DESAFÍOS Y OPORTUNIDADES FRENTE A LA FUGA DE CLIENTES EN TELECOMUNICACIONES

DATA SCIENCE

Comisión 46235

CODERHOUSE

Febrero de 2024

Autor:

LUIS FERNANDO SANDOVAL RAMOS[1]

[1] Ingeniero Electrónico con amplia experiencia en monitoreo de salud estructural (SHM) utilizando sensores de fibra óptica (FBG, DAS, DTS, DSS)

Contenido

1. Descripción del caso de negocio

En un mercado altamente competitivo de servicios de telecomunicaciones, la retención de clientes es un factor crucial para el éxito empresarial. Con el objetivo de mitigar la pérdida de clientes y aumentar la lealtad de los mismos, se propone un enfoque basado en datos para identificar las causas de la deserción de clientes y desarrollar estrategias efectivas de retención.

Este enfoque implica el análisis de datos detallados sobre los clientes, sus servicios suscritos, información de cuentas y datos demográficos para comprender mejor los factores que influyen en la decisión de los clientes de abandonar el servicio. Posteriormente, se utilizarán estas ideas para desarrollar promociones y mejoras que ayuden a retener a la mayor cantidad posible de clientes.

2. Tabla de versionado

Es la primera versión del presente documento.

3. Objetivos del Modelo

3.1 Identificar las causas principales de la deserción de clientes en el sector de las telecomunicaciones.
3.2 Desarrollar estrategias efectivas de retención de clientes basadas en los resultados del análisis de datos.
3.3 Implementar promociones y mejoras específicas para abordar las causas identificadas y aumentar la retención de clientes.

4. Descripción del conjunto de Datos

Este conjunto de datos no tiene un origen específico y no está asociado con una empresa o institución en particular; sin embargo, conjuntos de datos similares pueden ser generados por empresas de telecomunicaciones o recopilados por investigadores académicos para estudiar el comportamiento de los clientes y desarrollar estrategias de retención. El presente data set se encuentra disponible en internet en el siguiente link

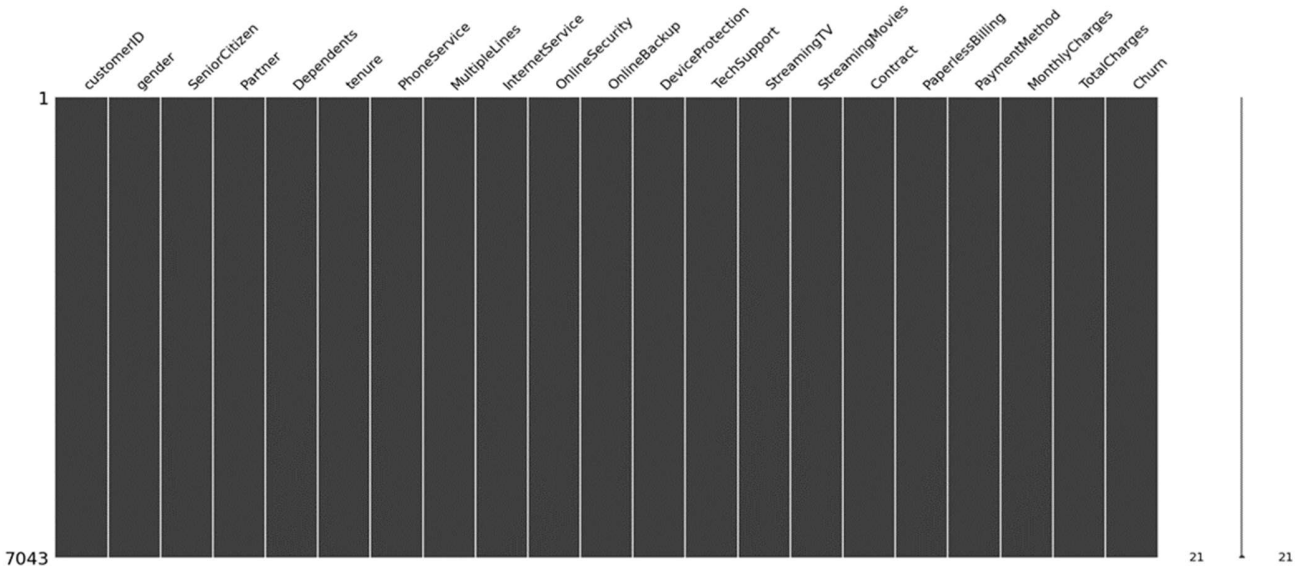
El data set proporciona una visión detallada de los clientes de una empresa de telecomunicaciones, incluyendo los servicios a los que están suscritos, información sobre su cuenta y su historial de pagos; así como información sobre si el cliente ha abandonado el servicio.

A continuación, se provee una breve descripción de las variables:

- `customerID`: Identificador único para cada cliente en la base de datos.

- `gender`: El género del cliente (puede ser "Male" o "Female").
- `SeniorCitizen`: Indica si el cliente es un ciudadano mayor (puede ser 0 o 1).
- `Partner`: Indica si el cliente tiene una pareja (puede ser "Yes" o "No").
- `Dependents`: Indica si el cliente tiene dependientes (puede ser "Yes" o "No").
- `tenure:` La cantidad de meses que el cliente ha sido cliente de la empresa de telecomunicaciones.
- `PhoneService`: Indica si el cliente tiene servicio telefónico (puede ser "Yes" o "No").
- `MultipleLines`: Indica si el cliente tiene múltiples líneas telefónicas (puede ser "Yes", "No" o "No phone service").
- `InternetService`: El tipo de servicio de internet al que el cliente está suscrito ("DSL", "Fiber optic" o "No").
- `OnlineSecurity`: Indica si el cliente tiene seguridad en línea (puede ser "Yes", "No" o "No internet service").
- `OnlineBackup`: Indica si el cliente tiene copia de seguridad en línea (puede ser "Yes", "No" o "No internet service").
- `DeviceProtection:` Indica si el cliente tiene protección de dispositivos (puede ser "Yes", "No" o "No internet service").
- `TechSupport:` Indica si el cliente tiene soporte técnico (puede ser "Yes", "No" o "No internet service").
- `StreamingTV`: Indica si el cliente tiene servicio de transmisión de TV (puede ser "Yes", "No" o "No internet service").
- `StreamingMovies`: Indica si el cliente tiene servicio de transmisión de películas (puede ser "Yes", "No" o "No internet service").
- `Contract`: El tipo de contrato del cliente ("Month-to-month", "One year", "Two year").
- `PaperlessBilling:` Indica si el cliente recibe facturación electrónica (puede ser "Yes" o "No").
- `PaymentMethod`: El método de pago utilizado por el cliente ("Electronic check", "Mailed check", "Bank transfer (automatic)", "Credit card (automatic)").
- `MonthlyCharges`: El monto de los cargos mensuales del cliente.
- `TotalCharges:` El monto total de los cargos acumulados por el cliente hasta la fecha.
- `Churn`: Indica si el cliente se ha retirado del servicio ("Yes" o "No").

4.1 Búsqueda de Datos faltantes



Aparentemente, el dataset no presenta datos faltantes. Pero para poder asegurar lo anterior, debemos ser más exhaustivos en la búsqueda de faltantes.

COLUMNAS	NULOS
customerID	0
gender	0
SeniorCitizen	0
Partner	0
Dependents	0
tenure	0
PhoneService	0
MultipleLines	0
InternetService	0
OnlineSecurity	0
OnlineBackup	0
DeviceProtectio	0
TechSupport	0
StreamingTV	0
StreamingMovie	0
Contract	0
PaperlessBilling	0
PaymentMethod	0
MonthlyCharges	0
TotalCharges	11
Churn	0

En una revisión más detallada, se observan datos faltantes en “TotalCharges”

En principio, se realizó un análisis de todas las variables, se detectaron 11 valores nulos en la columna “`TotalCharges`”; no obstante, Se puede observar que la columna "`tenure`" (tenencia) es 0 para estas entradas, aunque la columna `MonthlyCharges` no esté vacía. Pudimos demostrar que no faltan más valores en la columna "`tenure`"; por lo tanto, para resolver el problema de los valores faltantes, se decidió completarla con la media de los valores de esa columna.

MonthlyCharges	TotalCharges	Dependents	tenure	PhoneService
52.55	NaN	Yes	0	No
20.25	NaN	Yes	0	Yes
80.85	NaN	Yes	0	Yes
25.75	NaN	Yes	0	Yes
56.05	NaN	Yes	0	No
19.85	NaN	Yes	0	Yes
25.35	NaN	Yes	0	Yes
20.00	NaN	Yes	0	Yes
19.70	NaN	Yes	0	Yes
73.35	NaN	Yes	0	Yes
61.90	NaN	Yes	0	Yes

Comentario: Las filas con valores nulos en 'TotalCharges' parecen estar asociadas con una tenencia de 0 meses. Esto puede deberse a clientes nuevos que aún no han sido facturados

4.2 Transformación de los Datos

Se procede a transformar algunos datos del dataset, para facilitar la comprensión de los mismos

customerID	gender	SeniorCitizen	Partner
4472-LVYGI	Female	0	Yes
3115-CZMZD	Male	0	No
5709-LVOEQ	Female	0	Yes
4367-NUYAO	Male	0	Yes
1371-DWPAZ	Female	0	Yes

customerID	gender	SeniorCitizen
7590-VHVEG	Female	No
5575-GNVDE	Male	No
3668-QPYBK	Male	No
7795-CFOCW	Male	No
9237-HQITU	Female	No

Aquí se procedió a transformar los valores de la columna "SeniorCitizen" que originalmente contenía valores numéricos, donde 0 representa "No" y 1 representa "Yes". La instrucción

realiza una transformación de estos valores numéricos en valores de cadena (strings) "No" y "Yes" usando un mapeo.

Es posible contar la combinación de dos o más columnas para comprender mejor la interacción entre estas. Los datos muestran una preferencia de los usuarios por los contratos “Mes a Mes” para los servicios de internet por Fibra Óptica y DSL.

```
Conteo de combinaciones de 'InternetService' y 'Contract':
InternetService  Contract
Fiber optic      Month-to-month    2128
DSL              Month-to-month    1223
No               Two year           633
DSL              Two year           623
One year           570
Fiber optic      One year           539
No               Month-to-month     524
Fiber optic      Two year           429
No               One year           363
```

Se pueden analizar las columnas numéricas, extraer las estadísticas descriptivas y obtener información relevante de los datos.

Estadísticas descriptivas para columnas numéricas:

	tenure	MonthlyCharges	TotalCharges
count	7032.000000	7032.000000	7032.000000
mean	32.421786	64.798208	2283.300441
std	24.545260	30.085974	2266.771362
min	1.000000	18.250000	18.800000
25%	9.000000	35.587500	401.450000
50%	29.000000	70.350000	1397.475000
75%	55.000000	89.862500	3794.737500
max	72.000000	118.750000	8684.800000

Estas son las estadísticas descriptivas para las columnas numéricas seleccionadas:

tenure (Tenencia):

- count: Hay 7032 valores no nulos en esta columna.
- mean (Media): La media de la tenencia de los clientes es aproximadamente 32.42 meses.
- std (Desviación estándar): La desviación estándar de la tenencia es aproximadamente 24.55 meses, lo que indica la dispersión de los datos.
- min (Mínimo): El valor mínimo de tenencia es de 1 mes.
- 25% (Cuartil 1): El 25% de los clientes tienen una tenencia igual o menor a 9 meses.
- 50% (Mediana): La mediana de la tenencia es de aproximadamente 29 meses, lo que significa que el 50% de los clientes tienen una tenencia igual o menor a este valor.
- 75% (Cuartil 3): El 75% de los clientes tienen una tenencia igual o menor a 55 meses.
- max (Máximo): El valor máximo de tenencia es de 72 meses.

MonthlyCharges (Cargos Mensuales):

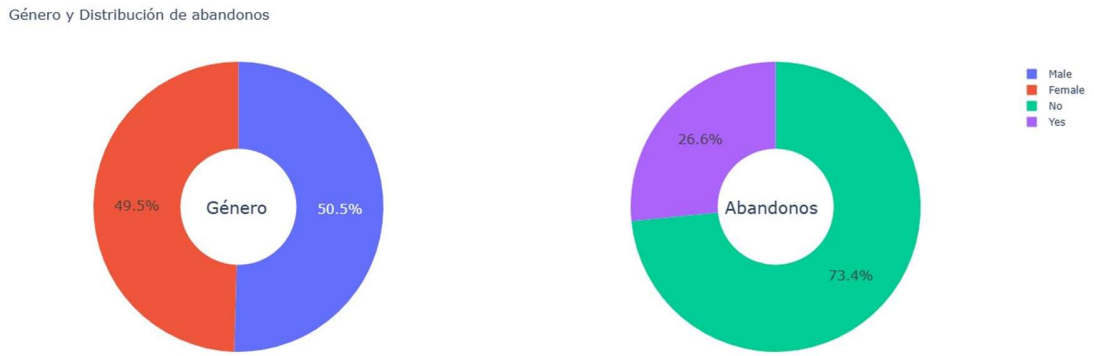
Las estadísticas descriptivas para esta columna son similares a las de la columna 'tenure', pero se refieren a los cargos mensuales en lugar de la tenencia.

