UNIVERSIDAD DEL AZUAY

FACULTAD DE CIENCIAS DE LA ADMINISTRACIÓN

ESCUELA DE INGENIERÍA DE SISTEMAS Y TELEMÁTICA

IMPACTO DE LA APLICACIÓN DE ALGORITMOS DE MINERÍA DE DATOS EN VARIABLES DE CONTAMINACIÓN DEL AIRE

TRABAJO DE TITULACIÓN PREVIO A LA OBTENCIÓN DEL TÍTULO DE

INGENIERO DE SISTEMAS Y TELEMÁTICA

AUTOR: JOHN JAVIER ORTEGA GUAMAN

DIRECTOR: ING. MARCOS PATRICIO ORELLANA CORDERO.

CUENCA, ECUADOR

2018

Dedicatoria

Este trabajo está dedicado a mi familia, principalmente a mis hermanos y mi tía Rosi, quienes, con sus consejos siempre me han incentivado y apoyado en el ámbito académico y desarrollo profesional.

Agradecimientos

Quiero agradecer a toda mi familia por el apoyo que me han brindado durante todos estos años, principalmente a mi abuela Alejandrina, quien ha sido más que una madre hasta el último de sus días y me ha dejado los regalos más grandes de esta vida: humildad y respeto. A mis tías, Rosa, Isabel, Celia, Luz y Susana, por su ayuda y confianza depositada en mi a pesar de tantos desaciertos que he cometido; a mi madre Gloria que, aunque físicamente no ha estado conmigo, ha sido un pilar fundamental en mi desarrollo personal y académico; a mis hermanos Oscar y Erika, quienes han estado a mi lado de forma incondicional en todo momento. Agradecer también a mi director de tesis, Ing. Marcos Orellana Cordero, por haberme brindado su tiempo y sus conocimientos para el correcto desarrollo de este proyecto.

Índice

Índice de Ilustraciones

Índice de Tablas

Índice de Ecuaciones

Resumen

Este trabajo se enfocó en el descubrimiento del mejor algoritmo de Minería de Datos para el análisis de las variables de contaminación del aire, dichos datos fueron recopilados de manera sistemática por una estación de monitoreo de la calidad del aire. Se evaluó el comportamiento y la eficiencia de los algoritmos para el análisis de 5 variables ambientales recogidas de la ciudad por un sistema de monitoreo. 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). Luego de obtener los resultados preliminares de todos los algoritmos evaluados, se determinó que k-means es el mejor algoritmo. Utilizando este algoritmo se realizó un análisis exhaustivo con el cual se identificó distintos patrones de comportamiento entre varios contaminantes en diferentes etapas del día. Además, se obtuvo que O3 es el contamínate que influye en mayor medida sobre el resto de contaminantes.

Abstract

This work focused on discovering the best data mining algorithm for the analysis of air pollution variables. The behavior and efficiency of the algorithms for the analysis of five environmental variables of the city were evaluated, these were collected by a monitoring system. The study variables were the main pollution generators: Ozone, Carbon Monoxide, Sulfur Dioxide, Nitrogen Dioxide and Particulate Material 2,5um. After evaluating all the algorithms, it was determined that k-means was the best algorithm. Different behavior patterns were identified among several pollutants at different stages of the day and it was found that ozone was the most influential pollutant.

UNIVERSIDAD DEL AZUAY
Dpto. Idiomas

Ing. Paúl Arpi
Traductor

Introducción.

En el presente trabajo se realiza el análisis de 5 algoritmos no descriptivos de minería de datos: K-means, X-means, Cobweb, Expectation Maximization y DBSCAN; en variables de contaminación del aire, para determinar cuál es el mejor para procesar este tipo de datos y determinar patrones de comportamiento, a través de la herramienta RapidMiner. Antes de aplicar cualquier proceso de minería de datos es importante asegurarse de la integridad de los mismos para garantizar la confiabilidad de los resultados por ello, se describe el procesamiento de los datos que han sido recolectados a través de una estación de monitoreo que registra los valores de los contaminantes cada segundo y los almacena en una base de datos, misma que debe ser reestructurada para obtener el *dataset* deseado y aplicar los métodos de minería. Luego se evalúa cada uno de los algoritmos, y una vez obtenido el mejor algoritmo en variables de contaminación de aire, se utiliza dicho algoritmo para determinar factores de correlación entre los contaminantes.

1.1. Objetivos

1.1.1. Objetivo general

Identificar los algoritmos relevantes que determinen las relaciones entre las variables de contaminación atmosférica.

1.1.2. 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.
- Identificar los algoritmos relevantes que determinen las relaciones entre las variables de contaminación atmosférica.
- Analizar los resultados obtenidos de la aplicación de los algoritmos seleccionados.

1.2. Justificación

En la actualidad existe un proyecto en la ciudad el cual se encarga de medir y publicar los niveles de contaminación a través de una página web. El presente trabajo tiene como propósito mejorar dicho proyecto implementando algoritmos de minería de datos, ya que la estación de monitoreo envía los datos medidos del aire cada segundo, lo que genera grandes volúmenes de información que deben ser procesados para extraer conocimiento. Para esto, es necesario analizar un conjunto de algoritmos de minearía de datos y determinar cuál es el algoritmo idóneo para el trabajo.

1.3. Alcance y resultados esperados

Culminado este proyecto se pretende determinar cuál o cuáles son los algoritmos óptimos para el análisis de contaminantes en el aire, e identificar los patrones de comportamiento de las variables.

Se debe obtener un informe técnico, matrices y documentos de evaluación resultantes de la evaluación de algoritmos.

2. Contaminación del Aire.

La contaminación del aire ha sido calificada por la OMS (Organización Mundial de la Salud) como uno de los grandes causantes de cáncer en los seres humanos, además, en

varios estudios se ha demostrado que la contaminación ambiental es la principal causa de muertes por problemas respiratorios y cardiovasculares (Fernandez-Camacho, y otros, 2015).

El constante crecimiento del tráfico vehicular y la industrialización son unas de las principales fuentes de contaminación de las ciudades. Todos los vehículos de combustión interna generan gases contaminantes, a esto se le suma los gases que son despedidos por las industrias que se encuentran en las cercanías de las zonas pobladas, además, la falta o el incumpliendo de la normativa para regular y controlar la composición de los combustibles incrementa considerablemente el impacto que tienen sobre el medio ambiente. (Fernandez-Camacho, y otros, 2015).

2.1. Variables de contaminación del aire.

Dentro del gran número de elementos que afectan la calidad del medio ambiente se encuentran cinco principales que aportan a la contaminación del aire. A continuación, se presenta una breve descripción de cada uno de ellos:

2.1.1. Material Particulado 2,5um (PM 2,5).

Son partículas microscópicas que se originan de fuentes primarias o pueden derivarse de la condensación de gases y se desplazan largas distancias a través del aire. Ingresan fácilmente al cuerpo humano a través de las vías respiratorias y pueden llegar al torrente sanguíneo (EMOV-EP, 2015). Algunas de las fuentes de este contaminante son las centrales eléctricas, escapes de los vehículos e incendios forestales. (Oficina-de-calidad-del-aire-y-radiacion, 2015).

2.1.2. Ozono (O3).

El ozono es un gas que dependiendo donde se encuentra puede ser beneficioso o perjudicial. Cuando este gas se encuentra en la atmosfera nos protege de los rayos ultra violetas emitidos por el sol, pero cuando este se encuentra en el nivel del suelo es un problema ya que afecta directamente al sistema respiratorio de las personas (Oficina-de-Calidad-del-Aire-y-Radiación, 2015).

2.1.3. Dióxido de Nitrógeno (NO2)

Este contaminante se forma a partir de emisiones a nivel del suelo relacionadas con la quema de combustibles fósiles para los vehículos, plantas industriales, centrales eléctricas, etc. Aporta a la formación de ozono al nivel de suelo y está relacionado con varias afecciones al sistema respiratorio (Office-of-air-and-radiation, Air Quality Guide for Nitrogen Dioxide, 2011).

2.1.4. Monóxido de Carbono(CO).

Es un gas inodoro e incoloro que se forma cuando el carbono en el combustible de los vehículos no se quema completamente. Los niveles de monóxido de carbono se elevan en las temporadas de frío ya que las temperaturas bajas inhiben la combustión (Office-of-air-and-radiation, Air Quality Index A guide to air quality and your health, 2003).

2.1.5. Dióxido de azufre (SO2).

Es un gas incoloro y reactivo que se produce cuando se queman combustibles que contienen azufre como el petróleo y el carbón, materiales utilizados en grandes cantidades por las industrias y las centrales eléctricas (Office-of-air-and-radiation, Air Quality Index A guide to air quality and your health, 2003).

2.2. Variables meteorológicas.

Las variables meteorológicas son fenómenos que se producen en la atmósfera dentro de las cuales se encuentran: la temperatura y la presión barométrica, que deben ser mencionadas ya que afectan a los valores de medición de los contaminantes. La temperatura es una medida que refleja la velocidad de movimiento o agitación de las moléculas en el aire, mientras que la presión barométrica (también conocida como presión atmosférica) no es otra cosa más que la fuerza que ejerce una columna de aire sobre la superficie terrestre (Crowe, 2011). Estos dos factores alteran los valores de los contamines atmosféricos y deben ser corregidos según como lo establece la Norma Técnica Ambiental Ecuatoriana en el artículo 4.1.2.3 del anexo 4 – Norma de Calidad de Aire Ambiente.

2.3. Proyectos relacionados

Tal y como se mencionó con anterioridad, uno de los propósitos de este trabajo es mejorar los sistemas de monitoreo de calidad del aire en las grandes ciudades. Además, este trabajo será realizado tomando como guía principal la forma en la que el proyecto de control y monitoreo de calidad del Aire de la ciudad de Cuenca realiza la recolección y tratamiento de los datos, por lo que es necesario conocer acerca del mismo.

2.3.1. Proyecto del IERSE para monitorear la calidad del aire.

El Instituto de Régimen Seccional del Ecuador – Adscrito al decanato de investigaciones de la universidad del Azuay (IERSE) cuenta con un proyecto el cual se encarga de gestionar y publicar la información acerca de los niveles de contaminación del aire a través de un índice general de la calidad del aire (IGCA). El IGCA se genera en base a las normativas Texto Unificado de Legislación Ambiental Secundaria (TULAS) y

Environmental Protection Agency (EPA). La información es publicada a través de una página WEB y de varios medios de fácil acceso para los interesados en conocer dicha información. Se hace uso de los estándares de *Sensor Observation Service* (SOS) del *Open Geospatial Consortium* (OGC) para el acceso estructurado a la información registrada por sensores (Sellers Walden, 2012).

A lo largo de los años, en dicho proyecto se ha recolectado una gran cantidad de información con la cual se busca determinar patrones de comportamiento de los contaminantes a través del uso de técnicas de minería de datos.

2.4. Minería de datos.

Si bien este trabajo se enfoca en el análisis de algoritmos de minería de datos, es necesario primero entender que es la minería y cuáles son los distintos tipos de algoritmos que existen.

La minería de datos es el proceso de descubrir una estructura en los datos a través del uso de algoritmos diseñados específicamente para trabajar con grandes volúmenes de información. La estructura puede presentar formas tan complejas o simples, dependiendo del análisis que se realice y de los datos utilizados. Es evidente que con el paso de los años la información que se genera alrededor del mundo crece exponencialmente como consecuencia de la digitalización y la globalización. La corporación internacional de datos (IDC por sus siglas en inglés) predice que para el año 2020 la toda la información digital llegará a los cuarenta mil *exabytes* (Roiger, 2017). En la actualidad la información que se genera diariamente supera los cinco *exabytes* (Bifet, 2013), por lo tanto, se presenta escenarios en donde es indispensable el uso de la minería de datos.

Dentro de la minería de datos existen dos principales tareas, descriptivas y predictivas. Las tareas descriptivas se enfocan en detectar patrones de comportamiento y las relaciones existentes sobre los atributos de los datos. Mientras que las tareas predictivas plantean un posible escenario futuro con un modelo entrenado en base a los datos históricos (Riquelme, Ruiz, & Gilbert, 2006).

2.4.1. Técnicas de minería de datos.

Trabajar con grandes volúmenes de información supone un reto para cualquiera que desea analizar grandes lotes de información, debido a que no pueden ser tratados con los métodos tradicionales. Por fortuna, hoy en día se cuenta con una gran variedad de métodos que han sido desarrollados precisamente para lidiar con bases de datos tan grandes.

Según los autores Medina y Gómez (2014) algoritmos en la minería de datos se clasifican de la siguiente manera:

- Exploración: Utilizada para encontrar relaciones sistemáticas entre las variables cuando se tiene poco o nada de conocimiento sobre los resultados próximos. Este método solo funciona como una primera etapa para la predicción.
    - De asociación
    - Agrupamiento
- Estadísticos: Utilizados cuando se trata de encontrar una función matemática que mejor relacione los datos. Busca corregir datos faltantes en una muestra, clasificar o realizar procesos de predicción.
    - Regresión Lineal
    - Regresión Logística

    - Redes Bayesianas
    - BayesNaive
- Clasificación de Datos: Basado en el aprendizaje inductivo.
    - Árboles de Decisión
    - Reglas de decisión
- Aprendizaje Automático: Combina la inteligencia artificial con la estadística para establecer una noción general de inferencia.
    - Lógica de Inferencia Difusa
    - Redes Neuronales
- Serie Temporal: Sirve para generar predicciones de tiempo en valores continuos, basándose únicamente en los datos que se utilizaron para crear el modelo. (Medina & Gómez, 2014)

Previamente a la implementación de cualquier modelo es necesario organizar la información: antes, durante y después del análisis. La creación o selección de un modelo depende de los datos, por lo que datos desorganizados podrían ser motivo de la selección de un modelo erróneo. Los datos deben ser revisados constantemente para identificar cualquier anomalía durante el proceso y corregirlas a tiempo. Evaluar los resultados obtenidos y asegurarse de que las conclusiones posean cierta concordancia con el problema planteado (Snee, 2015).

