TÍTULO

CURVAS DE POTENCIA EÓLICA
MODELADO Y EVALUACIÓN DEL RENDIMIENTO PREDICTIVO

AUTOR

Javier Moreno Arispón

Esta edición electrónica ha sido realizada en 2024

Tutor	Dr. D. Antonio Jesús Sánchez Fuentes
Institución	Universidad Internacional de Andalucía
Curso	*Máster de Formación Permanente en Big Data (2022/23)*
©	Javier Moreno Arispón
©	De esta edición: Universidad Internacional de Andalucía
Fecha documento	2023

Proyecto Fin de Máster

Máster en Big data

Curvas de Potencia Eólica: Modelado y Evaluación del Rendimiento Predictivo

Autor: Javier Moreno Arispón

Tutor: Antonio Jesús Sánchez Fuentes

Universidad Internacional de Andalucía
UNIA

Sevilla, 2023

Proyecto Fin de Máster
Máster en Big Data

Curvas de Potencia Eólica: Modelado y Evaluación del Rendimiento Predictivo

Autor:
Javier Moreno Arispón

Tutor:
Antonio Jesús Sánchez Fuentes

Universidad Internacional de Andalucía
UNIA
Sevilla, 2023

Proyecto Fin de Carrera: Curvas de Potencia Eólica: Modelado y Evaluación del Rendimiento Predictivo

Autor: Javier Moreno Arispón

Tutor: Antonio Jesús Sánchez Fuentes

El tribunal nombrado para juzgar el Proyecto arriba indicado, compuesto por los siguientes miembros:

Presidente:

Vocales:

Secretario:

Acuerdan otorgarle la calificación de:

Sevilla, 2023

El Secretario del Tribunal

A mi familia

A Lucía

Agradecimientos

En este viaje, no puedo evitar sentir una profunda gratitud hacia aquellos que han estado a mi lado. A mi familia y amigos, quienes han sido mi refugio constante; a Lucía, quien me ha aguantado con paciencia y amor inquebrantable; y a mi tutor, por brindarme la oportunidad y embarcarse en estos momentos finales conmigo. Este proyecto es el resultado final de algo que marcará el inicio de nuevas aventuras y logros, y les agradezco de corazón por ser parte esencial de esta emocionante y desafiante travesía.

Javier Moreno Arispón

Sevilla, 2023

Resumen

En la industria de la energía eólica, la curva de potencia es un elemento esencial que vincula la velocidad del viento a la potencia activa de los aerogeneradores. Esta relación es fundamental en diversas aplicaciones cruciales, como la selección de aerogeneradores, la estimación del factor de capacidad, el pronóstico de la energía eólica y el seguimiento del rendimiento de los parques eólicos. Para lograr una implementación efectiva de estas aplicaciones, es esencial contar con un modelo de curva de potencia preciso que se ajuste a las condiciones específicas de cada parque eólico

En este proyecto, se propone el desarrollo de un algoritmo que aborde varios aspectos críticos en el procesamiento y modelado de datos relacionados con la energía eólica. Este algoritmo llevará a cabo un análisis exhaustivo de los datos, seleccionará los inputs óptimos, detectará valores atípicos, los procesará y aplicará técnicas de filtrado. A continuación, utilizará diversas técnicas de aprendizaje automático para modelar la curva de potencia. Además, categorizará las predicciones resultantes en tres niveles de rendimiento: alto, medio y bajo. Esto permitirá a los usuarios anticiparse a posibles problemas y aprovechar oportunidades de mejora en la operación de los aerogeneradores.

La experimentación se ha llevado a cabo con un conjunto de datos anónimos, proporcionados de manera confidencial por un propietario de sistemas SCADA. Este conjunto de datos comprende 29 variables diferentes recopiladas a lo largo de 9 meses. Los datos corresponden a aerogeneradores del modelo AE-52 de 850 kW de potencia, fabricados por la empresa MADE.

Los resultados obtenidos en este proyecto indican que el modelado de la curva de potencia desarrollado podría resultar altamente beneficioso para los usuarios cuando se implemente en entornos de producción. El algoritmo proporciona una aproximación precisa de la curva de potencia, y las predicciones alcanzan niveles de acierto superiores al 95% en la categoría de "Alto Rendimiento". Esto demuestra su utilidad en la monitorización y gestión eficiente de parques eólicos.

Abstract

In the wind energy industry, the power curve is an essential element that connects wind speed to the active power of wind turbines. This relationship is crucial in various critical applications, such as wind turbine selection, capacity factor estimation, wind energy forecasting, and wind farm performance monitoring. To achieve effective implementation of these applications, it is essential to have an accurate power curve model that aligns with the specific conditions of each wind farm.

In this project, the development of an algorithm is proposed to address several critical aspects in the processing and modeling of wind energy-related data. This algorithm will conduct a comprehensive data analysis, select optimal inputs, detect outliers, process and filter them, and then apply machine learning techniques to model the power curve. Furthermore, it will categorize the resulting predictions into three performance levels: high, medium, and low. This will enable users to anticipate potential issues and leverage improvement opportunities in wind turbine operation.

The experimentation has been conducted using anonymous data provided confidentially by a SCADA system owner. This dataset comprises 29 different variables collected over a period of 9 months. The data corresponds to wind turbines of the AE-52 model with a capacity of 850 kW, manufactured by MADE.

The results obtained in this project indicate that the developed power curve modeling could be highly beneficial for users when implemented in production environments. The algorithm provides an accurate approximation of the power curve, and the predictions achieve accuracy levels exceeding 95% in the "High Performance" category. This demonstrates its utility in the monitoring and efficient management of wind farms.

Índice

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

1 PRELIMINARES

Ya no queda una piedra en pie porque el viento lo derribó, ya no queda nada de ayer, porque el viento se lo llevó

- Robe -

En la última década, la tecnología de energía eólica ha demostrado ser una fuente de energía renovable prometedora. Su crecimiento constante se debe a su competitividad económica y su respeto por el medio ambiente. No obstante, la variabilidad en la velocidad del viento plantea desafíos importantes en la gestión de la energía, lo que a su vez impacta la confiabilidad de la red eléctrica. Para afrontar este problema, se ha centrado la atención en las curvas de potencia de los aerogeneradores, que muestran la relación entre la velocidad del viento y la potencia generada. Estas curvas, que son no lineales, resultan cruciales para lograr una integración eficiente de la energía eólica en el sistema eléctrico y para supervisar el estado de los aerogeneradores. [5]

El presento proyecto se dedica a enfrentar los desafíos asociados con la monitorización de aerogeneradores y la estimación precisa de las curvas de potencia eólica. El objetivo de esta investigación es contribuir al desarrollo de soluciones que mejoren la confiabilidad y la eficiencia de la generación de energía eólica, aspecto fundamental para un futuro más sostenible en la producción de energía. [1]

1.1. Curva de potencia

La curva de potencia de un aerogenerador representa la potencia eléctrica producida por la turbina frente a las diferentes "velocidades de viento" a la altura del buje que pueden presentarse. Una aproximación general de esta "relación no lineal" podría expresarse como [7]:

$$P = \frac{1}{2}\rho A C_p(\lambda, \beta) v^3 \qquad (1)$$

donde:

- ρ representa la densidad del aire.
- A es el diámetro del rotor y determina la denominada "área de barrido", que es la superficie virtual que dibuja el rotor perpendicularmente al flujo del viento.
- C_p denota el coefiente de potencia, indica con qué eficiencia el aerogenerador convierte la energía del viento en electricidad. Este coeficiente es función de la relación entre la "velocidad punta" λ y el "ángulo de paso" de las palas β.
- υ representa la velocidad del viento a la altura del buje

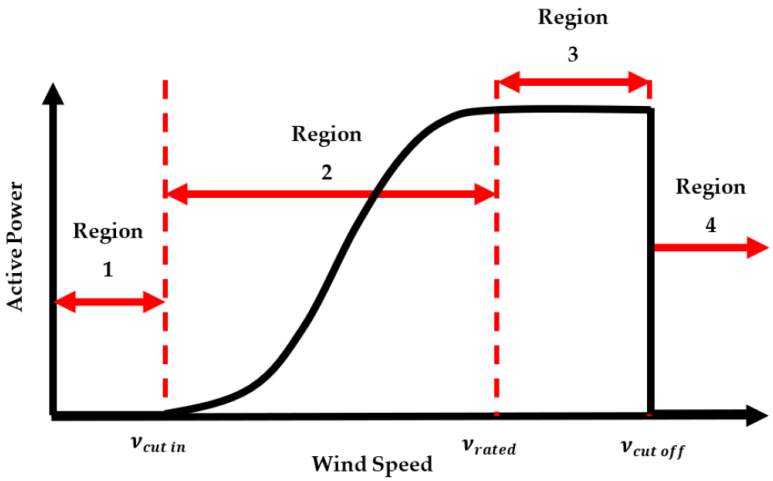

Figura 1. Curva de potencia teórica e intervalos de operación

En el gráfico de arriba (1) se presenta la curva de potencia teórica de un aerogenerador, que se detallará en el siguiente apartado. Además, se ilustran los diferentes rangos de funcionamiento en función de las "velocidades del viento". A continuación, se ofrece una breve explicación de lo que cada uno de estos intervalos representa:

1. A una velocidad de viento considerablemente baja, el aerogenerador no producirá energía; dicho punto de inflexión en la curva donde la turbina empieza a producir se dará a la "velocidad de arranque" o "conexión" (), que generalmente se sitúa alrededor de los 3-4 m/s.
2. A partir de la υ_{cut-in} el aerogenerador comenzará a generar energía de manera gradual hasta alcanzar la velocidad "nominal" o "υ_{rated}", velocidad a la cual el aerogenerador alcanzará su "Potencia nominal".
3. Por encima de la "velocidad nominal" el aerogenerador generará potencia a un ritmo constante, P_{rated} (Potencia nominal), esta será la potencia máxima de la turbina.
4. Una vez alcanzada la "velocidad" de corte del aerogenerador, $\upsilon_{cutt-off}$, se desconectará como medida preventiva, ya que velocidades superiores podrían suponer un riesgo para su integridad estructural. Por lo general, los aerogeneradores cuentan con mecanismos de autoregulación que los detienen a velocidades de viento cercanas a los 25 m/s.

1.2. Curva de potencia teórica

Principalmente, la curva de potencia es certificada por el fabricante, y el proceso de certificación está establecido por la CEI, “Comisión Electrotécnica internacional” [6]. El llamado procedimiento se trata de un proceso experimental en el que se mide simultáneamente la velocidad del viento y la potencia de salida, en un entorno experimental a condiciones meteorológicas y topográficas ideales.

La curva de potencia teórica se utiliza comúnmente en una amplia variedad de aplicaciones, como la "predicción de la energía eólica", que es el foco de estudio en este proyecto, así como en la "selección del modelo óptimo de aerogenerador". Además, se emplea como punto de "referencia" en los procesos de "caracterización del rendimiento de la turbina" para evaluar el impacto del deterioro del aerogenerador a lo largo de su ciclo de vida.

Es importante destacar que cada fabricante de aerogeneradores desarrolla diferentes modelos, y para cada modelo, se define su propia curva de potencia teórica. Entre los fabricantes más destacados en la industria se encuentran "Vestas", "Siemens Gamesa", "Goldwind" o "GE".

En el presente proyecto, se empleará un conjunto de datos que incluye información relevante de la turbina eólica MADE AE-52 de 850 kW. Además de los datos recopilados, se dispondrá de la correspondiente curva teórica de potencia asociada a esta turbina. La turbina MADE AE-52 se caracteriza por tener una capacidad de generación de 850 kilovatios, lo que la sitúa dentro de la categoría de turbinas eólicas de tamaño medio [22].

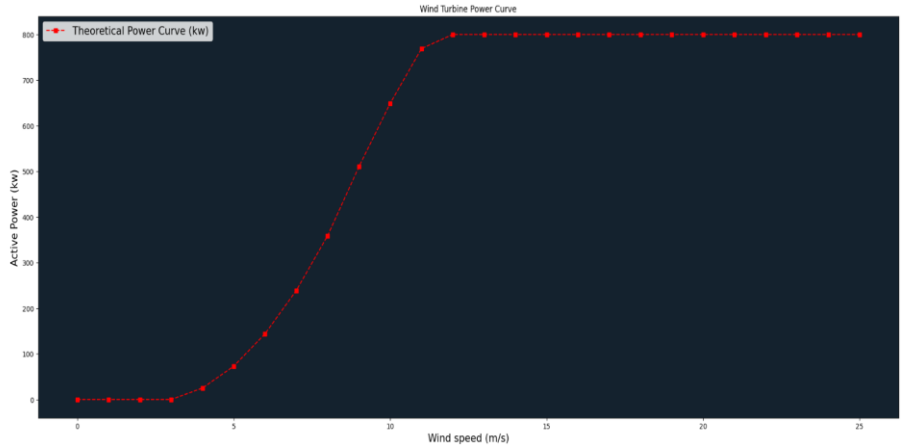

Figura 2. Curva de potencia teórica de una turbina MADE AE-52, 800kW

Esta elección es de gran importancia, ya que permitirá llevar a cabo un análisis detallado del rendimiento de esta turbina específica en condiciones reales de operación. La disponibilidad de la curva teórica de potencia proporciona una referencia fundamental para comparar y evaluar la eficiencia de la turbina en función de los datos recopilados en el campo. Este enfoque permitirá identificar posibles desviaciones entre el rendimiento real y el previsto, así como detectar anomalías o fallos que puedan afectar su operatividad.

1.3. Curva de potencia real

Adicionalmente, señalar que “curva de potencia teórica" proporcionada por el fabricante y calculada según el protocolo establecido por la CEI no toma en consideración el impacto del "desgaste" o "deterioro" del aerogenerador. Esto conduce a una discrepancia con la "curva de potencia real", obtenida a partir de los datos del sistema SCADA al analizarlas conjuntamente. La principal causa de esta discrepancia puede atribuirse a las fluctuaciones en los valores de la curva de potencia para velocidades del viento idénticas. Por lo tanto, la aplicación directa del enfoque convencional en situaciones como la "detección de anomalías y fallos" podría no ser apropiada. En su lugar, podría ser necesario adoptar un enfoque más refinado.

En el siguiente gráfico, se presenta el conjunto de datos de una turbina eólica real, que será utilizado en este estudio, consta de 29 variables y abarca un período de casi 10 meses. A partir de este conjunto de datos ya se pueden inferir las características que debería tener la curva de potencia real del aerogenerador seleccionado para este estudio.

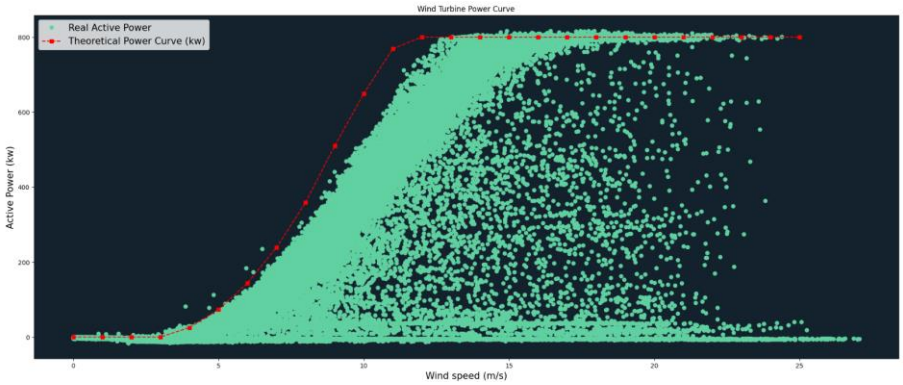

Figura 3. MADE AE-52, 800kW. Curva de potencia teórica vs real

2 APLICACIONES DE LA CURVA DE POTENCIA

Que algo no funcione como tú esperabas no quiere decir que sea inútil.

- Thomas Alva Edison -

Los modelos que se fundamentan en curvas de potencia precisas resultan sumamente valiosos en diversas aplicaciones dentro de la industria de la energía eólica. Estas aplicaciones abarcan principalmente la selección de aerogeneradores, la estimación del factor de capacidad, la evaluación y pronóstico de la energía eólica, así como la supervisión del estado de los sistemas. A continuación, se presenta una breve revisión de estas aplicaciones.

