UNIVERSIDAD TECNOLÓGICA DE PANAMÁ

FACULTAD DE INGENIERÍA DE SISTEMAS COMPUTACIONALES

MAESTRIA EN ANALITICA DE DATOS

“Predicción de la Producción de Energía Eólica Basada en Condiciones Meteorológicas”

Presentado por:
Ing. Jesús González Castillo

Profesor
Juan Marcos Castillo, PhD

PROYECTO FINAL DE MODELOS PREDICTIVOS

2025

Introducción

La transición hacia fuentes de energía renovable ha impulsado el uso de la energía eólica como una alternativa sostenible y eficiente. A medida que la demanda de electricidad crece y la preocupación por el cambio climático aumenta, la generación de energía limpia se vuelve una prioridad global. La energía eólica, en particular, ha mostrado un gran potencial debido a su bajo impacto ambiental y su capacidad para complementar otras fuentes renovables como la solar.

Sin embargo, la variabilidad en la velocidad del viento y las condiciones meteorológicas afectan la capacidad de generación de las turbinas eólicas, lo que representa un desafío para la estabilidad de la red eléctrica. La predicción precisa de la producción de energía eólica es fundamental para optimizar su integración en los sistemas eléctricos, reducir costos operacionales y mejorar la eficiencia de la planificación energética.

En este contexto, la implementación de modelos predictivos basados en series de tiempo se presenta como una solución clave para anticipar la generación de energía y optimizar su gestión. Mediante el uso de algoritmos avanzados de aprendizaje automático y técnicas estadísticas, es posible desarrollar modelos que proporcionen estimaciones más precisas, permitiendo a las empresas eléctricas y operadores de redes adoptar estrategias más eficientes para la distribución y almacenamiento de la energía eólica.

Justificación

El desarrollo de métodos precisos para predecir la producción de energía eólica es esencial para optimizar su integración en la red eléctrica y garantizar un suministro confiable. La capacidad de anticipar la generación eólica reducirá la dependencia de fuentes convencionales contaminantes y mejorará la estabilidad de la red, evitando desbalances y costos adicionales. Asimismo, una gestión eficiente de la energía eólica mediante modelos predictivos avanzados permitirá disminuir las emisiones de carbono y facilitar la transición hacia un sistema energético sostenible. Además, la mejora en las predicciones beneficiará a los inversores y empresas del sector, optimizando la planificación de infraestructuras y la inversión en nuevas plantas eólicas, e impulsará la integración con otras fuentes renovables como la solar. La creciente digitalización del sector energético y el acceso a grandes volúmenes de datos en tiempo real abre nuevas oportunidades para aplicar técnicas de machine learning e inteligencia artificial en la predicción de la producción de energía eólica, mejorando la eficiencia del sector y acelerando su sostenibilidad.

Además, el acceso a datos en tiempo real provenientes de sistemas SCADA ha permitido la creación de modelos más robustos, incorporando variables como velocidad, dirección y temperatura del viento. Este estudio utilizará el conjunto de datos Wind Turbine SCADA,

que contiene variables clave como la velocidad y dirección del viento, la temperatura, la humedad y la potencia generada, con el fin de desarrollar modelos predictivos más precisos que optimicen la gestión de la energía eólica.

Definición del Problema

El problema principal radica en la alta variabilidad de la producción de energía eólica debido a factores meteorológicos y operacionales. Esta variabilidad dificulta la planificación y la estabilidad de la red eléctrica, ya que una mala predicción puede generar problemas de suministro o exceso de generación. Un desbalance en la oferta de energía puede ocasionar costos adicionales para los operadores de la red y afectar la confiabilidad del sistema eléctrico.

Además, la falta de predicciones precisas puede llevar a un sobredimensionamiento o subdimensionamiento en la infraestructura de almacenamiento y transmisión de energía. Un sobredimensionamiento podría resultar en inversiones innecesarias en infraestructura, mientras que un subdimensionamiento podría generar pérdidas de energía o ineficiencias en la distribución. Por lo tanto, es fundamental desarrollar modelos que permitan estimar de manera más confiable la generación eólica en distintos horizontes temporales, minimizando la incertidumbre y optimizando la toma de decisiones en la industria energética.

Análisis Predictivo

Determinación de la base de Datos:

La base de datos utilizada en este análisis se obtuvo del conjunto de datos 'Wind Turbine SCADA Dataset', disponible en Kaggle: https://www.kaggle.com/code/kerneler/starter-wind-turbine-scada-dataset-eaa6e30d-2.

Este conjunto de datos contiene mediciones de la producción de energía de un aerogenerador, específicamente del archivo 'T1.csv'. El archivo incluye las siguientes variables:

- **'Date/Time':** Fecha y hora de la medición (tipo de datos: fecha/hora).
- **'LV ActivePower (kW)':** Potencia activa real generada por el aerogenerador en kilovatios (tipo de datos: numérico). Esta es la variable objetivo de nuestro análisis.
- **'Wind Speed (m/s)':** Velocidad del viento en metros por segundo (tipo de datos: numérico).
- **'Theoretical_Power_Curve (KWh)':** Curva de potencia teórica generada por el aerogenerador en kilovatios (tipo de datos: numérico).
- **'Wind Direction (°)':** Dirección del viento en grados (tipo de datos: numérico).

	Date/Time	LV ActivePower (kW)	Wind Speed (m/s)	Theoretical_Power_Curve (KWh)	Wind Direction (°)
0	01 01 2018 00:00	380.047791	5.311336	416.328908	259.994904
1	01 01 2018 00:10	453.769196	5.672167	519.917511	268.641113
2	01 01 2018 00:20	306.376587	5.216037	390.900016	272.564789
3	01 01 2018 00:30	419.645905	5.659674	516.127569	271.258087
4	01 01 2018 00:40	380.650696	5.577941	491.702972	265.674286

El archivo 'T1.csv' contiene un total de 50530 filas de datos, con mediciones tomadas cada 10 minutos.

Se observó que los datos reales de potencia activa ('LV ActivePower (kW)') presentan un nivel significativo de ruido, lo cual es común en mediciones de este tipo. No se encontraron valores faltantes en el conjunto de datos. La variable 'Wind Direction (°)' fue excluida del análisis debido a su baja correlación con la variable objetivo y su limitada relevancia teórica para la predicción de la potencia activa.

Limpieza y preprocesamiento

Se eliminan valores atípicos y datos inconsistentes. Pasos a Realizar.

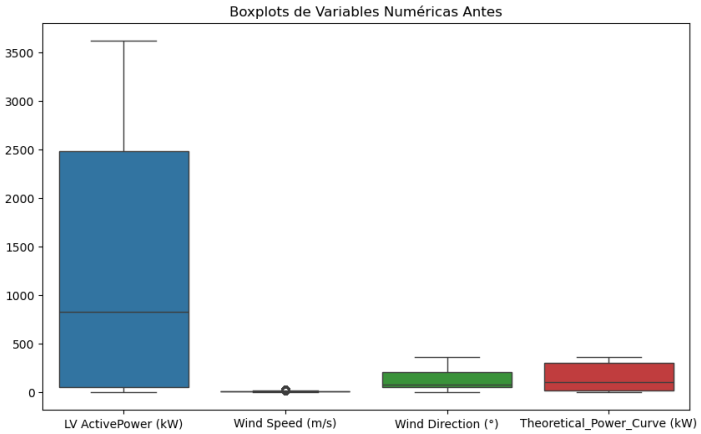
- **Manejo de valores nulos**
  - Identificar Valores Nulos en el dataset
  - Decidir cómo manejar los valores nulos (eliminar filas, imputar valores, etc.) se adjunta tabla con verificación que no hubo valores nulos.

