

PROYECTO CIENCIA DE DATOS TELCO CHURN

Andrés A. Capo Plaza

CODERHOUSE

Comisión 42390

INTRODUCCIÓN DEL PROYECTO



Se descargó un DataSet de Kaggle sintético el cual muestra los datos de 7.043 clientes, de una empresa dedicada a prestación de servicio de telecomunicaciones y streaming a través del cual se describen diversas variables que pueden tener los clientes repartidas en 21 columnas.



El principal objetivo de esta investigación es utilizar Machine Learning para predecir qué clientes abandonarán la empresa. Esto proporcionará información valiosa para la toma de decisiones comerciales, permitiendo la implementación de estrategias efectivas de retención de aquellos clientes con mayor probabilidad de darse de baja.

ANÁLISIS GRÁFICO

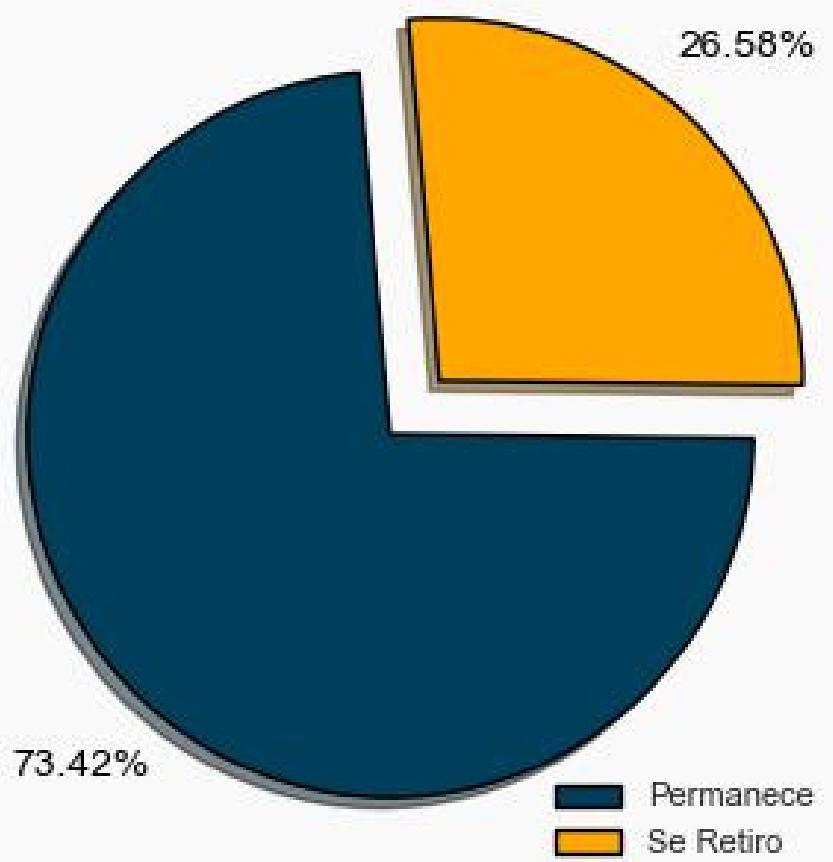
Por medio de los gráficos se pretende describir el comportamiento de los clientes así como sus preferencias, determinar si existe algún patrón que permita inferir el motivo de la baja.

En los siguientes gráficos se analizan las variables de manera segmentada si el cliente se dio o no la baja.

