Predicción de la velocidad del viento para la generación de energía eólica usando métodos de aprendizaje supervisado

Edwar Alejandro Londoño Ramírez Londoño, Sebastián Giraldo Zuluaga
Especialización Analítica y Ciencia de Datos
Universidad de Antioquia

RESUMEN El constante avance de la computación ha permitido que cada día surjan nuevas formas de encontrar solución a los diferentes problemas que se presenta. Es por ello que los modelos basados en datos con técnicas de machine learning han tomado gran relevancia en el momento de buscar alternativas que brinden soluciones de buena calidad en tiempos aceptables. En este trabajo se utiliza diferentes modelos de regresión que permitan pronosticar la velocidad del viento, la cual es insumo para determinar la potencia de salida en los generadores eólicos. Los datos utilizados son descargados de la página de The National Renewable Energy Laboratory (NREL) y se ajustan 4 modelos validados con 3 medidas de error.

PALABRAS CLAVE Pronóstico, Velocidad del Viento, Machine Learning, Análisis de Datos, Regresión Lineal, Máquina de Soporte Vectorial, árbol de decisión, Random Forest.

I. INTRODUCCION

En los últimos años el calentamiento global ha sido un aspecto de interés a nivel mundial dado que ha venido en aumento y tiene gran impacto en los ecosistemas. Es por ello, que los países están buscando diferentes alternativas para emitir menos CO2 al medio ambiente, entre ellas, incorporando en la matriz energética el uso de energías renovables como la generación solar y eólica que permitan atacar la problemática. La predicción de la velocidad del viento es un aspecto importante a tener en cuenta para estimar con mayor certeza la potencia de salida en los generadores eólicos, lo cual permite tener menores desbalances de potencia en los sistemas eléctricos y que este tipo de generación pueda participar en mercados eléctricos con una oferta de mayor firmeza, contribuyendo así con un menor costo en la operación del sistema eléctrico.

Los métodos de aprendizaje supervisado pueden ser utilizados para mejorar la precisión de estas predicciones y optimizar la producción de energía. Varias investigaciones han sido llevadas a cabo para determinar los modelos que mejor desempeño tienen en el momento de obtener las predicciones. En la literatura se pueden encontrar diferentes estudios donde se proponen modelos econométricos como VAR, ARMA, ARIMA, entre otros [1]; y otros modelos que usas redes neuronales y modelos de machine learning [2].

En este artículo, se presenta una introducción al uso de métodos de aprendizaje supervisado para la predicción de la velocidad del viento usando algoritmos como: Regresión Lineal (RL), Máquinas de Soporte Vectorial (MSV), Árboles de Decisión (AD) y RandomForest (RF).

II. ESTADO DEL ARTE

El machine learning es una de las tendencias globales en el campo de la predicción, es así que esta herramienta ha sido utilizando en múltiples investigaciones para resolver no solo problemas de predicción sino también de clasificación. En [3] las técnicas utilizadas son la RL donde el objetivo es encontrar los coeficientes que minimizan la suma de los residuos al cuadrado resultantes de las observaciones y la MSV donde se requiere encontrar los pesos que minimicen la función de pérdida. Se concluye que las predicciones con mayor precisión se dan con MSV como se evidencia en la Figura 1.

Observed Value	Linear Regression LR	GP-Poly Kernel (*exp*=1.0)	GP-Poly Kernel (*exp*=1.15)	GP-Poly Kernel (*exp*=1.16)	SVM Poly Kernel (*exp*=1.17)
7.2	6.6909	7.1789	7.0999	7.0938	7.0876
7.2	6.2863	7.1298	7.0016	6.9915	6.9812
7.2	6.1005	7.1964	7.0519	7.0392	7.0262
6.7	5.8483	7.1537	6.9228	6.9022	6.8811
6.7	5.542	7.1185	6.7479	6.7156	6.6826
MAE	0.9064	0.14346	0.14346	0.13866	0.1407
RMSE	0.934803	0.157003	0.157003	0.156036	0.157362

Figura 1. Resultados RL y MSV [3]

Adicionalmente, en [2] presentan el estado del arte, tendencias y desafíos para la predicción de la potencia eólica usando herramientas como: Redes Neuronales (NN), MSV, K Vecino más Cercano (KVMC) y RF, métodos con los cuales se puede hacer clasificación, regresión y agrupamiento.

