UNIVERSIDAD DEL AZUAY

FACULTAD DE CIENCIAS DE LA ADMINISTRACIÓN

ESCUELA DE INGENIERÍA DE SISTEMAS Y TELEMÁTICA

EVALUACIÓN DE LAS HERRAMIENTAS DE MINERÍA DE DATOS EN VARIABLES DE CONTAMINACIÓN ATMOSFÉRICA

TRABAJO DE GRADO PREVIO A LA OBTENCIÓN DEL TÍTULO DE INGENIERO DE SISTEMAS Y TELEMÁTICA.

AUTOR: MARCIA ALEXANDRA MATUTE RIVERA.

DIRECTOR: ING. CHESTER ANDREW SELLERS WALDEN M. Sc.

CUENCA, ECUADOR

2018

DEDICATORIA

El presente trabajo de tesis va dedicado principalmente a Dios por darme salud y fuerza de voluntad para afrontar las dificultades presentes en esta vida.

A mis padres y hermanas por el apoyo incondicional en todo momento, por sus consejos, sus valores y el arduo trabajo que han realizado guiándome por el buen camino hacia el éxito.

A mi esposo e hija por ser mi inspiración, admiración y fortaleza, sobre todo por la motivación constante y ser pilar fundamental en mi vida.

AGRADECIMIENTO.

Quiero agradecer a los profesores de la Universidad del Azuay, especialmente a mi director de tesis Ing. Chester Sellers por haberme guiado y apoyado durante esta investigación que me condujo a la obtención del título.

A la Ing. Belén Arias por compartir sus conocimientos y consejos en el desarrollo de esta tesis.

Al Ing. Juan Lima por haber colaborado en proceso de tratamiento de limpieza de los datos utilizados para las pruebas en esta tesis.

A mi familia por la confianza otorgada, por sus palabras de aliento, este nuevo logro es en gran parte gracias a ustedes.

Contenido

Índice de Tablas

Índice de Imágenes.

RESUMEN.

Este estudio se orientó a la evaluación de las herramientas de minería de datos: RapidMiner, KNime Weka, IBM SPSS Modeler y SAS para minería de datos en variables de contaminación atmosférica: Ozono (O_3), Monóxido de Carbono (CO), Dioxido de Azufre (SO_2), Dióxido de Nitrógeno (NO_2) y Material Particulado 2,5um ($PM_{2,5}$), estos datos fueron recogidos y almacenados por los sistemas de monitoreo continuo de la calidad del aire del municipio de Cuenca. Para el tratamiento de los datos las herramientas aplican el algoritmo k-means y funciones matemáticas como conversión de valores para descubrir correlaciones entre los datos; con los resultados obtenidos en cada herramienta se realizó un análisis de matriz comparativa con indicadores como: funcionalidad, usabilidad, rendimiento, soporte de tareas auxiliares, entre otros, de acuerdo a los resultados obtenidos la mejor herramienta para minería de datos en variables de contaminación atmosférica es RapidMiner.

ABSTRACT

This study was oriented to the evaluation of data mining tools RapidMiner, KNime Weka, IBM SPSS Modeler and SAS in variables of atmospheric pollution such as Ozone (O_3), Carbon Monoxide (CO), Sulfur Dioxide (SO_2), Dioxide of Nitrogen (NO_2) and Particulate Material 2.5um ($PM_{2,5}$). These data were collected and stored by the continuous air quality monitoring systems of the municipality of Cuenca. The tools applied the k-means algorithm and mathematical functions as conversion of values to process the data and discover correlations between them. With the results obtained in each tool, a comparative matrix analysis was performed with indicators such as functionality, usability, performance, support of auxiliary tasks and others. According to the obtained results, the best tool for data mining in air pollution variables was RapidMiner.

UNIVERSIDAD DEL AZUAY
Dpto. Idiomas

Translated by

Ing. Paul Arpi

Capítulo I: Introducción

Introducción

Este trabajo se orienta a investigar las herramientas de minería de datos más apropiadas para el análisis sobre datos recopilados de manera sistemática por la estación de monitoreo continuo de la calidad del aire del municipio de Cuenca. Se analizará el comportamiento de las 5 variables ambientales recogidas en la ciudad de Cuenca por dicho sistema. Las variables de estudio son los principales generadores de la contaminación del aire: Ozono (O_3), Monóxido de Carbono (CO) Dióxido de Azufre (SO_2), Dióxido de Nitrógeno (NO_2) y Material Particulado $_{2,5}$ um ($PM_{2.5}$). Para descubrir los patrones de comportamiento es necesario establecer correlaciones entre las variables de estudio según una cronología y frecuencia determinada además de la influencia de otras variables como son las meteorológicas. Para la implementación del proyecto será necesario experimentar con varias herramientas de minería de datos esto derivará en el desarrollo de una matriz que permita analizar y descubrir cuál es la herramienta más apropiada para procesar los datos de los contaminantes mencionados.

1.1. Definición de Objetivos

Objetivo general:

- Evaluar herramientas y técnicas de minería de datos aplicables a los agentes de contaminación atmosférica obtenidos de la estación de monitoreo continuo en la ciudad de Cuenca, determinando patrones de comportamiento entre las variables estudiadas.

Objetivos específicos:

- Conocer los proyectos que ha emprendido el IERSE y otros autores en cuanto a la captura, proceso y visualización de la calidad de aire en la ciudad de Cuenca.
- Elaborar un marco teórico sobre las técnicas de minería de datos aplicables al problema.
- Evaluar la capacidad de las herramientas de minería de datos para el análisis de la información recolectada.
- Establecer la mejor herramienta de minería de datos aplicable al problema.

1.1. Justificación.

El Instituto de Estudios de Régimen Seccional del Ecuador (IERSE) ha establecido un índice de calidad del aire, el cual lo ha publicado a través de una web personalizada para el efecto. Es parte esencial que en esta fase del proyecto se produzca lo que se denomina como el descubrimiento de conocimiento de la base de datos (KDD, Knowledge Discovery in Databases); este proceso tiene que ver con la aplicación de técnicas de minería de datos (DM, Data Mining) para el análisis científico de la información recolectada.

Por lo que, este trabajo se orienta al análisis de las herramientas de minería de datos más apropiadas para realizar el análisis de datos recopilados por la estación de monitoreo. Las herramientas aplican algoritmos y funciones matemáticas para descubrir patrones de comportamiento y relaciones entre los datos, lo que significa evaluar estas herramientas tomando en cuenta factores, como: gama de algoritmos disponibles, usabilidad, seguridad, viabilidad, velocidad, entre otras; la importancia de este trabajo radica en que permitirá descubrir patrones de comportamiento que derivan en conocimiento para realizar alertas tempranas que coadyuven a la toma de decisiones por parte de organismos de regulación municipal.

1.2. Alcance y Limitaciones.

Este trabajo tiene como alcance evaluar herramientas de minería de datos utilizando variables de contaminación atmosférica en la ciudad de Cuenca y obtener una matriz comparativa que nos indique cuál es la mejor herramienta para procesar los datos de las variables de contaminación atmosférica.

Capítulo II: Marco Teórico

Introducción.

En este capítulo se describe información sobre antecedentes y proyectos realizados en el IERSE en contraste con otros escenarios con estudios similares, un marco teórico sobre las técnicas de minería de datos y su aplicación en problemas de contaminación ambiental como Ozono (O_3), Monóxido de Carbono (CO) Dióxido de Azufre (SO_2), Dióxido de Nitrógeno (NO_2) y Material Particulado $_{2,5}$ um ($PM_{2.5}$).

2.1. Antecedentes

En la actualidad es posible almacenar altos volúmenes de datos generados por diferentes sistemas o fuentes de información, es así que las diferentes empresas u organizaciones tienen acceso a una extensa cantidad de datos relacionados a sus actividades habituales dentro de esto información útil para proyectos científicos. Dichos datos prometen a las empresas una visión clara de sus actividades, como están realizando estas actividades e incluso ayudan en el proceso de toma de decisiones, no obstante la información almacenada en las bases de datos son de difícil entendimiento para personas no especializadas, por lo que es necesario un proceso de análisis y tratamiento de los datos recogidos.

Existen diversas formas con las que podemos recoger información empezando desde registros, entrevistas, observación directa o por medio de software especializado para la adquisición de datos como puede ser un sistema de monitoreo y control de la calidad del aire.

En este trabajo nos enfocaremos en el tratamiento de los datos recogidos y almacenados por los sistemas de monitoreo continuo de la calidad del aire, este sistema se especializa por poseer un gran número de registros de información y son utilizados por personas expertas que a su vez son capaces de entender los efectos de los contaminantes en la salud de las personas.

Con el constante desarrollo económico y social de las grandes urbes se presentan incrementos en los niveles de contaminación debido a mayor demanda de energía y consumo precipitado de combustible fósil. La calidad de aire tiende a disminuir y tiene relación directa con problemas de salud de las personas convirtiéndose en un problema de particular interés para las autoridades de grandes ciudades y de los ciudadanos en

general. A pesar de los esfuerzos por apalear los efectos nocivos de la contaminación, los resultados son cada vez más evidentes en la salud de las personas y en el medio ambiente. Como lo enuncia Gaitán, Cancino, & Behrentz en su artículo "Análisis de la calidad del aire en Bogotá": el crecimiento se ve afectado por una mayor demanda de energía así como un acelerado consumo de combustibles fósiles (Gaitán, Cancino, & Behrentz, 2007). Al ser este un problema social; se han realizado varias iniciativas por evidenciar los fenómenos de la contaminación, en específico sobre la contaminación atmosférica.

El monitoreo de la calidad de aire no es una práctica nueva, se han venido implementando redes fijas de monitoreo de la calidad del aire, las siguientes ciudades pertenecientes a los diferentes países son un ejemplo: la zona del Valle de México, La ciudad de Beijing (China), existe una red de monitoreo ambiental en Londres, Red de Monitoreo de la Calidad del Aire de Bogotá, Redes de Monitoreo de Calidad del Aire para ciudades de Bolivia, etc. (Gaitán, Cancino, & Behrentz, 2007)

Estos sistemas de monitoreo por medio de sensores recogen información y la almacenan en una bases de datos, dichos datos son difíciles de interpretar para personas no especializadas por lo que para su utilización e interpretación es necesario un proceso de tratamiento y análisis de los datos allí recogidos, este proceso se llama minería de datos. Los trabajos existentes sobre el análisis de contaminación del aire implementan diferentes técnicas de minería de datos, es por ello indispensable convertir grandes cantidades de datos que existen en información que nos proporcionen experiencia y conocimiento, y que a su vez, esta información ayude en la toma de decisiones adecuadas para cada problemática que se afronta.

2.2.Trabajos relacionados.

Así como las grandes urbes han conseguido la implementación de redes de monitoreo y técnicas de análisis de datos, la Universidad del Azuay, a través de su centro de investigación IERSE "Instituto de Estudios de Régimen Seccional del Ecuador", impulsó el proyecto de calidad del aire por medio del GAD municipal de la ciudad de Cuenca y la Empresa EMOV-EP. De acuerdo a los datos recopilados mediante el sistema de monitoreo atmosférico, las concentraciones de las partículas totales en suspensión ocasionalmente están por encima de los niveles de las normas permisibles y los niveles de los gases en el aire, tales como Dióxido de Azufre (SO2), Dióxido de Nitrógeno (NO2)

y Monóxido de Carbono (CO) aunque están por debajo de las normas de la calidad del aire, tienen una tendencia creciente y en algunos casos se acercan o rebasan los valores permisibles. (Dominguez, Torres, Ortiz, Prijodko, & Korc, 2011). Mediante la aplicación de los programas y proyectos sobre la evaluación de la situación actual del recurso aire, fue posible identificar los problemas provocados por la contaminación de carácter antropogénico en la ciudad de Cuenca (Sellers, 2012), esto para priorizar las líneas de acción y mitigar el deterioro de la calidad atmosférica. (Díaz, 2015)

Los proyectos hasta ahora desarrollados han centrado sus esfuerzos en la captura, tratamiento y visualización de las variables monitoreadas por la estación; al punto de generar una escala que califica el aire que se respira en la ciudad de Cuenca de acuerdo a las normas nacionales (TULSMA) y normas internacionales (EPA) y esta información se publica a través de una página Web.

Con el pasar de los años este proyecto ha ido recolectando información con la que se aspira obtener patrones de comportamiento entre los contaminantes atmosféricos mediante la aplicación de técnicas de minería de datos.

2.3.Minería de datos aplicada a la Calidad del Aire.

Es importante mencionar que la contaminación atmosférica se entiende como la introducción de materias o productos nocivos en la atmosfera que es la que provee de oxigeno vital que necesitamos, estos contaminantes intervienen en las características naturales del ambiente y tienen influencia negativa en la salud de todos los seres vivos. La contaminación atmosférica es provocada por causas naturales así como también los ocasiona el ser humano (antrópicas), es decir, causas naturales como los gases y cenizas de la actividad volcánica, incendios forestales, polvo, etc, y entre las causas de contaminación por parte del hombre está el uso de vehículos automotores, el uso de pinturas, productos en aerosol, las plantas de energía, fabricas etc. (Martínez & Díaz, 2004)

La calidad del aire se mide mediante redes de control y monitoreo permanente del aire, ya sean estas públicas y/o privadas, estas redes están compuestas por una o varias estaciones de control en las cuales se miden diversos contaminantes y parámetros meteorológicos de forma continua. Estos datos se almacenan y envían a un centro de control para su gestión.

Los principales contaminantes en el aire son:

Ozono (O3).- el Ozono troposférico se encuentra a nivel del suelo, es un gas irritante que afecta a los ojos, nariz, garganta, ocasiona tos y problemas pulmonares, también afecta a la producción de cultivos. El O_3 se genera por la combinación del óxido de nitrógeno NO_x y compuestos orgánicos volátiles *COV* que son generados por la combustión de gasolina. Estos compuestos forman una neblina que es visible en zonas muy contaminadas, esta neblina es también llamada smog fotoquímico. (Sellers, 2017), (Gaitán, Cancino, & Behrentz, 2007), (Martínez & Díaz, 2004) , (Torres, 2002).

Monóxido de Carbono (CO).- es uno de los principales contaminantes del aire, es producido por los tubos de escape de los vehículos que usan gasolina, el *CO* es un gas venenoso que se forma por la quema parcial de gasolina petróleo, madera, etc. Puede ocasionar graves daños a la salud incluso provocar la muerte ya que al reducir el oxígeno en la sangre, en el cerebro y corazón. (Sellers, 2017), (Gaitán, Cancino, & Behrentz, 2007), (Martínez & Díaz, 2004).

Dióxido de Azufre (SO2), es un gas irritante que afecta al sistema respiratorio, además el SO_2 al entrar en contacto con la humedad del aire forma el llamado ácido sulfúrico H_2SO_4 que tiene efectos negativos en los suelos, ríos y en las plantas, afecta a los edificios y monumentos ya que provocan la corrosión del material (Sellers, 2017), (Gaitán, Cancino, & Behrentz, 2007), (Martínez & Díaz, 2004).

Dióxido de Nitrógeno (NO2).- es un gas irritante, puede llegar a ser toxico en altas concentraciones y existe un alto riesgo de aumentar infecciones pulmonares. El NO_2 se forma de la combustión en altas temperaturas como en el motor de los vehículos, la gasolina, plantas industriales y de manera natural como es la actividad volcánica (Sellers, 2017), (Gaitán, Cancino, & Behrentz, 2007), (Martínez & Díaz, 2004).

Material Particulado $_{2,5}$um ($PM_{2.5}$).- Son la combinación de partículas líquidas y sólidas que se hallan en suspensión en el aire. Al $PM2{,}5$ se le llaman partículas finas o ultra finas ya que su diámetro es igual o menor a 2,5 μm, dichas partículas pueden ser excesivamente dañinas ya que al ser tan finas pueden moverse con mayor facilidad por las vías respiratorias y penetrar en los pulmones, estas partículas causan irritación, en los ojos, nariz y garganta . (Fundación para la Salud Geoambiental, 2013), (Sellers, 2017), (Martínez & Díaz, 2004).

Resulta que la Calidad del Aire es uno de esos campos en que los datos son de difícil entendimiento y tratamiento por otros medios que no sean la minería de datos, y esto lo lleva a convertirse en uno de los campos donde es más útil e interesante aplicar técnicas de análisis como es la minería de datos.

Para llevar a cabo un trabajo de minería de datos es importante tener un área de estudio de la que deseemos recoger información y poder aplicar allí diversas técnicas, modelos, incluso usar herramientas (Software) de minería de datos que se pueden adquirir en el mercado.

2.4. Minería de Datos

Hoy en día gracias a los avances de la tecnología se puede almacenar grandes cantidades de información en bases de datos. Contar con una base de datos y las herramientas necearías para su análisis ayudan a descubrir información importante que sea de interés y a su vez obtener conocimiento por medio del análisis de estos datos. No obstante una base de datos con gran volumen de información se vuelve una limitante para análisis manuales en atención a lo cual se han desarrollado tecnologías especializadas que faciliten su tratamiento. Una de estas tecnologías es la minería de datos que proporciona mecanismos útiles para extraer información de interés de una base de datos.

En este capítulo se enfocará en la definición de la minería de datos y sus fundamentos principales. Un aspecto fundamental es que para la aplicación de las técnicas de minería de datos debe producirse lo que se denomina como el descubrimiento de conocimiento de la base de datos (KDD, Knowledge Discovery in Databases). De la misma forma, se revisan modelos, tareas, procesos metodológicos y técnicas más empleadas en la minería de datos.

2.4.1. ¿Qué es la minería de Datos?

En la actualidad la minería de datos cumple un papel importante como tecnología de apoyo para la exploración, análisis, compresión y aplicación del conocimiento adquirido de grandes volúmenes de datos, asimismo ayuda a la identificación de tendencias y comportamientos que proporcionan una mejor compresión de los fenómenos que nos rodean y contribuyen a la mejora en toma de decisiones.

La definición de Minería de Datos puede variar entre los diferentes investigadores ya sean estadísticos, analistas de datos u otros. Según Hand, Witten y Sumathi, la minería de datos se define como: "el procesamiento y análisis de conjuntos de datos, para extraer información de importancia." (Díaz & Suárez, 2009)

Analizar grandes volúmenes de datos de forma tradicional (manualmente) es imposible por lo que se necesita de técnicas y herramientas informáticas que sean capaces de realizar este análisis de forma automática y ayuden a los expertos a exponer dicha información de una forma entendible y que las personas no expertas puedan entenderla. Entonces se entiende que: "la minería de datos es la extracción de información de grandes cantidades de datos y esta información a su vez debe ser válida y útil para descubrir reglas o patrones significativos que ayuden en el momento de la toma de decisiones".

La minería de datos es un proceso dentro de un proceso más amplio conocido como extracción de conocimiento en bases de datos (KDD), que es descubrir patrones de comportamiento en grandes volúmenes de datos.

La minería de datos tiene la capacidad de informar sobre eventos que son importantes y que hasta el momento son desconocidos, por ejemplo, Informes de tráfico de red, Predicción de incendios forestales, Detección y prevención de fraude, tendencia de abandono de clientes, etc. permitiendo realizar pronósticos fiables haciendo posible emprender acciones correctivas o de control en ambientes con un riesgo de impacto mínimo. (De la Puente, 2010)

2.4.2. Técnicas de minería de datos.

Las técnicas de minería de datos permiten realizar tareas para: identificar patrones de comportamiento que expliquen y/o ayuden a entender los datos, tareas que permiten predecir valores futuros de variables que son de interés.

En la minería de datos existen diferentes técnicas como:

- **Sistemas de agrupamiento**: Consiste en obtener grupos o conjuntos con similitudes a partir de los datos recogidos. Esta técnica se la conoce como

segmentación o clustering y es una técnica no supervisada[1]. (Beltrán, 2010), (García, 2013)

- **Reglas de asociación:** Es una técnica no supervisada, su objetivo es identificar relaciones similares existentes en grupos de datos específicos. (Microsoft, 2016), (García, 2013), (Beltrán, 2010)
- **Correlaciones**: esta técnica proporcionan información sobre el grado de similitud entre variables de distintos conjuntos de datos. (Beltrán, 2010), (García, 2013)
- **Clasificación:** Se la denomina también técnicas de predicción, cuyo objetivo es predecir la clase a la que pertenecen nuevos objetos a partir de las variables restantes. (Beltrán, 2010), (García, 2013)
- **Regresión**: Su objetivo es predecir una o más variables continuas, basándose en otros atributos del conjunto a partir de la evolución de otra variable continua. (Beltrán, 2010), (García, 2013)

Para poder aplicar las técnicas de minería de datos se debe de tener un objetivo claro ya que las herramientas de minería de datos funcionan de mejor manera cuando se sabe que buscar en concreto.

2.4.3. ¿Qué es el Descubrimiento de conocimiento de la base de Datos (KDD, Knowledge Discovery in Databases)?

Es un tipo de inducción al conocimiento, para poder decidir que técnica es la más apropiada para cada situación, es necesario diferenciar que información deseamos extraer de los datos que almacenamos (Beltrán, 2010). Entonces para cada nivel de abstracción el conocimiento contenido en los datos se puede clasificar en distintas categorías y cada una requerirá una técnica más o menos avanzada para su recuperación como se detalla en (Daedalus S.A., 2002):

• **Conocimiento profundo.-** información que se encuentra almacenada en los datos, pero que es imposible de recuperar a menos que se tenga alguna clave que ayude a orientarse en la búsqueda.

[1] **Técnica no supervisada**: esta técnica es también conocida como técnica de descubrimiento del conocimiento, como su nombre lo dice es utilizada para el descubrimiento de patrones de comportamiento que están ocultos en grandes volúmenes de datos.

- **Conocimiento oculto.-** se refiere a la información desconocida y que es potencialmente útil y puede recuperarse mediante la aplicación de técnicas de minería de datos.

- **Conocimiento evidente.-** es aquella información que se puede recuperar fácilmente mediante una simple consulta en la base de datos por medio de sentencias SQL.

Esta tesis se enfoca en el descubrimiento del conocimiento oculto, se analizará la información de la base de datos que posee la estación de monitoreo continuo de la calidad de aire de la ciudad de Cuenca.

Como es evidente, las técnicas de minería de datos son herramientas que facilitan el descubrimiento de la información, es por ello que a continuación nos adentraremos en el tema.

2.5. Proceso de la minería de datos – Metodología CRISP-DM

La metodología para el desarrollo de este proyectos es CRISP-DM ((Cross Industry Standard Process for Data Mining), ya que es considerado uno de los principales modelos utilizados en el ámbito académico para el desarrollo de proyectos de minería de datos según una encuesta realizada en el 2014 por KDnuggets. (Piatetsky-Shapiro & Mayo, 2017)

CRISP-DM es un método que ayuda a orientar el trabajo de minería de datos, es decir, describe las fases, las tareas y como se relacionan cada una de las fases que normalmente se siguen en un proyecto, CRISP- DM presenta las fases como ciclo de vida de un proyecto. (Piatetsky-Shapiro & Mayo, 2017)

Este modelo tiene 6 fases, algunas de estas fases son bidireccionales permitiendo regresar y revisar las fases anteriores de ser necesario.

CRISP-DM es un método flexible y puede personalizarse creando así un modelo que se adapte a las necesidades de cada proyecto.

La siguiente imagen muestra las fases del modelo de referencia CRISP-DM:

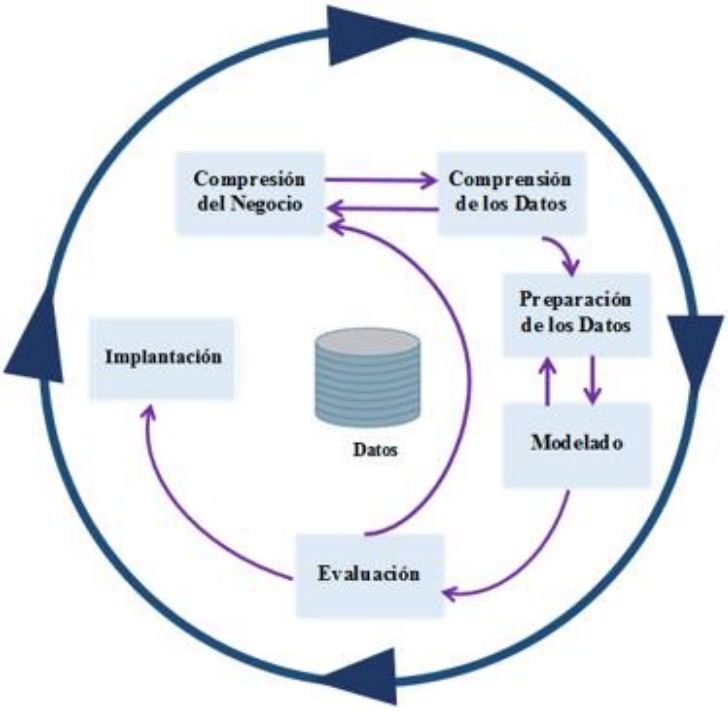


Imagen 1. Fases del modelo de referencia CRISP-DM

(Fuente: [CRISP-DM, 2000])

Comprensión del Negocio.- en esta fase se describe los objetivos y requerimientos para el desarrollo del proyecto, esta es la fase más importante, ya que, si no tenemos claro los objetivos de la empresa o no entendemos la problemática que existe no podremos obtener resultados favorables al emplear estos algoritmos. (Rodríguez Rojas, 2010)

En esta fase debemos realizar algunas tareas como:

- Establecer los objetivos del negocio.
- Establecer los requisitos del proyecto.
- Establecer los objetivos de la minería de datos.
- Generar el plan del proyecto.

