Aplicación de Data Mining para el consumo energético

Marietha Córdova Delgado
Facultad de Ingeniería
Universidad del Pacífico
Lima, Perú
m.cordovad@alum.up.edu.pe

Diego Medina Manrique
Facultad de Ingeniería
Universidad del Pacífico
Lima, Perú
dr.medinam@alum.up.edu.pe

Sebastian Gomez Mejia
Facultad de Ingeniería
Universidad del Pacífico
Lima, Perú
sm.gomezm@alum.up.edu.pe

Diego Quiñones Vivas
Facultad de Ingeniería
Universidad del Pacífico
Lima, Perú
da.quinonesv@alum.up.edu.pe

I. Introducción

La energía es un factor clave para el desarrollo económico, social y ambiental de un país [1]. Actualmente, cerca del 80% de la energía primaria consumida a nivel mundial proviene de combustibles fósiles, cuya combustión es principal fuente de emisiones de dióxido de carbono (CO_2) [2]. Según la Agencia Internacional de Energía (IEA), las emisiones energéticas llegaron a 37.4 gigatoneladas de CO_2 en 2023, impulsadas por la recuperación económica postpandemia y el aumento del consumo eléctrico en sectores de alta demanda energética (industria pesada, centros de datos, transporte eléctrico) [3].

En el año 2000 la comercialización de fuentes renovables cobró impulso con la expansión de tecnologías como la energía solar y eólica [4]. Como resultado, solo en 2024, estas fuentes representaron el 96% del total de la nueva capacidad eléctrica instalada a nivel mundial [4]. Sin embargo, la velocidad de transición a fuentes sostenibles varía entre países desarrollados y en desarrollo por las diferencias en la capacidad para invertir en infraestructura renovable, almacenamiento, eficiencia energética y electrificación [5].

Se estima que el 75% del crecimiento de la demanda energética al 2030 provendrá de economías emergentes cuyo acceso a tecnologías limpias es limitado [6]. Esto aumenta la dependencia de combustibles fósiles e incrementa el riesgo de generar mayores emisiones de CO_2. Sin embargo, es necesario reducir las emisiones 43% para mantener estable el calentamiento global al 2030 [7]. Esta contradicción entre demanda energética proyectada y metas climáticas evidencia la urgencia de comprender mejor los patrones de consumo energético a nivel mundial.

El comportamiento energético de los países está influenciado por factores socioeconómicos (PBI o población), y por dimensiones ambientales (CO_2) [1] [2] [4]. No obstante, el análisis de estos datos presenta desafíos como: alta dimensionalidad [8], datos faltantes [9], multicolinealidad [10] y redundancias entre variables [11]. Por lo que es necesario usar técnicas de minería de datos para asegurar una interpretación consistente y libre de sesgos [12].

La presente investigación se enfocará en el uso de técnicas de Minería de Datos (DM) y Machine Learning (ML) aplicadas a la base de datos *World Energy Consumption* [13] con el objetivo principal de:

> *Identificar patrones y factores asociados al consumo energético y emisiones de CO_2 mediante técnicas de minería de datos y machine learning, para aportar a la comprensión de la transición energética mundial.*

Y los siguientes objetivos específicos:

1) Realizar limpieza e imputación de valores faltantes en la base de datos de consumo energético.
2) Visualizar la evolución temporal del consumo energético y de las emisiones de CO_2.
3) Explorar correlaciones entre consumo energético, PBI, población y emisiones.
4) Identificar clústeres de países con patrones energéticos similares.
5) Evaluar modelos de predicción de consumo energético o emisiones a futuro.

Algunos conceptos clave para la investigación son:

1) **Knowledge Discovery in Databases (KDD):** Proceso metodológico para la extracción de conocimiento que incluye selección, limpieza, transformación y análisis de datos.
2) **Winsorización:** Técnica que reduce el impacto de valores extremos reemplazándolos por límites superiores o inferiores definidos, sin eliminar los registros.
3) **PCA:** Técnica de reducción de dimensionalidad que permite simplificar las variables originales en componentes principales que concentran la mayor variabilidad de los datos.
4) **Clustering:** Procesos y métodos que permiten agrupar países según patrones similares de consumo, producción y perfil energético.
5) **K-Means:** Algoritmo no supervisado que agrupa las observaciones en grupos (clústeres) basados en distancia, asignando cada punto al centroide más cercano.
6) **Exponential Smoothing (Holt-Winters):** Método de pronóstico que asigna mayor peso a los datos más recientes para modelar tendencias y realizar predicciones a partir de series temporales.

II. Estado del Arte

El preprocesamiento y la calidad de los datos son fundamentales para el análisis de datos energéticos. En Liu et al.

se analizaron los datos del monitoreo de múltiples regiones operativas del sistema eléctrico. El estudio indica que la etapa del preprocesamiento tiene el mayor impacto en la precisión de modelos energéticos; ya que un mal manejo de valores nulos, de valores atípicos y de redundancia entre variables puede generar grandes desviaciones en la calidad de los resultados y condiciona la eficiencia de los algoritmos de predicción o de clustering [16].

