Tecnológico de Monterrey

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:

11 de septiembre de 2025

Abstract **- El presente trabajo busca exponer un modelo de regresión lineal para predecir la producción de energía eléctrica neta de una central eléctrica de ciclo combinado (CCPP) a plena carga. El objetivo es estimar la producción horaria de energía eléctrica (EP) a partir de variables ambientales como la temperatura (T), la presión ambiente (AP), la humedad relativa (RH) y el vacío de escape (V). El modelo fue entrenado utilizando un algoritmo de descenso de gradiente estocástico (SGD) implementado desde cero y comparado con una implementación equivalente de la librería sklearn para validación.**

Se empleó un conjunto de datos de 9568 puntos recopilados durante 6 años. Se aplicó un preprocesamiento de estandarización a todas las variables para mejorar la convergencia del modelo. Los datos se dividieron en conjuntos de entrenamiento, validación y prueba para evaluar el rendimiento. Los resultados muestran una alta precisión predictiva, con un coeficiente de determinación (R2) de 0.9392 y un error cuadrático medio (MSE) de 0.0608 en el conjunto de prueba, lo que demuestra la capacidad del modelo para capturar la relación entre las variables ambientales y la producción de energía. Este trabajo reafirma la aplicabilidad de los métodos de aprendizaje automático para el monitoreo y la optimización del rendimiento en centrales eléctricas.

Keywords **– Machine Learning, Regresión Lineal, Descenso de Gradiente, Python, Análisis de Datos, Predicción de Energía, Central Eléctrica de Ciclo Combinado (CCPP), Rendimiento del Modelo, Métricas de Evaluación, Error Cuadrático Medio (MSE), Coeficiente de Determinación, Estandarización de Datos, Ingeniería de Características, Preprocesamiento de Datos**

1. INTRODUCCIÓN

1.1 Descripción del problema

La creciente demanda global de energía eléctrica hace imperativo optimizar la eficiencia de las infraestructuras de generación. Las centrales eléctricas de ciclo combinado (CCPP), al integrar turbinas de gas y de vapor, representan una tecnología de alta eficiencia. Sin embargo, su rendimiento está intrínsecamente ligado a las condiciones ambientales circundantes. Factores como la temperatura, la presión, la humedad y el vacío de escape tienen un impacto directo en la cantidad de energía que la planta puede generar. Modelar esta relación de manera precisa no es solo un desafío técnico, sino una necesidad operativa para la toma de decisiones, la gestión de recursos y el mantenimiento predictivo. Un modelo robusto y fiable permitiría a los operadores anticipar la producción de energía bajo diversas condiciones climáticas, identificar anomalías en el rendimiento y planificar estrategias para maximizar la producción.

1.2 Objetivo del análisis

El objetivo principal de este análisis es desarrollar un modelo de regresión lineal capaz de predecir la producción de energía eléctrica neta y horaria de una central de ciclo combinado a plena carga. Este modelo utilizará cuatro variables ambientales como predictores: la temperatura (AT), el vacío de escape (V), la presión ambiente (AP) y la humedad relativa (RH). Para lograrlo, se seguirá un enfoque de aprendizaje automático que incluye:

- Implementación de un modelo desde cero: Se construirá un modelo de regresión lineal utilizando un algoritmo de descenso de gradiente (Gradient Descent) para comprender los

mecanismos subyacentes del entrenamiento.
- Validación del modelo: La implementación manual se validará y comparará con un modelo de referencia de la biblioteca `scikit-learn` para asegurar la precisión y fiabilidad de los resultados.
- Evaluación del rendimiento: Se medirán las métricas de desempeño del modelo, como el Error Cuadrático Medio (MSE) y el Coeficiente de Determinación (R2), para cuantificar su capacidad predictiva en un conjunto de datos de prueba no visto previamente.

2. DESCRIPCIÓN DEL CONJUNTO DE DATOS

2.1 Origen y estructura de datos

El conjunto de datos utilizado en este estudio proviene de una Central Eléctrica de Ciclo Combinado (CCPP) operada a plena carga entre los años 2006 y 2011. Este dataset fue recopilado con el objetivo de modelar la producción de energía eléctrica de la planta. Contiene un total de 9568 puntos de datos, donde cada punto representa mediciones horarias promedio de las variables ambientales de entrada y la energía eléctrica de salida. La estructura del archivo CSV incluye una fila de encabezados y un total de 9568 filas con cinco columnas que corresponden a las variables de interés

2.2 Variables utilizadas en los modelos

El modelo de regresión se basa en un conjunto de cinco variables. Cuatro de ellas son variables independientes (características o *features*) que influyen en la producción de energía, mientras que una es la variable dependiente (objetivo o *target*).