Comprensión de los Datos.- se inicia con la identificación de los datos, es decir, que datos se necesitan, en donde los podemos encontrar y en que formatos se encuentran, para luego prepararlos e incluirlos en bases de datos en un formato adecuado. Según (Beltrán, 2010) esta es una de las tareas más difíciles de la minería de datos.

En esta fase se realizan algunas tareas como:

- Recopilar información inicial.
- Descripción de los datos.

- Exploración de los datos.
- Verificación la calidad de los datos.

Preparación de los datos.- en esta fase se preparan los datos obtenidos en la fase anterior, adaptándolos para aplicar técnicas de minería de datos, se filtran los datos, se eliminan los datos incorrectos, no válidos y desconocidos, se aplican estrategias para manejar valores ausentes, normalización de los datos, colocar los datos en el formato adecuado etc. (Rodríguez Rojas, 2010)

En esta fase debemos realizar algunas tareas como:

- Selección de los datos.
- Limpieza de los datos.
- Construcción de los datos.
- Integración de los datos.
- Formateo de los datos.

Modelado.- en esta fase antes de proceder a aplicar Data Mining a los datos, es esencial tener una idea de qué es lo que interesa averiguar (pregunta de investigación), tener los datos correctos, tener conocimiento de la técnica y qué herramientas se necesitan para obtenerlos. (Rodríguez Rojas, 2010)

En esta fase debemos realizar algunas tareas como:

- Seleccionar la técnica del modelado.
- Diseño del plan de prueba.
- Construcción del modelo.
- Evaluar el modelo.

Evaluación.- tras aplicar la técnica con la herramienta elegida, hay que saber interpretar los resultados o patrones obtenidos para saber los que son significativos y cómo podarlos para extraer únicamente los resultados útiles. (Rodríguez Rojas, 2010)

En esta fase debemos realizar algunas tareas como:

- Evaluar los resultados.
- Revisar el proceso
- Establecer las siguientes acciones.

Implementación.- Una vez obtenido los resultados que son útiles para nuestro problema, hay que identificar las acciones que se deben tomar, pensar en cómo proceder para llevarlas a cabo e implementar estos procesos. Tras la implementación se debe evaluar el plan para la implementación y para ello hay que observar los resultados, los costos y los beneficios que estos pueden generar. (Rodríguez Rojas, 2010)

2.6. Conclusión.

Es necesario contar con una metodología que sirva de guía para desarrollar correctamente un proyecto, particularmente la metodología CRISP se enfoca en el análisis de datos aplicada a la minería de datos lo que es la opción que se ha seleccionado para el desarrollo de este proyecto, entender los datos no es suficiente se debe empezar por conocer el negocio y la problemática que conlleva y así establecer la fuente de los datos.

Capítulo III: Herramientas de Minería de Datos

Introducción

En este capítulo se profundizará en la investigación relevante sobre las herramientas que permiten aplicar minería de datos en la información recolectada sobre los principales contaminantes que influyen en la calidad de aire de la ciudad de Cuenca. En esta sección se detalla información como: el tipo de licencia, cantidad de datos que soporta, sus características, ventajas y desventajas.

3.1.¿Qué son las Herramientas de minería de Datos?

Son herramientas de software para el desarrollo de modelos de minería de datos tanto libres como comerciales. En la actualidad existe una amplia gama de herramientas se utilizan como apoyo en la extracción de la información de grandes volúmenes de datos.

Las herramientas de minería de datos se pueden clasificar en tres grandes grupos como son:

- **Técnicas de visualización,** ayudan al descubrimiento manual de información, aquí se muestran tendencias, agrupamientos de datos, etc. (De la Puente, 2010) (Valcárcel, 2004).
- **Técnicas de verificación,** se conoce de antemano un modelo y se desea saber si los datos disponibles se ajustan a él. Las medidas que se utilizan en esta técnica son: (De la Puente, 2010) (Valcárcel, 2004).
    - **Soporte:** porcentaje de demandas que cumplen sus condiciones.
    - **Precisión:** porcentaje de casos en los que la regla se cumple.
- **Método de descubrimiento:** se busca un modelo desconocido, en los que se han de encontrar patrones e información potencialmente importante. (De la Puente, 2010) (Valcárcel, 2004).

3.2.1. WEKA

WEKA, acrónimo de Waikato Environment for Knowledge Analysis, es un conjunto de librerías JAVA para la extracción de conocimiento desde bases de datos. Está constituido por una serie de paquetes de

código abierto con diferentes técnicas de preprocesado, clasificación, agrupamiento, asociación y visualización.

Se trata de un software desarrollado en la Universidad de Waikato con sede en Nueva Zelanda y al ser de código abierto se encuentra bajo licencia GNU-GPL lo cual ha impulsado que sea una de las suites más utilizadas en el área en los últimos años. (University of Waikato, 2016)

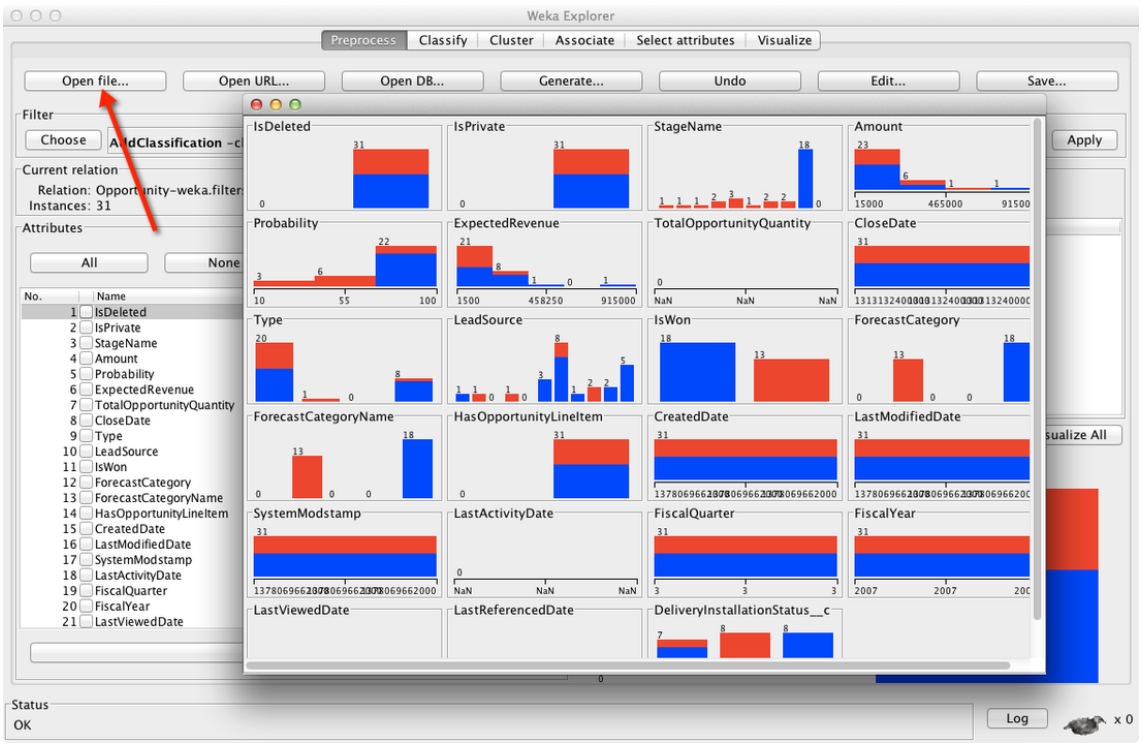


Imagen 2. Interfaz de la herramienta Weka

(Fuente: www.dataminingcrm.com ,2017)

3.2.1.1. Licencia de WEKA

Weka es un software de código abierto emitido bajo la Licencia Pública General de GNU y puede ser de acceso y distribución pública.

3.2.1.2. Características:

- Está implementado en Java, es muy portable.
- Posee una amplia gama de técnicas para procesamiento de datos.
- Permite combinar modelos.
- Interfaz gráfica amigable fácil de usar.
- Acceso a base de datos por medio de conexión JDBC (Java Database Conectivity).

3.2.1.3. Sistemas Operativos

WEKA al ser un software multiplataforma tiene compatibilidad con sistemas operativos como:

- Windows.
- Mac OS.

3.2.1.4. Áreas de Implementación

Weka al igual que las otras herramientas que aplican técnicas de minería de datos da solución a problemas de una variedad de áreas, algunas de ellas son:

Área (Casos de Uso)	**Solución**
Costumer Intelligence (Inteligencia del cliente)	• Percepción de la marca • Análisis de los clientes • Pronóstico de Abandono
Tecnología	• Detección y prevención de fraude. • Detección de Fallas.
Costumer Services (Servicio al cliente)	• Satisfacción del cliente. • Quejas y reclamos
Finanzas	• Límites de crédito. • Clasificación de Socios. • Movimiento de efectivo.
Fabricación	• Análisis de la calidad de fabricación y consumo de energía.

3.2.1.5. Fortalezas

- Soporta varias tareas de minería de datos.
- Permite abrir un conjunto de instancias desde un archivo, URL, Base de datos o generar datos artificiales.
- Extensibilidad permite añadir nuevas funcionalidades.
- Ofrece 4 interfaces:
    - **Explorer**.- Es una interfaz gráfica intuitiva para el usuario, en esta se puede hacer uso directo de los paquetes de Weka, acceso directo a los algoritmos.
    - **Experimenter**.- permite hacer comparaciones cuidadosas de la ejecución de los algoritmos sobre un conjunto de datos.

- **Knowledg Flow**.- es una interfaz que permite arrastrar y soltar los componentes necesarios para leer archivo de datos, aplicar filtros, algoritmos y visualizar la información de la forma más conveniente para el usuario.
- **Simple CLI.-** permite trabajar desde la consola e invocar mediante Java a los paquetes de Weka.

- Brinda al cliente una opción de trabajo que se acomode a sus necesidades.

3.2.1.6. Debilidades.

- No es posible realizar modelado de secuencias.
- No es posible realizar minería de datos multi-relacional.

Conclusión.

Este capítulo presento las herramientas que según el cuadrante mágico de Gartner son las aprobadas por científicos para el análisis de datos y son las que tienen mayor influencia en el mercado. Es importante tener una idea clara de lo que se necesita y se busca en una herramienta de software para minería de datos, ya que, existe variedad de software dedicados al tema de análisis de datos y se debe seleccionar la herramienta adecuada para cada problema, realizando pruebas y analizando los costos y beneficios. También es fundamental tener una lista de 5 o menos herramientas, pues de esta manera podremos involucrarnos lo suficiente y dedicarle la respectiva evaluación a cada una de ellas.

3.2.Software para aplicar minería de datos.

Existe una gran variedad de herramientas de minería de datos disponibles según la necesidad de cada escenario y su adaptación al problema.

Según el Magic Quadrant for Data Science Platforms existen herramientas de minería de datos que fueron evaluadas para proporcionar mayor facilidad en la toma de decisiones para la empresa.

Según Gartner en su artículo "Magic Quadrant for Data Science Platforms" publicado en febrero 14 de 2017 (Gartner, Inc., 2017), este cuadrante mágico evalúa a los proveedores de plataformas de ciencia de los datos, sometiéndolos a criterios de inclusión haciendo de estos los más representativos y relevantes en cuanto a términos de ejecución y visión.

La metodología para la selección de proveedores de plataformas de ciencia de los datos utilizada por los analistas (Alexander Linden, Peter Krensky, Jim Hare, Carlie J. Idoine,

Svetlana Sicular, Shubhangi Vashisth) incluye un proceso de calificación de pila que examina los productos de los proveedores en tres escenarios como son: refinamiento de la producción, exploración comercial y prototipos avanzados, además evalúan capacidades críticas para puntuar las plataformas en los escenarios mencionados, algunas de estas capacidades son:

- Accesos a los datos
- Preparación de los datos
- Exploración y visualización de los datos
- Automatización
- Interfaz de usuario
- Aprendizaje automático
- Flexibilidad, extensibilidad
- Rendimiento y escalabilidad
- Gestión de modelos
- Coherencia.

Según el Cuadrante Mágico las herramientas de minería de datos que cumplen con los requisitos de esta plataforma son: Knime, RapidMiner, SAS e IBM como se muestra en la figura a continuación.

Imagen 3. Cuadrante Mágico para Plataformas de Ciencia de Datos

Imagen tomada de fuente https://www.gartner.com/doc/reprints?id=1-3TK9NW2&ct=170215&st=sb , 2017

Como podemos observar las herramientas de minería de datos más relevantes se encuentran en el cuadrante de Líderes.

Según Gartner los Líderes están fuertemente presentes en el mercado y tiene una participación significativa, además de tener agilidad para responder ante las cambiantes condiciones del mercado.

A continuación se presentan dichas herramientas de Minería de datos de proveedores de software que para el propósito de pruebas buscamos ya sea software libre (open source), freeware o versiones de prueba de herramientas de datamining.

3.2.2. KNIME

KNIME ® "Konstanz Information Miner" es una plataforma de análisis desarrollada originalmente en el departamento de Catedra en Bioinformática y Minería de la Información en la Universidad de Constanza en Alemania con supervisión del profesor Michael Berthold, KNIME.com AG está ubicada actualmente en Zúrich, Suiza.

KNIME Analytics es una plataforma de código abierto, escrita en Java y está basada en Eclipse, posee una amplia gama de algoritmos, con interfaz de usuario amigable, escalable con total funcionalidad e intuitivo de aprender.

Esta plataforma fue desarrollada para abarcar grandes volúmenes de datos, transformar, analizar y crear modelos de exploración visual para descubrir información potencial que se encuentra oculta en los datos y ayuda a predecir eventos futuros en diversas áreas de trabajo como: Finanzas, Medios Sociales, Ventas, entre otras; algunos ejemplo son Análisis de los clientes, Pronóstico de Abandono, Detección y prevención de fraude, Detección de Fallas.

KNIME es una plataforma que no tiene limitaciones en el tamaño de procesamiento de la máquina, si esta posee disco duro y memoria suficiente, puede procesar información de un proyecto con millones de filas de datos y está disponible para todos. (KNIME AG, 2017)

3.2.1.1. Licencia de KNIME

- KNIME Analytics Platform es publicada bajo licencia de Open Source GPLv3
- Las extensiones de colaboración de KNIME que tienen licencia comercial son las siguientes:

	KNIME Analytics Platform	**KNIME TeamSpace**	**KNIME Server Lite**	**KNIME WebPortal**	**KNIME Server**
# Filas de datos	Sin Limites	Sin Limites	Sin Limites	Sin Limites	Sin Limites
Caducidad de la Licencia	Nunca	Después de 12 meses	Después de 12 meses	Después de 12 meses	Después de 12 meses

Atención al Cliente	En línea / Foro	En línea / Foro	En línea / Foro	En línea / Foro	Contactar KNIME
Cargo annual de Licencia	Gratuito	2.363$ por usuario	2.363$ por usuario	2.363$ por usuario	2.363$ por usuario

3.2.1.2. Características:

- Conectores para todos los principales formatos de archivo y bases de datos.
- Soporte para una gran cantidad de tipos de datos: XML, JSON, imágenes, documentos y muchos más.
- Mezcla y transformación de datos nativos y en la base de datos.
- Funciones matemáticas y estadísticas.
- Algoritmos avanzados de aprendizaje predictivo y máquina.
- Control de flujo de trabajo.
- Combinación de herramientas para Python, R, SQL, Java, Weka y muchos más.
- Vistas e informes interactivos de datos.

(KNIME AG, 2017)

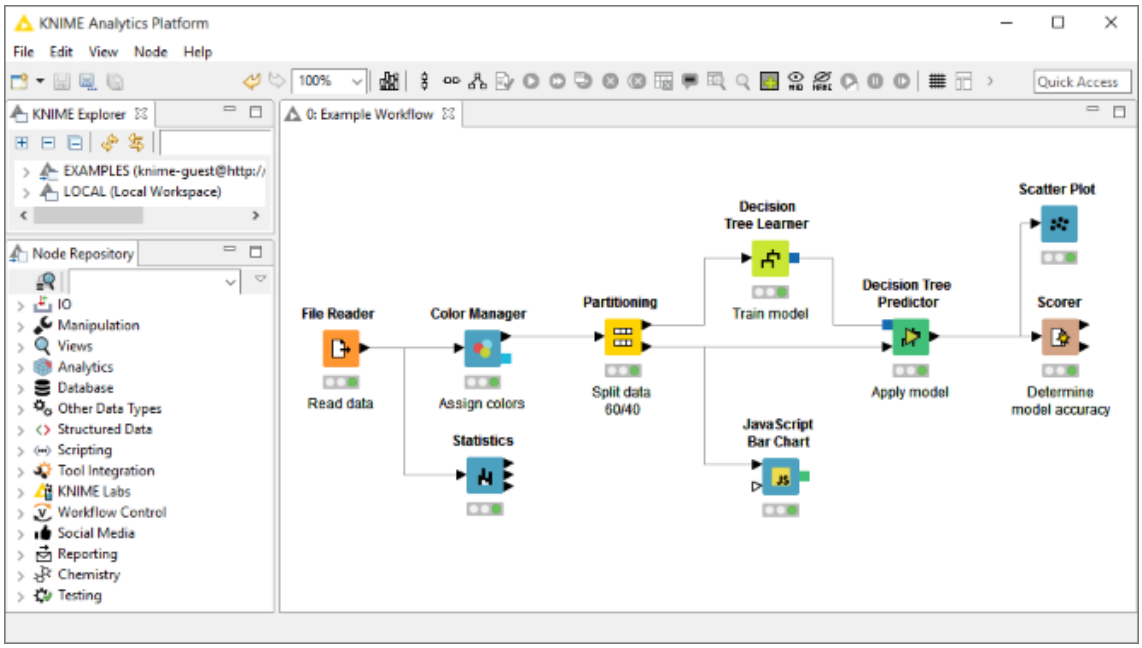


Imagen 4. Arboles de decisión modelado en la herramienta KNime.

Fuente (https://www.knime.org/knime-analytics-platform ,2017)

3.2.1.3. Sistemas Operativos

KNIME es multiplataforma y tiene compatibilidad con sistemas operativos como:

- Windows.
- Linux.
- Mac OS.

3.2.1.4. Áreas de Implementación

KNIME se usa para dar solución a problemas de una variedad de áreas, algunas de ellas son:

Área (Casos de Uso)	**Solución**
Customer Inteligence (Inteligencia del cliente)	• Análisis de abandono. • Predicción de abandono.
Social Media (Medios de comunicación Social)	• Análisis de sentimientos de medios Sociales. • Análisis de seguidores.
Finanzas	• Puntuación crediticia. • Clasificación crediticia. • Riesgo del cliente.
Fabricación	• Predicción del uso de la energía.
Cuidado de la salud	• Enumeración de la biblioteca química. • Detección de valores atípicos en reclamos médicos.
Ventas al por menor	• Recomendación sobre la música en las redes sociales. • Análisis de cesta de mercado y motores de recomendación.
Gobierno	• Informes de tráfico de red. • Predicción de incendios forestales

3.2.1.5. Fortalezas

- Bajo costo total de propiedad (TCO) y facilidad de uso.
- Acceso solido a datos, permite preparar los datos, transformar, verificar la calidad de los datos, etc.
- Calificación de ventas positivas y constante mejoras de sus productos.
- Tiene una comunidad activa y fuertes asociaciones con proveedores de software independiente e integradores de sistemas. (Gartner, Inc., 2017)

3.2.1.6. Debilidades

Según (Gartner, Inc., 2017) los encuestados mencionan las siguientes dificultades al usar KNIME.

- Puede ser difícil de escalar y compartir en las implementaciones empresariales debido a la necesidad de implementar componentes adicionales a KNIME Analytics Platform.
- Capacidad limitada en exploración y visualización de datos ya que su interfaz de usuario aunque esta mejorada aún se sigue realizando cambios para optimizarla..
- El servicio de soporte para América Latina, Asia, Australia entre otros es entregado a través de una red certificada de socios locales o a distancia.
- Problemas con la gestión del modelo, en la administración y documentación de grandes flujos de trabajo.

3.2.3. RapidMiner

RapidMiner Studio fue desarrollado inicialmente por la Universidad de Dortmud, con sede en Massachusetts, EE.UU. RapidMiner es una herramienta que posee un potente entorno de programación visual, aplica técnicas de minería de datos crear flujos de trabajo analíticos predictivos, extrae estadísticas e información clave, construye y entrega mejores modelos, según el portal internacional de Minería de Datos KDnuggets (Piatetsky-Shapiro & Mayo, 2017), en una encuesta realizada en 2015 RapidMiner obtuvo el segundo lugar como herramienta más utilizada . Esta herramienta ofrece una variedad de algoritmos de preparación de datos y de aprendizaje automático que permite maximizar la productividad

de la ciencia de los datos, agilizando la transformación, el desarrollo y la validación de los datos, además permite incluir algoritmos pertenecientes a la herramienta WEKA (Waikato Environment for Knowledge Analysis) para apoyar a los proyectos en los que se requiera procesar grandes volúmenes de datos. (RapidMiner, Inc, 2017).

3.2.3.1. Licencia de RapidMiner

	EDUCATIVO	**Gratis**	**PEQUEÑA**	**MEDIO**	**GRANDE**
# Filas de datos	Ilimitado	10,000	100,000	1,000,000	
Caducidad de la Licencia	Después de 12 meses	Nunca	Después de 12 meses	Después de 12 meses	Después de 12 meses
Atención al Cliente	Comunidad	Comunidad	Empresa	Empresa	Empresa
Cargo anual de Licencia	Gratis	Gratis	$ 2,500 anual	$ 5,000 Anual	$ 10,000 anuales

3.2.3.2. Características:

- Está desarrollado en Java.
- Es multiplataforma.
- Representación interna de los procesos de análisis de datos en ficheros XML.
- Accede a más de 40 tipos de archivos, incluidos SAS, ARFF, Stata y vía URL
- Incluye más de 60 tipos de archivos y formatos para datos estructurados y no estructurados.
- Puede usarse de diversas maneras:
    - A través de un GUI.
    - En línea de comandos.
    - En batch (lotes)
    - Desde otros programas, a través de llamadas a sus bibliotecas.
- Permite conexiones de bases de datos JDBC, incluyendo Oracle, IBM DB2, Microsoft SQL Server, MySQL, Postgres, etc
- Incluye gráficos y herramientas de visualización de datos.

- Los diagramas creados pueden ser con facilidad copiados y pegados en otras aplicaciones o a su vez exportar en formatos PNG, SVG, JPEG, EPS o PDF.
- Dispone de un módulo de integración

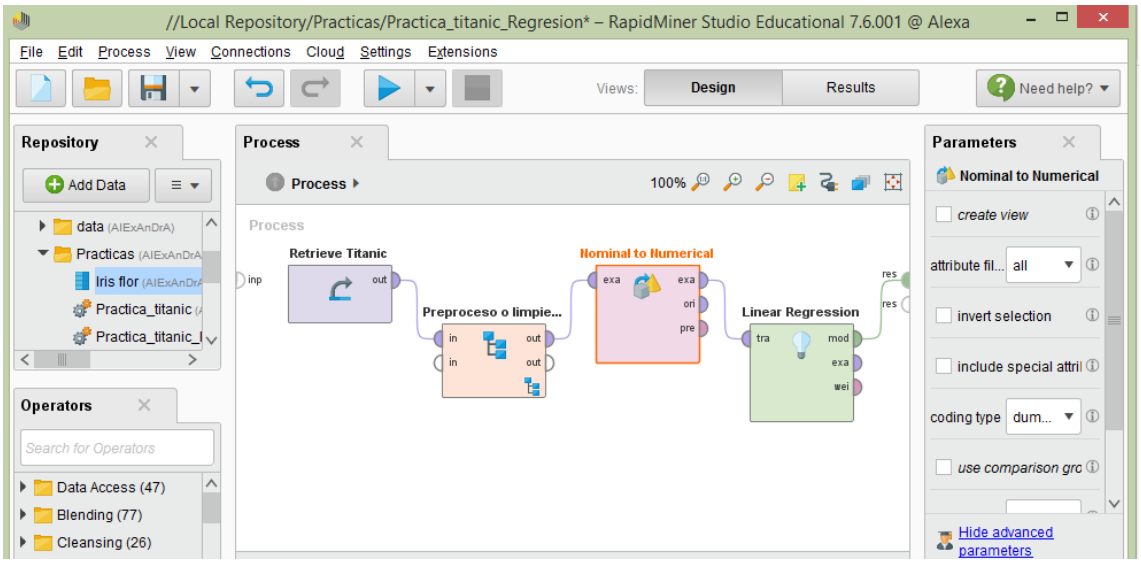


Imagen 5. Arboles de Decisión con RapidMiner

(Fuente: RapidMiner.com, 2017)

3.2.3.3. Sistemas Operativos

RapidMiner es multiplataforma y tiene compatibilidad con sistemas operativos como:

- Windows.
- Linux.
- Mac OS.

3.2.3.4. Áreas de Implementación

RapidMiner se usa para dar solución a problemas de una variedad de áreas, algunas de ellas son:

Área (Casos de Uso)	**Solución**
Customer Inteligence (Inteligencia del cliente) Cuidado de la salud	• Pronóstico de ventas en ciencias de la vida. • Minería de textos en salud los deseos del consumidor sobre el producto.
Social Media (Medios de comunicación Social)	• Segmentación de clientes con RapidMiner y QlikView en Telecomunicaciones.

