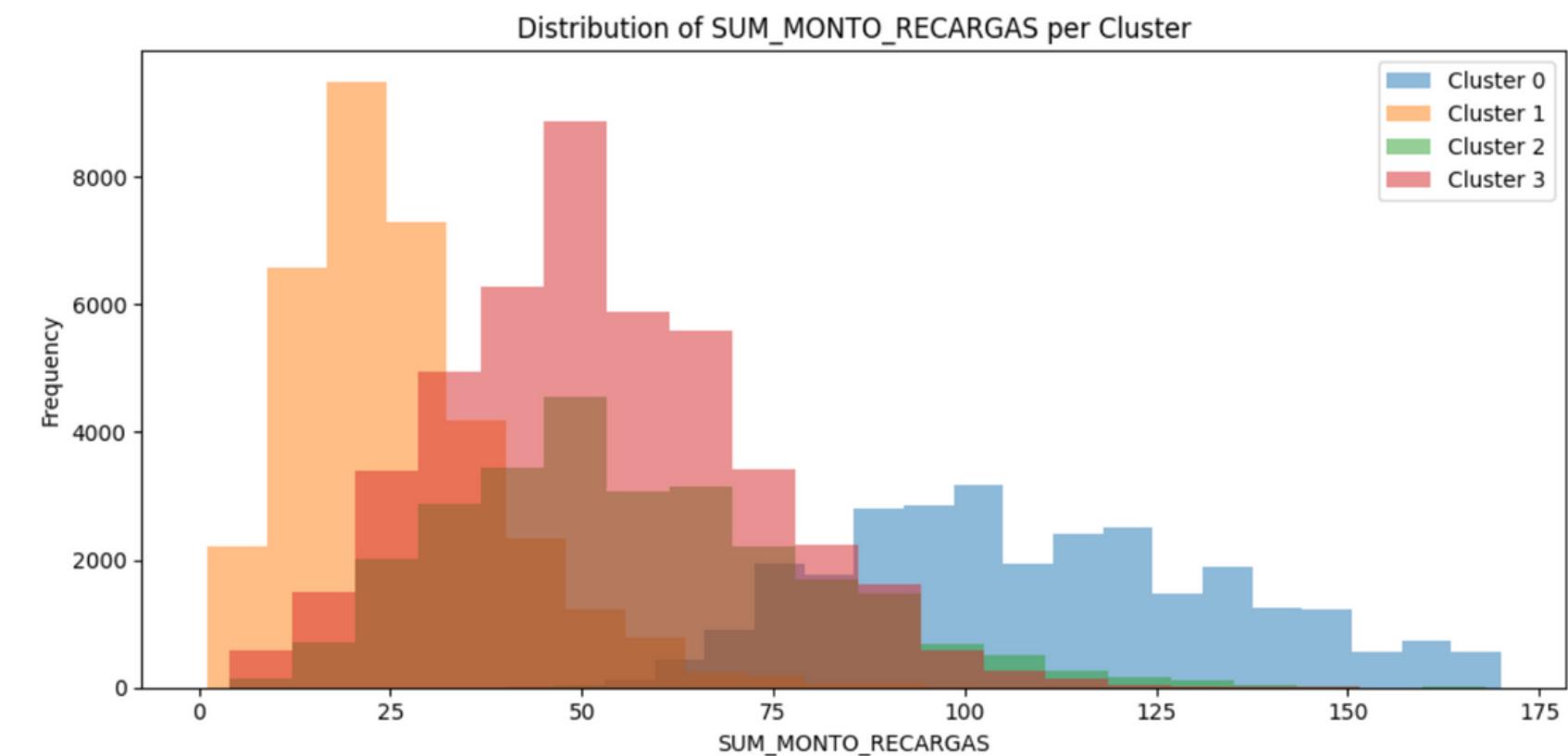
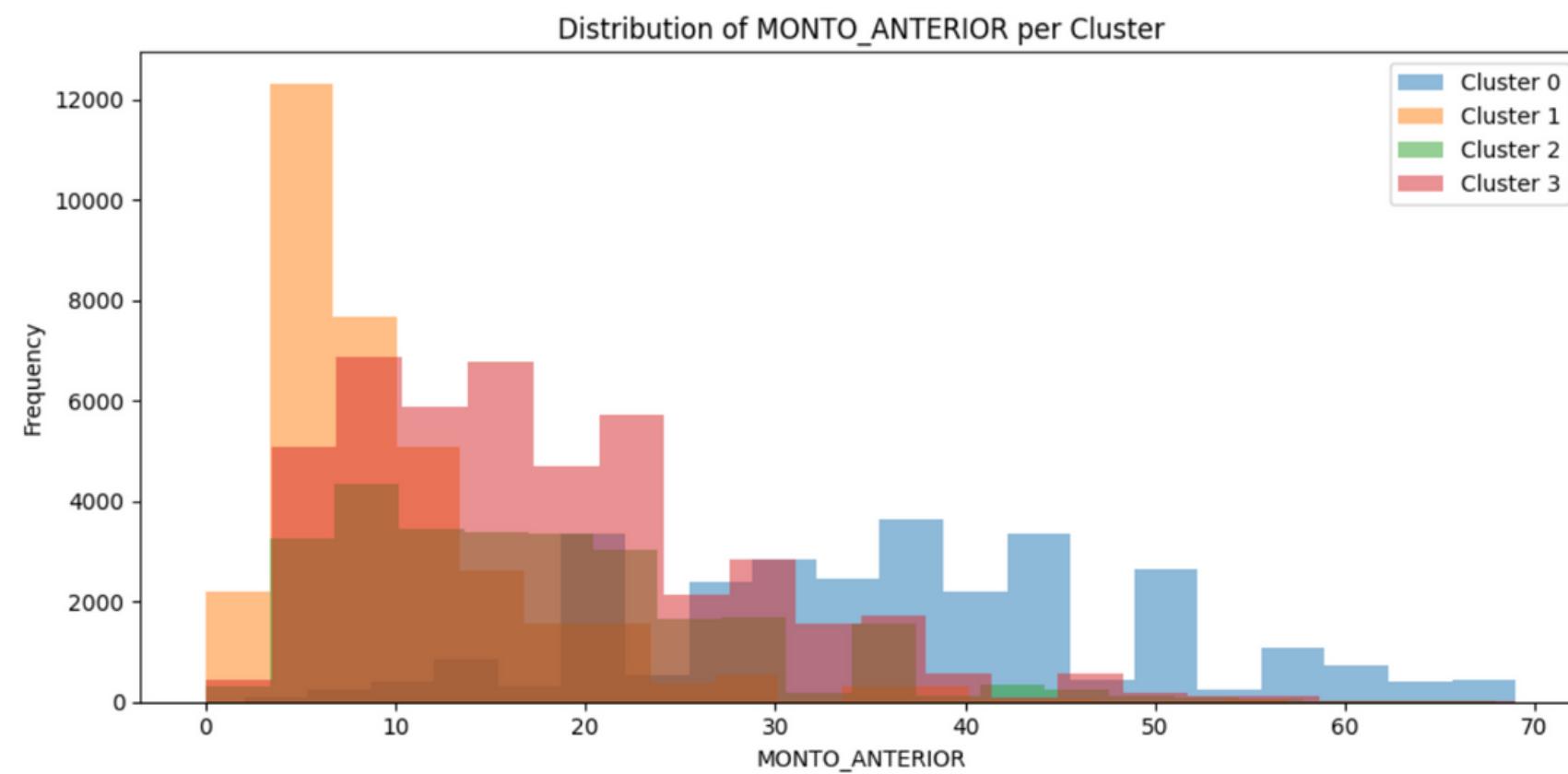
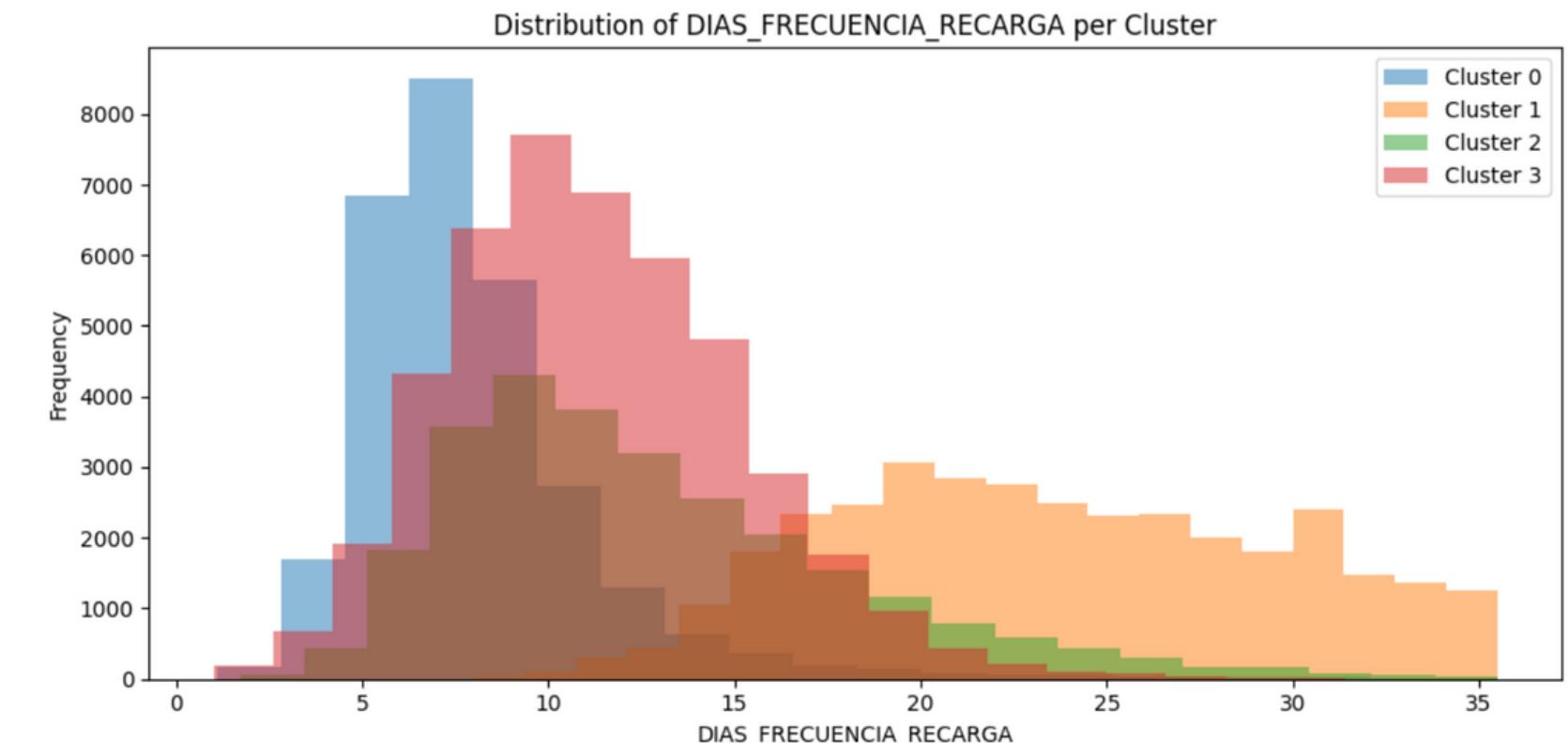


Realizamos un análisis de los cluster para determinar los tipos de usuario, los cuales son:

- Cluster 0 Recargadores Prodigiosos
- Cluster 1 Recargadores Cautelosos
- Cluster 2 Comunicadores Globales
- Cluster 3 Usuarios Estándar

