Modelo de Regresión para la Predicción del Consumo Energético en la Industria Siderúrgica

A. Cristobal
Pontificia Universidad Católica del Perú
Lima, Perú
a20203634@pucp.edu.pe

J. Moreno
Pontificia Universidad Católica del Perú
Lima, Perú
josue.moreno@pucp.edu.pe

E. Bravo
Pontificia Universidad Católica del Perú
Lima, Perú
elvis.bravo@pucp.edu.pe

P. Caceres
Pontificia Universidad Católica del Perú
Lima, Perú
20186862@pucp.edu.pe

***Resumen*—Este estudio aborda la predicción del consumo energético en una planta siderúrgica inteligente de pequeña escala ubicada en Corea del Sur, utilizando técnicas de regresión supervisada. El conjunto de datos utilizado incluye variables operativas registradas en tiempo real por sensores industriales, como consumo energético, factores de potencia y condiciones de operación. Se entrenaron cuatro modelos de regresión: `Random Forest`, `XGBoost`, `Ridge Regression` y `Red Neuronal (MLP)`, para estimar el consumo energético total a partir de las variables explicativas. La evaluación del desempeño se realizó mediante métricas como el error cuadrático medio (MSE) y el error absoluto medio (MAE), tanto en los conjuntos de entrenamiento y prueba como en la validación cruzada. Los resultados destacaron la superioridad del modelo `Random Forest` en términos de precisión predictiva. Este análisis demuestra el potencial del aprendizaje automático para optimizar la eficiencia energética en entornos industriales, contribuyendo a la sostenibilidad en procesos de manufactura avanzada. Todo los archivos se encuentran en el siguiente gidhub: https://github.com/Elvis1219BS/AM-Proyect.git**

***Index Terms*—regresión, consumo energético, sostenibilidad, Random Forest, aprendizaje automático.**

I. Introducción

Este trabajo se enmarca en el curso *Aprendizaje de Máquina por Alumnos* de la Pontificia Universidad Católica del Perú, y tiene como objetivo aplicar técnicas de regresión supervisada para predecir el consumo energético en entornos industriales inteligentes. En particular, se trabaja con un conjunto de datos reales provenientes de una planta siderúrgica a pequeña escala ubicada en Corea del Sur, donde sensores industriales recolectan información operativa en tiempo real.

El dataset incluye variables relevantes como consumo energético, factores de potencia, condiciones atmosféricas internas, y parámetros eléctricos, todos medidos en intervalos regulares. Estos datos ofrecen una oportunidad valiosa para evaluar la capacidad de algoritmos de aprendizaje automático en la predicción del comportamiento energético en un contexto industrial.

La predicción precisa del consumo energético es crucial para optimizar la eficiencia operativa, reducir costos y mitigar el impacto ambiental. A través del uso de modelos de regresión —incluyendo enfoques lineales y no lineales— se busca no solo estimar con precisión el consumo, sino también identificar los factores que más influyen en su variabilidad. Esta información puede ser utilizada para desarrollar estrategias de gestión energética más sostenibles, alineadas con los objetivos de la industria 4.0 y la transición hacia operaciones inteligentes basadas en datos.

II. Metodología

El desarrollo del presente estudio se estructuró en una serie de etapas metodológicas propias del enfoque de aprendizaje supervisado, orientadas a la predicción del consumo energético a partir de datos operativos capturados en tiempo real. A continuación, se describen las fases clave del proceso:

A. Carga y preprocesamiento de datos

La base de datos fue importada en formato `.csv` y sometida a un proceso de depuración inicial. Este incluyó la verificación de registros duplicados, detección de valores faltantes y validación de tipos de datos. Las variables categóricas fueron transformadas mediante codificación `one-hot`, mientras que las variables numéricas fueron normalizadas para asegurar una escala común en el entrenamiento de los modelos. La variable objetivo seleccionada fue el consumo energético (`Usage_kWh`), y el resto de columnas se utilizaron como variables explicativas.

B. División del conjunto de datos

Para garantizar una evaluación objetiva del rendimiento de los modelos, se realizó una partición del conjunto de datos en dos subconjuntos: entrenamiento (80 %) y prueba (20 %), empleando la función `train_test_split` de la biblioteca `scikit-learn`. Esta división permite entrenar los modelos sobre una muestra representativa y validar su capacidad de generalización en datos no vistos.

