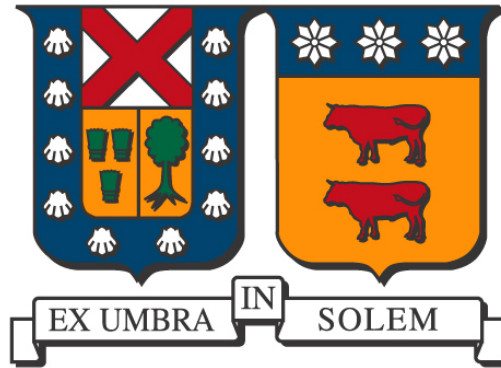


UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA
DEPARTAMENTO DE ELECTRÓNICA
VALPARAÍSO - CHILE



“DISEÑO E IMPLEMENTACIÓN DE UN DETECTOR
INTELIGENTE DE ANOMALÍAS EN CARTILLAS DE
INSPECCIÓN DE LÍNEAS DE TRANSMISIÓN
ELÉCTRICA”

FABRICIO CAMILO ROSALES SOTO

MEMORIA DE TITULACIÓN PARA OPTAR AL TÍTULO DE INGENIERO CIVIL
TELEMÁTICO

PROFESOR GUÍA: MAURICIO ALEJANDRO ARAYA LÓPEZ

PROFESOR CORREFERENTE: ROGER SCHURCH BRANDT

MARZO - 2022

Agradecimientos

Este trabajo no hubiera sido posible de no ser por quienes me han apoyado todos estos años. Desde mi familia, hasta mis amigos y conocidos.

En especial quisiera agradecer a mis padres, quienes me dieron la vida y que, junto con mis abuelos, me hicieron quién soy, con sus excelentes valores y principios, virtudes y defectos, luces y sombras, me enseñaron lo que es ser humano. También a mi pareja por su apoyo y compañía incondicional, preguntándome y animándome todas las semanas para terminar este trabajo. Por otro lado a mis variados grupos de amigos: los del colegio, con los cuales las estupideces y risas, nunca han faltado; los de la universidad y en especial a “*Los Mor-Mon(es)*”, con sus memes y esas juntas tan tranquilas que los caracteriza; y a los del trabajo *a.k.a.* “*La Consultoría*”, que me han ayudado a crecer bastante tanto personal como profesionalmente.

Adicionalmente dedico este trabajo a mi tata Q.E.P.D., quien siempre me preguntaba cómo me iba en la universidad y al momento de estar en este proceso, falleció a causa de la pandemia. Ojalá hubieras vivido lo suficiente para ver a tu nieto terminar este proceso, pero sé que de alguna forma estuviste presente igualmente.

Resumen

En la industria del transporte de energía eléctrica, se han propuesto diferentes metodologías para detectar preventiva o predictivamente la falla de un aislador eléctrico. Esto se puede producir debido al deterioro en su capacidad de aislación por la formación de una capa de contaminación. La consecuencia de esta pérdida puede llevar a una inestabilidad e incluso al fallo del sistema eléctrico. Sin embargo esto puede ser evitado con un oportuno mantenimiento, es decir, un lavado de aislación. En este trabajo se analizan datos correspondientes a aisladores ubicados en torres de alta tensión, con el objetivo de generar un sistema de detección de anomalías que sirva como complemento para la planificación del mantenimiento. Esto permite crear un indicador no subjetivo y en base a una metodología que indique la urgencia de mantenimiento. El sistema final considera dos técnicas para la detección de anomalías: Z-Score, para encontrar anomalías de tipo valores extremos dentro de una torre de alta tensión; e Isolation Forest, beneficiándose de la capacidad de aislación de puntos específicos al ser entrenado con la totalidad de datos. Los resultados obtenidos muestran que utilizando la información proveída por el detector de anomalías se logra una mejor división de la clase anómala original, la cual corresponde a la de mantenimiento con urgencia (es decir, está más cerca de la inminente falla).

Palabras claves: detección de anomalía, aprendizaje de máquinas, aislador eléctrico, efecto corona, alta tensión

Abstract

In the electrical energy transport industry, different methodologies have been proposed to preventively or predictively detect the failure of an electrical insulator. This can occur due to deterioration in its insulation capacity due to the formation of a layer of contamination. The consequence of this loss can lead to instability and even failure of the electrical system. However, this can be avoided with timely maintenance, that is, an insulation wash. In this work, data corresponding to insulators located in high voltage towers are analyzed, with the aim of generating an anomaly detection system that serves as a complement for maintenance planning. This allows the creation of a non-subjective indicator and based on a methodology that indicates the urgency of maintenance. The final system considers two techniques for the detection of anomalies: Z-Score, to find extreme value anomalies inside a high voltage tower; and Isolation Forest, benefiting from the ability to isolate specific points when trained with all the data. The results obtained show that using the information provided by the anomaly detector, a better division of the original anomalous class is achieved, which corresponds to the urgent maintenance class (i.e., it is closer to the imminent failure).

Keywords: anomaly detection, machine learning, electrical insulator, corona effect, high voltage

Glosario

- Aislador eléctrico: Elemento mecánico que sujeta la línea eléctrica a la estructura.
- Cartilla: Cartilla de inspección o plantilla utilizada en la toma de datos en terreno por los técnicos a cargo de la empresa.
- Conteo descarga SD: Hace referencia al conteo de descarga para la sección superior o derecha de la estructura de alta tensión, ocurrida en sus aisladores.
- Conteo descarga MM: Hace referencia al conteo de descarga para la sección media de la estructura de alta tensión, ocurrida en sus aisladores.
- Conteo descarga IZ: Hace referencia al conteo de descarga para la sección inferior o izquierda de la estructura de alta tensión, ocurrida en sus aisladores.
- Conteo de descargas: Indicador de la cantidad de descargas parciales ocurridas en cierto sector (aislador), entregado por la CoroCam 6D.
- CNN: Convolutional Neural Network o red neuronal convolucional, usada principalmente en imágenes o cuando hay matrices.
- Cosine Kernel: Función Kernel basada en coseno.
- Descarga de fase: Variable relacionada con la pérdida de la capacidad aislante del aislador eléctrico, descarga parcial.
- Dimensión: Término usado para referirse al tamaño del espacio del problema dado sus variables, por ej. tres variables implican tres dimensiones.
- Grid Search: Técnica usada para encontrar hiperparámetros óptimos de un modelo.
- Hiperparámetro: Característica externa del modelo, que no depende de los datos.
- Kernel: Función utilizada para mapear datos a una dimensión superior, con el fin de separar estos más fácilmente.
- KNN: K-Nearest Neighbor o k vecino más cercano, algoritmo de aprendizaje supervisado.
- Linear Kernel: Función Kernel con base lineal.
- ML: Machine Learning o Aprendizaje de Máquinas.
- NN: Neural Network o Red Neuronal, técnica utilizada para resolver diferentes problemas en Machine Learning.

- P75: Percentil 75.
- PCA: Principal Component Analysis o Análisis de Componentes Principales, método estadístico usado en reducción de dimensionalidad.
- Polynomial Kernel: Función Kernel con base polinómica.
- RBF Kernel: Función Kernel con base radial.
- SVM: Support Vector Machine, técnica de Machine Learning de aprendizaje supervisado.
- Sigmoid Kernel: Función Kernel basada en tangente hiperbólica.
- Torre de alta tensión: Estructura alta que sostiene al aislador eléctrico.
- UV: Ultra Violeta.

Índice general

1. Introducción	1
1.1. Propósito	2
1.2. Alcance	2
1.3. Definición del problema	2
1.3.1. Metodología de inspección y mantenimiento	3
1.3.2. Objetivos de la memoria	3
1.3.3. Requerimientos	4
1.4. Estructura de la memoria	5
2. Machine Learning y aisladores	7
2.1. Machine Learning	7
2.1.1. Etapas de Machine Learning	9
2.2. Detección de anomalías	10
2.2.1. Z-Score	10
2.2.2. Local Outlier Factor	11
2.2.3. Isolation Forest	11
2.3. Aisladores eléctricos	11
2.3.1. Contaminación y descargas	11
2.3.2. Efecto corona	12
2.4. Trabajo relacionado	12
2.4.1. Basado en ultrasonido	12
2.4.2. Basado en imágenes del espectro visible	13
2.4.3. Basado en imágenes del espectro UV	13
2.4.4. Basado en mediciones en laboratorio	14
2.5. Cartilla de inspección	15

3. Definición del problema	17
3.1. Identificación de requerimientos	17
3.2. Arquitectura del sistema	20
3.2.1. Esquema contextual del sistema	20
3.2.2. Esquema general del sistema	20
3.2.3. Esquema para el desarrollo del sistema	21
4. Diseño de la solución	23
4.1. Preparación del conjunto de datos	23
4.1.1. Adquisición de datos	23
4.1.2. Limpieza de datos	23
4.1.3. Extracción de características y almacenamiento	24
4.2. Resumen y visualización	25
4.2.1. Análisis local	25
4.2.2. Análisis global	27
4.3. Entrenamiento del modelo	30
4.3.1. <i>Clustering</i>	31
4.4. Detección de anomalías	34
4.4.1. Z-Score	34
4.5. Evaluación del modelo	37
4.6. Implementación	43
4.6.1. Detector de anomalías	43
4.6.2. Integración en Excel	43
5. Conclusiones	45
5.1. Conclusiones generales	45
5.2. Conclusiones específicas	46
5.3. Proyecciones	47
Anexos	49
A.1. Resumen estadístico 14 cartillas	49
A.2. t-SNE con diferente <i>perplexity</i>	50
A.2.1. Variables categóricas y numéricas	50
A.2.2. Solo variables numéricas	51
A.2.3. Solo con descargas de fase	52

Índice de figuras

2.1. Ejemplo de muestras en cartilla de inspección	8
2.2. Diagrama del proceso de Machine Learning utilizado, elaboración propia	9
2.3. Variables asociadas a cada aislador eléctrico	16
3.1. Ejemplo de estructura de cartilla de inspección	17
3.2. Diagrama de contexto del problema, elaboración propia	19
3.3. Diagrama general del sistema, elaboración propia	20
3.4. Diagrama de digitalización de datos, elaboración propia	21
4.1. Ejemplo de cartilla archivo 10068897_Informe Corona LLTT SL-AS.xlsx	25
4.2. Resumen estadístico de pandas de 10068897_Informe Corona LLTT SL-AS.xlsx	25
4.3. Matriz de correlación de 10068897_Informe Corona LLTT SL-AS.xlsx	26
4.4. Distribución de conteo de descargas de 10068897_Informe Corona LLTT SL-AS.xlsx	27
4.5. Distribución de variables ambientales de 10068897_Informe Corona LLTT SL-AS.xlsx	27
4.6. Resumen estadístico global variables cuantitativas	28
4.7. Resumen estadístico global variables cualitativas	28
4.8. Distribución global de conteo de descargas	28
4.9. Distribución global de temperatura y humedad relativa	29
4.10. Matriz de correlación global	29
4.11. Humedad vs temperatura global	30
4.12. Conteo de descargas vs humedad relativa global	30
4.13. Visualización mediante t-SNE incluyendo variables categóricas, con parámetro <i>perplexity</i> = 10	31
4.14. Visualización mediante t-SNE sin variables categóricas, con parámetro <i>perplexity</i> = 30	31
4.15. PCA utilizando <i>kernel rbf</i>	32
4.16. PCA utilizando <i>kernel poly</i>	32

4.17. PCA utilizando <i>kernel sigmoid</i>	33
4.18. PCA utilizando <i>kernel cosine</i>	33
4.19. PCA utilizando <i>kernel linear</i>	33
4.20. histograma anomalías Z-Score	34
4.21. histograma Z-Score con el valor respectivo de conteo de descarga	34
4.22. Datos anómalos Z-Score y temperatura	35
4.23. Datos anómalos Z-Score y humedad	35
4.24. histograma anomalías LoF	35
4.25. Datos anómalos LoF y temperatura	36
4.26. Datos anómalos LoF y humedad	36
4.27. histograma anomalías IForest	36
4.28. Datos anómalos IForest y temperatura	37
4.29. Datos anómalos IForest y humedad	37
4.30. conteo de descargas anómalas Z-Score	37
4.31. conteo de descargas anómalas Local Outlier Factor	38
4.32. conteo de descargas anómalas Isolation Forest	38
4.33. histograma valores normales Z-Score	39
4.34. histograma valores normales IForest	39
4.35. histograma valores normales LoF	40
4.36. Boxplot	40
4.37. Comparación descarga de fase, etiquetas nuevas versus originales	42
4.38. detector de anomalías en la cartilla de inspección	44
5.1. Proyección del diseño de la solución predictiva	47
A.2. t-SNE con <i>perplexity</i> 10	50
A.3. t-SNE con <i>perplexity</i> 50	50
A.4. t-SNE con <i>perplexity</i> 80	50
A.5. t-SNE sin variables categóricas con <i>perplexity</i> 20	51
A.6. t-SNE sin variables categóricas con <i>perplexity</i> 50	51
A.7. t-SNE sin variables categóricas con <i>perplexity</i> 80	51
A.8. t-SNE utilizando descargas de fase con <i>perplexity</i> 10	52
A.9. t-SNE utilizando descargas de fase con <i>perplexity</i> 30	52
A.10. t-SNE utilizando descargas de fase con <i>perplexity</i> 50	52
A.11. t-SNE utilizando descargas de fase con <i>perplexity</i> 80	53

Índice de tablas

4.1. Resumen estadístico conjuntos de anomalías	38
4.2. Medidas de desempeño del detector de anomalía, grupos EPRI A-B y EPRI C (anómalo)	41
4.3. Medidas de desempeño del detector de anomalía, grupos EPRI A, EPRI B y EPRI C (anómalo, re-etiquetado)	41
4.4. Cantidad de datos y sus etiquetas, original y nueva	43
A.1. Resumen estadístico temperatura y humedad 14 cartillas individuales	49

Capítulo 1

Introducción

Chilquinta energía es una compañía encargada de distribuir energía eléctrica en varias comunas de la región de Valparaíso, comprendiendo un área de cobertura de $11.496[km^2]$. La empresa debe preocuparse de brindar un servicio de calidad con una baja tasa de interrupciones del suministro en términos de tiempo y frecuencia, pero este objetivo se dificulta con la aparición de fenómenos eléctricos que afectan a componentes de las red de transmisión de energía.

La infraestructura encargada de la transmisión consta de varios elementos que permiten su operación, entre los cuales se encuentran el cable para el transporte de electricidad en alta tensión y el aislador. Este último tiene la función de soportar la tensión mecánica y eléctrica entre sus terminales, siendo el sujetador del cable a la estructura sostenedora (típicamente una torre de alta tensión), al mismo tiempo de evitar una conexión a tierra u otros conductores (como el material de la torre).

El aislador eléctrico puede ver degradada su capacidad de aislante debido a factores externos caóticos, como pueden ser el polvo, la humedad, la temperatura, el viento que trae partículas o el humo, es decir, agentes asociados al medio ambiente o como consecuencias de la acción humana. En general, la capa de suciedad que se acumula en su superficie provocará la aparición de descargas parciales, que sin una acción oportuna pueden desencadenar contorneo en el aislador, provocando cortes de energía.

Actualmente, esta problemática se resuelve de manera relativamente manual, en donde a través de la observación de fotografías y bajo criterio humano, se determina la programación de mantenimientos preventivos para asegurar el correcto funcionamiento de los aisladores. Esto resulta en un proceso poco eficiente e inexacto, generando una gestión inadecuada de recursos.

Con el objetivo de mejorar este proceso, Chilquinta ha propuesto un desafío para generar una solución de carácter predictivo mediante técnicas de Machine Learning. Sin embargo, construir una solución factible utilizando esta metodología dependerá enormemente de los datos que se tengan a la fecha, tanto su cantidad como formato.

1.1. Propósito

El presente trabajo tiene como propósito entregar una herramienta a la empresa que sirva como guía, para el complemento en la toma de decisiones a la hora de planificar los mantenimientos. Para llegar a ello se realiza una investigación del estado del arte para enfrentar este tipo de problemáticas. Luego se contrasta este con los datos a ser aportados por la empresa, en conjunto a su manera de realizar el proceso de mantenimiento. Una vez comprendido y analizado lo anterior, se exploran alternativas de solución para llegar a la herramienta que se requiera, junto con la construcción de la que mayor aporte pueda realizar a la empresa al corto y mediano plazo, en miras de un sistema predictivo.

1.2. Alcance

El alcance de esta memoria es contribuir al mejoramiento del proceso de planificación de inspecciones y/o mantenimiento de la empresa Chilquinta, dentro del contexto de aisladores eléctricos en torres de alta tensión, desarrollando un detector de anomalías que aproveche datos recopilados por el equipo técnico en terreno. La herramienta se perfila como una solución tipo *plug-and-play*, integrándose en la cartilla de inspección que utiliza la empresa a la fecha de esta memoria.

Para lograr esto, se conforma un equipo multidisciplinario compuesto por:

- Carlos Marcellini Yus: memorista de Ing. en Diseño de Productos, encargado de la experiencia de usuario, enfocándose en recursos humanos y estandarización de procedimientos, diseñando un manual de inspección unificado.
- Fabricio Rosales Soto: memorista de Ing. Civil Telemática, responsable de diseñar, implementar e integrar la herramienta, mediante uso de metodologías de minería de datos y técnicas de aprendizaje de máquinas.
- Ignacio Reyes Retamales: memorista de Ing. Eléctrica, encargado de evaluar los factores y fenómenos eléctricos que se producen, proponiendo una serie de mejoras en el proceso de inspección de la empresa, incluyendo el instrumento de registro (cartilla).

1.3. Definición del problema

Chilquinta se presenta ante el equipo con la siguiente pregunta: ¿Cómo podemos establecer una metodología de análisis sustentada en el método científico, cuyo resultado sean planes de limpieza de aislación focalizados para cada una de las líneas de transmisión? Por lo tanto, debe incluir: aspectos de

diseño de la instalación, zonas geográficas por donde atraviesan condiciones ambientales, mediciones coronográficas, nivel de contaminación, corrientes de fuga y resultados de las inspecciones visuales.

