Anomaly Detection System for Electrical Measurements Using Machine Learning

Gamarra, Sergio[1], Paz Fabiola [1] y Cognetta Analía [1]

[1] Facultad de Ingeniería, Universidad Nacional de Jujuy, Argentina
37730337@fi.unju.edu.ar

Abstract. In the electricity industry, distributors face a major challenge: identifying and reducing non-technical losses—that is, energy supplied but not billed due to fraud, unauthorized connections, or meter malfunctions. These losses affect the profitability of distributors, the stability of the power grid, and fairness among users. In Argentina, provincial distributors are responsible for the operation, maintenance, and metering of the electricity supply in each region. Within their structure, Energy Control departments must monitor consumption, analyze deviations, and coordinate inspections to detect irregularities. The incorporation of machine learning techniques offers an innovative and more effective solution to address this problem, as they can automate the identification of suspicious cases, reduce analysis time, and improve the accuracy of detecting anomalous consumption. This paper presents a system for detecting anomalies in electricity consumption based on machine learning techniques: LightGBM, and XGBoost, capable of processing large volumes of historical data and identifying complex patterns in consumption series. An experimental study is conducted to compare these techniques and select the one that best detects situations deviating from typical consumption behavior. The best model is integrated into an interactive web prototype that allows users to view and analyze results, thereby increasing the operational efficiency of the energy control department and improving its ability to respond to fraud or technical failures in the meters.

Keywords: Machine Learning, Anomaly Detection, Fraud Detection; LightGBM, XGBoost.

Sistema de Detección de Anomalías en Mediciones Eléctricas mediante Aprendizaje Automático

Resumen. En la industria eléctrica, las distribuidoras enfrentan un gran desafío: identificar y reducir las pérdidas no técnicas, es decir, la energía suministrada pero no facturada debido a fraudes, conexiones irregulares o fallas en medidores. Estas pérdidas afectan la rentabilidad de las distribuidoras, la estabilidad del sistema eléctrico y la equidad entre los usuarios. En Argentina, las distribuidoras provinciales son responsables de la operación, mantenimiento y medición del suministro eléctrico en cada región. Dentro de su estructura, las áreas de Control de Energía deben monitorear consumos,

analizar desviaciones y coordinar inspecciones para detectar irregularidades. La incorporación de técnicas de aprendizaje automático ofrece una solución innovadora y más eficaz para enfrentar esta problemática, dado que pueden automatizar la identificación de casos sospechosos, reducir el tiempo de análisis y mejorar la precisión en la detección de consumos anómalos. En este trabajo, se presenta un sistema para la detección de anomalías en el consumo de energía eléctrica basado en técnicas de aprendizaje automático: LightGBM y XGBoost, capaces de procesar grandes volúmenes de datos históricos, identificar patrones complejos en las series de consumo. Se realiza un estudio experimental para comparar estas técnicas y elegir la que mejor detecte situaciones que se desvíen del comportamiento de consumo habitual. El mejor modelo se integra en un prototipo web interactivo que permite visualizar y analizar los resultados, aumentando la eficiencia operativa del área de control de energía y mejorando su capacidad de respuesta ante fraudes o fallas técnicas en los medidores.

Palabras clave: Aprendizaje Automático, Detección de Anomalías, Detección de Fraudes; LightGBM, XGBoost.

1 Introducción

La Empresa Jujeña de Energía S.A. (EJESA), encargada de distribuir energía eléctrica en la provincia de Jujuy, Argentina, presta servicio desde 1996 a una amplia base de usuarios residenciales, comerciales e industriales (EJESA, 2022). La empresa cuenta con aproximadamente 250000 clientes vigentes, de los cuales solo una pequeña proporción (alrededor de 1.500) corresponde a tarifas T2 y T3 (usuarios de mediana y gran demanda). Aunque estos clientes consumen entre el 20 % y 25 % de la energía total, se dispone de planes de inspección específicos para ellos. Incluso, en una de las administraciones, se realizan inspecciones conjuntas con la toma mensual de lectura. El resto de los usuarios pertenece a la tarifa T1, correspondiente a consumos residenciales y pequeños comercios. La mayoría de la energía medida corresponde a estos usuarios, lo que dificulta cualquier tipo de revisión manual caso por caso debido a su volumen.

Hoy en día, las inspecciones provienen de diversas fuentes, como las observaciones realizadas por los lecturistas durante la toma de lectura, las denuncias de usuarios o terceros, los análisis manuales de consumos históricos mediante hojas de cálculo con macros y las revisiones visuales de curvas de consumo. Este enfoque tradicional exige mucho tiempo y recursos humanos, limitado la capacidad de análisis frente al creciente volumen de datos generados por los sistemas de medición. Uno de los desafíos más importantes que enfrenta EJESA es detectar de manera eficiente y precisa las irregularidades en el consumo eléctrico, ya sea por situaciones de fraude o por fallas en los medidores, que impiden un correcto registro del consumo. Ambos escenarios impactan negativamente en la economía de la empresa, afectan la equidad entre usuarios y reducen la eficiencia del sistema.

