Universidad Nacional de Loja

Facultad de la Energía, las Industrias y los de Recursos Naturales No Renovables

Carrera de Computación

Modelo híbrido de predicción de fallas para los aerogeneradores de la Central Eólica Villonaco utilizando inteligencia artificial y datos del sistema SCADA

Hybrid Fault Prediction Model for Wind Turbines at the Villonaco Wind Farm Employing Artificial Intelligence and SCADA System Data.

Trabajo de Integración Curricular, previo a la obtención del título de Ingeniero en Ciencias de la Computación

AUTOR:
Wagner Cristhoper Castillo Castro

DIRECTOR:
Ing. Pablo Fernando Ordoñez Ordoñez, Mg. Sc.

Loja - Ecuador
2024

Certificación

Loja, 29 de mayo de 2024

Ing. Pablo Fernando Ordoñez Ordoñez Mg.Sc.
DIRECTOR DE TRABAJO DE INTEGRACIÓN CURRICULAR

CERTIFICO:

Que he revisado y orientado todo el proceso de elaboración del Trabajo de Integración Curricular denominado: **Modelo híbrido de predicción de fallas para los aerogeneradores de la Central Eólica Villonaco utilizando inteligencia artificial y datos del sistema SCADA**, previo a la obtención del título de **Ingeniero en Ciencias de la Computación,** de la autoría del estudiante **Wagner Cristopher Castillo Castro**, con **cédula de identidad** Nro. **0705011898**, una vez que el trabajo cumple con todos los requisitos exigidos por la Universidad Nacional de Loja, para el efecto, autorizo la presentación del mismo para su respectiva sustentación y defensa.


Firmado electrónicamente por:
PABLO FERNANDO ORDONEZ ORDONEZ

Ing. Pablo Fernando Ordoñez Ordoñez Mg.Sc.
DIRECTOR DEL TRABAJO DE INTEGRACIÓN CURRICULAR

Autoría

Yo, **Wagner Cristopher Castillo Castro**, declaro ser autor del presente Trabajo de Integración Curricular y eximo expresamente a la Universidad Nacional de Loja y a sus representantes jurídicos de posibles reclamos y acciones legales por el contenido del mismo. Adicionalmente acepto y autorizo a la Universidad Nacional de Loja la publicación de mí Trabajo de Integración Curricular en el Repositorio Digital Institucional – Biblioteca Virtual.

Firma: 

Cédula de identidad: 0705011898

Fecha: 29 de mayo de 2024

Correo electrónico: wagner.castillo@unl.edu.ec

Teléfono: (+593) 99 148 3745

Carta de autorización por parte del autor para consulta, reproducción parcial o total, y/o publicación electrónica del texto completo, del Trabajo de Integración Curricular.

Yo, **Wagner Cristopher Castillo Castro**, declaro ser autor del Trabajo de Integración Curricular denominado**: Modelo híbrido de predicción de fallas para los aerogeneradores de la Central Eólica Villonaco utilizando inteligencia artificial y datos del sistema SCADA**, como requisito para optar el título de **Ingeniero en Ciencias de la Computación**, autorizo al sistema Bibliotecario de la Universidad Nacional de Loja para que con fines académicos muestre la producción intelectual de la Universidad, a través de la visibilidad de su contenido en el Repositorio Institucional.

Los usuarios pueden consultar el contenido de este trabajo en el Repositorio Institucional, en las redes de información del país y del exterior con las cuales tenga convenio la Universidad.

La Universidad Nacional de Loja, no se responsabiliza por el plagio o copia del Trabajo de Integración Curricular que realice un tercero.

Para constancia de esta autorización, en la ciudad de Loja, a los veintinueve días del mes de mayo de dos mil cuatro.

Firma: Firmado electrónicamente por: WAGNER CRISTHOPER CASTILLO CASTRO

Autor: Wagner Cristhoper Castillo Castro
Cédula: 0705011898
Dirección: Loja
Correo electrónico: wagner.castillo@unl.edu.ec
Teléfono: (+593) 99 148 3745

DATOS COMPLEMENTARIOS:

Director del Trabajo de Integración Curricular: Ing. Pablo Fernando Ordoñez Ordoñez, Mg.Sc.

Dedicatoria

En memoria de mi padre, José Arturo, quien durante el periodo en que se realiza este trabajo dejó su presencia en este mundo, dejando un vacío en la vida de mi querida madre y la mía. Gracias por todo lo que hiciste por nosotros, sin tu presencia, nuestra vida habría sido diferente. Querido papá espero seguir siendo un buen hijo para ti, siempre estarás junto a mí con tus recuerdos, aunque ya no te encuentres físicamente a nuestro lado.

Este trabajo está dedicado a la memoria de mi padre y al futuro de mi amada hija, Sophia Castillo, un regalo en mi vida, luz en mi oscuridad, su llegada fue mi fortaleza en el momento en que flaquearon mis fuerzas y a mi añorado padre que deseó haber tenido la oportunidad de ser el primero en tomar en brazos a mi hija. Mis pilares en este mundo, que me han enseñado lo que vale una familia, un amor perfecto y la persistencia para lograrlo.

Wagner Cristhoper Castillo Castro

Agradecimiento

A Dios, por permitirme alcanzar la última etapa de mi carrera profesional bajo su bendición y protección.

A mi madre, Yenny Castro, pilar inquebrantable que nunca me ha fallado y que siempre me ha sostenido. Sin su apoyo incondicional, las circunstancias adversas habrían sido insuperables. Gracias por ser una parte esencial de mi vida, por apoyarme en cada paso del camino y por fortalecerme con tu amor y dedicación constante.

A mi amada compañera, Verónica Placencia, con quien he enfrentado desafíos y celebrados logros. Nuestra hija Sophia es preciosa, y sin duda alguna, será aún más maravillosa a medida que crezca, reflejando la bondad y fortaleza que hemos sembrado en ella.

Expreso mi sincero agradecimiento al Ing. Pablo Fernando Ordoñez y al Ing. Jorge Luis Maldonado por su incansable labor como docentes y tutores durante todo el proceso de este trabajo, su ayuda inestimable ha sido fundamental y valiosa en cada etapa de este trabajo.

También quiero expresar mi gratitud a la Universidad Nacional de Loja, Facultad de Energía, las Industrias y los Recursos Naturales no Renovables, y en especial a los docentes que conforman la carrera de Computación, debido a que cada uno de ellos ha sido una fuente invaluable de inspiración y conocimiento en mi desarrollo académico.

A todos aquellos que de una u otra manera contribuyeron en este camino, ¡gracias! Este logro no habría sido posible sin su apoyo y aliento constante.

Finalmente, agradezco a mi familia y amigos, quienes siempre han creído en mí y me han brindado su amor incondicional. Su constante estímulo ha sido una fuerza motivadora en cada paso que he dado.

¡Gracias a todos!

Wagner Cristhoper Castillo Castro

Índice de contenido

Índice de tablas:

Índice de figuras:

Índice de ecuaciones:

Índice de anexos:

Glosario de términos:

Términos generales

- **UNL.** Universidad Nacional de Loja.
- **TIC.** Trabajo de integración curricular.
- **IA (Inteligencia Artificial).** Simulación de procesos de inteligencia humana por sistemas informáticos.
- **ML (Machine Learning, Aprendizaje Automático).** Subcampo de la IA que permite a las máquinas aprender de datos y mejorar con la experiencia.
- **SCADA (Supervisory Control And Data Acquisition).** Sistema para controlar procesos a distancia y adquirir datos de equipos en el campo.
- **O&M (Operación y Mantenimiento).** Actividades realizadas para mantener y operar infraestructuras y sistemas eficientemente.
- **CEV.** Central Eólica Villonaco.
- **RSL.** Revisión Sistemática de Literatura.

Términos de modelado predictivo y evaluación

- **Matriz de confusión (Confusion Matrix).** Muestra el número de predicciones correctas e incorrectas para cada clase.
- **Verdaderos Positivos (VP).** Instancias correctamente identificadas como positivas.
- **Verdaderos Negativos (VN).** Instancias correctamente identificadas como negativas.
- **Falsos Positivos (FP).** Instancias incorrectamente identificadas como positivas.
- **Falsos Negativos (FN).** Instancias incorrectamente identificadas como negativas.
- **Validación cruzada (Cross-validation).** Evalúa modelos dividiendo los datos para pruebas repetidas.
- **Índice de desbalance (Imbalance Ratio).** Proporción entre clases en datos.
- **Precisión (Precision).** Verdaderos positivos sobre predicciones positivas.
- **Recuperación (Recall).** Verdaderos positivos sobre total positivos reales.
- **Puntuación F1 (F1 Score).** Balance entre precisión y recuperación.
- **Sensibilidad (Sensitivity).** Igual a recuperación; verdaderos positivos sobre positivos reales.
- **Especificidad (Specificity).** Verdaderos negativos sobre negativos reales.
- **Exactitud (Accuracy).** Predicciones correctas sobre total de casos.

- **Curva ROC (Receiver Operating Characteristic).** Gráfico que muestra el rendimiento de un modelo de clasificación a todos los umbrales de clasificación.
- **Área bajo la Curva (AUC).** Medida de la capacidad del modelo para distinguir entre clases.

Técnicas de mejora de datos

- **SMOTE (Synthetic Minority Over-sampling Technique).** Técnica de sobremuestreo que crea ejemplos sintéticos de la clase minoritaria.
- **SMOTE-ENN (SMOTE with Edited Nearest Neighbors).** Combina SMOTE con limpieza de ejemplos mediante ENN.
- **ADASYN (Adaptive Synthetic Sampling).** Variante de SMOTE que genera muestras sintéticas adaptativas para clases minoritarias.
- **SMOTE-TOMEK.** Combina SMOTE con limpieza de ejemplos usando enlaces Tomek.
- **Bordeline-SMOTE.** Versión de SMOTE enfocada en ejemplos de la clase minoritaria cercanos a la frontera de decisión.

Técnicas de combinación

- **Bagging (Bootstrap Aggregating).** Método de ensamble que mejora la estabilidad y precisión de las máquinas de aprendizaje.
- **Stacking (Stacked Generalization).** Método de ensamble que combina múltiples modelos de predicción.
- **Boosting.** Técnica de ensamble que construye modelos de manera secuencial intentando corregir los errores del modelo anterior.
- **Voting Ensemble.** Técnica de ensamble que realiza predicciones basadas en la mayoría de votos de modelos diversos.
- **Blending Ensemble.** Método similar al stacking pero utiliza un conjunto de validación para entrenar el modelo de segundo nivel.
- **Weighted Average Ensemble.** Método de ensamble que asigna diferentes pesos a las predicciones de los modelos antes de combinarlas.

Técnicas de aprendizaje de máquina

- **GMM.** Modelo de mezcla gaussiana.
- **MSSM.** Semisupervisado de Mahalanobis.
- **BAS-SVM.** Algoritmo de búsqueda de antenas de escarabajo basado en SVM.

- **MPE.** Entropía de permutación multiescala.
- **CNN.** Redes neuronales convolucionales.
- **BiGRU-AM.** Unidades recurrentes bidireccionales con mecanismo de atención.
- **XGBoost.** Extreme Gradient Boosting.
- **LSTM.** Red neuronal de aprendizaje profundo.
- **DT.** Árbol de decisión.
- **RF.** Bosque Aleatorio.
- **LR.** Clasificador logístico.
- **ML BlackBox.** Aprendizaje automático basado en Black box.
- **FNN.** Red neuronal de alimentación directa de dos capas.
- **NBM.** Modelo de comportamiento normal.
- **AE.** Autoencoder.
- **SVR.** Regresión por vectores de soporte.
- **NARX.** Red neuronal Autorregresiva no lineal con entradas exógenas.
- **RL.** Regresión logística.
- **FCNN.** Red neuronal totalmente conectada.
- **ANN.** Redes neuronales artificiales.
- **DNN.** Redes neuronales profundas.
- **GP.** Proceso gaussiano.
- **SVM.** Máquina de vectores soporte.
- **GANs.** Redes generativas adversariales.
- **Wk-NNs.** Vecinos más cercanos k ponderados.
- **BT.** Árbol potenciado.
- **RBT.** Árbol reforzado RUS.
- **RotF.** Bosque rotatorio.
- **FA.** Algoritmo Firefly.
- **CP.** Mapa del Caos.
- **ET.** ExtraTree o Árboles Extremadamente Aleatorizados.
- **ELM.** Máquina de Aprendizaje Extremo.
- **CART.** Árboles de clasificación y regresión.
- **CB.** Clasificadores Bayesianos.
- **kNN.** K-vecinos más próximos.

- **RUSBoost.** Refuerzo por submuestreo aleatorio.
- **PCA.** Análisis de componentes principales.
- **NNs.** Redes neuronales.
- **DEL.** Aprendizaje Extremadamente Profundo.
- **FFNN.** Red Neuronal de Propagación Hacia Adelante.
- **CFNN.** Red Neuronal de Propagación en Cascada.
- **NN-EL.** Aprendizaje extremo basado en redes neuronales.
- **H-K-means.** K-means jerárquico.
- **R-NN-EL.** Aprendizaje Extremo basado en Redes Neuronales Reducido.
- **FS.** Síntesis difusa.
- **FTL.** Aprendizaje por transferencia basado en características.
- **K-means.** Agrupación de K-means.
- **DAE.** Autocodificador profundo.
- **LoMST-CUSUM.** Árbol de expansión mínima local combinado y suma acumulativa de datos de series temporales multivariadas.
- **WHC-LOF.** Agrupamiento jerárquico de Ward combinado y detección de novedades con factor de valores atípicos local.
- **NBM-LI.** Modelo de comportamiento normal con entradas retardadas.
- **CCA.** Análisis de correlación canónica.
- **KCPD.** Detección de puntos de cambio basada en núcleos.
- **t-SNE.** Incrustación estocástica de vecinos t-distribuida.
- **LSTM-AE.** Autoencoders de Memoria a Corto y Largo Plazo.

Componentes de turbinas eólicas

- **Caja de cambios (Gearbox).** Ajusta la velocidad de rotación del rotor para el generador.
- **Sistema de orientación (Yaw).** Orienta la turbina hacia el viento.
- **Tren de transmisión (Drive Train).** Transmite la fuerza del rotor al generador.
- **Torre (Tower).** Sostiene la turbina elevándola con la finalidad de captar una mayor cantidad de viento.
- **Sistema de control y frenado.** Gestiona las tareas de operación y seguridad de la turbina correspondiente.
- **Generador (Generator).** Convierte la energía de tipo mecánica a energía eléctrica.

- **Rotor (Rotor).** Es un componente central que realiza la tarea de giro conjuntamente con el viento.
- **Palas (Blades).** Realiza la tarea de capturar el viento y hacer girar el rotor.
- **Sistema hidráulico (Hydraulic System).** Facilita el movimiento y las configuraciones pertinentes de la operación como la orientación o el pitch.
- **Convertidor (Converter).** Convierte la energía eléctrica generada para su compatibilidad con la red eléctrica.
- **Sistema de pitch (Pitch System).** Ajusta el ángulo de las palas para controlar la velocidad del rotor y maximizar la eficiencia.
- **Rodamientos (Bearings).** Permiten el movimiento suave de partes giratorias como el rotor.
- **Transformador (Transformer).** Aumenta la tensión de la electricidad generada para su transmisión a larga distancia.
- **Grupo hidráulico (Hydraulic Group).** Conjunto de componentes que operan el sistema hidráulico para ajustes de pitch, frenado, y otros movimientos necesarios.
- **Planetary Gearbox (Caja de cambios planetaria).** Un tipo específico de caja de cambios que distribuye la carga a través de varios ejes para una mayor eficiencia y capacidad de carga.

Otros términos

- **DOI (Digital Object Identifier).** Identificador único para documentos digitales.
- **IDE (Entorno de Desarrollo Integrado).** Aplicación software que proporciona servicios comprehensivos para el desarrollo de software.

1. Título

Modelo híbrido de predicción de fallas para los aerogeneradores de la Central Eólica Villonaco utilizando inteligencia artificial y datos del sistema SCADA.

Hybrid Fault Prediction Model for Wind Turbines at the Villonaco Wind Farm Employing Artificial Intelligence and SCADA System Data.

2. Resumen

La energía eólica ha surgido como una alternativa prometedora frente al uso de combustibles fósiles, convirtiéndose en una de las fuentes de energía renovables de mayor crecimienyto a nivel mundial. A pesar de sus ventajas, la eficiencia de la energía eólica se ve limitada por los costos asociados a la operación y mantenimiento (O&M), los cuales representan una parte significativa del gasto total en un parque eólico. El objetivo de este trabajo de integración curricular (TIC) es implementar un modelo híbrido que combina técnicas de inteligencia artificia (IA) y datos del sistema SCADA (Supervisory Control and Data Acquisition) para realizar la clasificación de fallas en los aerogeneradores de la Central Eólica Villonaco (CEV), se debe determinar si los datos registrados corresponden a datos de falla o no. Por lo tanto, el resultado principal es un modelo híbrido basado en Redes Neuronales Artificiales (ANN), Random Forest (RF) y la técnica de combinación Blending Ensemble denominado Blending-ANN-RF, cuya ejecución se llevó a cabo bajo la adaptación de la metodología CRISP-DM, la cual se aplicó para orientar los experimentos con varios tipos algoritmos en conjunto con técnicas de sobremuestreo y en las diferentes técnicas de combinación, donde la técnica Blending emerge con especial relevancia. Esta estrategia logró un rendimiento sobresaliente en evaluaciones específicas, tales como las derivadas de la matriz de confusión, en la validación cruzada y el test de Wilcoxon, además de evidenciar un tiempo de entrenamiento reducido. Estos resultados respaldan la eficacia y eficiencia del modelo híbrido implementado, ratificando la idoneidad de la técnica Blending para mejorar las capacidades individuales de los algoritmos ANN y RF. Así mismo, se han identificado áreas prometedoras para futuras investigaciones, que incluyen la exploración de nuevos modelos que aprovechan los datos SCADA y componentes de un aerogenerador más estudiados. Estas líneas de investigación futuras prometen contribuir aún más al campo de la predicción de fallas en aerogeneradores.

Palabras Clave: *algoritmos de aprendizaje automático, predicción de fallas, técnicas de sobremuestreo, técnicas de combinación de algoritmos.*

Abstract

Wind energy has emerged as a promising alternative to the use of fossil fuels, becoming one of the fastest-growing sources of renewable energy worldwide. Despite its advantages, the efficiency of wind energy is limited by the costs associated with operation and maintenance (O&M), which represent a significant portion of the total expenditure in a wind farm. The objective of this curricular integration project (TIC) is to implement a hybrid model that combines artificial intelligence (AI) techniques and data from the SCADA (Supervisory Control and Data Acquisition) system to classify faults in the Central Eólica Villonaco (CEV). It is necessary to determine whether the recorded data corresponds to fault data or not. Therefore, the main result is a hybrid model based on Artificial Neural Networks (ANN), Random Forest (RF), and the Blending Ensemble technique called Blending-ANN-RF, whose execution was carried out under the adaptation of the CRISP-DM methodology, which was applied to guide experiments with various algorithms in conjunction with oversampling techniques and different combination techniques, where the Blending technique emerges with special relevance. This strategy achieved outstanding performance in specific evaluations, such as those derived from the confusion matrix, cross-validation, and the Wilcoxon test, in addition to demonstrating reduced training time. These results support the effectiveness and efficiency of the implemented hybrid model, confirming the suitability of the Blending technique to enhance the individual capabilities of the ANN and RF algorithms. Furthermore, promising areas for future research have been identified, including the exploration of new models that leverage SCADA data and more studied components of a wind turbine. These lines of future research promise to further contribute to the field of wind turbine failure prediction.

Keywords: *machine learning algorithms, fault prediction, oversampling techniques, algorithm combination techniques.*

3. Introducción

El presente trabajo de integración curricular (TIC) se enfocó en determinar un modelo híbrido de inteligencia artificial (IA) para la predicción de fallas en aerogeneradores con el uso de datos SCADA (Supervisión, Control y Adquisición de Datos) de la Central Eólica Villonaco (CEV). Con el propósito de encontrar un modelo optimo que pueda reconocer fallas, se plantea responder a la pregunta de investigación: ¿Qué probabilidad de acierto tendrá un modelo híbrido de predicción de fallas para los aerogeneradores de la CEV utilizando inteligencia artificial y datos del sistema SCADA? Este enfoque se fundamentó en la combinación de diversas técnicas y modelos de IA adaptados a las condiciones específicas de la CEV, utilizando datos del sistema SCADA provenientes de los registros generados por los aerogeneradores. Aunque el propósito fundamental de este modelo híbrido no incluye la implementación directa de medidas preventivas, se reconoce que la detección temprana de fallas proporciona información valiosa, esto permite a los técnicos encargados de los aerogeneradores tomar medidas preventivas de manera proactiva para reducir tiempos de inactividad y minimizar los costos asociados a operaciones y mantenimiento (O&M).

En Ecuador la CEV se destaca como uno de los proyectos más ambiciosos en términos de generación de energía renovable, ubicada en la provincia de Loja, en la sierra sur del país, cuenta con 11 aerogeneradores del tipo GW70/1500, con una capacidad de 1.5 MW cada uno [1]. Esta impresionante instalación sienta las bases para la implementación de avances tecnológicos que mejoren su eficiencia operativa. En este contexto, el objetivo es identificar un modelo híbrido para identificar fallas de manera efectiva, su implementación podría, indirectamente, influir en la toma de decisiones anticipadas e informadas, lo que potencialmente podría conducir a mejoras en la eficiencia operativa y en la reducción de los costos asociados a los tiempos de inactividad y a las reparaciones no planificadas, como se puede observar en [2]. Además, el presente TIC no solo promueve el uso de algoritmos de machine learning (ML), técnicas de sobremuestreo y técnicas de combinación aplicadas a la predicción de fallas, sino que también brinda una investigación sobre una gran cantidad de técnicas para la predicción de fallas, componentes de un aerogenerador más estudiados e investigaciones de otros autores relevantes.

Los modelos de predicción de fallas en aerogeneradores es un campo ampliamente investigado, la particularidad de la CEV, ubicada a una altitud de aproximadamente dos mil setecientos metros sobre el nivel del mar (msnm), presenta desafíos distintivos [1], donde esta elevada altitud genera condiciones únicas que demandan un enfoque especializado y

adaptado a las peculiaridades de la CEV, lo que a su vez puede beneficiar a todo el sector de la energía eólica.

La misión de este trabajo se enmarcó en un contexto de continuo desarrollo en el campo de la predicción de fallas en aerogeneradores, tal como se evidencia en el Anexo 1. Revisión Sistemática de Literatura (RSL), la cual planteó una sólida base de trabajos previos que han contribuido al avance de técnicas y modelos en la detección temprana de eventuales fallas, no obstante, las condiciones particulares de la CEV sugieren un desafío particular, por consiguiente, esta investigación se alinea con trabajos previos al extender y acoger las metodologías existentes para abordar todas estas peculiaridades y contribuir a la confiabilidad y eficiencia de la CEV.

Se ha establecido como objetivo principal la identificación de un modelo híbrido basado en IA, aplicando datos proporcionados por el sistema SCADA de los aerogeneradores de la CEV, para ello, se estableció dos objetivos específicos, el primero se basó en efectuar una RSL acerca de modelos predictivos basados en IA y datos SCADA para la detección de fallas en aerogeneradores, la cual permitió adquirir un profundo conocimiento de técnicas y modelos, componentes de un aerogenerador más estudiados y los autores más relevantes, por otro lado, el segundo objetivo específico se orientó en la implementación de un modelo híbrido, por lo cual se experimenta con varios tipos de algoritmos de ML en conjunto con técnicas de sobremuestreo y finalmente con las técnicas de combinación entre los algoritmos más relevantes. Al cumplir estos objetivos se espera contribuir en la generación de conocimiento científico en el área de la predicción de fallas en aerogeneradores, con ello, el alcance de este TIC se centró en la implementación de un modelo híbrido en el contexto particular de la CEV.

Las pruebas se centraron en diversas métricas considerando aquellas derivadas de la matriz de confusión: Precisión (P), Recuperación (R), Puntuación F1(PF1), Sensibilidad(S), Especificidad(ES), Exactitud(EX), Área bajo la curva(AUC), test de Wilcoxon y el tiempo de entrenamiento, no obstante, es importante indicar que el estudio no abordó aspectos relacionados con la operación de la CEV que se encuentren más allá de la predicción de fallas, como por ejemplo, la gestión de recursos humanos o aspectos legales, además, es importante señalar que la investigación presentó ciertas limitaciones como basarse en datos accesibles y disponibles hasta la fecha de corte de la investigación que toma el período de enero a diciembre de 2020, lo cual implica que los cambios o nuevos ajustes en las condiciones de operación, políticas o tecnologías posteriores a la fecha indicada no se reflejan en este estudio.

El presente TIC se ha estructurado en varias secciones que proveen una interpretación clara y organizada del desarrollo del proyecto, el Marco Teórico plantea una base sólida para comprender las teorías fundamentales relacionadas con los aspectos considerados importantes y esenciales para este estudio, por otra parte, la sección de Metodología indica las técnicas y procedimientos aplicados en la investigación, los Resultados presentan todos los hallazgos adquiridos que se dividen de acuerdo a los dos objetivos específicos en donde, el primero especifica las técnicas y modelos encontrados haciendo énfasis en los componentes de aerogenerador más estudiados en la predicción de fallas, en el segundo se evalúan los algoritmos de ML o técnicas de IA, y posteriormente los dos algoritmos con mejores resultados son considerados parte de las técnicas de combinación para generar un modelo híbrido con la finalidad de adquirir aún mejores resultados, todo ello se evalúa con diversas evaluaciones, como con métricas derivadas de la matriz de confusión y el test de Wilcoxon, a continuación, se encuentra la Discusión, donde se interpreta cada una de las fases del TIC a partir de la perspectiva del autor y finalmente, se detallan las Conclusiones y Recomendaciones derivadas del TIC, apoyadas por la Bibliografía y los Anexos que atribuyen un soporte documental y de evidencia científica del trabajo realizado.

4. Marco teórico

4.1. Antecedentes

4.1.1. Energía renovable y su importancia

Hoy en día, el mundo se encuentra envuelto en una lucha constante contra la contaminación mundial debido al uso de recursos naturales no renovables como gases de efecto invernadero y combustibles fósiles, por lo cual el mundo se enfrenta a problemas como el incremento de la contaminación global y la escasez, en vista de que estos recursos no pueden ser reemplazados a una igual velocidad de la que son consumidos para producir energía, la implementación de recursos naturales renovables para la generación de energía nace como una excelente alternativa para solucionar estos inconvenientes y convertirse en una fuente de energía más limpia y amigable con el medio ambiente [3].

La aplicación de energía renovable ofrece múltiples ventajas, considerando una elevada estabilidad en los precios de energía eléctrica, debido a que tienden a mantenerse estables o incluso reducirse con el paso del tiempo beneficiando a los consumidores y la economía a nivel global, esto a su vez contribuye beneficios ambientales significativos ya que, la energía producida es asimilada como energía limpia en contraste con el uso de recursos naturales no renovables [4], [5].

Existen múltiples tipos de energía considerada como renovable que se estiman rentables en términos de costo y eficiencia en contraste con los combustibles fósiles, sobresaliendo tres de ellos en particular, en primer lugar, la energía solar destaca por ser muy rentable en vista de la considerable disminución de los costos de los paneles solares y los avances en término de tecnología de almacenamiento de energía [5], en segundo lugar, la energía eólica ha experimentado una reducción importante de precios en la actualidad, lo cual genera que sea cada vez más competitiva frente a los combustibles fósiles y se considera relativamente económica y accesible para los países en vías de desarrollo [6], para finalizar, la energía hidroeléctrica es otra fuente muy rentable que se puede aprovechar en lugares que disponen de ríos y cascadas donde dichas fuentes de energía renovable atribuyen ventajas económicas y medioambientales representando una alternativa esperanzadora para satisfacer eventuales necesidades energéticas de forma sostenible y muy rentable [7].

4.1.2. Desarrollo de la energía eólica

El aumento de generación de energía eólica en las últimas dos décadas ha sido considerable en vista del atractivo en términos de producción y ha contribuido significativamente la economía a nivel mundial al reducir costos de producción, reducir la huella de carbono y promover la transición en rumbo a una matriz energética mayormente sostenible y amigable con el medio ambiente, además, su crecimiento se ve potenciado por los avances tecnológicos, elevadas inversiones económicas y considerables beneficios de orden ambiental [8].

4.2. Energía eólica

4.2.1. Energía eólica a nivel mundial

Global Wind Energy Council (GWEC) corresponde a una organización sin fines de lucro dedicada a facilitar y fomentar el desarrollo de energía eólica alrededor del mundo [9], con sede en Bruselas capital de Bélgica a partir del año 2005, donde su principal actividad es interactuar conjuntamente con gobiernos, empresas y otras entidades con la finalidad de motivar en la implementación de energía eólica como una fuente de energía limpia, sostenible y amigable con la naturaleza.

Una de las tareas fundamentales de GWEC es la estructuración de expedientes relacionados a la industria eólica alrededor del mundo siendo conocidos por otorgar un análisis muy detallado de datos actualizados y perspectivas del crecimiento, inversión, tendencias y múltiples factores de vital importancia dentro del sector eólico basándose en la información actualizada y recopilada en su informe del año 2022, de este modo, se reconocen los siguientes elementos clave de acuerdo a [9] que se describen a continuación:

- En el año 2021, se rastreó un aumento del 1,8% en la capacidad global de energía eólica encontrándose apenas un punto porcentual por debajo del récord registrado en el año 2020, de esta manera, durante el segundo año de la pandemia de COVID-19 se logró agregar aproximadamente un 94 GW de capacidad eólica en las actuales instalaciones sumando 93,6 GW ese año, por lo tanto, la capacidad acumulada mundial de energía eólica alcanzó los 837 GW representando un incremento interanual del 12% que destacó la capacidad de adaptación y eventual crecimiento del sector a pesar de enfrentar desafíos a nivel general.

- En el año 2022, se registró un acontecimiento significativo en la energía eólica con la instalación de 94 GW de capacidad en todo el mundo, de los cuales 21 GW fueron destinados a proyectos ubicados en alta mar donde dicho incremento evidenció un interés determinante y de expansión del uso de la energía eólica tanto en tierra como a nivel del mar.
- Los combustibles fósiles continúan dominando la industria energética mundial pese a ciertos obstáculos como la pandemia de COVID-19 y aunque se han realizado esfuerzos para adaptarse a fuentes de energía más ecológicas, dicha dependencia sigue siendo esencial en muchas naciones.
- La industria eólica confronta a una serie de adversidades que tendrían un impacto sustancial en su desarrollo y en su capacidad para competir a nivel económico y figuran las interrupciones de la cadena de suministro y el creciente coste de los productos considerados básicos y las materias primas.

4.2.2. Energía eólica en Ecuador

Ecuador ha emprendido importantes reformas para mejorar su sistema eléctrico y superar los inconvenientes detectados en las décadas de 1980 y 1990 dando como resultado el reconocimiento de la importancia de diversificar su energética comprometiéndose a disminuir drásticamente su dependencia a los combustibles fósiles y a impulsar su potencial de energía hidroeléctrica y otras fuentes de energía renovables no convencionales [10].

En Ecuador, durante el año 2006 se identificó que su sistema energético se estructuraba principalmente de dos fuentes que son las energías fósiles y las energías renovables, así, de acuerdo a la información disponible aproximadamente el 86% de la matriz energética se basaba en fuentes de energía fósil, mientras que el 10% correspondía a fuentes de energía renovable [11], por otra parte, en los últimos años, según se describe en [12], Ecuador ha logrado generar una capacidad de 21.5 MW de energía eólica, lo cual representa un importante avance en el territorio nacional proyectando que en los próximos años se logre una capacidad esperada de 200 MW.

En los últimos años, el país ha optado por la generación de energía eléctrica través de parques eólicos destacando el Parque Eólico Villonaco de la ciudad de Loja, el primero en su territorio continental, el cual comenzó a operar en 2013 con una capacidad ponderada de 16.5 MW [12], además, en 2021 se construyó un nuevo parque eólico en Ambocas, Loja, con una capacidad nominal de 10 MW, con ello, se

estima que este parque contribuya a aumentar la cantidad, así como la calidad de la energía generada en la ciudad [13].

Ecuador del mismo modo ha aumentado sus esfuerzos en el ámbito de la generación de energía eólica implementando proyectos en las Islas Galápagos donde se está aprovechando en la isla San Cristóbal la generación de energía que oscila los 2.4 MW y en la isla Baltra una capacidad de 2.25 MW [14], dicho éxito ha impulsado el desarrollo de nuevas iniciativas eólicas como por ejemplo el proyecto eólico de Mina de Huascachaca, Guarapamba y Ducal [15].

4.2.3. Energía eólica en Loja

La provincia de Loja, localizada en Ecuador, dispone de una población aproximada de 511.184 habitantes y se extiende sobre una superficie de 11.026 km2, lo que oscila el 4% del territorio nacional, dicha región resalta por su abundancia en recursos naturales renovables que la convierte en una localidad altamente favorecida por la naturaleza en este ámbito [16].

Según estudios técnicos, se ha identificado que el potencial eólico en la región se concentra en áreas específicas, donde la velocidad anual promedio del viento supera los 10 m/s, dicho hallazgo demuestra que existen zonas idóneas para el desarrollo de proyectos eólicos con condiciones de viento favorables que permiten una generación eficiente de energía limpia, por consiguiente, en la ciudad de Loja, se destaca la presencia del parque Eólico Villonaco, el cual dispone de 11 aerogeneradores y mantiene una capacidad de generación aproximada de 71,94 GWh al año, cabe recalcar que este parque eólico se distingue por su ubicación a una altitud de 2700 metros sobre el nivel del mar, lo que lo convierte en uno de los pocos parques en el mundo situados a una altura tan elevada [17], [18].

4.3. Aerogeneradores

4.3.1. Funcionamiento de los aerogeneradores

Los aerogeneradores transforman la energía cinética proveniente del viento en electricidad mediante un proceso eficiente que consiste en que el viento impulsa las palas del aerogenerador, cuyo movimiento transfiere energía al rotor conectado a un respectivo generador, que convierte esta energía mecánica en eléctrica, así, dicha electricidad se canaliza para alimentar redes eléctricas, sistemas de almacenamiento o consumo local con diseños que varían en tamaño y que pueden ser de eje horizontal o vertical donde los aerogeneradores pueden ser instalados tanto en tierra firme como

en mar abierto, optimizando el uso de varias condiciones del viento para aumentar la generación de energía renovable [19], [20], [21].

4.3.2. Componentes principales de los aerogeneradores

Según [19], un aerogenerador se compone de muchos componentes básicos que se combinan para convertir la energía cinética del viento en energía eléctrica, estos componentes se explican a continuación

- **Blades (Palas):** Son las principales piezas aerodinámicas de un aerogenerador; se asemejan en su forma a las alas de un avión y están optimizadas para aprovechar la energía del viento y transformarla en energía mecánica, además, giran en respuesta a la corriente de viento, que se transfiere al rotor del aerogenerador.
- **Rotor (Rotor):** La función del rotor, situado en el extremo del eje del aerogenerador, es transmitir el movimiento de rotación de las palas al sistema de producción de energía, se utiliza una caja de engranajes u otro sistema de transmisión para transferir este movimiento, aumentando la velocidad de rotación y dirigiéndolo hacia el generador.
- **Generator (Generador):** Un conjunto de bobinas e imanes componen el generador de la turbina eólica, que está unido al rotor y funciona según el principio de inducción electromagnética; cuando el rotor gira debido a la fuerza del viento, los imanes crean un campo magnético variable que fluye por las bobinas del generador, este cambio en el campo magnético hace que una corriente eléctrica fluya a través de las bobinas de acuerdo con el principio de inducción electromagnética, convirtiendo la energía mecánica del movimiento rotatorio en energía eléctrica útil.
- **Control and braking system (Sistema de control y frenado):** Los dos objetivos principales del sistema de control y frenado son regular la velocidad de rotación del rotor y detener el aerogenerador en caso de emergencia o durante el mantenimiento; además, el sistema utiliza un sistema de control electrónico para supervisar y modificar la velocidad del rotor y garantizar que las palas estén colocadas correctamente, de esta forma el sistema de control vigila otros parámetros del aerogenerador, como la temperatura, la producción de energía, la orientación de las palas y la transmisión de datos,

entre otros, y el sistema de frenado se activa cuando el aerogenerador debe detenerse, ya sea para un mantenimiento planificado o en caso de emergencia.

- **Tower (Torre):** Es la estructura vertical que soporta el rotor y eleva el aerogenerador a una altura ideal que es crucial para que las palas alcancen alturas donde los vientos favorables consiguen mejorar la captación de energía eólica y aumentar la producción y eficiencia de energía eléctrica.

4.3.3. Fallas en aerogeneradores

Según estudios como [22], analizan las fallas más comunes en los sistemas aerogeneradores que son susceptibles de sufrir averías y repercuten en su funcionamiento tales como el diseño exacto del aerogenerador, el entorno de funcionamiento y el mantenimiento pueden afectar su fiabilidad y rendimiento a largo plazo.

- **Blades (Palas):** La eficacia de la captación de energía se ve directamente afectada por ciertos factores como la acumulación de hielo, deformaciones estructurales, daños superficiales, acción del viento, desgaste por objetos externos y la fatiga de los materiales que someten a los aerogeneradores a condiciones meteorológicas severas.
- **Gearbox (Caja de Engranajes):** El uso continuado logra desgastar los engranajes de la caja de cambios que reducen la eficacia de la transmisión, ya que los problemas de sellado o lubricación de igual manera pueden crear fugas de aceite que ponen en peligro la lubricación apropiada de los componentes.
- **Braking system (Sistema de Frenado):** Las fallas en el sistema de frenado de emergencia hacen imposible la labor de detener el aerogenerador en situaciones extremadamente peligrosas.
- **Tower (Torre):** Las condiciones atmosféricas que provocan la corrosión y la erosión afectan a la torre de la turbina eólica porque la corrosión daña a la integridad estructural y la erosión puede debilitar el material y aumentar la probabilidad de fallo, de igual manera, se debe a cambios en la composición del suelo, erosión y condiciones geotécnicas que generan el desplazamiento o hundimiento de los cimientos de la torre que pone en peligro la estabilidad estructural de la turbina eólica.

- **Generator and Electrical System (Generador y Sistema eléctrico):** Son circunstancias de funcionamiento extremas, defectos de fabricación o problemas con el sistema de control que hacen que el generador y los componentes eléctricos desencadenen un cortocircuito y se sobrecalienten provocando daños permanentes y la pérdida de generación de energía.
- **Guidance and Control System (Sistema de Orientación y Control):** Los sistemas de monitorización y control como SCADA, son esenciales para la detección temprana de anomalías y fallos que pueden resultar en la incapacidad de identificar y mitigar inconvenientes que comprometen la operatividad del aerogenerador, por otra parte, el sistema de guiado es fundamental en la captura eficiente del viento que puede experimentar errores provocados por defectos en los sensores, actuadores o algoritmos de control y estos fallos impactan de forma directa a la producción de energía y a la vida útil del aerogenerador.

4.4. Sistemas SCADA

Según lo expresado por los autores [22], [23], entre las distintas variables críticas que el sistema SCADA recoge y registra en tiempo real se encuentran la velocidad del viento, temperatura, producción de energía, tensión, corriente y vibraciones que proporcionan información muy detallada sobre el rendimiento y el estado operativo de los aerogeneradores obtenida a través de sensores y dispositivos colocados estratégicamente en diversos componentes del aerogenerador donde dicha captura y registro continuo permite monitorizar y evaluar en tiempo real el funcionamiento del aerogenerador.

4.4.1. Principales funciones de un sistema SCADA

En [24], se destacan las principales funciones de un sistema SCADA que incluyen:

- **Supervisión:** Permite al operador seguir en tiempo real las modificaciones en el funcionamiento diario de la planta y la evolución de las variables de control que ayuda en la toma de decisiones optimizando las tareas de mantenimiento y el análisis estadístico de fallos incrementando la fiabilidad y eficiencia del sistema.
- **Control:** Es la capacidad de activar o desactivar equipos a distancia y de forma manual gracias a un sistema SCADA que de igual manera ofrece al

operador la posibilidad de efectuar acciones de control y modificar el curso del proceso en caso de que se produzcan circunstancias anómalas.

- **Adquisición de datos:** Se encarga de recopilar, procesar, almacenar y presentar sistemáticamente los datos de los equipos de campo, incluida la recopilación de datos en tiempo real de factores importantes como vibraciones, tensión, corriente, temperatura, velocidad del viento, producción de energía, etc.
- **Generación de reportes:** El sistema SCADA permite crear informes exhaustivos con representaciones gráficas, análisis estadísticos, predicciones, gestión de la producción y datos necesarios para la gestión financiera y administrativa a partir de los datos recopilados.
- **Representación de señales de alarmas:** En caso de averías o situaciones inusuales, ofrece al operador una representación clara y eficaz de las señales de alarma, que pueden ser auditivas mediante alarmas sonoras o visuales mediante indicadores en el panel de control.

4.4.2. Correlación de variables SCADA

Un coeficiente de correlación, que oscila entre -1 y +1 e indica el tipo de vínculo entre las variables, es el resultado de un proceso estadístico denominado análisis de correlación, que establece si dos variables están correlacionadas.

- Un signo negativo indica una relación negativa.
- Un signo positivo indica una relación positiva o directa.
- Un valor nulo señala la ausencia de tendencia entre las variables.

La magnitud indica la fuerza de la relación. Cuanto más cercano esté el valor a los extremos del intervalo (1 o -1), más fuerte será la tendencia de las variables, o menor será la dispersión entre los puntos.

Existen varios métodos para realizar un análisis de correlación y uno de ellos es el coeficiente de correlación de Pearson, el cual es apropiado para datos SCADA debido a la relación lineal entre dos variables continuas, lo cual concuerda con la naturaleza de los datos SCADA, los cuales suelen mantener mediciones continuas a través del tiempo, también, ofrece una medida cuantitativa de la fuerza y la relación entre las variables.

4.4.3. Coeficiente de correlación de Pearson

El coeficiente de correlación de Pearson es un método que evalúa la relación lineal entre dos variables continuas y es muy útil cuando los datos siguen una distribución normal y existe una relación entre las variables. Su fórmula se presenta en la Ecuación 1.

Ecuación 1. Coeficiente de correlación de Pearson.

$$XY = \frac{\sum_{i=1}^{n}(Xi - \bar{X})(Yi - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(Xi - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Yi - \bar{Y})^2}} \quad (1)$$

Donde:

- Xi y Yi son los valores individuales de las dos variables
- $\bar{X}$ y $\bar{Y}$ son las medias de las dos variables, respectivamente.

Esta fórmula se compone de dos partes principales:

- **Covarianza:** Se multiplican las desviaciones emparejadas de *X* y *Y*, y se suman los productos para medir cómo cambian juntos, lo que se denomina covarianza.
- **Denominador:** Se calculan las desviaciones estándar de *X* y *Y* para ver cuánto se dispersan alrededor de sus medias. Al dividir la covarianza por el producto de estas desviaciones, se obtiene el coeficiente de correlación de Pearson, que estandariza la medida de la correlación entre *X* y *Y*.

4.4.4. Mapa de calor y matriz de correlación

Un mapa de calor representa visualmente los resultados de la correlación entre variables utilizando métodos como el coeficiente de correlación de Pearson que esencialmente, muestra los resultados de una matriz de correlación, pero utiliza colores para resaltar las diferentes intensidades de correlación entre las variables, así este gráfico permite identificar diferentes patrones y la relación dentro de los datos, por otra parte, los colores intensos suelen indicar correlaciones más fuertes, mientras que los colores más suaves indican una correlación más débil.

Un ejemplo de cómo funciona la matriz de correlación y su mapa de calor derivado se puede apreciar así: la matriz de correlación se representa como se muestra en la Figura 1, mientras que su correspondiente mapa de calor se visualiza como se muestra en la Figura 2.

```
[[80  1 58 75 77 26 76 91 10 13]
 [28 87  7  1 75 94 27 20 40 90]
 [29 22 53 52 57 94 27  2 67 80]
 [ 7 58 34 27 60 23 48 81 86 55]
 [49 31 73 28 57 96 27 19 46 15]
 [99 30 39 36 46 77 37 52  6 80]
 [73 86 96 94 89 36 42 16 68 57]
 [63 38 41 81 70 11 84 22  5 57]
 [62 22 11 42 61 87 81 30 63 89]
 [20 90 16 38 97  4 45 59 78 76]]
```

Figura 1. Matriz de correlación.

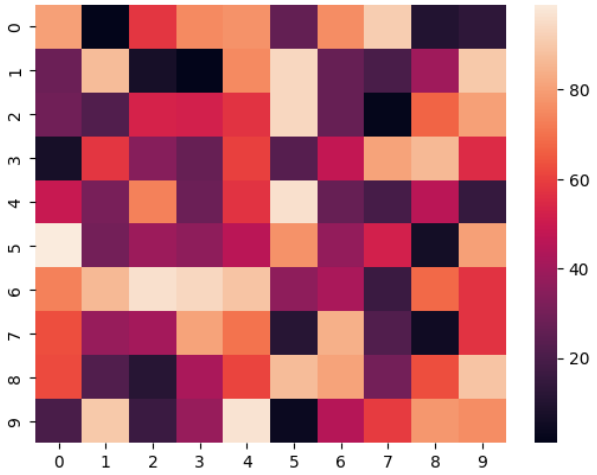

Figura 2. Mapa de calor derivado de la matriz de correlación

El mapa de calor muestra, según la escala de colores en la barra ubicada a la derecha, que los números que superan 60 tienden a ser más claros, mientras que los números inferiores tienden a ser más oscuros, lo cual permite visualizar la intensidad de la correlación entre las variables, proporcionando una representación intuitiva de su relación y un ejemplo de su aplicación se puede observan en [25].

4.5. Inteligencia artificial

4.5.1. Introducción a la inteligencia artificial

La inteligencia artificial representa un área de investigación dedicada al desarrollo de tecnologías y sistemas avanzados con la capacidad de ejecutar actividades que a nivel tradicional demandarían la inteligencia propia de los seres humanos, por lo general, esta área de estudio se apoya en la elaboración y aplicación de algoritmos y modelos matemáticos avanzados, los cuales dotan a las máquinas de la capacidad de aprender de manera autónoma, razonar, tomar decisiones informadas y abordar problemas complejos y gracias a su habilidad para analizar enormes volúmenes de información, reconocer patrones y ajustarse a nuevos contextos es lo

que hace que estos sistemas puedan ejecutar tareas de alta complejidad con notable precisión y eficiencia.

La IA ha revolucionado múltiples industrias y se utiliza en diversos campos de estudio, dispone de capacidades muy amplias que actualmente se aplican en sectores como odontología, medicina, agricultura, manufactura, seguridad, comercio electrónico, entre otros, además, su implementación ha propiciado avances significativos en estas áreas, permitiendo mejoras importantes en términos de eficiencia, precisión, automatización y en la capacidad para realizar tomas de decisiones basadas en el análisis inteligente de datos, a continuación, se presentan algunos ejemplos de cómo la IA está siendo aplicada de manera efectiva en los campos antes mencionados.

- **Odontología:** Se ha integrado de manera innovadora, abarcando especialidades como la radiología, periodoncia, rehabilitación oral, ortodoncia y odontología forense, dichas aplicaciones revolucionan la manera de diagnosticar y tratar las afecciones bucodentales, ofreciendo una precisión sin precedentes en la interpretación de imágenes radiológicas, evaluación de riesgos periodontales, planificación de tratamientos de rehabilitación y ortodoncia, así como en la identificación de individuos en contextos forenses [26].
- **Oncología:** En este campo la IA permite respaldar decisiones clínicas que incluyen, diagnostico automatizado de cáncer, capacidades de predicción mejoradas, actualización de datos en tiempo real, atención personalizada, aumento de eficiencia y mejora de resultados [27].
- **Agricultura:** Se utiliza para revolucionar la gestión de las explotaciones agrícolas proporcionando métodos para optimizar la producción, el uso de los recursos y la sostenibilidad medioambiental mediante un análisis en profundidad de las condiciones del suelo, estas técnicas ayudan a los agricultores a tomar decisiones mejor fundamentadas y aumentan la resistencia de los cultivos a las fluctuaciones climáticas y del mercado, así como su productividad [28].
- **Manufactura:** Al identificar con precisión defectos superficiales en productos terminados que son invisibles al ojo humano, la IA redefine el control de calidad y mejora la calidad de los productos, al tiempo que agiliza los procedimientos de producción y reduce los residuos y los costes de rechazo [29].

- **Seguridad:** En un entorno digital cada vez más complejo y vulnerable, proporciona soluciones avanzadas de detección y tratamiento de amenazas que permiten a las organizaciones adoptar un enfoque más proactivo del riesgo, mejorando su capacidad para reconocer, detener y responder a los ciberataques y salvaguardando la integridad de los sistemas de información y la privacidad de los datos [30].

