Análisis de modelos supervisados para la predicción de la Generación de una planta de Energía Solar

Claudia Lissette Gutiérrez Díaz[1]

Palabras Clave
Selección de modelos supervisados, Energía Solar, Métricas de desempeño, Diseño de experimentos
Fecha: 29 de noviembre de 2023

[1] *Lic. Ciencias Computacionales, Facultad de Ciencias Físico - Matemáticas, UANL*

Índice

1. Introducción

La predicción de energía solar desempeña un papel fundamental en la planificación y optimización de sistemas de generación de energía limpia y sostenible. El aprendizaje supervisado, una rama del aprendizaje automático, se ha convertido en una herramienta esencial en este contexto. En este estudio, exploramos diversos modelos de regresión, incluyendo Gradient Boosting, Light Gradient Boosting Machine, Bagging, Random Forest, XGBoost, AdaBoost, con el objetivo de seleccionar el más adecuado para la predicción de la generación de energía solar.

Para evaluar el rendimiento de estos modelos, empleamos métricas de evaluación de errores como el Error Cuadrático Medio (MSE), la Raíz del Error Cuadrático Medio (RMSE), el Error Absoluto Medio (MAE), el Coeficiente de Determinación (R^2) y el Error Porcentual Absoluto Medio (MAPE). Estas métricas nos permiten comparar y seleccionar el modelo que mejor se ajuste a nuestros datos y necesidades específicas.

Por último, con uno o más modelos seleccionados, haremos el diseño de experimentos de cada modelo seleccionado, para encontrar los parámetros más adecuados para obtener la menor diferencia entre los dados reales y los pronosticados.

2. Descripción del conjunto de datos de la Planta de Energía Solar

El conjunto de datos utilizado en este estudio se recopiló de una planta de energía solar ubicada en Villa de Arista, en el estado de San Luis Potosí, México. El objetivo de este conjunto de datos es analizar y predecir la generación de energía en kilovatios-hora (kWh) a partir de diversas condiciones ambientales y considerando que el tope de energía es de 30,000 kWh debido a que esta es la capacidad de los paneles solares.

A continuación, se describen las variables seleccionadas, así como en la figura 1 se muestran sus respectivos histogramas y en la tabla 1, la estadística descriptiva de los mismos.

- **Generación de Energía (kWh):** Esta es la variable dependiente en nuestro conjunto de datos y representa la cantidad de energía generada por la planta solar en kilovatios-hora en intervalos de una hora.
- **Condiciones Ambientales:** Las siguientes variables independientes representan las condiciones ambientales que podrían influir en la generación de energía solar:

1. **Anio:** Representa el año de los datos.
2. **Mes:** Indica un código único del mes de los datos.
3. **NumMes:** Indica el número del mes en el año de los datos. Va del 1 al 12.
4. **NumSemana:** Indica el número de semana en el año de los datos. Va del 1 al 52.
5. **NumDiaAnio:** Indica el número del día en el año de los datos. Va del 1 al 366.
6. **Dia:** Indica un código único del día de los datos.
7. **Hora:** Indica un código único de la hora de los datos.
8. **NumHora:** Indica el número de hora en el día de los datos. Va del 0 al 24.
9. **MesDia:** Es una combinación concatenada del número del mes y del número del día en formato de 4 dígitos.
10. **Temperatura:** Es la Temperatura ambiental en grados Celsius.
11. **ProbabilidadLluvia:** Porcentaje de lluvia esperado para esa hora.
12. **HumedadRelativa:** Porcentaje de la humedad relativa presentada en el aire.
13. **VelocidadViento:** Velocidad del viento en kilómetros por hora.
14. **DireccionViento:** Dirección del viento en grados.
15. **CoberturaNubes:** Porcentaje de nubes, niebla o bruma que cubren el cielo.
16. **IndiceUV:** Índice de radiación ultravioleta.
17. **CodCondCielo:** Una variable que describe las condiciones del cielo, que cuyas descripciones pueden ser despejado, poco nuboso, nublado, medio nublado, cielo nublado o cielo cubierto.
18. **CodDirViento:** La dirección del viento categorizada en términos de puntos cardinales (por ejemplo, norte, sur, este, oeste).
19. **VelocidadRafaga:** Velocidad de ráfaga del viento en kilómetros por hora.
20. **DPT:** Punto de rocío en grados Celsius.
21. **Generacion_prev_day:** Generación de la hora previa en kilowatts por hora.

