Evaluación de los modelos CatBoost, LightGBM y XGBoost para la identificación de consumos anómalos asociados a pérdidas no técnicas en sistemas de distribución de energía eléctrica

Anderson López Chaverra

Asesor

Esneider de Jesús Pineda Martinez

Universidad Nacional Abierta y a Distancia UNAD

Escuela de Escuela de Ciencias Básicas, Tecnología e Ingeniería ECBTI

Especialización en Ciencia de Datos y Analítica

2026

Nota de Aceptación

_____

Nombre Director de Trabajo de Grado

_____ _____

Jurado Jurado

Resumen

Las pérdidas no técnicas (PNT) de energía eléctrica asociadas principalmente a fraudes y errores de medición siguen impactando la rentabilidad de los operadores de red latinoamericanos. La literatura coincide en que los algoritmos de gradient boosting alcanzan los excelentes resultados de detección; no obstante, esos estudios se han realizado en contextos de datos y condiciones operativas diferentes a las de la empresa distribuidora colombiana analizada en este proyecto. En consecuencia, el presente trabajo replica tres modelos de clasificación tal como: CatBoost (CGB), LightGBM (LGB) y XGBoost (XGB), utilizando un dataset histórico que incluye lecturas mensuales, así como variables técnicas de los equipos de medida**.** Tras aplicar limpieza, normalización, balanceo de clases y generación de variables derivadas, cada algoritmo será optimizado y evaluado con F1 Score, Recall y Precisión mediante validación cruzada estratificada. El estudio busca comprobar si los altos niveles de desempeño mostrados en otros entornos se reproducen en la realidad local, y determinar qué modelo ofrece la mejor capacidad predictiva para futuras implementaciones de analítica de pérdidas.

Palabras clave**:** Pérdidas no técnicas, Detección de fraude, Aprendizaje automático, Gradient Boosting.

Abstract

Non-technical losses (NTL) of electrical energy, primarily associated with fraud and measurement errors, continue to impact the profitability of power network operators in Latin America. The literature agrees that gradient boosting algorithms achieve excellent detection results; however, these studies have been conducted in contexts with different data and operational conditions compared to those of the Colombian distribution company analyzed in this project. Consequently, this work replicates three classification models: CatBoost (CGB), LightGBM (LGB), and XGBoost (XGB), using a historical dataset that includes monthly readings as well as technical variables from the metering equipment. After applying data cleaning, normalization, class balancing, and the generation of derived features, each algorithm will be optimized and evaluated with F1 Score, Recall and Precision, through stratified cross-validation. The study aims to verify whether the high-performance levels shown in other environments can be replicated in the local context, and to determine which model provides the best predictive capability for future loss analytics implementations.

Keywords**:** Non-technical losses, Fraud detection, Machine learning, Gradient Boosting.

Tabla de Contenido

Lista de Figuras

Lista de Tablas

Introducción

Las pérdidas no técnicas (PNT) representan uno de los desafíos más significativos para la rentabilidad y la sostenibilidad de los operadores de red en el sector eléctrico a nivel mundial, y particularmente en América Latina y Colombia (Moreno Gallón, 2022). Estas pérdidas, que se asocian principalmente a fraudes, errores de medición y conexiones ilegales, impactan directamente la calidad del servicio y la viabilidad económica de las empresas distribuidoras (Arias Marín & Gómez Bravo, 2019; Mohammadi et al., 2020). La magnitud de este problema ha impulsado la búsqueda de soluciones innovadoras y eficientes para su detección y mitigación.

Tradicionalmente, la identificación de PNT dependía en gran medida de inspecciones físicas. No obstante, con la evolución de los medidores inteligentes y la capacidad de manejar grandes cantidades de datos de consumo, el uso de técnicas de minería de datos ha demostrado ser una estrategia altamente eficaz. (Loya Ayala et al., 2019; Quezada Mateo, 2017; Trejos Ramírez, 2014). La literatura académica reciente ha explorado extensamente el uso de diversos modelos de clasificación para identificar patrones de consumo anómalos que sugieran la presencia de PNT (Bohani et al., 2021; El-Hamouz, 2018; Imtiaz & Shaikh, 2020; Iskandar et al., 2021).

Dentro de este panorama, los algoritmos de Gradient Boosting (GB), específicamente XGBoost (XGB), LightGBM (LGB) y CatBoost (CGB), han demostrado un rendimiento superior en la detección de PNT en diversos contextos internacionales (Singh & Tiwari, 2024; Sun et al., 2022; Tang et al., 2023; Wang et al., 2021). Estos modelos, basados en el concepto de aprendizaje por ensamble, construyen secuencialmente múltiples árboles de decisión, corrigiendo los errores de los árboles predecesores, lo que les permite capturar relaciones complejas y no lineales en los datos de consumo energético (Khan et al., 2021; Mahfouz et al., 2021; Majumder

et al., 2022). La eficacia de los GBMs para manejar conjuntos de datos desbalanceados, un problema común en la detección de fraude donde los casos de PNT son una minoría, también es un factor clave en su adopción (Nabipour et al., 2022). Otros enfoques como las Redes Neuronales Profundas (Deep Learning), incluyendo CNNs y Autoencoders (Ferreira et al., 2021; Fuster Gelabert, 2023; Guilherme et al., 2022; Khan et al., 2022; Li et al., 2021), así como las Máquinas de Soporte Vectorial (SVM) (Ismail et al., 2018), también han mostrado resultados prometedores, consolidando el aprendizaje automático como la principal herramienta en esta área.

En el contexto colombiano, la Unidad de Planeación Minero-Energética (UPME), como entidad central en la planificación del sector, ha resaltado la importancia de gestionar eficientemente las pérdidas en el sistema eléctrico nacional (UPME, 2023). Informes y estudios de la UPME han analizado la situación actual de la eficiencia operativa y los desafíos asociados a la reducción de pérdidas, incluyendo las no técnicas, en las redes de distribución (UPME, 2023b; UPME, 2022). La necesidad de modernizar la medición y combatir el fraude eléctrico es un tema recurrente en las discusiones sobre la optimización del balance energético nacional, donde la UPME proporciona datos y análisis estadísticos relevantes (UPME, 2021; UPME, 2020).

A pesar de los buenos resultados de estos algoritmos, su efectividad puede variar según las particularidades del mercado y la infraestructura eléctrica local. Por lo tanto, el presente estudio propone evaluar el rendimiento de los modelos CatBoost, LightGBM y XGBoost mediante el análisis de métricas de desempeño, con el fin de determinar el clasificador más eficaz para la identificación de consumos anómalos asociados a PNT en un sistema de distribución de energía eléctrica en Colombia.

Planteamiento del Problema

Las Pérdidas No Técnicas (PNT) se configuran como una brecha comercial crítica más que como un fenómeno físico. Esta discrepancia entre la energía inyectada y la facturada, derivada de intervenciones irregulares en la red y anomalías en la gestión de la medida, compromete directamente el flujo de caja y la sostenibilidad operativa de las empresas del sector.

Esta problemática genera una presión financiera constante sobre las empresas distribuidoras, particularmente en América Latina y Colombia (Arias Marín & Méndez Peñaloza, 2022; Llagua Arévalo, 2023; Quezada Mateo, 2017; Trejos Ramírez, 2014; Tonato et al., 2023). Esta problemática genera una presión financiera constante sobre las empresas distribuidoras, ya que los costos asociados a la energía no facturada y al deterioro de la infraestructura deben ser gestionados en entornos regulatorios cada vez más exigentes.

Tradicionalmente, la detección de estas PNT se ha basado en inspecciones físicas reactivas, un método que, si bien sigue siendo vital, es costoso y de eficiencia limitada para cubrir la vasta población de usuarios (Arias Marín & Méndez Peñaloza, 2019; Loya Ayala et al., 2019). La principal limitación radica en la falta de precisión para priorizar las órdenes de inspección, lo que deriva en un alto porcentaje de visitas fallidas y un uso ineficiente de los recursos operativos en campo.