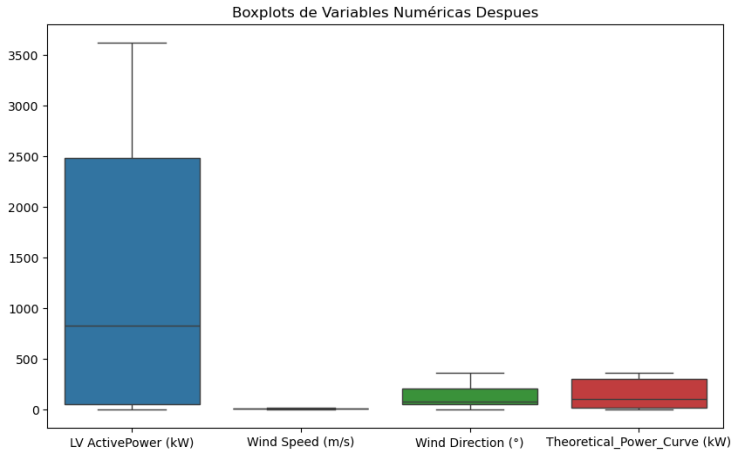
COLUMA	VAL NULOS
Date/Time	0
LV ActivePower (kW)	0
Wind Speed (m/s)	0
Theoretical_Power_Curve (KWh)	0
Wind Direction (°)	0

- **Conversión de la columna Date/Time**
  - Convertir la columna Date/Time a un formato de fecha y hora adecuado
  - Nuevo Formato ("%d %m %Y %H:%M")

- **Manejo de Unidades de Potencia**
  - Convertir la columna Theoretical_Power_Curve (KWh) a kilovatios (kW) para que sea comparable con LV ActivePower (kW).
  - Dividir los valores de Theoretical_Power_Curve (KWh) por 10 (ya que son mediciones cada 10 minutos).

- **Manejo de Valores Atípicos**
  - Identificar valores atípicos en las columnas numéricas.
  - Decidir cómo manejar los valores atípicos (eliminarlos, transformarlos, etc.).
  - Detección de outliers usando boxplots.
  - Verificar si las columnas numéricas son de tipo int54
  - Detectar y manejar los outliers usando el método IQR
  - Manejar los Outlier de la columna LV ActiverPower (Kw) por medio de Power Diference utilizando la columna Theoretical_power_Curve (kw)





- **Eliminación de columnas**
  - Se elimina la columna Power_Difference que no es relevante solo fue utilizada para el manejo de Outliers.

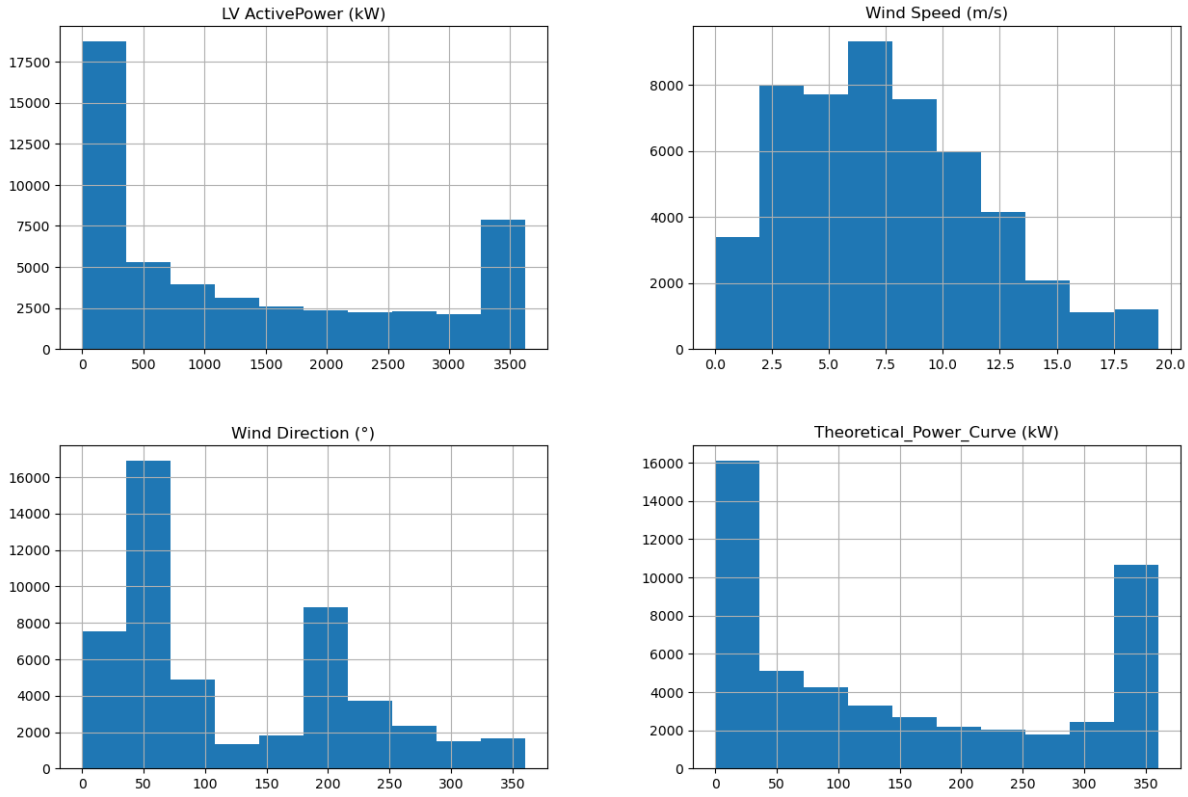
EDA (Análisis Exploratorio de Datos)

Se ha llevado a cabo un análisis descriptivo de las variables presentes en el conjunto de datos, identificando su distribución, tendencias y patrones.

- **Estadística Descriptiva de cada Columna**

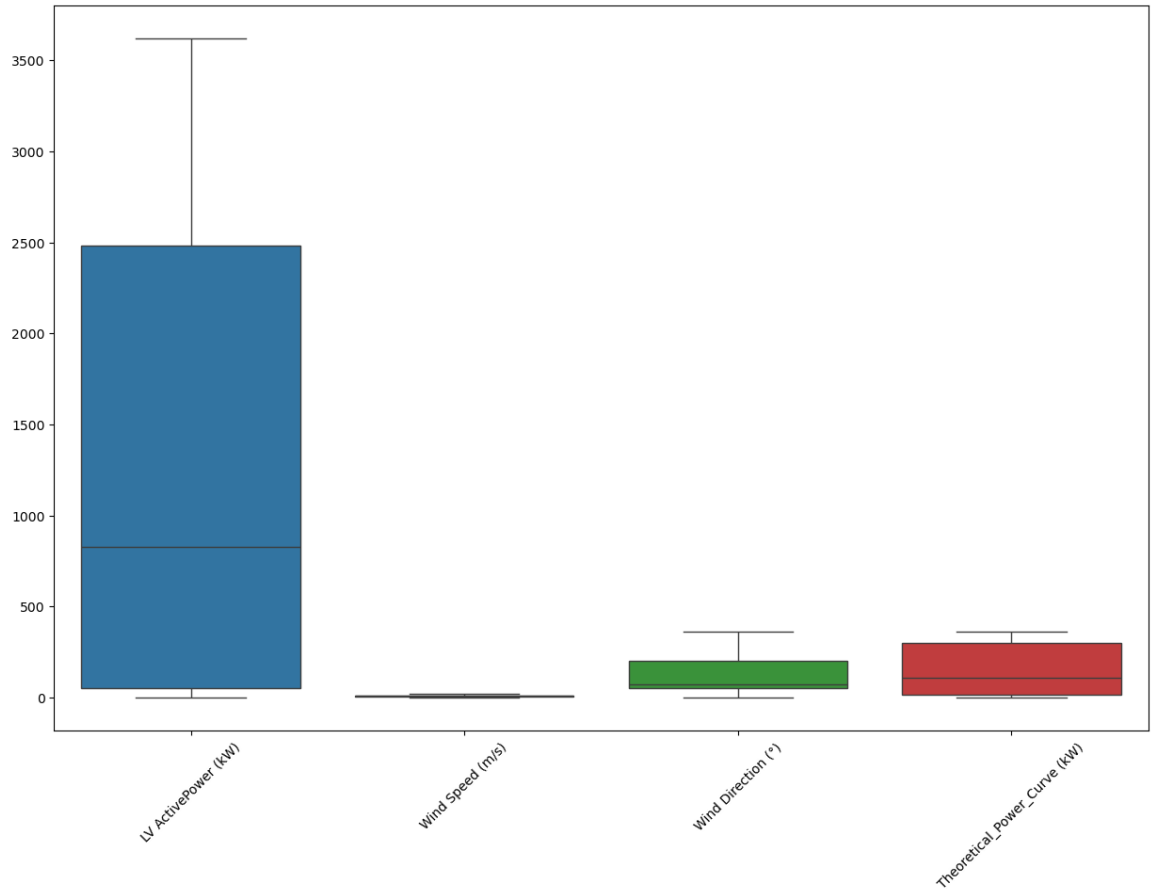
	LV ActivePower (kW)	Theoretical_Power_Curve (kW)	Wind Speed (m/s)	Wind Direction (°)
Cantidad	50530.000000	50530.000000	50530.000000	50530.000000
Promedio	1307.684332	149.217546	7.547758	123.687559
Desviación Estándar	1312.459242	136.801824	4.195567	93.443736
Mínimo	-2.471405	0.000000	0.000000	0.000000
25%	50.677890	16.132817	4.201395	49.315437
50%	825.838074	106.377628	7.104594	73.712978
75%	2482.507568	296.497246	10.300020	201.696720
Máximo	3618.732910	360.000000	19.447957	359.997589

