

# Aplicación de Data Mining para el consumo energético

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

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

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

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

## I. INTRODUCCIÓN

La energía y sus implicancias han sido objeto de debate e inquietud en la comunidad científica a lo largo de la historia debido a su importancia en el desarrollo económico, social y ambiental de un país [1].

La literatura evidencia una estrecha relación entre este recurso y el crecimiento económico, en específico en el Producto Bruto Interno (PBI); además sugiere que la transición hacia fuentes renovables contribuye de manera positiva tanto al desarrollo económico como a la sostenibilidad ambiental [2].

Actualmente, las cuatro quintas partes de la energía primaria consumida a nivel mundial provienen de combustibles fósiles (carbón, petróleo y gas); las cuales, al quemarse para generar energía, producen altas cantidades de dióxido de carbono (CO<sub>2</sub>) [3]. Sin embargo, en los últimos años se observa una tendencia creciente hacia fuentes renovables junto con procesos de electrificación de los sistemas productivos. Por ejemplo, en 2024 la energía solar y eólica se expandieron hasta el 96% del total de nuevas capacidades eléctricas instaladas a nivel mundial [4].

Sin embargo, la transición a energías sostenibles no es la misma tanto para economías desarrolladas y no desarrolladas en términos económicos. Los países con mayor desarrollo económico, al poseer niveles altos de PBI per cápita, pueden pagar el costo de transición. A diferencia de otros países con menor desarrollo cuya transferencia de consumo de energía a fuentes menos contaminantes se realiza mucho más lento en la medida que su economía le permita [2].

El análisis del consumo energético mundial permite identificar patrones históricos, desigualdades entre países y tendencias hacia la transición energética. Por lo mencionado, esta 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*, la cual integra

información histórica sobre variables energéticas, ambientales y socioeconómicas por país y año [5].

Algunos conceptos clave para la investigación son:

- 1) **KDD (Knowledge Discovery in Databases):** Proceso que incluye la selección, transformación, limpieza, minería de datos y evaluación de patrones en grandes volúmenes de información.
- 2) **K-Means:** Algoritmo no supervisado que agrupa datos en clústeres según su similitud, asignando cada punto al centro más cercano.
- 3) **PCA:** Técnica de reducción de dimensionalidad que permite simplificar los datos conservando las variables más informativas.
- 4) **Clustering:** Algoritmo no supervisado que agrupa países con características energéticas similares.

La aplicación de Data Mining (DM) y Machine Learning (ML) se centrará en descubrir patrones ocultos entre variables que pueden estar asociadas a la transición energética.

Además, se aplicará el proceso de KDD que incluirá:

- Limpieza y transformación de los datos energéticos.
- Análisis exploratorio de correlaciones entre consumo, PBI y emisiones.
- Técnicas de DM: agrupamiento de países según su matriz energética y reglas de asociación entre consumo y emisiones.
- Modelos predictivos: estimaciones de consumo y emisiones futuras mediante regresión o series temporales.

En ese sentido, el presente proyecto se traza como objetivo principal:

Identificar patrones globales y factores asociados al consumo energético y emisiones de CO<sub>2</sub> 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<sub>2</sub>.
- 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.

Con este fin, se expondrán primero el Estado del Arte sobre investigaciones relevantes que hayan abordado este mismo problema o similares. Posteriormente, se presentará el diseño del experimento, el cual comprende una descripción del conjunto de datos a usar y la metodología empleada. Finalmente, se explicarán las tareas realizadas en la etapa de experimentación y sus correspondientes resultados.

## II. ESTADO DEL ARTE

La predicción del consumo energético ha ganado relevancia debido a la creciente necesidad de optimizar el uso de energía y reducir las emisiones de gases de efecto invernadero. Con el aumento de los datos disponibles a través de redes inteligentes y medidores inteligentes, los modelos predictivos se han vuelto herramientas esenciales para la gestión energética. Sin embargo, la calidad de los datos juega un papel fundamental en la precisión de las predicciones, ya que los datasets suelen contener valores nulos, outliers y una alta dimensionalidad. Por lo tanto, el preprocesamiento de datos es una fase crítica para asegurar que los modelos sean precisos y eficaces.

