Diseño y evaluación de un sistema de predicción de abandono de clientes usando el dataset Telco Customer Churn

Sneyder Buitrago González, y Daniel Ramírez Cárdenas

ÍNDICE

I. DESCRIPCIÓN DEL PROBLEMA

El proyecto tiene como objetivo desarrollar un sistema de predicción del abandono de clientes (*churn*) para un operador de telecomunicaciones, con el fin de que la empresa pueda priorizar estrategias de retención basadas en las probabilidades individuales de abandono. Para ello se utiliza el conjunto de datos *Telco Customer Churn* (Kaggle) [1], que contiene 7,043 registros y alrededor de 21 variables relacionadas con características demográficas, contractuales, de servicios y de facturación. La variable objetivo, `Churn`, es binaria y representa si un cliente ha abandonado o no el servicio. En la muestra analizada, la proporción de clientes que abandonaron es de aproximadamente 26.5 % ($\approx 1{,}869$ positivos frente a $\approx 5{,}174$ negativos), lo que sugiere un desbalance moderado que debe considerarse tanto en la evaluación de modelos como en el diseño del experimento.

El análisis exploratorio de los datos revela patrones relevantes para la formulación del problema y el diseño del pipeline de preprocesamiento. En términos demográficos, la distribución por género es aproximadamente balanceada, mientras que la proporción de adultos mayores (`SeniorCitizen`) es reducida, con cerca de 1,100 casos. La variable `tenure`, que representa la permanencia del cliente, muestra concentraciones notables tanto en valores muy bajos (incluyendo un pico en `tenure`=0) como en valores altos (alrededor de 72 meses), lo que sugiere diferencias significativas en el comportamiento según el tiempo de relación con la empresa. En cuanto a las variables económicas, `MonthlyCharges` presenta una distribución multimodal que abarca desde valores bajos cercanos a 18 hasta más de 120, mientras que `TotalCharges` se encuentra fuertemente sesgada hacia la derecha, con valores que superan los 8,000. Además, se observan algunos valores faltantes (alrededor de 11 registros), usualmente asociados a clientes con `tenure`=0.

En el plano de los servicios contratados, la mayoría de los clientes dispone de `PhoneService`, y una fracción importante utiliza Internet por fibra óptica. Los servicios adicionales, como protección de dispositivos, respaldo de datos, soporte técnico o streaming, exhiben frecuencias diversas pero aportan información potencialmente discriminante. Desde el punto de vista contractual, la modalidad de contrato *month-to-month* es la más común y el método de pago *Electronic check* es el predominante, factores que, según la literatura, suelen estar asociados a mayores tasas de abandono, lo que motiva su análisis detallado.

Considerando la naturaleza de los datos y el objetivo del estudio, la estrategia metodológica se fundamenta en el aprendizaje supervisado para clasificación binaria. El preprocesamiento de los datos incluirá la conversión de la variable `TotalCharges` a formato numérico y el tratamiento de los valores faltantes mediante imputación por mediana condicional o, cuando sea apropiado, mediante la reconstrucción de su valor a partir de la relación entre `MonthlyCharges` y `tenure`. Asimismo, las variables binarias, como `Partner` o `PhoneService`, se transformarán en representaciones 0/1, mientras que las categóricas multiclase, como `InternetService`, `PaymentMethod` o `Contract`, se codificarán mediante *one-hot encoding*. Las variables numéricas relevantes, como `tenure`, `MonthlyCharges` y `TotalCharges`, se escalarán cuando el modelo lo requiera (por ejemplo, en SVM o redes neuronales), manteniendo su escala original en modelos basados en árboles.

