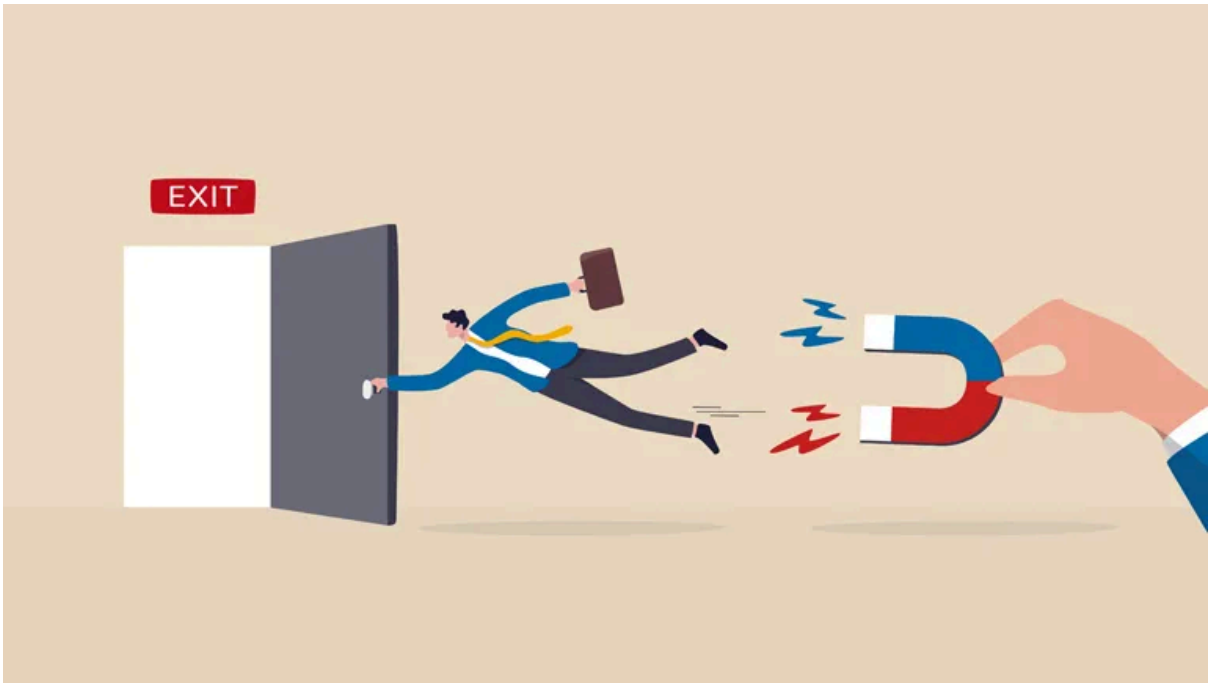
Trabajo final: Minería de Datos: Predicción del Churn:



Clara Pinto, Carlos Goyache y Julen García

1. Introducción:

En un entorno empresarial cada vez más competitivo, la retención de clientes se ha convertido en una prioridad estratégica para las organizaciones. La pérdida de clientes, conocida como *churn*, representa un desafío significativo para empresas de sectores como las telecomunicaciones, los servicios financieros, el comercio electrónico y otros donde la recurrencia del cliente es fundamental. Comprender los factores que llevan a un cliente a abandonar un servicio y anticiparse a su comportamiento es esencial para reducir costos, mejorar la satisfacción del cliente y aumentar los ingresos a largo plazo.

La predicción de *churn* mediante técnicas de análisis de datos y aprendizaje automático permite a las empresas identificar patrones de comportamiento asociados con la fuga de clientes y tomar decisiones informadas para prevenirla. Este trabajo tiene como objetivo desarrollar un modelo predictivo capaz de identificar clientes de una empresa de telecomunicaciones con alta probabilidad de abandono, utilizando un conjunto de datos históricos y diversas variables relevantes. Para ello, se utilizarán distintas técnicas de preprocesamiento de datos, se explorarán distintos algoritmos de clasificación, se evaluará su rendimiento y se analizará la importancia de las variables que influyen en la decisión del cliente.

2. Análisis de los datos:

En primer lugar, realizaremos un análisis exploratorio de los datos disponibles con el objetivo de identificar las técnicas de preprocesamiento más adecuadas y así optimizar el aprovechamiento del conjunto de datos.

2.1. Variables del problema:

El conjunto de datos utilizado en este trabajo está compuesto por diferentes tipos de variables, cada una con un papel clave en la predicción del *churn*. Estas variables pueden clasificarse en dos categorías principales:

Variables numéricas:

- Tenure: Número de meses que el cliente ha estado con la compañía.
- MonthlyCharges: Monto que el cliente paga mensualmente por el servicio.
- TotalCharges: Monto total que el cliente ha pagado desde que inició el servicio.

Variables categóricas:

- Gender: Género del cliente.
    - Female
    - Male
- SeniorCitizen: Indica si el cliente es una persona mayor.
    - 0
    - 1
- Partner: Indica si el cliente tiene pareja.
    - Yes
    - No
- Dependents: Indica si el cliente tiene personas a su cargo.
    - Yes
    - No
- PhoneService: Indica si el cliente tiene servicio telefónico.
    - Yes
    - No
- MultipleLines: Indica si el cliente tiene múltiples líneas telefónicas.
    - No phone service
    - No
    - Yes
- InternetService: Tipo de servicio de internet contratado.
    - No
    - DSL
    - Fiber optic
- OnlineSecurity: Indica si el cliente tiene servicio de seguridad en línea.
    - Yes
    - No

    - No internet service
- OnlineBackup: Indica si el cliente tiene servicio de respaldo en línea.
    - Yes
    - No
    - No internet service
- DeviceProtection: Indica si el cliente tiene protección para dispositivos.
    - Yes
    - No
    - No internet service
- TechSupport: Indica si el cliente tiene soporte técnico.
    - Yes
    - No
    - No internet service
- StreamingTV: Indica si el cliente tiene servicio de televisión en streaming.
    - Yes
    - No
    - No internet service
- StreamingMovies: Indica si el cliente tiene servicio de películas en streaming.
    - Yes
    - No
    - No internet service
- Contract: Tipo de contrato del cliente.
    - Month-to-month
    - One year
    - Two year
- PaperlessBilling: Indica si el cliente recibe la factura de manera electrónica.
    - Electronic check
    - Mailed check
    - Bank transfer (automatic)
    - Credit card (automatic)
- Churn: Es una variable que indica si un cliente ha cancelado su servicio o ha dejado de ser cliente de la empresa.
    - Yes
    - No

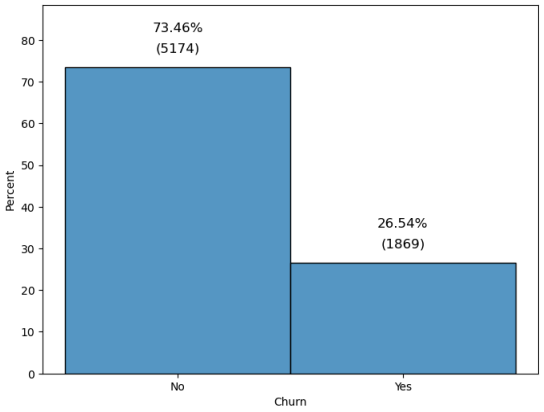
Al analizar el contenido de las variables del conjunto de datos, se identificaron 11 valores nulos en la variable "TotalCharges". Estos valores deben ser tratados adecuadamente antes de aplicar cualquier técnica de modelado, ya que podrían afectar la calidad y precisión de los resultados.

2.2. Distribución de las clases:

Con el fin de evaluar la complejidad del problema de clasificación, analizamos la distribución de las clases para detectar posibles desbalances. Esta visualización nos permite identificar

si existe una distribución desigual entre las clases, lo cual podría influir en el rendimiento del modelo y en la necesidad de aplicar técnicas de balanceo de clases.

Tras el análisis, vimos que nuestro conjunto de datos está compuesto por 7043 ejemplos, y que la distribución de las clases era la siguiente:



Como podemos observar, nuestro dataset está desbalanceado. Esto nos sugirió que probablemente la clasificación correcta de la variable “Churn” iba a ser una tarea complicada. Más concretamente, tenemos que el Imbalanced Ratio (IR) es 2.77.

2.3. Generación de variables:

Antes de comenzar a realizar un análisis exploratorio de las variables, creamos algunas nuevas variables a partir de las que ya tenemos.

Estas nuevas variables pueden ayudarnos a descubrir patrones relevantes que no son evidentes en la base de datos original, y pueden mejorar significativamente el rendimiento de los modelos predictivos o facilitar la interpretación del comportamiento de los clientes según el problema que estemos abordando.

Para ello, generamos las siguientes variables:

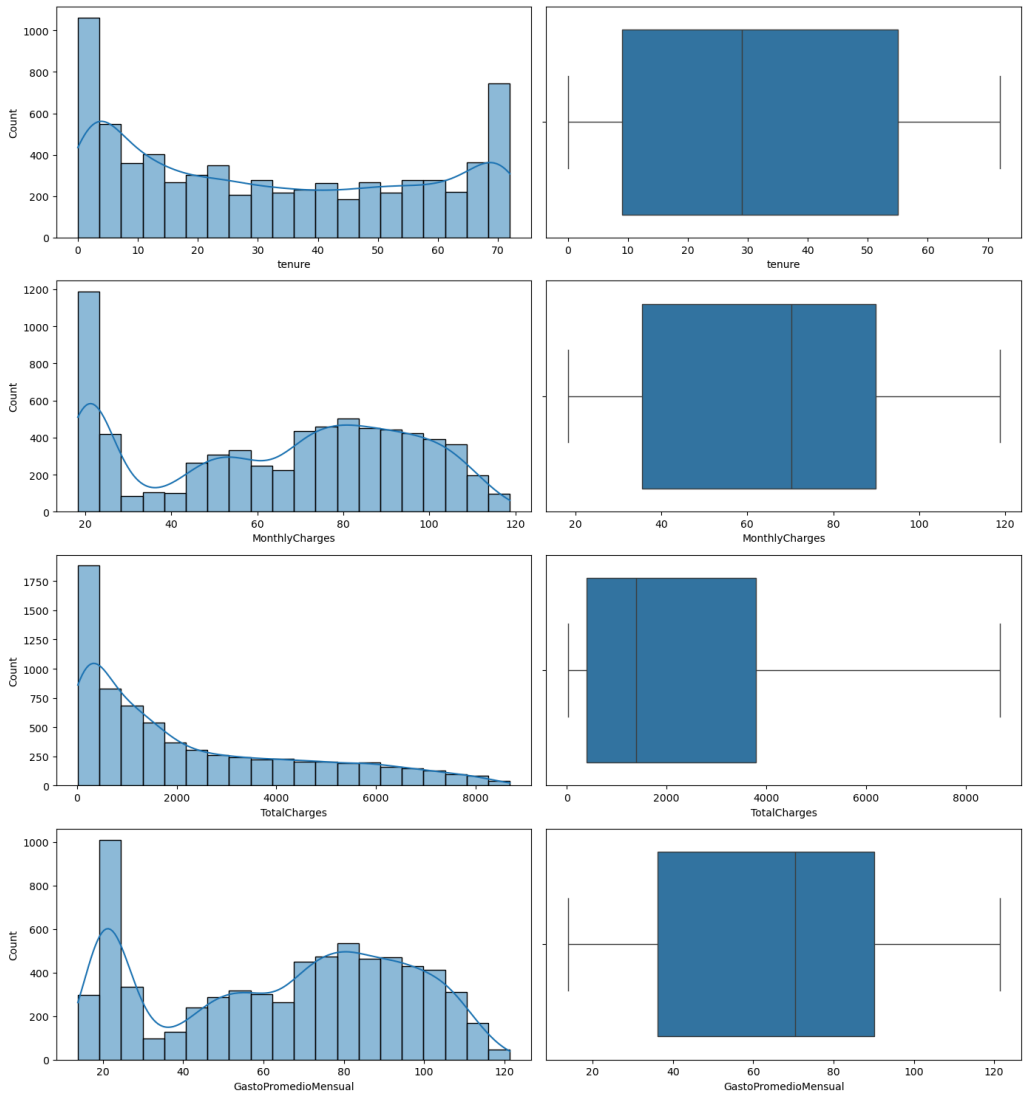
1. **NumServiciosContratados**
   Cuenta del total de servicios activos por cliente.
   Variables base: PhoneService, MultipleLines, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies.
2. **NumServiciosDeSeguridad**
   Suma de servicios relacionados con seguridad y soporte.
   Incluye: OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport.
3. **NumServiciosDeEntretenimiento**
   Total de servicios de entretenimiento contratados.
   Variables: StreamingTV + StreamingMovies.

4. **AntigüedadAlta**
   Binaria: 1 si tenure > 24, 0 si no.
   Sirve para segmentar clientes nuevos de fidelizados.
5. **GastoPromedioMensual**
   Cálculo: TotalCharges / tenure (cuando tenure > 0), sino TotalCharges.
   Permite observar si el cliente ha tenido un comportamiento de gasto constante o variable.
6. **TieneDependientesYPartner**
   Binaria: 1 si Partner y Dependents son "Yes", 0 si no.
   Puede indicar mayor estabilidad o compromiso con el servicio.

2.4. Distribuciones de las variables:

Para conocer a qué tipo de datos nos estamos enfrentando, es necesario conocer las distribuciones de las variables.