- **Visualizaciones de Histogramas**



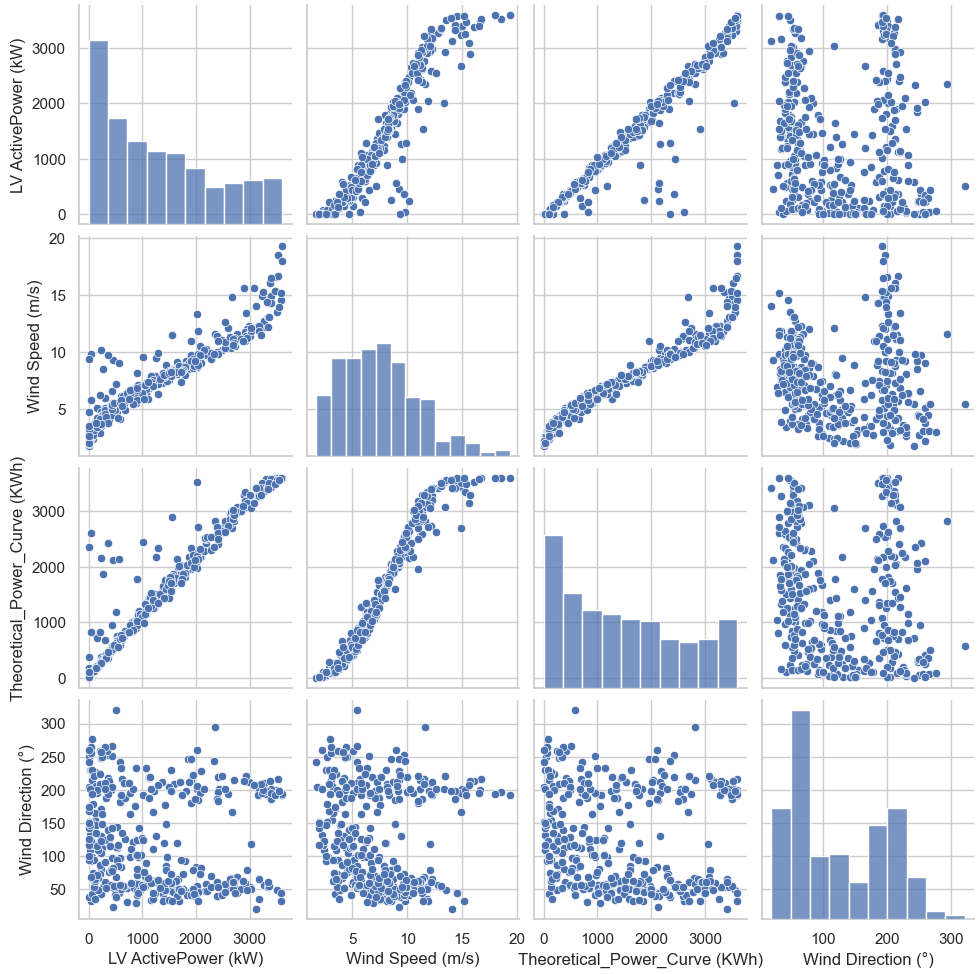
  - **LV ActivePower (kW):** Se observa una alta frecuencia de generación de baja potencia, con un pico significativo alrededor de los 3500 kW, lo que indica momentos de alta producción de energía. Esto sugiere que la generación de energía es variable, con períodos de baja y alta producción.
  - **Wind Speed (m/s):** La mayor concentración de datos se encuentra entre 5 y 10 m/s, lo que indica que esta es la velocidad del viento más común en el conjunto de datos. Esta información es crucial para entender la disponibilidad de viento y su impacto en la generación de energía.
  - **Wind Direction (°):** Se observan dos picos prominentes, uno cerca de los 100 grados y otro cerca de los 300 grados. Esto sugiere que el viento tiende a soplar con mayor frecuencia en estas dos direcciones, lo cual es relevante para la orientación óptima de las turbinas eólicas.
  - **Theoretical Power Curve (kW):** El gráfico muestra la distribución de la potencia que se debería generar en condiciones ideales, basada en las especificaciones del fabricante de las turbinas eólicas. Aunque debería reflejar una relación directa entre la velocidad del viento y la potencia, la similitud observada con el gráfico de la potencia activa real (LV ActivePower) sugiere que el parque eólico está operando de manera eficiente, con la generación real siguiendo de cerca la tendencia teórica. Las desviaciones entre ambas curvas, aunque esperadas debido a factores como la turbulencia y las condiciones climáticas, son mínimas, lo que indica un buen rendimiento del sistema