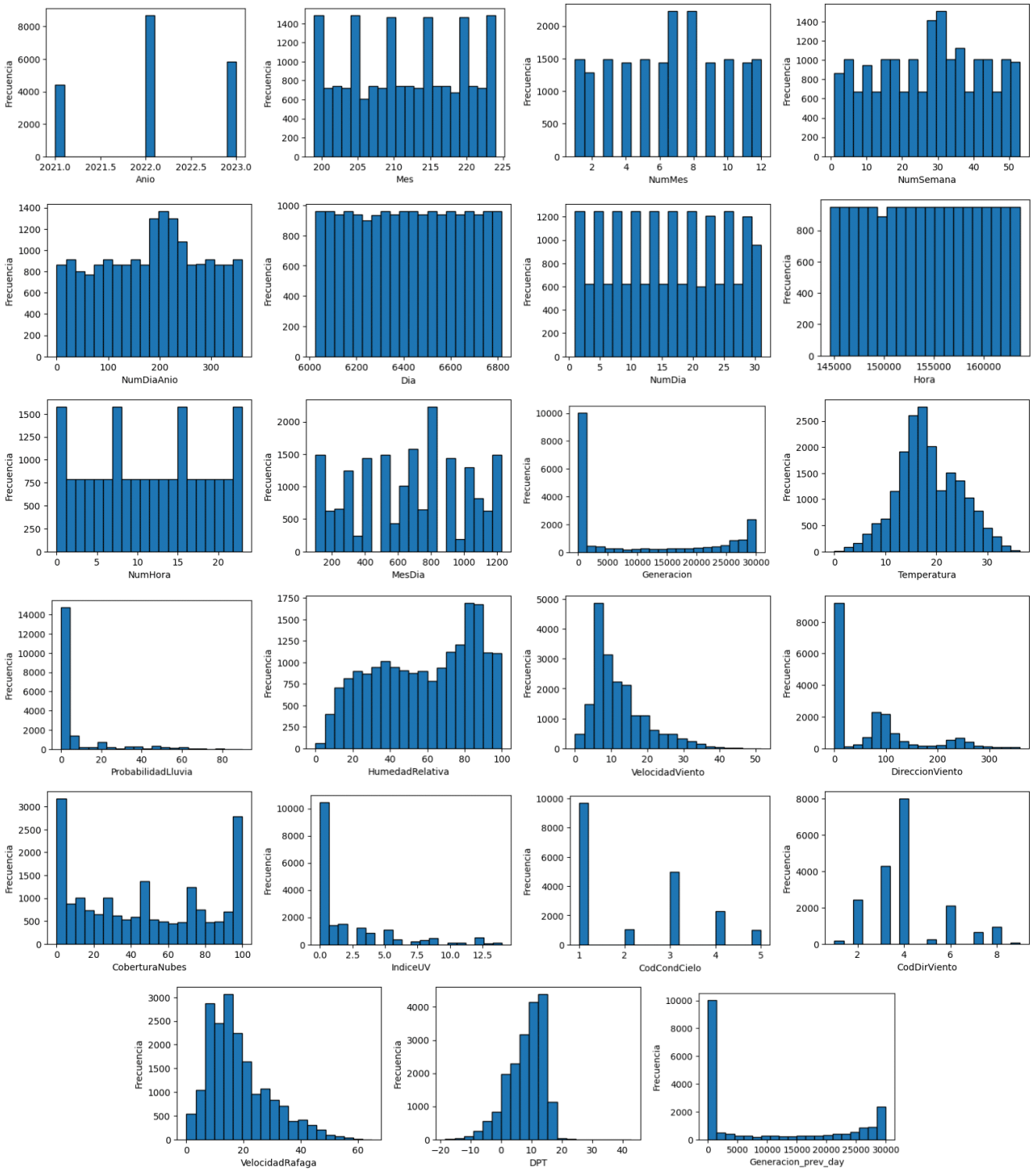

Figura 1. Histograma de las variables regresoras y dependientes.

- **Preparación del Conjunto de Datos:** Se realizaron procesos de limpieza de datos y se eliminaron valores vacíos.

Este conjunto de datos preparado es el que utilizamos para realizar la experimentación para encontrar el mejor modelo y parámetros para la predicción de la generación futura.

Cuadro 1. Muestra de las condiciones ambientales de Villa de Arista ya normalizadas