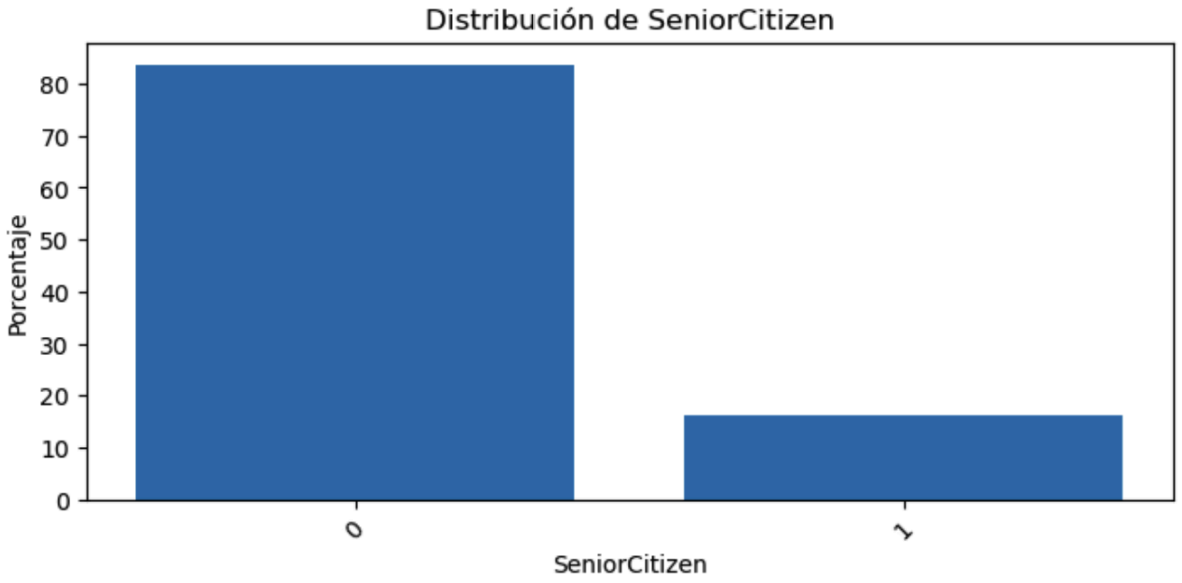
Variables numéricas:



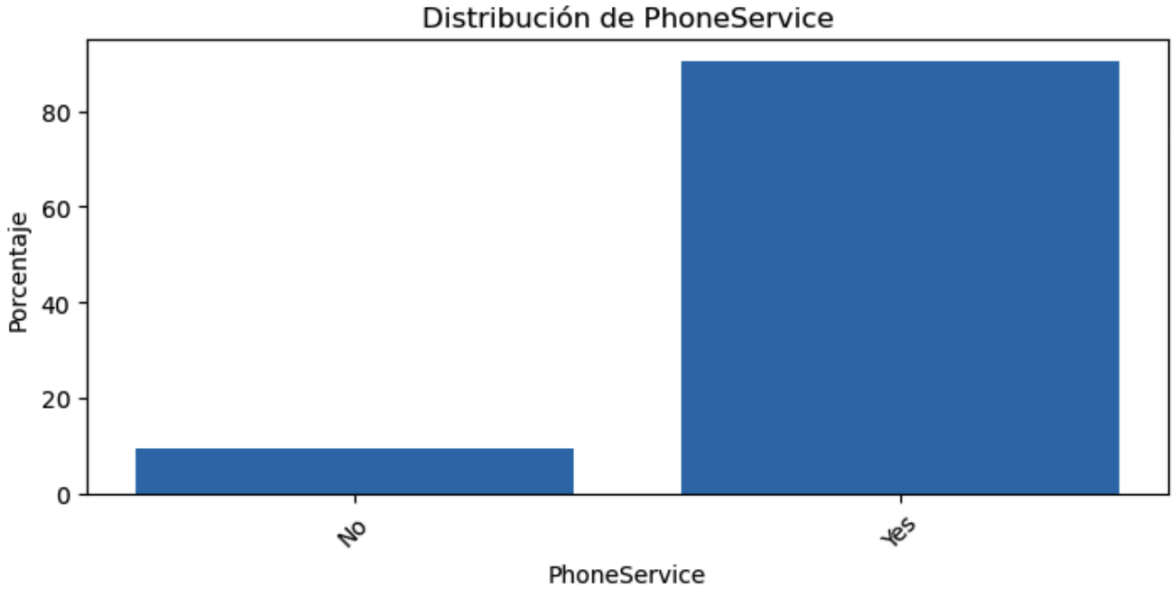
Podemos comprobar que ninguna de las distribuciones se ajusta a una curva normal. Asimismo, a través de los diagramas de caja (box plots) no se identificaron valores atípicos según el criterio del rango intercuartil.

Variables categóricas:

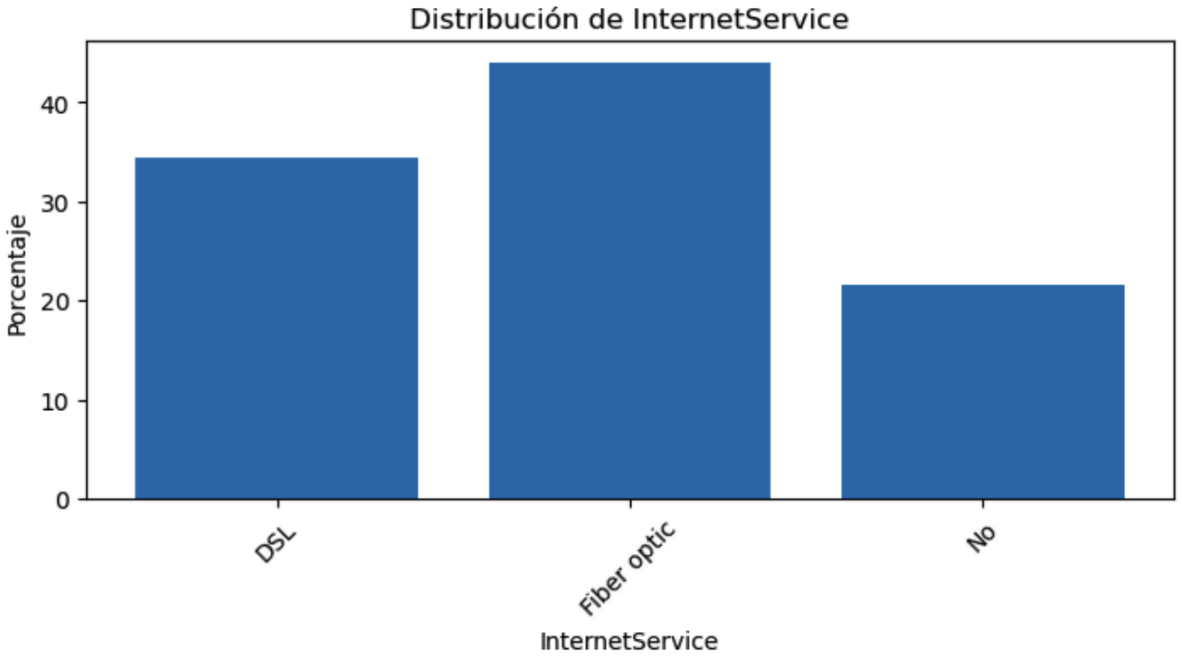
Destacan las siguientes distribuciones:



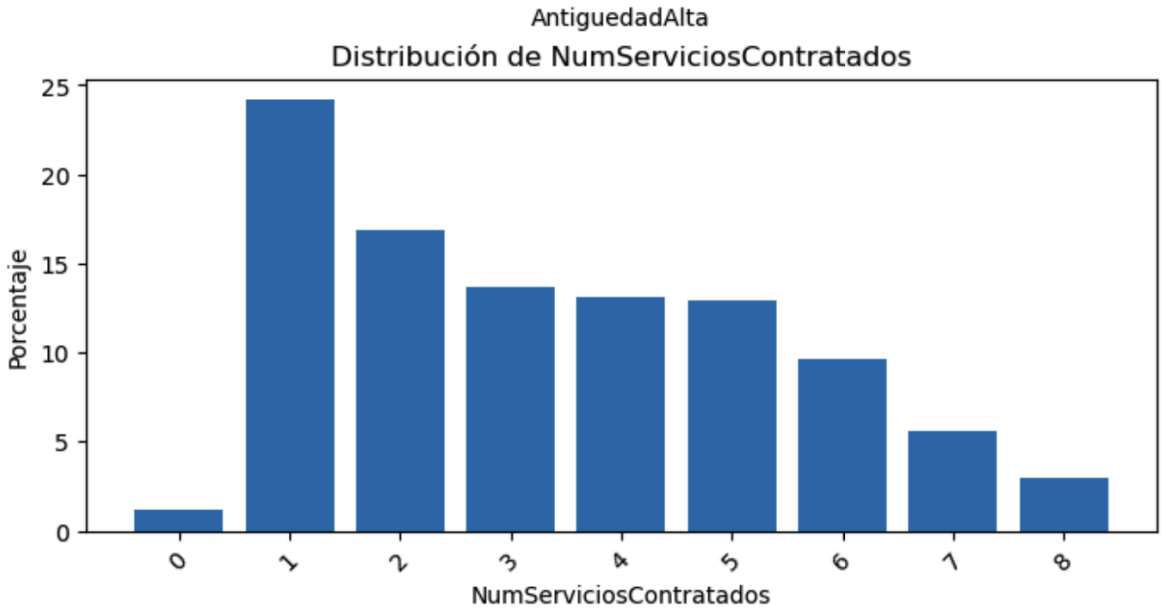
Como podemos observar, la mayoría de clientes no son senior.



Como era de esperar, al tratarse de una empresa de telecomunicaciones, la mayoría tiene servicio telefónico contratado.



Algo similar ocurre con los servicios de internet contratados.

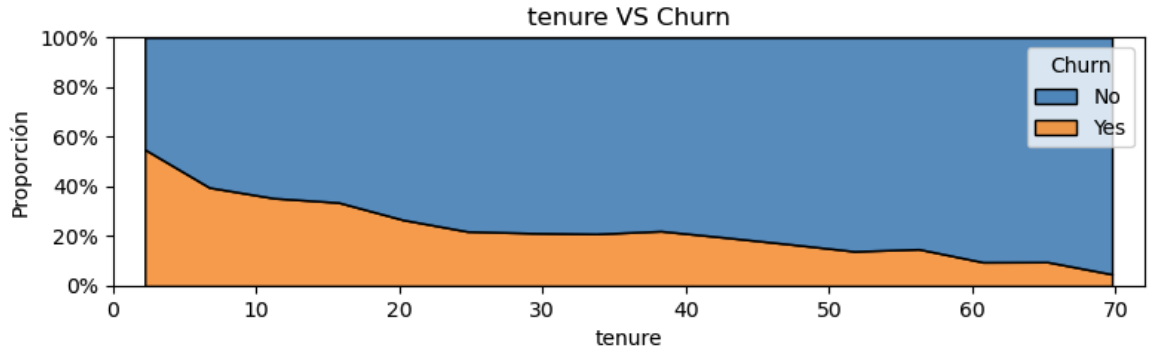


Como podemos observar, son muy pocos los clientes que no tienen servicios contratados. No obstante, a medida que aumenta el número de servicios, disminuye la cantidad de clientes que los poseen.

2.5. Relación entre las variables de entrada y de salida:

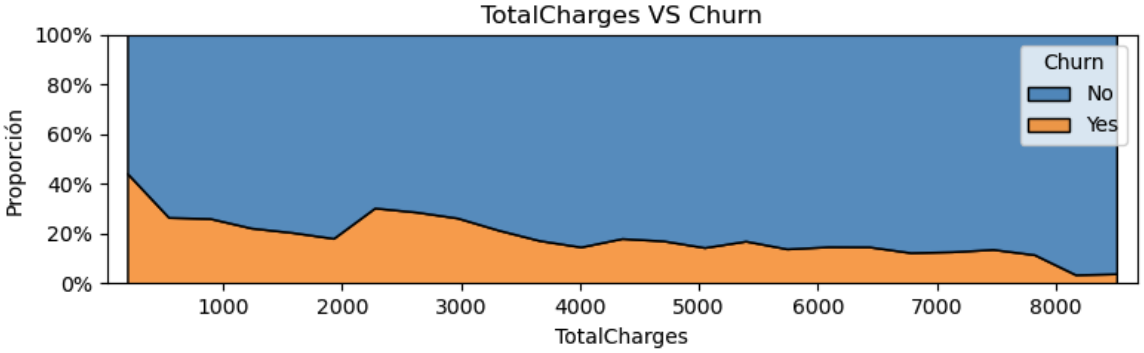
Para poder identificar posibles patrones entre las variables de entrada y de salida, se visualizan las relaciones entre ellas. Se identificaron los siguientes patrones:

Tenure vs Churn:



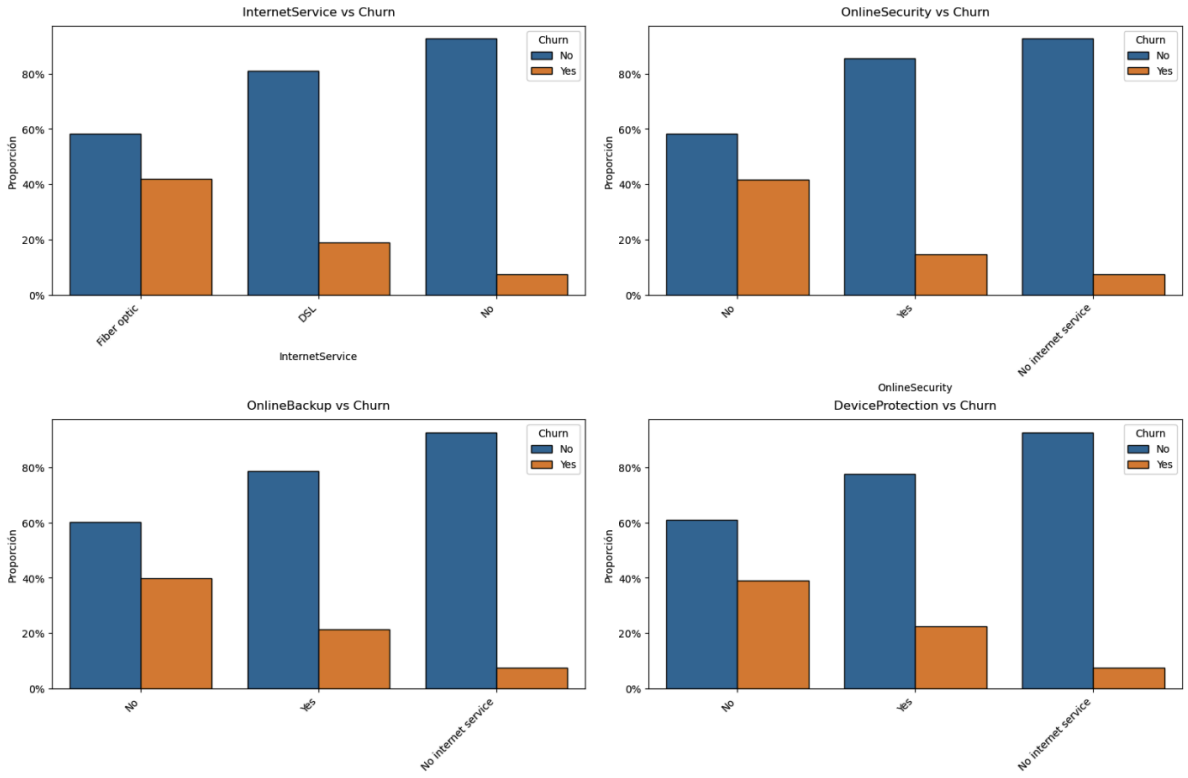
Se observa que los clientes con menor antigüedad en la empresa presentan una mayor proporción de abandono (churn). A medida que aumenta el tiempo que llevan como clientes, dicha proporción disminuye considerablemente. Esto sugiere que los clientes más recientes son más propensos a dejar la empresa, mientras que aquellos con una relación de largo plazo tienden a mostrar mayor lealtad.

MonthlyCharges vs Churn:



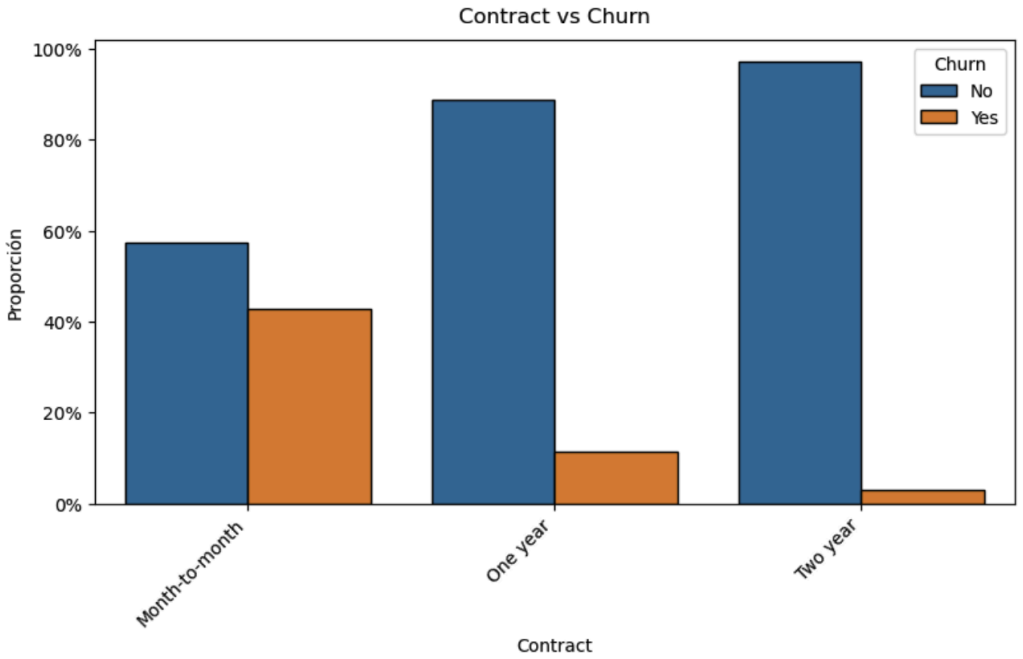
Existe una relación clara entre el monto mensual que paga un cliente y su probabilidad de abandono. Se observa una mayor proporción de churn entre aquellos clientes con cargos mensuales bajos. En contraste, los clientes que se encuentran en rangos altos y medios de pago muestran una menor tendencia al abandono. Esto sugiere que quienes pagan menos al mes son más propensos a dejar el servicio.

Servicios de internet vs Churn:



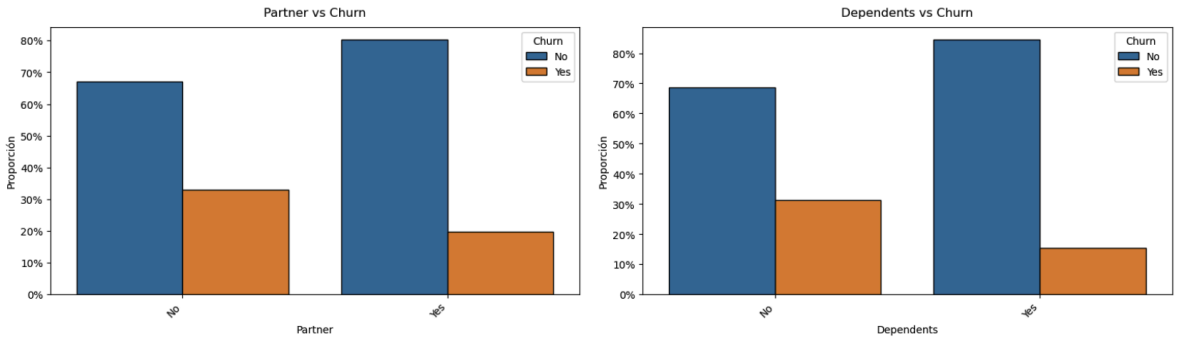
Se identifica un patrón claro entre los clientes con servicio de Internet: aquellos que no cuentan con servicios de seguridad tienden a abandonar con mayor frecuencia que quienes sí los tienen. Esto podría deberse a que los clientes sin estos servicios son más propensos a experimentar problemas relacionados con la seguridad o el funcionamiento del servicio, lo que puede generar insatisfacción y llevarlos a tomar la decisión de irse. En cambio, quienes cuentan con servicios de seguridad pueden sentirse más protegidos y satisfechos, lo que favorece su permanencia.

Contract vs Churn:



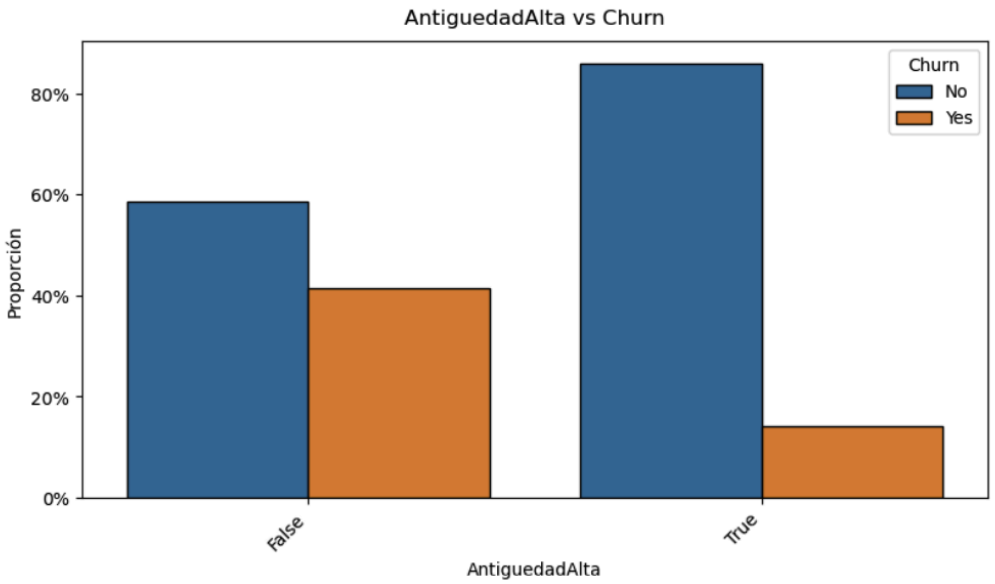
Esta visualización examina la relación entre el tipo de contrato (mensual, anual o bianual) y la probabilidad de abandono del servicio. Se observa que los clientes con contrato mensual presentan una proporción significativamente mayor de churn. En contraste, aquellos con contratos de uno o dos años muestran tasas de abandono considerablemente más bajas, lo que sugiere que los compromisos a largo plazo contribuyen a una mayor retención de clientes.

Partner y Dependants vs Churn:



Se analiza cómo el estado civil (tener pareja) y la presencia de dependientes económicos influyen en la probabilidad de abandono del servicio. Se observa que los clientes que no tienen pareja ni dependientes presentan una mayor proporción de churn en comparación con aquellos que sí los tienen. Esto sugiere que factores personales como la estabilidad familiar o la responsabilidad económica pueden estar asociados a una mayor permanencia, posiblemente debido a decisiones compartidas o una mayor necesidad de continuidad en los servicios contratados.

AntigüedaAlta vs Churn:



Se puede observar que los clientes con más de dos años de antigüedad tienden a ser más fieles a la empresa. Esto podría deberse a que, con el tiempo, han desarrollado mayor

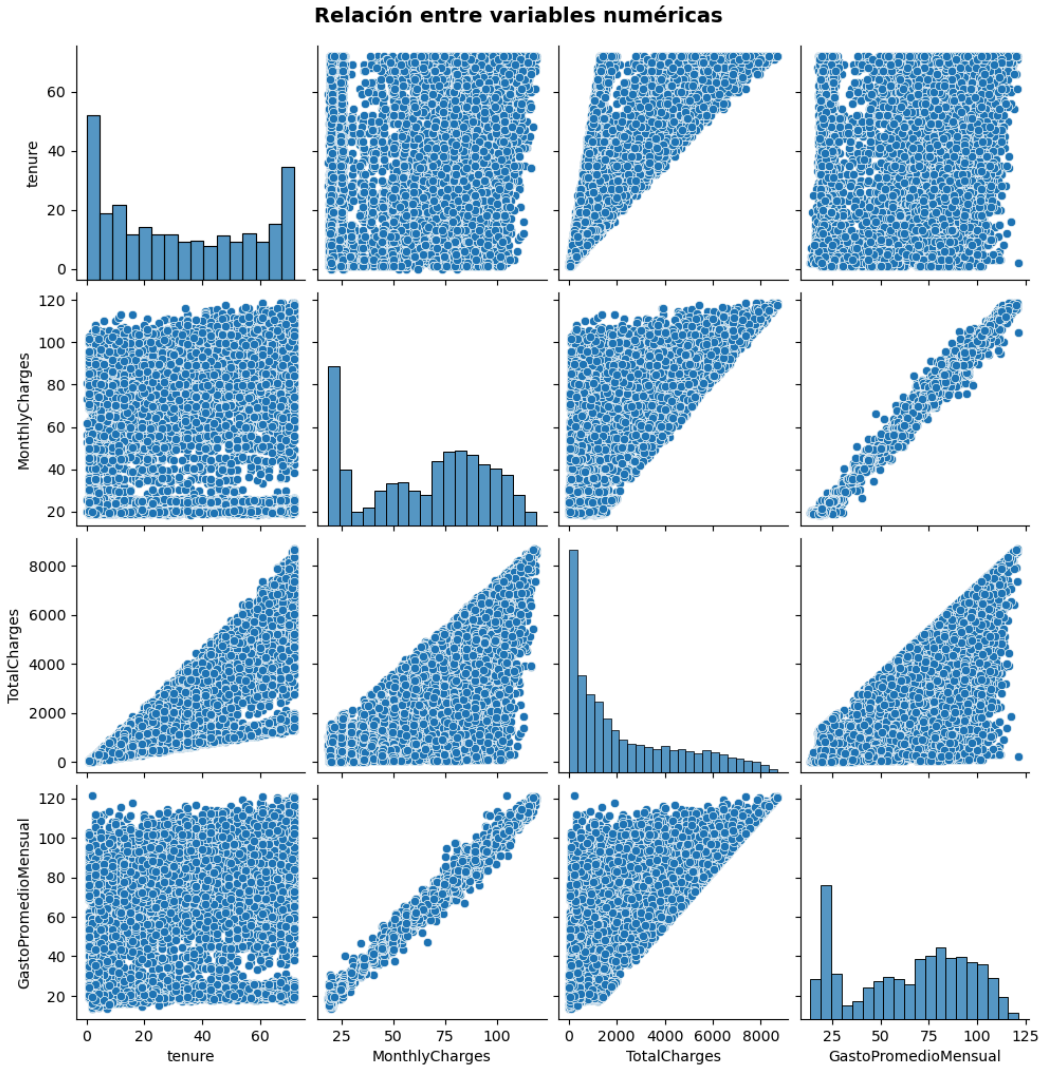
confianza en el servicio y se sienten satisfechos con la experiencia ofrecida, lo que reduce su probabilidad de abandono.