Con la creciente implementación de medidores inteligentes y disponibilidad de grandes volúmenes de datos de consumo de energía, ha surgido la oportunidad de emplear técnicas avanzadas Machine Learning para una detección más proactiva y eficiente de consumos anómalos (Fuster Gelabert, 2023; Moreno Gallón, 2022). La literatura académica ha demostrado que algoritmos de Gradient Boosting Machines (GBM) como XGBoost, LightGBM y CatBoost ofrecen un rendimiento superior en la detección de PNT en diversos contextos (Macao Sánchez

& Pujota Cuasapaz, 2022). Estos modelos son efectivos por su capacidad de manejar datos desbalanceados (donde los casos de fraude son minoría), capturar relaciones complejas y no lineales en los patrones de consumo, y su robustez ante valores atípicos y ruido. Sin embargo, el problema técnico que persiste es la incertidumbre sobre cuál de estas arquitecturas logra el equilibrio óptimo entre sensibilidad y precisión bajo las dinámicas de fraude específicas de un operador de red en Colombia.

En consecuencia, este trabajo busca resolver la ineficiencia en la identificación de anomalías mediante la evaluación comparativa de estos modelos. El objetivo es determinar el algoritmo que maximice la detección de fraudes reales, minimizando los falsos positivos que encarecen la operación, proporcionando así una herramienta de analítica predictiva que sustente la toma de decisiones estratégicas y mejore la efectividad de los planes de reducción de pérdidas del operador de red.

Justificación

La justificación de este proyecto radica en la crítica necesidad de abordar eficientemente las pérdidas no técnicas en el sector eléctrico. Las PNT no solo impactan negativamente los ingresos de los operadores de red, afectando su rentabilidad y capacidad de inversión, sino que también pueden generar desbalances de carga, inestabilidad en la red, sobrecargas y una disminución en la calidad del servicio para los usuarios legítimos. Además, desde una perspectiva social, el fraude energético implica una transferencia de costos hacia los consumidores honestos, distorsionando las tarifas y promoviendo la inequidad. La Unidad de Planeación Minero-Energética (UPME) de Colombia ha enfatizado repetidamente la importancia de gestionar eficientemente estas pérdidas para asegurar la viabilidad a largo plazo del sector eléctrico nacional.

La adopción de técnicas de Machine Learning, específicamente los algoritmos de Gradient Boosting, representa una evolución significativa respecto a los métodos tradicionales de detección. Estos modelos permiten pasar de un enfoque reactivo (inspecciones físicas) a uno proactivo y basado en evidencia, analizando grandes volúmenes de datos para identificar patrones de consumo anómalos que indican la presencia de PNT. La eficacia de los GBM ha sido ampliamente documentada en la literatura, lo que sugiere su potencial para ser una herramienta poderosa en el contexto colombiano. Sin embargo, es esencial verificar si estos resultados se mantienen en el contexto operativo y con los datos específicos de los operadores de red local. Este proyecto busca replicar y evaluar estos modelos en el entorno local, contribuyendo a la toma de decisiones para mitigar el impacto de las PNT.

Objetivos

Objetivo General

Evaluar el rendimiento de los modelos CatBoost, LightGBM y XGBoost, mediante el análisis de métricas de desempeño, para determinar el clasificador más eficaz en la identificación de consumos anómalos asociados a pérdidas no técnicas en un sistema de distribución de energía eléctrica.

Objetivos Específicos

Validar la calidad y consistencia de los datos históricos de consumo eléctrico y atributos técnicos, mediante procesos de depuración, tratamiento de valores atípicos, balanceo de clases y construcción de variables derivadas, con el fin de asegurar su idoneidad para la evaluación de modelos de clasificación.

Argumentar el desempeño de los modelos, mediante la comparación de métricas como F1-Score, Recall y Precisión, calculadas a partir de modelos ajustados por optimización de hiperparámetros, con el fin de sustentar su capacidad para detectar consumos anómalos.

Recomendar el modelo de clasificación más apropiado, mediante la jerarquización integral de resultados y la consideración de criterios de eficiencia computacional, con el fin de respaldar la toma de decisiones técnicas en una empresa distribuidora de energía eléctrica.

Marco Teórico

El presente marco teórico aborda los conceptos fundamentales relacionados con las pérdidas no técnicas (PNT) en sistemas de distribución eléctrica y la aplicación de algoritmos de clasificación de Machine Learning para su detección. Se estructura en tres secciones principales: la problemática de las PNT, la evolución de los métodos de detección y el análisis de los algoritmos de Machine Learning más relevantes.

Pérdidas No Técnicas (PNT) en Sistemas Eléctricos de Distribución

Las pérdidas en los sistemas de distribución de energía eléctrica se clasifican en técnicas y no técnicas. Las pérdidas técnicas son inherentes al funcionamiento físico de la red (ej., efecto Joule en conductores, pérdidas en transformadores), y su minimización está ligada a la arquitectura y diseño de la red (UPME, 2023b). Por otro lado, las pérdidas no técnicas (PNT) son aquellas que no tienen un origen físico en la red y se deben principalmente a factores externos o actividades irregulares. Estas incluyen el hurto de energía (conexiones ilegales, manipulación de medidores), errores en la medición o facturación, y fallas administrativas (Arias Marín & Gómez Bravo, 2019; Mohammadi et al., 2020).

El impacto de las PNT es multifacético. Financieramente, representan una disminución directa de los ingresos para los operadores de red, afectando su rentabilidad y capacidad de inversión (Moreno Gallón, 2022). Operacionalmente, pueden provocar desbalances en la carga, inestabilidad en la red, sobrecargas en transformadores y líneas, y una menor calidad del servicio para los usuarios legítimos (Imtiaz & Shaikh, 2020). Desde una perspectiva social y de equidad, el fraude energético implica una transferencia de costos hacia los consumidores que sí pagan sus facturas, distorsionando las tarifas y generando inequidad (Iskandar et al., 2021).

En Colombia, la Unidad de Planeación Minero-Energética (UPME) ha señalado que la disminución de las pérdidas representa un reto clave para asegurar la viabilidad a largo plazo del sector eléctrico. Esta entidad lleva a cabo evaluaciones continuas sobre el desempeño operativo de las empresas distribuidoras de energía. Asimismo, analiza el impacto de las pérdidas no técnicas (PNT) en la balanza energética del país, utilizando los datos presentados en sus informes estadísticos.

Evolución de los Métodos de Detección de PNT

Históricamente, la detección de PNT se basaba en métodos reactivos, como las inspecciones in situ impulsadas por denuncias o sospechas (Quezada Mateo, 2017). Si bien estas inspecciones siguen siendo una parte vital del proceso, su costo y eficiencia limitada para cubrir grandes poblaciones de usuarios han impulsado el desarrollo de técnicas más sofisticadas (Trejos Ramírez, 2014).

Con la popularización de la infraestructura de medición avanzada, se ha generado una gran cantidad de datos de consumo de energía con alta granularidad temporal (Loya Ayala et al., 2019). Esta riqueza de datos ha sentado las bases para la aplicación de enfoques basados en analítica de datos, minería de datos e inteligencia artificial (El-Hamouz, 2018; Bohani et al., 2021). Estos métodos permiten identificar patrones de consumo anómalos o comportamientos inusuales que podrían indicar la presencia de PNT, transformando la detección en un proceso más proactivo y basado en evidencia. La literatura ha explorado diversas técnicas de aprendizaje supervisado y no supervisado para este fin (Ruiz Suárez, 2022; Ismail et al., 2018).

Modelos de Clasificación Basados en Gradient Boosting para la Detección de PNT

La detección de PNT, cuando se aborda desde una perspectiva de datos de consumo, se formula comúnmente como un problema de clasificación. El objetivo es etiquetar como fraude -

no fraude a un consumidor. La literatura académica es vasta en la aplicación de modelos de Machine Learning (ML) y Deep Learning (DL) para este propósito (Singh & Tiwari, 2024).