Frente a esta situación, el uso de técnicas de aprendizaje automático se presenta como una alternativa prometedora. Estas herramientas son capaces de procesar grandes volúmenes de datos históricos, identificar patrones complejos y crear modelos predictivos capaces de detectar posibles consumos irregulares con

mayor precisión. Su efectividad ha sido ampliamente probada en otros sectores, como las finanzas, telecomunicaciones y salud, especialmente en tareas de detección de fraudes y anomalías (Lee et al., 2020; Shalev Shwartz & Ben-David, 2014).

Por lo tanto, en este trabajo se plantea el desarrollo de un sistema de detección de consumos irregulares que utilice modelos de aprendizaje automático. El objetivo principal es detectar posibles casos de fraude o fallas en medidores. A su vez, se propone el desarrollo de un prototipo web que permita al personal de EJESA interactuar con el modelo predictivo de forma simple, rápida y eficaz.

La estructura del trabajo es la siguiente: en la Sección 2 se abordan los conceptos fundamentales de las técnicas de aprendizaje automático; en la Sección 3 se detalla el prototipo del sistema; en la Sección 4 se describen los experimentos y se discuten los resultados. Por último, en la Sección 5 se presentan las conclusiones del trabajo.

2 Conceptos Fundamentales

2.1 Aprendizaje Automático

El aprendizaje automático es una rama de la inteligencia artificial (IA) que permite a los sistemas aprender a partir de los datos sin ser programados explícitamente para cada tarea. Su objetivo es desarrollar algoritmos capaces de identificar patrones y realizar predicciones basadas en la experiencia previa (Shalev-Shwartz & Ben-David, 2014). En este sentido, combina conceptos de estadística, probabilidad y optimización matemática para construir modelos que generalicen comportamientos observados.

Según Mitchell (1997), el aprendizaje automático es un programa informático que "aprende" de la experiencia E respecto a una clase de tareas T y una medida de desempeño P, si su desempeño en T (medido por P) mejora con la experiencia E. Esta definición es ampliamente aceptada y resume la esencia del aprendizaje a partir de datos.

Los modelos seleccionados para el trabajo son los modelos basados en árboles de decisión que ofrecen distintos enfoques para abordar problemas de clasificación binaria como la detección de irregularidades en consumos eléctricos.

Gradient Boosting Machine (GBM) propuesto por Friedman (2001), GBM pertenece a la familia de los métodos de boosting. Construye árboles secuencialmente, donde cada nuevo árbol corrige los errores de los anteriores. Este enfoque minimiza una función de pérdida mediante gradientes, ajustando los modelos hasta alcanzar un equilibrio entre sesgo y varianza (Haykin, et al., 2009).

Extreme Gradient Boosting (XGBoost) es una implementación optimizada del método Gradient Boosting, desarrollada por Chen y Guestrin en el año 2016. Su objetivo principal es mejorar la eficiencia computacional y la capacidad de generalización del modelo mediante varias optimizaciones técnicas. Entre sus principales características se encuentran: La incorporación de términos de regularización L1 (Lasso) y L2 (Ridge) en la función objetivo para controlar la complejidad del modelo y prevenir el sobreajuste; Un algoritmo sparsity-aware que

gestiona eficientemente los valores faltantes y datos dispersos; El uso de paralelización en la construcción de árboles y una estrategia de weighted quantile sketch para el manejo de grandes volúmenes de datos.

Gracias a estas mejoras, XGBoost se ha consolidado como uno de los algoritmos más utilizados en la práctica para problemas de clasificación y regresión estructurados, destacándose por su precisión y velocidad (Chen et al., 2016).

Light Gradient Boosting Machine (LightGBM) es una implementación moderna del algoritmo Gradient Boosting desarrollada por Microsoft Research (Ke et al., 2017). A diferencia de XGBoost, LightGBM emplea un enfoque basado en histogramas para acelerar la búsqueda de divisiones óptimas y reducir el uso de memoria. Además, utiliza una estrategia de crecimiento denominada leaf-wise, en la cual los nodos con la mayor reducción de pérdida se dividen primero. Esta técnica permite un ajuste más fino y eficiente en comparación con el crecimiento nivelado (level-wise) utilizado por otros algoritmos de boosting.

LightGBM también ofrece soporte nativo para variables categóricas y paralelización distribuida, lo que lo convierte en una herramienta escalable para grandes conjuntos de datos. Sin embargo, el crecimiento leaf-wise puede aumentar el riesgo de sobreajuste si no se controla la profundidad máxima de los árboles (Ke et al., 2017).

2.2 Metodología CRISP-DM

La metodología Cross Industry Standard Process for Data Mining (CRISP-DM) es el estándar más utilizado a nivel internacional para estructurar proyectos de minería de datos y aprendizaje automático. Su éxito radica en su carácter flexible e iterativo, que permite adaptar cada fase a las necesidades del proyecto y mantener la trazabilidad entre los objetivos de negocio y los resultados técnicos (Galán Cortina, 2016). Según Chapman et al. (2000), CRISP-DM se caracteriza por su independencia de herramientas y su aplicabilidad en distintos dominios, lo que favorece su adopción tanto en entornos académicos como industriales. Asimismo, diversos autores destacan su vigencia frente a otras metodologías como SEMMA o KDD, debido a su mayor claridad en la comunicación entre equipos técnicos y áreas de gestión (Wirth et al., 2000; Galán Cortina, 2016).