2.1 Selección de aerogenerador

Durante la fase de desarrollo de un parque eólico específico, la elección de las características de los aerogeneradores en los que invertir se convierte en una tarea crítica, especialmente en las primeras etapas del proyecto. La curva de potencia de un aerogenerador puede desempeñar un papel fundamental al comparar diferentes tipos de aerogeneradores y, en última instancia, en la selección del aerogenerador más adecuado entre las opciones disponibles [4]. Para garantizar la optimización de la eficiencia en los sistemas de parques eólicos, es esencial llevar a cabo una evaluación exhaustiva y seleccionar con precisión las propiedades de los aerogeneradores que se ajusten al régimen de vientos del emplazamiento en cuestión. Esto implica considerar varios aspectos, como el tamaño del aerogenerador, la compatibilidad con el lugar elegido, el historial de disponibilidad y confiabilidad, así como las garantías ofrecidas, entre otros factores. En términos generales, el tamaño del aerogenerador está estrechamente relacionado con la cantidad de energía que se debe generar. Sin embargo, es importante tener en cuenta dos limitaciones principales:

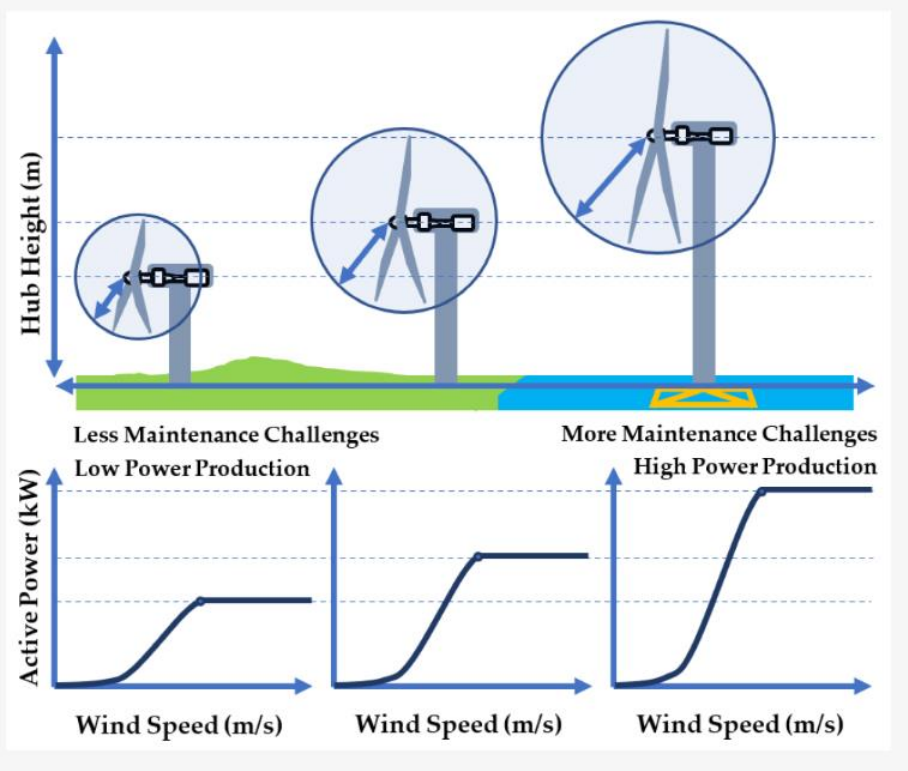

Figura 4. Altura del “hub” de la turbina vs Desafíos en el mantenimiento

1. Por un lado, es innegable que los aerogeneradores de mayores dimensiones y mayor altura tienen la capacidad de generar una cantidad superior de energía en comparación con sus contrapartes más pequeñas y de menor altura. No obstante, estos aerogeneradores más grandes pueden conllevar costos adicionales y retrasos en el mantenimiento, como, por ejemplo, la sustitución de componentes clave. Uno de los desafíos más destacados en este aspecto radica en la falta de infraestructura adecuada para elevar cargas pesadas hasta la cima de estas torres elevadas.
2. Por otro lado, los aerogeneradores de menor tamaño pueden parecer más fáciles de mantener; sin embargo, podrían generar ingresos de producción más reducidos debido a sus torres de menor altura o a una eficiencia general inferior.

Existen otras limitaciones que pueden incidir en la elección entre estos compromisos. Por ejemplo, la disponibilidad de superficie para el desarrollo del proyecto de parques eólicos y la ubicación seleccionada, ya sea en tierra o en alta mar, son factores críticos. Los proyectos en alta mar a menudo requieren el uso de aerogeneradores más grandes, mientras que los parques eólicos en tierra pueden enfrentar limitaciones en cuanto al espacio disponible, lo que, a su vez, restringe las opciones disponibles.

2.2 Estimación del factor de capacidad

El factor de capacidad de una turbina eólica, FC, se define como la relación entre la potencia promedio de salida de la turbina durante un período de tiempo y su potencia nominal.

En estos análisis, la "velocidad del viento" suele estimarse mediante el uso de la distribución de Weibull. Esta información, que requiere la curva de potencia del aerogenerador, es fundamental para diversos propósitos,

como la "optimización de dimensionamiento y costos", la "adecuación óptima turbina-emplazamiento" y la "evaluación de posibles ubicaciones", entre otros [2].

2.3 Evaluación y predicción de energía eólica

La energía eólica está experimentando un sólido crecimiento tanto en Europa como en todo el mundo. A pesar de ello, los costos de operación y mantenimiento siguen representando aproximadamente el 25-30% del "costo nivelado de la electricidad" (LCOE) en aerogeneradores terrestres y marinos. En consecuencia, numerosos operadores buscan reducir aún más estos costos y maximizar el tiempo de actividad de sus parques eólicos.

Para lograrlo, los parques eólicos de vanguardia se someten a una monitorización continua y remota desde un centro de control, lo que contribuye a disminuir los costos de mantenimiento al identificar de manera temprana fallos operativos y daños en desarrollo. Esto, a su vez, permite tomar decisiones informadas y responder de manera eficiente. La automatización en la monitorización de las condiciones, que incluye diagnósticos y pronósticos, se convierte en un elemento fundamental para implementar estrategias de mantenimiento basado en condiciones altamente efectivas.

Los aerogeneradores más modernos cuentan con una extensa red de sensores que supervisan tanto sus subsistemas internos como las condiciones ambientales circundantes. Estos avanzados sensores son capaces de recolectar una impresionante cantidad de datos, alcanzando cientos de gigabytes diarios para cada una de las múltiples variables y características que se evalúan. Esta información abarca desde la producción de energía hasta variables de estado térmico y electromecánico, respuestas de vibración, calidad del aceite y condiciones ambientales.

Además de estos sistemas de sensores especializados, los datos obtenidos a través del sistema de "supervisión y adquisición de datos en tiempo real" (SCADA, por sus siglas en inglés) se encuentran generalmente disponibles para enriquecer el proceso de monitoreo. Se ha propuesto que estos datos SCADA [9] puedan utilizarse como alternativas económicas a los sistemas de sensores específicos para la detección de fallos. Estos registros suelen incluir información crucial, como la generación activa de energía, la velocidad del rotor, temperaturas de componentes y condiciones ambientales, registradas en intervalos de 10 minutos o incluso minutales.

3 ANOMALÍAS Y FALLOS

Sólo si eres una anomalía del sistema puedes llegar a cambiarlo.

- Fernando de la Rosa -

En términos generales, las series temporales que se obtienen a partir de los datos del SCADA de un parque eólico suelen contener un significativo número de datos anómalos, principalmente derivados de fallos o irregularidades tanto en las turbinas como en el propio sistema SCADA. Aunque existen múltiples factores que pueden contribuir a desviaciones en la curva de potencia, este estudio se centrará en aquellos que tienen un impacto más significativo en las condiciones normales de operación del aerogenerador. En la mayoría de los casos, estas anomalías pueden atribuirse a la degradación o a un mal funcionamiento de los sensores de la turbina, errores en los sistemas y dispositivos de comunicación, condiciones ambientales adversas o deficiencias en el diseño de la turbina, así como a posibles fallos de fábrica, como desalineamientos (ángulo "yaw"), problemas en la orientación de las palas ("pitch system") o cuestiones aerodinámicas relacionadas con las palas de los aerogeneradores [8].

3.1 Principales anomalías y fallos más recurrentes

En base a la literatura científica revisada, y de acuerdo a la experiencia en el sector, las anomalías y fallos más recurrentes [3] en un aerogenerador a tener en consideración son:

- Agrupación de valores inferiores.
- Reducción de potencia o “Power Curtailment”.
- Dato de “Velocidad del viento” congelado.
- Dispersión de datos.
- Formación de hielo o residuos en las palas

3.1.1 Agrupación de valores inferiores

En situaciones donde existe un recurso eólico adecuado para que una turbina genere energía, a veces se producen periodos de inactividad, caracterizados por la presencia de valores de potencia prácticamente nulos. Estos intervalos, que podríamos denominar "valores horizontales", indican una subutilización de la turbina a pesar de las condiciones eólicas favorables.

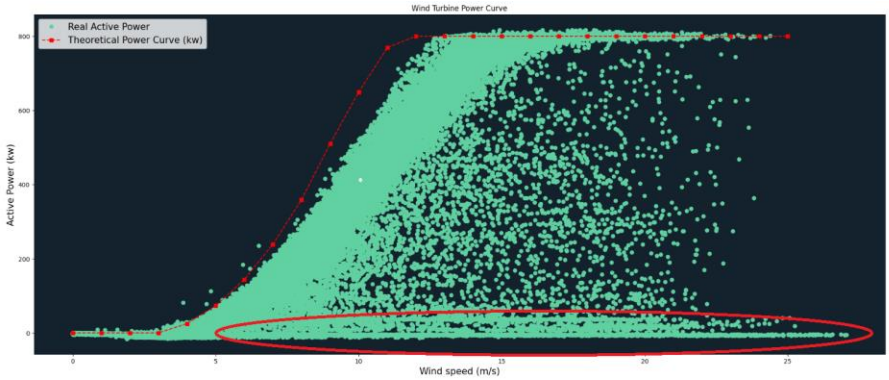

Figura 5 Tipo 1: Agrupación de valores inferiores

Causa raíz

- Instrumento de medición de potencia dañado.
- Fallo en el flujo de comunicaciones de los SCADAs que componen el Sistema.
- Activación de alarmas que resultan en la indisponibilidad de la turbina y, como consecuencia, su estado operativo cambia de "Marcha/Listo".

Solución propuesta

- Seleccionar exclusivamente los datos correspondientes a situaciones en las que la turbina exhiba un CS_StatusCat de 100 (se explicará en más detalle en capítulos posteriores), que representa su estado operativo mientras está en funcionamiento.
- Para la detección y manejo de valores atípicos, se aplicarán técnicas de eliminación de outliers, como cuantiles, el método Z-Score o, en este proyecto en particular, se implementará el algoritmo Isolation Forest.

3.1.2 Limitación de potencia de potencia

También conocido como “Power Curtailment” o “Derated”, estos sucesos son fácilmente identificables ya que la potencia activa permanece constante frente a fluctuaciones en la velocidad del viento.

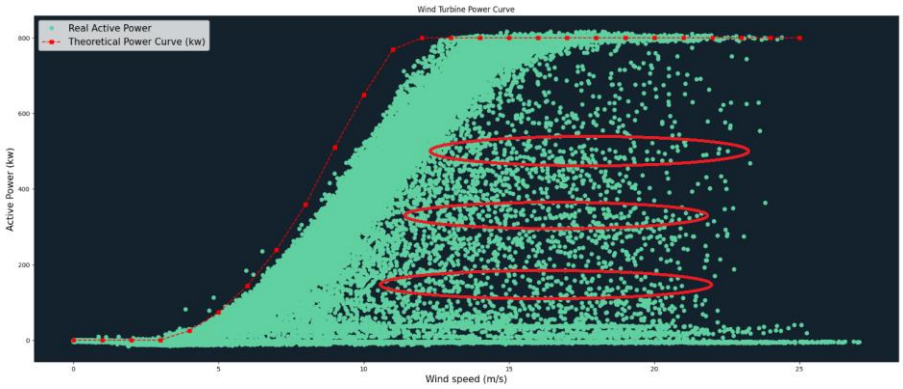

Figura 6 Tipo 2: Reducción de potencia

Causa raíz

- La disminución de la potencia se origina principalmente debido a "acciones de control" deliberadas. En otras palabras, se emiten instrucciones para limitar la generación de energía del aerogenerador a un nivel inferior al máximo posible. Por ejemplo, en el caso de la península ibérica, la mayoría de estas instrucciones son gestionadas por la Red Eléctrica Española (REE) con el propósito de asegurar la confiabilidad y calidad del suministro eléctrico en la red de transporte de alta tensión sin incidentes.
- Otra causa potencial de reducción de potencia radica en fallas en los sensores de carga.

Solución propuesta

- Cuando la turbina es regulada, el valor de CS_StatusCat cambia a un estado diferente de 100, que corresponde al estado de "Regulación". Por lo tanto, al filtrar nuevamente utilizando esta variable, se pueden identificar y eliminar los valores atípicos asociados al "Curtailment".
- En situaciones en las que no se cuente con esta variable, se puede emplear una estrategia alternativa dividiendo la curva en intervalos de, por ejemplo, 10 kW y luego aplicar técnicas de cuantiles a cada intervalo. De hecho, esta técnica se utiliza en este estudio.

3.1.3 Dato de “Velocidad del viento” congelado

En este caso los datos se apilarán verticalmente porque los datos que se están leyendo están congelados (no refrescan). En el gráfico se observarán aumentos de producción para el mismo recurso eólico. A continuación, se modificará el dataset para mostrar la forma que debería tener la curva de potencia para este escenario:

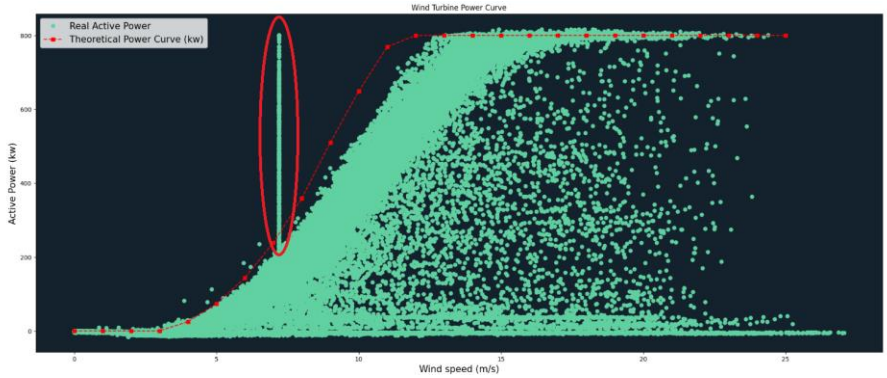

Figura 7. Tipo 3: Velocidad de viento congelada

Tal y como se indicará más adelante, la curva de potencia elegida parece no haber tenido incidencias con el anemométro ni cortes de comunicación relevantes, y por ello está libre de este tipo de anomalías.

Existen una gran variedad de razones que podrían desencadenar esta casuística, sin embargo, las más comunes son:

Causa raíz

- Fallo en los sensores del anemómetro o calibración ineficiente.
- Fallo en los instrumentos de medición de potencia.
- Fallo en los equipos de comunicación. Un corte en el flujo de comunicaciones.

Solución propuesta

- Este tipo de valor atípico puede ser complicado de abordar. Si la velocidad del viento parece estar congelada, esto podría indicar un problema con el anemómetro de la turbina o, en su ausencia, con la torre meteorológica en el emplazamiento (si se estuvieran utilizando sus datos). En estos casos, que difieren de una pérdida total de comunicación (ya que los demás datos son precisos), una posible solución sería descartar los datos que presenten esta anomalía. Otra opción, aunque más arriesgada debido a las particularidades geográficas de cada emplazamiento, sería reemplazar los datos de velocidad del viento utilizando información de aerogeneradores en la misma ubicación.
- En contraste, cuando se trata de una falla total de comunicación en el flujo de datos entre los sistemas SCADA, la variable CS_StatusCat asumiría el valor 107 y, adicionalmente, todos los datos se verían congelados o serían nulos. En este escenario, el proceso de filtrado sería considerablemente más sencillo y directo.

3.1.4 Dispersión de datos

En esta problemática se recogen todos aquellos datos distribuidos a lo largo de la curva aleatoriamente

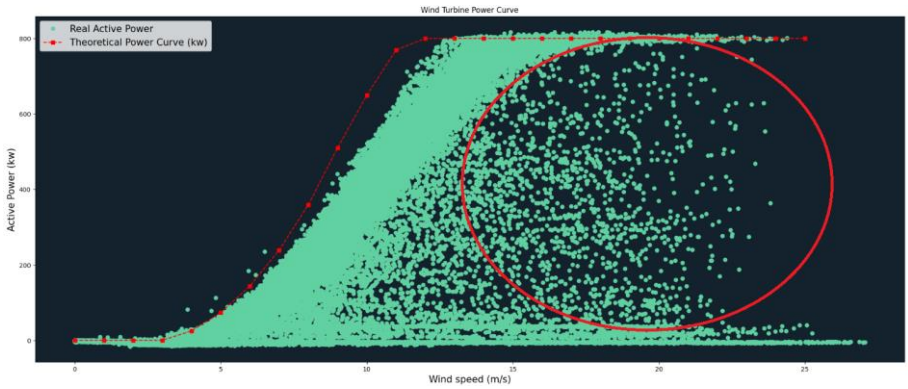

Figura 8. Tipo 3: Dispersión de datos

Causa raíz

- Perturbaciones causadas por el "efecto estela" [22] o por la "rampa de carga del aerogenerador", que se manifiestan durante el encendido o apagado del aerogenerador.
- Deterioro en los instrumentos de medición.
- Interferencia en las comunicaciones durante la transmisión de señales.