C. Selección y entrenamiento de modelos

Se optó por utilizar algoritmos de regresión de alto rendimiento, incluyendo *Random Forest Regressor*, *XGBoost*

Regressor, *Ridge Regression* y *Red Neuronal (MLP)*. Estos modelos fueron seleccionados debido a su capacidad para manejar relaciones no lineales complejas y su robustez frente al sobreajuste. Los cuatro modelos fueron entrenados sobre el conjunto de entrenamiento utilizando los hiperparámetros por defecto, con la posibilidad de realizar ajustes adicionales mediante técnicas de optimización de hiperparámetros como *Grid Search* o *Random Search*.

D. Evaluación del desempeño

El desempeño de los modelos fue evaluado mediante dos métricas fundamentales en problemas de regresión: el error cuadrático medio (MSE), que penaliza grandes desviaciones entre los valores predichos y reales, y el error absoluto medio (MAE), que mide la magnitud promedio de los errores sin considerar su dirección. Además, se implementó validación cruzada estratificada (`KFold`, $k = 5$) para estimar la estabilidad del rendimiento y evitar sesgos derivados de una única partición de los datos.

E. Interpretación de resultados

Finalmente, se evaluó la importancia relativa de las variables predictoras utilizando los mecanismos internos de los modelos basados en árboles, lo que permitió identificar los factores más influyentes en la predicción del consumo energético. Esta fase proporciona información clave para diseñar estrategias de eficiencia energética, optimizar procesos y priorizar las variables críticas en futuras iteraciones de modelado o implementación de mejoras.

III. Análisis Exploratorio de Datos (EDA)

A. Exploración inicial

El análisis exploratorio comenzó con una auditoría de calidad de los datos, verificando la presencia de valores nulos, registros duplicados y formatos inconsistentes. La inspección reveló que el conjunto de datos se encontraba limpio, sin registros faltantes ni redundancias evidentes, lo cual facilitó un flujo de trabajo eficiente hacia las siguientes fases del análisis.

Posteriormente, se realizó una caracterización univariada de las variables disponibles, agrupándolas por tipo: numéricas continuas, categóricas nominales y temporales. La **Tabla I** resume esta clasificación preliminar de atributos.

Cuadro I
Clasificación simplificada de las variables por tipo de dato

Variable	Tipo de Dato
`Usage_kWh`	Numérica
`Lagging_Current_Reactive.Power_kVarh`	Numérica
`Leading_Current_Reactive_Power_kVarh`	Numérica
`CO2(tCO2)`	Numérica
`Lagging_Current_Power_Factor`	Numérica
`Leading_Current_Power_Factor`	Numérica
`NSM`	Numérica
`WeekStatus`	Categórica
`Day_of_week`	Categórica

Para las variables numéricas, se analizaron sus distribuciones mediante histogramas con curvas de densidad y se exploraron valores atípicos con boxplots. Estos gráficos permitieron identificar sesgos, concentraciones y posibles transformaciones requeridas. La **Figura 1** muestra un ejemplo de esta visualización.

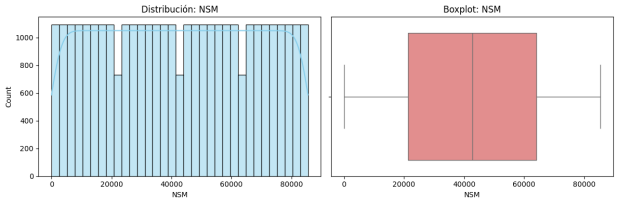

Figura 1. Distribuciones y valores atípicos en variables numéricas representadas mediante histogramas y boxplots.

En cuanto a las variables categóricas, se analizaron mediante gráficos de barras para evaluar la distribución de frecuencias y detectar posibles desequilibrios o categorías dominantes. La **Figura 2** presenta un ejemplo ilustrativo de este análisis.

