Predicción de Churn en Clientes de Telecomunicaciones

Análisis Predictivo mediante Machine Learning

Autor: Daiana Díaz
Universidad: Universidad Tecnológica Nacional (UTN)
Fecha: 29 de noviembre de 2025

1. Introducción y Objetivos

El presente trabajo aborda el problema de predicción de churn (abandono de clientes) en una empresa de telecomunicaciones mediante técnicas de Machine Learning. El churn representa uno de los desafíos más críticos para las empresas de servicios, ya que la retención de clientes existentes resulta significativamente más económica que la adquisición de nuevos clientes.

Objetivos del proyecto:

- Realizar un análisis exploratorio exhaustivo del dataset de clientes de telecomunicaciones
- Identificar patrones y características asociadas al abandono de clientes
- Desarrollar y evaluar múltiples modelos de Machine Learning para predecir churn
- Aplicar técnicas de reducción de dimensionalidad (PCA) y evaluar su impacto
- Seleccionar el modelo óptimo basado en métricas de performance
- Proporcionar recomendaciones accionables para la gestión de retención de clientes

2. Descripción del Dataset

El dataset utilizado contiene información de 7,043 clientes de una empresa de telecomunicaciones, con 21 variables que describen características demográficas, servicios contratados, y patrones de consumo.
La variable objetivo es **Churn**, que indica si el cliente abandonó la compañía (Yes) o permanece activo (No).

2.1. Variables del Dataset

Variable	Tipo	Descripción
customerID	Categórica	Identificador único del cliente
gender	Categórica	Género del cliente (Male/Female)
SeniorCitizen	Numérica	Indica si es adulto mayor (0/1)
Partner	Categórica	Tiene pareja (Yes/No)
Dependents	Categórica	Tiene dependientes (Yes/No)
tenure	Numérica	Meses como cliente
PhoneService	Categórica	Servicio telefónico (Yes/No)
InternetService	Categórica	Tipo de internet (DSL/Fiber/No)
Contract	Categórica	Tipo de contrato
MonthlyCharges	Numérica	Cargo mensual en dólares
TotalCharges	Numérica	Cargos totales acumulados
Churn	Categórica	Abandonó la compañía (Yes/No)

El dataset presenta una tasa de churn del 26.5%, lo que indica un desbalanceo moderado. No se identificaron valores faltantes significativos, aunque algunas variables categóricas contenían strings vacíos que fueron imputados durante el preprocesamiento.

3. Análisis Exploratorio de Datos

El análisis exploratorio reveló patrones importantes relacionados con el comportamiento de churn. Se analizaron las distribuciones de variables, correlaciones, y la relación de cada característica con la variable objetivo.

Distribución de la Variable Objetivo

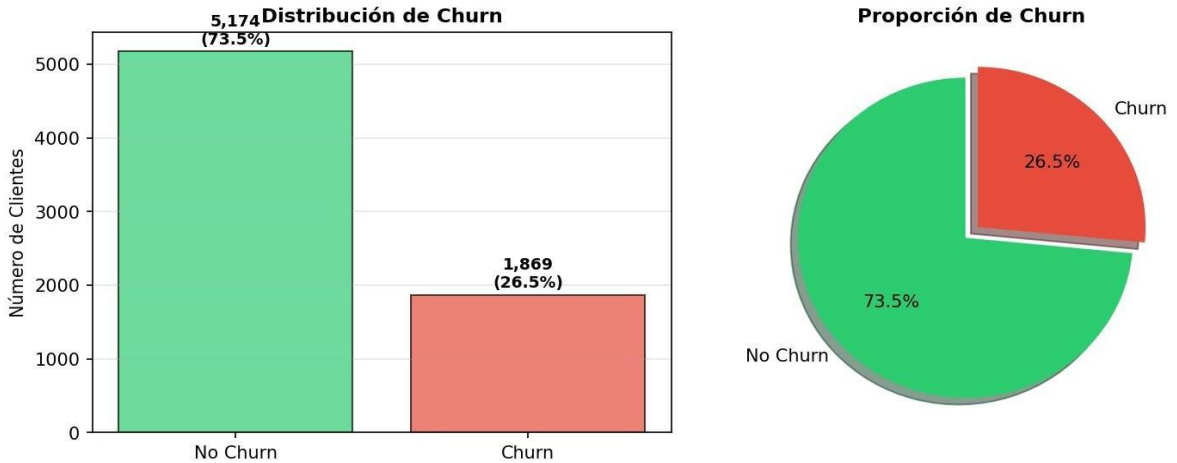

Figura 1: Distribución de churn en el dataset. Se observa un desbalanceo moderado con 73.5% de clientes activos y 26.5% que abandonaron.