2.4.2. Algoritmos utilizados en proyectos similares.

Como se mencionó en el punto anterior, existe una gran variedad de algoritmos de minería de datos. Algunos de estos algoritmos han sido implementados en proyectos relacionados con la contaminación, dichos casos pueden ayudar a la selección de los

modelos para el análisis por lo que a continuación se presenta un breve resumen de algunos de estos proyectos.

Se aplicó k-means y árboles de decisión para determinar la relación entre la presión del aire y la humedad con la contaminación del aire en la ciudad de Boushehr - Iran, tomando como referencia el nivel de polvo en el aire como contaminación.

k-means es uno de los métodos más simples no supervisados de clustering. La idea principal es separar un conjunto de datos definiendo k centroides, uno para cada clúster. Estos centroides deberán estar lo más alejado posible uno de otro. Luego cada dato es asignado al centroide más cercano. Cuando todos los datos se han asignado se vuele a calcular los centroides y a reasignar los datos a su centroide más cercano y se repite este proceso hasta que no existan modificaciones de los centroides.

Un árbol de decisión es uno de los algoritmos más populares en la minería de datos. Un árbol de decisión se construye en base a reglas de decisión y sirve como apoyo en la toma de decisiones, presentando las posibles consecuencias de cada opción, incluyendo costos y utilidades (Sahafizadeh & Ahmadi, 2009).

Colombia también presenta otro caso, en el cual se realizó un estudio sobre las variables meteorológicas en Manizales utilizando varios tipos de algoritmos a través de WEKA, una herramienta de minería de datos que incorpora un gran número de algoritmos y es de código abierto.

Dentro de los algoritmos de Agrupamiento se utilizó una implementación de *K-means* conocida como *SimpleK-means*. Como método de árbol de decisión se utilizó el algoritmo *REPTree, Bagging* y C4.5 implementado como el algoritmo J48 de WEKA.

LeastMedSq como un algoritmo de Regresión Lineal. El algoritmo con el que se obtuvo el mejor resultado fue Bagging (Duque, 2011).

España por su lado presenta también dos casos de la aplicación de algoritmos de minería de datos:

En el primero, el autor utiliza el clasificador *Fuzzy Lattice Reasoning*(FLR) para predecir los niveles de ozono en la ciudad de Valencia. Este método presenta dos variaciones, FLR con evaluación positiva lineal y una evaluación positiva con función sigmoidea. Comparando estos métodos con el método c4.5 sin podar, se obtiene una mejora de un 9% en la exactitud utilizando un *dataset* que ha sido pre procesado, y un 7% con un *dataset* crudo, es decir sin eliminar aquellos datos nulos o anómalos (Athanasiadis & Kaburlasos, 2006).

En el segundo, con los datos obtenidos por el centro de Andaluz del Medio Ambiente, se utilizó varios algoritmos con de minería de datos para analizar y medir los intercambios de dióxido de carbono y metano en el ecosistema, basándose en la clasificación de los datos. Se utilizaron los siguientes algoritmos: BayesNaive, Stacking, OneR y J48. De los resultados obtenidos, tanto BayesNaive, Stacking y J48 ofrecen resultados correctamente clasificados superiores a un 80%, siendo Stacking el algoritmo con el mejor resultado (González, 2013). La validación de los resultados fue realizada en base a los datos históricos.

En otro estudio se realizó el análisis para establecer la relación de PM2.5 y el tráfico en las ciudades, se utilizaron los métodos de reglas de asociación a través de Apriori, K-

means como método de clustering y J48 como clasificación. Los cuales muestran resultados similares (Du, 2016).

El continente asiático también ha sido escenario de varios estudios de minería de datos y contaminación. Swethal Raipure en la ciudad Pune implementa el algoritmo ID3 para analizar los contaminantes en el medio ambiente. ID3 es básicamente un algoritmo de árbol de decisión que será el encargado de generar predicciones sobre un área en particular. Una de las ventajas de éste método es que trabaja sin complicaciones sobre un conjunto de datos donde existan atributos perdidos (Raipure & Mehetre, 2015).

Y por último en Delhi, una de las mayores ciudades metropolitanas de la región sur de Asia en donde su nivel de contaminación se ha elevado considerablemente, tanto que en el año 2014 la OMS determino que Delhi es la ciudad más contaminada del mundo, en el año 2016 se realizó un estudio con el fin de determinar las tendencias de la contaminación y hacer predicciones. Para esto se utilizó el método de serie temporal con un *dataset* que involucraba las siguientes variables de contaminación: SO2, NO2, CO y O3. Esta información ha sido recolectada desde el año 2011 hasta el 2015 (Taneja, Sharma, Oberoi, & Navoria, 2016).

2.5. Metodología

La metodología utilizada en este proyecto es CRISP-DM. Pese a que existen varias metodologías, CRISP es ampliamente utilizada en los proyectos de minería de datos debido a su flexibilidad y facilidad de entendimiento como se explica a continuación.

Esta metodología está dividida en cuatro niveles de abstracción en los cuales las tareas se encuentran ordenadas jerárquicamente desde lo más general hasta lo más específico, organizando el desarrollo del proyecto en 6 fases como se observa en la figura 7, las

cuales no necesariamente son rígidas. En cada fase se plantean las tareas generales que a su vez se proyectan sobre tareas específicas. Por último, se describen las acciones que deben ser desarrolladas (O. Rodríguez, n.d.).

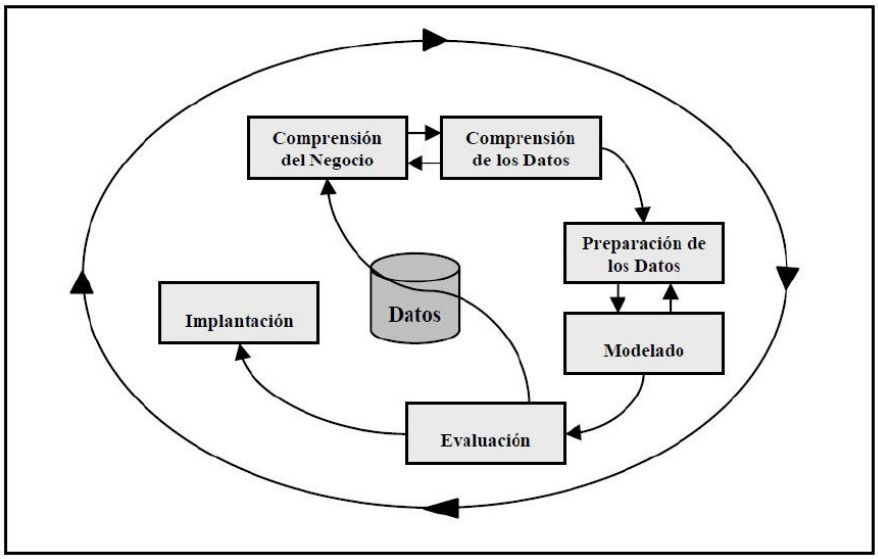


Ilustración 1: Metodología CRISP-DM (O. Rodríguez, n.d.)

2.5.1. Fases del modelo

A continuación, se presenta una breve descripción de cada una de las fases del modelo y las principales tareas dentro de cada fase, estas tareas al igual que las fases son flexibles y han de adaptarse a las necesidades del proyecto.

2.5.1.1. Compresión del negocio o problema.

Es necesario conocer el problema antes de buscar la solución. Comprender el problema ayuda a la recolección de los datos y a una mejor interpretación de los mismos, así mismo ayuda a la elección correcta de las herramientas de trabajo para obtener resultados confiables (O. Rodríguez, n.d.).

Principales Tareas:

- Determinar los objetivos del negocio.
- Valoración de la situación.
- Determinar los objetivos de la minería de datos.

2.5.1.2. Compresión de los datos.

Comprende la recolección inicial de datos, analizar, identificar y establecer la estructura y relaciones entre los mismos que permitan definir las primeras suposiciones (O. Rodríguez, n.d.).

Principales Tareas:

- Recolección de datos iniciales.
- Descripción de los datos.
- Exploración de datos.
- Verificación de la calidad de los datos.

2.5.1.3. Preparación de los datos.

En esta fase de preparación los datos son sometidos a procesos de selección, limpieza y/o transformación de formatos, quedando así listos para las distintas técnicas y herramientas a implementarse (O. Rodríguez, n.d.).

Principales Tareas:

- Selección de datos
- Limpieza de datos
- Estructura de los datos
- Integración de los datos

- Formateo de los datos

2.5.1.4. Modelado.

Aquí se seleccionan las técnicas de modelado más adecuadas dependiendo de los siguientes criterios:

- Apropiada al problema
- Disponer de datos adecuados
- Cumplir los requisitos del problema
- Tiempo adecuado para obtener un modelo
- Conocimiento de la técnica

Se procede con la generación del modelo dependiendo de las características y propiedades de los datos y el nivel de precisión deseado (O. Rodríguez, n.d.).

Principales Tareas:

- Selección de la técnica de modelado
- Generación del plan de prueba
- Construcción del modelo
- Evaluación del modelo

2.5.1.5. Evaluación.

Se evalúa el modelo en base a los criterios de éxito del problema. Es necesario tener en cuenta que la fiabilidad de los resultados obtenidos se aplica únicamente a los datos con los que se realizó las pruebas. Una buena práctica es revisar el proceso para identificar errores y repetir los pasos necesarios (O. Rodríguez, n.d.).

Principales Tareas:

- Evaluación de los resultados
- Proceso de revisión

2.5.1.6. Implementación.

Una vez el modelo haya sido construido y validado se procede con la implementación (O. Rodríguez, n.d.).

Principales Tareas:

- Plan de implementación
- Informe final

2.6. Conclusión

La elección del método para la minería de datos depende en gran medida de los datos con los que se va a trabajar, pero antes de entender los datos es necesario entender el negocio o problema para identificar los factores internos y externos que afectan al proyecto y determinar la fuente de los datos. La metodología CRISP proporciona las pautas necesarias para el correcto desarrollo de los proyectos de minería de datos y por su flexibilidad se adapta perfectamente a este proyecto. Además, luego de analizar varios casos del uso de minería de datos en variables de contaminación del aire, se puede observar como varían los resultados dependiendo tanto de las variables de contaminación involucradas, así como de la localidad en donde se realizó el estudio.

3. Recopilación y generación de datos

El proyecto de calidad del aire del IERSE alimenta su base de datos a través de una estación automática de monitoreo que registra los contaminantes atmosféricos antes mencionados. La estación de monitoreo también captura los datos de las siguientes

variables meteorológicas: precipitación, radiación solar global, velocidad y dirección del viento; temperatura y humedad relativa (EMOV-EP, 2015).

Ilustración 2: Estación de Monitoreo (EMOV-EP, 2015)

Las unidades de cada elemento se muestran en la tabla 1, estas medidas serán transformadas a microgramos por metro cúbico para unificar la base de datos y posteriormente normalizarlos ya que de otro modo sería muy difícil representar los resultados debido a la gran diferencia de rangos en los que se maneja cada contamínate.

Contaminante	Símbolo	Datalogger	Unidad
Ozono	O3	ppb	Partes por Billón
Monóxido de Carbono	CO	ppm	Partes por Millón
Dióxido de Nitrógeno	NO2	ppb	Partes por Billón
Dióxido de Azufre	SO2	ppb	Partes por Billón
Material Particulado 2,5 um	PM2.5	ug/m3	Microgramos por metro cubico

Tabla 1: Elementos y unidad de medida registrados (Sellers, 2013)

La información es almacenada cada segundo por lo que al término del día la base de datos obtiene 86400 mil registros nuevos. En una primera instancia los datos son capturados cada segundo y luego de un minuto se almacena los registros en servidor local de la EMOV en formato XLXS y RAW. La tabla 2 muestra la estructura que mantiene esta recopilación, donde el primer campo representa el nombre del contamínate, el segundo es el intervalo promedio, el tercer campo muestra la fecha de la recolección y los dos últimos muestran el valor redondeado y el valor exacto obtenido por el sensor respectivamente.