Entre los diversos modelos estudiados, las investigaciones recientes coinciden en el excelente desempeño de los algoritmos de Gradient Boosting (GB). Estos modelos de ensemble, que incluyen XGBoost (XGB), LightGBM (LGB) y CatBoost (CGB), han demostrado consistentemente una alta precisión y robustez en la detección de PNT en diversos estudios (Sun et al., 2022; Tang et al., 2023; Wang et al., 2021). La principal ventaja de estos modelos se basa en su capacidad para:

- Manejo de datos desbalanceados: El fraude es un fenómeno raro, lo que da lugar a bases de datos con un número limitado de casos positivos. Los GB, a menudo combinados con técnicas de remuestreo o ponderación de clases, son eficaces en estos escenarios (Nabipour et al., 2022).
- Capturar relaciones no lineales y complejas: Los patrones de fraude no suelen ser simples y pueden involucrar interacciones complejas entre múltiples variables (Khan et al., 2021; Majumder et al., 2022). Los GB construyen una serie de árboles de decisión donde cada nuevo árbol corrige los errores del anterior, permitiendo modelar estas complejidades de forma efectiva (Mahfouz et al., 2021).
- Robustez frente a valores atípicos y ruido: Aunque el preprocesamiento de datos es crucial, los GBMs son relativamente robustos a la presencia de ruido y valores atípicos en los datos (Li et al., 2021).

A continuación, se presentarán los tres modelos más utilizados de Gradient Boosting para la detección de pérdidas no técnicas (PNT): XGBoost (XGB), LightGBM (LGB) y CatBoost (CGB). Cada uno de estos modelos destaca por su capacidad para manejar datos desbalanceados,

capturar relaciones no lineales y ser robustos frente a valores atípicos, lo que los hace altamente efectivos en la identificación de consumos anómalos asociados a fraudes. En los siguientes subtemas, se detallarán las características y ventajas específicas de cada uno.

CatBoost (CGB)

CatBoost (CGB) es un algoritmo de Gradient Boosting desarrollado por Yandex, especialmente diseñado para manejar variables categóricas de manera eficiente. A diferencia de otros algoritmos como XGBoost y LightGBM, CatBoost no requiere un preprocesamiento exhaustivo de las variables categóricas, ya que tiene una técnica incorporada que las transforma de manera automática en un formato que puede ser utilizado en los modelos. Esta característica lo convierte en una opción atractiva cuando se trabaja con bases de datos que contienen un alto número de variables categóricas, como es común en el caso de los sistemas de detección de pérdidas no técnicas (PNT), donde las características como la ubicación, tipo de usuario, y comportamiento histórico de consumo son clave (Prokhorenkova et al., 2018).

Además, CatBoost es conocido por su robustez y rendimiento en tareas de clasificación, debido a su capacidad para manejar de manera efectiva tanto datos numéricos como categóricos. Utiliza un enfoque llamado ordered boosting, que ayuda a reducir el sobreajuste al optimizar el proceso de construcción de los árboles, especialmente en datos desbalanceados, como los encontrados en la detección de fraudes. Su capacidad para manejar grandes volúmenes de datos y su eficiencia en términos de tiempo de entrenamiento, junto con la calidad de las predicciones, lo convierte en una de las opciones preferidas en muchos problemas de machine learning, incluidos los relacionados con la detección de PNT (Li & Lin, 2020).

LightGBM (LGB)

LightGBM (LGB) es un algoritmo de Gradient Boosting desarrollado por Microsoft, diseñado para ser más eficiente y escalable en comparación con otros modelos de boosting. A diferencia de otros métodos como XGBoost y CatBoost, LightGBM emplea una técnica de histogram-based learning que permite agrupar los datos en histogramas para reducir el costo computacional durante el entrenamiento, lo que lo hace particularmente adecuado para grandes volúmenes de datos. Esta característica le otorga ventajas en tareas de clasificación y regresión cuando se enfrentan a conjuntos de datos de gran tamaño, como es el caso de la detección de pérdidas no técnicas (PNT), donde se manejan grandes cantidades de datos de consumo eléctrico (Ke et al., 2017).

Además, LightGBM es conocido por su capacidad para manejar datos desbalanceados de manera efectiva, gracias a su algoritmo de Leaf-wise growth que optimiza la precisión de los árboles de decisión en cada iteración. Esto lo convierte en una opción robusta para la detección de PNT, donde los datos suelen estar desbalanceados debido a la baja frecuencia de fraudes en comparación con los datos legítimos. Su velocidad de entrenamiento superior y su capacidad para evitar el sobreajuste, especialmente en escenarios con gran cantidad de características, hacen que LightGBM sea un modelo altamente eficiente y preciso para la clasificación de consumos anómalos (Li et al., 2020).

XGBoost (XGB)

XGBoost (XGB) es un algoritmo desarrollado por Tianqi Chen, XGBoost utiliza una técnica de boosting de árboles de decisión que se centra en optimizar la precisión del modelo al corregir los errores de los árboles anteriores en cada iteración. Este enfoque ha demostrado ser particularmente efectivo en tareas de detección de pérdidas no técnicas (PNT), donde los datos a

menudo son complejos y contienen patrones no lineales. La capacidad de XGBoost para manejar tanto datos numéricos como categóricos, junto con su habilidad para evitar el sobreajuste mediante regularización, lo convierte en una opción ideal para la clasificación de consumos anómalos (Chen & Guestrin, 2016).

Una de las características clave de XGBoost es su capacidad para manejar grandes volúmenes de datos de manera eficiente, lo que es especialmente útil en la detección de fraudes, donde las bases de datos pueden ser enormes y contener miles de registros de consumo. Además, XGBoost incorpora un enfoque de regularización L1 y L2 que ayuda a reducir la complejidad del modelo y mejora su capacidad de generalización. Esto es esencial cuando se trabaja con datos desbalanceados, como los de PNT, donde los casos de fraude son mucho menos frecuentes que los casos legítimos (Nabipour et al., 2022). Gracias a estas ventajas, XGBoost ha sido un modelo líder en competencias de machine learning y en aplicaciones industriales de clasificación de datos.

Métricas de Rendimiento

La matriz de confusión es un artefacto fundamental para evaluar el rendimiento de los modelos de clasificación. Aunque no es una métrica en sí misma, permite visualizar cómo se distribuyen las predicciones correctas e incorrectas del modelo, facilitando el cálculo de diversas métricas de rendimiento clave. Esta tabla muestra los resultados de las predicciones comparados con los valores reales, proporcionando cuatro categorías: verdaderos positivos (TP), verdaderos negativos (TN), falsos positivos (FP) y falsos negativos (FN). A partir de estos valores, se derivan métricas como:

Precisión

Mide la proporción de casos predichos como positivos que realmente son positivos. Es útil cuando se quiere minimizar los falsos positivos.

$$Precisión = \frac{TP}{TP + FP}$$

Recall

Mide la proporción de casos positivos que el modelo detecta correctamente. Es crucial cuando se quiere evitar los falsos negativos, es decir, no perder ningún caso positivo.

$$Recall = \frac{TP}{TP + FN}$$

F1-Score

Es la media armónica entre Precision y Recall, proporcionando un balance entre estas dos métricas. Es útil cuando se busca un equilibrio entre la capacidad del modelo de detectar correctamente los casos positivos y evitar los falsos positivos.

$$F1 = 2\ x\ \frac{Precisión\ x\ Recall}{Precisión + Recall}$$

Ética y Gobierno de Datos en las PNT

En el contexto del desarrollo de modelos de clasificación para la detección de pérdidas no técnicas (PNT) en el sector eléctrico, el uso de datos sensibles y personales de los clientes requiere un manejo responsable y ético. De acuerdo con la Ley 1581 de 2012, que regula la protección de los datos personales en Colombia, se deben seguir principios claros y estrictos para garantizar la confidencialidad, seguridad, transparencia y responsabilidad en el tratamiento de la información. En este sentido, el consentimiento informado para el uso de los datos personales se otorga por parte del cliente en el momento en que solicita el servicio y realiza su registro ante la

empresa distribuidora, instancia en la cual se le informa sobre las finalidades del tratamiento de sus datos.

