

**UNIVERSIDAD TECNOLÓGICA DE PANAMÁ**

**FACULTAD DE INGENIERÍA DE SISTEMAS DE COMPUTACIÓN**

**MAESTRÍA EN ANALÍTICA DE DATOS**

**CURSO:**

Modelos predictivos

**PROYECTO FINAL**

Análisis predictivo de la deserción de clientes en el sector de telecomunicaciones utilizando técnicas de aprendizaje automático

**FACILITADOR:**

Juan Marcos Castillo, PhD

**ESTUDIANTE:**

Kely Feng, 8-981-2182

**AÑO LECTIVO:**

2025

ÍNDICE

[Introducción 3](#_Toc203848535)

[I. Justificación 4](#_Toc203848536)

[Ii. Antecedentes 5](#_Toc203848537)

[Iii. Definición del problema 7](#_Toc203848538)

[Iv. Análisis predictivo 7](#_Toc203848539)

[V. Gráficas de visualización 11](#_Toc203848540)

[Vi. Resultados del modelo 18](#_Toc203848541)

[Conclusiones 20](#_Toc203848542)

[Recomendaciones y futuros estudios 21](#_Toc203848543)

[Referencias bibliográficas 22](#_Toc203848544)

[Anexos 23](#_Toc203848545)

# **INTRODUCCIÓN**

El sector de las telecomunicaciones atraviesa una etapa de profundas transformaciones impulsadas tanto por el avance tecnológico como por la creciente competencia en el mercado. En las últimas décadas, la liberalización del sector, la digitalización de los servicios y la aparición de nuevos actores especializados han modificado radicalmente las dinámicas comerciales tradicionales. Actualmente, los consumidores tienen acceso a una variedad cada vez mayor de opciones y ofertas, lo que les permite comparar servicios, precios y beneficios de manera sencilla y, en consecuencia, cambiar de proveedor con mayor libertad y frecuencia.

Esta situación ha elevado de manera significativa la relevancia de la retención de clientes para las empresas del sector. La deserción de clientes, conocida como *churn*, se refiere al fenómeno por el cual los usuarios dejan de utilizar un producto o servicio y optan por las alternativas que les ofrecen otros proveedores. Este fenómeno puede estar motivado por múltiples razones, como la insatisfacción con la calidad del servicio, la disponibilidad de mejores precios, el surgimiento de nuevas necesidades, la falta de atención personalizada, o simplemente por incentivos atractivos ofrecidos por la competencia. La gestión eficiente de la rotación de clientes se ha convertido en un factor estratégico clave, ya que mantener la base de clientes resulta, en la mayoría de los casos, menos costoso que captar nuevos usuarios.

En este contexto, la capacidad de anticipar el comportamiento de los clientes y predecir cuáles tienen una mayor propensión a abandonar el servicio representa una ventaja competitiva fundamental para las empresas de telecomunicaciones. El desarrollo de modelos predictivos, apoyados en técnicas avanzadas de análisis de datos y aprendizaje automático, permite a las organizaciones identificar patrones de comportamiento, comprender mejor las causas de la deserción y actuar proactivamente para mitigar su impacto. Este proyecto se propone contribuir a este objetivo mediante el análisis exhaustivo del fenómeno de *churn*, utilizando como base el dataset público **Telco Customer *Churn***, que integra información demográfica, contractual y de consumo de servicios de miles de clientes. La investigación abordará desde la preparación y limpieza de los datos, pasando por el análisis exploratorio y descriptivo, hasta la construcción, validación e interpretación de modelos de predicción que puedan ser aplicados de manera práctica en el entorno empresarial.

# **JUSTIFICACIÓN**

La gestión y prevención de la deserción de clientes en telecomunicaciones es un tema de gran relevancia tanto para la industria como para el ámbito académico. En un mercado donde los clientes disponen de múltiples alternativas, retener a los usuarios existentes se ha transformado en una de las prioridades fundamentales de las empresas. La lealtad del consumidor se ha visto debilitada por factores como la homogeneidad de los servicios, la facilidad para cambiar de proveedor y la creciente sensibilidad al precio. Así, la rotación o churn de clientes ya no es un fenómeno esporádico, sino un reto permanente que incide directamente en los ingresos, la reputación y la posición competitiva de las compañías.

El costo que representa perder un cliente suele ser mucho mayor que el de mantenerlo. Diversos estudios han demostrado que incrementar las tasas de retención, incluso en un pequeño porcentaje, puede generar impactos significativos en la rentabilidad de la empresa. Se estima, por ejemplo, que elevar la retención en apenas un 5% puede traducirse en aumentos del 25% o más en las ganancias, lo cual resalta la importancia de las estrategias orientadas a conservar a los clientes actuales. Sin embargo, muchas organizaciones todavía invierten la mayor parte de su presupuesto en captar nuevos clientes, dejando en segundo plano la fidelización de quienes ya forman parte de su cartera. Esta realidad ha llevado a reconsiderar las estrategias de gestión de clientes, dando mayor peso al análisis de datos.