4.5.2. Técnicas de inteligencia artificial aplicadas a la predicción de fallas

Como se ve en el **Anexo 1. Revisión Sistemática de Literatura**, el uso de la IA en la identificación y previsión de averías en los aerogeneradores se ha convertido en un campo muy estudiado, lo que demuestra la sorprendente diversidad de modelos utilizados en este campo; en los artículos seleccionados se encontraron más de 100 técnicas diferentes, lo que indica no solo el creciente interés por la aplicación concreta de la IA, sino también la riqueza y diversidad de enfoques desarrollados para abordar el problema de la detección temprana de averías en los aerogeneradores.

Las diversas técnicas de la IA muestran la creatividad y eficacia con que se está aplicando la tecnología en el mantenimiento y optimización de aerogeneradores, la amplitud y profundidad de los modelos presentados en este campo ilustran la complejidad de abordar la tarea de predecir y diagnosticar fallas en estos sistemas.

4.5.3. Aprendizaje automático (Machine Learning)

El principal objetivo del aprendizaje automático (machine learning, ML), un subcampo de la inteligencia artificial, es dotar a las máquinas de la capacidad de aprender de los datos por sí mismas sin tener que ser programadas explícitamente, el ML se centra en crear modelos y algoritmos que permitan a las máquinas hacer predicciones y tomar decisiones basadas en patrones y conocimientos encontrados en los datos [31].

En el contexto de la predicción de fallas en aerogeneradores, ML es esencial para analizar enormes conjuntos de datos históricos recopilados de los aerogeneradores, con el fin de encontrar patrones y relaciones importantes que ayuden a identificar posibles fallas en una fase temprana y tomar medidas preventivas.

A continuación, se describen algunos modelos de ML y en la Tabla 1 se destacan las ventajas y desventajas de cada uno de ellos.

- **Redes neuronales artificiales (ANN):** Están formados por capas de neuronas enlazadas que procesan y transfieren información; también sirven para analizar patrones complicados en los datos y prever la aparición de errores, su diseño está influido por el funcionamiento del cerebro humano.
- **Máquinas de vector soporte (SVM):** Mediante el proceso de encontrar el mejor hiperplano para dividir los datos en clases distintas para la predicción de fallas, se puede utilizar una SVM para categorizar los datos e identificar si un aerogenerador está funcionando según lo previsto o de forma óptima, o si hay indicios de un posible fallo.
- **Bosques aleatorios (Random Forest):** Es un algoritmo de aprendizaje automático basado en ensambles que combina múltiples árboles de decisión, cada uno de ellos se entrena con una muestra aleatoria de datos y se extrae la predicción final.
- **Red neuronal convolucional (CNN):** Especializada en el empleo de capas convolucionales para capas de agrupamiento que conservan características significativas tras detectar patrones locales y reducir la dimensionalidad en datos matriciales, como las fotografías.
- **K-vecinos más próximos (kNN):** Corresponde a método muy simple de aprendizaje supervisado que clasifica nuevos puntos de datos de acuerdo a la mayoría de sus k vecinos más cercanos en el espacio de características.
- **Red neuronal profunda (DNN):** Similar a una red neuronal convencional, pero tiene capas ocultas adicionales que permiten el aprendizaje de representaciones de datos aún más complejas.
- **Redes neuronales de memoria a corto y largo plazo (LSTM):** Se refiere a un tipo particular de red neuronal recurrente (RNN) que se utiliza para simular relaciones a largo plazo en secuencias y trata el problema del gradiente de fuga.
- **Autoencoder (AE):** Una red neuronal que intenta aprender una representación comprimida de los datos de entrada a través de una estructura de codificación, útil para reducción de dimensionalidad y generación de datos.
- **Árbol de decisión (DT):** Un método de aprendizaje supervisado que crea un modelo en forma de estructura de árbol, donde cada nodo interno representa una característica, y cada hoja una etiqueta.

- **Extreme gradient boosting (XGBoost):** Un algoritmo de aprendizaje automático basado en árboles de decisión que enfatiza la velocidad y el rendimiento al mejorar secuencialmente los árboles anteriores.
- **Modelo de comportamiento normal (NBM):** Un modelo estadístico que asume que los datos se distribuyen según una distribución normal, lo que permite realizar inferencias basadas en esta distribución.

Tabla 1. Ventajas y desventajas de modelos de aprendizaje automático

Modelo	Ventaja	Desventaja
ANN	Capacidad para aprender patrones complejos.	Dificultad para interpretar cómo y por qué se llega a una predicción o decisión.
	Adaptable a diferentes tipos de datos.	Propenso a sobreajuste.
SVM	Efectivo en espacios de alta dimensión.	Pueden ser computacionalmente costosas.
	Robusto ante sobreajuste en casos de dimensiones mayores a muestras.	Requiere selección cuidadosa del kernel.
RF	Maneja bien datos heterogéneos y faltantes.	Mayor complejidad computacional.
	Menos propensos al sobreajuste debido a la naturaleza de ensamble.	Menos efectivo en extrapolación fuera del rango de datos de entrenamiento.
CNN	Altamente efectivas en tareas de visión por computadora.	Pueden ser computacionalmente costosas.
	Gran capacidad para aprender jerarquías de características complejas en imágenes.	Requieren grandes cantidades de datos para evitar el sobreajuste.
kNN	Fácil de entender.	Sensible a datos ruidosos.
	No hace suposiciones sobre la distribución de los datos.	Puede ser costoso con grandes conjuntos de datos al almacenar todo.
DNN	Capacidad para aprender representaciones complejas.	Alto riesgo de sobreajuste.
	Buen rendimiento en datos no estructurados o no lineales.	Pueden ser computacionalmente costosas.
LSTM	Efectivas en modelar dependencias en secuencias.	Mayor tiempo de entrenamiento.
	Puede capturar dependencias a largo plazo.	Complejo y difícil de entrenar
AE	Reduce la dimensionalidad y es útil en la extracción de características.	Sensible a la calidad y cantidad de datos de entrada.
	Puede aprender representaciones no lineales de los datos.	Requiere ajuste fino de parámetros y arquitectura de la red.
DT	Fácil de interpretar.	Sensible a variaciones en los datos.
	No requiere normalización de datos.	
XGBoost	Alto rendimiento y velocidad.	Puede ser propenso a overfitting sin una cuidadosa regularización.
	Maneja bien diversas distribuciones de datos.	
NBM	Simple de implementar.	Puede ser demasiado simplista.
	Eficaz con suposiciones de normalidad.	Rendimiento limitado si las suposiciones de normalidad no se cumplen.

4.5.4. Modelos híbridos de inteligencia artificial

Los modelos híbridos de IA representan la convergencia de múltiples técnicas y enfoques de IA diseñados específicamente para abordar problemas específicos de manera más efectiva. Al integrar las fortalezas únicas de diversas metodologías, estos modelos crean soluciones más robustas y confiables, superando así las limitaciones individuales de cada técnica. Esta sinergia permite alcanzar una precisión superior en tareas como la predicción de fallas, mediante la combinación estratégica de diferentes tecnologías de IA.

Estos avances en la aplicación de modelos y algoritmos híbridos no solo han demostrado ser efectivos en áreas como la simulación de riego de superficie y la predicción en la producción de energía, sino que también han encontrado aplicaciones significativas en otros campos. En el **Anexo 1. Revisión Sistemática de Literatura**, del informe de la RSL, son examinados varios enfoques híbridos empleados para la predicción de fallas en aerogeneradores, también, de dichos enfoques se incluyen DNN/GMM, Ensemble W-kNN, R-NN-EL y LSTM-AE, por otra parte, en la Tabla 2 se detalla ventajas y desventajas asociadas al uso de estos algoritmos híbridos, basándose en las investigaciones extraídas de [32], [33], [34].

- **NN/GMM:** Integra una red neuronal profunda (DNN) con un modelo de mezcla gaussiana (GMM), este tipo de red se especializa en el aprendizaje profundo para capturar patrones complejos, mientras que el GMM se utiliza para modelar distribuciones de datos mediante la combinación de varias distribuciones gaussianas [34], con ello, la unión de ambas técnicas puede ser importante de implementarse en tareas de clasificación o reconocimiento de patrones.
- **NN/GMM:** Integra una red neuronal profunda (DNN) con un modelo de mezcla gaussiana (GMM), este tipo de red se especializa en el aprendizaje profundo para capturar patrones complejos, mientras que el GMM se utiliza para modelar distribuciones de datos mediante la combinación de varias distribuciones gaussianas [34], con ello, la unión de ambas técnicas puede ser importante de implementarse en tareas de clasificación o reconocimiento de patrones.
- **Ensemble W-kNN:** Combina técnicas de ensamblado con el algoritmo kNN, además de los modelos de ensamblado, como Bagging o Boosting que se utilizan para combinar múltiples clasificadores kNN para mejorar la precisión

y generalización del modelo según [34], permitiendo mejorar la gestión de conjuntos de datos complejos y el rendimiento de la predicción

- **R-NN-EL:** Se incorpora una red neuronal recurrente (RNN) con regularización Elastic Net (EL), son efectivas para modelar secuencias temporales, y la regularización Elastic Net que evita el sobreajuste y mejora la generalización del modelo [33], de esta manera, dicha combinación puede 22 ser útil para tareas de predicción en series temporales con datos relativamente ruidosos o con una elevada dimensionalidad.
- **LSTM-AE:** Se fusiona una red neuronal LSTM con un autoencoder, logrando que sea especializada en modelar dependencias a largo plazo en secuencias, mientras que, los autoencoders son útiles para la reducción de dimensionalidad y la extracción de características relevantes de los datos [33], así dicha combinación puede ser efectiva para tareas como la reconstrucción de secuencias o representación de datos secuenciales.

Tabla 2. Ventajas y desventajas de modelos híbridos de IA

Modelo híbrido	Ventaja	Desventaja
DNN/GMM	Puede capturar relaciones complejas en los datos.	DNN puede ser propenso a sobreajuste con conjuntos de datos pequeños.
	GMM es útil para modelar distribuciones de datos.	Requiere una cantidad considerable de datos para el entrenamiento eficaz.
Ensemble W-kNN	Mejora la precisión de la clasificación al combinar múltiples clasificadores kNN.	Requiere más recursos computacionales debido al ensamblaje de múltiples modelos.
	Menor susceptibilidad al ruido en los datos debido a la combinación de múltiples clasificadores.	Puede ser menos efectivo si los clasificadores base (kNN) tienen un desempeño similar.
R-NN-EL	RNN es efectiva para modelar dependencias temporales en secuencias.	RNN puede tener dificultades con secuencias largas (problema de memoria a largo plazo).
	Regularización Elastic Net ayuda a evitar el sobreajuste.	Mayor complejidad de ajuste de hiperparámetros con la regularización adicional.
LSTM-AE	LSTM es efectiva para modelar dependencias a largo plazo en secuencias.	Puede ser complejo de entrenar y requiere ajuste fino de hiperparámetros.
	Autoencoder ayuda en la extracción de características útiles y la reducción de dimensionalidad.	LSTM puede tener problemas para manejar patrones complejos en datos de series temporales muy ruidosos.

4.5.5. Técnicas de combinación de algoritmos de ML

Las técnicas de combinación de algoritmos en ML para implementar modelos híbridos se centran en la integración de múltiples algoritmos o modelos de

aprendizaje automático para aprovechar sus fortalezas individuales y mejorar el rendimiento general del modelo, de este modo, dichas estrategias se utilizan para combinar las predicciones de varios modelos y generar un modelo de mayor precisión o estabilidad, así varias de estas técnicas incluyen: Bagging, Stacking, Boosting, Voting, Blending, Weighted Average Ensemble, donde cada técnica presenta varias características que pueden ser seleccionadas de acuerdo a las necesidades específicas del problema y la disponibilidad de los datos.

- **Bagging (Bootstrap Aggregating):** Corresponde a una estrategia de combinación de modelos que se fundamenta en la creación de múltiples conjuntos de datos de entrenamiento mediante el muestreo con reemplazo, así de esta manera cada conjunto es implementado para entrenar un modelo base, y a continuación, las predicciones de estos modelos se combinan para obtener una predicción final, cuya representación se puede observar en la Figura 3.

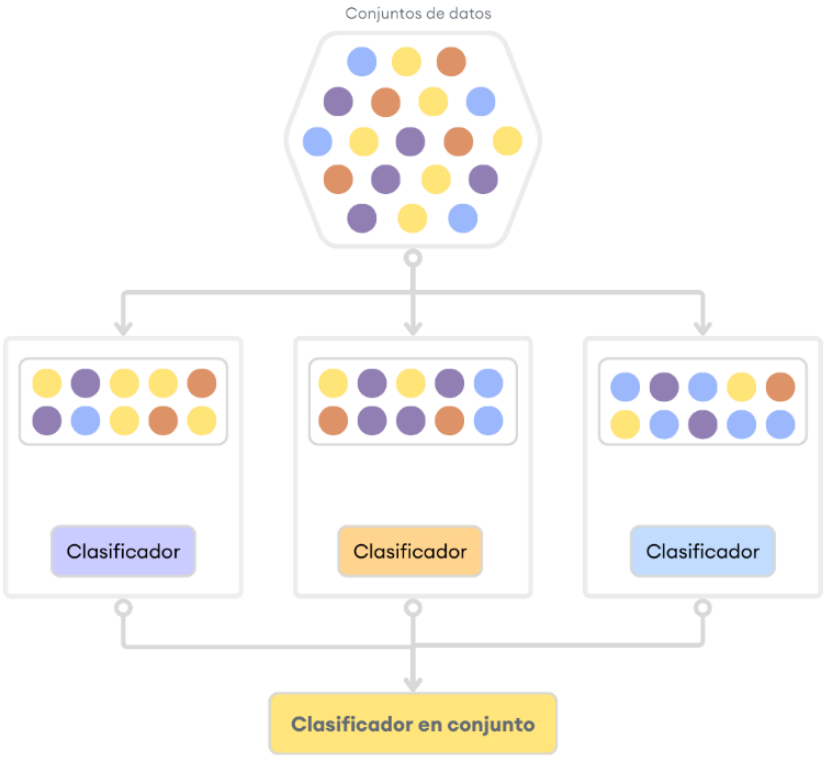

Figura 3. Arquitectura de Bagging

Esta técnica es especialmente eficaz cuando se trata de reducir la variabilidad y mejorar la estabilidad del modelo, siendo útil en problemas propensos al sobreajuste, no obstante, puede aumentar los requisitos computacionales y tecnológicos debido a la necesidad de entrenar varios

modelos, esta arquitectura es introducida en [35] y su aplicación es ejemplificada en [33], [36], [37].

- **Stacking (Stacked Generalization):** Combina la predicción de varios modelos base utilizando un modelo adicional conocido como meta-aprendiz, para el cual, primeramente, se entrenan modelos base independientes con diferentes algoritmos, seguidamente, las predicciones de estos modelos se implementan como entradas para el meta-aprendiz que efectúa la predicción final, tal como lo indica la Figura 4 que ilustra la arquitectura de Stacking.

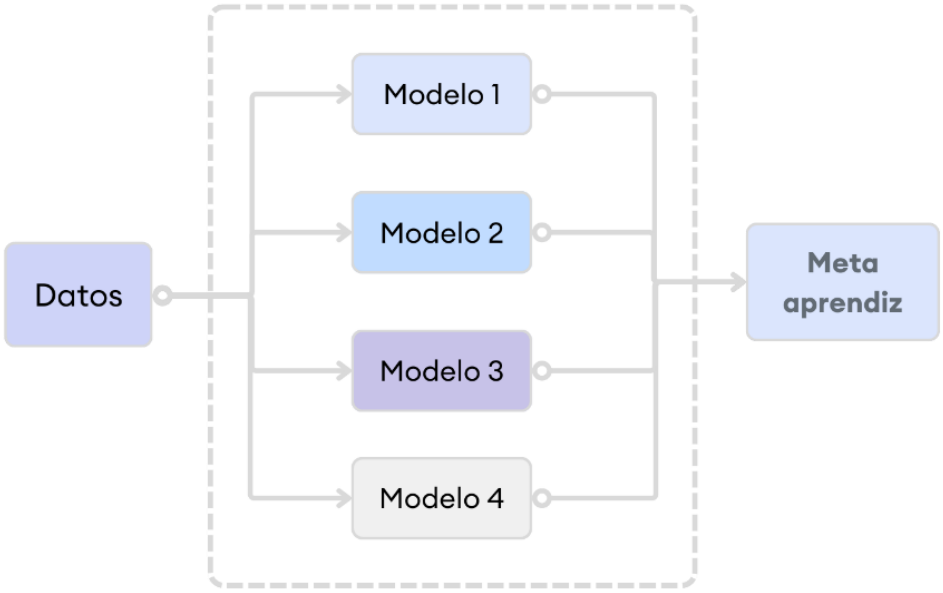

Figura 4. Arquitectura de Stacking

Stacking es muy versátil y puede aplicarse a cualquier algoritmo de aprendizaje con el fin de aprovechar las todas las fortalezas de diversos modelos y mejorar la precisión, no obstante, su implementación puede ser computacionalmente costosa, debido a la implicación del entrenamiento de múltiples modelos, esta técnica fue propuesta en [38] y se evidencia en [36], [37], [39].

- **Boosting:** Construye varios modelos de forma secuencial donde cada uno va corrigiendo los errores del modelo anterior, de este modo, todo modelo débil resultante se enfoca en los casos mal clasificados por los modelos anteriores, mejorando progresivamente su rendimiento y dicha arquitectura se puede observar en la Figura 5.

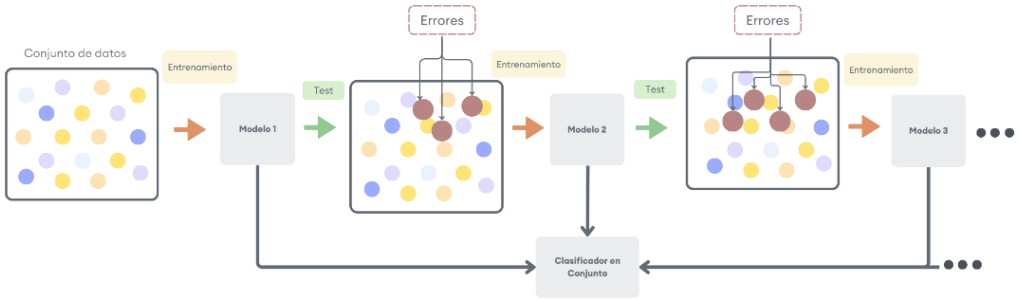

Figura 5. Arquitectura de Boosting

AdaBoost y Gradient Boosting son ejemplos comunes de esta técnica, sin embargo, aunque es considerada eficiente para reducir el sesgo y mejorar la precisión, es sensible al ruido y valores no comunes en los datos, esta técnica puede ser observada en [33], [36] y en variantes como Gradient Boosting como en [33], [37], [40].

- **Voting:** Combina las predicciones de varios modelos individuales, los cuales pueden ser de dos tipos, el primero Hard Voting, donde la clase predicha es seleccionada por mayoría de votos, y la segunda Soft Voting, donde se asignan pesos a las predicciones y se promedian para calcular la predicción final, por otra parte, Voting es simple de implementar y puede mejorar la precisión en comparación con modelos individuales, pero puede no ser muy efectivo si los modelos base se encuentran altamente correlacionados, así lo indica la Figura 6 que representa la arquitectura de Voting en el modo Hard Voting.

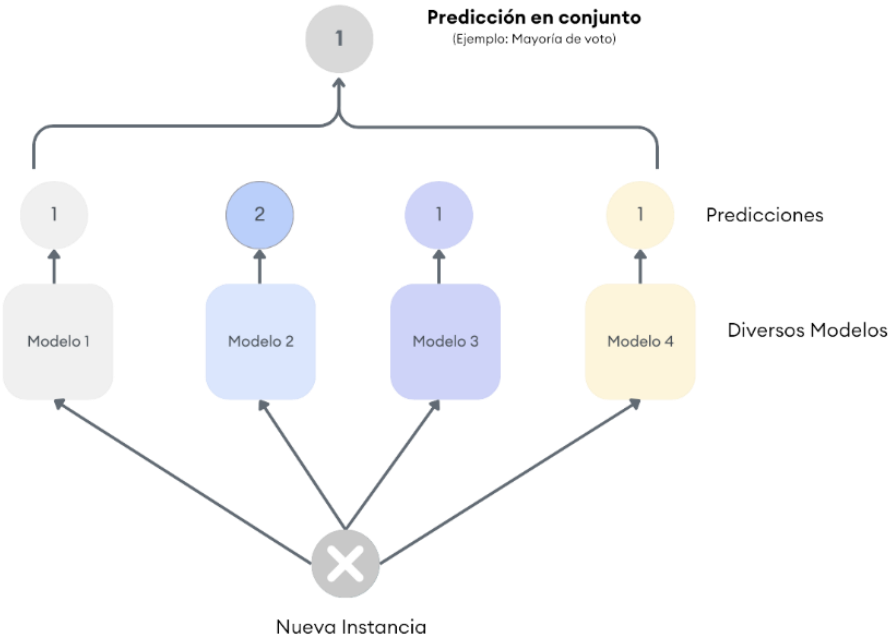

Figura 6. Arquitectura de Voting

Esta técnica puede ser sensible a modelos más ajustados; si alguno de los modelos implementados en la votación se encuentra mal configurado, podría afectar negativamente la precisión del clasificador, también, se puede observar en variantes como weighted majority voting en enfoques como [37], [41].

- **Blending Ensemble:** Es similar a Stacking, con la particularidad que utiliza dos conjuntos de datos separados para entrenar el modelo base y el modelo final, dichos modelos base se entrenan en un conjunto de datos y sus predicciones se utilizan como entrada para el modelo final, que se entrena en otro grupo de datos tal como lo indica la Figura 7 que presenta la arquitectura de Blending.

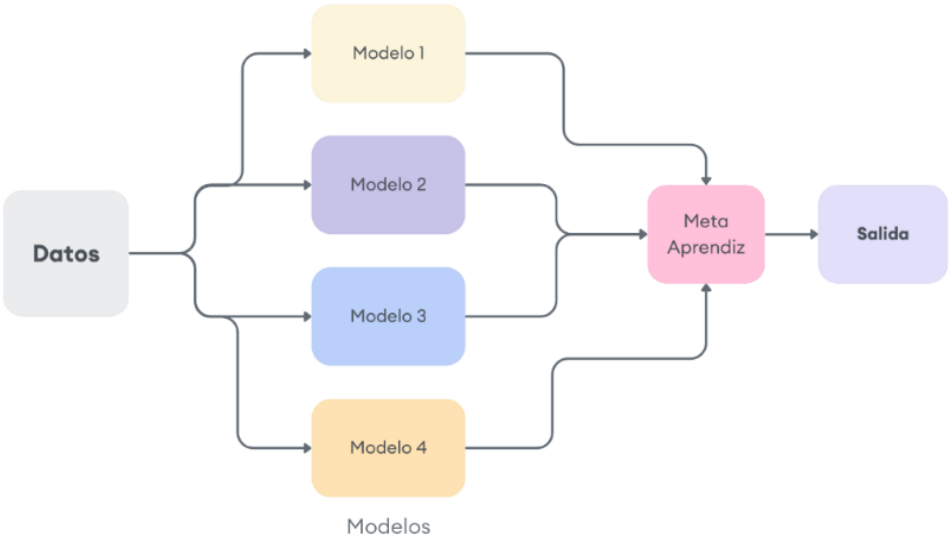

Figura 7. Arquitectura de Blending

Esta técnica puede reducir el riesgo de sobreajuste en comparación con Stacking, no obstante, puede aumentar la complejidad del proceso y se puede localizar en diferentes propósitos como lo indica [42].

- **Weighted Average Ensemble:** Combina las predicciones asignando pesos a cada modelo base, con dichos pesos se determinan según la confianza en la precisión de cada modelo permitiendo asignar importancia diferenciada a modelos considerados más confiables, mejorando así la precisión y la estabilidad del modelo final con una calibración adecuada de los pesos, tal como lo indica la Figura 8 que exhibe la arquitectura de Weighted Average Ensemble.

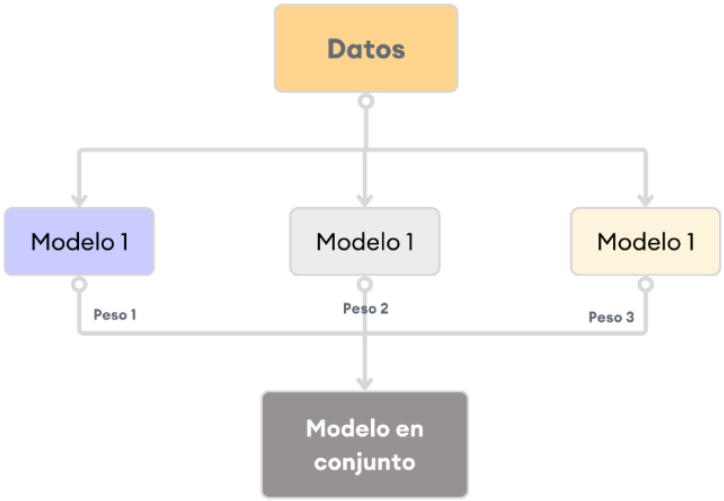

Figura 8. Arquitectura de Weighted Average Ensemble

Esta técnica se presenta con diversas variaciones, como se ilustra en el estudio [43] o en estudios similares, como el realizado en [44] en donde se propone configurar los pesos de las instancias a través de un factor aplicado a un SVM estimado.

4.6. Técnicas de sobremuestreo

Las técnicas de sobremuestreo son implementadas en el procesamiento de conjuntos de datos desequilibrados, donde una clase tiene significativamente menos ejemplos que otras, dichas técnicas se aplican con la finalidad de generar ejemplos sintéticos de la clase minoritaria para equilibrar el conjunto de datos y por lo general, se usan en el contexto del aprendizaje automático para abordar el desafío que representa el desequilibrio de clases, seguidamente, en la Figura 9 se ilustra el funcionamiento del sobremuestreo y se presenta varias de las técnicas de sobremuestreo más comunes utilizadas para enfrentar el desequilibrio de datos y en la Tabla 3 se detalla ventajas y desventajas de dichas técnicas.

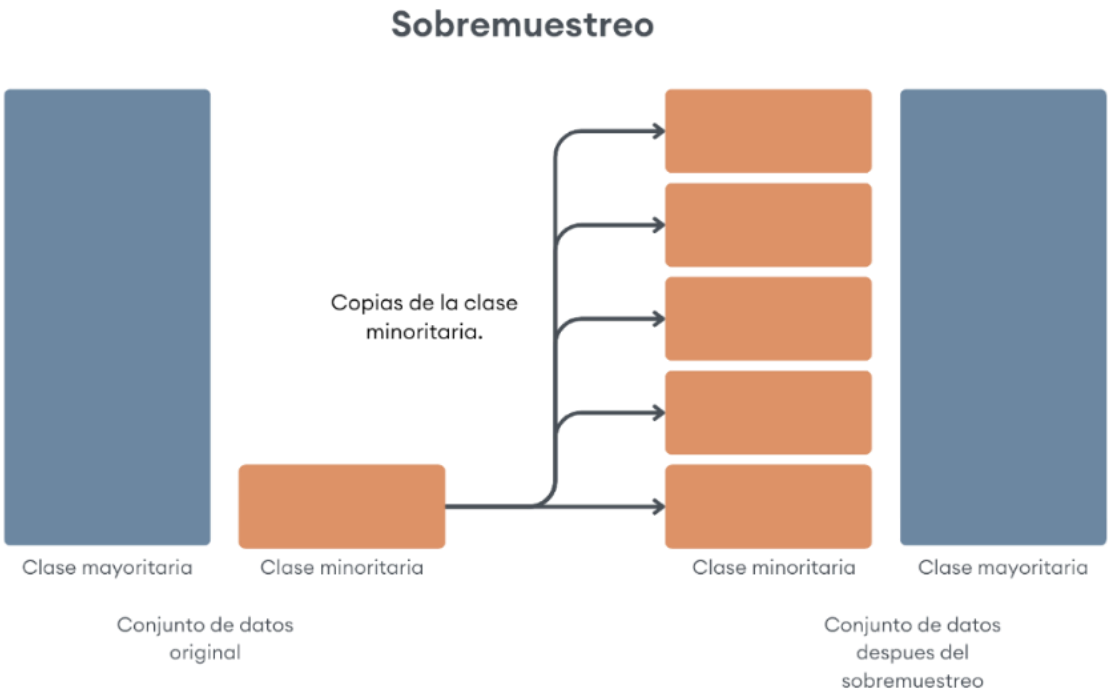

Figura 9. Representación del sobremuestreo

4.6.1. SMOTE (Synthetic Minority Over-Sampling Technique)

Este método, conocido como Synthetic Minority Over-Sampling Technique (SMOTE), se utiliza para aumentar el tamaño de la clase minoritaria generando patrones sintéticos [45], cuya idea principal detrás de SMOTE es generar nuevas muestras de datos sintéticas distribuidas a lo largo del segmento de línea que conecta ejemplos de la clase minoritaria existentes, además, con el paso de los años, se han propuesto múltiples variaciones de SMOTE que han demostrado su efectividad en estudios como en [46].

4.6.2. SMOTE-ENN (Synthetic Minority Over-sampling Technique) y ENN (Edited Nearest Neighbors).

SMOTE-ENN es una técnica que combina dos modelos de preprocesamiento de datos, por un lado, SMOTE (Synthetic Minority Oversampling Technique) y ENN (Edited Nearest Neighbors), dicha fusión tiene como finalidad disminuir la complejidad de los datos al eliminar instancias ruidosas y contrarrestar el solapamiento entre las clases, además, su aplicación se encuentra documentada y evidenciada en artículos como [36], [47], [48].

4.6.3. ADASYN (Adaptive Synthetic Sampling)

Es una técnica de preprocesamiento de datos utilizada para abordar el problema del desequilibrio de clases en conjuntos de datos de aprendizaje automático, puede por otra parte ser implementado en diversos entornos y es

adecuado para aplicaciones del mundo real con conjuntos de datos desequilibrados, no obstante, ADASYN no solo reduce el sesgo de aprendizaje causado por la distribución inicial de datos, sino que también ajusta de forma adaptativa el límite de decisión para enfocarse en muestras más difíciles de aprender [49], dicha información de la técnica es descrita en detalle en [50] y aplicada en múltiples artículos como [36], [50], [51], [52], [53], [54].

4.6.4. SMOTE-Tomek (SMOTE + Tomek Links)

Es una técnica consolidada que combina dos enfoques específicos para abordar el desbalance de clases en un conjunto de datos, además, dicha combinación aprovecha el sobremuestreo sintético de la clase minoritaria a través de SMOTE, en conjunto con la limpieza de la clase mayoritaria mediante Tomek Links que juntas estas estrategias contribuyen a crear un espacio de características más coherente y muy bien definido, donde las fronteras entre las clases se vuelven más nítidas, lo que a su vez, puede simplificar una clasificación más precisa y eficaz, mejorando la habilidad del modelo para distinguir entre las diferentes clases, incluso cuando hay un desequilibrio considerable en su distribución y su aplicación se puede conocer en [36], [55].

4.6.5. Borderline-SMOTE (Borderline Synthetic Minority Over-sampling Technique)

Borderline Synthetic Minority Over-sampling Technique (BorderlineSMOTE) es una variante especializada de la técnica SMOTE que se encuentra diseñada para mejorar el manejo de las instancias minoritarias que se localizan cerca de la frontera entre las clases, dicha técnica identifica ejemplos de la clase minoritaria ubicados en proximidad a la frontera con la clase mayoritaria, y genera instancias sintéticas a raíz de estos puntos críticos. La variante BorderlineSMOTE es ampliamente útil cuando las instancias fronterizas desempeñan un papel crítico en la tarea de clasificación cuando tratan de enfocarse en estas áreas complejas y desafiantes, la técnica es capaz de ofrecer una mejora en el rendimiento de la clasificación en comparación con las técnicas de sobremuestreo más convencionales [56] y esta técnica se describe originalmente en [57] y su aplicación se puede observar en [52], [58], [56].

Tabla 3**.** Ventajas y desventajas de técnicas de sobremuestreo

Técnica	Ventaja	Desventaja
SMOTE	Aumenta el tamaño de la clase minoritaria generando ejemplos sintéticos.	Puede generar ejemplos sintéticos que se encuentren en zonas de ruido o sobreajuste.
	Ayuda a mitigar el desequilibrio de clases sin eliminar muestras.	Sensible a valores atípicos y puede generar muestras sintéticas poco representativas.
SMOTE-ENN	Combina SMOTE con la eliminación de ejemplos ruidosos.	Puede ser computacionalmente costoso al requerir la ejecución de SMOTE y ENN secuencialmente.
	Mejora la calidad de los datos generados al remover ejemplos sintéticos que podrían ser ruido.	Puede eliminar ejemplos relevantes para la clase minoritaria.
ADASYN	Ajusta la densidad de muestreo sintético basado en la dificultad de aprendizaje de cada ejemplo.	Sensible a la calidad de las características y a la distribución inicial de datos.
	Da más énfasis a la generación de ejemplos en áreas más difíciles de aprender.	Puede generar sesgos al favorecer ciertos ejemplos o clases más que otros.
SMOTE-Tomek	Combina SMOTE con el algoritmo Tomek para eliminar ejemplos sintéticos cercanos a la frontera de decisión.	Puede eliminar ejemplos relevantes para la clase minoritaria si se aplican restricciones excesivas.
	Mejora la calidad de los ejemplos sintéticos generados por SMOTE.	
Borderline-SMOTE	Se enfoca en las muestras cercanas al límite de decisión entre clases, generando ejemplos sintéticos solo para estas instancias.	Sensible a la calidad de los datos y la selección de los ejemplos fronterizos.
	Reduce el riesgo de generar ejemplos sintéticos en regiones ruidosas o menos informativas.	

4.7. Medidas de desempeño del modelo de predicción

Las medidas de desempeño del modelo de predicción son métricas utilizadas para evaluar qué tan bien un modelo de aprendizaje automático realiza predicciones sobre un conjunto de datos. Estas medidas proporcionan información sobre la precisión, la calidad y la capacidad predictiva del modelo. Algunas de las medidas de desempeño más comunes incluyen:

4.7.1. Validación cruzada

En el campo del aprendizaje automático, la validación cruzada (Cross-validation) es un método crucial para medir la eficacia y la capacidad de generalización de un modelo, dicho método consiste en segmentar el conjunto de datos en varios grupos menores y emplearlos alternativamente como datos de entrenamiento y de prueba, de esta forma, es factible evaluar el modelo con todo el conjunto de datos, manteniendo separados los datos de entrenamiento y prueba en

cada paso según [59], seguidamente, se representa la validación cruzada en la Figura 10.

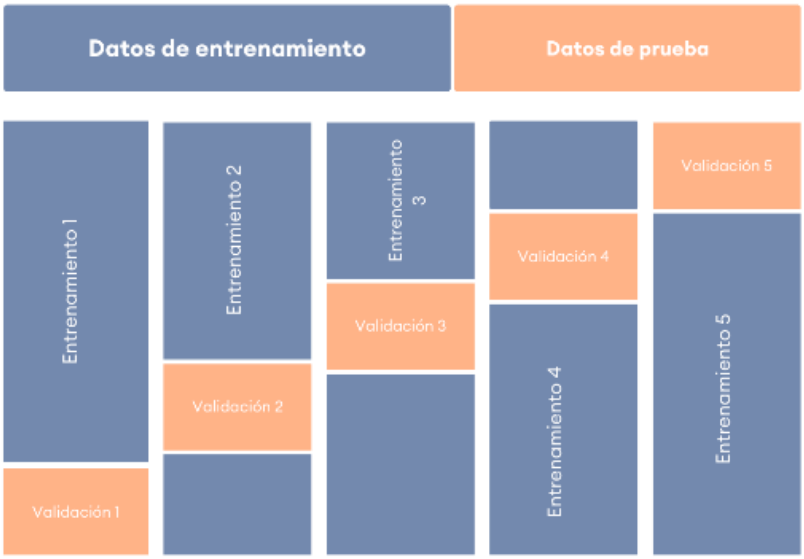

Figura 10. Representación de la validación cruzada. Adaptado de [60].

En este ejemplo el conjunto de entrenamiento se divide en cinco pliegues y en cada uno de ellos se considera uno para validar el modelo mientras los restantes se utilizan para efectuar el entrenamiento, todo el proceso se repite en cinco ocasiones, utilizando cada pliegue como conjunto de validación una vez, con ello, cuando se finaliza, se emplea la media de los resultados obtenidos en las cinco iteraciones como una estimación más robusta del rendimiento del modelo.

Entre las ventajas más notables de utilizar la validación cruzada, como se evidencia en [61], se destaca:

- **Evaluación de rendimiento:** La validación cruzada proporciona una visión detallada del desempeño del modelo en datos que no ha visto anteriormente, esto se realiza mediante la implementación de diversos grupos de datos de prueba, así, es posible obtener una valoración más sólida y confiable de su capacidad.
- **Prevención del sobreajuste:** La aplicación de la validación cruzada permite identificar si el modelo está adaptándose en exceso a los datos de entrenamiento, también puede presentar el caso de un rendimiento destacado en el entrenamiento, pero deficiente en la prueba siendo un indicador de sobreajuste.
- **Eficiencia:** A pesar de que la validación cruzada puede demandar una cantidad significativa de recursos computacionales (en especial con un alto

número de divisiones), tiene la ventaja de aprovechar todo el conjunto de datos para el entrenamiento y la prueba, lo cual resulta especialmente beneficioso en ocasiones donde la cantidad de datos es limitada.

4.7.2. Índice de desbalance

El índice de desbalance o IR (Imbalance Ratio) se emplea para cuantificar el desequilibrio entre las clases en un conjunto de datos, asimismo, es relevante cuando la distribución de clases no es uniforme, permitiendo evaluar la disparidad entre ellas, donde un índice de desbalance de 1 indica que las clases están equilibradas, cuanto mayor sea este índice, más desequilibrado estará el conjunto de datos y esto quede demostrado en la Ecuación 2 que describe formalmente este índice y su cálculo.

Ecuación 2. Índice de desbalance

$$Índice\ de\ desbalance\ = \frac{Número\ de\ muestras\ de\ la\ clase\ mayoritaria}{Número\ de\ muestras\ de\ la\ clase\ minoritaria} \quad (2)$$

4.7.3. Matriz de confusión

La matriz de confusión o "Confusion Matrix" se utiliza a menudo en problemas de clasificación para evaluar el rendimiento de un algoritmo, dicha matriz contrasta las clases predichas por un modelo con las clases reales [62], así, en una clasificación binaria, la matriz de confusión se estructura de 4 valores distintos detallados en la Figura 11.

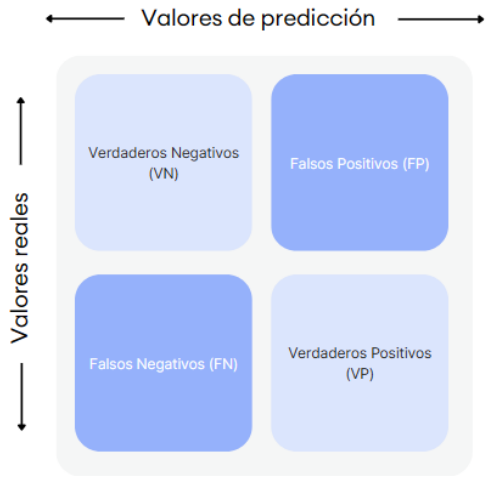

Figura 11. Matriz de confusión.

- **Verdaderos positivos (VP):** Casos en los que la clase positiva fue correctamente identificada como positiva.
- **Verdaderos negativos (VN):** Casos en los que la clase negativa fue correctamente identificada como negativa.

- **Falsos positivos (FP):** Casos en los que la clase negativa fue incorrectamente identificada como positiva.
- **Falsos negativos (FN):** Casos en los que la clase positiva fue incorrectamente identificada como negativa.

La matriz de confusión es útil para calcular diversas métricas importantes, como la precisión, sensibilidad, especificidad, y la puntuación F1 [62]. A diferencia de la precisión general, que puede ser insuficiente para comprender completamente el rendimiento del modelo, estas métricas ofrecen una visión más detallada y completa del comportamiento del algoritmo en diferentes aspectos de la clasificación. En la Tabla 4, se detallan ventajas y desventajas de usar estas medidas de desempeño.

Tabla 4. Ventajas y desventajas de las medidas de desempeño

Medida	Ventaja	Desventaja
Validación Cruzada	Mejor estimación del rendimiento del modelo al usar varias divisiones de los datos.	Puede ser computacionalmente costosa.
	Reduce la variabilidad en el cálculo del rendimiento del modelo al estimar los resultados de las diferentes iteraciones.	Necesita más tiempo de ejecución, especialmente en grupos de datos grandes.
	Aprovecha mejor el conjunto de datos disponible, aumentando su uso tanto para entrenamiento como para validación.	Puede ser menos efectiva en grupos de datos considerablemente pequeños, donde se limita la cantidad de muestras para validación.
Índice de Desbalance	Identifica y cuantifica la disparidad entre las clases, permitiendo comprender el desequilibrio en los datos.	No indica específicamente la efectividad de las estrategias de manejo del desequilibrio, solo señala su presencia.
	Ayuda a tomar decisiones informadas sobre qué medidas tomar para abordar el desequilibrio, como técnicas de remuestreo o ponderación de clases.	La elección de técnicas de manejo del desequilibrio puede ser un proceso empírico que depende del contexto y la naturaleza de los datos.
Matriz de Confusión	Proporciona información específica del rendimiento del modelo para cada clase.	Visión limitada en problemas de clases desbalanceadas, donde se pueden subestimar métricas clave como la precisión.
	Permite estimar métricas de desempeño como precisión, recuperación, puntuación F1, entre otras, que son trascendentales para comprender el comportamiento del modelo.	No proporciona directamente una sola métrica para detallar el rendimiento del modelo.

4.7.4. Métricas de evaluación especificas

Las métricas de evaluación específicas, que se derivan de la matriz de confusión, desempeñan un papel fundamental en la medición de la calidad y el rendimiento de un modelo de clasificación [62], [63], dichas métricas ofrecen una visión detallada de cómo el modelo se comporta en diferentes variables, como la capacidad de identificar correctamente las clases y el equilibrio entre la detección de verdaderos positivos y la reducción de falsos positivos

- **Precisión (Precision)**

  Mide la proporción de identificaciones positivas que fueron realmente correctas siendo útil cuando las consecuencias de los falsos positivos son importantes, este índice y su cálculo se describen formalmente en la Ecuación 3 detallada a continuación.

Ecuación 3. Precisión

$$Precisión = \frac{Verdaderos\ Positivos\ (VP) + Falsos\ Positivos\ (FP)}{Verdaderos\ Positivos\ (VP)} \quad (3)$$

- **Recuperación (Recall)**

  También conocida como sensibilidad, es la proporción de positivos reales que se identificaron correctamente y es fundamental cuando es importante capturar todos los casos positivos relevantes, este índice y su cálculo se detallan formalmente en la Ecuación 4 descrita a continuación

Ecuación 4. Recuperación

$$Recuperación = \frac{Verdaderos\ Positivos\ (VP) + Falsos\ Negativos\ (FN)}{Verdaderos\ Positivos\ (VP)} \quad (4)$$

- **Puntuación F1 (F1 Score)**

  Es la media armónica de la precisión y la recuperación que ofrece un equilibrio entre estas dos métricas y es útil cuando se quiere un balance entre precisión y recuperación, este índice y su cálculo se describen formalmente en la Ecuación 5 detallada a continuación.

Ecuación 5. Puntuación F1

$$Puntuación\ F1 = 2\ x \frac{Precisión\ x\ Recuperación}{Precisión\ x\ Recuperación} \quad (5)$$

- **Sensibilidad (Sensitivity)**

  También conocida como recuperación y, por lo tanto, comparte la misma descripción y aplicaciones que la recuperación, por ende, se estima

de la misma manera que la recuperación, este índice y su cálculo de detallan formalmente en la Ecuación 6 descrita a continuación.

Ecuación 6. Sensibilidad

$$Sensibilidad = \frac{Verdaderos\ Positivos\ (VP) + Falsos\ Negativos\ (FN)}{Verdaderos\ Positivos\ (VP)} \quad (6)$$

- **Especificidad (Specificity):**

La especificidad mide la proporción de negativos reales que se identificaron correctamente siendo fundamental en contextos donde es vital minimizar los falsos positivos, este índice y su cálculo de detallan formalmente en la Ecuación 7 descrita a continuación.

Ecuación 7. Especificidad

$$Especificidad = \frac{Verdaderos\ Negativos\ (VN)}{Verdaderos\ Negativos\ (VN) + Falsos\ positivos\ (FP)} \quad (7)$$

- **Exactitud (Accuracy)**

La exactitud es la proporción de predicciones correctas entre el número total de casos y es una métrica común para la evaluación general del rendimiento del modelo, este índice y su cálculo de detallan formalmente en la Ecuación 8 descrita a continuación.

Ecuación 8. Exactitud

$$Exactitud = \frac{(VP) + (VN)}{(VP) + (VN) + (FP) + (FN)} \quad (8)$$

Donde:

- VP = Verdaderos Positivos
- VN = Verdaderos Negativos
- FP = Falsos Positivos
- FN = Falsos Negativos

En la Tabla 5, se detallan ventajas y desventajas de usar estas métricas de evaluación específica.

Tabla 5. Ventajas y desventajas de las métricas de evaluación específica.

Métrica	Ventaja	Desventaja
Precisión	Es una métrica sencilla de entender y calcular: muestra la proporción de predicciones adecuadas sobre el total de predicciones.	No es adecuada en casos de clases desbalanceadas, ya que, puede ser engañosa y no reflejar correctamente el rendimiento del modelo.
	Útil cuando las clases se encuentran equilibradas.	Puede proporcionar una visión sesgada en problemas donde la precisión no es la única métrica relevante.
Recuperación	Se centra en la capacidad del modelo para identificar adecuadamente las instancias de una clase, minimizando los falsos negativos.	Podría no ser ideal cuando la cantidad de falsos positivos es crítica, ya que, se puede ignorar este aspecto.
	Es útil en problemas donde es fundamental minimizar los falsos negativos.	
Puntuación F1	Combina precisión y recuperación en una sola métrica, lo que la hace importante en casos donde se busca un balance entre ambas.	Puede ser más difícil de interpretar que medidas individuales como la precisión o la recuperación.
	Se adapta muy bien a conjuntos de datos desbalanceados.	
Sensibilidad	Enfocada en la capacidad del modelo para identificar correctamente las instancias positivas, disminuyendo los falsos negativos.	No considera los falsos positivos, lo que podría ocasionar un problema considerando la magnitud del contexto.
	Importante en problemas donde minimizar los falsos negativos es crítico.	
Especidad	Estima la capacidad del modelo para identificar correctamente las instancias negativas, disminuyendo los falsos positivos.	No considera los falsos negativos, lo que puede representar un problema en ciertas situaciones.
	Útil en problemas donde minimizar los falsos positivos es crucial.	
Exactitud	Mide la proporción de predicciones correctas en relación con todas las predicciones realizadas.	Es engañosa en conjuntos de datos desbalanceados, donde no se visualiza de forma correcta el rendimiento del modelo.
	Fácil de interpretar y calcular.	

4.7.5. Curva ROC

La Curva ROC (Receiver Operating Characteristic) es empleada para evaluar la capacidad de un modelo de clasificación en discriminar entre las clases positivas y negativas [64], su grafica curva se traza representando la sensibilidad en función de (1 - especificidad) para diversos umbrales de decisión, dichas especificaciones se presentan en la Figura 12 que plantea un ejemplo gráfico de la apariencia típica de una Curva ROC.

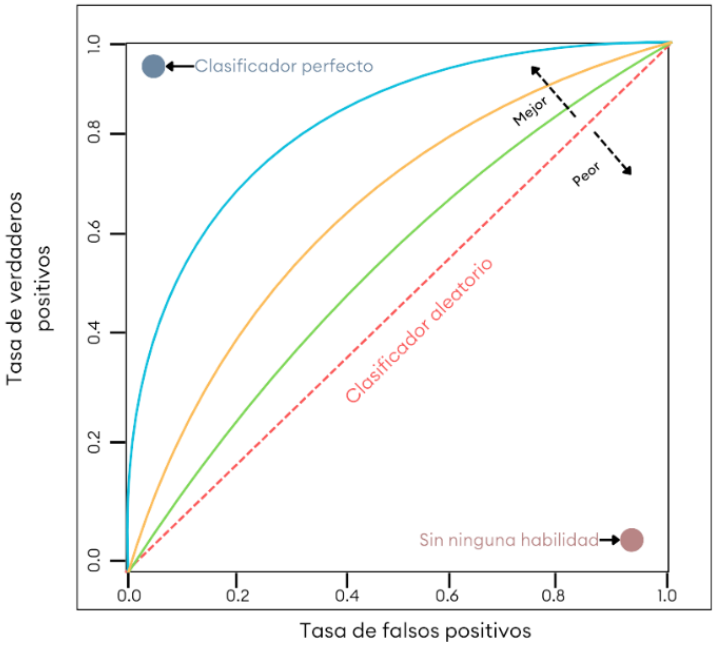

Figura 12. Curva ROC. Adaptado de [65].

