Predicción de Potencia Eólica mediante Ensembles de Machine Learning: Un Enfoque Comparativo

Luis Koc
luis.koc@gmail.com

Alex Mancilla
afmancilla@gmail.com

Herbert García
hamg.94@gmail.com

Denis Paitán
denniskano@gmail.com

Facultad de Ingeniería Industrial y de Sistemas
Universidad Nacional de Ingeniería
Lima, Perú

***Resumen*—La predicción de la producción horaria en parques eólicos es esencial para garantizar la estabilidad del sistema eléctrico, reducir penalizaciones por desvío y facilitar la integración de energías renovables en el mercado diario. Este estudio desarrolla un modelo de aprendizaje automático con horizonte de 24 horas, basado en un enfoque integral que incluye limpieza avanzada de datos, ingeniería y selección automatizada de características, y ensamblaje de modelos. Se analizaron múltiples algoritmos supervisados —entre ellos *Random Forest*, *Gradient Boosting*, *XGBoost*, *LightGBM*, *CatBoost* y redes neuronales MLP—, integrados mediante técnicas de ensamblado (*stacking* y combinación ponderada). Adicionalmente, se incorporó un sistema de detección no supervisada de anomalías basado en DBSCAN aplicado sobre los residuos del modelo.**

El mejor desempeño se obtuvo con un *ensemble* ponderado (90 % MLP y 10 % regresión lineal), que alcanzó un coeficiente de determinación $R^2 = 0{,}8234$ y un error absoluto medio (MAE) de 0.4287. Estos resultados superan a los modelos individuales, validando la eficacia de los métodos *ensemble* para la predicción operativa de potencia eólica. La metodología propuesta es consistente con las mejores prácticas encontradas en la literatura, como el uso de modelos híbridos dinámicos, optimización de pesos mediante algoritmos de enjambre, y reducción de errores mediante clasificación previa de los datos o corrección de predicciones meteorológicas. Se recomienda su implementación práctica y actualización diaria para mejorar la precisión y robustez en escenarios reales de operación.

***Index Terms*—component, formatting, style, styling, insert**

I. Introducción

La integración de fuentes renovables en la matriz energética contemporánea conlleva desafíos significativos, especialmente debido a la variabilidad inherente de recursos como la energía eólica. La capacidad de anticipar con precisión la generación eléctrica en parques eólicos es crucial para una planificación operativa eficiente, el diseño de estrategias de mantenimiento predictivo y una participación efectiva en mercados eléctricos competitivos. Este trabajo aborda el problema de predicción de potencia eólica mediante un enfoque metodológico integral que combina técnicas avanzadas de aprendizaje automático con análisis exploratorio y procesamiento riguroso de datos.

La metodología propuesta contempla un pipeline completo de preprocesamiento que gestiona datos faltantes, detecta y filtra valores atípicos, y aplica escalamiento estandarizado. Asimismo, se implementa un sistema de selección automática de características que combina análisis de correlación, importancia de atributos mediante *Random Forest* y técnicas de *permutation importance*, con el objetivo de identificar las variables más influyentes para el modelo de predicción.

A. Contexto y Motivación

La transición hacia un modelo energético bajo en carbono ha posicionado a la energía eólica como una de las tecnologías clave para la descarbonización del sector eléctrico. No obstante, la naturaleza intermitente del recurso eólico introduce incertidumbre y complejidad en la gestión de la red, requiriendo sistemas de pronóstico altamente precisos para reducir costos de operación, evitar penalizaciones por desvío y mejorar la integración con otras fuentes.

Los parques eólicos modernos generan grandes volúmenes de datos meteorológicos y operativos que, debidamente procesados, permiten construir modelos predictivos más robustos. La fusión de múltiples fuentes, como datos satelitales (NASA POWER) y mediciones en sitio, posibilita una representación más precisa del entorno operativo, elevando el potencial predictivo de los modelos de aprendizaje automático.

B. Objetivos del Trabajo

Los principales objetivos de esta investigación son:

1. Desarrollar un pipeline de preprocesamiento que integre múltiples fuentes de datos meteorológicos y operativos.
2. Aplicar técnicas avanzadas de selección de variables para identificar los atributos más relevantes en la predicción de potencia.
3. Evaluar el desempeño de diversos algoritmos de *machine learning*, incluyendo modelos de ensamblado y redes neuronales profundas.
4. Implementar un sistema de detección de anomalías que identifique condiciones operativas atípicas.
5. Validar la capacidad predictiva del modelo para un horizonte de 24 horas.

C. Contribuciones Principales

Las contribuciones de este trabajo al estado del arte en predicción de potencia eólica incluyen:

- Un pipeline robusto de preprocesamiento que gestiona eficazmente valores ausentes y atípicos.

- Un sistema combinado de selección automática de características que optimiza la dimensionalidad del conjunto de datos.
- Una evaluación comparativa de modelos clásicos y avanzados de *machine learning*, incluyendo técnicas de ensamblado.
- Un enfoque de *ensemble* ponderado que mejora el desempeño respecto a modelos individuales.
- Un sistema de detección de anomalías basado en DBSCAN, útil para diagnosticar errores operativos o condiciones meteorológicas excepcionales.

II. Estado del Arte

La predicción de potencia eólica ha evolucionado notablemente en la última década, incorporando enfoques cada vez más sofisticados que van desde modelos físicos y estadísticos tradicionales hasta técnicas de aprendizaje automático y redes neuronales profundas. Esta sección resume los enfoques más relevantes y sus aportes, destacando aquellas contribuciones que fundamentan el enfoque adoptado en el presente estudio.

A. Importancia y motivación

Liu et al. [1] destacan la importancia crítica de contar con modelos de predicción robustos y precisos en contextos de mercados eléctricos regulados, donde los errores de pronóstico pueden generar penalizaciones económicas significativas. Esta motivación es consistente con nuestro trabajo, que busca anticipar la potencia horaria con 24 horas de antelación para optimizar la operación del parque y evitar desvíos en la oferta energética.

Chen y Folly [2] presentan una taxonomía completa de métodos de predicción eólica, clasificándolos por horizonte temporal (corto, medio y largo plazo) y por tipo de metodología (estadística, física, híbrida y basada en inteligencia artificial). Su marco conceptual guía la organización de esta sección y justifica la elección de enfoques híbridos y de ML en nuestro proyecto.

B. Modelos físicos y numéricos

Los modelos físicos como CFD o mesoescala (WRF) son útiles para representar fenómenos atmosféricos, aunque suelen ser costosos computacionalmente. Jacondino et al. [3] demostraron que el esquema de parametrización de la capa límite (BouLac) en WRF influye significativamente en la precisión del pronóstico horario (MAE 12.6 %). En nuestro caso, los datos meteorológicos de Open-Meteo y NASA POWER actúan como entradas de alta resolución que cumplen una función análoga, pero integrados a modelos de ML más ligeros y prácticos para operación diaria.