El Análisis de Componentes Principales (PCA) ha demostrado ser una herramienta efectiva para reducir la dimensionalidad, eliminar características redundantes y mejorar la calidad de los modelos generados. En Parhizkar et al. se demostró la utilidad del PCA en sistemas energéticos complejos, ya que permite eliminar ruido, reducir redundancias y mejorar la eficiencia del clustering y predicción [14]. Asimismo, Zhang et al. propone el uso de PCA robusto (RPCA) para identificar que factores macroeconómicos influyen en el consumo energético total de China. El estudio determino que el PIB y la urbanización son los factores más significativos en el crecimiento del consumo energético; y la aplicación de RPCA redujo de manera eficaz la inestabilidad generada por datos ruidosos o incompletos [17].

Por otro lado, la clusterización se utilizó para identificar patrones de consumo energético. Métodos como K-means y *clusterwise* regression son útiles para agrupar datos con patrones similares de consumo energético. Li et al. utilizaron datos de consumo energético en edificios y sistemas urbanos para identificar clases de comportamiento energético demostrando que la combinación PCA y K-means mejora la separación de grupos y facilita la interpretación de patrones energéticos homogéneos [18]. Además, los resultados indican que la clusterización mejora significativamente la precisión de los modelos predictivos cuando se utiliza en conjunto con modelos de regresión [14].

En cuanto a tareas de aprendizaje automático orientadas al pronóstico, diversos estudios han empleado modelos basados en series temporales para estimar tendencias energéticas a futuro. Taylor (2003) adaptó el método de Holt-Winters de suavizamiento exponencial para la predicción de la demanda eléctrica [19]. Este tipo de enfoques resulta útil cuando el consumo energético presenta ciclos regulares y tendencia a largo plazo, características comunes en series energéticas reales.

La revisión muestra que los avances más significativos en el campo incluyen el rol crucial del preprocesamiento en la calidad analítica, la utilidad de PCA y RPCA para manejar datos energéticos de alta dimensionalidad, la eficacia del clustering para clasificar perfiles energéticos, la importancia de integrar variables energéticas con factores socioeconómicos y ambientales; y la incorporación creciente de técnicas de ML para el pronóstico temporal. Sin embargo, se necesita más investigación en cuanto al análisis entre países, en lugar de estudios sectoriales, que combinen preprocesamiento, reducción dimensional, clustering y predicción; y que incluyan consumo, emisiones y economía en un mismo análisis.

En ese sentido, A diferencia de los estudios revisados, que se enfocan en predicción o análisis sectoriales, este proyecto integra variables energéticas, socioeconómicas y ambientales en un único análisis global, además aplica PCA para reducir dimensionalidad, utiliza K-means para identificar perfiles energéticos entre países y emplea un modelo de suavizamiento exponencial para explorar la evolución temporal del consumo. El proyecto aporta un análisis claro y completo que integra distintas dimensiones energéticas, permitiendo identificar patrones que no se observan con métodos más simples o centrados solo en predicción.

III. Diseño del experimento

A. Descripción del conjunto de datos

Dataset: *World Energy Consumption* de Kaggle [13].

Dimensiones: El dataset inicial contenía 22,000 registros y 129 variables. Tras preprocesamiento quedaron 8,089 registros y 44 variables (2 categóricas y 42 numéricas).

Variables: consumo total y per cápita; producción energética por fuente (carbón, petróleo, gas, nuclear, hidroeléctrica, solar, eólica, entre otras); emisiones de CO_2; indicadores socioeconómicos (PBI) y demográficos (población), variaciones anuales de consumo y participación porcentual por tipo de energía.

Tareas de Machine Learning: Se realizaron tres tareas:

- Reducción de dimensionalidad (PCA) para la identificación de variables con mayor contribución al comportamiento energético.
- Clustering (k-means) para agrupar países según patrones similares de consumo, producción y perfil energético.
- Pronóstico temporal mediante Exponential Smoothing (Holt-Winters) para estimar tendencias de consumo a futuro.

No se realizó división de muestras en conjuntos de entrenamiento, validación y prueba debido a la naturaleza exploratoria y no supervisada de las dos primeras tareas y a que el propio modelo Holt-Winters realiza validación mediante particiones temporales.

Análisis exploratorio de los datos: Se realizó lo siguiente:

1) Diagrama de cajas (boxplots) sobre las variables numéricas para la evaluación de outliers.
2) Mapa de calor de correlaciones para evaluar la relación entre variables numéricas.
3) Estadísticas descriptivas por variable: media, mediana, std, skewness, kurtosis y porcentaje de nulos.
4) Algunas tablas resumen finales.