	count	mean	std	min	25%	50%	75%	max
Anio	18945.00	2022.07	0.73	2021.00	2022.00	2022.00	2023.00	2023.00
Mes	18945.00	211.51	7.52	199.00	205.00	212.00	218.00	224.00
NumMes	18945.00	6.62	3.32	1.00	4.00	7.00	9.00	12.00
NumSemana	18945.00	27.53	14.64	1.00	15.00	29.00	39.00	53.00
NumDiaAnio	18945.00	183.26	100.08	0.00	100.00	191.00	263.00	360.00
Dia	18945.00	6422.06	228.82	6026.00	6223.00	6423.00	6620.00	6817.00
NumDia	18945.00	15.72	8.82	1.00	8.00	16.00	23.00	31.00
Hora	18945.00	154117.24	5491.94	144601.00	149337.00	154136.00	158872.00	163608.00
NumHora	18945.00	11.50	6.92	0.00	6.00	12.00	18.00	23.00
MesDia	18945.00	677.53	332.10	101.00	411.00	712.00	924.00	1231.00
Generacion	18945.00	9748.95	12091.66	0.00	0.00	371.39	23562.55	29966.14
Temperatura	18945.00	18.43	6.06	0.00	14.40	18.00	22.90	36.40
ProbabilidadLluvia	18945.00	5.75	13.82	0.00	0.00	0.00	2.00	90.00
HumedadRelativa	18945.00	57.84	26.93	0.00	35.00	60.00	82.00	100.00
VelocidadViento	18945.00	11.98	7.40	0.00	7.00	9.40	15.10	51.50
DireccionViento	18945.00	68.36	84.67	0.00	0.00	47.00	102.00	359.00
CoberturaNubes	18945.00	46.56	34.98	0.00	12.00	45.00	77.00	100.00
IndiceUV	18945.00	2.09	3.30	0.00	0.00	0.00	3.00	14.00
CodCondCielo	18945.00	2.15	1.30	1.00	1.00	1.00	3.00	5.00
CodDirViento	18945.00	4.04	1.58	1.00	3.00	4.00	4.00	9.00
VelocidadRafaga	18945.00	18.03	10.76	0.00	10.30	15.00	24.00	65.00
DPT	18945.00	8.42	5.88	-18.30	5.00	9.70	13.00	43.00
Generacion_prev_day	18945.00	9748.95	12091.66	0.00	0.00	371.39	23562.55	29966.14

3. Metodologías y mediciones

El aprendizaje supervisado, de acuerdo a (Igual & Seguí, 2017), está compuesto por algoritmos que aprenden de un conjunto de entrenamiento de ejemplos etiquetados (ejemplares) para generalizar al conjunto de todas las entradas posibles. Ejemplos de técnicas de aprendizaje supervisado: *logistic regression*, *support vector machines*, *decision trees*, *random forest*, etc.

3.1 Métodos empleados

Para este estudio, utilizaremos la librería `scikit-learn` de Python para realizar la comparación de los modelos. A continuación, se describen los métodos que analizaremos:

1. **Ada Boost**: (Freund & Schapire, 1997) es un metaestimador que comienza ajustando un regresor en el conjunto de datos original y, a continuación, ajusta copias adicionales del regresor en el mismo conjunto de datos, pero en el que los pesos de las instancias se ajustan en función del error de la predicción actual. De este modo, los regresores posteriores se centran más en los casos difíciles.

2. **Bagging**: (Breiman, 1996) es un metaestimador de conjunto que ajusta regresores base cada uno en subconjuntos aleatorios del conjunto de datos original y luego agrega sus predicciones individuales (ya sea por votación o por promedio) para formar una predicción final.

3. **Gradient Boosting**: (Friedman, 2001) este estimador construye un modelo aditivo por etapas y permite optimizar funciones de pérdida diferenciables arbitrarias. En cada etapa se ajusta un árbol de regresión sobre el gradiente negativo de la función de pérdida dada.

4. **Random Forest**: (Breiman, 2001) es un metaestimador que ajusta una serie de árboles de decisión clasificatorios en varias submuestras del conjunto de datos y utiliza el promedio para mejorar la precisión predictiva y controlar el sobreajuste.

5. **XGBoost**: (Zhang et al., 2020) combina una estrategia de precompletado no supervisada con un enfoque de aprendizaje automático supervisado, en forma de refuerzo de gradiente extremo.

6. **LGBM**: «LGBM Documentation», 2023 utiliza algoritmos basados en histogramas, que agrupan los valores de las características continuas (atributos) en intervalos discretos. Esto acelera el entrenamiento y reduce el uso de memoria.

3.2 Métricas para Medir Modelos de Regresión

Las métricas para medir modelos de regresión son herramientas fundamentales para evaluar qué tan bien un modelo se ajusta a los datos y qué precisión tiene en la predicción de valores numéricos. Algunas de las métricas más comunes para medir modelos de regresión incluyen:

1. **Error Cuadrático Medio (MSE)**: El MSE calcula la media de los errores al cuadrado entre las predicciones del modelo y los valores reales. Cuanto menor sea el MSE, mejor será el modelo. Sin embargo, es sensible a valores atípicos, ya que castiga más los errores grandes.

2. **Raíz del Error Cuadrático Medio (RMSE)**: El RMSE es simplemente la raíz cuadrada del MSE. Proporciona una medida del error en la misma unidad que la variable dependiente y es más interpretable.