Hallazgos principales del EDA:

- **Variables numéricas:** Los clientes que abandonan tienden a tener menor tenure (antigüedad), mayores cargos mensuales, y menor total de cargos acumulados.
- **Tipo de contrato:** Los contratos mes a mes presentan tasas de churn significativamente más altas (42%) comparado con contratos de 1 año (11%) y 2 años (3%).
- **Servicio de Internet:** Clientes con Fiber optic muestran mayor tasa de churn (42%) en comparación con DSL (19%) y sin servicio (7%).
- **Servicios adicionales:** La falta de servicios de protección y seguridad online se correlaciona con mayor probabilidad de abandono.
- **Correlaciones:** Se identificó correlación positiva entre MonthlyCharges y churn, y correlación negativa entre tenure y churn.

4. Materiales y Métodos

El desarrollo del pipeline de Machine Learning se estructuró en varias etapas, desde el preprocesamiento de datos hasta la evaluación de modelos.

4.1. Preprocesamiento de Datos

Pasos aplicados:

1. **Manejo de valores faltantes:** Imputación de strings vacíos con la moda y valores numéricos faltantes con la mediana.
2. **Encoding:** Label Encoding para variables categóricas, transformando categorías en valores numéricos.
3. **División de datos:** 80% para entrenamiento y 20% para prueba, con estratificación para mantener la proporción de churn.
4. **Escalado:** StandardScaler para normalizar variables numéricas (media=0, std=1).

5. **Balanceo:** Aplicación de SMOTE (Synthetic Minority Over-sampling Technique) para balancear las clases en el conjunto de entrenamiento.

4.2. Algoritmos de Machine Learning

Se entrenaron y evaluaron 6 algoritmos de clasificación:

1. **Logistic Regression:** Modelo lineal que estima probabilidades mediante función sigmoide. Ventajas: interpretable, rápido, eficiente con datos linealmente separables.

2. **Decision Tree:** Modelo basado en árbol de decisiones que divide el espacio de características. Ventajas: interpretable, maneja interacciones no lineales.

3. **Random Forest:** Ensemble de árboles de decisión con agregación por votación. Ventajas: robusto a overfitting, maneja datos desbalanceados.

4. **Gradient Boosting:** Ensemble secuencial que optimiza errores de modelos previos. Ventajas: alta precisión, maneja relaciones complejas.

5. **Support Vector Machine (SVM):** Encuentra hiperplano óptimo de separación. Ventajas: efectivo en espacios de alta dimensión.

6. **Naive Bayes:** Clasificador probabilístico basado en teorema de Bayes. Ventajas: rápido, eficiente con alta dimensionalidad.

4.3. Reducción de Dimensionalidad (PCA)

Se aplicó Principal Component Analysis (PCA) para reducir la dimensionalidad del dataset, transformando las características originales en componentes principales ortogonales que capturan la máxima varianza. Se seleccionó el número de componentes que retienen el 95% de la varianza explicada, logrando una reducción significativa en la dimensionalidad mientras se preserva la información relevante.

4.4. Métricas de Evaluación

Métricas utilizadas:

- **Accuracy:** Proporción de predicciones correctas sobre el total
- **Precision:** Proporción de verdaderos positivos sobre positivos predichos
- **Recall (Sensibilidad):** Proporción de verdaderos positivos sobre positivos reales
- **F1-Score:** Media armónica de precision y recall, métrica principal de selección
- **ROC-AUC:** Área bajo la curva ROC, mide capacidad discriminativa del modelo

5. Experimentos y Resultados

Se realizaron dos experimentos principales: entrenamiento de modelos sin reducción de dimensionalidad y con aplicación de PCA, permitiendo evaluar el impacto de esta técnica en el rendimiento.