Servicio Financiero	• Análisis de sentimiento en PayPal con RapidMiner.
Fabricación	• RapidMiner Data Mining para análisis de riesgo en cadenas de suministro.

3.2.3.5. Fortalezas

- Plataforma amplia ya que se utiliza para una variedad de áreas y casos de uso, posee una amplia gama de algoritmos.
- Facilidad de uso y aprendizaje, posee una interfaz amigable.
- Velocidad en la creación de modelos

(Gartner, Inc., 2017)

3.2.3.6. Debilidades

- Capacidad limitada en el uso de los datos en la edición gratuita.
- Costo total de propiedad TCO impredecible.
- Información escasa acerca de cómo usar operadores específicos.

(Gartner, Inc., 2017)

3.2.4. IBM SPSS Modeler

IBM SPSS es una empresa que ofrece soluciones de análisis y según el cuadrante mágico de Garnet es nuevamente líder en análisis predictivo, tiene su sede en Armonk, Nueva York,EE.UU.

IBM SPSS Modeler es software de minería de datos y aprendizaje automático, cuyo nombre originalmente era Clementine y este nombre fue cambiando en el 2009 luego de la adquisición de IBM[2] de SPSS[3] obtuvo su nombre actual. Dicho software permite construir modelos de predicción entre otras tareas de análisis, además permite utilizar la gama de algoritmos ya sean de minería de datos o algoritmos

[2] **IBM** (International Business Machines Corporation) se traduce como "Maquina de Negocios Internacionales "

[3] **SPSS** (Statistical Package for the Social Sciences) se traduce como "Paquete estadístico para las ciencias sociales"

estadísticos sin la complejidad de programar. Se considera que IBM SPSS Modeler es uno de los mejores proveedores de software que aplica técnicas de minería de datos, es escalable y flexible adaptándose a las necesidades de cada negocio. (IBM, 2017)

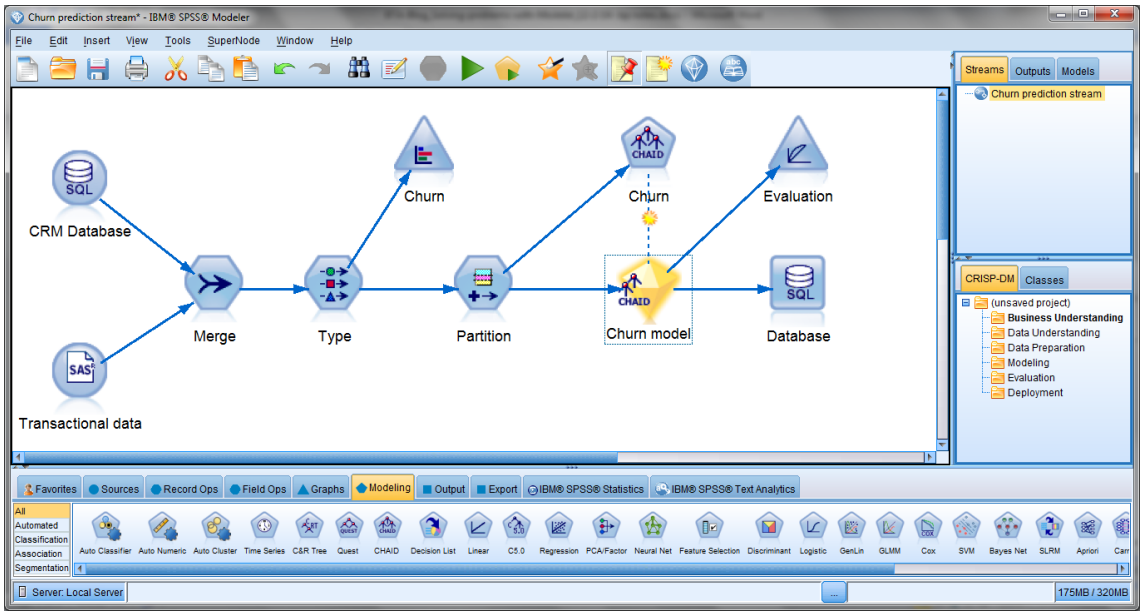


Imagen 6. Automatización de modelos predictivos SPSS Modeler

(Fuente: TPS Analitics for succes ,2017)

SPSS Modeler proporciona el análisis predictivo para ayudar a descubrir las tramas de datos, aumento de la exactitud de predicción y mejorar la toma de decisiones. (IBM, 2017)

3.2.4.1. Licencia de IBM

	Gratis	PROFESIONAL	PREMIUM	GOLD
# Filas de datos				
Caducidad de la Licencia	Nunca	Después de 12 meses	Después de 12 meses	Después de 12 meses
Atención al Cliente		Empresa	Empresa	Empresa
Cargo anual de Licencia	Gratis	Sin especificar	Sin especificar	Sin especificar

3.2.4.2. Características:

- Soporta varias fuentes de datos como es leer datos de archivos planos, hojas de cálculo y principales bases de datos relacionales.

- Posee un interfaz gráfica fácil de usar e intuitiva para visualizar proceso de minería de datos.
- Otorga modelos predictivos más precisos
- Modelo automatizado.
- Variedad de métodos algorítmicos.
- Excelente análisis de texto.

3.2.4.3. Sistemas Operativos

IBM es multiplataforma y tiene compatibilidad con sistemas operativos como:

- Windows.
- Mac OS.
- Linux.
- Unix.

3.2.4.4. Áreas de Implementación

IBM se usa para dar solución a problemas de una variedad de áreas, algunas de ellas son:

Área (Casos de Uso)	**Solución**
Mercado y Producto	• Pronostico de ventas. • Percepción de la marca • Análisis de los clientes
Tecnología	• Intrusión • Detección y prevención de fraude. • Fallas
Costumer Services (Servicio al cliente)	• Satisfacción del cliente. • Quejas y reclamos
Finanzas	• Límites de crédito. • Lavado de dinero. • Movimiento de efectivo.
Fabricación	• Análisis de la calidad de fabricación

3.2.4.5. Fortalezas

- Se centra en el soporte de tecnologías de código abierto.
- Acepta una amplia gama de tipos de datos, variedad de base de datos.

- Amplitud de modelos, transparencia en los flujos de trabajo.

3.2.4.6. Debilidades

- Problemas con los costos totales de propiedad TCO, confusión sobre las ofertas de productos.
- Problemas con el soporte al cliente, dificultad para encontrar los enlaces correctos.
- Información inaccesible acerca de los tipos de licencias y sus costos.

3.2.5. SAS Enterprise Miner.

SAS Enterprice Miner desarrollado por SAS Institute apareció por primera vez en el año de 1917, es una herramienta avanzada de minería de datos que brinda soluciones para crear modelos de predicción en grandes volúmenes de datos, además tiene varias características y funcionalidades para que analistas puedan modelar sus datos. Posee una interfaz gráfica sencilla que permite al usuario crear flujos de proceso mediante la selección de nodos específicos y estos son arrastrados al área de trabajo. Tiene variedad de algoritmos como: redes neuronales, regresión lineal, arboles de decisión, entre otros. (Predictive Analytics Today, 2017)

Imagen 7. Visualización de datos en SAS Enterprise Miner.

(Fuente: Sas.com ,2017)

3.2.5.1. Licencia de SAS

SAS Enterprise se encuentra bajo licencia de software propietario.

3.2.5.2.Características.

- Interfaz gráfica fácil de usar.
- Modelado predictivo y descriptivo
- Capacidades de alto rendimiento.
- Una manera rápida, fácil y autosuficiente para que los usuarios de negocios generen modelos.
- Procesamiento escalable
- Preparación, resumen y exploración de datos sofisticados.
  (SAS Institut Inc., 2017)

3.2.5.3. Sistemas Operativos

SAS es multiplataforma y tiene compatibilidad con sistemas operativos como:

- Windows.
- Mac OS.
- Linux.

(SAS Institut Inc., 2017)

Capítulo IV. Metodología del trabajo y experimentación

Introducción

En este capítulo se detallará la información para la experimentación con los datos recogidos por la red que monitorea la calidad de aire en la ciudad de Cuenca, así como también se explica de donde se extrajeron los datos, cómo se extrajeron los datos, cómo se encuentra estructurada la base de datos y que información contiene esta base.

Se realizan pruebas de: capacidad de procesamiento de datos, tiempo de ejecución de algoritmos, aplicación de fórmulas matemáticas para conversión de valores, entre otros, mediante el uso de las herramientas de minería de datos mencionadas en el capítulo anterior para procesar los datos, aplicara minería de datos y analizar los resultados que resultan de dichas herramientas y proceder a evaluar que herramienta es la más adecuada.

4.1. Recopilación de la Información.

Este trabajo se desarrolló sobre los datos de contaminación atmosférica de la ciudad de Cuenca. La ciudad de Cuenca está localizada en la República del Ecuador, y pertenece al cantón Cuenca provincia del Azuay. "La capital cuencana se caracteriza por estar localizada en un valle de origen glacial con una cota (altura) promedio de 2550 msnm, presenta una morfología irregular y compleja característica de los valles interandinos, tiene una temperatura promedio de 16.1◦C y una presión barométrica local promedio de 751.89 mmHg." (Sellers & Espinoza, 2017; Sellers C. A., 2013)

Como se ha mencionado anteriormente, en este trabajo se emplea la metodología de CRISP para aplicar minería de datos.

4.1.1. Comprensión del negocio

En la ciudad de Cuenca existe un sistema de monitoreo continuo de la calidad del aire, este sistema mediante sensores recolecta información sobre los niveles de los gases en el aire, tales como Ozono(O_3), Dióxido de Azufre (SO_2), Dióxido de Nitrógeno (NO_2) y Monóxido de Carbono (CO) y Material Particulado($PM_{2.5}$).

La Empresa Municipal de Movilidad Tránsito y Trasporte de la ciudad de Cuenca (EMOV-EP) cuenta con una estación de monitoreo continuo en tiempo real con sensores que miden los contaminantes atmosféricos a intervalos de 1 segundo y tiene una cobertura de 4km de radio, dicha estación se encuentra ubicada en la zona alta del edificio de la alcaldía. **(Sellers, 2017)**

En la siguiente tabla se exponen los elementos monitoreados por la estación.

Contaminante	Símbolo	Datalogger	Unidad
Ozono	O_3	ppb	Partes por Billón
Dióxido de Azufre	SO_2	ppb	Partes por Billón
Dióxido de Nitrógeno	NO_2	ppb	Partes por Billón
Monóxido de Carbono	CO	ppm	Partes por Millón
Material Particulado2,5um	$PM_{2.5}$	ug/m3	Microgramos por metro cubico

Tabla 1 Contaminantes y unidades de medida.

Fuente (Articulo "Publicación de contaminantes atmosféricos de la ciudad de Cuenca" Sellers, C. & Espinoza, C. 2017)

Debido a la existencia de gran cantidad de información de estos contaminantes se busca aplicar técnicas de minería de datos con el fin de que ayuden a determinar relaciones y/o patrones de comportamiento de estas variables en la contaminación atmosférica.

4.1.2. Comprensión de los datos.

Los datos son capturados por los sensores de la estación de monitoreo y son transmitidos a un servidor local cada minuto, se almacenan y se muestran en archivos con extensión XLS, CSV, aquí se registra la hora y fecha de la observación, el identificador del procedimiento, el objeto de interés, el fenómeno registrado, un valor registrado del contaminante y un número identificador de la observación entre otros.

En las tablas 2 y 3 se muestra la estructura de los campos de la base de datos.

time_stamp	procedure_id	feature_of_interest_id	phenomenon_id
16/09/2017 0:01	urn:ogc:object:feature:Sensor:sensor-2	foi_2001	urn:ogc:def:phenomenon:OGC:1.0.30:O3
16/09/2017 0:01	urn:ogc:object:feature:Sensor:sensor-3	foi_3001	urn:ogc:def:phenomenon:OGC:1.0.30:CO
16/09/2017 0:01	urn:ogc:object:feature:Sensor:sensor-4	foi_4001	urn:ogc:def:phenomenon:OGC:1.0.30:NO2
16/09/2017 0:01	urn:ogc:object:feature:Sensor:sensor-5	foi_5001	urn:ogc:def:phenomenon:OGC:1.0.30:SO2
16/09/2017 0:01	urn:ogc:object:feature:Sensor:sensor-6	foi_6001	urn:ogc:def:phenomenon:OGC:1.0.30:PM2_5
16/09/2017 0:02	urn:ogc:object:feature:Sensor:sensor-2	foi_2001	urn:ogc:def:phenomenon:OGC:1.0.30:O3
16/09/2017 0:02	urn:ogc:object:feature:Sensor:sensor-3	foi_3001	urn:ogc:def:phenomenon:OGC:1.0.30:CO
16/09/2017 0:02	urn:ogc:object:feature:Sensor:sensor-4	foi_4001	urn:ogc:def:phenomenon:OGC:1.0.30:NO2
16/09/2017 0:02	urn:ogc:object:feature:Sensor:sensor-5	foi_5001	urn:ogc:def:phenomenon:OGC:1.0.30:SO2
16/09/2017 0:02	urn:ogc:object:feature:Sensor:sensor-6	foi_6001	urn:ogc:def:phenomenon:OGC:1.0.30:PM2_5
16/09/2017 0:03	urn:ogc:object:feature:Sensor:sensor-2	foi_2001	urn:ogc:def:phenomenon:OGC:1.0.30:O3
16/09/2017 0:03	urn:ogc:object:feature:Sensor:sensor-3	foi_3001	urn:ogc:def:phenomenon:OGC:1.0.30:CO
16/09/2017 0:03	urn:ogc:object:feature:Sensor:sensor-4	foi_4001	urn:ogc:def:phenomenon:OGC:1.0.30:NO2
16/09/2017 0:03	urn:ogc:object:feature:Sensor:sensor-5	foi_5001	urn:ogc:def:phenomenon:OGC:1.0.30:SO2
16/09/2017 0:03	urn:ogc:object:feature:Sensor:sensor-6	foi_6001	urn:ogc:def:phenomenon:OGC:1.0.30:PM2_5
16/09/2017 0:04	urn:ogc:object:feature:Sensor:sensor-2	foi_2001	urn:ogc:def:phenomenon:OGC:1.0.30:O3
16/09/2017 0:04	urn:ogc:object:feature:Sensor:sensor-3	foi_3001	urn:ogc:def:phenomenon:OGC:1.0.30:CO
16/09/2017 0:04	urn:ogc:object:feature:Sensor:sensor-4	foi_4001	urn:ogc:def:phenomenon:OGC:1.0.30:NO2
16/09/2017 0:04	urn:ogc:object:feature:Sensor:sensor-5	foi_5001	urn:ogc:def:phenomenon:OGC:1.0.30:SO2
16/09/2017 0:04	urn:ogc:object:feature:Sensor:sensor-6	foi_6001	urn:ogc:def:phenomenon:OGC:1.0.30:PM2_5
16/09/2017 0:05	urn:ogc:object:feature:Sensor:sensor-2	foi_2001	urn:ogc:def:phenomenon:OGC:1.0.30:O3
16/09/2017 0:05	urn:ogc:object:feature:Sensor:sensor-3	foi_3001	urn:ogc:def:phenomenon:OGC:1.0.30:CO

Tabla 2. Estructura de los campos de la base de datos

(Fuente: IERSE, 2017)

numeric_value	observation_i	valor	valor_sin_factor	id_mysql
204.153.055.898	10293011	204.153.055.898	204.153.055.898	15565133
0.789981	10293009	0.789981	0.789981	15565129
234.809.131.769	10293010	234.809.131.769	234.809.131.769	15565131
301.132.514	10293013	301.132.514	301.132.514	15565135
7.2	10293012	7.2	7.2	15565134
230.873.838.344	10293016	230.873.838.344	230.873.838.344	15565140
0.789981	10293014	0.789981	0.789981	15565136
234.451.850.316	10293015	234.451.850.316	234.451.850.316	15565138
30.034.695.092	10293018	30.034.695.092	30.034.695.092	15565142
9.7	10293017	9.7	9.7	15565141
247.549.803.894	10293021	247.549.803.894	247.549.803.894	15565147
0.789981	10293019	0.789981	0.789981	15565143
228.001.979.875	10293020	228.001.979.875	228.001.979.875	15565145
29.929.953.348	10293023	29.929.953.348	29.929.953.348	15565149
11.7	10293022	11.7	11.7	15565148
236.269.003.669	10293026	236.269.003.669	236.269.003.669	15565154
0.80143	10293024	0.80143	0.80143	15565150
228.001.979.875	10293025	228.001.979.875	228.001.979.875	15565152
27.416.151.492	10293028	27.416.151.492	27.416.151.492	15565156
11.7	10293027	11.7	11.7	15565155
221.201.778.325	10293031	221.201.778.325	221.201.778.325	15565161
0.80143	10293029	0.80143	0.80143	15565157

Tabla 3. Estructura de los campos de la base de datos

(Fuente: IERSE, 2017)

Los datos obtenidos para el propósito de este trabajo se encuentran en formato .CSV y contiene un total de 215.436 registros de cada contaminante tomados desde la fecha 16/09/2017, hora: 0:01:00 am hasta 16/10/2017, hora: 9:23:00 am, dicha base de datos fue proporcionada por el centro de investigación IERSE.

5. Preparación de los datos

En esta etapa se preparan los datos obtenidos, seleccionando los que son de interés y descartando los que no, es importante entender previamente como se encuentran los datos y aplicar los procedimientos necesarios para filtrar, manejar valores ausentes o nulos y eliminarlos si es el caso, se debe normalizar los datos y colocar los datos en el formato adecuado para luego poder aplicar minería de datos.

El archivo que contiene los datos cuenta con varios parámetros por lo que se procedió a importar los datos a una base de datos para seleccionar los necesarios y ponerlo en un formato adecuado para ser utilizados en las diferentes herramientas de minería de datos.

Empezamos seleccionando un método para poder importar los datos que se encuentran en un archivo CSV.

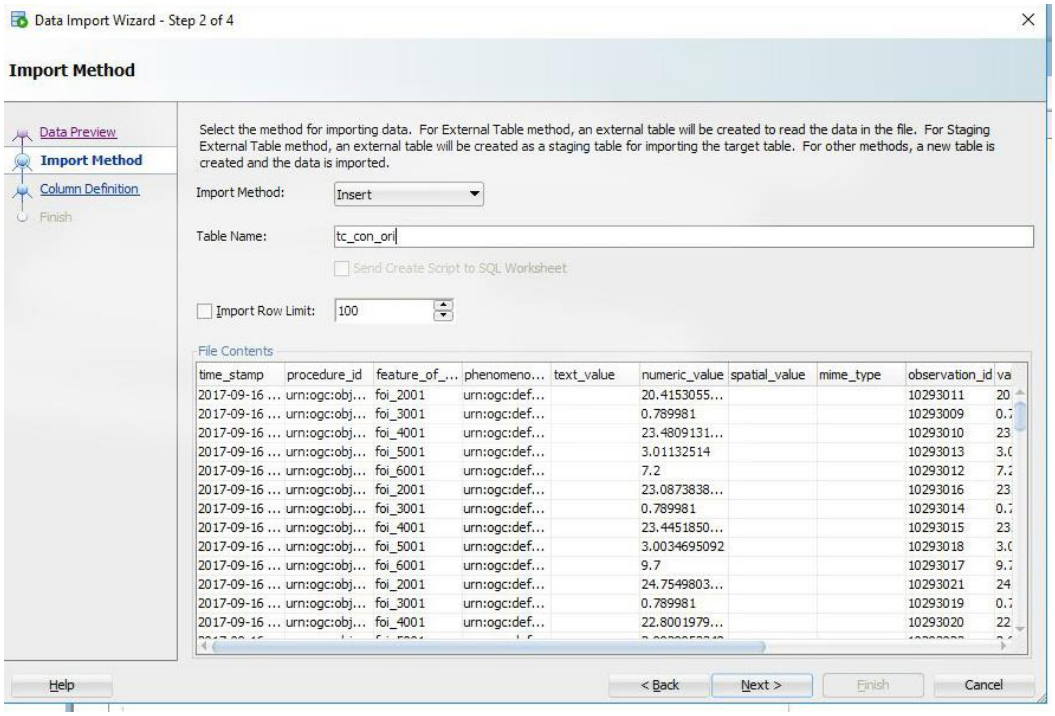


Imagen 8. Selección del método para importar los datos.
(Fuente: LIDI, 2017)

En este proceso se seleccione las columnas para importar del conjunto de datos y podemos organizarlas en el orden que necesitemos.

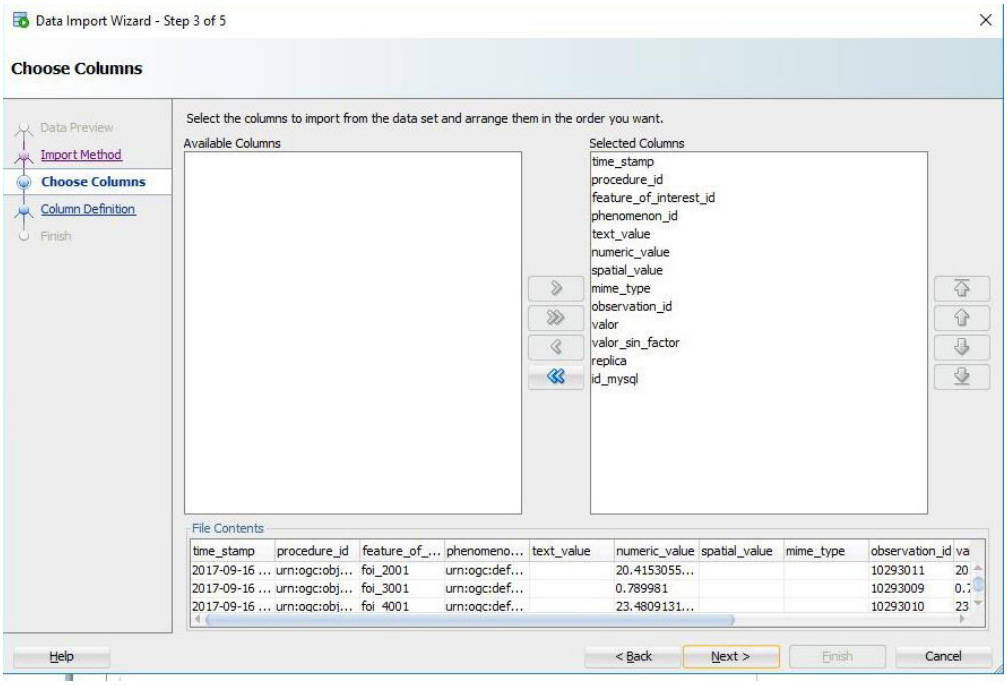


Imagen 9. Selección de las columnas para la nueva tabla
(Fuente: LIDI, 2017)

Para cada dato de la columna a la izquierda se define los valores de la base de datos que va a ser creada con la importación de los datos.

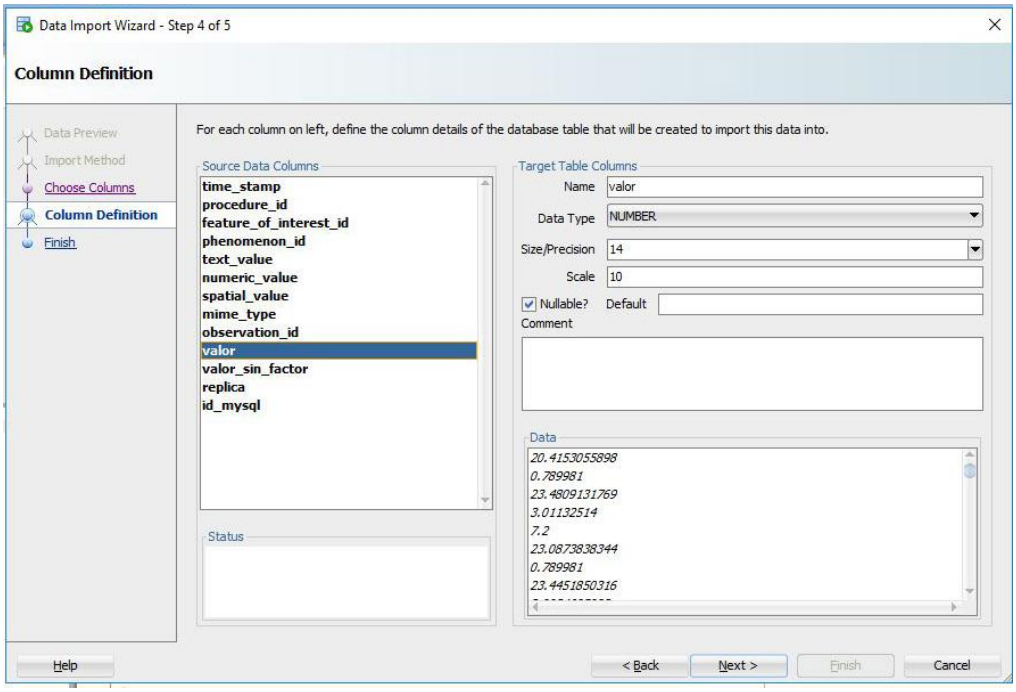


Imagen 10. Definición de los datos para cada columna

(Fuente: LIDI, 2017)

Finalmente tendremos los datos importados en una base de datos y listos para realizar los procedimientos necesarios para filtrar los valores de las variables de contaminación atmosférica (O_3, SO_2, NO_2, CO, $PM_{2.5}$) que serán objeto de nuestro estudio.

Para la construcción de los datos se procede a realizar un ***pivot*** (trasponer las filas a columnas) de esta forma se prepara la tabla para su proceso esto solo sirve para ordenar y procesar los datos de mejor manera.