B. Metodología

La metodología sigue el proceso KDD, que organiza el trabajo en entendimiento del problema, selección y recopilación de datos, preprocesamiento, transformación y minería de datos e interpretación de resultados. Los dos primeros puntos fueron desarrollados anteriormente, por lo en este apartado se explicará preprocesamiento y los algoritnmos a aplicar. El análisis se implementó en Python utilizando librerías como *pandas*, *numpy*, *scikit-learn* y *statsmodels*.

1) *Preprocesamiento*
   - **Eliminación Inconsistencias:** Se eliminó el registro donde *year = ' Chile'* ya que no corresponde a ningún año válido. Se descartó que sea un posible desplazamiento de las columnas a la derecha.
   - **Tratamiento de duplicados:** Tras la revisión, se confirmó que no existían duplicados, por lo que este procedimiento no aplica.
   - **Manejo de valores nulos:** El dataset abarca periodos históricos amplios y tecnologías que no estuvieron disponibles en todos los años, lo que genera un alto volumen de valores faltantes. Estrategias implementadas:
     a) **Interpolación lineal:** Para variable de energía con valores conocidos en años consecutivos. Si un registro presenta información del 2000 y 2002 pero no del 2001 entonces se interpola para estimar dicho valor correspondiente.
     b) **Imputación por propagación:** Para *population*, *gdp* y *energía*. Se rellenaron los valores faltantes con el dato válido anterior (*backward fill - bfill*) o posterior (*forward fill - ffill*) más cercano. Esta técnica preserva la continuidad temporal de la serie y evita introducir valores inconsistentes entre países.
     c) **Selección de periodos temporales:** Se trabajó únicamente con datos posteriores al año 2000, ya que antes de esa fecha las energías renovables aún no tenían presencia comercial significativa [4] y el dataset presenta una gran cantidad de valores nulos en ese periodo.
     d) **Imputación de país usando *iso_code:*** Si *iso_code = PER* se completa automaticamente *country = Perú* para mantener integridad referencial.
     e) **Eliminación por cantidad de nulos:** Se elimina las columnas con más del 30% de valores faltantes como en Mallala et al. [20]. Este enfoque asegura la conservación de variables con suficiente información para análisis estadístico y modelos posteriores.
   - **Outliers:** Se aplicaron métodos basados en el rango intercuartílico (IQR) y winsorización.
   - **Transformación:** Para variables con distribución aproximadamente normal se usó estandarización (z-score). Para variables con distribuciones no normales se usó reescalamiento (MinMax/robusto)
2) *Algoritmos empleados y criterio de selección*
   a) **PCA:** Seleccionado por su capacidad de reducir dimensionalidad, eliminar redundancias y mejorar la interpretabilidad, siguiendo prácticas destacadas en estudios previos [14], [17].
   b) **k-means:** Método que agrupar países según patrones energéticos [18]. Se utilizó el método del codo para determinar el número óptimo de clústeres.
   c) **Exponential Smoothing:** Modelo adecuado para capturar tendencias suavizadas y variaciones estructurales en series de consumo energético. [19]
3) *Métricas de evaluación*
   a) **Para PCA:** Porcentaje de varianza explicada para determinar cuántos componentes retener.
   b) **Para k-means:** Inercia (SSE) y método del codo, además del coeficiente de Silhouette para validar la separación entre clústeres.
   c) **Para Exponential Smoothing:**No usamos una métrica puesto que estamos prediciendo valores futuros que no estan registrados en la base de datos.
4) *Optimización de hiperparámetros*
   a) **Para PCA:** se seleccionó el número óptimo de componentes asegurando al menos el 80% de varianza explicada. [20]
   b) **Para k-means:** se evaluaron valores de k desde 2 hasta 10 mediante el método del codo y Silhouette.
   c) **Para Exponential Smoothing:** se ajustaron los parámetros de suavizamiento (α, β, γ) mediante búsqueda exhaustiva controlada (grid search).

IV. Experimentacion y Resultados

Esta sección presenta los resultados obtenidos a lo largo del proyecto, haciendo enfasis en las tres tareas principales: reducción de dimensionalidad mediante PCA, agrupamiento con k-means y pronóstico temporal mediante Exponential Smoothing.

A. Preprocesamiento

Se calcularon medidas como media, mediana, desviación estándar, skewness, kurtosis y porcentaje de valores nulos. Esto permitió identificar variables con alta asimetría y presencia significativa de ceros estructurales, especialmente en aquellas relacionadas a energías renovables en décadas previas.

El mapa de calor reveló ocho pares de variables con correlación mayor a 0.7, clasificadas como redundantes. Para mitigar la multicolinealidad se conservaron únicamente las variables más representativas: *greenhouse_gas_emissions_per_capita*, *electricity_generation_per_capita*, *renewables_electricity_per_capita* y *electricity_demand_per_capita* (Véase Anexo 1).

