



DIVISIÓN DE ESTUDIOS DE POSGRADO E INVESTIGACIÓN

OPCIÓN I.- TESIS

TRABAJO PROFESIONAL

**“DESARROLLO DE UN MODELO INTELIGENTE
PARA EL DIAGNÓSTICO DE FALLAS EN
GENERADORES DE POTENCIA”**

**QUE PARA OBTENER EL GRADO DE
MAESTRA EN INGENIERÍA INDUSTRIAL**

PRESENTA

I.Q. Quetzalli Amaya Sánchez

DIRECTOR DE TESIS

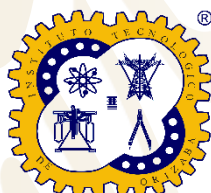
Dr. Alberto Alfonso Aguilar Lasserre

CO-DIRECTOR DE TESIS

M.I.I. Marco Julio del Moral Argumedo

CO-DIRECTOR EXTERNO DE TESIS

Dr. Gustavo Arroyo Figueroa



ORIZABA, VERACRUZ, MÉXICO

ABRIL 2024

Orizaba, Veracruz, **20/marzo/2024**
Dependencia: **División de Estudios de
Posgrado e Investigación**
Asunto: **Autorización de Impresión**
OPCION: I

C. AMAYA SÁNCHEZ QUETZALLI
Candidato(a) a Grado de Maestro(a) en:
INGENIERÍA INDUSTRIAL
PRESENTE

De acuerdo con el reglamento de Titulación vigente de los Centros e Institutos Tecnológicos Federales del Tecnológico Nacional de México, de la Secretaría de Educación Pública, y habiendo cumplido con todas las indicaciones que la Comisión Revisora le hizo respecto a su Trabajo Profesional titulado:

**"DESARROLLO DE UN MODELO INTELIGENTE PARA EL DIAGNÓSTICO DE FALLAS EN
GENERADORES DE POTENCIA"**

Comunico a usted que este Departamento concede su autorización para que proceda a la impresión del mismo.

ATENTAMENTE
Excelencia en Educación Tecnológica®
CIENCIA – TÉCNICA - CULTURA®



OFELIA LANDETA ESCAMILLA
JEFE DE LA DIVISIÓN DE ESTUDIOS
DE POSGRADO E INVESTIGACIÓN

CSR/magh



OG-13-F06





EDUCACIÓN
SECRETARÍA DE EDUCACIÓN PÚBLICA



TECNOLÓGICO
NACIONAL DE MÉXICO

Instituto Tecnológico de Orizaba
División de Estudios de Posgrado e Investigación

Orizaba Veracruz, **22/febrero/2024**
Asunto: **Revisión de trabajo escrito**

C. OFELIA LANDETA ESCAMILLA
ENCARGADA DE LA DIVISIÓN DE ESTUDIOS
DE POSGRADO E INVESTIGACIÓN
P R E S E N T E.-

Los que suscriben, miembros del jurado, han realizado la revisión de la Tesis del (la) C.:

AMAYA SÁNCHEZ QUETZALLI

La cual lleva el título de:

DESARROLLO DE UN MODELO INTELIGENTE PARA EL DIAGNÓSTICO DE FALLAS EN
GENERADORES DE POTENCIA

y concluyen que se acepta.

ATENTAMENTE
Excelencia en Educación Tecnológica®
CIENCIA – TÉCNICA - CULTURA®

PRESIDENTE: **DR. ALBERTO AGUILAR LASSERRE**

SECRETARIO: **MII JULIO DEL MORAL ARGUMEDO**

VOCAL: **DR. GUSTAVO ARROYO FIGUEROA**

VOCAL SUP.: **DR RAMIRO MEZA PALACIOS**

FIRMA

FIRMA

FIRMA

FIRMA

TA-09-F18



Av. Oriente 9 Núm.852, Colonia Emiliano Zapata. C.P. 94320 Orizaba, Veracruz.
Tel. 01 (272)1105360 e-mail: dir_orizaba@tecnm.mx | tecnm.mx | orizaba.tecnm.mx



2024
Felipe Carrillo
PUERTO
GOBIERNO DEL ESTADO DE VERACRUZ
GOBIERNO DEL INSTITUTO TECNOLÓGICO NACIONAL DE MÉXICO

Declaración de originalidad y cesión de derechos

Orizaba, Veracruz, el día 08 del mes de abril del año 2024.

El(la) que suscribe

C. Quetzalli Amaya Sánchez

Declaro que esta tesis, que tiene una extensión de 180 cuartillas, ha sido escrita por mí y constituye el registro escrito del trabajo de la tesis titulada

"DESARROLLO DE UN MODELO INTELIGENTE PARA EL DIAGNÓSTICO DE FALLAS EN GENERADORES DE POTENCIA"

del programa: Maestría en Ingeniería Industrial bajo la asesoría y dirección del (la) Dr. Alberto Alfonso Aguilar Lasserre y M.I.I. Marco Julio del Moral Argumedo no ha sido sometida en ninguna otra institución previamente.

Todos los datos y las referencias a materiales ya publicados están debidamente identificados con su respectivo crédito e incluidos en las notas bibliográficas y en las citas que se destacan como tal y, en los casos que así lo requieran, cuento con las debidas autorizaciones de quienes poseen los derechos patrimoniales. Por lo tanto, me hago responsable de cualquier litigio o reclamación relacionada con derechos de propiedad intelectual, exonerando de toda responsabilidad al Tecnológico Nacional de México campus Orizaba.

También declaro que, al presentar esta tesis, cedo los derechos del trabajo al Tecnológico Nacional de México campus Orizaba para su difusión, con fines académicos y de investigación, bajo las regulaciones propias de la institución y que si existe algún acuerdo de confidencialidad de la información lo haré saber en forma escrita para que se omitan las secciones correspondientes.

Los usuarios de la información no deben reproducir el contenido textual, gráficas o datos del trabajo sin el permiso expreso del autor y del director del trabajo. Este puede ser obtenido escribiendo a la siguiente dirección: depi_orizaba@tecnm.mx. Si el permiso se otorga, el usuario deberá dar el agradecimiento correspondiente y citar la fuente de este.



Quetzalli Amaya Sánchez

AGRADECIMIENTOS

Quiero expresar mi más sincero agradecimiento a todas las personas que contribuyeron de manera significativa en la realización de esta tesis. Sus invaluable aportes, apoyo y guía fueron fundamentales en cada etapa de este proceso.

En primer lugar, quiero expresar mi profunda gratitud a mi asesor, el Dr. Lasserre, por su guía experta, sus conocimientos y dedicación incansable a lo largo de este proyecto. Además, agradezco sinceramente su motivación constante. Su continua confianza en mis habilidades y su elogio constante de mi trabajo han sido un impulso invaluable para mi confianza y motivación durante este proceso.

También debo agradecer al Mtro. Marco Julio, por su valioso aporte y asesoramiento adicional. Su perspectiva única y sus sugerencias han enriquecido enormemente este trabajo.

Agradezco especialmente al Dr. Gustavo Arroyo por su valiosa retroalimentación, la cual ha contribuido significativamente a mejorar la calidad de esta investigación. Su participación activa en la gestión de la base de datos y su colaboración directa con el Instituto Nacional de Electricidad y Energías Limpias (INEEL) han sido pilares fundamentales para el desarrollo de este trabajo.

Agradezco a mi compañero Oscar Vélez, por su conocimiento en programación y su asesoramiento experto en el desarrollo del sistema inteligente. Su habilidad técnica y su disposición para compartir su experiencia han sido fundamentales para el éxito de este proyecto.

A mis respetados profesores de la maestría, cuya experiencia, enseñanzas y orientación han sido fundamentales en mi formación académica y profesional.

A mis amigos de maestría, quienes compartieron conmigo momentos académicos. Agradezco su compañerismo y apoyo incondicional.

Por último, agradezco al CONACYT, por el apoyo económico brindado.

A todas estas personas e instituciones, les estoy sinceramente agradecido por su contribución a este logro académico.

DEDICATORIAS

Con humildad y gratitud, dedico este trabajo a aquellos que han estado a mi lado en este viaje académico, compartiendo alegrías, superando desafíos y celebrando triunfos.

A mis amados padres, cuyo amor incondicional y apoyo constante han sido mi mayor fortaleza y guía en cada paso que he dado. Todo lo que he logrado en la vida se lo debo a ustedes. Gracias por ser mi mayor inspiración y por creer en mí incluso cuando yo dudaba de mí mismo. Su dedicación y sacrificio han sido el impulso detrás de mis logros. Los amo y admiro profundamente.

A mi abuelita chula, hermosa, preciosa, por su amor, sabiduría y cariño infinito. Su presencia ha sido una luz en mi camino, y agradezco sinceramente sus oraciones constantes a Dios, las cuales han sido un pilar espiritual importante en mi vida.

A mi hermano y mis adorables sobrinos, quienes son mi motivación y han llenado mi vida de alegría y momentos magníficos.

A mi extraordinaria familia, por su amor y apoyo absoluto. No puedo más que sentirme bendecido por tenerlos a mi lado, alentándome y aplaudiendo cada uno de mis logros. En particular, agradezco de manera especial a mi tía Cari, mi tío Saul y mi prima Joyce, su generosidad y hospitalidad han sido un regalo invaluable, permitiéndome ser parte de su familia y creando recuerdos que atesoraré para siempre.

A mis queridos amigos, por su amistad incondicional y por la hermosa presencia que han tenido en mi vida. Gracias por compartir conmigo momentos inolvidables. En especial, quiero agradecer a Gustavo, Frank, Adán y Germán. Cada uno de ustedes han contribuido a hacer de estos años los mejores momentos de mi vida. Gracias por su compañía, sus consejos, su paciencia, su apoyo y por ser la razón de tantas carcajadas. Los llevo siempre en mi corazón.

A mí mismo, por mi dedicación, mi perseverancia y mi determinación para alcanzar mis metas. Esta tesis es un recordatorio de mi capacidad para lograr grandes cosas cuando me comprometo con mis sueños.

RESUMEN

El aumento constante en la demanda de energía ha impulsado el crecimiento de los sistemas eléctricos de potencia, generando mayor complejidad en su operación y control. Los generadores eléctricos, esenciales para la conversión de energía mecánica a eléctrica, desempeñan un papel crucial en la seguridad y calidad del servicio de la red eléctrica. Ante la diversidad de fallas asociadas al envejecimiento y condiciones operativas, se vuelve crucial contar con herramientas que respalden las decisiones sobre mantenimiento: reparación, reacondicionamiento o reemplazo de estos generadores (Cinvestav, 2022). En este escenario, las tendencias actuales han resaltado el control tolerante a fallas mediante métodos de aprendizaje automático, brindando la capacidad de identificar fallos en los generadores en servicio y facilitando la optimización de costos, confiabilidad, disponibilidad y seguridad en la gestión de generadores de potencia (Ramirez Mongui & Dávila Arias, 2022).

En este documento se presenta el diseño y desarrollo de un modelo inteligente para diagnosticar las fallas que ocurren en generadores de potencia del sistema eléctrico nacional utilizando diferentes indicadores para determinar el nivel de degradación de los componentes del generador, principalmente el análisis de descargas parciales.

Para lograr esto, se emplea Anaconda y Python para establecer un entorno de trabajo, creando una base de datos previamente preprocesada con 196 casos y 57 características relacionadas con diversas fallas. Esta base de datos se utiliza para entrenar y evaluar diez algoritmos de clasificación de aprendizaje automático, destacándose el rendimiento óptimo del algoritmo Gradient Boosting. A través de la implementación de la Búsqueda Grid, se optimizan los hiperparámetros del algoritmo, resultando en un modelo con una precisión balanceada del 93%, un índice Kappa de 0.90, un índice de Jaccard de 0.86 y un coeficiente de correlación de Matthews de 0.91. La curva ROC presenta un macro promedio de 96% y la curva PR un micro promedio de 89%. Estos resultados confirman la eficacia del modelo en la clasificación de fallas en generadores de potencia.

Finalmente, la implementación del modelo culmina con la creación de una interfaz de usuario denominada Sistema Inteligente para el Diagnóstico de Fallas en Generadores de Potencia (SIDG), simplificando la interacción con el modelo y ofreciendo una solución accesible y eficiente a los usuarios finales.

ABSTRACT

The constant increase in energy demand has propelled the growth of power electrical systems, resulting in increased complexity in their operation and control. Electric generators, which are essential for converting mechanical to electrical energy, play a crucial role in the safety and quality of service in the electrical grid. Given the variety of failures associated with aging and operational conditions, it becomes crucial to have tools that support decisions regarding maintenance: repair, refurbishment, or replacement of these generators (Cinvestav, 2022). In this scenario, current trends have emphasized fault-tolerant control through machine learning methods, providing the ability to identify faults in generators in service and facilitating the optimization of costs, reliability, availability, and security in power generator management (Ramirez Mongui & Dávila Arias, 2022).

This document presents the design and development of an intelligent model to diagnose faults occurring in power generators of the national electrical system using different indicators to determine the level of degradation of generator components, primarily through partial discharge analysis.

To achieve this, Anaconda and Python are employed to establish a working environment, creating a previously preprocessed database with 196 cases and 57 features related to various faults. This database is used to train and evaluate ten machine learning classification algorithms, with the optimal performance of the Gradient Boosting algorithm standing out. Through the implementation of Grid Search, the hyperparameters of the algorithm are optimized, resulting in a model with a balanced accuracy of 93%, a Kappa index of 0.90, a Jaccard index of 0.86, and a Matthews correlation coefficient of 0.91. The ROC curve shows a macro-average of 96%, and the PR curve shows a micro-average of 89%. These results confirm the effectiveness of the model in classifying faults in power generators

Finally, the implementation of the model concludes with the creation of a user interface named the Intelligent System for Power Generator Fault Diagnosis (SIDG), simplifying the interaction with the model and providing an accessible and efficient solution to end-users.

CONTENIDO

AGRADECIMIENTOS.....	i
DEDICATORIAS.....	ii
RESUMEN.....	iii
ABSTRACT.....	iv
INTRODUCCIÓN.....	1
1. CAPÍTULO 1. GENERALIDADES	3
1.1 Introducción.....	3
1.2 Planteamiento del problema.....	3
1.3 Objetivos.....	4
1.3.1 Objetivo general	4
1.3.2 Objetivos específicos.....	4
1.4 Justificación	5
1.5 Hipótesis	6
1.6 Descripción de la empresa en colaboración.....	6
1.6.1 Misión.....	7
1.6.2 Visión	7
1.6.3 Política.....	7
1.6.4 Objetivos.....	8
1.7 Estado del arte.....	8
1.8 Conclusión	17
CAPÍTULO 2. MARCO TEÓRICO	18
2.1 Introducción.....	18
2.2 Generadores de potencia.....	18
2.2.1 Clasificación	18
2.2.2 Componentes principales.....	19
2.2.3 Ciclo de vida.....	25
2.2.4 Índice de Salud	27
2.2.5 Mecanismos de envejecimiento y modos de falla	28
2.2.6 Técnicas de diagnóstico.....	33
2.3 Inteligencia Artificial.....	37
2.3.1 Lógica Difusa	38
2.3.2 Machine Learning.....	40

2.4	Evaluación de un modelo de clasificación.....	61
2.4.1	Generalidades	61
2.4.2	Métricas de evaluación	62
2.4.3	Matriz de confusión.....	64
2.4.4	Curva PR	64
2.4.5	Curva ROC-AUC	65
2.5	Conclusión	66
CAPÍTULO 3. METODOLOGÍA.....		67
3.1	Introducción	67
3.2	Metodología básica	67
3.3	Análisis general.....	70
3.3.1	Definir el problema bajo el dominio	70
3.3.2	Búsqueda de soluciones previas	71
3.3.3	Determinar tipo de problema.....	71
3.3.4	Generación de supuestos	72
3.3.5	Propuesta	72
3.4	Obtención de datos.....	72
3.4.1	Determinar datos necesarios	73
3.4.2	Capturar u obtener datos.....	73
3.4.3	Crear entorno de trabajo	74
3.4.4	Datos.....	76
3.5	Exploración	77
3.5.1	Análisis de atributos	77
3.5.2	Aplicar estadísticas	80
3.5.3	Visualización	82
3.5.4	Información	84
3.6	Pre-procesamiento	84
3.6.1	Limpieza de datos.....	84
3.6.2	Ingeniería de características.....	85
3.6.3	Selección de características	88
3.6.4	Conocimiento	88
3.7	Elección algorítmica	88
3.7.1	Probar distintos algoritmos.....	89
3.7.2	Análisis de métricas.....	90

3.7.3	Seleccionar el mejor algoritmo	124
3.7.4	Experimento	124
3.8	Ajuste del modelo	128
3.8.1	Implementar búsqueda en rejilla.....	128
3.8.2	Entrenar modelo	129
3.8.3	Evaluar el modelo	129
3.8.4	Interpretación del modelo.....	133
3.8.5	Análisis de errores	136
3.8.6	Modelo.....	138
3.9	Conclusión	139
CAPÍTULO 4. RESULTADOS		140
4.1	Introducción	140
4.2	Implementación	140
4.2.1	Evaluación de las predicciones del modelo.....	140
4.2.2	Desarrollo de la interfaz gráfica	147
4.2.3	Sistema	148
4.3	Conclusión	154
CONCLUSIONES.....		155
RECOMENDACIONES		158
REFERENCIAS BIBLIOGRÁFICAS		159

ÍNDICE DE FIGURAS

Figura 2.1 Generador conectado en estrella	20
Figura 2.2 Generador conectado en delta	20
Figura 2.3 Generador de polos salientes y de rotor cilíndrico.....	21
Figura 2.4 Conmutador y montaje de portaescobillas	22
Figura 2.5 Excitatriz sin escobillas.....	23
Figura 2.6 Sistema de excitación estático.....	24
Figura 2.7 Etapas del ciclo de vida: a) Prepuesta en servicio, b) Puesta, y c) Fin de vida.....	26
Figura 2.8 El diagrama de bloques para desarrollar la metodología de lógica difusa muestra las tres etapas que constituyen el control. Tomando como entradas X (3A), se lleva a cabo el proceso conforme al método (3B), obteniendo una respuesta de salidas Y (3C)	40
Figura 2.9 Metodología ML	42
Figura 2.10 Red neuronal natural	44
Figura 2.11 RNA de propagación hacia adelante	45
Figura 2.12 SVM binario, el hiperplano (línea continua), el margen (entre dos líneas discontinuas) y los vectores de soporte (puntos encerrados en un círculo en la línea discontinua)	48
Figura 2.13 Conjunto de datos antes de LDA (izquierda) y los datos después de LDA (derecha)	51
Figura 2.14 Comparación entre LDA (izquierda) y QDA (derecha).....	52
Figura 2.15 Árbol de decisión de clasificación multiclase.....	53
Figura 2.16 Procedimiento del algoritmo Random Forest	56
Figura 2.17 Comparación de clasificadores en ML.....	60
Figura 2.18 Estructura de la matriz de confusión.....	64
Figura 2.19 Curva PR.....	65
Figura 2.20 Curva ROC.....	66
Figura 3.1 Diagrama de flujo de la metodología.....	69
Figura 3.2 Anaconda Navigator	75
Figura 3.3 Interfaz de Jupyter Notebook.....	76
Figura 3.4 Resultado del análisis estadístico	81
Figura 3.5 Mapa de calor de las variables numéricas.....	82
Figura 3.6 Distribución de frecuencia Diagnóstico	83
Figura 3.7 Gráfico circular Diagnóstico	83

Figura 3.8 Aplicación de One-Hot Enconding	86
Figura 3.9 Mapa de calor de la base de datos.....	87
Figura 3.10 Datos de entrenamiento y prueba.....	90
Figura 3.11 Matrices de confusión SVM	92
Figura 3.12 Curva ROC y AUC por SVM	93
Figura 3.13 Curva PR por SVM.....	94
Figura 3.14 Matrices de confusión Naive Bayes.....	96
Figura 3.15 Curva ROC y AUC por Naive Bayes	96
Figura 3.16 Curva PR por Naive Bayes	97
Figura 3.17 Matrices de confusión LDA.....	99
Figura 3.18 Curva ROC y AUC por LDA.....	100
Figura 3.19 Curva PR por LDA	100
Figura 3.20 Matrices de confusión QDA	102
Figura 3.21 Curva ROC y AUC por QDA	103
Figura 3.22 Curva PR por QDA.....	104
Figura 3.23 Matrices de confusión Decision Tree.....	106
Figura 3.24 Curva ROC y AUC por Decision Tree	106
Figura 3.25 Curva PR por Decision Tree	107
Figura 3.26 Matrices de confusión Random Forest	109
Figura 3.27 Curva ROC y AUC por Random Forest	110
Figura 3.28 Curva PR por Random Forest	110
Figura 3.29 Matrices de confusión Gradient Boosting.....	112
Figura 3.30 Curva ROC y AUC por Gradient Boosting	113
Figura 3.31 Curva PR por Gradient Boosting	114
Figura 3.32 Matrices de confusión XG Boost.....	116
Figura 3.33 Curva ROC y AUC por XGBoost.....	116
Figura 3.34 Curva PR por XGBoost	117
Figura 3.35 Matrices de confusión K-NN	119
Figura 3.36 Curva ROC y AUC por K-NN.....	120
Figura 3.37 Curva PR por K-NN.....	120
Figura 3.38 Matrices de confusión Artificial Neural Network.....	122
Figura 3.39 Curva ROC y AUC por Artificial Neural Network	123

Figura 3.40 Curva PR por Neural Network.....	124
Figura 3.41 Curva ROC - Macro Promedios.....	126
Figura 3.42 Curva PR - Micro Promedios.....	127
Figura 3.43 Código búsqueda en rejilla.....	128
Figura 3.44 Matrices de confusión Gradient Boosting optimizado.....	131
Figura 3.45 Curva ROC y AUC por Gradient Boosting optimizado	132
Figura 3.46 Curva PR por Gradient Boosting optimizado	132
Figura 3.47 Importancia de características Gradient Boosting optimizado	134
Figura 3.48 Primer árbol del modelo optimizado.....	135
Figura 4.1 Datos nuevos preprocesados para su predicción.....	142
Figura 4.2 Matrices de confusión de las predicciones del modelo.....	144
Figura 4.3 Curva ROC y AUC de las predicciones del modelo.....	144
Figura 4.4 Curva PR de las predicciones del modelo.....	145
Figura 4.5 Ícono de SIDG	148
Figura 4.6 Ventana principal de SIDG	149
Figura 4.7 Ventana Acerca de SIDG.....	149
Figura 4.8 Menú desplegable correspondiente a Enfriamiento	150
Figura 4.9 Menú desplegable correspondiente a Tipo	151
Figura 4.10 Diagnóstico clase 4	151

ÍNDICE DE TABLAS

Tabla 1.1 Revisión de la literatura aplicando técnicas de control y monitoreo	10
Tabla 1.2 Revisión de la literatura aplicando técnicas de control y monitoreo (continuación)....	11
Tabla 1.3 Revisión de la literatura aplicando RNA.....	12
Tabla 1.4 Revisión de la literatura aplicando LD.....	12
Tabla 1.5 Revisión de la literatura aplicando ML	13
Tabla 1.6 Revisión de la literatura aplicando ML (continuación).....	14
Tabla 1.7 Revisión de la literatura aplicando DL	15
Tabla 1.8 Revisión de la literatura aplicando DL (continuación).....	16
Tabla 2.1 Valores generales de un problema de clasificación.....	62
Tabla 2.2 Descripción de las métricas de evaluación.....	63
Tabla 3.1 Columnas del archivo y la característica que representa.....	78

Tabla 3.2 Columnas del archivo y la característica que representa (continuación).....	79
Tabla 3.3 Conteo de clases y categorías	79
Tabla 3.4 Algoritmos de aprendizaje supervisado para clasificar fallas en un generador de potencia	89
Tabla 3.5 Reporte de clasificación del algoritmo SVM	91
Tabla 3.6 Reporte de clasificación del algoritmo Naive Bayes.....	95
Tabla 3.7 Reporte de clasificación del algoritmo LDA.....	98
Tabla 3.8 Reporte de clasificación del algoritmo QDA	101
Tabla 3.9 Reporte de clasificación del algoritmo Decision Tree	105
Tabla 3.10 Reporte de clasificación del algoritmo Random Forest.....	108
Tabla 3.11 Reporte de clasificación del algoritmo Gradient Boosting.....	111
Tabla 3.12 Reporte de clasificación del algoritmo XG Boost	115
Tabla 3.13 Reporte de clasificación del algoritmo K – NN	118
Tabla 3.14 Reporte de clasificación del algoritmo Artificial Neural Network.....	121
Tabla 3.15 Resumen de análisis de métricas	125
Tabla 3.16 Ajuste de hiperparámetros en el algoritmo Gradient Boosting	129
Tabla 3.17 Reporte de clasificación del algoritmo Gradient Boosting optimizado.....	130
Tabla 3.18 Instancias equivocadas del modelo optimizado.....	137
Tabla 4.1 Reporte de clasificación de las predicciones del modelo Gradient Boosting.....	143
Tabla 4.2 Predicciones equivocadas del modelo desarrollado	146
Tabla 4.3 Recomendaciones para abordar las fallas en generadores de potencia	150
Tabla 4.4 Comparación del presente estudio con otras investigaciones similares que aplican ML	153

ÍNDICE DE ECUACIONES

Ecuación 1	49
Ecuación 2	52
Ecuación 3	52
Ecuación 4	54
Ecuación 5	54
Ecuación 6	54

ABREVIATURAS

AUC	Area Under the Curve
CFE	Comisión Federal de Electricidad
DL	Deep Learning
IA	Inteligencia Artificial
INEEL	Instituto Nacional de Electricidad y Energías Limpias
IS	Índice de Salud
K-NN	K-Nearest Neighbors
LDA	Linear Discriminant Analysis
LD	Lógica Difusa
MEM	Mercado Eléctrico Mayorista de México
ML	Machine Learning
PR	Presicion – Recall
QDA	Quadratic Discriminant Analysis
RNA	Red Neuronal Artificial
ROC	Receiver Operating Characteristic
SEN	Sistema Eléctrico Nacional
SIDG	Sistema Inteligente para el Diagnóstico de fallas en Generadores de Potencia
SVM	Support Vector Machines

INTRODUCCIÓN

Los sistemas eléctricos de potencia desempeñan un papel fundamental en el desarrollo y bienestar de la sociedad contemporánea, siendo las plantas generadoras el punto de partida donde la energía mecánica se transforma en energía eléctrica. En México, la Comisión Federal de Electricidad (CFE) juega un papel destacado al ser responsable de la generación del 54% de la energía del país, gestionando un total de 2,275 subestaciones de potencia. El restante 46% de la energía proviene de plantas generadoras privadas (CFE, 2023).

Un generador eléctrico se define como una máquina eléctrica rotatoria encargada de convertir la energía mecánica en energía eléctrica. Dada su función fundamental en las centrales de generación eléctrica, estos equipos exigen un elevado nivel de confiabilidad para asegurar la seguridad de la red y proporcionar una calidad de servicio óptima (Cinvestav, 2022). La falla de un generador en servicio puede generar inestabilidad en el sistema, con consecuencias que van desde la desactivación de subsistemas interconectados hasta la interrupción del suministro eléctrico, resultando en pérdidas económicas y, en situaciones críticas, incluso pérdidas humanas. Entre estas fallas, el cortocircuito entre aspiraciones del devanado del estator, causado por el envejecimiento del aislamiento, destaca como la más prevalente, representando más del 25% de las incidencias en generadores. Las descargas parciales, responsables del 22% de las fallas en devanados, se perfilan como una causa raíz significativa, subrayando la importancia de su cuidado en términos de operación y mantenimiento (Yang, Li, Ye, & Song, 2019).

Los generadores necesitan mantenimiento regular para asegurar su correcto funcionamiento. La falta de mantenimiento acelera su deterioro, aumentando la posibilidad de necesitar reemplazo. Aunque estos equipos están diseñados con confiabilidad, incluso pequeños problemas no detectados a tiempo pueden llevar a una falla total. De acuerdo con (QuimiNet, 2021), los costos de mantenimiento de generadores eléctricos varían entre 5,000.00 y 1,000,000,000.00 MXN, dependiendo del tipo de fallo, la pieza afectada, el modelo o tamaño del generador. Por ello, el monitoreo y diagnóstico de fallas en generadores de gran escala en las redes eléctricas está adquiriendo mayor importancia para prolongar su vida útil.

Una técnica de diagnóstico ampliamente estudiada es el análisis de descargas parciales, que se centra en la evaluación de la descarga parcial y su interconexión en cada análisis. Sin embargo, la interpretación de estos análisis suele depender en gran medida de la experiencia de expertos en

activos eléctricos para determinar con precisión la presencia de fallas en el equipo. Para mejorar el diagnóstico convencional, es esencial contar con mecanismos que simplifiquen la interpretación de los resultados de las técnicas de diagnóstico y permitan la clasificación de generadores y sus componentes según sus condiciones. Este enfoque facilitaría la comunicación de la condición de los generadores a los responsables de los activos, agilizando la toma de decisiones sobre acciones de mantenimiento, rehabilitación o reemplazo (Ramirez Mongui & Dávila Arias, 2022).

Este trabajo presenta el desarrollo de un sistema inteligente destinado al diagnóstico de fallas en generadores de potencia de centrales hidroeléctricas y termoeléctricas en el sistema eléctrico nacional (SEN). La metodología se basa en el análisis de descargas parciales y otras características críticas en la operación de los generadores, empleando técnicas de Machine Learning (ML). La colaboración en este proyecto se lleva a cabo en conjunto con el INEEL (Instituto Nacional de Electricidad y Energías Limpias), reconocido por su enfoque en la investigación para abordar las necesidades del sector eléctrico y energético del país, teniendo como usuario principal la CFE.

La estructura de este trabajo se inicia con la introducción a la problemática asociada a las fallas en generadores de potencia, resaltando la importancia de la detección temprana de mecanismos de degradación como una estrategia clave para prolongar su vida útil. Asimismo, se exponen los objetivos y la hipótesis de la investigación, mientras se revisa la literatura existente, destacando investigaciones previas que han abordado el diagnóstico y la detección de fallas en equipos eléctricos similares.

Dentro del marco teórico, se presentan los fundamentos técnicos y teóricos que respaldan este trabajo, explorando aspectos clave del funcionamiento de los generadores y el análisis de las descargas eléctricas, así como las técnicas de inteligencia artificial y algoritmos de aprendizaje automático.

En la metodología, se siguen una serie de etapas para desarrollar el modelo inteligente, el cual se optimiza para obtener un mejor rendimiento y se valida mediante métricas específicas.

Finalmente, se culmina con la implementación del modelo, del cual se deriva un sistema interactivo capaz de colaborar con expertos de la CFE, facilitando su aplicación efectiva en entornos prácticos.

CAPÍTULO 1. GENERALIDADES

1.1 Introducción

Este capítulo aborda la problemática asociada a las fallas en generadores de potencia, destacando la crucial importancia de la detección temprana de mecanismos de degradación como estrategia para extender su vida útil. Asimismo, se exponen los objetivos, la hipótesis, y se revisa la literatura existente, resaltando investigaciones previas que han abordado el diagnóstico y detección de fallas en equipos eléctricos.

1.2 Planteamiento del problema

Los generadores de potencia son esenciales para el funcionamiento eficiente de las plantas de generación. Aunque son componentes altamente confiables, la detección temprana de mecanismos de degradación es crucial para prevenir fallos y extender su vida útil. La gestión del ciclo de vida abarca etapas como especificación, montaje, puesta en marcha, operación, mantenimiento y desmantelamiento, siendo vital contar con herramientas adecuadas, especialmente en las fases finales, para evaluar el estado de salud de los generadores. Estas herramientas fundamentan decisiones sobre el mantenimiento: reparación, reacondicionamiento o reemplazo, guiadas por criterios de optimización de costos, confiabilidad, disponibilidad y seguridad.

La evaluación del estado o índice de salud (IS) implica identificar marcadores e índices que cuantifiquen la degradación de los componentes del generador. Las estrategias de IS incluyen inspección visual, pruebas eléctricas y mecánicas, el uso de técnicas en línea y/o fuera de línea, y el análisis de parámetros de operación y mantenimiento, respaldados por estrategias basadas en condiciones según estándares actuales y conocimiento experto (CIGRE, 2024).

Los generadores emplean sistemas de enfriamiento por aire, agua o hidrógeno, enfrentando esfuerzos eléctricos, mecánicos y térmicos debido a la operación cíclica vinculada al control del mercado eléctrico mayorista de México (MEM). La identificación de fallas, centrada en el análisis de descargas parciales, se basa en fenómenos que se consideran normales si fueron previstos en el diseño y anormales si no lo fueron. Este análisis ofrece información crucial para el diagnóstico de

generadores, destacando descargas a la ranura, delaminación entre aislamiento y conductor, entre bobinas en cabezales, superficiales/tracking por contaminación, y descargas internas en huecos del aislamiento principal.

Aunque el análisis de descargas parciales es una técnica bien establecida, la interpretación aún depende en gran medida de la evaluación de expertos en activos eléctricos. Por tanto, se propone la implementación de algoritmos de Inteligencia Artificial (IA) para mejorar la precisión del diagnóstico, ya que estos métodos ofrecen flexibilidad en el manejo de relaciones lineales y no lineales, resistencia al ruido y a la incertidumbre, y alto rendimiento en la identificación de fallas.

1.3 Objetivos

El propósito fundamental de la investigación es alcanzar los siguientes objetivos.

1.3.1 Objetivo general

Diseñar y desarrollar un modelo inteligente que permita diagnosticar las fallas que ocurren en generadores de potencia de centrales hidroeléctricas y termoeléctricas del SEN utilizando diferentes indicadores, como el análisis de descargas parciales, para determinar la clase de degradación de los componentes del generador.

1.3.2 Objetivos específicos

- Generar una base de datos de fallas en generadores asociada al análisis de descargas parciales, validada por los expertos en generadores.
- Diseñar y desarrollar un modelo inteligente basados en algoritmos de ML.
- Validar el desempeño del modelo inteligente mediante el porcentaje de precisión, la matriz de confusión, curva ROC-AUC y curva PR.
- Demostrar la capacidad de diagnóstico del modelo inteligente mediante la comparación de otros algoritmos de ML desarrollados para tal fin.
- Crear un sistema interactivo para facilitar la interacción y uso práctico del modelo inteligente, proporcionando una herramienta eficiente para los usuarios finales.

1.4 Justificación

Los generadores desempeñan un papel esencial en los sistemas eléctricos de potencia. Con la creciente complejidad y la elevada demanda de los sistemas industriales, la tolerancia a la degradación del rendimiento se reduce significativamente. Esto, a su vez, incrementa la importancia de detectar e identificar de manera temprana cualquier anomalía o posible falla. Implementar una operación tolerante a fallas se vuelve crucial para minimizar la degradación del rendimiento, reducir costos y prevenir situaciones peligrosas (Ruíz Mondragón, Mora Flórez, & Pérez Londoño, 2009).

A lo largo del tiempo, los generadores han experimentado cambios significativos. Los modelos antiguos, caracterizados por su baja capacidad y aislamiento a base de asfalto, utilizaban sistemas de enfriamiento con aire y operaban con paros y arranques mínimos, sosteniendo una carga prácticamente constante que les confería una vida útil de 40 a 50 años. En contraste, los generadores modernos, de mayor capacidad y con sistemas de aislamiento y enfriamiento mejorados, operan con gradientes eléctricos, térmicos, mecánicos y magnéticos más elevados. Aunque están sujetos a paros y arranques en diversas condiciones de potencia activa y reactiva, su vida útil se ha reducido a alrededor de 30 años.

No obstante, se ha comprobado que, mediante la detección temprana y la eliminación de los mecanismos de deterioro con un mantenimiento eficiente, la vida útil de los generadores modernos puede prolongarse. Este enfoque no solo optimiza el rendimiento, sino que también resulta en ahorros considerables, dado que los generadores representan uno de los componentes más costosos del sistema eléctrico (Letal, Sedding, & Sasic, 2018).

En este contexto, el IS emerge como una herramienta integral que fusiona toda la información proveniente de las técnicas de diagnóstico, ofreciendo un valor numérico único que refleja el estado general del activo en análisis y de cada uno de sus componentes clave (Cerón A. , Echeverry, Aponte, & Romero, 2015). Además, su utilidad se extiende a la gestión de activos, proporcionando respaldo en la toma de decisiones, orientando la planificación de estrategias de mantenimiento, sirviendo como indicador de riesgo y facilitando la estimación de la vida residual (Jürgensen, Godin, & Hilber, 2017).

Con las recientes tendencias hacia la digitalización, ha surgido un interés creciente en la implementación de estrategias de control tolerante a fallas basadas en métodos de aprendizaje inteligente. Estos métodos se destacan por su capacidad para determinar el IS de una máquina en funcionamiento, lo que posibilita la estimación del momento propicio para llevar a cabo su mantenimiento. Este enfoque permite tomar decisiones oportunas con el fin de prevenir fallas inesperadas, reducir los tiempos de inactividad y minimizar los costos totales a lo largo de su ciclo de vida.

Surge así la relevancia del mantenimiento con el innovador enfoque del aprendizaje automático (ML) para analizar y detectar áreas que requieren acciones de mantenimiento urgentes (Campus MVP, 2021). El uso de tecnologías como la inteligencia artificial (IA) y el aprendizaje automático (ML) posibilita mejoras sustanciales en la disponibilidad, control, gestión y operación de los activos.

1.5 Hipótesis

El modelo inteligente tendrá un alto desempeño en el diagnóstico de fallas en generadores de potencia del sistema eléctrico nacional utilizando diferentes indicadores, principalmente el análisis de descargas parciales, para determinar la degradación de sus componentes.

1.6 Descripción de la empresa en colaboración

El Instituto Nacional de Electricidad y Energías Limpias (INEEL) es un centro de investigación con más de 45 años de trayectoria, ubicado en Cuernavaca, Morelos, México. Este instituto se ha consolidado como una entidad dedicada a abordar las necesidades del sector eléctrico y energético. En el marco del presente proyecto, se establece una colaboración con el INEEL, cuyo enfoque principal es proporcionar soluciones para el sector energético de México. La CFE será el principal beneficiario de la propuesta desarrollada en este trabajo.

El INEEL destaca por su experiencia y conocimientos, llevando a cabo actividades de innovación, investigación y desarrollo tecnológico. Su labor se orienta a ofrecer soluciones integrales que abarcan toda la cadena de valor del sector energético, con especial énfasis en energías limpias y renovables, electricidad y tecnologías transversales.

A través de su experiencia, el INEEL contribuye a la confiabilidad, sustentabilidad, eficiencia y seguridad del SEN, respaldando así la transición energética soberana en México.

Según (Instituto Nacional de Electricidad y Energías Limpias, 2023), la misión, la visión, las políticas y los objetivos de esta institución se presentan a continuación.

1.6.1 Misión

La misión fundamental del INEEL es llevar a cabo investigación aplicada y desarrollo tecnológico para respaldar la soberanía energética y fomentar el desarrollo sostenible en los ámbitos de electricidad y energías limpias. Además, se compromete a impulsar el desarrollo sustentable de la industria energética a través de la innovación.

1.6.2 Visión

La visión del INEEL se centra en alcanzar el reconocimiento como líder en investigación aplicada y desarrollo tecnológico en los campos de electricidad y energías limpias. Aspira a recibir reconocimientos por sus contribuciones únicas que fortalecen la competitividad industrial. Estas contribuciones se materializan a través de la formación de ecosistemas de innovación y la creación de valor, integrando a los sectores gubernamentales, industriales y académicos. Este enfoque conjunto tiene como objetivo propiciar y potenciar la sustentabilidad de la industria energética.

1.6.3 Política

El compromiso del INEEL es cumplir con las necesidades y superar las expectativas de sus clientes a través de la innovación, eficacia y mejora continua de sus procesos. Este compromiso se lleva a cabo en concordancia con el marco normativo, legal y regulatorio aplicable en aspectos como la gestión de la calidad, el desarrollo sustentable, la igualdad laboral, la gestión ambiental, así como la seguridad y salud en el trabajo.

1.6.4 Objetivos

Los objetivos que el INEEL busca alcanzar son los siguientes:

1. Realizar investigación aplicada, desarrollo tecnológico e innovación con el propósito de fortalecer operativa y tecnológicamente a las Empresas Productivas del Estado, organismos públicos, empresas mixtas y compañías de capital nacional que participan en su cadena de suministro.
2. Fortalecer la vinculación del Instituto con instituciones, dependencias y entidades relacionadas con el sector energético, para respaldar el desarrollo de tecnologías críticas.
3. Contribuir a la formación de especialistas que puedan abordar las necesidades derivadas de la transición energética y el fortalecimiento del sector.
4. Garantizar la viabilidad financiera y el desarrollo sostenido de la Institución, con el fin de respaldar la Transición Energética Soberana en los ámbitos científico y tecnológico.

1.7 Estado del arte

En esta sección, la revisión de la literatura recopila información de investigaciones que han abordado el diagnóstico y detección de fallas en generadores de potencia, así como en sus componentes clave y otros equipos eléctricos. Se focaliza en el empleo de técnicas de IA y métodos específicos dentro del ámbito eléctrico.

La evolución del diagnóstico de fallas y la protección tiene raíces tan antiguas como las propias máquinas. Inicialmente, tanto fabricantes como usuarios de máquinas eléctricas confiaron en protecciones básicas como sobrecorriente, sobretensión y fallas a tierra para garantizar la seguridad y el funcionamiento confiable. No obstante, a medida que las funciones desempeñadas por estas máquinas se volvieron más complejas, surgió la necesidad de mejorar en el ámbito del diagnóstico de fallas. En la actualidad, diagnosticar las fallas desde su inicio se ha vuelto crucial, ya que el tiempo de inactividad no programado de la máquina puede alterar los plazos y ocasionar considerables pérdidas financieras (Nandi, Toliyat, & Li, 2006).

Se han propuesto diversos enfoques para llevar a cabo la detección y el diagnóstico de fallas, y en este sentido, (Hurtado Cortés, Villarreal López, & Villarreal López, 2016) realiza la siguiente clasificación:

- Métodos basados en modelos matemáticos del proceso: Estrategias que emplean modelos formulados a partir del conocimiento de las dinámicas involucradas en el proceso. Se fundamentan en la obtención de diferencias entre las salidas reales del proceso y las predicciones de un modelo del proceso, permitiendo inferir la presencia de una falla.
- Métodos a partir de modelos obtenidos de datos históricos del proceso: Estos enfoques implican construir modelos utilizando datos recopilados durante el funcionamiento previo del sistema. Se busca aprender patrones y comportamientos del sistema a partir de datos históricos, y se utilizan técnicas como redes neuronales o modelos difusos para construir modelos que puedan generalizar a partir de estos datos.
- Métodos a partir de datos del proceso: Estos enfoques también parten de grandes conjuntos de datos históricos, pero su perspectiva difiere al no buscar la obtención de un modelo del proceso. En su lugar, se centran en resolver problemas de clasificación. Se ha propuesto el uso de clasificadores difusos, análisis de componentes principales, redes neuronales artificiales, máquinas con vectores de soporte, funciones de base radial, entre otras.

Se ha explorado la literatura para analizar desarrollos y aplicaciones en la detección y diagnóstico de fallas en generadores mediante el empleo de técnicas de control y monitoreo. Un ejemplo significativo es el trabajo de (Yang, Li, Ye, & Song, 2019), quienes proponen una metodología de procesamiento de señales jerárquicas para monitorear la condición de los generadores y diagnosticar fallas. Esta metodología se basa en datos de forma de onda eléctrica sin procesar en redes eléctricas. Para su implementación, desarrollan un algoritmo jerárquico que localiza el sitio de la falla y monitorea la condición de los generadores en redes eléctricas, validándolo en una red estándar IEEE bajo diversos escenarios. Los resultados indican la capacidad de monitorear múltiples generadores utilizando un número limitado de sensores de forma de onda, lo que sugiere una reducción en los costos de mantenimiento y una mejora en la confiabilidad del sistema.

En una línea similar, (Campuzano Martínez, 2016) presenta un diagnóstico del aislamiento de cuatro generadores en una Central Geotérmica en México. Utilizando técnicas de evaluación en línea y fuera de línea, fundamenta su evaluación en el parámetro de descargas parciales. Los resultados obtenidos permiten la identificación de mecanismos de deterioro en los devanados, guiando la ejecución de acciones de mantenimiento y fortaleciendo la confiabilidad de los generadores de la central.

En la Tabla 1.1 se concentra la aportación de artículos antes mencionados donde hacen uso de técnicas de control y monitoreo para la detección y diagnóstico de fallas de generadores.

Tabla 1.1 Revisión de la literatura aplicando técnicas de control y monitoreo

Título del artículo	Autor(es)	Aportación	Técnica
Monitoreo de condición y diagnóstico de fallas de generadores en redes eléctricas	(Yang, Li, Ye, & Song, 2019)	Metodología de procesamiento de señales jerárquicas para el monitoreo de la condición del generador y el diagnóstico de fallas	Técnicas de monitoreo y algoritmo jerárquico
Diagnóstico de generadores eléctricos de potencia con técnicas de monitoreo en línea y fuera de línea	(Campuzano Martínez, 2016)	Metodología de diagnóstico de generadores de potencia empleando técnicas de monitoreo en línea y fuera de línea, y así determinar tanto el origen como el nivel de afectación del sistema aislante para tomar acciones de mantenimiento	Técnicas de monitoreo en línea y fuera de línea

Fuente: Elaboración propia

Además, (Medrano Hurtado, Pérez Tello, Gómez Sarduy, & Vera Pérez, 2016) proponen una metodología para la detección de fallas en rodamientos de un generador síncrono mediante el análisis de señales de vibración. Estas señales son capturadas a través de transductores de aceleración, como acelerómetros piezoeléctricos y acústicos, utilizando micrófonos omnidireccionales. El procesamiento de las señales de vibración se realiza mediante el algoritmo de densidad espectral de potencia. Al analizar las señales de vibración resultantes de rodamientos tanto sin fallas como con fallas artificiales, logran identificar de manera temprana las fallas en los rodamientos y establecer diagnósticos.

Por otro lado, en su estudio, (Flores & Asiaín, 2011) emplean una metodología para el diagnóstico de fallas en máquinas eléctricas rotatorias utilizando la técnica de espectros de frecuencia de bandas laterales de las corrientes de fase. Para validar esta técnica, establecieron sistemas de prueba en laboratorio para verificar las magnitudes de las bandas generadas por fallas tanto mecánicas como eléctricas. Los resultados obtenidos demuestran la eficacia de esta técnica en la detección de la presencia de fallas en las máquinas eléctricas.

En la Tabla 1.2 se concentra la aportación de artículos antes mencionados donde hacen uso de técnicas de control y monitoreo para la detección y diagnóstico de fallas de generadores.

Tabla 1.2 Revisión de la literatura aplicando técnicas de control y monitoreo (continuación)

Título del artículo	Autor(es)	Aportación	Técnica
Nueva metodología de diagnóstico de fallas en rodamientos en una máquina síncrona mediante el procesamiento de señales vibro-acústicas empleando análisis de densidad de potencia	(Medrano Hurtado, Pérez Tello, Gómez Sarduy, & Vera Pérez, 2016)	Metodología de diagnóstico empleando señales de vibración para procesar las señales de vibración empleando el algoritmo conocido como densidad espectral de potencia	Algoritmo de densidad espectral de potencia
Diagnóstico de fallas en máquinas eléctricas rotatorias utilizando la técnica de espectros de frecuencia de bandas laterales	(Flores & Asiaín, 2011)	Metodología para el diagnóstico de fallas en máquinas eléctricas rotatorias utilizando la técnica de espectros de frecuencia de bandas laterales de las corrientes de fase	Técnica de espectros de frecuencia de bandas laterales

Fuente: Elaboración propia

El uso de técnicas de IA para abordar la detección y el diagnóstico de fallas constituye un campo activo de investigación en la actualidad. Un ejemplo de esto es el trabajo de (Altaf, Al-Anbuky, & GholamHosseini, 2014), que presentan el diseño de una RNA distribuida supervisada capaz de identificar múltiples tipos de fallas y la ubicación del evento de falla en una red de motores industriales. Los resultados de la simulación demuestran que la técnica es efectiva en la identificación del tipo y la ubicación de eventos de falla dentro de una red de motores distribuidos.

La Tabla 1.3 concentra la contribución del artículo mencionado donde utiliza RNA para la detección y diagnóstico de fallas en motores.

Tabla 1.3 Revisión de la literatura aplicando RNA

Título del artículo	Autor(es)	Aportación	Técnica
Diagnóstico de fallas en una red de motores distribuidos usando una red neuronal artificial	(Altaf, Al-Anbuky, & GholamHosseini, 2014)	Una red neuronal artificial distribuida supervisada capaz de identificar múltiples tipos de fallas, así como la ubicación del evento de falla dentro de una red de motores industriales	RNA

Fuente: Elaboración propia

Los sistemas basados en Lógica Difusa (LD) han demostrado eficacia en la resolución de problemas de detección y diagnóstico de fallas. Una de las principales virtudes de estos sistemas reside en la interpretabilidad del diagnóstico obtenida al observar las diversas reglas activadas durante el proceso de inferencia.

En este contexto, (Melih, Veysi, Hakan, & Muhamment, 2022) proponen una metodología innovadora para analizar fallas en generadores de barcos. Integran una versión mejorada del método de toma de decisiones multicriterio, enriquecida con LD, logrando así desarrollar un mecanismo de decisión para priorizar las fallas del generador.

En la Tabla 1.4 se concentra la aportación del artículo antes mencionado donde hacen uso de la técnica LD para la detección y diagnóstico de fallas en generadores.

Tabla 1.4 Revisión de la literatura aplicando LD

Título del artículo	Autor(es)	Aportación	Técnica
Método mejor - peor mejorado difuso tipo 2 de intervalo para la evaluación de fallas de generadores diesel de barcos.	(Melih, Veysi, Hakan, & Muhamment, 2022)	Nueva metodología denominada método de toma de decisiones multicriterio mejor - peor enriquecida con LD de tipo 2.	LD

Fuente: Elaboración propia

En los últimos tiempos, la tecnología de IA ha experimentado un desarrollo acelerado, y se ha observado un interés creciente en la investigación de métodos basados en algoritmos de ML para la detección de fallas en maquinaria. En una investigación, (Quintanilha, y otros, 2021) emplean un algoritmo Random Forest para detectar y clasificar diversas fallas en un sistema de generación de energía en embarcaciones posicionadas dinámicamente, mejorando considerablemente la precisión y velocidad en comparación con métodos basados en reglas.

(Murad, y otros, 2021) presentan una comparación entre dos algoritmos de ML, Máquina de Soporte Vectorial (SVM) y K-Vecinos Más Cercanos (K-NN), basada en el análisis probabilístico de componentes principales para el diagnóstico efectivo de fallas de interruptor abierto y reducir el tiempo de duración del diagnóstico. Los métodos de diagnóstico de fallas se comparan mediante simulaciones y resultados experimentales utilizando un controlador de arreglo de puertas programable en campo. Tanto la simulación del sistema desarrollado como los resultados experimentales demuestran la eficiencia y precisión superiores del método de diagnóstico de fallas basado en SVM en comparación con el método basado en K-NN.

En la Tabla 1.5 se concentra la aportación de artículos antes mencionados donde hacen uso de algoritmos de ML para la detección y diagnóstico de fallas de equipos eléctricos.

Tabla 1.5 Revisión de la literatura aplicando ML

Título del artículo	Autor(es)	Aportación	Técnica
Detector / clasificador de fallas para generadores de energía de anillo cerrado usando aprendizaje automático	(Quintanilha, y otros, 2021)	Modelo para detectar y clasificar varios tipos de fallas en un sistema de generación de energía de embarcaciones	Random Forest
Diagnóstico de fallas de interruptor abierto del inversor multinivel de puente H en cascada en generadores de energía distribuida mediante algoritmos de aprendizaje automático	(Murad, y otros, 2021)	Modelos basados en el análisis probabilístico de componentes principales para el diagnóstico efectivo de fallas de interruptor abierto	SVM y K-NN

Fuente: Elaboración propia

Por su parte, (Cui, Weng, & Zhang, 2022) se presentan un modelo de ML estructurado en tres etapas diseñado para el diagnóstico de fallas en rodamientos de turbinas. Este enfoque se valida utilizando datos reales provenientes de conjuntos de datos diversificados, específicamente centrados en señales de vibración no estacionarias, obteniendo buenos resultados en el diagnóstico de falla de rodamientos de aerogeneradores.

(González, Piedrahita, & Castrillón, 2020) crean un conjunto de datos compuesto por 300 registros reales que contenían 10 variables independientes y una variable dependiente. Este esfuerzo permite el desarrollo de una metodología para predecir la parada de las máquinas generadoras en una central hidroeléctrica implementando la plataforma Weka y el algoritmo de clasificación J48, logrando predecir fácilmente futuros paros de máquinas.

En la Tabla 1.6 se concentra la aportación de artículos antes mencionados donde hacen uso de algoritmos de ML para la detección y diagnóstico de fallas de equipos eléctricos.