Parameter	Average Interval	Date	Value	Raw Value
OZONE	001m	01/08/2012 9:13	9.5125	9.5124998
OZONE	001m	01/08/2012 9:14	9.09375	9.09375
OZONE	001m	01/08/2012 9:15	9.2375	9.23750019
OZONE	001m	01/08/2012 9:16	10.1812	10.1812496
OZONE	001m	01/08/2012 9:17	10.4187	10.4187498
OZONE	001m	01/08/2012 9:18	10.4125	10.4125003
OZONE	001m	01/08/2012 9:19	10.7688	10.7687501
OZONE	001m	01/08/2012 9:20	10.5125	10.5124998
OZONE	001m	01/08/2012 9:21	10.9625	10.9624996
OZONE	001m	01/08/2012 9:22	11.8062	11.8062496
OZONE	001m	01/08/2012 9:23	11.4563	11.4562501
OZONE	001m	01/08/2012 9:24	10.875	10.875
OZONE	001m	01/08/2012 9:25	11.05	11.0500001
OZONE	001m	01/08/2012 9:26	11.1125	11.1125001
OZONE	001m	01/08/2012 9:27	9.925	9.92500019
OZONE	001m	01/08/2012 9:28	8.99375	8.99374961
OZONE	001m	01/08/2012 9:29	9.49375	9.49374961
OZONE	001m	01/08/2012 9:30	9.9625	9.96249961
OZONE	001m	01/08/2012 9:31	9.26875	9.26875019
OZONE	001m	01/08/2012 9:32	9.075	9.0749998

Tabla 2: Estructura los datos recolectados por la estación de monitoreo (Sellers, 2013)

Los datos registrados son ingresados en una geodatabase PostGIS utilizando instrucciones SQL(Structured Query Language), esta base de datos se encuentra instalada en los servidores de la universidad del Azuay y para conectarse se hace uso de una Red Privada Virtual (VPN) para garantizar la seguridad e integridad de los datos (Sellers, 2013).

En la base de datos los elementos son homogenizados a partes por millón (ppm) a excepción del Material Particulado el cual se mantendrá en microgramos por metro cúbico (ug/m3). Todos los datos serán redondeados a 3 decimales y se depura la base de datos para eliminar datos y registros erróneos como aquellos que se generan al momento de ajustar los equipos o a fallas eléctricas y que alteran de forma directa los resultados de cualquier análisis (Sellers Walden, 2012).

Luego se realiza una corrección a los datos, utilizando la presión barométrica y la temperatura de la localidad en donde se la realizó la recolección de los datos (Tulas, Libro VI). Para esto se hace uso de la siguiente formula:

$$Cc = Co * \frac{760 mmHg}{Pbl mmHg} * \frac{(273 + t°C)°K}{298°K}$$

Ecuación 1: Formula de Corrección de temperatura y presión atmosférica (Sellers, 2013)

Donde:

Cc: Concentración corregida

Co: Concentración Observada

Pbl: Presión barométrica local, en milímetros por mercurio

tC: Temperatura local, en grados centígrados.

4. Compresión de los datos

En esta sección, se analizará la estructura de los datos que han sido recolectados a través de la estación de monitoreo para entender la naturaleza de los mismos, y de esta manera, facilitar la elección de métodos de procesamiento.

La base de datos fue proporcionada por el IERSE, la cual contiene 215436 registros tomados de 18 días; cuyos valores cumplen con la restricción de no ser cero o muy cercanos a cero, además, se observa que los valores nulos son escasos. Para la selección de la muestra se utilizó un método no probabilístico, un muestreo intencional u opinático en el cual un experto usa su criterio para seleccionar datos que sean representativos para el caso de análisis (Lagares & Puerto, 2001).

time_stamp	procedure_id	feature_of_interest_id	phenomenon_id
9/16/2017 0:01	urn:ogc:object:feature:Sensor:sensor-2	foi_2001	urn:ogc:def:phenomenon:OGC:1.0.30:O3
9/16/2017 0:01	urn:ogc:object:feature:Sensor:sensor-3	foi_3001	urn:ogc:def:phenomenon:OGC:1.0.30:CO
9/16/2017 0:01	urn:ogc:object:feature:Sensor:sensor-4	foi_4001	urn:ogc:def:phenomenon:OGC:1.0.30:NO2
9/16/2017 0:01	urn:ogc:object:feature:Sensor:sensor-5	foi_5001	urn:ogc:def:phenomenon:OGC:1.0.30:SO2
9/16/2017 0:01	urn:ogc:object:feature:Sensor:sensor-6	foi_6001	urn:ogc:def:phenomenon:OGC:1.0.30:PM2_5
9/16/2017 0:02	urn:ogc:object:feature:Sensor:sensor-2	foi_2001	urn:ogc:def:phenomenon:OGC:1.0.30:O3
9/16/2017 0:02	urn:ogc:object:feature:Sensor:sensor-3	foi_3001	urn:ogc:def:phenomenon:OGC:1.0.30:CO
9/16/2017 0:02	urn:ogc:object:feature:Sensor:sensor-4	foi_4001	urn:ogc:def:phenomenon:OGC:1.0.30:NO2
9/16/2017 0:02	urn:ogc:object:feature:Sensor:sensor-5	foi_5001	urn:ogc:def:phenomenon:OGC:1.0.30:SO2
9/16/2017 0:02	urn:ogc:object:feature:Sensor:sensor-6	foi_6001	urn:ogc:def:phenomenon:OGC:1.0.30:PM2_5
9/16/2017 0:03	urn:ogc:object:feature:Sensor:sensor-2	foi_2001	urn:ogc:def:phenomenon:OGC:1.0.30:O3
9/16/2017 0:03	urn:ogc:object:feature:Sensor:sensor-3	foi_3001	urn:ogc:def:phenomenon:OGC:1.0.30:CO
9/16/2017 0:03	urn:ogc:object:feature:Sensor:sensor-4	foi_4001	urn:ogc:def:phenomenon:OGC:1.0.30:NO2
9/16/2017 0:03	urn:ogc:object:feature:Sensor:sensor-5	foi_5001	urn:ogc:def:phenomenon:OGC:1.0.30:SO2
9/16/2017 0:03	urn:ogc:object:feature:Sensor:sensor-6	foi_6001	urn:ogc:def:phenomenon:OGC:1.0.30:PM2_5
9/16/2017 0:04	urn:ogc:object:feature:Sensor:sensor-2	foi_2001	urn:ogc:def:phenomenon:OGC:1.0.30:O3
9/16/2017 0:04	urn:ogc:object:feature:Sensor:sensor-3	foi_3001	urn:ogc:def:phenomenon:OGC:1.0.30:CO
9/16/2017 0:04	urn:ogc:object:feature:Sensor:sensor-4	foi_4001	urn:ogc:def:phenomenon:OGC:1.0.30:NO2
9/16/2017 0:04	urn:ogc:object:feature:Sensor:sensor-5	foi_5001	urn:ogc:def:phenomenon:OGC:1.0.30:SO2
9/16/2017 0:04	urn:ogc:object:feature:Sensor:sensor-6	foi_6001	urn:ogc:def:phenomenon:OGC:1.0.30:PM2_5
9/16/2017 0:05	urn:ogc:object:feature:Sensor:sensor-2	foi_2001	urn:ogc:def:phenomenon:OGC:1.0.30:O3
9/16/2017 0:05	urn:ogc:object:feature:Sensor:sensor-3	foi_3001	urn:ogc:def:phenomenon:OGC:1.0.30:CO
9/16/2017 0:05	urn:ogc:object:feature:Sensor:sensor-4	foi_4001	urn:ogc:def:phenomenon:OGC:1.0.30:NO2
9/16/2017 0:05	urn:ogc:object:feature:Sensor:sensor-5	foi_5001	urn:ogc:def:phenomenon:OGC:1.0.30:SO2
9/16/2017 0:05	urn:ogc:object:feature:Sensor:sensor-6	foi_6001	urn:ogc:def:phenomenon:OGC:1.0.30:PM2_5

Tabla 3: Estructura de la Base de datos - Parte 1 (IERSE, 2017)

text_value	numeric_value	spatial_value	mime_type	observation_id	valor	valor_sin_factor	replica	id_mysql
	20.41530559			10293011	20.4153056	20.41530559	t	1.6E+07
	0.789981			10293009	0.789981	0.789981	t	1.6E+07
	23.48091318			10293010	23.4809132	23.48091318	t	1.6E+07
	3.01132514			10293013	3.01132514	3.01132514	t	1.6E+07
	7.2			10293012	7.2	7.2	t	1.6E+07
	23.08738383			10293016	23.0873838	23.08738383	t	1.6E+07
	0.789981			10293014	0.789981	0.789981	t	1.6E+07
	23.44518503			10293015	23.445185	23.44518503	t	1.6E+07
	3.003469509			10293018	3.00346951	3.003469509	t	1.6E+07
	9.7			10293017	9.7	9.7	t	1.6E+07
	24.75498039			10293021	24.7549804	24.75498039	t	1.6E+07
	0.789981			10293019	0.789981	0.789981	t	1.6E+07
	22.80019799			10293020	22.800198	22.80019799	t	1.6E+07
	2.992995335			10293023	2.99299533	2.992995335	t	1.6E+07
	11.7			10293022	11.7	11.7	t	1.6E+07
	23.62690037			10293026	23.6269004	23.62690037	t	1.6E+07
	0.80143			10293024	0.80143	0.80143	t	1.6E+07
	22.80019799			10293025	22.800198	22.80019799	t	1.6E+07
	2.741615149			10293028	2.74161515	2.741615149	t	1.6E+07
	11.7			10293027	11.7	11.7	t	1.6E+07
	22.12017783			10293031	22.1201778	22.12017783	t	1.6E+07
	0.80143			10293029	0.80143	0.80143	t	1.6E+07
	22.51813368			10293030	22.5181337	22.51813368	t	1.6E+07
	2.898727765			10293033	2.89872777	2.898727765	t	1.6E+07
	11.7			10293032	11.7	11.7	t	1.6E+07

Tabla 4: Estructura de la Base de datos - Parte 2 (IERSE, 2017)

Las tablas 3 y 4 muestran la estructura de base de datos proporcionada por el IERSE, en donde se tiene la siguiente información: la fecha y la hora de la recolección, el identificador de cada sensor utilizado para medir los valores de los contaminantes respectivamente, un código de identificación de la muestra, el nombre del contaminante, el valor medido, además de tres campos sin datos que son omitidos en el proceso de medición y que para este caso no representa un problema ya que no son necesarios para el análisis.

4.1. Variabilidad de los datos.

Para garantizar la confiabilidad de los resultados es necesario que el análisis y las pruebas sean realizadas en distintos escenarios, por lo tanto, las variables de estudio deben presentar una variación considerable en sus valores. A continuación, se muestra

una serie de ilustraciones que muestran los valores históricos de las variables de contaminación a lo largo del año 2015 tomadas del informe anual de calidad del aire de la ciudad de Cuenca.

Registros de la estación automática: Material Particulado 2.5um (PM2.5)

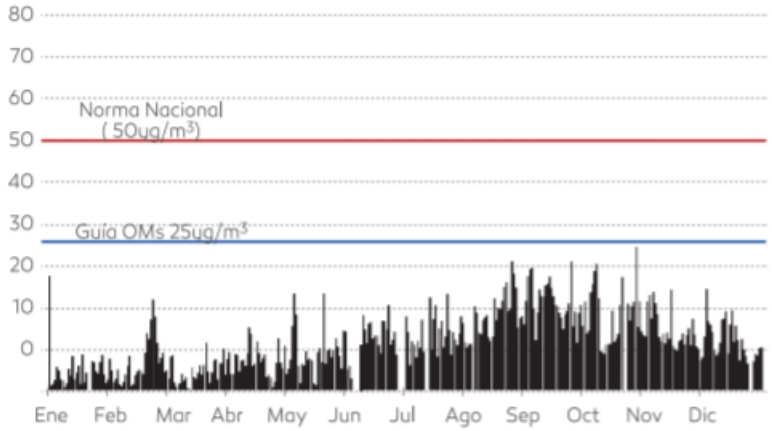


Ilustración 3: Registros de Material Particulado 2.5 um (PM2.5) del año 2015 (EMOV-EP, 2015).

Registros de la estación automática: Monóxido de Carbono (CO)

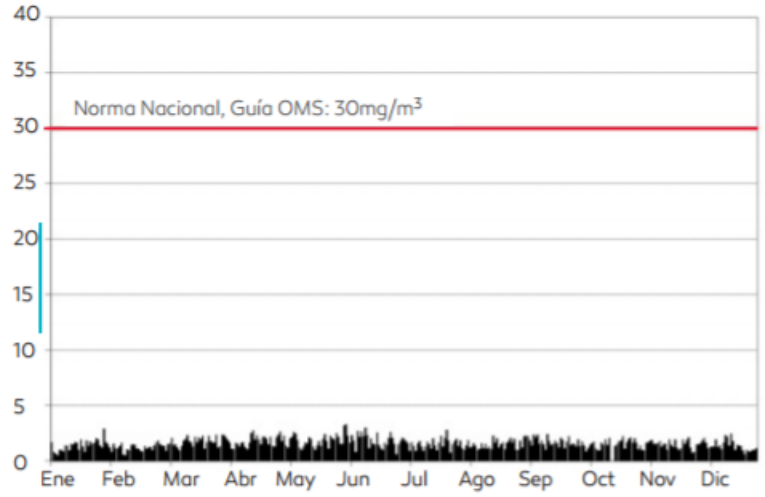


Ilustración 4: Registros de Monóxido de Carbono (CO) del año 2015 (EMOV-EP, 2015).

