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: 21 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 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.

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:

- **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. **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. **Humedad Relativa (%):** Porcentaje de humedad presentada en el aire.
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. **Condición del Cielo:** Una variable que describe las condiciones del cielo, que cuyas descripciones pueden ser despejado, poco nuboso, nublado, medio nublado, cielo nublado o cielo cubierto.

- **Preparación del Conjunto de Datos:** Se realizaron procesos de limpieza de datos; se quitó la estacionalidad por el método de *diferenciación* (2023); se normalizaron los valores utilizando el estándar de mínimos y máximos de la librería `sklearn.preprocessing.MinMaxScaler`; se redujeron las variables usando la técnica de *Análisis de Componentes Principales, (PCA por sus siglas en inglés)* usando la librería `sklearn.decomposition.PCA`; y por último, se eliminaron valores vacíos.

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

Generación	Semana	Día	Hora	Hum. Relativa	Cobertura Nubes	Cond. Cielo	Dir. Viento
0.5045	0.8846	0.8666	0.6322	0.4880	0.0100	0.0000	0.8750
0.2410	0.5000	0.4972	0.4601	0.5178	0.5000	0.5000	0.3750
0.5332	0.9230	0.9111	0.1913	0.6666	0.2500	0.5000	0.2500
0.5001	0.9230	0.9055	0.6493	0.3154	0.5800	0.0000	0.6250

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.

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. **Bayes Ridge**: (Saqib, 2021) es un tipo de modelización condicional en la que la media de una variable se describe mediante una combinación lineal de otras variables, con el objetivo de obtener la probabilidad posterior de los coeficientes de regresión y, en última instancia, permitir la predicción en función de los valores observados de los regresores.

4. **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.

5. **Decision Tree**: (Suthaharan & Suthaharan, 2016) un modelo de aprendizaje supervisado que mapea jerárquicamente un dominio de datos en un conjunto de respuestas. Divide un dominio de datos (nodo) recursivamente en dos subdominios tales que los subdominios tienen una mayor ganancia de información que el nodo que se dividió.

6. **Lasso**: (Emmert-Streib & Dehmer, 2019) se trata de un método de análisis de regresión que realiza tanto la selección de variables como la regularización con el fin de mejorar la precisión de la predicción y la interpretabilidad del modelo de regresión estadística.

7. **Linear Regression**: (Granados, 2016) ajusta modelos lineales o linealizables entre una variable dependiente y más de una variables independientes.

8. **PLS Regression**: (Abdi, 2010) es una técnica reciente que combina y generaliza características del análisis de componentes principales (PCA) y de la regresión lineal múltiple.

9. **PCA Linear Regression**: (Ergon et al., 2014) para obtener vectores de puntuación ortogonales.

10. **Ridge Regression**: (Hoerl & Kennard, 1970) como medio de estimar los coeficientes de regresión con un error cuadrático medio menor que sus homólogos de mínimos cuadrados cuando los predictores están correlacionados.

11. **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.

12. **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.

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 un porcentaje de entrenamiento del 80 % de los datos de manera que el 20 % restante, corresponde al final de los datos, simulando una predicción de la generación futura. Compararemos cada uno de los modelos mencionados con cada una de las métricas mencionadas en las secciones pasadas.

