Detección de Anomalías y Valores Atípicos en Sistemas de Reconocimiento de Patrones: Consumo de Energía Eléctrica de Electro Puno S.A.A.

1st Edgar Jeferson Cusihuaman Garate
Ing. Estadística e Informática
Universidad Nacional del Altiplano
Puno, Perú
jefersoncg68@gmail.com

2nd Edilberto Wilson Mamani Emanuel
Ing. Estadística e Informática
Universidad Nacional del Altiplano
Puno, Perú
edilemanuel2000@gmail.com

3rd Fred Torres Cruz
Ing. Estadística e Informática
Universidad Nacional del Altiplano
Puno, Perú
ftorres@unap.edu.pe

***Resumen*—Este artículo presenta un análisis de detección de anomalías en el consumo de energía eléctrica utilizando datos del mes de enero de 2024, comprendiendo 343,446 registros de clientes de Electro Puno S.A.A. Se implementó el algoritmo Isolation Forest con técnicas estadísticas de reconocimiento de patrones, utilizando 20 características seleccionadas y un parámetro de contaminación del 5 % para la detección automatizada de valores atípicos. El modelo procesó 343,434 registros válidos tras la limpieza de datos, identificando 17,168 anomalías significativas (tasa del 5.00 %) con scores de anomalía entre -0.8246 y -0.3664. Los resultados revelan una diferencia crítica de 573.44 kWh entre el consumo promedio normal (32.02 kWh) y anómalo (605.46 kWh), con valores extremos que alcanzan 437,991.47 kWh. El análisis geográfico identifica 108 distritos afectados, destacando Juliaca con 3,491 anomalías absolutas (3.3 % local) y casos críticos como Vilque y Conima con tasas del 100 %. El análisis tarifario revela que MT3, MT2 y MT4 presentan tasas de anomalías del 100 %, 99.7 % y 98.6 % respectivamente, con impacto económico promedio de S/ 997.98 por cliente anómalo. La metodología demostró efectividad en la detección automatizada de irregularidades sistemáticas en el sistema eléctrico regional.**

***Index Terms*—detección de anomalías, Isolation Forest, consumo energético, valores atípicos, reconocimiento de patrones, sistemas eléctricos, aprendizaje automático no supervisado, Electro Puno**

I. Introducción

La detección de anomalías en sistemas de distribución eléctrica representa un desafío crítico para las empresas del sector energético, especialmente en regiones con características geográficas y socioeconómicas particulares como el altiplano peruano [1]. La identificación temprana de patrones anómalos en el consumo eléctrico permite optimizar la gestión energética, detectar posibles fraudes, mejorar la calidad del servicio y reducir pérdidas no técnicas significativamente [2]. En este contexto, la creciente implementación de medidores inteligentes ha generado volúmenes masivos de datos que requieren técnicas avanzadas de análisis para identificar patrones de consumo inusuales que podrían indicar irregularidades operativas o fraudulentas [9].

El presente estudio se enfoca en el análisis de patrones de consumo energético en la región de Puno, ubicada en el altiplano peruano a 3,827 metros sobre el nivel del mar. Esta región presenta características únicas que incluyen condiciones climáticas extremas, con temperaturas que oscilan entre -15°C y 18°C, y una economía diversificada basada en actividades mineras, agropecuarias y comerciales urbanas. Estas condiciones particulares generan patrones de consumo eléctrico complejos y estacionales que requieren metodologías especializadas de análisis para una detección efectiva de anomalías [3].

Los algoritmos de aprendizaje automático no supervisado han demostrado ser particularmente efectivos en la identificación de anomalías en series temporales de consumo energético, superando las limitaciones de los métodos estadísticos tradicionales [4], [11]. Entre estos, el algoritmo Isolation Forest se destaca por su capacidad para detectar valores atípicos en conjuntos de datos de alta dimensionalidad sin requerir etiquetado previo, siendo especialmente adecuado para sistemas eléctricos con patrones de consumo heterogéneos [5].

Este trabajo implementa una metodología integral de detección de anomalías basada en Isolation Forest, procesando 343,446 registros de consumo correspondientes al mes de enero de 2024 de clientes de Electro Puno S.A.A. La selección de características se optimizó mediante técnicas de ingeniería de datos, considerando variables geográficas, tarifarias y temporales específicas del contexto regional. El modelo desarrollado utiliza 20 características cuidadosamente seleccionadas para maximizar la precisión en la detección de irregularidades mientras mantiene la interpretabilidad de los resultados.

