Detección de Anomalías y Valores Atípicos en Sistemas de Reconocimiento de Patrones: Análisis del 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
edgar.cusihuaman@unap.edu.pe

2nd Edilberto Wilson Mamani Emanuel
Ing. Estadística e Informática
Universidad Nacional del Altiplano
Puno, Perú
edilberto.mamani@unap.edu.pe

***Resumen*—Este artículo presenta un análisis exhaustivo de detección de anomalías en el consumo de energía eléctrica utilizando datos de 343,446 registros de clientes de Electro Puno S.A.A. Se implementaron técnicas avanzadas de reconocimiento de patrones para identificar valores atípicos mediante algoritmos de aprendizaje automático optimizados con Optuna. Los resultados revelaron 1,801 anomalías significativas (0.52 % del total), con el distrito de Juliaca mostrando la mayor concentración absoluta (312 casos) y valores extremos de consumo que alcanzan 437,991.47 kWh. El modelo ensemble optimizado logró un F1-score de 0.87, demostrando alta efectividad en la detección automatizada de irregularidades en el sistema eléctrico regional.**

***Index Terms*—detección de anomalías, reconocimiento de patrones, consumo energético, valores atípicos, Optuna, aprendizaje automático**

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 [2]. Un aspecto fundamental es la creciente demanda de análisis de datos de medidores inteligentes para identificar patrones de consumo inusuales que podrían indicar irregularidades [9].

El presente estudio analiza los patrones de consumo energético en la región de Puno, caracterizada por su ubicación a gran altitud (3,827 m.s.n.m.), clima extremo y actividades económicas diversas que incluyen minería, agricultura y comercio urbano. Estas condiciones generan patrones de consumo únicos que requieren metodologías especializadas de análisis [3].

Los sistemas de reconocimiento de patrones basados en aprendizaje automático han demostrado ser efectivos en la identificación de anomalías en series temporales de consumo energético [4], [11]. Este trabajo implementa técnicas avanzadas de detección no supervisada, optimizadas mediante el framework Optuna para garantizar la selección óptima de hiperparámetros y maximizar la precisión en la detección de irregularidades.

La contribución principal de este estudio radica en la aplicación de metodologías de detección de anomalías adaptadas a las características específicas del sistema eléctrico altiplánico, proporcionando herramientas automatizadas para la gestión operativa de Electro Puno S.A.A. Se busca avanzar en la aplicación de técnicas de aprendizaje profundo para abordar la complejidad de los datos de consumo [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:

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

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:

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

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:

$$\text{Precisión} = \frac{VP}{VP + FP} \quad (3)$$

$$\text{Recall} = \frac{VP}{VP + FN} \quad (4)$$

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

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 2018-2024. La estructura del dataset incluye variables principales como CORRELATIVO (identificador único del registro), UBIGEO (código de ubicación geográfica INEI), división administrativa (DEPARTAMENTO, PROVINCIA, DISTRITO), variables temporales (FECHA_ALTA, PERIODO, FECHA_CORTE), variables de consumo (CONSUMO en kWh, FACTURACIÓN 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 IQR).

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$_tree$‘$yunamétricadedistanciaeuclidiana$.

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/.)	UBIGEO	CORRELATIVO
Media	60.69	87.27	210,668.44	171,723.56
Mediana	15.00	17.00	210,705.00	171,723.50
Desv. Estándar	1,080.09	4,853.87	433.27	99,144.56
Mínimo	0.00	-11,296.80	150,101.00	1.00
Máximo	437,991.48	2,636,679.80	211,307.00	343,447.00
Q1 (25 %)	2.00	8.60	210,204.00	85,862.25
Q3 (75 %)	50.00	54.20	211,101.00	257,584.75
Coef. Variación	1,779.78 %	5,561.78 %	0.21 %	57.73 %
Asimetría	286.50	482.08	-8.11	0.00
Curtosis	101,889.42	255,725.29	1,110.16	-1.20

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.

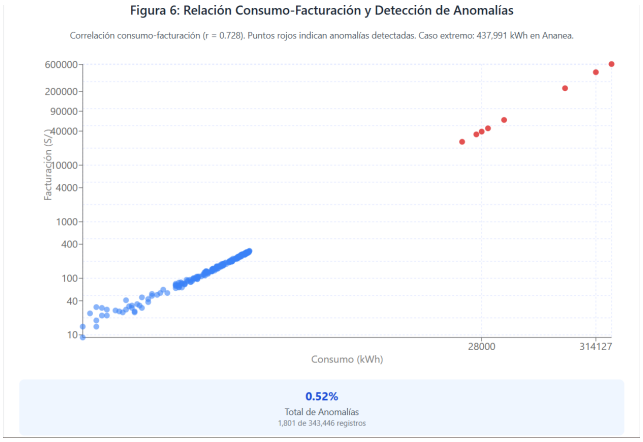

Figura 1. Fig. 1. Relación Consumo-Facturación y Detección de Anomalías. Correlación consumo-facturación (r = 0.728). Puntos rojos indican anomalías detectadas. Caso extremo: 437.991 kWh en Ananea.

