Application of CRISP-DM Methodology in the Analysis of Dissolved Gases in Dielectric oil of Electrical Transformers in the Ecuadorian Electrical Sector

Aplicación de la Metodología CRISP-DM en el Análisis de Gases Disueltos en Aceite Dieléctrico de Transformadores Eléctricos del Sector Eléctrico Ecuatoriano

C.A. Molina[1] 0009-0009-2655-0813 F.V. Bonilla[2] 0000-0001-6542-9666

[1]*Ponticifia Universidad Católica del Ecuador, Quito, Ecuador*
E-mail: camolinab@puce.edu.ec
[2]*Universidad Internacional del Ecuador, Quito, Ecuador*
E-mail: febonillave@uide.edu.ec

Abstract

This study addresses the application of the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology in the analysis of dissolved gases in oil of power transformers, being this a critical component in electrical systems. The adoption of this six-phase structured method allowed a comprehensive evaluation of the condition of the transformer units of the Ecuadorian electrical system based on the analysis of investment and expansion data of the sector, as well as the study of 1 099 DGA (Dissolved Gas Analysis) profiles obtained from a population of 153 transformers located in the different regions of continental Ecuador. The findings described in this work have the potential to significantly improve investment and maintenance strategies and policies. In addition, the adoption of automation techniques in the DGA classification process is proposed, using supervised learning models to enhance the reliability and efficiency of the public energy service. The results suggest that this approach not only improves the diagnosis within the maintenance activities, but also provides a solid basis to draw a roadmap towards a predictive asset management, resulting in a substantial improvement of the reliability of the national power system.

Index terms— Machine Learning, Random Forest, CRISM-DM, Data Science, EDA, ETL, DGA, PCA, Overfitting, Transformer.

Resumen

Este estudio aborda la aplicación de la metodología Cross-Industry Standard Process for Data Mining (CRISP-DM) en el análisis de los gases disueltos en aceite de transformadores de potencia, siendo este un componente crítico en los sistemas eléctricos. La adopción de este método estructurado de seis fases permitió evaluar de forma integral la condición de las unidades de transformación del sistema eléctrico ecuatoriano a partir del análisis de datos de inversión y expansión del sector, así como del estudio de 1 099 perfiles DGA (Dissolved Gas Analysis) obtenidos de una población de 153 transformadores ubicados en las distintas regiones de Ecuador continental. Los hallazgos descritos en este trabajo tienen el potencial de mejorar significativamente las estrategias y políticas de inversión y mantenimiento. Además, se propone la adopción de técnicas de automatización en el proceso de clasificación DGA, utilizando modelos de aprendizaje supervisado para potenciar la confiabilidad y eficiencia del servicio público de energía. Los resultados sugieren que este enfoque no solo mejora el diagnóstico dentro de las actividades de mantenimiento, sino que también proporciona una base sólida para trazar una hoja de ruta hacia una gestión predictiva de los activos, lo que se traduce en una mejora sustancial de la confiabilidad del sistema eléctrico nacional.

Palabras clave— Aprendizaje de Máquina, Bosques Aleatorios, CRISM-DM, Ciencia de Datos, EDA, ETL, DGA, PCA, Sobreajuste, Transformador.

Recibido: 21-04-2024, Aprobado tras revisión: 13-06-2024
Forma sugerida de citación: Molina, C.; Bonilla, V. (2024). "Aplicación de la metodología CRISP-DM en el análisis de gases disueltos en aceite dieléctrico de transformadores eléctricos del sector eléctrico ecuatoriano". Revista Técnica ***"energía"***. No. 21, Issue I, Pp. 12-21
ISSN On-line: 2602-8492 - ISSN Impreso: 1390-5074
Doi: https://doi.org/10.37116/revistaenergia.v21.n1.2024.635

1. INTRODUCCIÓN

En el período 2014 – 2023, los sistemas eléctricos de potencia de Ecuador se han caracterizado por una notable expansión, y con ello, la necesidad de incrementar su capacidad de potencia de transformación en subestaciones eléctricas y patios de elevación [1], [2]. Con esta expansión también se observó la introducción de unidades de transformación provenientes de fabricantes poco conocidos en la infraestructura del sistema eléctrico ecuatoriano, en comparación con equipos previamente establecidos, suscitando la necesidad de realizar un análisis más profundo que permitan verificar la calidad de manufactura y de los materiales empleados durante su construcción. Frente a esta situación y considerando la velocidad, volumen, variedad, veracidad y valor de los datos obtenidos a partir de la concentración de gases disueltos en el aceite aislante para el diagnóstico de transformadores eléctricos, son cada vez más crecientes los estudios que apuntan a la adopción de tecnologías computacionales como el Machine Learning para dar solución al procesamiento de información. En este marco, varios investigadores han demostrado de manera efectiva la precisión y eficacia de modelos de clasificación automática para la detección de la condición de transformadores eléctricos, empleando algoritmos de aprendizaje supervisado [3] - [7].

Con base en el estado del arte en relación con la aplicación de técnicas de Machine Learning para el diagnóstico en transformadores eléctricos, en [3] se propone el uso del algoritmo de bosques aleatorios (Random Forest) como alternativa al análisis de gases disueltos (DGA). Este modelo previamente entrenado y validado a partir de 128 muestras de aceite, es capaz de clasificar cuatro categorías de diagnóstico distintas: con un 100% para el diagnóstico de descarga de alta energía, un 77% para descarga de baja energía, un 60% para estado normal y un 97% para el estado de sobrecalentamiento. En [4], se implementa una estrategia de preprocesamiento basada en la técnica de bootstrap con el propósito de mitigar el desequilibrio intrínseco de las clases en el conjunto de datos, facilitando así una evaluación más equilibrada de los modelos predictivos. Posteriormente, se procede con la aplicación de algoritmos de programación genética para identificar y extraer las características más significativas. En la fase final de este estudio, fueron entrenados y evaluados tres distintos modelos de clasificación: Redes Neuronales Artificiales (ANN), Máquinas de Vectores de Soporte (SVM) y Vecinos más Cercanos (KNN). Por otro lado, los estudios propuestos en [5], [6], que se distinguen por alcanzar elevados índices de precisión en la detección y clasificación de diagnósticos en transformadores eléctricos, acusan sus resultados a la capacidad de los algoritmos SVM para procesar datos dinámicos y no lineales, dada la habilidad que tienen estos modelos para ser entrenados utilizando hiperplanos de separación de gran margen.