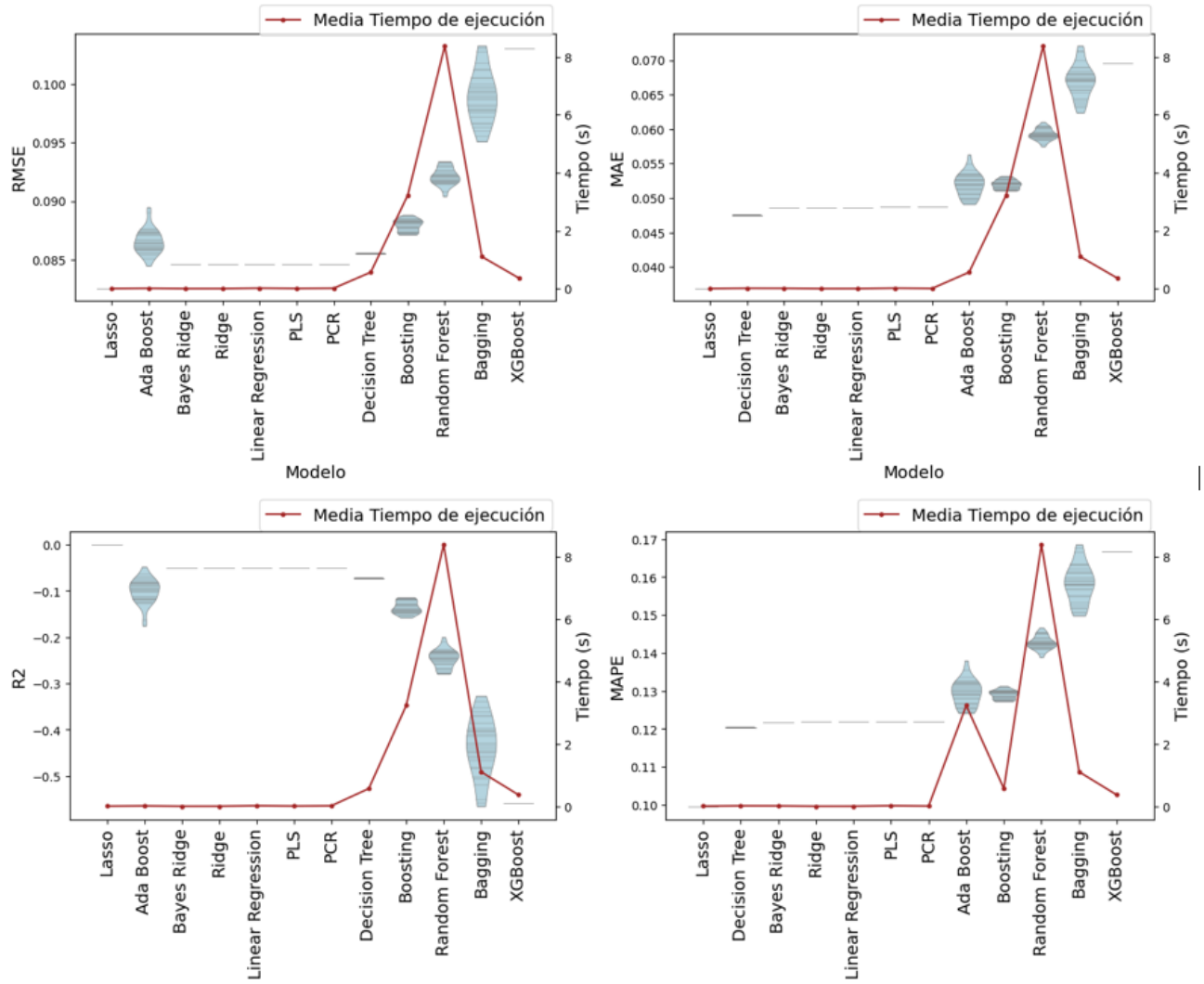


Figura 1. Comparación entre las métricas RMSE, MAE, R^2 y MAPE.

Al analizar las comparativas que se muestran en la figura 1, organizadas de mejor a peor desempeño, se destaca que el modelo *Lasso* lidera en todas las métricas. Sin embargo, modelos como *Decision Tree* y *Bayesian Ridge*, se ponen en segundo y tercer lugar, con muy poca variación de los datos. Nos interesa analizar a profundidad estos 3 modelos, para seleccionar el mejor.

4.1 Diseño de experimentos

Para elegir al mejor modelo, de los tres preseleccionados, vamos a variar sus diferentes parámetros y evaluar por separado el mejor de cada uno.