Conclusión:

En conjunto, los patrones identificados permiten perfilar a los clientes con mayor riesgo de abandono. Los factores que más contribuyen al churn se relacionan con la falta de compromiso a largo plazo, niveles bajos de interacción o valor percibido, y menor estabilidad personal. Específicamente, los clientes con contratos mensuales, sin servicios complementarios como seguridad, con menores pagos mensuales y sin vínculos personales como pareja o dependientes, conforman los segmentos más vulnerables. Esta información resulta clave para diseñar estrategias de retención más efectivas, enfocadas en ofrecer mayor valor, generar confianza desde las primeras etapas del ciclo del cliente y promover compromisos más duraderos.suscribirse a servicios adicionales.

2.6. Interacción entre las variables numéricas:

Para tratar de entender cómo interactúan las variables numéricas entre sí, visualizaremos una matriz de pares que las enfrente una a una.

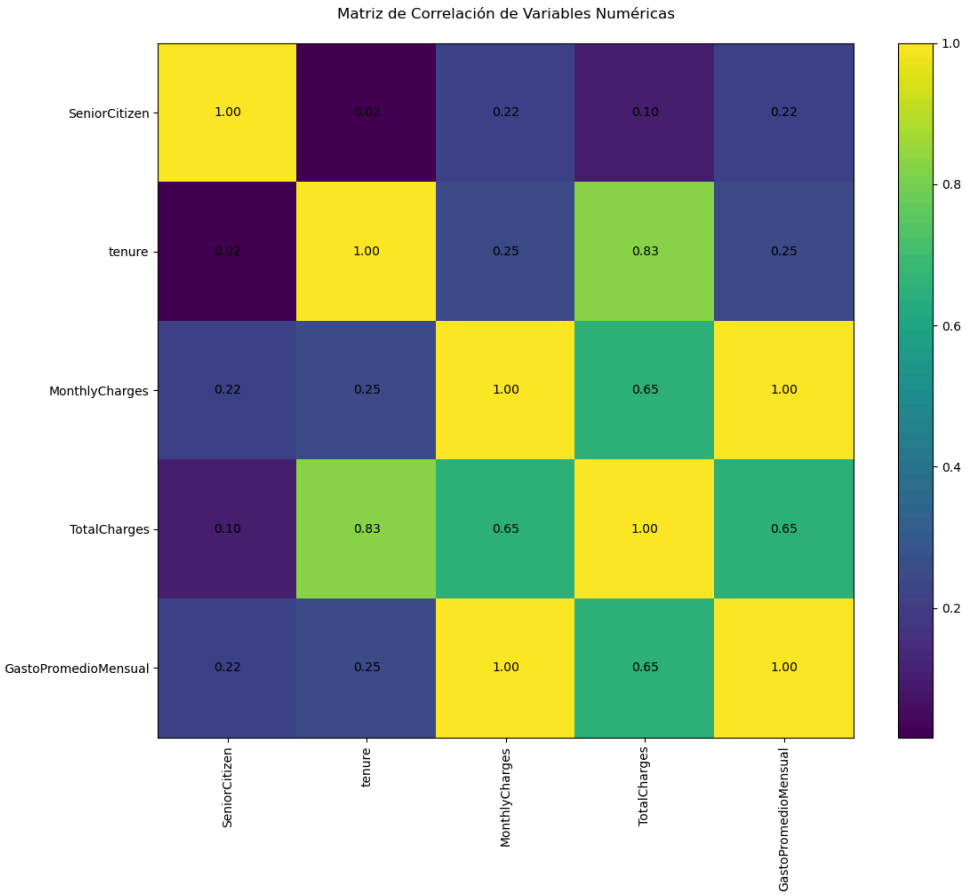


No se observa una correlación significativa entre la antigüedad del cliente (tenure) y el monto mensual que paga (MonthlyCharges), lo que indica que el tiempo como cliente no predice directamente el pago mensual. En contraste, existe una relación lineal positiva clara entre la antigüedad y el importe total facturado (TotalCharges), ya que a mayor tiempo con la empresa, mayor es el gasto acumulado.

Asimismo, el total facturado muestra una correlación positiva con el cargo mensual, aunque con mayor dispersión, dado que depende tanto del pago mensual como del tiempo de permanencia. El gasto promedio mensual (GastoPromedioMensual) presenta una correlación positiva con el total facturado, reflejando que un mayor gasto medio se traduce en un mayor total acumulado.

El gasto promedio mensual (GastoPromedioMensual) presenta una fuerte correlación positiva con la factura mensual (MonthlyCharges), reflejando que un mayor gasto medio se traduce en un mayor monto mensual. Esto tiene sentido, ya hemos creado la variable GastoPromedioMensual a partir de MonthlyCharges.

Por otro lado, la antigüedad no influye de manera clara sobre el gasto promedio mensual, ya que la correlación es prácticamente nula y la dispersión alta. Finalmente, se evidencia una correlación positiva entre el cargo mensual y el gasto promedio, lo que indica que a mayores cargos mensuales corresponde un mayor gasto medio.



A través de la matriz de correlación, podemos confirmar y visualizar las relaciones observadas entre las variables numéricas.

3. Modelo base:

Para comenzar a analizar el efecto de las distintas técnicas de minería de datos en la mejora de las predicciones, es necesario establecer un modelo de referencia sobre el cual aplicar las mejoras posteriores.

El modelo base será lo más sencillo posible, utilizando únicamente el preprocesamiento mínimo necesario para permitir que los algoritmos de aprendizaje automático funcionen correctamente. De esta manera, podremos medir de forma clara el impacto de cada técnica adicional aplicada en fases posteriores.

A continuación analizaremos cual es el preprocesamiento mínimo para que un algoritmo de aprendizaje automático funcione con nuestro conjunto de datos.

3.1. Tratamiento de valores faltantes:

Como se ha mencionado anteriormente, hemos encontrado valores nulos en la variable "TotalCharges". Para que nuestro algoritmo de aprendizaje se pueda ejecutar, necesitamos tratar de alguna manera estos valores.

Como primera solución, utilizaremos un imputador por la media de la variable.

3.2. Transformación de variables: categórica a numérica:

En nuestro problema tenemos algunas variables categóricas. Estas deben ser convertidas a numéricas, ya que más adelante, cuando necesitemos aplicar un modelo de aprendizaje automático éstas no podrán ser tratadas. Como primera solución, comenzaremos utilizando la codificación ordinaria por su sencillez.

3.3. Modelo de Aprendizaje Automático:

Como modelo base para las pruebas utilizaremos K-Nearest Neighbours con sus valores por defecto.

3.4. Rendimiento del Modelo Base:

Para poder medir el rendimiento base del modelo, se particionará el dataset en 4:

- Conjunto de entrenamiento
- Conjunto de validación 1: conjunto para seleccionar las técnicas de preprocesamiento en futuros pasos.
- Conjunto de validación 2: conjunto para seleccionar el umbral en futuros pasos.
- Conjunto de test: conjunto para medir el rendimiento del modelo final.

Dado que en este problema nuestro interés está en detectar la mayor cantidad de clientes que se van a ir sin tener muchos falsos positivos, utilizaremos el área bajo la curva precisión-recall (AUC-PR) como medida independiente del umbral.

Estos fueron los resultados el modelo base:

- AUC-PR (Entrenamiento): 71.39%
- AUC-PR (Validación 1): 52.58%

El descenso significativo en el AUC-PR al pasar del entrenamiento a la validación indica que el modelo presenta sobreajuste: aprende bien el patrón de los datos de entrenamiento, pero pierde precisión al enfrentarse a nuevos datos, lo que compromete su utilidad práctica.

Matriz de Confusión – Entrenamiento:

	No Churn (0)	Churn (1)
Predicho 0	92.7%	41.3%
Predicho 1	7.3%	58.7%

Matriz de Confusión – Validación 1:

	No Churn (0)	Churn (1)
Predicho 0	88.2%	58.8%
Predicho 1	11.8%	41.2%

Podemos ver que la mayor dificultad que está teniendo el modelo es la identificación de los clientes que van a abandonar.

El modelo identifica correctamente a los clientes que no abandonan con alta precisión (92.7% en entrenamiento, 88.2% en validación).

Sin embargo, tiene un desempeño más débil al identificar a los clientes que sí abandonan:

- 41.2% de acierto en validación significa que más de la mitad de los clientes que efectivamente se van no son detectados por el modelo.
- Esto es preocupante, ya que el objetivo principal del modelo es justamente anticipar el churn para actuar a tiempo.

Conclusiones:

- Rendimiento desigual: El modelo favorece la predicción de clientes que no abandonan, lo que podría estar relacionado con un desbalance de clases (más clientes permanecen que abandonan).
- Impacto en decisiones comerciales: Si el modelo no detecta con suficiente precisión a los clientes en riesgo de churn, se pierde la oportunidad de implementar estrategias de retención eficaces.

4. Mejoras en el preprocesamiento:

En esta sección, exploraremos cómo diversas técnicas de preprocesamiento pueden influir en el rendimiento del modelo base.

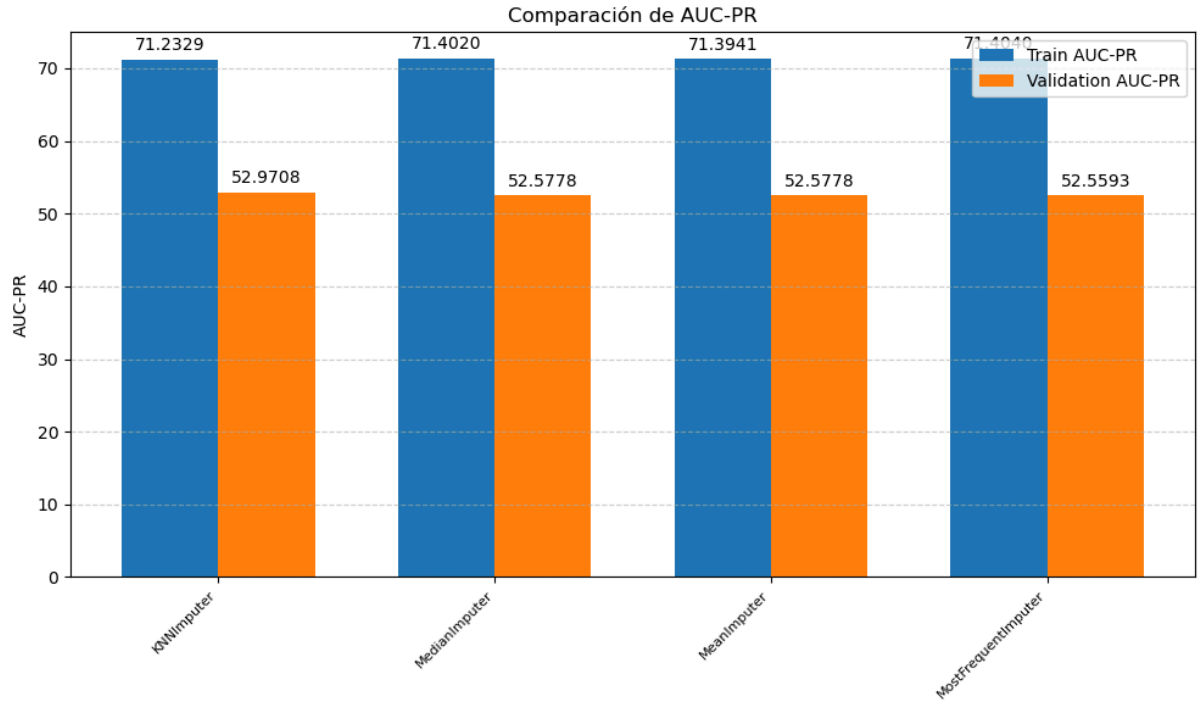
El objetivo principal es identificar las estrategias de preprocesamiento más efectivas que permitan maximizar el área bajo la curva de precisión-recall.

4.1. Imputación de valores faltantes:

En este apartado probaremos cómo afecta cada uno de los métodos de imputación de valores faltantes. Probaremos los siguientes métodos:

- Imputación por la media
- Imputación por la mediana
- Imputación por la moda
- Imputación por KNN

Estos fueron los resultados:



No se observa una gran diferencia entre el uso de distintos imputadores. Esto puede deberse a la baja cantidad de valores faltantes en el conjunto de datos, o a que la variable con valores faltantes no tiene un impacto significativo en la tarea de clasificación. Aun así, KNNImputer muestra un rendimiento algo superior, por ello, continuaremos con él.

