UNIVERSIDAD NACIONAL DEL CALLAO

FACULTAD DE INGENIERÍA ELÉCTRICA Y ELECTRÓNICA

ESCUELA PROFESIONAL DE INGENIERÍA ELÉCTRICA

"IMPLEMENTACIÓN DE APLICATIVO BASADO EN ALGORITMO DE MACHINE LEARNING PARA LA IDENTIFICACIÓN AUTOMÁTICA DE CAUSA RAÍZ DE FALLAS EN LÍNEAS DE TRANSMISIÓN DEL GRUPO ISA PERÚ"

TESIS PARA OPTAR EL TÍTULO PROFESIONAL DE

INGENIERO ELECTRICISTA

AUTOR

Bach. JUAN ROGGER LLACZA PORRAS

ASESOR

Mg. Ing. JESSICA ROSARIO MEZA ZAMATA

Callao, 2022

PERÚ

HOJA DE REFERENCIA DEL JURADO Y APROBACIÓN

PRESIDENTE	**:**	**Ing. Fredy Adán Castro Salazar**
SECRETARIO	**:**	**Mg. Ing. Pedro Antonio Sánchez Huapaya**
VOCAL	**:**	**Mg. Ing. Jorge Elías Moscoso Sánchez**
ASESOR	**:**	**Mg. Ing. Jessica Rosario Meza Zamata**

DEDICATORIA

A mis padres Juan y Elva, con su constante esfuerzo y sacrificio, pude concretar mis objetivos.

AGRADECIMIENTO

Expreso mi agradecimiento primeramente a Dios, por brindarme las oportunidades de crecimiento en el aspecto personal y profesional.

A mis padres Elva y Juan, que con su constante apoyo incondicional y fe en mi persona, contribuyeron a lograr mis objetivos.

A mi alma mater y a mis profesores quienes con maestría compartieron conocimiento y enseñanzas sobre la carrera de Ing. Eléctrica.

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

RESUMEN

Se investigó mediante un diseño experimental cuasi experimental, la implementación de una aplicativo basado en algoritmo de clasificación de Machine Learning, para la identificación automática de la causa raíz de las fallas en líneas de transmisión del grupo ISA Perú. La muestra estuvo constituida por los archivos COMTRADE de fallas generados por los relés de protección de la línea de transmisión del grupo ISA Perú, en la cual se haya confirmado la causa raíz de la falla, en una ventana de tiempo del 2020 al 2021, principalmente. Los objetivos específicos fueron desarrollar un aplicativo basado en lenguaje de programación Python para transformar los datos obtenidos del archivo COMTRADE de las fallas en líneas de transmisión y posteriormente desarrollar un aplicativo basado en algoritmo de clasificación de aprendizaje supervisado de Machine Learning, para la identificación automática de la causa raíz de fallas en líneas de transmisión del grupo ISA Perú. El estudio se realizó en el área de Mantenimiento de Líneas de Transmisión de Red de Energía del Perú. Para la ingeniería de variables, se utilizaron características contextuales (tiempo y zona geográfica), características en el dominio del tiempo y de la frecuencia. Para el análisis exploratorio de datos, se utilizaron diagramas de caja. Se eligió el algoritmo k-nearest neighbors para realizar el entrenamiento del modelo de aprendizaje supervisado de Machine Learning. Los resultados de las métricas de evaluación fueron para el conjunto de muestras de test, una exactitud del 93%, con una sensibilidad del 97% para descarga atmosférica, 96% para humedad y contaminación y 85% para quema de vegetación.

ABSTRACT

It was investigated through a quasi-experimental experimental design on the implementation of an application based on a Machine Learning classification algorithm, for the automatic identification of the root cause of failures in transmission lines of the ISA Peru group. The sample consisted of the COMTRADE files of faults generated by the protection relays of the transmission line of the ISA Peru group, in which the root cause of the fault has been confirmed, within a time window mainly from 2020 to 2021. The specific objectives were to develop an application based on the Python programming language to transform the data obtained from the COMTRADE file of faults in transmission lines and subsequently develop an application based on Machine Learning supervised learning classification algorithm, for automatic identification. of the root cause of failures in transmission lines of the ISA Peru group. The study was carried out in the area of Maintenance of Transmission Lines of Red de Energía del Perú. For the engineering of variables, contextual characteristics (time and geographical area), characteristics in the time and frequency domain were used. For exploratory data analysis, box plots were used. The k-nearest neighbors algorithm was chosen to perform the training of the Machine Learning supervised learning model. The results of the evaluation metrics were for the set of test samples, an accuracy of 93% was obtained, with a recall of 97% for atmospheric discharge, 96% for humidity and pollution, and 85% for vegetation burning.

INTRODUCCIÓN

La energía eléctrica se ha convertido en una base fundamental para el desarrollo tecnológico. El sistema eléctrico de potencia está conformado por los sectores de generación de la energía eléctrica (a través de recursos renovables en centrales hidráulicas, centrales solares y centrales eólicas y recursos no renovables en centrales a vapor, central de biomasa y centrales de ciclo combinado); el sector de transmisión está conformada por subestaciones eléctricas (del tipo de transformación y de maniobras) y líneas de transmisión (con estructuras tales como torres metálicas y postes de madera) en alta y extra alta tensión y el sector de distribución, el cual está conformada por subestaciones, líneas de media tensión y redes de baja tensión.

Cada sector se encuentra susceptible a que los equipos que conforman parte de la transmisión de la energía eléctrica se vean afectados por fallas de origen interno o externo, los cuales repercuten en la operación del sistema eléctrico de potencia, disminuyendo la confiabilidad, en aquellos casos en el que se aísla el componente en falla del sistema eléctrico por la actuación del sistema de protección de la línea de transmisión. Ante eso, las empresas que conforman los sectores eléctricos plantean estrategias de mantenimiento centrado en confiabilidad (MCC). A pesar de ello, hay modos de falla que dependen de circunstancias externas (ambientales y circunstanciales). En el caso específico de las líneas de transmisión de alta y extra alta tensión, tales modos de fallas de origen externo (del tipo ambiental) son frecuentemente contaminación y humedad en cadena de aisladores, descargas atmosféricas y quema de vegetación (que disminuye la resistividad dieléctrica del aire), entre otros. Ante

estos eventos las empresas del sector de transmisión de energía eléctrica, ejecutan actividades de mantenimiento correctivo de emergencia para recuperar la disponibilidad de la línea de transmisión y la confiabilidad de la operación del sistema eléctrico de potencia.

I. PLANTEAMIENTO DEL PROBLEMA

En la operación del Sistema Eléctrico Interconectado Nacional, se presentan fallas en líneas de transmisión de alta y extra alta tensión, que originan indisponibilidad en la línea de transmisión afectada. La causa raíz de la falla en la línea de transmisión puede abarcar distintos tipos, tales como humedad y contaminación (predominantemente en zona costera) en la cadena de aisladores de suspensión y anclaje de las torres de alta tensión, descargas atmosféricas, y quema de vegetación alrededor de la línea de transmisión (las partículas reducen la rigidez dieléctrica del aire). Una vez ocurrida la falla, el área de Mantenimiento de Líneas de Transmisión, solicita la distancia corregida al punto de falla, calculada por el área de Protecciones a partir de las oscilografías de la falla en ambos extremos de la línea de transmisión. Con la información de la distancia corregida el punto de falla, se elige un rango de estructuras (±5% de la distancia de falla corregida) para realizar la inspección de la línea de transmisión.

1.1 Descripción de la realidad problemática

Cuando se presenta falla en la línea de transmisión, se desconoce la causa raíz que originó la falla. En ciertos casos, las condiciones contextuales tales como hora, estación del año y mapa de descargas atmosféricas en tiempo real, durante el evento de falla, dan un indicio de la causa raíz y es posible realizar un intento de energización de la línea de transmisión, previa verificación que el punto de falla no corresponda a una zona poblada.

En los casos más severos, las condiciones contextuales no permiten estimar la causa raíz de la falla o el intento de energización resultó sin éxito. En estos escenarios, el área de Mantenimiento de Líneas de Transmisión programa a un

equipo de inspectores para que el desplazamiento a la línea de transmisión, y realizar la inspección en un rango del ±5% de la distancia de falla corregida (rango de estructuras de la línea de transmisión). Esto trae como consecuencia un mayor tiempo de indisponibilidad de la línea de transmisión, mayores gastos en la reposición del servicio hasta incluso con la posibilidad de pago de compensación por interrupción del suministro eléctrico. Tener implementado un aplicativo que identifique automáticamente la causa raíz de la falla en la línea de transmisión, a partir de las características en el dominio del tiempo y de la frecuencia (de las formas de ondas de tensión y corriente de la falla) y características contextuales (hora, mes del año, región, estación del año en que ocurrió la falla), puede proveer información a el equipo inspector de la línea de transmisión para que se encuentre apropiadamente equipado y preparado para buscar signos de fallas característicos durante la inspección, así como reponer el servicio de la línea de transmisión en los casos donde no está permitida a declararse disponible sin realizar una inspección completa para resolver la incertidumbre acerca de la causa de falla y daño sobre los elementos de la línea de transmisión.

1.2 Formulación del problema

1.2.1 Problema general

¿La implementación de un aplicativo basado en algoritmo de Machine Learning identificará automáticamente la causa raíz de fallas en líneas de transmisión del grupo ISA Perú?

1.2.2 Problemas específicos

P.E.1 ¿El desarrollo de un aplicativo basado en lenguaje de programación Python, extraerá características contextuales, características en el dominio del tiempo y de la frecuencia de los archivos COMTRADE de las fallas en líneas de transmisión?

P.E.2 ¿El entrenamiento de un modelo de Machine Learning basado en algoritmo de clasificación de aprendizaje supervisado, identificará automáticamente la causa raíz de fallas en líneas de transmisión del grupo ISA Perú?

1.3 Objetivos

1.3.1 Objetivo general

Implementar un aplicativo basado en algoritmo de Machine Learning para la identificación automática de la causa raíz de fallas en líneas de transmisión del grupo ISA Perú.

1.3.2 Objetivos específicos

O.E.1 Desarrollar un aplicativo basado en lenguaje de programación Python, para la extracción características contextuales, características en el dominio del tiempo y de la frecuencia de los archivos COMTRADE de las fallas en líneas de transmisión.

O.E.2 Realizar el entrenamiento de un modelo de Machine Learning basado en algoritmo de clasificación de aprendizaje supervisado, para la identificación automática de la causa raíz de fallas en líneas de transmisión del grupo ISA Perú.

1.4 Justificación

La presente investigación tiene justificación práctica debido a que la implementación del aplicativo identificará automáticamente la causa raíz de la falla en la línea de transmisión del grupo ISA Perú. Tiene justificación económica, debido a que permitirá disminuir el valor de una posible penalidad económica por interrupción de suministro de energía eléctrica. Tiene justificación tecnológica debido a que, en el desarrollo del aplicativo, se emplean técnicas de analítica de datos y aprendizaje supervisado de Machine Learning.

1.5 Limitantes de la investigación

En la investigación se tiene como límites los siguientes puntos:

- Límite sobre la cantidad de archivos COMTRADE de fallas en líneas de transmisión. Esto se debe a que, en ciertos eventos de falla, no ha sido posible identificar la causa raíz mediante inspección en la línea de transmisión, por lo que se dispone de una menor cantidad de archivos COMTRADE, para el aprendizaje del modelo de Machine Learning.
- Para las fallas con causa raíz de quema de vegetación, se dispone de una menor cantidad de fallas para el periodo seleccionado de estudio (2020 al 2021). Por lo que, para el adecuado aprendizaje del modelo de Machine Learning, fue necesario buscar eventos de falla desde el 2010.

II. MARCO TEÓRICO

2.1 Antecedentes del estudio

Flores (2021) realizó un estudio titulado “Identificación de causa raíz de fallas por descargas eléctricas en líneas de transmisión” en la ciudad de Quito, Ecuador.

Propuso una metodología para la identificación de la causa raíz de las desconexiones en las líneas de transmisión de alta tensión en Ecuador. Para tal fin, utilizó parámetros eléctricos de la falla, tales como fases falladas, punto de incepción de falla, nivel de tensión eléctrica e impedancia de falla y parámetros no eléctricos como la condición ambiental en el instante de la falla. Estos parámetros fueron utilizados como datos de entrada para el entrenamiento del algoritmo de aprendizaje supervisado de Machine Learning Vecinos Más Cercanos (k-nearest neighbors). Para la validación de los resultados utilizó la matriz de confusión para identificar el nivel de precisión de clasificación del modelo entrenado. El mejor resultado que obtuvo fue del 93.75% de precisión de clasificación, para el caso específico de descargas atmosféricas como causa raíz de la falla en la línea de transmisión.

Adauto (2021) realizó un estudio sobre “Aplicación de inteligencia artificial en la detección de fallas en los motores eléctricos de corriente continua de imán permanente” en la ciudad de Huancayo, Perú.

El objetivo de la investigación fue identificar el aumento de confiabilidad del procedimiento de reconocimiento y localización de fallas en motores eléctricos de corriente continua de imán permanente, a través de la aplicación de un modelo de Machine Learning. La conclusión del estudio fue que no se logró incrementar la fiabilidad y eficiencia en la utilización de métodos de diagnóstico

de fallas mediante la aplicación de modelos de Machine Learning, en la cual se dispone de poca cantidad de datos de los motores en estudio.

Gallegos (2020) realizó un estudio sobre “Identificación de fallas en sistemas eléctricos de potencia basado en reconocimiento de patrones”, en la ciudad de Quito, Ecuador.

El objetivo de la investigación fue el de lograr una alta precisión de identificación de falla en líneas de transmisión, a través de un modelo entrenado de Machine Learning para el reconocimiento de patrones. La conclusión del estudio fue que el modelo de Machine Learning basado en de algoritmo k-nearest neighbors, permite identificar con éxito el tipo de falla en la línea de transmisión, en un tiempo de 3 s, a través del reconocimiento de patrones.

Reveco (2019) realizó un estudio sobre “Análisis predictivo de activos mineros para obtención de intervalo de falla mediante algoritmos de Machine Learning”, en la ciudad de Santiago de Chile, Chile.

El objetivo del estudio es generar un modelo predictivo basado en algoritmo de Machine Learning, para la identificación de la ventana de tiempo entre fallas en motores Diesel de unidades de camiones del sector minero. La conclusión del estudio fue que los tres modelos revisados (complementarios) en conjunto logran obtener una predicción general del comportamiento de los motores Diesel de las unidades de camiones del sector minero. Los modelos de detección de anomalía y predicción de RUL (Remaining Useful Life) fueron sugeridos para implementación en las actividades del proceso minero, por en cambio el modelo de clasificación de falla no aporta valor al proceso debido a que no logra una alta precisión para la clasificación de causas de fallas.

Bautista (2018) realizó un estudio titulado “Identificación de 11 tipos de fallas en líneas de transmisión de alta tensión utilizando redes neuronales” en la ciudad de Bucaramanga, Colombia.

El objetivo general del estudio fue la identificación de 11 probables tipos de fallas (haciendo referencia a los lazos de impedancia de falla) en las líneas eléctricas de alta tensión, aplicando algoritmo de redes perceptrón multicapa de redes neuronales. La muestra estuvo conformada por una base de datos de fallas simuladas en el programa ATP Draw. En las simulaciones se consideraron la variación de parámetros como valores de resistencia de falla, punto de falla en la línea de transmisión y ángulos de tensión y corriente de fases. Para la validación de resultados utilizaron los porcentajes de la precisión de la matriz de confusión y de la raíz del error cuadrático medio (RMSE). El resultado de la investigación fue que el modelo entrenado tiene una precisión mayor al 98% en la clasificación correcta de los tipos de fallas (lazos de impedancia) en las líneas de transmisión.

Hurtado, Villareal y Villareal (2016) realizaron un estudio sobre “Detección y diagnóstico de fallas mediante técnicas de inteligencia artificial, un estado del arte” en la ciudad de Bogotá, Colombia.

El objetivo del estudio fue organizar, clasificar y comparar técnicas de inteligencia artificial para la detección y diagnóstico de fallas. La conclusión del estudio fue que se debe centrar en la mejora de la robustez y adaptabilidad en las técnicas de inteligencia artificial, como parte de futuras estrategias para la detección de fallas y diagnóstico.

Minnaar, Niccols y Gaunt (2015) realizaron un estudio sobre “Automating transmisión line fault root causa analysis” en la ciudad Eskom Holdings, Sudáfrica.

El objetivo del estudio fue establecer una base operacional para clasificación de fallas de acuerdo con su causa raíz. La identificación de la causa raíz de la falla fue resuelta mediante un modelo entrenado con algoritmo de aprendizaje supervisado de Machine Learning. Para ello utilizan los siguientes parámetros contextuales como hora, mes, región, densidad promedio de descarga atmosférica y parámetros de forma de onda como tensión nominal, tipo de falla, corriente de componente de secuencia negativa a medio ciclo del inicio de la falla, constante de tiempo de la corriente de falla de secuencia positiva, constante de tiempo de la corriente de falla de secuencia negativa, máxima relación de tensión de secuencia negativa de falla a su respectiva pre-falla. La muestra estuvo conformada por 2672 fallas en las líneas de transmisión de 220, 275 y 400 kV del sistema eléctrico de Eskom en Sudáfrica, registradas en una ventana de tiempo del 1995 al 2008. Para la identificación de la causa raíz de la falla, se entrenó un modelo mediante el algoritmo denominado “Vecino más cercano” (k-nearest neighbors) de aprendizaje supervisado de Machine Learning, para a partir de las características contextuales y de forma de onda de la falla, clasificar automáticamente la causa raíz de la falla. Para la validación de resultados utilizó el porcentaje de precisión de la matriz de confusión y medida F. El resultado fue que el mejor porcentaje de precisión de clasificación fue del 90% mediante el empleo de solo características contextuales, sin embargo, esta manera de clasificación no utiliza las características de las formas de onda de falla.

Combinando las características contextuales y de formas de onda (hora, corriente de secuencia negativa a medio ciclo del inicio de falla, región, constante de tiempo de corriente de falla de secuencia positiva, mes y fases en falla), se obtiene una precisión de clasificación del 86%. Por en cambio, si solo se utilizan características de forma de onda, la precisión de clasificación más alta fue del 80%. El estudio comparó los desempeños de clasificación (mediante medida F) de los algoritmos: vecino más cercano (k-nearest neighbors), árbol de decisiones, red neuronal de base radial y clasificadores Naive Bayes. El mejor desempeño se dio mediante el clasificador de k - vecinos más cercanos, el cual obtuvo un desempeño de 12 a 14 % más alto que los otros clasificadores. El estudio concluye que el algoritmo clasificador vecino es idóneo para la clasificación de la causa raíces de fallas en líneas de transmisión.

Pianeta (2015) realizó un estudio sobre “Modelo adaptativo de inteligencia artificial para detección selectiva de fallas de alta impedancia en líneas de transmisión de dos terminales de doble circuito”, en la ciudad de Medellín, Colombia.

Su objetivo fue el diseño, elaboración y validación de un modelo de inteligencia artificial sobre protección adaptativa que sea sensible a los cambios de la topología del sistema eléctrico y cambios en el despacho de carga para una operación selectiva en la ocurrencia de una falla en líneas de transmisión de circuitos paralelos de dos terminales. El estudio concluye que la identificación de la región ideal de actuación de un relé de protección de distancia mediante la aplicación de redes neuronales es más seguro y confiable, en los variados

estados de operación del sistema eléctrico. La región adaptativa se adapta a estos diferentes valores de potencia activa y reactiva del sistema eléctrico.

Barrera, Meléndez, Kulkarni y Santoso (2012) realizaron un estudio sobre "Feature análisis and automatic classification of short-circuit faults resulting from external causes".

El objetivo del estudio fue la identificación de la causa raíz de fallas en redes de distribución a partir de un conjunto de características basadas en el tiempo y características basadas en la forma de onda de tensión y corriente de falla. El estudio se enfoca en identificar las causas raíces de origen externo, tales como contacto con vegetación, contacto con animales, descargas atmosféricas y fallas en cables. La muestra estuvo conformada por 181 eventos de falla reales documentadas, entre el periodo del 2002 al 2006, en redes de distribución de 12.47 kV. Las formas de onda de corriente y tensión fueron muestreadas a una tasa de 128/256 muestras por ciclo. Para la elaboración del conjunto de reglas de clasificación de causa raíz de falla, se utilizaron análisis multivariado de varianza y algoritmo de extracción de reglas CN2. El resultado fue una tasa de clasificación del 93.4% para el enfoque que combina características basadas en tiempo y características de formas de onda de tensión y corriente.

2.2 Bases teóricas

2.2.1 Clasificación de fallas en líneas de transmisión

Angulo (2015) indica que las fallas en líneas de transmisión se clasifican por los siguientes criterios:

- **Por la duración:**
  - **Transitorias:** donde la extinción de la falla se da sin alguna intervención o por actuación de la función 79 de los relés de protección (recierre exitoso).
  - **Permanentes:** provoca la indisponibilidad de la línea de transmisión y se requiere ejecutar un mantenimiento correctivo para la recuperación de la disponibilidad de la línea de transmisión.
- **Por la forma:**
  - **Paralelo:** son los lazos de falla: cortocircuito de fase-fase, fase-tierra, fase-fase-tierra y cortocircuito trifásico.
  - **Serie:** ausencia de una o dos fases (conductores).
  - **Evolutiva**: consiste en que la falla inicia con un determinado lazo de falla y evoluciona a otro lazo de falla (afectando a más elementos).
- **Por la simetría** de la forma de onda:
  - **Simétrica:** son fallas en donde no se tienen presentes las componentes de secuencia negativa y cero (cortocircuito trifásico).
  - **Asimétrica:** son fallas en donde se tienen presentes las tres componentes de secuencia (positiva, negativa y cero).

2.2.2 Teorema de Fortescue de componentes simétricas

Anderson (1995) el teorema indica que todo sistema de "n" fasores desbalanceados, se puede descomponer en "n" fasores balanceados. Estos fasores balanceados presentan un desfase definido por la ecuación (2.1):

$$\theta_k = k\left(\frac{2\pi}{n}\right) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (2.1)$$

Donde "k" se define como el número de secuencia y "n" el número de fasores.

Teorema de Fortescue aplicado a redes trifásicas

Un sistema trifásico de fasores desbalanceados puede ser descompuesto en tres sistemas trifásicos de secuencia balanceados, los cuales son denominados componente de secuencia positiva, componente de secuencia negativa y componente de secuencia cero.

Componente de secuencia positiva: conjunto de fasores trifásicos con magnitudes idénticas y desfasados por 120°, giran en mismo sentido al del sistema trifásico desbalanceado. Se representa por el subíndice 1.

Componente de secuencia negativa: conjunto de fasores trifásicos con magnitudes idénticas y desfasados por 120°, giran en sentido contrario al del sistema trifásico desbalanceado. Se representa por el subíndice 2.

Componente de secuencia cero: conjunto de fasores trifásicos con magnitudes idénticas y no tienen ángulo de desfase entre ellas. Esta secuencia se representa por el subíndice 0.

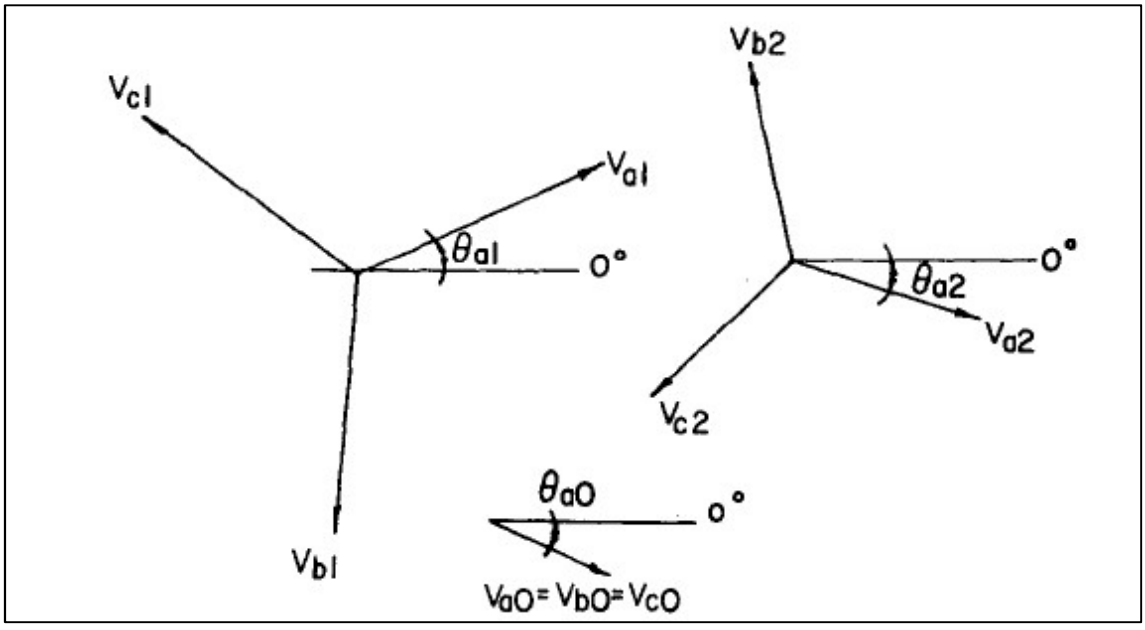


Figura N° 2.1: Representación de las componentes de secuencias (Fuente: [1])

Se define el operador “a”, como aquella cantidad compleja que produce el desfase entre fasores de las componentes de secuencia positiva y negativa. En la ecuación (2.2) se observa en forma cuadrangular:

$$a = -\frac{1}{2} + j\frac{\sqrt{3}}{2} \quad \ldots\ldots (2.2)$$

En la ecuación (2.3) se observa el operador a en forma polar:

$$a = 1\angle 120° \quad \ldots\ldots (2.3)$$

De manera analítica, los fasores de tensión de un sistema desbalanceado se representan con sus componentes de secuencia, a través de. la ecuación (2.4):

$$V_a = V_{a0} + V_{a1} + V_{a2}$$
$$V_b = V_{b0} + V_{b1} + V_{b2} \quad \ldots\ldots (2.4)$$
$$V_c = V_{c0} + V_{c1} + V_{c2}$$

Tomando en consideración que los fasores de secuencia cero son iguales y los tres fasores de la componente de secuencia positiva están desfasados por 120°, se puede tomar uso del operador “a” para representar los fasores de tensión V_{b1} y V_{c1}, mediante V_{a1}, tal como se observan en las ecuaciones (2.5) y (2.6), respectivamente:

$$V_{b1} = a^2 \cdot V_{a1} \quad \ldots\ldots (2.5)$$

$$V_{c1} = a \cdot V_{a1} \quad \ldots\ldots (2.6)$$

$$V_{a0} = V_{b0} = V_{c0} \quad \ldots\ldots (2.7)$$

A partir de las ecuaciones (2.5), (2.6) y (2.7), redefinimos las ecuaciones (2.4), en la ecuación (2.8):

$$V_a = V_{a0} + V_{a1} + V_{a2}$$

$$V_b = V_{a0} + a^2 \cdot V_{a1} + a \cdot V_{a2} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots. (2.8)$$

$$V_c = V_{a0} + a \cdot V_{a1} + a^2 \cdot V_{a2}$$

Las ecuaciones (2.7) se pueden expresar de manera matricial, tal como se observa en la ecuación (2.8):

$$\begin{bmatrix} V_a \\ V_b \\ V_c \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & a^2 & a \\ 1 & a & a^2 \end{bmatrix} \begin{bmatrix} V_{a0} \\ V_{a1} \\ V_{a2} \end{bmatrix} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2.9)$$

De la ecuación (2.9) se puede calcular las componentes de secuencia positiva, negativa y cero de tensión, a partir de los fasores de tensión del sistema desbalanceado, tal como se observa en la ecuación (2.10):

$$\begin{bmatrix} V_{a0} \\ V_{a1} \\ V_{a2} \end{bmatrix} = \frac{1}{3} \begin{bmatrix} 1 & 1 & 1 \\ 1 & a & a^2 \\ 1 & a^2 & a \end{bmatrix} \begin{bmatrix} V_a \\ V_b \\ V_c \end{bmatrix} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots. \ldots\ldots\ldots (2.10)$$

Lo mismo se aplica a los fasores de corriente del sistema desbalanceado, tal como se observan en las ecuaciones (2.11) y (2.12):

$$\begin{bmatrix} I_a \\ I_b \\ I_c \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & a^2 & a \\ 1 & a & a^2 \end{bmatrix} \begin{bmatrix} I_{a0} \\ I_{a1} \\ I_{a2} \end{bmatrix} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots. \ldots (2.11)$$

$$\begin{bmatrix} I_{a0} \\ I_{a1} \\ I_{a2} \end{bmatrix} = \frac{1}{3} \begin{bmatrix} 1 & 1 & 1 \\ 1 & a & a^2 \\ 1 & a^2 & a \end{bmatrix} \begin{bmatrix} I_a \\ I_b \\ I_c \end{bmatrix} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots.. (2.12)$$

En la **Figura N° 2.2** se observa el diagrama fasorial del sistema trifásico desbalanceado con sus respectivas componentes de secuencia:

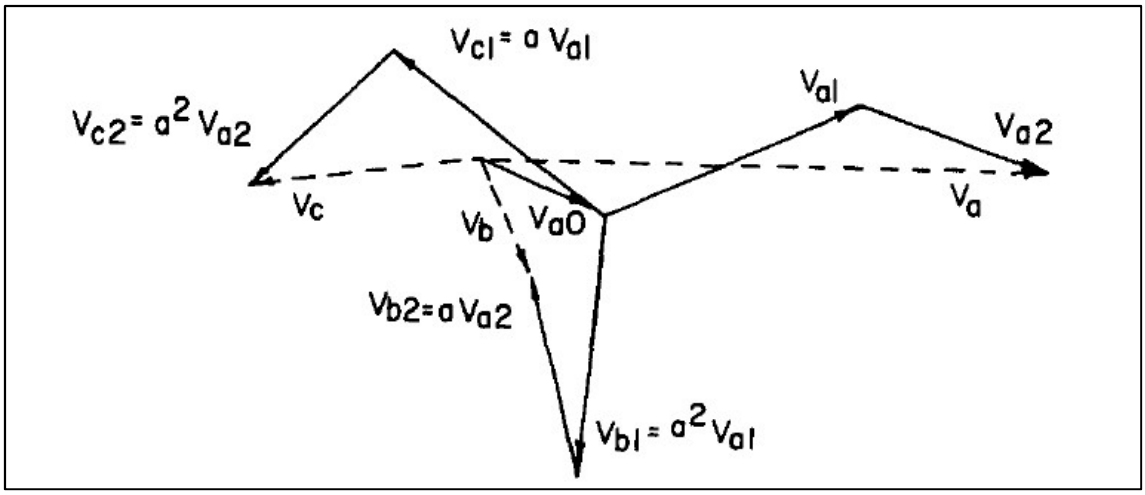


Figura N° 2.2: Diagrama fasorial del sistema trifásico desbalanceado (Fuente: [1])

2.2.3 Tipos de falla en líneas de transmisión

a) Falla monofásica a tierra:

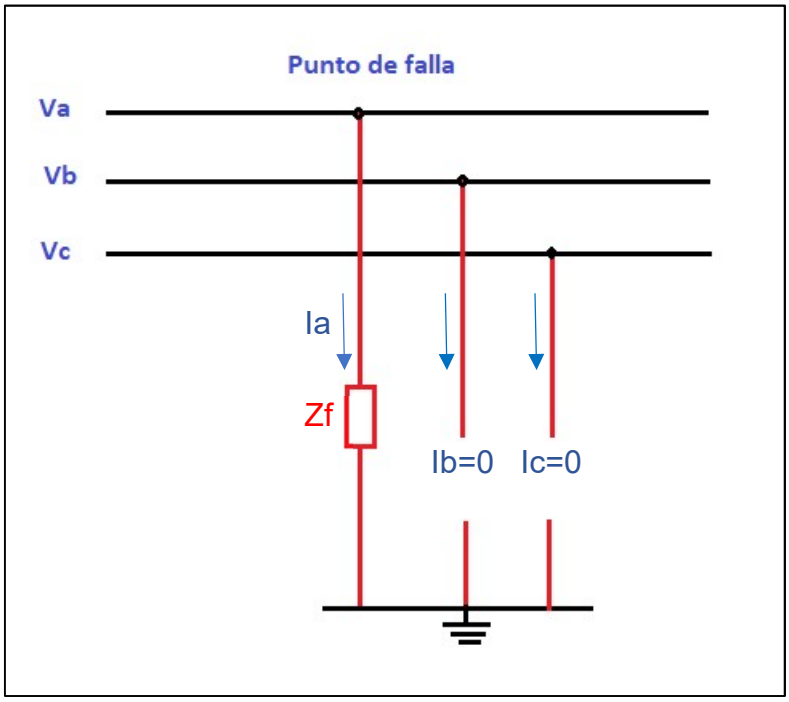


Figura N° 2.3: Diagrama de falla monofásica a tierra (Fuente: Propia del autor)

