Tecnológico de Monterrey

Instituto Tecnológico y de Estudios Superiores de Monterrey

Campus Querétaro

Momento de Retroalimentación: Módulo 2

Evidencia 1. Portafolio de análisis: 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:

15 de septiembre de 2025

Abstract **- El presente reporte muestra un modelo de regresión lineal para predecir la producción de energía eléctrica neta (EP) en una central de ciclo combinado (CCPP) a plena carga. El objetivo es estimar la producción horaria de energía a partir de cuatro variables ambientales: temperatura (T), presión ambiente (AP), humedad relativa (RH) y vacío de escape (V). Para lograrlo, se implementó y entrenó un modelo de regresión lineal con un algoritmo de descenso de gradiente estocástico (SGD) desde cero, y sus resultados se validaron comparándolos con una implementación de la librería scikit-learn.**

El modelo se entrenó con un conjunto de datos de 9,568 puntos, que se pre procesaron mediante estandarización para optimizar la convergencia. Los resultados muestran una alta capacidad predictiva. La implementación inicial desde cero obtuvo un coeficiente de determinación (R^2) de 0.9159 y un error cuadrático medio (MSE) de 0.0896 en el conjunto de prueba. Tras un proceso de optimización, ajustando los hiper parámetros (epochs y learning rate), el modelo mejorado superó a la implementación de la librería con un R^2 de 0.9337 y un MSE de 0.0706. Estos resultados confirman la aplicabilidad del 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:
   - Entrenamiento:
     - $MSE = 0.0900$
     - $R^2 = 0.9093$
   - Validación:
     - $MSE = 0.0905$
     - $R^2 = 0.9089$
   - Prueba:
     - $MSE = 0.0896$
     - $R^2 = 0.9159$

b) Implementación de regresión lineal con uso de framework:
   - Entrenamiento:
     - $MSE = 0.0768$
     - $R^2 = 0.9232$
   - Validación:
     - $MSE = 0.0750$
     - $R^2 = 0.9232$
   - Prueba:
     - $MSE = 0.0717$
     - $R^2 = 0.9249$

4.3 Gráficas obtenidas de desempeño de los modelos

a) Modelo de regresión lineal desde cero:

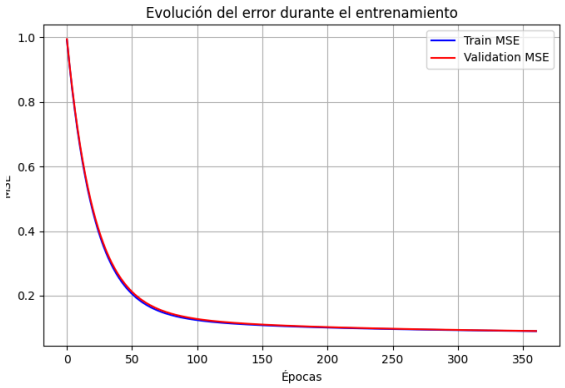

Figura 1. Evolución del MSE del conjunto de entrenamiento y validación

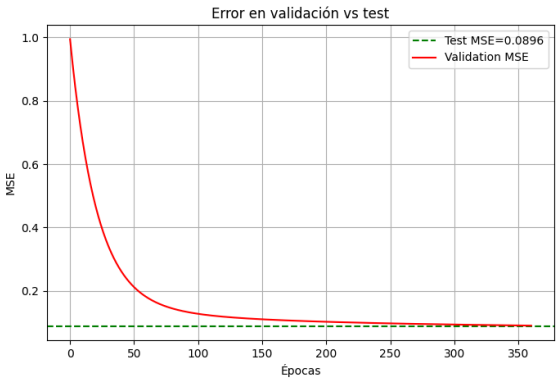

Figura 2. Comparativa MSE del conjunto de validación y prueba

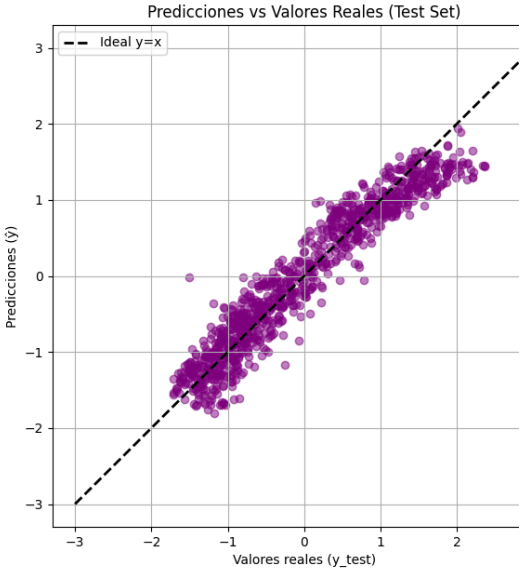

Figura 3. Predicciones hechas con parámetros resultantes del modelo

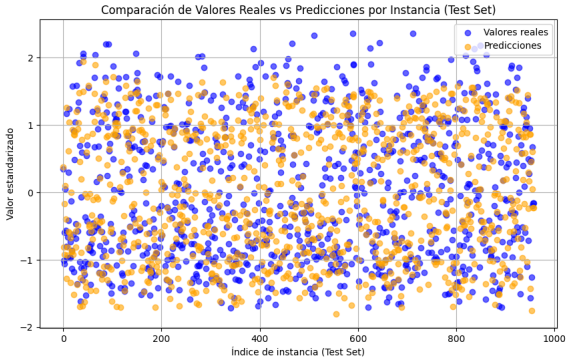

Figura 4. Comparativa de valores reales contra predicciones en torno a las instancias

b) Modelo de regresión lineal con uso de framework:

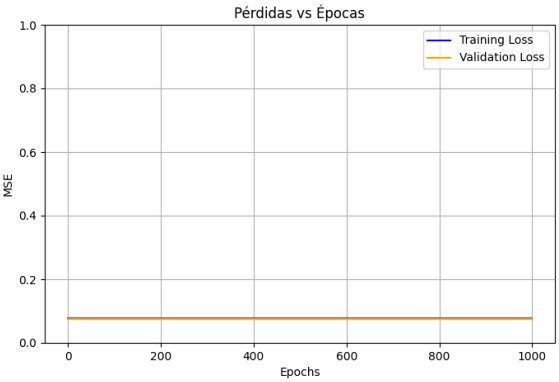

Figura 5. Comparativa de MSE entre conjunto de entrenamiento y validación

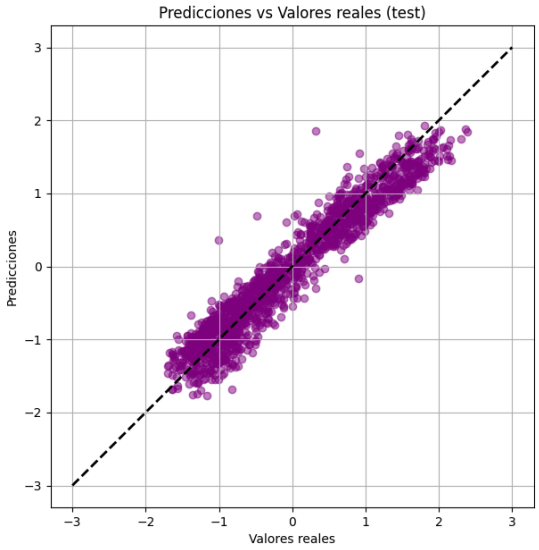

Figura 6. Predicciones hechas con los parámetros obtenidos del modelo

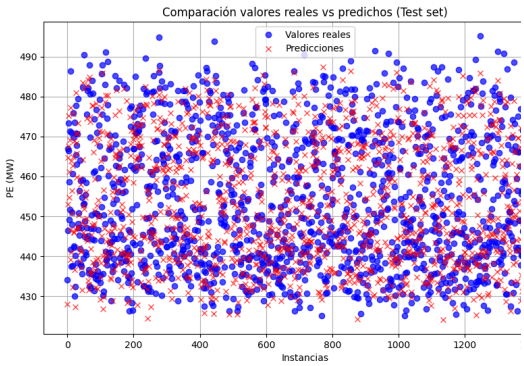

Figura 7. Comparativa de valores reales contra predicciones en torno a las instancias