En la curva Roc, el eje *x* muestra la tasa de falsos positivos (proporción de negativos reales clasificados erróneamente), mientras que el eje *y* refleja la tasa de verdaderos positivos (proporción de positivos reales identificados correctamente), así, un clasificador adecuado se sitúa en el punto donde la tasa de verdaderos positivos es 1 y la de falsos positivos es 0, especificando una clasificación sin errores.

En contraste, un clasificador aleatorio efectúa predicciones sin criterio ubicándose en la línea diagonal, siendo cualquier modelo útil aquel que se localice por encima de esta línea, así, en la Figura 12 se observan tres curvas, cada una representa un modelo distinto, siendo la curva más cercana al clasificador aceptable lo que indica el modelo con mayor rendimiento cuyas palabras "mejor" y "peor" se utilizan para indicar la dirección hacia la cual el rendimiento del modelo mejora o empeora respectivamente, de esta manera, el análisis se fundamenta en la metodología presentada en [62] donde se introducen y explican los conceptos de la curva ROC.

- **Área bajo la Curva (AUC)**

  Es una métrica que cuantifica el rendimiento general de la curva ROC, donde, un AUC de 1 significa que el modelo tiene una capacidad aceptable para distinguir entre las clases, mientras que un AUC de 0,5 indica que la capacidad de discriminación del modelo es semejante a adivinar al azar, es decir, el AUC resume la curva ROC en una única

expresión numérica, proporcionando una forma fácil y efectiva de comparar diferentes modelos. La curva ROC es una representación gráfica de la capacidad de discriminación de un modelo a lo largo de diferentes umbrales, mientras que el AUC es una métrica numérica que encapsula esa capacidad en una sola cifra, seguidamente, la Tabla 6 resume ventajas y desventajas de usar la curva ROC.

Tabla 6. Ventajas y desventajas de la curva ROC.

Ventaja	Desventaja
Muestra cómo fluctúa la sensibilidad y especificidad al evaluar la capacidad de discriminación del modelo en diferentes puntos de corte.	Podría no proporcionar una imagen clara del rendimiento del modelo (casos de extremo desbalance).
No está limitada a un umbral específico y proporciona una visión más general del rendimiento del modelo en la clasificación.	No considera las consecuencias o costos asociados con los falsos positivos y falsos negativos, lo que puede ser considerable en ciertos contextos.
Permite comparar diferentes modelos según su rendimiento general, sin considerar el umbral de decisión.	La curva ROC no proporciona un valor específico para la evaluación del modelo, requiere análisis adicional para estimar el rendimiento general.

4.7.6. Tiempo de entrenamiento en modelos de ML

El tiempo de entrenamiento es la cantidad de tiempo que necesita un algoritmo de ML para familiarizarse y adaptarse a un conjunto de datos; en este caso, el conjunto de datos es SCADA, por lo que el tiempo puede afectar la eficiencia del proceso de modelado.

Hay varios factores clave que afectan al tiempo de entrenamiento de los algoritmos de ML, que son los siguientes:

- **Tamaño del conjunto de datos:** Generalmente, un conjunto de datos más grande requiere mayor costo computacional y, por ende, más tiempo de entrenamiento.
- **Complejidad del modelo:** Algunos algoritmos de ML, como las SVM o las DNN, pueden requerir un tiempo de entrenamiento considerablemente mayor en comparación con otros algoritmos.
- **Recursos de hardware:** La disponibilidad y la calidad de CPU, GPU y RAM pueden influir significativamente en el tiempo de entrenamiento.

- **Optimización y técnicas de entrenamiento:** El uso de técnicas de optimización, como la selección de características y la normalización, puede mejorar la eficiencia del entrenamiento.
- **Escala e implementación:** En un entorno de producción, el tiempo de entrenamiento también puede verse afectado por factores como la distribución y paralelización del entrenamiento, la administración de recursos y otros desafíos operativos.
- **Google colab como herramienta:** Google Colab, un entorno de desarrollo gratuito basado en Jupyter Notebook, es ampliamente utilizado gracias a su acceso gratuito a GPUs y TPUs, además en su versión de pago permite el uso de GPUs más potentes.

4.7.7. Test de Wilcoxon

El Test de Wilcoxon, una técnica estadística no paramétrica, se utiliza para comparar dos conjuntos de datos emparejados, como en la evaluación del rendimiento de dos algoritmos. Ofrece dos variantes: la prueba de rango con signo y la prueba de rango de suma de Wilcoxon, enfocadas en analizar hipótesis sobre medianas y distribuciones sin asumir normalidad estadística. Este test es fundamental en estudios como el de [66], que examina modelos de clasificación y emplea puntuaciones de ranking para evaluar de manera rápida y aproximada la significancia de las diferencias observadas.

EI valor P en el Test de Wilcoxon mide la probabilidad de encontrar una diferencia bajo la hipótesis nula. Un valor P bajo (generalmente <0,05) sugiere que la diferencia no es aleatoria, 10 que puede llevar a rechazar la hipótesis nula. Sin embargo, es crucial valorar el contexto y la relevancia práctica de estas diferencias, no solo el valor P, para una interpretación correcta de los resultados.

- **Prueba de rango con signo de Wilcoxon:** Se aplica cuando se tienen dos conjuntos de datos relacionados, como mediciones repetidas en el mismo grupo. La prueba compara las medianas de las diferencias entre los pares para determinar si son cero.
- **Prueba de rango de suma de Wilcoxon (también conocida como prueba de Mann-Whitney U):** Se utiliza para comparar dos grupos independientes y determinar si proceden de la misma población o de poblaciones con igual

mediana. Es una alternativa al t-test cuando no se cumplen las suposiciones de normalidad.

Al plantear las hipótesis para el Test de Wilcoxon, se considera lo siguiente:

- Para ambas versiones, la hipótesis nula asume que no hay diferencia significativa entre las medianas de las diferencias (prueba de rango con signo) o entre las medianas de las poblaciones (prueba de rango de suma de Wilcoxon).
- La hipótesis alternativa indica que hay una diferencia significativa entre las medianas, ya sea en el contexto de diferencias emparejadas o en grupos independientes.

La Tabla 7 detalla las ventajas y desventajas de usar el test de Wilcoxon.

Tabla 7. Ventajas y desventajas del test de Wilcoxon.

Ventaja	Desventaja
No requiere normalidad.	Menos poder estadístico que algunas pruebas paramétricas cuando los datos cumplen con los supuestos.
Resistente a datos atípicos.	
Útil para muestras pequeñas.	No puede manejar datos categóricos o nominales directamente.
Sensible a los cambios en la mediana, no solo a los cambios en la media.	Requiere que las observaciones sean independientes y pareadas.

4.7.8. Curva de calibración

Una curva de calibración ayuda a determinar la confiabilidad de las predicciones de un modelo, especialmente en modelos de clasificación. Según se explica en [67], esta curva estima el valor de probabilidad asociado a una predicción, aunque puede ser difícil determinar su fiabilidad, la calibración ajusta las probabilidades predichas para que reflejen mejor la proporción de datos reales.

De esta manera, se puede afirmar que un modelo está apropiadamente calibrado para cualquier valor de p, donde la clasificación predicha con una confianza (probabilidad) de p es correcta el $100 * p$ por ciento del tiempo.

Ecuación 9. Curva de calibración

$$P\left(\hat{Y} = \hat{Y} \middle| \hat{P} = p\right) = p \quad (9)$$

Donde:

- $p \in [0,1]$

Por ejemplo, al seleccionar observaciones con una probabilidad de predicción del 70%, se espera que el porcentaje de estas observaciones

correctamente clasificadas sea del 70%, este principio se generaliza a todo el rango de probabilidades [0, 1], lo que resulta en una diagonal perfecta.

La calibración de un modelo se refleja en la proximidad entre los valores de proporción empírica y de confianza, es decir, cuanto más se aproxime la curva obtenida a la diagonal, así, la curva de calibración se sitúa por encima de la diagonal si el modelo tiende a infravalorar las probabilidades y por debajo si las sobrevalora, esto se puede observar en la Figura 13.

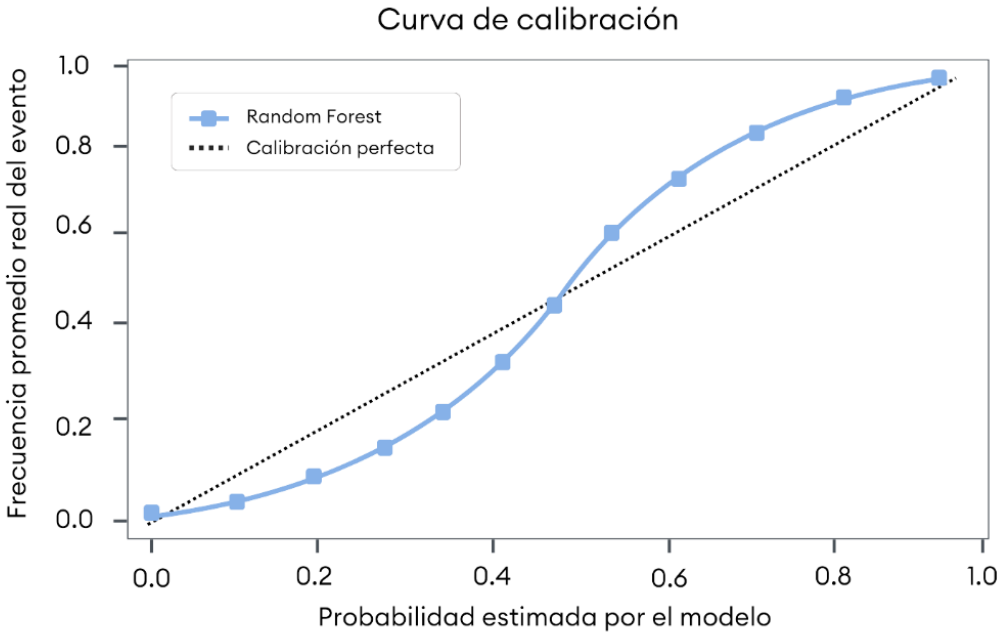

Figura 13. Curva de calibración. Adaptado de [67].

La gráfica de la Figura 13, muestra la curva de calibración de un modelo de bosque aleatorio (random forest). El patrón de la curva indica que, para valores bajos, el modelo tiende a sobrevalorar sus probabilidades y, para valores altos, a infravalorarlas. Por otra parte, su uso se puede encontrar en artículos como [68].

4.7.9. Curva precisión-recuperación

La curva precisión-recuperación o PR, es una medida útil para evaluar la efectividad de la predicción en situaciones de desequilibrio de clases, es decir, cuando una clase es mucho más predominante que otra.

La precisión mide la proporción de ejemplos clasificados como positivos que realmente lo son, mientras que la recuperación evalúa la proporción de ejemplos positivos que el modelo identifica correctamente. En la Figura 14, se muestra la curva de PR, la cual ofrece una representación visual de la relación entre la precisión y la recuperación del modelo.

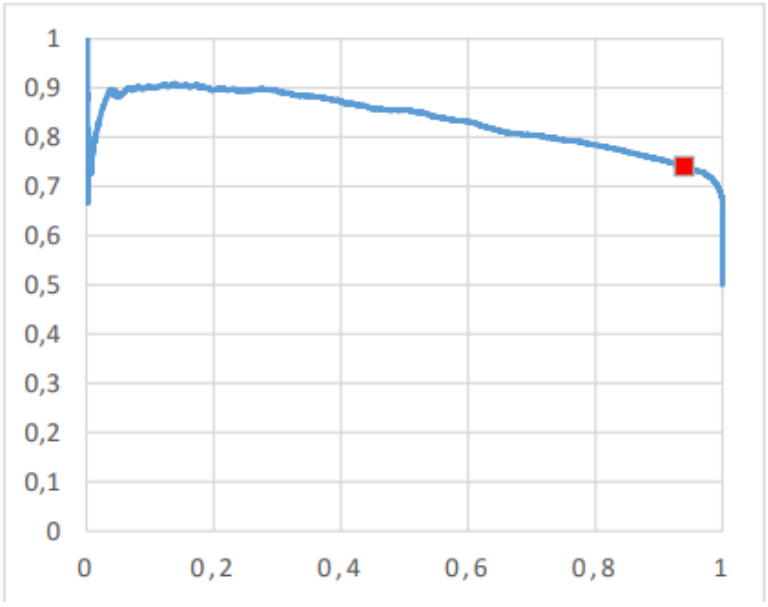

Figura 14. Curva precisión-recuperación [69].

La curva PR grafica la precisión en el eje vertical *y* la recuperación en el horizontal, siendo el punto óptimo, resaltado por la puntuación F1 (representado como un cuadro pequeño de color rojo), la cual fusiona ambos aspectos de manera equilibrada, donde cada punto en la curva corresponde a un umbral de clasificación diferente a medida que aumenta el umbral de clasificación, la precisión tiende a incrementar mientras que la recuperación disminuye, y viceversa lo cual permite observar cómo varían la precisión y la recuperación según el umbral de clasificación.

Por otra parte, la curva PR difiere de la curva ROC en varios aspectos. Por ejemplo, según se expone en [70], mientras que las curvas ROC representan la tasa de verdaderos positivos frente a la tasa de falsos positivos, las curvas PR muestran la precisión frente a la recuperación. En situaciones de eventos raros con baja prevalencia, una curva ROC puede sobreestimar el rendimiento de un modelo debido a la predominancia de la clase negativa, mientras que la curva PR puede ofrecer una evaluación más precisa en tales casos. El uso de la curva PR se puede encontrar en artículos como [71], [72].

4.8. Lenguajes de programación para la implementación de IA

Hoy en día, la IA está cobrando cada vez más importancia y está creciendo rápidamente, sobre todo para hacer frente a todos los retos tecnológicos, sociales y económicos del mundo, esto se debe a que muchos lenguajes de programación ofrecen métodos y bibliotecas específicamente adaptados a este ámbito.

A continuación, se enumeran algunos de los lenguajes de programación más destacados utilizados en la implementación de aplicaciones de IA:

- **Python:** La documentación oficial de este lenguaje de programación se encuentra en su **sitio web principal**, que ofrece recursos esenciales para desarrolladores y especialistas en IA, como demuestran artículos como [73], [74], es bien conocido por su popularidad y amplia aplicación en el campo de la inteligencia artificial, su sintaxis clara y legible simplifica enormemente el proceso de desarrollo e implementación de algoritmos de IA.
- **R:** Centrado en el análisis estadístico y la manipulación de datos, R se ha convertido en un lenguaje de programación muy popular, esto se debe en parte a la abundancia de paquetes y bibliotecas disponibles para la visualización de datos, ML y la estadística, además, su **sitio web principal** proporciona documentación completa para desarrolladores y usuarios, y se utiliza en diversos contextos, como se describe en [75], [76].
- **Java:** Java se emplea ampliamente en el entorno empresarial y también se utiliza en el desarrollo de aplicaciones de IA, ofrece acceso a varias bibliotecas y frameworks, como Deeplearning4j, que facilitan la creación de aplicaciones fiables de aprendizaje automático y procesamiento de grandes volúmenes de datos, la documentación oficial de Java está disponible en su **sitio web principal** y sirve como manual básico para implementar soluciones de IA, en [77], [78], se observan algunos ejemplos.
- **C++:** Este lenguaje de programación de propósito general conocido por sus cualidades de control y eficiencia de bajo nivel que lo hacen idóneo en el desarrollo de aplicaciones de IA que requieren un adecuado rendimiento, cuya documentación ofrece una base de conocimientos esenciales para la implementación eficaz de algoritmos de IA que puede encontrarse en la biblioteca de **documentación de Microsoft** como un recurso fundamental evidenciado en [79], [80].
- **Matlab:** Es un entorno de desarrollo y programación diseñado específicamente para aplicaciones técnicas y científicas especialmente en el campo de la IA donde su uso se justifica por la extensa variedad de funciones y técnicas disponibles para el análisis de datos, implementación de algoritmos de aprendizaje automático y simulación de sistemas complejos, por otra parte, su documentación oficial es una

fuente invaluable de información para el servicio de la comunidad científica y técnica involucrada en la IA y se encuentra disponible en su **sitio web principal**, tal como se puede observar en [76], [81], [82].

4.9. Trabajos relacionados

La concentración de múltiples investigaciones en la detección de fallas y el monitoreo de aerogeneradores evidencia una fuerte inclinación hacia el uso de técnicas de IA y el análisis de datos provenientes del sistema SCADA donde dichos estudios destacan la importancia de la detección temprana de anomalías para prevenir daños mayores y mejorar la eficiencia operativa en centrales eólicas, además, presentan un objetivo común al utilizar la IA como herramienta principal para anticipar y prevenir fallas en aerogeneradores y mejorar su capacidad predictiva al realizar un monitoreo continuo en busca de una operación más eficiente.

4.9.1. Using SCADA data for wind turbine condition monitoring - A review

El documento proporciona una evaluación exhaustiva de las aplicaciones de la IA para la toma de decisiones basada en datos en la industria eólica bajo un enfoque principal en el monitoreo basado en condición (CBM) y de igual manera incluye el análisis del rendimiento siempre que sea relevante para O&M [83], además, implementa métricas bibliométricas para analizar la literatura científica incluyendo la frecuencia de palabras clave relevantes y la evolución temática de las publicaciones. Los algoritmos que conforman el modelo incluyen técnicas de ML como redes neuronales, transformaciones lineales no supervisadas, y modelos de aprendizaje profundo como redes neuronales feedforward y redes neuronales de memoria a corto plazo, por otra parte, el documento utiliza una técnica de combinación de modelos de ensamble híbrido para predecir la producción de energía eólica [83]. Los tipos de datos utilizados incluyen datos SCADA y parámetros meteorológicos para el pronóstico de la producción de energía eólica, cuya relación radica en el enfoque hacia la optimización del mantenimiento y la operación de parques eólicos mediante la utilización de datos del sistema SCADA 44 el desarrollo de modelos predictivos para predecir y prevenir fallas en los aerogeneradores.

4.9.2. Scientometric review of artificial intelligence for operations maintenance of wind turbines: The past, present and future

En este artículo se listan varias maneras de como a lo largo de la historia se han venido controlando las fallas en los aerogeneradores, conforme la tecnología

ha avanzado es notorio que en una gran parte de las centrales eólicas han optado por crear algoritmos predictivos para crear un sistema de respuestas ante cualquier incidente [84], de este modo se demuestra que los parques que sacarán un mayor provecho de la energía eólica serán los que implemente prevención antes que corrección, ya que, los rubros por reparaciones son más elevados que por mantenimiento preventivo [84], con ello, dicho estudio comparte el enfoque hacia la prevención de fallas a través de modelos predictivos basados en IA y datos del sistema SCADA, a través de los cuales se puede llevar a una operación más eficiente y rentable de los aerogeneradores en las centrales eólicas.

4.9.3. A methodology for performance assessment at system level—Identification of operating regimes and anomaly detection in wind turbines

Se presenta una propuesta de una metodología para la detección de anomalías en aerogeneradores basada en el análisis de datos históricos y evaluación del rendimiento, donde su enfoque se centra en identificar períodos críticos de bajo rendimiento y es aplicable tanto a aerogeneradores terrestres como marinos, por otra parte, se aborda la revisión de modelos de mantenimiento basados en condiciones para la energía renovable marina, con énfasis en aerogeneradores, y se discuten modelos predictivos, técnicas de diagnóstico y pronóstico de fallas, y enfoques basados en datos para el monitoreo del estado de los aerogeneradores y la detección de fallas [85] además, se propone una metodología para evaluar el rendimiento de los aerogeneradores a nivel del sistema utilizando datos reales del sistema SCADA, que implica el uso de PCA y K-means para identificar regímenes operativos y evaluar el rendimiento. Se alinean en su interés por identificar anomalías, evaluar el rendimiento de los aerogeneradores y proponer estrategias para mejorar su funcionamiento, utilizando datos del sistema SCADA y técnicas de análisis avanzadas [85].

4.9.4. An Anomaly Detection Approach Based on Machine Learning and SCADA Data for Condition Monitoring of Wind Turbines

Se presenta un enfoque de detección de anomalías utilizando aprendizaje automático y datos SCADA para monitorear la condición de los aerogeneradores, su enfoque utiliza modelos NARX para estimar señales de temperatura, la distancia de Mahalanobis como indicadores evaluados, y la transformada wavelet para eliminar el ruido de las señales [86], así comparten la finalidad de utilizar datos del

sistema SCADA y técnicas avanzadas de aprendizaje automático para detectar anomalías en aerogeneradores, con el objetivo de prevenir problemas mayores y mejorar la eficiencia en el monitoreo y mantenimiento de parques eólicos.

4.9.5. Anomaly detection and condition monitoring of wind turbine gearbox based on LSTM-FS and transfer learning

El documento propone un método novedoso para predecir el estado operativo de los engranajes de aerogeneradores utilizando una combinación de redes neuronales de tipo LSTM (long short-term memory), síntesis difusa (FS) y transfer learning basado en características [87], cuyo método tiene como objetivo detectar posibles fallas en otros engranajes de aerogeneradores haciendo uso de un número limitado de muestras de datos defectuosos. La viabilidad y precisión del método se verifican utilizando datos reales de monitoreo de aerogeneradores, es por ello que se propone un método para el mantenimiento basado en el estado operativo de los aerogeneradores utilizando calibración de estado multidimensional basada en LSTM, selección de características, análisis estadístico y transfer learning [87], su método se aplica a datos de monitoreo de dos aerogeneradores con información defectuosa, y los resultados se comparan con registros de fallas, así ambos buscan mejorar la eficacia en el monitoreo y mantenimiento de los sistemas eólicos mediante el uso de técnicas de IA y validación con datos reales.

4.9.6. Anomaly detection for wind turbine damaged due to lightning strike

El documento de investigación trata sobre el uso de un modelo de aprendizaje automático y datos del sistema de supervisión y adquisición de datos (SCADA) para detectar anomalías en las palas de turbinas eólicas causadas por impactos de rayos, su objetivo del estudio es aumentar la disponibilidad de las turbinas eólicas al permitir una detección más rápida y reinicio después de los impactos de rayos [88]. El estudio presenta un método para la detección de anomalías en los datos de turbinas eólicas utilizando aprendizaje automático, cuya metodología implica el preprocesamiento de los datos mediante normalización, conversión de ventana deslizante y conversión de promedio móvil, su rendimiento se evalúa utilizando matrices de confusión y curvas ROC, por otra parte, destaca la importancia de incluir características temporales en el modelo para una mejor detección de anomalías, así, ambos se centran en tratar de mejorar la eficiencia de detección de problemas en los aerogeneradores utilizando datos del sistema SCADA y técnicas avanzadas de aprendizaje automático [88]. Aunque el artículo

se enfoca en la detección de anomalías específicas debido a impactos de rayos, y el presente trabajo busca predecir fallas en general, ambos son ejemplos de cómo la IA y el análisis de datos pueden mejorar la operatividad y la eficiencia de las centrales eólicas.

4.9.7. Condition monitoring and anomaly detection of wind turbine based on cascaded and bidirectional deep learning networks

Presenta un modelo de detección de fallas en turbinas eólicas basado en una red neuronal convolucional y unidades recurrentes bidireccionales con mecanismo de atención, así, el modelo se compara con otros y se muestra superior en la detección temprana de fallas en turbinas eólicas, no obstante, propone un método para el monitoreo de condiciones y detección de anomalías en turbinas eólicas utilizando redes de aprendizaje profundo en cascada y bidireccionales [89], por otra parte, se propone un modelo CNN-BiGRU-AM para la detección de anomalías en turbinas eólicas utilizando datos SCADA, el cual se entrena en datos de operación normal y se utiliza para predecir las temperaturas del rodamiento y del aceite, en ambos enfoques se reconocen la importancia de identificar problemas de aerogeneradores en etapas tempranas, aunque el artículo señala la dificultad de hacerlo basándose únicamente en ciertos tipos de tendencias, así mismo se utilizan datos SCADA para entrenar el modelo.

4.9.8. Evaluation of anomaly detection of an autoencoder based on maintenance information and SCADA-data

Se presenta el desarrollo y evaluación de un modelo de Autoencoder (AE) para la detección de anomalías y la predicción de fallas en sistemas de turbinas eólicas, donde se discute el uso de datos SCADA y reportes de servicio para monitorear la salud de las turbinas eólicas, detectar fallas con anticipación y reducir los costos de mantenimiento [90], , también se aborda la preparación de los datos de entrada para el entrenamiento del autoencoder, incluyendo la imputación de datos faltantes, la escala de datos y el filtrado de modos operativos

4.9.9. Fault detection and anti-icing technologies in wind energy conversion systems A review

Se discute la implementación de deshielo ultrasónico y sistemas de monitoreo activo integrados para las palas de aerogeneradores que también explora el uso de IA para la detección y diagnóstico de fallas en sistemas de energía eólica

[91], [92], por otra parte, el artículo también discute el uso de diferentes materiales y tecnologías para el deshielo y la detección de fallas en sistemas de energía eólica.

4.9.10. Fleet-based early fault detection of wind turbine gearboxes using physics-informed deep learning based on cyclic spectral coherence

Se presenta una metodología para la detección automatizada de fallas en sistemas mecánicos utilizando aprendizaje profundo y procesamiento de señales [92], así, se demuestra la efectividad del método en la detección de fallas y en la provisión de información útil para el mantenimiento, de esta manera, se compara con el análisis de envolvente tradicional mostrando la superioridad de la metodología propuesta.

4.9.11. Real-time condition monitoring and fault detection of components based on machine-learning reconstruction model

Se propone un método de monitoreo de condiciones y detección de fallas en tiempo real para turbinas eólicas que utiliza un modelo de reconstrucción basado en aprendizaje automático [93], dicho método se enfoca en anomalías de alta temperatura y emplea técnicas de procesamiento de datos como limpieza de datos y selección de características.

4.9.12. Reduced neural network based ensemble approach for fault detection and diagnosis of wind energy converter systems

El estudio propone un método de diagnóstico y detección de fallas en sistemas de conversión de energía eólica basado en redes neuronales, la eficacia del método se demuestra en términos de precisión, memoria y tiempo computacional, lo que indica el excelente rendimiento y resistencia de los paradigmas de diagnósticos sugeridos, las medidas de evaluación utilizadas en este artículo son comparables.

4.9.13. An Artificial Intelligence Neural Network Predictive Model for Anomaly Detection and Monitoring of Wind Turbines Using SCADA Data

El documento se centra en el desarrollo y evaluación de un modelo predictivo de red neuronal de IA para la detección de anomalías y monitoreo de turbinas eólicas utilizando datos SCADA, dicho estudio compara diferentes modelos de redes neuronales artificiales y demuestra que el Modelo 2 es el mejor predictor de fallas en las turbinas eólicas, alcanzando una precisión del 99.8% [94].

El modelo utiliza datos históricos de supervisión y adquisición de datos (SCADA) recopilados de nueve turbinas eólicas durante un período de diez años, con una frecuencia de muestreo de aproximadamente 0.0016 Hz [94], además, el estudio incluye la combinación de técnicas de aprendizaje automático, como el uso de algoritmos de redes neuronales artificiales y la inclusión de datos de temperatura ambiental, interna y externa, así como valores de desviación estándar de los datos de los sensores para mejorar la precisión del modelo.

4.9.14. Combination of Thermal Modelling and Machine Learning Approaches for Fault Detection in Wind Turbine Gearboxes

Se compara el modelado de redes térmicas y enfoques de aprendizaje automático para la detección de fallas en cajas de engranajes de turbinas eólicas utilizando datos de temperatura, dicho estudio sugiere el potencial de establecer valores de umbral de monitoreo de condiciones para cajas de engranajes de turbinas eólicas operativas [95], [96].

4.9.15. Deep learning for automated drivetrain fault detection

El documento presenta un sistema de aprendizaje profundo basado en datos para la detección de fallas en turbinas eólicas a gran escala utilizando el procesamiento de redes neuronales convolucionales en entradas de señales de vibración complejas para aprender de expertos en diagnóstico humano y proporcionar detección temprana y robusta de fallas en rodamientos del rotor, planetarios y de caja de engranajes helicoidales. El estudio se basa en datos de 251 fallas reales de rodamientos de turbinas eólicas y cuantifica con precisión el rendimiento del modelo de diagnóstico a nivel de flota [96], también indica que utiliza métricas de evaluación como el puntaje de probabilidad logarítmica negativa y el puntaje de AUC para evaluar el rendimiento de los modelos, dichos algoritmos que conforman el modelo incluyen regresión logística, redes neuronales completamente conectadas y redes neuronales convolucionales profundas [96], además, el documento implementa una técnica de combinación de datos que implica el muestreo aleatorio de pares de valores de datos de tiempo para cada turbina y los datos usados en el estudio incluyen datos de vibración de turbinas eólicas y la salida de expertos humanos en el diagnóstico de fallas.

4.9.16. Deep Learning for fault detection in wind turbines

El documento revisa el uso de redes neuronales artificiales y aprendizaje profundo para la detección de fallas en turbinas eólicas categorizando los diferentes

enfoques y destacando el potencial del aprendizaje profundo para mejorar el monitoreo de la salud de las turbinas eólicas [97], por otra parte, los algoritmos utilizados incluyen CNN, SVM y autoencoders basados en redes neuronales, además, se emplean métricas de evaluación como la P, R y PF1 para medir el rendimiento de los modelos y los datos utilizados abarcan desde datos de vibración hasta datos SCADA aplicando técnicas de combinación de datos para mejorar la precisión de los modelos.

4.9.17. Enabling Co-lnnovation for a Successful Digital Transformation in Wind Energy Using a New Digital Ecosystem and a Fault Detection Case Study

El documento trata sobre la evaluación de diferentes algoritmos de detección de fallas en turbinas eólicas utilizando datos de SCADA y mástil meteorológico, se utilizan métricas de evaluación como el coste total de predicción y el coste de predicción promedio para comparar el rendimiento de los algoritmos. Los algoritmos que conforman el modelo incluyen Normal Behaviour Models (NBM), Combined Local Minimum Spanning Tree and Cumulative Sum of Multivariate Time Series Data (LoMST-CUSUM), Combined Ward Hierarchical Clustering and Novelty Detection with Local Outlier Factor (WHC-LOF), Normal Behaviour Model with Lagged Inputs (NBM-LI), Canonical Correlation Analysis (CCA), y Kernel Change-Point Detection (KCPD) [98]. Se utilizan datos de SCADA y se aplican técnicas de combinación de modelos más allá del aprendizaje supervisado, incorporando los últimos resultados de la investigación en detección de anomalías.

4.9.18. Exploring Quantum Machine Learning and Feature Reduction Techniques for Wind Turbine Pitch Fault Detection

Presenta el uso de algoritmos de aprendizaje automático, incluyendo núcleos cuánticos, para el diagnóstico de fallas en sistemas de turbinas eólicas. Utiliza métricas de evaluación como precisión, recuperación, puntuación F1y desviación estándar para evaluar el rendimiento de los modelos de clasificación. Los algoritmos que conforman el modelo incluyen SVM-RBF, PCA y AE como técnicas de reducción de características [99]. Presenta el uso de algoritmos de aprendizaje automático, incluyendo núcleos cuánticos, para el diagnóstico de fallas en sistemas de turbinas eólicas. Utiliza métricas de evaluación como precisión, recuperación, puntuación F1y desviación estándar para evaluar el rendimiento de

los modelos de clasificación. Los algoritmos que conforman el modelo incluyen SVM-RBF, PCA y AE como técnicas de reducción de características.

4.9.19. SCADA-based wind turbine anomaly detection using Gaussian process models for wind turbine condition monitoring purposes

El documento considera la magnitud de los modelos de curvas de potencia en la estimación del rendimiento de aerogeneradores cuya detección de anomalías y la monitorización de condiciones implementa métricas de evaluación como el ajuste de modelos de procesos gaussianos y copulas a datos de curvas de potencia así como la precisión en la estimación de incertidumbres para la detección de posibles fallos en los aerogeneradores [100], dicho modelo se compone de algoritmos de procesos gaussianos y copulas para el análisis de datos de curvas de potencia y utiliza la técnica de combinación de Gaussian Process (GP) para la evaluación de la incertidumbre en la estimación del rendimiento de los aerogeneradores [100], cuyos datos implementados incluyen curvas de potencia de aerogeneradores capturados a partir de datos SCADA y otros elementos de supervisión y control de aerogeneradores.

4.9.20. Tandem Connectionist Anomaly Detection: Use of Faulty Vibration Signals in Feature Representation Learning

El documento propone un sistema de detección de anomalías basado en la conexión tandem DNN/GMM para monitorear máquinas rotativas utilizando señales de vibración y datos defectuosos de máquinas no objetivo para mejorar la detección de anomalías para la máquina objetivo, así, dicho sistema se evalúa utilizando métricas como las curvas ROC y el AUC para estimar su rendimiento en la detección de anomalías. Los algoritmos que conforman el modelo incluyen un extractor de características basado en DNN y un verificador de estado normal basado en GMM [101], cuyo documento también utiliza datos de vibración de componentes de turbinas eólicas para demostrar la efectividad del sistema propuesto. Además, se emplea una técnica de combinación de DNN y GMM para mejorar la detección de anomalías, y se evalúa el desempeño del sistema en condiciones dependientes e independientes de la máquina, así como en diferentes tipos de máquinas [101].

4.9.21. Wind turbine generator controller signals supervised machine learning for shaft misalignment fault detection A doubly fed induction generator practical case study

El artículo aborda diversos modelos y técnicas para el diagnóstico de fallas y monitoreo de condiciones en sistemas mecánicos y eléctricos, incluyendo el uso de imágenes térmicas, el método de medición de deformación, el monitoreo de deformación del marco y algoritmos de IA como redes neuronales y máquinas de vectores de soporte [102]. También cubre la aplicación de estas técnicas en turbinas eólicas y generadores de inducción. Así mismo, proporciona una revisión exhaustiva del estado actual de la investigación en este campo. Compara el rendimiento diagnóstico de diferentes algoritmos de aprendizaje automático, como árbol de decisión, Random Forest, Catboost, XGBoost, SVM y clasificador logístico [102]. Los resultados muestran que las señales del bucle del controlador pueden distinguir con precisión entre condiciones de operación saludables y defectuosas, logrando la mayor precisión con el modelo SVM. El análisis de diferentes características seleccionadas para el entrenamiento indica que ciertas señales contribuyen más al proceso de decisión.

4.9.22. Wind turbines anomaly detection based on power curves and ensemble learning

El documento trata sobre el desarrollo de una herramienta en línea para monitorear la eficiencia de turbinas eólicas basada en datos SCADA originales y técnicas de aprendizaje automático, centrándose en un enfoque de aprendizaje en conjunto. Se utilizan diferentes algoritmos de aprendizaje automático, como vecinos más cercanos ponderados (Wk-NN), árbol potenciado (BT), árbol RUSBoosted (RBT), máquinas de vectores de soporte (SVM), bosque aleatorio (RF) y bosque de rotación (RotF), para la detección de anomalías en el comportamiento de las turbinas eólicas [34]. Se emplean métricas como precisión, sensibilidad, especificidad, puntuación f1, precisión y coeficiente de correlación de Matthews (MCC) para evaluar el rendimiento de los modelos. Además, se utiliza un enfoque de aprendizaje en conjunto para mejorar los resultados de los clasificadores individuales. El modelo se construye a partir de datos de curvas de potencia de un parque eólico en tierra en Brasil, y se busca proporcionar una herramienta para la detección en línea de condiciones anormales en las turbinas eólicas, con el objetivo de mejorar la eficiencia y reducir los costos de operación y mantenimiento [34].

4.9.23. A Data-Mining Approach for Wind Turbine Fault Detection Based on SCADA Data Analysis Using Artificial Neural Networks

El documento aborda la detección de fallas en turbinas eólicas utilizando datos históricos de SCADA y Redes Neuronales Artificiales utilizando métricas de evaluación como P, R, PF1, S, ES, EX y AUC para evaluar la capacidad del modelo en la detección de fallas [103]. Los algoritmos que conforman el modelo incluyen técnicas de clustering, combinación de K-Means y el uso de la distancia de Mahalanobis, también, se emplean datos de SCADA registrados cada 10 minutos, así como registros de intervenciones de mantenimiento [103], además, se hace referencia a la importancia de la selección de variables y la combinación de técnicas de disminución de datos como el Análisis de Componentes Principales (PCA) para la implementación de un sistema de monitoreo de turbinas eólicas.

4.9.24. A Small-Sample Wind Turbine Fault Detection Method With Synthetic Fault Data Using Generative Adversarial Nets

El documento se centra en la utilización de redes generativas adversariales (GANs) para generar datos sintéticos de fallas en turbinas eólicas con el objetivo de mejorar la detección de fallas en estos equipos, por otra parte, para evaluar la efectividad del método propuesto se incorporan métricas de evaluación que abordan la generación de datos desde tres aspectos importantes los cuales son detalle, estadísticas y modificación de GANs [104], así mismo, el modelo de detección de fallasse compone de algoritmos de IA como ANN, SVM y DT y emplea una técnica de combinación de datos que incluye la generación de datos sintéticos de fallas y la extensión de datos reales con ruido aleatorio que verifican la efectividad de la detección de fallas cuya información es utilizada en el estudio y son datos de supervisión, control y adquisición de datos (SCADA) recopilados de un parque eólico en el norte de China.

4.9.25. Combining SCADA and vibration data into a single anomaly detection model to predict wind turbine component failure

Se presenta un listado de referencias relacionadas con el monitoreo de condiciones y detección de fallas en turbinas eólicas que incluyen técnicas de aprendizaje automático y enfoques basados en datos, así mismo, se explora la combinación de datos SCADA y vibración para predecir fallas en componentes de turbinas eólicas que presentan dos estudios de caso que detallan la metodología empleada [105], de igual manera, se describe una metodología para detectar anomalías en componentes de turbinas eólicas mediante un Modelo de Comportamiento Normal (NBM) y un clasificador de SVM que presentan dos casos

de estudio que abordan el monitoreo de condiciones y la detección de fallas en turbinas eólicas, en el primer caso se utiliza un modelo de red neuronal de alimentación directa de dos capas para predecir la temperatura del aceite de la caja de cambios, mientras que en el segundo caso se aplica un algoritmo de bosque aleatorio para representar el comportamiento normal de la variable objetivo [105], además, se usan métricas de evaluación como el Root Mean Square Error (RMSE) y la desviación estándar para contrastar el rendimiento del modelo de detección de anomalías con enfoques estándar observados en la literatura, finalmente, la técnica de combinación utilizada integra un grupo de modelos de comportamiento normal que emplean diferentes indicadores de diagnóstico de diversas fuentes de datos para detectar la misma falla cuyos tipos de datos utilizados incluyen información de SCADA, vibración y temperatura para predecir fallas en componentes de turbinas eólicas.

4.9.26. Fault Detection Based on a Combined Approach Of FA-CP-ELM with Application to Wind Turbine System

Se presenta un método novedoso para la detección de fallas en sistemas de turbinas eólicas que utiliza un enfoque combinado de modelado matemático, modelos de predicción y algoritmos de optimización que se prueba aplicando datos de simulación y muestra resultados prometedores para la detección de fallas usando métricas como el error de entrenamiento, error de predicción y capacidad de aproximación de la máquina de aprendizaje extremo (ELM) para evaluar el modelo propuesto [106], cuyos algoritmos que conforman el modelo incluyen el algoritmo de Firefly, algoritmo de optimización de caos y ELM, finalmente, dicho estudio utiliza datos de simulación para verificar la efectividad del método propuesto mediante el reconocimiento de la presencia de falsas alarmas en la detección de fallas y la necesidad de abordar este problema [106].

4.9.27. Fault detection by an ensemble framework of Extreme Gradient Boosting (XGBoost) in the operation of offshore wind turbines

El documento se centra en la detección de fallos en turbinas eólicas offshore utilizando datos SCADA y técnicas de aprendizaje automático como XGBoost y LSTM para desarrollar un modelo predictivo para la detección temprana de fallos, dicho estudio se basa en datos históricos SCADA recopilados de una turbina eólica offshore en Escocia que incluyen parámetros eléctricos, presión, temperatura y mediciones atmosféricas [107], este modelo predictivo se evalúa utilizando ciertas

métricas como la puntuación de información mutua, coeficiente de correlación, métricas de rendimiento como el error residual, la comparación entre las predicciones del modelo y las observaciones reales [107], finalmente se describe el uso de la técnica de selección de parámetros univariados para disminuir el número de parámetros de entrada en el modelo predictivo.

4.9.28. Imbalanced Classification Based on Minority Clustering SMOTE with Wind Turbine Fault Detection Application

El documento detalla el uso del algoritmo MC-SMOTE para abordar el problema de grupos de datos desequilibrados en la detección de fallas especialmente en la formación de hielo en las palas de turbinas eólicas donde se discuten los resultados experimentales que muestran que MC-SMOTE supera a SMOTE en términos de precisión, recuperación, F-measure y G-mean, de igual manera, se menciona la influencia de problemas de incertidumbre en la clasificación desequilibrada y la detección de fallas sugiriéndolo como un posible tema de investigación a futuro, además, se proporciona una pseudo-código para el algoritmo donde se discuten los criterios de evaluación para conjuntos de datos desequilibrados y se presentan propiedades de conjuntos de datos de referencia [108], por otra parte, el documento compara el rendimiento de MC-SMOTE con otros algoritmos de sobremuestreo utilizando conjuntos de datos de referencia y evalúa su efectividad en la detección de fallas por formación de hielo en las palas de turbinas eólicas cuyos resultados muestran que MC-SMOTE supera a otros algoritmos en la mayoría de los casos siendo adecuado para procesar datos distribuidos de manera desigual dentro del espacio de características.

4.9.29. Predictive maintenance of abnormal wind turbine events by using machine learning based on condition monitoring for anomaly detection

Se presenta el uso de modelos de aprendizaje automático para el monitoreo de condiciones de turbinas eólicas utilizando datos SCADA. Describe el algoritmo KNN, el modelo RF y la Red Neuronal Profunda, y su aplicación al conjunto de datos obtenido de turbinas eólicas [109], El conjunto de datos incluye características relacionadas con el viento, la potencia, la temperatura y otros parámetros. Se discuten el preprocesamiento de datos, la limpieza, la normalización, el preprocesamiento de ventanas de tiempo, la ampliación de datos, la selección de características y la extracción de características [109], El documento

también presenta los resultados experimentales y la evaluación del rendimiento de los modelos de aprendizaje automático.

4.9.30. Research on Fault Detection for Three Types of Wind Turbine Subsystems Using Machine Learning

Se discute el uso de modelos de SVM y Regresión de Vectores de Soporte (SVR) para la detección de fallas en sistemas de turbinas eólicas. Así mismo, abarca el proceso de clasificación de imágenes utilizando SVM y la construcción de matrices de confusión para la evaluación del rendimiento. Además, proporciona información sobre la extracción de características de imágenes y el preprocesamiento para la detección de fallas utilizando modelos SVM y SVR [110]. Se discute varios enfoques de diagnóstico de fallas para el sistema de paso de las turbinas eólicas, incluyendo el uso de modelos de comportamiento normal, algoritmos basados en observadores y técnicas de aprendizaje automático como máquinas de vectores de soporte y redes convolucionales profundas. Utiliza métricas de evaluación como precisión, recuperación, especificidad, tasa de detección negativa y tasa de falsas alarmas para comparar la precisión de detección entre los modelos de redes neuronales convolucionales, SVM y SVR para fallas en el generador, convertidor y sistema de paso de las turbinas eólicas [110]. Utiliza datos de radar charts generados correspondientes a diferentes estados de operación del generador, convertidor y sistema de paso de las turbinas eólicas para la detección de fallas.

4.9.31. Wind Turbine Fault Detection Using Highly Imbalanced Real SCADA Data

Se realiza la detección de fallas en turbinas eólicas utilizando datos SCADA y técnicas de aprendizaje automático discutiendo las fases de un sistema eléctrico trifásico y los informes de alarmas generados, por otra parte, se describe una metodología de detección de fallas, análisis de datos, preprocesamiento y etiquetado, además, se explican la modelización de datos y la reducción de dimensiones utilizando Análisis de Componentes Principales (PCA), así como técnicas para abordar datos desequilibrados como el sobremuestreo aleatorio y la ampliación de datos [111], cuyas técnicas se utilizan para abordar el desequilibrio en el conjunto de datos y mejorar el rendimiento de los algoritmos de clasificación.

Se realiza la detección de fallas en turbinas eólicas utilizando datos SCADA y técnicas de aprendizaje automático discutiendo las fases de un sistema eléctrico

trifásico y los informes de alarmas generados, por otra parte, se describe una metodología de detección de fallas, análisis de datos, preprocesamiento y etiquetado, además, se explican la modelización de datos y la reducción de dimensiones utilizando Análisis de Componentes Principales (PCA), así como técnicas para abordar datos desequilibrados como el sobremuestreo aleatorio y la ampliación de datos [111], cuyas técnicas se utilizan para abordar el desequilibrio en el conjunto de datos y mejorar el rendimiento de los algoritmos de clasificación. Las métricas de evaluación utilizadas incluyen precisión (acc), valor predictivo positivo (ppv), R, PF1 y tasa de falsos positivos (fpr), por otra parte, los algoritmos que conforman el modelo son KNN, SVM y RUSBoost, de igual modo, implementa datos de monitoreo del sistema SCADA de una turbina eólica, así como registros de alarmas generadas durante el mismo período [111].

5. Metodología

Para desarrollar el presente TIC, se empleó recursos y procedimientos para realizar un proceso estructurado. Así mismo, se buscó obtener conocimientos esenciales para proponer una solución al problema de estudio; en consecuencia, la sección **5.1** describe el área de estudio donde se realizó el proyecto y la sección **5.2** expone el procedimiento seguido para consumar cada objetivo planteado.

5.1. Área de estudio

El presente TIC se desarrolló en la ciudad de Loja, Ecuador, en la Universidad Nacional de Loja, específicamente en la carrera de Ingeniería en Computación, perteneciente a la facultad de la Energía, las Industrias y los Recursos Naturales No Renovables y se realizó durante el periodo académico de abril 2023 a septiembre 2023.

El enfoque primordial del presente TIC consistió en determinar un modelo híbrido de IA para predecir fallas de los aerogeneradores de la CEV, para esto se ocupan los datos SCADA del año 2020, en síntesis, se tomaron en cuenta dos etapas clave, primero se realizó una RSL con el objetivo de profundizar en la predicción de fallas en aerogeneradores, segundo se realizaron experimentos con diferentes algoritmos, técnicas de sobremuestreo y técnicas de combinación para identificar el modelo híbrido con mejor desempeño.

5.1.1. Materiales

Los diferentes recursos ocupados para el desarrollo del presente TIC se detallan a continuación:

5.1.1.1. Recurso humano

La Tabla 8 detalla el recurso humano involucrado y sus respectivas responsabilidades.

Tabla 8. Recurso humano

Recurso	Responsabilidad
Wagner Cristhoper Castillo Castro	Estudiante a cargo de la ejecución y desarrollo del TIC.
Pablo Fernando Ordoñez Ordoñez	Director del TIC, responsable de la orientación, supervisión y apoyo, asegurando la calidad y el progreso adecuado.
Jorge Luis Maldonado Correa	Docente encargado de asesorar y proporcionar los datos del sistema SCADA de la CEV.

5.1.1.2. Recursos científicos

Los recursos científicos empleados durante el proceso del TIC se presentan en la Tabla 9.

Tabla 9. Recursos científicos

Recurso	Descripción
Estudio del estado del arte	Se empleó la técnica de la RSL para buscar y analizar información bibliográfica relevante sobre la predicción de fallas en aerogeneradores utilizando datos SCADA e IA. Los resultados detallados de esta revisión se presentan en el **Anexo 1. Revisión Sistemática de Literatura.**
Metodología de Bárbara Kitchenham y Bacca	Esta metodología permitió realizar la RSL de manera rigurosa, garantizando los resultados basados en el problema de investigación. Además, la metodología propone elementos de pensamiento conceptual para facilitar las tareas del investigador [112].
Bases de datos grises	Las bases de datos grises proporcionaron acceso a toda la información científica o técnica que no ha sido publicada comercialmente, los cuales pueden ser producidos por diversas entidades como organizaciones gubernamentales, instituciones académicas, grupos de investigación, empresas privadas, entre otras.
Adaptación de la metodología CRISP-DM	La adaptación de esta metodología permitió guiar los diferentes algoritmos de ML, incluyendo las técnicas de combinación para obtener los modelos híbridos. La descripción de esta implementación se presenta en: **Flujograma para el desarrollo de los algoritmos.**

5.1.1.3. Recursos de hardware y software

Los recursos de hardware y software facilitaron el desarrollo del TIC, permitiendo el uso de recursos computacionales, científicos y académicos, tanto públicos como privados en las diversas actividades.

La Tabla 10, presenta los recursos de hardware empleados y la Tabla 11 detalla los recursos de software utilizados.

Tabla 10. Recursos de hardware

Recurso	Descripción
Laptop	Dispositivo informático usado para una variedad de tareas.