Registros de la estación automática: Dióxido de Azufre (SO2)

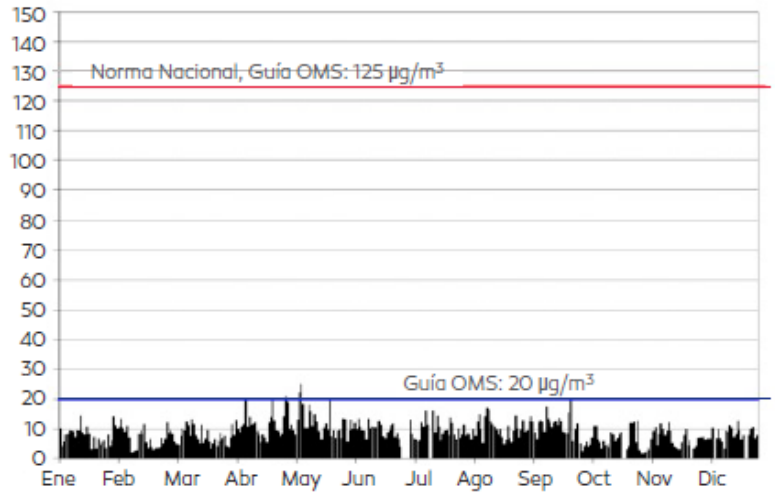


Ilustración 5: Registros de Dióxido de Azufre (SO2) del año 2015 (EMOV-EP, 2015).

Registros de la estación automática: Dióxido de Nitrógeno (NO2)

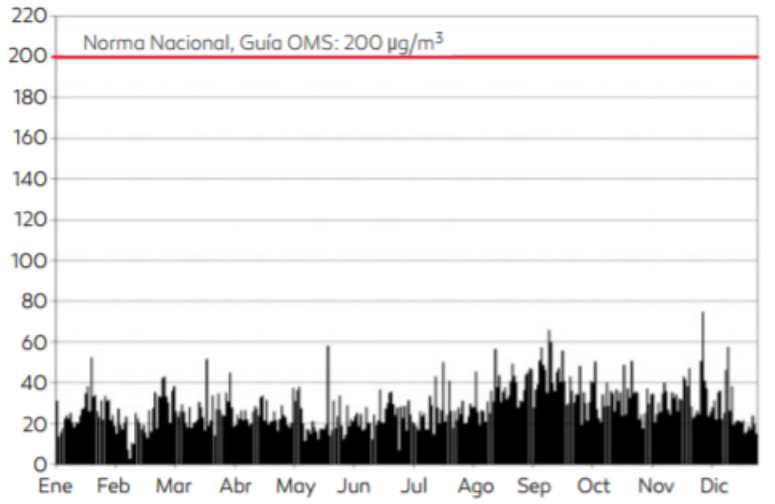


Ilustración 6: Registros de Dióxido de Nitrógeno (NO) del año 2015 (EMOV-EP, 2015).

Registros de la estación automática: Ozono (03)

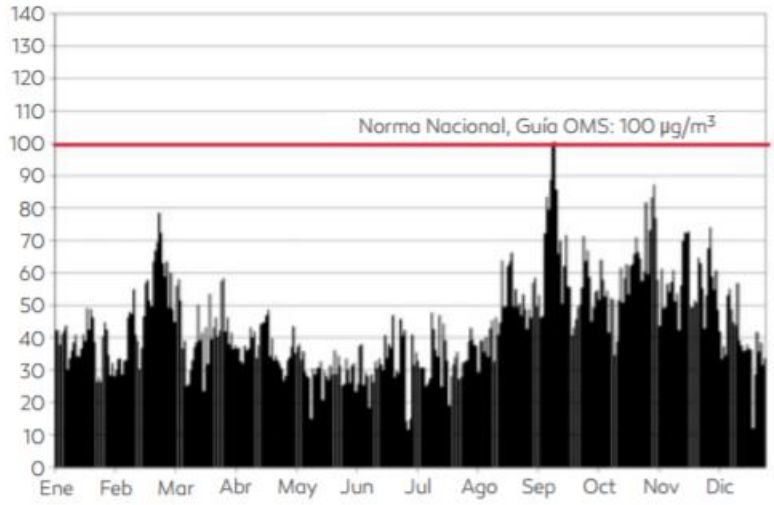


Ilustración 7: Registros de Ozono (O3) del año 2015 (EMOV-EP, 2015).

4.2. Conclusión

Como se observa en los gráficos anteriores, todas las variables presentan concentraciones que por lo general están por debajo de los límites que determina OMS y la norma Nacional, sin embargo, los datos de todas las variables presentan una variación considerable a lo largo de todo el año, por lo tanto, presenta varios escenarios de trabajo y aporta credibilidad a los resultados.

5. Preparación de los datos

En esta sección se describe el procesamiento de los datos para obtener una base de datos con la estructura de datos adecuada para aplicar los métodos de minería, además, se presenta el proceso de limpieza de datos para eliminar registros con datos anómalos y/o nulos que impidan la ejecución de los algoritmos o afecten a los resultados.

Como se observó en la tabla 3, a cada contaminante le corresponde un registro dentro de la base de datos, formando cinco tuplas diferentes por cada sensor de la estación de

monitoreo. Dentro de la herramienta que se utilizará, esta estructura de la base de datos no está permitida, por lo tanto, es necesario combinar estos cinco registros en una sola tupla tomando como atributos los valores de cada contaminante junto con la fecha y hora en la que realizó la medición. Al final de este proceso quedan 43760 registros tal y como se observa en la ilustración 8.

```
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;
select count(*) from tc_envcon;
```

	COUNT(*)
1	43760

Ilustración 8: Sentencia SQL para restructurar la base de datos y contar el número de registros (LIDI, 2017)

La Tabla 5 muestra la estructura de final del *dataset*, en donde únicamente se conserva la fecha y la hora en la que se realizó la medición contenidas en el campo CA_TIME_STAMP, y los valores medidos de cada contaminante en su columna correspondiente.

CA_TIME_STAMP	O3	CO	NO2	SO2	PM2_5
16-Sep-17	20.4153056	0.789981	23.4809132	3.01132514	7.2
16-Sep-17	23.0873838	0.789981	23.445185	3.00346951	9.7
16-Sep-17	24.7549804	0.789981	22.800198	2.99299533	11.7
16-Sep-17	23.6269004	0.80143	22.800198	2.74161515	11.7

Tabla 5: Estructura de la base de datos luego de aplicar el proceso de reestructuración (LIDI, 2017)

Luego de reestructurar el *dataset* se realiza una limpieza de los datos, eliminando todos los registros que contengan al menos un dato nulo ya que en la siguiente etapa este tipo de datos dificultan el procedimiento, ya sea por que alteran los resultados o simplemente algunos algoritmos no aceptan un *dataset* con este tipo de datos.

Con una simple consulta SQL a la base de datos se puede obtener el número exacto de registros que contienen al menos un dato nulo. La ilustración muestra la sentencia utilizada junto al total de registros con datos nulos.

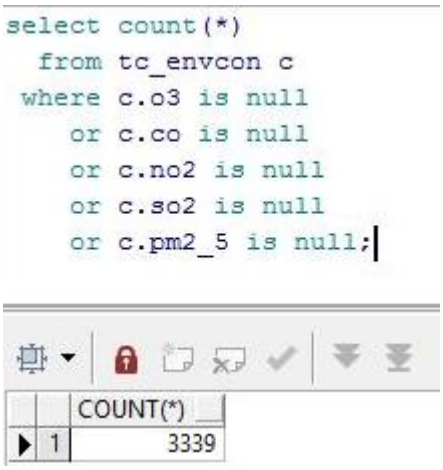


Ilustración 9: Sentencia SQL para contar número de registros con datos nulos (LIDI, 2017).

A continuación, utilizando otra sentencia SQL a la base de datos se procede a eliminar los registros con datos nulos. La ilustración 10 muestra la sentencia que elimina todos los registros que contienen al menos un dato nulo.

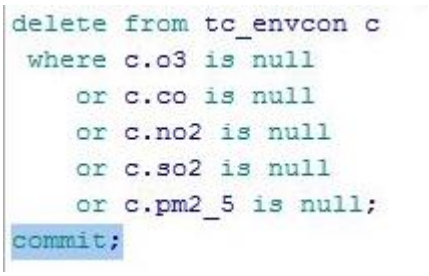


Ilustración 10: Sentencia SQL para eliminar los registros don datos nulos (LIDI, 2017).

Los datos son registrados cada segundo, pero debido a que la variabilidad en ese periodo es despreciable se tomará un valor promedio cada minuto. La literatura recomienda que este período no sobrepase la media hora, por lo tanto, un minuto es suficiente para cumplir este requerimiento. Al final se obtiene un *dataset* con 23575 registros.

5.1. Conclusión

Luego de aplicar una serie de operaciones sobre los datos recolectados por la estación de monitoreo se ha obtenido un *dataset* con la estructura deseada y con datos depurados que cumple tanto las necesidades del problema, así como de los requerimientos y normas establecidas por la teoría en la minería de datos.

6. Modelado

A continuación, se presenta las herramientas utilizadas para la preparación y minado de los datos junto con los algoritmos que serán evaluados, además, se plantea el proceso de pruebas a los que serán sometidos los algoritmos.

Con los datos obtenidos se evidencia que no existe una variable dependiente en el *dataset*, por lo tanto, los métodos supervisados o de entrenamiento se descartan, quedando únicamente los métodos no supervisados (clustering). Existe una gran variedad de algoritmos de clustering, sin embargo, en este trabajo se analizará

únicamente cinco algoritmos, que son algunos de los más representativos de esta categoría, como se apreció en algunos de los casos de minería de datos revisados anteriormente, además, se encuentran ya implementados de forma gratuita en RapidMiner. Estos algoritmos son: k-means, X-means, Expectation Maximization (EM), Cobweb y DBSCAN.

Luego de obtener la base de datos y seleccionar los algoritmos, se realizan las pruebas de cada uno de los algoritmos para analizar los resultados. Para esta tarea se utilizará la herramienta RapidMiner, la cual es una herramienta gratuita y cuenta con una gran variedad de algoritmos ya implementados, además, es una aplicación multiplataforma desarrollada en java, por lo tanto, puede ser ejecutada en cualquier sistema operativo. RapidMiner presenta una interfaz totalmente gráfica, por lo que es amigable con el usuario y muy fácil de utilizar. Si bien la versión libre está limitada para el manejo de un máximo de 10000 registros, existe la posibilidad de obtener una licencia educacional de forma gratuita con la cual permite manejar un número ilimitado de registros por un año, tiempo suficiente para llevar a cabo este proyecto. La ilustración muestra que se ha realizado la activación de esta licencia con éxito.

Ilustración 11: Licencia Educacional activada (RapidMiner, 2017).

La información obtenida de cada prueba será almacenada en una matriz comparativa que servirá como soporte para el análisis de resultados**.** Todas estas pruebas se realizarán en un mismo equipo para mantener un mismo escenario para todas las pruebas y evitar el sesgo.

Características de equipo:

Procesador: Intel Core I7 7500U 2.7–2.9 GHz

Memoria RAM (Random Access Memory): 12 GB

Sistema Operativo: Windows 10 Home Edition.

GPU: NVIDIA GeForce 940MX 2GB

Disco Duro: 2 TB a 54000 RPM

Una vez definidos todos los requerimientos, se evalúa cada uno de los algoritmos utilizando el mismo *dataset* con 23575 registros. Antes de aplicar cualquier procedimiento de minería de datos se revisa el estado de la base de datos para verificar

que no existan datos anómalos que afecten el análisis y distorsionen los resultados. Para ello *RapidMiner* cuenta con una opción que permite visualizar las estadísticas del *dataset*.

La ilustración 12 muestra dicha herramienta, en la cual es posible visualizar el nombre de los atributos, el tipo de dato, el número de datos perdidos en cada columna, los valores máximos y mínimos de cada atributo, así como de su promedio; todo esto con el fin de validar los datos antes de procesarlos.

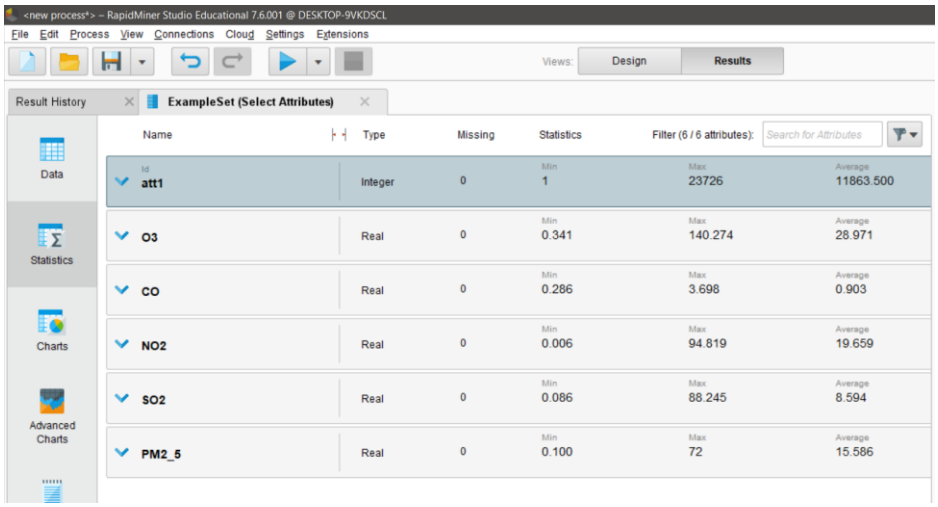


Ilustración 12: Interfaz de RapidMIner para visualizar las estadísticas de la base de datos.