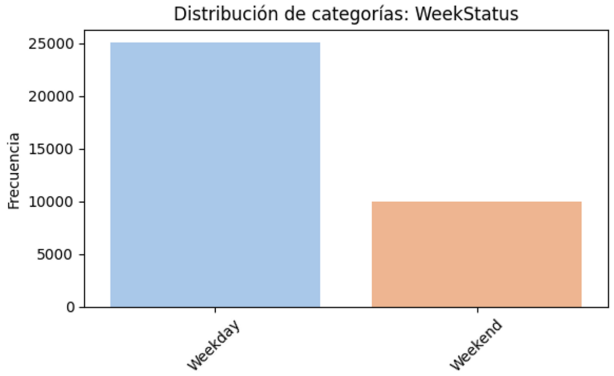

Figura 2. Frecuencia de categorías para variables cualitativas.

B. Análisis temporal

Dado que el conjunto de datos incluye un atributo de fecha y hora (`date`), se realizó un análisis temporal para evaluar patrones de consumo energético a lo largo del día, la semana y el mes.

Se derivaron variables como la hora del día y el día del mes, y se calcularon promedios de consumo para identificar ciclos operativos. La **Figura 3** revela picos de actividad energética durante horas laborales, mientras que la **Figura 4** evidencia fluctuaciones a lo largo del mes que podrían estar vinculadas con turnos de producción o mantenimiento.

C. Análisis de correlación

Se examinó la correlación entre variables numéricas mediante el coeficiente de Pearson, construyendo una matriz de correlación visualizada como un mapa de calor. Esta herramienta permitió detectar relaciones lineales fuertes (colinealidad) que podrían afectar la estabilidad de los modelos de regresión.

La **Figura 5** muestra la matriz resultante. En base a este análisis, se excluyeron variables altamente correlacionadas como `Lagging_Current_Reactive.Power_kVarh` y

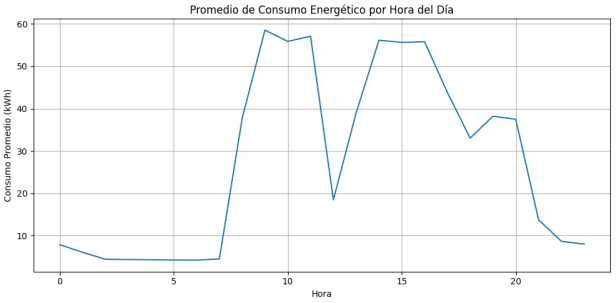

Figura 3. Consumo promedio de energía por hora del día. Se observan picos durante el horario productivo.

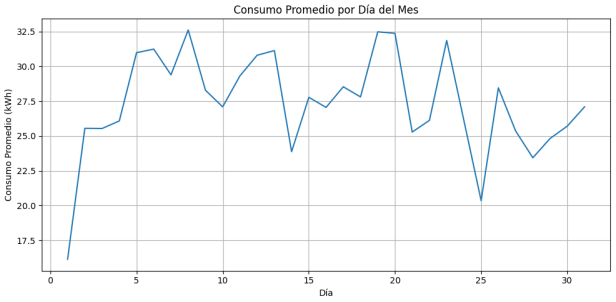

Figura 4. Variación promedio del consumo energético según el día del mes.

`Leading_Current_Power_Factor` para reducir la redundancia y mejorar la eficiencia del modelo.

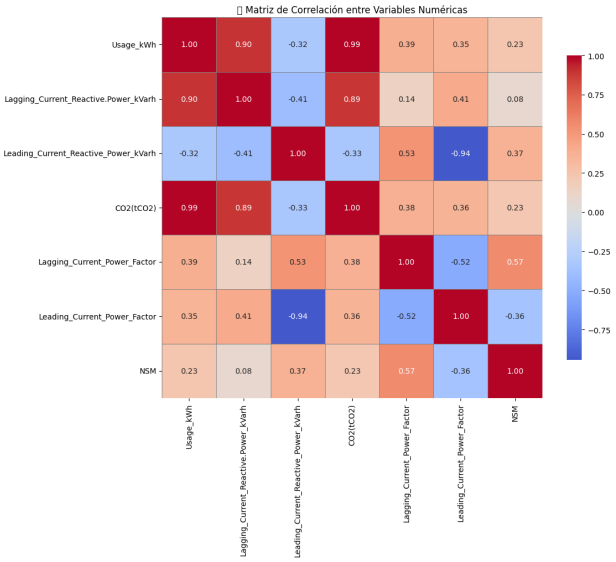

Figura 5. Matriz de correlación entre variables numéricas. Se destacan relaciones superiores a 0.85.