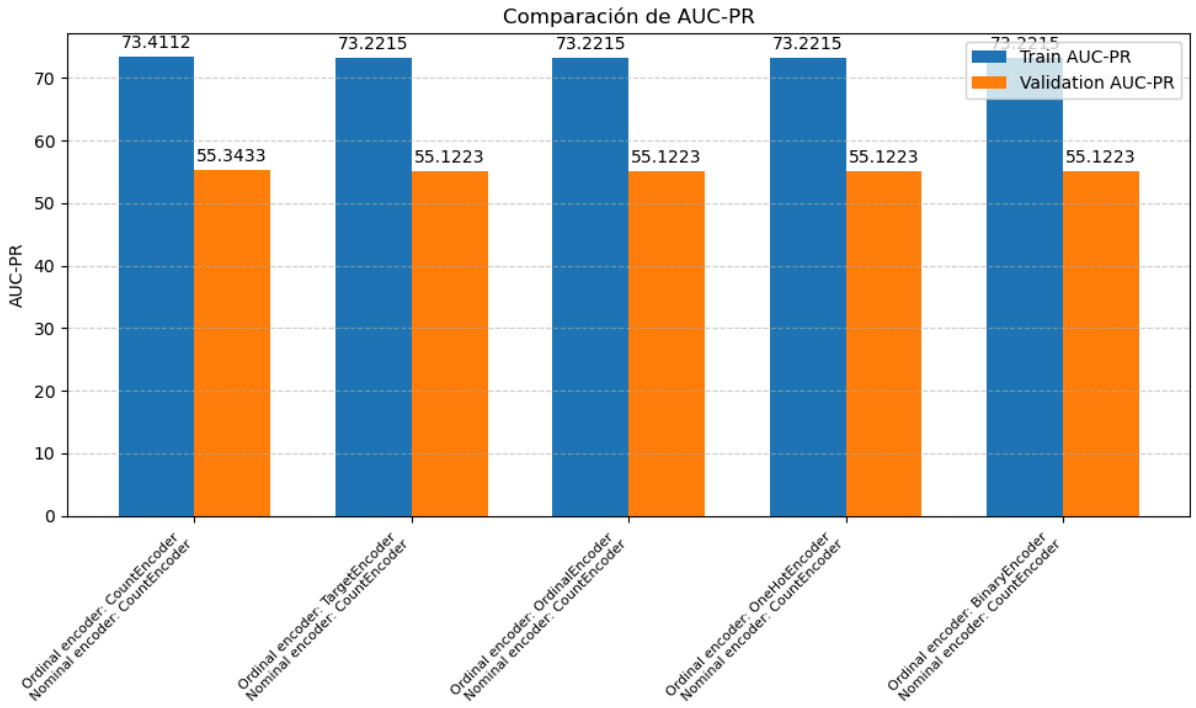
4.2. Codificación de las variables categóricas:

A continuación analizaremos el efecto de distintos métodos para codificar variables categóricas en variables numéricas. Probaremos los siguientes métodos:

- Codificación por conteo
- Codificación ordinal
- Codificación binaria
- Codificación One Hot Encoding
- Codificación basada en la salida del problema de clasificación

Hay que tener en cuenta que dentro de las variables categóricas tenemos dos clases, las ordinales y las nominales. Por tanto, utilizaremos distintas codificaciones para cada una de estas.

Los resultados fueron los siguientes:



Podemos observar que el codificador que ofrece mejores resultados, tanto para las variables categóricas **ordinales** como **nominales**, es el **CountEncoder**. Este comportamiento es coherente con el uso de un modelo **K-Nearest Neighbors (KNN)**, ya que este tipo de codificación facilita la distinción entre valores frecuentes y poco frecuentes. Esto sugiere que el hecho de que una categoría sea habitual tiene un peso importante en el proceso de clasificación.

En concreto, al codificar las categorías según su frecuencia de aparición en el conjunto de datos, la diferencia numérica entre una categoría común y una rara se vuelve significativa. Esto ayuda al algoritmo de vecinos más cercanos a capturar mejor la relación entre los

ejemplos, ya que las distancias reflejan, de forma implícita, la representatividad de cada valor.

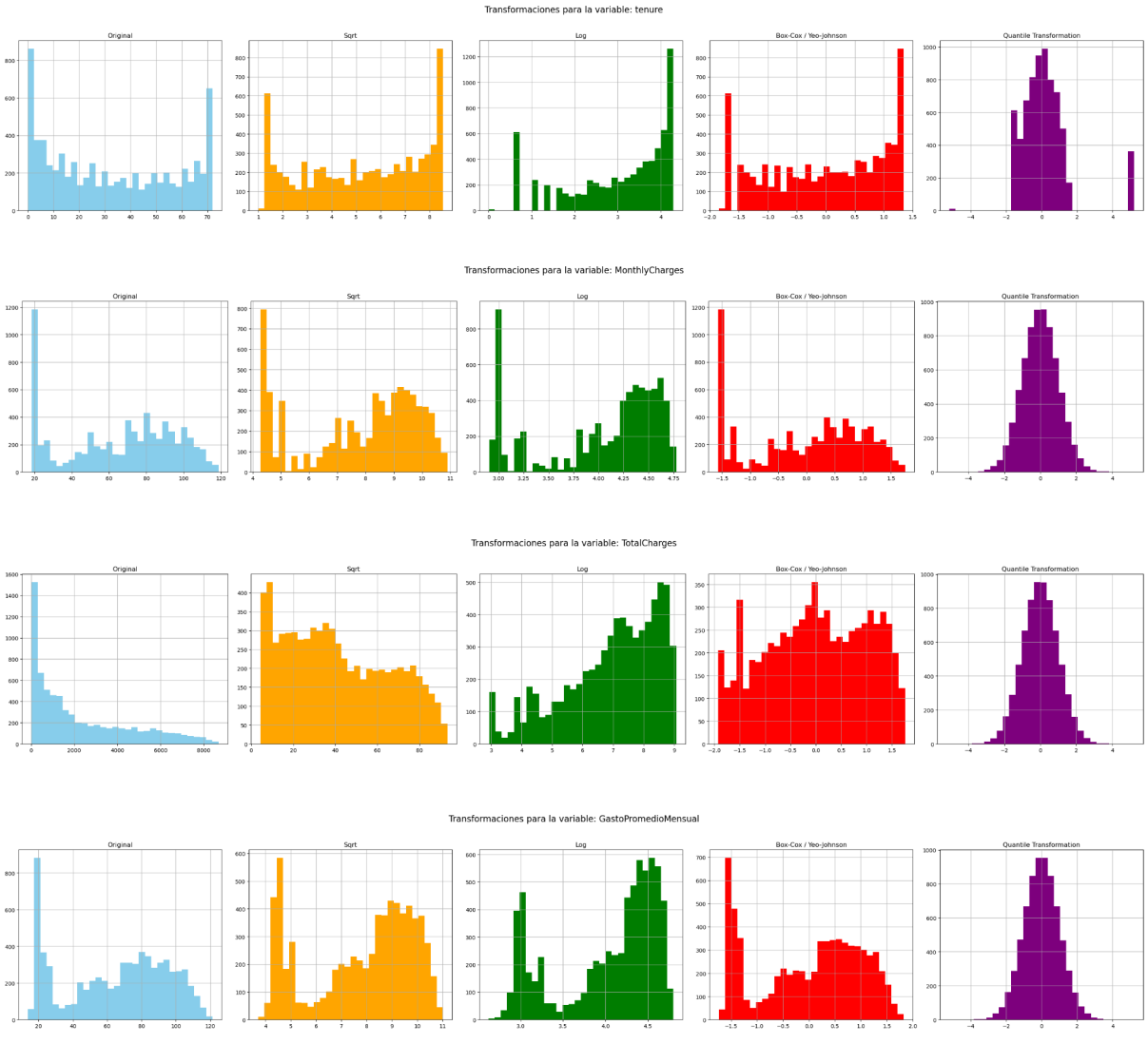
Además de facilitar esta distinción, CountEncoder presenta otra ventaja importante: **no incrementa la dimensionalidad del dataset**. A diferencia de otros métodos como OneHotEncoder, no genera columnas adicionales, lo que implica un menor **coste computacional** y una mayor eficiencia, especialmente en conjuntos de datos con muchas categorías.

4.3. Transformación de variables:

Durante el análisis exploratorio se observó que las variables numéricas no siguen una distribución normal. Por ello, aplicaremos diversas transformaciones con el objetivo de aproximarlas a una distribución normal. Las transformaciones consideradas son:

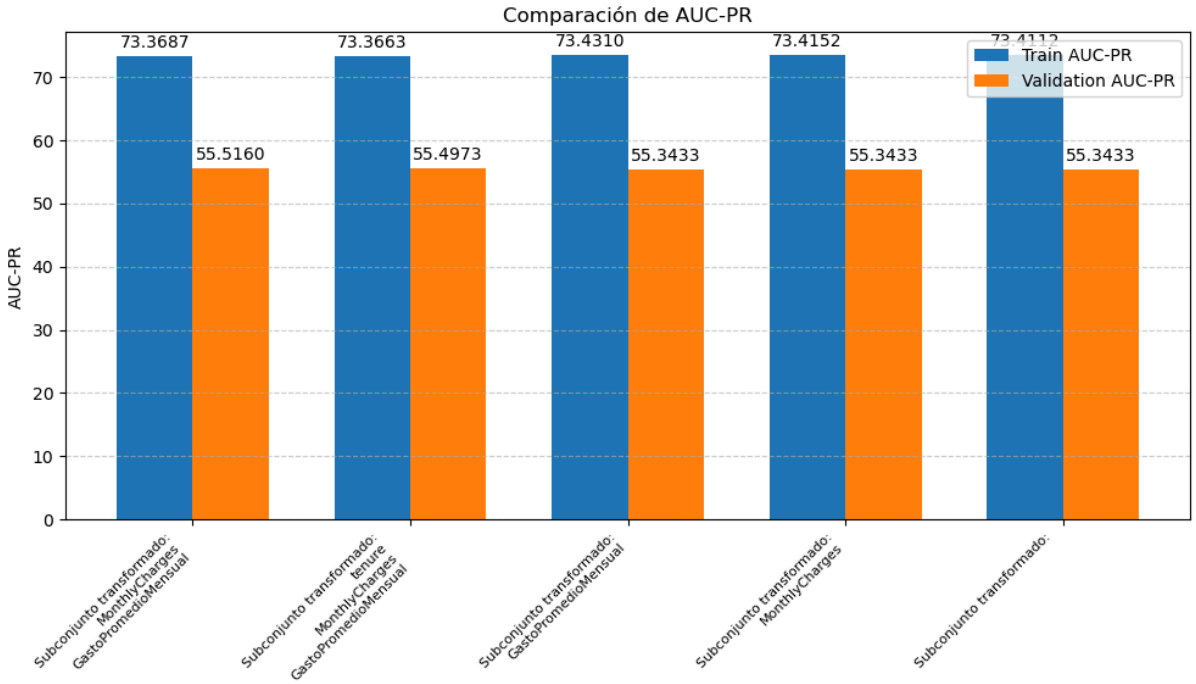
- Transformación logarítmica
- Transformación de raíz cuadrada
- Transformación Box-Cox o Yeo-Johnson (según el dominio de la variable)
- Transformación basada en cuantiles (Quantile Transformation)

Evaluaremos visualmente el efecto de cada una de estas transformaciones sobre la distribución de las variables numéricas:



Como se puede observar, la transformación basada en cuantiles es la que mejor aproxima las variables numéricas a una distribución normal. Para evaluar su impacto en el rendimiento del modelo, probaremos distintas combinaciones en el preprocesamiento. En concreto, analizaremos si aplicar o no esta transformación de forma individual para cada variable mejora la capacidad de clasificación.

Estos fueron los resultados:



Dado que transformar las variables MonthlyCharges y GastoPromedioMensual contribuye a una ligera mejora en el AUC-PR, decidimos mantener dicha transformación en el preprocesamiento. Esta mejora probablemente se deba a que el modelo KNN tiende a funcionar mejor cuando las variables numéricas siguen una distribución aproximadamente normal.

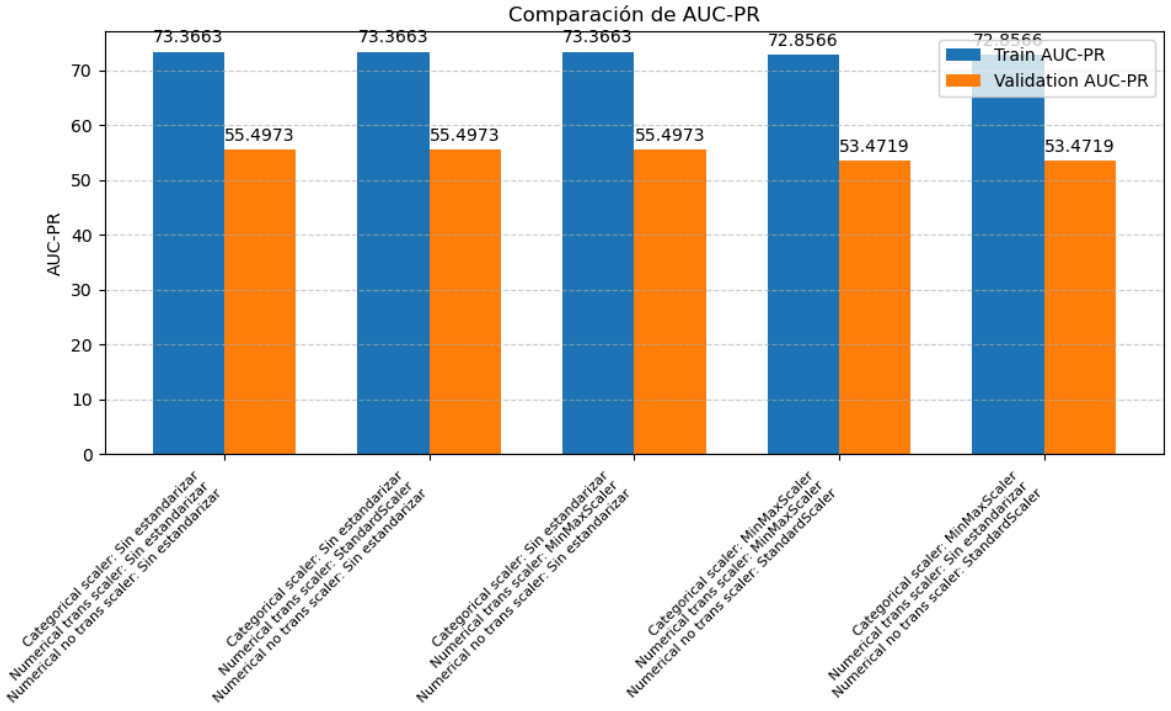
Además, trabajar con variables normalizadas facilita procesos posteriores como la detección de outliers o la aplicación de técnicas de escalado, mejorando así la estabilidad y el rendimiento general del modelo.