El proceso de pruebas se realizó bajo la siguiente estructura:

Se cargan los datos a través del operador "*Read CSV*", posterior a esto, se realiza una conversión de los datos con la ayuda de los operadores "*Nominal to Date*" para el filtrado de los datos por hora. Luego, se unifica el *dataset* transformado el CO de mg/m3 a ug/m3 y se normaliza los valores de los contaminantes para trabajar en una misma escala y mejorar la comprensión de los resultados. Se aplica un filtro para tomar únicamente los datos en ciertas horas del día a través del operador "*Select Attributes"*

que genera un *dataset* únicamente con los valores de los contaminantes para aplicar el método de *clustering* correspondiente. De todos los clústeres obtenidos se toma el clúster con mayor número de elementos, en el cual se realizará un análisis de frecuencia para establecer el factor de correlación de cada uno de los elementos. Este proceso se encuentra representado en la siguiente ilustración:

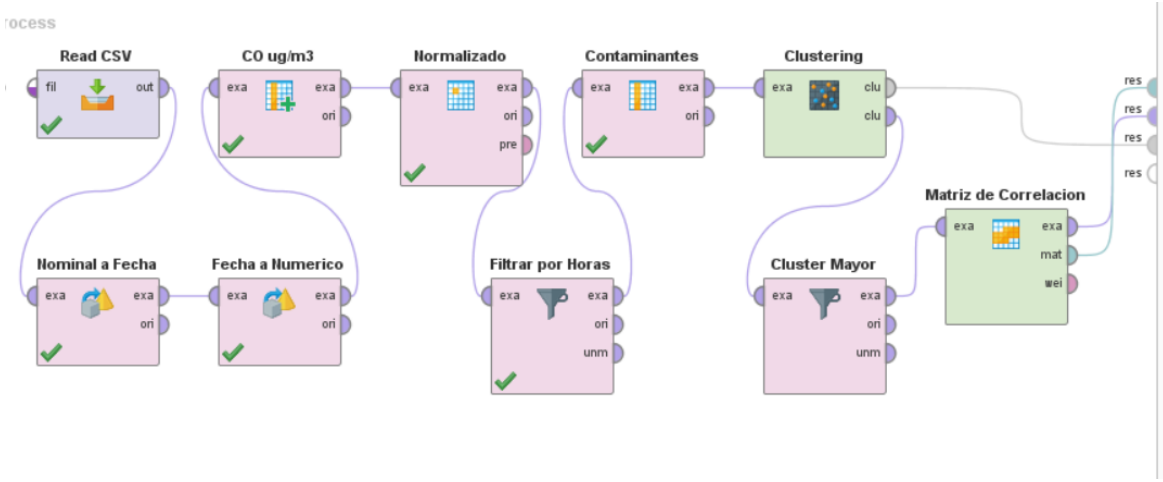


Ilustración 13: Diagrama del proceso de ejecución

Una de las falencias de *RapidMiner* es su limitada capacidad para representar los resultados mediante gráficos, por esto se optó por utilizar el software *Minitab*, que es una herramienta para el procesamiento estadístico de los datos, con la cual se realizará las gráficas para visualizar el comportamiento de los datos y compararlos con los factores de correlación obtenidos en *RapidMiner*.

6.1. Conclusión

Es fundamental conocer la herramienta de minería de datos, ya que el mal uso de la misma o la mala interpretación de los resultados podría llevar a conclusiones erróneas y a una mala toma de decisiones que perjudican tanto a empresa como a usuarios.

7. Evaluación

En esta sección los algoritmos son evaluados siguiendo el protocolo establecido anteriormente y los resultados obtenidos serán analizados para determinar cuál es el mejor algoritmo en base a los criterios que se definen a continuación.

7.1. Criterios de evaluación

Dentro de la minería de datos, el tiempo es uno de los factores más importantes a tener en cuenta, por lo tanto, el tiempo de ejecución de cada algoritmo será considerado como un criterio de evaluación. Puesto que se pretende determinar patrones de comportamiento de las variables de contaminación del aire, los factores de correlación obtenidos serán otro criterio de evaluación de los algoritmos.

7.2. Ejecución de Pruebas

A continuación, se presenta una breve explicación del funcionamiento de cada algoritmo seguido de los resultados obtenidos luego de ejecutar las pruebas utilizando el esquema establecido en el modelado.

7.2.1. K-means

En este algoritmo, el usuario determina el número de clúster para construir el modelo, luego, el algoritmo determina los centroides de cada clúster de forma aleatoria y los elementos serán asignados al clúster cuya distancia con el centroide sea la menor. Para la siguiente iteración, se recalcula los centroides y reasigna los elementos. Este proceso se repite hasta que la variabilidad de los centroides en cada iteración sea mínima o nula (Anu, Divyadharshini, & Science, 2017).

Resultados

El tiempo de ejecución del algoritmo con 4 clústeres fue de 500 ms (milisegundos), y se obtuvo los valores de los centroides por contaminante, además del número de elementos en cada clúster como se muestra en la tabla 6. La ilustración 14 muestra el tamaño de cada clúster siendo el clúster dos el más grande con el 50% de los datos.

	Clúster 0	Clúster 1	Clúster 2	Clúster 3
O3	17,09	65,03	35,38	11,20
NO2	43,09	10,76	7,36	22,79
SO2	26,71	5,31	4,96	13,42
PM2_5	56,68	44,75	12,63	15,10
CO	33,98	15,74	12,31	22,04
Ítems	1765	1710	6098	3359
%	14%	13%	47%	26%

Tabla 6: Centroides y número de elementos por clúster obtenidos con K-means

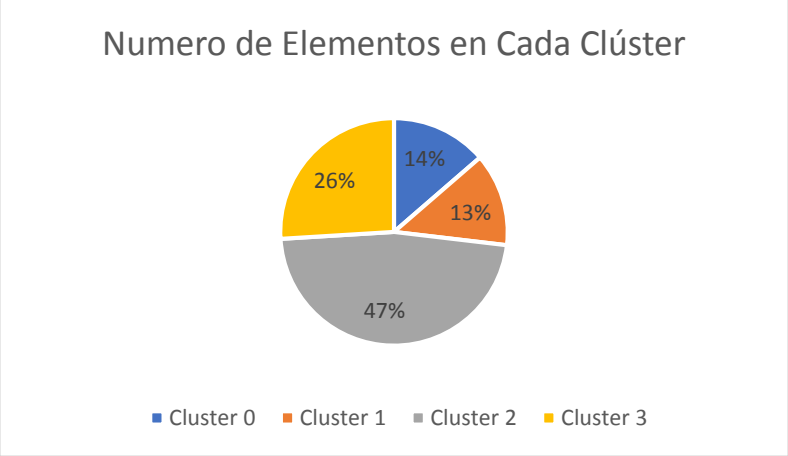


Ilustración 14: Distribución de los elementos en cada clúster obtenida con K-means

Los resultados muestran una dependencia de algunos contaminantes con otros en ciertos niveles, por ejemplo, existe una relación indirecta del O3 con el resto de contaminantes cuando este se encuentra en valores altos o bajos, pero cuando el O3 se encuentra en

niveles medios y cercanos a cero se pierde esta relación. En la siguiente ilustración se muestra el comportamiento de los componentes en cada clúster:

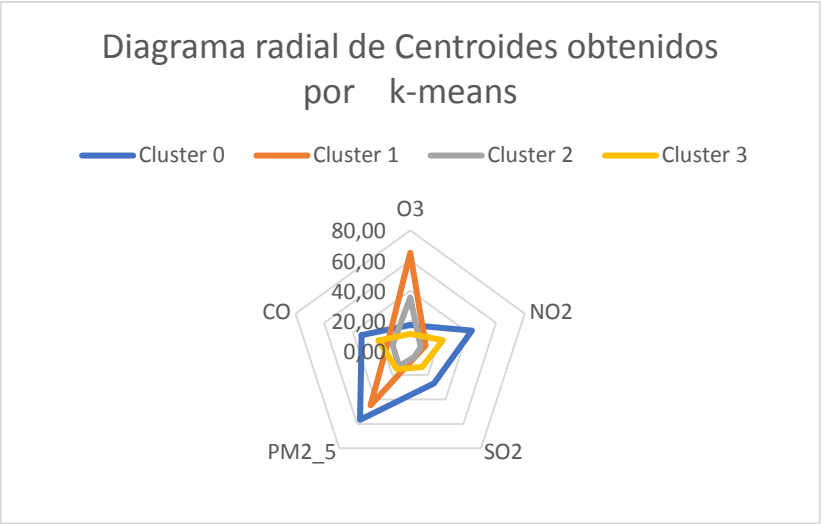


Ilustración 15: Diagrama Radial de los centroides por contaminante obtenido con K-means

Se utiliza el operador "*Correlation Matrix"* sobre los datos del clúster más grande para obtener el factor de correlación de cada uno de los elementos, de esta se forma analiza de manera más precisa el comportamiento de un contamínate frente a otro. En las siguientes ilustraciones se muestra los factores de correlación obtenidos entre los contaminantes junto a su respectiva gráfica de series de tiempo, en donde el eje vertical muestra el valor medido de los contaminantes y el eje horizontal muestra el índice los contamínate.

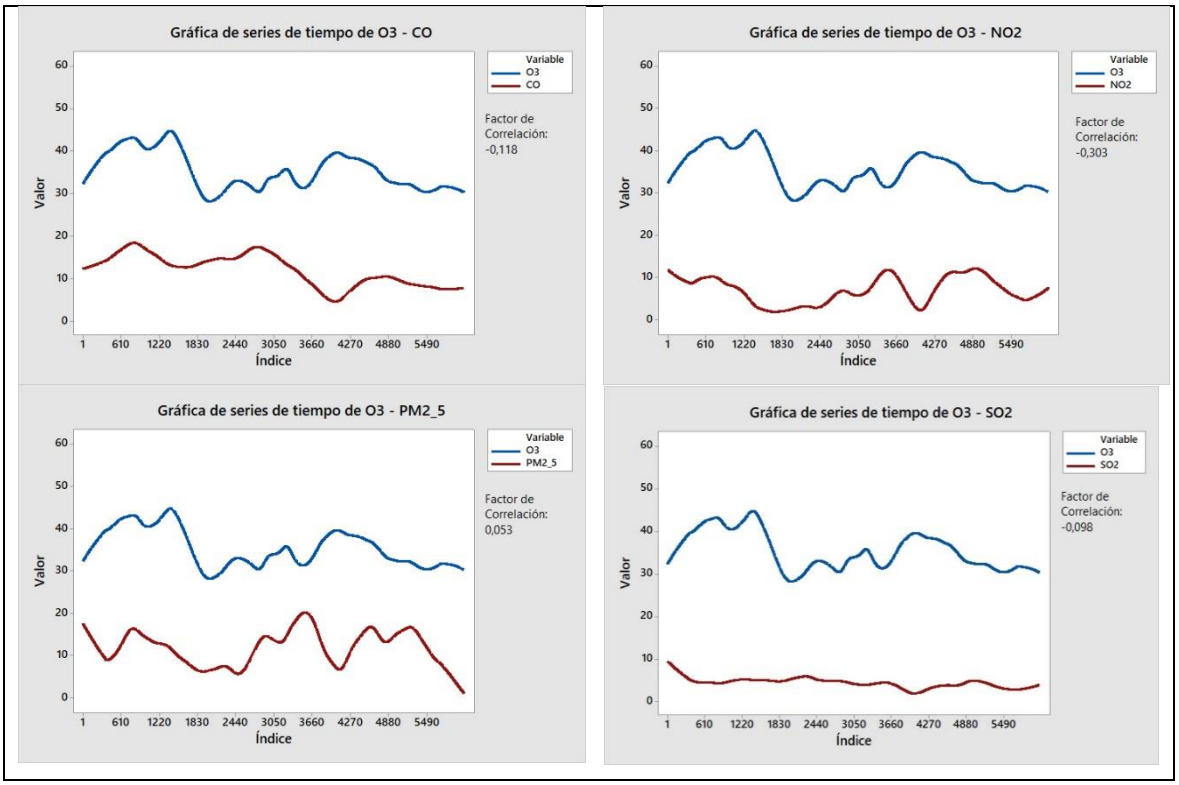


Ilustración 16: Gráficas de serie de tiempo y factor de correlación de O3 obtenidos con k-means

En la ilustración 16 se observa que el O3 presenta una correlación considerable únicamente con el NO2 con un factor de correlación de -0.3, mientras que para los demás contaminantes dicho factor es cercano a cero. Si bien el O3 presenta factores de correlación bajos con el CO y PM2_5, esto no quiere decir que no exista un patrón de comportamiento ya que como se ven las gráficas, existen tramos en los cuales las líneas presentan un comportamiento directo y otros con un comportamiento indirecto, lo cual puede hacer que el valor de factor de correlación disminuya.