Tabla 11. Recursos de software

Recurso	Descripción
Scopus	Base de datos bibliográfica para buscar y acceder a artículos científicos, libros y patentes.
Web of Science	Base de datos bibliográfica y herramienta de análisis de citas para investigar y seguir la literatura científica.
Google Scholar	Motor de búsqueda académico para encontrar artículos científicos, libros y documentos relevantes.
Mendeley	Gestor de referencias bibliográficas y plataforma para colaboración académica.
Python	Lenguaje de programación utilizado en análisis de datos, inteligencia artificial.
Visual Estudio Code	Entorno de desarrollo integrado (IDE) para escribir, depurar y editar código de programación.
Lucidchart	Lucidchart es una plataforma en línea para crear diagramas y visualizaciones de datos de forma sencilla y colaborativa.
Canva	Canva es una herramienta en línea que permite crear diseños gráficos de forma sencilla y visualmente atractiva
Zoom	Plataforma de comunicación en tiempo real de audio y video para realizar reuniones virtuales.

Recurso	Descripción
Microsoft Office	Suite de aplicaciones con herramientas como Word y Excel.
OneDrive	Servicio de almacenamiento en la nube que permite la creación, edición y visualización de documentos desde diferentes dispositivos.
Google Colab Pro	Plataforma colaborativa en la nube para proyectos de aprendizaje automático en Python, con notebooks interactivos y acceso a recursos de hardware.
Google Drive	Google Drive es un servicio de almacenamiento en la nube y compartición de archivos.
GitHub	GitHub es una plataforma en línea para alojar y gestionar proyectos de desarrollo de software con Git, permitiendo la colaboración y revisión de código.

5.1.1.4. Insumos

La Tabla 12 presenta los insumos empleados.

Tabla 12. Insumos

Recurso	Descripción
Internet	Utilizado para la investigación, acceso a técnicas, comunicación, etc.

5.1.1.5. Datos SCADA

Los datos SCADA registrados y empleados provienen de los aerogeneradores de la CEV, la cual se encuentra ubicada en la provincia de Loja, ciudad de Loja. El conjunto de datos está conformado por la información recopilada durante el año 2020, que incluyen un total de 71 variables monitoreadas por cada 10 minutos a lo largo de los 366 días del año. La Tabla 13 presenta las variables recopiladas.

Tabla 13. Variables registradas del sistema SCADA de la CEV.

N°	Nombre de variable	Descripción	Unidad
1	Identificador Único	Número de identificación única para cada registro.	(sin unidad)
2	Tiempo Redondeado	Tiempo de registro redondeado o agrupado.	(Tiempo)
3	Velocidad Media del Viento	Velocidad promedio del viento durante un período determinado.	(m/s)
4	Velocidad Máxima del Viento	Velocidad máxima registrada del viento.	(m/s)
5	Velocidad Mínima del Viento	Velocidad mínima registrada del viento.	(m/s)
6	Potencia Activa Media en la Red	Promedio de potencia activa suministrada a la red.	(W)
7	Potencia Activa Máxima en la Red	Máxima potencia activa suministrada a la red.	(W)
8	Potencia Activa Mínima en la Red	Mínima potencia activa suministrada a la red.	(W)
9	Temperatura Máxima de los Condensadores del Generador	Temperatura máxima alcanzada por los condensadores en el generador.	(°C)

N°	Nombre de variable	Descripción	Unidad
10	Potencia Reactiva Media del Convertidor	Promedio de potencia reactiva en el convertidor.	(VAR)
35	Temperatura Máxima del Inductor de CA	Temperatura máxima en el inductor de corriente alterna.	(°C)
11	Potencia Reactiva Máxima del Convertidor	Máxima potencia reactiva en el convertidor.	(VAR)
12	Potencia Reactiva Mínima del Convertidor	Mínima potencia reactiva en el convertidor.	(VAR)
13	Velocidad Media del Generador	Velocidad promedio del generador.	(RPM)
14	Velocidad Máxima del Generador	Velocidad máxima alcanzada por el generador.	(RPM)
15	Velocidad Mínima del Generador	Velocidad mínima alcanzada por el generador.	(RPM)
16	Temperatura Ambiente Media	Temperatura ambiente promedio.	(°C)
17	Temperatura Ambiente Máxima	Temperatura ambiente máxima registrada.	(°C)
18	Temperatura Ambiente Mínima	Temperatura ambiente mínima registrada.	(°C)
19	Ángulo Medio de la Pala	Ángulo promedio de las palas del aerogenerador.	(°)
20	Ángulo Máximo de la Pala	Ángulo máximo alcanzado por las palas del aerogenerador.	(°)
21	Ángulo Mínimo de la Pala	Ángulo mínimo alcanzado por las palas del aerogenerador.	(°)
22	Posición Media de la Nacelle	Posición promedio de la nacelle o góndola del aerogenerador.	(°)
23	Tensión Media de la Fase 1 en la Red	Promedio de tensión en la fase 1 de la red.	(V)
24	Tensión Media de la Fase 2 en la Red	Promedio de tensión en la fase 2 de la red.	(V)
25	Tensión Media de la Fase 3 en la Red	Promedio de tensión en la fase 3 de la red.	(V)
26	Corriente Media de la Fase 1 en la Red	Promedio de corriente en la fase 1 de la red.	(A)
27	Corriente Media de la Fase 2 en la Red	Promedio de corriente en la fase 2 de la red.	(A)
28	Corriente Media de la Fase 3 en la Red	Promedio de corriente en la fase 3 de la red.	(A)
29	Aceleración Máxima de la Nacelle	Aceleración máxima registrada en la nacelle.	(m/s^2)
30	Temperatura Máxima de los Devanados	Temperatura máxima alcanzada en los devanados del generador.	(°C)
31	Temperatura Máxima de la Caja Superior	Temperatura máxima en la caja superior del aerogenerador.	(°C)
32	Temperatura Máxima de la Nacelle	Temperatura máxima registrada en la nacelle.	(°C)
33	Temperatura Máxima del IGBT	Temperatura máxima del transistor bipolar de puerta aislada.	(°C)

N°	Nombre de variable	Descripción	Unidad
34	Temperatura Máxima de los Condensadores de Enlace de CC	Temperatura máxima en los condensadores del enlace de corriente continua.	(°C)
35	Temperatura Máxima del Inductor de CA	Temperatura máxima en el inductor de corriente alterna.	(°C)
36	Temperatura Máxima del Rectificador	Temperatura máxima del rectificador.	(°C)
37	Temperatura Máxima del IGBT del Chopper	Temperatura máxima del IGBT en el chopper o cortador.	(°C)
38	Temperatura Máxima del Inductor de CC	Temperatura máxima en el inductor de corriente continua.	(°C)
39	Temperatura Máxima del IGBT de Aumento de Tensión	Temperatura máxima en el IGBT utilizado para aumentar la tensión.	(°C)
40	Temperatura Máxima del Motor de Pitch 1	Temperatura máxima en el motor de pitch para la pala 1.	(°C)
41	Temperatura Máxima del Motor de Pitch 2	Temperatura máxima en el motor de pitch para la pala 2.	(°C)
42	Temperatura Máxima del Motor de Pitch 3	Temperatura máxima en el motor de pitch para la pala 3.	(°C)
43	Temperatura Máxima del Condensador de Pitch 1	Temperatura máxima en el condensador de pitch para la pala 1.	(°C)
44	Temperatura Máxima del Condensador de Pitch 2	Temperatura máxima en el condensador de pitch para la pala 2.	(°C)
45	Temperatura Máxima del Condensador de Pitch 3	Temperatura máxima en el condensador de pitch para la pala 3.	(°C)
46	Temperatura Máxima de la Caja de Pitch 1	Temperatura máxima en la caja de control de pitch para la pala 1.	(°C)
47	Temperatura Máxima de la Caja de Pitch 2	Temperatura máxima en la caja de control de pitch para la pala 2.	(°C)
48	Temperatura Máxima de la Caja de Pitch 3	Temperatura máxima en la caja de control de pitch para la pala 3.	(°C)
49	Temperatura Máxima del Convertidor de Pitch 1	Temperatura máxima del convertidor de pitch para la pala 1.	(°C)
50	Temperatura Máxima del Convertidor de Pitch 2	Temperatura máxima del convertidor de pitch para la pala 2.	(°C)
51	Temperatura Máxima del Convertidor de Pitch 3	Temperatura máxima del convertidor de pitch para la pala 3.	(°C)
52	Temperatura Máxima de la Fuente de Alimentación de Pitch 1	Temperatura máxima de la fuente de alimentación para el sistema de pitch de la pala 1.	(°C)
53	Temperatura Máxima de la Fuente de Alimentación de Pitch 2	Temperatura máxima de la fuente de alimentación para el sistema de pitch de la pala 2.	(°C)
54	Temperatura Máxima de la Fuente de Alimentación de Pitch 3	Temperatura máxima de la fuente de alimentación para el sistema de pitch de la pala 3.	(°C)

N°	Nombre de variable	Descripción	Unidad
55	Modo de Operación	Estado o modo de operación actual del aerogenerador.	(sin unidad)
56	Energía Producida	Cantidad total de energía producida.	(kWh)
57	Energía Consumida	Cantidad total de energía consumida por el aerogenerador.	(kWh)
58	Tiempo de Encendido	Tiempo total que el aerogenerador ha estado encendido.	(Horas)
59	Tiempo de Operación sin Errores	Tiempo total en el que el sistema de generación de energía eólica ha estado en estado OK.	(Horas)
60	Tiempo de Errores	Tiempo total en el que el sistema ha estado en estado de error.	(Horas)
61	Tiempo con Condiciones Ambientales Favorables	Tiempo total en el que las condiciones ambientales han sido adecuadas para la operación.	(Horas)
62	Tiempo con Condiciones Ambientales Desfavorables	Tiempo total en el que las condiciones ambientales no han sido adecuadas para la operación.	(Horas)
63	Tiempo de Servicio/Mantenimiento	Tiempo total dedicado al servicio o mantenimiento.	(Horas)
64	Tiempo en Parada Controlada de la Red	Tiempo total en el que el aerogenerador ha estado parado debido a controles de la red eléctrica.	(Horas)
65	Tiempo Total de Producción de Energía	Tiempo total dedicado a la producción de energía.	(Horas)
66	Energía Producida Hoy	Cantidad de energía producida en el día actual.	(kWh)
67	Energía Producida Ayer	Cantidad de energía producida en el día anterior.	(kWh)
68	Energía Producida Dos Días Antes	Cantidad de energía producida dos días antes.	(kWh)
69	Energía Producida Tres Días Antes	Cantidad de energía producida tres días antes.	(kWh)
69	Energía Producida Tres Días Antes	Cantidad de energía producida tres días antes.	(kWh)
70	Identificación de la Turbina Eólica	Modelo o tipo de aerogenerador.	(Modelo)
71	Descripción del Fallo	Bit en caso de fallo (1) o comportamiento normal (0).	(bit)

Número (N°).

5.2. Procedimiento

Con el propósito de ejecutar el objetivo general "Desarrollar un modelo híbrido basado en inteligencia artificial para predecir fallas en los aerogeneradores de la Central Eólica Villonaco utilizando datos del sistema SCADA" establecido en el presente TIC, se implementó el siguiente procedimiento:

5.2.1. Objetivo 1: Realizar una revisión sistemática de literatura acerca de modelos predictivos basados en inteligencia artificial para la detección de fallas en aerogeneradores.

Este objetivo se desarrolla a través de la metodología propuesta por Torres [112], la cual se basa en un método adaptado de Barbara Kitchenham y Bacca. Para ello en la fase de planeación se realizó la descripción detallada del problema de investigación, el planteamiento de las preguntas de investigación, la elaboración del mentefacto conceptual y la definición de los criterios de inclusión y exclusión utilizados en la RSL, con el propósito de identificar documentos pertinentes y relevantes para el estudio. Se llevó a cabo un análisis minucioso de los documentos seleccionados, extrayendo la información relevante y necesaria para responder de manera precisa las preguntas de investigación planteadas, con el fin de determinar cuáles son los modelos y técnicas de IA más eficaces y adecuadas para el desarrollo del modelo híbrido destinado a la detección de fallas en aerogeneradores **(ver sección 6.1.1. Resultado 1: Técnicas y modelos de inteligencia artificial más usados aplicados a la predicción de fallas de aerogeneradores)**, determinar cuáles son los componentes más estudiados **(ver sección 6.1.2. Resultado 2: Componentes del aerogenerador más estudiados en la predicción de fallas)** y así mismo, determinar cuáles son los artículos más relevantes **(ver sección 6.1.3. Resultado 3: Artículos más importantes sobre modelos de predicción de fallas en aerogeneradores en los últimos años).** Finalmente, se elaboró el informe final de la RSL **(ver Anexo 1. Revisión Sistemática de Literatura).** En donde, se presentó los hallazgos clave de manera clara y objetiva, incluyendo los resultados más relevantes y su respectivo análisis, acompañados de las conclusiones derivadas de dichos resultados**.**

5.2.2. Objetivo 2: Implementar un modelo híbrido para la predicción de fallas en los aerogeneradores del parque Eólico Villonaco.

Se analizaron las variables más relevantes con respecto al conjunto SCADA de la CEV registrado en el año 2020, considerando la variable objetivo que permitiera identificar aquellas que pudieran detectar fallas en cualquier componente del aerogenerador **(ver sección 6.2.1. Resultado 4: Análisis exploratorio de los datos SCADA).** Por otra parte, se analizó la distribución de clases y se mostró el índice de desbalance a tratar y cómo influyeron las técnicas de sobremuestreo al balancear las clases **(ver sección 6.2.2. Resultado 5: Distribución de clases y división de datos).** En donde, previamente se realizó una distinción entre las clases 0 y 1 para los datos

que pertenecían a un comportamiento normal y cuales no **(ver sección 5.2.2.1. Distinción entre las clases 1 y 0: Identificación de fallas y comportamiento normal).**

A través de la adaptación de la metodología CRISP-DM **(ver sección 5.2.2.2. Flujograma para el desarrollo de los algoritmos),** se buscó combinar las técnicas seleccionadas y reconocidas durante la RSL con el fin de determinar un modelo híbrido apropiado. Para ello se realizó una comparación de algoritmos de ML y técnicas de sobremuestreo **(ver sección 6.2.3. Resultado 6: Algoritmos de ML y técnicas de sobremuestreo),** en donde, se llevó a cabo un análisis de los algoritmos con mejores desempeños mediante cinco experimentos.

Se comparó distintas técnicas de combinación de algoritmos de ML con el objetivo de identificar un modelo híbrido **(ver sección 6.2.4. Resultado 7: Técnicas de Combinación),** para aprovechar las fortalezas de diferentes algoritmos y reducir sus debilidades para construir un modelo más robusto.

Se aplicó el test de Wilcoxon **(ver sección 6.2.5. Resultado 8: Test de Wilcoxon)** con el fin de comprobar si existían o no diferencias estadísticamente significativas entre los algoritmos seleccionados.

Finalmente se muestra un resumen de todos los experimentos y resultados en **(ver sección 6.2.6. Resumen de resultados),** y se expusieron las capacidades del modelo híbrido seleccionado en términos de predicción de fallas y probabilidad de aciertos. **(ver sección 6.2.7. Resultado 9: Modelo híbrido Blending-ANN-RF).**

5.2.2.1. Distinción entre las clases 1 y 0: Identificación de fallas y comportamiento normal.

La distinción entre las clases 1 y 0 se puede entender de una manera sencilla:

- La clase 1 se asigna a todos los registros que indican fallas.
- La clase 0 se asigna a los registros que representan un comportamiento normal, es decir, donde no se han detectado fallas.

En la clase 1 se registran todos los datos que señalan fallas identificados por los sistemas SCADA dentro de los componentes del aerogenerador, la cual también es denominada clase minoritaria. Por otra parte, la clase 0 emplea los datos que reflejan un funcionamiento estándar del aerogenerador, asimismo denominada como clase mayoritaria o datos de comportamiento normal.

Para comprender las clases 1 y 0, se presenta la cantidad de registros de cada aerogenerador de la CEV durante el año 2020. A continuación, se detallan los datos correspondientes en la Figura 15.

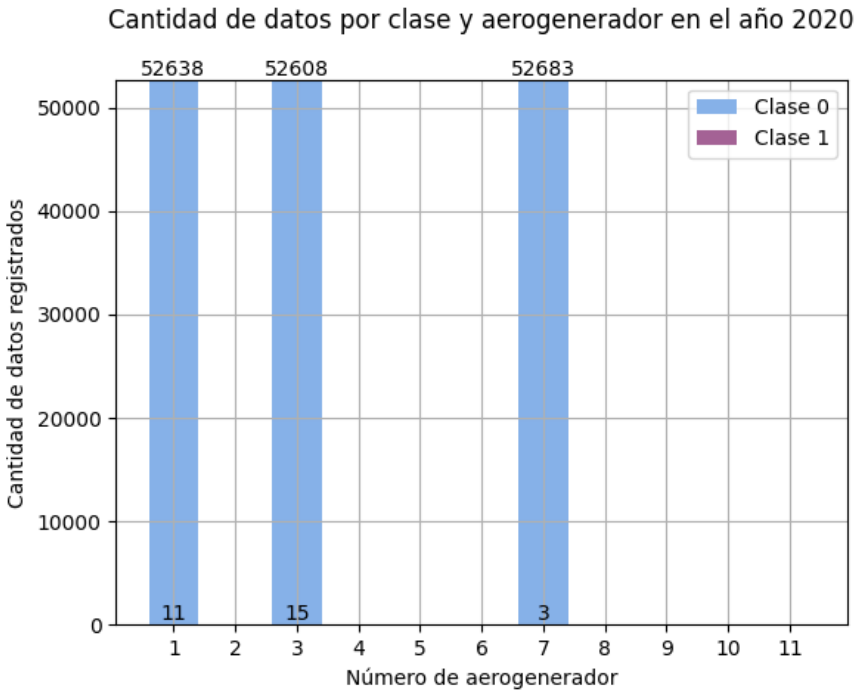

Figura 15. Número de registros por clase y aerogenerador durante el año 2020.

El conjunto de datos incluye registros para los aerogeneradores 1, 3 y 7 de los 11 aerogeneradores presentes. Se observa una cantidad considerable de datos para la clase 0, aproximadamente de 52,600 para cada uno de los tres aerogeneradores, mientras que para la clase 1 solo se cuenta con 11, 15 y 3 registros respectivamente. Esto muestra un desbalance significativo que debe abordarse. Además, se estima un funcionamiento bastante normal y sin complicaciones durante el año 2020.

5.2.2.2. Flujograma para el desarrollo de los algoritmos.

La metodología utilizada para guiar los distintos algoritmos se basa en una adaptación de CRISP-DM, cuya estructura detallada se encuentra en [113]. Esta metodología es reconocida por su eficacia y estructura bien definida. Por lo tanto, la Figura 16 presenta el flujograma que ilustra la metodología adaptada de CRISP-DM, aplicada específicamente para orientar cada algoritmo seleccionado del presente trabajo.

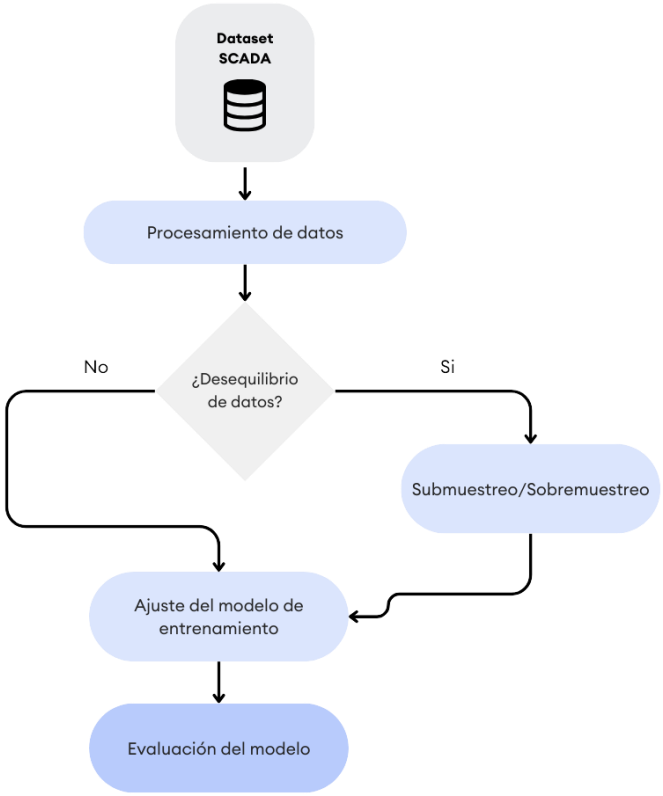

Figura 16. Flujograma metodológico utilizado.

En primer lugar, se realizó la importación de los datos SCADA provenientes de los aerogeneradores de la CEV. Posteriormente, se procedió con el preprocesamiento de los datos, enfocándose en el análisis de la distribución de las clases, diferenciando entre datos que representaban el comportamiento normal y aquellos correspondientes a fallas. Este proceso fue fundamental para el posterior entrenamiento de los algoritmos, dada la notable disparidad entre la gran cantidad de datos de comportamiento normal y la escasez de datos correspondientes a fallas. Este desequilibrio caracterizado como un desequilibrio extremo, condujo a la necesidad de reducir la extensa cantidad de datos relacionados con el comportamiento normal (clasificados como la clase mayoritaria) en comparación con los datos que representaban fallas (clasificados como la clase minoritaria).

Para abordar este problema, se aplicó dos técnicas de reducción de datos, descritas a continuación:

- **Intercambio de datos de fallas entre aerogeneradores:** Este proceso es posible gracias a la homogeneidad tecnológica entre cada aerogenerador en la CEV. Además, los datos utilizados corresponden al mismo año para

todos los aerogeneradores. En este proceso, se llevó a cabo la identificación del aerogenerador que presentaba un mayor número de fallas para emplearlo como receptor de las incidencias de los demás aerogeneradores. Específicamente, el proceso implicó intercambiar los datos de comportamiento normal en la misma fecha, hora y minuto en que se produjo una falla en un aerogenerador distinto. Por consiguiente, el conjunto de datos resultante contiene un único aerogenerador que reporta todas las fallas, mientras que el resto muestra únicamente datos de comportamiento normal. La representación de este proceso se puede visualizar en la Figura 17.

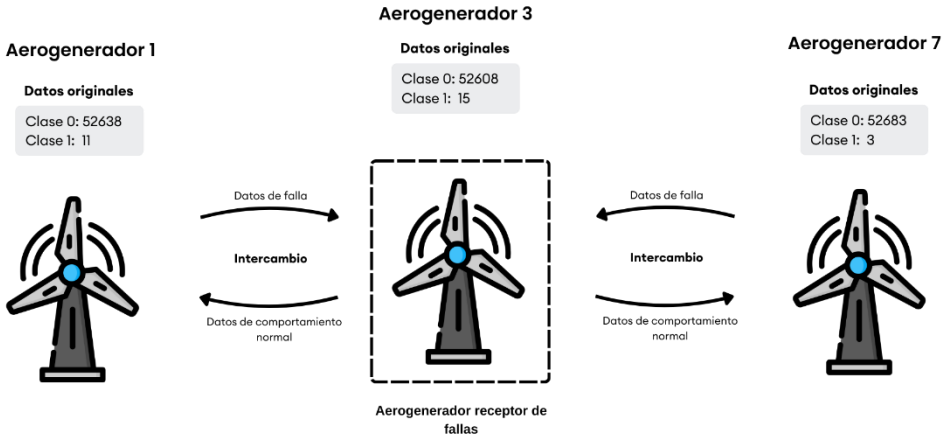

Figura 17. Intercambio de fallas entre aerogeneradores

En la Figura 17, se observa cómo los aerogeneradores 1 y 7 transfieren las fallas al aerogenerador con más incidencias, el aerogenerador 3. De esta manera, solo el aerogenerador 3 conserva datos de la clase 1, mientras que los demás no. Además, para mitigar el desequilibrio de clases, se priorizará el uso exclusivo de los datos del aerogenerador 3.

- **Reducción de la clase mayoritaria por ventana de tiempo:** A pesar de la efectividad del intercambio de datos entre aerogeneradores, persiste un notable desequilibrio. Para abordar esta disparidad, se implementó una segunda técnica que implica la delimitación temporal de cada falla registrada en los datos del aerogenerador 3. De este modo, se aplica una ventana de tiempo alrededor de cada evento de falla durante un periodo asignado, por ejemplo, 3 horas. Durante este lapso, únicamente se conservan los datos que ocurrieron dentro de las 3 horas previas y

posteriores al evento de falla. Este proceso se representa de manera más clara en la Figura 18.

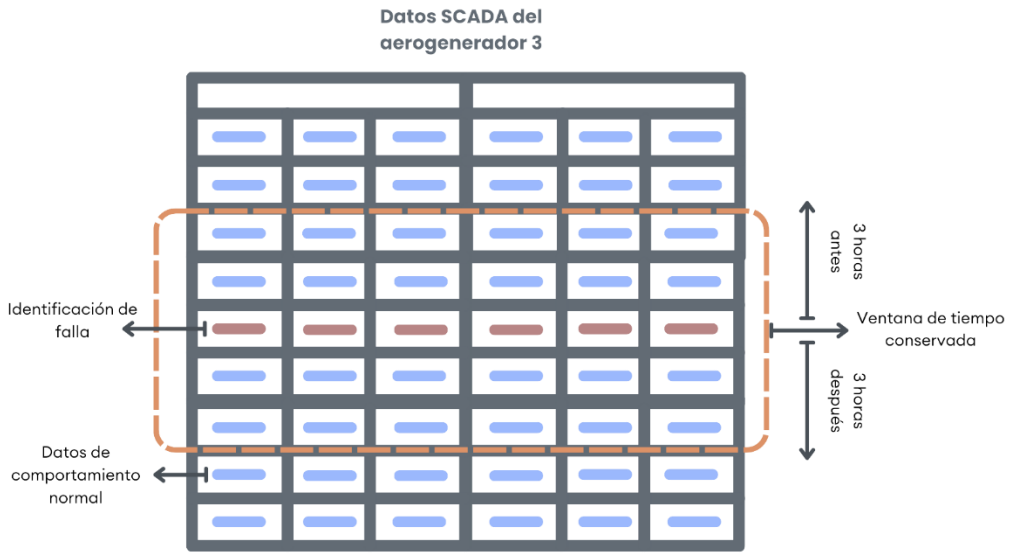

Figura 18. Reducción de la clase mayoritaria por ventana de tiempo

Como se muestra únicamente se emplean los datos del aerogenerador 3, donde primero se localiza la falla y luego se conservan los datos de comportamiento normal durante un período de 3 horas antes y después del evento. Los datos conservados se representan mediante líneas punteadas de color naranja, mientras que los demás datos de comportamiento normal no se utilizan. Naturalmente, este proceso redujo efectivamente el desbalance entre las clases

Luego de estos procesos, se llevaron a cabo ajustes específicos para separar los datos en conjuntos de características (X) y la variable objetivo (y). Después de esto, se normalizan las características para mejorar la convergencia del modelo y reducir la sensibilidad a la escala de las características. Luego, se establece un entrenamiento del 70% y una evaluación del 30%, permitiendo utilizar una parte de los datos para entrenar el modelo y la otra parte para evaluar su rendimiento y generalización en datos no vistos.

Este paso permitió la aplicación de técnicas de sobremuestreo, como SMOTE, con el propósito de balancear las clases existentes en el conjunto de datos. Posteriormente, tras la implementación de la técnica de sobremuestreo, se procedió a afinar el modelo de entrenamiento seleccionado con el objetivo de obtener las métricas más óptimas, considerando las particularidades inherentes

a cada algoritmo utilizado, los cuales poseen distintas naturalezas y características.

Una vez concluido el entrenamiento del modelo, se llevó a cabo su evaluación empleando diversas técnicas, que incluyeron la utilización de la matriz de confusión, el informe de clasificación, métricas de evaluación específicas y demás. Estas estrategias de evaluación proporcionaron una comprensión detallada del desempeño y la capacidad predictiva del modelo entrenado.

5.2.3. Enfoque metodológico empleado

Para ofrecer una visión concisa del enfoque seguido en este trabajo, se detalla en la Figura 19 una representación secuencial del mismo.

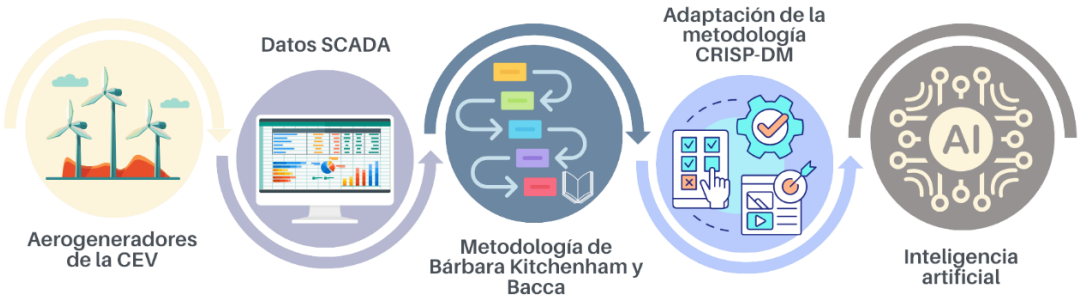

Figura 19. Enfoque metodológico ocupado

Desde una perspectiva general, en primer lugar, todo parte con los aerogeneradores de la CEV, donde se usan los datos SCADA del año 2020 como punto de partida. Luego, se procede a analizar una amplia gama de autores que han investigado la predicción de fallas en aerogeneradores, para lo cual se emplea la metodología de Barbara Kitchenham y Bacca.

Consecutivamente se ocupa la adaptación de la metodología CRISP-DM, que permite guiar y experimentar con los algoritmos y técnicas de combinación. Finalmente, se identifica un modelo híbrido de IA adecuado para la predicción de fallas.

6. Resultados

Los resultados de los objetivos específicos del TIC se presentarán en esta sección.

6.1. Objetivo 1: Realizar una revisión sistemática de literatura acerca de modelos predictivos basados en inteligencia artificial para la detección de fallas en aerogeneradores.

6.1.1. Resultado 1: Técnicas y modelos de inteligencia artificial más usados aplicados a la predicción de fallas de aerogeneradores.

La Figura 20 ofrece una representación gráfica extraída del **Anexo 1. Revisión Sistemática de Literatura**, que destaca la prevalencia de las técnicas y modelos predominantes utilizadas en esta área. En síntesis, se observa que SVM cuenta con 11 apariciones y ANN con 8 apariciones. Además, existe una categoría denominada "Otros" que incluye 21 elementos encontrados. Es esencial mencionar que esta categoría agrupa elementos con una única repetición y sin una conexión evidente con los demás elementos.

La Tabla 14 muestra las técnicas y modelos más utilizadas con sus respectivos autores y la Figura 20 ofrece un resumen de los mismos.

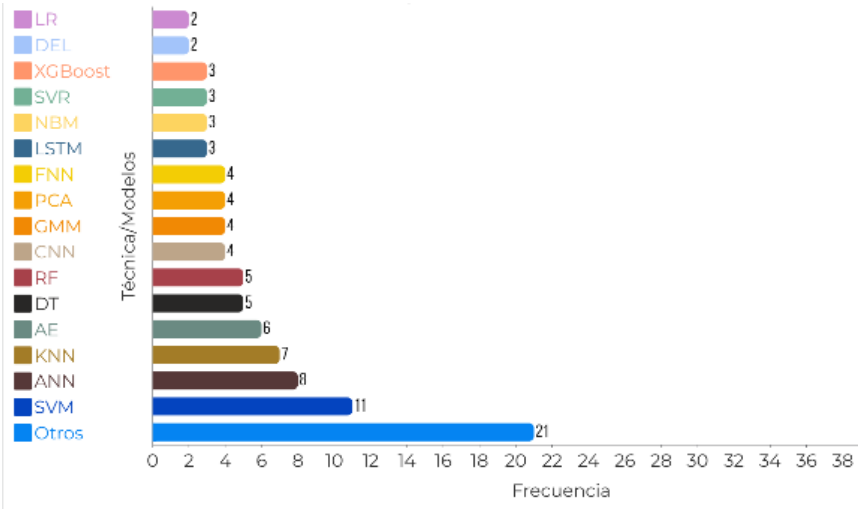

Figura 20. Frecuencias de las técnicas y modelos más usados según la RSL.

Tabla 14. Técnicas y modelos más usados

Técnicas / Modelos	Referencias	Total
LR	[96], [102]	2
DEL	[106], [33]	2
XGBoost	[102], [107], [114]	3
SVR	[93], [110], [115]	3
NBM	[90], [105], [98]	3
LSTM	[107], [87], [87]	3
FNN	[33], [96], [105], [33]	4
PCA	[85], [99], [111], [116]	4

Técnicas / Modelos	Referencias	Total
GMM	[88], [100], [101], [115]	4
CNN	[87], [89], [96], [110]	4
RF	[34], [102], [105], [109], [116]	5
DT	[102], [104], [34], [108], [116], [34]	5
AE	[90], [92], [99], [117], [118], [119]	6
KNN	[33], [34], [85], [109], [111], [117], [118]	7
ANN	[97] [94], [101], [103], [104], [109], [33], [120]	8
SVM	[34], [91], [102], [104], [105], [108], [109], [110], [111], [114], [116]	11
Otros	MSSM, MPE [91]; BiGRU-AM [89]; Catboost [102]; ML (Black Box) [95]; NARX [86]; GANs [104]; FA, CP [106]; CB [108]; RUSBoost [111]; Bagging, Boosting, Random subspace [33]; FS, FTL [87]; LoMST-CUSUM, WHC-LOF, CCA, KCPD [98]; t-SNE[118].	21

6.1.2. Resultado 2: Componentes del aerogenerador más estudiados en la predicción de fallas

La Figura 21 presenta los componentes más estudiados del aerogenerador para la predicción de fallas en donde, se destaca a Wind Turbine con una frecuencia de 18 y Gearbox con una frecuencia de 9, de un total de 51 elementos. Se presenta el campo "Otros", que incluye los componentes con una única frecuencia. La Tabla 15 muestra las referencias que respaldan dicha información.

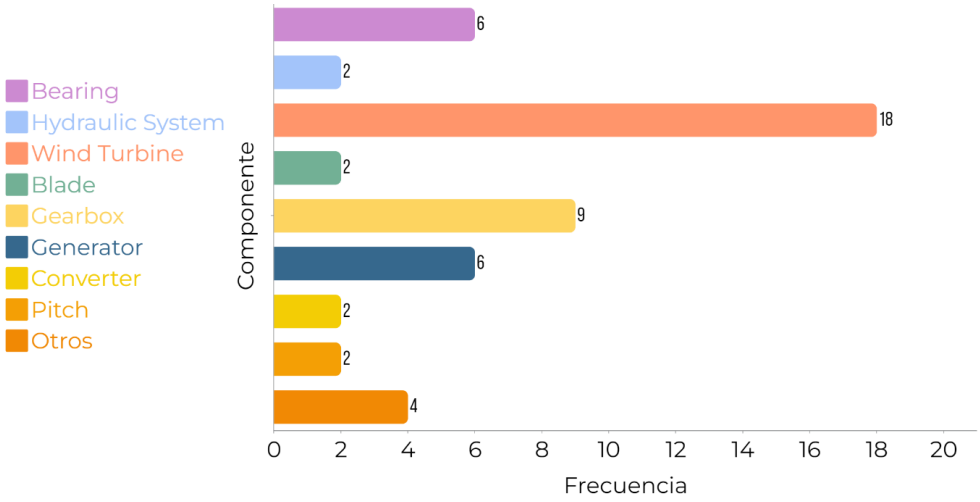

Figura 21. Frecuencias de los componentes más estudiados.

Tabla 15. Componentes más estudiados.

Componentes	Referencias	Total
Hydraulic System	[98], [119]	2
Blade	[108], [115]	2
Converter	[33], [110]	2
Pitch	[99], [110]	2

Componentes	Referencias	Total
Otros	Planetary [96]; Yaw [100]; Transformer [98]; Rotor [119].	4
Bearing	[89], [93], [105], [107]	6
Generator	[93], [98], [103], [105], [107], [110]	6
Gearbox	[86], [89], [92], [95], [98], [103], [105], [111], [119]	9
Wind Turbine	[34], [85], [87], [90], [91], [94], [97], [101], [102], [104], [106], [109], [114], [115], [116], [117], [118], [120]	18

6.1.3. Resultado 3: Artículos más importantes sobre modelos de predicción de fallas en aerogeneradores en los últimos años.

Se identifican los artículos más importantes identificados desde el año 2018 hasta junio del 2023. La Figura 22 organiza los artículos de la siguiente manera: aquellos con menos de 20 citaciones se representan mediante círculos azules, mientras que los artículos con un rango de citaciones entre 20 y 50 se muestran con círculos amarillos y los artículos con más de 50 citaciones se representan con círculos anaranjados. De los cuales, destacan cuatro documentos con más de 50 citas. La Tabla 16 detalla estos últimos.

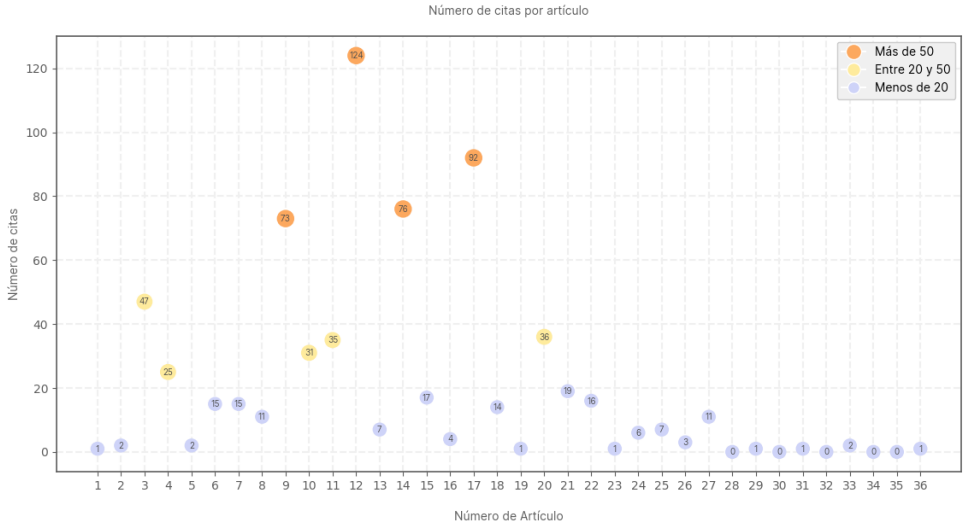

Figura 22. Documentos citados

Tabla 16. Documentos más importantes

Referencia	Título	Cant. Cit.
[93]	Real-time condition monitoring and fault detection of components based on machine-learning reconstruction model	73
[97]	Deep Learning for fault detection in wind turbines	124
[100]	SCADA-based wind turbine anomaly detection using Gaussian process models for wind turbine condition monitoring purposes	76
[104]	A Small-Sample Wind Turbine Fault Detection Method With Synthetic Fault Data Using Generative Adversarial Nets	92

Cantidad de citaciones (Cant. Cit).

6.2. Objetivo 2: Implementar un modelo híbrido para la predicción de fallas en los aerogeneradores del parque Eólico Villonaco.

6.2.1. Resultado 4: Análisis exploratorio de los datos SCADA

Este estudio se enfocó en predecir fallas en los aerogeneradores, lo cual implicó la necesidad de determinar la relación entre las variables del conjunto SCADA para identificar posibles asociaciones. Para ello, se utilizaron las 71 variables registradas en el conjunto de datos SCADA de la CEV correspondientes al año 2020. Este análisis correlacional suele ser representado visualmente mediante un mapa de calor, el cual indica las variables más relevantes mediante una escala de colores. La Figura 23 ilustra dicho análisis.

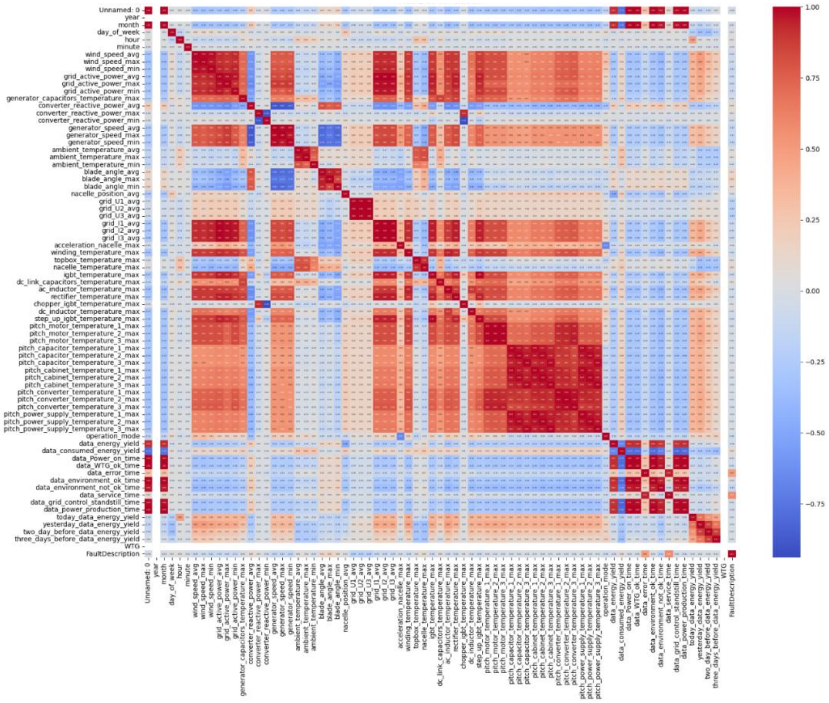

Figura 23. Análisis correlacional de las variables del sistema SCADA en el año 2020

El análisis correlacional permitió examinar la relación entre todas las variables SCADA. Al interpretar el gráfico, se pudo observar en la barra vertical que los valores cercanos a cero indicaron una falta de relación entre las variables. Los valores negativos, representados en azul, señalaron una correlación inversa, mientras que los valores positivos, mostrados en rojo, indicaron una correlación directa.

Esta observación reveló que la mayoría de las variables en el conjunto de datos SCADA mostraron múltiples relaciones entre sí, con el color rojo predominando sobre el azul. Esto indicó una tendencia hacia correlaciones positivas, sugiriendo que muchos aspectos del sistema estaban interrelacionados de manera

directa, proporcionando una perspectiva valiosa sobre la interconexión de las variables.

No obstante, en relación a determinar las variables más relevantes para el análisis de fallas en el presente trabajo, es importante destacar que las variables SCADA estuvieron sujetas a un filtro previo. Aunque un aerogenerador de la CEV registra más de 100 variables, solo se utilizaron 71 en el análisis. Esta selección se debió a dos motivos principales: primero, no todas las variables eran pertinentes para el análisis de fallas y, segundo, los datos completos no fueron proporcionados por los propietarios, Gensur. Por lo tanto, solo se dispuso de 71 variables.

Por otra parte, un análisis de variables es ocupado para identificar aquellas más relevantes con respecto a una variable objetivo. En un escenario hipotético en el que se necesite llevar a cabo un análisis de fallas para determinar las variables más importantes relacionadas con un componente específico del aerogenerador, como el convertidor, se podría utilizar una técnica como el análisis correlacional o el PCA (Análisis de Componentes Principales) sobre todas las variables disponibles. Esto permitiría identificar aquellas variables del conjunto total (71 variables) que presentan una correlación estrecha con el convertidor, reduciendo así el conjunto a un grupo más selecto de, por ejemplo, 25 variables relevantes.

Sin embargo, en principio el propósito era realizar una clasificación general de fallas en todo el aerogenerador, lo que implicaba que el propio aerogenerador era la variable objetivo. Esto significaba que todas las variables eran importantes y relevantes para el análisis y la comprensión de los algoritmos posteriores. Por lo cual, esto conllevaba a la necesidad de utilizar todas las variables disponibles en el conjunto de datos SCADA de los aerogeneradores de la CEV.

6.2.2. Resultado 5: Distribución de clases y división de datos

La Figura 24 presenta la distribución de clases luego de aplicar las técnicas de intercambio de fallas entre aerogeneradores y reducción de la clase mayoritaria por ventana de tiempo, obteniendo como resultado el siguiente índice de desbalance.

$$IR = \frac{|4420|}{|29|} = 152.4137$$

Este índice indica que hay 152 elementos en promedio por cada falla registrada, por lo cual, para mejorar este desbalance se puede observar en la gráfica conjunta como el sobremuestreo crea nuevas instancias de la clase minoritaria para equilibrarse conjunto a la clase mayoritaria.

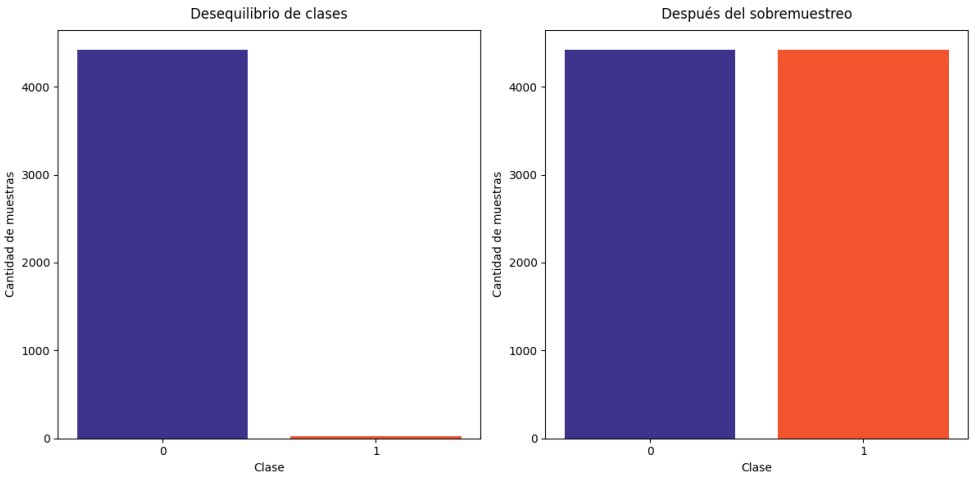

Figura 24. Desequilibrio de clases antes y después del sobremuestreo

Luego del proceso de sobremuestreo, comúnmente se dividen los datos en entrenamiento (70%) y evaluación (30%), las cantidades de esta división se presentan en la Figura 25.

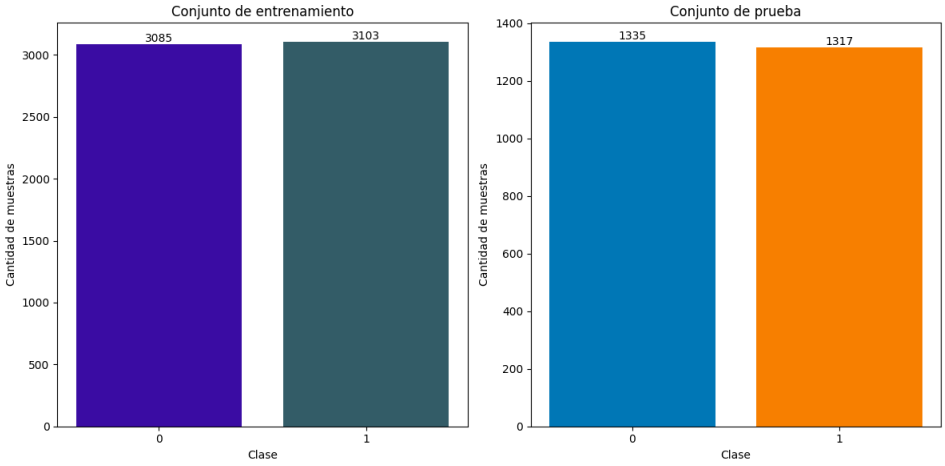

Figura 25. División de los datos en el conjunto de entrenamiento y prueba

Una observación vital encontrada de las técnicas de sobremuestreo es que aplicarlo luego de etiquetar los datos en fallas y comportamiento normal, crea nuevas instancias de la clase minoritaria que llevan al sobreajuste, lo cual fue notorio al realizar diferentes experimentos, con varios algoritmos de ML y diversos umbrales de la estrategia de sobremuestreo, umbrales como: 0.02, 0.05, 0.10, 0.25, 0.35, 0.50 76 y “automático”, no condujeron a resultados distintos. Además, se probó con la validación cruzada, y se experimentó con diferentes cantidades de pliegues: 10, 25, 100, 250, 1000. Sin embargo, cada evaluación se mostró invariable conduciendo a los mismos resultados que se buscaba evadir.

Ante estos resultados inusuales, la opción más viable fue aplicar el sobremuestreo después de la división de datos (70% para entrenamiento, 30% para prueba), aplicando el sobremuestreo al conjunto de prueba en menor proporción, por lo que la clase minoritaria obtuvo nuevos datos sintéticos para evaluar a los diferentes algoritmos y no condujo al sobreajuste, aunque la cantidad de datos de la clase minoritaria se redujo. La Figura 26 representa el conjunto ocupado para evaluar los algoritmos.