4.4. Normalización de las variables:

En este apartado, se visualizará el efecto de la normalización sobre las variables **numéricas** y **categóricas**. Para las variables numéricas, se probarán los siguientes métodos de normalización:

- **Estandarización mediante la media y la desviación estándar (Z-score)**.
- **Estandarización mediante el mínimo y el máximo (min-max)**.

En cuanto a las variables categóricas, se aplicará únicamente la **normalización min-max** a las variables categóricas ordinales, dado que el uso del Z-score podría distorsionar los patrones de **rareza** que resultan cruciales para modelos como **KNN**.



El mejor rendimiento en validación (55.49%) se obtiene sin aplicar escalamiento, lo cual es contraintuitivo en modelos basados en distancia como KNN, donde normalmente se espera que el escalamiento mejore el rendimiento.

Tanto el escalado con StandardScaler como con MinMaxScaler reducen el AUC-PR de validación.

Posible explicación:

- Una o más variables numéricas podrían tener una escala naturalmente dominante y, por tanto, están aportando una señal fuerte al modelo. Al aplicar escalamiento, esa influencia se neutraliza y pierde relevancia, lo que disminuye la capacidad predictiva del modelo.
- Esto es especialmente importante en KNN, ya que se basa en cálculos de distancia: reducir la escala de variables informativas perjudica la calidad de las predicciones.

Conclusiones:

- Mejor estrategia: Para este conjunto de datos, el modelo KNN funciona mejor sin normalizar las variables numéricas, lo que indica que hay variables con escalas significativas y gran poder predictivo.

- Normalizar no siempre ayuda: Aunque escalar datos suele ser una buena práctica, en este caso reduce la capacidad del modelo de detectar churn, evidenciado por la caída en AUC-PR de validación.

Por ello, decidimos no incluir el escalado en el preprocesamiento.

4.5. Detección de Outliers:

En este apartado, se analizará el efecto de la **detección de outliers** en las variables numéricas. Se separán las variables numéricas en 2 categorías: las variables transformadas en el paso anterior y las no transformadas. Para cada una de estas categorías, se probarán distintos detectores. En concreto, se probarán los siguientes métodos de detección:

- Sin tratamiento de outliers (Passthrough)
- Detección y tratamiento de outliers mediante:
    - Media ± Desviación Estándar (Mean ± Std)
- Rango Intercuartílico (IQR)

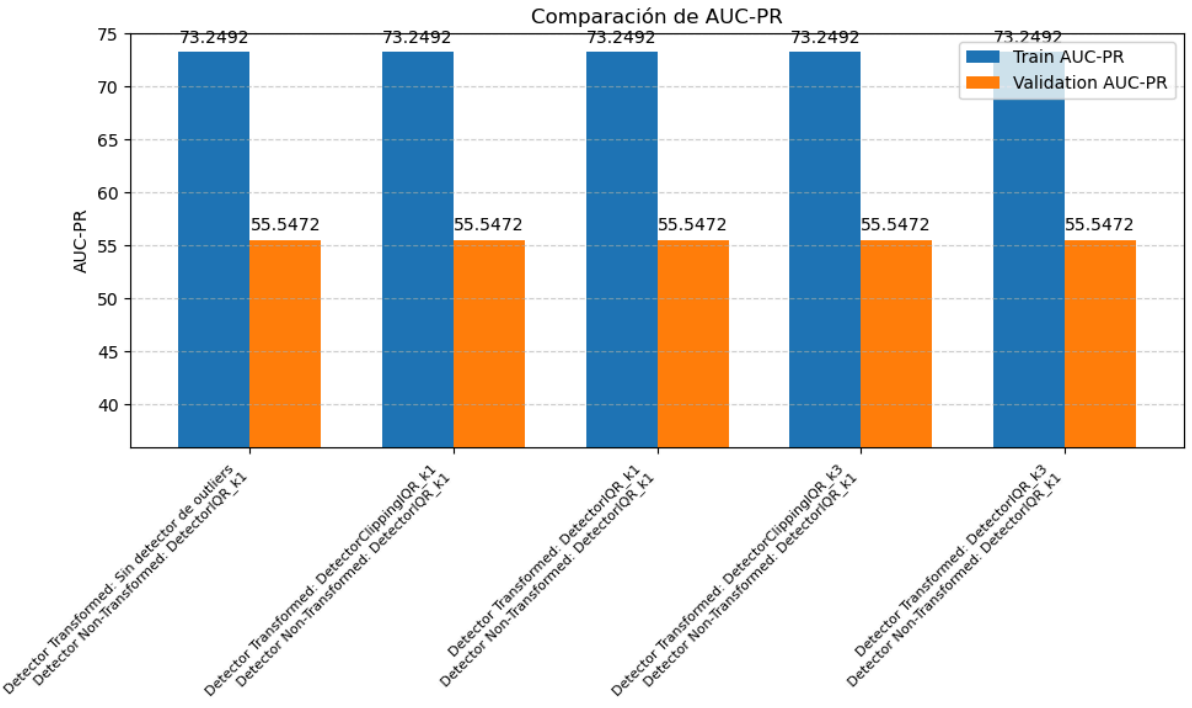
Cada uno de estos métodos se evaluará utilizando valores de *k* iguales a 1, 2 y 3.

Además, para cada tipo de detector de outliers, se probarán dos estrategias de tratamiento:

- Sustituir los valores atípicos por la media o la mediana, según corresponda.
- Realizar un recorte (*clipping*) de los valores atípicos a los límites máximos y mínimos definidos por el método seleccionado.

Las estrategias de clipping buscan minimizar el impacto de los valores extremos y mejorar la calidad del conjunto de datos para posteriores análisis y modelados.

Estos fueron los resultados obtenidos:



Interpretación:

- El tratamiento de outliers IQR con k=1 sobre las variables transformadas genera una pequeña mejora en el rendimiento de validación (sube de 55.49% a 55.54%).
- En general las diferencias son menores (±0.1), lo que indica que los valores atípicos no son tan extremos ni numerosos como para afectar en gran cantidad el rendimiento.

Conclusiones:

- Impacto limitado del tratamiento de outliers: Las diferencias de rendimiento en AUC-PR son menores (±0.1), lo que indica que el conjunto de datos no está significativamente afectado por valores extremos.
- Ligera ventaja para Mean±Std: Esta técnica muestra el mayor AUC-PR en validación, aunque la mejora es mínima.

4.6. Selección de variables:

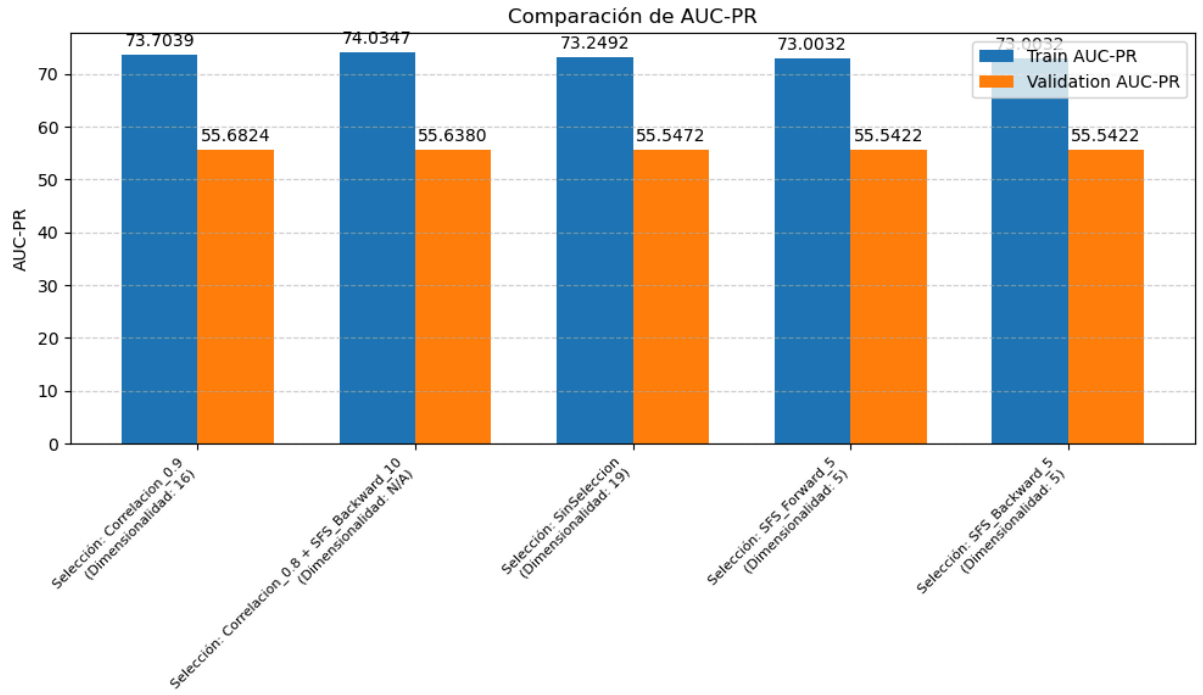
Cuando se tienen muchos ejemplos de alta dimensionalidad, aplicar técnicas de minería de datos y entrenar modelos de aprendizaje resulta muy costoso. Para reducir el coste computacional, trataremos de identificar las variables más importantes en nuestro problema.

Para ello vamos a utilizar distintas técnicas:

- Filtro de correlación
- Filtro ANOVA
- Sequential Forward Selection (SFS)
- Sequential Backward Elimination (SBE)
- Selección de variables mediante algoritmos genéticos
- Filtro de correlación + Wrapper
- Filtro ANOVA + Wrapper

Con estos métodos, nuestro objetivo es obtener el mayor rendimiento posible con el menor coste computacional posible.

Estos fueron los resultados obtenidos:



Interpretación:

- El filtro basado en la correlación con un umbral de 0.9 no solo consigue reducir el coste computacional al eliminar variables, sino que además aumenta el rendimiento
- Esto sugiere que antes variables redundantes podrían estar confundiendo al clasificador, y por ello el rendimiento era peor.

Conclusión:

- Continuaremos utilizando el filtro de correlación con un umbral de 0.9.

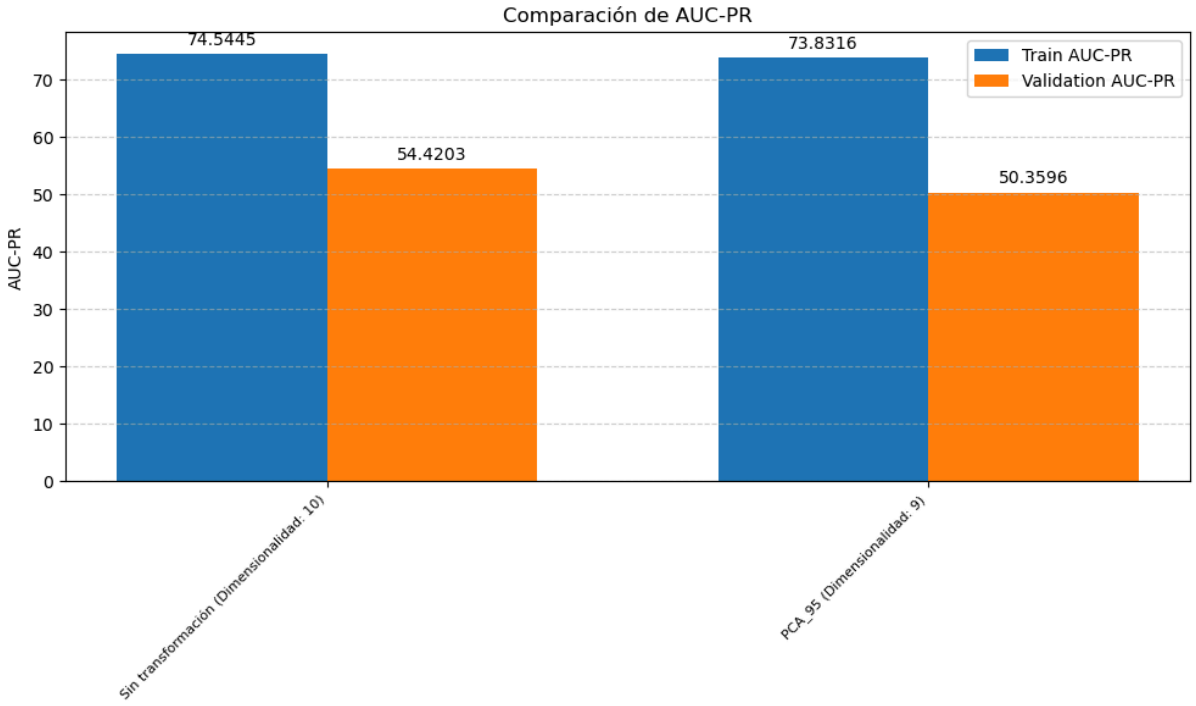
4.7. Análisis de Componentes Principales:

Otra técnica utilizada para reducir la dimensionalidad es el Análisis de Componentes Principales (PCA). Esta técnica estadística es utilizada para aumentar la separabilidad de los datos.

Para probar que tal se adapta esta técnica a nuestro problema, crearemos una nueva pipeline con el siguiente flujo:

- Preprocesamiento de los datos
- Selección de variables
- Normalización de los datos, para que la media sea 0 y la desviación estándar 1 (necesario para aplicar PCA)
- Aplicamos la transformación PCA
- Entrenamiento del clasificador base

Estos fueron los resultados obtenidos:



Interpretación:

- El modelo sin reducción de dimensionalidad obtiene mejor desempeño en validación, lo cual sugiere que la reducción con PCA puede estar eliminando características útiles para la predicción.