Existen 3 aproximaciones principales en la predicción de la potencia eólica que son: método físico basado en predicciones meteorológicas numéricas, con alto costo

computacional y es ampliamente utilizado en la industria; método estadístico dividido en series de tiempo que tiene como base regresiones lineales, lo cual puede perder precisión ante las perturbaciones no lineales que presenta el viento e inteligencia computacional que tiene la capacidad de modelar fenómenos no lineales y sistemas complejos; y finalmente método híbrido el cual es menos utilizado. En la Figura 2 se observa como viene aumentando la tendencia en la investigación usando las diferentes técnicas de machine learning, evidenciando que las RN son las más utilizadas.

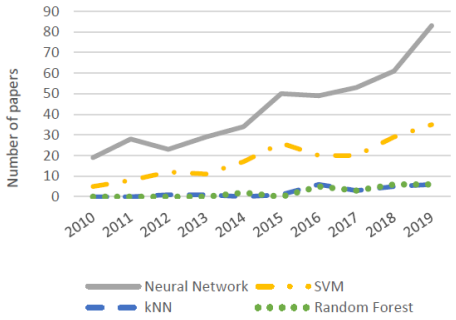

Figura 2. Investigación en métodos de predicción [2]

Por otro lado, en [1] mencionan que el pronóstico de la velocidad del viento se divide en 4 categorías: pronóstico de muy corto plazo, corto plazo, mediano plazo y largo plazo. Se realiza una comparación de los métodos estadísticos convencionales ARMA y ARIMA que contribuyen a la predicción de la velocidad del viento a corto plazo, usando modelos matemáticos para estimar parámetros y hacer predicciones basadas en valores pasados. Se compara con Redes Neuronales Artificiales (RNA) y MSV, donde ambas tienen ventajas y desventajas, pero las segundas tienen un rendimiento mejor en comparación con los métodos estadísticos convencionales. Las métricas de error utilizadas fueron Root Mean Squared Error (RMSE) y Mean Absolute Precentage Error (MAPE). En la Figura 3 y la Figura 4 se muestran los errores de cada modelo.

Method	2-hour ahead		24-hour ahead	
	RMSE	MAPE	RMSE	MAPE
ARMA (16,1)	0.822	7%	2.89	60%
ARIMA (2,1,2)	0.690	6%	2.97	57%

Figura 3. Error de los métodos estadísticos convencionales [1]

Method		Linear SVM	Polynomial SVM	RBF SVM	ANN
2-hour ahead	RMSE	0.565	0.552	1.571	0.723
	MAPE	5%	5%	18%	7%
6-hour ahead	RMSE	1.932	1.833	1.877	1.969
	MAPE	23%	25%	27%	28%
12-hour ahead	RMSE	2.432	1.954	2.395	2.127
	MAPE	45%	41%	40%	42%
24-hour ahead	RMSE	2.838	2.235	2.558	2.365
	MAPE	37%	34%	33%	33%
36-hour ahead	RMSE	2.824	2.629	3.474	2.749
	MAPE	36%	38%	38%	35%

Figura 4. Error de los métodos de machine learning [1]

Otras técnicas de machine learning las cuales son basadas en el majeo eficiente de los datos son utilizadas en [4] para predecir la potencia eólica de salida en 3 turbinas, una de Francia, otra de Turquía e información tomada de Kaggle. Es importante mencionar que esto difiere un poco de la salida esperada en el presente trabajo debido a que no es la velocidad del viento, pero esta variable está estrechamente relacionada con la potencia de salida, la cual presenta una alta correlación como se observa en la Figura 5 que corresponde a los datos del generador en Francia usado para verificar la eficiencia de los modelos en [4].

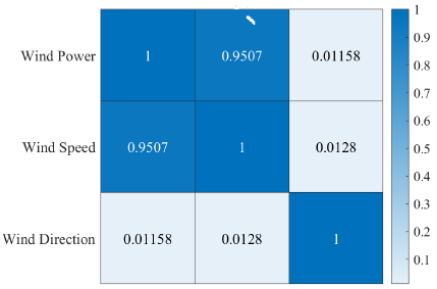

Figura 5. Matriz de Correlación turbina en Francia [4]

Los autores investigaron el desempeño de modelos para pronóstico uni-variado con la variable potencia eólica con datos de serie de tiempo. Se utilizó Optimización Bayesiana (OB) para encontrar los hiper-parámetros de Proceso de Regresión Gausiana (PRG), MSV con diferentes kernel y Ensemble Learning (ES) (Boosted trees y Baggged trees). Adicionalmente, se incorporaron medidas de rezago de manera dinámica para mejorar las predicciones. Las medidas de error utilizadas fueron RMSE, MAE y R^2.