El auge del análisis predictivo, especialmente con la llegada de técnicas de aprendizaje automático, ha marcado un antes y un después en la forma de abordar la retención. Los modelos basados en datos permiten detectar patrones de comportamiento que suelen anticipar el abandono, como cambios en los hábitos de consumo, demoras en los pagos, consultas frecuentes a soporte o reducción del uso de servicios contratados. Así, las empresas pueden pasar de una estrategia reactiva, respondiendo al churn una vez que ocurre, a una proactiva, interviniendo antes de que el cliente tome la decisión de marcharse. El uso de modelos como la regresión logística y árboles de decisión ha demostrado su eficacia tanto en la literatura como en la práctica empresarial, facilitando la identificación temprana de clientes en riesgo y la implementación de campañas focalizadas y oportunas.

Desde una perspectiva académica, este proyecto representa una oportunidad para fortalecer competencias en análisis de datos, modelado predictivo y análisis de comportamiento del cliente, integrando conocimientos teóricos y habilidades prácticas en un caso real. Trabajar con datos reales del sector permite experimentar con todo el ciclo analítico, desde la exploración y limpieza de datos, la selección y evaluación de modelos, hasta la interpretación y comunicación de resultados. En términos profesionales, el proyecto tiene una aplicación inmediata, ya que los modelos desarrollados pueden servir de base para la implementación de soluciones de negocio reales, mejorar la toma de decisiones y contribuir a la formación de perfiles más competitivos en el mercado laboral.

Por último, abordar el churn en telecomunicaciones a través del análisis de datos es relevante no solo por su impacto económico, sino también por sus implicaciones sociales. Al mejorar la eficiencia de las estrategias de retención, las empresas pueden ofrecer servicios más ajustados a las necesidades de los usuarios, optimizar sus recursos y aportar a la sostenibilidad del sector. Así, este proyecto no solo responde a una problemática actual del mercado, sino que también se alinea con las tendencias globales de digitalización, personalización y uso responsable de la información, contribuyendo a un desarrollo más eficiente y orientado al cliente en la industria de las telecomunicaciones.

# **ANTECEDENTES**

La predicción de la deserción de clientes, conocida como churn, ha sido durante años un área prioritaria de investigación tanto en el mundo académico como en la industria de las telecomunicaciones. La importancia de este fenómeno radica en su impacto directo sobre los ingresos, la rentabilidad y la sostenibilidad de las empresas, en un contexto donde la competencia es cada vez más intensa y los clientes cuentan con múltiples alternativas para elegir. En las primeras etapas de este campo, los métodos tradicionales basados en estadística, como la regresión logística y los árboles de decisión, se empleaban para analizar variables básicas como el tiempo de permanencia del cliente, la frecuencia de uso del servicio y el historial de pagos. Estos enfoques proporcionaban un primer marco de referencia útil, pero resultaban limitados frente a la creciente complejidad y dinamismo de los mercados actuales.

El avance de la digitalización y el crecimiento exponencial en la generación de datos han abierto nuevas posibilidades para abordar la predicción de *churn* desde una perspectiva mucho más sofisticada. Actualmente, la integración de técnicas de aprendizaje automático, tales como random forest, gradient boosting y máquinas de soporte vectorial, ha permitido superar las limitaciones de los modelos clásicos al capturar relaciones no lineales y patrones ocultos entre las variables. A su vez, el auge del aprendizaje profundo ha impulsado el uso de redes neuronales, capaces de procesar grandes volúmenes de datos heterogéneos, mejorando significativamente la capacidad de anticipar el comportamiento de los clientes y aumentar la precisión en la identificación de quienes presentan mayor riesgo de abandono.

Las investigaciones recientes subrayan que la efectividad de los modelos de predicción de *churn* depende en gran medida de la calidad y el tratamiento de los datos. Por ejemplo, la selección de atributos relevantes, como el tipo de contrato, los servicios adicionales contratados, la existencia de quejas, la antigüedad del cliente y el comportamiento de facturación, ha demostrado ser fundamental para alimentar los algoritmos y mejorar sus resultados. Además, el desbalance de clases es un reto recurrente en este campo, ya que normalmente el porcentaje de clientes que abandonan el servicio es considerablemente menor al de los que permanecen. Para abordar este problema, se han implementado técnicas como el sobremuestreo, el submuestreo y el uso de métricas de evaluación ajustadas, permitiendo que los modelos sean más sensibles y robustos frente a la detección de casos minoritarios.