Solución propuesta

- La cuantificación de este tipo de errores resulta bastante desafiante, por lo que quizás sea más adecuado recurrir a las técnicas mencionadas anteriormente para tratar los valores atípicos. En este sentido, se aplicarán métodos de eliminación de outliers, como el enfoque basado en cuantiles, el método Z-Score o el algoritmo Isolation Forest, una de las técnicas más usadas en ML para afrontar estos desafíos.

3.1.5 Acumulación de hielo o residuos en las palas

Frecuentemente, se puede observar un desplazamiento proporcional de los puntos a lo largo de la curva, lo cual suele indicar eventos como la formación de hielo en las palas o la acumulación de residuos en las mismas.

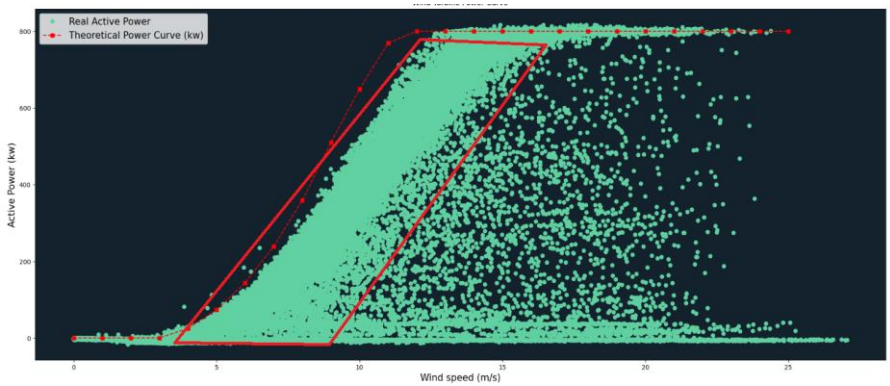

Figura 9. Tipo 5: Formación de hielo o residuos en las palas

Causa raíz

- Estos fenómenos están principalmente condicionados por factores ambientales y la ubicación geográfica específica del parque eólico.
- Problemas en el funcionamiento de los sensores.
- Mantenimiento inadecuado, incluyendo reparaciones insuficientes, falta de lubricación y limpieza insuficiente de las palas.

Solución propuesta

- Z-score, cuantiles o Isolation Forest, entre otras técnicas.

4 ANÁLISIS, PROCESAMIENTO Y FILTRADO DE DATOS

Una de las cosas mas agradables de la vida: ver como se filtra el sol entre las hojas

-Mario Benedetti-

El procesamiento y filtrado de datos en la curva de potencia de una turbina eólica implica la manipulación y refinamiento de los datos recopilados para obtener una representación precisa del rendimiento de la turbina. Esto es esencial para comprender y modelar la relación entre la velocidad del viento y la potencia de salida de la turbina. A continuación, se describen algunos métodos clave utilizados en este proceso.

4.1 Fases del análisis y procesamiento de datos

Fase 1. Identificación de "Inputs" Relevantes y Preparación de Datos

1. **Identificación de Inputs Relevantes:** Antes de abordar el modelado de la curva de potencia, es crucial realizar un análisis inicial para determinar cuáles son los inputs más relevantes. Esto se puede lograr mediante técnicas como la matriz de correlación, que ayudarán a identificar las variables más influyentes en la generación de energía de la turbina.
2. **Fase de Limpieza de Datos:**

- Identificación de Outliers: La primera etapa consiste en identificar y eliminar los valores atípicos en los datos. Estos valores anómalos pueden surgir debido a condiciones inusuales, fallos en la turbina o problemas en los sensores. Detectar y tratar los outliers contribuye a garantizar que los datos se ajusten mejor a la curva de potencia ideal.
- Filtrado de Anomalías: Además de los outliers, los datos pueden contener anomalías que también deben ser tratadas. Por ejemplo, cuando la velocidad del viento está por encima o por debajo de los límites de velocidad de corte, la potencia debe ser cero. Si se registra una potencia no nula en tales condiciones, se considera una anomalía que debe corregirse.

3. **Corrección de Datos Faltantes:** En esta etapa, se abordan los datos faltantes en el conjunto histórico recopilado por el SCADA. Estas lagunas pueden surgir debido a problemas en los sensores o a interrupciones en la comunicación en el flujo de datos. Los datos faltantes se estiman o interpolan para asegurar que la serie temporal de datos esté completa y coherente.

Fase 2. Modelado de la curva de potencia

El modelado de la curva de potencia implica la elección de técnicas apropiadas (ML, Deep learning,etc), la consideración de premisas clave y la evaluación constante del rendimiento del modelo. Un enfoque sólido en esta etapa es fundamental para optimizar la operación y el rendimiento de las turbinas eólicas en tiempo real y en el futuro. En el próximo capítulo se explicará con más detalle esta fase.

Fase 3. Resultados, conclusiones, análisis de mejora y trabajos a futuro

La Fase 3 abarca los resultados, conclusiones, mejoras y futuros trabajos. En esta etapa, se presentan los resultados obtenidos, se extraen conclusiones, se analiza cómo mejorar el proyecto y se identifican áreas para futuras investigaciones y desarrollos. Esta fase se dividirá en 2 capítulos distintos.

4.2 Análisis de datos

La generación eficiente de energía eólica depende en gran medida de la precisión de la curva de potencia de un aerogenerador. Antes de modelar esta curva, es esencial realizar un análisis de datos exhaustivo.

El análisis de datos permite comprender la naturaleza de los datos, identificar valores atípicos y seleccionar las características más relevantes. Además, ayuda a evaluar la calidad de los datos y a elegir el enfoque de modelado adecuado. Las visualizaciones de datos, como gráficos de dispersión y correlación, desempeñan un papel clave al revelar patrones y relaciones entre las variables.

El análisis de datos, por tanto, es una fase fundamental en la modelación de la curva de potencia en la energía eólica, garantizando resultados precisos y confiables para la generación de energía sostenible.

4.2.1 Exploración y selección de los datos de entrada

La selección de variables adecuadas desempeña un papel crítico en el modelado de una curva de potencia de una turbina eólica. Identificar los inputs más relevantes y descartar los menos significativos es esencial para construir modelos precisos que ayuden a comprender y predecir el rendimiento de la turbina. En este contexto, se exploran diversas técnicas y enfoques para seleccionar los inputs óptimos que impulsarán la calidad y la eficacia del modelo de curva de potencia. Este proceso de exploración y selección de datos de entrada es fundamental para garantizar que el análisis resultante sea valioso y confiable, ya que no todas las variables pueden contribuir de manera significativa al modelado de la curva de potencia. A continuación, se presentan algunas de las técnicas más apropiadas para llevar a cabo esta selección de inputs de manera efectiva y eficiente

4.2.1.1 Análisis de correlación

En primer lugar, se diseñará una matriz de correlación adecuada para el conjunto de datos. Esto es fundamental en el modelado de una curva de potencia, ya que permite identificar las relaciones entre las variables predictoras y la variable objetivo (en este caso, la potencia). La matriz proporciona información sobre la fuerza y dirección de estas relaciones, lo que resulta esencial para la selección de las variables más relevantes en el modelado y para comprender su impacto en la potencia de la turbina eólica. Este proceso orienta la toma de decisiones y facilita la construcción de modelos más precisos.

En el notebook de Colab se ha desarrollado un código que calcula y muestra gráficamente la correlación entre diferentes variables numéricas en un conjunto de datos. Se enfoca en encontrar la relación de estas variables con una variable específica llamada "potencia" (ActivePower) y visualiza las correlaciones utilizando un gráfico de barras, lo que facilita la comprensión de cómo las variables predictoras se relacionan con la variable objetivo.

Para desarrollar este notebook, se ha empleado una amplia gama de bibliotecas de Python e información que exponen comunidades expertas en el sector, cada biblioteca y técnicas está adaptada a un campo de aplicación específico. Todas las subfunciones empleadas en este estudio se encuentran documentadas en los sitios web de las respectivas bibliotecas y webs utilizadas [10-21].

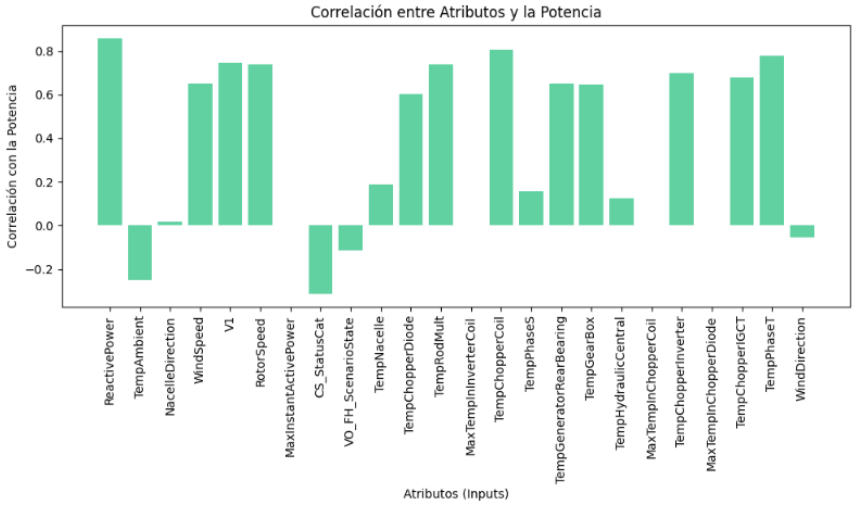

Figura 10. Matriz de correlación, “ActivePower” variable objetivo

Para clarificar los resultados obtenidos será de vital importancia comentar:

- **Altura de las barras**: La altura de cada barra representa la magnitud de la correlación entre una variable predictora y la potencia. Barras más altas indican una correlación más fuerte.
- **Dirección de la correlación**: La dirección de la correlación se determina por la orientación de la barra. Las barras hacia arriba indican una correlación positiva (aumento de la variable predictora, aumento de la potencia), mientras que las barras hacia abajo representan una correlación negativa (aumento de la variable predictora, disminución de la potencia).

Este gráfico es útil para identificar las relaciones más significativas entre las variables predictoras y la potencia, lo que puede ser valioso en el análisis de datos y la toma de decisiones.

Selección de “inupts”

Para identificar las variables con un mayor impacto en las predicciones, se emplea un algoritmo que define un umbral de correlación. Este algoritmo selecciona las variables que exhiben una correlación relevante con la variable objetivo "ActivePower". Además, se brinda al usuario la oportunidad de descartar variables adicionales antes de mostrar las variables definitivamente elegidas.

```
Variables seleccionadas para entrenar los datos en el modelado de la curva de potencia con técnicas de ML:
1. ReactivePower
2. WindSpeed
3. V1
4. RotorSpeed
5. TempChopperDiode
6. TempRodMult
7. TempChopperCoil
8. TempGeneratorRearBearing
9. TempGearBox
10. TempChopperInverter
11. TempChopperIGCT
12. TempPhaseT
Ingrese el número de la variable que desea descartar (0 para finalizar): 1
ReactivePower ha sido marcada para descartar.
Ingrese el número de la variable que desea descartar (0 para finalizar): 3
V1 ha sido marcada para descartar.
Ingrese el número de la variable que desea descartar (0 para finalizar): 0
Variables seleccionadas:
1. WindSpeed
2. RotorSpeed
3. TempChopperDiode
4. TempRodMult
5. TempChopperCoil
6. TempGeneratorRearBearing
7. TempGearBox
8. TempChopperInverter
9. TempChopperIGCT
10. TempPhaseT
```

Figura 11. Matriz de correlación. Selección de variables con umbral y Control de Usuario

4.2.1.2 Experiencia en el sector

La experiencia en el dominio de la energía eólica y la física detrás del funcionamiento de las turbinas pueden proporcionar información valiosa sobre qué características son más relevantes. A menudo, los expertos en el campo pueden identificar características críticas de manera más efectiva. Por ello se implementa un apartado extra al código donde se le dará la opción al usuario de introducir variables que considera cruciales en el entrenamiento, como la WindDirection o la NacelleDirection.

```
¿Desea agregar variables adicionales? (s/n): s
Variables disponibles para agregar:
1. LocalTimestamp
2. Path
3. SiteName
4. Timestamp
5. ActivePower
6. ReactivePower
7. TempAmbient
8. NacelleDirection
9. V1
10. MaxInstantActivePower
11. CS_StatusCat
12. VO_FH_ScenarioState
13. TempNacelle
14. MaxTempInInverterCoil
15. TempPhaseS
16. TempHydraulicCentral
17. MaxTempInChopperCoil
18. MaxTempInChopperDiode
19. WindDirection
Ingrese el número de la variable adicional que desea agregar (0 para finalizar): 8
NacelleDirection ha sido agregada.
```

Figura 12. Selección de Variables Basada en Experiencia

4.2.1.3 Selección basada en modelos de ML

Se podría entrenar un modelo de aprendizaje automático con todas las características disponibles y luego utilizar técnicas de selección de características específicas del modelo, como coeficientes de regresión o puntuaciones de importancia, para identificar las características más relevantes.

Uno de los métodos más comunes en esta categoría es el uso de modelos de regresión lineal, como por ejemplo una "Regresión Lineal Múltiple". Este modelo se utiliza para predecir una variable de salida continua (en este caso, "ActivePower") basándose en múltiples variables de entrada (las columnas seleccionadas por el usuario). Es una técnica ampliamente utilizada en estadísticas y aprendizaje automático debido a su simplicidad y capacidad para modelar relaciones lineales entre las variables de entrada y la variable objetivo.

Sin embargo, también es importante destacar que la regresión lineal tiene limitaciones, ya que asume una relación lineal entre las variables de entrada y la variable objetivo, lo que puede no ser válido en todos los casos.

Por lo tanto, se recomienda evaluar diferentes modelos y técnicas si la relación entre las variables no es lineal o si se busca un rendimiento más avanzado.

Regresión Lineal

En el algoritmo las características se enumeran en orden descendente de importancia (Ver figura 13). Las características en la parte superior tienen el mayor impacto en el modelo, mientras que las de la parte inferior tienen menos importancia.

```
Variables disponibles para incluir en el modelo:
1. LocalTimestamp
2. Path
3. SiteName
4. Timestamp
5. ActivePower
6. ReactivePower
7. TempAmbient
8. NacelleDirection
9. WindSpeed
10. V1
11. RotorSpeed
12. MaxInstantActivePower
13. CS_StatusCat
14. VO_FH_ScenarioState
15. TempNacelle
16. TempChopperDiode
17. TempRodMult
18. MaxTempInInverterCoil
19. TempChopperCoil
20. TempPhaseS
21. TempGeneratorRearBearing
22. TempGearBox
23. TempHydraulicCentral
24. MaxTempInChopperCoil
25. TempChopperInverter
26. MaxTempInChopperDiode
27. TempChopperIGCT
28. TempPhaseT
29. WindDirection
```

```
Importancia de las características:
                     Feature    Importance
18           TempChopperIGCT  3.807604e+01
4                 RotorSpeed  1.319526e+01
10           TempChopperCoil  3.426669e+00
16       TempChopperInverter  3.329124e+00
11                TempPhaseS  2.043710e+00
3                         V1  1.709674e+00
2                  WindSpeed  1.070596e+00
12  TempGeneratorRearBearing  4.903628e-01
14      TempHydraulicCentral  4.158784e-01
1           NacelleDirection  2.347431e-01
5       MaxInstantActivePower  2.020606e-14
9       MaxTempInInverterCoil  7.105427e-15
15       MaxTempInChopperCoil  3.552714e-15
17      MaxTempInChopperDiode  0.000000e+00
19                TempPhaseT -4.447411e-02
20             WindDirection -5.385304e-02
13                TempGearBox -4.929901e-01
8                 TempRodMult -1.746587e+00
0                 TempAmbient -3.209710e+00
6                 TempNacelle -4.746351e+00
7           TempChopperDiode -3.757060e+01
```

Figura 13. ML, selección de variables con un modelo de "Regresión lineal"

La columna "Importancia" muestra los coeficientes del modelo, que indican la dirección (positiva o negativa) y la magnitud de la influencia de cada característica en la variable objetivo. Un coeficiente positivo significa que a medida que la característica aumenta, la variable objetivo también tiende a aumentar, y viceversa.

Utilizando esta información, se podría identificar las características más relevantes para las predicciones y tomar decisiones informadas sobre qué variables incluir o excluir en tu modelo. También puedes se puede utilizar estos resultados para entender cómo cada característica contribuye al resultado final y para la interpretación del modelo.