De la **Figura N° 2.3**, se verifica lo siguiente:

$$V_a = Z_f \cdot I_a \quad \ldots\ldots (2.13)$$

$$I_b = I_c = 0 \quad \ldots\ldots (2.14)$$

Reemplazamos la ecuación (2.14) en (2.12) y se obtiene la ecuación (2.15):

$$\begin{bmatrix} I_{a0} \\ I_{a1} \\ I_{a2} \end{bmatrix} = \frac{1}{3} \begin{bmatrix} 1 & 1 & 1 \\ 1 & a & a^2 \\ 1 & a^2 & a \end{bmatrix} \begin{bmatrix} I_a \\ 0 \\ 0 \end{bmatrix} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2.15)$$

Realizamos la multiplicación de la matriz y se obtiene la ecuación (2.16):

$$\begin{bmatrix} I_{a0} \\ I_{a1} \\ I_{a2} \end{bmatrix} = \frac{1}{3} \begin{bmatrix} I_a \\ I_a \\ I_a \end{bmatrix} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2.16)$$

De la ecuación (2.16) se desprende lo siguiente:

$$I_{a0} = I_{a1} = I_{a2} = \frac{I_a}{3} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2.17)$$

Reemplazamos la ecuación (2.13) en (2.4):

$$V_{a0} + V_{a1} + V_{a2} = Z_f \cdot I_a \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2.18)$$

Reemplazamos la ecuación (2.17) en (2.18):

$$V_{a0} + V_{a1} + V_{a2} = 3Z_f \cdot I_{a1}. \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2.19)$$

De la ecuación (2.17) se observa que las componentes de secuencia positiva, negativa y cero de corriente son iguales, por lo que se interpreta que las redes de secuencia están conectadas en serie. De la ecuación (2.19) se observa que la suma de las componentes de secuencia positiva, negativa y cero de tensión es igual a tres veces la falla, por lo que al circuito de las redes de secuencia se le agrega la cantidad de $3Z_f$. En la **Figura N° 2.4** se observa la conexión de red de secuencia para una falla monofásica a tierra:

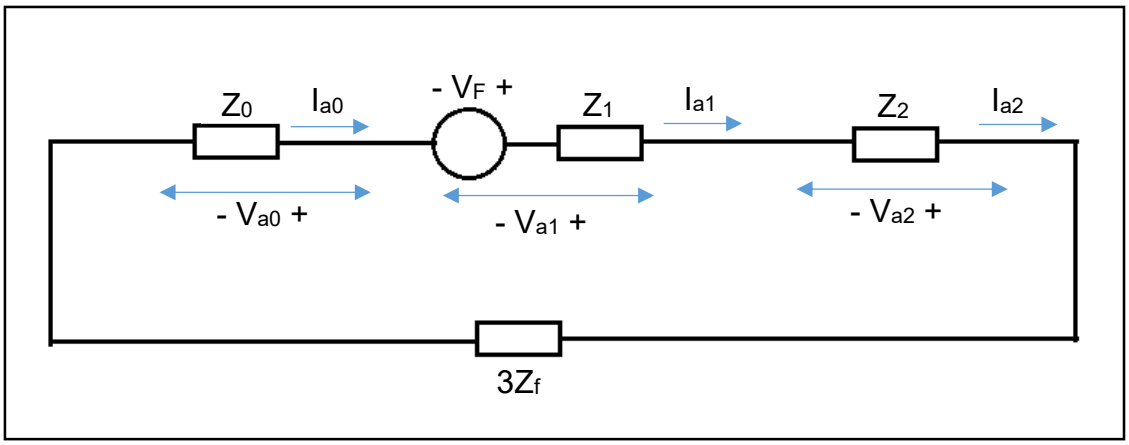


Figura N° 2.4: Conexión red secuencia para falla monofásica a tierra [Fuente: Propia del autor]

Finalmente, de la **Figura N° 2.4,** se concluye la siguiente ecuación (2.20):

$$I_{a0} = I_{a1} = I_{a2} = \frac{V_F}{Z_0 + Z_1 + Z_2 + 3Z_F} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2.20)$$

b) Falla bifásica:

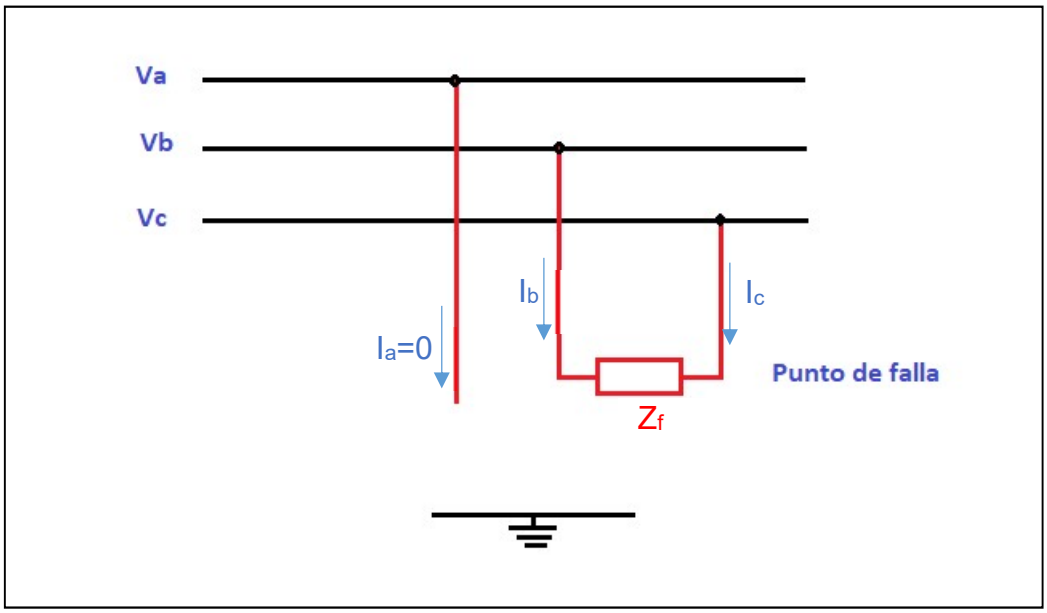


Figura N° 2.5: Diagrama de falla bifásica (Fuente: Propia del autor)

De la **Figura N° 2.5**, se verifica lo siguiente:

$$I_b = -I_c \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2.21)$$

$$I_a = 0 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2.22)$$

$$V_b - V_c = I_b \cdot Z_f \ldots\ldots\ldots (2.23)$$

Reemplazamos las ecuaciones (2.21) y (2.22) en (2.12) y se obtiene:

$$\begin{bmatrix} I_{a0} \\ I_{a1} \\ I_{a2} \end{bmatrix} = \frac{1}{3} \begin{bmatrix} 1 & 1 & 1 \\ 1 & a & a^2 \\ 1 & a^2 & a \end{bmatrix} \begin{bmatrix} 0 \\ I_b \\ -I_c \end{bmatrix} \ldots\ldots\ldots (2.24)$$

De la resolución de la ecuación (2.24), se obtiene lo siguiente:

$$\begin{bmatrix} I_{a0} \\ I_{a1} \\ I_{a2} \end{bmatrix} = j \frac{I_b}{\sqrt{3}} \begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix} \ldots\ldots\ldots (2.25)$$

De la ecuación (2.25), se desprende que:

$$I_{a1} = -I_{a2} \ldots\ldots\ldots (2.26)$$

$$I_{a0} = 0 \ldots\ldots\ldots (2.27)$$

Ahora reemplazamos la ecuación (2.8) y (2.11) en (2.23), se obtiene lo siguiente:

$$Z_f \cdot (I_{a0} + a^2 I_{a1} + a I_{a2}) = (V_{a0} + a^2 \cdot V_{a1} + a \cdot V_{a2}) - (V_{a0} + a \cdot V_{a1} + a^2 \cdot V_{a2}) \ldots (2.28)$$

De la resolución de la ecuación (2.27) se obtiene lo siguiente:

$$Z_f \cdot I_{a1} = V_{a1} - V_{a2} \ldots\ldots\ldots (2.29)$$

De la ecuación (2.27) se observa que la corriente de secuencia cero es cero y de la ecuación (2.26) se deduce la conexión de red de secuencia:

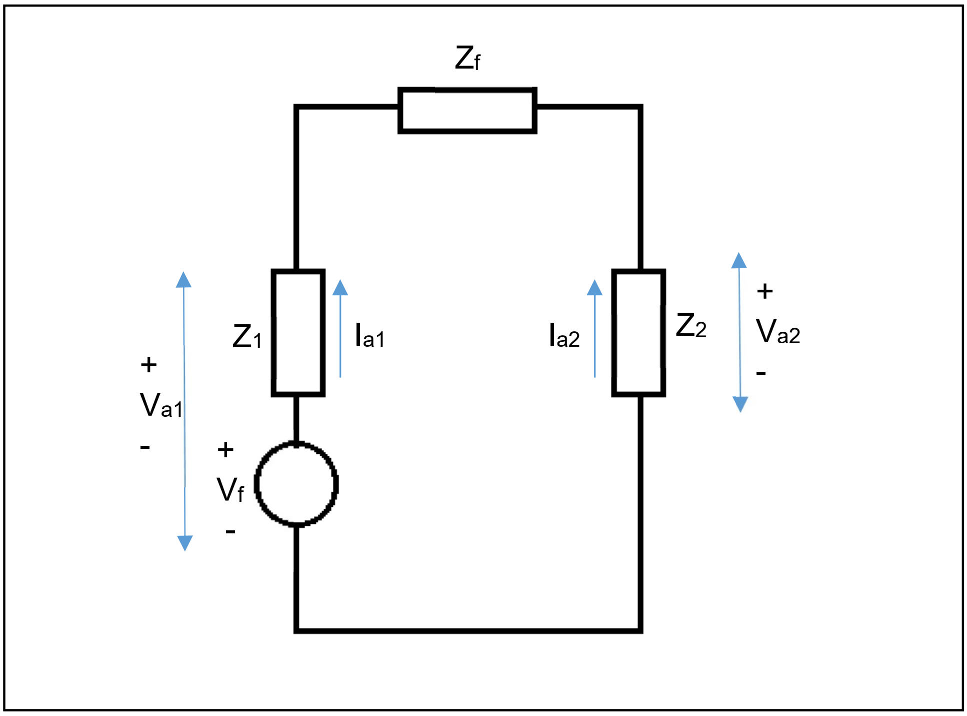


Figura N° 2.6: Conexión red secuencia falla bifásica (Fuente: Propia del autor)

De la **Figura N° 2.6**, se desprende la ecuación (2.30):

$$I_{a1} = \frac{V_f}{Z_1 + Z_2 + Z_f} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2.30)$$

c) Falla bifásica a tierra:

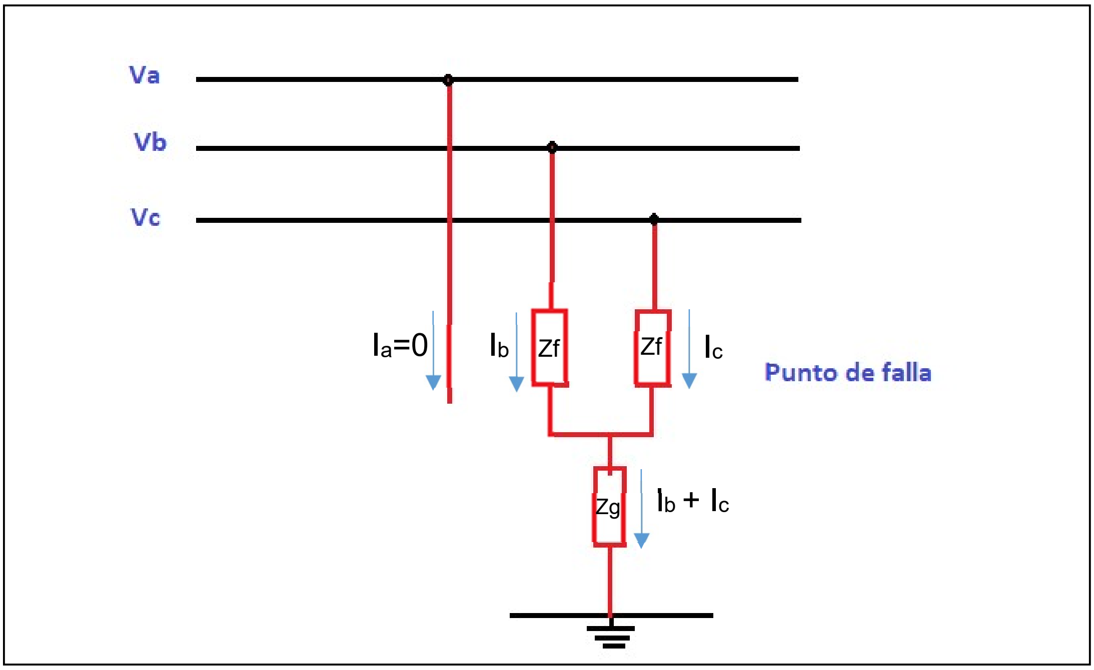


Figura N° 2.7: Diagrama de falla bifásica a tierra (Fuente: Propia del autor)

De la **Figura N° 2.7**, se deducen las siguientes ecuaciones:

$$V_b = (Z_f + Z_g) \cdot I_b + Z_g \cdot I_c \quad \ldots \ldots (2.31)$$

$$V_c = (Z_f + Z_g) \cdot I_c + Z_g \cdot I_b \quad \ldots \ldots (2.32)$$

$$I_a = 0 \quad \ldots \ldots (2.33)$$

A partir de la ecuación (2.8), se calcula $V_b - V_c$:

$$V_b - V_c = j\sqrt{3}(V_{a1} - V_{a2}) \quad \ldots \ldots (2.34)$$

De las ecuaciones (2.31) y (2.32) se calcula también $V_b - V_c$:

$$V_b - V_c = Z_f(I_b - I_c) \quad \ldots \ldots (2.35)$$

De la ecuación (2.11), se calcula $I_b - I_c$:

$$I_b - I_c = j\sqrt{3}(I_{a1} - I_{a2}) \quad \ldots \ldots (2.36)$$

Reemplazamos las ecuaciones (2.34) y (2.36) en (2.35):

$$V_{a1} - Z_f I_{a1} = V_{a2} - Z_f I_{a2} \quad \ldots \ldots (2.37)$$

De la ecuación (2.8) se puede calcular lo siguiente:

$$V_b + V_c = 2V_{a0} - (V_{a1} + V_{a2}) \quad \ldots \ldots (2.38)$$

De la suma de las ecuaciones (2.31) y (2.32) se obtiene:

$$V_b + V_c = Z_f[2I_{a0} - (I_{a1} + I_{a2})] + Z_g[4I_{a0} - 2(I_{a1} + I_{a2})] \quad \ldots \ldots (2.39)$$

De la resolución de las ecuaciones (2.38) con (2.39), se obtiene:

$$V_{a0} - Z_f I_{a0} - 3Z_g I_{a0} = V_{a1} - Z_f I_{a1} \quad \ldots \ldots (2.40)$$

De la ecuación (2.37) se deduce que, si se le agrega una impedancia de falla a la secuencia positiva y negativa, estas son iguales (conexión en paralelo). A su vez, de la ecuación (2.40) se deduce que si a la secuencia cero se le agrega la impedancia $Z_f + 3Z_g$, será idéntica a la ecuación (2.37), por lo que se debe conectar en paralelo, tal como se indica en la **Figura N° 2.8**:

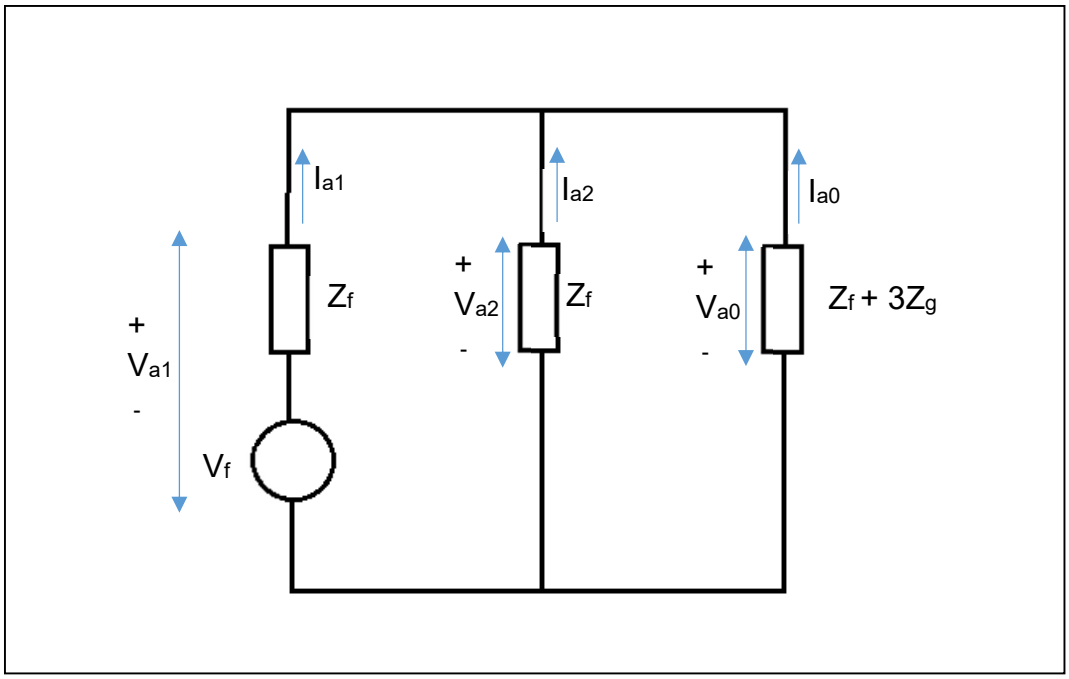


Figura N° 2.8: Conexión red secuencia falla bifásica a tierra (Fuente: Propia del autor)

De la Figura N° 2.8, se deduce la ecuación (2.41):

$$I_{a1} = \frac{V_f}{Z_1 + Z_f + \frac{(Z_2 + Z_f)(Z_0 + Z_f + 3Z_g)}{Z_0 + Z_2 + 2Z_f + 3Z_g}} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2.41)$$

d) Falla trifásica

Una falla trifásica al ser simétrica carece de componentes de secuencia negativa y cero, por ende, solo contiene componente de secuencia positiva.

La corriente de falla se calcula con la ecuación (2.42):

$$I_{a1} = I_a = \frac{V_f}{Z_1 + Z_f} \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots . \ldots . (2.42)$$

2.2.4 Método de cálculo de fasores de tensión y corriente

Rahman (1988) realizó un estudio comparativo de algoritmos para la protección digital diferencial de transformadores de potencia. El estudio concluye que los algoritmos transformada discreta de Fourier, filtro de respuesta de impulso finito, transformada rectangular y filtro de mínimos cuadrados son capaces de distinguir entre un evento de falla interna y evento de corriente inrush. En el caso de una falla interna, los algoritmos proveen una decisión confiable de disparo dentro de un ciclo desde el instante de ocurrencia de la falla interna.

Transformada discreta de Fourier

TRANSENER (2002) indicó un método de cálculo de las cantidades fasoriales tomando como base la Transformada discreta de Fourier, el cual se tomó como referencia.

Sea $\boldsymbol{s}$ una función periódica de entrada (onda de tensión o corriente) de periodo N y $\boldsymbol{s(n)}$ como el valor de la muestra $\boldsymbol{n}$ registrada. En las ecuaciones (2.43) y (2.44) se definen el cálculo de las componentes de seno y coseno, en el instante de la muestra $\boldsymbol{k}$:

$$S_{seno} = \frac{\sqrt{2}}{N} \sum_{p=1}^{N} s(k-p+1) \cdot \sin\left(\frac{2\pi p}{N}\right) \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots . (2.43)$$

$$S_{coseno} = \frac{\sqrt{2}}{N} \sum_{p=1}^{N} s(k-p+1) \cdot \cos\left(\frac{2\pi p}{N}\right) \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots (2.44)$$

La cantidad $\boldsymbol{N}$ se define como la cantidad de muestras registradas en un periodo de 1 ciclo.

A partir de las ecuaciones (2.43) y (2.44) se define el fasor $\boldsymbol{S_k}$ para la estampa de tiempo correspondiente a la muestra $\boldsymbol{k}$:

$$S_k = S_{seno} + jS_{coseno} \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots . (2.45)$$

2.2.5 Machine Learning

Khana y Awad (2015) definieron el Machine Learning como una rama de la inteligencia artificial que sistemáticamente aplica algoritmos para sintetizar las relaciones subyacentes entre datos e información. Indicaron que Arthur Samuel en 1959 definió el Machine Learning como un campo de estudio que dota a las computadoras la habilidad para aprender sin ser explícitamente programada. El Machine Learning tiene la habilidad de caracterizar relaciones subyacentes dentro de una extensa matriz de datos de tal manera que permite solucionar problemas en analítica de datos, reconocimiento de comportamiento de patrones y evolución de información. Su característica computacional es generalizar la experiencia del entrenamiento y genera una hipótesis que estima la función objetivo. El objetivo del Machine Learning es predecir futuros eventos o escenarios que son desconocidos para la computadora.

Maini (2017) indica que la inteligencia artificial consiste en diseñar un agente inteligente que percibe su ambiente y toma decisiones para maximizar las oportunidades de lograr su objetivo. Los fundamentos de la inteligencia artificial incluyen matemáticas, lógica, filosofía, probabilidad, lingüística, neurociencia y teoría de decisión; estos campos se encuentran dentro de la inteligencia artificial. El algoritmo de Machine Learning permite identificar patrones en un conjunto de datos observados, para la construcción de modelos que explican la realidad del

conjunto de datos y predicen sin tener reglas y modelos programados explícitamente para dicho fin.

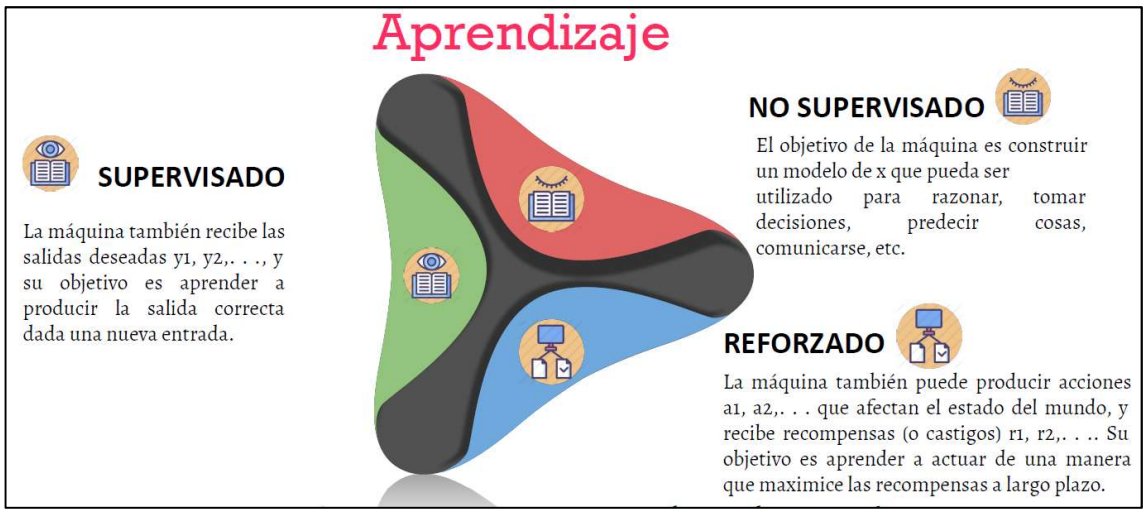


Figura N° 2.9: Tipos de Machine Learning (Fuente: [2])

En la **Figura N° 2.9** se observan los tipos de aprendizaje dentro de Machine Learning.

Aprendizaje supervisado

Rama del Machine Learning en la cual se tiene identificado el resultado de salida, a partir una información de entrada para el entrenamiento del modelo.

Algoritmo k-nearest neighbors

Khanna y Awad (2015), definieron el algoritmo k-nearest neighbors como una metodología de clasificación que identifica un grupo de k objetos en el conjunto de entrenamiento que están más cercanos al objeto del test y asigna una etiqueta basada en la clase más dominante. Los tres elementos fundamentales de este enfoque son:

- Un conjunto de objetos existente etiquetados.
- Una distancia métrica para calcular la distancia entre objetos.
- El número de vecinos más cercanos (k).

En el estudio citado indicaron que para realizar la clasificación de un objeto sin etiqueta, la distancia entre este y los objetos etiquetados son calculados y también se identifican los k-vecinos más cercanos (k-nearest neighbors). Las etiquetas de clase de los objetos vecinos más cercanos sirven como una referencia para la clasificación de los objetos sin etiqueta.

Métodos de cálculos de distancia en algoritmo k-nearest neighbors

Según Zijie Zhu (2020), el cálculo de las distancias se realiza de la siguiente manera:

- Distancia Euclideana

  Corresponde a la distancia más cercana entre dos puntos. Se calcula mediante la ecuación (2.46):

  $$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2.46)$$

- Distancia Manhattan

  En referencia a la calle Manhattan, donde el desplazamiento de un punto a otro se realizada a través de giros en ángulos rectos. Se calcula mediante la ecuación (2.47):

  $$d(x,y) = \sum_{i=1}^{n}|x_i - y_i| \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2.47)$$

- Distancia Chebyshev

  Corresponde al máximo valor del valor absoluto de la diferencia de las coordenadas entre dos puntos. Se calcula mediante la ecuación (2.48):

  $$d(x,y) = max_i|x_i - y_i| \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2.48)$$

Donde “n” es el número de características del conjunto de datos.

Análisis de componentes principales

Es un método de reducción de dimensionalidad de un conjunto de datos, realiza la transformación de un amplio conjunto de características a uno más reducido, el cual contiene la mayor cantidad de información del conjunto de datos.

Tipo de datos estadísticos

- Características categóricas: también denominados como cualitativos. Son clases que se excluyen mutuamente entre sí y pueden carecer o no de un orden lógico. Se subdividen en binario (1 ó 0) y nominal (elementos discretos).
- Características numéricas: tienen asignado un orden lógico. Se subdividen en discretos (no son posibles de medir, pero si de contar), ordinales (correspondiente a calificación) y continuos (representan mediciones).

Matriz de confusión

Una matriz que permite visualizar el desempeño de un algoritmo de clasificación de aprendizaje supervisado, para la comparación de la clasificación predicha contra la clasificación real a través de categorías como falso positivo (FP), verdadero positivo (VP), falso negativo (FN) y verdadero negativo (VN).

Tabla N° 2.1: Matriz de confusión.

		Predicho	
		Positivo	**Negativo**
Real	**Positivo**	Verdadero positivo (VP)	Falso Negativo (FN)
	Negativo	Falso positivo (FP)	Verdadero Negativo (VN)

Fuente: Elaboración propia.

Donde:

VP: resultado identificado correctamente como positivo.

VN: resultado identificado correctamente como negativo.

FP: resultado identificado incorrectamente como positivo.

FN: resultado identificado incorrectamente como negativo.

Indicadores de matriz de confusión

- **Exactitud:** en inglés "accuracy"**.** Es la tasa de las predicciones correctas realizadas por el modelo de Machine Learning. Hace referencia a la proximidad del resultado hacia el valor verdadero. En la ecuación (2.49) se observa la fórmula de cálculo para una matriz binaria:

$$Exactitud = \frac{VP + VN}{VP + VN + FP + FN} \dots \dots (2.49)$$

- **Precisión:** en inglés "precisión". Es la tasa de dispersión del conjunto de datos. Se gana más precisión en cuando la dispersión sea menor. En la ecuación (2.50) se observa la fórmula de cálculo para una matriz binaria:

$$Precisión = \frac{VP}{VP + FP} \dots \dots (2.50)$$

- **Sensibilidad:** en inglés "recall". Es la tasa de casos positivos que se identificaron correctamente por el modelo de Machine Learning. En la ecuación (2.51) se observa la fórmula de cálculo para una matriz binaria:

$$Sensibilidad = \frac{VP}{VP + FN} \dots \dots (2.51)$$

- **F1 score:** es utilizado cuando la cantidad de datos en las clases o categorías son desbalanceadas. En la ecuación (2.52) se observa la fórmula de cálculo para una matriz binaria:

$$F1\ score = \frac{2 * Sensibilidad * Precisión}{Sensibilidad + Precisión} \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots . (2.52)$$

Causas de fallas de origen externo en líneas de transmisión

Se indican las causas de origen externo, que se presentan con mayor cantidad de frecuencia en las fallas en líneas de transmisión:

a) Humedad y contaminación.

Las líneas de transmisión aéreas son vulnerables a la contaminación por parte del entorno externo, tales contaminaciones principalmente son del tipo industrial, marina y desértica. La contaminación de origen marino se presenta generalmente en zonas costeras del país, aunque por la acción del viento se puede prolongar su alcance a zonas más lejanas. La contaminación industrial se presenta por la operación de empresas de los rubros petroquímico, química, cementera, ladrilleras, metalúrgica, entre otros, las cuales emiten las partículas contaminación. En zonas desérticas, se facilita la contaminación de los aisladores con partículas de polvo por acción del viento y la ausencia de vegetación.

El aumento de la humedad relativa del ambiente también facilita la reducción del nivel de aislamiento de la cadena de aisladores, ocasionando una falla en la línea de transmisión, por humedecer la capa de partículas contaminantes.

El proceso de la descarga eléctrica en la cadena de aisladores por contaminación se da por la formación de una capa con las partículas contaminantes, luego con el aumento de la humedad relativa del medio ambiente, aumenta la conductividad de la capa de partículas, con el consecuente incremento de la corriente de fuga. Posteriormente producto del flujo de corriente de fuga en cadena de aisladores, la capa de partículas contaminantes se seca formando una banda seca, a través de la cual se expandirá la corriente de fuga, ampliando la banda seca y concluyendo en una descarga a través de toda la cadena de aisladores.

En la **Figura N° 2.10** se observa una cadena de aislador contaminado:

Figura N° 2.10: Cadena de aislador contaminado (Fuente: [3])

b) Quema de vegetación.

Se presentan cuando se realiza quema de caña en la zona costera o cuando hay incendio en zonas de vegetación aledañas a la línea de transmisión.

En la Figura N° 2.11, se observa una falla de causa raíz de quema de caña en L-2232.

Figura N° 2.11: Quema de caña en L-2232 que ocasionó falla (Fuente: [4]).

c) Descarga atmosférica.

Es una causa raíz de falla de origen natural. La ocurrencia de las descargas atmosféricas en las líneas de transmisión depende de la condición climática y de la zona geográfica de la línea de transmisión.