3. **Error Absoluto Medio (MAE)**: El MAE calcula el promedio de los valores absolutos de los errores entre las predicciones y los valores reales. Es menos sensible a valores atípicos en comparación con el MSE.

4. **Coeficiente de Determinación (R^2)**: El R^2 proporciona una medida de la bondad del ajuste del modelo y cuánta variabilidad en los datos es explicada por el modelo. Un valor cercano a 1 indica un buen ajuste, mientras que un valor cercano a 0 indica que el modelo no explica mucha variabilidad.

5. **Error Porcentual Absoluto Medio (MAPE)**: El MAPE mide el error promedio como un porcentaje de los valores reales. Es útil para comprender el error relativo del modelo en comparación con el tamaño de los valores reales.

Para este estudio analizaremos cada una de las métricas y tomaremos una decisión en base a ello.

4. Selección del mejor modelo

Tomaremos en cuenta para la selección de los datos futuros de prueba un número aleatorio entre 5 y 30 días; mientras que los datos de entrenamiento restantes serán porcentajes aleatorios entre 70 %, 80 %, 90 % y 100 %.

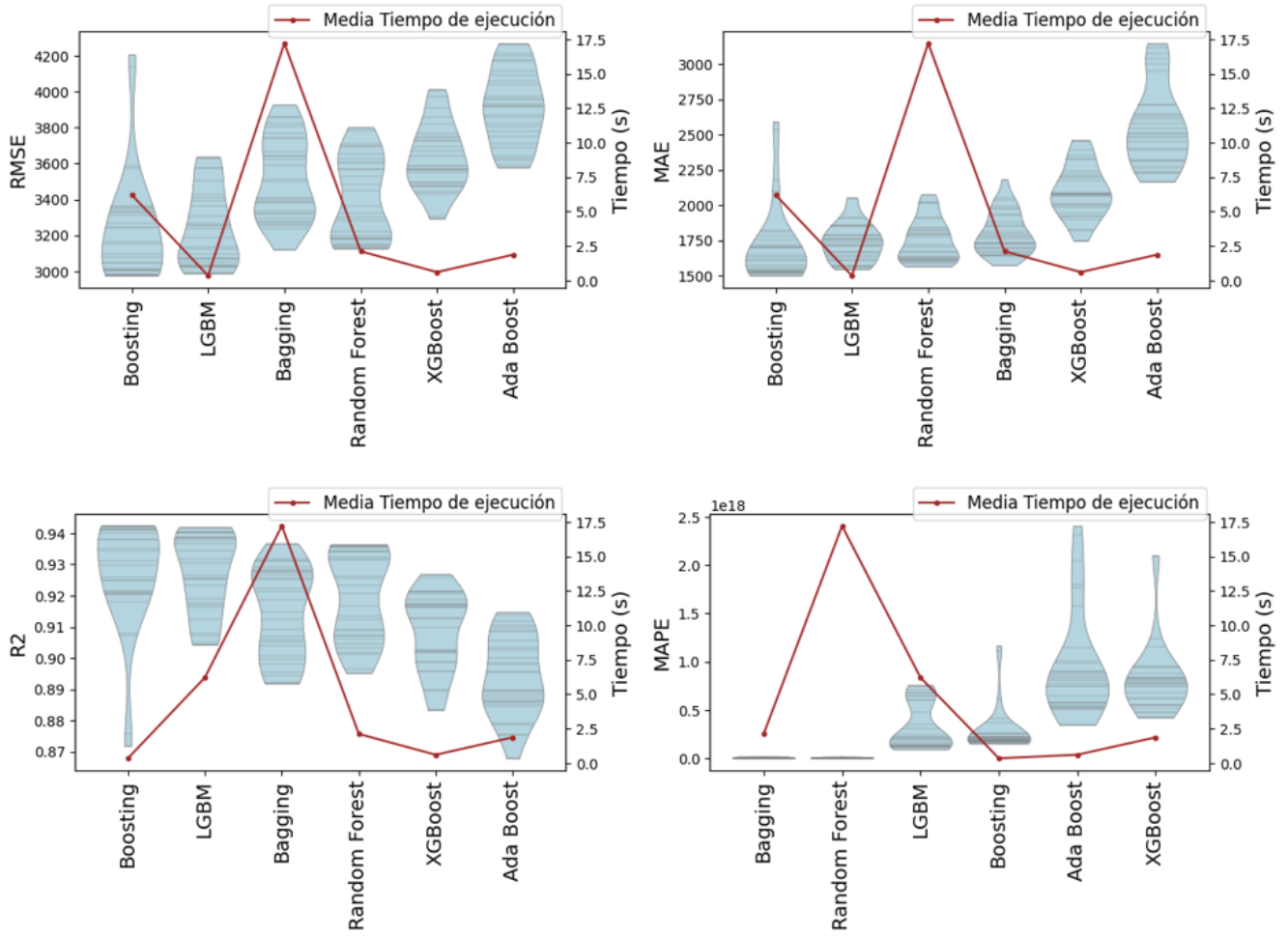

Figura 2. Comparación entre las métricas RMSE, MAE, R^2 y MAPE.