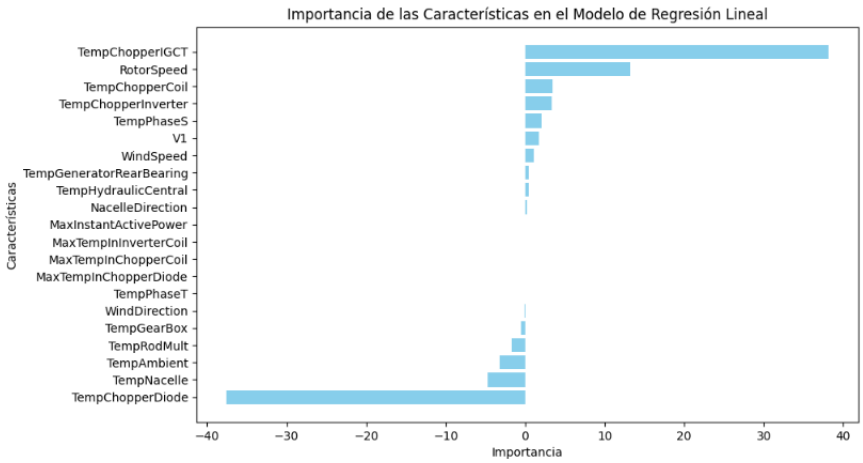

Figura 14. ML, selección de variables con un modelo de "Regresión lineal". Gráfico de barras

Hacer hincapié en que las variables que tienen un mayor peso en "la variable objetivo" es la TempChopperIGCT y TempChopperDiode, tal y como se puede ver de manera más clara en la figura 13. Todo parece apuntar a que "TempChopperIGCT" es una variable que representa la temperatura de un componente llamado "Chopper

IGCT" en un aerogenerador. Este componente es un dispositivo electrónico utilizado para controlar la velocidad o la potencia en sistemas de conversión de energía, como inversores o convertidores. Es aconsejable en estos casos, ponerse en contacto con el fabricante o ponerse en contacto con un experto para entender cada una de las variables que nutrirán el algoritmo.

4.2.1.4 Otras técnicas

En el contexto del modelado de curvas de potencia en aerogeneradores, se exploran diversas técnicas de selección de características además de la matriz de correlación y la selección basada en modelos de ML. Algunas de estas técnicas incluyen:

- **Análisis de Componentes Principales (PCA):** PCA es una técnica de reducción de dimensionalidad que proyecta los datos originales en un nuevo espacio de características, lo que facilita la identificación de variables relevantes.
- **Selección de Características por Pruebas Estadísticas:** Emplea pruebas como ANOVA o prueba t para evaluar la relación entre las variables predictoras y la variable objetivo, seleccionando aquellas con puntuaciones significativas.
- **Búsqueda "Hacia Atrás y Hacia Adelante:** Estas estrategias implican añadir o eliminar variables de forma iterativa, evaluando el rendimiento del modelo en cada paso. Hacia adelante comienza con un modelo vacío y agrega variables, mientras que hacia atrás parte de todas las variables y las elimina una por una.
- **Técnicas de Eliminación Recursiva de Características (RFE):** Estas técnicas entrenan modelos con todas las variables y eliminan iterativamente las menos importantes hasta lograr el conjunto deseado.
- **Análisis de Correlación No Lineal:** Además de la correlación lineal, se consideran técnicas que detectan correlaciones no lineales, como la correlación de rango de Spearman o la información mutua.
- **Validación Cruzada:** Utiliza técnicas de validación cruzada para evaluar cómo diferentes conjuntos de variables afectan el rendimiento del modelo en términos de métricas de evaluación como el error cuadrático medio (MSE) o el coeficiente de determinación (R^2).
- **Algoritmos de Selección de Características Incorporados**: Algunos algoritmos de aprendizaje automático, como LASSO (Least Absolute Shrinkage and Selection Operator) o Ridge Regression, incorporan selección de características como parte de su proceso de entrenamiento.

La elección de la técnica de selección de características dependerá de la naturaleza de los datos y los objetivos específicos del modelado. Es común combinar varias técnicas y evaluar cuál funciona mejor para un caso particular.

4.2.2 Identificación y filtrado de "outliers"

El preprocesamiento de datos se utiliza para filtrar los datos de viento cuando la turbina funciona en condiciones anormales, como las limitaciones de producción o daños en la integrad física de la máquina, aliviando así los efectos adversos de estos valores atípicos en las fases de entrenamiento del modelo de la curva de potencia. Sin embargo, no se puede garantizar que se detecten y procesen diferentes tipos de valores atípicos durante la fase de preprocesamiento de datos. Por lo tanto, algunas anomalías ocultas seguirán presentes en los datos. En consecuencia, la distribución de errores en los datos de modelización de la curva de potencia eólica será asimétrica. En la actualidad, se puede concluir a partir de la revisión de las técnicas de modelización de la curva de potencia eólica que pocos modelos han considerado adecuadamente este problema. En este Cuaderno se estudiarán algunos ejemplos sencilos, pero igual de efectivos, para preparar los datos para el posterior modelado de la curva con técnicas de aprendizaje automático

En esta segunda fase, por tanto, se identificará y eliminarán los valores atípicos en los datos. Estos valores atípicos pueden surgir debido a condiciones inusuales, fallos en la turbina o problemas en los sensores. La detección de outliers ayuda a garantizar que los datos se ajusten mejor a la curva de potencia ideal. Se utilizarán una serie de técnicas en este apartado y se irán explicando cada una de ellas detalladamente:

4.2.2.1 Diagramas de Caja – "Boxplots"

Los diagramas de caja son gráficos que representan la distribución de los datos y muestran visualmente los valores atípicos. Los valores que caen fuera de los "bigotes" del diagrama se consideran atípicos y pueden ser identificados fácilmente. He aquí un resumen de lo que representa:

- La "caja" es la diferencia entre el tercer cuartil (Q3) y el primer cuartil (Q1). El 50% de los datos se encuentran dentro de esta caja, y la línea en el medio de la caja representa la mediana (Q2), que es el valor que divide el conjunto de datos en dos mitades iguales.

- Bigotes (Whiskers): Los bigotes se extienden desde la caja hacia afuera y representan el rango de los datos dentro de un cierto límite. Los límites suelen calcularse como 1.5 veces el IQR. Cualquier valor que esté más allá de estos límites se considera un valor atípico (outlier) y se muestra como un punto individual en el gráfico.

- Valores Atípicos (Outliers): Los valores atípicos se muestran como puntos individuales que están por encima o por debajo de los bigotes. Estos son valores que se desvían significativamente de la mayoría de los datos y pueden indicar la presencia de observaciones inusuales o errores en los datos.

- Línea Media (Median Line): La línea en el centro de la caja representa la mediana (Q2), que es el valor medio de los datos.

- Límites Superior e Inferior: Los límites superior e inferior de la caja suelen representar el tercer cuartil (Q3) y el primer cuartil (Q1), respectivamente.

A continuación, se presentarán únicamente algunas variables clave, como 'ActivePower', 'WindSpeed', 'WindDirection', 'NacelleDirection', con el fin de simplificar la representación y resaltar la utilidad de esta técnica de manera más evidente.

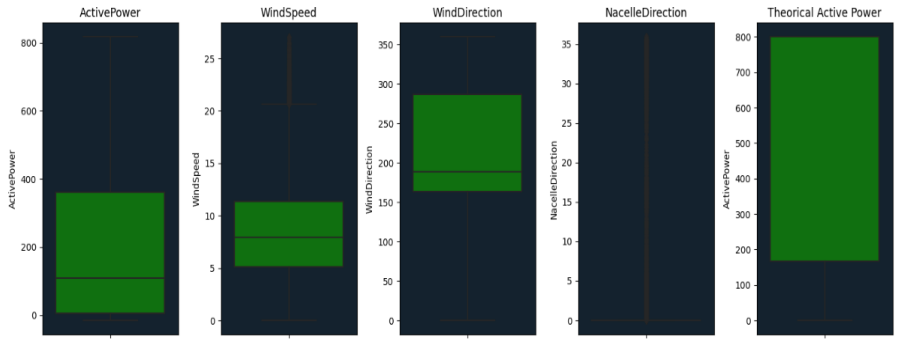

Figura 15. Diagrama de cajas para 'ActivePower', 'WindSpeed', 'WindDirection', 'NacelleDirection'

De esta primera técnica se puede obtener información relevante:

- A pesar de estar estudiando una turbina de 800kW y estar representando casi un año de datos, en el primer gráfico se puede observar cómo la mediana se concentra en los 100 kW de potencia para vientos que rondan los 5-10 m/s.
- Los bigotes parecen marcar valores coherentes, valores comprendidos entre 0-800 kW(nominal de la turbina) y viento entre 0-23 m/s, condiciones climáticas a priori posibles teniendo en cuenta la localización geográfica del emplazamiento.

- El tercer cuartil (Q3) para la Potencia Activa y Velocidad de viento es un claro de reflejo de que la turbina elegida para este presente estudio se alejará bastante de la curva teórico que la define.

4.2.2.2 Gráfico de dispersión – "Scatter plot"

Es una herramienta de visualización de datos que se utiliza para representar la relación entre dos variables numéricas. Cada punto en el gráfico de dispersión representa un par de valores correspondientes a las dos variables en estudio (WindSpeed y ActivePower, inputs que tendrán una mayor correlación).

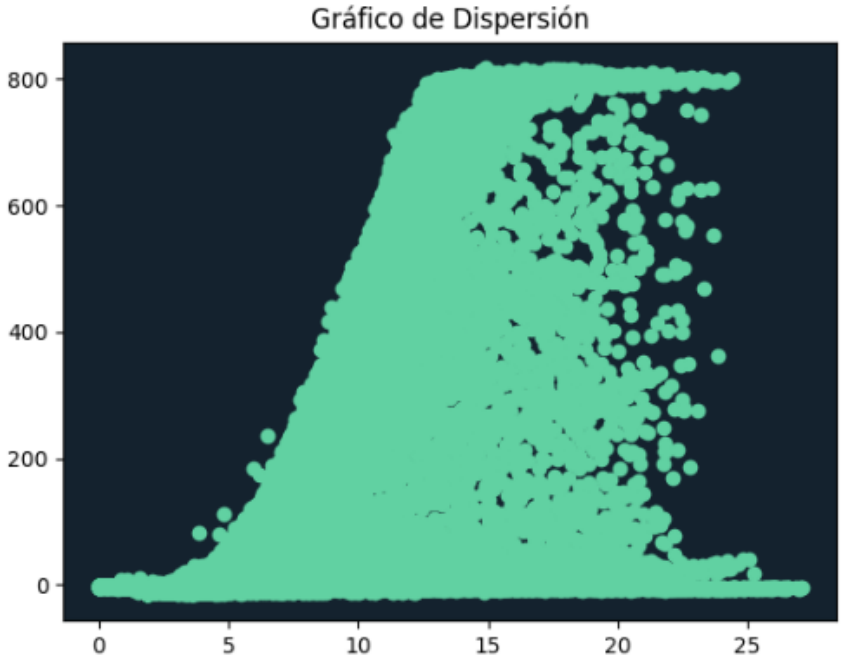

Figura 16. Gráfico de dispersion de 'ActivePower' y 'WindSpeed'- 'ScatterPlot'

Para este caso en particular:

- Se observan muchos datos anómalos en la curva estudiada, eso quiere decir que se tendrán que emplear técnicas de filtrado de outliers antes de usar las técnicas de predicción
- A pesar de que la curva presenta muchos outliers, La curva de potencia real se puede deducir en el gráfico.

4.2.2.3 Estudio de casos prácticos

En el contexto de la curva de potencia de una turbina eólica, a veces se pueden identificar outliers comparando los datos con los límites físicos y tecnológicos conocidos. Por ejemplo, si se registra potencia cuando la velocidad del viento es cero o está por encima de la velocidad de corte, se considera un outlier. En posteriores capítulos se dotará al algoritmo con la capacidad suficiente para definirle la potencia nominal y la potencia mínima con la que se modelará la curva.

Seguidamente se mostrará lo sencillo que resulta este filtrado y el impacto negativo que podría llegar a tener en las predicciones si no se lleva a cabo correctamente:

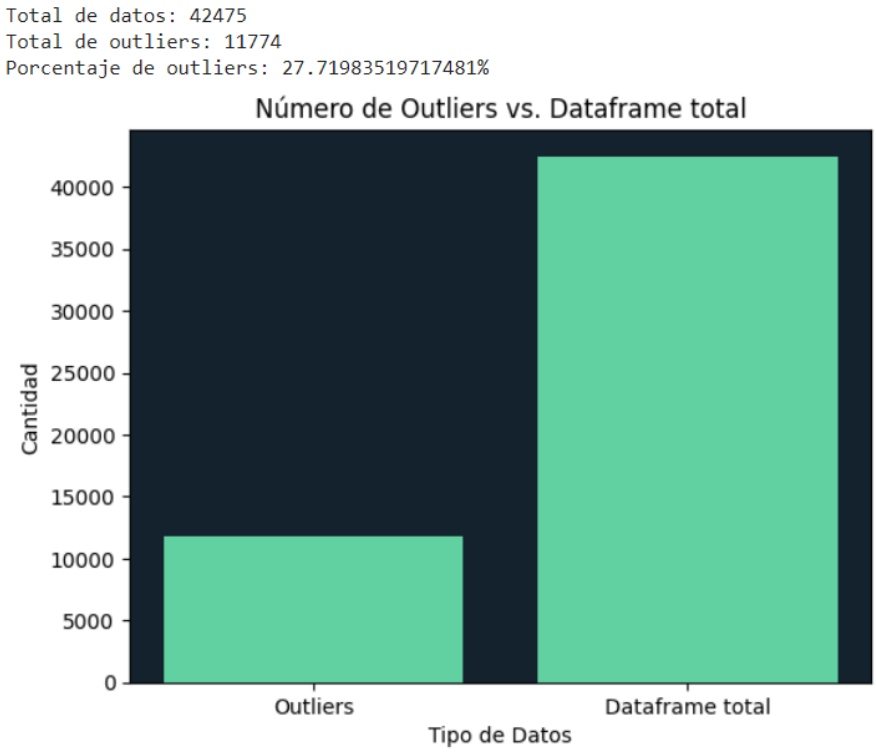

Figura 17. Filtrado de datos basado en la experiencia

Donde se ha tenido en cuenta:

- La potencia mínima será de 10 kW, para evitar el proceso de arranque y parada de la turbina que resulta bastante impredecible.
- La potencia máxima es 800 kW, valor máximo de potencia permitido, que se corresponde con la nominal de la turbina.

Sin duda se utilizará este filtrado en esta primera etapa, sin embargo, antes se estudiará un caso particular que se presenta en este dataset.

Adicionalmente, los datos obtenidos del SCADA tienen una peculiaridad muy útil en el presente proyeto. La turbina posee una variable que indica el estado operacional en el que se encuentra, en concreto el CS_StatusCat. Este estado operacional lo calcula el SCADA en función a las alarmas activas que presenta.

En el algoritmo que se está desarrollando se quiere descartar toda anomalía que pueda presentar la turbina, es decir, cuando la turbina esté trabajando en "condiciones anormales" de operación (alarmas activas, regulaciones, paradas por operadores o mantenimiento, etc). Este escenario se dará cuando la turbina tenga un CS_StatusCat=100 ("Running"), información obtenida del proveedor de datos.

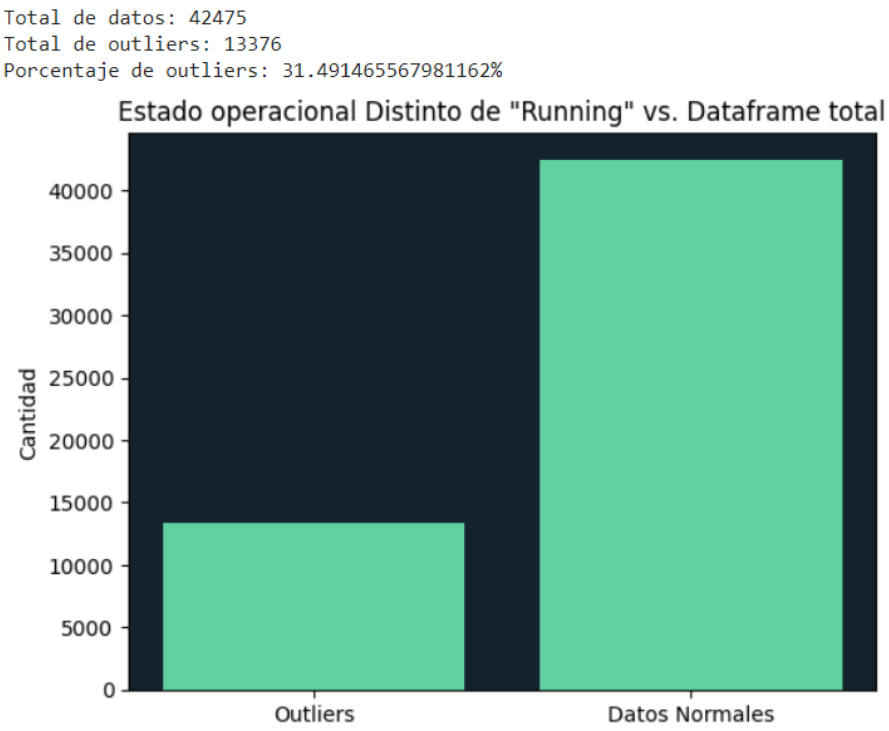

Figura 18. Filtrado de datos basado en la experiencia usando variables que definen el estado operacional

- Aplicando este filtrado se descartaría el 32% de los datos totales. Este método podría resultar muy agresivo, no obstante, se estaría garantizando un alto grado de fiabilidad en los resultados finales obtenidos al eliminar toda situación anómala en la operación de la turbina.