- **Lasso:** utilizaremos la librería `sklearn.linear_model.Lasso`, y se hará variación del siguiente parámetro:
  - **`alpha`:** [0.001, 0.01, 0.1, 1.0, 10.0].
- **Decision Tree:** utilizaremos la librería `sklearn.tree.DecisionTreeRegressor`, y se hará variación de los siguientes parámetros:
  - **`max_depth`:** [None, 2, 10, 20, 30].
  - **`min_samples_split`:** [2, 5, 10].
  - **`min_samples_leaf`:** [1, 2, 4].

- **`max_features`:** [‘auto’, ‘sqrt’, ‘log2’].

- **Bayesian Ridge:** utilizaremos la librería `sklearn.linear_model.BayesianRidge`, y se hará variación de los siguientes parámetros:

  - **`alpha_1`:** [1e-6, 1e-5, 1e-4].
  - **`alpha_2`:** [1e-6, 1e-5, 1e-4].
  - **`lambda_1`:** [1e-6, 1e-5, 1e-4].
  - **`lambda_2`:** [1e-6, 1e-5, 1e-4].

Aplicando la librería de `sklearn.model_selection.GridSearchCV`, se hará la selección de la mejor combinación de hiperparámetros de cada modelo, dando como resultado los siguientes parámetros:

- **Lasso:** `alpha`: 0.001.
- **Decision Tree:** `max_depth`: 2, `max_features`: ’sqrt’, `min_samples_leaf`: 1, `min_samples_split`: 10.
- **Bayesian Ridge:** `alpha_1`: 1e-06, `alpha_2`: 0.0001, `lambda_1`: 0.0001, `lambda_2`: 1e-06.

De acuerdo con los resultados de los mejores parámetros por modelo, podemos realizar una simulación del 20% de los datos que se desean predecir, y comparando nuestros resultados, con la realidad. Como podemos observar en la figura 2, no hay una diferencia relevante y apreciable con este tipo de gráfico, es consideramos necesario mostrarlo, para visualizar los residuales entre cada uno de los modelos.

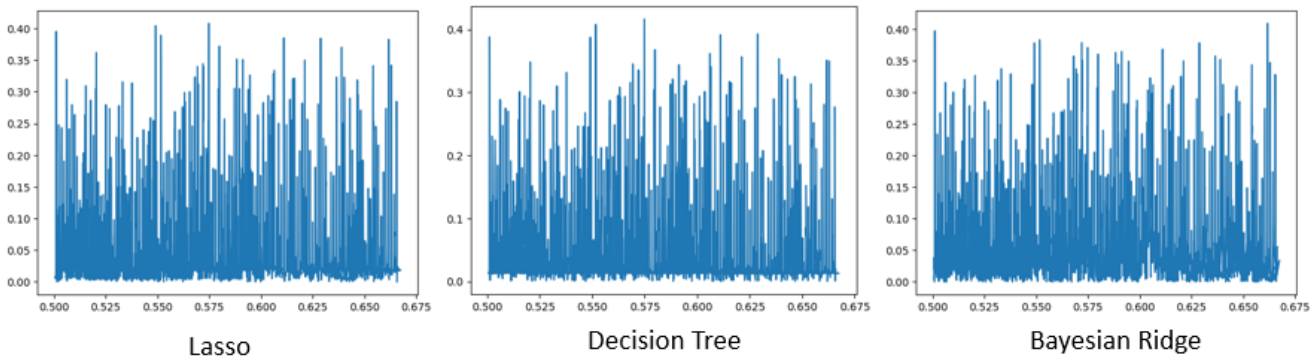


Figura 2. Diferencia entre la predicción real y la calculada por modelo *Lasso*, *Decision Tree* y *Bayesian Ridge*.