La información utilizada en este proyecto, como la clasificación del cliente, código de identificación, consumos históricos, datos socioeconómicos y detalles del servicio, es esencial para entrenar y evaluar los modelos de clasificación. No obstante, para efectos de este trabajo de grado, estos datos serán previamente codificados, de manera que no permitan la identificación directa de los clientes, garantizando así la protección de su privacidad y el cumplimiento de la Ley 1581 de 2012.

Metodología

El presente proyecto se desarrolló bajo un enfoque aplicado, orientado a la construcción y evaluación de modelos de machine learning para la detección de consumos anómalos asociados a pérdidas no técnicas en un sistema de distribución de energía eléctrica. La metodología adoptada se basa en el modelo CRISP-DM (Cross-Industry Standard Process for Data Mining), el cual estructura el proceso analítico en fases secuenciales e iterativas, garantizando coherencia entre el problema planteado, el tratamiento de los datos, el modelado y la evaluación de resultados.

A continuación, se describen las fases metodológicas implementadas.

Comprensión del Problema

En esta fase se identificó la problemática asociada a las pérdidas no técnicas en los sistemas de distribución de energía eléctrica, las cuales representan impactos económicos y operativos significativos para las empresas distribuidoras. Estas pérdidas se manifiestan, entre otros aspectos, a través de patrones anómalos de consumo eléctrico, difíciles de detectar mediante métodos tradicionales.

Con base en este contexto, se definió el problema como una tarea de clasificación binaria, orientada a identificar registros con comportamiento anómalo a partir de históricos de consumo y atributos técnicos. Asimismo, se estableció la necesidad de evaluar diferentes algoritmos de *machine learning* que permitieran maximizar la capacidad de detección de estos eventos, priorizando métricas que reflejen un equilibrio entre detección correcta y control de errores.

Comprensión de los Datos

El conjunto de datos utilizado en el proyecto está conformado por registros históricos de consumo eléctrico, asociados a usuarios del sistema de distribución de energía. Inicialmente, el

conjunto de datos cuenta con 461.509 registros y 11 variables, las cuales incluyen tanto series temporales de lecturas de consumo como atributos técnicos del suministro.

Las variables de entrada originales se agrupan en dos categorías principales. En primer lugar, se encuentran las variables numéricas, correspondientes a las lecturas mensuales de consumo eléctrico, que permiten analizar la evolución temporal del comportamiento de cada usuario. En segundo lugar, se incluyen variables categóricas y técnicas, tales como estrato, ubicación, tipo de servicio y características del medidor, las cuales aportan contexto operativo y socioeconómico al análisis.

En esta etapa, se realizó una revisión general de la estructura de los datos, el tipo de variables y la distribución de la variable objetivo, con el fin de identificar posibles inconsistencias y evaluar la necesidad de aplicar transformaciones posteriores. Asimismo, se evidenció que la información contenida en las series temporales de consumo presenta patrones que pueden ser aprovechados mediante la generación de variables derivadas, orientadas a capturar comportamientos anómalos y características estadísticas relevantes del consumo eléctrico.

Estas variables derivadas, junto con los procesos de limpieza y balanceo de clases, se describen en detalle en la siguiente fase de la metodología.

Tabla 1

Variables de Entrada del Conjunto de Datos

Ítem	Atributo	Descripción
1	PRODUCTO	Identificador único del cliente o del punto de suministro.
2	FACT_CATEGORIA	Categoría tarifaria o tipo de facturación del cliente (p. ej., residencial, comercial, industrial, oficial, entre otros).

Ítem	Atributo	Descripción
3	TIPO_POBLACION	Clasificación del entorno del usuario según su ubicación (urbano o rural).
4	COD_CIRCUITO	Código del circuito eléctrico al que se encuentra conectado el cliente.
5	COD_CONEXION	Identificador del punto de conexión asociado al suministro (trasformador).
6	ZONA_REGION	Zona o región operativa donde se localiza el cliente dentro del sistema de distribución.
7	PRO_PLAN_COMERCIAL	Plan o segmento comercial asignado al cliente según la clasificación interna de la empresa.
8	CICLO	Ciclo de lectura/facturación asociada al cliente.
9	UBICACION	Municipio u ubicación administrativa donde se encuentra el cliente.
10	ANORMAL	Variable objetivo del modelo: etiqueta que indica la ocurrencia de consumo anómalo (p. ej., 1 = anómalo, 0 = normal).
11–70	LECTURAS	Serie temporal de consumos/lecturas periódicas registradas (5 años).

Preparación de los Datos

En esta fase se realizó el proceso de depuración, transformación y enriquecimiento del conjunto de datos con el fin de garantizar su idoneidad para el entrenamiento y evaluación de los modelos de clasificación. En primer lugar, se definió la ventana temporal de análisis correspondiente a las lecturas mensuales de consumo eléctrico y se verificó la consistencia de los

valores numéricos, realizando conversiones de tipo cuando fue necesario para asegurar que los cálculos estadísticos se ejecutaran correctamente.

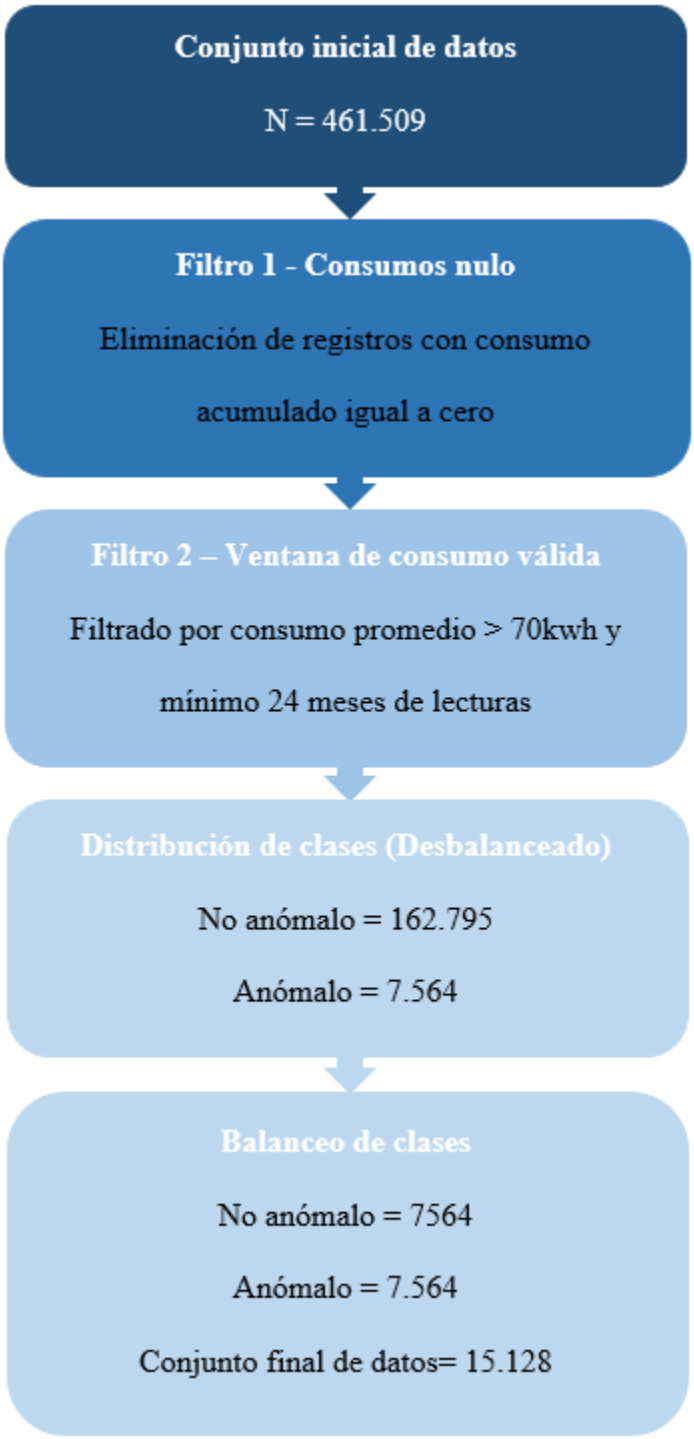
Posteriormente, se aplicaron criterios de filtrado orientados a eliminar registros sin información útil o con series temporales insuficientes para caracterizar el comportamiento del consumo. En particular, se excluyeron: (i) los registros cuya suma de consumos es igual a cero dentro de la ventana de tiempo analizada, (ii) los registros con menos de 24 meses de lecturas válidas, y (iii) aquellos cuyo consumo promedio en la ventana válida fue inferior a 70 kWh. Como resultado de este proceso de depuración, la muestra pasó de 461.509 registros iniciales a 170.332 registros finales.