De lo anterior, y considerando reuniones con la contraparte de este desafío, se desprenden los siguientes problemas específicos a resolver:

1. Falta de una metodología científica de análisis establecida de variables para realizar limpieza de aisladores.
2. Carencia de conocimiento si las mantenciones se sobreestiman o subestiman.
3. Inexistencia de un modelado sobre la salud de los aisladores (como nivel de contaminación o si tiene fisuras), que considere los casos particulares de la línea a la que pertenecen, su ubicación y el comportamiento del ambiente.
4. Ausencia de una herramienta moderna para la interpretación de los datos tomados en terreno de inspecciones.

1.3.1. Metodología de inspección y mantenimiento

A la fecha, la forma de realizar los procesos, relacionados con el mantenimiento de los aisladores, se pueden resumir en las siguientes etapas:

1. Inspección en terreno: Un técnico de la empresa se dirige a las dependencias para tomar datos del aislador, entre los cuales están: ambientales (temperatura y humedad), conteo de descargas por fase e imágenes UV (mediante Corocam), voltaje de la línea, entre otros. Toda la información recabada se reúne en una plantilla de mantenimiento en formato Excel.
2. Análisis Excel de clasificación: Un ingeniero de mantenimiento analiza los resultados de la inspección, y en base a rúbricas, en conjunto con su experiencia, realiza un plan de mantenimiento para las líneas de transmisión y sus aisladores, cuya ejecución será encargada al personal de mantenimiento.

Por lo tanto, su sistema de planificación de mantenimiento se basa en los datos de la cartilla de inspección, sin ninguna automatización del proceso, ya sea para tratar los datos, como para analizarlos.

1.3.2. Objetivos de la memoria

General

Implementar una herramienta preventiva del estado de los aisladores, basado en técnicas de Machine Learning, para brindar apoyo efectivo en la toma de decisiones de mantenimiento.

Específicos

1. **Unificar** el formato de los datos de variables eléctricas, ambientales y geográficas relevantes, al momento de la recolección.
2. **Diseñar y desarrollar** un sistema capaz de diagnosticar el estado de los aisladores, en cuanto a necesidad de mantenimiento.
3. **Integrar** la solución al proceso de la empresa, generando el menor impacto en el quehacer de los trabajadores.
4. **Validar** el funcionamiento del sistema, mediante métricas y retroalimentación de la empresa.

1.3.3. Requerimientos

El sistema a desarrollar debe poder cumplir con los siguientes requerimientos.

Funcionales

El sistema debe:

1. Contar con información de los aspectos técnicos y variables de la infraestructura relativa al desafío, como posición, condiciones ambientales, mediciones coronográficas, corrientes de fuga, resultados de inspecciones e información sobre mantenciones previas para ser usadas como variables en el modelo.
2. Ser capaz de adquirir datos desde las cartillas de inspección.
3. Contar con una etapa de *curatoría* de los datos para corregir posibles errores y crear una base de datos con un formato único.
4. Proveer una interfaz gráfica de interacción al usuario.
5. Sugerir si se realiza o no una limpieza a algún aislador (es decir, la sección de línea), para apoyar la toma de decisión del usuario.

No funcionales

El sistema debe:

1. Poseer una interfaz accesible, sencilla e intuitiva al usuario.
2. Ejecutarse de manera rápida, demorándose mucho menos que un humano en realizar sugerencias.

3. Tener un Uptime de al menos un 95 %, dado que es crucial la disponibilidad para el usuario.
4. Ser confiable, en cuanto a seguridad de la información.
5. Estandarizar el formato de las plantillas, reduciendo las posibles subjetividades de los técnicos.

De ambiente

Para no interferir en los procedimientos del día a día de la empresa, y como una transición suave de un proceso transformación digital, se opta por integrar la solución al ambiente del equipo de mantenimiento, es decir, lo desarrollado se implementará en excel a través del *plugin* PyXLL. En consecuencia, los requerimientos del sistema, para ser ejecutado, son:

- **Hardware:** computador con características que soporte Microsoft Excel.
- **Software:** Python 3.8 o superior y Microsoft Excel con plugin PyXLL.

1.4. Estructura de la memoria

Este escrito está estructurado de la siguiente manera: El capítulo 2 se centra en explicar la teoría que se utiliza para la realización de la memoria, el trabajo desarrollado en problemas similares por otros autores y un análisis de la cartilla de inspección. En el capítulo 3, se define el problema y se revisará más en detalle la información proveída por la empresa, finalizando con el esquema del sistema. En el capítulo 4, se desarrolla el diseño de la solución y se implementa, analizando gráficas y resultados obtenidos. A partir de ello se muestra la integración del sistema al proceso de la empresa. En el capítulo 6, se muestra un conjunto de conclusiones, generales y específicas, en torno al problema y solución desarrollados, además de las proyecciones del trabajo de memoria.

Capítulo 2

Machine Learning y aisladores

A lo largo de este capítulo, se presentará el marco teórico que sustenta el desarrollo de esta memoria y la composición del problema, abarcando desde Machine Learning hasta la teoría detrás del fenómeno eléctrico del problema.

2.1. Machine Learning

Aprendizaje de Máquinas o **Machine Learning**, es la programación de computadoras para optimizar un criterio de rendimiento futuro utilizando datos de ejemplo o experiencias pasadas. Tenemos el **modelo** definido en base a algunos parámetros y el **aprendizaje**, el cual es la ejecución de un programa de computadora para optimizar los parámetros del modelo, utilizando los datos de entrenamiento o la experiencia pasada [1].

Dentro del contexto de esta memoria, es que existe un conjunto de N muestras llamado **conjunto de entrenamiento o *training set***, usado para sintonizar o afinar los parámetros del modelo adaptativo. Estas muestras corresponden a cada registro de las condiciones del aislador eléctrico en estudio, observado como fila en la cartilla de inspección utilizada por la empresa (analizada en 2.5), y cada muestra posee columnas representando variables, también llamadas como dimensión o característica (*feature*), que cuantifican o categorizan alguna propiedad del problema. En la Figura 2.1 se puede apreciar un ejemplo de registro de muestras.

Estr.	Circuito		N° Registro	Tipo Aislación	Distancia Obj-cámara [m]	Temp. Amb. [°C]	Humedad Relativa [%]	Descarga por fase			Ruido		Daño Visible	Nivel de Contaminación			Rango EPRI				OBSERVACIONES
	1	2						I-Z	M-M	S-D	SI	NO		SI	NO	1	2	3	A	B	
1			000_2020_04_28_12_38_51	Polimero/Loza	20		20%	33	33	33	x		x			x			x	No se verifican descargas por efecto corona, tampoco se detectan anomalías que sugieran una intervención mayor. Aislación con nivel de contaminación alto, se recomienda realizar limpieza a corto plazo.	
2			000_2020_04_28_13_04_13	Polimero	23		20%	50	16	83	x		x			x			x	No se verifican descargas por efecto corona, tampoco se detectan anomalías que sugieran una intervención mayor. Aislación con nivel de contaminación alto, se recomienda realizar limpieza a corto plazo.	
3			000_2020_04_28_13_14_12	Polimero	23		20%	16	16	33	x		x			x			x	No se verifican descargas por efecto corona, tampoco se detectan anomalías que sugieran una intervención mayor. Aislación con nivel de contaminación alto, se recomienda realizar limpieza a corto plazo.	
4			000_2020_04_28_13_19_30	Polimero	30		20%	0	0	16	x		x			x			x	No se verifican descargas por efecto corona, tampoco se detectan anomalías que sugieran una intervención mayor. Aislación con nivel de contaminación alto, se recomienda realizar limpieza a corto plazo.	

Figura 2.1: Ejemplo de muestras en cartilla de inspección

Como el interés de este trabajo es conocer si será necesario realizar limpieza o no de un aislador, se necesita verificar qué tipo de problema o aprendizaje se requiere utilizar. Si dentro de los datos existe alguna variable que pueda servir de etiqueta, como por ejemplo si el aislador está sucio o está limpio, es posible utilizar técnicas de clasificación. En tal caso, la variable correspondiente a la etiqueta de interés sería llamada *target* y sería un problema de **aprendizaje supervisado** o *Supervised Learning*.

El concepto de aprendizaje puede ser visto como el resultado de la ejecución de un algoritmo de machine learning utilizado para lograr una función $f(x)$ con entrada x , que puede ser un vector de múltiples dimensiones, y una salida y , la cual puede depender también de parámetros θ de dicha función. Para el caso del ejemplo de clasificación mencionado anteriormente, el modelo recibiría el conjunto de entrenamiento como entrada para la fase de aprendizaje, que debe contar explícitamente con una variable *target* a utilizar, representando cuál es la salida esperada en función de las entradas.

Luego de la etapa de entrenamiento, se verifica el desempeño del modelo a través del conjunto de pruebas o *test set*, evaluando la habilidad que tiene para categorizar correctamente nuevas muestras nunca antes vistas, es decir, su habilidad para **generalizar** el fenómeno plasmado en el modelo.

Tal y como se ha visto, la clasificación no es más que relacionar valores de entrada con un número finito de categorías discretas cuya asignación es conocida, sin embargo, para el caso que se deba relacionar estos con un espacio continuo de salida (también conocido), como un intervalo que represente el nivel de suciedad del aislador o una escala con la urgencia de limpieza, se estará frente a un problema de **regresión**. Este corresponde a la otra forma de aprendizaje supervisado cuyo *target* es una variable numérica continua.

Si no existe alguna etiqueta de utilidad o valores de referencia para establecer relaciones con la entrada, es decir, no existen valores que puedan ser usados como *target*, contando solo con los datos de entrada x , es posible que se esté frente a un problema de **aprendizaje no supervisado** o *Unsupervised Learning*. El objetivo en este tipo de problemas puede ser descubrir grupos en base a relaciones entre los datos de entrada, conocido como *clustering*, o determinar la distribución de los datos en el espacio de entrada, conocido como *density estimation*. Adicionalmente, es posible utilizar técnicas para realizar

proyecciones de datos con muchas dimensiones en dos o tres con el objetivo de visualizar estos [2].

Finalmente, existe un tercer tipo de aprendizaje llamado **Reinforcement Learning**, o aprendizaje reforzado, el cual se basa en encontrar acciones adecuadas a tomar para maximizar una recompensa, el cual dado el contexto del desafío, no será tomado en cuenta para el desarrollo.

2.1.1. Etapas de Machine Learning

Para el desarrollo de la memoria se utiliza la forma de trabajo típica de un proceso de Machine Learning, es decir:

1. Definir el problema: identificar de qué trata la problemática y definir la tarea a realizar usando Machine Learning (algoritmos de entrenamiento, modelos posibles, tipo de entrada/salida y de aprendizaje).
2. Construcción del conjunto de datos: incluyen adquirir, inspeccionar, resumen estadístico y visualización de datos.
3. Entrenamiento del modelo: Se definirá qué tipo de modelo utilizar, en base a la caracterización de los datos, ajustando configuraciones o parámetros que influyan en el comportamiento del modelo (si corresponde).
4. Evaluación del modelo entrenado: utilizando métricas que ayuden a dilucidar el desempeño del modelo entrenado, dependiendo de sus características.

Las etapas anteriores se utilizan para llegar a la etapa de inferencia, con la utilización del modelo entrenado. Este proceso no es lineal y está en constante monitoreo y evaluación, como se ilustra en la Figura 2.2.

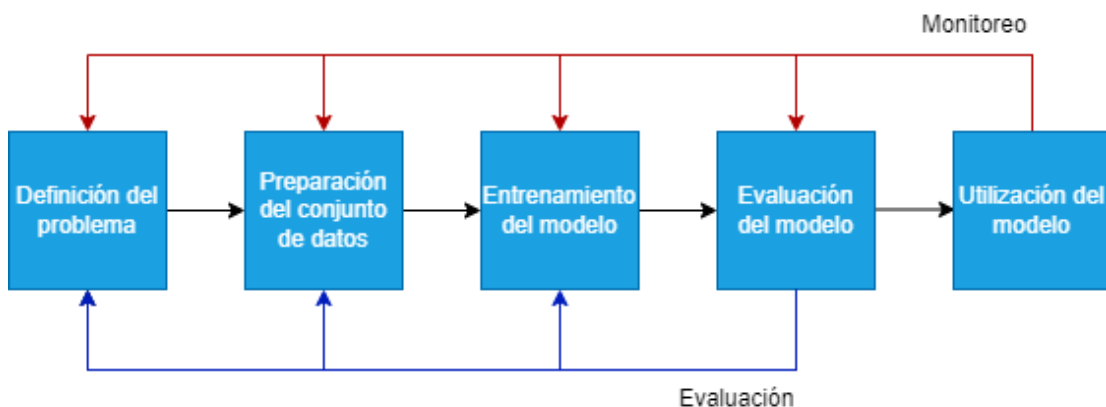


Figura 2.2: Diagrama del proceso de Machine Learning utilizado, elaboración propia

2.2. Detección de anomalías

Una anomalía es un dato con características muy diferentes al resto, el cual puede indicar un comportamiento anormal del sistema. Alternativamente, las anomalías pueden ser errores que deben ser descartados para obtener estadísticas confiables. Podemos diferenciar entre dos tipos de anomalías: de valor extremo (si los datos distribuyen como una gaussiana, se encontrarían en la cola) y las aisladas (se encuentran mezcladas con los datos normales de alguna forma, no en la cola de la gaussiana) [3].

La detección de anomalías no es categorizada como un método supervisado, donde se tiene el problema de separación de dos clases (normales y anomalías), esto debido a que generalmente no existen suficientes instancias que puedan ser etiquetadas como anomalías y no tienen un patrón que pueda ser fácilmente capturado por un clasificador. Lo común es modelar los datos normales para identificar los que se encuentran fuera de esa clasificación, por ello se le llama alternativamente como clasificación de una clase. El problema principal es que los datos muchas veces no están etiquetados y las anomalías se encuentran mezcladas con estos.

Existen métodos que primero realizan una estimación de la densidad (*density estimation*) y luego revisan la existencia de instancias con baja probabilidad de estar dentro de la estimación de densidad. Sin embargo, dada la naturaleza de las anomalías, no existen garantías que la distinción (aplicando algún modelo de detección) será perfecta, es decir, siempre es posible que exista un grado de contaminación de los datos.

En general, es posible dividir los métodos en dos grandes grupos: basados en densidad, que trabajan bajo el supuesto de que las instancias normales ocurren en regiones densas, mientras que las anomalías en regiones dispersas; y basados en distancia, donde se asume que una instancia normal es cercana a sus vecinos y la anomalía se encuentra alejada.

2.2.1. Z-Score

Es la forma más básica (desde la estadística) de encontrar anomalías de valores extremos, asumiendo que los datos distribuyen normal, se realiza la siguiente transformación:

$$Z_i = \frac{x_i - \mu}{\sigma}$$

Con Z_i siendo el valor Z-Score correspondiente a x_i , donde μ y σ son el promedio y la desviación estándar, respectivamente. Valores altos y negativos, corresponderán a valores ubicados en los extremos de la distribución normal.

2.2.2. Local Outlier Factor

Este método de detección de anomalías compara la densidad de la vecindad de una instancia con el promedio de densidad de las de sus vecinos. Sea $d_k(\mathbf{x})$ la distancia entre la instancia $\mathbf{x}$ y su vecino más cercano k , y sea $N(\mathbf{x})$ el conjunto de entrenamiento de instancias que están en la vecindad de $\mathbf{x}$, por ejemplo, sus k vecinos cercanos. Con $d_k(\mathbf{s})$ para $\mathbf{s} \in N(\mathbf{x})$, comparamos $d_k(\mathbf{x})$ con el promedio de $d_k(\mathbf{s})$ para un $\mathbf{s}$:

$$\text{LOF}(\mathbf{x}) = \frac{d_k(\mathbf{x})}{\sum_{\mathbf{s} \in N(\mathbf{x})} d_k(\mathbf{s}) / |N(\mathbf{x})|}$$

Si $\text{LOF}(\mathbf{x})$ es cercano a 1, $\mathbf{x}$ no es una anomalía. A medida que este se hace más grande, la probabilidad de que sea una anomalía se incrementa.

2.2.3. Isolation Forest

Este método se basa en árboles binarios, midiendo individualmente la susceptibilidad de las instancias a ser aisladas, donde las anomalías serán aquellos datos con una susceptibilidad alta.

Su funcionamiento se basa en generar varios árboles binarios aleatorios donde las instancias, ubicadas en un espacio, son recursivamente particionadas (de forma aleatoria y diferente para cada árbol), donde las anomalías se encontrarán más cerca de la raíz y los datos normales más lejos. Un número pequeño de anomalías producirá un número pequeño de particiones. Si en un bosque (*forest*) de árboles aleatorios se repite constantemente que cierto dato se encuentra más cerca de la raíz, es muy probable que sea una anomalía. Para lograr ello, se obtiene el promedio del camino (*path*) desde la raíz hasta el dato en particular, y luego se realiza un *ranking* de acuerdo al tamaño del promedio del *path* de todos los puntos, donde los anómalos serán los que tengan menor valor [4].

2.3. Aisladores eléctricos

En el contexto de líneas de alta tensión o redes de transporte de energía eléctrica, los aisladores cumplen la función de evitar la fuga de energía en el paso de ésta hacia tierra. Con el tiempo estos acumulan una capa de contaminación que dependiendo de las condiciones del medio ambiente, puede deteriorar su capacidad de aislación. El proceso de acumulación de contaminación puede desencadenar la producción de descargas en los aisladores.

2.3.1. Contaminación y descargas

Las fuentes de contaminación que pueden afectar al sistema de aislación incluyen: productos industriales que contienen sales solubles, contaminación proveniente de las carreteras, excremento de ave,

arena de desiertos o dunas, o sal marina, es decir, sal llevada desde el agua de mar por vientos hasta $15 - 30[km]$ al interior o más lejos.

Los aisladores en servicio con el tiempo acumulan capas ligeras de contaminación, que empiezan poco a poco a construir debido a condiciones del medioambiente (días secos y húmedos), una capa mas robusta que es capaz de comenzar a conducir corrientes de no más de varios miliamperios, resultando en el calentamiento de la superficie del aislador. Con el tiempo esta se deteriora cada vez más permitiendo la producción de arcos eléctricos cada vez más grandes, culminando en contorneos [5].