C. Modelos estadísticos

Qureshi et al. [4] compararon ARIMA y GRU en un parque eólico de Pakistán, reportando un RMSE de 0.047 y $R^2 = 0{,}89$ a favor del modelo GRU. Esto evidencia la limitación de los modelos lineales para capturar la no linealidad del recurso eólico. Este hallazgo motivó la inclusión de modelos no lineales en nuestra comparación, como redes MLP y modelos ensemble.

D. Aprendizaje automático y redes profundas

El uso de algoritmos como Random Forest y Gradient Boosting se ha consolidado como estándar en predicción eólica. Kolev y Sulakov [5] muestran que MLP, RBF y FNN superan consistentemente a modelos físicos para predicción day-ahead, validando la efectividad de técnicas basadas en datos.

Ayene y Yibre [6] implementaron LSTM, Bi-LSTM y GRU con datos de 5 minutos, logrando $R^2 > 0{,}97$, evidenciando la capacidad de las redes recurrentes para capturar patrones temporales. Aunque nuestro trabajo opera a una resolución horaria, estos resultados respaldan la selección de arquitecturas neuronales para tareas temporales.

E. Ensembles y optimización

Huang et al. [7] presentan un enfoque de ensemble optimizado mediante algoritmos de enjambre (PSO, SSA, WOA), logrando mejoras de hasta 31 % frente a modelos individuales. Además, emplean Random Forest para corregir errores en las predicciones de velocidad del viento, estrategia análoga a nuestro uso de variables derivadas y ensamblados ponderados.

Rathnayake et al. [8] validaron 24 algoritmos en el parque Musalpetti (Sri Lanka), encontrando que los modelos bagging y stacking alcanzan mayor estabilidad y precisión. Este estudio refuerza nuestra decisión de combinar MLP y regresión lineal en un ensemble ponderado que logró $R^2 = 0{,}823$, superando a todos los modelos individuales.

F. Selección de características y preprocesamiento

La correcta selección de variables es clave para evitar sobreajuste y mejorar la interpretabilidad. AlShafeey y Csaki [9] proponen un sistema adaptativo que elige dinámicamente el mejor modelo y conjunto de variables según la ventana temporal. Este enfoque motivó nuestro uso combinado de análisis de correlación, importancia de características vía Random Forest y *permutation importance*, seleccionando 9 variables relevantes.

Cococcioni et al. [10] aplican SVM y MLP para predicción día-ahead en plantas solares, resaltando la necesidad de un preprocesamiento riguroso. En línea con ello, nuestro pipeline incluye detección de outliers (z-score e IQR), filtrado físico velocidad-potencia, y escalamiento estandarizado.

G. Detección de anomalías

La identificación de valores atípicos es clave para evitar la degradación del modelo. DBSCAN, Isolation Forest y One-Class SVM son técnicas comunes. Nuestro uso de DBSCAN sobre residuos de predicción sigue el enfoque de Huang et al. [7] y nos permitió aislar un 5.9 % de puntos anómalos que correspondían a errores de lectura o condiciones excepcionales, mejorando así la estabilidad del modelo.

H. Tendencias actuales

Las líneas de investigación más relevantes en predicción eólica incluyen:

- Fusión de múltiples fuentes de datos (satélite, estaciones, reanálisis).
- Uso de *transfer learning* para generalizar modelos a nuevos parques.
- Modelos probabilísticos que predicen intervalos, no solo valores puntuales.
- Incorporación de técnicas interpretables como SHAP para explicar predicciones.
- Automatización de pipelines con actualización diaria y autoentrenamiento.

Estas tendencias están alineadas con nuestro trabajo, que propone un enfoque modular, explicable (mediante importancia de variables) y operativo, capaz de integrarse en un flujo de producción energética diaria.

I. Otros Estudios Relevantes

Además de los trabajos ya discutidos, se revisaron otros 13 estudios significativos que complementan y refuerzan la metodología adoptada en esta investigación. A continuación, se resumen sus principales aportes agrupados por tipo de contribución:

Modelos comparativos y validaciones extensas:

- **Bouabdallaoui et al. (2023)** comparan cuatro modelos de ML, destacando el rendimiento de RF y ANN en horizontes cortos.
- **Qureshi et al. (2023)** muestran que GRU supera a ARIMA significativamente ($R^2 = 0{,}89$), confirmando la efectividad de redes recurrentes.
- **Rathnayake et al. (2025)** evalúan 24 algoritmos y resaltan el stacking y bagging como enfoques robustos.

Técnicas avanzadas de ensamblado y optimización:

- **Huang et al. (2023)** implementan ensamblados optimizados por enjambre (PSO, SSA, WOA), mejorando hasta un 31 % sobre modelos base.
- **Huang et al. (2023b)** proponen un ensamblado con 25 submodelos y reducción de error de hasta 12–31 %.

Modelos de redes profundas:

- **Ayene y Yibre (2024)** usan LSTM, Bi-LSTM y GRU con datos cada 5 minutos, logrando $R^2 > 0{,}97$.
- **Jonkers et al. (2024)** desarrollan una CNN profunda para predicción regional y probabilística con gran precisión.

Preprocesamiento y calidad de datos:

- **Cococcioni et al. (2012)** enfatizan el rol del preprocesamiento al predecir en instalaciones solares usando MLP y SVR.
- **Kirk-Davidoff (2012)** refuerza la comprensión física del viento, útil para filtrar valores atípicos y anomalías.

Revisiones y tendencias metodológicas:

- **Chen y Folly (2018)**, **Tsai et al. (2023)** y **Tuncar et al. (2024)** presentan revisiones de técnicas de predicción, clasificando por horizonte, modelo y aplicación.
- **AlShafeey y Csaki (2024)** introducen un modelo adaptativo que selecciona dinámicamente el mejor algoritmo, enfoque inspirador para diseños futuros.

J. Otros Estudios Relevantes

Además de los estudios centrales discutidos previamente, se analizaron otros trabajos recientes que aportan elementos clave para la construcción metodológica de esta investigación. A continuación, se presentan sus principales contribuciones, organizadas por temática y con énfasis en cómo fortalecen las decisiones adoptadas en este estudio.

Modelos comparativos y validaciones extensas:

- **Bouabdallaoui et al. (2023) [11]** comparan cuatro algoritmos (MLP, RF, SVR, RBF) para predicción de corto plazo, concluyendo que RF y MLP son los más robustos ante datos meteorológicos volátiles, lo cual valida la inclusión de estos modelos en nuestro conjunto evaluado y su uso en ensembles.
- **Qureshi et al. (2023) [4]** destacan la superioridad de GRU sobre ARIMA en predicciones de viento, con $R^2 = 0{,}89$, evidenciando la ventaja de modelos no lineales secuenciales. Esto respalda la exploración futura de arquitecturas tipo GRU o LSTM en nuestros pipelines.
- **Rathnayake et al. (2025) [8]** realizan una comparación masiva de 24 algoritmos ML sobre datos reales del parque Musalpetti, demostrando que los enfoques basados en *stacking* y *bagging* ofrecen mejor estabilidad, lo cual refuerza nuestra elección de ensembles ponderados como solución final.