A partir de este conjunto depurado, la distribución de la variable objetivo evidenció 162.768 registros clasificados como no anómalos y 7.564 registros clasificados como anómalos. Dado el desbalance entre clases, se aplicó un procedimiento de balanceo, ajustando la clase mayoritaria para igualar la cantidad de registros de la clase minoritaria. En consecuencia, se construyó un conjunto final balanceado de 15.128 registros, compuesto por 7.564 no anómalos y 7.564 anómalos. Este submuestreo se realizó para eliminar el sesgo hacia la clase mayoritaria y maximizar la sensibilidad de los modelos ante el fraude, permitiendo que los algoritmos identifiquen patrones críticos de la clase minoritaria sin comprometer la detección por el desbalance original.

El proceso de depuración y balanceo del conjunto de datos se resume en la **Figura 1**, donde se ilustra la reducción progresiva de la muestra a partir de la aplicación de criterios de filtrado y balanceo de clases.

Figura 1

Proceso de Depuración y Balanceo del Conjunto de Datos.



Finalmente, con el propósito de capturar patrones relevantes del comportamiento del consumo y mejorar la capacidad predictiva de los modelos, se calcularon variables derivadas a partir de las series temporales de lecturas. Estas variables resumen propiedades estadísticas (tendencia central, dispersión y extremos) e incorporan indicadores de variación del consumo, lo que permite representar con mayor precisión el comportamiento del usuario en el tiempo. La definición y selección de estas variables derivadas se realizó tomando como referencia los criterios propuestos por Arias Marín y Méndez Peñaloza (2022), orientados a la caracterización de series de consumo y al soporte de la identificación de comportamientos anómalos. La **Tabla 2** presenta el conjunto de variables derivadas implementadas.

Tabla 2

Variables Derivadas Calculadas a Partir del Histórico de Consumo

Ítem	Atributo	Descripción
71	MEDIA	Promedio del consumo en la ventana válida de lecturas.
72	MEDIANA	Mediana del consumo en la ventana válida de lecturas.
73	MAX	Valor máximo de consumo registrado en la ventana válida.
74	MIN	Valor mínimo de consumo registrado en la ventana válida.
75	RANGO	Diferencia entre el consumo máximo y el consumo mínimo (MAX − MIN).
76	DESV_EST	Desviación estándar del consumo en la ventana válida (medida de dispersión).
77	COEF_VAR	Coeficiente de variación del consumo (DESV_EST / MEDIA), indicador de variabilidad relativa.

Ítem	Atributo	Descripción
78	MEDIATRIZ	Semirango del consumo: RANGO / 2.
79	NUM_CEROS	Número de períodos dentro de la ventana válida en los que el consumo fue igual a cero.
80	AU_CON	Cantidad de períodos en los que el consumo fue mayor que la primera lectura válida (comparación respecto al valor inicial).
81	DI_CON	Cantidad de períodos en los que el consumo fue menor que la primera lectura válida (comparación respecto al valor inicial).
82	AU_CON_MEDIA	Cantidad de períodos en los que el consumo fue mayor que la MEDIA del registro.
83	DI_CON_MEDIA	Cantidad de períodos en los que el consumo fue menor que la MEDIA del registro.
84	AU_CON_MEDIANA	Cantidad de períodos en los que el consumo fue mayor que la MEDIANA del registro.
85	DI_CON_MEDIANA	Cantidad de períodos en los que el consumo fue menor que la MEDIANA del registro.

Entorno Computacional y Herramientas de Implementación

Las pruebas se ejecutaron en un computador portátil con procesador Intel Core i5-210H, 16 GB de memoria RAM, almacenamiento SSD de 0.5 TB y una GPU NVIDIA RTX 3050 con 6 GB de memoria dedicada (VRAM), además de 7.8 GB de memoria compartida. El entorno de

ejecución correspondió al sistema operativo Windows 11, utilizando Python versión 3.13.5 dentro de un entorno Anaconda/venv.

Para el desarrollo y entrenamiento de los modelos se emplearon las librerías scikit-learn, CatBoost, LightGBM y XGBoost, en sus versiones 1.6.1, 1.2.8, 4.6.0 y 3.1.2, respectivamente. El entrenamiento se realizó en CPU, habilitando paralelismo de hasta 12 hilos (n_jobs = 12). Con el fin de garantizar la reproducibilidad de los resultados, se fijó una semilla aleatoria común mediante el parámetro random_state = 42.

Modelado

En la fase de modelado se entrenaron tres modelos de clasificación basados en técnicas de gradient boosting: CatBoost, LightGBM y XGBoost. Estos modelos fueron seleccionados debido a su reconocida capacidad para manejar datos tabulares, capturar relaciones no lineales complejas y ofrecer un alto desempeño en problemas de clasificación binaria, particularmente en escenarios asociados a la detección de comportamientos anómalos.

El proceso de entrenamiento se realizó sobre el conjunto de datos balanceado, utilizando tanto variables originales como variables derivadas obtenidas en las etapas previas de preparación de los datos e ingeniería de características. Para cada algoritmo se definieron y ajustaron los hiperparámetros más relevantes asociados al proceso de aprendizaje y a la complejidad del modelo, tales como el número de iteraciones, la profundidad de los árboles y la tasa de aprendizaje. El ajuste de estos hiperparámetros tuvo como objetivo optimizar el desempeño predictivo de los modelos y reducir el riesgo de sobreajuste. Dicho proceso se realizó de forma secuencial, ajustando cada parámetro de mayor a menor impacto y validando la métrica de F1 en cada etapa; este procedimiento se aplicó específicamente para determinar los valores óptimos del número de iteraciones, la profundidad de los árboles y el learning rate.

La **Tabla 3** presenta un resumen de los hiperparámetros principales considerados para cada algoritmo y la configuración final utilizada durante el entrenamiento.