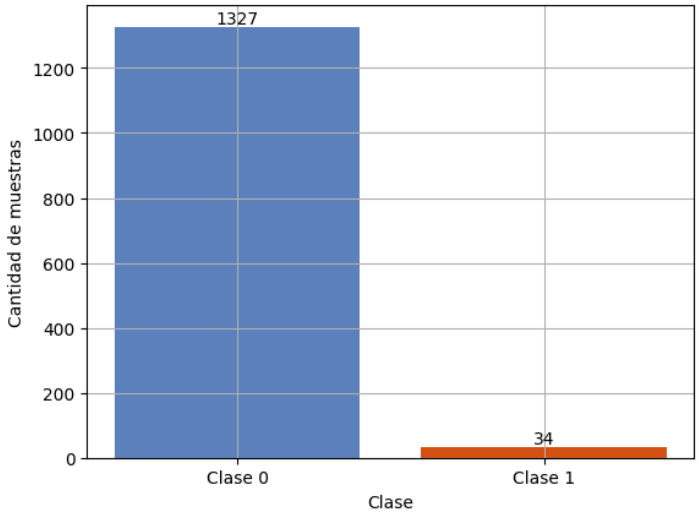

Figura 26. Conjunto de evaluación

6.2.3. Resultado 6: Algoritmos de ML y técnicas de sobremuestreo.

La Tabla 17 muestra seis algoritmos de ML y cinco técnicas de sobremuestreo, seleccionados principalmente mediante el análisis de la RSL. Además, para evaluar los resultados de las combinaciones se empleó la matriz de confusión y el cálculo de métricas específicas como: P, R, PF1, S, ES, EX y AUC en las clases (1 y 0). Estas métricas proporcionaron una visión completa del rendimiento de cada algoritmo y permitieron compararlos para identificar los mejores resultados.

Tabla 17. Algoritmos de ML y técnicas de sobremuestreo seleccionadas

Algoritmo de ML	Red Neuronal Artificial (ANN)
	Deep Neural Network (DNN)
	Support Vector Machine (SVM)
	k-Nearest Neighbors (KNN)
	Random Forest (RF)
	ExtraTree (ET)
Técnica de sobremuestreo	Synthetic Minority Over-sampling Technique (SMOTE)
	Adaptive Synthetic Sampling (ADASYN)
	SMOTE + Edited Nearest Neighbors (SMOTE-ENN)
	SMOTE + Tomek links (SMOTE-Tomek)
	Borderline Synthetic Minority Over-sampling Technique (Borderline-SMOTE)

Se llevó a cabo cinco experimentos con el propósito de comparar y evaluar el rendimiento de diversos algoritmos de ML en conjunto con diferentes técnicas de sobremuestreo. Los resultados de estos experimentos se presentan en la Tabla 18, Tabla 19, Tabla 20, Tabla 21, Tabla 22 y el resumen de los experimentos en la Tabla 24.

6.2.3.1. Experimento 1: Evaluación con la técnica SMOTE

En este experimento se evaluó la técnica SMOTE junto con los algoritmos de ML: ANN, DNN, SVM, KNN, RF y ET, la Tabla 18 exhibe los resultados de las métricas de evaluación específicas (P, R, PF1, S, ES, EX, AUC) para los diferentes algoritmos.

Tabla 18. Resultados de evaluación con la técnica SMOTE.

A	C	P	R	PF1	S	ES	EX	AUC
ANN	0	1.00	0.99	1.00	1.00	0.99	0.99	1.00
	1	0.87	1.00	0.93				
DNN	0	1.00	0.98	0.99	0.97	0.98	0.98	1.00
	1	0.62	0.97	0.76				
SVM	0	1.00	0.99	1.00	1.00	0.99	0.99	1.00
	1	0.83	1.00	0.91				
KNN	0	1.00	1.00	1.00	0.91	1.00	0.99	1.00
	1	0.86	0.96	0.89				
RF	0	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	1	0.92	1.00	0.96				
ET	0	1.00	1.00	1.00	0.94	1.00	0.99	1.00
	1	0.84	0.94	0.89				

Algoritmo (A); Clase (C); Precisión (P); Recuperación (R); Puntuación F1 (PF1); Sensibilidad (S); Especificidad (ES); Exactitud (EX); Área bajo la curva (AUC).

Se tomó en cuenta P, R y PF1 como las métricas más relevantes para la clase 1, por lo cual se puede comprender algunos aspectos sustanciales para los algoritmos más notables de este experimento.

- ANN exhibió 0.87 en P, 1.00 en R y 0.93 en PF1.
- RF presentó 0.92 en P, 1.00 en R y 0.96 en PF1.
- SVM mostró 0.83 en P, 1.00 en R y 0.91 en PF1.

Estos algoritmos cuentan con un equilibrio favorable entre P y R, lo que sugiere una proporción adecuada de aciertos. Por otra parte, algoritmos como DNN, KNN y ET mostraron en las mismas métricas un desempeño regular y menor en el caso de DNN con 0.71 en PF1.

En función de lo observado, la Figura 27 y Figura 28 presenta las matrices de confusión de los dos mejores algoritmos que sobresalen con la técnica SMOTE, siendo ANN y RF.

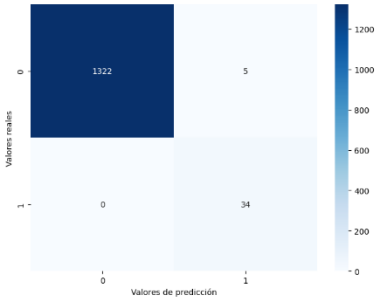

Figura 27. Matriz de confusión de ANN con SMOTE

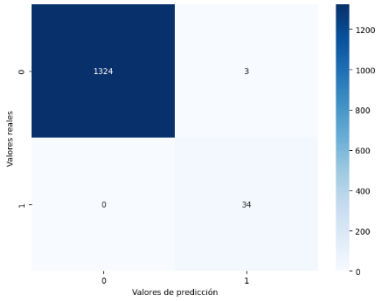

Figura 28. Matriz de confusión de RF con SMOTE

La matriz de confusión se etiqueta de la siguiente manera. El eje *x* presenta los valores de predicción y el eje *y* los valores reales. La etiqueta 0 corresponde a comportamiento normal y 1 a falla.

Se observó con ANN que:

- Predijo 1322 instancias de comportamiento normal.
- Predijo 34 instancias de fallas.
- Cometió 5 errores con instancias de comportamiento normal.
- No cometió errores con instancias de falla.

Se observó con RF que:

- Predijo 1324 instancias de comportamiento normal.
- Predijo 34 instancias de fallas.
- Cometió 3 errores con instancias de comportamiento normal.
- No cometió errores con instancias de falla.

6.2.3.2. Experimento 2: Evaluación con la técnica ADASYN

En este experimento se evaluó la técnica ADASYN junto con los algoritmos de ML: ANN, DNN, SVM, KNN, RF y ET. La Tabla 19 exhibe los resultados de las métricas de evaluación específicas (P, R, PF1, S, ES, EX, AUC) para los diferentes algoritmos.

Tabla 19. Resultados de evaluación con la técnica ADASYN.

A	C	P	R	PF1	S	ES	EX	AUC
ANN	0	1.00	0.99	1.00	1.00	0.99	0.99	1.00
	1	0.83	1.00	0.91				
DNN	0	1.00	0.96	0.98	0.91	0.96	0.96	0.99
	1	0.37	0.91	0.53				
SVM	0	1.00	0.99	1.00	1.00	0.99	0.99	1.00
	1	1.0	0.83	0.91				
KNN	0	1.00	1.00	1.00	0.91	1.00	0.99	1.00
	1	0.84	0.91	0.87				
RF	0	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	1	0.97	1.00	0.99				
ET	0	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	1	0.87	1.00	0.93				

Algoritmo (A); Clase (C); Precisión (P); Recuperación (R); Puntuación F1 (PF1); Sensibilidad (S); Especificidad (ES); Exactitud (EX); Área bajo la curva (AUC).

P, R y PF1 se consideraron las métricas más significativas para la clase 1, lo que permitió comprender varios elementos clave de los algoritmos más notables de este experimento.

- RF exhibió 0.97 en P, 1.00 en R y 0.99 en PF1.
- ET presentó 0.87 en P, 1.00 en R y 0.93 en PF1.

Se destaca que estos algoritmos mostraron un rendimiento apropiado, lo que evidencia una discriminación adecuada para las clases de intereses con una clasificación correcta. Además, se observó que SVM presento resultados igualmente destacables, aunque con un rendimiento ligeramente inferior, obteniendo 0.91 en PF1.

En cambio, para las métricas (S, ES, EX y AUC) se puede observar que RF y ET también lograron valores óptimos subrayando la capacidad para discriminar idóneamente entre las clases.

En base a esta observación, la Figura 29 y Figura 30 presentan las matrices de confusión de los dos mejores algoritmos que sobresalen con la técnica ADASYN, siendo RF y ET.

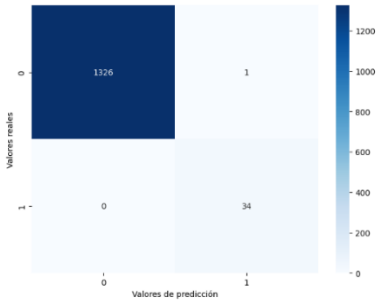

Figura 29. Matriz de confusión de RF con ADASYN

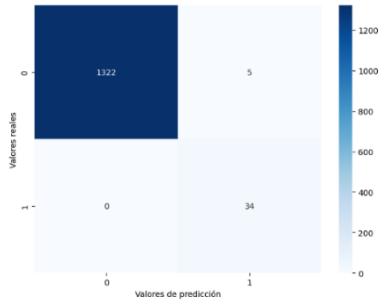

Figura 30. Matriz de confusión de ET con ADASYN

La matriz de confusión se etiqueta de la siguiente manera. El eje *x* presenta los valores de predicción y el eje *y* los valores reales. La etiqueta 0 corresponde a comportamiento normal y 1 a falla.

Se observó con RF que:

- Predijo 1326 instancias de comportamiento normal.
- Predijo 34 instancias de fallas.
- Cometió 1 error con instancias de comportamiento normal.

Se observó con ET que:

- Predijo 1322 instancias de comportamiento normal.
- Predijo 34 instancias de fallas.
- Cometió 5 errores con instancias de comportamiento normal.

- No cometió errores con instancias de falla.
- No cometió errores con instancias de falla.

6.2.3.3. Experimento 3: Evaluación con la técnica SMOTE-ENN

La técnica SMOTE-ENN se evaluó en este experimento junto con los algoritmos de ML: ANN, DNN, SVM, KNN, RF y ET. Los resultados de las evaluaciones específicas (P, R, PF1, S, ES, EX, AUC) de los distintos algoritmos se muestran en la Tabla 20.

Tabla 20. Resultados de evaluación con la técnica SMOTE-ENN.

A	C	P	R	PF1	S	ES	EX	AUC
ANN	0	1.00	0.99	1.00	1.00	0.99	0.99	1.00
	1	0.74	1.00	0.85				
DNN	0	1.00	0.97	0.98	0.91	0.97	0.97	0.99
	1	0.45	0.91	0.60				
SVM	0	1.00	0.99	1.00	1.00	0.99	0.99	1.00
	1	0.76	1.00	0.86				
KNN	0	1.00	0.99	0.99	0.91	0.99	0.99	0.95
	1	0.67	0.91	0.78				
RF	0	1.00	1.00	1.00	0.91	1.00	1.00	0.99
	1	0.94	0.91	0.93				
ET	0	1.00	0.99	1.00	0.85	0.99	0.99	0.98
	1	0.78	0.85	0.82				

Algoritmo (A); Clase (C); Precisión (P); Recuperación (R); Puntuación F1 (PF1); Sensibilidad (S); Especificidad (ES); Exactitud (EX); Área bajo la curva (AUC).

Los resultados más significativos fueron tomados en cuenta en consideración de la clase 1, enfocándose en P, R y PF1, lo que simplifico la detección de los algoritmos más apreciables de este experimento.

- RF exhibió 0.94 en P, 0.91 en R y 0.93 en PF1.
- SVM presentó 0.76 en P, 1.00 en R y 0.86 en PF1.
- ANN mostró 0.74 en P, 1.00 en R y 0.85 en PF1.

Se observo que RF exhibe resultados notables con 0.93 en PF1, sugiriendo una capacidad equilibrada y adecuada para identificar las instancias de la clase 1. Asimismo, SVM y ANN presentaron rendimientos considerables y similares en sus clasificaciones, con valores de 0.86 y 0.85 en PF1 respectivamente.

La Figura 31 y Figura 32 presenta las matrices de confusión de los dos mejores algoritmos que sobresalen con la técnica SMOTE-ENN, siendo SVM y RF.

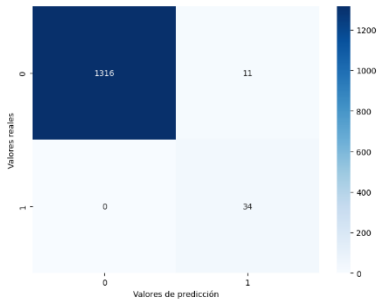

Figura 31. Matriz de confusión de SVM con SMOTE-ENN

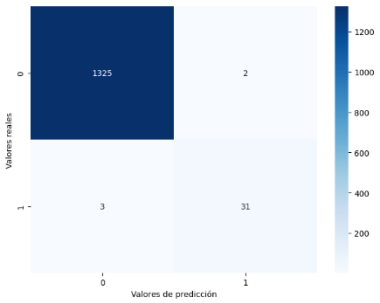

Figura 32. Matriz de confusión de RF con SMOTE-ENN

La matriz de confusión se etiqueta de la siguiente manera. El eje *x* presenta los valores de predicción y el eje *y* los valores reales. La etiqueta 0 corresponde a comportamiento normal y 1 a falla.

Se observó con SVM que:

- Predijo 1316 instancias de comportamiento normal.
- Predijo 34 instancias de fallas.
- Cometió 11 errores con instancias de comportamiento normal.
- No cometió errores con instancias de falla.

Se observó con RF que:

- Predijo 1325 instancias de comportamiento normal.
- Predijo 31 instancias de fallas.
- Cometió 2 errores con instancias de comportamiento normal.
- Cometió 3 errores con instancias de falla.

6.2.3.4. Experimento 4: Evaluación con la técnica SMOTE-Tomek

En este experimento, la técnica SMOTE-Tomek fue evaluada en conjunto con los algoritmos de ML: ANN, DNN, SVM, KNN, RF y ET. Los resultados de las evaluaciones especificas (P, R, PF1, S, ES, EX, AUC) para los distintos algoritmos se muestran en la Tabla 21.

Tabla 21. Resultados de evaluación con la técnica SMOTE-Tomek.

A	C	P	R	PF1	S	ES	EX	AUC
ANN	0	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	1	0.87	1.00	0.93				
DNN	0	1.00	0.99	1.00	1.00	0.99	0.99	1.00
	1	0.72	1.00	0.84				
SVM	0	1.00	0.99	1.00	1.00	0.99	0.99	1.00
	1	0.83	1.00	0.91				
KNN	0	1.00	1.00	1.00	0.91	1.00	0.99	1.00
	1	0.86	0.91	0.89				
RF	0	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	1	0.92	1.00	0.96				
ET	0	1.00	1.00	1.00	0.94	1.00	0.99	1.00
	1	0.84	0.94	0.89				

Algoritmo (A); Clase (C); Precisión (P); Recuperación (R); Puntuación F1 (PF1); Sensibilidad (S); Especificidad (ES); Exactitud (EX); Área bajo la curva (AUC).

La clase 1 fue objeto de los resultados más significativos centrándose en P, R y PF1, lo que facilito la identificación de los algoritmos más destacados. A continuación, se describen las principales observaciones.

- ANN exhibió 0.87 en P, 1.00 en R y 0.93 en PF1.
- RF presentó 0.92 en P, 1.00 en R y 0.96 en PF1.

Se observó que ANN y RF obtuvieron resultados satisfactorios para identificar y clasificar apropiadamente la clase 1, señalándolos como enfoques útiles. Adicionalmente, es evidente que SVM y KNN presentan resultados próximos en las mismas métricas, con valores PF1 de 0.91 y 0.89, respectivamente.

En función de lo identificado, la Figura 33 y Figura 34 presenta las matrices de confusión de los dos mejores algoritmos que sobresalen con la técnica SMOTE-Tomek, siendo ANN y RF.

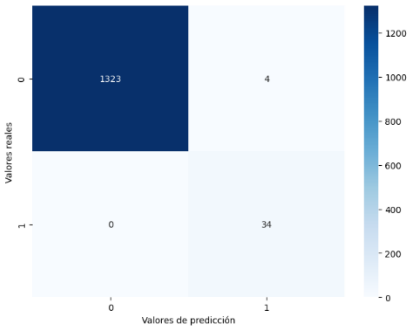

Figura 33. Matriz de confusión de ANN con SMOTE-Tomek

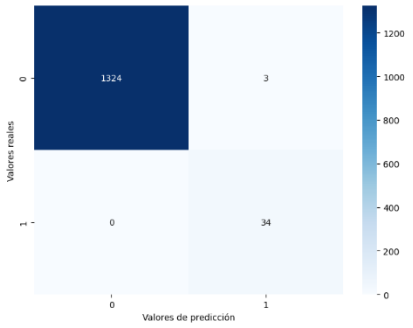

Figura 34. Matriz de confusión de RF con SMOTE-Tomek

La matriz de confusión se etiqueta de la siguiente manera. El eje *x* presenta los valores de predicción y el eje *y* los valores reales. La etiqueta 0 corresponde a comportamiento normal y 1 a falla.

Se observó con ANN que:

- Predijo 1323 instancias de comportamiento normal.
- Predijo 34 instancias de fallas.
- Cometió 4 errores con instancias de comportamiento normal.
- No cometió errores con instancias de falla.

Se observó con RF que:

- Predijo 1324 instancias de comportamiento normal.
- Predijo 34 instancias de fallas.
- Cometió 3 errores con instancias de comportamiento normal.
- No cometió errores con instancias de falla.

6.2.3.5. Experimento 5: Evaluación con la técnica Borderline-SMOTE

En este experimento los algoritmos de ML: ANN, DNN, SVM, KNN, RF y ET se utilizaron en conjunto con la técnica de sobremuestreo Borderline-SMOTE para su evaluación. Los resultados de las evaluaciones particulares para cada uno de los algoritmos se muestran en la Tabla 22.

Tabla 22. Resultados de evaluación con la técnica Bordeline-SMOTE.

A	C	P	R	PF1	S	ES	EX	AUC
ANN	0	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	1	0.85	1.00	0.92				
DNN	0	1.00	0.98	0.99	1.00	0.98	0.98	1.00
	1	0.55	1.00	0.71				
SVM	0	1.00	0.99	1.00	1.00	0.99	0.99	1.00
	1	0.83	1.00	0.91				
KNN	0	1.00	1.00	1.00	0.82	1.00	0.99	1.00
	1	0.85	0.82	0.84				
RF	0	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	1	0.89	1.00	0.94				
ET	0	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	1	0.89	1.00	0.94				

Algoritmo (A); Clase (C); Precisión (P); Recuperación (R); Puntuación F1 (PF1); Sensibilidad (S); Especificidad (ES); Exactitud (EX); Área bajo la curva (AUC).

Los resultados del experimento permitieron captar ciertos aspectos remarcables que se analizan a continuación, con un énfasis en la clase 1 y considerando P, R y PF1.

- ANN exhibió 0.85 en P, 1.00 en R y 0.92 en PF1.
- RF presentó 0.89 en P, 1.00 en R y 0.94 en PF1.
- ET mostró 0.89 en P, 1.00 en R y 0.94 en PF1.

Tras analizar los datos, se evidenció que RF y ET mostraron un rendimiento idéntico sin diferencias perceptibles incluso teniendo en cuenta las métricas S, ES, EX y AUC. Además, ANN indicó un promedio aceptable con 0.92 en PF1. Estos algoritmos fueron considerados como las elecciones en el marco del experimento.

En adición, es esencial mencionar que SVM y KNN mostraron resultados semejantes, con valores PF1 de 0.91 y 0.84, respectivamente.

En el sentido del análisis, la Figura 35, Figura 36 y Figura 37 presenta las matrices de confusión de los tres mejores algoritmos que sobresalieron con la técnica SMOTE-Tomek: RF, ET y ANN.

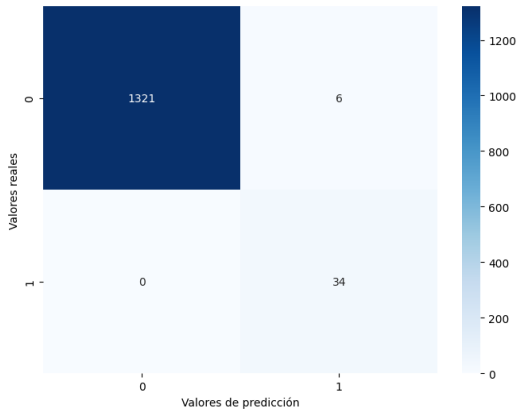

Figura 35. Matriz de confusión de ANN con Borderline-SMOTE

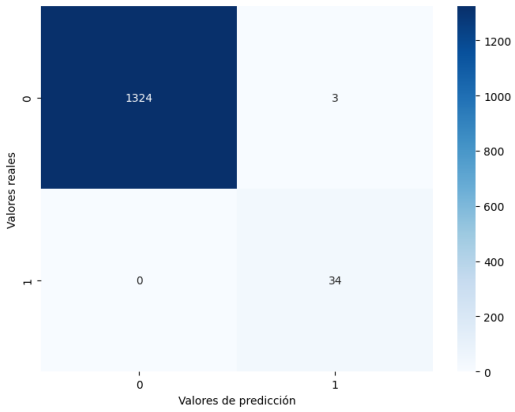

Figura 36. Matriz de confusión de RF con Borderline-SMOTE

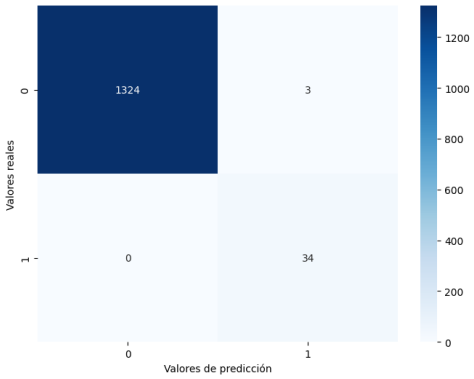

Figura 37. Matriz de confusión de ET con Borderline-SMOTE

La matriz de confusión se etiqueta de la siguiente manera. El eje *x* presenta los valores de predicción y el eje *y* los valores reales. La etiqueta 0 corresponde a comportamiento normal y 1 a falla.

Se observó con ANN que:

- Predijo 1321 instancias de comportamiento normal.
- Predijo 34 instancias de fallas.
- Cometió 6 errores con instancias de comportamiento normal.
- No cometió errores con instancias de falla.

Se observó con RF y ET que:

- Predijo 1324 instancias de comportamiento normal.
- Predijo 34 instancias de fallas.
- Cometió 3 errores con instancias de comportamiento normal.
- No cometió errores con instancias de falla.

6.2.3.6. Puntuaciones destacadas de los algoritmos de ML evaluados

Los resultados se basan en las métricas P, R y PF1, que se eligieron por su capacidad para proporcionar una comprensión directa y completa del rendimiento del modelo, especialmente en situaciones en las que existe un

desequilibrio de clases. Asimismo, en estudios como [34], [97], [99], [103], [108], [110], [111] estás métricas son esenciales para permitir a los investigadores comparar la efectividad de diferentes técnicas de sobremuestreo y algoritmos de ML en términos de su capacidad no solo para clasificar correctamente las instancias sino también para manejar el desbalance de clases de manera efectiva.

A continuación, se detallan las mejores puntuaciones de los algoritmos por cada experimento en la Tabla 23.

Tabla 23. Mejores puntuaciones de algoritmos de ML evaluados.

Técnica	Algoritmo	Clase	Precisión	Recuperación	Puntuación F1
SMOTE	ANN	0	1.00	0.99	1.00
		1	0.87	1.00	0.93
	RF	0	1.00	1.00	1.00
		1	0.92	1.00	0.96
ADASYN	RF	0	1.00	1.00	1.00
		1	0.97	1.00	0.99
	ET	0	1.00	1.00	1.00
		1	0.87	1.00	0.93
SMOTE-ENN	SVM	0	1.00	0.99	1.00
		1	0.76	1.00	0.86
	RF	0	1.00	1.00	1.00
		1	0.94	0.91	0.93
SMOTE-tomek	ANN	0	1.00	1.00	1.00
		1	0.87	1.00	0.93
	RF	0	1.00	1.00	1.00
		1	0.92	1.00	0.96
Borderline-SMOTE	ANN	0	1.00	1.00	1.00
		1	0.85	1.00	0.92
	RF	0	1.00	1.00	1.00
		1	0.89	1.00	0.94
	ET	0	1.00	1.00	1.00
		1	0.89	1.00	0.94

Se destaca que a RF y ANN emergen como los algoritmos que mantienen consistentemente altos niveles para ambas clases en la mayoría de las técnicas de sobremuestreo evaluadas (ver Figura 38). Estos algoritmos demostraron ser más robustos en la clasificación de conjuntos de datos desequilibrados.

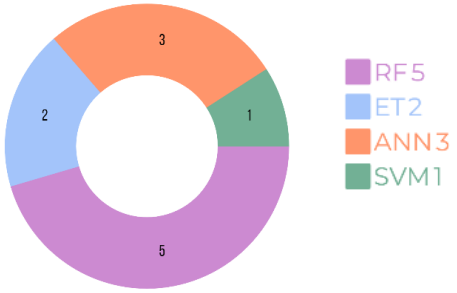

Figura 38. Algoritmos evaluados con mejores puntuaciones.

6.2.3.7. Resumen de experimentos

Tras realizar los experimentos utilizando diversas técnicas de sobremuestreo (SMOTE, ADASYN, SMOTE-ENN, SMOTE-Tomek y Borderline-SMOTE), y evaluar los algoritmos de ML (ANN, DNN, SVM, KNN, RF y ET). En la Tabla 24 se resume la selección de los mejores algoritmos por cada experimento realizado, donde cabe destacar que por cada experimento se seleccionaron dos algoritmos.

Tabla 24. Resumen de los experimentos entre las técnicas de sobremuestreo y los algoritmos seleccionados

		Técnicas de sobremuestreo				
		SMOTE	ADASYN	SMOTE-ENN	SMOTE-Tomek	Borderline-SMOTE
Algoritmo	ANN	X			X	X
	DNN					
	SVM			X		
	KNN					
	RF	X	X	X	X	X
	ET		X			X

Esta tabla esencialmente resume los resultados de la Tabla 23, que exhibe los puntajes utilizados para evaluar el rendimiento.

6.2.4. Resultado 7: Técnicas de combinación

Entre los diversos resultados observados, se destaca la estrategia de potenciar las métricas de evaluación mediante la aplicación de técnicas de combinación de algoritmos, dando lugar a un modelo híbrido. En la Tabla 25, se presenta las estrategias para combinar los algoritmos ANN y RF.

Tabla 25. Técnicas de combinación

Número	**Técnica de combinación**	**Nombre completo**
1	Bagging	Bootstrap Aggregating
2	Stacking	Stacked Generalization
3	Boosting	Boosting
4	Voting	Voting Ensemble
5	Blending	Blending Ensemble
6	Weighted Average Ensemble	Weighted Average Ensemble

7.

6.2.4.1. Evaluación de las técnicas de combinación

Las técnicas de combinación permiten elevar la precisión y la capacidad de generalización de los modelos, aun cuando estos tienden a tener un desequilibrio entre clases, por lo que experimentar con diferentes técnicas de combinación fue fundamental para determinar cuál de ellas se adaptaba mejor al problema en cuestión.

La Tabla 26 recopila los resultados obtenidos al evaluar las estrategias de combinación entre ANN y RF, los cuales se fusionaron mediante distintas estrategias con el objetivo de analizar su rendimiento al trabajar en conjunto.

Tabla 26. Resultados de evaluación de técnicas de combinación

A	C	P	R	PF1	S	ES	EX	AUC	VC	TE
Blending	0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	100.00%	0.03 s
	1	1.00	1.00	1.00						
Boosting	0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	100.00%	0.10 s
	1	1.00	1.00	1.00						
Stacking	0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	99.93%	0.62 s
	1	1.00	1.00	1.00						
Voting	0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	99.92%	8.87 s
	1	0.94	1.00	0.97						
Bagging	0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	99.92%	10.91 s
	1	1.00	1.00	1.00						
Weighted Average Ensemble	0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	99.82%	9.74 s
	1	1.00	1.00	1.00						

Algoritmo (A); Clase (C); Precisión (P); Recuperación (R); Puntuación F1 (PF1); Sensibilidad (S); Especificidad (ES); Exactitud (EX); Área bajo la curva (AUC); Validación Cruzada (VC); Tiempo de Entrenamiento (TE).

Tras examinar los datos de la Tabla 26, se extrajeron algunas observaciones notables.

- **Rendimiento general:** En términos de P, R, PF1, S, ES y EX, todas las técnicas examinadas mostraron un rendimiento particular para las clases 1 y 0 y la proximidad de estos valores a 1 indicó una capacidad de diferenciación sustancial en los modelos.
- **AUC:** Todas las técnicas de combinación obtuvieron un AUC elevado, lo que sugirió una capacidad discriminativa en la clasificación binaria. Un AUC cercano a 1 indicó un rendimiento sólido en la distinción entre las clases 1 y 0.
- **Validación Cruzada:** Utilizando esta técnica, se evaluó la fiabilidad del rendimiento del modelo en busca de resultados más representativos y generalizables a través de diversos conjuntos de datos, en consecuencia, todas las técnicas de combinación mostraron un rendimiento notable en VC con una precisión aceptable en la clasificación de datos en varias divisiones de entrenamiento y prueba.
- **Tiempo de entrenamiento:** Se observó la diferencia entre los tiempos de entrenamiento de las diferentes técnicas, por una parte, Blending y Boosting emplearon un tiempo reducido (0,03 y 0,10 segundos, respectivamente)

mientras que Bagging y Weighted Average Ensemble necesitaron más tiempo (10,91 y 9,74 segundos, respectivamente), dicha apreciación fue indicativa al tener en cuenta las preferencias temporales y la necesidad de respuestas rápidas frente a un mayor rendimiento previsto, de manera que, algunas de ellas con tiempos de respuesta más cortos como Boosting y Blending pudieron ser mejores cuando se requerían respuestas rápidas, así, en contraste, si el tiempo de entrenamiento no fuera una preocupación importante ciertas técnicas como Bagging y Weighted Average Ensemble serían opciones apropiadas

6.2.4.2. Matriz de confusión y curva ROC de las técnicas de combinación

Se presentan la matriz de confusión y curva ROC de las técnicas de combinación, estas se exhiben en el mismo orden que en la Tabla 25 y la organización es la siguiente:

- Figura 39 y Figura 40: Técnica Blending.
- Figura 41 y Figura 42: Técnica Boosting.
- Figura 43 y Figura 44: Técnica Stacking.
- Figura 45 y Figura 46: Técnica Voting.
- Figura 47 y Figura 48: Técnica Bagging.
- Figura 49 y Figura 50: Técnica Weighted Average Ensemble.

Los gráficos inferiores destacan algunos aspectos interesantes. En esencia cada una de las figuras cuentan con los mismos resultados sin diferencias discernibles tanto en la matriz de confusión como en la curva ROC y en consecuencia la interpretación solo se ve necesaria una vez.

Las figuras se comprenden de la siguiente manera:

- **Matriz de confusión**: El eje *x* presenta los valores de predicción y el eje *y* los valores reales. La etiqueta 0 corresponde a comportamiento normal y 1 a falla.
- **Curva ROC:** El eje *x* representa la tasa de falsos positivos y el eje *y* la tasa de verdaderos positivos. El AUC es el valor cuantitativo del modelo. La línea azul es una línea de referencia de un clasificador aleatorio, mientras que la línea amarilla corresponde a la curva del modelo.

Se encontró la siguiente interpretación. Para las matrices de confusión de las técnicas de combinación se observó que:

- Predijeron 1327 instancias de comportamiento normal.
- Predijeron 34 instancias de fallas.
- No cometieron errores con instancias de comportamiento normal.
- No cometieron errores con instancias de falla.

Para las curvas ROC se observó que:

- Las curvas ROC contaron con un AUC de 1.00 lo que sugirió una capacidad de discriminación destacada entre las clases.

1. Blending

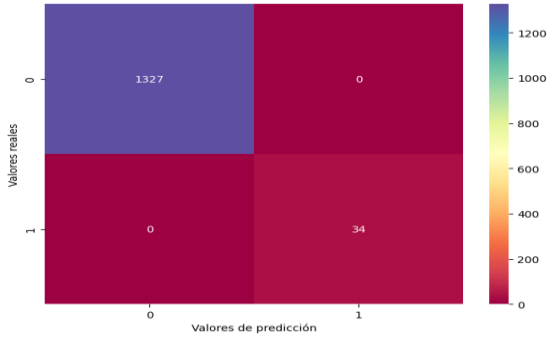

Figura 39. Matriz de confusión de la técnica Blending

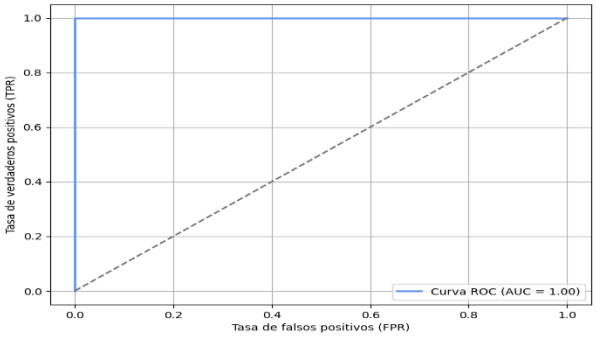

Figura 40. Curva ROC de la técnica Blending

2. Boosting

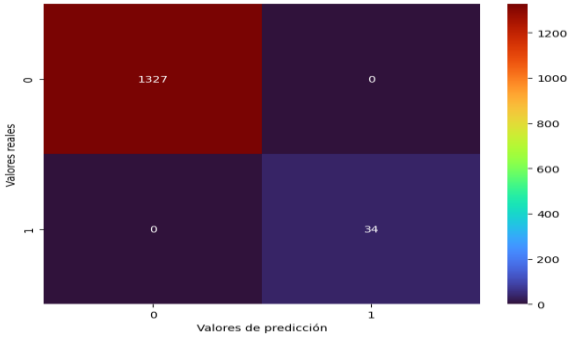

Figura 41. Matriz de confusión de la técnica Boosting

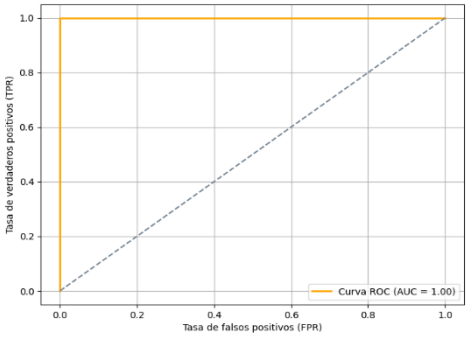

Figura 42. Curva ROC de la técnica Boosting

3. Stacking

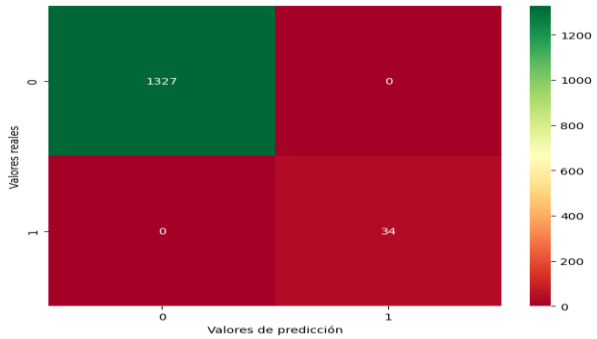

Figura 43. Matriz de confusión de la técnica Stacking

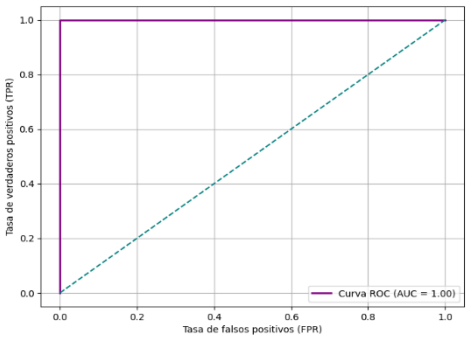

Figura 44. Curva ROC de la técnica Stacking

4. Voting

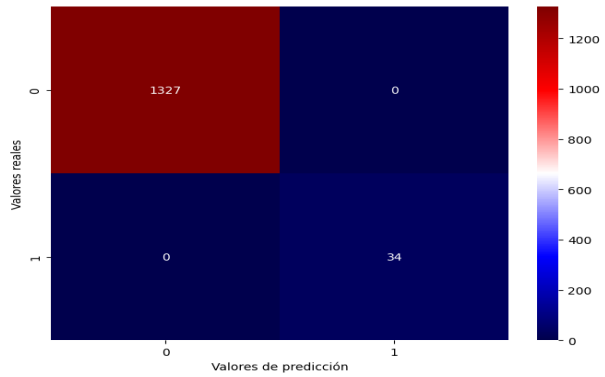

Figura 45. Matriz de confusión de la técnica Voting

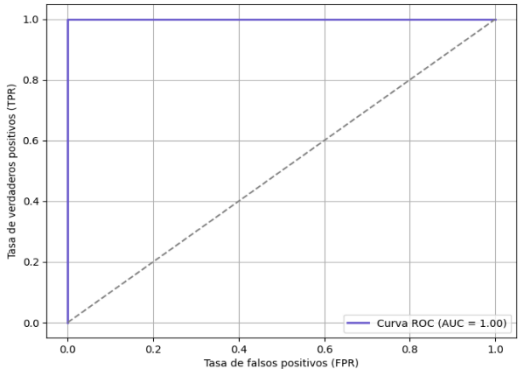

Figura 46. Curva ROC de la técnica Voting

5. Bagging

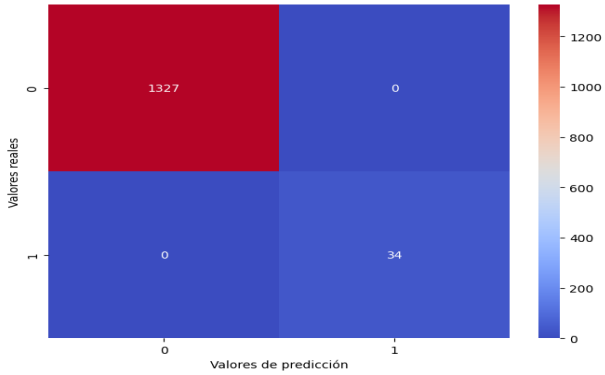

Figura 47. Matriz de confusión de la técnica Bagging

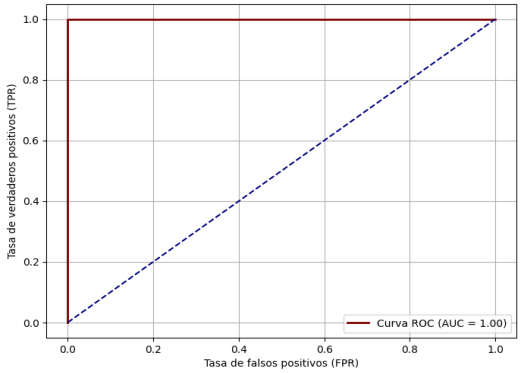

Figura 48. Curva ROC de la técnica Bagging

6. Weighted Average Ensemble

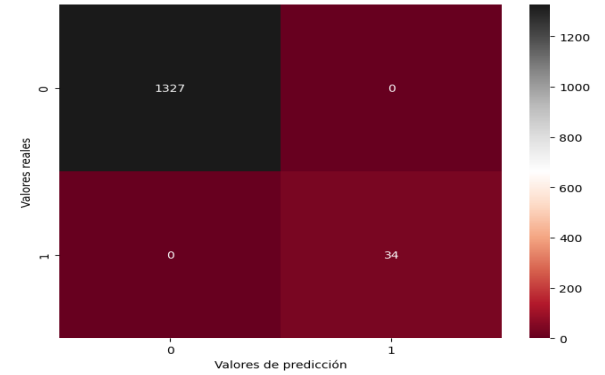

Figura 49. Matriz de confusión de la técnica Weighted Average Ensemble

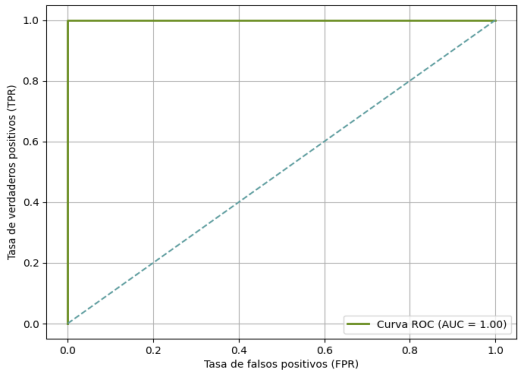

Figura 50. Curva ROC de la técnica Weighted Average Ensemble

6.2.5. Resultado 8: Test de Wilcoxon

La Tabla 27 presenta los resultados del test de Wilcoxon de las técnicas de combinación. Estos resultados cuentan con la media aritmética de los valores obtenidos del test de Wilcoxon por cada técnica evaluada, lo que permitió ver el nivel de diferencia general entre los modelos evaluados.

Tabla 27. Resultados del test de Wilcoxon

Técnica	Algoritmos evaluados	Test de Wilcoxon	Media
Stacking	ANN-RF	0.1441	0.3132
	Meta Model (ExtraTrees) - RF	0.1408	
	Meta Model (ExtraTrees) - ANN	0.6547	
Bagging	ANN - RF	0.0679	0.1503
	Bagging - RF	0.3173	
	Bagging - ANN	0.0656	
Voting	ANN - RF	0.0679	0.1503
	Voting - RF	0.3173	
	Voting - ANN	0.0656	
Weighted Average Ensemble	ANN - RF	0.0625	0.0934
	WAE - RF	0.1088	
	WAE - ANN	0.1088	
Boosting	ANN - RF	0.0679	0.0653
	Boosting - RF	0.0656	
	Boosting - ANN	0.0625	
Blending	ANN - RF	0.0679	0.0653
	Blending (Logistic regression) – RF	0.0656	
	Blending (Logistic regression) - ANN	0.0625	

El test de Wilcoxon mostró un rendimiento similar entre los algoritmos evaluados, la técnica Stacking destacó con una media de 0,3132, seguido de cerca por Bagging y Voting, con una puntuación media de 0,1503. En contraste, Weighted Average Ensemble, Boosting y Blending obtuvieron puntuaciones ligeramente inferiores con medias de 0,0934, 0,0653 y 0,0653, respectivamente.

Es relevante mencionar que todas las técnicas cumplieron con el criterio del test de Wilcoxon, lo que indica que no hubo variaciones estadísticamente relevantes en los resultados, esta perspectiva respaldó las capacidades comparativas de las técnicas examinadas.

6.2.6. Resumen de resultados.

Los experimentos mostraron que los algoritmos RF y ANN lograron puntuaciones destacadas con las técnicas de sobremuestreo, apareciendo en cinco y tres de los cinco experimentos realizados, respectivamente. Es importante recalcar que en cada experimento se eligieron los dos algoritmos con las mejores puntuaciones.

El experimento con las técnicas de combinación mostró que Blending y Boosting exhibieron un rendimiento optimo en términos de P, R, PF1, S, ES, EX, y AUC. Adicionalmente, Blending y Boosting tienen tiempos de entrenamiento muy reducidos, con 0.03 y 0.10 segundos respectivamente.

El test de Wilcoxon utilizó un umbral de significancia de 0.05, por lo que un valor cercano a 1 sugirió poca o ninguna diferencia significativa, mientras que uno cercano a 0 indicó una diferencia significativa. Todos los valores de las técnicas de combinación superaron este umbral, lo que implicó que no existieron diferencias estadísticamente significativas entre los algoritmos evaluados.

Teniendo en cuenta los resultados de las evaluaciones, tanto los modelos obtenidos por Boosting como por Blending resultaron ser técnicas eficaces, pero con una ligera ventaja en tiempo de entrenamiento para la técnica Blending y sin diferencias significativas según el test de Wilcoxon.

La elección de un modelo híbrido se fundamentó en la técnica Blending, de acuerdo con las evaluaciones realizadas y dos conceptos esenciales explicados en la Tabla 28.

Tabla 28. Complejidad vs Interpretabilidad

Complejidad del modelo	Interpretabilidad
Es el número de algoritmos independientes que intervienen en la técnica de combinación y la forma en que se combinan sus predicciones, dentro de este marco, técnicas como Bagging y Voting utilizaban combinaciones independientes de muchos algoritmos base, por otro lado, técnicas como Stacking y Blending utilizaban algoritmos más complejos que podían aprender a combinar las predicciones de los modelos base.	Esta idea estaba relacionada con la facilidad para entender y explicar el proceso de toma de decisiones de los modelos, por consiguiente, técnicas como Bagging y Voting produjeron modelos más interpretables al votar o promediar las predicciones de los algoritmos subyacentes. En cambio, técnicas más complejas como Stacking y Blending pueden haber producido resultados menos claros en términos de interpretación debido a su proceso de aprendizaje más complejo para integrar las predicciones de los modelos base.

6.2.7. Resultado 9: Modelo híbrido Blending-ANN-RF

En base a los diversos experimentos y evaluaciones, se identificó que la técnica Blending sobresalió entre las demás como un modelo adecuado y equilibrado. Como resultado, la Figura 51 presenta el funcionamiento del modelo híbrido denominado Blending-ANN-RF.

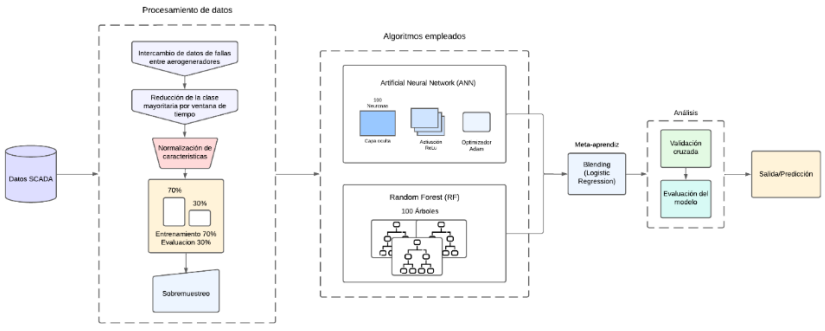

Figura 51. Modelo Blending-ANN-RF

Se pueden visualizar las etapas que componen el modelo híbrido, desde la utilización de los datos SCADA hasta la predicción. A continuación, se explica brevemente cada una de las etapas:

- **Datos SCADA:** El primer paso fue utilizar los datos SCADA de la CEV.
- **Procesamiento de datos:** Se aplicaron a los datos una serie de transformaciones y técnicas para prepararlos para el análisis y modelización, entre las que se incluyen:
    - **Intercambio de datos de falla entre aerogeneradores:** A través de este proceso, la información fue intercambiada entre los aerogeneradores, de forma que aquellos con menos fallas transmitieron las suyas al aerogenerador con más, unificándolas en un único aerogenerador y teniendo en cuenta el momento exacto en que ocurrieron.
    - **Reducción de la clase mayoritaria por ventana de tiempo:** Dentro de un margen de tres horas antes y después de encontrar una falla, únicamente se guardaron los datos cercanos a la misma.
    - **Normalización de características:** Se mejoró la capacidad de generalización del modelo al escalar las características de los datos al establecer los rangos entre 1 y 0.
    - **Entrenamiento 70% y evaluación 30%:** El 30% de los datos se ocupó para la evaluación y el 70% restante para el aprendizaje del modelo.
    - **Sobremuestreo:** Esta técnica aumentó la cantidad de datos de la clase minoritaria mediante la creación de muestras sintéticas que ayudaron a contrarrestar el desequilibrio de clases y optimizar el rendimiento del modelo.
- **Algoritmos empleados:** Se utilizaron los algoritmos ANN y RF:
    - **ANN:** Se aplicó una capa oculta estructurada de cien neuronas cuya función de unidad lineal rectificada (ReLU) introduce no linealidades en el modelo y captura con una mayor precisión las relaciones complejas entre los distintos datos y el optimizador Adam disminuye la función de pérdida durante el tiempo de entrenamiento de la red neuronal.

- **RF:** Se utilizó la cantidad de cien árboles de decisión que conformaron el bosque de árboles aleatorios.

- **Meta-aprendiz:** Con la finalidad de obtener una predicción final se implementó un algoritmo adicional (Regresión logística) que cumplió el papel de meta-aprendiz para promediar las predicciones efectuadas por los algoritmos base (ANN y RF).
- **Análisis:** La estimación del rendimiento del modelo se dividió en dos etapas que se detallan a continuación:
  - **Validación cruzada:** Permitió realizar una validación del rendimiento del modelo y estimar la capacidad de generalización considerando datos no vistos.
  - **Evaluación del modelo:** Proporcionó una comprensión mucho más completa respecto a la capacidad del modelo tomando en cuenta a métricas como P, R, PF1, S, ES, EX y AUC.
- **Salida / Predicción:** Permitió conocer el resultado final generado por el modelo.