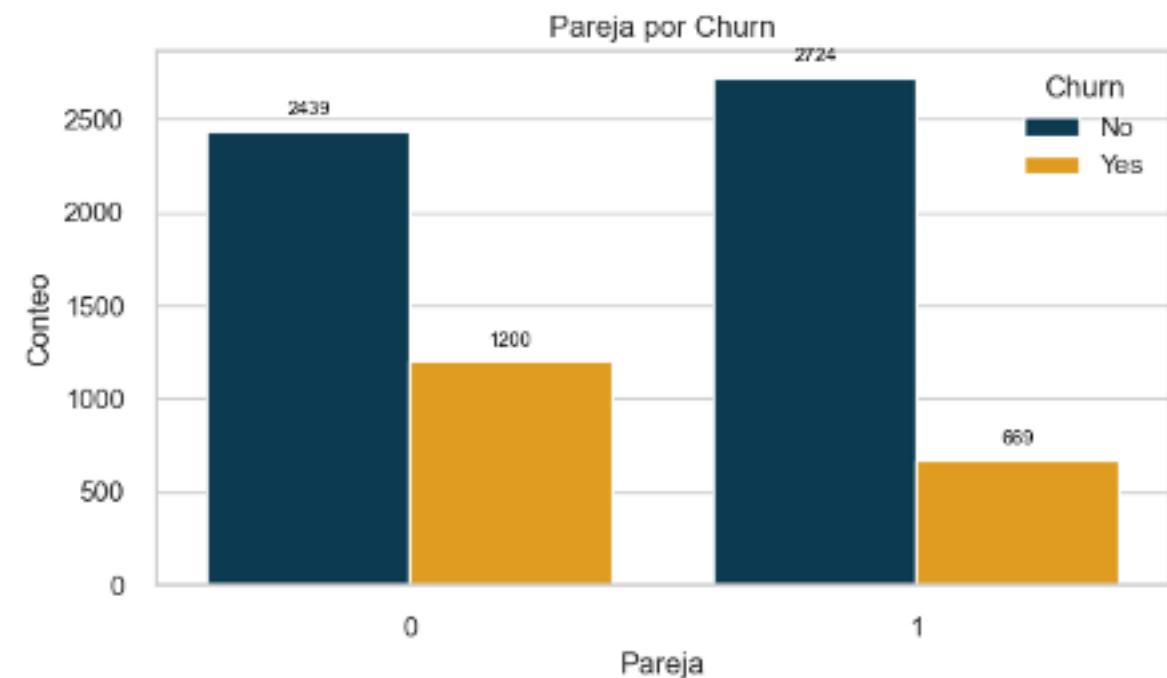
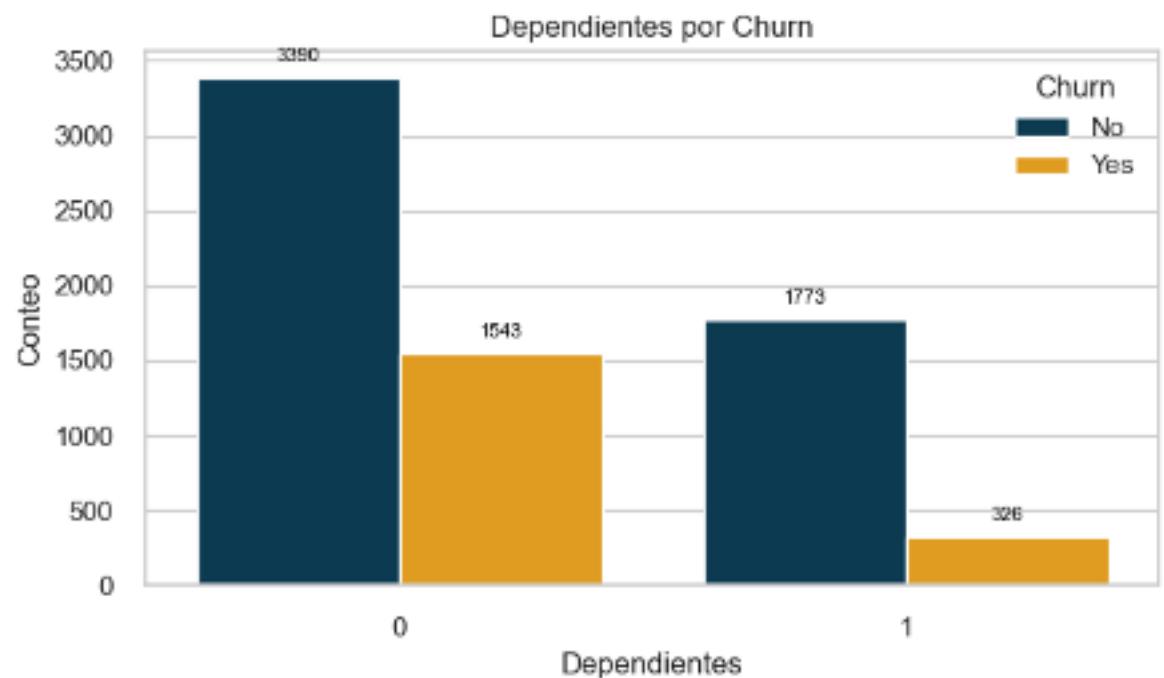
