**BENEMÉRITA UNIVERSIDAD AUTÓNOMA DE PUEBLA**

**COMPLEJO REGIONAL CENTRO - SAN JOSÉ CHIAPA**

# Análisis de información meteorológica y mediciones de campo eléctrico en el área de la ciudad de Puebla.

Que para obtener el título de

**Licenciado en Ingeniería en Sistemas y Tecnologías de la Información Industrial**

**P R E S E N T A**

Edilberto Vazquez Luna

**ASESORES DE TESIS**

M.C. Diana Ivone Tapia López
Dra. María Claudia Denicia Carral

Puebla, Pue., Abril de 2023

# AGRADECIMIENTOS

Quiero expresar mi más sincero agradecimiento a mis asesoras, M.C. Diana Ivone Tapia López y Dra. María Claudia Denicia Carral, por su invaluable ayuda y orientación durante el desarrollo de mi tesis "Análisis de información meteorológica y mediciones de campo eléctrico en el área de la ciudad de Puebla". Gracias a su orientación y conocimientos en el campo de la Minería de Datos, pude realizar un análisis riguroso y detallado de las mediciones de campo eléctrico. Su guía y apoyo fueron fundamentales en el éxito de mi investigación. Asimismo, quisiera expresar mi agradecimiento a todos los profesores de la carrera, quienes con sus enseñanzas y guía me brindaron las herramientas y conocimientos necesarios para poder llevar a cabo esta investigación. De nuevo, les agradezco por su tiempo, paciencia y dedicación en guiarme a lo largo de este proceso.

Asimismo, quiero agradecer al Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE) por proporcionarme los datos necesarios para mi investigación. Sin su colaboración, habría sido imposible llevar a cabo este estudio y obtener los resultados que me llenan de satisfacción. Agradezco al personal del INAOE por su dedicación y compromiso en el mantenimiento y disponibilidad de estos datos, su apoyo y contribución han sido fundamentales en el éxito de mi investigación.

# Resumen

La minería de datos se puede utilizar en cualquier trabajo en el que se requiera hacer descubrimiento/reconocimiento de patrones. Hoy en día, este método es aplicado en el campo de la meteorología con propósitos de predicción. La correcta predicción de los datos con suficiente tiempo de anticipación puede llegar a prevenir desastres futuros, establecer medidas de protección en la infraestructura y alertar a la población. El objetivo de este trabajo es realizar un análisis de los datos meteorológicos y mediciones de campo eléctrico en zonas del estado de Puebla, utilizando técnicas computacionales para correlacionar datos, encontrar patrones y estimar tendencias de comportamiento, para descubrir conocimiento sobre los cambios climáticos de la zona. Para el análisis de los datos se realizó previamente el proceso de Extracción Transformación y Carga por sus siglas en ingles ETL. La transformación de los datos se realiza con el lenguaje de programación Golang (GO), en tanto que el análisis de los datos se lleva a cabo con el software RapidMiner y Power BI, mientras que para la aplicación de modelos de minería de datos se utilizó el software Weka.

**Palabras clave:** Minería de Datos, descargas eléctricas atmosféricas, mediciones de campo eléctrico.

Contenido

Lista de Figuras ..... VI

Lista de Tablas..... VII

1. Introducción ..... 1

1.1 Antecedentes del Proyecto ..... 1

1.2 Planteamiento del Problema ..... 1

1.3 Objetivos ..... 3

1.4 Pregunta de Investigación ..... 4

1.5 Justificación ..... 4

1.6 Alcances ..... 5

1.7 Impacto Socioeconómico ..... 5

1.8 Organización de la tesis ..... 5

2. Estado del arte ..... 6

2.1. Métodos para la predicción del clima ..... 6

2.2. Trabajos existentes ..... 7

3. Marco teórico ..... 10

3.1. Campo Eléctrico ..... 10

3.2. Descarga Atmosférica ..... 10

3.3. Daños ocasionados por descargas eléctricas atmosféricas ..... 12

3.4. Monitor de Campo Eléctrico Atmosférico EFM-100 ..... 12

3.4.1. Funcionamiento del EFM 100 ..... 12

3.5. Detección de nube de tormenta ..... 13

3.6. Detección de rayos ..... 14

3.7. Vantage VUE Weather Station ..... 15

3.7.1. Funcionamiento de la Vantage VUE Weather Station ..... 15

3.8. Minería de datos ..... 16

3.8.1. Definición de minería de datos ..... 16

3.8.2. Técnicas de minería de datos. ..... 16

3.8.3. Arboles de decisión ..... 17

3.8.4. Estructura básica de un árbol de decisión ..... 17

3.9. Proceso KDD ..... 19

4. Marco metodológico ..... 21

4.1. Selección ..... 21

4.2 Diccionario de datos ..... 22

4.2. Preprocesamiento ..... 24

4.2.1. Preprocesamiento del conjunto datos meteorológicos ..... 25

4.2.2. Preprocesamiento del conjunto Campo eléctrico ..... 27

4.2.3. Preprocesamiento del conjunto Rayos detectados ..... 28

4.3. Transformación ..... 30

4.4. Minería de datos ..... 32

4.4.1. Análisis de las variables ..... 33

4.4.2. Experimento 1: Aplicación del modelo J48 al conjunto Rayos-meteorológicos 37

4.4.3. Experimento 2: Aplicación del modelo J48 al conjunto Rayos-meteorológicos discretizado ..... 39

4.4.4 Experimento 3: Aplicación del modelo J48 al conjunto Rayos-campo-eléctrico 40

4.4.5. Experimento 4: Aplicación del modelo J48 al conjunto Rayos-meteorológicos dividido por rangos mediante la variable distancia (distance) ..... 41

4.4.5. Experimento 5: Aplicación del modelo J48 al conjunto Rayos-campo-eléctrico dividido por rangos mediante la variable distancia (distance) ..... 45

5. Conclusiones y trabajo a futuro ..... 50

6. Bibliografía ..... 52

# I. Lista de Figuras

Ilustración 1: Rango de detección de rayos de los medidores de campo eléctrico ..... 3
Ilustración 2: Comportamiento del campo eléctrico ..... 10
Ilustración 3: Tipos de Descargas Eléctricas Atmosféricas ..... 11
Ilustración 4: Monitor de campo eléctrico atmosférico EFM-100. Tomada de (BOLTEK CORPORATION, 2016) ..... 13
Ilustración 5: Detección de una nube de tormenta. Tomada de (BOLTEK CORPORATION, 2016) ..... 13
Ilustración 6: Comportamiento del campo eléctrico ante una descarga atmosférica. Tomada de (BOLTEK CORPORATION, 2016) ..... 14
Ilustración 7: Estación meteorológica Vantage VUE. Tomada de (Davis Instruments, 2011) ..... 15
Ilustración 8: Registro capturado por la estación meteorológica ..... 16
Ilustración 9: Estructura de un árbol de decisión ..... 18
Ilustración 10: Proceso KDD ..... 19
Ilustración 11: Captura del conjunto “Datos-meteorológicos” ..... 21
Ilustración 12: Conjuntos “Rayo detectados” y “Campo eléctrico” ..... 22
Ilustración 13: Proceso con el que se obtuvieron los conjuntos finales ..... 25
Ilustración 14: 10 primeras líneas de un archivo CSV del conjunto Datos-meteorológicos ..... 25
Ilustración 15: Conjunto “Datos-meteorológicos” preprocesado ..... 26
Ilustración 16: 10 primeras líneas de un archivo EFM del conjunto “Campo-eléctrico” ..... 27
Ilustración 17: Conjunto “Campo-eléctrico” preprocesado ..... 28
Ilustración 18: 10 primeras del archivo EFMEvents.log del conjunto “Rayos-detectados” ..... 28
Ilustración 19: Conjunto “Rayos-detectados” preprocesado ..... 29
Ilustración 20: Archivo ARFF obtenido del conjunto “Rayos-meteorológicos” ..... 31
Ilustración 21: Archivo ARFF del conjunto "Rayos-campo-eléctrico" ..... 32
Ilustración 22: Rayos detectados de 2016 a 2019 ..... 33
Ilustración 23: Cantidad de rayos detectados por KM ..... 34
Ilustración 24: Conteo de rayos por A) Temperatura, B) Sensación térmica, C) Punto de rocío, D) índice de calor ..... 36
Ilustración 25: Conteo de rayos por A) Humedad, B) Velocidad promedio del viento, C) Presión barométrica, D) Lluvia ..... 36
Ilustración 26: Conteo de rayos detectados y no detectados por campo eléctrico ..... 37
Ilustración 27: Selección de variables ..... 49

# II. Lista de Tablas

Tabla 1: Ubicación de Medidores de Campo Eléctrico y Estación Meteorológica ..... 2
Tabla 2: Revisión de trabajos relacionados ..... 9
Tabla 3: Clasificación de algoritmos de minería de datos ..... 17
Tabla 4: Conjuntos de datos ..... 21
Tabla 5: Variables del conjunto de “Datos-meteorológicos” ..... 23
Tabla 6: Variables del conjunto “Campo-eléctrico” ..... 24
Tabla 7: Variables del conjunto de “Rayos-detectados” ..... 24
Tabla 8: Variables obtenidas de preprocesar el conjunto “Datos-meteorológicos” ..... 26
Tabla 9: Variables obtenidas de procesar el conjunto “Campo-eléctrico” ..... 28
Tabla 10: Variables obtenidas de procesar el conjunto “Rayos-detectados” ..... 29
Tabla 11: Conjuntos finales obtenidos ..... 30
Tabla 12: Variables obtenidos del conjunto "Rayos-meteorológicos" ..... 32
Tabla 13: Variables obtenidos del conjunto "Rayos-campo-eléctrico" ..... 32
Tabla 14: Métricas obtenidas de cada variable en relación con los rayos del conjunto “Rayos-meteorológicos” ..... 35
Tabla 15: Resultados experimento 1 ..... 38
Tabla 16: Resultados experimento 2 ..... 39
Tabla 17: Resultados experimento 2 ..... 41
Tabla 18: Resultados Experimento 4 rango [0 - 9.5] ..... 42
Tabla 19: Resultados Experimento 4 rango [9.5 - 19] ..... 43
Tabla 20: Resultados Experimento 4 rango [19 - 28.5] ..... 44
Tabla 21: Resultados Experimento 4 rango [28.5 - 38] ..... 44
Tabla 22: Resultados Experimento 5 rango [0 - 9.5] ..... 45
Tabla 23: Resultados Experimento 5 rango [9.5 - 19] ..... 46
Tabla 24: Resultados Experimento 5 rango [19 - 28.5] ..... 46
Tabla 25: Resultados Experimento 5 rango [28.5 - 38] ..... 47
Tabla 26: Mejores reglas obtenidas de cada experimento ..... 48
Tabla 27: Precisión general de cada experimento ..... 48

# Capítulo 1

# 1. Introducción

## 1.1 Antecedentes del Proyecto

Una descarga atmosférica comúnmente llamada rayo, es el fenómeno ocurrido de cargas opuestas de nubes al suelo, suelo a las nubes o entre nubes, siendo la primera la más común con un 90% las cuales son comunes en todo el mundo. La magnitud de una descarga atmosférica puede provocar grandes daños, entre los que pueden destacarse:

- Accidentes laborales para los trabajadores.
- Paradas en líneas de producción o servicios.
- Daños en los equipos electrónicos.
- Daños en la infraestructura.
- Incendios
- Pérdida en la red de datos y comunicaciones.

Teniendo esto en cuenta, la comunidad científica se ha dado a la tarea de aplicar modelos de minería de datos a los datos meteorológicos recaudados a través de los años con el objetivo de generar predicciones para evitar desastres futuros. Con esto en mente se planteó que el objetivo del proyecto sea determinar que variables y relaciones entre estas influyen en la generación de una descarga eléctrica para las zonas del estado de Puebla.

## 1.2 Planteamiento del Problema

El Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE), se ha dado la tarea de colocar diversas estaciones climatológicas y medidores de campo eléctrico atmosférico, ubicados en diferentes zonas del Estado de Puebla, como se puede ver en la Tabla 1, las cuales recaban datos continuamente sobre los cambios climáticos de las zonas, que posteriormente pueden ser analizados para diferentes propósitos.

Gracias a la implementación de estaciones climatológicas y medidores de campo eléctrico los cuales recopilan información de su entorno, y mediante la combinación de estos dos conjuntos de datos generados, se obtiene nueva información que puede ser sometida a un análisis estadístico o diferentes modelos de minería de datos con los cuales se puede

determinar que variables o reglas contribuyen a la generación de una descarga eléctrica. Mediante estas variables o reglas se pueden tomar decisiones y acciones alrededor de estos eventos naturales como tormentas, rayos (descargas atmosféricas), inundaciones, huracanes etc.

Para este proyecto se utilizan los datos del medidor de campo eléctrico y estación meteorológica ubicados en Tonantzintla, Puebla. Una vez que se hace la correcta extracción y preprocesamiento de los datos, estos se utilizan para realizar predicciones, y estimación de tendencias de comportamiento de los campos y las descargas eléctricas atmosféricas que se producen, en relación con las condiciones climatológicas detectadas por la estación en ese momento. El número de estaciones se puede observar en la tabla 1, y el área de detección de rayos de todos los medidores de campo eléctrico se puede observar en la ilustración 1.

| **Ubicación** | **Equipos** |
|---|---|
| **INAOE, Sta. Ma. Tonantzintla** | Medidor de campo y Estación Meteorológica |
| **Gran Telescopio Milimétrico, Sierra Negra** | Medidor de campo y Estación Meteorológica |
| **BUAP CRC, San José Chiapa** | Medidor de campo y Estación Meteorológica |
| **UPAEP, Puebla** | Estación Meteorológica |
| **Instituto Tecnológico de Puebla, Puebla** | Estación Meteorológica |
| **Coatzacoalcos** | Medidor de campo y Estación Meteorológica |

***Tabla 1: Ubicación de Medidores de Campo Eléctrico y Estación Meteorológica***

Para este proyecto se utilizaron los datos del medidor del campo y estación meteorológica del INAOE ubicados en Sta. Ma. Tonantzintla.

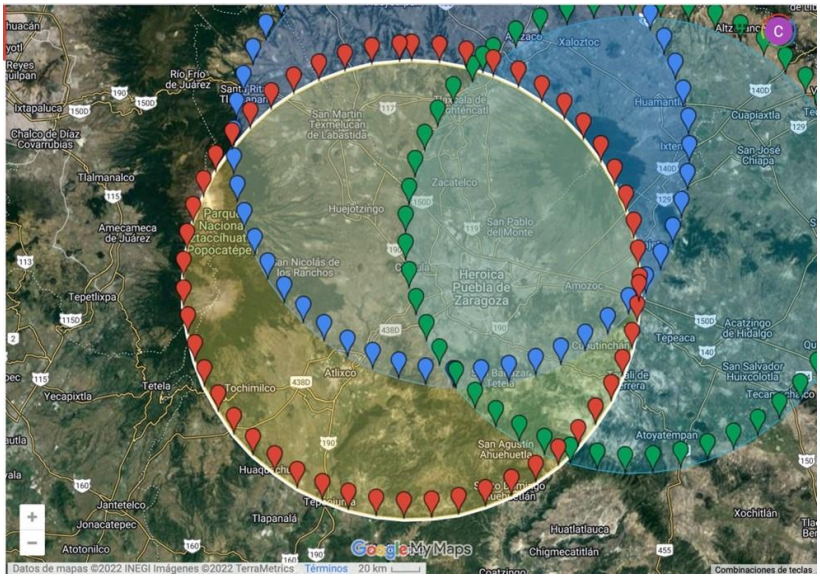


***Ilustración 1: Rango de detección de rayos de los medidores de campo eléctrico***

## 1.3 Objetivos

Los objetivos de este trabajo de tesis son los siguientes:

**Objetivo general**

Realizar el análisis de la información meteorológica y mediciones de campo eléctrico del estado de Puebla, utilizando técnicas computacionales para correlacionar las variables y estimar tendencias de comportamiento de los campos eléctricos y las descargas atmosféricas que estos producen.

**Objetivos específicos**

- Realizar la limpieza, procesamiento, carga y análisis, del conjunto de datos de la estación meteorológica y medidor de campo eléctrico.
- Aplicar métodos computacionales, para hacer un análisis preciso de los datos, con el objetivo de descubrir patrones y predecir tendencias de comportamiento.
- Aplicar el modelo de minería de datos y análisis de la información generada, con el objetivo cuando se dan ciertas condiciones climatológicas en predecir en base a los patrones arrojados por el algoritmo, cuando es más probable que suceda una descarga eléctrica atmosférica.

## 1.4 Pregunta de Investigación

¿Es posible establecer bajo qué correlación de variables meteorológicas y niveles de campo eléctrico puede suceder una descarga eléctrica?

## 1.5 Justificación

Con la aplicación de diferentes técnicas computacionales es posible que se puedan encontrar patrones que den información sobre las condiciones climatológicas comunes que originan una descarga eléctrica comúnmente llamada rayo, que es el fenómeno ocurrido de cargas opuestas de nubes al suelo, suelo a las nubes o entre nubes, siendo la primera la más común en todo el mundo, con un 90% de frecuencia. La magnitud de una descarga eléctrica puede provocar grandes daños, entre los que pueden destacarse:

- Accidentes laborales para los trabajadores.
- Paradas en líneas de producción o servicios.
- Daños en los equipos electrónicos.
- Daños en la infraestructura.
- Incendios
- Pérdida en la red de datos y comunicaciones.

Con el conocimiento previamente descubierto a través de este trabajo, se puede aportar información relevante, sobre las descargas eléctricas atmosféricas, la cual estará disponible para que las organizaciones puedan tomar sus propias decisiones sobre qué hacer en caso de una posible descarga, como destinar recursos a sistemas de prevención contra descargas eléctricas atmosféricas o minimizar los daños causados como la reparación de equipos y al mismo tiempo mantener la integridad física de los trabajadores, y minimizar o eliminar las pérdidas económicas causadas por las descargas atmosféricas.

## 1.6 Alcances

- Selección de las Técnicas o algoritmos más adecuados para el tratamiento de los datos meteorológicos.
- Definición de un modelo para el tratamiento de datos meteorológicos que permita hacer predicciones de descargas eléctricas atmosféricas en el estado de Puebla.
- La predicción de las descargas eléctricas atmosféricas en las zonas de Cholula a partir del conjunto de datos analizados.
- Artículo con los resultados de la investigación.

## 1.7 Impacto Socioeconómico

Con el conocimiento previamente descubierto, se pretende aportar información relevante, de cuando es más probable la generación de una descarga eléctrica, esta información estará disponible para que las organizaciones puedan tomar sus propias decisiones sobre qué hacer en caso de una posible descarga eléctrica. Como destinar recursos a sistemas de prevención o minimizar los daños causados como la reparación de equipos y al mismo tiempo mantener la integridad física de los trabajadores, y minimizar o eliminar las pérdidas económicas causadas por las descargas atmosféricas.

## 1.8 Organización de la tesis

La tesis se encuentra organizada de la siguiente forma:

En el capítulo II se lleva a cabo una revisión del estado del arte a través de la descripción de los trabajos existentes que sirven de base y referencia a este trabajo de investigación. En el capítulo III se presentan los fundamentos teóricos que dan soporte a este trabajo; se describe como se genera una descarga eléctrica, así como las herramientas utilizadas con las que se obtuvieron los datos y la metodología y modelos de minería de datos que se usaron para este proyecto. En el capítulo IV se describen las fases de la metodología desarrollada en este trabajo y los resultados obtenidos tras el procesamiento y análisis de los datos. Finalmente, en el capítulo V se detallan las conclusiones del proyecto y el trabajo futuro que se puede realizar a partir de este proyecto.

# Capítulo 2

# 2. Estado del arte

## 2.1. Métodos para la predicción del clima

Actualmente existen diferentes métodos para la predicción del clima. El método utilizado depende de la cantidad y tipo de información disponible, la dificultad que requiere la situación y la precisión con la que se requiera realizar el pronóstico (P.Kalaiyarasi & Mrs.A.Kalaiselvi, 2018).

A continuación, se mencionan 3 métodos para la predicción del clima.

1. **Predicción sinóptica del clima:** El centro meteorológico, proporciona una tabla sinóptica para cada día. Dentro de un tiempo específico, se observan diferentes parámetros climáticos. Se recopilan diferentes tipos de datos y el estudio de estos desde miles de estaciones meteorológicas
2. **Predicción numérica del clima:** La capacidad de la computadora para predecir el clima se conoce como predicción numérica del clima. Se utilizan complejos programas de cómputo, que procesan datos en supercomputadoras y proporcionan predicciones de las variables meteorológicas, como temperatura, viento, humedad, presión atmosférica y precipitación de la lluvia.
3. **Predicción estadística del clima:** Los métodos estadísticos puros se utilizan para predecir el clima, junto con los métodos numéricos. Utiliza registros pasados de parámetros climáticos para predecir la ocurrencia futura.

Para este trabajo se hizo uso de la predicción estadística del clima, debido a la precisión que otorga al usar registros históricos que permiten hacer una mejor predicción del clima junto con la predicción numérica la cual se beneficia de la utilización de programas de cómputo y las supercomputadoras las cuales se encargan de predecir los patrones o variables que pueden influir a futuro.

## 2.2. Trabajos existentes

Existen múltiples trabajos en la literatura que se han dedicado al estudio de las condiciones ambientales que originan una descarga eléctrica para tratar de predecir estos eventos. Los más relevantes para el objetivo de este trabajo, se presentan en la Tabla 2.

| Trabajo | Autores | Conjunto de datos | Técnicas usadas | Precisión |
|---|---|---|---|---|
| **Lightning prediction using satellite atmospheric sounding data and feed-forward artificial neural network.** | Elton Rafael Alves, Carlos Tavares da Costa Jr, Marcio Nirlando Gomes Lopes, Brígida Ramati Pereira da Rocha, Jose Alberto Silva de Sá. (2017) | Datos de sondeo del satélite NOAA-19. | Artificial Neural Network (ANN). | 95.6% |
| **Artificial Neural Network Model for the Prediction of Thunderstorms over Kolkata.** | Litta A. J, Sumam Mary Idicula, C. Naveen Francis. (2012) | Datos meteorológicos recopilados del departamento meteorológico indio (IMD). | Artificial Neural Network (ANN). | 76% |
| **Prediction of Severe Thunderstorms applying Neural Network using RSRW Data.** | Himadri Chakrabarty, Sonia Bhattacharya. (2014) | Datos meteorológicos recopilados del departamento meteorológico indio (IMD). | Artificial Neural Network (ANN-MLP). | 70% |
| **Application of K Nearest Neighbor Technique to Predict Severe Thunderstorms.** | Himadri Chakrabarty, Sonia Bhattacharya. (2015) | Datos meteorológicos recopilados del departamento meteorológico indio (IMD). | K-Nearest Neighbor (K-NN). | 88.21% |
| **Cumulonimbus prediction using artificial neural network back propagation with radiosonde indeces.** | Agie Wandala Putra, Chidchanok Lursinsap. (2014) | Datos obtenidos de una radiosonda, que derivo en 52 índices de inestabilidad atmosférica. | Back Propagation Neural Network (BPN). | 74.8% para BPN y 82.5% para BPN-PCA |

| | | | | |
|---|---|---|---|---|
| **An Efficient Approach towards Thunderstorm Detection Using Saliency Map.** | Rasika P. Gawande, Namrata D. Ghuse. (2012) | Imágenes satelitales obtenidas de National Oceanic and Atmospheric Administration (NOAA). | K- Means Clustering, Haar Wavelet Transformation. | 94% |
| **Application of Machine Learning Technique to Predict Severe Thunderstorms using upper air data.** | Dyuti Chatterjee, Himadri Chakrabarty. (2015) | Datos meteorológicos recopilados del departamento meteorológico indio (IMD). | K-Nearest Neighbor (K-NN). | 72.16% |
| **Lightning Forecasting Using ANN-BP & Radiosonde.** | Leong Yeng Weng, Jamaludin Bin Omar, Yap Keem Siah, Syed Khaleel Ahmed, Izham Bin Zainal Abidin, Noradlina Abdullah. (2010) | Lecturas obtenidas de una radiosonda del Departamento Meteorológico de Malasia. | Artificial Neural Network (ANN-BP). | 67% |
| **A Modular Neural Network Approach For Locating Cloud-To-Ground Lightning Strokes.** | S. A. Emamghoreishi, R. Moini, S. H. H. Sadeghi. (2002) | Campo electromagnético. | Artificial Neural Network (ANN-MLP). | El error relativo es inferior al 5% |
| **Detection of Lightning Pattern Changes UsingMachine Learning Algorithms.** | Aimée Booysens, Serestina Viriri. (2014) | Datos obtenidos del Lightning Imaging Sensor (LIS) abordo del Tropical Rainfall Measuring Mission (TRMM). | Naıve Bayes, Decision Trees, K-Means Clustering, Support Vector Machines (SVM). | 63% para conjuntos de datos grandes y 73% para conjuntos de datos pequeños |
| **A lightning motion prediction technology based on spatial clustering method.** | Guo Juntian, Gu ShanQiang, Feng Wanxing. (2011) | Datos obtenidos de la base de datos de Lightning Locating System of SGEPRI (State Grid Electric. | Spatial Clustering. | 75% |

| | | | | |
|---|---|---|---|---|
| | | Power Research Institute) en China. | | |
| **Artificial Neural Network Based Technique for Lightning Prediction.** | Dalina Johari, Titik Khawa Abdul Rahman, Ismail Musirin. (2008) | Datos obtenidos de Malaysian Meteorological Services (MMS) y Tenaga.<br><br>Nasional Berhad (TNB) Research and Development. | Artificial Neural Network (ANN-BP). | Rms error 0.41 con un valor de R de 0.99997 |
| **Application Study of Machine Learning in Lightning Forecasting.** | Taorong Qiu, Shanshan Zhang, Hou Zhou, Xiaoming Bai, Ping Liu. (2013) | Datos obtenidos de Jiangxi.<br><br>Meteorological.<br><br>Administration. | Support Vector.<br><br>Machine (SVM), Rough Set (RS) | 71% |

***Tabla 2: Revisión de trabajos relacionados***

Como se puede observar en la tabla 2 en la columna conjuntos de datos, todos los datos se extrajeron de estaciones o centros meteorológicos externos. Antes de que se aplicaran directamente los modelos de minería de datos se realizó un proceso anterior que es la limpieza y preprocesamiento de los datos ya que por lo general los datos que se extraen de estos centros y estaciones meteorológicas viene con ruido en los datos. Se tomo de ejemplo el trabajo de (Agie Wandala Putra, CUMULONIMBUS PREDICTION USING ARTIFICIAL NEURAL, 2014) en los cuales se utilizaron técnicas como limpieza, selección, normalización, discretización y reducción de dimensionalidad algunos de los campos que se utilizaron fueron “Temperatura”, “Humedad”, “Viento”, y “Punto de roció”. Una vez que se tenían los datos en un formato adecuado se procedía a la aplicación del modelo de minería de datos, teniendo como entrada el conjunto de datos limpios y preprocesados, una vez que se aplicaba el modelo al conjunto de datos se obtenía como salida el conjunto de las reglas de asociación, variables, o datos más relevantes que el algoritmo arrojó. Una vez que se obtenían los resultados se comparaba la precisión que tuvo el modelo contra otros modelos o los mismos resultados de aplicar el algoritmo varias veces, todo con el objetivo de determinar qué modelo o iteración de este arrojaba mejores predicciones o resultados.

# Capítulo 3

# 3. Marco teórico

## 3.1. Campo Eléctrico

Es un área en el espacio que rodea a una carga o cargas eléctricas (positiva o negativa). Un campo eléctrico puede generarse alrededor de una sola carga o de varias cargas generando un campo resultante como se muestra en la ilustración 1.

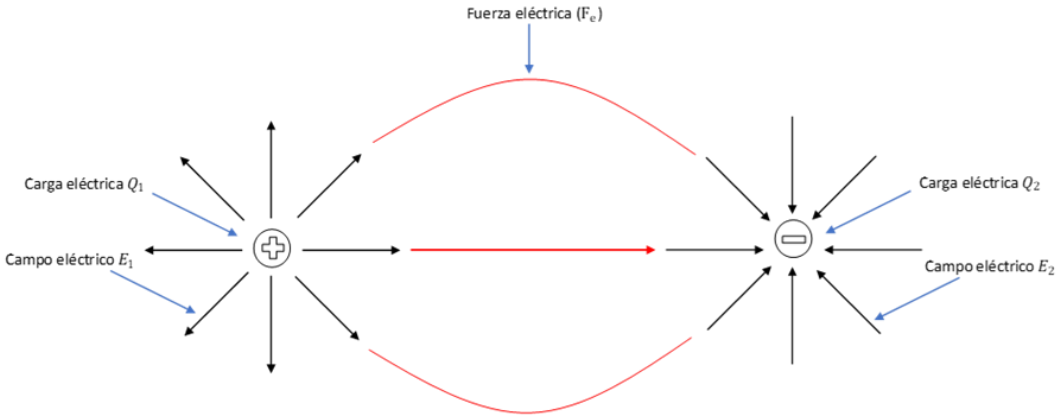


***Ilustración 2: Comportamiento del campo eléctrico***

En una definición formal, el campo eléctrico (E) se define como la cantidad de fuerza eléctrica *($F_e$)* por carga ejercida (Q) en un punto en el espacio.

Fórmula para calcular el campo eléctrico $E = \frac{F_e}{Q}$

El campo eléctrico generado por las nubes de tormenta generalmente se mide en volts por metros (E = Volts/metro o también como Newtons/Coulomb). Los campos eléctricos que se generan en tormentas eléctricas normalmente se miden en miles de volts por metro (E = kV/m).

## 3.2. Descarga Atmosférica

Las descargas atmosféricas comúnmente llamados rayos son descargas eléctricas que se producen en la atmosfera de la tierra. El fenómeno se da debido a la acumulación de cargas eléctricas opuestas entre las nubes, el suelo o el aire. El fenómeno ocurre cuando la diferencia de cargas positiva y negativa son lo suficientemente fuerte en el campo eléctrico, entonces la descarga ocurre, la electricidad fluye y se libera, igualando momentáneamente las cargas (Re, 2019).

Generalmente las descargas eléctricas atmosféricas ocurren de las nubes al suelo en terreno plano, pero también pueden ocurrir de nubes a nubes, suelo a las nubes (esto suele ocurrir en montañas o estructuras altas). El 90% de los rayos que se detectan son de nube negativa hacia suelo positivo como se observa en las ilustraciones 2-7. Esto se debe a que bajo una nube de tormenta el campo eléctrico de la superficie se invierte (SIERRA TAMARA & ESCOBAR MORALES, 2004).

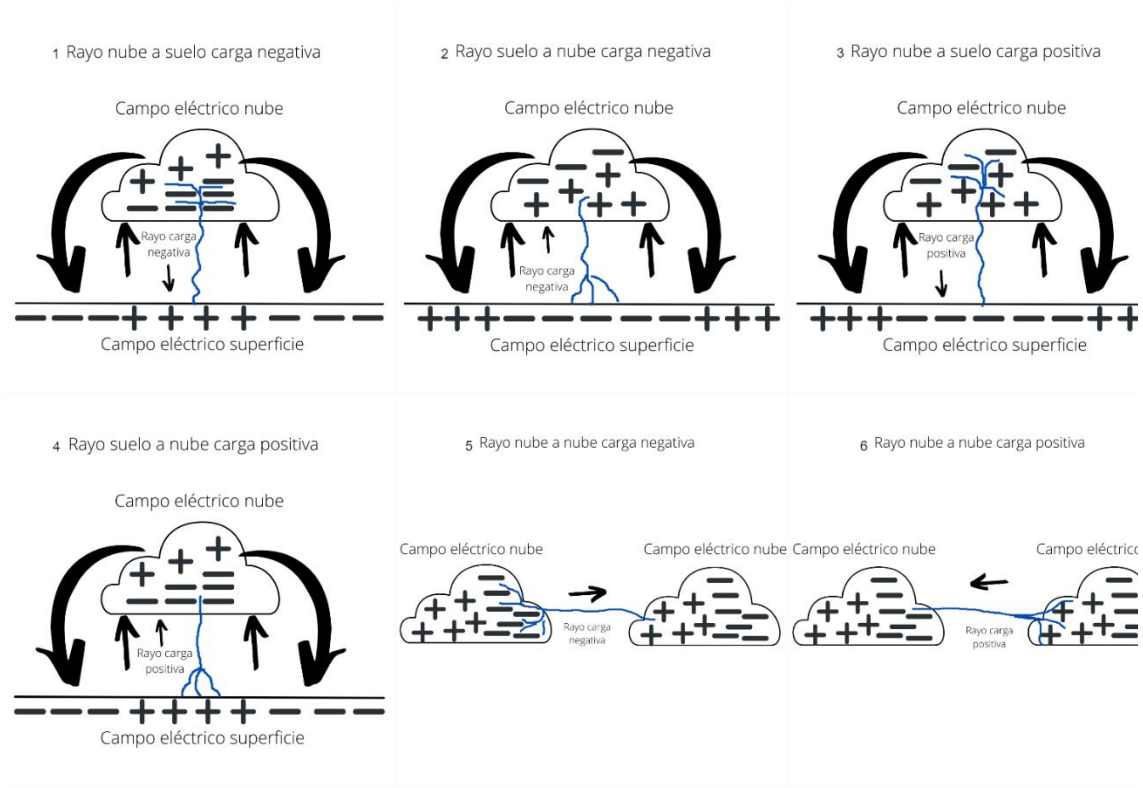


***Ilustración 3: Tipos de Descargas Eléctricas Atmosféricas***

Las descargas eléctricas atmosféricas (rayos) se asocian comúnmente a las tormentas, lluvias o vientos intensos, sin embargo, estos se pueden generar también durante condiciones de tormentas de nieve, arena, erupciones volcánicas, huracanes o incendios forestales.

El aumento de temperatura del aire a lo largo del recorrido de la descarga eléctrica (también llamado canal de descarga o rutina principal de paso) permite que el aire se expanda rápidamente, generando ondas sonoras asociadas a truenos, que se suelen confundir como rayos.

## 3.3. Daños ocasionados por descargas eléctricas atmosféricas

Las descargas eléctricas atmosféricas pueden llegar a ocasionar daños económicos y sociales a los sistemas de telecomunicaciones, distribución de energía y vuelos generados por rayos directos e indirectos que conducen a pérdidas monetarias. La mayoría de estos desastres son causados por descargas eléctricas atmosféricas de nube a suelo (Rafael Alves, Tavares da Costa, Gomes Lopes, Pereira da Rochaa, & Silva de Sá, 2017).

Ejemplos de desastres causados por las descargas atmosféricas:

- Accidentes laborales para los trabajadores.
- Paradas en líneas de producción o servicios.
- Daños en los equipos electrónicos.
- Daños en la infraestructura.
- Incendios.
- Pérdida en la transferencia de datos y comunicaciones.

## 3.4. Monitor de Campo Eléctrico Atmosférico EFM-100

### 3.4.1. Funcionamiento del EFM 100

El molino de campo eléctrico EFM-100, detecta el campo eléctrico mediante la exposición repetida y el blindaje de una serie de electrodos de detección. Un molino de campo eléctrico utiliza un rotor mecánico para proteger y exponer alternativamente varias placas de detección a un campo eléctrico. Cuando las placas de detección se exponen al campo eléctrico, una carga eléctrica se extrae del suelo a las placas a través de una resistencia de detección. Cuando las placas de detección están protegidas del campo, la carga vuelve a tierra, nuevamente través de la resistencia de detección. Esta carga móvil es una corriente eléctrica que se mide como un voltaje de CA (corriente alterna) a través de la resistencia de detección. El tamaño del voltaje es proporcional al tamaño del campo eléctrico aplicado a las placas. (BOLTEK CORPORATION, 2016).

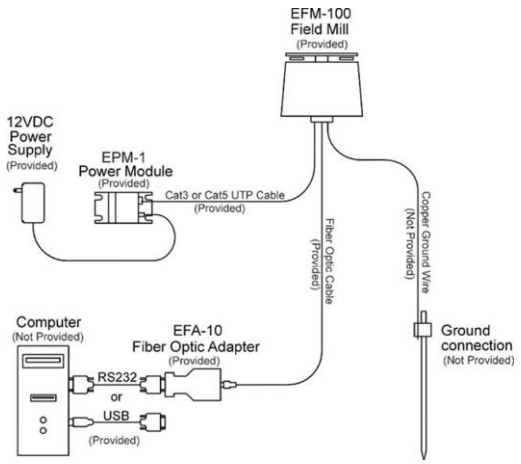


***Ilustración 4: Monitor de campo eléctrico atmosférico EFM-100. Tomada de (BOLTEK CORPORATION, 2016)***

## 3.5. Detección de nube de tormenta

1. A menudo, la primera indicación de una nube de tormenta que se aproxima es una lectura de campo positivo seguida de una inversión de polaridad del campo a negativo a medida que la nube se mueve por encima.
2. Con la nube de tormenta directamente encima, la polaridad del campo se invierte a negativo.
3. Una vez que la nube ha pasado sobre el campo a menudo se invierte de nuevo a positivo antes de decaer a una lectura normal de campo eléctrico de tiempo justo de aproximadamente 0,1 kV / m. Esta inversión de campo al final de la tormenta se ha referido como el "Final de la Oscilación de tormenta" (End of Storm Oscillation, ESO).

Este comportamiento se puede observar en la ilustración 4.

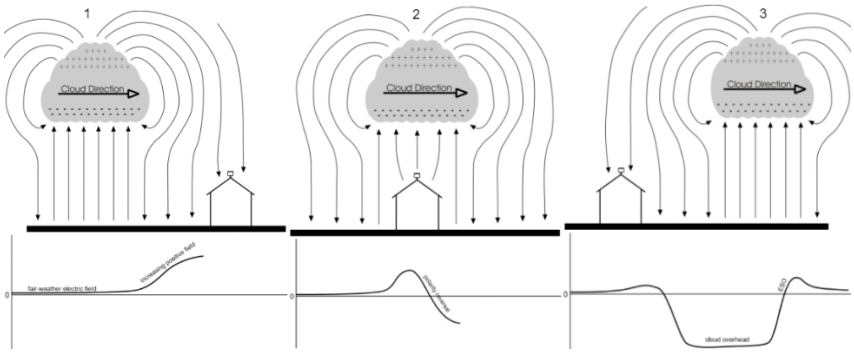


***Ilustración 5: Detección de una nube de tormenta. Tomada de (BOLTEK CORPORATION, 2016)***

## 3.6. Detección de rayos

Los cambios escalonados en la magnitud del campo eléctrico indican un rayo como se observa en la ilustración 5. Esto es debido a que un rayo cercano crea picos más escalonados en la lectura del campo eléctrico, un rayo cercano produce un mayor cambio en el campo eléctrico que un rayo lejano (BOLTEK CORPORATION, 2016). El medidor de campo eléctrico puede detectar un Rayo a 24 millas (38.6243 Km) de distancia.

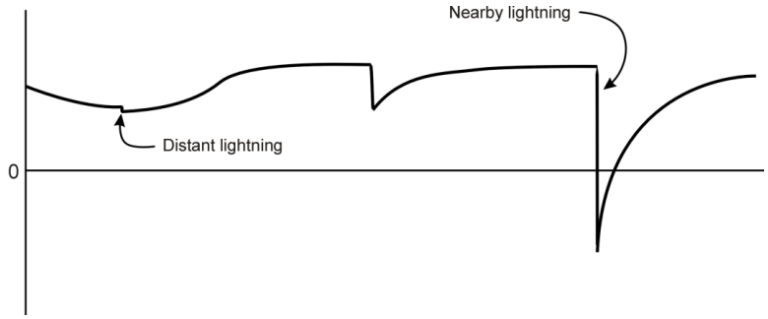


***Ilustración 6: Comportamiento del campo eléctrico ante una descarga atmosférica. Tomada de*** *(BOLTEK CORPORATION, 2016)*

Toda la información recopilada por el monitor de campo eléctrico se almacena en el mismo dispositivo, la información que genera se guarda en 2 archivos diferentes.

El primer archivo se guarda con el nombre EFMEvents.log el cual guarda los registros de los eventos del dispositivo, así como los rayos que se detectaron y la distancia de estos. El dispositivo hace un registro con la fecha, hora y el evento que registro un ejemplo sería el siguiente “02/29/2016 22:08:46 INAOE parque Lightning Detected at 35 km”.

El segundo archivo se guarda con el nombre de la ubicación de la estación y la fecha en la que se creó ese archivo y con la extensión efm ejemplo “INAOE parque-01012017.efm”. Este archivo contiene los registros del nivel del campo eléctrico de la siguiente manera “12:23:12+0.13,0”, cada campo está separado por una coma. El primer campo es la hora, el segundo el nivel y la polaridad del campo y el ultimo indica el estado del rotor, cero indica que no hubo fallo y uno que si hubo.

## 3.7. Vantage VUE Weather Station

### 3.7.1. Funcionamiento de la Vantage VUE Weather Station

El conjunto de sensores integrados (ISS) inalámbrico Vantage Vue recopila datos meteorológicos externos y los envía de forma inalámbrica a una consola a través de una radio de baja potencia. La ISS funciona con energía solar e incluye una batería de respaldo. El Vantage Vue ISS contiene un colector de lluvia, un sensor de temperatura / humedad, un anemómetro y una veleta. El sensor de temperatura / humedad está montado en un escudo de radiación pasivo para minimizar el impacto de la radiación solar en las lecturas del sensor. El anemómetro mide la velocidad del viento y la veleta mide la dirección del viento. El Módulo de interfaz de sensor (SIM) está alojado dentro de la ISS y comprende el "cerebro" del sistema Vantage Vue y el transmisor de radio. La tarjeta SIM recopila datos meteorológicos externos de los sensores de la ISS y los transmite a su consola Vantage Vue (Davis Instruments, 2011). La ilustración 6 indica las partes que componen a la estación meteorológica

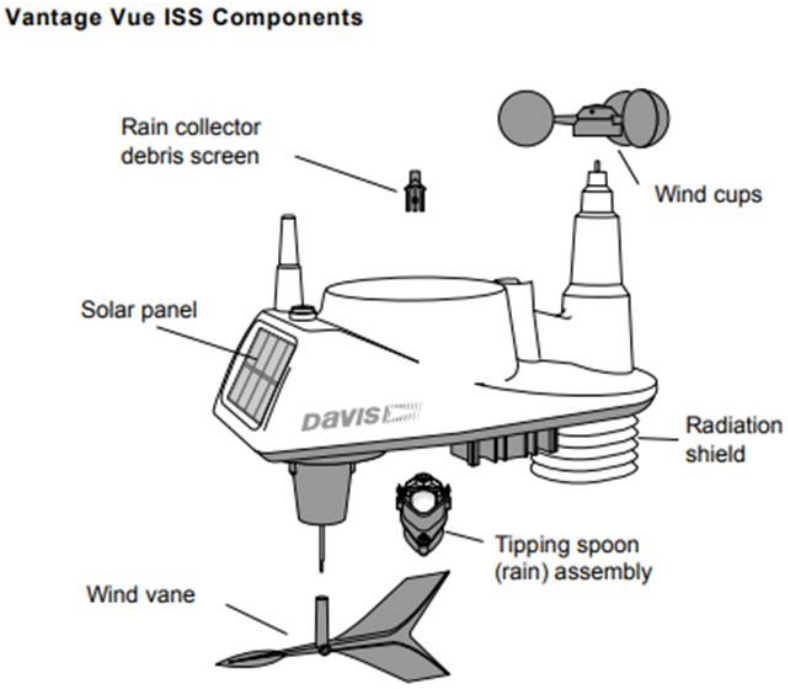


***Ilustración 7: Estación meteorológica Vantage VUE. Tomada de (Davis Instruments, 2011)***

La estación meteorológica guarda los datos en un archivo con extensión .csv (comma-separated values) el cual contiene diversos valores meteorológicos, los registros contienen

la fecha, hora y los campos separados por punto y coma, los valores en decimal se ponen con una coma en lugar de un punto como se muestra en la ilustración 7:

```
Fecha (America/Mexico_City);Tempin (ºC);Temp (ºC);Chill (ºC);Dewin (ºC);Dew (ºC);Heatin (ºC);Heat (ºC);Humin (%);Hum (%);Wspdhi (km/h);Wspdavg (km/h);Wdiravg (º);Bar (mmHg);Rain (mm);Rainrate (mm/h);
2019-01-29 14:00:00;26,7;20,8;20,6;2,3;2,2;4,9;18,9;29;29;22,7;9,7;145;763,9;0;0;
```

***Ilustración 8: Registro capturado por la estación meteorológica***

## 3.8. Minería de datos

### 3.8.1. Definición de minería de datos

Esta disciplina se puede definir como el descubrimiento de patrones y conocimiento a través de grandes cantidades de datos, almacenados en bases de datos, data warehouses, o cualquier otro medio de almacenamiento de información (Coronado Arjona, Bianchi Rosado, Vivas Burgos, & Perera Collí, 2017).

Antes de la aplicación del algoritmo de minería se realiza un proceso previo conocido como ETL (Extract, Transform and Load). El proceso consiste en extraer los datos de una o varias fuentes, limpiarlos y/o reformatearlos y cargarlos en otra base de datos o herramienta, esto se hace con el objetivo de preparar los datos de manera más homogénea para su análisis.

### 3.8.2. Técnicas de minería de datos.

Las técnicas de minería de datos se pueden clasificar en dos tipos: supervisados y no supervisados (Coronado Arjona, Bianchi Rosado, Vivas Burgos, & Perera Collí, 2017).

- **Algoritmos supervisados:** Los algoritmos supervisados, predicen un dato desconocido a partir de un conjunto de datos conocido. A partir de datos con una etiqueta conocida se induce un modelo que relaciona dicha etiqueta con los atributos descriptivos. Los algoritmos supervisados requieren de una fase de entrenamiento (construcción de un modelo usando un subconjunto de datos con etiqueta conocida) y prueba (prueba del modelo sobre el resto de los datos).

- **Algoritmos no supervisados:** Los algoritmos no supervisados se caracterizan por descubrir modelos o características significativas a partir de los datos de entrada. Estos algoritmos realizan tareas descriptivas como el descubrimiento de patrones y

tendencias en los datos actuales. El descubrimiento de estos patrones sirve para llevar a cabo acciones y obtener un beneficio científico o de negocio de ellas.

Los algoritmos se clasifican tal y como se muestran en la Tabla 3.

| **Supervisados** | **No Supervisados** |
|---|---|
| Arboles de decisión | Reglas de asociación |
| Redes neuronales | K-means |
| K-NN (k Nearest Neighbour) | Clustering |
| Clasificación de Naive Bayes | Segmentación |
| SVM (Support Vector Machines) | Detección de anomalías |
| Regresión lineal | Análisis de componentes principales |

***Tabla 3: Clasificación de algoritmos de minería de datos***

### 3.8.3. Arboles de decisión

Un árbol de decisión es un conjunto de condiciones o reglas organizadas en una estructura jerárquica, de tal manera que la decisión final se puede determinar siguiendo las condiciones que se cumplen desde la raíz hasta alguna de sus hojas (VIZCAINO GARZON, 2008). Tiene una estructura similar a un diagrama de flujo, donde un nodo interno representa una característica (variable) de un conjunto de ejemplos, la rama representa una regla (valor) que puede tomar esa característica y el nodo hoja (clase a predecir) representa el resultado.

### 3.8.4. Estructura básica de un árbol de decisión

**Nodo raíz (nodo de decisión superior):** Representa a toda la población o muestra y esto se divide en dos o más conjuntos homogéneos.

**Ramas:** Brindan los posibles caminos (valores) que se tienen de acuerdo con la decisión tomada.

**Nodos internos:** Tras la primera división se encuentran estos nodos, que vuelven a dividir el conjunto de ejemplos en función de las clases (variables).

**Nodos hoja:** Nodos sin división que se ubican en la parte inferior del esquema y representan el valor de la clase que devolverá el árbol.

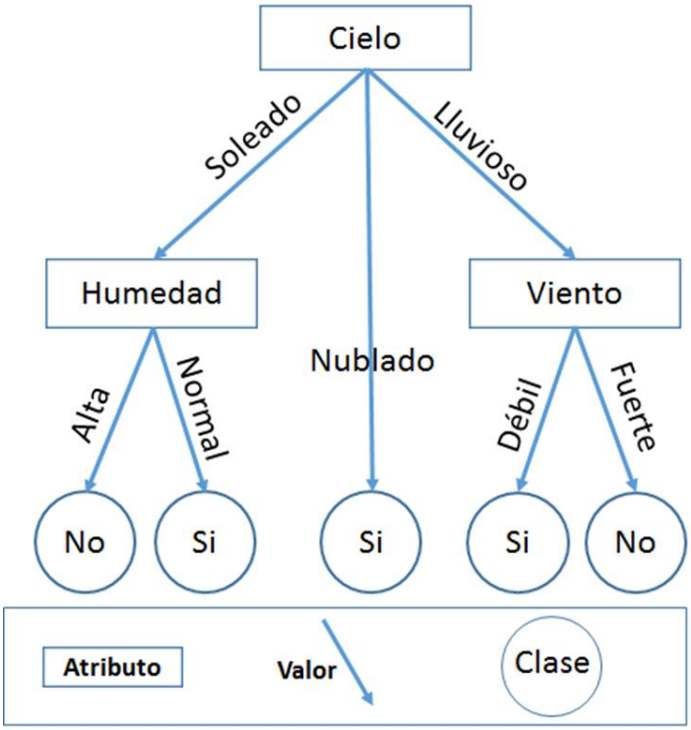


***Ilustración 9: Estructura de un árbol de decisión***

- Cada nodo que no sea una hoja representa un variable.
- Las ramas que partan del nodo etiquetado con la variable están etiquetadas con cada uno de los posibles valores de la variable.
- Cada hoja corresponde a un valor de la clasificación.

## 3.9. Proceso KDD

El proceso KDD (Knowledge Discovery in Databases) es un proceso en el cual se analizan las reglas o patrones, en base a la extracción de conocimiento. Este proceso puede ser iterativo y se puede aplicar las veces que sea necesario para obtener la información necesaria.

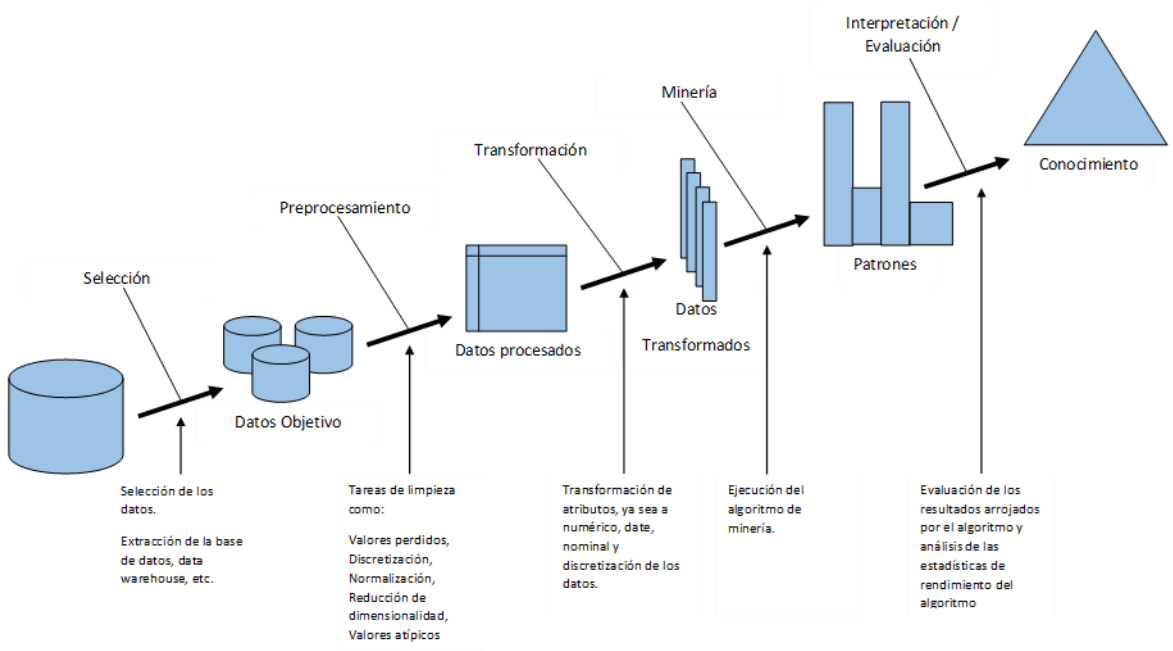


***Ilustración 10: Proceso KDD***

1. **Selección:** Tarea cuya función es la de recopilar los datos más relevantes del repositorio seleccionado.
2. **Preprocesamiento de datos:** Fase en la cual se tratan con los datos y se aplican tareas de limpieza, normalización, discretización, reducción de dimensionalidad, etc.
3. **Transformación de datos:** Fase en la cual los datos son transformados a un formato apropiado nominal, numérico, date, etc. Y a un formato de archivo especifico como CSV o ARFF, para la herramienta que ejecutara el modelo de minería
4. **Minería de datos:** En esta fase se entrenan el modelo de minería con el conjunto previamente preprocesado y transformado. Para evaluar el desempeño de los algoritmos se utilizan algunos métodos como validación cruzada (cross validation) o división de datos (data splitting o train-test split), para posteriormente compara las

métricas arrojadas, por todos los algoritmos, como la precisión, error cuadrático medio, coeficiente de correlación, matriz de confusión etc.

5. **Interpretación/Evaluación:** Para esta fase las métricas arrojadas por los diferentes algoritmos se comparan para determinar que algoritmo arroja los datos más precisos, dependiendo de la variable a evaluar/predecir y se selecciona el que haya realizado el mejor análisis estadístico de los datos para poder evaluar las reglas que arrojo y poder tomar una decisión.

Este capítulo sirvió como introducción para entender los conceptos que se usaron para este trabajo como que es una descarga atmosférica como se genera esta, también se describe como funciona el algoritmo que se usó para el proceso de minería de datos, así como el proceso KDD el cual se implementó para llevar acabo la parte práctica de este trabajo de investigación.

# Capítulo 4

# 4. Marco metodológico

Los resultados que se muestran a continuación son el resultado del conjunto de técnicas que se aplicaron durante el proceso KDD, así como los resultados obtenidos durante cada etapa.

## 4.1. Selección

Los conjuntos de datos con los que se trabajaron se extrajeron de la estación meteorológica y el del medidor del campo eléctrico los cuales fueron proporcionados por el Instituto Nacional de Astrofísica, Óptica y Electrónica.

Los conjuntos proporcionados fueron los siguientes.

| Conjunto | Extensión | Codificación | Archivos | Descripción | Intervalo |
|---|---|---|---|---|---|
| **Datos-meteorológicos** | CSV | UTF-16 LE | 10 | Registros de los datos meteorológicos | 2019 |
| **Campo-eléctrico** | EFM | UTF-8 | 188 | Registros de las mediciones de campo eléctrico | 2016 a 2019 |
| **Rayos-detectados** | LOG | ANSI | 1 | Registros de eventos del medidor de campo eléctrico | 2016 a 2019 |

***Tabla 4: Conjuntos de datos***

A continuación, se muestra en las ilustraciones 11 y 12 una sección de cada conjunto obtenido.

| Nombre | Fecha de modificación | Tipo | Tamaño |
|---|---|---|---|
| Weathercloud Pque Tecnolgico INAOE 2019-01.csv | 01/11/2021 08:54 p. m. | Archivo CSV | 48 KB |
| Weathercloud Pque Tecnolgico INAOE 2019-02.csv | 01/11/2021 10:41 p. m. | Archivo CSV | 425 KB |
| Weathercloud Pque Tecnolgico INAOE 2019-03.csv | 04/10/2019 01:19 p. m. | Archivo CSV | 429 KB |

***Ilustración 11: Captura del conjunto "Datos-meteorológicos"***

| Nombre | Fecha de modificación | Tipo | Tamaño |
|---|---|---|---|
| EFMEvents.log | 07/12/2021 09:36 p. m. | Documento de te... | 16,807 KB |
| INAOE parque-01012017.efm | 07/12/2021 09:41 p. m. | Archivo EFM | 30,878 KB |
| INAOE parque-01022017.efm | 07/12/2021 09:41 p. m. | Archivo EFM | 30,883 KB |

***Ilustración 12: Conjuntos "Rayo detectados" y "Campo eléctrico"***

Como se puede observar en la ilustración 12 los conjuntos ***Rayo-detectados y Campo-eléctrico*** se encuentran en la misma carpeta ya que ambos conjuntos son proporcionados por el medidor de campo eléctrico.

## 4.2 Diccionario de datos

También se identificaron las variables que posee cada conjunto como los que se describen en las tablas 5, 6, 7. Lo cual permitió generar un diccionario de datos para cada conjunto.

| **Conjunto: Datos meteorológicos** | | | **Numero de registros: 18,718** | |
|---|---|---|---|---|
| **Atributo** | **Tipo** | **Valores/Clases** | **Unidad de medida** | **Descripción** |
| **Fecha** | Date | En-29, 2019 a Oct 4, 2019 | Time-stamp | Fecha y hora en la que se registró el evento |
| **Tempin** | Numérico | 0º a 60º | ºC (grados Celsius) | Temperatura interna de la estación meteorológica |
| **Temp** | Numérico | -40º a 65º | ºC (grados Celsius) | Temperatura externa |
| **Chill** | Numérico | -79º a 57º | ºC (grados Celsius) | Sensación térmica |
| **Dewin** | Numérico | -76º a 54º | ºC (grados Celsius) | Punto de rocío interno de la estación meteorológica |
| **Dew** | Numérico | -76º a 54º | ºC (grados Celsius) | Punto de rocío externo |
| **Heatin** | Numérico | -40º a 74º | ºC (grados Celsius) | Índice de calor interno de la estación meteorológica |
| **Heat** | Numérico | -40º a 74º | ºC (grados Celsius) | Índice de calor externo |

| | | | | |
|---|---|---|---|---|
| **Humin** | Numérico | 0% a 100% | % (unidad de medida en porcentaje) | Humedad interna de la estación meteorológica |
| **Hum** | Numérico | 0% a 100% | % (unidad de medida en porcentaje) | Humedad exterior |
| **Wspdhi** | Numérico | 3 a 100 km/h | km/h (kilómetros/hora) | Velocidad del viento |
| **Wspdavg** | Numérico | 3 a 100 km/h | km/h (kilómetros/hora) | Velocidad media del viento |
| **Wdiravg** | Numérico | 0º a 360º | Grados | Dirección del viento |
| **Bar** | Numérico | 410 a 820 | mmHg (milímetros de mercurio) | Presión barométrica |
| **Rain** | Numérico | 0 a 199.99 | mm (milímetro de agua caída) | Lluvia |
| **Rainrate** | Numérico | 0.04 a 40 | mm/hr (milímetro hora de agua caída) | Promedio de la lluvia |

***Tabla 5: Variables del conjunto de "Datos-meteorológicos"***

| **Conjunto: Campos eléctricos** | | | **Numero de Registros: 115,551,015** | |
|---|---|---|---|---|
| **Atributo** | **Tipo** | **Valores/Clases** | **Unidad de medida** | **Descripción** |
| **Hora** | Time | Feb 29, 2016 a Nov 12, 2019 | Time stamp | hora en la que se registró el evento |

| | | | | |
|---|---|---|---|---|
| **Nivel del Campo Eléctrico** | Numérico | -20 a 20 kv/m | kv/m (kilovolts/metro) | Nivel y polaridad del campo |
| **Fallo del Molino/Rotor** | Binario | 0 a 1 | | Indica con el valor de 0 si funciona correctamente y 1 si falló el molino del medidor de campo eléctrico, |

***Tabla 6: Variables del conjunto "Campo-eléctrico"***

| **Conjunto: Rayos detectados** | | | **Numero de registros: 249,486** | |
|---|---|---|---|---|
| **Dato** | **Tipo** | **Valores/Clases** | **Unidad de medida** | **Descripción** |
| **Fecha** | Date | Feb 29, 2016 a Oct 4 2019 | Time-stamp | Fecha y hora en la que se registró el evento |
| **Rayo Detectado** | Nominal categórico | Lightning Detected | | Si se detectó un rayo |
| **Distancia** | Numérico | 0 a 38 km | km (kilómetros) | Distancia a la que cayó el rayo |

***Tabla 7: Variables del conjunto de "Rayos-detectados"***

## 4.2. Preprocesamiento

El preprocesamiento de datos se realizó con un programa hecho con el lenguaje GO, se realizó en este lenguaje ya que permite hacer uso de la concurrencia y el uso de todos los núcleos del procesador de forma sencilla, lo cual permite que el preprocesamiento de los datos sea mucho más rápido y eficiente.

En la ilustración 13 se puede observar el proceso que se llevó a cabo para obtener los conjuntos finales con los cuales se trabajó la parte de minería de datos.

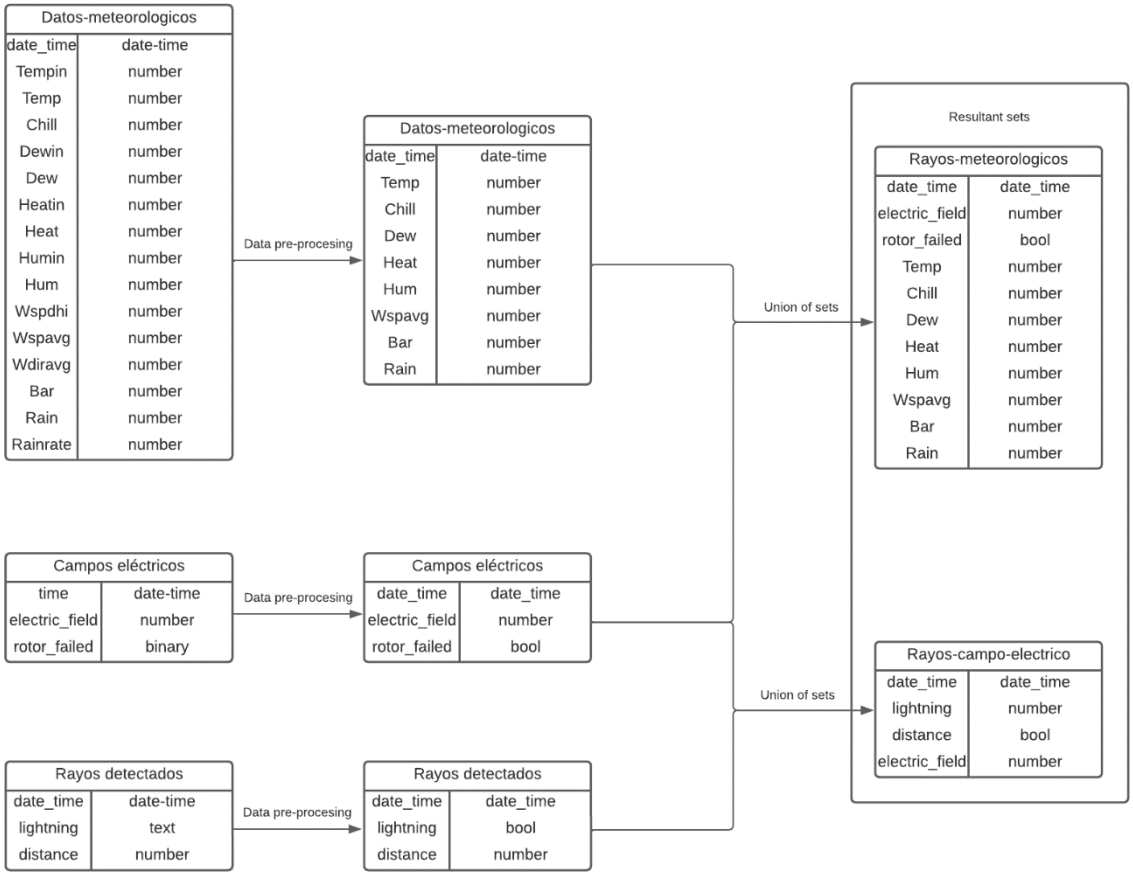


***Ilustración 13: Proceso con el que se obtuvieron los conjuntos finales***

### 4.2.1. Preprocesamiento del conjunto datos meteorológicos

El conjunto **Datos-meteorológicos** están conformado por 10 archivos CSV cada uno contienen 16 variables. Los registros se generan cada 10 minutos, como se observa en la ilustración 14.

```
Fecha (America/Mexico_City);Tempin (°C);Temp (°C);Chill (°C);Dewin (°C);Dew (°C);Heatin (°C);Heat (°C);Humin (%);Hum (%);Wspdhi (km/h);Wspdavg (km/h);Wdiravg (°);Bar (mmHg);Rain (mm);Rainrate (mm/h);
2019-01-29 13:50:00;26,5;20,7;20,6;;;;;34;;;;63;763,9;0;0;
2019-01-29 14:00:00;26,7;20,8;20,6;;2,2;;18,9;29;29;22,7;9,7;145;763,9;0;0;
2019-01-29 14:10:00;26,7;20,6;20,6;;1,7;;18,3;29;29;22,7;13;78;763,8;0;0;
2019-01-29 14:20:00;26,7;20,3;20,6;;2,2;;18,3;28;30;27,4;13;76;763,9;0;0;
2019-01-29 14:30:00;26,8;20,6;20;;3,3;;18,9;29;32;22,7;9,7;51;763,6;0;0;
2019-01-29 14:40:00;26,7;21,2;21,1;;4,4;;18,9;28;33;20,9;11,2;194;763,6;0;0;
2019-01-29 14:50:00;25,5;21,3;18,9;;3,9;;19,4;30;32;25,9;11,2;161;763,6;0;0;
2019-01-29 15:00:00;25,9;21,3;21,1;;3,3;;19,4;29;31;28,8;14,4;160;763,5;0;0;
2019-01-29 15:10:00;26,3;21,3;21,1;;3,3;;19,4;30;31;19,4;11,2;175;763,3;0;0;
2019-01-29 15:20:00;26,6;21,6;20,6;;3,9;;19,4;28;31;24,1;11,2;151;763,5;0;0;
```

***Ilustración 14: 10 primeras líneas de un archivo CSV del conjunto Datos-meteorológicos***

Para facilitar la manipulación del conjunto se unieron los 10 archivos en uno solo archivo CSV, se cambiaron los nombres de las cabeceras y se eliminaron los valores vacíos. Se hizo una reducción de dimensionalidad, la cual consiste en seleccionar las variables

óptimas para la variable a predecir. La selección se hizo con la asesoría del especialista encargado de las estaciones meteorológicas y medidores de campo eléctrico. Las variables seleccionadas se pueden observar en la ilustración 15 y la definición de cada una en la tabla 8.

```
date,temp,chill,dew,heat,hum,wspdavg,bar,rain
2019-01-29 14:00:00,20.8,20.6,2.2,18.9,29.0,9.7,763.9,0.0
2019-01-29 14:10:00,20.6,20.6,1.7,18.3,29.0,13.0,763.8,0.0
2019-01-29 14:20:00,20.3,20.6,2.2,18.3,30.0,13.0,763.9,0.0
2019-01-29 14:30:00,20.6,20.0,3.3,18.9,32.0,9.7,763.6,0.0
2019-01-29 14:40:00,21.2,21.1,4.4,18.9,33.0,11.2,763.6,0.0
2019-01-29 14:50:00,21.3,18.9,3.9,19.4,32.0,11.2,763.6,0.0
2019-01-29 15:00:00,21.3,21.1,3.3,19.4,31.0,14.4,763.5,0.0
2019-01-29 15:10:00,21.3,21.1,3.3,19.4,31.0,11.2,763.3,0.0
2019-01-29 15:20:00,21.6,20.6,3.9,19.4,31.0,11.2,763.5,0.0
2019-01-29 15:30:00,21.1,20.6,3.3,18.9,31.0,13.0,763.5,0.0
```

***Ilustración 15: Conjunto "Datos-meteorológicos" preprocesado***

| Atributo | Descripción |
|---|---|
| **date** | Fecha y hora en la que se registró el evento |
| **temp** | Temperatura externa |
| **chill** | Sensación térmica |
| **dew** | Punto de roció externo |
| **heat** | Índice de calor externo |
| **hum** | Humedad exterior |
| **wspdavg** | Velocidad media del viento |
| **bar** | Presión barométrica |
| **rain** | Lluvia |

***Tabla 8: Variables obtenidas de preprocesar el conjunto "Datos-meteorológicos"***

### 4.2.2. Preprocesamiento del conjunto Campo eléctrico

El conjunto de **Campo eléctrico** se compone de 188 archivos (archivos con extensión efm). Los archivos tienen registros cada segundo a diferencia del conjunto **Datos-meteorológicos** que son cada 10 minutos. El medidor de campo eléctrico realiza la medición que le corresponde 10 veces, para identificar las variaciones que puedan presentarse entre los milisegundos, por eso se tiene la hora, minuto y segundo repetidos para varios registros, pero con diferentes niveles de campo eléctrico como se observa en la ilustración 16.

```
00:00:00,+0.04,0
00:00:00,+0.04,0
00:00:00,+0.04,0
00:00:00,+0.04,0
00:00:00,+0.05,0
00:00:00,+0.04,0
00:00:00,+0.05,0
00:00:00,+0.04,0
00:00:00,+0.04,0
00:00:00,+0.05,0
```

***Ilustración 16: 10 primeras líneas de un archivo EFM del conjunto "Campo-eléctrico"***

Igual que en el conjunto anterior se unieron todos los archivos en uno solo archivo CSV, para su manipulación y se eliminó la columna 3 ya que solo indica si el rotor del medidor de campo eléctrico funciona y no es información relevante para saber si se detectó una descarga eléctrica. Posteriormente se agregó la fecha en la primera columna la cual viene dada en el nombre de los archivos como se muestra en la ilustración 12. Para reducir el número de registros con la misma hora, se obtuvo el promedio del campo eléctrico de cada segundo, simplificando de 10 registros por segundo a 1, obteniendo así un nuevo archivo con las 2 columnas que se agregaron: fecha y promedio del campo eléctrico como se muestra en la ilustración 17 y la tabla 9.

```
date, electric_field
2019-01-07 12:27:52,-0.02
2019-01-07 12:27:53,-0.02
2019-01-07 12:27:54,-0.02
2019-01-07 12:27:55,-0.02
2019-01-07 12:27:56,-0.01
2019-01-07 12:27:57,-0.01
2019-01-07 12:27:58,-0.01
2019-01-07 12:27:59,-0.01
2019-01-07 12:28:00,-0.02
2019-01-07 12:28:01,-0.02
```

***Ilustración 17: Conjunto "Campo-eléctrico" preprocesado***

| Atributo | Descripción |
|---|---|
| **date** | Fecha y hora en la que se registró el evento |
| **electric_field** | Nivel del campo eléctrico |

***Tabla 9: Variables obtenidas de procesar el conjunto "Campo-eléctrico"***

### 4.2.3. Preprocesamiento del conjunto Rayos detectados

El conjunto **Rayos detectados** contiene un solo archivo con los eventos registrados por el software del medidor de campo eléctrico, como se puede observar en la ilustración 18 contiene la fecha y hora en la que se registró un rayo y la distancia a la que se detectó, también contiene registros de cuando se activa una alarma, cuando se inició el software, o cuando éste falló.

```
02/29/2016 19:36:25 SYSTEM EFM Software Started.
02/29/2016 19:36:34 Configuration Saved
02/29/2016 22:08:46 INAOE parque Lightning Detected at 35 km.
02/29/2016 22:08:46 INAOE parque Lightning Alarm Active.
02/29/2016 22:10:12 INAOE parque Lightning Detected at 28 km.
02/29/2016 22:12:27 INAOE parque Lightning Detected at 26 km.
02/29/2016 22:13:00 INAOE parque Lightning Detected at 25 km.
02/29/2016 22:14:40 INAOE parque Lightning Detected at 27 km.
02/29/2016 22:15:49 INAOE parque Lightning Detected at 20 km.
02/29/2016 22:16:48 INAOE parque Lightning Detected at 25 km.
```

***Ilustración 18: 10 primeras del archivo EFMEvents.log del conjunto "Rayos-detectados"***

Para seleccionar la información más importante del archivo, los cuales son la fecha y hora en que se detectó un rayo, la distancia a la que se detectó respecto al medidor, se hizo uso

de expresiones regulares, para formar un patrón de búsqueda que ayudara a extraer las características ya mencionadas. Se creó un nuevo archivo CSV el cual contiene 4 columnas con las siguientes características extraídas, como se puede observar en la ilustración 19 y la tabla 10.

```
date,date_round,Llightning,distance
2016-02-29 22:08:46,2016-02-29 22:10:00,true,35
2016-02-29 22:10:12,2016-02-29 22:10:00,true,28
2016-02-29 22:12:27,2016-02-29 22:10:00,true,26
2016-02-29 22:13:00,2016-02-29 22:10:00,true,25
2016-02-29 22:14:40,2016-02-29 22:10:00,true,27
2016-02-29 22:15:49,2016-02-29 22:20:00,true,20
2016-02-29 22:16:48,2016-02-29 22:20:00,true,25
2016-02-29 22:17:57,2016-02-29 22:20:00,true,31
2016-02-29 22:19:14,2016-02-29 22:20:00,true,20
```

***Ilustración 19: Conjunto "Rayos-detectados" preprocesado***

| Dato | Descripción |
|---|---|
| **date** | Fecha y hora en la que se registró el evento |
| **date_round** | Fecha redondeada para poder unir este conjunto con el conjunto de datos meteorológicos |
| **lightning** | Si se detectó un rayo |
| **distance** | Distancia a la que cayó el rayo |

***Tabla 10: Variables obtenidas de procesar el conjunto "Rayos-detectados"***

El campo date_round se creó con la finalidad de poder correlacionar el conjunto de la estación meteorológica con el de rayos detectados; este campo se creó con la fecha extraída y se redondeó el tiempo diez minutos hacia arriba y 10 minutos hacia abajo.

## 4.3. Transformación

Esta fase consistió en transformar los datos a un formato adecuado para la herramienta en la que se realizaría la fase de minería que en este caso fue el software Weka. Para la transformación de los conjuntos se utilizó el software Rapidminer ya que provee algoritmos para la discretización y normalización de los datos; se utilizó el mismo software para la unión y obtención de los conjuntos resultantes.

Para poder aplicar los algoritmos de Minería de datos se crearon 2 conjuntos **Rayos-meteorológicos** y **Rayos-campo-eléctrico** como se observa en la tabla 11. Cada conjunto resultó en un archivo ARFF como se observa en las ilustraciones 20 y 21 y un diccionario de datos como se observa en la tabla 12 y 13

| Conjunto | Descripción | Número de variables | Número de registros |
|---|---|---|---|
| **Rayos-meteorológicos** | Unión conjuntos Rayos detectados con los Datos meteorológicos mediante los campos “date_round” y “date” | 11 | 41,501 |
| **Rayos- campo-eléctrico** | Unión conjuntos Rayos detectados y Campo eléctrico mediante los campos “date” de cada conjunto | 4 | 41,501 |

***Tabla 11: Conjuntos finales obtenidos***

```
@RELATION LightningsWeatherCloudData

@ATTRIBUTE 'date' DATE "yyyy.MM.dd HH:mm:ss Z"
@ATTRIBUTE 'lightning' {'true','false'}
@ATTRIBUTE 'distance' real
@ATTRIBUTE 'temp' real
@ATTRIBUTE 'chill' real
@ATTRIBUTE 'dew' real
@ATTRIBUTE 'heat' real
@ATTRIBUTE 'hum' real
@ATTRIBUTE 'wspdavg' real
@ATTRIBUTE 'bar' real
@ATTRIBUTE 'rain' real

@DATA
"2019.10.02 17:30:00 -0500",'false',0.0,20.0,20.0,16.1,20.6,79.0,11.2,762.2,0.6
"2019.10.02 17:40:00 -0500",'false',0.0,19.7,19.4,16.1,20.0,81.0,6.5,762.1,0.8
"2019.10.02 17:50:00 -0500",'false',0.0,19.3,18.9,16.1,20.0,81.0,9.7,761.9,0.8
"2019.10.02 18:00:00 -0500",'false',0.0,19.3,19.4,16.1,20.0,82.0,9.7,761.8,0.8
"2019.10.02 18:10:00 -0500",'false',0.0,19.2,18.9,16.1,20.0,81.0,9.7,761.7,0.8
"2019.10.02 18:20:00 -0500",'false',0.0,19.2,18.9,16.1,20.0,81.0,9.7,761.8,0.8
"2019.10.02 18:30:00 -0500",'false',0.0,19.3,19.4,16.1,20.0,81.0,9.7,761.8,0.8
"2019.10.02 18:40:00 -0500",'false',0.0,19.0,18.9,16.1,19.4,82.0,9.7,761.9,0.8
"2019.10.02 18:50:00 -0500",'false',0.0,18.7,18.3,15.6,19.4,83.0,11.2,761.9,0.8
"2019.10.02 19:30:00 -0500",'false',0.0,18.2,18.3,15.6,18.3,84.0,7.9,762.5,0.8
"2019.10.02 19:50:00 -0500",'false',0.0,13.9,13.9,11.7,13.9,85.0,13.0,763.3,5.8
"2019.10.02 20:00:00 -0500",'false',0.0,13.6,13.9,12.2,13.9,90.0,1.4,763.5,9.6
"2019.10.02 20:10:00 -0500",'false',0.0,13.9,13.9,12.8,13.9,93.0,1.4,763.6,11.8
"2019.10.02 20:20:00 -0500",'false',0.0,14.4,14.4,12.8,14.4,91.0,3.2,763.9,12.6
"2019.10.02 20:30:00 -0500",'false',0.0,14.4,14.4,12.8,14.4,91.0,3.2,764.0,12.8
```