Tabla 1.6 Revisión de la literatura aplicando ML (continuación)

Título del artículo	Autor(es)	Aportación	Técnica
Un marco de extracción de características y aprendizaje automático para el diagnóstico de fallas en rodamientos	(Cui, Weng, & Zhang, 2022)	Modelo de ML de tres etapas para el diagnóstico de fallas en rodamientos de aerogeneradores	k-NN
Predicción de parada de máquinas generadoras en una central hidroeléctrica por medio de minería de datos	(González, Piedrahita, & Castrillón, 2020)	Modelo de ML para predecir paros de máquinas generadoras.	Bayesiano J48

Fuente: Elaboración propia

La detección de fallas en máquinas mediante algoritmos Deep Learning (DL) están siendo objeto de estudios activos. Un ejemplo destacado se encuentra en el trabajo de (Hweon Ki, Song Hyun, & Chang Lak, 2022), donde presentan un método para aprender el sonido del motor de un generador diesel instalado en una planta de energía nuclear después de la puesta en marcha. Utilizan un algoritmo de DL no supervisado y un novedoso enfoque para predecir fallas en el generador diesel. La implementación de dos algoritmos de DL, demuestra que el algoritmo de detección de anomalías no supervisado logra una precisión significativamente superior.

Además, (Li, Yu, Yu, & Wang, 2022) proponen un método activo de tolerancia a fallas con acción incremental para el sistema de control de fallas del sensor del generador de vapor de paso, basado en el algoritmo de DL por refuerzo, Deep Q – Network. Los resultados revelan que el algoritmo propuesto simplifica el sistema de detección de fallas, aunque se señalan algunas limitaciones.

En la Tabla 1.7 se concentra la aportación de artículos antes mencionados donde hacen uso de algoritmos de DL para la detección y diagnóstico de fallas de generadores.

Tabla 1.7 Revisión de la literatura aplicando DL

Título del artículo	Autor(es)	Aportación	Técnica
Propuesta de un nuevo método para el aprendizaje de sonidos de generadores diesel y detección de sonidos anormales utilizando un algoritmo de aprendizaje profundo no supervisado	(Hweon Ki, Song Hyun, & Chang Lak, 2022)	Método para aprender el sonido del motor de un generador diesel con un algoritmo de DL y un nuevo método para predecir la falla	Codificador automático de CNN y clasificación binaria
Sistema de control tolerante a fallas para generador de vapor de un solo paso basado en algoritmo de aprendizaje de refuerzo	(Li, Yu, Yu, & Wang, 2022)	Método activo de tolerancia a fallas con acción incremental para el sistema de control con fallas del sensor del generador de vapor de un solo paso	Deep Q-Network

Fuente: Elaboración propia

Por otro lado, (Gensheng & Jingquan, 2022) realizan un estudio comparativo de métodos de diagnóstico de fallas basado en DL aplicados a máquinas rotativas (motores eléctricos) en plantas de energía nuclear. Se seleccionaron cuatro modelos de DL: Red Neuronal Profunda de Avance (DFNN), Red Neuronal Convolucional (CNN), Red Neuronal de Unidad Recurrente Aislada (GRU) y Red Neuronal Convolucional Recurrente (CRNN). Los resultados revelan que el modelo CRNN alcanza una precisión destacada y un rendimiento superior en todos los casos de prueba.

En la Tabla 1.8 se concentra la aportación del artículo anterior donde hacen uso de algoritmos de DL para la detección y diagnóstico de fallas de generadores.

Tabla 1.8 Revisión de la literatura aplicando DL (continuación)

Título del artículo	Autor(es)	Aportación	Técnica
Un estudio comparativo de métodos de diagnóstico de fallas basados en aprendizaje profundo para máquinas rotativas en plantas de energía nuclear	(Gensheng & Jingquan, 2022)	Estudio comparativo de métodos de diagnóstico de fallas basados en DL	DFNN, CNN, GRU y CRNN

Fuente: Elaboración propia

Los esfuerzos en el campo del diagnóstico están orientados a abordar diversos problemas, algunos de naturaleza general y otros más específicos. A pesar de los avances, persisten desafíos sin resolver o parcialmente abordados, como la formulación de reglas heurísticas efectivas y la validación de resultados en condiciones de información incompleta o errónea.

En párrafos anteriores, se ha estructurado y comparado parte de la literatura relacionada con el monitoreo y diagnóstico de equipos eléctricos, considerando diversas características que podrían considerarse deseables en un sistema de detección y diagnóstico de fallas. Estos análisis revelan que el empleo de técnicas de IA para abordar la detección y diagnóstico de fallas constituye un área de investigación dinámica, como evidencia la amplia literatura disponible que presenta propuestas metodológicas y aplicaciones derivadas de su aplicación.

La evolución de las contribuciones al conocimiento para abordar el diagnóstico del IS de un generador de potencia se ha desarrollado a lo largo del tiempo. Inicialmente, se emplearon métodos convencionales, tanto en línea como fuera de línea, para el monitoreo y evaluación de la condición del equipo, destacando el seguimiento en línea por su capacidad para mantener la operación continua del generador. Sin embargo, ambos métodos presentan la desventaja de un proceso de implementación tedioso.

Con el avance de las tecnologías, surgieron enfoques adicionales basados en la utilización de sensores, sistemas de recopilación y tratamiento de señales para evaluar la condición de salud de los equipos. A pesar de estos avances, persisten deficiencias que requieren más investigación, como la confiabilidad de la recopilación de señales de sensores y la precisión en el tratamiento y análisis de datos para lograr diagnósticos más completos y precisos.

Con la popularización de las técnicas de IA, se ha ampliado la investigación en el monitoreo de condición y diagnóstico para los generadores. La LD se ha destacado por su idoneidad en el análisis cualitativo de sistemas complejos, aunque presenta dificultades en la predicción de tendencias de falla, y la precisión del diagnóstico depende de la selección adecuada de parámetros. Por otro lado, las RNA son altamente aplicables en sistemas donde extraer conocimiento de especialistas es complejo, pero su principal limitación radica en la necesidad de grandes cantidades de datos para garantizar la confiabilidad del diagnóstico de fallas.

Las estrategias futuras de detección y diagnóstico deben centrarse en mejorar la robustez y adaptabilidad, permitiendo la continua evolución del sistema. Además, se busca alcanzar diagnósticos interpretables por el operador, identificando las ubicaciones de las fallas y sus posibles causas (Hurtado Cortés, Villarreal López, & Villarreal López, 2016).

En el contexto de las tendencias actuales de digitalización, el uso del mantenimiento predictivo para monitorear las condiciones y evaluar el rendimiento de los equipos se presenta como una estrategia prometedora. Por lo que, se hace necesario contar con modelos inteligentes que permitan diagnosticar fallas en los equipos eléctricos de manera oportuna, apoyando la toma de decisiones relacionadas con la operación y el mantenimiento de los equipos del SEN.

1.8 Conclusión

La investigación sobre el diagnóstico de fallas en generadores de potencia ha evolucionado a lo largo del tiempo, destacando la importancia de la detección temprana de degradaciones para prevenir fallos y prolongar la vida útil de estos componentes.

La revisión de la literatura ha destacado avances significativos en el campo, desde métodos convencionales hasta enfoques basados en IA, sin embargo, se debe mejorar en robustez y adaptabilidad. Para superar este desafío, se propone diseñar y desarrollar un modelo inteligente basado en algoritmos de ML para diagnosticar fallas en generadores, que contribuya a mejorar la eficiencia y confiabilidad de los generadores de potencia en el SEN.

CAPÍTULO 2. MARCO TEÓRICO

2.1 Introducción

Este capítulo presenta los fundamentos técnicos y teóricos esenciales empleados en la presente investigación. Comienza con una exposición de las generalidades asociadas a los generadores de potencia, abordando aspectos clave de su funcionamiento y relevancia en el contexto de la generación eléctrica. Posteriormente, se profundiza en el análisis de descargas parciales, destacando su importancia como indicador de posibles fallas en los generadores.

Finalmente, el capítulo concluye con la exploración de técnicas de IA y algoritmos de ML. Esta última sección se centra en proporcionar una comprensión integral de las herramientas computacionales utilizadas para potenciar el diagnóstico de fallas en generadores de potencia, destacando la relevancia de la IA en este contexto específico.

2.2 Generadores de potencia

Un generador eléctrico constituye un dispositivo capaz de transformar energía mecánica en energía eléctrica, desempeñando un papel fundamental en los sistemas eléctricos de potencia. Dada su relevancia, se impone la necesidad de asegurar un elevado nivel de confiabilidad, fundamental para salvar la integridad de la red y proporcionar un servicio eléctrico de calidad óptima.

2.2.1 Clasificación

2.2.1.1 Generadores de corriente alterna

Existen dos categorías principales de generadores de corriente alterna: los generadores síncronos y los generadores de inducción, también conocidos como generadores asíncronos. Los generadores síncronos son dispositivos cuya corriente de campo magnético se alimenta mediante una fuente externa de potencia de corriente alterna. Por otro lado, los generadores de inducción son máquinas cuya corriente de campo magnético se suministra a través de inducción magnética, es decir, mediante acción transformadora en sus devanados de campo (Enríquez Harper, 2004).

2.2.1.2 Generadores de corriente directa

Los generadores de corriente continua son dispositivos que transforman la energía mecánica en energía eléctrica de corriente continua. En su mayoría, estos generadores presentan similitudes con los generadores de corriente alterna en términos de tener voltajes y corrientes de corriente alterna en su interior, sin embargo, la característica distintiva de los generadores de corriente continua radica en la presencia de un mecanismo que convierte los voltajes internos de corriente alterna en voltajes de corriente continua en sus terminales de salida (Enríquez Harper, 2004).

2.2.2 Componentes principales

De acuerdo con las investigaciones de (Enríquez Harper, 2004), los elementos fundamentales de un generador incluyen el estator, el rotor, el sistema de enfriamiento, la excitatriz, el conmutador, las chumaceras, los componentes mecánicos y el regulador. A continuación, se presenta una breve descripción de cada uno de estos componentes.

2.2.2.1 Estator

A continuación, se detallan los aspectos cruciales a tener en cuenta como parte de los elementos que componen el estator de un generador.

- Elementos mecánicos: Los elementos mecánicos de un generador abarcan la carcasa, el núcleo, las bobinas y la caja de terminales, siendo estos fundamentales para su funcionamiento y estructura.
- Sistemas de conexión en estrella: Los devanados del estator de un generador suelen conectarse en forma de estrella. La Figura 2.1 ilustra el esquema de un generador con esta configuración en estrella, destacando la disposición de las conexiones.
- Sistema de conexión en delta: La conexión en delta se logra al unir las terminales 1 a 6, 2 a 4 y 3 a 5, mientras que las terminales de línea se conectan a 1, 2 y 3. Aunque esta configuración resulta en un voltaje menor en comparación con la conexión en estrella, se traduce en un aumento de la corriente de línea. La Figura 2.2 presenta el esquema de un generador conectado en delta para una visualización clara de esta disposición.

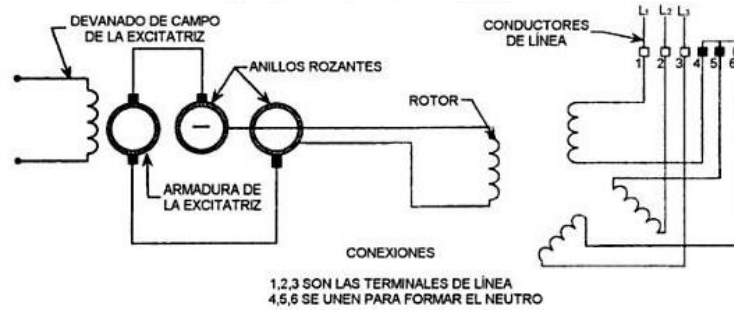


Figura 2.1 Generador conectado en estrella

Fuente: (Enríquez Harper, 2004)

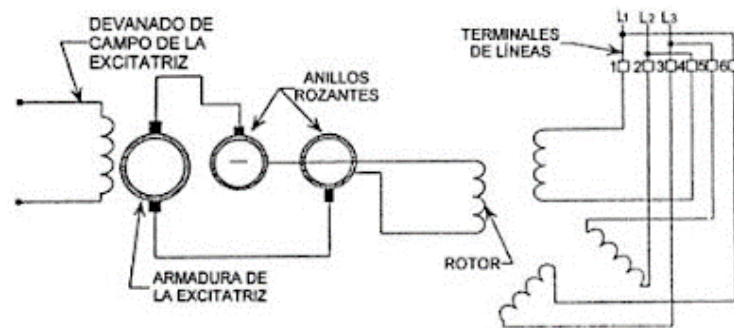


Figura 2.2 Generador conectado en delta

Fuente: (Enríquez Harper, 2004)

2.2.2.2 Rotor

Para generar el campo magnético en el rotor, se emplean polos compuestos por conjuntos de láminas de hierro magnético, diseñados para generar corrientes circulantes. Estos polos cuentan con conductores de cobre enrollados alrededor del hierro y son excitados por una corriente continua. Los polos del rotor se disponen en pares, posicionados a 180° uno del otro. En el rotor, se alojan las bobinas del devanado de campo, las cuales inducen el voltaje en el devanado de armadura, donde se encuentran las bobinas que determinan si el generador es monofásico o trifásico.

La Figura 2.3 presenta la clasificación de los generadores según el tipo de rotor para una mejor comprensión visual.

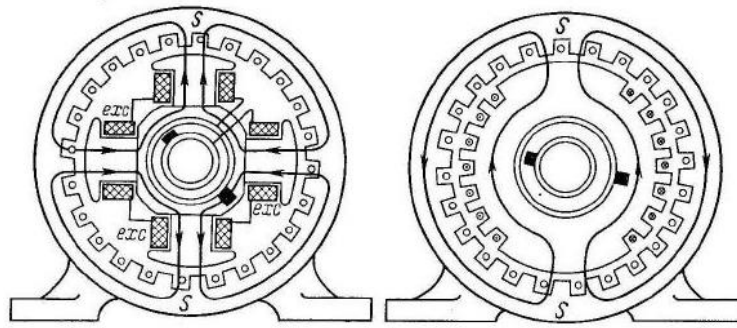


Figura 2.3 Generador de polos salientes y de rotor cilíndrico

Fuente: (Enríquez Harper, 2004)

2.2.2.3 Sistemas de enfriamiento

Los métodos de enfriamiento complementarios aplicados en los generadores de corriente alterna son el enfriamiento por aire, la combinación de aire y agua con intercambiador de calor, y el enfriamiento mediante el uso de agua con intercambiador de calor.

A continuación, se proporciona una descripción detallada de cada uno de estos métodos.

- **Enfriamiento por aire:** Los generadores que adoptan este método de enfriamiento utilizan aire del entorno a temperatura ambiente como medio de disipación térmica. El aire fluye a través del estator y el rotor gracias a ventiladores ubicados en ambos extremos del rotor. El aire caliente se expulsa por la parte posterior del generador, completando así un ciclo de circulación.
- **Intercambiador de calor aire-aire:** En contraste con los generadores de enfriamiento natural, aquellos con un intercambiador de calor aire-aire recirculan constantemente el mismo aire a través del estator. Este método presenta la ventaja adicional de mantener limpios los aislantes, ya que el aire se renueva de manera constante, eliminando la necesidad de utilizar filtros de aire en el sistema.
- **Intercambiador de calor aire-agua:** Diferente al sistema de intercambio de calor aire-aire, un generador con un intercambiador de calor aire-agua canaliza el calor proveniente del rotor y el estator a través de un enfriador compuesto por varios tubos de cobre perforados dispuestos alrededor del diámetro exterior de dichos tubos. Este método requiere una fuente de agua para el enfriamiento, que circula a través de estos intercambiadores. Esta

configuración tiene la ventaja de prevenir la introducción de contaminantes en los devanados del generador.

2.2.2.4 Excitatriz

Los generadores de potencia incorporan diversas excitatrices con el propósito de generar la corriente de excitación requerida en el rotor y, finalmente, inducir la corriente eléctrica en el estator. Cada variedad de excitatriz exhibe atributos distintivos y aplicaciones particulares, y la elección entre ellas está condicionada por factores como las dimensiones del generador, la aplicación específica y los criterios de regulación de la tensión.

A continuación, se describen los tipos principales de excitatrices empleados en generadores de potencia.

a) Excitatriz de corriente continua

En sistemas más compactos, como grupos electrógenos de menor escala o plantas de emergencia, se emplean sistemas de excitación basados en pequeños generadores de corriente continua directamente acoplados al eje del generador principal. Estos generadores de corriente continua, en realidad, generan corriente alterna y la rectifican mediante un conmutador equipado con escobillas deslizantes, como se ilustra en la Figura 2.4, la cual muestra las partes del conmutador y el montaje del portaescobillas.

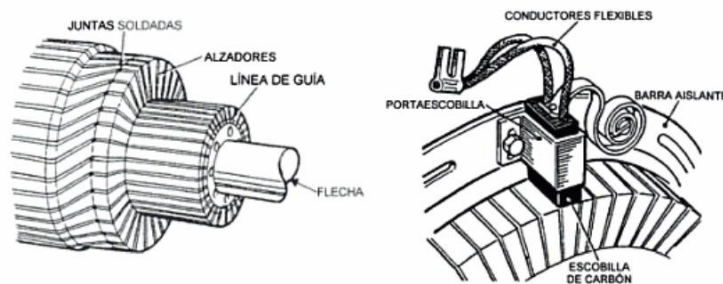


Figura 2.4 Conmutador y montaje de portaescobillas

Fuente: (Enríquez Harper, 2004)

Los sistemas de excitación basados en generadores de corriente continua plantean un desafío significativo debido a su necesidad de mantenimiento más frecuente en comparación con otros sistemas de excitación, como los rotatorios sin escobillas o aquellos de naturaleza estática. Esta limitación ha llevado a que, para generadores de dimensiones relativamente amplias, no se recomienda su empleo en servicios de emergencia. Además, desde una perspectiva operativa, estos sistemas presentan inconvenientes adicionales que, en última instancia, podrían afectar la calidad de la forma de onda.

b) Excitatriz sin escobillas

Otro método de excitación, conocido como excitatriz sin escobillas, opera bajo el mismo principio mencionado anteriormente, pero con la característica distintiva de prescindir de las escobillas y el conmutador. En esencia, la excitatriz se configura como un generador de corriente alterna con polos de campo estacionarios; el voltaje generado en corriente alterna en los devanados rotatorios se rectifica mediante diodos ubicados en la estructura giratoria. La tensión resultante de los rectificadores rotatorios se aplica directamente a los devanados de campo del generador, y el regulador de voltaje controla la corriente de campo para lograr el voltaje deseado en las terminales. La Figura 2.5 ilustra el funcionamiento de la excitatriz sin escobillas.

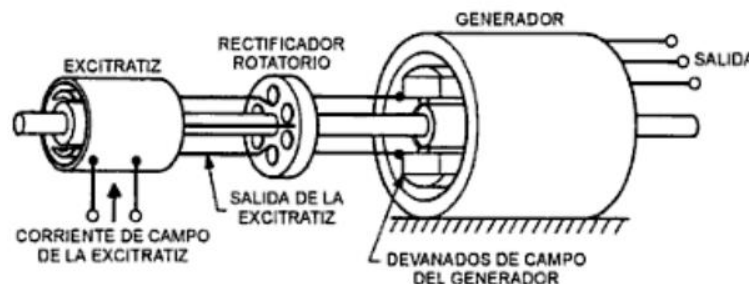


Figura 2.5 Excitatriz sin escobillas

Fuente: (Enríquez Harper, 2004)

c) Excitatriz estática

Un sistema de excitación adicional, conocido como excitatriz estática, prescinde de la excitatriz rotatoria, manteniendo los anillos rozantes y las escobillas, y obtiene la potencia para el campo directamente desde la salida del generador.

El sistema de control de voltaje gestiona la corriente de campo para regular el voltaje del generador y garantiza una respuesta más rápida en comparación con otros sistemas de excitación.

La Figura 2.6 ilustra un ejemplo de sistema de excitación estática.

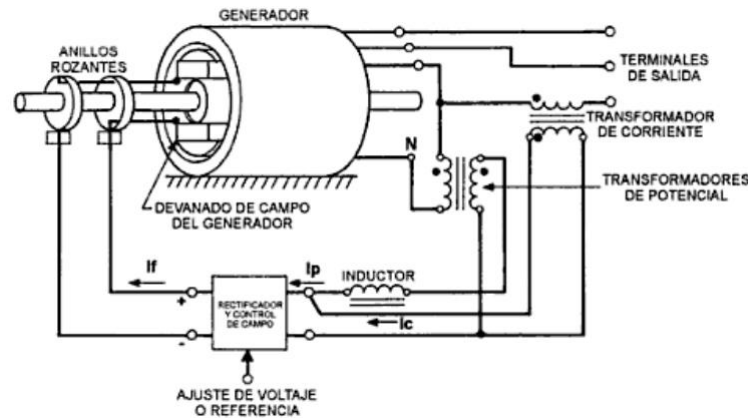


Figura 2.6 Sistema de excitación estático

Fuente: (Enríquez Harper, 2004)

2.2.2.5 Chumaceras y componentes mecánicos

Las chumaceras, cojinetes o rodamientos desempeñan un papel crucial en cualquier máquina rotativa, ya que sostienen los rotores. Estos no solo proporcionan soporte a un eje de rotación, sino que también minimizan la fricción durante el movimiento rotativo. La clasificación de las chumaceras depende de la orientación del eje, distinguiéndose entre horizontales y verticales. Asimismo, se categorizan en hidrodinámicas, ideales para bajas velocidades, e hidrostáticas, diseñadas para altas velocidades y cargas pesadas.

En cuanto a los componentes mecánicos, se incluyen los anillos colectores, especialmente en generadores con excitación estática. Estos anillos facilitan la conexión eléctrica entre un componente fijo, el sistema de excitación y un componente rotativo, los devanados del rotor. Los anillos colectores se sitúan en la flecha del rotor, en el lado de excitación, y se conectan a los devanados del rotor mediante guías aisladas.

2.2.2.6 Regulador

La principal tarea de un regulador consiste en emplear las señales de retroalimentación proporcionadas por los transformadores de instrumento, de potencial y corriente, para mantener el voltaje del generador dentro de los niveles requeridos para su funcionamiento óptimo.

Además de esta función de regulación, los reguladores también desempeñan un papel crucial en la protección contra cambios repentinos o bruscos en la carga, así como contra valores de voltaje que puedan resultar excesivos. Este doble asegura propósito tanto la estabilidad operativa como la salvaguardia del generador frente a condiciones adversas (Enríquez Harper, 2004).

2.2.3 Ciclo de vida

El conocimiento del ciclo de vida de un generador tiene un impacto positivo en su gestión, ya que la identificación oportuna de los momentos críticos para el mantenimiento permite mitigar riesgos, reducir costos y mantener todo el sistema en óptimo funcionamiento.

Conforme a la clasificación propuesta por (Parraga Acosta, 2019), el ciclo de vida de un generador abarca tres etapas fundamentales: la etapa de prepuesta en servicio, la etapa de operación y la etapa de fin de vida. La Figura 2.7 ilustra detalladamente las actividades asociadas con cada una de estas fases.

En términos generales, el inicio del ciclo de vida de un generador implica la selección del equipo, considerando factores económicos, energéticos, ambientales, instalación y mantenimiento, rendimiento, propiedades operativas y tecnológicas, así como características eléctricas y mecánicas. Durante el montaje del equipo, se deben tener en cuenta aspectos cruciales relacionados con las condiciones de la infraestructura de la planta para preservar su rendimiento y prevenir posibles fallas. La puesta en marcha involucra pruebas de operación inicial para iniciar su funcionamiento.

En cuanto al mantenimiento, no existe una única metodología definitiva para establecer los requisitos de mantenimiento de cada equipo. No obstante, algunos datos iniciales proporcionados por el fabricante, el departamento de mantenimiento, los operadores, el área de ingeniería y los resultados del análisis de criticidad son fundamentales para planificar las tareas de mantenimiento.

Las actividades de mantenimiento incluyen la inspección del estado actual del equipo, la preservación de su estado teórico y la restauración de cualquier degradación. El monitoreo constante es esencial para obtener diagnósticos oportunos. Tras realizar el mantenimiento y analizar el índice de salud, se puede tomar decisiones informadas sobre el equipo, como ajustes, sustitución de piezas o reemplazo completo.

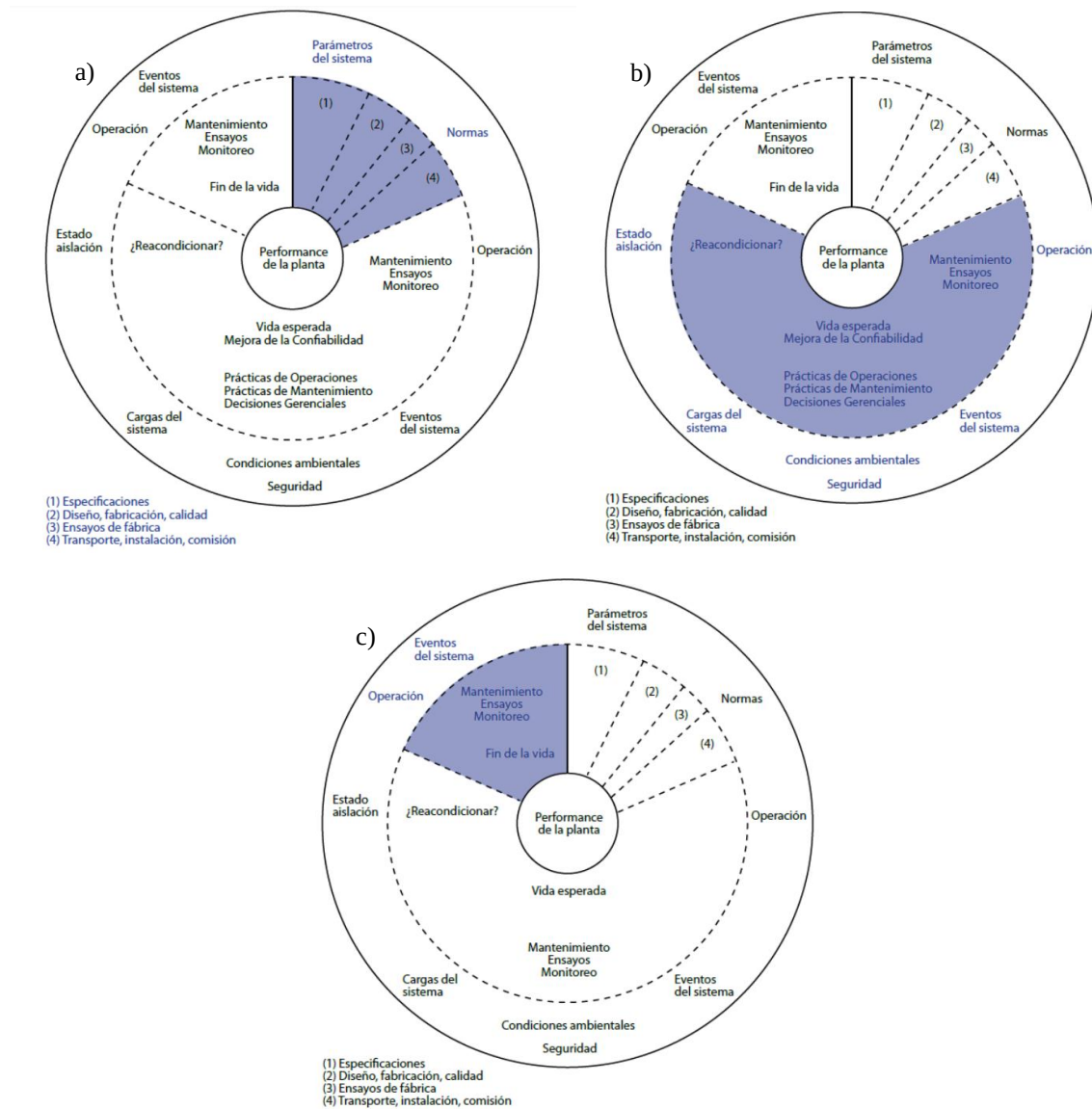


Figura 2.7 Etapas del ciclo de vida: a) Prepuesta en servicio, b) Puesta, y c) Fin de vida

Fuente: (Parraga Acosta, 2019)

2.2.4 Índice de Salud

La evaluación del estado o índice de salud (IS) constituye el procedimiento mediante el cual se identifican marcadores e índices con el fin de determinar y cuantificar el grado de degradación de los componentes de los generadores. Las estrategias de IS aplicadas a generadores de potencia abarcan diversas técnicas, tales como inspección visual, pruebas eléctricas y mecánicas, la implementación de enfoques en línea y/o fuera de línea, el análisis de parámetros de operación y mantenimiento, y la aplicación de estrategias. Basado en condiciones respaldadas por estándares actuales y conocimiento experto.

El índice de salud posibilita a las compañías eléctricas categorizar sus generadores de potencia en tres grupos distintos: 1) aquellos que requieren inversión inmediata de capital para renovación o reemplazo, 2) los que necesitan renovación o mantenimiento en un futuro próximo, pero no de manera inmediata, y 3) aquellos que se encuentran en buen estado y no exigen ninguna intervención. Además, el índice de salud se revela como una herramienta valiosa en la planificación de estrategias de mantenimiento para generadores que presentan signos de deterioro acelerado, los cuales no se corresponden con su tasa natural de envejecimiento (Cerón A. F., Echeverry, Aponte, & Romero, 2015).

2.2.4.1 Metodología general

Conforme a lo señalado por (Strategy, 2020), la metodología siguiente presenta de manera concisa el procedimiento para implementar la estrategia del IS.

1. Identificación de modos de falla: En esta etapa inicial, resulta crucial comprender la degradación del generador, sus componentes y los modos de falla asociados, siendo esencial para la correcta aplicación de los procedimientos de evaluación de la condición. En este contexto, se destaca la importancia de identificar los modos de degradación, así como comprender la naturaleza y las consecuencias de las fallas en el activo. Además, se busca determinar, cuando sea posible, el intervalo de tiempo hasta que el activo alcance el punto de falla.
2. Determinación de tareas de condición: El siguiente paso consiste en identificar las tareas de mantenimiento basadas en condición, las cuales facilitarán un diagnóstico preciso de la

condición del activo, permitiendo mitigar la ocurrencia de los modos de falla previamente identificados.

3. Cálculo del indicador y diagnóstico: En esta fase, los resultados obtenidos de las tareas de mantenimiento basadas en condición se evalúan de acuerdo con los parámetros de calificación establecidos por normas técnicas de institutos reconocidos como IEEE (Instituto de Ingenieros Eléctricos y Electrónicos), IEC (Comisión Electrotécnica Internacional) y CIGRE (Consejo Internacional de Grandes Redes Eléctricas). Los datos cuantitativos obtenidos permitirán el cálculo del índice de salud del activo, consolidando en un único indicador la condición global del equipo y cuantificando la degradación a largo plazo del generador.
4. Evaluación de la salud del activo: El último paso implica estimar la vida útil remanente del activo, identificando aquellos generadores que se encuentran al final de su vida útil y presentan un alto riesgo de experimentar fallas generalizadas. Esto puede requerir inversiones de capital para el reacondicionamiento o reemplazo del activo. La determinación de la criticidad del equipo dentro de la organización y el conocimiento de la salud del activo permiten establecer criterios apropiados para definir acciones como reposiciones inmediatas para el final de la vida útil, inicio de procesos de reposición, evaluación de reacondicionamiento, mejora de condiciones de operación, aumento de frecuencias de mantenimiento, o la implementación de inspecciones y mantenimientos preventivos o predictivos.

2.2.5 Mecanismos de envejecimiento y modos de falla

Los elementos esenciales de los generadores experimentan un deterioro progresivo debido a las tensiones eléctricas, mecánicas, térmicas y ambientales a las que están expuestas en condiciones normales de funcionamiento. Este proceso de envejecimiento se ve acelerado por un mantenimiento deficiente y por operar en condiciones severas o anormales (Klempner & Kerszenbaum, 2004).

2.2.5.1 Mecanismos de deterioro del estator

El rompimiento dieléctrico en el estator surge a partir de la formación de arborescencias, mayormente desencadenadas por la actividad de descargas parciales (Bruetsch, Tari, Froehlich, Weiers, & Vogelsang, 2008). En los devanados del estator de los generadores, las cavidades internas en el aislamiento pueden originarse por defectos de fabricación o por el deterioro del aislamiento debido a la delaminación. Este fenómeno se produce cuando los generadores experimentan frecuentes paros y arranques, provocando un calentamiento y enfriamiento más rápido de los conductores en comparación con las capas aislantes. Esta disparidad en las tasas de contracción y enfriamiento puede generar tensiones mecánicas, resultando en la separación del aislamiento de la superficie del conductor (García, y otros, 1998).

En los cabezales de los devanados del estator de generadores termoeléctricos, se han identificado cuatro mecanismos de deterioro. El primero de ellos está relacionado con la deficiente graduación del campo eléctrico en esta zona, debido a la gran cantidad de bobinas y la considerable longitud de los cabezales y puentes de conexión, lo que resulta en la disminución de las distancias dieléctricas. Estas condiciones propician la generación de descargas parciales externas en esta área.

El segundo mecanismo de deterioro se localiza en las uniones entre bobinas, que presentan soldaduras de baja calidad con pocos puntos de contacto. Dichas soldaduras defectuosas generan puntos calientes, pueden ocasionar la apertura de los conductores y, en última instancia, provocar la quema del aislamiento en esta región (García, y otros, 1998).

El tercer mecanismo de deterioro implica la apertura de las resistencias equipotenciales que conectan los conductores de las bobinas con los conductos de enfriamiento. Si estas resistencias se desconectan, se generan diferencias de potencial entre los conductores y los ductos, dando lugar a descargas parciales externas.

Finalmente, el cuarto mecanismo de deterioro se vincula con la fatiga de los subconductores. Este fenómeno se manifiesta cuando algunos cabezales se aflojan y comienzan a vibrar a frecuencias cercanas a 60 o 120 Hz. A estas frecuencias, el cobre de los subconductores experimenta fatiga, fracturándose y abriéndose, lo que permite que la corriente nominal circule a través de ellos (Klempner & Kerszenbaum, 2004). Esto desencadena arcos eléctricos con altas temperaturas, quemando el aislamiento adyacente y formando trayectorias de carbonización que se extienden hacia las zonas más cercanas con potencial de tierra, resultando en una falla dieléctrica.

2.2.5.2 Mecanismos de deterioro de los componentes del rotor

El principal mecanismo de deterioro en los rotores de polos lisos se presenta en forma de cortocircuitos entre aspiraciones (Tayner, Ran, Penman, & Sedding, 2008). Con el tiempo, la cantidad de cortocircuito aumenta, generando un desequilibrio térmico y electromagnético significativo. Este desequilibrio conlleva niveles elevados de vibración o tensiones inducidas notables en la flecha del rotor. Además, los cortocircuitos entre aspiraciones pueden dañar el aislamiento ubicado debajo de los anillos de retención, provocando caídas a tierra. En generadores de tipo abierto, a pesar de que el aire de refrigeración atraviesa filtros, partículas de contaminación y humedad logran infiltrarse en el interior del generador, mezclándose con partículas de carbón generadas por la fricción de las escobillas. Esta mezcla contaminante se deposita en los devanados del rotor, ocasionando cortocircuitos entre aspiraciones, especialmente en la zona de cabezales.

En relación con las fallas mecánicas, la desalineación entre los ejes del generador y la turbina puede originarse por un acoplamiento deficiente, una alineación incorrecta de los cojinetes o una inclinación del eje. Esta desalineación resulta en el fallo prematuro de cojinetes, ejes y acoplamientos, además de generar temperaturas elevadas y pérdida de aceite en los cojinetes. También provoca incrementos en los niveles de vibraciones radiales y axiales, deterioro anticipado del acoplamiento y ruptura o aflojamiento de tornillos, pernos y pasadores guía.

El desbalance del rotor del generador puede atribuirse a errores en las tolerancias de diseño y fabricación, imprecisiones durante el mecanizado y ensamblaje, flechas flexionadas, cambio de componentes del rotor durante operaciones de mantenimiento, desgaste irregular durante la operación del generador, acumulación de material durante su funcionamiento o distorsión del rotor debido a gradientes de temperatura.

2.2.5.3 Mecanismos de deterioro del núcleo del estator

El desajuste del núcleo puede causar de forma intermitente la separación entre las barras guía y las hendiduras de cola de milano de las laminaciones, lo que puede resultar en la formación de arcos y quemaduras en esa área, conocida como electroerosión. Además, la falta de fijación del núcleo puede llevar al desprendimiento de laminaciones, las cuales pueden incrustarse en los devanados del estator, provocando así una falla dieléctrica (Galaz, 1996).

Los esfuerzos mecánicos y térmicos pueden aumentar en los extremos del núcleo debido al campo magnético axial durante el funcionamiento de los generadores, especialmente en condiciones de sobreexcitación o subexcitación. Bajo estas circunstancias, el deterioro del aislamiento entre las laminaciones puede dar lugar a cortocircuitos, generando altas temperaturas que ocasionan la fusión de las laminaciones o la degradación del aislamiento de las bobinas, resultando en una falla a tierra. Por último, los impactos y rozamientos en los paquetes de laminación, ocurridos durante el ensamblaje del rotor o la colocación de las cuñas, pueden generar cortocircuitos entre las laminaciones y, en consecuencia, causar sobrecalentamientos significativos.

2.2.5.4 Mecanismos de deterioro de las chumaceras y de los componentes mecánicos

Las cifras proporcionadas por los principales fabricantes de rodamientos revelan que el 36% de las fallas se atribuyen a una lubricación deficiente, el 34% resulta de una operación inadecuada y el 16% de las fallas se deben al manejo incorrecto de los rodamientos (Kruger, 2009). Por otro lado, los problemas en los anillos colectores se pueden rastrear hasta cuatro causas principales: desgaste irregular, presencia de impurezas en la superficie, formación de manchas y aparición de huellas de pincel. Estos problemas generan diversas consecuencias, entre las que se incluyen la pérdida simultánea de contacto, la erosión eléctrica y la pérdida de redondez, el desequilibrio del rotor, la creación de una zona de alta fricción en el anillo, la inestabilidad en las escobillas, la contaminación, la alta fricción en los portaescobillas, la respuesta lenta de los resortes y la corrosión debajo de las escobillas (Castillo & Gazcón, 2017).

2.2.5.5 Mecanismos de deterioro del sistema de enfriamiento

El proceso de deterioro en el sistema de enfriamiento indirecto con aire se produce mediante la obstrucción de los conductos de refrigeración debido a la acumulación de sustancias contaminantes. Esta amalgama contaminante se forma a partir del polvo de carbón generado por las escobillas, la humedad y las partículas del entorno, que logran infiltrarse en el interior de la máquina. Esta contaminación puede obstruir los conductos de refrigeración, generando áreas con temperaturas elevadas en el núcleo. Además, los intercambiadores de calor experimentan desgaste

y corrosión química en sus componentes metálicos, lo que puede propiciar la formación de microorganismos y la aparición de corrosión en las tuberías y demás elementos.

En generadores refrigerados con hidrógeno, las vibraciones en los cabezales del estator pueden afectar la soldadura de los subconductores de las bobinas, ocasionando su apertura y provocando arcos eléctricos. En generadores refrigerados con agua e hidrógeno, las fallas más comunes incluyen fugas de agua refrigerante, entrada de humedad, fracturas en las tuberías y obstrucción de los conductos. La vibración normal de una máquina en funcionamiento también puede generar resonancias en las tuberías del sistema de enfriamiento, lo que puede dar lugar a fallas por fatiga en una tubería y fugas de agua de refrigeración. Por último, los conductos de refrigeración con aberturas pequeñas pueden bloquearse debido a la presencia de óxidos de cobre (Tayner, Ran, Penman, & Sedding, 2008).

2.2.5.6 Mecanismos de deterioro del sistema de excitación

La disfunción en un sistema de excitación también se conoce como pérdida de campo o excitación. En tales casos, el generador absorbe la potencia reactiva del sistema y empieza a operar como un generador de inducción, con una velocidad superior a la velocidad síncrona.

En los sistemas de excitación sin escobillas, una falla común se manifiesta en los diodos del rectificador rotatorio (Pranupa, Kiran Kurmar, & Nagaraja, 2018). Estos diodos pueden experimentar fallos de circuito abierto o cortocircuito. Cuando se produce un circuito abierto en un diodo, el alternador principal no puede suministrar el voltaje nominal sin sobrecargar el excitador. Por otro lado, un diodo en cortocircuito representa una de las condiciones de falla más críticas, ya que exige un aumento significativo en la corriente de campo del excitador para mantener el voltaje del alternador. Si un rectificador experimenta un cortocircuito, fluye una corriente excepcionalmente alta en el devanado del inducido del excitador, generando un calentamiento excesivo y potencialmente causando una falla en el excitador. En caso de que se produzca un circuito abierto en un rectificador, el regulador de voltaje aumentará de manera considerable la excitación para mantener el nivel operativo, lo que podría desencadenar una falla en el regulador.

2.2.6 Técnicas de diagnóstico

2.2.6.1 Monitoreo de descargas parciales

Una descarga parcial se define como una descarga eléctrica o chispa que conecta parcialmente el aislamiento de un conductor. Específicamente, la chispa se genera entre el conductor y el aislamiento, sin completar la descarga hacia otro elemento conductor, lo que justifica el término parcial. Esto implica que no hay una ruptura completa del aislamiento, ya que el resto del sistema aislante continúa desempeñando su función.

La ocurrencia de descargas parciales se da en pequeñas cavidades presentes en el sistema aislante que rodea al conductor. Estos huecos pueden originarse por fabricación inadecuada, un proceso de impregnación mal ejecutado o como resultado de los efectos provocados por esfuerzos eléctricos, mecánicos, térmicos o químicos. Sin importar la causa de la formación de agujeros en el sistema aislante, en el interior de estos orificios existe un medio gaseoso con una constante dieléctrica inferior en comparación con el medio aislante circundante. Esta diferencia provoca un aumento del campo eléctrico en el interior del agujero, generando la ionización transitoria del gas hasta quebrar su rigidez dieléctrica y producir la descarga parcial (Galván Barrera, 2010).

Según (Galván Barrera, 2010), las descargas parciales se pueden categorizar de la siguiente manera:

- Internas: Estas descargas parciales se generan en cavidades gaseosas dentro de un sistema aislante o en la interfaz entre los electrodos y el sistema aislante. También pueden ocurrir en cavidades ubicadas en la transición entre dos aislantes con propiedades dieléctricas distintas.
- Superficiales: Se producen en la superficie de aislantes sólidos o en el aire, siempre que la componente tangencial del campo eléctrico, paralelo a la superficie del material aislante, alcance niveles suficientemente elevados.
- Corona: Este tipo de descargas ocurre en áreas limitadas por un medio gaseoso, donde se intensifica el campo eléctrico debido a la presencia de elementos metálicos afilados o partículas metálicas pequeñas. Comúnmente se manifiestan en zonas cercanas a los cabezales de las bobinas o en la salida de las barras, por lo que se procura que esta zona tenga una forma lo más ovalada posible.

- **Arborescentes:** Este tipo de descargas parciales es el más peligroso y puede considerarse como descargas parciales de tipo interno. Inicialmente, se generan en un solo punto y luego se ramifican. Las ramificaciones presentan una estructura hueca donde el gas presente se ioniza, generando descargas parciales de considerable amplitud. La arborescencia continúa creciendo rápidamente hasta perforar el aislamiento en un tiempo extremadamente breve.

Las mediciones de descargas parciales pueden llevarse a cabo tanto fuera de línea como en línea. En las mediciones fuera de línea, la máquina se desconecta de la red, la conexión de neutro se abre y el bobinado se energiza mediante un transformador. Cada fase puede evaluarse individualmente, y es posible determinar las tensiones de ignición/extinción para diferentes actividades de descarga. No obstante, los ensayos fuera de línea exponen las secciones del bobinado cercanas al neutro a la máxima tensión y no reflejan los efectos térmicos y mecánicos de la corriente de carga. Por otro lado, las mediciones en línea requieren condensadores de conexión previamente instalados y permiten estudiar aspectos como la dependencia de corriente de la actividad de descarga o la influencia de las vibraciones. Ambos enfoques tienen sus ventajas y desventajas, y ninguna metodología puede reemplazar completamente a la otra (Gross & Soller, s.f.).

2.2.6.2 Bobina exploradora

El Sistema de Monitoreo de Flujo de Rotor (RFM) emplea una sonda de inducción, también conocida como sonda de flujo o bobina exploradora, que consiste en una pequeña bobina de alambre fino montada en una varilla y ubicada en el entrehierro del generador. La varilla, flexible y ajustable, facilita la extracción o inserción del rotor. La señal proveniente de la sonda de flujo se dirige a través de un analizador de forma de onda y un programa de software de análisis computarizado.

Los resultados del análisis comprenden la integración de la señal para la construcción de la forma de onda de la densidad de flujo en el entrehierro, la medición de la amplitud de los picos de tensión, el cálculo del promedio de la amplitud, las comparaciones entre polos, y la presentación tabular y gráfica del análisis. Este método posibilita la detección de cortocircuitos entre aspiraciones, suministrando datos específicos acerca del número y la ubicación del cortocircuito, con la capacidad de identificar el polo correspondiente.

2.2.6.3 Monitoreo de la corriente en el neutro

Los generadores eléctricos trifásicos son diseñados y fabricados con la finalidad de lograr un equilibrio electromagnético óptimo, buscando un flujo magnético sinusoidal y minimizar el contenido armónico en la salida del generador. A pesar de este esfuerzo por alcanzar un equilibrio perfecto, la perfección electromagnética nunca se logra por completo. En consecuencia, durante las condiciones normales de operación, persiste un cierto nivel de tensión y corriente en el neutro, a pesar de que se espera que ambos valores sean ideales y alcancen 0.

2.2.6.4 Excentricidad

En los generadores, la excentricidad del rotor conlleva a la creación de un entrehierro no uniforme, generando una fuerza desequilibrada de atracción magnética entre el rotor y el estator. Las considerables fuerzas radiales que actúan sobre el rotor también impactan los cojinetes, pudiendo provocar daños o fallos si estas fuerzas no se mantienen a niveles bajos.

La monitorización en línea del entrehierro se realiza periódicamente midiendo la corriente capacitiva modulada entre un sensor ubicado en el estator y los polos del rotor. El sensor funciona como una de las placas de un capacitor, mientras que las otras placas son representadas por los polos del rotor que pasan frente al sensor. El espacio de aire entre el rotor y el estator sirve como dieléctrico del condensador. En estas condiciones, la señal de corriente modulada es inversamente proporcional a las dimensiones del entrehierro. A través del procesamiento electrónico, es posible convertir esta señal en una respuesta de salida lineal que representa la dimensión precisa del entrehierro.

2.2.6.5 Monitoreo de vibraciones

La vibración se refiere al movimiento del periódico u oscilatorio de un objeto. En el caso de una máquina rotatoria, este fenómeno engloba el movimiento periódico de los rotores, la carcasa, la tubería y el sistema de cimentación. Por lo general, estas vibraciones son de amplitud reducida, requiriendo equipos sensibles para su detección. La medición en línea de estas vibraciones se lleva a cabo mediante sensores piezoeléctricos, los cuales emiten una señal proporcional al

desplazamiento de la sección instrumentada. Estos sensores piezoeléctricos adoptan una forma cilíndrica con una sección roscada para su fijación (ISO 10816-5, 2000).

En el contexto de una máquina rotatoria, existen dos enfoques para medir las vibraciones. Uno de ellos implica el uso de sensores de velocidad o aceleración ubicados en la superficie del pedestal de la chumacera, una parte estacionaria que generalmente proporciona un indicador confiable del nivel o la severidad de la vibración para cada tipo de máquina. La segunda opción implica la medición relativa de la vibración de la flecha con respecto a una parte fija, sustancialmente el cuerpo de la chumacera (ISO 10816-1, 1995).

2.2.6.6 Monitoreo de la temperatura de los cojinetes

La temperatura del metal en los rodamientos del rotor se encuentra directamente vinculada a la temperatura de entrada del aceite lubricante en los cojinetes. Diversos factores, como la vibración, la alineación, la condición del aceite, las precargas de los rodamientos y el flujo de aceite, también influyen en las temperaturas de los cojinetes. La supervisión de las temperaturas del metal en las mitades superior e inferior del rodamiento, o en las pastillas individuales en un cojinete de pastillas basculantes, se realiza mediante termopares insertados en el material del rodamiento. Normalmente, se instalan uno o dos sensores en la mitad inferior del rodamiento.

2.2.6.7 Monitoreo de las condiciones del sistema de enfriamiento

En generadores refrigerados por aire, se realiza un monitoreo integral que abarca la medición de la cantidad de agua necesaria para el enfriamiento, así como la temperatura del agua tanto en la entrada como en la salida del enfriador. Además, se registra la presión del agua en ambas etapas, y se supervisa la temperatura en los devanados del estator.

En generadores refrigerados con hidrógeno, se lleva a cabo un seguimiento detallado de la temperatura, pureza, presión y humedad presentes en el sistema de hidrógeno.

Para los generadores refrigerados con agua e hidrógeno, el monitoreo se extiende a la temperatura y flujo tanto en la entrada como en la salida, tanto para el agua como para el hidrógeno. Además, se evalúa la conductividad del agua, mientras que en el hidrógeno se supervisa la pureza,

presión y humedad. Este enfoque abarcador asegura un control eficiente de las condiciones de refrigeración en cada tipo de generador.

2.2.6.8 Monitores de las condiciones del sistema de excitación

Los sistemas de excitación estática de los generadores incorporan un registrador de eventos y una función de autodiagnóstico que posibilitan la detección e identificación de posibles fallos internos en el sistema de control (CFE W4101-16, 2014). En lo que respecta a los demás componentes, tanto eléctricos como mecánicos, se realiza un análisis exhaustivo de las fallas y disparos recurrentes, evaluando además el tiempo requerido para su reparación. En el caso específico de los sistemas de excitación sin escobillas, se dispone de un equipo portátil con aplicaciones de software diseñadas para la configuración, diagnóstico, mantenimiento y pruebas del sistema (CFE VF000-72, 2012).

2.3 Inteligencia Artificial

La denominación científica de "homo sapiens" que la humanidad ha atribuido a sí misma refleja la apreciación de las capacidades mentales tanto en el contexto de la vida diaria como en la formación de la propia identidad. Por otro lado, la Inteligencia Artificial se dedica a comprender entidades inteligentes, buscando no solo adquirir conocimientos sobre los seres humanos, como lo hacen la Filosofía y la Psicología, sino también construir y comprender estas entidades inteligentes (Ponce, 2010).

La Inteligencia Artificial, según (Haugeland, 1985), se define como "la fascinante tarea de lograr que las computadoras piensen... máquinas con mente en su sentido más amplio y literal". Otra definición, propuesta por (Luger y Stubblefield, 1993), describe la IA como "la rama de la ciencia de la computación dedicada a la automatización de la conducta inteligente". En términos más amplios, la IA representa la disciplina científica encargada del diseño de sistemas de computación inteligente. Estos sistemas exhiben características que emula la inteligencia observada en el comportamiento humano, abarcando áreas como la comprensión del lenguaje, el aprendizaje, el razonamiento y la resolución de problemas.

La IA es un ámbito diverso y dinámico que se despliega a través de diversas ramas, áreas y técnicas, cada una abordando enfoques específicos y aplicaciones concretas. Destacan entre ellas Lógica Difusa, Machine Learning y Redes Neuronales Artificiales, que constituyen pilares fundamentales de la IA, contribuyendo significativamente a su desarrollo y aplicación en diversos contextos. Es importante señalar que RNA, aunque es una disciplina autónoma, forma parte integral del amplio espectro de ML. Estas disciplinas proporcionan herramientas y metodologías únicas para enfrentar desafíos particulares.

2.3.1 Lógica Difusa

La Lógica Difusa constituye un conjunto de principios matemáticos fundamentados en grados de membresía o pertenencias, cuya finalidad radica en modelar información. Este proceso de modelado se lleva a cabo mediante reglas lingüísticas que aproximan una función mediante la relación entre las entradas y salidas del sistema, conocido como composición. Una característica destacada del sistema de control basado en lógica difusa es su capacidad para prescindir de un conocimiento previo del modelo matemático del sistema real. En este enfoque, el sistema puede conceptualizarse como una "caja negra" a la cual se le suministran entradas, ya través de su funcionamiento, la planta genera la salida deseada. Esta particularidad simplifica la implementación del control difuso, ya que se adapta eficazmente a sistemas complejos sin la necesidad de una comprensión detallada de su estructura interna (Ponce, 2010).

2.3.1.1 Estructura del modelo difuso

Según lo señalado por (Ponce, 2010), la estructura básica de un controlador difuso se compone de los siguientes elementos:

- Reglas: Estas directrices determinan la acción de control que debe llevarse a cabo y son extraídos del conocimiento de un experto. Dichas reglas se configuran en forma de relaciones, siendo la lógica difusa una metodología que se sustenta en las relaciones. Estas relaciones se establecen mediante la aplicación de reglas "SI-ENTONCES", permitiendo así la modelación de aspectos cualitativos del conocimiento humano y de los procesos de razonamiento sin la necesidad de una formalización precisa.

- Difusificador: Actúa como el vínculo entre las entradas del sistema en su formato real y su representación en términos difusos. Cada entrada debe ser transformada a una forma que pueda ser empleada por las reglas del controlador difuso.
- Desdifusificador: Encargado de convertir un valor difuso, resultante de la aplicación de las reglas, en una salida concreta y real. Este componente juega un papel esencial al proporcionar una salida que puede ser interpretada y utilizada en el contexto del sistema de control difuso.

Estos elementos, en conjunto, constituyen la infraestructura fundamental que permite el funcionamiento y la toma de decisiones del controlador difuso.