- **Graficos de Boxplot**

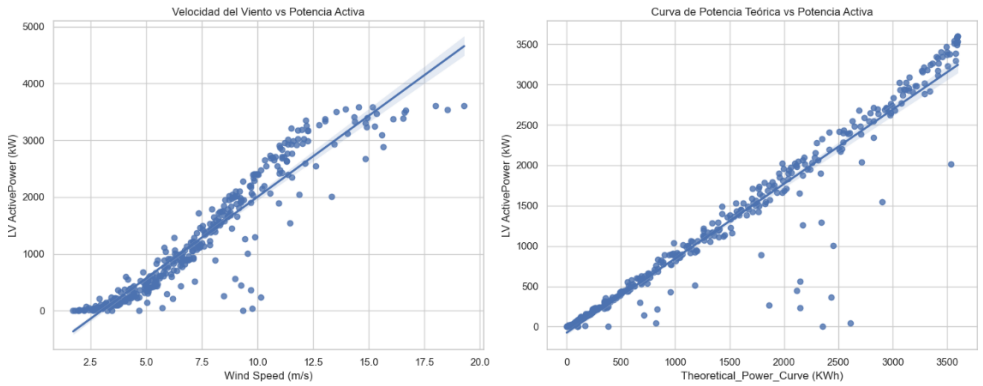


- **LV ActivePower (kW):** Este boxplot revela una distribución marcadamente asimétrica en la potencia activa generada, con una mediana baja que indica una predominancia de baja producción energética, mientras que la presencia de numerosos valores atípicos superiores señala la ocurrencia de picos significativos de alta generación, reflejando la variabilidad inherente en la producción de energía eólica.
- **Wind Speed (m/s):** La velocidad del viento muestra una distribución estrecha y concentrada, con una mediana baja y una caja pequeña, lo que sugiere que las velocidades de viento más frecuentes son bajas y que la variación en la velocidad del viento es limitada en este conjunto de datos.
- **Wind Direction (°):** La dirección del viento presenta una distribución más amplia, indicando una mayor variabilidad en la dirección del viento, y muestra una cantidad considerable de valores atípicos, lo que sugiere que, aunque hay direcciones predominantes, existen variaciones significativas.
- **Theoretical Power Curve (kW):** Este boxplot muestra una distribución similar a la de la potencia activa real, aunque con menos valores atípicos extremos, lo que refuerza la idea de que la generación real de energía sigue de cerca la tendencia de la curva de potencia teórica, y al igual que el LV Active power, muestra una cantidad considerable de valores atípicos.

- **Diagrama de Dispersión (Scatter Plots)**

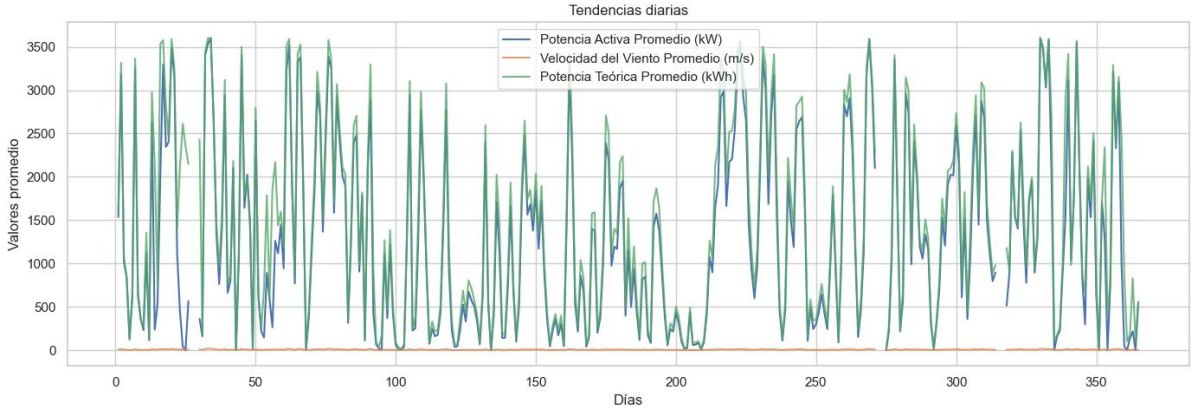


- **LV ActivePower (kW) vs. Wind Speed (m/s):** La relación entre la potencia activa generada y la velocidad del viento es claramente positiva y fuerte, mostrando que a medida que aumenta la velocidad del viento, la potencia generada también aumenta, aunque de manera no lineal.
- **LV ActivePower (kW) vs. Theoretical Power Curve (kW):** Se observa una correlación muy fuerte entre la potencia activa generada y la curva de potencia teórica, lo que indica que la generación real de energía sigue de cerca el modelo teórico, sugiriendo una operación eficiente del sistema.
- **Wind Speed (m/s) vs. Theoretical Power Curve (kW):** Esta relación ilustra la curva de potencia teórica en acción, demostrando cómo la potencia teórica aumenta con la velocidad del viento, siguiendo una curva característica que define el rendimiento ideal de la turbina.
- **Wind Direction (°) vs. Otras Variables:** Las relaciones entre la dirección del viento y las otras variables son menos directas, sin mostrar correlaciones fuertes aparentes, lo que sugiere que la dirección del viento, por sí sola, no determina directamente la potencia generada o la velocidad del viento, aunque puede influir en la eficiencia de la turbina.



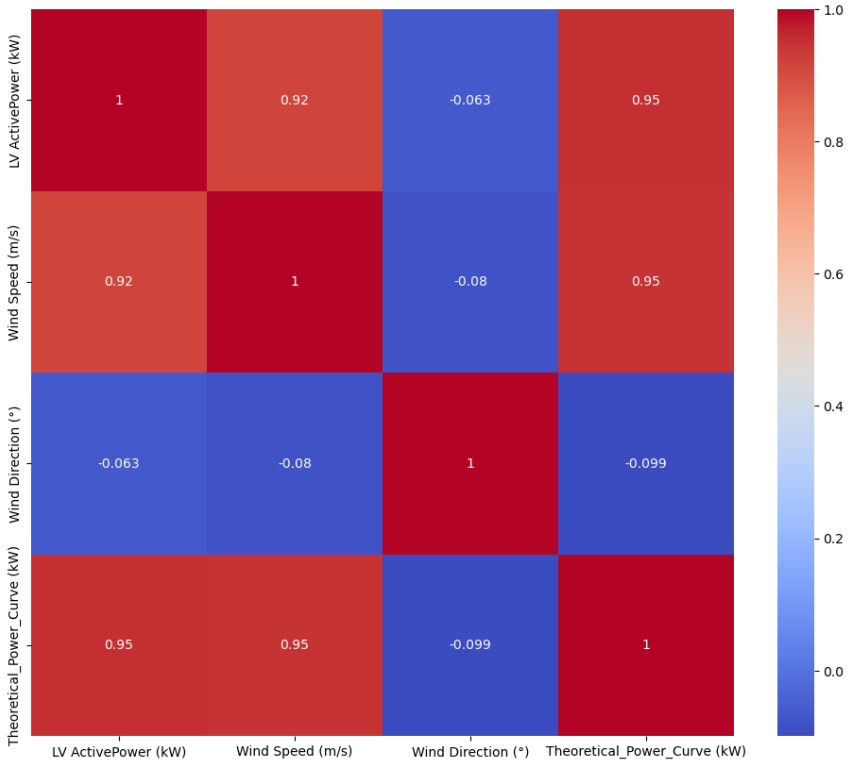
- **Velocidad del Viento vs Potencia Activa:** Este gráfico muestra una relación positiva y no lineal entre la velocidad del viento y la potencia activa generada, donde la potencia aumenta con la velocidad del viento, aunque con una zona de baja potencia para velocidades bajas y una meseta superior que indica un límite de producción, sugiriendo que la generación de energía eólica sigue un patrón predecible, pero con variaciones.
- **Curva de Potencia Teórica vs Potencia Activa:** Este gráfico revela una fuerte correlación positiva entre la curva de potencia teórica y la potencia activa real, indicando que la generación de energía del sistema se adhiere estrechamente al modelo teórico, lo que sugiere una alta eficiencia operativa, aunque con ligeras dispersiones debido a factores externos que afectan la producción real.

- **Tendencias Diarias**



El gráfico muestra las tendencias diarias de la potencia activa promedio (kW), la velocidad del viento promedio (m/s) y la potencia teórica promedio (kWh). El eje x representa la serie de tiempo. El eje y muestra los valores promedio. La línea azul indica la potencia activa promedio, la línea naranja representa la velocidad del viento promedio y la línea verde muestra la potencia teórica promedio. Podemos observar que la potencia activa y teórica fluctúan considerablemente a lo largo del año, generalmente siguiendo un patrón similar, mientras que la velocidad del viento se mantiene relativamente baja y estable durante la mayor parte del año con pequeños picos ocasionales.