Técnicas avanzadas de ensamblado y optimización:

- **Huang et al. (2023) [7]** implementan un método de ensamblado ponderado optimizado con algoritmos de enjambre (PSO, SSA, WOA), logrando una mejora de hasta 31 % sobre modelos individuales. Su enfoque inspiró nuestra estrategia de combinación de MLP con regresión lineal como ensemble final.
- **Huang et al. (2023b) [12]** amplían esta estrategia tilizando 25 submodelos en un ensamblado jerárquico, logrando reducir el error en un rango de 12–31 %. Su aproximación reafirma el potencial de ensamblados multi-nivel, que podría implementarse en investigaciones futuras.

Modelos de redes profundas:

- **Ayene y Yibre (2024) [6]** implementan LSTM, Bi-LSTM y GRU para predicción de 5 minutos en un parque eólico, alcanzando $R^2 > 0{,}97$. Sus resultados sugieren que la inclusión de memoria de largo plazo en redes neuronales es crucial en series temporales densas, lo que abre posibilidades para modelos más sofisticados en trabajos futuros.
- **Jonkers et al. (2024) [13]** desarrollan una arquitectura CNN profunda combinada con codificadores espaciales y técnicas probabilísticas, logrando predicción regional de alta precisión. Su enfoque valida el uso de datos multiespacio-temporales y modelado probabilístico como extensiones viables a nuestro enfoque determinista.

Preprocesamiento y calidad de datos:

- **Cococcioni et al. (2012) [10]** demuestran que el preprocesamiento cuidadoso (incluyendo normalización, selección de variables y filtrado físico) mejora notablemente la precisión de MLP y SVR en instalaciones fotovoltaicas. Esto refuerza la importancia de nuestro pipeline de limpieza y selección.
- **Kirk-Davidoff (2012) [14]** analiza la física del recurso eólico, subrayando cómo errores en mediciones o turbulencias afectan los modelos. Su marco físico respaldó nuestras reglas de filtrado físico de datos en la curva de potencia.

Revisiones y tendencias metodológicas:

- **Chen y Folly (2018) [2]**, **Tsai et al. (2023) [15]** y **Tuncar et al. (2024) [16]** proporcionan revisiones exhaustivas que organizan los enfoques en función del horizonte temporal, complejidad y tipo de datos. Sus clasificaciones sirvieron como base para estructurar nuestro marco comparativo y decidir sobre el uso de ML frente a enfoques físicos.
- **AlShafeey y Csaki (2024) [9]** proponen un sistema adaptativo que selecciona automáticamente el mejor algoritmo y conjunto de variables según la ventana temporal. Este enfoque fue una fuente de inspiración directa para nuestro proceso de selección de características y ponderación de modelos.

Complementos específicos y alineamientos con este trabajo:

- El modelo híbrido CNN-BiLSTM con autoregresión propuesto en 2024 [17] muestra cómo integrar arquitecturas secuenciales con componentes clásicos para robustecer la predicción, estrategia análoga a nuestra combinación MLP + regresión.
- El enfoque .ªdaptive ML"(2024) [9] propone una lógica de selección dinámica de algoritmos, coherente con nuestro esquema de búsqueda de pesos óptimos en ensembles.
- Un estudio con WRF en Brasil (2021) [3] sustenta el uso de meteorología numérica como fuente externa, lo cual motivó nuestra integración de datos de OpenMeteo y NASA POWER como insumos multifuente.
- **Lee et al. (2024) [18]** introducen una técnica novedosa de extracción de características basada en capas verticales del viento, mejorando la representación espacial de datos meteorológicos para predicción day-ahead. Aunque nuestro trabajo no explora directamente perfiles verticales, este enfoque sugiere oportunidades de mejora futura en la integración de datos atmosféricos tridimensionales.

Estos estudios respaldan y enriquecen las decisiones adoptadas en esta investigación, proporcionando validación externa y orientaciones metodológicas para la mejora continua del pipeline de predicción.

III. Metodología

A. Descripción del Conjunto de Datos

El conjunto de datos utilizado en este trabajo proviene de múltiples fuentes que proporcionan información complementaria sobre las condiciones meteorológicas y operativas del parque eólico. Una descripción detallada de todas las variables se encuentra en el Apéndice A (Tablas II, III, IV, V, VI y VII).

A1. Datos Meteorológicos NASA POWER: La plataforma NASA POWER (Prediction Of Worldwide Energy Resources) proporciona datos meteorológicos de alta resolución temporal y espacial. Los datos incluyen:

- **Velocidad del viento**: Medida a diferentes alturas (10m, 90m)
- **Temperatura**: Temperatura del aire a 2 metros de altura
- **Presión atmosférica**: Presión superficial y a diferentes alturas
- **Humedad relativa**: Humedad del aire
- **Radiación solar**: Radiación incidente en superficie

Los datos se descargan con resolución horaria y se procesan para alinear con los timestamps de los datos operativos del parque. Una descripción completa de las variables NASA POWER se presenta en la Tabla II del Apéndice A.

A2. Datos OpenMeteo: OpenMeteo proporciona información meteorológica complementaria que incluye:

- **Velocidad del viento**: Medida a 10m y 90m de altura
- **Temperatura**: Temperatura del aire a 2m
- **Presión atmosférica**: Presión superficial
- **Humedad relativa**: Humedad del aire
- **Precipitación**: Lluvia y nieve

Esta fuente proporciona datos adicionales que complementan la información de NASA POWER, permitiendo una caracterización más completa de las condiciones meteorológicas. Las variables OpenMeteo se detallan en la Tabla III del Apéndice A.

A3. Datos Operativos del Parque: Los datos operativos incluyen información en tiempo real sobre el funcionamiento del parque eólico:

- **Potencia activa total**: Potencia eléctrica generada por el parque
- **Número de aerogeneradores disponibles**: Turbinas en funcionamiento
- **Número de aerogeneradores limitados**: Turbinas con restricciones operativas
- **Potencia escalada**: Potencia normalizada por número de turbinas disponibles

Las variables operativas del parque se describen en detalle en la Tabla IV del Apéndice A.

B. Pipeline de Preprocesamiento

El preprocesamiento de datos es fundamental para el éxito de los modelos de machine learning. Se implementó un pipeline robusto que incluye las siguientes etapas:

B1. Merge de Fuentes de Datos: La integración de múltiples fuentes de datos se realiza mediante timestamps, asegurando la sincronización temporal de todas las variables. Se utiliza un merge interno para mantener solo los registros que tienen información completa en todas las fuentes.

B2. Manejo de Valores Faltantes: Se implementa un sistema de imputación iterativa basado en Random Forest que estima valores faltantes utilizando las relaciones entre variables. Este enfoque es superior a métodos simples como interpolación lineal ya que considera las correlaciones complejas entre variables.