2.3.2. Efecto corona

El efecto corona o descarga corona es un fenómeno no lineal presente en la fase inicial de descargas eléctricas, en el cual existe un flujo de energía eléctrica desde un conductor hacia un medio ionizado. Se trata de descargas localizadas independientes y próximas a las superficies de electrodos bajo los efectos de campos eléctricos de alta intensidad, provocando estrés en la capacidad aislante del medio circundante, que se manifiesta principalmente como una luminosidad violeta, ruido e interferencias de radiofrecuencia altas [6]. En el caso de los aisladores eléctricos, se puede producir dentro del aislador, en el conductor o en la interfáz del aislador.

2.4. Trabajo relacionado

2.4.1. Basado en ultrasonido

En [7] se desarrolla un clasificador embebido para identificar la degradación en aisladores de polímero, utilizando como entrada señales ultrasonido emitidas por estos en una línea de tensión. Existe una etapa de preprocesamiento en donde se utiliza la transformada discreta de Wavelet, para descomponer la señal y extraer características.

Luego del preprocesamiento y extracción de características, se tiene otra etapa que corresponde a la de los algoritmos de clasificación, que en este caso se utiliza Naive Bayes y KNN, los cuales resultan en un accuracy promedio de al menos 91 % y 93 % para cada uno respectivamente.

El enfoque interesante de este trabajo radica en base a la forma de obtener las características y el dataset utilizado: 1464 patrones divididos en 5 clases, las cuales son: condición normal (sin efecto corona), y 4 estados de degradación; oxidado, con heces de pájaro, contaminación por sales, y destellando (descargas moderadas), además cada patrón está compuesto de 5000 puntos.

2.4.2. Basado en imágenes del espectro visible

Una arquitectura basada en redes neuronales convolucionales (CNN) en cascada es lo propuesto por los autores en [8], los cuales cuentan con casi 2000 imágenes de zonas de aisladores en alta resolución como entrada, divididas en dos clases: defectuoso y no defectuoso, donde con ello se busca poder predecir cuando un aislador se encuentra en estado defectuoso.

El sistema está compuesto por dos CNN, una para encontrar dentro de la imagen al aislador (ILN) y otra para detectar el estado de este (DDN), puestas en cascada en el mismo orden. Las CNN van con las respectivas capas de convolución, *pooling* y *fully connected*, aplicando filtros sobre regiones locales de la entrada, obteniendo mapas de características de la imagen. Para probar el modelo en cascada se utilizan tres modelos preentrenados: VGG-16 [9], ResNet-50 y ResNet-101 [10].

Los resultados obtenidos del rendimiento de ambas redes ILN y DDN son prometedores, al tener en el mejor de los casos para ILN utilizando VGG-16 una precisión de 0.904, recall de 0.966 y F1 score de 0.934, mientras que para DDN utilizando ResNet-101 una precisión de 0.91, recall de 0.958 y F1 score de 0.933.

2.4.3. Basado en imágenes del espectro UV

Los autores en [11] proponen un sistema más completo que recibe como entrada video del sensor ultravioleta, el cual será analizado *frame a frame* pasando principalmente por 4 etapas: adquisición, preprocesamiento, clasificación e interpretación, con el objetivo de detectar si existen o no efectos de descargas parciales.

En la primera etapa de adquisición se recibe el flujo de video y este se descompone cuadro por cuadro, realizando una transformación a escala de grises.

En la siguiente etapa de preprocesamiento, a la imagen en escala de grises se le aplica un filtro mediano para reducir el ruido y preservar las esquinas lo mejor posible, además se utiliza un kernel de tamaño 3 dentro de los parámetros del filtro mediano. Luego se procede a segmentar la imagen y aplicar un umbral, obteniendo finalmente una imagen binaria. A continuación se procede a extraer características, implementando una medida de qué tan bien la descarga corona está representada en cada imagen, lo cual se logra a través del conteo de los píxeles no cero.

Continuando con la etapa de clasificación, en este punto se prueba utilizando aprendizaje supervisado mediante 4 variaciones de SVM: Linear SVM y 3 SVM con diferente *kernel* (RBF, Polynomial y Linear). Adicionalmente se utiliza Nearest Neighbors con 2 variantes: KNN y RadiusNN.

Finalmente la etapa de interpretación es en la cual existe una interfaz humano-máquina para mostrar las mediciones y resultados, donde el usuario decidirá si lo concluido por el proceso es significativo o no.

El *dataset* utilizado para las pruebas consistió en 2 videos de resolución 240x320 y 280x352, con 444 y 330 muestras de cuadros respectivamente. Los resultados muestran una precisión superior a 0.94 para dos técnicas: SVM variante RBF y NN variante RadiusNN, el resto no tuvo buen rendimiento.

2.4.4. Basado en mediciones en laboratorio

Bajo condiciones de laboratorio y mediciones del campo eléctrico, los autores en [12] construyen un *dataset* de 42 características para 36 muestras de las pruebas realizadas, con el objetivo de predecir el voltaje que dará comienzo al efecto corona (*onset*) y el que provocará el primer arco eléctrico (*breakdown*).

El sistema se constituye de etapas de adquisición de datos y preprocesamiento de variables directamente eléctricas, entre ellas la fuerza, energía y gradiente del campo eléctrico, utilizando para su modelado aprendizaje supervisado mediante SVM, complementando con *grid search* para optimización de parámetros. Además se construyen en base a las 42 características (dimensiones) diferentes *sub-datasets* mediante reducción de dimensionalidad basada en correlaciones.

El entrenamiento del algoritmo fue realizado con 8 muestras experimentales, para luego ser probado con las 36 de pruebas que ya se tenían, obteniendo para el *onset* un error relativo máximo de 6 % y 19 % para 42 y 27 dimensiones, respectivamente. También se obtiene bajo este método para el *onset* un error porcentual absoluto promedio de 1.78 % y 2.56 % para 42 y 27 dimensiones. Los resultados para el *breakdown* muestran para 40, 37 y 31 dimensiones un error porcentual absoluto promedio de 2.08 %, 1.77 % y 2.69 %, respectivamente.

Conclusiones

La metodología más usada en la literatura es el aprendizaje supervisado, siendo SVM la técnica más popular, siempre bajo el supuesto que el fenómeno del efecto corona (sea o no en aisladores) puede ser predecido en base al comportamiento del pasado.

Además, todo indica que la forma general de abordar esta clase de problemas siempre trae consigo las mismas etapas: adquisición, preprocesamiento, extracción de características, modelado a través de alguna técnica de Machine Learning e interpretación de los resultados obtenidos. Esto motivará a que el trabajo siga una estructura similar.

Finalmente esta investigación servirá para demarcar las bases de las condiciones, el enfoque y la meta alcanzable, en función de la cantidad y calidad de datos con los que se cuenten y el conjunto de herramientas de análisis, dependiente del tipo de variable a utilizar.

2.5. Cartilla de inspección

La cartilla de inspección es el documento oficial (archivo Excel) que utiliza la empresa para registrar los datos en las salidas a terreno del equipo técnico. El objetivo principal es aportar de manera conclusiva en la toma de decisiones, respecto al requerimiento de mantención en los aisladores. El documento es llenado por el técnico a cargo, quien identifica el valor (cualitativo/cuantitativo) de las variables en terreno, para luego ser observado y analizado por el ingeniero de mantenimiento.

La información de la cartilla puede ser dividida en dos grandes grupos, como se puede ver en la Figura 3.1: por un lado, características generales de la línea de transmisión, y por otro, las mediciones específicas a los aisladores eléctricos que son parte de esta.

La información relativa a la línea de transmisión, como un todo, se agruparán bajo la etiqueta de *metadatos*: Instalación, OM, Año en servicio, Clima, Voltaje, Tipo de contaminación, Demanda y Viento, entre otros. Las mediciones respecto de cada aislador eléctrico estarán registradas luego del apartado de *metadatos*, donde dependiendo del año de la cartilla, tendrá mas o menos columnas, las cuales llamaremos variables o características, ya que son el insumo principal de este trabajo.

En consecuencia, cada aislador eléctrico tendrá número de estructura, circuito asociado, temperatura ambiente, humedad relativa y descarga por fase, entre otros. La Figura 2.3 muestra un análisis detallado, donde se tienen las columnas:

- **Variables:** Corresponden a los datos ingresados a la cartilla de inspección.
- **Tipo de variable:** Clasificación si la variable corresponde a un dato cuantitativo o cualitativo.
- **Dependencia:** Corresponde a si la variable depende de otro valor medible o del criterio del inspector en terreno.
- **Medida:** Unidad de medida de la variable.
- **Definición:** Alude a una descripción de la variable medida.
- **Significancia:** Qué representa la variable y por qué se mide.
- **Valor de importancia:** Estimado de cómo afecta este tipo de dato en la decisión de realizar mantenimiento o acelerar las limpiezas periódicas de la línea.

Además, existe la columna registro, la cual enlaza el aislador con su video correspondiente, tomado con la CoroCam 6D, cámara orientada a la captura de proyección de imágenes UV y utilizada para la detección de señales de efecto corona en tiempo real.

Variable	Tipo	Dependencia	Medida	Definición	Significancia	Valor de importancia
Clima	Cualitativa	Independiente	Tipo de clima	Condición meteorológica registrada el día de medición. (Nublado, soleado/despejado, neblina)	Resume aproximadamente las variables Meteorológicas recurrentes bajo un concepto.	Bajo
Contaminación (Según ambiente)	Cualitativa	Independiente	Tipo de contaminación	Conjunto de circunstancias o factores físicos, químicos y biológicos que rodean al objeto en particular e influyen en su desarrollo y comportamiento.	Indica la disposición de un aislador a contaminarse.	Alto
Material aislador	Cualitativa	Independiente	Tipo de material	El material del cual está compuesto el aislador.	Afecta en la hidrofiliadad de la superficie. Afecta en su vida útil general.	Medio
Ruido	Cualitativa	Independiente	Tipo de ruido	Sonido característico de las descargas generadas en el aislador.	Indicador de la existencia de descargas parciales en la superficie del aislador.	Alto
Daño visible	Cualitativa	Independiente	Tipo de daño	Daños superficiales (Grietas o picados) apreciados visualmente en la superficie del aislador.	Dependiendo de la intensidad, acusa mantenimiento necesario a corto plazo.	Alto
Descargas	Cuantitativa	Dependiente	Cantidad de descargas	La magnitud relacionada con los impulsos de DP permitidos en un objeto de ensayo en una tensión específica, siguiendo un procedimiento de ensayo y un condicionamiento específicos.	La magnitud de la cuenta de DP por lo general está relacionada con el nivel de contaminación o daño en la estructura	Alto
Temperatura	Cuantitativa	Independiente	Grados Celsius	Grado o nivel térmico de un cuerpo o de la atmósfera	Incide en la densidad del aire que rodea el aislador	Bajo
Humedad	Cuantitativa	Independiente	% de Humedad	Relación entre la cantidad de vapor de agua que tiene una masa de aire y la máxima que podría tener.	Tiene relación con la cantidad de corrientes de fuga	Bajo
Nivel contaminación	Cuantitativa	Dependiente	% de Contaminación	Clasificación de la contaminación basada en inspección visual	Tiene relación con el requerimiento de limpieza	Alto
EPRI	Cuantitativa	Dependiente	Nivel de EPRI	Clasificación basada en manual de Electric Power Research Institute	Dependiendo el rango de clasificación se decide qué decisión tomar con respecto a la urgencia de mantenimiento	Alto
Tiempo	Cuantitativa	Independiente	Días diferentes de medición	Referencia de días de inspección	Variabilidad de los datos recolectados	Bajo

Figura 2.3: Variables asociadas a cada aislador eléctrico

Capítulo 3

Definición del problema

A continuación se verá un análisis más detallado, desde un punto de vista técnico y metodológico, para llegar al diseño e implementación de la solución. Además, se definirán desde los requerimientos a cumplir, hasta la metodología de trabajo.

3.1. Identificación de requerimientos

La empresa basa su proceso de inspección y análisis de los aisladores eléctricos en archivos excel, llamados cartilla de inspección, siendo esta la forma de almacenar sus datos. Por cada rutina de inspección de las líneas de transmisión, se genera una cartilla nueva con el formato presentado en la Figura 3.1.

CHILQUINTA energía															Plan de Mantenimiento Reporte de Inspección Descarga Corona				Clasificación Periodo de Limpieza/Lavado de Aislación						
Instalación		LT 44KV LAS VEGAS LOS ANDES				OM		10069891							A	No es necesario									
Año en servicio		1925				Clima		Despejado/Soleado							B	Podría ser necesario									
Voltaje [kV]		44				Tipo de contaminación		Agrícola							C	Realizar a corto plazo									
Demanda [MVA]		9,2 C2-6,7// 23-04				Viento [km/h]		---							D	Realizar urgente									
Inspector		JRF				Permiso		PQ3784, 3834							X	bajo nivel de confianza, resultado engañosos									
Equipo		CoroCAM D6				Fecha		22 y 23 Abril 2020																	
Nivel de contaminación: (1) menor 40%; (2) entre 40% y 70% y (3) mayor a 70%															Informe				JRF						
Estr.		Circuito		Hº Registro		Tipo Aislación		Distancia Obj-cámara [m]		Temp. Amb. [°C]		Humedad Relativa [%]		Descarga por fase		Ruido		Daño Visible		Nivel de Contaminación		Rango EPRI		OBSERVACIONES	
1		2												I-Z		M-M		S-D		SI		NO			
																				1		2		3	
																				A		B		C	
																						D			

Figura 3.1: Ejemplo de estructura de cartilla de inspección

Al revisar en detalle los archivos entregados por la empresa, se han encontrado discrepancias o incoherencias en las cartillas, entre los cuales están:

- No existe un apartado conclusivo con relación al estado de los aisladores, por lo que no satisface el objetivo principal de la inspección.

- Hay variables cuya definición, cálculo e influencia en la problemática no son claras.
- Los campos de porcentaje de contaminación y Rango EPRI, son establecidas bajo un criterio arbitrario y poco normado.
- El Rango EPRI tiene relación con el conteo de las descargas captadas por la cámara, sin embargo el número de descargas ingresados en la cartilla no concuerda con el rango asignado.
- La cantidad de variables significativas del tipo cualitativo conforman el 50 % de los datos, por lo que para formular una relación matemática entre las variables y su influencia en los aisladores requiere de un análisis exploratorio previo.
- Los datos medidos bajo el uso de equipo específico (temperatura, humedad y cantidad de descargas) corresponden solo a un 30 %, por lo que la exactitud del resto de datos obtenidos está afectada por la percepción y juicio subjetivo de quien las registra. Además, la medición no está exenta de errores de calibración, uso e interpretación de las mediciones.
- Una tasa no menor de repetición de valores iguales de una misma variable a lo largo de la línea, lo cual puede ser un reflejo de una toma de datos inexacta y replicada a lo largo del documento. Esto puede influir en la precisión de las conclusiones, al igual que la correlación de las variables.
- Los ambientes se clasifican en tres categorías: Agrícola, Industrial/minero y Smog, de los cuales no se detallan sus características y quedan a criterio del inspector en terreno. Mientras que en la normativa 60815 se recomienda una clasificación en cinco categorías: Desierto, Costa, Industrial, Agrícola e interiores, las cuales poseen una descripción clara de la categoría y una aplicación de acuerdo al tipo de monitoreo con el que se cuente.
- El ruido es señal de actividad eléctrica en la superficie de los aisladores, pero también puede estar presente en otros equipos eléctricos, por lo que el dato ingresado puede estar alterado, es decir, se puede dar el caso en que en la cartilla hay circuitos que posean un bajo nivel de descargas con ruido, y viceversa.

Además, al tiempo del desarrollo de esta etapa de la memoria, solo se cuenta con 1 cartilla de inspección por línea de transmisión, correspondiente a un periodo de inspección, siendo en total 1076 registros repartidos en 28 cartillas, es decir, no existen posibilidades de trabajar con series de tiempo¹.

¹Referido a datos con una marca temporal continua, que permite ver la evolución de cierta propiedad en el tiempo.

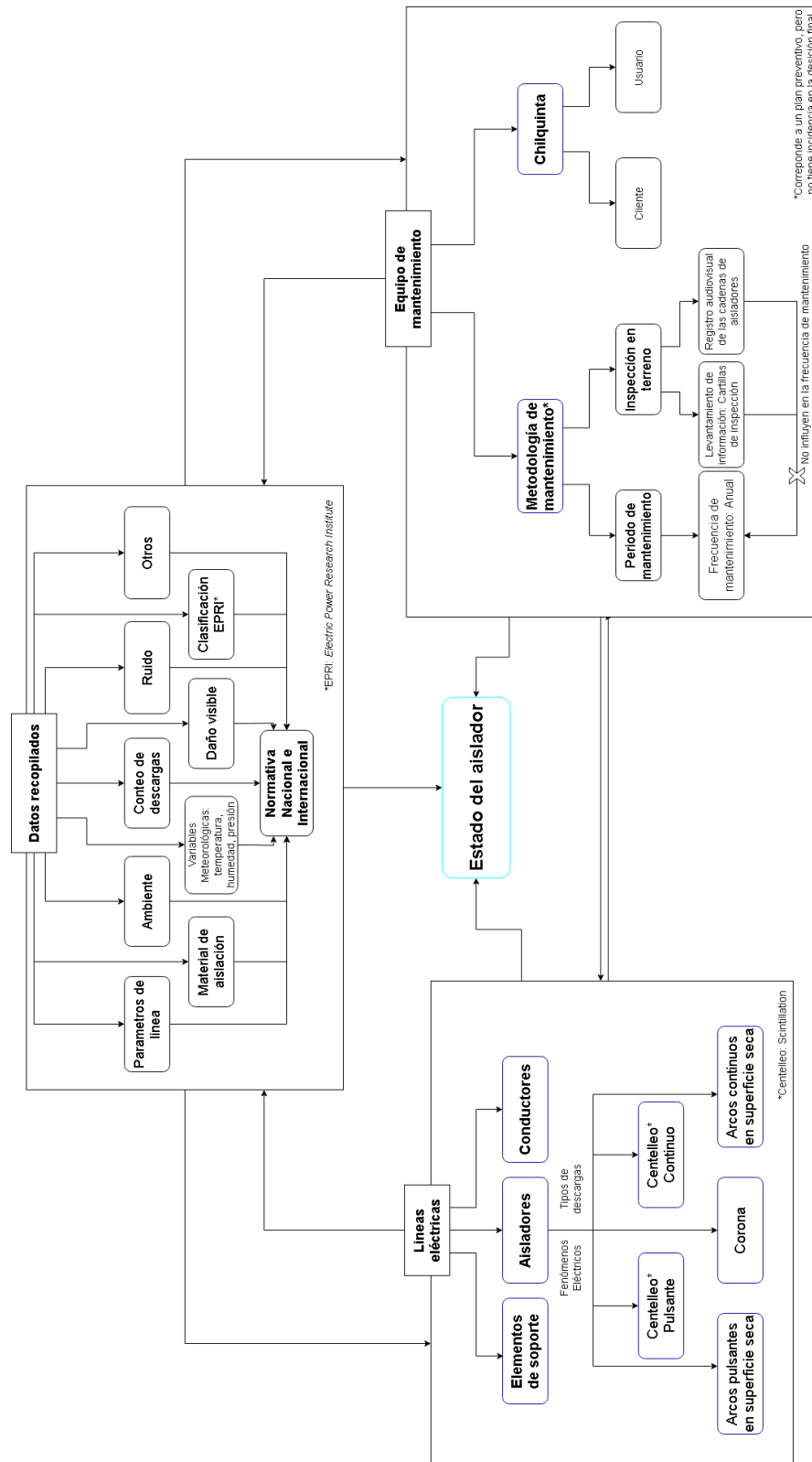


Figura 3.2: Diagrama de contexto del problema, elaboración propia

3.2. Arquitectura del sistema

A continuación se presenta, a nivel de diagramas, las diferentes áreas que abarca el sistema, y cómo este se desenvuelve.