2.3.1.2 Metodología

La Lógica Difusa implica un proceso que consta de tres etapas cruciales para lograr el resultado deseado. Inicialmente, en la primera etapa, se lleva a cabo un proceso en el cual las variables exhiben un grado de incertidumbre metalingüística. En este sentido, el rango de valores (universo de discurso) de cada variable se categoriza mediante conjuntos difusos. Cuando los sensores registran estas variables, los valores son sometidos a un proceso de fusión, durante el cual se ajustan a un rango de pertenencia entre cero (0) y uno (1). Este proceso busca determinar en qué medida el valor adquirido pertenece a un conjunto difuso, y los conjuntos difusos están caracterizados por funciones de membresía ajustadas al punto de operación óptima del sistema.

La segunda etapa implica la formulación de reglas lingüísticas (inferencia) que actúan como directrices para que el sistema adopte un comportamiento más adecuado, idóneo o deseado, de acuerdo con el modelo de referencia o los objetivos del usuario. Se evalúa el grado de pertenencia de cada variable en un conjunto de reglas de inferencia, las cuales fueron establecidas en colaboración con un experto. Este conjunto de reglas de inferencia determina una consecuencia, asignando un grado de pertenencia a un conjunto difuso que caracteriza las salidas del sistema.

Finalmente, la tercera etapa, conocida como desfusificación, consiste en transformar el grado de pertenencia, derivado de la consecuencia de la regla de inferencia, en un valor nítido o real. Para lograr esto, se han ajustado previamente las funciones de membresía de cada salida con el propósito de obtener un valor cuantificable. En última instancia, el controlador difuso proporciona valores nítidos o reales como resultado de las reglas lingüísticas previamente estructuradas, permitiendo

que el sistema interprete las instrucciones y ejecute las acciones correspondientes (Ponce, 2010). La Figura 2.8 ilustra el diagrama esquemático del controlador difuso.

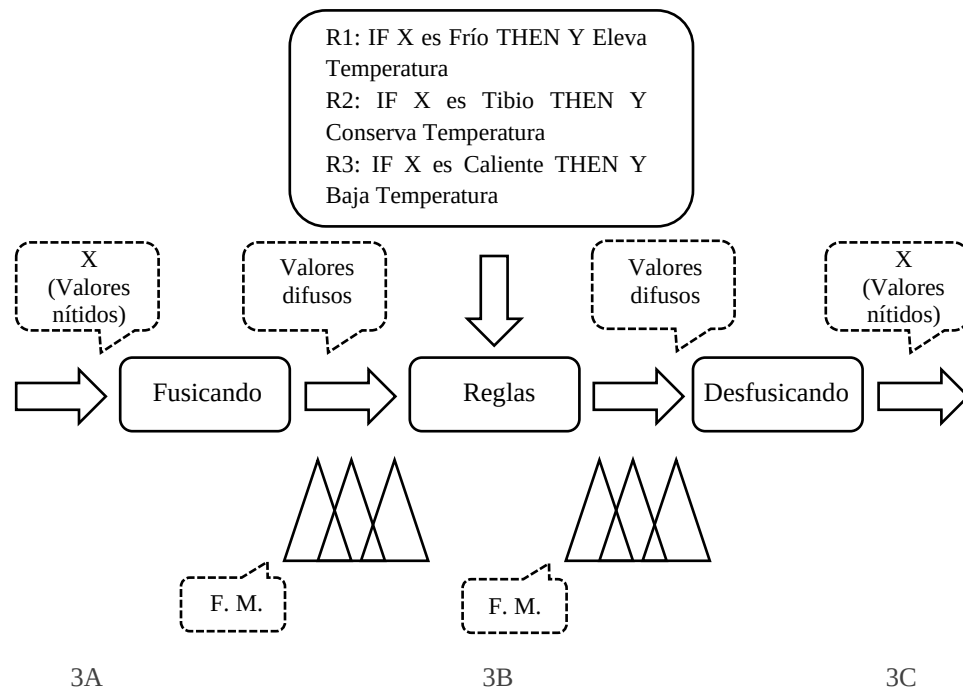


Figura 2.8 El diagrama de bloques para desarrollar la metodología de lógica difusa muestra las tres etapas que constituyen el control. Tomando como entradas X (3A), se lleva a cabo el proceso conforme al método (3B), obteniendo una respuesta de salidas Y (3C)

Fuente: (Ponce, 2010)

2.3.2 Machine Learning

Machine Learning, conocido como Aprendizaje Automático en español, se basa fundamentalmente en un conjunto de algoritmos matemáticos desarrollados en la década de 1980. Este campo representa una parte crucial de la Inteligencia Artificial y constituye la ciencia de capacitar a las computadoras o máquinas para aprender de manera autónoma a partir de datos, realizando análisis automático sin intervención humana. La disciplina abarca diversos algoritmos matemáticos diseñados para tomar decisiones o, de manera más precisa, prever resultados en función de conjuntos específicos de datos.

La denominación Machine Learning fue acuñada por Arthur Samuel, un pionero estadounidense en ciencias de la computación e inteligencia artificial, en 1959 mientras estaba en International Business Machines Corporation (IBM). Samuel definió Machine Learning como el campo de estudio que confiere a las computadoras la capacidad de aprender sin una programación explícita. En una perspectiva más moderna, en 1997, Tom Mitchell proporcionó una definición actualizada de Machine Learning, expresándolo como "Un programa de computadora se considera que aprende de la experiencia E con respecto a una clase de tareas T y una medida de rendimiento P , si su desempeño en las tareas de T , medido por P , mejora con la experiencia E " (Xiao, 2022).

2.3.2.1 Metodología

De acuerdo a (Géron, 2019), el desarrollo de un proyecto de Machine Learning sigue una metodología estructurada, desde la definición del problema hasta el despliegue del modelo.

En la fase inicial, se comprenden los objetivos y se formula el problema específico en términos de ML. La recopilación de datos, paso crítico, es seguida de un análisis exploratorio y limpieza para manejar valores atípicos. Posteriormente, el preprocesamiento implica seleccionar características relevantes y transformar los datos según sea necesario.

La división del conjunto de datos en entrenamiento y prueba prepara para la siguiente etapa, donde se selecciona el algoritmo de ML y se ajustan sus hiperparámetros. El proceso de aprendizaje y validación cruzada se realiza con el conjunto de entrenamiento. La evaluación del modelo se ejecuta en el conjunto de pruebas, utilizando métricas adecuadas. Se ajusta el modelo si es necesario y se procede al despliegue en producción.

La documentación y comunicación son esenciales para registrar el proceso y las decisiones clave. Los resultados se presentan formalmente, compartiendo el conocimiento adquirido. Este ciclo de vida puede ser iterativo, permitiendo ajustes y mejoras a medida que se obtenga más información o se identifiquen áreas de oportunidad.

En la Figura 2.9 se ilustra la metodología general de un proyecto de ML.

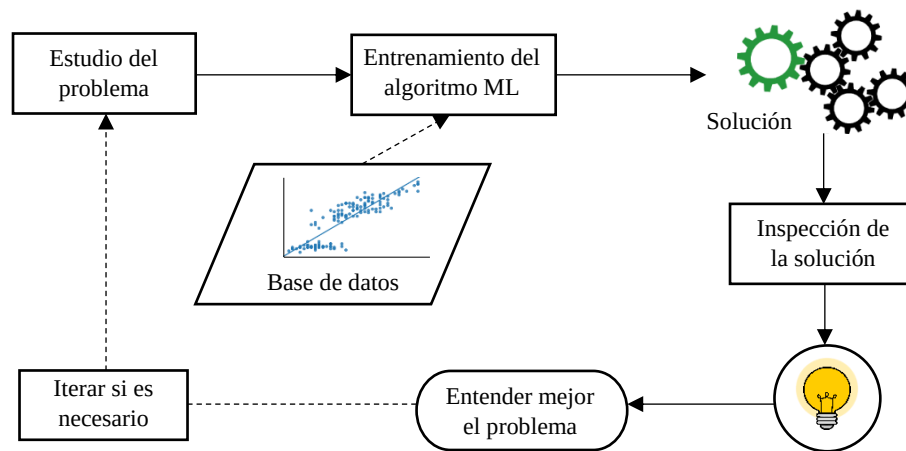


Figura 2.9 Metodología ML

Fuente: (Géron, 2019)

2.3.2.2 Clasificación

Según (Xiao, 2022), Machine Learning se puede clasificar en cuatro categorías principales:

- **Aprendizaje supervisado:** Durante el aprendizaje supervisado, los modelos se entrenan con datos que tienen etiquetas. En este proceso, los algoritmos ajustan continuamente los parámetros de los modelos para minimizar el error entre la salida calculada y la salida deseada para una entrada específica. Se utiliza comúnmente para tareas de clasificación y regresión.
- **Aprendizaje no supervisado:** En el aprendizaje no supervisado, los modelos se alimentan con datos sin etiquetas. Los algoritmos analizan los datos y los agrupan según sus características. Esta categoría se aplica combinada a tareas de agrupamiento y asociación. El agrupamiento implica dividir los datos en grupos, y la asociación se refiere a descubrir reglas que describen la mayoría de los datos.
- **Aprendizaje semi supervisado:** En el aprendizaje semi supervisado, se utilizan conjuntos de datos que contienen tanto ejemplos etiquetados como no etiquetados. Esta metodología es valiosa cuando no es posible etiquetar todos los datos disponibles. La práctica común es agrupar los datos mediante un algoritmo de aprendizaje no supervisado y luego emplear los datos etiquetados existentes para asignar etiquetas al resto de los datos no etiquetados.

- Aprendizaje por refuerzo: En el aprendizaje por refuerzo, los algoritmos aprenden a través de la interacción con un entorno, determinando qué acciones generan la máxima recompensa acumulativa.

Machine Learning aborda de manera eficiente desafíos que exigen ajustes manuales extensos o reglas detalladas. Asimismo, resuelve con eficacia problemas complejos que superan las soluciones efectivas de enfoques tradicionales, exhibiendo una capacidad única para adaptarse dinámicamente a entornos cambiantes. Además, proporciona percepciones valiosas en la comprensión de problemas complejos y extensos conjuntos de datos (Géron, 2019).

2.3.2.3 Redes Neuronales Artificiales

Las Redes Neuronales Artificiales se definen como sistemas de mapeo no lineales, cuya estructura se fundamenta en los principios observados en los sistemas nerviosos de humanos y animales. Estas redes consisten en un conjunto extenso de procesadores simples interconectados mediante conexiones ponderadas. Dichas unidades de procesamiento, denominadas neuronas, reciben entradas de otros nodos y generan salidas escalares simples que dependen de la información local almacenada internamente o recibida a través de las conexiones con pesos. Las RNA tienen la capacidad de realizar diversas funciones complejas, en función de la configuración de sus conexiones (Ponce, 2010).

2.3.2.3.1 Fundamento biológico

Las redes neuronales artificiales son representaciones computacionales de las redes neuronales biológicas presentes en el cerebro humano. El concepto de redes neuronales fue formulado por el neurofisiólogo estadounidense Warren McCulloch y el lógico estadounidense Walter Pitts en 1943. Una red neuronal biológica se compone de una red de neuronas interconectadas. La neurona biológica típicamente consta de un cuerpo celular, dendritas y un axón. El cuerpo celular, también conocido como soma, da lugar a las dendritas y el axón, filamentos que se extienden desde él. Mientras las dendritas suelen extenderse unos pocos cientos de micrómetros desde el soma, el axón puede llegar a tener hasta un metro de longitud. En el extremo del axón se encuentran las terminales axónicas, y cada terminal está conectada a otra neurona a través de la sinapsis. En el caso de cada neurona, las dendritas sirven como entradas desde otras neuronas, y el

axón funciona como la salida hacia otras neuronas (Xiao, 2022). En la Figura 2.10 se muestran los elementos de una red neuronal natural.

Una característica distintiva de las neuronas es su capacidad para la comunicación. Concretamente, las dendritas y el cuerpo celular reciben señales de entrada, las integran y combinan, para luego emitir señales de salida. El axón se encarga de transmitir estas señales a los terminales axónicos, desde donde se distribuye la información (Ponce, 2010).

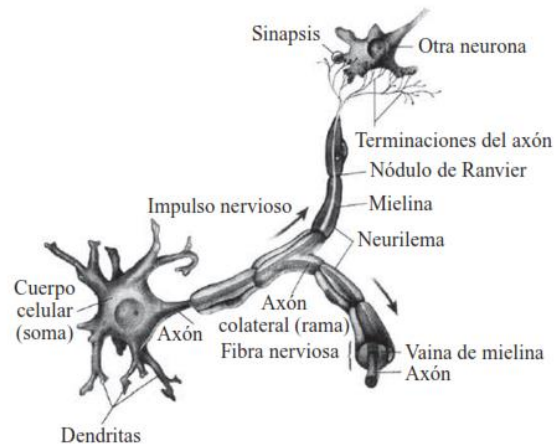


Figura 2.10 Red neuronal natural

Fuente: (Ponce, 2010)

2.3.2.3.2 Topologías de las redes neuronales artificiales

A semejanza de las neuronas biológicas, las neuronas artificiales poseen tanto una entrada como una salida. Estas neuronas artificiales reciben señales de entrada, las ponderan mediante pesos, las suman y, posteriormente, las someten a una función de activación no lineal para obtener la salida. La función de activación de un nodo determina la salida de dicho nodo en función de una entrada o de un conjunto de entradas. Al igual que las redes de neuronas biológicas, las redes neuronales artificiales están conformadas por neuronas artificiales interconectadas, generalmente distribuidas en tres capas: una capa de entrada, una capa de salida y una capa oculta. La razón por la que se limita a una sola capa oculta es que la complejidad computacional se incrementa a medida que se añaden más capas (Xiao, 2022).

De acuerdo con (Ponce, 2010), existen dos topologías ampliamente utilizadas, las cuales se diferencian en la forma en que establecen las conexiones:

- Redes de propagación hacia adelante: En este tipo de redes, el flujo de la información va exclusivamente desde las entradas hasta las salidas, extendiéndose a través de varias capas de unidades. No se establecen conexiones de retroalimentación. La Figura 2.11 ilustra una red de propagación hacia adelante con 4 capas, con x entradas y y salida.
- Redes recurrentes: Estas redes incluyen conexiones de retroalimentación, lo que puede conducir a un proceso evolutivo hacia un estado estable, donde no hay cambios en el estado de activación de las neuronas.

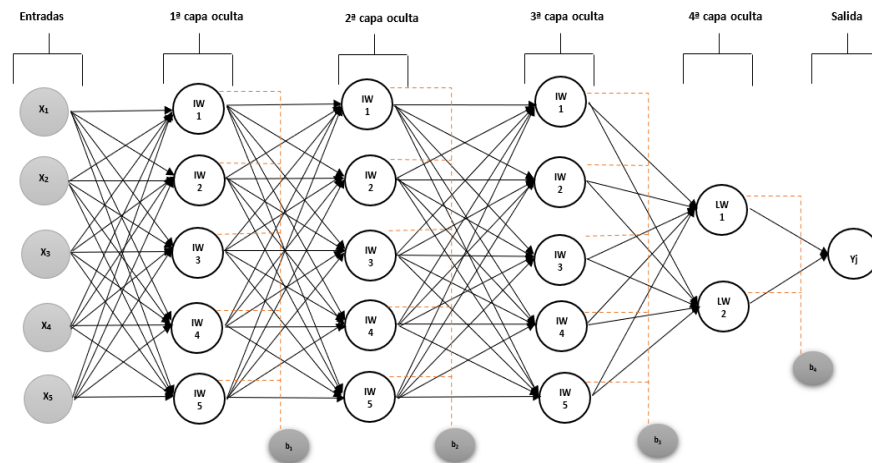


Figura 2.11 RNA de propagación hacia adelante

Fuente: Elaboración propia

2.3.2.3.3 Metodología

Para entrenar una red neuronal artificial, se inician los pesos de manera aleatoria y se suministra a la red un conjunto de datos de entrenamiento con entradas y salidas específicas. Cada vez que la red genera una salida a partir de las entradas, se utiliza una función de pérdida para compararla con la salida deseada, y se retroalimentan las diferencias, denominadas errores, a la red para ajustar los pesos en consecuencia. Este proceso, conocido como retropropagación, implica el ajuste iterativo de los pesos mediante el descenso de gradiente, un método que calcula el gradiente de los errores con respecto a los pesos y los ajusta para minimizar los errores. Este ciclo se repite hasta que los pesos de la red neuronal alcanzan estabilidad. El entrenamiento de una red neuronal artificial se reduce esencialmente a un problema de optimización.

El descenso de gradiente es un algoritmo de optimización iterativo ampliamente empleado para localizar mínimos locales en una función. La estrategia consiste en dar pasos sucesivos en la dirección opuesta al gradiente, que indica la pendiente más pronunciada. Tras el entrenamiento, la red puede recibir datos no vistos como entrada, y proporcionará una salida predicha. Aunque las redes neuronales artificiales demandan extenso entrenamiento y un considerable tiempo para ello, una vez entrenadas, son capaces de generar resultados con rapidez (Xiao, 2022).

2.3.2.3.4 Ventajas y desventajas

En relación con las ventajas y desventajas del empleo de Redes Neuronales Artificiales en comparación con otras técnicas, (Ponce, 2010) las exponen de la siguiente manera:

Ventajas de las RNA frente a otros sistemas de procesamiento de información:

- Las RNA tienen la capacidad de sintetizar algoritmos mediante un proceso de aprendizaje.
- No es necesario poseer conocimientos detallados de los aspectos matemáticos para utilizar la tecnología neuronal; basta con estar familiarizado con los datos específicos del trabajo.
- La resolución de problemas no lineales constituye una fortaleza destacada de las RNA.
- Las RNA exhiben robustez; incluso en el caso de fallos en algunos elementos de procesamiento, la red continúa su funcionamiento. Esta característica contrasta con la programación tradicional.

Desventajas de las RNA:

- Las RNA requieren ser entrenadas para cada problema, y encontrar la arquitectura adecuada implica realizar múltiples pruebas. El proceso de entrenamiento puede extenderse considerablemente, consumiendo varias horas de la unidad central de procesamiento.
- Dado que las redes se entrenan en lugar de ser programadas, exigen una gran cantidad de datos.
- Las RNA presentan un aspecto complejo para un observador externo que busque realizar modificaciones. Incorporar nuevo conocimiento implica ajustar las interacciones entre numerosas unidades para que su efecto unificado sintetice este conocimiento.

2.3.2.4 Algoritmos de clasificación

Los algoritmos constituyen el núcleo esencial del aprendizaje automático, representando el código que guía a la computadora en la búsqueda y ajuste de ponderaciones de posibles respuestas en función de la retroalimentación recibida.

En la actualidad, existen diversos algoritmos de Machine Learning ampliamente reconocidos, muchos de los cuales están integrados en bibliotecas populares de codificación para la ciencia de datos. La construcción de un modelo básico de ML implica la exploración de múltiples algoritmos precargados para determinar cuál se adapta de manera óptima a los datos. Cada modelo exhibe sus propias características distintivas, fortalezas, debilidades, arquitectura y un enfoque singular para ponderar los resultados (Norman, 2019).

2.3.2.4.1 Support Vector Machines

Una Máquina de Vectores de Soporte, SVM por sus siglas en inglés, destaca como un modelo de Machine Learning altamente eficiente y adaptable, con la capacidad de ejecutar tareas de clasificación, regresión, e incluso identificación de valores atípicos tanto en contextos lineales como no lineales. Este enfoque se encuentra entre las opciones más favorecidas en el ámbito de ML. Su rendimiento se destaca especialmente en la clasificación de conjuntos de datos complejos, especialmente aquellos de dimensiones pequeñas o medianas (Géron, 2019).

La Figura 2.12 ilustra de manera clara un ejemplo elemental de un problema de clasificación binaria mediante el uso de SVM.

Este modelo es entrenado con datos y configura un hiperplano para efectuar la separación entre los dos conjuntos de datos. En este escenario, el hiperplano toma la forma de una línea recta. SVM ajusta la posición del hiperplano con el objetivo de maximizar los márgenes entre ambos conjuntos de datos. Los puntos de datos situados en estos márgenes son denominados vectores de soporte. Este caso representa un problema de clasificación lineal bidimensional de naturaleza sencilla. No obstante, es importante señalar que SVM también demuestra su eficacia en la resolución de problemas de clasificación no lineales tridimensionales, donde el hiperplano podría adoptar la forma de una superficie curva plana o incluso estructuras más complejas (Xiao, 2022).

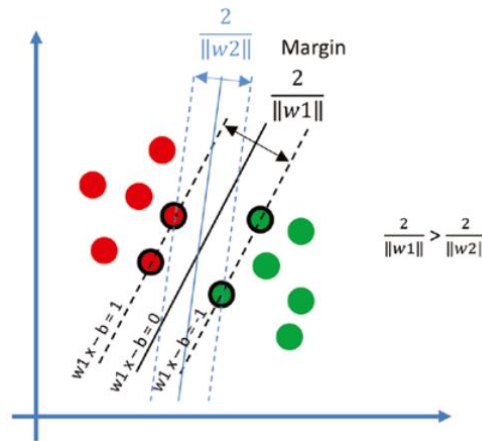


Figura 2.12 SVM binario, el hiperplano (línea continua), el margen (entre dos líneas discontinuas) y los vectores de soporte (puntos encerrados en un círculo en la línea discontinua)

Fuente: (Géron, 2019)

Las SVM destacan en el aprendizaje automático con ventajas clave. Son efectivas en espacios de grandes dimensiones, incluso cuando las dimensiones superan el número de muestras. Además, su capacidad para utilizar un subconjunto de puntos de entrenamiento en la función de decisión, conocidos como vectores de soporte, contribuye a una eficiencia notable en términos de memoria. La versatilidad de las SVM se muestra al especificar funciones del Kernel para la decisión, permitiendo opciones comunes o personalizadas.

Sin embargo, las SVM tienen limitaciones. Evitar el sobreajuste es esencial cuando las características superan significativamente las muestras. Además, cabe destacar que las SVM no generan estimaciones de probabilidad directas; para ello, se recurre a una costosa validación cruzada quíntuple (Pedregosa F. , y otros, 2011).

2.3.2.4.2 Naive Bayes

Naive Bayes constituye otro algoritmo de aprendizaje supervisado ampliamente utilizado que emplea el teorema de Bayes, incorporando la suposición ingenua de las probabilidades de las características en un conjunto de datos específico. La suposición ingenua implica que cada característica es independiente de las demás. De acuerdo con el teorema de Bayes, la probabilidad de que ocurra una clase (C) para un conjunto de datos con características dadas (X) se puede calcular mediante la Ecuación 1.

$$P(C|X) = \frac{P(C) * P(X|C)}{P(X)} \quad \text{Ecuación 1}$$

Donde:

$P(C|X)$ es la posterior, la probabilidad de la clase (C) dadas las características (X).

$P(C)$ es el prior, la probabilidad de clase (C).

$P(X|C)$ es la verosimilitud, la probabilidad de las características (X) para la clase dada (C).

$P(X)$ es la probabilidad previa de las características (X).

El proceso de entrenamiento de un clasificador Naive Bayes implica calcular las probabilidades $P(C)$ de todas las clases para las características y la evidencia dada $P(X)$. Estas probabilidades $P(C)$ y $P(X)$ son constantes para un conjunto de datos específico y se almacenan para su uso futuro. Dada una muestra X , se calcula la probabilidad $P(X|C)$, y posteriormente, se puede calcular la probabilidad posterior $P(C|X)$ (Xiao, 2022).

A pesar de sus suposiciones simplificadas, los clasificadores de Naive Bayes destacan por su eficacia en aplicaciones del mundo real, como la clasificación de documentos y el filtrado de correo no deseado. Además, requieren una cantidad mínima de datos de entrenamiento para estimar los parámetros necesarios. En cuanto a la velocidad, son notablemente rápidos en comparación con métodos más complejos, gracias a la independencia en la estimación de distribuciones de características condicionales.

Aunque su rendimiento como clasificador es reconocido, es importante tener en cuenta que Naive Bayes puede ser un estimador deficiente, por lo que se aconseja no tomar las salidas de probabilidad de predict_proba con excesiva seriedad (Pedregosa F. , y otros, 2011).

2.3.2.4.3 Dimensionality Reducion

En desafíos de Machine Learning, la abundancia de características por instancia de entrenamiento, no solo ralentiza el proceso de entrenamiento, sino que también presenta el desafío conocido como la "maldición de la dimensionalidad".

Afortunadamente, en el mundo real, a menudo es factible reducir significativamente el número de características, convirtiendo problemas inicialmente difíciles en enfoques más manejables.

La reducción de dimensionalidad, a pesar de su capacidad para acelerar el entrenamiento, conlleva la pérdida de información y puede afectar el rendimiento del sistema, complicando los flujos de trabajo y dificultando el mantenimiento. Se recomienda explorar el entrenamiento con datos originales antes de considerar la reducción de dimensionalidad, especialmente si la lentitud del entrenamiento se convierte en un problema. Aunque en algunos casos esta reducción mejora el rendimiento al eliminar ruido y detalles innecesarios, su impacto principal radica en la aceleración del entrenamiento.

Además de su utilidad para facilitar el entrenamiento, la reducción de dimensionalidad resulta invaluable para la visualización de datos. Al disminuir el número de dimensiones, es posible representar de manera condensada un conjunto de entrenamiento de alta dimensionalidad en gráficos, proporcionando percepciones cruciales al identificar visualmente patrones, como agrupaciones (Géron, 2019).

2.3.2.4.3.1 Linear Discriminant Analysis

El Análisis Discriminante Lineal, LDA por sus siglas en inglés, figura como una de las técnicas de reducción de dimensiones más prevalentes en Machine Learning. Su objetivo es proyectar un conjunto de datos en un espacio de dimensiones inferiores para lograr una mejor separación entre clases y reducir los costos computacionales asociados.

En la Figura 2.13 se presenta una ilustración del funcionamiento de LDA. Para un conjunto de datos específico que abarca tres clases y se mide en dos características (X_1 y X_2), inicialmente mezcladas, la separación resulta desafiante. LDA aborda este desafío al buscar nuevos ejes, representados por las líneas discontinuas, y proyectar los datos en estos nuevos ejes (LDA_1 y LDA_2). Ajusta estos ejes considerando las medias y variaciones de cada grupo de datos, mejorando así la separación entre las clases.

Tras la aplicación de LDA, es posible visualizar los datos a lo largo de los nuevos ejes, lo que mejora significativamente la separación. En esencia, LDA realiza una proyección lineal de los datos de una dimensión a otra (Xiao, 2022).

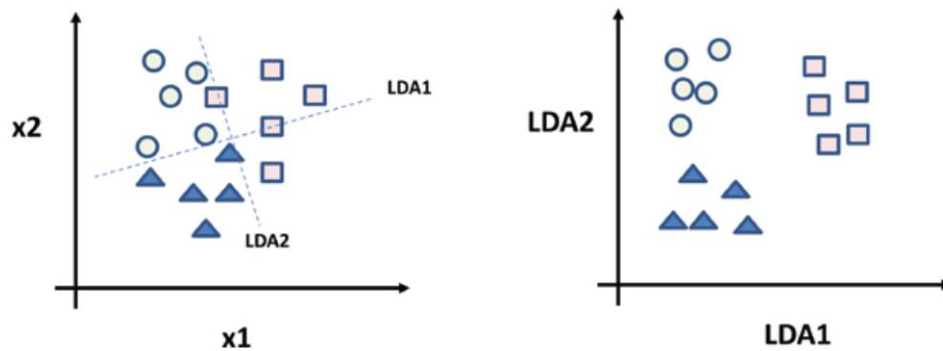


Figura 2.13 Conjunto de datos antes de LDA (izquierda) y los datos después de LDA (derecha)

Fuente: (Xiao, 2022)

2.3.2.4.3.2 Quadratic Discriminant Analysis

El Análisis Discriminante Cuadrático, conocido como QDA por sus siglas en inglés, es una técnica de Machine Learning empleada tanto para la clasificación como para la reducción de dimensionalidad. Esta metodología se aplica en problemas de clasificación bajo la premisa de que las distribuciones de las características para cada clase siguen distribuciones normales multivariadas con covarianzas distintas.

La diferencia fundamental entre el análisis discriminante lineal y el análisis discriminante cuadrático radica en las suposiciones acerca de las covarianzas de las clases. Mientras que LDA asume que todas las clases comparten la misma matriz de covarianza, QDA permite que cada clase tenga su propia matriz de covarianza. Esta característica confiere a QDA una mayor flexibilidad en la modelización de la relación entre las características y las clases, al no imponer la restricción de igualdad en las covarianzas.

Estos clasificadores resultan atractivos debido a que poseen soluciones de fácil cálculo, son inherentemente multiclase, han demostrado su eficacia en la práctica y no requieren ajuste de hiperparámetros.

La Figura 2.14 ilustra un ejemplo donde se representan los elipsoides de covarianza y los límites de decisiones aprendidas por LDA y QDA. Los elipsoides exhiben la doble desviación estándar para cada clase. Con LDA, la desviación estándar es constante para todas las clases, mientras que con QDA cada clase tiene su propia desviación estándar. La fila inferior demuestra

que el LDA solo puede aprender límites lineales, mientras que el QDA puede aprender límites cuadráticos, otorgándole mayor flexibilidad.

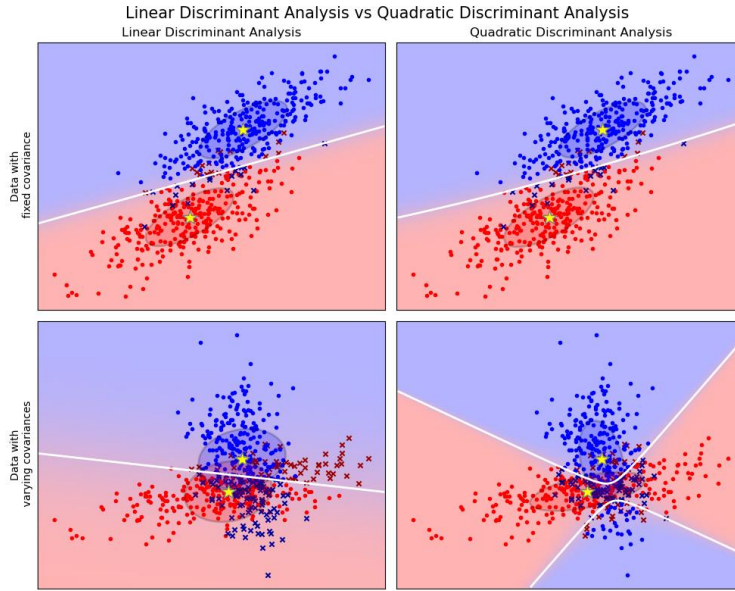


Figura 2.14 Comparación entre LDA (izquierda) y QDA (derecha)

Fuente: (Pedregosa F. , y otros, 2011)

Según (Pedregosa F. , y otros, 2011), tanto LDA como QDA pueden ser derivados a partir de modelos probabilísticos simples que describen la distribución condicional de clase de los datos. Esto logra modelando, $P(X|y = k)$ para cada clase k , la distribución condicional de las características. Posteriormente, las predicciones se obtienen mediante la aplicación de la regla de Bayes a cada muestra de entrenamiento $x \in R^d$ como se muestra en la Ecuación 2. Y se elige la clase k que maximiza la probabilidad posterior.

$$P(y = k|x) = \frac{P(x|y = k) P(y = k)}{P(x)} = \frac{P(x|y = k) P(y = k)}{\sum_l P(x|y = l) * P(y = l)} \quad \text{Ecuación 2}$$

De manera más específica, en el análisis discriminante lineal y cuadrático, $P(x|y)$ se modela la distribución condicional de las características para cada clase como una distribución gaussiana multivariada con densidad como se muestra en la Ecuación 3, donde d es el número de características.

$$P(x|y = k) = \frac{1}{(2\pi)^{d/2} |\Sigma_k|^{1/2}} \exp \left(-\frac{1}{2} (x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k) \right) \quad \text{Ecuación 3}$$

2.3.2.4.4 Decision Tree

Un árbol de decisión representa uno de los enfoques más prevalentes en el aprendizaje supervisado no paramétrico y se aplica tanto a problemas de clasificación como de regresión. A partir de este árbol, árbol invertido, se derivan conjuntos de reglas que fundamentan la toma de decisiones. Cada decisión se fundamenta en estas reglas, donde un nodo del árbol representa la variable de consulta y la rama refleja el valor correspondiente a esa variable.

La toma de decisiones se inicia desde la raíz del árbol, avanzando por las ramas hasta llegar a las hojas. De este modo, cada rama se convierte en una regla "si-entonces". Los nodos hoja, por su parte, abarcan todos los resultados posibles dentro del conjunto de datos. La complejidad de las reglas y del modelo aumenta a medida que el árbol se adentra en profundidad (Xiao, 2022).

La finalidad de un algoritmo de entrenamiento para árboles de decisión radica en construir un árbol basado en un conjunto de datos y, en última instancia, generar un conjunto de reglas para realizar predicciones ante nuevas muestras. En la Figura 2.15 se presenta un ejemplo de visualización de un árbol de decisión de clasificación multiclase de un conjunto de datos Iris.

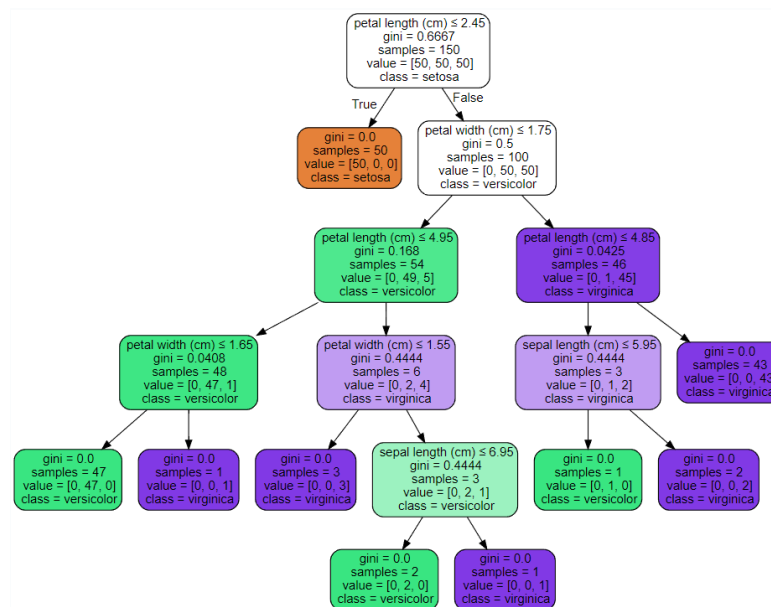


Figura 2.15 Árbol de decisión de clasificación multiclase

Fuente: (Pedregosa F. , y otros, 2011)

De acuerdo a (Pedregosa F. , y otros, 2011), dado un conjunto de vectores de entrenamiento $x_i \in R^n$, $i=1, \dots, i$ y un vector de etiqueta $y \in R^l$, un árbol de decisión divide de manera recursiva el espacio de características, agrupando las muestras con etiquetas similares.

Para un nodo dado m , se representa los datos como Q_m con n_m muestras. Para cada candidato $\theta = (j, t_m)$ de división que consiste en una característica j y un umbral t_m , los datos se dividen en dos subconjuntos $Q_m^{left}(\theta)$ y $Q_m^{right}(\theta)$ de acuerdo a las condiciones de la Ecuación 4 y 5.

$$Q_m^{left}(\theta) = \{(x, y) | x_j \leq t_m\} \quad \text{Ecuación 4}$$

$$Q_m^{right}(\theta) = Q_m \setminus Q_m^{left}(\theta) \quad \text{Ecuación 5}$$

La calidad de una división candidata en un nodo. m se calcula mediante una función de impureza o función de pérdida $H()$ como se muestra en la Ecuación 6, cuya elección depende de la tarea a resolver, ya sea clasificación o regresión.

$$G(Q_m, \theta) = \frac{n_m^{left}}{n_m} H(Q_m^{left}(\theta)) + \frac{n_m^{right}}{n_m} H(Q_m^{right}(\theta)) \quad \text{Ecuación 6}$$

Se seleccionan los parámetros que minimizan la impureza. Este proceso se repite de manera recursiva para los subconjuntos $Q_m^{left}(\theta^*)$ y $Q_m^{right}(\theta^*)$ hasta alcanzar la profundidad máxima permitida $n_m < \min_{samples}$ o $n_m = 1$.

Los árboles de decisión ofrecen varias ventajas, como su fácil comprensión y visualización intuitiva, así como la mínima preparación de datos, eliminando la necesidad de normalización y manipulación extensiva. La complejidad logarítmica en la predicción de datos, su capacidad para manejar variables numéricas y categóricas, y su naturaleza de modelo de caja blanca son rasgos notables. Además, su idoneidad para problemas de múltiples salidas y la capacidad de validación mediante pruebas estadísticas aportan confiabilidad.

Sin embargo, presentan desventajas, como la propensión al sobreajuste, mitigada mediante técnicas como la poda o la restricción de la profundidad del árbol. La inestabilidad, que genera árboles diferentes ante pequeñas variaciones en los datos, se atenúa mediante el uso de árboles en conjuntos. Las predicciones no son uniformes ni continuas, limitando la capacidad de extrapolación (Pedregosa F. , y otros, 2011).

2.3.2.4.5 Ensemble Learning

En el ámbito de Machine Learning, los métodos de conjunto se fundamentan en la idea de combinar las predicciones de varios modelos predictivos con el fin de lograr resultados más precisos que los obtenidos por un único predictor. Este concepto, conocido como aprendizaje en conjunto, se materializa mediante algoritmos de conjunto. La esencia de este enfoque guarda similitud con el principio de la sabiduría de la multitud, que destaca la capacidad de combinar respuestas provenientes de múltiples individuos para abordar interrogantes complejos. En este contexto, se sugiere que, en muchas situaciones, la respuesta colectiva supera la de un individuo experto. Así, el aprendizaje en conjunto se refiere a la práctica de fusionar las predicciones de un grupo de modelos predictivos para mejorar la precisión en comparación con un predictor individual. El conjunto de estos predictores se denomina simplemente conjunto, y la técnica que emplea este enfoque se conoce como método de conjunto en el ámbito del aprendizaje automático (Géron, 2019).

2.3.2.4.5.1 Random Forest

Random Forest es un algoritmo que emplea varios árboles de decisión, ya que un solo árbol podría resultar insuficiente para ciertas aplicaciones. Este método genera de manera aleatoria un conjunto de árboles de decisión, cada uno operando con un subconjunto aleatorio de datos. Posteriormente, Random Forest combina las salidas de cada árbol de decisión individual para obtener la salida final. Este algoritmo de aprendizaje en conjunto es versátil y aplicable tanto a problemas de clasificación como de regresión (Xiao, 2022).

El algoritmo del Bosque Aleatorio introduce mayor aleatoriedad durante el crecimiento de los árboles; en lugar de buscar la característica óptima al dividir un nodo, se busca la mejor entre un subconjunto aleatorio de características. Esto conduce a una mayor diversidad de árboles, lo que, a su vez, implica un intercambio de un mayor sesgo por una menor variación, generalmente resultando en un modelo globalmente más eficaz (Géron, 2019).

La Figura 2.16 presenta una representación visual del procedimiento del algoritmo Random Forest. La utilización de varios árboles de decisión en este enfoque contribuye a mitigar el sobreajuste, al tiempo que mejora el rendimiento global del modelo.

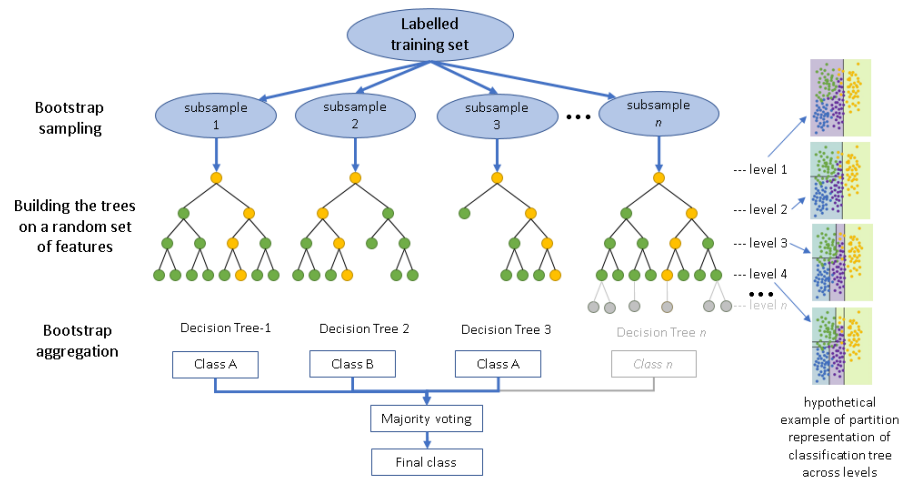


Figura 2.16 Procedimiento del algoritmo Random Forest

Fuente: (Kumar, 2021)

El algoritmo Random Forest aborda eficazmente el riesgo de sobreajuste interno a los árboles de decisión individuales al incorporar múltiples árboles no correlacionados. Esta estrategia reduce la varianza general y el error de predicción, mejorando así la generalización y robustez del modelo. Su versatilidad se evidencia al abordar problemas tanto de regresión como de clasificación con alta precisión, y su capacidad para manejar la imputación de valores faltantes refuerza su utilidad en la manipulación de datos incompletos.

Sin embargo, los bosques aleatorios no están exentos de desventajas. El proceso puede volverse intensivo en tiempo al trabajar con grandes conjuntos de datos, ya que el cálculo de predicciones para cada árbol individual puede ser computacionalmente costoso. Además, la necesidad de procesar conjuntos de datos extensos requiere una cantidad considerable de recursos para el almacenamiento de información. Aunque los bosques aleatorios ofrecen precisión, la interpretación de predicciones se vuelve más compleja debido a su naturaleza compuesta en comparación con un solo árbol de decisión.

Random Forest cuenta con tres hiperparámetros claves que deben ajustarse previo al proceso de entrenamiento. Estos son el tamaño del nodo, la cantidad de árboles en el bosque y la cantidad de características que se deben muestrear (IBM, 2023).

2.3.2.4.5.2 Gradient Boosting

Un modelo de Gradient Boosting se compone de varios árboles de decisión individuales que se entrenan de manera secuencial, donde cada nuevo árbol se esfuerza por corregir los errores de los árboles anteriores. La predicción de una nueva observación se calcula mediante la agregación de las predicciones de todos los árboles individuales que componen el modelo.

De acuerdo a (Pedregosa F. , y otros, 2011), el procedimiento Gradient Boosting para la clasificación guarda similitudes con el caso de regresión. No obstante, la suma de los árboles $F_M(x_i) = \sum_m h_m(x_i)$ no conduce a una predicción homogénea, ya que no puede representar directamente una clase, dado que los árboles realizan predicciones de valores continuos. La conversión del valor predicho $F_M(x_i)$ a una clase o probabilidad está vinculada a la función de pérdida. En el caso de la pérdida logarítmica, la probabilidad de que x_i pertenezca a la clase positiva se modela mediante la función sigmoidea $p(y_i = 1|x_i) = \sigma(F_M(x_i))$ donde σ es la salida del modelo.

Para problemas de clasificación multiclase, se construyen árboles distintos k (uno por cada clase k) en cada M iteración. La probabilidad de que x_i pertenezca a la clase k se modela mediante un softmax de la $F_{M,k}(x_i)$ valores. Es relevante señalar que, incluso en una tarea de clasificación, el h_m subestimador sigue siendo un regresor, no un clasificador, debido a que está entrenado para predecir gradientes (negativos), los cuales son siempre cantidades continuas.

Las fortalezas de Gradient Boosting son notables. En primer lugar, destaca por su capacidad automática para seleccionar predictores. Además, tiene la flexibilidad de manejar tanto predictores numéricos como categóricos sin necesidad de recurrir a la creación de variables. Un atributo distintivo es que exige considerablemente menos limpieza y preprocesamiento de datos en comparación con otros métodos de aprendizaje estadístico, simplificando así la fase de preparación de datos.

Sin embargo, los inconvenientes de los bosques aleatorios también deben ser considerados. Al fusionar varios árboles, se sacrifica la interpretabilidad que caracteriza a los modelos basados en un solo árbol. En el contexto de predictores continuos, los bosques aleatorios pierden parte de su información al categorizarlos en el momento de la división de los nodos.

Además, presentan limitaciones en la capacidad de extrapolación fuera del rango de los predictores observados en los datos de entrenamiento. Por su destacado rendimiento, Gradient Boosting ha surgido como el algoritmo predilecto para el manejo de datos tabulares, propiciando el desarrollo de diversas implementaciones. Cada una de estas variantes presenta características específicas que las vuelven más idóneas según el contexto de uso (Amat Rodrigo, 2023).

2.3.2.4.5.3 XG Boost

La técnica de aprendizaje supervisado conocida como Extreme Gradient Boosting, XG Boost, se basa también en árboles de decisión. Este algoritmo implica la formación secuencial de árboles de decisión en un ensamblado llamado Classification and Regression Trees (CART). La incorporación gradual de árboles tiene como objetivo aprender de los resultados anteriores y corregir los errores generados, utilizando el método de gradiente descendente hasta que no se pueda rectificar más el error.

Además, el XG Boost implementa características como el procesamiento en paralelo, la poda de árboles, la gestión de valores faltantes y la regularización, con el fin de prevenir el sobreajuste y el sesgo del modelo (Chen & Guestrin, 2016).

El proceso de ajuste de un conjunto de entrenamiento con XG Boost implica una serie de pasos. Inicialmente, se realiza una predicción inicial en el conjunto de datos. Luego, se calculan los residuos residuales en la diferencia entre el valor predicho y los valores observados. Un árbol de decisión se construye utilizando estos residuos, utilizando una puntuación de similitud de los mismos. La similitud de los datos en una hoja y la ganancia de similitud de la división posterior se determinan y se comparan para establecer una entidad y un umbral para un nodo. El valor de salida de cada hoja se calcula considerando los residuos, especialmente en el caso de la clasificación, donde se emplea el registro de momios y las probabilidades para calcular los valores. La salida del árbol se convierte en los nuevos residuos del conjunto de datos, utilizados para construir otro árbol. Este proceso se repite hasta que los residuales dejan de disminuir o se alcanza el número de iteraciones especificado. Es relevante destacar que cada árbol subsiguiente aprende de los árboles previos sin tener asignado el mismo peso, a diferencia del funcionamiento del Random Forest (ArcGISPro, 2023).

Según (Espinoza Zúñiga, 2020), el algoritmo XG Boost presenta diversas ventajas, entre las que se destacan su capacidad para manejar grandes volúmenes de datos con múltiples variables, la habilidad para gestionar valores perdidos, la obtención de resultados altamente precisos y una ejecución rápida y eficiente. No obstante, se han identificado algunas desventajas significativas. En primer lugar, el XG Boost consume recursos computacionales considerables en bases de datos extensas, por lo que se recomienda una evaluación detallada de las variables más informativas antes de aplicar la técnica en conjuntos de datos de gran escala.

Además, es crucial ajustar adecuadamente los parámetros del algoritmo para minimizar el error de precisión y prevenir el sobreajuste del modelo, especialmente cuando se trabaja con un número considerable de árboles. Finalmente, es importante tener en cuenta que el XG Boost solo opera con vectores numéricos, lo que implica la necesidad de convertir previamente los tipos de datos no numéricos a formatos numéricos antes de la aplicación del algoritmo.

2.3.2.4.6 K-Nearest Neighbors

K-Nearest Neighbors (K-NN) es un algoritmo que emplea K puntos más cercanos para determinar la clasificación de un conjunto de datos (Xiao, 2022). El aprendizaje supervisado basado en vecinos se presenta en dos modalidades: clasificación para datos con etiquetas discretas y regresión para datos con etiquetas continuas.

La esencia de los métodos del vecino más cercano radica en identificar un número predeterminado de muestras de entrenamiento más cercanas al nuevo punto y predecir la etiqueta a partir de ellas. Este número de muestras puede ser una constante definida por el usuario o variar según la densidad local de puntos. En términos generales, la distancia puede ser medida mediante diversas métricas, siendo la distancia euclidiana estándar la opción más común. Los métodos basados en vecinos se reconocen como métodos de aprendizaje automático no generalizadores, ya que simplemente retienen toda la información de entrenamiento.

A pesar de su simplicidad, los vecinos más cercanos han demostrado su eficacia en numerosos problemas de clasificación y regresión, incluido el reconocimiento de dígitos escritos a mano y el análisis de escenas en imágenes de satélite. Dada su naturaleza no paramétrica, este método suele destacar en situaciones de clasificación caracterizadas por límites de decisión altamente irregulares (Pedregosa F. , y otros, 2011).

Finalmente, en la Figura 2.17, se presenta una comparación visual de los clasificadores previamente abordados. La finalidad de este ejemplo es destacar la naturaleza de los límites de decisión de diversos clasificadores. Los gráficos representan los puntos de entrenamiento mediante colores sólidos, mientras que los puntos de prueba se presentan semitransparentes, y en la sección inferior derecha del gráfico exhibe la precisión de la clasificación en el conjunto de prueba.

En la elección de un algoritmo, resulta crucial tener en cuenta que, en ausencia de cualquier suposición sobre los datos, no hay razón para favorecer un modelo sobre otro. Este fenómeno es conocido como el teorema de No Free Lunch. Para determinados conjuntos de datos, un modelo lineal podría ser el más adecuado, mientras que, para otros conjuntos, una red neuronal podría ofrecer un rendimiento superior. No existe un modelo que tenga una garantía previa de ser superior. La única forma de conocer con certeza cuál es el modelo más adecuado implica evaluar todas las opciones disponibles (Géron, 2019).

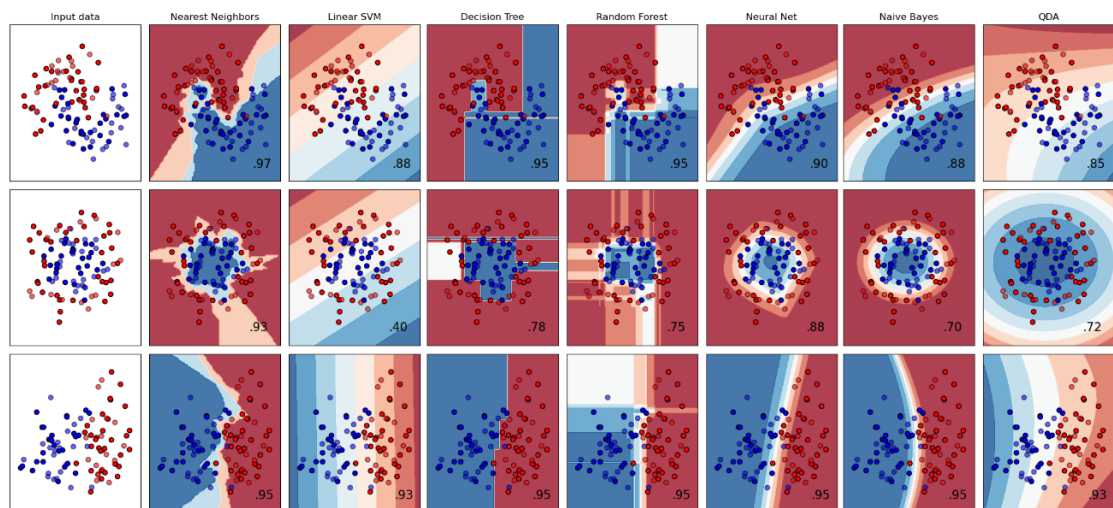


Figura 2.17 Comparación de clasificadores en ML

Fuente: (Pedregosa F. , y otros, 2011)

2.4 Evaluación de un modelo de clasificación

Para abordar exitosamente un problema de clasificación, es necesario identificar el algoritmo más adecuado a través de la aplicación de una o más métricas de evaluación del rendimiento en diversos algoritmos candidatos. Las métricas más utilizadas para evaluar la eficacia de los clasificadores son la tasa de errores y la tasa de aciertos. No obstante, numerosos experimentos han evidenciado que cuando las probabilidades a priori no están equilibradas, ambas medidas resultan inapropiadas, ya que no tienen en cuenta distintos tipos de errores. En consecuencia, estas métricas exhiben un marcado sesgo a favor de la clase mayoritaria (García Jiménez, 2010).

Por este motivo, es importante emplear diversas métricas para evaluar de manera precisa la capacidad de clasificación de un modelo. Algunos ejemplos reconocidos incluyen la matriz de confusión y la curva ROC-AUC. Estas métricas ofrecen una visión más clara del rendimiento del algoritmo, especialmente en situaciones de desequilibrio en la base de datos.

2.4.1 Generalidades

Para evaluar el rendimiento del modelo, se dispone de un conjunto de datos, el cual se divide en un número de instancias positivas (P) y un número de instancias negativas (N), cuya suma constituye la población total (POP). A continuación, se somete este conjunto de datos a un algoritmo de clasificación, se registran las salidas generadas y se comparan con las etiquetas reales de cada clase. De este proceso se obtienen cuatro medidas fundamentales: TP, TN, FP y FN.

En la Tabla 2.1 se especifican los valores generales de un problema de clasificación.