Otra tendencia importante identificada en la literatura reciente es la combinación de modelos tradicionales y avanzados en sistemas híbridos, lo que permite aprovechar las fortalezas de diferentes técnicas y adaptarse a distintos escenarios de datos. Por ejemplo, se han desarrollado enfoques que integran modelos de machine learning con redes neuronales profundas o que combinan métodos supervisados y no supervisados para mejorar la segmentación y la personalización de las intervenciones.

Además de la precisión, la interpretabilidad de los modelos ha cobrado un papel relevante en la adopción de soluciones de predicción de *churn* en la industria. Herramientas como LIME y SHAP han sido propuestas para ayudar a los equipos de negocio a entender cómo los algoritmos llegan a sus predicciones, facilitando la toma de decisiones basada en información clara y comprensible para los gestores. Esta explicabilidad resulta especialmente valiosa cuando se busca justificar acciones personalizadas de retención o asignar recursos de manera eficiente.

Finalmente, la integración de modelos predictivos en los sistemas de gestión empresarial no solo permite anticipar el abandono, sino también automatizar respuestas, personalizar campañas de retención y optimizar la asignación de incentivos para los clientes de alto riesgo. De acuerdo con los estudios más recientes, las empresas que logran implementar sistemas inteligentes de predicción y acción temprana en *churn* consiguen ventajas competitivas sostenibles, incrementando tanto la satisfacción como la lealtad del cliente.

# **DEFINICIÓN DEL PROBLEMA**

En el sector de las telecomunicaciones, la pérdida de clientes representa uno de los principales desafíos estratégicos y financieros para las empresas, debido al alto costo asociado a la captación de nuevos usuarios y a la creciente facilidad con la que los clientes pueden cambiar de proveedor. A pesar de los esfuerzos realizados para mejorar la calidad del servicio y desarrollar estrategias de fidelización, las tasas de deserción continúan siendo elevadas, impactando negativamente la rentabilidad y la estabilidad del negocio.

Actualmente, muchas empresas del sector carecen de herramientas suficientemente precisas para anticipar qué clientes tienen una mayor probabilidad de abandonar el servicio, lo que limita la eficacia de sus acciones preventivas y de retención. La gran cantidad y variedad de datos generados por los usuarios, aunque ofrece un gran potencial analítico, presenta desafíos relacionados con la selección, el procesamiento y la interpretación de la información relevante para la toma de decisiones.

En este contexto, surge la necesidad de desarrollar un modelo predictivo basado en técnicas de análisis de datos y aprendizaje automático, capaz de identificar, a partir de información demográfica, contractual y de consumo, los factores que influyen en la decisión de abandono y predecir de manera anticipada qué clientes presentan mayor riesgo de *churn*. La solución a este problema permitirá a las empresas optimizar recursos, personalizar campañas de retención y mejorar significativamente sus indicadores de lealtad y rentabilidad.

# **ANÁLISIS PREDICTIVO**

1. **Determinación de la base de datos**

Para el presente estudio se seleccionó el dataset Telco Customer *Churn* de la plataforma Kaggle, el cual recopila datos reales de clientes de una compañía de telecomunicaciones. Este conjunto de datos fue elegido por su relevancia y estructura, ya que incluye tanto variables numéricas (como los cargos mensuales y el tiempo de permanencia del cliente) como categóricas (por ejemplo, tipo de contrato, servicios adicionales y método de pago). La variable objetivo, *churn*, indica si un cliente abandonó (“Yes”) o no (“No”) la compañía, lo que convierte este problema en uno de clasificación binaria clásico y altamente representativo en aplicaciones reales de análisis predictivo.

Previo al análisis, se revisó la estructura de los datos mediante comandos como *df.info()* y *df.head()*, para identificar el número de registros, las columnas disponibles y los tipos de datos, así como la existencia de valores nulos o inconsistentes. Esto permitió asegurar que la base de datos cumplía con las condiciones mínimas de calidad y representatividad para su utilización en modelos de *machine learning*.

1. **Limpieza y preprocesamiento**

El pre-procesamiento es una fase fundamental en cualquier proyecto de ciencia de datos, ya que garantiza que los modelos reciban información limpia y coherente. Se llevaron a cabo las siguientes tareas:

* Eliminación de columnas irrelevantes: La columna “customerID” fue eliminada por no aportar valor predictivo al modelo.
* Conversión de tipos de datos: La columna “TotalCharges” fue convertida a formato numérico. Aquellos valores imposibles de convertir (por ejemplo, celdas vacías o registros erróneos) se transformaron en NaN para su posterior tratamiento.
* Gestión de valores nulos: Se identificaron registros con valores faltantes en “TotalCharges”. Estos fueron imputados con la media de la columna, reduciendo la pérdida de datos y manteniendo la consistencia estadística del dataset.
* Filtrado de registros atípicos: Se eliminaron registros con “tenure” igual a cero, ya que representan clientes recién ingresados o posibles errores en la base de datos.
* Recodificación de variables: Para facilitar la interpretación y el análisis, la variable “SeniorCitizen” fue transformada de formato 0/1 a “No”/“Yes”, lo que ayuda en las visualizaciones y en el entendimiento por parte de los usuarios no técnicos.
* Estandarización de variables numéricas: Se identificaron las columnas numéricas relevantes para posteriores etapas de modelado y escalado, específicamente “tenure”, “MonthlyCharges” y “TotalCharges”.

1. **Análisis descriptivo**

El análisis exploratorio de datos (EDA) es esencial para descubrir patrones y relaciones dentro del dataset antes de construir cualquier modelo predictivo. Para profundizar en el entendimiento de los datos, se realizaron diversas visualizaciones y resúmenes estadísticos, entre las que destacan:

* Distribución de género y *churn*: Los gráficos de pastel permitieron observar la proporción de hombres y mujeres en la base de clientes, así como la tasa global de abandono (*churn*), un indicador clave para empresas de servicios.
* Variables numéricas: Histogramas, curvas de densidad (KDE) y boxplots para “tenure”, “MonthlyCharges” y “TotalCharges” mostraron que los clientes con baja antigüedad presentan una mayor tasa de *churn*, y quienes abandonan suelen tener cargos mensuales más altos.
* Variables categóricas: Gráficos de barras para las variables contractuales y de servicios (como tipo de contrato, streaming, método de pago, soporte técnico, seguridad online) revelaron que los clientes con contratos mensuales y sin servicios adicionales tienen mayor propensión a abandonar la compañía.
* Matriz de correlación: Utilizando LabelEncoder para codificar variables categóricas, se construyó una matriz de correlación para identificar las relaciones lineales entre todas las variables y la variable objetivo “*Churn*”. Las correlaciones más fuertes permitieron resaltar

Entre los hallazgos más relevantes destacan:

* Los clientes con contratos mensuales presentan tasas de *churn* significativamente superiores en comparación con quienes tienen contratos de mayor plazo (anual o bianual).
* Los usuarios con cargos mensuales más elevados son más propensos a abandonar la compañía, mientras que los que permanecen suelen tener antigüedad y cargos totales mayores.
* Variables como el uso de servicios adicionales (por ejemplo, soporte técnico o seguridad online), la modalidad de facturación electrónica y la existencia de dependientes o pareja mostraron asociaciones con la probabilidad de *churn*.
* No se observaron diferencias marcadas de *churn* por género, pero sí por otras variables contractuales y de servicios.

1. **Selección de variables**

La selección de variables es un paso crítico que afecta directamente la eficiencia y precisión de los modelos predictivos. Para ello, se combinaron dos estrategias:

* **Análisis de correlación:** Se priorizaron aquellas variables con mayor correlación absoluta (positiva o negativa) respecto al *churn*, como “MonthlyCharges”, “PaperlessBilling”, “SeniorCitizen”, “PaymentMethod”, “Contract”, “Tenure”, “OnlineSecurity”, “TechSupport”, “TotalCharges”, “OnlineBackup”, “DeviceProtection”.
* **Importancia según Random Forest:** Se entrenó un modelo Random Forest y se visualizó la importancia de cada variable. Los resultados confirmaron que “Contract”, “tenure”, “MonthlyCharges”, “OnlineSecurity”, “TechSupport”, y “InternetService” son las variables más determinantes para explicar el abandono de clientes.

1. **Selección de modelos**

El problema abordado corresponde a una clasificación supervisada, ya que el objetivo es predecir si un cliente de telecomunicaciones abandonará o no la compañía, de acuerdo con la variable objetivo *Churn*. Para ello, se seleccionaron diversos algoritmos de clasificación ampliamente utilizados y robustos en la literatura y la industria. Se implementaron y compararon los siguientes modelos en Python:

* **Regresión Logística:** Modelo base ampliamente utilizado por su sencillez, interpretabilidad y eficiencia. Permite identificar la influencia y el peso de cada variable predictora sobre la probabilidad de abandono.
* **Árbol de Decisión:** Modelo flexible y visual, útil para identificar reglas explícitas de decisión y segmentar a los clientes según diferentes características. Es especialmente útil cuando se busca interpretar el razonamiento detrás de la predicción.
* **Random Forest:** Algoritmo ensemble robusto y versátil, ideal para datos tabulares con múltiples variables. Ofrece una alta precisión, reduce el riesgo de sobreajuste respecto a los árboles simples y permite medir la importancia relativa de cada variable en la predicción del *churn*, como se muestra en el análisis anterior.
* **K-Nearest Neighbors (KNN):** Este modelo predice el *churn* en función de los clientes más similares, aportando una perspectiva complementaria a los métodos anteriores.
* **Máquinas de Vectores de Soporte (SVM):** Algoritmo potente para separar clases incluso en escenarios con fronteras complejas. Es especialmente útil cuando las relaciones no son lineales y requiere que las variables estén correctamente escaladas.