2.3 Conceptual

Manejo del aplicativo basado en algoritmo de clasificación de Machine Learning que identifica la causa raíz de fallas en líneas de transmisión

En la **Figura N° 2.13** se observa el diagrama de flujo del aplicativo basado en algoritmo de clasificación de Machine Learning para clasificación de causa raíz de falla en líneas de transmisión:

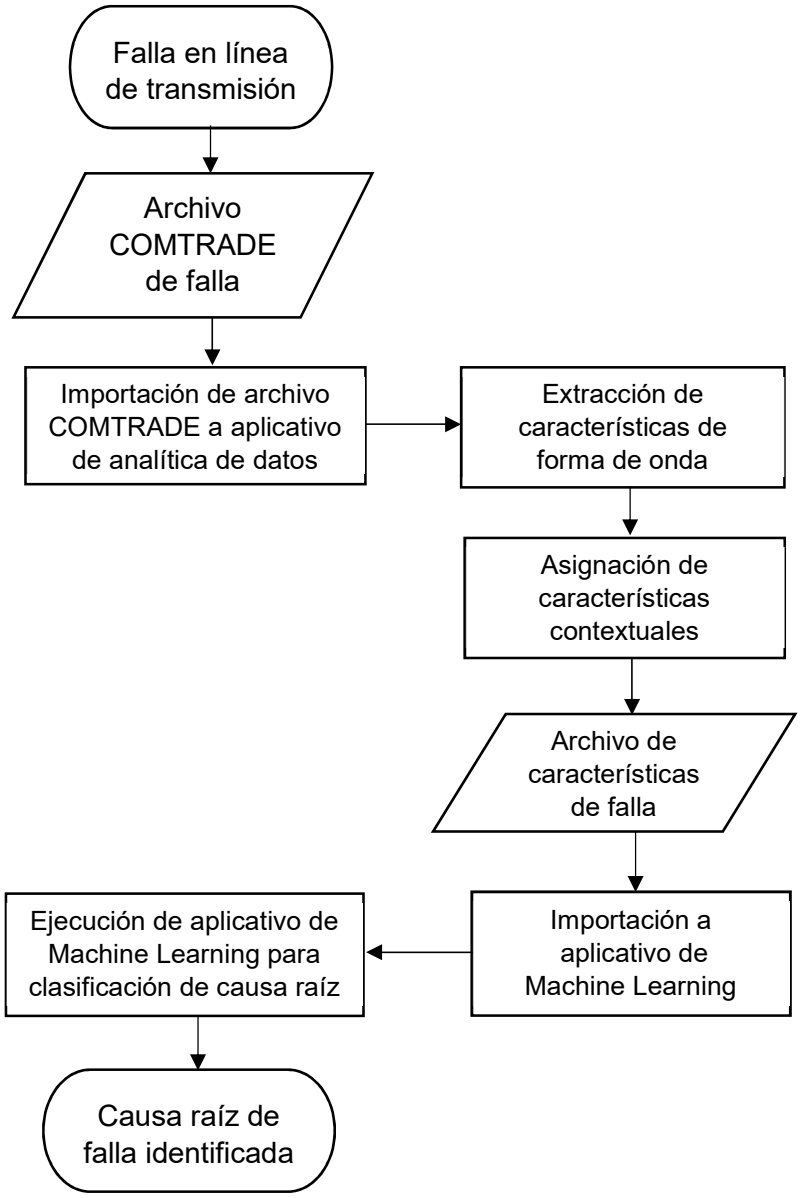


Figura N° 2.13: Diagrama de flujo de aplicativo de Machine Learning.

2.4 Definición de términos básicos

COMTRADE: De sus siglas en inglés “Commom Format for Transient Data Exchange” (formato común para intercambio de datos transitorios). Es un formato estandarizado con la norma IEEE Std. C37.111, la cual define un formato común, para el almacenamiento de eventos transitorios del sistema eléctrico.

Archivo “.DAT”: Archivo (necesario) definido en la norma IEEE Std. C37.111. Contiene los valores de datos escalados del evento muestreado de señales analógicas y digitales del evento transitorio, en formato Binario o ASCII. Por cada muestra también se tiene el número de muestra y la estampa de tiempo.

Archivo “.CFG”: Archivo (necesario) definido en la norma IEEE Std. C37.111. Está en formato ASCII y contiene los siguientes datos: nombre de subestación, año de revisión de estándar COMTRADE, cantidad total de canales analógicos y digitales, cantidad de canales analógicos, cantidad de canales digitales, nombres de canales (con unidades, relación de transformación y tipo de valores primarios o secundarios), frecuencia, tasa de muestreo, estampa de tiempo de la primera muestra, estampa de tiempo del disparo del relé de protección, tipo de archivo de datos (binario o ASCII) y factor multiplicador de estampa de tiempo.

Archivo “.HDR”: Archivo (opcional) definido en la norma IEEE Std. C37.111. El archivo está en formato ASCII y contiene información suplementaria provista para el usuario final, para una mejor comprensión del registro del evento de falla.

Oscilografía: conjunto de archivos basados en el estándar COMTRADE.

Falla: Condición transitoria del sistema eléctrico de potencia, caracterizada por altas magnitudes de corriente que afectan de manera importante la vida útil de los equipos del sistema eléctrico de potencia.

Sistema eléctrico de potencia: Conjunto de equipos primarios y secundarios que participan en el proceso de generación, transmisión y distribución de la energía eléctrica.

Relé de protección: dispositivo que monitorea continuamente las señales analógicas de corriente y tensión a fin de detectar una condición de falla en la zona de protección para aislarlo del sistema eléctrico, mediante un envío de señal de disparo al(los) interruptor(es), con la finalidad de minimizar el daño a los equipos del sistema eléctrico de potencia.

Confiabilidad: Probabilidad de que el sistema de protección actuará correctamente cuando sea requerido. La confiabilidad tiene dos enfoques: el sistema debe operar en la presencia de una falla dentro de su zona de protección (dependabilidad) y no debe operar para fallas presentes fuera de su zona de protección o en ausencia de falla en su zona de protección (seguridad).

Sensitividad: habilidad del sistema de protección el cual consiste en la detección de mínimo cambio por encima del valor límite (pick-up) de una función de protección.

Selectividad: habilidad del sistema de protección el cual consiste en mantener la máxima continuidad del suministro de energía eléctrica, a través del aislamiento de solo aquellos componentes afectados por una falla, del sistema eléctrico de potencia.

Interruptor de potencia: equipo que a través de sus contactos internos establece e interrumpe el flujo de corriente eléctrica en condición de régimen

permanente del sistema eléctrico. El interruptor también es capaz de interrumpir corrientes de falla.

Zona de protección: término utilizado en sistemas de protección para hacer referencia a que un conjunto de equipos forma parte del alcance (selectividad) de monitoreo de señales de corriente y tensión por parte de los relés de protección para el despeje de fallas en el menor tiempo posible.

Resistencia de falla: es la resistencia que se presenta en el punto de falla debido la caída de tensión a través de un arco eléctrico o debido a otra resistencia en el lazo de falla.

Lazo de falla: Se denomina a la impedancia presentada en condición de falla en las fases afectadas. Puede ser el tipo fase-fase, fase-neutro y 3-fases.

Líneas de transmisión: Conjunto de equipos que permiten la transmisión de energía eléctrica e interconecta subestaciones eléctricas. Están conformados por estructuras de soporte (metálica o madera), conductores de aleación de aluminio, cadena de aisladores de anclaje o suspensión, puesta a tierra y cable de guarda.

Subestación eléctrica: Nodo en el que se conectan las líneas de transmisión. Está conformado por equipos primarios (interruptor de potencia, seccionadores de línea, seccionador de barra, seccionador de tierra, seccionador de enlace, transformador de corriente, transformador de tensión, transformador de potencia, autotransformador, transformador zig-zag, banco de condensadores, compensador estático de potencia reactiva, reactor de línea, reactor de barra), equipos secundarios (relé de protección, registrador de falla, contador de

energía, GPS, controlador de bahía, servidor SAS, RTU, equipo de onda portadora, multiplexor) y equipos de servicios auxiliares (banco de baterías y grupo electrógeno).

Transformador de corriente: equipo primario que reproduce proporcionalmente los valores de corriente primaria (lado de alta tensión) en su lado secundario en base a una relación de transformación, a un nivel en que puedan manejar los equipos de protección, control y medición.

Transformador de tensión: equipo primario que reproduce proporcionalmente los valores de tensión primaria (lado de alta tensión) en su lado secundario en base a una relación de transformación, a un nivel en que puedan manejar los equipos de protección, control y medición.

Componentes simétricas: sistema mediante el cual un sistema desbalanceado trifásico, se puede descomponer en 3 sistemas balanceados: componente de secuencia positiva, componente de secuencia negativa y componente de secuencia cero.

Componente de secuencia positiva: sistema trifásico de fasores con la misma magnitud y desfasado en 120°, que giran en sentido antihorario.

Componente de secuencia negativa: sistema trifásico de fasores con la misma magnitud y desfasado en 120°, que giran en sentido horario.

Componente de secuencia cero: sistema trifásico de fasores con la misma magnitud y en fase, que no giran.

Tasa de muestreo: es la cantidad de muestras registradas por unidad de tiempo. Las muestras se capturan de señales continuas en un instante de tiempo para generar señales discretas.

Estampa de tiempo: dato que contiene la fecha, hora, minuto y segundo, generalmente asociado a una señal digital o analógica (cantidad discreta).

Forma de onda: conjunto de señales discretas que representan las magnitudes de las señales de corriente y tensión.

Característica de forma de onda: Parámetro cuantitativa que identifica una particularidad de la forma de onda.

Causa raíz de falla: se denomina causa raíz a aquello que originó la condición de falla en la línea de transmisión.

Grupo ISA Perú: grupo empresarial conformada por las filiales Red de Energía de Perú, Consorcio Transmantaro e ISA Perú, los cuales tienen en concesión líneas de transmisión y subestaciones para la administración de la operación y mantenimiento de sus equipos.

Algoritmo: Conjunto de comandos escritos en un determinado lenguaje de programación, los cuales tienen un sentido lógico, para la ejecución de una acción.

Instancia: Un objeto caracterizado por un vector de características del cual el modelo es entrenado para uso de predicción o generalización.

Clasificador: Método que recepciona una nueva entrada como una instancia sin etiqueta de una observación o característica y lo clasifica en una clase o categoría a la cual pertenece.

Conjunto de datos: Colección de datos donde cada columna corresponde a un atributo y cada fila corresponde a un registro del conjunto de datos.

Muestra: también denominado ejemplo. Es cada fila del conjunto de datos.

Característica: también denominado variable. Es cada columna del conjunto de datos.

Dimensión: Conjunto de atributos o variables que define una propiedad. Las funciones principales de la dimensión son el filtrado, clasificación y agrupamiento.

Matriz de confusión: Una matriz que permite visualizar el desempeño de un algoritmo de clasificación, para la comparación de la clasificación predicha contra la clasificación real a través de categorías como falso positivo, verdadero positivo, falso negativo y verdadero negativo.

Modelo: Estructura que resume un conjunto de datos para clasificación o descripción.

Hiperparámetro: es un parámetro empleado para controlar el proceso de aprendizaje.

Overfitting: consiste cuando el modelo de clasificación se centró demasiado en aprender del conjunto de datos de entrenamiento, que cuando se lo evalúa con nuevos datos, clasifica de manera errónea.

III. HIPÓTESIS Y VARIABLES

3.1 Hipótesis

Las hipótesis se detallan a continuación:

3.1.1 Hipótesis general

La implementación de un aplicativo basado en algoritmo de Machine Learning identificará automáticamente la causa raíz de fallas en líneas de transmisión del grupo ISA Perú.

3.1.2 Hipótesis específicas

H.E.1 El desarrollo de un aplicativo basado en lenguaje de programación Python, extraerá las características contextuales, características en el dominio del tiempo y de la frecuencia de los archivos COMTRADE de las fallas en líneas de transmisión.

H.E.2 El entrenamiento de un modelo de Machine Learning basado en algoritmo de clasificación de aprendizaje supervisado, identificará automáticamente la causa raíz de fallas en líneas de transmisión del grupo ISA Perú.

3.2 Definición conceptual de variables

Se describe de la siguiente manera:

X: Aplicativo basado en algoritmo de clasificación de Machine Learning

Y: Causa raíz de falla en líneas de transmisión

3.3 Operacionalización de variables

Por su naturaleza, todas las variables identificadas son del tipo cualitativo. Por su dependencia, la variable Y es dependiente y la variable X es independiente.

Tabla N° 3.1: Operacionalización de variables.

Variables	Dimensiones	Indicadores
Independiente: Aplicativo basado en algoritmo de Machine Learning	Aplicativo para la extracción de características de forma de onda de archivos COMTRADE	a) Valores de características de forma de onda de tensión b) Valores de característica de forma de onda de corriente
	Asignación de características contextuales correspondientes al tiempo de ocurrencia de falla	a) Hora de evento b) Temporada del año c) Región
Dependiente: Causa raíz de falla en líneas de transmisión.	Aplicativo para la clasificación automática de causa raíz de fallas en líneas de transmisión	a) Exactitud de matriz de confusión b) Precisión de matriz de confusión c) Sensibilidad de matriz de confusión d) F1 score de matriz de confusión
	Causa raíz de falla identificada	a) Exactitud de matriz de confusión b) Precisión de matriz de confusión c) Sensibilidad de matriz de confusión d) F1 score de matriz de confusión

Fuente: Elaboración propia.

IV. DISEÑO METODOLÓGICO

4.1 Tipo y diseño de investigación

4.1.1 Tipo de investigación

El tipo de investigación es aplicada, por el motivo de que parte de una base de teórica sobre análisis de componentes simétricas de fallas y algoritmo de clasificación de Machine Learning, para la implementación de un programa que identifique automáticamente la causa raíz de la falla en la línea de transmisión.

4.1.2 Diseño de la investigación

El diseño de la investigación es del tipo experimental, retrospectivo y longitudinal. Es del tipo experimental cuasi experimental debido a que se realizará un análisis de las características de la forma de onda de tensión y corriente de falla, para que el algoritmo de Machine Learning asocie los patrones identificados en las características a una causa raíz de falla. Es del tipo retrospectivo y longitudinal debido a que se utilizará una base de datos de fallas de las líneas de transmisión en concesión al grupo ISA Perú, en una ventana de tiempo del 2020 al 2021.

4.2 Método de investigación

El presente trabajo de investigación fue diseñado bajo el planteamiento metodológico del enfoque cuantitativo y analítico.

4.3 Población y muestra

4.3.1 Población del estudio

La población del estudio está conformada por las líneas de transmisión en niveles de tensión de 138, 220 y 500 kV en concesión del grupo ISA Perú.

4.3.2 Muestra del estudio

Debido a la naturaleza del estudio, se requiere que el modelo de Machine Learning realice el entrenamiento con la mayor cantidad de eventos de falla, por lo que la muestra corresponde a la cantidad total de la población. La muestra está conformada por los archivos COMTRADE (oscilografía), capturada por los relés de protección por la falla en la línea de transmisión, las cuales son el objeto de análisis.

4.4 Lugar de estudio

El presente estudio se realizará en el área de Mantenimiento de Líneas de Transmisión de Red de Energía del Perú S.A., en la cual gestiona información sobre la causa raíz confirmada de eventos de fallas en líneas de transmisión.

4.5 Técnicas e instrumentos para la recolección de datos

El plan de recolección de datos se elaborará en base a lo siguiente:

- Fuentes teóricas

Se utilizó técnicas documentales para la relación de datos bibliográficos sobre método de componentes simétricas, transformada discreta de Fourier, algoritmo de aprendizaje supervisado de Machine Learning, publicados físicamente y en formato digital.

- Método para la recolección de datos

Se recolectó una base de datos de fallas de causa raíz de origen externo en líneas de transmisión del grupo ISA Perú.

- Medición de variables

Para la medición de variable dependiente, se utilizarán las métricas de la matriz de confusión: exactitud, precisión, sensibilidad y F1 score.

4.6 Análisis y procesamiento de datos

El problema de identificación (clasificación) de causa raíz en líneas de transmisión es definido como un problema multiclase con 3 clases.

4.6.1 Recolección de datos

A. Recopilación de archivos COMTRADE de fallas en líneas de transmisión

El área de Protecciones del Departamento de Operaciones de Red de Energía del Perú S.A., cuenta con un portal web, en la que se almacena los archivos COMTRADE de las fallas en las líneas de transmisión, este repositorio está organizado por año, filial del grupo ISA Perú, línea de transmisión y fecha de ocurrida la falla.

Tabla N° 4.1: Cantidad de archivos COMTRADE por causa raíz de falla.

Causa raíz	2010 - 2019	2020	2021	Sub-total
Descarga atmosférica	-	58	67	125
Humedad y contaminación	-	11	78	89
Quema de vegetación	56	18	17	91
			Total	305

Fuente: Elaboración propia.

En la Tabla N° 4.1, se observan la cantidad de archivos COMTRADE de las fallas en la que se confirmó la causa raíz de la falla en la línea de transmisión correspondió a “descarga atmosférica”, “quema de vegetación” y “humedad y contaminación”. También se observa que con respecto a la

causa raíz de quema de vegetación, se cuenta con 35 fallas en los años 2020 y 2021, esta cantidad es insuficiente con respecto a la cantidad de archivos COMTRADE de las causas descarga atmosférica y humedad y contaminación, para que el modelo de Machine Learning realice el proceso de aprendizaje. Por ello se ha recopilado archivos COMTRADE de fallas del periodo correspondiente del 2010 al 2019, para la causa raíz de quema de vegetación. De esta manera, se logró recopilar una cantidad de 91 archivos COMTRADE para la causa raíz de quema de caña.

B. Calidad de información

Se ha validado la correspondencia de los archivos COMTRADE con la causa raíz, tomando como sustento las siguientes fuentes de información:

- Informe fotográfico de inspección por falla en la línea de transmisión.
- Informe final de perturbaciones, elaborado por el área de Protecciones del Departamento de Operaciones.
- Base de datos de fallas de las Áreas de Mantenimiento de Líneas de Transmisión.

4.6.2 Preparación de datos

A. Ingeniería de variables

Minnaar (2015) consideró en su estudio, las características de tiempo y espacio en la que ocurrieron las fallas, denominadas características contextuales y características de corriente y tensión en el dominio de la frecuencia (fasores) para la identificación automática de la causa raíz de fallas en líneas de transmisión.

Wang (2015) consideró en su estudio características en el dominio del tiempo para el diagnóstico de fallas en rodamientos.

En el presente estudio, se han agrupado las variables bajo las siguientes clases:

- **Características contextuales:**

  Mes: correspondiente al tiempo de ocurrido la falla. El tipo de dato es numérico discreto.

  Hora: correspondiente al tiempo de ocurrido la falla. El tipo de dato es numérico discreto.

  code_día: correspondiente al horario diurno o nocturno en que se presentó la falla. Tipo de dato original es categórico nominal, se convirtió a numérico discreto: nocturno (1) y diurno (2).

  code_estación: correspondiente a la estación del año en la que ocurrió la falla. El tipo de dato original es categórico nominal, se convirtió a numérico discreto: verano (1), otoño (2), invierno (3), primavera (4).

  code_región: correspondiente a la zona geográfica predominantemente a la que corresponde la línea de transmisión. El tipo de dato es categórico nominal, se convirtió a numérico discreto: costa (1), sierra (2) y selva (3).

  tipo_falla: correspondiente al lazo de falla (falla monofásica a tierra, falla bifásica aislada, falla bifásica a tierra y falla trifásica). El tipo de dato es categórico nominal, se convirtió a numérico discreto: falla

monofásica a tierra (1), falla bifásica aislada (2), falla bifásica a tierra (3) y falla trifásica (4).

Tensión_kV: tensión nominal de línea a línea de la línea de transmisión. El tipo de dato es numérico discreto.

- **Variables en el dominio de la frecuencia:**

  dif_i0max: es el máximo cambio de corriente de secuencia cero, entre la muestra “n” y “n-1”. El tipo de dato es numérico continua.

  dif_i1max: es el máximo cambio de corriente de secuencia positiva, entre la muestra “n” y “n-1”. El tipo de dato es numérico continua.

  dif_i2max: es el máximo cambio de corriente de secuencia negativa, entre la muestra “n” y “n-1”. El tipo de dato es numérico continua.

  Vrel_max0: es la relación entre la máxima tensión de secuencia cero de la ventana de tiempo de falla, respecto al promedio de la tensión de secuencia cero de la ventana de tiempo de pre-falla. El tipo de dato es numérica continuo.

  Vrel_max2: es la relación entre la máxima tensión de secuencia negativa de la ventana de tiempo de falla, respecto al promedio de la tensión de secuencia negativa de la ventana de tiempo de pre-falla. El tipo de dato es numérico continuo.

  I0_half_cycle: magnitud de corriente de secuencia cero a medio ciclo desde la muestra inicial de falla. El tipo de dato es numérico continuo.

  I0_one_cycle: magnitud de corriente de secuencia cero a un ciclo desde la muestra inicial de falla. El tipo de dato es numérico continuo.

I0_one_cycle_half: magnitud de corriente de secuencia cero a un ciclo y medio desde la muestra inicial de falla. El tipo de dato es numérico continuo.

I0_two_cycle: magnitud de corriente de secuencia cero a dos ciclos desde la muestra inicial de falla. El tipo de dato es numérico continuo.

I2_half_cycle: magnitud de corriente de secuencia negativa a medio ciclo desde la muestra inicial de falla. El tipo de dato es numérico continuo.

I2_one_cycle: magnitud de corriente de secuencia negativa a un ciclo desde la muestra inicial de falla. El tipo de dato es numérico continuo.

I2_one_cycle_half: magnitud de corriente de secuencia negativa a un ciclo y medio desde la muestra inicial de falla. El tipo de dato es numérico continuo.

I2_two_cycle: magnitud de corriente de secuencia negativa a dos ciclos desde la muestra inicial de falla. El tipo de dato es numérico continuo.

Irel_max0: es la relación entre la máxima corriente de secuencia cero de la ventana de tiempo de falla, respecto al promedio de la corriente de secuencia cero de la ventana de tiempo de pre-falla. El tipo de dato es numérico continuo.

Irel_max1: es la relación entre la máxima corriente de secuencia positiva de la ventana de tiempo de falla, respecto al promedio de la corriente de secuencia positiva de la ventana de tiempo de pre-falla. El tipo de dato es numérico continuo.

Irel_max2: es la relación entre la máxima corriente de secuencia negativa de la ventana de tiempo de falla, respecto al promedio de la corriente de secuencia negativa de la ventana de tiempo de pre-falla. El tipo de dato es numérico continuo.

Const time T0: es el tiempo transcurrido desde el inicio de la falla hasta el tiempo correspondiente al valor que da como resultado del producto de 0.63 con la diferencia entre la máxima corriente de falla y la corriente de pre-falla de secuencia cero. El tipo de dato es numérico continuo.

Const time T1: es el tiempo transcurrido desde el inicio de la falla hasta el tiempo correspondiente al valor que da como resultado del producto de 0.63 con la diferencia entre la máxima corriente de falla y la corriente de pre-falla de secuencia positiva. El tipo de dato es numérico continuo.

Const time T2: es el tiempo transcurrido desde el inicio de la falla hasta el tiempo correspondiente al valor que da como resultado del producto de 0.63 con la diferencia entre la máxima corriente de falla y la corriente de pre-falla de secuencia negativa. El tipo de dato es numérico continuo.

- **Variables en el dominio de tiempo:**

  iR_rms: corresponde al valor de la raíz media cuadrática de la ventana de tiempo de falla de la señal de corriente de la fase “R”. Se calcula con la ecuación (4.1):

$$señal{-}phase_{rms} = \left[\frac{1}{n}\sum_{i=1}^{n} {x_i}^2\right]^{1/2} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (4.1)$$

Donde “n” es la cantidad de muestras de la ventana de tiempo de falla. El tipo de dato es numérico continuo.

iS_rms: corresponde al valor de la raíz media cuadrática de la ventana de tiempo de falla de la señal de corriente de la fase “S”. Se calcula con la ecuación (4.1). El tipo de dato es numérico continuo.

iT_rms: corresponde al valor de la raíz media cuadrática de la ventana de tiempo de falla de la señal de corriente de la fase “T”. Se calcula con la ecuación (4.1). El tipo de dato es numérico continuo.

vR_rms: corresponde al valor de la raíz media cuadrática de la ventana de tiempo de falla de la señal de tensión de la fase “R”. Se calcula con la ecuación (4.1). El tipo de dato es numérico continuo.

vS_rms: corresponde al valor de la raíz media cuadrática de la ventana de tiempo de falla de la señal de tensión de la fase “S”. Se calcula con la ecuación (4.1). El tipo de dato es numérico continuo.

vT_rms: corresponde al valor de la raíz media cuadrática de la ventana de tiempo de falla de la señal de tensión de la fase “T”. Se calcula con la ecuación (4.1). El tipo de dato es numérico continuo.

iR_root: corresponde al valor de la raíz de la serie de tiempo de falla de la señal de corriente de la fase “R”. Se calcula con la ecuación (4.2):

$$señal{-}phase_{root} = \left[\frac{1}{n}\sum_{i=1}^{n} |x_i|^{1/2}\right]^2 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (4.2)$$

Donde “n” es la cantidad de muestras de la ventana de tiempo de falla.

El tipo de dato es numérico continuo.

iS_root: corresponde al valor de la raíz de la serie de tiempo de falla de la señal de corriente de la fase “S”. Se calcula con la ecuación (4.2). El tipo de dato es numérico continuo.

iT_root: corresponde al valor de la raíz de la serie de tiempo de falla de la señal de corriente de la fase “T”. Se calcula con la ecuación (4.2). El tipo de dato es numérico continuo.

vR_root: corresponde al valor de la raíz de la serie de tiempo de falla de la señal de tensión de la fase “R”. Se calcula con la ecuación (4.2). El tipo de dato es numérico continuo.

vS_root: corresponde al valor de la raíz de la serie de tiempo de falla de la señal de tensión de la fase “S”. Se calcula con la ecuación (4.2). El tipo de dato es numérico continuo.

vT_root: corresponde al valor de la raíz de la serie de tiempo de falla de la señal de tensión de la fase “T”. Se calcula con la ecuación (4.2). El tipo de dato es numérico continuo.

iR_std: corresponde al valor de la desviación estándar de la serie de tiempo de falla de la señal de corriente fase “R”. Se calcula con la ecuación (4.3):

$$señal{-}phase_{desv-std} = \left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - x_{abs-mean})^2\right]^{1/2} \quad \ldots\ldots\ldots\ldots\ldots (4.3)$$

Donde “n” es la cantidad de muestras de la ventana de tiempo de falla y $x_{abs-mean}$ es el valor promedio absoluto y se calcular con la ecuación (4.4):

$$x_{abs-mean} = \frac{1}{n}\sum_{i=1}^{n}|x_i| \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (4.4)$$

El tipo de dato es numérico continuo.

iS_std: corresponde al valor de la desviación estándar de la serie de tiempo de falla de la señal de corriente fase “S”. Se calcula con la ecuación (4.3). El tipo de dato es numérico continuo.

iT_std: corresponde al valor de la desviación estándar de la serie de tiempo de falla de la señal de corriente fase “T”. Se calcula con la ecuación (4.3). El tipo de dato es numérico continuo.

vR_std: corresponde al valor de la desviación estándar de la serie de tiempo de falla de la señal de tensión fase “R”. Se calcula con la ecuación (4.3). El tipo de dato es numérico continuo.

vS_std: corresponde al valor de la desviación estándar de la serie de tiempo de falla de la señal de tensión fase “S”. Se calcula con la ecuación (4.3). El tipo de dato es numérico continuo.

vT_std: corresponde al valor de la desviación estándar de la serie de tiempo de falla de la señal de tensión fase “T”. Se calcula con la ecuación (4.3). El tipo de dato es numérico continuo.

vR_fac_crest: corresponde al factor de cresta de la serie de tiempo de falla de la señal de tensión fase “R”. Se calcula con la ecuación (4.5):

$$señal{-}phase_{fac-crest} = \frac{x_{max}}{x_{rms}} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (4.5)$$

Donde x_{max} corresponde al máximo valor absoluto de la serie de tiempo de señal y se calcula con la ecuación (4.6) y x_{rms} se calcula con la ecuación (4.1):

$$x_{max} = max|x_i| \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots. (4.6)$$

El tipo de dato es numérico continuo.

vS_fac_crest: corresponde al factor de cresta de la serie de tiempo de falla de la señal de tensión fase “S”. Se calcula con la ecuación (4.5). El tipo de dato es numérico continuo.

vT_fac_crest: corresponde al factor de cresta de la serie de tiempo de falla de la señal de tensión fase “T”. Se calcula con la ecuación (4.5). El tipo de dato es numérico continuo.

iR_fac_imp: corresponde al factor de impulso de la serie de tiempo de falla de la señal de corriente fase “R”. Se calcula con la ecuación (4.7):

$$señal\text{—}phase_{fac-imp} = \frac{x_{max}}{x_{abs-mean}} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (4.7)$$

El tipo de dato es numérico continuo.

iS_fac_imp: corresponde al factor de impulso de la serie de tiempo de falla de la señal de corriente fase “S”. Se calcula con la ecuación (4.7). El tipo de dato es numérico continuo.

iT_fac_imp: corresponde al factor de impulso de la serie de tiempo de falla de la señal de corriente fase “T”. Se calcula con la ecuación (4.7). El tipo de dato es numérico continuo.

4.6.3 Extracción de datos y consolidación de base de datos

A. Aplicativo de analítica de datos para la extracción de características

Se implementó un aplicativo de analítica de datos, para la extracción de características contextuales, características en el dominio del tiempo y características en el dominio de la frecuencia de los archivos COMTRADE.

El aplicativo está basado en lenguaje de programación Python, utilizando el IDE Pycharm 2021.2.3 (Community Edition).