III. IMPLEMENTACIÓN DEL MODELO

A. *Dataset*

El conjunto de datos usados fue descargado de la página de The National Renewable Energy Laboratory (NREL), el cual se especializa en la investigación y desarrollo de energía renovable, eficiencia energética, integración de sistemas de energía y transporte sostenible.

Son datos obtenidos a través de reanalisis que provee el estado de las variables meteorológicas del pasado. Estos datos son construidos con mediciones u observaciones realizadas y pronóstico de las variables en el pasado, basado en modelos de predicción modernos.

Las variables usadas para hacer la predicción de la velocidad del viento son descargadas con base en las coordenadas (11.77, -72.44) [5]. Este lugar está ubicado en la zona norte de Colombia donde se tiene diferentes proyectos de generación eólica. Se descargaron los datos desde el 01/01/2019 hora 1, hasta el 31/12/2021 hora 24, para un total de 26280 registros. A continuación, se listan las variables tomadas: Temperatura (°C), Punto de Rocío (°C), Ozono (ud), Humedad Realtiva (%), Presión (mbar), Precipitaciones (cm), Dirección del viento (grados), Velocidad del viento (m/seg).

Después de descargar la información en archivos .csv, se cargas los datos a un DataFrame de Python para realizar análisis exploratorio y limpieza de datos.
En la exploración de datos, se identifica que no hay datos faltantes, no hay datos duplicados y se extraen las columnas que se listaron anteriormente para realizar el modelo. No se incluye datos de tiempo dado que no se va a hacer análisis temporal para esta primera parte de la monografía. En la Figura 6 se observa el número de registros con un total de 26280 y 8 columnas.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 26280 entries, 0 to 26279
Data columns (total 8 columns):
 #   Column              Non-Null Count  Dtype  
---  ------              --------------  -----  
 0   Temperature         26280 non-null  float64
 1   Dew Point           26280 non-null  float64
 2   Ozone               26280 non-null  float64
 3   Relative Humidity   26280 non-null  float64
 4   Pressure            26280 non-null  int64  
 5   Precipitable Water  26280 non-null  float64
 6   Wind Direction      26280 non-null  int64  
 7   Wind Speed          26280 non-null  float64
dtypes: float64(6), int64(2)
memory usage: 1.6 MB
```

Figura 6. Información del DataFrame

En la Figura 7 se observan los datos normalizados, donde la variable de dirección del viento es la que más valores atípicos presenta.

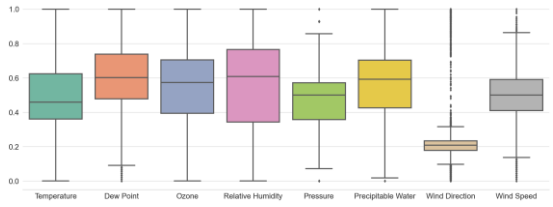

Figura 7. Normalización de los datos

Se realiza detección de datos atípicos usando el algoritmo LOF con 7 vecinos, encontrando sólo 26 registros. Se eliminan estos registros quedando un total de 26280.
Al validar la correlación de las variables, se observa que la humedad relativa presenta una relación inversa con respecto a la temperatura y directa con respecto al punto de rocío, siendo la primera una relación fuerte. También se presenta relación directa entre el punto de rocío y las precipitaciones, Figura 8.

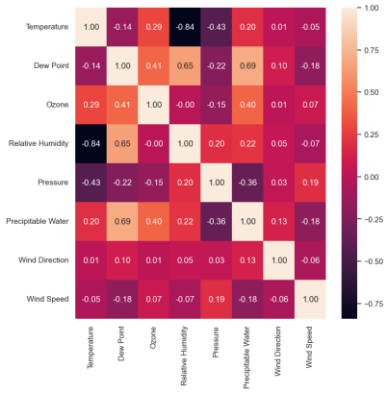

Figura 8. Correlación de las variables

Para la implementación de los modelos se hace una división del dataset en 80% para entrenamiento y un 20% para las pruebas.

B. Modelo de Regresión Lineal

Es una técnica estadística que tiene como objetivo predecir los valores de una variable continua dependiente con base en los valores de una o varias variables independientes.
Se realizó el ajuste con los datos de entrenamiento, luego se realiza una validación cruzada con 10 pliegues para evaluar el rendimiento del modelo, en la Figura 9, se observa que el ajuste no es bueno.

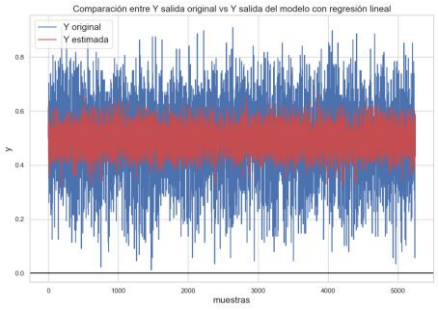

Figura 9. Predicción con la Regresión Lineal

C. Máquina de Soporte Vectorial

Inicialmente con una muestra de 1000 registros, se ajustaron 5 MSV con kernel: lineal, polinómico grado dos, grado 3, rbf y sigmoide con el objetivo de validar cual se ajusta mejor a los datos. Los resultados de la Figura 10, muestra que el mejor ajuste del R^2, tanto para training y test en el **rbf**.