Las fases de estas metodologías son las siguientes: 1- Comprensión del negocio (Business Understanding), 2- Comprensión de los datos (Data Understanding); 3- Preparación de los datos (Data Preparation), 4- Modelado (Modeling), 5- Evaluación (Evaluation) , y 6- Despliegue (Deployment).

2.3 Herramientas de Desarrollo y Entorno Tecnológico

En este trabajo el lenguaje Python se consolidó como el estándar en ciencia de datos por su sintaxis simple, su ecosistema maduro y la gran cantidad de bibliotecas disponibles (McKinney, 2010). Entre las principales bibliotecas utilizadas en esta tesis se destacan:

- pandas: para manipulación y análisis de datos estructurados.

- scikit-learn: para preprocesamiento, división de datos, métricas y modelado base (Géron, 2022).
- TSFEL: para la extracción automática de características estadísticas, temporales y espectrales a partir de series de consumo eléctrico (Barandas et al., 2020).
- Optuna: para la optimización de hiperparámetros mediante búsqueda bayesiana y poda temprana (Akiba et al., 2019).
- XGBoost y LightGBM: bibliotecas de gradient boosting altamente eficientes, diseñadas para tareas de clasificación binaria y datasets de gran volumen (Chen et al., 2016; Ke et al., 2017).

El backend se desarrolló con Flask, un microframework de Python para servicios RESTful, y el frontend con React y Tailwind CSS, garantizando un diseño modular y adaptable a futuras integraciones con sistemas corporativos.

3 Sistema Propuesto

En esta sección se describe el proceso de desarrollo del sistema para la detección de anomalías en el consumo de energía eléctrica. Para ello, se llevan a cabo las diferentes fases de la metodología.

3.1 Comprensión de los datos

Los datos de esta investigación fueron provistos por el área de Control de Energía de EJESA e incluyen historial de consumo y datos administrativos de servicios eléctricos activos. La muestra abarca clientes de la categoría tarifaria T1 (T1R residenciales, T1G comercios pequeños, T1TS residenciales con tarifa social), que concentra la mayoría de los usuarios de la empresa. Las principales fuentes de datos consideradas fueron:

- Historial mensual de consumo eléctrico (25 meses previos a la inspección).
- Información comercial del servicio (estado, grupo tarifario, sucursal, etc.).
- Observaciones del personal de lectura en la última toma de lectura.
- Registros de consumo de agua, cuando estaban disponibles.

La incorporación de los registros de consumo de agua responde al esquema de facturación vigente en la provincia de Jujuy. De acuerdo con el marco regulatorio de la actividad eléctrica provincial y el contrato de concesión de EJESA, la facturación del servicio eléctrico se realiza de manera unificada con los servicios de agua potable y saneamiento (Provincia de Jujuy, 1996; Provincia de Jujuy y Empresa Jujeña de Energía S.A., 1996). En consecuencia, los sistemas comerciales de la empresa disponen, para determinados usuarios, de información asociada al consumo de agua. En este trabajo, dicha información se incorpora como una variable contextual complementaria, ya que puede aportar señales indirectas sobre el nivel de ocupación o actividad del inmueble y contribuir al análisis del comportamiento del consumo eléctrico.

El conjunto de datos consolidado contiene 51.834 registros, cada uno correspondiente a un servicio eléctrico. Cada fila combina variables numéricas como consumos mensuales, consumos de agua y la variable objetivo *target*. La variable *target* indica si un servicio fue clasificado como irregular, ya sea por fraude o por desperfecto en el medidor, lo que implica la presencia de un consumo que no fue bien facturado. Las variables categóricas incluyen el estado del servicio o sucursal, grupo tarifario (T1R, T1G y T1TS), estado del servicio y observaciones del lecturista.

Los servicios regulares mantienen un consumo promedio más alto y estable a lo largo del tiempo, con un patrón de estacionalidad evidente. En cambio, los servicios irregulares muestran una caída progresiva en el consumo, con quiebres abruptos que podrían reflejar manipulaciones, desconexiones o fallas en el medidor. Este análisis comparativo de las variables numéricas muestra diferencias consistentes entre ambas clases.

Se revisaron los valores faltantes de cada variable. Esta evaluación es fundamental para detectar posibles problemas de calidad que puedan afectar el rendimiento del modelo y que deberán ser corregidos durante el preprocesamiento. Los mayores porcentajes de valores faltantes se observan en las variables asociadas al consumo de agua (9,44 % y 9,27 %). Esta situación se debe a que, si bien la factura de energía eléctrica y agua se encuentra unificada para determinados usuarios, no todos los servicios eléctricos cuentan con información disponible de consumo de agua. Así, estos nulos reflejan la estructura del sistema y no errores. En los consumos eléctricos, los nulos se concentran en los meses más antiguos (consumo del mes 8 al 25) porque algunos servicios son recientes y aún no completan los 25 meses de historial. En estos casos, el porcentaje de nulos se mantiene por debajo del 4 % y se tratará durante el preprocesamiento. La variable de observaciones del lecturista presenta solo 7 nulos (0,01 %), que también se considerarán en la limpieza para mantener la integridad del conjunto.