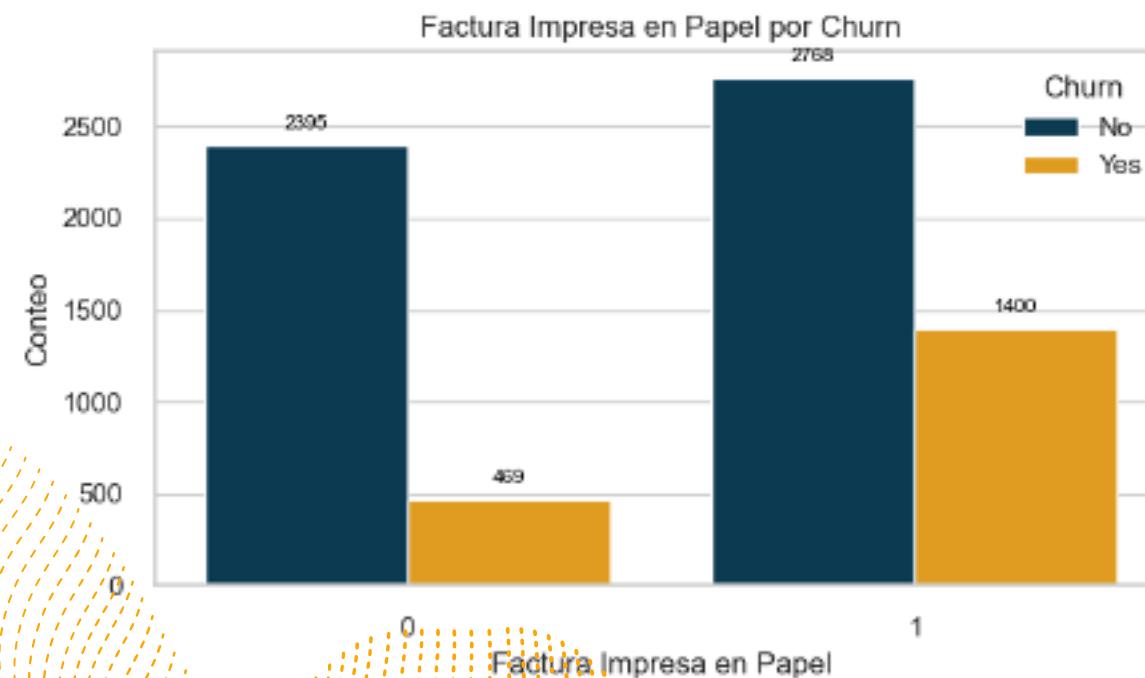
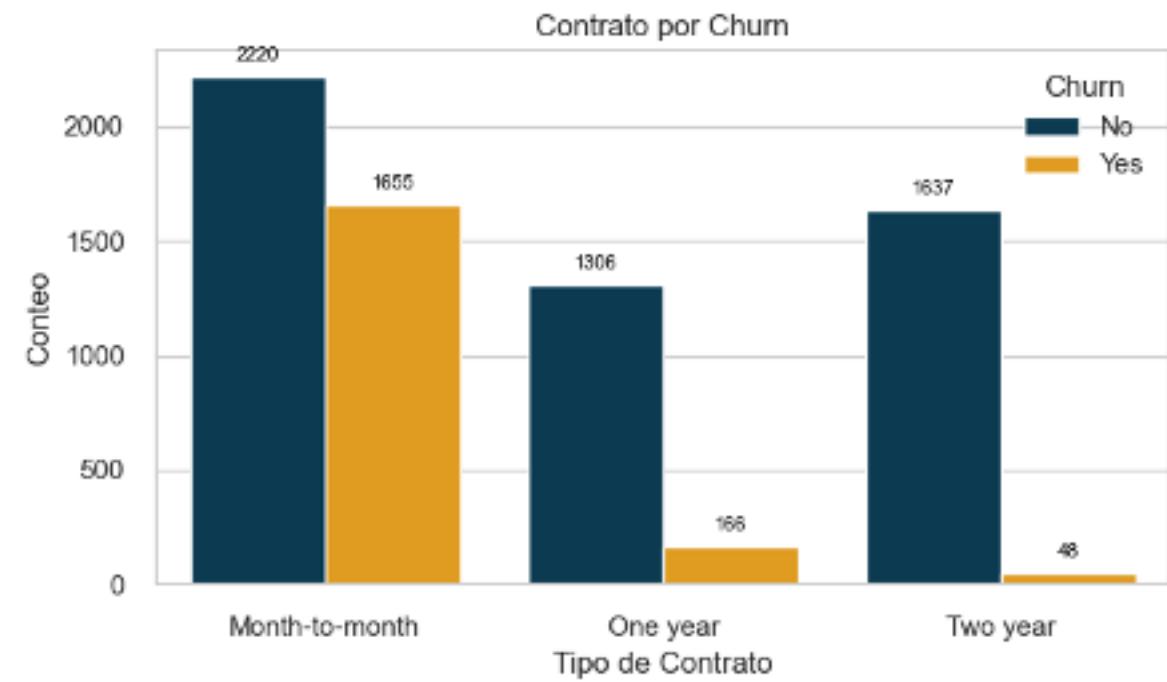