Sin embargo, a pesar de los progresos significativos en la incorporación de técnicas de aprendizaje automático para la evaluación del estado operativo de transformadores eléctricos, se identifican limitaciones críticas en la literatura, tales como la carencia de una metodología investigativa específicamente delineada para el análisis sistemático de datos, el sobreajuste de los modelos, así como, la insuficiencia y la heterogeneidad de los datos disponibles para el modelamiento y validación. Este conjunto de condiciones resalta la necesidad de un enfoque más estructurado y riguroso en la investigación, que no solo aborde la diversificación y representación de los datos para mejorar las métricas de los modelos, sino que también establezca un marco metodológico claro y específico para la exploración y análisis de datos en el dominio de la ingeniería eléctrica en el sector.

En este contexto, en el estado del arte en torno a la ciencia de datos señala que la implementación de técnicas de aprendizaje automático, sin la adhesión de metodologías estructuradas como CRISP-DM (acrónimo de Cross-Industry Standard Process for Data Mining), restringe significativamente la eficacia del Machine Learning en ofrecer análisis detallados y soluciones dedicadas [8] - [12]. Bajo este enfoque, es posible no solo mejorar la precisión y relevancia de los modelos de clasificación, sino también facilitar el desarrollo de estrategias innovadoras que puedan contribuir a la mejora de procesos clave, como la automatización del diagnóstico y planificación de tareas de mantenimiento.

Este estudio propone una metodología alineada con los principios y avances de la cuarta revolución industrial, aplicado al DGA, que se fundamenta en la implementación de las seis etapas especificadas por la metodología CRISP-DM, cuyo enfoque se centra en la extracción de conocimientos significativos relacionados con variables diagnósticas pertinentes a uno de los componentes más críticos y de mayor valor económico dentro de los sistemas eléctricos de potencia. Con esta orientación no solo se aspira a mejorar la confiabilidad y la vida útil de los transformadores eléctricos mediante un mantenimiento predictivo más efectivo, sino también a optimizar los recursos y la gestión de activos dentro de la infraestructura eléctrica ecuatoriana, promoviendo así una mayor eficiencia y sostenibilidad en el sector. En el contexto de la actual situación energética en Ecuador, este estudio propone estrategias orientadas a reformular los criterios de adquisición de infraestructura eléctrica, encaminadas a mejorar la rentabilidad a largo plazo de las inversiones. Finalmente, este trabajo plantea la adopción de sistemas de automatización en actividades de diagnóstico, con el objeto de mejorar los tiempos de respuesta en el análisis de resultados, acciones que apuestan a mejorar las estrategias de mantenimiento y la confiabilidad del servicio público de energía eléctrica.

La Sección 2 presenta la importancia de la aplicación de la metodología CRISP-DM en el análisis de gases disueltos. En la Sección 3 se detalla la aplicación de la metodología propuesta para el análisis de la condición de las unidades de transformación de potencia del sistema eléctrico ecuatoriano. La Sección 4 discute los resultados obtenidos. Finalmente, las conclusiones y recomendaciones se presentan en la Sección 5.

2. ANÁLISIS DE GASES DISUELTOS DGA Y LA METODOLOGÍA CRISP-DM

En la década de 1960, se encontró que la presencia de hidrógeno en aceites dieléctricos bajo estrés indicaba fallos en transformadores (ver Tabla 1), originando el método DGA, ahora un enfoque confiable y económico para identificar varias anomalías [13]. Avances científicos y tecnológicos han refinado este método, permitiendo clasificar las fallas en siete categorías esenciales [14].

Tabla 1: Clasificación de fallas en transformadores

ETIQUETA	DESCRIPCIÓN DEL TIPO DE FALLA
PD	Descargas Parciales
T1	Falla Térmica T < 300ºC
T2	Falla Térmica 300ºC < T < 700ºC
T3	Falla Térmica T > 700ºC
DT	Falla Térmica y Eléctrica
D1	Falla Eléctrica de Bajo Nivel
D2	Falla Eléctrica de Alto Nivel

Para la determinación de categorías de fallos en transformadores eléctricos, las guías técnicas de mantenimiento estándar recurren a los criterios de validación establecidos en las publicaciones y guías establecidas para el efecto en [13] - [15]. Estos criterios, sumado con la experiencia adquirida en el diagnóstico DGA, han sido sintetizados en la Fig. 1.

El procedimiento indicado en la Fig. 1 comienza con el establecimiento de límites de valoración obtenidos de las referencias bibliográficas [15] - [17]. Se calculan la producción y las tasas de cambio de la generación de gases en el fluido aislante del transformador. A continuación, las mediciones se categorizan en un Estado DGA que varía entre 1 y 3, proporcionando una escala que determina el nivel de atención requerida. Finalmente, esta categorización permite clasificar la condición del equipo, así como emitir diagnósticos y planes de mantenimiento específicos. Este proceso tradicional no solo valida y aprueba los resultados, sino que también establece la necesidad de adoptar la metodología CRISP-DM.

En este contexto, la metodología CRISP-DM, constituye un marco de trabajo estructurado y ampliamente validado, diseñado para la ejecución de proyectos de minería de datos, aprendizaje automático e inteligencia de negocios. CRISP-DM facilita la extracción de conocimientos relevantes y la identificación de patrones no evidentes en conjuntos de datos variados, contribuyendo así la toma de decisiones informadas y la generación de valor estratégico [9], [12].

