Aprendizaje Supervisado

Claudia Lissette Gutiérrez Díaz[1]*

Palabras Clave
Regresión Lineal, Bayesian Ridge, Árboles de Decisión, Random Forest, Reforzamiento Adaptativo (Adaptive Boosting), Bagging, Gradiente Descendiente, MSE, RMSE, MAE, R^2, MAPE

Fecha: 30 de octubre 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 Regresión Lineal, Bayesian Ridge, Árboles de Decisión, Random Forest, Adaptive Boosting, Bagging y Gradiente Descendiente, 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.

Nuestra investigación se centra en la comparación de los datos reales de generación de energía solar con las predicciones realizadas por los diferentes modelos. El objetivo es identificar el modelo que ofrezca las predicciones más precisas y consistentes.

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 es la capacidad de los paneles solares. Se realizaron procesos de filtrado, diferenciación y normalización para obtener un conjunto de datos preparado de la siguiente manera:

- **Generación de Energía (kWh):** Esta es la variable dependiente en nuestro conjunto de datos y representa la cantidad de energía generada por la planta de energía solar en kilovatios-hora en un intervalos de una hora.

- **Condiciones Ambientales:** Las siguientes variables independientes representan las condiciones ambientales que podrían influir en la generación de energía solar:

  1. **Número de Semana:** Indica el número de semana, del 1 al 52 a la cual pertenece el dato.
  2. **Número de Día:** Indica el número de día, del 1 al 366 a la cual pertenece el dato.
  3. **Número de Hora:** Indica el número de hora, del 1 al 24 a la cual pertenece el dato.
  4. **Probabilidad de Lluvia (%):** La probabilidad de precipitación en forma de lluvia en porcentaje.
  5. **Cobertura de Nubes (%):** Porcentaje de nubes o niebla o bruma que cubren el cielo.
  6. **Dirección Categórica del Viento:** La dirección del viento categorizada en términos de puntos cardinales (por ejemplo, "norte", "sur", "este", "oeste").
  7. **Temperatura del Punto de Rocío (°C):** La temperatura a la cual el aire se satura y la humedad relativa es del 100 %.
  8. **Velocidad del Viento (km/h):** La velocidad del viento en kilómetros por hora.
  9. **Temperatura (°C):** La temperatura ambiente en grados Celsius.
  10. **Índice UV:** El índice de radiación ultravioleta que mide la intensidad de la radiación solar.
  11. **Condición del Cielo:** Una variable que describe las condiciones del cielo, que podría ser una categoría como "despejado", "poco nuboso", "nublado", "medio nublado", "cielo nublado", "cielo cubierto".

- **Preparación del Conjunto de Datos:** Las variables independientes fueron sometidas a un proceso de filtrado mediante el método de selección de características estadísticas para identificar las más relevantes para la predicción de la generación de energía solar. Además, se eliminó la estacionalidad mediante el método de diferenciación. Los datos independientes se normalizaron para garantizar la consistencia en las escalas.

Semana	Día	Hora	Prob.Lluvia	%Nubes	Dir.Viento	DPT	Vel.Viento	Temp.	IndiceUV	CondCielo
0.5000	0.5028	0.0000	0.5280	0.7000	0.3750	0.5072	0.5590	0.4474	0.4444	0.7500
0.5000	0.5028	0.0001	0.5528	0.7000	0.3750	0.5072	0.5842	0.4737	0.4444	0.7500
0.5000	0.5028	0.0001	0.5342	0.7000	0.3750	0.5072	0.5590	0.4737	0.4444	0.2500
0.5000	0.5028	0.0002	0.5031	0.7000	0.3750	0.4928	0.5842	0.4474	0.4444	0.2500

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

Este conjunto de datos preparado se utiliza para realizar análisis estadísticos y modelado predictivo, con el propósito de entender cómo las condiciones ambientales afectan la generación en la planta de energía solar.

3. Metodologías y mediciones

Los métodos supervisados son un enfoque fundamental en el aprendizaje automático y se utilizan para predecir o clasificar datos en función de ejemplos de entrenamiento previamente etiquetados. En un contexto supervisado, se dispone de un conjunto de datos de entrenamiento que consta de pares de entrada y salida (o etiquetas), y el objetivo es aprender una función que mapee las entradas a las salidas de manera precisa. De acuerdo a **(Chakraborty et al., 2023)**, la predicción de energías solares ha aumentado de auge y se han ido mejorando las metodologías de predicciones, pero es necesario iniciar desde las bases estadísticas, usando modelos como árboles de decisión, random forest, regresiones lineales y logísticas, etc.