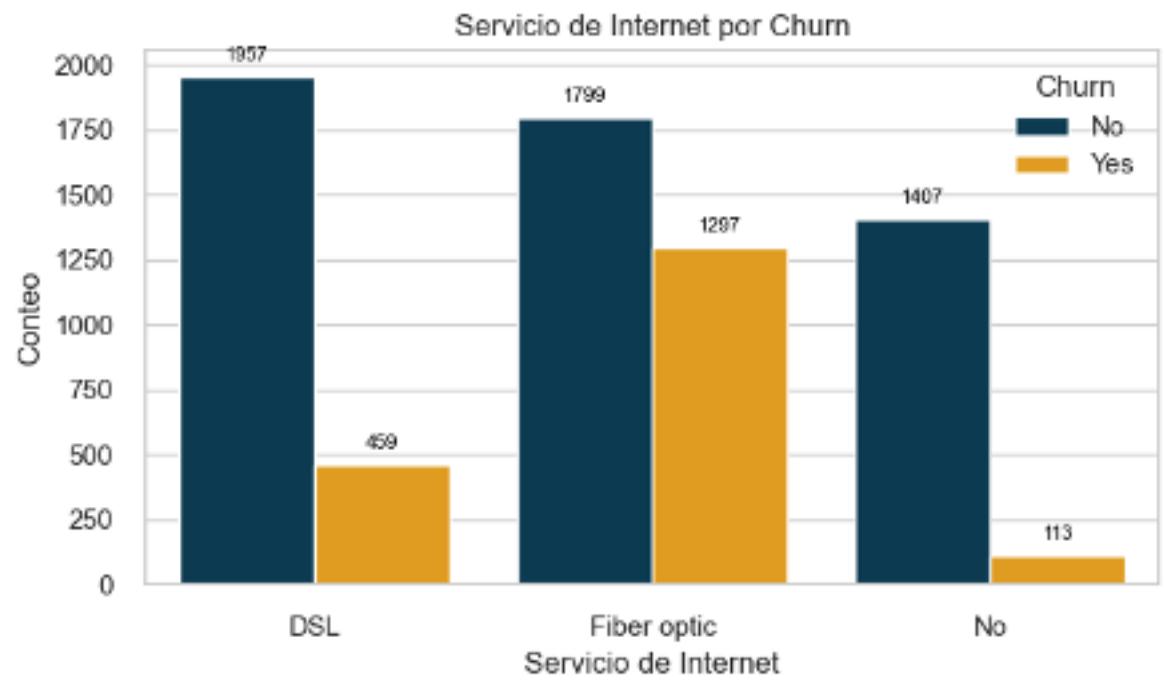
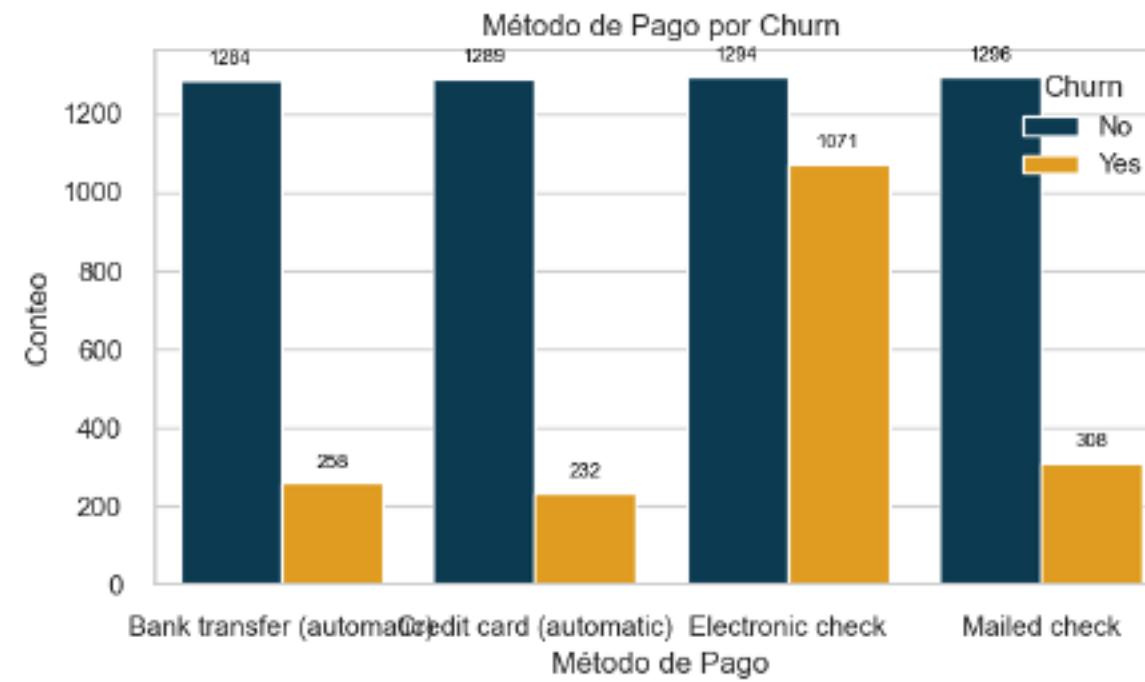
DISTRIBUCIÓN DE LA ROTACIÓN

En base a la cantidad real de clientes que se decidió trabajar, se muestra que los datos están desbalanceados, teniendo un mayor porcentaje de las personas que deciden quedarse dentro de la empresa.

Gráfico de Rotación

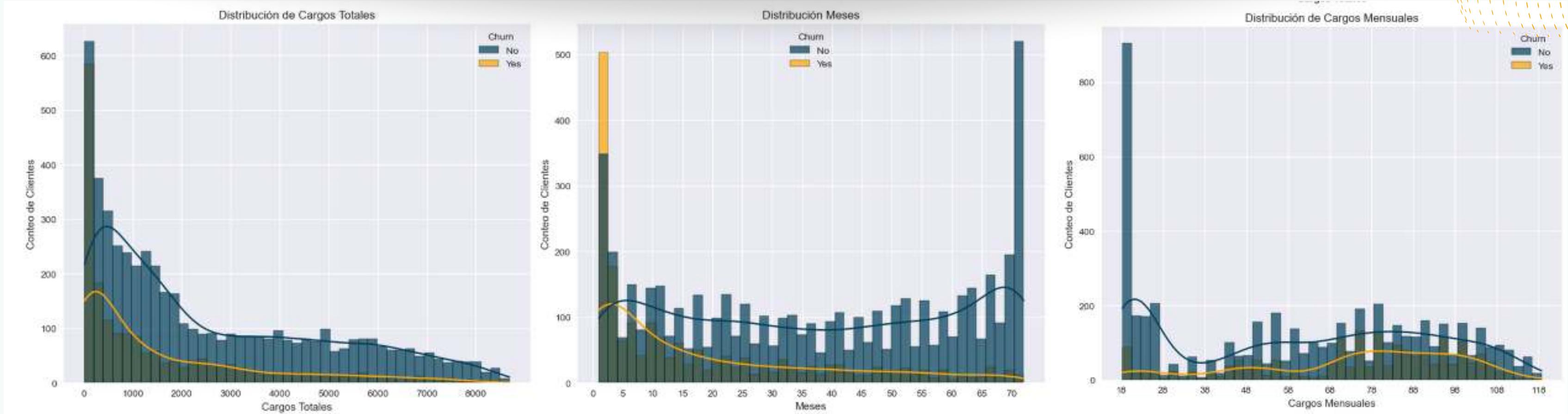


A través de estos graficos de barra se muestra el conteo de las variables más relevantes de aquellos clientes que se dieron o no la baja de la empresa



- Churn = Rotación
- El "0" representa NO y el "1" SI

Por medio de estos histogramas podemos evidenciar donde están acumulados mas los datos.



Los histogramas, diseñados para visualizar claramente la distribución de clientes que se dan de baja o permanecen en la empresa, revelan patrones significativos. En el primer gráfico a la izquierda, la densidad de personas con gastos totales entre 18.80 y 1,500 sugiere la presencia de un grupo importante. En el gráfico central, los picos en las esquinas indican una concentración mensual, posiblemente relacionada con contratos mensuales que podrían influir en las bajas. Esta tendencia parece disminuir con el tiempo. En el último gráfico a la derecha, se destaca un aumento en las bajas entre los rangos de 70 a 98, señalando un segmento crítico que merece especial atención, esto podría estar relacionado con la calidad de servicio que se contrata.

HIPOTESIS

En base a los gráficos anteriormente presentados surgen varias interrogantes:

Una de ellas es la alta tasa de clientes que se dan la baja cuando han contratado servicio de internet de Fibra óptica, en comparación con los que prefieren usar DSL; esto puede sugerir dos escenarios: el primero que el servicio de fibra óptica es muy costoso, o el servicio no es tan bueno en comparación con la competencia.

Otra de estas conclusiones es que quizá el contenido que se ofrece esta muy orientado a niños, lo que puede explicar la baja en personas que no tienen dependientes, y/o están solteros.

Se observa también según el análisis gráfico que es posible que la facturación NO impresa (Digital) tenga un impacto en la baja del servicio. Así mismo la opción de pago por cheque electrónico y el contrato del servicio mes a mes parecen ser opciones que son sensibles a la baja, se sugiere revisar las promociones que se ofrecen por tiempo de contratación.

K-MEANS

Algoritmo no supervisado de Machine Learning.

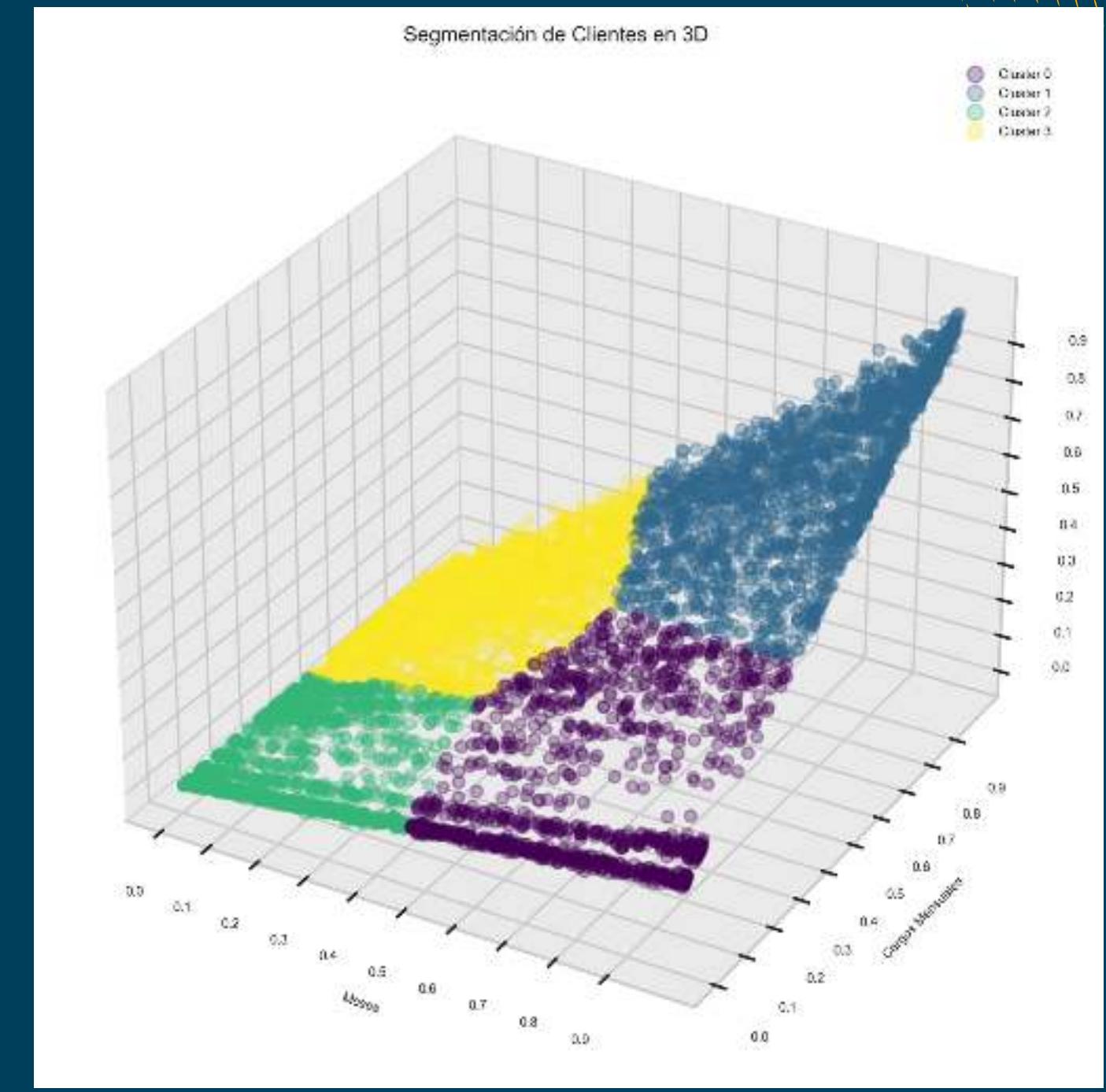
Por medio de este algoritmo se pretende agrupar o segmentar a los clientes en base a sus características para comprender mejor su comportamiento. Para ello se usaron las variables numéricas (meses de permanencia, cargos mensuales y cargos totales).

K-MEANS

Se realizaron varias pruebas para determinar el número óptimo de segmentos para agrupar a los clientes. Se tomó como referencia la Media Aritmética (Promedio) para agrupar los cliente como se muestra en el resumen a continuación:

Promedio por Segmento:

Grupo	Meses	Cargos Mensuales	Cargos Totales
0	53	34.55	1816.96
1	59	93.08	5529.65
2	10	31.88	303.59
3	15	80.82	1245.43



MODELOS DE CLASIFICACIÓN

Algoritmo supervisado de clasificación:

- Árbol de Decisión
- Bosque Aleatorio
- Regresión Logística
- XGBoost
- SVM

Estos algoritmos fueron usados para buscar el modelo mas robusto y de esta manera hacerlo útil a nuestro propósito, buscar el mejor modelo que prediga si un cliente se da o no la baja del servicio. Estos fueron optimizados y sujetos a varias técnicas.

Técnicas de Muestreo Usadas:

- Smote
- SmoTomekLink

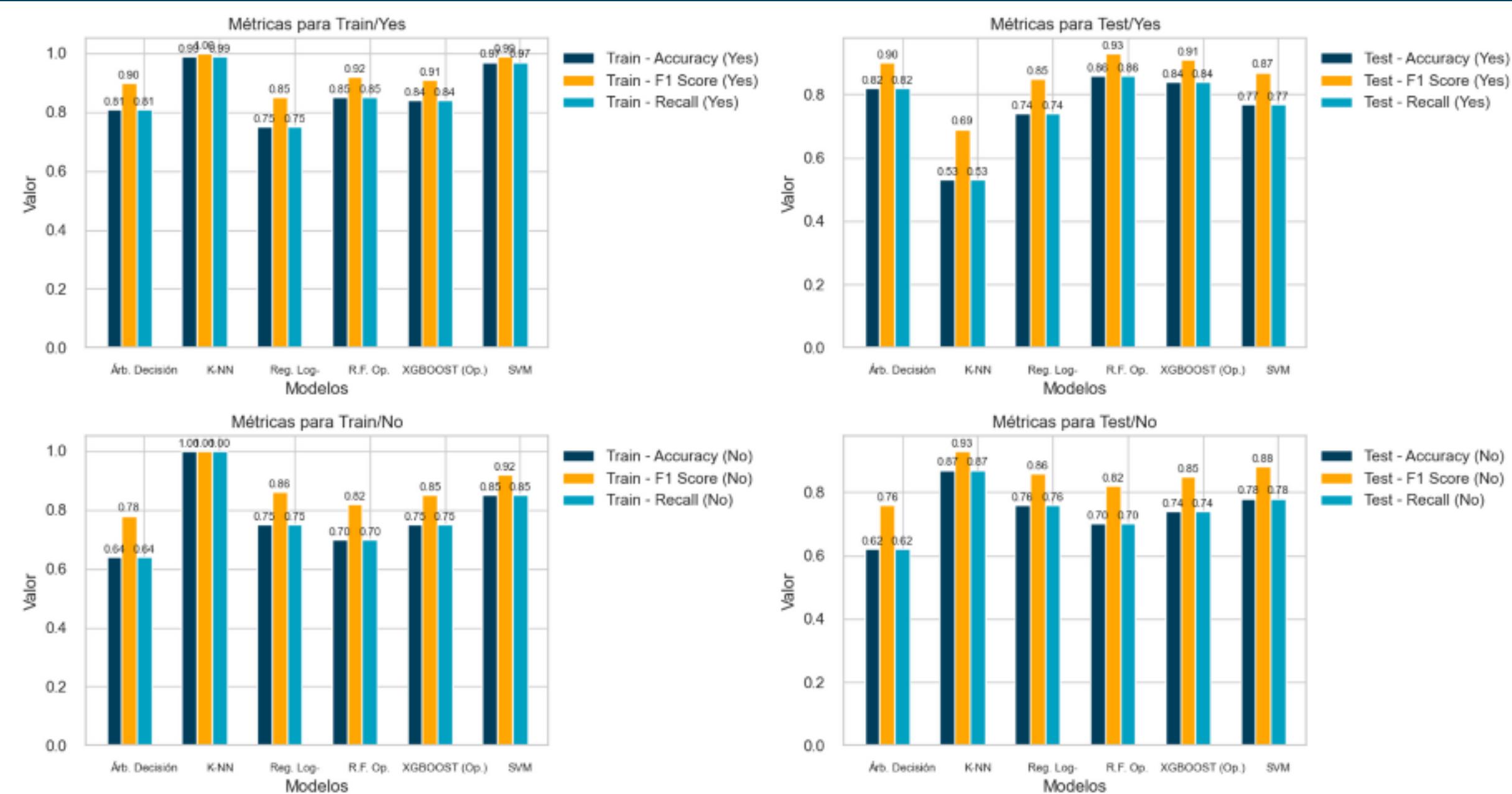
Técnicas de Optimización Usadas:

- GridSearchCV / K-Fold
- GridSearchCV / Halving
- GridSearchCV / Stratified K-Fold

El mayor desafío de entrenar estos modelos estuvo en entrenarlos con la mayor cantidad de personas que se dieron la baja siendo esto un reto ya que como se muestra en gráficos anteriores estos representan el 26% del Total.

El objetivo era obtener los mejores resultados en métricas de Recall y F1 Score

MODELOS DE CLASIFICACIÓN



Este gráfico muestra el performance de los modelos tanto para los datos que se usaron en Entrenamiento (Train) como en Prueba (Test)

En la imagen de la izquierda se pueden apreciar el rendimiento que tuvieron los modelos de clasificación siendo los que tienen el mejor rendimiento:

- SVM (Suport Vector Machine)
- XGBoost
- Random Forest Op.
-

Las Métricas a la que le dimos relevancia fueron al Recall y al F1 Score.

CONCLUSIÓN

Con el conjunto de datos explorados en este proyecto de DATA SCIENCE, se logró obtener valiosas perspectivas sobre las características del servicio y el comportamiento de los clientes, brindándonos una comprensión más profunda del funcionamiento interno de la empresa. En un mercado que evoluciona constantemente con la llegada de nuevos y atractivos servicios, especialmente en el ámbito de las transmisiones, donde cada vez más empresas se suman a la competencia, surge un desafío crucial para las compañías: retener y fidelizar a sus clientes. Por lo tanto, resulta fundamental mantenerse informado sobre los gustos y preferencias de la audiencia.

Si bien las técnicas de Machine Learning contribuirán significativamente a fortalecer nuestro poder de predicción, es importante destacar que a lo largo del proyecto, el análisis estadístico, especialmente en el ámbito descriptivo, nos permitió identificar patrones de comportamiento que podrían estar ejerciendo influencia en la decisión de los clientes de dar de baja el servicio. Este enfoque integral nos proporciona una visión completa y enriquecedora para abordar los desafíos presentes y futuros en el sector.

Durante el desarrollo de este proyecto, se examinó minuciosamente el comportamiento de 7.043 clientes a partir de una base de datos con 22 columnas. Utilizando el algoritmo de K-Means, los clientes fueron categorizados en 4 segmentos, considerando variables clave como el tiempo de permanencia, el costo mensual y el gasto total. Además, se implementaron cinco modelos de clasificación con el objetivo principal de determinar si un cliente optaría por quedarse o abandonar la empresa. Siendo los que mejores resultados obtuvieron en primer lugar el SVM, el XGBoost (Algoritmo de Boosting) y el Random Forest o Bosque Aleatorio (Algoritmo de Bagging). Estos modelos fueron sometidos a pruebas rigurosas, y se realizaron ajustes para optimizar su rendimiento, garantizando así la obtención de resultados de alta calidad. Sin embargo estos modelos se vieron afectados por el desbalanceo entre los clientes que se fueron que representan un 27% y los que se quedaron 73%. Alimentar estos modelos con más datos de clientes con la característica de haberse dado la baja o saber el motivo por el cual lo hicieron puede ser nuestro próximo desafío.