Tabla 2.1 Valores generales de un problema de clasificación

Métrica	Nombre	Traducción	Descripción
POP	Population	Población	Número de elementos por clase
P	Positive	Positivo	Número de muestras positivas
N	Negative	Negativo	Número de muestras negativas
TP	True Positive	Verdadero Positivo	Detección de la condición cuando está presente
TN	True Negative	Verdadero Negativo	No detección de la condición cuando no está presente
FP	False Positive	Falso Positivo	Detección de la condición cuando no está presente (error estadístico Tipo I)
FN	False Negative	Falso Negativo	No detección de la condición cuando está presente (error estadístico Tipo II)

2.4.2 Métricas de evaluación

La Tabla 2.2 presenta las fórmulas junto con breves descripciones de las métricas utilizadas en esta investigación para la evaluación del desempeño de los algoritmos de clasificación en contextos caracterizados por desequilibrios en la distribución de las clases. Estas métricas posibilitan una evaluación exhaustiva del rendimiento de un modelo en cada clase de manera individual, lo cual resulta fundamental en situaciones donde las clases presentan disparidades.

Tabla 2.2 Descripción de las métricas de evaluación

Métrica	Traducción	Descripción	Fórmula
Accuracy (Acc)	Exactitud	Número de predicciones correctas de todas las predicciones	$Acc = \frac{TP + TN}{TP + TN + FP + FN}$
Balanced Accuracy	Exactitud Balanceada	Mide el rendimiento donde las clases están desequilibradas	$Balanced_Acc = \frac{TPR + TNR}{2}$
Cohen's Kappa Index	Índice Kappa de Cohen	Mide la concordancia entre las predicciones y las etiquetas reales	$Kappa = \frac{(P_{obs} - P_{exp})}{(1 - P_{exp})}$
Error Rate (Err)	Tasa de Error	Número de predicciones incorrectas de todas las predicciones	$Err = \frac{FP + FN}{TP + TN + FP + FN}$
F1-Score	Medida F1	Promedio armónico de la Precisión y Recall	$F1 = 2 * \frac{PPV * TPR}{PPV + TPR}$
False Positive Rate (FPR)	Tasa de Falsos Positivos	Proporción de todos los negativos que aun así salieron positivos	$FPR = \frac{FP}{FP + TN} = 1 - TNR$
False Negative Rate (FNR)	Tasa de Falsos Negativos	Proporción de todos los positivos que la prueba detecta como negativo	$FNR = \frac{FN}{TP + FN} = 1 - TPR$
Jaccard Index	Índice de Jaccard	Mide la similitud y diversidad de los conjuntos de muestra	$J(y, \hat{y}) = \frac{ y \cap \hat{y} }{ y + \hat{y} - y \cap \hat{y} }$
Matthews Correlation Coefficient	Coeficiente de correlación de Matthews	Mide la calidad de las predicciones	$MCC = \frac{(TP * TN - FP * FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$
Negative Predictive Value (NPV)	Valor Predictivo Negativo	Proporción de negativos que corresponden a ausencia de condición	$NPV = \frac{TN}{TN + FN}$
Positive Predictive Value (PPV)	Valor Predictivo Positivo	Precisión, proporción de positivos que corresponden a la presencia de condición	$PPV = \frac{TP}{TP + FP}$
True Negative Rate (TNR)	Tasa de Verdaderos Negativos	Especificidad, mide la proporción de negativos que son correctamente clasificados	$TNR = \frac{TN}{TN + FP}$
True Positive Rate (TPR)	Tasa de Falsos Positivos	Sensibilidad o Recall, mide la proporción de positivos que son identificados como tales	$TPR = \frac{TP}{TP + FN}$

2.4.3 Matriz de confusión

Las matrices de confusión constituyen una herramienta valiosa para proporcionar una descripción integral del rendimiento de un modelo predictivo al aplicar un conjunto de pruebas. Asimismo, sirven como base para una variedad extensa de medidas de rendimiento que pueden poner de relieve distintos aspectos del desempeño del modelo predictivo. Entre las medidas más fundamentales se encuentran la Tasa de Verdaderos Positivos (TPR), Tasa de Verdaderos Negativos (TNR), Tasa de Falsos Negativos (FNR) y Tasa de Falsos Positivos (FPR) (Kelleher, Mac Name, & D'Arcy, 2015).

En el caso de problemas de clasificación binaria, la matriz de confusión suele representarse como una matriz de 2×2 , mientras que, para problemas de clasificación con múltiples clases, se adopta combinada una matriz de confusión de dimensiones $n \times n$, donde n es el número de clases. La estructura de la matriz de confusión se visualiza en la Figura 2.18.

		Estimado por el modelo	
		Negativo (N)	Positivo (P)
Real	Negativo (N)	TN	FP
	Positivo (P)	FN	TP

Figura 2.18 Estructura de la matriz de confusión

Fuente: (Kelleher, Mac Name, & D'Arcy, 2015)

2.4.4 Curva PR

La curva Precision-Recall es un indicador esencial para evaluar el rendimiento predictivo, especialmente en situaciones con desequilibrios notables entre las clases. Un área amplia bajo esta curva indica un rendimiento elevado tanto en precisión como en recuperación, donde la alta precisión se asocia con una baja tasa de falsos positivos y la alta recuperación con una baja tasa de falsos negativos. Puntuaciones elevadas en ambas dimensiones señalan que el clasificador genera resultados precisos y recupera la mayoría de los resultados positivos (Pedregosa F. , y otros, 2011). En la Figura 2.19 se visualiza un ejemplo de la curva Precision-Recall.

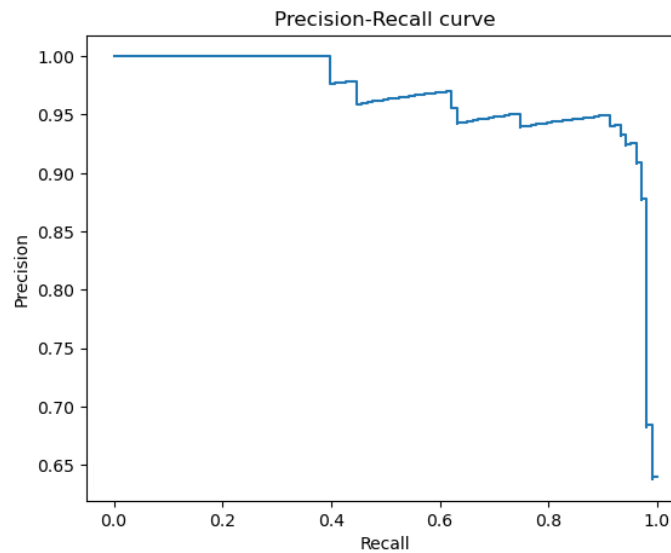


Figura 2.19 Curva PR

Fuente: (Pedregosa F. , y otros, 2011)

2.4.5 Curva ROC-AUC

Según (Japkowicz & Shah, 2011), la Curva ROC (Receiver Operating Characteristic) es una herramienta gráfica efectiva para evaluar el rendimiento de un algoritmo de aprendizaje en diferentes umbrales de decisión. En la Figura 2.20 se muestra la representación gráfica, el eje x muestra la Tasa de Falsos Positivos (FPR), mientras que el eje y refleja la Tasa de Verdaderos Positivos (TPR) de un clasificador. La TPR indica la sensibilidad del clasificador, y la FPR es complementaria a la especificidad del clasificador. La Curva ROC examina la relación entre la sensibilidad y la especificidad del clasificador.

Dado que tanto TPR como FPR cumplen con $0 \leq TPR \leq 1$ y $0 \leq FPR \leq 1$, la Curva ROC se representa en un cuadrado unitario. El punto (0,0) representa un clasificador trivial que etiqueta todas las instancias como negativas, con TPR y FPR iguales a cero. En el otro extremo, el punto (1,1) representa un clasificador que etiqueta todas las instancias como positivas, con valores unitarios tanto de TPR como de FPR. La diagonal que conecta estos puntos [(0,0) y (1,1)] tiene $TPR = FPR$ en todos los puntos, indicando que los clasificadores a lo largo de esta diagonal se consideran aleatorios, asignando etiquetas de manera aleatoria.

Los clasificadores por encima de esta diagonal superan a los aleatorios, mientras que los debajo presentan un rendimiento inferior. Los puntos (1,0) y (0,1) son los extremos restantes del espacio ROC. El punto (1,0) indica un FPR de 1 y un TPR de 0, lo que implica que el clasificador comete errores en todas sus predicciones. En contraste, el punto (0,1) representa el clasificador ideal, que acierta en todos los aspectos positivos y no comete errores negativos.

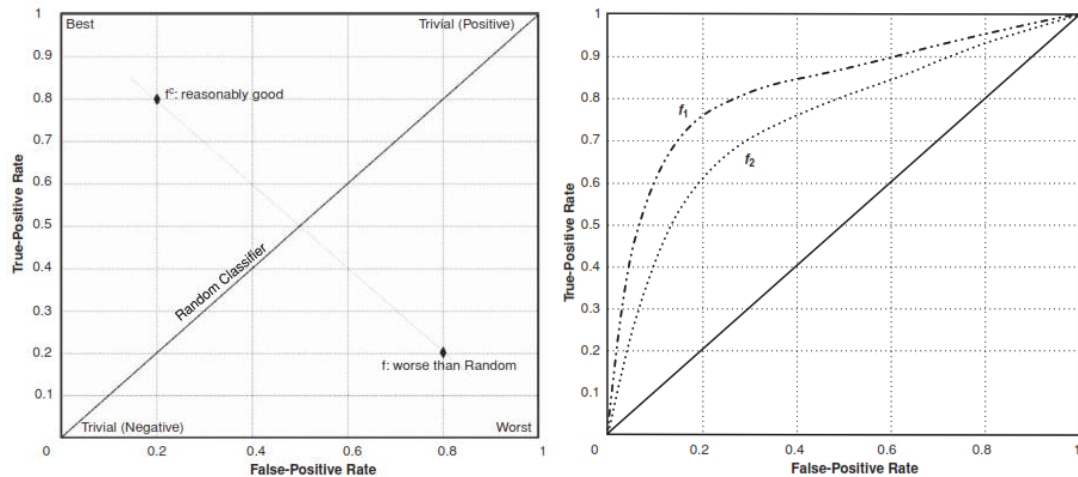


Figura 2.20 Curva ROC

Fuente: (Japkowicz & Shah, 2011)

Para la elección entre la Curva ROC y la Curva PR, es crucial considerar la naturaleza del análisis deseado. La Curva PR debe preferirse cuando el análisis de falsos positivos es más relevante que el de falsos negativos, mientras que la Curva ROC resulta más apropiada en el caso contrario (Géron, 2019).

2.5 Conclusión

Antes de adentrarnos en la metodología propuesta, es fundamental abordar las principales características de los generadores de potencia, así como explorar el análisis de descargas parciales como un método crucial para la detección de posibles fallos en estos equipos eléctricos. Asimismo, la descripción de los algoritmos de clasificación de ML y las métricas de evaluación asociadas ofrece un conocimiento esencial para la comprensión y desarrollo del presente trabajo.

CAPÍTULO 3. METODOLOGÍA

3.1 Introducción

En este capítulo, se exploran los conocimientos y algoritmos de Machine Learning (ML) previamente discutidos en el marco teórico, con el objetivo de desarrollar un modelo inteligente para diagnosticar fallas en generadores de potencia del sistema eléctrico nacional. Este enfoque involucra la utilización de diversos indicadores para evaluar el nivel de degradación de los componentes del generador.

El capítulo se inicia con una descripción general de la metodología, proporcionando una visión panorámica del desarrollo del modelo inteligente. Posteriormente, se detallan las actividades realizadas en cada etapa, comenzando con un análisis exhaustivo de la problemática en estudio y la formulación de una propuesta de solución. Continuando con la adquisición de los datos necesarios sobre las fallas de los generadores de potencia, permitiendo un análisis de sus atributos y la aplicación de procesos de limpieza y pre-procesamiento de datos. El siguiente paso implica la implementación de diez algoritmos de clasificación de ML, cuyo rendimiento será comparado y evaluado a través de métricas apropiadas. Finalmente, se procede a ajustar el modelo para su posterior implementación a un sistema.

3.2 Metodología básica

Cuando se emplea ML con el fin de permitir que un sistema aprenda a partir de datos y se obtengan resultados de predicción satisfactorias, resulta fundamental adherirse a un proceso claramente definido. Esto garantiza una orientación efectiva durante el desarrollo y proporciona una guía para abordar situaciones en las que los resultados no alcancen las expectativas. En este sentido, el desarrollo del presente trabajo de investigación se sustenta en una metodología básica de un proceso de ML. En la Figura 3.1 que se presenta a continuación, se ilustra el diagrama de flujo de esta metodología, ofreciendo una visión visual de los pasos que se seguirán en el transcurso de la investigación.

Según lo señalado por (Géron, 2019), la metodología consta de siete etapas que tienen como objetivo principal transformar una problemática en un sistema funcional. En este contexto, se ha aplicado esta metodología para desarrollar un sistema de diagnóstico de fallas en generadores de potencia. A continuación, se detallan las etapas de esta metodología:

- b) **Análisis General:** En esta etapa, se enfoca en comprender los objetivos del proyecto, delimitar la problemática de interés, conocimiento del proceso, buscar posibles soluciones y generar una propuesta de solución.
- c) **Obtención de Datos:** Aquí se determina y recopilan los datos necesarios para abordar la problemática.
- d) **Exploración:** En este punto, se exploran los datos para obtener información. Esto se logra mediante un análisis de atributos, la aplicación de la estadística, la visualización de datos y la identificación de posibles transformaciones.
- e) **Pre-procesamiento:** Esta etapa comprende la limpieza de datos, la selección y escalado de atributos, y la aplicación de ingeniería de características, con el propósito de extraer conocimiento y mejorar la comprensión de los datos, generando una base de datos de aprendizaje consolidada.
- f) **Elección algorítmica:** En esta fase, se seleccionan algoritmos pertinentes para abordar la problemática y se evalúa su rendimiento utilizando diversas métricas. Este análisis permite elegir el algoritmo más adecuado y confiable.
- g) **Ajuste del modelo:** En este punto se calibran los parámetros del modelo para asegurarse de que sea la mejor opción para resolver el problema. Para ello, se implementa un algoritmo de optimización, se entrena y evalúa el modelo.
- h) **Implementación:** La última etapa consiste en convertir el modelo en un sistema completo capaz de interactuar con expertos y usuarios, permitiendo su aplicación en situaciones prácticas.

Estas siete etapas conforman la metodología empleada en este trabajo de investigación para desarrollar el sistema de diagnóstico de fallas en generadores de potencia.

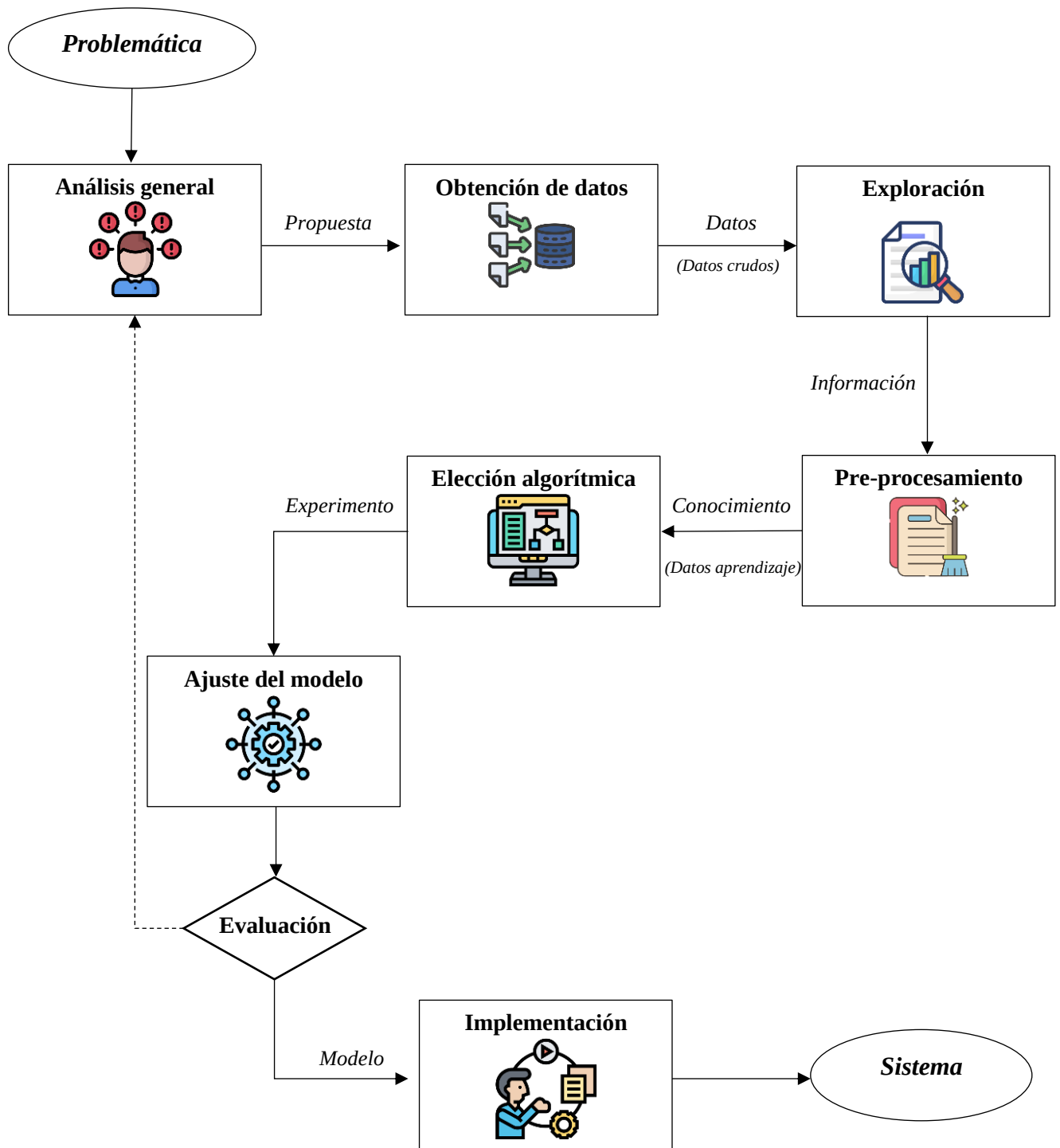


Figura 3.1 Diagrama de flujo de la metodología

3.3 Análisis general

En esta fase inicial, se lleva a cabo un análisis integral de la problemática con el objetivo de obtener una perspectiva global del proceso, estableciendo así los límites y alcances de la misma. Asimismo, se busca una comprensión profunda de los objetivos del proyecto, conocimiento del proceso, proporcionando un marco de referencia esencial y delineando el espacio en el que se buscarán soluciones. En esta etapa inicial, también se conceptualiza un plan preliminar que servirá como guía para alcanzar los objetivos propuestos.

Dentro de este análisis de la problemática, se abordan cuestiones claves, tales como: ¿cuál es la naturaleza del problema?, ¿cuáles son las soluciones existentes?, ¿cuál es el proceso manual para resolver el problema?, y ¿cuáles serán los indicadores utilizados para medir el rendimiento de la solución? Estos interrogantes ayudan a aclarar la problemática y a definir la dirección del proyecto. Para llevar a cabo este análisis general, se recurre a la información proporcionada en los dos primeros capítulos de este trabajo, donde se define de manera detallada tanto el problema en sí como el espacio de las soluciones pertinentes.

3.3.1 Definir el problema bajo el dominio

El problema que se aborda en esta investigación se centra en el diagnóstico de los tipos de fallas que pueden afectar a los generadores de potencia en centrales hidroeléctricas y termoeléctricas del sistema eléctrico nacional. Se busca lograr este diagnóstico utilizando diversos indicadores que permitan cuantificar el nivel de degradación de los componentes de los generadores, sin depender de una inspección visual realizada por expertos en el ámbito eléctrico y energético de CFE.

En la actualidad, la detección y diagnóstico de fallas en generadores en el sistema eléctrico nacional se lleva a cabo mediante la aplicación de técnicas tanto en línea como fuera de línea. Estas técnicas se basan principalmente en la monitorización de parámetros relacionados con las descargas parciales y en la inspección visual. Estas técnicas han demostrado su eficacia en la identificación de los mecanismos de deterioro que afectan a los componentes de los generadores.

Sin embargo, en este estudio se propone una aproximación innovadora al emplear algoritmos de ML, los cuales tienen la capacidad de aprender de forma autónoma los valores de los indicadores que permiten distinguir los diferentes tipos de fallas que pueden presentar los generadores. Este enfoque promete ofrecer una solución más eficiente para el diagnóstico de fallas en generadores de potencia, reduciendo la dependencia de inspecciones visuales y el análisis de los expertos en generadores, mejorando la capacidad predictiva de fallas.

3.3.2 Búsqueda de soluciones previas

En el estado del arte, se presentan soluciones previas para abordar el desafío del diagnóstico de fallas en generadores de potencia. Algunos autores han implementado técnicas de IA como la lógica difusa y las redes neuronales artificiales. La técnica LD se ha empleado exitosamente para el análisis cualitativo de sistemas complejos. Sin embargo, presenta dificultades al tratar de analizar la tendencia de la falla, y la precisión del diagnóstico depende en gran medida de la selección adecuada de los parámetros y el conjunto de reglas expertas que capturen adecuadamente el conocimiento humano de un dominio específico. Por otro lado, las RNA son altamente aplicables en sistemas donde resulta complicado extraer el conocimiento de expertos. No obstante, una de sus principales limitaciones radica en la necesidad de disponer de una gran cantidad de datos para su entrenamiento, con el fin de garantizar la fiabilidad del diagnóstico de fallas.

Las nuevas estrategias de detección y diagnóstico deben enfocarse en mejorar tanto la robustez como la adaptabilidad del sistema, permitiendo su evolución continua. Además, es fundamental que estos diagnósticos sean interpretables por el operador del generador, de manera que se facilite su comprensión y toma de decisiones. Para esto, los algoritmos de ML son más adecuados cuando se trata de problemas con conjuntos de datos, patrones complejos o cuando se busca la automatización del aprendizaje a partir de datos.

3.3.3 Determinar tipo de problema

El problema inherente al diagnóstico de fallas en generadores de potencia se cataloga como un problema de clasificación multiclase, dado que el diagnóstico del generador se divide en múltiples categorías de fallas. En consecuencia, las métricas de rendimiento (desempeño) de los

algoritmos de ML se basan en una combinación de medidas que incluyen el F1-score, Recall, la exactitud, la precisión, curva ROC y curva PR.

En este contexto, se considerará un nivel de significancia mínimo del 10% en la evaluación de resultados. Esta elección se efectúa con el propósito de obtener resultados más confiables y robustos que permitan una evaluación precisa y significativa del rendimiento de los algoritmos de diagnóstico de fallas en generadores de potencia.

3.3.4 Generación de supuestos

En el marco de la clasificación de las fallas en generadores, se parte del supuesto de que los indicadores de descargas parciales y los factores de disipación albergan una cantidad sustancial de información. Esta información se considera suficiente para descifrar el grado de degradación de los componentes del generador y, asimismo, para identificar con precisión el tipo de falla que se manifiesta en el equipo.

3.3.5 Propuesta

Se propone el desarrollo de un modelo inteligente basado en algoritmos de ML con la capacidad de clasificar de manera precisa las fallas que afectan a los generadores de potencia. Este diagnóstico se llevará a cabo mediante la evaluación de indicadores que cuantifican el grado de degradación de sus componentes. El modelo aspira a ofrecer un rendimiento óptimo y ser de fácil utilización para los expertos de la CFE. El cumplimiento integral de todos los objetivos establecidos en este trabajo de investigación constituye un requisito esencial.

3.4 Obtención de datos

En esta segunda etapa, se procede a la identificación y recopilación de los datos necesarios para abordar la problemática planteada. La precisión en la adquisición de datos se convierte en un elemento crítico para asegurar la integridad y confiabilidad del estudio. Adicionalmente, es de suma importancia comprender la naturaleza de los datos y los procesos con los que se va a trabajar.

En este caso, nos encontramos con datos tabulares que incluyen tanto valores categóricos como numéricos. Esta comprensión resulta fundamental, ya que influye en la elección de los algoritmos que serán empleados para resolver la problemática en cuestión.

3.4.1 Determinar datos necesarios

Dado que este proyecto se centra en la aplicación de ML, resulta altamente preferible contar con una cantidad significativa de datos, es decir, un mayor número de filas en un archivo tabular. Esta abundancia de datos permitirá que los algoritmos de aprendizaje automático adquieran un mayor nivel de precisión en sus resultados. Asimismo, se hace necesario reunir una amplia gama de características, representadas como columnas en un archivo tabular, para evaluar y comprender el impacto de cada una de ellas en la etiqueta de salida.

En el contexto de esta tesis, se requieren datos específicos, que incluyen información sobre las descargas parciales, los valores de los factores de disipación, los tipos de sistemas de enfriamiento, y los tipos de generadores de potencia. Estos datos deben estar debidamente clasificados y validados por expertos en el campo. Además, es fundamental contar con información detallada sobre los protocolos de análisis aplicados en la obtención y procesamiento de estos datos.

3.4.2 Capturar u obtener datos

Los datos son obtenidos de generadores eléctricos nacionales y datos de generadores reportados en la bibliografía. Las mediciones de las descargas parciales se realizan tanto en línea como fuera de línea. Para las mediciones fuera de línea, se emplea un circuito de detección directa que involucra la inserción de una impedancia en serie con un capacitor de bloqueo. La corriente generada por las descargas parciales atraviesa esta impedancia y se transforma en una señal de tensión, la cual se amplifica y envía a un detector de descargas parciales para su posterior medición.

En el caso de las mediciones en línea, se utilizan sistemas de Análisis de Generadores en Línea junto con diversos tipos de sensores, como sensores capacitivos, inductivos y tipo antena. Estos sensores se acoplan a las terminales de línea, las terminales neutras, los cables de las salidas de fase en generadores de baja capacidad y a los capacitores supresores de transitorios, según el tipo de sensor requerido. Para evaluar el nivel de descargas parciales en un sistema aislante, se

suelen seguir las recomendaciones establecidas por el CRIEPI (Central Research Institute of Electric Power Industry).

La base de datos crudos utilizada en este estudio es proporcionada por expertos en generadores eléctricos del INEEL. Esta base de datos contiene información relevante en las siguientes columnas: la ubicación de la planta donde se encuentra instalado el generador, la unidad correspondiente, el tipo de sistema de enfriamiento, los valores de los factores de disipación, las mediciones de las descargas parciales y la clasificación de las fallas específicas que se presentan en los generadores de potencia.

3.4.3 Crear entorno de trabajo

Los entornos de trabajo desempeñan un papel fundamental en la gestión del proceso de ML, ya que facilitan la creación de flujos de trabajo de aprendizaje automático que son reproducibles, auditables y portátiles en diversos destinos de procesamiento. En este contexto específico, se elige Python¹ como el lenguaje de programación primario para la implementación de la metodología básica de ML. Para gestionar de manera efectiva los paquetes y entornos necesarios, se opta por emplear Anaconda como gestor de paquetes, respaldado por su interfaz gráfica, Anaconda Navigator, que se utiliza como herramienta integral para la gestión de entornos de trabajo. Y la utilización de Notebooks permiten crear modelos de manera interactiva y fácilmente.

3.4.3.1 Anaconda Navigator

La gestión del entorno de trabajo se lleva a cabo mediante Anaconda Navigator, una herramienta que no solo facilita el inicio de aplicaciones, sino que también proporciona la mayoría de los módulos esenciales necesarios para iniciar el análisis de datos en el contexto de ML. En la Figura 3.2 se presenta la ventana principal de este navegador de Anaconda como referencia visual.

Para habilitar la funcionalidad de paquetes y entornos dentro de Anaconda Navigator, es necesario llevar a cabo la instalación de módulos específicos que permitan su ejecución.

¹ Python es uno de los lenguajes más usados en la actualidad para Machine Learning.

En el contexto de este proyecto, se han identificado y utilizado los siguientes módulos de Python de relevancia:

- Jupyter: Este módulo posibilita la creación y el uso de documentos digitales interactivos.
- Pandas: Se emplea para la manipulación de datos tabulares, facilitando la organización y procesamiento de la información.
- Numpy: Es esencial en tareas de cómputo matemático, proporcionando funcionalidades que son fundamentales en la mayoría de los análisis de datos y operaciones matemáticas.
- Scikit-learn (Sklern): Este módulo es una herramienta esencial en el ámbito de Machine Learning y Ciencia de Datos, ya que ofrece una amplia gama de algoritmos y utilidades para la construcción de modelos predictivos.
- Matplotlib: Se utiliza para la generación de gráficos y figuras, lo que facilita la visualización de resultados y la presentación de datos de manera efectiva.

La incorporación de estos módulos en el entorno de trabajo es de gran utilidad para el desarrollo y ejecución de las tareas relacionadas con ML.

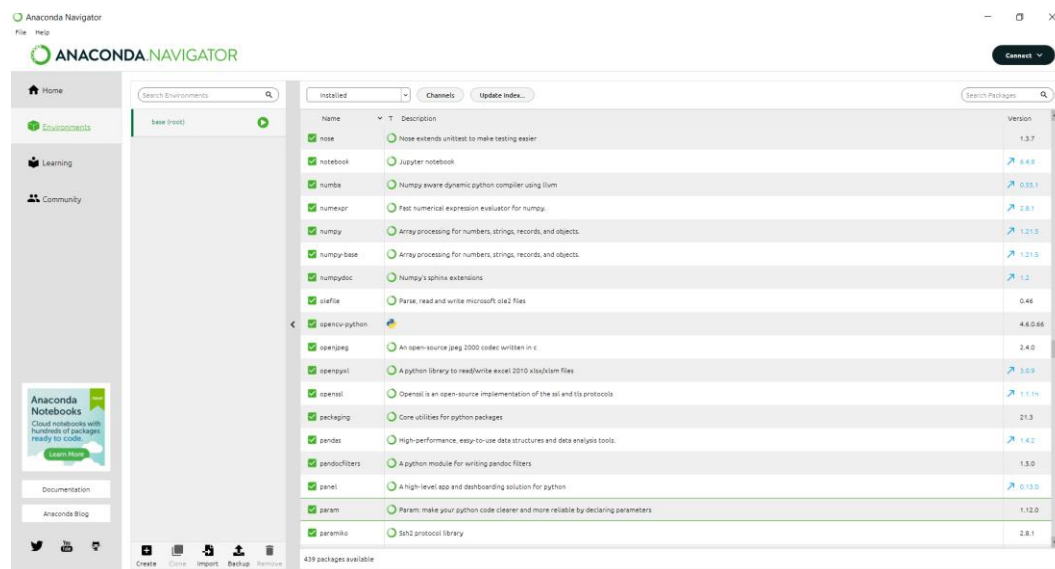


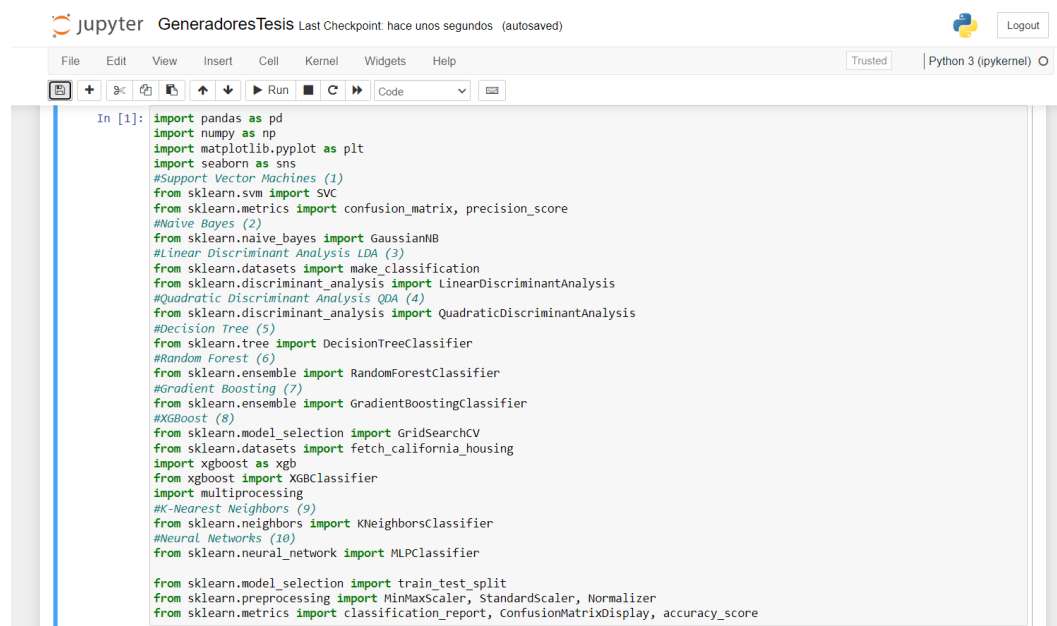
Figura 3.2 Anaconda Navigator

3.4.3.2 Jupyter Notebooks

La ejecución del código Python requerido para el desarrollo de este trabajo se lleva a cabo a través de la interfaz de Jupyter Notebook, tal como se ilustra en la Figura 3.3.

El documento Jupyter Notebook se estructura en distintas celdas, cada una con opciones variadas que permiten definir su contenido, que puede ser un título, texto descriptivo o código ejecutable. Cabe destacar que cada celda puede ejecutarse de manera individual, y los resultados se muestran de forma inmediata. Estos resultados pueden incluir datos numéricos, tablas, figuras y gráficos, dependiendo de la naturaleza de la tarea. Además de estas opciones, las bibliotecas específicas posibilitan la importación de imágenes o videos, la inclusión de enlaces a recursos en línea, e incluso la incorporación de una página web completa dentro del propio documento.

En el contexto de este trabajo en particular, se ha optado por importar un archivo con extensión csv. Este archivo contiene la base de datos que se utilizará en las pruebas y análisis relacionados con los generadores de potencia.



The screenshot shows the Jupyter Notebook interface with the title 'GeneradoresTesis'. The top bar includes a 'Logout' button and a 'Python 3 (ipykernel)' indicator. The main area displays a code cell with the following imports:

```
In [1]: import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
#Support Vector Machines (1)
from sklearn.svm import SVC
from sklearn.metrics import confusion_matrix, precision_score
#Naive Bayes (2)
from sklearn.naive_bayes import GaussianNB
#Linear Discriminant Analysis LDA (3)
from sklearn.datasets import make_classification
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
#Quadratic Discriminant Analysis QDA (4)
from sklearn.discriminant_analysis import QuadraticDiscriminantAnalysis
#Decision Tree (5)
from sklearn.tree import DecisionTreeClassifier
#Random Forest (6)
from sklearn.ensemble import RandomForestClassifier
#Gradient Boosting (7)
from sklearn.ensemble import GradientBoostingClassifier
#XGBoost (8)
from sklearn.model_selection import GridSearchCV
from sklearn.datasets import fetch_california_housing
import xgboost as xgb
from xgboost import XGBClassifier
import multiprocessing
#K-Nearest Neighbors (9)
from sklearn.neighbors import KNeighborsClassifier
#Neural Networks (10)
from sklearn.neural_network import MLPClassifier

from sklearn.model_selection import train_test_split
from sklearn.preprocessing import MinMaxScaler, StandardScaler, Normalizer
from sklearn.metrics import classification_report, confusionMatrixDisplay, accuracy_score
```

Figura 3.3 Interfaz de Jupyter Notebook

3.4.4 Datos

Como parte final de esta etapa, se genera un archivo en formato csv que alberga información relacionada con 18 características derivadas de pruebas llevadas a cabo por expertos en generadores. Estas características están asociadas a diferentes tipos de fallas en los generadores de potencia. El archivo abarca un total de 196 ejemplos, los cuales se van a convertir en conocimiento a través de la aplicación del entorno de trabajo computacional previamente instalado y configurado.

3.5 Exploración

La fase de exploración de datos comprende un análisis minucioso y una visualización detallada de los datos con el propósito de adquirir un nivel de familiaridad con ellos. Este proceso permite obtener información significativa y detectar patrones que puedan merecer un análisis más profundo. Además de cumplir con esta función, la exploración de datos también desempeña un papel fundamental en la estimación de si los datos disponibles son suficientes y pertinentes para la construcción de un modelo.

En esencia, este enfoque busca proporcionar una comprensión integral de los datos, lo que a su vez respalda la formulación de un plan de trabajo más experto y contribuye a la generación de un análisis más enriquecedor y valioso.

3.5.1 Análisis de atributos

En el contexto de este estudio, una característica se define como un atributo de entrada esencial que se utiliza en el proceso de entrenamiento del modelo. Estas características desempeñan un papel crítico al permitir que el modelo identifique patrones significativos, lo que resulta fundamental para la realización de predicciones precisas. Por consiguiente, es obligatorio que las características seleccionadas sean altamente relevantes para la problemática abordada en la tesis.

En este trabajo en particular, se hace necesario compilar un conjunto de datos que incluya detalles de mediciones relacionadas con las descargas parciales. Estas mediciones se consideran una vía efectiva para cuantificar el nivel de degradación de los componentes del generador.

En términos prácticos, se emplean datos tabulares para representar la información, donde cada columna representa un atributo específico, cada fila corresponde a una muestra y se adjunta una etiqueta que indica la clase a la que pertenece cada una de las muestras. Esta estructura de datos tabulares facilita el proceso de análisis y modelado, proporcionando una base sólida para el desarrollo del modelo de diagnóstico de fallas en generadores.

3.5.1.1 Atributos para la clasificación

La base de datos tabular destinada a este problema en particular consta de un total de 18 atributos, 196 muestras y una etiqueta denominada Diagnóstico (Tipo de Falla). No obstante, no todos estos atributos resultan pertinentes para la tarea de predicción, ya que carecen de información relevante en este contexto. En consecuencia, se procede a descartar cuatro de estos atributos: Planta, Unidad, Fecha, e Inspección visual. Cada uno de estos atributos se relaciona con el nombre de la planta a la que está asociado el generador, la unidad específica de la central, la fecha en que se llevó a cabo la prueba y las observaciones visuales realizadas en el generador, respectivamente.

Las Tablas 3.1 y 3.2 se presenta de manera detallada las columnas seleccionadas que serán objeto de análisis en términos de sus características y posteriormente serán utilizadas como entrada para los algoritmos de ML previamente elegidos. Esta selección cuidadosa de atributos optimiza el proceso de modelado y contribuye a la obtención de resultados más precisos y relevantes en el diagnóstico de fallas en generadores de potencia.

Tabla 3.1 Columnas del archivo y la característica que representa

Columna	Tipo	Característica
Enfriamiento	Categórica	Tipo de enfriamiento usado en el generador
Tipo	Categórica	Tipo de generador, voltaje y carga
FD_tand_A	Numérica	Factor de disipación $\tan \delta$ (0.2Vn) Ø A
FD_tand_B	Numérica	Factor de disipación $\tan \delta$ (0.2Vn) Ø B
FD_tand_C	Numérica	Factor de disipación $\tan \delta$ (0.2Vn) Ø C
FD_Dtand_A	Numérica	Factor de disipación $\Delta \tan \delta$ (0.6Vn-0.2Vn)) Ø A
FD_Dtand_B	Numérica	Factor de disipación $\Delta \tan \delta$ (0.6Vn-0.2Vn)) Ø B
FD_Dtand_C	Numérica	Factor de disipación $\Delta \tan \delta$ (0.6Vn-0.2Vn)) Ø C

Tabla 3.2 Columnas del archivo y la característica que representa (continuación)

Columna	Tipo	Característica
DP_VnCrT_Ap	Numérica	Descargas parciales (nC) $V_n/\sqrt{3}$ Ø A semiciclo (+)
DP_VnCrT_An	Numérica	Descargas parciales (nC) $V_n/\sqrt{3}$ Ø A semiciclo (-)
DP_VnCrT_Bp	Numérica	Descargas parciales (nC) $V_n/\sqrt{3}$ Ø B semiciclo (+)
DP_VnCrT_Bn	Numérica	Descargas parciales (nC) $V_n/\sqrt{3}$ Ø B semiciclo (-)
DP_VnCrT_Cp	Numérica	Descargas parciales (nC) $V_n/\sqrt{3}$ Ø C semiciclo (+)
DP_VnCrT_Cn	Numérica	Descargas parciales (nC) $V_n/\sqrt{3}$ Ø C semiciclo (-)

3.5.1.2 Conteo de clases y categorías

En la Tabla 3.3 se muestra el conteo de cada clase perteneciente a la etiqueta Diagnóstico (Tipo de Falla), donde se observa que la Clase 1 es la que tiene el mayor número de datos y la Clase 3 el menor número de datos.

Tabla 3.3 Conteo de clases y categorías

Clase	Categoría	Cantidad
0	Descargas parciales internas en huecos del aislamiento principal	43
1	Descargas en la zona de graduación de campo eléctrico en la interface de las pinturas conductora y semiconductora. Se incluyen las descargas por efecto de mala graduación por contaminación superficial	52
2	Descargas entre bobinas en cabezales o entre cabezales y dedos de sujeción del núcleo	33
3	Descargas parciales por delaminación del aislamiento principal y conductor de cobre	31
4	Descargas a la ranura (descargas entre la laminación del núcleo y la superficie de los bastones)	37

Generalmente, se considera que una base de datos es balanceada cuando todas las clases tienen aproximadamente la misma cantidad de muestras o están dentro de un rango razonable de diferencia. Algunos investigadores utilizan un umbral del 5 – 10% como referencia para determinar si una base de datos es suficientemente balanceada. Esto significa que, si la diferencia en la proporción de muestras entre clases es menor a dicho umbral, se considera que la base de datos es balanceada. Sin embargo, este valor es arbitrario y puede ajustarse según las necesidades específicas de la problemática (García García, Reding Bernal, & López Alvarenga, 2013).

Para determinar si la base de datos de este trabajo es balanceada o no, se establece un umbral de referencia del 10%. Se verifica que las cinco clases cumplan con la condición de tener una diferencia menor o igual a 19.6 datos entre ellas. La mayoría de las clases cumplen con esta condición, sin embargo, se observa una diferencia de 21 datos entre la clase 1 y la clase 3.

Este desequilibrio sustancial en una sola clase indica que la base de datos se considera desbalanceada. La presencia de una única clase con un desequilibrio significativo puede tener un impacto adverso en el rendimiento del modelo y su capacidad de generalización. Es importante que todas las clases en una base de datos estén equilibradas o dentro de un rango razonable de diferencia entre sí para garantizar que el modelo de ML no esté sesgado hacia una clase en particular y pueda aprender de manera efectiva todas las clases del problema. Si una clase tiene un desequilibrio sustancial, el modelo puede tener dificultades para aprenderla y, en consecuencia, su capacidad para hacer predicciones precisas en esa clase específica podría verse comprometida. Por lo que, cuando se evalúan algoritmos entrenados y probados con una base de datos desbalanceada, es importante utilizar métricas de evaluación que sean apropiadas y sensibles al desequilibrio de clases. Las métricas tradicionales como la exactitud (accuracy) pueden ser engañosas en este contexto, ya que un modelo puede parecer tener un buen rendimiento al predecir la clase mayoritaria en una base de datos desbalanceada, incluso si no está haciendo un buen trabajo en las clases minoritarias.

3.5.2 Aplicar estadísticas

En este apartado, se lleva a cabo una exploración de las características numéricas de los datos que conforman la base de datos. La Figura 3.4 presenta los resultados del análisis estadístico, que

incluye información sobre la cantidad de datos, la media, la desviación estándar, los valores máximos y mínimos, así como los cuantiles.

```
df.describe()
```

	FD_Dtand_A	FD_Dtand_B	FD_Dtand_C	DP_VnCrT_Ap	DP_VnCrT_An	DP_VnCrT_Bp	DP_VnCrT_Bn	DP_VnCrT_Cp	DP_VnCrT_Cn	Diagnostico
count	196.000000	196.000000	196.000000	196.000000	196.000000	196.000000	196.000000	196.000000	196.000000	196.000000
mean	0.941286	0.923931	0.953884	14.304490	14.247092	14.985765	15.094643	14.952245	15.271837	1.831633
std	0.758014	0.763781	0.804225	19.269028	19.114699	20.231753	20.232587	20.195916	20.611282	1.427646
min	0.060000	0.100000	0.100000	0.550000	0.600000	0.450000	0.500000	0.200000	0.300000	0.000000
25%	0.417500	0.420000	0.400000	2.100000	2.075000	2.275000	2.675000	2.200000	2.400000	1.000000
50%	0.620000	0.610000	0.600000	6.900000	6.600000	7.050000	7.000000	7.000000	6.700000	2.000000
75%	1.312500	1.210000	1.292500	18.000000	18.000000	20.000000	20.000000	18.000000	18.000000	3.000000
max	3.800000	4.000000	4.100000	100.000000	98.000000	100.000000	100.000000	100.000000	100.000000	4.000000

Figura 3.4 Resultado del análisis estadístico

La evaluación de las relaciones entre las variables independientes y la variable dependiente es fundamental por varias razones. En primer lugar, esta valoración permite identificar qué características guardan una relación más sólida con la variable dependiente, lo que facilita la concentración en las que tienen un mayor impacto en la predicción, lo que a su vez puede mejorar la eficacia del modelo. En segundo lugar, comprender cómo las variables independientes se relacionan con la variable dependiente es esencial para interpretar el modelo y sus predicciones, ya que las características altamente correlacionadas aportan información valiosa para la toma de decisiones del modelo. Llevar a cabo una evaluación de las variables independientes en relación con la variable dependiente es una práctica crucial para la selección y utilización efectiva de las características, mejorar la interpretación del modelo y optimizar su rendimiento.

En consecuencia, se lleva a cabo un análisis de correlación de Kendall para examinar las relaciones entre las variables numéricas y la variable categórica, Diagnóstico. La correlación de Kendall, siendo una medida no paramétrica, resulta apropiada para evaluar la asociación entre variables clasificadas ordinalmente. Su aplicación es pertinente cuando se dispone de una variable dependiente categórica y variables independientes de diversa índole.

En la Figura 3.5, se presenta el mapa de calor de las variables cuantitativas con respecto a la variable Diagnóstico, donde se visualiza los coeficientes de correlación de Kendall entre las variables correspondientes.

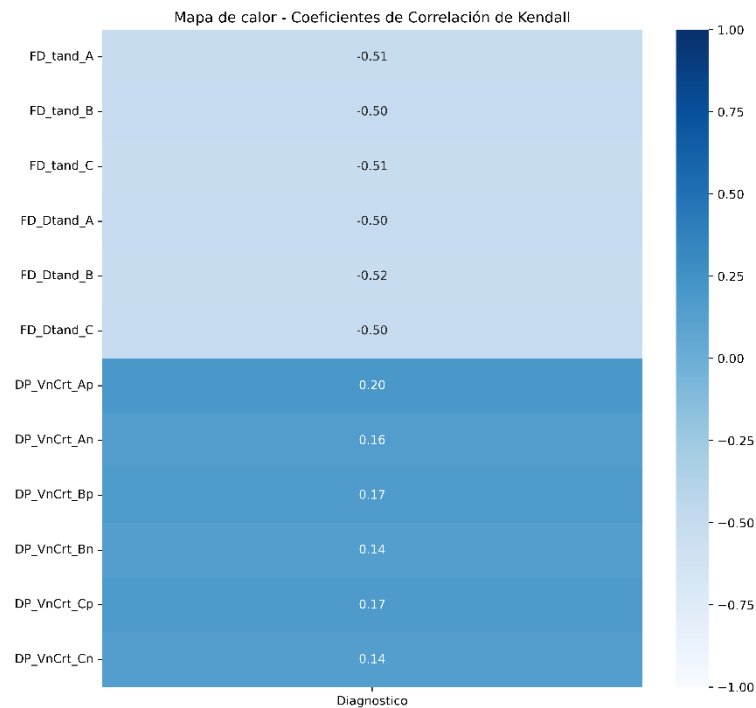


Figura 3.5 Mapa de calor de las variables numéricas

El mapa de calor revela la intensidad de la relación entre variables, representando valores en un rango de -1 a 1, donde los tonos de blanco a azul intenso indican la magnitud de la correlación. Coeficientes cercanos a 1 denotan una correlación positiva, mientras que aquellos cercanos a -1 indican una correlación negativa. En este contexto, las variables asociadas con factores de disipación (FD) presentan una correlación negativa significativa con la variable Diagnóstico, sugiriendo que un aumento en estas variables tiende a disminuir la variable Diagnóstico. En contraste, las variables relacionadas con descargas parciales (DP) presentan una correlación positiva, aunque más moderada, con la variable Diagnóstico, indicando que incrementos en estas variables tienden a aumentar la variable Diagnóstico.

3.5.3 Visualización

En cuanto a la visualización de los datos, los gráficos estadísticos son una herramienta fundamental para representar de manera visual la evolución y relaciones entre diversas variables estadísticas. En la Figura 3.6 se presenta la distribución de frecuencia de los datos relacionados con el diagnóstico de generadores, proporcionando una representación gráfica que facilita la comprensión de la distribución y patrones de estos datos.

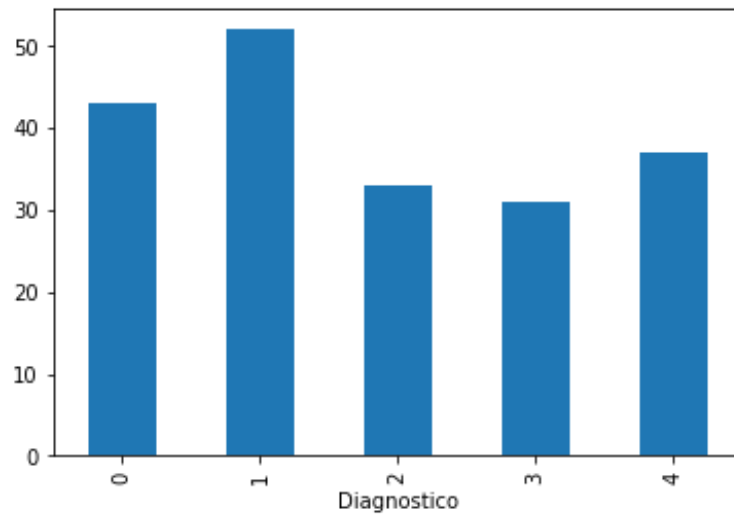


Figura 3.6 Distribución de frecuencia Diagnóstico

La Figura 3.7 presenta un gráfico circular que visualiza de manera efectiva las proporciones y los porcentajes de las distintas clases de falla de la variable Diagnóstico. Este gráfico circular brinda una representación visual clara de la distribución de las clases de falla en la base de datos.

Este análisis visual destaca claramente el desequilibrio que existe en la base de datos, en la que las clases más predominantes son la clase 1 y la clase 0, mientras que las clases minoritarias comprenden la clase 4, la clase 2 y la clase 3. Esto se relaciona con el diagnóstico de fallas en generadores, dado que las clases 0 y 1 son las fallas menos severas y se presentan con mayor frecuencia. Esta observación es esencial para comprender la composición de la base de datos y sus implicaciones en el proceso de modelado y predicción.

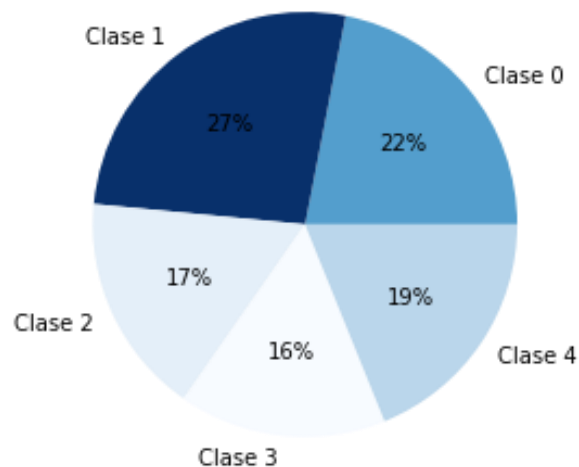


Figura 3.7 Gráfico circular Diagnóstico

3.5.4 Información

Una vez completado el proceso de exploración de datos, se obtiene información relevante acerca de la distribución de datos en cada clase, así como estadísticas descriptivas de las columnas numéricas. Esta información resulta fundamental para llevar a cabo el pre-procesamiento de los datos antes de iniciar su proceso de entrenamiento.

La exploración de datos no solo nos proporciona una comprensión detallada de la naturaleza de cada dato, sino que también nos ofrece claridad sobre las posibles transformaciones que pueden aplicarse. Por ejemplo, en algunos casos es necesario convertir variables categóricas en representaciones vectoriales adecuadas. Esta fase de exploración nos brinda información valiosa para determinar qué tipos de algoritmos y arquitecturas son más apropiados para abordar el problema en cuestión. Además, es esencial entender la relación entre las variables antes de desarrollar un modelo inteligente para lograr resultados más efectivos.

3.6 Pre-procesamiento

El pre-procesamiento de datos se revela como una fase crucial en el proceso, ya que abarca una serie de técnicas relacionadas con el análisis de datos, tales como la preparación, imputación y reducción de datos. Su fundamental radica en la mejora de la calidad de la base de datos, lo cual se traduce en una obtención de conocimiento más efectiva y en una preparación adecuada de los datos antes de introducirlos en los algoritmos de aprendizaje automático.

La fase de preparación de datos involucra diversos aspectos, como la transformación de datos, la normalización, la integración de diferentes fuentes de datos, la eliminación de ruido y la imputación de valores faltantes. Cada una de estas acciones desempeña un papel crucial en la creación de un conjunto de datos limpios y coherente que sea apto para alimentar los algoritmos de manera efectiva.

