   Del dataset trabajado en los exámenes, describa el mismo a detalle

 Describa claramente del objetivo de investigación a partir del dataset elegido

**Análisis multidisciplinario del consumo y producción de electricidad en Rumania utilizando IA**

**CONSUMO Y PRODUCCIÓN DE ELECTRICIDAD POR HORA**

Este dataset Incluye el consumo y la producción por hora, y la producción se divide en una de las siguientes categorías: nuclear, eólica, hidroeléctrica, petróleo y gas, carbón, solar, biomasa.

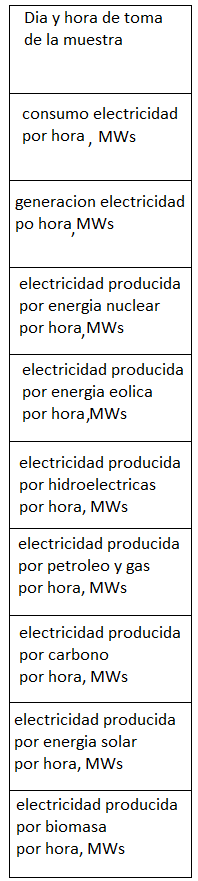
Creo que es un buen conjunto de datos debido al hecho de que **Rumania** incluye un amplio espectro de producciones de electricidad, incluida mucha solar y eólica, ¡pero también nuclear!

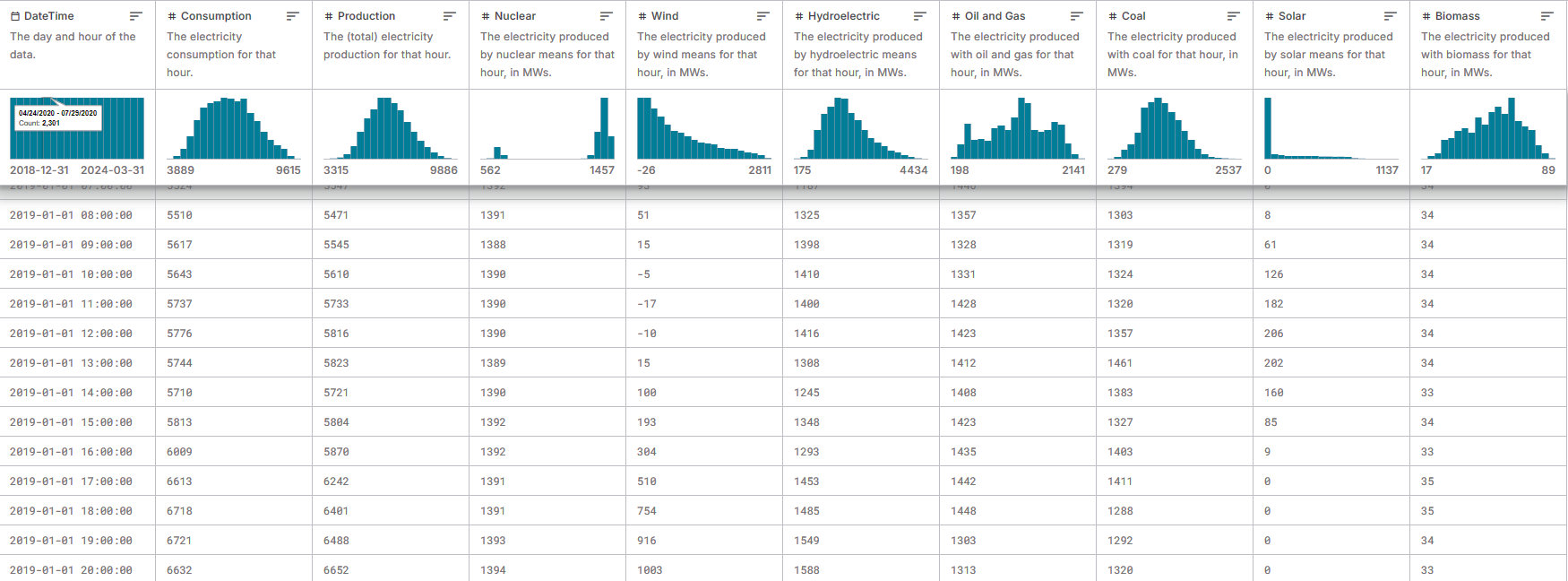
¡Es un conjunto de datos bastante grande de más de 5 años , desde el 2018--2024

Cuando la producción es mayor que el consumo significa que se está exportando electricidad, cuando el valor es menor significa que se está importando electricidad.