--Crear tabla cruzada (piv ...

SQL | Output | Statistics

```
create table tc_envcon as(
select *
  from (select regexp_substr(s.phenomenon_id, '[^:]+$') ec_phenomenon,
               s.valor ec_value,
               to_date(substr(s.time_stamp, 0, 19), 'yyyy-mm-dd HH24:MI:SS') ec_time_stamp
          from tc_con_ori s)
pivot(sum(ec_value)
   for ec_phenomenon in('O3' as "O3",
                        'CO' as "CO",
                        'NO2' as "NO2",
                        'SO2' as "SO2",
                        'PM2_5' as "PM2_5")))
 order by ec_time_stamp
;

alter table tc_envcon add constraint tc_envcon_pk primary key(ec_time_stamp);
commit;
```

Imagen 11. Uso del operador Pivot y generamos la estructura de datos

(Fuente: LIDI, 2017)

Utilizamos el operador ***pivot*** para pasar de una visualización compleja de datos como se muestra en la tabla 4.

time_stamp	phenomenon_id	value
16/09/2017 0:01	urn:ogc:def:phenomenon:OGC:1.0.30:O3	20,41530559
16/09/2017 0:01	urn:ogc:def:phenomenon:OGC:1.0.30:CO	0,789981
16/09/2017 0:01	urn:ogc:def:phenomenon:OGC:1.0.30:NO2	23,4809131769
16/09/2017 0:01	urn:ogc:def:phenomenon:OGC:1.0.30:SO2	3,01132514
16/09/2017 0:01	urn:ogc:def:phenomenon:OGC:1.0.30:PM2_5	7,2
16/09/2017 0:02	urn:ogc:def:phenomenon:OGC:1.0.30:O3	23,0873838344
16/09/2017 0:02	urn:ogc:def:phenomenon:OGC:1.0.30:CO	0,789981
16/09/2017 0:02	urn:ogc:def:phenomenon:OGC:1.0.30:NO2	23,4451850316
16/09/2017 0:02	urn:ogc:def:phenomenon:OGC:1.0.30:SO2	30,034695092
16/09/2017 0:02	urn:ogc:def:phenomenon:OGC:1.0.30:PM2_5	9,7
16/09/2017 0:03	urn:ogc:def:phenomenon:OGC:1.0.30:O3	24,7549803894
16/09/2017 0:03	urn:ogc:def:phenomenon:OGC:1.0.30:CO	0,789981
16/09/2017 0:03	urn:ogc:def:phenomenon:OGC:1.0.30:NO2	22,8001979875
16/09/2017 0:03	urn:ogc:def:phenomenon:OGC:1.0.30:SO2	29,929953348
16/09/2017 0:03	urn:ogc:def:phenomenon:OGC:1.0.30:PM2_5	11,7

Tabla 4. Datos originales. (Fuente: IERSE, 2017)

A una visualización de datos ordenada específicamente por contamínate atmosférico como se muestra en la tabla 5.

EC_TIME_STAMP	O3	CO	NO2	SO2	PM2_5
16/09/2017 0:01:00	20.4153055898	0.789981	23.4809131769	3.01132514	7.2
16/09/2017 0:02:00	23.0873838344	0.789981	23.4451850316	3.0034695092	9.7
16/09/2017 0:03:00	24.7549803894	0.789981	22.8001979875	2.9929953348	11.7

Tabla 5. Estructura de datos generada usando pivot.

(Fuente: LIDI, 2017)

Con la reestructuracion del conjunto de datos se realizan los procedimientos de limpieza de datos en los que se preocede a la eliminación de los registros que contengan datos nulos, este paso se realiza ya que si existen datos nulos podrían alterar los resultados. A continuación se muestra las sentencias SQL utilizadas para la eliminación de datos nulos.

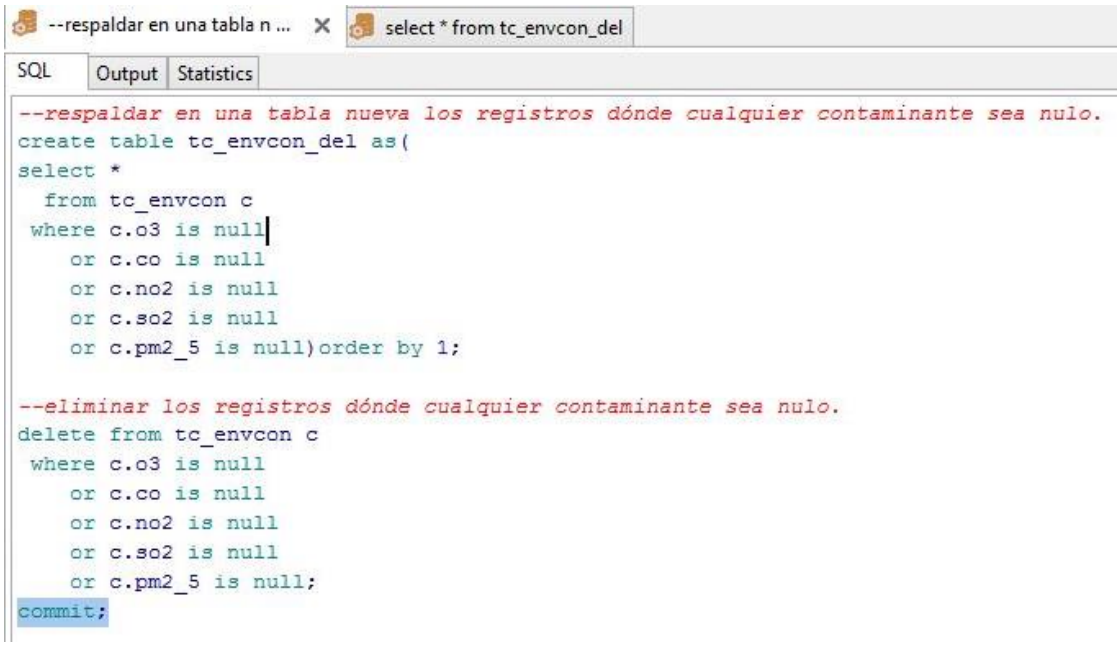


Imagen 12. Respaldo y eliminación de valores nulos encontrados

(Fuente: LIDI, 2017)

Una vez revisado y eliminado los registros que contienen valores nulos, se realiza una última selección en la que se extrae únicamente los datos que tienen mediciones durante todo el día, así evitamos que hayan datos faltantes para aplicar minería de datos.

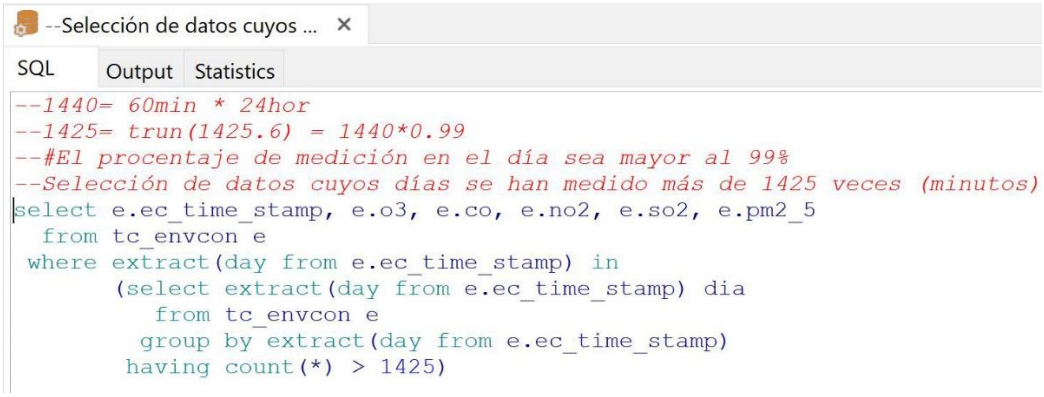


Imagen 13. Selección de los datos cuyas medidas en el día sean mayor al 99%

(Fuente: LIDI, 2017)

Conclusión.

Al término del proceso de preparación de los datos, se eliminó datos nulos y se obtuvo una base de datos adecuadamente estructurada con 25.877 registros, cantidad de datos aptos que cumplen con las exigencias del proyecto y podemos proceder con la siguiente fase para las respectivas pruebas.

6. Modelado

Una vez concluido la fase de preparación de los datos, tenemos la información lista para ser analizada, es aquí donde aplicaremos minería de datos, haciendo uso del algoritmo K-Means (método no supervisado), ya que en los datos de los contaminantes no existe una variable dependiente, y al no especificar un modelo para los datos en base a un conocimiento previo quedan descartados los métodos de entrenamiento o supervisados.

El algoritmo K-means genera los centros de clúster aleatoriamente para un número de clusters predefinido por el usuario. Realiza una agrupación transparente que asigna un vector de datos a exactamente un clúster[4]. El algoritmo finaliza cuando las asignaciones del clúster ya no cambian.

El algoritmo de agrupamiento utiliza la distancia euclidiana en los atributos seleccionados. Este algoritmo no normaliza los datos, por ello si es necesario, se debe considerar utilizar el "Normalizador[5]" como paso de preproceso.

En esta fase procedemos a realizar las pruebas con los datos y aplicamos el algoritmo k-means con la ayuda de las herramientas de minería de datos mencionadas en el capítulo 3, que previamente deben estar instaladas en la PC que vayamos a trabajar.

Para este trabajo de evaluación de las herramientas de minería de datos se utilizó una PC con las siguientes características;

[4] **Clúster**: grupo que tienen un cierto grado de homogeneidad al compartir, en distinta medida, una serie de características semejantes. (Clave, 2017)

[5] **Normalizador**: es un operador que se utiliza para la normalización de datos es decir, es una técnica de preprocesamiento utilizada para reescalar los valores de los atributos para que estén en un rango específico. La normalización de los datos se aplica cuando se tiene atributos de diferentes escalas y unidades de medida. (RapidMiner, 2017)

Sistema.

Procesador: Intel® Core (TM) i7 -3630QM CPU @ 2.40 GHz

Memoria instalada (RAM): 8.00 GB

Sistema Operativo: Windows 8.1 Pro

Tipo de Sistema: Sistema operativo de 64 bits, procesador x64.

Se instalaron 5 herramientas de minería de datos (RapidMiner, KNime,Weka,IBM SPSS Modeler, SAS Entrerprise Miner); unas con versión de prueba 30 días, otras con licencia educacional y también versión libre, con el fin de poder llevar a cabo las pruebas necesarias para seleccionar la herramienta que presente mejores resultados al procesar los datos de contaminación atmosférica.

Con los programas instalados y los datos listos se procedió a la creación de modelos para el flujo de datos, aplicación de algoritmos, visualización de datos, en cada una de las herramientas de minería de datos con la *dataset* que contiene los 25.877 registros.

Modelado de los datos.

Las pruebas se realizaron en base a la siguiente estructura:

Primero se leen los datos que fueron preparados previamente, dependiendo de la herramienta se carga directamente la base de datos como un archivo normal y en otras herramientas se las realiza por medio de operadores como: File Reader, Reader (csv,Excel, ect, se elige según el tipo de archivo), luego se trasforma los datos del campo CO miligramos/m3 a microgramos/m3 (es el único campo que tiene otra unidad de medida) para poder trabajar con los datos en una misma unidad de medida; a continuación se normaliza todos los datos para trabajar en una escala de 0 a 100 y por medio de un filtro se selecciona los datos a los que se va aplicar minería de datos; en este estudio se aplicará Clustering (K-Means), se seleccionó este algoritmo ya que según el estudio denominado "**Impacto de la aplicación de algoritmos de minería de datos en variables de contaminación del aire**" realizado por J. Ortega , K-Means es el mejor algoritmo para aplicar minería de datos en variables de contaminación del aire, ya que presenta tiempos de ejecución bastante cortos frente a otros algoritmos como: X-Means, CobWeb DBSCAM (Ortega Guaman & Orellana Cordero, 2018). Una vez aplicado el algoritmo de clustering y con los resultados de este proceso se procede a seleccionar el clúster que

contiene el mayor número de datos para realizar el análisis de frecuencia y establecer el factor de correlación entre cada uno de los contaminantes estudiados. Se elige el Clúster con mayor número de datos ya que aquí se puede encontrar las tendencias en el comportamiento de los contaminantes. A continuación se presenta las pruebas del modelo en las diferentes herramientas de minería de datos.

Modelado en la herramienta KNIME.

Esta es una herramienta gratuita, cuenta con una amplia gama de algoritmos y permite procesar un número ilimitado de datos. Esta herramienta tiene un área de trabajo sutilmente intuitiva, cuenta con un repositorio de nodos que permite crear un modelo fácilmente empezando desde cargar archivos de variadas extensiones para procesamiento de datos, aplicar fórmulas matemáticas para conversión de datos de ser necesario, algoritmos, visualización de resultados por medio de gráficos, etc. A continuación se muestra la interfaz de trabajo de esta herramienta.

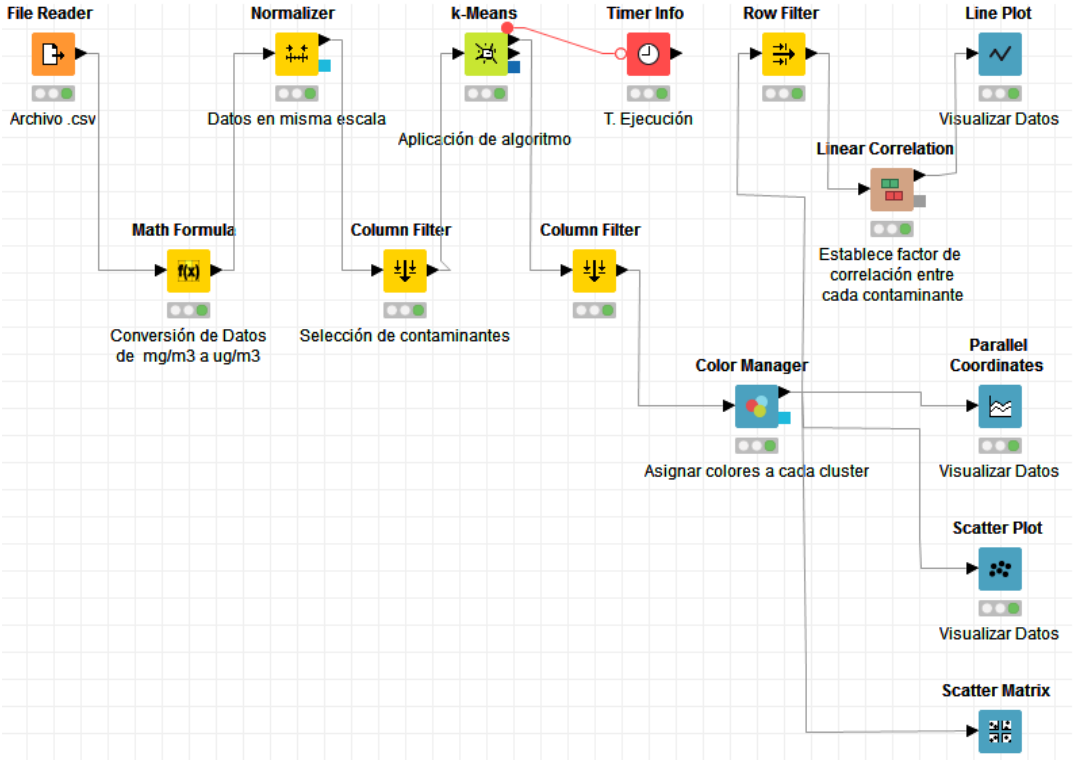


Imagen 14. Modelado con aplicación del algoritmo K-Means en la herramienta KNIME.

Fuente: (Marcia Matute, 2017)

KNIME permite la configuración de los nodos según las necesidades del usuario.

Presentación de Resultados.

Tras la aplicación del algorito K-Means se puede visualizar los siguientes resultados.

Cluster name="cluster_0" size="8804"
Cluster name="cluster_1" size="10906"
Cluster name="cluster_2" size="1957"
Cluster name="cluster_3" size="4210"

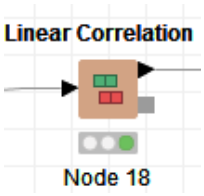
Row ID	O3	NO2	SO2	PM2_5	CO_ug
cluster_0	32.418	7.017	5.247	12.39	11.571
cluster_1	8.041	20.829	9.229	12.873	17.216
cluster_2	62.566	12.563	5.325	43.928	16.309
cluster_3	14.073	42.849	19.606	46.98	32.834

Tabla 6. Centroides, resultado de la aplicación de algoritmo K.Means, Herramienta KNIME.

Fuente: (Marcia Matute, 2017)

Para visualizar los resultados se debe dar click derecho en cada uno de los nodos para ver las tablas o graficos de los datos que contiene cada nodo.

Tabla matriz de correlación.



Row ID	O3	NO2	SO2	PM2_5	CO_ug
O3	1	0.015	0.061	-0.093	-0.075
NO2	0.015	1	0.359	0.15	0.546
SO2	0.061	0.359	1	0.049	0.223
PM2_5	-0.093	0.15	0.049	1	0.131
CO_ug	-0.075	0.546	0.223	0.131	1

Tabla 7. Correlación entre cada contaminante. Herramienta KNIME.

Fuente: (Marcia Matute, 2017)

Color Manager:

Asigna colores a una columna numérica seleccionada, permitiendo identificar los diferentes clústeres por un color en específico.

Row ID	O3	NO2	SO2	PM2_5	CO_ug	Cluster
25784	2.962	30.961	6.332	20.417	41.667	cluster_1
4096	55.42	30.95	3.825	12.778	20.333	cluster_0
21414	8.815	30.95	5.226	12.083	11	cluster_1
22992	2.813	30.95	9.573	17.639	17.333	cluster_1
25760	3.192	30.95	6.175	22.639	40.333	cluster_1
20688	11.38	30.938	15.929	71.667	25	cluster_3
23480	12.657	30.938	27.724	25.833	25	cluster_1
2471	45.909	30.924	5.356	40.833	25.333	cluster_2
5400	51.362	30.924	2.774	33.333	28	cluster_2
16728	14.729	30.924	5.484	11.389	16	cluster_1
20747	27.115	30.924	19.638	49.028	17	cluster_3

Tabla 8. Visualización de los datos pertenecientes a un clúster específico. Herramienta KNIME.

Fuente: (Marcia Matute, 2017)

Coordenadas paralelas:

Traza los datos en Coordenadas Paralelas. Este tipo de graficas se utiliza para trazar multiples datos numéricos, este tipo de visualización permite ver si existe relación entre múltiples variables. (Ingenio Virtual, 2017)

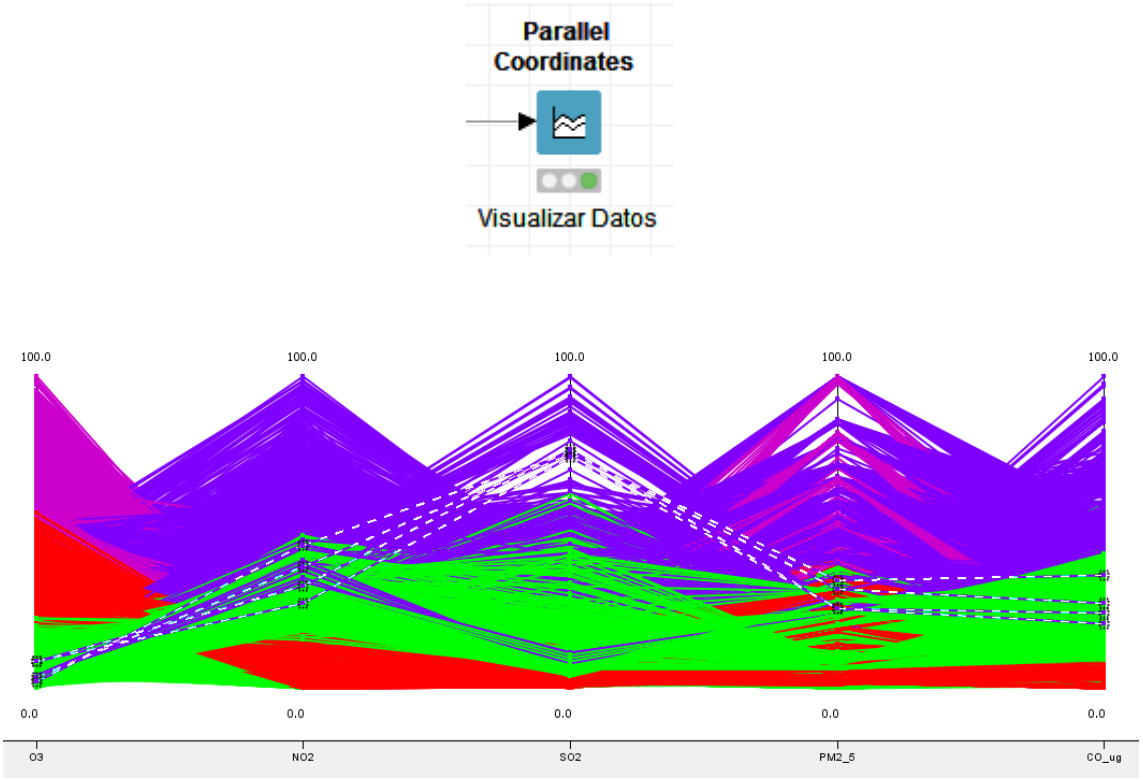


Imagen 15. Gráfico de Coordenadas paralelas. Herramienta KNIME.

Fuente: (Marcia Matute, 2017)

En la imagen 15 se presenta la graficas de los valores de 5 contaminantes atmosféricos, seleccionado un punto en el eje de las y podemos observar el comportamiento del contamínate en relación con los demás. Sin embargo este tipo de grafica es útil cuando se realiza con poco datos, como podemos observar en la imagen 15 resulta difícil de interpretar ya que tiene demasiados datos.

Gráfico de dispersión:

Crea un diagrama de dispersión de dos atributos seleccionados por el usuario. Cada punto de datos se muestra como un punto en su lugar correspondiente, en función de los valores de los atributos seleccionados.

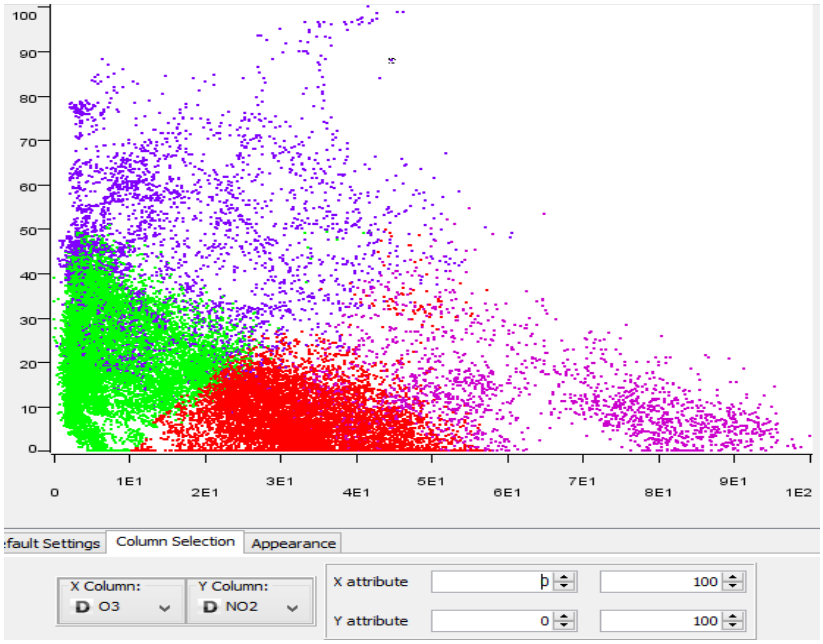
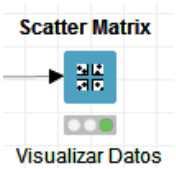


Imagen 16. Gráfico de dispersión.

Fuente: (Marcia Matute, 2017)

Matriz de dispersión:

Muestra graficas de dispersión de todas las variables en una sola matriz, permite la comparación de varios conjuntos de datos para buscar patrones de comportamiento y posibles relaciones entre las variables. (ArcGIS, 2016)



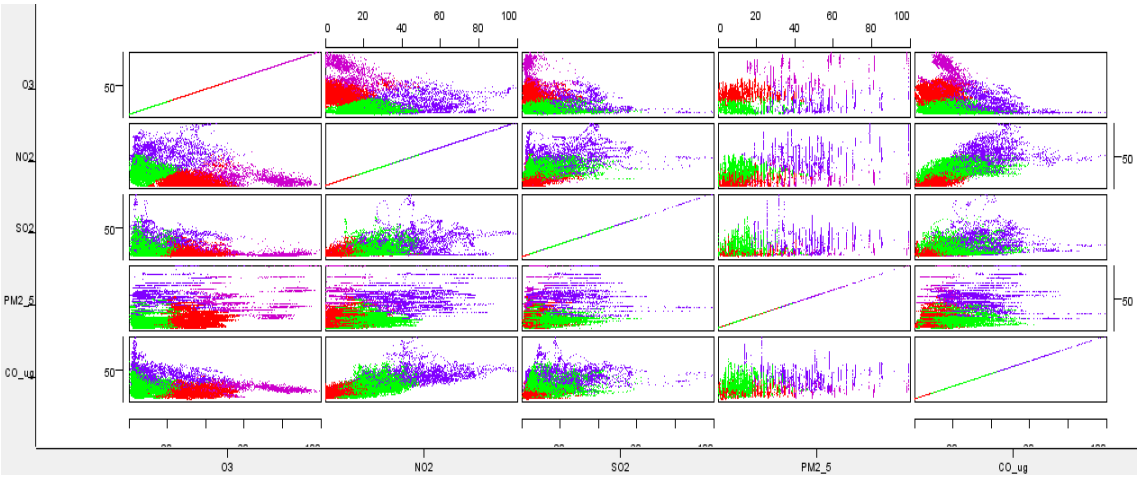


Imagen 17. Gráfico Matriz de dispersión, comparación entre contaminantes.