Al realizar diagramas de boxplots, incluso después del preprocesamiento aún se encontraron ciertos valores extremos. Esto se explica porque no todos los outliers pueden ser eliminados sin perder información valiosa, y algunos permanecen al encontrarse dentro de límites aceptables para el análisis (Véase Anexo 2).

El experimento se estructuró en tres componentes:

- PCA para reducción de dimensionalidad.
- K-means para agrupación no supervisada.
- Exponential Smoothing para pronóstico temporal.

Debido a que las dos primeras tareas son no supervisadas, no se realizó división en conjuntos de entrenamiento y prueba

B. Tarea 1: Reducción de Dimensionalidad (PCA)

El PCA permitió condensar las 44 variables en 4 componentes principales, explicando el 82.6% de la varianza acumulada. Este número se seleccionó aplicando un umbral mínimo del 85% de varianza explicada [**?**]. Este resultado permitió condensar la información en dimensiones más representativas, facilitando la interpretación de los patrones y preparando los datos para un modelo posterior de Clusterización (Véase Anexo 3)

Componente	**Varianza Explicada**	**Variables dominantes**
PC1	41.67%	electricity demand, electricity generation
PC2	23.02%	low carbon electricity
PC3	11.93%	gas electricity
PC4	8.86%	gas electricity

C. Tarea 2: Clustering con k-means

Se aplicó k-means sobre los componentes principales de la tarea anterior. El objetivo es agrupar los países en perfiles energéticos diferenciados, ofreciendo una visión más ordenada del conjunto de datos. Para determinar el número óptimo de grupos en el clustering se aplicó el método del codo, que consiste en graficar la inercia (suma de las distancias internas de cada clúster) frente al número de posibles clústeres. El punto de quiebre o "codo" de la curva indicó que el valor más adecuado era k = 3 (Véase Anexo 4).

Entonces el modelo agrupó a los países en tres perfiles energéticos diferenciados:

- Clúster 1: países con baja intensidad energética y menor producción eléctrica.
- Clúster 2: países con alta producción y demanda eléctrica.
- Clúster 3: países en transición con crecimiento moderado y mayor participación renovable.

El gráfico sobre los dos primeros componentes muestra una separación clara entre clústeres, confirmando que la reducción de dimensionalidad fue efectiva para mejorar la segmentación. El problema que se tuvo posteriormente es que la cantidad de datos por cluster estaba desbalanceado, el cluster 2 presento una mayor cantidad de registros frente a los otros grupos (Véase Anexo 5).

D. Tarea 3: Exponential Smoothing

Para cada país y para cada tipo de energía se aplicó un modelo de Suavizamiento Exponencial de Holt-Winters (sin estacionalidad), utilizando tendencia aditiva. Antes del ajuste, se verificó que cada serie tuviera al menos dos valores válidos y que el máximo fuera positivo. De no cumplirse esta condición, se asignaron pronósticos igual a cero para evitar resultados artificiales. Con el modelo ajustado, se generaron proyecciones para los cinco años posteriores al último dato disponible de cada país, restringiendo los valores pronosticados a ser no negativos.

En síntesis, al aplicar filtros para evitar resultados artificiales algunas tablas presentan valores iguales a cero porque no existía evidencia histórica que permitiera generar un pronóstico confiable. En contraste, países con series completas y magnitudes energéticas significativas, como China, sí muestran proyecciones coherentes y proporcionales a su tamaño y a sus tendencias recientes.

V. Discusión

A. Relación entre PCA y clustering

Los resultados muestran que la aplicación de PCA no modifica de manera significativa el desbalanceo presente en los datos. Esto se debe a que PCA es una técnica de reducción de dimensionalidad que transforma las variables originales en nuevas componentes principales, pero no altera la distribución de las clases ni modifica la proporción entre ellas. En otras palabras, aunque PCA cambia la representación geométrica de los datos, el desbalanceo sigue estando presente porque este proviene de la frecuencia relativa de cada clase, no de la dimensionalidad del espacio. Por lo tanto, las estrategias específicas para tratar el desbalanceo deben aplicarse independientemente de si se utiliza PCA o no.

B. Conexión con la literatura

El análisis de clustering ha segmentado la base de datos de países en tres grupos (Cluster 0, Cluster 1 y Cluster 2), permitiendo identificar patrones energéticos, económicos y de transición distintivos. El Cluster 0 se distingue como el grupo de las Potencias Globales y Consumidores Dominantes, caracterizándose por los valores más altos en cifras absolutas, especialmente en PIB, Consumo de Energía Primaria y Generación Eléctrica. Este grupo de economías gigantes, sin embargo, muestra las cuotas más bajas de energías limpias y de bajas emisiones de carbono, indicando una alta dependencia fósil que subraya el desafío de la descarbonización a gran escala.