3.2 Preprocesamiento de Datos

Esta fase implica una serie de transformaciones destinadas a mejorar la calidad del conjunto de datos, corregir inconsistencias detectadas y construir nuevas variables que puedan aportar valor predictivo. En particular, se abordan los siguientes aspectos clave:

- División del conjunto de datos para entrenamiento y prueba: se realizó una división del conjunto de datos en dos subconjuntos: uno para entrenamiento (80 %) y otro para prueba (20 %). En total, se obtuvieron 41.467 registros para entrenamiento y 10.367 registros para prueba.
- Tratamiento de valores faltantes: se optó por reemplazar los valores nulos con la media del mismo registro, calculada sobre los consumos válidos de ese servicio. De esta manera, se mantiene el perfil individual de consumo sin introducir sesgos externos.
- Limpieza y estandarización de variables categóricas: se aplicó conversión de texto a mayúsculas; eliminación de tildes y caracteres especiales; corrección ortográfica básica; y unificación de etiquetas con significado equivalente.
- Generación de variables derivadas: se generaron nuevas variables que representan mejor el comportamiento de consumo de los clientes. Estas

nuevas variables enriquecen la capacidad del modelo para identificar patrones fuera de lo habitual.

- Selección de atributos relevantes para el modelo: mediante label encoding y one-hot encoding, se realizó una selección de variables para identificar las de mayor capacidad predictiva, eliminando atributos redundantes y variables irrelevantes.
- Balanceo de la clase minoritaria (target = 1): se utiliza un parámetro que aumenta el peso de la clase minoritaria durante el entrenamiento.

3.3 Modelado

En esta sección se entrenan dos modelos supervisados con configuraciones por defecto y se enfrentan en una primera comparación directa. A partir de sus resultados en Precisión, Recall, F1-Score y Accuracy, se elige el de mejor desempeño. Este modelo, el de mayor poder predictivo, será puesto a prueba en un entorno operativo real en la sección siguiente. Dado que es un problema de clasificación binaria con fuerte desbalance entre clases (la mayoría de los servicios son normales), se adopta como métrica principal el F1-score, que equilibra precisión y recall, es decir, falsos positivos y falsos negativos.

Se seleccionaron LightGBM y XGBoost por su buen desempeño en clasificación con fuerte desbalance, demostrado en la detección de fraude. Para el entrenamiento se empleó una configuración experimental orientada a optimizar hiperparámetros y maximizar el rendimiento y su capacidad de generalización. Se utilizó Optuna, una biblioteca de optimización basada en Tree-structured Parzen Estimator (TPE), dado que estudios recientes (Akiba et al., 2019) muestran que estos enfoques encuentran configuraciones competitivas en menos tiempo computacional y son adecuados para problemas con múltiples hiperparámetros y clases desbalanceadas. La configuración incluyó:

- Número de iteraciones: 50 por modelo.
- Validación cruzada estratificada de 5 pliegues.
- Métrica de optimización: F1_score para la clase 1 (servicio irregular).
- Espacio de búsqueda definido manualmente, considerando parámetros sensibles como *n_estimators, learning_rate, max_depth, num_leaves y scale_pos_weight* (en LightGBM) *o max_depth, learning_rate, min_child_weight y subsample* (en XGBoost), entre otros.

Los resultados obtenidos se resumen en la Tabla 1. Los modelos LightGBM y XGBoost tuvieron rendimientos similares, con un F1-CV promedio cercano a 0,81 y buen F1 para la Clase 1. XGBoost fue levemente el más equilibrado para esta clase, con F1-CV = 0,81, pero con una precisión de 89,3 % y recall de 71,2 %, detectando muchos casos positivos sin generar demasiados falsos.

Tabla 1 - Comparación de métricas de desempeño de los modelos para el target=1 (clase 1).

Modelo	F1 CV Promedio	F1 (Test)	Precision	Recall	Accuracy
LightGBM (optuna)	0,8089	0,7911	0,9084	0,7006	0,9874
XGBoost (optuna)	0,8122	0,7925	0,8936	0,7119	0,9873

Además del análisis cuantitativo, se realizó un análisis visual del comportamiento de los modelos entrenados. Se utilizaron las siguientes métricas: Curva AUC ROC (Receiver Operating Characteristic), Curva Precision-Recall (AP) y Matriz de confusión.

