 Artículo que resuma lo realizado (min 3 hojas)

**Análisis de Datos de Consumo y Producción de Electricidad Utilizando Técnicas de Machine Learning**

**Introducción**

En este artículo, presentamos un análisis exhaustivo del consumo y la producción de electricidad utilizando técnicas de machine learning. El objetivo es comprender patrones en los datos, aplicar métodos de reducción de dimensionalidad, y realizar un proceso de aprendizaje no supervisado para identificar patrones ocultos. Nuestro análisis se basa en un dataset que contiene información detallada sobre el consumo y la producción de diferentes fuentes de energía en intervalos horarios.

**Dataset de Energía**

El dataset utilizado en este análisis contiene registros horarios de consumo y producción de electricidad, con las siguientes columnas:

* **DateTime**: Fecha y hora de la medición.
* **Consumo**: Consumo total de electricidad (en megavatios).
* **Produccion**: Producción total de electricidad (en megavatios).
* **Nuclear**: Producción de electricidad de plantas nucleares.
* **Wind**: Producción de electricidad de fuentes eólicas.
* **Hydroelectric**: Producción de electricidad de plantas hidroeléctricas.
* **Oil and Gas**: Producción de electricidad de plantas de petróleo y gas.
* **Coal**: Producción de electricidad de plantas de carbón.
* **Solar**: Producción de electricidad de fuentes solares.
* **Biomass**: Producción de electricidad de fuentes de biomasa.

**Objetivos del Análisis**

1. **Preprocesamiento y Normalización de Datos**: Mejorar la calidad de los datos para un análisis efectivo.
2. **Selección del Clasificador y Evaluación del Modelo**: Utilizar un algoritmo de machine learning para predecir el consumo de electricidad basado en las diferentes fuentes de producción.
3. **Reducción de Dimensionalidad con PCA**: Simplificar el dataset manteniendo la mayor parte de la variabilidad.
4. **Aprendizaje No Supervisado con K-Means**: Identificar patrones y agrupaciones naturales en los datos sin usar la variable de clase.

**Preprocesamiento y Balanceo de Datos**

**Objetivo:** Mejorar la calidad de los datos para un análisis efectivo.

**Pasos de Preprocesamiento**

1. **Conversión de Fechas**: Convertimos la columna de fechas al formato datetime para facilitar las operaciones temporales. Utilizamos la función pd.to\_datetime().
2. **Normalización de Datos**: Las características numéricas se normalizaron usando StandardScaler para asegurar que todas contribuyan de manera equitativa en el modelo.
3. **Ingeniería de Características**: Extrajimos nuevas características a partir de la columna de fechas, como hora del día y mes, para identificar patrones temporales.

**Justificación:**

* La normalización es esencial para mejorar la convergencia de los algoritmos de machine learning y asegurar un peso equitativo de todas las características.
* La ingeniería de características permite capturar información relevante que puede mejorar el rendimiento del modelo.

**Selección del Clasificador**

**Clasificador Seleccionado**: **Random Forest Regressor**

**Justificación**

* **Manejo de Datos Numéricos**: Random Forest maneja eficientemente datos numéricos, modelando relaciones no lineales complejas.
  + Referencia: Breiman, L. (2001). "Random Forests." Machine Learning 45, pp. 5-32. [DOI:10.1023/A:1010933404324]
* **Robustez ante Overfitting**: Combina múltiples árboles para reducir la variabilidad y mejorar la precisión.
  + Referencia: Liaw, A., & Wiener, M. (2002). "Classification and Regression by RandomForest." R News, Vol. 2/3, pp. 18-22. [ISBN: 3-900051-07-0]
* **Importancia de las Características**: Proporciona medidas de importancia de características, útiles para identificar variables relevantes.
  + 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]

**Evaluación del Modelo**

**Primera Ejecución**

**80/20 Split**:

* **MSE**: 0.1440
* **R²**: 0.8584

**50/50 Split**:

* **MSE**: 0.1640
* **R²**: 0.8370

**Conclusión**:

* El modelo Random Forest demuestra un rendimiento robusto con altas puntuaciones de R² y bajos valores de MSE en ambas divisiones. Esto indica que el modelo explica bien la variabilidad en el consumo de electricidad.

**Splits en 100 Asignaciones**

**Proceso**:

* Utilizamos Stratified K-Fold para realizar 100 splits, asegurando una evaluación robusta del rendimiento del modelo.

**Mediana del MSE**: [Valor calculado] **Mediana del R²**: [Valor calculado]

**Aplicación de Componentes Principales (PCA)**

**Objetivo**: Reducir la dimensionalidad del dataset mientras se mantiene la mayor parte de la variabilidad.

**Pasos del PCA**

1. **Estandarización de Datos**: Normaliza los datos para tener media 0 y desviación estándar 1.
2. **Cálculo de la Matriz de Covarianza**: Calcula la matriz de covarianza de los datos estandarizados.
3. **Descomposición en Valores Propios**: Realiza la descomposición en valores propios (eigenvalue decomposition) de la matriz de covarianza para encontrar los eigenvalores y eigenvectores.
4. **Selección de Componentes Principales**: Los eigenvectores correspondientes a los mayores eigenvalores forman los componentes principales. Los eigenvalores representan la cantidad de varianza explicada por cada componente.

**Resultados del PCA**:

* **Componentes: 8**: MSE: 0.1825, R²: 0.8205
* **Componentes: 7**: MSE: 0.1785, R²: 0.8245
* **Componentes: 6**: MSE: 0.2005, R²: 0.8029
* **Componentes: 5**: MSE: 0.2171, R²: 0.7865
* **Componentes: 4**: MSE: 0.2537, R²: 0.7505
* **Componentes: 3**: MSE: 0.3173, R²: 0.6880

**Conclusión**:

* La mejor relación entre MSE y R² se observa con 7 componentes. A medida que se reducen los componentes, la precisión disminuye, indicando pérdida de información.

**Aprendizaje No Supervisado**

**Algoritmo Utilizado**: **K-Means Clustering**

**Resultados del K-Means:**

* Realizamos clustering con 3, 5 y 7 clusters, visualizando los resultados en el espacio de las dos primeras componentes principales.

**Conclusión**:

* Los clusters identificados ayudan a comprender la distribución y separación de los datos en el espacio reducido por PCA, revelando patrones ocultos.