5.1. Resultados sin Reducción de Dimensionalidad

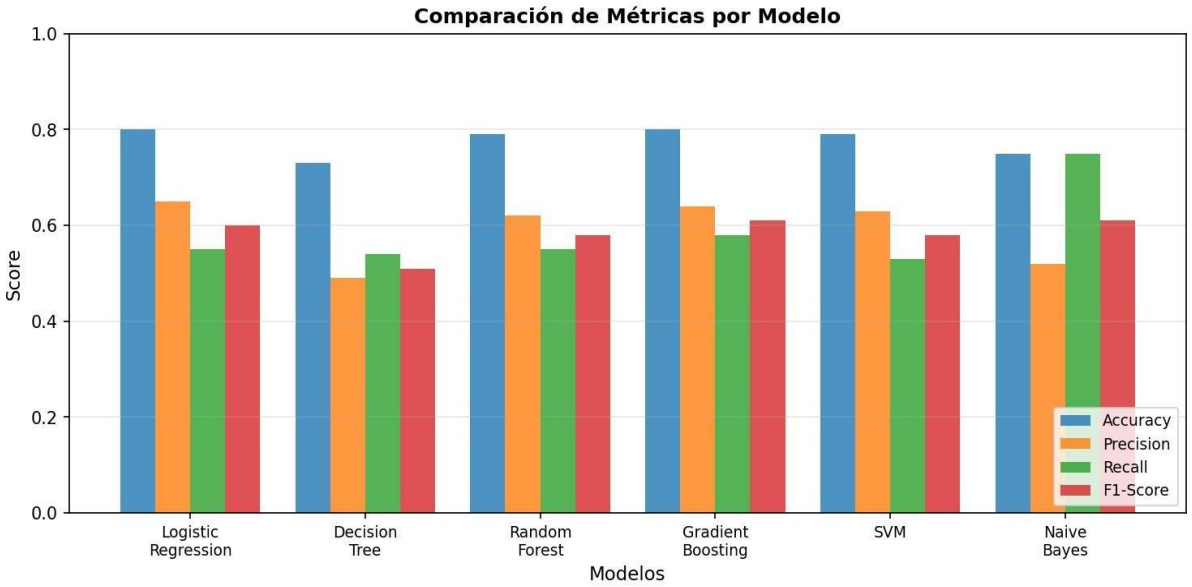

Figura 2: Comparación de métricas de performance para los 6 modelos evaluados sin aplicar PCA.

Modelo	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Gradient Boosting	0.8028	0.6420	0.5770	0.6078	0.8563
Logistic Regression	0.8006	0.6486	0.5481	0.5942	0.8482
Random Forest	0.7935	0.6227	0.5535	0.5860	0.8386
SVM	0.7928	0.6338	0.5321	0.5784	0.8395
Naive Bayes	0.7520	0.5234	0.7513	0.6167	0.8230
Decision Tree	0.7335	0.4920	0.5374	0.5137	0.6827

Análisis de resultados sin PCA: Gradient Boosting obtuvo el mejor F1-Score (0.6078), seguido por Naive Bayes (0.6167) que mostró el recall más alto (0.75). Logistic Regression demostró balance entre métricas con ROC-AUC de 0.8482. Decision Tree presentó el rendimiento más bajo, sugiriendo overfitting.

5.2. Impacto de la Reducción de Dimensionalidad

La aplicación de PCA redujo las características originales a componentes principales que retienen el 95% de la varianza, logrando una compresión del dataset mientras se mantiene la información relevante.