3.2.1. Esquema contextual del sistema

En la Figura 3.2 se muestra el esquema contextual en el cual la solución a desarrollar estará inmersa, donde se pueden diferenciar tres grandes bloques influyendo en el fenómeno eléctrico de interés:

1. **Datos recopilados:** Corresponden al máximo de variables que pueden ser extraídas desde la cartilla de inspección, rellena por el técnico en terreno. A su vez este monitoreo (manual por ahora), debe cumplir con las normativas nacionales e internacionales.
2. **Equipo de mantenimiento:** Por parte de la empresa, que actúa como cliente y usuario, existe una metodología de mantenimiento establecida, la cual no influye (al momento de realizar este levantamiento) en la frecuencia de mantenimiento.
3. **Líneas eléctricas:** Son la parte física que interactúa con los aisladores eléctricos, que son el principal objetivo de este trabajo, en específico, el fenómeno corona y las descargas de fase.

3.2.2. Esquema general del sistema

El sistema a desarrollar y su interacción con el usuario se presenta en la Figura 3.3

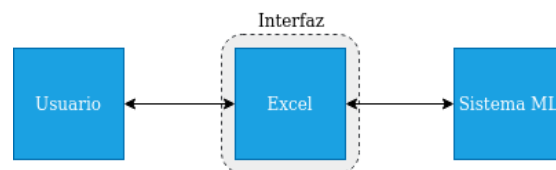


Figura 3.3: Diagrama general del sistema, elaboración propia

donde:

- Usuario: representa a la persona que interactuará (solo lectura) con el sistema, siendo esta un técnico o ingeniero del área de mantenimiento.
- Excel: se trata de la interfaz de donde el sistema desarrollado obtendrá los datos, además tendrá una pequeña etapa de procesamiento estadístico automático, para mostrar información al usuario.
- Sistema ML: la solución desarrollada, siguiendo los lineamientos de la sección 2.1.1, es decir, el modelo de inferencia.

3.2.3. Esquema para el desarrollo del sistema

Sistema alternativo de apoyo al desarrollo de la solución, pero que no necesariamente irán en la solución final.

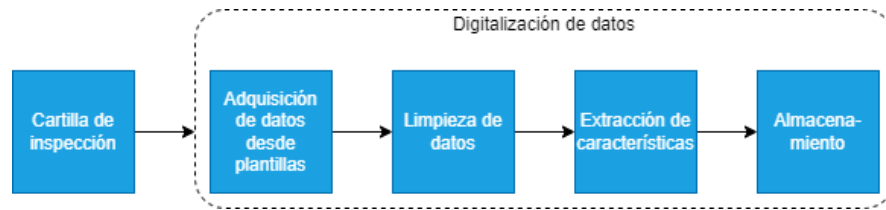


Figura 3.4: Diagrama de digitalización de datos, elaboración propia

En general, podemos llamar a las etapas de apoyo como Transformación de datos, ya que se encargan de tomar los datos de las cartillas (o plantillas) de inspección, corregir formatos, imputar o eliminar datos según corresponda, extraer características y almacenar estos datos a ser consumidos en la etapa de aprendizaje.

Capítulo 4

Diseño de la solución

A continuación se describirá el proceso seguido para llegar al prototipo de solución, siguiendo la metodología genérica de Machine Learning del capítulo anterior, abarcando las etapas de preparación del conjunto de datos, entrenamiento del modelo y evaluación del modelo.

4.1. Preparación del conjunto de datos

Para preparar el conjunto de datos es necesario construir un paquete que ayudará al traspaso de estos, desde un excel a un formato legible por Python y adecuado para entrenar algún modelo. Esta etapa considera la construcción del sistema de apoyo, presentado en la sección 3.2.3, del desarrollo. Además, contará con etapas (que son parte de un análisis exploratorio de datos) de resumen estadístico y visualización de datos. Al final de este proceso, tendremos las distintas planillas listas para ser utilizadas por la etapa de entrenamiento.

4.1.1. Adquisición de datos

Mediante el uso de la biblioteca Pandas¹ se realiza la transformación de datos desde los archivos Excel. Cabe destacar que, por la variación de formato y nombres de las columnas, fue necesario hacer uso de expresiones regulares simples, para que así se pudiera realizar de una forma más general.

4.1.2. Limpieza de datos

En esta etapa se realizan correcciones de formato y nombres de propiedades de la cartilla, entre las cuales podemos destacar:

¹ Biblioteca de Python utilizada para manejo de datos, estructurados como tabla.

- Registro: se puede encontrar como registro o n° de registro, su utilidad radica en que posee una marca temporal que entrega la información de cuándo se realizó la medición, en caso que no se cuente con este dato, el valor por defecto es la fecha puesta en los metadatos de la cartilla.
- Tipo de contaminación: se puede encontrar como contaminación o polvo.
- Temperatura: la columna se puede encontrar con los nombres de t° o temperatura ambiente (o alguna abreviación de esta), su valor es puesto en grados o solo el número.
- Humedad: la columna se puede encontrar con los nombres de humedad o humedad relativa (o alguna abreviación de esta), donde su valor es puesto en decimal o porcentaje.
- Aislante: se puede encontrar como tipo de aislación o aislación, se corrigen nombres como porcelana y cerámica como loza.

El resto de correcciones corresponden a faltas ortográficas o rellenado con la moda, esto último se realiza con el objetivo de no perder información, además que se observa en todas las cartillas, respecto a temperatura o humedad, que casi todos los registros poseen valores similares.

4.1.3. Extracción de características y almacenamiento

Se procede a realizar la eliminación de columnas de variables binarias complementarias, como circuito, que tenía solo dos posibilidades que no podían ocurrir al mismo tiempo, lo mismo sucede con Ruido (si o no) y Daño visible (si o no). Finalmente solo tendremos las características de pertenece al circuito 1 (verdadero o falso), existencia de ruido (verdadero o falso) y presencia de daño visible (verdadero o falso).

Por otro lado, se realiza una normalización de la variable descarga de fase, debido a que no todas las mediciones con la CoroCam son realizadas a la misma distancia, lo que influye en el valor del conteo de descarga. Este valor sufre un decaimiento proporcional al cuadrado de la distancia, por lo tanto se realiza la siguiente corrección

$$\text{Conteo}_{norm} = \text{Conteo}_{cartilla} \cdot \frac{d_{cam-obj}^2}{20^2},$$

es decir, se utilizará una distancia de 20 metros como referencia para todos los conteos de descarga.

El almacenamiento se deja, por el momento, como simplemente juntar todos los datos de las cartillas y guardarlo en formato Pickle².

²Biblioteca de Python usada para serializar objetos, permitiendo ser guardados y cargados en una futura utilización.

4.2. Resumen y visualización

A continuación se realiza un análisis exploratorio de datos, con el objetivo de obtener un resumen estadístico y una visualización de los datos, para poder dilucidar el modelo a utilizar. En primera instancia se analiza un conjunto de cartillas individualmente, y posteriormente se consideran la totalidad de datos, como un solo conjunto.

4.2.1. Análisis local

Se procede a preprocesar las variables y a obtener las estadísticas básicas de las cartillas, como se puede ejemplificar en las Figuras 4.1 y 4.2

	Circuito1	TempAmbiente	HumedadRelativa	faseIZ_norm	faseMM_norm	faseSD_norm	tipo_Polimero	tipo_Vidrio
0	1	18.0	0.79	150.0	300.0	300.0	0	1
1	0	18.0	0.79	1400.0	1466.0	1350.0	0	1
2	1	18.0	0.79	16.0	66.0	200.0	0	1
3	0	18.0	0.79	16.0	16.0	16.0	0	1
4	1	18.0	0.79	650.0	3300.0	3750.0	0	1
...	...	...	...	...	...	...	...	...
155	0	16.0	0.85	216.0	83.0	83.0	1	0
156	1	16.0	0.85	83.0	66.0	83.0	0	1
157	0	16.0	0.85	150.0	33.0	33.0	1	0
158	1	16.0	0.85	2350.0	1850.0	685.0	0	1
159	0	16.0	0.85	100.0	16.0	16.0	1	0

Figura 4.1: Ejemplo de cartilla archivo 10068897_Informe Corona LLTT SL-AS.xlsx

	Circuito1	TempAmbiente	HumedadRelativa	faseIZ_norm	faseMM_norm	faseSD_norm	tipo_Polimero	tipo_Vidrio
count	160.00000	160.000000	160.000000	160.000000	160.000000	160.000000	160.000000	160.000000
mean	0.50000	16.875000	0.769125	2068.431250	2213.718750	1927.731250	0.312500	0.675000
std	0.50157	2.147838	0.127700	5089.386051	4817.761048	3951.840217	0.464968	0.469845
min	0.00000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.00000	16.000000	0.790000	50.000000	50.000000	50.000000	0.000000	0.000000
50%	0.50000	18.000000	0.790000	150.000000	158.000000	150.000000	0.000000	1.000000
75%	1.00000	18.000000	0.850000	1162.500000	1712.000000	2004.000000	1.000000	1.000000
max	1.00000	18.000000	0.850000	30416.000000	37800.000000	29150.000000	1.000000	1.000000

Figura 4.2: Resumen estadístico de pandas de 10068897_Informe Corona LLTT SL-AS.xlsx

donde se observa la inexistencia de las columnas: número de registro (ID de video CoroCam), el número de la estructura (ID de estructura), distancia objeto-camara (utilizada previamente para normalizar), descargas de fase (originales), ruido, daño visible y nivel de contaminación, estas últimas tres

debido a que son tomadas subjetivamente y su aporte radica en aumentar las dimensiones al problema.

Tampoco se consideran las variables EPRI.A, EPRI.B, EPRI.C y EPRI.D, ya que son justamente las que se desean evitar, al ser parte de la metodología antigua que quieren cambiar, además que la mayoría en las cartillas son del mismo tipo.

En total la cartilla de ejemplo posee 160 datos, con un promedio de 17° y 77 % aproximadamente de temperatura y humedad relativa (respectivamente), descargas parciales con mucha dispersión y solo dos tipos de aisladores: vidrio y polímero.

En la siguiente Figura 4.3 se puede apreciar la matriz de correlación para la cartilla de ejemplo



Figura 4.3: Matriz de correlación de 10068897_Informe Corona LLTT SL-AS.xlsx

de donde se puede apreciar que las descargas de fase tienen una correlación fuerte entre si, en específico para las que se encuentran al lado de la otra, es decir, la descarga de fase inferior con la media tiene una mayor correlación que con la superior. Esto nos puede estar dando luces que la degradación del aislador es proporcional y causados por el mismo fenómeno.

Por otro lado, la temperatura ambiente y la humedad relativa no tienen mayor correlación, aunque

teóricamente a cierto nivel de humedad relativa debería darse más fuerte el efecto corona, pero esta relación puede no ser lineal.

La correlación realizada anteriormente tiene un supuesto, el cual es que los datos tienen una distribución normal, por lo que a continuación se muestran gráficos de barra de estas variables

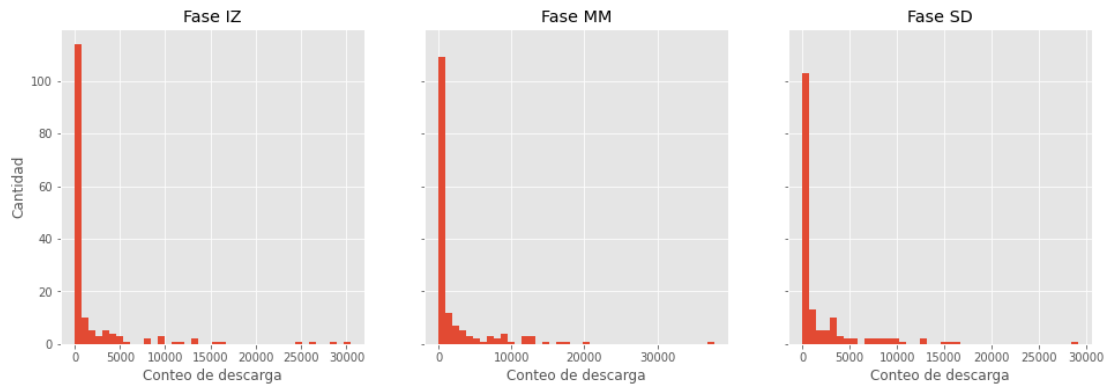


Figura 4.4: Distribución de conteo de descargas de 10068897_Informe Corona LLTT SL-AS.xlsx

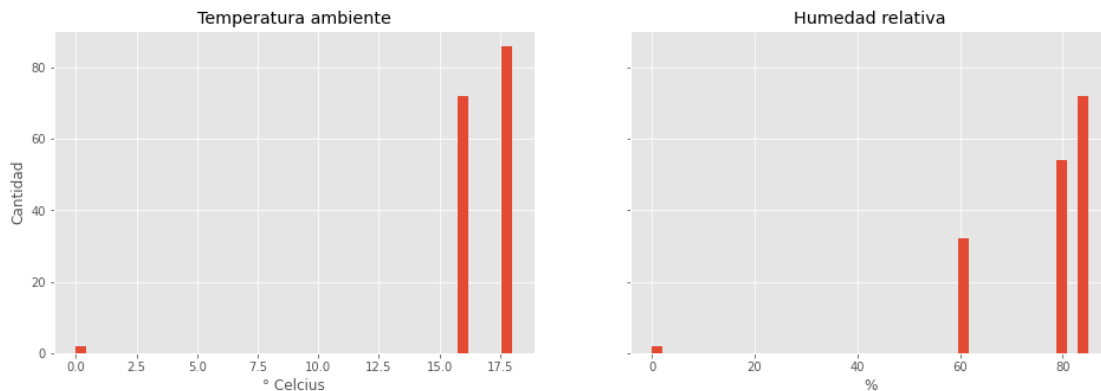


Figura 4.5: Distribución de variables ambientales de 10068897_Informe Corona LLTT SL-AS.xlsx

Es claro que los datos de una sola cartilla no son suficientes para este tipo de análisis. A modo de ejemplo, se deja en el anexo A.1 más datos sobre otras cartillas (y la cantidad de registros que poseen), con lo cual se hace más visible la necesidad de trabajar con la totalidad de datos.

4.2.2. Análisis global

Luego de realizar el mismo tratamiento a las variables que en la sección anterior, se procede a mostrar los resúmenes estadísticos, en base a 1076 registros obtenidos de la totalidad de las cartillas de inspección, en las Figuras 4.6 y 4.7 (ahora separando entre cualitativas y cuantitativas)

	temperatura	humedad	descFaseIZ	descFaseMM	descFaseSD
count	1076.000000	1076.000000	1076.000000	1076.000000	1076.000000
mean	17.681970	0.532731	450.976766	448.768587	504.417286
std	3.294685	0.195342	2167.007101	2039.980564	2075.491127
min	9.000000	0.200000	0.000000	0.000000	0.000000
25%	15.000000	0.330000	0.000000	16.000000	16.000000
50%	18.000000	0.550000	33.000000	33.000000	33.000000
75%	20.300000	0.625000	116.000000	116.000000	116.000000
max	24.000000	0.910000	30416.000000	37800.000000	29150.000000

Figura 4.6: Resumen estadístico global variables cuantitativas

	circuito1	tipo_polimero	tipo_vidrio	tipo_loza	tipo_loza/polimero
count	1076.0	1076	1076	1076	1076
unique	2.0	2	2	2	2
top	0.0	0	0	0	0
freq	609.0	811	878	717	822

Figura 4.7: Resumen estadístico global variables cualitativas

donde se puede apreciar ahora que efectivamente las descargas de fase tienen una dispersión mayor, pero esto puede estar siendo provocado por valores extremos, ya que al mirar al P75 se tienen descargas de fase del orden de los 100 conteos. Al mismo tiempo es curioso como, en comparación al resumen particular presentado en la Figura 4.2, los promedios de las tres descFase comienzan a converger en valores cercanos o similares, y que el P75 sea el mismo, lo cual puede ser debido a la ley de los grandes números [13].