En este contexto, el Análisis de Componentes Principales (PCA) es una herramienta robusta utilizada para reducir la dimensionalidad de los datos y mejorar la calidad de los modelos predictivos. PCA es eficaz para eliminar características redundantes y ruidosas, lo que mejora la precisión de los modelos sin perder información crucial [6]. La literatura sugiere que PCA permite identificar las variables más relevantes para el modelo, lo cual es esencial cuando se trabajan con grandes volúmenes de datos que contienen variables como temperatura, humedad o el tipo de energía utilizada [6]. Este enfoque se complementa con el manejo de outliers en cinco etapas que incluye la eliminación de ruido y la identificación de outliers sistemáticos [7]. Este proceso es clave para asegurar que los datos sean de alta calidad antes de ser introducidos en modelos predictivos.

La correcta gestión de datos faltantes y la reducción de dimensionalidad son consideradas las principales dificultades en el preprocesamiento de datos de consumo energético. El preprocesamiento es una fase crítica que define el éxito del proceso analítico, ya que los problemas de valores nulos y outliers pueden afectar seriamente los resultados predictivos [8]. Para abordar estos desafíos, es esencial aplicar técnicas de reducción de dimensionalidad como PCA, que no solo mejora la eficiencia computacional, sino que también facilita la identificación de patrones relevantes en los datos [8].

En cuanto a la predicción del consumo energético a gran escala, la literatura sugiere un enfoque basado en PCA robusto (RPCA) para identificar los factores clave que influyen en el consumo energético, como el PIB, la población, la estructura industrial, la estructura de consumo de energía y la urbanización [9]. Este modelo de predicción energética es fundamental para la toma de decisiones estratégicas, ya que permite a las instituciones gestionar problemas como las emisiones de contaminantes y de carbono, y planificar la inversión en energías renovables.

Por otro lado, en relación a la clusterización, la literatura sugiere que métodos como K-means y clusterwise regression son útiles para agrupar datos con patrones similares de consumo energético. La literatura sugiere usar clusterización para mejorar la precisión de las predicciones de demanda máxima y consumo total de energía al identificar grupos de edificios con comportamientos energéticos similares [10]. Esta técnica, combinada con PCA, permite reducir la dimensionalidad del dataset y mejorar la eficiencia computacional. Además, los resultados indican que la clusterización mejora significativamente la precisión de los modelos predictivos cuando se utiliza en conjunto con modelos de regresión [6].

En resumen, el preprocesamiento de datos y el uso de PCA son esenciales para mejorar la precisión de las predicciones del consumo energético, especialmente cuando se trabaja con datasets complejos que contienen valores faltantes, outliers y dimensionalidad alta. A pesar de los avances en este campo, sigue siendo crucial optimizar las técnicas de imputación y gestión de outliers para asegurar la calidad de los datos y mejorar la eficiencia de los modelos predictivos en el contexto de la gestión energética global.

## III. DISEÑO DEL EXPERIMENTO

### A. Descripción del conjunto de datos

- Medio de obtención: El dataset utilizado corresponde a World Energy Consumption, disponible en la plataforma Kaggle [11].
- Dimensión del dataset: Inicialmente, el conjunto contaba con aproximadamente 22,000 registros y 129 variables, correspondientes a distintos países y años. Las variables incluyen indicadores de consumo y producción total y per cápita, participación de fuentes energéticas (carbón, petróleo, gas, nuclear, hidroeléctrica, solar, eólica, entre otras), emisiones de CO<sub>2</sub>, intensidad energética y datos socioeconómicos relacionados con el PIB y la población. Tras el proceso de depuración, la base quedó reducida a 8,089 registros y 44 variables.
- Tarea de ML: El objetivo del análisis no fue la predicción supervisada, sino la identificación de patrones mediante reducción de dimensionalidad (PCA) y clustering. Por esta razón, no se realizó una división en conjuntos de entrenamiento, validación y prueba, ya que todo el procesamiento se aplicó sobre el dataset completo.
- Análisis exploratorio de los datos: Se realizaron principalmente 2, diagrama de cajas para la evaluación de outliers y un mapa de calor para medir la correlación. Por un lado,