Para seleccionar el mejor modelo, se utilizaron métricas estándar de clasificación binaria como exactitud (accuracy), precisión (precision), recall, F1-score y la matriz de confusión. Estas métricas permiten evaluar tanto la capacidad global de predicción como el desempeño específico en la detección de clientes propensos a abandonar.

La comparación de modelos permite identificar no solo cuál tiene el mejor desempeño predictivo, sino también cuál ofrece interpretabilidad y robustez para el negocio, facilitando la toma de decisiones estratégicas orientadas a la retención de clientes.

# **GRÁFICAS DE VISUALIZACIÓN**

Con el propósito de identificar patrones relevantes en las variables del conjunto de datos, se generaron distintas visualizaciones que permiten analizar la distribución y relación entre variables sociodemográficas, económicas y características de servicio, en función de la deserción de clientes. Este enfoque es fundamental para comprender los factores que inciden en la decisión de los usuarios de mantenerse o abandonar la compañía de telecomunicaciones.

* **Distribución de género**

La visualización de la variable “gender” revela que la proporción de clientes hombres y mujeres es muy similar, sin diferencias apreciables en la tasa de *churn* entre ambos grupos. Esto indica que el género no representa un factor determinante en la deserción de clientes dentro del sector analizado, y su impacto como predictor resulta bajo.



* **Impacto de la variable adulto mayor en la tasa de *churn***

La gráfica de adultos mayores muestra que los clientes clasificados como “senior citizen” presentan una tasa de *churn* ligeramente superior respecto al resto de la población. Sin embargo, la magnitud de esta diferencia no es suficiente para considerarla una variable crítica, aunque puede adquirir mayor relevancia cuando se analiza en combinación con otros factores, como los servicios contratados o el tipo de contrato.



* **Relación entre pareja y/o dependientes y tasa de deserción**

El análisis de la distribución de *churn* en función de la presencia de pareja o dependientes indica que los clientes que viven en pareja o tienen personas a su cargo tienden a mostrar una menor tasa de deserción. Esto sugiere que los usuarios con mayor estabilidad familiar o responsabilidades económicas presentan un mayor compromiso con la compañía, por lo que estas variables pueden ser útiles para segmentar campañas de retención.

A graph of a couple of pink and purple bars

AI-generated content may be incorrect.A graph showing different colored bars

AI-generated content may be incorrect.

* **Relación entre antigüedad del cliente vs tasa de deserción**

La gráfica de “tenure” pone en evidencia que los clientes con menor tiempo de permanencia (meses con la compañía) tienen una probabilidad mucho mayor de *churn*. A medida que aumenta la antigüedad, disminuye de forma notable la tasa de abandono, reafirmando la importancia de la fidelización temprana y el seguimiento a nuevos usuarios.

A diagram of a chart

AI-generated content may be incorrect.

* **Impacto del tipo de contrato en el *churn***

El análisis del tipo de contrato evidencia una diferencia sustancial: los clientes con contrato mensual registran una tasa de *churn* considerablemente superior frente a quienes poseen contratos anuales o bianuales. Los contratos a mayor plazo actúan como barrera de salida, incrementando la estabilidad y reduciendo la rotación de usuarios.

A graph of a contract and a contract

AI-generated content may be incorrect.

* **Relación entre facturación electrónica y el *churn***

La visualización de la variable de facturación electrónica muestra que los clientes que optan por la facturación electrónica presentan una tasa de *churn* levemente mayor, posiblemente asociada a un perfil más digital y menos fidelizado, aunque el impacto de esta variable es secundario respecto a otras.

A graph of a bar graph

AI-generated content may be incorrect.

* **Relevancia del método de pago en la tasa de deserción**

La variable de método de pago indica que los usuarios que utilizan métodos automáticos de pago (transferencias o tarjetas) tienen una menor probabilidad de *churn* respecto a quienes pagan por cheque electrónico o de forma manual, sugiriendo que la automatización de pagos puede facilitar la retención.

A graph of payment method and payment method

AI-generated content may be incorrect.

* **Cargos mensuales y totales vs *churn***

Las gráficas de cargos mensuales y cargos totales demuestran que los clientes con cargos mensuales elevados tienden a desertar con mayor frecuencia, mientras que aquellos con cargos totales altos suelen corresponder a usuarios antiguos y con menor *churn*. El nivel de gasto mensual es, por tanto, un factor de riesgo relevante.