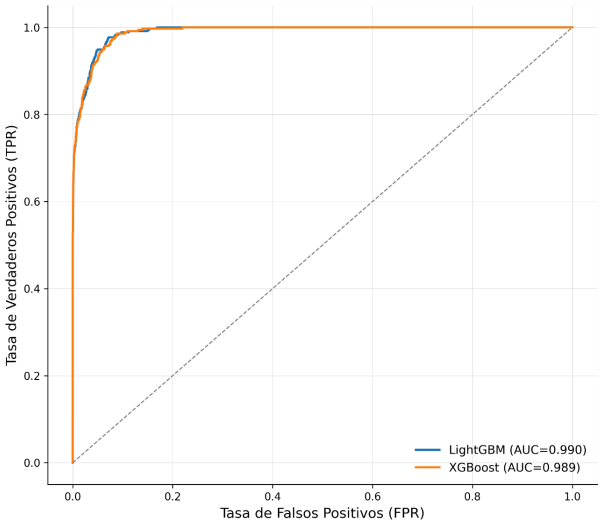

Fig. 1 - Curvas ROC de los modelos.

El análisis de las curvas ROC indica que ambos modelos tienen un desempeño excelente, con AUC cercanos a 1 y gran capacidad de discriminación. LightGBM logra un AUC de 0.990 y XGBoost de 0.989; aunque LightGBM tiene una ligera ventaja, la diferencia es mínima y su rendimiento es comparable. Por ello, la elección del modelo puede basarse en criterios adicionales, como costo computacional o estabilidad del entrenamiento. El análisis de la curva Precision–Recall muestra un buen desempeño de ambos modelos en la detección de la clase minoritaria, con Average Precision (AP) de 0.858 para XGBoost y 0.864 para LightGBM. Aunque LightGBM logra un valor ligeramente mayor, la diferencia es pequeña, lo que indica un comportamiento similar. Estos resultados coinciden con los de la curva ROC y confirman la robustez de ambos enfoques en la clasificación.

Las matrices de confusión de la Fig. 2 muestran cuántos servicios se clasificaron bien y mal. Con LightGBM se obtienen 9988 verdaderos negativos, 106 falsos positivos, 25 falsos negativos y 248 verdaderos positivos. XGBoost presenta 9983 verdaderos negativos, 102 falsos positivos, 30 falsos negativos y 252 verdaderos positivos. XGBoost detecta ligeramente más anomalías y reduce algo las falsas alarmas, mientras que LightGBM deja pasar menos anomalías sin detectar. En general ambos modelos logran muchos verdaderos positivos (fraudes o medidores dañados detectados) y muy pocos falsos positivos (inspecciones innecesarias).

Dado que el área de Control de Energía funciona como una unidad de negocio independiente con recursos limitados, es clave equilibrar dos riesgos:

- Falsos negativos (irregularidades no detectadas): generan pérdidas por energía no facturada que no puede recuperarse.

- Falsos positivos (inspecciones innecesarias): implican costos logísticos y de personal sin ingresos asociados, reduciendo la rentabilidad.

En tareas de detección de anomalías la minimización de falsos negativos resulta crítica, se observa que el modelo LightGBM muestra una ligera ventaja en este aspecto, aunque ambos modelos presentan un desempeño general comparable.

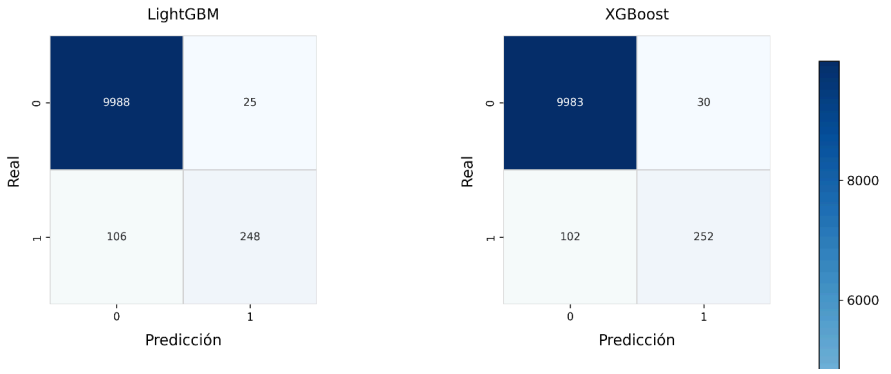

Fig. 2 - Matrices de confusión de los modelos optimizados.

3.4 Evaluación Operativa y Despliegue del Modelo Predictivo

Se evalúa el desempeño del modelo predictivo desarrollado, se valida que cumpla con los objetivos del negocio y su integración dentro de un entorno web funcional. El propósito es analizar el comportamiento del modelo en un contexto operativo real, compararlo con la herramienta que utilizan actualmente STD (llamado Sistema de Toma de Decisiones) para la detección de fraudes y medidores dañados, y describir su implementación dentro del prototipo web diseñado para el área de Control de Energía de EJESA.

3.5 Evaluación con datos operativos

Para validar la utilidad del modelo en un contexto real, se analizaron los resultados de las inspecciones efectivamente realizadas entre marzo y agosto de 2025, clasificadas según la herramienta que originó su detección: el modelo de aprendizaje automático (XGBoost) y la herramienta tradicional de análisis STD.