en el análisis inicial de outliers se emplearon diagramas de caja sobre las variables numéricas. Este procedimiento permitió detectar y tratar valores atípicos de acuerdo con la distribución de cada variable. Una curiosidad observada es que, incluso después de aplicar los criterios de depuración, algunos diagramas aún mostraban la presencia de ciertos valores fuera del rango esperado. 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.

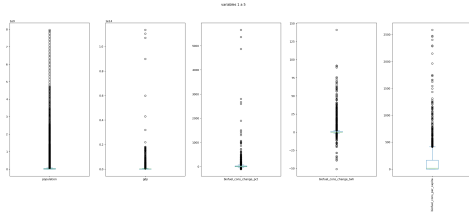


Fig. 1: Diagrama de cajas de las cinco primeras variables

En cuanto a la correlación, se construyó un mapa de calor para evaluar la relación entre variables numéricas. Se identificaron seis pares con correlaciones superiores a 0.7, lo que las catalogó como altamente redundantes: 'hydroshareelec' y 'renewableshareelec' (0.95), 'lowcarbonshareelec' y 'renewableshareelec' (0.91), 'hydroshareelec' y 'lowcarbonshareelec' (0.85), 'hydroelecpercapita' y 'renewableelecpercapita' (0.76), 'electricitygeneration' y 'primaryenergyconsumption' (0.73); y 'hydroelectricity' y 'renewableelectricity' (0.71).

Para evitar problemas de multicolinealidad, se conservaron las variables 'renewableshareelec', 'renewableelecpercapita', 'electricitygeneration' y 'primaryenergyconsumption', eliminándose las demás del análisis.

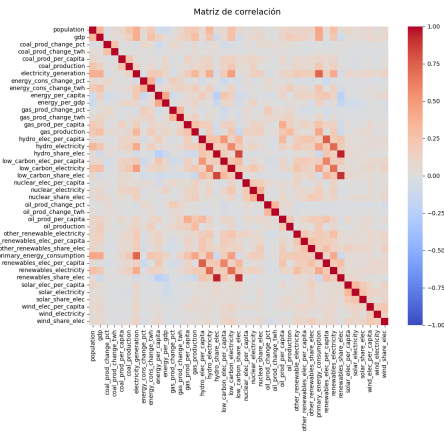


Fig. 2: Matriz de correlación de variables numéricas

## B. Metodología

Con el fin de organizar el trabajo se utilizará la metodología Knowledge Discovery in Databases (KDD). En este apartado se tratarán las secciones de entendimiento del problema, selección

y recopilación de datos; y preprocesamiento de los datos. Luego, en la parte de experimentación, se plantean dos tareas de minería de datos: la primera es un análisis exploratorio para identificar las características más relevantes con PCA y la segunda consta de un modelo de clustering para agrupar las variables por medio de K-means.

1) *Entendimiento del problema:* El presente estudio aborda el problema del análisis y predicción del consumo energético mundial, considerando tanto el crecimiento de la demanda como los cambios en la producción y distribución de energía en diferentes regiones del mundo. Este enfoque es relevante en el contexto actual de la sostenibilidad y la transición hacia energías renovables, dado que la dependencia de los combustibles fósiles y la presión sobre los recursos naturales plantean importantes retos ambientales, económicos y sociales. Se plantea entonces la aplicación de técnicas de Minería de Datos y Aprendizaje Automático con el fin de identificar patrones de consumo y producción de energía, así como de proyectar tendencias que apoyen la toma de decisiones en políticas energéticas y desarrollo sostenible.