Al analizar las comparativas que se muestran en la figura 2, organizadas de mejor a peor desempeño, se destaca que el modelo *Boosting* lidera en todas las métricas. Sin embargo, el modelo *LGBM*, que en 3 de las cuatro métricas está en segundo lugar, presenta menor variación con respecto del primero, por lo que hemos decidido seleccionarlo como el mejor modelo. Para este caso, despreciamos el MAPE, por tener posiciones muy diferentes que el resto.

4.1 Diseño de experimentos

Habiendo seleccionado el modelo *LGBM*, partiremos de las medias de sus métricas de desempeño, mostradas en el cuadro 2 para realizar la selección de diferentes parámetros e intentar mejorar sus números. Se usará la librería `scikit-learn` `LGBMRegressor`, para realizar la experimentación de cada combinación de parámetros.

- **`num_leaves`:** Valores aleatorios enteros [2 - 1024].
- **`subsample`:** Valores aleatorios flotantes [0.05 - 1].
- **`colsample_bytree`:** Valores aleatorios flotantes [0.05 - 1].
- **`min_data_in_leaf`:** Valores aleatorios enteros [1 - 100].
- **`Dias`:** Tamaño en días que se van a predecir [5, 7, 10, 15, 30].
- **`Porcentajes`:** Porcentaje de datos restantes del pasado que se usará para entrenamiento [.7, .8, .9].

Haciendo una prueba con hiperparámetros aleatorios y todas las combinaciones entre los días y porcentajes de entrenamiento, decidimos primero evaluar cuántos días es más óptimo predecir de acuerdo a este modelo y qué porcentaje del pasado usar.

Cuadro 2. Valores promedio obtenidos de la selección del mejor modelo de *LGBM*.

RMSE	MAE	R^2	MAPE	Tiempo
3246.366	1741.633	0.926	3.360E+17	0.340

Si observamos la figura 3, podemos observar que las pruebas al predecir 5 días son las que tienden a arrojar un menor MAE. Si observamos la figura 4, vemos que las distribuciones de los 5 días son mejores. Aplicando una prueba de Wilcoxon para comparar las distribuciones no paramétricas de cada prueba con respecto a sus MAEs arrojados, dio con un 95 % de confianza que la prueba de predicción de 5 días usando el 70 % de los datos es menor a todas las demás.