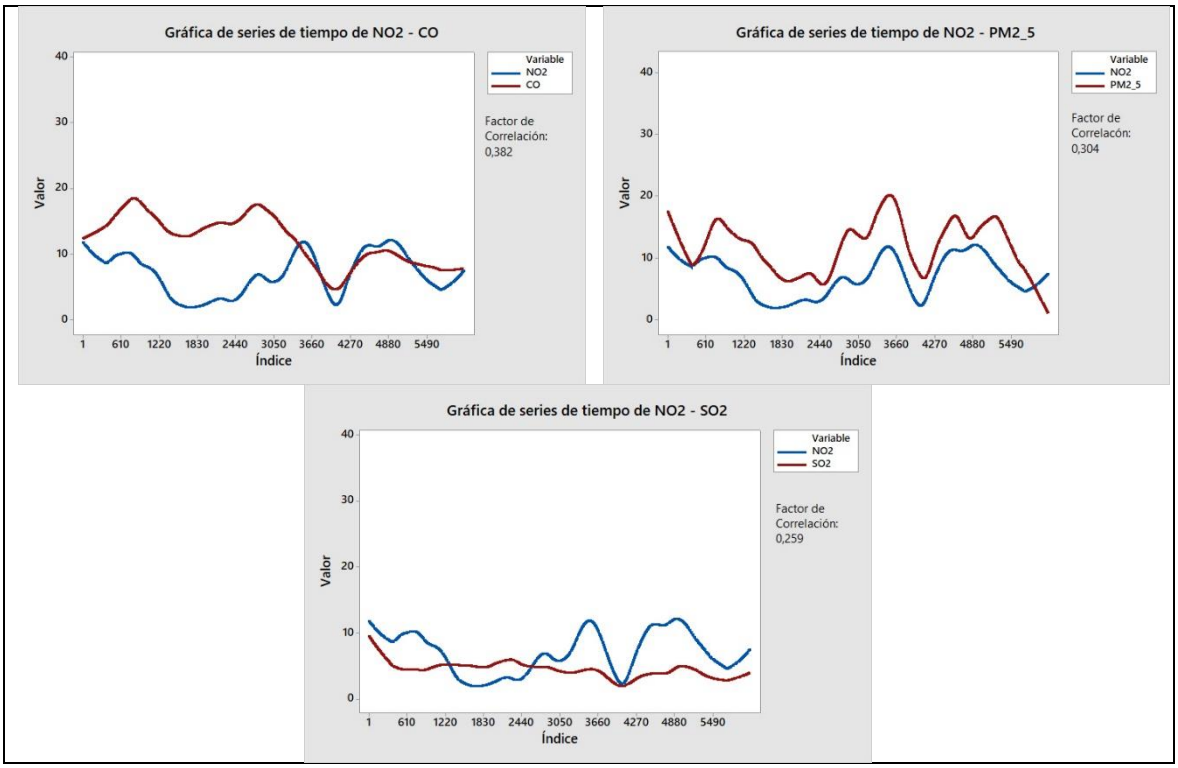


Ilustración 17: Gráficas de serie de tiempo y factores de correlación de NO2 obtenidos con K-means

De la ilustración anterior se observa que el NO2 presenta un patrón de comportamiento de forma directa con el CO, PM2_5 y CO ya que se sus factores de correlación con dichos contaminantes son cercanos a 0.3, además, las gráficas de series de tiempo muestran curvas con comportamiento similar en todo el tramo. En la gráfica con el SO2 puede parecer que esto no se cumpla debido a los grandes picos y valles que presenta el NO2, pero el comportamiento de SO2, si bien no cambia en gran medida, sigue el patrón establecido por NO2.

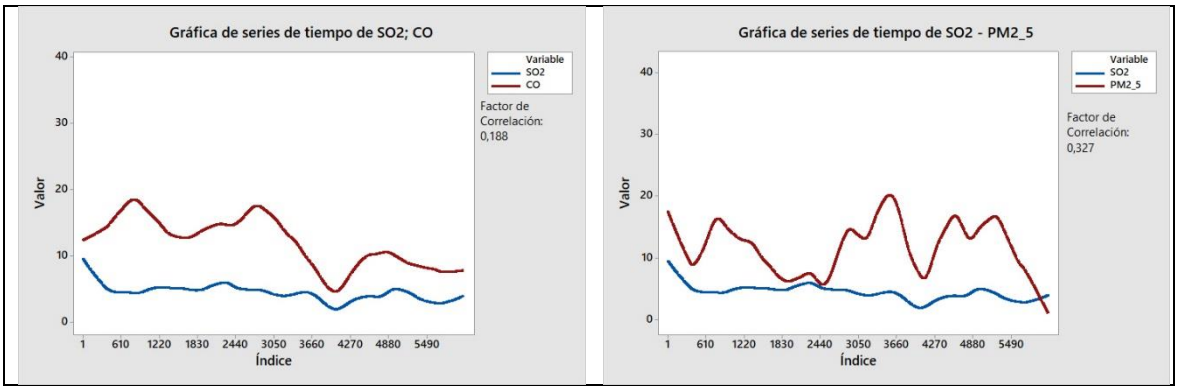


Ilustración 18: Gráficas de series de tiempo y factores de correlación de SO2 obtenidos con K-means

En la ilustración 18 se observa que, pese a que el factor de correlación con PM2_5 es mucho mayor al de CO, las gráficas no reflejan dichos valores. La gráfica de series de tiempo con CO presenta un patrón mucho más armónico con el SO2 frente al del PM2_5.

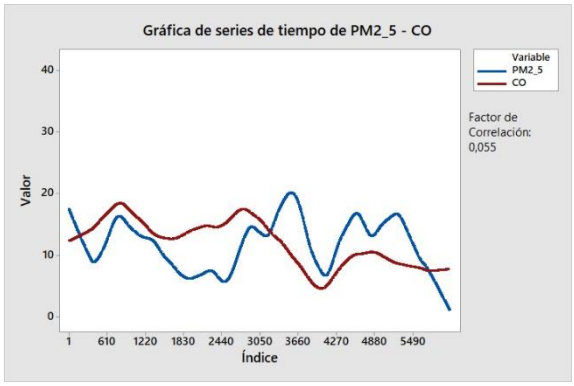


Ilustración 19: Gráfica de serie de tiempo y factor de correlación de PM_5 y CO obtenido con K-means

La ilustración 19 presenta el factor de correlación de PM2_5 y el CO junto a su gráfica de series de tiempo, en la cual se puede observar tramos en los cuales las líneas describen un comportamiento similar y otros en los que para nada dichas curvas se relacionan, esto se refleja en el factor de correlación tan bajo obtenido.

7.2.2. **X-means**

A este método se le puede considerar como una extensión del *k-means*, ya que recibe como parámetro un rango en el cual se determina el valor de *k* con el cual se obtiene los

mejores resultados aplicando *k-means*. *X-means,* en cada iteración del algoritmo determina si es necesario agregar un nuevo clúster junto con la posición del nuevo centroide en caso de ser necesario. Para esto, realiza una subdivisión de cada clúster, es decir, ejecuta un 2-means en cada clúster y, en el clúster donde la distancia entre centroides sea amplia será ubicado un nuevo centroide. Una vez que el algoritmo se haya ejecutado para todos los valores de *k* dentro del rango solicitado, se devolverá el mejor resultado obtenido (Dan Pelleg, 2015).

Resultados

Este algoritmo fue ejecutado en un rango de 2 a 6 clústeres, siendo 4 el número óptimo de clústeres con un tiempo de ejecución de 1766 ms. La tabla 7 muestra los centroides y el número de elementos por clúster. Además, la figura 23 representa el tamaño de los clústeres, siendo el más grande el clúster 1 con el 60% de los datos.

	Clúster 0	Clúster 1	Clúster 2	Clúster 3
O3	65,39	28,70	30,86	9,28
NO2	9,74	10,71	39,93	32,84
SO2	3,75	6,28	23,59	23,43
PM2_5	36,32	12,36	68,47	30,99
CO	15,92	13,69	30,38	31,96
Ítems	1712	7849	1138	2233
%	13%	61%	9%	17%

Tabla 7: Centroides y número de elementos por clúster obtenidos con X-means

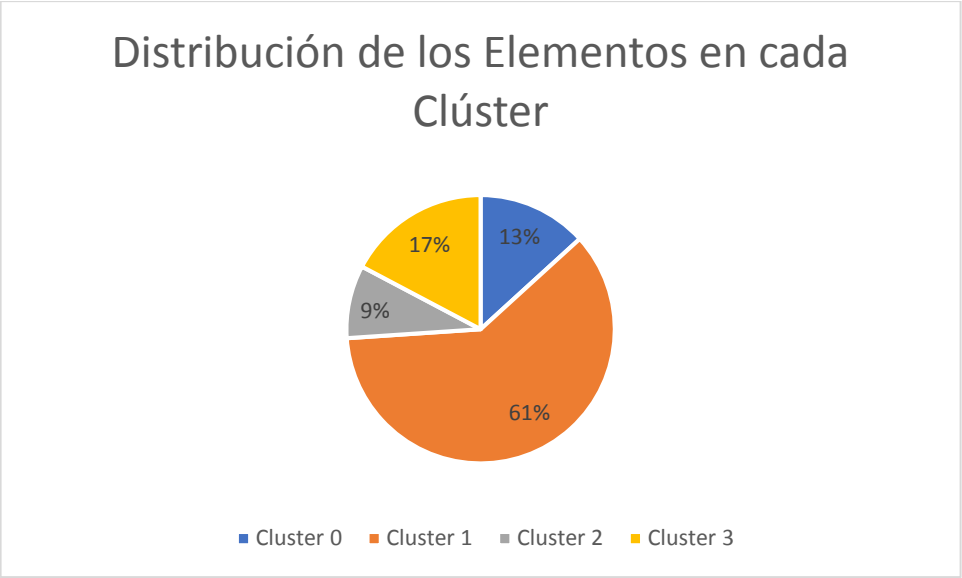


Ilustración 20: Distribución de los elementos por clúster obtenido con X-means

Los resultados obtenidos son muy similares a los de K-means, a tal punto que se podría decir que los patrones se mantienen. A continuación, se presenta gráficamente el comportamiento de los componentes en cada clúster:

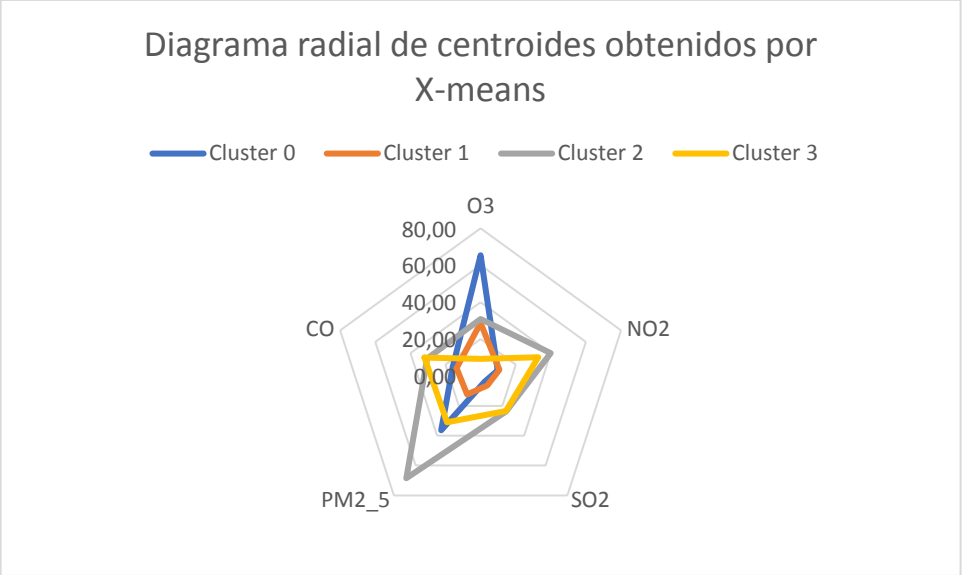


Ilustración 21: Diagrama Radial de centroides por contaminante obtenido con X-means

A continuación, las ilustraciones de la 22 a la 25 muestran los factores de correlación obtenidos entre cada contaminante junto a sus respectivas gráficas de series de tiempo que validan los resultados.

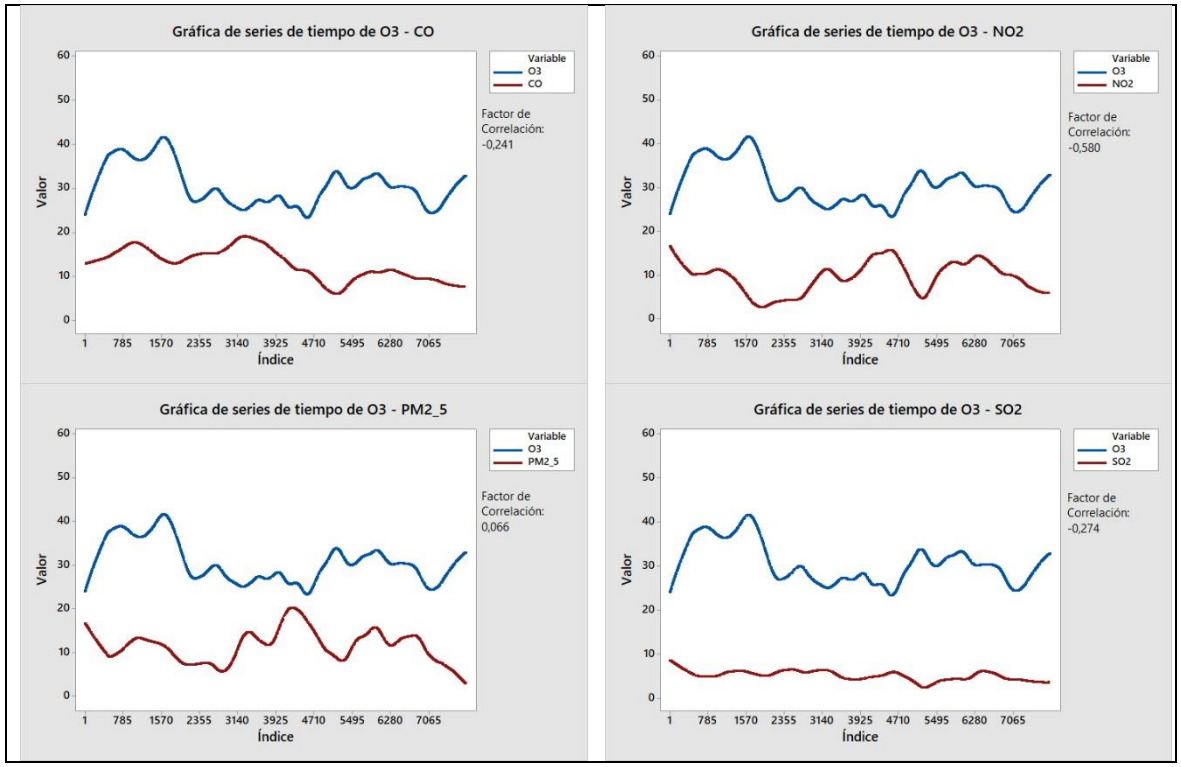


Ilustración 22: Gráficas de series de tiempo y factores de correlación de O3 obtenidos con X-means

En la ilustración 22 se observa que los factores de correlación de O3 obtenidos con este algoritmo son mayores a los obtenidos con k-means, además, la correlación con NO2 se mantiene como la más fuerte en relación a los demás contaminantes con un factor de correlación de -0.5.