- Variables de entrada (Features):
    - Temperatura (AT): Temperatura ambiente en grados Celsius (°C).
    - Vacío de escape (V): Vacío de escape en cm de mercurio (cm Hg), relacionado con el rendimiento de la turbina de vapor.
    - Presión ambiente (AP): Presión ambiente en milibares (mbar).
    - Humedad relativa (RH): Humedad relativa en porcentaje (%).
- Variable de salida (Target):
    - Producción de energía eléctrica neta (PE): La energía eléctrica neta producida por la planta, medida en megavatios (MW).

2.3 Preprocesamiento y transformación de datos

Para asegurar una convergencia eficiente del modelo y optimizar su rendimiento, el conjunto de datos fue sometido a un proceso de preprocesamiento estándar. La técnica principal aplicada fue la estandarización de características. Cada variable (tanto las de entrada como la de salida) se transformó restando su media y dividiendo por su desviación estándar. Esta transformación asegura que todas las variables tengan una media de cero y una desviación estándar de uno.

La estandarización de datos se define mediante la ecuación:

$$Z = \frac{X - \mu}{\sigma}$$

Donde x es el valor original, μ es la media de la variable y σ es la desviación estándar de la variable.

Este paso es crucial para algoritmos como el descenso de gradiente, ya que evita que las variables con rangos de valores mayores dominen la función de costo, permitiendo que el algoritmo converja de manera más rápida y estable. Una vez estandarizados, los datos se dividieron en conjuntos de entrenamiento (70%), validación (20%) y prueba (10%) para el entrenamiento y la evaluación del modelo.

3. IMPLEMENTACIÓN DE MODELOS

3.1 Separación de datos

Para garantizar una evaluación imparcial del modelo y evitar el sobreajuste (overfitting), el conjunto de datos se dividió en tres subconjuntos distintos: entrenamiento, validación y prueba. Esta metodología es crucial en el aprendizaje automático para simular el rendimiento del modelo con datos nuevos y no vistos. La división se realizó de la siguiente manera:

Conjunto de Entrenamiento (70% de los datos): Utilizado para entrenar el modelo. El algoritmo de descenso de gradiente ajusta los parámetros (pesos y sesgo) iterativamente con el objetivo de minimizar la función de costo en este subconjunto.

Conjunto de Validación (20% de los datos): Utilizado para monitorear el progreso del entrenamiento. En cada época, se calcula el error del modelo en este conjunto. La curva de error de validación ayuda a detectar el sobreajuste; si el error de entrenamiento sigue disminuyendo pero el de validación comienza a aumentar, es una señal de que el modelo está memorizando el ruido de los datos de entrenamiento. Se utiliza una métrica de error en este conjunto como criterio de parada para el entrenamiento.

Conjunto de Prueba (10% de los datos): Utilizado para la evaluación final del modelo una vez que el entrenamiento ha concluido. Este conjunto de datos se mantiene completamente separado y el modelo no tiene acceso a él durante el entrenamiento. Esto permite obtener una estimación objetiva y realista del rendimiento del modelo en un escenario de producción con datos completamente nuevos.

La separación de los datos en estos tres conjuntos garantiza que la evaluación del modelo sea robusta y que las métricas de rendimiento finales reflejen su capacidad de generalización.

4. EVALUACIÓN DEL MODELO

4.1 Métricas utilizadas

Para evaluar el rendimiento y la precisión del modelo de regresión lineal, se utilizaron dos métricas clave ampliamente reconocidas en el ámbito del aprendizaje automático: el Error Cuadrático Medio (MSE) y el Coeficiente de Determinación.

Estas métricas proporcionan una visión completa de la capacidad del modelo para predecir los valores de la producción de energía eléctrica.