De forma más gráfica, se mostrará en un gráfico de dispersión lo que supondría este primer filtrado por variable CS_StatusCat en la forma de la curva resultante:

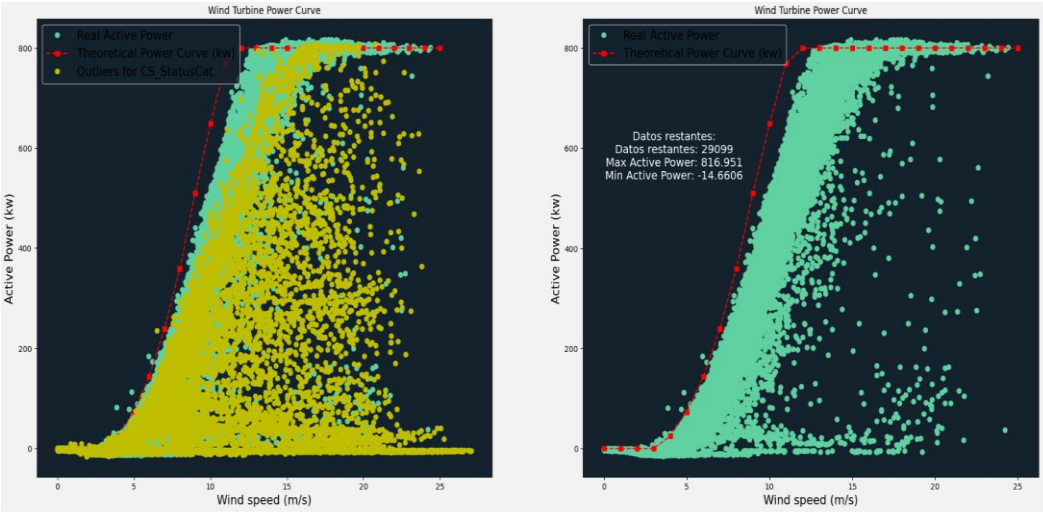

Figura 19. Filtrado de datos basado en la experiencia usando variables que definen el estado operacional. Gráfico de dispersión y resultados

- Este primer proceso de filtrado parece eliminar puntos inusuales en la curva, como por ejempl, los ya mencionados por “power curtailment”.
- Con este filtro, se logra una aproximación más cercana a la curva teórica propuesta por el tecnólogo.
- Es importante señalar que, a pesar de este filtrado inicial, aún no se han eliminado los valores extremadamente bajos o altos de potencia. Tras este primer filtrado, el conjunto de datos consta de 29,099 registros.

4.2.2.4 Filtrado con Técnicas estadísticas

Se pueden aplicar pruebas estadísticas formales, como la prueba de Grubbs o la prueba de Dixon, para detectar valores atípicos en función de criterios estadísticos. Estas pruebas establecen umbrales para identificar valores que se desvían significativamente de la distribución.

Z-score

En este apartado se utilizará el método "Z-score", que es una medida estadística que ayuda a cuantificar la relación de un valor de datos con respecto a la media y la dispersión de una distribución, lo que facilita la identificación de valores inusuales o atípicos en un conjunto de datos.

En una curva de potencia de una turbina eólica, este enfoque detectará valores atípicos en la velocidad del viento ('WindSpeed') que son estadísticamente inusuales en comparación con la mayoría de los datos. Estos outliers podrían deberse a condiciones meteorológicas extremas, fallas en los sensores o cualquier otra causa que provoque mediciones excepcionalmente altas o bajas de la velocidad del viento en relación con la media de la serie de datos.

Es importante destacar que esta técnica es bastante agresiva en la detección de outliers, ya que considera como tales a los valores que están más allá de 3 desviaciones estándar de la media. Dependiendo de la naturaleza de los datos y las necesidades específicas, es posible que se desee ajustar el umbral de corte (3 en este caso) para ser más o menos estricto en la identificación de outliers. Un valor más alto hará que la técnica sea menos sensible a los valores atípicos, mientras que un valor más bajo hará que sea más sensible.

En esta sección del algoritmo se establecen varias variables de entrada de gran relevancia:

- Desv se define como 3. Esto significa que los valores que se encuentren a más de 3 desviaciones estándar de la media serán considerados como valores atípicos u outliers.
- NominalPower se establece en 800 y servirá como la potencia nominal, además de definir la potencia máxima.
- CutinActivePower se fija en 10. Como se discutió en capítulos anteriores, se considerarán como mínimo 10 kW de potencia en la curva para evitar situaciones impredecibles.
- StepAP se establece en 10. Esto significa que el dataframe se dividirá en pequeños dataframes que abarcarán intervalos de 10 kW de potencia.
- El dataframe ya ha sido previamente filtrado en función de la variable operacional de la turbina.

Cabe destacar, como cabe esperar por los inputs definidos, que esta parte del código eliminará aquellos valores máximos y mínimos de potencia que se mencionó anteriormente.

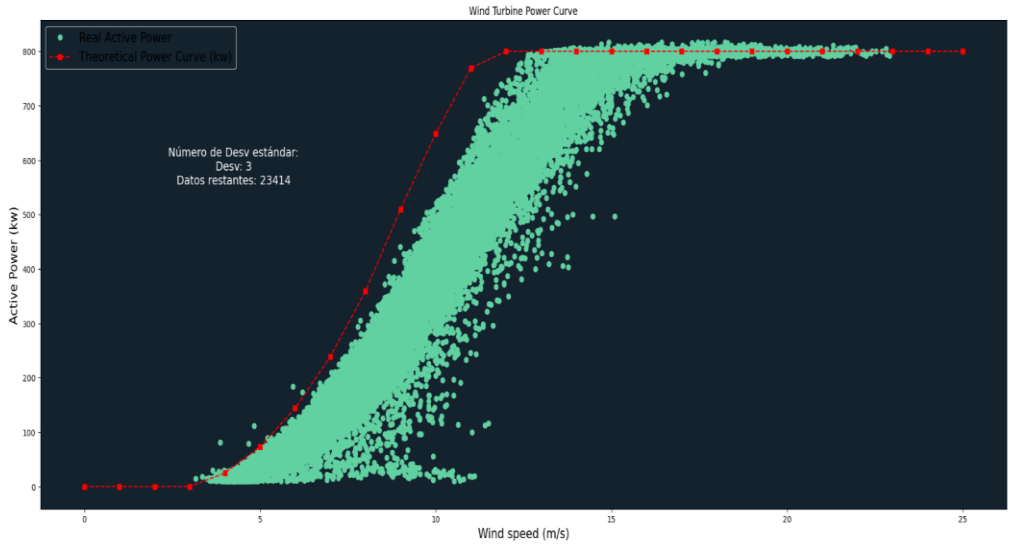

Figura 20. Filtrado de outliers. Análisis Z-score

- Según se puede apreciar en el gráfico que se muestra arriba (Figura 20), los resultados parecen seguir de cerca la curva de potencia teórica, pero sin realizar un sobreajuste en la real. Sin embargo, la curva sigue siendo bastante gruesa, esto podría llevar a predicciones erróneas.

Adicionalmente, se llevará a cabo un recorrido de la función para diferentes valores de desviación estándar, con el propósito de evaluar cómo afecta la variación de este parámetro, ya sea al aumentarlo o disminuirlo.

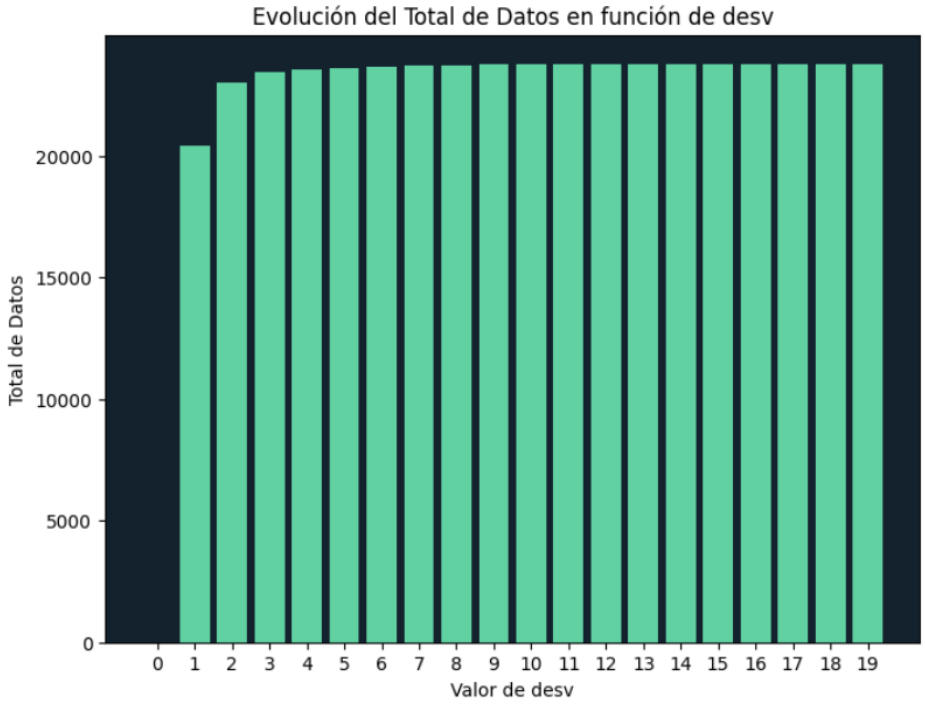

Figura 21. Filtrado de outliers. Análisis Z-score, peso del umbral “desv”

- Tras analizar el gráfico obtenido en la etapa anterior, parece evidente que las desviaciones estándar por encima de 3 apenas contribuirán significativamente a los resultados obtenidos.

Cuantiles

Una alternativa al método previamente utilizado podría ser el uso de cuantiles, lo cual presenta ventajas significativas. Este enfoque es robusto y adaptable a diferentes distribuciones de datos. Además, conserva más información útil, lo que resulta en una reducción de errores. Es generalizable a distintos conjuntos de datos y fácil de interpretar, lo que fomenta la consistencia en el análisis de datos.

Los cuantiles son valores que dividen un conjunto de datos ordenados en partes iguales, lo que permite entender la distribución de los datos y encontrar valores atípicos. Por ejemplo, el cuantil 50% es la mediana, que separa la mitad de los datos por encima y por debajo. Los cuantiles son útiles en estadísticas y análisis de datos para resumir la información y detectar valores extremos.

A continuación, se han desarrollado 2 funciones muy similares a las usadas anteriormente con el método z-score, pero adaptadas a este nuevo método. La primera función utiliza cuantiles para definir un rango dentro del cual se consideran los valores como no outliers (Qmin 0.25 y Qmax 0.75), y luego filtra el DataFrame original para retener solo esos valores dentro de ese rango. Esto es útil para eliminar valores extremos que pueden distorsionar el análisis de datos

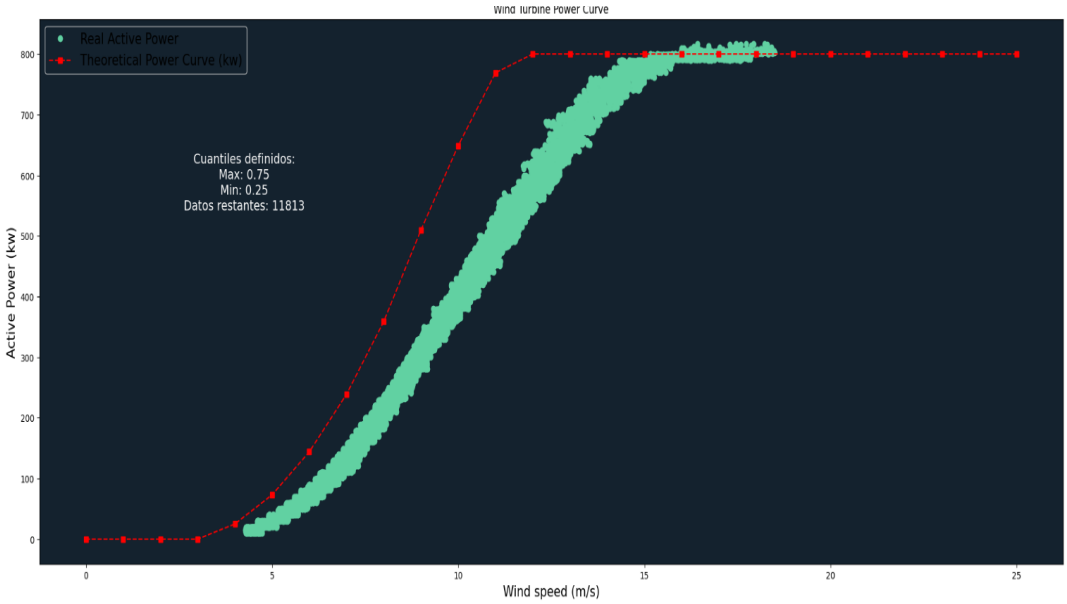

Figura 22. Filtrado de outliers. Análisis con Cuantiles

- Inicialmente, al analizar los resultados, se puede inferir que el conjunto de datos está experimentando un fenómeno de sobreajuste, lo cual se refleja en el grosor de la curva. Este suceso podría plantear ciertas preocupaciones en cuanto a la idoneidad de las predicciones, dado que este tipo de enfoque no se muestra particularmente versátil y adaptable, características esenciales en un contexto real.

Z-score vs Cuantiles

La elección entre las dos funciones de filtrado de datos, una basada en puntuaciones Z (Z-scores) y la otra basada en cuantiles, depende de la naturaleza de los datos y los requisitos específicos de tu análisis. Ambos métodos tienen ventajas y desventajas:

- Filtrado por Puntuaciones Z (Z-scores)

Ventajas: Es efectivo para identificar outliers en función de la desviación estándar. Permite un control preciso sobre la cantidad de datos que se eliminan ajustando el valor de Des (umbral de Z-score). Es adecuado cuando los datos siguen una distribución normal o aproximadamente normal.

Desventajas: Puede eliminar valores que son atípicos pero relevantes para ciertos análisis. Menos robusto ante distribuciones de datos altamente asimétricas o con colas pesadas.

- Filtrado por Cuantiles:

Ventajas: Es robusto y menos sensible a distribuciones de datos no normales o con colas pesadas. Conserva más información útil, incluyendo valores atípicos que podrían ser importantes. Es fácil de interpretar y generalizable a diferentes tipos de datos.

Desventajas: No permite un control tan preciso sobre la cantidad de datos que se eliminan como las puntuaciones Z.

En general, si los datos tienen una distribución normal o se asemejan a una distribución normal, el filtrado por puntuaciones Z puede ser adecuado y dará un mayor control sobre la cantidad de datos eliminados. Sin embargo, si los datos son asimétricos o tienes razones para conservar valores atípicos, el filtrado por cuantiles es una opción más robusta.

La elección dependerá del objetivo final del usuario que modele el algoritmo y del conocimiento sobre la distribución de datos. En este caso se probarán ambos métodos para determinar cuál podría ser más eficaz para este caso específico

4.2.2.5 Filtrado con Técnicas de ML

Los métodos de Aprendizaje Automático pueden ser valiosos para la identificación automatizada de outliers en grandes conjuntos de datos. A continuación, se muestran algunos ejemplos de métodos de Aprendizaje Automático que se utilizan comúnmente para la detección de anomalías:

- **Isolation Forest:** Este algoritmo crea un bosque de árboles de decisión aleatorios y aísla los outliers al identificar las instancias que requieren menos divisiones en los árboles para ser separadas. Los datos que requieren menos divisiones se consideran más propensos a ser outliers.
- **Local Outlier Factor (LOF):** LOF compara la densidad local de un punto de datos con la densidad local de sus vecinos. Si un punto tiene una densidad significativamente menor en comparación con sus vecinos, se considera un outlier.

- **One-Class SVM (Support Vector Machine**): Este método entrena un modelo para clasificar si un punto de datos pertenece a una sola clase (la clase "normal") o es un outlier. Puede ser efectivo cuando se tiene un conjunto de datos desequilibrado con una gran cantidad de datos normales y pocos outliers.
- **DBSCAN (Density-Based Spatial Clustering of Applications with Noise):** Aunque principalmente se utiliza para la agrupación, DBSCAN también puede identificar outliers como ruido en el conjunto de datos. Detecta puntos que no pueden ser agrupados en ningún clúster y los etiqueta como outliers.
- **Autoencoders:** En el aprendizaje profundo, los autoencoders pueden utilizarse para reconstruir los datos de entrada y luego calcular el error de reconstrucción. Los datos que tienen un alto error de reconstrucción se consideran outliers.
- **K-Means:** Si bien K-Means es una técnica de agrupación, también se puede usar para detectar outliers midiendo la distancia de cada punto al centroide más cercano. Los puntos que están muy lejos de su centroide se consideran outliers.