D. Codificación de variables categóricas

Para incorporar adecuadamente las variables categóricas al proceso de modelado, se aplicó la técnica de *one-hot encoding* mediante la función `get_dummies` de `pandas`. Este procedimiento transforma cada categoría en una columna binaria, eliminando la ambigüedad asociada al tratamiento ordinal o numérico de valores nominales.

Este paso resultó fundamental para que los algoritmos de regresión pudieran procesar correctamente las variables `WeekStatus` y `Day_of_week`, generando representaciones explícitas como `Day_of_week_Monday`, `WeekStatus_Weekend`, entre otras.

El conjunto de datos resultante integró tanto variables numéricas estandarizadas como categóricas codificadas, optimizando su compatibilidad con los modelos de aprendizaje automático aplicados en fases posteriores.

IV. Modelado Predictivo

Para la construcción del modelo de predicción del consumo energético, se adoptó un enfoque de aprendizaje supervisado basado en algoritmos de regresión, entrenados sobre datos preprocesados provenientes de sensores industriales.

A. División del conjunto de datos

El conjunto fue dividido en subconjuntos de entrenamiento (80 %) y prueba (20 %), utilizando la función `train_test_split` de la biblioteca `scikit-learn`. Esta estrategia permite evaluar el rendimiento del modelo en datos no vistos, garantizando una estimación más confiable de su capacidad de generalización.

B. Selección y entrenamiento de modelos

Se entrenaron y compararon cuatro modelos de regresión de alto desempeño:

- **Random Forest Regressor:** Un modelo de ensamblado que combina múltiples árboles de decisión construidos sobre subconjuntos aleatorios de datos y variables. Su robustez frente al sobreajuste y su capacidad para capturar relaciones no lineales lo hacen idóneo para entornos industriales ruidosos.
- **XGBoost Regressor:** Un algoritmo de *gradient boosting* que construye árboles secuenciales minimizando una función de pérdida regularizada. Destaca por su precisión, manejo de datos faltantes y eficiencia computacional.
- **Ridge Regression:** Un modelo de regresión lineal que incorpora regularización L2, penalizando los coeficientes de las variables para evitar el sobreajuste. Este modelo es especialmente útil cuando existen muchas variables correlacionadas y ayuda a mejorar la estabilidad del modelo.
- **Red Neuronal (MLP):** Un modelo basado en redes neuronales de tipo perceptrón multicapa (MLP), que emplea capas ocultas para aprender representaciones no lineales complejas de los datos. Su capacidad para modelar relaciones complejas lo convierte en una opción potente para tareas de regresión, aunque es más susceptible al sobreajuste sin un adecuado ajuste de hiperparámetros.

Los cuatro modelos fueron entrenados utilizando un conjunto de variables numéricas estandarizadas y variables categóricas codificadas mediante *one-hot encoding*. Se utilizaron

los hiperparámetros por defecto como punto de partida, reservando la optimización avanzada para futuras iteraciones.

C. Validación cruzada

Para asegurar la robustez de los resultados y reducir la varianza asociada a la partición del conjunto de prueba, se aplicó validación cruzada de k-folds con $k = 5$. En cada iteración, se calcularon los promedios de desempeño mediante la raíz del error cuadrático medio (RMSE), proporcionando una estimación más confiable de la capacidad de generalización de los modelos.

En la **Figura 6** se presentan los resultados de la validación cruzada para los modelos evaluados, mostrando los valores promedio de RMSE en las particiones de entrenamiento y prueba.

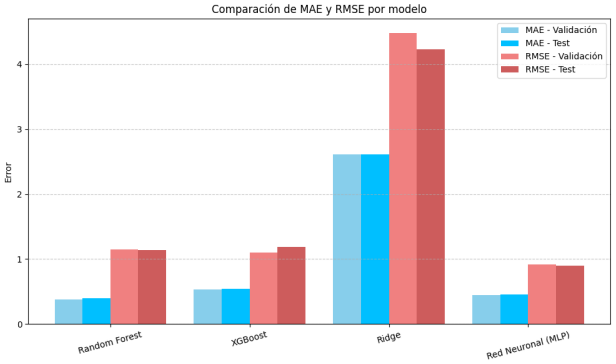

Figura 6. Resultados de la validación cruzada ($k = 5$).