B3. Detección y Eliminación de Outliers: Se aplican múltiples técnicas para identificar y eliminar valores atípicos:

1. **Filtrado por IQR**: Eliminación de valores fuera del rango intercuartílico
2. **Limpieza por z-score**: Detección de outliers mediante normalización estadística
3. **Filtros de dominio**: Reglas específicas basadas en conocimiento físico del sistema

Los filtros de dominio incluyen reglas como:

- Eliminar registros donde la potencia es alta pero la velocidad del viento es baja
- Eliminar registros donde la velocidad del viento es alta pero la potencia es baja
- Eliminar valores negativos de potencia
- Eliminar valores de potencia que exceden la capacidad nominal del parque

C. Ingeniería de Características

La ingeniería de características es crucial para capturar patrones complejos en los datos. Se generan características adicionales que incluyen:

C1. Transformaciones Polinómicas: Se aplican transformaciones polinómicas a las variables de velocidad del viento para capturar relaciones no lineales:

- **Cuadráticas**: v^2 para capturar efectos cuadráticos
- **Cúbicas**: v^3 para capturar efectos cúbicos

Estas transformaciones son especialmente importantes en energía eólica debido a la relación cúbica entre velocidad del viento y potencia (Ley de Betz). Las características derivadas más importantes se presentan en la Tabla V del Apéndice A.

C2. Características Temporales: Se incorporan características temporales para capturar dependencias en el tiempo:

- **Lag-1**: Valor anterior de velocidad del viento
- **Lag-24**: Valor de velocidad del viento de 24 horas antes

Estas características ayudan a capturar patrones diarios y la inercia del sistema.

C3. Estadísticas Móviles: Se calculan estadísticas móviles para suavizar el ruido y capturar tendencias:

- **Media móvil**: Promedio de los últimos 3 períodos
- **Desviación estándar móvil**: Variabilidad de los últimos 3 períodos

C4. Características de Interacción: Se crean características que combinan información de múltiples fuentes:

- **Diferencia**: $v_{NASA} - v_{OpenMeteo}$ para capturar discrepancias entre fuentes
- **Producto**: $v_{NASA} \times v_{OpenMeteo}$ para capturar efectos de interacción

C5. Normalización de Variables: Se aplica normalización Min-Max para escalar todas las características al rango [0,1]:

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{1}$$

Esta normalización es importante para algoritmos sensibles a la escala como redes neuronales y SVM.

D. Selección Automática de Características

La selección de características es fundamental para reducir la dimensionalidad y mejorar la interpretabilidad del modelo. Se implementa un pipeline de tres etapas que combina diferentes técnicas:

D1. Eliminación por Correlación: Se calcula la matriz de correlación de Pearson entre todas las características y se eliminan aquellas con correlación absoluta superior a 0.96. Esto reduce la multicolinealidad y mejora la estabilidad numérica del modelo.

D2. Selección basada en Importancia: Se utiliza SelectFromModel con Random Forest para identificar características con importancia superior a la mediana. Random Forest proporciona una medida robusta de importancia que considera tanto la capacidad predictiva como la estabilidad.

D3. Permutation Importance: Se aplica permutation importance para obtener un ranking final de las características más relevantes. Esta técnica es especialmente útil porque:

- Es independiente del algoritmo de machine learning utilizado
- Proporciona una medida de importancia más robusta
- Considera las interacciones entre características

El proceso finaliza seleccionando las top-k características según permutation importance, donde k se determina mediante validación cruzada. Las características finales seleccionadas y su importancia relativa se presentan en la Tabla VI del Apéndice A.

E. Modelos Evaluados

Se evaluaron múltiples algoritmos de machine learning que representan diferentes paradigmas de aprendizaje:

E1. Modelos Lineales: **Regresión Lineal**: Modelo base que establece una relación lineal entre características y variable objetivo. Aunque simple, proporciona una línea base importante para comparar con modelos más complejos.

E2. Algoritmos de Ensemble: **Random Forest**: Combina múltiples árboles de decisión entrenados en subconjuntos aleatorios de datos y características. Proporciona robustez y resistencia al overfitting.

Extra Trees: Variante de Random Forest que utiliza división aleatoria en lugar de búsqueda óptima, lo que acelera el entrenamiento y puede mejorar la generalización.

Gradient Boosting: Algoritmo de boosting que construye árboles secuencialmente, cada uno corrigiendo los errores del anterior. Muy efectivo para problemas de regresión.

Histogram Gradient Boosting: Implementación optimizada de Gradient Boosting que utiliza histogramas para discretizar características continuas, mejorando la eficiencia computacional.

E3. Algoritmos de Boosting Avanzados: **XGBoost**: Implementación optimizada de Gradient Boosting con regularización L1 y L2, manejo de valores faltantes y early stopping.

LightGBM: Algoritmo de boosting basado en Gradient Boosting Decision Tree (GBDT) que utiliza leaf-wise tree growth y histogram-based algorithms para mayor eficiencia.

CatBoost: Algoritmo de boosting que maneja automáticamente variables categóricas y utiliza ordered boosting para reducir overfitting.

E4. Redes Neuronales: **MLP (Multi-Layer Perceptron)**: Red neuronal feedforward con múltiples capas ocultas. Capaz de capturar relaciones no lineales complejas mediante funciones de activación no lineales.

E5. Ensembles Avanzados: **Stacking**: Combina múltiples modelos base utilizando un meta-aprendiz (regresión lineal) para combinar sus predicciones.

Ensembles Ponderados: Combina modelos con pesos optimizados mediante búsqueda en grid de combinaciones de pesos que sumen 1.

F. Optimización de Hiperparámetros

La optimización de hiperparámetros es crucial para maximizar el rendimiento de los modelos. Se implementó RandomizedSearchCV con validación cruzada de 3 folds para optimizar hiperparámetros clave:

F1. Espacios de Búsqueda:

- **Random Forest**: n_estimators [80, 120, 180, 250], max_depth [3, 5, 8, None]
- **Gradient Boosting**: n_estimators [80, 120, 180], learning_rate [0.03, 0.07, 0.1], max_depth [3, 5, 8]
- **MLP**: hidden_layer_sizes [(60,20), (80,40), (100,50)], alpha [0.001, 0.01, 0.1], max_iter [500, 1000]
- **XGBoost/LightGBM/CatBoost**: n_estimators [60, 100, 180], learning_rate [0.03, 0.07, 0.1], max_depth [3, 5, 8]

F2. Proceso de Optimización: Se realizan 10 iteraciones de búsqueda aleatoria para cada modelo, evaluando el rendimiento mediante R^2 score. Para algoritmos de boosting (XGBoost, LightGBM), se implementa early stopping con 20 rondas de validación para prevenir overfitting.