El procedimiento de uso es el siguiente:

- Ejecución del aplicativo desde el entorno de PyCharm.
- Importación del archivo COMTRADE:

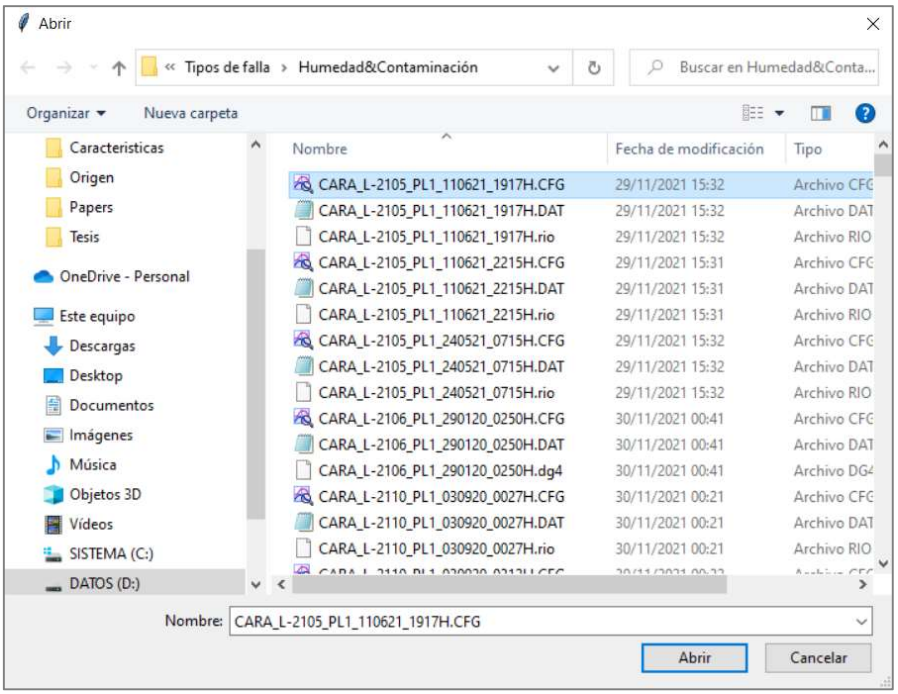


Figura N° 4.1: Importación del archivo .CFG (Fuente: propia del autor)

- Selección automática de muestra inicial y final de falla para generar el gráfico de la forma de onda de corriente de falla del archivo COMTRADE:

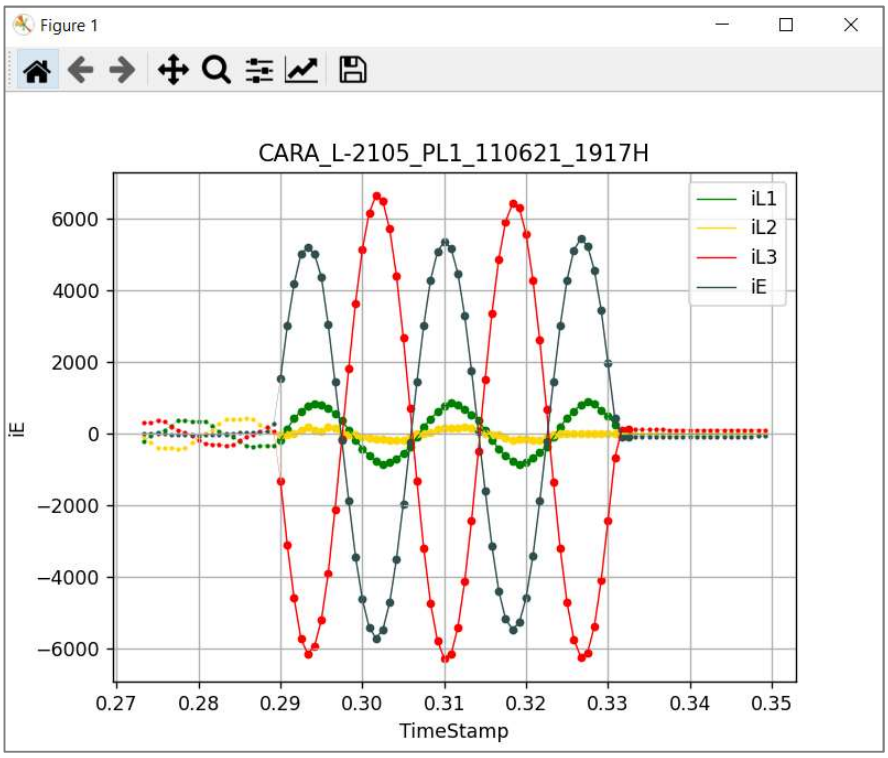


Figura N° 4.2: Corrientes de falla del archivo COMTRADE (Fuente: propia del autor)

- Extracción y exportación de características a formato Microsoft Excel (.XLSX)

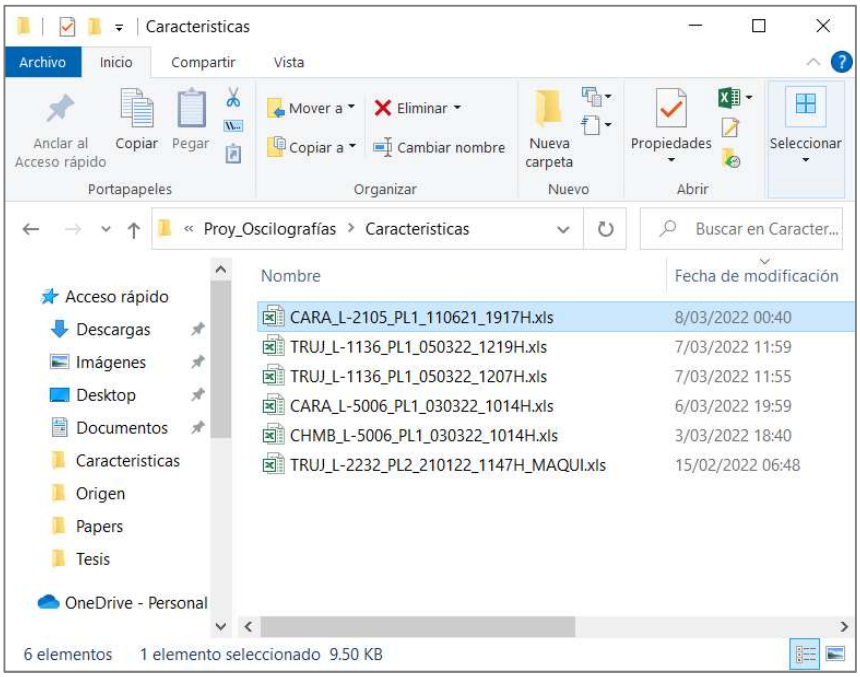


Figura N° 4.3: Exportación de características del archivo COMTRADE (Fuente: propia del autor)

B. Consolidación de base de datos

Se realizó la extracción de las características mediante el aplicativo de analítica de datos se realizó para los 305 archivos COMTRADE que conforman la muestra del estudio. Posteriormente se consolidó el conjunto de características de cada archivo COMTRADE en un archivo Microsoft Excel (.XLSX) y se agregó una columna adicional para indicar la causa raíz de falla. A partir de este conjunto de datos se realizó la exploración de datos y entrenamiento del modelo de Machine Learning.

C. Exploración de datos

- **Resumen de parámetros estadísticos de las características del conjunto de datos**

Se cargó la base de datos en Google Colaboratory para obtener los parámetros estadísticos del conjunto de datos, así como los percentiles 25%, 50% y 75%:

Tabla N° 4.2: Parámetros estadísticos del conjunto de datos

	count	mean	std	min	25%	50%	75%	max
Mes	305.0	6.65	3.48	1.00	4.00	7.00	10.00	12.00
Hora	305.0	11.59	6.14	0.00	6.00	13.00	16.00	23.00
code_día	305.0	1.61	0.49	1.00	1.00	2.00	2.00	2.00
code_estación	305.0	2.61	1.16	1.00	2.00	3.00	4.00	4.00
code_región	305.0	1.58	0.73	1.00	1.00	1.00	2.00	3.00
Tensión kV	305.0	233.02	93.60	138.00	220.00	220.00	220.00	500.00
tipo_falla	305.0	1.33	0.70	1.00	1.00	1.00	1.00	4.00
dif_i0max	305.0	62.20	75.83	0.26	19.30	39.58	67.81	444.65
dif_i1max	305.0	82.55	81.48	1.46	30.55	54.42	100.47	444.61
dif_i2max	305.0	83.82	78.57	3.24	33.85	59.10	105.90	451.15
Vrel_max0	305.0	183.77	172.77	0.91	73.50	126.44	244.23	1138.53
Vrel_max2	305.0	81.43	66.40	0.97	42.95	62.52	102.06	648.78
I0_half_cycle	305.0	348.73	524.90	0.51	66.69	164.65	414.29	4751.11
I0_one_cycle	305.0	724.23	1037.12	0.49	144.30	352.53	798.01	7130.56
I0_one_cycle_half	305.0	814.04	1161.71	0.71	165.92	394.47	912.52	8232.99
I0_two_cycle	305.0	806.78	1131.20	1.42	183.68	378.57	943.10	8378.16
I2_half_cycle	305.0	154.38	319.81	1.46	37.50	75.14	166.07	4076.16
I2_one_cycle	305.0	310.61	558.58	3.61	78.52	150.37	333.68	5456.74
I2_one_cycle_half	305.0	334.39	568.08	3.37	89.65	164.51	355.70	4927.81
I2_two_cycle	305.0	333.42	577.23	3.38	96.75	160.03	353.40	5606.47
Irel_max0	305.0	909.08	1222.94	3.25	244.61	469.91	1046.96	9143.87
Irel_max1	305.0	19.66	202.06	0.92	2.53	4.40	9.22	3532.30
Irel_max2	305.0	381.47	611.30	4.10	123.64	218.43	387.50	5606.47
Const time T0	305.0	0.39	0.29	0.06	0.20	0.30	0.50	2.00
Const time T1	305.0	0.37	0.29	0.00	0.20	0.30	0.50	2.00
Const time T2	305.0	0.39	0.29	0.06	0.20	0.30	0.50	2.00
iR_rms	305.0	942.28	1515.08	27.99	166.17	302.67	1155.39	10777.37
iS_rms	305.0	1046.34	1485.70	21.44	163.07	356.49	1329.26	10744.93
iT_rms	305.0	1013.32	1500.68	10.67	190.73	329.02	1366.67	11117.26
vR_rms	305.0	112807.77	58593.41	13511.58	75662.37	116247.62	129525.10	299822.40
vS_rms	305.0	108960.62	55856.41	9168.57	74544.87	114255.13	127885.45	294256.40

Fuente: elaboración propia.

Tabla N° 4.3: Parámetros estadísticos del conjunto de datos

	count	mean	std	min	25%	50%	75%	max
vT_rms	305.0	109408.28	55447.02	9245.85	71928.04	118672.15	129150.88	302364.59
iR_root	305.0	717.07	1154.60	16.09	123.07	234.20	920.15	8215.25
iS_root	305.0	798.91	1127.81	13.05	124.46	262.90	1020.57	8516.10
iT_root	305.0	766.61	1125.55	4.87	142.61	254.16	1074.14	8206.61
vR_root	305.0	91687.36	48143.29	6149.95	62006.46	94481.68	106688.97	248452.47
vS_root	305.0	88644.37	46023.41	5224.74	60940.76	93078.82	105387.57	242876.56
vT_root	305.0	88852.71	45808.28	5769.46	58081.78	97811.94	105594.66	247777.23
iR_std	305.0	922.51	1477.13	20.93	166.76	302.12	1158.37	10359.25
iS_std	305.0	1025.11	1440.22	21.50	163.06	353.48	1310.96	10484.48
iT_std	305.0	989.41	1447.57	10.63	186.46	326.28	1341.18	10422.72
vR_std	305.0	113164.00	58655.49	13529.56	75964.58	116768.84	130096.08	299375.26
vS_std	305.0	109308.92	56027.16	8731.50	74682.54	115014.36	128086.61	294847.38
vT_std	305.0	109763.76	55489.66	8169.58	72162.87	119045.42	129478.92	300872.59
vR_fac_crest	305.0	1.75	0.72	1.37	1.44	1.49	1.68	7.10
vS_fac_crest	305.0	1.76	0.75	1.39	1.45	1.51	1.68	6.47
vT_fac_crest	305.0	1.79	0.75	1.37	1.45	1.50	1.71	7.28
iR_fac_imp	305.0	2.11	0.60	1.54	1.78	2.00	2.26	9.48
iS_fac_imp	305.0	2.13	0.54	1.56	1.84	2.00	2.31	6.67
iT_fac_imp	305.0	2.13	0.57	1.54	1.84	2.01	2.25	9.19

Fuente: elaboración propia.

En la **Tabla N° 4.2** y **Tabla N° 4.3**, se pueden observar el valor promedio (mean), valor mínimo (min), valor máximo (max), desviación estándar (std) y percentiles 25%, 50% y 75% de las características contextuales, características en el dominio del tiempo y características en el dominio de la frecuencia.

- **Exploración de características por causa raíz**

A continuación, se utilizó el método de estadística descriptiva, diagrama de caja, para observar la distribución de los datos por cada causa raíz de falla:

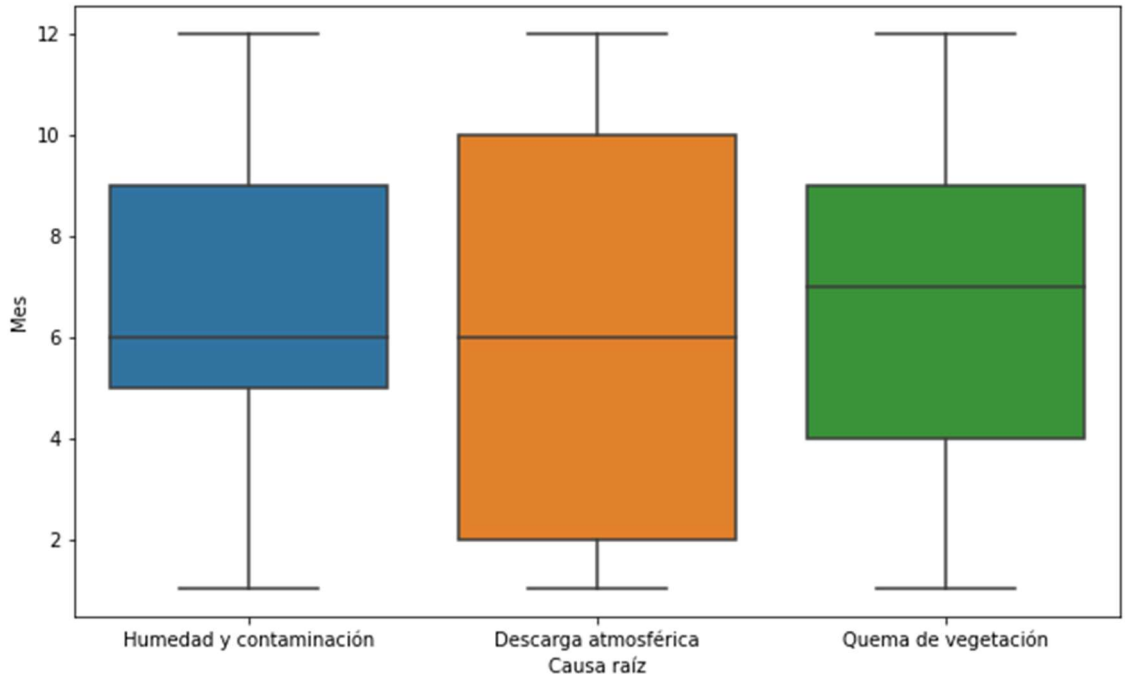


Figura N° 4.4: Diagrama de caja de la característica ‘mes’ (Fuente: propia del autor).

En la **Figura N° 4.4**, se observa que las fallas por humedad y contaminación se presentan predominantemente en los meses de mayo a septiembre. Las fallas por descarga atmosférica se presentan desde febrero hasta octubre y las fallas de quema de vegetación se presentan desde abril a septiembre.

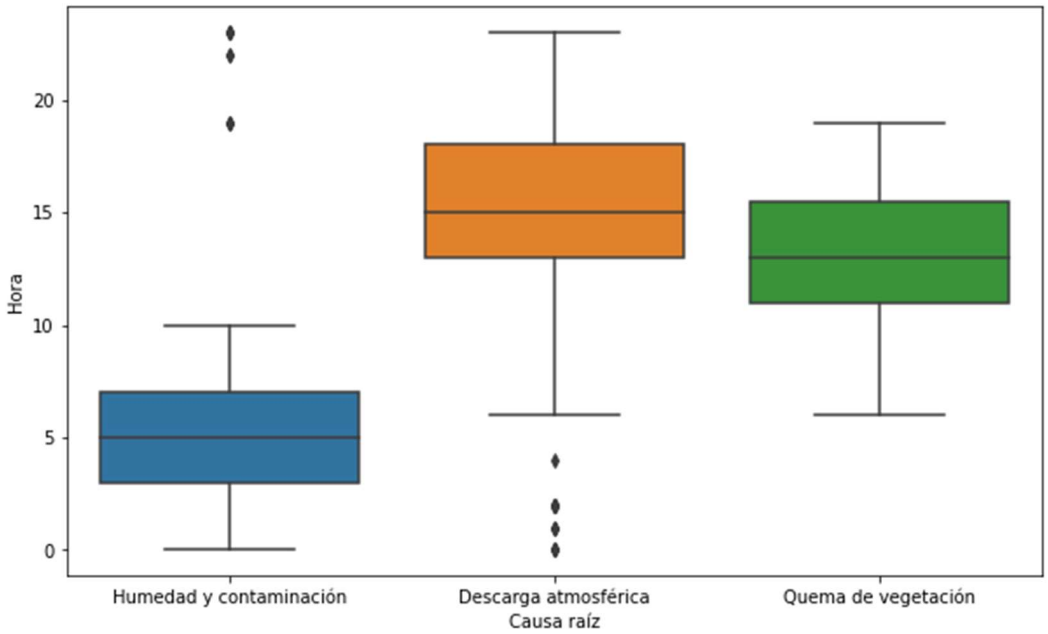


Figura N° 4.5: Diagrama de caja de la característica 'hora' (Fuente: propia del autor).

En la **Figura N° 4.5**, se observa que las fallas por humedad y contaminación se presentan entre las 03:00 y 07:00 horas. Las fallas por descarga atmosférica se presentan entre las 13:00 y 18:00 horas y las fallas de quema de vegetación se presentan entre las 12:00 y 16:00 horas.

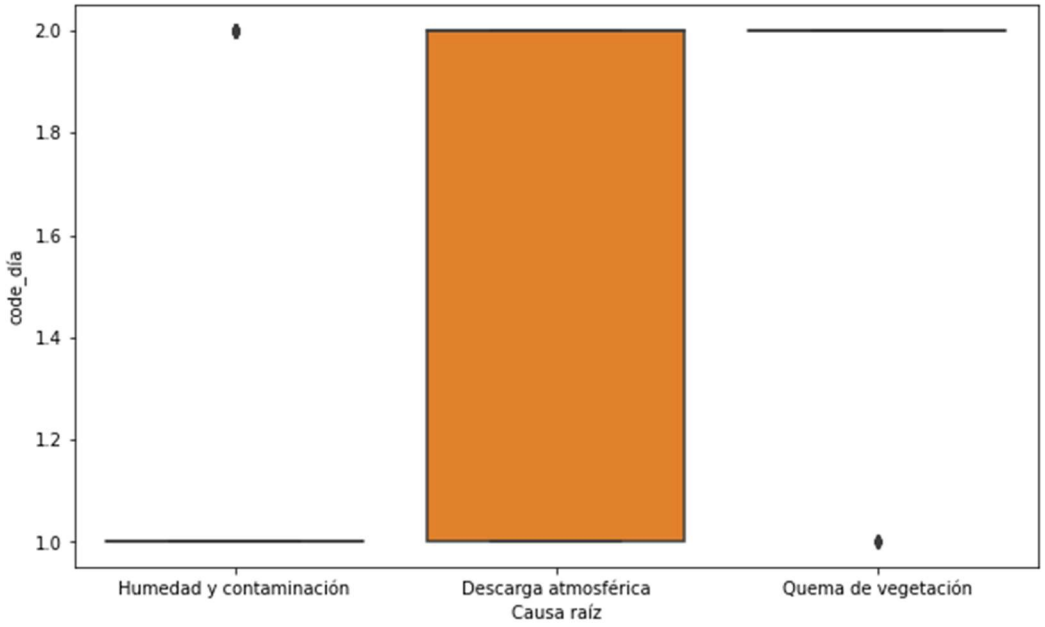


Figura N° 4.6: Diagrama de caja de la característica 'code_día' (Fuente: propia del autor).

En la **Figura N° 4.6**, se observa que las fallas por humedad y contaminación son correspondientes predominantemente al valor 1 (horario nocturno), las fallas por descarga atmosférica son correspondientes al valor 1 y 2 (horario diurno y nocturno) y las fallas de quema de vegetación son correspondientes al valor 2 (horario diurno).

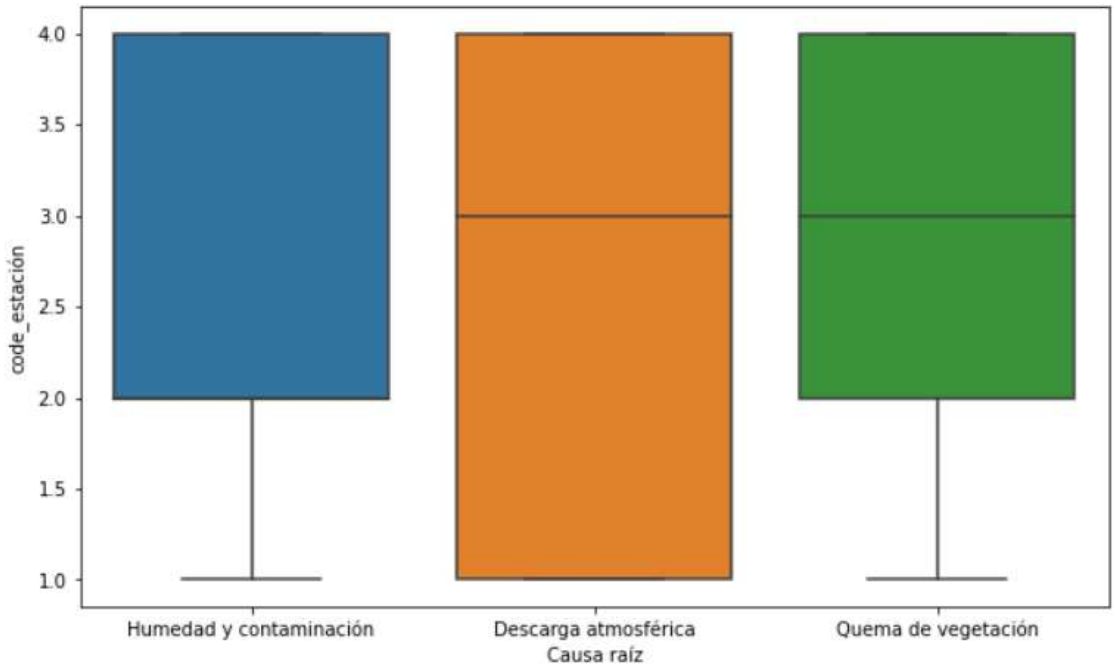


Figura N° 4.7: Diagrama de caja de la característica 'code_estación' (Fuente: propia del autor).

En la **Figura N° 4.7**, se observa que las fallas por humedad y contaminación se presentan predominantemente en valores del 2 al 4 (otoño, invierno y primavera), las fallas por descarga atmosférica se presentan en todo el año, pero predominantemente en el valor 3 (invierno). Las fallas de quema de vegetación se presentan predominantemente en valores del 2 al 4 (otoño, invierno y primavera).

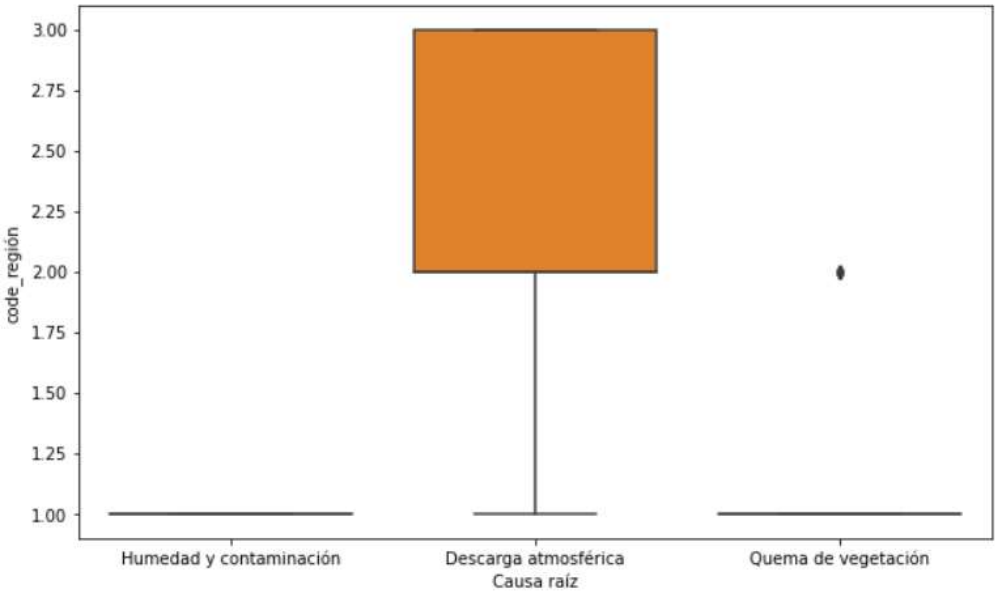


Figura N° 4.8: Diagrama de caja de la característica 'code_región' (Fuente: propia del autor).

En la **Figura N° 4.8**, se observa que las fallas por humedad y contaminación se presentan solo en el valor 1 (región costa), las fallas por descarga atmosférica se presentan en los valores 2 y 3 (región sierra y selva correspondientemente). Las fallas de quema de vegetación se presentan predominantemente en el valor 1 (región costa).

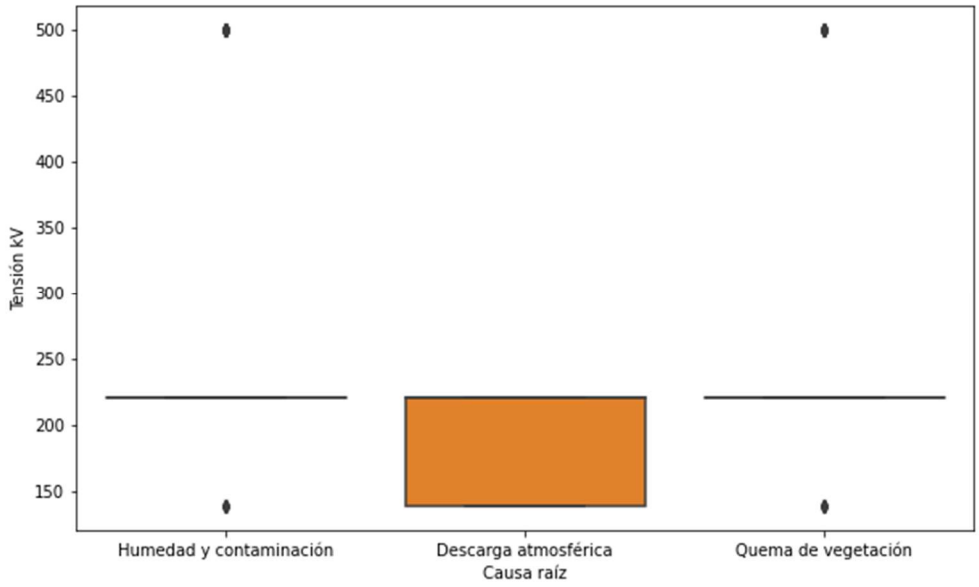


Figura N° 4.9: Diagrama de caja de la característica 'Tensión kV' (Fuente: propia del autor).

En la **Figura N° 4.9**, se observa que las fallas por humedad y contaminación se presentan predominantemente en el nivel de tensión de 220 kV, las fallas por descarga atmosférica se presentan en 138 y 220 kV. Las fallas de quema de vegetación se presentan predominantemente en 220 kV.

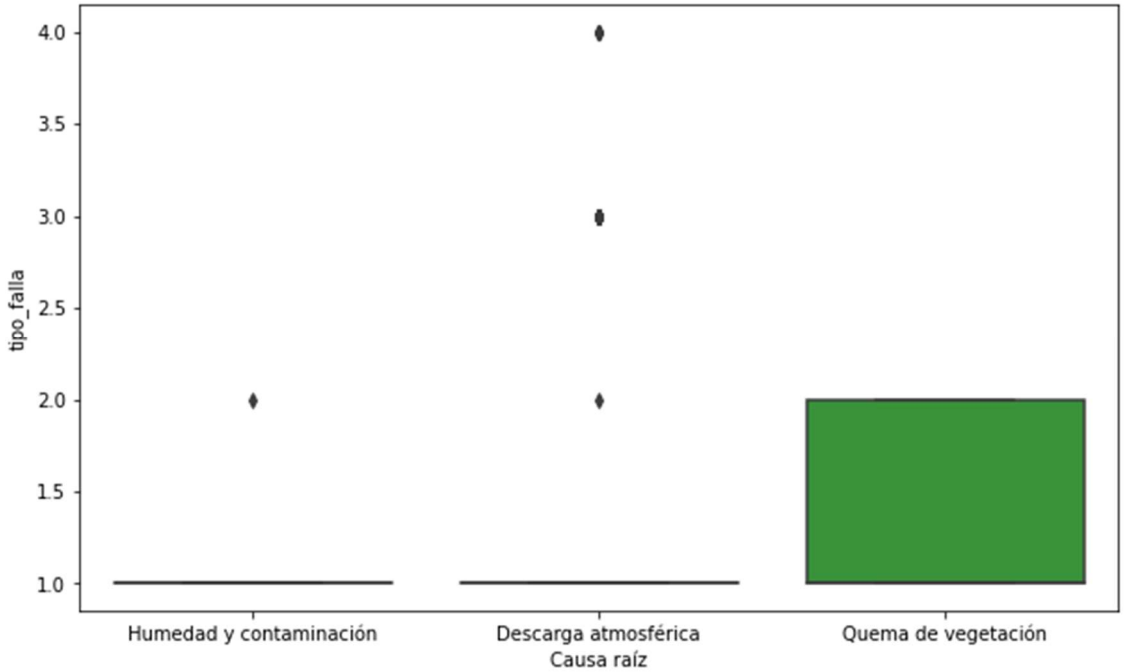


Figura N° 4.10: Diagrama de caja de la característica 'tipo_falla' (Fuente: propia del autor).

En la **Figura N° 4.10**, se observa que las fallas por humedad y contaminación se presentan predominantemente en el valor 1 (falla monofásica a tierra). Las fallas por descarga atmosférica se presentan predominantemente en el valor 1 (falla monofásica a tierra), Las fallas de quema de vegetación se presentan predominantemente en valores del 1 y 2 (falla monofásica a tierra y falla bifásica aislada).

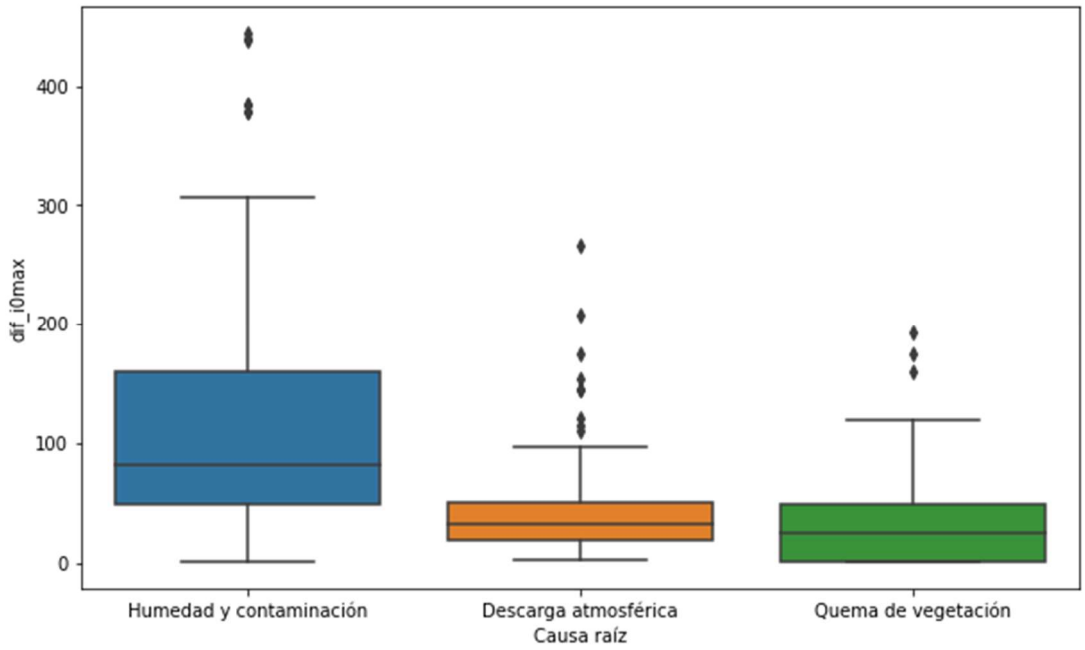


Figura N° 4.11: Diagrama de caja de la característica 'dif_i0max' (Fuente: propia del autor).

En la **Figura N° 4.11,** se verificó que la característica 'dif_i0max' permite diferenciar la causa raíz de humedad y contaminación de las otras causas raíces.

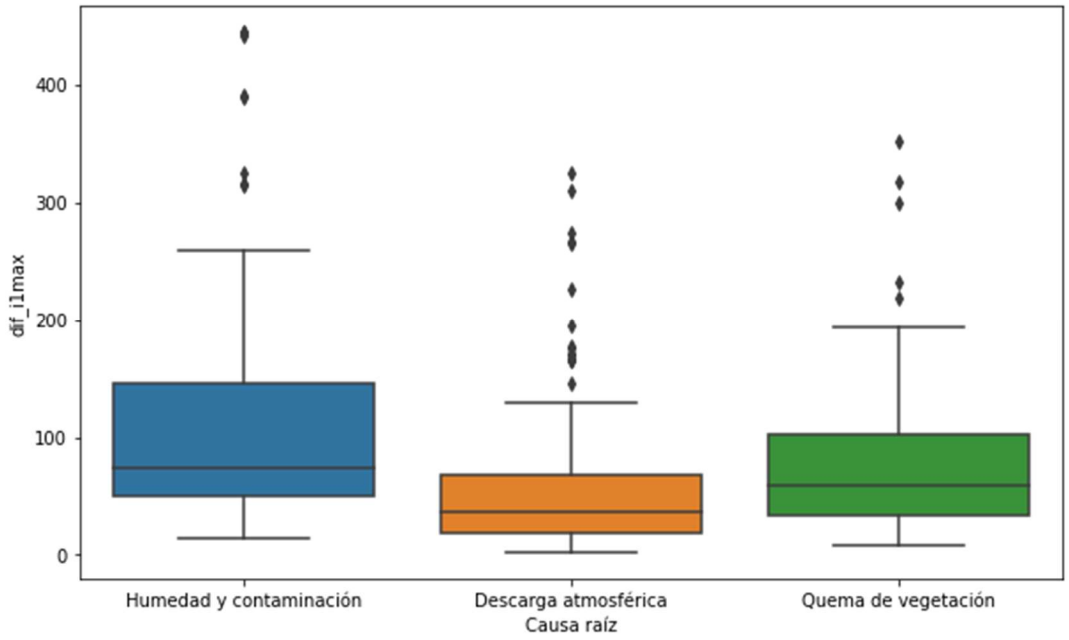


Figura N° 4.12: Diagrama de caja de la característica 'dif_i1max' (Fuente: propia del autor).