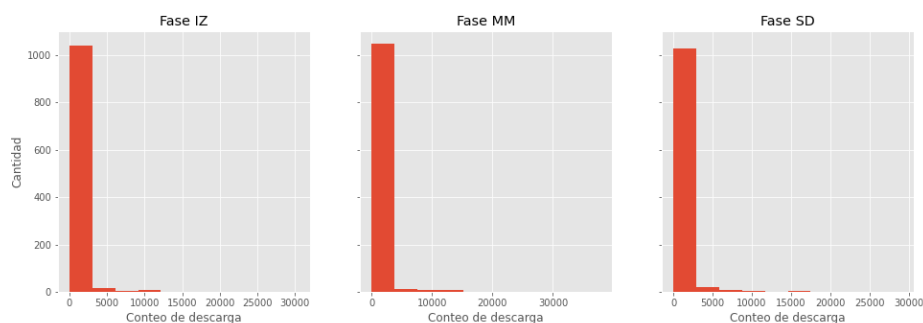


Figura 4.8: Distribución global de conteo de descargas

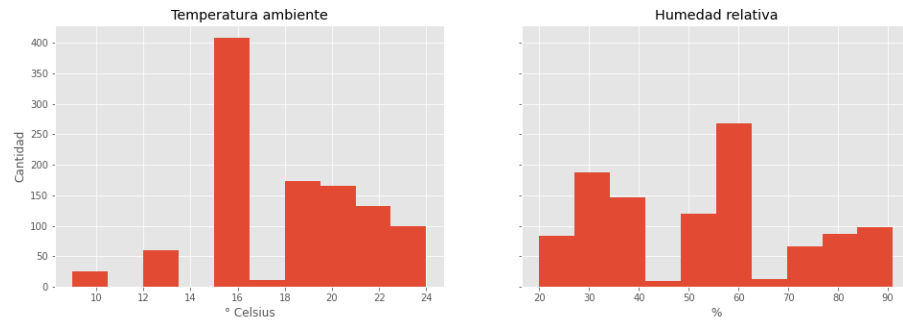


Figura 4.9: Distribución global de temperatura y humedad relativa

En las Figuras 4.8 y 4.9 se puede ver la distribución de los conteos de descargas y humedad, donde en los primeros es posible observar (más claramente que antes) que existen anomalías alterando los gráficos, las cuales habrá que trabajar en un paso posterior. En cambio, en las gráficas de distribución de temperatura y humedad, no se aprecian anomalías a simple vista, ya que es posible visualizar de mejor manera una suerte de distribución normal, que si tuvieramos más datos se podría afirmar con fuerza que existe. Para efectos de este trabajo, y como decisión de diseño ante la falta de un mayor volumen de datos para probarlo (con mayor fuerza), se asume que las variables cuantitativas distribuyen de forma normal.

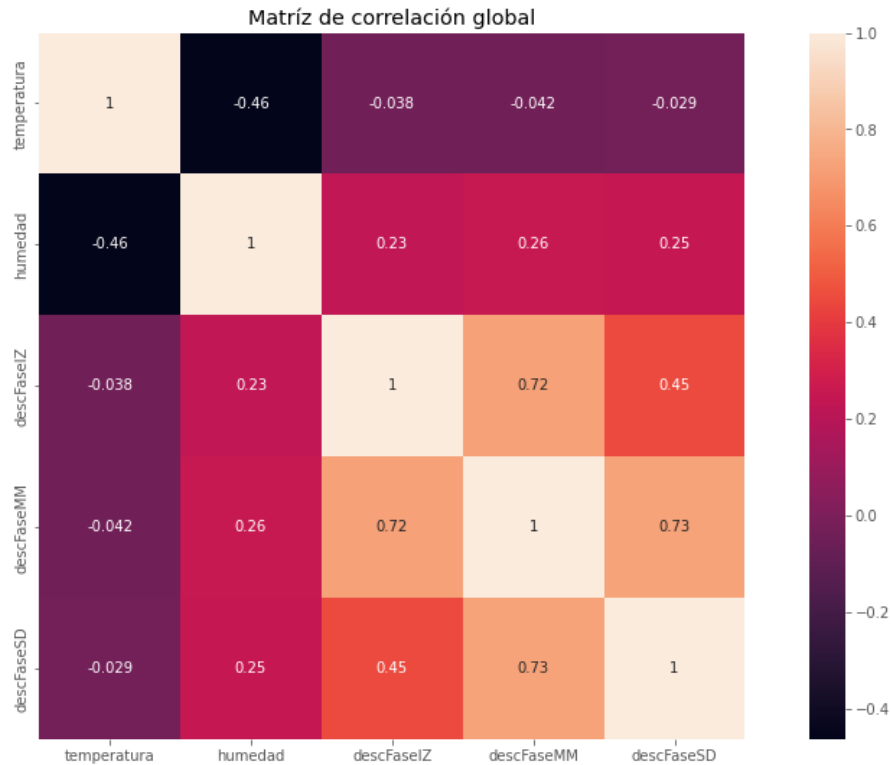


Figura 4.10: Matriz de correlación global

En la matriz de correlación con la totalidad de datos, presentada en la Figura 4.10, se reafirma la relación entre las tres mediciones de conteo de descarga de fase dicha anteriormente, además se observa una relación inversa entre la humedad y temperatura, la cual se puede apreciar más en detalle en la Figura 4.11, donde se traza en azul la tendencia lineal de los puntos.

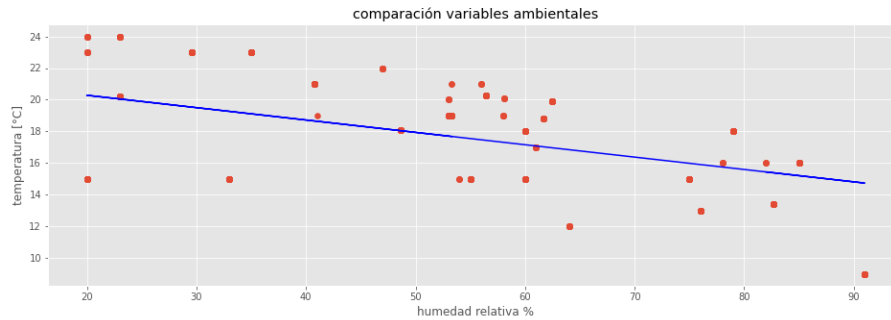


Figura 4.11: Humedad vs temperatura global

La teoría desde la disciplina de electricidad dice que a mayor humedad ambiental, mayores posibilidades de producirse el efecto corona, es por ello que si analizamos la correlación entre humedad y las diferentes descargas de fase, veremos que están entre 0.23 y 0.26, lo cual en primera instancia muestra que sea cual sea el fenómeno que esté afectando (sea el efecto corona o no), podemos decir que es el mismo que produce que la correlación sea similar. En segunda instancia, a través de la Figura 4.12, podemos verificar que si bien la correlación no es muy fuerte, existe una tendencia a que se produzcan descargas más intensas (en magnitud) a mayor % de humedad relativa.

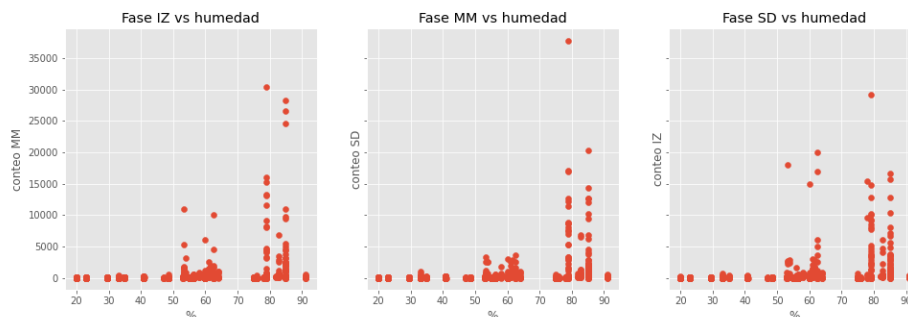


Figura 4.12: Conteo de descargas vs humedad relativa global

4.3. Entrenamiento del modelo

Dados los resultados anteriores, es posible observar que no existe dato alguno que pueda ser utilizado como *ground truth* o verdad fundamental, ya que la etiqueta EPRI es lo que se tiene en duda y, de alguna

forma, se debe despejar.

En consecuencia, realizar un modelado a través de aprendizaje supervisado no es opción, por lo que se procederá a probar si existe alguna disposición natural de los datos a agruparse, en base a las descargas de fase juntas (o no) a las variables ambientales, es decir, se optará por un modelado mediante aprendizaje no supervisado utilizando técnicas de *Clustering*.

4.3.1. *Clustering*

Antes de utilizar cualquier algoritmo de *clustering*, se realizará una visualización en dos dimensiones de la totalidad de datos mediante t-SNE. Esta técnica sirve para ver si la similitud de los datos es capaz de retener alguna estructura local, al mismo tiempo que revela alguna estructura global importante. [14].

A continuación se presentan los principales resultados de aplicar t-SNE, tanto para el total de datos, incluyendo las variables categóricas y quitando estas, es decir, para las variables cuantitativas solamente.

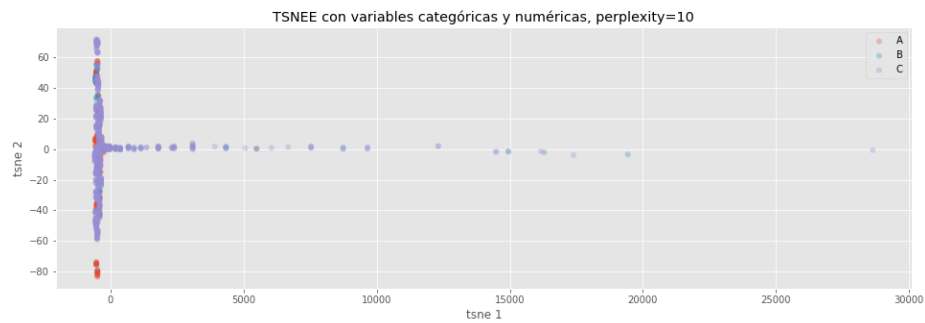


Figura 4.13: Visualización mediante t-SNE incluyendo variables categóricas, con parámetro *perplexity* = 10

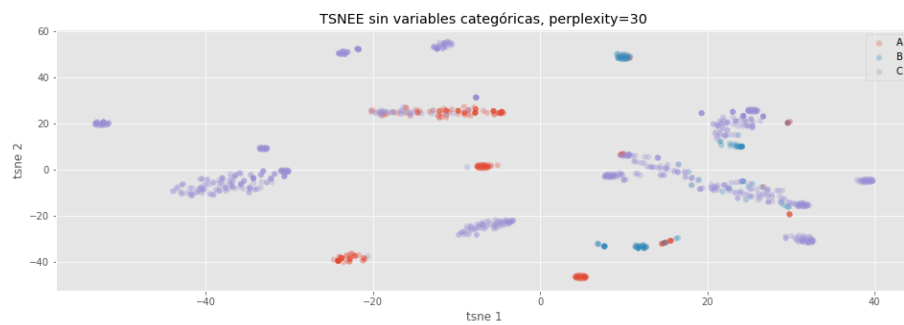


Figura 4.14: Visualización mediante t-SNE sin variables categóricas, con parámetro *perplexity* = 30

Para la visualización con variables categóricas, presentado en la Figura 4.13, se utilizaron las variables de descarga de fase, si presenta ruido, circuito al que pertenece, medidas ambientales, tipo de

material del aislador, clima al que está expuesto y el tipo de contaminación que le rodea, cabe destacar que estas dos últimas variables están descritas de manera generalizada para todos los datos de la cartilla, no considerando si, por ejemplo, la línea parte en un ambiente con contaminación agrícola y termina en otro con contaminación salina. Para la visualización sin variables categóricas, referidas también como variables cuantitativas, se utilizaron las descargas de fase y las medidas ambientales.

Es posible apreciar que no existen indicios de agrupamiento para el conjunto de variables categóricas y numéricas (cuantitativas), pero ocurre lo contrario cuando solo se consideran estas últimas, donde se ve una estructura más definida con distintos cúmulos. Pese a que esta última se ve prometedora, es necesario mencionar que las letras de la leyenda hacen referencia a la etiqueta EPRI del dato asociado, por lo tanto, la técnica no encontró ninguna forma de agrupar los datos correspondientes a etiquetas iguales y separar los distintos. En el anexo A.2 se dejan las figuras de la aplicación de t-SNE con diferentes variables y parámetro *perplexity*.

En consecuencia, se tiene cinco dimensiones (descargas de fase IZ, MM y SD, mas temperatura y humedad) para realizar *clustering*, por lo que se procede a reducir dimensiones utilizando diferentes *kernels* con PCA.

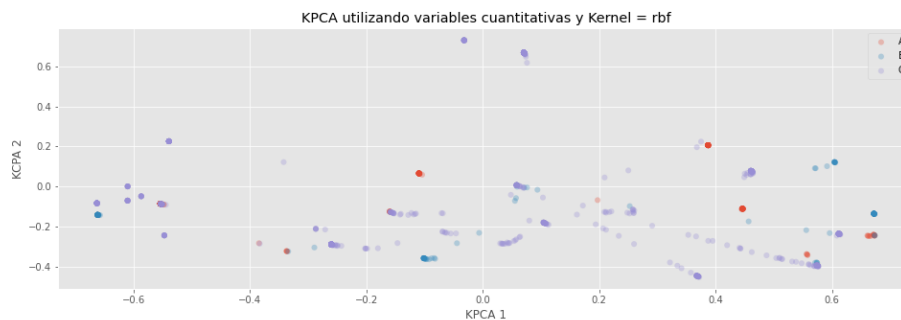


Figura 4.15: PCA utilizando *kernel rbf*

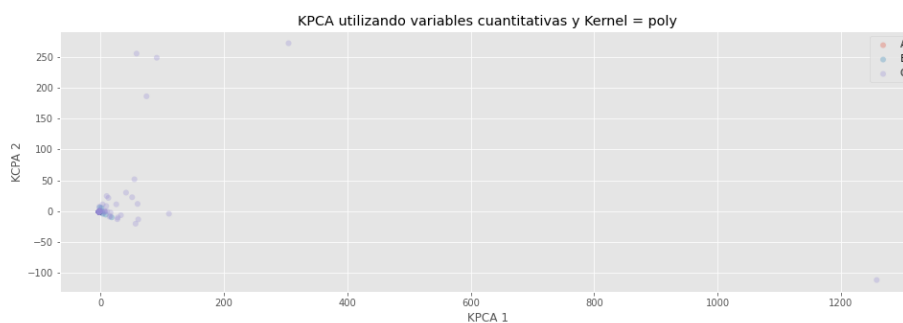
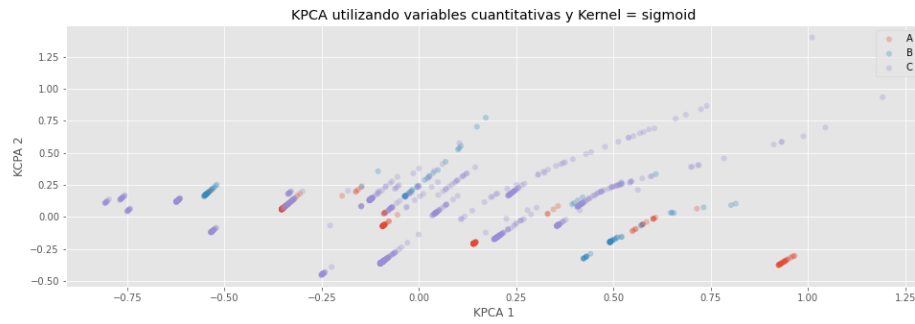
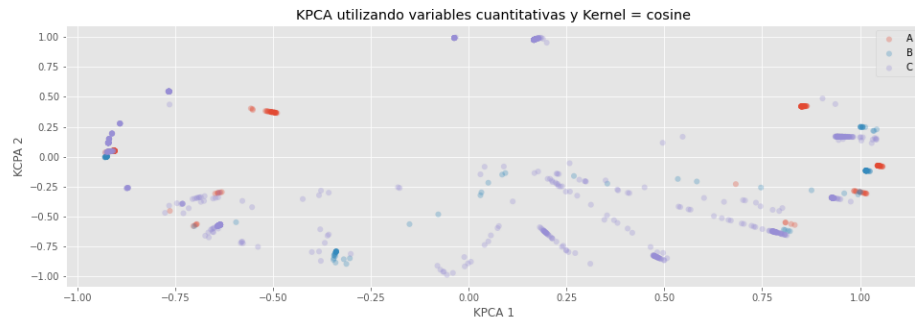
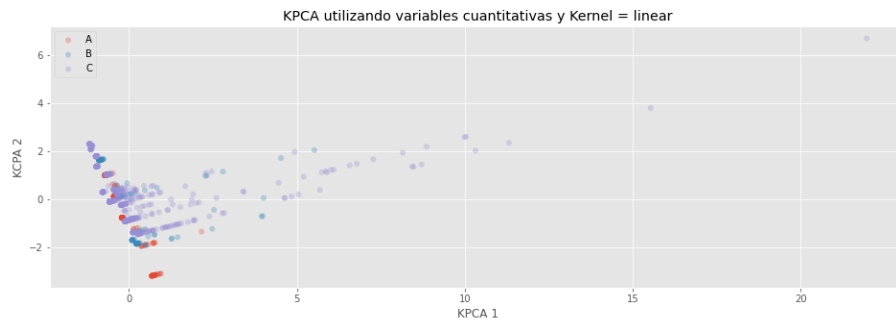


Figura 4.16: PCA utilizando *kernel poly*

Figura 4.17: PCA utilizando *kernel sigmoid*Figura 4.18: PCA utilizando *kernel cosine*Figura 4.19: PCA utilizando *kernel linear*

De las Figuras de la 4.15 a la 4.18, presentando el resultado de aplicar PCA con diferentes bases, se aprecia que no existe una estructura definida de donde se puedan obtener *clusters*, esto debido a que son muy pocos datos. Es por ello, y considerando que el interés principal de este trabajo es establecer una metodología (de análisis sustentada en el método científico), que se procederá alternativamente a desarrollar un detector de anomalías, el cual podrá discriminar entre dato normal y anomalía.

4.4. Detección de anomalías

A continuación se procederá a utilizar tres técnicas para detectar anomalías, comenzando por Z-Score, que realizará detección de anomalías de valores extremos, y terminando por una que se basa en la construcción de árboles. En esta sección no se utilizarán las variables ambientales, ya que su aporte es marginal dado que no tienen mucha variabilidad, además es físicamente muy poco probable (por no decir imposible) que exista una anomalía en la temperatura o humedad relativa, caso contrario para las variables de conteo de descargas, las cuales serán el centro del análisis siguiente.

4.4.1. Z-Score

Para decidir si es anomalía o no, se utiliza un umbral de Z-Score= 3, el cual es un valor típico usado en la literatura, también conocido como $\sigma = 3$.