En contraste, el Cluster 1 representa a los Líderes de la Transición Energética, cuya característica dominante es la alta penetración de energías renovables y de bajas emisiones de carbono, superando a los otros grupos en sostenibilidad del mix eléctrico. Aunque sus valores de consumo y PIB son los más bajos en términos absolutos, este cluster sirve como referencia de las trayectorias de transición exitosas. Finalmente, el Cluster 2 agrupa a las Economías Intermedias con Alta Eficiencia, presentando niveles moderados o intermedios en consumo y PIB, pero destacándose por poseer la Intensidad Energética más baja. Este rasgo implica que son los países más eficientes, requiriendo menos energía para generar una unidad de riqueza económica, y se encuentran en una etapa de transición energética intermedia, sin la dependencia del Cluster 0 ni la madurez del Cluster 1.

C. Limitaciones

Algunas limitaciones que tuvimos durante el desarrollo del trabajo fueron: La base de datos contiene una gran cantidad de variables, la presencia de una amplia variedad de valores nulos

e incompletos en sus series históricas limitó el uso de toda la información (p. ej., datos detallados sobre la producción de energía por tipo en años anteriores o para países con menor registro histórico). Esto obligó a realizar una rigurosa imputación y selección de variables con alta completitud temporal.

El periodo de estudio abarca múltiples décadas que incluyen eventos históricos disruptivos como crisis energéticas, recesiones globales y la reciente pandemia de COVID-19. Estos acontecimientos generan puntos de inflexión y outliers que pueden introducir un sesgo de periodo en los modelos.

El modelo de pronóstico seleccionado prioriza la simplicidad, interpretabilidad y escalabilidad a múltiples países. Sin embargo, esta elección conlleva la limitación de que el modelo no captura adecuadamente las no-linealidades complejas, las interacciones dinámicas de alta complejidad. Por lo tanto, el modelo opera bajo el supuesto de que las tendencias históricas de las variables predictoras se mantendrán, haciendo que el pronóstico sea más robusto a corto plazo y menos preciso para proyecciones a largo plazo.

VI. Conclusiones y trabajos futuros

El proyecto ha cumplido satisfactoriamente con sus objetivos iniciales mediante la aplicación rigurosa de técnicas de Data Mining sobre la base de datos global de consumo energético, lo cual ha permitido extraer patrones clave y validar hipótesis fundamentales. La primera conclusión de peso es la ratificación del vínculo económico-energético: el crecimiento del Producto Interno Bruto (PIB) se mantiene como el factor causal dominante en la demanda de electricidad y energía primaria en la mayoría de las regiones analizadas, lo que establece la planificación de capacidad sobre un sólido pilar macroeconómico. Paralelamente, el análisis confirma que la transición hacia fuentes renovables es una tendencia global e irreversible en clara aceleración. El aumento robusto en la cuota de generación eléctrica proveniente de fuentes como la solar y la eólica subraya la urgencia de adaptar las infraestructuras de red y almacenamiento para gestionar eficientemente la intermitencia inherente a estas fuentes limpias.

Un hallazgo metodológico crucial fue la heterogeneidad regional de los datos, demostrando que las políticas, economías y contextos geográficos resultan en patrones de consumo y adopción tecnológica significativamente distintos. Esta realidad hizo que la segregación del análisis por país fuera imprescindible, permitiendo al enfoque de Data Mining generar un pronóstico baseline diferenciado y específico para cada jurisdicción. La principal contribución del proyecto reside en la entrega de este modelo predictivo de referencia, que, a pesar de sus limitaciones de simplicidad, es una herramienta práctica y valiosa capaz de proyectar la demanda de electricidad en unidades tangibles (GWh/TWh), facilitando la evaluación inicial de escenarios futuros y la identificación de posibles desviaciones en las tendencias históricas de consumo.

Finalmente, para superar las limitaciones de simplicidad del modelo baseline y abordar la complejidad intrínseca del sector, se establecen recomendaciones claras para la investigación futura. Se subraya la necesidad de que futuras iteraciones del proyecto profundicen el análisis, integrando explícitamente variables de política regulatoria y migrando hacia el uso de modelos de Machine Learning más avanzados (tales como Redes Neuronales o Gradient Boosting). Estos enfoques de mayor complejidad permitirán capturar las no-linealidades, las interacciones dinámicas entre variables y, consecuentemente, mejorar significativamente la precisión del pronóstico, especialmente en proyecciones a largo plazo.