En la **Figura N° 4.12,** se verificó que la característica ‘dif_i1max’ permite diferenciar moderadamente entre las causas raíces de falla.

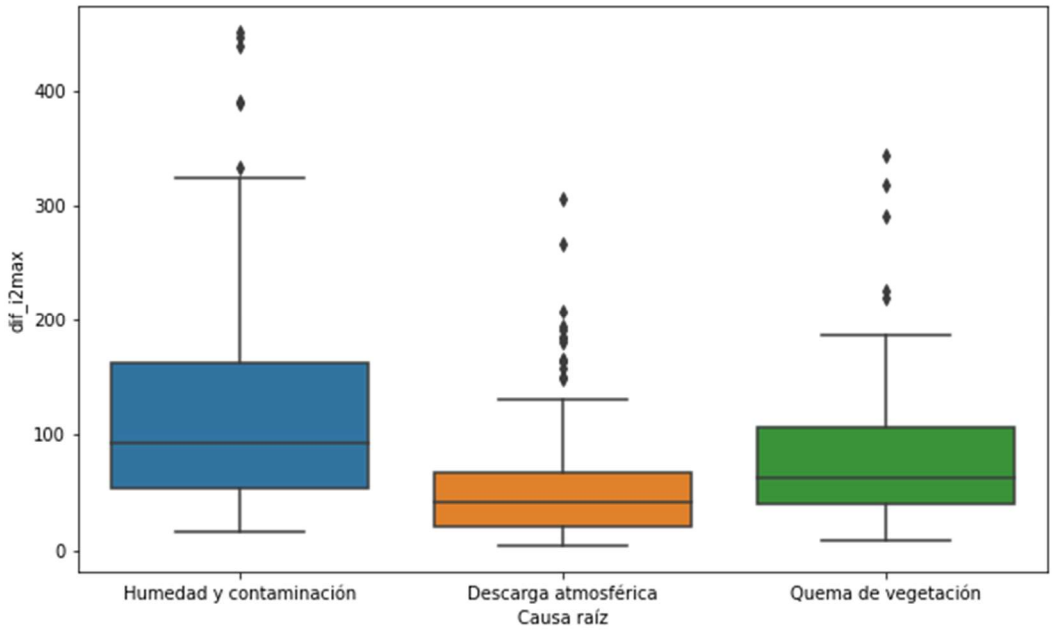


Figura N° 4.13: Diagrama de caja de la característica ‘dif_i2max’ (Fuente: propia del autor).

En la **Figura N° 4.13,** se verificó que la característica ‘dif_i2max’ permite diferenciar moderadamente entre las causas raíces de falla.

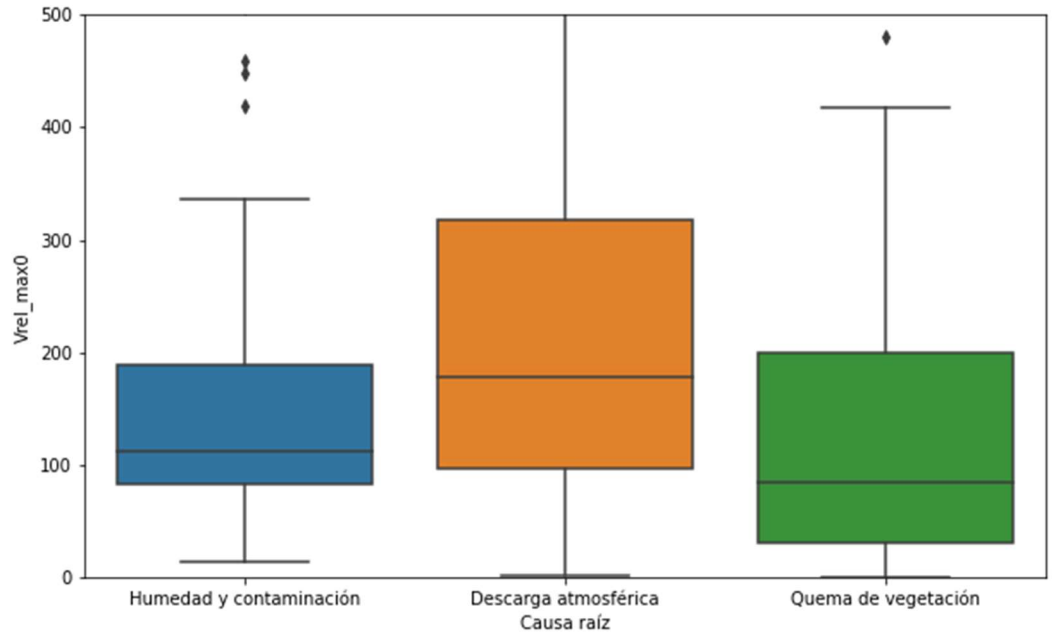


Figura N° 4.14: Diagrama de caja de la característica ‘Vrel_max0’ (Fuente: propia del autor).

En la **Figura N° 4.14,** se verificó que la característica 'Vrel_max0' permite diferenciar moderadamente entre las causas raíces de falla.

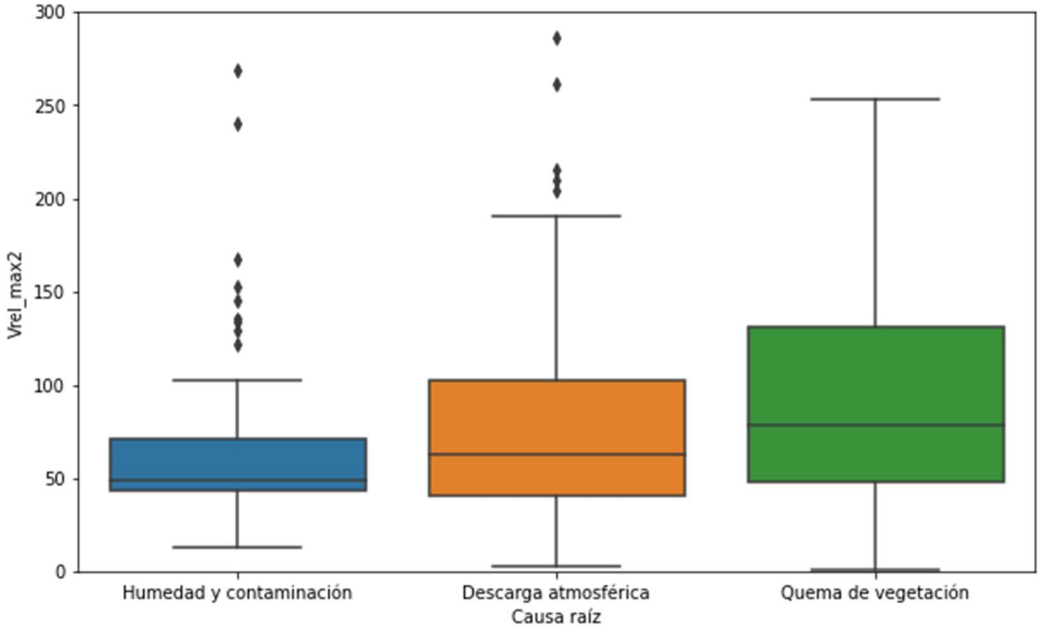


Figura N° 4.15: Diagrama de caja de la característica 'Vrel_max2' (Fuente: propia del autor).

En la **Figura N° 4.15,** se verificó que la característica 'Vrel_max2' permite diferenciar moderadamente entre las causas raíces de falla.

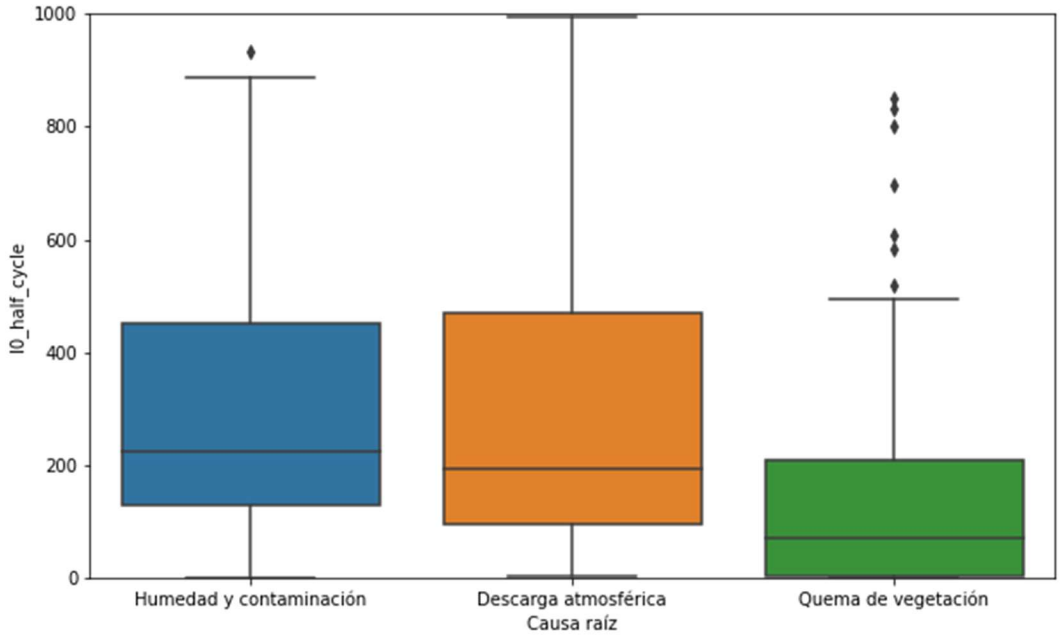


Figura N° 4.16: Diagrama de caja de la característica 'I0_half_cycle' (Fuente: propia del autor).

En la **Figura N° 4.16,** se verificó que la característica ‘I0_half_cycle’ permite diferenciar la causa raíz de quema de vegetación, respecto a las otras.

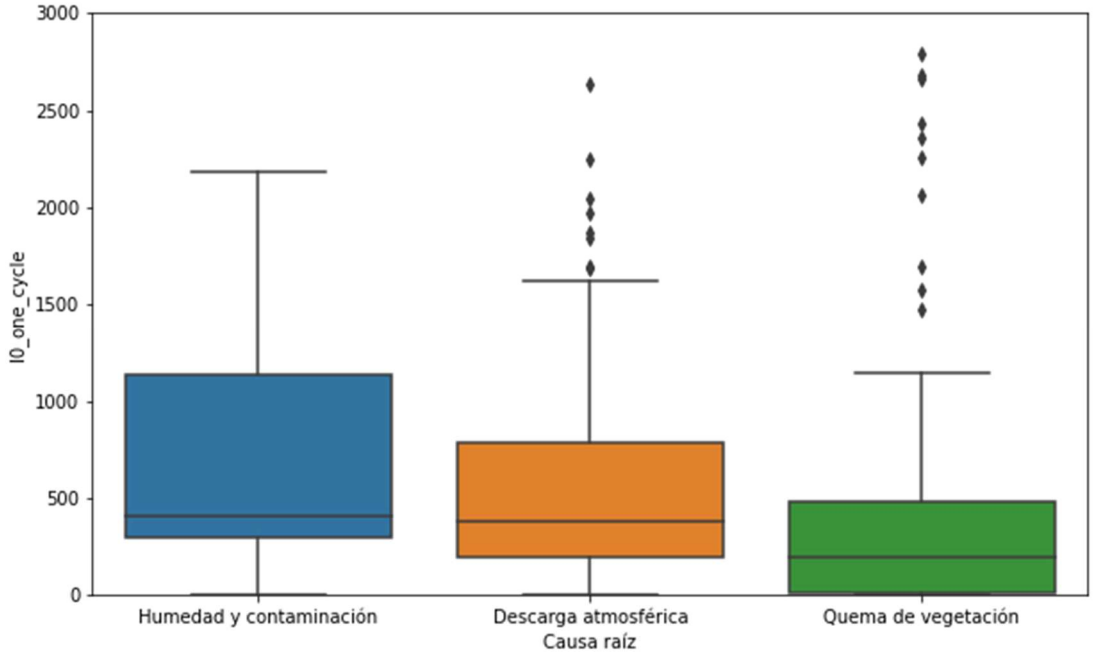


Figura N° 4.17: Diagrama de caja de la característica ‘I0_one_cycle’ (Fuente: propia del autor).

En la **Figura N° 4.17,** se verificó que la característica ‘I0_one_cycle’ permite diferenciar moderadamente entre las causas raíces de falla.

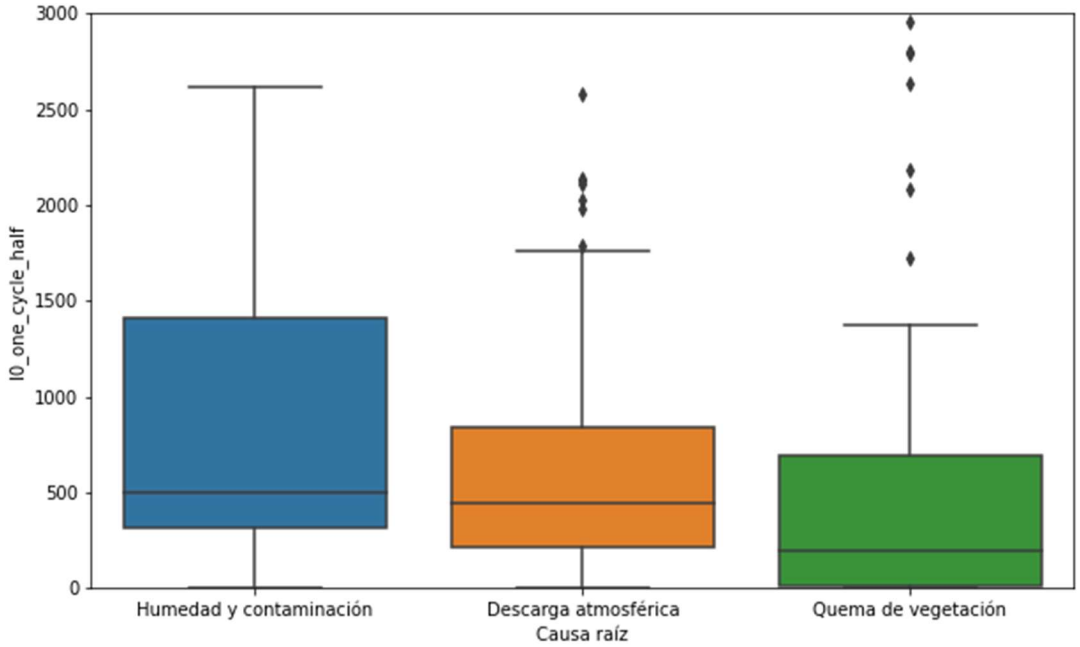


Figura N° 4.18: Diagrama de caja de la característica ‘I0_one_cycle_half’ (Fuente: propia del autor).

En la **Figura N° 4.18,** se verificó que la característica ‘I0_one_cycle_half’ permite diferenciar moderadamente entre las causas raíces de falla.

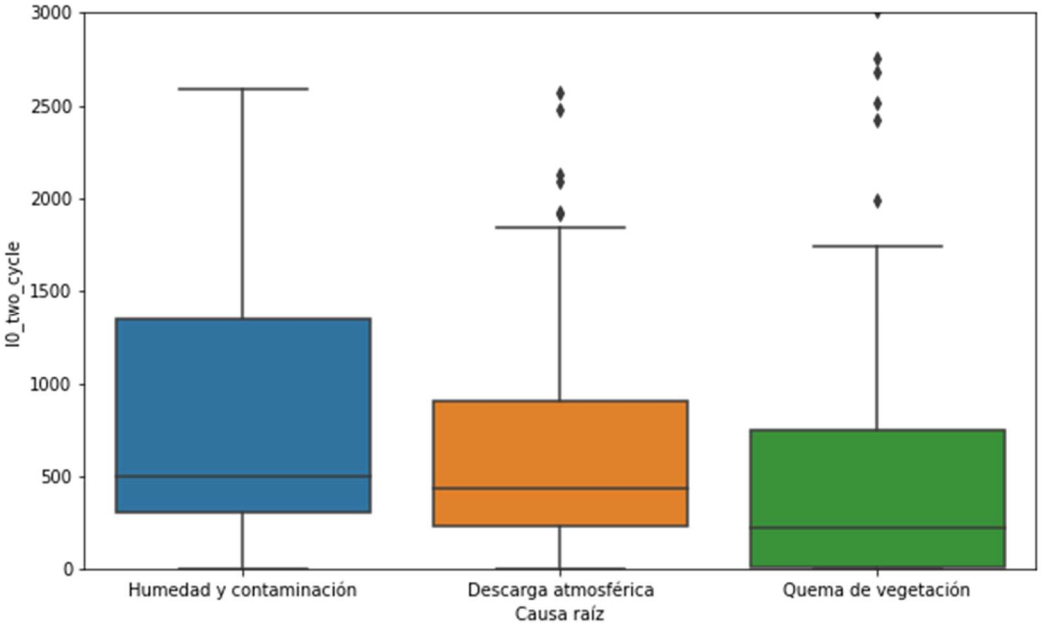


Figura N° 4.19: Diagrama de caja de la característica ‘I0_two_cycle’ (Fuente: propia del autor).

En la **Figura N° 4.19,** se verificó que la característica ‘I0_two_cycle’ permite diferenciar moderadamente entre las causas raíces de falla.

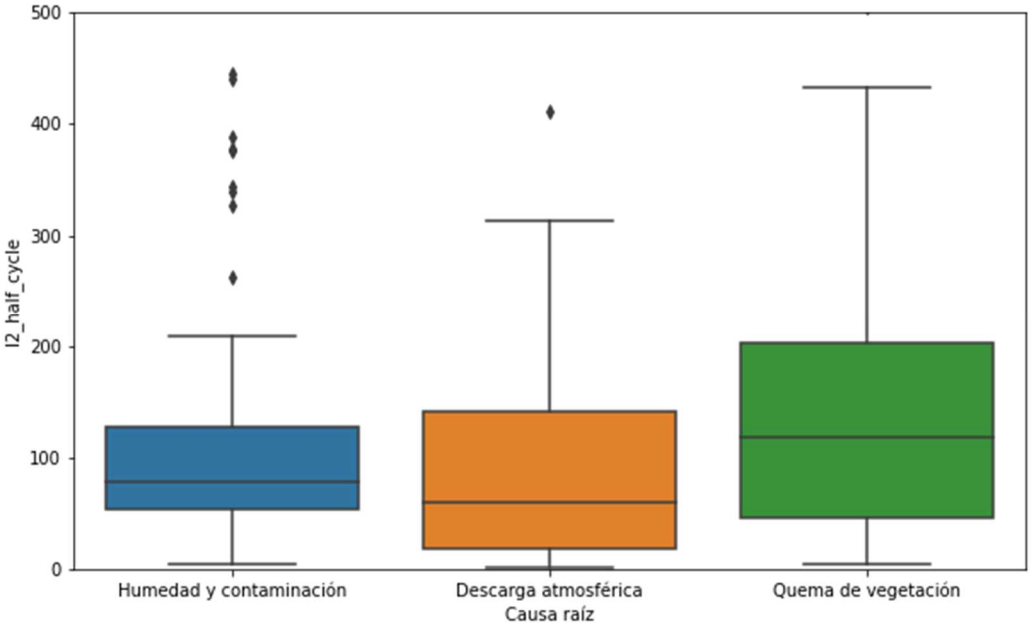


Figura N° 4.20: Diagrama de caja de la característica ‘I2_half_cycle’ (Fuente: propia del autor).

En la **Figura N° 4.20,** se verificó que la característica ‘I2_half_cycle’ permite diferenciar moderadamente entre las causas raíces de falla.

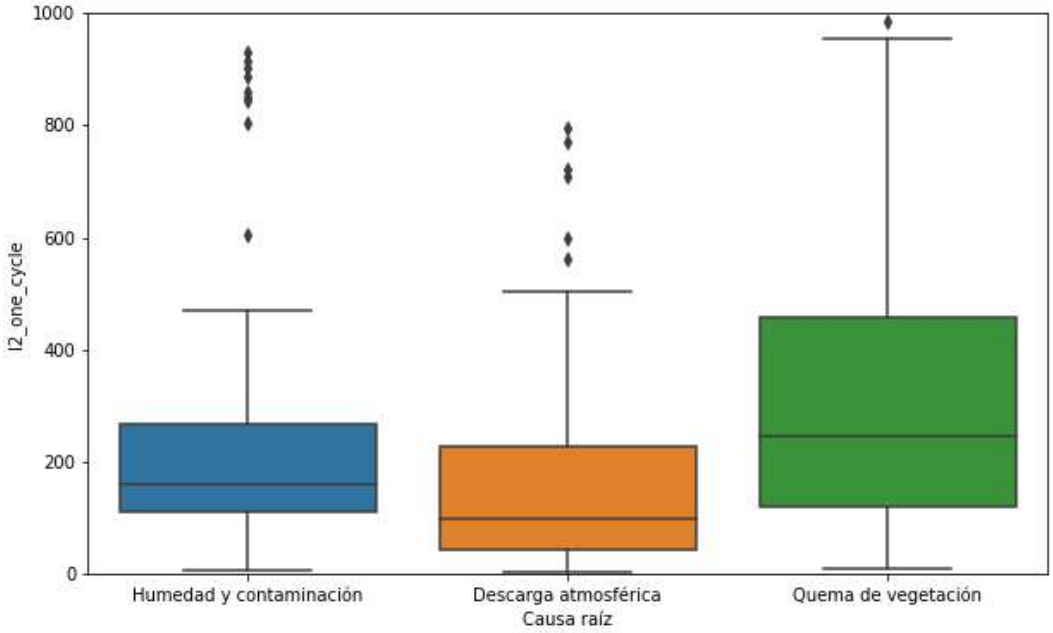


Figura N° 4.21: Diagrama de caja de la característica ‘I2_one_cycle’ (Fuente: propia del autor).

En la **Figura N° 4.21,** se verificó que la característica ‘I2_one_cycle’ permite diferenciar moderadamente entre las causas raíces de falla.

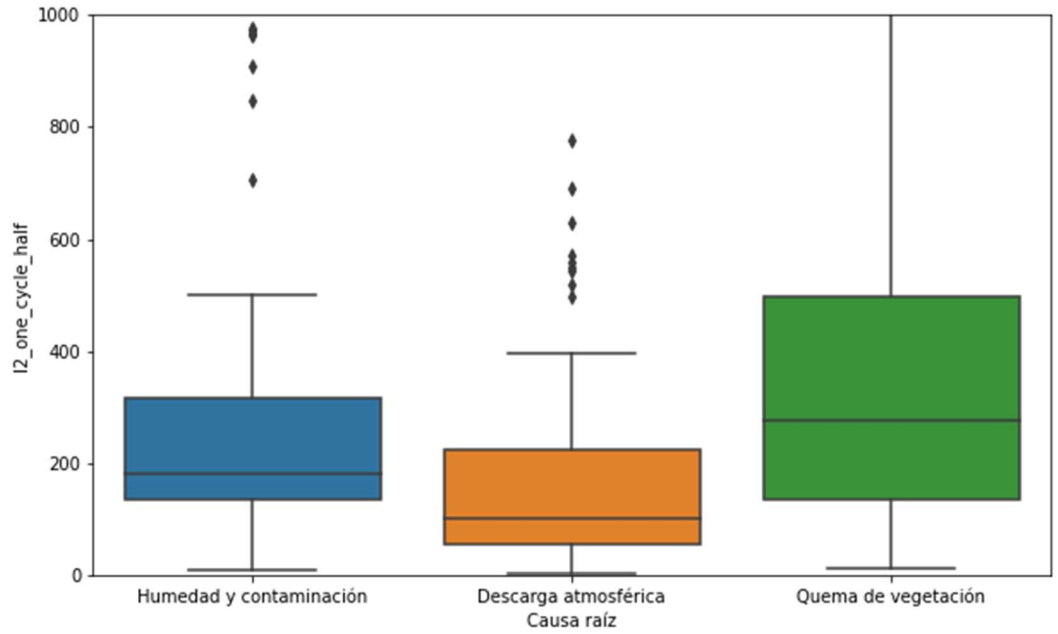


Figura N° 4.22: Diagrama de caja de la característica ‘I2_one_cycle_half’ (Fuente: propia del autor).

En la **Figura N° 4.22,** se verificó que la característica ‘I2_one_cycle_half’ permite diferenciar moderadamente entre las causas raíces de falla.

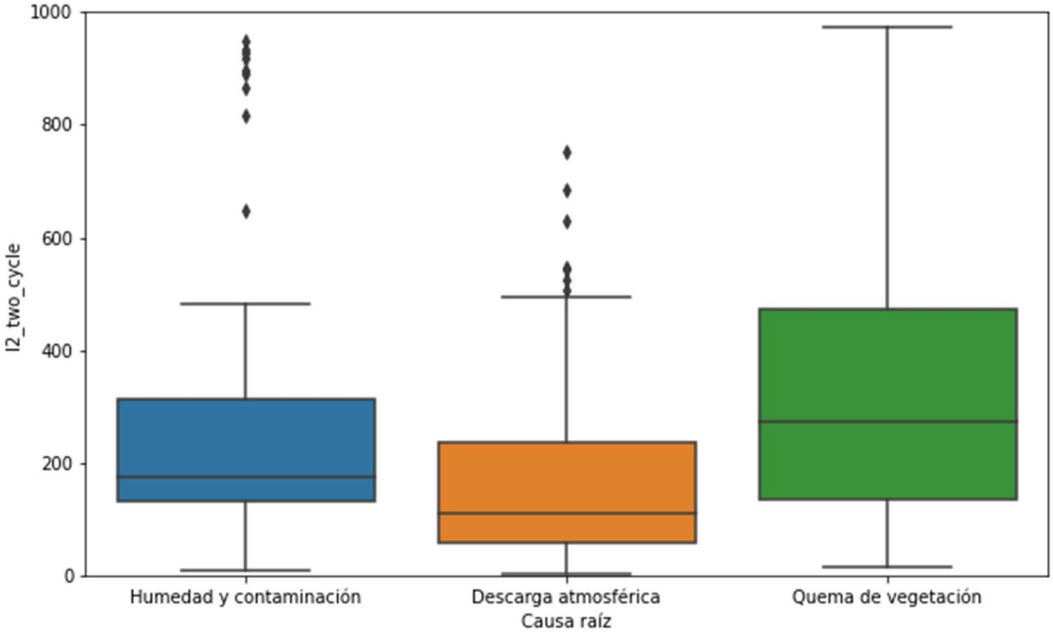


Figura N° 4.23: Diagrama de caja de la característica ‘I2_two_cycle’ (Fuente: propia del autor).

En la **Figura N° 4.23,** se verificó que la característica ‘I2_two_cycle’ permite diferenciar moderadamente entre las causas raíces de falla.

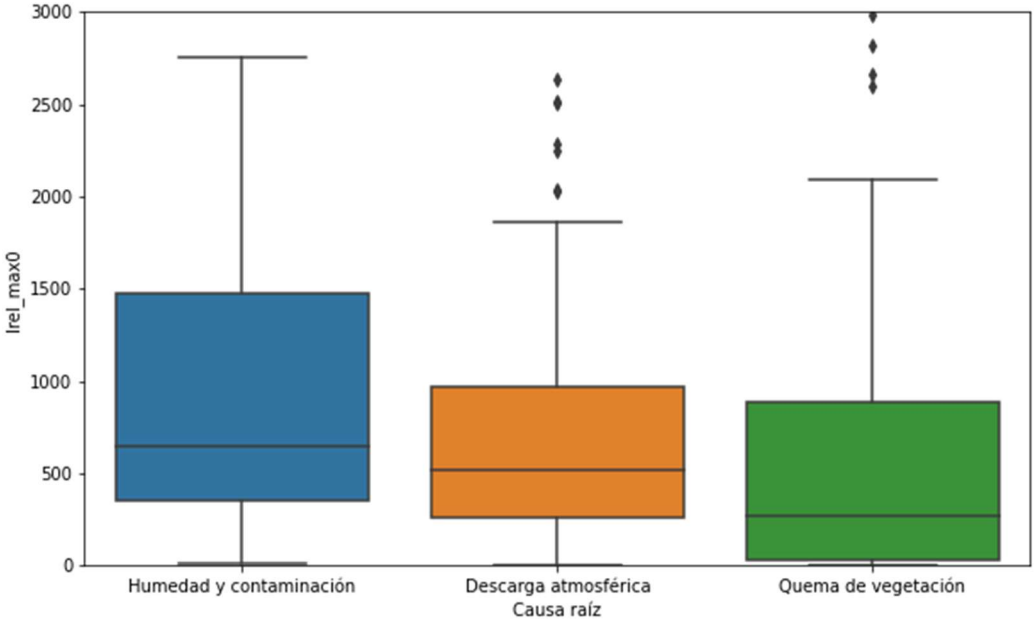


Figura N° 4.24: Diagrama de caja de la característica ‘Irel_max0’ (Fuente: propia del autor).

En la **Figura N° 4.24,** se verificó que la característica ‘Irel_max0’ permite diferenciar moderadamente entre las causas raíces de falla.

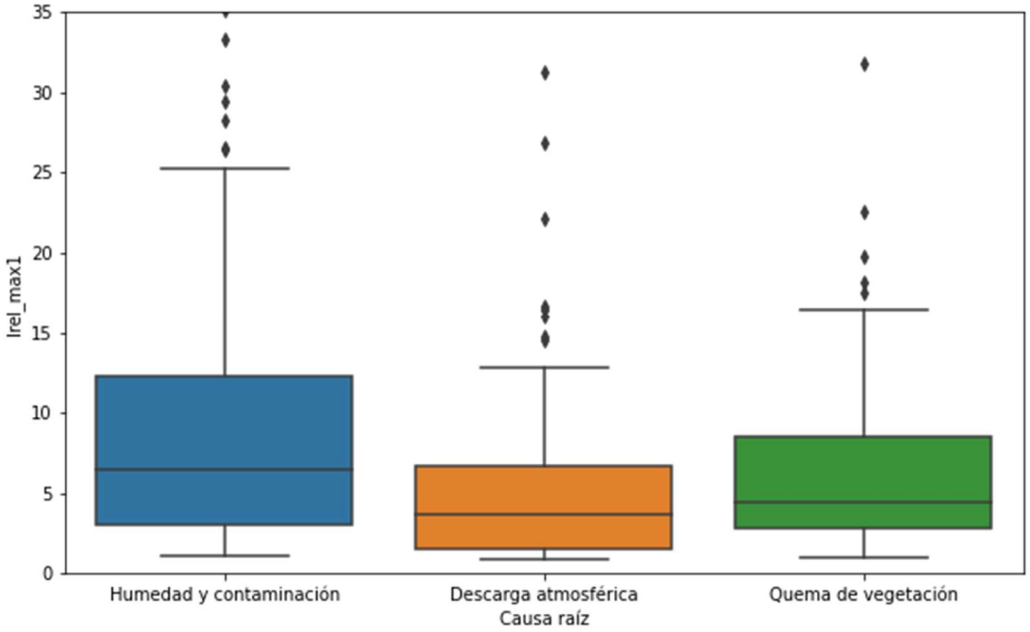


Figura N° 4.25: Diagrama de caja de la característica ‘Irel_max1’ (Fuente: propia del autor).

En la **Figura N° 4.25,** se verificó que la característica ‘Irel_max1’ permite diferenciar ligeramente entre las causas raíces de falla.