6.2.7.1. Predicción de fallas y probabilidad de aciertos

Se analizó la capacidad predictiva del modelo híbrido Blending-ANN-RF en referencia a la detección de fallas, examinando su capacidad en términos de probabilidad de aciertos y predicción de fallas.

1. **Predicción de fallas**

   La capacidad de un modelo para anticipar o prevenir la presencia de fallas se denomina como predicción de fallas y se evaluaron varios enfoques para determinar la capacidad del modelo.

   a. **Comparativa entre la predicción y el valor real.**

      Contrastar los valores reales y los previstos otorgó una manera de detectar tendencias, desviaciones y el grado de ajuste entre ellos.

      En relación a lo antes expuesto, en la Figura 52 se logró contrastar las predicciones con los datos reales distinguiéndose en el eje de las ordenadas los dos tipos de clases que son 1 (Fallas) y 0 (Comportamiento normal) y en el eje de las abscisas se visualizó el número de muestras evaluadas donde las predicciones del modelo se

plasmaron con una "x" de color rojo y los datos reales con un círculo azul.

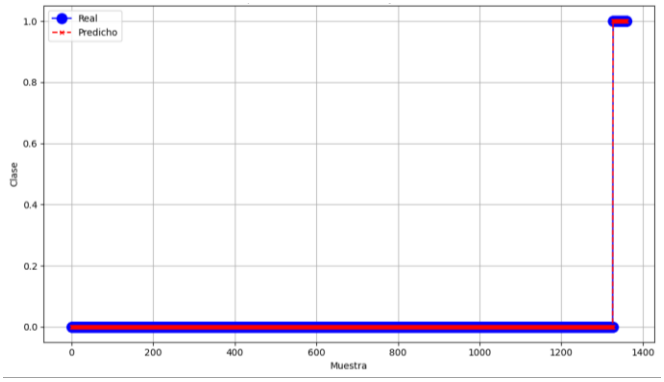

Figura 52. Comparativa entre la predicción y el valor real del modelo Blending-ANN-RF

Al contrastar las predicciones del modelo con los datos reales se visualizó que cada predicción se ubicaba cerca de los datos reales correspondientes sugiriendo que el modelo fue coherente en su capacidad de reconocer entre las diferentes clases de interés sin evidenciar discrepancias perceptibles.

b. Curva de calibración

Para valorar el rendimiento del modelo en términos de su capacidad de predicción se empleó la curva de calibración que ilustra el contraste entre las probabilidades previstas por el modelo y la frecuencia real de los sucesos observados.

La Figura 53 representó la curva de calibración indicando en el eje *x* la probabilidad promedio establecida por el modelo y en el eje *y* la frecuencia promedio real del evento.

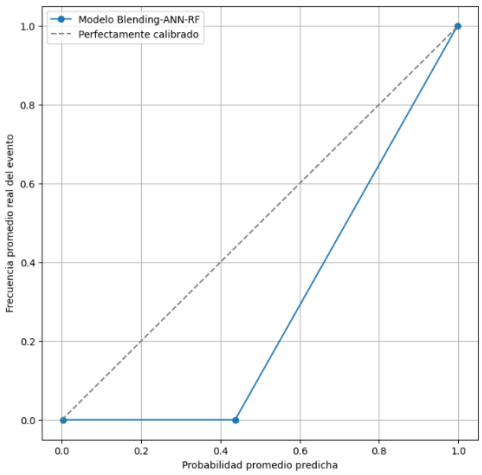

Figura 53. Curva de calibración del modelo Blending-ANN-RF

Tres elementos clave identificados por la curva de calibración se explican a continuación:

- En la coordenada de origen (0,0), se exhibió que cuando el modelo predijo una probabilidad de 0 dichos valores reales también fueron 0, indicando una predicción acertada de datos de comportamiento normal.
- En las coordenadas (0, 0.44) se pudo observar que cuando el modelo predecía una probabilidad del 0.44% de falla, los datos reales indicaban un comportamiento normal, sugiriendo una sobrevaloración para este suceso en particular.
- En el vértice (1,1) se logró revelar que cuando el modelo predijo una falla, los datos reales efectivamente reflejaron la presencia de una falla, lo cual indicó una correspondencia adecuada entre los eventos observados y la predicción del modelo.

En la Figura 53 se proporcionó una visualización de la distinción entre los datos normales y las fallas, lo que subrayó la importancia de ajustar la precisión del modelo para predecir el comportamiento normal, como se hizo hincapié para el segundo punto.

2. Probabilidad de aciertos

La probabilidad de aciertos corresponde al valor de confianza en las predicciones de un modelo, señalando la confianza del mismo en sus predicciones sobre la ocurrencia de eventos reales, especialmente cuanto mayor sea la probabilidad de aciertos también deberá ser la confianza en la precisión de las predicciones del modelo.

a. Validación cruzada

Para evaluar el rendimiento del modelo, la validación cruzada permitió dividir los datos en varias particiones y alternar entre el uso de una parte como datos de entrenamiento y otra como datos de prueba, evitando así que el modelo memorizara los datos y aprendiera de patrones generalizables.

La Figura 54 presenta los puntajes de evaluación de la validación cruzada de los algoritmos base de la técnica Blending,

donde en el eje *x* se representa las iteraciones, mientras que el eje *y* muestra las puntuaciones obtenidas, utilizando la precisión como criterio de evaluación.

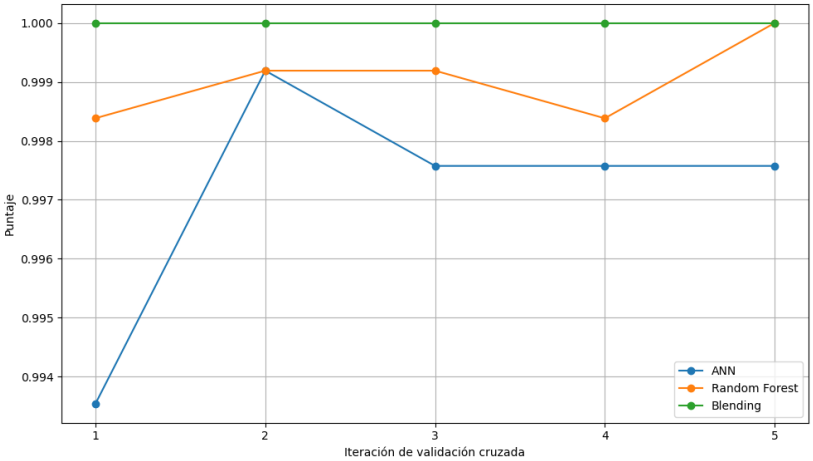

Figura 54. Validación cruzada del modelo Blending-ANN-RF

Según los resultados del modelo Blending-ANN-RF, ANN presentó la precisión más baja en la primera iteración, mientras que RF alcanzó su máximo en la iteración final, igualando la precisión consistente de Blending a lo largo de todas las iteraciones.

Hay que señalar que se realizaron cinco iteraciones en la validación cruzada, elección que resultó suficiente para obtener resultados consistentes, como lo demostró la constante precisión de la técnica de Blending desde la primera iteración.

b. Curva precisión-recuperación

Para evaluar el rendimiento del modelo en situaciones de desequilibrio de clases, la curva de precisión-recuperación resultó especialmente útil, ilustrando la identificación de instancias positivas y la prevención de falsos positivos.

En la Figura 55, se representó en el eje *x* la recuperación y en el eje *y* la precisión, además, PF1 evidenció un equilibrio entre recuperación y precisión con el objetivo de minimizar los falsos positivos y negativos y el AUC proporcionó una medida cuantitativa del rendimiento del modelo.

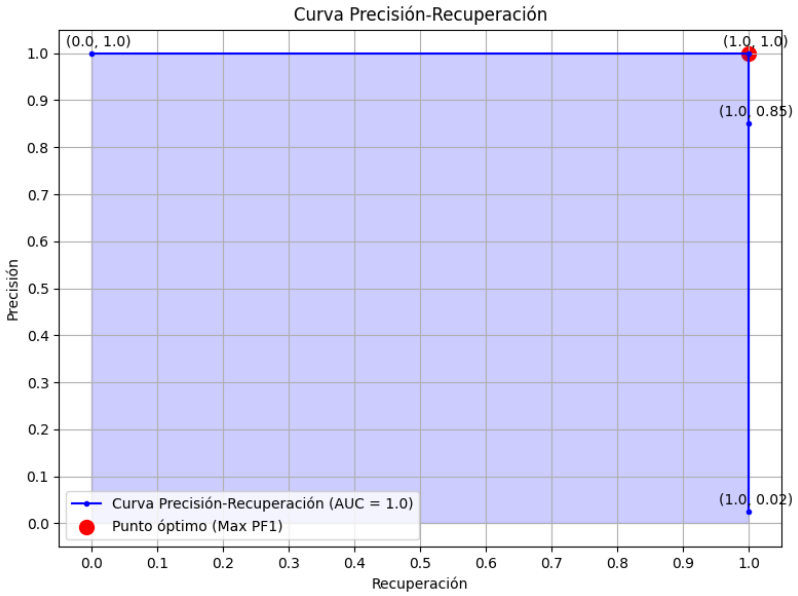

Figura 55. Curva precisión-recuperación

Se evidenció un equilibrio en el análisis de precisión y recuperación al identificar correctamente los casos positivos y las instancias que eran realmente positivas, el modelo fue capaz de identificar la mayoría de las instancias positivas minimizando los falsos positivos, lo que se reflejó en la curva que se inclinó hacia la esquina superior derecha.

El caso ideal en que la precisión y la recuperación alcanzaron su valor máximo estuvo representado por el punto óptimo de PF1, que se encontró en las coordenadas [1,1], donde el modelo clasificó correctamente las instancias positivas sin falsos positivos.

En última instancia, el AUC fue igual a 1, esta métrica cuantificó el rendimiento global del modelo, donde un AUC de 1 indicó que el modelo pudo distinguir entre clases positivas y negativas con un grado de precisión aceptable.

7. Discusión

En respuesta a la pregunta de investigación sobre la probabilidad de acierto de un modelo híbrido de predicción de fallas para los aerogeneradores de la CEV utilizando IA y datos del sistema SCADA, se implementó un modelo híbrido utilizando la técnica de combinación Blending, la cual usa de base a los algoritmos ANN y RF, los cuales son los algoritmos más destacados entre los experimentos realizados con las técnicas de sobremuestreo, de esta forma, la evaluación de este modelo mostró una probabilidad de acierto sobresaliente en las diversas evaluaciones.

En relación a los algoritmos más empleados descubiertos en la RSL, ANN y SVM ocupan este lugar, donde estudios como [121], [122], [123], los describen como algoritmos competentes por su adaptabilidad en diferentes contextos de análisis de fallas en aerogeneradores, demostrando ser versátiles y efectivos en la predicción de fallas, siendo reconocidos en investigaciones como la de [124], que además de estos algoritmos, identifica a SVR, GPR, DBN y K-NN como modelos prominentes. Por otra parte, aunque SVM no formo parte de uno de los algoritmos base para el modelo híbrido, autores como [125] lo señalan como un algoritmo que resalta por su ventaja en el manejo de datos de fallas con desequilibrios significativos.

La principal limitación enfrentada fue el desbalance de clases en los datos SCADA de la CEV, donde la clase mayoritaria superaba significativamente a la minoritaria, el procesamiento de datos y las técnicas de sobremuestreo equilibraron este desbalance; en contraste autores como [121], [122], [123], [125], cuentan un conjunto de datos considerablemente grande, pero a pesar de esta limitación el modelo empleado exhibió resultados notables en medidas como P, R, PF1, S, ES, EX, VC y AUC con resultados cercanos la unidad, mientras que autores como [126] presentan resultados como un AUC de 0.73 con el uso del algoritmo RF.

En función de lo expuesto anteriormente, el tratamiento para mejorar el desequilibrio de clases permitió mejorar y obtener un IR igual 152.41, considerado un desbalance pronunciado, pero no excesivo ni preocupante, en comparación estudios como [108], que examina diferentes conjuntos de datos, conjuntos denominados como “breast-0” y “pima” cuentan con un IR de 1.7 y 1.9 respectivamente, lo que indica poco o ningún desbalance y conjuntos como “ecoli-om” y “abalone-12” cuentan con un IR de 15.8 y 14.7 respectivamente, considerándose un desequilibrio considerado moderado, estos resultados son atribuibles en parte a una amplia recolección de datos.

A pesar del grado de desproporción entre clases manejado, el modelo híbrido Blending-ANN-RF demostró resiliencia, logrando resultados consistentes en las diversas evaluaciones realizadas, lo cual indicó su destreza frente a la clase minoritaria que fue sustancialmente menos representada, por lo que la reducción del dataset mediante el intercambio de datos de fallas entre aerogeneradores y la reducción de la clase mayoritaria por ventana de tiempo, fueron técnicas clave que contribuyeron a tratar el desequilibrio y obtener mejores resultados, mismos que indicaron la posibilidad de sobreajuste, por lo que estrategias como la validación cruzada y el ajuste de hiperparámetros se implementaron para mitigar el riesgo y mejorar la capacidad de generalización del modelo a nuevas instancias, en síntesis los resultados de las diferentes evaluaciones realizadas reflejan valores apropiados y proporcionan indicadores de precisión acorde al conjunto de datos SCADA de la CEV.

Para comprender el nivel de diferencia general entre las técnicas de combinación, el test de Wilcoxon mostró que no había diferencias estadísticamente significativas, lo que enfatizó la estabilidad de los modelos, sumado a eso con el fin de indagar con mayor detalle la capacidad del modelo híbrido seleccionado, el análisis de la predicción de fallas y probabilidad de aciertos reveló que el modelo Blending-ANN-RF contaba con capacidades remarcables. En adición a su elección, se consideró factores como la complejidad del modelo, la interpretabilidad y la dificultad de aplicación, considerándose decisivo para un equilibrio entre el tiempo de procesamiento, los recursos disponibles, la comprensión de las decisiones y la precisión del modelo

El análisis integral del presente trabajo permitió identificar cierta limitación que podrían influir en la generalización de los resultados a un contexto más amplio y sugirió posibles mejoras para investigaciones futuras, donde a pesar de los esfuerzos metodológicos y analíticos, se reconocen aspectos que podrían impactar la aplicabilidad de estos hallazgos en escenarios más diversos. En congruencia, una de las principales limitaciones radicó en la falta de validación externa en condiciones operativas reales, si bien se emplearon técnicas como la validación cruzada, la ausencia de validaciones adicionales en entornos operativos reales pudo reducir la aplicabilidad directa de los resultados a situaciones prácticas; por lo tanto, considerar y abordar esta limitación podría contribuir significativamente a fortalecer la aplicabilidad y la generalización de los resultados obtenidos a una población más amplia de escenarios operativos en el campo de la predicción de fallas en aerogeneradores mediante modelos híbridos.

Finalmente, en relación a la pregunta de investigación, ¿Qué probabilidad de acierto tendrá un modelo híbrido de predicción de fallas para los aerogeneradores de la Central Eólica Villonaco (CEV) utilizando inteligencia artificial y datos del sistema SCADA?, el presente trabajo identifico al modelo Blending-ANN-RF con resultados cercanos a la unidad en las métricas P, R, PF1, S, ES, EX, AUC, VC y pasó el test de Wilcoxon. Junto a esto, evaluaciones en la predicción de fallas, como la comparativa entre la predicción y el valor real y curva de calibración permitieron observar la capacidad del modelo para hacer una distinción apropiada entre clases, asimismo la probabilidad de aciertos reveló en la validación cruzada que el modelo alcanza una precisión máxima en la primera iteración y la curva de precisión-recuperación identificó un puntaje de 1.00 en PF1 el cual es un valor que refleja el equilibrio entre P y R.

8. Conclusiones

8.1. Generales

- Se logró responder satisfactoriamente a la pregunta de investigación '¿Qué probabilidad de acierto tiene un modelo híbrido de predicción de fallas para los aerogeneradores de la Central Eólica Villonaco utilizando inteligencia artificial y datos del sistema SCADA?' a través del desarrollo y la implementación de un modelo híbrido. Este modelo integró algoritmos como ANN y RF, a través de la técnica de combinación Blending, la cual destacó particularmente demostrando una eficiencia notable en términos de tiempo de entrenamiento, y alcanzando un rendimiento sobresaliente en la validación cruzada, lo que indicó un alto grado de exactitud en la predicción de fallas. Además, los resultados del test de Wilcoxon confirmaron la ausencia de diferencias estadísticas significativas en su rendimiento en comparación con sus algoritmos base, lo que subraya la efectividad de Blending en este contexto. Estos resultados destacan no solo la viabilidad del modelo híbrido en el contexto específico de la CEV, sino también su potencial aplicabilidad en escenarios similares de predicción de fallas en aerogeneradores.
- La RSL reveló que SVM y ANN fueron los algoritmos más aplicados para predecir fallas en aerogeneradores alcanzando el 30.56% y 27.78%, respectivamente, mientras que los componentes del aerogenerador más estudiados fueron el aerogenerador en sí mismo con el 50% de apariciones y la caja de cambios con el 25%, dichos resultados reflejan el enfoque y aceptación en el campo de estudio de los artículos analizados
- La investigación resaltó los artículos más influyentes en el campo de predicción de fallas en aerogeneradores con IA y datos SCADA como lo indica con [97] y [104] que lideraron con un 57.41% y 42.59% de citaciones respectivamente llegando a ser estudios muy cruciales por sus innovadoras metodologías y su impacto en la aplicación efectiva de modelos híbridos que subrayan la necesidad de continuar la investigación en este entorno para mejorar la precisión y confiabilidad en la predicción de fallas.
- Se analizó seis algoritmos distintos (ANN, DNN, SVN, KNN, RF, ET) en combinación con cinco técnicas de sobremuestreo, las cuales son SMOTE, ADASYN, SMOTE-ENN, SMOTE-Tomek, y Borderline-SMOTE, cuyos

resultados arrojaron que ANN y RF sobresalen en las métricas P, R y FP1 en los cinco experimentos realizados donde cada uno destacó los dos mejores algoritmos.

- Se evaluaron seis técnicas de combinación para los algoritmos ANN y RF con el objetivo de estructurar el modelo híbrido más adecuado centrado en las métricas P, R, PF1, S, ES, ES, EX, AUC, VC, TE y el test de Wilcoxon donde todas las técnicas superaron las valoraciones que permitieron identificar finalmente al modelo híbrido Blending-ANN-RF como un modelo competente, asimismo, se probaron sus capacidades en la predicción de fallas y probabilidad de aciertos donde se registró una capacidad adecuada del mismo.

8.2. Técnicas

- Google Colab Pro resultó esencial para potenciar la velocidad y eficiencia en el entrenamiento y procesamiento de los algoritmos(ANN, DNN, SVM, KNN, ET, RF) y técnicas de combinación (Bagging, Blending, Boosting, Stacking, Voting, Weighted Average Ensemble) que gracias a su provisión de recursos computacionales avanzados como GPUs y TPUs aplicados para manejar grandes conjuntos de datos de forma eficiente y acelerar significativamente el entrenamiento de modelos logran superar las capacidades de los entornos locales y un ejemplo destacado fue el modelo híbrido Blending-ANN-RF que logró un periodo de entrenamiento notablemente reducido en contraste a sus homónimos que de igual manera alcanzó máximas puntuaciones en métricas como P, R y PF1 cuyos resultados fueron corroborados por análisis detallados como el test de Wilcoxon, análisis de predicción de fallas y probabilidades de acierto.
- La selección de herramientas tecnológicas, como *Python 3.10.12*, *Sklearn 1.2.2* y *Keras 2.15.0* puede ser vital para quienes busquen replicar los resultados de este estudio, junto con un aspecto fundamental del proceso de desarrollo: la estandarización en la división de los datos, asignando el 30% para pruebas, 70% para entrenamiento y utilizando una semilla de 42 para asegurar la consistencia en los resultados.
- El ajuste fino de los algoritmos conjuntamente con la validación cruzada y la optimización de hiperparámetros jugaron un rol fundamental en el rendimiento, de igual manera, el procesamiento de datos fue importante donde técnicas como la reducción de datos por ventana de tiempo e intercambio de datos resultaron

ser vitales para reducir la tasa de desbalanceo de las clases, aplicadas al conjunto de 157,929 registros SCADA de la CEV que redujeron significativamente la tasa de desbalanceo de clases pasando de un IR de 5445.827 a uno más manejable de 152.413 demostrando su eficacia ante la desproporción de datos.

- El desequilibrio de clases fue un desafío frecuente en el aprendizaje automático que logró sesgar los resultados del modelo hacia las clases predominantes, así la librería *Imbalanced-learn (0.10.1)* al implementar técnicas de sobremuestreo como SMOTE, ADASYN, SMOTE-ENN, SMOTE-Tomek y Borderline-SMOTE, demostró ser primordial en equilibrar la distribución de clases y ajustando adecuadamente la proporción de las clases minoritarias llegó a facilitar la reducción del sesgo hacia la clase mayoritaria con la creación de muestras sintéticas de la clase minoritaria para optimizar las capacidades predictivas y generalización de los algoritmos.
- La librería *sklearn 1.2.2* fue determinante para la mayoría de las tareas implementadas en el campo de la IA como el consumo de diferentes algoritmos, técnicas de preprocesamiento, evaluación a la matriz de confusión o validación cruzada, adicionalmente librerías como *scipy 1.11.4* que de igual manera fueron imperativas al permitir efectuar el test de Wilcoxon para evaluar los distintos modelos híbridos, no obstante, a pesar de la existencia de alternativas como *statsmodels*, *scipy* se destacó por su documentación, eficiencia y aceptación en el área del análisis científico permitiendo efectuar comparaciones estadísticas confiables entre las técnicas de combinación, considerándose como una decisión que impactó de forma predominante la calidad y confiabilidad del análisis estadístico, asegurando la integridad y validez de los resultados.

9. Recomendaciones

Se ofrecen algunas recomendaciones y orientaciones en virtud de continuar trabajando en este campo sobre la base de los resultados del actual TIC:

Los resultados obtenidos en este TIC sugieren varias líneas de investigación y experimentación para trabajos futuros como puede ser la exploración de los modelos y técnicas que la RSL identifica como más usados en la predicción de fallas en aerogeneradores, donde este punto de partida abre una ventana de oportunidades para efectuar un proceso experimental con los algoritmos expuestos logrando de esta manera adaptarlos y potenciarlos en contextos similares.

La RSL hace énfasis en la existencia de diversos enfoques netamente en modelos híbridos que constituyen una base sólida para la experimentación y desarrollo de soluciones prometedoras, al mismo tiempo que se identifican los elementos principales sobre los cuales se enfoca la predicción de fallas y con dicha información los investigadores pueden guiar la selección y la estructuración de algoritmos que ya han sido utilizados con éxito optimizando así el abordaje de la predicción de fallas en contextos específicos.

Los resultados obtenidos en el segundo objetivo del presente TIC destacan la efectividad de técnicas de sobremuestreo como SMOTE-Tomek para ANN y ADASYN para RF que indican que la aplicación de dichas técnicas resultaría altamente beneficiosa en problemas similares abriendo la puerta a resultados prometedores en investigaciones futuras, por lo tanto, se aconseja la exploración de estas técnicas en contextos análogos para validar y expandir el conocimiento actual.

Con la finalidad de determinar la aplicabilidad de las técnicas de sobremuestreo o submuestreo en contextos particulares y establecer cuáles son las más efectivas en diversos escenarios se considera elemental que los futuros estudios efectúen un análisis exhaustivo del desequilibrio de clases donde una selección informada de estas técnicas puede optimizar de gran manera la eficacia de los modelos predictivos.

En cuanto al entorno de desarrollo es imprescindible evaluar alternativas de plataformas en la nube como Google Colab, AWS o Azure que precisan de factores como la disponibilidad de recursos y la facilidad de integración con técnicas de aprendizaje automático, los cuales permiten determinar qué entorno es más adecuado y al realizar esta elección estratégica será posible maximizar las capacidades tecnológicas disponibles logrando un desarrollo eficiente de los modelos.

Se sugiere investigar estrategias avanzadas en conjuntos de datos grandes como el intercambio de fallas entre aerogeneradores y la reducción de la clase mayoritaria por ventana de tiempo, ejecutando en práctica estas tácticas, las métricas de rendimiento de los algoritmos podrían mejorar considerablemente.

Finalmente, se sugiere ampliar el estudio del presente TIC para evaluar los modelos y técnicas probadas para la predicción de fallas en diferentes sistemas de maquinaria como maquinaria industrial y turbinas hidroeléctricas con la finalidad de validar la robustez y versatilidad de los modelos examinados, de igual manera, se aconseja investigar su infraestructura y enfoque metodológico para la detección de anomalías en contextos mucho más amplios, como pueden ser en procesos industriales y redes eléctricas, lo que ayudaría a integrar soluciones para la supervisión y el mantenimiento predictivo en toda una serie de industrias.

10. Bibliografía

[1] G. Erazo and J. Maldonado, "Análisis de la producción de energía de la Central Eólica Villonaco," Universidad Técnica Particular de Loja, Loja, 2017. Accessed: Jun. 15, 2023. [Online]. Available: https://dspace.utpl.edu.ec/handle/123456789/17842

[2] Y. Zhao, D. Li, A. Dong, D. Kang, Q. Lv, and L. Shang, "Fault prediction and diagnosis of wind turbine generators using SCADA data," *Energies (Basel)*, vol. 10, no. 8, pp. 1–17, Aug. 2017, doi: 10.3390/en10081210.

[3] E. E. Hernández Parada, "Decisiones de inversión en proyectos de energía renovable no convencional," Universidad de el Salvador, 2016. Accessed: Jun. 15, 2023. [Online]. Available: https://ri.ues.edu.sv/id/eprint/12888/

[4] L. Dan and D. Siobhan, "La energía renovable como resguardo frente a la fluctuación de precios de los combustibles: aprovechamiento de beneficios," Sep. 2008. Accessed: Jun. 15, 2023. [Online]. Available: https://policycommons.net/artifacts/1185357/renewable-energy-as-a-hedge-against-fuel-price-fluctuation/1738480/

[5] L. A. Bird, K. S. Cory, and B. G. Swezey, "Renewable Energy Price-Stability Benefits in Utility Green Power Programs," pp. 1–36, Aug. 2008, doi: 10.2172/936506.

[6] M. Bobadilla, "Desarrollo de la Energía Eólica en Panamá," *El Tecnológico*, vol. 27, no. 1, pp. 1–2, Apr. 2017, Accessed: Jun. 16, 2023. [Online]. Available: https://revistas.utp.ac.pa/index.php/el-tecnologico/article/view/1284

[7] S. Medina and A. K. Venegas, "Energías renovables un futuro optimo para Colombia," *Punto de Vista*, vol. 9, no. 13, pp. 1–16, 2018, Accessed: Jun. 16, 2023. [Online]. Available: https://dialnet.unirioja.es/servlet/articulo?codigo=6540491

[8] H. Habibi, I. Howard, and S. Simani, "Reliability improvement of wind turbine power generation using model-based fault detection and fault tolerant control: A review," *Renew Energy*, vol. 135, pp. 877–896, May 2019, doi: 10.1016/j.renene.2018.12.066.

[9] "GWEC | GLOBAL WIND REPORT 2022," 2022. Accessed: Jun. 16, 2023. [Online]. Available: https://gwec.net/wp-content/uploads/2022/03/GWEC-GLOBAL-WIND-REPORT-2022.pdf

[10] A. Vera Vera, N. Balderramo Vélez, G. Pico Mera, E. Rodríguez Indarte, and M. L. Dávila Cedeño, "Realidad actual del sector eléctrico ecuatoriano," *Revista de Investigaciones en Energía, Medio Ambiente y Tecnología: RIEMAT ISSN: 2588-0721*, vol. 4, no. 1, pp. 1–5, Jul. 2019, doi: 10.33936/riemat.v4i1.1939.

[11] J. Macuy and M. A. Sotomayor Páez, "Proyecto de factibilidad para la implementación de una empresa productora y comercializadora de paneles fotovoltaicos en Ecuador," Trabajo de Titulación que se presenta como requisito para el título de Ingeniería en Ciencias en Empresariales con Especialización en Dirección y Planeación Comercial , Universidad de

Especialidades Espíritu Santo UEES, Guayaquil, 2011. Accessed: Jun. 16, 2023. [Online]. Available: http://repositorio.uees.edu.ec/123456789/698

[12] M. Ayala and G. Riba, "Confronting Urban Electricity Demand with Wind Energy Supply: Case Study in Ecuador," *Preprints.org*, pp. 1–12, Jul. 2019, doi: https://www.preprints.org/manuscript/201907.0098/v1.

[13] O. Cabeza-Gras and V. Jaramillo-García, "Wind energy system in ambocas-ecuador: Distributed generation and energy quality," *Renewable Energy and Power Quality Journal*, vol. 19, pp. 609–613, Sep. 2021, doi: 10.24084/repqj19.361.

[14] Á. G. Párraga Palacios *et al.*, "Producción de energía eólica en Ecuador," *Ciencia Digital*, vol. 3, no. 3, pp. 22–32, Jul. 2019, doi: 10.33262/cienciadigital.v3i3.610.

[15] D. F. Galarza Anguisaca and D. I. Román Puga, "Metodología para el análisis factibilidad de generación de energía eléctrica a partir de energía eólica: caso de estudio en la parroquia Yangana, provincia de Loja," Universidad Politécnica Salesiana del Ecuador, Cuenca, 2021. Accessed: Jun. 16, 2023. [Online]. Available: http://dspace.ups.edu.ec/handle/123456789/20022

[16] S. L. Parra Chalán, "Factores que inciden en el crecimiento poblacional en los años 2016 al 2021," Instituto Tecnológico Sudamericano, Loja, 2022. Accessed: Jun. 16, 2023. [Online]. Available: http://dspace.tecnologicosudamericano.edu.ec/jspui/handle/123456789/476

[17] M. Ayala, J. Maldonado, E. Paccha, and C. Riba, "Wind power resource assessment in complex terrain: Villonaco case-study using computational fluid dynamics analysis," in *Energy Procedia*, Elsevier Ltd, Feb. 2017, pp. 41–48. doi: 10.1016/j.egypro.2016.12.127.

[18] J. G. Castillo Armijos, "Análisis técnico y económico de la producción de energía en el Parque Eólico Villonaco," Universidad Técnica Particular de Loja, Loja , 2016. Accessed: Jun. 16, 2023. [Online]. Available: https://dspace.utpl.edu.ec/handle/123456789/14551

[19] J. F. Manwell, J. G. McGowan, and A. L. Rogers, *Wind Energy Explained: Theory, Design and Application*, Second Edition. Wiley, 2010. Accessed: Jun. 16, 2023. [Online]. Available: http://ee.tlu.edu.vn/Portals/0/2018/NLG/Sach_Tieng_Anh.pdf

[20] J. C. Castañeda Ramírez, "Energía Eólica en la Universidad Michoacana de San Nicolás de Hidalgo," *Ciencia Nicolaita*, no. 73, pp. 1–18, May 2018, doi: https://doi.org/10.35830/cn.vi73.368.

[21] S. Santiago Pradillo, "Situación actual y perspectivas de la energía eólica marina en Europa," Proyecto Fin de Carrera/Grado, Escuela Técnica Superior De Ingenieros De Minas y Energía, 2018. Accessed: Jun. 16, 2023. [Online]. Available: https://oa.upm.es/52964/

[22] A. Rodríguez Penin, *Sistemas SCADA*, Tercera Edición. Alfaomega Grupo Editor, 2011. Accessed: Jun. 17, 2023. [Online]. Available: https://instipp.edu.ec/Libreria/libro/Sistemas%20SCADA%20(%20PDFDrive%20).pdf

[23] J. Tautz-Weinert and S. J. Watson, “Using SCADA data for wind turbine condition monitoring - A review,” *IET Renewable Power Generation*, vol. 11, no. 4, pp. 382–394, Mar. 2017, doi: 10.1049/iet-rpg.2016.0248.

[24] M. E. García Villacís, “Sistema SCADA para el proceso de paletizado L4 de envases de cristal en la Empresa Cristalería del Ecuador S. A. Cridesa de Guayaquil,” Universidad Técnica de Ambato, Ambato, 2014. Accessed: Jun. 17, 2023. [Online]. Available: https://repositorio.uta.edu.ec/handle/123456789/8550

[25] R. Banik, P. Das, S. Ray, and A. Biswas, “Prediction of electrical energy consumption based on machine learning technique,” *Electrical Engineering*, vol. 103, no. 2, pp. 909–920, Apr. 2021, doi: 10.1007/S00202-020-01126-Z/METRICS.

[26] S. Uribe, “Futuro de la inteligencia artificial en Odontología,” *Odontología Sanmarquina*, vol. 24, no. 3, pp. 305–307, Jul. 2021, doi: 10.15381/os.v24i3.20726.

[27] A. Loaiza-Bonilla, “La inteligencia artificial en oncología: contexto actual y una visión hacia la próxima década,” *Medicina (B Aires)*, vol. 43, no. 4, pp. 1–8, Jan. 2022, doi: https://doi.org/10.56050/01205498.1642.

[28] J. S. Bonilla Segovia, F. A. Dávila Rojas, and M. W. Villa Quishpe, “Estudio del uso de técnicas de inteligencia artificial aplicadas para análisis de suelos para el sector agrícola,” *Recimundo*, vol. 5, no. 1, pp. 4–19, Jan. 2021, doi: 10.26820/recimundo/5.(1).enero.2021.4-19.

[29] L. E. Aparicio Pico, P. Devia Lozano, and O. J. Amaya Marroquin, “Aplicación de Deep Learning para la identificación de defectos superficiales utilizados en control de calidad de manufactura y producción industrial: Una revisión de la literatura,” *Ingeniería*, vol. 28, no. 1, pp. 1–20, Nov. 2022, doi: 10.14483/23448393.18934.

[30] A. Hernández Yeja, J. De la Rosa Pasteur, and O. Rodríquez Huice, “Aplicación de técnicas de Inteligencia Artificial en la Seguridad Informática: un estudio,” *Inteligencia Artificial*, vol. 51, pp. 1–8, 2013, [Online]. Available: http://journal.iberamia.org/

[31] A. Moreno Ribas *et al.*, *Aprendizaje automático*, Edicions UPC. Politext , 1994. Accessed: Jun. 17, 2023. [Online]. Available: https://upcommons.upc.edu/handle/2099.3/36157

[32] H. Chen, H. Liu, X. Chu, Q. Liu, and D. Xue, “Anomaly detection and critical SCADA parameters identification for wind turbines based on LSTM-AE neural network,” *Renew Energy*, vol. 172, pp. 829–840, Jul. 2021, doi: 10.1016/j.renene.2021.03.078.

[33] K. Dhibi, M. Mansouri, K. Bouzrara, H. Nounou, and M. Nounou, “Reduced neural network based ensemble approach for fault detection and diagnosis of wind energy converter systems,” *Renew Energy*, vol. 194, pp. 778–787, 2022, doi: 10.1016/j.renene.2022.05.082.

[34] S. R. Moreno, L. S. Coelho, H. V. H. Ayala, and V. C. Mariani, “Wind turbines anomaly detection based on power curves and ensemble learning,” *IET Renewable Power Generation*, vol. 14, no. 19, pp. 4086–4093, 2020, doi: 10.1049/iet-rpg.2020.0224.

[35] L. Breiman, “Bagging Predictors,” *Mach Learn*, vol. 24, pp. 123–140, 1996, doi: https://doi.org/10.1007/BF00058655.

[36] S. Chatterjee and Y. C. Byun, “Highly imbalanced fault classification of wind turbines using data resampling and hybrid ensemble method approach,” *Eng Appl Artif Intell*, vol. 126, pp. 1–14, Nov. 2023, doi: 10.1016/j.engappai.2023.107104.

[37] M. A. Ganaie, M. Hu, A. K. Malik, M. Tanveer, and P. N. Suganthan, “Ensemble deep learning: A review,” *Eng Appl Artif Intell*, vol. 115, pp. 1–47, Oct. 2022, doi: 10.1016/j.engappai.2022.105151.

[38] D. H. Wolpert, “Stacked generalization,” *Neural Networks*, vol. 5, no. 2, pp. 1–57, Jan. 1992, doi: 10.1016/S0893-6080(05)80023-1.

[39] Z. Tang *et al.*, “Fault Diagnosis of Wind Turbine Generators Based on Stacking Integration Algorithm and Adaptive Threshold,” *Sensors*, vol. 23, no. 13, Jul. 2023, doi: 10.3390/s23136198.

[40] A. Kolios, M. Richmond, S. Koukoura, and B. Yeter, “Effect of weather forecast uncertainty on offshore wind farm availability assessment,” *Ocean Engineering*, vol. 285, pp. 1–14, Oct. 2023, doi: 10.1016/j.oceaneng.2023.115265.

[41] Z. Xu *et al.*, “Fault diagnosis of wind turbine bearing using a multi-scale convolutional neural network with bidirectional long short term memory and weighted majority voting for multi-sensors,” *Renew Energy*, vol. 182, pp. 615–626, Jan. 2022, doi: 10.1016/j.renene.2021.10.024.

[42] J. Huertas-Tato, R. Aler, I. M. Galván, F. J. Rodríguez-Benítez, C. Arbizu-Barrena, and D. Pozo-Vázquez, “A short-term solar radiation forecasting system for the Iberian Peninsula. Part 2: Model blending approaches based on machine learning,” *Solar Energy*, vol. 195, pp. 685–696, Jan. 2020, doi: 10.1016/j.solener.2019.11.091.

[43] R. Chen, J. Zhu, X. Hu, H. Wu, X. Xu, and X. Han, “Fault diagnosis method of rolling bearing based on multiple classifier ensemble of the weighted and balanced distribution adaptation under limited sample imbalance,” *ISA Trans*, vol. 114, pp. 434–443, Aug. 2021, doi: 10.1016/j.isatra.2020.12.034.

[44] W. Lee, C. H. Jun, and J. S. Lee, “Instance categorization by support vector machines to adjust weights in AdaBoost for imbalanced data classification,” *Inf Sci (N Y)*, vol. 381, pp. 92–103, Mar. 2017, doi: 10.1016/j.ins.2016.11.014.

[45] R. M. Valdovinos Rosas, “Técnicas de Submuestreo, Toma de Decisiones y Análisis de Diversidad en Aprendizaje Supervisado con Sistemas Múltiples de Clasificación,” Universitat Jaume I, Castelló de la Plana, 2006.

[46] D.-S. Huang, X.-P. Zhang, and G.-B. Huang, “Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning,” in *Lecture Notes in Computer Science*, Springer-

Verlag Berlin Heidelberg., 2005, pp. 1–1127. Accessed: Jun. 17, 2023. [Online]. Available: https://link.springer.com/chapter/10.1007/11538059_91

[47] H. Zhang, R. Wang, R. Pan, and H. Pan, "Imbalanced fault diagnosis of rolling bearing using enhanced generative adversarial networks," *IEEE Access*, vol. 8, pp. 185950–185963, 2020, doi: 10.1109/ACCESS.2020.3030058.

[48] T. Sasada, Z. Liu, T. Baba, K. Hatano, and Y. Kimura, "A resampling method for imbalanced datasets considering noise and overlap," in *Procedia Computer Science*, Elsevier B.V., 2020, pp. 420–429. doi: 10.1016/j.procs.2020.08.043.

[49] H. He, Y. Bai, E. García, and S. Li, "ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning," Hong Kong: IEEE, Aug. 2008, pp. 1–7. doi: 10.1109/IJCNN.2008.4633969.

[50] H. Haibo, B. Yang, A. Edwardo, E. Garcia, and L. Shutado, "ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning," *IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pp. 1322–8, Aug. 2008, doi: http://dx.doi.org/10.1109/IJCNN.2008.4633969.

[51] Z. Qing, Q. Zeng, H. Wang, Y. Liu, T. Xiong, and S. Zhang, "ADASYN-LOF Algorithm for Imbalanced Tornado Samples," *Atmosphere (Basel)*, vol. 13, no. 4, Apr. 2022, doi: 10.3390/atmos13040544.

[52] J. Liu, H. Yang, J. He, Z. Sheng, and S. Chen, "Unbalanced Fault Diagnosis Based on an Invariant Temporal-Spatial Attention Fusion Network," *Comput Intell Neurosci*, vol. 2022, pp. 1–15, 2022, doi: 10.1155/2022/1875011.

[53] W. Chen, H. Zhou, L. Cheng, and M. Xia, "Wind Turbine Blade Icing Diagnosis Using Convolutional LSTM-GRU With Improved African Vultures Optimization," *IEEE Open Journal of Instrumentation and Measurement*, vol. 1, pp. 1–9, Dec. 2022, doi: 10.1109/ojim.2022.3217850.

[54] S. Sun, W. Hu, Y. Liu, T. Wang, and F. Chu, "Matching contrastive learning: An effective and intelligent method for wind turbine fault diagnosis with imbalanced SCADA data," *Expert Syst Appl*, vol. 223, pp. 1–12, Aug. 2023, doi: 10.1016/j.eswa.2023.119891.

[55] C. Cai *et al.*, "Review of Data-Driven Approaches for Wind Turbine Blade Icing Detection," *Sustainability*, vol. 15, no. 2, p. 1617, Jan. 2023, doi: 10.3390/su15021617.

[56] Q. H. Doan, D.-K. Thai, and N. L. Tran, "A hybrid model for predicting missile impact damages based on k-nearest neighbors and Bayesian optimization," *Journal of Science and Technology in Civil Engineering (STCE) - NUCE*, vol. 14, no. 3, pp. 1–14, Aug. 2020, doi: 10.31814/stce.nuce2020-14(3)-01.

[57] H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning," in *Advances in Intelligent Computing. ICIC*, H. Han,

W.-Y. Wang, and B.-H. Mao, Eds., Berlin, Heidelberg: Springer, 2005, pp. 878–887. doi: https://doi.org/10.1007/11538059_91.

[58] S. Maldonado, J. López, and C. Vairetti, "An alternative SMOTE oversampling strategy for high-dimensional datasets," *Applied Soft Computing Journal*, vol. 76, pp. 380–389, Mar. 2019, doi: 10.1016/j.asoc.2018.12.024.

[59] D. P. Kroese, Z. I. Botev, T. Taimre, and R. Vaisman, *Data Science and Machine Learning Mathematical and Statistical Methods*. CRC Press, 2022. Accessed: Jul. 31, 2023. [Online]. Available: https://books.google.com.ec/books?id=F7zADwAAQBAJ

[60] "Capítulo 29 Validación cruzada | Introducción a la ciencia de datos." Accessed: Apr. 16, 2024. [Online]. Available: https://rafalab.dfci.harvard.edu/dslibro/cross-validation.html

[61] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning Data Mining, Inference, and Prediction*, Second Edition. Springer, 2009. Accessed: Jul. 31, 2023. [Online]. Available: https://link.springer.com/book/10.1007/978-0-387-21606-5

[62] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit Lett*, vol. 27, no. 8, pp. 861–874, Jun. 2006, doi: 10.1016/j.patrec.2005.10.010.

[63] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf Process Manag*, vol. 45, no. 4, pp. 427–437, Jul. 2009, doi: 10.1016/j.ipm.2009.03.002.

[64] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognit*, vol. 30, no. 7, pp. 1145–1159, Jul. 1997, doi: 10.1016/S0031-3203(96)00142-2.

[65] J. Brownlee, *Imbalanced Classification with Python Choose Better Metrics, Balance Skewed Classes, and Apply Cost-Sensitive Learning*, V1.2. Machine Learning Mastery, 2020.

[66] F. Wilcoxon, "Individual Comparisons by Ranking Methods," vol. 1, no. 6, pp. 80–83, 1945, doi: https://doi.org/10.2307/3001968.

[67] J. Amat Rodrigo, "cienciadedatos.net." Zenodo, Oct. 2023. doi: 10.5281/zenodo.10006330.

[68] Z. Huang *et al.*, "Machine learning-based survival prediction nomogram for postoperative parotid mucoepidermoid carcinoma," *Sci Rep*, vol. 14, no. 1, Dec. 2024, doi: 10.1038/s41598-024-58329-8.

[69] D. C. Piñol and E. J. M. Reyes, "Automatic Handgun Detection in X-ray Images using Bag of Words Model with Selective Search," Mar. 2019, doi: https://doi.org/10.48550/arXiv.1903.01322.

[70] J. Davis and M. Goadrich, "The relationship between Precision-Recall and ROC curves," in *Proceedings of the 23rd International Conference on Machine Learning*, in ICML '06. New York, NY, USA: Association for Computing Machinery, 2006, pp. 233–240. doi: 10.1145/1143844.1143874.

[71] B. Ozenne, F. Subtil, and D. Maucort-Boulch, “The precision-recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases,” *J Clin Epidemiol*, vol. 68, no. 8, pp. 855–859, Aug. 2015, doi: 10.1016/j.jclinepi.2015.02.010.

[72] J. Miao and W. Zhu, “Precision–recall curve (PRC) classification trees,” *Evol Intell*, vol. 15, no. 3, pp. 1545–1569, Sep. 2022, doi: 10.1007/s12065-021-00565-2.

[73] S. Barber *et al.*, “Best Practice Data Sharing Guidelines for Wind Turbine Fault Detection Model Evaluation,” *Energies (Basel)*, vol. 16, no. 8, Apr. 2023, doi: 10.3390/en16083567.

[74] C. Correa-jullian, S. Cofre-martel, G. S. Martin, E. L. Droguett, G. de N. P. Leite, and A. Costa, “Exploring Quantum Machine Learning and Feature Reduction Techniques for Wind Turbine Pitch Fault Detection,” *Energies (Basel)*, vol. 15, no. 8, Apr. 2022, doi: 10.3390/en15082792.

[75] R. Islam, S. Tanweer, T. Nafis, I. Hussain, and S. N. Qurashi, “PROGNOSIS OF CARDIOVASCULAR DISEASE USING PRINCIPAL COMPONENT ANALYSIS AND SUPPORT VECTOR MACHINE CLASSIFICATION ALGORITHMS IN THE R STUDIO ENVIRONMENT,” *J Theor Appl Inf Technol*, vol. 102, no. 6, pp. 2287–2296, Mar. 2024.

[76] M. M. Jibril *et al.*, “New random intelligent chemometric techniques for sustainable geopolymer concrete: low-energy and carbon-footprint initiatives,” *Asian Journal of Civil Engineering*, vol. 25, no. 2, pp. 2287–2305, Feb. 2024, doi: 10.1007/s42107-023-00908-7.

[77] D. Jia, L. Wang, N. Valencia, J. Bhimani, B. Sheng, and N. Mi, “Learning-Based Dynamic Memory Allocation Schemes for Apache Spark Data Processing,” *IEEE Transactions on Cloud Computing*, vol. 12, no. 1, pp. 13–25, Jan. 2024, doi: 10.1109/TCC.2023.3329129.

[78] Bostjan Kaluza, *Machine Learning in Java*. UK: Packt Publishing Ltd, 2016.

[79] H. Yu and B. M. Wilamowski, “C++ implementation of neural networks trainer,” *Proceedings - 2009 International Conference on Intelligent Engineering Systems, INES 2009*, pp. 257–262, 2009, doi: 10.1109/INES.2009.4924772.

[80] K. Kolodiazhnyi, *Hands-on Machine Learning with C++: Build, train, and deploy end-to-end machine learning and deep learning pipelines*. Packt Publishing Ltd, 2020.

[81] M. Aly, T. Kasem, and R. Shalaby, “Grid Fault Detection of DFIG Wind Farms using a High-Fidelity Model and Machine Learning,” *NILES 2022 - 4th Novel Intelligent and Leading Emerging Sciences Conference, Proceedings*, pp. 270–275, Jan. 2022, doi: 10.1109/NILES56402.2022.9942421.

[82] A. A. Jaber and R. Bicker, “A simulation of non-stationary signal analysis using wavelet transform based on LabVIEW and Matlab,” in *Proceedings - UKSim-AMSS 8th European Modelling Symposium on Computer Modelling and Simulation, EMS 2014*, Institute of Electrical and Electronics Engineers Inc., 2014, pp. 138–144. doi: 10.1109/EMS.2014.38.

[83] J. Tautz-Weinert and S. J. Watson, “Using SCADA data for wind turbine condition monitoring - A review,” *IET Renewable Power Generation*, vol. 11, no. 4. Institution of Engineering and Technology, pp. 382–394, Mar. 15, 2017. doi: 10.1049/iet-rpg.2016.0248.

[84] J. Chatterjee and N. Dethlefs, “Scientometric review of artificial intelligence for operations & maintenance of wind turbines: The past, present and future,” *Renewable and Sustainable Energy Reviews*, vol. 144. Elsevier Ltd, Jul. 01, 2021. doi: 10.1016/j.rser.2021.111051.

[85] J. Urmeneta, J. Izquierdo, and U. Leturiondo, “A methodology for performance assessment at system level—Identification of operating regimes and anomaly detection in wind turbines,” *Renew Energy*, vol. 205, pp. 281–292, 2023, doi: 10.1016/j.renene.2023.01.035.