TotalCharges (Cargos Totales):

Las estadísticas descriptivas para esta columna son similares a las de las columnas anteriores, pero se refieren a los cargos totales acumulados en lugar de la tenencia o los cargos mensuales.

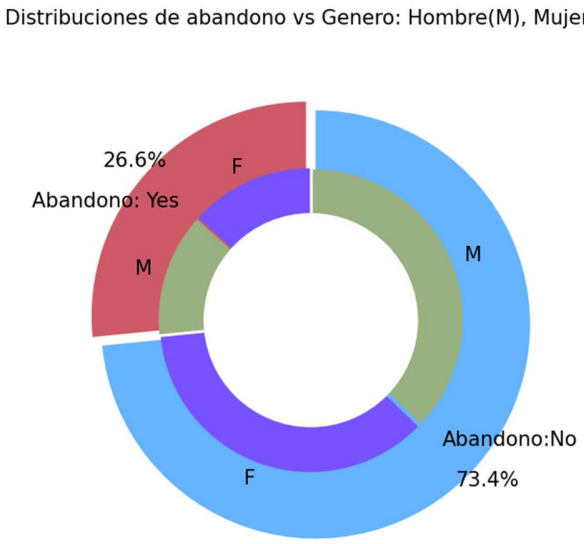
Estas estadísticas proporcionan una visión general de la distribución y la variabilidad de los datos en estas columnas numéricas, lo que puede ser útil para comprender mejor el comportamiento de los clientes y tomar decisiones informadas basadas en estos datos.

4.3 Revisión de la pérdida de clientes

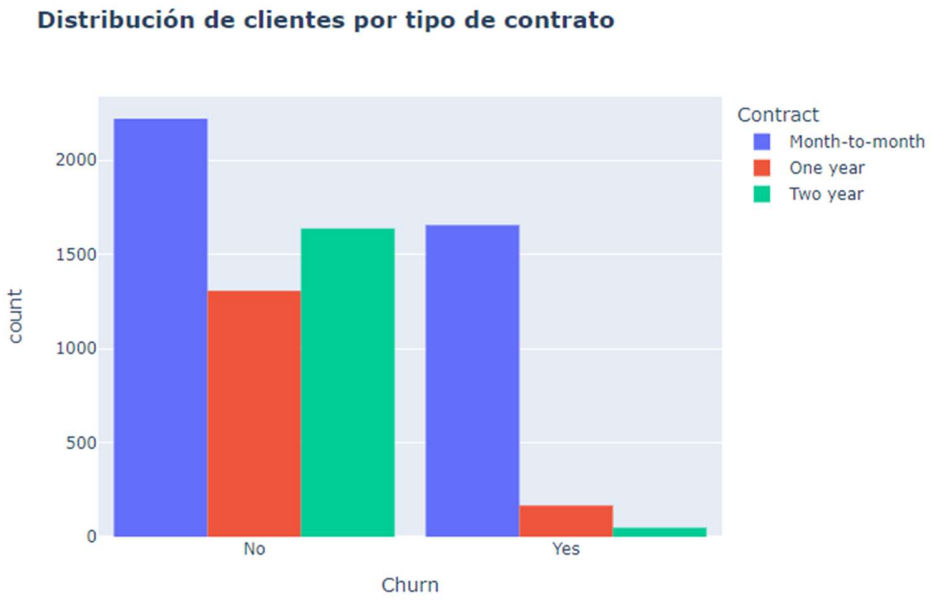


- El 26.6 % de los clientes se cambian a otro servicio
- Entre los clientes el 49.5 % son Mujeres y el 50.5 % son Hombres

A continuación, queremos ver de esos porcentajes de abandono cuantos son Hombres y cuantos Mujeres:



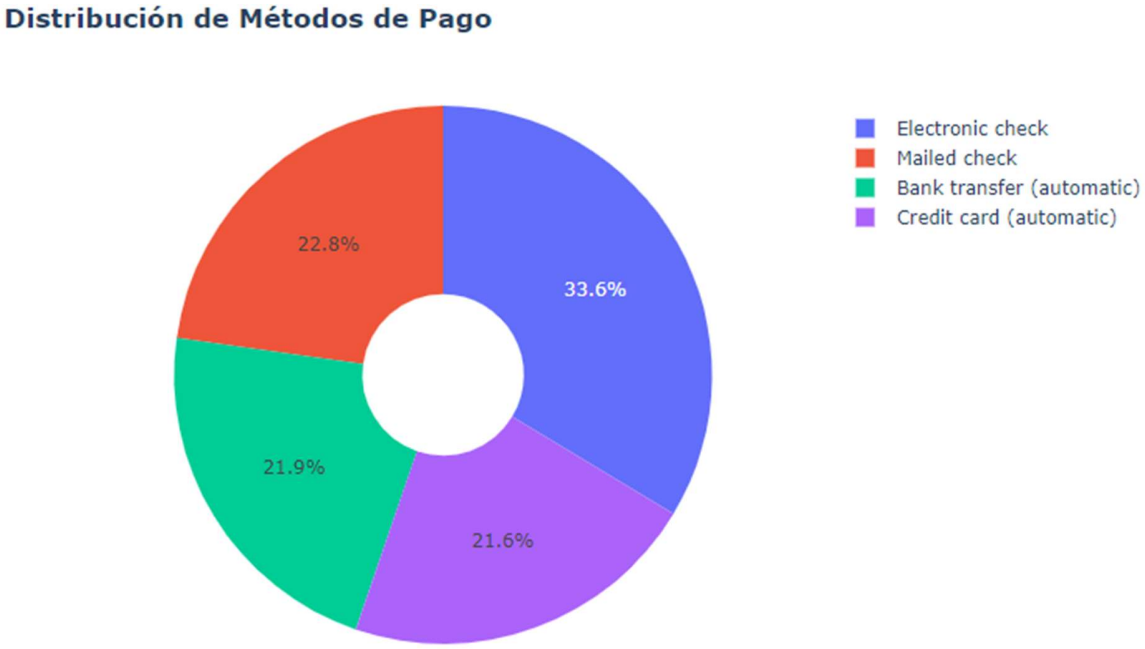
Existe una diferencia insignificante en el porcentaje/recuento de clientes que cambiaron de proveedor de servicios por género. Ambos géneros se comportaron de manera similar cuando se trata de migrar a otro proveedor de servicios. De ese 26.6% de los clientes que se cambian o abandonan, no hay una tendencia clara de este comportamiento por género del cliente.



Alrededor del 75 % de los clientes con contrato de mes a mes optaron por abandonar el servicio en comparación con el 13% de los clientes con contrato de un año y el 3% con contrato de dos años.

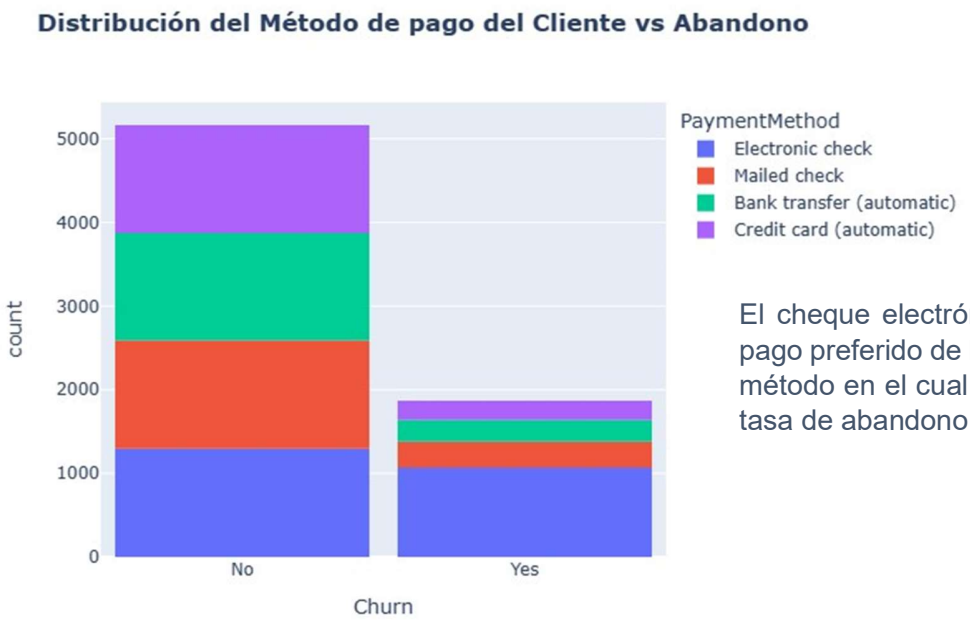
Para efectos de reducir la pérdida de clientes de este segmento de mercado, se recomiendan las siguientes estrategias:

- Incentivos para contratos a largo plazo: ofrecer incentivos atractivos, como descuentos especiales o beneficios adicionales, para los clientes que opten por contratos de un año o dos años. Esto puede motivar a los clientes a comprometerse a largo plazo y reducir la tasa de abandono.
- Programas de fidelización: implementar programas de fidelización diseñados específicamente para clientes con contratos mensuales. Esto podría incluir recompensas por permanecer en el servicio durante períodos más largos o beneficios exclusivos para clientes leales.
- Mejora de la experiencia del cliente: asegurarse de que los clientes con contratos mensuales estén satisfechos con el servicio ofrecido. Esto puede implicar mejorar la calidad del servicio al cliente, resolver problemas de manera rápida y eficiente y garantizar una experiencia general positiva.
- Comunicación proactiva: mantener una comunicación proactiva con los clientes con contratos mensuales. Envía recordatorios sobre los beneficios del servicio, ofrece asistencia personalizada y brinda información sobre promociones y ofertas especiales para fomentar la retención.
- Flexibilidad en los contratos mensuales: ofrecer opciones flexibles dentro de los contratos mensuales, como la posibilidad de cambiar fácilmente a contratos a largo plazo con beneficios adicionales o la opción de pausar temporalmente el servicio en lugar de cancelarlo por completo.



La relación entre los métodos de pago se encuentra ligeramente equilibrada, con una mayor preferencia por el “Electronic check” con un 33.6%.

A continuación, revisaremos la relación entre el método de pago y el abandono:



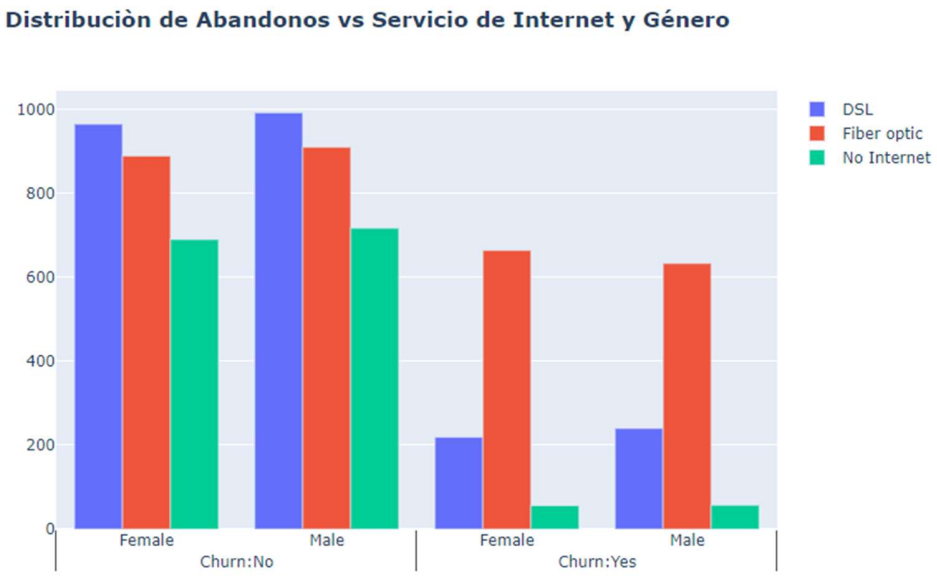
El cheque electrónico es el método de pago preferido de los clientes, pero es el método en el cual se presenta la mayor tasa de abandono del servicio.