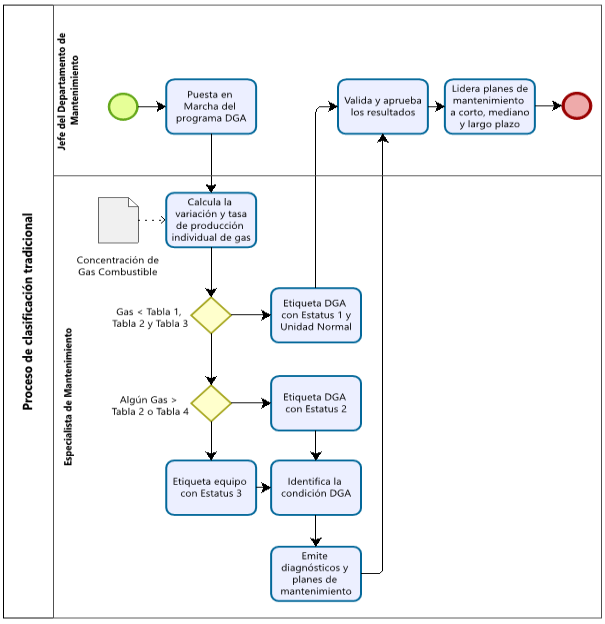

Figura 1: Proceso tradicional de clasificación DGA para transformadores eléctricos

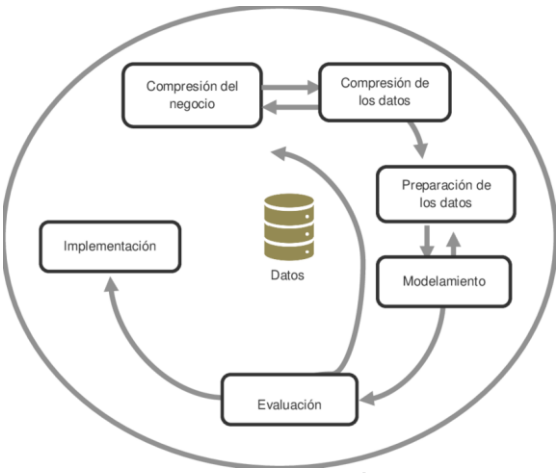

Figura 2: Fases de la metodología CRISP-DM

En el marco de CRISP-DM (ver Fig. 2), como se describe en [18], el ciclo de vida del proyecto abarca seis fases interrelacionadas que se siguen de manera secuencial. Las flechas representan las trayectorias más comunes entre estas etapas, indicando que el flujo del proyecto puede ser iterativo y permite movimientos flexibles entre las fases según sea requerido para el ajuste del análisis.

La aplicación de la metodología CRISP-DM al proceso tradicional de clasificación de la condición en transformadores eléctricos, permite un enfoque de gestión avanzada y estructurada al proceso DGA, favoreciendo así la toma de decisiones basadas en el análisis de los datos, lo que constituye en una oportunidad para el establecimiento de estrategias de mantenimiento a nivel gerencial a corto, mediano y largo plazo.

3. APLICACIÓN DE LA METODOLOGÍA CRISP-DM EN EL ANÁLISIS DE LA CONDICIÓN DE UNIDADES DE TRANSFORMACIÓN DEL SECTOR ELÉCTRICO ECUATORIANO

La aplicación de la metodología CRISP-DM comprende 6 fases, las cuales se aplican en los siguientes numerales:

3.1. Fase de comprensión del negocio

Para la explicación de esta fase, se utilizan los datos de la Estadística Anual y Multianual del Sector Eléctrico Ecuatoriano, debido a su valor estratégico para describir la situación actual y definir los objetivos de la minería de datos. En este contexto, en 2023, se identificó que la infraestructura del Sistema Nacional de Transmisión (SNT) en Ecuador estaba compuesta por 91 transformadores, distribuidos en 56 subestaciones fijas y 4 subestaciones móviles, con una capacidad de transformación máxima de 15 855,55 MVA; lo que representa un incremento del 81,86% respecto al 2014 [2].

Asimismo, durante el mismo período de análisis, la infraestructura de subtransmisión eléctrica de las empresas distribuidoras experimentó un crecimiento significativo, registrando 374 subestaciones con una capacidad combinada de transformación de 8 796,92 MVA, lo que representa un aumento del 53,46% en su capacidad máxima. las empresas generadoras y autogeneradoras también mostraron una importante evolución en cuanto a su capacidad de transformación de potencia. Las empresas generadoras reportaron un incremento del 73,87% en su capacidad máxima, mientras que las autogeneradoras experimentaron un aumento del 31,33%. Este crecimiento, que refleja una evolución histórica del sector energético ecuatoriano, resalta la relevancia e interés que tiene el estudio profundo del estado de las unidades de transformación existentes en el sector [2].

Otro indicador que refleja esta expansión de la capacidad de transformación en toda la cadena de valor del sistema eléctrico ecuatoriano es el nivel de inversión registrado entre 2009 y 2018 (ver Fig. 3). Esta asignación de recursos económicos contribuyó significativamente a la consecución de la soberanía energética ecuatoriana, permitiendo alcanzar una cobertura del servicio eléctrico del 97,05% para el 2018. No obstante, con estas cifras, resulta esencial evaluar el estado actual de esta infraestructura para determinar la efectividad y sostenibilidad a largo plazo de las estrategias de inversión implementadas, lo que permitirá evaluar la necesidad de ajustar dichas estrategias para lograr futuros objetivos en el ámbito energético [1].

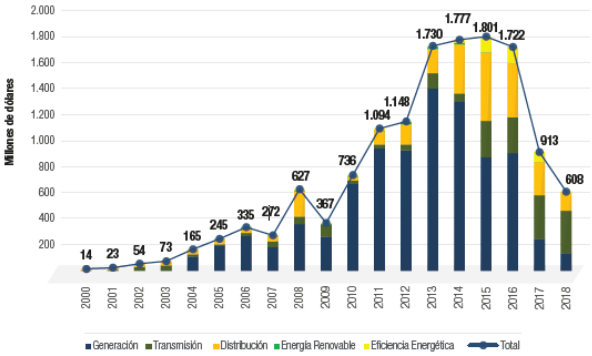

Figura 3: Inversión anual por etapa funcional en el sector eléctrico ecuatoriano

En este contexto, surge la necesidad de realizar una evaluación sistemática del estado de los transformadores instalados en el sistema eléctrico ecuatoriano, empleando técnicas DGA, como herramienta primaria de diagnóstico, con lo cual, se vislumbra una oportunidad significativa para el descubrimiento de conocimiento a partir de patrones y características relevantes. Este enfoque no solo facilita la selección precisa de una técnica de aprendizaje automático, sino que también enriquecerá el proceso de toma de decisiones con insights valiosos para la planificación estratégica y la gestión óptima de activos tanto para empresas públicas y privadas del sector eléctrico.