Estos son solo algunos ejemplos de métodos de Aprendizaje Automático que pueden utilizarse para detectar outliers en conjuntos de datos. La elección del método depende de la naturaleza de tus datos y de los patrones de outliers que se estén tratando de identificar.

Isolation Forest

En el presente proyecto se usará Isolation Forest para detectar anomalías en el conjunto de datos. Esta técnica basa en la idea de que los outliers son puntos de datos que pueden aislarse más fácilmente que los datos normales en un conjunto de datos multidimensional. El algoritmo crea múltiples árboles de decisión de forma aleatoria y utiliza la profundidad de los puntos en estos árboles para calcular puntajes de anomalía. Los puntos con puntajes de anomalía más altos se consideran outliers.

Ventajas de Isolation Forest:

- Eficiente y escalable para conjuntos de datos grandes.
- No requiere conocimiento previo de la distribución de datos.
- Adecuado para datos multidimensionales.
- Efectivo en la detección de outliers incluso en conjuntos de datos desequilibrados.

Para realizar este proceso de filtrado, asumiremos que el 50% de los datos en el dataframe están contaminados, lo que significa que la mitad de los datos en el dataframe se consideran valores atípicos. En este caso específico, dado que el dataframe se divide en más de 50 dataframes más pequeños, esto implica que en cada subconjunto se supondrá que el 50% de sus datos también son valores atípicos.

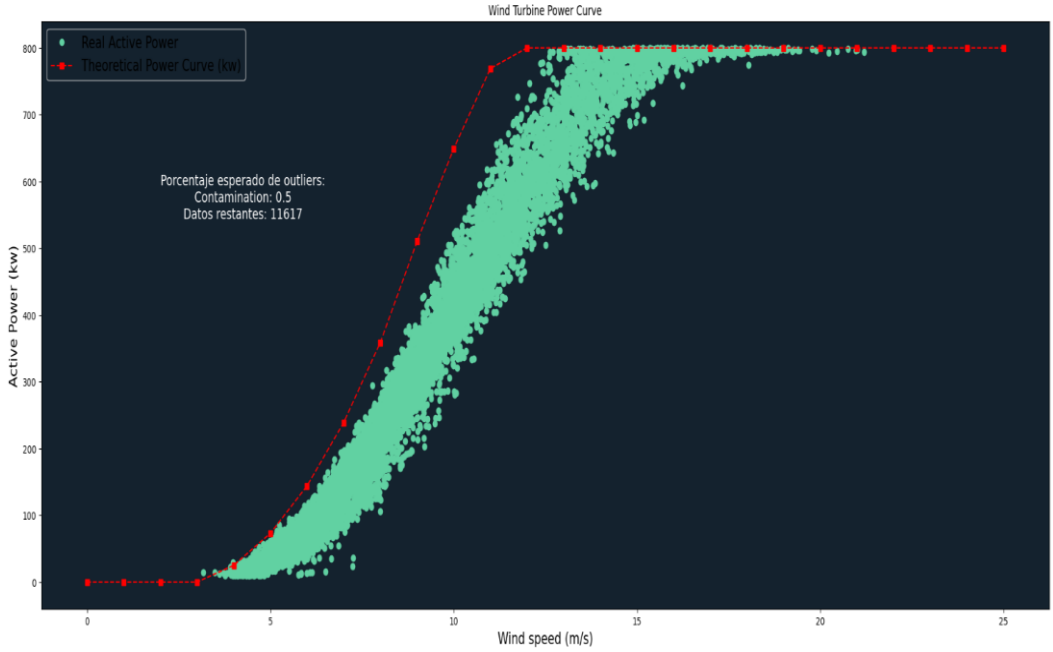

Figura 23. Filtrado de outliers. ML, Isolation Forest

- Como era de esperar de una técnica de aprendizaje automático, los resultados obtenidos parecen ser significativamente mejores que los obtenidos mediante las metodologías anteriores. El ancho de la curva se ha reducido sin caer en un sobreajuste como sucedió con el método de cuantiles, y parece que no hay valores atípicos que tengan un impacto significativo en los datos que componen la curva.

Finalmente, tras aplicar todos estos métodos de filtrado de outliers, se le da al usuario la capacidad de decidir, imprimiendo por pantalla las opciones, qué método decide usar en su estudio.

```
Elija el método que desea elegir para eliminar los outliers:
1. Isolation Forest
2. Cuantiles
3. Z-Score
Por favor, ingrese el número correspondiente al DataFrame que desea utilizar (1, 2 o 3): 1
```

Figura 24. Filtrado de outliers. Algoritmo que permite al usuario elegir la metodología a seguir en el proceso de modelado.

5 MODELADO DE LA CURVA DE POTENCIA

En el campo de la generación de energía eólica, el modelado preciso de la curva de potencia de un aerogenerador es de vital importancia para maximizar su eficiencia y rendimiento. Para lograr esta precisión, se recurre a técnicas de Machine Learning (ML) que permiten construir modelos predictivos capaces de estimar la potencia generada en función de diversas variables ambientales. Estos modelos emplean algoritmos de regresión que analizan patrones en los datos históricos de operación de la turbina, como la velocidad del viento, la dirección del viento y otros parámetros relevantes.

En este contexto, este proyecto explorará las diversas técnicas de ML utilizadas para el modelado de la curva de potencia en aerogeneradores, en concreto KNN, DTR y ETR, destacando su importancia en la mejora del rendimiento de las instalaciones eólicas y su contribución al aprovechamiento óptimo de la energía renovable. Además, se examinarán ejemplos prácticos y consideraciones clave en la implementación de estos modelos, brindando una visión completa de su aplicación en la industria eólica.

5.1 Estrategias implementadas para el modelado

En el marco de este proyecto, a continuación, se precisará qué estrategia se ha elegido para el modelado de la curva de potencia, de entre todas las técnicas que se han discutido. Es importante destacar que estos parámetros son ajustables, y, además, algunas de las funciones empleadas disponen de diversos hiperparámetros que tienen el potencial de optimizar aún más el desempeño del algoritmo.

- Inputs para el modelado: Se empleó una matriz de correlación con un umbral de +/- 0.6, lo que resultó en la identificación de 12 variables de entrada. Además, se incluyeron las variables "WindDirection" y "NacelleDirection" en el conjunto de entrada, a pesar de su correlación aparentemente baja, debido a su posible utilidad en las predicciones."

```
inputs_with_methods
```

```
['ReactivePower',
 'WindSpeed',
 'V1',
 'RotorSpeed',
 'TempChopperDiode',
 'TempRodMult',
 'TempChopperCoil',
 'TempGeneratorRearBearing',
 'TempGearBox',
 'TempChopperInverter',
 'TempChopperIGCT',
 'TempPhaseT',
 'WindDirection',
 'NacelleDirection',
 'MaxTempInInverterCoil',
 'ActivePower']
```

Figura 25. Variables de entrada elegidas en el estudio

- Se realiza un filtrado basado en el estado operacional "En Marcha" (CS_StatusCat = 100).
- Para la detección de valores atípicos, se implementa el método Isolation Forest debido a su robustez en los resultados y su capacidad para prevenir el sobreajuste de la curva. Se considera una "contaminación" del 50% de los datos. Además, el dataframe original se divide en subconjuntos de 10 kW cada uno, a los cuales se les aplica esta técnica de detección de valores atípicos (consulte la Figura 23).
- Se establece un umbral de potencia mínima de 10 kW y una potencia nominal de 800 kW.

5.2 Métricas de Evaluación para Analizar la Calidad de las Predicciones

Las métricas desempeñan un papel fundamental en la evaluación de modelos de aprendizaje automático, especialmente cuando se trata de modelar una curva de potencia. Estas métricas permiten medir cuán precisa y efectiva es la representación del comportamiento de la curva obtenida por el modelo en comparación con los valores reales. En este contexto, las métricas proporcionan una base objetiva para evaluar la calidad del modelo y determinar su capacidad para predecir la potencia de manera precisa. A continuación, se profundizará en detalle en algunas de las métricas clave y su importancia en el proceso de evaluación del modelo:

Mean Absolute Error (MAE - Error Absoluto Medio):

- Representa el promedio de las diferencias absolutas entre las predicciones del modelo y los valores reales.
- Cuanto menor sea el MAE, mejor será el modelo. Indica cuán cerca están las predicciones del modelo de los valores reales en promedio.
- Esta métrica es más resistente a los valores atípicos, por lo que la convierte en una elección adecuada cuando se desea una evaluación del rendimiento del modelo que sea menos influenciada por observaciones extremas en los datos (Ver ecuación (2)). En este escenario, los valores atípicos ya han sido eliminados, lo que hace que esta métrica sea muy apropiada y válida.

  Por ejemplo, un MAE de 30 kW querrá decir que, de media, las predicciones tienen un error de +/- 30kW

$$mae = \frac{1}{n}\sum|y_i - \bar{y}| \qquad (2)$$

Mean Squared Error (MSE - Error Cuadrático Medio):

- Representa el promedio de las diferencias cuadráticas entre las predicciones del modelo y los valores reales.
- Es más sensible a errores grandes debido al término de cuadrado. Cuanto menor sea el MSE, mejor será el modelo.
- La unidad de medida es el cuadrado de la unidad de medida de la variable objetivo original, lo que puede dificultar la interpretación.

$$mse = \frac{1}{n} \sum (y_n - \bar{y})^2 \qquad (3)$$

Root Mean Squared Error (RMSE - Raíz del Error Cuadrático Medio):

- Es la raíz cuadrada del MSE. Se utiliza para proporcionar una medida del error en la misma escala que los datos originales. Es una medida que tiene la misma unidad de medida que la variable objetivo original, lo que facilita la interpretación.
- Combina la sensibilidad a los errores grandes de MSE con la facilidad de interpretación de MAE.
- Al igual que MSE, RMSE puede verse afectada por valores atípicos, ya que los errores se elevan al cuadrado antes de tomar la raíz cuadrada.

$$mse = \sqrt{\frac{1}{n} \sum (y_n - \bar{y})^2} \qquad (4)$$

Accuracy on Training set (Precisión en el conjunto de entrenamiento):

Esta métrica se utiliza comúnmente en problemas de clasificación y representa la precisión del modelo en predecir correctamente las etiquetas en el conjunto de entrenamiento.

Puede ser útil para evaluar el rendimiento del modelo en los datos de entrenamiento, pero no siempre es un buen indicador de la capacidad del modelo para generalizar a nuevos datos.

Accuracy on Testing set (Precisión en el conjunto de pruebas):

Similar a la métrica de precisión en el conjunto de entrenamiento, esta métrica evalúa la precisión del modelo en el conjunto de prueba, que representa datos no vistos previamente por el modelo.

Es una métrica importante para evaluar qué tan bien se desempeña el modelo en datos nuevos y desconocidos.

5.3 Técnicas usadas en el modelado

5.3.1 K-Vecinos más próximos (KNN)

El modelo "KNN" es el modelo no paramétrico más simple, este ha tenido éxito en multitud de aplicaciones, incluido el modelado de curvas de potencia de turbinas eólicas. El principio fundamental de esta metodología es encontrar un número predeterminado de muestras de aprendizaje cercanas, "vecinos" (en términos de distancia), a la muestra considerada y posteriormente estimar la predicción. En general, la medida métrica de distancia es la distancia euclidiana estándar. Asimismo, podrá especificarse el número de vecinos más cercanos a examinar; este valor se denomina "k".

Además de K, existen otros hiperparámetros que se pueden ajustar, como la métrica de distancia utilizada (por ejemplo, distancia Euclidiana o Manhattan), el peso de los vecinos (uniforme o ponderado), y otros parámetros relacionados con el rendimiento y la eficiencia del algoritmo.

Por ejemplo, si K = 3, el algoritmo buscará los 3 puntos de entrenamiento más cercanos al punto de prueba y calculará una predicción basada en la media (o mediana) de los valores de destino de esos 3 vecinos.

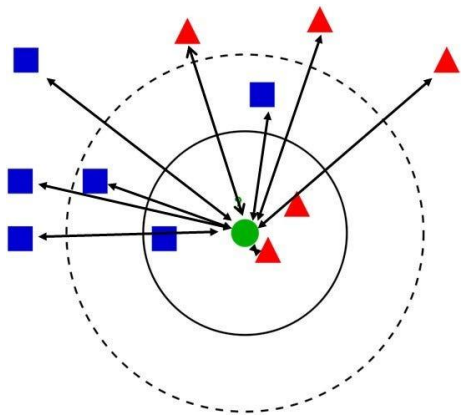

Figura 26. Método "K-Vecinos más próximos"

Resultados en las predicciones

Method	Mean Absolute Error	Mean Squared Error	Root Mean Squared Error	Accuracy on Training set	Accuracy on Testing set
KNeighborsRegressor	30.595733	2064.881928	45.440972	0.97721	0.964168

Tabla 1. Métricas obtenidas en el modelado con "KNeighborsRegressor"

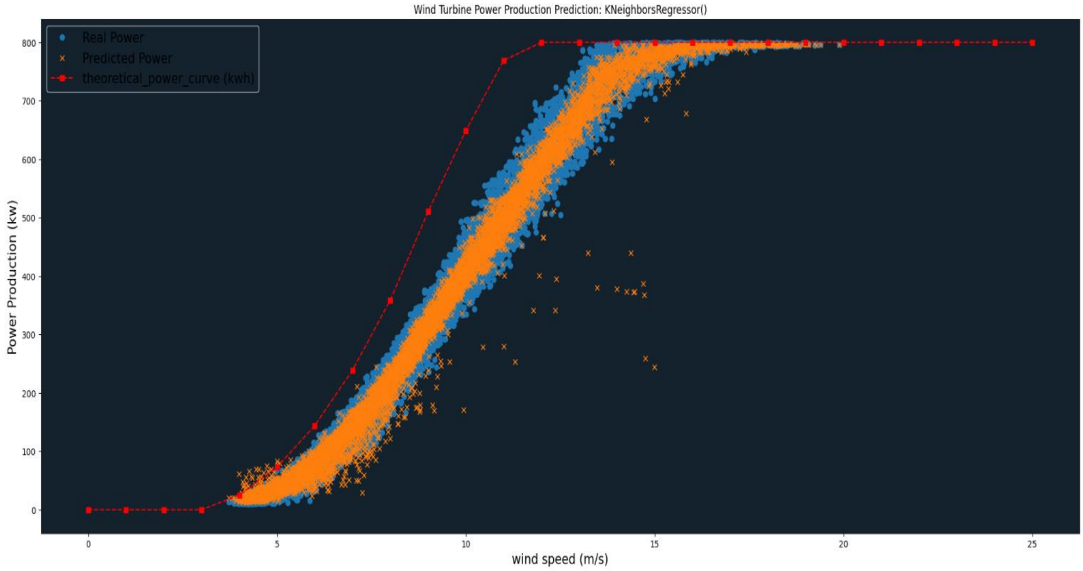

Figura 27. Resultados de Predicciones Utilizando el DataFrame Completo, ya filtrado, con el Modelo "KNeighborsRegressor"

5.3.2 Regresión por Árboles de Decisión

Es una herramienta de aprendizaje automático utilizada para realizar tareas de regresión, lo que significa que se utiliza para predecir valores numéricos continuos en función de un conjunto de variables de entrada. Esta técnica se basa en la construcción de un árbol de decisión que divide los datos en subconjuntos cada vez más pequeños, de manera que se pueda realizar una predicción precisa en cada "hoja" del árbol.

A continuación, se explicará de forma simplificada cómo funciona el "DecisionTreeRegressor":

- **Construcción del Árbol**: El proceso comienza con un único nodo que contiene todos los datos de entrenamiento. Luego, el algoritmo busca la característica (variable de entrada) y el umbral que mejor separan los datos en dos grupos. Esto se hace seleccionando la característica y el umbral que minimizan el error de regresión, que suele ser el error cuadrático medio (MSE).

- **División de los Datos**: Una vez que se elige la característica y el umbral, los datos se dividen en dos subconjuntos en función de si cumplen o no con la condición establecida por la característica y el umbral seleccionados. Esto crea dos nodos hijos conectados al nodo padre.
- **Recursión:** El proceso se repite recursivamente en cada nodo hijo, dividiendo los datos nuevamente en función de las características y umbrales que minimizan el error de regresión en cada paso. Este proceso continúa hasta que se alcanza un criterio de parada, como una profundidad máxima del árbol o un número mínimo de muestras por hoja.
- **Predicción**: Para hacer una predicción para una nueva muestra, se sigue el camino desde el nodo raíz hasta una hoja del árbol, y se toma el valor promedio de las muestras en esa hoja como la predicción final.