Fuente: (Marcia Matute, 2017)

Temporizador.

Este nodo contiene información de tiempo de ejecución individual y agregada para todos los nodos del flujo de trabajo.

Row ID	Name	Execution Time	Execution Time since Start
Node 1	File Reader	328	670
Node 3	k-Means	3082	6208
Node 5	Color Manager	1	1
Node 7	Scatter Plot	102	102
Node 9	Normalizer	315	654

Tabla 9. Tiempo de ejecución (milisegundos) de cada nodo (proceso).

Fuente: (Marcia Matute, 2017)

Conclusión.

La herramienta dispone de una amplia gama de algoritmos, permite aplicar procesos de limpieza de datos, posee una ventana de información sobre cada operador que contiene en su repositorio brindando así una guía para el usuario facilitando el desarrollo y configuración del modelo. Permite guardar la información generada del modelo en archivos .csv, arff entre otros. La versión de Knime Analytics Platform es gratuita.

Modelado en la herramienta RapidMiner.

La versión de prueba de 30 días de esta herramienta permite procesar hasta 10.000 registros, razón por la que se procedió a instalar una versión con licencia académica por un año, que permite procesar un número ilimitado de registros lo que es conveniente para poder evaluar la capacidad y tiempo de ejecución de la herramienta. La interfaz de trabajo de esta herramienta un tanto parecida a la herramienta KNIME permite ir armando y configurando el modelo de acuerdo a las necesidades del usuario. A continuación se muestra la interfaz de trabajo de esta herramienta.

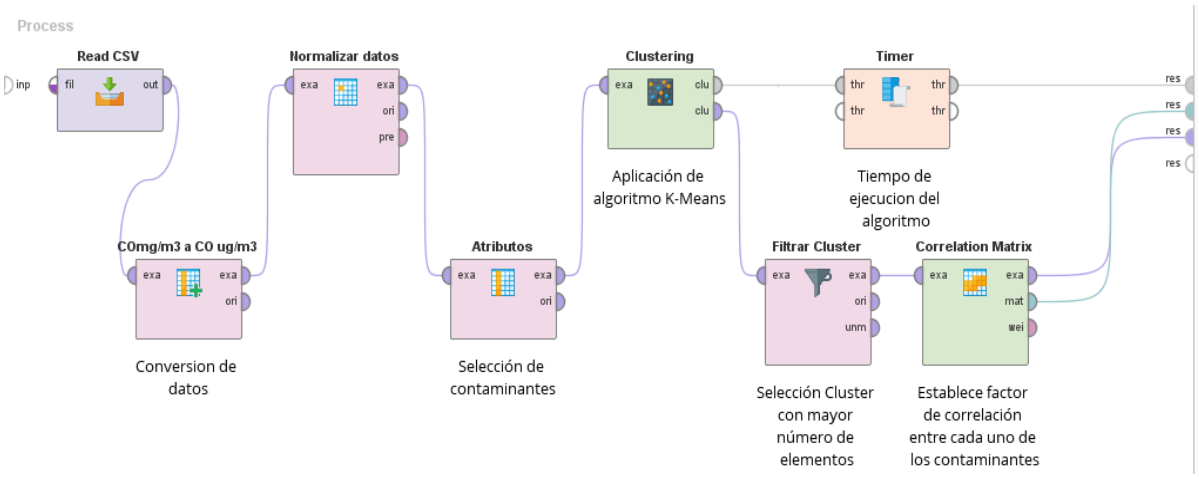


Imagen 18. Modelado con aplicación algoritmo K-Means en la herramienta RapidMiner.

Fuente: (Marcia Matute, 2017)

Presentación de Resultados.

Esta herramienta muestra dos interfaces: la primera que es el área de trabajo en la que se realizan los modelos y la segunda es una vista de los resultados, en esta se visualizan todos los resultados del proceso en tablas y también se puede visualizar los datos por medio de las gráficas.

La herramienta RapidMiner presenta un tiempo de ejecución de 2 segundos para todo el modelo y un tiempo de ejecución de 500 milisegundos para el algoritmo lo que es un excelente tiempo considerando el volumen de datos que se está procesando, los tiempos de ejecución de otros algoritmos alcanzan tiempos de hasta 2 minutos con tan solo un aproximado de doce mil registros.

RapidMiner presenta los resultados por medio de las siguientes interfaces:

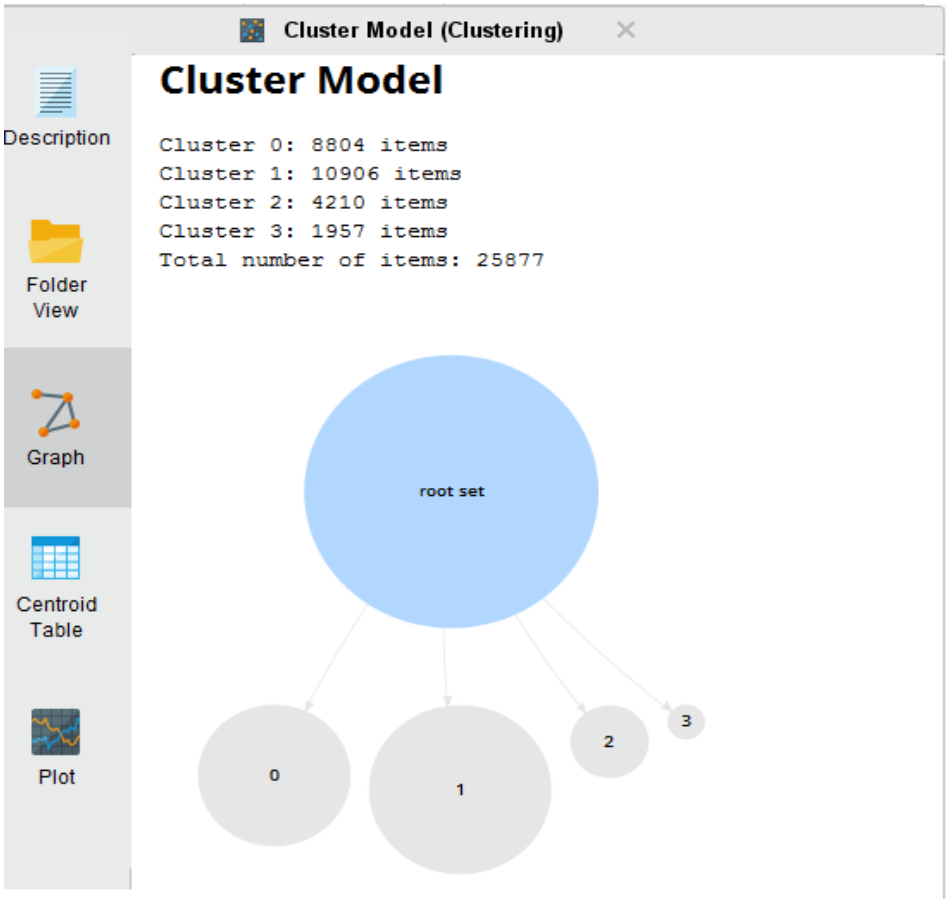


Imagen 19.Centroides, resultado de la aplicación de algoritmo K.Means, Herramienta RapidMiner.

Fuente: (Marcia Matute, 2017)

La imagen 19 describe que datos se asignaron a cada clúster e informa cuantos registros contiene cada uno de ellos y en la tabla 12 se presenta el valor de los centroides de cada contaminante.

Tabla de Centroides.

Attribute	cluster_0	cluster_1	cluster_2	cluster_3
O3	32.418	8.041	14.073	62.566
NO2	7.017	20.829	42.849	12.563
SO2	5.247	9.229	19.606	5.325
PM2_5	12.390	12.873	46.980	43.928
CO_ug	11.571	17.216	32.834	16.309

Tabla 10. Centroides, resultado de la aplicación de algoritmo K.Means, Herramienta RapidMiner.

Fuente: (Marcia Matute, 2017)

Para exportar o guardar los resultados de estas tablas se debe agregar al modelo un operador "Write" y seleccionar el tipo de extensión que se quiere ya sea CSV, Excel,ARFF, etc.

Tabla matriz de correlación.

En la siguiente tabla se muestra los valores de correlación que existe entre las variables de contaminación del aire, si los valores son altos y positivos existe una correlación positiva, si son muy bajos presentan una relación indirecta y si son cercanos a cero pierden relación.

Attributes	O3	NO2	SO2	PM2_5	CO_ug
O3	1	0.015	0.061	-0.093	-0.075
NO2	0.015	1	0.359	0.150	0.546
SO2	0.061	0.359	1	0.049	0.223
PM2_5	-0.093	0.150	0.049	1	0.131
CO_ug	-0.075	0.546	0.223	0.131	1

Tabla 11. Correlación entre contaminantes, herramienta RapidMiner.

Fuente: (Marcia Matute, 2017)

Se puede observar en la tabla 11 que O3 frente a PM2,5 y CO_ug tiene una factor de correlación inversa bajo, como se muestra en la **imagen 21** , también esta NO2 en la **imagen 22** frente al SO2 y CO_ ug, tiene factor de correlación positiva de 0.35 y 0.54 respectivamente lo que significa que existe un comportamiento similar entre estos contaminantes. Además los resultados muestran que existe una dependencia entre algunos contaminantes, existen espacios en los que se presentan un relación directa y en otros relación indirecta (**imagen 23)** de SO2 con O3 y PM2,5 cuando los valores de este están entre valores altos o muy bajos, pero si se acercan a cero se pierde dicha relación.

Gráficos de las correlaciones entre cada contaminante.

RapidMiner ofrece una variada gama de estilos de gráfico para la visualización de datos, es importante seleccionar la opción que presente los resultados de la forma más clara posible, en este caso de análisis lo que se busca es encontrar patrones de comportamiento

y relación entre las variables de contaminación atmosférica, por lo que se puede presentar los datos en graficas de series de tiempo como también en graficas de dispersión.

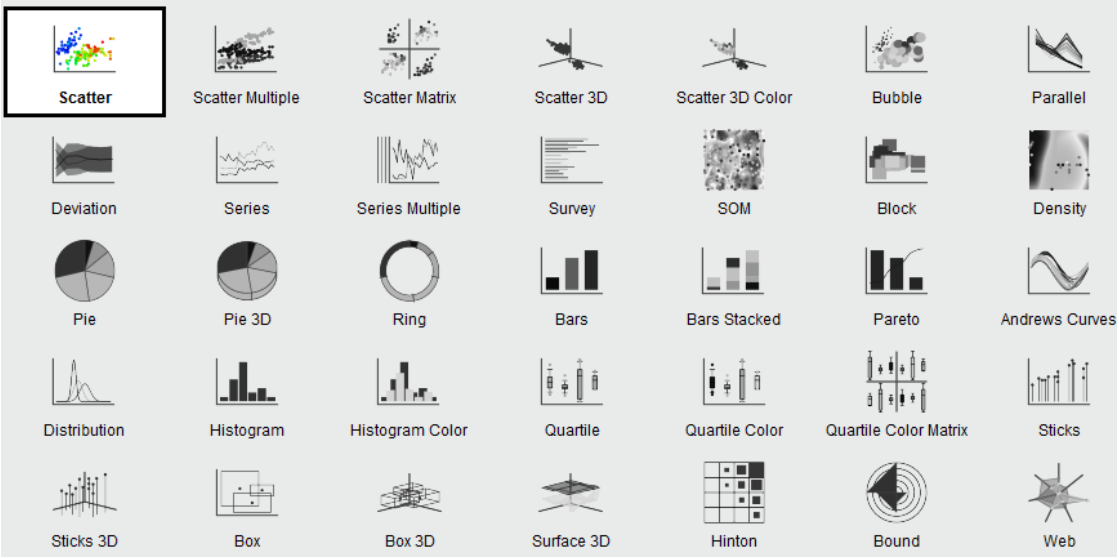


Imagen 20. Estilo de Gráficos en herramienta RapidMiner.

(Fuente: Herramienta RapidMiner)

En las siguientes imágenes utilizando graficas de series de tiempo se puede observar comportamiento de los contaminantes.

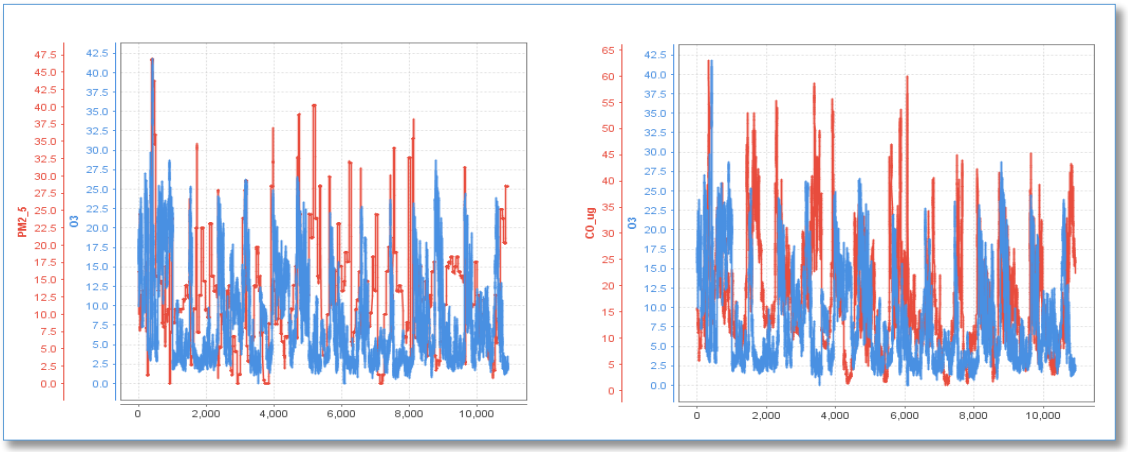


Imagen 21. Correlación con Contaminante O3, Grafico con herramienta RapidMiner.

Fuente: (Marcia Matute, 2017)

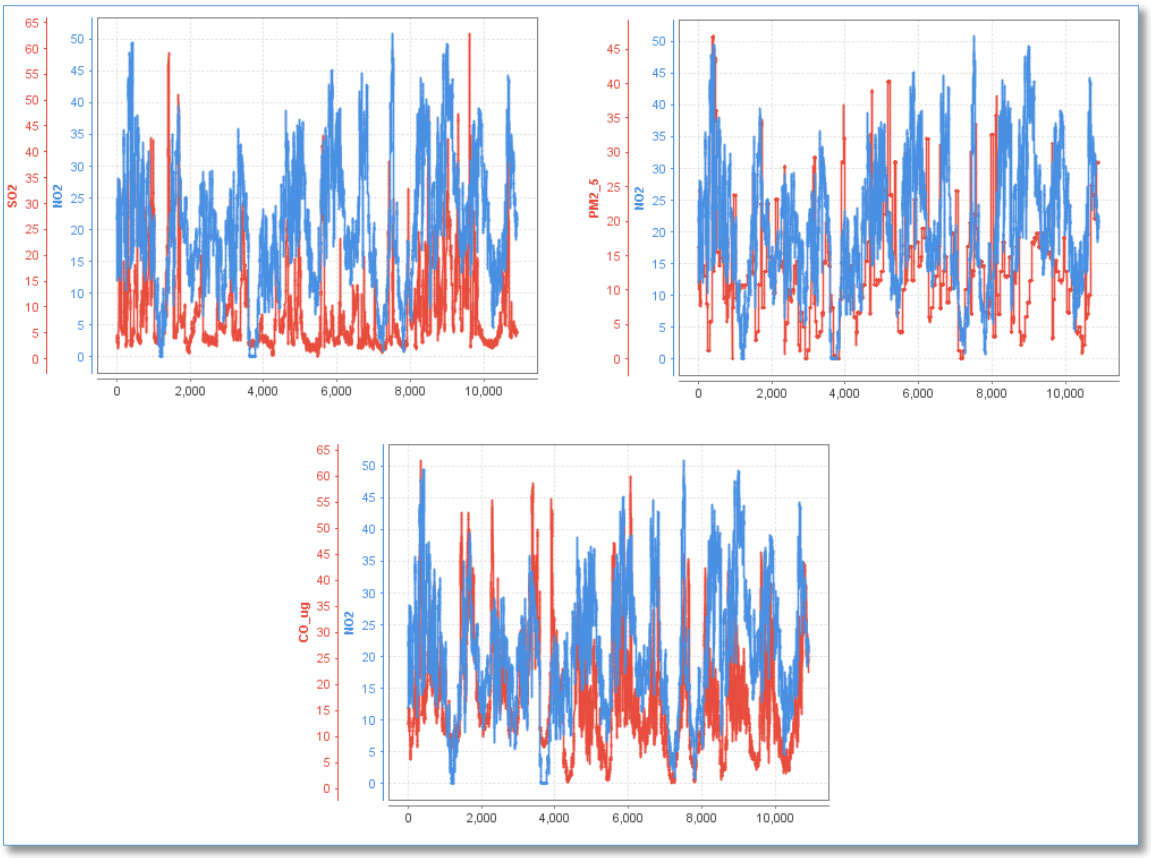


Imagen 22. Correlación con Contaminante NO2, Grafico con herramienta RapidMiner.

Fuente: (Marcia Matute, 2017)

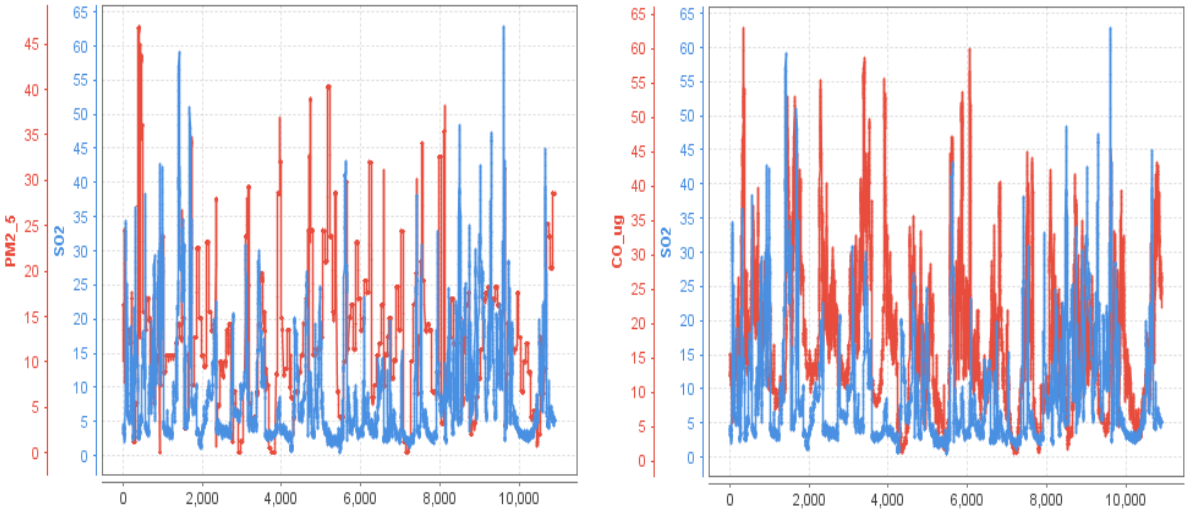


Imagen 23.Correlación con Contaminante SO2, Grafico con herramienta RapidMiner.

Fuente: (Marcia Matute, 2017)

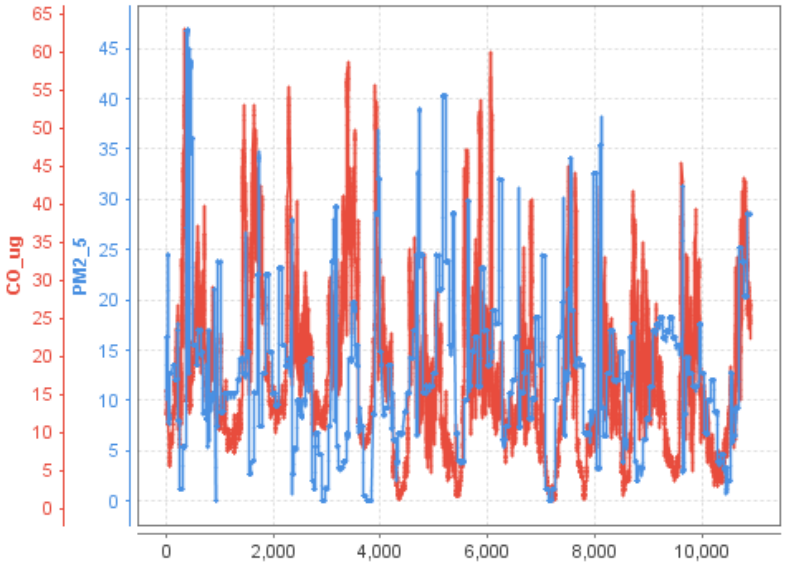


Imagen 24. Correlación con Contaminante PM2,5, Grafico con herramienta RapidMiner.

Fuente: (Marcia Matute, 2017)

Conclusión.

La herramienta soporta archivos con varias extensiones, permite visualizar los datos ingresados y presenta un informe de valores perdidos o erróneas en caso de existir, permite aplicar el número de filtros necesarios, aplicar fórmulas matemáticas para conversión de valores, segmentar la información, agregar condiciones, manipular la información de filas y columnas, sin dañar la base de datos original.

Posee una amplia gama de algoritmos y gráficas para representar los datos y visualizar los resultados. Mediante las gráficas de series de tiempo se puede revisar la existencia de correlación o patrones de comportamiento.

No existe forma de exportar los resultados en conjunto, se debe agregar un operador para cada resultado que se quiera exporta y utilizar en otras herramientas o ya sea para realizar los respectivos reportes.

Modelado en la herramienta WEKA

Weka es una herramienta de software gratuita, su interfaz presenta varios ambientes de trabajo, una de ella es “Explorer” y es la que utilizamos para el propósito de pruebas de

esta investigación ya que en esta interfaz se puede realizar aplicación de diferentes algoritmos sobre los datos de entrada. A diferencia de otras herramientas en Weka se selecciona la forma como introducir los datos con los que se va a trabajar, sea por medio de un fichero, conexión a una base de datos o si se introduce directamente la información en la aplicación.

En esta herramienta una vez que cargamos el archivo nos muestra un resumen de lo que este contiene como: el número de las variables existentes, valores perdidos, cantidad de datos existentes, resumen estadístico de los valores de cada variable. A continuación se muestra la interfaz de trabajo de esta herramienta.

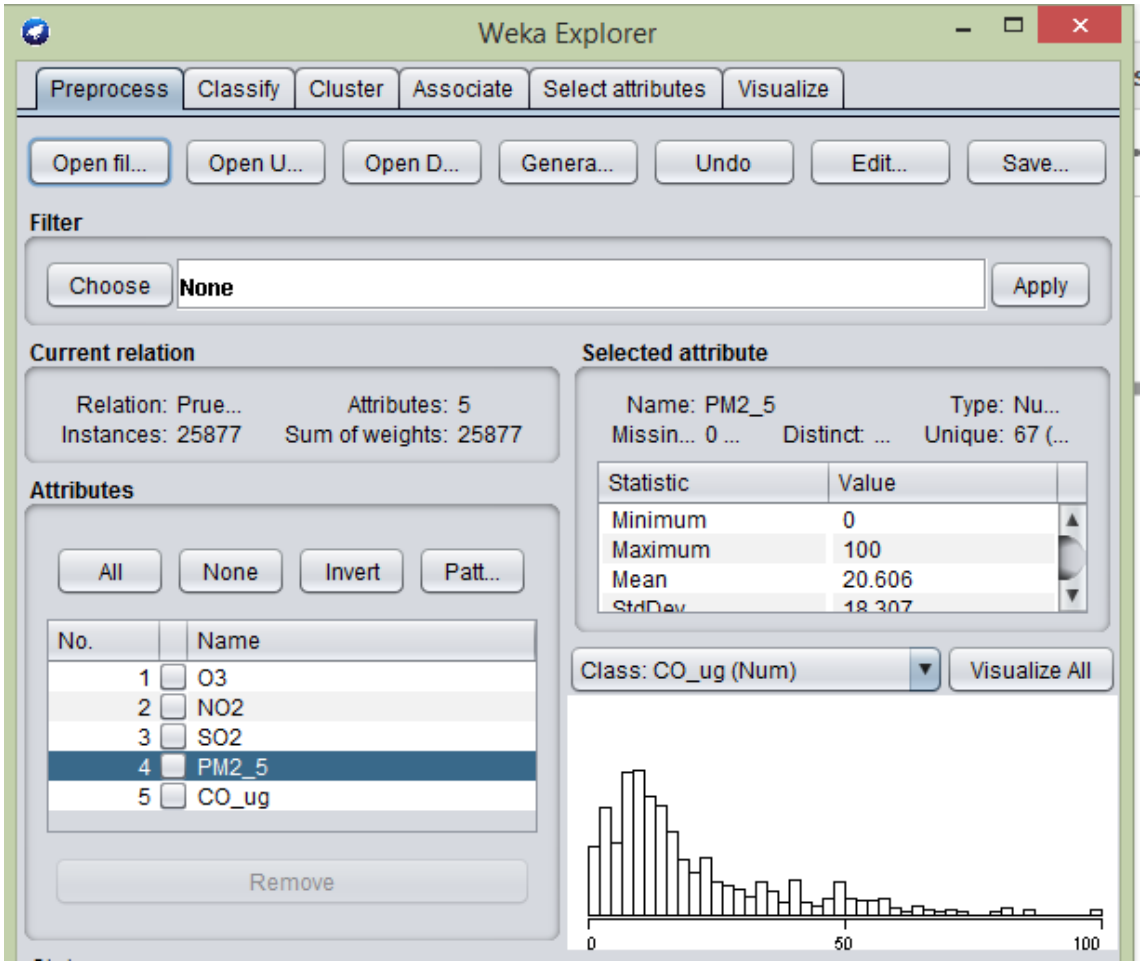


Imagen 25. Interfaz de trabajo. Herramienta WEKA.