A nivel general, el modelo basado en aprendizaje automático alcanzó una eficacia del 20,11%, frente al 13,37% obtenido por STD, lo que representa una mejora relativa del 51 % en la capacidad de detección de casos positivos. Además, el modelo permitió incrementar el volumen total de inspecciones (1.711 frente a 1.092) sin sacrificar la tasa de aciertos. Los resultados por administración se presentan en la Tabla 2

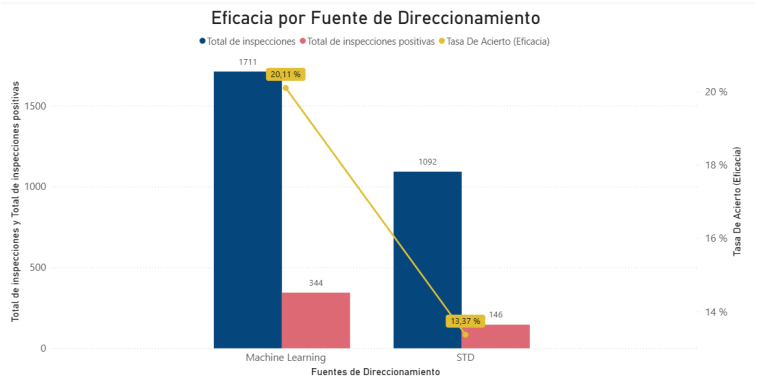

Fig. 3 - Cantidad de inspecciones realizadas y eficacia por herramientas

Tabla 2 - Resultados operativos por administración

Administración	Herramienta	Inspecc. Totales	Inspecc Positivas	Eficacia %
San Salvador	Modelo de Aprendizaje Automático	739	80	10,83
	STD	279	25	8,96
Valles y Yungas	Modelo de Aprendizaje Automático	972	264	27,16
	STD	735	120	16,33

En términos regionales, la administración Valles y Yungas registró los mejores resultados operativos, con una eficacia del 27,16%, seguida por San Salvador con un 10,83%. Si bien estas variaciones se explican por las particularidades socioeconómicas y operativas de cada zona, en ambos casos se verifica una clara superioridad del enfoque basado en aprendizaje automático frente al método tradicional.

3.6 Analisis e interpretacion de resultados

El aumento simultáneo del volumen de inspecciones y de la eficacia alcanzada por el modelo predictivo constituye un hallazgo de gran relevancia para la gestión operativa del área de Control de Energía. El enfoque basado en aprendizaje automático logró una eficacia del 20,11%, frente al 13,37% de la herramienta tradicional, acompañado además de un mayor número de inspecciones totales.

4 .Despliegue del modelo en un prototipo funcional

El prototipo permite ejecutar el modelo fácilmente, sin conocimientos técnicos ni de programación, y convierte un proceso antes manual y distribuido en una herramienta centralizada, accesible y automatizada. Con el prototipo web, el proceso se simplifica y descentraliza: los supervisores descargan los archivos TXT de su sucursal desde GESCO INFORMES y los cargan directamente en el sistema, sin pasos intermedios. Estos TXT no incluyen todos los servicios de la provincia, sino que están segmentados por sucursal para evitar archivos demasiado grandes.

El sistema ofrece una solapa para cargar los TXT generados por GESCO INFORMES. En esta etapa: el usuario selecciona el archivo de su sucursal; el backend lo procesa automáticamente; se validan formatos y estructuras para su análisis. El usuario ingresa a la solapa de Análisis de Consumo, donde se muestra, de forma clara y centralizada, toda la información de los servicios junto con la proyección generada por el modelo predictivo.

Esta solapa, ilustrada en la Fig. 4 , ofrece: - Filtros dinámicos por tarifa, localidad, administración, estado del servicio y nivel de probabilidad de irregularidad, para acotar el análisis en segundos. - Tarjetas individuales por servicio, que permiten un análisis rápido, visual e intuitivo de cada caso.

Cada tarjeta de la Fig. 5 incluye: Información comercial clave del servicio (tarifa, dirección, localidad y número de servicio). Un gráfico de consumo eléctrico de los últimos 25 meses, más los dos últimos consumos de agua, ideal para detectar quiebres, caídas abruptas o patrones de consumo atípicos de un solo vistazo. La probabilidad estimada de irregularidad, calculada automáticamente por el modelo predictivo. Un botón de acceso directo a Google Maps, que permite ubicar con precisión geográfica el servicio en el mapa.

Las métricas se pueden observar en las Figs 6-7 y permiten comprender el comportamiento global del modelo y brindan transparencia sobre su nivel de confiabilidad para la toma de decisiones.