3.1 Métodos Supervisados

Para este estudio, utilizaremos la librería («scikit-learn: Machine Learning in Python», s.f.) de Python para realizar la comparación de los modelos. A continuación, se describen los métodos que analizaremos:

1. **Regresión Lineal (LR)**: Se utiliza para modelar y predecir relaciones lineales entre una variable dependiente y una o más variables independientes. En su forma más simple, busca encontrar la línea recta que mejor se ajusta a los datos, minimizando la suma de las diferencias entre las observaciones reales y las predicciones. Esta técnica es ampliamente utilizada para tareas de predicción y estimación, y su simplicidad e interpretabilidad la hacen valiosa en la modelización de fenómenos con relaciones lineales. La regresión lineal se puede extender para abordar problemas multivariados y complejos al considerar múltiples predictores y utilizar variantes como la regresión lineal múltiple, la regresión polinómica y la regresión regularizada.

2. **Bayesian Ridge (BR)**: Es un modelo de regresión bayesiana que combina el enfoque de regresión lineal con la teoría bayesiana. A diferencia de la regresión lineal tradicional, Bayesian Ridge no solo proporciona estimaciones de los coeficientes de regresión, sino que también modela la incertidumbre asociada con estos coeficientes. Esto lo convierte en una herramienta poderosa para la regresión, ya que puede manejar datos ruidosos y proporcionar estimaciones robustas. Además, permite incorporar información previa sobre los coeficientes, lo que es especialmente útil en situaciones en las que se dispone de conocimiento previo sobre las relaciones entre las variables.

3. **Árboles de Decisión (AD)**: Tiene una estructura de árbol jerárquica, que consta de un nodo raíz, ramas, nodos internos y nodos hoja. Se utilizan para ajustar una curva senoidal con observaciones ruidosas adicionales. Como resultado, aprenden regresiones lineales locales que aproximan la curva senoidal.

4. **Random Forest (RF)**: Combina métodos de aprendizaje en conjunto con el marco de árboles de decisión para crear múltiples árboles de decisión dibujados de manera aleatoria a partir de los datos, promediando los resultados para generar un nuevo resultado que a menudo resulta en predicciones/clasificaciones sólidas.

5. **Reforzamiento adaptativo (RA, Adaptive Boosting o AdaBoost)**: Es un algoritmo de ensamblado en aprendizaje automático que se destaca por su capacidad para mejorar la precisión de los modelos combinando múltiples clasificadores débiles o regresores débiles. Lo que hace que sea único es su enfoque en ejemplos difíciles; en cada iteración, asigna un mayor peso a los ejemplos mal clasificados, lo que permite que el algoritmo se centre en las instancias más desafiantes. Finalmente, este combina los clasificadores o regresores débiles en un modelo fuerte mediante votación ponderada, lo que resulta en un modelo altamente preciso y generalizable.

6. **Bagging**: Es un meta-estimador de conjunto que ajusta regresores base en subconjuntos aleatorios del conjunto de datos original y luego agrega sus predicciones individuales (ya sea por votación o promediando) para formar una predicción final. Este tipo de meta-estimador se utiliza típicamente como una forma de reducir la varianza de un estimador tipo ”caja negra” (por ejemplo, un árbol de decisión) al introducir aleatoriedad en su procedimiento de construcción y luego formar un conjunto a partir de él.

7. **Gradiente Descendente (GD)**: El gradiente descendente es una técnica de optimización utilizada para entrenar modelos de regresión y clasificación. Se ajustan los parámetros del modelo de manera iterativa minimizando una función de costo. Puede utilizarse con diversos algoritmos, como el descenso de gradiente estocástico (SGD) o el descenso de gradiente por lotes.

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. Seleccionando el mejor modelo

Para la selección del mejor modelo, tomaremos en cuenta un porcentaje de entrenamiento del 70 % de los datos. Compararemos cada uno de los modelos mencionados con cada una de las métricas mencionadas en las secciones pasadas.

Al analizar las comparativas en la **figura 1**, organizadas de mejor a peor desempeño, se destaca que el modelo *Random Forest* lidera en todas las métricas, aunque con el compromiso de un mayor tiempo de cálculo. No obstante, si consideramos la eficiencia computacional y deseamos un sólido rendimiento, los modelos de *Bagging* y *Gradiente* se presentan como alternativas competitivas en el segundo lugar. Para esta ocasión, dado que la cantidad de datos abarca casi dos años y el tiempo de cálculo no es una preocupación primordial en este contexto, hemos elegido el modelo ***Random Forest*** como nuestra preferencia.