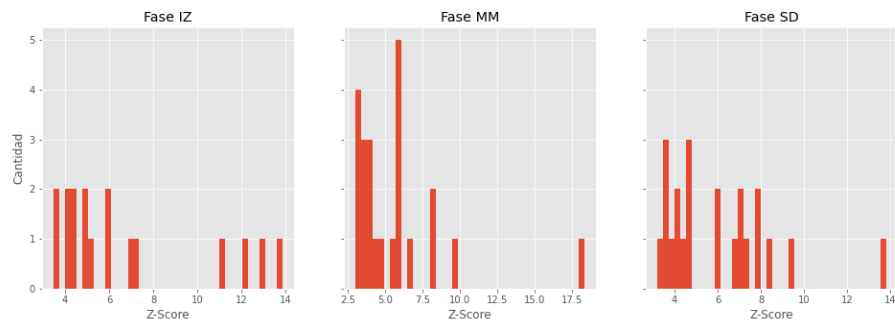


Figura 4.20: histograma anomalías Z-Score

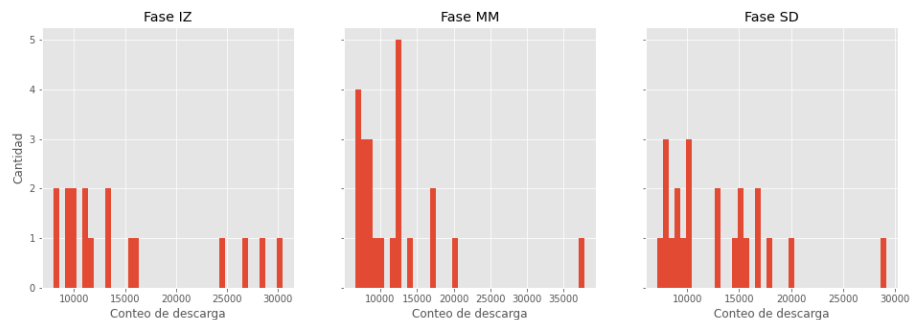


Figura 4.21: histograma Z-Score con el valor respectivo de conteo de descarga

En la Figura 4.20 se presenta el histograma con las anomalías detectadas usando esta técnica con su Z-Score, y en 4.21 el respectivo valor de conteo de descarga que le pertenece. Es posible observar que, aún así con un umbral alto, se detectan 36 datos anómalos del total usando esta técnica.

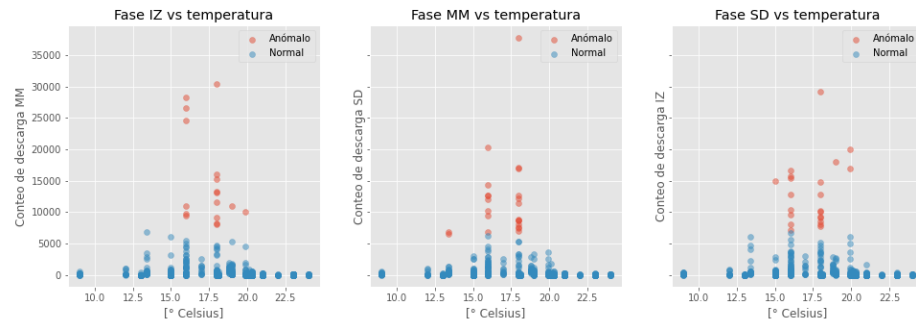


Figura 4.22: Datos anómalos Z-Score y temperatura

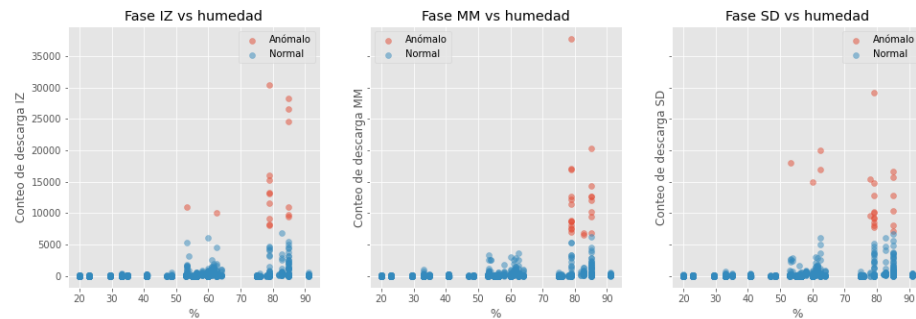


Figura 4.23: Datos anómalos Z-Score y humedad

Al graficar las descargas de fase marcadas como anomalías y las normales, versus la temperatura y humedad, se obtienen las Figuras 4.22 y 4.23, respectivamente. En estas se puede observar como esta técnica identifica efectivamente los datos cuyo valor de conteo de descargas es extremo.

Local Outlier Factor

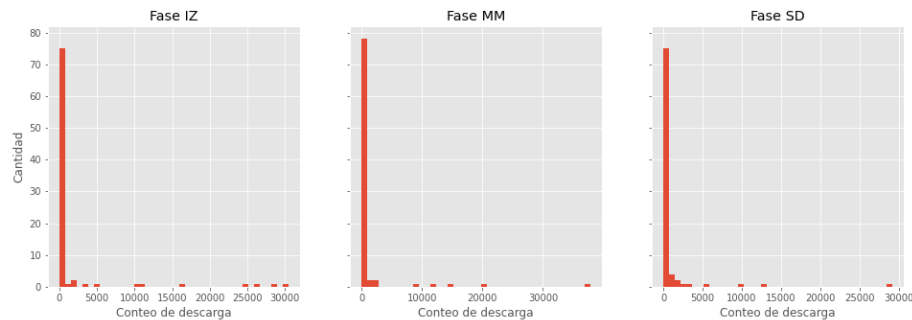


Figura 4.24: histograma anomalías LoF

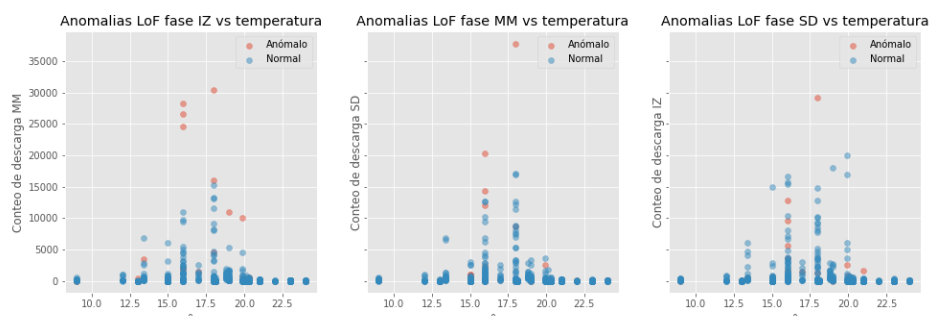


Figura 4.25: Datos anómalos LoF y temperatura

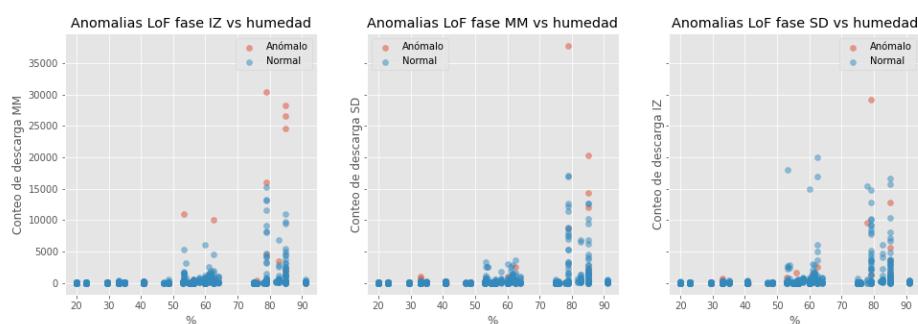


Figura 4.26: Datos anómalos LoF y humedad

En la Figura 4.24 se aprecia la distribución en histogramas de los datos marcados como anomalías por esta técnica, la cual tiene un principio distinto a Z-Score, donde se ve que existen diferencias. Estas se pueden apreciar también comparando con los gráficos de conteo de descargas versus temperatura y humedad, Figuras 4.25 y 4.26 respectivamente.

Isolation Forest

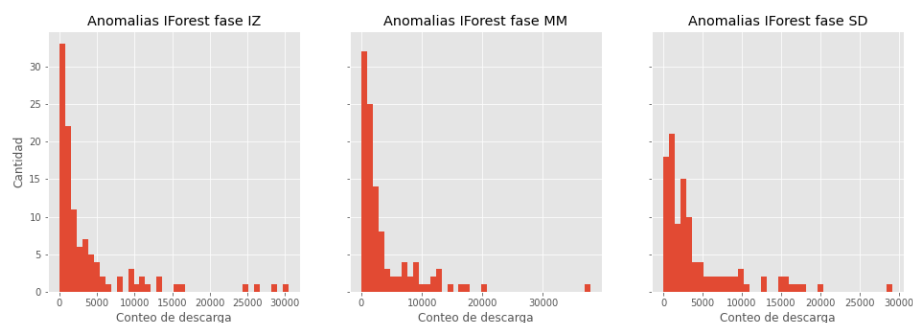


Figura 4.27: histograma anomalías IForest

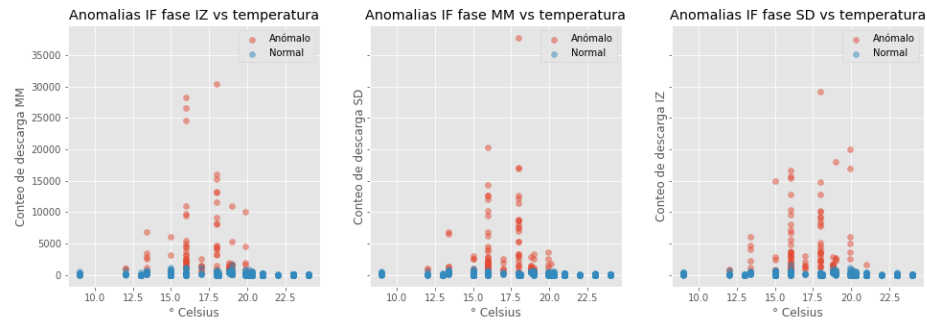


Figura 4.28: Datos anómalos IForest y temperatura

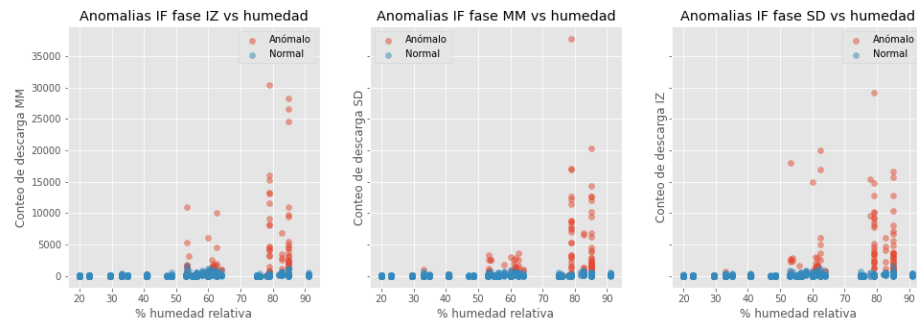


Figura 4.29: Datos anómalos IForest y humedad

En la Figura 4.27 se puede apreciar la distribución de las anomalías encontradas por esta técnica, la cual identificó una mayor cantidad de datos como anomalías. También en 4.28 y 4.23 se aprecian las anomalías versus temperatura y humedad, respectivamente, donde es notorio que de las tres técnicas presentadas, esta parece ser la más estricta.

4.5. Evaluación del modelo

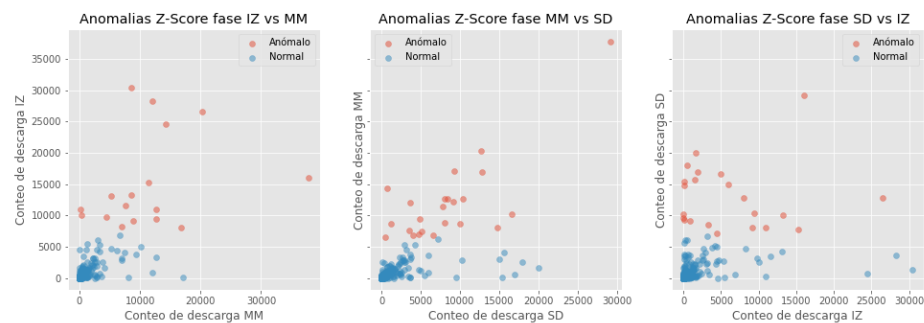


Figura 4.30: conteo de descargas anómalas Z-Score

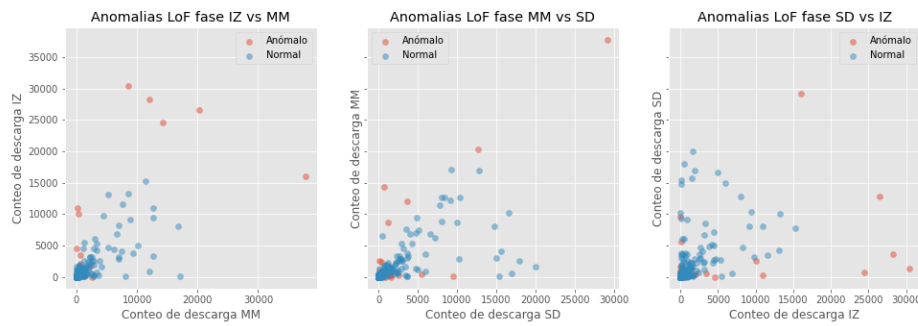


Figura 4.31: conteo de descargas anómalas Local Outlier Factor

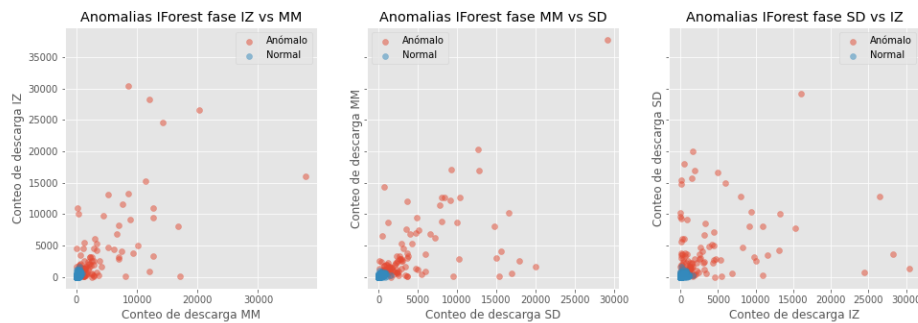


Figura 4.32: conteo de descargas anómalas Isolation Forest

En las Figuras 4.30, 4.31 y 4.32 se puede apreciar las diferentes descargas de fase graficadas, en las que se evidencia la capacidad de Z-Score para encontrar anomalías de valores extremos y también el (aparentemente) buen desempeño de Isolation Forest para aislar anomalías. Por otro lado Local Outlier Factor puede estar teniendo un mal desempeño debido a que el algoritmo se basa en mediciones de distancia para estimar densidad, y al tener datos sin una estructura muy definida, puede verse afectado en su efectividad.

Técnica	Descarga de fase	Mínimo	Máximo	Promedio	Desv. Estándar
Z-Score	IZ	33	30.416	8.386,94	8.290,21
	MM	83	37.800	8.537,14	7.199,51
	SD	183	29.150	9.176,92	6.412,44
Local Outlier Factor	IZ	0	30.416	1.897,44	6.136,06
	MM	0	37.800	1.223,11	5.003,29
	SD	0	29.150	872,54	3.574,73
Isolation Forest	IZ	16	30.416	3.798,69	5.863,19
	MM	16	37.800	3.822,97	5.376,12
	SD	16	29.150	4342,81	5.154,40

Tabla 4.1: Resumen estadístico conjuntos de anomalías

En la tabla 4.1 se puede apreciar un resumen estadístico de los distintos conjuntos anómalos que se forman al utilizar las diferentes técnicas, cabe destacar que tanto Z-Score, LoF e IForest logran encontrar 36 (3,35 %), 87 (8,09 %) y 108 (10,04) valores anómalos, respectivamente.

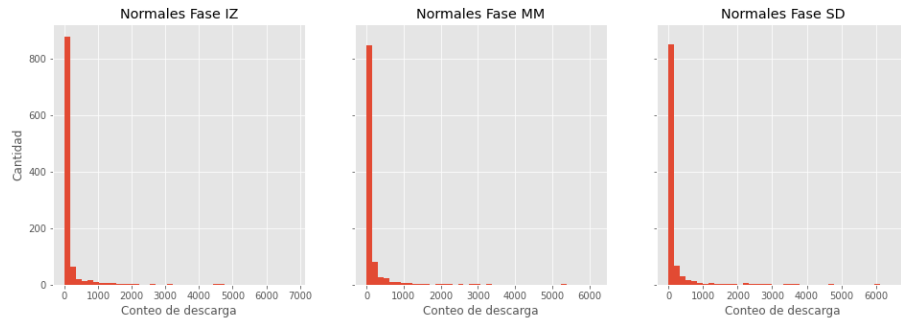


Figura 4.33: histograma valores normales Z-Score

Como se puede apreciar en la Figura 4.33, la distribución de conteo de descargas sin las anomalías detectadas por Z-Score logra reducir la cola superior, siendo una técnica robusta ante valores muy extremos.

Utilizando Isolation Forest se logran encontrar 108 anomalías, las cuales una vez descartadas hacen que la distribución anteriormente presentada sea máximo hasta un conteo de descargas de 1600 aproximadamente, como se presenta en la Figura 4.34.

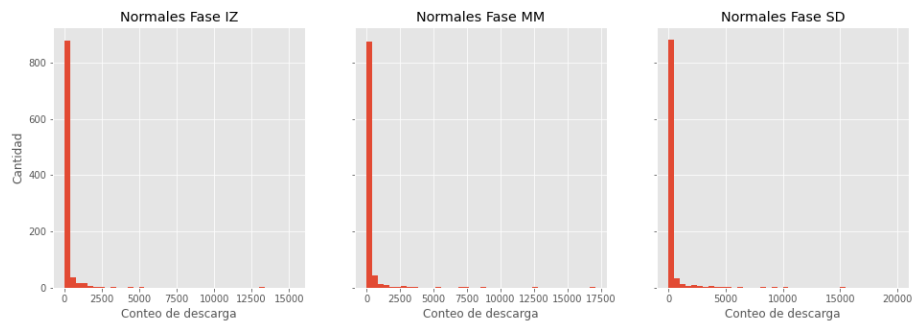


Figura 4.34: histograma valores normales IForest

Local Outlier Factor, como se muestra a continuación, no reduce la dispersión de los datos sin las anomalías, por lo que es posible que no tenga buen desempeño.