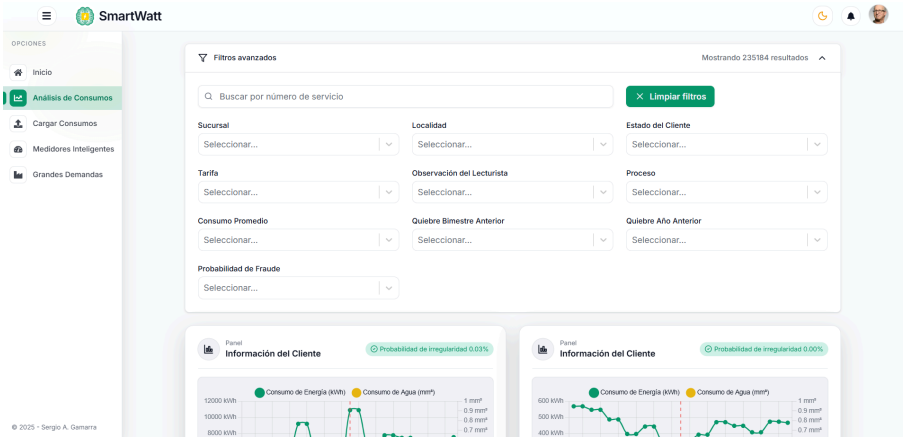

Fig. 4 - Pantalla de Análisis de Consumo con filtros dinámicos y datos de los servicios