Tabla 3

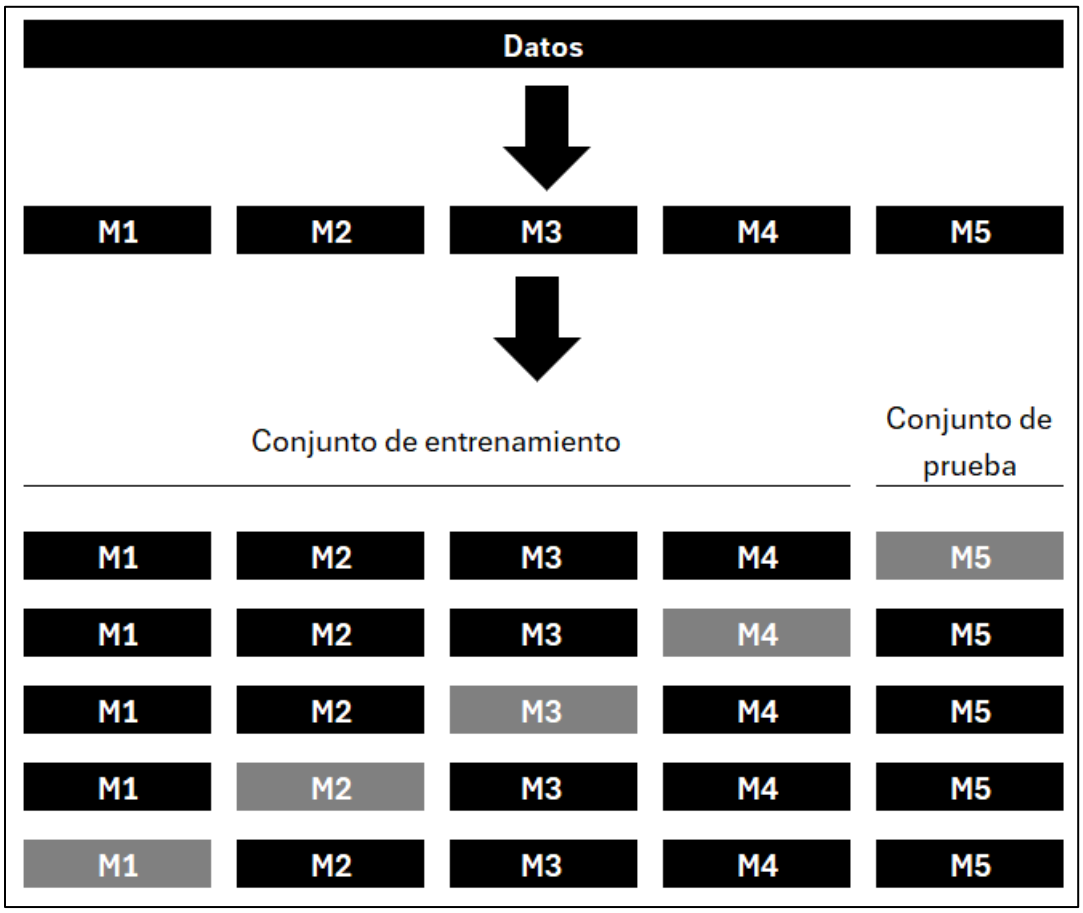
Resumen de Hiperparámetros Utilizados por Modelo

Modelo	Iteraciones	Profundidad	Learning_rate
CGB	100	7	0.1
LGB	50	5	0.1
XGB	50	5	0.1

Con el fin de obtener una estimación robusta del desempeño de los modelos y minimizar la dependencia de una única partición de los datos, se empleó un esquema de validación cruzada k-fold, con k = 5. En este enfoque, el conjunto de datos se divide en cinco subconjuntos de tamaño similar (M1–M5). En cada iteración, uno de estos subconjuntos se utiliza como conjunto de validación, mientras que los cuatro restantes conforman el conjunto de entrenamiento. Este procedimiento se repite de manera iterativa hasta que cada subconjunto ha sido utilizado una vez como conjunto de validación. El esquema general del proceso de validación cruzada empleado se ilustra en la **Figura 2**.

Figura 2

Esquema del Proceso de Validación Cruzada k-fold (k = 5) Aplicado Durante el Entrenamiento de los Modelos.



Adicionalmente, durante la fase de entrenamiento se registró el tiempo de entrenamiento de cada modelo, con el propósito de analizar su eficiencia computacional y evaluar su viabilidad para una eventual implementación en un entorno operativo real. Este análisis permite complementar la evaluación del desempeño predictivo con una perspectiva práctica relacionada con los recursos computacionales requeridos por cada algoritmo.

Evaluación

La evaluación del desempeño de los modelos se llevó a cabo utilizando métricas adecuadas para problemas de clasificación binaria con énfasis en la detección de eventos de interés. En particular, se emplearon las métricas F1-score, recall y precisión, las cuales permiten analizar de manera integral la capacidad de los modelos para identificar consumos anómalos y controlar errores de clasificación.

Las métricas fueron calculadas como el promedio de los resultados obtenidos en los pliegues de la validación cruzada. Adicionalmente, se realizó un proceso de optimización del umbral de decisión, seleccionando el valor que maximiza el F1-score, con el fin de adaptar la salida probabilística de los modelos a las características del problema analizado.

Selección del Modelo

Finalmente, se realizó una comparación integral de los modelos entrenados, considerando conjuntamente las métricas de desempeño predictivo y la eficiencia computacional. Este análisis permitió jerarquizar los modelos evaluados e identificar aquel que presenta el mejor compromiso entre capacidad de detección de consumos anómalos y costo computacional.

El modelo seleccionado constituye una herramienta potencial de apoyo a la toma de decisiones técnicas en la empresa distribuidora de energía eléctrica, contribuyendo a la identificación temprana de pérdidas no técnicas y a la optimización de los recursos operativos.

Resultados

En esta sección se presentan los resultados obtenidos a partir de la evaluación de los modelos CatBoost, LightGBM y XGBoost aplicados a la detección de consumos anómalos asociados a pérdidas no técnicas en un sistema de distribución de energía eléctrica. La evaluación se realizó mediante un esquema de validación cruzada k-fold (k = 5), lo que permitió estimar el desempeño de los modelos de manera robusta y reducir la dependencia de una única partición de los datos. El análisis considera métricas de desempeño predictivo y una métrica complementaria relacionada con la eficiencia computacional, con el fin de soportar la selección del modelo más adecuado.

Desempeño Predictivo de los Modelos

El desempeño de los modelos se analizó utilizando las métricas F1-score, recall y precisión, calculadas como el promedio de los cinco pliegues de la validación cruzada. Estas métricas permiten evaluar de forma integral la capacidad de los modelos para clasificar correctamente los registros, considerando tanto la detección de casos positivos como el control de errores de clasificación.

Los resultados del F1-score, presentados en la **Tabla 4**, muestran que XGBoost alcanza el mayor valor promedio (0.7572), seguido de CatBoost (0.7518) y LightGBM (0.7496). Este comportamiento indica que XGBoost logra un mejor equilibrio entre precisión y recall en la tarea de clasificación. En cuanto a la métrica de recall (**Tabla 5**), XGBoost vuelve a destacar con el valor promedio más alto (0.7777), evidenciando una mayor capacidad para identificar correctamente los casos positivos, aspecto especialmente relevante en la detección de consumos anómalos. Por su parte, LightGBM y CatBoost presentan valores ligeramente inferiores, aunque con resultados consistentes a lo largo de los pliegues.

Tabla 4

Resultados de la Validación Cruzada para Métrica de F1 Score

Modelo	F1 (M1)	F1 (M2)	F1 (M3)	F1 (M4)	F1 (M5)	Media
CGB	0.7569	0.7474	0.7579	0.7477	0.7494	0.7518
LGB	0.7591	0.7399	0.7571	0.7452	0.7467	0.7496
XGB	0.7659	0.7526	0.7616	0.7560	0.7502	0.7572

Tabla 5

Resultados de la Validación Cruzada para la Métrica de Recall

Modelo	Rec (M1)	Rec (M2)	Rec (M3)	Rec (M4)	Re (M5)	Media
CGB	0.7759	0.7706	0.7759	0.7534	0.7546	0.7661
LGB	0.7918	0.7587	0.7746	0.7580	0.7566	0.7679
XGB	0.7957	0.7792	0.7805	0.7772	0.7559	0.7777

Respecto a la precisión, los resultados consignados en la **Tabla 6** indican que CatBoost obtiene el mayor valor promedio (0.7383), seguido muy de cerca por XGBoost (0.7380), lo que sugiere un desempeño similar de ambos modelos en la reducción de falsos positivos. LightGBM presenta un valor promedio inferior en esta métrica, manteniendo no obstante un comportamiento estable durante la validación cruzada.