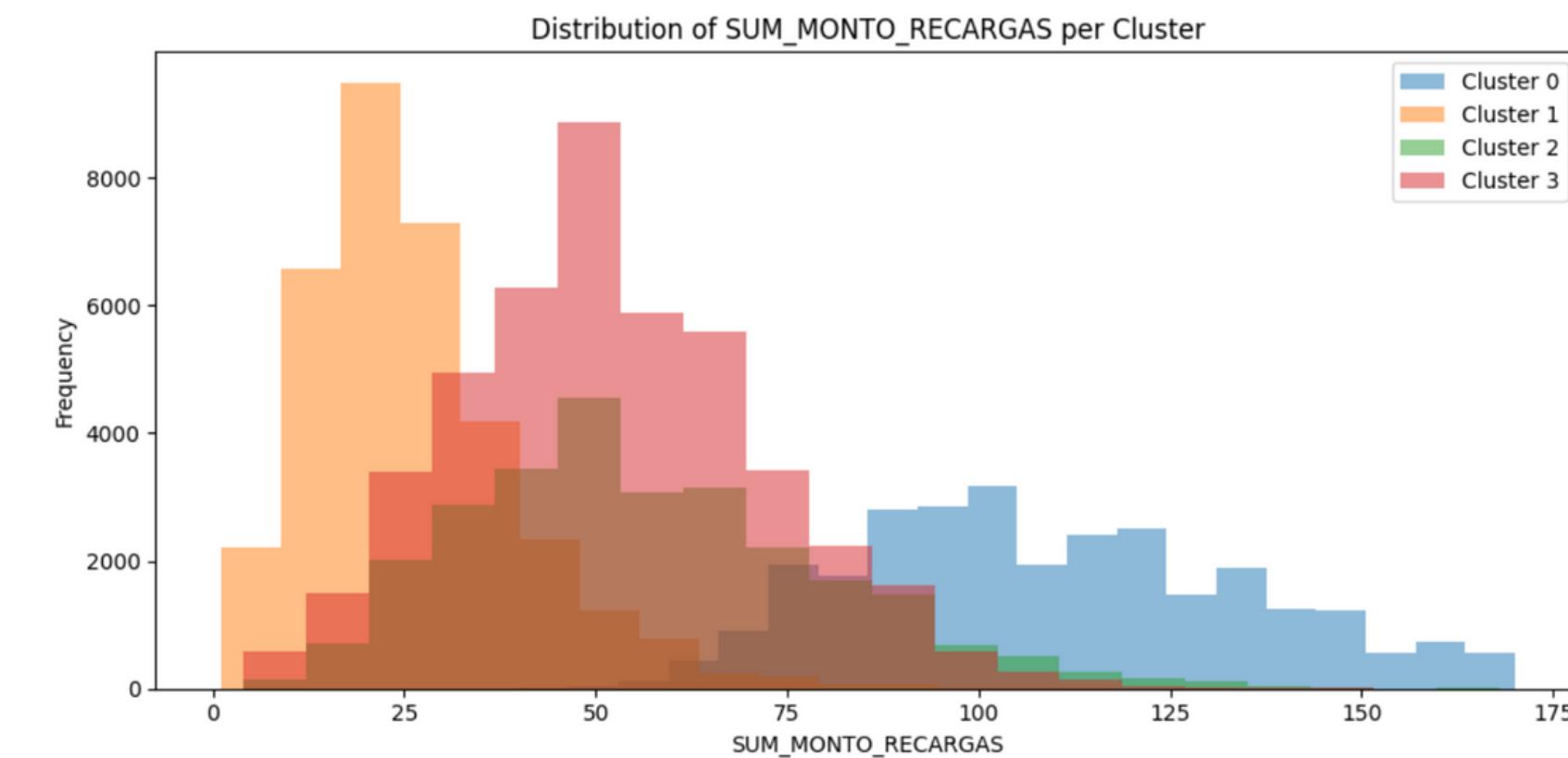
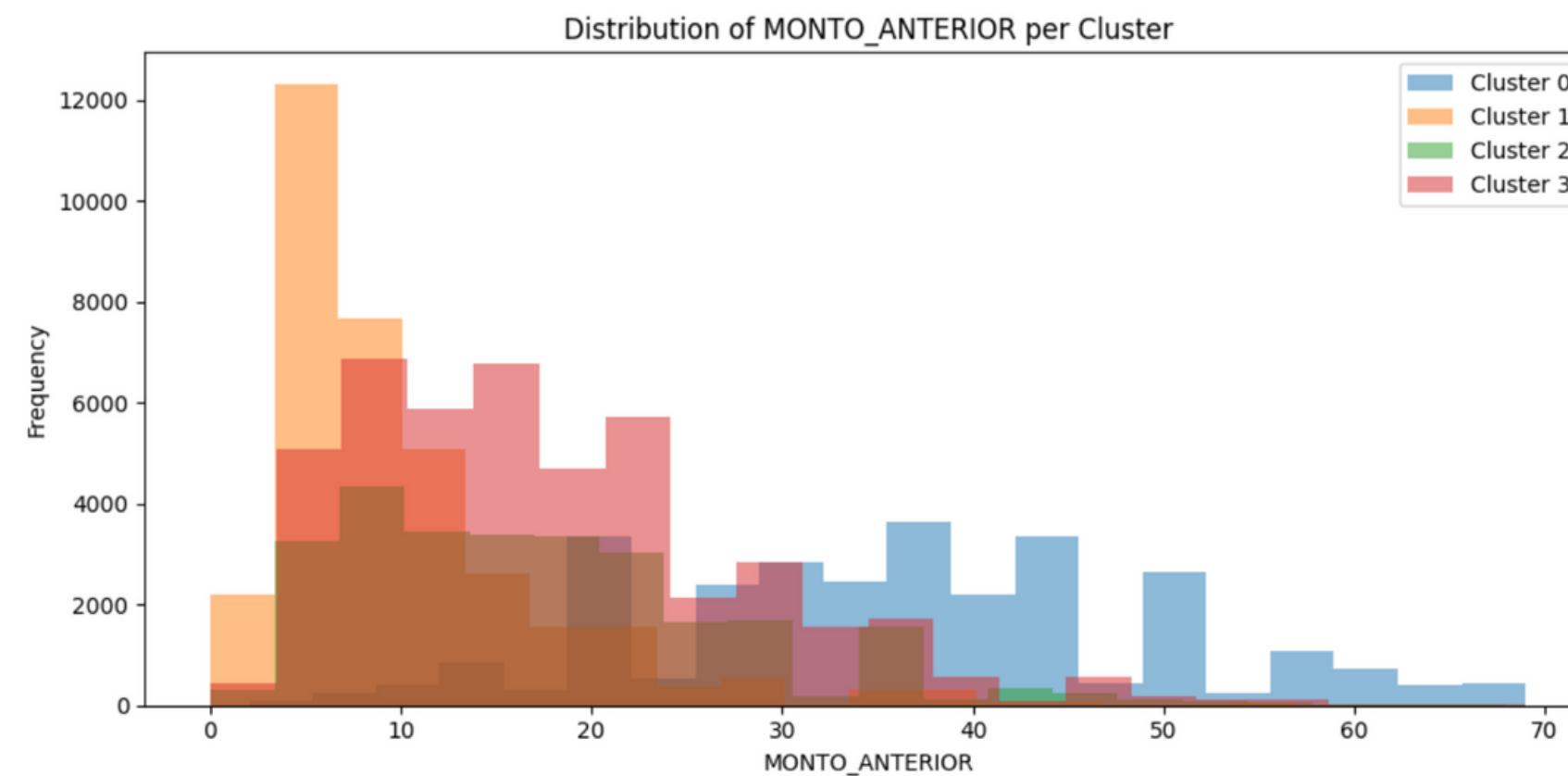
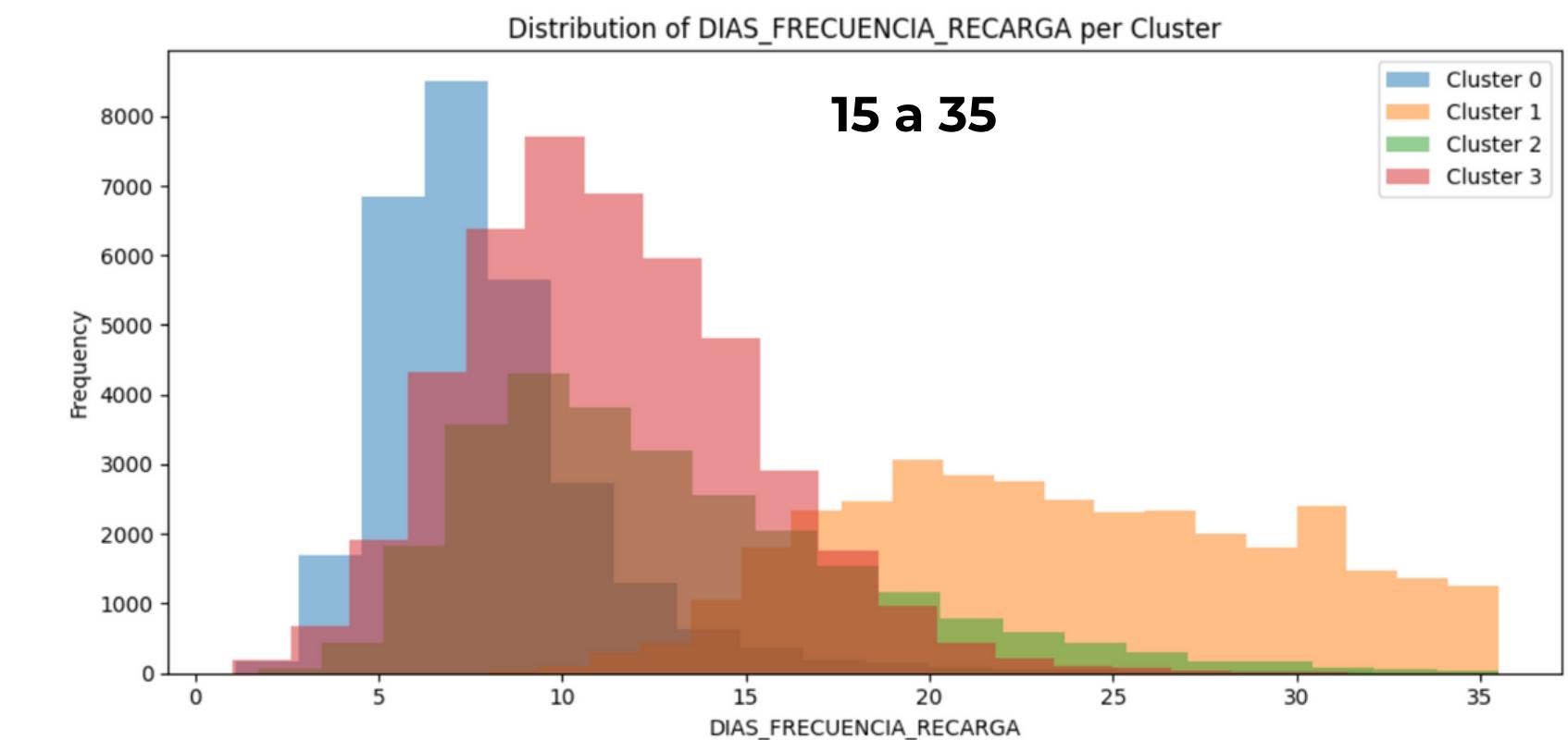
Cluster 0: Recargadores Prodigiosos

Usuarios destacados por realizar recargas frecuentes y de cuantías considerables. Sobresalen por la rapidez con la que efectúan sus recargas, evidenciando una marcada preferencia por montos substanciales.