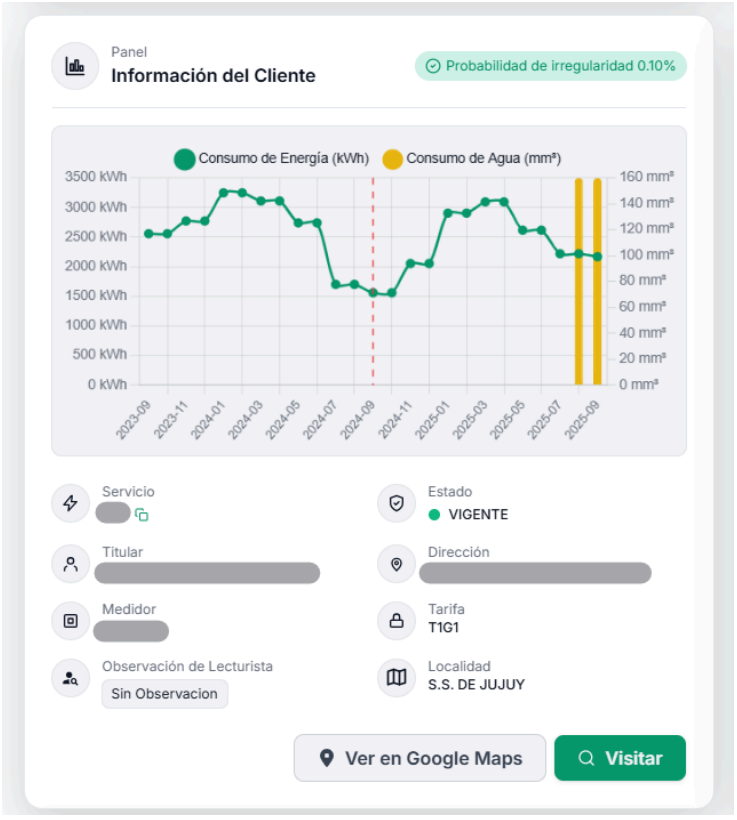

Fig. 5 - Vista de tarjeta de servicio con consumo, ubicación y probabilidad

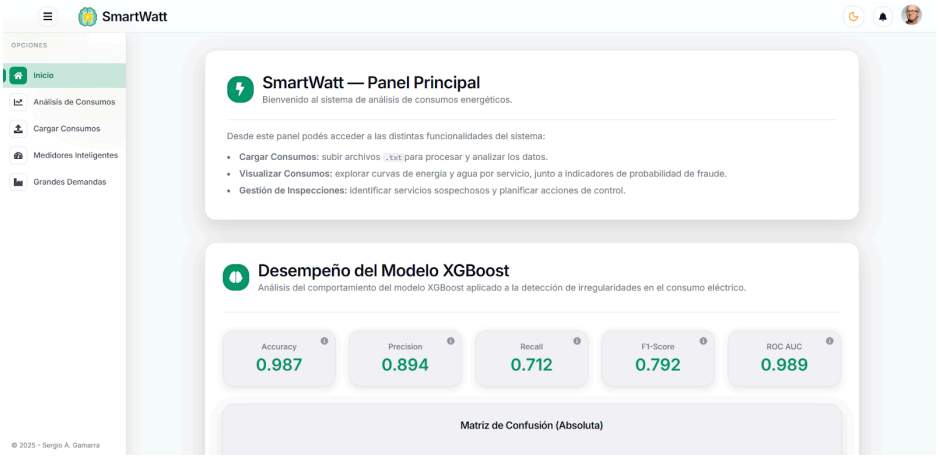

Fig. 6. Pantalla de Inicio con las métricas generales del modelo

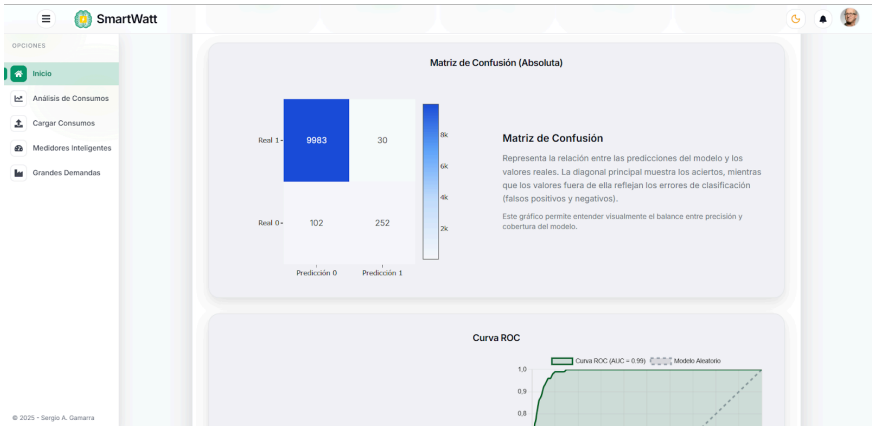

Fig. 7. Pantalla de Inicio con métricas del modelo y curvas ROC y Precision-Recall

5 Conclusiones

Esta investigación y desarrollo surgen de una necesidad de negocio concreta: optimizar las inspecciones, mejorar la asignación de recursos operativos y reducir las pérdidas no técnicas de la distribuidora eléctrica. La metodología CRISP-DM permitió abordar el problema de forma ordenada, integrando conocimiento del negocio, análisis y preparación de datos, modelado, evaluación y despliegue, y concretando una solución de aplicación real en la empresa. El desbalance de clases y la complejidad de los datos exigieron tareas de limpieza, tratamiento de valores faltantes, estandarización de variables categóricas y creación de nuevas variables a partir de historiales de consumo eléctrico y de agua, mejorando

la calidad del dataset y habilitando un modelado robusto. La evaluación con datos reales de inspecciones validó el impacto del modelo en el negocio. Los resultados muestran que el enfoque de aprendizaje automático supera ampliamente a la herramienta tradicional (STD), con una mejora relativa del 51% en la eficacia de detección, incluso con más inspecciones, confirmando su solidez técnica y utilidad operativa y económica.

El prototipo web funcional integró el modelo en una herramienta para el área de Control de Energía, permitiendo cargar datos reales, visualizar probabilidades de irregularidad, analizar historiales de consumo, acceder a información comercial y geolocalizar servicios, lo que facilita la planificación de inspecciones y la toma de decisiones diaria.

Se demuestra así que las técnicas de aprendizaje automático mejoran sustancialmente la detección de fraudes y medidores dañados, optimizan el uso de recursos operativos y reducen las pérdidas no técnicas, convirtiendo grandes volúmenes de datos históricos en información útil, accionable y orientada al negocio, fortaleciendo los procesos de control y aportando valor real a la organización.

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. En Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 2623–2631). ACM. https://doi.org/10.1145/3292500.3330701.

Barandas, M., Folgado, D., Bota, P., Santos, S., Abreu, M., Silva, C., & Gamboa, H. (2020). TSFEL: Time Series Feature Extraction Library. SoftwareX, 11, 100456. https://doi.org/10.1016/j.softx.2020.1004.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Wirth, R., Shearer, C. (2000). CRISP-DM 1.0: Step-by-step data mining guide. SPSS Inc.

Chen, T., Guestrin, C. (2016). XGBoost: A scalable tree boosting system. En Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). ACM. https://doi.org/10.1145/2939672.2939785.

Chiang, J.Y., Ouyang, C.L., Lee, J.Y., Chang, C.H. (2024). Binary classification with imbalanced data: A comparative study using resampling techniques and machine learning models. Entropy, 26(1), 15. https://doi.org/10.3390/e26010015.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189–1232.

Galán Cortina, C. (2016). Minería de datos: Metodología CRISP-DM aplicada a la gestión de información educativa [Trabajo fin de máster, Universidad de Oviedo]. Repositorio de la Universidad de Oviedo.

Géron, A. (2022). Hands-on machine learning with Scikit-Learn, Keras and TensorFlow (3rd ed.). O'Reilly Media.

Haykin, S. (2009). Neural networks and learning machines (3rd ed.). Prentice Hall.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. En Advances in Neural Information Processing Systems (Vol. 30, pp. 3146–3154).

McKinney, W. (2010). Data structures for statistical computing in Python. En S. van der Walt & J. Millman (Eds.), Proceedings of the 9th Python in Science Conference (pp. 51–56).
Mitchell, T. M. (1997). Machine learning. McGraw-Hill.
Quinlan, J. R. (1986). Induction of decision trees. Machine Learning, 1(1), 81–106.
Rokach, L., & Maimon, O. (2014). Data mining with decision trees: Theory and applications (2nd ed.). World Scientific Publishing.
Shalev Shwartz, S., Ben David, S. (2014). Understanding machine learning: From theory to algorithms. Cambridge University Press.
Wirth, R., Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. En Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining (pp. 29–39).
Provincia de Jujuy. (1996). Ley N.º 4888. Marco Regulatorio de la Actividad Eléctrica de la Provincia de Jujuy. Legislatura de la Provincia de Jujuy.
Provincia de Jujuy y Empresa Jujeña de Energía S.A (1996) Contrato de concesión del servicio público de distribución de energía eléctrica de la Provincia de Jujuy. Documento institucional.