3.2. Fase de entendimiento de los datos

La fase de entendimiento de los datos dentro del modelo CRISP-DM, esencial por su profundidad analítica, se orienta a generar un diagnóstico del estado actual de las unidades de transformación en el sistema eléctrico ecuatoriano. Este análisis toma como referencia 1 099 registros de cromatografía de gases de aceite dieléctrico obtenidas de una población de 153 individuos ubicados en las distintas regiones de Ecuador continental (costa, sierra y oriente), incorporando a este estudio características significativas como el tipo y nivel de cargabilidad registrada en los transformadores del sistema eléctrico ecuatoriano, así como, las estrategias y políticas de mantenimiento. Este enfoque permite una comprensión holística de los factores que influyen en su rendimiento y longevidad.

Este análisis empieza por identificar la distribución de la variable objetivo en términos porcentuales.

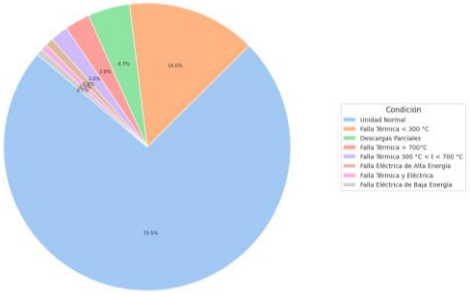

Figura 4: Condición operativa de la población de transformadores eléctricos bajo estudio

Este primer análisis revela que: el 73,50% de las unidades han sido clasificadas como Unidad Normal – UN, lo cual indica un funcionamiento correcto y valida la efectividad de las prácticas de operación y mantenimiento actuales. Sin embargo, existe un 26,50% de las muestras que presentan distintos grados de problemas operativos que requieren atención. Por otro lado, el patrón identificado en la Fig. 4, resalta un desequilibrio en la clase objetivo, el cual, requiere de tratamiento previo al proceso de entrenamiento de los modelos de clasificación automática.

En la Tabla 2, se muestran los patrones, tendencias y características estadísticas principales de las variables de interés.

Tabla 2: Resumen de los estadísticos principales

Variable	Min	Max	Media	Mediana	Std	Skew	Kurtosis
AGE	0,079	49,079	23,89	27,02	13,11	-0,196	-1,37
O2/N2	0	0,488	0,133	0,115	0,103	0,890	0,64
H2	0	3 673	10,94	0	112,34	31,59	1 028,95
CH4	0	1 037	28,019	12	50,65	8,96	151,28
C2H6	0	592	42,740	6	90,19	3,19	11,01
C2H4	0	1 068	6,924	1	45	18,82	397,09
C2H2	0	200	0,343	0	6,19	30,81	987,51
CO	0	1 821	298,36	200	272,14	1,63	3,24
CO2	29	27 237	1 951,35	1 550	1 917.85	4,85	41
O2	0	45 533	5 424,09	2 939	5 905,71	2,07	6,31
N2	1 980	133 852	44 290,19	40 700	26 718.14	0,654	-0,100

El EDA (acrónimo de: Exploratory Data Analysis), revela una amplia dispersión en los valores mínimos y máximos que se ve reflejado en los rangos de las variables, lo cual señala la necesidad de aplicar técnicas de escalamiento. Asimismo, la variabilidad en las concentraciones de gases combustibles es consistente con el perfil operativo de transformadores eléctricos, donde una baja producción de gases es una característica de condiciones normales en su operación. La asimetría (sesgo) y la curtosis (medida de colas pesadas) en la distribución de datos, apuntan a que, si bien los valores altos son indicadores de eventos de fallo, caracterizados por la liberación de energía y la gasificación activa, estos constituyen anomalías poco frecuentes en el comportamiento general de los equipos.

Al analizar la variable AGE, se observa una dispersión en las edades de los transformadores que varía desde aproximadamente un mes hasta casi 50 años. La media es superior a la mediana, lo cual indica una distribución con una tendencia hacia unidades más jóvenes, como se muestra en la Fig. 5.

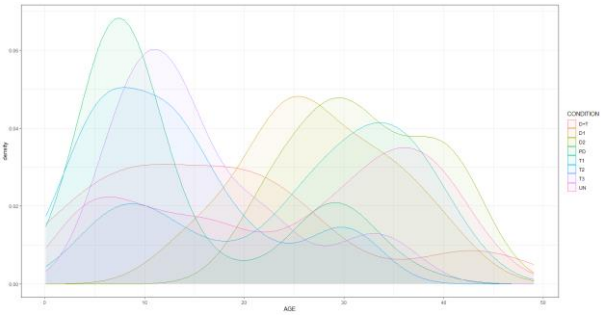

Figura 5: Análisis de densidades de la edad de los equipos

La gráfica ilustra una distribución bimodal en las edades de los transformadores eléctricos, reflejando dos subconjuntos distintos dentro de la población bajo estudio. Este patrón bimodal corresponde a periodos específicos de adquisición y puesta en marcha de nuevos equipos, en coherencia con las etapas de inversión y modernización de la infraestructura eléctrica durante los años 2008 a 2023, identificada previamente en la fase de comprensión del negocio. La presencia de dos modas en la distribución indica la coexistencia de transformadores de mayor longevidad y los más recientemente incorporados al sistema eléctrico ecuatoriano.

Frente a este análisis, la primera moda coincide con una mayor incidencia de condiciones adversas, identificadas como PD, D+T, T2 y T3, en equipos más nuevos. Esta tendencia marca una potencial problemática en la esperanza de vida de unidades recién integradas, considerando que, la segunda moda observada en la Fig. 5, alberga la mayor concentración de transformadores eléctricos en estado de condición normal, lo cual subraya su robustez y confiabilidad técnica del equipamiento y tecnología utilizada hasta antes del año 2008.