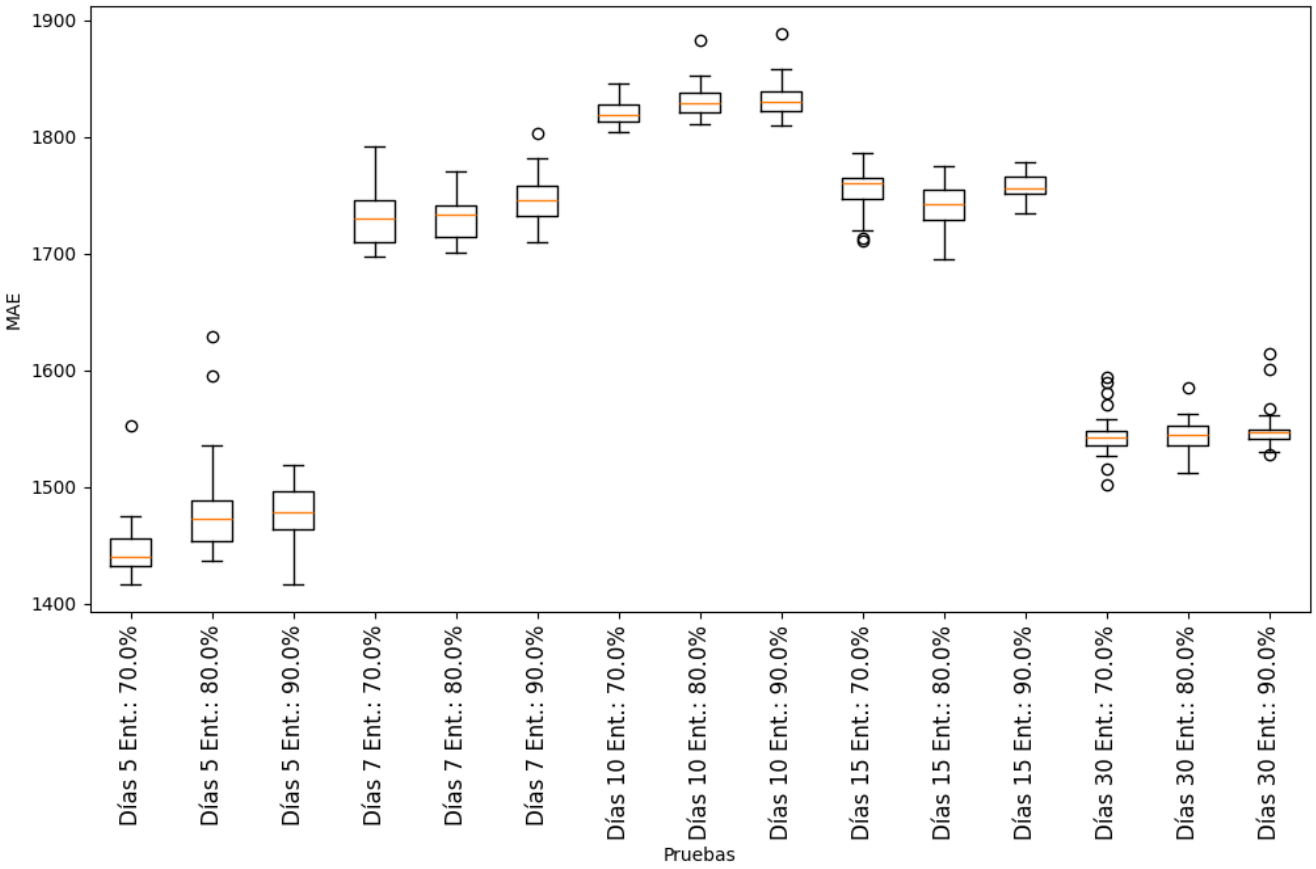

Figura 3. Comparación de los MAEs entre las diferentes combinaciones de Días y Porcentajes de entrenamiento.

Una vez obtenidas las variables del día y del porcentaje de entrenamiento óptimos, usamos la librería `Optune` para determinar los mejores hiperparámetros, los cuales dieron los siguientes resultados:

- **`num_leaves`:** 500.
- **`subsample`:** 0.10698460631792395.
- **`colsample_bytree`:** 0.7272836809565294.
- **`min_data_in_leaf`:** 85.
- **`Dias`:** 5.
- **`Porcentajes`:** 0.7.

Una vez aplicados los parámetros, de acuerdo al cuadro 3, vemos una mejoría en las métricas de desempeño, por lo que lo consideramos un caso existoso. Vemos en la figura 5 una comparación entre la Generación real versus la predicha por el modelo, lo cual se ve ajustada de acuerdo a la realidad.

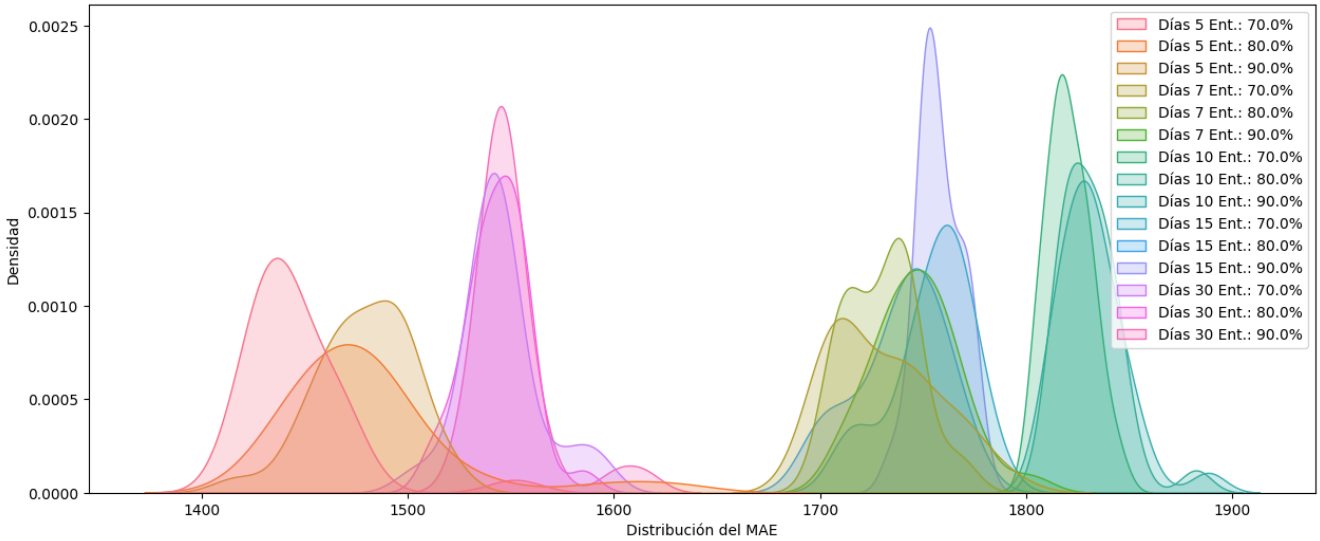

Figura 4. Comparación de las densidades de los MAEs entre las diferentes combinaciones de Días y Porcentajes de entrenamiento.