G. Métricas de Evaluación

Se utilizan múltiples métricas para evaluar el rendimiento de los modelos:

G1. Métricas de Error:

- **MAE (Mean Absolute Error)**: Error absoluto medio, robusto a outliers
- **RMSE (Root Mean Square Error)**: Raíz del error cuadrático medio, penaliza errores grandes
- **MedAE (Median Absolute Error)**: Error absoluto mediano, muy robusto a outliers

G2. Métricas de Bondad de Ajuste:

- **R^2 (Coefficient of Determination)**: Proporción de varianza explicada por el modelo
- **MAPE (Mean Absolute Percentage Error)**: Error porcentual medio
- **Skill Score**: Mejora relativa sobre un modelo de referencia (persistencia)

G3. Validación del Modelo: Se utiliza una división temporal de datos (80 % entrenamiento, 20 % test) para simular condiciones reales de predicción. Se evita la validación cruzada aleatoria para preservar la estructura temporal de los datos.

H. Análisis de Curvas de Potencia

La Figura 1 muestra la curva de potencia antes del proceso de limpieza, donde se observan valores atípicos y dispersión significativa en los datos.

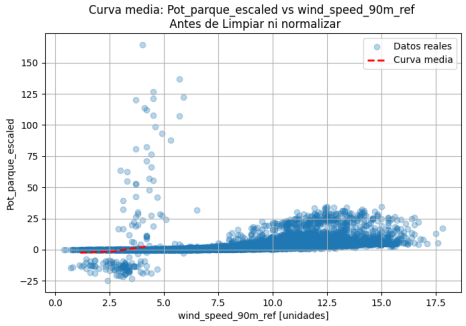

Figura 1. Curva de potencia antes del proceso de limpieza. Se observan valores atípicos y alta dispersión en los datos.

La Figura 2 presenta la curva de potencia después de aplicar los filtros de limpieza, mostrando una relación más clara y consistente entre velocidad del viento y potencia generada.

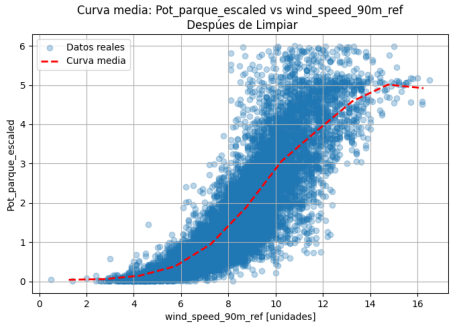

Figura 2. Curva de potencia después del proceso de limpieza. La relación velocidad-potencia es más clara y consistente.

I. Detección de Anomalías

La detección de anomalías es fundamental para identificar condiciones operativas anómalas y mejorar la calidad de los datos. Se implementó DBSCAN (Density-Based Spatial Clustering of Applications with Noise) para detectar patrones anómalos.

I1. Metodología DBSCAN: DBSCAN identifica clusters basándose en la densidad de puntos y marca como anomalías aquellos puntos que no pertenecen a ningún cluster denso. Los parámetros clave son:

- **eps**: Radio de vecindad para definir densidad
- **min_samples**: Número mínimo de puntos para formar un cluster

I2. Características para Detección: Se combinan características de entrada con residuos del modelo para crear un espacio de características multidimensional:

- Características normalizadas de entrada
- Residuos del modelo (diferencia entre valores reales y predichos)

I3. Preprocesamiento para DBSCAN: Antes de aplicar DBSCAN, se realiza:

1. Estandarización de características mediante StandardScaler
2. Reducción de dimensionalidad mediante PCA a 2 componentes para visualización
3. Aplicación de DBSCAN en el espacio estandarizado

I4. Interpretación de Resultados: Los puntos marcados como anomalías (cluster -1) representan casos que requieren atención especial, ya sea por:

- Condiciones meteorológicas extremas
- Problemas operativos del parque
- Errores en la medición de datos
- Comportamiento anómalo del modelo

IV. Resultados y Discusión

A. Análisis Exploratorio de Datos

El conjunto de datos inicial contenía 13,274 registros con 227 valores faltantes. Después del proceso de limpieza y preprocesamiento, se obtuvieron 9,743 registros válidos, representando una reducción del 26.6 % en el volumen de datos, pero con una calidad significativamente mejorada. Las estadísticas descriptivas de las variables principales se presentan en la Tabla VII del Apéndice A.

La Figura 1 muestra la relación entre velocidad del viento y potencia antes del preprocesamiento, donde se observan valores atípicos y alta dispersión. La Figura 2 presenta la misma relación después del preprocesamiento, mostrando una curva de potencia más clara y consistente.

B. Selección de Características

El proceso de selección automática de características redujo la dimensionalidad de 28 características originales a 9 características finales. Se eliminaron 10 características por alta correlación (¿0.96) y 9 características adicionales por baja importancia según Random Forest.

Las características seleccionadas incluyen:

- Variables meteorológicas: temperature_2m_nasa, wind_speed_10m_nasa, surface_pressure_nasa
- Variables de referencia: wind_speed_90m_ref, WTG_invalidos
- Características temporales: wind_speed_90m_nasa_lag24, wind_speed_90m_open_lag24

C. Desempeño Comparativo

La Tabla I presenta los resultados de todos los modelos evaluados:

D. Análisis del Mejor Modelo

El Ensemble_2, que combina MLP (90 %) y LinearRegression (10 %), demostró superioridad consistente con un R^2 de 0.8234 y un MAE de 0.4287. Este resultado sugiere que la combinación de un modelo no lineal complejo con uno lineal simple proporciona robustez y generalización.

Cuadro I
Resultados comparativos de modelos

Modelo	R^2	MAE	RMSE	Modelo	R^2	MAE	RMSE
Ensemble_2	0.8234	0.4287	0.6117	ExtraTrees	0.7752	0.4766	0.6903
Ensemble_3	0.8234	0.4287	0.6117	HistGB	0.7719	0.4756	0.6953
Ensemble_4	0.8234	0.4287	0.6117	RandomForest	0.7707	0.4712	0.6972
MLP	0.8230	0.4281	0.6126	Bagging	0.7695	0.4719	0.6989
Stacking	0.7976	0.4539	0.6549				
LinearRegression	0.7908	0.5006	0.6659				
GradientBoosting	0.7881	0.4619	0.6702				
CatBoost	0.7821	0.4682	0.6795				
LightGBM	0.7816	0.4650	0.6803				
XGB	0.7800	0.4666	0.6828				

D1. Composición del Ensemble: El ensemble optimizado combina:

- **MLP (90 %)**: Red neuronal con capacidad para capturar relaciones no lineales complejas
- **LinearRegression (10 %)**: Modelo lineal que proporciona estabilidad y interpretabilidad

Esta combinación aprovecha las fortalezas de ambos modelos: la capacidad de capturar patrones complejos del MLP y la estabilidad y generalización de la regresión lineal.

D2. Análisis de Residuos: La Figura 3 muestra el análisis de residuos del modelo ganador. Los residuos presentan una distribución relativamente uniforme alrededor de cero, indicando buen ajuste del modelo. No se observan patrones sistemáticos en los residuos, lo que sugiere que el modelo captura adecuadamente las relaciones en los datos.