4.4 Diagnósticos de la implementación del modelo desde cero

Con base en las métricas y gráficas mostradas con anterioridad, se puedo ver que con base en tanto el error cuadrático medio (MSE) y el coeficiente de determinación, la implementación del modelo que presentó un mejor desempeño, con diferencia mínima, fue la que empleó el uso de un framework, ya que logra explicar 0.9% más de la variabilidad de los datos para el conjunto de datos de prueba, en comparación con la implementación desde cero, por ende, es que se optó por elegir a esta última implementación para realizar un serie de mejoras que logren mejorar su desempeño en torno a una serie de diagnósticos a los que se le someterá.

Para poder realizar un diagnóstico integral de la implementación del modelo de regresión lineal desde cero, no es suficiente considerar las métricas del MSE y de coeficiente de correlación, sino que es necesario evaluar el grado de sesgo y varianza de la misma implementación para poder determinar el grado de ajuste del modelo.

En primer lugar, el nivel de sesgo del modelo refleja el error sistemático de las predicciones, en otras palabras, es la capacidad del modelo de poder capturar la relación real entre las variables. Para poder realizar su diagnóstico, se consideraron los valores de coeficientes de determinación de los conjuntos de entrenamiento y validación. Para poder cuantificar el sesgo del modelo, esto se puede lograr mediante el cálculo del promedio aritmético de ambos coeficientes de determinación, donde un valor de promedio aritmético mayor o igual de 0.8, indica un grado bajo de sesgo; un valor entre 0.6 y 0.8, indica un nivel medio de sesgo; y un valor menor a 0.6 indica un nivel alto de sesgo. Por ende, se puede visualizar el nivel de sesgo de la implementación desde cero en torno a la siguiente tabla:

TABLA 1. Nivel de sesgo de implementación desde cero

Conjunto	R^2	Promedio de coeficientes	Nivel de sesgo
Entrenamiento	0.0900	0.9091	Bajo
Validación	0.0905		

Por otro lado, el grado de varianza refleja la sensibilidad del modelo a los datos de entrenamiento y su capacidad de generalización. Para su diagnóstico se comparan los valores absolutos de las diferencias entre el del coeficiente de determinación y del error cuadrático medio. En términos de cuantificación del grado de varianza, es prudente considerar intervalos de valores entre las métricas a utilizar, obteniendo que una diferencia absoluta de coeficientes de determinación menor a 0.1 y una diferencia absoluta de errores cuadráticos medios menor a 20%, se puede considerar un grado de varianza bajo; si se tiene una diferencia absoluta de coeficientes de determinación entre a 0.1 y 0.2, y una diferencia absoluta de errores cuadráticos medios entre 20% y 50%, se considera un grado medio de varianza; mientras que si se tiene una diferencia absoluta de coeficientes de determinación mayor a 0.2 y una diferencia absoluta de errores cuadráticos medios mayor a 50%, se considera como un grado alto de varianza.

Las ecuaciones que describen la cuantificación de la varianza del modelo son las siguientes:

$$Gap_{R^2} = \left| R^2_{train} - R^2_{val} \right|$$

$$Gap_{MSE} = \left| \frac{MSE_{train} - MSE_{val}}{MSE_{train}} \right|$$

En términos de la presente implementación desde cero del modelo de regresión lineal, su grado de varianza se cuantifica como se muestra en la siguiente tabla:

TABLA 2. Nivel de varianza de implementación desde cero

Métrica	Entrenamiento	Validación	Diferencia absoluta	Nivel de varianza
MSE	0.0900	0.0905	0.0005	Bajo
R^2	0.9093	0.9089	0.0004	

Con base en los diagnósticos anteriores, se puede observar que tanto el nivel de varianza como el de sesgo del modelo son bajos. Por lo tanto, esto indica que el ajuste del modelo se encuentra dentro del escenario de fitting adecuado, lo cual se refleja también en la gráfica que muestra la evolución del MSE en los conjuntos de entrenamiento y prueba. Estos resultados permiten concluir que se trata de un modelo de regresión lineal robusto.