Cuadro 3. Valores promedio antes y después del diseño de experimentos.

RMSE	MAE	R^2	MAPE	Tiempo
3246.366	1741.633	0.926	3.360E+17	0.340
2600.208	1515.945	0.954	3.809E+17	0.583

Como dato añadido, se muestra en la figura 6, la importancia de cada variable usada en el modelo, donde vemos que las características principales, fueron el número de la hora, la generación de la hora anterior, la cobertura de las nubes y la temperatura.

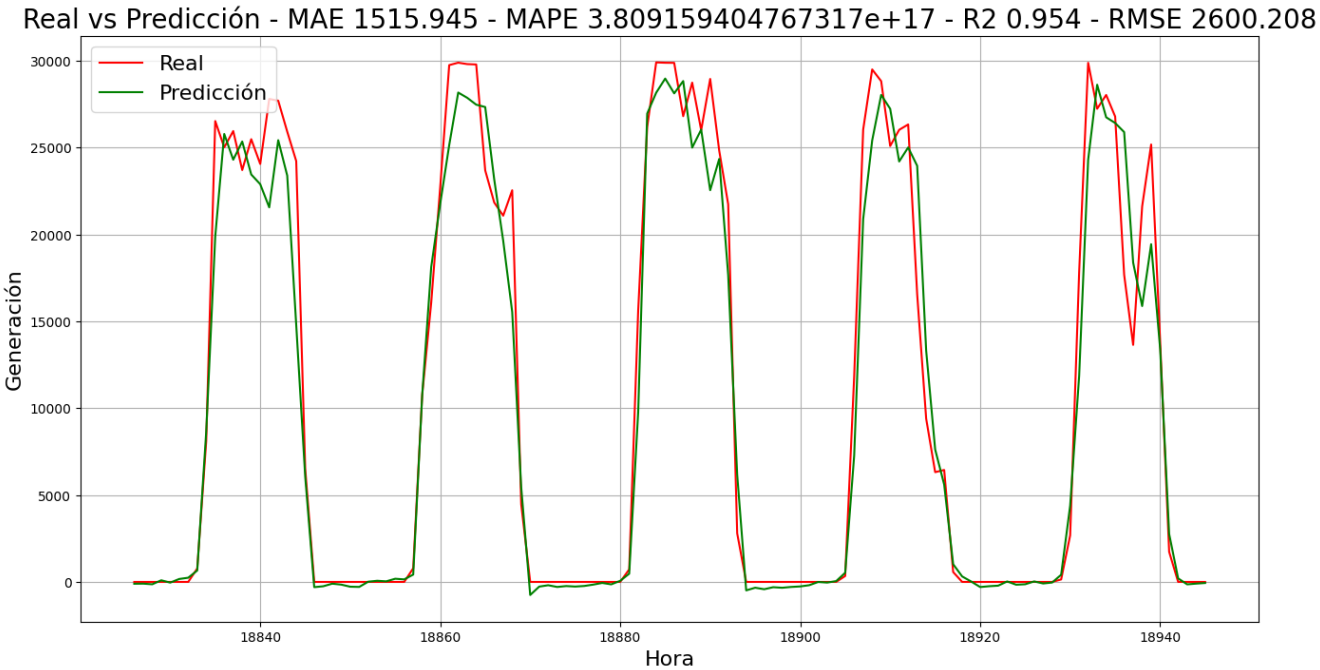

Figura 5. Comparación de Generación de Energía Real vs Energía predecida.