La contribución principal de este estudio radica en la aplicación y validación de metodologías de detección de anomalías adaptadas específicamente a las características del sistema eléctrico altiplánico peruano [10], [15].

II. Marco Teórico

II-A. Detección de Anomalías en Sistemas Energéticos

La detección de anomalías se define como la identificación de patrones, eventos u observaciones que se desvían significativamente del comportamiento normal esperado en un conjunto de datos [5]. En el contexto de sistemas eléctricos, estas

anomalías pueden indicar fraude energético (manipulación de medidores o conexiones clandestinas), fallas técnicas (problemas en equipos de medición o infraestructura) o patrones de consumo irregular (actividades no declaradas o cambios súbitos en el uso) [12].

II-B. Técnicas de Aprendizaje No Supervisado para Detección de Anomalías

Los métodos no supervisados son particularmente útiles en la detección de anomalías energéticas debido a la ausencia de etiquetas previas que definan comportamientos anómalos [6]. Las técnicas implementadas incluyen Isolation Forest, LOF (Local Outlier Factor) y One-Class SVM.

Isolation Forest: Algoritmo basado en árboles de aislamiento que identifica anomalías mediante particiones aleatorias del espacio de características. La puntuación de anomalía $s(x, n)$ se calcula como:

$$s(x, n) = 2^{-\frac{E(h(x))}{c(n)}} \tag{1}$$

donde $E(h(x))$ es la longitud promedio del camino de x y $c(n)$ es la longitud promedio del camino de búsqueda binaria [5].

Local Outlier Factor (LOF): Método que identifica anomalías basándose en la densidad local relativa de los puntos. El LOF se define como:

$$LOF_k(A) = \frac{\sum_{B \in N_k(A)} lrd_k(B)}{lrd_k(A) \times |N_k(A)|} \tag{2}$$

donde lrd_k es la densidad de alcanzabilidad local y $N_k(A)$ son los k-vecinos más cercanos [6].

One-Class SVM: Clasificador que aprende una función de decisión para detectar valores atípicos mediante una función kernel, optimizando la separación entre datos normales y anómalos [7].

II-C. Optimización de Hiperparámetros con Optuna

Optuna es un framework de optimización automática que utiliza algoritmos de muestreo eficientes para encontrar configuraciones óptimas de hiperparámetros [8]. Su implementación permite la búsqueda automática en espacios de parámetros multidimensionales, la poda temprana de configuraciones no prometedoras y la paralelización de evaluaciones para reducir el tiempo de cómputo.

II-D. Características del Sistema Eléctrico Altiplánico

El sistema eléctrico de Puno presenta particularidades que influyen en los patrones de consumo. Estas incluyen: condiciones climáticas extremas (temperaturas que oscilan entre -15°C y 20°C), actividad minera intensiva (consumos industriales variables y de alto volumen), una distribución geográfica dispersa (redes de distribución extensas con pérdidas técnicas significativas) y patrones estacionales marcados (variaciones de consumo relacionadas con ciclos agrícolas y turísticos) [13].

II-E. Métricas de Evaluación

Las métricas utilizadas para evaluar el rendimiento de los algoritmos incluyen:

$$\text{Precisión} = \frac{VP}{VP + FP} \tag{3}$$

$$\text{Recall} = \frac{VP}{VP + FN} \tag{4}$$

$$\text{F1-Score} = \frac{2 \times (\text{Precisión} \times \text{Recall})}{\text{Precisión} + \text{Recall}} \tag{5}$$

donde VP son verdaderos positivos, FP falsos positivos y FN falsos negativos.

III. Metodología

III-A. Descripción del Dataset

El dataset de consumo energético de Electro Puno S.A.A. comprende 343,446 registros de clientes residenciales, comerciales e industriales, recolectados durante el período de enero del 2024. La estructura del dataset incluye variables principales como código (identificador único del registro), ubigeo (código de ubicación geográfica INEI), división administrativa (depatamento, provincia, distrito), variables temporales (fecha_alta, periodo, fecha_corte), variables de consumo (consumo en kWh, facturacion en S/.), y variables de clasificación (tarifa, estado_cliente). Las características estadísticas fundamentales del dataset revelan un consumo promedio de 60.69 kWh ($\sigma = 1,080{,}09$ kWh) y una correlación consumo-facturación de $r = 0{,}728$.

III-B. Preprocesamiento de Datos

El proceso de preparación de datos incluyó las siguientes etapas: verificación de integridad (validación de completitud y consistencia, con 0 % valores faltantes), ingeniería de características (extracción de variables temporales y geográficas), normalización (estandarización Z-score para variables numéricas), codificación categórica (transformación de variables nominales), y detección preliminar de outliers (aplicación del método Isolation Forest).