- **Análisis de correlación**



1. **Correlación entre LV ActivePower (kW) y Wind Speed (m/s) → 0.92:** La potencia generada por la turbina está fuertemente correlacionada con la velocidad del viento. Este resultado confirma que un incremento en la velocidad del viento lleva a un aumento en la potencia generada. Dado que la correlación no es 1.0, también hay otros factores que afectan la potencia generada, como las condiciones del equipo y posibles pérdidas mecánicas.

2. **Correlación entre LV ActivePower (kW) y Theoretical_Power_Curve (kW) → 0.95:** La relación extremadamente alta indica que la potencia real generada sigue muy de cerca la curva teórica proporcionada por el fabricante. Tenido en cuenta que la correlación no es perfecta, lo cual podría indicar pérdidas de eficiencia, desgaste de la turbina o diferencias en condiciones ambientales reales versus las ideales del fabricante.

3. **Correlación entre Wind Speed (m/s) y Theoretical_Power_Curve (kW) → 0.95**: Este resultado confirma que la curva teórica de potencia está basada principalmente en la velocidad del viento. No sorprende que sean casi idénticos, ya que los valores teóricos de generación de energía se calculan directamente con la velocidad del viento.

4. **Correlación entre Wind Direction (°) y las demás variables → Baja correlación (-0.063 a -0.099):** A diferencia de las otras variables, la dirección del viento no parece afectar significativamente la potencia generada o la velocidad del viento. Esto sugiere que la turbina puede estar bien

diseñada para ajustarse a diferentes direcciones del viento sin afectar demasiado la producción. Si la turbina estuviera mal alineada con los vientos predominantes, se esperaría una correlación más fuerte.

Implicaciones para Modelos Predictivos

1. Variables clave para la predicción de potencia
    - Dado que la potencia generada (LV ActivePower) está altamente correlacionada con la velocidad del viento y la potencia teórica, estas variables deberían tener alta relevancia en un modelo predictivo.
    - Wind Direction parece ser menos útil para la predicción, aunque podría explorarse en modelos más complejos.

2. Posible multicolinealidad en modelos de regresión
    - Como la velocidad del viento y la potencia teórica están muy correlacionadas, incluir ambas en un modelo de regresión podría causar problemas de multicolinealidad.

3. Evaluación de eficiencia real vs. teórica
    - Como la potencia real no coincide exactamente con la teórica, se podría investigar si hay factores externos (temperatura, humedad, mantenimiento de la turbina) que expliquen la diferencia.

Selección de Variables

La selección de variables es un paso crucial en la construcción de modelos predictivos, ya que permite reducir la dimensionalidad del problema, mejorar la interpretabilidad del modelo y minimizar la presencia de ruido en los datos. Para determinar qué variables utilizar en el modelo de predicción de energía eólica, se realizaron varios análisis exploratorios y estadísticos.

Criterios para la selección de variables

Se emplearon los siguientes métodos para identificar las variables más relevantes:

- **Análisis de correlación:** Se evaluó la relación entre cada variable independiente y la variable objetivo (*LV ActivePower (kW)*) mediante el coeficiente de correlación.
- **Importancia de características:** Se entrenó un modelo preliminar basado en *Random Forest* para evaluar la importancia de cada variable en la predicción de la potencia generada.

Variables Seleccionadas

Tras este análisis, se seleccionaron las siguientes variables predictoras:

1. **Wind Speed (m/s):** Es la variable más influyente en la generación de energía, ya que la potencia producida por la turbina depende directamente de la velocidad del viento.

2. **Theoretical_Power_Curve (KWh):** Representa la potencia teórica estimada por el fabricante con base en la velocidad del viento. Esta variable permite comparar la eficiencia real de la turbina con la esperada.

Variable Objetivo

La variable a predecir es:

- **LV ActivePower (kW):** Representa la cantidad de energía generada por la turbina en cada intervalo de 10 minutos.

Variable No Seleccionada: **Wind Direction (°)**

Inicialmente se consideró la inclusión de la variable Wind Direction (°) debido a su posible influencia en la generación de energía, ya que las turbinas ajustan su orientación para maximizar la eficiencia. Sin embargo, tras realizar el análisis de correlación, se observó que Wind Direction (°) tenía una correlación muy baja con la variable objetivo, lo que sugiere que no aporta significativamente al modelo predictivo.

Selección del Modelo

Para la predicción de la generación de energía eólica, se evaluaron varios enfoques de modelado con el fin de determinar cuál ofrece la mejor precisión y generalización. La elección de los modelos se basó en su capacidad para manejar relaciones lineales y no lineales, su interpretabilidad y su desempeño en datos similares.

Modelos Evaluados

Se compararon los siguientes modelos:

1. Regresión Lineal

Este modelo se utilizó como referencia inicial debido a su simplicidad y capacidad para modelar relaciones lineales entre las variables predictoras y la variable objetivo. Su ecuación general es:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon$$

Donde:

- Y es la potencia generada (*LV ActivePower*).
- X_n son las variables predictoras seleccionadas.
- β_n son los coeficientes estimados del modelo.
- ε representa el término de error.

Dado que la relación entre la velocidad del viento y la potencia generada no es completamente lineal, este modelo tiene ciertas limitaciones en la predicción de energía eólica.

2. Random Forest Regressor

Para capturar relaciones no lineales y mejorar la precisión, se utilizó un modelo de *Random Forest*, un algoritmo de aprendizaje supervisado basado en ensamble de árboles de decisión.

Las principales ventajas de este modelo incluyen:

- Capacidad para modelar relaciones no lineales.
- Robustez frente a la multicolinealidad y valores atípicos.
- Reducción del sobreajuste mediante la agregación de múltiples árboles de decisión.

Evaluación del Desempeño de los Modelos

Para comparar el rendimiento de los modelos, se utilizaron las siguientes métricas:

- **Error Absoluto Medio (MAE):** Mide el error promedio absoluto entre los valores reales y predichos.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

- **Error Cuadrático Medio (MSE):** Penaliza los errores grandes al elevarlos al cuadrado.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- **Coeficiente de Determinación (R^2):** Indica qué porcentaje de la variabilidad en la variable objetivo es explicada por el modelo.

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

- **Raíz cuadrada** del **error cuadrático medio (RMSE)**: Es una medida que penaliza los errores grandes más que los pequeños debido al **cuadrado** de las diferencias.

$$MSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Resultados y Elección del Mejor Modelo

Para determinar el mejor modelo de predicción de la potencia activa generada por la turbina (**LV ActivePower (kW)**), se probaron varios algoritmos de regresión. Inicialmente, se entrenaron modelos de **Regresión Lineal** y **Random Forest**, obteniendo los siguientes resultados:

Modelo	**R^2**	**MAE**	**MSE**	**RMSE**
Regresión Lineal	0.90	171.86	162531.30	403.15
Random Forest	0.87	201.29	227851.13	477.38

Si bien la **Regresión Lineal** presentó un desempeño aceptable, el modelo **Random Forest** mostró una menor capacidad predictiva.

Para explorar más alternativas, se utilizó la librería de phyton **LazyPredict**, que permite evaluar múltiples algoritmos de Machine Learning de manera automatizada. El desempeño de los modelos se comportó de la siguiente manera:

- **Rendimiento de Alto Nivel (R-cuadrado ~ 0.91):**
  Modelos como GaussianProcessRegressor, GradientBoostingRegressor, HistGradientBoostingRegressor, XGBRegressor y LGBMRegressor demuestran un rendimiento excepcional. Estos modelos alcanzan altos valores de R-cuadrado (alrededor de 0.91), lo que indica un ajuste fuerte a los datos, y valores bajos de RMSE (alrededor de 400), lo que significa predicciones precisas. Son modelos poderosos que pueden capturar relaciones complejas en los datos.