3.6.1 Limpieza de datos

La limpieza de datos representa una fase de vital importancia en la preparación de los datos crudos antes de su aplicación en el contexto de ML. Los datos sin procesar pueden estar sujetos a

una variedad de errores que potencialmente introducen ruido y afectan la precisión de los modelos resultantes. En este sentido, es necesario llevar a cabo acciones específicas, como la estandarización, normalización, eliminación de datos incompletos o duplicados, así como la supresión de datos irrelevantes o valores faltantes.

A pesar de que la base de datos proporcionada por el INEEL ha sido validada por expertos, es fundamental aplicar un escalamiento de datos, es decir, ajustar las características de un conjunto de datos para que tengan una escala común o similar, lo que permite eliminar cualquier sesgo presente. Este proceso es esencial para facilitar la optimización durante el entrenamiento de modelos de aprendizaje automático, asegurando que las distintas magnitudes de las características no tengan un impacto negativo en el rendimiento del modelo.

3.6.2 Ingeniería de características

La ingeniería de características abarca la extracción y transformación de variables a partir de datos sin procesar, permitiendo así su utilización en procesos de entrenamiento y predicción en el contexto de ML. En el caso de las características categóricas, es importante destacar que no pueden emplearse directamente en la mayoría de los algoritmos de ML, debido a que muchos de estos algoritmos se fundamentan en operaciones matemáticas que requieren de datos numéricos. Además, es crucial realizar una codificación adecuada de las categorías, ya que los algoritmos pueden malinterpretar las relaciones numéricas si no se efectúa esta codificación de forma correcta.

Un ejemplo de esto es la codificación de la variable categórica Enfriamiento, donde asignar valores numéricos (por ejemplo, Aire como 1, Aire forzado como 2, Hidrógeno como 3 e Hidrógeno a una presión de 1.2 kg/cm² como 4) puede llevar al algoritmo a asumir relaciones incorrectas entre las categorías, como que la categoría Aire es más similar con la categoría Aire forzado que con la categoría Hidrógeno, lo cual sería erróneo. Por este motivo, las características categóricas deben someterse a una transformación adecuada para convertirlas en características numéricas. Una de las técnicas más empleadas para llevar a cabo esta conversión es el denominado One-Hot Encoding (codificación one-hot), que transforma cada dato categórico en un vector con n elementos, donde n representa la cantidad de categorías únicas.

En este trabajo, dado que la base de datos incluye dos columnas categóricas se aplica la técnica mencionada anteriormente para transformar dichas características en valores numéricos.

La Figura 3.8 ilustra la implementación de la técnica de One-Hot Encoding para las columnas Enfriamiento y Tipo, generando variables ficticias (dummy) para representar cada una de las categorías de manera adecuada.

Las variables dummy que representan las categorías de la variable Enfriamiento son: Enfriamiento_AIRE, Enfriamiento_HIDRÓGENO, Enfriamiento_HIDRÓGENO hidrógeno a una presión de 1.2 kg/cm2, y Enfriamiento_AIRE FORZADO.

Por otro lado, las variables dummy que representan las categorías de la variable Tipo son: Tipo_375 MW / 17 KV, Tipo_208 MVA / 16 KV, Tipo_218 MVA / 15 KV, Tipo_153 MVA / 16 KV, Tipo_125 MVA / 16.5 KV, Tipo_341 MVA / 13.8 KV, Tipo_49 MVA / 13.8 KV, Tipo_269 MVA / 18 KV, Tipo_339 MVA / 18 KV, Tipo_212 MVA / 18 KV, Tipo_183 MVA / 13.8 KV, Tipo_337 MVA / 18 KV, Tipo_103 MVA / 11.5 KV, Tipo_365 MVA / 21 KV, Tipo_143 MVA / 13.8 KV, Tipo_389 MVA / 20 KV, Tipo_193 MVA / 13.8 KV, Tipo_187 MVA / 16 KV, Tipo_140 MVA / 13.8 KV, Tipo_113 MVA / 13.8 KV, Tipo_177 MVA / 16.5 KV, Tipo_236 MVA / 16.5 KV, Tipo_91.5 MVA / 13.8 KV, Tipo_207 MVA / 18 KV, Tipo_215 MVA / 16.5 KV, Tipo_110 MVA / 13.8 KV, Tipo_126 MVA / 13.8 KV, Tipo_188 MVA / 16.5 KV, Tipo_211 MVA / 16.5 KV, Tipo_40 MVA / 13.8 KV, Tipo_105 MVA / 13.8 KV, Tipo_33 MVA / 13.8 KV, Tipo_45 MVA / 13.8 KV, Tipo_37.5 MVA / 13.8 KV, Tipo_43 MVA / 11.5 KV, Tipo_51.6 MVA / 13.8 KV, Tipo_49 MVA / 11.5 KV, Tipo_50 MVA / 13.8 KV, Tipo_35 MVA / 13.8 KV, Tipo_60 MVA / 13.8 KV, y Tipo_25 MVA / 13.8 KV.

df										
	Enfriamiento_AIRE	Enfriamiento_HIDRÓGENO	Enfriamiento_HIDRÓGENO hidrógeno a una presión de 1.2 kg/cm2	Enfriamiento_AIRE FORZADO	Tipo_375 MW / 17 KV	Tipo_208 MVA / 16 KV	Tipo_218 MVA / 15 KV	Tipo_153 MVA / 16 KV	Tipo_125 MVA / 16.5 KV	Tipo_341 MVA / 13.8 KV
0	1	0	0	0	1	0	0	0	0	0
1	1	0	0	0	1	0	0	0	0	0
2	1	0	0	0	0	1	0	0	0	0
3	1	0	0	0	0	1	0	0	0	0
4	1	0	0	0	0	0	1	0	0	0
...	...	...	...	...	...	...	...	...	...	...
191	0	0	0	1	0	0	0	0	0	0
192	0	0	0	1	0	0	0	0	0	0
193	0	0	0	1	0	0	0	0	0	0
194	0	0	0	1	0	0	0	0	0	0
195	0	0	0	1	0	0	0	0	0	0

196 rows × 58 columns

Figura 3.8 Aplicación de One-Hot Encoding

Una vez que se han codificado las categorías de las variables Enfriamiento y Tipo, se procede a realizar un análisis de correlación entre las características contenidas en la base de datos y la variable Diagnóstico. La Figura 3.9 presenta un mapa de calor que visualiza las correlaciones de Kendall de las características con respecto a la variable Diagnóstico, proporcionando así una representación gráfica de la relación existente.

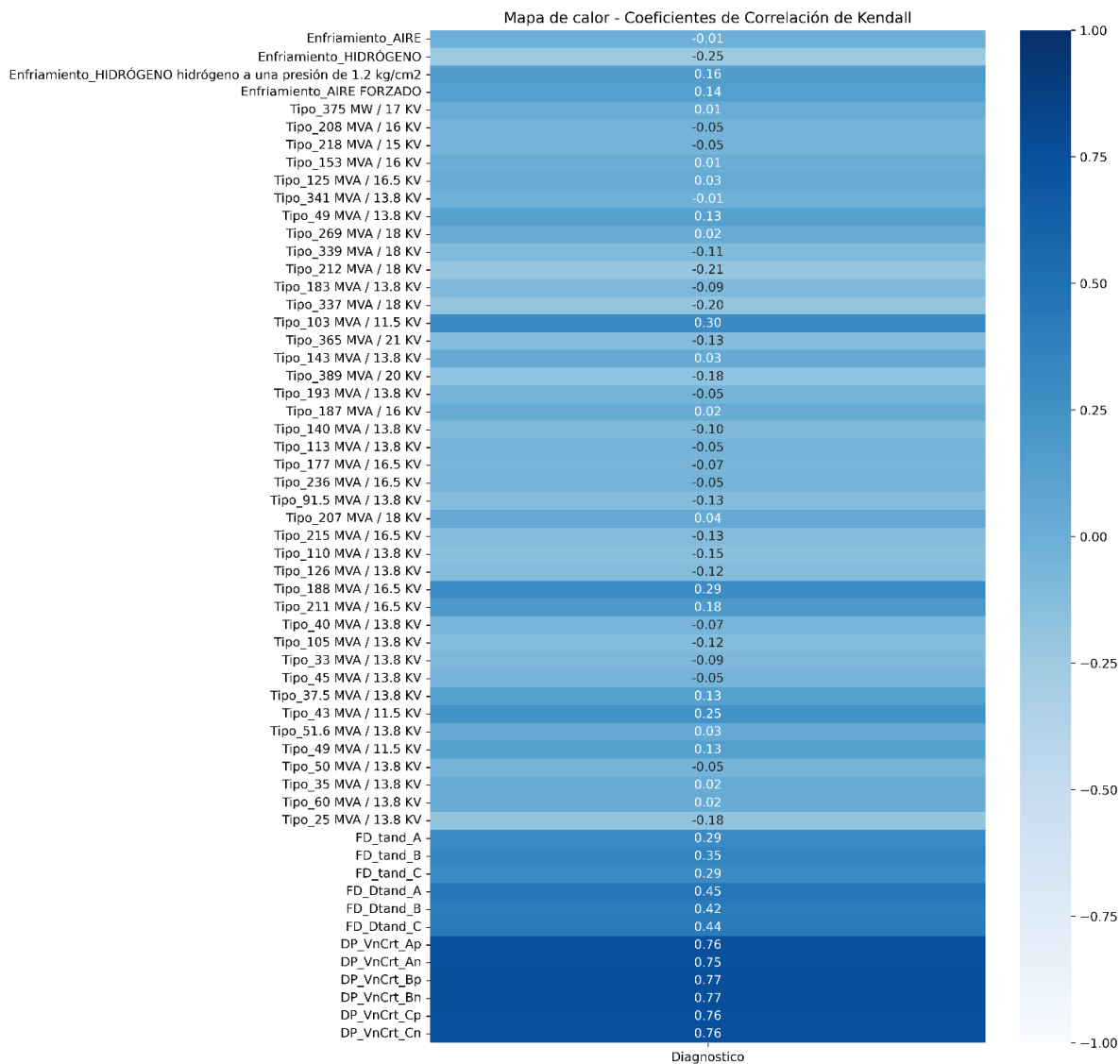


Figura 3.9 Mapa de calor de la base de datos

De la figura anterior, se observa que las características asociadas a las descargas parciales muestran correlaciones que oscilan entre 0.75 y 0.77, presentando tonalidades de azul más intensas en el mapa. Por otro lado, las características relacionadas con los factores de disipación exhiben correlaciones que varían entre 0.29 y 0.45, lo que se refleja en tonalidades azules de intensidad

moderada. La variación en los coeficientes de correlación de Kendall, según se observa en el mapa de calor de las variables numéricas, puede atribuirse a la inclusión de más características en el análisis, ya que se considera la relación conjunta de todas las características y la variable Diagnóstico.

En última instancia, las categorías de la variable Tipo exhiben correlaciones que fluctúan entre -0.18 y 0.29, generando tonalidades de azul de intensidad débil. Por otro lado, las categorías de la variable Enfriamiento revelan correlaciones que varían entre -0.25 y 0.16, predominando en su mayoría tonalidades más claras de azul en el mapa.

3.6.3 Selección de características

Las características en ML es una funcionalidad que tiene como objetivo reducir el número de características mediante el uso de filtros, generando subconjuntos de características más relevantes. Los algoritmos de selección de características implementan métodos que permiten eliminar una o varias características de un conjunto de datos determinado.

3.6.4 Conocimiento

El conocimiento obtenido durante la etapa de pre-procesamiento se convierte en una base de datos tabular en la que los valores se encuentran normalizados y las características categóricas han sido convertidas de manera efectiva en representaciones numéricas. Esta transformación optimiza la capacidad de generalización de los algoritmos de ML. Como último paso, se crea un archivo en formato csv que contiene la base de datos de aprendizaje, lista para ser utilizado como entrada en el proceso de entrenamiento de los algoritmos de clasificación.

3.7 Elección algorítmica

En un proyecto que implica ML, la selección del algoritmo está intensamente vinculada con los objetivos establecidos. Para tomar la decisión acertada respecto al algoritmo que mejor se ajusta a las necesidades del proyecto, resulta esencial evaluar cuidadosamente los datos disponibles, así como su tipo. De esta manera, la información que se incorpora al modelo puede ser procesada de manera eficaz y eficiente.

En el contexto de este trabajo, el objetivo principal consiste en la clasificación de las fallas presentes en generadores de potencia. Esto implica la capacidad de predecir a qué categoría pertenece una muestra específica en función de los datos disponibles.

3.7.1 Probar distintos algoritmos

Los algoritmos de clasificación de aprendizaje supervisado probados para este trabajo se muestran en la Tabla 3.4. Donde se muestra el clasificador con los hiperparámetros con los que se van a entrenar.

Tabla 3.4 Algoritmos de aprendizaje supervisado para clasificar fallas en un generador de potencia

Algoritmo	Clasificador
Support Vector Machines	SVC(probability=True, random_state=42)
Naive Bayes*	GaussianNB()
Linear Discriminant Analysis*	LinearDiscriminantAnalysis()
Quadratic Discriminant Analysis*	QuadraticDiscriminantAnalysis()
Decision Tree	DecisionTreeClassifier(random_state=42, max_depth=5, min_samples_split=2)
Random Forest	RandomForestClassifier(random_state=42, max_depth=5, n_estimators=50, min_samples_split=2)
Gradient Boosting	GradientBoostingClassifier(random_state=42, max_depth=5, n_estimators=50, min_samples_split=2)
XGBoost	XGBClassifier(random_state=42, max_depth=5, n_estimators=50)
K – Nearest Neighbors*	KNeighborsClassifier()
Artificial Neural Network	MLPClassifier(alpha=1, max_iter=1000, random_state=42)

* Hiperparámetros utilizados por defecto en el algoritmo

En la Figura 3.10 se muestra la división de los datos de entrenamiento y prueba, en donde el 80% de los datos son para el entrenamiento y el porcentaje restante para probar los algoritmos.

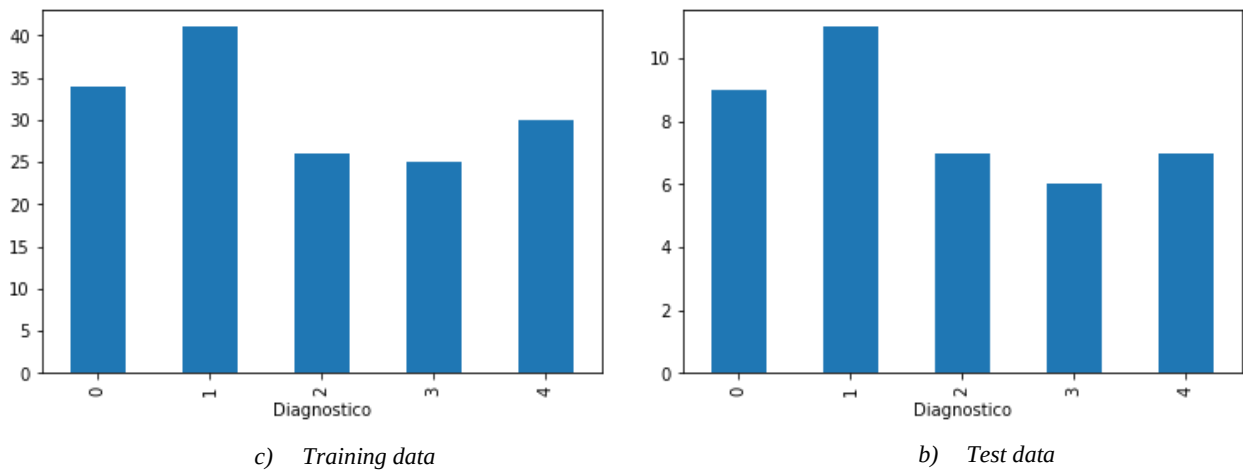


Figura 3.10 Datos de entrenamiento y prueba

3.7.2 Análisis de métricas

En esta sección, se realiza una comparación del desempeño de los algoritmos de clasificación multiclase. Estos algoritmos son evaluados utilizando métricas específicas que consideran el desequilibrio presente en la distribución de las clases. De esta manera, se proporciona una evaluación más justa y equitativa del rendimiento de los modelos.

El análisis de métricas de evaluación en ML es un proceso fundamental para evaluar el rendimiento y la efectividad de un modelo de aprendizaje automático. Cuando se entrena un modelo de ML, es crucial comprender cuán bien está funcionando en datos de prueba o en situaciones del mundo real. Las métricas de evaluación proporcionan una medida cuantitativa de qué tan bien el modelo está haciendo sus predicciones. Se suele expresar en un intervalo de valores comprendido entre 0 y 1, en el cual una puntuación de 1 denota un modelo perfecto que realiza clasificaciones precisas.

3.7.2.1 Rendimiento Support Vector Machines

En la Tabla 3.5 se muestra los resultados del reporte de clasificación del algoritmo SVM. Donde, en la clase 0 el 90% de las predicciones positivas de la falla tipo 0 son correctas e identifica el 100% de todos los casos reales de la falla tipo 0, lo que significa un rendimiento sólido para esta clase. En la clase 1, el 67% de las predicciones positivas de la falla tipo 1 son correctas e identifica el 73% de todos los casos reales de la falla tipo 1. En la clase 2, el modelo no hace ninguna predicción correcta ni logra identificar ningún solo caso real de la falla tipo 2, lo que significa que el rendimiento es extremadamente deficiente para esa clase. En la clase 3, el 100% de las predicciones positivas de la falla tipo 3 son correctas e identifica el 33% de todos los casos reales de la falla tipo 3. Y en la clase 4, el 38% de las predicciones positivas de la falla tipo 4 son correctas e identifica el 86% de todos los casos reales de la falla tipo 4.

El algoritmo SVM presenta una exactitud global del 0.62, indicando que el 62% de todas las predicciones hechas por el modelo son correctas.

Tabla 3.5 Reporte de clasificación del algoritmo SVM

Support Vector Machines					
	Clase 0	Clase 1	Clase 2	Clase 3	Clase 4
Precision	0.90	0.67	0.00	1.00	0.38
Recall	1.00	0.73	0.00	0.33	0.86
F1-score	0.95	0.70	0.00	0.50	0.52
Support	9	11	7	6	7
Accuracy	0.62				

Otras métricas para evaluar la clasificación con desequilibrio de clases se detallan a continuación. El algoritmo SVM presenta una exactitud balanceada con valor de 0.58, lo que indica que el modelo tiene en general un bajo rendimiento. También, muestra un valor de 0.52 para el índice Kappa de Cohen lo que sugiere una concordancia baja entre las predicciones del modelo y las etiquetas reales. Un índice de Jaccard de 0.45 indica una similitud baja entre las predicciones y las etiquetas reales. Por su parte, un valor de 0.56 para el coeficiente de correlación de Matthews sugiere una estimación moderada entre las predicciones y las etiquetas reales.

Además, el algoritmo muestra una especificidad con valor de 1 para la clase 0, lo que significa que el modelo identifica muy bien esa clase. La clase 1 tiene un valor de 0.73, indicando una especificidad moderada para esa clase. En la clase 2 el modelo no identifica correctamente esa clase, por lo que tiene un valor de especificidad de 0. Para la clase 3, el valor 0.33 indica una especificidad baja para esa clase. Y, la clase 4 tiene una alta especificidad de 0.86 para esa clase.

Los resultados indican un rendimiento bajo del modelo en diferentes métricas. La precisión equilibrada, el índice Kappa de Cohen y el coeficiente de precisión de Matthews sugieren un rendimiento bajo, mientras que el índice de Jaccard indica una similitud moderada entre las predicciones y las etiquetas reales. La especificidad varía según la clase, con algunas clases bien identificadas y otras con un rendimiento más limitado.

En situaciones que involucran problemas multiclase, es esencial analizar las posibles confusiones entre las diferentes categorías de fallas. Para abordar esta evaluación, se recurre al empleo de una matriz de confusión. En la Figura 3.11 se muestra las matrices de confusión, tanto normalizada como no normalizada, del algoritmo SVM. Se puede observar en la matriz no normalizada que un total de 15 elementos están mal clasificados, de las cuales, una falla tipo 3 es clasificada como falla tipo 0; una falla tipo 2, dos fallas tipo 3 y una falla tipo 4 son clasificadas como fallas tipo 1; y, tres fallas tipo 1, seis fallas tipo 2 y una falla tipo 3 son clasificadas como falla tipo 4.

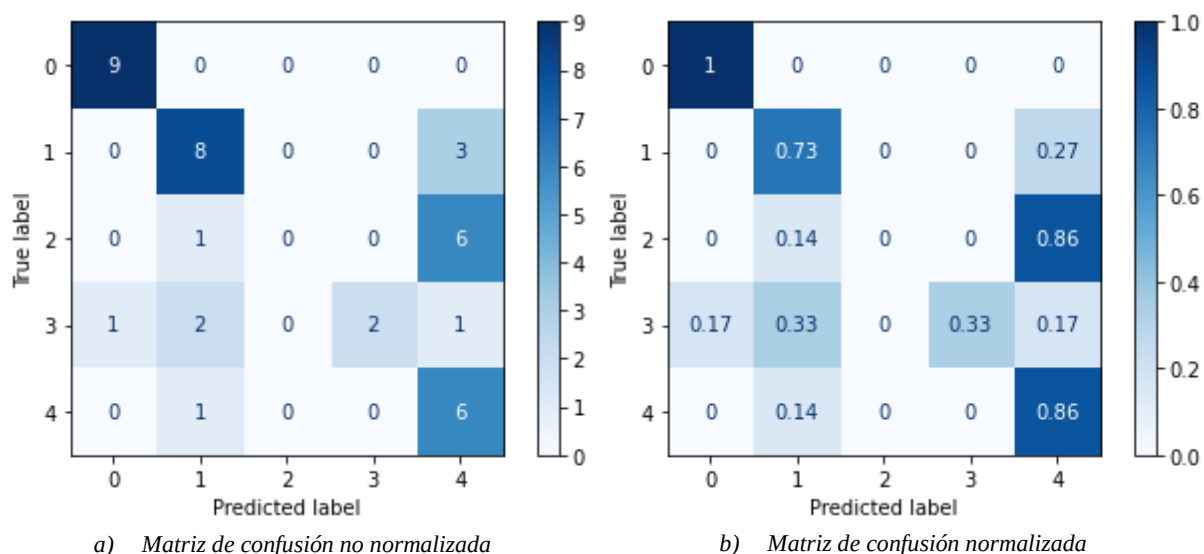


Figura 3.11 Matrices de confusión SVM

Con respecto a las métricas gráficas, en la Figura 3.12 se presenta la curva ROC del algoritmo SVM. Donde se muestra que el AUC más bajo es para la clase 2 con un valor de 0.50, el AUC más alto es para la clase 0 con un valor de 0.99, y un macro promedio de 74%, lo que indica un rendimiento moderado del modelo en la tarea de clasificación. Se observa que las líneas trazadas no están exactamente sobre la esquina superior izquierda, lo que indica que este algoritmo no presenta un excelente desempeño en esta métrica.

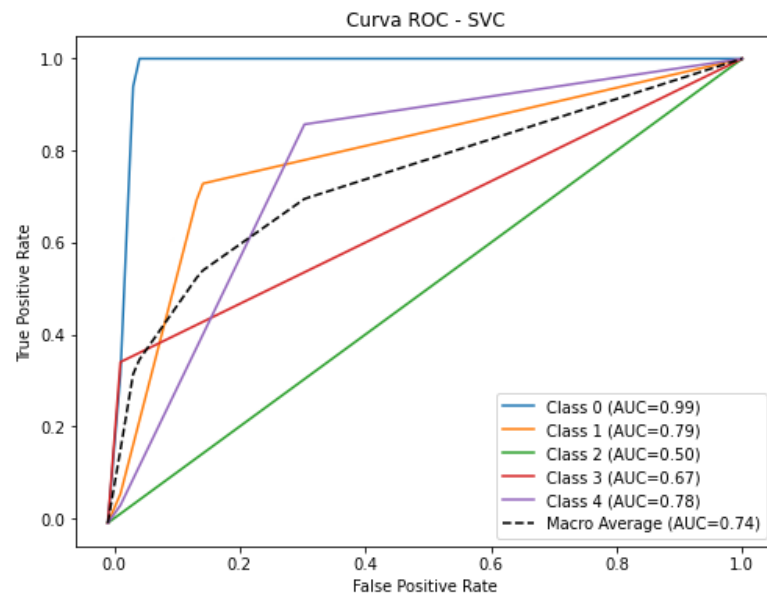


Figura 3.12 Curva ROC y AUC por SVM

Finalmente, en la Figura 3.13 se muestra la curva PR del algoritmo SVM. Donde, el PR más bajo pertenece a la clase 2 con un valor de 0.17, el PR más alto es para la clase 0 con un valor de 0.90, y un micro promedio de 0.48, lo que sugiere que el modelo es capaz de identificar correctamente el 48% de los ejemplos que etiquetó como positivos, es decir, no hay una buena precisión en las predicciones positivas.

En general, el algoritmo SVM presenta un bajo rendimiento en la clasificación de fallas en generadores de potencia.

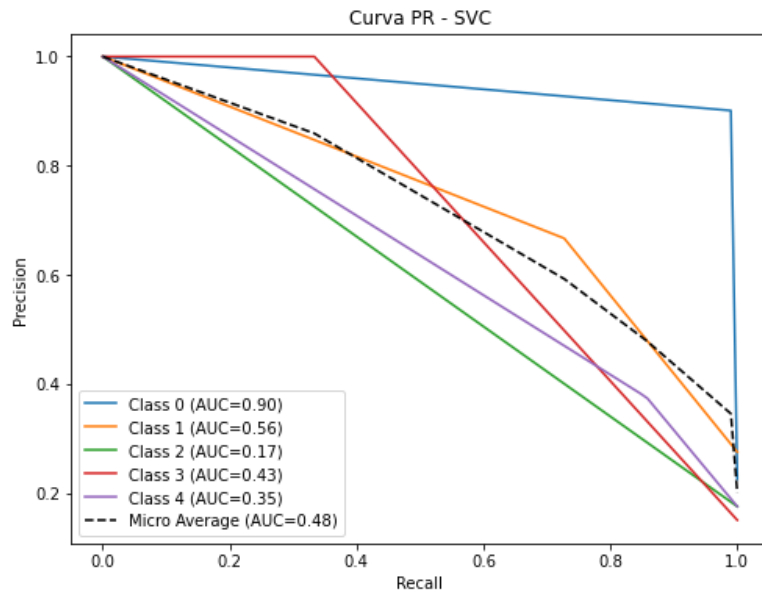


Figura 3.13 Curva PR por SVM

3.7.2.2 Rendimiento Naive Bayes

En la Tabla 3.6 se muestra los resultados del reporte de clasificación del algoritmo Naive Bayes. Donde, en la clase 0 todas de las predicciones positivas de la falla tipo 0 son correctas, sin embargo, identifica el 67% de todos los casos reales de la falla tipo 0. Para la clase 1, el 75% de las predicciones positivas de la falla tipo 1 son correctas e identifica el 55% de todos los casos reales de la falla tipo 1. En la clase 2, el 78% de las predicciones positivas de la falla tipo 2 son correctas, pero identifica todos los casos reales de la falla tipo 2. En la clase 3, el 50% de las predicciones positivas de la falla tipo 3 son correctas e identifica el 83% de todos los casos reales de la falla tipo 3. Por último, la clase 4 presenta una clasificación perfecta para la falla tipo 4. En general, el algoritmo Naive Bayes presenta una exactitud del 0.78, indicando que el 78% de todas las predicciones hechas por el modelo son correctas.

Otras métricas para evaluar la clasificación con desequilibrio de clases se detallan a continuación. El algoritmo Naive Bayes presenta una exactitud balanceada de 0.78, lo que indica que el modelo tiene en general un rendimiento regular. También, muestra un valor de 0.72 para el índice Kappa de Cohen lo que sugiere una concordancia sustancial entre las predicciones del modelo y las etiquetas reales. Un índice de Jaccard de 0.63 indica una similitud considerable entre las predicciones y las etiquetas reales.

Por su parte, un valor de 0.73 para el coeficiente de correlación de Matthews sugiere una evaluación sustancial entre las predicciones y las etiquetas reales. Además, el algoritmo muestra una especificidad con valor de 0.66 para la clase 0, lo que significa una especificidad moderada. La clase 1 tiene un valor de 0.54, indicando una especificidad baja para esa clase. En la clase 2 tiene un valor de 1, indicando una alta especificidad para esa clase. Para la clase 3, el valor 0.84 indica una especificidad alta para esa clase. Y, la clase 4 tiene una alta especificidad de 1 para esa clase.

Los resultados en general son relativamente positivos y sugieren que el modelo tiene un buen rendimiento en la tarea de clasificación multiclase, con una precisión balanceada sólida, un índice Kappa de Cohen sustancial y otras métricas que indican una buena concordancia y similitud entre las predicciones y las etiquetas reales. Sin embargo, en este trabajo se considera un nivel de significancia mínimo del 10% en la evaluación de resultados.

Tabla 3.6 Reporte de clasificación del algoritmo Naive Bayes

	Naive Bayes				
	0	1	2	3	4
<i>Precision</i>	1.00	0.75	0.78	0.50	1.00
<i>Recall</i>	0.67	0.55	1.00	0.83	1.00
<i>F1-score</i>	0.80	0.63	0.88	0.62	1.00
<i>Support</i>	9	11	7	6	7
<i>Accuracy</i>	0.78				

Por otro lado, en la Figura 3.14 se muestra las matrices de confusión, tanto normalizada como no normalizada, del algoritmo Naive Bayes. Se puede observar en la matriz no normalizada que un total de 9 elementos están mal clasificados, de las cuales, una falla tipo 0 y una falla tipo 3 son clasificadas como falla tipo 1; dos fallas tipo 1 son clasificadas como fallas tipo 2; y, dos fallas tipo 0 y tres fallas tipo 1 son clasificadas como falla tipo 3.

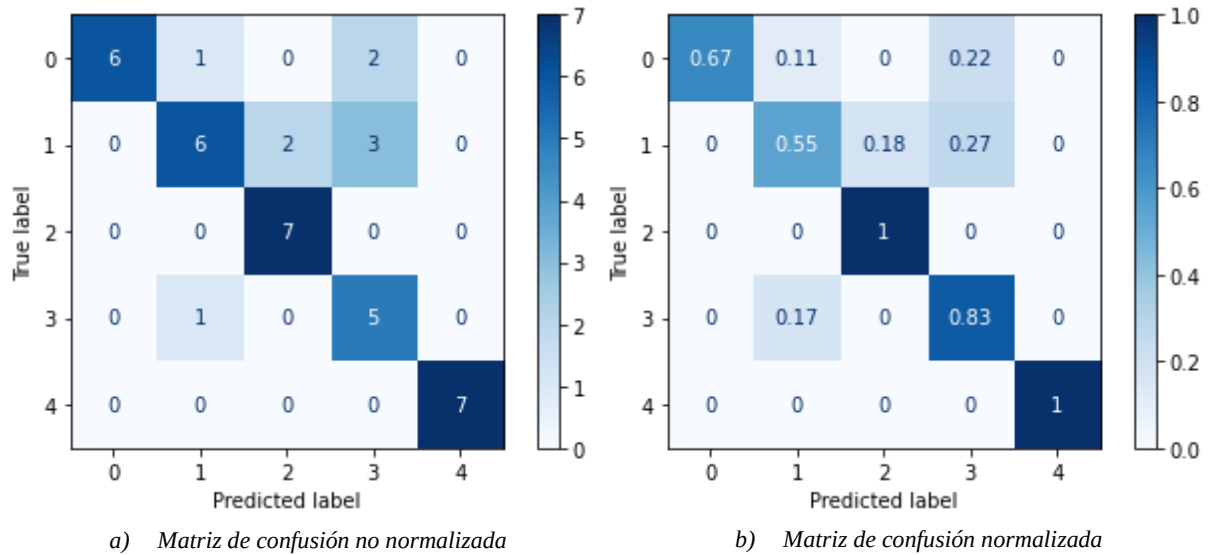


Figura 3.14 Matrices de confusión Naive Bayes

En la Figura 3.15 se muestra la curva ROC del clasificador Naive Bayes. Donde se puede observar que el AUC más bajo es para la clase 1 con un valor de 0.74 y el AUC más alto es para la clase 4 con un valor de 1. Un valor del 88% en el macro promedio sugiere que el modelo es capaz de alcanzar tasas de verdaderos positivos relativamente altas mientras mantiene tasas de falsos positivos razonablemente bajas.

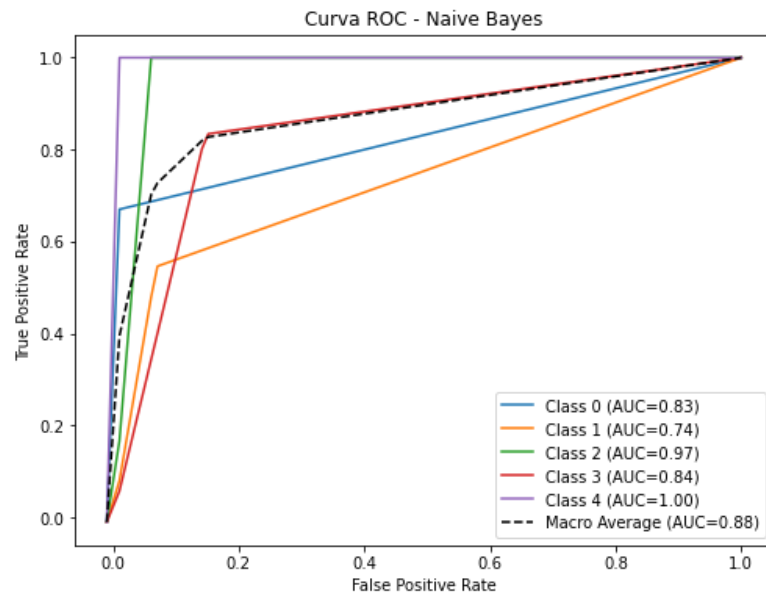


Figura 3.15 Curva ROC y AUC por Naive Bayes

Por otro lado, en la Figura 3.16 se muestra la curva PR del algoritmo Naive Bayes. Donde, el PR más bajo pertenece a la clase 3 con un valor de 0.44 y el PR más alto es para la clase 4 con

un valor de 1. Un micro promedio de 70% en la curva PR indica un rendimiento moderado en la clasificación de fallas en generadores de potencia.

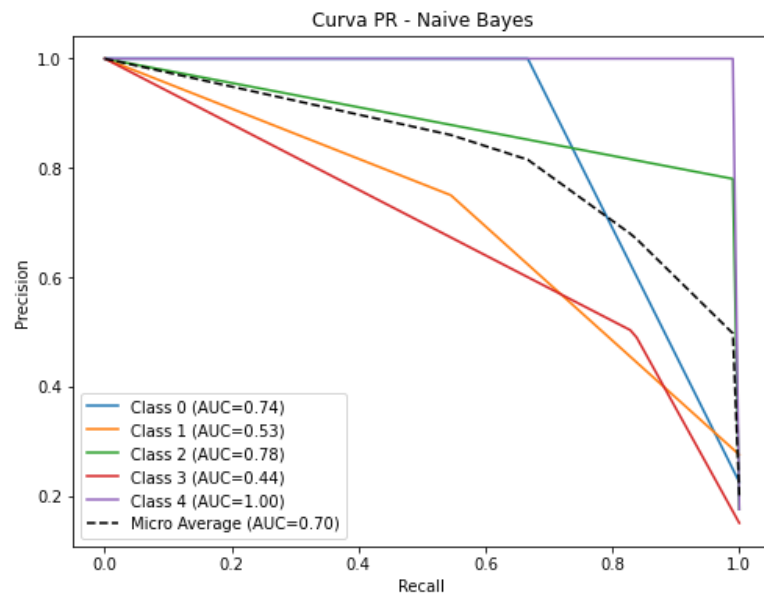


Figura 3.16 Curva PR por Naive Bayes

3.7.2.3 Rendimiento Linear Discriminant Analysis

Con respecto al algoritmo Linear Discriminant Analysis, en la Tabla 3.7 se muestra los resultados del reporte de clasificación. Donde, en la clase 0 todas las predicciones positivas de la falla tipo 0 son correctas, pero identifica el 89% de todos los casos reales de la falla tipo 0. Para la clase 1, el 73% de las predicciones positivas de la falla tipo 1 son correctas, al igual que los casos reales de la misma falla. En la clase 2, el 71% de las predicciones positivas de la falla tipo 2 son correctas e identifica el 71% de todos los casos reales de la falla tipo 2. En la clase 3, el 83% de las predicciones positivas de la falla tipo 3 son correctas, al igual que los casos reales de la misma falla. Para la clase 4, el 88% de las predicciones positivas de la falla tipo 4 son correctas e identifica todos los casos reales de la falla tipo 4.

En general, el algoritmo LDA presenta una exactitud del 0.82, indicando que el 82% de todas las predicciones hechas por el modelo son correctas, por lo que se considera un buen rendimiento.

Tabla 3.7 Reporte de clasificación del algoritmo LDA

Linear Discriminant Analysis					
	<i>0</i>	<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>
<i>Precision</i>	1.00	0.73	0.71	0.83	0.88
<i>Recall</i>	0.89	0.73	0.71	0.83	1.00
<i>F1-score</i>	0.94	0.73	0.71	0.83	0.93
<i>Support</i>	9	11	7	6	7
<i>Accuracy</i>	0.82				

Otras métricas para evaluar la clasificación con desequilibrio de clases se detallan a continuación. El algoritmo LDA presenta una exactitud balanceada con valor de 0.83, lo que indica que el modelo tiene en general un buen rendimiento. También, muestra un valor de 0.78 para el índice Kappa de Cohen lo que sugiere una concordancia sustancial entre las predicciones del modelo y las etiquetas reales. Un índice de Jaccard de 0.70 indica una similitud considerable entre las predicciones y las etiquetas reales. Por su parte, un valor de 0.78 para el coeficiente de correlación de Matthews sugiere una estimación sustancial entre las predicciones y las etiquetas reales.

Además, el algoritmo muestra una especificidad de 0.88 en la clase 0, lo que indica una alta especificidad. La clase 1 tiene un valor de 0.73, indicando una especificidad moderada para esa clase. En la clase 2 un valor de 0.71, indicando una especificidad moderada para esa clase. Para la clase 3, el valor 0.84 indica una especificidad alta para esa clase. Y, la clase 4 tiene una alta especificidad de 1, lo que significa que el modelo identifica muy bien esa clase.

Los resultados son muy positivos y sugieren que el modelo tiene un buen rendimiento en la tarea de clasificación multiclase, con una alta precisión balanceada, un índice Kappa de Cohen sustancial y otras métricas que indican una buena concordancia, similitud y calidad en las predicciones en todas las clases.

En la Figura 3.17 se muestra las matrices de confusión, tanto normalizada como no normalizada, del algoritmo LDA. Se puede observar en la matriz no normalizada que un total de 7 elementos están mal clasificados, de las cuales, una falla tipo 0 y dos fallas tipo 2 son clasificadas como falla 1; dos fallas tipo 1 son clasificadas como falla 2; otra falla tipo 1 es clasificada como falla 3; y una falla tipo 3 es clasificada como falla 4.

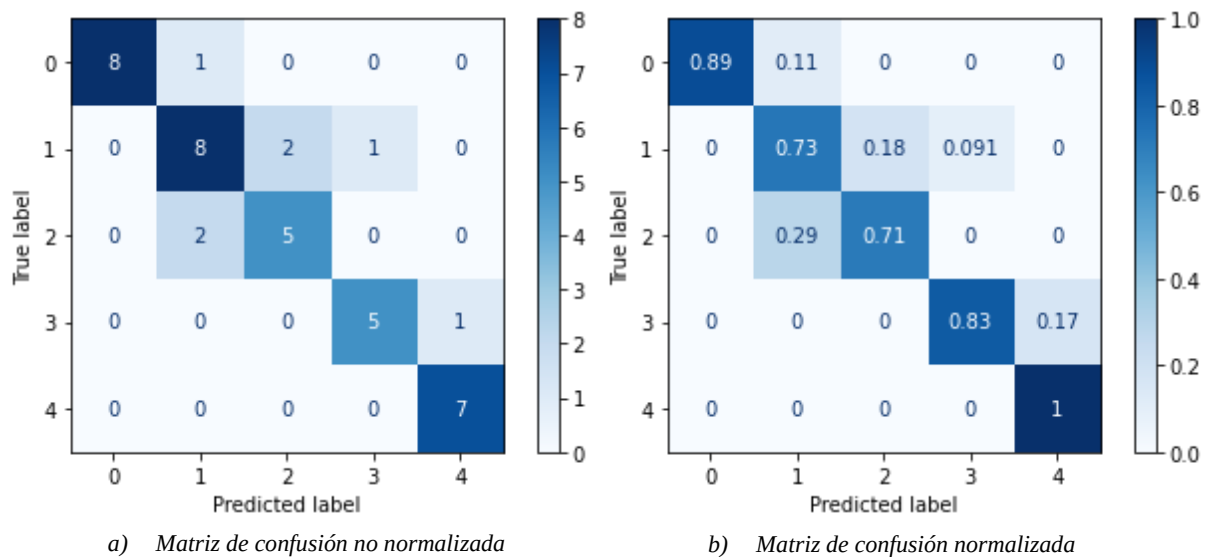


Figura 3.17 Matrices de confusión LDA

Por otro lado, en la Figura 3.18 se muestra la curva ROC del clasificador LDA. Donde se puede observar que el AUC más bajo es para la clase 1 con un valor de 0.81 y el AUC más alto es para la 4 con un valor de 0.99, y un macro promedio de 89%. Gráficamente, las líneas trazadas están cerca de la esquina superior izquierda, lo que indica que el algoritmo presenta un buen rendimiento.

Y en la Figura 3.19 se muestra la curva PR del algoritmo LDA. Donde, el PR más bajo pertenece a la clase 2 con un valor de 0.56 y el PR más alto es para la clase 0 con un valor de 0.91. En la curva PR, un micro promedio de 73% sugiere que el modelo tiene una precisión razonable en las predicciones positivas.

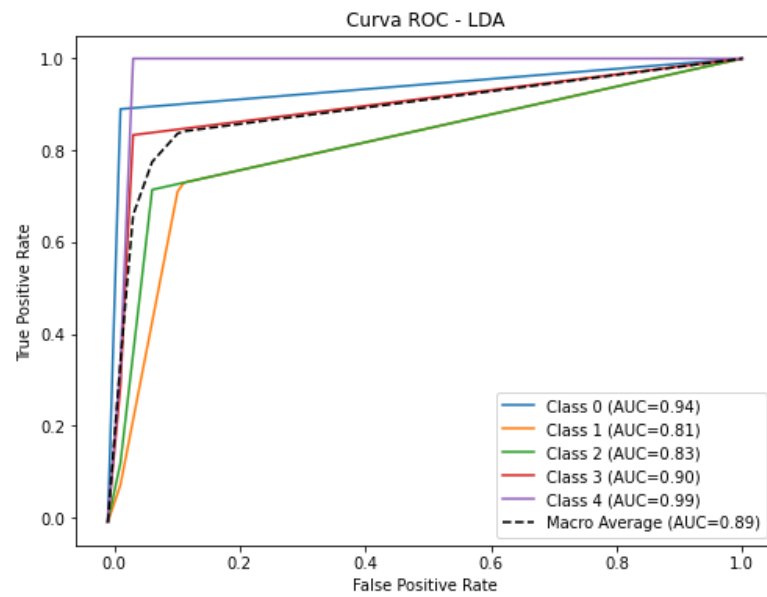


Figura 3.18 Curva ROC y AUC por LDA

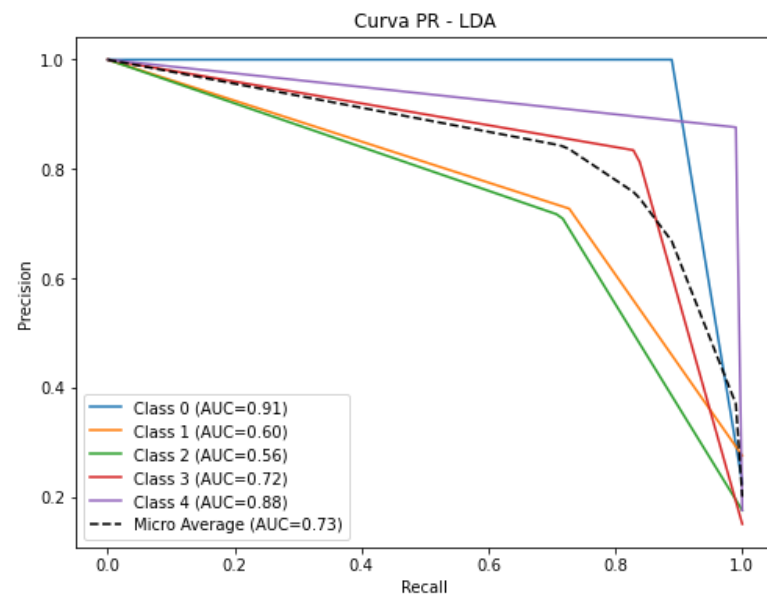


Figura 3.19 Curva PR por LDA

3.7.2.4 Rendimiento Quadratic Discriminant Analysis

Por su parte, el reporte de clasificación del algoritmo Quadratic Discriminant Analysis se muestra en la Tabla 3.8. Donde, en la clase 0 el 89% de las predicciones positivas de la falla tipo 0 son correctas e identifica el 89% de los casos reales de dicha falla. En la clase 1, el 64% de las predicciones positivas de la falla tipo 1 son correctas, al igual que los casos reales de la misma falla. En la clase 2, el 71% de las predicciones positivas de la falla tipo 2 son correctas, al igual que los casos reales de la misma falla. En la clase 3, el 67% de las predicciones positivas de la falla tipo 3 son correctas e identifica el 67% de los casos reales de dicha falla. Por último, la clase 4 presenta una clasificación perfecta para la falla tipo 4.

El algoritmo QDA presenta una exactitud del 0.78, indicando que el 78% de todas las predicciones hechas por el modelo son correctas. El rendimiento de este algoritmo no es tan bueno debido a que es recomendable usarlo para una base de datos grande.

Tabla 3.8 Reporte de clasificación del algoritmo QDA

Quadratic Discriminant Analysis					
	0	1	2	3	4
<i>Precision</i>	0.89	0.64	0.71	0.67	1.00
<i>Recall</i>	0.89	0.64	0.71	0.67	1.00
<i>F1-score</i>	0.89	0.64	0.71	0.67	1.00
<i>Support</i>	9	11	7	6	7
<i>Accuracy</i>	0.78				

Otras métricas para evaluar la clasificación con desequilibrio de clases se detallan a continuación. El algoritmo QDA presenta una exactitud balanceada con valor de 0.78, lo que indica que el modelo tiene en general un rendimiento regular. También, muestra un valor de 0.72 para el índice Kappa de Cohen lo que sugiere una concordancia sustancial entre las predicciones del modelo y las etiquetas reales. Un índice de Jaccard de 0.63 indica una similitud baja entre las predicciones y las etiquetas reales. Por su parte, un valor de 0.72 para el coeficiente de correlación de Matthews sugiere una estimación sustancial entre las predicciones y las etiquetas reales.

Además, el algoritmo muestra una especificidad de 0.88 en la clase 0, lo que indica una alta especificidad. La clase 1 tiene un valor de 0.64, indicando una especificidad baja para esa clase. En la clase 2 un valor de 0.71, indicando una especificidad moderada para esa clase. Para la clase 3, el valor 0.66 indica una especificidad baja para esa clase. Y, la clase 4 tiene una alta especificidad de 1, lo que significa que el modelo identifica muy bien esa clase.

Los resultados sugieren que el modelo tiene un rendimiento moderado en la tarea de clasificación multiclase, con una alta precisión balanceada, un índice Kappa de Cohen sustancial y otras métricas que indican una moderada concordancia, similitud y calidad en las predicciones en todas las clases.

Por otro lado, en la Figura 3.20 se muestra las matrices de confusión, tanto normalizada como no normalizada, del algoritmo QDA. Se puede observar en la matriz no normalizada que un total de 9 elementos están mal clasificados, de las cuales, una falla tipo 3 es clasificada como falla tipo 0; una falla tipo 0, dos fallas tipo 2 y una falla tipo 3 son clasificadas como falla tipo 1; dos fallas tipo 1 son clasificadas como falla tipo 2; y dos fallas más tipo 1 clasificadas como falla tipo 3. En general, este algoritmo no tiene buen rendimiento.

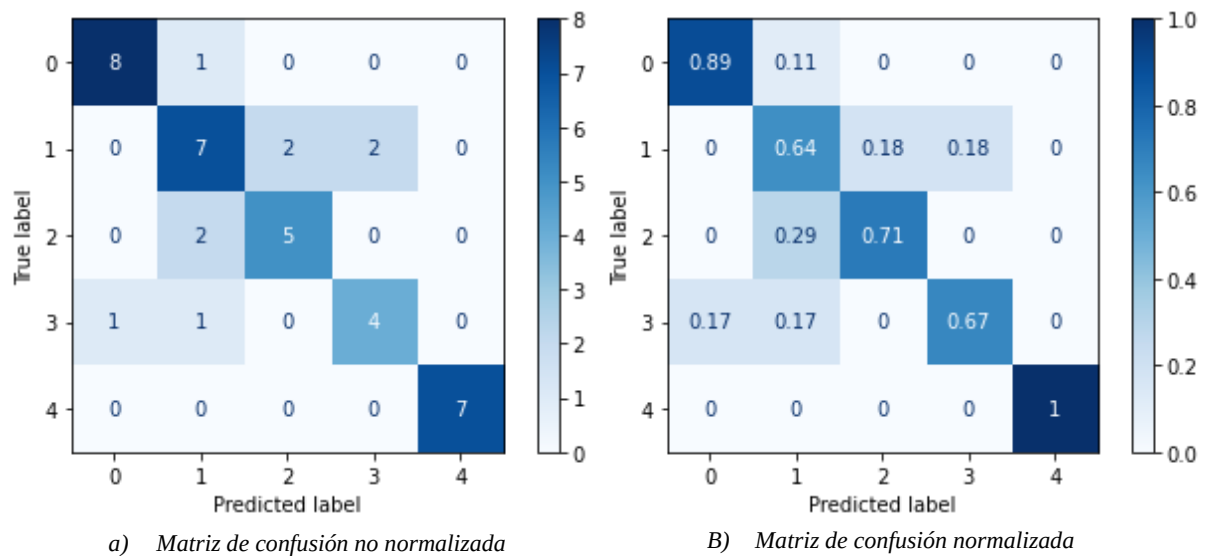


Figura 3.20 Matrices de confusión QDA

En la Figura 3.21 se muestra la curva ROC del clasificador QDA. Donde se puede observar que el AUC más bajo es para la clase 1 con un valor de 0.75 y el AUC más alto es para la clase 4 con un valor de 1. Un macro promedio del 86% indica un rendimiento sólido del modelo en términos de la capacidad de separar las clases.

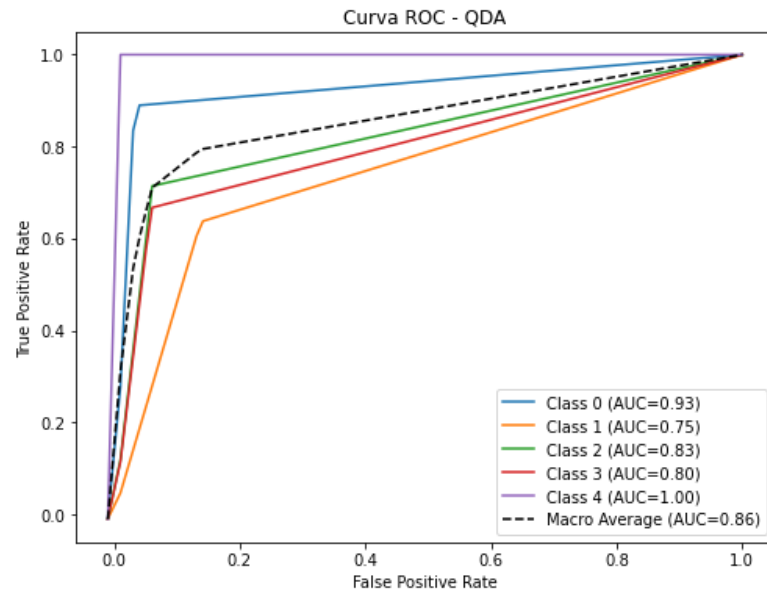


Figura 3.21 Curva ROC y AUC por QDA

Finalmente, en la Figura 3.22 se muestra la curva PR del algoritmo QDA. Donde, el PR más bajo pertenece a la clase 3 con un valor de 0.49 y el PR más alto es para la clase 4 con un valor de 1. En la curva PR, un micro promedio de 67% sugiere que el modelo tiene una precisión moderada en las predicciones positivas.

En general, el algoritmo QDA muestra un rendimiento regular en la clasificación de fallas en generadores de potencia.