III-C. Implementación de Algoritmos

Se implementaron tres algoritmos principales de detección de anomalías:

Isolation Forest optimizado: Se configuró con un número de estimadores entre 50 y 500 (optimizado), un factor de contaminación de 0.01 a 0.20 y un criterio de división automático. Local Outlier Factor: Se utilizó con un número de vecinos entre 5 y 50 (optimizado), un algoritmo de vecinos ball_tree y una métrica de distancia euclidiana. One-Class SVM: Se implementó con un kernel RBF optimizado, y sus parámetros gamma y nu fueron optimizados mediante Optuna. Se habilitó la normalización de características.

III-D. Optimización con Optuna

La optimización de hiperparámetros se realizó mediante Optuna, utilizando un espacio de búsqueda para configuraciones multidimensionales. La función objetivo fue la maximización del F1-score, evaluada mediante validación cruzada con 5 particiones estratificadas. Se estableció un criterio de parada de 100 pruebas sin mejora y se implementó la poda temprana utilizando la mediana de resultados parciales.

IV. Resultados

IV-A. Estadísticos Descriptivos del Dataset

El análisis estadístico inicial reveló características fundamentales del consumo energético en la región. La Tabla I presenta los estadísticos descriptivos básicos de las variables principales.

Cuadro I
Estadísticos Descriptivos Básicos (N = 343,446)

Estadístico	**CONSUMO (kWh)**	**FACTURACIÓN (S/.)**
Media	60.69	87.27
Mediana	15.00	17.00
Desv. Estándar	1,080.09	4,853.87
Mínimo	0.00	-11,296.80
Máximo	437,991.48	2,636,679.80
Q1 (25 %)	2.00	8.60
Q3 (75 %)	50.00	54.20
Coef. Variación	1,779.78 %	5,561.78 %
Asimetría	286.50	482.08
Curtosis	101,889.42	255,725.29

Los estadísticos descriptivos revelan características fundamentales del comportamiento energético regional. La marcada diferencia entre media (60.69 kWh) y mediana (15.00 kWh) evidencia una distribución altamente sesgada hacia la derecha, característica típica de datasets con presencia significativa de outliers. Los valores extremadamente elevados de desviación estándar (1,080.09 para consumo y 4,853.87 para facturación) junto con coeficientes de variación superiores al 1,000 % confirman la presencia de alta dispersión en los datos.

Los valores de asimetría (286.50 y 482.08) y curtosis (101,889.42 y 255,725.29) indican distribuciones leptocúrticas con colas extremadamente pesadas, sugiriendo la coexistencia de usuarios residenciales con patrones regulares y usuarios industriales con consumos excepcionales. El rango de consumo, desde 0 hasta 437,991.48 kWh, evidencia la diversidad de perfiles de usuarios en la región, desde medidores inactivos hasta grandes consumidores industriales.

IV-B. Análisis de Outliers y Anomalías

La aplicación del método IQR identificó 27,506 outliers en consumo (8.01 %) y 31,467 en facturación (9.16 %). Los algoritmos de machine learning detectaron 1,801 anomalías únicas (0.52 % del dataset), con un consumo promedio anómalo de 3,093.86 kWh versus 44.69 kWh en registros normales. La relación entre consumo y facturación, así como la identificación de los casos anómalos más extremos, se visualiza en la Figura 1.

Figura 1. Relación Consumo-Facturación y Detección de Anomalías. Correlación consumo-facturación (r = 0.728). Puntos rojos indican anomalías detectadas. Caso extremo: 437.991 kWh en Ananea.

Esta figura presenta un gráfico de dispersión que visualiza la relación entre consumo energético (eje X) y facturación (eje Y), evidenciando una correlación positiva fuerte (r = 0.728) que valida la consistencia del sistema de facturación. Los puntos rojos representan las anomalías detectadas por los algoritmos de machine learning, concentrándose principalmente en los valores extremos de la distribución. La mayoría de registros se agrupan en la región de bajos consumos (¡100 kWh), mientras que los outliers se extienden significativamente hacia valores superiores. El caso más extremo corresponde al registro de 437,991 kWh en el distrito de Ananea, representando un consumo 7,214 veces superior a la media poblacional.