El DecisionTreeRegressor tiene la ventaja de ser fácil de comprender y visualizar, pero tiende a ser propenso al sobreajuste si no se controla adecuadamente. Para mitigar el sobreajuste, se pueden aplicar técnicas como la poda del árbol o limitar su profundidad máxima.

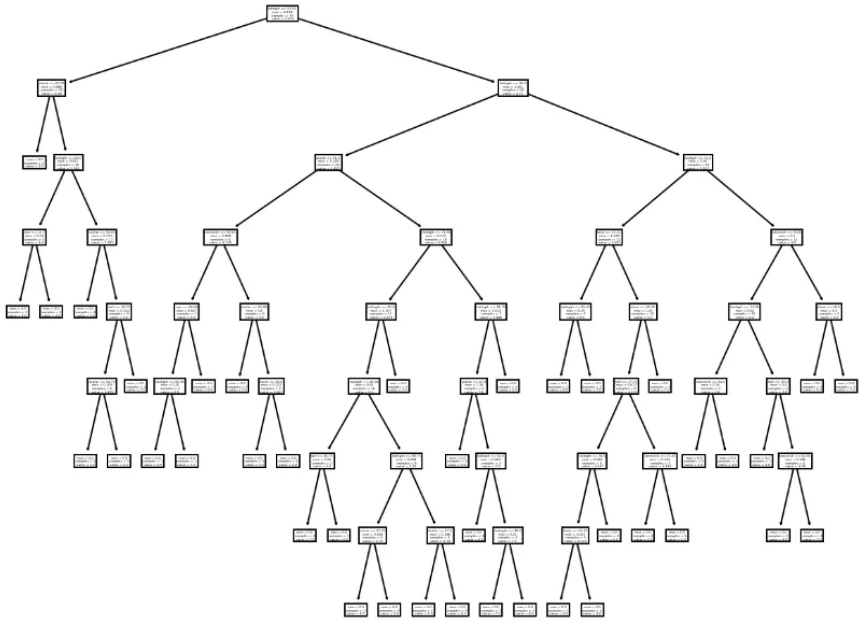

Figura 28. Método “Regresión por Árboles de Decisión”

Resultados en las predicciones

Method	Mean Absolute Error	Mean Squared Error	Root Mean Squared Error	Accuracy on Training set	Accuracy on Testing set
DecisionTreeRegressor	14.698545	624.267606	24.985348	1.00000	0.989167

Tabla 2. Métricas obtenidas en el modelado con “DecisionTreeRegressor”

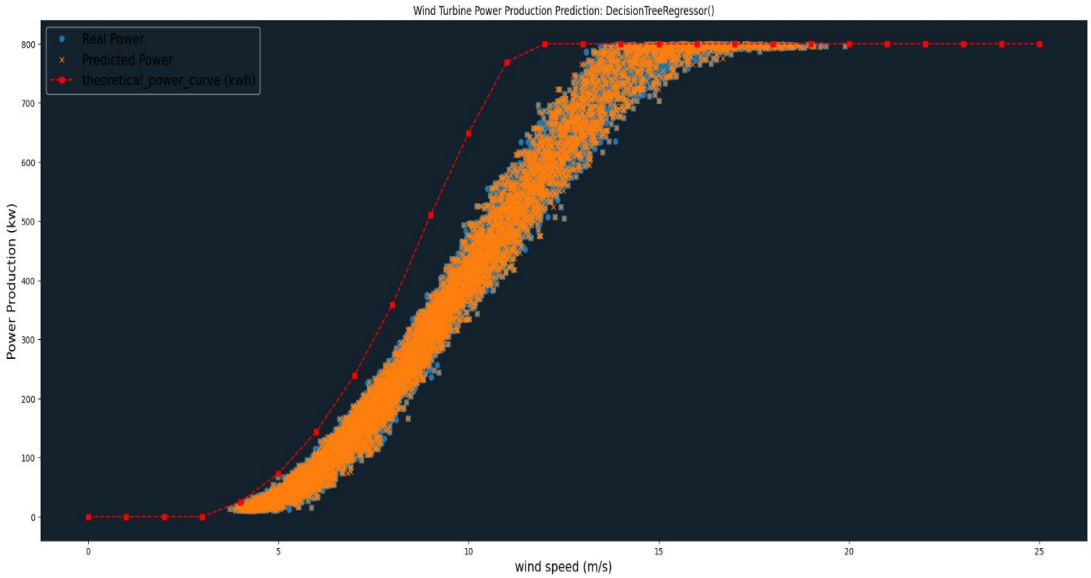

Figura 29. Resultados de Predicciones Utilizando el DataFrame Completo, ya filtrado, con el Modelo " DecisionTreeRegressor "

5.3.3 Regresión de Árboles extra

Este algoritmo es una variante de los Árboles de Decisión y pertenece a la familia de métodos de ensamblado de modelos, específicamente a los "Ensembles" basados en árboles.

La característica distintiva de esta técnica radica en su alta aleatoriedad durante la construcción de árboles. A diferencia de los árboles de decisión estándar, que seleccionan las mejores características y umbrales en función de criterios como la reducción del error cuadrático medio (MSE), en este caso se toma un enfoque más aleatorio y diversificado. En cada nodo de división del árbol:

- Se eligen características al azar de todo el conjunto de características disponibles.
- Se seleccionan umbrales al azar para esas características.
- Se construyen múltiples árboles de decisión de esta manera simultáneamente.

Esta aleatoriedad adicional en la construcción de árboles tiene varios beneficios:

- Reduce el riesgo de sobreajuste, ya que los árboles resultantes son más diversos y menos propensos a ajustarse demasiado a los datos de entrenamiento.
- Aumenta la robustez del modelo al reducir la influencia de datos atípicos y el ruido en los datos.
- Permite una construcción de modelos más rápida en comparación con árboles de decisión estándar.

En la etapa de predicción, “ExtraTreesRegressor” realiza un promedio de las predicciones de todos los árboles construidos, lo que produce una predicción final. Este proceso de ensamblado ayuda a mejorar la estabilidad y la precisión de las predicciones.

Resultados en las predicciones

Method	Mean Absolute Error	Mean Squared Error	Root Mean Squared Error	Accuracy on Training set	Accuracy on Testing set
ExtraTreesRegressor	10.062754	267.431532	16.353334	1.00000	0.995359

Tabla 3. Métricas obtenidas en el modelado con “ExtraTreesRegressor”

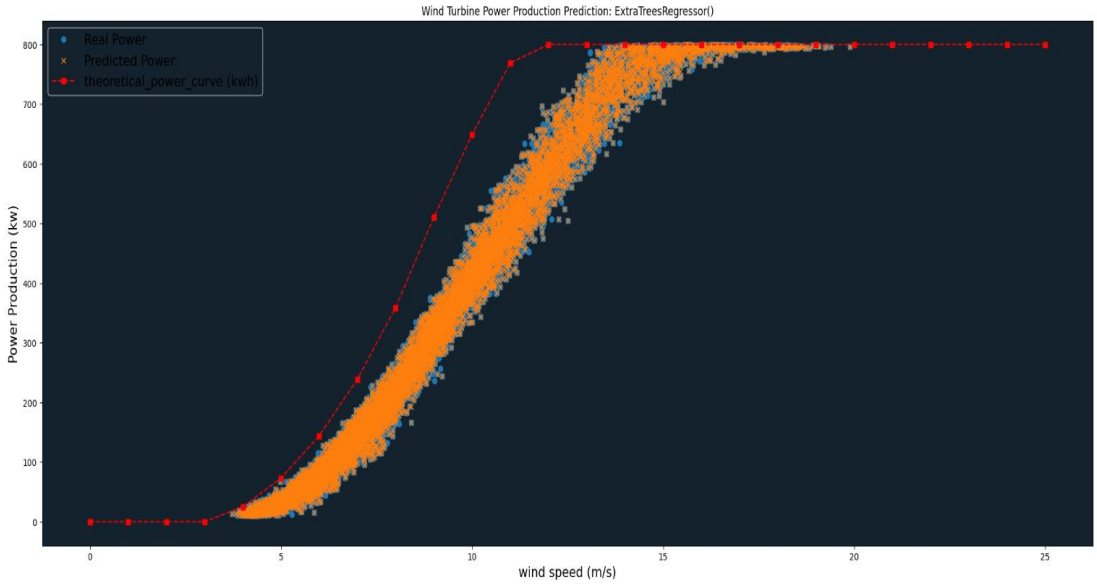

Figura 30. Resultados de Predicciones Utilizando el DataFrame Completo, ya filtrado, con el Modelo " ExtraTreesRegressor "

5.3.4 Otras técnicas

Además de los algoritmos previamente mencionados (KNeighborsRegressor, DecisionTreeRegressor y ExtraTreesRegressor), existen diversas alternativas en el campo de Machine Learning que se pueden considerar adecuadas para abordar el modelado de la curva de potencia en aerogeneradores:

- **Random Forest Regressor**: Este método es similar al ExtraTreesRegressor, ya que se basa en conjuntos de árboles de decisión. Sin embargo, utiliza un enfoque diferente para mejorar la precisión del modelo.
- **Support Vector Machine (SVM) Regressor**: El SVM puede ser especialmente efectivo cuando se trata de modelar relaciones no lineales en los datos de la curva de potencia.
- **Gradient Boosting Regressor:** Algoritmos como GradientBoostingRegressor son altamente eficaces para mejorar la precisión del modelo al combinar múltiples modelos más simples.
- **Redes Neuronales (Deep Learning**): Si bien las redes neuronales pueden capturar relaciones complejas en los datos, generalmente requieren un gran volumen de datos y una afinación cuidadosa de los hiperparámetros.
- **Regresión Lineal:** A pesar de su simplicidad, la regresión lineal puede servir como punto de partida para establecer una línea base y compararla con modelos más avanzados.
- **Regresión LASSO y Ridge:** Estos métodos de regularización son útiles para evitar el sobreajuste en modelos de regresión lineal.
- **XGBoost y LightGBM**: Estas bibliotecas de gradient boosting, XGBoost y LightGBM, son altamente optimizadas y se destacan en competencias de Machine Learning.

La elección del algoritmo más apropiado dependerá de diversos factores, como el tamaño del conjunto de datos, la complejidad de la relación entre las variables, los objetivos de rendimiento y los recursos computacionales disponibles. Se recomienda llevar a cabo pruebas y experimentos con varios algoritmos y técnicas de selección de características para determinar cuál se adapta mejor a un conjunto de datos específico.

5.4 Evaluación del rendimiento de la turbina eólica

La eficiencia en la generación de energía eólica es una preocupación clave en la industria de las energías renovables. Para garantizar un rendimiento óptimo de las turbinas eólicas, es esencial evaluar y monitorear continuamente su producción. En este contexto, el uso de modelos de aprendizaje automático se ha convertido en una herramienta fundamental para predecir y analizar el rendimiento de las turbinas. Este proyecto presenta un código que permite analizar y visualizar cómo diferentes modelos de aprendizaje automático predicen el rendimiento de las turbinas eólicas, categorizándolo en niveles de alto, medio y bajo rendimiento. Esta evaluación es esencial para identificar oportunidades de mejora, anticiparse a problemas y garantizar una generación de energía eficiente y sostenible a partir de fuentes eólicas.

El algortimo desarrollado en este apartado evalúa cómo los modelos de aprendizaje automático predicen el rendimiento de una turbina eólica y lo categoriza en tres niveles: Alto Rendimiento, Rendimiento Medio y Bajo Rendimiento.

Utiliza la precisión de las predicciones para determinar el rendimiento de la turbina, donde valores positivos indican que la turbina está produciendo por encima de las predicciones, y valores negativos indican un rendimiento por debajo de las expectativas. Adicionalmente, permite al usuario definir límites de rendimiento y seleccionar modelos de aprendizaje automático a analizar. Y crea gráficos que muestran la distribución del rendimiento de la turbina según cada modelo, utilizando barras y gráficos de pastel para visualizar los resultados.

La función desplegada permite establecer los umbrales de evaluación de rendimiento de la turbina. En este caso particular, se ha definido que si la predicción supera el 90% del valor real, se considerará que la turbina está funcionando a un nivel de rendimiento alto. Si la predicción se encuentra entre el 70% y el 90% del valor real, se clasificará como un rendimiento medio, mientras que si la predicción es inferior al 70% del valor real, se considerará un bajo rendimiento. Es importante destacar que cuando la predicción es menor que el valor real, siempre se considerará que la turbina está funcionando a un nivel de alto rendimiento.

La categorización del rendimiento de la turbina es fundamental para identificar problemas y oportunidades de mejora en la generación de energía eólica, lo que puede conducir a un funcionamiento más eficiente y rentable de las turbinas.

Resultado de las evaluaciones

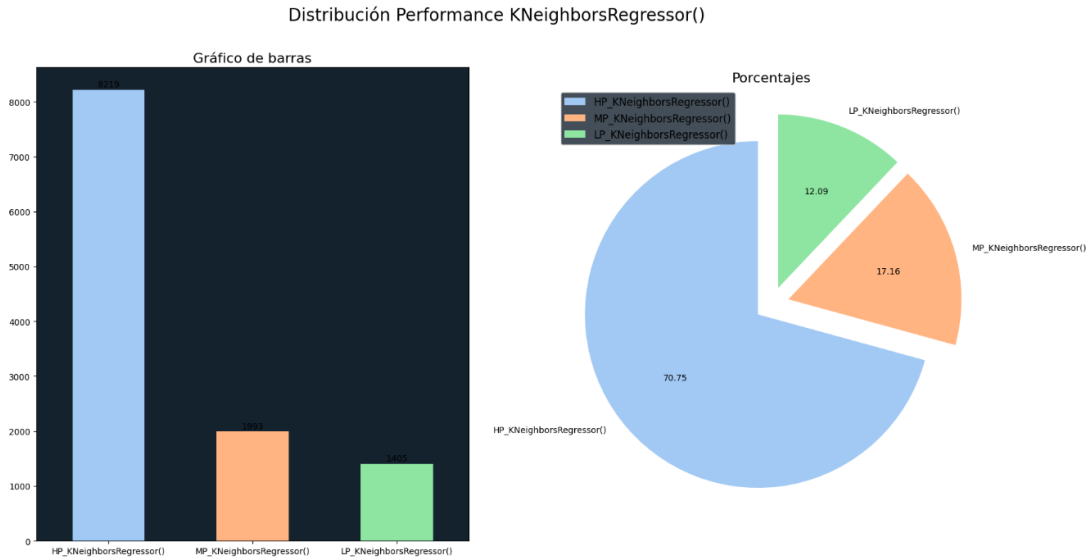

Figura 31. Evaluación del rendimiento de la turbina. Modelo " KNeighborsRegressor"

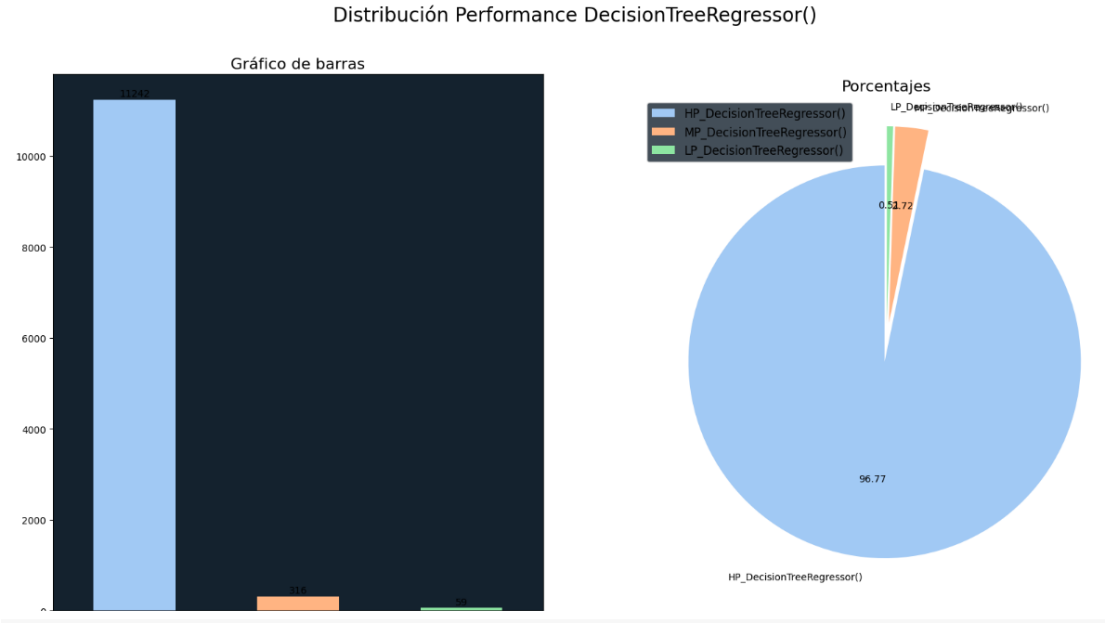

Figura 32. Evaluación del rendimiento de la turbina. Modelo " DecisionTreeRegressor "