Una explicación a esta situación es la incorporación de equipos en el sistema eléctrico ecuatoriano sin una evaluación técnica previa y rigurosa del desempeño histórico de las marcas o fabricantes, de conformidad con estándares nacionales e internacionales. A esta problemática se suma la experiencia limitada del personal encargado de las adquisiciones en la definición de especificaciones técnicas más detalladas y robustas, lo que contribuye a la selección de equipos con una aparente predisposición a obsolescencia programada. Este hallazgo, incide directamente en la expectativa de vida operativa de transformadores eléctricos, la cual se estima en 25 años, vista únicamente en términos de la despolimerización del papel aislante, producto de los efectos del pirólisis [19]. Este análisis conduce a una investigación de las relaciones entre las variables en estudio. Para este propósito se utilizó un mapa de calor para las variables predictoras, una técnica estadística esencial que proporciona una representación visual de las correlaciones de Pearson, facilitando la identificación de colinealidades.

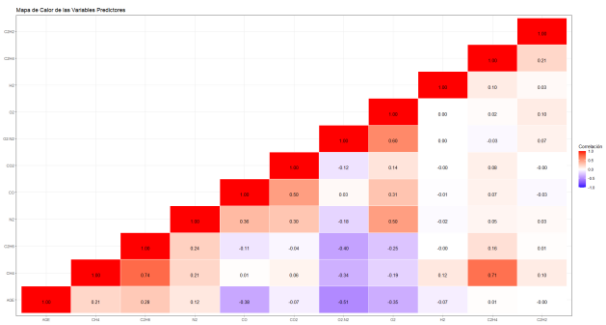

Figura 6: Análisis de correlación

La Fig. 6, muestra notables correlaciones entre metano (CH_4) y etano (C_2H_6), así como entre metano (CH_4) y etileno (C_2H_4). Estos vínculos entre gases son de particular interés, pues permiten identificar fallas térmicas T1 y T2 respectivamente. Si bien, desde la perspectiva de la ciencia de datos, la presencia de fuertes correlaciones recalca la necesidad de considerar la interdependencia entre variables al construir modelos analíticos, desde el campo de la ingeniería de mantenimiento, la presencia combinada de estos gases ha sido reconocida como un indicador esencial para la identificación y clasificación de anomalías térmicas en transformadores eléctricos [14] - [16].

La fase del entendimiento de los datos culmina con la aplicación de la técnica de Análisis de Componentes Principales (PCA), el cual confirma las correlaciones previamente identificadas entre metano y etano, así como entre metano y etileno. Por otro lado, contrariamente a las suposiciones empíricas más comunes sobre el diagnóstico de transformadores, este análisis revela una correlación menos pronunciada entre la producción de gases y la antigüedad del equipo. Esto implica que, si bien la edad del transformador juega un rol en la generación de gases con el tiempo, hay otros factores más influyentes, como son fallas del tipo eléctrico o térmico, que son factores clave en los procesos de gasificación.

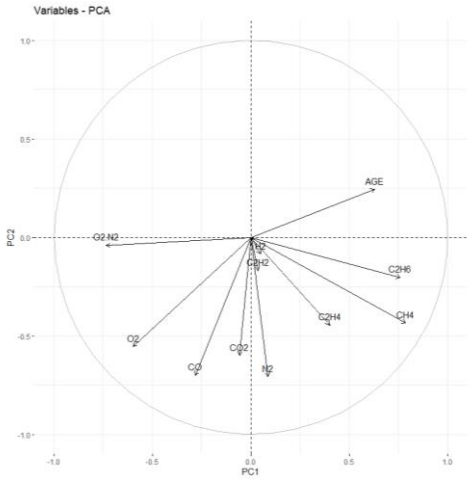

Figura 7: Análisis de correlación utilizando PCA

Finalmente, el gráfico PCA (Fig. 7) denota una marcada relación entre acetileno (C_2H_2) y el nitrógeno (N_2). Este vínculo inesperado es una observación que requiere una exploración más detallada en investigaciones futuras para esclarecer su naturaleza y significado.

3.3. Fases de preparación de los datos y modelamiento analítico

Partiendo del conocimiento de los datos, la fase de preparación de datos consistió en la aplicación de un proceso ETL (acrónimo de: Extract, Transform, Load), cuyas fases se describen a continuación:

a) Depuración de registros duplicados,

b) Escalamiento de las variables predictoras,

c) Identificación y manejo adecuado de valores atípicos (outliers),

d) Transformación de variables categóricas y,

e) Selección de las variables predictoras relevantes.

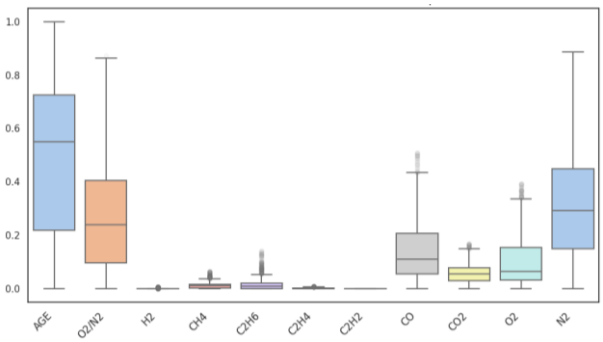

Figura 8: Inspección de los datos preparados para el modelamiento analítico

Este proceso implicó la reducción de 9 registros duplicados (0,82% de la totalidad de la información) y cuyo tratamiento de outliers permitió mejorar el sesgo como resultado de la reducción significativa de valores extremos, mejorando así la calidad de los datos para análisis posteriores (ver Fig.8). Para este tratamiento se aplicó un método por el cual los valores que excedían tres desviaciones estándar de la mediana fueron sustituidos por la propia mediana. Esta técnica se basa en el principio de que, en una distribución normal, se espera que el 99,7% de los datos se encuentre dentro de tres desviaciones estándar de la media, lo cual resultó ser un enfoque adecuado para este estudio. Además, la aplicación de este método se justificó por las siguientes razones:

a) Los valores atípicos se determinaron como manifestaciones legítimas de variabilidad en las condiciones de los datos, y no como errores de muestreo.

b) La mediana, por su resistencia a los valores extremos en comparación con la media, ofrece una medida más confiable del centro de la distribución de datos.

c) Aunque inicialmente se empleó la técnica convencional basada en el rango intercuartílico (IQR) para identificar valores atípicos, este método no resultó ser el más apropiado para este estudio en particular, ya que redujo la precisión de los modelos analíticos.

Una vez preparados los datos, se definieron las características predictoras esenciales para el modelado analítico (ver Tabla 3).