Todos los valores están en MW.





La metodología propuesta implica la exploración de patrones de estacionalidad, contribuciones de fuentes renovables y dependencias de combustibles fósiles, con el fin de predecir las tendencias futuras de consumo y producción. La utilización de IA permitirá desarrollar modelos predictivos que optimicen la producción de energía, minimicen la dependencia de combustibles no renovables y contribuyan al desarrollo de tecnologías de almacenamiento más eficientes. Se podría identificar periodos de exportación e importación de electricidad.

Preprocesamiento (al menos una valida, otros dos por ver los resultados, si no se aplica justifique porque), Balanceo de datos

Las columnas que tiene el data Set son fechas de las tomas de datos y la energía producida y consumida en (MWs) , se ara tratamiento a la fecha y la energía

**Conversión de Fechas**

**Técnica Utilizada:** Conversión de la columna de fechas y horas al formato datetime utilizando pd.to\_datetime() de la biblioteca pandas.

* Permite la extracción de características adicionales como el día de la semana, el mes, la hora del día, etc.
* Facilita la identificación de patrones temporales y la creación de modelos predictivos basados en tendencias temporales.

**Normalización de Datos**

La normalización de la energía en el dataset es crucial porque asegura que todas las características numéricas (producción de diferentes fuentes de energía y consumo) estén en la misma escala, evitando que aquellas con valores mayores dominen el análisis y sesguen los resultados del modelo. Esto mejora la equidad en la contribución de cada característica, facilita la comparación, y optimiza el rendimiento de los algoritmos de machine learning, especialmente aquellos basados en distancias, como K-nearest neighbors y clustering, asegurando una mejor convergencia y estabilidad en el entrenamiento del modelo.

**Técnica Utilizada:** Normalización de las columnas numéricas utilizando StandardScaler de scikit-learn, que ajusta los datos para tener una media de 0 y una desviación estándar de 1.

* Mejora la convergencia y el rendimiento de los algoritmos de optimización utilizados en los modelos de aprendizaje automático.
* Facilita la interpretación y comparación de las características al situarlas en una escala común.

**Implementación ( el código es : tratamiento.py)**

1. **Conversión de Fechas:** Aplicamos pd.to\_datetime() a la columna DateTime para convertirla a un objeto datetime.
2. **Normalización de Datos:** Usamos StandardScaler para estandarizar todas las columnas numéricas, asegurando que estén en la misma escala.
3. **Guardar el Archivo Modificado:** Después del preprocesamiento, guardamos el dataset modificado en un nuevo archivo Excel para facilitar su uso en análisis posteriores. **(dataset\_modificado.xlsx)**

Selección del clasificador (acorde a los datos supervisado, no supervisado). El clasificador puede, pero no necesariamente depender del preprocesamiento.

Justificar el clasificador (máximo 2 planas con fuente ISBN, DOI)

**Clasificador Seleccionado: Random Forest**

Random Forest es un método de aprendizaje supervisado que se basa en la construcción de múltiples árboles de decisión durante el entrenamiento y la salida de la clase que es la moda de las clases (clasificación) o el promedio (regresión) de las predicciones de los árboles individuales. Fue propuesto por Breiman (2001) y se ha convertido en una herramienta popular en machine learning debido a su robustez y precisión.