2) *Selección y recopilación de datos:* En esta investigación se utilizó el dataset World Energy Consumption, disponible en Kaggle. Esta base de datos recopila información histórica y actualizada sobre el consumo y la producción de energía a nivel global, abarcando diferentes fuentes como carbón, petróleo, gas, nuclear, hidroeléctrica y renovables. El conjunto contiene miles de registros organizados por país y año, incluyendo indicadores de producción, consumo, emisiones de CO<sub>2</sub>, y participación de cada fuente energética en la matriz global.

Las variables principalmente se dividen en los siguientes grupos:

Consumo energético total y per cápita: valores agregados que muestran la cantidad de energía utilizada por cada país y región.

Fuentes de energía específicas: indicadores de producción y consumo desagregados por tipo (carbón, petróleo, gas, nuclear, hidroeléctrica, solar, eólica, etc.).

Emisiones y sostenibilidad: datos sobre emisiones de gases de efecto invernadero y participación de fuentes renovables en el mix energético.

Variables socioeconómicas derivadas: características calculadas como la intensidad energética respecto al PIB o la proporción de consumo por habitante, que permiten evaluar eficiencia y desarrollo.

### 3) Limpieza y preprocesamiento:

- **Eliminación de duplicados:** El primer paso fue la verificación de registros duplicados en la base de datos. Tras la revisión, se confirmó que no existían duplicados, por lo que este procedimiento no generó cambios en la cantidad de registros.
- **Manejo de valores nulos:** El tratamiento de valores faltantes se realizó en dos niveles. A nivel de columnas, se estableció un umbral del 0.7 de valores nulos. Todas aquellas variables que superaban este porcentaje de datos faltantes fueron eliminadas por su bajo aporte al análisis. Y a nivel de filas, se dejaron de lado los registros

con más del 0.5 de valores nulos, ya que contenían muy poca información útil para ser considerados. Como resultado de este proceso, se obtuvo un conjunto de datos más consistente, compuesto por 44 variables finales, con información suficiente para los siguientes pasos del análisis.

- El siguiente paso consistió en la detección y manejo de valores atípicos, considerando únicamente las variables numéricas. Se aplicaron dos metodologías según el comportamiento de cada variable: Para aquellas con distribución normal o cercana a la normalidad, los outliers se identificaron utilizando el criterio de  $\pm 3$  desviaciones estándar. Para las variables no normales, se empleó el método de los cuartiles e IQR (rango intercuartílico), detectando valores fuera del rango  $[Q1 - 1.5 \cdot IQR, Q3 + 1.5 \cdot IQR]$ .

El tratamiento de estos outliers permitió eliminar los registros que perjudicaban el comportamiento de cada variable. Tras este proceso, la base de datos quedó conformada por 8,089 registros.

- Finalmente, se aplicaron técnicas de normalización para asegurar que todas las variables se encontraran en escalas comparables: Para las variables con comportamiento normal o cercano a la normalidad se utilizó la estandarización (media 0, desviación estándar 1). Para las variables con distribuciones no normales, se aplicó un reescalamiento, con el fin de reducir el impacto de magnitudes muy distintas entre variables. Este proceso dejó la base de datos lista para los procedimientos de reducción de dimensionalidad (PCA) y análisis de clustering, garantizando un tratamiento homogéneo de las variables.

## IV. EXPERIMENTACIÓN Y RESULTADOS

### A. Tarea 1. PCA

Se aplicó el Análisis de Componentes Principales (PCA) con el objetivo de reducir la complejidad del conjunto de datos y eliminar redundancias entre variables. A través de este procedimiento, la base de datos pasó de 44 variables originales a 5 componentes principales, manteniendo un alto porcentaje de la varianza explicada. Este resultado permitió condensar la información en dimensiones más representativas, facilitando la interpretación de los patrones y preparando los datos para un agrupamiento más robusto.