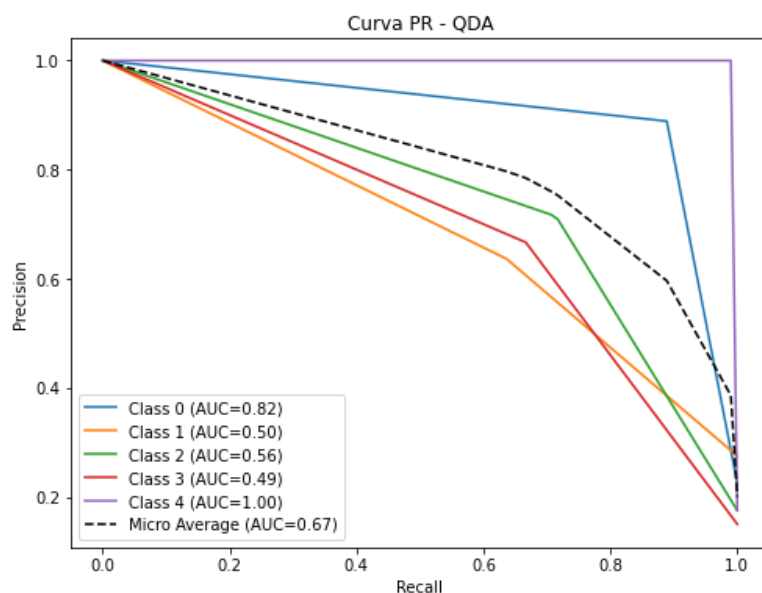


Figura 3.22 Curva PR por QDA

3.7.2.5 Rendimiento Decision Tree

En la Tabla 3.9 se muestra el reporte de clasificación del algoritmo Decision Tree. Donde, la clase 0 presenta una clasificación perfecta para la falla tipo 0. En la clase 1, todas las predicciones positivas de la falla tipo 1 son correctas, pero identifica el 55% de todos los casos reales de la falla tipo 1. En la clase 2, el 64% de las predicciones positivas de la falla tipo 2 son correctas, pero identifica el 100% de todos los casos reales de esa misma falla. En la clase 3, todas las predicciones positivas de la falla tipo 3 son correctas, pero identifica el 67% de todos los casos reales para esa clase. Por último, en la clase 4 el 60% de las predicciones positivas de la falla tipo 4 son correctas e identifica el 86% de todos los casos reales de la falla tipo 4. En general, el algoritmo Decision Tree presenta una exactitud del 0.80, indicando que el 80% de todas las predicciones hechas por el modelo son correctas.

Otras métricas para evaluar la clasificación con desequilibrio de clases se detallan a continuación. El algoritmo Decision Tree presenta una exactitud balanceada de 0.81, lo que indica que el modelo tiene en general un buen rendimiento. También, muestra un valor de 0.75 para el índice Kappa de Cohen lo que sugiere una concordancia sustancial entre las predicciones del modelo y las etiquetas reales. Un índice de Jaccard de 0.66 indica una similitud considerable entre las predicciones y las etiquetas reales.

Por su parte, un valor de 0.77 para el coeficiente de correlación de Matthews sugiere una estimación sustancial entre las predicciones y las etiquetas reales. Además, el algoritmo muestra una especificidad de 1 en la clase 0, lo que indica que es capaz de identificar correctamente las instancias de la clase negativa. La clase 1 tiene un valor de 0.54, indicando una especificidad baja para esa clase. En la clase 2 un valor de 1, indicando una especificidad alta para esa clase. Para la clase 3, el valor 0.66 indica una especificidad baja para esa clase. Y, la clase 4 tiene una alta especificidad de 0.86.

Los resultados son positivos y sugieren que el modelo tiene un buen rendimiento en la tarea de clasificación multiclase, con una alta precisión balanceada, un índice Kappa de Cohen sustancial y otras métricas que indican una buena concordancia, similitud y calidad en las predicciones en todas las clases.

Tabla 3.9 Reporte de clasificación del algoritmo Decision Tree

Decision Tree					
	0	1	2	3	4
<i>Precision</i>	1.00	1.00	0.64	1.00	0.60
<i>Recall</i>	1.00	0.55	1.00	0.67	0.86
<i>F1-score</i>	1.00	0.71	0.78	0.80	0.71
<i>Support</i>	9	11	7	6	7
<i>Accuracy</i>	0.80				

Con respecto a la métrica matriz de confusión, en la Figura 3.23 se muestra las matrices de confusión, tanto normalizada como no normalizada, de dicho algoritmo. Se puede observar en la matriz no normalizada que un total de 8 elementos están mal clasificados, de los cuales, tres fallas de tipo 1 es clasificado como falla de tipo 2, dos fallas de tipo 1 clasificado como falla de tipo 4, dos fallas de tipo 3 están clasificadas como falla de tipo 4 y una falla de tipo 4 clasificada como una falla de tipo 2.

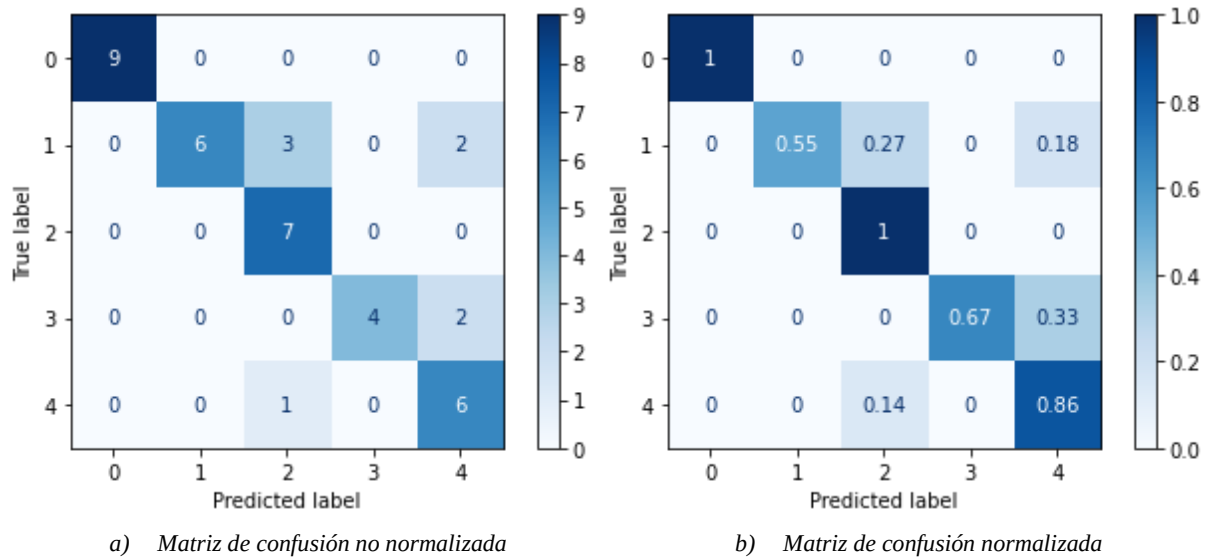


Figura 3.23 Matrices de confusión Decision Tree

Por otra parte, en la Figura 3.24 se muestra la curva ROC del algoritmo Decision Tree. Donde, el AUC más bajo es para la clase 1 con un valor de 0.77 y el AUC más alto es para la clase 0 con un valor de 1. Un macro promedio de 88% indica un rendimiento sólido del modelo en términos de la capacidad de separar las clases, pero se puede observar que las líneas trazadas no están exactamente sobre la esquina superior izquierda, lo que indica que este algoritmo no presenta un excelente desempeño.

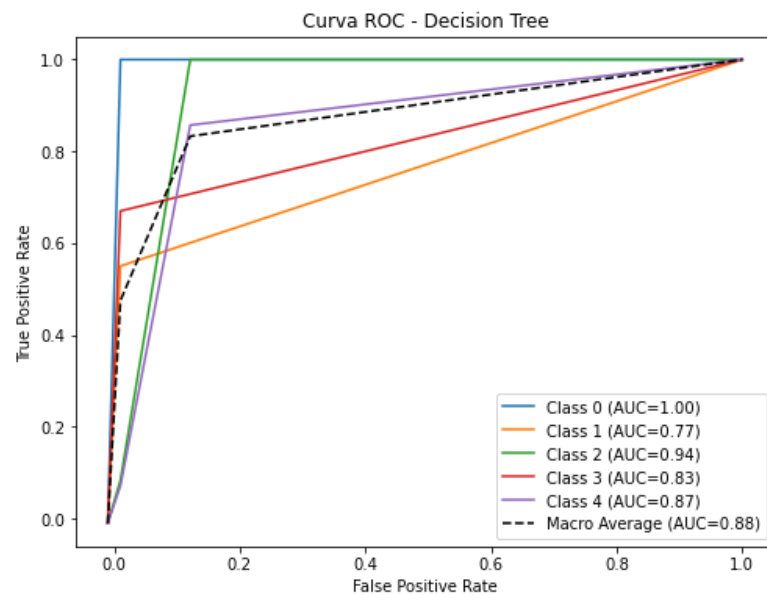


Figura 3.24 Curva ROC y AUC por Decision Tree

En la Figura 3.25 se muestra la curva PR del algoritmo Decision Tree. Donde, el PR más bajo es para la clase 4 con un valor de 0.54 y el PR más alto es para la clase 0 con un valor de 1. Un micro promedio de 71% indica un rendimiento moderado del modelo en la precisión y en la identificación de ejemplos positivos. Además, se observa que las líneas trazadas no están cerca de la esquina superior derecha, lo que indica que este algoritmo no presenta buen desempeño.

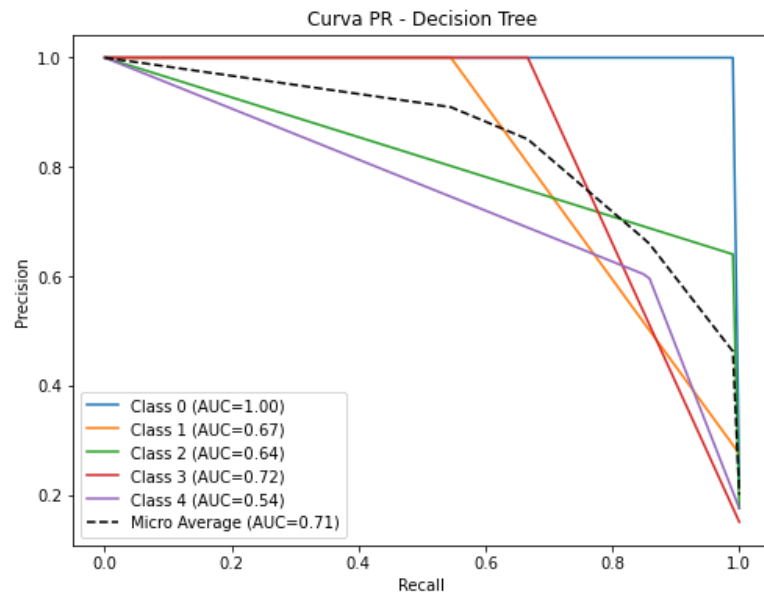


Figura 3.25 Curva PR por Decision Tree

3.7.2.6 Rendimiento Random Forest

En la Tabla 3.10 se muestra su reporte de clasificación del algoritmo Random Forest. Donde, en la clase 0 el 90% de las predicciones positivas para la falla 0 son correctas e identifica todos los casos reales de la misma. Para la clase 1, el 80% de las predicciones positivas para la falla 1 son correctas e identifica el 73% de todos los casos reales de la falla tipo 1. Para la clase 2, el 71% de las predicciones positivas para la falla tipo 2 son correctas e identifica el 71% de todos los casos reales de la falla tipo 2, lo que significa que el modelo detecta la mayoría de los casos reales de esta clase. Para la clase 3 todas las predicciones positivas para la falla tipo 3 son correctas, sin embargo, identifica el 83% de los casos de la misma. Por último, para la clase 4 el 88% de las predicciones positivas de la falla 4 son correctas e identifica todos los casos de la falla tipo 4.

En general, el algoritmo Random Forest presenta una exactitud del 0.85, indicando que el 85% de todas las predicciones hechas por el modelo son correctas. Indicando un buen rendimiento.

Tabla 3.10 Reporte de clasificación del algoritmo Random Forest

	Random Forest				
	0	1	2	3	4
<i>Precision</i>	0.90	0.80	0.71	1.00	0.88
<i>Recall</i>	1.00	0.73	0.71	0.83	1.00
<i>F1-score</i>	0.95	0.76	0.71	0.91	0.93
<i>Support</i>	9	11	7	6	7
<i>Accuracy</i>	0.85				

Otras métricas para evaluar la clasificación con desequilibrio de clases se detallan a continuación. El algoritmo Random Forest presenta una exactitud balanceada de 0.85, lo que indica que el modelo tiene un buen rendimiento. También, muestra un valor de 0.81 para el índice Kappa de Cohen lo que sugiere una concordancia sustancial entre las predicciones del modelo y las etiquetas reales. Un índice de Jaccard de 0.74 indica una similitud considerable entre las predicciones y las etiquetas reales. Por su parte, un valor de 0.81 para el coeficiente de correlación de Matthews sugiere una estimación sustancial entre las predicciones y las etiquetas reales.

Además, el algoritmo muestra una especificidad de 1 en la clase 0, lo que indica que es capaz de identificar correctamente las instancias de la clase negativa. La clase 1 tiene un valor de 0.72, indicando una especificidad moderada para esa clase. En la clase 2 un valor de 0.71, indicando una especificidad moderada para esa clase. Para la clase 3, el valor 0.84 indica una especificidad moderada para esa clase. Y, la clase 4 tiene una alta especificidad de 1.

Los resultados son positivos y sugieren que el modelo tiene un buen rendimiento en la tarea de clasificación multiclase, con una alta precisión balanceada, un índice Kappa de Cohen sustancial y otras métricas que indican una buena concordancia, similitud y calidad en las predicciones en todas las clases.

Por otro lado, en la Figura 3.26 se muestra las matrices de confusión, tanto normalizada como no normalizada, del clasificador del algoritmo Random Forest.

Se puede observar en la matriz no normalizada que 6 elementos se clasificaron mal, de los cuales, una falla de tipo 3 es clasificada como falla tipo 0, dos fallas de tipo 2 son clasificadas como fallas de tipo 1, dos fallas de tipo 1 son clasificadas como fallas de tipo 2, y una falla de tipo 1 es clasificada como falla de tipo 4.

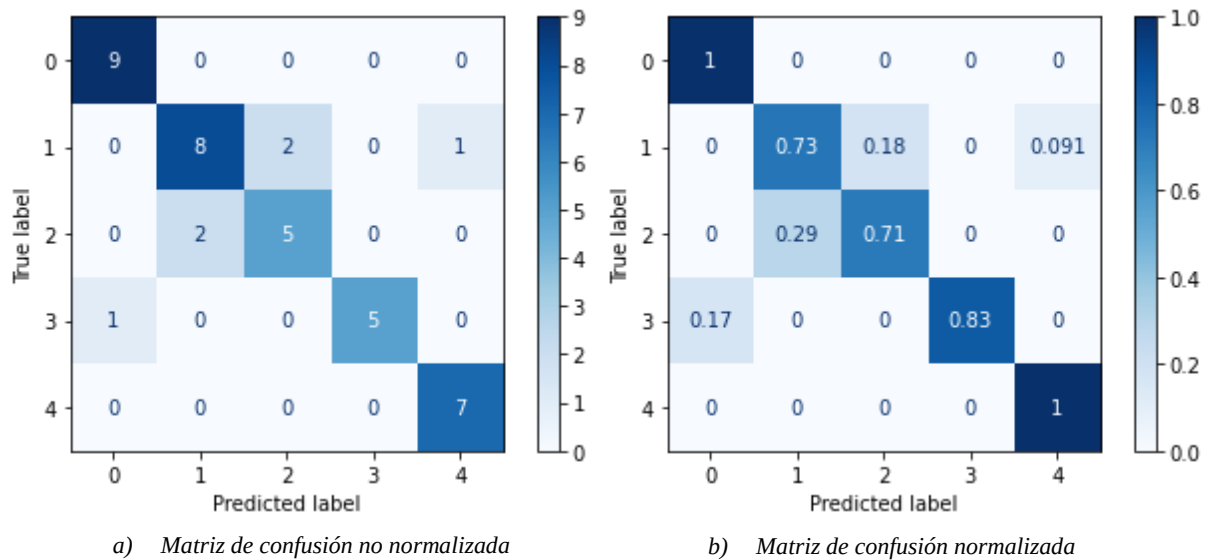


Figura 3.26 Matrices de confusión Random Forest

Con respecto a las métricas gráficas, en la Figura 3.27 se muestra la curva ROC del algoritmo Random Forest. Donde, el AUC más bajo es para la clase 1 y 2 con un valor de 0.83, y el AUC más alto es para la clase 0 y 4 con un valor de 0.99. Además, un macro promedio de 91% que indica un rendimiento sólido del modelo en términos de la capacidad de separar las clases. Gráfica y estadísticamente presenta un buen desempeño.

En la Figura 3.28 se muestra la curva PR del algoritmo Random Forest. Donde, el PR más bajo es para la clase 2 con un valor de 0.56, y el PR más alto es para la clase 0 con un valor de 0.90. Un micro promedio de 77% indica un rendimiento moderado del modelo en la precisión y en la identificación de ejemplos positivos.

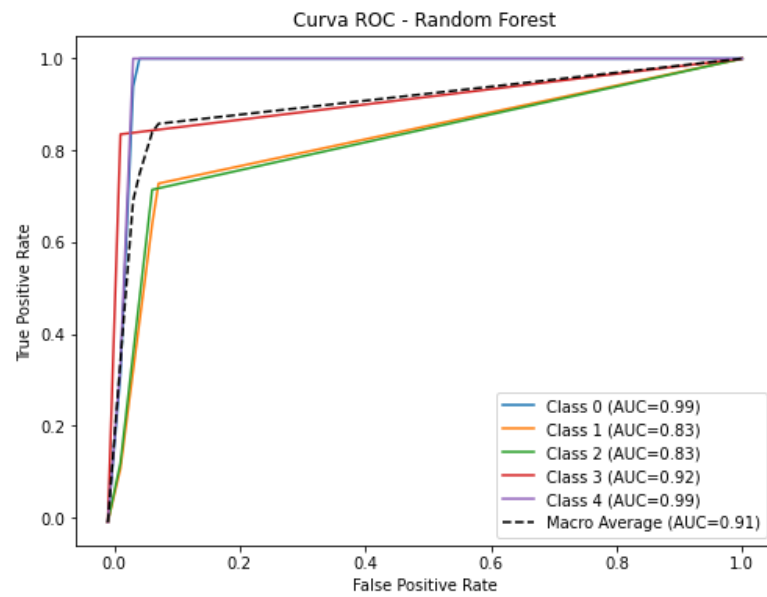


Figura 3.27 Curva ROC y AUC por Random Forest

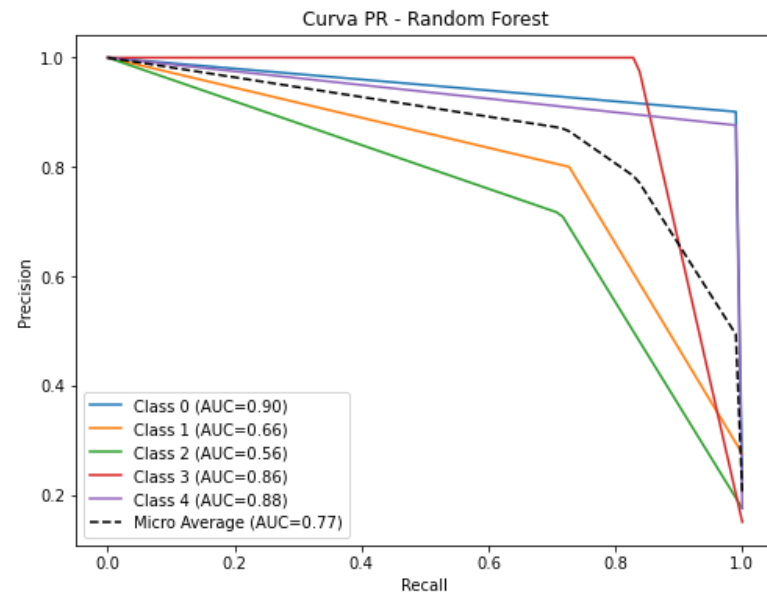


Figura 3.28 Curva PR por Random Forest

3.7.2.7 Rendimiento Gradient Boosting

Por otro lado, en la Tabla 3.11 se muestra el reporte de clasificación del algoritmo Gradient Boosting. En donde, la clase 0 presenta una clasificación perfecta para la falla tipo 0, es decir, no hay falsos positivos ni falsos negativos. En la clase 1, el 82% de las predicciones positivas para la falla tipo 1 son correctas e identifica el 82% de todos los casos reales. En la clase 2, el 62% de las predicciones positivas de la falla tipo 2 son correctas e identifica el 71% de todos los casos reales. Al igual que la clase 0, la clase 3 presenta una clasificación perfecta para la falla de tipo 3. Por último, para la clase 4 todas las predicciones positivas para la falla 4 son correctas, sin embargo, identifica el 86% de todos los casos reales de la misma.

En general, el algoritmo Gradient Boosting presenta una exactitud de 0.88, indicando que el 88% de todas las predicciones hechas por el modelo son correctas. Con altos valores de precisión y exhaustividad para la mayoría de las clases, se considera un buen rendimiento.

Tabla 3.11 Reporte de clasificación del algoritmo Gradient Boosting

Gradient Boosting					
	0	1	2	3	4
<i>Precision</i>	1.00	0.82	0.62	1.00	1.00
<i>Recall</i>	1.00	0.82	0.71	1.00	0.86
<i>F1-score</i>	1.00	0.82	0.67	1.00	0.92
<i>Support</i>	9	11	7	6	7
<i>Accuracy</i>	0.88				

Otras métricas para evaluar la clasificación con desequilibrio de clases se detallan a continuación. El algoritmo Gradient Boosting presenta una exactitud balanceada de 0.88, lo que indica que el modelo tiene un buen rendimiento. También, muestra un valor de 0.84 para el índice Kappa de Cohen lo que sugiere una concordancia sustancial entre las predicciones del modelo y las etiquetas reales. Un índice de Jaccard de 0.78 indica una similitud considerable entre las predicciones y las etiquetas reales.

Por su parte, un valor de 0.84 para el coeficiente de correlación de Matthews sugiere una estimación sustancial entre las predicciones y las etiquetas reales. Además, el algoritmo muestra una especificidad de 1 en la clase 0, lo que indica que es capaz de identificar correctamente las instancias de la clase negativa. La clase 1 tiene un valor de 0.82, indicando una especificidad relativamente alta para esa clase. En la clase 2 un valor de 0.71, indicando una especificidad moderada para esa clase. Para la clase 3, el valor 1 indica una especificidad alta para esa clase. Y, la clase 4 tiene una especificidad relativamente alta de 0.86.

Los resultados son positivos y sugieren que el modelo tiene un buen rendimiento en la tarea de clasificación multiclase, con una alta precisión balanceada, un índice Kappa de Cohen sustancial y otras métricas que indican una buena concordancia, similitud y calidad en las predicciones en todas las clases.

Con respecto a la matriz de confusión, en la Figura 3.29 se muestra las matrices de confusión del algoritmo Gradient Boosting, tanto normalizada como no normalizada. Se puede observar en la matriz no normalizada que 5 elementos son mal clasificados, de los cuales, dos fallas tipo 1 se clasifican como falla tipo 2, dos fallas tipo 2 se clasifican como falla tipo 1, y una falla de tipo 4 se clasifica como falla de tipo 2. Este algoritmo presenta un buen rendimiento.

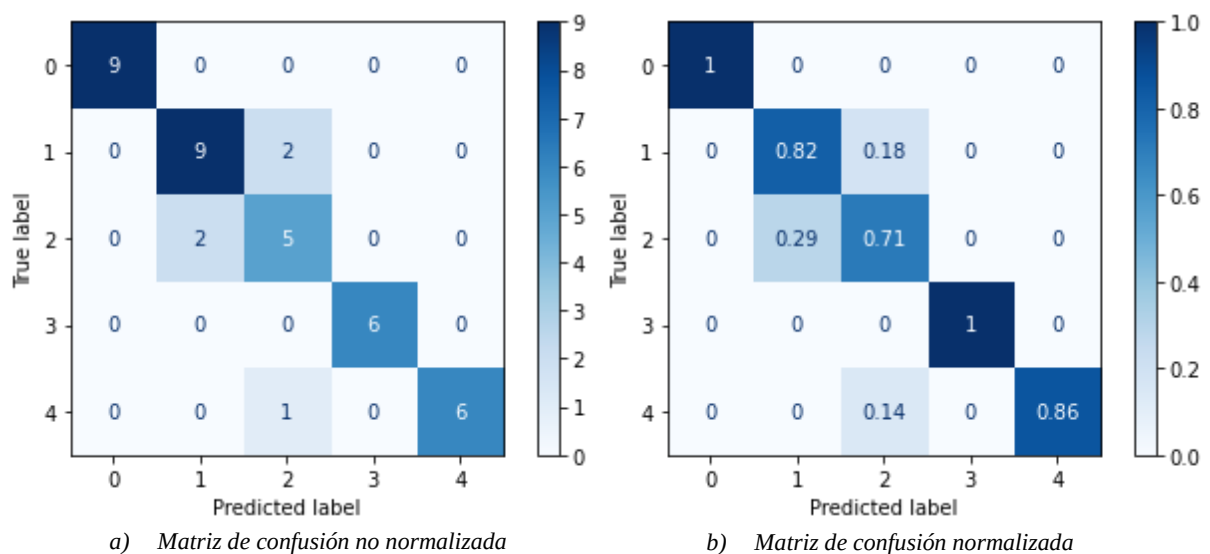


Figura 3.29 Matrices de confusión Gradient Boosting

Ahora, en la Figura 3.30 se muestra la curva ROC del algoritmo Gradient Boosting. Donde, el AUC más bajo es para la clase 2 con un valor de 0.81 y el AUC más alto es para la clase 0 y 3 con un valor de 1. Además, un macro promedio de 92% que indica un rendimiento sólido del modelo en términos de la capacidad de separar las clases. Gráfica y estadísticamente presenta un buen desempeño.

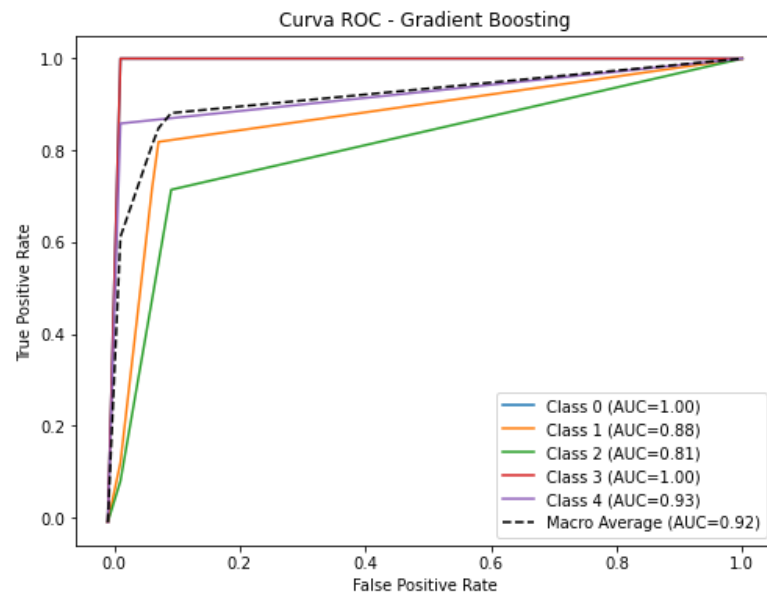


Figura 3.30 Curva ROC y AUC por Gradient Boosting

Finalmente, en la Figura 3.31 se muestra la curva PR del algoritmo Gradient Boosting. Donde, el PR más bajo es para la clase 2 con un valor de 0.50, y el PR más alto es para la clase 0 y 3 con un valor de 1. Un micro promedio de 82% indica un rendimiento sólido del modelo en la precisión y en la identificación de ejemplos positivos.

En general, el algoritmo Gradient Boosting muestra un rendimiento relativamente bueno en la clasificación de fallas en generadores de potencia.

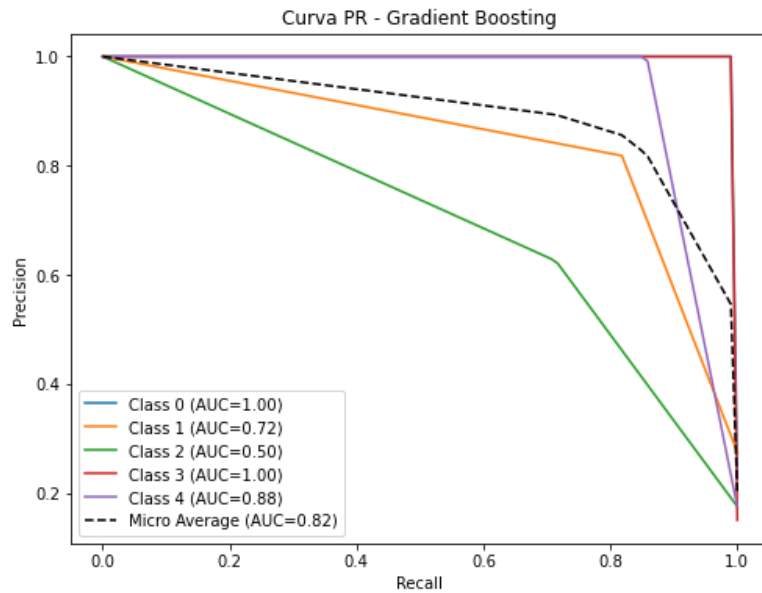


Figura 3.31 Curva PR por Gradient Boosting

3.7.2.8 Rendimiento XG Boost

Con respecto al algoritmo XG Boost, en la Tabla 3.12 se muestra su reporte de clasificación. Donde, en la clase 0 indica una clasificación perfecta para la falla de tipo 0, es decir, no hay falsos positivos ni falsos negativos, el resultado es excelente. Para la clase 1, el 89% de las predicciones positivas para la falla 1 son correctas, pero identifica el 73% de todos los casos reales. Para la clase 2, el 62% de las predicciones positivas para la falla tipo 2 son correctas e identifica el 71% de todos los casos reales, lo que significa que el modelo detecta la mayoría de los casos reales de esta clase. Al igual que la clase 0, la clase 3 presenta una clasificación perfecta para la falla tipo 3. Por último, en la clase 4 el 75% de las predicciones positivas para la falla tipo 4 son correctas e identifica el 86% de todos los casos reales.

En general, el algoritmo XG Boost presenta una exactitud del 0.85, indicando que el 85% de todas las predicciones hechas por el modelo son correctas, por lo que se considera un buen rendimiento.

Otras métricas para evaluar la clasificación con desequilibrio de clases se detallan a continuación. El algoritmo XG Boost presenta una exactitud balanceada de 0.86, lo que indica que el modelo tiene un buen rendimiento.

También, muestra un valor de 0.81 para el índice Kappa de Cohen lo que sugiere una concordancia sustancial entre las predicciones del modelo y las etiquetas reales. Un índice de Jaccard de 0.74 indica una similitud considerable entre las predicciones y las etiquetas reales. Por su parte, un valor de 0.81 para el coeficiente de correlación de Matthews sugiere una estimación sustancial entre las predicciones y las etiquetas reales.

Además, el algoritmo muestra una especificidad de 1 en la clase 0, lo que indica que es capaz de identificar correctamente las instancias de la clase negativa. La clase 1 tiene un valor de 0.72, indicando una especificidad moderada para esa clase. En la clase 2 un valor de 0.71, indicando una especificidad moderada para esa clase. Para la clase 3, el valor 1 indica una especificidad alta para esa clase. Y, la clase 4 tiene una especificidad relativamente alta de 0.86.

Los resultados son positivos y sugieren que el modelo tiene un buen rendimiento en la tarea de clasificación multiclase, con una alta precisión balanceada, un índice Kappa de Cohen sustancial y otras métricas que indican una buena concordancia, similitud y calidad en las predicciones en todas las clases.

Tabla 3.12 Reporte de clasificación del algoritmo XG Boost

	XGBoost				
	0	1	2	3	4
<i>Precision</i>	1.00	0.80	0.62	1.00	0.86
<i>Recall</i>	1.00	0.73	0.71	1.00	0.86
<i>F1-score</i>	1.00	0.76	0.67	1.00	0.86
<i>Support</i>	9	11	7	6	7
<i>Accuracy</i>	0.85				

En la Figura 3.32 se muestra las matrices de confusión del algoritmo XG Boost, tanto normalizada como no normalizada. Se puede observar en la matriz no normalizada que 6 elementos son mal clasificados, de los cuales, dos fallas de tipo 2 se clasifican como falla tipo 1; dos fallas de tipo 1 y una falla de tipo 4 se clasifican como falla de tipo 2; y una falla de tipo 1 se clasifica como falla de tipo 4.

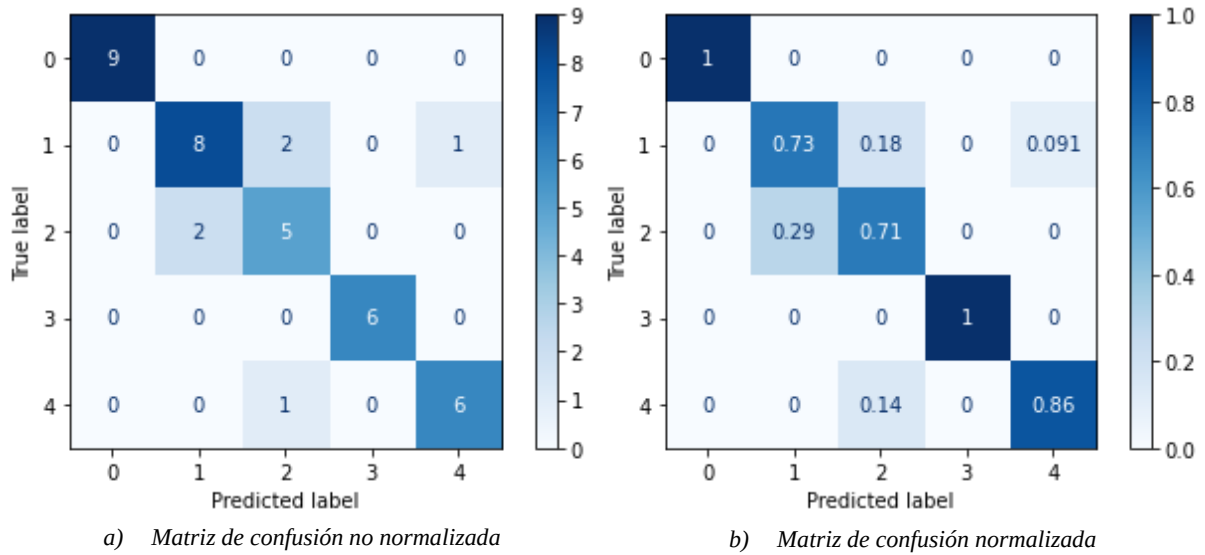


Figura 3.32 Matrices de confusión XG Boost

Por otro lado, en la Figura 3.33 se muestra la curva ROC del algoritmo XG Boost. Donde, el AUC más bajo es para la clase 2 con un valor de 0.81 y el AUC más alto es para la clase 0 y 3 con un valor de 1. Además, un macro promedio de 91% que indica un rendimiento sólido del modelo en términos de la capacidad de separar las clases. Gráfica y estadísticamente presenta un buen desempeño.

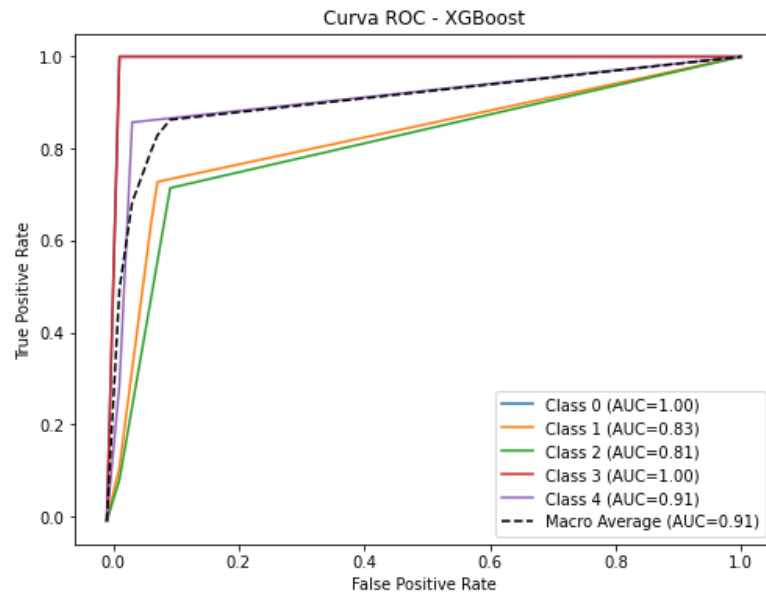


Figura 3.33 Curva ROC y AUC por XGBoost

En la Figura 3.34 se muestra la curva PR del clasificador del algoritmo XGBoost. Donde, el PR más bajo es para la clase 2 con un valor de 0.50, y el PR más alto es para la clase 0 y 3 con un valor de 1. Un micro promedio de 78% indica un rendimiento moderado del modelo en la precisión y en la identificación de ejemplos positivos.

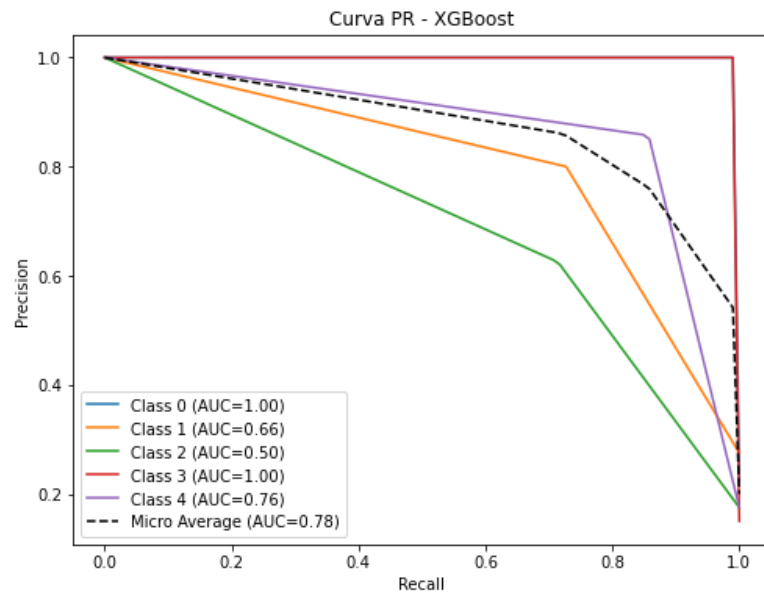


Figura 3.34 Curva PR por XGBoost

En general, el algoritmo XGBoost muestra un rendimiento relativamente bueno en la clasificación de fallas en generadores de potencia.

3.7.2.9 Rendimiento K – Nearest Neighbors

En la Tabla 3.13 se muestra el reporte de clasificación del algoritmo K – Nearest Neighbors. Donde, en la clase 0 todas las predicciones positivas de la falla tipo 0 son correctas, sin embargo, solo identifica el 78% de todos los casos reales en esa falla. En la clase 1, el 57% de las predicciones positivas de la falla 1 son correctas e identifica el 36% de todos los casos reales. En la clase 2, el 42% de las predicciones positivas de la falla tipo 2 son correctas e identifica el 71% de todos los casos reales de esa misma. En la clase 3, el 67% de las predicciones positivas de la falla 3 son correctas, pero identifica todos los casos reales. Por último, en la clase 4 el 60% de las predicciones positivas de la falla 4 son correctas, pero identifica el 43% de todos los casos reales de la misma.

En general, el algoritmo K – Nearest Neighbors presenta una exactitud del 0.62, indicando que el 62% de todas las predicciones hechas por el modelo son correctas, por lo que se considera un mal rendimiento.

Tabla 3.13 Reporte de clasificación del algoritmo K – NN

K – Nearest Neighbors					
	<i>0</i>	<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>
<i>Precision</i>	1.00	0.57	0.42	0.67	0.60
<i>Recall</i>	0.78	0.36	0.71	1.00	0.43
<i>F1-score</i>	0.88	0.44	0.53	0.80	0.50
<i>Support</i>	9	11	7	6	7
<i>Accuracy</i>	0.62				

Otras métricas para evaluar la clasificación con desequilibrio de clases se detallan a continuación. El algoritmo K – NN presenta una exactitud balanceada de 0.66, lo que indica que el modelo tiene un rendimiento bajo. También, muestra un valor de 0.53 para el índice Kappa de Cohen lo que sugiere una concordancia baja entre las predicciones del modelo y las etiquetas reales. Un índice de Jaccard de 0.45 indica una similitud baja entre las predicciones y las etiquetas reales. Por su parte, un valor de 0.55 para el coeficiente de correlación de Matthews sugiere una estimación baja entre las predicciones y las etiquetas reales. Además, el algoritmo muestra una especificidad de 0.78 en la clase de 0, lo que indica una especificidad moderada para identificar correctamente las instancias de la clase negativa. La clase 1 tiene un valor de 0.36, indicando una especificidad muy baja para esa clase. En la clase 2 un valor de 0.71, indicando una especificidad moderada para esa clase. Para la clase 3, el valor 1 indica una especificidad alta para esa clase. Y, la clase 4 tiene una especificidad baja de 0.43.

Los resultados no son positivos y sugieren que el modelo tiene un rendimiento bajo en la tarea de clasificación multiclase, con una baja precisión balanceada, un índice Kappa de Cohen sustancial y otras métricas que indican una moderada concordancia, similitud y calidad en las predicciones en todas las clases.

Por otra parte, en la Figura 3.35 se muestran las matrices de confusión, tanto normalizada como no normalizada, del algoritmo K-NN. Se puede observar en la matriz no normalizada que 15 elementos son mal clasificados, de los cuales; una falla de tipo 0, una falla de tipo 2 y una falla de tipo 4 se clasifican como falla de tipo 1; cuatro fallas de tipo 1 y tres fallas de tipo 4 se clasifican como falla de tipo 2; una falla de tipo 0 y dos fallas de tipo 1 se clasifican como falla de tipo 3; y, una falla de tipo 1 y una falla de tipo 2 se clasifican como falla de tipo 4. Por lo que este algoritmo no presenta un buen rendimiento.

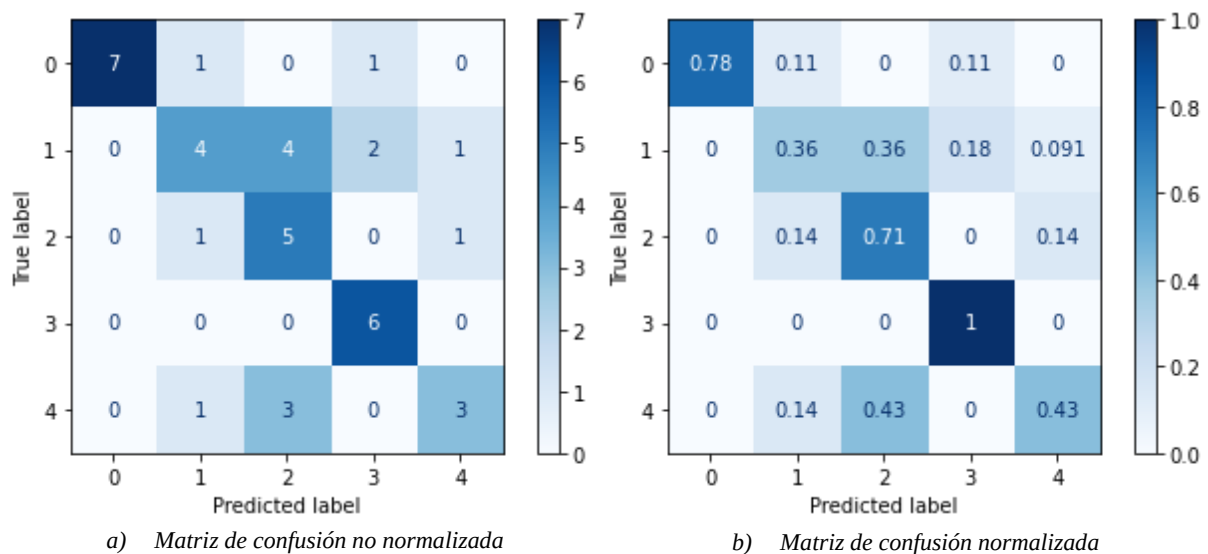


Figura 3.35 Matrices de confusión K-NN

Con respecto a las métricas gráficas, en la Figura 3.36 se muestra la curva ROC del algoritmo K-NN. Donde, el AUC más bajo pertenece a la clase 1 con un valor de 0.63, y el AUC más alto es para la clase 3 con un valor de 0.96. Además, un macro promedio de 78% que indica un rendimiento moderado del modelo en términos de la capacidad de separar las clases.

En la Figura 3.37 se muestra la curva PR del algoritmo K-NN. Donde, el PR más bajo es para la clase 2 con un valor de 0.35, y el PR más alto es para la clase 0 con un valor de 0.83. Un micro promedio de 52% indica un rendimiento bajo del modelo en la precisión y en la identificación de ejemplos positivos.

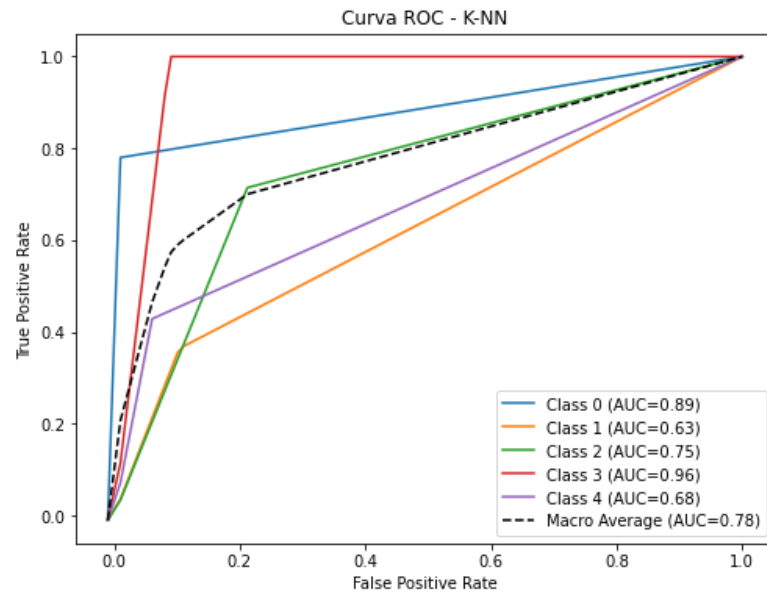


Figura 3.36 Curva ROC y AUC por K-NN

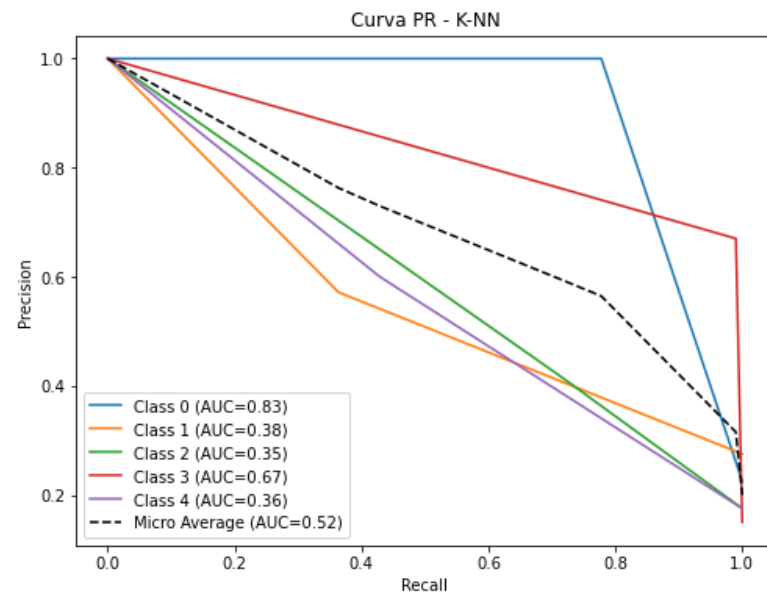


Figura 3.37 Curva PR por K-NN

3.7.2.10 Rendimiento Artificial Neural Network.

En la Tabla 3.14 se presenta el reporte de clasificación del algoritmo Artificial Neural Network. Donde, la clase 0 presenta una clasificación perfecta para la falla tipo 0. En la clase 1, el 55% de las predicciones positivas de la falla tipo 1 son correctas e identifica el 55% de todos los casos reales de la misma. En la clase 2, el 50% de las predicciones positivas de la falla tipo 2 son correctas e identifica el 14% de todos los casos reales de esa misma. En la clase 3, el 80% de las predicciones positivas de la falla tipo 3 son correctas, sin embargo, identifica el 67% de todos los casos reales de la misma. Por último, en la clase 4 el 54% de las predicciones positivas de la falla tipo 4 son correctas, pero identifica todos los casos reales de la misma.

En general, el algoritmo Artificial Neural Network presenta una exactitud del 0.68, indicando que el 68% de todas las predicciones hechas por el modelo son correctas, lo que significa que el algoritmo tiene un bajo rendimiento. Lo anterior se debe a que una red neural artificial necesita de un conjunto de datos grande para poder entrenar y aprender. En este caso, no se cuenta con un conjunto de datos grande de los generadores de potencia.

Tabla 3.14 Reporte de clasificación del algoritmo Artificial Neural Network

	Artificial Neural Network				
	<i>0</i>	<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>
<i>Precision</i>	1.00	0.55	0.50	0.80	0.54
<i>Recall</i>	1.00	0.55	0.14	0.67	1.00
<i>F1-score</i>	1.00	0.55	0.22	0.73	0.70
<i>Support</i>	9	11	7	6	7
<i>Accuracy</i>	0.68				

Otras métricas para evaluar la clasificación con desequilibrio de clases se detallan a continuación. El algoritmo Artificial Neural Network presenta una exactitud balanceada de 0.67, lo que indica que el modelo tiene un rendimiento muy bajo. También, muestra un valor de 0.59 para el índice Kappa de Cohen lo que sugiere una concordancia baja entre las predicciones del modelo y las etiquetas reales. Un índice de Jaccard de 0.51 indica una similitud baja entre las

predicciones y las etiquetas reales. Por su parte, un valor de 0.60 para el coeficiente de correlación de Matthews sugiere una estimación baja entre las predicciones y las etiquetas reales. Además, el algoritmo muestra una especificidad de 1 en la clase de 0, lo que indica una especificidad alta para identificar correctamente las instancias de la clase negativa. La clase 1 tiene un valor de 0.54, indicando una especificidad muy baja para esa clase. En la clase 2 un valor de 0.14, indicando una especificidad muy baja para esa clase. Para la clase 3, el valor 0.66 indica una especificidad baja para esa clase. Y, la clase 4 tiene una especificidad alta de 1.

Los resultados no son positivos y sugieren que el modelo tiene un rendimiento bajo en la tarea de clasificación multiclase, con una baja precisión balanceada, un índice Kappa de Cohen sustancial y otras métricas que indican una moderada concordancia, similitud y calidad en las predicciones en todas las clases.

Por otro lado, en la Figura 3.38 se muestra las matrices de confusión, tanto normalizada como no normalizada, del algoritmo Artificial Neural Network. Se puede observar en la matriz no normalizada que 13 elementos son mal clasificados, de los cuales; tres fallas de tipo 2 y dos fallas de tipo 3 son clasificadas como fallas de tipo 1; una falla de tipo 1 es clasificada como falla de tipo 2; otra falla de tipo 1 es clasificada como falla de tipo 3; y, tres fallas de tipo 1 y tres fallas de tipo 2 son clasificadas como fallas de tipo 4. Por lo que dicho algoritmo no presenta un buen rendimiento.

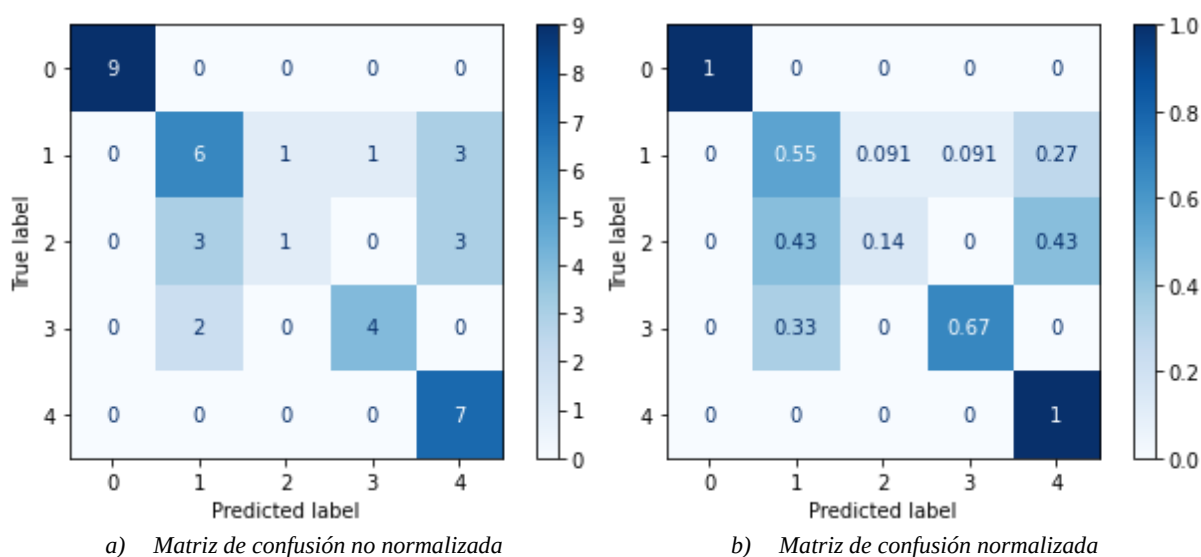


Figura 3.38 Matrices de confusión Artificial Neural Network

En la Figura 3.39 se muestra la curva ROC del algoritmo Artificial Neural Network. Donde, el AUC más bajo es para la clase 2 con un valor de 0.56, y el AUC más alto es para la clase 0 con un valor de 1. Además, un macro promedio de 79% que indica un rendimiento moderado del modelo en términos de la capacidad de separar las clases.