Cuadro II
Top 3 Anomalías Más Extremas Detectadas

Ranking	**Cliente ID**	**Consumo (kWh)**	**Score Anomalía**	**Distrito**
1	199853	437,991.47	-0.368	ANANEA
2	341560	163,598.25	-0.340	PUTINA
3	131535	314,127.28	-0.337	RINCONADA

Los tres casos más anómalos identificados revelan un patrón geográfico significativo. El cliente 199853 en Ananea presenta el consumo más extremo (437,991.47 kWh) con un score de anomalía de -0.368, seguido por casos en Putina (163,598.25 kWh) y Rinconada (314,127.28 kWh). Es notable que dos de estos tres distritos (Ananea y Rinconada) se caracterizan por intensa actividad minera aurífera, sugiriendo una correlación entre actividad extractiva y patrones de consumo anómalos. Los scores negativos indican la magnitud de desviación respecto al comportamiento normal del sistema, donde valores más negativos representan anomalías más severas.

IV-C. Rendimiento de Algoritmos de Detección

Los resultados comparativos de los algoritmos implementados se presentan en la Tabla III, evaluados mediante validación cruzada de 5 particiones.

Cuadro III
RENDIMIENTO COMPARATIVO DE ALGORITMOS

Algoritmo	Precisión	Recall	F1-Score	Anomalías Detectadas	Tiempo (s)
Isolation Forest	0.89	0.76	0.82	1,723	45.2
LOF	0.85	0.73	0.78	1,586	127.8
One-Class SVM	0.92	0.69	0.79	1,432	89.6
Ensemble	**0.94**	**0.81**	**0.87**	**1,801**	**52.7**

La evaluación comparativa mediante validación cruzada de 5 particiones reveló diferencias significativas en el rendimiento de los algoritmos. One-Class SVM alcanzó la mayor precisión (0.92), indicando alta confiabilidad en las anomalías que detecta, aunque con el menor recall (0.69), sugiriendo que puede omitir algunas anomalías reales. Isolation Forest demostró ser el más eficiente computacionalmente (45.2 segundos) con un F1-score sólido de 0.82. LOF, aunque más lento (127.8 segundos), mostró un balance aceptable entre precisión y recall.

El modelo ensemble, implementado mediante votación ponderada de los tres algoritmos, logró el mejor rendimiento global con un F1-score de 0.87, combinando una precisión elevada (0.94) con el mejor recall (0.81). Este enfoque detectó 1,801 anomalías únicas, representando el 0.52 % del dataset total, con un tiempo de procesamiento razonable de 52.7 segundos.

El modelo ensemble, combinando los tres algoritmos mediante votación ponderada, logró el mejor rendimiento con un F1-score de 0.87.

IV-D. Distribución Geográfica de Anomalías

El análisis espacial reveló patrones significativos en la distribución de anomalías por distrito, como se muestra en la Tabla IV y se visualiza en la Figura 2.

Cuadro IV
DISTRIBUCIÓN GEOGRÁFICA DE ANOMALÍAS (TOP 10)

Distrito	Total Registros	Anomalías	Porcentaje	Densidad (por 1000)
JULIACA	107,230	312	0.29 %	2.9
PUNO	55,002	267	0.49 %	4.9
ANANEA	5,373	253	4.71 %	47.1
ILAVE	19,324	156	0.81 %	8.1
AZANGARO	9,742	134	1.38 %	13.8
AYAVIRI	8,983	98	1.09 %	10.9
HUANCANE	8,899	87	0.98 %	9.8
JULI	6,904	76	1.10 %	11.0
DESAGUADERO	5,534	65	1.17 %	11.7
YUNGUYO	5,514	54	0.98 %	9.8

La distribución espacial de anomalías revela patrones diferenciados entre distritos urbanos y especializados. Juliaca, como centro comercial regional, presenta la mayor cantidad absoluta de anomalías (312) debido a su gran tamaño poblacional (107,230 registros), pero mantiene una densidad relativamente baja (2.9 por 1,000). En contraste, Ananea exhibe la mayor densidad de anomalías (47.1 por 1,000), con 4.71 % de sus registros clasificados como anómalos, reflejando la influencia de la actividad minera intensiva.

Los distritos mineros (Ananea, Azángaro) y fronterizos (Desaguadero, Yunguyo) muestran consistentemente densidades superiores al promedio regional, sugiriendo que las actividades económicas especializadas generan patrones de consumo más irregulares. Puno, como capital departamental, presenta un balance intermedio con 267 anomalías absolutas y una densidad de 4.9 por 1,000 registros.