Para la etapa de modelado, se propone comparar diferentes enfoques de aprendizaje supervisado con el fin de equilibrar interpretabilidad, capacidad predictiva y diversidad algorítmica. La regresión logística se utilizará como modelo base por su simplicidad y capacidad explicativa. Los modelos basados en árboles, como Random Forest y Gradient Boosting, permitirán capturar interacciones no lineales y manejar variables mixtas sin necesidad de normalización. Adicionalmente, se considerarán SVM con kernel radial, apropiadas para estructuras de decisión no lineales, y redes neuronales multicapa, que podrán explorar relaciones más complejas entre las variables. Esta combinación metodológica garantiza un análisis comparativo robusto y permite identificar el mejor compromiso entre rendimiento predictivo y explicabilidad.

II. ESTADO DEL ARTE

En la literatura reciente que emplea exactamente el archivo `WA_Fn-UseC_-Telco-Customer-Churn.csv` (Telco Customer Churn) se observa una convergencia metodológica: ambos estudios modelan el problema como clasificación supervisada binaria y comparan familias de algoritmos lineales, basados en árboles y de tipo ensamblado o redes neuronales

[1]–[3]. El estudio identificado como SDPIT (2024) realiza un flujo reproducible (limpieza de registros con `TotalCharges` vacíos, codificación de categóricas, particionado entrenamiento/prueba) y compara regresión logística, árboles, Random Forest y SVM; sus resultados muestran AUCs entre ≈0.81 y ≈0.85, con regresión logística en la cumbre según su experimento [2]. Pawar et al. (GIJET, 2024) evalúan además XGBoost y redes neuronales aplicando K-fold cross-validation para estabilizar estimaciones, y reportan que ensamblados y ANN alcanzan mayores *accuracy* (por ejemplo ANN ≈85.6 % en la configuración reportada), mientras que regresión logística y Random Forest se ubican en rangos de 79–82 % según el ajuste [3].

Ambos trabajos usan métricas derivadas de la matriz de confusión y AUC; a continuación se presentan de forma separada las definiciones matemáticas básicas utilizadas en dichos estudios.

$$\text{Precision} \;=\; \frac{TP}{TP+FP}$$

$$\text{Recall (TPR)} \;=\; \frac{TP}{TP+FN}$$

$$F1 \;=\; 2\cdot\frac{\text{Precision}\cdot\text{Recall}}{\text{Precision}+\text{Recall}}$$

$$\text{AUC} \;=\; P\big(\text{score}(x^{+}) > \text{score}(x^{-})\big)$$

De forma sucinta, los trabajos que usan exactamente el mismo dataset coinciden en que (i) el preprocesamiento de `TotalCharges` y de las categóricas es crítico, (ii) los métodos de ensamblado y las redes mejoran *accuracy* con *tuning*, y (iii) la regresión logística sigue siendo competitiva en AUC bajo ciertas condiciones empíricas. Estas observaciones justifican comparar modelos lineales, basados en árboles y redes en la etapa experimental de este informe [4], [5].

REFERENCIAS

[1] I. S. Data, “Telco customer churn dataset,” https://www.kaggle.com/datasets/blastchar/telco-customer-churn, 2017, wA_Fn-UseC_-Telco-Customer-Churn.csv. Consultado en octubre de 2025.

[2] Highlights in Science, Engineering and Technology (SDPIT) authors, “Telco customer churn prediction,” https://www.researchgate.net/publication/381544028_Telco_Customer_Churn_Prediction, 2024, estudio experimental usando el dataset WA_Fn-UseC_-Telco-Customer-Churn (Kaggle). Consultado en octubre de 2025.

[3] D. Pawar, Y. Sabla, N. Tayal, and S. Nainan, “Predicting telecom customer churn: An in-depth evaluation of machine learning algorithms,” *Global International Journal of Engineering and Technology (GIJET)*, 2024, estudio experimental usando WA_Fn-UseC_-Telco-Customer-Churn (Kaggle). Consultado en octubre de 2025.

[4] A. Barsotti *et al.*, “A decade of churn prediction techniques in the telco industry: survey and perspectives,” *SN Computer Science*, 2024.

[5] W. Verbeke, D. Martens, C. Mues, and B. Baesens, “Social network analysis for customer churn prediction,” *Applied Soft Computing*, vol. 14, pp. 431–446, 2014.