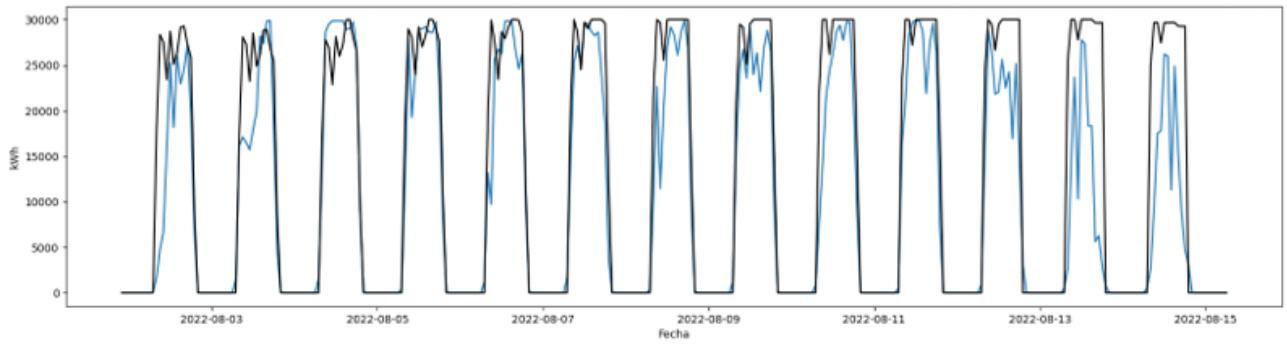
Otra de las maneras en las que podemos comparar los resultados entre los modelos, es convirtiendo los datos a la escala original y regresando su estacionalidad. Para este ejercicio, vamos a tomar en cuenta 3 ajustes extras:

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.

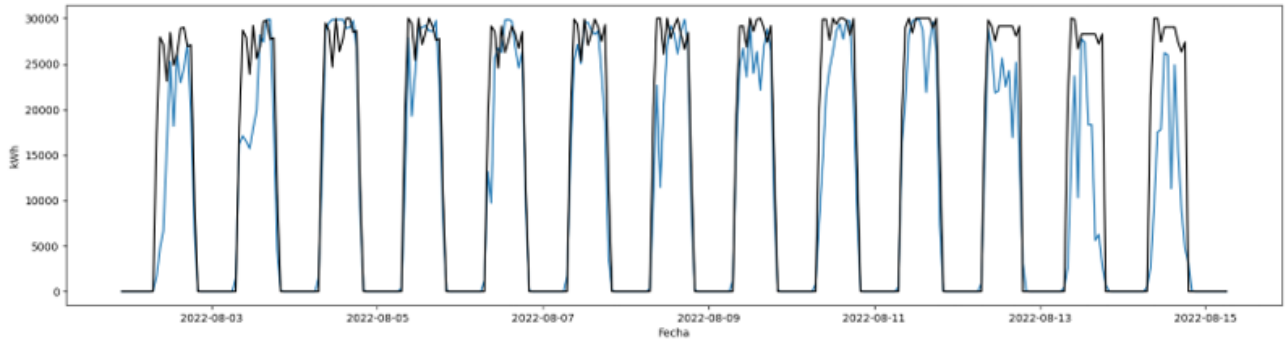
2. De acuerdo a la capacidad de la planta, los paneles tienen una capacidad contratada de 30,000 kwh, por lo que el valor predicho no puede superar este valor, así que si éste rebasa el dicho número, se reemplazará por 30,000.

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.

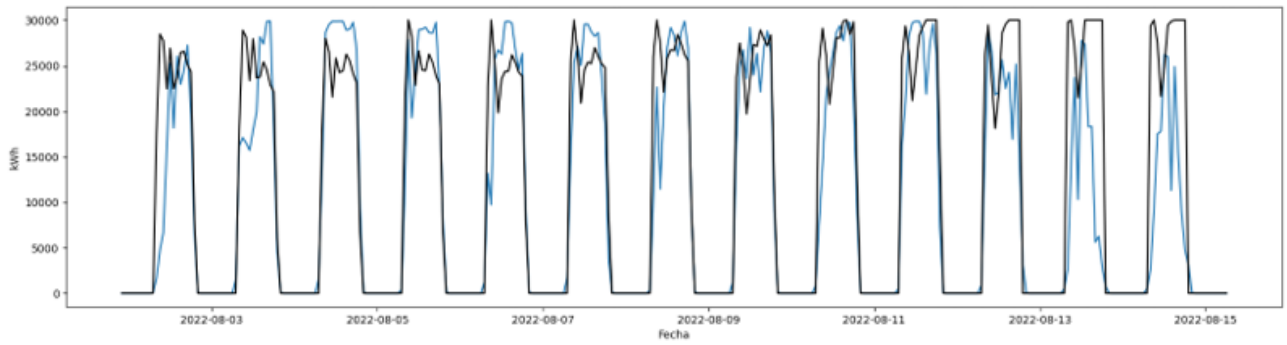
Como se aprecia en la figura 3, para la cual decidimos reducir la cantidad de registros mostrados y facilitar su visualización, los 3 modelos se ajustan a los datos, se respeta su estacionalidad y comportamiento diario, cuando en los datos originales, se detecta una baja en la energía, los modelos lo respetan y lo imitan. Con esto, seguimos sin poder concluir con exactitud, cuál modelo es el más apropiado y vencedor, por lo que a continuación se mostrará una comparación empalmada de la densidad de los residuales.