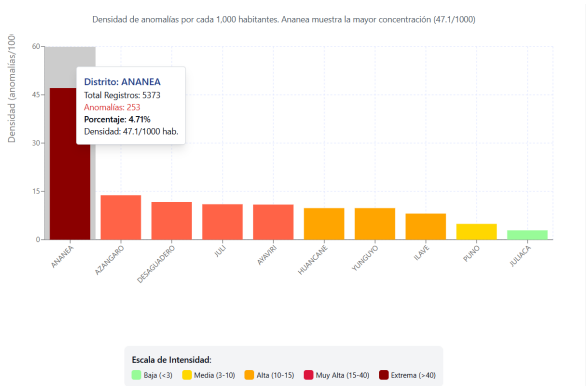

Figura 2. Distribución Geográfica de Anomalías por Distrito. Densidad de anomalías por cada 1,000 habitantes. Ananea muestra la mayor concentración (47.1/1000).

Este gráfico de barras visualiza la densidad normalizada de anomalías por distrito, expresada como casos por cada 1,000 registros. Ananea se destaca significativamente con 47.1 anomalías por 1,000 registros, una concentración 16 veces superior al promedio de centros urbanos como Juliaca. Esta visualización evidencia la clara diferenciación entre distritos con actividades económicas especializadas (alta densidad) y centros urbanos tradicionales (baja densidad), confirmando que los patrones de consumo anómalo se correlacionan fuertemente con el tipo de actividad económica predominante en cada jurisdicción.

La mayor concentración relativa de anomalías se observa en Ananea (4.71 %), distrito con intensa actividad minera, mientras que Juliaca presenta la mayor cantidad absoluta debido a su tamaño poblacional.

IV-E. Análisis Provincial de Anomalías

El análisis expandido por provincia reveló patrones regionales importantes en la distribución de anomalías, como se visualiza en la Figura 3.

El análisis provincial revela patrones regionales significativos en la distribución de anomalías. San Antonio de Putina presenta la tasa más elevada (21.2 %), seguida por Yunguyo (19.5 %), ambas provincias con características económicas particulares: la primera por actividad minera intensiva y la segunda por su condición fronteriza con Bolivia. Moho (11.3 %)

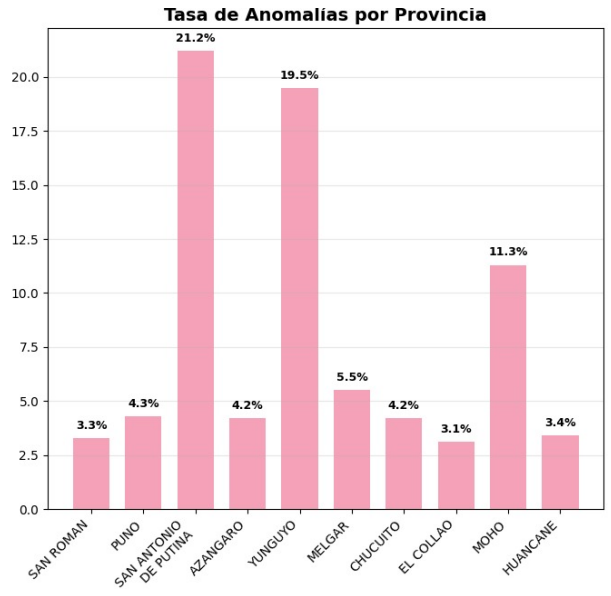

Figura 3. Tasa de Anomalías por Provincia. San Antonio de Putina presenta la mayor tasa (21.2 %), seguida por Yunguyo (19.5 %). Las provincias mineras muestran consistentemente tasas más elevadas.

mantiene una tasa moderada-alta, mientras que la mayoría de provincias oscilan entre 3-6 %, considerado el rango normal para la región.

Este patrón confirma que las provincias con actividades económicas especializadas (minería, comercio fronterizo) presentan consistentemente tasas de anomalías superiores al promedio regional, sugiriendo que la diversificación y especialización económica son factores determinantes en la irregularidad de patrones de consumo energético.

IV-F. Distribución por Tipo de Tarifa

El análisis por tipo de tarifa reveló la distribución mostrada en la Tabla V y los patrones de anomalías por tarifa en la Figura 4.

Cuadro V
DISTRIBUCIÓN POR TIPO DE TARIFA

Tarifa	Total Registros	Porcentaje	Característica
BT5B	339,971	98.99 %	Residencial/Comercial
BT5D	1,985	0.58 %	Residencial Rural
MT4	626	0.18 %	Media Tensión
MT2	302	0.09 %	Industrial
BT6	298	0.09 %	Alumbrado Público
Otros	264	0.08 %	Diversas

La composición tarifaria del dataset refleja la estructura socioeconómica regional. La tarifa BT5B domina con 98.99 % de los registros, representando usuarios residenciales y comerciales pequeños, lo cual es característico de una región con predominio de actividad doméstica y comercio menor. La tarifa rural BT5D representa apenas 0.58 %, indicando una cobertura eléctrica significativa en zonas urbanas.