Tabla 3: Variables para el entrenamiento de modelos analíticos

ATRIBUTO	DESCRIPCIÓN DEL PREDICTOR
AGE	Edad del Equipo
O_2/N_2	Relación oxígeno – nitrógeno
H_2	Hidrógeno
CH_4	Metano
C_2H_6	Etano
C_2H_4	Etileno
C_2H_2	Acetileno
CO_2	Dióxido de Carbono
$CONDITION\_CODE$	Clase Objetivo (Etiqueta)

Previo al proceso de modelado analítico, se adoptaron estrategias esenciales enfocadas en la precisión, equidad y generalización, cuya necesidad emerge a partir del proceso de entendimiento de los datos:

a) Se seleccionaron algoritmos de clasificación supervisada especialmente dedicados por su tolerancia a los efectos del desbalance de clases y cualquier sesgo aún presente en la distribución de las variables predictoras.

b) Se equilibró la clase de interés utilizando técnicas avanzadas de generación de datos sintéticos mediante la técnica Synthetic Minority Over-sampling Technique (SMOTE).

c) Se dividieron los datos, destinando el 70% para entrenamiento y el 30% restante para evaluación, siendo esta una cantidad significativa para una validación precisa, previniendo así los efectos del overfitting.

d) El afinamiento de los hiperparámetros se llevó a cabo a través de la aplicación de la técnica de búsqueda en cuadrícula (Grid Search), utilizando validación cruzada con cinco pliegues, lo que permitió, además de reducir el riesgo de sobreajuste, mejorar la generalización de los modelos al evaluar su rendimiento con distintas particiones de los datos.

3.4. Fase de evaluación de los modelos de clasificación

Esta sección está dedicada a la validación de los modelos de clasificación automática previamente entrenados. Para esta evaluación, se utilizó métricas de precisión (accuracy), sensibilidad (recall) y puntuación F1 (F1-Score) (ver Tabla 4).

Tabla 4: Evaluación de los modelos de aprendizaje automático

MODELO	ACCURACY	RECALL	F1-SCORE
Árbol de Decisión	81,34%	81.35%	82,19%
Bosques Aleatorios	91,13%	91,13%	91,12%
SVM	85,93%	85,93%	85,91%
KNN	71,87%	71,87%	77,07%

El modelo de clasificación entrenado a partir del algoritmo de bosques aleatorios obtuvo una precisión, sensibilidad y puntuación F1 superiores al 91%, lo que lo convierte en una propuesta sólida para la detección de fallas en los transformadores eléctricos dentro del contexto energético ecuatoriano.

Otra evidencia positiva en torno a los resultados obtenidos se muestra en la curva característica operativa del receptor (ROC) ilustrada en la Fig. 9. Con un área bajo la curva (AUC) de 0,97, se destaca la notable capacidad del modelo de bosques aleatorios para diferenciar entre las distintas categorías operativas de las muestras DGA de los transformadores eléctricos analizados en este estudio. Un AUC cercano a 1 indica que el modelo es altamente sensible, con una elevada probabilidad de detectar correctamente las clases. Al mismo tiempo, el modelo demuestra una alta especificidad, como lo evidencia una baja tasa de falsos positivos, validando las métricas obtenidas en la Tabla 5.

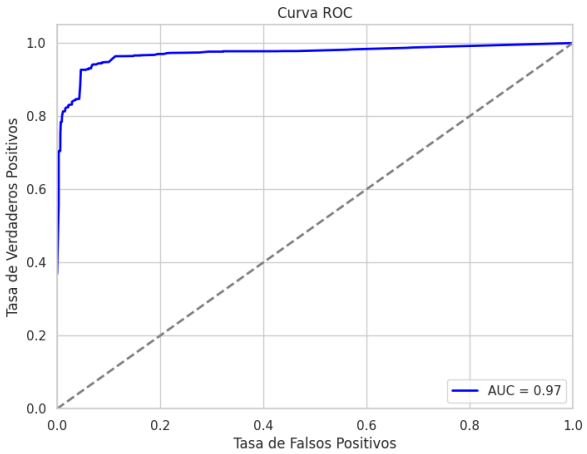

Figura 9: Curva ROC y AUC del modelo Bosques Aleatorios

Finalmente, en la Tabla 6 se presenta la matriz de confusión para el modelo de bosques aleatorios, destacándose por su competencia al identificar correctamente la condición operativa normal (UN) en 227 instancias; sin embargo, la precisión al clasificar ciertas condiciones de falla varía. Estas diferencias en la precisión de la identificación de categorías operativas de transformadores eléctricos pueden estar vinculadas a una representación insuficiente de datos en el conjunto de entrenamiento para las clases D+T, D1 y D2, más que a una deficiencia intrínseca del modelo. Sin embargo, a pesar de estos retos, el modelo ha probado su eficacia al reconocer estados operativos normales, lo que se refleja en el valor de AUC en la curva ROC, demostrando una excelente capacidad de discriminación general antes que específica.

Tabla 5: Matriz de confusión del modelo de bosques aleatorios

	D+T	D1	D2	PD	T1	T2	T3	UN
D+T	1	0	0	0	0	1	0	0
D1	0	1	1	0	0	0	0	0
D2	0	0	1	0	0	0	0	1
PD	0	0	0	14	0	0	0	2
T1	0	0	0	0	43	0	0	8
T2	0	0	0	0	1	6	0	1
T3	0	0	0	0	1	0	5	0
UN	1	0	0	4	7	0	1	227

3.5. Implementación

La Fig. 10 presenta un proceso automatizado de clasificación DGA para transformadores eléctricos diseñado para mejorar y optimizar las actividades de mantenimiento. Este sistema ha demostrado ser capaz de reducir significativamente el tiempo necesario para el procedimiento de clasificación tradicional, alcanzando una eficiencia incrementada de hasta un 92% con el volumen de datos manejados en este estudio. La ventaja principal de esta metodología no reside solo en la celeridad del proceso, sino en evolución funcional del especialista de mantenimiento, quien se convierte en un en un supervisor de la clasificación DGA. Este cambio permite que los especialistas aprovechen el tiempo ahorrado en el desarrollo estrategias de mantenimiento más enfocadas y personalizadas para cada unidad de transformación eléctrica que requiera atención.