### B. Tarea 2. Clustering

Posteriormente, sobre los componentes principales obtenidos con PCA, se realizó un análisis de clustering que permitió agrupar los registros en 3 clusters de comportamiento similar. Estos grupos reflejan diferencias claras en los perfiles energéticos y socioeconómicos, ofreciendo una visión más ordenada del conjunto de datos. Los resultados muestran que el clustering fue efectivo para identificar patrones comunes entre observaciones, lo que abre la posibilidad de comparaciones entre los distintos grupos y sugiere lineamientos diferenciados para futuros análisis o aplicaciones prácticas.

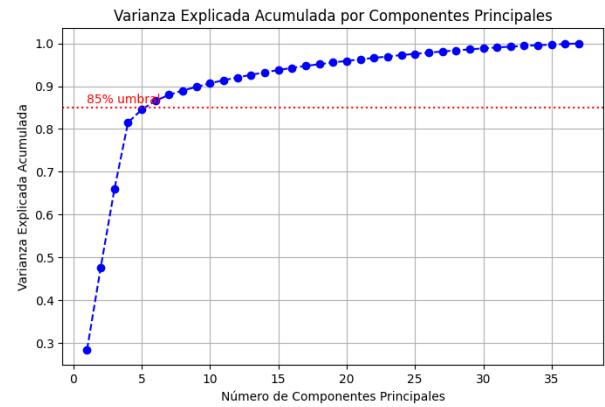


Fig. 3: Componentes del modelo PCA

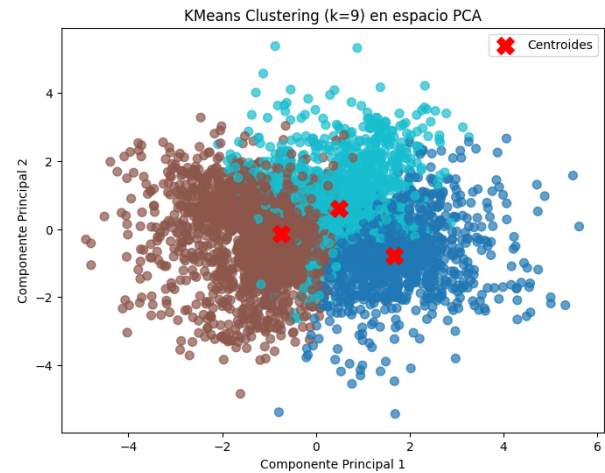


Fig. 4: Clusters respecto a los dos componentes

Es importante mencionar que, 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  $k = 3$ , por lo que finalmente se trabajó con tres clases que permitieron representar de manera equilibrada la estructura del dataset.

## 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] 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>
- [3] Ritchie, H., & Rosado, P. (2017, 2 octubre). Fossil fuels. *Our World In Data*. <https://ourworldindata.org/fossil-fuels>
- [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] Owid. (s.f.). *energy-data/owid-energy-codebook.csv* at master · owid/energy-data. GitHub. <https://github.com/owid/energy-data/blob/master/owid-energy-codebook.csv>

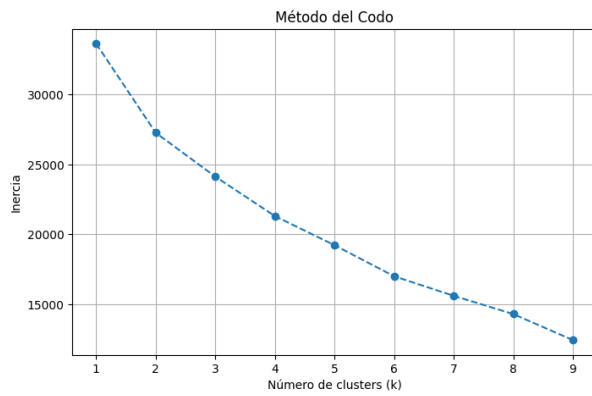


Fig. 5: Método del codo para el Clustering

- [6] 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>
- [7] 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>
- [8] 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>
- [9] 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>
- [10] 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>
- [11] World energy consumption. (2023, 26 noviembre). Kaggle. <https://www.kaggle.com/datasets/pralabhpoudel/world-energy-consumption>