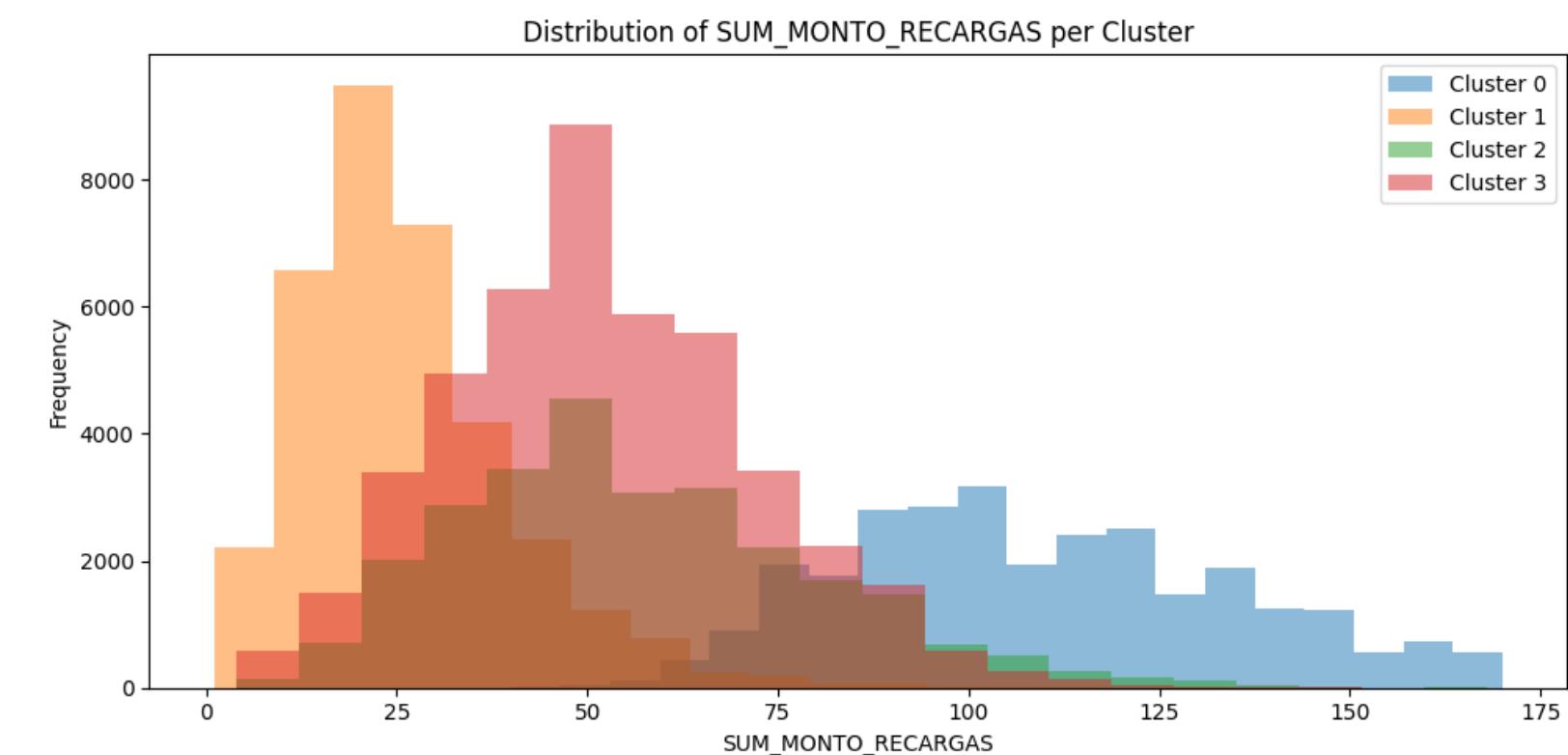
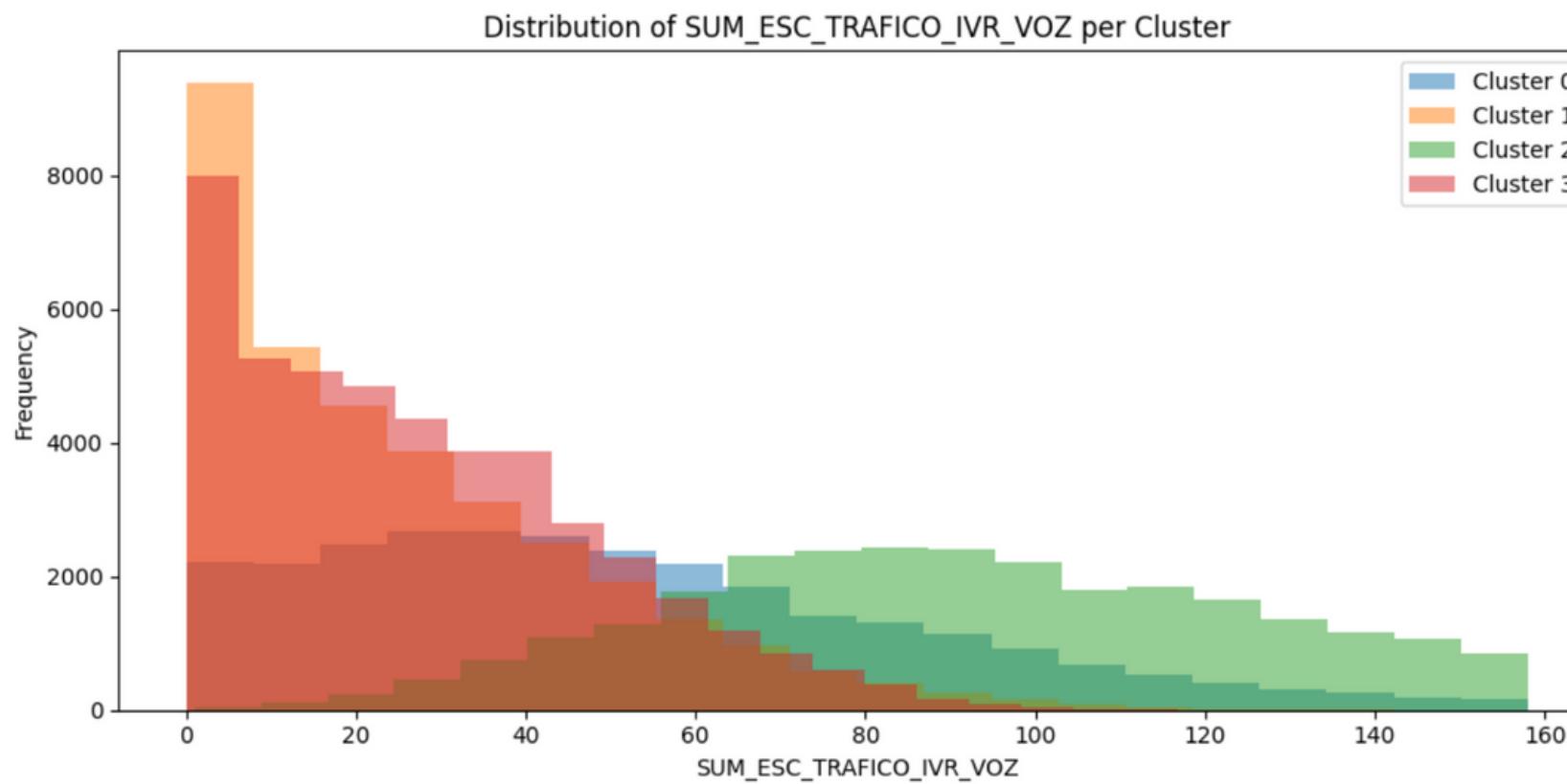
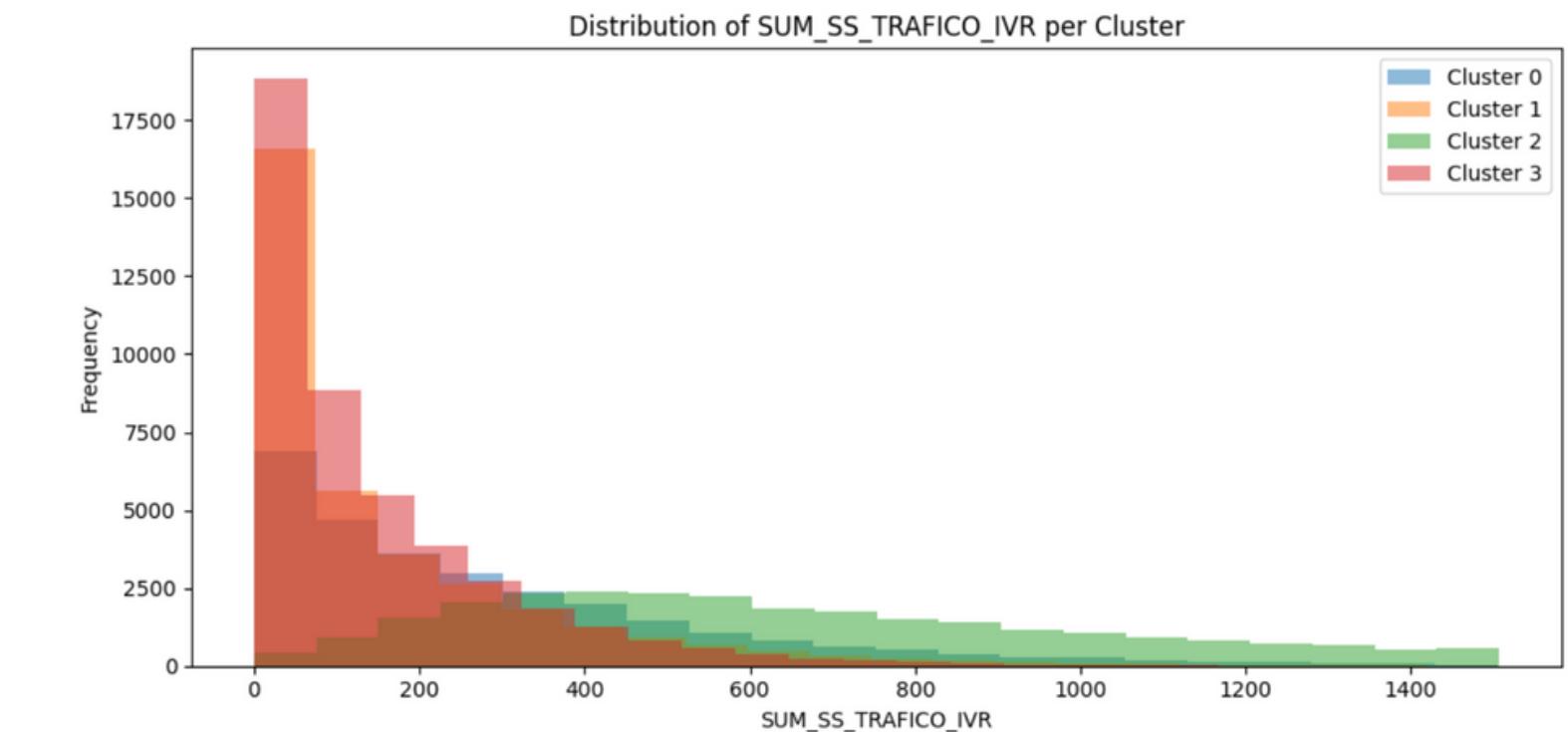
Cluster 1 Recargadores Cautelosos

Usuarios que se distinguen por su escasa frecuencia de recargas y montos moderados. Se caracterizan por mantener intervalos extensos entre recargas, alcanzando en promedio más de un mes entre cada transacción.



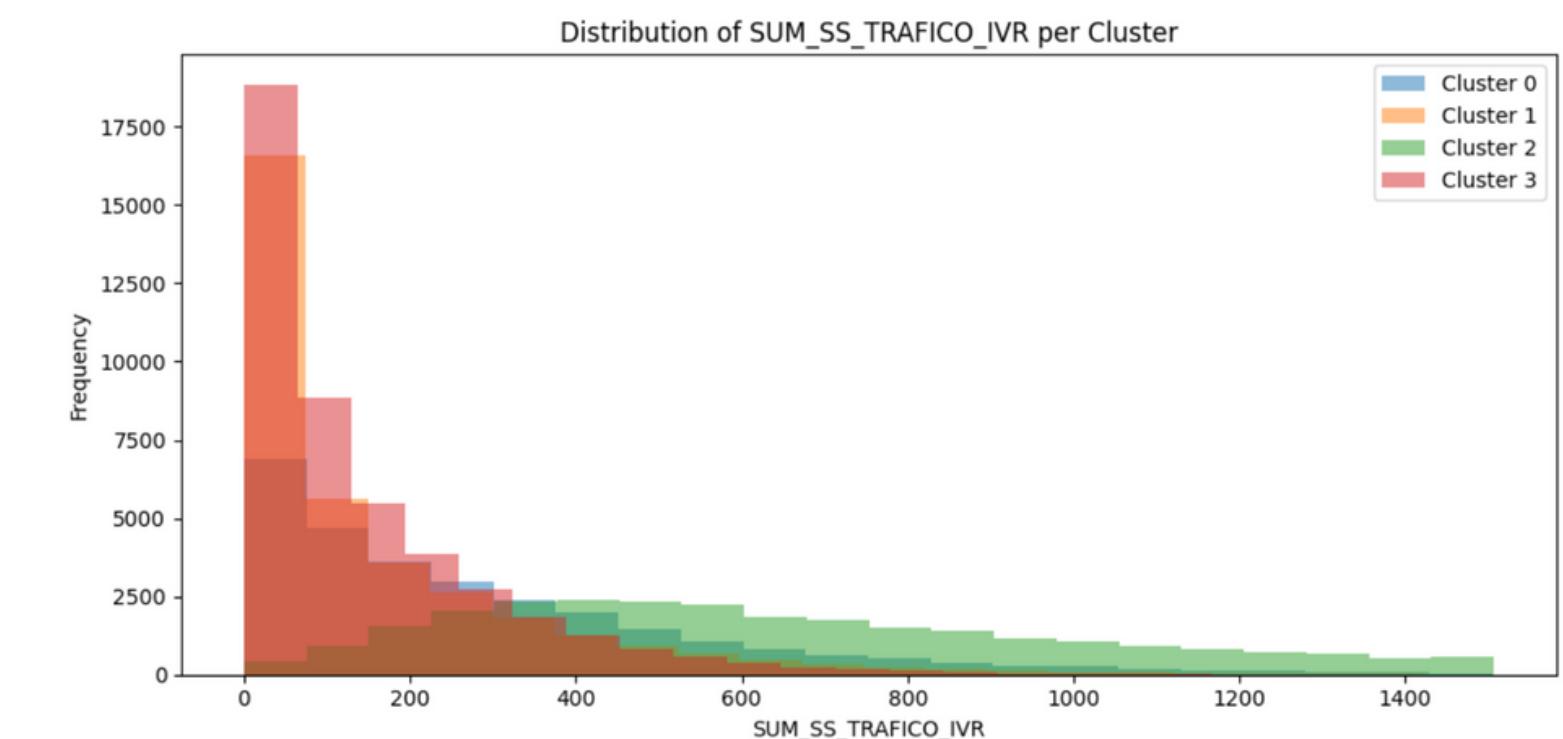
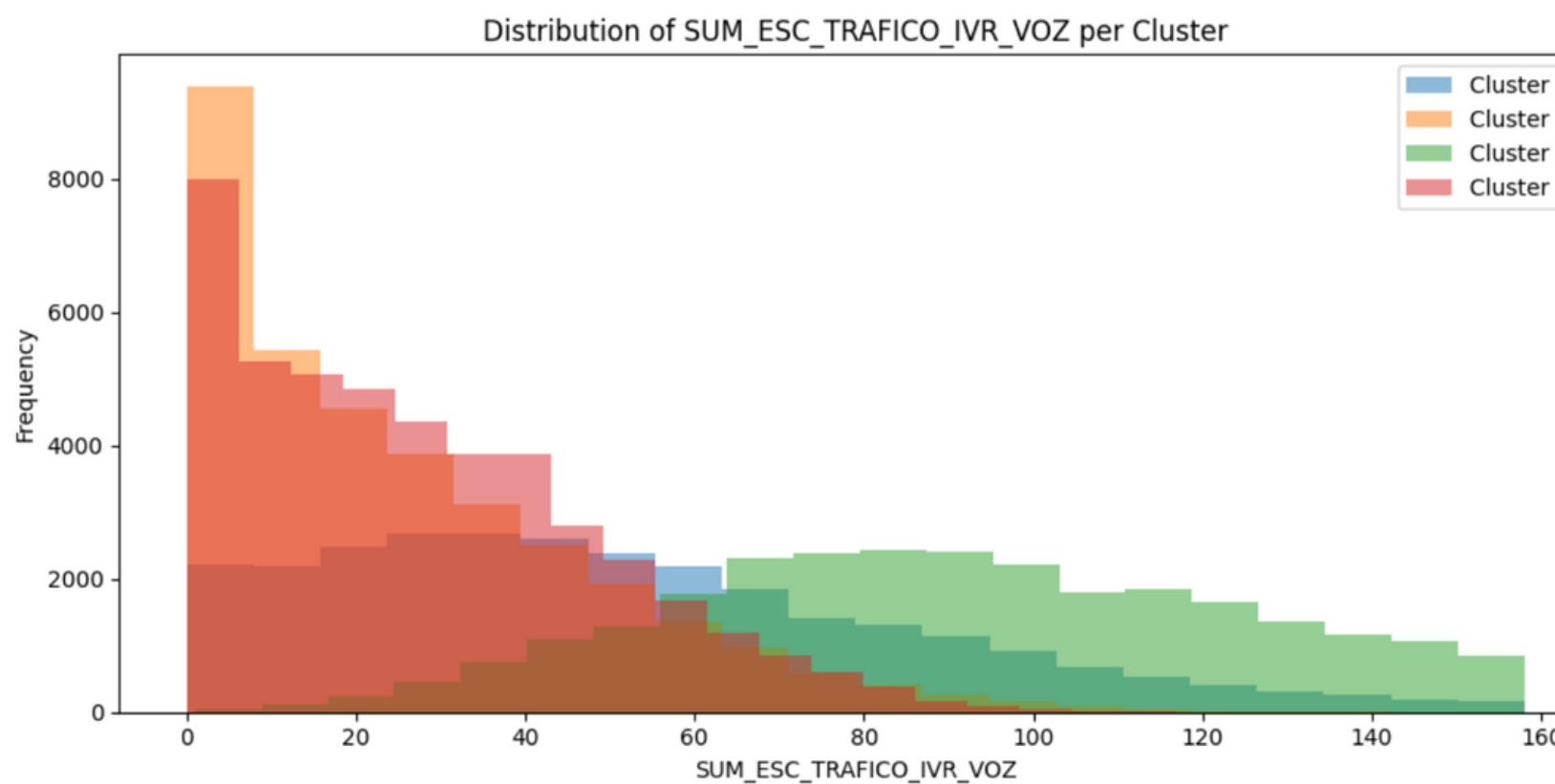
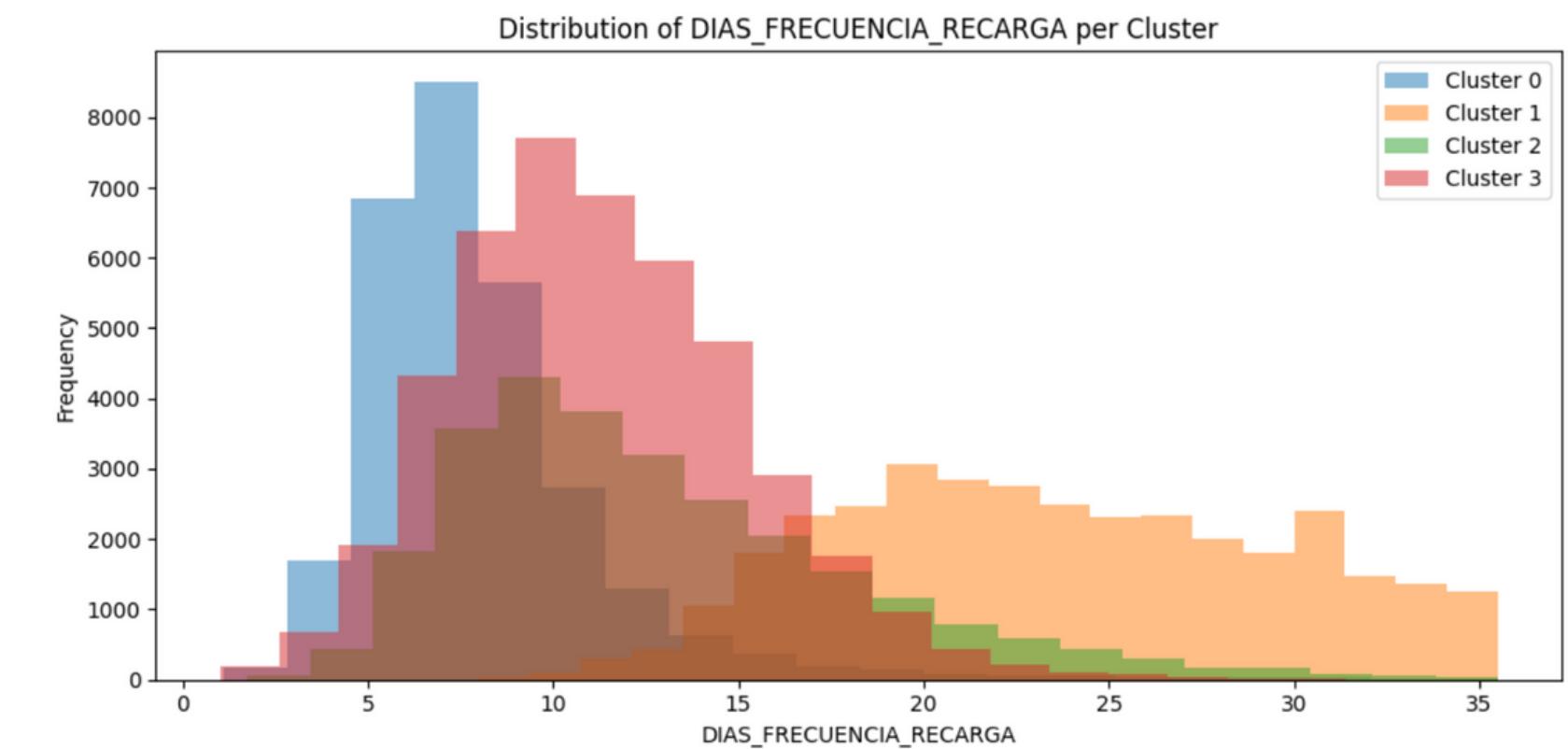
Cluster 2 Comunicadores Globales