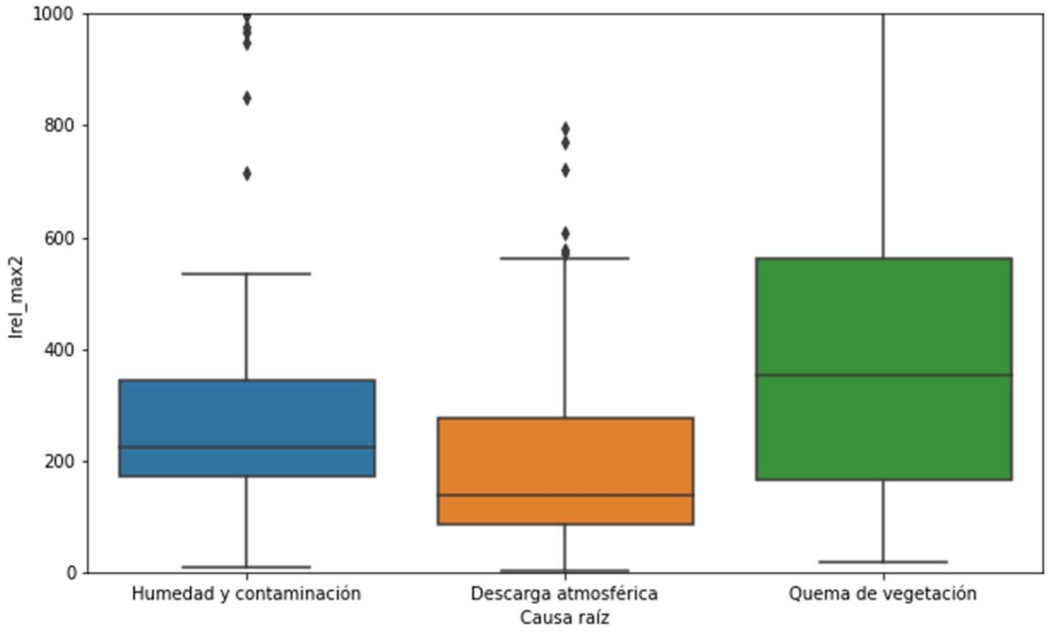


Figura N° 4.26: Diagrama de caja de la característica ‘Irel_max2’ (Fuente: propia del autor).

En la **Figura N° 4.26,** se verificó que la característica ‘Irel_max2’ permite diferenciar moderadamente entre las causas raíces de falla.

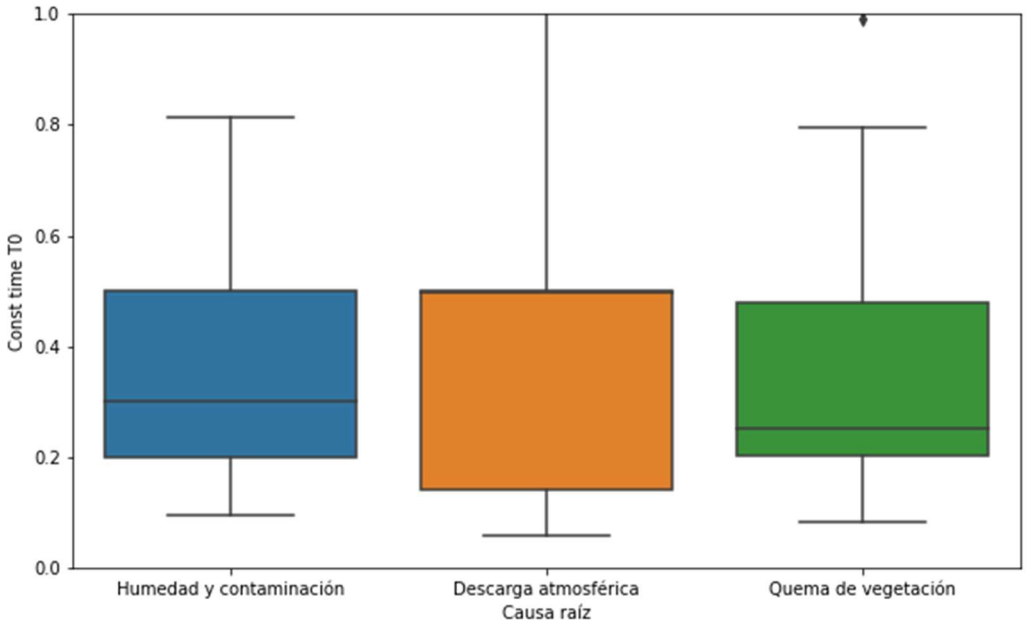


Figura N° 4.27: Diagrama de caja de la característica ‘Const time T0’ (Fuente: propia del autor).

En la **Figura N° 4.27,** se verificó que la característica ‘Const time T0’ muestra diferencias sutiles entre las causas raíces de falla.

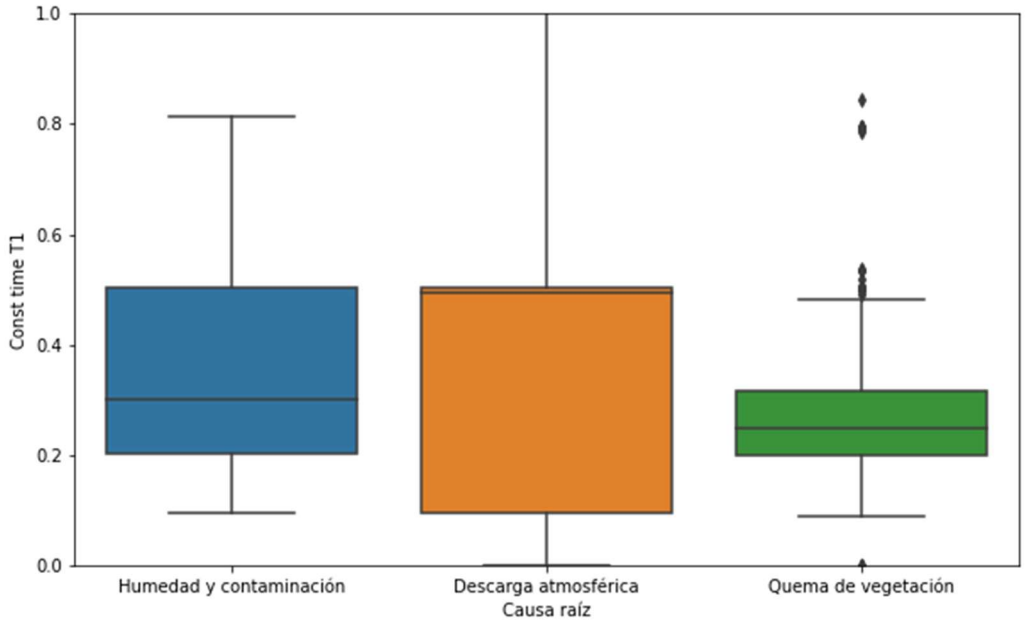


Figura N° 4.28: Diagrama de caja de la característica ‘Const time T1’ (Fuente: propia del autor).

En la **Figura N° 4.28,** se verificó que la característica permite diferenciar moderadamente entre las causas raíces de falla.

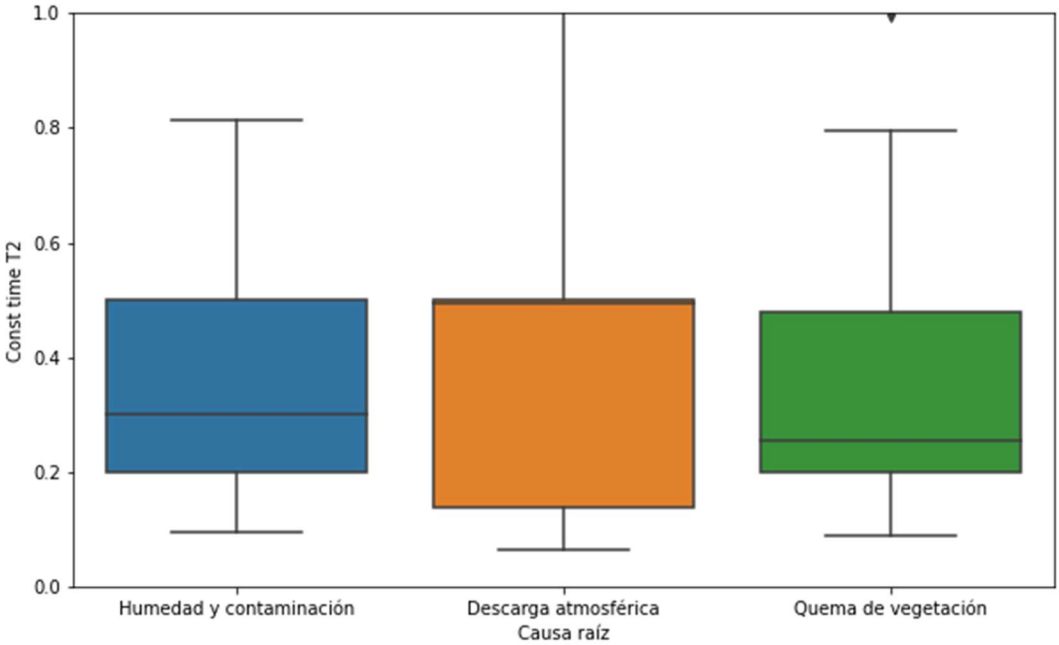


Figura N° 4.29: Diagrama de caja de la característica 'Const time T2' (Fuente: propia del autor).

En la **Figura N° 4.29,** se verificó que la característica 'Const time T2' muestra diferencias sutiles entre las causas raíces de falla.

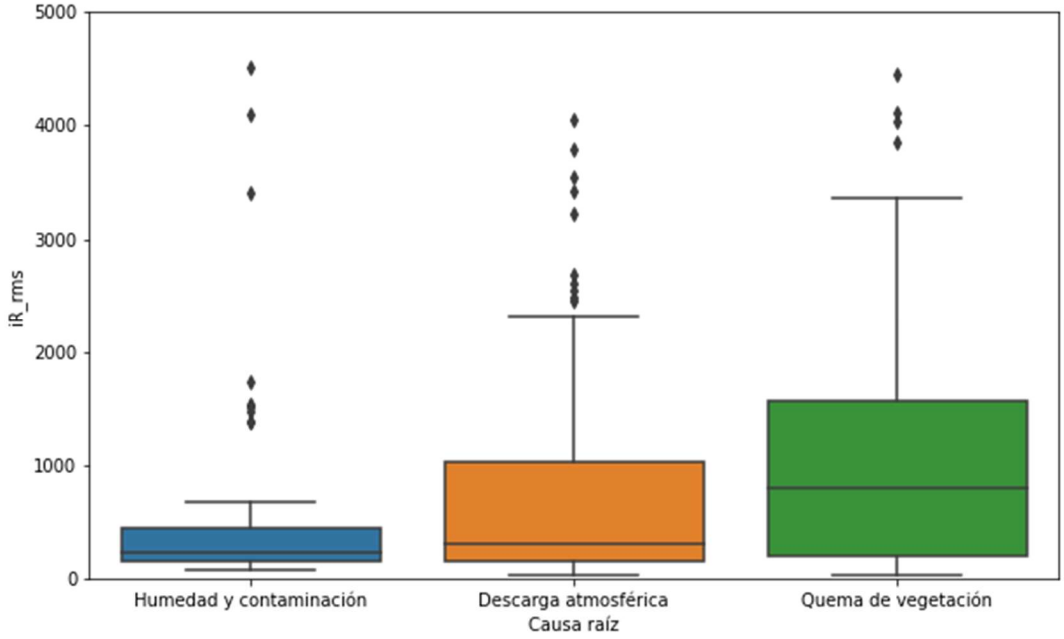


Figura N° 4.30: Diagrama de caja de la característica 'iR_rms' (Fuente: propia del autor).

En la **Figura N° 4.30,** se verificó que la característica ‘iR_rms’ permite diferenciar moderadamente entre las causas raíces de falla.

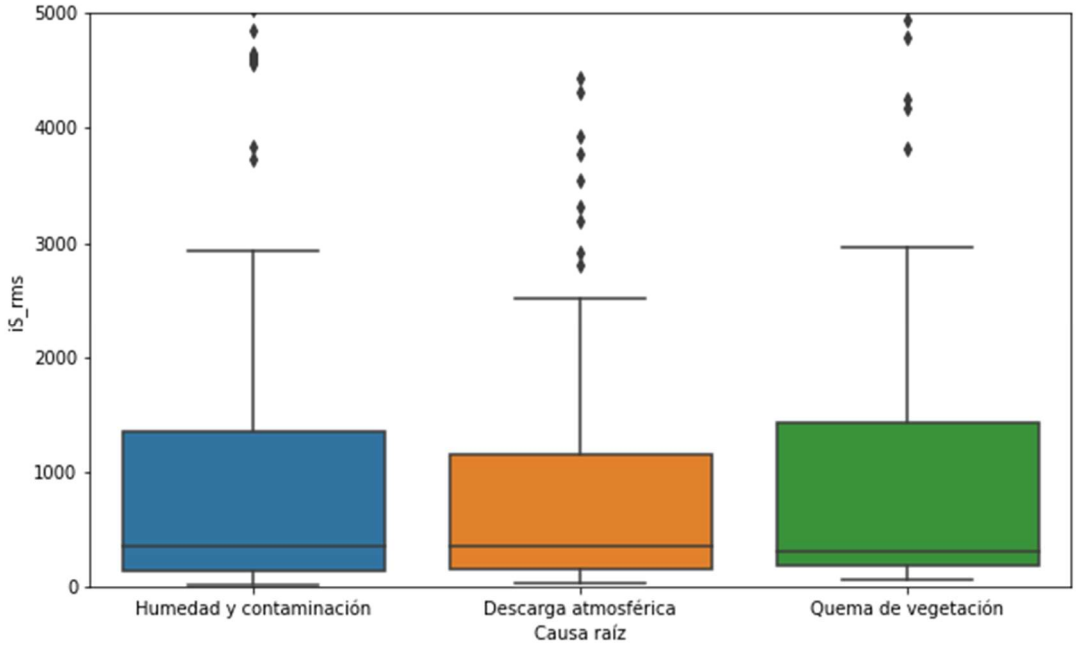


Figura N° 4.31: Diagrama de caja de la característica ‘iS_rms’ (Fuente: propia del autor).

En la **Figura N° 4.31,** se verificó que la característica ‘iS_rms’ muestra diferenciales sutiles entre las causas raíces de falla.

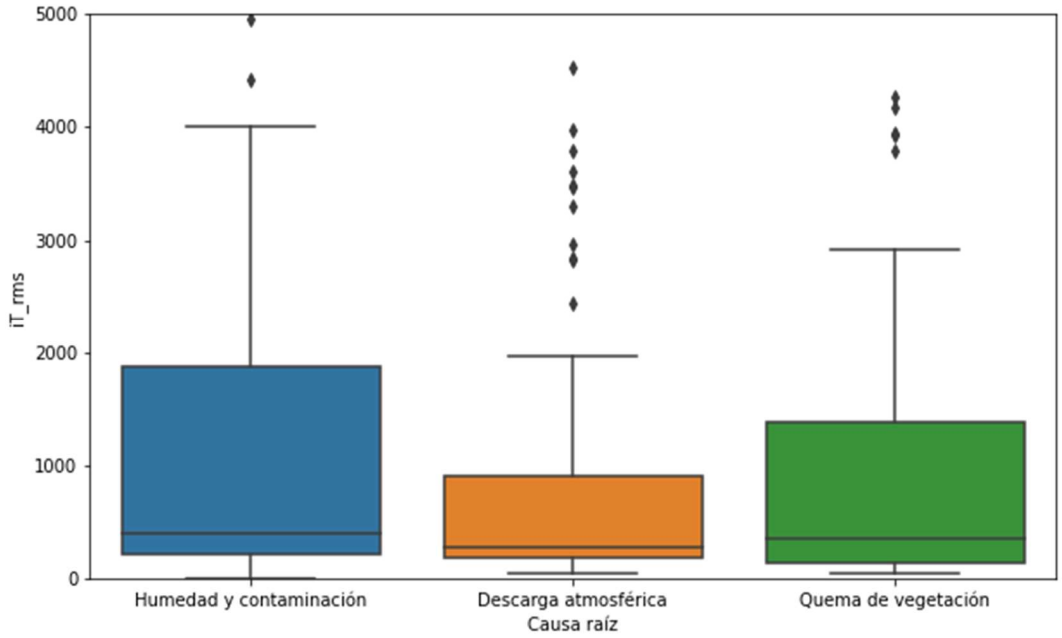


Figura N° 4.32: Diagrama de caja de la característica ‘iT_rms’ (Fuente: propia del autor).

En la **Figura N° 4.32,** se verificó que la característica ‘iT_rms’ permite diferenciar moderadamente entre las causas raíces de falla.

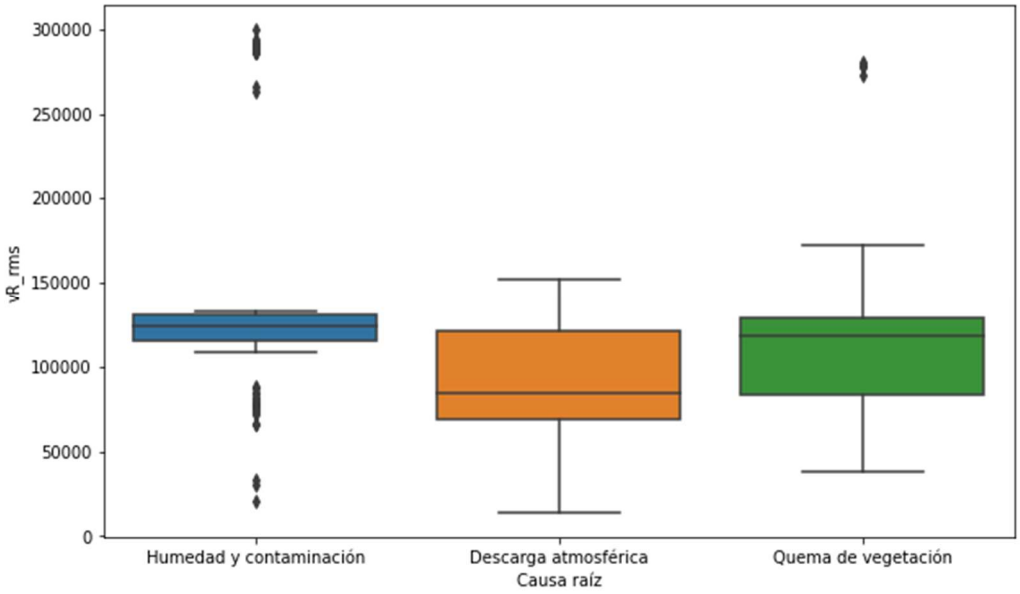


Figura N° 4.33: Diagrama de caja de la característica ‘vR_rms’ (Fuente: propia del autor).

En la **Figura N° 4.33,** se verificó que la característica ‘vR_rms’ permite diferenciar la causa raíz de humedad y contaminación respecto a las otras causas.

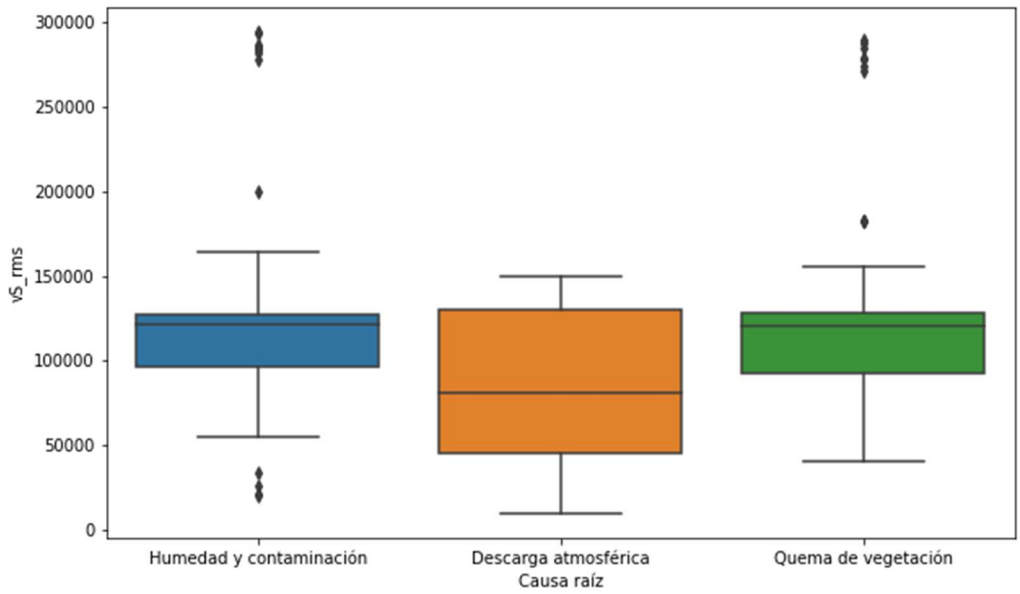


Figura N° 4.34: Diagrama de caja de la característica ‘vS_rms’ (Fuente: propia del autor).

En la **Figura N° 4.34,** se verificó que la característica ‘vS_rms’ permite diferenciar moderadamente entre la causa raíz de descarga atmosférica de las otras causas raíces.

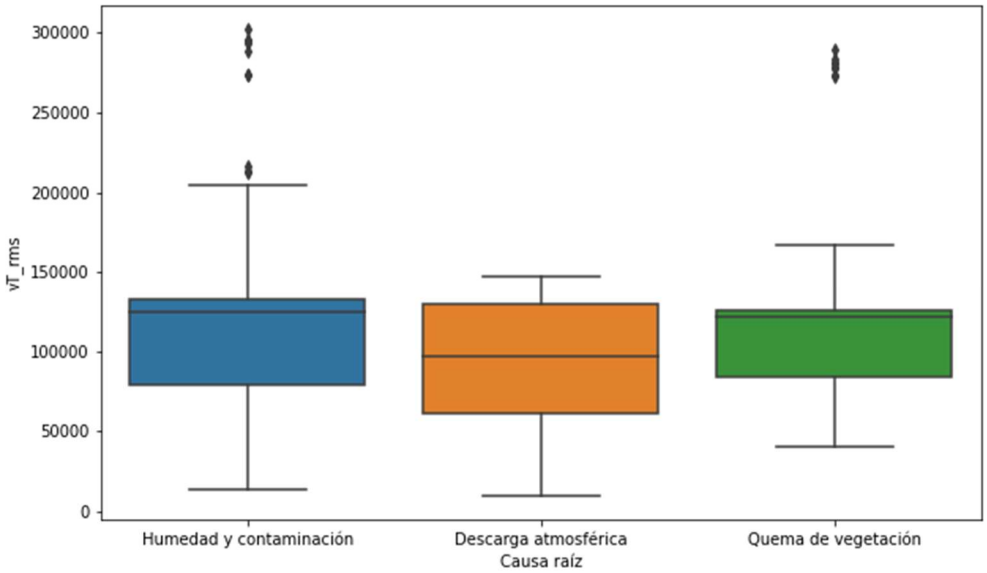


Figura N° 4.35: Diagrama de caja de la característica ‘vT_rms’ (Fuente: propia del autor).

En la **Figura N° 4.35,** se verificó que la característica ‘vT_rms’ muestra diferencias sutiles entre las causas raíces.

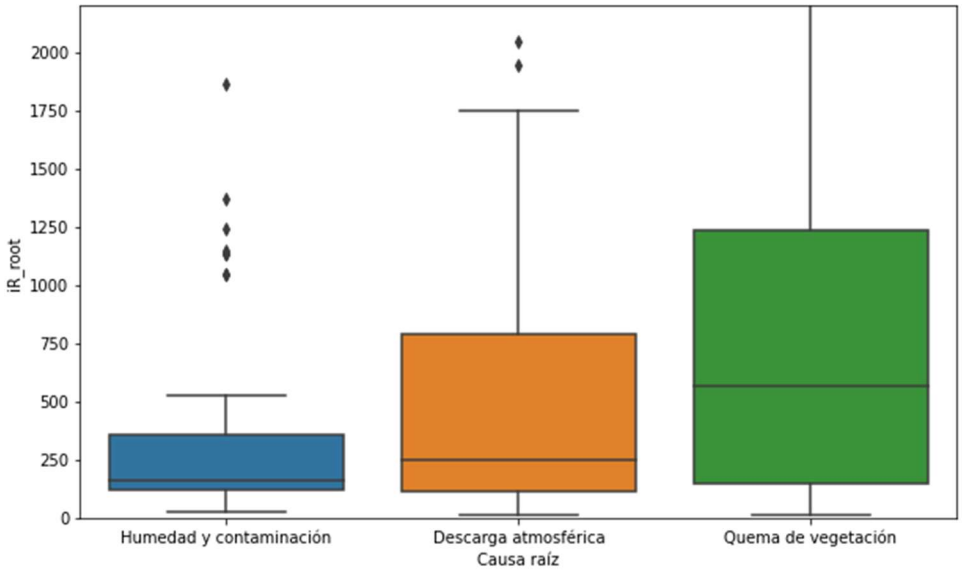


Figura N° 4.36: Diagrama de caja de la característica ‘iR_root’ (Fuente: propia del autor).

En la **Figura N° 4.36,** se verificó que la característica ‘iR_root’ permite diferenciar moderadamente entre las causas raíces de falla.

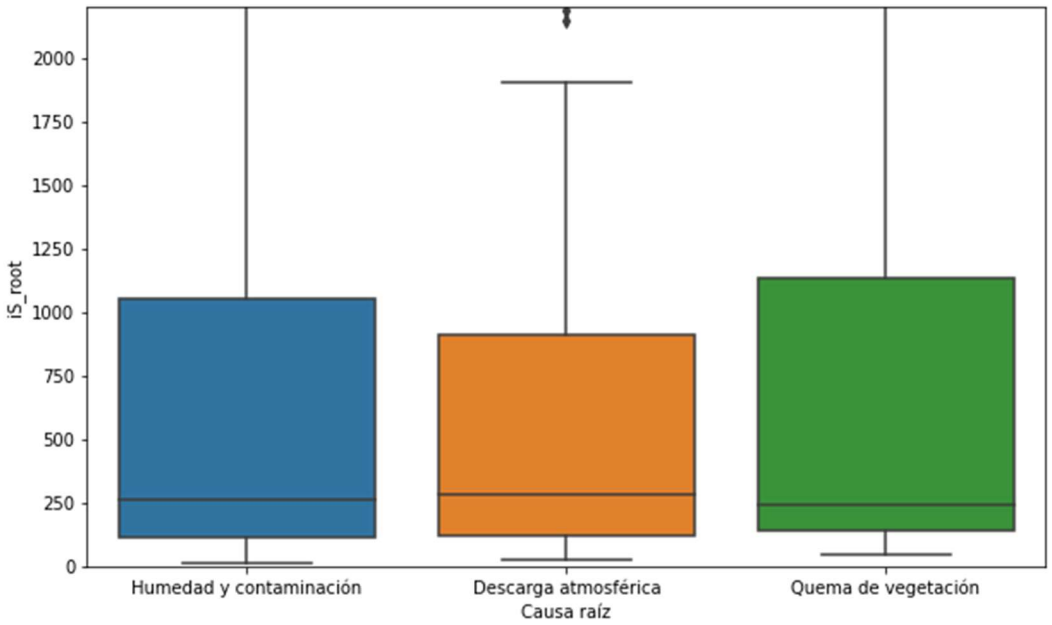


Figura N° 4.37: Diagrama de caja de la característica ‘iS_root’ (Fuente: propia del autor).

En la **Figura N° 4.37,** se verificó que la característica ‘iS_root’ muestra diferencias sutiles entre las causas raíces de falla.

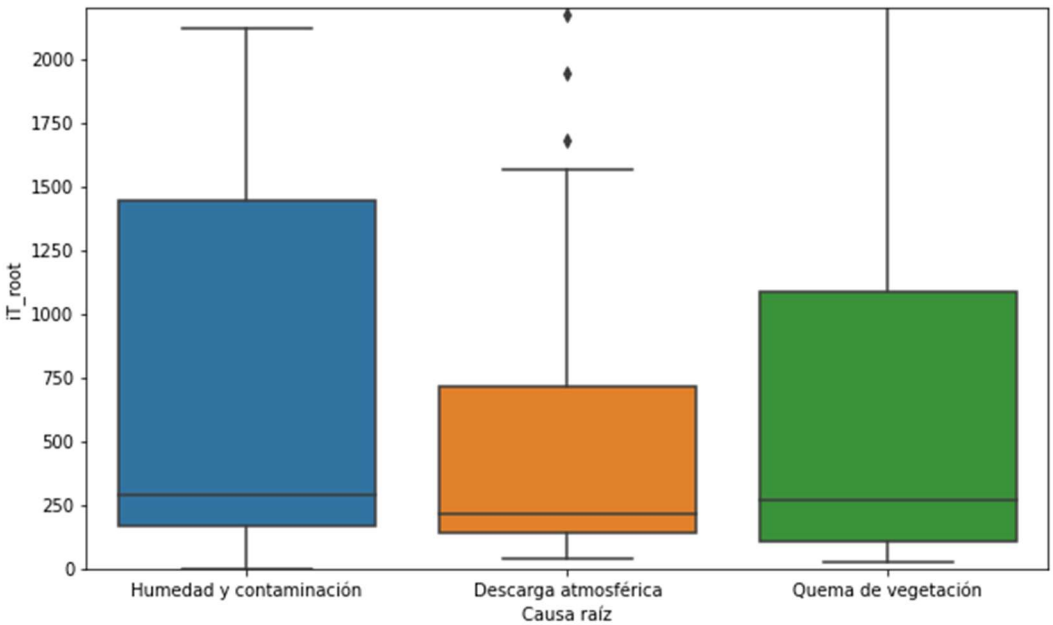


Figura N° 4.38: Diagrama de caja de la característica ‘iT_root’ (Fuente: propia del autor).

En la **Figura N° 4.38,** se verificó que la característica ‘iT_root’ permite diferenciar moderadamente entre las causas raíces de falla.

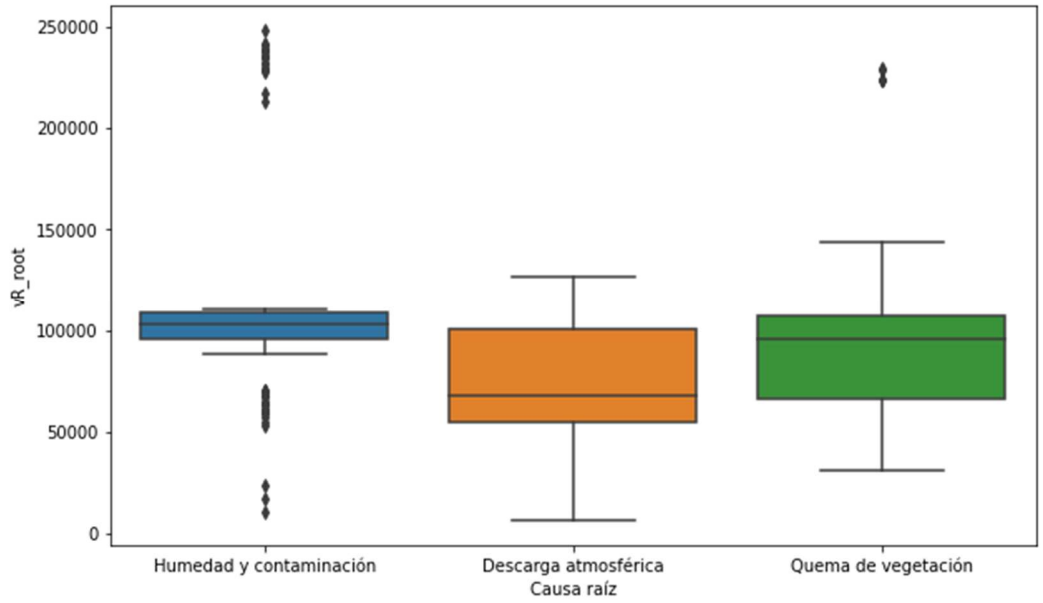


Figura N° 4.39: Diagrama de caja de la característica ‘vR_root’ (Fuente: propia del autor).

En la **Figura N° 4.39,** se verifica que la característica ‘vR_root’ permite diferenciar claramente la causa raíz de humedad y contaminación de las otras causas.