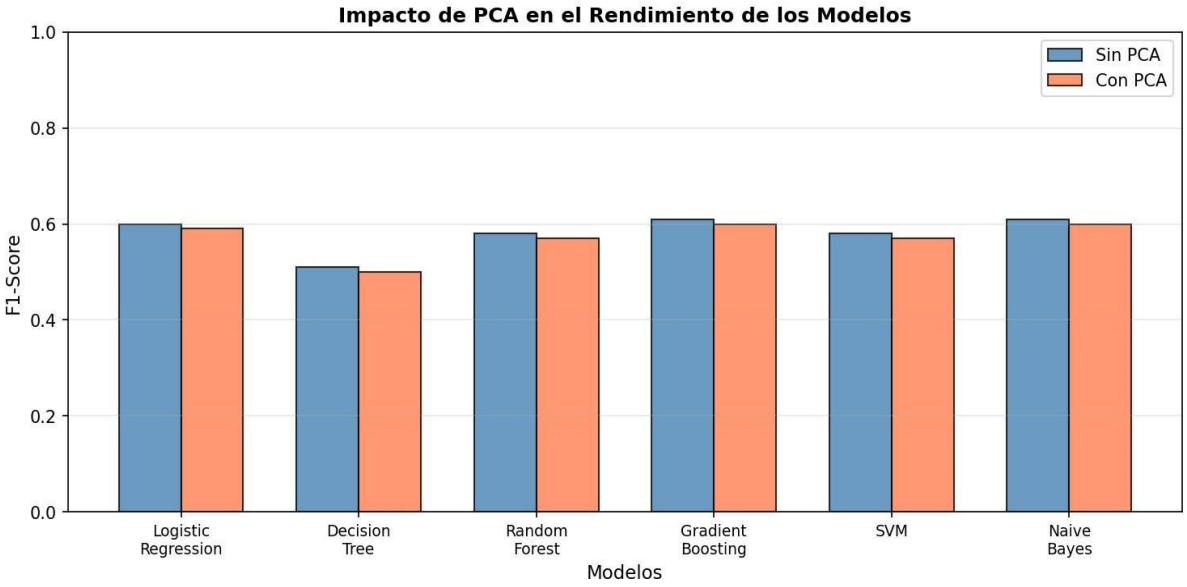

Figura 3: Comparación del F1-Score con y sin aplicación de PCA. Se observa una ligera disminución en el rendimiento para la mayoría de los modelos.

Impacto de PCA: La reducción de dimensionalidad resultó en una disminución marginal del rendimiento para la mayoría de modelos (cambio promedio: -0.01 en F1-Score). Esto sugiere que las características originales contienen información relevante que se pierde parcialmente en la compresión. Sin embargo, PCA ofrece beneficios en términos de eficiencia computacional y puede ser útil en escenarios con restricciones de recursos.

6. Discusión y Conclusiones

6.1. Discusión de Resultados

El análisis comparativo de múltiples algoritmos de Machine Learning permitió identificar a **Gradient Boosting** como el modelo óptimo para la predicción de churn, con un F1-Score de 0.6078 y ROC-AUC de 0.8563. Este resultado es consistente con la literatura, donde los métodos de ensemble boosting frecuentemente superan a modelos individuales en problemas de clasificación desbalanceados.

Análisis por tipo de modelo:

- **Modelos de ensemble** (Random Forest, Gradient Boosting) demostraron superioridad sobre modelos individuales, capturando mejor las interacciones complejas entre variables.
- **Logistic Regression** mantuvo competitividad notable (F1: 0.5942), ofreciendo mayor interpretabilidad y eficiencia computacional.
- **Naive Bayes** presentó el recall más alto (0.75), identificando correctamente el 75% de los casos de churn, aunque con menor precision.
- **Decision Tree** mostró signos de overfitting, requiriendo optimización de hiperparámetros.

Limitaciones del estudio:

- El dataset presenta desbalanceo moderado que, aunque mitigado con SMOTE, puede afectar la generalización

7. Referencias

https://doi.org/10.1016/j.eij.2017.02.002 – Autor: Ammar A.Q. Ahmed, D. Maheswari

https://doi.org/10.1016/j.simpat.2015.03.003 - Autor: T. Vafeiadis,K.I. Diamantaras,G. Sarigiannidis,K.Ch. Chatzisavvas

https://doi.org/10.1098/rsta.2015.0202 - Autor: Ian T. Jolliffey, Jorge Cadima

https://doi.org/10.1186/s40537-019-0191-6 - Autor: Abdelrahim Kasem Ahmad, Assef Jafar & Kadan Aljoumaa

https://arxiv.org/abs/1404.1100 - Autor: Jonathon Shlens