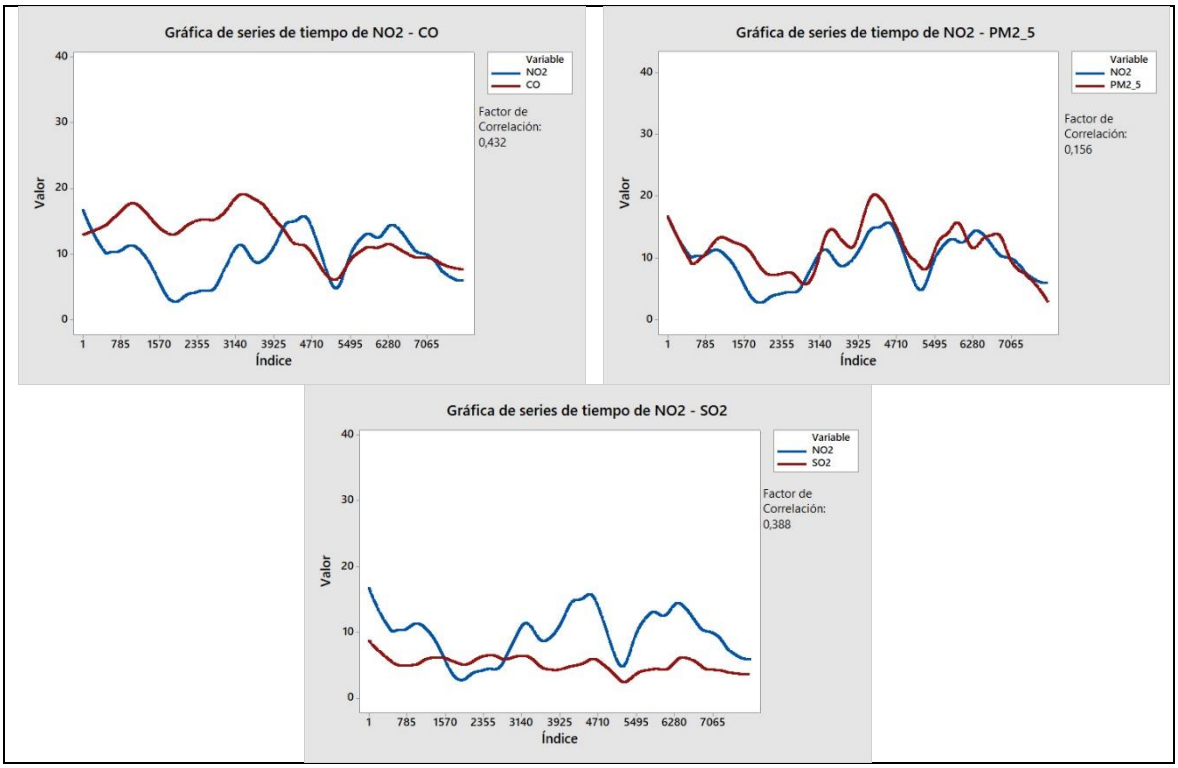


Ilustración 23: Gráficas de series de tiempo y factores de correlación de NO2 obtenidos con X-means

La gráfica anterior presenta los factores de correlación de NO2, en donde se observa que existe una relación considerable con el CO y el SO2 ya que sus factores de correlación rodean el 0.4. No obstante si se toma en cuenta las curvas de NO2 y PM2_5 se puede apreciar que también existe una fuerte correspondencia entre estos dos factores incluso mayor que las 2 anteriores, pero su factor de correlación presenta un valor bajo.

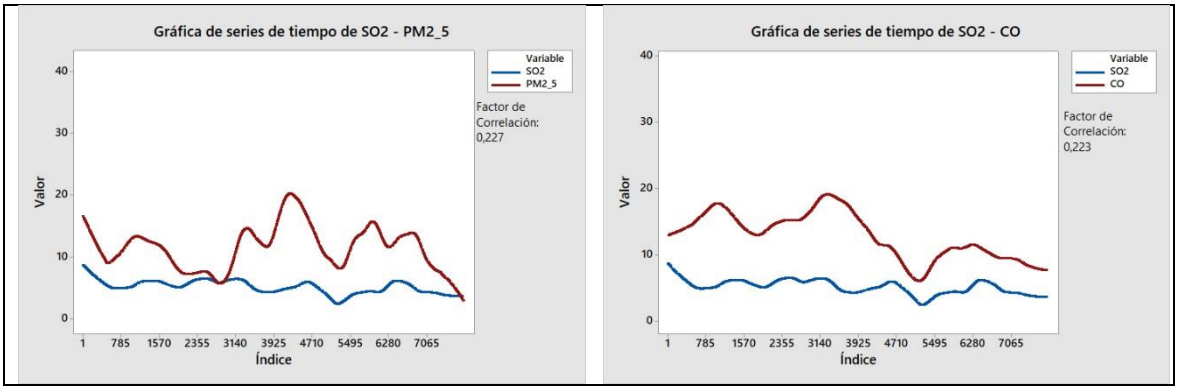


Ilustración 24: Gráficas de Series de tiempo y factores de correlación de SO2 obtenidos con X-means

En la figura anterior se muestra los factores de correlación del SO2 con el PM2_5 y el CO obtenidos con x-means. El factor de correlación en ambos casos es de 0.22 lo que significa que existe una leve correlación entre dichos contamines, sin embargo, considerando las gráficas de series de tiempo se esperaría que el factor de correlación con el CO fuese mayor ya que sus curvas se comportan manera más armónica.

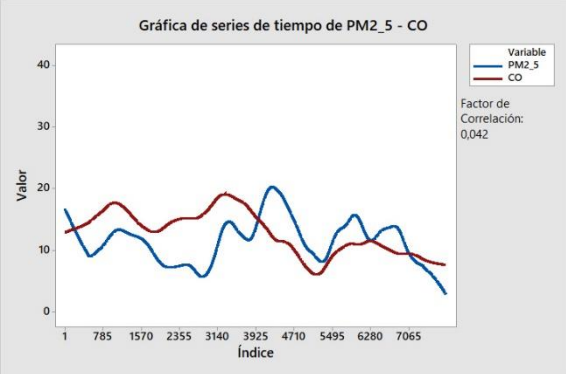


Ilustración 25: Gráfica de series de tiempo y factor de correlación entre PM2_5 y CO obtenidos con X-means

La anterior ilustración muestra los resultados obtenidos entre el PM2_5 y el CO, mismos que son similares a los obtenidos con k-means, tanto en el factor de correlación, así como en las gráficas de series de tiempo.

7.2.3. **Expectation Maximization**

El algoritmo de maximización de probabilidad utiliza parámetros de distribución para calcular la probabilidad de que un elemento pertenezca a un clúster. Es un modelo incremental, el cual utiliza una validación cruzada para determinar el número óptimo de clústeres. EM consta principalmente de dos fases: la *E-Step* o etapa de probabilidad, en la cual se calcula la probabilidad de pertenencia de cada elemento y se asigna los elementos a su respectivo clúster; Y la *M-Step* o etapa de maximización, en donde se reestima los parámetros de distribución para maximizar la probabilidad de pertenencia de los elementos (Anu et al., 2017).

Resultados

El tiempo de ejecución de este algoritmo fue de 1.88 minutos, obteniendo 4 como el numero óptimo de clústeres. la tabla 8 muestra los valores de los centroides por contamínate y el número de elementos en cada clúster, y la ilustración 26 representa el tamaño de los clústeres, siendo el más grande el clúster 0, con el 40% de los datos.

	Clúster 0	Clúster 1	Clúster 2	Clúster 3
O3	34,01	5,22	21,39	44,28
NO2	7,46	27,71	31,61	13,51
SO2	3,93	20,92	20,60	6,21
PM2_5	10,09	17,67	44,94	29,27
CO	12,51	28,34	26,94	15,22
Ítems	5375	1673	2600	3284
%	42%	13%	20%	25%

Tabla 8: Centroides y número de elementos por cada clúster obtenido con EM

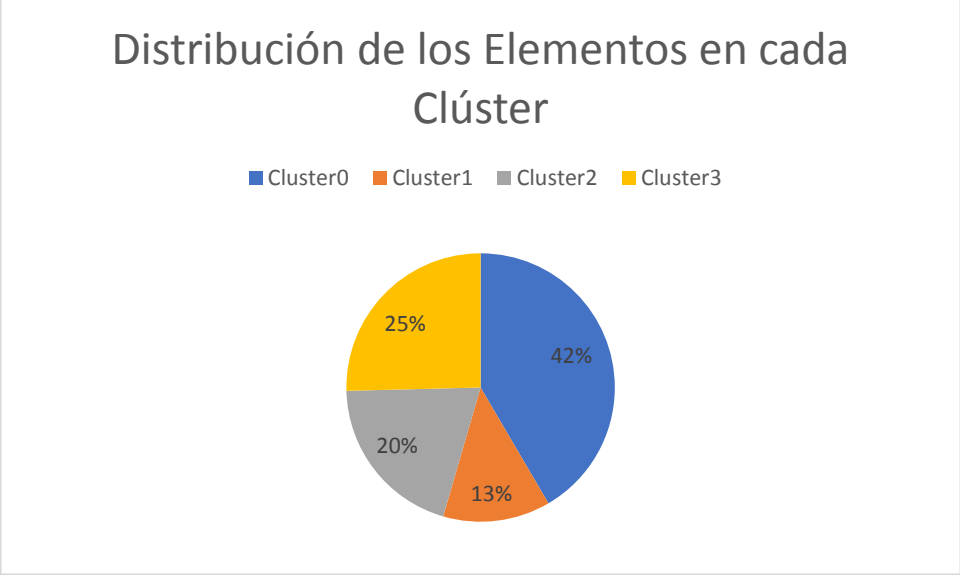


Ilustración 26: Número de elementos en cada clúster obtenidos con EM

Los resultados son similares a los valores obtenidos por K-means y X-means, por lo tanto, describen un patrón de comportamiento similar de los contaminantes, tal y como se observa en la ilustración 27.

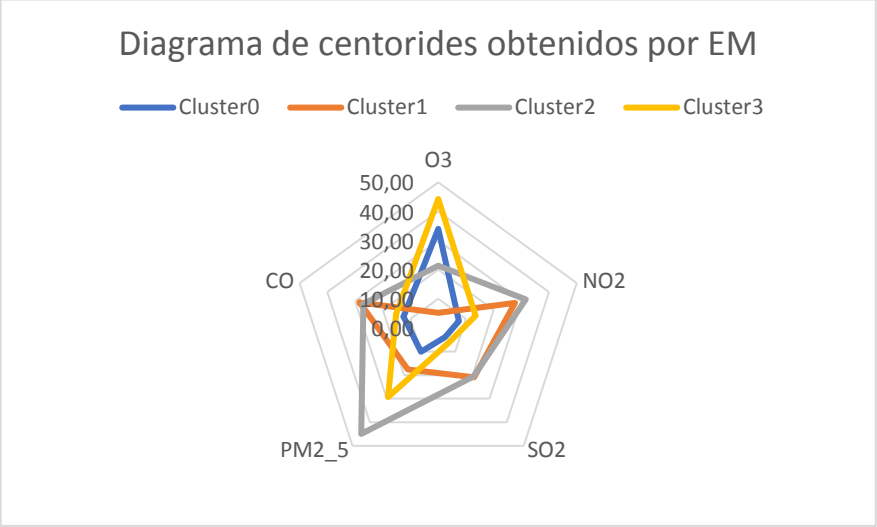


Ilustración 27: Diagrama Radial de centroides por contamínate obtenidos con EM

Utilizando los datos del clúster 0, se obtiene los factores de correlación entre cada uno de los contaminantes. Las siguientes ilustraciones muestran dichos valores junto a sus respectivas gráficas de series de tiempo.