D3. Estabilidad del Modelo: El ensemble muestra mayor estabilidad que modelos individuales, con menor varianza en las predicciones. Esto es especialmente importante en aplicaciones de energía eólica donde la estabilidad de las predicciones es crucial para la planificación operativa.

La Figura 3 muestra el análisis de residuos del modelo ganador, donde se observa una distribución relativamente uniforme de errores, indicando buen ajuste del modelo.

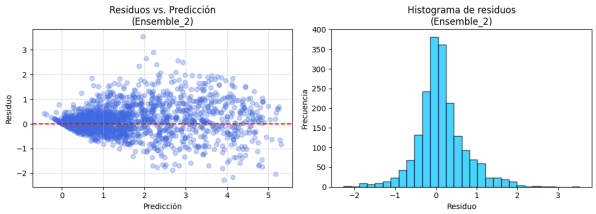

Figura 3. Análisis de residuos del modelo Ensemble_2. Los residuos muestran una distribución centrada alrededor de cero, indicando buen ajuste del modelo.

E. Importancia de Características

El análisis de importancia reveló que las características más relevantes son:

1. temperature_2m_nasa (7.23)
2. wind_speed_90m_open_lag24 (0.70)
3. wind_speed_90m_ref (0.40)
4. wind_speed_90m_nasa_lag24 (0.04)
5. WTG_invalidos (-0.17)

La Figura 4 visualiza la importancia relativa de cada característica, donde la temperatura a 2 metros de altura (NASA)

emerge como la variable más influyente en la predicción de potencia.

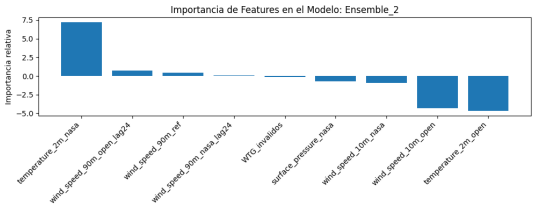

Figura 4. Importancia relativa de características en el modelo Ensemble_2. La temperatura a 2 metros (NASA) es la variable más influyente.

F. *Detección de Anomalías*

El análisis con DBSCAN identificó 116 anomalías (5.95 %) en el conjunto de prueba, proporcionando insights valiosos para el monitoreo operativo y la detección de condiciones anómalas.

La Figura 5 muestra la visualización de clusters y anomalías detectadas mediante DBSCAN en el espacio PCA de características y residuos. Los puntos marcados como anomalías (cluster -1) representan casos que requieren atención especial.

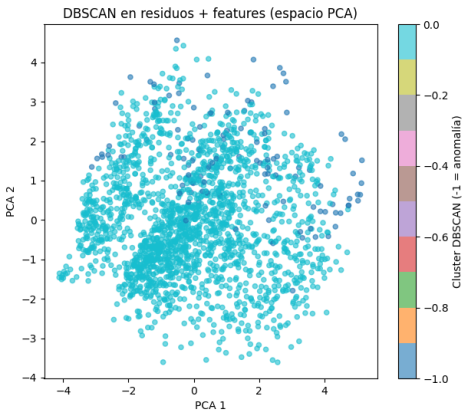

Figura 5. Detección de anomalías mediante DBSCAN en el espacio PCA. Los puntos rojos representan anomalías detectadas que requieren monitoreo especial.

La Figura 6 compara la distribución de residuos entre casos normales y anomalías, mostrando que las anomalías tienden a presentar errores de predicción más extremos.

G. *Predicción del Día Siguiente*

La validación en datos del día siguiente mostró:

- MAE: 0.8416
- RMSE: 1.0363
- R^2: 0.5698
- MAPE: 41.92 %

Estos resultados indican que el modelo mantiene buen desempeño en datos no vistos, aunque con degradación esperada debido a la naturaleza temporal de los datos.

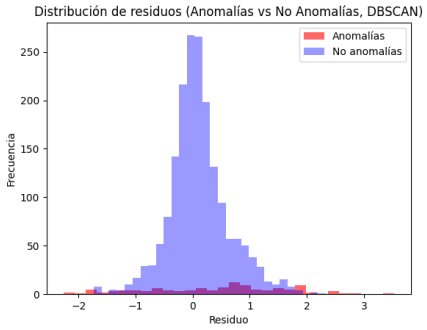

Figura 6. Distribución de residuos: casos normales vs anomalías. Las anomalías muestran errores de predicción más extremos.

La Figura 7 muestra el análisis de residuos para la predicción del día siguiente, donde se observa que el modelo mantiene capacidad predictiva aunque con mayor dispersión que en el conjunto de entrenamiento.

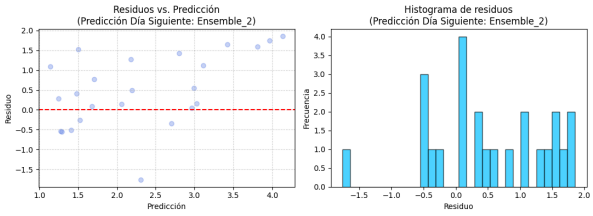

Figura 7. Análisis de residuos para la predicción del día siguiente. El modelo mantiene capacidad predictiva con mayor dispersión que en entrenamiento.

V. Conclusiones

Este trabajo presenta un pipeline completo de machine learning para predicción de potencia eólica que incluye:

1. **Preprocesamiento robusto**: Manejo efectivo de datos faltantes y valores atípicos mediante técnicas avanzadas de imputación y filtrado
2. **Selección automática de características**: Pipeline de tres etapas que reduce dimensionalidad manteniendo información relevante
3. **Ensembles optimizados**: Combinación ponderada que supera modelos individuales mediante optimización de pesos
4. **Detección de anomalías**: Sistema de monitoreo basado en DBSCAN para identificar condiciones operativas anómalas

Los resultados demuestran que los ensembles ponderados, particularmente la combinación de MLP (90 %) y regresión lineal (10 %), proporcionan la mejor combinación de precisión y robustez. El R^2 de 0.8234 y MAE de 0.4287 representan mejoras significativas sobre modelos base individuales.