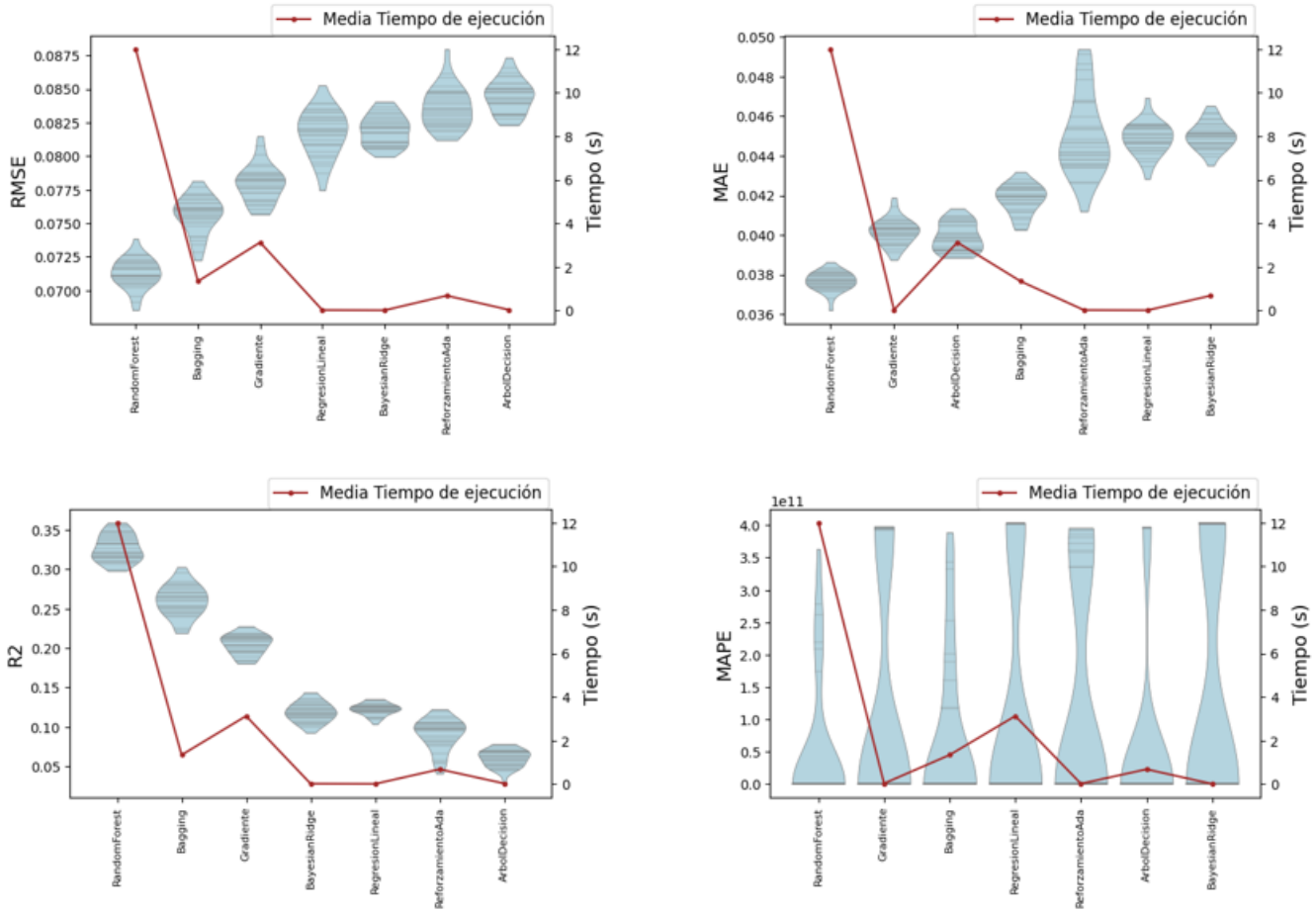

figura 1. Comparación entre las métricas RMSE, MAE, R^2 y MAPE.