D. Resultados

Los resultados promedios obtenidos de la validación cruzada se presentan a continuación:

- **Random Forest Regressor:** RMSE promedio = 1.2411 ± 0.0883
- **XGBoost Regressor:**RMSE promedio = 1.3236 ± 0.0582
- **Ridge Regressor:** RMSE promedio = 4.8101 ± 0.1703
- **MLP Regressor:** RMSE promedio = 0.9286 ± 0.0885

A pesar de que la red neuronal presentó métricas ligeramente superiores en términos de MAE y RMSE, se optó por el modelo Random Forest debido a su mayor interpretabilidad y facilidad de ajuste. El modelo XGBoost se descartó al momento de realizar la validación cruzada, por tener un RMSE de 1.3236 mayor que el de Random Forest 1.24. El modelo de Random Forest obtuvo un MAE menor tanto en validación como en prueba, lo que indica un menor error promedio absoluto por predicción. Además, un valor bajo de RMSE de la validación cruzada, por lo que su rendimiento fue altamente consistente entre los conjuntos de validación y test y demuestra una buena capacidad de generalización. Este modelo también ofrece ventajas prácticas como la capacidad de evaluar la importancia de las variables, lo que resulta útil para obtener información explicativa adicional sobre el comportamiento del sistema modelado. Por estas razones, Random Forest se considera la opción más confiable y manejable en este caso.

Cuadro II
THE MODELS PERFORMANCE.

Models	Opt. Parameters	Training		Testing	
		RMSE	MAE	RMSE	MAE
XGBoost	max_depth = 6	1.103	1.191	0.541	0.00
RF	max_depth = 20	1.151	0.382	1.134	0.400
Ridge	alpha = 0.01	4.478	2.611	4.222	2.612
MLP	HiddenL = (1,64)	0.914	0.448	0.902	0.448

E. Evaluación e interpretación

Se complementó la evaluación con un análisis interpretativo visual. En primer lugar, se exploró la importancia de las variables utilizando los mecanismos internos deL modelo de mejor performance (Ramdon Forest). La **Figura 7** presentan las puntuaciones de importancia relativa asignadas. Variables como la emisión de CO_2 (`CO2`) y el factor de potencia (`Lagging_Current_Power_Factor`) se destacan como las más influyentes en la predicción del consumo energético.

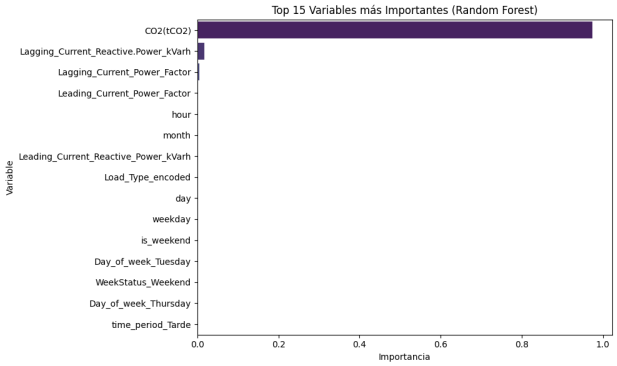

Figura 7. Importancia relativa de las variables predictoras según Random Forest.

Además, se analizaron los gráficos de dispersión entre los valores reales y los valores predichos. Este análisis facilita la evaluación de la precisión individual de las predicciones. Las **Figuras 8** muestra el modelo de mejor performance genera predicciones cercanas a la línea ideal ($y = \hat{y}$), indicando una alta concordancia.

Finalmente, se representaron los residuos (errores de predicción) para el modelo de RF. Un patrón aleatorio de los residuos, como se observa en las **Figuras 9**, sugiere una buena especificación del modelo y ausencia de sesgos sistemáticos.

V. CONCLUSIONES

Este estudio demuestra la aplicabilidad y efectividad de los algoritmos de aprendizaje supervisado para predecir el consumo energético en una planta siderúrgica inteligente, utilizando datos recolectados en tiempo real mediante sensores industriales.