References

[1] Carpintero, O., & Frechoso, F. (2023). Energía, sostenibilidad y transición: nuevos desafíos y problemas pendientes. Arbor, 199(807), a687. https://doi.org/10.3989/arbor.2023.807001

[2] Ritchie, H., & Rosado, P. (2017, 2 octubre). Fossil fuels. Our World In Data. https://ourworldindata.org/fossil-fuels

[3] International Energy Agency. (2024). Global Energy and CO_2 Status Report 2024. https://www.iea.org/reports/global-energy-review-2025/co2-emissions

[4] International Renewable Energy Agency [IRENA] (2025, 26 marzo). Record-Breaking Annual Growth in Renewable Power Capacity. https://www.irena.org/News/pressreleases/2025/Mar/Record-Breaking-Annual-Growth-in-Renewable-Power-Capacity

[5] Tambini, K., & Vergara, V. (2024). El impacto del consumo de energía en el crecimiento económico: Un análisis con datos de panel. Desafíos Economía y Empresa, 004, 99-114. https://doi.org/10.26439/ddee2024.n04.6247

[6] International Energy Agency. (2023). World Energy Outlook 2023. https://www.iea.org/reports/world-energy-outlook-2023

[7] IPCC. (2023). AR6 Synthesis Report: Climate Change 2023. https://www.ipcc.ch/report/ar6/syr/

[8] Pan, H, Yin, Z & Jiang, X. (2022). High-Dimensional Energy Consumption Anomaly Detection: A Deep-Learning Based Method. Energies (MDPI).

[9] Duarte, O, Duarte, J & Rosero-Garcia, J. (2024). Data Imputation in Electricity Consumption Profiles through Autoencoders. Journal/Mathematics (MDPI).

[10] Al-Essa LA, Ebrahim EA & Mergiaw YA (2024). Bayesian regression modeling and inference of energy drivers — Frontiers in Energy Research.

[11] Mateos, G., & Giannakis, G. B. (2013). Load Curve Data Cleansing and Imputation via Sparsity and Low Rank.

[12] Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques (3rd ed.). Morgan Kaufmann.

[13] World energy consumption. (2023, 26 noviembre). Kaggle. https://www.kaggle.com/datasets/pralabhpoudel/world-energy-consumption o Owid (s/f). GitHub. https://github.com/owid/energy-data/blob/master/owid-energy-codebook.csv

[14] Parhizkar, T., Rafieipour, E., & Parhizkar, A. (2020). Evaluation and improvement of energy consumption prediction models using principal component analysis based feature reduction. Journal of Cleaner Production, 279, 123866. https://doi.org/10.1016/j.jclepro.2020.123866

[15] Zhao, T., Sun, Y., Chai, Z., & Li, K. (2022). An outlier management framework for building performance data and its application to the power consumption data of building energy systems in non-residential buildings. Journal of Building Engineering, 65, 105688. https://doi.org/10.1016/j.jobe.2022.105688

[16] Liu, X., Ding, Y., Tang, H., & Xiao, F. (2021). A data mining-based framework for the identification of daily electricity usage patterns and anomaly detection in building electricity consumption data. Energy & Buildings, 231, 110601. https://doi.org/10.1016/j.enbuild.2020.110601

[17] Zhang, L., Ge, R., & Chai, J. (2019). Prediction of China's energy consumption based on robust principal component analysis and PSO-LSSVM optimized by the Tabu search algorithm. Energies, 12(1), 196. https://doi.org/10.3390/en12010196

[18] Li, K., Ma, Z., Robinson, D., Lin, W., & Li, Z. (2020). A data-driven strategy to forecast next-day electricity usage and peak electricity demand of a building portfolio using cluster analysis, Cubist regression models and Particle Swarm Optimization. Journal Of Cleaner Production, 273, 123115. https://doi.org/10.1016/j.jclepro.2020.123115

[19] Taylor, J. W. (2003). Short-term electricity demand forecasting using double seasonal exponential smoothing. Journal of the Operational Research Society, 54(8), 799–805. https://doi.org/10.1057/palgrave.jors.2601589

[20] Mallala, B., Ahmed, A. I. U., Pamidi, S. V., Faruque, M. O., & Reddy, R. (2025). Forecasting global sustainable energy from renewable sources using random forest algorithm. Results in Engineering, 25, 103789.

[21] Jolliffe, I. (2011). Principal component analysis. In International encyclopedia of statistical science (pp. 1094-1096). Springer, Berlin, Heidelberg.