Tabla 6

Resultados de la Validación Cruzada para la Métrica de Precisión

Modelo	Pre (M1)	Pre (M2)	Pre (M3)	Pre (M4)	Pre (M5)	Media
CGB	0.7388	0.7255	0.7406	0.7421	0.7442	0.7383
LGB	0.7291	0.7220	0.7403	0.7329	0.7371	0.7323
XGB	0.7381	0.7227	0.7437	0.7359	0.7446	0.7380

Eficiencia Computacional

Adicionalmente al desempeño predictivo, se analizó el tiempo de entrenamiento de cada modelo como una métrica complementaria, cuyos resultados se presentan en la **Tabla 7**. Los valores obtenidos muestran que LightGBM presenta el menor tiempo de entrenamiento promedio (0.8728 segundos), seguido de XGBoost (1.1294 segundos), mientras que CatBoost registra el mayor tiempo de entrenamiento (9.7518 segundos). Estos resultados reflejan diferencias significativas en la complejidad computacional de los algoritmos y son relevantes para evaluar su viabilidad en escenarios operativos.

Tabla 7

Resultados del Tiempo de Entrenamiento de los Modelos

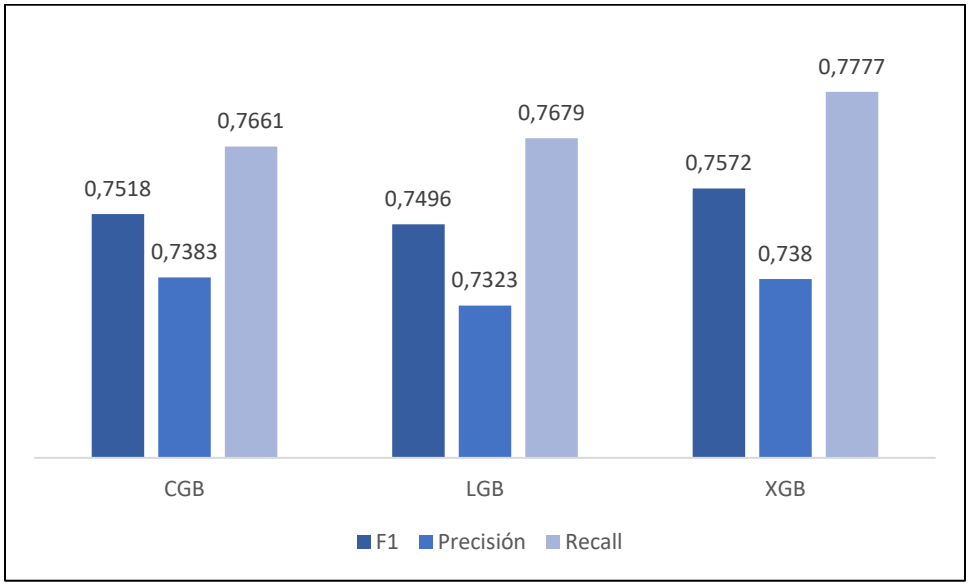
Modelo	Tiempo [s]
CGB	9.7518
LGB	0.8728
XGB	1.1294

Análisis Comparativo

Como se observa en la **Figura 3**, los tres modelos presentan un desempeño global muy cercano en F1, precisión y recall. En F1-score, los resultados son: CGB = 0,7518, LGB = 0,7496 y XGB = 0,7572, donde XGBoost alcanza el valor más alto, aunque con una diferencia marginal frente a CatBoost y LightGBM. En precisión, CatBoost registra el mayor valor (0,7383), seguido muy de cerca por XGBoost (0,7380) y LightGBM (0,7323), lo que indica un comportamiento prácticamente equivalente entre CGB y XGB para esta métrica. En recall, XGBoost obtiene el mejor desempeño (0,7777), seguido por LightGBM (0,7679) y CatBoost (0,7661). En conjunto, estas diferencias confirman que, en términos predictivos, no existe una superioridad marcada, sino variaciones pequeñas en el balance entre detección (recall) y precisión.

Figura 3

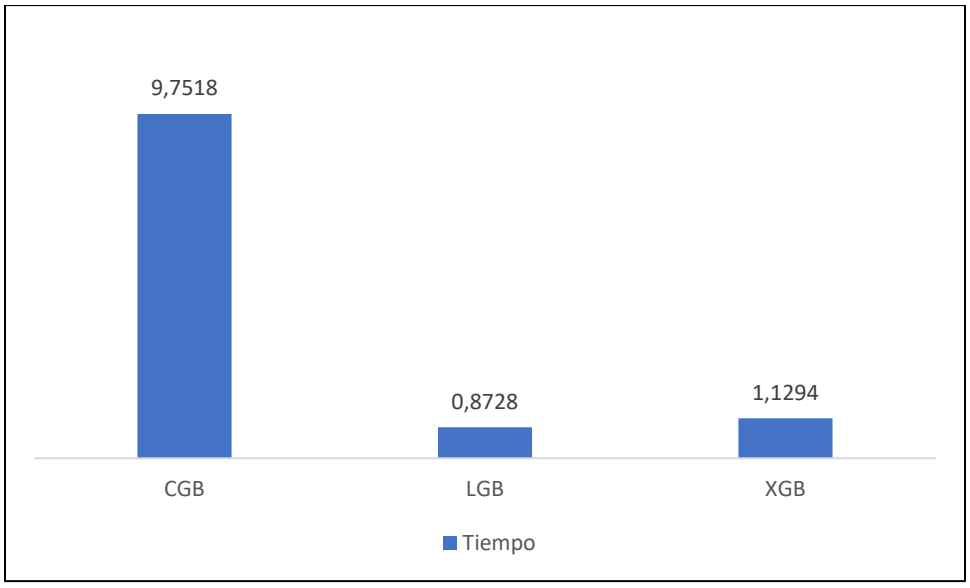
Comparación de Métricas Promedio (F1-score, Precisión y Recall) para CatBoost, LightGBM y XGBoost



No obstante, la **Figura 4** evidencia una diferencia clara en términos de eficiencia computacional: CatBoost presenta un tiempo significativamente mayor (≈ 9,7 s), mientras que LightGBM registra el menor tiempo (≈ 0,8 s) y XGBoost se ubica en un punto intermedio (≈ 1,1 s). En consecuencia, dado que las métricas de desempeño son comparables, el tiempo de entrenamiento se convierte en un criterio técnico-práctico determinante para la implementación. Bajo este enfoque, XGBoost se perfila como la alternativa más conveniente por su mejor balance entre desempeño y costo computacional, mientras que LightGBM constituye una opción preferente cuando la organización requiere ciclos de entrenamiento más ágiles y rápida iteración del modelo. Por su parte, CatBoost, aunque competitivo en desempeño, resulta menos favorable cuando el tiempo de procesamiento es un factor crítico.

Figura 4

Comparación del Tiempo de Entrenamiento (Segundos) para CatBoost, LightGBM y XGBoost



La implementación de estos modelos permite la transición de una gestión de PNT reactiva hacia un enfoque predictivo, impactando directamente en la optimización del OPEX de

la empresa. Al alcanzar una precisión cercana al 74%, el operador de red garantiza que la mayoría de las órdenes de servicio generadas culminen en el hallazgo efectivo de una anomalía, lo que reduce drásticamente el gasto en cuadrillas y logística desperdiciada en inspecciones fallidas o falsos positivos. Simultáneamente este cambio de paradigma favorece la recuperación de ingresos mediante el fortalecimiento del balance energético. El nivel de recall alcanzado por XGBoost indica que la organización podría capturar casi el 78% de los fraudes existentes en la muestra analizada, traduciéndose de forma inmediata en una reducción del índice de pérdidas y en un incremento de la energía facturada.

Conclusiones

Los resultados obtenidos demuestran que la aplicación de técnicas de limpieza de datos, balanceo de clases y generación de variables derivadas a partir de los históricos de consumo eléctrico permitió conformar un conjunto de datos adecuado para el entrenamiento de modelos de clasificación. La estabilidad observada en las métricas de desempeño a lo largo de la validación cruzada evidencia que el tratamiento previo de los datos fue determinante para garantizar la calidad del proceso de modelado y la confiabilidad de los resultados.