El análisis exploratorio inicial permitió identificar patrones operativos significativos, tales como variaciones cíclicas horarias y semanales, así como relaciones de alta colinealidad entre algunas variables numéricas. Estas correlaciones fueron gestionadas mediante la eliminación selectiva de variables

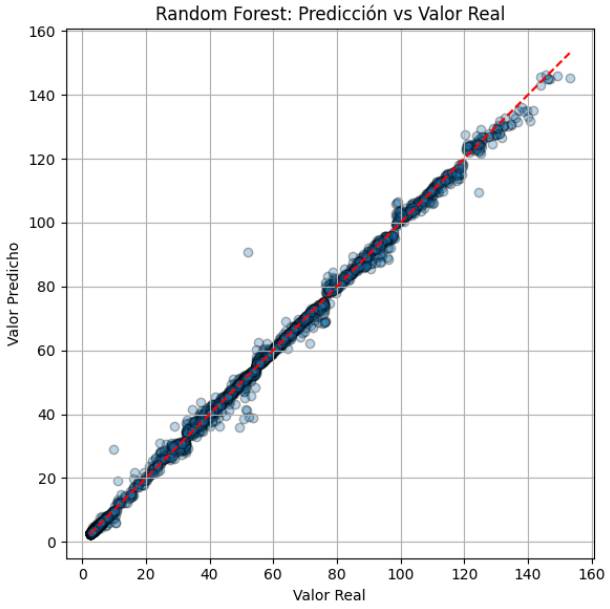

Figura 8. Valores reales vs. predichos por el modelo Random Forest.

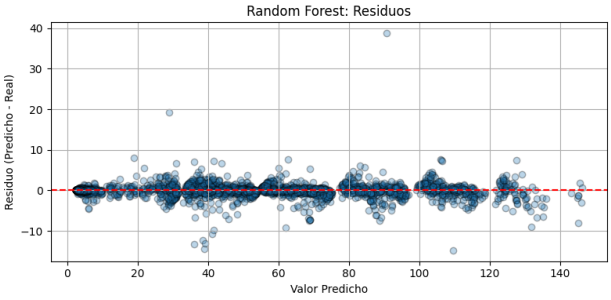

Figura 9. Distribución de residuos del modelo Random Forest.

redundantes, lo que mejoró la eficiencia del modelo y mitigó posibles efectos de multicolinealidad.

Se entrenaron cuatro modelos de regresión de alto rendimiento: `Random Forest`, `XGBoost`, `Ridge Regression` y `Red Neuronal (MLP)`. Los modelos `Random Forest (RF)` y `Red Neuronal (MLP)` presentaron métricas sobresalientes, siendo `RF` el de mayor rendimiento, con una raíz cuadrática media (RMSE) inferior al de `MLP` ($RMSE = 1{,}13$) y un error absoluto medio (MAE) de $MAE = 0{,}40$. Estos resultados validan la capacidad predictiva del modelo basado en árboles de decisión y su adecuada adaptación al contexto industrial.

El análisis de importancia de características permitió identificar las variables más determinantes en la estimación del consumo energético. Entre ellas destacan la emisión de CO_2 (`CO2`) y el factor de potencia (`Lagging_Current_Power_Factor`). Esta información proporciona una base analítica clave para la toma de decisiones operativas, permitiendo diseñar estrategias de eficiencia energética orientadas a la optimización del uso de recursos.

Como líneas de trabajo futuro, se plantean las siguientes extensiones:

- Aplicar técnicas de ajuste de hiperparámetros, como la *Bayesian Optimization*, para mejorar aún más el desempeño predictivo.
- Ampliar el conjunto de variables predictoras, incluyendo factores exógenos como condiciones climáticas, precios de energía o demandas externas.

En conclusión, los resultados obtenidos respaldan el uso del aprendizaje automático como herramienta para la optimización energética en entornos industriales inteligentes, contribuyendo tanto a la sostenibilidad como a la mejora de la gestión operativa basada en datos.

Referencias

[1] UCI Machine Learning Repository, "Steel Industry Energy Consumption Data Set," [Online]. Available: https://archive.ics.uci.edu/dataset/851/steel+industry+energy+consumption. [Accessed: 18-Jun-2025].

[2] H. K. Ahn, S. C. Lee, J. H. Kim and C. S. Lee, "Energy consumption forecasting with machine learning in steel industry," *Building Research & Information*, vol. 49, no. 7, pp. 744–759, 2021. [Online]. Available: https://doi.org/10.1080/09613218.2020.1809983