Este grupo constituye un subconjunto del Cluster 3, destacándose como usuarios "promedio" en términos de recargas. Sin embargo, su singularidad radica en el uso frecuente de mensajes internacionales, mostrando una marcada preferencia por esta modalidad de comunicación. Aunque mantienen una prevalencia de minutos de llamadas nacionales, también acumulan una cantidad significativa de mensajes de voz en sus buzones.

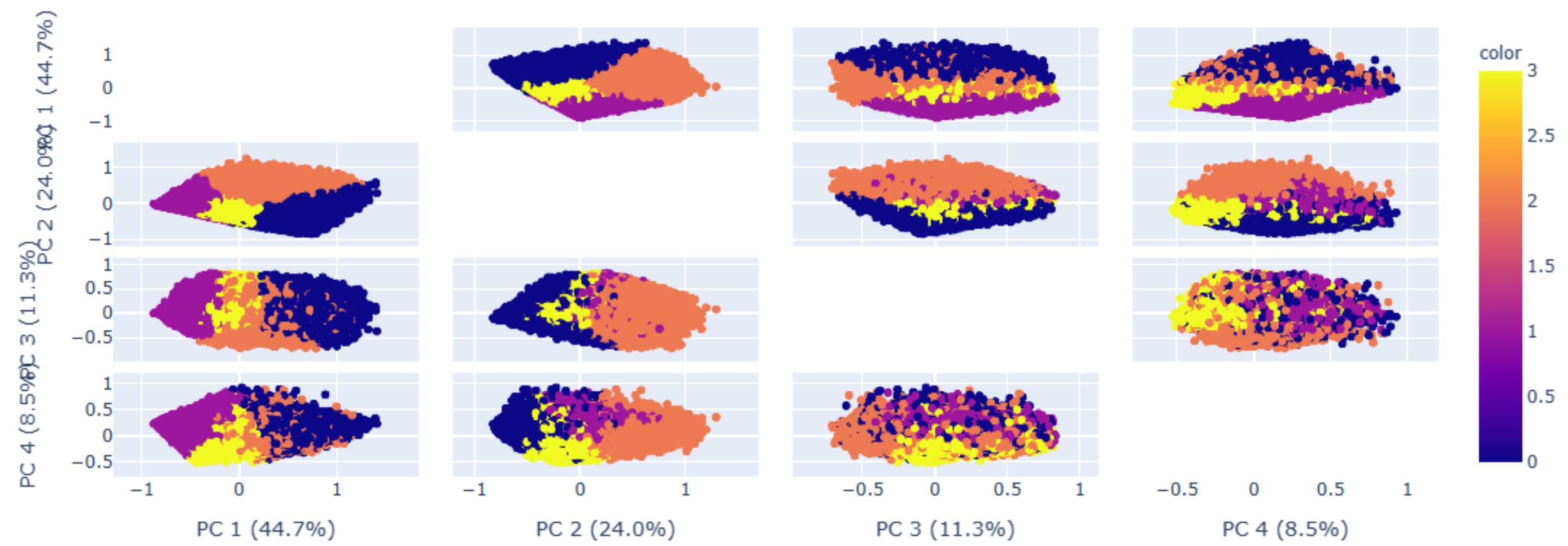
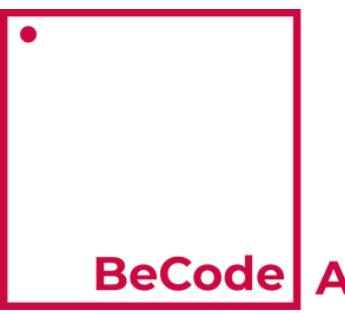


Cluster 3 Usuarios Estándar

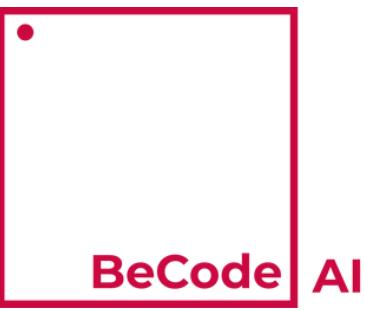
1. Usuarios promedio que recargan aproximadamente cada diez días, mostrando esporádico interés en los servicios de mensajería internacional y buzón de voz. Estos individuos recargan con moderación, ajustándose a sus necesidades y evitando gastos innecesarios.



PCA



Torch

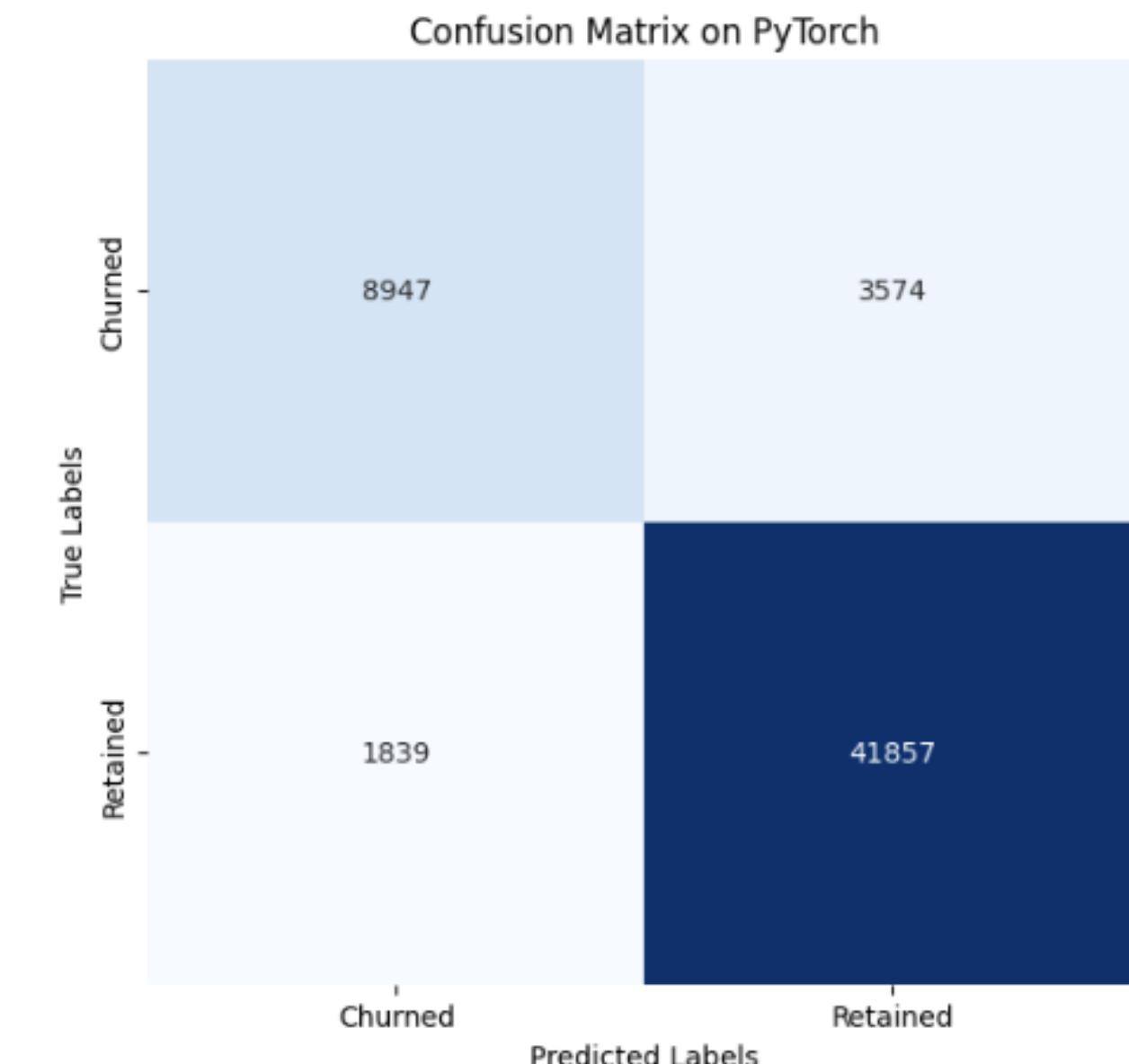


El primer modelo que escogimos fue Torch con la intención de implementar redes neuronales, se implemento con una sola capa oculta.