[86] Y. Cui, P. Bangalore, and L. B. Tjernberg, “An anomaly detection approach based on machine learning and scada data for condition monitoring of wind turbines,” in *2018 International Conference on Probabilistic Methods Applied to Power Systems, PMAPS 2018 - Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2018. doi: 10.1109/PMAPS.2018.8440525.

[87] Y. Zhu, C. Zhu, J. Tan, Y. Tan, and L. Rao, “Anomaly detection and condition monitoring of wind turbine gearbox based on LSTM-FS and transfer learning,” *Renew Energy*, vol. 189, pp. 90–103, 2022, doi: 10.1016/j.renene.2022.02.061.

[88] T. Matsui, K. Yamamoto, and J. Ogata, “Anomaly detection for wind turbine damaged due to lightning strike,” *Electric Power Systems Research*, vol. 209, 2022, doi: 10.1016/j.epsr.2022.107918.

[89] L. Xiang, X. Yang, A. Hu, H. Su, and P. Wang, “Condition monitoring and anomaly detection of wind turbine based on cascaded and bidirectional deep learning networks,” *Appl Energy*, vol. 305, 2022, doi: 10.1016/j.apenergy.2021.117925.

[90] M.-A. Lutz *et al.*, “Evaluation of anomaly detection of an autoencoder based on maintenace information and SCADA-data,” *Energies (Basel)*, vol. 13, no. 5, 2020, doi: 10.3390/en13051063.

[91] C. Choe Wei Chang, T. Jian Ding, T. Jian Ping, M. Ariannejad, K. Chia Chao, and S. B. Samdin, “Fault detection and anti-icing technologies in wind energy conversion systems: A review,” *Energy Reports*, vol. 8, pp. 28–33, 2022, doi: 10.1016/j.egyr.2022.10.234.

[92] F. Perez-Sanjines, C. Peeters, T. Verstraeten, J. Antoni, A. Nowé, and J. Helsen, “Fleet-based early fault detection of wind turbine gearboxes using physics-informed deep learning based on cyclic spectral coherence,” *Mech Syst Signal Process*, vol. 185, 2023, doi: 10.1016/j.ymssp.2022.109760.

[93] C. Yang, J. Liu, Y. Zeng, and G. Xie, “Real-time condition monitoring and fault detection of components based on machine-learning reconstruction model,” *Renew Energy*, vol. 133, pp. 433–441, 2019, doi: 10.1016/j.renene.2018.10.062.

[94] A. Amini, J. Kanfoud, and T.-H. Gan, "An Artificial Intelligence Neural Network Predictive Model for Anomaly Detection and Monitoring of Wind Turbines Using SCADA Data," *Applied Artificial Intelligence*, vol. 36, no. 1, 2022, doi: 10.1080/08839514.2022.2034718.

[95] B. Corley, S. Koukoura, J. Carroll, and A. McDonald, "Combination of thermal modelling and machine learning approaches for fault detection in wind turbine gearboxes," *Energies (Basel)*, vol. 14, no. 5, 2021, doi: 10.3390/en14051375.

[96] M. Bach-Andersen, B. Rømer-Odgaard, and O. Winther, "Deep learning for automated drivetrain fault detection," *Wind Energy*, vol. 21, no. 1, pp. 29–41, 2018, [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85037532713&partnerID=40&md5=b01dc51a03c60ac57be2c9b333d3647c

[97] G. Helbing and M. Ritter, "Deep Learning for fault detection in wind turbines," *Renewable and Sustainable Energy Reviews*, vol. 98, pp. 189–198, 2018, doi: 10.1016/j.rser.2018.09.012.

[98] S. Barber *et al.*, "Enabling Co-Innovation for a Successful Digital Transformation in Wind Energy Using a New Digital Ecosystem and a Fault Detection Case Study," *Energies (Basel)*, vol. 15, no. 15, 2022, doi: 10.3390/en15155638.

[99] C. Correa-jullian, S. Cofre-martel, G. S. Martin, E. L. Droguett, G. de N. P. Leite, and A. Costa, "Exploring Quantum Machine Learning and Feature Reduction Techniques for Wind Turbine Pitch Fault Detection," *Energies (Basel)*, vol. 15, no. 8, Apr. 2022, doi: 10.3390/en15082792.

[100] R. K. Pandit and D. Infield, "SCADA-based wind turbine anomaly detection using Gaussian process models for wind turbine condition monitoring purposes," *IET Renewable Power Generation*, vol. 12, no. 11, pp. 1249–1255, 2018, doi: 10.1049/iet-rpg.2018.0156.

[101] T. Hasegawa, J. Ogata, M. Murakawa, and T. Ogawa, "Tandem Connectionist Anomaly Detection: Use of Faulty Vibration Signals in Feature Representation Learning," in *2018 IEEE International Conference on Prognostics and Health Management, ICPHM 2018*, Institute of Electrical and Electronics Engineers Inc., 2018. doi: 10.1109/ICPHM.2018.8448450.

[102] A. Al-Ajmi, Y. Wang, and S. Djurović, "Wind turbine generator controller signals supervised machine learning for shaft misalignment fault detection: A doubly fed induction generator practical case study," *Energies (Basel)*, vol. 14, no. 6, 2021, doi: 10.3390/en14061601.

[103] A. Santolamazza, D. Dadi, and V. Introna, "A data-mining approach for wind turbine fault detection based on scada data analysis using artificial neural networks," *Energies (Basel)*, vol. 14, no. 7, 2021, doi: 10.3390/en14071845.

[104] J. Liu, F. Qu, X. Hong, and H. Zhang, “A Small-Sample Wind Turbine Fault Detection Method With Synthetic Fault Data Using Generative Adversarial Nets,” *IEEE Trans Industr Inform*, vol. 15, no. 7, pp. 3877–3888, 2019, doi: 10.1109/TII.2018.2885365.

[105] A. Turnbull, J. Carroll, and A. McDonald, “Combining SCADA and vibration data into a single anomaly detection model to predict wind turbine component failure,” *Wind Energy*, vol. 24, no. 3, pp. 197–211, 2021, doi: 10.1002/we.2567.

[106] W. Yu, S. Huang, and J. Wang, “Fault Detection Based on a Combined Approach of FA-CP-ELM with Application to Wind Turbine System,” *Journal of Electrical Engineering and Technology*, vol. 16, no. 1, pp. 547–557, 2021, doi: 10.1007/s42835-020-00561-z.

[107] P. Trizoglou, X. Liu, and Z. Lin, “Fault detection by an ensemble framework of Extreme Gradient Boosting (XGBoost) in the operation of offshore wind turbines,” *Renew Energy*, vol. 179, pp. 945–962, 2021, doi: 10.1016/j.renene.2021.07.085.

[108] H. Yi, Q. Jiang, X. Yan, and B. Wang, “Imbalanced Classification Based on Minority Clustering Synthetic Minority Oversampling Technique with Wind Turbine Fault Detection Application,” *IEEE Trans Industr Inform*, vol. 17, no. 9, pp. 5867–5875, 2021, doi: 10.1109/TII.2020.3046566.

[109] H. Chen, J.-Y. Hsu, J.-Y. Hsieh, H.-Y. Hsu, C.-H. Chang, and Y.-J. Lin, “Predictive maintenance of abnormal wind turbine events by using machine learning based on condition monitoring for anomaly detection,” *Journal of Mechanical Science and Technology*, vol. 35, no. 12, pp. 5323–5333, 2021, doi: 10.1007/s12206-021-1105-z.

[110] Z. Liu, C. Xiao, T. Zhang, and X. Zhang, “Research on fault detection for three types of wind turbine subsystems using machine learning,” *Energies (Basel)*, vol. 13, no. 2, 2020, doi: 10.3390/en13020460.

[111] C. Velandia-Cardenas, Y. Vidal, and F. Pozo, “Wind turbine fault detection using highly imbalanced real scada data,” *Energies (Basel)*, vol. 14, no. 6, 2021, doi: 10.3390/en14061728.

[112] P. V. Torres-Carrion, C. S. Gonzalez-Gonzalez, S. Aciar, and G. Rodriguez-Morales, “Methodology for systematic literature review applied to engineering and education,” in *IEEE Global Engineering Education Conference, EDUCON*, IEEE Computer Society, May 2018, pp. 1364–1373. doi: 10.1109/EDUCON.2018.8363388.

[113] P. Chapman *et al.*, *CRISP-DM 1.0. Step-by-step data mining guide*, 1.0. CRISP-DM Consortium, 2000.

[114] C. Tutivén, H. Andérica, Á. Encalada, C. Benalcazar-Parra, and Y. Vidal, “Wind Turbine Multi-Fault Detection and Classification using Machine Learning Techniques,” in *International Conference on Structural Health Monitoring of Intelligent Infrastructure: Transferring Research into Practice, SHMII*, International Society for Structural Health Monitoring of Intelligent Infrastructure, ISHMII, 2021, pp. 1191–1198. [Online]. Available:

https://www.scopus.com/inward/record.uri?eid=2-s2.0-85130706047&partnerID=40&md5=5fa648b6aa3e50cc0d0b1d09c2fa17c4

[115] T. Matsui, K. Yamamoto, and J. Ogata, "Anomaly Detection Using a SCADA Feature Extractor and Machine Learning to Detect Lightning Damage on Wind Turbine Blades," *IEEJ Transactions on Electrical and Electronic Engineering*, vol. 17, no. 6, pp. 945–951, 2022, doi: 10.1002/tee.23599.

[116] M. Ayman, M. Othman, N. Mahmoud, Z. Tamer, M. Sayed, and Y. M. I. Hassan, "Fault Detection in Wind Turbines using Deep Learning," in *MIUCC 2022 - 2nd International Mobile, Intelligent, and Ubiquitous Computing Conference*, S. N. R. S. E. S. E. Bahaa-Eldin A. AbdelRaouf A., Ed., Institute of Electrical and Electronics Engineers Inc., 2022, pp. 272–278. doi: 10.1109/MIUCC55081.2022.9781749.

[117] A. Jothis, B. Nishau, Y. Abdullahr, S. Suliaman, and N. Salam, "An Adaptive Anomaly Detection and Fault Diagnosis in Wind Turbine," in *Proceedings of 2021 IEEE International Women in Engineering (WIE) Conference on Electrical and Computer Engineering, WIECON-ECE 2021*, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 172–175. doi: 10.1109/WIECON-ECE54711.2021.9829658.

[118] Z. Jiang, Y. Cao, H. Hong, and Q. Yang, "Fault detection and diagnosis of wind turbine gearbox based on acoustic analysis," in *Proceedings - 2021 International Conference on Power System Technology: Carbon Neutrality and New Type of Power System, POWERCON 2021*, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 2047–2052. doi: 10.1109/POWERCON53785.2021.9697513.

[119] J. E. Urrea Cabus, Y. Cui, and L. B. Tjernberg, "An Anomaly Detection Approach Based on Autoencoders for Condition Monitoring of Wind Turbines," in *2022 17th International Conference on Probabilistic Methods Applied to Power Systems, PMAPS 2022*, Institute of Electrical and Electronics Engineers Inc., 2022. doi: 10.1109/PMAPS53380.2022.9810575.

[120] F. Bilendo, H. Badihi, N. Lu, P. Cambron, and B. Jiang, "An Intelligent Data-Driven Machine Learning Approach for Fault Detection of Wind Turbines," in *2021 6th International Conference on Power and Renewable Energy, ICPRE 2021*, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 444–449. doi: 10.1109/ICPRE52634.2021.9635340.

[121] J. Maldonado, S. Martín, E. Artigao, and E. Gómez, "Using SCADA data for wind turbine condition monitoring: A systematic literature review," *Energies (Basel)*, vol. 13, no. 12, Jun. 2020, doi: 10.3390/en13123132.

[122] A. Stetco *et al.*, "Machine learning methods for wind turbine condition monitoring: A review," *Renew Energy*, vol. 133, pp. 620–635, Apr. 2019, doi: 10.1016/j.renene.2018.10.047.

[123] Y. Li, S. Liu, and L. Shu, “Wind turbine fault diagnosis based on Gaussian process classifiers applied to operational data,” *Renew Energy*, vol. 134, pp. 357–366, Apr. 2019, doi: 10.1016/j.renene.2018.10.088.

[124] I. M. Black, M. Richmond, and A. Kolios, “Condition monitoring systems: a systematic literature review on machine-learning methods improving offshore-wind turbine operational management,” *International Journal of Sustainable Energy*, vol. 40, no. 10, pp. 923–946, 2021, doi: 10.1080/14786451.2021.1890736.

[125] P. Qian, D. Zhang, X. Tian, Y. Si, and L. Li, “A novel wind turbine condition monitoring method based on cloud computing,” *Renew Energy*, vol. 135, pp. 390–398, May 2019, doi: 10.1016/j.renene.2018.12.045.

[126] R. Espinar, L. Tutor, and R. P. Mejías, “Modelos de Clasificación con datos no balanceados,” Universidad de Sevilla , Sevilla, 2018. Accessed: Jul. 31, 2023. [Online]. Available: https://idus.us.es/bitstream/handle/11441/77518/Espinar%20Lara%20Roc%C3%ADo%20TFG.pdf?sequence=1&isAllowed=y

[127] E. Lander, “Nueva generación de sensores electroquímicos basados en película delgada para la medida de la concentración de CO2 en conducciones de gas natural y biogás.,” Universidad de Navarra, 2014. Accessed: May 23, 2023. [Online]. Available: https://dadun.unav.edu/handle/10171/37323

[128] B. G. Guerrero Hoyos, F. D. J. Vélez Macías, and D. E. Morales Quintero, “Energía eólica y territorio: sistemas de información geográfica y métodos de decisión multicriterio en La Guajira (Colombia),” *Ambiente y Desarrollo*, vol. 23, no. 44, pp. 1–20, Feb. 2020, doi: 10.11144/javeriana.ayd23-44.eets.

[129] A. Martínez, A. Bayod, and M. Pérez, “La industria de la energía eólica en España,” *Boletín Económico de ICE*, pp. 19–30, 2002, Accessed: May 27, 2023. [Online]. Available: https://dialnet.unirioja.es/servlet/articulo?codigo=271253

[130] A. Azamar Alonso and Y. García Beltrán, “Diagnóstico y riesgos de la energía eólica en México,” *Revista de Geografía Agrícola*, no. 67, pp. 1–19, Dec. 2021, doi: 10.5154/r.rga.2021.67.02.

[131] U. Quirama Estrada, J. Sepúlveda Aguirre, M. Morelo Machado, C. Mosquera Romaña, and L. C. Valle Beleño, “Beneficios económicos de la energía renovable en Colombia,” *Administración & Desarrollo*, vol. 52, no. 2, pp. 171–183, Dec. 2022, doi: 10.22431/25005227.vol52n2.9.

[132] K. R. A. Sampaio and V. Batista, “O atual cenário da produção de energia eólica no Brasil: Uma revisão de literatura,” *Research, Society and Development*, vol. 10, no. 1, pp. 1–8, Jan. 2021, doi: 10.33448/rsd-v10i1.12107.

[133] M. Liton Hossain, A. Abu-Siada, and S. M. Muyeen, “Methods for advanced wind turbine condition monitoring and early diagnosis: A literature review,” *Energies (Basel)*, vol. 11, no. 5, pp. 1–14, 2018, doi: 10.3390/en11051309.

[134] E. Martínez, E. Macias, J. Blanco, and J. Díez, “Averías en aerogeneradores: detección a partir de datos SCADA,” *Revista DYNA*, vol. 89, no. 5, p. 1, 2014, doi: https://doi.org/10.6036/7063.

[135] E. Martinez, E. Jimenez, J. Blanco, and J. Díez, “Failure detection and prediction in wind turbines by using Scada data,” *DYNA Energía y Sostenibilidad* , vol. 2, no. 1, pp. 1–10, 2013, doi: https://doi.org/10.6036/ES5708.

[136] A. Amini, J. Kanfoud, and T. H. Gan, “An Artificial Intelligence Neural Network Predictive Model for Anomaly Detection and Monitoring of Wind Turbines Using SCADA Data,” *Applied Artificial Intelligence*, vol. 36, no. 1, pp. 1–15, 2022, doi: 10.1080/08839514.2022.2034718.

[137] M. Neshat *et al.*, “Hybrid Neuro-Evolutionary Method for Predicting Wind Turbine Power Output,” Apr. 2020. doi: https://doi.org/10.48550/arXiv.2004.12794.

[138] S. M. Zubiría, *Contemporary Didactics-Mentefacts*, 1st ed. Bogotá, Colombia, 1997.

[139] A. Zaher, S. D. J. McArthur, D. G. Infield, and Y. Patel, “Online wind turbine fault detection through automated SCADA data analysis,” *Wind Energy*, vol. 12, no. 6, pp. 574–593, 2009, doi: 10.1002/we.319.

[140] H. Zhao, H. Liu, W. Hu, and X. Yan, “Anomaly detection and fault analysis of wind turbine components based on deep learning network,” *Renew Energy*, vol. 127, pp. 825–834, Nov. 2018, doi: 10.1016/j.renene.2018.05.024.

[141] A. Stetco *et al.*, “Machine learning methods for wind turbine condition monitoring: A review,” *Renewable Energy*, vol. 133. Elsevier Ltd, pp. 620–635, Apr. 01, 2019. doi: 10.1016/j.renene.2018.10.047.

[142] M. Schlechtingen and I. Ferreira Santos, “Comparative analysis of neural network and regression based condition monitoring approaches for wind turbine fault detection,” *Mechanical Systems and Signal Processing*, vol. 25, no. 5. pp. 1849–1875, Jul. 2011. doi: 10.1016/j.ymssp.2010.12.007.

[143] A. Stetco *et al.*, “Machine learning methods for wind turbine condition monitoring: A review,” *Renew Energy*, vol. 133, pp. 620–635, Apr. 2019, doi: 10.1016/j.renene.2018.10.047.

[144] K. Leahy, L. Hu, I. Konstantakopoulos, C. Spanos, and A. Agogino, “Diagnosing Wind Turbine Faults Using Machine Learning Techniques Applied to Operational Data,” Ottawa, ON, Canada: IEEE International Conference on Prognostics and Health Management (ICPHM), Jun. 2016, pp. 1–8. doi: 10.1109/ICPHM.2016.7542860.

[145] J. Chatterjee and N. Dethlefs, “Scientometric review of artificial intelligence for operations & maintenance of wind turbines: The past, present and future,” *Renewable and Sustainable Energy Reviews*, vol. 144, pp. 1–31, Jul. 2021, doi: 10.1016/j.rser.2021.111051.

[146] Y. S. Afridi, K. Ahmad, and L. Hassan, “Artificial intelligence based prognostic maintenance of renewable energy systems: A review of techniques, challenges, and future research directions,” *Int J Energy Res*, vol. 46, no. 15, pp. 21619–21642, Dec. 2022, doi: 10.1002/er.7100.

11. Anexos

Anexo 1. Revisión Sistemática de Literatura

Modelos de predicción de fallas en aerogeneradores mediante inteligencia artificial y datos SCADA. Una revisión de literatura.

1. Resumen

En las últimas décadas, la energía eólica ha crecido significativamente, estableciéndose como una fuente de energía renovable crítica y en continuo desarrollo. A pesar de su potencial, los elevados costos de operación y mantenimiento (O&M) comprometen la eficiencia de la generación eólica, constituyendo una fracción considerable del coste total de generación en los parques eólicos. En este contexto, es evidente que la detección de fallas en los aerogeneradores puede contribuir significativamente a reducir los costos asociados a la O&M. La industria utiliza ampliamente los modernos sistemas SCADA (Supervisory Control and Data Acquisition) para monitorear los aerogeneradores recopilando una gran cantidad de datos operativos. Sin embargo, la optimización de las actividades de O&M mediante la aplicación de técnicas de inteligencia artificial (IA) puede mejorar la eficacia de la detección de anomalías y la prevención de fallas

Para avanzar en este campo, esta Revisión Sistemática de la Literatura (RSL) se centró en determinar técnicas de IA con datos SCADA para mejorar la predicción de fallas en aerogeneradores. Siguiendo la metodología adaptada de los planteamientos de Barbara Kitchenham y Bacca, se plantean tres preguntas clave: la primera busca identificar las técnicas y modelos de IA más recientes aplicados en la predicción de fallas en aerogeneradores; la segunda pregunta pretende determinar los componentes más estudiados de los aerogeneradores; y la tercera pregunta se dirige a reconocer los artículos más relevantes y citados. Es esencial mencionar que las principales conclusiones incluyen un análisis de las técnicas de IA más utilizadas, con SVM (30,56%), ANN (22,22%) y KNN (19,44%) a la cabeza, así como de los componentes más estudiados, destacando a la caja de cambios (17,3%) y los rodamientos (13,5%). Además, un examen meticuloso de 36 artículos científicos publicados entre enero de 2018 y junio de 2023 constituyó la base de este análisis.

Palabras Clave: *aerogenerador, datos SCADA, modelo híbrido, inteligencia artificial, predicción de fallas.*

1.2. Abstract

In recent decades, wind energy has significantly grown, establishing itself as a critical and continuously developing renewable energy source. Despite its potential, high operation and maintenance (O&M) costs compromise the efficiency of wind power generation, constituting a considerable fraction of the total generation cost in wind farms. In this context, it is evident that detecting faults in wind turbines can significantly contribute to reducing O&M-associated costs. The industry widely utilizes modern Supervisory Control and Data Acquisition (SCADA) systems to monitor wind turbines by collecting vast operational data. However, optimizing O&M activities by applying artificial intelligence (AI) techniques can enhance the effectiveness of anomaly detection and failure prevention.

To advance in this field, this Systematic Literature Review (RSL) focused on determining AI techniques with SCADA data to enhance the prediction of failures in wind turbines. Following the methodology adapted from the approaches of Barbara Kitchenham and Bacca, we posed three key questions: the first one seeks to identify the most recent AI techniques and models applied in predicting failures in wind turbines; the second question aims to determine the most studied components of wind turbines; and the third question is directed towards recognizing the most relevant and cited articles. It is essential to mention that the main conclusions include an analysis of the most used AI techniques, with SVM (30.56%), ANN (22.22%), and KNN (19.44%) leading the way, as well as the most studied components, highlighting gearboxes (17.3%) and bearings (13.5%). Additionally, a meticulous examination of 36 scientific articles published between 2018 and June 2023 formed the basis for this analysis.

Keywords: w*ind turbine, SCADA data, hybrid model, artificial intelligence, fault prediction.*

2. Introducción

En la sociedad moderna, la generación de energía desempeña un papel fundamental. Históricamente, la explotación de fuentes de energía fósil como el carbón mineral, el petróleo y el gas natural ha sido crucial para el progreso industrial, económico y social. Sin embargo, factores como el calentamiento global, la dependencia energética y la limitación de los recursos han impulsado el desarrollo de las energías renovables [127]. En este contexto, la energía eólica ha adquirido un gran impacto y se ha catalogado como la fuente de energía renovable de mayor crecimiento en el mundo, con una tasa de crecimiento anual del 30% en las últimas dos décadas [8], [128].

Un ejemplo destacado del crecimiento de la energía eólica se puede observar en países como España, donde la energía eléctrica ha experimentado una amplia difusión y un crecimiento notable en los últimos años. Esto ha permitido el desarrollo de un sector industrial que actualmente exporta y establece acuerdos de desarrollo y cooperación con empresas de diversos países [129]. En México, los parques eólicos tienen un valor estratégico importante para el desarrollo energético y son valiosos para el ámbito empresarial [130]. Por otro lado, en Colombia, el gobierno busca incentivar el desarrollo de la energía eólica y ha implementado una metodología para identificar y evaluar las zonas con potencial eólico [125].

La energía eólica presenta una amplia variedad de beneficios. Uno de los aspectos destacados es que se trata de una fuente de energía renovable y limpia, lo que implica que no emite gases de efecto invernadero ni otros contaminantes [130], [131], [132]. Además, resulta económicamente favorable, especialmente en áreas con vientos fuertes y constantes. También contribuye a diversificar la matriz energética de un país, reduciendo la dependencia de los combustibles fósiles [131], [132].

Otro aspecto relevante es su capacidad para funcionar como una fuente de energía descentralizada, lo que significa que puede generarse en pequeña escala y cerca del lugar donde se necesita consumir la energía. Además, es autónoma, lo que implica que puede utilizarse en áreas remotas o aisladas donde no hay acceso a la red eléctrica [133]. Sin embargo, a pesar de los múltiples beneficios que ofrece la energía eólica, existen problemas relacionados con las tareas de operación y mantenimiento (O&M) de los aerogeneradores que pueden afectar estos beneficios. Estos problemas pueden afectar a diversos componentes, como se puede observar en la Figura 1.

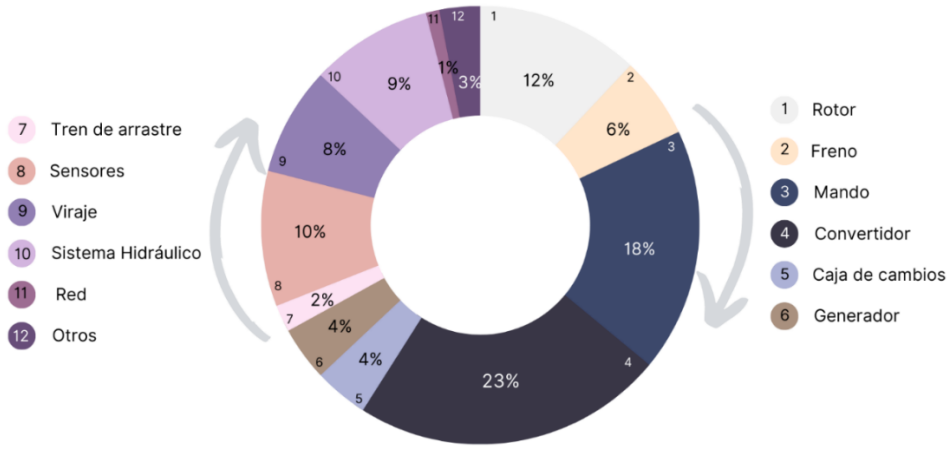

Figura 1. Medición de la frecuencia de averías en los elementos de los aerogeneradores. Adaptado de [133].

Según los resultados recopilados, se encontró que el convertidor (converter) es el componente más propenso a fallas, representando el 23% de las averías registradas. Le sigue el control (control) con un 18%, y el rotor (rotor) con un 12%. Otros problemas reportados incluyen el sistema hidráulico (hydraulic system) con un 9%, el sistema de orientación (yaw) (8%), y los sensores (10%). El freno, la caja de cambios (gearbox) y el generador (generator) presentaron una incidencia similar de fallas, con un 6%, 4% y 4% respectivamente. Otros componentes, como el tren de transmisión (drive train), el sistema hidráulico y otros elementos, representaron un porcentaje menor de las averías, con un 2%, 9% y 3% respectivamente, y la interconexión con la red eléctrica (grid) tuvo la menor incidencia de fallas, con solo un 1%.

Para mejorar la estabilidad de los aerogeneradores y abordar posibles problemas, el uso de datos SCADA (Supervisory Control and Data Acquisition) se ha propuesto como una solución alternativa para el monitoreo de estos sistemas [134], [135]. Los datos recopilados por el sistema SCADA, mediante el uso de inteligencia artificial (IA), pueden emplearse para predecir y detectar posibles fallas en los componentes principales de un aerogenerador.

Existen numerosos estudios que utilizan la combinación de datos SCADA e IA para predecir fallas en los aerogeneradores. Esta combinación permite la detección temprana de fallas, mejora la eficiencia, reduce los costos y entre otros beneficios. Por ejemplo, en el estudio [136], se presenta un modelo predictivo basado en una red neuronal de IA que utilizó datos SCADA para predecir la temperatura de salida del generador del aerogenerador con una precisión del 99,8% detectada hasta dos meses antes.

Sin embargo, es importante destacar que el uso de modelos híbridos puede ofrecer ventajas superiores a otros modelos de IA. Por ejemplo, en [137], se propone un enfoque de

machine learning (ML) compuesto por un algoritmo neuroevolutivo híbrido, el cual logra una predicción eficaz de la producción de energía en parques eólicos que supera a modelos de IA tradicionales e incluso a sus homólogos, dicha evidencia hace énfasis en la relevancia de los modelos híbridos en el campo de la IA aplicada a la energía eólica, marcando un hito importante en la búsqueda de alternativas más eficientes y efectivas.

Dada la relevancia y el potencial de estos hallazgos, es esencial delinear las contribuciones clave que emergen de esta Revisión Sistemática de Literatura (RSL). En este contexto, las principales contribuciones de este estudio son las siguientes:

- Determinación de las técnicas y modelos de IA más recientes aplicados a la predicción de fallas.
- Identificación de los componentes del aerogenerador que más se estudian en la predicción de fallas.
- Identificación de los artículos más importantes de predicción de fallas en aerogeneradores en los últimos años (enero de 2018 hasta junio de 2023)
- Realización de un análisis bibliométrico de los documentos seleccionados.
- Identificación de modelos netamente híbridos.
- Comparación de los resultados obtenidos con los hallazgos de otros estudios relevantes, proporcionando un marco comparativo para evaluarlo frente a otros enfoques.

La revisión actual se desarrolla de la siguiente manera: comienza con la metodología, donde se detalla el proceso utilizado en la RSL, ilustrando las diferentes fases del artículo con un mapa conceptual y definiendo las preguntas de investigación, esta sección comprende tres subsecciones: planificación, realización de la revisión e informe de revisión, se concluye con la sección de discusión, donde se discuten los hallazgos, y finalmente, en la sección conclusiones, se presentan las conclusiones extraídas de esta revisión.

3. Metodología

Con la finalidad de responder a las preguntas de investigación detalladas posteriormente, se realizó la RSL apoyada de la metodología propuesta por Torres [112], que a su vez está basada en un método adaptado de los enfoques desarrollados por Barbara Kitchenham y Bacca. La Figura 2, esquematiza detalladamente los pasos que conforman el proceso llevado a cabo en esta RSL.

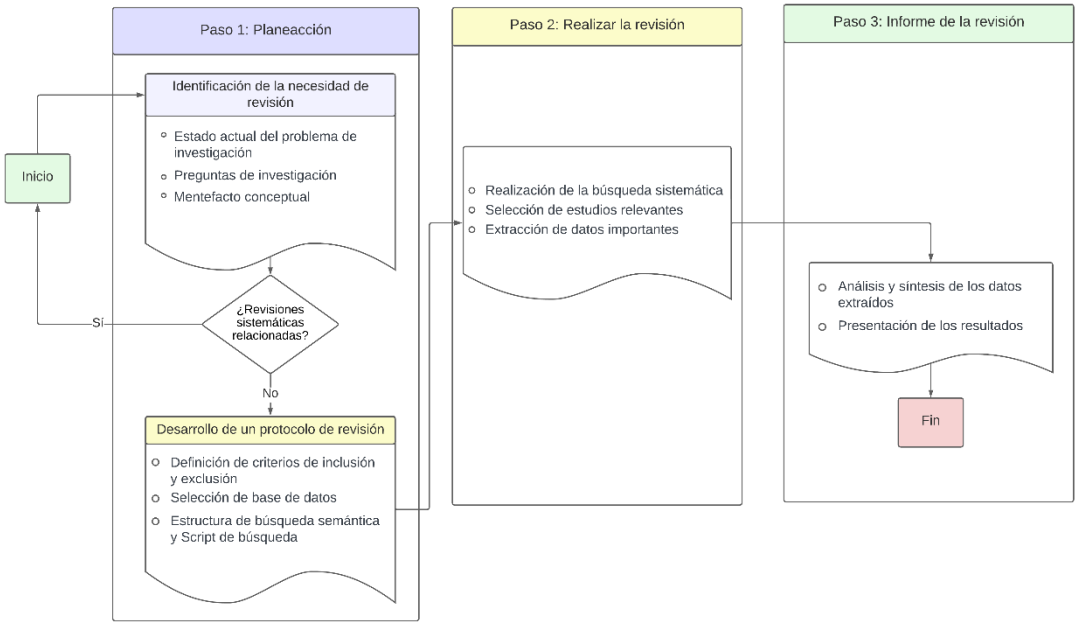

Figura 2. Diagrama de flujo de la metodología. Adaptado de [121].

3.1. Descripción de las fases de RSL

3.1.1. Fase de planeación

En esta fase, se realiza la descripción detallada del problema de investigación, el planteamiento de las preguntas de investigación, la elaboración del mentefacto conceptual y la definición de los criterios de inclusión y exclusión utilizados en la RSL.

3.1.1.1. Identificación del problema de investigación

La generación de energía eólica ha experimentado un crecimiento acelerado en las últimas décadas, sin embargo, este avance conlleva desafíos, especialmente en los costos asociados con O&M en los aerogeneradores, los cuales pueden aumentar debido a fallas inesperadas, haciendo indispensable la capacidad de detectar y prever problemas futuros para reducir los gastos de O&M y garantizar la eficiencia energética de los aerogeneradores.

En este sentido, la RSL se centra en comprender los enfoques tratados por los autores, como las técnicas y modelos empleados, componentes de los aerogeneradores más estudiados y los artículos más relevantes. El propósito es ampliar el conocimiento sobre soluciones de IA aplicables a los aerogeneradores para mejorar la eficiencia en la predicción y manejo de fallas.

3.1.1.2. Planteamiento de preguntas de investigación (PI):

En la investigación científica, la formulación de preguntas de investigación es crucial para dirigir el enfoque y dirección del estudio. Por lo cual se establecieron las siguientes preguntas clave para guiar y orientar la RSL:

- PI1: ¿Cuáles son las técnicas y modelos de inteligencia artificial más usados aplicados a la predicción de fallas de aerogeneradores?
- PI2: ¿Cuáles son los componentes del aerogenerador que más se estudian en la predicción de fallas?
- PI3: ¿Cuáles son los artículos más importantes (usualmente más citados) de predicción de fallas en aerogeneradores en los últimos años?

3.1.1.3. Mentefacto conceptual

El mentefacto conceptual concebida por [138], se diseñó con el objetivo de mejorar la comprensión y facilitar el aprendizaje efectivo, este grafico es de uso frecuente en contextos pedagógicos para la representación visual de conceptos clave que ayudan así a estructurar y organizar el conocimiento de forma clara y se compone de varias categorías que organizan la información jerárquicamente:

- **Supraordenada**: Representa un concepto general o principal alrededor del cual se organiza la información.
- **Isoordenada**: Incorpora conceptos que son igualmente relevantes y que se relacionan entre sí.
- **Infraordenada**: Integra conceptos mucho más específicos o subordinados que se ubican dentro de la jerarquía superordinada.
- **Exclusiones**: Vincula conceptos que no forman parte del conjunto o que sencillamente no se incorporan bajo el contexto principal.

La Figura 3, que incorpora estas categorías se ha demostrado como una solución eficaz para guiar tanto la búsqueda bibliográfica como el planteamiento del problema de investigación.

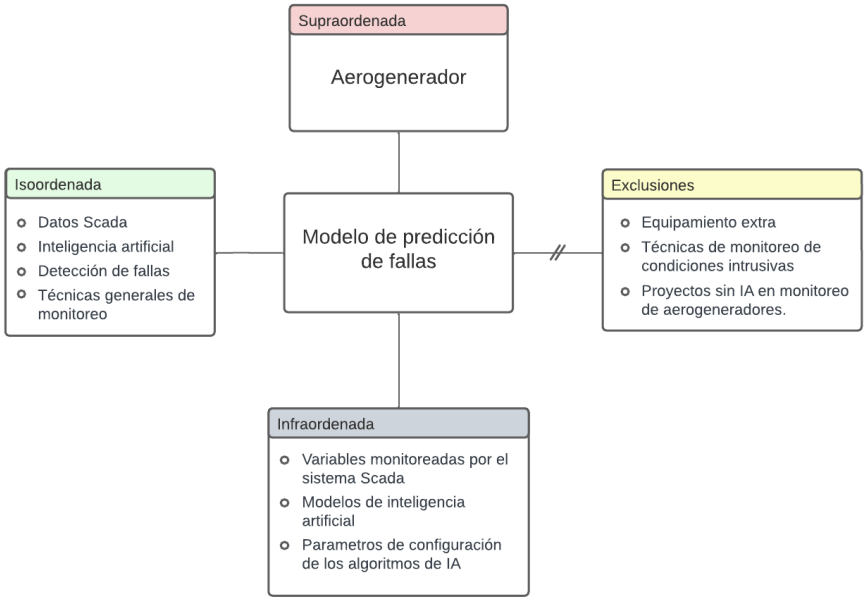

Figura 3. Mentefacto conceptual.

Su análisis e interpretación es la siguiente:

- En la parte superior donde se ubica la categoría "Aerogenerador" se establece como el concepto raíz o principal que se enfoca en el "Modelo de predicción de fallas".

  En la parte derecha donde se ubican las exclusiones se señalan aquellos conceptos relacionados, pero no directamente pertinentes o acordes a la búsqueda específica.
- En la sección del lado izquierdo se localizan las isoordenadas que incluyen conceptos clave que han sido esenciales para la búsqueda bibliográfica.
- En la parte inferior se ubican las infraordinadas que detallan las variables esenciales para identificar modelos relevantes de IA.

3.1.1.4. Definición de los criterios de inclusión y exclusión.

Con la finalidad de incorporar información relevante y/o desestimar aquella que no aporta en la RSL, se determinaron criterios de inclusión y exclusión para la selección de artículos que debido a su importancia se detallan a continuación.

3.1.1.4.1. Criterios de Inclusión

- Artículos (Articles, Conferences Papers, Reviews) publicados en reconocidas bases de datos científicas
- Artículos publicados entre enero de 2018 – junio de 2023

- Artículos publicados en el idioma Ingles

3.1.1.4.2. Criterios de exclusión

- Revistas que no se consideran en publicaciones científicas.
- Patentes, Libros, Editoriales, Reportes técnicos, Catálogos.
- Excluir todos los campos del conocimiento que sean Física, Astronomía, Profesiones sanitarias y Ciencia de materiales.

3.1.1.5. Selección de base de datos y fuentes de información relevantes.

Se utilizó la plataforma Scopus (Ver enlace: https://www.scopus.com/) en la adquisición de artículos y trabajos de investigación relevantes, la cual es una base de datos bibliográfica de amplia cobertura multidisciplinaria que integra una extensa selección de revistas científicas y conferencias de alta calidad, que ofrece la opción de búsqueda avanzada para refinar y optimizar los resultados, dicha elección aseguró el acceso a una variedad de fuentes confiables y actualizadas.

3.1.1.6. Estructura semántica de búsqueda

Para efectuar una búsqueda precisa y efectiva en la base de datos Scopus, se desarrolló un script que emplea una estructura semántica definida mediante un mentefacto y apoyada por un tesauro científico, dicha metodología asegura resultados más exactos y eficientes, así en la Tabla 1 se detalla la estructura de la búsqueda semántica que se encuentra organizada en cuatro niveles.

Tabla 1. Estructura de búsqueda semántica

Nivel 1	+ Wind Turbine	wind W/1 (turbine OR farm OR power)
Nivel 2	+ Artificial Intelligence	AND ("machine learning” OR “Artificial intelligence" OR "AI")
Nivel 3	+ Fault Prediction	AND TITLE (detection W/1 (fault OR failure OR anomaly OR diagnosis)
Nivel 4	+ SCADA	AND (SCADA)

El primer nivel se centra en la categoría principal o supraordenada, el siguiente nivel es enfocado en la exploración de investigaciones acordes a la IA, por otra parte, el tercer nivel identifica los estudios relacionados con la detección de fallas y el nivel final se basa en la búsqueda dentro de los sistemas SCADA.

Es importante recalcar que los cuatro niveles se encuentran estrictamente relacionados a las preguntas de investigación, donde cada nivel asume los términos sinónimos proporcionados por el tesauro, permitiendo considerar los

más relacionados, así, en la Tabla 2 se presenta la cadena de búsqueda elaborada para efectuar la investigación en Scopus.

Tabla 2. Cadena de búsqueda

Cadena de búsqueda
TITLE-ABS-KEY (wind W/1 (turbine OR farm OR power) AND ("machine learning" OR "Artificial intelligence" OR "AI")) AND TITLE (detection W/1 (fault OR failure OR anomaly OR diagnosis)) AND (scada)

La cadena de búsqueda presentada en la Tabla 2 incluye todos los términos que obedecen a la estructura de búsqueda semántica presentada en la Tabla 1, no obstante, la base de datos Scopus cumple con precisión los criterios de inclusión y exclusión previamente establecidos y es por ello que la Tabla 3 presenta la cadena de búsqueda que incorpora dichos criterios y arroja un total de treinta y nueve resultados considerando varios tipos de documentos entre ellos artículos, conferencias y revistas.

Tabla 3. Cadena de búsqueda con los criterios de inclusión y exclusión

Cadena de búsqueda
TITLE-ABS-KEY (wind W/1 (turbine OR farm OR power) AND ("machine learning" OR "Artificial intelligence" OR "AI")) AND TITLE (detection W/1 (fault OR failure OR anomaly OR diagnosis)) AND (scada) AND (LIMIT-TO (PUBYEAR , 2023) OR LIMIT-TO (PUBYEAR , 2022) OR LIMIT-TO (PUBYEAR , 2021) OR LIMIT-TO (PUBYEAR , 2020) OR LIMIT-TO (PUBYEAR , 2019) OR LIMIT-TO (PUBYEAR , 2018)) AND (EXCLUDE (SUBJAREA , "PHYS") OR EXCLUDE (SUBJAREA , "HEAL") OR EXCLUDE (SUBJAREA , "MATE")) AND (EXCLUDE (SRCTYPE , "k"))

3.1.2. Fase de realización de la revisión

3.1.2.1. Realización de la búsqueda sistemática de la literatura utilizando los criterios de búsqueda definidos en la fase de planificación.

La primera cadena búsqueda detallada anteriormente en la Tabla 2 arroja un grupo de resultados de alrededor de cincuenta y cuatro documentos, como se muestra en la Figura 4.

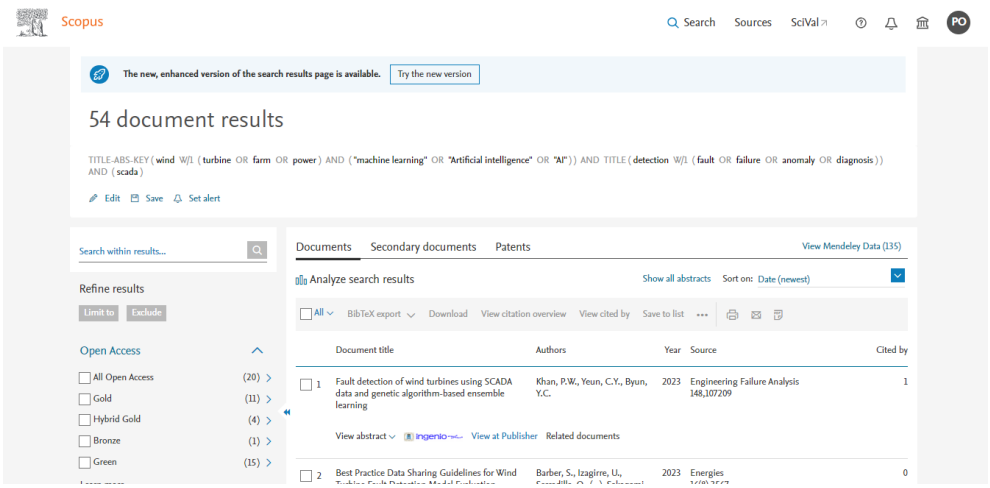

Figura 4. Resultado de búsqueda en Scopus

3.1.2.2. Selección de los artículos relevantes según los criterios de inclusión y exclusión definidos en la fase de planificación.

La selección de documentos relevantes para la RSL se realizó obedeciendo los criterios de inclusión y exclusión establecidos en la metodología que durante el proceso fueron ajustados en la base de datos de Scopus para refinar la búsqueda y tras la modificación de la cadena de búsqueda y la aplicación de estos criterios fueron identificados un total de treinta y seis documentos considerados pertinentes para la presente investigación.

En la Tabla 4, se presenta un resumen de los documentos seleccionados.

Tabla 4. Resumen de los documentos seleccionados

N°	Ref	T	AU	A	DOI
1	[88]	Anomaly detection for wind turbine damaged due to lightning strike	Matsui, T. and Yamamoto, K. and Ogata, J.	2022	10.1016/j.epsr.2022.107918
2	[91]	Fault detection and anti-icing technologies in wind energy conversion systems: A review	Choe Wei Chang, C. and Jian Ding, T. and Jian Ping, T. and Ariannejad, M. and Chia Chao, K. and Samdin, S.B.	2022	10.1016/j.egyr.2022.10.234
3	[89]	Condition monitoring and anomaly detection of wind turbine based on cascaded and bidirectional deep learning networks	Xiang, L. and Yang, X. and Hu, A. and Su, H. and Wang, P.	2022	10.1016/j.apenergy.2021.117925
4	[107]	Fault detection by an ensemble framework of Extreme Gradient Boosting (XGBoost) in the operation of offshore wind turbines	Trizoglou, P. and Liu, X. and Lin, Z.	2021	10.1016/j.renene.2021.07.085

N°	Ref	T	AU	A	DOI
5	[102]	Wind turbine generator controller signals supervised machine learning for shaft misalignment fault detection: A doubly fed induction generator practical case study	Al-Ajmi, A. and Wang, Y. and Djurović, S.	2021	10.3390/en14061601
6	[95]	Combination of thermal modelling and machine learning approaches for fault detection in wind turbine gearboxes	Corley, B. and Koukoura, S. and Carroll, J. and McDonald, A.	2021	10.3390/en14051375
7	[105]	Combining SCADA and vibration data into a single anomaly detection model to predict wind turbine component failure	Turnbull, A. and Carroll, J. and McDonald, A.	2021	10.1002/we.2567
8	[90]	Evaluation of anomaly detection of an autoencoder based on maintenace information and SCADA-data	Lutz, M.-A. and Vogt, S. and Berkhout, V. and Faulstich, S. and Dienst, S. and Steinmetz, U. and Gück, C. and Ortega, A.	2020	10.3390/en13051063
9	[93]	Real-time condition monitoring and fault detection of components based on machine-learning reconstruction model	Yang, C. and Liu, J. and Zeng, Y. and Xie, G.	2019	10.1016/j.renene.2018.10.062
10	[86]	An anomaly detection approach based on machine learning and scada data for condition monitoring of wind turbines	Cui, Y. and Bangalore, P. and Tjernberg, L.B.	2018	10.1109/PMAPS.2018.8440525
11	[96]	Deep learning for automated drivetrain fault detection	Bach-Andersen, M. and Rømer-Odgaard, B. and Winther, O.	2018	
12	[97]	Deep Learning for fault detection in wind turbines	Helbing, G. and RiTICer, M.	2018	10.1016/j.rser.2018.09.012
13	[101]	Tandem Connectionist Anomaly Detection: Use of Faulty Vibration Signals in Feature Representation Learning	Hasegawa, T. and Ogata, J. and Murakawa, M. and Ogawa, T.	2018	10.1109/ICPHM.2018.8448450

N°	Ref	T	AU	A	DOI
14	[100]	SCADA-based wind turbine anomaly detection using Gaussian process models for wind turbine condition monitoring purposes	Pandit, R.K. and Infield, D.	2018	10.1049/iet-rpg.2018.0156
15	[110]	Research on fault detection for three types of wind turbine subsystems using machine learning	Liu, Z. and Xiao, C. and Zhang, T. and Zhang, X.	2020	10.3390/en13020460
16	[104]	A Small-Sample Wind Turbine Fault Detection Method With Synthetic Fault Data Using Generative Adversarial Nets	Liu, J. and Qu, F. and Hong, X. and Zhang, H.	2019	10.1109/TII.2018.2885365
17	[34]	Wind turbines anomaly detection based on power curves and ensemble learning	Moreno, S.R. and Coelho, L.S. and Ayala, H.V.H. and Mariani, V.C.	2020	10.1049/iet-rpg.2020.0224
18	[106]	Fault Detection Based on a Combined Approach of FA-CP-ELM with Application to Wind Turbine System	Yu, W. and Huang, S. and Wang, J.	2021	10.1007/s42835-020-00561-z
19	[108]	Imbalanced Classification Based on Minority Clustering Synthetic Minority Oversampling Technique with Wind Turbine Fault Detection Application	Yi, H. and Jiang, Q. and Yan, X. and Wang, B.	2021	10.1109/TII.2020.3046566
20	[103]	A data-mining approach for wind turbine fault detection based on scada data analysis using artificial neural networks	Santolamazza, A. and Dadi, D. and Introna, V.	2021	10.3390/en14071845
21	[111]	Wind turbine fault detection using highly imbalanced real scada data	Velandia-Cardenas, C. and Vidal, Y. and Pozo, F.	2021	10.3390/en14061728
22	[94]	An Artificial Intelligence Neural Network Predictive Model for Anomaly Detection and Monitoring of Wind Turbines Using SCADA Data	Amini, A. and Kanfoud, J. and Gan, T.-H.	2022	10.1080/08839514.2022.2034718

N°	Ref	T	AU	A	DOI
23	[109]	Predictive maintenance of abnormal wind turbine events by using machine learning based on condition monitoring for anomaly detection	Chen, H. and Hsu, J.-Y. and Hsieh, J.-Y. and Hsu, H.-Y. and Chang, C.-H. and Lin, Y.-J.	2021	10.1007/s12206-021-1105-z
24	[33]	Reduced neural network based ensemble approach for fault detection and diagnosis of wind energy converter systems	Dhibi, K. and Mansouri, M. and Bouzrara, K. and Nounou, H. and Nounou, M.	2022	10.1016/j.renene.2022.05.082
25	[99]	Exploring Quantum Machine Learning and Feature Reduction Techniques for Wind Turbine Pitch Fault Detection	Correa-jullian, C. and Cofre-martel, S. and Martin, G.S. and DrogueTIC, E.L. and Leite, G.N.P. and Costa, A.	2022	10.3390/en15082792
26	[87]	Anomaly detection and condition monitoring of wind turbine gearbox based on LSTM-FS and transfer learning	Zhu, Y. and Zhu, C. and Tan, J. and Tan, Y. and Rao, L.	2022	10.1016/j.renene.2022.02.061
27	[85]	A methodology for performance assessment at system level—Identification of operating regimes and anomaly detection in wind turbines	Urmeneta, J. and Izquierdo, J. and Leturiondo, U.	2023	10.1016/j.renene.2023.01.035
28	[92]	Fleet-based early fault detection of wind turbine gearboxes using physics-informed deep learning based on cyclic spectral coherence	Perez-Sanjines, F. and Peeters, C. and Verstraeten, T. and Antoni, J. and Nowé, A. and Helsen, J.	2023	10.1016/j.ymssp.2022.109760
29	[98]	Enabling Co-Innovation for a Successful Digital Transformation in Wind Energy Using a New Digital Ecosystem and a Fault Detection Case Study	Barber, S. and Lima, L.A.M. and Sakagami, Y. and Quick, J. and Latiffianti, E. and Liu, Y. and Ferrari, R. and Letzgus, S. and Zhang, X. and Hammer, F.	2022	10.3390/en15155638
30	[115]	Anomaly Detection Using a SCADA Feature Extractor and Machine Learning to Detect Lightning Damage on Wind Turbine Blades	Matsui, T. and Yamamoto, K. and Ogata, J.	2022	10.1002/tee.23599

N°	Ref	T	AU	A	DOI
31	[117]	An Adaptive Anomaly Detection and Fault Diagnosis in Wind Turbine	Jothis, A. and Nishau, B. and Abdullahr, Y. and Suliaman, S. and Salam, N.	2021	10.1109/WIECON-aECE54711.2021.9829658
32	[120]	An Intelligent Data-Driven Machine Learning Approach for Fault Detection of Wind Turbines	Bilendo, F. and Badihi, H. and Lu, N. and Cambron, P. and Jiang, B.	2021	10.1109/ICPRE52634.2021.9635340
33	[114]	Wind Turbine Multi-Fault Detection and Classification using Machine Learning Techniques	Tutivén, C. and Andérica, H. and Encalada, Á. and Benalcazar-Parra, C. and Vidal, Y.	2021	
34	[118]	Fault detection and diagnosis of wind turbine gearbox based on acoustic analysis	Jiang, Z. and Cao, Y. and Hong, H. and Yang, Q.	2021	10.1109/POWERCON53785.2021.9697513
35	[119]	An Anomaly Detection Approach Based on Autoencoders for Condition Monitoring of Wind Turbines	Urrea Cabus, J.E. and Cui, Y. and Tjernberg, L.B.	2022	10.1109/PMAPS53380.2022.9810575
36	[116]	Fault Detection in Wind Turbines using Deep Learning	Ayman, M. and Othman, M. and Mahmoud, N. and Tamer, Z. and Sayed, M. and Hassan, Y.M.I.	2022	10.1109/MIUCC55081.2022.9781749

Número (N°); Referencia (Ref); Título (T); Autores (AUT); Año (A); Digital Object Identifier (DOI).