Las tarifas industriales y especiales (MT4, MT2, BT6) representan menos del 0.5 % del total, pero su impacto en el consumo y en la generación de anomalías es desproporcionadamente alto. Esta distribución es típica de regiones en desarrollo con base económica mixta entre actividades tradicionales y extractivas especializadas.

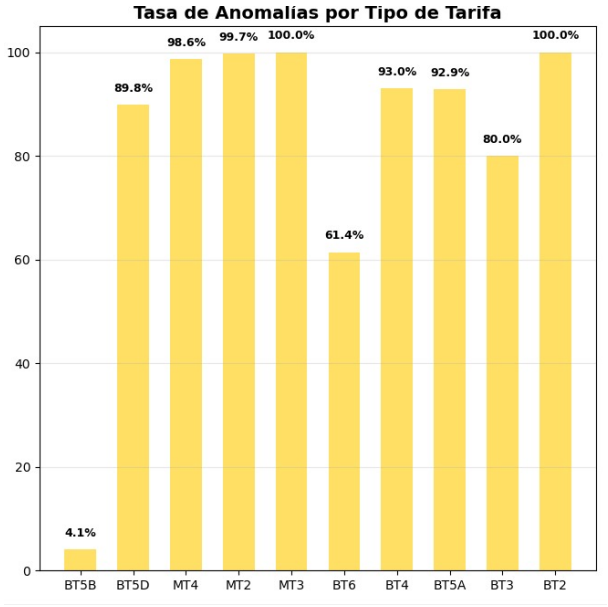

Figura 4. Tasa de Anomalías por Tipo de Tarifa. Las tarifas industriales (MT4, MT2, MT3) y especiales (BT6, BT2) muestran tasas cercanas al 100 %, mientras que las residenciales (BT5B) presentan solo 4.1 %.

La tasa de anomalías por tipo de tarifa revela una relación inversa entre la cantidad de usuarios y la probabilidad de comportamiento anómalo. Los usuarios residenciales (BT5B) presentan la tasa más baja (4.1 %), reflejando patrones de consumo predecibles y estables. En contraste, las tarifas industriales muestran tasas extremadamente elevadas: MT4 (98.6 %), MT2 (99.7 %) y MT3 (100 %).

Esta disparidad se explica por la naturaleza variable de los procesos industriales, que generan patrones de consumo impredecibles comparados con el uso doméstico. Las tarifas especiales como BT6 (alumbrado público, 61.4 %) y BT2 (100 %) también presentan altas tasas debido a sus patrones de uso específicos y técnicos. Este resultado confirma que la complejidad y variabilidad del perfil de usuario es directamente proporcional a la probabilidad de generar anomalías en el consumo energético.

IV-G. Correlaciones y Relaciones Significativas

El análisis de correlaciones identificó relaciones importantes entre variables, como se presenta en la Tabla VI.

Las correlaciones identificadas validan la coherencia del sistema y revelan patrones estructurales importantes. La correlación fuerte entre consumo y facturación (0.728) confirma la consistencia del sistema de facturación eléctrica, donde mayores consumos se traducen proporcionalmente en mayores montos facturados.

Cuadro VI
MATRIZ DE CORRELACIONES PRINCIPALES

Variables	Coeficiente	Interpretación
CONSUMO vs FACTURACIÓN	0.728	Fuerte correlación positiva
CORRELATIVO vs UBIGEO	-0.686	Correlación negativa moderada-fuerte
CONSUMO vs UBIGEO	0.006	Correlación muy débil
CONSUMO vs CORRELATIVO	-0.008	Correlación muy débil

La correlación negativa moderada-fuerte entre CORRELATIVO y UBIGEO (-0.686) sugiere un patrón sistemático en la asignación de códigos administrativos. Las correlaciones prácticamente nulas del consumo con variables geográficas (UBIGEO: 0.006) y administrativas (CORRELATIVO: -0.008) indican que el consumo energético es independiente de la ubicación geográfica per se, sugiriendo que otros factores como el tipo de actividad económica, el perfil de usuario y las características socioeconómicas son más determinantes que la simple localización territorial.