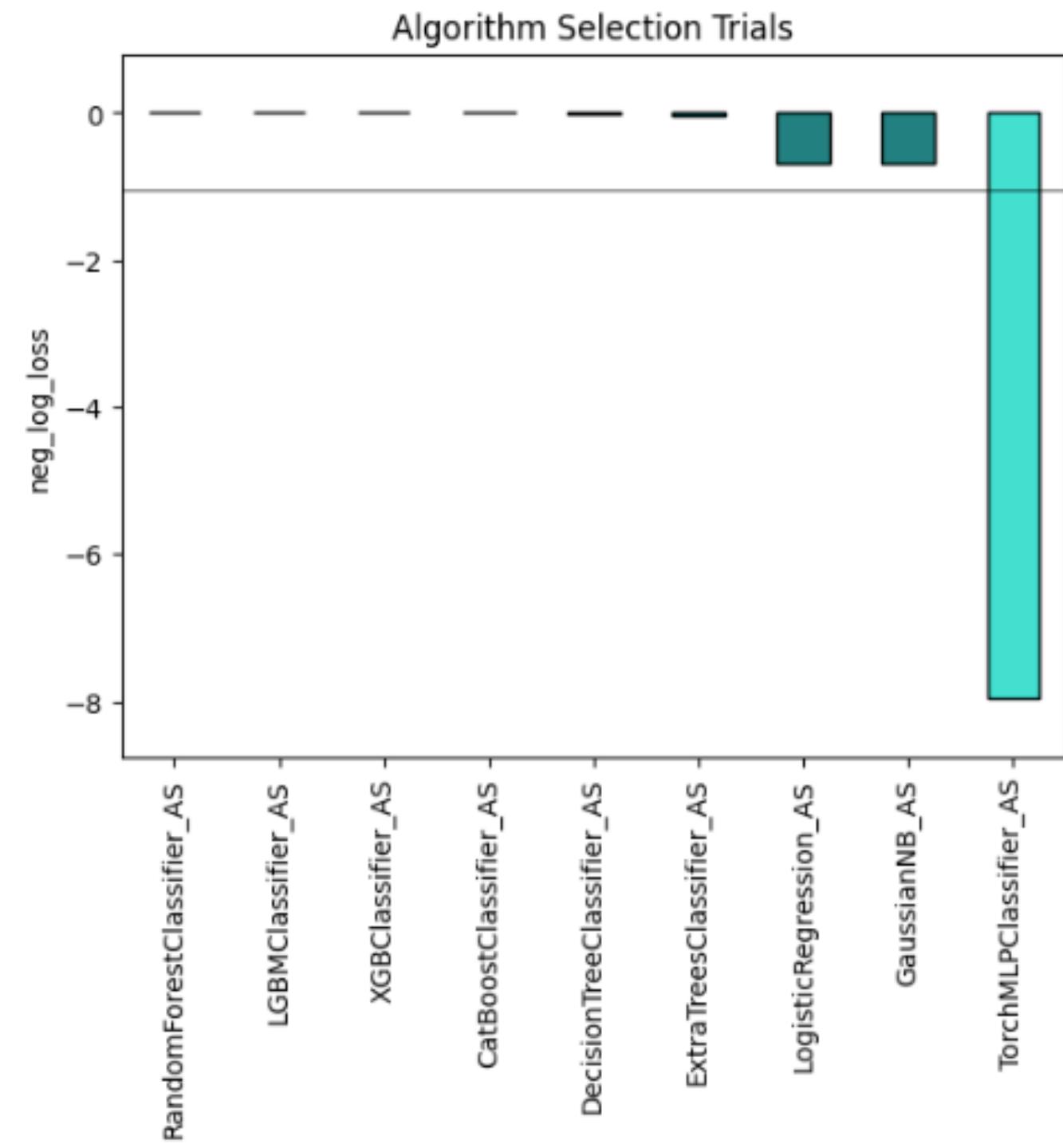
Los resultados que se obtuvieron con Torch fueron bastante buenos con una precisión del 0.90

- Binary Cross Entropy
- 1 “Capa Oculta” y 10 neuronas
- 32 samples por época
- 16 epochs

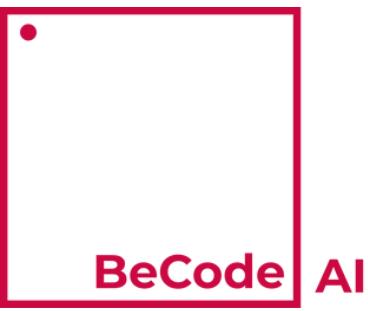
Accuracy score: 0.90
Churned precision: 0.83
Churned recall: 0.71
Retained precision: 0.92
Retained recall: 0.96



Auto ML



LGBM

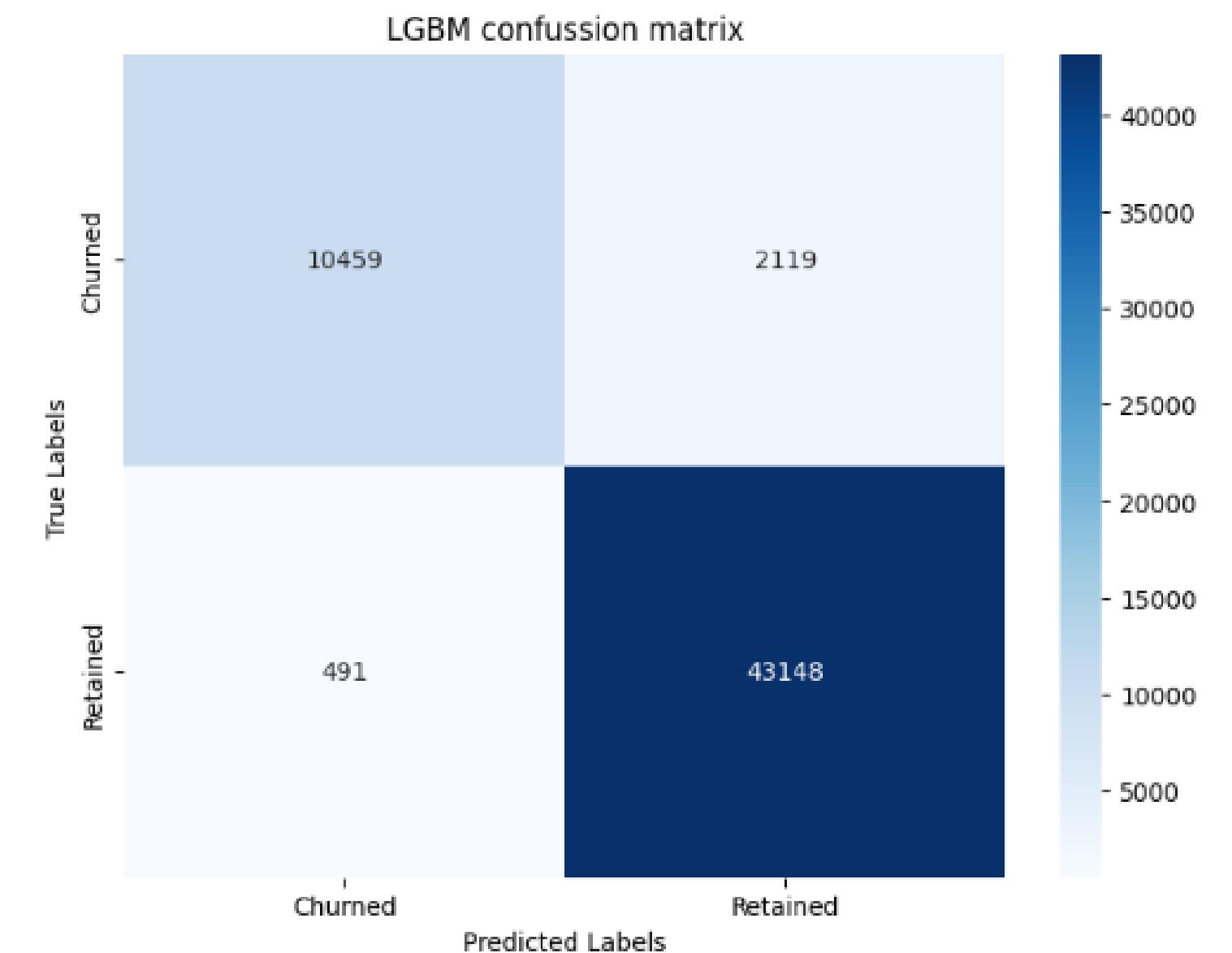


El segundo modelo que utilizamos fue LGBM (Light Gradient Boosting Machine) el cual es un algoritmo de aprendizaje automático basado en árboles de decisión .

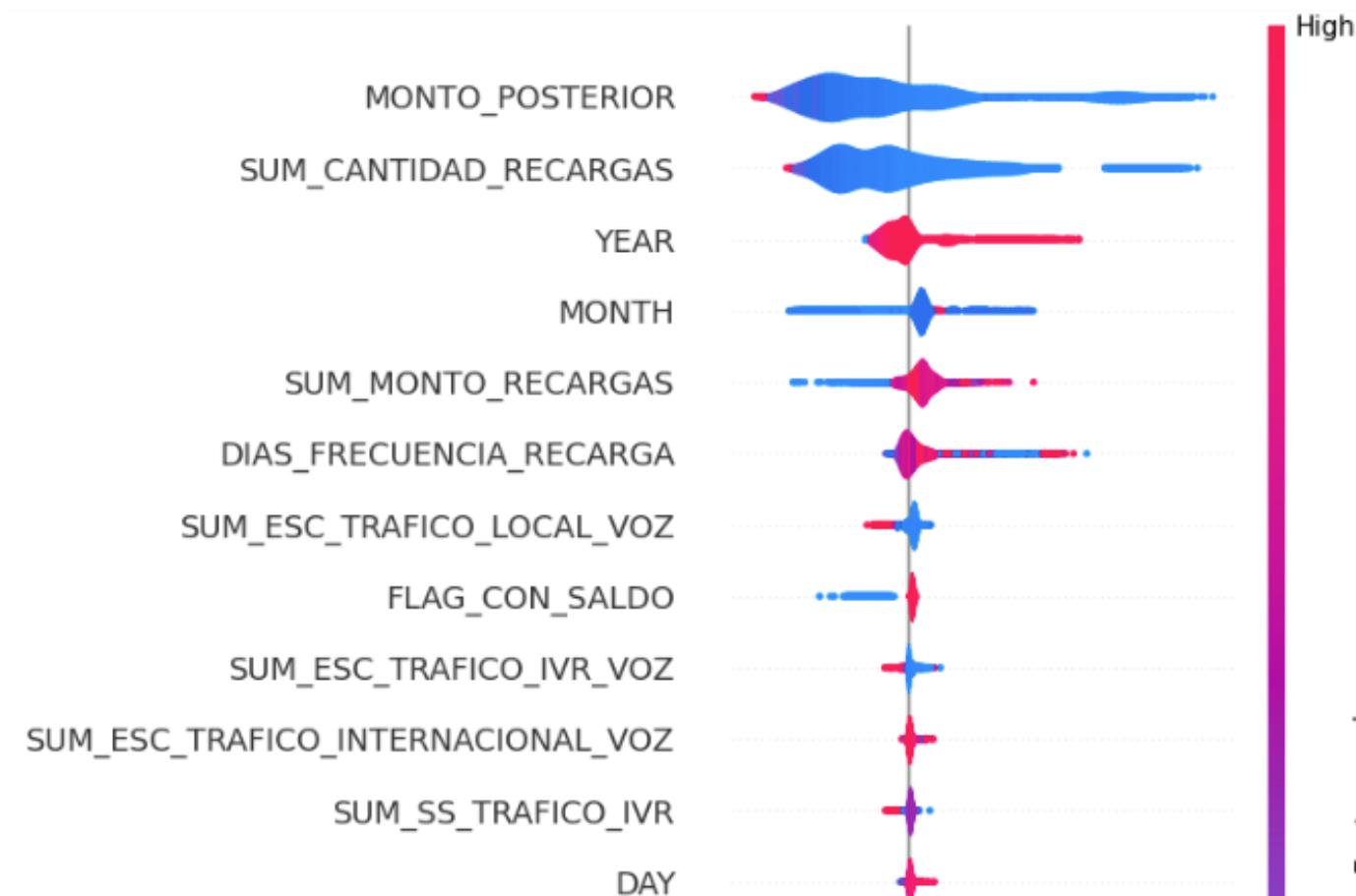
Este modelo fue uno de los recomendados por ADS, tras analizarlo y probarlo obtuvimos una precisión de 0.95.

- Captura relaciones no lineales
- Eficiencia
- Interpretabilidad
- Menos datos necesarios

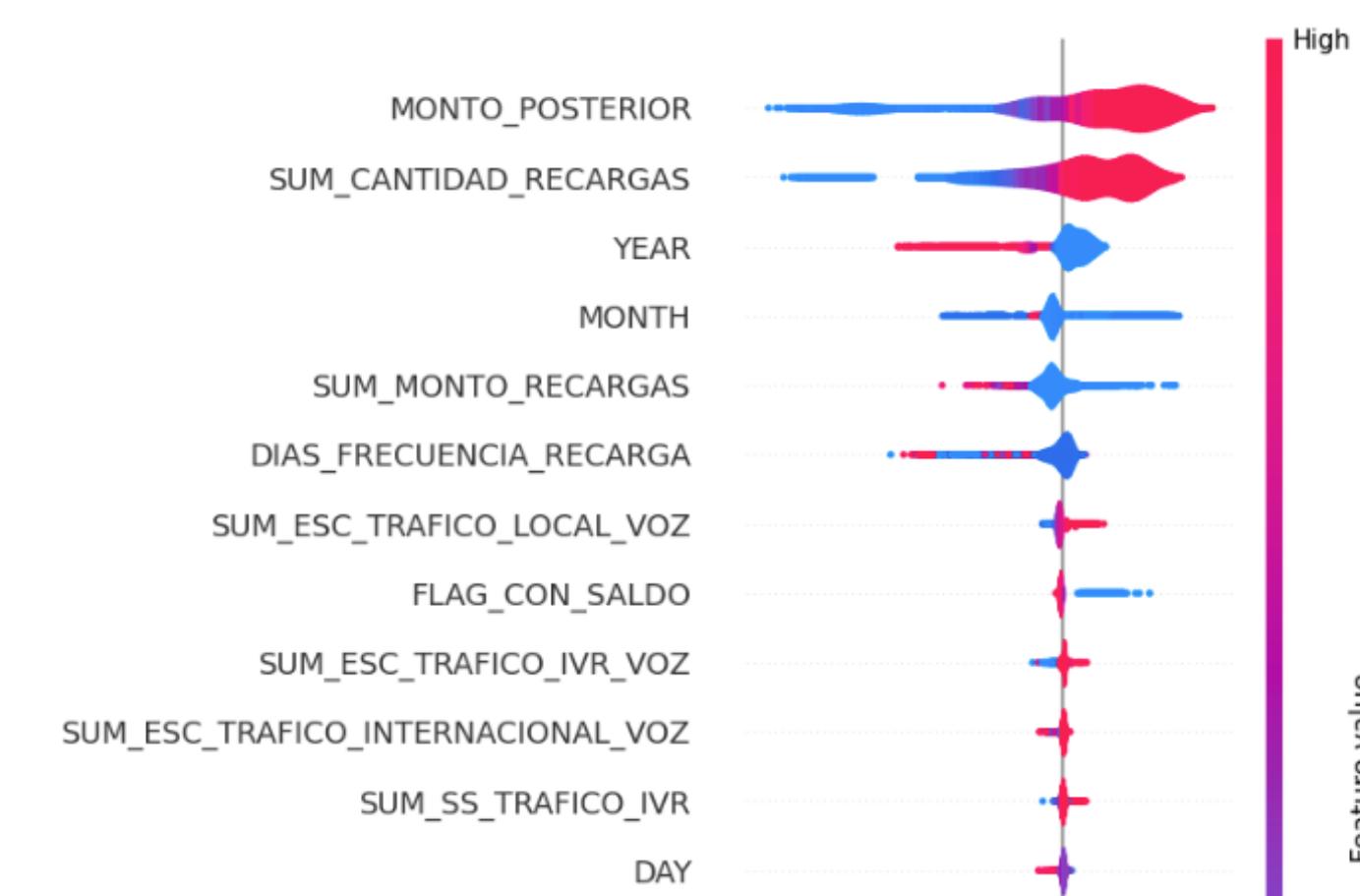
Accuracy score: 0.95
Churned precision: 0.96
Churned recall: 0.83
Retained precision: 0.95
Retained recall: 0.99