El análisis comparativo de CatBoost, LightGBM y XGBoost evidencia un desempeño global muy similar entre los tres modelos, con un equilibrio comparable entre la capacidad de detección y el control de falsas alarmas ($F1 \approx 0{,}75$). Esta evaluación se fundamentó en modelos ajustados mediante un muestreo y selección secuencial de los hiperparámetros de mayor impacto. Dado que no se observan diferencias sustanciales en el rendimiento predictivo, el tiempo de entrenamiento se consolida como un criterio técnico-práctico relevante para orientar la selección e implementación del modelo en un entorno operativo. En conjunto, los resultados respaldan la pertinencia de los algoritmos de gradient boosting como herramientas efectivas para apoyar la detección de consumos anómalos asociados a pérdidas no técnicas en sistemas de distribución de energía eléctrica.

Al integrar las métricas de desempeño predictivo con el análisis del tiempo de entrenamiento, se concluye que la elección del modelo no debe fundamentarse únicamente en su capacidad de clasificación, sino también en criterios de eficiencia computacional y viabilidad de implementación. Bajo este enfoque, XGBoost se perfila como la alternativa más conveniente para su aplicación práctica al mantener un rendimiento competitivo y estable, con un costo computacional equilibrado que favorece su escalabilidad. Por su parte, LightGBM constituye una

opción igualmente válida cuando la organización requiere ciclos de entrenamiento más ágiles y una rápida iteración del modelo, especialmente en escenarios donde la actualización frecuente del clasificador sea un factor prioritario para la empresa distribuidora de energía eléctrica.

Recomendaciones

Para trabajos futuros, se recomienda ampliar el conjunto de variables del análisis incorporando información técnica y operativa adicional, como eventos de mantenimiento, historial de inspecciones en campo, características de la red de distribución y variables socioeconómicas del entorno. Con ello, se busca fortalecer la capacidad explicativa del modelo y validar su utilidad práctica como herramienta de apoyo a la toma de decisiones dentro de una empresa distribuidora de energía eléctrica.

Desde la perspectiva de la organización involucrada, se sugiere evaluar la implementación del modelo seleccionado como herramienta de soporte para la toma de decisiones, integrándolo con los sistemas de información existentes. Esta integración permitiría priorizar inspecciones en campo, optimizar la asignación de recursos y reducir los costos asociados a las pérdidas no técnicas, incrementando la eficiencia operativa y maximizando el impacto de la solución desarrollada.

Como línea de profundización, se recomienda incorporar herramientas de Sistemas de Información Geográfica (GIS) al proceso de análisis, con el fin de incluir la dimensión espacial del consumo eléctrico. La integración de información georreferenciada facilitaría la identificación de patrones espaciales de consumo atípico, la detección de zonas críticas con mayor incidencia de pérdidas no técnicas y el análisis de relaciones entre el comportamiento del consumo y la ubicación dentro de la red de distribución, fortaleciendo así la capacidad analítica y predictiva de los modelos.

Referencias Bibliográficas

Almakhles, D. J., & Alharbi, A. S. (2022). *A Deep Learning Approach for Non-Technical Loss Detection in Smart Grids*. Energies, 15(1), 382.

Arias Marín, C. E., & Méndez Peñaloza, G. F. (2022). *Metodología basada en minería de datos para la detección de pérdidas no técnicas en sistemas de distribución de energía eléctrica*. Universidad Politécnica Salesiana.

Bohani, F. A., Suliman, A., Saripuddin, M., Sameon, S. S., Md Salleh, N. S., & Nazeri, S. (2021). A comprehensive analysis of supervised learning techniques for electricity theft detection. *Journal of Electrical and Computer Engineering*, 2021.

Chen, J., Chen, H., Chen, Y., & Li, C. (2021). *A Hybrid Electricity Theft Detection Method Based on Stacking Ensemble Learning*. Sensors, 21(10), 3469.

Fuster Gelabert, P. (2023). *Deep Learning Aplicado a la Detección de Robos de Energía (Trabajo de Fin de Máster)*. Universitat de les Illes Balears.

Khan, A., Ma, Y., Yan, Z., & Alahmad, M. (2020). Electricity Theft Detection in Smart Grids: A Machine Learning Approach. *IEEE Access*, 8, 140027-140040.

Li, Y., Wang, B., Zhang, R., & Liu, Q. (2023). An improved CatBoost model for electricity theft detection based on feature engineering. *Journal of Electrical Engineering & Technology*, 18(2), 793-804.

Llagua Arévalo, J. L. (2023). *Detección de pérdidas no técnicas en clientes especiales con telemedición, basada en inteligencia artificial con aplicación en la empresa eléctrica Ambato*. Escuela Politécnica Nacional.

Macao Sánchez, R. A., & Pujota Cuasapaz, E. J. (2022). *Predicción del hurto de energía eléctrica a través del uso de la inteligencia artificial mediante algoritmos de Machine*

Learning para CNEL EP Unidad de Negocios Santo Domingo. Universidad Técnica Estatal de Quevedo.

Majumder, S., Majumdar, A., & Das, S. (2022). Ensemble Learning Based Non-Technical Loss Detection in Smart Grids. *IEEE Transactions on Smart Grid*, 13(1), 393-402.

Moreno Gallón, D. A. (2022). *Análisis de los registros de un operador de red eléctrica nacional para la búsqueda de pérdidas no técnicas (Proyecto de grado para aspirar al título de Ingeniero Electricista)*. Universidad Tecnológica de Pereira.

Orozco-Castañeda, P., Rodriguez-Resendiz, J., & Sanchez-Mendiola, J. I. (2020). Detection of non-technical losses in smart grids using machine learning algorithms. *Energies*, 13(22), 5940.

Sharma, S., & Majumdar, A. (2020). Unsupervised detection of non-technical losses via recursive transform learning. *IEEE Transactions on Power Delivery*.

Singh, P., & Tiwari, S. K. (2024). A review on electricity theft detection techniques using machine learning and deep learning in smart grid. *Sustainable Energy Technologies and Assessments*, 55, 102928.

Song, Y., Fan, R., Yang, H., & Liu, C. (2021). An improved XGBoost model for non-technical loss detection in power grids. *Applied Energy*, 290, 116744.

Sun, X., Huang, D., Li, Q., & Xu, Y. (2022). Non-technical loss detection in smart grids using federated learning and ensemble methods. *Applied Energy*, 314, 118933.

Tang, Y., Luo, S., Tan, S., & Li, C. (2023). Electricity Theft Detection in Smart Grid Using a Graph Convolutional Network and Attention Mechanism. *IEEE Transactions on Smart Grid.*

Tonato, M., Mazón, A., & Asanza, V. (2023). Clasificación de Pérdidas No Técnicas Basadas en el Aprendizaje Automático en Macrodatos. *Revista Politécnica*.

UPME. (2023). Informe de gestión del sector energético colombiano: Desafíos en la reducción de pérdidas no técnicas.

UPME. (2022). Estudio de optimización de redes de distribución para la mitigación de pérdidas en Colombia.

UPME. (2021). Análisis de la eficiencia operativa en la distribución de energía eléctrica: Foco en indicadores de pérdidas.

UPME. (2020). Boletín estadístico minero-energético: Incidencia de las pérdidas no técnicas en el balance energético nacional.

UPME. (2024). Propuestas para la modernización de la medición y la lucha contra el fraude eléctrico en Colombia.

Wang, B., Zhang, S., Li, H., & Wang, Y. (2021). Non-technical loss detection in smart grid using a novel hybrid deep learning model. *Journal of Cleaner Production*, 295, 126462.

Zhou, Q., Li, W., Sun, Y., & Han, M. (2024). A comprehensive review of non-technical loss detection in smart grids: Challenges, methods, and future directions. *Energy Reports*, 10, 438-456.