- **Rendimiento Sólido (R-cuadrado ~ 0.90):**
  Un grupo significativo de modelos, incluidos MLPRegressor, SVR y diversos modelos lineales (como Lasso, Ridge y LinearRegression), muestran un buen rendimiento. Aunque ligeramente menos precisos que los modelos de alto nivel, aún proporcionan predicciones confiables. Los modelos lineales de este grupo suelen ofrecer ventajas en cuanto a interpretabilidad.

- **Rendimiento Moderado (R-cuadrado 0.82-0.89):**
  Modelos como RandomForestRegressor, BaggingRegressor, ExtraTreesRegressor, KNeighborsRegressor, HuberRegressor, RANSACRegressor, AdaBoostRegressor y DecisionTreeRegressor muestran una disminución considerable en el rendimiento. Estos modelos pueden ser menos adecuados si la precisión de las predicciones es crítica. Para los modelos basados en árboles, es posible que estén sobreajustados. KNeighborsRegressor es muy sensible a la escala de las características y a los valores atípicos.

- **Rendimiento Pobre (R-cuadrado < 0.80 o Negativo):**
  Modelos como TweedieRegressor, ElasticNetCV, DummyRegressor, KernelRidge y QuantileRegressor muestran un rendimiento pobre. El DummyRegressor sirve como una línea base, mostrando el rendimiento de un modeloque predice la media. Los valores negativos de R-cuadrado para KernelRidge y QuantileRegressor indican que estos modelos son muy inapropiados y rinden peor que simplemente predecir la media. ElasticNetCV es una regresión lineal regularizada, y su bajo rendimiento indica que un modelo lineal podría no ser la mejor opción para estos datos.

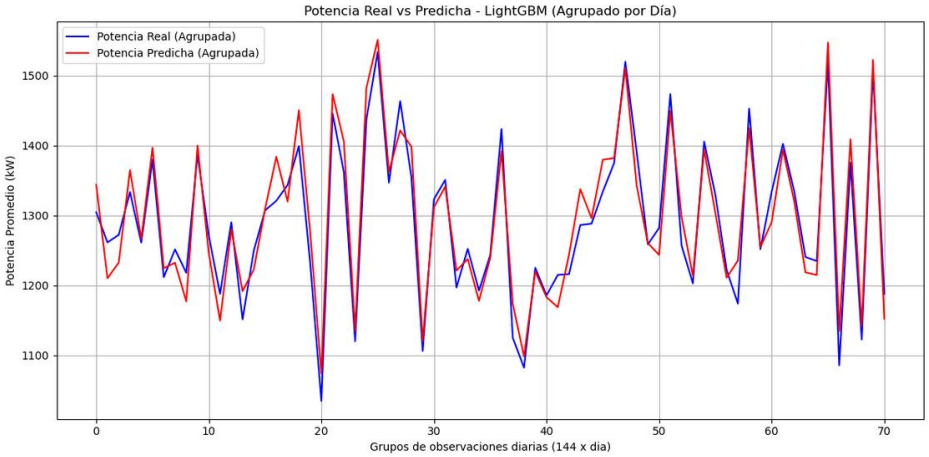
Model	R-Squared	Adjusted R-Squared	RMSE
GaussianProcessRegressor	0.91	0.91	399.04
GradientBoostingRegressor	0.91	0.91	399.93
HistGradientBoostingRegressor	0.91	0.91	400.08
XGBRegressor	0.91	0.91	400.33
LGBMRegressor	0.91	0.91	400.48
MLPRegressor	0.90	0.90	405.56
SVR	0.90	0.90	413.32
LassoLars	0.90	0.90	413.55
LassoCV	0.90	0.90	413.55
Lasso	0.90	0.90	413.55
LassoLarsIC	0.90	0.90	413.56
Lars	0.90	0.90	413.56
LassoLarsCV	0.90	0.90	413.56
LarsCV	0.90	0.90	413.56
LinearRegression	0.90	0.90	413.56
OrthogonalMatchingPursuitCV	0.90	0.90	413.56
TransformedTargetRegressor	0.90	0.90	413.56
RidgeCV	0.90	0.90	413.56
BayesianRidge	0.90	0.90	413.56
Ridge	0.90	0.90	413.56
NuSVR	0.90	0.90	413.56
SGDRegressor	0.90	0.90	414.09
OrthogonalMatchingPursuit	0.90	0.90	417.87
AdaBoostRegressor	0.90	0.90	418.02
LinearSVR	0.90	0.90	422.36
RANSACRegressor	0.89	0.89	424.35
PassiveAggressiveRegressor	0.89	0.89	426.13
HuberRegressor	0.89	0.89	426.19
KNeighborsRegressor	0.89	0.89	434.89
RandomForestRegressor	0.87	0.87	477.28
BaggingRegressor	0.86	0.86	486.56
ElasticNet	0.85	0.85	500.28
ExtraTreesRegressor	0.85	0.85	505.85
ExtraTreeRegressor	0.83	0.83	543.90
DecisionTreeRegressor	0.82	0.82	556.76
TweedieRegressor	0.79	0.79	602.50
ElasticNetCV	0.75	0.75	648.91
DummyRegressor	-0.00	-0.00	1306.39
KernelRidge	-0.10	-0.10	1369.99
QuantileRegressor	-0.13	-0.13	1385.94

Tabla: resultado del LazyPredict

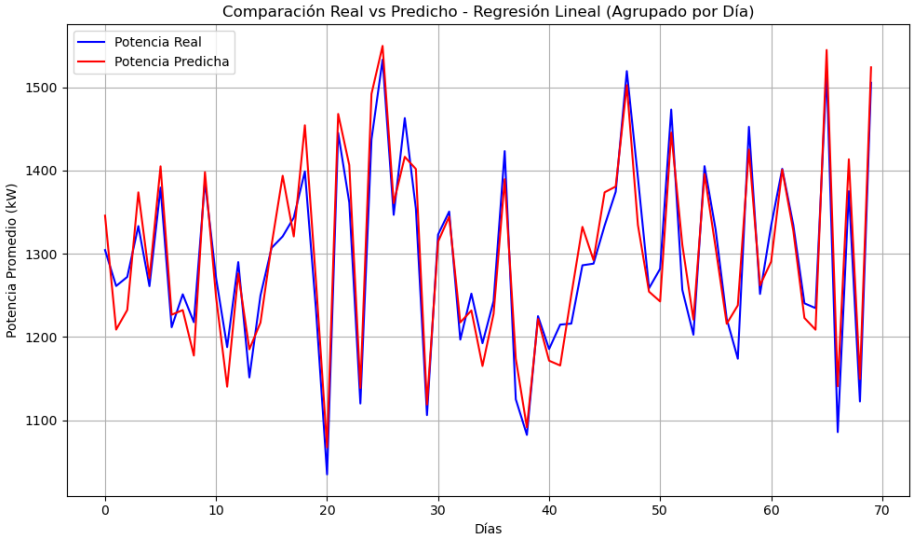
Análisis:

Modelo	R^2	MAE	MSE	RMSE
Regresión Lineal	0.90	171.86	162531.30	413.56
Random Forest	0.87	201.29	227851.13	477.28
LightGBM	0.91	166.33	159868.87	399.83

