Instituto Tecnológico y de Estudios Superiores de Monterrey

Campus Querétaro

Momento de Retroalimentación: Módulo 2

Análisis Preliminar del Modelo de Predicción del Desempeño Energético Neto de Plantas de Ciclo Combinado

Alumno:

Tomás Pérez Vera – A01028008

TC3006C.101

Inteligencia Artificial Avanzada Para la Ciencia de Datos I

Fecha:

31 de agosto de 2025

1. Introducción

1.1 Descripción del problema

El problema por resolver es la **predicción del desempeño energético** de una planta de energía de ciclo combinado (CCPP) que opera a carga máxima. La eficiencia y la producción de este tipo de plantas no son constantes y se ven directamente afectadas por las **condiciones ambientales**, como la temperatura, la presión, la humedad y el vacío de escape. La capacidad de predecir con precisión la producción de energía neta por hora (EP) es crucial para **optimizar la operación**, planificar el mantenimiento, y gestionar de manera eficiente los costos de combustible. El objetivo es construir un modelo que pueda aprender la compleja relación entre estas variables ambientales y la producción de energía.

1.2 Objetivo del análisis

El objetivo principal de este análisis es **desarrollar e implementar un modelo de regresión lineal** para predecir la producción de energía eléctrica neta por hora (EP) de una CCPP. El modelo utilizará datos históricos para **entrenar** un algoritmo que pueda generalizar y hacer predicciones precisas sobre nuevas condiciones ambientales. Se busca demostrar que las variables de entrada seleccionadas son suficientes para estimar la variable objetivo, y que el modelo es robusto, medido por métricas de rendimiento como el **error cuadrático medio (MSE)** y el **coeficiente de determinación (R^2)**.

2. Descripción del conjunto de datos

2.1 Origen y estructura de los datos

El conjunto de datos contiene **9568 puntos de datos** de una CCPP recopilados durante 6 años (2006-2011). Los datos se organizan en un archivo Folds5x2_pp_copy.csv con una estructura tabular. Cada fila representa un registro de una hora y las columnas corresponden a las variables ambientales y la producción de energía neta.

2.2 Variables utilizadas en el modelo

El modelo utiliza un conjunto de **cuatro variables de entrada (características)** para predecir una **variable de salida (objetivo)**:

- **Variables de entrada:**
  - **Temperatura ambiente (AT):** Temperatura del aire.
  - **Vacío de escape (V):** Presión del aire de escape.
  - **Presión ambiente (AP):** Presión barométrica.

  - **Humedad relativa (RH):** Porcentaje de humedad en el aire.
- **Variable de salida (a predecir):**
  - **Salida de energía eléctrica neta (PE):** La cantidad de electricidad generada por la planta en esa hora.

2.3 Preprocesamiento y transformación de datos

Antes de entrenar el modelo, los datos fueron preprocesados para mejorar la estabilidad y el rendimiento del algoritmo. La **estandarización** fue la técnica principal aplicada. Esta transformación ajusta los valores de las características y el objetivo para que tengan una media de cero y una desviación estándar de uno. Esto evita que una característica con un rango de valores mayor (por ejemplo, la presión) domine el cálculo del gradiente y el aprendizaje del modelo, asegurando que todas las variables contribuyan de manera equitativa. La fórmula utilizada para la estandarización es:

$$X = \frac{(X - \mu)}{\sigma}$$

3. Implementación del modelo

3.1 Separación de datos (Entrenamiento/Prueba)

El conjunto de datos se dividió en dos subconjuntos para evaluar el rendimiento del modelo de manera justa. El **conjunto de entrenamiento** se utilizó para que el algoritmo aprendiera los parámetros óptimos, mientras que el **conjunto de prueba** se reservó para evaluar el modelo con datos que nunca había visto.

- **Conjunto de entrenamiento:** Incluye el **70% de los datos**. Se utiliza para entrenar el modelo de regresión lineal a través del algoritmo de descenso de gradiente.
- **Conjunto de prueba:** Incluye el **30% restante de los datos**. Se utiliza para probar el modelo final y calcular su precisión con la métrica R^2, verificando su capacidad para generalizar en nuevos datos.

4. Evaluación del modelo

4.1 Métricas utilizadas

Una vez que el modelo ha sido entrenado, es crucial evaluar su rendimiento para entender qué tan bien predice la producción de energía en nuevos datos. La evaluación se realiza utilizando métricas que cuantifican la diferencia entre los valores predichos por el modelo y los valores reales.

En la implementación, se utilizaron dos métricas principales para evaluar el modelo de regresión:

- **Error Cuadrático Medio (MSE):** Esta métrica mide el error promedio de las predicciones del modelo. Se calcula promediando los cuadrados de las diferencias entre los valores predichos y los valores reales. Al elevar al cuadrado los errores, el MSE penaliza de forma más severa los errores grandes. Tu función cost_function calcula el MSE, que se utiliza para visualizar la convergencia del modelo durante el entrenamiento. Un valor de MSE bajo indica que el modelo tiene un buen rendimiento. La fórmula es:

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(\hat{y} - y)^2$$

- **Coeficiente de Determinación (R2):** También conocido como R-cuadrado, esta métrica indica la proporción de la varianza en la variable dependiente (EP) que puede ser explicada por las variables independientes (AT, V, AP, RH) en el modelo. El valor de R2 varía de 0 a 1.
    - Un **valor de 1** indica que el modelo explica perfectamente la variabilidad de la variable de salida (las predicciones coinciden exactamente con los valores reales).
    - Un **valor de 0** indica que el modelo no explica nada de la variabilidad, lo que significa que no hay relación entre las variables de entrada y la de salida.