```
Linear    Training:  0.43077246180368056 Test: 0.481537042903168
Poly 2    Training:  0.63996579234072 Test: 0.5955246210692464
Poly 3    Training:  0.6147118719712223 Test: 0.5200149855776526
rbf       Training:  0.8261262893704697 Test: 0.6713648658813858
sigmoide Training:  -6391550.187644434 Test: -5420084.8967929715
```

Figura 10. MSV con diferentes kernel para una muestra de 100 registros

Con el anterior resultado se aplica GridSearch a la MSV con kernel rbf, con los parámetros y el rango que se observa en la Tabla 1. En la Figura 11 se observa la predicción del SVR mostrando mejores resultados que la RL.

Tabla 1. GridSearch MSV con kernel rbf

Parámetro	Rango	Valor Final
C	0.1, 1, 10, 100, 100	1
Gamma	1, 0.1, 0.01, 0.01, 0.001, 'auto', 'scale'	'scale'

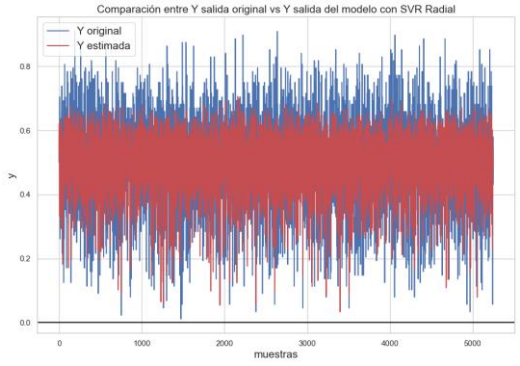

Figura 11. Predicción con el SVR

D. Árboles de Decisión

Los árboles de decisión funcionan dividiendo el conjunto de datos en subconjuntos, donde cada división representa un nodo en el árbol y las ramas representan las características. Su predicción se basa en el valor promedio del subconjunto. En la Tabla 2 se muestra el rango de parámetros con el que se ejecutó el GridSearch para el AD y los valores elegidos. En la Figura 12, se observa que el modelo queda limitado a valores de 0.6, y no se tiene una adecuada representación de la salida.

Tabla 2. GridSearch Árbol de decisión

Parámetro	Rango	Valor Final
max_depth	2, 4, 6, 8, 10	8
min_samples_split	15, 30, 45, 60	30
min_samples_leaf	15, 30, 45, 60	60
max_features	auto', 'sqrt', 'log2'	'scale'

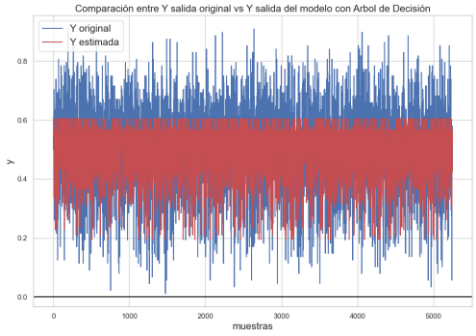

Figura 12. Predicción con el Árbol de decisión

E. Random Forest

Este es un método que combina múltiples árboles de decisión para mejorar el rendimiento y reducir el sobreajuste. En la Tabla 3 se muestran los valores de GridSearch y la Figura 13 el ajuste realizado, donde se observa mejor aproximación con respecto a los otros modelos.

Tabla 3. GridSearch Random Forest

Parámetro	Rango	Valor Final
n_estimators	20	20
max_features	2, 4, 6	2
max_depth	3, 10, 20	20

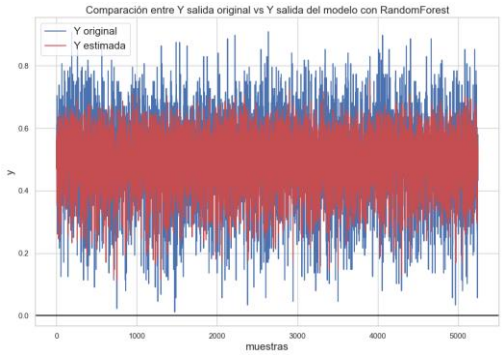

Figura 13. Predicción con Random Forest

F. Métricas de Desempeño

Para medir el desempeño de los diferentes modelos, se utilizaron 3 métricas tanto para training como para test que son: R^2, Mean Absolte Error (MAE) y el Root Mean Squared Error (RMSE). En la Tabla 4 se observan las métricas obtenidas utilizando validación cruzada con 10 pliegues. El modelo con mejor desempeño tanto en train y test es el Random Forest.