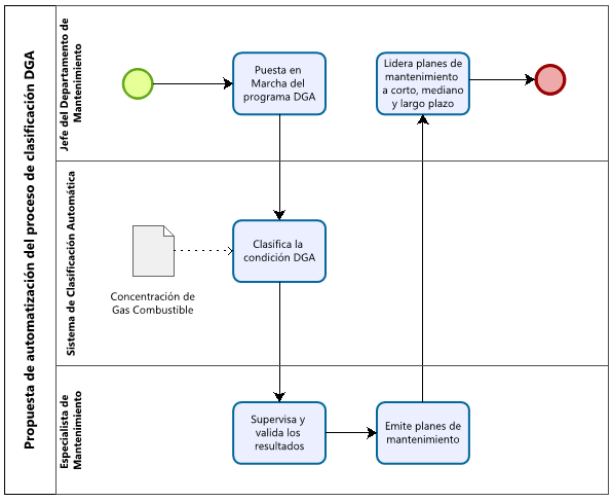

Figura 10: Propuesta de automatización del proceso de clasificación DGA para transformadores eléctricos

La comparación entre el proceso tradicional de clasificación DGA y la metodología automatizada propuesta revela diferencias fundamentales en la operatividad y la asignación de responsabilidades. En el enfoque convencional, detallado en la Fig. 1, el proceso es más lineal y manual, requiriendo una serie de pasos que van desde el cálculo de la producción de gas hasta la emisión de un diagnóstico. Contrariamente, la Fig. 10 muestra un sistema automatizado que integra y procesa datos con mínima intervención humana, dejando al especialista de mantenimiento un papel de supervisión y análisis crítico. Esta optimización del flujo de trabajo facilita una asignación estratégica de recursos, donde los especialistas dedican sus esfuerzos al planeamiento de protocolos de mantenimiento y a la gestión del activo, en lugar de ocuparse de tareas de cálculo y clasificación que pueden ser ejecutadas por sistemas automatizados.

4. DISCUSIÓN Y RESULTADOS

La aplicación de la metodología CRISP-DM, no solo culminó en la generación de un modelo de clasificación para la detección de fallas en los transformadores eléctricos con una precisión, sensibilidad y puntuación F1 superiores al 91%, sino que también permitió obtener insights valiosos sobre la condición operativa de estos equipos. Se develó que, aunque el 73,5% de las muestras se encuentran en estado normal, un significativo 26,5% requiere atención. Estos datos son esenciales para comprender la salud global del parque de transformadores y orientar estrategias de inversión y programas de mantenimiento, con el objetivo de reducir los costos de energía adicional debido a contingencias que podrían provocar restricciones de carga o salidas no programadas de estos equipos.

Un hallazgo importante encontrado en las fases de comprensión del negocio y entendimiento de datos es la identificación de subpoblaciones en la variable de edad. Se descubrió que las unidades de transformación recientemente integradas a la infraestructura eléctrica ecuatoriana muestran una mayor incidencia de condiciones: PD, D+T, T2 y T3. La aparente predisposición de los equipos más nuevos hacia una obsolescencia prematura señala la necesidad de revisar las estrategias y políticas de adquisición para mitigar este riesgo, garantizando una mayor durabilidad y confiabilidad de los activos adquiridos.

El uso de PCA en la fase de entendimiento de los datos reveló una correlación inesperada entre el acetileno (C_2H_2) y el nitrógeno (N_2). Este vínculo sugiere la posibilidad de procesos subyacentes no convencionales o no completamente entendidos, que requieren una investigación más profunda. Comprender esta correlación desde la perspectiva química de las fallas en transformadores eléctricos podría aportar nuevas perspectivas en el diagnóstico y la comprensión de estos eventos, cambiando potencialmente el enfoque del mantenimiento y la evaluación de estos equipos.

Por otro lado, los resultados mostrados en la Tabla 5 indican que, mientras las referencias bibliográficas de

[5], [6], [20], tienden a favorecer el algoritmo SVM para tareas de automatización del diagnóstico de la condición de transformadores eléctricos basado en el análisis de gases disueltos en aceite; el modelo de bosques aleatorios muestra un mejor desempeño para la diversidad de datos recopilados en este estudio. Este modelo destaca por su eficacia en un enfoque basado en reglas y el paradigma de la sabiduría de las masas, frente a otras metodologías de aprendizaje supervisado, que no podrían ser generalizadas efectivamente dadas las características y condiciones operativas específicas del sector eléctrico ecuatoriano. El modelo entrenado con el algoritmo de bosques aleatorios no solo ofrece una mayor exactitud en la clasificación, sino que también demuestra un balance óptimo entre la precisión y la detección de clases positivas, un aspecto esencial en escenarios donde los errores de predicción pueden tener consecuencias críticas.

En contraste con los resultados obtenidos en [3], la implementación del modelo CRISP-DM, en conjunción con una muestra más amplia y diversa de datos DGA de transformadores eléctricos, ha permitido expandir los cuatro estados de diagnóstico a las siete categorías esenciales. Esta mejora en la metodología también ha reducido el sobreajuste y ha mejorado el rendimiento del modelo de clasificación automática, demostrando la importancia de un enfoque holístico en la ciencia de datos aplicada al diagnóstico de los transformadores eléctricos.

Como resultado final de este estudio, se propone la implementación de un proceso automatizado de clasificación DGA para transformadores eléctricos, facilitando la optimización de estrategias de mantenimiento y mejorando la confiabilidad del servicio público de energía eléctrica.

5. CONCLUSIONES Y RECOMENDACIONES

En este trabajo presenta la aplicación de la metodología CRISP-DM al tradicional análisis de gases disueltos en transformadores eléctricos del sector eléctrico ecuatoriano. Esta metodología, que abarcó desde la comprensión del negocio hasta la implementación de un modelo de machine learning para automatizar la clasificación DGA, representa un avance significativo para el sector energético.

Desde la perspectiva de la ciencia de datos, los resultados obtenidos permitirán a los agentes públicos y privados en generación, transmisión y distribución de energía eléctrica reformular políticas de mantenimiento, abordando problemáticas que van desde la mejora en la metodología de adquisición hasta el desarrollo y aplicación de estrategias innovadoras en los procesos de diagnóstico y planificación de tareas de mantenimiento.