Conclusión:

- Dado que utilizar PCA no mejora los resultados, continuaremos sin usarlo.

4.8. Eliminación del ruido:

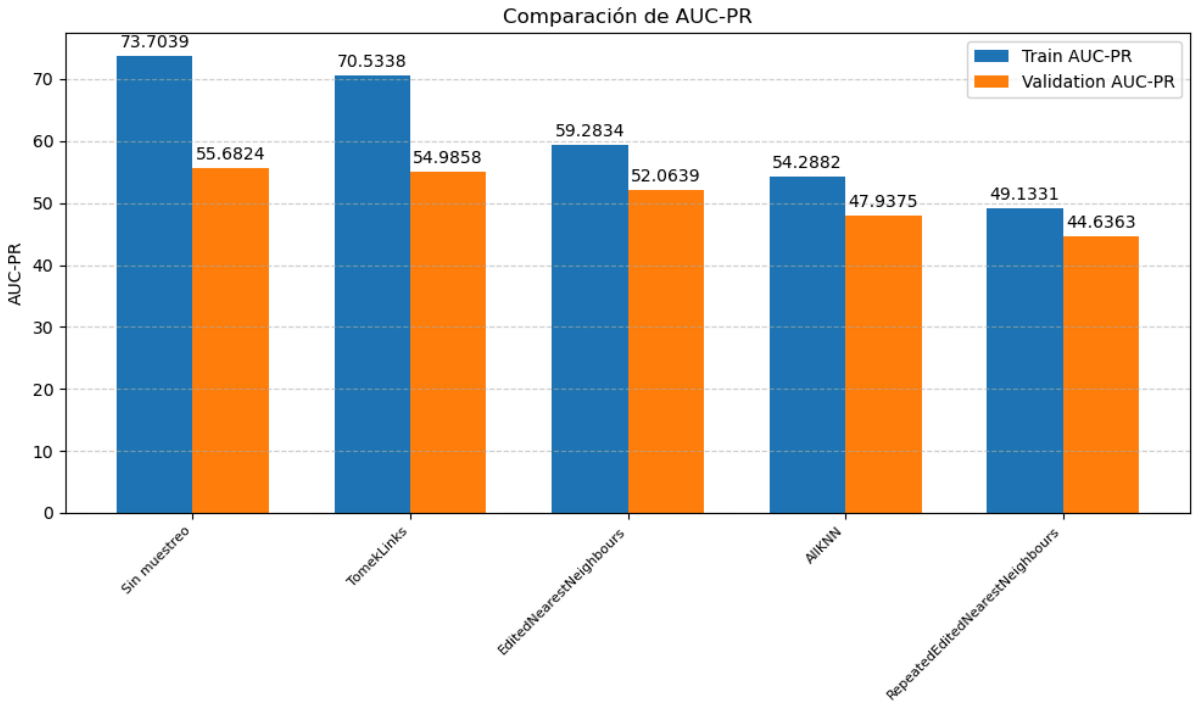
El ruido representa un desafío importante en los problemas de clasificación, ya que puede confundir al modelo durante el proceso de aprendizaje y dificultar la correcta predicción de ejemplos pertenecientes a distintas clases.

Para mitigar este problema, se evaluará el efecto de diferentes técnicas de eliminación de ruido, concretamente:

- Edited Nearest Neighbours (ENN)
- Repeated Edited Nearest Neighbours (RENN)
- All KNN
- Tomek Links

Estas técnicas buscan limpiar el conjunto de datos eliminando o corrigiendo instancias conflictivas, con el objetivo de mejorar el rendimiento y la generalización de los modelos de clasificación.

Estos fueron los resultados obtenidos:



Dado que la eliminación de ruido no ha demostrado mejorar el rendimiento, se ha decidido no utilizar estas técnicas en el flujo de trabajo.

La falta de mejora podría estar relacionada con la propia naturaleza del conjunto de datos. Por ejemplo, en datasets donde las clases están altamente entremezcladas y no existe una separación clara entre ellas, resulta muy difícil identificar y eliminar instancias ruidosas de manera efectiva.

4.9. Desbalanceo del problema:

Uno de los mayores problemas presentes en nuestro dataset es el desbalanceo. Para tratar de balancearlo, aplicaremos distintas técnicas de muestreo e hibridaciones:

Muestreo:

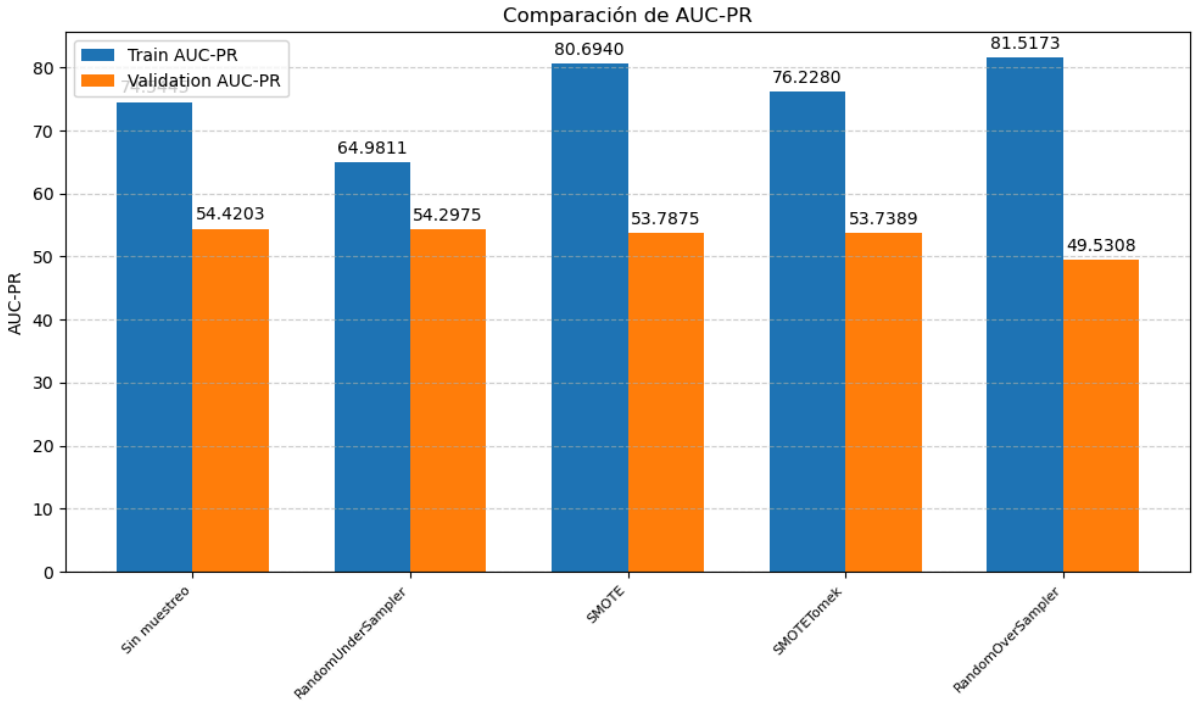
- Random Under Sampling
- Random Over Sampling
- SMOTE

Hibridaciones:

- SMOTE + ENN

- SMOTE + Tomek

Estos fueron los resultados obtenidos:



Interpretación:

- RandomUnderSampler reduce mucho el AUC-PR en entrenamiento, pero mantiene la validación casi igual. Es decir, reduce el overfitting sin pérdida notable en validación.
- SMOTE y SMOTETomek aumentan el AUC-PR en entrenamiento, pero no mejoran en validación, lo cual sugiere overfitting a los datos sintéticos.
- RandomOverSampler muestra el mayor AUC-PR en entrenamiento, pero el peor en validación, indicando un sobreajuste significativo.

Conclusión:

- Por la misma razón que motivó la falta de mejora al aplicar técnicas de eliminación de ruido, el balanceo del dataset tampoco ha producido una mejora en el rendimiento.
- Por tanto, no se aplicará ninguna técnica de remuestreo ni métodos híbridos en el flujo de trabajo.

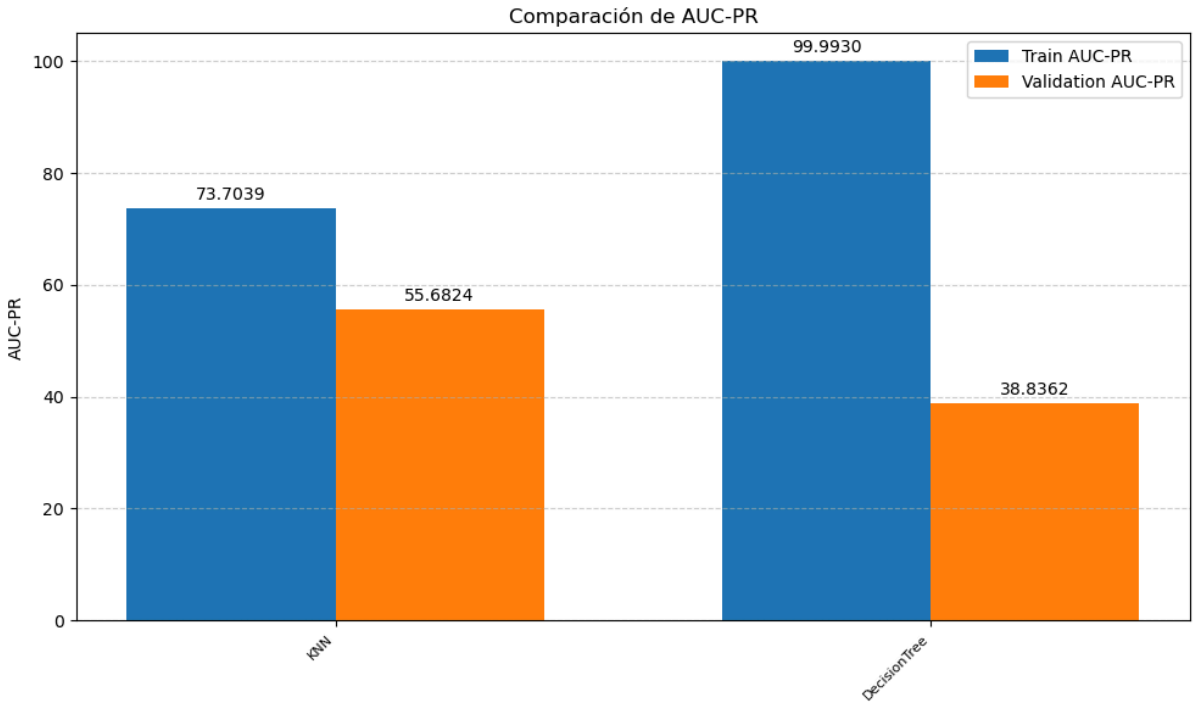
5. Elección del modelo:

En este apartado se compararán dos modelos de aprendizaje automático: **K-Nearest Neighbors (KNN)** y **árbol de decisión (DecisionTreeClassifier)**, con el objetivo de determinar cuál se ajusta mejor al problema de predicción de deserción de clientes (*churn*).

El flujo del preprocesamiento de datos es el decidido en los apartados anteriores:

- **Generación de variables adicionales** mediante ingeniería de características personalizada.
- **Codificación de variables categóricas** con CountEncoder.
- **Tratamiento de valores faltantes** en variables numéricas con imputación por la media.
- **Transformación de variables numéricas** mediante QuantileTransformer, para una mejor distribución.
- **Detección y tratamiento de outliers** en variables no transformadas, empleando el método del rango intercuartíl (IQR).
- **Eliminación de variables altamente correlacionadas**, aplicando un umbral de correlación del 0.9.
- **Evaluación del rendimiento** con la métrica **AUC-PR**.

Los resultados obtenidos fueron los siguientes:



Como se observa en la gráfica, el árbol de decisión alcanza una AUC-PR de 96.2% en entrenamiento, pero su rendimiento cae drásticamente en validación (38.8%), lo que indica un claro sobreajuste. En cambio, el modelo KNN muestra un rendimiento más equilibrado,

con una AUC-PR de 53.6% en validación, lo que lo convierte en una opción más confiable para generalizar a nuevos datos.

Conclusión:

A pesar de que el árbol de decisión ha mostrado un rendimiento superior en entrenamiento, su alto grado de sobreajuste lo descalifica como modelo robusto para este caso. Por ello, se elige avanzar con el modelo KNN, que ha demostrado una mejor capacidad de generalización y un balance razonable entre complejidad y rendimiento predictivo.

Es importante destacar que el preprocesamiento ha sido optimizado específicamente para KNN, lo que también ha podido favorecer su mejor comportamiento.

6. Ajuste de Hiper Parámetros del Modelo:

En esta sección se llevará a cabo el ajuste del modelo K-Nearest Neighbors (KNN) utilizando una búsqueda aleatoria de hiper parámetros de 100 iteraciones, con el objetivo de optimizar su rendimiento.

El ajuste se realizará sobre la secuencia de pasos empleada en el apartado anterior, pero esta vez con el clasificador K-Nearest Neighbors. Se evaluarán de forma conjunta los siguientes parámetros:

- Número de vecinos considerados en la predicción (valores impares entre 3 y 60).
- Tipo de ponderación de los vecinos, ya sea uniforme o basada en la distancia.
- Métrica de distancia utilizada para calcular la cercanía entre puntos, pudiendo ser distancia euclidiana o Manhattan.
- Umbral de correlación para filtrar variables altamente correlacionadas, con 1000 valores posibles entre 0.6 y 0.9.

Para la validación, se utilizará un método de Hold Out estratificado, garantizando que cada combinación de parámetros sea evaluada de forma robusta. La métrica empleada para medir el desempeño será el área bajo la curva Precision-Recall. La validación se realizó con los juntos de entrenamiento y validación 1 unidos.