3.1.2.3. Extracción de datos relevantes de los artículos seleccionados.

La Tabla 5 agrupa las técnicas y modelos utilizados en la predicción de fallas de aerogeneradores cuyos datos han sido extraídos de cada documento analizado, donde se dio prioridad a aquellas técnicas y modelos que según lo planteado en los artículos presentaron un rendimiento superior o una notable relevancia.

Tabla 5. Recopilación de técnicas/modelos

N°	Ref	T/M	Nombres completos
1	[88]	GMM	Modelo de mezcla gaussiana (GMM).
2	[91]	MSSM, BAS-SVM, MPE	Algoritmo de aprendizaje supervisado de Mapeo Semisupervisado de Mahalanobis (MSSM), Algoritmo de búsqueda de antenas de escarabajo basado en SVM (BAS-SVM), Entropía de permutación multiescala (MPE)

N°	Ref	T/M	Nombres completos
3	[89]	CNN-BiGRU-AM	Combinación de redes neuronales convolucionales (CNN), Unidades recurrentes bidireccionales con mecanismo de atención (BiGRU-AM).
4	[107]	XGBoost, LSTM	Modelo de Extreme Gradient Boosting (XGBoost), Red neuronal de aprendizaje profundo llamada Long Short-Term Memory (LSTM).
5	[102]	DT, RF, Catboost, XGBoost, SVM, LR	Árbol de decisión (DT), Bosque Aleatorio (RF), Catboost, Extreme Gradient Boosting (XGBoost), Máquina de vectores soporte (SVM), Clasificador logístico (LR)
6	[95]	ML (Black Box)	Aprendizaje automático basado en Black box (ML (Black Box))
7	[105]	FNN, RF, SVM, NBM	Red neuronal de alimentación directa de dos capas (FNN), Algoritmo de bosques aleatorios (RF), Clasificador de máquina de vectores de soporte de una sola clase (SVM), Modelo de comportamiento normal (NBM)
8	[90]	NBM, AE	Modelo de comportamiento normal (NBM), Autoencoder (AE)
9	[93]	SVR	Regresión por vectores de soporte (SVR)
10	[86]	NARX	Red neuronal Autoregresiva no lineal con entradas exógenas (NARX)
11	[96]	LR, FCNN, CNN	Regresión logística (RL), Red neuronal totalmente conectada (FCNN), Red neuronal convolucional (CNN)
12	[97]	ANN	Redes neuronales artificiales (ANN)
13	[101]	DNN, GMM: DNN/GMM	Redes neuronales profundas (DNN), Modelos de mezcla gaussiana (GMM).
14	[100]	GP	Proceso gaussiano (GP)
15	[110]	CNN, SVM, SVR	Red neuronal convolucional (CNN), Máquina de vectores soporte (SVM), Regresión por vectores de soporte (SVR)
16	[104]	GANs, ANN, SVM, DT	Redes generativas adversariales (GANs), Redes neuronales artificiales (ANN), Máquina de vectores soporte (SVM), Árbol de decisión (DT)
17	[34]	W-kNNs, BT, RBT, RF, RotF, SVM: Ensemble W-kNN	Vecinos más cercanos k ponderados (Wk-NNs), Árbol potenciado (BT), Árbol reforzado RUS (RBT), Bosque aleatorio (RF), Bosque rotatorio (RotF), Máquinas de vectores soporte (SVM), Conjunto W-kNN
18	[106]	FA, CP, ELM: FA-CP-ELM	Algoritmo Firefly (FA), Mapa del Caos (CP) y Máquina de Aprendizaje Extremo (ELM)
19	[108]	SVM, CART, CB	Máquina de vectores soporte (SVM), Árboles de clasificación y regresión (CART), Clasificadores Bayesianos (CB)
20	[103]	ANN	Redes neuronales artificiales (ANN)
21	[111]	kNN, SVM, RUSBoost, PCA	K-vecinos más próximos (kNN), Máquina de vectores soporte (SVM), Refuerzo por submuestreo aleatorio (RUSBoost), Análisis de componentes principales (PCA)

N°	Ref	T/M	Nombres completos
22	[94]	ANN	Redes neuronales artificiales (ANN)
23	[109]	RF, SVM, KNN, DNN	Bosque aleatorio (RF), Máquina de vectores soporte (SVM), K-vecinos más próximos (kNN), Red Neuronal Profunda (DNN).
24	[33]	NNs, Bagging, Boosting, Random subspace: DEL, ANN, FFNN, CFNN: NN-EL, H-K-means, R-NN-EL	Redes neuronales (NNs), Bagging, Boosting, Subespacio aleatorio, Aprendizaje Extremadamente Profundo (DEL), Redes neuronales artificiales (ANN), Red Neuronal de Propagación Hacia Adelante (FFNN), Red Neuronal de Propagación en Cascada (CFNN), Aprendizaje extremo basado en redes neuronales (NN-EL), K-means jerárquico (H-K-means), Aprendizaje Extremo basado en Redes Neuronales Reducido (R-NN-EL)
25	[99]	PCA, AE	Análisis de componentes principales (PCA), Autoencoder (AE)
26	[87]	LSTM, FS, CNN, FTL: LSTM-FS	Redes neuronales de memoria a corto y largo plazo (LSTM), Síntesis difusa (FS), Redes neuronales convolucionales (CNN), Aprendizaje por transferencia basado en características (FTL)
27	[85]	PCA y K-means	Análisis de componentes principales (PCA), Agrupación de K-means (K-means)
28	[92]	DAE	Autocodificador profundo (DAE)
29	[98]	NBM, LoMST-CUSUM, WHC-LOF, CCA, KCPD	Modelo de comportamiento normal (NBM), Árbol de expansión mínima local combinado y suma acumulativa de datos de series temporales multivariadas (LoMST-CUSUM), Agrupamiento jerárquico de Ward combinado y detección de novedades con factor de valores atípicos local (WHC-LOF), Modelo de comportamiento normal con entradas retardadas (NBM-LI), Análisis de correlación canónica (CCA), Detección de puntos de cambio basada en núcleos (KCPD)
30	[115]	GMM, SVR	Modelo de mezcla gaussiana (GMM), Regresión por vectores de soporte (SVR)
31	[117]	AE, kNN	Autoencoder (AE), K-vecinos más próximos (kNN)
32	[120]	ANN	Redes neuronales artificiales (ANN)
33	[114]	SVM, XGBoost	Máquina de vectores soporte (SVM), Extreme Gradient Boosting (XGBoost)
34	[118]	AE, K-means, t-SNE	Autoencoder (AE), Agrupación de K-means (K-means), Incrustación estocástica de vecinos t-distribuida (t-SNE)
35	[119]	AE	Autoencoder (AE)
36	[116]	SVM, RF, DT, LSTM-AE, PCA	Máquina de vectores soporte (SVM), Bosque aleatorio (RF), Árbol de decisión (DT), Autoencoders de Memoria a Corto y Largo Plazo (LSTM-AE), Análisis de componentes principales (PCA)

Número (N°); Referencia (Ref); Técnicas/Modelos (T/M).

A partir de los datos recopilados, se pueden identificar algunas tendencias interesantes. Se observa la cantidad de 103 técnicas y modelos recolectados.

- **Identificación de modelos híbridos**

La Tabla 6 presenta los modelos netamente híbridos identificados de la Tabla 5. Estos combinan diversas técnicas y modelos para optimizar el rendimiento y la precisión en el aprendizaje automático. Cada uno se caracteriza por una composición única, lo que facilita la explotación de las ventajas específicas de cada uno en la resolución de desafíos particulares.

Tabla 6. Modelos híbridos identificados

N°	Ref	Modelo híbrido	Nombre de la combinación
1	[101]	DNN/GMM	Redes neuronales profundas (DNN), Modelos de mezcla gaussiana (GMM).
2	[91]	BAS-SVM	Algoritmo de búsqueda de antenas de escarabajo basado en SVM (BAS-SVM).
3	[89]	CNN-BiGRU-AM	Combinación de redes neuronales convolucionales (CNN), Unidades recurrentes bidireccionales con mecanismo de atención (BiGRU-AM).
4	[34]	Ensemble W-kNN	Vecinos más cercanos k ponderados (Wk-NNs).
5	[106]	FA-CP-ELM	Algoritmo Firefly (FA), Mapa del Caos (CP) y Máquina de Aprendizaje Extremo (ELM)
6	[33]	CFNN	Red Neuronal de Propagación en Cascada (CFNN).
7	[33]	NN-EL	Aprendizaje extremo basado en redes neuronales (NN-EL).
8	[33]	R-NN-EL	Aprendizaje Extremo basado en Redes Neuronales Reducido (R-NN-EL)
9	[87]	LSTM-FS	Redes neuronales de memoria a corto y largo plazo (LSTM), Síntesis difusa (FS).
10	[116]	LSTM-AE	Autoencoders de Memoria a Corto y Largo Plazo (LSTM-AE).

Número (N°); Referencia (Ref).

3.1.2.4. Extracción de componentes del aerogenerador que más se estudian en la predicción de fallas.

La Tabla 7 proporciona un desglose detallado de los componentes de aerogeneradores más estudiados, esenciales para su operación y directamente vinculados a su rendimiento y fiabilidad. Al analizarla, se observan diferentes tendencias en la investigación de fallas en distintos componentes.

Con un total de cincuenta y dos elementos listados, se destacan aquellos que han recibido mayor atención investigativa, facilitando así la identificación de áreas prioritarias para la detección y prevención de fallas. Este enfoque permite centrar los esfuerzos de investigación en los componentes más críticos,

optimizando las estrategias para mejorar la eficiencia y seguridad de los aerogeneradores.

Tabla 7. Componentes del aerogenerador más estudiados

N°	Ref	Componentes estudiados
1	[88]	Blade
2	[91]	Wind Turbine
3	[89]	Bearing, Gearbox
4	[107]	Generator, Bearing
5	[102]	Wind Turbine
6	[95]	Gearbox
7	[105]	Gearbox, Generator, Bearing
8	[90]	Wind Turbine
9	[93]	Generator, Bearing
10	[86]	Gearbox
11	[96]	Bearing, Planetary,
12	[97]	Wind Turbine
13	[101]	Wind Turbine
14	[100]	Yaw
15	[110]	Generator, Converter, Pitch Sistema
16	[104]	Wind Turbine
17	[34]	Wind Turbine
18	[106]	Wind Turbine
19	[108]	Blade
20	[103]	Gearbox, Generator
21	[111]	Gearbox
22	[94]	Wind Turbine
23	[109]	Wind Turbine
24	[33]	Converter
25	[99]	Pitch
26	[87]	Wind Turbine
27	[85]	Wind Turbine
28	[92]	Gearbox
29	[98]	Gearbox, Generator, Bearing, Transformer, Hydraulic Group
30	[115]	Wind Turbine
31	[117]	Wind Turbine
32	[120]	Wind Turbine
33	[114]	Wind Turbine
34	[118]	Wind Turbine
35	[119]	Rotor, Gearbox, Hydraulic System
36	[116]	Wind Turbine

Número (N°); Referencia (Ref).

3.1.2.5. Extracción de la cantidad de citas de los documentos seleccionados.

El número de citas que recibe un artículo constituye un indicador clave de su importancia y relevancia dentro del ámbito académico, en este sentido, se realizó un conteo de las citas de los documentos seleccionados en este trabajo.

La Tabla 8 presenta un recuento de las citas de dichos documentos, organizadas en orden ascendente, en donde se remarca que los cuatro últimos documentos de la lista emergen como los más citados siendo [93], [100], [34], [97] acumulando respectivamente 73, 76, 92, y 124 citas, subrayando la relevancia de estos trabajos y contribución a la investigación en sus respectivos campos, este análisis refleja su rol fundamental en enriquecer y guiar el desarrollo de futuras investigaciones.

Tabla 8. Cantidad de citaciones de los documentos seleccionados.

N°	Ref	Cantidad de citaciones
1	[92]	0
2	[115]	0
3	[120]	0
4	[118]	0
5	[119]	0
6	[88]	1
7	[108]	1
8	[109]	1
9	[98]	1
10	[117]	1
11	[116]	1
12	[91]	2
13	[102]	2
14	[114]	2
15	[87]	3
16	[104]	4
17	[33]	6
18	[101]	7
19	[99]	7
20	[90]	11
21	[85]	11
22	[106]	14
23	[95]	15
24	[105]	15
25	[94]	16
26	[110]	17
27	[111]	19
28	[107]	25
29	[86]	31
30	[96]	35
31	[103]	36

N°	Ref	Cantidad de citaciones
32	[89]	47
33	[93]	73
34	[100]	76
35	[34]	92
36	[97]	124

Número (N°); Referencia (Ref).

3.1.3. Fase de informe de la revisión

3.1.3.1. Análisis y síntesis de los datos extraídos de los artículos seleccionados.

3.1.3.1.1. Tipos de documentos:

La clasificación de los documentos científicos examinados en la presente investigación revela una distribución entre artículos, contribuciones a conferencias y revisiones, así, en la Figura 5 se ilustra la proporción de cada tipo de documento dentro del conjunto de datos analizado, lo cual muestra que los artículos corresponden a la mayoría con un total de 27 documentos que representa el 75% del total, por otra parte, las contribuciones a conferencias suman 8 documentos que es equivalente al 22.2% del grupo, también, las revisiones se encuentran representadas por un único documento que corresponde al 2.8% del total, dicha distribución subraya la predominancia de artículos en el corpus de la investigación que destacan la importancia de las publicaciones en revistas como fuente primaria de conocimiento científico en el campo de estudio.

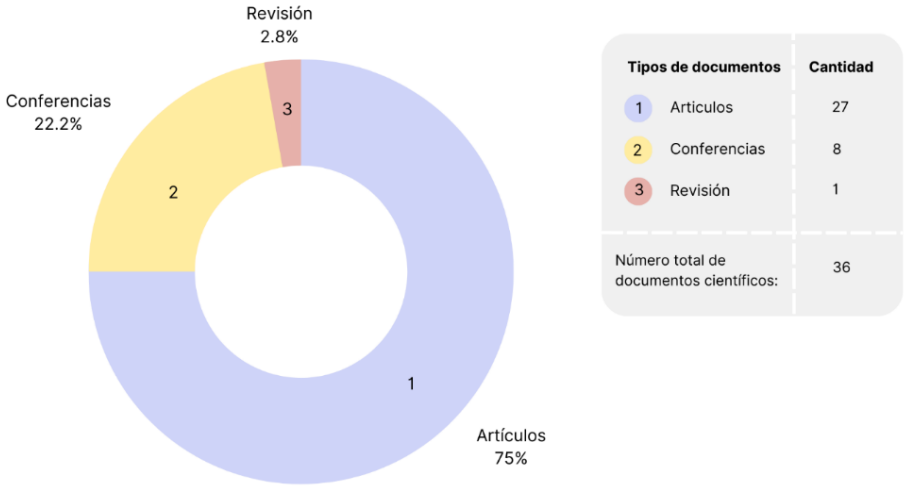

Figura 5. Tipos de documentos

3.1.3.1.2. Editoriales

Las editoriales desempeñan un papel esencial dentro del ecosistema de la investigación científica responsabilizándose de la publicación y propagación de documentos académicos siendo esencial para el proceso investigativo, ya que estas entidades se encargan de seleccionar, evaluar y publicar una variedad de contribuciones científicas abarcando desde artículos y ponencias en conferencias hasta revisiones y otros formatos, cuya misión principal es garantizar tanto la calidad como la difusión extensiva de estos trabajos asegurando su alineación con los estándares de exactitud académica.

En el contexto de este estudio, la Figura 6 ilustra el papel de las editoriales involucradas donde a la derecha del gráfico se puede observar cómo se enumeran las distintas editoriales que han aportado a esta investigación otorgando una panorámica de las editoriales participantes, por otra parte, al lado izquierdo se muestra el número de documentos publicados por cada una que proporciona una perspectiva detallada sobre su aportación al cuerpo de conocimiento revisado, así dicha representación simplifica la apreciación del significativo impacto y la importancia de las editoriales en el proceso de diseminación del conocimiento científico en el área estudiada.

Dentro de esta investigación son destacadas las editoriales como la Sociedad Coreana de Ingenieros Mecánicos que disponen de 9 documentos, el Instituto de Ingenieros Eléctricos y Electrónicos Inc. con 7 documentos y MDPI AG de 6 documentos que subrayan su contribución primordial al conjunto de datos analizado.

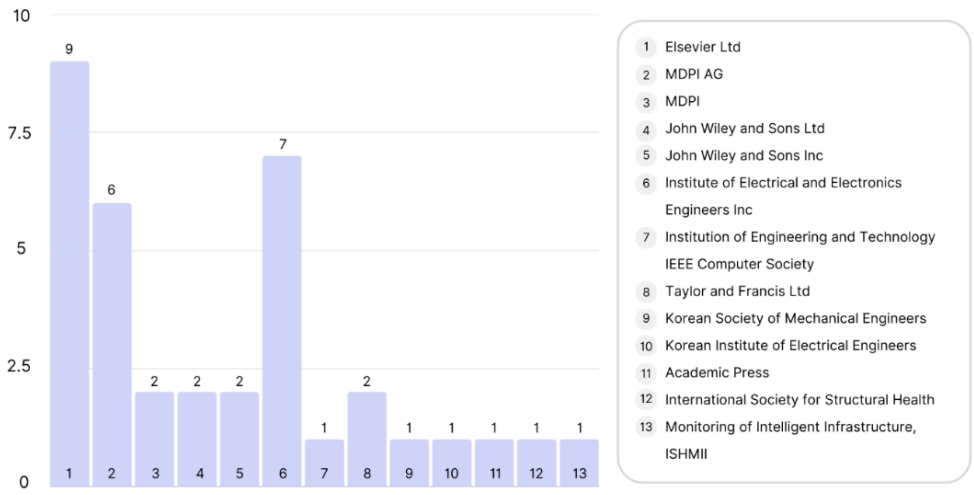

Figura 6. Editoriales

3.1.3.1.3. Documentos más citados

La Figura 7 presenta el nivel de impacto y la influencia de los trabajos más destacados en el ámbito académico acorde con la presente investigación, así, dicha representación hace énfasis en una visión detallada sobre los documentos que han sido ampliamente reconocidos y valorados por la comunidad científica basándose en el número de citas recibidas, lo que proporciona un indicador objetivo del impacto de cada documento en la literatura científica, por consiguiente, una inspección minuciosa del gráfico permite identificar aquellos trabajos que han ejercido más influencia y recibido un mayor reconocimiento siendo esencial para comprender el contexto global y las aportaciones fundamentales en el área de estudio de este proyecto.

La organización del gráfico estructura primeramente los artículos con menos de 20 citas que se marcan con círculos azules, aquellos con citas de entre 20 y 50 aparecen con círculos amarillos y los documentos con un número superior de citas se destacan con círculos naranjas, además cada círculo indica el número exacto de citas recibidas permitiendo una evaluación precisa del impacto.

Finalmente, para una identificación clara de qué artículo corresponde a cada cifra de citas se implementa una numeración del 1 al 36 donde dichos números se agrupan de manera directa con los documentos listados en la Tabla 4, donde se detallan todos los artículos examinados y aquella correlación entre números y documentos específicos simplifica el entendimiento de la procedencia de cada cita en el gráfico ofreciendo una guía visual para navegar a través de las contribuciones con mayor cantidad de influencia en el estudio.

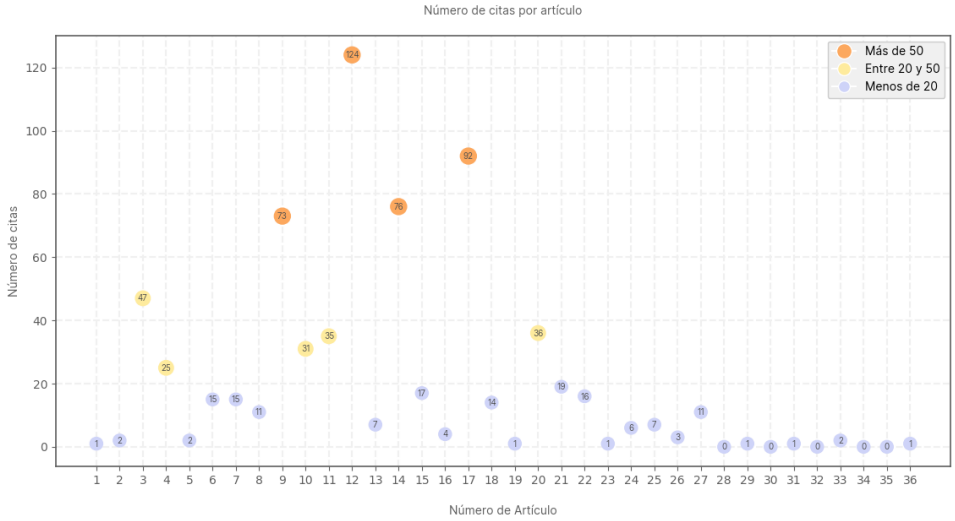

Figura 7. Documentos citados

3.1.3.1.4. Bases de datos indexadas

Las bases de datos indexadas desempeñan un papel primordial para facilitar un acceso rápido y eficiente a la información optimizando de manera significativa la eficacia en la búsqueda de documentos relevantes, así, en el contexto de la presente investigación se recurrió predominantemente a la base de datos Scopus para la recolección de información, no obstante, es importante señalar que los documentos seleccionados para este estudio también se encuentran disponibles en otras bases de datos de renombre como IEEE Xplore, ScienceDirect, Web of Science y Google Scholar garantizando una amplia cobertura y accesibilidad a la información científica pertinente tal como lo detalla la Tabla 9, que indica los documentos que están indexados en estas plataformas adicionales.

Tabla 9. Bases de datos indexadas

N°	Ref	Scopus	IEEE Xplorer	Science@ Direct	Web of Science	Google Scholar
1	[88]	x		x	x	x
2	[91]	x		x	x	x
3	[89]	x		x	x	x
4	[107]	x		x	x	x
5	[102]	x			x	x
6	[95]	x			x	x
7	[105]	x			x	x
8	[90]	x			x	x
9	[93]	x		x	x	x
10	[86]	x	x		x	x
11	[96]	x			x	x
12	[97]	x		x	x	x

N°	Ref	Scopus	IEEE Xplorer	Science@ Direct	Web of Science	Google Scholar
13	[101]	x	x		x	
14	[100]	x			x	x
15	[110]	x			x	x
16	[104]	x	x		x	x
17	[34]	x			x	x
18	[106]	x			x	x
19	[108]	x	x		x	x
20	[103]	x			x	x
21	[111]	x			x	x
22	[94]	x			x	x
23	[109]	x			x	x
24	[33]	x		x	x	x
25	[99]	x			x	x
26	[87]	x		x	x	x
27	[85]	x		x	x	x
28	[92]	x		x	x	x
29	[98]	x			x	x
30	[115]	x			x	x
31	[117]	x	x		x	x
32	[120]	x	x			x
33	[114]	x				x
34	[118]	x	x			x
35	[119]	x	x		x	x
36	[116]	x	x			x
Total		**36**	**9**	**10**	**32**	**35**

Número (N°); Referencia (Ref).

Los resultados revelan que Google Scholar lidera la lista como la base de datos que contiene la mayor cantidad de documentos relevantes para este estudio, que suman así un total de 35 siguiéndole de cerca Web of Science con 32 documentos identificados, por otra parte, ScienceDirect e IEEE Xplore presentan cifras semejantes en términos de documentos disponibles con recuentos de 10 y 9 respectivamente donde dichos resultados se encuentran sintetizados de forma visual en la Figura 8, proporcionando así una representación clara y concisa de la distribución de documentos por base de datos.

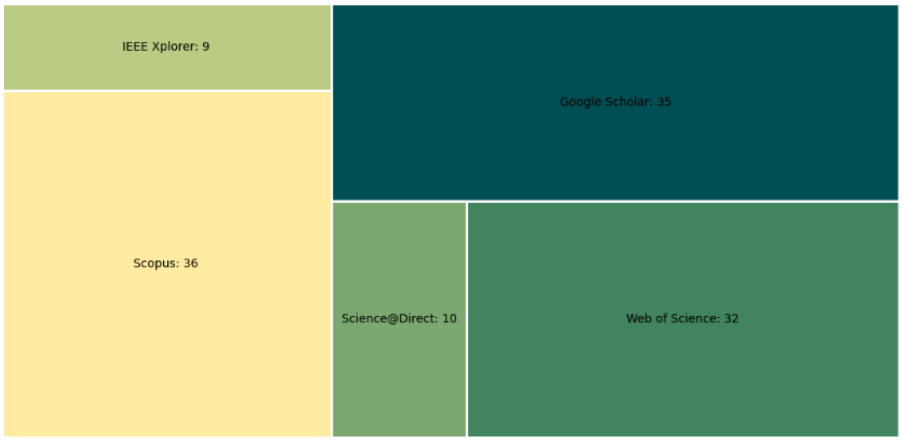

Figura 8. Documentos indexados

3.1.3.1.5. Relación entre los documentos consultados

Técnicas como Citation Gecko facilitan la gestión de referencias y la búsqueda de literatura científica relevante permitiendo efectuar búsquedas avanzadas en bases de datos académicas y otras fuentes de información científica como lo muestra Citation Gecko en la Figura 9, la cual usa los documentos seleccionados para extraer referencias y crear una red de citas visual.

En esta herramienta los nodos de color amarillo representan los documentos utilizados en la investigación actual que se conocen como semillas y se conectan con nodos grises que simbolizan las referencias, es decir, documentos relacionados que pueden ser relevantes para la investigación, por otra parte, la red visual facilita el análisis de las conexiones entre los documentos semilla y las referencias donde los nodos más grandes indican una mayor cantidad de conexiones con otras semillas, variando su tamaño según el número de estas conexiones, por último, los nodos grises más grandes o con más conexiones pueden ser especialmente relevantes para la investigación.

La Figura 9 destaca cinco agrupaciones principales donde cada una representa un grupo de documentos relacionados.

Figura 9. Visualización de la red de conexiones en Citation Gecko

La Tabla 10 proporciona información detallada sobre los cinco artículos más destacados, incluyendo el número de citas recibidas por cada uno, matizando que el número de citas de estos artículos supera el total de citas de los documentos seleccionados en esta RSL, como se ilustra en la Figura 7.

Tabla 10. Documentos más relevantes de Citation Gecko

Referencia	Título	Citaciones
[139]	Online wind turbine fault detection through automated SCADA data analysis	299
[83]	Using SCADA data for wind turbine condition monitoring a review	220
[140]	Anomaly detection and fault analysis of wind turbine components based on deep learning network	189
[141]	Machine learning methods for wind turbine condition monitoring: A review	448
[142]	Comparative analysis of neural network and regression based condition monitoring approaches for wind turbine fault detection	292

3.1.3.2. Presentación de los resultados de la RSL.

3.1.3.2.1. PI1: ¿Cuáles son las técnicas y modelos de inteligencia artificial más usados aplicados a la predicción de fallas de aerogeneradores?

La Tabla 11 categoriza las técnicas y modelos utilizados en los estudios revisados, organizándolos según sus similitudes y conexiones. Esta destaca la presencia de 94 técnicas/modelos distintos, mientras que la Tabla 5 señala un total de 103. La razón detrás de esta diferencia radica en el criterio de selección adoptado; en la cuenta de 94, solo se incluyen técnicas/modelos únicos mencionados en los artículos, omitiendo cualquier repetición. Esto significa que, si un estudio emplea varias variantes de un mismo algoritmo, estas variantes se consideran como una única entrada. Por lo tanto, la variación numérica se explica por este enfoque de contabilización, que busca prevenir el recuento múltiple de algoritmos ya citados por los mismos investigadores.

Tabla 11. Frecuencias de técnicas y modelos

Técnicas / Modelos	Referencias	Total
LR	[96], [102]	2
DEL	[106], [33]	2
XGBoost	[102], [107], [114]	3
SVR	[93], [110], [115]	3
NBM	[90], [105], [98]	3
LSTM	[107], [87], [116]	3
FNN	[33], [96], [105], [33]	4
PCA	[85], [99], [111], [116]	4
GMM	[88], [100], [101], [115]	4
CNN	[87], [89], [96], [110]	4
RF	[34], [102], [105], [109], [116]	5
DT	[102], [104], [34], [108], [116], [34]	5
AE	[90], [92], [99], [117], [118], [119]	6
KNN	[33], [34], [85], [109], [111], [117], [118]	7
ANN	[97] [94], [101], [103], [104], [109], [33], [120]	8
SVM	[34], [91], [102], [104], [105], [108], [109], [110], [111], [114], [116]	11
Otros	MSSM, MPE [91]; BiGRU-AM [89]; Catboost [102]; ML (Black Box) [95]; NARX [86]; GANs [104]; FA, CP [106]; CB [108]; RUSBoost [111]; Bagging, Boosting, Random subspace [33]; FS, FTL [87]; LoMST-CUSUM, WHC-LOF, CCA, KCPD [98]; t-SNE[118].	21

La Figura 10 proporciona un resumen visual de las técnicas y modelos identificados, resaltando las categorías más prevalentes. Notablemente, la

categoría ‘Otros’ incluye aquellos modelos mencionados solamente una vez y representa la diversidad de enfoques empleados. Entre las técnicas y modelos con frecuencias de mención más altas se encuentran: la Máquina de Vectores de Soporte (SVM) citada 11 veces (30.56%), Redes Neuronales Artificiales (ANN) 8 veces (22.22%), K Vecinos más Cercanos (KNN) 7 veces (19.44%), y Árboles de Decisión (DT) 5 veces (13.88%). Cada uno de estos representa un porcentaje significativo dentro del total de los 36 documentos examinados, destacando su relevancia y recurrencia en el campo de estudio.

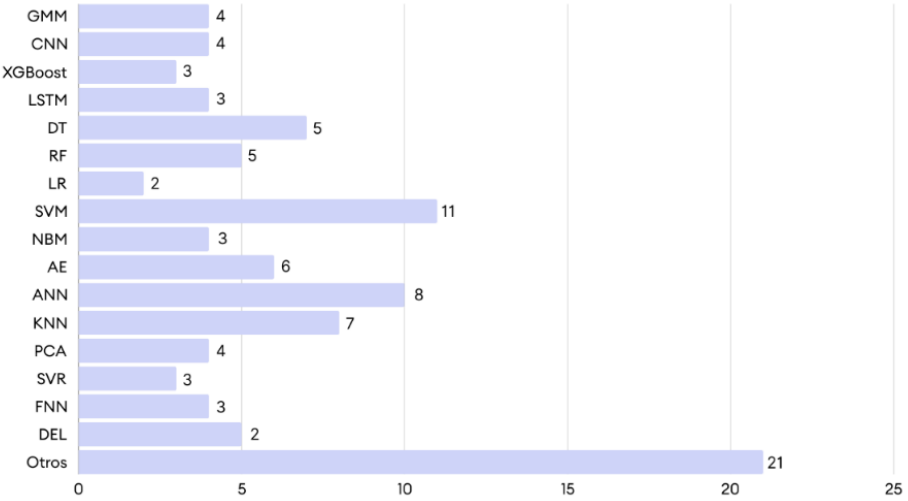

Figura 10. Modelos y técnicas encontrados

3.1.3.2.2. PI2: ¿Cuáles son los componentes del aerogenerador que más se estudian en la predicción de fallas?

La Tabla 12 muestra la frecuencia de los componentes de aerogeneradores más estudiados según lo identificado en los documentos seleccionados. Estos son destacados por los autores como focos principales de estudio y experimentación, siendo fundamentales para probar las técnicas y modelos descritos en la investigación.

Tabla 12. Componentes más estudiados

Componente	Referencias	Total
Hydraulic System	[98], [119]	2
Blade	[108], [115]	2
Converter	[33], [110]	2
Pitch	[99], [110]	2
Otros	Planetary [96]; Yaw [100]; Transformer [98]; Rotor [119].	4
Bearing	[89], [93], [105], [107], [96], [98]	6
Generator	[93], [98], [103], [105], [107], [110]	6

Componente	Referencias	Total
Gearbox	[86], [89], [92], [95], [98], [103], [105], [111], [119]	9
Wind Turbine	[34], [85], [87], [90], [91], [94], [97], [101], [102], [104], [106], [109], [114], [115], [116], [117], [118], [120]	18

Se destaca la coincidencia de elementos de la Tabla 7 con la Tabla 12, la cual contabiliza un total de 52 elementos de los 36 documentos analizados, esta información se encuentra resumida en la Figura 10, en donde se enfatiza que la predicción de fallas en aerogeneradores lidera con un 50% (18 veces), seguida por la caja de cambios con un 25% (9 veces), los rodamientos con un 19.44% (7 veces) y el generador con un 16.67% (6 veces). Finalmente, el campo “Otros” engloba a aquellos que se repiten una sola vez.

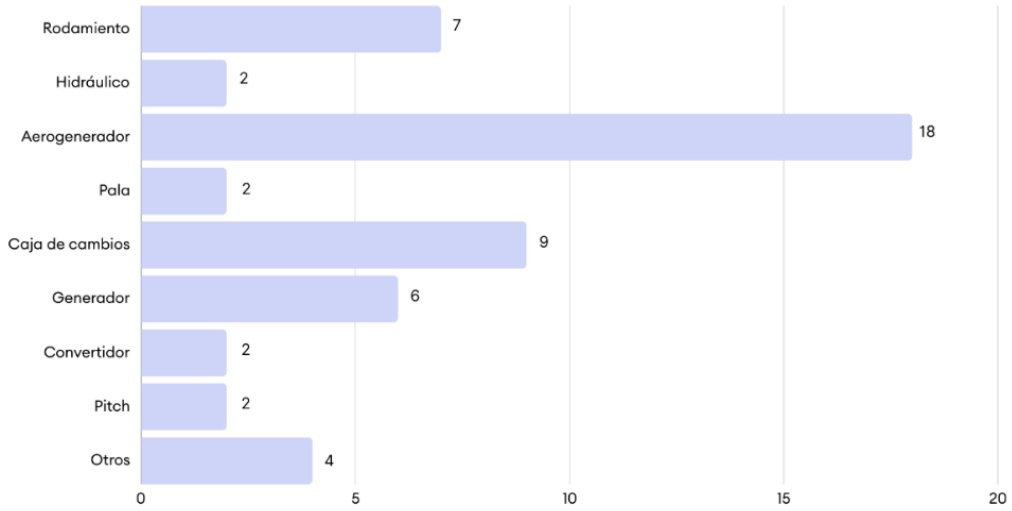

Figura 11. Frecuencia de componentes principales

3.1.3.2.3. PI3: ¿Cuáles son los artículos más importantes (usualmente más citados) sobre modelos híbridos de predicción de fallas en aerogeneradores en los últimos años?

La Figura 7 presenta los artículos más relevantes, destacando los que tienen mayor cantidad de citas, de esto modo se identifica cinco de ellos en color amarillo, dentro del rango de 20 a 50 referencias obtenidas, los que superan las 50 referencias se muestran en rojo, con un total de cuatro artículos, considerados los más notables.

La Tabla 13 resume los artículos más importantes, donde [97] sobresale con 124 citas, [104] con 92 citas, [100] con 76 citas y [93] con 73

citas, este conteo de referencias evidencia su relevancia en el campo de su estudio.

Tabla 13. Documentos más importantes

Ref	Título	Cantidad de citaciones
[97]	Deep Learning for fault detection in wind turbines	124
[104]	A Small-Sample Wind Turbine Fault Detection Method With Synthetic Fault Data Using Generative Adversarial Nets	92
[100]	SCADA-based wind turbine anomaly detection using Gaussian process models for wind turbine condition monitoring purposes	76
[93]	Real-time condition monitoring and fault detection of components based on machine-learning reconstruction model	73

1. Discusión de los resultados y conclusiones.

1.1. Discusión

En la presente RSL se analizaron diferentes técnicas y modelos utilizados para la predicción de fallas en aerogeneradores con base en la literatura científica seleccionada, se encontró que las técnicas más comúnmente utilizadas son ANN y SVM, representando 30.56% y 22.22% respectivamente. Estos resultados guardan relación con [121], [123], [143]. En [143], se menciona que aproximadamente en dos tercios de los casos en los que se utiliza el enfoque de aprendizaje automático, se emplean modelos de clasificación, principalmente ANN, SVM y DT, y el tercio restante se inclina hacia modelos de regresión. [123] destaca que ANN y SVM son los algoritmos de aprendizaje automático más populares y que han experimentado un rápido desarrollo en la última década. Adicionalmente, [121] enfatiza que ANN y SVM son las técnicas de IA más recurrentes, representando el 39% y el 27% de las aplicaciones revisadas.

El estudio [124] amplía el espectro de modelos populares de aprendizaje automático para la predicción de fallas en aerogeneradores, incluyendo ANN, SVR, GPR, DBN y K-NN, mientras que [144] recalca las SVM como una solución efectiva para otorgar tratamiento a datos de fallas desequilibradas, optimizando la precisión en las predicciones, de igual forma, [145] demuestra que las SVM aplicadas como modelo clasificador ofrecen los mejores resultados en la detección temprana de fallas en subcomponentes de aerogeneradores, con ello, este estudio de igual modo indica que SVR, RNN y DBN son los modelos menos tomados en consideración, con porcentajes del 23%, 21% y 15% respectivamente, lo cual coincide con los hallazgos de la presente

RSL, donde la SVR oscila con un escaso 8.33% de aplicación, haciendo énfasis en su limitado uso en contraste con otras técnicas.

En esta investigación se determinó que las predicciones de fallas en aerogeneradores radican principalmente en tres componentes los cuales corresponden al aerogenerador (50%), la caja de cambios (25%) y los rodamientos (19.44%), así, este hallazgo coincide con estudios previos como [121], que recalca que alrededor del 26% de la literatura objeto de estudio se concentra en la caja de cambios, a pesar de no ser el componente que mayores fallas presenta, es el que más contribuye en un valor aproximado del 20% del tiempo total de inactividad de los aerogeneradores, por otro lado, [97] detalla una tabla resumen de investigaciones que también plantea a la caja de cambios como el componente más estudiado, seguido de los rodamientos y bajo esa misma línea [146] resume trabajos importantes en el mantenimiento y pronóstico de fallas de componentes de energías renovables, haciendo énfasis en las técnicas aplicadas con mayor frecuencia a aerogeneradores, de este modo, dichos estudios comprueban la importancia de establecer esfuerzos en predecir y prevenir fallas en estos componentes considerados críticos para la eficiencia y durabilidad de los aerogeneradores.

1.2. Conclusiones

Con base en la literatura científica revisada por los diversos autores, se determina que los modelos y técnicas mayormente utilizadas para la predicción de fallas en aerogeneradores son SVM con 30.56% de apariciones, ANN con un 22.22% y KNN con un 19.44%, dichas técnicas se han presentado de la revisión de 36 documentos seleccionados y de 103 técnicas identificadas de las cuales se pueden comprender como enfoques que reflejan su capacidad y efectividad en atender el problema de predicción de fallas en aerogeneradores.

Los documentos más citados en la presente RSL corresponden a [97] con 124 citaciones y [104] con 92 citaciones de los cuales se han recibido un importante número de citas que destacan su gran relevancia y reconocimiento en el campo de la detección de fallas en aerogeneradores, no obstante, la mayoría de los documentos revisados en la RSL cuentan con más de 5 citaciones, como se detalla en la Figura 7 que indica su relevancia e impacto en el campo de la predicción de fallas en los aerogeneradores y demostrando que estos documentos han sido reconocidos y referenciados por otros investigadores y expertos en el campo confirmando su calidad y aporte en el área.

La predicción de fallas en aerogeneradores es una temática de gran interés y ampliamente investigada, tal como se detalla en el marco de la presente investigación

donde se identificó que la predicción aplicada a los aerogeneradores contiene un 34.6% de los artículos seleccionados, de igual modo, se observó que la caja de cambios (gearbox) representa un 17.3% de los estudios analizados, mientras que los rodamientos (Bearing) oscilan un 13.5% del total de los artículos seleccionados. Estos resultados otorgan la atención y relevancia que se ha dedicado a la predicción de fallas en aerogeneradores principalmente en lo que respecta a la caja de cambios y los rodamientos que son componentes críticos para el funcionamiento óptimo de los aerogeneradores y su monitoreo y predicción de fallas son fundamentales para prevenir elevados y costosos tiempos de inactividad y maximizar la producción de energía.

Los modelos híbridos identificados en Tabla 6 aprovechan las fortalezas individuales de las técnicas que los componen permitiendo abordar desafíos específicos y acoplarse a los requisitos particulares de cada aplicación, así, dichos modelos se benefician de la combinación de múltiples enfoques para adquirir un mejor rendimiento y alta precisión en el aprendizaje automático, además, la Figura 10 muestra la frecuencia de las técnicas y modelos de IA obtenidos en la RSL proporcionando una referencia útil al considerar la frecuencia de las técnicas y modelos de IA mencionados en la RSL para obtener una idea de las tendencias y enfoques más populares en el desarrollo de modelos híbridos que sirve como una primera guía para explorar eventuales combinaciones y enfoques en la construcción de nuevos modelos híbridos.

Anexo 2. Repositorios de GitHub

En los siguientes enlaces, se puede encontrar el código basado en las bibliotecas de Sklearn y Keras, en dichos enlaces, se trabaja con los Notebooks de Python para implementar los diversos algoritmos de machine learning y técnicas de combinación que forman parte del presente TIC.

Nota: Estos repositorios solo accesibles para el personal descrito en el acuerdo de confidencialidad. **Ver Anexo 4.**

- **GitHub: Computación UNL:**
    - https://github.com/Computacion-UNL/prediction_failure
- **GitHub: Wagner Cristhoper Castillo Castro:**
    - https://github.com/wagnercastillo/FailurePrediction-WTG

Anexo 3. Certificado de validación

Yo, Jorge Luis Maldonado Correa, director del proyecto de investigación MONITOREO INTELIGENTE DE LOS DATOS DEL SISTEMA SCADA DE LA CENTRAL EÓLICA VILLONACO PARA LA PREDICCIÓN DE FALLAS INCIPIENTES EN LOS AEROGENERADORES, código 36-DI-FEIRNNR-2021, informo que:

- Se proporcionó asesoría e información técnica al Sr. Wagner Cristhoper Castillo Castro, portador de la cédula de identidad Nro. 0705011898, para la elaboración de su trabajo de titulación denominado "Modelo híbrido de predicción de fallas para los aerogeneradores de la Central Eólica Villonaco utilizando inteligencia artificial y datos del sistema SCADA"
- Conozco los resultados alcanzados por el estudiante y puedo indicar que el desempeño de los algoritmos propuestos es similar al alcanzado por nuestro grupo de investigación.

Es todo cuanto puedo certificar en honor a la verdad, por lo que autorizo al interesado hacer usos del presente.

Loja, 25 de marzo de 2024

Firmado electrónicamente por:
JORGE LUIS MALDONADO CORREA

Ing. Jorge Luis Maldonado Correa
Docente de la carrera de Ingeniería Electromecánica

Anexo 4. Certificado de confidencialidad

Loja, 25 de marzo del 2024

Ing. Pablo Fernando Ordoñez Ordoñez
DIRECTOR DE LA CARRERA DE INGENIERÍA EN SISTEMAS/COMPUTACIÓN.

Certificado

Mediante la presente, en mi calidad de director del Trabajo de Integración Curricular titulado "**Modelo híbrido de predicción de fallas para los aerogeneradores de la Central Eólica Villonaco utilizando inteligencia artificial y datos del sistema SCADA**", llevado a cabo por el Sr. **Wagner Cristhoper Castillo Castro**, identificado con la cédula de ciudadanía Nro. **0705011898**, deseo hacer constar lo siguiente:

Todos los algoritmos desarrollados y los datos del sistema SCADA utilizados para la predicción de fallas han sido manejados bajo estricta confidencialidad y, por lo tanto, no pueden ser divulgados, salvo para mi persona y el Sr. **Wagner Cristhoper Castillo Castro**.

Esta declaración se hace conforme a la verdad y en cumplimiento de mis responsabilidades como director de carrera. Estoy disponible para cualquier aclaración adicional que pueda necesitarse.

Atentamente,

Ing. Pablo Fernando Ordoñez Ordoñez, Mg. Sc.
Director de la Carrera de Ingeniería en Sistemas/Computación

Anexo 5. Anteproyecto del trabajo de titulación

El anteproyecto de trabajo de titulación se puede visualizar accediendo al siguiente enlace.

- **Enlace:** Anteproyecto de trabajo de titulación

Anexo 6. Certificado de traducción del resumen

Loja, 29 de mayo del 2024

David Andrés Araujo Palacios
TRADUCTOR E INTÉRPRETE DE IDIOMAS (INGLÉS-ESPAÑOL-INGLÉS)

CERTIFICO:

Que se ha realizado la traducción de español a inglés del resumen derivado del trabajo de Integración curricular denominado **"Modelo híbrido de predicción de fallas para los aerogeneradores de la Central Eólica Villonaco utilizando inteligencia artificial y datos del sistema SCADA."** de autoría del estudiante **Wagner Cristhoper Castillo Castro** portador de la cédula de identidad número **0705011898**, estudiante de la **Carrera de Ingeniería en Computación** de la Universidad Nacional de Loja, bajo la dirección del **Ing. Pablo Fernando Ordóñez Ordóñez Mg. Sc.**

Es todo cuanto puedo certificar en honor a la verdad, facultando al interesado hacer uso del presente como considere.

David Andrés Araujo Palacios
Registro: MDT-3104-CCL-252098