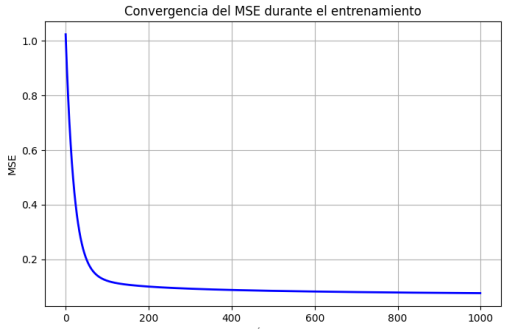
4.2 Resultado de métricas

- **Coeficiente de Determinación (R2):** 0.9231
- **Error Cuadrático Medio (MSE):** 0.075

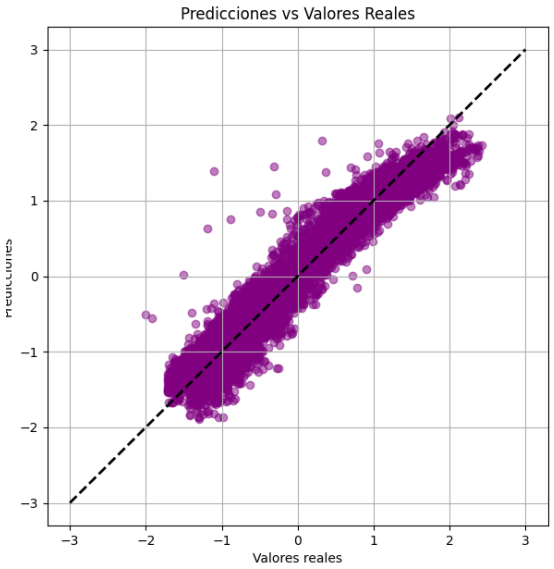
El modelo de regresión lineal ha demostrado un rendimiento excelente, logrando explicar más del 92% de la variabilidad en la producción de energía eléctrica de la planta. Esto confirma que las variables ambientales utilizadas son cruciales para predecir el desempeño de la CCPP. Un MSE final muy bajo indica que, en promedio, las predicciones del modelo se ajustan de manera precisa a los valores reales, lo que lo hace muy confiable para su uso.

4.3 Gráficas de desempeño del modelo

- Pérdida de entrenamiento



- Comparación de predicciones y valores reales en prueba



Como se puede ver en ambas gráficas, el modelo de regresión lineal ha sido entrenado y evaluado con éxito. La gráfica de convergencia del MSE muestra que el **error disminuyó consistentemente**, indicando que el modelo aprendió de manera eficiente. Por otro lado, la gráfica de "Predicciones vs Valores Reales" revela un **alto grado de precisión**, ya que los puntos se agrupan cerca de la línea ideal. Sin embargo, se nota una mayor dispersión en los valores extremos, lo que sugiere que el modelo podría tener más dificultad para predecir en esas condiciones. A pesar de esto, el modelo demuestra una **gran capacidad de generalización** y es una herramienta confiable.

5. Conclusiones y recomendaciones

5.1 Desempeño general del modelo

El modelo de regresión lineal desarrollado para predecir la producción de energía neta de la CCPP mostró un **excelente desempeño general**. Con un **coeficiente de determinación (R2) de 0.9231**, el modelo logró explicar más del 92% de la variabilidad en la producción de energía a

partir de las condiciones ambientales. Esto valida la hipótesis de que variables como la temperatura, presión, humedad y el vacío son los **factores clave** que influyen en el rendimiento de la planta. El bajo **Error Cuadrático Medio (MSE) final de 0.075** confirma que las predicciones del modelo se acercan mucho a los valores reales, lo que lo hace una herramienta **altamente confiable** para la optimización y planificación de las operaciones de la planta.

5.2 Posibles mejoras futuras

A pesar de su sólido desempeño, el modelo puede ser mejorado. La ligera dispersión observada en los valores extremos sugiere que un modelo lineal podría no capturar toda la complejidad de los datos. Para mejorar la precisión, especialmente en condiciones ambientales fuera del rango promedio, se recomienda considerar los siguientes enfoques:

- **Explorar modelos no lineales:** Se pueden implementar algoritmos de aprendizaje automático más avanzados como **Árboles de Decisión, Bosques Aleatorios (Random Forest) o Redes Neuronales**. Estos modelos tienen la capacidad de aprender relaciones complejas y no lineales, lo que podría reducir el error en los valores extremos y mejorar el rendimiento general.
- **Validación cruzada K-fold:** Aunque se usó una división 70/30, una validación cruzada más robusta, como **K-fold**, proporcionaría una evaluación más completa y confiable del modelo, al asegurar que el rendimiento no dependa de una sola división específica de los datos.

Conclusión

El análisis realizado demuestra que el **aprendizaje automático es una herramienta poderosa** para predecir el rendimiento de una planta de energía de ciclo combinado. El modelo de regresión lineal es un excelente punto de partida que proporciona predicciones muy precisas. Las futuras mejoras, a través del uso de modelos más sofisticados, podrían llevar la precisión a niveles aún mayores, permitiendo una gestión más eficiente y rentable de la planta de energía.

Bibliografía

Ibm. (2025, 15 abril). CCPA Compliance. IBM. https://www.ibm.com/think/topics/ccpa-compliance

Tfekci, P. & Kaya, H. (2014). Combined Cycle Power Plant [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5002N.