VII. ANEXOS

Fig. 1: Anexo 1:Matriz de correlación de variables numéricas

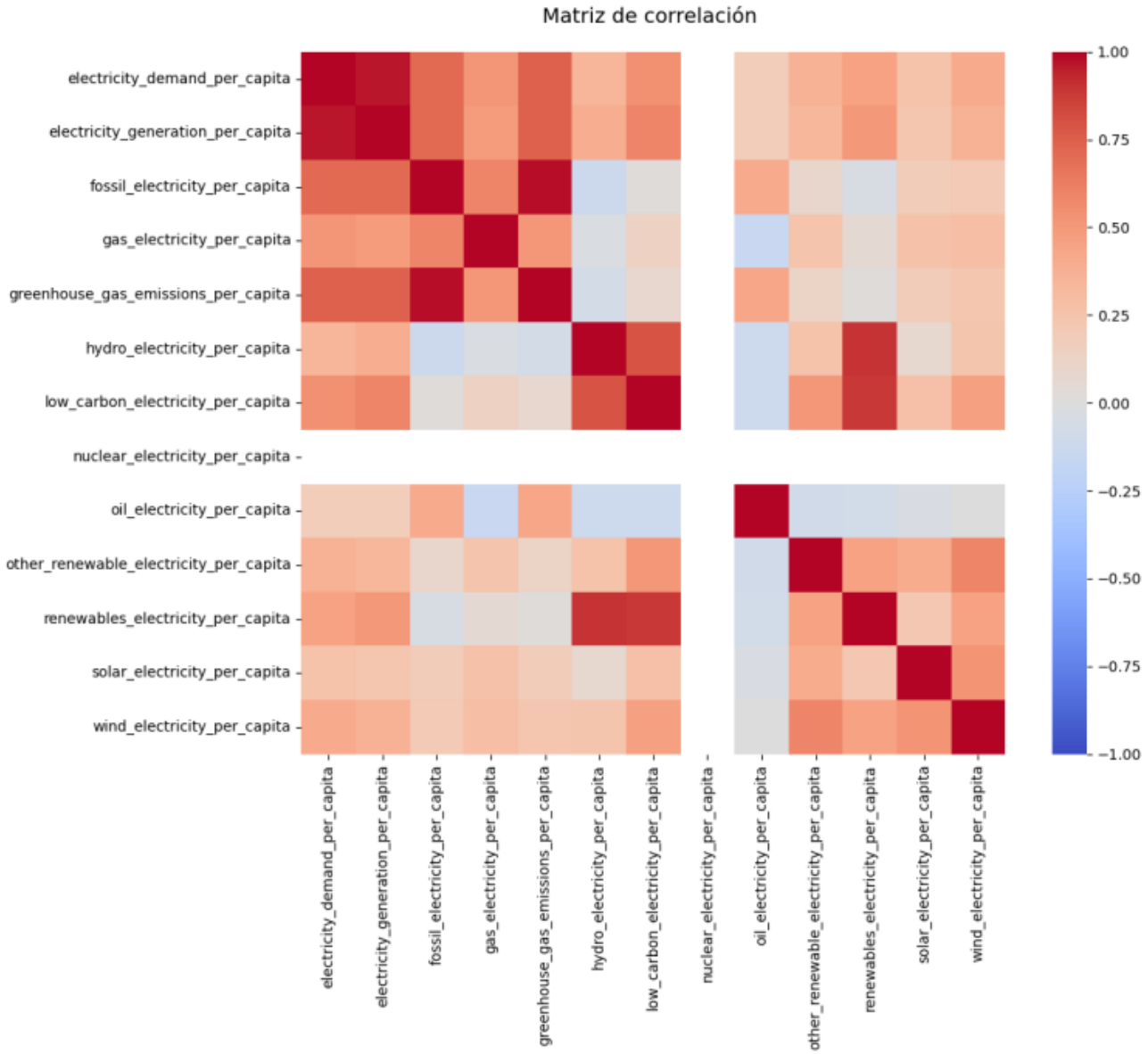

Fig. 2: Anexo 2: Diagrama de cajas de las primeras 10 variables después del preprocesamiento