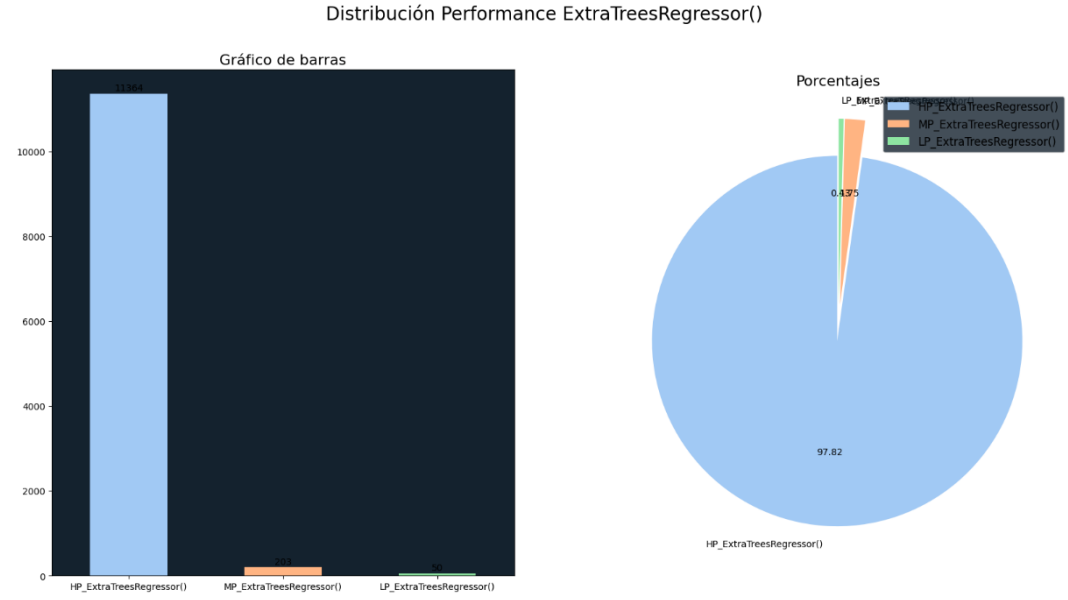

Figura 33. Evaluación del rendimiento de la turbina. Modelo " ExtraTreesRegressor "

6 DISCUSIÓN Y ANÁLISIS DE RESULTADOS

Resumidamente, los resultados obtenidos en el presente proyecto son altamente alentadores. Durante su ejecución, se llevó a cabo una meticulosa selección de las variables de entrada, basada en un criterio bien definido y relevante para los objetivos del proyecto. Es importante destacar que este proceso de selección no condujo a un sobreajuste de la curva de potencia, lo que es fundamental para garantizar la precisión de las predicciones.

Uno de los logros más destacados de este proyecto es la capacidad de realizar predicciones precisas y confiables. Estas predicciones se han diseñado para categorizar y evaluar el rendimiento de manera efectiva. Esta categorización podría ser de gran utilidad para los técnicos y profesionales involucrados en el monitoreo y mantenimiento de las turbinas.

Es fundamental destacar que este proyecto estuvo enfocado en la búsqueda de la optimización y el incremento de la eficiencia en la operación de las turbinas. La precisión en la clasificación del rendimiento desempeña un papel crítico al identificar áreas de mejora y al facilitar la toma de decisiones en tiempo real. Los resultados obtenidos en esta investigación, más allá de su promisorio carácter, representan un avance significativo hacia el aumento de la eficiencia energética y la mejora de la rentabilidad en las actividades relacionadas con la energía eólica.

Finalmente, en este capítulo se comentará de forma más detallada los resultados obtenidos en los capítulos anteriores.

6.1 Resultados. Selección de variables de entrada

En el análisis de resultados, uno de los primeros pasos cruciales fue la creación de una matriz de correlación diseñada específicamente para el conjunto de datos. Esta matriz desempeñó un papel fundamental en la modelización de la curva de potencia al ayudar a la identificación de las relaciones entre las variables predictoras y la variable objetivo, que en este caso es la potencia de la turbina eólica. La información proporcionada por esta matriz reveló la intensidad y la dirección de estas relaciones, lo que resultó de gran utilidad en la selección de las variables más relevantes en el modelo y en la comprensión de cómo influyen en la potencia generada por la turbina. Este proceso fue esencial para tomar decisiones y garantizar la construcción de modelos más precisos.

Siguiendo esta metodología y estableciendo un umbral de correlación predefinido, se logró reducir el conjunto de variables de entrada a 12. Además, se otorga al usuario la flexibilidad de decidir qué variables de entrada incluir o excluir en el modelo. Este enfoque se caracteriza por su gran claridad y facilidad de comprensión, aunque cabe señalar que existen métodos más avanzados disponibles.

Si bien sería ideal mejorar el código para que itere automáticamente sobre las variables de entrada a lo largo de todo el proceso y, basándose en los resultados de las predicciones, sugiera las variables más relevantes, en esta fase inicial, se considera que este análisis inicial de selección de variables con criterio es plenamente válido.

Este enfoque proporciona una base sólida y comprensible para avanzar en la optimización del modelo y la toma de decisiones.

6.2 Resultados. Procesamiento y filtrado de datos

En este capítulo, se han presentado diversas técnicas para abordar el filtrado de outliers. Sin lugar a dudas, el primer paso es el filtrado basado en el "estado operacional" de la turbina, y esto resulta crítico para obtener resultados finales de calidad, ya que permite eliminar una gran cantidad de valores atípicos y dar sentido a la forma de la curva.

Además, aunque este proceso de filtrado se aplica en el modelado de la curva, el filtrado basado en límites de potencia también desempeña un papel fundamental. La configuración de estos límites puede variar según el modelo de la turbina, su potencia nominal o las necesidades específicas del usuario. La exclusión de valores por debajo de 10 kW ha demostrado ser una decisión acertada, especialmente dado el comportamiento impredecible asociado con el arranque y la parada de una turbina eólica.

Por último, entre las técnicas consideradas, se ha optado por utilizar el método de "Isolation Forest" para el filtrado de outliers. Esta elección se debe a su capacidad excepcional para identificar de manera eficiente valores inusuales sin hacer suposiciones concretas sobre la distribución de los datos. El uso de este enfoque mejora significativamente la calidad de los datos, lo que a su vez conduce a un modelado más preciso y confiable de la curva de potencia. Además, se destaca por su capacidad para preservar la forma natural de la curva, su facilidad de interpretación y su idoneidad para lidiar con datos multidimensionales, como las curvas de potencia que dependen de diversas variables ambientales y operativas.

El modelado de la curva se realiza con un conjunto de datos que consta de 11617 observaciones, seleccionadas de un total de 42476.

6.3 Resultados. Modelado de la curva

En el proceso de modelado de la curva, se ha evaluado varias técnicas y, finalmente, se ha optado por utilizar el algoritmo ExtraTreesRegressor debido a que proporciona las predicciones más precisas en comparación con otras opciones.

En un análisis inicial de las métricas de rendimiento de diferentes modelos, ExtraTreesRegressor sobresale claramente en términos de precisión. Comparándolo con dos otros métodos, sus métricas demuestran un rendimiento superior:

Method	Mean Absolute Error	Mean Squared Error	Root Mean Squared Error	Accuracy on Training set	Accuracy on Testing set
ExtraTreesRegressor	10.062754	267.431532	16.353334	1.00000	0.995359
DecisionTreeRegressor	14.698545	624.267606	24.985348	1.00000	0.989167
KNeighborsRegressor	30.595733	2064.881928	45.440972	0.97721	0.964168

Tabla 4. Métricas obtenidas en las técnicas estudiadas

Este análisis inicial respalda la elección del “ExtraTreesRegressor” debido a su capacidad para generar predicciones altamente precisas.

Además, una vez que el modelo ha sido entrenado y aplicado a la totalidad de los datos, que han sido previamente filtrados para asegurar que la turbina esté en condiciones normales de operación, se observa que las predicciones resultantes son muy acertadas. Esto indica que el modelo no solo es preciso en la fase de entrenamiento y validación, sino que también es efectivo al enfrentar datos en situaciones reales, lo que es esencial para su utilidad práctica.

En resumen, la elección del ExtraTreesRegressor se basa en su destacado rendimiento en métricas de precisión y en su capacidad para generar predicciones precisas en condiciones de operación reales, lo que lo convierte en una opción sólida para el modelado de la curva de potencia en este contexto.

6.4 Resultados. Evaluación del rendimiento de la turbina

En este último capítulo, se ha implementado una categorización de la turbina en tres niveles de rendimiento basados en las predicciones obtenidas durante el proceso de modelado. Esta categorización permitirá llevar a cabo un monitoreo proactivo y anticipar posibles problemas en la operación de la turbina.

Los resultados obtenidos en esta fase han superado las expectativas iniciales del proyecto. Este algoritmo, mediante mejoras continuas, la incorporación de nuevas técnicas y la refinación de detalles, podría convertirse en un indicador clave de rendimiento (KPI) para la turbina en tiempo real.

De hecho, al analizar la figura 33, se puede apreciar que casi el 98% de las predicciones se alinean con el rendimiento esperado de la turbina. Solo en un pequeño conjunto de 50 muestras las predicciones parecen distanciarse significativamente de los valores reales.

Es importante destacar que, en condiciones de operación reales, esta discrepancia estará más equilibrada, lo que indica un funcionamiento óptimo. Por ejemplo, cuando la turbina presente alarmas que afecten a su producción o esté siendo regulada o desalineada, el indicador proporcionará una alerta a los operadores, señalando que algo inusual está ocurriendo con la turbina.

En resumen, este enfoque de categorización y monitoreo ha demostrado ser altamente efectivo en el seguimiento y mantenimiento proactivo de la turbina, y su potencial como KPI en tiempo real sugiere que podría convertirse en una herramienta esencial para optimizar la operación y la eficiencia en el campo de la energía eólica.

7 TRABAJOS FUTUROS

En el contexto del modelado de la curva de potencia mediante técnicas de aprendizaje automático, se presentan una serie de desafíos y oportunidades que orientarán las futuras investigaciones y mejoras en la herramienta analítica. En este entorno de vanguardia, se considera la búsqueda de soluciones innovadoras y la optimización de los procesos de generación de energía eólica como fundamentales. Se han identificado varios aspectos clave que merecen atención especial y que marcarán la dirección de las investigaciones futuras.

Cabe destacar que esta propuesta de trabajo se presentará formalmente a una empresa proveedora de SCADA, con el objetivo de implementar estas mejoras en el entorno de producción. Esto no solo permitirá una colaboración valiosa, sino que también contribuirá a la eficiencia y la calidad de la generación de energía eólica en las instalaciones que gestione el software.

En primer lugar, se contempla la posibilidad de habilitar la configuración de todos los parámetros y atributos que definen cada algoritmo de aprendizaje automático como "inputs" en el entorno de producción del algoritmo. Esta medida otorgaría a los clientes mayor independencia, permitiéndoles externalizar proyectos analíticos o utilizar equipos especializados en análisis de datos para adaptar la herramienta a sus necesidades específicas.

Adicionalmente, se busca dotar a la herramienta de la capacidad de determinar de manera automática la técnica más apropiada y los parámetros de configuración óptimos en función de los "inputs" previamente definidos. Esto simplificaría la toma de decisiones y optimizaría el proceso de modelado, mejorando la utilización de los recursos disponibles.

Una vez establecida la técnica y los parámetros óptimos, se procederá a la implementación de la función de predicción de la telemedida de tiempo real (TTR) en el sistema SCADA. Esta información, combinada con la técnica previamente establecida en el proyecto de "rendimiento" de la turbina, permitirá al cliente determinar si la turbina está operando por debajo de su capacidad potencial.

Es importante destacar que la implementación de esta predicción no se limitará únicamente a los datos del SCADA, sino que se extenderá a las señales enviadas a REE (Red Eléctrica Española). Existen Protocolos de Operación (PO) en los que es esencial regular la producción, y cuando se realiza dicha regulación, es necesario proporcionar la Potencia Producible a REE. Esto se debe a que REE es responsable de compensar a los productores por la energía que dejan de generar debido a discrepancias con las predicciones y las ventas acordadas el Mercado Elétrico.

Por lo general, el algoritmo que calcula la potencia producible realiza interpolaciones basadas en la curva de potencia teórica, suponiendo que se trata del escenario óptimo, sin considerar el desgaste de la turbina ni las condiciones ambientales ideales. Sin embargo, según la documentación revisada, esta suposición no siempre se cumple. En algunas ocasiones, los resultados de producción superan la curva de potencia teórica. Este tipo de predicciones será más común en turbinas que llevan poco tiempo en operación que han recibido un mantenimiento adecuado. Sin embargo, esto no es aplicable al modelo seleccionado para el proyecto actual.

En otras palabras, se enviará el valor calculado por el algoritmo convencional de la potencia producible siempre que la predicción sea menor que este. En el caso de que la predicción supere la curva de potencia teórica, se enviará el valor de la predicción. Esto podría resultar en beneficios económicos significativos para el productor.

Además, se contemplará la posibilidad de integrar el historial de alarmas en las técnicas de aprendizaje automático utilizadas con el objetivo de predecir las alarmas que podrían surgir como resultado de las predicciones. Esta integración permitirá anticipar y abordar problemas operativos de manera proactiva.

En el ámbito de las mejoras adicionales, se planea la incorporación de técnicas de deep learning al proyecto. Esto abrirá la puerta a la exploración de modelos más complejos y la identificación de patrones más profundos en los datos, lo que mejoraría la precisión de las predicciones.

Por otro lado, considerando que algunos SCADAs pueden realizar reinicios automáticos en las turbinas, se aspira a lograr una integración efectiva de ambas herramientas. Esto implicaría que el SCADA informe automáticamente al usuario cuando exista una alta probabilidad de que la turbina deba reiniciarse, proporcionando detalles sobre las variables de entrada y las alarmas relacionadas con esta predicción.

Finalmente, se plantea la programación periódica y escalonada de la actualización de la curva de potencia con los nuevos datos de entrenamiento. Esta práctica garantizará que el modelo se mantenga actualizado y se adapte a cambios en las condiciones operativas y ambientales a lo largo del tiempo.

Estos esfuerzos futuros tienen como objetivo mejorar la flexibilidad, precisión y utilidad de la herramienta de modelado de la curva de potencia, permitiendo a los clientes optimizar la operación de aerogeneradores y aprovechar al máximo su capacidad de generación de energía.

REFERENCIAS

[1] Badihi, H.; Bilendo, F.; Lu, N. Wind Turbine Anomaly Detection Based on SCADA Data. Handb. Smart Energy Syst, 2022

[2] Ditkovich, Y.; Kuperman, A.; Byalsky, M.; Yahalom, A. A Generalized Approach to Estimating Capacity Factor of Fixed Speed Wind Turbines. IEEE Trans. Sustain. Energy, 2012

[3] Gill, S.; Galloway, S.; Stephen, B. Wind Turbine Condition Assessment Through Power Curve Copula Modeling. IEEE Trans. Sustain. Energy, 2011

[4] Hu, S.-Y.; Cheng, J.-H. Performance evaluation of pairing between sites and wind turbines. Renew. Energy, 2007

[5] IRENA. Renewable Power Generation Costs in 2021; International Renewable Energy Agency: Abu Dhabi, United Arab Emirates, 2022

[6] Lydia, M.; Kumar, S.S.; Selvakumar, A.I.; Kumar, G.E.P. A comprehensive review on wind turbine power curve modeling techniques. Renew. Sustain. Energy Rev, 2014

[7] Lydia, M.; Selvakumar, A.I.; Kumar, S.S.; Kumar, G.E.P. Advanced Algorithms for Wind Turbine Power Curve Modeling. IEEE Trans. Sustain. Energy, 2013

[8] Shen, X.; Fu, X.; Zhou, C. A Combined Algorithm for Cleaning Abnormal Data of Wind Turbine Power Curve Based on Change Point Grouping Algorithm and Quartile Algorithm. IEEE Trans. Sustain. Energy, 2018

[9] Tautz-Weinert, J.; Watson, S. Using SCADA data for wind turbine condition monitoring—A review. IET Renew. Power Gener, 2016

[10] https://docs.python.org/es/3.10/library/os.html

[11] https://github.com/

[12] https://kaggle.com/

[13] https://matplotlib.org/

[14] https://numpy.org/

[15] https://pandas.pydata.org/

[16] https://scikit-learn.org/stable/

[17] https://scipy.org/

[18] https://seaborn.pydata.org/

[19] https://stackoverflow.com/

[20] https://www.python.org/

[21] https://www.reddit.com/

[22] https://www.thewindpower.net/turbine_es_50_made_ae-52.php

[23]https://www.vectorenewables.com/es/recursos/blog/sabes-lo-que-es-el-efecto-estela-en-un-parque-eolico#:~:text=El%20'efecto%20estela'%20es%20el,a%20los%20parques%20e%C3%B3licos%20vecinos