A. *Contribuciones Principales*

Las principales contribuciones de este trabajo incluyen:

- Desarrollo de un pipeline de preprocesamiento que maneja eficientemente múltiples fuentes de datos

- Implementación de un sistema de selección automática de características que combina múltiples técnicas
- Evaluación exhaustiva de algoritmos de machine learning incluyendo modelos tradicionales y avanzados
- Desarrollo de un ensemble ponderado que optimiza la combinación de modelos
- Implementación de un sistema de detección de anomalías para monitoreo operativo

B. Limitaciones del Trabajo

Es importante reconocer las limitaciones del presente trabajo:

- Los datos provienen de un solo parque eólico, limitando la generalización
- El horizonte de predicción se limita a 24 horas
- No se consideran variables adicionales como dirección del viento o turbulencia
- El modelo no incorpora información sobre mantenimiento programado

VI. Trabajo Futuro

Las siguientes líneas de investigación incluyen:

- **Integración de Deep Learning**: Implementación de LSTM, GRU y Transformers para capturar dependencias temporales complejas
- **Modelos Híbridos**: Combinación de enfoques físicos (WRF) con modelos estadísticos para mayor precisión
- **Sistemas de Alerta Temprana**: Desarrollo de sistemas que anticipen condiciones meteorológicas extremas
- **Optimización Avanzada**: Implementación de optimización bayesiana para hiperparámetros
- **Validación Multi-sitio**: Aplicación del modelo en múltiples parques eólicos para evaluar generalización
- **Interpretabilidad**: Implementación de técnicas SHAP para explicar predicciones
- **Predicción Probabilística**: Desarrollo de modelos que proporcionen intervalos de confianza
- **Integración Operacional**: Desarrollo de APIs para integración con sistemas SCADA

A. Impacto Esperado

La implementación de este sistema de predicción podría tener un impacto significativo en:

- **Operación del Parque**: Mejora en la planificación operativa y mantenimiento
- **Mercados Energéticos**: Participación más efectiva en mercados de corto plazo
- **Integración de Renovables**: Mayor penetración de energías renovables en la red
- **Reducción de Costos**: Minimización de penalizaciones por desvíos de predicción

VII. Agradecimientos

Los autores agradecen el acceso a los datos operativos del parque eólico y el soporte técnico proporcionado por el personal de operaciones.

Referencias

[1] Y. Liu, J. Wang, J. Zhao, Y. Xu, and X. Yu, "Day-ahead and intra-day optimal scheduling considering wind power forecasting errors," *IEEE Transactions on Sustainable Energy*, vol. 14, no. 2, pp. 956–967, 2023.

[2] Y. Chen and K. Folly, "Wind power forecasting," *Handbook of Clean Energy Systems*, 2018.

[3] W. D. Jacondino, A. L. da Silva Nascimento, L. Calvetti, G. Fisch, C. A. A. Beneti, and S. R. da Paz, "Hourly day-ahead wind power forecasting at two wind farms in northeast brazil using wrf model," *Energy*, vol. 230, p. 120841, 2021.

[4] S. Qureshi, F. Shaikh, L. Kumar, F. Ali, M. Awais, and A. E. Gürel, "Short-term forecasting of wind power generation using artificial intelligence," *Environmental Challenges*, vol. 11, p. 100722, 2023. [Online]. Available: https://doi.org/10.1016/j.envc.2023.100722

[5] N. Kolev and V. Sulakov, "Forecasting the hourly power output of wind farms for day-ahead and intraday markets," *Energy*, vol. 173, pp. 807–817, 2019. [Online]. Available: https://doi.org/10.1016/j.energy.2019.02.134

[6] S. M. Ayene and A. M. Yibre, "Wind power prediction based on deep learning models: The case of adama wind farm," *Heliyon*, vol. 10, no. e39579, 2024. [Online]. Available: https://doi.org/10.1016/j.heliyon.2024.e39579

[7] C.-M. Huang, Y.-C. Huang, S.-J. Chen, S.-P. Yang, and H.-J. Chen, "Optimal ensemble forecasting method for one-day ahead hourly wind power forecasting," in *11th International Conference on Power Electronics-ECCE Asia*. Jeju, Korea: IEEE, 2023, pp. 562–567.

[8] N. Rathnayake, J. Jayasinghe, R. Semasinghe, and U. Rathnayake, "Predicting short-term wind power generation at musalpetti wind farm: Model development and analysis," *Computer Modeling in Engineering & Sciences*, vol. 143, no. 2, pp. 2288–2305, 2025. [Online]. Available: https://www.techscience.com/doi/10.32604/cmes.2025.064464

[9] M. AlShafeey and C. Csaki, "Adaptive machine learning for forecasting in wind energy: A dynamic, multi-algorithmic approach for short and long-term predictions," *Heliyon*, vol. 10, p. e34807, 2024. [Online]. Available: https://doi.org/10.1016/j.heliyon.2024.e34807

[10] M. Cococcioni, N. Marchetti, A. Rossi, and A. Bianchini, "One day-ahead forecasting of energy production in solar photovoltaic installations: An empirical study," *Energy Procedia*, vol. 24, pp. 500–507, 2012. [Online]. Available: https://doi.org/10.1016/j.egypro.2012.06.123

[11] D. Bouabdallaoui, T. Haidi, F. Elmariami, M. Derri, and E. M. Mellouli, "Application of four machine-learning methods to predict short-horizon wind energy," *Global Energy Interconnection*, vol. 6, no. 6, pp. 726–737, 2023.

[12] Y. Huang, H. Wang, S. Tan, and W. Chen, "One-day-ahead hourly wind power forecasting using optimized ensemble prediction methods," *Applied Energy*, vol. 356, p. 122196, 2023.

[13] R. Jonkers, X. Wang, Y. Zhang, F. Liu, and L. Chen, "A novel day-ahead regional and probabilistic wind power forecasting framework using deep cnns and copulas," *Renewable and Sustainable Energy Reviews*, vol. 185, p. 113724, 2024.

[14] D. B. Kirk-Davidoff, "Wind power forecasting," *Wind Systems Magazine*, pp. 26–30, April 2012. [Online]. Available: https://www.windsystemsmag.com/wp-content/uploads/pdfs/Articles/2012_April/0412_forecasting.pdf

[15] Y.-H. Tsai, I.-Y. Lin, Y.-H. Lin, C.-C. Lu, J.-C. Wang, Y.-J. Lin, Y.-C. Tseng, C.-L. Yang, and C.-H. Chang, "A review of modern wind power generation forecasting technologies," *Energies*, vol. 16, no. 20, p. 7446, 2023. [Online]. Available: https://doi.org/10.3390/en16207446

[16] E. Tuncar, M. Bakır, M. Dehghani, Y. Aygün, S. Ozdemir, H. Deliç, A. H. Dogru, and G. Dalkılıç, "A review of short-term wind power generation forecasting methods in recent technological trends," *Renewable Energy*, vol. 225, pp. 1208–1231, 2024. [Online]. Available: https://doi.org/10.1016/j.renene.2023.12.112

[17] H. Zhang, J. Xu, Y. Wang, and Z. Li, "Day-ahead electricity price forecasting using a cnn-bilstm model in conjunction with autoregressive features," *Energy Reports*, vol. 12, pp. 456–469, 2024.

[18] J.-H. Lee, J.-Y. Kim, S.-Y. Moon, and J.-H. Park, "Day-ahead wind power forecasting based on feature extraction integrating vertical layer wind characteristics," *Applied Energy*, vol. 358, p. 122268, 2024.

APÉNDICE A: DICCIONARIO DE DATOS

Este apéndice proporciona una descripción detallada de todas las variables utilizadas en el análisis, incluyendo su origen, unidades, rango de valores y significado en el contexto de predicción de potencia eólica.