Lasso



Decision Tree



Bayesian Ridge

Figura 3. Comparación de Generación de Energía Real vs Energía predictiva por modelo *Lasso*, *Decision Tree* y *Bayesian Ridge*.

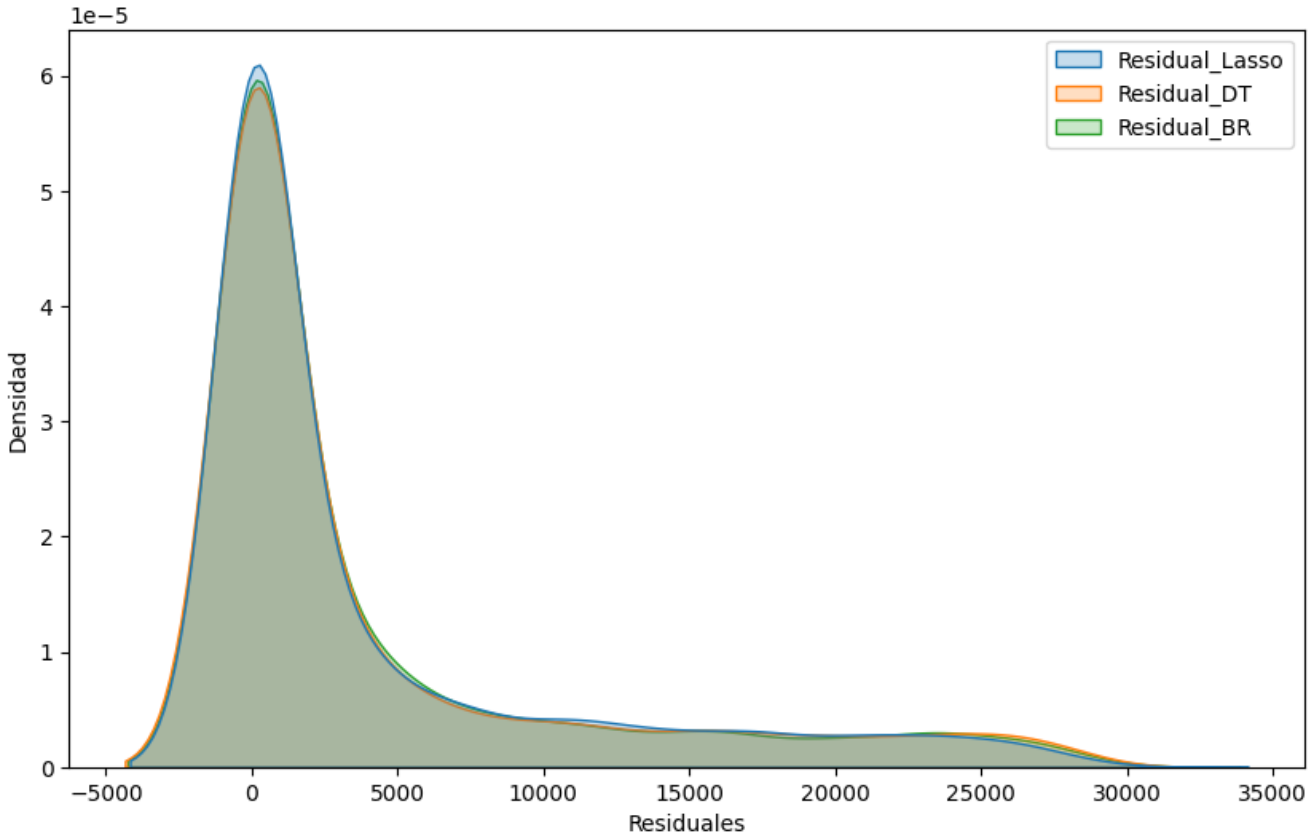


Figura 4. Densidad de residuales por modelo *Lasso*, *Decision Tree* y *Bayesian Ridge*.

De acuerdo a la figura 4, vemos que casi no hay diferencia entre los modelos entre su distribución de residuales, sin embargo, aquí sí podemos apreciar una mejor precisión en el modelo *Lasso*, al tener más valores concentrados en el 0.

Por último, si observamos el cuadro

Cuadro 2. Métricas de desempeño de los modelos *Lasso*, *Decision Tree* y *Bayesian Ridge*.

Método	MAE	MAPE	RMSE	R^2
Lasso	**4045.0314**	2.8160	**8062.4584**	0.4979
Decision Tree	4176.5080	**2.5153**	8375.2318	**0.4582**
Bayesian Ridge	4093.7367	2.7509	8206.5882	0.4798

Viendo las métricas del cuadro 2, hay un empate entre el modelo *Lasso* y *Decision Tree*, combinado con la comparación de la densidad entre los modelo, vamos a concluir como ganador al modelo *Lasso*.

5. Conclusión

Al comparar los modelos *Lasso*, *Decision Tree* y *Bayesian Ridge* para predecir la generación de energía solar, se observa que, aunque el modelo *Lasso* mostró un rendimiento ligeramente superior, las diferencias entre los modelos no son lo suficientemente significativas como para descartar alternativas. Se concluye que cualquiera de los tres modelos puede ser utilizado para la predicción.

A pesar de la elección de un “mejor”modelo, ninguno de los modelos mostró métricas de desempeño buenas. Esto sugiere que hay margen para mejorar la precisión mediante la exploración de características más avanzadas.

En términos de aprendizaje futuro, se sugiere considerar estrategias más avanzadas de Aprendizaje Automático, como Aprendizaje Profundo u otras técnicas, para obtener una precisión aún mejor en la predicción de la generación de energía solar.