El mejor modelo obtenido será utilizado en las siguientes etapas del análisis.

Mejores hiperparámetros encontrados

Tras el proceso de ajuste del modelo K-Nearest Neighbors (KNN), se han identificado los siguientes hiper parámetros como óptimos:

- **Umbral de correlación para el filtrado de variables altamente correlacionadas**: 0.72
- **Tipo de ponderación de los vecinos**: uniforme
- **Número óptimo de vecinos considerados**: 39
- **Métrica de distancia utilizada**: Manhattan

Mejor desempeño obtenido

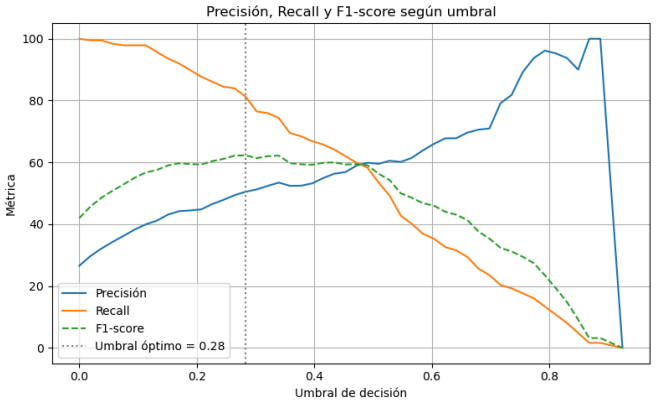
- **AUC-PR (validación hold-out estratificado)**: **65.01%**

7. Elección del Umbral:

Por defecto, los modelos de clasificación binaria predicen la clase positiva (en este caso, clientes que abandonan) cuando la probabilidad estimada supera un umbral de 0.5. Sin embargo, este valor no siempre es el óptimo, especialmente en contextos con clases desbalanceadas como el churn, donde es más importante detectar a los clientes que efectivamente van a abandonar, incluso a costa de tener más falsos positivos.

Objetivo:

Ajustar el umbral de decisión del modelo KNN entrenado, para encontrar el punto que entregue el mejor equilibrio entre precisión y recall, evaluando el F1-score como métrica resumen. Para ello, utilizamos el conjunto de entrenamiento de validación 2.



Interpretación:

- En el gráfico generado, se observa cómo varían la precisión, el recall y el F1-score al modificar el umbral de clasificación. La línea punteada vertical indica el umbral que maximiza el F1-score, correspondiente a un valor de 0.256. Con este umbral, se obtiene un F1-score del 62.48%.
- Este punto representa el mejor equilibrio entre captar la mayor cantidad posible de clientes que abandonan (alto recall), sin sacrificar demasiada precisión.

Conclusión:

- El uso del umbral ajustado de 0.256 en lugar del estándar 0.5 permite mejorar la capacidad del modelo para detectar churn, lo que es altamente deseable en contextos de negocio donde prevenir la pérdida de clientes es prioritario, incluso si implica contactar a algunos clientes que no iban a abandonar.
- El F1-score en los datos de test fue del 56.98%.

8. Interpretabilidad del modelo:

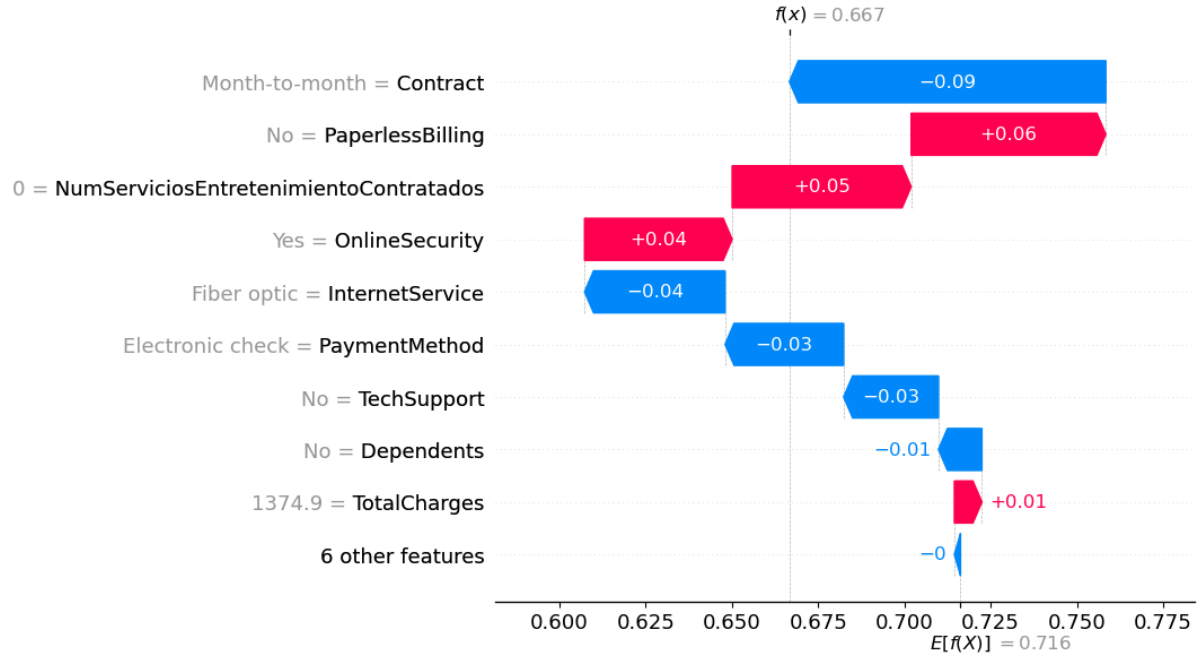
La **interpretabilidad** de un modelo es crucial para entender cómo cada variable influye en las decisiones de clasificación. Este entendimiento no solo ayuda a mejorar la transparencia del modelo, sino que también facilita la detección de posibles problemas, como el **Data Leakage**.

8.1. Interpretabilidad Local:

Una de las técnicas más efectivas para obtener interpretaciones a nivel local son los **SHAP values**. Con estos, podemos visualizar cómo las distintas variables de un ejemplo afectan a la probabilidad final de un modelo.

Cliente que se queda:

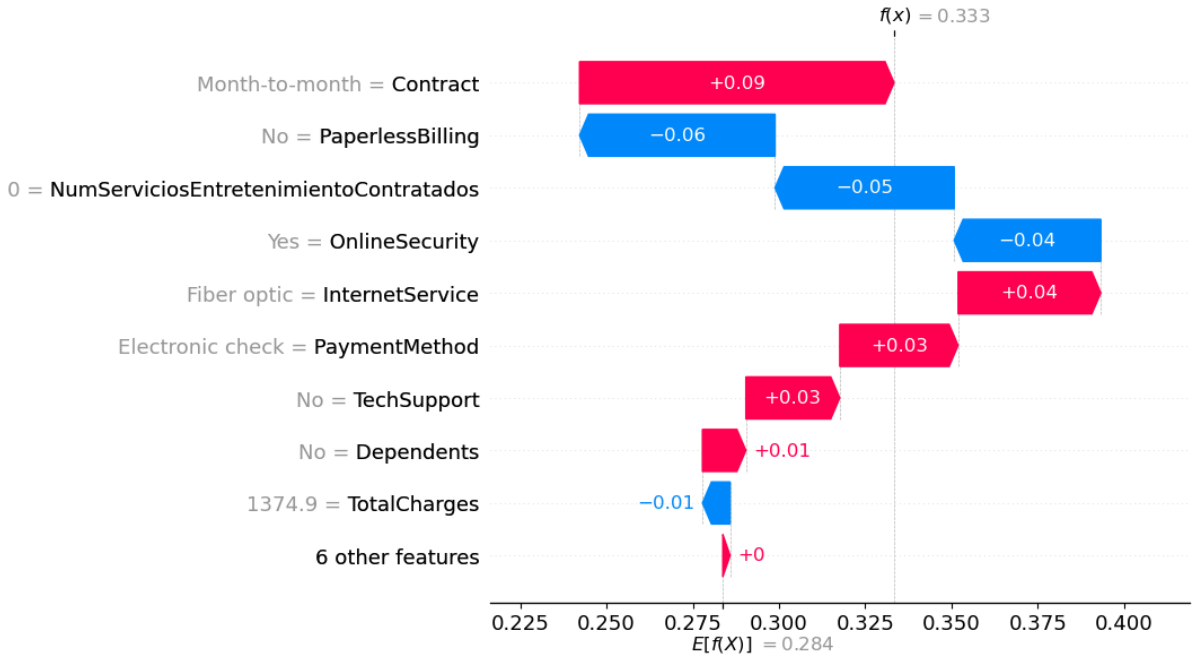
En este caso, interpretaremos la predicción de un ejemplo que se va a ir de la empresa.



Interpretación:

- La predicción del modelo para este cliente fue menor que el promedio, debido principalmente a factores como tener un contrato mes a mes, pagar con cheque electrónico, y no contar con soporte técnico. Estos factores disminuyen la probabilidad predicha. En cambio, tener facturación en papel, servicios de seguridad y sin servicios de entretenimiento aumentaron ligeramente la predicción.

Cliente que se va:



Interpretación:

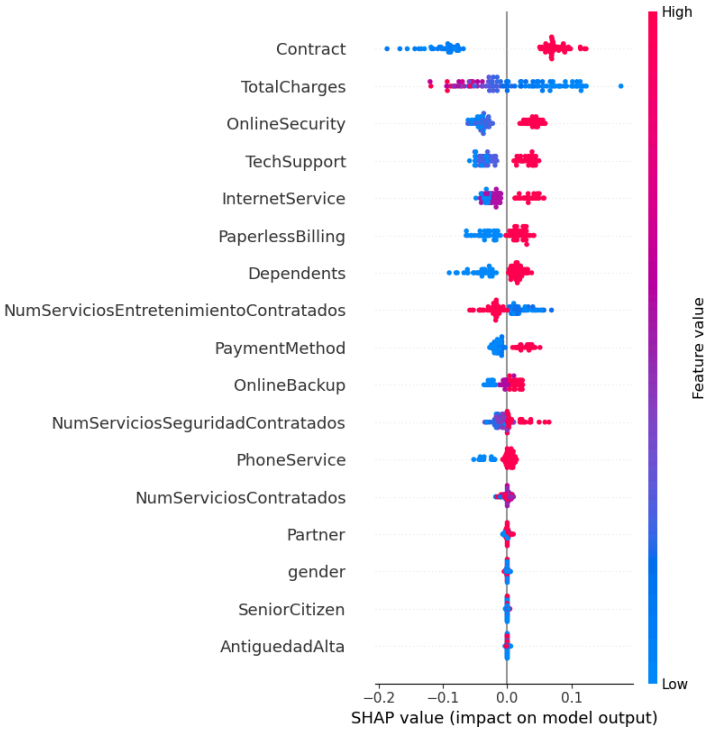
- La predicción del modelo para este cliente indica una alta probabilidad de abandono, influenciada principalmente por factores como tener un contrato mes a mes, pagar mediante medios electrónicos, utilizar conexión de fibra óptica, no contar con soporte técnico y no tener dependientes. Estos elementos están asociados a un mayor riesgo de que el cliente decida cancelar el servicio.

Conclusión (de ambos):

- La alta probabilidad de abandono en ambos casos está influenciada por características específicas del cliente asociadas a un mayor riesgo. Esto refleja la necesidad de analizar individualmente los perfiles y sugiere que ciertos grupos de clientes podrían requerir estrategias personalizadas de retención para reducir el churn.

8.2. Interpretabilidad Global:

Para tratar de entender qué variables y valores son los que más afectan a la clasificación final del churn, visualizaremos una gráfica resumen de los valores SHAP:



Interpretación (*):

El análisis de los valores SHAP revela que ciertas variables tienen un impacto significativo en la probabilidad de abandono de los clientes. En particular, los contratos de tipo mes a mes se asocian con una mayor probabilidad de baja, reflejando una menor vinculación con el servicio. Por otro lado, los clientes con cargos totales más elevados tienden a mantenerse, posiblemente debido a un mayor uso de los servicios. Además, la falta de servicios como seguridad en línea y soporte técnico incrementa el riesgo de abandono. Finalmente, no utilizar factura electrónica también se relaciona con una mayor propensión a dejar el servicio. Estas variables ofrecen pistas valiosas para enfocar estrategias de retención.

(*) Para interpretar correctamente el resumen de valores SHAP, es importante entender que en las variables categóricas, los valores altos indican una mayor frecuencia del atributo correspondiente, mientras que los valores bajos reflejan una menor frecuencia (por el tipo de codificación utilizado). Esta relación nos permite identificar qué categorías tienen un mayor impacto en la predicción del modelo.

9. Conclusión:

El modelo predictivo desarrollado permite identificar con precisión los perfiles de clientes con mayor probabilidad de abandono. A través del uso de técnicas de aprendizaje automático y la interpretación de resultados mediante SHAP, se logra una comprensión clara de los factores que influyen en la decisión del cliente. Entre las variables más relevantes se encuentran el tipo de contrato, el monto de facturación, la contratación de servicios adicionales y el uso de herramientas como la factura electrónica o el soporte técnico.

Los clientes más propensos a abandonar tienen contratos mensuales, pagan montos bajos, no contratan servicios complementarios y no cuentan con dependientes. Estos hallazgos permiten orientar estrategias de retención más efectivas, como incentivar contratos a largo plazo, promover servicios de valor agregado y aplicar acciones de fidelización temprana.

En conjunto, este análisis aporta una base sólida para la toma de decisiones estratégicas enfocadas en reducir el churn y fortalecer la relación con los clientes.