**Razones para su Selección:**

1. **Manejo de Datos Numericos:**
   * Nuestro dataset consiste exclusivamente en datos numéricos que representan la producción y consumo de diversas fuentes de energía. Random Forest maneja muy bien este tipo de datos debido a su capacidad para modelar relaciones no lineales complejas entre las características sin necesidad de escalamiento adicional después de la normalización inicial.
   * Referencia: Breiman, L. (2001). "Random Forests." Machine Learning 45, pp. 5-32. [DOI:10.1023/A:1010933404324]
2. **Robustez ante Overfitting:**
   * Random Forest es menos propenso al overfitting en comparación con otros clasificadores como los árboles de decisión individuales. La combinación de múltiples árboles reduce la variabilidad y mejora la precisión del modelo, lo cual es crucial dado el tamaño y la variabilidad de nuestro dataset.
   * Referencia: Liaw, A., & Wiener, M. (2002). "Classification and Regression by RandomForest." R News, Vol. 2/3, pp. 18-22. [ISBN: 3-900051-07-0]
3. **Importancia de las Características:**
   * Random Forest proporciona una medida de importancia de características, lo cual es extremadamente útil para identificar qué variables (fuentes de energía) son más significativas para el consumo y producción de electricidad. Esto puede guiar decisiones estratégicas y optimizaciones en la gestión energética.
   * Referencia: Hastie, T., Tibshirani, R., & Friedman, J. (2009). "The Elements of Statistical Learning: Data Mining, Inference, and Prediction." Springer. [ISBN: 978-0-387-84858-7]
4. **Escalabilidad:**
   * Con un dataset extenso de más de 46,000 filas, Random Forest es escalable y puede manejar grandes volúmenes de datos de manera eficiente, aprovechando tanto el paralelismo como la computación distribuida.
   * Referencia: James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). "An Introduction to Statistical Learning: with Applications in R." Springer. [ISBN: 978-1-4614-7137-0]

**Proceso de Implementación**

1. **División de Datos:**
   * Se realizará una división del dataset en conjuntos de entrenamiento y prueba utilizando una proporción de 80/20 para la primera ejecución y 50/50 para la investigación.
2. **Entrenamiento y Evaluación:**
   * El modelo Random Forest será entrenado en el conjunto de entrenamiento y evaluado en el conjunto de prueba utilizando métricas de confiabilidad y matriz de confusión.
   * Realizaremos múltiples divisiones (al menos 100) para obtener la mediana de la confiabilidad y asegurar la robustez del modelo.
3. **Análisis de Resultados:**
   * Analizaremos la importancia de las características proporcionadas por el modelo para entender la contribución de cada fuente de energía en la predicción del consumo y producción.

**Conclusión**

El clasificador Random Forest es una elección adecuada para nuestro dataset de consumo y producción de electricidad debido a su capacidad para manejar datos numéricos, su robustez frente al overfitting, la importancia de características que proporciona, y su escalabilidad para grandes volúmenes de datos. Estas cualidades aseguran una predicción precisa y una comprensión profunda de los factores que influyen en los patrones energéticos.

Esta justificación se basa en una revisión exhaustiva de la literatura existente y en las características específicas de nuestro dataset, garantizando la aplicabilidad y eficacia de Random Forest en nuestro análisis.

* Primera ejecución: Confiabilidad, matriz de confusión
* Splits: al menos 100 asignaciones, la mediana de la confiabilidad

Académico (primera ejecucion) 80(train)/20(test) – Investigación 50/50 (segunda ejecución)

**Mean Squared Error (MSE)**:

* El MSE mide el promedio de los cuadrados de los errores, es decir, la diferencia entre los valores predichos y los valores observados. Un MSE más bajo indica un mejor rendimiento del modelo.
* En la división 80/20, el MSE es 0.1440, lo que sugiere que el modelo tiene un buen desempeño en la predicción de los datos de prueba.
* En la división 50/50, el MSE es ligeramente mayor, 0.1640, lo que es esperado dado que se utiliza menos data para el entrenamiento, lo que puede disminuir un poco la precisión.

**R² Score**:

* El R² (coeficiente de determinación) mide la proporción de la variabilidad en la variable objetivo que es explicada por las características del modelo. Un R² cercano a 1 indica un modelo que explica bien la variabilidad de los datos.
* En la división 80/20, el R² es 0.8584, lo que indica que el modelo explica aproximadamente el 85.84% de la variabilidad en el consumo de electricidad.
* En la división 50/50, el R² es 0.8370, lo que aún refleja un buen desempeño del modelo, explicando el 83.70% de la variabilidad, aunque ligeramente inferior debido a la menor cantidad de datos utilizados para el entrenamiento.