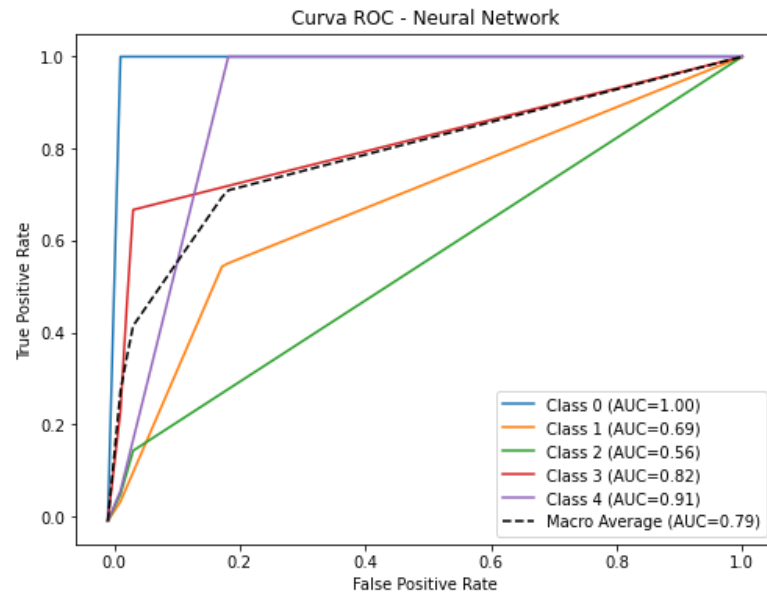


Figura 3.39 Curva ROC y AUC por Artificial Neural Network

Finalmente, en la Figura 3.40 se muestra la curva PR del algoritmo Artificial Neural Network. Donde, el PR más bajo es para la clase 2 con un valor de 0.22, y el PR más alto es para la clase 0 con un valor de 1. Un micro promedio de 55% indica un bajo rendimiento del modelo en la precisión y en la identificación de ejemplos positivos.

Estadísticamente, el algoritmo Artificial Neural Network presenta un bajo rendimiento de clasificación de fallas en generadores de potencia.

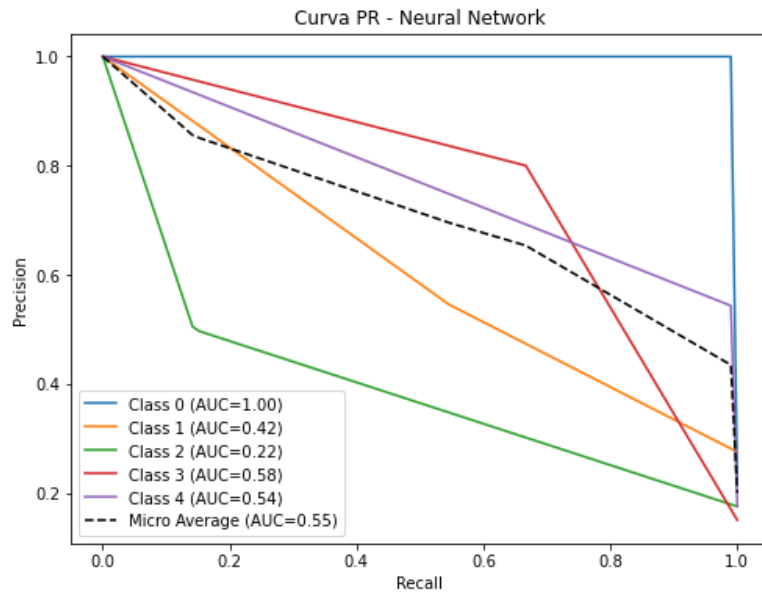


Figura 3.40 Curva PR por Neural Network

3.7.3 Seleccionar el mejor algoritmo

En la Tabla 3.15, se condensan los valores resultantes de cada clasificador en relación con las métricas de evaluación mencionadas anteriormente. Se observa que el algoritmo resaltado en negrita es el que logra el mejor rendimiento. Asimismo, en la Figura 3.41 se representa la curva ROC de los macro promedios de los clasificadores, mientras que la Figura 3.42 presenta la curva PR de los micro promedios de dichos clasificadores.

Una vez analizado las métricas de evaluación de los algoritmos de clasificación, el algoritmo Gradient Boosting presenta los mejores resultados de clasificación, prediciendo mejor las clases de tipo 3 y tipo 4 consideradas por expertos las fallas más graves en los generadores de potencia, por lo que es preferible tener una menor tasa de falsos negativos en este tipo de fallas.

3.7.4 Experimento

Se debe desarrollar un experimento con el algoritmo ganador, Gradient Boosting, en el cual se debe modificar los hiperparámetros del algoritmo para encontrar los mejores valores y asegurar el mejor rendimiento.

Tabla 3.15 Resumen de análisis de métricas

Algoritmo	Precision por clase					Recall por clase					F1-score por clase					Especificidad por clase					Índice Kappa de Cohen	Índice Jaccard	Correlación Matthews	Exactitud balanceada
	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4				
Support Vector Machines	0.90	0.67	0.00	1.00	0.38	1.00	0.73	0.00	0.33	0.86	0.95	0.70	0.00	0.50	0.52	1.00	0.73	0.00	0.33	0.86	0.52	0.45	0.56	0.58
Naive Bayes	1.00	0.75	0.78	0.50	1.00	0.67	0.55	1.00	0.83	1.00	0.80	0.63	0.88	0.62	1.00	0.66	0.54	1.00	0.84	1.00	0.72	0.63	0.73	0.78
Linear Discriminant Analysis	1.00	0.73	0.71	0.83	0.88	0.89	0.73	0.71	0.83	1.00	0.94	0.73	0.71	0.83	0.93	0.88	0.73	0.71	0.84	1.00	0.78	0.70	0.78	0.83
Quadratic Discriminant Analysis	0.89	0.64	0.71	0.67	1.00	0.89	0.64	0.71	0.67	1.00	0.89	0.64	0.71	0.67	1.00	0.88	0.64	0.71	0.66	1.00	0.72	0.63	0.72	0.78
Decision Tree	1.00	1.00	0.64	1.00	0.60	1.00	0.55	1.00	0.67	0.86	1.00	0.71	0.78	0.80	0.71	1.00	0.54	1.00	0.66	0.86	0.75	0.66	0.77	0.81
Random Forest	0.90	0.80	0.71	1.00	0.88	1.00	0.73	0.71	0.83	1.00	0.95	0.76	0.71	0.91	0.93	1.00	0.72	0.71	0.84	1.00	0.81	0.74	0.81	0.85
Gradient Boosting	1.00	0.82	0.62	1.00	1.00	1.00	0.82	0.71	1.00	0.86	1.00	0.82	0.67	1.00	0.92	1.00	0.82	0.71	1.00	0.86	0.84	0.78	0.84	0.88
XGBoost	1.00	0.80	0.62	1.00	0.86	1.00	0.73	0.71	1.00	0.86	1.00	0.76	0.67	1.00	0.86	1.00	0.72	0.71	1.00	0.86	0.81	0.74	0.81	0.86
K – Nearest Neighbors	1.00	0.57	0.42	0.67	0.60	0.78	0.36	0.71	1.00	0.43	0.88	0.44	0.53	0.80	0.50	0.78	0.36	0.71	1.00	0.43	0.53	0.45	0.55	0.66
Artificial Neural Network	1.00	0.55	0.50	0.80	0.54	1.00	0.55	0.14	0.67	1.00	1.00	0.55	0.22	0.73	0.70	1.00	0.54	0.14	0.66	1.00	0.59	0.51	0.60	0.67

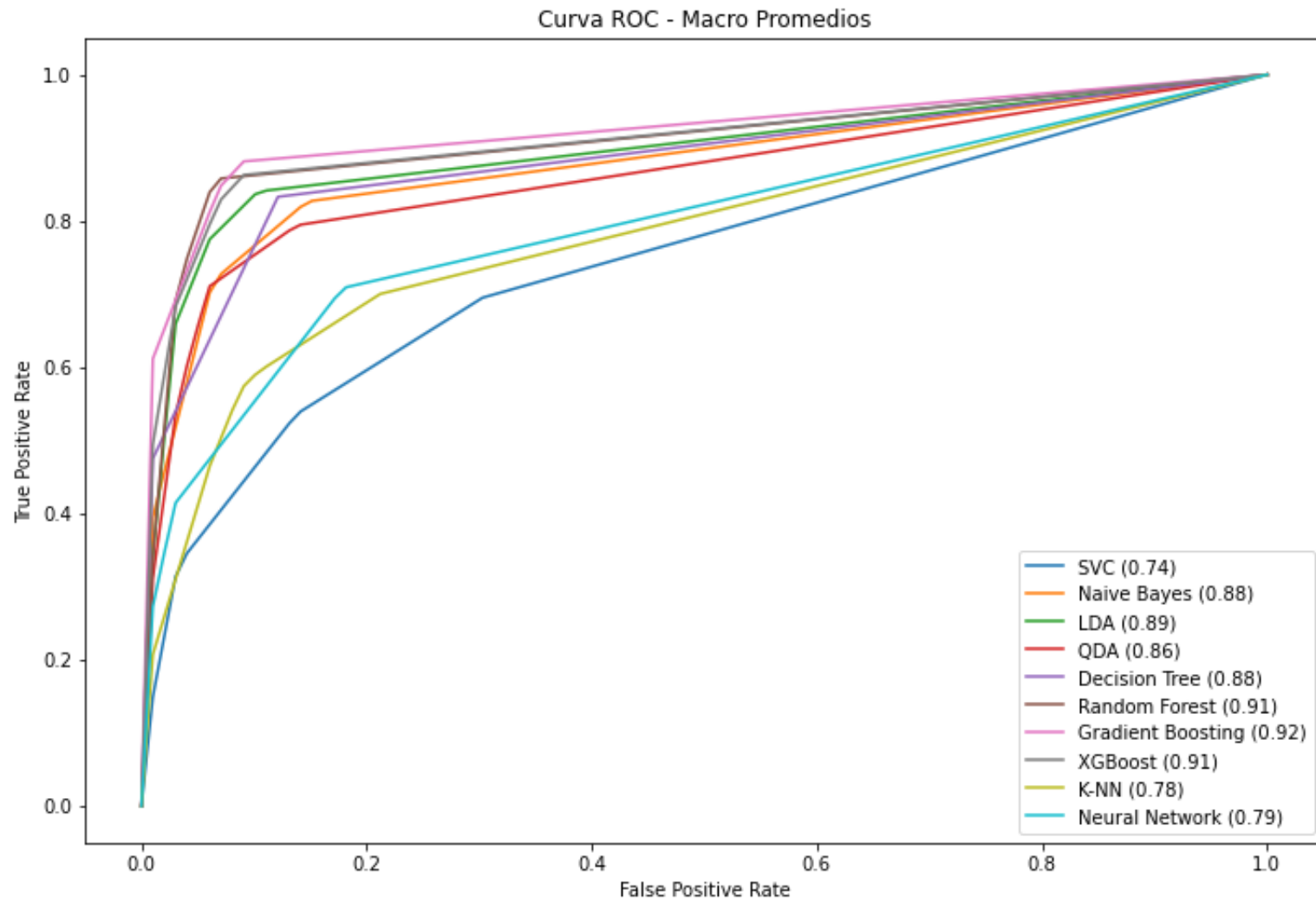


Figura 3.41 Curva ROC - Macro Promedios

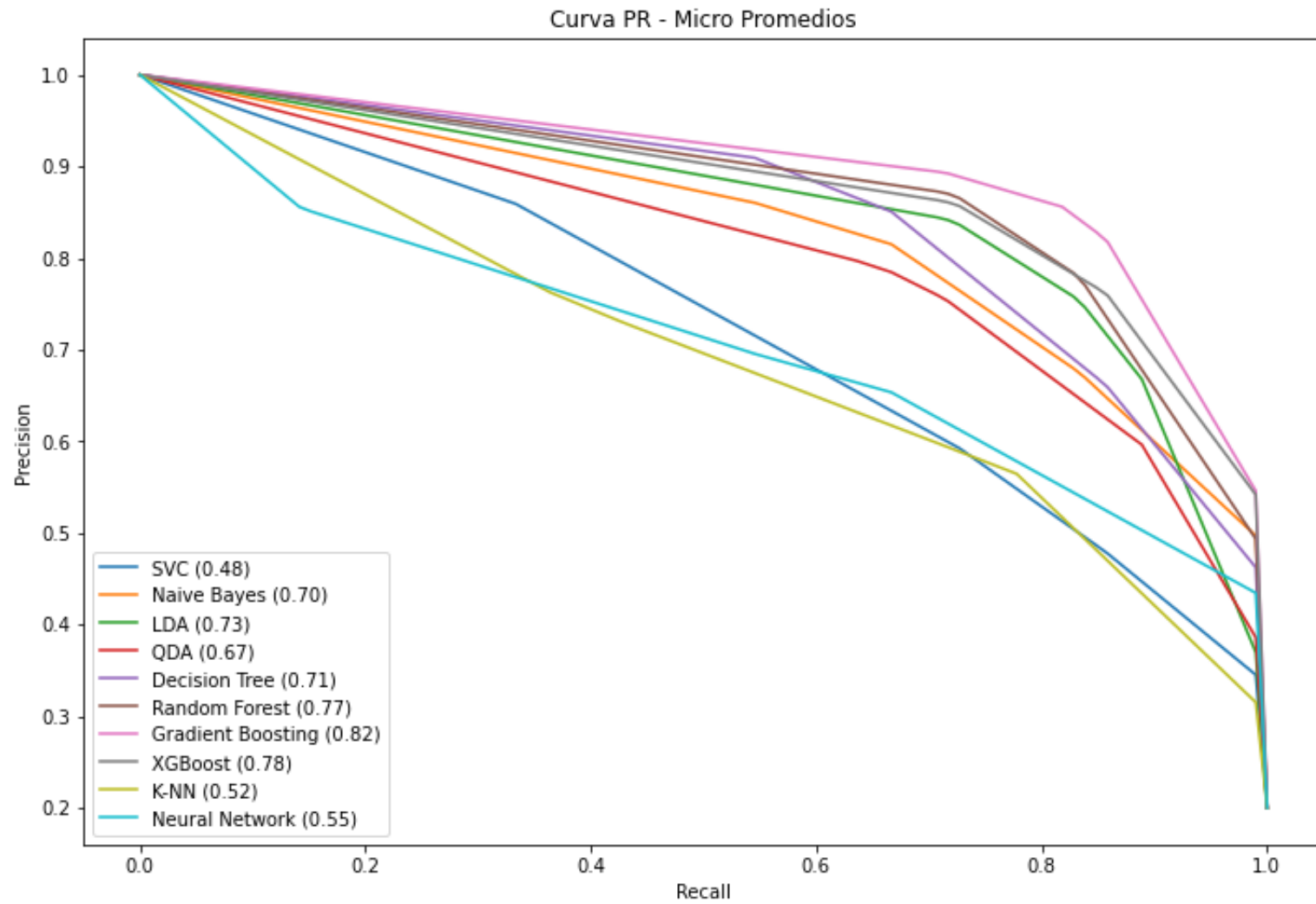


Figura 3.42 Curva PR - Micro Promedios

3.8 Ajuste del modelo

El ajuste de modelo se refiere al proceso de encontrar la combinación óptima de parámetros para un modelo de ML con el fin de lograr un rendimiento óptimo en un problema específico. Esto implica experimentar con diferentes valores de hiperparámetros y evaluar cómo afecta el rendimiento del modelo en un conjunto de datos de validación. El ajuste de modelo es un paso crucial, ya que determina en gran medida la calidad de las predicciones que el modelo puede realizar. En este trabajo, se implementa la técnica Búsqueda en rejilla para encontrar la combinación óptima de los hiperparámetros del algoritmo Gradient Boosting, y así obtener el mejor rendimiento del modelo de clasificación de fallas en generadores de potencia.

3.8.1 Implementar búsqueda en rejilla

La búsqueda en rejilla, o Grid search en inglés, es una técnica utilizada en ML para encontrar la mejor combinación de hiperparámetros de un modelo. Los hiperparámetros son configuraciones ajustables que no se aprenden automáticamente durante el entrenamiento del modelo y afectan su comportamiento y rendimiento. La búsqueda en rejilla es una estrategia sistemática para probar diferentes valores de hiperparámetros de manera exhaustiva dentro de un rango determinado.

3.8.1.1 Ajuste de hiperparámetros

El proceso de búsqueda en rejilla implica identificar los hiperparámetros del modelo que se desean ajustar y establecer un conjunto de posibles valores para cada uno. En la Figura 3.43 se muestra el código de la técnica búsqueda en rejilla. Para este análisis específico, se generaron un total de 990 combinaciones con un tiempo de entrenamiento de aproximadamente 6 horas.

```
#Búsqueda Grid
from sklearn.model_selection import GridSearchCV

parameters = {
    "max_features": ['sqrt', 'log2', None],
    "max_depth": [None],
    "n_estimators": range(2, 1000, 30),
    "min_samples_split": range(2, 100, 10)
}

clf1 = GridSearchCV(clf, parameters, cv = 5)
```

Figura 3.43 Código búsqueda en rejilla

En la Tabla 3.16 se presenta los hiperparámetros que se optimizan para mejorar el rendimiento del algoritmo Gradient Boosting.

Tabla 3.16 Ajuste de hiperparámetros en el algoritmo Gradient Boosting

Hiperparámetros	Descripción	Mejor estimador
max_features	El número máximo de características que se consideran al tomar decisiones sobre divisiones en un nodo del árbol.	‘sqrt’
max_depth	El número de divisiones que cada árbol de decisión puede realizar.	None
n_estimators	El número de árboles de decisión en el bosque aleatorio.	122
min_samples_split	El número mínimo de muestras necesarias para dividir un nodo interno.	22

El estimador ‘sqrt’ indica al modelo seleccionar automáticamente la raíz cuadrada del número total de características. Por su parte, el estimador None indica al modelo no establecer un límite máximo en la profundidad del árbol, es decir, el árbol puede crecer hasta que se cumplan ciertas condiciones de parada o hasta que se alcance un nivel de pureza deseado en las hojas del árbol.

3.8.2 Entrenar modelo

Luego de aplicar la técnica búsqueda en rejilla, se entrena nuevamente el algoritmo Gradient Boosting con los hiperparámetros optimizados, utilizando la misma división de datos de entrenamiento (80%) y prueba (20%).

3.8.3 Evaluar el modelo

En esta sección, se evalúa el rendimiento del algoritmo de clasificación Gradient Boosting optimizado. A través de un análisis de métricas específicas que consideran el desequilibrio presente

en la distribución de las clases. De esta manera, se proporciona una evaluación más justa y equitativa del rendimiento de los modelos.

En la Tabla 3.17 se muestra el reporte de clasificación del algoritmo Gradient Boosting optimizado, en donde, la clase 0 indica una clasificación perfecta para la falla de tipo 0, es decir, no hay falsos positivos ni falsos negativos. Para la clase 1, el 90% de las predicciones positivas para la falla 1 son correctas e identifica el 82% de todos los casos reales de esa falla. Para la clase 2, el 75% de las predicciones positivas para la falla tipo 2 son correctas e identifica el 86% de los casos reales de la misma. Al igual que la clase 0, la clase 3 y la clase 4 presentan una clasificación perfecta para las fallas tipo 3 y tipo 4, respectivamente. El algoritmo Gradient Boosting optimizado indica que el 93% de todas las predicciones hechas por el modelo son correctas, indicando un buen rendimiento.

Tabla 3.17 Reporte de clasificación del algoritmo Gradient Boosting optimizado

	Gradient Boosting optimizado				
	0	1	2	3	4
<i>Precision</i>	1.00	0.90	0.75	1.00	1.00
<i>Recall</i>	1.00	0.82	0.86	1.00	1.00
<i>F1-score</i>	1.00	0.86	0.80	1.00	1.00
<i>Support</i>	9	11	7	6	7
<i>Accuracy</i>	0.93				

Otras métricas para evaluar la clasificación con desequilibrio de clases se detallan a continuación. El algoritmo Gradient Boosting presenta una exactitud balanceada de 0.93, lo que indica que el modelo tiene un alto rendimiento. También, muestra un valor de 0.90 para el índice Kappa de Cohen lo que sugiere una concordancia sustancial entre las predicciones del modelo y las etiquetas reales.

Un índice de Jaccard de 0.86 indica una similitud relativamente alta entre las predicciones y las etiquetas reales. Por su parte, un valor de 0.91 para el coeficiente de correlación de Matthews sugiere una estimación sustancial entre las predicciones y las etiquetas reales.

Además, el algoritmo muestra una especificidad de 1 en la clase 0, lo que indica que es capaz de identificar correctamente las instancias de la clase negativa. La clase 1 tiene un valor de 0.82, indicando una especificidad relativamente alta para esa clase. En la clase 2 un valor de 0.86, indicando una especificidad relativamente alta para esa clase. Al igual que la clase 0, las clases 3 y 4 presentan una especificidad con valor de 1, lo que indica una especificidad alta.

Los resultados son muy positivos y sugieren que el modelo tiene un alto rendimiento en la tarea de clasificación multiclase, con una alta precisión balanceada, un índice Kappa de Cohen sustancial y otras métricas que indican una buena concordancia, similitud y calidad en las predicciones en todas las clases.

Con respecto a la métrica matriz de confusión, en la Figura 3.44 se muestra las matrices de confusión, tanto normalizada como no normalizada, del algoritmo Gradient Boosting optimizado. Se puede observar en la matriz no normalizada que 3 elementos se clasificaron mal, de los cuales, dos elementos de clase 1 se clasifican como clase 2, y un elemento de clase 2 se clasifica como clase 1.

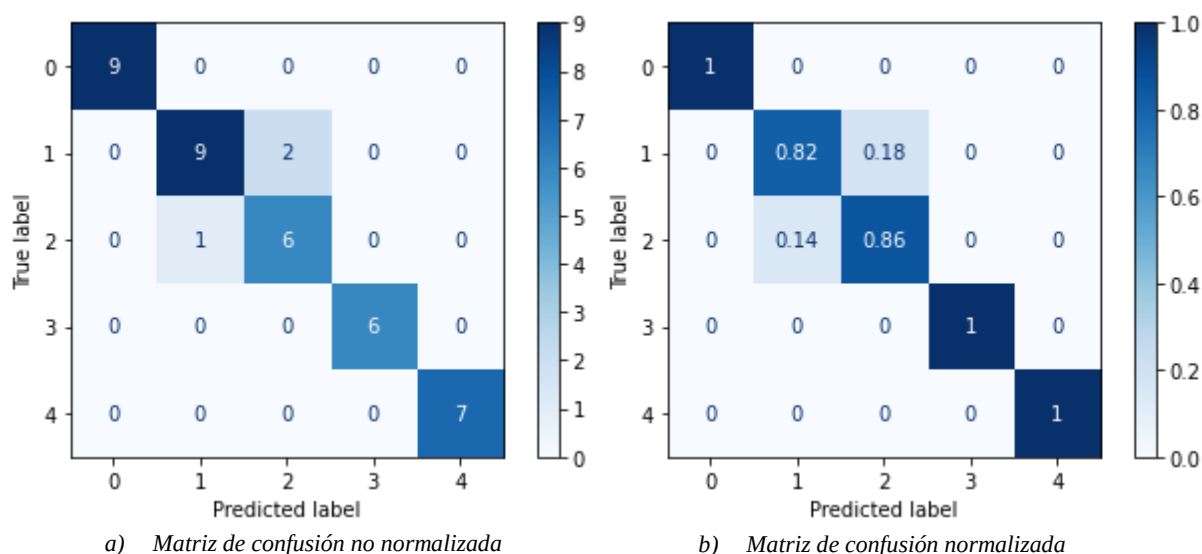


Figura 3.44 Matrices de confusión Gradient Boosting optimizado

Por otra parte, en la Figura 3.45 se muestra la curva ROC del algoritmo Gradient Boosting optimizado. Donde, el AUC más bajo es para la clase 1 con un valor de 0.89 y el AUC más alto es para la clase 0, 3 y 4 con un valor de 1. Además, un macro promedio de 96% que indica un rendimiento sólido del modelo en términos de la capacidad de separar las clases. Gráfica y estadísticamente presenta un buen desempeño.

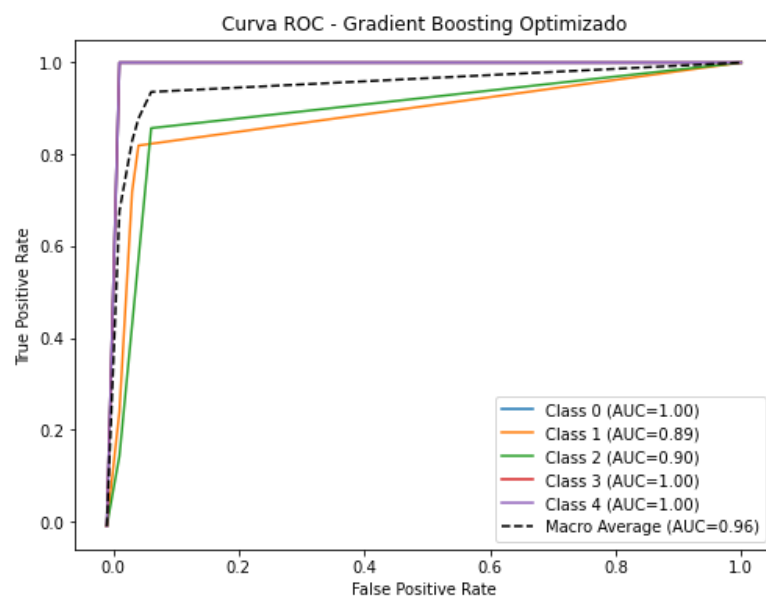


Figura 3.45 Curva ROC y AUC por Gradient Boosting optimizado

En la Figura 3.46 se muestra la curva PR del algoritmo Gradient Boosting optimizado. Donde, el PR más bajo es para la clase 2 con un valor de 0.67, y el PR más alto es para la clase 0, 3 y 4 con un valor de 1. Un micro promedio de 89% indica que el modelo presenta un rendimiento relativamente alto en la precisión y en la identificación de ejemplos positivos.

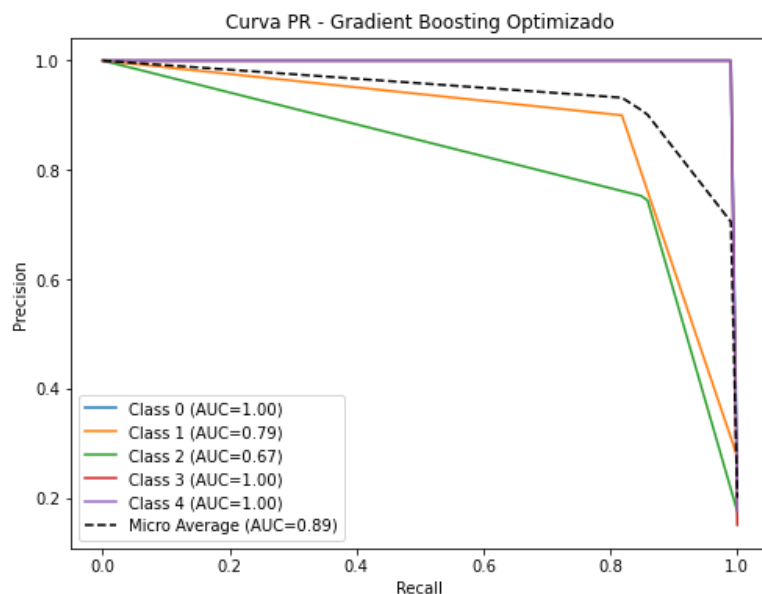


Figura 3.46 Curva PR por Gradient Boosting optimizado

En general, la técnica búsqueda en rejilla encuentra los valores óptimos de los hiperparámetros del algoritmo Gradient Boosting, arrojando resultados con un alto rendimiento en la clasificación de fallas en generadores de potencia, con altos valores en las métricas de evaluación en la mayoría de las clases.

3.8.4 Interpretación del modelo

La interpretación de un modelo de Gradient Boosting optimizado conlleva la comprensión de cómo el modelo toma sus decisiones y de qué manera influyen diversos factores en las predicciones resultantes. Un aspecto fundamental en este proceso radica en la identificación de las características más relevantes para el modelo.

La Figura 3.47 muestra la gráfica de importancia de características del clasificador Gradient Boosting optimizado. Esta visualización permite apreciar cuán significativa es cada característica en el modelo de aprendizaje automático optimizado. Las importancias de las características reflejan la contribución de cada una en la toma de decisiones del modelo.

Se observa que las características Enfriamiento_AIRE FORZADO y FD_Dtand_B presentan una importancia relativa superior a 0.10, lo que las coloca por encima de las demás características en términos de relevancia.

Esto sugiere que estas dos características son las más influyentes en el modelo Gradient Boosting optimizado, ya que ejercen un impacto más significativo en sus decisiones. Cabe destacar que la importancia potencialmente elevada de la característica relacionada con el tipo de enfriamiento podría deberse a que la mayoría de los datos en la base corresponden a este tipo de enfriamiento.

Las características subsiguientes se refieren a los factores de disipación y muestran una importancia relativa que oscila entre 0.06 y 0.08. Por otro lado, las características relacionadas con el tipo de generador y el enfriamiento tipo hidrógeno a una presión de 1.2 kg/cm² muestran un valor de importancia relativa muy bajo en las decisiones del modelo. Esto podría atribuirse a la escasez de datos pertenecientes a estas categorías, lo que resulta en una falta de información suficiente para el entrenamiento del modelo.

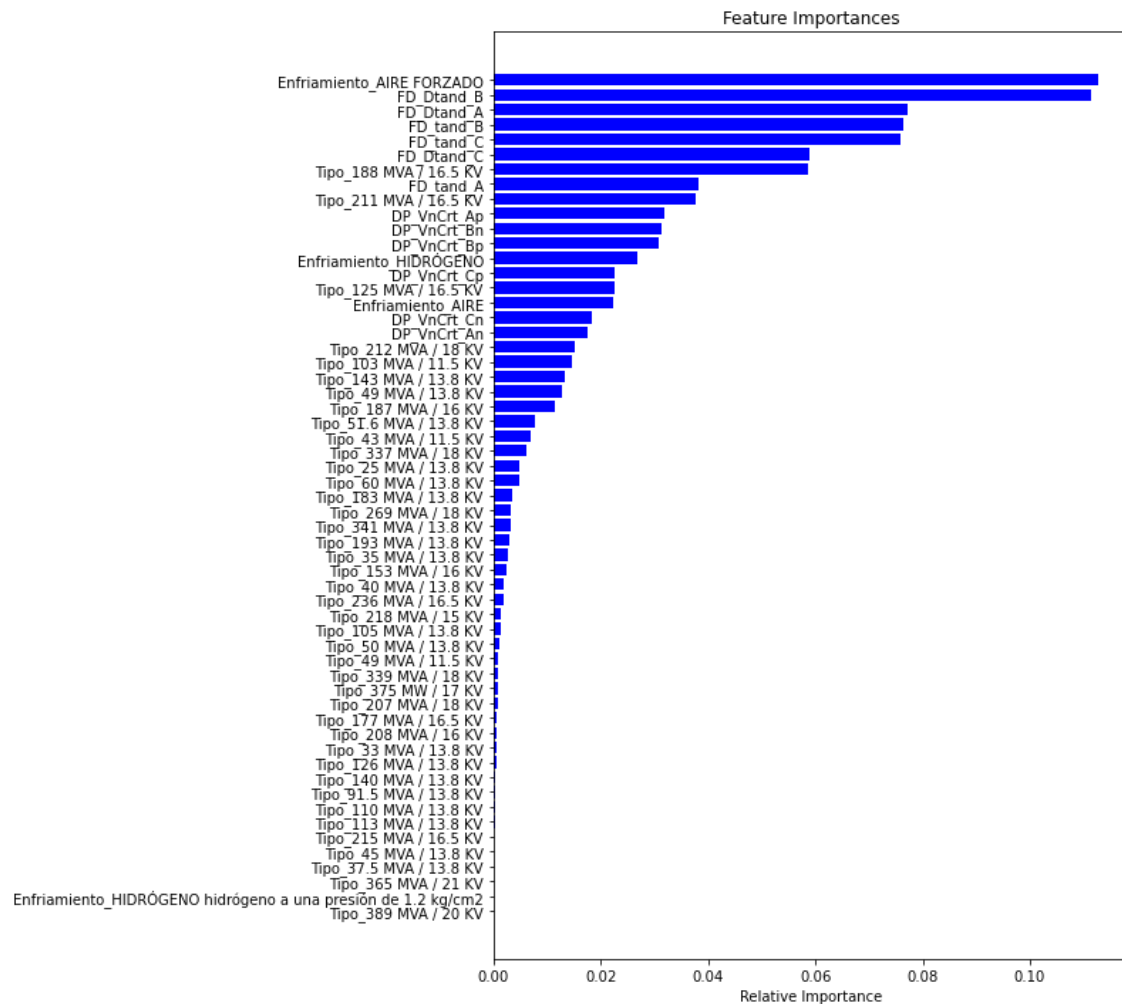


Figura 3.47 Importancia de características Gradient Boosting optimizado

3.8.4.1 Visualización del árbol de decisión

Un paso crítico en la interpretación de un modelo de Gradient Boosting es la visualización de árboles individuales para comprender cómo se toman decisiones en cada uno de ellos. Esta técnica permite identificar las características que desempeñan un papel fundamental en la división de los datos en clases o etiquetas.

El árbol de decisión divide las muestras en función de condiciones, asignando valores de destino. Este proceso se repite a medida que desciende por el árbol hasta llegar a las hojas. En la Figura 3.48 se presenta el primer árbol de decisión del modelo Gradient Boosting optimizado con un puntaje F1 macro de 93.14%.

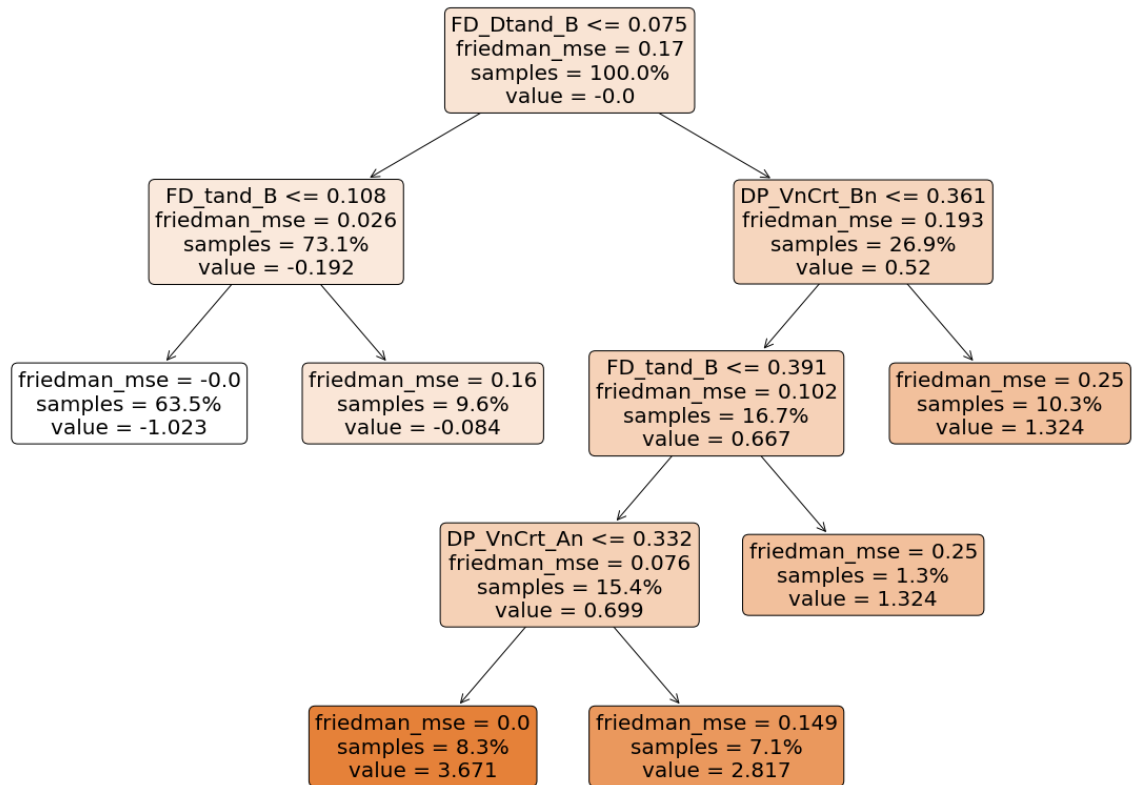


Figura 3.48 Primer árbol del modelo optimizado

En el nodo raíz, todas las muestras del conjunto de entrenamiento se dividen inicialmente en función de la condición $FD_Dtand_B \leq 0.075$. El 73.1% de las muestras que cumplieron con la condición del nodo raíz se dirigen hacia el nodo izquierdo que presenta la condición $FD_tand_B \leq 0.108$. En cambio, el 26.9% de muestras que no cumplieron la condición del nodo raíz ni la condición del nodo izquierdo se dirigen hacia el nodo derecho que presenta la condición $DP_VnCrT_Bn \leq 0.361$. Posteriormente, el 16.7% de muestras que cumplen con la condición del nodo derecho DP_VnCrT_Bn se dirigen hacia el nodo izquierdo, el cual presenta la condición $FD_tand_B \leq 0.391$.

Ahora, el 15.4% de las muestras que cumplen con la condición FD_tand_B se dirigen hacia el nodo izquierdo que presenta la condición $DP_VnCrT_An \leq 0.332$. De este último nodo se desglosan las hojas del árbol, en donde el 8.3% de las muestras que cumplen con la condición DP_VnCrT_An se dirigen a la hoja izquierda, la cual presenta un valor de destino de 3.671. Por otro lado, el 7.1% de las muestras que no cumplen con la condición DP_VnCrT_An ni la condición del nodo izquierdo se dirigen hacia el nodo derecho, el cual presenta un valor de destino de 2.817.

Los colores del árbol de decisión proporcionan información visual sobre la calidad de las divisiones en el árbol. En este caso, las hojas presentan tonos más oscuros lo que indica una mayor pureza, es decir, las muestras son predominantemente de una sola clase. Asimismo, la métrica MSE de Friedman mide la impureza de un nodo. Cuanto menor sea el MSE, menor es la impureza y mejor es la separación de clases.

Además, las características que se visualizan en el árbol de decisión son las relacionadas con los factores de disipación y las descargas parciales, lo que indica que son esenciales para comprender la toma de decisiones del modelo.

Es importante mencionar que un árbol individual de Gradient Boosting no proporciona directamente la visualización de las etiquetas de clase para un problema de clasificación, dado que los árboles en Gradient Boosting están diseñados para trabajar en conjunto y mejorar la predicción combinada del conjunto. Es decir, las predicciones de clases son resultado de las decisiones de múltiples árboles.

3.8.5 Análisis de errores

Un análisis de errores en ML hace referencia al proceso de examinar y comprender las discrepancias entre las predicciones del modelo y los valores reales en un conjunto de datos.

En la Tabla 3.18 se presenta las tres instancias que el modelo Gradient Boosting optimizado ha predicho de manera incorrecta.

La discrepancia observada entre el diagnóstico real y las predicciones del modelo en tres instancias puede explicarse, en parte, por el tamaño limitado del conjunto de datos de entrenamiento. Es posible que el modelo no haya tenido suficientes ejemplos para adquirir un entendimiento completo y preciso de situaciones particulares. Además, las características Enfriamiento y Tipo exhiben una amplia diversidad de categorías, muchas de las cuales cuentan con un número muy reducido de ejemplos disponibles en el conjunto de datos. Esta particularidad en la distribución de categorías plantea una dificultad en la modelización, ya que la escasez de ejemplos en algunas categorías puede limitar la capacidad del modelo para generalizar de manera efectiva en esas situaciones específicas.

Tabla 3.18 Instancias equivocadas del modelo optimizado

Características		Instancia 6	Instancia 14	Instancia 83
Enfriamiento		Enfriamiento_AIRE (0.091883)	Enfriamiento_AIRE (0.098965)	Enfriamiento_AIRE FORZADO (0.042816)
Tipo		Tipo_153 MVA / 16 KV (0.091883)	Tipo_125 MVA / 16.5 KV (0.098965)	Tipo_187 MVA / 16 KV (0.042816)
FD_tand_A		0.077182	0.059379	0.021836
FD_tand_B		0.077182	0.061358	0.021408
FD_tand_C		0.074425	0.0574	0.02098
FD_Dtand_A		0.118529	0.096986	0.01156
FD_Dtand_B		0.118529	0.095996	0.011132
FD_Dtand_C		0.116691	0.097976	0.012845
DP_VnCrT_Ap		0.404285	0.376068	0.385348
DP_VnCrT_An		0.404285	0.376068	0.385348
DP_VnCrT_Bp		0.367532	0.405757	0.406756
DP_VnCrT_Bn		0.367532	0.405757	0.406756
DP_VnCrT_Cp		0.404285	0.395861	0.428164
DP_VnCrT_Cn		0.404285	0.415654	0.428164
Diagnóstico	Real	2	1	1
	Predicho	1	2	2

Por otro lado, al analizar detenidamente los valores de las características en la base de datos, específicamente para las muestras de la clase 1 y la clase 2, y en particular para las instancias en las que se produjeron predicciones incorrectas, se destaca la notable similitud en la mayoría de las características. Por ejemplo, al examinar la instancia 6, se evidencia que los valores de las características presentan características distintivas de ambas clases, con la excepción de tres

atributos relacionados con los factores de disipación. En el caso de la instancia 14, las clases solo coinciden en dos características: Enfriamiento y Tipo. Finalmente, en el caso de la instancia 83, los valores de las características cumplen con las particularidades de ambas clases, excepto en dos características relacionadas con las descargas parciales.

Este hallazgo revela una alta similitud entre las instancias de la clase 1 y la clase 2 en términos de sus características, lo que puede generar un desafío adicional para el modelo en la tarea de distinguir entre estas clases. Específicamente, las similitudes en las características en las instancias mencionadas, pueden haber contribuido a las predicciones incorrectas, ya que la ambigüedad en la distinción entre estas clases dificulta la tarea de clasificación.

3.8.6 Modelo

El clasificador Gradient Boosting optimizado exhibe un rendimiento superior en las métricas de evaluación en la mayoría de las clases, destacando una exactitud balanceada del 93%, lo que representa una mejora significativa en comparación con el clasificador Gradient Boosting no optimizado.

Además, es relevante destacar que el modelo optimizado demuestra una capacidad sobresaliente en la clasificación de las clases 0, 3 y 4. Estas dos últimas clases son consideradas por expertos como las fallas más críticas en los generadores de potencia. Desde una perspectiva económica, minimizar la tasa de falsos negativos en la detección de estas fallas es de suma importancia.

En consecuencia, considerando el desempeño general y la adaptación del modelo a la base de datos de este estudio, se ha determinado que el modelo Gradient Boosting optimizado es la elección más adecuada, el cual será implementado en las etapas posteriores. Los hiperparámetros seleccionados para este modelo son los siguientes: `max_depth = None`, `max_features = 'sqrt'`, `n_estimators = 122`, `min_samples_split = 22`, y `random_state = 42`.

3.9 Conclusión

En este capítulo, se abordan las primeras seis etapas de la metodología propuesta para desarrollar un modelo inteligente destinado al diagnóstico de fallas en generadores de potencia mediante el empleo de algoritmos de ML. La problemática se analiza en detalle, y se formula la propuesta de solución, que consiste en la creación de un modelo inteligente. La base de datos comprende 196 casos y abarca 14 características vinculadas a diversos tipos de fallas en generadores de potencia. También, se establece un entorno de trabajo en Anaconda Navigator utilizando Python como lenguaje de programación. La etapa de exploración incluye un análisis estadístico descriptivo para comprender la naturaleza de los datos y seleccionar los algoritmos más adecuados. Y, en la etapa de pre-procesamiento, se convierten variables categóricas a representaciones numéricas, y se normaliza la base de datos, generando un archivo con 57 características relacionadas con el diagnóstico.

Posteriormente, se divide la base de datos en conjuntos de entrenamiento y prueba, y se evalúan 10 algoritmos de clasificación. Entre ellos, el algoritmo Gradient Boosting muestra el mejor rendimiento, según métricas de evaluación apropiadas para bases de datos desbalanceadas. Se realiza un ajuste mediante la implementación de búsqueda en rejilla para optimizar los hiperparámetros del algoritmo. Como resultado, se obtiene un modelo Gradient Boosting optimizado con una exactitud balanceada del 93%.

CAPÍTULO 4. RESULTADOS

4.1 Introducción

Este capítulo se enfoca en la Implementación, que constituye la fase final de la metodología propuesta para el desarrollo de un modelo inteligente diseñado para diagnosticar fallas en generadores de potencia mediante la aplicación de algoritmos de ML. Por lo que se va a transformar el modelo obtenido en el capítulo anterior en un sistema interactivo capaz de colaborar con expertos en generadores de potencia, posibilitando así su aplicación en escenarios prácticos.

4.2 Implementación

La implementación se refiere al proceso crucial en el cual el modelo de ML desarrollado se lleva desde el entorno de desarrollo y entrenamiento hacia un entorno de aplicación práctica. La implementación implica la configuración del sistema para que pueda recibir datos, procesarlos y generar predicciones o clasificaciones pertinentes.

En este caso, el modelo se convierte en una herramienta práctica y efectiva para el diagnóstico de fallas en generadores de potencia, desempeñando un papel crucial en la mejora de la eficiencia y la toma de decisiones en el campo de la generación de energía.

4.2.1 Evaluación de las predicciones del modelo

Antes de implementar el modelo desarrollado, es necesario evaluar las predicciones generadas por el modelo con nuevos datos, con el fin de garantizar su eficacia y precisión para diagnosticar fallas en generadores de potencia en un entorno operativo real.

4.2.1.1 Datos nuevos

Los datos utilizados para evaluar las predicciones del modelo Gradient Boosting fueron proporcionados nuevamente por el INEEL. Es importante destacar que estos datos fueron recopilados en campo y no han sido previamente validados por expertos, lo que podría implicar la presencia de posibles fuentes de ruido e incertidumbre elevada en los mismos.

La nueva colección de datos abarca un total de 71 casos vinculados a diversas fallas en los generadores de potencia. Sin embargo, se procedió a excluir algunos de estos casos, específicamente aquellos caracterizados por un sistema de enfriamiento hidroeléctrico, dado que el modelo previamente desarrollado no fue entrenado para abordar esta categoría particular de la variable Enfriamiento. Además, se identificaron casos con información faltante en algunas características, los cuales también fueron excluidos. Como resultado, el conjunto de datos en formato CSV finalmente contiene 50 casos, donde se observe una prevalencia de situaciones correspondientes a fallas de tipo 0 y tipo 1.

4.2.1.2 Preprocesamiento de los datos nuevos

Al realizar predicciones con el modelo Gradient Boosting sobre nuevos datos, se requiere aplicar las mismas transformaciones que se llevaron a cabo en el conjunto de datos de entrenamiento y prueba a los datos crudos. Este proceso abarca la normalización y la generación de variables dummy para las características categóricas, asegurando así la coherencia en la preparación de los datos y la congruencia con el preprocesamiento aplicado durante la fase de entrenamiento y evaluación del modelo.

Es fundamental que los nuevos datos abarquen todas las características presentes en el conjunto de datos utilizado para el entrenamiento del modelo. En el caso de que falte alguna de estas características en los datos nuevos, es necesario añadir las características faltantes asignándoles valores cero, con el propósito de llevar a cabo las predicciones de manera efectiva. Además, es esencial garantizar que las características de los datos nuevos mantengan tanto el mismo nombre como el mismo orden que las presentes en el conjunto de datos de entrenamiento.

En la Figura 4.1 se muestran los primeros casos del conjunto de datos nuevos preprocesados para su predicción.

	Enfriamiento_AIRE	Enfriamiento_HIDRÓGENO	Enfriamiento_HIDRÓGENO hidrógeno a una presión de 1.2 kg/cm2	Enfriamiento_AIRE FORZADO	Tipo_375 MW / 17 KV	Tipo_208 MVA / 16 KV	Tipo_218 MVA / 15 KV	Tipo_153 MVA / 16 KV	Tipo_125 MVA / 16.5 KV	Tipo_341 MVA / 13.8 KV
0	0.000000	0.309445	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0
1	0.000000	0.022929	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0
2	0.000000	0.000000	0.004742	0.000000	0.0	0.0	0.0	0.0	0.0	0.0
3	0.000000	0.006724	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0
4	0.000000	0.063188	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0
5	0.000000	0.301021	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0
6	0.000000	0.057154	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0
7	0.000000	0.058335	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0
8	0.000000	0.211917	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0
9	0.000000	0.260687	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0
10	0.000000	0.228874	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0
11	0.000000	0.260516	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0
12	0.000000	0.143889	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0
13	0.000000	0.123799	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0
14	0.000000	0.026683	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0

Figura 4.1 Datos nuevos preprocesados para su predicción

4.2.1.3 Análisis de métricas de predicción

La evaluación de las predicciones del modelo desarrollado se lleva a cabo mediante el análisis de las mismas métricas utilizadas en apartados anteriores. A continuación, se procede a la interpretación detallada de los resultados obtenidos para cada una de estas métricas.

En la Tabla 4.1 se muestra el reporte de clasificación de las predicciones del modelo Gradient Boosting. Donde, en la clase 0 todas las predicciones positivas para la falla tipo 0 son correctas, sin embargo, identifica el 88% de todos los casos reales de la misma. Para la clase 1, el 88% de las predicciones positivas para la falla 1 son correctas e identifica el 88% de todos los casos reales. Para la clase 2, el 71% de las predicciones positivas para la falla tipo 2 son correctas, pero identifica todos los casos reales de la misma falla. Finalmente, la clase 3 y la clase 4 indica una clasificación perfecta para la falla de tipo 3 y 4, respectivamente, es decir, no hay falsos positivos ni falsos negativos en esas clases.

Las predicciones generadas por el modelo Gradient Boosting presenta una exactitud de 0.92, lo que implica que el 92% de las predicciones realizadas por el modelo concuerdan correctamente con los resultados reales. Este resultado señala un rendimiento generalmente sólido del modelo en la tarea de clasificación de fallas en generadores de potencia.

Tabla 4.1 Reporte de clasificación de las predicciones del modelo Gradient Boosting

	Predicciones del modelo				
	0	1	2	3	4
<i>Precision</i>	1.00	0.88	0.71	1.00	1.00
<i>Recall</i>	0.88	0.88	1.00	1.00	1.00
<i>F1-score</i>	0.94	0.88	0.83	1.00	1.00
<i>Support</i>	17	17	5	4	7
<i>Accuracy</i>	0.92				

Analizando otras métricas, las predicciones del modelo presentan una exactitud balanceada del 92%, lo que indica predicciones precisas en todas las categorías. También, muestra un valor de 0.89 para el índice Kappa de Cohen lo que sugiere una concordancia sustancial entre las predicciones del modelo y las etiquetas reales. Un índice de Jaccard de 0.85 indica una similitud relativamente buena entre las predicciones y las etiquetas reales. Por otra parte, un valor de 0.89 para el coeficiente de correlación de Matthews sugiere una estimación sustancial entre las predicciones y las etiquetas reales. Con respecto a la especificidad, las predicciones muestran una especificidad de 0.88 en las clases 0 y 1, lo que indica que es capaz de identificar correctamente el 88% las instancias negativas en dichas clases. Las clases 2, 3 y 4 presentan una especificidad alta con valor de 1, lo que indica que el modelo tiene buena capacidad para evitar clasificar erróneamente casos negativos como positivos.

Por lo tanto, el modelo presenta en general un buen rendimiento en la tarea de clasificación multiclase, con una precisión balanceada y especificidad aceptable, una correlación de Matthews sustancial y otras métricas refuerzan la confiabilidad del modelo, proporcionando una medida de la concordancia más allá del acuerdo aleatorio.

Con respecto a la métrica matriz de confusión, en la Figura 4.2 se muestra las matrices de confusión, tanto normalizada como no normalizada, de las predicciones del modelo desarrollado. Se puede observar en la matriz no normalizada que 4 elementos se clasificaron mal, de los cuales, dos elementos de clase 0 se clasifican como clase 1, y dos elementos de clase 1 se clasifica como clase 2.