- Error Cuadrático Medio (MSE): El MSE es una de las métricas de error más comunes para problemas de regresión. Mide la diferencia promedio al cuadrado entre los valores predichos por el modelo y los valores reales. Al elevar al cuadrado las diferencias, esta métrica penaliza más los errores grandes, lo que la hace muy útil para identificar modelos con grandes desviaciones en algunas predicciones. Un valor de MSE cercano a cero indica un modelo con un error de predicción bajo, mientras que

un valor más alto sugiere un menor rendimiento.

La ecuación que describe el MSE es la siguiente:

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

Donde n es el número de observacione y_i es el valor real y $\hat{y}_i$ es el valor predicho.

- Coeficiente de Determinación: El coeficiente de determinación, es una métrica que representa la proporción de la varianza en la variable dependiente que es predecible a partir de las variables independientes. Se utiliza para medir qué tan bien los valores predichos se ajustan a los valores reales. El valor del coeficiente de determinación va de 0 a 1. Un valor de 1 indica que el modelo explica toda la variabilidad de la variable de respuesta, mientras que un valor de 0 indica que el modelo no explica la variabilidad de los datos. Un coeficiente de determinación más alto, cercano a 1, significa que el modelo es más preciso.

  La ecuación que describe el coeficiente de determinación es:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

  Donde SS_{res} es la suma de los cuadrados residuales (la suma de los errores al cuadrado) y SS_{tot} es la suma de los cuadrados totales (la variabilidad de los datos reales).

4.2 Resultados de métricas utilizadas

a) Implementación de regresión lineal desde cero:
   - $MSE = 0.0896$
   - $R^2 = 0.9159$

b) Implementación de regresión lineal con uso de framework:
   - $MSE = 0.0717$
   - $R^2 = 0.9249$

4.3 Gráficas obtenidas de desempeño de los modelos

a) Modelo de regresión lineal desde cero:

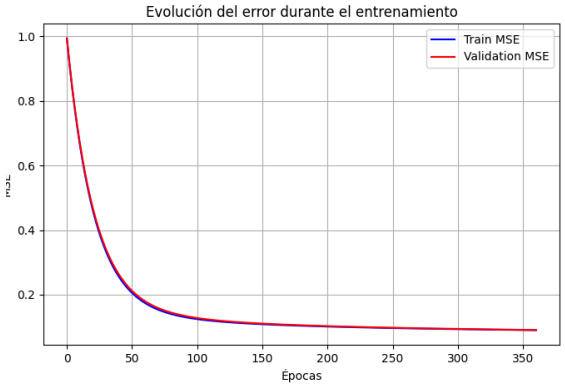

Figura 1. Evolución del MSE del conjunto de entrenamiento y validación

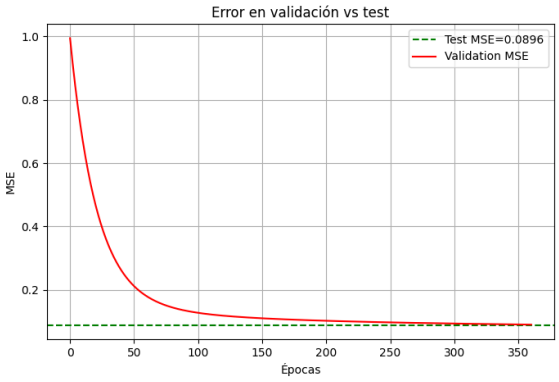

Figura 2. Comparativa MSE del conjunto de validación y prueba

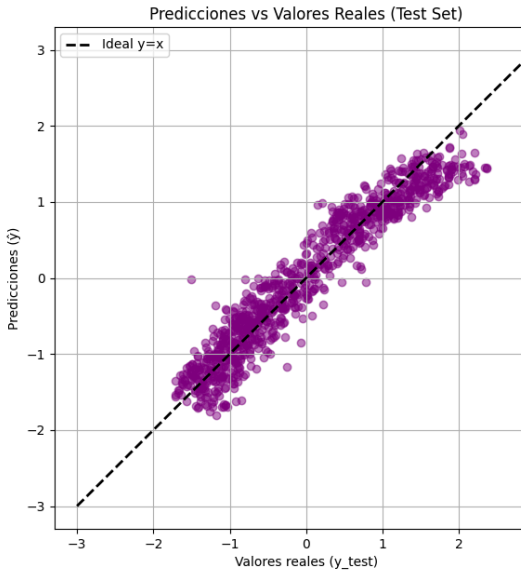

Figura 3. Predicciones hechas con parámetros resultantes del modelo

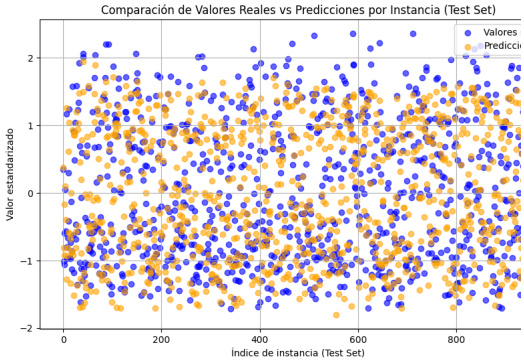

FIgura 4. Comparativa de valores reales contra predicciones en torno a las instancias

b) Modelo de regresión lineal con uso de framework:

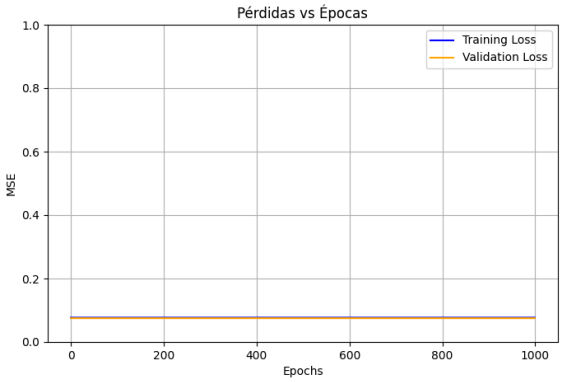

FIgura 5. Comparativa de MSE entre conjunto de entrenamiento y validación

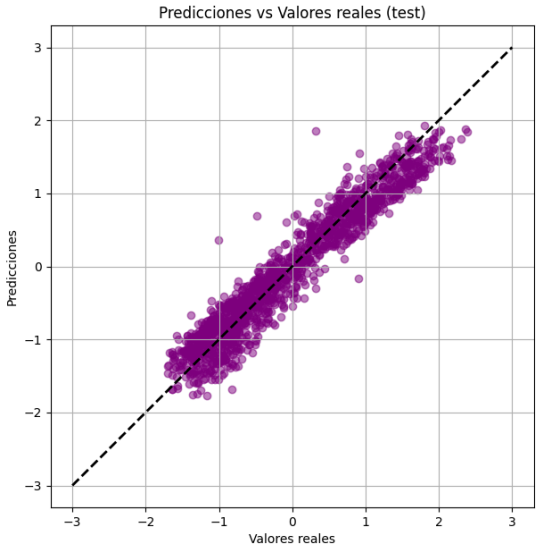

Figura 6. Predicciones hechas con los parámetros obtenidos del modelo

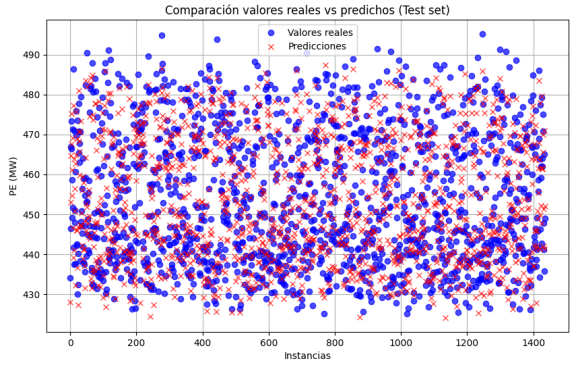

Figura 7. Comparativa de valores reales contra predicciones en torno a las instancias

4.4 Interpretación de gráficas obtenidas

Con base en las gráficas obtenidas para cada caso de implementación, tanto en la implementación desde cero como la implementación con el uso de framework, se puede observar que en ambos casos los modelos logran entrar en el escenario de fitting, dado que los errores cuadráticos medios en tanto en el conjunto de datos de entrenamiento y datos de validación, no presentan importantes diferencias entre sí.

De igual forma, se puede concluir que ambos modelos logran un poder de predicción óptimo para el fenómeno a modelar, ya que para el caso de la implementación del modelo desde cero, este logra obtener un coeficiente de determinación de 0.9159, lo cual implica que el modelo logra explicar el 91.59% de la varianza de los datos, mientras que el modelo con implementación con el uso de framework, logra obtener un coeficiente de determinación de 0.9249, lo que significa que este explica el 92.49% de la varianza en los datos, teniendo así una mejora del 0.9% con respecto a la implementación desde cero.

5. CONCLUSIONES Y RECOMENDACIONES

5.1 Desempeño general de los modelos

Los resultados obtenidos confirman que la regresión lineal es una herramienta altamente efectiva para modelar la relación entre las variables ambientales y la producción de energía eléctrica en una central de ciclo combinado. Ambos modelos, tanto la implementación manual como la versión de `Scikit-learn`, demostraron un rendimiento sobresaliente en la predicción de la variable objetivo, con coeficientes de determinación (R2) superiores a 0.91. Esto indica que más del 91% de la variabilidad en la producción de energía puede ser explicada por las cuatro variables de entrada (AT, V, AP, RH). Este hallazgo no solo valida la hipótesis inicial, sino que también subraya la importancia de monitorear estas variables para predecir el rendimiento de la planta.

Si bien ambas implementaciones fueron exitosas, el modelo de `Scikit-learn` superó a la implementación manual con un menor Error Cuadrático Medio (MSE), 0.0717 frente a 0.0896. Esta diferencia, aunque pequeña, resalta la superioridad de los frameworks optimizados. Sus algoritmos de optimización interna están diseñados para una convergencia más rápida y precisa, lo que se traduce en un modelo final más robusto. La implementación manual, aunque didácticamente invaluable para comprender los fundamentos del aprendizaje automático, demuestra la complejidad de alcanzar la misma eficiencia sin las optimizaciones de un framework.

5.2 Posibles mejoras futuras

Para continuar mejorando la precisión y la robustez del modelo, se recomiendan las siguientes acciones:

- Exploración de Modelos Más Complejos: Si bien la regresión lineal funcionó bien, otros modelos como las Máquinas de Vectores de Soporte (SVM), Árboles de Decisión, o Redes Neuronales podrían capturar relaciones no lineales en los datos, lo que podría reducir aún más el error residual.
- Ingeniería de Características: Se podrían crear nuevas variables a partir de las existentes. Por ejemplo, la interacción entre la temperatura y la humedad podría tener un efecto significativo en el rendimiento de la planta que no es capturado por un modelo lineal simple.
- Uso de un Conjunto de Datos Más Amplio y Diverso: Incluir datos de la planta en diferentes modos de operación (por ejemplo, cargas parciales) o de un período de tiempo más prolongado podría aumentar la generalidad y robustez del modelo. Además, considerar otras variables operativas, como la carga de la turbina o el flujo de combustible, podría proporcionar una visión más completa y mejorar la precisión.
- Validación Cruzada por K-folds: El dataset fue proporcionado con particiones para 5x2 fold cross-validation, lo cual no fue completamente implementado. Adoptar este

método en futuras iteraciones permitiría obtener una estimación del rendimiento del modelo más estable y menos sesgada que una única división de datos.

Estas mejoras no solo optimizarán el modelo, sino que también contribuirían a una comprensión más profunda de la dinámica operativa de la central eléctrica de ciclo combinado.

CONLUSIÓN

El análisis realizado confirma que las variables ambientales son un factor determinante en la producción de energía de una central eléctrica de ciclo combinado. El modelo de regresión lineal, tanto en su implementación manual como en la optimizada con el framework `Scikit-learn`, demostró una notable precisión predictiva. Con un coeficiente de determinación (R2) superior a 0.91, ambos modelos lograron explicar más del 90% de la variabilidad en la producción de energía, lo que valida su utilidad para el monitoreo y la estimación del rendimiento de la planta.

Si bien el modelo manual fue crucial para comprender los principios del descenso de gradiente, la versión de `Scikit-learn` mostró un rendimiento ligeramente superior, con un MSE más bajo (0.0717). Esto resalta la importancia de utilizar herramientas y librerías especializadas que han sido optimizadas para la eficiencia y precisión. En resumen, este proyecto no solo ha creado un modelo predictivo robusto, sino que también ha demostrado la viabilidad de aplicar técnicas de *machine learning* para abordar desafíos de optimización en el sector energético, sentando las bases para futuras mejoras y análisis más avanzados.

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.