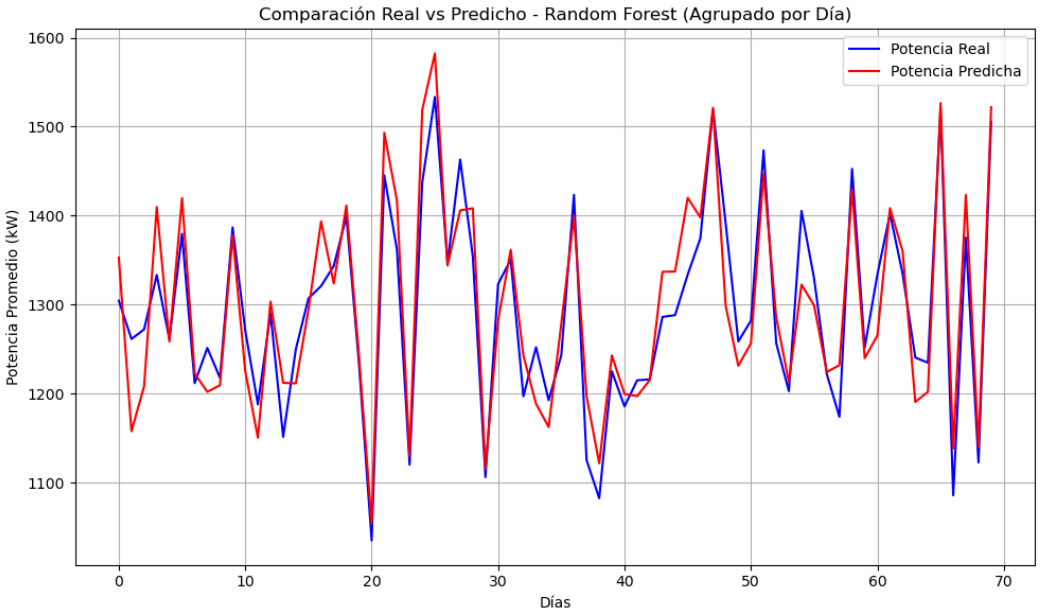
- **LightGBM** es claramente el modelo con mejor desempeño en general, ya que:
  - Tiene el **mayor R^2** (0.91), lo que indica que explica el 91% de la variabilidad en los datos.
  - Registra los **errores más bajos** en todas las métricas (MAE, MSE y RMSE), lo que indica mayor precisión.



- **Regresión Lineal** se comporta bastante bien, incluso supera a Random Forest en todas las métricas.
  - Esto sugiere que la relación entre las variables es mayormente lineal, y no tan compleja como para requerir modelos más complejos como Random Forest.



- **Random Forest** fue el modelo con peor desempeño relativo en este caso, con el menor R^2 y mayores errores. Puede deberse a:
    - Sobreajuste o subajuste.
    - Parámetros no optimizados.
    - Redundancia o ruido en las variables.



El modelo **LightGBM** es el más prometedor según las métricas, ya que tiene el mejor balance entre R^2 y error.

Aplicación de Machine Learning para la Predicción de la Potencia Eólica

Etapa ML	¿Dónde? (indica localización de la etapa en el código py)
1. Preparación del dataset	pd.read_csv(...)
2. Selección de variables (X, y)	X = df[['Wind_Speed', 'Theoretical_Power_Curve']]
3. División del conjunto (train/test)	train_test_split(...)
4. Selección de modelo ML	lgb_model = LGBMRegressor(...)
5. Entrenamiento del modelo	lgb_model.fit(...)
6. Predicción	y_pred_lgb = lgb_model.predict(...)
7. Evaluación del modelo	r2_score(...), mean_absolute_error(...), etc.
8. Exportación del modelo entrenado	joblib.dump(...)

1. Descripción del Problema

En este proyecto, el objetivo es predecir la potencia activa generada por una turbina eólica (LV ActivePower (kW)) utilizando variables predictoras como la velocidad del viento (Wind Speed (m/s)) y la curva de potencia teórica (Theoretical Power Curve (kW)). Esta predicción es fundamental para optimizar el uso de la energía eólica.

2. Preparación de los Datos

Antes de entrenar el modelo de Machine Learning, el conjunto de datos fue limpiado y preprocesado:

- Se eliminaron las columnas irrelevantes.
- Se manejaron valores faltantes y se realizaron transformaciones necesarias en los nombres de las columnas para facilitar su uso.
- Se seleccionaron las variables predictoras (Wind Speed (m/s) y Theoretical Power Curve (kW)) y la variable objetivo (LV ActivePower (kW)).

3. División de los Datos

El conjunto de datos fue dividido en dos subconjuntos:

- **Entrenamiento (80%)**: para entrenar el modelo.
- **Prueba (20%)**: para evaluar el rendimiento del modelo.

4. Selección del Modelo: LightGBM

El modelo seleccionado para este proyecto es **LightGBM (Light Gradient Boosting Machine)**, debido a su eficiencia y precisión en tareas de predicción numérica, especialmente con grandes conjuntos de datos.

Parámetros del modelo:

- **Número de estimadores**: 100 árboles
- **Tamaño de las hojas**: 31
- **Tasa de aprendizaje**: 0.05
- **Fracción de características**: 0.9
- **Objetivo**: Regresión

5. Entrenamiento y Evaluación del Modelo

El modelo fue entrenado utilizando el conjunto de entrenamiento y evaluado con el conjunto de prueba. Las métricas clave utilizadas para evaluar el rendimiento fueron:

- **MAE (Mean Absolute Error)**
- **MSE (Mean Squared Error)**
- **RMSE (Root Mean Squared Error)**
- **R^2 (Coeficiente de determinación)**

Resultados del Modelo LightGBM:

- **R^2**: 0.91
- **MAE**: 166.33
- **MSE**: 159,868.87
- **RMSE**: 400.48

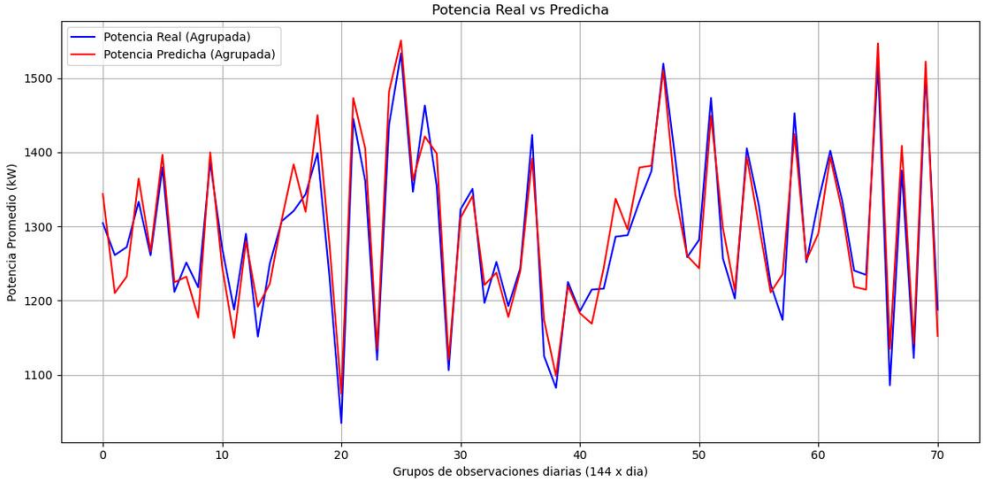
Estos resultados muestran que el modelo es muy preciso, con un **R^2** cercano a 1, lo que indica que el modelo puede explicar la mayor parte de la variabilidad de la potencia activa generada.