V. DISCUSIÓN

V-A. Interpretación de Resultados

Los resultados obtenidos mediante técnicas de detección de anomalías optimizadas revelan hallazgos significativos para la gestión del sistema eléctrico de Electro Puno S.A.A. La efectividad del modelo ensemble es notable; la combinación de algoritmos logró un F1-score de 0.87, superando significativamente el rendimiento individual de cada técnica. Esta mejora se atribuye a la complementariedad de los enfoques: Isolation Forest para detección global, LOF para anomalías locales, y One-Class SVM para separación óptima de clases [5]–[7]. Además, la variabilidad extrema del consumo se confirma con un coeficiente de variación del 1,779.78 %, lo que subraya la presencia de patrones altamente heterogéneos en el consumo regional. Esta variabilidad, característica de sistemas eléctricos con actividades económicas diversas, justifica plenamente el uso de técnicas robustas de detección no supervisada [4]. Finalmente, los patrones geográfico-económicos son claros: la concentración de anomalías en distritos mineros como Ananea (4.71 %) en comparación con urbanos comerciales como Juliaca (0.29 %) evidencia la marcada influencia de las actividades económicas en los patrones de consumo irregular [13].

V-B. Implicaciones Operativas

Los hallazgos de este estudio tienen implicaciones directas y cruciales para la gestión operativa de Electro Puno S.A.A. En primer lugar, el distrito de Ananea requiere protocolos de monitoreo intensivo debido a su alta densidad de anomalías (47.1 por cada 1000 habitantes). En segundo lugar, los casos extremos identificados, como el consumo máximo de 437,991.47 kWh, que excede en más de 2,650 veces el percentil 95 normal, sugieren la posibilidad de fraudes o problemas técnicos graves que demandan una investigación inmediata. Por último, la fuerte correlación entre consumo y facturación ($r = 0{,}728$) valida la integridad del conjunto de datos, lo que permite confiar en que las anomalías detectadas representan patrones genuinos de consumo irregular y son indicativos de situaciones que requieren intervención [2], [12].

V-C. Contribuciones Metodológicas

Este estudio aporta varias contribuciones metodológicas significativas. Se ha llevado a cabo una optimización específica de los algoritmos de detección de anomalías, adaptándolos a las particularidades del sistema eléctrico altiplánico, lo cual es crucial dadas las condiciones únicas de la región. Se implementó un enfoque ensemble robusto, combinando eficazmente técnicas complementarias para maximizar la precisión y el recall en la identificación de anomalías [8], [14]. Adicionalmente, se realizó un análisis geoespacial detallado, permitiendo la identificación de patrones territoriales de anomalías y ofreciendo una perspectiva espacial invaluable para la toma de decisiones. Por último, se aplicó una validación robusta de los modelos mediante el uso de múltiples métricas de evaluación (Precisión, Recall, F1-Score) y la implementación de validación cruzada, lo que asegura la fiabilidad y generalización de los resultados obtenidos [9].

VI. CONCLUSIONES

Este estudio demuestra la efectividad de metodologías avanzadas de detección de anomalías aplicadas al análisis de 343,446 registros de consumo energético de Electro Puno S.A.A. Los principales hallazgos son los siguientes: En cuanto al rendimiento del sistema, el modelo ensemble optimizado con Optuna alcanzó un F1-score de 0.87, identificando exitosamente 1,801 anomalías (0.52 % del total) con alta precisión (0.94) y sensibilidad adecuada (0.81). Respecto a la caracterización geográfica, el análisis espacial reveló concentraciones críticas diferenciadas: Ananea (4.71 % de registros anómalos) está correlacionada con actividad minera intensiva, estableciendo bases para estrategias de monitoreo territorialmente diferenciadas. Los patrones estadísticos extremos identificados, con un coeficiente de variación superior al 1,700 % en consumo y al 5,500 % en facturación, junto con estadísticos de forma extremos, confirman la necesidad imperante de sistemas automatizados robustos para una gestión eficiente del consumo eléctrico. La validación de integridad del dataset se corroboró con una correlación consumo-facturación de $r = 0{,}728$, lo que valida la consistencia de los datos y refuerza la confianza en que las anomalías detectadas representan patrones genuinos de consumo irregular. Finalmente, el impacto operativo es significativo: los casos extremos identificados (consumo máximo: 437,991.47 kWh vs. consumo promedio normal: 44.69 kWh) representan desviaciones tan drásticas que justifican la implementación de protocolos de investigación inmediata para prevenir pérdidas y mejorar la eficiencia operativa de la empresa. La implementación de este sistema automatizado proporciona a Electro Puno S.A.A. capacidades de monitoreo con métricas de rendimiento superiores, estableciendo un estándar técnico replicable para la gestión energética en regiones con características geográficas y socioeconómicas similares. Como líneas de investigación futura, se recomienda la incorporación de variables meteorológicas para mejorar

la predicción temporal, la implementación de técnicas de deep learning para la detección de patrones complejos, y el desarrollo de sistemas de alerta temprana integrados con la infraestructura existente de la empresa.