Tabla 4. Métricas de desempeño de los modelos

Modelo	Tipo	R2	MAE	RMSE
Rregresión Lineal	Train	0.12	0.1	0.13
	Test	0.11	0.1	0.13
SVR-rbf	Train	0.42	0.082	0.1
	Test	0.37	0.086	0.11
Arbol de Decisión	Train	0.33	0.089	0.11
	Test	0.29	0.082	0.11
RandomForest	Train	0.51	0.073	0.097
	Test	0.4	0.083	0.1

IV. CONCLUSIONES

Los modelos de machine learning con aprendizaje supervisado usados no mostraron buenas métricas para la predicción de la velocidad del viento con los datos proporcionados. Sin embargo, el modelo con mejor comportamiento es el RandomForest donde en train y test se obtuvieron valores de 0.51 y 0.4 respectivamente para el R^2, en las demás métricas también mostró mejor desempeño obteniendo valores menores con respecto a los otros modelos.

La Máquina de Soporte Vectorial tiene un mayor costo computacional en el momento de hacer evaluación de los hyper-parámetros.

Tener una base de datos con series de tiempo podía causar muchos sesgos al aplicar algoritmos de modelos de aprendizaje supervisado como los que se probaron, sin embargo, probar otros métodos como las Redes Neuronales podría ayudar a mejorar las métricas de desempeño y ajustar los datos a un buen modelo.

REFERENCES

[1] S. M. R. H. Shawon, M. A. Saaklayen, and X. Liang, “Wind Speed Forecasting by Conventional Statistical Methods and Machine Learning Techniques,” *2021 IEEE Electr. Power Energy Conf. EPEC 2021*, pp. 304–309, 2021, doi: 10.1109/EPEC52095.2021.9621686.

[2] K. L. Jorgensen and H. R. Shaker, “Wind Power Forecasting Using Machine Learning: State of the Art, Trends and Challenges,” *2020 8th Int. Conf. Smart Energy Grid Eng. SEGE 2020*, pp. 44–50, 2020, doi: 10.1109/SEGE49949.2020.9181870.

[3] M. E. K. Ali, M. Z. Hassan, A. B. M. S. Ali, and J. Kumar, “Prediction of Wind Speed Using Real Data: An Analysis of Statistical Machine Learning Techniques,” *Proc. - 2017 4th Asia-Pacific World Congr. Comput. Sci. Eng. APWC CSE 2017*, pp. 259–264, 2018, doi: 10.1109/APWConCSE.2017.00051.

[4] A. Alkesaiberi, F. Harrou, and Y. Sun, “Efficient Wind Power Prediction Using Machine Learning Methods: A Comparative Study,” *Energies*, vol. 15, no. 7, 2022, doi: 10.3390/en15072327.

[5] NREL, “Variables Climatológicas,” 2023. https://nsrdb.nrel.gov/data-viewer.