A pesar de las métricas moderadas, el ejercicio de aprendizaje supervisado se considera satisfactorio para este análisis, sirviendo como punto de partida y proporcionando una base para futuras investigaciones y mejoras en el análisis de datos energéticos.

Referencias

Abdi, H. (2010). Partial least squares regression and projection on latent structure regression (PLS Regression). *Wiley interdisciplinary reviews: computational statistics*, *2*(1), 97-106.

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

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

Emmert-Streib, F., & Dehmer, M. (2019). High-dimensional LASSO-based computational regression models: regularization, shrinkage, and selection. *Machine Learning and Knowledge Extraction*, *1*(1), 359-383.

Ergon, R., Granato, D., & Ares, G. (2014). Principal component regression (PCR) and partial least squares regression (PLSR). *Mathematical and statistical methods in food science and technology Wiley Blackwell, Chichester*, 121-42.

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.

Granados, R. M. (2016). Modelos de regresión lineal múltiple. *Granada, España: Departamento de Economía Aplicada, Universidad de Granada*.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55-67.

Howell, E. (2023). *Demystifying Stationarity in Time Series Analysis*. https://www.youtube.com/watch?v=621MSxpYv60&ab_channel=EgorHowell

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

Saqib, M. (2021). Forecasting COVID-19 outbreak progression using hybrid polynomial-Bayesian ridge regression model. *Applied Intelligence*, *51*(5), 2703-2713.

Suthaharan, S., & Suthaharan, S. (2016). Decision tree learning. *Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning*, 237-269.

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.