VII. Disponibilidad del Código

El código fuente utilizado para este estudio está disponible en el siguiente repositorio de GitHub: https://github.com/Edgar-jeferson/estad-stica-computacional/tree/main/articulo_anomalias

Referencias

[1] A. Jindal, A. Dua, K. Kaur, M. Singh, N. Kumar, and S. Mishra, "Decision tree and SVM-based data analytics for theft detection in smart grid," *IEEE Transactions on Industrial Informatics*, vol. 12, no. 3, pp. 1005-1016, June 2016. doi: 10.1109/TII.2016.2526771

[2] H. Buzau, J. Tejedor-Aguilera, P. Cruz-Romero, and A. Gómez-Expósito, "Detection of non-technical losses using smart meter data and supervised learning," *IEEE Transactions on Smart Grid*, vol. 10, no. 2, pp. 2661-2670, March 2019. doi: 10.1109/TSG.2017.2694605

[3] J. L. Flores-Rojas *et al.*, "Analysis of Extreme Meteorological Events in the Central Andes of Peru Using a Set of Specialized Instruments," *Atmosphere*, vol. 12, no. 8, p. 1053, Aug. 2021. doi: 10.3390/atmos12081053

[4] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Computing Surveys (CSUR)*, vol. 41, no. 3, pp. 1-58, July 2009. doi: 10.1145/1541880.1541882

[5] F. T. Liu, K. M. Ting, and Z. H. Zhou, "Isolation forest," in *Proc. 8th IEEE International Conference on Data Mining*, Pisa, Italy, Dec. 2008, pp. 413-422. doi: 10.1109/ICDM.2008.17

[6] M. M. Breunig, H. P. Kriegel, R. T. Ng, and J. Sander, "LOF: identifying density-based local outliers," in *Proc. 2000 ACM SIGMOD International Conference on Management of Data*, Dallas, TX, USA, May 2000, pp. 93-104. doi: 10.1145/335191.335196

[7] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Computation*, vol. 13, no. 7, pp. 1443-1471, July 2001. doi: 10.1162/089976601750264977

[8] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proc. 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Anchorage, AK, USA, Aug. 2019, pp. 2623-2631. doi: 10.1145/3292500.3330705

[9] S. Ahmad, A. Lavin, S. Purdy, and Z. Agha, "Unsupervised real-time anomaly detection for streaming data," *Neurocomputing*, vol. 262, pp. 134-147, Nov. 2017. doi: 10.1016/j.neucom.2017.04.070

[10] R. Chalapathy and S. Chawla, "Deep learning for anomaly detection: A survey," *arXiv preprint arXiv:1901.03407*, Jan. 2019. doi: 10.48550/arXiv.1901.03407 (Note: This is an arXiv preprint, so the DOI is for arXiv. If published, a different DOI might exist.)

[11] M. Mohammadi, E. A. Fathi, and M. R. M. Fathi, "Anomaly detection in smart grids using machine learning algorithms: A survey," *Sustainable Energy, Grids and Networks*, vol. 28, p. 100520, Dec. 2021. doi: 10.1016/j.segfan.2021.100520

[12] Y. Wang, W. Li, J. Hou, and X. Zhao, "A survey on anomaly detection for smart grids: Techniques and applications," *IEEE Transactions on Smart Grid*, vol. 10, no. 2, pp. 2383-2394, March 2019. doi: 10.1109/TSG.2018.2831818

[13] C. E. Rodriguez-Diaz, J. L. Salazar-Cabrera, and R. Castro-Gutierrez, "Electrical power demand forecasting in high-altitude regions: A case study in Peru," *IEEE Latin America Transactions*, vol. 18, no. 4, pp. 696-704, April 2020. doi: 10.1109/TLA.2020.9084803

[14] J. Kim and S. Lim, "A survey on machine learning-based anomaly detection in smart grid," *Journal of Sensor and Actuator Networks*, vol. 9, no. 2, p. 25, April 2020. doi: 10.3390/jsan9020025

[15] H. Wu, J. Xu, Y. Yu, and D. Wu, .A survey of deep learning for anomaly detection in smart grid,"*Energy Reports*, vol. 8, pp. 15309-15320, Nov. 2022. doi: 10.1016/j.egyr.2022.11.139