Fuente: (Marcia Matute, 2017)

Para el preprocesamiento como el filtrado de los datos se elige fácilmente de la opción ***Filter*** en una de las pestañas en la interfaz, se configura según las necesidades y se aplica a los datos, de igual manera se procede con la aplicación del algoritmo seleccionando la pestaña ***Cluster***, elegir el algoritmo SimpleKMeans, configurarlo y aplicar a los datos. Si

bien esta herramienta permite procesar información de una forma rápida sin necesidad de creación de modelos, no permitió aplicar fórmulas matemáticas para conversión del contaminante CO, por lo que se debe tener la base de datos unificada antes de cargarla en la herramienta, no permite mostrar en tablas o archivo de texto el factor de correlación una vez aplicado el algoritmo, filtrar clúster de mayor tamaño y/o exportar los resultados obtenidos.

Presentación de Resultados.

La herramienta WEKA presenta un tiempo de ejecución del algoritmo de 0.36 segundos, una vez aplicado el algoritmo se presenta los siguientes resultados:

Ventana Clúster de Salida.

La imagen 26 muestra un informe de los resultados, valores de los Centroides, resultado de la aplicación de algoritmo K.Means, según la herramienta que se utilice varia la asignación de los datos a cada clúster, sin embargo los valores de los centroides de cada contaminante son idénticos a los valores obtenidos en las otras herramientas.

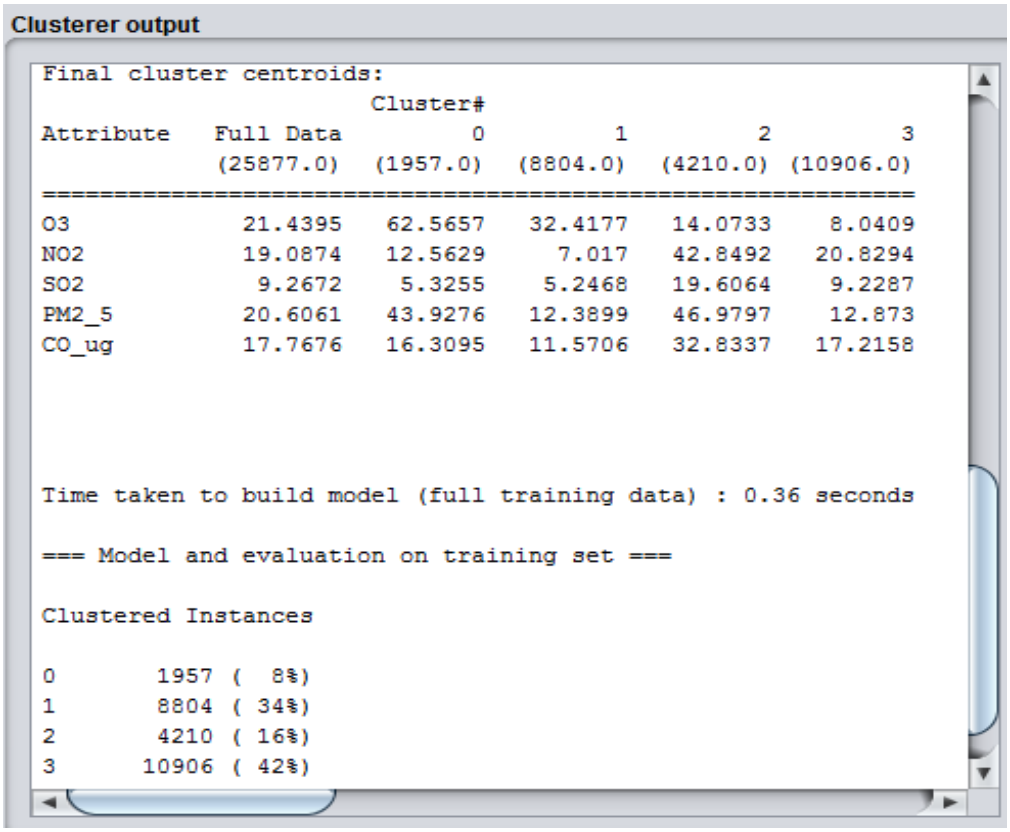


Imagen 26. Ventana de resultados valores de los centroides y porcentaje de datos asignado a cada clúster. Herramienta WEKA.

Fuente: (Marcia Matute, 2017)

Gráficos de visualización de los datos.

WEKA no posee variedad de gráficos, presenta únicamente la gráfica de barras del comportamiento de los contaminantes y la gráfica de la matriz de relación entre los contaminantes, lo que limita la exploración de los resultados.

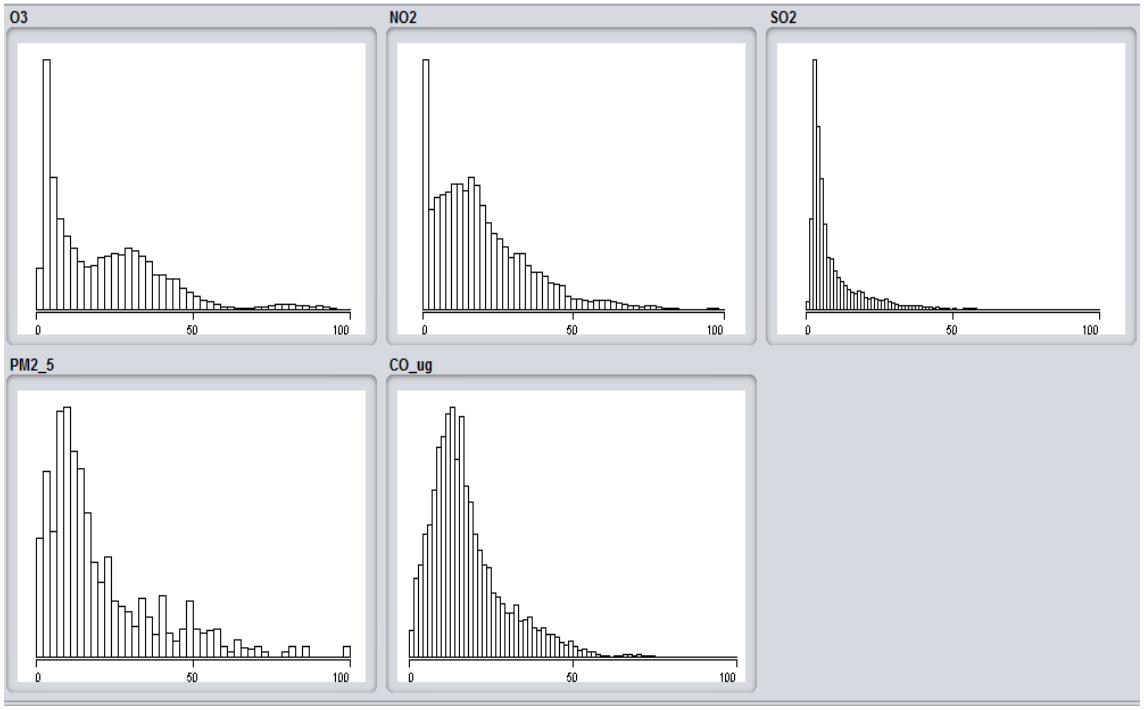


Imagen 27. Gráfico del comportamiento de cada contaminante.

Fuente: (Marcia Matute, 2017)

En la imagen 28 se muestra la pestaña de visualización, dicha pestaña presenta gráficos en 2D que son interactivos y relacionan pares de atributos, es decir, permite explorar en la gráfica la existencia de todas las asociaciones o correlaciones posibles entre cada atributo; pulsando en cada cuadro podemos ver la ampliación de la gráfica.

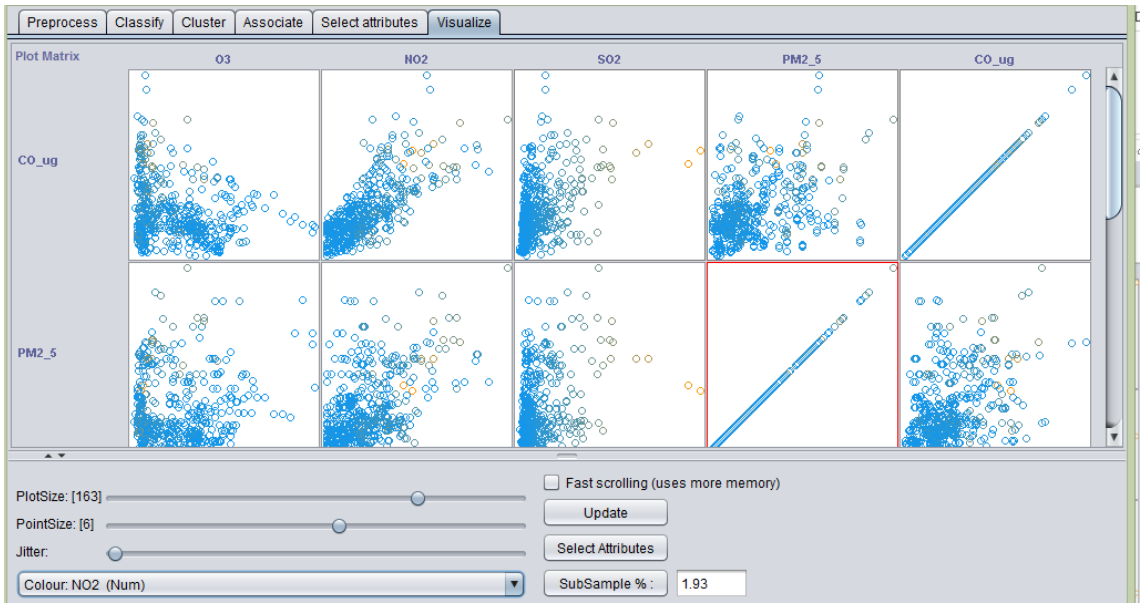


Imagen 28. Visualización de resultados. Gráficos herramienta WEKA.

Fuente: (Marcia Matute, 2017)

La imagen 29 muestra la gráfica individual de la correlación entre dos contaminantes, la herramienta WEKA permite almacenar esta información únicamente en formato (.arff).

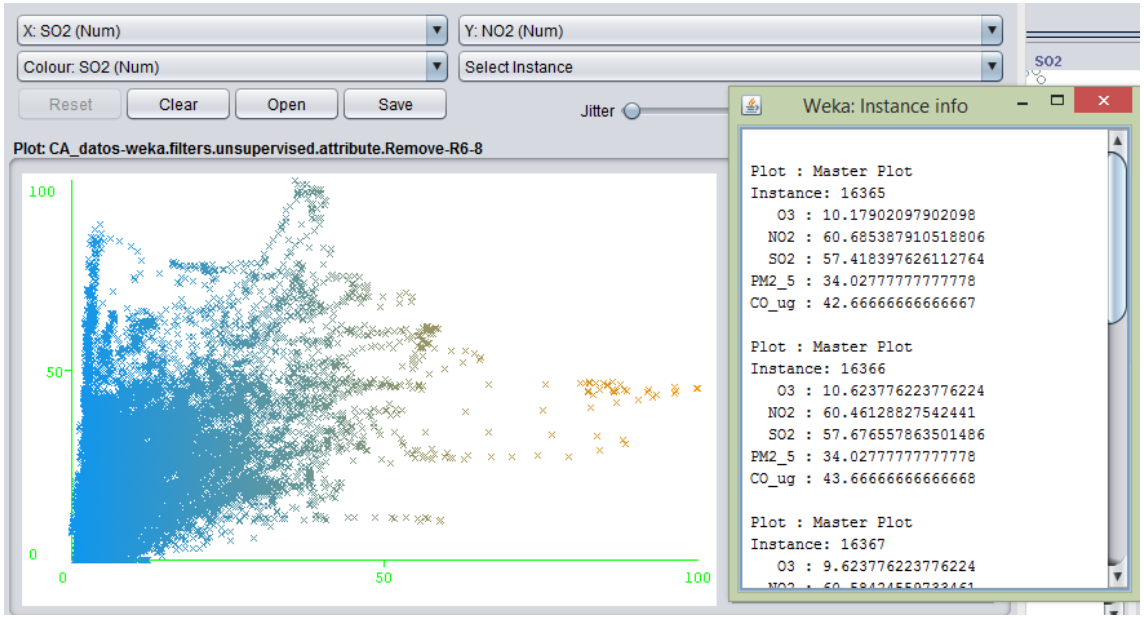


Imagen 29. Gráfico de correlación entre dos contaminantes SO2 y NO2. Herramienta WEKA.

Fuente: (Marcia Matute, 2017)

Conclusión.

La herramienta Weka permite identificar los campos que poseen valores erróneos y que no sean de utilidad un tema muy importante a la hora de procesar los datos esto ayuda a evitar obtener resultados erróneos o con sesgo.

Los tiempos de ejecución son bastante rápidos en comparación con los tiempos obtenidos en las otras herramientas de minería de datos (KNime, RapidMiner), que se evaluaron, su interfaz es intuitiva y de fácil de usar, por otra parte tiene limitadas gráficas para representar los valores y resultados, limitada capacidad para aplicar varios filtros a los datos, es decir permite aplicar un filtro a la vez.

Si bien el software permite interactuar con una variedad de fuentes de datos, sin embargo permite únicamente guardar los resultados en extensión .arff.

Capítulo V: Evaluación y Resultados.

Introducción

Es claro que existen una amplia gama de herramientas de minería de datos por lo que es preciso analizarlos y realizar su respectiva comparación. En este capítulo se presenta la matriz de calidad de software para las herramientas de minería de datos que se evaluaron en el capítulo anterior. Se califica las herramientas tomando en cuenta factores, como: gama de algoritmos disponibles, usabilidad, seguridad, viabilidad, velocidad, entre otras.

5.1. Metodología para evaluar las herramientas.

Dependiendo del campo donde se vaya a emplear un software es importante realizar una adecuada evaluación y este estudio de evaluación comparativo entre las herramientas de minería de datos estudiadas en los capítulos anteriores ayudan a entender qué elementos son importantes y que se evaluó.

Para empezar existen algunas metodologías para evaluar calidad de software entre ellas esta “Once Factores de McCall”, Método QSOS (Qualification and Selection of Open Source Software) para la evaluación del Software y los criterios de la metodología “ISO/IEC 25010”, estas metodologías evalúan áreas de trabajo del software en las que se define aspectos como: la funcionalidad, usabilidad, fiabilidad y rendimiento entre otras.

Acerca del Método QSOS, este modelo está diseñado para seleccionar y calificar productos de software en especias de código abierto. Este método consta de la siguiente metodología: Definición, Evaluación, Calificación, y selección. (QSOS, 2017)

Definición: en este paso se define varios elementos de la tipología como: Familia de Software, Tipos de Licencia, Tipos de comunidades. (Capitulo 3)

Evaluación: aquí se realiza la recolección de información acerca del producto en cuestión, se elabora la tarjeta de identificación del software (Tabla 17), así como se realiza la elaboración de la hoja de evaluación del software (Tabla 19).

Calificación: en este paso se definen los indicadores que traduzcan las necesidades y restricciones relacionadas, este paso se realiza las respectivas entrevista con las personas que van a utilizar el software (Tablas 14,15 y 16), con los resultados obtenidos se especifica los pesos que le asigna a cada aspecto del software y estos pesos son utilizados para calcular el puntaje final y ver qué software se ajusta más a las necesidades del usuario.

Selección: en este último paso se identifica el software que cumple y satisfaga los requerimientos del usuario.

Para el paso **calificación** y con el propósito de evaluar la mejor herramienta de minería de datos en variables de contaminación atmosférica se realizó una entrevista con personas de diferentes áreas de trabajo (con conocimientos técnicos: alto y medio) que van a participar en el uso de esta herramienta y se desarrolló la respectiva encuesta basada en los criterios de evaluación de software del modelo de McCall y de las Métricas usadas por QSOS, se utilizó los siguientes criterios:

Funcionalidad.- la funcionalidad del software ayuda a evaluar lo bien que la herramienta se adapta a diferentes dominios de problemas de minería de datos, ayuda a evaluar si la herramienta realiza correctamente la tarea encomendada. (Carey, Marjaniemi , Collier, & Sautter)

Criterios de evaluación

- Variedad de algoritmos.
- Metodología prescrita (Guía paso a paso).
- Validación del modelo.
- Flexibilidad del tipo de datos.

- Algoritmos modificables.
- Reportes.
- Exportación de Modelos.

Usabilidad.- Un problema con el uso de herramientas de minería fáciles de usar es su potencial mal uso, por ello es necesario saber qué nivel de conocimientos técnicos (bajo, medio, alto) debe poseer el usuario para manejar la herramienta. La herramienta no solo debe ser fácil de usar, debe guiar al usuario hacia la extracción correcta de datos y el usuario debe sentirse cómodo al trabajar con esta. (Carey, Marjaniemi , Collier, & Sautter)

Criterios de evaluación

- Interfaz de usuario amigable.
- Visualización de datos.
- Informe de errores.
- Historial de acciones.

Rendimiento.- Se refiere a la capacidad de una herramienta para manejar las diferentes fuentes de datos de una forma eficiente y si la herramienta responde con la velocidad apropiada. Aquí se toma en cuenta los siguientes criterios: (Carey, Marjaniemi , Collier, & Sautter)

Criterios de evaluación.

- Variedad de la plataforma.
- Acceso heterogéneo a los datos (interactúa el software con una variedad de fuentes de datos).
- Tamaño de los datos.
- Eficiencia.
- Interoperabilidad.

Soporte de tareas auxiliares.- permite al usuario realizar los procesos necesarios de limpieza, manipulación, transformación, visualización y otras tareas de datos que admiten la extracción de datos, ya que ningún conjunto de datos puede estar totalmente listo para la minería. (Carey, Marjaniemi , Collier, & Sautter)

Criterios de evaluación.

- Limpieza de datos.

- Remplazar valores.
- Filtrado de datos.
- Eliminación de registros.
- Manejo de espacios en blanco.
- Manipulación de metadatos.
- Resultado Feedback.

Una vez que tenemos los criterios establecidos se asigna peso a estos para poder establecer que tan importante es cada uno de ellos. (QSOS, 2017)

Peso(Nivel de Importancia)	**Calificación**
Muy Alto	5
Alto	4
Medio	3
Bajo	2
Muy Bajo	1

Tabla 12. Nivel de importancia por criterio

Fuente: (Marcia Matute, 2017)

Resultados de la encuesta.

Indica el nivel de importancia de los criterios de Funcionalidad en una herramienta.

Variedad de algoritmos	Metodología Prescrita	Validación del modelo creado	Flexibilidad del tipo de datos	Algoritmos modificables	Generación de Informes de resultados	Exportación de Modelos	**Funcionalidad**
5	5	5	4	4	5	4	
5	5	5	5	4	5	4	
3	3	3	2	2	2	3	
5	4	4	5	3	4	3	**25%**
3	5	4	5	3	4	5	
4	5	4	5	4	5	5	
5	5	5	5	2	5	4	

4,29	**4,57**	**4,29**	**4,43**	**3,14**	**4,29**	**4**

Tabla 13. Respuestas criterios de Funcionalidad.

Fuente: (Marcia Matute, 2017)

Indica el nivel de importancia de los criterios de Usabilidad en una herramienta

Interfaz de usuario amigable	Facilidad de uso	Visualización de datos	Informe de errores	Historial de acciones	Variedad de dominio	**Usabilidad**
4	4	5	5	5	5	
5	5	4	5	4	4	
3	3	5	3	2	3	
4	5	5	4	3	3	**26%**
5	5	4	4	4	5	
4	4	5	5	4	5	
5	3	5	3	3	4	
4,29	**4,14**	**4,71**	**4,14**	**3,57**	**4,14**	

Tabla 14. Respuestas criterios de Usabilidad.

Fuente: (Marcia Matute, 2017)

Indica el nivel de importancia de los Rendimiento en una herramienta.

Variedad de la plataforma	Acceso a datos heterogéneo	Tamaño de datos que puede procesar	Eficiencia	Interoperabilidad	Robustez	**Rendimiento**
4	4	5	5	4	4	
5	5	5	5	5	5	
1	2	2	2	1	2	**26%**
2	4	5	5	5	5	
5	4	4	5	4	5	
5	5	5	5	5	5	

4	5	5	5	4	5
Pesos promedio por criterio					
3,71	**4,14**	**4,43**	**4,57**	**4**	**4,43**

Tabla 15. Respuestas criterios de Rendimiento.

Fuente: (Marcia Matute, 2017)

Indica el nivel de importancia de los criterios de Soporte de tareas auxiliares en una herramienta.

Limpieza de datos	Sustitución de valores	Transformación de valores	Filtrado de datos	Manejo de espacios en blanco	Manipulación de metadatos	Resultados	**Soporte de tareas Auxiliares**
4	4	4	4	4	3	4	
4	4	5	5	4	5	5	
1	2	2	2	2	3	1	
3	3	4	5	3	3	4	
5	4	5	4	4	5	5	**23%**
5	5	5	5	4	5	5	
3	4	4	5	2	3	4	
Pesos promedio por criterio							
3,57	**3,71**	**4,14**	**4,29**	**3,29**	**3,86**	**4,00**	

Tabla 16.Respuestas criterios de Soporte de tareas auxiliares

Fuente: (Marcia Matute, 2017)

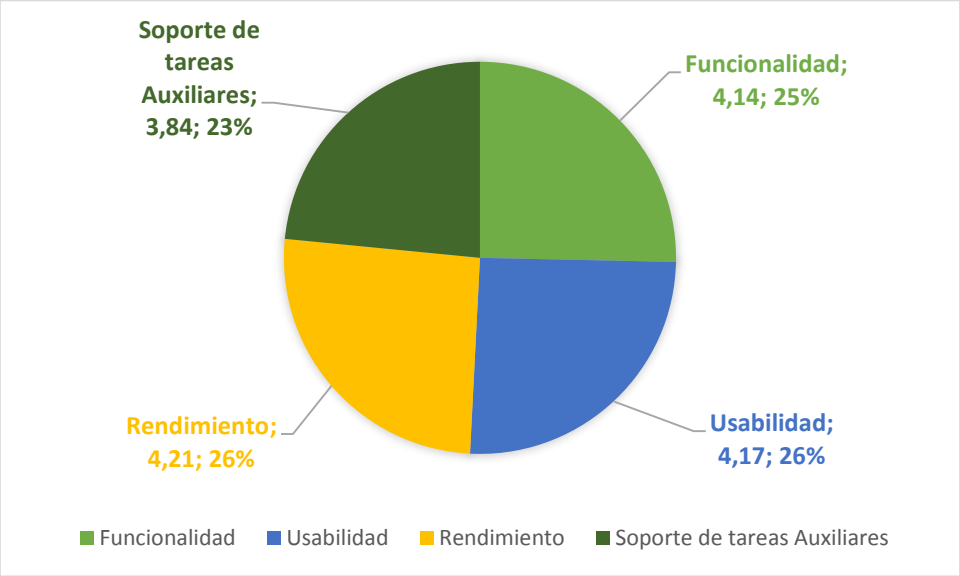


Imagen 30. Resumen encuesta sobre Herramientas de minería de datos.

Fuente: (Marcia Matute, 2017)

Con los datos resultantes de esta encuesta se procede a llenar la matriz de evaluación de software de minería de datos.

Además se tienen en cuenta que no deben faltar las generalidades que caracterizan al software, donde se incluyen parámetros como nombre de la aplicación, versión, país de origen, licencia, creador, entre otros. La siguiente tabla nos presenta dicha información.