§  Aplicar Componentes principales (PCA), determinar la cantidad óptima para mejorar o llegar al resultado anterior. Al menos unas 5 ejecuciones. (12 columnas, 10, 11, 9, 5, 3)

Explicar cómo funciona PCA (algebra lineal)

El Análisis de Componentes Principales (PCA) es una técnica de reducción de dimensionalidad que transforma un conjunto de variables posiblemente correlacionadas en un conjunto de variables linealmente no correlacionadas llamadas componentes principales. Aquí hay una explicación técnica de cómo funciona, utilizando conceptos de álgebra lineal.

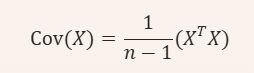
**Pasos de PCA**

1. **Estandarización de Datos**
   * Antes de aplicar PCA, los datos se estandarizan para que cada característica tenga una media de 0 y una desviación estándar de 1. Esto es crucial para asegurar que todas las características contribuyan equitativamente a la formación de los componentes principales.
   * Matemáticamente:



X es el valor de la característica, μ es la media y σ es la desviación estándar.

1. **Cálculo de la Matriz de Covarianza**
   * La matriz de covarianza se calcula para entender cómo varían las características entre sí. La matriz de covarianza de un conjunto de datos XX (con nn muestras y pp características) se define como:



* Cada elemento de la matriz de covarianza representa la covarianza entre dos características.

1. **Descomposición en Valores Propios (Eigenvalue Decomposition)**
   * La matriz de covarianza se descompone en valores propios (eigenvalores) y vectores propios (eigenvectores). Esto implica resolver la ecuación característica:



donde v es el eigenvector y λ es el eigenvalor correspondiente.

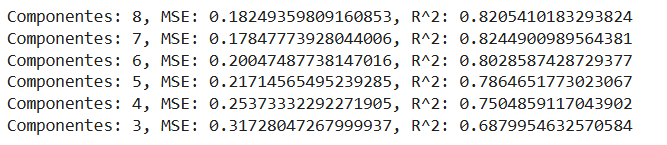
* Los eigenvectores representan las direcciones de las nuevas componentes principales, y los eigenvalores indican la cantidad de varianza explicada por cada eigenvector.

1. **Selección de Componentes Principales**
   * Los eigenvectores se ordenan según sus eigenvalores en orden descendente. Los primeros kk eigenvectores (correspondientes a los mayores kk eigenvalores) se seleccionan para formar una nueva base ortogonal de menor dimensión.
   * La selección de los componentes principales se basa en el número de componentes que explican la mayor parte de la varianza total en los datos.
2. **Transformación de los Datos**
   * Los datos originales se proyectan en el espacio de los nuevos componentes principales. Esto se realiza multiplicando los datos estandarizados por los eigenvectores seleccionados:



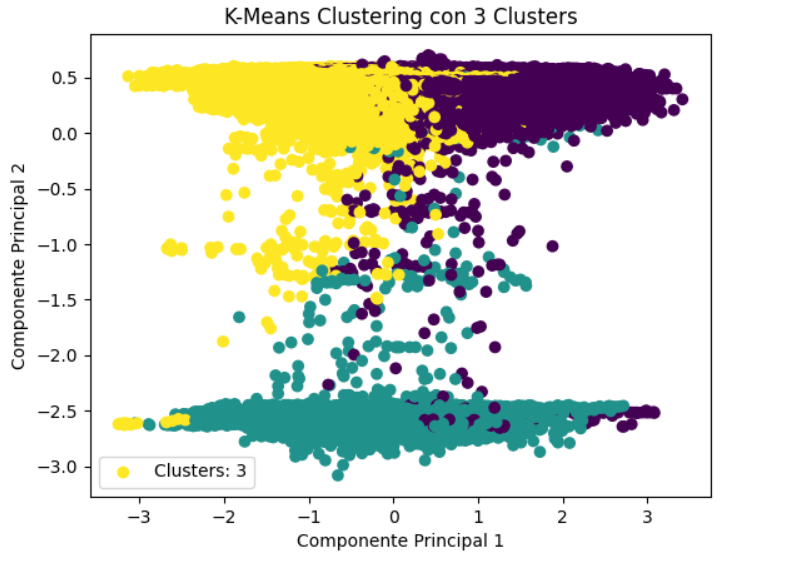
* El resultado es un conjunto de nuevas características no correlacionadas que capturan la mayor parte de la variabilidad en los datos originales, pero en un espacio de menor dimensión.