5. Modelo de Regresión Random Forest

Teniendo en cuenta la discusión presentada en el artículo de (Ahmad et al., 2018), es evidente que la utilización de este modelo para predecir la generación de energía solar conlleva la necesidad de disponer de una cantidad significativa de datos, lo que, a menudo, se traduce en un alto consumo de recursos y tiempo. En nuestro caso particular, al no contar con un extenso historial de datos, es posible que este modelo no sea la elección más adecuada, y que sus resultados no alcancen el nivel óptimo esperado. Profundizemos un poco más en este modelo.

5.1 Características

Las características clave del modelo de regresión Random Forest incluyen:

- **Ensamblado de Árboles de Regresión**: El modelo se basa en la combinación de múltiples árboles de regresión, donde cada árbol se entrena en una submuestra aleatoria del conjunto de datos de entrenamiento.

- **Promedio de Predicciones**: Para realizar una predicción de regresión, el modelo promedia las predicciones de todos los árboles individuales, lo que mejora la precisión y reduce el riesgo de sobreajuste.
- **Selección Aleatoria de Características**: Cada árbol de regresión selecciona aleatoriamente un subconjunto de características en cada división, lo que introduce aleatoriedad y diversidad en el modelo.
- **Lidiar con Datos Ruidosos y No Lineales**: El modelo es eficaz en la mitigación de problemas de ruido en los datos y puede capturar relaciones no lineales entre las variables predictoras y la variable objetivo.
- **Tolerancia a sobreajuste**: Debido a su naturaleza de ensamblado y aleatoriedad, es menos propenso al sobreajuste que un solo árbol de regresión.
- **Interpretabilidad y Visualización**: Aunque los modelos Random Forest son menos interpretables que un solo árbol, aún es posible analizar la importancia relativa de las características en el proceso de toma de decisiones del modelo.

5.2 Comparación Pruebas vs. Predicción

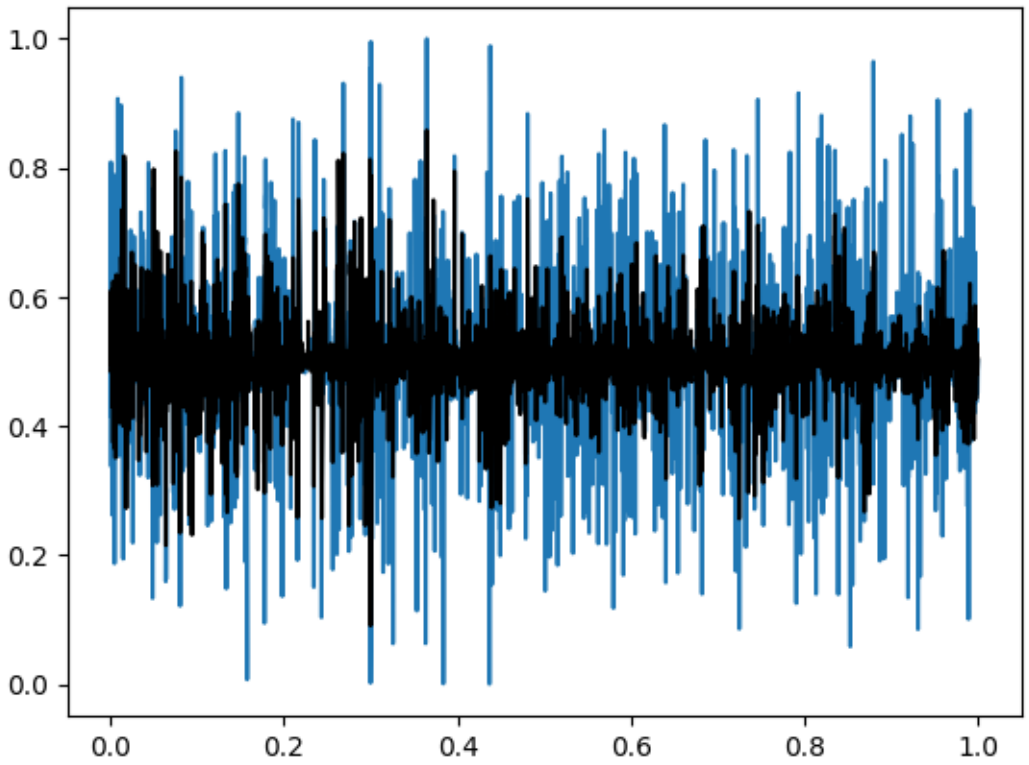

figura 2. Comparación la muestra de prueba y su predicción.