Criterios Generales	Herramientas de Data Mining				
Nombre de la aplicación	**RapidMiner Studio**	**KNIME**	**Weka**	**IBM SPSS Modeler**	**SAS Enterprise Miner**
Tipo	Análisis y minería de datos.	Minería de datos	Aprendizaje automático y minería de datos	Minería de datos y análisis predictivo	Gestión de datos y análisis predictivo.
Versión	8.1	3.1.1	3.8.1		14.3
País de origen	Massachusetts, EE.UU	Constanza, Alemania	Nueva Zelanda	Reino Unido	Carolina del Norte
Creador	Rapid-I	KNIME AG	Universidad de Waikato	IBM Corp.	SAS Institute
Licencia	*AGPL(Affero General Public License) *Educacional	Licencia Pública General de GNU	Licencia Pública General de GNU.	Software propietario	Software propietario
Lenguajes de programación para su desarrollo	Java	Java	Java	Java	Lenguaje de programación C.
Página Web de la aplicación	https://my.rapidminer.com	https://www.knime.com/	https://www.cs.waikato.ac.nz/ml/index.html	https://www.ibm.com/products/spss-modeler	https://www.sas.com/en_us/insights/analytics/predictive-analytics.html
Página de descarga	https://my.rapidminer.com/nexus/account/index.html#downloads	https://www.knime.com/downloads	https://www.cs.waikato.ac.nz/ml/weka/downloading.html	https://dal05.objectstorage.softlayer.net/v1/AUTH_14da1a30-d001-4a2c-9060-cadaf68ff44d/yp-ngp-spss-dal05/ModelerSub/Subscription/20180	http://ftp.sas.com/techsup/download/hotfix/HF2/A9S.html

				228/ModelerSub_Setup.exe	
Sistema Operativo	Windows 32 y 64 bits	Windows 32 y 64 bits	Windows 32 y 64bits	Windows	Windows
	Linux Requiere: Java 8	Linux 32 y 64 bits	Linux Requiere: Java 9	Linux	Linux
	Mac Requiere: Mac OS 10.8+	Mac	Mac OS X	Unix	Unix
				Mac OS X	OpenVMS Alpha
Motor de base de datos soportado	Java	Java	Java	Java	Java
Documentación Técnica	Disponible en la página https://docs.rapidminer.com/	Incluida en la instalación. Disponible en la página https://www.knime.com/documentation.	Disponible en la página https://www.cs.waikato.ac.nz/ml/weka/documentation.html	Disponible en la pagina https://www.ibm.com/es-es/marketplace/spss-modeler/resources#product-header-top	Disponible en la página http://support.sas.com/documentation/
Actualizaciones	8.1v. 6 de febrero de 2018	3.1.1v. 29 de enero de 2016	3.8.1v. Octubre 2016	18.v. 1 junio de 2017	14.3v. 01 de marzo de 2018
Soporte Técnico	Tipos de soporte: comunitario y empresarial.	Tipo de soporte: Comunitario y empresarial.	Tipos de soporte: comunitario y foros en vivo ofrecidos por la empresa Pentaho.	Tipo de soporte:	Tipo de soporte: Asistente en línea.
	https://docs.rapidminer.com/latest/support/	https://www.knime.com/knime-community	https://www.cs.waikato.ac.nz/ml/weka/help.html	https://www.ibm.com/support/home/?lnk=msu	http://support.sas.com/software/products/enterprise-miner/index.html#s1=3

Tabla 17. Tarjeta descriptiva de software

5.1.1. Análisis de las herramientas de minería de datos

Para determinar que herramienta software de Data Mining cumple en su mayoría los criterios evaluados se realizaron una tabla comparativa entre los productos RapidMiner, Weka, Knime utilizando los factores de McCall. Los otros productos como SAS Enterprise Miner e IBM SPSS Modeler no se incluyen en esta tabla ya que debido a las limitaciones en las versiones libres de estas herramientas no se pudo completar la etapa de experimentación y no se obtuvo resultados que ayudaran en esta investigación.

Una vez que los criterios han sido ponderados con respecto al conjunto de necesidades específicas, las herramientas ahora se pueden calificar para comparación.

Los criterios dentro de la categoría Funcionalidad, Usabilidad, Rendimiento y Soporte de tareas auxiliares tienen pesos asignados para que el peso total dentro de cada categoría sea igual a 1.00 o 100% según la importancia del criterio que se busca posea la herramienta.

La asignación de pesos está basada en el modelo QSOS (Capitulo 5,"Metodologia") y se califica las herramientas usando la siguiente escala:

Calificación	Peso
Cumple totalmente	**4**
Cumple medianamente	**3**
Cumple parcialmente	**2**
No cumple	**1**

Tabla 18. Escala de Calificación

Fuente: (Marcia Matute, 2017)

EVALUACIÓN DE LAS HERRAMIENTAS DE MINERÍA DE DATOS							
Criterio	**Peso**	**RapidMiner**		**KNIME**		**Weka**	
Funcionalidad	**0,25%**	**Calificación**	**Puntaje**	**Calificación**	**Puntaje**	**Calificación**	**Puntaje**
Variedad de algoritmos	0,35	4	1,4	4	1,4	4	1,4
Metodología prescrita.	0,05	2	0,1	3	0,15	3	0,15
Validación del modelo.	0,10	4	0,4	4	0,4	1	0,1
Flexibilidad del tipo de datos.	0,30	4	1,2	4	1,2	4	1,2
Reportes	0,15	3	0,45	2	0,3	2	0,3
Exportación de Modelos	0,05	4	0,2	4	0,2	3	0,15
Puntuación de categoría			**3,75**		**3,65**		**3,3**
Usabilidad	**0,26%**	**Calificación**	**Puntaje**	**Calificación**	**Puntaje**	**Calificación**	**Puntaje**
Interfaz de usuario amigable	0,20	4	0,8	4	0,8	4	0,8
Facilidad de Uso	0,20	3	0,6	3	0,6	4	0,8
Visualización de datos	0,35	3	1,05	2	0,7	2	0,7
Informe de errores	0,10	2	0,2	3	0,3	1	0,1

Historial de acciones	0,15	3	0,45	1	0,15	1	0,15
Puntuación de categoría			**3,1**		**2,55**		**2,55**
Rendimiento	**0,26%**	**Calificación**	**Puntaje**	**Calificación**	**Puntaje**	**Calificación**	**Puntaje**
Variedad de la plataforma	0,05	4	0,2	4	0,2	4	0,2
Acceso heterogéneo a los datos	0,10	4	0,4	4	0,4	4	0,4
Tamaño de los datos.	0,30	4	1,2	4	1,2	4	1,2
Eficiencia/ tiempo de respuesta	0,30	3	0,9	3	0,9	4	1,2
Interoperabilidad	0,05	4	0,2	4	0,2	2	0,1
Robustez.	0,20	4	0,8	4	0,8	4	0,8
Puntuación de categoría			**3,7**		**3,7**		**3,9**
Soporte de tareas auxiliares	**0,23%**	**Calificación**	**Puntaje**	**Calificación**	**Puntaje**	**Calificación**	**Puntaje**
Limpieza de datos	0,10	4	0,4	4	0,4	2	0,2
Remplazar valores	0,05	4	0,2	4	0,2	2	0,1
Filtrado de datos	0,15	4	0,6	4	0,6	3	0,45

Eliminación de registros	0,05	4	0,2	4	0,2	3	0,15
Manejo de espacios en blanco	0,05	4	0,2	4	0,2	2	0,1
Manipulación de metadatos	0,30	4	1,2	4	1,2	3	0,9
Resultado Feedback	0,30	4	1,2	4	1,2	2	0,6
Puntuación de categoría			**4**		**4**		**2,5**

Tabla 19. Evaluación de las herramientas de minería de datos.

Fuente: (Marcia Matute, 2017)

Se realizó la evaluación de las herramientas con una escala discreta de calificación, se plantea indicadores desde "1" que pertenece a que **no cumple** con el criterio, es decir, la herramienta no posee dicha funcionalidad, hasta el indicador "4" que corresponde a que **cumple totalmente**, quiere decir, que la herramienta posee dicha funcionalidad. Se asignó los valores a cada criterio según los resultados obtenidos en el proceso de experimentación.

5.1.2. Análisis de resultados de la evaluación de herramientas de minería de datos.

En este punto se identifica aspectos relacionados con la evaluación y comparación realizada a los softwares de minería de datos RapidMiner, Knime y Weka, se analiza los valores obtenidos en los criterios de Funcionalidad, Usabilidad, Rendimiento y Soporte de tareas auxiliares, se presenta el análisis estadístico que ayuda en el proceso de definición de la herramienta adecuada para el desarrollo de minería de datos en variables de contaminación atmosférica.

Se presenta el siguiente resumen de resultados.

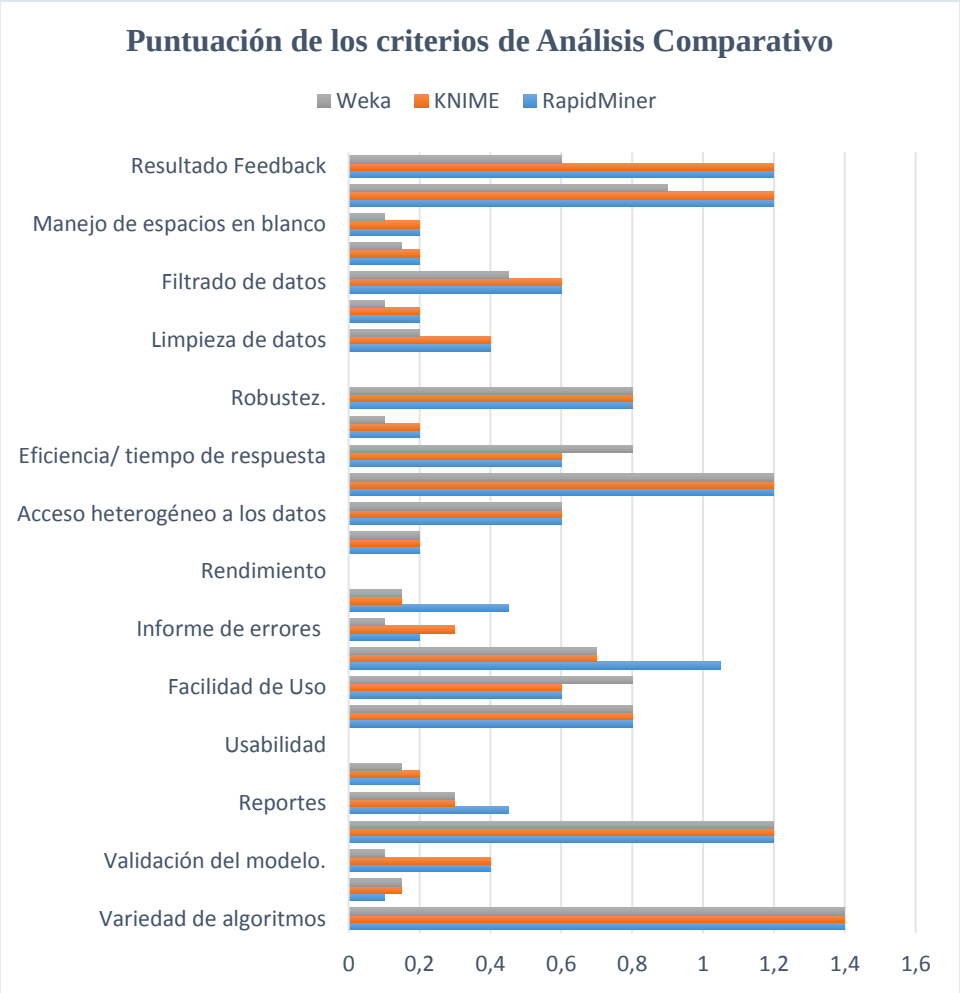


Imagen 31. Comparación entre las Herramientas Minería de datos.

En la imagen 31 se observa que las herramientas RapidMiner y Knime en la mayoría de criterios van a la par, sin embargo RapidMiner a Knime en presentación de "Reportes", "Visualización de datos", otro punto a favor RapidMiner es el tiempo de respuesta, se presentaron 3018 milisegundos para Knime y 1907 milisegundos para RapidMiner.

La herramienta Weka aunque presenta un tiempo de ejecución menor a las otras herramientas, obtuvo 63% en Soporte de tareas auxiliares a diferencia de las otras dos que obtuvieron 100% en esta categoría.

CRITERIOS	PUNTAJE RapidMiner	% RapidMiner	PUNTAJE Knime	% Knime	PUNTAJE Weka	% Weka
Funcionalidad	3,75	94%	3,65	91%	3,3	83%
Usabilidad	3,1	78%	2,55	64%	2,55	64%

Rendimiento	3,7	93%	3,7	93%	3,9	98%
Soporte de tareas auxiliares	4	100%	4	100%	2,5	63%
PROMEDIO	**3,64**	**91%**	**3,48**	**87%**	**3,06**	**77%**

Tabla 20. Promedio. Fuente:

(Marcia Matute, 2017)

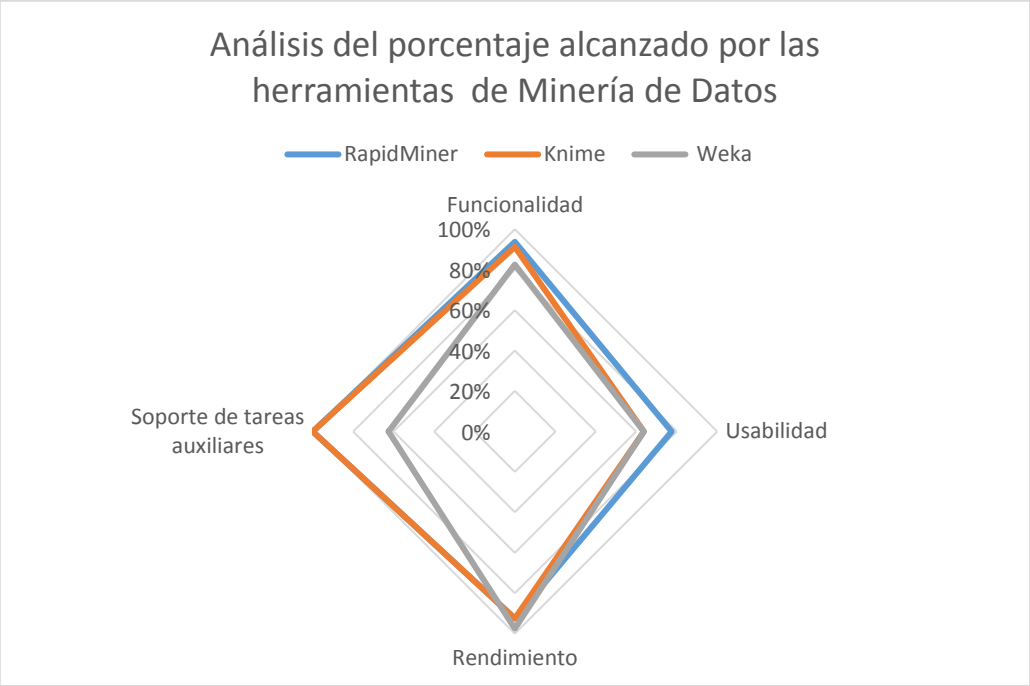


Imagen 32. Porcentaje resultante de la evaluación de las herramientas.

Fuente: (Marcia Matute, 2017)

La imagen 32 muestra que la herramienta RapidMiner y Knime tienen porcentajes parecidos en Funcionalidad, Rendimiento y Soporte a tareas auxiliares pero RapidMiner supera a las herramientas Weka y Knime en Usabilidad.

5. CONCLUSIONES.

Se seleccionaron cinco herramientas de minería de datos como son: Knime, RapidMiner, IBM SPSS Modeler, SAS Enterprise Miner y también Weka, se eligieron estas herramientas basados en el cuadrante mágico de Gartner de febrero del 2017.

En este estudio se evaluó el funcionamiento de las herramientas KNime, RapidMiner y Weka, se indaga en sus características principales, ventajas y desventajas, adquiriendo una comprensión general de los softwares estudiados en este trabajo de investigación.

La experimentación y evaluación de las herramientas realizadas con un conjunto de datos de aproximadamente 25877 registros; da como resultado que la herramienta para minería de datos en variables de contaminación atmosférica que cumple con la mayoría de los requisitos es RadipMiner, dado que supera en un 14% a la herramienta Weka y en un 4% a la herramienta KNime, con un tiempo de ejecución de 500 milisegundos además se confirma que RapidMiner es la plataformas de ciencia de datos y aprendizaje automático líder por quinto año consecutivo en el cuadrante mágico según Gartner en su reporte de febrero de 2018.

No obstante la superioridad de RapidMiner ante Knime estadísticamente no es realmente significativa, la herramienta KNime es de software libre pero tiene problemas con la gestión del modelo, en la administración y documentación de grandes flujos de trabajo, además de la capacidad limitada en exploración y visualización de datos, característica esencial para poder observar e interpretar fácilmente la información, RapidMiner se puede obtener de forma gratuita con licencia educacional permitiendo trabajar con volumen de datos ilimitados debiendo ser renovada cada año, por ello es necesario tener en cuenta el costo de implementación de la herramienta con el fin de que la herramienta se ajuste favorablemente a las necesidades del usuario.

Con este análisis se puede concluir que uso de técnicas de Minería de Datos en especial clustering con series de tiempo es una alternativa viable para encontrar relaciones entre variables de contaminación atmosférica y pronosticar comportamientos de las mismas.

El objetivo plantado de evaluar herramientas de minería de datos en variables de contaminación atmosférica se efectuó adecuadamente en las herramientas seleccionadas, adicionalmente al no tener la base de datos adecuadamente estructurada, fue necesario aplicar procesos de limpieza de datos de datos para obtener un conjunto de datos óptimo para las respectivas pruebas dentro de las herramientas.

6. RECOMENDACIONES.

En vista de que la selección adecuada de la herramienta de minería de datos es una etapa crítica en la que podemos adquirir una herramienta posiblemente costosa y que no cumpla con nuestros requerimientos se recomienda primeramente entender el negocio y la problemática del mismo, comprender los datos para saber que se desea conseguir con esta información y ponerlos en un formato adecuado para su posterior procesamiento.

En cuanto a la evaluación de las herramientas se recomienda no evaluar más de 5 herramientas, de esta forma no se sacrificaría el factor tiempo enfocándonos en evaluar herramientas que posiblemente no aporten a nuestro objetivo y podremos dedicar más tiempo a productos que realmente nos ayuden.

Por último no se debe apresurar a manejar las herramientas sin un previo conocimiento sobre las herramientas ya que el manejo inadecuado podría acarrear resultados erróneos.

7. REFERENCIAS BIBLIOGRÁFICAS.

ArcGIS. (2016). Obtenido de Fundamentos de los gráficos de matriz de dispersión: http://desktop.arcgis.com/es/arcmap/10.3/map/graphs/scatter-plot-matrix-graphs.htm

Beltrán, M. B. (2010). *Minería de datos.* Recuperado el 2017, de http://bbeltran.cs.buap.mx/NotasMD.pdf

Carey, B., Marjaniemi , C., Collier, K., & Sautter, D. (s.f.). *A Methodology for Evaluating and Selecting Data Mining Software.* doi:10.1109/HICSS.1999.772607

Clave. (2017). *Diccionario Clave.* Obtenido de http://clave.smdiccionarios.com/app.php

Daedalus S.A. (2002). *ISSUU.* Obtenido de Minería de Datos: https://issuu.com/hmorenop/docs/mineria-de-datos---daedalus-white-paper

De la Puente, M. (2010). Gestión del conocimiento y Minería de Datos. *Consultora de Ciencias de la Información*.

Díaz, A., & Suárez, Y. (2009). Herramientas de minería de datos. *Revista Cubana de Ciencias Informaticas*, 3(3-4).

Frontline Systems Inc. . (2017). *XLMiner*. Obtenido de https://www.xlminer.com/

Fundación para la Salud Geoambiental. (2013). *Material Particulado*. Obtenido de http://www.saludgeoambiental.org/material-particulado

Gaitán, M., Cancino, J., & Behrentz, E. (2007). Analysis of Bogota's Air Quality. *Revista de Ingeniería*, 82.

García, F. (2013). Aplicación de técnicas de minería de datos a datos obtenidos por el centro Andaluz del Medio Ambiente(CEAMA). *Uiversidad de Granada*.

Gartner, Inc. (14 de Febreso de 2017). *Gartner*. Obtenido de Magic Quadrant for Data Science Platforms: https://www.gartner.com/doc/reprints?id=1-3TK9NW2&ct=170215&st=sb

IBM. (2017). *IBM SPSS Modeler*. Obtenido de https://www.ibm.com/us-en/marketplace/spss-modeler.

Ingenio Virtual. (2017). *Ingenio Virtual*. Obtenido de http://www.ingeniovirtual.com/tipos-de-graficos-y-diagramas-para-la-visualizacion-de-datos/

KNIME AG. (2017). *Plataforma de análisis KNIME*. Obtenido de https://www.knime.com/products

Lima, J. F. (2017).

Martínez, E., & Díaz, Y. (2004). *Contaminación Atmosférica* (Vol. 45). Univ de Castilla La Mancha. Recuperado el Julio de 2017

Microsoft. (Noviembre de 2016). *Conceptos de Minería de datos*. Obtenido de https://msdn.microsoft.com/es-es/library/ms175595(v=sql.120).aspx

Ortega Guaman, J., & Orellana Cordero, M. (2018). *Universidad del Azuay.* Obtenido de Dspace Universidad del Azuay "Trabajos de grado": http://dspace.uazuay.edu.ec/handle/datos/7800

Piatetsky-Shapiro, G., & Mayo, M. (2017). *KDnuggets™*. Obtenido de CRISP-DM, still the top methodology for analytics, data mining, or data science projects: https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html

Predictive Analytics Today. (2017). *Predictive Analytics Today*. Obtenido de https://www.predictiveanalyticstoday.com/sas-enterprise-miner/

QSOS. (2017). *Metodo QSOS*. Obtenido de http://www.qsos.org/method

RapidMiner. (2017). *RapidMiner.* Obtenido de https://docs.rapidminer.com/

RapidMiner, Inc. (2017). *RapidMiner Studio*. Obtenido de https://rapidminer.com/products/studio/

Rodríguez Rojas, D. (2010). *Metodología para el Desarrollo de Proyectos en Minería de Datos CRISP-DM.* Recuperado el Septiembre de 2017, de http://www.oldemarrodriguez.com/yahoo_site_admin/assets/docs/Documento_CRISP-DM.2385037.pdf

SAS Institut Inc. (2017). *SAS Enterprise Miner*. Obtenido de https://www.sas.com/en_us/software/enterprise-miner.html

Sellers, I. (05 de 04 de 2017). *Publicación de los contaminantes atmosféricos de la estación de monitoreo en tiempo real de la ciudad de Cuenca, utilizando servicion estándares OGC.* Obtenido de http://dspace.uazuay.edu.ec/bitstream/datos/2546/1/09734.pdf

Timarán, S. R., Hernández, I., Caicedo, S. J., Hidalgo, A., & Alvarado, J. C. (2106). *El proceso de descubrimiento de conocimiento en bases de datos. En Descubrimiento de patrones de desempeño académico con árboles de decisión en las competencias genéricas de la formación profesional.* Bogotá: Universidad Cooperativa de Colombia. Recuperado el 2017, de http://ediciones.ucc.edu.co/index.php/ucc/catalog/download/36/40/230-1.

Torres, W. H. (2002). BIOLOGÍA DE LAS ESPECIES DE OXÍGENO REACTIVAS. *Instituto de Fisiología Celular, Universidad Nacional Autónoma de México*.

University of Waikato. (2016). *Data Mining Software in Java*. Obtenido de http://www.cs.waikato.ac.nz/ml/weka/

Valcárcel, V. (2004). DATA MININGYEL DESCUBRIMIENTO DEL CONOCIMIENTO. *Revista de la Facultad de Ingeniería Industrial*, 83-86.

CONVOCATORIA

Por disposición de la Junta Académica de **Ingeniería de Sistemas y Telemática**, se convoca a los Miembros del Tribunal Examinador, a la sustentación del Protocolo del Trabajo de Titulación: **"EVALUACIÓN DE LAS HERRAMIENTAS DE MINERÍA DE DATOS EN VARIABLES DE CONTAMINACIÓN ATMOSFÉRICA"**, presentado por la estudiante **Marcia Alexandra Matute Rivera**, previa a la obtención del grado de **Ingeniera en Sistemas y Telemática**, para el día **LUNES 15 DE MAYO DE 2017 A LAS 08h00.** La sustentación se realizará en el **laboratorio del IERSE.**

Cuenca, 10 de mayo de 2017

Dra. Jenny Ríos Cuello
Secretaria de la Facultad

Ing. Chester Seller Walden ✓

Ing. Marcos Orellana Cordero ✓

Dr. Francisco Salgado Arteaga ✓

mjmr/

Comunicado 12/05/2017.

Oficio Nro. 062-2017-DIST-UDA

Cuenca, 12 de mayo de 2017

Señor Ingeniero
Oswaldo Merchán Manzano
DECANO DE LA FACULTAD DE CIENCIAS DE LA ADMINISTRACIÓN
Presente.-

De nuestras consideraciones:

La Junta Académica de la Escuela de Ingeniería de Sistemas y Telemática, reunida el día 12 de mayo del 2017, recibió el proyecto de tesis titulado "Evaluación de las herramientas de minería de datos en variables de contaminación atmosférica", presentado por Marcia Alexandra Matute Rivera estudiante de la Escuela de Ingeniería de Sistemas y Telemática, y revisado por el Ing. Chester Sellers, previo a la obtención del título de Ingeniero de Sistemas y Telemática.

Por lo expuesto, y de conformidad con el Reglamento de Graduación de la Facultad, recomendamos como director y responsable de aplicar cualquier modificación al diseño del trabajo de graduación posterior al Ing. Chester Sellers y como miembros del Tribunal a Francisco Salgado Ph.D. e Ing. Marcos Orellana.