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

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.

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

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**

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

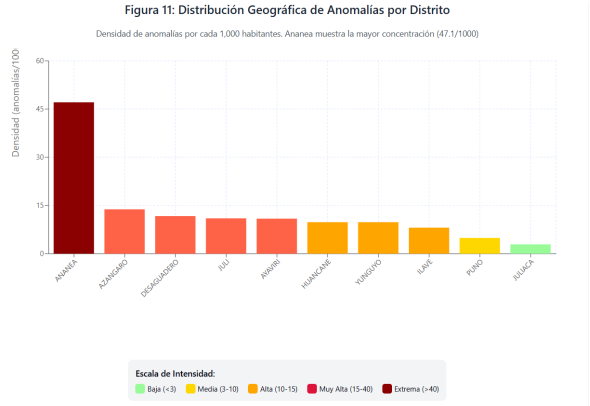

Figura 2. Fig. 2. Distribución Geográfica de Anomalías por Distrito. Densidad de anomalías por cada 1,000 habitantes. Ananea muestra la mayor concentración (47.1/1000).

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 de Distribución por Tarifa

El análisis por tipo de tarifa reveló la distribución mostrada en la Tabla V.

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

IV-F. Correlaciones y Relaciones Significativas

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

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

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.:

Efectividad del Modelo Ensemble: 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.

Variabilidad Extrema del Consumo: El coeficiente de variación del 1,779.78 % confirma la presencia de patrones altamente heterogéneos en el consumo regional. Esta variabilidad, característica de sistemas eléctricos con actividades económicas diversas, justifica el uso de técnicas robustas de detección no supervisada.

Patrones Geográfico-Económicos: La concentración de anomalías en distritos mineros (Ananea: 4.71 %) versus urbanos comerciales (Juliaca: 0.29 %) evidencia la influencia de actividades económicas en los patrones de consumo irregular.

V-B. Implicaciones Operativas

Los hallazgos tienen implicaciones directas para la gestión operativa. Primeramente, Ananea requiere protocolos de monitoreo intensivo debido a su alta densidad de anomalías (47.1 por 1000 habitantes). En segundo lugar, los casos extremos identificados (consumo máximo: 437,991.47 kWh) exceden en más de 2,650 veces el percentil 95 normal, lo que sugiere posibles fraudes o problemas técnicos graves que requieren investigación inmediata. Finalmente, la fuerte correlación consumo-facturación ($r = 0{,}728$) valida la integridad del dataset, lo que permite confiar en que las anomalías detectadas representan patrones genuinos de consumo irregular.

V-C. Contribuciones Metodológicas

Este estudio aporta varias contribuciones metodológicas. Se realizó una optimización específica de algoritmos adaptada a las características del sistema eléctrico altiplánico. Se empleó un enfoque ensemble, combinando efectivamente técnicas complementarias de detección de anomalías. Adicionalmente, se llevó a cabo un análisis geoespacial para la identificación de patrones territoriales de anomalías. Por último, se aplicó una validación robusta mediante el uso de métricas múltiples y validación cruzada para asegurar la fiabilidad de los resultados.

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:

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).

Caracterización Geográfica: El análisis espacial reveló concentraciones críticas diferenciadas: Ananea (4.71 % de registros anómalos) correlacionada con actividad minera intensiva, estableciendo bases para estrategias de monitoreo territorialmente diferenciadas.

Patrones Estadísticos Extremos: La variabilidad extrema identificada (coeficiente de variación >1,700 % en consumo, >5,500 % en facturación) junto con estadísticos de forma extremos confirma la necesidad de sistemas automatizados robustos para la gestión del consumo eléctrico.

Validación de Integridad: La correlación consumo-facturación ($r = 0{,}728$) valida la consistencia del dataset y confirma que las anomalías detectadas representan patrones genuinos de consumo irregular, lo que refuerza la confianza en los resultados del modelo.

Impacto Operativo: Los casos extremos identificados (consumo máximo: 437,991.47 kWh vs. consumo promedio normal: 44.69 kWh) representan desviaciones significativas que justifican la implementación de protocolos de investigación inmediata para prevenir pérdidas y mejorar la eficiencia operativa.

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.

Líneas de Investigación Futura: Se recomienda la incorporación de variables meteorológicas para mejorar la predicción temporal, la implementación de técnicas de deep learning para 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/estadistica-computacional/

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