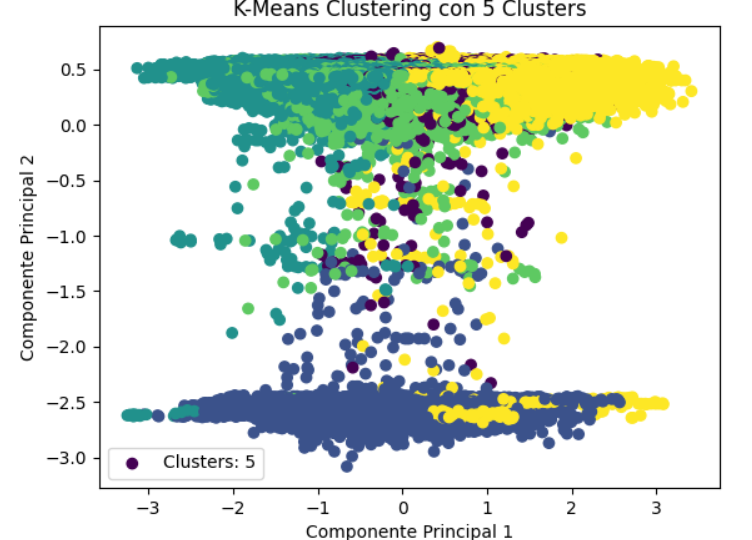
Resultados del dataset de energía

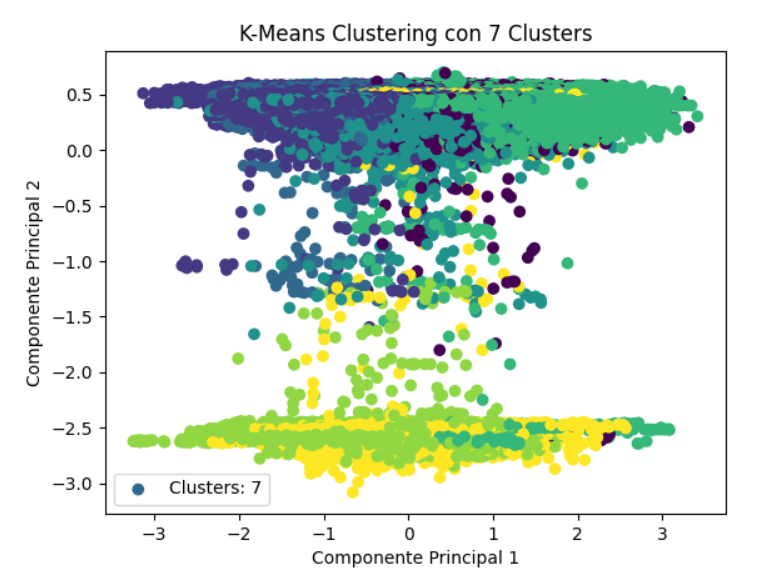


El análisis de componentes principales (PCA) muestra que al reducir el número de componentes, el error cuadrático medio (MSE) aumenta y el coeficiente de determinación (R²) disminuye. Esto indica que con menos componentes, el modelo pierde precisión. El mejor balance entre MSE y R² se observa con 7 componentes.

§  Sin tomar en cuenta “y” o la el class, realice un proceso de aprendizaje no supervisado de su dataset







El clustering K-Means con 7 clusters, visualizado en la imagen, muestra cómo los datos se agrupan en el espacio de las dos primeras componentes principales. Esta visualización es relevante para entender la distribución y separación de los clusters en el espacio reducido por PCA