**A graph of a distribution of electricity

AI-generated content may be incorrect.A graph of a graph of a person

AI-generated content may be incorrect.**

* **Relación entre variables de servicios adicionales vs *churn***

Variables como servicio de internet, seguridad online, soporte técnico, protección de dispositivo y respaldo online, muestran que los clientes que no contratan servicios adicionales presentan tasas de *churn* significativamente superiores. La falta de servicios de valor agregado parece asociarse a una menor fidelidad y mayor predisposición al cambio.

A graph of different colored bars

AI-generated content may be incorrect.A graph of different colored bars

AI-generated content may be incorrect.

* **Análisis de matriz de correlación entre las variables y *churn***

La matriz de correlación permite identificar la fuerza y dirección de la relación lineal entre las diferentes variables del dataset y la variable objetivo, en este caso, el *churn*. Un valor de correlación cercano a 1 o -1 indica una fuerte relación positiva o negativa respectivamente, mientras que valores próximos a 0 sugieren una relación débil o nula. Se encontró lo siguiente:

* **Contract (Tipo de contrato):** Es la variable con la correlación negativa más alta respecto a *churn*. Esto significa que a mayor duración del contrato (por ejemplo, contratos de uno o dos años), la probabilidad de *churn* disminuye de manera significativa. Los clientes con contrato mensual son mucho más propensos a abandonar el servicio, lo que corrobora la importancia de las estrategias de fidelización a través de contratos de mayor plazo.
* **Tenure (Antigüedad):** Se observa una correlación negativa considerable entre tenure y *churn*. Los clientes con mayor tiempo en la compañía tienden a permanecer, mientras que la rotación es mucho mayor entre los usuarios recientes. Esto refuerza la necesidad de enfocarse en los primeros meses del ciclo de vida del cliente para reducir el *churn*.
* **MonthlyCharges (Cargos mensuales):** Existe una correlación positiva moderada entre los cargos mensuales y el *churn*. Los clientes con facturas mensuales más altas tienden a desertar con mayor frecuencia, posiblemente debido a una menor percepción de valor o sensibilidad al precio.
* **TotalCharges (Cargos totales):** Se identifica una ligera correlación negativa con *churn*, lo que es consistente con el hallazgo de que los clientes antiguos (con mayores cargos acumulados) son menos propensos a irse.
* **Servicios adicionales:** Variables como “OnlineSecurity”, “TechSupport”, “OnlineBackup” y “DeviceProtection” muestran una correlación negativa con *churn*, aunque de menor magnitud que tenure y contrato. Esto implica que los clientes que contratan servicios complementarios presentan una mayor fidelidad.
* **Método de pago:** Esta variable exhibe una correlación baja pero significativa con *churn*, siendo los métodos automáticos (débito/transferencia) asociados a menor deserción en comparación con los pagos manuales o electrónicos.
* **Variables demográficas:** El género, la condición de adulto mayor, la presencia de pareja o dependientes muestran correlaciones cercanas a cero. Esto indica que, al menos desde una perspectiva lineal, su impacto en la probabilidad de *churn* es bajo.
* **Relación entre variables predictoras:** La matriz revela también fuertes correlaciones entre ciertos servicios adicionales (por ejemplo, entre OnlineSecurity, OnlineBackup y DeviceProtection), lo que sugiere que muchos clientes contratan paquetes completos de servicios, y la falta de estos incrementa la vulnerabilidad al *churn*.



# **RESULTADOS DEL MODELO**

Tras la etapa de preprocesamiento y análisis exploratorio, se entrenaron y evaluaron cinco modelos de clasificación: Regresión Logística, Árbol de Decisión, Random Forest, K-Nearest Neighbors (KNN) y SVM. Cada modelo fue ajustado y probado utilizando el conjunto de datos preparado, reservando el 30% de las observaciones para la validación.

Las métricas clave consideradas para comparar los modelos fueron: accuracy, precisión, recall, F1-score y la matriz de confusión. A continuación, se presentan los resultados más relevantes para cada algoritmo:

**Tabla 1**

*Comparación de métricas principales por modelo*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Modelo** | **Accuracy** | **Precision** | **Recall** | **F1-score** |
| Árbol de Decisión | 0.74 | 0.74 | 0.74 | 0.74 |
| Random Forest | 0.79 | 0.79 | 0.79 | 0.79 |
| Regresión Logística | 0.81 | 0.80 | 0.81 | 0.80 |
| KNN | 0.78 | 0.76 | 0.78 | 0.76 |
| SVM | 0.81 | 0.80 | 0.81 | 0.80 |