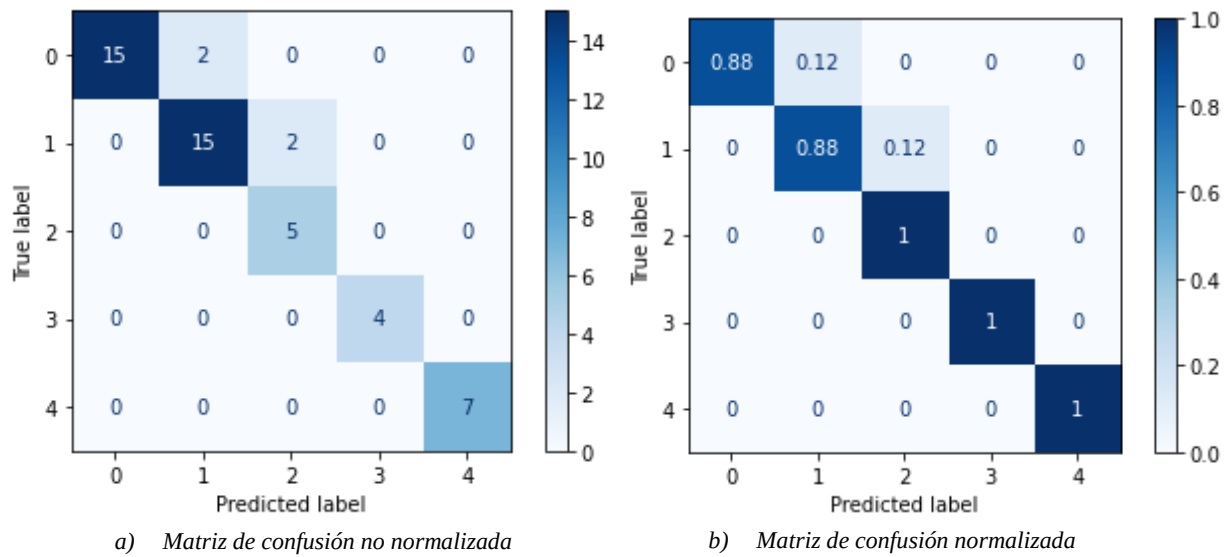


Figura 4.2 Matrices de confusión de las predicciones del modelo

Por otra parte, en la Figura 4.3 se muestra la curva ROC para las predicciones del modelo. Donde, el AUC más bajo es para la clase 1 con un valor de 0.94 y el AUC más alto es para las clases 3 y 4 con un valor de 1. Además, un macro promedio de 97% que indica un rendimiento sólido del modelo en términos de la capacidad de separar las clases. Gráfica y estadísticamente presenta un buen desempeño.

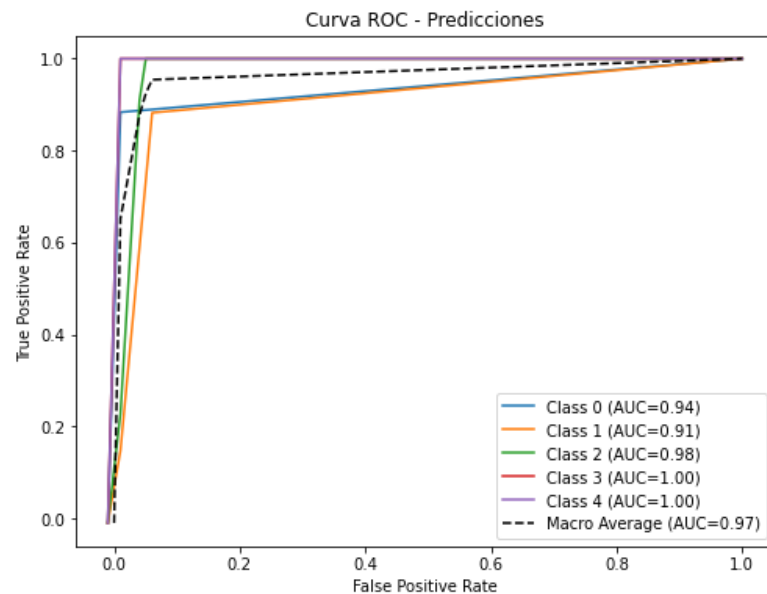


Figura 4.3 Curva ROC y AUC de las predicciones del modelo

En la Figura 4.4 se muestra la curva PR de las predicciones del modelo. Donde, el PR más bajo es para la clase 2 con un valor de 0.71, y el PR más alto es para las clases 3 y 4 con un valor de 1. Un micro promedio de 89% indica que el modelo presenta un rendimiento relativamente bueno en la precisión y en la identificación de ejemplos positivos.

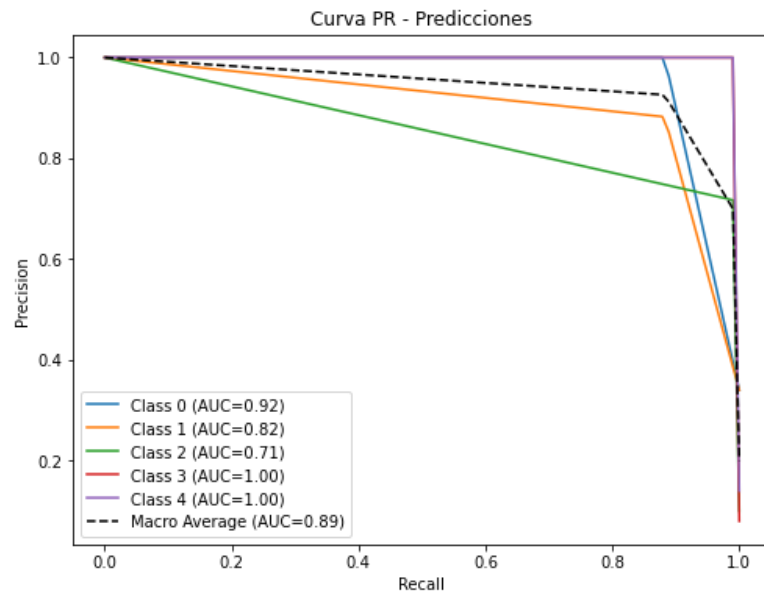


Figura 4.4 Curva PR de las predicciones del modelo

La evaluación del rendimiento del modelo de clasificación desarrollado para identificar cinco tipos de fallas en generadores de potencia revela resultados alentadores y prometedores. Las métricas proporcionan una visión integral del desempeño del modelo en distintos aspectos, destacando su capacidad para abordar múltiples clases y mantener un equilibrio en la precisión de las predicciones.

Estos resultados fortalecen la confianza en la aplicabilidad y eficacia del modelo en entornos del mundo real, proporcionando una herramienta valiosa para el diagnóstico preciso de fallas en generadores de potencia.

4.2.1.4 Análisis de errores de predicción

En la Tabla 4.2 se presentan los casos que el modelo predice incorrectamente. Es importante examinar detenidamente estas instancias con el fin de comprender las discrepancias entre las predicciones del modelo y las etiquetas reales del conjunto de datos nuevos.

Tabla 4.2 Predicciones equivocadas del modelo desarrollado

Instancia	Etiqueta real	Predicción
24	1	2
25	1	2
30	0	1
37	0	1

Las discrepancias en las etiquetas reales, específicamente en la clasificación de la clase 1 como clase 2 por parte del modelo, se atribuyen a la similitud en la mayoría de sus características. La distinción entre estas clases resulta más compleja, representando un desafío adicional para el modelo, tal como se detalla en el capítulo anterior. Con respecto a las discrepancias restantes, donde el modelo clasifica instancias de clase 0 como instancias de clase 1, revela la complejidad en la tarea de clasificación, especialmente cuando las características críticas comparten similitudes entre estas clases.

La instancia 30 muestra un enfriamiento de tipo Aire, el cual está presente en el conjunto de datos de entrenamiento de la clase 1 pero ausente en la clase 0. No obstante, dicha instancia presenta un tipo de generador 33 MVA / 13.8 KV que pertenece a casos de clase 0. Esto influye mucho en la clasificación del modelo, sin embargo, la característica Enfriamiento Aire posee mayor prioridad que la característica Tipo 33 MVA / 13.8 KV, según el gráfico de importancia de características. Además, la instancia presenta valores que se atribuyen a ambas clases en las características relacionadas con los factores de disipación y las descargas parciales. En un escenario similar, la instancia 37 presenta un enfriamiento de tipo Aire perteneciente a casos de clase 1, y un tipo de generador 126 MVA / 13.8 KV perteneciente a casos de clase 0, que, de acuerdo al gráfico de importancia de características, la característica Enfriamiento Aire tiene una ponderación relativa mayor que la característica Tipo 126 MVA / 13.8 KV.

En consecuencia, el modelo Gradient Boosting clasifique erróneamente las instancias 30 y 37 como pertenecientes a la clase 1, así como las instancias 24 y 25 como pertenecientes a la clase 2. Estas situaciones destacan la necesidad de una comprensión detallada de las relaciones entre las características para mejorar la robustez del modelo en escenarios de alta complejidad.

Este análisis destaca la necesidad corroborar la validación de los datos nuevos por parte de los expertos para disminuir la incertidumbre, y continuar mejorando el modelo para abordar de manera más efectiva la clasificación, particularmente cuando la variabilidad en ciertas características puede conducir a interpretaciones confusas.

4.2.2 Desarrollo de la interfaz gráfica

Para la creación del interfaz gráfico del modelo de diagnóstico de fallas en generadores de potencia mediante el empleo del algoritmo Gradient Boosting, se ha decidido utilizar el entorno de desarrollo de Python junto con la biblioteca Tkinter, debido a la disposición de elementos y su facilidad de navegación. Este proceso se inicia incorporando elementos visuales, como etiquetas, cuadros de texto, menús desplegables y botones, mediante la configuración de widgets proporcionados por Tkinter. El diseño resultante es sencillo pero funcional, amigable para el usuario final y facilitador tanto de la introducción de datos como de la interpretación de resultados.

En particular, se da relevancia a la inclusión de funciones de validación en los campos de entrada, que incluyen mensajes de advertencia y la verificación de campos obligatorios. Esto se realiza con el propósito de prevenir la introducción de valores no válidos por parte del usuario y asegurar la inclusión de toda la información necesaria.

Una vez completada la elaboración del interfaz gráfico, se procede a establecer la conexión con el modelo de diagnóstico basado en el algoritmo Gradient Boosting. En etapas posteriores, se definen áreas específicas en la interfaz para la presentación de resultados, utilizando etiquetas dinámicas para exhibir el diagnóstico resultante, el tipo de falla detectada y las recomendaciones correspondientes. La entrada del usuario se recopila y se utiliza para realizar predicciones mediante el modelo previamente entrenado.

Con el fin de garantizar la funcionalidad correcta del interfaz, se realizan pruebas, incluyendo la validación de campos de entrada y la verificación de la coherencia entre los resultados presentados en la interfaz y las predicciones generadas por el modelo de diagnóstico. Una vez confirmado su adecuado funcionamiento, se procede con el empaquetado del código para convertirlo en una aplicación de escritorio compatible con sistemas operativos Windows. Este enfoque integral en el diseño, validación y verificación asegura la robustez y eficacia del interfaz en la implementación práctica del modelo de diagnóstico propuesto.

4.2.3 Sistema

La implementación exitosa de esta metodología resulta en un interfaz gráfico intuitivo y funcional, proporcionando una herramienta efectiva para la interacción de los usuarios finales con el modelo de diagnóstico desarrollado. La Figura 4.5 muestra el ícono correspondiente a la aplicación de escritorio denominada Sistema Inteligente para el Diagnóstico de Fallas en Generadores de Potencia (SIDG).

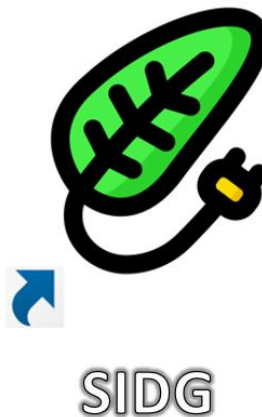


Figura 4.5 Ícono de SIDG

En la Figura 4.6, se presenta la ventana principal del sistema inteligente diseñado para el diagnóstico de fallas en generadores de potencia. Los campos de entrada abarcan 14 características, donde Enfriamiento y Tipo se presentan como menús desplegables para seleccionar opciones dentro de sus respectivas categorías. Las 12 variables restantes corresponden a factores de disipación y descargas parciales, y se requiere ingresar los valores del caso a predecir en los cuadros de texto designados.

Una vez que se ha seleccionado la categoría correspondiente para las primeras dos variables y se han ingresado valores numéricos para las otras doce, se debe hacer clic en el botón Enviar. De manera automática, junto a la etiqueta Diagnóstico, se revelará el número de la clase a la que pertenece la falla. Inmediato a la etiqueta Tipo de falla, se mostrará el nombre específico de la falla detectada en el generador. Asimismo, al lado de la etiqueta Recomendaciones, se visualizará la sugerencia correspondiente para abordar la falla identificada.



Enfriamiento	FD_tand_A	DP_VnCrt_Ap	Diagnóstico:
Enfriamiento_AIRE FORZADO			
	FD_tand_B	DP_VnCrt_An	Tipo de falla:
Tipo			Recomendaciones:
Tipo_375 MW / 17 KV	FD_tand_C	DP_VnCrt_Bp	
			Enviar
	FD_Dtand_A	DP_VnCrt_Bn	
	FD_Dtand_B	DP_VnCrt_Cp	
	FD_Dtand_C	DP_VnCrt_Cn	

Figura 4.6 Ventana principal de SIDG

En la Figura 4.7, se muestra la interfaz de la ventana del botón Acerca de SIDG, la cual brinda información detallada sobre el autor del sistema y sus colaboradores en el contexto de la tesis.

Acerca de SIDG

Tecnológico Nacional de México Campus Orizaba

Autor: Quetzalli Amaya Sánchez
 Director: Dr. Alberto Alfonso Aguilar Lasserre
 Codirector: M.I.I. Marco Julio Del Moral Argumedo

Instituto Nacional de Electricidad y Energías Limpias
 Asesor externo: Dr. Gustavo Arroyo Figueroa

Figura 4.7 Ventana Acerca de SIDG

En cuanto a las recomendaciones, la Tabla 4.3 detalla las comisiones específicas para abordar cada uno de los cinco tipos distintos de fallas.

Tabla 4.3 Recomendaciones para abordar las fallas en generadores de potencia

Clase	Recomendaciones
0	Monitoreo en línea o evaluación periódica anual. Si solo hay evidencia de afectación en la cita de protección contra humedad, realizar proceso de limpieza y aplicar barniz aislante para restituir el afectado.
1	Efectuar limpieza en la zona de graduación. Revisar y limpiar las zonas afectadas en el próximo mantenimiento. Sustituir cinta de protección contra humedad en zonas afectadas.
2	Reparación de las conexiones entre bobinas.
3	Si se registra incremento en magnitud de más de 30 % considerar cambio de devanado.
4	Reemplazo de devanado del estator. Rehabilitar devanado.

La Figura 4.8 muestra el menú desplegable correspondiente a la variable Enfriamiento, donde se pueden visualizar las cuatro categorías asociadas a esta variable.

Sistema Inteligente para el Diagnóstico de Fallas en Generadores de Potencia (SIDG)

Enfriamiento

Enfriamiento_AIRE FORZADO

Enfriamiento_AIRE

Enfriamiento_HIDRÓGENO

Enfriamiento_HIDRÓGENO hidrógeno a una presión de 1.2 kg/cm2

Enfriamiento_AIRE FORZADO

FD_tand_A

DP_VnCrT_Ap

DP_VnCrT_An

FD_tand_C

DP_VnCrT_Bp

FD_Dtand_A

DP_VnCrT_Bn

FD_Dtand_B

DP_VnCrT_Cp

FD_Dtand_C

DP_VnCrT_Cn

Diagnóstico:

Tipo de falla:

Recomendaciones:

Enviar

Acerca de SIDG

Figura 4.8 Menú desplegable correspondiente a Enfriamiento

La Figura 4.9 presenta el menú desplegable correspondiente a la variable "Tipo", en el cual se pueden observar las 41 categorías asociadas a esta variable.

Sistema Inteligente para el Diagnóstico de Fallas en Generadores de Potencia (SIDG)

Tipo_375 MVA / 17 KV
Tipo_208 MVA / 16 KV
Tipo_218 MVA / 15 KV
Tipo_153 MVA / 16 KV
Tipo_123 MVA / 16.5 KV
Tipo_341 MVA / 13.8 KV
Tipo_49 MVA / 13.8 KV
Tipo_269 MVA / 18 KV
Tipo_339 MVA / 18 KV
Tipo_212 MVA / 18 KV
Tipo_183 MVA / 13.8 KV
Tipo_337 MVA / 18 KV
Tipo_103 MVA / 11.5 KV
Tipo_365 MVA / 21 KV
Tipo_143 MVA / 13.8 KV
Tipo_389 MVA / 20 KV
Tipo_193 MVA / 13.8 KV
Tipo_187 MVA / 16 KV
Tipo_140 MVA / 13.8 KV
Tipo_113 MVA / 13.8 KV
Tipo_177 MVA / 16.5 KV
Tipo_236 MVA / 16.5 KV
Tipo_91.5 MVA / 13.8 KV
Tipo_207 MVA / 18 KV
Tipo_213 MVA / 16.5 KV
Tipo_110 MVA / 13.8 KV
Tipo_126 MVA / 13.8 KV
Tipo_188 MVA / 16.5 KV
Tipo_211 MVA / 16.5 KV
Tipo_40 MVA / 13.8 KV
Tipo_105 MVA / 13.8 KV
Tipo_33 MVA / 13.8 KV
Tipo_45 MVA / 13.8 KV
Tipo_37.5 MVA / 13.8 KV
Tipo_43 MVA / 11.5 KV
Tipo_51.6 MVA / 13.8 KV
Tipo_49 MVA / 11.5 KV
Tipo_50 MVA / 13.8 KV
Tipo_35 MVA / 13.8 KV
Tipo_60 MVA / 13.8 KV
Tipo_25 MVA / 13.8 KV

FD_tand_A DP_VnCrt_Ap Diagnóstico:
FD_tand_B DP_VnCrt_An Tipo de falla:
FD_tand_C DP_VnCrt_Bp Recomendaciones:
FD_Dtand_A DP_VnCrt_Bn Enviar
FD_Dtand_B DP_VnCrt_Cp
FD_Dtand_C DP_VnCrt_Cn

Figura 4.9 Menú desplegable correspondiente a Tipo

Finalmente, en la Figura 4.10 se muestra el diagnóstico de un caso perteneciente a la clase 4 en el conjunto de datos de predicción. Esto se realiza con el propósito de verificar la funcionalidad del sistema y confirmar la coherencia de los resultados obtenidos.

Sistema Inteligente para el Diagnóstico de Fallas en Generadores de Potencia (SIDG)

Enfriamiento: Enfriamiento_AIRE
Tipo: Tipo_188 MVA / 16.5 KV

FD_tand_A: 1.34 DP_VnCrt_Ap: 42
FD_tand_B: 1.22 DP_VnCrt_An: 36
FD_tand_C: 1.25 DP_VnCrt_Bp: 19
FD_Dtand_A: 2.2 DP_VnCrt_Bn: 20
FD_Dtand_B: 2.5 DP_VnCrt_Cp: 43
FD_Dtand_C: 2.6 DP_VnCrt_Cn: 31

Diagnóstico: 4
Tipo de falla: Descargas a la ranura (descargas entre la laminación del núcleo y la superficie de los bastones)
Recomendaciones: Reemplazo de devanado del estator. Rehabilitar devanado.

Enviar

Acerca de SIDG

Figura 4.10 Diagnóstico clase 4

La interfaz gráfica no solo cumple con los criterios de funcionamiento práctico, sino que también se alinea de manera coherente con las necesidades y expectativas de los usuarios finales. Este nivel de adaptabilidad y facilidad de uso contribuye de manera sustancial a la aplicabilidad práctica del modelo, especialmente en entornos operativos críticos como el sector eléctrico nacional.

4.2.3.1 Comparación del sistema con investigaciones similares

En esta sección, se lleva a cabo una comparación del sistema inteligente desarrollado en el marco de esta investigación. La comparativa se realiza con investigaciones afines que emplean algoritmos de ML para la detección de fallas en generadores de potencia. En la Tabla 4.4 se presenta un resumen consolidado de tres trabajos de investigación relevantes en relación con el presente estudio.

Esta investigación sobresale por varias razones. En primer lugar, destaca por la diversidad de características en la base de datos, proporcionando una visión completa del equipo. En segundo lugar, la elección del modelo Gradient Boosting mejora la precisión mediante la combinación de múltiples modelos más débiles, optimizando la eficiencia y contribuyendo a la robustez del sistema. Además, se destaca la amplitud y relevancia de las variables consideradas, especialmente la variable dependiente relacionada con distintos tipos de descargas parciales internas en el generador, lo que aumenta la capacidad diagnóstica del modelo. Por último, la investigación se distingue por su destacada tasa de exactitud en la clasificación de fallas en generadores de potencia, facilitando una evaluación más precisa y sofisticada de las condiciones operativas y de falla, respaldando así la efectividad del modelo propuesto.

Tabla 4.4 Comparación del presente estudio con otras investigaciones similares que aplican ML

Referencia	Aportación y técnica	Base de datos	Clasificación	Exactitud
(González, Piedrahita, & Castrillón, 2020)	Modelo Bayesiano J48 para predecir paros de máquinas generadoras	300 registros reales, 10 variables independientes referente a la operación de la máquina (temperatura, presión, caudal, clima, etc.) y una variable dependiente (paro de máquina)	Si: Disparo de máquina No: No disparo de máquina	94%
(Quintanilha, y otros, 2021)	Modelo Random Forest para detectar fallas en generadores de energía de anillo cerrado	Se generaron segmentos de datos de 128 muestras por cada una de las 18 señales asociadas a los distintos sensores y representadas como series de tiempo con 2000 muestras por segundo.	T01: Pérdida de detección de voltaje T02: Falla interna del regulador automático de voltaje (AVR) T03: AVR no funciona T04: Ganancia AVR defectuosa T05: Ganancia de derivación del AVR defectuosa T06: Pérdida parcial de detección de voltaje T07: Pérdida de detección de velocidad T08: Sobreproducción T09: Subproducción T10: Controlador de velocidad que no funciona T11: Controlador de velocidad defectuoso T12: Estante de combustible congelado	83%
(Cui, Weng, & Zhang, 2022)	Modelo k-NN de tres etapas para el diagnóstico de fallas en rodamientos de aerogeneradores	64 señales de vibración del rodamiento por cada condición operativa (cargas y velocidades de rotación) para cada modo de falla	Etapas 1 (100%) - Rodamientos normales - Rodamientos defectuosos Etapas 2 (80%) - Rodamientos de bola internas - Rodamientos de bolas externas Etapas 3 (90%) - Rodamiento con falla en la pista interna - Rodamiento con falla en la pista.	89%
Presente investigación, 2024	Sistema inteligente para el diagnóstico de fallas en generadores de potencia basado en un modelo Gradient Boosting	196 registros reales, 57 variables independientes referente al enfriamiento y tipo de generados, factores de disipación y descargas parciales; y una variable dependiente (diagnóstico)	0: Descargas parciales internas en huecos del aislamiento principal 1: Descargas en la zona de graduación de campo eléctrico en la interface de las pinturas conductora y semiconductor 2: Descargas entre bobinas en cabezales o entre cabezales y dedos de sujeción del núcleo 3: Descargas parciales por delaminación del aislamiento principal y conductor de cobre 4: Descargas a la ranura	93%

4.3 Conclusión

Este capítulo se inicia con la evaluación de las predicciones del modelo utilizando 50 nuevos datos, con el propósito de garantizar su eficacia y precisión en el diagnóstico de fallas en generadores de potencia. Para ello, se implementa el mismo proceso de preprocesamiento aplicado al conjunto de datos de entrenamiento y prueba, que comprende la conversión de variables categóricas a variables dummy y la normalización. El resultado de esta evaluación revela que el modelo alcanza una exactitud balanceada del 92%, indicando un rendimiento sólido en términos de la capacidad para clasificar correctamente las clases.

Dado el buen desempeño del modelo, se procede a la creación de la interfaz gráfica utilizando la biblioteca tkinter de Python. Este paso es fundamental para facilitar la interacción y uso práctico del modelo, proporcionando una herramienta accesible y eficiente para los usuarios finales. Además, la comparación de nuestro sistema con investigaciones similares que utilizan algoritmos de ML refuerza la contribución única y valiosa de esta investigación al ámbito del diagnóstico de fallas en generadores de potencia.

CONCLUSIONES

En el transcurso de esta investigación, se lograron avances significativos en la comprensión y abordaje de las problemáticas asociadas a las fallas en generadores de potencia en centrales hidroeléctricas y termoeléctricas del sistema eléctrico nacional (SEN). Los resultados obtenidos y los objetivos cumplidos reflejan la contribución sustancial de este estudio al campo de la gestión eficiente de sistemas eléctricos de potencia.

A continuación, se exponen las conclusiones más destacadas en relación con los objetivos establecidos para esta investigación.

✓ Desarrollo exitoso del modelo inteligente

Se logra diseñar y desarrollar con éxito un modelo inteligente basado en un algoritmo Gradient Boosting. Este modelo, enfocado en el análisis de descargas parciales y otras características críticas, se ha implementado con el propósito de diagnosticar fallas en generadores de potencia presentes en las centrales hidroeléctricas del sistema eléctrico nacional.

El modelo Gradient Boosting, configurado con hiperparámetros `max_depth=None`, `max_features='sqrt'`, `n_estimators=122`, `min_samples_split=22` y `random_state=42`, alcanza un rendimiento del 93%. Es importante destacar que las características más significativas y que contribuyen de manera significativa a la toma de decisiones del modelo son Enfriamiento_AIRE FORZADO, así como `FD_Dtand_B`, `FD_Dtand_A`, `FD_tand_B` y `FD_tand_C`.

✓ Generación y validación de la base de datos

La creación de una base de datos de fallas en generadores de potencia constituye una fase crucial en el desarrollo del modelo inteligente. Este proceso implica la normalización y la conversión de variables categóricas a representaciones numéricas, culminando en la generación de un archivo que almacena información de 196 registros reales. Estos registros están asociados con la variable Diagnóstico y 57 características, abarcando aspectos como el tipo de enfriamiento y voltaje del generador, factores de disipación y descargas parciales.

Es importante destacar que la base de datos presenta un desbalance, ya que se observa una diferencia significativa entre los casos de la clase mayoritaria y la minoritaria.

Así mismo, la exhaustiva validación realizada por expertos en generadores asegura la fiabilidad y pertinencia de los datos, aspectos fundamentales para el entrenamiento y evaluación del modelo inteligente. Este proceso consolida la utilidad de la base de datos en el desarrollo de un modelo preciso y confiable para el diagnóstico de fallas en generadores de potencia.

✓ Validación del desempeño del modelo inteligente

Para la validación del modelo Gradient Boosting, se aplicaron diversas métricas específicas destinadas a evaluar datos desbalanceados. El modelo exhibe una destacada capacidad para clasificar las fallas en generadores de potencia con una precisión balanceada del 93%. Igualmente, presenta una concordancia sustancial, reflejada en un índice Kappa de Cohen de 0.90, junto con una alta similitud, evidenciada en un índice de Jaccard de 0.86 y un coeficiente de correlación de Matthews de 0.91. La robustez del modelo se manifiesta en una especificidad notable, alcanzando valores de 1 para las clases mayoritarias y niveles relativamente altos para las clases minoritarias.

De la base de prueba, que comprende un total de 40 datos, la matriz de confusión revela que se han producido 3 elementos clasificados de manera incorrecta en las clases 1 y 2. Las curvas ROC y PR respaldan la solidez del modelo, exhibiendo AUC sólidos y un rendimiento elevado en precisión e identificación de ejemplos positivos. El macro promedio en la curva ROC del 96% y el micro promedio en la curva PR del 89%, subrayan la eficacia del modelo para separar las clases de manera efectiva.

Para validar el rendimiento sobresaliente del modelo, la evaluación de sus predicciones con datos nuevos muestra una precisión balanceada del 92%. Además, se observa un índice Kappa de Cohen de 0.89, un índice de Jaccard de 0.85 y un coeficiente de correlación de Matthews de 0.89. En cuanto a la especificidad, se destacan valores de 1 en la mayoría de las clases, indicando una capacidad significativa para evitar clasificaciones erróneas de casos negativos como positivos.

En relación con la base de datos nuevos, que abarca un total de 50 datos para su predicción, la matriz de confusión revela que se han producido 4 elementos clasificados de manera incorrecta en las clases 1 y 2. Por otro lado, las curvas ROC y PR exhiben valores de AUC altos, con un macro promedio en la curva ROC del 97% y un micro promedio en la curva PR del 89%.

En conjunto, estos resultados confirman la eficiencia del modelo Gradient Boosting en la clasificación de fallas en generadores de potencia, consolidándolo como una herramienta confiable

y precisa para la detección y diagnóstico de problemas en este contexto. Estos hallazgos refuerzan la confianza en la aplicabilidad del modelo en entornos del mundo real, proporcionando una herramienta valiosa para el diagnóstico preciso de fallas en generadores de potencia.

✓ Eficiencia del modelo inteligente

Al comparar el sistema inteligente desarrollado en esta investigación con trabajos afines que emplean algoritmos de Machine Learning para la detección de fallas en generadores de potencia, se destacan diversas razones que respaldan la singularidad y eficacia de este enfoque. La diversidad de características en la base de datos emerge como un factor distintivo, proporcionando una visión integral del equipo. La elección estratégica del modelo Gradient Boosting constituye otro aspecto fundamental, mejorando la precisión mediante la combinación de múltiples modelos más débiles, lo que optimiza la eficiencia y contribuye a la robustez del sistema.

La amplitud y relevancia de las variables consideradas, especialmente aquellas vinculadas a distintos tipos de descargas parciales internas en el generador, amplifican la capacidad diagnóstica del modelo, posibilitando una clasificación detallada y específica de posibles problemas. La destacada tasa de exactitud en la clasificación de fallas en generadores de potencia no solo valida el rendimiento del modelo, sino que también proporciona una herramienta que facilita una evaluación más precisa y sofisticada de las condiciones operativas y de falla.

✓ Creación de un sistema interactivo

La exitosa implementación del modelo ha culminado en la creación de una aplicación de escritorio denominada Sistema Inteligente para el Diagnóstico de Fallas en Generadores de Potencia (SIDG). Esta herramienta simplifica la interacción con el modelo, proporcionando una solución accesible y eficiente para los usuarios finales.

Este trabajo representa un avance significativo en la mejora de la confiabilidad y eficiencia en la gestión de generadores de potencia. La combinación de técnicas de diagnóstico y la aplicación de algoritmos de Machine Learning ofrece un enfoque prometedor para abordar los desafíos asociados a las fallas en generadores. La implementación práctica del SIDG se destaca como una herramienta valiosa que puede tener un impacto positivo en la operación y mantenimiento de generadores de potencia, contribuyendo así al avance continuo de la industria eléctrica.

RECOMENDACIONES

Es esencial reconocer las limitaciones inherentes a esta investigación con el propósito de perfeccionar el modelo inteligente de diagnóstico de fallas en generadores de potencia en investigaciones futuras. A continuación, se presentan las recomendaciones correspondientes.

- **Ampliar la base de datos:** Se recomienda ampliar la base de datos incluyendo muchos más casos reales en cada categoría de las variables Enfriamiento y Tipo, así como en las cinco clases diagnóstico. Dado que la actual base de datos presenta desbalance, esta expansión contribuirá a un mejor aprendizaje del modelo y, por ende, a resultados de predicción más precisos.
- **Validar el conjunto de datos nuevos:** Se sugiere realizar una validación exhaustiva en el conjunto de datos nuevos, ya que podría contener anomalías que podrían haber afectado el rendimiento del modelo en la predicción. La verificación de la integridad de la base de datos de predicción es esencial para asegurar la confiabilidad de los resultados.
- **Validar en condiciones de información incompleta o errónea:** Esta recomendación busca evaluar la robustez del sistema ante posibles escenarios del mundo real donde la disponibilidad de datos puede ser limitada o su calidad comprometida. La capacidad del modelo para lidiar con información incompleta o incorrecta es crucial para garantizar su aplicabilidad y eficacia en entornos prácticos, reforzando así su utilidad en situaciones más desafiantes.
- **Incorporar variables en tiempo real:** Para futuras investigaciones, se sugiere explorar la incorporación de más variables y datos en tiempo real. Esto puede potenciar aún más la capacidad predictiva del modelo, reduciendo la dependencia de la calidad de los datos y la experiencia de los expertos en la validación de los mismos.
- **Promover la aplicación del Sistema Inteligente para el Diagnóstico de Fallas en Generadores de Potencia (SIDG)** por operadores expertos de generadores de potencia del SEN para que sirva de apoyo a la toma de decisiones de mantenimiento de estos activos, se valide y se generen nuevos casos.

Estas recomendaciones buscan fortalecer y mejorar la robustez del modelo de diagnóstico de fallas en generadores de potencia, preparándolo para enfrentar desafíos y escenarios más complejos en futuras investigaciones.

REFERENCIAS BIBLIOGRÁFICAS

- Altaf, S., Al-Anbuky, A., & GholamHosseini, H. (2014). Fault diagnosis in a distributed motor network using Artificial Neural Network. *International Symposium on Power Electronics, Electrical Drives, Automation and Motion*, 190-197. Recuperado de <https://api.semanticscholar.org/CorpusID:15457795>.
- Amat Rodrigo, J. (Septiembre de 2023). *Ciencia de datos*. Obtenido de Gradient Boosting con Python: https://www.cienciadedatos.net/documentos/py09_gradient_boosting_python.html
- ArcGISPro. (2023). *ArcGIS Pro*. Obtenido de Cómo funciona el algoritmo XGBoost: [https://pro.arcgis.com/es/pro-app/latest/tool-reference/geoai/how-xgboost-works.htm#:~:text=XGBoost%20es%20un%20m%C3%A9todo%20de,\(refuerzo%20de%20gradientes%20extremo\)](https://pro.arcgis.com/es/pro-app/latest/tool-reference/geoai/how-xgboost-works.htm#:~:text=XGBoost%20es%20un%20m%C3%A9todo%20de,(refuerzo%20de%20gradientes%20extremo)).
- Bruetsch, R., Tari, M., Froehlich, K., Weiers, T., & Vogelsang, R. (2008). Insulation Failure Mechanisms of Power Generators. *IEEE Electrical Insulation Magazine*, Recuperado de <https://api.semanticscholar.org/CorpusID:11894658>.
- Cadena, J., Cadena, J., & Pérez Londoño, S. (2008). Aplicación de redes neuronales probabilísticas en la detección de fallas incipientes en transformadores. *Scientia Et Technica*, XIV(39), ISSN: 0122-1701, 48-53. Recuperado de <https://www.redalyc.org/articulo.oa?id=84920503010>.
- Campus MVP. (21 de Abril de 2021). Obtenido de Uso de la inteligencia artificial en la industria 4.0: mantenimiento predictivo: <https://www.campusmvp.es/recursos/post/ejemplo-de-uso-de-la-inteligencia-artificial-en-la-industria-4-0-mantenimiento-predictivo.aspx>
- Campuzano Martínez, I. R. (2016). Diagnóstico de generadores eléctricos de potencia con técnicas de monitoreo en línea y fuera de línea. *Información tecnológica*, 27(2), 11-20. Recuperado de <https://dx.doi.org/10.4067/S0718-07642016000200003>.
- Castillo, M., & Gazcón, F. (2017). Tipología de Fallas en Generadores Síncronos Hidráulicos de la Empresa de Energía del Pacífico. , *Trabajo de Grado, Universidad del Valle, Facultad de Ingeniería, Escuela de Ingeniería Eléctrica y Electrónica, Santiago de Cali*, Recuperado

de <https://bibliotecadigital.univalle.edu.co/entities/publication/4925bb94-bf81-4b2e-9130-d138e5df077d>.

Cerón, A. F., Echeverry, D. F., Aponte, G., & Romero, A. A. (2015). Índice de Salud para Transformadores de Potencia Inmersos en Aceite Mineral con Voltajes entre 69kV y 230kV usando Lógica Difusa. *Información tecnológica*, 26(2), 107-116. Recuperado de <https://dx.doi.org/10.4067/S0718-07642015000200013>.

Cerón, A., Echeverry, D., Aponte, G., & Romero, A. (2015). Índice de Salud para Transformadores de Potencia Inmersos en Aceite Mineral con Voltajes entre 69kV y 230kV usando Lógica Difusa. *Información Tecnológica*, Vol 26 (2), 107-116.

CFE. (2023). Obtenido de Historia de la CFE: <https://www.cfe.mx/nuestraempresa/pages/historia.aspx>

CFE VF000-72. (2012). Unidad de control de excitación para modernización de sistemas de excitación tipo brushless. *Especificación CFE VF000-72*, Recuperado de https://lapem.cfe.gob.mx/normas/carga_pagina.asp?pag=VF000-72.pdf.

CFE W4101-16. (2014). Sistema de excitación estático para generadores síncronos de centrales eléctricas. *Especificación CFE W4101-16*, Recuperado de https://lapem.cfe.gob.mx/normas/carga_pagina.asp?pag=W4101-16.pdf.

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (págs. 785-794). Recuperado de <https://dl.acm.org/doi/pdf/10.1145/2939672.2939785>.

CIGRE. (2024). Obtenido de <https://www.cigre.org.mx/>

Cinvestav. (27 de 10 de 2022). Obtenido de Sistemas Eléctricos de Potencia: <https://www.cinvestav.mx/Departamentos/Sistemas-Electricos-de-Potencia>

Cui, B., Weng, Y., & Zhang, N. (2022). A feature extraction and machine learning framework for bearing fault diagnosis. *Renewable Energy*, Volume 191, Pages 987-997. ISSN 0960-1481. Recuperado de <https://www.sciencedirect.com/science/article/pii/S0960148122005225>.

Díaz Ramírez, J. (2021). Aprendizaje Automático y Aprendizaje Profundo. *Ingeniare. Revista chilena de ingeniería*, vol. 29 N°2, 182-183. Recuperado de <https://www.scielo.cl/pdf/ingeniare/v29n2/0718-3305-ingeniare-29-02-180.pdf>.

- Enríquez Harper, G. (2004). *El libro práctico de los generadores, transformadores y motores eléctricos*. México: LIMUSA. Obtenido de https://books.google.es/books?hl=es&lr=&id=P62ebMavSIIC&oi=fnd&pg=PA1&dq=generadores+el%C3%A9ctricos&ots=HyZRHBcf6H&sig=MndTnziO8pAFezi4c0lLgu_MRfU#v=onepage&q&f=false
- Espinoza Zúñiga, J. J. (2020). Aplicación de algoritmos Random Forest y XGBoost en una base de solicitudes de tarjetas de crédito. *Ingeniería, investigación y tecnología*, 21(3), 00002., Recuperado de https://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S1405-77432020000300002.
- Flores, R., & Asiaín, T. I. (2011). Diagnóstico de fallas en máquinas eléctricas rotatorias utilizando la técnica de espectros de frecuencia de bandas laterales. *Información tecnológica*, 22(4), 73-84. Recuperado de <https://dx.doi.org/10.4067/S0718-07642011000400009>.
- Galaz, J. (1996). Un aspecto de mantenimiento poco difundido. El circuito magnético (núcleo). *RVP'96-GEN-01. Reunión de Verano de Potencia, Acapulco, México*, Recuperado de <https://www.redalyc.org/pdf/3374/337450992001.pdf>.
- Galván Barrera, R. P. (2010). Criterios de análisis de patrones gráficos de las mediciones de descargas parciales en línea para diagnóstico del devanado estáticos en generadores hidroeléctricos. *Facultas de Ingeniería Eléctrica y Electrónica*, 28-33. Recuperado de <https://repositorio.uncp.edu.pe/bitstream/handle/20.500.12894/3591/Galvan%20Barrera.pdf?sequence=1&isAllowed=y>.
- García García, J. A., Reding Bernal, A., & López Alvarenga, J. C. (2013). Cálculo del tamaño de la muestra en investigación en educación médica. *Investigación en educación médica*, 2(8), 217-224. Recuperado de http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S2007-50572013000400007&lng=es&tlng=es.
- García Jiménez, V. (2010). *Distribuciones de clases no balanceadas: Métricas, análisis de complejidad y algoritmos de aprendizaje*. Catellón. Recuperado de <https://www.tdx.cat/handle/10803/10491#page=1>: Tesis doctoral.
- García, A., Reyes, O., Robles, E., Esparza, M., López, J., Medina, A., & Uribe, M. (1998). Mecanismos de Deterioro de Generadores Eléctricos. *Boletín IIE*.

- Generac*. (3 de Junio de 2020). Obtenido de Planta de emergencia: tipos y características de estos generadores: <https://blog.generaclatam.com/planta-de-emergencia>
- Gensheng, Q., & Jingquan, L. (2022). A comparative study of deep learning-based fault diagnosis methods for rotating machines in nuclear power plants. *Annals of Nuclear Energy, Volume 178*, 109334, ISSN 0306-4549, Recuperado de <https://doi.org/10.1016/j.anucene.2022.109334>.
- Géron, A. (2019). *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow*. Sebastopol, CA 95472: O'Reilly Media, Inc. ISBN: 9781491962299.
- González, G., De Angelo, C., Forchetti, D., & Aligia, D. (2018). Diagnóstico de fallas en el convertidor del rotor en generadores de inducción con rotor bobinado. *Revista Iberoamericana De Automática E Informática Industrial*, 15(3), 297-308. Recuperado de <https://doi.org/10.4995/riai.2017.9042>.
- González, H., Piedrahita, J., & Castrillón, O. (2020). Predicción de parada de máquinas generadoras en una central hidroeléctrica por medio de minería de datos. *Información tecnológica*, 31(5), 215-222. Recuperado de https://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0718-07642020000500215&lng=en&nrm=iso&tlng=en.
- Gross, D. W., & Soller, M. (s.f.). Diagnóstico y monitoreo de descargas parciales en generadores y motores. *Electrical Testing Group*, Recuperado de https://www.inducor.com.ar/articulostecnicos/diagnostico_y_monitoreo_de_descargas_parciales_en_generadores_y_motores.pdf.
- Henao, H., Capolino, G. A., Fernández, M., Filippetti, F., Bruzzese, C., Strangas, E., . . . Hedayati, S. (2013). Trends in Fault Diagnosis for Electrical Machines. *Industrial Electronics Magazine*, 1-27. Recuperado de <https://riunet.upv.es/handle/10251/98423>.
- Hong-Bae, J. (2022). A Review on the Advanced Maintenance Approach for Achieving the Zero-Defect Manufacturing System. *Frontiers in Manufacturing Technology*. 2. 9209900., 1-15. Recuperado de https://www.researchgate.net/publication/361335418_A_Review_on_the_Advanced_Maintenance_Approach_for_Achieving_the_Zero-Defect_Manufacturing_System.

- Hurtado Cortés, L., Villarreal López, E., & Villarreal López, L. (2016). Detección y diagnóstico de fallas mediante técnicas de inteligencia artificial, un estado del arte. *Dyna*, 83 (199). ISSN: 0012-7353, 19-28. Recuperado de <https://www.redalyc.org/articulo.oa?id=49648868002>.
- Hweon Ki, J., Song Hyun, K., & Chang Lak, K. (2022). Proposal of a new method for learning of diesel generator sounds and detecting abnormal sounds using an unsupervised deep learning algorithm. *Nuclear Engineering and Technology*, ISSN 1738-5733, Recuperado de <https://doi.org/10.1016/j.net.2022.10.019>.
- IBM. (2023). IBM. Obtenido de What is random forest?: <https://www.ibm.com/topics/random-forest>
- IEEE. (2022). Obtenido de <https://phmconf.org/>
- Instituto Nacional de Electricidad y Energías Limpias. (2023). Obtenido de Gobierno de México: <https://www.gob.mx/ineel/que-hacemos>
- ISO 10816-1. (1995). Mechanical vibration -Evaluation of machine vibration by measurements on non-rotating parts, Part 1 General guidelines. Recuperado de <https://www.iso.org/obp/ui/#iso:std:iso:10816:-1:ed-1:v1:en>.
- ISO 10816-5. (2000). Mechanical vibration -Evaluation of machine vibration by measurements on non-rotating parts, Part 5 Machine sets in hydraulic power generating/pumping and pumping plants. Recuperado de <https://www.iso.org/obp/ui/#iso:std:iso:10816:-5:ed-1:v1:en>.
- Japkowicz, N., & Shah, M. (2011). *Evaluating Learning Algorithms A Classification Perspective*. New York: Cambridge University Press. Recuperado de <https://api.semanticscholar.org/CorpusID:31015609>.
- Jürgensen, J., Godin, A., & Hilber, P. (2017). Health Index as Condition Estimator for Power System Equipment: a Critical Discussion and Case Study. *24th International Conference & Exhibition on Electricity Distribution (CIRED)*, 12-15. Recuperado de <https://pdfs.semanticscholar.org/16d5/0366fea467b55a61565def7d184a4ee473db.pdf>.

- Kelleher, J., Mac Name, B., & D'Arcy, A. (2015). *Fundamentals of Machine Learning for Predictive Data Analytics. Algorithms, Worked Examples, and Case Studies*. London, England: Massachusetts Institute of Technology.
- Klempner, G., & Kerszenbaum, I. (2004). Operation and Maintenance of Large Turbo-generators. *IEEE Press Series on Power Engineering, Wiley-Inter-science, A John Wiley & Sons, Inc. Publications*.
- Kruger, W. (2009). Detectando Fallas en Rodamientos Utilizando Métodos de Pruebas Eléctricos y Mecánicos de Vibraciones. Recuperado de https://www.academia.edu/16699190/Detectando_fallas_en_rodamientos_utilizando_metodos_electrico_y_mecanico.
- Kumar, R. (23 de Julio de 2021). *Nerd For Tech*. Obtenido de Random Forest: <https://medium.com/nerd-for-tech/random-forest-sturdy-algorithm-d60b9f9140d4>
- Letal, J., Sedding, H., & Sasic, M. (2018). *Importance of Operating Parameters when Assessing the Condition of Machines On-Line*. Obtenido de CIGRE: <http://www.cigre.org>, CIGRE
- Li, C., Yu, R., Yu, W., & Wang, T. (2022). Fault-tolerant control system for once-through steam generator based on reinforcement learning algorithm. *Nuclear Engineering and Technology, Volume 54, Issue 9, ISSN 1738-5733, 3283-3292*. Recuperado de <https://doi.org/10.1016/j.net.2022.04.014>.
- Lima Álvarez, M., Ramírez Fiallo, O., Fernández García, S., & Pérez Barrios, R. J. (2011). Sistema para diagnóstico de fallas en transformadores a través de lógica difusa. *Revista Científica de Ingeniería Energética. Vol. XXIV, No. 3/2003, 34-41*. Recuperado de https://www.researchgate.net/publication/266484172_Sistema_para_diagnostico_de_fallas_en_transformadores_a_traves_de_logica_difusa.
- Medrano Hurtado, Z. Y., Pérez Tello, C., Gómez Sarduy, J., & Vera Pérez, M. (2016). Nueva metodología de diagnóstico de fallas en rodamientos en una máquina síncrona mediante el procesamiento de señales vibro-acústicas empleando análisis de densidad de potencia. *Ingeniería, Investigación y Tecnología, Volume 17, Issue 1, ISSN 1405-7743, 73-85*. Recuperado de <https://doi.org/10.1016/j.riit.2016.01.007>.

- Melih, Y., Veysi, B., Hakan, D., & Muhammet, G. (2022). An interval type-2 fuzzy enhanced best–worst method for the evaluation of ship diesel generator failures. *Engineering Failure Analysis*, Volume 138, 106428, ISSN 1350-6307, Recuperado de <https://doi.org/10.1016/j.engfailanal.2022.106428>.
- Moreno, A., Armengol, E., Béjar, J., Belanche, L., Cortés, U., Gavalda, R., . . . Sánchez, M. (1998). Aprendizaje automático. *Edicions UPC*, 6-15. Recuperado de <https://upcommons.upc.edu/bitstream/handle/2099.3/36157/9788483019962.pdf?sequence=1&isAllowed=y>.
- Murad, A., Zakiud, D., Evgeny, S., Khalid, M. C., Ahmad H., M., & Zhiyuan, C. (2021). Open switch fault diagnosis of cascade H-bridge multi-level inverter in distributed power generators by machine learning algorithms. *Energy Reports*, Volume 7, ISSN 2352-4847, 8929-8942. Recuperado de <https://doi.org/10.1016/j.egyr.2021.11.058>.
- Nandi, S., Toliyat, H., & Li, X. (2006). Condition monitoring and fault diagnosis of electrical motors—A review. *Energy Conversion, IEEE Transactions on*. 20., 719-729. Recuperado de https://www.researchgate.net/publication/3270486_Condition_Monitoring_and_Fault_Diagnosis_of_Electrical_Motors-A_Review/citation/download.
- Norman, A. T. (2019). *Aprendizaje automático en acción*. Tektime. Recuperado de: <https://www.perlego.com/book/1373133/aprendizaje-automtico-en-accin-un-libro-para-el-lego-gua-paso-a-paso-para-los-novatos-pdf>.
- Parraga Acosta, J. C. (2019). El índice de salud y su influencia en la gestión del ciclo de vida de los transformadores del sistema eléctrico de la Mina Cobriza. *Universidad Nacional del Centro de Perú UNCP*, 13-14. Recuperado de <https://repositorio.uncp.edu.pe/bitstream/handle/20.500.12894/5174/Parraga%20Acosta.pdf?sequence=1&isAllowed=y>.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 2825-2830. Recuperado de: <https://scikit-learn.org/stable/modules/tree.html#tree-algorithms>.

- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12, 2825-2830. Recuperado de <https://scikit-learn.org/stable/index.html>.
- Ponce, P. (2010). *Inteligencia artificial con aplicaciones a la ingeniería*. México: Alfaomega Grupo Editor, S. A. de C. V. Recuperado de <https://content.e-bookshelf.de/media/reading/L-18426071-262e96cc3a.pdf>.
- Pranupa, S., Kiran Kurmar, B., & Nagaraja, S. (2018). Detection of Rotating Diode Failure Condition & its Protection in Brushless Alternator. *International Journal of Engineering*, Recuperado de <https://api.semanticscholar.org/CorpusID:150017778>.
- QuimiNet. (2021). Obtenido de Precios de mantenimiento de generadores eléctricos: <https://www.quiminet.com/productos/mantenimiento-de-generadores-electricos-87140771412/precios.htm>
- Quintanilha, I. M., Elias, V. R., da Silva, F. B., Fonini, P. A., da Silva, E. A., Netto, S. L., . . . Andersen, R. B. (2021). A fault detector/classifier for closed-ring power generators using machine learning. *Reliability Engineering & System Safety*, Volume 212, 107614, ISSN 0951-8320, Recuperado de <https://doi.org/10.1016/j.ress.2021.107614>.
- Ramirez Mongui, J. J., & Dávila Arias, J. Y. (2022). Hacia el mantenimiento basado en condición de los dispositivos inteligentes en la industria 4.0. Recuperado de https://cimga.com/2022/Trabajos/012_TRA_COL_J_RAMIREZ_CIMGA2022.pdf.
- Ruíz Mondragón, B., Mora Flórez, J., & Pérez Londoño, S. (2009). Modelado del generador con aterramiento de alta reactancia para el análisis de fallas del estator. *Scientia Et Technica*, XV(42), ISSN: 0122-1701, 333-338. Recuperado de <https://www.redalyc.org/articulo.oa?Id=84916714062>.
- Sandoval, L. J. (2018). Algoritmos de aprendizaje automático para análisis y predicción de datos. *ITCA FEPADE*, 36-40. Recuperado de http://redicces.org.sv/jspui/bitstream/10972/3626/1/Art6_RT2018.pdf.
- Sheppard, J., Kaufman, M., & Wilmer, T. (2009). IEEE Standards for Prognostics and Health Management. In *IEEE Aerospace and Electronic Systems Magazine*, vol. 24, no. 9. DOI:

10.1109/MAES.2009.5282287, 34-41. Recuperado de <https://ieeexplore.ieee.org/document/5282287>.

Slomovitz, D., Trigo, L., & Fernández, D. J. (2000). Nuevas tecnologías estadísticas para la supervisión del estado de generadores usando medición de descargas parciales. *In Congreso CIER 2000, Buenos Aires, Argentina*, 1-5. Recuperado de <https://www.colibri.udelar.edu.uy/jspui/bitstream/20.500.12008/30039/1/STF00.pdf>.

Strategy. (28 de Julio de 2020). Obtenido de Salud de activos: una estrategia para la gestión óptima de transformadores de potencia y estimación de su vida útil: <https://www.strategy.com.co/salud-de-activos-gestion-optima/>

Tayner, P., Ran, L., Penman, J., & Sedding, H. (2008). Condition Monitoring of Rotating Electrical. Recuperado de <https://api.semanticscholar.org/CorpusID:106935756>.

Xiao, P. (2022). *Artificial Intelligence Programming with Python: From Zero to Hero*. New Jersey: Wiley.

Yang, B., Li, F., Ye, J., & Song, W. (2019). Condition Monitoring and Fault Diagnosis of Generators in Power Networks. *IEEE Power & Energy Society General Meeting (PESGM)*, 1-5. DOI: 10.1109/PESGM40551.2019.8973427. Recuperado de https://sensorweb.engr.uga.edu/wp-content/uploads/2019/02/Dropbox_yang2019condition.pdf.

Žarković, M., & Stojković, Z. (2017). Analysis of artificial intelligence expert systems for power transformer condition monitoring and diagnostics. *Electric Power Systems Research*, Volume 149, ISSN 0378-7796, 125-136. Recuperado de <https://doi.org/10.1016/j.epsr.2017.04.025>.