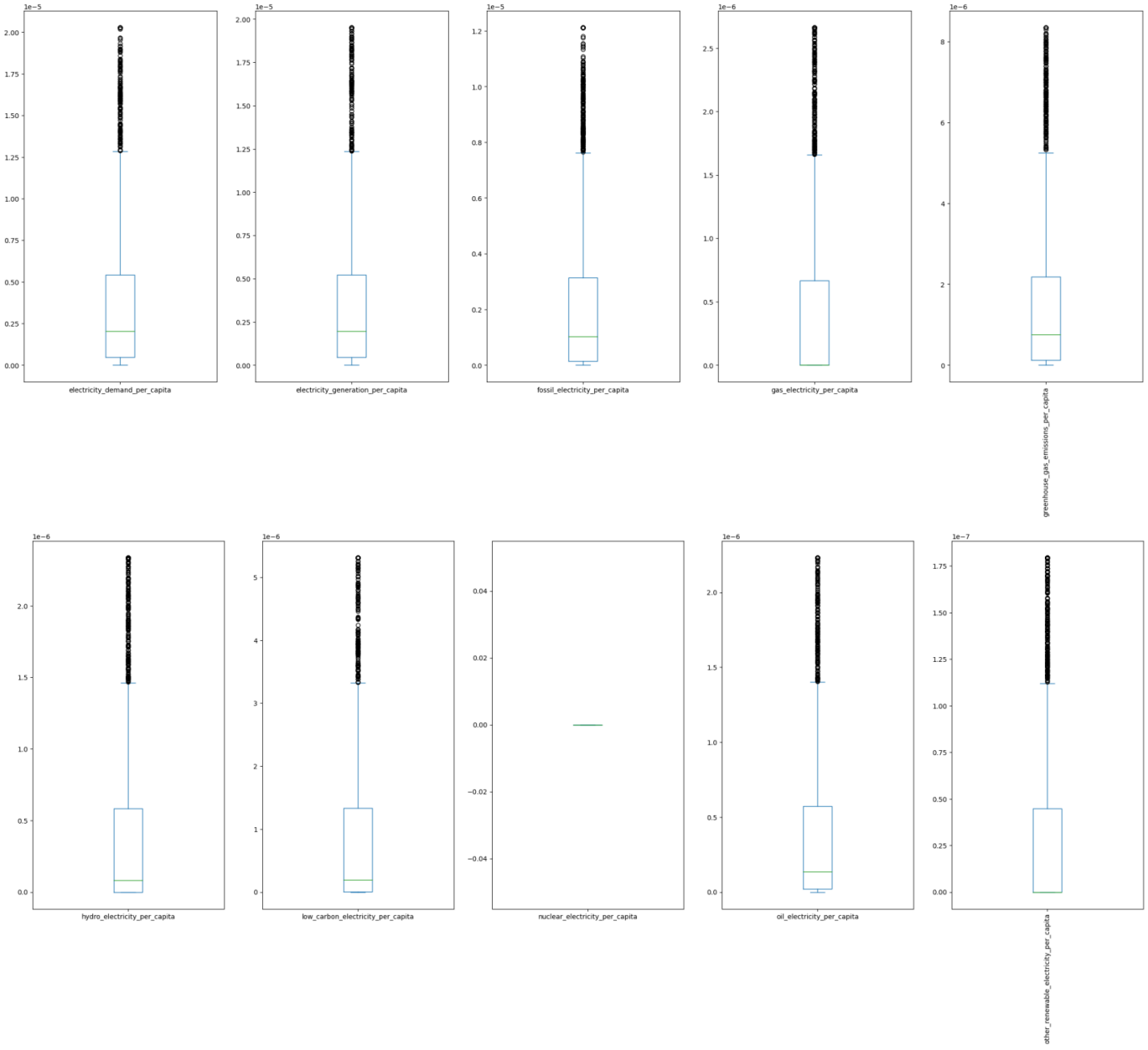

Fig. 3: Anexo 3: PCA

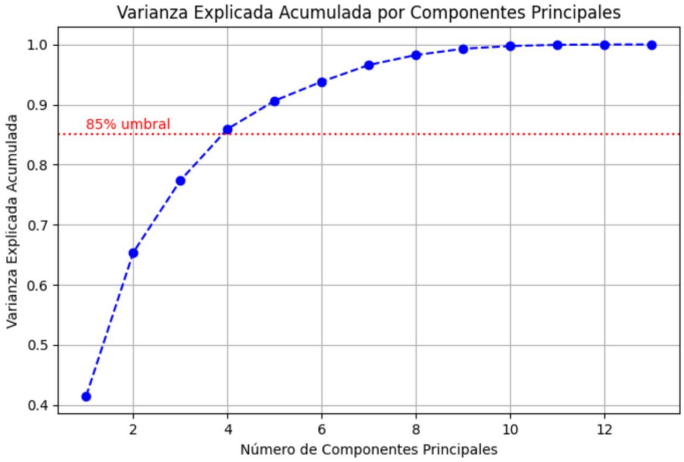

Fig. 4: Anexo 4: Método del codo

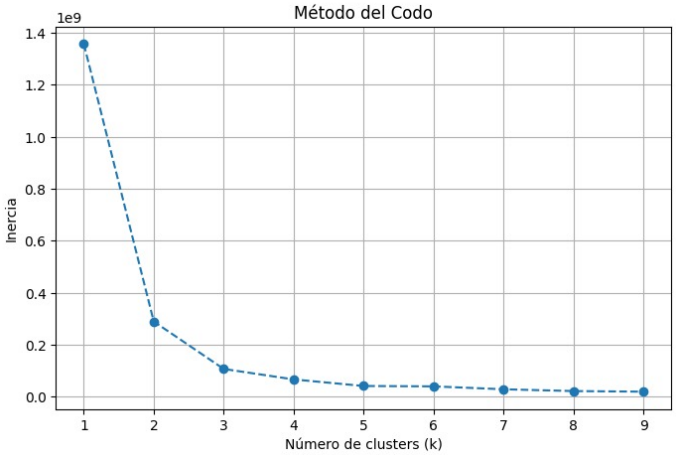

Fig. 5: Anexo 5: Clusters