6. Visualización de Resultados

Se realizaron varias visualizaciones para entender mejor el rendimiento del modelo:

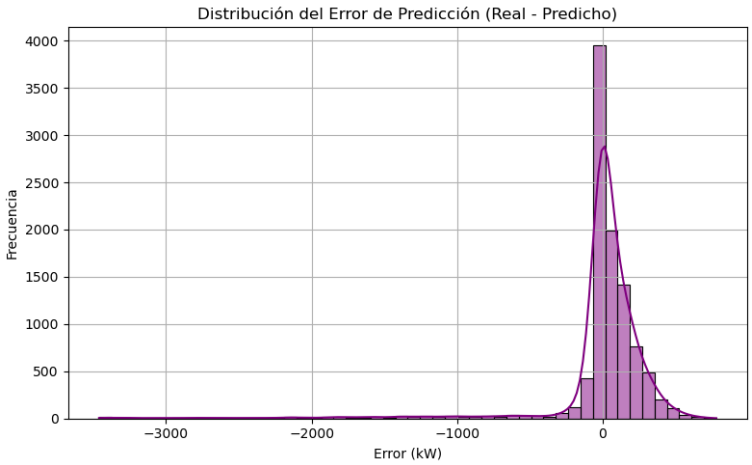
Comparación Real vs Predicho

Una gráfica de dispersión ilustra la relación entre la potencia real y la potencia predicha por el modelo. Para este gráfico, los valores se agruparon considerando que, según la definición del conjunto de datos, existe una observación cada 10 minutos, lo que equivale a 144 observaciones diarias. El modelo ajusta de manera precisa los valores observados, lo que evidencia su capacidad para capturar adecuadamente la variabilidad diaria de la potencia generada por la turbina. Las trayectorias de la línea azul (potencia real) y la línea roja (potencia predicha) siguen patrones muy similares a lo largo del tiempo, lo que sugiere un ajuste robusto. Esta interpretación visual se ve respaldada por las métricas obtenidas, donde LightGBM alcanza un R^2 de 0.91, un MAE de 166.33 kW y un RMSE de 399.83 kW, lo que refleja un rendimiento altamente confiable del modelo.



Distribución de los Errores

Se mostró una distribución de los errores, la cual indica que la mayoría de los errores de predicción son pequeños, lo cual es un buen indicador de que el modelo está funcionando correctamente.



7. Conclusiones del ML

El modelo LightGBM demuestra un excelente desempeño en la predicción de la potencia activa generada por las turbinas eólicas, con un R^2 de 0.91 y un error mínimo. Esto sugiere que **LightGBM** es un modelo adecuado para este tipo de predicción, y puede ser utilizado para optimizar el rendimiento de las turbinas eólicas en un entorno de producción.

8. Guardado del Modelo

El modelo entrenado ha sido guardado para futuras predicciones, entre los adjuntos se encuentra un archivo que explica como utilizar el modelo entrenado y generado.

Conclusiones

En este proyecto, se desarrolló un modelo predictivo para estimar la potencia activa generada por una turbina eólica en función de las variables de velocidad del viento y curva de potencia teórica. A través de un enfoque estructurado, se completaron varias etapas que permitieron alcanzar una solución efectiva para la predicción de la energía eólica.

1. **Preprocesamiento de Datos**: Se realizó una limpieza exhaustiva del conjunto de datos para eliminar valores faltantes y corregir posibles inconsistencias. Las columnas fueron adecuadamente nombradas para garantizar un fácil uso durante el modelado.
2. **Análisis Exploratorio**: Se ejecutaron análisis descriptivos, como la correlación entre las variables, para identificar las relaciones importantes. Las visualizaciones de dispersión y distribución mostraron que la velocidad del viento y la curva de potencia teórica son factores cruciales para predecir la potencia de la turbina.
3. **Modelado Predictivo con Machine Learning**: Se implementaron varios modelos de Machine Learning, siendo LightGBM el modelo seleccionado debido a su alta precisión y rendimiento. Este modelo logró un R^2 de 0.91, lo que indica una capacidad de predicción efectiva. Las métricas como el MAE (166.33) y RMSE (399.84) confirman que el modelo tiene un error mínimo y ajusta correctamente los valores observados.
4. **Evaluación y Comparación de Modelos**: Se evaluaron diferentes modelos, como Regresión Lineal y Random Forest, pero fue LightGBM el que presentó los mejores resultados en términos de precisión y capacidad de generalización.

En conclusión, este proyecto muestra cómo el uso adecuado de datos históricos, teoricos y técnicas de Machine Learning puede optimizar la predicción de la generación de energía eólica, contribuyendo a la toma de decisiones más informadas en la gestión de recursos energéticos.

Recomendaciones y Futuros Estudios

1. **Optimización y ajuste de hiperparámetros**: Aunque el modelo LightGBM ha demostrado ser efectivo, siempre hay espacio para la mejora. Se recomienda realizar un ajuste más exhaustivo de los hiperparámetros del modelo. Otras técnicas para optimizar los hiperparametros como

Grid Search o Randomized Search podrían ayudar a encontrar la combinación óptima como el número de estimadores, la profundidad de los árboles, la tasa de aprendizaje, y la fracción de características consideradas. Este proceso puede mejorar aún más el rendimiento predictivo y reducir el error.

2. **Exploración de más características o variables**: El modelo actual se basó únicamente en velocidad del viento y curva de potencia teórica para realizar las predicciones. Para mejorar la precisión del modelo, se recomienda ampliar las variables de entrada. Factores adicionales como temperatura**,** humedad y presión atmosférica podrían tener un impacto significativo en la potencia generada por las turbinas eólicas. Incluir estas variables permitiría al modelo capturar de manera más robusta las condiciones climáticas que afectan la generación de energía.

3. **Integración de modelos híbridos**: Otra recomendación interesante sería la creación de modelos híbridos que combinen múltiples técnicas, como la combinación de modelos de Machine Learning con modelos físicos de aerodinámica de turbinas. Esto permitiría aprovechar tanto el poder de predicción de los algoritmos de Machine Learning como las leyes físicas que rigen el comportamiento del viento y la conversión de la energía en las turbinas.

4. **Desarrollo de plataformas de predicción en tiempo real**: En el futuro, el modelo podría ser desplegado en plataformas de predicción en tiempo real para que los operadores de parques eólicos puedan ajustar su producción según las predicciones de energía. Esto sería útil para la optimización de la operación de las turbinas y la mejora de la gestión de la red eléctrica. La integración de predicciones a corto plazo con sistemas de control dinámico de las turbinas puede aumentar la eficiencia general de las plantas de energía eólica.

5. **Estudios sobre la eficiencia de las turbinas eólicas**: Finalmente, sería interesante expandir este estudio para evaluar la eficiencia de las turbinas eólicas mediante la comparación entre la potencia teórica y la potencia real generada. Esto podría ser útil para identificar oportunidades de mejora en el diseño de las turbinas o en su mantenimiento.

Bibliografía

1. **LightGBM Documentation** (2025). *LightGBM: Light Gradient Boosting Machine*. Enlace: https://lightgbm.readthedocs.io/en/latest/

2. **Python Software Foundation** (2025). *Python Programming Language*. Enlace: https://www.python.org/

3. **LightGBM GitHub Repository** (2025). *LightGBM: Light Gradient Boosting Machine*. Enlace: https://github.com/microsoft/LightGBM

4. **Scikit-learn Documentation** (2025). *Scikit-learn: Machine Learning in Python*. Enlace: https://scikit-learn.org/stable/

5. **Scikit-learn User Guide** (2025). *Supervised Learning, Model Evaluation, and Model Selection*. Enlace: https://scikit-learn.org/stable/user_guide.html

6. **LazyPredict Documentation** (2025). *LazyPredict: An easy-to-use Python library for quick and efficient model comparison*. Enlace: https://github.com/shankarpandala/lazypredict

Anexos

Enlace: https://github.com/JsGoz/UTP_MODELO_PREDICTIVO
En el enlace a GitHub se encontrará los siguientes archivos:

1. Base de datos
    - DataSetOriginal.csv
    - DataSetFinal.csv

2. Análisis descriptivo
    - Limpieza_Datos.ipnyb
    - EDA.ipnyb

3. Análisis Predictivo
    - Modelo_Predictivo.ipnyb

4. Resultados e Implementación del Modelo
    - MachineLearning.ipnyb
    - modelo_LightGBM.pkl

5. Documentación
    - ReporteFinal.pdf
    - PresentacionFinal.ppt
    - Como_uso_el_.pkl.pdf