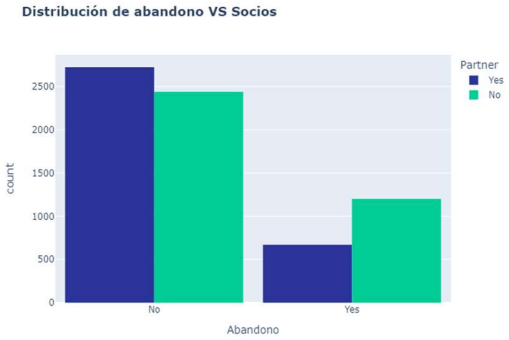
- Los principales clientes que se mudaron tenían cheque electrónico como método de pago.
- Los clientes que optaron por la transferencia automática con tarjeta de crédito o la transferencia automática bancaria y el cheque enviado por correo como método de pago tenían menos probabilidades de mudarse.

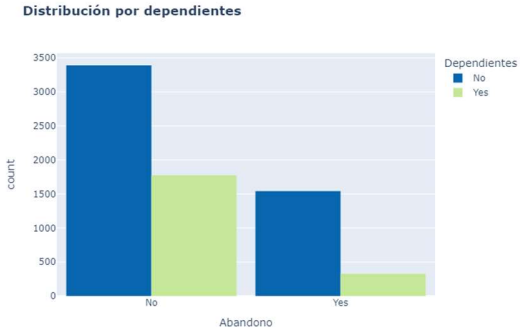
Siguiendo el mismo razonamiento respecto al tipo de contrato, se debe incentivar la permanencia en el servicio si se realizan pagos con cheque electrónico.

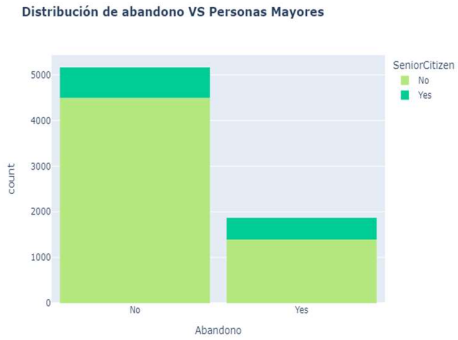
En el siguiente gráfico se muestra la distribución de abandono por los servicios de internet más demandados y por el género. Tal y como se muestra en las gráficas anteriores, la relación entre el abandono y el género no es relevante.

- Una gran cantidad de clientes eligen el servicio de Fibra óptica y también es evidente que los clientes que utilizan Fibra óptica tienen una alta tasa de abandono, esto podría sugerir una insatisfacción con este tipo de servicio de Internet.
- Los clientes que tienen servicio DSL son mayoritarios y tienen una menor tasa de abandono en comparación con el servicio de fibra óptica.



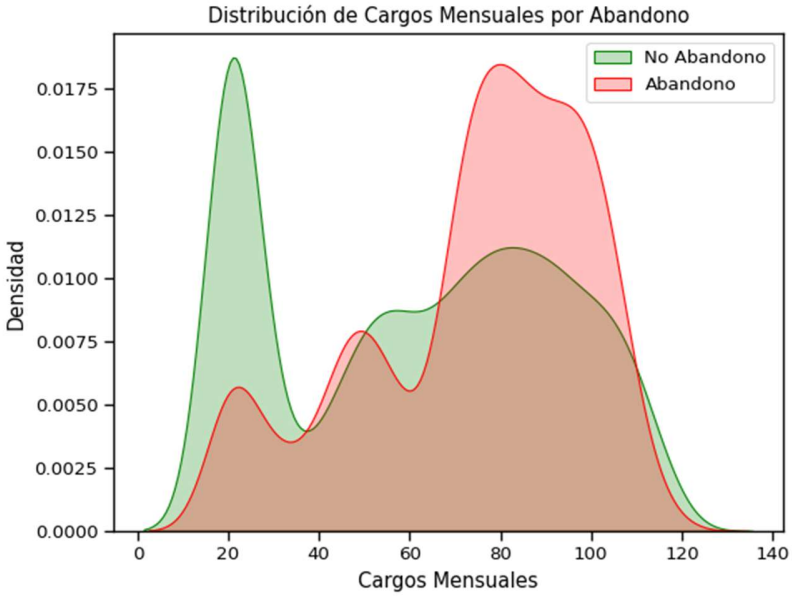




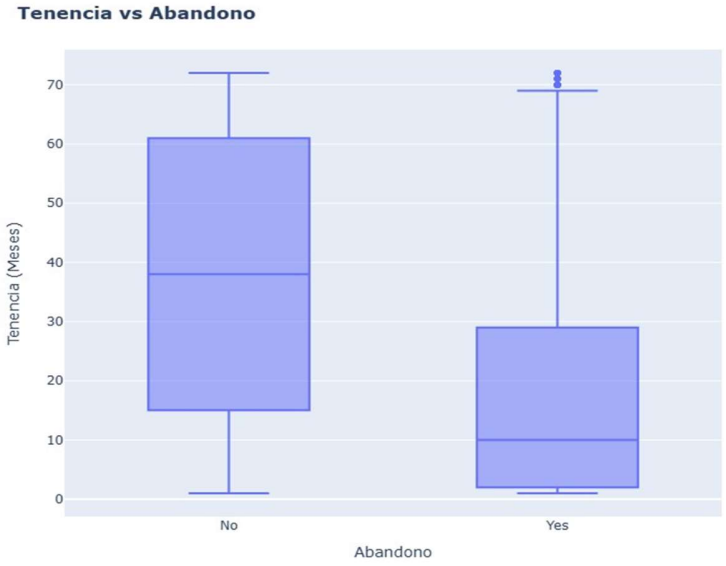


- Los abonados que tienen Socios abandonan menos el servicio
- Los abonados con Dependientes, abandonan menos el servicio
- Los Jubilados o personas mayores abandonan menos el servicio

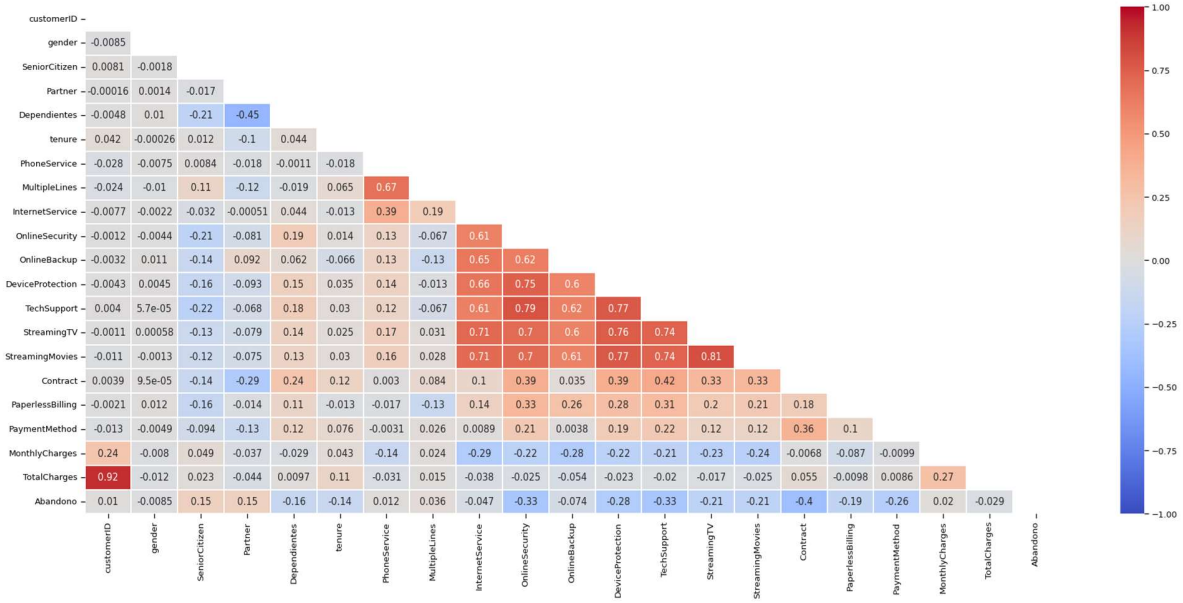
Los datos suministrados en el dataset muestran una mayor claridad en cuanto a la relación de los servicios contratados y la tasa de abandonos. En las siguientes gráficas se nota una alta tasa de abandono en los clientes que no cuentan con el servicio de “Online Security” y los que no adquieren el “Tech Support”; es posible que la percepción de calidad en comparación con el soporte y la seguridad en línea sean aspectos para analizar en la fidelización de los futuros clientes.



Los clientes que tienen cargos mensuales altos tienen una tasa de abandono mayor, los que tienen valores entre 65 y 110 USD están abandonando de forma recurrente.



La mayor tasa de abandono del servicio se presenta entre el mes 1 y el mes 29. Esto corresponde a aproximadamente el 50% de los usuarios que abandonan el servicio



Hay una relación modesta entre el abandono de clientes y la edad o si el servicio es compartido con parejas o dependientes o si el usuario es jubilado.

5. EDA – Exploratory Data Analysis

El Análisis Exploratorio de Datos (EDA, por sus siglas en inglés, Exploratory Data Analysis) es un enfoque fundamental en la ciencia de datos que implica explorar y comprender los datos antes de aplicar métodos de modelado o inferencia más avanzados. Consiste en un conjunto de técnicas y métodos para resumir, visualizar y analizar los datos con el objetivo de obtener información útil y detectar patrones, tendencias, anomalías y relaciones entre las variables.

Algunos de los aspectos principales del EDA incluyen:

- **Resumen estadístico inicial:** calcular estadísticas descriptivas básicas, como la media, la mediana, la desviación estándar, los valores mínimos y máximos, y los cuartiles, para comprender la distribución y la variabilidad de los datos.

- **Visualización de datos:** utilizar gráficos y visualizaciones para representar los datos de manera efectiva. Esto puede incluir histogramas, diagramas de dispersión, diagramas de caja, gráficos de barras, gráficos de líneas, mapas de calor y otros tipos de gráficos que ayuden a revelar patrones o relaciones en los datos.

- **Identificación de valores atípicos (outliers):** detectar y manejar valores atípicos que pueden ser errores de entrada, datos mal medidos o representar casos significativos que deben ser tenidos en cuenta en el análisis.

- **Exploración de relaciones entre variables:** analizar la correlación entre variables, identificar dependencias y relaciones lineales o no lineales, y explorar cómo las variables afectan a otras.

- **Manipulación y transformación de datos:** limpiar y preprocesar los datos para corregir errores, tratar con valores faltantes, normalizar variables y realizar otras transformaciones necesarias para preparar los datos para el modelado.

- **Extracción de características (Feature Engineering):** identificar y crear nuevas características que puedan ser útiles para el modelado, como la agregación de variables, la creación de variables dummy, la codificación de variables categóricas, etc.