**Fuente:** Elaboración propia a partir de los reportes de clasificación generados en Python.

**Análisis de resultados:**

* Regresión Logística y SVM obtuvieron el mejor desempeño global, con un f1-score weighted avg. de 0.80 y un accuracy de 0.81. Esto indica que ambos modelos son muy consistentes para predecir correctamente tanto a los clientes que abandonan como a los que permanecen, ponderando por la proporción real de cada clase.
* Random Forest también logró un desempeño robusto (f1-score weighted avg. de 0.79 y accuracy de 0.79), mostrando ser una alternativa sólida, especialmente para capturar relaciones no lineales y combinaciones de variables.
* K-Nearest Neighbors (KNN) alcanzó un f1-score weighted avg. de 0.76, lo que, si bien es aceptable, muestra un desempeño ligeramente inferior al de los modelos mencionados anteriormente.
* Árbol de Decisión obtuvo los valores más bajos (f1-score weighted avg. de 0.74), indicando que, aunque es interpretable, no es el modelo más eficiente para este problema.

La matriz de confusión muestra cómo se desempeñó cada modelo en la predicción de las dos clases:

* 0: Clientes que NO abandonan (No *Churn*)
* 1: Clientes que SÍ abandonan (*Churn*)

**Tabla 2**

*Matriz de confusión por modelo*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Modelo** | **Verdaderos No *Churn* (TN)** | **Falsos *Churn* (FP)** | **Falsos No *Churn* (FN)** | **Verdaderos *Churn* (TP)** |
| Árbol de Decisión | 1255 | 294 | 265 | 296 |
| Random Forest | 1377 | 172 | 262 | 299 |
| Regresión Logística | 1381 | 168 | 235 | 326 |
| KNN | 1391 | 158 | 315 | 246 |
| SVM | 1423 | 126 | 280 | 281 |

**Fuente:** Elaboración propia a partir de los reportes de clasificación generados en Python.

**Análisis de matriz de confusión:**

* Regresión Logística y SVM tienen la mayor cantidad de verdaderos negativos (TN), es decir, predicen muy bien a los clientes que permanecen, y también la mayor cantidad de verdaderos positivos (TP) entre los modelos probados, lo que resulta en mejor recall y precisión global, especialmente reflejado en el *weighted average*.
* Random Forest también predice bien la clase mayoritaria y logra un balance razonable entre TP y TN, lo que respalda su buen desempeño global.
* KNN muestra un número bajo de verdaderos positivos (246), lo que explica su menor recall para la clase de *churn*, aunque predice muchos verdaderos negativos (1391).
* Árbol de Decisión tiene un desempeño más bajo: logra menos TP y más FN, lo que afecta tanto el recall como el f1-score.

# **CONCLUSIONES**

El análisis predictivo realizado permitió abordar de manera integral la problemática del *churn* o deserción de clientes en el sector de telecomunicaciones. A partir de la exploración y limpieza de datos, se identificaron variables clave que influyen en la decisión de abandono, siendo especialmente relevantes la antigüedad del cliente (tenure), los cargos mensuales y totales, el tipo de contrato y el acceso a servicios adicionales como soporte técnico y seguridad en línea.

En la fase de modelado, se implementaron y compararon diversos algoritmos de clasificación, como Regresión Logística, Árbol de Decisión, Random Forest, K-Nearest Neighbors y SVM. Los resultados indicaron que los modelos de Regresión Logística y SVM ofrecieron el mejor desempeño general, logrando una weighted average accuracy del 81%, y un balance adecuado entre precisión y recall en ambas clases. Sin embargo, todos los modelos presentaron cierto grado de dificultad para identificar correctamente a los clientes propensos a abandonar, lo que sugiere la necesidad de seguir trabajando en técnicas para el manejo de clases desbalanceadas.

Desde una perspectiva práctica, estos hallazgos ofrecen a la empresa información accionable sobre los factores de mayor peso en la deserción. Por ejemplo, la identificación de los clientes con contratos mensuales, cargos elevados y poco tiempo de permanencia puede permitir el diseño de estrategias de retención más efectivas y focalizadas, como ofertas personalizadas o mejoras en el servicio al cliente. Asimismo, el análisis de importancia de variables puede orientar las campañas de marketing y optimizar los recursos destinados a prevenir el *churn*.

Desde el punto de vista académico, el proyecto demuestra la utilidad de los modelos supervisados para la predicción de eventos de negocio relevantes y valida el enfoque basado en técnicas de análisis exploratorio, selección de variables y comparación de diferentes algoritmos. Se recomienda para futuros estudios profundizar en el ajuste de hiperparámetros, probar modelos avanzados (como XGBoost o redes neuronales), y considerar técnicas de balanceo de clases (como SMOTE) para mejorar la capacidad predictiva en casos desbalanceados.