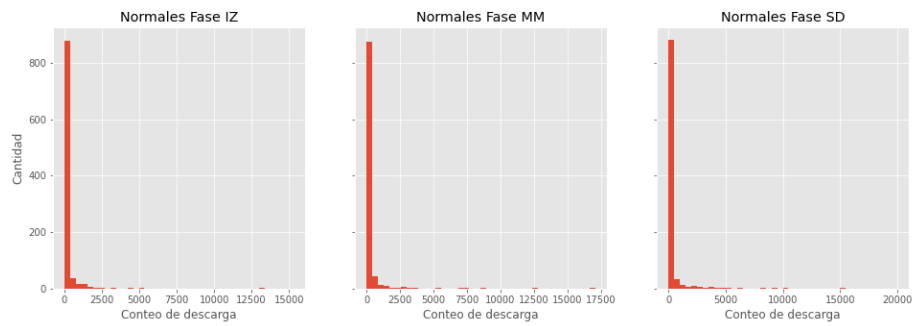


Figura 4.35: histograma valores normales LoF

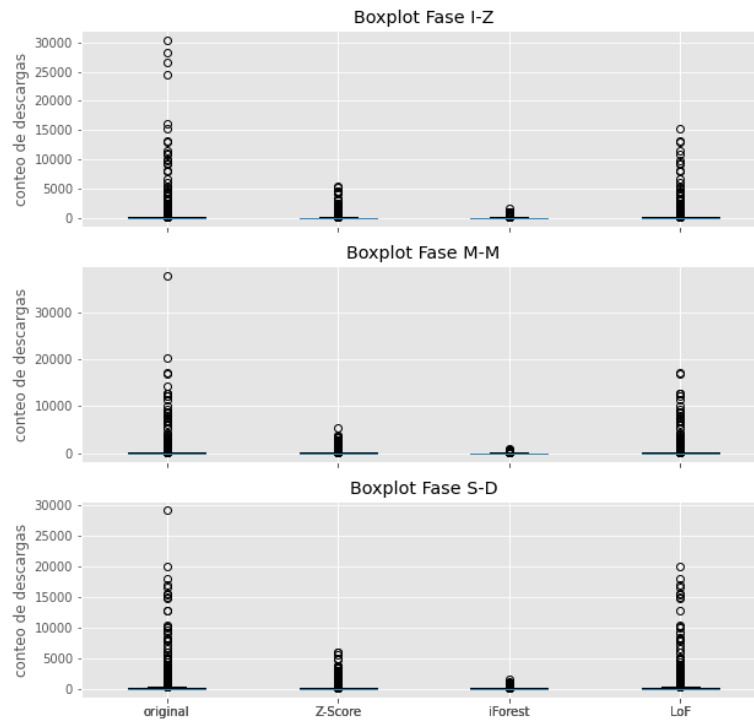


Figura 4.36: Boxplot

Si visualizamos los datos sin las anomalías en *boxplot* se obtiene lo que se muestra en la Figura 4.36, donde está más claro que tanto Z-Score, como Isolation Forest, reducen la dispersión al eliminar los datos que podemos considerar anómalos. Sobre la etiqueta “original” se puede observar lo disperso que se encontraban los datos con las anomalías.

Las aplicaciones de las anomalías contemplan desde la detección de fraude bancario, hasta usos en medicina, seguridad en redes o control de calidad, entre otros. Es esta última aplicación la que motiva a realizar el supuesto, que bajo condiciones de operación sin problemas, los aisladores normales son los calificados como EPRI A o EPRI B (no necesita lavado el aislador o podría ser necesario). En el caso de la calificación EPRI C (realizar lavado a corto plazo) y EPRI D (realizar lavado urgente), se consedaráran como el grupo de falla (anómalo), dado que si no se toma la acción recomendada, el sistema eléctrico perderá estabilidad y puede ocasionar un corte de energía.

En base a lo anterior y como primer acercamiento, se verificarán medidas de cohesión y separación de *cluster*, como medida de desempeño del detector de anomalía ante la falta de *ground truth*, debido a que (dado el fenómeno en cuestión y las variables) las anomalías de tipo valor extremo son más predominantes que las aisladas.

	Davies-Bouldin	Silhouette	Calinski Harabasz
iForest	1.076009	0.841220	497.597008
Original	4.067918	-0.164582	6.075134
LOF	2.968176	0.568030	20.700506

Tabla 4.2: Medidas de desempeño del detector de anomalía, grupos EPRI A-B y EPRI C (anómalo)

Como se puede observar, se obtienen diferentes indicadores (Davies-Bouldin, Silhouette y Calinski Harabasz) usados en el área como medida de desempeño para *clustering*, asumiendo que en sí aquí se tienen dos grupos: normales y anormales, obteniendo un mejor desempeño Isolation Forest. Además, es posible hacer una comparación entre el estado de los datos al comienzo de este trabajo versus al final, tal y como se ve en la tabla como “Original”, lo que reafirma la mejora de separación o proceso de etiquetado.

Considerando que se obtienen mejores indicadores utilizando Isolation Forest y que Z-Score es una regla de oro en estadística, se utilizarán estos dos como componentes del detector de anomalías como solución. La metodología será simple, si existe un registro con alguna característica de descarga parcial (IZ, MM o SD) anómala por Z-Score y, al mismo tiempo, también se marca como tal por Isolation Forest, entonces el dato se considerará como anómalo tipo EPRI C. Si se re-etiquetan los datos, se obtienen los siguientes resultados:

Etiqueta	Davies-Bouldin	Silhouette	Calinski Harabasz
Nueva	3.565275	-0.445923	6.129291
Original	5.474676	-0.420857	5.725213

Tabla 4.3: Medidas de desempeño del detector de anomalía, grupos EPRI A, EPRI B y EPRI C (anómalo, re-etiquetado)

Es posible observar que, calculando las medidas de desempeño especiales para *clustering*, la asig-

nación de etiquetas nuevas en comparación a la antigua, arroja mejores resultados. Para el índice de Davies-Bouldin se obtuvo un menor valor, indicando que los grupos están más compactos y los centros de cada uno más separados. El criterio de Calinski-Harabasz, también demuestra una mejora en la relación entre los miembros de un grupo en particular, en comparación al resto perteneciente a otros grupos.

Silhouette *score*, al contrario, nos indica que los grupos empeoraron su agrupación, esto puede deberse a la existencia de elementos solapados, ya que como hemos visto en los análisis anteriores: no existe una estructura definida para la separación de los *clusters*. Sin embargo, los valores son similares y reafirma, que en base a los valores de las descargas de fase, la forma de agrupación (EPRI) utilizada no es muy precisa en separar las categorías, lo cual reafirma el por qué no hemos utilizado técnicas de *clasificación* para la solución del problema, desconfiando en primera instancia del etiquetado EPRI.

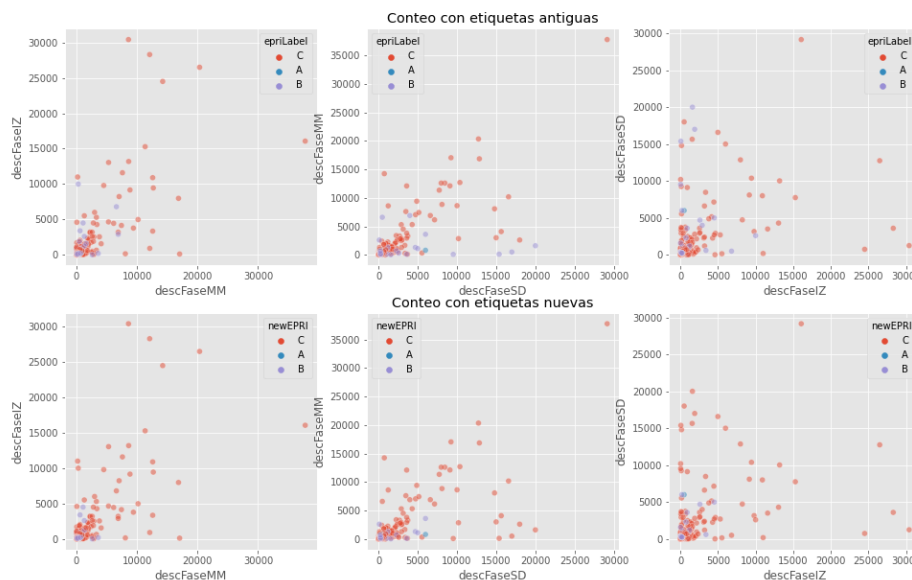


Figura 4.37: Comparación descarga de fase, etiquetas nuevas versus originales

Si se compara la gráfica de la descarga de fase contra otra distinta, con las viejas y nuevas etiquetas, ver Figura 4.37, es posible verificar que hay registros que han cambiado su coloración, en especial con registros asociados al EPRI B, el cual en teoría debiera ser el más cercano al EPRI C. Al mismo tiempo, por lógica no deberían existir registros EPRI A que se hayan re-etiquetado a EPRI C. Concretamente, se presenta en la siguiente tabla la evolución entre el etiquetado original y el nuevo.

Etiqueta	Original	Nueva
A	216	216
B	109	102
C	751	758

Tabla 4.4: Cantidad de datos y sus etiquetas, original y nueva

Por lo tanto, considerando que las anomalías nuevas a encontrar debieran ser pocas, los resultados son aceptables.

4.6. Implementación

Debido a la variabilidad encontrada en las cartillas de inspección, teniendo en cuenta que cada línea de transmisión tiene diferentes condiciones, y debido a los resultados anteriores presentados, se implementarán dos detectores de anomalía.

4.6.1. Detector de anomalías

Por un lado, Z-Score se desempeñó de manera robusta, pero al usarlo con la totalidad de datos de descargas de fase y tener una dispersión alta, su efectividad se ve reducida dada su dependencia estadística de medidas de tendencia central. Es por ello que se implementará como un detector de anomalías local, utilizando como entrada cada descarga de fase, es decir, existirán tres Z-Score, uno por cada descarga de fase para cada registro en particular.

Por el otro, Isolation Forest tuvo mejor desempeño que Local Outlier Factor, y considera más datos de entrada que solo la descarga de fase, por lo que se utilizará como un detector de anomalías global, siendo entrenado con la totalidad de datos de las cartillas.

Ambos detectores de anomalía trabajarán velando que no existan datos extraños, tanto a nivel local como global, generando una sugerencia de etiqueta EPRI C (esto debido a que en los datos existentes no hay registro de EPRI D) en caso que ambos detecten como anomalía el mismo registro, evitando el posible sesgo que puede tener Z-Score al mirar solo los datos de la cartilla de inspección de la que es parte, y lo mismo para Isolation Forest, al haber sido entrenado con una mayor cantidad de datos que posiblemente tengan mayor variabilidad.

4.6.2. Integración en Excel

Para la ejecución del detector de anomalías se utilizará el plugin PyXll para cargar el modelo y su comunicación con los datos del excel, además se habilitará una hoja solo de lectura donde se mostrarán los resultados de los detectores de anomalía.

	A	B	C	D	E	F	G	H	I	J	K
1											
2							Z-Score	I-Z	M-M	S-D	
3							Promedio	1979,965517	813,5344828	421,8793103	
4							Desviación	8036,446079	2602,732743	930,5531184	
5											
6											
7	Estructura	Circuito		Descarga por fase			Detector de anomalía global	Detector de anomalía particular			Sugerencia
8		1	2	I-Z	M-M	S-D		I-Z	M-M	S-D	
9	137	x		33	83	50	No	-0,242266979	-0,280679791	-0,399632544	
10	140	x		33	116	50	No	-0,242266979	-0,26800081	-0,399632544	
11	141	x		8200	2400	100	Si	0,773978251	0,609538387	-0,34590106	
12	142	x		483	1600	4200	Si	-0,186272079	0,302169141	4,060080628	EPRI C
13	143	x		300	250	1300	No	-0,209043338	-0,216516461	0,943654556	
14	144	x		43750	14200	2850	Si	5,197575405	5,143234761	2,60933056	EPRI C
15	150	x		100	83	100	No	-0,233929961	-0,280679791	-0,34590106	
16	151	x		33	33	116	No	-0,242266979	-0,299890369	-0,328706985	
17	152	x		50	33	33	No	-0,240151617	-0,299890369	-0,417901249	
18		Cartilla inspección	Anomalías	Clasificación EPRI	Información adicional						

Figura 4.38: detector de anomalías en la cartilla de inspección

En la hoja mostrada en la figura 4.38 se pueden apreciar las columnas de estructura, circuito y descarga por fase, las cuales son copiadas automáticamente desde la hoja general de la cartilla de inspección. Además se tienen cuatro columnas dedicadas a los detectores de anomalía, global y particular, y finalmente una sugerencia de etiqueta EPRI.

Por otro lado, se ha incluido un apartado que muestre información acerca del Z-Score general de la cartilla, incluyendo el promedio y desviación para cada variable a la que se le aplica este análisis, con el objetivo de aportar más información general de las condiciones de la línea al usuario.

Como se ha explicado anteriormente, se analizan por separado las tres variables de descargas de fase I-Z, M-M y S-D (llamadas variables eléctricas), y comparando cada registro con la distribución lograda por el conjunto de datos de la línea, donde se marcará como anomalía en base al *threshold* $|Z - Score| > 3$. Además a un costado del detector particular o local, se encuentra el campo del detector de anomalía global, que fue entrenado con la totalidad de datos aportados por la empresa al momento de realizar esta memoria.

Como valor agregado al corto plazo, la cartilla entregará una sugerencia de EPRI C al encontrar un registro anómalo para ambos detectores, es decir, un registro es marcado como anomalía tanto al ser comparado con el resto de datos de su cartilla en particular, y al ser comparado con la totalidad de datos (global). Cabe destacar que en los datos entregados por la empresa no se ha encontrado registro alguno de EPRI D, es por ello que se decide ignorarlo.

Capítulo 5

Conclusiones

En el presente capítulo se analizarán los objetivos definidos y la hipótesis definida, con el objetivo de verificar su cumplimiento, y las razones, en caso que corresponda, de su implementación.

5.1. Conclusiones generales

A lo largo de este trabajo, se ha demostrado que con los datos actuales no es posible modelar la periodicidad en el tiempo, con la que se deben limpiar los aisladores eléctricos dado a una falta de datos. Por otro lado, sí se evidencia que es posible construir un modelo multivariable que ayude a la empresa en sus estrategias de planificación de mantenimientos, reduciendo los criterios subjetivos debido al detector de anomalías y el etiquetado sugerido de EPRI C, el cual en parte cumple con la premisa de modelar la periodicidad de limpieza.

Al inicio de este trabajo se buscaba generar un sistema predictivo, lo cual no fue posible debido a que se contaba solo con una muestra temporal por cada registro de medición (es decir, datos insuficientes para trabajar con series de tiempo), es por ello que se ha trazado una ruta de seguimiento para la empresa, con el fin de que se llegue al sistema predictivo. Lo realizado no es más que parte del camino a recorrer, lo que se explica más en detalles en las proyecciones.

El detector de anomalías desarrollado, junto con la actualización de la cartilla y unificación de formato de esta, permite realizar una estimación menos subjetiva del real estado de los aisladores, ya no dependiendo tanto de la experiencia de la persona que los lee. Se ha desarrollado un sistema que no interfiere con los procedimientos de la empresa, es claro y conciso, sin mayores complejidades.

5.2. Conclusiones específicas

- A través del análisis de datos se pudieron encontrar inconsistencias en los valores dentro de las cartillas, insumo que fue utilizado para una reestructuración de esta y estandarización de formatos.
- El método de aprendizaje no supervisado Isolation Forest logra una mejor aislación de los datos anómalos, y junto con Z-Score logran realizar una corrección de etiquetado que hace que los grupos estén mejor cohesionados y diferenciados entre sí.
- La cartilla de inspección ha sido mejorada sustancialmente en comparación a como se encontraba cuando se comenzó este trabajo, impactando directamente en la mejora del procedimiento de inspección y posterior análisis del área de mantenimiento de Chilquinta.
- Con la solución realizada se han mejorado la calidad de los datos y se han creado herramientas de análisis más objetivas. También se ha trazado un camino que logrará la mejora de gestión de recursos, optimización de tiempos y mejor accesibilidad a la información.

Lo que se construyó en este trabajo es parte de una solución mayor, la cual se puede ver en la Figura 5.1, donde cada solución en particular ataca a un problema diferente, el cual es especialmente abordado por cada una de las disciplinas dialogando entre sí para lograr una cohesión del trabajo. A continuación se especifican más estas etapas:

1. Estandarizar procedimientos: Orientada a mejorar la manera y protocolos que sigue el personal de inspección de Chilquinta, teniendo como objetivo construir un manual de inspección.
2. Reestructurar cartilla de inspección: Orientada a estandarizar y unificar formatos en la cartilla de inspección, además de agregar información de apoyo para la clasificación EPRI y de los aisladores.
3. Detector de anomalías: Orientado a velar por la calidad de los datos y generar sugerencia en clasificación EPRI cuando existan registros extraños.
4. Centralizar fuente de datos históricos: Orientado al almacenamiento y visualización, centralizando la fuente de datos y mejorando la accesibilidad a estos.
5. Digitalizar recolección de información: Orientado al reemplazo de la cartilla de inspección por su homólogo digital, ingresando directamente la información a la base de datos y pudiendo ser consumida por el ingeniero de mantenimiento a cargo.