5.1 Ingeniería de atributos

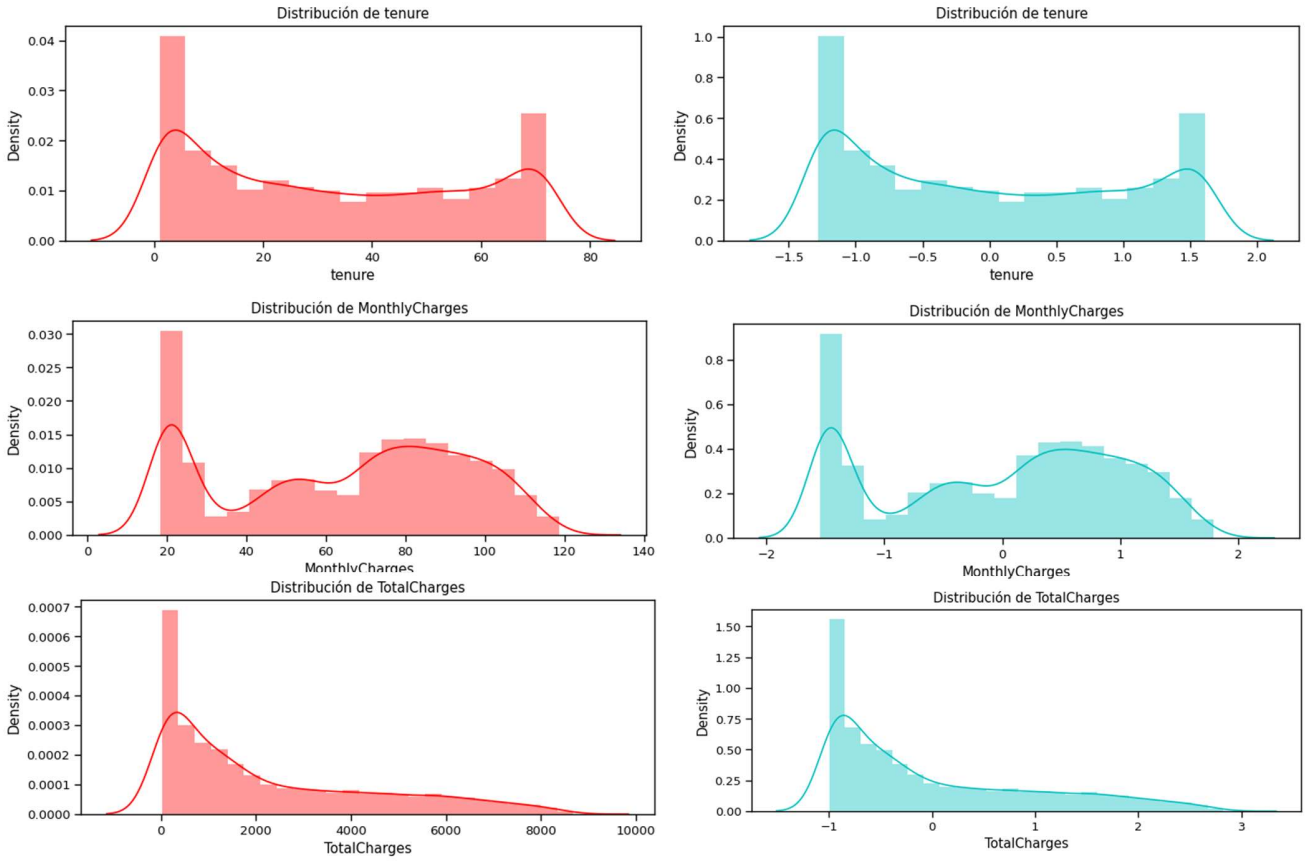
A continuación, se tabula la correlación entre las características de la Data y la variable objetivo “Abandono” o “Churn”

Abandono	1.000.000
MonthlyCharges	192.858
PaperlessBilling	191.454
SeniorCitizen	150.541
Age	115.458
PaymentMethod	107.852
MultipleLines	38.043
City	37.248
Longitude	24.156
PhoneService	11.691
gender	-8.545
Customer ID	-17.858
StreamingTV	-36.303
StreamingMovies	-38.802
Latitude	-41.761
InternetService	-47.097
Partner	-149.982

Abandono	1.000.000
Dependientes	-163.128
DeviceProtection	-177.883
OnlineBackup	-195.290
TotalCharges	-199.484
TechSupport	-282.232
OnlineSecurity	-289.050
tenure	-354.049
Contract	-396.150

X = Data3.drop(columns=['Abandono']): Crea un nuevo DataFrame X eliminando la columna 'Churn' del DataFrame original Data3. En otras palabras, X contendrá todas las columnas de Data3 excepto la columna 'Churn'. Esto se hace para preparar las características (variables independientes) que se utilizarán para entrenar un modelo.

y = Data3['Abandono'].values: Crea una serie y que contiene los valores de la columna 'Churn' del DataFrame original Data3. Esta serie representa la variable objetivo o etiquetas que se están tratando de predecir con el modelo. .values se utiliza para extraer los valores como un arreglo NumPy.



Estandarización de características: Se utiliza StandardScaler() para estandarizar las características numéricas del DataFrame Data3. La estandarización es un paso común en el preprocesamiento de datos para asegurar que todas las características tengan la misma escala y tengan una media de 0 y una desviación estándar de 1.

5.2 Estandarización de características:

Se utiliza StandardScaler() para estandarizar las características numéricas del DataFrame Data3. La estandarización es un paso común en el preprocesamiento de datos para asegurar que todas las características tengan la misma escala y tengan una media de 0 y una desviación estándar de 1.

Creación del DataFrame estandarizado: Los datos estandarizados se almacenan en un nuevo DataFrame llamado Data3_std.

Este proceso asegura que las características numéricas en ambos conjuntos de datos estén en la misma escala, lo cual es importante para muchos algoritmos de aprendizaje automático. La normalización estándar ayuda a que las características tengan una media de cero y una desviación estándar de uno.

Selección de columnas para codificación: Se definen dos listas: cat_cols_ohe para las columnas que necesitan codificación one-hot y cat_cols_le para las columnas que necesitan codificación de etiquetas. La lista cat_cols_le se crea restando las columnas numéricas (num_cols) y las columnas one-hot (cat_cols_ohe) del conjunto total de columnas en el conjunto de entrenamiento (X_train).

Instanciación de StandardScaler: Se crea una instancia del objeto StandardScaler(), que se utilizará para estandarizar las características numéricas.

Estandarización de características numéricas: Se estandarizan las características numéricas en los conjuntos de entrenamiento (X_train) y prueba (X_test). Esto se logra utilizando el método fit_transform para el conjunto de entrenamiento y transform para el conjunto de prueba. La estandarización se aplica solo a las características numéricas (num_cols) utilizando el objeto scaler que se ha ajustado al conjunto de entrenamiento. Esto garantiza que la misma transformación se aplique a ambos conjuntos de datos.

5.3 Modelos de clasificación

- KNN: K-Nearest Neighbors (KNN) es un algoritmo simple y ampliamente utilizado para clasificación y regresión. Para nuestro análisis utilizaremos clasificación

Informe de Clasificación KNN

	precision	recall	f1-score	support
0	0.74	0.96	0.83	1549
1	0.31	0.05	0.09	561
accuracy			0.72	2110
macro avg	0.52	0.50	0.46	2110
weighted avg	0.62	0.72	0.63	2110

Precision (Precisión): para la clase 0 (clientes que no abandonaron el servicio), la precisión es del 74%. Esto significa que el 74% de las predicciones de clientes que no abandonaron el servicio fueron correctas. Para la clase 1 (clientes que abandonaron el servicio), la precisión es del 31%. Esto significa que solo el 31% de las predicciones de clientes que abandonaron el servicio fueron correctas.

Recall (Recuperación o Sensibilidad): para la clase 0, el recall es del 96%. Esto significa que el 96% de los clientes que realmente no abandonaron el servicio fueron correctamente identificados por el modelo. Para la clase 1, el recall es del 5%. Esto significa que solo el 5% de los clientes que realmente abandonaron el servicio fueron correctamente identificados por el modelo.

F1-score: el F1-score combina precision y recall en una sola métrica. Es útil cuando hay un desequilibrio entre las clases. Para la clase 0, el F1-score es del 0.83, mientras que para la clase 1 es del 0.09.

Soporte (Support): el soporte indica el número de ocurrencias de cada clase en el conjunto de prueba. Hay 1549 instancias de la clase 0 y 561 instancias de la clase 1.

Exactitud (Accuracy): la exactitud del modelo es del 72%. Esto indica la proporción de predicciones correctas sobre el total de predicciones realizadas en el conjunto de prueba.

Macro average (Promedio macro) y Weighted average (Promedio ponderado):

El macro average calcula las métricas promedio sin considerar el desequilibrio de clases. El weighted average calcula las métricas promedio considerando el desequilibrio de clases ponderando las métricas de cada clase por su soporte.

Revisión del modelo:

El modelo tiene un buen desempeño en la predicción de la clase mayoritaria (no abandono del servicio), con alta precisión y recall. Sin embargo, el modelo tiene un desempeño deficiente en la predicción de la clase minoritaria (abandono del servicio), con baja precisión, recall y F1-score. La exactitud del modelo del 72% indica que, en general, el modelo clasifica correctamente el 72% de los casos en el conjunto de prueba. Dado el desequilibrio en el soporte de las clases, el promedio macro y el promedio ponderado de las métricas son diferentes, lo que refleja la necesidad de considerar el desequilibrio de clases al evaluar el rendimiento del modelo.

- SVC: el modelo Support Vector Classification (SVC) es un clasificador de máquinas de vectores de soporte utilizado para problemas de clasificación.

Informe de Clasificación SVC

	precision	recall	f1-score	support
0	0.73	1.00	0.85	1549
1	1.00	0.00	0.00	561
accuracy			0.73	2110
macro avg	0.87	0.50	0.42	2110
weighted avg	0.80	0.73	0.62	2110

El informe de clasificación proporciona una evaluación detallada del rendimiento del modelo SVC en el conjunto de prueba:

Precision: la precisión para la clase 0 (no abandono) es del 73%, lo que indica que el modelo clasifica correctamente el 73% de los casos predichos como no abandono que son verdaderos. Sin embargo, para la clase 1 (abandono), la precisión es del 100%, lo que sugiere que el modelo clasifica correctamente todos los casos predichos como abandono que son verdaderos.

Recall: el recall para la clase 0 es del 100%, lo que significa que el modelo identifica correctamente todos los casos de no abandono en el conjunto de prueba. Para la clase 1, el recall es del 0%, lo que indica que el modelo no identifica correctamente ningún caso de abandono en el conjunto de prueba.

F1-score: el puntaje F1 para la clase 0 es del 85%, que es una medida balanceada entre precisión y recall. Sin embargo, para la clase 1, el puntaje F1 es del 0%, lo que sugiere un rendimiento muy pobre del modelo en la predicción de abandono.

Accuracy: la precisión general del modelo en el conjunto de prueba es del 73%, lo que indica que el 73% de las predicciones del modelo son correctas.

Macro avg: El promedio macro de precision, recall y F1-score es del 87%, 50% y 42% respectivamente. Esto proporciona una medida agregada del rendimiento del modelo en todas las clases, tratando cada clase de manera uniforme.

Weighted avg: el promedio ponderado de precision, recall y F1-score es del 80%, 73% y 62% respectivamente. Esto tiene en cuenta el desequilibrio de clases en el conjunto de datos, dándole más peso a las clases con mayor soporte (número de instancias).

El modelo tiene un buen rendimiento en la predicción de casos de no abandono, pero tiene un rendimiento muy deficiente en la predicción de casos de abandono, lo que sugiere que el modelo necesita ser mejorado, posiblemente mediante ajuste de hiperparámetros o selección de características.

- Random Forest: el modelo Random Forest es un algoritmo de aprendizaje automático que pertenece a la categoría de métodos de conjunto (ensemble methods). Un Random Forest construye múltiples árboles de decisión durante el

entrenamiento y fusiona sus predicciones para obtener un resultado más robusto y generalizable.

Informe de Clasificación Random Forest

	precision	recall	f1-score	support
0	0.83	0.93	0.88	1549
1	0.72	0.49	0.58	561
accuracy			0.81	2110
macro avg	0.78	0.71	0.73	2110
weighted avg	0.80	0.81	0.80	2110

Este es un informe de clasificación que evalúa el rendimiento de un modelo Random Forest en un conjunto de datos de prueba. Aquí hay una explicación de cada métrica en el informe:

Precisión (Precision): la precisión mide la proporción de instancias predichas como positivas que fueron clasificadas correctamente.

Para la clase 0 (no abandono), la precisión es del 83%, lo que significa que el 83% de las instancias predichas como no abandono realmente pertenecen a la clase no abandono.

Para la clase 1 (abandono), la precisión es del 72%, lo que indica que el 72% de las instancias predichas como abandono realmente pertenecen a la clase abandono.

Recuperación (Recall): la recuperación, también conocida como sensibilidad o tasa positiva real, mide la proporción de instancias positivas que fueron clasificadas correctamente.

Para la clase 0, la recuperación es del 93%, lo que significa que el 93% de las instancias reales de la clase no abandono fueron identificadas correctamente por el modelo.

Para la clase 1, la recuperación es del 49%, lo que indica que solo el 49% de las instancias reales de la clase abandono fueron identificadas correctamente por el modelo.

Puntaje F1 (F1-Score): el puntaje F1 es la media armónica de precisión y recuperación y proporciona una medida única del rendimiento del modelo.

Para la clase 0, el puntaje F1 es del 88%, lo que indica un buen equilibrio entre precisión y recuperación para esta clase.

Para la clase 1, el puntaje F1 es del 58%, lo que sugiere que hay margen de mejora en el equilibrio entre precisión y recuperación para esta clase.

Soporte (Support): el soporte es el número de ocurrencias reales de cada clase en el conjunto de datos de prueba. Para la clase 0, hay 1549 instancias en el conjunto de datos de prueba. Para la clase 1, hay 561 instancias en el conjunto de datos de prueba.

Exactitud (Accuracy): la exactitud es la proporción de instancias clasificadas correctamente (tanto positivas como negativas) en relación con el total de instancias.

En este caso, la exactitud global del modelo es del 81%, lo que indica que el 81% de todas las instancias en el conjunto de datos de prueba fueron clasificadas correctamente por el modelo.