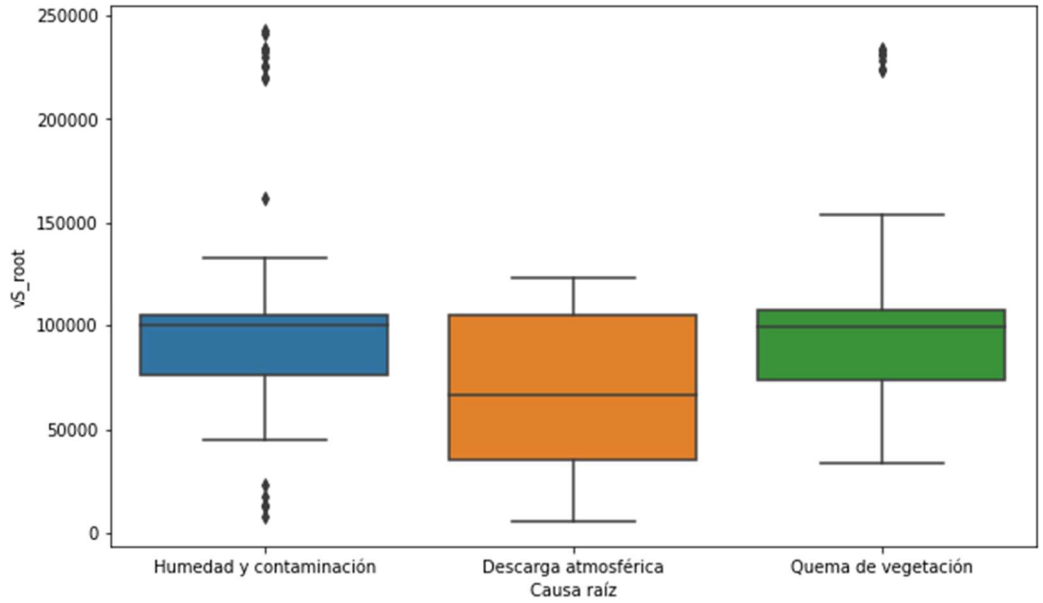


Figura N° 4.40: Diagrama de caja de la característica ‘vS_root’ (Fuente: propia del autor).

En la **Figura N° 4.40,** se verificó que la característica ‘vS_root’ permite diferenciar moderadamente la causa raíz de descarga atmosférica de las otras causas.

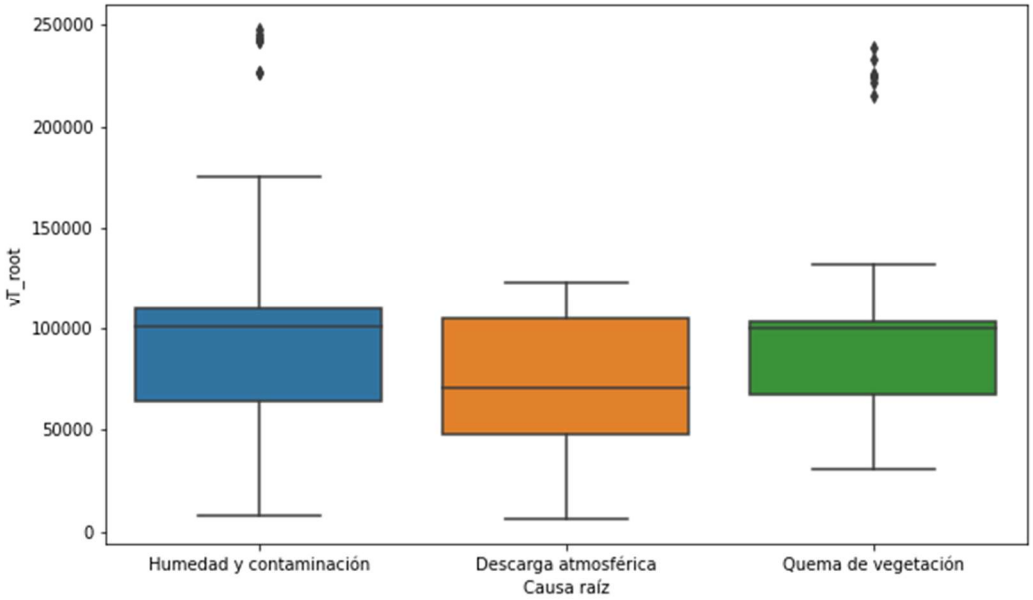


Figura N° 4.41: Diagrama de caja de la característica ‘vT_root’ (Fuente: propia del autor).

En la **Figura N° 4.41,** se verificó que la característica ‘vT_root’ muestra diferencias sutiles entre las causas raíces de falla.

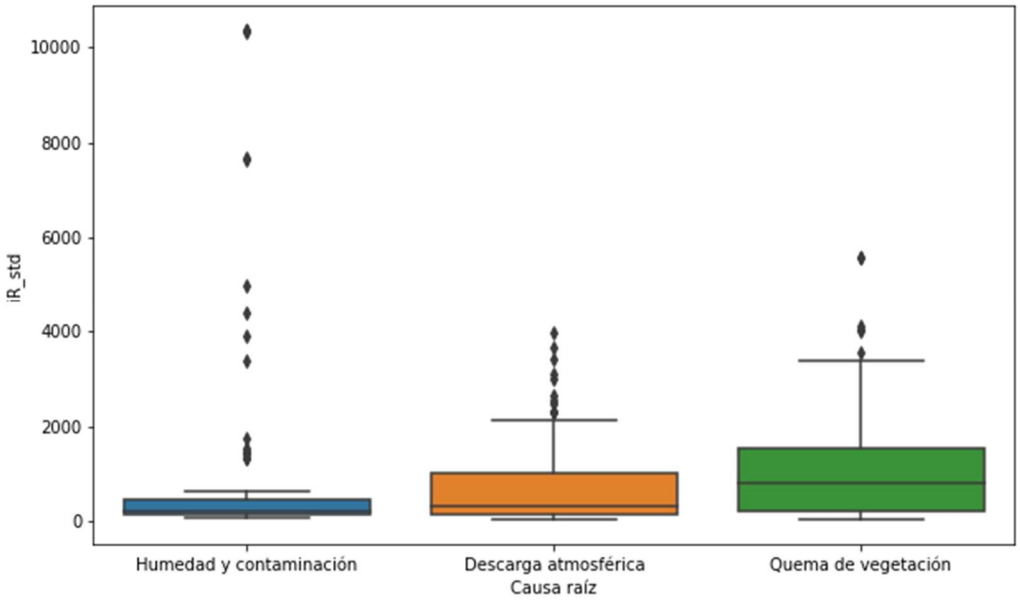


Figura N° 4.42: Diagrama de caja de la característica ‘iR_std’ (Fuente: propia del autor).

En la **Figura N° 4.42,** se verificó que la característica ‘iR_std’ permite diferenciar moderadamente entre las causas raíces de falla.

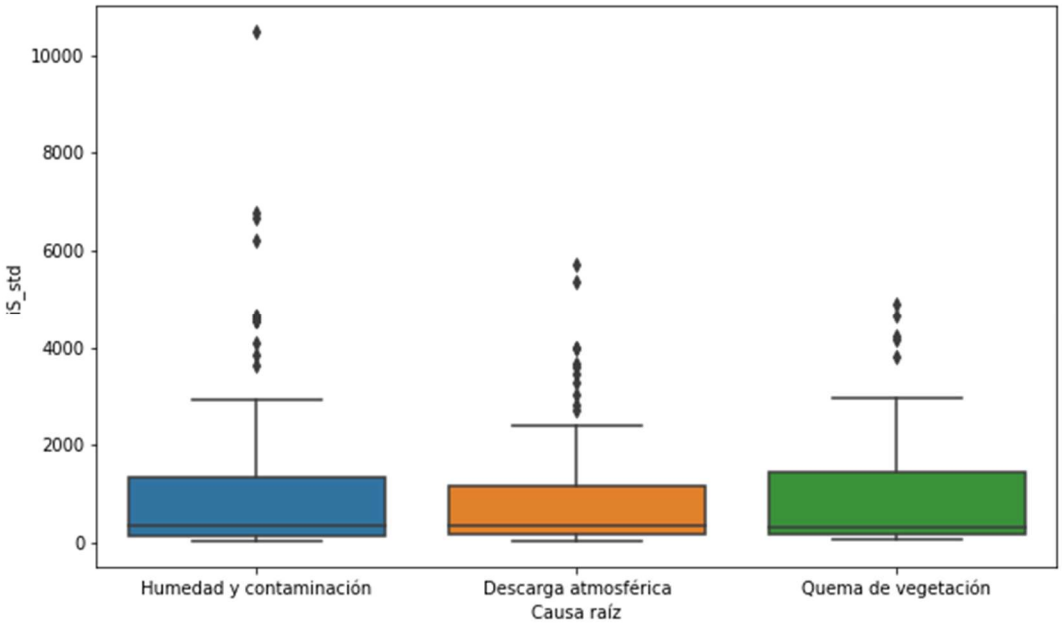


Figura N° 4.43: Diagrama de caja de la característica ‘iS_std’ (Fuente: propia del autor).

En la **Figura N° 4.43,** se verificó que la característica ‘iS_std’ muestra diferencias sutiles entre las causas raíces de falla.

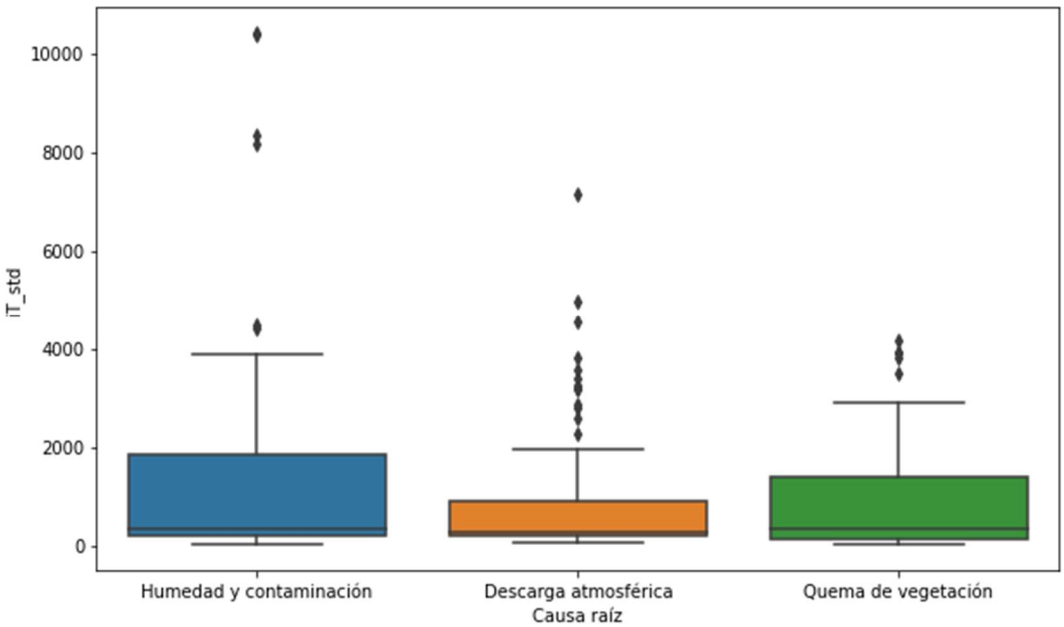
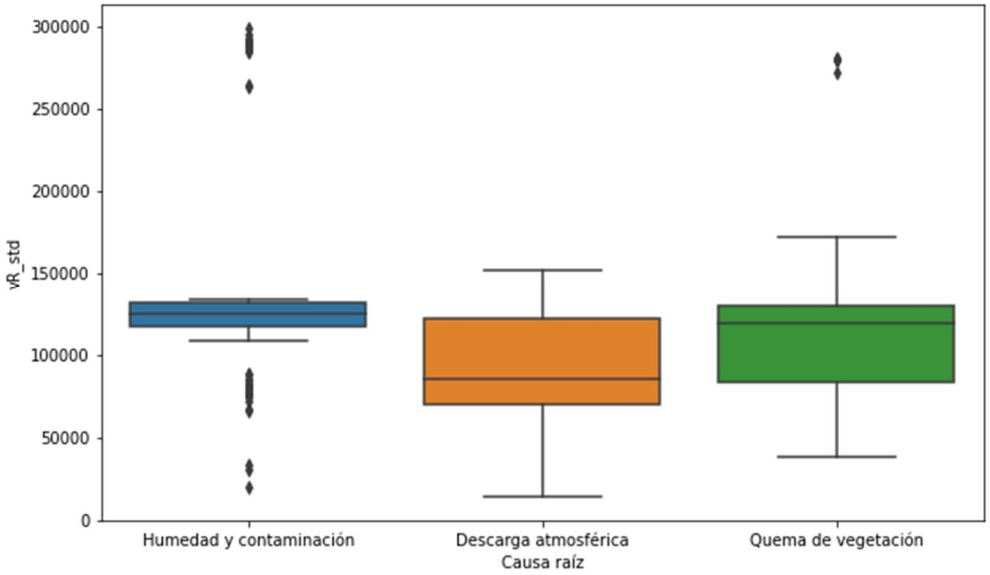


Figura N° 4.44: Diagrama de caja de la característica ‘iT_std’ (Fuente: propia del autor).

En la **Figura N° 4.44,** se verificó que la característica ‘iT_std’ permite diferenciar la causa raíz de descarga atmosférica de otras causas raíces.

Figura N° 4.45: Diagrama de caja de la característica ‘vR_std’ (Fuente: propia del autor).



En la **Figura N° 4.45,** se verificó que la característica ‘vR_std’ permite diferenciar claramente la causa raíz de humedad y contaminación de las otras causas.

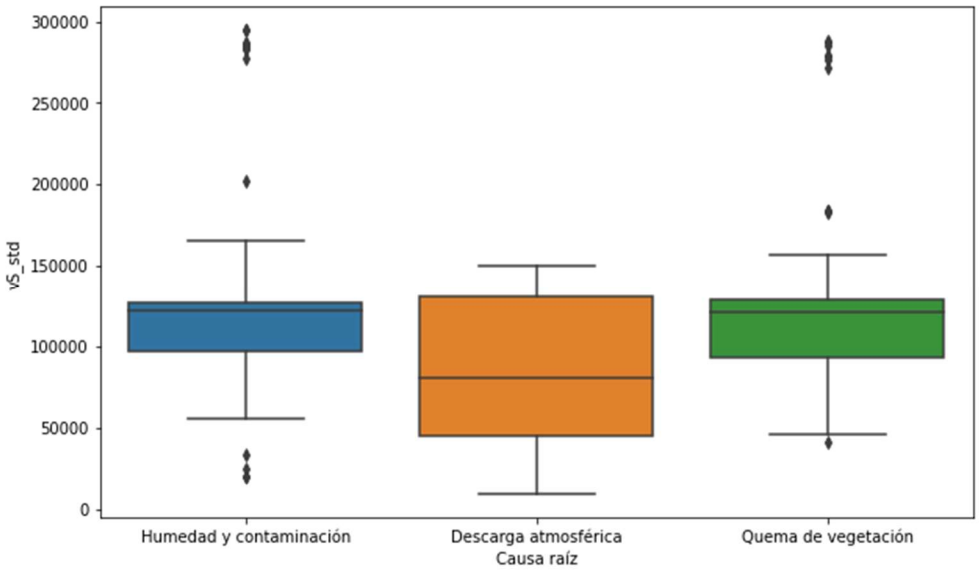


Figura N° 4.46: Diagrama de caja de la característica ‘vS_std’ (Fuente: propia del autor).

En la **Figura N° 4.46,** se verificó que la característica ‘vS_std’ permite diferenciar la causa raíz de descarga atmosférica de otras causas raíces.

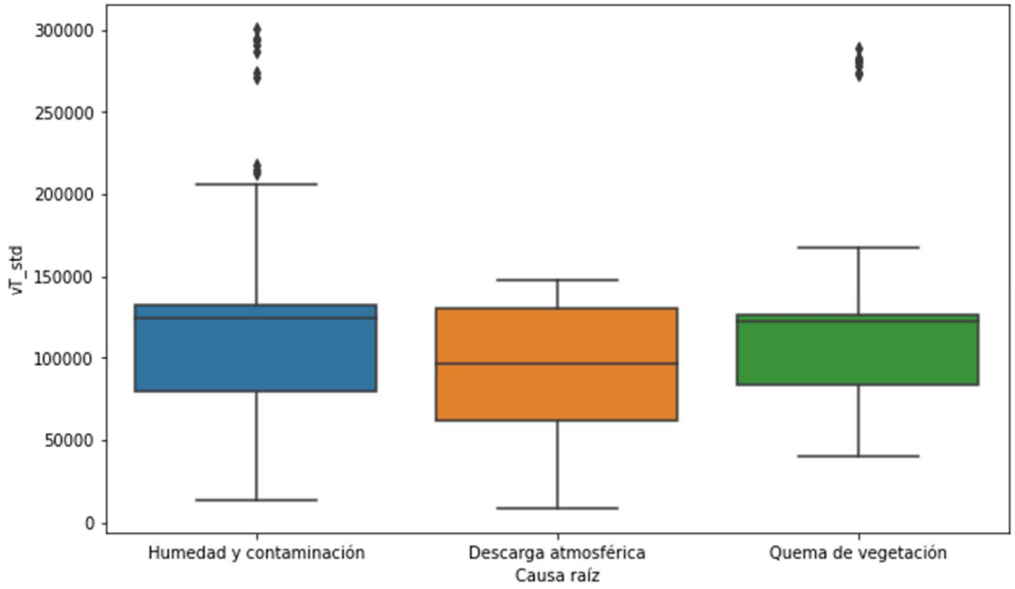


Figura N° 4.47: Diagrama de caja de la característica ‘vT_std’ (Fuente: propia del autor).

En la **Figura N° 4.47,** se verificó que la característica ‘vT_std’ permite diferenciar sutilmente entre las causas raíces.

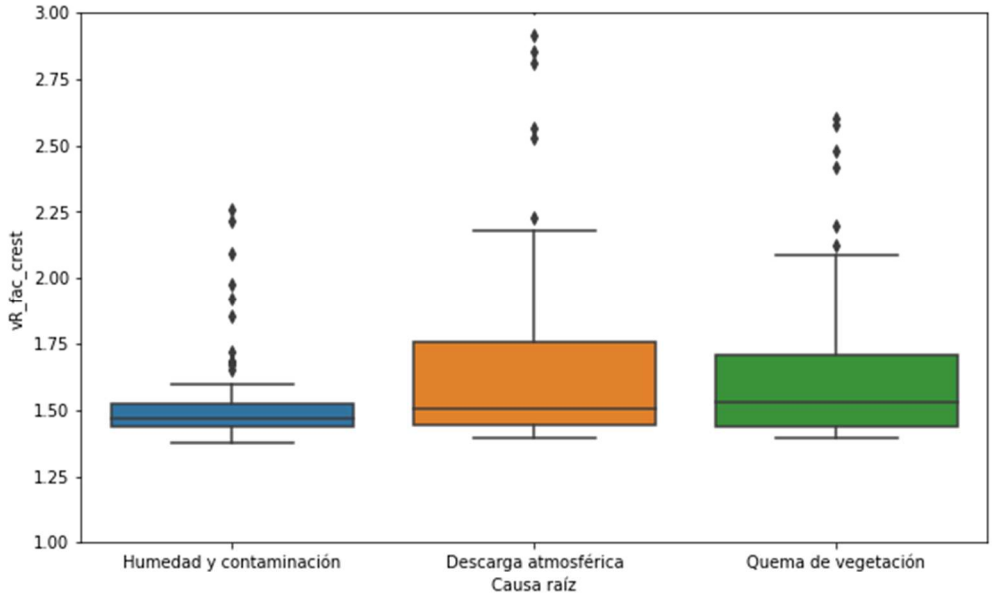


Figura N° 4.48: Diagrama de caja de la característica ‘vR_fac_crest’ (Fuente: propia del autor).

En la **Figura N° 4.48,** se verificó que la característica permite diferenciar la causa raíz de humedad y contaminación de las otras causas raíz.

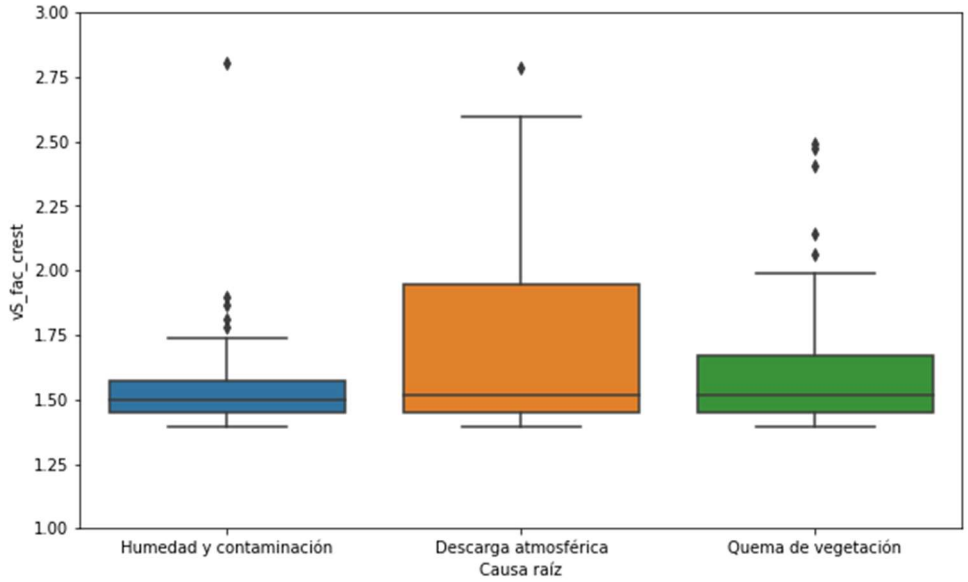


Figura N° 4.49: Diagrama de caja de la característica 'vS_fac_crest' (Fuente: propia del autor).

En la **Figura N° 4.49,** se verificó que la característica permite diferenciar la causa raíz de descarga atmosférica de las otras causas raíces de falla.

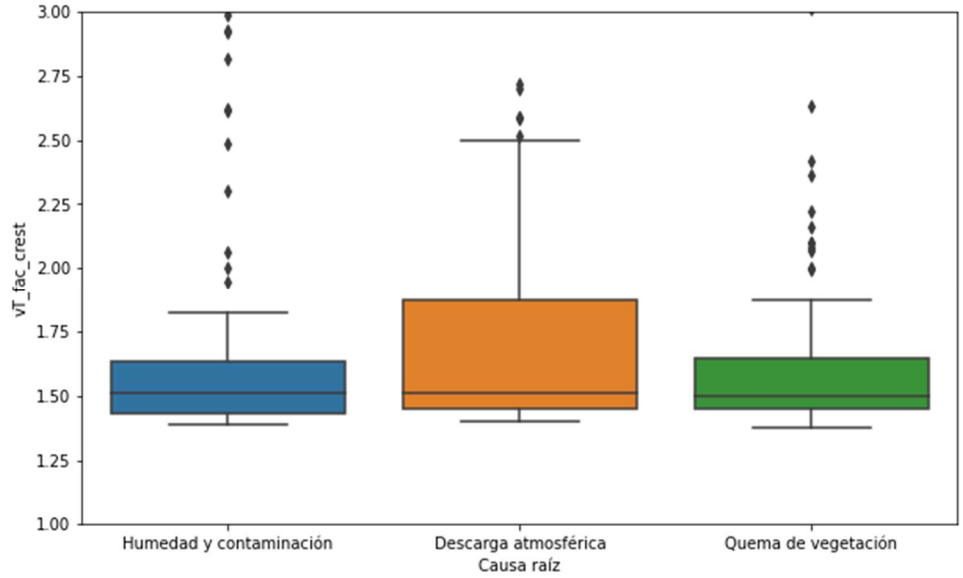


Figura N° 4.50: Diagrama de caja de la característica 'vT_fac_crest' (Fuente: propia del autor).

En la **Figura N° 4.50,** se verificó que la característica permite diferenciar la causa raíz de descarga atmosférica de las otras causas raíces de falla.

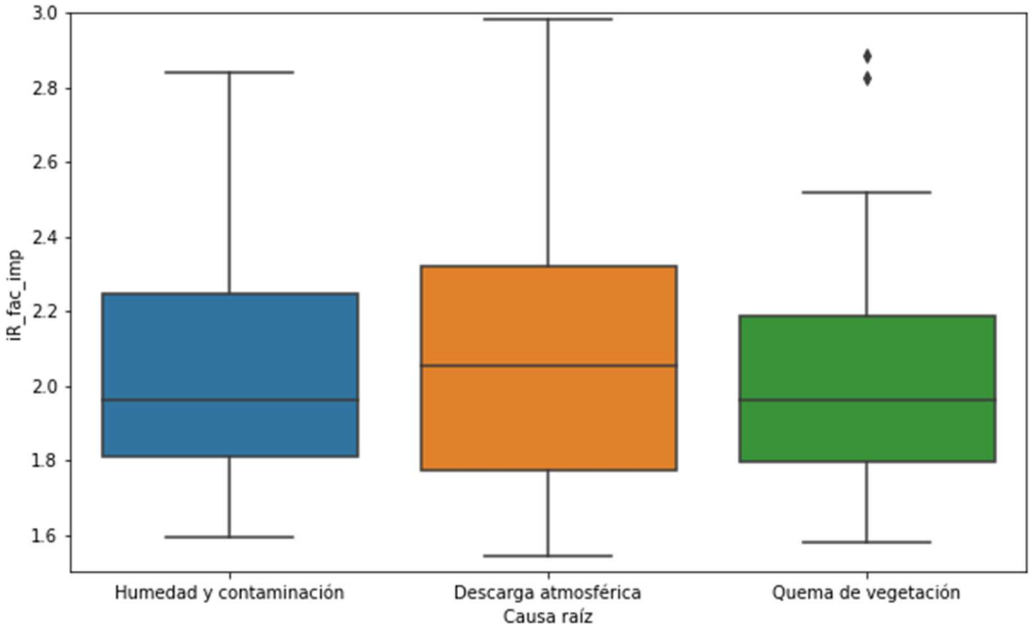


Figura N° 4.51: Diagrama de caja de la característica 'iR_fac_imp' (Fuente: propia del autor).

En la **Figura N° 4.51,** se verificó que la característica 'iR_fac_imp' muestra diferencias sutiles en las causas raíces.

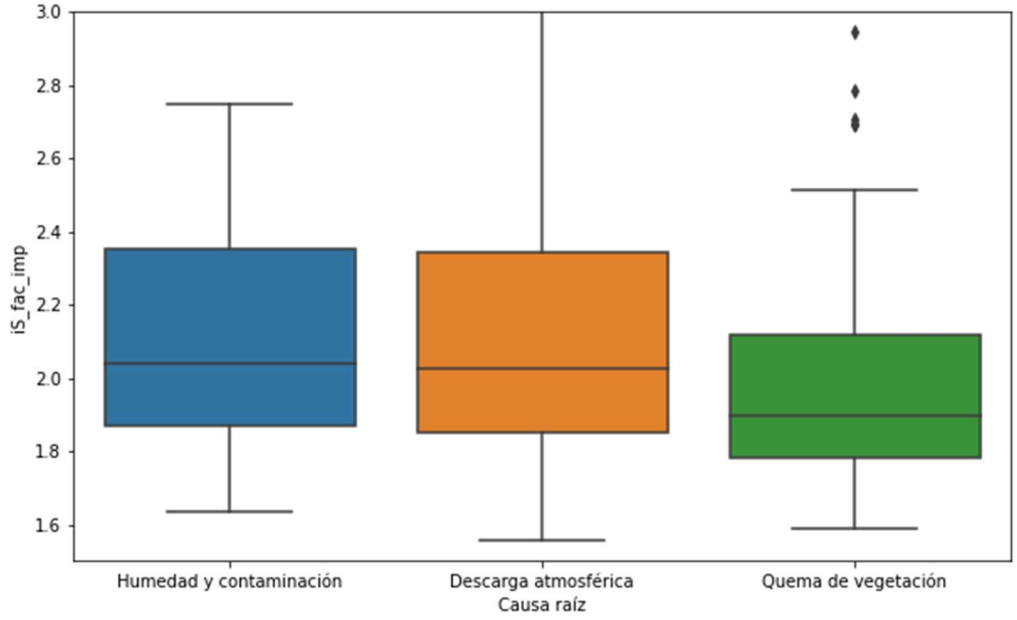


Figura N° 4.52: Diagrama de caja de la característica 'iS_fac_imp' (Fuente: propia del autor).

En la **Figura N° 4.52,** se verificó que la característica ‘iS_fac_imp’ muestra diferencias sutiles entre las causas raíces de falla.

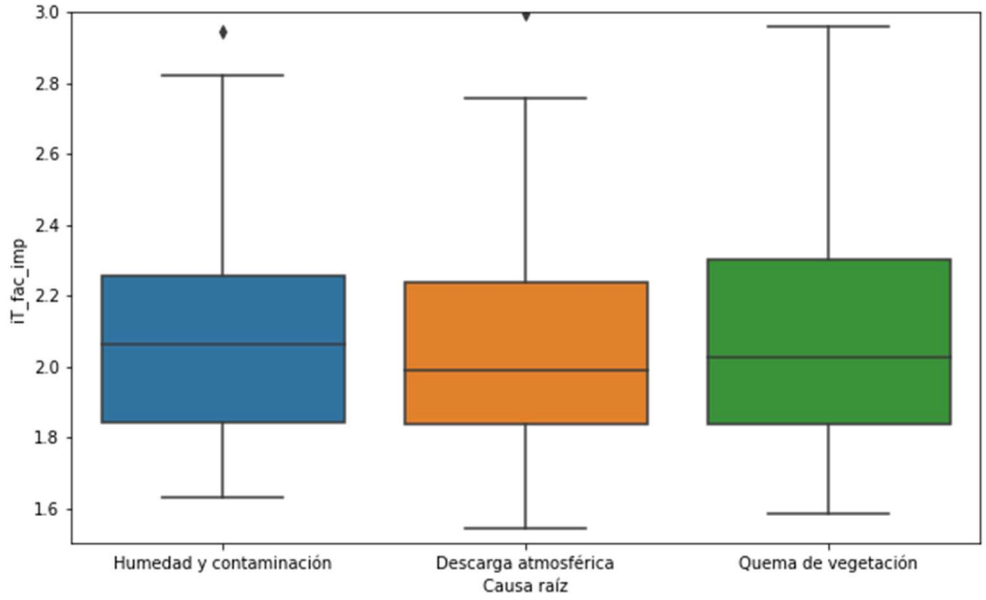


Figura N° 4.53: Diagrama de caja de la característica ‘iT_fac_imp’ (Fuente: propia del autor).

En la **Figura N° 4.53,** se verificó que la característica ‘iT_fac_imp’ muestra diferencias muy sutiles entre las causas raíces de falla.

4.6.4 Selección de algoritmo de aprendizaje supervisado y entrenamiento del modelo de Machine Learning

A. Selección de algoritmo de clasificación de aprendizaje supervisado

Minnaar, Niccols y Gaunt (2015) en el estudio sobre “Automating transmisión line fault root causa analysis” realizaron un estudio comparativo sobre el resultado de las métricas de exactitud y f1-score, de modelos entrenados con algoritmos de clasificación árbol de decisiones, redes neuronales, Naive Bayes y k-nearest neighbors. El estudio concluye que el

algoritmo con el mejor resultado de métricas de evaluación fue el algoritmo de k-nearest neighbors.

En este estudio, se empleó el algoritmo de clasificación de aprendizaje supervisado denominado k–nearest neighbors (k – vecinos más cercanos).

B. Entrenamiento del modelo de Machine Learning

Para el entrenamiento del modelo basado en el algoritmo de k–nearest neighbors, se implementó un programa en PyCharm 2021.2.3 empleando lenguaje de programación Python.