En conclusión, la integración de análisis estadístico, visualización y modelado predictivo no solo permite anticipar el comportamiento de los clientes, sino también aporta una base sólida para la toma de decisiones estratégicas que contribuyan a reducir la deserción y mejorar la rentabilidad del negocio.

# **RECOMENDACIONES Y FUTUROS ESTUDIOS**

Con base en los resultados obtenidos y el análisis realizado, se proponen las siguientes recomendaciones y líneas de trabajo para futuros estudios:

* Ampliar la exploración de técnicas de preprocesamiento: Se recomienda, para estudios futuros, evaluar el impacto de diferentes métodos de tratamiento de valores faltantes, codificación de variables categóricas (como One-Hot Encoding, ordinal encoding), y escalamiento de variables, a fin de comparar cómo influyen en la capacidad predictiva y la estabilidad de los modelos aplicados al problema de *churn*.
* Comparar una mayor diversidad de algoritmos: Si bien el presente estudio se centró en algoritmos clásicos de clasificación como Árboles de Decisión, Random Forest, Regresión Logística, KNN y SVM, se sugiere ampliar el análisis a modelos de boosting (XGBoost, LightGBM), ensambles y técnicas de aprendizaje profundo, para evaluar mejoras potenciales en métricas como precisión, recall o área bajo la curva ROC.
* Aplicar validación cruzada y ajuste de hiperparámetros: Se recomienda implementar validación cruzada k-fold y grid search para optimizar los hiperparámetros de los modelos, obteniendo así estimaciones más robustas de su desempeño y reduciendo el riesgo de sobreajuste o variabilidad debida a la partición de los datos.
* Estudiar el desbalance de clases: Dado que el fenómeno de *churn* suele presentar una distribución desigual (menos casos de abandono respecto a permanencia), es relevante explorar técnicas para balancear las clases, tales como sobremuestreo (SMOTE), submuestreo, o el uso de penalizaciones en los modelos.
* Profundizar en el análisis de interpretabilidad: Es recomendable complementar el análisis con técnicas de interpretabilidad como SHAP, LIME, o el análisis de importancia de variables, para entender con mayor claridad el aporte individual de cada predictor y facilitar la comunicación de resultados a audiencias no técnicas.
* Incorporar variables externas y datos longitudinales: Para futuras investigaciones, se sugiere integrar fuentes adicionales de información, como encuestas de satisfacción, registros de interacción con el cliente o series temporales, que permitan analizar la evolución del *churn* a lo largo del tiempo y bajo diferentes contextos.

# **REFERENCIAS BIBLIOGRÁFICAS**

Arabnia, H., Beikmohammadi, A., Imani, M., Joudaki, M. (2025). *Customer Churn Prediction: A Review of Recent Advances, Trends, and Challenges in Conventional Machine Learning and Deep Learning*. Research Gate. Recuperado el 30 de junio de 2025 de [https://www.researchgate.net/publication/390255151\_Customer\_*Churn*\_Prediction\_A\_Review\_of\_Recent\_Advances\_Trends\_and\_Challenges\_in\_Conventional\_Machine\_Learning\_and\_Deep\_Learning](https://www.researchgate.net/publication/390255151_Customer_Churn_Prediction_A_Review_of_Recent_Advances_Trends_and_Challenges_in_Conventional_Machine_Learning_and_Deep_Learning)

Chandrakala, D., Kumar, A. (2016). *A Survey on Customer Churn Prediction using Machine Learning Techniques*. International Journal of Computer Applications. Recuperado el 30 de junio de 2025 de <https://www.ijcaonline.org/archives/volume154/number10/kumar-2016-ijca-912237.pdf>

Kaggle. (s.f.). *Telco Customer Churn*. Kaggle. Obtenido de [https://www.kaggle.com/datasets/blastchar/telco-customer-*churn*](https://www.kaggle.com/datasets/blastchar/telco-customer-churn)

Peng, K., Peng, Y., Li, W. (2023). Research on customer *churn* prediction and model interpretability analysis. National Library of Medicine. Recuperado el 8 de julio de 2025 de <https://pmc.ncbi.nlm.nih.gov/articles/PMC10707658/>

Prabadevi, B., Shalini, R., Kavitha, B.R. (2023). *Customer churning analysis using machine learning algorithms*. Science Direct. Recuperado el 3 de julio de 2025 de <https://www.sciencedirect.com/science/article/pii/S2666603023000143>

# **ANEXOS**

A continuación, se listan los recursos complementarios y archivos utilizados en el desarrollo del presente proyecto. Estos se encuentran disponibles en el siguiente repositorio de GitHub: <https://github.com/kelyfeng/Modelos_predictivos_proyecto_final>