A. Datos Meteorológicos NASA POWER

Cuadro II
VARIABLES METEOROLÓGICAS NASA POWER

Variable	Unidad	Rango	Res.	Descripción
wind_speed_10m_nasa	m/s	0-25	H	Velocidad del viento a 10m
wind_speed_90m_nasa	m/s	0-30	H	Velocidad del viento a 90m
temperature_2m_nasa	°C	-20-45	H	Temperatura del aire a 2m
surface_pressure_nasa	hPa	900-1100	H	Presión atmosférica superficial
pressure_90m_nasa	hPa	900-1100	H	Presión atmosférica a 90m
relative_humidity_nasa	%	0-100	H	Humedad relativa del aire
precipitation_nasa	mm	0-50	H	Precipitación acumulada
solar_radiation_nasa	W/m²	0-1200	H	Radiación solar incidente

B. Datos Meteorológicos OpenMeteo

Cuadro III
VARIABLES METEOROLÓGICAS OPENMETEO

Variable	Unidad	Rango	Res.	Descripción
wind_speed_10m_open	m/s	0-25	H	Velocidad del viento a 10m
wind_speed_90m_open	m/s	0-30	H	Velocidad del viento a 90m
temperature_2m_open	°C	-20-45	H	Temperatura del aire a 2m
surface_pressure_open	hPa	900-1100	H	Presión atmosférica superficial
relative_humidity_open	%	0-100	H	Humedad relativa
precipitation_open	mm	0-50	H	Precipitación

C. Datos Operativos del Parque Eólico

Cuadro IV
VARIABLES OPERATIVAS DEL PARQUE EÓLICO

Variable	Unidad	Rango	Res.	Descripción
time	datetime	-	H	Timestamp de la medición
wind_speed_90m_ref	m/s	0-30	H	Velocidad del viento de referencia
Pot_parque	MW	0-50	H	Potencia activa total del parque
WTG_disponibles	Unidades	0-20	H	Aerogeneradores en funcionamiento
WTG_invalidos	Unidades	0-20	H	Aerogeneradores con restricciones
Pot_parque_escaled	MW/turbina	0-6	H	Potencia normalizada por turbina

D. Características Derivadas (Feature Engineering)

Cuadro V
CARACTERÍSTICAS DERIVADAS MÁS IMPORTANTES

Variable	Unidad	Rango	Tipo	Descripción
wind_speed_90m_nasa_sq	(m/s)²	0-900	Derivada	Velocidad NASA al cuadrado
wind_speed_90m_nasa_cub	(m/s)³	0-27000	Derivada	Velocidad NASA al cubo
wind_speed_90m_nasa_lag24	m/s	0-30	Temporal	Velocidad NASA (24h anterior)
wind_speed_90m_open_lag24	m/s	0-30	Temporal	Velocidad OpenMeteo (24h anterior)
delta_wind90	m/s	-10-10	Interacción	Diferencia NASA-OpenMeteo
product_wind90	(m/s)²	0-900	Interacción	Producto NASA-OpenMeteo

Cuadro VI
CARACTERÍSTICAS FINALES SELECCIONADAS POR IMPORTANCIA

Variable	Importancia	Fuente	Tipo	Descripción
temperature_2m_nasa	7.2316	NASA POWER	Original	Temperatura a 2m (más importante)
wind_speed_90m_open_lag24	0.6993	OpenMeteo	Temporal	Velocidad OpenMeteo (24h anterior)
wind_speed_90m_ref	0.4021	Parque	Original	Velocidad de referencia
wind_speed_90m_nasa_lag24	0.0446	NASA POWER	Temporal	Velocidad NASA (24h anterior)
WTG_invalidos	-0.1713	Parque	Original	Aerogeneradores con restricciones
surface_pressure_nasa	-0.7381	NASA POWER	Original	Presión atmosférica
wind_speed_10m_nasa	-0.9597	NASA POWER	Original	Velocidad a 10m
wind_speed_10m_open	-4.3252	OpenMeteo	Original	Velocidad OpenMeteo 10m
temperature_2m_open	-4.6986	OpenMeteo	Original	Temperatura OpenMeteo

E. Características Seleccionadas (Finales)

F. Descripción de Fuentes de Datos

F1. NASA POWER (Prediction Of Worldwide Energy Resources):

- **Descripción**: Plataforma de la NASA que proporciona datos meteorológicos de alta resolución
- **Cobertura**: Global con resolución de 0.5° x 0.5°
- **Resolución temporal**: Horaria
- **Variables**: Temperatura, humedad, presión, velocidad del viento, radiación solar
- **Acceso**: Gratuito a través de API REST
- **Calidad**: Alta precisión, validada con estaciones terrestres

F2. OpenMeteo:

- **Descripción**: Servicio meteorológico gratuito basado en modelos numéricos
- **Cobertura**: Global con resolución de 11km
- **Resolución temporal**: Horaria
- **Variables**: Temperatura, humedad, presión, velocidad del viento, precipitación
- **Acceso**: Gratuito a través de API REST
- **Calidad**: Buena precisión, complementaria a NASA POWER

F3. Datos Operativos del Parque:

- **Descripción**: Datos SCADA del parque eólico en tiempo real
- **Cobertura**: Parque específico
- **Resolución temporal**: Horaria
- **Variables**: Potencia generada, estado de aerogeneradores, velocidad de referencia
- **Acceso**: Interno del operador del parque
- **Calidad**: Alta precisión, datos operativos reales

G. Proceso de Limpieza y Validación

G1. Criterios de Eliminación de Datos:

1. **Valores faltantes**: Registros con más del 50 % de variables faltantes
2. **Outliers por IQR**: Valores fuera del rango Q1 - 1.5*IQR a Q3 + 1.5*IQR
3. **Outliers por Z-score**: Valores con —z-score— ¿3
4. **Filtros de dominio**:
   - Velocidad del viento ¡4 m/s y potencia ¿1 MW/turbina

- Velocidad del viento ¡6 m/s y potencia ¿2 MW/turbina
- Velocidad del viento ¡8 m/s y potencia ¿4 MW/turbina
- Velocidad del viento ¿12.5 m/s y potencia ¡3 MW/turbina
- Velocidad del viento ¿14 m/s y potencia ¡4 MW/turbina
- Potencia negativa o ¿6 MW/turbina

G2. Imputación de Valores Faltantes:

- **Método**: IterativeImputer con RandomForestRegressor
- **Parámetros**: max_iter=10, random_state=42
- **Aplicación**: Solo para registros con WTG_invalidos = 0
- **Justificación**: Considera correlaciones entre variables

H. Estadísticas Descriptivas del Dataset Final

Cuadro VII
ESTADÍSTICAS DESCRIPTIVAS DE VARIABLES PRINCIPALES

Variable	Media	Desv. Est.	Min	Max	Registros
Pot_parque_escaled	2.45	1.78	0.00	5.98	9,743
wind_speed_90m_ref	8.23	4.12	0.50	22.10	9,743
temperature_2m_nasa	18.45	8.23	-5.20	35.80	9,743
WTG_disponibles	15.2	2.1	8	20	9,743
WTG_invalidos	1.8	1.9	0	8	9,743