REFERENCIAS BIBLIOGRÁFICAS

[1] Ministerio de Energía y Recursos Naturales no Renovables del Ecuador, “Plan Maestro de Electricidad”.

[2] Agencia de Regulación y Control de Energía y Recursos Naturales No Renovables, “Estadística Anual y Multianual 2023 DEL SECTOR ELÉCTRICO ECUATORIANO”, Quito.

[3] A. Freire, J. Astudillo, C. Quinatoa, y F. Arias, “Interpretación de Gases Disueltos en Aceite Dieléctrico Mediante Bosques Aleatorios Para la Detección de Anomalías en Transformadores de Potencia”, Revista Técnica “energía”, vol. 19, no 2, pp. 90–98, ene. 2023, doi: 10.37116/revistaenergia.v19.n2.2023.544.

[4] W. H. Tang y Q. H. Wu, “Condition Monitoring and Assessment of Power Transformers Using Computational Intelligence”, Power Systems, vol. 58, 2011, doi: 10.1007/978-0-85729-052-6.

[5] X. Z. Wang, M. Z. Lu, y J. B. Huo, “Fault diagnosis of power transformer based on large margin learning classifier”, en Proceedings of the 2006 International Conference on Machine Learning and Cybernetics, 2006, pp. 2886–2891. doi: 10.1109/ICMLC.2006.259075.

[6] T. Liu y Z. Wang, “Design of power transformer fault diagnosis model based on support vector machine”, en 2009 International Symposium on Intelligent Ubiquitous Computing and Education, IUCE 2009, 2009, pp. 137–140. doi: 10.1109/IUCE.2009.59.

[7] R. J. Fiallos, “Dissolved gas content forecasting in power transformers based on Least Square Support Vector Machine (LSSVM)”, Latin American Journal of Computing, vol. IV, no 3, pp. 55–60, 2017.

[8] A. K. Oktavius, S. R. Manalu, Sasmoko, Y. Indrianti, y J. V. Moniaga, “Artificial Intelligence in Entrepreneurial Mindfulness Using CRISP-DM Method”, en Proceedings of the 2022 IEEE 7th International Conference on Information Technology and Digital Applications, ICITDA 2022, Institute of Electrical and Electronics Engineers Inc., 2022. doi: 10.1109/ICITDA55840.2022.9971384.

[9] Institute of Electrical and Electronics Engineers. y Institute of Electrical and Electronics Engineers. Columbia Section., Incorporation of both Pre-conceptual Schemas and Goal Diagrams in CRISP-DM, 2011.

[10] M. T. Hayat Suhendar y Y. Widyani, “Machine Learning Application Development Guidelines Using CRISP-DM and Scrum Concept”, en Proceedings of 2023 IEEE International Conference on Data and Software Engineering, ICoDSE 2023,

Institute of Electrical and Electronics Engineers Inc., 2023, pp. 168–173. doi: 10.1109/ICoDSE59534.2023.10291438.

[11] S. Maataoui, G. Bencheikh, y G. Bencheikh, "Predictive Maintenance in the Industrial Sector: A CRISP-DM Approach for Developing Accurate Machine Failure Prediction Models", en 2023 5th International Conference on Advances in Computational Tools for Engineering Applications, ACTEA 2023, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 223–227. doi: 10.1109/ACTEA58025.2023.10193983.

[12] J. Salcedo y K. (Consultant) McCormick, IBM SPSS modeler essentials: effective techniques for building powerful data mining and predictive analytics solutions. Packt Publishing, 2017.

[13] M. Duval, "Dissolved gas analysis: It can save your transformer", IEEE Electrical Insulation Magazine, vol. 5, no 6, pp. 22–27, nov. 1989, doi: 10.1109/57.44605.

[14] JWG D1/A2.47, "Advances in DGA interpretation", CIGRE Technical Brochure 771, no July, p. 15, 2013, doi: 10.1002/bapi.201390039.

[15] IEEE Std C57.104, IEEE Guide for the Interpretation of Gases Generated in Oil-Immersed Transformers,2019-06-13, vol. 2019. 2019.

[16] INTERNATIONAL ELECTROTECHNICAL y COMMISSION, "Mineral oil-filled electrical equipment in service - Guidance on the interpretation of dissolved and free gases analysis", IEC 60599, vol. 4, 2022.

[17] C. W. G. D1.01/A2.11, "Recent developments on the interpretation of dissolved gas analysis in transformers", CIGRE Brochure 296, vol. 296, no June, pp. 1–33, 2006.

[18] IBM, "IBM SPSS Modeler CRISP-DM Guide". [En línea]. Disponible en: https://www.ibm.com/docs/en/spss-modeler/18.0.0?topic=spss-modeler-crisp-dm-guide

[19] T. Committee of the IEEE Power y E. Society, "IEEE Guide for Loading Mineral-Oil-Immersed Transformers and Step-Voltage Regulators Sponsored by the Transformers Committee", 2012.

[20] L. V. Ganyun, C. Haozhong, Z. Haibao, y D. Lixin, "Fault diagnosis of power transformer based on multi-layer SVM classifier", Electric Power Systems Research, vol. 74, no 1, pp. 1–7, abr. 2005, doi: 10.1016/J.EPSR.2004.07.008.

Carlos Augusto Molina. - Nació en Quito en 1986. En 2024 recibió su título de Magister en Ciencias de la Información con mención en Data Science por la Pontificia Universidad Católica del Ecuador. Desde 2004 cuenta con experiencia en diagnóstico y mantenimiento de equipamiento de alto voltaje. Su campo de investigación está orientado al análisis de datos aplicado al mantenimiento centrado en la confiabilidad.

Félix Vladimir Bonilla. - Nació en Quito en 1978. En 2018 recibió su título de PhD en Ciencias en la Universidad Federal del Sur de Rusia. Se desempeño como decano de la Escuela de Ciencias de la Tierra, Energía y Ambiente de la Universidad de Investigación de Tecnología Yachay. Actualmente es profesor de la Escuela de Mecatrónica de la UIDE y del programa de Maestría en Ciencias de la Información de la PUCE. Su campo de investigación es la aplicación del aprendizaje automático y aprendizaje profundo en el análisis de señales.