Atentamente,

Ing. Marcos Orellana Cordero
Cordinador Escuela de Ingeniería de Sistemas y Telemática
Universidad del Azuay

UNIVERSIDAD DEL AZUAY

ACTA

SUSTENTACIÓN DE PROTOCOLO/DENUNCIA DEL TRABAJO DE TITULACIÓN

1.1 **Nombre del estudiante: Marcia Alexandra Matute Rivera**
1.2 **Director sugerido**: Ing. Chester Seller Walden
1.3 **Codirector (opcional)**: _____
1.4 **Tribunal:** Ing. Marcos Orellana Cordero/ Dr. Francisco Salgado Arteaga
1.5 **Título propuesto: "EVALUACIÓN DE LAS HERRAMIENTAS DE MINERÍA DE DATOS EN VARIABLES DE CONTAMINACIÓN ATMOSFÉRICA"**
1.6 **Resolución**:

1.6.1 Aceptado sin modificaciones ✓

1.6.2 Aceptado con las siguientes modificaciones:

_____

_____

_____

1.6.3 Responsable de dar seguimiento a las modificaciones: Ing. Chester Seller Walden
1.6.4 No aceptado
- Justificación:

_____
_____
_____

Tribunal

Ing. Chester Seller Walden

Ing. Marcos Orellana Cordero

Dr. Francisco Salgado Arteaga

Srta. Marcia Alexandra Matute Rivera

Dra. Jenny Ríos Coello
Secretario de Facultad

Fecha de sustentación: día **LUNES 15 DE MAYO DE 2017 A LAS 08h00**

UNIVERSIDAD DEL AZUAY

RÚBRICA PARA LA EVALUACIÓN DEL PROTOCOLO DE TRABAJO DE TITULACIÓN

1.1 Nombre del estudiante: Marcia Alexandra Matute Rivera
1.2 Director sugerido: Ing. Chester Seller Walden
1.3 Codirector (opcional):
1.4 Título propuesto: "EVALUACIÓN DE LAS HERRAMIENTAS DE MINERÍA DE DATOS EN VARIABLES DE CONTAMINACIÓN ATMOSFÉRICA"
1.5 Revisores (tribunal): Ing. Marcos Orellana Cordero/ Dr. Francisco Salgado Arteaga

1.6 Recomendaciones generales de la revisión:

	Cumple totalmente	Cumple parcialmente	No cumple	Observaciones (*)
Línea de investigación				
1. ¿El contenido se enmarca en la línea de investigación seleccionada?	✓			
Título Propuesto				
2. ¿Es informativo?	✓			
3. ¿Es conciso?	✓			
Estado del arte				
4. ¿Identifica claramente el contexto histórico, científico, global y regional del tema del trabajo?	✓			
5. ¿Describe la teoría en la que se enmarca el trabajo	✓			
6. ¿Describe los trabajos relacionados más relevantes?	✓			
7. ¿Utiliza citas bibliográficas?	✓			
Problemática y/o pregunta de investigación				
8. ¿Presenta una descripción precisa y clara?	✓			
9. ¿Tiene relevancia profesional y social?	✓			
Hipótesis (opcional)				
10.¿Se expresa de forma clara?	✓			
11.¿Es factible de verificación?	✓			
Objetivo general				
12.¿Concuerda con el problema formulado?	✓			
13.¿Se encuentra redactado en tiempo verbal infinitivo?	✓			
Objetivos específicos				

14.¿Concuerdan con el objetivo general?	✓			
15.¿Son comprobables cualitativa o cuantitativamente?	✓			
Metodología				
16.¿Se encuentran disponibles los datos y materiales mencionados?	✓			
17.¿Las actividades se presentan siguiendo una secuencia lógica?	✓			
18.¿Las actividades permitirán la consecución de los objetivos específicos planteados?	✓			
19.¿Los datos, materiales y actividades mencionadas son adecuados para resolver el problema formulado?	✓			
Resultados esperados				
20.¿Son relevantes para resolver o contribuir con el problema formulado?	✓			
21.¿Concuerdan con los objetivos específicos?	✓			
22.¿Se detalla la forma de presentación de los resultados?	✓			
23.¿Los resultados esperados son consecuencia, en todos los casos, de las actividades mencionadas?	✓			
Supuestos y riesgos				
24.¿Se mencionan los supuestos y riesgos más relevantes?	✓			
25.¿Es conveniente llevar a cabo el trabajo dado los supuestos y riesgos mencionados?	✓			
Presupuesto				
26.¿El presupuesto es razonable?	✓			
27.¿Se consideran los rubros más relevantes?	✓			
Cronograma				
28.¿Los plazos para las actividades son realistas?	✓			
Referencias				
29.¿Se siguen las recomendaciones de normas internacionales para citar?	✓			
Expresión escrita				
30.¿La redacción es clara y fácilmente comprensible?	✓			
31.¿El texto se encuentra libre de faltas ortográficas?	✓			

(*) Breve justificación, explicación o recomendación.

- Opcional cuando cumple totalmente,
- Obligatorio cuando cumple parcialmente y NO cumple.

.....

.....

.....

Ing. Chester Seller Walden

Ing. Marcos Orellana Cordero

Dr. Francisco Salgado Arteaga

Cuenca, 12 de mayo de 2017

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

De mi consideración,

Yo, **Chester Andrew Sellers Walden** informo que he revisado el protocolo de trabajo de titulación previo a la obtención del título de Ingeniera en Sistemas y Telemática, denominado **"Evaluación de las herramientas de minería de datos en variables de contaminación atmosférica"**, realizado por el estudiante **Marcia Alexandra Matute Rivera**, con código estudiantil 46666, 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, suscribo.

Atentamente

MSc. Chester Sellers

UNIVERSIDAD DEL AZUAY

Cuenca, 12 de mayo de 2017

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

De mi/ nuestra consideración,

Estimado Señor Decano, yo **Marcia Alexandra Matute Rivera** con C.I. **0105232011**, código estudiantil 46666; estudiante de la Carrera de Sistemas y Telemática, solicito muy comedidamente a usted y por su intermedio al Consejo de Facultad, la aprobación del protocolo de trabajo de titulación con el tema **"Evaluación de las herramientas de minería de datos en variables de contaminación atmosférica"** previo a la obtención del título de Ingeniera en Sistemas y Telemática para lo cual adjunto la documentación respectiva.

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

Atentamente:

Marcia Matute

Estudiante de la Carrera de Sistemas y Telemática

Edición autorizada de 10.000 ejemplares
Del 788.501 al 798.500 Nº 0795814

UNIVERSIDAD DEL
AZUAY

DOCTORA JENNY RIOS COELLO, SECRETARIA DE LA FACULTAD DE CIENCIAS DE LA ADMINISTRACION DE LA UNIVERSIDAD DEL AZUAY

C E R T I F I C A:

Que, la señorita **MATUTE RIVERA MARCIA ALEXANDRA** con código **46666,** alumna de la escuela de **INGENIERIA DE SISTEMAS Y TELEMATICA,** tiene aprobado más del 80% de los créditos de su malla de estudios.

Que, a la señorita **MATUTE RIVERA MARCIA ALEXANDRA** le falta aprobar las siguientes asignaturas para finalizar sus estudios:

INGENIERIA DE SOFTWARE II

CALIDAD DE SOFTWARE

PROYECTOS TELEMATICOS

SISTEMA DE INFORMACION GERENCIAL

PRODUCCION II

DISEÑO DE TRABAJO DE GRADUACION

Cuenca, 30 de marzo de 2017

Derecho No. **001-001-000155830**
mjmr.-

UNIVERSIDAD DEL
AZUAY
FACULTAD DE
ADMINISTRACION
SECRETARIA

Edición autorizada de 10.000 ejemplares
Del 788.501 al 798.500 Nº 0791604

UNIVERSIDAD DEL AZUAY

1. Datos generales

1.1 Nombre del estudiante: Marcia Alexandra Matute Rivera

1.1.1 Código: 46666

1.1.2 Contacto: 402418-0998263148-alexarivera2103@hotmail.com

1.2 Director sugerido: Chester Andrew Sellers Walden, MSc.

1.2.1 Contacto: 285237-0992636061-csellers@uazuay.edu.ec

1.3 Co-director sugerido:

1.3.1 Contacto:

1.4 Asesor metodológico: Ing. Francisco Salgado Arteaga

1.5 Aprobación: Junta académica

Consejo de Facultad.

1.6 Línea de Investigación de la carrera: Calidad del Aire

1.6.1 Código UNESCO: 1203 Informática de computadores

1203.17 Informática.

1.6.2 Tipo de trabajo: Proyecto de Investigación Aplicada.

1.7 Área de estudio: Innovación Tecnológica.

1.8 Título propuesto: Evaluación de las herramientas de minería de datos en variables de contaminación atmosférica.

1.9 Estado del proyecto: Investigación complementaria al proyecto "Sistema de monitoreo en tiempo real de calidad del aire en la ciudad de Cuenca". Se evaluarán herramientas de minería de datos que ayuden al análisis de las variables contaminantes atmosféricas, para el aporte científico de información derivada del sistema de monitoreo.

Edición autorizada de 10.000 ejemplares
Del 788.501 al 798.500
Nº 0797234

2. Contenido

2.1 Motivación de la investigación:

Esta investigación pretende analizar diferente herramientas de minería de datos que proveerán mejores resultados al análisis de los datos obtenidos por la estación de monitoreo automático de la calidad del aire del GAD municipal de Cuenca. Se procura analizar las mejores herramientas de minería de datos en variables de contaminación atmosférica. Con toda la información recopilada, se implementará nuevos algoritmos en dichas herramientas para obtener datos y realizar consultas interactivas respecto a los contaminantes del aire y su afectación sobre los habitantes en la ciudad de Cuenca.

La selección de este tema se debe primordialmente a la necesidad que existe de obtener y acceder a información sobre el índice de calidad de aire en la ciudad de Cuenca, con la finalidad de aplicar técnicas de minería de datos en las variables de la calidad del aire que ayuden a determinar patrones de comportamiento en la contaminación atmosférica.

2.2 Problemática:

La contaminación del aire tiene incidencia directa en la salud de las personas, por ello es necesario el desarrollo de mecanismos capaces de prevenir y mitigar sus impactos. Consciente de la importancia que genera este ámbito en la calidad de vida de los habitantes, el IERSE ha desarrollado proyectos que califican la calidad del aire en base a los datos generados por la estación base implementada por el Municipio Cuenca. Los proyectos hasta ahora desarrollados han centrado sus esfuerzos en la captura, tratamiento y visualización de las variables monitoreadas por la estación; al punto de generar una escala que califica el aire que se respira en la ciudad de Cuenca de acuerdo a las normas nacionales (TULSMA) y normas internacionales (EPA).

El problema radica en que se necesita de herramientas de minería de datos aplicadas a variables de contaminación atmosférica que proporcionen mejores resultados que ayuden a descubrir patrones de comportamiento que generarán el conocimiento suficiente y necesario para que puedan emitir alertas tempranas y se tomen decisiones por parte de los organismos de regulación y control.

2.3 Pregunta de investigación:

¿Qué técnicas de minería de datos se aplican al problema de la contaminación atmosférica?
¿Cómo están organizados los datos generados por la estación de monitoreo?
¿Cuáles son los patrones de comportamiento de las variables ambientales?
¿Cuál es la herramienta apropiada de minería de datos aplicable a variables de contaminación atmosférica?

2.3 Resumen:

Este trabajo se orientará al análisis de las herramientas de minería de datos más apropiadas para el análisis sobre datos recopilados de manera sistemática por la estación de monitoreo continuo de la calidad del aire del municipio de Cuenca. Se analizará el comportamiento de las 5 variables ambientales recogidas en la ciudad de Cuenca por dicho sistema. Las variables de estudio son los principales generadores de la contaminación del aire: Ozono (O3), Monóxido de Carbono (CO) Dióxido de Azufre (SO2), Dióxido de Nitrógeno (NO2) y Material Particulado 2,5um (PM2.5). Para descubrir los patrones de comportamiento es necesario establecer correlaciones entre las variables de estudio según una cronología y frecuencia determinada y la influencia de otras variables como son las meteorológicas. Para la implementación del proyecto será necesario experimentar con varias herramientas de DM, lo que determinará los mejores resultados que permitan elaborar indicadores de gestión y predicción.

2.4 Indagación exploratoria y base conceptual:

Con el constante desarrollo económico y social de las grandes urbes se presentan incrementos en los niveles de contaminación debido a mayor demanda de energía y consumo precipitado de combustible fósil. El deterioro de la calidad de aire tiene relación directamente con problemas de salud de la ciudadanía y esta se ha convertido en un problema de particular interés para los administradores de grandes ciudades y de los ciudadanos en general. A pesar de los esfuerzos por apalear los efectos nocivos de la contaminación, los resultados son cada vez más evidentes en la salud de las personas y en el medio ambiente. Como le enuncia Gaitán, Cancino, & Behrentz en su artículo "Análisis de la calidad del aire en Bogotá" : el crecimiento se ve afectado por una mayor demanda de energía así como un acelerado consumo de combustibles fósiles (Gaitán, Cancino, & Behrentz, 2007). Al ser este un problema social; se han realizado varias iniciativas por evidenciar los fenómenos de la contaminación, en particular de la contaminación atmosférica.

Así como las grandes urbes han conseguido la implementación de redes de monitoreo y técnicas de análisis de datos, la Universidad del Azuay, a través de su centro de investigación IERSE "Instituto de Estudios de Régimen Seccional del Ecuador", impulsó el proyecto de calidad del aire por medio del GAD[1] municipal de la ciudad de Cuenca; es así que en la ciudad de Cuenca mediante el diagnóstico del sistema de monitoreo continuo de calidad del aire realizado por el Equipo técnico EMOV[2] Cuenca, de acuerdo a los datos recopilados mediante el monitoreo atmosférico, las concentraciones de las partículas totales en suspensión están por encima de los niveles de las normas permisibles y los niveles de los gases en el aire, tales como dióxidos de azufre, óxidos de nitrógeno y monóxido de carbono, aunque están por debajo de las normas de la calidad del aire, tienen una tendencia creciente y en algunos casos se acercan o rebasan los valores permisibles. (Dominguez, Torres, Ortiz, Prijodko, & Korc, 2011). Mediante aplicación de los proyectos y programas en la evaluación de la situación actual del recurso aire, fue posible identificar los problemas provocados por la contaminación de carácter antropogénico en la ciudad de Cuenca(Sellers, 2012), esto para priorizar las líneas de acción y mitigar el deterioro de la calidad atmosférica. (Díaz, 2015)

[1] GAD (Gobierno Autónomo Descentralizado).

[2] EMOV (Empresa Municipal de Obras Viarias).

La evaluación de herramientas de minería de datos, al igual que en el caso de medición de la contaminación del aire en Cuenca necesita de un nivel de análisis temporal y agregado, es decir, que a través de técnicas y mecanismos de tratamiento de la información, se tracen patrones de comportamiento de los datos generados por la estación base. Para encontrar estos patrones se adoptará técnicas de minería de datos, es preciso convertir grandes volúmenes de datos generados por la estación de monitoreo continuo en conocimiento y sabiduría que sea útil en la toma de decisiones.

Según estudios existentes, los datos de las estaciones tienen una fuerte correlación entre la calidad del aire y la aplicación de técnicas de minería de datos; se afirma que se puede construir una red que etiquete la calidad del aire y las características observadas a través de múltiples fuentes (Zheng, Liu, & Hsieh, 2013). La minería de datos al ser una herramienta fundamental para el análisis y descubrimiento del conocimiento a partir de dichos datos, admite que es posible estructurar patrones de comportamiento, clasificación y predicción de valores y asociación de atributos entre los agentes que forman parte del índice de calidad del aire (Díaz & Suárez, 2009), y así dotar de las nociones básicas para una toma de decisión eficaz, efectiva y oportuna. El manejo de esta información permitirá usar el conocimiento recopilado sobre la materia y el reconocimiento de la naturaleza secuencial en la resolución de un problema (Snee, 2015).

2.5 Justificación.

El Instituto de Estudios de Régimen Seccional del Ecuador (IERSE) ha establecido un índice de calidad del aire, el cual lo ha publicado a través de una web personalizada para el efecto. Es parte esencial que en esta fase del proyecto se produzca lo que se denomina como el descubrimiento de conocimiento de la base de datos (KDD, Knowledge Discovery in Databases); este proceso tiene que ver con la aplicación de técnicas de minería de datos (DM, Data Mining) para el análisis científico de la información recolectada. Por lo que, este trabajo se orienta al análisis de las herramientas de minería de datos más apropiadas para estudios sobre datos recopilados por la estación de monitoreo. Se analizará el comportamiento de las variables ambientales recogidas utilizando técnicas de minería de datos para el análisis de estas variables. Las herramientas deben aplicar algoritmos y funciones matemáticas para descubrir patrones de comportamiento y relaciones entre los datos, lo que significa evaluar las herramientas tomando en cuenta factores, como: gama de algoritmos disponibles, usabilidad, seguridad, viabilidad, velocidad, entre otras; la importancia de este trabajo radica en que permitirá descubrir patrones de comportamiento que derivan en conocimiento para realizar alertas tempranas que coadyuven a la toma de decisiones por parte de organismos de regulación municipal.

2.6 Objetivo general:

Evaluar herramientas y técnicas de minería de datos aplicables a los agentes de contaminación atmosférica obtenidos de la estación de monitoreo continuo en la ciudad de Cuenca, determinando patrones de comportamiento entre las variables estudiadas.

2.7 Objetivos específicos:

- Conocer los proyectos que ha emprendido el IERSE y otros autores en cuanto a la captura, proceso y visualización de la calidad de aire en la ciudad de Cuenca.
- Elaborar un marco teórico sobre las técnicas de minería de datos aplicables al problema.
- Evaluar la capacidad de las herramientas de minería de datos para el análisis de la información recolectada.
- Determinar patrones de comportamiento entre las variables que influyen en la calidad del aire.
- Establecer la mejor herramienta de minería de datos aplicable al problema.

2.8 Metodología:

Este proyecto tendrá una metodología orientada a la investigación, en primera instancia será necesario la revisión de los proyectos desarrollados por el IERSE dentro del campo de la contaminación atmosférica; lo que involucra una revisión sistemática de literatura de estudios similares en contraste con los estudios desarrollados; esto incluye el levantamiento teórico necesario en conjunción con la indagación de trabajos desarrollados en otros escenarios, pero que persiguen un mismo objetivo. Luego de establecer las pautas teóricas requeridas para el proyecto, se procederá con un levantamiento de los datos que sirven de materia prima en la aplicación de las técnicas de minería de datos. Las técnicas de minería de datos se particularizan por la ejecución de algoritmos para un conjunto definido de datos, lo que se conoce científicamente como un conjunto de datos (dataset). Establecer la herramienta adecuada que se adhiera a la forma del problema requiere una configuración de parámetros, así como la medición de la efectividad en cada aplicación.

2.9 Alcances y resultados esperados:

Objetivo	Resultado
Conocer los proyectos que ha emprendido el IERSE y otros autores en cuanto a la captura, proceso y visualización de la calidad de aire en la ciudad de Cuenca	Estado del arte sobre los proyectos realizados en el IERSE y en contraste con otros escenarios con estudios similares
Elaborar un marco teórico sobre las técnicas de minería de datos aplicables al problema	Marco teórico sobre las técnicas de minería de datos y su aplicación en problemas de contaminación ambiental.
	Artículo de tipo revisión sistemática de literatura.

Evaluar la capacidad de las herramientas de minería de datos para el análisis de la información recolectada.	Matriz de evaluación de las herramientas que aplican los algoritmos de minería de datos a contaminantes atmosféricos. Trabajo de graduación pregrado.
Determinar patrones de comportamiento entre las variables que influyen en la calidad del aire	Matriz de relación entre las variables de contaminación atmosférica
Establecer la mejor herramienta de minería de datos aplicable al problema.	Cuadro comparativo entre Herramientas de Minería de Texto

2.10 Supuestos y riesgos:

Supuestos	Riesgos	Soluciones
Existe Información sobre las variables de contaminación a estudiarse en la actualidad	No existe información sobre las variables de contaminación a estudiarse en la actualidad	Usar datos existentes basada en otros estudios.
Se cuenta con las herramientas necesarias para realizar la investigación	No se cuenta con las herramientas necesarias para realizar la investigación	Realizar un listado, previo al desarrollo del proyecto, con las herramientas necesarias y asegurarse de su disponibilidad a lo largo de la investigación
La o las Herramientas elegidos satisfacen las necesidades del usuario.	Las herramientas elegidas no satisfacen las necesidades del usuario.	Efectuar una recolección de requerimientos minuciosa
La herramienta implementada de minería de datos en variables de contaminación atmosférica, otorga mejores resultados y ayuda en el análisis de la toma de decisiones.	Los resultados que otorga la herramienta no representan ninguna mejora en la toma de decisiones sobre las variables de contaminación del aire en la ciudad de Cuenca.	Efectuar una recolección de requerimientos minuciosa, mantener reuniones con el usuario, presentar avances y realizar pruebas

2.11 Presupuesto:

No requiere un presupuesto.

UNIVERSIDAD DEL AZUAY

2.12 Financiamiento:

No requiere financiamiento.

2.13 Esquema tentativo: Incluye el listado de temas y subtemas que nos guiarán en el desarrollo del trabajo. Este esquema podrá ser revisado durante el desarrollo del trabajo. Debe estar en relación directa con los objetivos específicos.

1. INTRODUCCIÓN

1.1 Antecedentes

1.2 Planteamiento del Problema

1.3 Justificación

1.4 Definición de Objetivos

2. CAPÍTULO 1.

2.1 Introducción

2.2 ¿Qué es Minería de Datos?

2.3 ¿Qué son las Herramientas de minería de Datos?

2.4 Software disponible para aplicar técnicas de minería de datos.

2.5 Algoritmos y técnicas de minería de datos aplicadas a datos obtenidos por la red de monitoreo de la ciudad de Cuenca.

3. CAPÍTULO 2

3.1 Evaluar la capacidad de las herramientas de minería de datos para el análisis de la información recolectada.

3.2 Determinar patrones de comportamiento entre las variables que influyen en la calidad del aire.

4. CAPÍTULO 3 – RESULTADOS.

4.1 Matriz de evaluación de las herramientas que aplican los algoritmos de minería de datos a contaminantes atmosféricos.

4.2 Cuadro comparativo entre Herramientas de Minería de Datos

5. CONCLUSIONES.

6. RECOMENDACIONES.

7. REFERENCIAS BIBLIOGRÁFICAS

Edición autorizada de 10.000 ejemplares
Del 788.501 al 798.500
Nº 0797250

2.14 Cronograma:

OBJETIVOS	ACTIVIDADES	MAYO (2017)					JUNIO (2017)				JULIO (2017)					AGOSTO (2017)				SEPTIEMBRE (2017)			
		1	8	15	22	29	5	12	19	26	3	10	17	24	3	7	14	21	28	4	11	18	25
	Aplicación de la minería de datos en el análisis de contaminantes atmosféricos																						
Conocer los proyectos que ha emprendido el IERSE y otros autores en cuanto a la captura, proceso y visualización de la calidad de aire en la ciudad de Cuenca	Revisión de bibliografía específica de los proyectos emprendidos por el IERSE																						
	Revisión de bibliografía específica de proyectos realizados en el campo de la contaminación atmosférica con técnicas de minería de datos																						
	Revisión de los registros generados por la estación de monitoreo																						
Elaborar un marco teórico sobre las técnicas de minería de datos aplicables al problema	Contrastar los estudios realizados en el área de contaminación atmosférica con las técnicas de minería de datos																						
	Revisión de bibliografía sobre la minería de datos y sus técnicas de aplicación																						
Evaluar la capacidad de las herramientas de minería de datos para el análisis de la información recolectada.	Revisión de las herramientas disponibles y su oportunidad en el proyecto																						
	Registro de los experimentos realizados con los factores de precisión y porcentaje de error																						
Determinar patrones de comportamiento entre las variables que influyen en la calidad del aire	Análisis de la correlación de las variables que influyen en la contaminación atmosférica																						
Establecer la mejor herramienta de minería de datos aplicable al problema.	Matriz de evaluación de las herramientas que aplican los algoritmos de minería de datos a contaminantes atmosféricos. Cuadro comparativo entre Herramientas de Minería de Datos.																						

2.15 Referencias:

UNIVERSIDAD DEL AZUAY

Aguilar, A. (2015). Informe de calidad de aire Cuenca. Cuenca.

Díaz, C. A. (2015). Aplicación de la Herramienta Informática R.

Dominguez, J., Torres, Y., Ortiz, E., Prijodko, V., & Korc, M. (2011). Diagnóstico del sistema de monitoreo de la calidad de aire. Cuenca.

Gaitán, M., Cancino, J., & Behrentz, E. (2007). Análisis del estado de la calidad del aire en Bogotá. Revista de Ingeniería Universidad de Los Andes, (26), 81–92. https://doi.org/10.3989/scimar.2007.71n3485

Li, S. T., & Shue, L. Y. (2004). Data mining to aid policy making in air pollution management. Expert Systems with Applications, 27(3), 331–340. https://doi.org/10.1016/j.eswa.2004.05.015

Sakarde, M. S., Choudhari, M. M., & Gode, M. S. (2014). Implementation of WSN Based Air Pollution Monitoring System using Data Mining Technique. International Journal of Computer Science and Mobile Computing, 3(6), 790–795.

Sellers, C. (2012). MONITOREO DE LA CIUDAD DE CUENCA , UTILIZANDO SERVICIOS.

Snee, R. D. (2015). A Practical Approach to Data Mining: I Have All These Data; Now What Should I Do? Quality Engineering, 27(4), 477–487. https://doi.org/10.1080/08982112.2015.1065322

Zheng, Y., Liu, F., & Hsieh, H. (2013). U-Air: when urban air quality inference meets big data. Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '13, 1436–1444. https://doi.org/10.1145/2487575.2488188

Suárez, Y. R., & Amador, A. D. (2009). Herramientas de minería de datos. Revista Cubana de Ciencias Informáticas, 3(3-4).

2.16 Firma de responsabilidad (Estudiante):

Marcia Alexandra Matute Rivera.

2.17Firma de responsabilidad:

Ing. Chester Andrew Sellers. **Ing. Francisco Salgado**

2.18 Fecha de Entrega.

15 de Mayo del 2017