5.3. Proyecciones

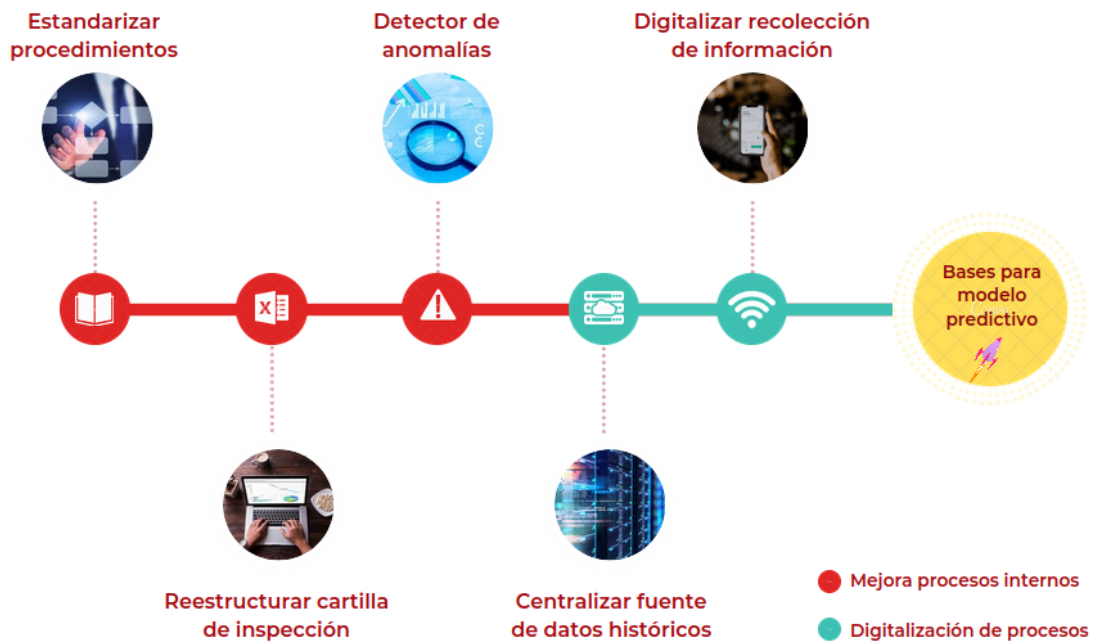


Figura 5.1: Proyección del diseño de la solución predictiva

Lo abordado por el equipo corresponde a las etapas de mejora de procesos internos y se dejará como trabajo a futuro y opcional la digitalización de procesos, lo cual puede propiciar un mejor escenario para lograr el objetivo de construir un modelo predictivo. A continuación se especifican más estas etapas:

Por otro lado, se espera que se implemente la nueva cartilla con el detector de anomalías y sea probado durante un período en las inspecciones próximas de la empresa. Esto debido a que al momento de realizar este trabajo, la empresa está entrando en un período de transición de dueños y los tiempos no se han ajustado a los del período de la memoria.

Anexos

A.1. Resumen estadístico 14 cartillas

En la siguiente tabla se puede apreciar un resumen estadístico para la temperatura y humedad de 14 cartillas

Archivo de cartilla asociado	Temp. promedio	Temp. desviación	Humedad promedio	Humedad desviación	Cantidad registros
10068894_Informe Corona LLTT_CC-BO.xlsx	19.0	0.0	0.53	0.028	18
10070670_Informe Corona LLTT_LC-EM.xlsx	-	-	0.67	0.047	63
10069891_Informe Corona LLTT_LV-LA 1-70.xlsx	19.2	2.4	0.60	0.086	136
10070047_Informe Corona LLTT_AM-LE 1-64.xlsx	15.7	1.3	0.53	0.027	37
10068897_Informe Corona LLTT_SL-AS.xlsx	17.1	1.0	0.78	0.093	158
10069823_Informe Corona LLTT_SP-MI E23-E24-E25 (AM).xlsx	16.0	0.0	0.78	0.000	6
10068899_Informe Corona LLTT_SP-PE.xlsx	22.5	1.1	0.28	0.064	41
10070673_Informe Corona LLTT_TA-SF.xlsx	23.0	0.0	0.34	0.000	27
10070049_Informe Corona LLTT_AM-SA 1-64.xlsx	15.7	1.3	0.53	0.027	37
10069822_Informe Corona LLTT_TO-CC.xlsx	15.0	0.0	0.54	0.000	4
10071388_Informe Corona LLTT_LC-EM (adicional).xlsx	12.0	0.0	0.64	0.000	12
10071385_Informe Corona LLTT_SA-TA (adicional).xlsx	18.3	0.9	0.60	0.014	34
10068889_Informe Corona LLTT_AC-CC.xlsx	20.7	0.8	0.43	0.050	34
10069890_Informe Corona LLTT_CH-CA 1-7.xlsx	15.0	0.0	0.75	0.000	8

Tabla A.1: Resumen estadístico temperatura y humedad 14 cartillas individuales

A.2. t-SNE con diferente *perplexity*

A.2.1. Variables categóricas y numéricas

Para generar los siguientes resultados se usó t-SNE con todas las variables, incluyendo categóricas.

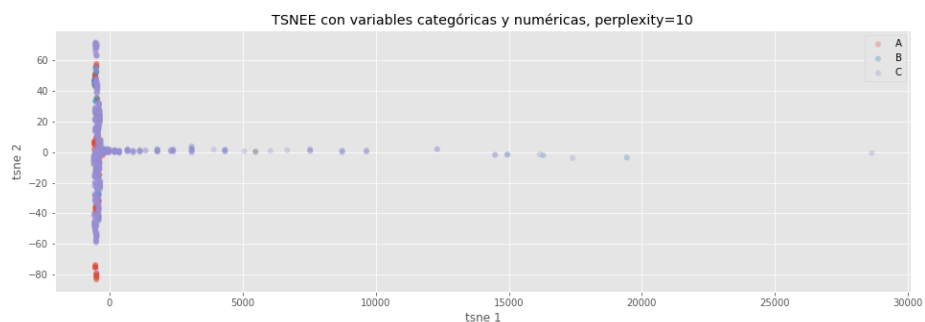


Figura A.2: t-SNE con *perplexity* 10

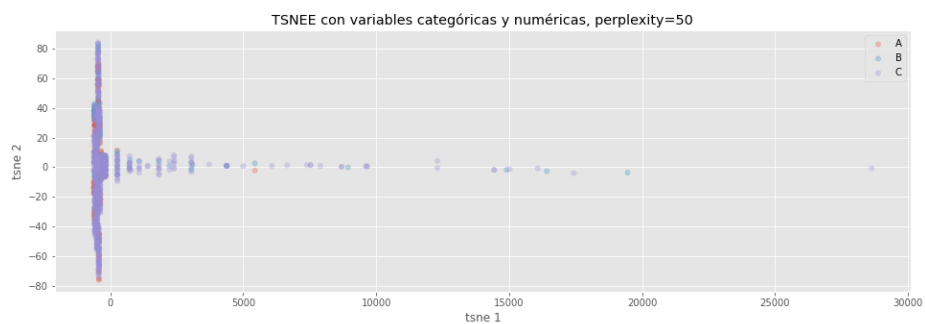


Figura A.3: t-SNE con *perplexity* 50

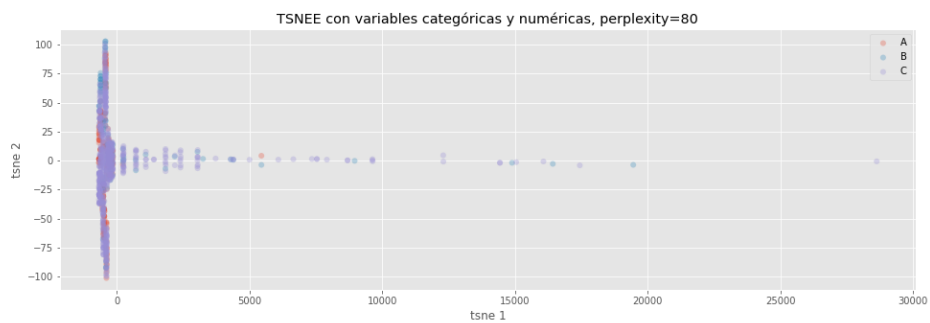


Figura A.4: t-SNE con *perplexity* 80

A.2.2. Solo variables numéricas

Las siguientes figuras presentan los resultados de t-SNE aplicado solo a descargas de fase y variables ambientales.

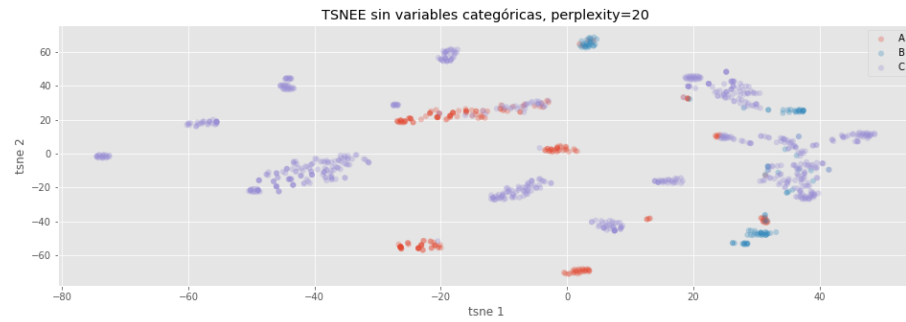


Figura A.5: t-SNE sin variables categóricas con *perplexity* 20

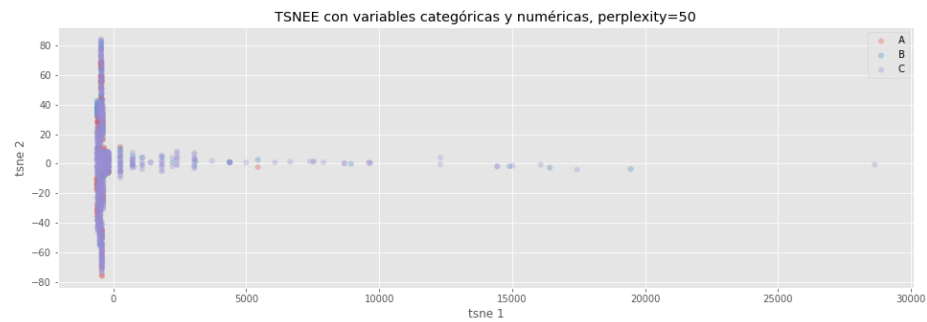


Figura A.6: t-SNE sin variables categóricas con *perplexity* 50

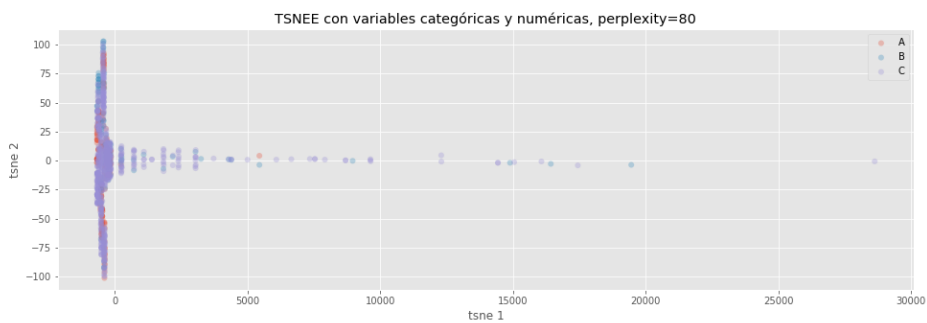
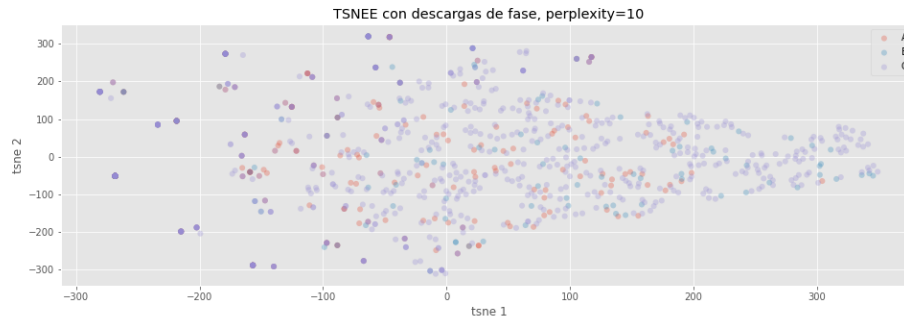
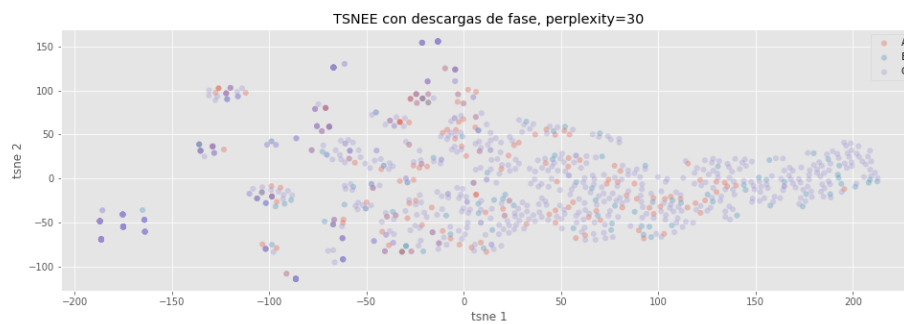
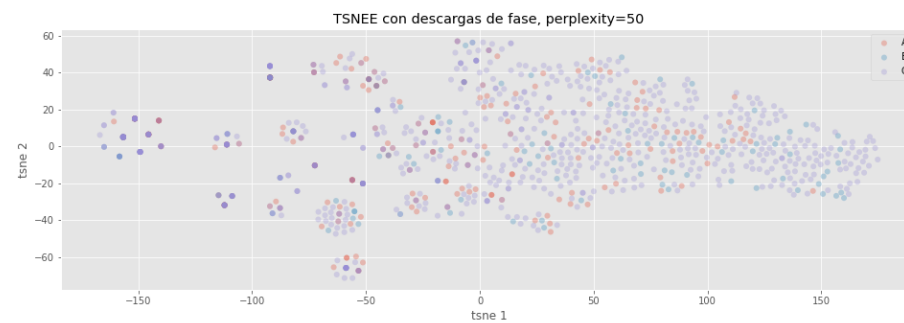


Figura A.7: t-SNE sin variables categóricas con *perplexity* 80

A.2.3. Solo con descargas de fase

Figura A.8: t-SNE utilizando descargas de fase con *perplexity* 10Figura A.9: t-SNE utilizando descargas de fase con *perplexity* 30Figura A.10: t-SNE utilizando descargas de fase con *perplexity* 50

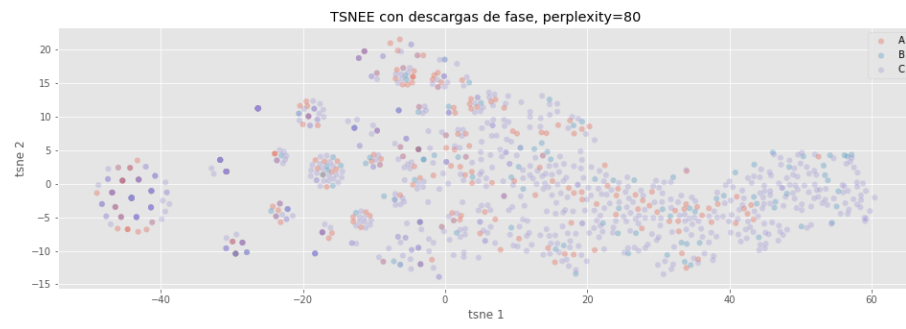


Figura A.11: t-SNE utilizando descargas de fase con *perplexity* 80

Referencias

- [1] Ethem Alpaydin, *Introduction to machine learning*. Cambridge, Massachusetts: The Mit Press, 2014, pp. 1–10.
- [2] C. M. Bishop, *Pattern Recognition And Machine Learning*. Springer, 2016.
- [3] C. C. Aggarwal, *Data mining : the textbook*. Cham Switzerland ; New York: Springer, 2016, pp. 237–242.
- [4] F. T. Liu, K. M. Ting, and Z.-H. Zhou, “Isolation Forest,” *2008 Eighth IEEE International Conference on Data Mining*, Dec. 2008, doi: 10.1109/icdm.2008.17.
- [5] E. Kuffel, W. S. Zaengl, and J. Kuffel, *High voltage engineering fundamentals*. Nueva Delhi (India) Amsterdam (Holanda) Elsevier, 2014.
- [6] E. Stracqualursi, R. Araneo, and S. Celozzi, “The Corona Phenomenon in Overhead Lines: Critical Overview of Most Common and Reliable Available Models,” *Energies*, vol. 14, no. 20, p. 6612, Oct. 2021, doi: 10.3390/en14206612.
- [7] R. Guerreiro Carvalho Cunha, E. T. da Silva, and C. M. S. Medeiros, “An Embedded Application to Identify Degradation in Energized Polymeric Insulators Using Machine Learning and Wavelet Transform,” *IEEE Xplore*, Nov. 01, 2018. <https://ieeexplore.ieee.org/document/8691920> (accessed Jan. 18, 2022).
- [8] X. Tao, D. Zhang, Z. Wang, X. Liu, H. Zhang, and D. Xu, “Detection of Power Line Insulator Defects Using Aerial Images Analyzed With Convolutional Neural Networks,” *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 50, no. 4, pp. 1486–1498, Apr. 2020, doi: 10.1109/TSMC.2018.2871750.
- [9] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, “Indoor Segmentation and Support Inference from RGBD Images.” [Online]. Available: https://cs.nyu.edu/~silberman/papers/indoor_seg_support.pdf

-
- [10] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object Detection via Region-based Fully Convolutional Networks," *arXiv.org*, 2016. <https://arxiv.org/abs/1605.06409>
- [11] C. Vidan and M. Maracine, "Corona Discharge Classification Based on UAV Data Acquisition," *IEEE Xplore*, May 01, 2017. <https://ieeexplore.ieee.org/document/7968634> (accessed Jan. 18, 2022).
- [12] Z. Qiu, J. Ruan, D. Huang, M. Wei, L. Tang, and S. Shu, "Corona onset and breakdown voltage prediction of rod-plane air gaps based on SVM algorithm," *IEEE Xplore*, Oct. 01, 2015. <https://ieeexplore.ieee.org/document/7352007> (accessed Jan. 18, 2022).
- [13] G. G. Roussas, *An introduction to probability and statistical inference*. Amsterdam ; Boston: Elsevier, Academic Press, 2015, p. 208.
- [14] L. Com and G. Hinton, "Visualizing Data using t-SNE Laurens van der Maaten," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008, [Online]. Available: <https://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf>