Universidad del Azuay

Facultad de Ciencias de la Administración

Escuela de Ingeniería de Sistemas y Telemática

ANÁLISIS DE COMPORTAMIENTO DE COMPONENTES ATMOSFÉRICOS Y VARIABLES METEOROLÓGICAS APLICANDO TÉCNICAS DE ASOCIACIÓN DE MINERÍA DE DATOS

Trabajo de graduación previo a la obtención del Título de Ingeniero en Sistemas y Telemática

Autor:
Jimmy Fernando Salto Sumba.

Directores:
Ing. Marcos Patricio Orellana Cordero

Cuenca – Ecuador
2020

DEDICATORIA

El presente trabajo se lo dedico principalmente a Dios, por ser la inspiración y brindarme la fuerza para continuar el del proceso de cumplir con mi objetivo. A mis padres, por estar conmigo, por su sacrificio, por enseñarme valores, apoyarme en todo momento y a que si caigo debo levantarme, además de ser las bases que me ayudaron a llegar hasta aquí. A mi hermano por estar siempre presente y su presencia en el transcurso de esta etapa de mi vida. A todas las personas que me apoyaron y han aportado de alguna manera para que este trabajo se realice con éxito.

AGRADECIMIENTO

Agradezco a Dios, que me dio el valor y la fuerza para continuar en este proceso por ser el apoyo y la fortaleza en todos los momentos de dificultad que se presentaron en el camino. Mi agradecimiento incuantificable a mis padres Luis y Ana quienes con su apoyo en todo momento me han permitido cumplir con esta meta en mi vida, de igual manera a mi hermano Danny, por su compañía en este proceso, así como a toda mi familia por su sacrificio y apoyo moral. Agradezco al ingeniero Marcos Orellana, tutor del proyecto por aportar con su conocimiento, así como a todos los docentes de la Universidad del Azuay por haber compartido sus conocimientos en el camino de la preparación de la profesión

ÍNDICE

FIGURAS

TABLAS

ECUACIONES

ANEXOS

RESUMEN

La relación entre los componentes atmosféricos y variables meteorológicas es importante para evaluar la calidad del aire y con ello evitar riesgos en la salud, sin embargo, encontrar una relación de asociación entre estos factores podría ser compleja por la cantidad de métodos de categorización que se pueden emplear, tales como, frecuencia, tamaño, binning, entre otros. Por lo tanto, el objetivo de este estudio fue la creación de una metodología que prepara los datos a través de un proceso de discretización y luego aplica técnicas de asociación de las posibles combinaciones entre las variables de estudio. Los resultados mostraron que la metodología empleada fue efectiva en la ubicación de patrones, que son de utilidad para el gestor ambiental para encontrar conocimiento.

Palabras clave: minería de datos, reglas de asociación, discretización, contaminantes atmosféricos, variables meteorológicas.

ABSTRACT

The relationship between atmospheric components and meteorological variables is important to assess air quality and thereby avoid health risks. However, finding an association relationship between these factors could be complex because of the number of categorization methods that can be used, such as frequency, size, binning, among others. Therefore, the objective of this study was the creation of a methodology that prepares the data through a process of discretization and then applies techniques of association of the possible combinations between the study variables. The results showed that the used methodology was effective in the location of patterns, which were useful for the environmental agent to find knowledge.

Keywords: Data mining, association rules, discretization, air pollutants, weather variables.

UNIVERSIDAD DEL AZUAY
Dpto. Idiomas

Translated by
Ing. Paúl Arpi

1. INTRODUCCIÓN

El fenómeno de la contaminación del aire se encuentra conectado con el desarrollo continuo de la sociedad. Las áreas urbanas se encuentran habitadas por un gran número de personas, por lo que la industria y el transporte se encuentran altamente desarrollados. En general, la contaminación emitida es una amenaza para los ciudadanos, dado que es uno de los principales causantes de problemas en la salud. Se hace necesaria entonces la regulación de los niveles de contaminantes del aire como una de las tareas más importantes para los gobiernos en desarrollo. Por lo tanto, es necesario que exista una adecuada gestión de la calidad del aire de parte de los gobernantes (Czechowski, Badyda, & Grzegorz, 2013; Martínez, Troncoso, Álvarez, & Riquelme, 2010).

Para evidenciar los problemas de contaminación del aire, es necesario recopilar datos que caractericen la calidad del aire. En las ciudades existe una monitorización constante de componentes atmosféricos y variables meteorológicas mediante sensores que recolectan una gran cantidad de datos. Luego, estos datos son utilizados en combinación tanto de técnicas de minería de datos como de *machine learning*, para descubrir patrones de comportamiento basados en grandes conjuntos de datos para generar conocimiento(Kalyankar & Alaspurkar, 2013). De esta forma, se crean modelos de predicción precisos a partir de grandes y complejos conjuntos de datos, con la finalidad de determinar hipótesis o pronosticar el comportamiento de la calidad del aire. Dichas predicciones resultan útiles para determinar cómo se encuentran los niveles de los contaminantes (Czechowski et al., 2013).

Se han realizado varios estudios relacionados tanto a nivel local como internacional. Un ejemplo que cabe destacar es el caso Estambul, provincia de Turquía, que ha implementado un sistema basado en módulos para realizar predicciones de contaminación del aire. El primer módulo recolecta datos de variables meteorológicas y mediciones diarias de los niveles de contaminación atmosférica de sitios web. El segundo módulo, conocido como módulo de pronóstico usa redes neuronales para pronosticar hasta tres días de niveles de contaminantes, y finalmente, el módulo web donde se visualizan los resultados categorizados en bueno, advertencia y peligroso (Kurt, Gulbagci, Karaca, & Alagha, 2008). En el caso de Franja de Gaza, donde se recopilaron datos durante un periodo de nueve años

[1977-1985] y posteriormente, con los resultados de las técnicas aplicadas sobre el conjunto de elementos, se realizó un análisis generando conocimiento útil para agricultoras, sector turístico, sectores hídricos y otros sectores involucrados con el comportamiento del aire (Kohail & Alaa, 2011). De igual manera, se pronosticó días polvorientos utilizando presión de aire, humedad y datos anteriores de días polvorientos (Sahafizadeh & Ahmadi, 2009). Así también, con el uso de redes neuronales, se determinó niveles de dióxido de azufre (SO2), monóxido de carbono (CO) y material particulado (PM10) con tres días de anticipación (Kurt & Oktay, 2010).

Estudios realizados en la ciudad de Cuenca determinaron el algoritmo óptimo para el análisis de componentes atmosféricos recolectados de una estación ubicada en el centro de la ciudad. Entre los estudios, existe un estudio que enuncia la técnica de minería de datos para encontrar relación entre componentes atmosféricos y variables meteorológicas. Así como también, la búsqueda de algoritmos óptimos para el análisis de variables de contaminación atmosférica (Andrade & Orellana, 2018; Ortega & Orellana, 2018).

A pesar de los estudios mencionados anteriormente, no existen aplicaciones de minería de datos que involucren tanto componentes atmosféricos como variables meteorológicas y entreguen resultados descriptivos y comprensibles basados en la discretización de esta variables y métodos de asociación. Un estudio que exploró estos métodos es de (Lanjewar & Shah, 2012) que aplicaron técnicas de asociación a un conjunto de datos de contaminantes del aire, de tal manera que, encontraron relaciones de asociación existentes entre estos componentes. Sin embargo, esta propuesta busca extender el estudio de (Lanjewar & Shah, 2012) extendiendo el estudio hacia los componentes atmosféricos y variables meteorológicas, con técnicas de discretización, y luego aplicar técnicas de asociación que determinen patrones relevantes dentro del conjunto de reglas de asociación.

Por lo tanto, el objetivo de este estudio es aplicar técnicas de asociación de minera de datos para analizar el comportamiento de componentes atmosféricos y variables meteorológicas. Esto se realizará a través de una discretización de estos componentes que proporcionen los resultados óptimos en cuanto a criterios de soporte, confianza y *lift*.

La zona de estudio es la ciudad de Cuenca, que cuenta con monitorización continua de componentes atmosféricos y variables meteorológicas para evaluar la calidad del aire. Sin embargo, en ciertos casos existe gran cantidad de información que no presenta relevancia o que presentan inconsistencias, por lo que, es necesario convertir estos datos en *sets* limpios para obtener resultados confiables, posteriormente, con los datos preprocesados se discretiza el conjunto de datos y se aplican técnicas de asociación, logrando obtener resultados que presenten relaciones de asociación relevantes.

2. MARCO TEÓRICO

Los datos recolectados a través de la Red de Monitoreo de Cuenca, registra datos a través de una estación automática instalada en el Centro Histórico. Esta infraestructura almacena tanto niveles de contaminación atmosférica como valores de variables meteorológicas en un rango de cuatro kilómetros a la redonda (Emov Ep, 2016). El conjunto de datos entregados a través del Municipio de Cuenca se describe a continuación:

2.1 Variables meteorológicas

Tabla 1*: Variables Meteorológicas.*

Variables meteorológicas	Siglas de representación	Unidad de medida
Humedad relativa	HR_AV	Valor Porcentual De Agua
Precipitación	PRECIP_SUM	Milímetros De Agua
Presión Atmosférica	PATM_AV	Hectopascal
Punto de rocío	DP_AV	Grados Centígrados
Radiación Global	RADGLOBAL_AV	Vatios Por Metro Cuadrado
Radiación UVA	UVA_AV	Vatios Por Metro Cuadrado
Radiación UVE	UVE_AV	Vatios Por Metro Cuadrado
Temperatura del aire	TEMPAIRE_AV	Grados Centígrados
Velocidad del viento	WINDSPEED_AV	Metros Por Segundo

Humedad Relativa

La Humedad Relativa se refiere al momento en que una masa de aire entra en saturación, es decir, no puede contener más vapor de agua y según las condiciones climáticas el restante de aire es transformado en agua líquida o hielo. Por ello, es conocido como el porcentaje de vapor presente en el aire y, puede variar a factores como lluvia, cercanía al mar o sectores verdes (Sánchez Zavaleta, 2016).

Temperatura del aire

Temperatura hace referencia a la rapidez del movimiento de partículas que conforman la materia, por lo tanto, mientras más agitación exista mayor será la temperatura. Es por ello que varía de acuerdo a varios factores, tales como, el día, la noche, ubicación geográfica y finalmente a la estación del año. Así pues, en verano se incrementa y en invierno disminuye (Jiménez, Capa, & Lozano, 2004).

Presión Atmosférica

La Presión Atmosférica indica la fuerza ejercida por el peso del aire debido a la gravedad. Es así que, puede variar de acuerdo a la altitud de la zona, es decir, mientras más elevada sea nuestra posición con respecto a la atmósfera, menor será la cantidad del aire disminuyendo la fuerza ejercida sobre el cuerpo (Jiménez et al., 2004).

Velocidad del Viento

El viento es el desplazamiento de aire de un punto a otro, generalmente es causado cuando dos puntos generan diferencia de presión o de temperatura. En primer lugar, cuando la presión del aire es distinta es propenso a moverse desde la zona con alta presión a la zona con baja presión. En segundo lugar, al elevarse la temperatura del aire superando la temperatura de su entorno su volumen aumenta, por lo que, disminuye su densidad. De esta manera, debido al efecto flotación, la masa de aire caliente asciende y es remplazada por otras masas de aire (Jiménez et al., 2004).

Precipitación

Una nube está conformada por diminutas gotas y cristales de hielo provenientes del vapor de agua causada por una masa de aire que al momento de ascender a través de la atmosfera es enfriado hasta llegar a la saturación. Por ello, al alcanzar la saturación es condensada en formas de gotas, por lo que, finalmente se transforma en estado líquido como llovizna o chubasco o, a su vez, en estado sólido como cristales de hielo (Jiménez et al., 2004).

Punto de Roció

El punto de roció expresa la cantidad de vapor de agua contenida en un gas o en el entorno, en otras palabras, es la temperatura a la que se solidifica el vapor de agua que contiene cierta cantidad de gas. De esta manera, se transforma en nubes o escarcha (Martines & Gómez, 2008) .

Radiación Global

La radiación es la transferencia de energía entre el Sol y la Tierra que se desplaza a través del espacio a manera de ondas que poseen determinadas cantidades de energía que son absorbidas por la superficie terrestre o cuerpos situados sobre ella (Jiménez et al., 2004).

Radiación UV

Como se conoce, el sol emite energía a grandes cantidades en longitudes de onda, esto, a su vez, afecta a la piel del ser humano causando quemaduras y otros efectos secundarios para la salud. Sin embargo, la capa de ozono sirve como filtro de radiación UV, no obstante, los tipos de radiación UVA-UVE no son absorbidas del todo por la capa de ozono, por lo que, el paso de estas radiaciones afecta a la piel y ojos del ser humano (Martines & Gómez, 2008).

Contaminantes atmosféricos

Tabla 2*: Contaminantes atmosféricos.*

Contaminantes atmosféricos	Siglas de Representación	Unidad de Medida
Dióxido de Nitrógeno	NO2	Microgramos Por Metro Cúbico
Dióxido de Azufre	SO2	Microgramos Por Metro Cúbico
Material Particulado	PM2.5	Microgramos Por Metro Cúbico
Monóxido de Carbono	CO	Microgramos Por Metro Cúbico
Ozono	O3	Microgramos Por Metro Cúbico

Dióxido de Nitrógeno (NO2)

El dióxido de nitrógeno (NO2) es el resultado de la combustión generada a altas temperaturas, tales como, vehículos, empresas eléctricas, calentadores o cocinas (Ubilla & Yohannessen, 2017). Es considerado nocivo para la salud, puesto que, al mezclarse con el aire su toxicidad aumenta causando problemas respiratorios. Por ejemplo: tos, secreción nasal y dolor de garganta (Uysal & Gunal, 2018).

Dióxido de Azufre (SO2)

El dióxido de nitrógeno (NO2) es la consecuencia de quema de combustibles fósiles, es decir, combustión de carbón con alto contenido de azufre y petróleo. Así pues, las principales fuentes de emanación son fundición de materiales, quema de desechos, y de igual manera, fuentes naturales como los volcanes (Ubilla & Yohannessen, 2017; Uysal & Gunal, 2018).

Material Particulado (PM2.5)

El Material Particulado (PM2.5) es originado por la suspensión de polvo, tierra, tormentas de viento, volcanes u otro material de carretera o agricultura, por ello, es considerado un contaminante complejo, puesto que está formado por diferentes componentes de diferentes emisiones (Ubilla & Yohannessen, 2017). Por lo tanto, es clasificado en función de su tamaño, así pues, PM10 hace referencia a partículas de menos de 10 micrómetros y PM 2.5 partículas de menos de 2.5 micrómetros (Uysal & Gunal, 2018).

Ozono (O3)

El Ozono (O3) generalmente se incrementan en verano, dado que, son generados por procesos químicos tales como compuestos orgánicos volátiles y dióxido de nitrógeno en medio de luz solar, por lo tanto, su presencia es común en ambientes urbanos y sectores industriales (Uysal & Gunal, 2018).

Monóxido de Carbono (CO)

El monóxido de carbono (CO) es generado por la combustión de combustibles de carbono, es así que, las principales emanaciones de monóxido de carbono son vehículos motorizados que utilizan combustible como gasolina o diésel y fabricas industriales que

empleen compuestos de carbono. Por otro lado, es conocido por su nivel de toxicidad para el ser humano, dado que, causa problemas cardiovasculares o incluso la muerte (Téllez, Rodríguez, & Fajardo, 2006; Uysal & Gunal, 2018).

3. TRABAJOS RELACIONADOS

Es necesario recalcar que, el constante aumento de la población junto con la actividad urbana implica que la calidad del aire se vea afectada por ciertos contaminantes, tales como, PM10, S02, CO, y N02. Estos contaminantes son considerados nocivos para la salud, es así que la continua exposición a este tipo de contaminación puede causar enfermedades respiratorias, cardiovasculares, así como también, deficiencias neurológicas e incluso partos prematuros. Por ello, con el uso de minería de datos es posible proporcionar y descubrir conocimiento de datos ambientales con el propósito de gestionar la calidad del aire (Du & Varde, 2016). Es así que, la mayoría de los estudios relacionados con el uso tanto de variables meteorológicas y contaminantes atmosféricos están orientados a evaluar la calidad del aire, así como, los riesgos que pueden ocasionar en la salud del ser humano. Tal es el caso de (Wu, Wang, Liu, & Li, 2018) donde utilizaron minería de datos con métodos de agrupamiento y posteriormente reglas de asociación para el control de la contaminación del aire en las zonas urbanas de las ciudades de China. Es necesario mencionar que dentro de este estudio se utilizan variables meteorológicas tales como: humedad, velocidad del viento y temperatura, con un periodo de tres meses. La humedad y la velocidad se agruparon en cinco niveles, y la temperatura en nueve niveles. Posteriormente, obtuvieron conjuntos de elementos frecuentes del conjunto de elementos inicial y finalmente utilizaron un algoritmo conocido como Kulczynski que, en otras palabras, realiza un filtrado de reglas de asociación que indican relaciones interesantes. Algo semejante ocurre en la investigación de (Toti et al., 2016) que utilizo reglas de asociación con el propósito de identificar efectos de la contaminación con respecto al asma. En primer lugar, establecieron criterios de corte con intención de limitar el número de reglas resultantes, de esta manera, se eliminó la redundancia de reglas encontradas. Así pues, encontraron 27 reglas que cumplen con los criterios definidos, de esta manera, se identificó que el Ozono es un componente que se encuentra previo a los ataques de asma, así también, se determinó que, la mezcla de otros componentes como SO2 NO, NO2, PM, resultan ser más dañinas para la salud.

Por otro lado, para la predicción de la calidad del aire (Cagliero et al., 2016) utilizaron contaminantes atmosféricos como PM10 y PM2.5. Posteriormente, fueron categorizados en código de colores, es decir, de verde a rojo, siendo el color verde no crítico y rojo como nivel

crítico o altamente severo. Por otro lado, las variables meteorológicas consideradas fueron temperatura, humedad, precipitación, viento y presión atmosférica, sin embargo, estos se discretizaron en niveles personalizados respectivamente. Finalmente, aplicaron reglas de asociación, logrando predecir la calidad del aire en regiones donde no existen estaciones de monitoreo. De la misma manera, como indica (Du & Varde, 2016) a través de la recolección de datos de tráfico vehicular y PM2.5 lograron predecir concentraciones de PM2.5. Por lo que, esto puede usarse para realizar estimaciones de la calidad del aire en resultados interpretables por el público, así como también, la de detección de problemas de salud y seguridad de los habitantes.

Es necesario mencionar que dentro de minería de datos no solo existe la posibilidad de evaluar la calidad del aire, tal es el caso del estudio de (Qin, Liu, Wang, Song, & Qu, 2015) donde recolectaron contaminantes atmosféricos como PM10, PM2.5, SO2, O3, CO, NO2. Luego, aplicaron técnicas de minería de datos con el propósito de analizar una proyección espacio temporal de la materia dentro de la región Jing-Jin-Ji en China. En otras palabras, analizaron la manera en la que se propaga la materia particulada en cuanto a distancia en un tiempo determinado, es así que, el resultado de las reglas de asociación indicó rutas del movimiento de la materia. En resumen, revelaron información de cómo y cuándo la materia particulada de una ciudad afectara a otra ciudad cercana.

Finalmente, a pesar de que ciertas investigaciones tienen el mismo objetivo de encontrar relaciones entre variables meteorológicas y componentes atmosféricos, las técnicas utilizadas pueden variar. Es así que, como menciona (Huang, Zhang, Wang, & Zhu, 2015) utilizaron conjuntos de datos de PM2.5 y conjuntos de variables meteorológicas recolectados por el Ministerio de Protección Ambiental y el Centro de Datos Meteorológicos en China. Sin embargo, utilizaron algoritmos de redes neuronales con la finalidad de pronosticar la contaminación del aire basados en PM2.5, de donde se infiere que, la única variable meteorológica que se relaciona con la PM2.5 es la humedad, por lo tanto, se dice que la humedad provoca una disolución de la PM2.5. No obstante, variables meteorológicas como la temperatura y velocidad del viento no presentaron correlaciones con PM2.5.De manera similar, con el uso de redes neuronales se analiza el enfoque de la contaminación en Delhi - India, esto con la ayuda de datos recolectados en un periodo de cuatro años (2011-2015), de

tal manera que, los resultados encontrados indican que tanto el NO2 y O3 se incrementarán un futuro, por otro lado, los niveles de SO2 y CO mantendrán su tendencia a causa del incremento del polvo de carreteras y construcciones (Uysal & Gunal, 2018).

4. MATERIALES Y MÉTODO

4.1Materiales

4.1.1 Zona de Aplicación

Cuenca, con 505.585 habitantes, se ubica como la tercera ciudad del Ecuador, localizada al sur del país (INEC, 2017). Se encuentra a una altura de 2.550 metros sobre el nivel del mar, entre las coordenadas 78°59’ – 79°01’ de longitud oeste y 2°52’ – 2°54’de latitud sur. Presenta un clima templado, con una temperatura anual de 15°C. Cuenta con pluviosidad anual entre 700-110 y 75% de humedad relativa. Una velocidad de viento de entre 4m/s y 5,5 m/s (Emov Ep, 2014; Emov Ep, 2016).

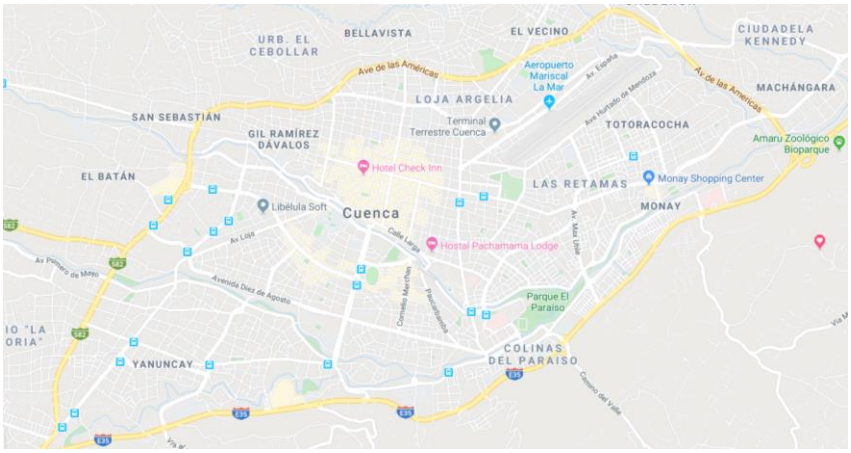

Ilustración 1: *Zona de estudio Ciudad de Cuenca.*

4.1.2 Datos

Los datos utilizados en el desarrollo del estudio competen a un periodo de tiempo de enero a noviembre del año 2018, e involucran mediciones tomadas en rangos de 10 minutos cada uno. Las variables que se utilizaron para el desarrollo del estudio son:

Contaminantes Atmosféricos:

- Dióxido de Nitrógeno (NO2).
- Dióxido de Azufre (SO2).

- Material Particulado (PM2,5).
- Monóxido de Carbono (CO).
- Ozono (O3).

Variables Meteorológicas:

- Humedad Relativa (HR_AV).
- Precipitación (PRECIP_SUM).
- Presión Atmosférica (PATM_AV).
- Punto del rocío (DP_AV).
- Radiación Global (RADGLOBAL_AV).
- Radiación ultravioleta A (UVA_AV).
- Radiación ultravioleta E (UVE_AV).
- Temperatura del Aire (TEMPAIRE_AV).
- Velocidad del viento (WINDSPEED_AV).

El conjunto de datos proporcionado originalmente contiene valores como máximo, mínimo y promedio de cada variable meteorológica. Por ello, se usará valores promedios. En base a que tanto valores máximos como mínimos representan situaciones poco comunes dentro del conjunto de datos (Andrade & Orellana, 2018). Por otro lado, ya que dentro del conjunto de datos existen atributos que tienen dependencia de otros, como es el caso del viento, el cual esta especificado en dos atributos como dirección del viento y la velocidad del viento. Una medida de aplicación estaría en la confluencia de estos dos atributos, lo que determinaría un índice final. El objetivo de este estudio no corresponde a la creación de nuevos atributos determinados por otros, por tanto, la dirección del viento es descartada desde el inicio del tratamiento de los datos.

4.1.3 Software

4.1.4 RapidMiner

RapidMiner es una plataforma software de ciencia de datos caracterizada por ser abierta y extensible. Un aspecto importante a mencionar es la funcionalidad que ofrece RapidMiner. Es decir, unifica todo el ciclo de vida de la ciencia de datos. Iniciando en la preparación de

los datos, así pues, ofrece soluciones rápidas a problemas como valores perdidos y atípicos. Posteriormente, aplica aprendizaje automático y finalmente implementa el modelo de predicción. Varias empresas utilizan productos y funcionalidades de RapidMiner con la finalidad de generar ingresos, reducir costos y evitar riesgos. Particularmente RapidMiner posee portabilidad, por ello, es ejecutado en varias plataformas, a más de ser una herramienta gratuita (Rapidminer, 2019).

RapidMiner es usado como herramienta para el análisis de contaminantes atmosféricos y variables meteorológicas (Andrade & Orellana, 2018). Permite discretizar el set de datos tanto de contaminantes atmosféricos y variables meteorológicas. En base al conjunto de datos resultante se aplica algoritmos de minería de datos. Por último, se evalúan los resultados de los algoritmos aplicados.

4.2 Método

La minería de datos analiza gran cantidad de datos para obtener información y conocimiento, que será utilizada en diferentes pronósticos. Durante los últimos años, la minería de datos ha tenido un enorme crecimiento, es decir, existen varios modelos estándar para la minería de datos, todos estos enfocados en pasos secuenciales, cada uno de estos pasos ayudan al desarrollo e implementación de tareas de minería de datos (Shafique & Haseeb, 2013). Por ello, en este estudio se utilizará el modelo CRISP-DM, no obstante, con la finalidad de construir y documentar las respectivas etapas del desarrollo se utilizó un Metamodelo de Ingeniería de Procesos de Software conocida como SPEM.

4.2.1 CRISP-DM

El modelo a utilizar en este estudio es CRISP-DM, puesto que proporciona una visión de ciclo de vida de un proyecto de minería de datos. Incluye fases del proyecto, tareas a realizar, y finalmente obtención de resultados (Ferna, 2010).

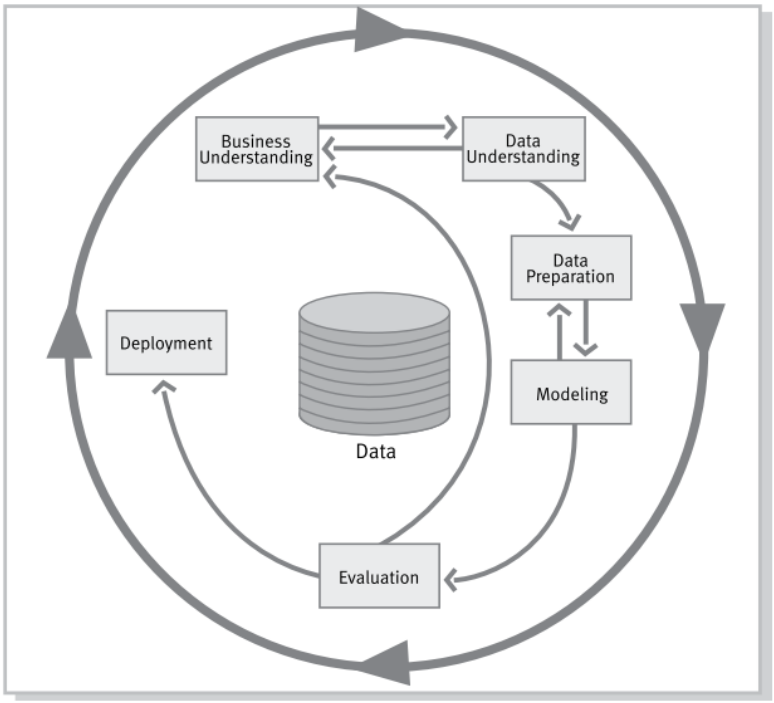

Ilustración 2: *Cross-Industry Standard Process for Data Mining (CRISP-DM) process model (Ncr et al., 2000).*

Se debe mencionar que el orden de las fases no es secuencial, esto quiere decir que, indican tan solo el flujo de los procesos más importantes y frecuentes entre sí. Sin embargo, durante el desarrollo del proyecto depende de qué tarea en particular se debe realizar a continuación. Por lo que, al ser un proceso cíclico la minería de datos no termina al implementarse la solución, por el contrario, se presentan nuevas incógnitas con mayor especificidad (Ferna, 2010). El ciclo de vida de CRISP-DM etapas: comprensión del negocio o problema, comprensión de los datos, preparación de los datos, modelado, evaluación e implementación.

4.2.1.1 Comprensión del problema

La etapa inicial de la metodología CRISP-DM se enfoca en comprender tanto los objetivos como requerimientos del problema desde el punto de vista comercial. De esta manera, este conocimiento es convertido en un problema de minería de datos, por lo que, se elabora y diseña un plan para lograr los objetivos (Azevedo & Santos, 2008; Wirth, 2000).Por ello, dentro del estudio fue necesario identificar el problema para diseñar una metodología adecuada.

4.2.1.2 Comprensión de los datos

En la etapa de comprensión de datos, inicialmente, se realiza la recopilación de los datos, posteriormente, se realizan ciertas actividades con la finalidad de familiarizarse con los datos. Por lo que, se logra identificar los primeros problemas con la calidad de los datos y, se determinan las primeras relaciones que ayudan en la formación de una hipótesis (Ferna, 2010; Hahsler, 2016; McNicholas, Murphy, & O'Regan, 2008).

Como se ha dicho, en el desarrollo de este estudio los datos considerados fueron contaminantes atmosféricos y variables meteorológicas, recolectados por una estación de aire en la ciudad de Cuenca. Así, por ejemplo, las variables meteorológicas se ilustran en la Tabla 1, y las variables meteorológicas en la Tabla 2.

4.2.1.3 Preparación de los datos

Dentro de la preparación de los datos se aplican técnicas de tratamiento de la información con la finalidad de construir un conjunto de datos limpio para para ser utilizado en etapas posteriores (Colina, 2017). Es necesario recalcar que, las tareas de preparación de datos se pueden realizar varias veces, por lo tanto, se involucran procesos de seleccionan de tablas, atributos y registros (Ferna, 2010; McNicholas et al., 2008). Esta etapa es importante para el estudio ya que fue necesario incluir procesos de transformación y limpieza de datos.

4.2.1.4 Modelado

Por lo general, existen distintas técnicas que brindan solución a problemas de minería de datos, por ello, se puede aplicar distintas técnicas de modelado. Así pues, los parámetros son ajustados de manera óptima (Ferna, 2010). Por otro lado, es necesario mencionar que ciertas técnicas poseen requerimientos específicos, por lo tanto, es común volver a la fase de preparación de los datos (Colina, 2017; Ferna, 2010). Este estudio se utilizó reglas de asociación, con la finalidad de descubrir conocimiento oculto del conjunto de datos de variables ambientales.

4.2.1.5 Evaluación

En esta etapa, es necesario evaluar el modelo realizado desde el punto de vista del análisis de datos, por lo tanto, es importante evaluar a profundidad cada uno de los pasos realizados que involucran al modelo (Ferna, 2010). Por lo general, al final de esta etapa se decide si los resultados obtenidos son utilizados (Colina, 2017). Por ello, el estudio utilizó diferentes métricas de evaluación con el propósito de mejorar los resultados.

4.2.1.6 Implementación

Por lo general, los resultados obtenidos representados como conocimiento son presentados de forma que el usuario final pueda utilizar o implementar, por lo tanto, la implementación puede repetirse en un proceso de minería de datos o presentarse a manera de informe que dé solución al problema planteado. Cualesquiera que fuese el caso, es importante tomar acciones con el propósito de utilizar los modelos creados (Colina, 2017; Ferna, 2010). Finalmente, el estudio propuesto indicó los resultados a través de graficas con la descripción respectiva de sus métricas.

4.2.2 Metamodelo de Ingeniería de Procesos de Software

Conviene subrayar que los procesos involucrados con la implementación o desarrollo de software requieren gestionar etapas de desarrollo, por ello, modelar los procesos de software es una forma de mejorar el desarrollo y la calidad de los resultados. Es así que, existen varios métodos para modelar procesos, sin embargo, aquellos que se enfocan en productos de estudio son los más idóneos. SPEM (*Software Process Engineering Metamodel*) es uno de ellos, por lo tanto, con el uso de SPEM se puede visualizar, construir y documentar etapas de desarrollo, dicho de otra manera, se puede indicar el conjunto de procesos de desarrollo con sus respectivas etapas. Dado que, cuenta con sintaxis y estructura para cada aspecto de los procesos con la finalidad de ser utilizada en gestión de procesos o análisis (Menéndez Domínguez & Castellanos Bolaños, 2015).

4.3 Factor de valor atípico local

Dentro de minería de datos, la detección de valores atípicos es fundamental para las aplicaciones, tales como, detección de fraudes financieros, distinción de irregularidades, o incluso pronósticos erróneos del clima (Bhatt, Dhakar, & Chaurasia, 2016; Yan, Cao, &

Rundensteiner, 2017). Dentro de un conjunto de datos, es común encontrar elementos que no satisfacen el comportamiento común de los datos, de ahí que, estos elementos son conocidos como anormalidades, desviaciones, discordancias, o anomalías (Aggarwal, 2013). Por lo tanto, la detección de valores atípicos se enfoca en encontrar valores que se alejan considerablemente de la distribución común de los datos. Es decir, un valor atípico es considerado como un punto de datos que es significativamente diferente a los elementos existentes (Jiawei, Micheline, & Pei, 2019).

Así, por ejemplo, con el uso un histograma de frecuencia como se observa en la Ilustración 3 contiene un rango inicial [50, 59] que cuenta con tres valores, el segundo rango [60, 69] siete valores, doce valores para el tercero [70, 79] y finalmente, un solo valor para el último rango [80, 89]. Por lo tanto, se observa que la mayoría de valores se concentran en los rangos de [50,79], de lo que se deduce que el valor del último rango, es considerado como valor atípico (Schubert, Zimek, & Kriegel, 2014; Wu et al., 2016).

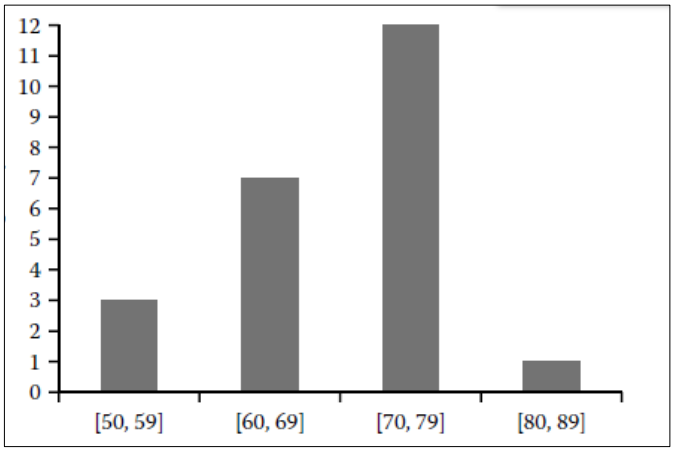

Ilustración 3*: Histograma de frecuencia (Schubert et al., 2014; Wu et al., 2016).*

Como se mencionó anteriormente, dentro de minería de datos existen varias técnicas para la detección de valores atípicos, uno de ellos es el método de detección basados en densidad. Dicho método evalúa la correspondencia o relación entre la densidad local alrededor de un valor y la densidad local alrededor de sus valores vecinos (Schubert et al., 2014).Por lo tanto, su enfoque principal es establecer un factor atípico local (LOF) a cada valor del conjunto de datos representando un grado de anormalidad (Diao, Liu, Meng, Ye, & He, 2015). De tal manera que, si un elemento está ubicado dentro de un grupo homogéneo el valor LOF

asignado será próximo a uno, no obstante, si la densidad entre el grupo local y la densidad en torno a sus grupos próximos es mayor, se le asigna un valor de LOF más elevado (Schubert et al., 2014; Wu et al., 2016).

4.4Discretización

Dentro de minería de datos no es conveniente en ocasiones trabajar con valores continuos o numéricos, ya sea porque es innecesario desde el punto de vista computacional, o, a su vez, porque no se pueda extraer información útil, así también, el no categorizar el conjunto de datos puede causar ineficacia y lentitud en la implementación de los algoritmos (Lavangnananda & Chattanachot, 2017). Por lo tanto, es necesario utilizar técnicas de tratamiento de la información, tal es el caso de la discretización. La discretización es considerada como el proceso de transformación de valores numéricos continuos a valores categóricos o nominales, esto quiere decir que, convierte un conjunto de datos cuantitativos en cualitativos, generando una división categórica en base a intervalos, con el propósito de determinar específicamente grupos que representen la mayor cantidad de información como sea posible (Geaur Rahman & Zahidul Islam, 2016; Lavangnananda & Chattanachot, 2017; Ramirez-Gallego et al., 2015; Ramírez-Gallego et al., 2016; Saeed et al., 2015).De manera más específica, se dice que, si se tiene un conjunto de datos continuos Aj el proceso de discretización, divide al conjunto de datos en grupos basados en puntos de cortes $Pj = \{Pj1, Pj2, \dots Pjq\}$, donde, $Pj1\ y\ Pjq$ son los limites superior e inferior de Aj respectivamente. Así pues, en base a los puntos de corte, los grupos pertenecientes de Aj se definen, tal que, $I = \{[Pj1, Pj2], (Pj2, Pj3], \dots, (Pjq - 1, Pjq]\}$ (Geaur Rahman & Zahidul Islam, 2016).

Por otro lado, conviene subrayar que una de las ventajas de la discretización es la minimización y reducción de datos, puesto que reduce grandes cantidades de datos a grupos categóricos. No obstante, estas ventajas se acompañan de un costo, significa que, el proceso de discretización genera pérdida de información, por lo que, uno de los aspectos principales es disminuir está perdida de información a través del uso de distintas técnicas de discretización (García, Ramírez-Gallego, Luengo, Benítez, & Herrera, 2016; Ramirez-Gallego et al., 2015). Así pues, ciertas técnicas presentan limitaciones, tal es el caso del

requerimiento de un valor de entrada equivalente al número de grupos que se desea crear, o, por otro lado, el número de elementos, esto es, el tamaño del grupo (Geaur Rahman & Zahidul Islam, 2016).

Los algoritmos de discretización más populares son los denominados no supervisados que constan de discretización por ancho igual y agrupación por frecuencia igual (Jiang & Zhao, 2012), sin embargo, por motivos de experimentación se utilizó dos técnicas adicionales conocidas como tamaño y, especificadas por el usuario como se indica a continuación:

Ancho igual

La técnica de ancho igual conocida como *binning* requiere como parámetro de entrada el número de grupos que se desea obtener, luego, analiza el valor máximo y mínimo del conjunto de datos con el propósito de fraccionar en el número de grupos especificado (Jiang & Zhao, 2012; Uysal & Gunal, 2018). De manera que, todos los rangos definidos posean la misma longitud de rango, sin embargo, existe la posibilidad que los grupos creados contengan un número variable de elementos entre sí (Ropero, Renooij, & Van Der Gaag, 2018). Dicho de otra manera, esta técnica distribuye los elementos de manera desequilibrada, por lo tanto, ciertos grupos contenderán demasiados elementos y, por otro lado, otros grupos pueden presentar ausencia de los mismos, razón por la cual se altera la construcción de una buena estructura de datos. Por ello, es considerado como buena práctica permitir que los rangos de los grupos sean de diferente longitud (Witten, Frank, & Hall, 2011).

Frecuencia igual

La discretización por frecuencia, de manera análoga a la técnica anterior recibe como parámetro de entrada el número de grupos que se desea obtener, así pues, a través de la técnica se divide el conjunto de datos en grupos basados en la distribución de elementos. Es decir, fracciona de tal manera que el número de elementos de los grupos sea equitativo entre sí (Jiang & Zhao, 2012). Por otro lado, es conocido como ecualización de histograma, por la forma en la que representa los resultados, en otras

palabras, la distribución de los elementos en cada uno de los rangos es ecuánime asemejándose a un histograma completamente plano (Witten et al., 2011).

Tamaño

Para la discretización a través de tamaño, es necesario definir un parámetro de entrada, es decir, el número de elementos que contendrá cada grupo. Así pues, la técnica agrupa los elementos en subconjuntos con el mismo número de datos especificados por el discretizador (Rapidminer, 2019).

Especificación de usuario

En ciertos casos, es necesario personalizar la discretización con el propósito de mejorar los resultados, por lo que, es necesario definir ciertas características de los grupos a generar. Por lo tanto, la técnica denominada especificación por usuario, funciona de tal manera que, el usuario especifica el límite superior de cada rango, puesto que, el límite inferior del mismo es determinado por el límite superior del rango anterior, así también, el límite inferior del rango inicial es el infinito negativo, de la misma manera, el infinito positivo puede ser o no utilizado como límite superior del último rango (Rapidminer, 2019).

4.5 Reglas de asociación

Las reglas de asociación, son utilizadas para descubrir elementos que se relacionan entre sí de manera frecuente. Por ejemplo, para determinar que productos compran los clientes a menudo de manera conjunta. Por ello, los artículos que se compran juntos se pueden colocar próximos entre ellos. Otra aplicación es el análisis de texto, es decir, clasificar y recuperar documentos (Ye, 2015).

La ventaja de los algoritmos de reglas de asociación, es que las asociaciones se pueden establecer entre cualesquiera de los atributos de un conjunto de datos. En otras palabras, buscan una serie de reglas, para cada una de las cuales se interpreta una conclusión diferente. Sin embargo, los algoritmos de asociación buscan patrones en grandes conjuntos de datos. Por ello, suelen ocupar un mayor tiempo de ejecución (IBM, 2019). Medidas como soporte,

confianza y *lift* son utilizados para encontrar subconjuntos de elementos de distintos conjuntos de elementos los cuales se detallan a continuación:

- **Soporte:**

Indica la proporción de registros de datos que contiene un conjunto (X) (Ye, 2015). Es decir, hace referencia al número de instancias en la que aparece A y B de manera conjunta en un conjunto de datos, dicho de otra manera, es la frecuencia con la que los elementos A y B se repiten correctamente (Aggarwal, 2013) .

- **Confianza:**

Es necesario recalcar que, si el soporte mide la frecuencia, la confianza mide la precisión de la regla de asociación. Así pues, la confianza de una regla A->B es el fraccionamiento de la frecuencia de los elementos A y B entre la frecuencia del elemento A, dicho de otra manera, la confianza se obtiene dividiendo el soporte(A->B) con respecto al soporte(A) como indica la Ecuación 1 (Aggarwal, 2013; Witten et al., 2011) .

$$Confianza(A \rightarrow B) = \frac{Soporte(A \rightarrow B)}{Soporte(A)} \quad (1)$$

- ***Lift***

El valor del *lift* es usado como medida de representación de una regla de asociación (McNicholas et al., 2008; Qin et al., 2015) . Esto quiere decir que, el *lift* interpreta distancia entre el soporte de una regla $(A \rightarrow B)$ fraccionado por su $(Soporte(A) * Soporte(B))$, en otras palabras, la magnitud en que A y B no son independientes entre sí, como se observa en la Ecuación 2 (McNicholas et al., 2008).

$$Lift(A \rightarrow B) = \frac{Soporte(A \rightarrow B)}{Soporte(A) * Soporte(B)} \quad (2)$$

Es necesario mencionar que el *lift* puede tomar ciertos valores, según (McNicholas et al., 2008) son los siguientes:

- Si el *lift* es superior a 1, indica que la regla de asociación aparece con frecuencia mayor a lo esperado lo que quiere decir que tiene un alto nivel de representación.
- Si el *lift* es cercano a 1, indica que la regla de asociación aparece de manera a lo esperado.
- Si el *lift* es inferior a 1, indica que la regla de asociación aparece con menor frecuencia a lo que se esperaba.

4.5.1 Apriori

El algoritmo Apriori es un procedimiento eficiente para formar conjuntos de elementos frecuentes. Por lo tanto, "Un conjunto de elementos puede ser un conjunto de elementos frecuentes, solo si todos sus subconjuntos son conjuntos de elementos frecuentes" (Ye, 2015).

Apriori inicialmente encuentra los *ítemsets* frecuentes, es decir, conjuntos de elementos que tienen mayor número de frecuencia, sin embargo, para que la generación de *itemsets* sea continua se debe establecer un umbral de soporte inicial, por ello, es necesario generar todos los *ítemsets* posibles y para cada uno de ellos definir su frecuencia, a continuación, los ítems candidatos serán generados en un inicio de 1 *ítem*, posteriormente de 2 *ítems* y así de manera continua, sin embargo, esta continua generación de conjunto de *ítems* demanda alto costo computacional. Por lo que, basados en el principio de montonocidad que menciona. Si un *itemset* es frecuente entonces todos los subconjuntos de este también son frecuentes. De manera análoga, si un *itemset* no es frecuente entonces cualquier conjunto que contenga a este *itemset* tampoco será frecuente, por ello, no es viable generar todos los *itemsets* posibles de una base de datos. Por lo tanto, la generación de *itemsets* es continua hasta que no existen *itemsets* que aprueben el soporte inicial definido, finalmente este proceso nos da como resultado un conjunto de datos con los ítems entre los cuales existe asociación (Ye, 2015).

5. METODOLOGÍA

Para realizar el presente estudio fue necesario seguir una serie de etapas que ayudarán a gestionar de mejor manera el proyecto. La etapa inicial se enfocó en la recopilación de los datos, etapa en la cual se almacenan datos acerca de componentes atmosféricos y variables meteorológicas en la ciudad de Cuenca-Ecuador. Posteriormente, se prepararon los datos, es decir, se rellenan datos si es que existiesen faltantes o simplemente se los descartan como valores atípicos que puedan posteriormente generar inconvenientes en la siguiente etapa. La discretización de los datos es fundamental en este estudio, por lo que, la manera en la que se discreticen los conjuntos de datos tanto de variables atmosféricas como de variables meteorológicas brindaron los resultados esperados. A continuación, se aplican técnicas de asociación, en este caso particular se utilizó la técnica Apriori, un algoritmo caracterizado por su facilidad de uso. Consecuentemente, se estableció métricas como soporte, confianza y lift para evaluar los resultados. Finalmente, se evaluaron los resultados en base a las métricas establecidas anteriormente, esto determina si es o no factible volver a discretizar los componentes atmosféricos o variables meteorológicas.

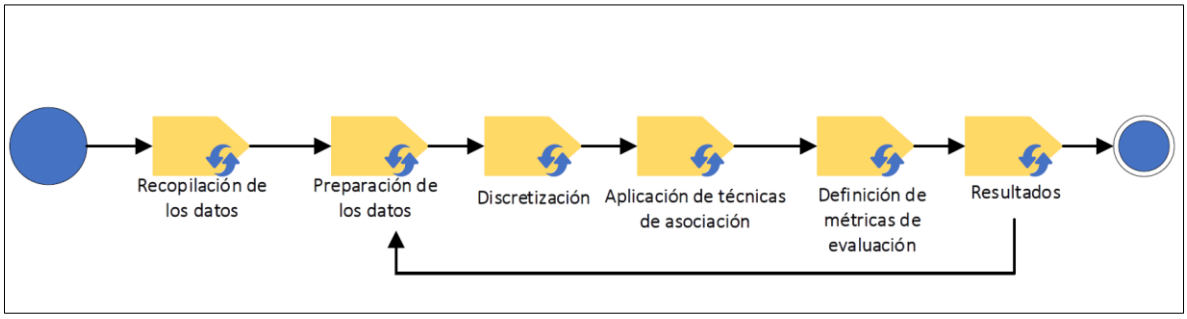

Ilustración 4: *Metodología de desarrollo del proyecto.*

5.1 Recopilación de los datos

Los datos recolectados a través de la Red de Monitoreo de Cuenca, registra datos a través de una estación automática instalada en el Centro Histórico. Esta infraestructura almacena tanto niveles de contaminación atmosférica como valores de variables meteorológicas en un rango de cuatro kilómetros a la redonda (Emov Ep, 2016). Los datos utilizados en el desarrollo del estudio competen a un periodo de tiempo de enero a noviembre

del año 2018, e involucran mediciones tomadas en rangos de 10 minutos cada uno. Los datos entregados por el Municipio de Cuenca son descritos en la sección de Marco Teórico.

5.2 Preparación de los datos.

Es importante mencionar que las estaciones de monitoreo encargadas de la recolección de los datos capturan valores erróneos por diferentes causas, tales como, errores en los sensores de las estaciones meteorológicas, interrupciones eléctricas o, por otro lado, que los niveles de contaminantes sean inferiores a los umbrales definidos para la recolección de los datos. Otro aspecto importante, es la recolección de contaminantes atmosféricos realizada en lapsos de 10 minutos para cada registro. No obstante, los registros pertenecientes a variables meteorológicas se recolectaban en lapso de 1 minuto, por lo que, genera discordancia y problemas al trabajar con el conjunto de datos. Así pues, en el estudio de (Andrade & Orellana, 2018) se da solución al problema seleccionando 10 registros pertenecientes a la misma fecha y hora de variables meteorológicas, posteriormente, se calcula el promedio de los registros seleccionados que, a su vez, son en un lapso de 10 minutos, de esta manera, coinciden tanto los registros de contaminantes atmosféricos como variables meteorológicas, finalmente, se realiza una vinculación entre los conjuntos de datos. Por otro lado, con el propósito de gestionar de mejor manera la preparación de los datos, se definió una secuencia de pasos a realizar como se muestra en la Ilustración 5.

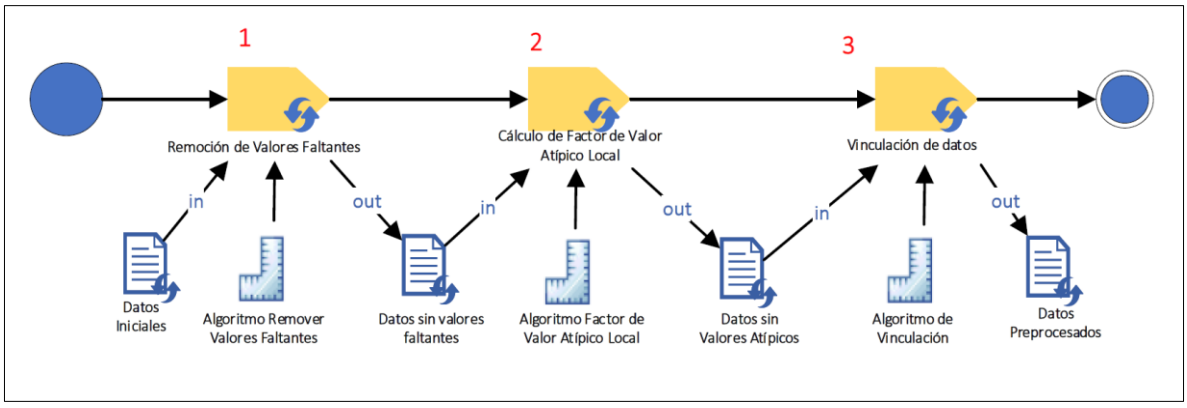

Ilustración 5: *Secuencia de procesos para la preparación de los datos.*

1. **Remoción de valores faltantes**

Datos Iniciales

Dentro del conjunto de datos existen atributos que tienen dependencia de otros, como es el caso del viento, el cual esta especificado en dos atributos: dirección del viento y velocidad del viento. Una medida de aplicación estaría en la confluencia de estos dos atributos, lo que determinaría un índice final. El objetivo de este estudio no corresponde a la creación de nuevos atributos determinados por otros, por tanto, la dirección del viento es descartada desde el inicio del tratamiento de los datos.

Algoritmo Remover Valores Faltantes

Para remover valores faltantes inicialmente se utilizó el operador *Select Attributes* que permitió seleccionar solo los atributos a utilizar en el desarrollo del estudio, por ello, el resto de atributos fueron eliminados descartando así la dirección del viento. Posteriormente, se utilizó el operador *Remove Missing Values* puesto que cuenta con la opción de seleccionar atributos sin valores faltantes.

Datos sin valores faltantes

Posterior a la selección de atributos se generó un conjunto de datos que satisfacen las condiciones especificadas anteriormente, tales como, remoción de faltantes y remoción de atributos que no son pertinentes para el desarrollo del estudio, de manera particular, la dirección del viento. Así también, es necesario mencionar que esta sección es importante dado que el resultado de este recurso se utilizó en etapas posteriores.

2. Cálculo de Factor de Valor Atípico Local

Algoritmo de Factor de Valor Atípico Local

Esta etapa del proceso fue considerada importante, dado que se identifican valores atípicos dentro del conjunto de datos los cuales afectan los resultados. Por ello, es necesario mencionar que tan solo los atributos de tipo numérico pueden presentar valores atípicos, por lo que, fue necesario realizar procedimientos previos con la finalidad de evitar errores en el uso de operadores. Es así que, el atributo hora de tipo *date* fue descartado en base a que causaba un error de formato en el operador de LOF, dado que este recibe como parámetro de entrada valores numéricos. Sin embargo, al

final del proceso existe una vinculación de datos donde se añade este atributo al conjunto de datos.

Así pues, con la ayuda del operador *Detect Outlier (LOF)* se calcularon estos valores. En primer lugar, se estableció los parámetros de entrada conocidos como valores *MinPts* que hacen referencia a los rangos: límite inferior y superior, definidos como 10 y 20 respectivamente, en otras palabras, son las distancias respectivas con las que se trabajó para realizar el cálculo. De esta manera, posterior al proceso de cálculo se obtuvo un nuevo conjunto de datos con una columna adicional nombrada *outlier* que contiene el factor de valor atípico calculado. Consecuentemente, se utilizó el operador *Filter Examples* con el propósito de descartar registros con valores atípicos, de manera que aquellos registros que no cumplieron con la condición de *outlier <= 1.6* fueron descartados.

Posterior a las etapas de remoción de faltantes y cálculo de LOF se obtuvieron 8,479 registros finales, que corresponde al 16% del total de los datos; esto quiere decir que el 84% de los datos se perdieron al momento de eliminar registros con campos faltantes y columnas con valores atípicos. Sin embargo, durante la experimentación se identificó que la mayor parte de los datos se pierden durante el proceso de remoción de faltantes como se indica en la Tabla 3. Lo que implica que al eliminar registros con campos faltantes se obtiene un total de 36,084 registros equivalente al 69% de los datos. Por otro lado, el proceso correspondiente a la depuración de los valores atípicos, a través de la técnica de LOF, dio como resultado la eliminación del 15% de los registros, de manera que, no es una perdida considerable de datos. No obstante, como se mencionó previamente gran cantidad de registros se perdían en la remoción de faltantes, por lo que, se decidió aplicar el método de (Peña, Ortega, & Orellana, 2019).

Tabla 3*: Limpieza de datos inicial.*

Mes	Número de datos iniciales	Número de datos sin valores faltantes	Número de elementos sin valores atípicos	Numero de Datos Resultantes	
				Número de datos iniciales	Número de datos limpios
Enero	4463	4085	929	4463	929

Febrero	4032	2880	594	4032	594
Marzo	4464	3795	850	4464	850
Abril	4320	4018	934	4320	934
Mayo	4464	3097	684	4464	684
Junio	4320	2188	465	4320	465
Julio	4464	4184	916	4464	916
Agosto	4464	4076	557	4464	557
Septiembre	4320	2644	589	4320	589
Octubre	4464	4076	933	4464	933
Noviembre	4320	754	178	4320	178
Diciembre	4464	287	58	4464	850
		36084	7687	52559	8479

El método de (Peña, Ortega, & Orellana, 2019) se enfoca en la ejecución de los siguientes pasos: En primer lugar, rellena los valores faltantes con ceros y, filtra datos definiendo un umbral mayor a cero. En segundo lugar, se aíslan cada una de las variables. En tercer lugar, se aplica el algoritmo LOF donde se establece el umbral correspondiente con el propósito de seleccionar valores que se encuentren bajo el umbral establecido. De manera más detallada, en el rellenado de faltantes utilizaron dos métodos regularizados, por una parte, la regresión Lasso y, por otro lado, la regresión de Ridge. Esto quiere decir que, inicialmente, definieron cierto número de puntos hacia adelante y hacia atrás dentro de un bloque de datos, de esta manera, se logra estimar el valor de un punto faltante dentro del conjunto de datos. Conviene subrayar que, la regresión de Lasso presenta superioridad frente a la regresión de Ridge en el caso de rellenado de datos faltantes. Así pues, el método seleccionado fue evaluado en base a la correlación (R2), error absoluto medio (MAE) y error cuadrático medio (RMSE) obteniendo buenos resultados en cuanto a términos de precisión y precisión entre espacios con distintas longitudes.

Finalmente, posterior a la aplicación del método (Peña et al., 2019) se obtuvo un nuevo conjunto de datos con un total de 52560 registros. Por lo que, se aplicó nuevamente el proceso de limpieza inicial como se especifica en la Tabla 4, de esta manera los resultados finales fueron 30622 registros totalmente limpios. Para ser más específicos, al eliminar valores faltantes se obtuvo 32517 registros que hace referencia al 62 % de

los datos, continuando con el proceso de limpieza se aplicó el algoritmo de LOF obteniendo 30622 registros totalmente limpios que se refiere al 58% de los datos. Cabe recalcar que la eliminación del 62% de los datos no afecta de manera directa al trabajo puesto que posterior a la limpieza se obtiene un conjunto de datos de calidad, sin datos faltantes y valores atípicos con los cuales se puede trabajar de manera óptima mejorando la calidad de los resultados.

Tabla 4: *Limpieza de datos final*.

Mes	Número de datos iniciales	Número de datos sin valores faltantes	Número de elementos sin valores atípicos	Numero de Datos Resultantes	
				Número de datos iniciales	Número de datos limpios
Enero	4464	3910	3707	4464	3707
Febrero	4032	2399	2284	4032	2284
Marzo	4464	3510	3388	4464	3388
Abril	4320	3666	3491	4320	3491
Mayo	4464	2738	2611	4464	2611
Junio	4320	2057	2022	4320	2022
Julio	4464	4464	3779	4464	3779
Agosto	4464	2163	2046	4464	2046
Septiembre	4320	2623	2492	4320	2492
Octubre	4464	3993	3873	4464	3873
Noviembre	4320	731	679	4320	679
Diciembre	4464	263	250	4464	250
		32517	30622	52560	30622

Datos sin Valores Atípicos

Finalmente, luego de aplicar los respectivos algoritmos tanto para la remoción de faltantes, así como también la eliminación valores atípicos, se obtuvo un conjunto de datos con 30622 registros que indica el 58% de los datos. Sin embargo, como se

mencionó anteriormente, fue necesario descartar la columna tipo *date*, dado que causaría un error de formato, por ello, en la siguiente etapa se da solución a esta operación realizada.

3. **Vinculación de datos**

Algoritmo de Vinculación

La siguiente etapa da solución a la desvinculación de la columna de tipo *date* que se realizó con el propósito de evitar errores de formato. Por lo tanto, a través del ID que contiene cada uno de los registros y con el uso del operador *Intersect* se logró unir los registros de datos preprocesados junto con los registros de datos iniciales correspondientes, así pues, para cada registro se adjuntó su registro de hora.

Datos Preprocesados

Posterior a los procesos de limpieza y vinculación de datos se logró unir los registros de datos preprocesados junto con los registros de datos iniciales correspondientes. Es así que finalmente se obtuvo un conjunto de datos con total de 30622 registros totalmente limpios con la finalidad de emplearlos en el modelado de datos.

5.3Modelado de datos

Como se mencionó en un inicio se utilizó como guía de desarrollo el modelo CRISP-DM, por ello, dentro del Modelado de Datos se incluye la etapa de Discretización y Aplicación de Técnicas de Asociación referentes a la metodología indicada. Así pues, en la Ilustración 6 se indica el modelo con cada una de las etapas que se utilizó para el desarrollo del estudio. No obstante, es necesario indicar de manera más específica cada uno de los operadores que se utilizaron en el transcurso del estudio que se indica en la sección de Anexos.

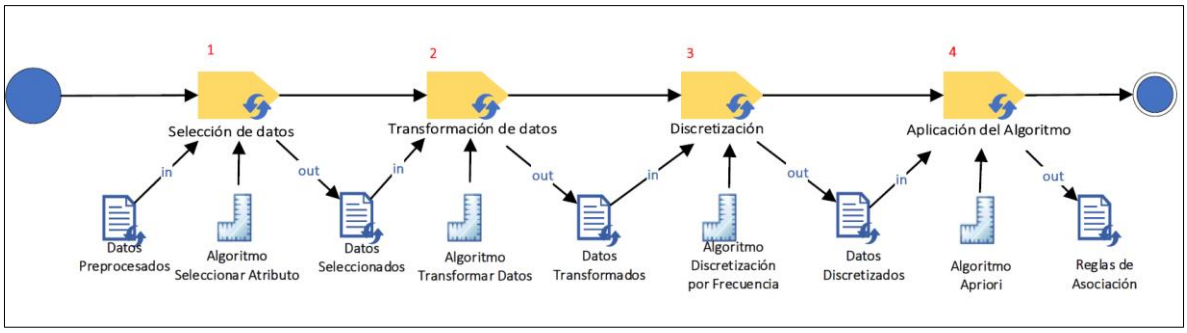

Ilustración 6: *Secuencia de procesos del desarrollo del proyecto.*

1. **Selección de los datos**

Algoritmo Seleccionar Atributo

De manera análoga a la etapa de preparación de los datos, se realizó una selección de atributos, dado que dentro del conjunto de datos existen atributos que tiene dependencia de otros, como es el caso de la dirección viento. Por lo que, a través del operador *Select Attributes*, se filtró solo atributos que nos servirán en el desarrollo del estudio, por lo tanto, el resto de atributos se eliminaron del conjunto de ejemplos descartando así la dirección del viento.

Posteriormente, durante la experimentación se decidió descartar los atributos de PRECIP_SUM (Precipitación), RADGLOBAL_AV (Radiación Global), a causa de que contienen demasiados valores iguales, es decir, si existen demasiados valores iguales dentro de una columna en la etapa de discretización el grupo categórico crece demasiado incluso doblando el tamaño máximo especificado por el usuario ocasionando errores. Del mismo modo, variables meteorológicas como UVA_AV (Radiación UVA), UVE_AV (Radiación UVE) se descartaron, dado que estas dos variables meteorológicas son similares lo que causaba un sesgo en los resultados finales, puesto que se presentaban en la mayoría de reglas de asociación.

Datos Seleccionados

Finalmente, la ejecución de los procesos anteriores da como resultado un conjunto de datos preprocesados, es así que en etapas futuras se prevé el riesgo de errores, como en la etapa de discretización, ya sea porque el formato de entrada es erróneo o, a su vez,

si existen grandes cantidades de valores iguales que afectan el proceso de discretización.

2. Transformación de los datos

Algoritmo Transformar Datos

Previo a la etapa de discretización, es necesario realizar una transformación de los datos, ya que es imposible discretizar atributos de tipo date, por lo tanto, fue necesario realizar un tratamiento previo de los datos. Ahora bien, el operador *Date to Numerical* nos permite seleccionar componentes de fecha u hora, tales como, hora, día, semana, trimestre. Por lo que, la extracción de estos componentes se lo realizó de manera relativa, dicho en otras palabras, es posible extraer el día relativo al mes, semana o incluso año. Ejemplificando, si se obtiene la fecha 25/abril/2020 el día relativo al mes seria 25, sin embargo, el día relativo al año seria 115 porque es el día 115 del año. De la misma manera, funciona la extracción de hora, por ello, para realizar el estudio se seleccionó la hora relativa al día. Finalmente, el resultado de esta transformación da como resultado atributos de tipo fecha convertidos en atributos de tipo numérico.

Datos Transformados

Como se mencionó en la etapa anterior el operador *Date to Numerical*, da como resultado un conjunto de datos procesados, es decir, los atributos de tipo fecha transformados en atributos de tipo numérico con los cuales se puede trabajar correctamente en la etapa de discretización.

3. Algoritmo de Discretización

La técnica de asociación se fundamenta en la agrupación de *items* que se presentan conjuntamente a través de varias transacciones. Cuando los *items* no corresponden a una unidad específica, sino al contrario son construidos por rangos que determinan esa unidad, es necesario utilizar una técnica denominada discretización. En el caso de las variables de contaminantes atmosféricos y variables meteorológicas se necesita conformar grupos con rangos determinados que representen estas unidades específicas.

Por ejemplo, el caso del ozono tiene una dispersión de datos numéricos en una unidad de tiempo, por lo que, es necesario establecer rangos entre el valor iniciar y final del contaminante para reducir el número de grupos representados; con esto, se podrá establecer la relación con otras variables. Para crear estos grupos, es necesario determinar el rango inicial y final de cada grupo, para lo cual existen varias técnicas que se pueden aplicar. Por ejemplo, si los rangos fuesen hitos proporcionales a la escala, un grupo podría contener más o menos registros que otro grupo, en otras palabras, los grupos estarían desbalanceados. Al contrario, si se determinan los rangos en función del número de registros, se obtendría un grupo más balanceado. Por lo tanto, es necesario describir las diferencias presentadas entre los diferentes tipos de discretización obtenidos como se muestra a continuación:

Ancho igual

En primer lugar, se discretizó a través de la técnica de *Binning*. Esta técnica se aplica a través de operador *Discretize by Binning*, el cual discretiza en base al número de subconjuntos especificado por el usuario. Se estableció ocho grupos de rangos de manera automática, sin embargo, el tamaño de cada grupo con respecto a otro es variable. En otras palabras, en los rangos iniciales existe gran afluencia de datos, no obstante, esta acumulación de datos fue disminuyendo, de tal manera, que el último rango cuenta con pocos elementos o ninguno. Para ilustrar mejor la distribución de elementos se muestra en la sección de anexos.

Tamaño

Con la finalidad de obtener mejores resultados se experimentó con la discretización por tamaño. Es así que, al tener 30.622 registros se fracciono por el número de registros por el número de grupos, en este caso ocho. De esta manera, se obtuvo un tamaño de 3.827elementos para cada grupo. Sin embargo, al generar los grupos tanto la distribución como el tamaño de elementos no fueron los esperados. En primer lugar, el operador de discretización creo nueve rangos, siendo los esperados ocho rangos. En segundo lugar, se observa que la distribución es equitativa en varios de los rangos, sin embargo, en el rango inicial, no se aprecian elementos. Esto quiere decir que, la

distribución de los datos no es la correcta o a su vez no es uniforme, por lo tanto, esta discretización también fue descartada.

Especificación de usuario

Por otra parte, se experimentó la discretización por especificación de usuario. No obstante, conviene subrayar que al especificar rangos muy grandes o muy pequeños sesga la información. Por lo cual, se estableció ocho rangos que abarquen la mayor cantidad de datos. Así, por ejemplo, el rango inicial [-∞,100] va desde el infinito negativo hasta 100, posteriormente se incrementó en intervalos de 100, finalizando en el rango ochocientos hasta el infinito [800, ∞]. Sin embargo, los resultados fueron aún peores que en los casos anteriores, es así que la distribución de los datos se centraba tan solo en los dos primeros rangos. Por lo que, similar a los casos anteriores esta técnica también fue descartada.

Frecuencia Igual

Luego de las experimentaciones realizadas con cada una de las técnicas de discretización, se seleccionó la discretización por frecuencia. Por lo tanto, con la ayuda del operador *Discretize by Frequency*, se estableció el número de grupos a crear, en este caso ocho grupos de rangos. Como se puede ver en la Ilustración 7 para el caso del ozono los grupos son balanceados. Similar comportamiento se logró el resto de variables de estudio.

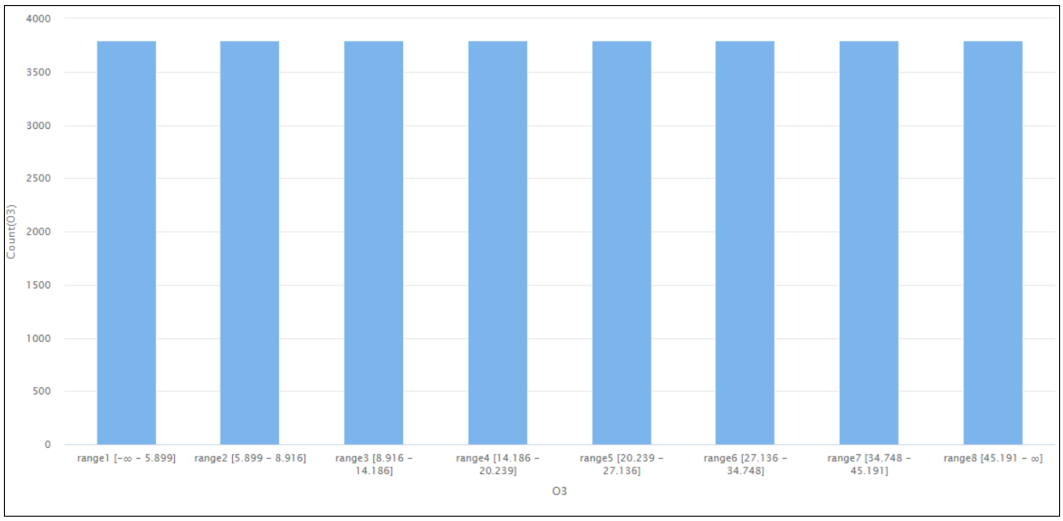

Ilustración 7: *Discretización por Frecuencia*.

Datos Transformados

La técnica de asociación se fundamenta en la agrupación de *itemsets* que se presentan conjuntamente a través de varias transacciones. Por ello, en el caso de las variables de contaminantes atmosféricos y variables meteorológicas fue necesario conformar grupos con rangos determinados que representen la mayor cantidad de información posible. Es así que, finalizada la etapa de discretización se obtuvo un conjunto de datos correctamente discretizado para la aplicación reglas de asociación.

4. Aplicación del Algoritmo

Algoritmo Apriori

Conviene mencionar el algoritmo Apriori brinda un procedimiento eficiente para formar conjuntos de elementos frecuentes. Por ello, se aplicó el operador *FP-Growth* que cumple la misma funcionalidad que Apriori dentro de RapidMiner. Sin embargo, para reducir el conjunto de elementos frecuentes, este operador recibe como parámetro un soporte mínimo de frecuencias, es decir, las ocurrencias frecuentes y no tan frecuentes. Así pues, luego de aplicar el algoritmo de elementos frecuentes fue necesario aplicar el operador *Create Association Rules* para generar las reglas de asociación. No obstante, para limitar el número de resultados obtenidos fue necesario

modular las métricas mencionadas en un inicio, tales como, soporte, confianza y *lift* que se encuentran descritas con mayor detalle en la sección de Evaluación.

Reglas de Asociación

La Tabla 7 muestra los resultados de la aplicación del algoritmo Apriori. La primera columna indica el numero de la regla, la segunda columna señala la premisa del resultado y finalmente se indica la conclusión de la regla en la tercera columna. Es importante indicar en cada regla los parámetros que justifican la regla, por lo tanto, se indica el soporte, confianza y lift respectivamente.

5.4Evaluación

La presente etapa se enfoca en la definición de métricas de evaluación con la finalidad de limitar y evaluar los resultados obtenidos. Así pues, como indica (Witten et al., 2011) al aplicar técnicas de asociación es posible generar grandes cantidades de reglas, sin embargo, el enfoque se lo debe realizar en aquellas que presenten un alto índice de confianza. Así también, como parte de la evaluación (Aggarwal, 2013) menciona que una regla de asociación es válida si cumple con criterios de soporte y confianza. Por ello, con el propósito obtener resultados de calidad las reglas obtenidas fueron evaluadas en base a tres métricas que son, soporte, confianza y *lift*.

En primer lugar, se estableció un soporte del 50%, confianza del 80% y *lift* de 80, no obstante, al ser un soporte demasiado elevado, no se obtuvo ningún resultado. En segundo lugar, se experimentó con soporte del 10%, junto a una confianza y *lift* del 80% respectivamente, así pues, se obtuvieron dos resultados como se observa en la Tabla 5. Si bien estos resultados demuestran resultados interesantes, no fueron suficientes para representar el conocimiento del conjunto de elementos.

Tabla 5: *Segunda evaluación reglas de asociación.*

No.	Premisa	Conclusión	Soporte	Confianza	*Lift*
1	TEMPAIRE_AV = range8 [18.950 - ∞], WINDSPEED_AV = range8 [3.150 - ∞]	HR_AV = range1 [-∞ - 44.500]	4%	81%	6.00
2	O3 = range1 [-∞ - 5.899], HR_AV = range8 [86.500 - ∞]	DP_AV = range8 [10.850 - ∞]	4%	81%	7.00

Para mejorar los resultados se estableció un soporte del 5%, una confianza del 80% y un lift de 80%. Los resultados se muestran en la Tabla 6. Si bien estos resultados demuestran reglas de asociación fuertes no son suficientes para representar el conocimiento de la base de datos, por lo que, posteriormente se disminuyó el soporte.

Tabla 6:*Tercera evaluación reglas de asociación.*

No.	Premisa	Conclusión	Soporte	Confianza	Lift
1	HR_AV = range1 [-∞ - 44.500], O3 = range8 [45.191 - ∞]	TEMPAIRE_AV = range8 [18.950 - ∞]	6%	81%	7
2	HORA = range5 [12.500 - 15.500], TEMPAIRE_AV = range8 [18.950 - ∞]	HR_AV = range1 [-∞ - 44.500]	5%	82%	7
3	HR_AV = range1 [-∞ - 44.500], HORA = range5 [12.500 - 15.500]	TEMPAIRE_AV = range8 [18.950 - ∞]	5%	83%	7
4	O3 = range8 [45.191 - ∞], TEMPAIRE_AV = range8 [18.950 - ∞]	HR_AV = range1 [-∞ - 44.500]	6%	85%	7

En base a los resultados anteriormente descritos, se modificó el parámetro de soporte a un 3% y el lift a 80%. A diferencia de la experimentación anterior que arrojó dos resultados reglas de asociación, el resultado luego del ajuste produjo diecisiete reglas.

6. RESULTADOS

Como se observa en la Ilustración 8, los resultados tienden a agruparse por rangos, es decir, como se observa los puntos de coloración amarilla representan las reglas que, si bien cumplen con el umbral establecido de un 3% presentan un bajo índice de confianza con una oscilación por debajo del 40%. Por otro lado, las reglas de coloración verde presentan incluso en ciertos casos un soporte superior al 5%, sin embargo, la confianza se encuentra por debajo del 60%. Por lo tanto, el estudio realizado se enfocó en las reglas de coloración azul dado que cumplen tanto con el umbral de soporte del 3% junto a una confianza del 80% o más.

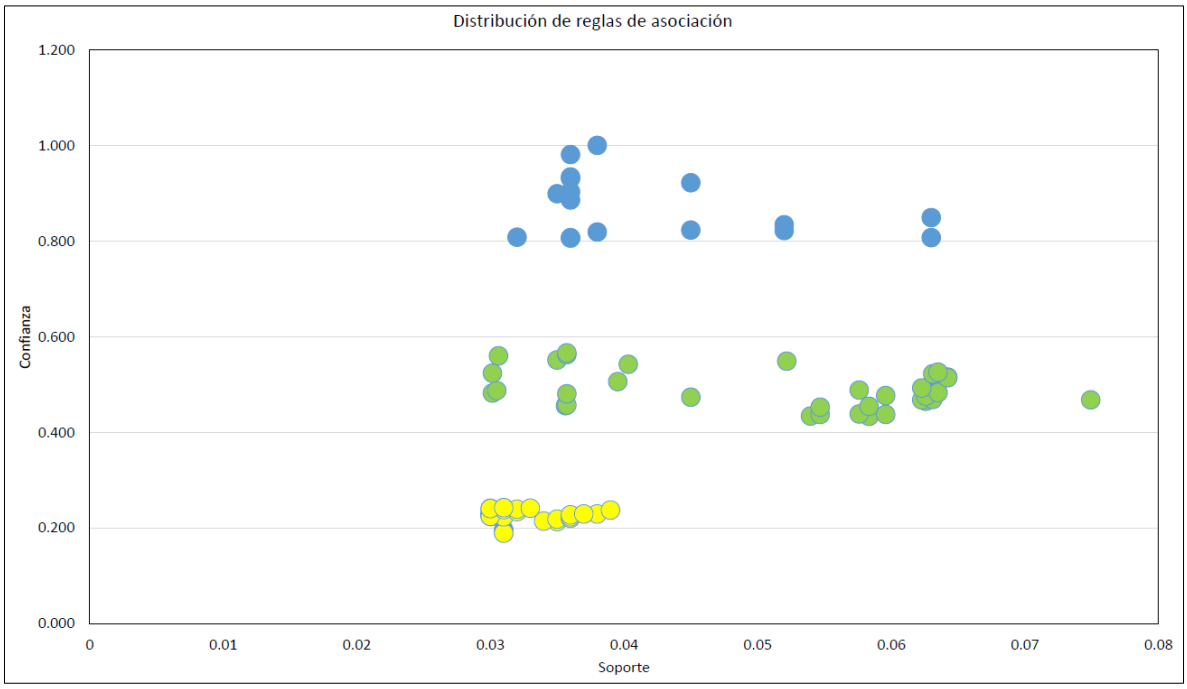

Ilustración 8: *Distribución de los resultados en base a soporte y frecuencia.*

Una vez definida la sección de evaluación es necesario describir e interpretar los resultados obtenidos en cada una de las diferentes reglas de asociación. Es necesario mencionar que los resultados obtenidos, por un lado, presentan la relación entre las variables meteorológicas y por otro lado, la relación entre contaminantes y variables meteorológicas. Así pues, los resultados se observan tanto en la Ilustración 9 y de manera más detallada en la Tabla 7. Mayor especificidad se puede encontrar en la sección anexos.

Tabla 7:*Resultados del operador Create Association Rules*.

No.	Premisa	Conclusión	Soporte	Confianza	*Lift*
1	TEMPAIRE_AV = range8 [18.950 - ∞], WINDSPEED_AV = range8 [3.150 - ∞]	HR_AV = range1 [-∞ - 44.500]	4%	81%	6.00
2	O3 = range1 [-∞ - 5.899], HR_AV = range8 [86.500 - ∞]	DP_AV = range8 [10.850 - ∞]	4%	81%	7.00
3	HR_AV = range1 [-∞ - 44.500], O3 = range8 [45.191 - ∞]	TEMPAIRE_AV = range8 [18.950 - ∞]	6%	81%	7.00
4	HR_AV = range1 [-∞ - 44.500], HORA = range4 [9.500 - 12.500]	O3 = range8 [45.191 - ∞]	3%	81%	6.00
5	TEMPAIRE_AV = range2 [11.850 - 12.750], HR_AV = range8 [86.500 - ∞]	DP_AV = range8 [10.850 - ∞]	4%	82%	7.00
6	HORA = range5 [12.500 - 15.500], TEMPAIRE_AV = range8 [18.950 - ∞]	HR_AV = range1 [-∞ - 44.500]	5%	82%	6.00
7	PATM_AV = range1 [-∞ - 749.328], TEMPAIRE_AV = range8 [18.950 - ∞]	HR_AV = range1 [-∞ - 44.500]	5%	82%	6.00
8	HR_AV = range1 [-∞ - 44.500], HORA = range5 [12.500 - 15.500]	TEMPAIRE_AV = range8 [18.950 - ∞]	5%	83%	7.00
9	O3 = range8 [45.191 - ∞], TEMPAIRE_AV = range8 [18.950 - ∞]	HR_AV = range1 [-∞ - 44.500]	6%	85%	6.00
10	HORA = range5 [12.500 - 15.500], O3 = range8 [45.191 - ∞], TEMPAIRE_AV = range8 [18.950 - ∞]	HR_AV = range1 [-∞ - 44.500]	4%	89%	7.00
11	WINDSPEED_AV = range1 [-∞ - 0.750], DP_AV = range8 [10.850 - ∞]	HR_AV = range8 [86.500 - ∞]	4%	90%	7.00
12	HR_AV = range1 [-∞ - 44.500], HORA = range5 [12.500 - 15.500], O3 = range8 [45.191 - ∞]	TEMPAIRE_AV = range8 [18.950 - ∞]	4%	90%	7.00
13	HR_AV = range1 [-∞ - 44.500], PATM_AV = range1 [-∞ - 749.328]	TEMPAIRE_AV = range8 [18.950 - ∞]	5%	92%	8.00
14	O3 = range1 [-∞ - 5.899], DP_AV = range8 [10.850 - ∞]	HR_AV = range8 [86.500 - ∞]	4%	93%	8.00
15	DP_AV = range1 [-∞ - 5.250], O3 = range8 [45.191 - ∞]	HR_AV = range1 [-∞ - 44.500]	4%	93%	7.00
16	HORA = range2 [3.500 - 7.500], DP_AV = range8 [10.850 - ∞]0	HR_AV = range8 [86.500 - ∞]	4%	98%	8.00
17	TEMPAIRE_AV = range2 [11.850 - 12.750], DP_AV = range8 [10.850 - ∞]	HR_AV = range8 [86.500 - ∞]	4%	100 %	8.00

AssociationRules

```
Association Rules
[TEMPAIRE_AV = range8 [18.950 - ∞], WINDSPEED_AV = range8 [3.150 - ∞]] --> [HR_AV = range1 [-∞ - 44.500]] (confidence: 0.806)
[O3 = range1 [-∞ - 5.899], HR_AV = range8 [86.500 - ∞]] --> [DP_AV = range8 [10.850 - ∞]] (confidence: 0.807)
[HR_AV = range1 [-∞ - 44.500], O3 = range8 [45.191 - ∞]] --> [TEMPAIRE_AV = range8 [18.950 - ∞]] (confidence: 0.807)
[HR_AV = range1 [-∞ - 44.500], HORA = range4 [9.500 - 12.500]] --> [O3 = range8 [45.191 - ∞]] (confidence: 0.808)
[TEMPAIRE_AV = range2 [11.850 - 12.750], HR_AV = range8 [86.500 - ∞]] --> [DP_AV = range8 [10.850 - ∞]] (confidence: 0.819)
[HORA = range5 [12.500 - 15.500], TEMPAIRE_AV = range8 [18.950 - ∞]] --> [HR_AV = range1 [-∞ - 44.500]] (confidence: 0.822)
[PATM_AV = range1 [-∞ - 749.328], TEMPAIRE_AV = range8 [18.950 - ∞]] --> [HR_AV = range1 [-∞ - 44.500]] (confidence: 0.823)
[HR_AV = range1 [-∞ - 44.500], HORA = range5 [12.500 - 15.500]] --> [TEMPAIRE_AV = range8 [18.950 - ∞]] (confidence: 0.834)
[O3 = range8 [45.191 - ∞], TEMPAIRE_AV = range8 [18.950 - ∞]] --> [HR_AV = range1 [-∞ - 44.500]] (confidence: 0.849)
[HORA = range5 [12.500 - 15.500], O3 = range8 [45.191 - ∞], TEMPAIRE_AV = range8 [18.950 - ∞]] --> [HR_AV = range1 [-∞ - 44.500]] (confidence: 0.886)
[WINDSPEED_AV = range1 [-∞ - 0.750], DP_AV = range8 [10.850 - ∞]] --> [HR_AV = range8 [86.500 - ∞]] (confidence: 0.899)
[HR_AV = range1 [-∞ - 44.500], HORA = range5 [12.500 - 15.500], O3 = range8 [45.191 - ∞]] --> [TEMPAIRE_AV = range8 [18.950 - ∞]] (confidence: 0.903)
[HR_AV = range1 [-∞ - 44.500], PATM_AV = range1 [-∞ - 749.328]] --> [TEMPAIRE_AV = range8 [18.950 - ∞]] (confidence: 0.922)
[O3 = range1 [-∞ - 5.899], DP_AV = range8 [10.850 - ∞]] --> [HR_AV = range8 [86.500 - ∞]] (confidence: 0.931)
[DP_AV = range1 [-∞ - 5.250], O3 = range8 [45.191 - ∞]] --> [HR_AV = range1 [-∞ - 44.500]] (confidence: 0.934)
[HORA = range2 [3.500 - 7.500], DP_AV = range8 [10.850 - ∞]] --> [HR_AV = range8 [86.500 - ∞]] (confidence: 0.981)
[TEMPAIRE_AV = range2 [11.850 - 12.750], DP_AV = range8 [10.850 - ∞]] --> [HR_AV = range8 [86.500 - ∞]] (confidence: 1.000)
```

Ilustración 9:*Reglas de asociación.*

La primera regla de asociación que se muestra a continuación, representa una oscilación de temperatura en un rango de 18.9°C o más, con una velocidad del viento en un rango de 3.1 m/h o superior. Esto en consecuencia produce vapor de agua en el aire con valores por debajo del 44.5% como se ilustra en la Ilustración 10. Es lógico deducir que la humedad se reduce conforme el calor aumenta y el viento se incrementa.

- [TEMPAIRE_AV = range8 [18.950 - ∞], WINDSPEED_AV = range8 [3.150 - ∞]] →[HR_AV = range1 [-∞ - 44.500]]

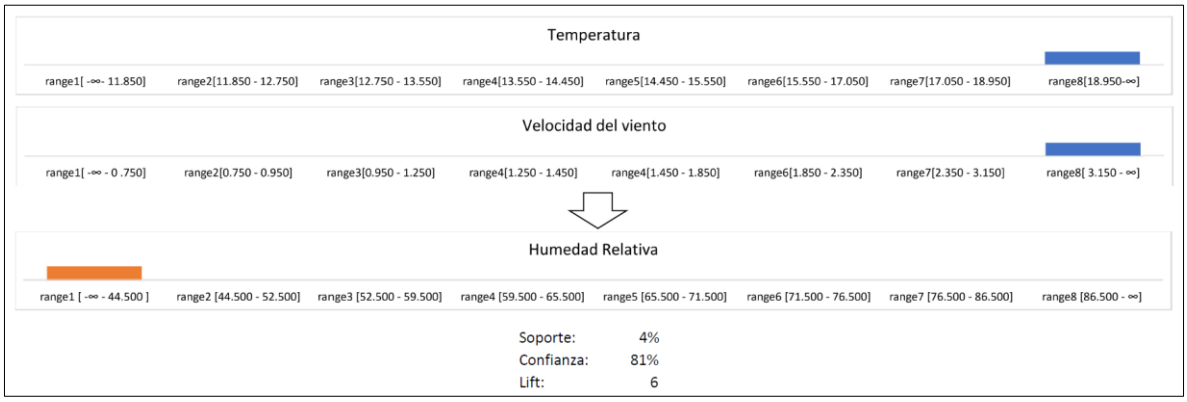

Ilustración 10*: Rangos de regla de asociación número 1.*

La segunda regla de asociación indica que si el ozono se encuentra por debajo de los 5.9 microgramos por metro cúbico y la humedad relativa es superior al 86.5%. Entonces, el contenido de vapor de agua presente en un gas supera los 10.85 °C. Por lo tanto, si el ozono disminuye a proporciones bajas y a su vez la humedad aumenta a su más rango elevado, el punto de roció se incrementa como se indica en la Ilustración 11.

- [O3 = range1 [-∞ - 5.899], HR_AV = range8 [86.500 - ∞]] →[DP_AV = range8 [10.850 - ∞]]

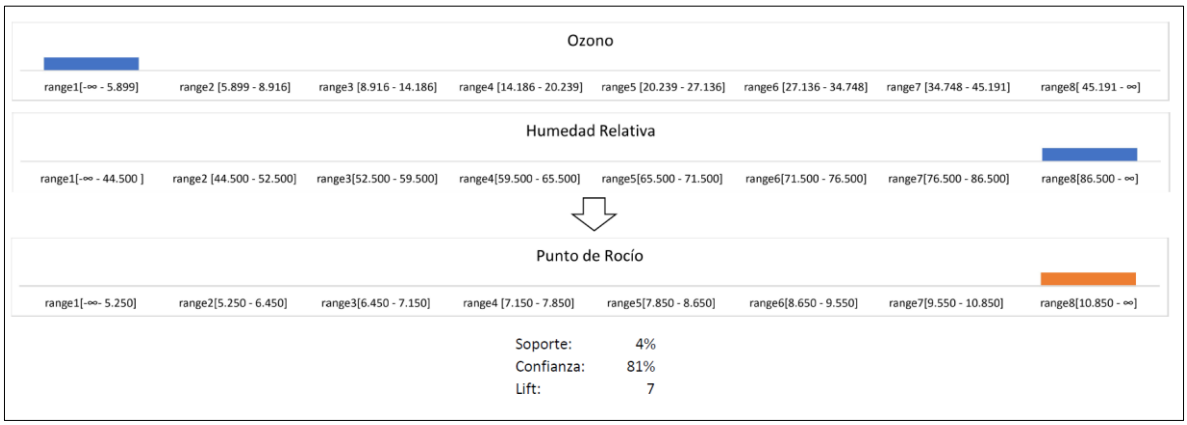

Ilustración 11*: Rangos de regla de asociación número 2.*

La tercera regla de asociación muestra que si en existe vapor de agua por debajo del 44.5% acompañado de ozono por encima de los 45 microgramos por metro cúbico, causa que la temperatura se eleve por encima de los 18.95° C. Por ello, si la humedad baja a su rango inferior y el ozono se incrementa, entonces el calor se eleva su rango superior como indica la Ilustración 12.

- [HR_AV = range1 [-∞ - 44.500], O3 = range8 [45.191 - ∞]]→[TEMPAIRE_AV = range8 [18.950 - ∞]]

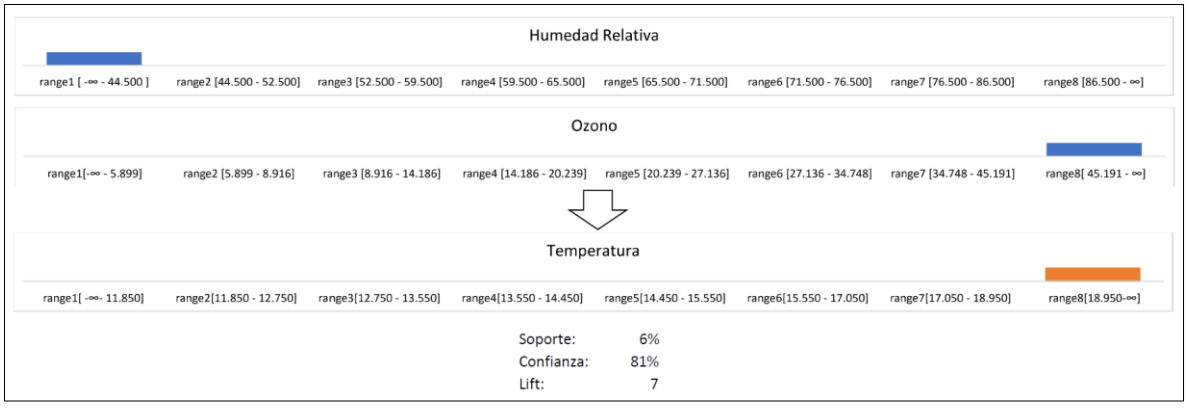

Ilustración 12*: Rangos de regla de asociación número 3.*

En cuanto a la cuarta regla muestra que entre las 9:50am - 12:50pm acompañado por un porcentaje de vapor de agua por debajo del 44.5%, infieren en un ozono por encima de los 45.19 microgramos por metro cúbico. Por ello, se deduce que a media mañana e inicio de

la tarde, si existe un bajo nivel de humedad el ozono se incrementa a su nivel más alto como muestra la Ilustración 13.

- [HR_AV = range1 [-∞ - 44.500], HORA = range4 [9.500 – 12.500]] →[O3 = range8 [45.191 - ∞]]

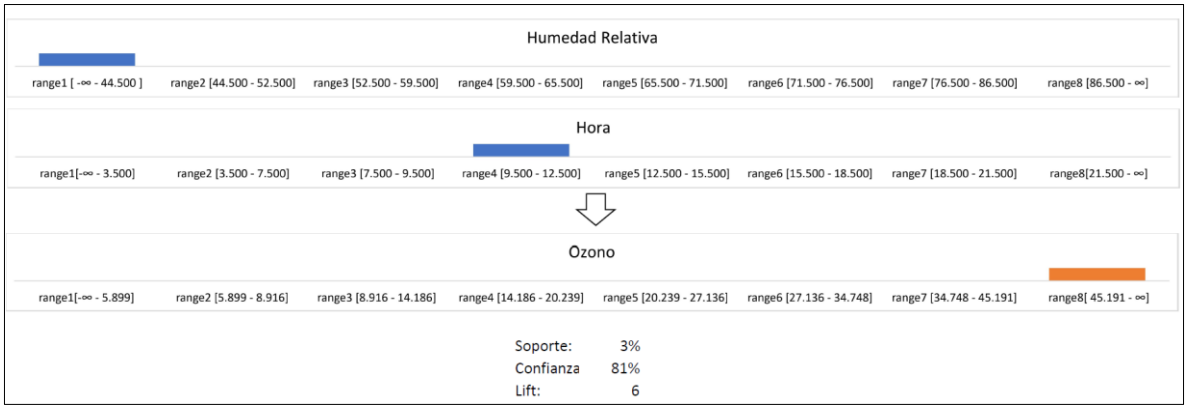

Ilustración 13: *Rangos de regla de asociación número 4.*

La quinta regla indica que una oscilación de temperatura entre los 11.85°C y los 12.75°C acompañado de humedad relativa de 86.5% o superior, genera un punto de roció entre los 10.85°C o más. Lo que quiere decir que, cuando existe alta presencia de humedad acompañada del frio del entorno, el punto de roció se eleva hacia su rango más alto como se indica en la Ilustración 14.

- [TEMPAIRE_AV = range2 [11.850 - 12.750], HR_AV = range8 [86.500 - ∞]] → [DP_AV = range8 [10.850 - ∞]]

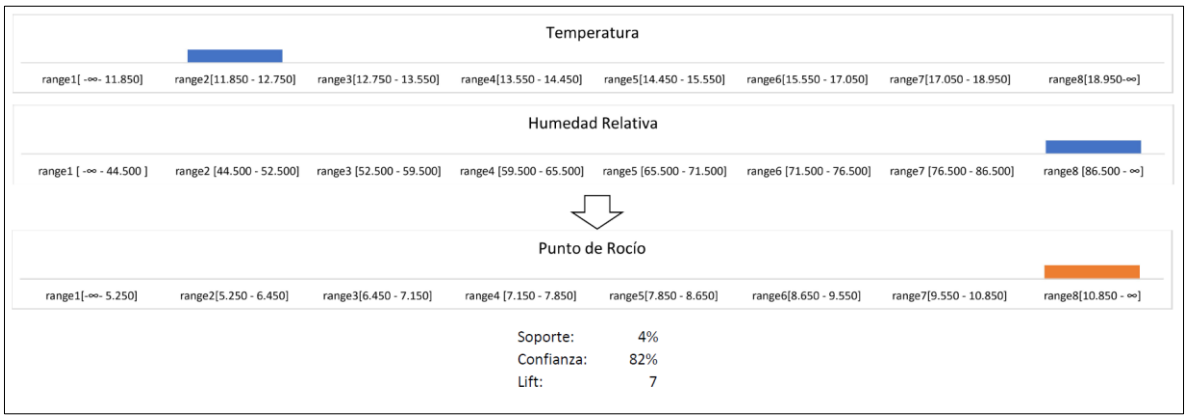

Ilustración 14*: Rangos de regla de asociación número 5.*

La sexta regla de asociación señala que dentro del horario entre las 12:50 pm y las 15:50 pm frente a una temperatura entre los 18.95°C o más, determina una humedad relativa por debajo del 44.5%. Por lo tanto, como indica en la Ilustración 15 si la temperatura aumenta durante la tarde, la humedad disminuye a su nivel más bajo.

- [HORA = range5 [12.500 - 15.500], TEMPAIRE_AV = range8 [18.950 - ∞]] →[HR_AV = range1 [-∞ - 44.500]]

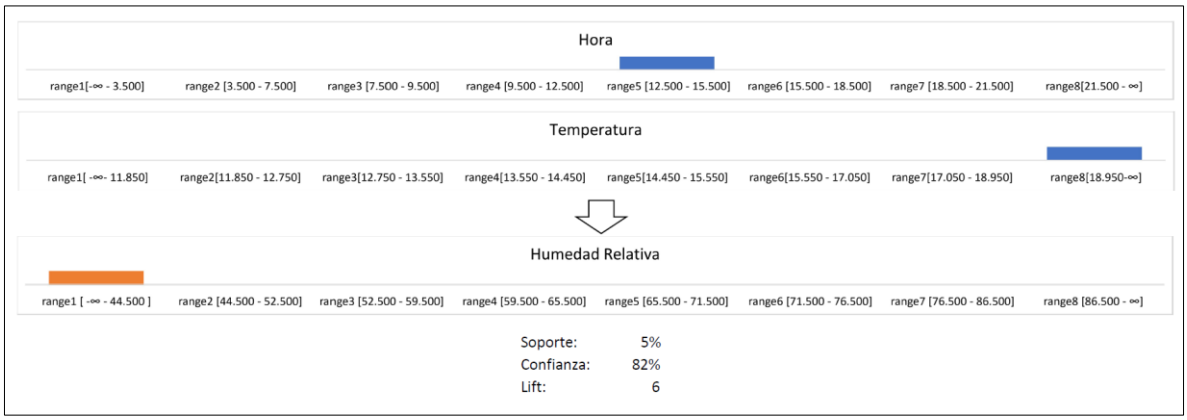

Ilustración 15: *Rangos de regla de asociación número 6.*

La séptima regla de asociación muestra que una oscilación de presión atmosférica por debajo de los 749.32 hectopascales junto a una temperatura por encima de los 18.95°C, infieren en una humedad relativa por debajo del 44.5%. Lo que determina que, si la fuerza ejercida por el peso disminuye y por el contrario el calor aumenta, la humedad se reduce como se observa en Ilustración 16.

- [PATM_AV = range1 [-∞ - 749.328], TEMPAIRE_AV = range8 [18.950 - ∞]] →[HR_AV = range1 [-∞ - 44.500]]

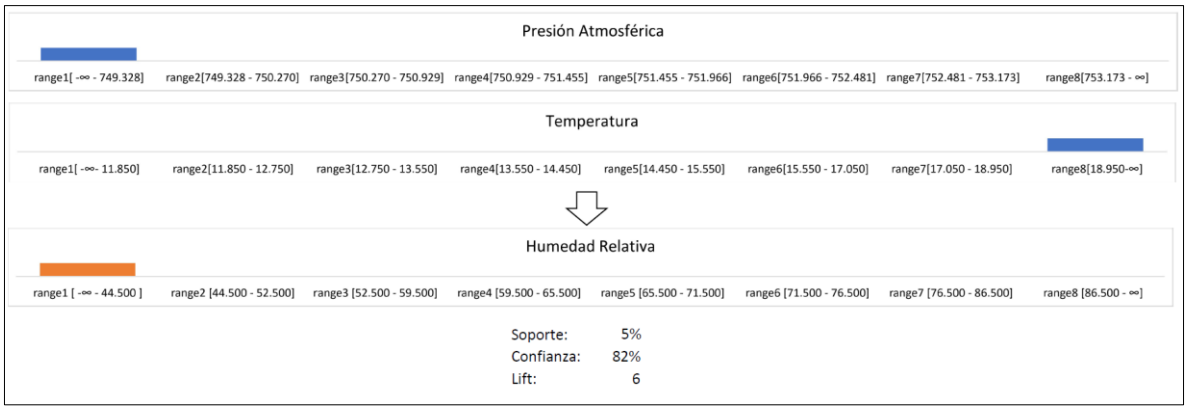

Ilustración 16: *Rangos de regla de asociación número 7.*

La octava regla de asociación señala que entre las 12:50 pm y 15:50 pm junto a una humedad relativa por debajo de los 44.5% determina en una temperatura por encima de los 18.95°C. Por ello, a inicios de la tarde con un bajo nivel de humedad, el calor se eleva a su nivel más alto como se aprecia en la Ilustración 17.

- [HR_AV = range1 [-∞ - 44.500], HORA = range5 [12.500 - 15.500]] →[TEMPAIRE_AV = range8 [18.950 - ∞]]

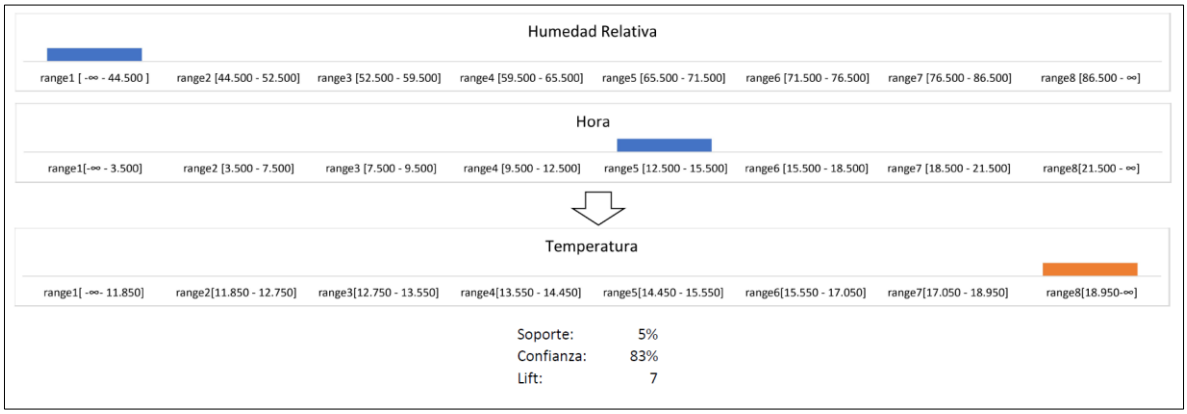

Ilustración 17: *Rangos de regla de asociación número 8.*

La novena regla muestra que el ozono por encima de los 45.2 microgramos por metro cúbico junto a una temperatura de 19°C o superior dan como resultado una humedad relativa inferior al 44.5%. Por lo tanto, se concluye en que si existe un alto nivel de ozono y una alta temperatura, la humedad disminuye como se observa en la Ilustración 18.

- [O3 = range8 [45.191 - ∞], TEMPAIRE_AV = range8 [18.950 - ∞]]→[HR_AV = range1 [-∞ - 44.500]]

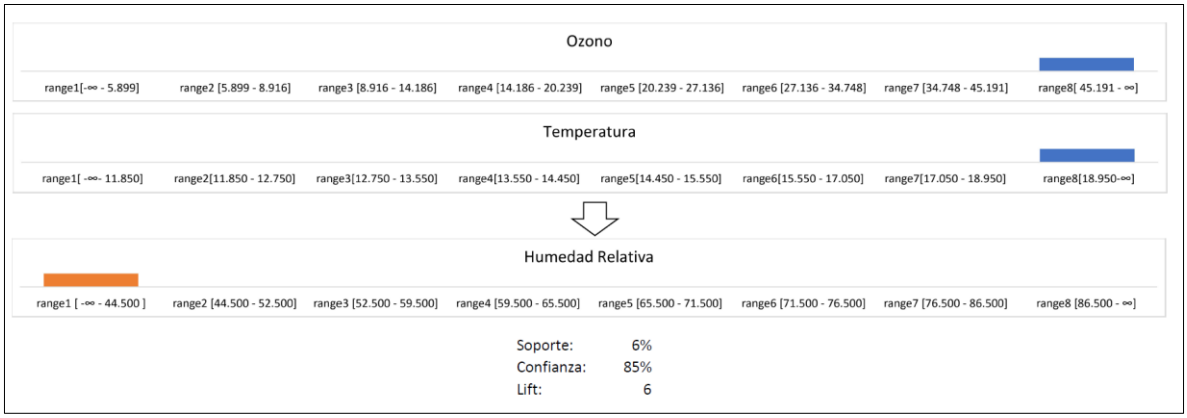

Ilustración 18*: Rangos de regla de asociación número 9.*

La décima regla indica que durante las 12:50 pm y 15:50 pm, frente a un ozono superior a los 45.2 microgramos por metro cúbico y una temperatura de 19°C o más, infieren en una humedad relativa por debajo de los 44.5%. Por lo tanto, como indica la Ilustración 19 durante la tarde con un alto grado de calor junto al contaminante ozono, la humedad disminuye.

- [HORA = range5 [12.500 - 15.500], O3 = range8 [45.191 - ∞], TEMPAIRE_AV = range8 [18.950 - ∞]]→[HR_AV = range1 [-∞ - 44.500]]

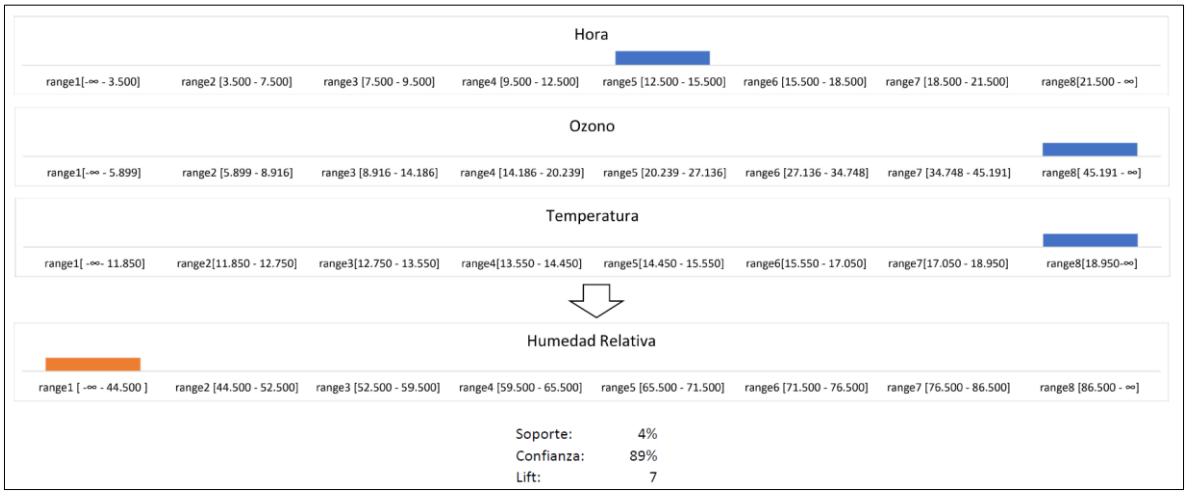

Ilustración 19: *Rangos de regla de asociación número 10.*

La undécima regla de asociación indica que cuando la velocidad del viento se encuentra por debajo de los 0.8 metros por hora y el punto de roció por encima de los 10.9 °C, causan una humedad del 86.5% o más. Por ello, cuando la velocidad del viento disminuye y el punto de roció se eleva, la humedad se incrementa a su nivel más alto como muestra la Ilustración 20.

- [WINDSPEED_AV = range1 [-∞ - 0.750], DP_AV = range8 [10.850 - ∞]] →[HR_AV = range8 [86.500 - ∞]]

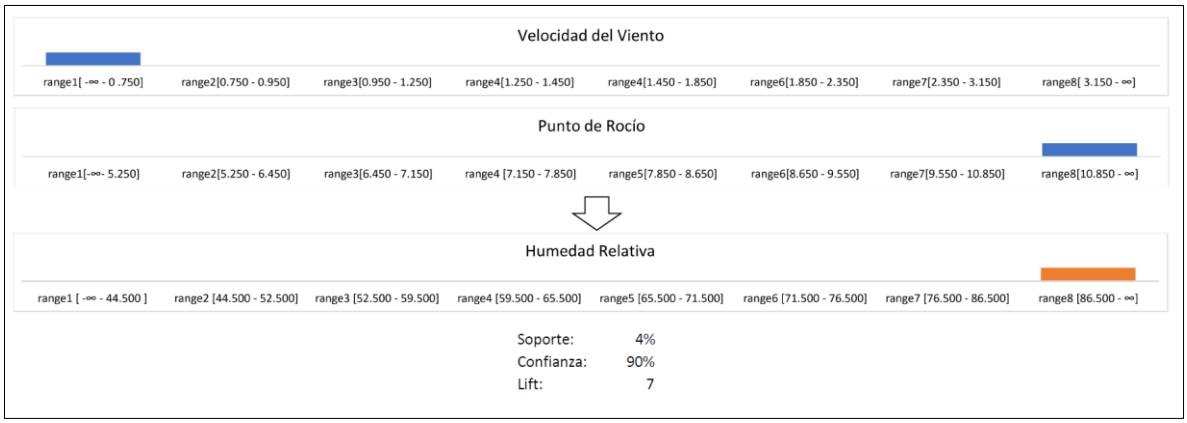

Ilustración 20: *Rangos de regla de asociación número 11.*

La duodécima regla de asociación muestra que una humedad relativa por debajo del 44.5% junto a un ozono por encima de los 45.2 microgramos por metro cúbico entre las 12:50 pm y 15:50 pm, dan como resultado una temperatura de 18.9°C o más. La Ilustración 21 indica que cuando la humedad relativa disminuye y el ozono se eleva durante la tarde, la temperatura aumenta a su rango superior.

- [HR_AV = range1 [-∞ - 44.500], HORA = range5 [12.500 - 15.500], O3 = range8 [45.191 - ∞]] →[TEMPAIRE_AV = range8 [18.950 - ∞]]

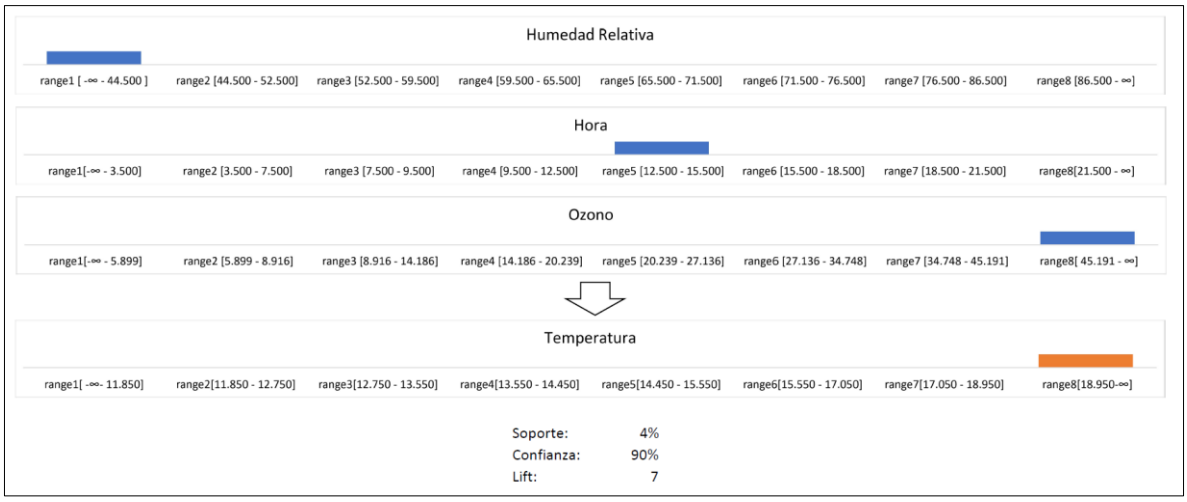

Ilustración 21: *Rangos de regla de asociación número 12.*

En cuanto a la decimotercera regla de asociación muestra que si existe humedad relativa por debajo del 44.5% y una presión atmosférica inferior a los 749.3 hectopascales, entonces la temperatura radica en los 19°C o más. En la Ilustración 22 se deduce que cuanto la humedad y la presión atmosférica se encuentren en sus rangos inferiores, el calor se eleva a su rango más alto.

- [HR_AV = range1 [-∞ - 44.500], PATM_AV = range1 [-∞ - 749.328]] --> [TEMPAIRE_AV = range8 [18.950 - ∞]]

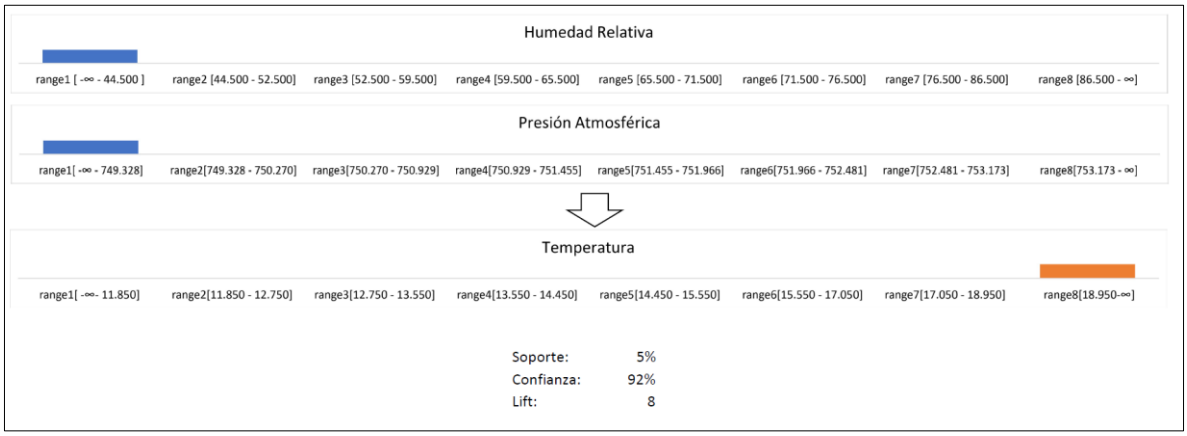

Ilustración 22: *Rangos de regla de asociación número 13.*

La decimocuarta regla de asociación indica que un ozono por debajo de los 5.9 microgramos por metro cúbico y un punto de roció superior a los 10.9 °C, genera como resultado una humedad relativa por encima del 86.5%. Es así que, cuando el ozono disminuye

y por el contrario el punto de roció se eleva, entonces la humedad relativa se incrementa como se muestra en la Ilustración 23.

- [O3 = range1 [-∞ - 5.899], DP_AV = range8 [10.850 - ∞]] →[HR_AV = range8 [86.500 - ∞]]

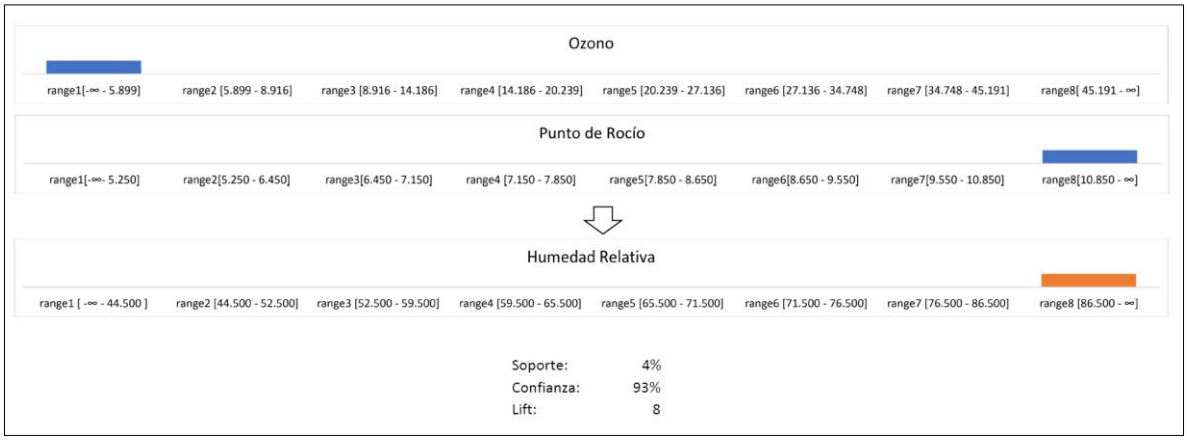

Ilustración 23*: Rangos de regla de asociación número 14.*

La decimoquinta regla de asociación es similar a la regla decimocuarta dado que, cuenta con los mismos componentes, sin embargo, los rangos en los que se infieren son diferentes es así que cuando el punto de roció se encuentra en los 5.25°C o menos junto a un ozono por encima de los 45.2 microgramos por metro cúbico, infieren en una humedad relativa por debajo del 44.5%. En la Ilustración 23 se observa que si se eleva el rango de temperatura y a su vez se disminuye el punto de rocío, entonces se reduce la humedad relativa.

- [DP_AV = range1 [-∞ - 5.250], O3 = range8 [45.191 - ∞]] →[HR_AV = range1 [-∞ - 44.500]]

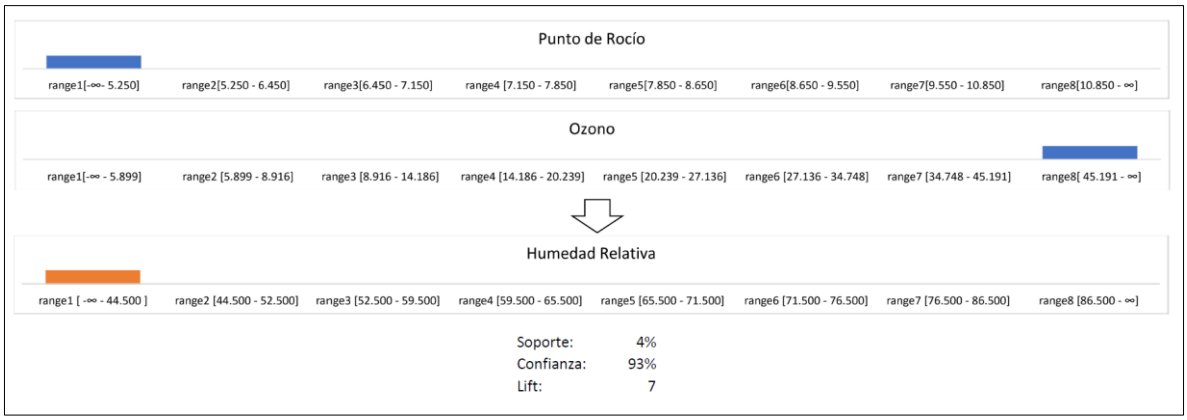

Ilustración 24*: Rangos de regla de asociación número 15.*

La decimosexta regla indica que durante las 3:50am y 7:50am junto a un punto de rocío superior a los 10.85°C, infieren en una humedad relativa superior al 86.5%. Por ello, la Ilustración 25 muestra que, a tempranas horas de la mañana en presencia de un alto nivel de punto de rocío la humedad se incrementa.

- [HORA = range2 [3.500 - 7.500], DP_AV = range8 [10.850 - ∞]] → [HR_AV = range8 [86.500 - ∞]]

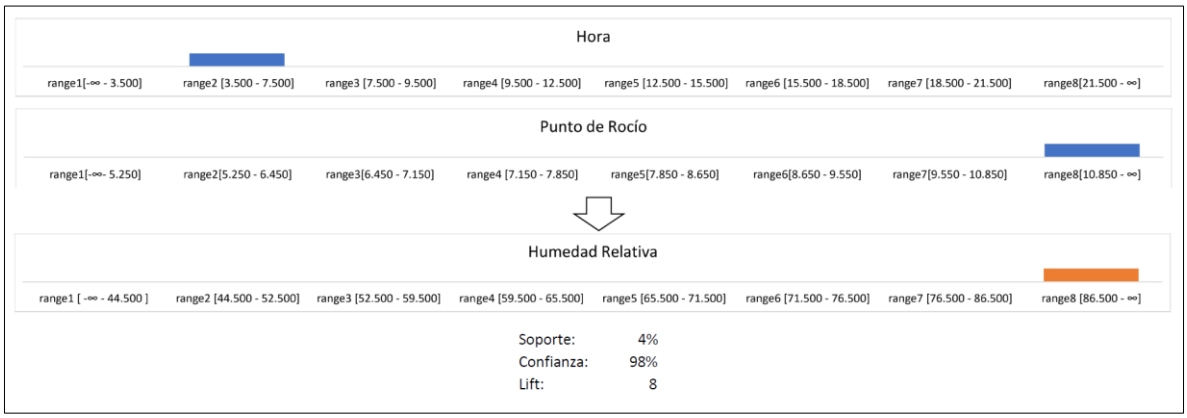

Ilustración 25*: Rangos de regla de asociación número 16.*

La regla decimoséptima muestra que una oscilación de temperatura entre los 11.9°C y 12.8°C y un punto de roció de 10.9°C o superior determinan una humedad relativa del 86.5% o más. Es importante mencionar que esta regla posee el valor más alto en cuanto a confianza con un valor del 100%. Por ello, es deducible que cuando el ambiente se torna frio y el punto de roció se eleva consecuentemente la humedad relativa se incrementa como se muestra en la Ilustración 26.

- [TEMPAIRE_AV = range2 [11.850 - 12.750], DP_AV = range8 [10.850 - ∞]] → [HR_AV = range8 [86.500 - ∞]]

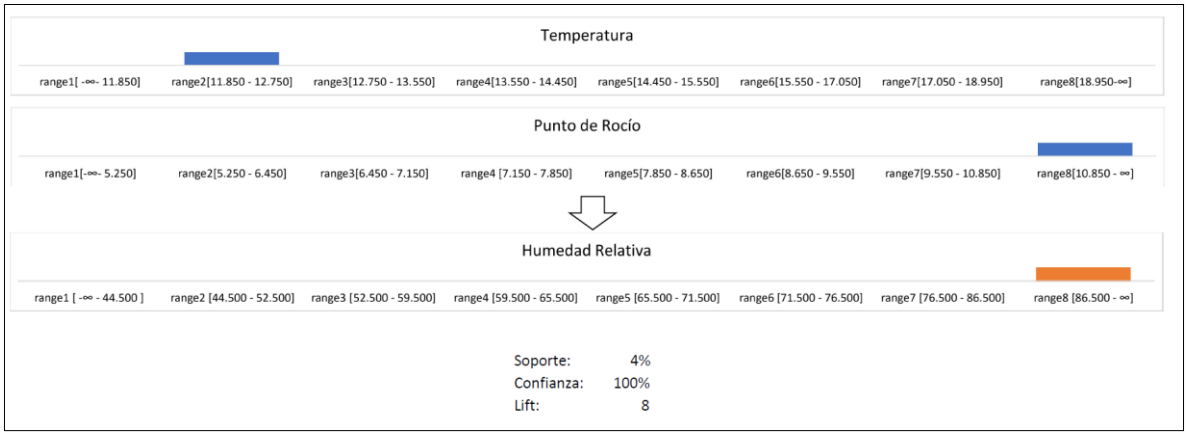

Ilustración 26: *Rangos de regla de asociación número 17.*

7. DISCUSIÓN

El objetivo principal de este estudio fue la aplicación de técnicas de asociación de minería de datos para analizar el comportamiento de contaminantes atmosféricos y variables meteorológicas. Se utilizó un conjunto de datos de enero a noviembre del año 2018. Durante la experimentación se probaron varios métodos de discretización con los valores continuos de las variables de estudio. Sin embargo, algunos métodos no resultaron eficaces para el descubrimiento de patrones, ya que en ciertos rangos se acumulaba mayor cantidad de datos y en otros rangos muy pocos o ninguno. Fue importante seleccionar la técnica adecuada de discretización, es así que se seleccionó el método de discretización frecuencia, puesto que este representa una división más equilibrada a nivel de cada rango.

Así también, es importante establecer parámetros mínimos adecuados en la selección de las reglas, por ello, en un principio se estableció un soporte de 50%, confianza del 80% y un lift del 80%, no obstante, no se obtuvieron los resultados esperados. Por lo que, se decidió reducir el valor del soporte al 10%, obteniendo así dos resultados. Si bien los resultados obtenidos son buenos, no representan en su totalidad el conocimiento del conjunto de datos, por lo que se disminuyó aún más el soporte a un 5%, lo que dio como resultado cuatro reglas de asociación. Por ello, finalmente se decidió establecer un soporte del 3% con lo cual el número de reglas aumento significativamente a diecisiete reglas.

La metodología aplicada fue eficaz en la búsqueda de patrones de comportamiento de los contaminantes atmosféricos y variables meteorológicas. Se concuerda con el comportamiento definido por (Othman, Ismail, & Latif, 2018) que indica que la temperatura es un factor meteorológico que influye en la formación de O3. El resultado obtenido en la regla de asociación número tres conincide con esta afirmación de manera particular el ozono en sus rangos más elevados, la temperatura tiende a elevarse. De manera análoga (Christy & Khanaa, 2016) menciona que cuando la temperatura se encuentra en niveles altos, es una excelente condición que conduce a la formación y acumulación de contaminantes. La correspondencia a los resultados obtenidos en la regla de asociación número tres concuerda nuevamente. Sin embargo,

lo interesante del presente estudio es que se encontraron patrones no usuales con relación a otros trabajos.

8. CONCLUSIONES

En este estudio se aplicó la técnica de asociación de minería de datos para analizar el comportamiento de contaminantes atmosféricos y variables meteorológicas. Es así que, para el desarrollo del estudio se utilizaron datos recolectados en la ciudad de Cuenca en un intervalo de meses de enero a noviembre del año 2018, cabe mencionar que la etapa de discretización fue uno de los pilares fundamentales dentro del proceso, dado que, fue necesario modular el número de grupos con la finalidad de obtener resultados eficientes posterior a la aplicación de la algoritmia.

Por ello, se determinó que la mejor manera de discretizar los datos previos a la aplicación de técnicas de asociación, es la técnica de discretización por frecuencia, puesto que realiza una distribución ecuánime de los elementos, definiendo rangos en base a las frecuencias de los elementos del conjunto de datos. Por otro lado, se descartaron las técnicas como *binning*, tamaño y especificación por usuario, puesto que presentaban una serie de incongruencias al momento de discretizar, lo que causaba un sesgo en los resultados finales.

Los parámetros de evaluación considerados para el desarrollo de este estudio fueron el soporte, confianza y, lift de cada regla de asociación, de esta manera, inicialmente se moduló el soporte desde el 100%, sin embargo, no se presentaban resultados. Por lo que, finalmente se estableció un soporte del 3%, acompañado de una confianza del 80% y, un lift de 80. De tal manera que los resultados que no presentaron relevancia fueron descartados, obteniendo así 17 reglas de asociación que describen el comportamiento de variables meteorológicas y contaminantes atmosférico. Así pues, se obtuvo conocimiento útil y comprensible para el control del aire.

9. BIBLIOGRAFÍA

Aggarwal, C. (2013). Data Mining - The Textbook. In *Journal of Chemical Information and Modeling* (Vol. 53). https://doi.org/10.1017/CBO9781107415324.004

Andrade, P., & Orellana, M. (2018). Aplicación de minería de datos en el análisis de contaminantes atmosféricos y variables meteorológicas. *Universidad Del Azuay*.

Azevedo, A., & Santos, M. F. (2008). KDD , SEMMA AND CRISP-DM : A PARALLEL OVERVIEW Ana Azevedo and M . F . Santos. *IADIS European Conference Data Mining*, 182–185. Retrieved from http://recipp.ipp.pt/handle/10400.22/136%0Ahttp://recipp.ipp.pt/bitstream/10400.22/136/3/KDD-CRISP-SEMMA.pdf

Bhatt, V., Dhakar, M., & Chaurasia, B. K. (2016). Filtered Clustering Based on Local Outlier Factor in Data Mining. *International Journal of Database Theory and Application*, *9*(5), 275–282. https://doi.org/10.14257/ijdta.2016.9.5.28

Cagliero, L., Cerquitelli, T., Chiusano, S., Garza, P., Ricupero, G., & Xiao, X. (2016). Modeling Correlations among Air Pollution-Related Data through Generalized Association Rules. *2016 IEEE International Conference on Smart Computing, SMARTCOMP 2016*. https://doi.org/10.1109/SMARTCOMP.2016.7501707

Christy, S., & Khanaa, V. (2016). Impacts of ambient air quality data analysis in urban and industrial area helps in policy making. *Journal of Chemical and Pharmaceutical Sciences*, *9*(3), 1088–1091.

Colina, A. M. (2017). *Retos y perspectivas de las competencias profesionales*. Samborondón: Universidad Ecotec.

Czechowski, P., Badyda, A., & Grzegorz, M. (2013). *DATA MINING SYSTEM FOR AIR QUALITY MONITORING NETWORKS*. *39*(4), 123–144. https://doi.org/10.2478/aep-2013-0041

Diao, Y., Liu, K. Y., Meng, X., Ye, X., & He, K. (2015). A Big Data Online Cleaning Algorithm Based on Dynamic Outlier Detection. *Proceedings - 2015 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery, CyberC 2015*, 230–234. https://doi.org/10.1109/CyberC.2015.68

Du, X., & Varde, A. S. (2016). Mining PM2.5 and traffic conditions for air quality. *2016 7th International Conference on Information and Communication Systems, ICICS 2016*, 33–38. https://doi.org/10.1109/IACS.2016.7476082

Emov Ep, R. D. M. D. L. C. D. A. D. C. (2014). *Resumen del inventario de emisiones atmosféricas del cantón Cuenca, año 2011. Cuenca-Ecuador.* (DECEMBER 2014).

Emov Ep, R. D. M. D. L. C. D. A. D. C. (2016). *Inventario de Emisiones Atmosféricas de Cuenca*. (December), 163. https://doi.org/10.13140/RG.2.2.17665.66405

Ferna, C. (2010). *A survey of data mining and knowledge discovery process models and methodologies*. *25*, 137–166. https://doi.org/10.1017/S0269888910000032

García, S., Ramírez-Gallego, S., Luengo, J., Benítez, J. M., & Herrera, F. (2016). Big data preprocessing: methods and prospects. *Big Data Analytics*, *1*(1), 1–22. https://doi.org/10.1186/s41044-016-0014-0

Geaur Rahman, M., & Zahidul Islam, M. (2016). Discretization of continuous attributes through low frequency numerical values and attribute interdependency. *Expert Systems with Applications*, *45*, 410–423. https://doi.org/10.1016/j.eswa.2015.10.005

Hahsler, M. (2016). Grouping Association Rules Using Lift. *Proceedings of the 11th INFORMS Workshop on Data Mining and Decision Analytics*. Retrieved from http://cran.r-project.org/

Huang, M., Zhang, T., Wang, J., & Zhu, L. (2015). A new air quality forecasting model using data mining and artificial neural network. *Proceedings of the IEEE International Conference on Software Engineering and Service Sciences, ICSESS*, *2015-Novem*, 259–262. https://doi.org/10.1109/ICSESS.2015.7339050

IBM. (2019). *Lift in an association rule*. 1–2. Retrieved from https://www.ibm.com/support/knowledgecenter/en/SSEPGG_9.5.0/com.ibm.im.model.doc/c_lift_in_an_association_rule.html

INEC. (2017). Resultados Censo de Población. *Población y Demografía*, 3–4.

Jiang, Y., & Zhao, L. (2012). Intelligent Decision Technologies. In *Smart Innovation, Systems and Technologies* (Vol. 15). https://doi.org/10.1007/978-3-642-29977-3

Jiawei, H., Micheline, K., & Pei, J. (2019). Data Mining Concepts and Techniques. In *Journal of Chemical Information and Modeling* (Vol. 53). https://doi.org/10.1017/CBO9781107415324.004

Jiménez, R. M. R., Capa, Á. B., & Lozano, A. P. (2004). *Meteorología Y Climatología*. Retrieved from https://cab.inta-csic.es/uploads/culturacientifica/adjuntos/20130121115236.pdf

Kalyankar, & Alaspurkar. (2013). Data Mining Technique to Analyse the Metrological Data. *International Journal of Advanced Research in Computer Science and Software Engineering*, *3*(2), 2277. Retrieved from www.ijarcsse.com

Kohail, S. N., & Alaa, A. M. (2011). Implementation of Data Mining Techniques for Meteorological Data Analysis (A case study for Gaza Strip). *International Journal of Information and Communication Technology Research*, *1*(3), 96–100.

Kurt, A., Gulbagci, B., Karaca, F., & Alagha, O. (2008). An online air pollution forecasting system using neural networks. *Environment International*, *34*(5), 592–598. https://doi.org/10.1016/j.envint.2007.12.020

Kurt, A., & Oktay, A. B. (2010). Forecasting air pollutant indicator levels with geographic models 3 days in advance using neural networks. *Expert Systems with Applications*, *37*(12), 7986–7992. https://doi.org/10.1016/j.eswa.2010.05.093

Lanjewar, U., & Shah, J. (2012). Air pollution monitoring & tracking system using mobile sensors and analysis of data using data mining. *International Journal of Advanced*

Computer Research, *2*(4, 6), 19–23.

Lavangnananda, K., & Chattanachot, S. (2017). Study of discretization methods in classification. *2017 9th International Conference on Knowledge and Smart Technology: Crunching Information of Everything, KST 2017*, 50–55. https://doi.org/10.1109/KST.2017.7886082

Li, A., & Xu, X. (2018). *A New PM2.5 Air Pollution Forecasting Model Based on Data Mining and BP Neural Network Model*. *65*(Cimns), 110–113. https://doi.org/10.2991/cimns-18.2018.25

Martines, E., & Gómez, A. (2008). Cálculo de la Temperatura de Punto de Rocío a Diferentes Valores de Presión. *Simposio de Metrología, 22 al 24 d*, 1. Retrieved from http://www.uan.edu.mx/es/comunicados/modelos-matematicos-aplicados-en-la-agricultura%0Aunmht://unmht/file.5/G:/Yunier/CIGB/HeberNem-S/Busquedas/MHT/Programa para calcular la Temperatura de Bulbo húmedo %29 Termodinámica Grupo 9.mht/

Martínez, B., Troncoso, A., Álvarez, F., & Riquelme, J. C. (2010). Mining quantitative association rules based on evolutionary computation and its application to atmospheric pollution. *Integrated Computer-Aided Engineering*, *17*(3), 227–242. https://doi.org/10.3233/ICA-2010-0340

McNicholas, P. D., Murphy, T. B., & O'Regan, M. (2008). Standardising the lift of an association rule. *Computational Statistics and Data Analysis*, *52*(10), 4712–4721. https://doi.org/10.1016/j.csda.2008.03.013

Menéndez Domínguez, V. H., & Castellanos Bolaños, M. E. (2015). SPEM: Software Process Engineering Metamodel. *Revista Latinoamericana de Ingenieria de Software*, *3*(2), 92. https://doi.org/10.18294/relais.2015.92-100

Ncr, P. C., Spss, J. C., Ncr, R. K., Spss, T. K., Daimlerchrysler, T. R., Spss, C. S., & Daimlerchrysler, R. W. (2000). *Crisp-dm 1.0*.

Ortega, J., & Orellana, M. (2018). Impacto de la aplicación de algoritmos de minería de datos en variables de contaminación del aire. *Universidad Del Azuay*, (2). https://doi.org/10.1051/matecconf/201712107005

Othman, Z. A., Ismail, N., & Latif, M. T. (2018). Association rules of temperature towards high and low ozone in Putrajaya. *Proceedings of the 2017 6th International Conference on Electrical Engineering and Informatics: Sustainable Society Through Digital Innovation, ICEEI 2017*, *2017-Novem*, 1–5. https://doi.org/10.1109/ICEEI.2017.8312438

Peña, M., Ortega, P., & Orellana, M. (2019). A novel imputation method for missing values in air pollutant time series data. *Universidad Del Azuay*, (June 2012).

Qin, S., Liu, F., Wang, C., Song, Y., & Qu, J. (2015). Spatial-temporal analysis and projection of extreme particulate matter (PM10 and PM2.5) levels using association rules: A case study of the Jing-Jin-Ji region, China. *Atmospheric Environment*, *120*, 339–350. https://doi.org/10.1016/j.atmosenv.2015.09.006

Ramirez-Gallego, S., Garcia, S., Mourino-Talin, H., Martinez-Rego, D., Bolon-Canedo, V., Alonso-Betanzos, A., … Herrera, F. (2015). Distributed Entropy Minimization Discretizer for Big Data Analysis under Apache Spark. *Proceedings - 14th IEEE International Conference on Trust, Security and Privacy in Computing and Communications, TrustCom 2015*, *2*, 33–40. https://doi.org/10.1109/Trustcom.2015.559

Ramírez-Gallego, S., García, S., Mouriño-Talín, H., Martínez-Rego, D., Bolón-Canedo, V., Alonso-Betanzos, A., … Herrera, F. (2016). Data discretization: Taxonomy and big data challenge. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *6*(1), 5–21. https://doi.org/10.1002/widm.1173

Rapidminer, C. (2019). *Discretizar por especificación de usuario*. *5*, 1–5. Retrieved from https://docs.rapidminer.com/latest/studio/operators/cleansing/binning/discretize_by_user_specification.html

Ropero, R. F., Renooij, S., & Van Der Gaag, L. C. (2018). Discretizing environmental data for learning Bayesian-network classifiers. *Ecological Modelling*, *368*, 391–403. https://doi.org/10.1016/j.ecolmodel.2017.12.015

Saeed, U., Sarim, M., Usmani, A., Mukhtar, A., Shaikh, A. B., & Raffat, S. K. (2015). Application of Machine learning Algorithms in Crime Classification and Classification Rule Mining. *Research Journal of Recent Sciences*, *4*(3), 106–114. Retrieved from www.isca.me

Sahafizadeh, E., & Ahmadi, E. (2009). Prediction of air pollution of Boushehr City using data mining. *2nd International Conference on Environmental and Computer Science, ICECS 2009*, 33–36. https://doi.org/10.1109/ICECS.2009.18

Sánchez Zavaleta, C. A. (2016). Evolution of the climate change concept and its impact in the public health of Peru. *Revista Peruana de Medicina Experimental y Salud Publica*, *33*(1), 128–138. https://doi.org/10.17843/rpmesp.2016.331.2014

Schubert, E., Zimek, A., & Kriegel, H. P. (2014). Local outlier detection reconsidered: A generalized view on locality with applications to spatial, video, and network outlier detection. *Data Mining and Knowledge Discovery*, *28*(1), 190–237. https://doi.org/10.1007/s10618-012-0300-z

Shafique, U., & Haseeb, Q. (2013). 済無No Title No Title. In *Journal of Chemical Information and Modeling* (Vol. 53). https://doi.org/10.1017/CBO9781107415324.004

Téllez, J., Rodríguez, A., & Fajardo, Á. (2006). Contaminación por monóxido de carbono: Un problema de salud ambiental. *Revista de Salud Publica*, *8*(1), 108–117. https://doi.org/10.1590/s0124-00642006000100010

Toti, G., Vilalta, R., Lindner, P., Lefer, B., Macias, C., & Price, D. (2016). Analysis of correlation between pediatric asthma exacerbation and exposure to pollutant mixtures with association rule mining. *Artificial Intelligence in Medicine*, *74*, 44–52. https://doi.org/10.1016/j.artmed.2016.11.003

Ubilla, C., & Yohannessen, K. (2017). Contaminación Atmosférica Efectos En La Salud

Respiratoria En El Niño. *Revista Médica Clínica Las Condes*, *28*(1), 111–118. https://doi.org/10.1016/j.rmclc.2016.12.003

Uysal, A. K., & Gunal, S. (2018). The impact of discretization method on the detection of six types of anomalies in datasets. *Belgian/Netherlands Artificial Intelligence Conference*, (Bnaic), 47–62.

Wirth, R. (2000). CRISP-DM : Towards a Standard Process Model for Data Mining. *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*, (24959), 29–39. https://doi.org/10.1.1.198.5133

Witten, I. H., Frank, E., & Hall, M. A. (2011). Data Mining. In *Data Mining*. https://doi.org/10.1016/C2009-0-19715-5

Wu, X., Wang, Q., Liu, Y., & Li, Y. (2018). Data-driven regional analysis of urban atmosphere pollution based on density clustering. *Proceedings of the 2017 IEEE 2nd Information Technology, Networking, Electronic and Automation Control Conference, ITNEC 2017*, *2018-Janua*(1), 412–417. https://doi.org/10.1109/ITNEC.2017.8284764

Wu, X., Wang, Q., Liu, Y., Li, Y., Kurt, A., Gulbagci, B., … Hall, M. A. (2016). Data Mining Theories, Algorithms, and Examples. *Journal of Chemical Information and Modeling*, *6*(3), 1689–1699. https://doi.org/10.1017/CBO9781107415324.004

Yan, Y., Cao, L., & Rundensteiner, E. A. (2017). Distributed Top-N local outlier detection in big data. *Proceedings - 2017 IEEE International Conference on Big Data, Big Data 2017*, *2018-Janua*, 827–836. https://doi.org/10.1109/BigData.2017.8257998

Ye, N. (2015). Data mining. In *IEEE Potentials* (Vol. 16). https://doi.org/10.1109/45.624335

10. ANEXOS

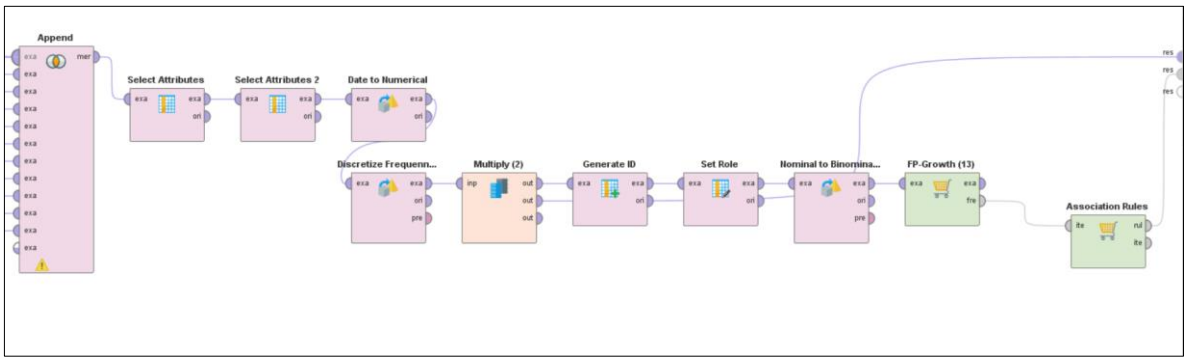

Ilustración 27*: Modelado de operadores.*

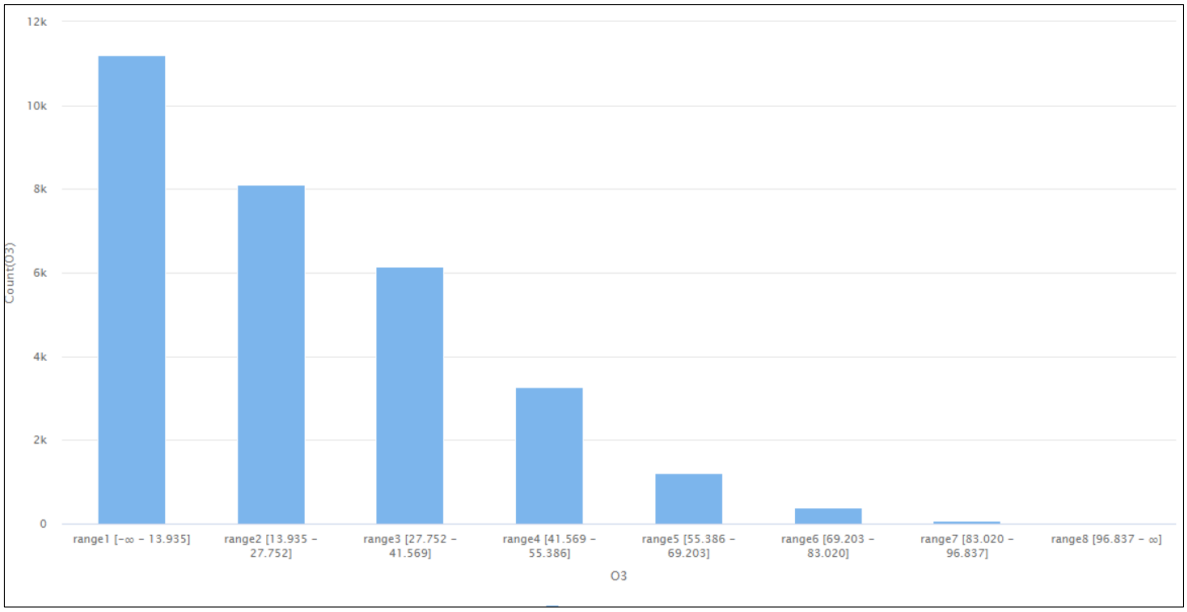

Ilustración 28: *Discretización por Binning.*

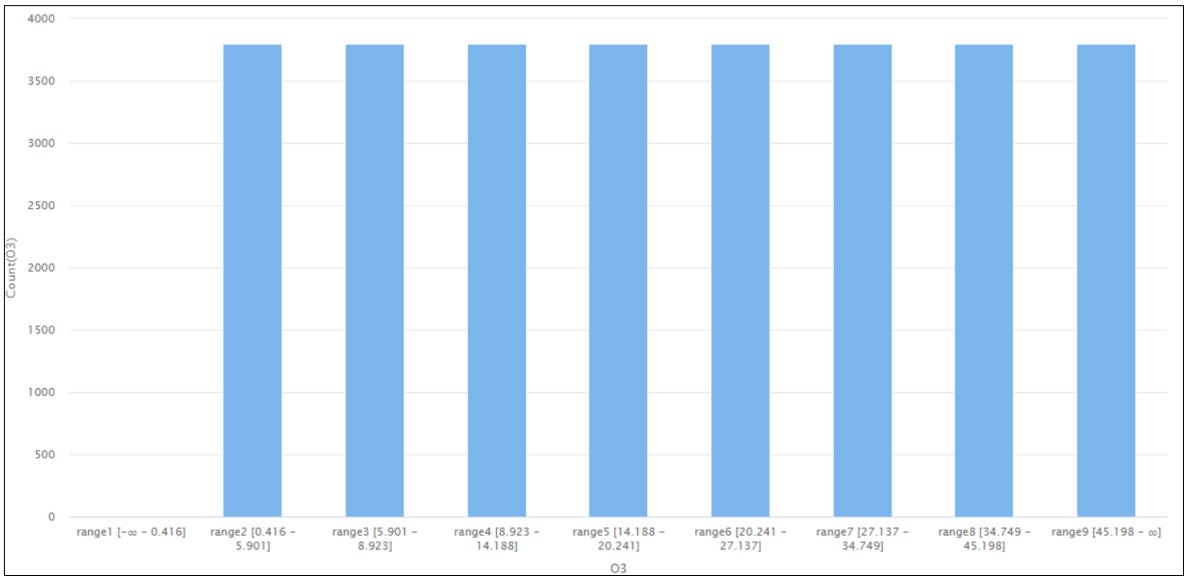

Ilustración 29: *Discretización por Tamaño (Size).*

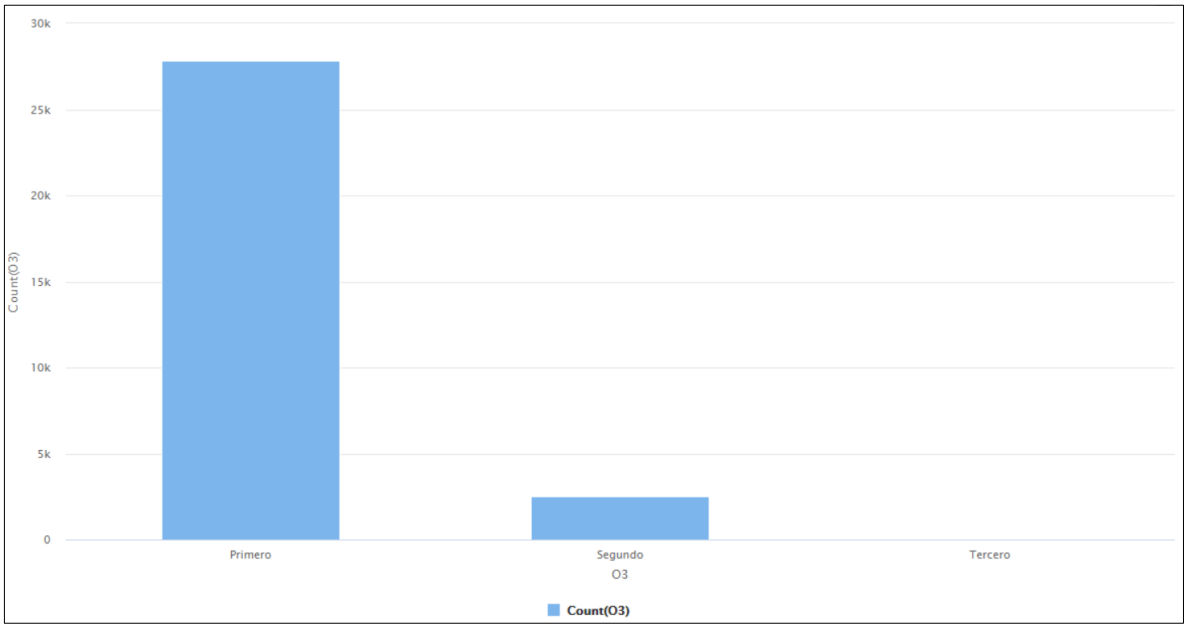

Ilustración 30: *Discretización por Especificación de Usuario.*

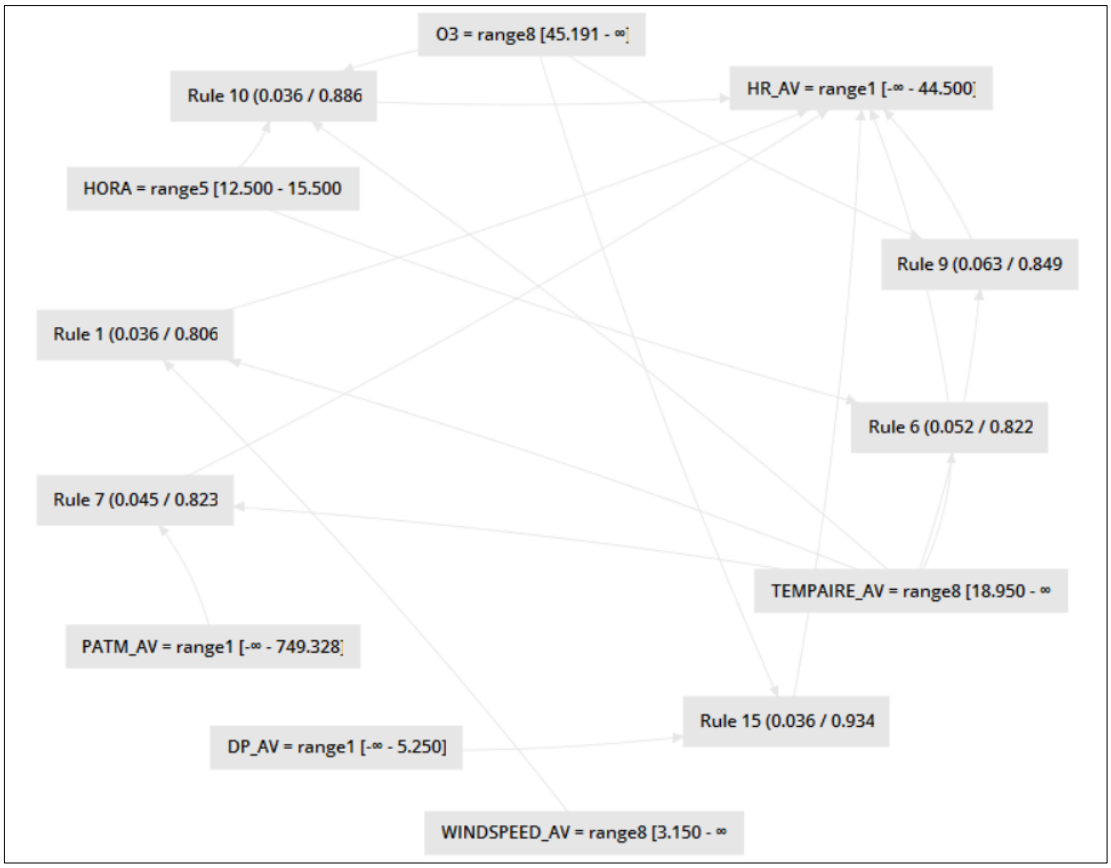

Ilustración 31: *Relación entre reglas de asociación de HR_AV = range1 [-∞ - 44.500].*

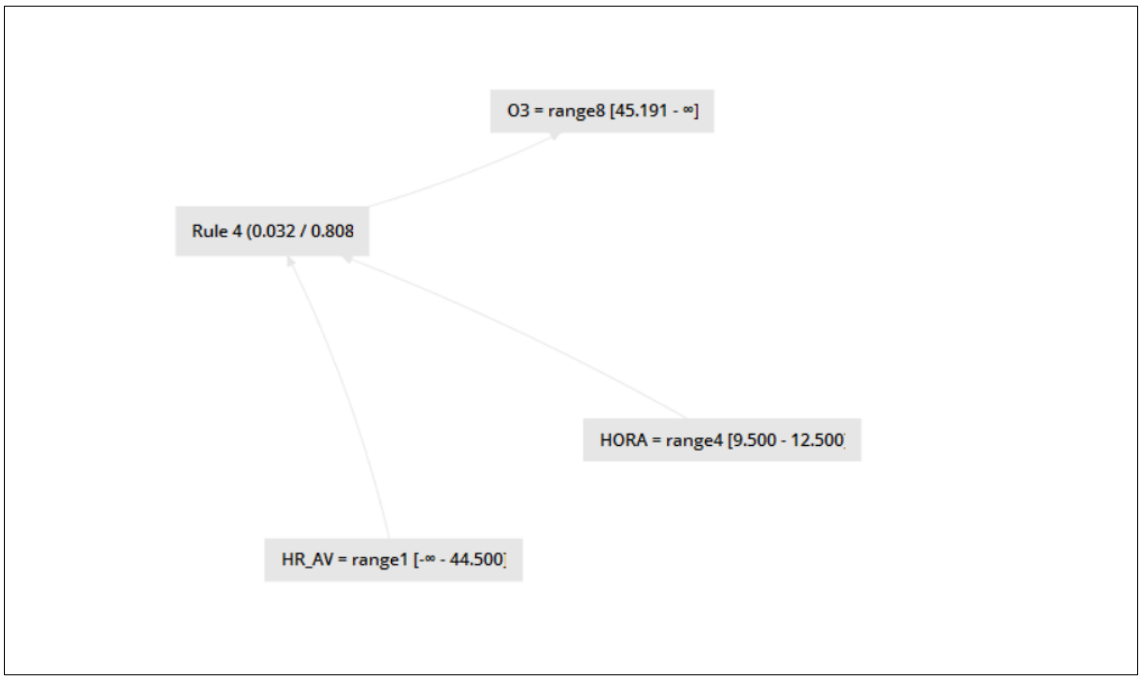

Ilustración 32: *Relación entre reglas de asociación de O3 = range8 [45.191 - ∞].*

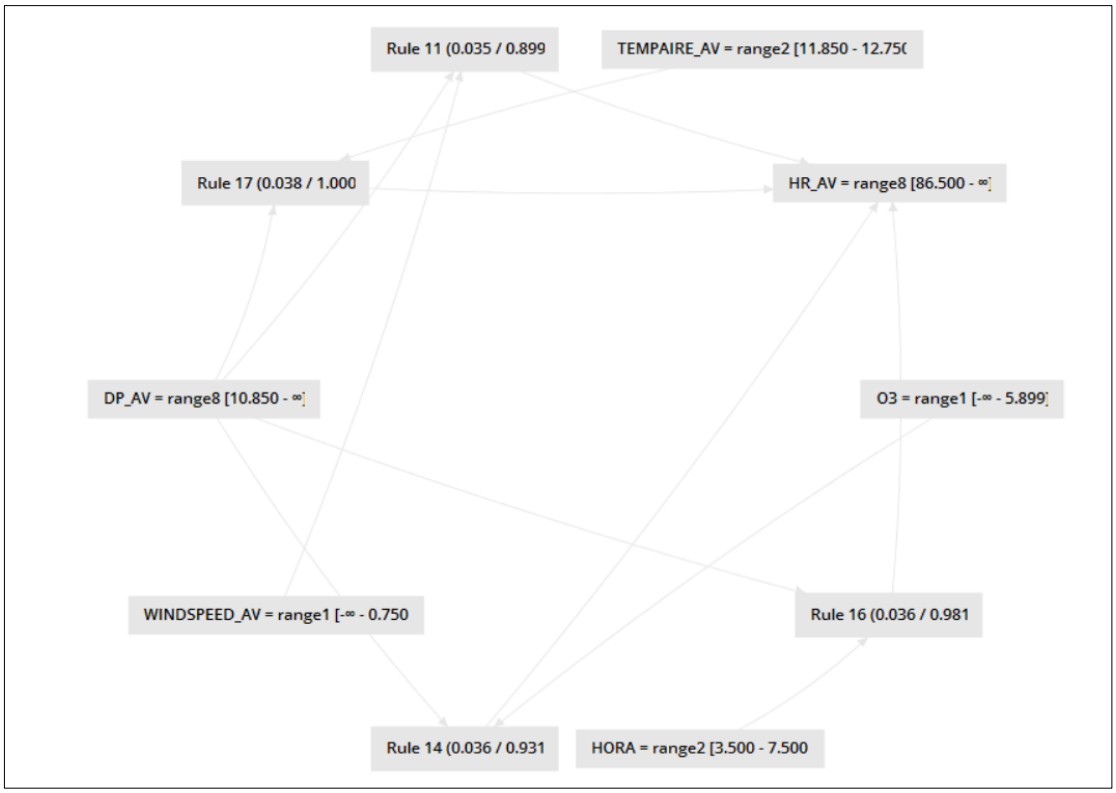

Ilustración 33: *Relación entre reglas de asociación HR_AV = range8 [86.500 - ∞].*

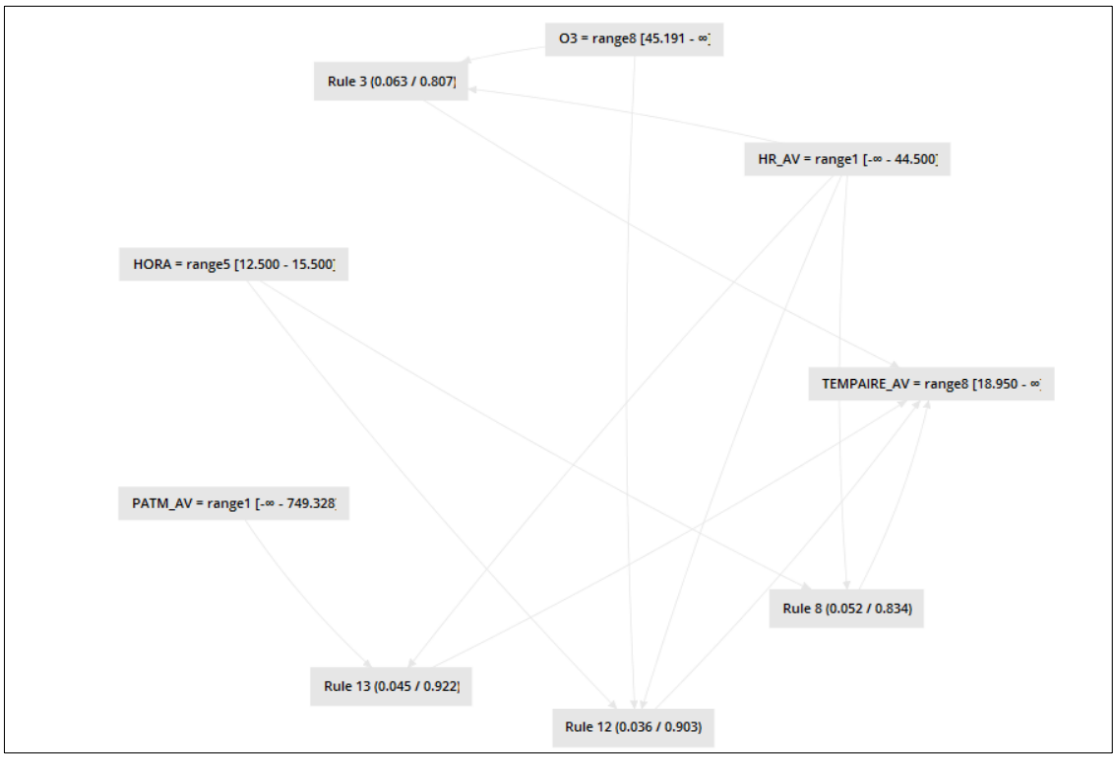

Ilustración 34: *Relación entre reglas de asociación TEMPAIRE_AV = range8 [18.950 - ∞].*

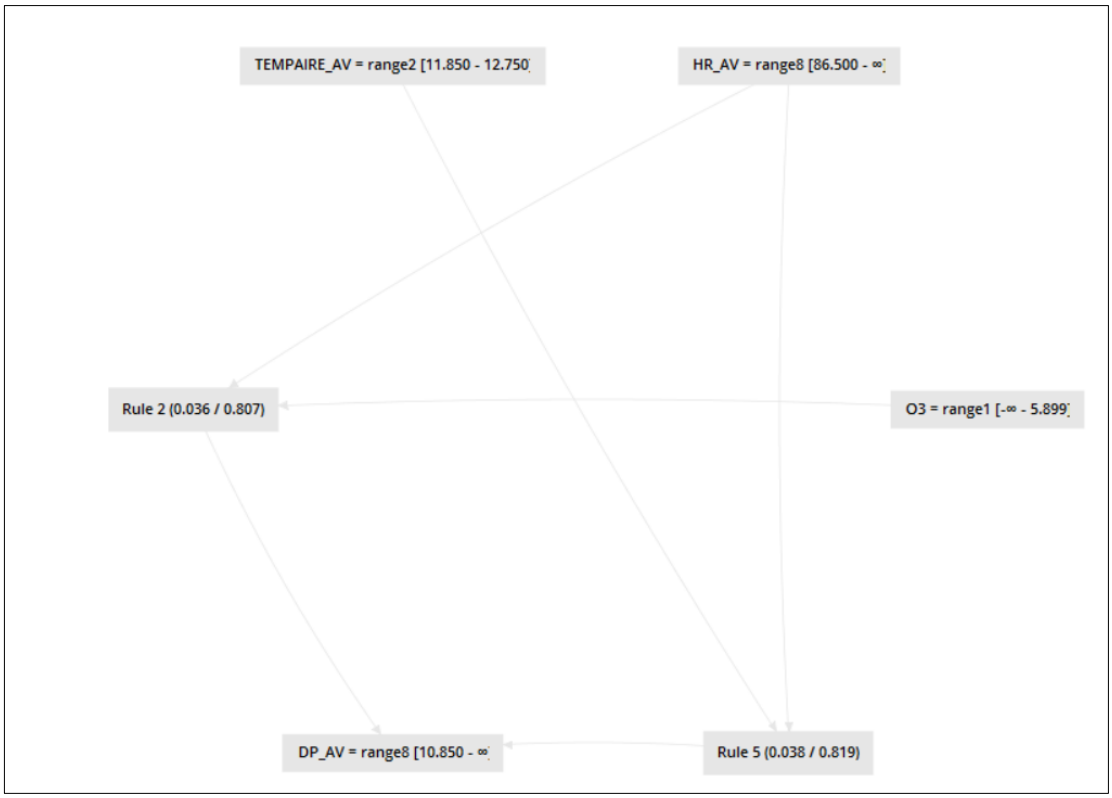

Ilustración 35: *Relación entre reglas de asociación DP_AV = range8 [10.850 - ∞].*

UNIVERSIDAD DEL AZUAY

Doctora María Elena Ramírez Aguilar, Secretaria de la Facultad de Ciencias de la Administración de la Universidad del Azuay

C E R T I F I C A:

Que, el Consejo de Facultad de Ciencias de la Administración, en sesión del 31 de julio de 2019, conoció y aprobó la solicitud para la realización del trabajo de titulación y el respectivo protocolo presentado por:

Estudiante: Jimmy Fernando Salto Sumba con código 74309
Tema: **Análisis de comportamiento de componentes atmosféricos y variables meteorológicas aplicando técnicas de asociación de minería de datos**
Previo a la obtención del título de **Ingeniero de Sistemas y Telemática**
Director: Ing. Marcos Orellana Cordero
Tribunal: Ing. Patricia Ortega Chasi e Ing. Esteban Crespo Martínez

Plazo de presentación del trabajo de titulación: El Consejo de Facultad resolvió establecer el plazo de seis meses para la presentación del trabajo de titulación concluido y calificado por el Director; este plazo se contará desde la fecha de aprobación del protocolo, esto es hasta el 31 de enero de 2020.

Cuenca, 1 de agosto de 2019

Dra. María Elena Ramírez Aguilar
Secretaria de la Facultad de Ciencias de la Administración

UNIVERSIDAD DEL AZUAY
Facultad de Ciencias de la Administración
SECRETARÍA

UNIVERSIDAD DEL AZUAY

CONVOCATORIA

Por disposición de la Junta Académica de la escuela de Ingeniería de Sistemas y Telemática se convoca a los Miembros del Tribunal Examinador, a la sustentación del Protocolo del Trabajo de Titulación: **Análisis de comportamiento de componentes atmosféricos y variables meteorológicas aplicando técnicas de asociación de minería de datos,** presentado por el estudiante Jimmy Fernando Salto Sumba con código 74309, previa a la obtención del título de Ingeniero de Sistemas y Telemática, para el día **Jueves, 27 de junio de 2019 a las 09:20**

Tomar en cuenta que posterior a la sustentación del Diseño del Trabajo de Titulación, por ningún concepto se puede realizar modificaciones ni cambios en los documentos; únicamente, en caso de diseño aprobado con modificación, el Director adjuntará al esquema un oficio indicando que se procede con los cambios sugeridos.

Cuenca, 24 de junio de 2019

Dra. María Elena Ramírez Aguilar
Secretaria de la Facultad

Ing. Marcos Orellana Cordero .....

Ing. Patricia Ortega Chasi .....

Ing. Esteban Crespo Martínez .....

Oficio Nro. 052-2019-DIST-UDA

Cuenca, 18 de junio de 2019

Ingeniero,
Oswaldo Merchán Manzano
DECANO DE LA FACULTAD DE CIENCIAS DE LA ADMINISTRACIÓN
UNIVERSIDAD DEL AZUAY

De nuestras consideraciones,

La Junta Académica de la Escuela de Ingeniería de Sistemas y Telemática, reunida el día 18 de junio del 2019, revisó la documentación del trabajo de titulación denominado**"ANÁLISIS DE COMPORTAMIENTO DE COMPONENTES ATMOSFÉRICOS Y VARIABLES METEOROLÓGICAS APLICANDO TÉCNICAS DE ASOCIACIÓN DE MINERÍA DE DATOS"**, por la/el estudiante **JIMMY FERNANDO SALTO SUMBA**, con código/s estudiantil **74309**, estudiante/s de la Escuela de Ingeniería de Sistemas y Telemática, y revisado por **MARCOS ORELLANA**, previo a la obtención del título de Ingeniero de Sistemas y Telemática.

La Junta Académica considera que la documentación cumple con las normas legales y reglamentarias de la Universidad y de la Facultad de Ciencias de la Administración y designa como miembros del tribunal a Patricia Ortega y Esteban Crespo, así por su digno intermedio, el conocimiento y aprobación por parte del Consejo de Facultad.

Atentamente,

Marcos Orellana Cordero
Coordinador de la Escuela de Ingeniería de Sistemas y Telemática
Universidad del Azuay

1. FECHA DE RECEPCIÓN DE PROTOCOLO: 17-06-2019 FIRMA

2. REVISIÓN DE ESTADO ACADÉMICO DEL ALUMNO:

NOMBRE: Jimmy Fernando Salto Sumba

CÓDIGO: 74309

CARRERA: Ingeniería en Sistemas y Telemática.

FECHA DE INICIO DE ESTUDIOS: 17 Sep/2013

FECHA CULMINACIÓN DE ESTUDIOS: No Termina — Tramitado homolog

HOMOLOGACIONES: NO CARRERA PROCEDENTE Tiene que pedir Homol. H. Cred. X

CONVALIDACIONES: NO UNIVERSIDAD PROCEDENTE:

FECHA DE ESTA REVISIÓN: 17 Junio/2019 FIRMA:

DE: DRA. MARÍA ELENA RAMÍREZ, SECRETARIA

ASUNTO: ENVÍO DE PROTOCOLO DE TRABAJO DE TITULACIÓN

PARA: JUNTA ACADÉMICA DE LA CARRERA DE Sistemas

TÍTULO A OTORGARSE: Ingeniero de sistemas y Telemática

Observación:

4 Homologación pendiente

Fecha de revisión: 17/06/2019

FIRMA:

TÍTULO DEL TRABAJO: Análisis de comportamiento de componentes atmosféricos y variables meteorológicas aplicando técnicas de asociación de minería de datos.

REALIZADO EN EL CURSO DE METODOLOGÍA: X SI ____ NO

FECHA DE APROBACIÓN DEL CONSEJO DE FACULTAD: 31 de julio de 2019

DIRECTOR: Ing. Marcos Orellana Cordero

TRIBUNAL: Ing. Patricia Ortega Chasi e Ing. Esteban Crespo Martínez.

UNIVERSIDAD DEL AZUAY

ACTA

SUSTENTACIÓN DE PROTOCOLO/DENUNCIA DEL TRABAJO DE TITULACIÓN

1. Nombre del estudiante: Jimmy Fernando Salto Sumba
2. Código: 74309
3. Director sugerido: Ing. Marcos Orellana Cordero
4. Codirector (opcional): _____
5. Tribunal: Ing. Patricia Ortega Chasi e Ing. Esteban Crespo Martínez
6. Título propuesto: **Análisis de comportamiento de componentes atmosféricos y variables meteorológicas aplicando técnicas de asociación de minería de datos**
7. Aceptado sin modificaciones: _____
8. Aceptado con las siguientes modificaciones:

_____

_____

_____

9. No aceptado
10. Justificación:

_____

_____

_____

Tribunal

Ing. Marcos Orellana Cordero	Ing. Patricia Ortega Chasi	Ing. Esteban Crespo Martínez

Sr. Jimmy Fernando Salto Sumba	Dra. María Elena Ramírez Aguilar Secretaria de la Facultad

UNIVERSIDAD DEL AZUAY

RÚBRICA PARA LA EVALUACIÓN DEL PROTOCOLO DE TRABAJO DE TITULACIÓN
(Tribunal)

1. Nombre del estudiante: Jimmy Fernando Salto Sumba
2. Código: 74309
3. Director sugerido: Ing. Marcos Orellana Cordero
4. Codirector (opcional):
5. Título propuesto: **Análisis de comportamiento de componentes atmosféricos y variables meteorológicas aplicando técnicas de asociación de minería de datos**
6. Revisores tribunal: Ing. Patricia Ortega Chasi e Ing. Esteban Crespo Martínez

	Cumple	No cumple
Problemática y/o pregunta de investigación		
1. ¿Presenta una descripción precisa y clara?	/	
2. ¿Tiene relevancia profesional y social?	/	
Objetivo general		
3. ¿Concuerda con el problema formulado?	/	
4. ¿Se encuentra redactado en tiempo verbal infinitivo?	/	
Objetivos específicos		
5. ¿Permiten cumplir con el objetivo general?	/	
6. ¿Son comprobables cualitativa o cuantitativamente?	/	
Metodología		
7. ¿Se encuentran disponibles los datos y materiales mencionados?	/	
8. ¿Las actividades se presentan siguiendo una secuencia lógica?	/	
9. ¿Las actividades permitirán la consecución de los objetivos específicos planteados?	/	
10. ¿Las técnicas planteadas están de acuerdo con el tipo de investigación?	/	
Resultados esperados		
11. ¿Son relevantes para resolver o contribuir con el problema formulado?	/	
12. ¿Concuerdan con los objetivos específicos?	/	
13. ¿Se detalla la forma de presentación de los resultados?	/	
14. ¿Los resultados esperados son consecuencia, en todos los casos, de las actividades mencionadas?	/	

Nota sobre 10 puntos: : 10

Ing. Marcos Orellana Cordero

Ing. Patricia Ortega Chasi

Ing. Esteban Crespo Martínez

Cuenca, 13 de junio del 2019

Ingeniero,
Oswaldo Merchán Manzano
DECANO DE LA FACULTAD DE CIENCIAS DE LA ADMINISTRACIÓN
UNIVERSIDAD DEL AZUAY

De mi consideración,

Estimado Señor Decano, yo **Jimmy Fernando Salto Sumba** con C.I. **0106046626**, código estudiantil 74309; estudiante de la Carrera de Ingeniería de Sistemas y Telemática, solicito encarecidamente a usted, la aprobación del protocolo de trabajo de titulación con el tema **"ANÁLISIS DE COMPORTAMIENTO DE COMPONENTES ATMOSFÉRICOS Y VARIABLES METEOROLÓGICAS APLICANDO TÉCNICAS DE ASOCIACIÓN DE MINERÍA DE DATOS"** previo a la obtención del título de Ingeniero de Sistemas y Telemática para lo cual adjunto la documentación respectiva.

Por la favorable acogida que brinde a la presente, anticipo mi agradecimiento.

Atentamente

Jimmy Fernando Salto Sumba
Estudiante de la Escuela de Ingeniería de Sistemas y Telemática
Universidad del Azuay

0906743

UNIVERSIDAD
DEL AZUAY

LA SECRETARIA DE LA FACULTAD DE CIENCIAS DE LA ADMINISTRACION DE LA UNIVERSIDAD DEL AZUAY

CERTIFICA:

Que, el señor JIMMY FERNANDO SALTO SUMBA, con número de cédula de identidad 0106046626, código de estudiante Nro. 74309, alumno de la carrera de INGENIERÍA DE SISTEMAS Y TELEMÁTICA, tiene aprobado el 89,96% de créditos de su malla curricular.

Cuenca, 28 de mayo de 2019

Dra. María Elena Ramírez Aguilar
SECRETARIA DE LA FACULTAD DE CIENCIAS DE LA ADMINISTRACION

UNIVERSIDAD DEL AZUAY
Facultad de Ciencias de la Administración
SECRETARÍA

Derecho Nro. Null

rgp.-

0904372

Cuenca, 13 de junio del 2019

Ingeniero,
Oswaldo Merchán Manzano
DECANO DE LA FACULTAD DE CIENCIAS DE LA ADMINISTRACIÓN
UNIVERSIDAD DEL AZUAY

De mi consideración,

Yo, **Marcos Patricio Orellana Cordero** informo que he revisado el protocolo de trabajo de titulación, elaborado previo a la obtención del título de Ingeniero de Sistemas y Telemática, **"ANÁLISIS DE COMPORTAMIENTO DE COMPONENTES ATMOSFÉRICOS Y VARIABLES METEOROLÓGICAS APLICANDO TÉCNICAS DE ASOCIACIÓN DE MINERÍA DE DATOS"**, realizado por el estudiante **Salto Sumba Jimmy Fernando con código estudiantil** 74309, protocolo que a mi criterio, cumple con los lineamientos y requerimientos establecidos por la carrera.

Por lo expuesto, me permito sugerir que sea considerado para la revisión y sustentación del mismo,

Sin otro particular, me suscribo.

Atentamente

Marcos Patricio Orellana Cordero

Universidad del Azuay

0906742

UNIVERSIDAD DEL AZUAY

UNIVERSIDAD DEL AZUAY

GUÍA PARA LA ELABORACIÓN Y PRESENTACIÓN DE LA DENUNCIA/PROTOCOLO DE TRABAJO DE TITULACIÓN

1. DATOS GENERALES

1.1 Nombre del estudiante:

1.1.1 Código: 74309

1.1.2 Contacto:

Telf.: 072879348

Cel.: 0991552027

Correo Electrónico: jimmy.salto@outlook.es

1.2 Director sugerido:

1.2.1 Contacto: Orellana Cordero, Marcos Ingeniero.

Teléfonos:

Telf.: 074128640

Cel.: 0999955611

Correo Electrónico: marore@uazuay.edu.ec

1.3 Co-director sugerido: (opcional).

1.4 Asesor metodológico: Daniela Elisabet Ballari

1.5 Tribunal designado:

1.6 Aprobación:

1.7 Línea de investigación de la carrera:

Código UNESCO:

1203 Ciencia de Los Ordenadores

1203.17 Informática

1.1.1 Tipo de Trabajo:

Proyectos Técnicos-Investigación

1.2 Área de estudio:

Minería de Datos

1.3 Título propuesto:

Análisis de comportamiento de componentes atmosféricos y variables meteorológicas aplicando técnicas de asociación de minería de datos.

1.4 Estado del proyecto:

Nuevo

2. CONTENIDO

2.1 Motivación de la investigación:

El fenómeno de la contaminación del aire se encuentra vinculado con el desarrollo continuo de la sociedad. Las áreas urbanas se encuentran habitadas por un gran número de personas, por lo tanto, la industria y el transporte se encuentran altamente desarrollados. En general, la contaminación emitida es una amenaza no regulada para los ciudadanos, causando problemas en la salud. Por lo que, la regulación de los niveles de contaminantes del aire se ha convertido en una de las tareas más importantes para los gobiernos en desarrollo, por lo tanto, es necesario que exista una adecuada gestión de la calidad del aire de parte de los gobernantes (Czechowski, Badyda, & Majewski, 2013; Martínez-Ballesteros, Troncoso, Martínez-Álvarez, & Riquelme, 2010).

Para evidenciar los problemas de contaminación del aire, es necesario recopilar datos que caractericen la calidad del aire. En las ciudades existe una monitorización constante de componentes atmosféricos y variables meteorológicas que genera gran cantidad de datos, los cuales son utilizados, en combinación tanto de técnicas de minería de datos como de *machine learning*, para descubrir patrones de comportamiento basados en grandes conjuntos de datos para generar conocimiento (Kalyankar & Alaspurkar, 2013). Así se crean modelos de predicción precisos a partir de grandes y complejos conjuntos de datos, con la finalidad de determinar hipótesis o pronosticar el comportamiento de la calidad del aire. Dichas predicciones resultan útiles para determinar cómo se encuentran los niveles de los contaminantes (Czechowski et al., 2013).

2.2 Problemática:

A pesar de los trabajos mencionados anteriormente, no existen aplicaciones de minería de datos que involucren tanto componentes atmosféricos como variables meteorológicas y que den como resultado una relación descriptiva y comprensible basados en la discretización de estas variables, esto implica utilizar métodos de asociación. El trabajo de Lanjewar & Shah (2012) exploró estos métodos con la aplicación de técnicas de asociación a un conjunto de datos de contaminantes del aire y consecuentemente determinó la relación de asociación que existe entre estos, aunque el mismo no cuenta con un análisis de las variables meteorológicas. El presente proyecto busca extender el trabajo de Lanjewar & Shah (2012) en combinación tanto de componentes atmosféricos como de variables meteorológicas, realizando la discretización de las mismas para la aplicación de técnicas de asociación que determinen relaciones descriptivas y comprensibles para los ciudadanos.

2.3 Pregunta de investigación:

¿Existen asociaciones relevantes que demuestren patrones de comportamiento entre componentes atmosféricos y variables meteorológicas?

2.4 Resumen

La relación entre los componentes atmosféricos y las variables meteorológicas es importante para evaluar la calidad del aire y con ello evitar riesgos en la salud, sin embargo, encontrar una relación de asociación entre estos factores depende del método de categorización que se emplee, es decir, frecuencia, tamaño, binning, etc. Por lo tanto, el objetivo de este trabajo es discretizar un conjunto de datos, tanto de variables atmosféricas como variables meteorológicas, y posteriormente aplicar técnicas de asociación para evaluar los resultados de combinaciones entre estas y encontrar patrones de comportamiento. De esta

manera, se espera formalizar combinaciones apropiadas para criterios de soporte, confianza y lift.

2.5 Estado del arte y marco teórico

Mucho se ha estudiado al respecto a nivel internacional acerca del uso de técnicas de minería de datos para el tratamiento tanto de componentes atmosféricos como de variables atmosféricas. Un ejemplo a destacar es el caso Estambul, provincia de Turquía, que ha implementado un sistema basado en módulos para realizar predicciones de contaminación del aire. El primer módulo recolecta datos de variables meteorológicas y mediciones diarias de niveles de contaminación atmosféricas de sitios web. El segundo módulo, conocido como módulo de pronóstico, usa redes neuronales para pronosticar hasta tres días de niveles de contaminantes. Finalmente, el tercer módulo, conocido como módulo web, se utiliza para visualizar los resultados categorizados en tres niveles de calidad (Kurt, Gulbagci, Karaca, & Alagha, 2008). El caso de Franja de Gaza, se recopilaron datos durante un periodo de nueve años [1977-1985] y posteriormente, con los resultados de las técnicas aplicadas sobre el conjunto de datos, se realizó un análisis generando conocimiento útil para agricultoras, sector turístico, sectores hídricos y otros sectores involucrados con el comportamiento del aire (Kohail & El-halees, 2011). De igual manera, se pronosticó días polvorientos utilizando presión de aire, humedad y datos anteriores de días polvorientos (Sahafizadeh & Ahmadi, 2009). Así también, con el uso de redes neuronales, se determinaron niveles de dióxido de azufre (SO_2), monóxido de carbono (CO) y material particulado (PM10) con tres días de anticipación (Kurt & Oktay, 2010).

Actualmente, dentro del contexto local, la ciudad de Cuenca cuenta con proyectos realizados tales como el análisis de patrones de comportamiento entre componentes atmosféricos y variables meteorológicas (Andrade & Orellana, 2018). Así como, la búsqueda de algoritmos óptimos para el análisis de variables de contaminación atmosférica (Ortega & Orellana, 2018).

2.6 Objetivo general

Aplicar técnicas de asociación de minería de datos para analizar el comportamiento de componentes atmosféricos y variables meteorológicas.

2.7 Objetivos específicos

1. Revisar estudios similares con técnicas de minería de datos.
2. Experimentar con diferentes niveles de discretización de componentes atmosféricos y variables meteorológicas.
3. Identificar patrones relevantes que determinen la asociación entre las variables atmosféricas y meteorológicas.
4. Exponer los resultados obtenidos de la técnica utilizada.

2.8 Metodología

Para realizar el presente trabajo es necesario seguir una serie de etapas que ayudarán a gestionar de mejor manera el proyecto. La etapa inicial se enfoca en la recopilación de los datos, etapa en la cual los datos son capturados por sensores de la estación meteorológica ubicada en el centro histórico de Cuenca. Posteriormente, se preparan los datos, es decir, se realiza una limpieza de los mismos, por ejemplo, se rellenan datos si es que existiesen faltantes o simplemente se los descartan como valores atípicos que puedan posteriormente generar inconvenientes en la siguiente etapa. Por otro lado, la discretización de

los datos es fundamental en este trabajo, por lo que, la manera en la que se discreticen los conjuntos de datos tanto de variables atmosféricas como de variables meteorológicas brindará los resultados esperados. A continuación, se aplica técnicas de asociación, en este caso en particular se utilizará la técnica *A priori*, algoritmo seleccionado en base a la bibliografía revisada, debido a que es el más utilizado para determinar reglas de asociación, además de ser caracterizado por su facilidad de uso. Consecuentemente se establecerán métricas como soporte, confianza y lift para evaluar los resultados. Finalmente, se evalúan los resultados en base a las métricas establecidas anteriormente, esto determina si es o no factible volver a discretizar los componentes atmosféricos o variables meteorológicas.

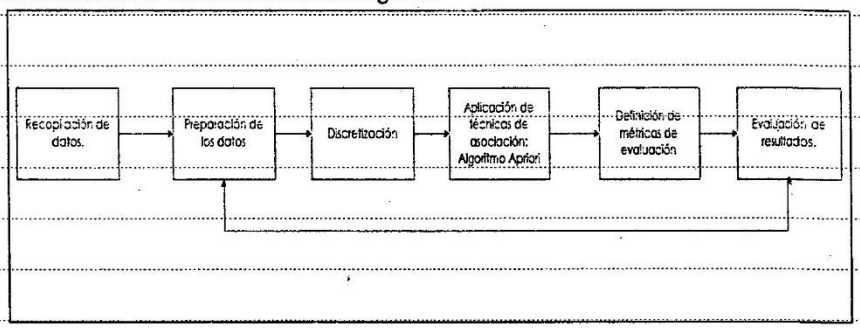

Ilustración 1 Metodología de desarrollo del proyecto.

2.8.1 Área de estudio

El área de estudio es la ciudad de Cuenca, con aproximadamente quinientos mil habitantes, se ubica como la tercera ciudad del Ecuador. Localizada al sur del país (INEC, 2017). Se encuentra a una altura de 2.550 metros sobre el nivel del mar, entre las coordenadas 78°59' – 79°01' de longitud oeste y 2°52' – 2°54'de latitud sur. Presenta un clima templado, es decir, temperatura anual de 15°C. Cuenta con pluviosidad anual entre 700-110 y 75% de humedad relativa. Una velocidad de viento de entre 4m/s y 5,5 m/s (Emov Ep, 2014; Emov Ep, 2016).

2.8.2 Materiales

2.8.2.1 Datos

Los datos fueron recolectados desde el año 2012 por la Red de Monitoreo de Cuenca, la cual registra datos a través de una estación automática instalada en el Centro Histórico, por lo tanto, esta infraestructura almacena tanto niveles de contaminación atmosférica como valores de variables meteorológicas (Emov Ep, 2016) en un rango de 4 kilómetros a la redonda.

Los datos utilizados en el desarrollo del trabajo competen a un periodo de tiempo entre el día 16 de septiembre del año 2018 y el día 14 de octubre del año 2018, e involucran mediciones tomadas en lapsos de 10 minutos. Las variables que se utilizarán para el desarrollo del trabajan son:

- Contaminantes Atmosféricos: Ozono (O3), Monóxido de Carbono (CO), Dióxido de nitrógeno (NO2), Dióxido de Azufre (SO2), Material Particulado (PM2,5).
- Variables Meteorológicas: Temperatura del Aire, Humedad Relativa, Punto del rocío, Presión Atmosférica, Radiación Global, Precipitación, Velocidad del viento, Radiación ultravioleta A, Radiación ultravioleta E.

Los conjuntos de datos tanto de componentes atmosféricos y variables meteorológicas proporcionados originalmente contienen valores como máximo, mínimo y promedio de cada variable. Por ello, se usará valores promedios, esto se debe a que, tanto valores máximos como mínimos representan situaciones poco comunes dentro del conjunto de datos (Andrade & Orellana, 2018). Por otro lado, del conjunto de datos se descartaron valores pertenecientes a la dirección del viento, dato que no es relevante para el desarrollo de este proyecto.

2.8.2.2 Software

2.8.2.2.1 RapidMiner

RapidMiner es una plataforma software de ciencia de datos caracterizada por ser abierta y extensible. Un aspecto importante a mencionar es la funcionalidad que ofrece RapidMiner, es decir, unifica todo el ciclo de vida de la ciencia de datos., iniciando en la preparación de los datos, dentro de esta etapa ofrece soluciones rápidas a problemas como valores perdidos y atípicos, posteriormente, aplica aprendizaje automático y finalmente, implementa un modelo de predicción. Varias empresas utilizan productos y funcionalidades de RapidMiner con la finalidad de generar ingresos, reducir costos y evitar riesgos. Particularmente RapidMiner posee portabilidad, por ello, es ejecutado en varias plataformas, a más de ser una herramienta gratuita (Rapidminer, 2019).

RapidMiner es usado como herramienta para el análisis de contaminantes atmosféricos y variables meteorológicas (Andrade & Orellana, 2018). Permite discretizar el set de datos tanto de contaminantes atmosféricos y variables meteorológicas. En base al conjunto de datos resultante se aplica algoritmos de minería de datos y se evalúan los resultados de los algoritmos aplicados (Rapidminer, 2019).

2.8.2.2.2 Python

Python es un lenguaje de alto nivel orientado a objetos, caracterizado por su semántica dinámica, por ello, es usado para el desarrollo ágil de aplicaciones. Cabe mencionar que su sintaxis simple es una ventaja, puesto que, reduce el costo de mantenimiento del software, otra ventaja en particular de Python es que posee extensas librerías permitiendo la reutilización de código (Python, 2019) esto lo hace atractivo para un ambiente de desarrollo de *machine learning*, en base a que posee una gran colección de paquetes de códigos de repositorios de código abierto, algunas de estas librerías son *numpy* para trabajar con imágenes y texto, *Lobrosa* para el procesamiento de archivos de audio, *Panda* para *machine learning*, además de una gran variedad de librerías para el procesamiento de datos.

2.8.3 Métodos

La minería de datos analiza gran cantidad de datos para obtener información y conocimiento, que será utilizada en diferentes pronósticos. Durante los últimos años, la minería de datos ha tenido un enorme crecimiento, es decir, existen varios modelos estándar para la minería de datos, todos estos enfocados en pasos secuenciales, cada uno de estos pasos ayudan al desarrollo e implementación de tareas de minería de datos (Shafique & Haseeb, 2014). Por ello, en este trabajo se utilizará el modelo CRISP-DM.

2.8.3.1 CRISP-DM

El modelo a utilizar en este trabajo es CRISP-DM, este modelo proporciona una visión de ciclo de vida de un proyecto de minería de datos. Incluye fases del proyecto, tareas a realizar, y finalmente obtención de resultados (Ferna, 2010). El ciclo de vida de CRISP-DM contiene las siguientes etapas: comprensión del negocio o problema, comprensión de los datos, preparación de los datos, modelado, evaluación e implementación.

Se debe mencionar que, la secuencia de las fases no es secuencial esto quiere decir que indican tan solo el flujo de los procesos más importantes y frecuentes entre sí, sin embargo, durante el desarrollo del proyecto depende de qué tarea en particular se debe realizar a continuación. Al ser un proceso cíclico la minería de datos no termina al implementarse la solución. Por el contrario, se presentan nuevas incógnitas con mayor especificidad (Ferna, 2010).

2.8.3.2 Reglas de asociación

Las reglas de asociación, son utilizadas para descubrir elementos que se relacionan entre sí de manera frecuente. Por ejemplo, para determinar que productos compran los clientes a menudo de manera conjunta, por lo tanto, los artículos que se compren juntos se pueden colocar próximos entre ellos en los estantes. Otra aplicación, es el análisis de texto, es decir, clasificar y recuperar documentos (Ye, 2015).

La ventaja de los algoritmos de reglas de asociación, es que las asociaciones se pueden establecer entre cualquier atributo de un conjunto de variables, en otras palabras, buscan una serie de reglas para cada una de las cuales se interpreta una conclusión diferente, por otra parte, los algoritmos de asociación buscan también patrones en grandes conjuntos de datos, por lo que, suelen ocupar un mayor tiempo de ejecución (IBM, 2019). Medidas como soporte, confianza y lift son utilizados para encontrar subconjuntos de elementos de distintos conjuntos de elementos.

- Soporte: Hace referencia a la proporción de registros de datos que contiene un conjunto (X) (Ye, 2015).
- Confianza: Mide la proporción de registros que contiene tanto un conjunto A como un conjunto C (Ye, 2015).
- *Lift*: Es la proporción basada en el soporte de un conjunto de productos frente al soporte teórico esperado (Aggarwal, 2015).

2.8.3.3 A priori

El algoritmo Apriori brinda un procedimiento eficiente para formar conjuntos de elementos frecuentes. Por lo tanto, "Un conjunto de elementos puede ser un conjunto de elementos frecuentes, solo si todos sus subconjuntos son conjuntos de elementos frecuentes" (Ye, 2015).

Apriori inicialmente encuentra los *itemsets* frecuentes, es decir, conjuntos de elementos que tienen mayor número de frecuencia, sin embargo, para que la generación de *itemsets* sea continua se debe establecer un umbral de soporte inicial, por ello, es necesario generar todos los *itemsets* posibles y para cada uno de ellos definir su frecuencia, a continuación, los ítems candidatos serán generados en un inicio de 1 ítem, posteriormente de 2 *items* y así de manera continua, sin embargo, esta continua generación de conjunto de *items* demanda alto costo computacional. Por lo que, basados en el principio de montonocidad que menciona, si un *itemset* es frecuente entonces todos los subconjuntos de este también son frecuentes, de manera análoga, si un *itemset* no es frecuente entonces cualquier conjunto que contenga a este *itemset* tampoco será

frecuente, por ello, no es viable generar todos los *itemsets* posibles de una base de datos. Por lo tanto, la generación de *itemsets* es continua hasta que no existen *itemsets* que aprueben el soporte inicial definido, finalmente este proceso nos da como resultado un conjunto de datos con los *items* entre los cuales existe asociación.

2.9 Alcances y resultados esperados

Dentro del alcance del proyecto se pretende realizar discretizaciones adecuadas tanto de componentes atmosféricos como de variables meteorológicas para posteriormente aplicar técnicas de asociación. De esta manera, se identificarán patrones relevantes que determinen relación de asociación y finalmente analizar resultados obtenidos de la aplicación de la técnica utilizada con la finalidad de generar conocimiento interpretable para los ciudadanos.

2.10 Supuestos y riesgos

Riesgo	Probabilidad	Solución
Datos atípicos, faltantes o irrelevantes que cambien los resultados de la investigación.	Alta	Realizar una preparación de los datos previo al desarrollo.
Recursos de procesamiento bajo que no satisfagan los requerimientos de procesamiento del algoritmo.	Baja	Incorporación de mayor número de equipos para satisfacer el procesamiento requerido.
Datos recolectados que no presentan la variabilidad suficiente, así como, inconsistencias presentadas dentro de las mismas.	Media	Realizar de mejor manera la etapa inicial, enfocándonos en la preparación y limpieza de los datos.
Mala estimación del tiempo para el desarrollo de la investigación.	Media	Justificar moratoria y solicitar una prórroga para la entrega de la investigación.

2.11 Presupuesto

Rubro-Denominación	Justificación	Costo (Detalle)	Cantidad	Costo Total
Remuneración	Remuneración por trabajos realizado mensualmente	800	4 meses	3200
Impresiones	Impresiones del documento.	50	1	50
Servicio de internet	Conexión a internet para consultas y retroalimentación de la investigación	25	4 meses	100
Total				3350

2.12 Financiamiento: Propio

2.13 Esquema tentativo

Capítulo 1: Introducción.
Capítulo 2: Estado del arte.
Capítulo 3: Aplicación de técnicas de asociación.
4.1 Discretización de datos.
4.2 Aplicación Algoritmo Apriori.
Capítulo 4: Análisis de resultados.
4.1 Evaluación de resultados.
Conclusiones y trabajos futuros
Bibliografía

2.14 Cronograma

Objetivo Especifico	Actividad	Semanas															
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Revisar estudios similares con técnicas de minería de datos.	Revisión de documentación, tutoriales a cerca del algoritmo a utilizar.																
	Aprendizaje de la técnica seleccionada para el desarrollo del proyecto.																
Experimentar con diferentes niveles de discretización de componentes atmosféricos y variables meteorológicas.	Categorizar componentes atmosféricos y variables meteorológicas.																
	Aplicar la algoritmia de Apriori sobre conjunto de datos discretizados.																
Identificar patrones relevantes que determinen la asociación entre las variables atmosféricas y meteorológicas.	Evaluación y Retroalimentación de los resultados obtenidos tras la aplicación del algoritmo Apriori.																
Exponer los resultados obtenidos de la técnica utilizada.	Evaluación de resultados de la aplicación del algoritmo Apriori sobre componentes atmosféricos y variables meteorológicas.																
Escritura del proyecto	Redacción y correcciones respectivas sobre el proyecto de investigación.																

2.15 Referencias

Czechowski, P., Badyda, A., & Majewski, G. (2013). Data mining system for air quality monitoring networks. *Archives of Environmental Protection*, *39*(4), 123–144. https://doi.org/10.2478/aep-2013-0041

Kalyankar, M. A., & Alaspurkar, S. J. (2013). Data Mining Technique to Analyse the Metrological Data. *International Journal of Advanced Research in Computer Science and Software Engineering*, *3*(2), 114–118. Retrieved from www.ijarcsse.com

Kohail, S. N., & El-halees, A. M. (2011). Implementation of Data Mining Techniques for Meteorological Data Analysis (A case study for Gaza Strip). *International Journal of Information and Communication Technology Research*, *1*(3), 96–100.

Kurt, A., Gulbagci, B., Karaca, F., & Alagha, O. (2008). An online air pollution forecasting system using neural networks. *Environment International*, *34*(5), 592–598. https://doi.org/10.1016/j.envint.2007.12.020

Lanjewar, U. M., & Shah, J. J. (2012). Air Pollution Monitoring & Tracking System Using Mobile Sensors and Analysis of Data Using Data Mining. *International Journal of Advanced Computer Research*, (4), 2277–7970.

Martínez-Ballesteros, M., Troncoso, A., Martínez-Álvarez, F., & Riquelme, J. C. (2010). Mining quantitative association rules based on evolutionary computation and its application to atmospheric pollution. *Integrated Computer-Aided Engineering*, *17*(3), 227–242. https://doi.org/10.3233/ICA-2010-0340

Ortega, J., & Orellana, M. (2018). IMPACTO DE LA APLICACIÓN DE ALGORITMOS DE MINERÍA DE DATOS EN VARIABLES DE CONTAMINACIÓN DEL AIRE.Universidad del Azuay. Recuperado de http://e-journal.uajy.ac.id/14649/1/JURNAL.pdf

Andrade, P., & Orellana, M. (2018). APLICACIÓN DE MINERIA DE DATOS EN EL ANÁLISIS DE CONTAMINANTES ATMOSFÉRICOS Y VARIABLES METEOROLÓGICAS. Universidad del Azuay. Retrieved from http://e-journal.uajy.ac.id/14649/1/JURNAL.pdf

Sahafizadeh, E., & Ahmadi, E. (2009). Prediction of air pollution of Boushehr City using data mining. *2nd International Conference on Environmental and Computer Science, ICECS 2009*, 33–36. https://doi.org/10.1109/ICECS.2009.18

Zhang, S., Zhang, C., & Yang, Q. (2003). Data preparation for data mining. *Applied Artificial Intelligence*, *17*(5–6), 375–381. https://doi.org/10.1080/713827180

Andrade, P., & Orellana, M. (2018). APLICACIÓN DE MINERIA DE DATOS EN EL ANÁLISIS DE CONTAMINANTES ATMOSFÉRICOS Y VARIABLES METEOROLÓGICAS. Universidad del Azuay.

Bauerle, L. (2019). Liderazgo. 1–7. https://rapidminer.com/us/#about-us-leadership!loading

Cagliero, L., Cerquitelli, T., Chiusano, S., Garza, P., Ricupero, G., & Xiao, X. (2016). Modeling Correlations among Air Pollution-Related Data through Generalized Association Rules. 2016 IEEE International Conference on Smart Computing, SMARTCOMP 2016. https://doi.org/10.1109/SMARTCOMP.2016.7501707

Emov Ep, R. D. M. D. L. C. D. A. D. C. (2014). Resumen del inventario de emisiones atmosféricas del cantón Cuenca, año 2011. Cuenca-Ecuador. (DECEMBER 2014).

Emov Ep, R. D. M. D. L. C. D. A. D. C. (2016). Inventario de Emisiones Atmosféricas de Cuenca. (Diciembre), 163. https://doi.org/10.13140/RG.2.2.17665.66405

Ferna, C. (2010). A survey of data mining and knowledge discovery process models and methodologies. 25, 137–166. https://doi.org/10.1017/S0269888910000032

INEC. (2017). Resultados Censo de Población. Población y Demografía, 3–4. Recuperado de http://www.ecuadorencifras.gob.ec/censo-de-poblacion-y-vivienda/

Aggarwal, C. (2015). Data Mining The Textbook. Recuperado de https://www.springer.com/la/book/9783319141411

Modeler, I. B. M. S., Anterior, S., & Realizar, A. (2019). Reglas de asociación. 5–8. https://www.ibm.com/support/knowledgecenter/es/SS3RA7_sub/modeler_mainhelp_client_ddita/clementine/nodes_associationrules.html

Ncr, P. C., Spss, J. C., Ncr, R. K., Spss, T. K., Daimlerchrysler, T. R., Spss, C. S., & Daimlerchrysler, R. W. (2000). Crisp-dm 1.0. https://www.the-modeling-agency.com/crisp-dm.pdf

Shafique, U., & Haseeb, Q. (2014). A Comparative Study of Data Mining Process Models (KDD , CRISP-DM and SEMMA). 12(1), 217–222.

Heroku, E. (2019). ¿ Qué es Python ? Resumen ejecutivo ¿ Qué es Python ? Resumen ejecutivo. 2–3. https://www.python.org/doc/

Ye, N. (2015). Data Mining Theories, Algorithms and Examples. In IEEE Potentials (Vol. 16).

Aggarwal, C. (2015). *Data Mining The Textbook*. Retrieved from https://www.springer.com/la/book/9783319141411

2.16 Firma de responsabilidad (estudiante)

2.17 Firma de responsabilidad (director sugerido)

2.18 Fecha de entrega:

13 de Junio del 2019