LGBM - Impacto en el modelo



Churned

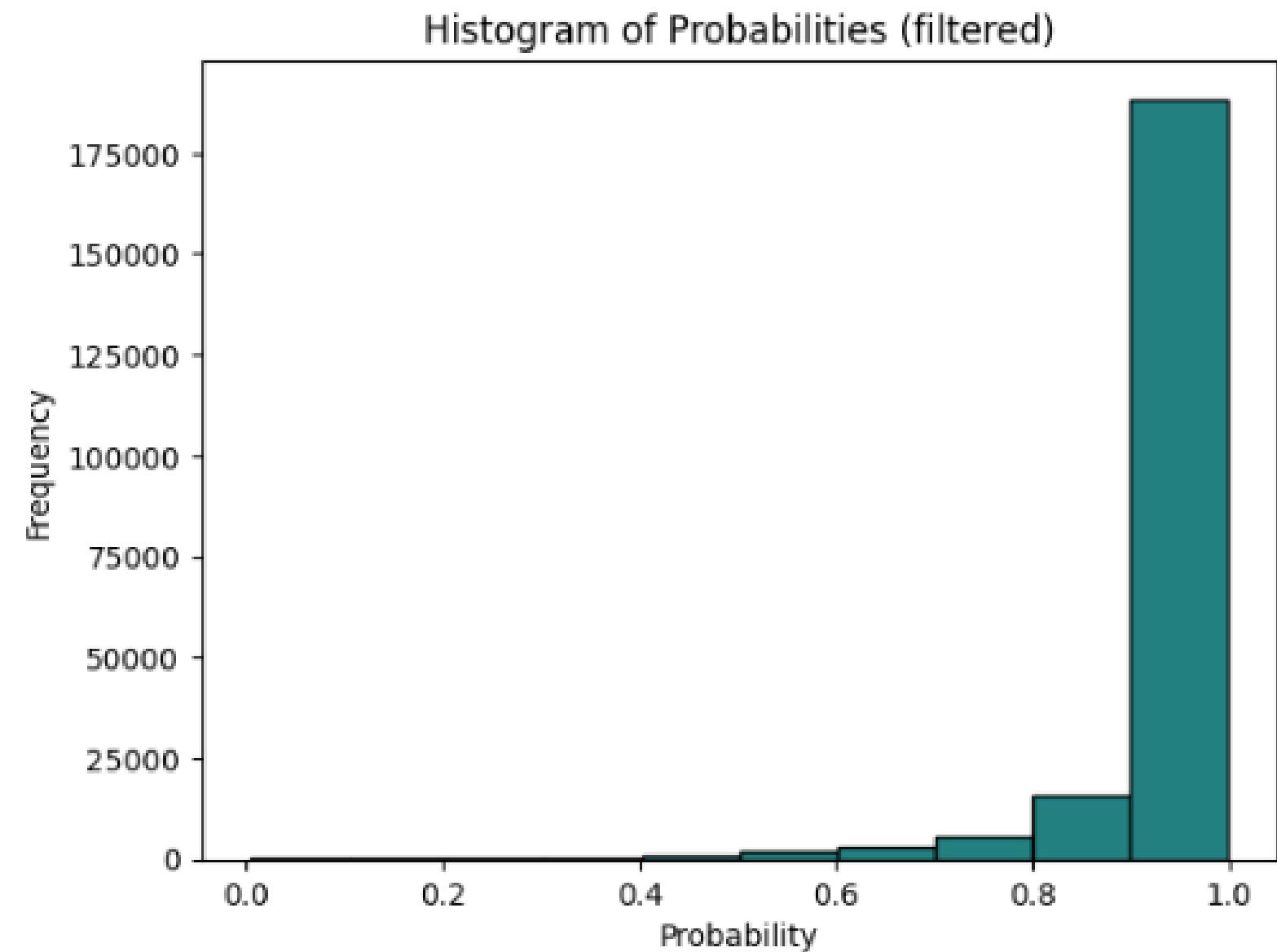


Cliente activo

Probabilidad de retener clientes activos

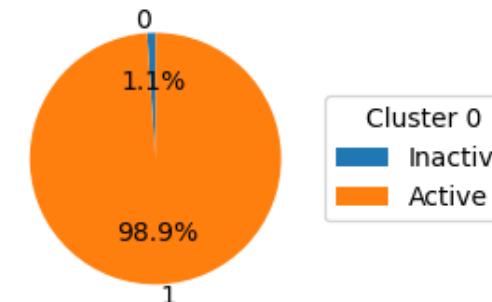
Realizamos un historgrama para obtener la probabilidad de retener a un cliente, dependiendo de la probabilidad se le asigna el nivel de riesgo:

- Bajo > 85%
- 85% > Moderado > 70%
- 70% > Alto > 50%
- 50% > Critico

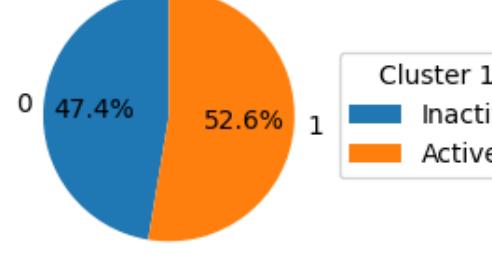


Probabilidad de retener clientes activos por Cluster

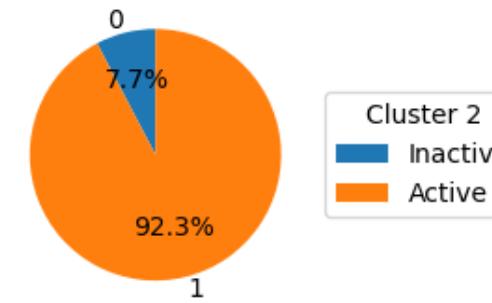
Histograms for Cluster 0



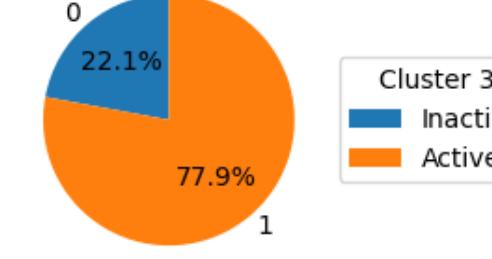
Histograms for Cluster 1



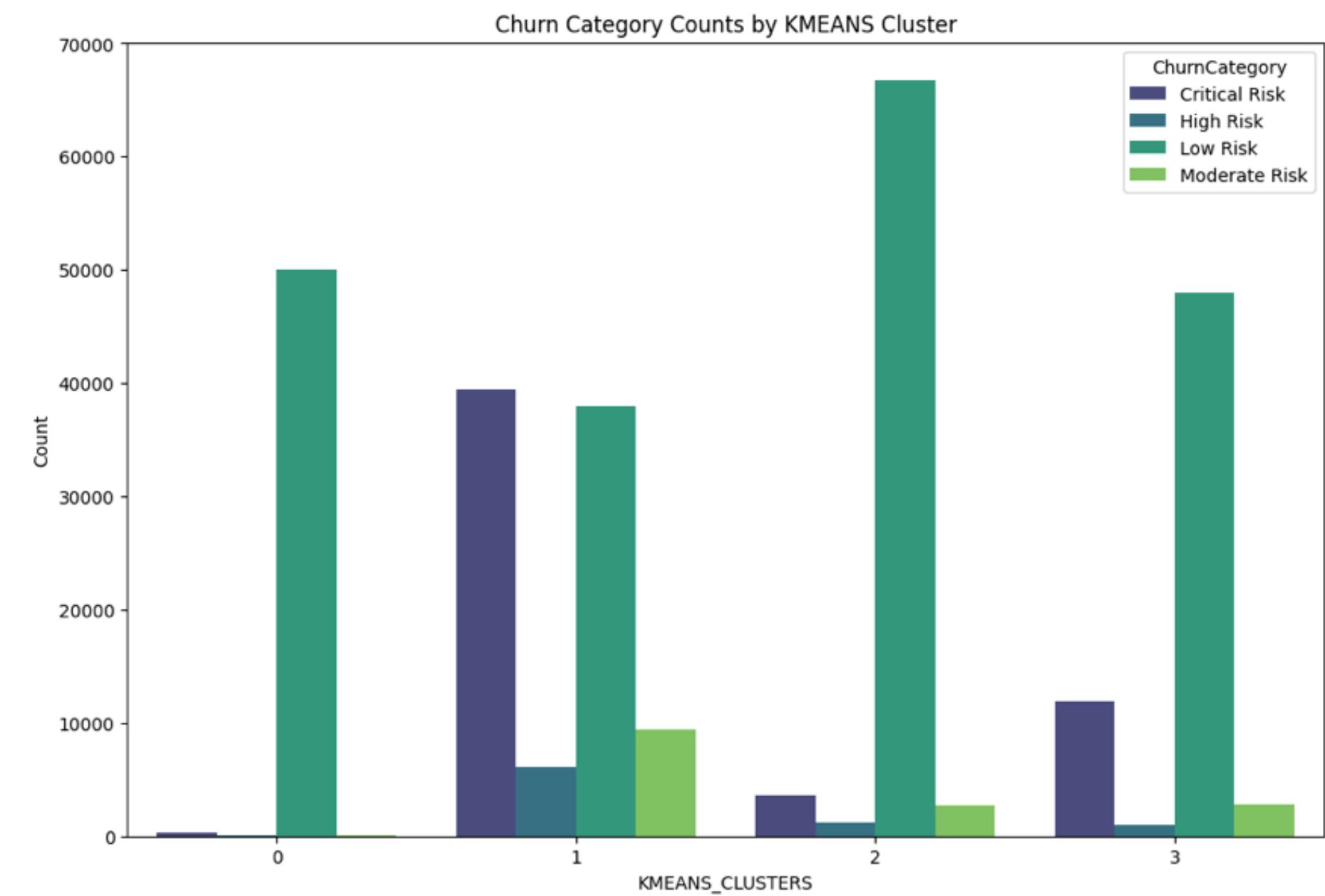
Histograms for Cluster 2



Histograms for Cluster 3

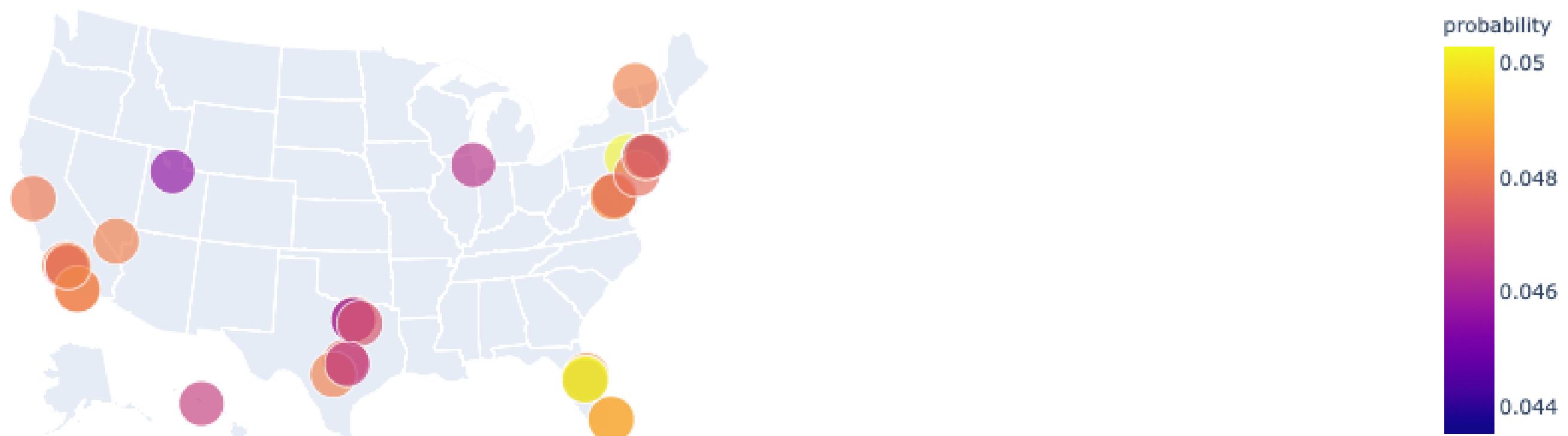


Pie Chart FLAG_CLIENTE_ACTIVO per cluster



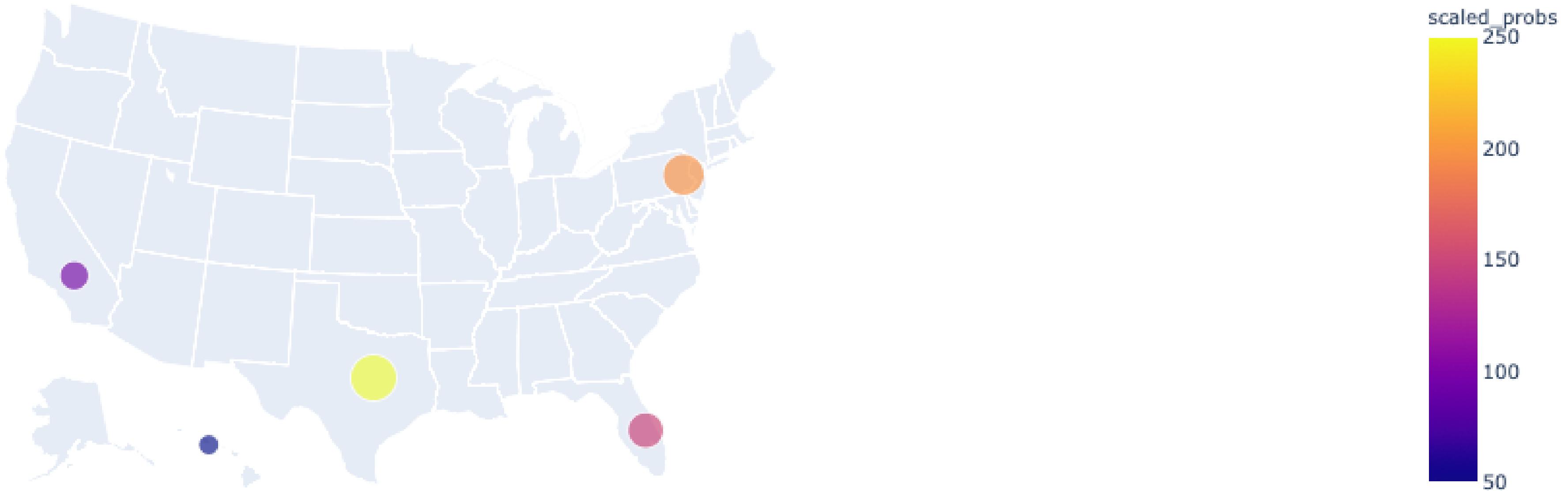
Probabilidad Churn por tiendas

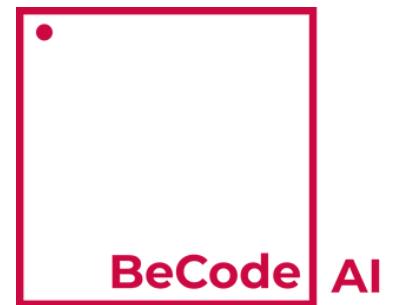
Mean Probability for Each Store



Regiones Churn

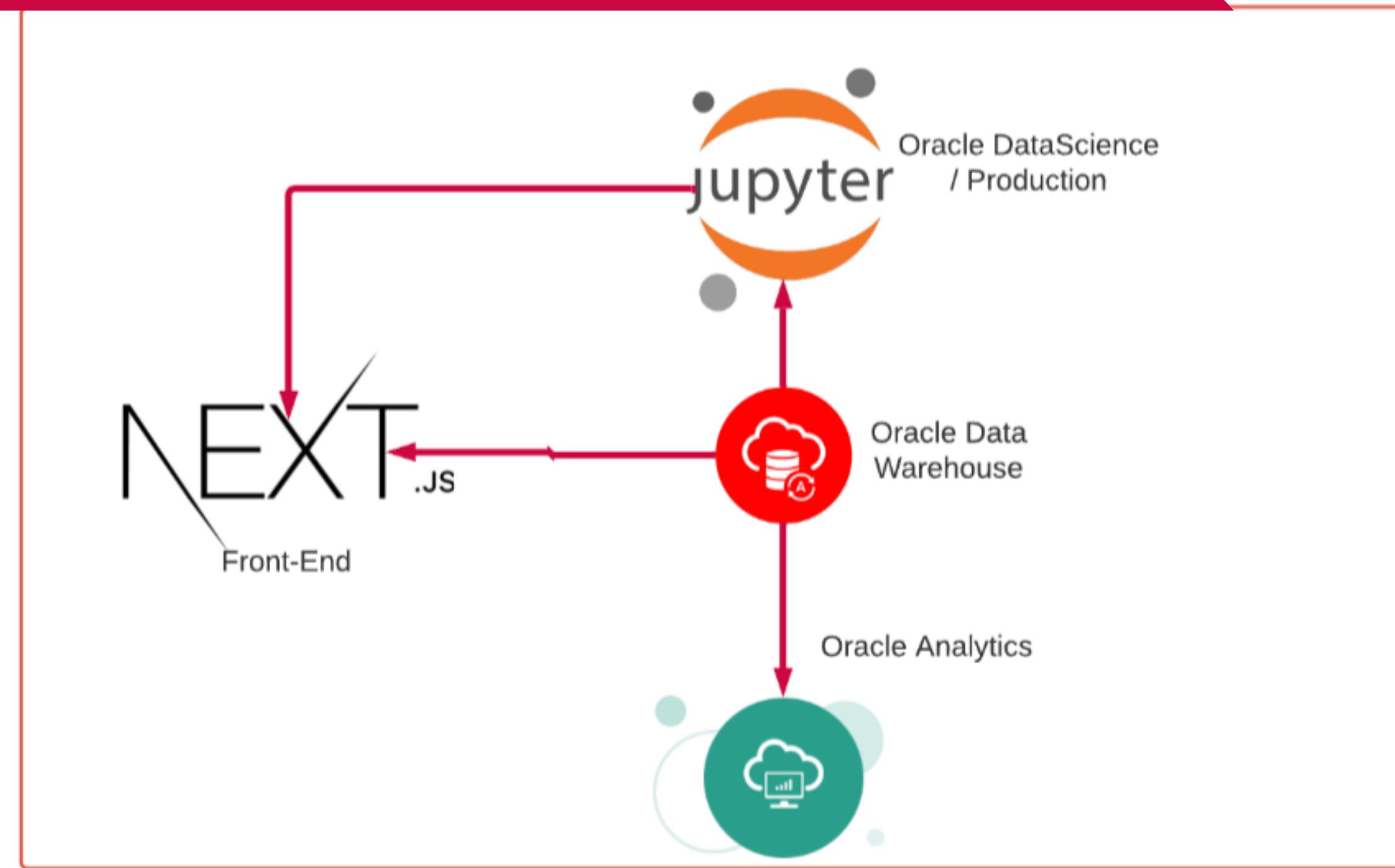
Probabilidad





Desarrollo Web

Arquitectura Plataforma Web



LLaMA

BeCode AI

POST <https://www.becode.software/api/llama/message> Send

Params Authorization Headers (9) **Body** • Pre-request Script Tests Settings Cookies

none form-data x-www-form-urlencoded raw binary GraphQL JSON [Beautify](#)

```
1 {  
2   "prompt": "This is a cluster of traveler users, who make use roaming services a lot. We must focus on generating marketing with promos regarding packages for travelers in america."  
3 }
```

Body Cookies Headers (6) Test Results Status: 200 OK Time: 12.65 s Size: 886 B Save as example ⚙️

Pretty Raw Preview Visualize JSON [Beautify](#)

```
1 {  
2   "message": "Here is a short SMS message as a promo for the clients of TELCO, targeting travelers who use roaming services a lot:\n\nHey Traveler! 🚀 Are you getting tired of high roaming bills while traveling in the US? 🛫 We have the perfect solution for you! 🎉 TELCO is offering exclusive roaming packages that will keep you connected without breaking the bank! 📱👍 Enjoy unlimited calls, texts, and data in the US, Canada, and Mexico for just $30/day! 🚀💰 Don't miss out! Contact us now to learn more and start traveling happy! 😊 #TelcoRoaming #TravelHappy #SaveMoney #StayConnected"  
3 }
```



BeCode AI

Plataforma Web

Front end

En el desarrollo de la página web, elegimos el framework React con la extensión Next debido a su capacidad para la creación de APIs y su alineación con el sistema de React, proporcionando acceso a diversas librerías. Next facilitó la visualización de gráficas, siendo crucial para presentar de manera efectiva los datos recopilados.

Aprovechando la versatilidad de Next, incorporamos Twilio para gestionar mensajes a través de WhatsApp, así como el envío y almacenamiento de imágenes y archivos, ampliando nuestras capacidades de comunicación y almacenamiento de manera integral. Para la gestión de correos electrónicos, implementamos Nodemailer, una herramienta esencial para enviar mensajes a través del correo mediante conexiones SMTP con nuestro servidor de correo bot, permitiéndonos enviar correos electrónicos de manera eficaz y ampliar nuestras opciones de comunicación.



Base de dato

Utilizamos dos plataformas diferentes para la gestión de la base de datos: Oracle y Supabase. Oracle desempeña un papel crucial en el almacenamiento de datos para el entrenamiento de modelos, principalmente utilizando archivos CSV y aprovechando la facilidad de establecer conexiones desde el notebook que gestiona Oracle. Por otro lado, Supabase se implementa para la autenticación de usuarios y el almacenamiento de datos más orientados a la plataforma, como la información de usuarios, historial de distribución de marketing y distribución de diferentes usuarios.



Back-end

Flask es una parte fundamental de nuestra arquitectura, integrándose con Python y proporcionando diversas bibliotecas para facilitar el análisis de datos. Actúa como un puente esencial al conectar la base de datos Oracle con la interfaz frontal de nuestra plataforma web, permitiendo gestionar eficientemente el flujo de datos a través de la aplicación.

En el contexto de Flask, implementamos un pipeline que simplifica la realización de predicciones con los datos recibidos. Este proceso se replica en el caso de Llama, donde recibimos información y solicitudes, generando respuestas esperadas o predicciones necesarias. La versatilidad de Flask como framework nos permite manejar eficazmente tanto las solicitudes de datos como las predicciones resultantes.



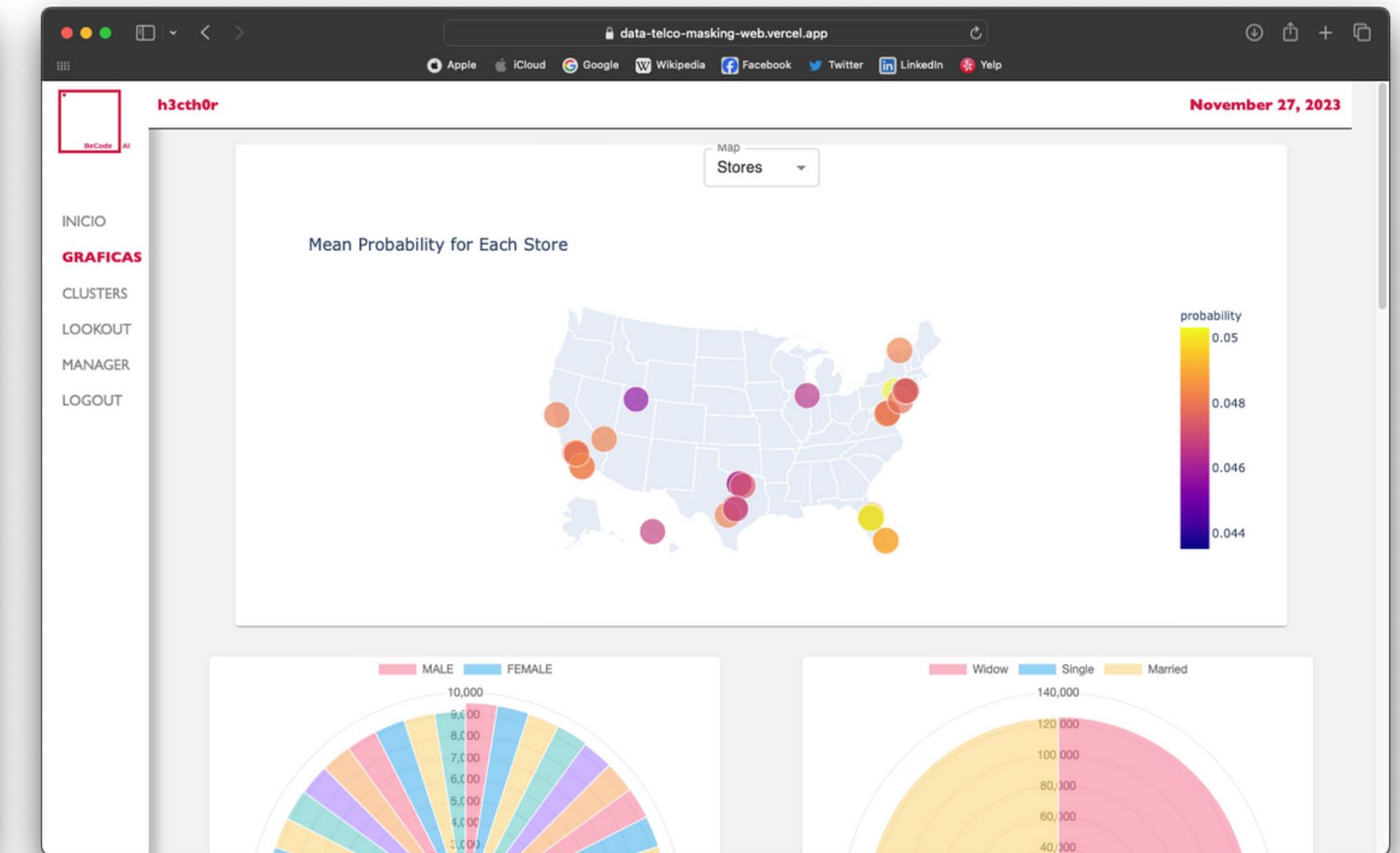
The screenshot shows the 'INICIO' tab of the application. At the top, there's a welcome message 'Welcome h3cth0r !' with a clapping hands emoji. Below it is a section titled 'Team Work' with three people emoji, containing a table of contributions:

ID	Cluster	Date	User	Type	Description
2	Comunicadores Globales	date	user_1	text	We hope this m...
5	Recargadores Cautelosos	date	h3cth0r	text	Ready to doubl...
7	Gasto Continuo	date	misaelchavez16	Text	Good news, yo...
0	Recargadores Prodigiosos	date	h3cth0r	text	Greeting Custo...
1	Comunicadores Globales	date	user_1	text	To claim your e...
3	Recargadores Cautelosos	date	German	text	Muchas ofertas
4	Usuarios Estándar	date	German	text	Muchas ofertas
8	Usuarios Estándar	date	German	text	Muchas ofertas

On the right, there's a section titled 'Your Contributions' with a bee emoji, showing two entries:

ID	Date	Type	Cluster
5	date	text	Recargadores Cautelosos
0	date	text	Recargadores Prodigiosos

The top right corner shows the date 'November 27, 2023'.



En el menú principal, el usuario encontrará dos tablas que le permitirán visualizar las contribuciones de marketing realizadas por su equipo de marketing.