Como se observa en la **figura 2**, hay un comportamiento similar entre lo que se prueba y lo predicho, pero como se mencionó anteriormente, no es un resultado óptimo como el que se espera. Sin embargo, para el ejercicio, vamos a continuar convirtiendo los datos a las unidades de energía. Al desnormalizar los datos y regresar su estacionalidad, vamos a considerar 3 ajustes extras.

1. La energía no puede ser negativa, por lo que si alguna predicción al sumarse con el registro de 24 horas antes da un número negativo, éste se reemplazará por un 0.
2. De acuerdo a la capacidad de la planta, los paneles tienen una capacidad contratada de 30,000 kwh, por lo que el valor predicho no puede superar este valor, así que si éste rebasa el dicho número, se reemplazará por 30,000.
3. Durante la noche, no hay generación, por lo que no debe haber valores en esas horas. Como referencia para este registro, se tomará el Índice UV como factor que indicará si debe haber o no generación en esa hora, si el Índice UV es 0, se reemplazará el valor predicho por 0.

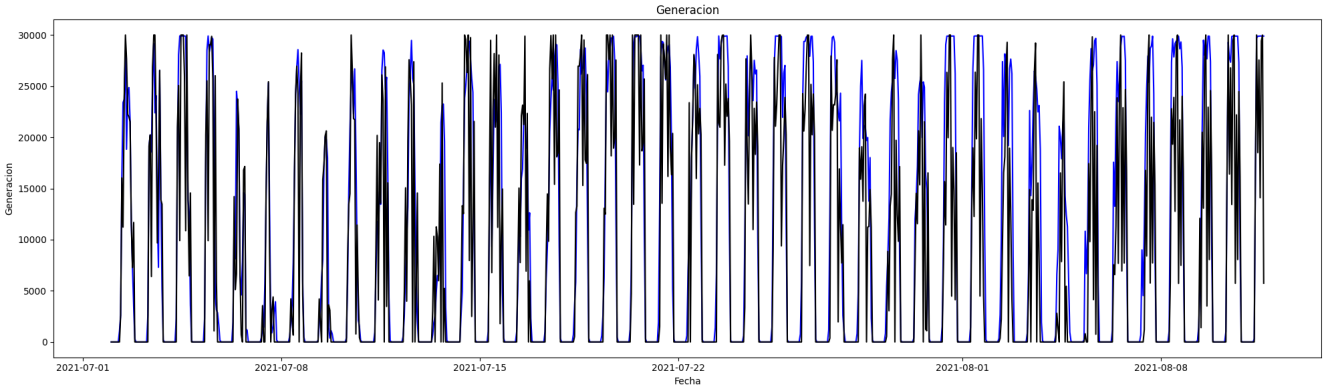

figura 3. Comparación de Generación de Energía Real vs Energía predictiva

Como se aprecia en la representación visual de la **figura 3**, que ha sido reducida para facilitar la comprensión, las áreas azules reflejan los valores reales, mientras que las áreas negras indican las predicciones del modelo. A simple vista, es evidente que las predicciones del modelo siguen de manera acertada el comportamiento esperado de los datos reales.

6. Conclusión

En conclusión, la aplicación del aprendizaje supervisado en la predicción de energía solar se revela como un enfoque valioso para la optimización y gestión de sistemas de generación de energía limpia. Hemos explorado una variedad de modelos de regresión, incluyendo Regresión Lineal, Bayesian Ridge, Árboles de Decisión, Random Forest, Adaptive Boosting, Bagging y Gradiente Descendiente, con el objetivo de seleccionar el modelo más adecuado para nuestras necesidades.

Mediante el análisis comparativo de datos reales y predicciones, respaldado por métricas de evaluación de errores como el MSE, RMSE, MAE, R^2 y MAPE, hemos identificado al modelo Random Forest como nuestra elección principal. Este modelo ha demostrado un rendimiento bueno al seguir con precisión el comportamiento esperado de la generación de energía solar, incluso en ausencia de un historial de datos extenso.

Referencias

Ahmad, M. W., Reynolds, J., & Rezgui, Y. (2018). Predictive modelling for solar thermal energy systems: A comparison of support vector regression, random forest, extra trees and regression trees. *Journal of cleaner production*, *203*, 810-821.

Chakraborty, D., Mondal, J., Barua, H. B., & Bhattacharjee, A. (2023). Computational solar energy–Ensemble learning methods for prediction of solar power generation based on meteorological parameters in Eastern India. *Renewable Energy Focus*, *44*, 277-294.

scikit-learn: Machine Learning in Python [Recuperado el 28 de octubre de 2023]. (s.f.). https://scikit-learn.org/