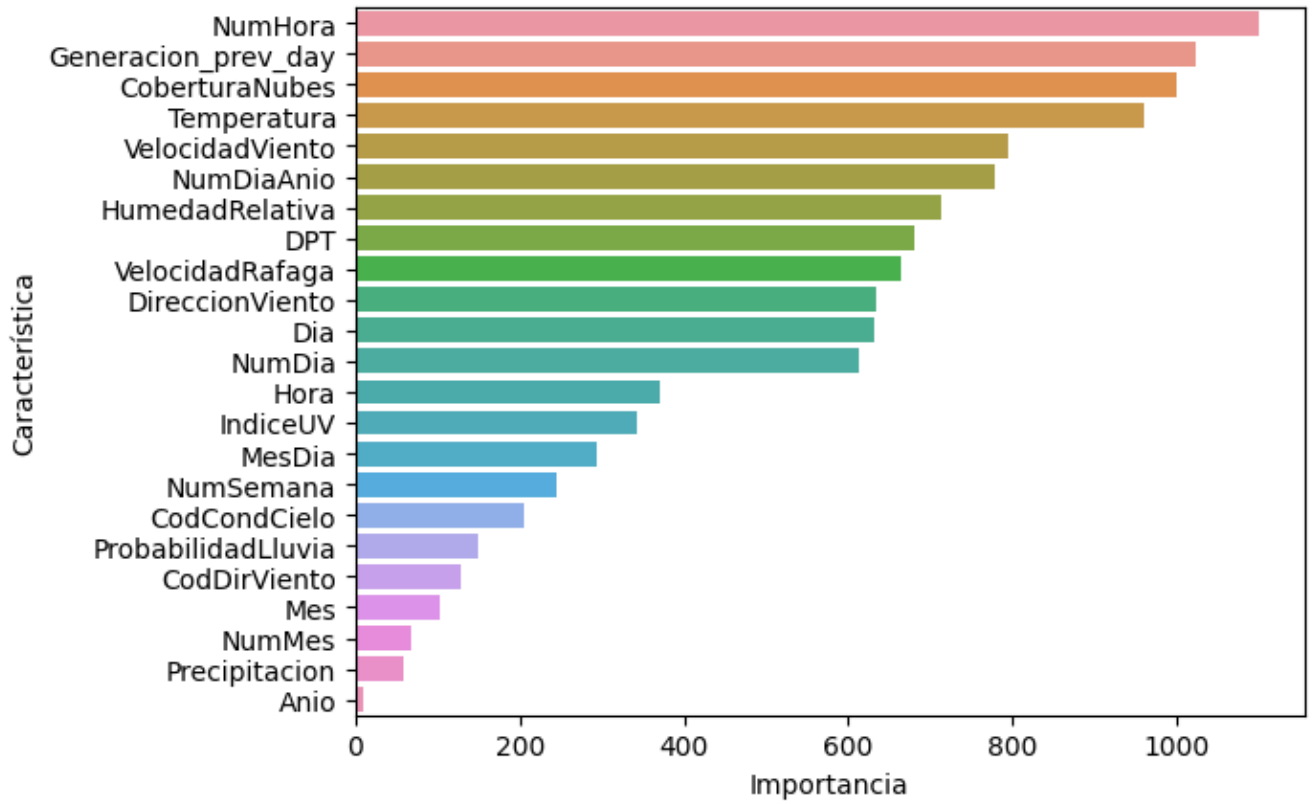

Figura 6. Importancia de las variables para el modelo *LGBM*.

5. Conclusión

Con base en la evaluación de diversos modelos supervisados, hemos concluido que el modelo LGBM (Light Gradient Boosting Machine) sobresale como el mejor entre ellos, respaldado por métricas de desempeño tales como MAE, R2, RMSE y MAPE.

Adicionalmente, llevamos a cabo un diseño de experimentos para optimizar los parámetros del modelo seleccionado. Este proceso nos condujo a la conclusión de que la configuración más óptima implica predecir 5 días utilizando el 70 % de los datos disponibles. Esta elección se fundamenta en la maximización del rendimiento del modelo en términos de precisión y generalización.

Consideramos que existen oportunidades para mejorar aún más nuestras predicciones. Una estrategia sería la exploración de técnicas no supervisadas, como K-Means, para identificar patrones y agrupamientos en los regresores. Integrar estos agrupamientos en el modelo supervisado podría ofrecer una mejora en la capacidad de predicción, especialmente al lidiar con diferentes comportamientos climáticos.

Además, se propone una mejora futura enfocada en la evaluación de regresores para estimar la probabilidad de error asociada con las predicciones climatológicas. Este enfoque permitirá una comprensión más completa y robusta de la confiabilidad de nuestro modelo.

En resumen, la combinación de un modelo supervisado eficaz como LGBM, optimizado a través de un diseño de experimentos, junto con la exploración de enfoques no supervisados y la evaluación de regresores para estimar la probabilidad de error, conforma una estrategia integral para mejorar aún más la calidad de las predicciones climáticas y proporcionar resultados más confiables en escenarios futuros para la predicción de la Generación de Energía.

Referencias

Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*, 123-140.

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5-32.

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, *55*(1), 119-139.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189-1232.

Igual, L., & Seguí, S. (2017). Supervised Learning. En *Introduction to Data Science: A Python Approach to Concepts, Techniques and Applications* (pp. 67-96). Springer International Publishing. https://doi.org/10.1007/978-3-319-50017-1_5

LGBM Documentation. (2023). lightgbm. https://lightgbm.readthedocs.io/en/latest/Features.html

Zhang, X., Yan, C., Gao, C., Malin, B. A., & Chen, Y. (2020). Predicting missing values in medical data via XGBoost regression. *Journal of healthcare informatics research*, *4*, 383-394.