***Ilustración 20: Archivo ARFF obtenido del conjunto "Rayos-meteorológicos"***

| **Atributo** | **Descripción** |
|---|---|
| **date** | Fecha y hora en la que se registró el evento |
| **lightning** | True indica si se detectó un rayo y false si no se detectó un rayo |
| **distance** | Distancia a la que cayó el rayo |
| **temp** | Temperatura externa |
| **chill** | Sensación térmica |
| **dew** | Punto de rocío externo |
| **heat** | Índice de calor externo |
| **hum** | Humedad exterior |
| **wspdavg** | Velocidad promedio del viento |
| **bar** | Presión barométrica |

| | |
|---|---|
| **rain** | Lluvia |

***Tabla 12: Variables obtenidos del conjunto "Rayos-meteorológicos"***

```
@RELATION RapidMinerData

@ATTRIBUTE 'date' DATE "yyyy.MM.dd HH:mm:ss Z"
@ATTRIBUTE 'lightning' {'false','true'}
@ATTRIBUTE 'distance' real
@ATTRIBUTE 'electric_field' real

@DATA
"2019.01.07 12:28:02 -0600",'false',0.0,-0.02,
"2019.01.07 12:32:34 -0600",'false',0.0,-0.06,
"2019.01.07 12:32:52 -0600",'false',0.0,-0.09,
"2019.01.07 12:40:50 -0600",'false',0.0,-0.03,
"2019.01.07 12:49:52 -0600",'false',0.0,0.02,
"2019.01.07 12:51:12 -0600",'false',0.0,0.02,
"2019.01.07 13:05:57 -0600",'false',0.0,-0.02,
"2019.01.07 13:12:30 -0600",'false',0.0,-0.02,
"2019.01.07 13:16:29 -0600",'false',0.0,-0.0,
"2019.01.07 13:30:23 -0600",'false',0.0,0.0,
"2019.01.07 13:36:24 -0600",'false',0.0,-0.01,
"2019.01.07 13:39:58 -0600",'false',0.0,0.02,
"2019.01.07 13:41:05 -0600",'false',0.0,0.01,
"2019.01.07 13:46:36 -0600",'false',0.0,0.02,
"2019.01.07 13:53:38 -0600",'false',0.0,0.01,
"2019.01.07 13:55:17 -0600",'false',0.0,0.02,
"2019.01.07 14:05:27 -0600",'false',0.0,0.01,
"2019.01.07 14:09:50 -0600",'false',0.0,-0.01,
"2019.01.07 14:15:40 -0600",'false',0.0,0.03,
"2019.01.07 14:16:17 -0600",'false',0.0,0.02,
"2019.01.07 14:17:52 -0600",'false',0.0,0.03,
"2019.01.07 14:20:24 -0600",'false',0.0,0.04,
```

***Ilustración 21: Archivo ARFF del conjunto "Rayos-campo-eléctrico"***

| Atributo | Descripción |
|---|---|
| **date** | Fecha y hora en la que se registró el evento |
| **lightning** | True indica si se detectó un rayo y false si no se detectó un rayo |
| **distance** | Distancia a la que cayó el rayo |
| **electric_field** | Nivel del campo |

***Tabla 13: Variables obtenidos del conjunto "Rayos-campo-eléctrico"***

## 4.4. Minería de datos

A continuación, se muestran todos los experimentos realizados y los resultados de haber aplicado los algoritmos de minería de datos. El algoritmo utilizado fue J48 el cual crea un

árbol de decisión, se utilizó este algoritmo debido a la capacidad que tiene para categorizar los datos de entrada además de que permite utilizar datos numéricos y categóricos y por su fácil legibilidad a la hora de entregar los resultados, además su costo computacional no requiere mucho.

### 4.4.1. Análisis de las variables

Antes de pasar directamente a la fase de minería de datos, se realizó un análisis de cada variable en los conjuntos obtenidos con relación a los rayos de forma individual. El análisis se realizó con las herramientas Power BI y Rapidminer ya que estas herramientas permiten analizar los conjuntos antes de aplicar los algoritmos y obtener información relevante para poder contrastar con los resultados que arrojen los modelos de minería.

Utilizando el conjunto **Rayos-detectados** se puede observar en la ilustración 22 que la cantidad de descargas eléctricas atmosféricas detectadas aumentó para el 2017, mientras que para el 2018 y 2019 disminuyó considerablemente en comparación al 2016.

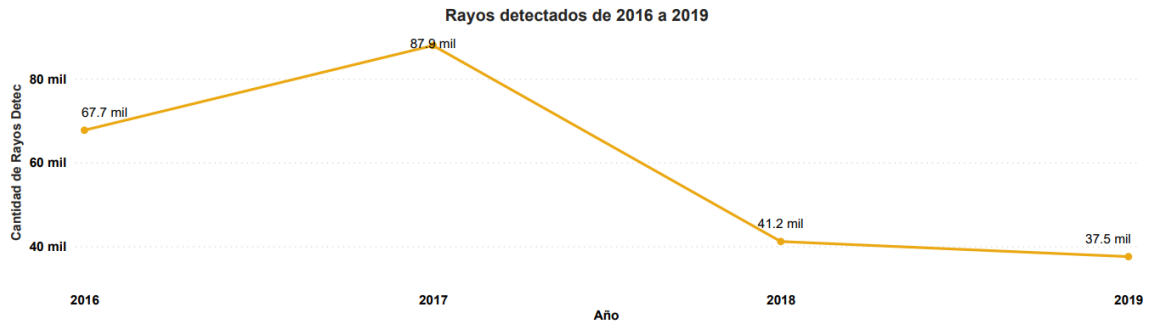


***Ilustración 22: Rayos detectados de 2016 a 2019***

En la ilustración 23 se observa que el medidor de campo eléctrico detectó una mayor cantidad de descargas eléctricas atmosféricas mientras más se fue acercando al límite de detección que son 38 km, mientras que en el radio del kilómetro 32 y 36 no detectó ninguna descarga eléctrica.

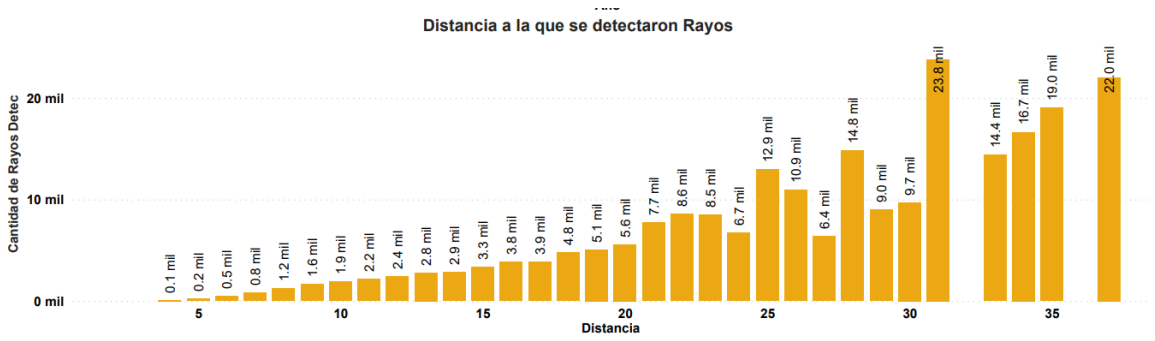


***Ilustración 23: Cantidad de rayos detectados por KM***

Para el caso del conjunto de **Rayos-meteorológicos** se obtuvieron los valores mínimos, maximos y promedio para cada variable en relacion a los rayos detectados y no detectados como se muestra en la tabla 13. Tambien se graficó cada variable en relacion a la cantidad de rayos detectados y rayos no detectados como se muestra en las ilustraciones 22 y 23.

En la ilustracion 24 se puede observar que para la generacion de una descarga electrica atmosferica, los grados celcius suelen estar entre los 12 y 24 grados, para las variables temperatura, sensacion térmica, punto de rocio y calor. Este comportamiento indica que a temperaturas bajas y altas, no es probable que suceda una descarga electrica, lo mismo aplica para las variables de “Sensacion Térmica”, “Punto de Rocío”, y “Calor”. En la ilustracion 25 la variable a tomar es la “Humedad” en la cual se empezo a detectar una mayor cantidad apartir de 75 grados”. Tambien se puede observar que la generacion de descargas electricas aumenta considerablemente a partir de los 12 grados con los picos mas altos entre los 14 y 15 grados. Estos rangos de valores los retomaremos mas adelante para contrastarlos con los resultados arrojados al aplicar el algoritmo de mineria.

| Atributo | Rayos Detectados valores | Rayos no detectados valores | Unidad de medida |
|---|---|---|---|
| **temp (Temperatura)** | MIN (11.9), MAX (37), AVG (26.7) | MIN (3), MAX (29.3), AVG (17.6) | ºC (grados Celsius) |
| **chill (Sensación térmica)** | MIN (9.4), MAX (27.2), AVG (15.8) | MIN (2.8), MAX (29.4), AVG (17.4) | ºC (grados Celsius) |

| | | | |
|---|---|---|---|
| **dew (Punto de rocío)** | MIN (1.1), MAX (17.2), AVG (13.3) | MIN (-11.7), MAX (17.8), AVG (9.1) | ºC (grados Celsius) |
| **heat (Calor)** | MIN (12.2), MAX (26.1), AVG (16.7) | MIN (2.8), MAX (27.8), AVG (17.2) | ºC (grados Celsius) |
| **hum (Humedad)** | MIN (26), MAX (97), AVG (82.1) | MIN (9), MAX (97), AVG (61) | % (unidad de medida en porcentaje) |
| **wspdavg (Velocidad promedio del viento)** | MIN (0), MAX (32), AVG (11.5), AVG (11.57) | MIN (0), MAX (35), AVG (6) | km/h (kilómetros/hora) |
| **bar (Presión barométrica)** | MIN (760.2), MAX (769.1), AVG (764) | MIN (759), MAX (772.4), AVG (765.8) | mmHg (milímetros de mercurio) |
| **rain (Lluvia)** | MIN (0), MAX (43.8), AVG (9.9) | MIN (0), MAX (43.8), AVG (0.7) | mm (milímetro de agua caída) |

***Tabla 14: Métricas obtenidas de cada variable en relación con los rayos del conjunto "Rayos-meteorológicos"***

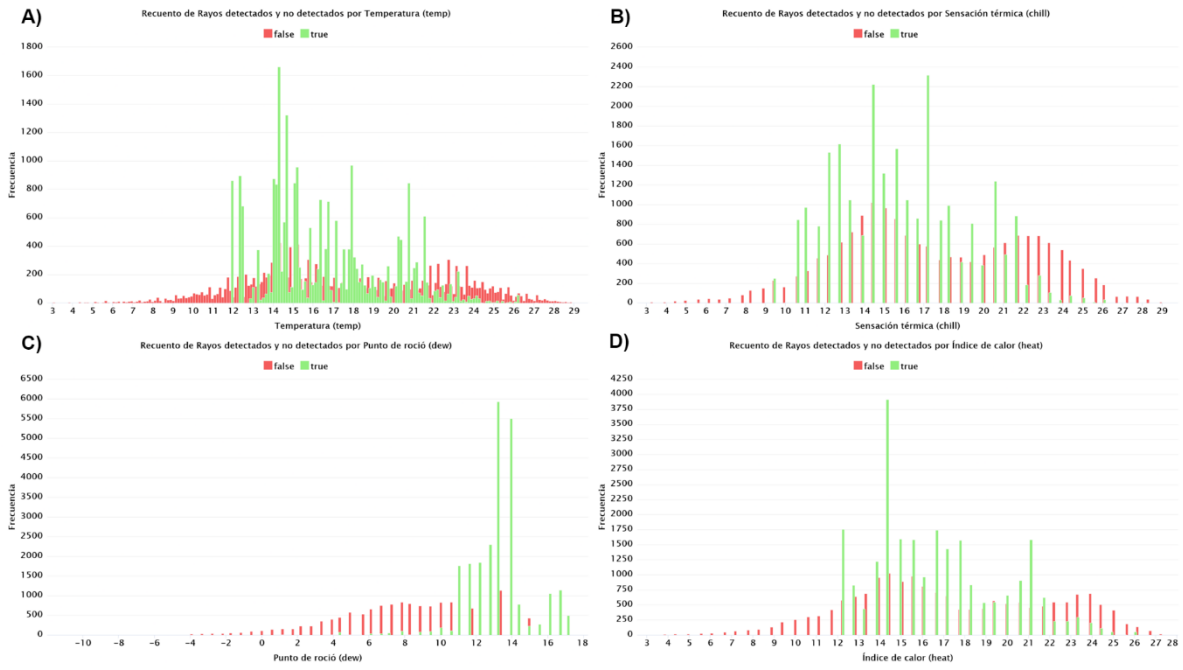


***Ilustración 24: Conteo de rayos por A) Temperatura, B) Sensación térmica, C) Punto de rocío, D) índice de calor***

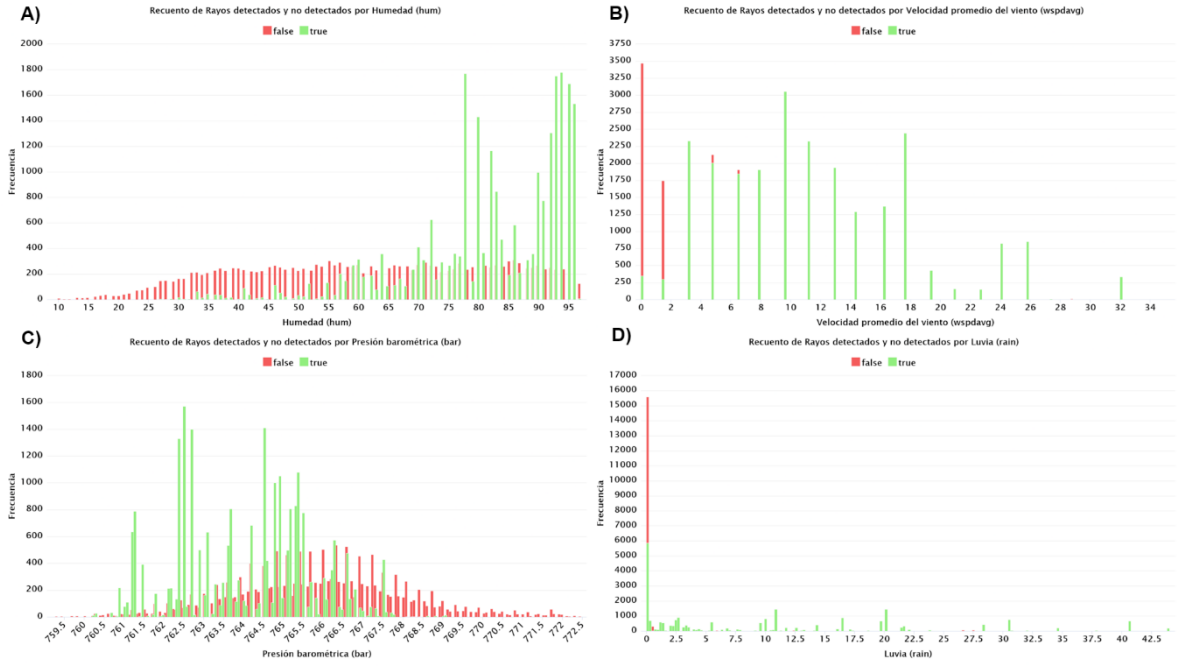


***Ilustración 25: Conteo de rayos por A) Humedad, B) Velocidad promedio del viento, C) Presión barométrica, D) Lluvia***

Para el conjunto **Rayos-campo-eléctrico**, se realizó el mismo proceso de obtener los valores mínimos, máximos y promedio de la tabla 14. También se graficó la cantidad de

rayos detectados y no detectados por el campo eléctrico como se observa en la ilustración 26, en donde verde indica los rayos detectados y rojo los rayos no detectados.

| Atributo | Rayos Detectados valores | Rayos no detectados valores | Unidad de medida |
|---|---|---|---|
| **electric_field (Campo eléctrico)** | MIN (-11.16), MAX (11), AVG (1.2) | MIN (-10.7), MAX (11), AVG (0.06) | kv/m (kilovolts/metro) |
| **distance (Distancia)** | MIN (4), MAX (37), AVG (27) | MIN (0), MAX (0), AVG (0) | km (kilómetros) |

***Tabla 14: Métricas obtenidas de cada variable en relación con los rayos del conjunto "Rayos-campo-eléctrico"***

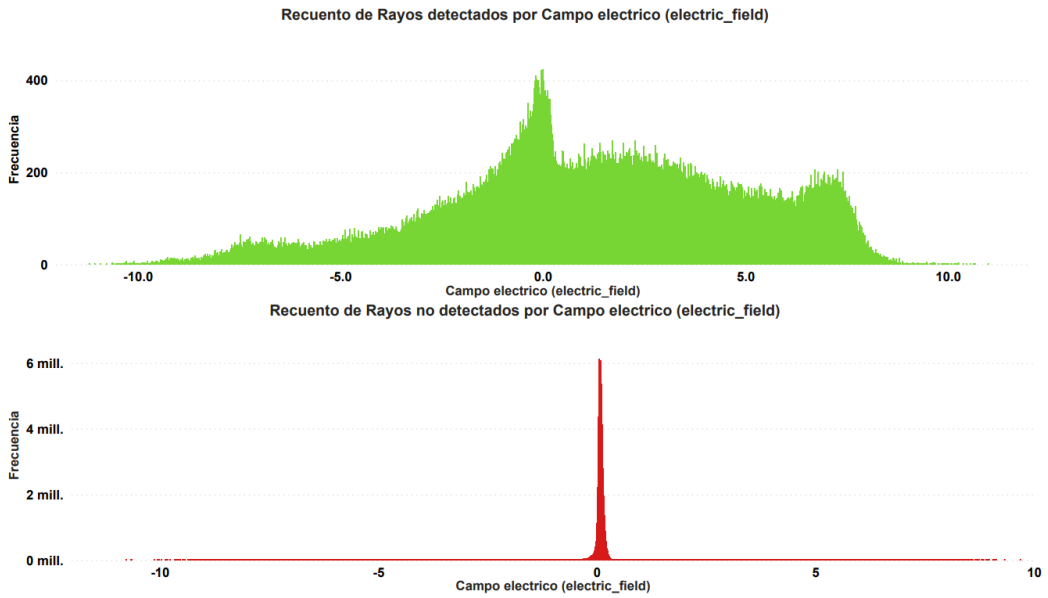


***Ilustración 26: Conteo de rayos detectados y no detectados por campo eléctrico***

Terminado el análisis de los conjuntos, para comprender mejor la distribución de los datos, se procedió a la aplicación de los algoritmos de Minería de datos.

### 4.4.2. Experimento 1: Aplicación del modelo J48 al conjunto Rayos-meteorológicos

Para el primer experimento se aplicó el algoritmo J48 al conjunto **Rayos-meteorológicos**. Con el objetivo de encontrar reglas que permitan predecir una descarga eléctrica mediante

variables meteorológicas. Antes de pasar directamente a la aplicación del modelo se balancearon los datos con Rapidminer, ya que se tienen más instancias de rayos no detectados en la variable lightning como se muestra en la tabla 15. Una vez realizado el experimento se obtuvieron los resultados presentados en la Tabla 15.

| | | | |
|---|---|---|---|
| **Numero de instancias** | 35186 | | |
| **Número de variables** | 9 (temp, chill, dew, heat, hum, wspavg, bar, rain, lightning) | | |
| **Algoritmo** | J48 | | |
| **Test** | Cross validation, Folds (10) | | |
| **F-Measure** | 0.908 | | |
| **Mejores reglas** | | Precisión | Confianza |
| **rain > 0.6 y heat <= 21.1 y wspdavg > 1.4 y hum > 96: false (33)** | | 1 | 0.00093 |
| **rain <= 0.6 y dew > 11.7 y bar <= 761.5: true (922.0/44.0)** | | 0.95 | 0.026 |
| **rain > 0.6 y heat <= 21.1 y wspdavg > 1.4 y hum <= 96: true (13264.0/684.0)** | | 0.94 | 0.37 |
| **rain <= 0.6 y dew <= 11.7: false (12735.0/810.0)** | | 0.93 | 0.36 |
| **rain > 0.6 y heat > 21.1: false (193.0/17.0)** | | 0.91 | 0.0054 |

***Tabla 15: Resultados experimento 1***

Las 2 mejores reglas que se obtuvieron para el experimento 1 con el conjunto de **Rayos-meteorológicos** fueron la numero tres y cuatro.

Regla 3 indica cuando se puede generar un rayo: Lluvia tiene que ser mayor a 0.6 el calor menor o igual a 21.1 grados la velocidad del viento mayor a 1.4 km/h y la humedad menor o igual a 96 grados dando un total de 13264 reglas encontradas con este patrón.

Regla 4 indica cuando no se puede generar un rayo: Lluvia tiene que ser menor a 0.6 el punto de rocío tiene que ser menor o igual a 11.7 dando un total de 12735 reglas encontradas con este patrón.

### 4.4.3. Experimento 2: Aplicación del modelo J48 al conjunto Rayos-meteorológicos discretizado

Se aplicó el mismo modelo, pero esta vez con el conjunto discretizado, esto con el objetivo de ver como se comportaba el modelo ante el conjunto discretizado y que tipo de reglas se obtenían. Cada variable se discretizó en 3 rangos (min, med, max).

| | | | |
|---|---|---|---|
| **Numero de instancias** | 35186 | | |
| **Número de variables** | 9 (temp, chill, dew, heat, hum, wspavg, bar, rain, lightning) | | |
| **Algoritmo** | J48 | | |
| **Test** | Cross validation, Folds (10) | | |
| **F-Measure** | 0.908 | | |
| **Mejores reglas** | | Precisión | Confianza |
| **rain = [3.5 - ∞] y hum = [86.5 - ∞] y wspdavg = [5.6 - 12.1] y heat = [18.6 - ∞]: false (1.0)** | | 1 | 0.00002908 |
| **rain = [3.5 - ∞] y hum = [86.5 - ∞] y wspdavg = [12.1 - ∞]: true (3190.0/11.0)** | | 0.99 | 0.09 |
| **rain = [3.5 - ∞] y hum = [86.5 - ∞] y wspdavg = [5.6 - 12.1] y heat = [-∞ - 14.7]: true (2245.0/43.0)** | | 0.98 | 0.06 |
| **rain = [-∞ - 0.1] y dew = [-∞ - 11.4]: false (11669.0/619.0)** | | 0.94 | 0.33 |
| **rain = [-∞ - 0.1] y dew = [-∞ - 11.4]: false (11669.0/619.0) y dew = [11.4 - 13.6] y wspdavg = [-∞ - 5.6]: false (1800.0/139.0)** | | 0.92 | 0.05 |

***Tabla 16: Resultados experimento 2***

Como se puede observar en la Tabla 16, el F-measure bajó, pero la precisión para las 5 mejores reglas aumentó, no solo eso si no que no hubo mucha variación en las reglas que arrojó el algoritmo con el conjunto discretizado respecto a los resultados del experimento 1.

Las 2 mejores reglas que se obtuvieron para el experimento 2 con el conjunto de **Rayos-meteorológicos-discretizado** fueron la numero dos y cuatro.

Regla 2 indica cuando se puede generar un rayo: Lluvia tiene que ser mayor o igual a 3.5 la humedad mayor o igual a 86.5 grados y la velocidad del viento mayor o igual a 12.1 km/h dando un total de 3190 reglas encontradas con este patrón.

Regla 4 indica cuando no se puede generar un rayo: Lluvia tiene que ser menor a 0 o mayor a 0.1 el punto de rocío tiene que ser menor a 0 o mayor o igual a 11.4 dando un total de 11669 reglas encontradas con este patrón.

La mejor regla obtenida para este experimento es: Lluvia tiene que empezar apartir

### 4.4.4 Experimento 3: Aplicación del modelo J48 al conjunto Rayos-campo-eléctrico

El objetivo de este experimento fue predecir cuándo es más probable que suceda una descarga atmosférica dependiendo del nivel del campo eléctrico. Para este experimento se realizó el mismo procedimiento que el experimento 1, se balancearon los datos y se aplicó el algoritmo J48, pero esta vez para el conjunto “Rayos detectados” y “Campos eléctricos”.

| | |
|---|---|
| **Numero de instancias** | 465890 |
| **Número de variables** | 2 (electric_field, lightning) |
| **Algoritmo** | J48 |
| **Test** | Cross validation, Folds (10) |
| **F-Measure** | 0.93 |

| **Mejores reglas** | Precisión | Confianza |
|---|---|---|
| **electric_field > -0.1 y electric_field > 0.2 y electric_field > 0.3: true (142629.0/4988.0)** | 0.96 | 0.3 |
| **electric_field > -0.1 y electric_field <= 0.2: false (218787.0/10796.0)** | 0.95 | 0.46 |
| **electric_field <= -0.1 y electric_field <= -0.54: true (72738.0/5138.0)** | 0.92 | 0.1 |
| **electric_field <= -0.1 y electric_field > -0.54 y electric_field <= -0.26: true (11130.0/2734.0)** | 0.75 | 0.02 |
| **electric_field > -0.1 y electric_field > 0.2 y electric_field <= 0.3: false (10190.0/2579.0)** | 0.74 | 0.02 |

***Tabla 17: Resultados experimento 2***

Para este conjunto de datos no se discretizaron los datos, ya que solo se tiene una variable sobre el que se realizará la generación de reglas que es el campo eléctrico (electric_field). Además de que el modelo ya está generando rangos en cada rama que generada.

Las 2 mejores reglas que se obtuvieron para el experimento 2 con el conjunto de **Rayos-campo-eléctrico** fueron la numero uno y dos.

Regla 1 indica cuando se puede generar un rayo: El campo eléctrico tiene que ser mayor a –0.1, mayor a 0.2 y mayor a 0.3 dando un total de 142629 reglas encontradas con este patrón.

Regla 4 indica cuando no se puede generar un rayo: El campo eléctrico tiene que ser mayor a –0.1 y menor a 0.2 dando un total de 218787 reglas encontradas con este patrón.

### 4.4.5. Experimento 4: Aplicación del modelo J48 al conjunto Rayos-meteorológicos dividido por rangos mediante la variable distancia (distance)

El objetivo de este experimento es ver que reglas arroja el modelo cuando se divide el data set en cuatro dependiendo de la variable distancia (distance). Esto con la finalidad de saber si las reglas previamente entregadas por el algoritmo J48 variaban dependiendo de la

distancia a la que se detectó una descarga eléctrica. La división se realizó en 4 rangos [0 – 9.5], [9.5 - 19], [19 – 28.5], [28.5 - 38] como se puede observar en las tablas 18, 19, 20, 21.

| | | | |
|---|---|---|---|
| **Numero de instancias** | 600 | | |
| **Número de variables** | 9 (temp, chill, dew, heat, hum, wspavg, bar, rain, lightning) | | |
| **Algoritmo** | J48 | | |
| **Test** | Cross validation, Folds (10) | | |
| **F-Measure** | 0.96 | | |
| **Mejores reglas** | | Precisión | Confianza |
| **rain <= 0.6 y wspdavg > 9.7 y heat <= 21.1 y bar > 763.1 y dew <= 11.7: false (25.0/1.0)** | | 0.96 | 0.04 |
| **rain <= 0.6 y wspdavg > 9.7 y heat <= 21.1 y bar > 763.1 y dew > 11.7: true (18.0/1.0)** | | 0.94 | 0.04 |
| **rain <= 0.6 y wspdavg <= 9.7: false (219.0/3.0)** | | 0.98 | 0.3 |
| **rain > 0.6 y heat <= 19.4 y wspdavg > 3.2 y bar <= 766: true (243.0/1.0)** | | 0.99 | 0.4 |
| **rain <= 0.6 y wspdavg > 9.7 y heat > 21.1: false (36.0)** | | 1 | 0.06 |

***Tabla 18: Resultados Experimento 4 rango [0 - 9.5]***

| | |
|---|---|
| **Numero de instancias** | 7610 |
| **Número de variables** | 9 (temp, chill, dew, heat, hum, wspavg, bar, rain, lightning) |

| | | | |
|---|---|---|---|
| **Algoritmo** | J48 | | |
| **Test** | Cross validation, Folds (10) | | |
| **F-Measure** | 0.93 | | |
| **Mejores reglas** | | Precisión | Confianza |
| **rain > 0 y wspdavg > 1.4 y temp > 22.1: false (85.0)** | | 1 | 0.01 |
| **rain <= 0 y wspdavg <= 7.9 y bar > 762.6: false (2343.0/35.0)** | | 0.98 | 0.30 |
| **rain > 0 y wspdavg > 1.4 y temp <= 22.1 y bar <= 765.5: true (2948.0/104.0)** | | 0.96 | 0.38 |
| **rain <= 0 y wspdavg > 7.9 y dew > 9.4 y heat <= 21.1 y bar <= 764.6: true (366.0/23.0)** | | 0.93 | 0.04 |
| **rain <= 0 y wspdavg > 7.9 y dew <= 9.4: false (628.0/53.0)** | | 0.91 | 0.08 |

***Tabla 19: Resultados Experimento 4 rango [9.5 - 19]***

| | | | |
|---|---|---|---|
| **Numero de instancias** | 17530 | | |
| **Número de variables** | 9 (temp, chill, dew, heat, hum, wspavg, bar, rain, lightning) | | |
| **Algoritmo** | J48 | | |
| **Test** | Cross validation, Folds (10) | | |
| **F-Measure** | 0.9 | | |
| **Mejores reglas** | | Precisión | Confianza |
| **rain <= 0.6 y dew <= 11.7 y wspdavg <= 7.9: false (4234.0/49.0)** | | 0.98 | 0.24 |

| | | |
|---|---|---|
| **rain <= 0.6 y dew > 11.7 y hum > 85: false (791.0/15.0)** | 0.98 | 0.04 |
| **rain > 0.6 y heat > 21.1: false (99.0/3.0)** | 0.96 | 0.005 |
| **rain > 0.6 y heat <= 21.1 y wspdavg > 1.4: true (6867.0/348.0)** | 0.94 | 0.39 |
| **rain <= 0.6 y dew <= 11.7 y wspdavg > 7.9 y dew <= 8.3: false (1201.0/93.0)** | 0.92 | 0.06 |

***Tabla 20: Resultados Experimento 4 rango [19 - 28.5]***

| | | | |
|---|---|---|---|
| **Numero de instancias** | **22074** | | |
| **Número de variables** | 9 (temp, chill, dew, heat, hum, wspavg, bar, rain, lightning) | | |
| **Algoritmo** | J48 | | |
| **Test** | Cross validation, Folds (10) | | |
| **F-Measure** | 0.91 | | |
| **Mejores reglas** | | Precisión | Confianza |
| **dew <= 10.6 y heat <= 15.6: false (3186.0/6.0)** | | 0.99 | 0.99 |
| **dew > 10.6 y wspdavg > 1.4 y rain <= 0.6 y heat > 15.6 y bar > 762.1 y hum > 56 y bar > 766.7: false (216.0/1.0)** | | 0.99 | 0.99 |
| **dew <= 10.6 y heat > 15.6 y wspdavg > 14.4 y temp > 22.7: false (386.0/16.0)** | | 0.95 | 0.96 |
| **dew > 10.6 y rain > 0.6 y heat <= 21.1: true (7876.0/420.0)** | | 0.94 | 0.94 |
| **dew > 10.6 y wspdavg > 1.4 y rain <= 0.6 y heat <= 15.6: false (537.0/35.0)** | | 0.93 | 0.93 |

***Tabla 21: Resultados Experimento 4 rango [28.5 - 38]***

Como se puede observar en las reglas generadas, las reglas no tienden a variar mucho dependiendo de la distancia a la que se detectó el rayo.

Las 2 mejores reglas que se obtuvieron para el experimento 4 con el conjunto de **Rayos-meteorológicos** divididos por rangos fueron.

Rango [28.5 - 38] regla 4 indica cuando se puede generar un rayo: El punto de rocío tiene que ser mayor a 10.6 la lluvia tiene que ser mayor a 0.6 y el calor tiene que ser menor o igual a 21.1 dando un total de 7876 reglas encontradas con este patrón.

Rango [28.5 - 38] regla 1 indica cuando no se puede generar un rayo: El punto de rocío tiene que ser menor o igual a 10.6 y el calor tiene que ser menor o igual a 15.6 dando un total de 3186 reglas encontradas con este patrón.

### 4.4.5. Experimento 5: Aplicación del modelo J48 al conjunto Rayos-campo-eléctrico dividido por rangos mediante la variable distancia (distance)

Se realizo el mismo proceso que el experimento 4, pero aplicado al conjunto “Rayos detectados” y “Campo eléctrico”, como se observa en las tablas 22, 23, 24, 25. Debido a que el árbol solo arrojo 4 reglas para cada rango, solo se selecciona la mejor de las 4.

| | | | |
|---|---|---|---|
| **Numero de instancias** | 8842 | | |
| **Número de variables** | 9 (electric_field, lightning) | | |
| **Algoritmo** | J48 | | |
| **Test** | Cross validation, Folds (10) | | |
| **F-Measure** | 0.94 | | |
| **Mejores reglas** | | Precisión | Confianza |
| **electric_field > 0.3: true (2511.0/83.0)** | | 0.96 | 0.96 |

***Tabla 22: Resultados Experimento 5 rango [0 - 9.5]***

| | | | |
|---|---|---|---|
| **Numero de instancias** | 66142 | | |
| **Número de variables** | 9 (electric_field, lightning) | | |
| **Algoritmo** | J48 | | |
| **Test** | Cross validation, Folds (10) | | |
| **F-Measure** | 0.93 | | |
| **Mejores reglas** | | Precisión | Confianza |
| **electric_field > -0.11 y electric_field > 0.2 y electric_field > 0.29: true (20240.0/730.0)** | | 0.96 | 0.96 |

***Tabla 23: Resultados Experimento 5 rango [9.5 - 19]***

| | | | |
|---|---|---|---|
| **Numero de instancias** | 66142 | | |
| **Número de variables** | 9 (electric_field, lightning) | | |
| **Algoritmo** | J48 | | |
| **Test** | Cross validation, Folds (10) | | |
| **F-Measure** | 0.93 | | |
| **Mejores reglas** | | Precisión | Confianza |
| **electric_field > -0.1 y electric_field > 0.2 y electric_field > 0.29: true (50366.0/1831.0)** | | 0.96 | 0.96 |

***Tabla 24: Resultados Experimento 5 rango [19 - 28.5]***

| | | | |
|---|---|---|---|
| **Numero de instancias** | 227286 | | |
| **Número de variables** | 9 (electric_field, lightning) | | |
| **Algoritmo** | J48 | | |
| **Test** | Cross validation, Folds (10) | | |
| **F-Measure** | 0.93 | | |
| **Mejores reglas** | | Precisión | Confianza |
| **electric_field > -0.11 y electric_field > 0.22 y electric_field > 0.3: true (69756.0/2437.0)** | | 0.96 | 0.96 |

***Tabla 25: Resultados Experimento 5 rango [28.5 - 38]***

Para este experimento las reglas obtenidas arrojaron resultados similares donde indica que un rayo se empieza a generar cuando el campo eléctrico es mayor a 0.2

A continuación, se muestran las mejores reglas obtenidas de cada experimento en la tabla 27.

| Experimento | No. Regla | Regla | Precisión | Confianza |
|---|---|---|---|---|
| **1** | 3 | rain > 0.6 y heat <= 21.1 y wspdavg > 1.4 y hum <= 96: true (13264.0/684.0) | 0.94 | 0.37 |
| **1** | 4 | rain <= 0.6 y dew <= 11.7: false (12735.0/810.0) | 0.93 | 0.36 |
| **2** | 2 | rain = [3.5 - ∞] y hum = [86.5 - ∞] y wspdavg = [12.1 - ∞]: true (3190.0/11.0) | 0.99 | 0.09 |
| **2** | 2 | rain = [-∞ - 0.1] y dew = [-∞ - 11.4]: false (11669.0/619.0) | 0.94 | 0.33 |
| **3** | 1 | electric_field > -0.1 y electric_field > 0.2 y electric_field > 0.3: true (142629.0/4988.0) | 0.96 | 0.3 |
| **3** | 4 | electric_field <= -0.1 y electric_field > -0.54 y electric_field <= -0.26: true (11130.0/2734.0) | 0.75 | 0.02 |

| | | | | |
|---|---|---|---|---|
| **4** | Rango [28.5 - 38] regla 1 | dew <= 10.6 y heat <= 15.6: false (3186.0/6.0) | 0.99 | 0.99 |
| **4** | Rango [28.5 - 38] regla 4 | dew > 10.6 y rain > 0.6 y heat <= 21.1: true (7876.0/420.0) | 4 | Rango [28.5 - 38] regla 4 |
| **5** | Rango [28.5 - 38] regla 1 | electric_field > -0.11 y electric_field > 0.22 y electric_field > 0.3: true (69756.0/2437.0) | 0.96 | 0.96 |

***Tabla 26: Mejores reglas obtenidas de cada experimento***

En la tabla 27 se muestra la precisión general que se obtuvo en cada experimento

| **Experimento** | **Precisión general** |
|---|---|
| **1** | 0.9 |
| **2** | 0.87 |
| **3** | 0.93 |
| **4 rango [0-9.5]** | 0.96 |
| **4 rango [9.5-19]** | 0.93 |
| **4 rango [19-28.5]** | 0.90 |
| **4 rango [28.5-38]** | 0.91 |
| **5 rango [0-9.5]** | 0.94 |
| **5 rango [9.5-19]** | 0.93 |
| **5 rango [19-28.5]** | 0.93 |
| **5 rango [28.5-38]** | 0.93 |

***Tabla 27: Precisión general de cada experimento***

Al analizar los resultados para el conjunto **Rayos-meteorológicos**. Se observó que en la mayoría de los experimentos la variable rain era el nodo principal del cuál se desprendían las demás ramas, también las variables más predominantes a la hora de detectar una descarga eléctrica son “rain (lluvia)”, “dew (punto de rocío)”, “wspdavg (velocidad promedio del viento)”, “bar (presión atmosférica)”, para comprobar esto se realizó una selección de variables obteniendo el siguiente resultado como se observa en la siguiente ilustración y

efectivamente se obtuvieron estas 4 variables las cuales predominan mayormente al momento de generarse una descarga atmosférica.

```
=== Run information ===

Evaluator:    weka.attributeSelection.CfsSubsetEval -P 1 -E 1
Search:       weka.attributeSelection.BestFirst -D 1 -N 5
Relation:     RapidMinerData-weka.filters.unsupervised.attribute.Remove-R9-10
Instances:    35186
Attributes:   9
              temp
              chill
              dew
              heat
              hum
              wspdavg
              bar
              rain
              lightning
Evaluation mode:    10-fold cross-validation


=== Attribute selection 10 fold cross-validation (stratified), seed: 1 ===

number of folds (%)  attribute
           0(  0 %)    1 temp
           0(  0 %)    2 chill
          10(100 %)    3 dew
           0(  0 %)    4 heat
           0(  0 %)    5 hum
          10(100 %)    6 wspdavg
          10(100 %)    7 bar
          10(100 %)    8 rain
```

***Ilustración 27: Selección de variables***

# Capítulo 5

# 5. Conclusiones y trabajo a futuro

Con las reglas arrojadas después de aplicar el algoritmo J48 el cual crea un árbol de decisiones con reglas basadas en los datos. Se pudieron identificar ciertos patrones como:

- La precisión más relevante se obtuvo en el primer experimento, como se observa en la tabla 26 con la regla número tres con una precisión de 0.94 y una confianza de 0.37, ya que los resultados fueron los que más se acercaron a los resultados de las gráficas analizadas en la sección **4.4.1. Análisis de las variables.**
- Para el conjunto "Rayos-Campo-eléctrico" se observó que una descarga eléctrica atmosférica puede empezar a generarse cuando el campo eléctrico es superior a 0.1 o menor a -0.1.
- Con esta información se puede concluir que una descarga eléctrica atmosférica se puede generar a partir de -0.1 o 0.1, y la descarga más común ocurre cuando el campo eléctrico es positivo, esta descarga se podría determinar cómo una descarga de nube a suelo.
- Como se puede observar las variables determinantes en las reglas en las que puede empezar a generarse una descarga eléctrica atmosférica son: rain (lluvia), dew (punto de rocío), wspdavg (velocidad promedio del viento y bar (presión atmosférica). También puede observarse que para que ocurra una descarga atmosférica los valores de las variables dew (punto de roció), heat (calor), chill (sensación térmica) deben estar en un rango de 11 y 22, esto contrasta con el análisis previo que se realizó en la sección **4.4.1. Análisis de las variables**, en la cual se observó que un rayo se puede generar entre los rangos 12 y 24 en estas mismas variables.

Con las conclusiones tomadas anteriormente se lograron responder las preguntas planteadas en la sección 1.4 las cuales fueron.

- ¿Es posible establecer bajo qué correlación de variables meteorológicas y niveles de campo eléctrico puede suceder una descarga eléctrica?

Las respuestas a esta pregunta es que si se puede establecer una correlación con las variables meteorológicas usadas para este proyecto y así poder predecir cuándo es más probable que suceda una descarga eléctrica.

Al final de este proyecto se lograron cumplir los objetivos planteados los cuales fueron:

- Realizar la limpieza, procesamiento, carga y análisis de cada conjunto extraído.
- Se hizo uso de métodos computacionales y se obtuvieron patrones y tendencias de comportamiento.
- Se aplico el modelo de minería el cual fue un árbol de decisiones, y se obtuvieron reglas con las cuales se hizo un análisis para determinar cuando era más probable que suceda una descarga atmosférica en ciertas condiciones climatológicas.

Con estos puntos cumplidos se alcanzó el objetivo general, que fue la recolección el procesamiento de los datos y el análisis de los datos arrojados por el algoritmo de minería (árbol de decisión), con el cual se pudieron detectar patrones de comportamiento los cuales pueden indicar la generación de una descarga eléctrica atmosférica

El trabajo a futuro de este proyecto se pretende analizar la información de las demás estaciones meteorológicas y realizar la aplicación de más algoritmos de minería para poder contrastar la información con otras técnicas diferentes de minería de datos.

# 6. Bibliografía


VIZCAINO GARZON , P. A. (2008). *APLICACIÓN DE TÉCNICAS DE INDUCCIÓN DE ÁRBOLES DE DECISIÓN A PROBLEMAS DE CLASIFICACIÓN MEDIANTE EL USO DE WEKA.* BOGOTÁ.

Agie Wandala Putra, C. L. (2014). *CUMULONIMBUS PREDICTION USING ARTIFICIAL NEURAL.*

Agie Wandala Putra, C. L. (2015). Cumulonimbus prediction using artificial neural network back propagation with radiosonde indeces.

Aimée Booysens, S. V. (2014). Detection of Lightning Pattern Changes UsingMachine Learning Algorithms.

BOLTEK CORPORATION. (2016). Installation/Operators Guide for model EFM-100C. Canada.

Coronado Arjona, Bianchi Rosado, Vivas Burgos, & Perera Collí. (2017). Estudio comparativo de técnicas de minería de datos para la predicción de rutas de huracanes. *TECNOLOGÍA EDUCATIVA REVISTA CONAIC*, 10.

Davis Instruments. (2011). Vantage VUE Integrated Sensor Suite Installation Manual.

Dyuti Chatterjee, D. H. (2015). Application of Machine Learning Technique to. *International Journal of Scientific & Engineering Research*.

Elton Rafael Alves, C. T. (2017). Lightning prediction using satellite atmospheric sounding data and feed-forward artificial neural network. *Journal of Intelligent & Fuzzy Systems*.

Emamghoreishi, S., Moini, R., & Sadeghi, S. (2002). A modular neural network approach for locating cloud-to-ground lightning strokes. *IEEE*.

Himadri Chakrabarty, S. B. (2014). Prediction of Severe Thunderstorms Applying Neural Network using RSRW Data. *International Journal of Computer Applications*.

Himadri Chakrabarty, S. B. (2015). Application of K-Nearest Neighbor Technique to Predict Severe Thunderstorms. *International Journal of Computer Applications*.

Johari, D., Rahman, T. K., & Musirin, I. (2008). Artificial Neural Network Based Technique for Lightning Prediction. *IEEE*.

Juntian, G., ShanQiang, G., & Wanxing, F. (2011). A lightning motion prediction technology based on spatial clustering method.

Litta A. J, S. M. (2012). Artificial Neural Network Model for the Prediction of Thunderstorms over Kolkata. *International Journal of Computer Applications*.

Olaiya , F., & Adesesan, B. (febrero de 2012). Application of Data Mining Techniques in Weather Prediction and Climate. *Modern Education and Computer Science PRESS*, 9.

P.Kalaiyarasi, & Mrs.A.Kalaiselvi. (2018). Data Mining Techniques Using To Weather Prediction. *International Journal of Computer Science Trends and Technology*, 6.

Rafael Alves, E., Tavares da Costa, C., Gomes Lopes, M. N., Pereira da Rochaa, B. R., & Silva de Sá, J. A. (2017). Lightning prediction using satellite. *Journal of Intelligent & Fuzzy Systems*, 15.

Rasika P. Gawande, N. D. (2012). An Efficient Approach towards Thunderstorm Detection Using Saliency Map. *International Journal of Science and Research*.

Re, S. (2019). Descargas Atmosféricas – Preparación y Protección. 2.

SIERRA TAMARA, L. J., & ESCOBAR MORALES, G. A. (2004). *INVESTIGACIONES SOBRE DESCARGAS ELÉCTRICAS ATMOSFERICAS.*

sitio big data. (14 de 12 de 2019). *sitiobigdata.com*. Obtenido de https://sitiobigdata.com/2019/12/14/arbol-de-decision-en-machine-learning-parte-1/#:~:text=Ganancia%20de%20informaci%C3%B3n%3A,acuerdo%20con%20sus%20clasificaci%C3%B3n%20objetivo.

Taorong Qiu, S. Z. (2013). Application Study of Machine Learning in Lightning Forecasting.

Weng, L. Y., Omar, J. B., Siah, Y. K., Ahmed, S. K., Abidin, I. B., & Abdullah, N. (2010). Lightning Forecasting Using ANN-BP & Radiosonde. *IEEE*.