Dado el panorama del ajuste del modelo, entonces es prudente considerar que las mejoras a realizar en esta implementación, giran en torno al ajuste de los hiper parámetros del modelo.

4.5 Mejoras aplicadas al modelo

Dado el panorama del ajuste del modelo, entonces es prudente considerar que las mejoras a realizar en esta implementación, giran en torno al ajuste de los hiper parámetros del modelo.

Por ende, es preciso mencionar que los hiper parámetros del modelo con los que se obtuvieron las primeras métricas de este, fue con base en la implementación de 360 *epochs* y un *learning rate* de 0.01. Con base en estos valores, de manera heurística, se determinó que los hiper parámetros que lograron tener mejorar el desempeño de la implementación de la regresión lineal desde cero, fue con un número de 1000 *epochs* y un *learning rate* de 0.03

4.6 Métricas de la implementación de un modelo regresión lineal desde cero mejorada

- Entrenamiento:
  - $MSE = 0.0721$
  - $R^2 = 0.9274$
- Validación:
  - $MSE = 0.0697$
  - $R^2 = 0.9299$
- Prueba:
  - $MSE = 0.0706$
  - $R^2 = 0.9337$

4.7 Gráficas de la implementación de un modelo regresión lineal desde cero mejorada

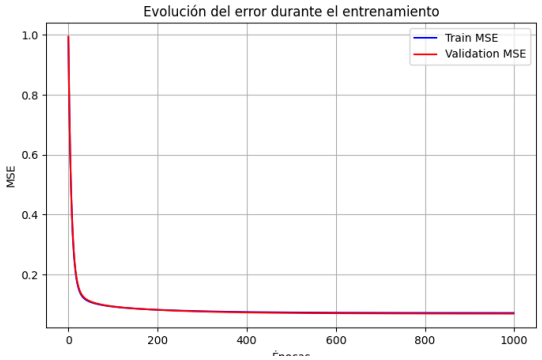

Figura 8. Evolución del MSE del conjunto de entrenamiento y validación

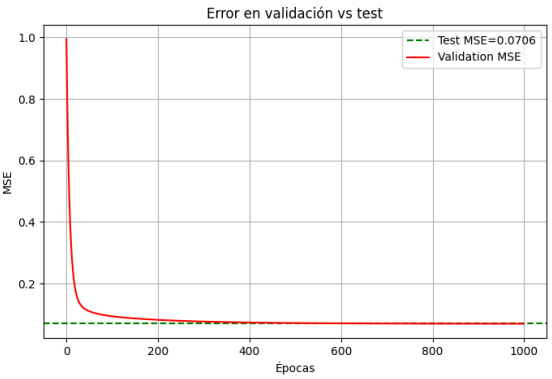

Figura 9. Comparativa MSE del conjunto de validación y prueba

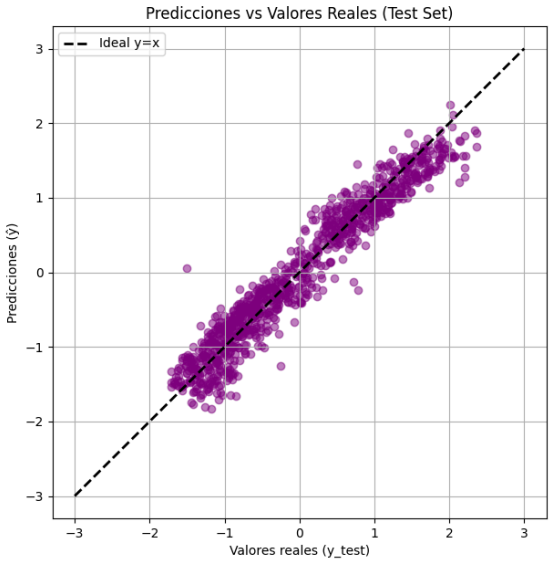

Figura 10. Predicciones hechas con los parámetros obtenidos del modelo

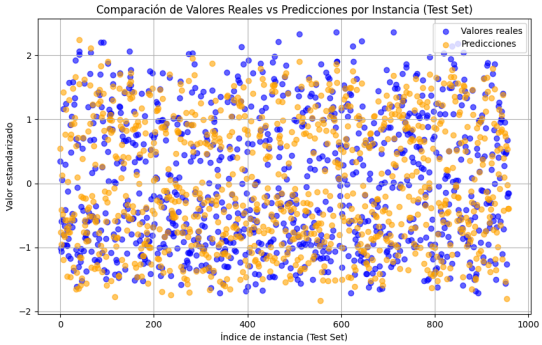

Figura 10. Comparativa de valores reales contra predicciones en torno a las instancias

5. CONCLUSIONES

5.1 Desempeño general de modelos

El análisis de desempeño de los modelos de regresión lineal indica que ambas implementaciones presentan un alto nivel de precisión predictiva. La implementación desde cero logró un MSE de 0.0896 y un R^2 de 0.9159 en el conjunto de prueba, mientras que la implementación utilizando el framework de sklearn alcanzó un MSE de 0.0717 y un R^2 de 0.9249 en el mismo conjunto.

Estos resultados demuestran que ambos modelos son capaces de capturar de manera efectiva la relación entre las variables ambientales (temperatura, vacío de escape, presión ambiente y humedad relativa) y la producción de energía eléctrica neta de la planta. Las diferencias observadas, aunque pequeñas, muestran que la implementación con framework presenta un desempeño ligeramente superior, logrando explicar un porcentaje mayor de la varianza de los datos y obtener un error más bajo en el conjunto de prueba.

Adicionalmente, los diagnósticos de sesgo y varianza de la implementación desde cero revelan niveles bajos de ambos, lo que indica que el modelo tiene un ajuste adecuado y no presenta problemas de *underfitting* u *overfitting* . Las curvas de evolución del MSE en entrenamiento y validación también reflejan un comportamiento estable y convergente a lo largo del proceso de entrenamiento.

5.2 Análisis de mejoras realizadas

Para optimizar el desempeño del modelo implementado desde cero, se realizaron ajustes a los hiper parámetros del algoritmo de descenso de gradiente. Inicialmente, el modelo se entrenó con 360 epochs y un *learning rate* de 0.01. Tras un análisis heurístico y pruebas adicionales, se determinó que un incremento a 1000 epochs y un learning rate de 0.03 permitió mejorar la convergencia del modelo y reducir el error de predicción.

Con estos ajustes, la implementación mejorada alcanzó un MSE de 0.0706 y un R^2 de 0.9337 en el conjunto de prueba, superando incluso los resultados obtenidos por la implementación con framework. Las gráficas de evolución del MSE muestran una disminución constante del error durante el entrenamiento y validación, y las comparaciones entre predicciones y valores reales evidencian una aproximación más cercana a los datos observados.

Este análisis demuestra que pequeñas modificaciones en los hiper parámetros pueden tener un impacto significativo en el desempeño del modelo, mejorando su precisión y capacidad de generalización.

CONLUSIÓN

El análisis confirma que los modelos de regresión lineal son herramientas altamente efectivas para predecir la producción de energía eléctrica neta en una CCPP a partir de variables ambientales clave. La implementación desde cero y la versión con scikit-learn demostraron ser robustas, presentando bajos niveles de sesgo y varianza, lo que asegura su fiabilidad.

Este trabajo subraya la importancia de la optimización de hiper parámetros. Si bien la implementación inicial desde cero mostró un desempeño sólido, los ajustes realizados permitieron superar los resultados de la implementación con librería, logrando una mayor precisión y capacidad predictiva. El modelo mejorado alcanzó un MSE de 0.0706 y un R^2 de 0.9337 en el conjunto de prueba, demostrando que una configuración adecuada puede maximizar el rendimiento. En síntesis, este trabajo valida la efectividad del aprendizaje automático en el análisis y optimización del rendimiento energético, proporcionando un modelo confiable para la gestión de plantas de ciclo combinado.

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.