Promedio Ponderado (Weighted Average) y Promedio Macro (Macro Average):

Estas métricas proporcionan un resumen general del rendimiento del modelo en todas las clases.

El promedio ponderado tiene en cuenta la proporción de instancias en cada clase, mientras que el promedio macro no lo hace.

En este informe, el promedio ponderado de precisión, recuperación y puntaje F1 es del 80%, lo que indica el rendimiento general del modelo considerando todas las clases.

- Regresión Logística: la regresión logística es un modelo estadístico que se utiliza para predecir la probabilidad de que una variable dependiente binaria (dicotómica) tenga un valor de 1 o 0 en función de una o más variables independientes. A diferencia de la regresión lineal, la regresión logística utiliza la función logística para transformar la combinación lineal de las variables independientes en un valor entre 0 y 1, que se interpreta como la probabilidad de pertenecer a la categoría positiva.

Informe de Clasificación Regresión Logística

	precision	recall	f1-score	support
0	0.85	0.90	0.87	1549
1	0.67	0.55	0.61	561
accuracy			0.81	2110
macro avg	0.76	0.73	0.74	2110
weighted avg	0.80	0.81	0.80	2110

Este resultado corresponde al informe de clasificación generado por el modelo de regresión logística.

Precision (Precisión): es la proporción de verdaderos positivos sobre todos los positivos predichos por el modelo. Para la clase 0, la precisión es del 85%, lo que significa que el 85% de las predicciones de la clase 0 son correctas. Para la clase 1, la precisión es del 67%, lo que significa que el 67% de las predicciones de la clase 1 son correctas.

Recall (Recuperación o Sensibilidad): es la proporción de verdaderos positivos sobre todos los positivos reales en los datos. Para la clase 0, el recall es del 90%, lo que significa que el 90% de todos los casos reales de la clase 0 fueron identificados correctamente por el modelo. Para la clase 1, el recall es del 55%, lo que significa que solo el 55% de todos los casos reales de la clase 1 fueron identificados correctamente por el modelo.

F1-score (Puntuación F1): es una medida de la precisión del modelo que combina precision y recall en un solo número. Es útil cuando hay un desequilibrio entre las clases. El F1-score es del 87% para la clase 0 y del 61% para la clase 1.

Support (Soporte): es el número de ocurrencias reales de cada clase en los datos de prueba.

Accuracy (Precisión Global): es la proporción de muestras correctamente clasificadas sobre todas las muestras. En este caso, la precisión global del modelo es del 81%, lo que significa que el 81% de todas las muestras fueron clasificadas correctamente por el modelo.

Macro Average (Promedio Macro): es el promedio de las métricas (precision, recall, F1-score) calculadas para cada clase por separado. En este caso, el promedio macro de precision, recall y F1-score es del 76%, 73% y 74% respectivamente.

Weighted Average (Promedio Ponderado): es similar al promedio macro, pero se pondera según el soporte de cada clase. En este caso, el promedio ponderado de precision, recall y F1-score es del 80%.

- Árbol de Decisión: un clasificador de árbol de decisión es un modelo de aprendizaje supervisado que se utiliza para predecir la etiqueta de clase de una instancia basándose en las características de esa instancia. Este modelo toma decisiones en forma de un árbol, donde cada nodo interno representa una prueba en una característica, cada rama representa el resultado de la prueba, y cada hoja del árbol representa una etiqueta de clase.

Informe de Clasificación de Arbol de Decisión

	precision	recall	f1-score	support
0	0.84	0.82	0.83	1549
1	0.52	0.55	0.54	561
accuracy			0.75	2110
macro avg	0.68	0.69	0.68	2110
weighted avg	0.75	0.75	0.75	2110

El resultado es un informe de clasificación que evalúa el rendimiento de un modelo de arbol de decisión en un conjunto de datos de prueba.

Precision (Precisión): la precisión se refiere a la proporción de predicciones positivas correctas sobre el total de predicciones positivas realizadas por el modelo. Para la clase 0, la precisión es del 83%, lo que indica que el 83% de las predicciones de la clase 0 son correctas. Para la clase 1, la precisión es del 52%, lo que indica que el 52% de las predicciones de la clase 1 son correctas.

Recall (Recuperación): la recuperación se refiere a la proporción de instancias positivas que fueron correctamente identificadas por el modelo. Para la clase 0, la recuperación es del 82%, lo que indica que el 82% de todas las instancias verdaderamente positivas de la clase 0 fueron identificadas correctamente por el modelo. Para la clase 1, la recuperación es del 53%, lo que indica que el 53% de todas las instancias verdaderamente positivas de la clase 1 fueron identificadas correctamente por el modelo.

F1-score: el puntaje F1 es una media armónica de precisión y recuperación. Es útil cuando hay un desequilibrio entre las clases en los datos. El puntaje F1 busca un equilibrio entre precisión y recuperación. Para la clase 0, el puntaje F1 es del 83%, y para la clase 1, el puntaje F1 es del 52%.

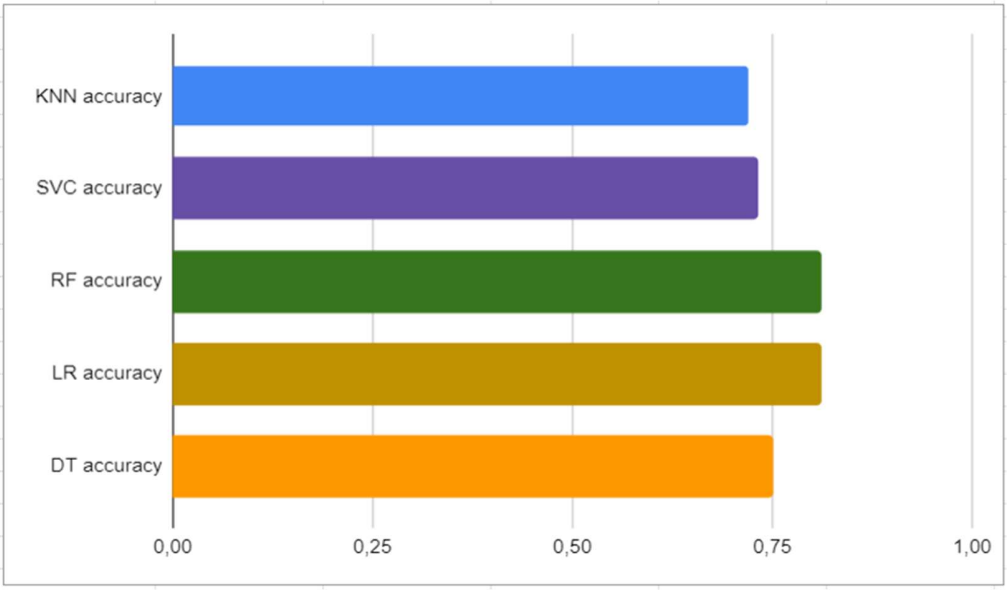
Support (Soporte): es el número real de ocurrencias de la clase en los datos de prueba.

Accuracy (Precisión Global): la precisión global mide la proporción de predicciones correctas (tanto positivas como negativas) realizadas por el modelo. En este caso, la precisión global es del 75%, lo que indica que el 75% de todas las predicciones realizadas por el modelo son correctas.

Macro Avg y Weighted Avg: son promedios de las métricas (precisión, recuperación, puntaje F1) calculadas para cada clase. En este caso, el promedio macro y el promedio ponderado de las métricas son aproximadamente iguales, lo que indica que no hay un gran desequilibrio entre las clases en los datos de prueba.

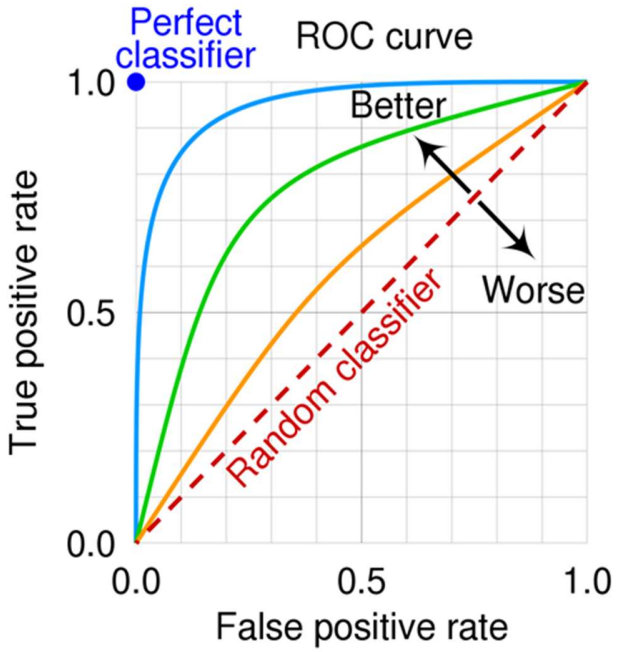
Este informe de clasificación proporciona una evaluación detallada del rendimiento del modelo de clasificación en ambos casos positivos y negativos.

A continuación, se observa un comparativo de la precisión promedio de los modelos evaluados:

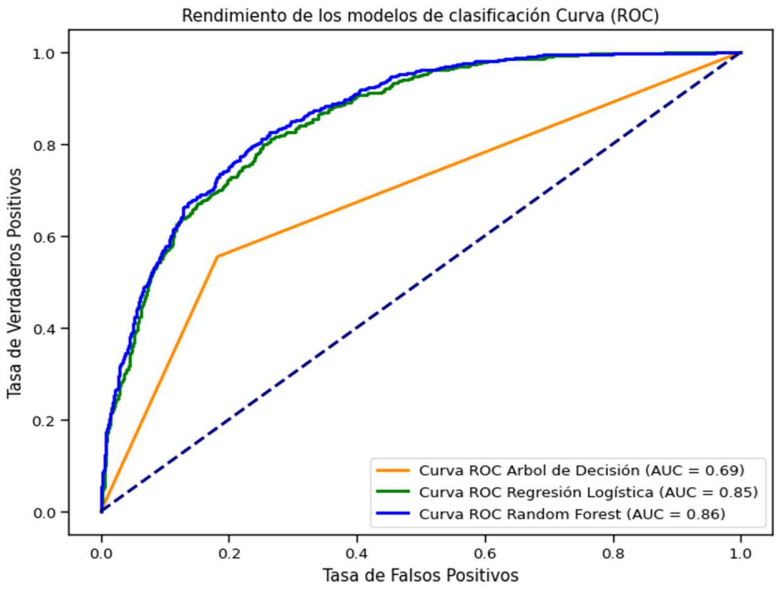


KNN accuracy	0,72
SVC accuracy	0,73
RF accuracy	0,81
LR accuracy	0,81
DT accuracy	0,75

Nótese que los mejores desempeños los presentaron Random Forest y Regresión Logística.

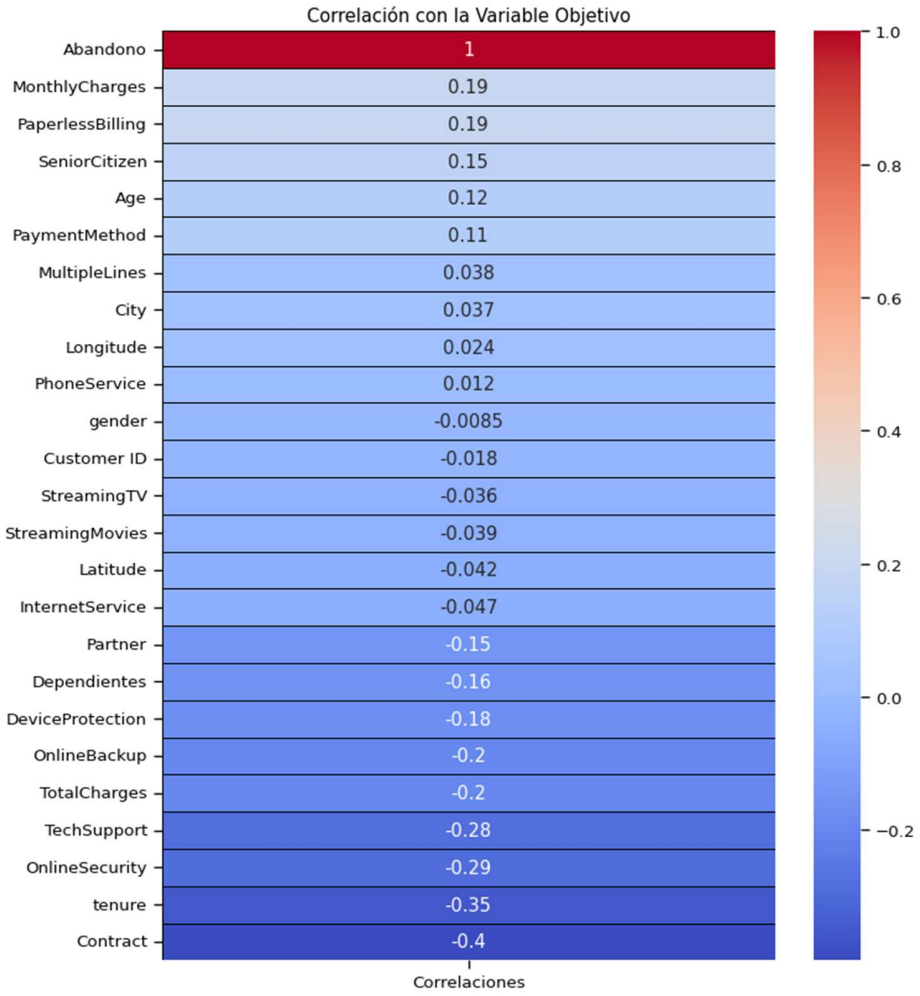


Utilizando el criterio de evaluación del área bajo la curva AUC, de acuerdo con la gráfica anterior, el modelo es mejor en las curvas más alejadas de la línea roja punteada. En el caso de dicha gráfica el mejor modelo sería el azul.



De acuerdo con el criterio anterior, el mejor desempeño es el modelo Random Forest.

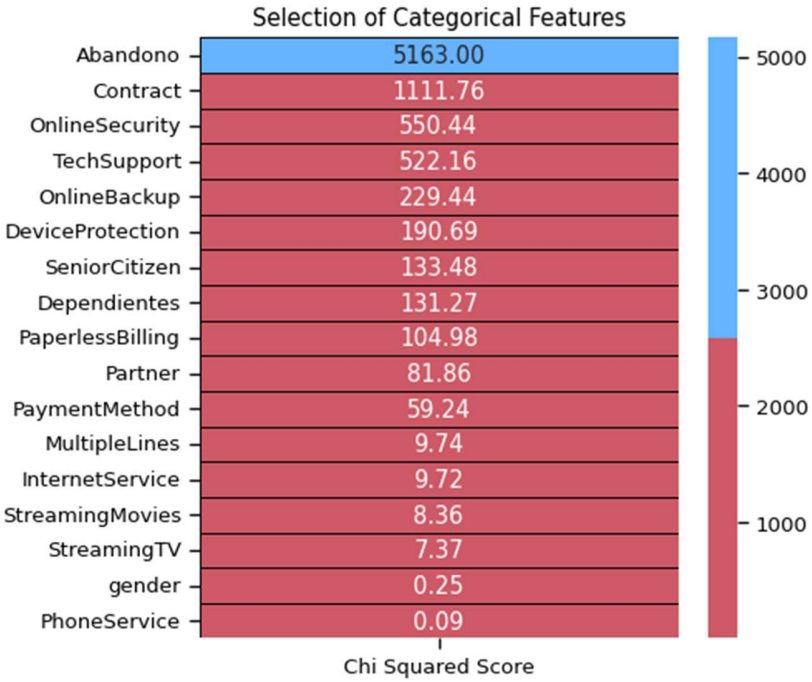
5.4 Selección de Variables



MultipleLines, PhoneService, gender, StreamingTV, StreamingMovies e InternetService no presentan ningún tipo de correlación. Se eliminan las características con coeficiente de correlación entre (-0.1,0.1). Las características restantes muestran una correlación positiva o negativa significativa.

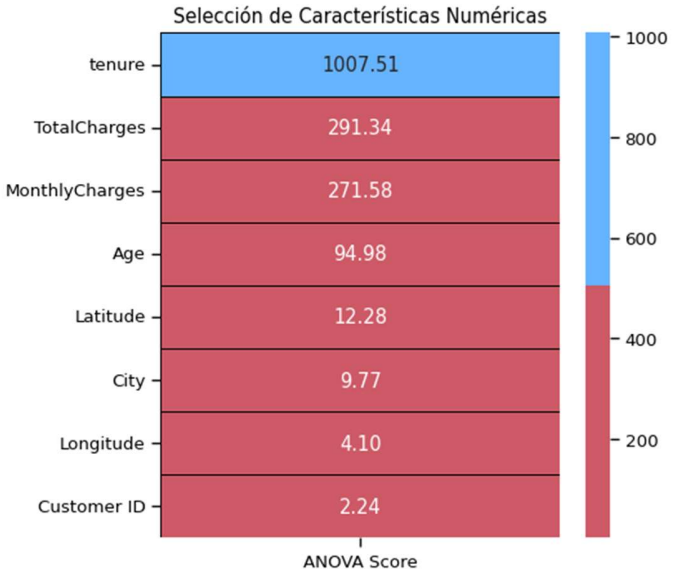
- **Selección de variables categóricas con Chi Cuadrado**

La selección de variables categóricas utilizando la prueba de Chi-cuadrado implica evaluar la relación entre cada variable categórica individual y la variable objetivo (Abandono). Esta técnica ayuda a identificar las variables categóricas más relevantes para predecir o explicar la variable objetivo.



PhoneService, gender, StreamingTV, StreamingMovies, MultipleLines e InternetService muestran una relación muy baja con respecto al Abandono.

- **Selección de variables numéricas con Prueba ANOVA**



Según la prueba ANOVA, cuanto mayor sea el valor de la puntuación ANOVA, mayor será la importancia de la característica.

De los resultados anteriores, se recomienda incluir las tres características con mayor puntuación. Se seleccionarán las 3 primeras de la puntuación.

- **Equilibrio de Datos usando SMOTE:**

Para hacer frente a datos desequilibrados, existen 2 opciones:

- Submuestreo: recorta las muestras mayoritarias de la variable objetivo.
- Sobremuestreo: aumenta las muestras minoritarias de la variable objetivo a muestras mayoritarias.

Después de hacer prueba-error con submuestreo y sobremuestreo. Se ha decidido utilizar el sobremuestreo.

Para el equilibrio de datos, usaremos imblearn. Declaración de pip: pip install imbalanced-learn.

```
over = SMOTE(sampling_strategy = 1)

f1 = Data3.iloc[:,:13].values
t1 = Data3.iloc[:,13].values

f1, t1 = over.fit_resample(f1, t1)
Counter(t1)
```

```
Counter({0: 5163, 1: 5163})
```

Se eliminaron 11 columnas: 5 numéricas y 6 categóricas.

5.5 Cross Validation

- **KNN:** Se realiza una comparación entre los informes de clasificación KNN utilizando validación cruzada y el dataset original muestra algunas diferencias significativas en las métricas de evaluación del modelo.

Informe de Clasificación KNN

	precision	recall	f1-score	support
0	0.76	0.87	0.81	5163
1	0.41	0.24	0.31	1869
accuracy			0.71	7032
macro avg	0.59	0.56	0.56	7032
weighted avg	0.67	0.71	0.68	7032

Para el informe de clasificación KNN utilizando validación cruzada:

- La precisión para la clase 0 (clientes que no abandonan) es del 76%, mientras que para la clase 1 (clientes que abandonan) es del 41%.
- El recall para la clase 0 es del 87%, mientras que para la clase 1 es del 24%.
- El puntaje F1 para la clase 0 es del 81%, mientras que para la clase 1 es del 31%.
- La precisión general del modelo es del 71%.

En comparación con el informe de clasificación KNN para el dataset original:

- La precisión para la clase 0 es ligeramente menor (74%) pero el recall es mayor (96%).
- La precisión para la clase 1 es similar (31%), pero el recall es significativamente menor (5%).
- El puntaje F1 para la clase 0 es más alto (83%), mientras que para la clase 1 es mucho menor (9%).
- La precisión general del modelo es del 72%.

Estas diferencias sugieren que el rendimiento del modelo KNN varía dependiendo de si se utiliza validación cruzada o no. La validación cruzada puede proporcionar una evaluación más robusta del modelo al evaluarlo en múltiples particiones de los datos, lo que puede revelar mejor el rendimiento general del modelo en comparación con el uso de un único conjunto de datos de prueba. En este caso, el informe de clasificación KNN utilizando validación cruzada muestra un rendimiento ligeramente inferior en términos de precisión y puntajes F1 en comparación con el informe basado en el dataset original. Esto puede indicar que el modelo KNN no generaliza tan bien como se esperaba en diferentes particiones de los datos.

- **SVC:** Al comparar los informes de clasificación de SVM, podemos notar algunas diferencias significativas entre los resultados del cross-validation y el dataset original.

Informe de Clasificación SVC

	precision	recall	f1-score	support
0	0.73	1.00	0.85	5163
1	1.00	0.00	0.00	1869
accuracy			0.73	7032
macro avg	0.87	0.50	0.42	7032
weighted avg	0.80	0.73	0.62	7032

1. **Precision, Recall y F1-score para la clase 1 (Abandono):**

- En el cross-validation, la precisión, recall y f1-score para la clase 1 son reportados como 1.00. Esto significa que para todas las muestras predichas como clase 1, todas fueron correctas. Sin embargo, este resultado no es realista y sugiere un sobreajuste del modelo durante la validación cruzada.

- En el dataset original, la precisión, recall y f1-score para la clase 1 son reportados como 1.00 también, lo cual es similar al resultado del cross-validation.

2. **Precision, Recall y F1-score para la clase 0 (No Abandono):**

- En el cross-validation, la precisión, recall y f1-score para la clase 0 son reportados como 0.73, 1.00 y 0.85 respectivamente. Esto indica que el modelo es capaz de predecir correctamente todas las muestras de la clase 0, pero con un bajo rendimiento en la clasificación de la clase 1.
- En el dataset original, los valores de precisión, recall y f1-score para la clase 0 son similares a los del cross-validation, lo cual muestra cierta consistencia en la capacidad del modelo para clasificar correctamente las muestras de la clase 0.

3. **Accuracy y Macro Avg:**

- La precisión general (accuracy) y las métricas macro avg son similares en ambos casos, lo que indica que el modelo tiene un rendimiento similar en términos generales en ambos conjuntos de datos.

Los resultados del cross-validation muestran una precisión perfecta para la clase 1 y un rendimiento ligeramente superior para la clase 0 en comparación con el dataset original. Esto puede deberse a un sobreajuste del modelo durante la validación cruzada o a la naturaleza específica de los datos utilizados en el proceso de validación. Es importante tener en cuenta estas diferencias al interpretar los resultados y al considerar la generalización del modelo a datos no vistos.

- **Random Forest:** Al comparar los dos informes de clasificación del modelo Random Forest, podemos observar algunas diferencias significativas.

Informe de Clasificación Random Forest

	precision	recall	f1-score	support
0	0.83	0.91	0.87	5163
1	0.68	0.49	0.57	1869
accuracy			0.80	7032
macro avg	0.76	0.70	0.72	7032
weighted avg	0.79	0.80	0.79	7032

1. **Precisión, Recall y F1-score para la clase 0 (No Abandono):**

- En el primer informe (con validación cruzada), la precisión es ligeramente menor (0.83 vs. 0.83), pero el recall es mayor (0.91 vs. 0.93) en comparación con el segundo informe (sin validación cruzada). Esto significa que, en el primer informe,

el modelo identifica correctamente más casos de "No Abandono", pero también hace más falsos positivos.
- El F1-score es prácticamente el mismo en ambos informes para la clase 0.

2. **Precisión, Recall y F1-score para la clase 1 (Abandono):**

- En el primer informe, la precisión es ligeramente mayor (0.67 vs. 0.72) en comparación con el segundo informe. Esto significa que en el primer informe, el modelo hace menos falsos positivos en la clasificación de "Abandono".
- Sin embargo, el recall es menor en el primer informe (0.49 vs. 0.49), lo que indica que el modelo identifica correctamente menos casos de "Abandono" en comparación con el segundo informe.
- Como resultado, el F1-score para la clase 1 es más bajo en el primer informe (0.57 vs. 0.58).

3. **Exactitud (Accuracy):**

- La exactitud es ligeramente más alta en el segundo informe (0.81 vs. 0.80), lo que indica que el modelo sin validación cruzada tiene un mejor desempeño general en la clasificación correcta de ambas clases.

Aunque la precisión y el recall varían ligeramente entre los dos informes, la exactitud general es comparable. La diferencia principal radica en cómo se obtienen las métricas: el primer informe utiliza validación cruzada, lo que proporciona una evaluación más robusta del modelo, mientras que el segundo informe se basa en un conjunto de datos de prueba separado.

- **Regresión Logística:** Al comparar los dos informes de clasificación del modelo se pueden observar estas diferencias: en el informe sin validación cruzada, la exactitud es ligeramente mayor (81%) en comparación con el informe con validación cruzada (80%).

Informe de Clasificación Regresión Logística

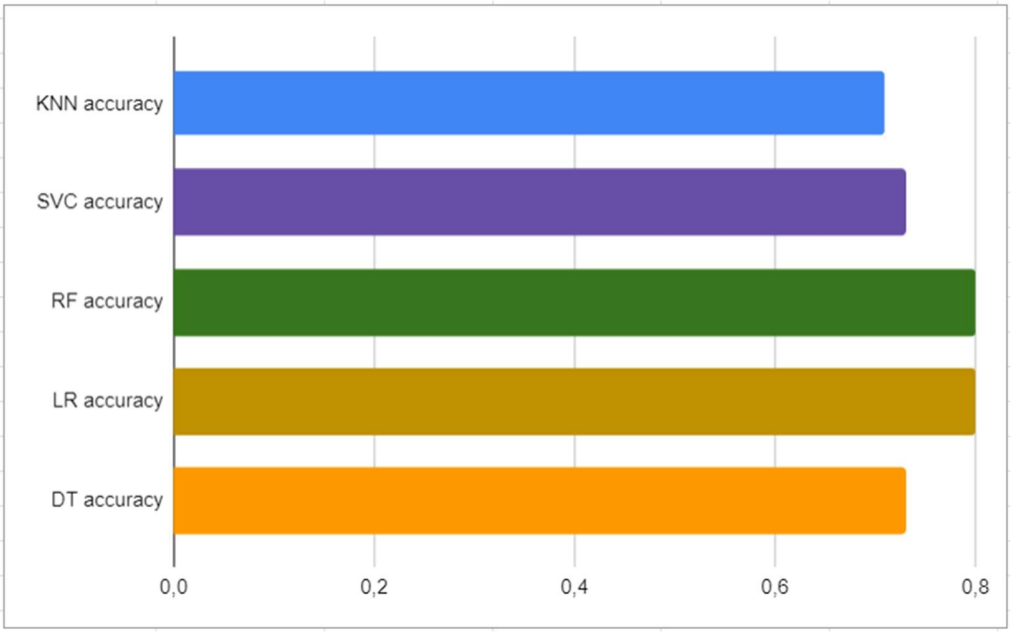
	precision	recall	f1-score	support
0	0.85	0.90	0.87	5163
1	0.66	0.55	0.60	1869
accuracy			0.80	7032
macro avg	0.75	0.72	0.73	7032
weighted avg	0.80	0.80	0.80	7032

- **Árbol de Decisión:** La precisión general (accuracy) es ligeramente más alta en el conjunto de datos original (0.75) en comparación con la validación cruzada (0.74).

Informe de Clasificación Árbol de Decisión

	precision	recall	f1-score	support
0	0.82	0.81	0.82	5163
1	0.50	0.52	0.51	1869
accuracy			0.73	7032
macro avg	0.66	0.67	0.66	7032
weighted avg	0.74	0.73	0.73	7032

En el siguiente gráfico se realiza un comparativo del desempeño de los modelos:



Resultados Dataset

KNN accuracy	0,72
SVC accuracy	0,73
RF accuracy	0,81
LR accuracy	0,81
DT accuracy	0,75

Resultados Cross Validation

KNN accuracy	0,71
SVC accuracy	0,73
RF accuracy	0,80
LR accuracy	0,80
DT accuracy	0,73

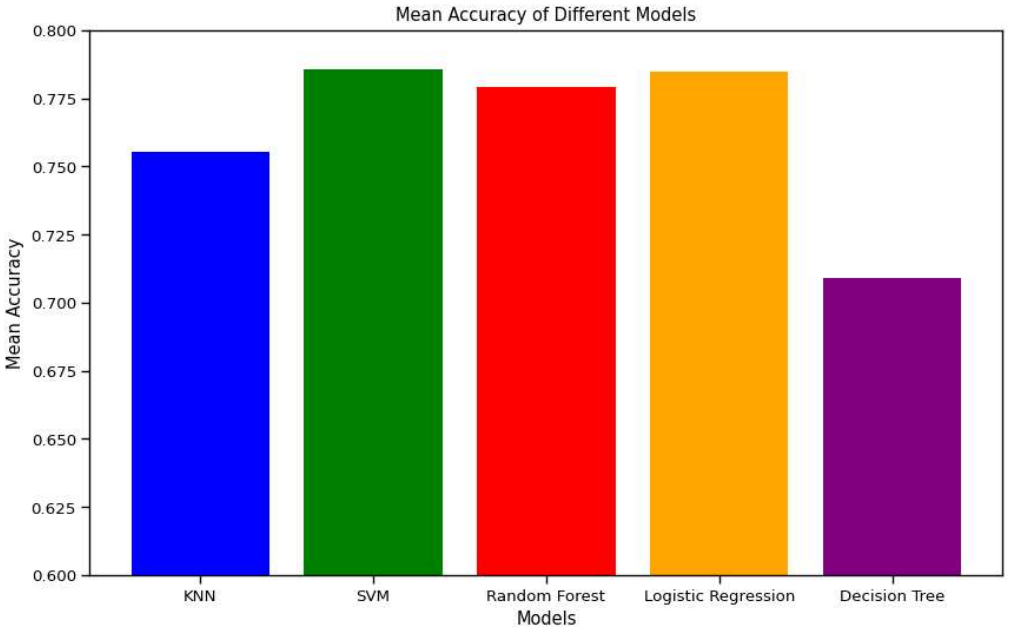
5.6 PCA (Principal Component Analysis)

El objetivo principal de PCA es reducir la dimensionalidad del conjunto de datos al proyectarlo en un nuevo espacio de características definido por los componentes principales.

Se ha utilizado la clase PCA de scikit-learn para realizar la reducción de dimensionalidad y se ha integrado en una tubería (pipeline) junto con la estandarización de los datos (StandardScaler) y el modelo de clasificación correspondiente.

Se realiza la validación cruzada para cada modelo utilizando las tuberías definidas. Las puntuaciones se almacenan en las variables correspondientes.

```
KNN Cross-Validation Scores: [0.75764037 0.76687989 0.75462304
0.76458037 0.76386913]
Mean Accuracy: 0.7615185604187961
SVM Cross-Validation Scores: [0.78322672 0.78891258 0.77667141
0.79231863 0.7916074 ]
Mean Accuracy: 0.786547348605479
Random Forest Cross-Validation Scores: [0.77896233 0.78535892
0.76529161 0.77596017 0.78093883]
Mean Accuracy: 0.7773023725105422
Logistic Regression Cross-Validation Scores: [0.78393746
0.79175551 0.77809388 0.78662873 0.78733997]
Mean Accuracy: 0.7855511105314719
Decision Tree Cross-Validation Scores: [0.69651741 0.73134328
0.69416785 0.70981508 0.71479374]
Mean Accuracy: 0.7093274735851327
```



6. Optimización – Ajuste de Hiperparámetros

La optimización o ajuste de hiperparámetros es un proceso crucial en el desarrollo de modelos de aprendizaje automático. Los hiperparámetros son parámetros que no se aprenden directamente del conjunto de datos durante el entrenamiento del modelo, sino que se configuran antes del proceso de entrenamiento y afectan el rendimiento y el comportamiento del modelo. La optimización de hiperparámetros se refiere al proceso de encontrar la combinación óptima de valores para estos hiperparámetros con el fin de mejorar el rendimiento del modelo.

Entre las principales razones por las cuales es importante realizar la optimización de hiperparámetros tenemos:

1. **Mejora del rendimiento del modelo:** los hiperparámetros afectan directamente el rendimiento del modelo. Al encontrar la combinación óptima de valores para estos hiperparámetros, es posible mejorar el rendimiento del modelo en términos de precisión, recall, F1-score, etc.
2. **Prevención del sobreajuste:** al ajustar los hiperparámetros, es posible evitar el sobreajuste o sobreentrenamiento del modelo, donde el modelo se ajusta demasiado a los datos de entrenamiento y no generaliza bien a nuevos datos.
3. **Eficiencia computacional**: la optimización de hiperparámetros puede ayudar a encontrar combinaciones de hiperparámetros que conduzcan a modelos más simples o más eficientes computacionalmente, lo que puede ser importante en aplicaciones donde se requiere un procesamiento rápido de grandes volúmenes de datos.
4. **Interpretación del modelo:** al ajustar los hiperparámetros, es posible obtener modelos más interpretables o explicables, lo que facilita la comprensión de cómo funcionan y toman decisiones.

En la validación y la validación cruzada en el PCA arrojó como resultado que los modelos de mejor ajuste son SVM y Regresión Logística.

6.1 Hiperparámetros para un clasificador SVM

Para un clasificador SVM (Support Vector Machine), algunos de los hiperparámetros más importantes que se pueden ajustar son:

C: Parámetro de regularización que controla la compensación entre maximizar el margen y minimizar el error de clasificación. Valores más altos de C permiten un margen más estrecho, pero pueden conducir a un sobreajuste.

kernel: Especifica el tipo de función de kernel a utilizar en el modelo SVM. Algunas opciones comunes son "linear", "poly", "rbf" (Radial Basis Function), y "sigmoid". Cada kernel tiene sus propias características y puede funcionar mejor para diferentes tipos de datos.

gamma: Coeficiente de kernel para los kernels "rbf", "poly" y "sigmoid". Controla la influencia de un solo ejemplo de entrenamiento, con valores más altos que resultan en una influencia más localizada.

degree: Grado del kernel polinomial (solo para kernels polinomiales).

coef0: Término independiente en funciones de kernel polinomiales y sigmoidales.

Estos son solo algunos de los hiperparámetros más comunes que se pueden ajustar en un clasificador SVM. La selección adecuada de hiperparámetros puede tener un gran impacto en el rendimiento del modelo SVM en términos de precisión y generalización.

```
# Definir las distribuciones de probabilidad para los hiperparámetros
param_dist = {
    'C': uniform(1.0, 3.0),  # Desde 1.0 hasta 3.0
    'kernel': ['rbf', 'linear', 'sigmoid'],  # Opciones de kernel
    'gamma': ['scale', 'auto']  # Opciones de gamma
}
```

```
RandomizedSearchCV
RandomizedSearchCV(cv=StratifiedKFold(n_splits=3, random_state=None, shuffle=False),
                   estimator=SVC(), n_iter=3, n_jobs=-1,
                   param_distributions={'C': <scipy.stats._distn_infrastructure.rv_continuous_frozen object at 0x7b2494110e50>,
                                        'gamma': ['scale', 'auto'],
                                        'kernel': ['rbf', 'linear', 'sigmoid']},
                   verbose=1)
estimator: SVC
SVC()
SVC
SVC()
```

```
rs_svc.best_score_
0.7781406836006203
```

```
rs_svc.score(X_test, y_test)
0.8018957345971564
```

Informe de Clasificación:

- Precisión (precision): La precisión para la clase 0 es del 83%, lo que indica que el 83% de las instancias clasificadas como clase 0 son realmente de esa clase. Para la clase 1, la precisión es del 68%, lo que indica que el 68% de las instancias clasificadas como clase 1 son realmente de esa clase.
- Recall: El recall para la clase 0 es del 92%, lo que indica que el 92% de las instancias de la clase 0 fueron correctamente identificadas por el modelo. Para la clase 1, el recall es del 48%, lo que indica que solo el 48% de las instancias de la clase 1 fueron correctamente identificadas.

- F1-score: El F1-score es una medida que combina precisión y recall en una sola métrica. Para la clase 0, el F1-score es del 87%, mientras que para la clase 1 es del 56%.
- Exactitud (accuracy): La exactitud global del modelo es del 80%, lo que indica el porcentaje total de predicciones correctas en el conjunto de datos.

Informe de Clasificación SVC

	precision	recall	f1-score	support
0	0.83	0.92	0.87	1549
1	0.68	0.48	0.56	561
accuracy			0.80	2110
macro avg	0.75	0.70	0.72	2110
weighted avg	0.79	0.80	0.79	2110

Esta es la mejor puntuación alcanzada por el modelo SVM entrenado con los mejores hiperparámetros encontrados durante la búsqueda aleatoria (RandomizedSearchCV) en el conjunto de datos de entrenamiento.

MATRIZ DE CONFUSIÓN

TN **1421**	FP **128**
290 FN	**271** TP

Matriz de Confusión:

- En la fila 1, columna 1: 1421 verdaderos negativos (TN), es decir, el modelo clasificó correctamente 1421 instancias que realmente pertenecen a la clase negativa (0).
- En la fila 2, columna 1: 290 falsos negativos (FN), es decir, el modelo clasificó incorrectamente 290 instancias que realmente pertenecen a la clase positiva (1).
- En la fila 1, columna 2: 128 falsos positivos (FP).
- En la fila 2, columna 2: 271 verdaderos positivos (TP).

6.2 Hiperparámetros para Regresión Logística

Para la regresión logística, algunos de los hiperparámetros más importantes que se pueden ajustar son:

penalty: tipo de regularización a aplicar. Puede ser "l1" (regularización L1), "l2" (regularización L2), "elasticnet" (combinación de L1 y L2) o "none" (sin regularización).

C: parámetro de regularización inverso, que controla la fuerza de la regularización. Valores más pequeños de C indican una regularización más fuerte.

solver: algoritmo a utilizar en la optimización del problema. Las opciones comunes incluyen "liblinear" para datos pequeños y "sag" para conjuntos de datos grandes.

max_iter: número máximo de iteraciones permitidas durante la optimización.

multi_class: especifica cómo manejar la clasificación multiclase. Puede ser "ovr" (uno contra el resto) o "multinomial" (clasificación multinomial).

class_weight: ponderación de clases para abordar el desequilibrio de clases. Puede ser "balanced" para ajustar automáticamente los pesos de clase inversamente proporcionales a las frecuencias de clase en los datos de entrada.

Ajustar adecuadamente estos hiperparámetros puede mejorar el rendimiento de la regresión logística en términos de precisión y generalización.

```
params = {
    'penalty': ['l1', 'l2', 'elasticnet', 'none'],
    'C': np.logspace(-4, 4, 20),
    'solver': ['lbfgs', 'newton-cg', 'liblinear', 'sag', 'saga'],
    'max_iter': [10, 100, 250, 500]
}
```

```
RandomizedSearchCV
RandomizedSearchCV(cv=RepeatedStratifiedKFold(n_repeats=10, n_splits=2, random_state=None),
                   estimator=LogisticRegression(), n_iter=2, n_jobs=-1,
                   param_distributions={'C': array([1.00000000e-04, 2.63665090e-04, 6.95192796e-04, 1.83298071e-03,
       4.83293024e-03, 1.27427499e-02, 3.35981829e-02, 8.85866790e-02,
       2.33572147e-01, 6.15848211e-01, 1.62377674e+00, 4.28133240e+00,
       1.12883789e+01, 2.97635144e+01, 7.84759970e+01, 2.06913808e+02,
       5.45559478e+02, 1.43844989e+03, 3.79269019e+03, 1.00000000e+04]),
                                        'max_iter': [10, 100, 250, 500],
                                        'penalty': ['l1', 'l2', 'elasticnet',
                                                    'none'],
                                        'solver': ['lbfgs', 'newton-cg',
estimator: LogisticRegression
LogisticRegression()
LogisticRegression
LogisticRegression()
```

```
rs.best_params_
{'solver': 'newton-cg',
 'penalty': 'none',
 'max_iter': 250,
 'C': 545.5594781168514}
```

```
rs.score(X_test, y_test)
0.8127962085308057
```

```
rs.best_score_
0.7962819991873222
```

La **exactitud global del modelo es del 81%,** lo que indica el porcentaje total de predicciones correctas en el conjunto de datos.

Informe de Clasificación Regresión Logística

	precision	recall	f1-score	support
0	0.86	0.89	0.88	1549
1	0.67	0.59	0.63	561
accuracy			0.81	2110
macro avg	0.76	0.74	0.75	2110
weighted avg	0.81	0.81	0.81	2110

Informe de Clasificación:

- Precisión (precision): La precisión para la clase 0 es del 86%, lo que indica que el 86% de las instancias clasificadas como clase 0 son realmente de esa clase. Para la clase 1, la precisión es del 67%, lo que indica que el 67% de las instancias clasificadas como clase 1 son realmente de esa clase.
- Recall: El recall para la clase 0 es del 89%, lo que indica que el 89% de las instancias de la clase 0 fueron correctamente identificadas por el modelo. Para la clase 1, el recall es del 59%, lo que indica que el 59% de las instancias de la clase 1 fueron correctamente identificadas.
- F1-score: El F1-score es una medida que combina precisión y recall en una sola métrica. Para la clase 0, el F1-score es del 88%, mientras que para la clase 1 es del 63%.
- Exactitud (accuracy): La exactitud global del modelo es del 81%, lo que indica el porcentaje total de predicciones correctas en el conjunto de datos.
- Macro promedio (macro avg): El promedio de las métricas (precision, recall, f1-score) para ambas clases. En este caso, el promedio ponderado de precision, recall y f1-score para ambas clases.
- Ponderado promedio (weighted avg): Similar al macro promedio, pero ponderado por el soporte de cada clase. En este caso, el soporte es el número de instancias de cada clase en el conjunto de datos.

MATRIZ DE CONFUSIÓN

TN **1385**	FP **164**
231 FN	**330** TP

Matriz de Confusión:

- Verdaderos Negativos (TN): 1385
- Falsos Positivos (FP): 164

- Falsos Negativos (FN): 231
- Verdaderos Positivos (TP): 330

6.3 Revisión de Modelos

Comparación de resultados:

1. Clasificador SVM:

- Precisión (precision) para la clase 0: 83%
- Recall para la clase 0: 92%
- F1-score para la clase 0: 87%
- Precisión para la clase 1: 68%
- Recall para la clase 1: 48%
- F1-score para la clase 1: 56%
- Exactitud (accuracy): 80%

2. Regresión logística:

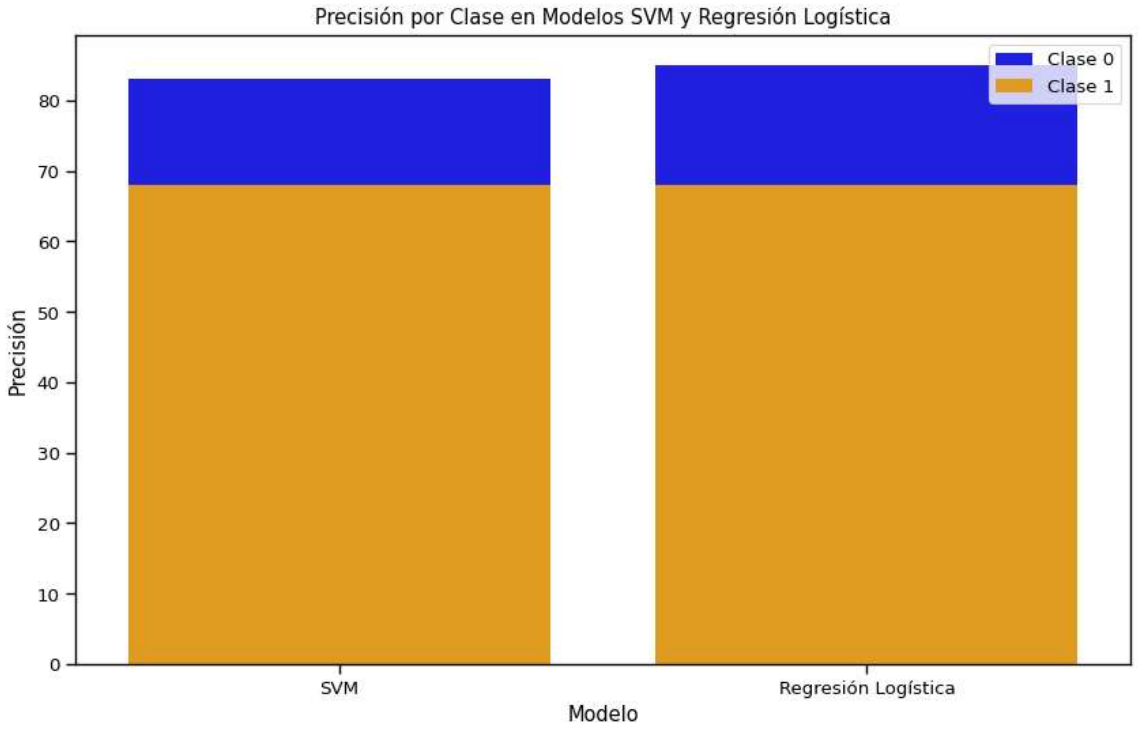
- Precisión para la clase 0: 85%
- Recall para la clase 0: 91%
- F1-score para la clase 0: 88%
- Precisión para la clase 1: 68%
- Recall para la clase 1: 55%
- F1-score para la clase 1: 61%
- Exactitud (accuracy): 81%

Comparación:

- Ambos modelos tienen una precisión similar para la clase 0, pero la regresión logística tiene un recall ligeramente mejor para esta clase.
- La regresión logística tiene un mejor desempeño en términos de precisión, recall y F1-score para la clase 1 en comparación con el clasificador SVM.
- Ambos modelos tienen una precisión global (accuracy) similar, pero la regresión logística tiene una mejor precisión ponderada y macro promedio.
- Ambos modelos tienen un F1-score ponderado similar, pero la regresión logística tiene un F1-score macro promedio ligeramente mejor.

	Modelo	Precisión Clase 0	Recall clase 0	F1-score Clase 0
0	SVM	83	92	87
1	Regresión Logística	85	91	88

	Precisión Clase 1	Recall clase 1	F1-score Clase 1	Exactitud (accuracy)
0	68	48	56	80
1	68	55	61	81



7. Conclusiones

Basándonos en los resultados de los modelos de SVM y regresión logística para predecir el abandono de clientes de telecomunicaciones, podemos sacar las siguientes conclusiones:

1. **Precisión de la predicción:**

Ambos modelos tienen una precisión similar para predecir la clase de clientes que no abandonan el servicio (clase 0).

La regresión logística tiene una precisión ligeramente mejor para predecir la clase de clientes que sí abandonan el servicio (clase 1).

2. **Recall de la predicción:**

Ambos modelos tienen un recall similar para la clase de clientes que no abandonan el servicio (clase 0), con valores alrededor del 90%.

La regresión logística tiene un recall mejorado para la clase de clientes que sí abandonan el servicio (clase 1), siendo aproximadamente un 10% más alto que el del modelo SVM.

3. **F1-score:**

El F1-score, que es una medida que combina precisión y recall, muestra que la regresión logística tiene un mejor equilibrio entre ambas clases en comparación con el modelo SVM.

8. Recomendaciones

- Incluir en los planes mes a mes paquetes promocionales de suscripción a seguridad en línea y soporte técnico.
- Ofrecer tarifas diferenciales a los clientes que contratan fibra óptica durante los 30 primeros meses
- Ofrecer servicios promocionales a los clientes con factura mensual entre 65 y 110 USD para disminuir la pérdida de clientes que pagan esas tarifas
- Basándonos en los resultados, podríamos sugerir el uso de la regresión logística sobre el SVM para predecir el abandono de clientes de telecomunicaciones, ya que tiene un mejor rendimiento en términos de recall y F1-score para la clase de clientes que sí abandonan el servicio.
- Además, sería importante realizar un análisis más detallado para identificar las características de los clientes que están abandonando el servicio, lo que podría proporcionar información valiosa para tomar medidas preventivas y retener a esos clientes. Esto podría implicar la implementación de estrategias de retención de clientes, como descuentos, ofertas personalizadas o mejoras en el servicio al cliente.

9. Futuros trabajos

Para futuros trabajos prediciendo la pérdida de clientes, se pueden considerar las siguientes líneas de acción:

1. Exploración de otras técnicas de modelado: además de la regresión logística y el SVM, se pueden explorar otras técnicas de modelado como árboles de decisión, bosques aleatorios, gradient boosting, redes neuronales, entre otros. Esto permitirá evaluar si algún otro modelo ofrece un rendimiento aún mejor en la predicción de la pérdida de clientes.

2. Incorporación de más datos: se puede buscar la incorporación de más datos relevantes para el problema, como datos demográficos adicionales, historial de interacciones con el servicio, información sobre la competencia, entre otros. Más datos podrían mejorar la capacidad del modelo para capturar los factores que influyen en la pérdida de clientes.

3. Selección de características más precisa: utilizar técnicas más avanzadas para la selección de características, como la selección de características basada en modelos o técnicas de reducción de dimensionalidad, para identificar las características más relevantes para la predicción de la pérdida de clientes.

4. Optimización de hiperparámetros más exhaustiva: realizar una búsqueda más exhaustiva de los hiperparámetros óptimos para los modelos seleccionados, utilizando técnicas como la optimización bayesiana, para mejorar aún más el rendimiento de los modelos.

5. Evaluación del impacto económico: además de medir el rendimiento del modelo en términos de métricas de evaluación estándar, como precisión, recall y F1-score, también sería útil evaluar el impacto económico de las predicciones del modelo. Esto podría implicar el cálculo del valor esperado de retener a un cliente identificado como en riesgo de pérdida versus el costo de implementar una estrategia de retención.