Se utilizó la función 'train_test_split' de la librería de Scikit-learn versión 1.0.2, para la división del conjunto de datos en dos grupos: conjunto de muestras de entrenamiento y conjunto de muestras de test. Se ajustó el hiper-parámetro 'test_size' de la función a 0.3:

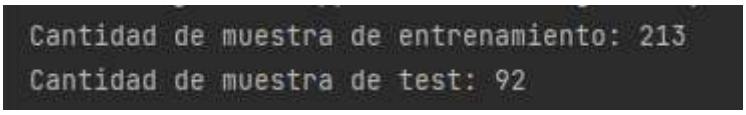


Figura N° 4.54: Cantidad de datos de muestra de entrenamiento y test (Fuente: propia del autor)

C. Ajuste de hiper-parámetros

Se utilizó la función 'GridSearchCV' de la librería Scikit-learn versión 1.0.2. Se encontró que el mejor método de distancia fue 'chebyshev'. Se ajustó el número de vecinos más cercano igual a 7 y una reducción de componentes principales igual a 8.

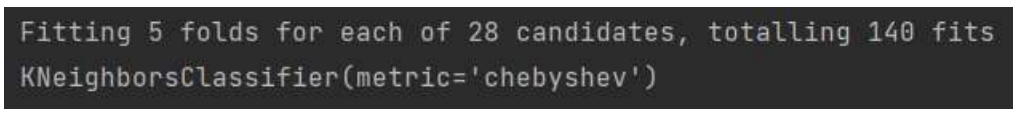


Figura N° 4.55: Determinación del mejor método de cálculo de distancia mediante GridSearchCV (Fuente: propia del autor)

4.6.5 Evaluación del modelo de clasificación de Machine Learning

A. Matriz de consistencia

- **Modelo de clasificación entrenado con el conjunto de muestra de entrenamiento**

  En la **Figura N° 4.56**, se observa la matriz de confusión del modelo de clasificación que entrenó con el conjunto de muestra de entrenamiento:

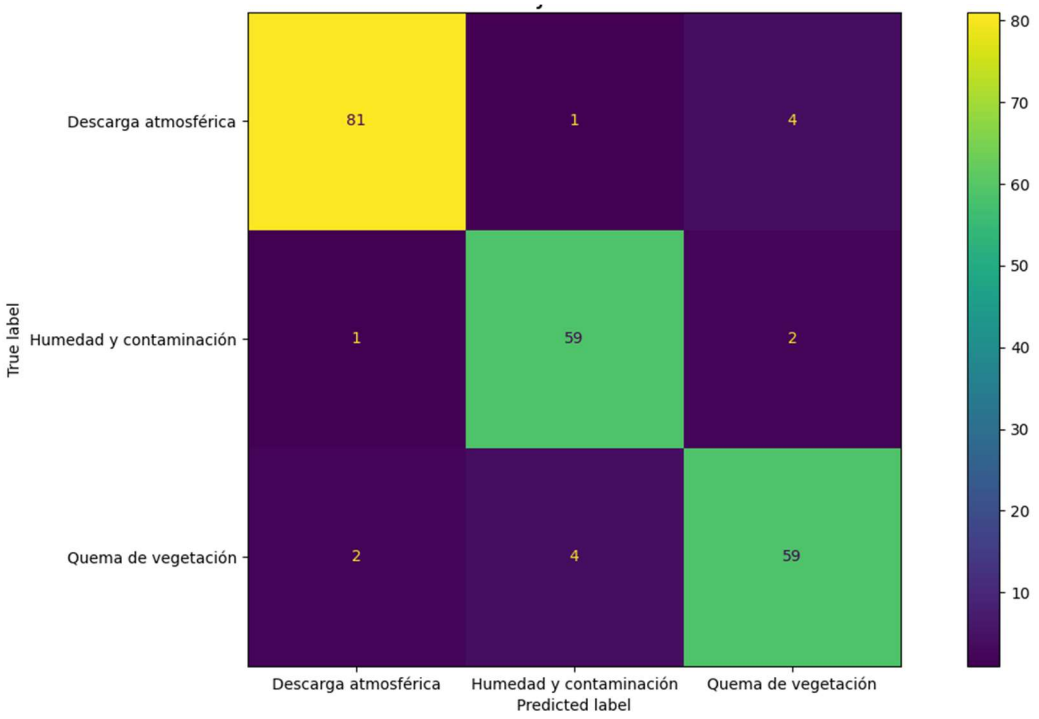


Figura N° 4.56: Matriz de confusión del modelo de clasificación a partir del conjunto de muestra de entrenamiento (Fuente: propia del autor)

En la **Figura N° 4.56**, de las 86 muestras de descarga atmosférica 81 fueron identificadas correctamente como descarga atmosférica y 5 muestras fueron identificadas incorrectamente. De las 62 muestras de humedad y contaminación, 59 muestras fueron identificadas correctamente como humedad y contaminación y 3 muestras fueron clasificada incorrectamente. De las 65 muestras de quema de

vegetación, 59 muestras fueron identificadas correctamente como quema de vegetación y 6 fueron clasificadas incorrectamente.

- **Modelo de clasificación puesto a prueba con el conjunto de muestra test**

En la **Figura N° 4.57**, se observa la matriz de confusión del modelo de clasificación puesto a prueba con el conjunto de muestra test:

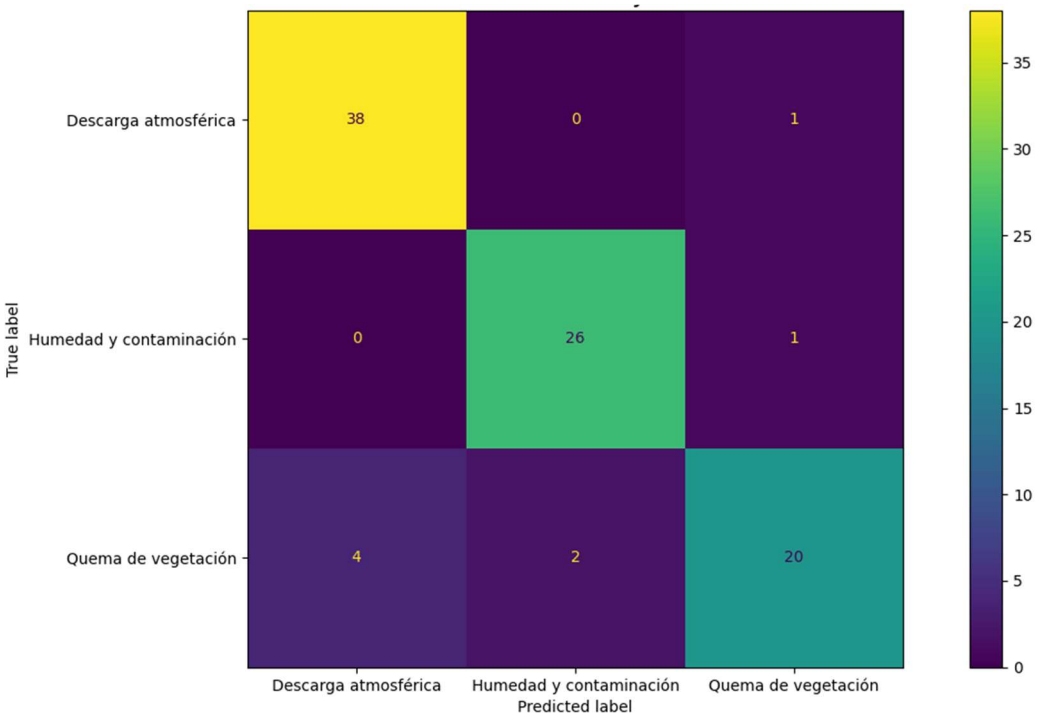


Figura N° 4.57: Matriz de confusión del modelo de clasificación a partir del conjunto de muestra test (Fuente: propia del autor)

En la **Figura N° 4.57**, se puede observar la matriz de consistencia de la muestra de test. De las 39 muestras de descarga atmosférica, 38 fueron identificadas correctamente como descarga atmosférica y 1 muestra fue identificada incorrectamente. De las 27 muestras de humedad y contaminación, 26 muestras fueron identificadas correctamente como humedad y contaminación y 1 muestra fue

clasificada incorrectamente. De las 26 muestras de quema de vegetación, 20 muestras fueron identificadas correctamente como quema de vegetación y 6 fueron clasificadas incorrectamente.

B. Métricas de evaluación

- **Modelo de clasificación entrenado con el conjunto de muestra de entrenamiento**

Reporte de clasificación a partir del conjunto de datos de entrenamiento:

	precision	recall	f1-score	support
Descarga atmosférica	0.94	0.91	0.92	86
Humedad y contaminación	0.90	0.98	0.94	62
Quema de vegetación	0.89	0.85	0.87	65
accuracy			0.91	213
macro avg	0.91	0.91	0.91	213
weighted avg	0.91	0.91	0.91	213

Figura N° 4.58: Reporte de clasificación del modelo a partir del conjunto de muestra de entrenamiento (Fuente: propia del autor)

De la **Figura N° 4.58**, se observa el reporte de clasificación del modelo entrenado a partir de la muestra de entrenamiento. La exactitud del modelo es de 91%. La sensibilidad para la causa raíz de descarga atmosférica, humedad y contaminación y quema de vegetación fue de 91%, 98% y 85% respectivamente.

- **Modelo de clasificación puesto a prueba con el conjunto de muestra test**

Reporte de clasificación a partir del conjunto de datos test:

	precision	recall	f1-score	support
Descarga atmosférica	0.97	0.97	0.97	39
Humedad y contaminación	0.87	0.96	0.91	27
Quema de vegetación	0.96	0.85	0.90	26
accuracy			0.93	92
macro avg	0.93	0.93	0.93	92
weighted avg	0.94	0.93	0.93	92

Figura N° 4.59: Reporte de clasificación del modelo a partir del conjunto de muestra test (Fuente: propia del autor)

De la **Figura N° 4.59**, se observa el reporte de clasificación del modelo a partir del conjunto de muestra test. exactitud del modelo es del 93%. La sensibilidad para la causa raíz de descarga atmosférica, humedad y contaminación y quema de vegetación fue de 97%, 96% y 85% respectivamente.

V. RESULTADOS

5.1 Resultados descriptivos

Con la implementación del aplicativo (basado en lenguaje de programación Python) para la extracción de características del archivo COMTRADE de fallas en líneas de transmisión, se logró consolidar un conjunto de datos de características contextuales, características en el dominio del tiempo y de la frecuencia, el cual se utilizó para generar el conjunto de muestra de entrenamiento (70%) y el conjunto de muestra test (30%).

Tabla N° 5.1: Parámetros estadísticos de características contextuales.

	Cantidad	Promedio	Desv std	Val mín	Cuartil Q1	Cuartil Q2	Cuartil Q3	Val máx
Mes	305.0	6.65	3.48	1.0	4.0	7.0	10.0	12.0
Hora	305.0	11.59	6.14	0.0	6.0	13.0	16.0	23.0
code_día	305.0	1.61	0.49	1.0	1.0	2.0	2.0	2.0
code_estación	305.0	2.61	1.16	1.0	2.0	3.0	4.0	4.0
code_región	305.0	1.58	0.73	1.0	1.0	1.0	2.0	3.0
Tensión kV	305.0	233.02	93.60	138.0	220.0	220.0	220.0	500.0
tipo_falla	305.0	1.33	0.70	1.0	1.0	1.0	1.0	4.0

Fuente: elaboración propia.

En la **Tabla N° 5.1**, se observa el valor promedio, desviación estándar, valor mínimo, cuartil Q1, cuartil Q2, cuartil Q3 y valor máximo de las características contextuales del conjunto de datos de fallas en líneas de transmisión (305 muestras).

Tabla N° 5.2: Parámetros estadísticos de características en el dominio de la frecuencia de las formas de onda de tensión y corriente de falla.

	Cantidad	Promedio	Desv std	Val mín	Cuartil Q1	Cuartil Q2	Cuartil Q3	Val máx
dif_i0max	305.0	62.20	75.83	0.26	19.30	39.58	67.81	444.65
dif_i1max	305.0	82.55	81.48	1.46	30.55	54.42	100.47	444.61
dif_i2max	305.0	83.82	78.57	3.24	33.85	59.10	105.90	451.15
Vrel_max0	305.0	183.77	172.77	0.91	73.50	126.44	244.23	1138.53
Vrel_max2	305.0	81.43	66.40	0.97	42.95	62.52	102.06	648.78
I0_half_cycle	305.0	348.73	524.90	0.51	66.69	164.65	414.29	4751.11
I0_one_cycle	305.0	724.23	1037.12	0.49	144.30	352.53	798.01	7130.56
I0_one_cycle_half	305.0	814.04	1161.71	0.71	165.92	394.47	912.52	8232.99
I0_two_cycle	305.0	806.78	1131.20	1.42	183.68	378.57	943.10	8378.16
I2_half_cycle	305.0	154.38	319.81	1.46	37.50	75.14	166.07	4076.16
I2_one_cycle	305.0	310.61	558.58	3.61	78.52	150.37	333.68	5456.74
I2_one_cycle_half	305.0	334.39	568.08	3.37	89.65	164.51	355.70	4927.81
I2_two_cycle	305.0	333.42	577.23	3.38	96.75	160.03	353.40	5606.47
Irel_max0	305.0	909.08	1222.94	3.25	244.61	469.91	1046.96	9143.87
Irel_max1	305.0	19.66	202.06	0.92	2.53	4.40	9.22	3532.30
Irel_max2	305.0	381.47	611.30	4.10	123.64	218.43	387.50	5606.47
Const time T0	305.0	0.39	0.29	0.06	0.20	0.30	0.50	2.00
Const time T1	305.0	0.37	0.29	0.00	0.20	0.30	0.50	2.00
Const time T2	305.0	0.39	0.29	0.06	0.20	0.30	0.50	2.00

Fuente: elaboración propia.

En la **Tabla N° 5.2**, se observa el valor promedio, desviación estándar, valor mínimo, cuartil Q1, cuartil Q2, cuartil Q3 y valor máximo de las características en el dominio de la frecuencia de las formas de onda de tensión y corriente, del conjunto de datos de fallas en líneas de transmisión (305 muestras).

Tabla N° 5.3: Parámetros estadísticos de características en el dominio del tiempo de las formas de onda de tensión y corriente de falla.

	Cantidad	Promedio	Desv std	Val mín	Cuartil Q1	Cuartil Q2	Cuartil Q3	Val máx
iR_rms	305.0	942.28	1515.08	27.99	166.17	302.67	1155.39	10777.37
iS_rms	305.0	1046.34	1485.70	21.44	163.07	356.49	1329.26	10744.93
iT_rms	305.0	1013.32	1500.68	10.67	190.73	329.02	1366.67	11117.26
vR_rms	305.0	112807.77	58593.41	13511.58	75662.37	116247.62	129525.10	299822.40
vS_rms	305.0	108960.62	55856.41	9168.57	74544.87	114255.13	127885.45	294256.40
vT_rms	305.0	109408.28	55447.02	9245.85	71928.04	118672.15	129150.88	302364.59
iR_root	305.0	717.07	1154.60	16.09	123.07	234.20	920.15	8215.25
iS_root	305.0	798.91	1127.81	13.05	124.46	262.90	1020.57	8516.10
iT_root	305.0	766.61	1125.55	4.87	142.61	254.16	1074.14	8206.61
vR_root	305.0	91687.36	48143.29	6149.95	62006.46	94481.68	106688.97	248452.47
vS_root	305.0	88644.37	46023.41	5224.74	60940.76	93078.82	105387.57	242876.56
vT_root	305.0	88852.71	45808.28	5769.46	58081.78	97811.94	105594.66	247777.23
iR_std	305.0	922.51	1477.13	20.93	166.76	302.12	1158.37	10359.25
iS_std	305.0	1025.11	1440.22	21.50	163.06	353.48	1310.96	10484.48
iT_std	305.0	989.41	1447.57	10.63	186.46	326.28	1341.18	10422.72
vR_std	305.0	113164.00	58655.49	13529.56	75964.58	116768.84	130096.08	299375.26
vS_std	305.0	109308.92	56027.16	8731.50	74682.54	115014.36	128086.61	294847.38
vT_std	305.0	109763.76	55489.66	8169.58	72162.87	119045.42	129478.92	300872.59
vR_fac_crest	305.0	1.75	0.72	1.37	1.44	1.49	1.68	7.10
vS_fac_crest	305.0	1.76	0.75	1.39	1.45	1.51	1.68	6.47
vT_fac_crest	305.0	1.79	0.75	1.37	1.45	1.50	1.71	7.28
iR_fac_imp	305.0	2.11	0.60	1.54	1.78	2.00	2.26	9.48
iS_fac_imp	305.0	2.13	0.54	1.56	1.84	2.00	2.31	6.67
iT_fac_imp	305.0	2.13	0.57	1.54	1.84	2.01	2.25	9.19

Fuente: elaboración propia.

En la **Tabla N° 5.3**, se observa el valor promedio, desviación estándar, valor mínimo, cuartil Q1, cuartil Q2, cuartil Q3 y valor máximo de las características en el dominio del tiempo de las formas de onda de tensión y corriente, del conjunto de datos de fallas en líneas de transmisión (305 muestras).

5.2 Resultados inferenciales

Con el entrenamiento del modelo de Machine Learning basado en algoritmo k-nearest neighbors de clasificación de aprendizaje supervisado, se realizó el aprendizaje con el conjunto de muestra de entrenamiento (70% del total de muestras) y se sometió a prueba con el conjunto de muestra test (30% del total de muestras).

Reporte de clasificación a partir del conjunto de datos test:

	precision	recall	f1-score	support
Descarga atmosférica	0.97	0.97	0.97	39
Humedad y contaminación	0.87	0.96	0.91	27
Quema de vegetación	0.96	0.85	0.90	26
accuracy			0.93	92
macro avg	0.93	0.93	0.93	92
weighted avg	0.94	0.93	0.93	92

Figura N° 5.1: Reporte de clasificación del modelo a partir del conjunto de muestra de test (Fuente: propia del autor)

En la **Figura N° 5.1**, se observa el resultado de las métricas de evaluación del modelo de clasificación puesto a prueba con el conjunto de muestra test. La exactitud del modelo (accuracy) es del 93%. La sensibilidad para la clase descarga atmosférica es 97%, la sensibilidad (recall) para la clase humedad y contaminación es 96% y sensibilidad para la clase quema de vegetación es 85%. La precisión para la clase descarga atmosférica es 97%, la precisión para la clase humedad y contaminación es 87% y sensibilidad para la clase quema de vegetación es 96%. La métrica f1-score para la clase descarga atmosférica es 97%, para la clase humedad y contaminación es 91% y para la clase quema de

vegetación es 90%. Del resultado de la métrica F1-score para cada causa raíz, se deduce que es coherente con el resultado de las métricas de exactitud y precisión.

5.3 Otro tipo de resultados estadísticos

En la **Tabla N° 5.4**, se puede observar un resumen de los resultados de las métricas de evaluación del modelo de clasificación para el conjunto de muestra de entrenamiento y conjunto de muestra test.

Tabla N° 5.4: Tabla comparativa de resultado de métricas de evaluación.

Métrica de evaluación		Entrenamiento (%)	Test (%)
Precisión global del modelo		91	93
Descarga atmosférica	Sensibilidad	91	97
	Precisión	94	97
	f1-score	92	97
Humedad y contaminación	Sensibilidad	98	96
	Precisión	90	87
	f1-score	94	91
Quema de vegetación	Sensibilidad	85	85
	Precisión	89	96
	f1-score	87	90

Fuente: elaboración propia.

De la **Tabla N° 5.4**, se observó que el resultado de las métricas de evaluación del modelo de clasificación para los conjuntos de muestra de entrenamiento y test, tienen valores muy similares, lo que significa que el modelo entrenado no realizó overfitting. Del resultado de las métricas de evaluación para el conjunto de muestras test, se deduce que el modelo tiene una alta sensibilidad para la

identificación de la causa raíz (habilidad para identificar la causa raíz) y una alta precisión (grado de certeza o confianza de que el resultado es correcto).

VI. DISCUSIÓN DE RESULTADOS

6.1 Contrastación y demostración de la hipótesis con los resultados

Hipótesis general

La implementación de un aplicativo basado en algoritmo de Machine Learning logró identificar automáticamente la causa raíz de fallas en líneas de transmisión del grupo ISA Perú.

Hipótesis específica

H.E.1 El desarrollo de un aplicativo basado en lenguaje de programación Python, logró extraer las características contextuales, características en el dominio del tiempo y de la frecuencia de los archivos COMTRADE de las fallas en líneas de transmisión.

H.E.2 El entrenamiento del modelo de Machine Learning basado en algoritmo de clasificación de aprendizaje supervisado, logró identificar automáticamente la causa raíz de fallas en líneas de transmisión del grupo ISA Perú, con una exactitud del 93%.

6.2 Contrastación de los resultados con otros estudios similares

Flores (2021) realizó un estudio titulado “Identificación de causa raíz de fallas por descargas eléctricas en líneas de transmisión”, en la cual el modelo entrenado obtuvo una exactitud de 93.75%. Por otra parte, Minnaar, Niccols y Gaunt (2015) realizaron un estudio sobre “Automating transmisión line fault root causa analysis”, en la que el modelo entrenado obtuvo una exactitud de 90%. Por otra parte, Reveco (2019) realizó un estudio sobre “Análisis predictivo de activos mineros para obtención de intervalo de falla mediante algoritmos de Machine Learning”, en la que el modelo entrenado obtuvo una exactitud de clasificación

de falla del 90.54%. Por otra parte, Adauto (2021) realizó un estudio sobre “Aplicación de inteligencia artificial en la detección de fallas en los motores eléctricos de corriente continua de imán permanente”, en la que el modelo obtuvo una exactitud del 98% y f1-score de 11.59%, lo cual implica que el modelo no tiene clasifica correctamente. Por otra parte, Gallegos (2020) realizó un estudio sobre “Identificación de fallas en sistemas eléctricos de potencia basado en reconocimiento de patrones”, en la que la mejor exactitud obtenida por el modelo fue del 93.33% Por otra parte, Bautista (2018) realizó un estudio titulado “Identificación de 11 tipos de fallas en líneas de transmisión de alta tensión utilizando redes neuronales”, en la que el mejor resultado de exactitud del modelo entrenado, fue del 99.8% para el tramo 4. Por otra parte, Barrera, Meléndez, Kulkarni y Santoso (2012) realizaron un estudio sobre “Feature análisis and automatic classification of short-circuit faults resulting from external causes”, en la que el resultado de exactitud de clasificación fue del 93.4%.

En el presente estudio, el modelo entrenado obtuvo una exactitud del 93%.

Se verificó que la exactitud del modelo entrenado en el presente estudio guarda una alta relación con las exactitudes obtenidas en los modelos de los estudios precedentes.

VII. CONCLUSIONES

En la investigación realizada se llegó a la conclusión que la implementación del aplicativo basado en algoritmo de Machine Learning logró identificar automáticamente la causa raíz de fallas en líneas de transmisión del grupo ISA Perú.

1) Los resultados de la investigación confirman que la implementación del aplicativo basado en lenguaje de programación Python, transformó los datos obtenidos del archivo COMTRADE de las fallas en líneas de transmisión, logrando caracterizar las causas raíces de falla, en base a características contextuales, características en el dominio del tiempo y características en el dominio de la frecuencia.
2) Los resultados de la investigación confirman que el modelo de Machine Learning basado en algoritmo k-nearest neighbors de clasificación de aprendizaje supervisado (basado en lenguaje de programación Python), identificó automáticamente la causa raíz de fallas en líneas de transmisión del grupo ISA Perú, con una exactitud del 93%.

VIII. RECOMENDACIONES

Al comprobarse que la implementación del aplicativo basado en algoritmo de Machine Learning logró identificar automáticamente la causa raíz de fallas en líneas de transmisión del grupo ISA Perú, se da como primera recomendación incluir en el estudio más causas raíces de fallas en líneas de transmisión (acercamiento de vegetación, nevada, caída de conductor u otros), teniendo la consideración la aplicación de algoritmos que trabajen con conjunto de datos desbalanceados o balanceados.

1) Se recomienda incluir dentro de la ingeniería de variables la resistencia de la falla. Dicha resistencia de falla es posible de ser calculada a partir de los archivos COMTRADE generados en ambos extremos de la línea de transmisión.
2) Se recomienda explorar el entrenamiento del modelo de Machine Learning con otros algoritmos de aprendizaje supervisado, tales como random forest, regresión logística y máquina de soporte vectorial, con la finalidad de evaluar el resultado de las métricas.

IX. REFERENCIAS BIBLIOGRÁFICAS

[1] ANDERSON, M. Paul. Analysis of Faulted Power Systems. Estados Unidos. The Institute of Electrical and Electronics Engineers. 1995.

[2] DATA SCIENCE RESEARCH PERU. Diapositivas del curso de Especialización de Análisis de datos con Python. Lima Perú. Octubre, 2020.

[3] VYT CONTRATISTAS. Informe de inspección ligera L-2232. La Libertad Perú. Abril, 2019.

[4] VYT CONTRATISTAS. Informe de inspección por falla en L-2232. La Libertad Perú. Agosto, 2021.

[5] VYT CONTRATISTAS. Informe de inspección por falla en L-2232. La Libertad Perú. Mayo, 2019.

[6] KINDERMAN, Geraldo. Curto-Circuito. Porto Algre Brasil. 1997.

[7] PIANETA ESCUDERO, Alfredo Miguel. Modelo adaptativo de inteligencia artificial para detección selectiva de fallas de alta impedancia en líneas de transmisión de dos terminales de doble circuito. Medellín Colombia. Octubre 2015.

[8] ARAUZ GALLEGOS, Jonathan Fernando. Identificación de fallas en sistemas eléctricos de potencia basado en reconocimiento de patrones. Quito Ecuador. Febrero 2020.

[9] HURTADO Luini, VILLAREAL Edwin, VILLAREAL Luís. Detección y diagnóstico de fallas mediante técnicas de inteligencia artificial, un estado del arte. Bogotá Colombia. Julio, 2016.

[10] ADAUTO ARANA, Ricardo Michael. Aplicación de la inteligencia artificial en la detección de fallas en los motores eléctricos de corriente continua de imán permanente. Huancayo Perú. Mayo, 2021.

[11] REVECO DIAZ, María Ignacia. Análisis predictivo de activos mineros para obtención de intervalo de falla mediante algoritmos de Machine Learning. Santiago de Chile. 2019.

[12] ANDERSON, M. Paul. Power System Protection. Estados Unidos. The Institute of Electrical and Electronics Engineers. 1999.

[13] KHANA Rahul y AWAD Mariette. Efficient Learning Machines: Theories, Concepts and Applications for Engineers and System Designers. Abril, 2015.

[14] MAINI Vishal y SABRI Samer. Machine Learning for Humans. Agosto, 2017.

[15] ANGULO, Germán. Separata de curso Operación y Configuración equipo de prueba Omicron CMC 356. Lima Perú, 2021.

[16] TRANSENER. Fundamentos de Protección de Transmisión. Lima Perú. 2002.

[17] lEEE. Standard Common format for transient data exchange (COMTRADE) for power systems. 2013.

[18] COSTELLO David y ZIMMERMAN Karl. Determining the Faulted Phase. Diciembre, 2015.

[19] WANG Xiang, ZHENG Yuan, ZHAO Zhenzhou y WANG Jinping. Bearing Fault Diagnosis Based on Statistical Locally Linear Embedding. China. 2015.

[20] KHANNA Rahul y AWAD Mariette. Efficient Learning Machines: Theories, Concepts, adn Applications for Engineers and System Designers. Abril, 2015.

[21] ZIJIE Zhu, MENGTIAN Zhang. K-Nearest Neighbors (KNN) Classification with Different Distance Metrics. Mayo, 2020.

ANEXOS

- Anexo 1: Código del aplicativo basado en algoritmo k-nearest neighbors de clasificación de aprendizaje supervisado de Machine Learning.

Anexo 1

Código del aplicativo basado en algoritmo k-nearest neighbors de clasificación de aprendizaje supervisado de Machine

```
import pandas as pd
import numpy as np
from sklearn.neighbors import KNeighborsClassifier
from tkinter import filedialog as fd
from sklearn.decomposition import PCA
import matplotlib.pyplot as plt
np.seterr(divide='ignore', invalid='ignore')

# Establece el ajuste para ver todo el contenido de un DataFrame
pd.set_option('display.max_rows', None)
pd.set_option('display.max_columns', None)
pd.set_option('display.width', None)
pd.set_option('display.max_colwidth', None)
pd.set_option('display.colheader_justify', 'center')

#Importación del conjunto de datos
df =
pd.read_excel('D:\Proy_Oscilografías\Origen\db_master_rev1_tesis.xlsx'
)
#Se asigna la columna de la variable objetivo
y = df['Causa raíz']

#Se define las características predictoras (variables independiente)
predictoras =
['Mes','Hora','code_día','code_estación','code_región','Tensión
kV','tipo_falla','dif_i0max','dif_i1max','dif_i2max','Vrel_max0','Vrel
_max2','I0_half_cycle','I0_one_cycle','I0_one_cycle_half','I0_two_cycl
e','I2_half_cycle','I2_one_cycle','I2_one_cycle_half','I2_two_cycle','
Irel_max0','Irel_max1','Irel_max2','Const time T0','Const time
T1','Const time
T2','iR_rms','iS_rms','iT_rms','vR_rms','vS_rms','vT_rms','iR_root','i
S_root','iT_root','vR_root','vS_root',
'vT_root','iR_std','iS_std','iT_std','vR_std','vS_std','vT_std','vR_fa
c_crest','vS_fac_crest','vT_fac_crest','iR_fac_imp','iS_fac_imp','iT_f
ac_imp']

#Se genera el dataframe de observaciones
X = df[predictoras]
X = pd.DataFrame(X,columns=X.columns)

#Se divide el conjunto de datos en sub-conjunto de entrenamiento y
test
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.3, random_state=123)
print("Cantidad de muestra de entrenamiento:",len(X_train))
print("Cantidad de muestra de test:",len(X_test))

#Se realiza el escalado de cada atributo
from sklearn.preprocessing import MinMaxScaler
sc = MinMaxScaler()
sc.fit(X_train)
X_train = sc.transform(X_train)
X_train = pd.DataFrame(X_train,columns=X.columns)

#Se aplica el análisis de componentes principales
n_pca = 8
pca = PCA(n_components=n_pca)
pca.fit(X_train)
```

```
X_train = pca.transform(X_train)

#Se realiza el entrenamiento del algoritmo de clasificación KNN.
knn = KNeighborsClassifier(metric=' chebyshev', n_neighbors=7)
# Entrenamiento del modelo
knn.fit(X_train,y_train)
# Predecimos la data de entrenamiento
train_pred=knn.predict(X_train)

#Se aplica el escalado al subconjunto de muestra de test
X_test = sc.transform(X_test)
#Se aplica la reducción de dimensionalidad
X_test = pca.transform(X_test)
# Predecimos la data del test
test_pred=knn.predict(X_test)

# Calculando las principales métricas de evaluación
from sklearn import metrics
from sklearn.metrics import classification_report, confusion_matrix,
accuracy_score, ConfusionMatrixDisplay
result_train = confusion_matrix(y_train, train_pred)
result_test = confusion_matrix(y_test, test_pred)
print("\nMatriz de confusión de muestra entrenamiento:")
print(result_train)

disp =
ConfusionMatrixDisplay(confusion_matrix=result_train,display_labels=kn
n.classes_)
disp.plot()

plt.title("Matriz de confusión del conjunto de datos de
entrenamiento",fontsize = 18)
plt.show()

disp2 =
ConfusionMatrixDisplay(confusion_matrix=result_test,display_labels=knn
.classes_)
disp2.plot()

plt.title("Matriz de confusión del conjunto de datos Test",fontsize =
18)
plt.show()

result_train = classification_report(y_train, train_pred)
print("\nReporte de clasificación a partir del conjunto de datos de
entrenamiento:")
print (result_train)
accuracy_global_train = metrics.accuracy_score(y_train,train_pred)
print("Accuracy:",np.round(accuracy_global_train,2)*100,"%")

print("\nMatriz de confusión de muestra test:")
print(result_test)
result_test = classification_report(y_test, test_pred)
print("\nReporte de clasificación a partir del conjunto de datos
test:")
print (result_test)
accuracy_global_test = accuracy_score(y_test,test_pred)
print("Accuracy:",np.round(accuracy_global_test,2)*100,"%")
```

Nota: La propiedad del código corresponde a Red de Energía del Perú S.A.