En la pestaña de gráficos, se pueden explorar diversos gráficos que presentan datos de manera integral, incluyendo integraciones con nuestros modelos conectados.

Clusters

ID	Low	Moderate	High	Critical	Label	Actions
2	37917	9378	6130	39404	Recargadores Cautelosos	<i>info</i> <i>edit</i>
3	66739	2692	1225	3564	Comunicadores Globales	<i>info</i> <i>edit</i>
4	47936	2766	977	11883	Usuarios Estándar	<i>info</i> <i>edit</i>
1	50036	119	37	281	Recargadores Prodigiosos	<i>info</i> <i>edit</i>

Recargadores Cautelosos

Description

Personas que utilizan las recargas que realizan para realizar llamadas

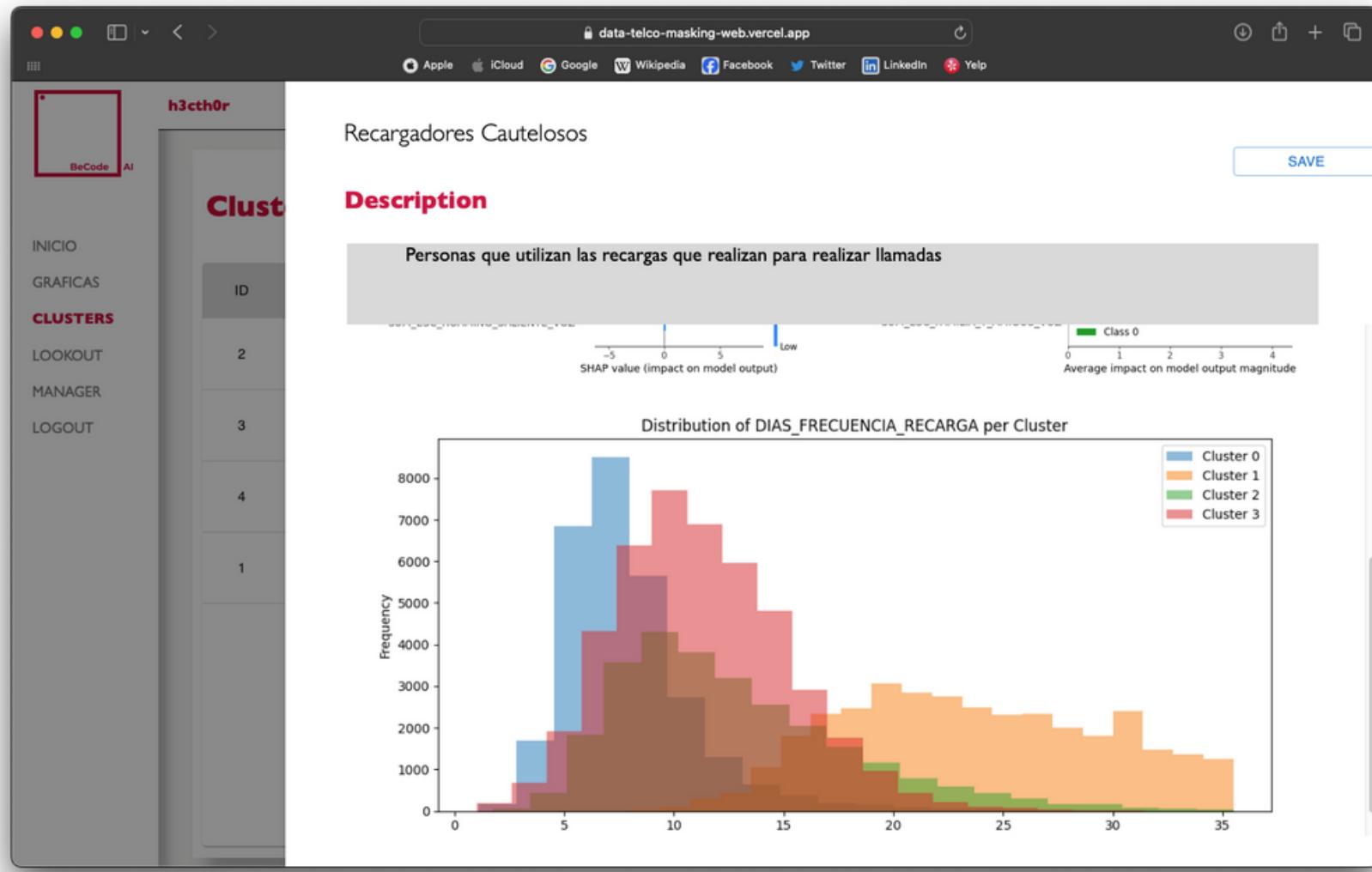
Feature	Mean	Std	Min	25%	50%	75%	Max
MONTO_ANTERIOR	4.98	7.50	0.00	0.00	2.00	7.00	112.00
DAY	15.91	8.74	1.00	8.00	16.00	23.00	31.00

SHAP Feature Importance Plot:

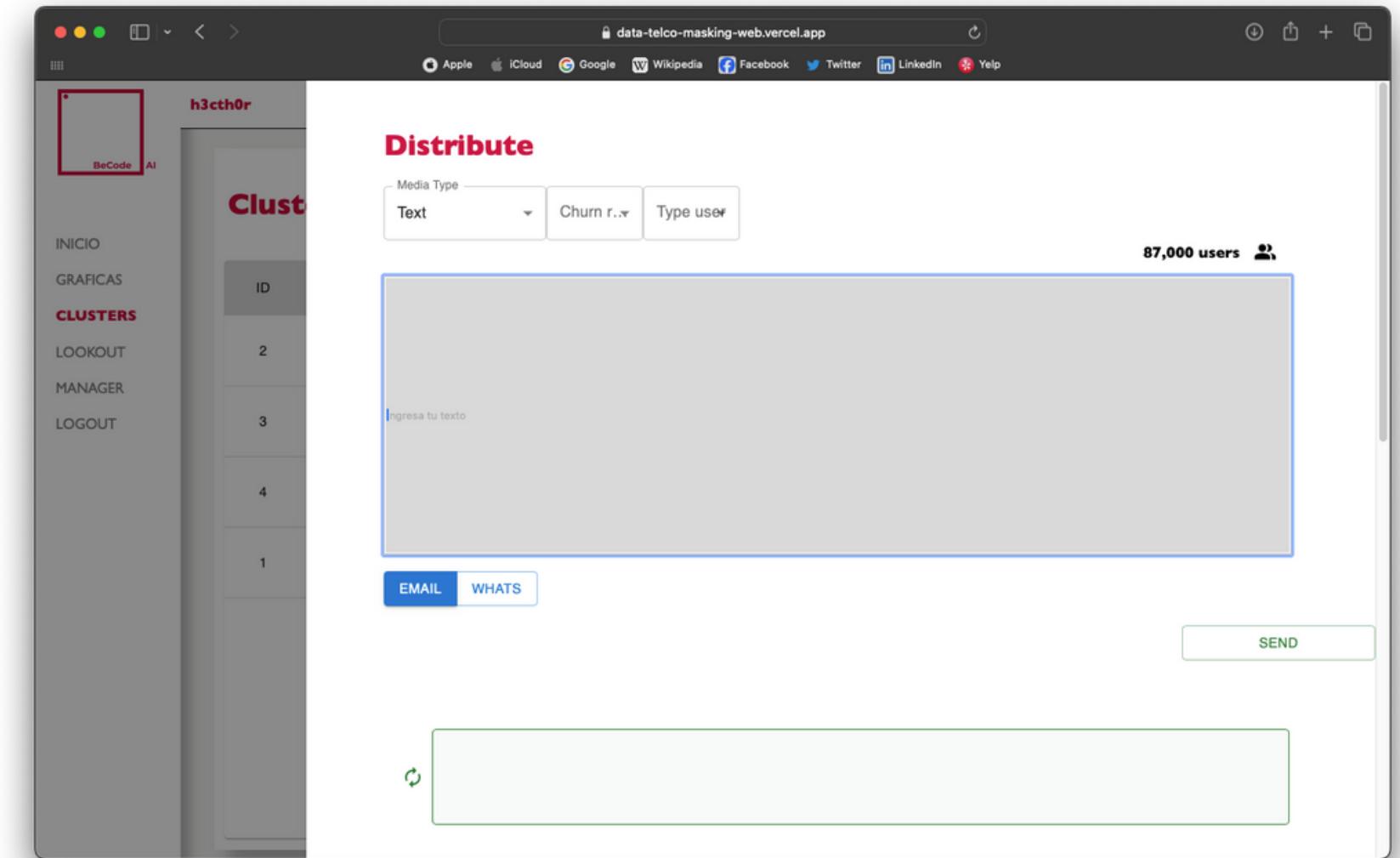
- Y-axis: DIAS_FRECUENCIA_RECARGA, SUM_CANTIDAD_RECARGAS, SUM_MONTO_RECARGAS, MONTO_POSTERIOR, SUM_ESC_TRAFICO_IVR_VOZ, SUM_ESC_TRAFICO_LOCAL_VOZ, SUM_SS_TRAFICO_IVR, SUM_ESC_TRAFICO_INTERNAZIONAL_VOZ, SUM_ESC_ROAMING_ENTRANTE_VOZ, SUM_ESC_FAMILY_Y_AMIGOS_VOZ, SUM_ESC_ROAMING_SALIENTE_VOZ.
- X-axis: SHAP value (impact on model output magnitude), ranging from -5 to 5.
- Legend: Class 1 (blue), Class 3 (purple), Class 2 (red), Class 0 (green).
- Plot shows that Class 1 (blue) has the highest impact across most features, particularly DIAS_FRECUENCIA_RECARGA and SUM_CANTIDAD_RECARGAS.

En la tabla de clusters, se exhiben los clusters identificados en nuestro modelo de segmentación de usuarios, proporcionando detalles como el nombre de su etiqueta y la cantidad de personas en cada uno, clasificadas según sus niveles de riesgo.

En la opción de informe para cada cluster, se presenta información más detallada del mismo, permitiendo al usuario llevar a cabo un análisis propio. Se brinda la posibilidad de editar el nombre del cluster de acuerdo con la evaluación personal del usuario.



Aquí se exhiben gráficos adicionales que también son presentados en el informe.



En la sección de distribución, dentro de este modal, se facilita la distribución de marketing a los clientes mediante el uso de filtros que permiten enfocarse en usuarios específicos. La solución puede distribuir contenido de marketing a través de medios como email y WhatsApp.

The screenshot shows the "Lookout" tab of the application. On the left, there's a sidebar with navigation links: INICIO, GRAFICAS, CLUSTERS, LOOKOUT (which is highlighted in red), MANAGER, and LOGOUT. The main area has a header with the date "November 27, 2023". A search bar at the top contains the ID "PX38513830". Below it, there's a large circular profile picture of a person labeled "Client". To the right of the profile are several data fields: Name (Elizabeth), Phone (+1654123987), LastName (Taylor), email (elizabethtaylor@example.com), Location (Dallas, TX), and Type of User (Recargadores Cautelosos). A yellow box highlights the Churn Risk: Moderate - 0.2357. At the bottom, there's a placeholder box for a recommendation.

En la pestaña "Lookout", es posible buscar a un usuario por su DNI para obtener información específica, que incluye datos calculados mediante nuestros modelos, como la probabilidad de Churn y el tipo de usuario o el cluster al que pertenece. Esta herramienta está diseñada especialmente para que el equipo de soporte pueda abordar situaciones específicas de manera efectiva.

The screenshot shows the "MANAGER" tab of the application. On the left, there's a sidebar with navigation links: INICIO, GRAFICAS, CLUSTERS, LOOKOUT, MANAGER (which is highlighted in red), and LOGOUT. The main area has a header with the date "November 27, 2023". The title "Team Members" is displayed above a table. The table has columns: User, Team_ID, Role, number of post, and Actions. The data in the table is as follows:

User	Team_ID	Role	number of post	Actions
German	hg432	dsc	16	...
misaelchavez16	hg432	marketing	1	...
h3cth0r	hg432	manager	2	...
user_1	hg432	marketing	2	...

Se han incorporado herramientas de gestión de equipos y usuarios para transformar esto en una plataforma plenamente funcional.

BeCode AI

Conclusiones

Conclusiones del proyecto

Nuestra entrega considera varios equipos de marketing para un mismo set de clusters, pues consideramos que diferentes regiones pueden interpretar de manera diferente los resultados. Sin embargo, se podría escalar a futuro para almacenar diferentes modelos para diferentes empresas.

De igual manera, consideramos que a futuro se podrían implementar las siguientes propuestas:

- Agregar módulos de análisis. Una manera dinámica de agregar diferentes gráficas a su propio análisis, seleccionando los atributos y clusters a tomar en cuenta.
- Diferentes maneras de clusterizar. Proponer diferentes maneras de clusterizar: usando herramientas como UMAP o PCA, o una simple selección de atributos para alterar los clusters resultados, y dar la opción al equipo de marketing de escoger el que más se acomode a sus objetivos.
- Utilizar un mejor LLM. Usar un mejor LLM, como LLaMA con 70 billones de parámetros, para poder tener una mejor precisión a la hora de hacer recomendaciones de mensajes y promociones

Conclusiones personales

Este proyecto fue bastante retador. Debido a la naturaleza del mismo, tuvimos que adentrarnos a hacer un análisis de datos profundo. Esto fue complicado, ya que el mismo dataset con el que estuvimos trabajando contaba con una serie de comportamientos inusuales.

Es así que nos dimos cuenta de la importancia que tiene el proceso de análisis de datos, pues fue la parte del proyecto en la que más nos vimos trabajando. Desde entender por completo nuestro dataset y sus comportamientos, hasta ser capaces de tomar decisiones en base a nuestro conocimiento y evidencias empíricas.

En general, estamos muy orgullosos de lo que pudimos lograr en este proyecto, pues fuimos capaces de tomar los conocimientos que ya teníamos y aplicarlos a un entorno nuevo. Fuimos capaces de aprender y tomar decisiones con respecto a los temas de la materia por nuestra cuenta.