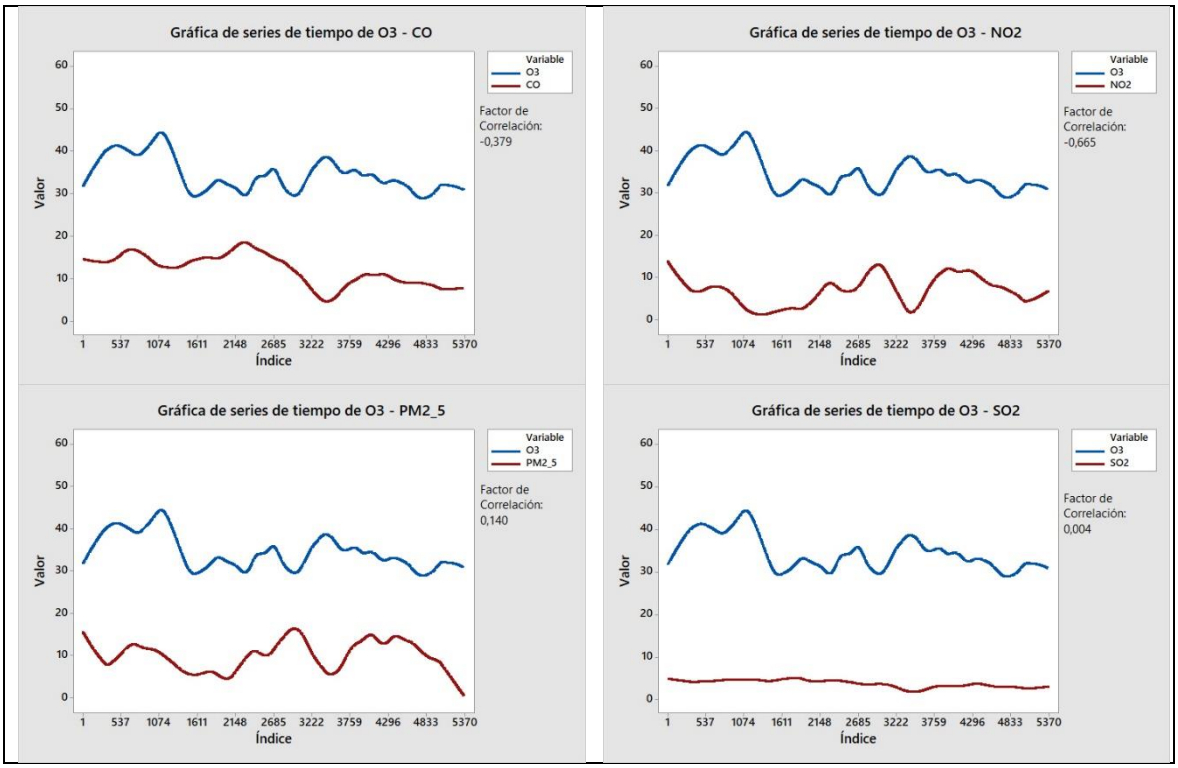


Ilustración 28: Diagramas de series de tiempo y factores de correlación de O3 obtenidos con EM

En la ilustración 28 se observa los factores de correlación de O3 obtenidos con el algoritmo EM junto a sus gráficas de series de tiempo. En este caso los valores de los factores son aún mayores a los obtenidos con los dos algoritmos anteriores, pero aún se mantiene el factor de correlación con NO2 como factor más alto frente a los otros valores. Entre el O3 y el SO tanto la gráfica como el factor de correlación indican que no existe correlación alguna, pues se tiene un factor de correlación igual a cero, además, el SO2 tiene un comportamiento linear uniforme, mientras que el O3 sufre varias alteraciones.

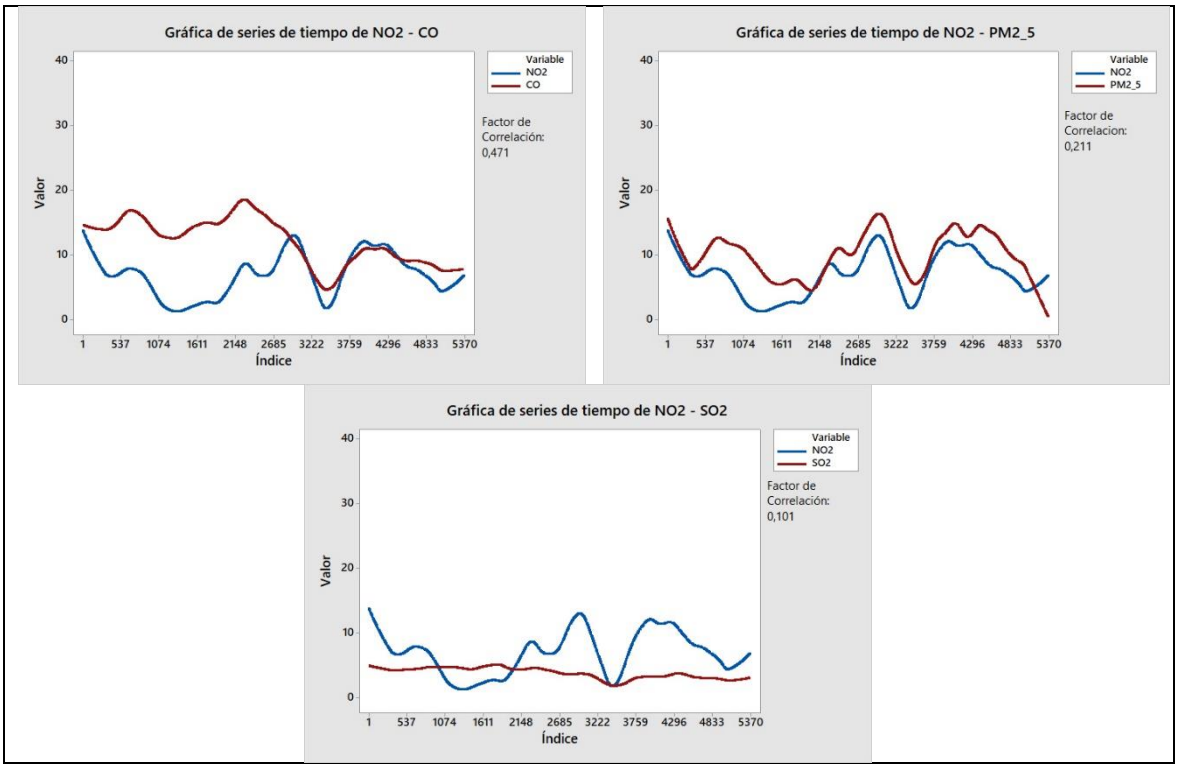


Ilustración 29: Diagramas de series de tiempo y factores de correlación de NO2 obtenidos con EM

En la gráfica anterior se muestra los factores de correlación de NO2 obtenidos, donde el valor del factor de correlación con el PM2_5 si bien aumenta con relación a lo obtenido con el algoritmo anterior, no concuerda con lo que describen las gráficas de series de tiempo, ya que dichas gráficas presentan un comportamiento muy similar.

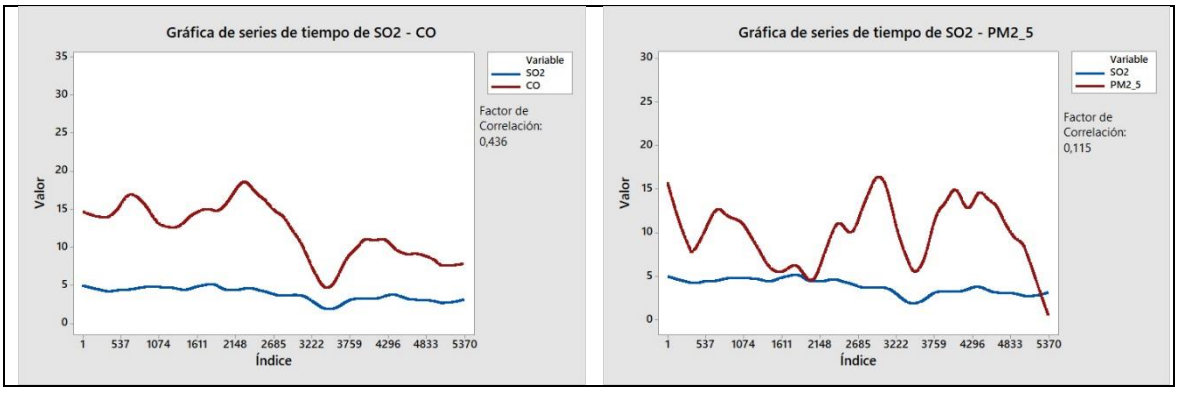


Ilustración 30: Diagramas de series de tiempo y factores de correlación de SO2 obtenidos con EM

En la gráfica anterior se observa los factores de correlación de SO2 obtenidos frente a sus respectivas de series de tiempo, donde se observa que existe una correlación con el CO mucho mayor que con el PM2_5, debido a que le PM2_5 presenta variaciones mucho mayores a las de CO frente al comportamiento uniforme que mantiene el SO2.

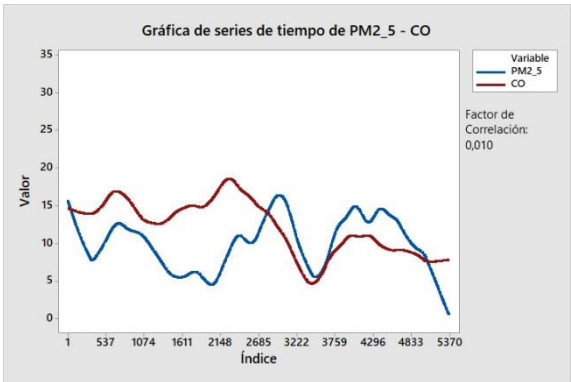


Ilustración 31: Gráfica de series de tiempo y factor de correlación entre PM2_5 y CO obtenidos con EM

En esta última ilustración de los resultados de este algoritmo se muestra el factor de correlación entre el PM2_5 frente a su respectiva gráfica de series de tiempo. Al igual que en todos los casos anteriores, los resultados son muy similares tanto en su valor de factor de correlación, así como el comportamiento que describen las gráficas de series de tiempo.

7.2.4. **Cobweb**

Es un algoritmo de clustering jerárquico de orden divisivo. El resultado de aplicar este método es un árbol donde la raíz constituye la totalidad de los datos y las hojas representan la segmentación de los clústeres. El algoritmo comienza con un único nodo raíz, y en cada iteración, se agregan nuevas instancias hasta obtener un numero de clústeres con una densidad apropiada. La nueva instancia puede ser asignada a un nodo existente, dividir un nodo o generar uno nuevo nodo que alberge dicha instancia (D. Rodríguez, Cuadrado, & Sicilia, 2007).

Resultados

Este algoritmo tardó 6001 ms en completar su ejecución; a diferencia de los anteriores únicamente devuelve el número de clústeres y el número de elementos por cada clúster. En donde el clúster 0 posee el mayor número de elementos con un 80 por ciento de los datos como se observa en la tabla 9 y en la ilustración 32.

	Clúster 0	Clúster 1	Clúster 2	Clúster 3
Ítems	10431	491	1000	1010
%	81%	4%	8%	8%

Tabla 9: Número de elementos por cada clúster obtenido con Cobweb

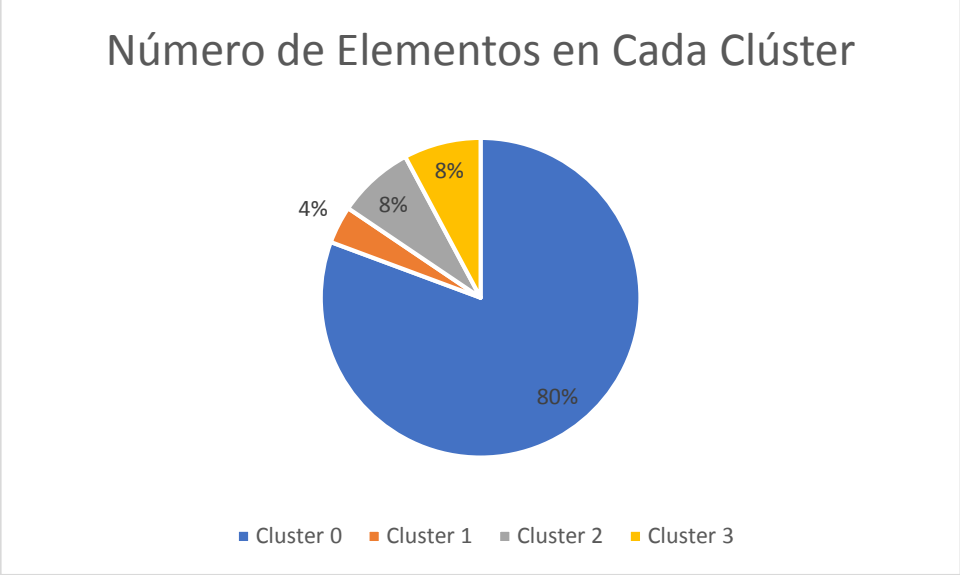


Ilustración 32: Número de elementos por clúster obtenido con Cobweb.

Sobre el clúster más grande (clúster 0), se obtiene los factores de correlación entre cada uno de los contaminantes. A continuación, las ilustraciones muestran dichos valores junto a sus respectivas gráficas de series de tiempo.

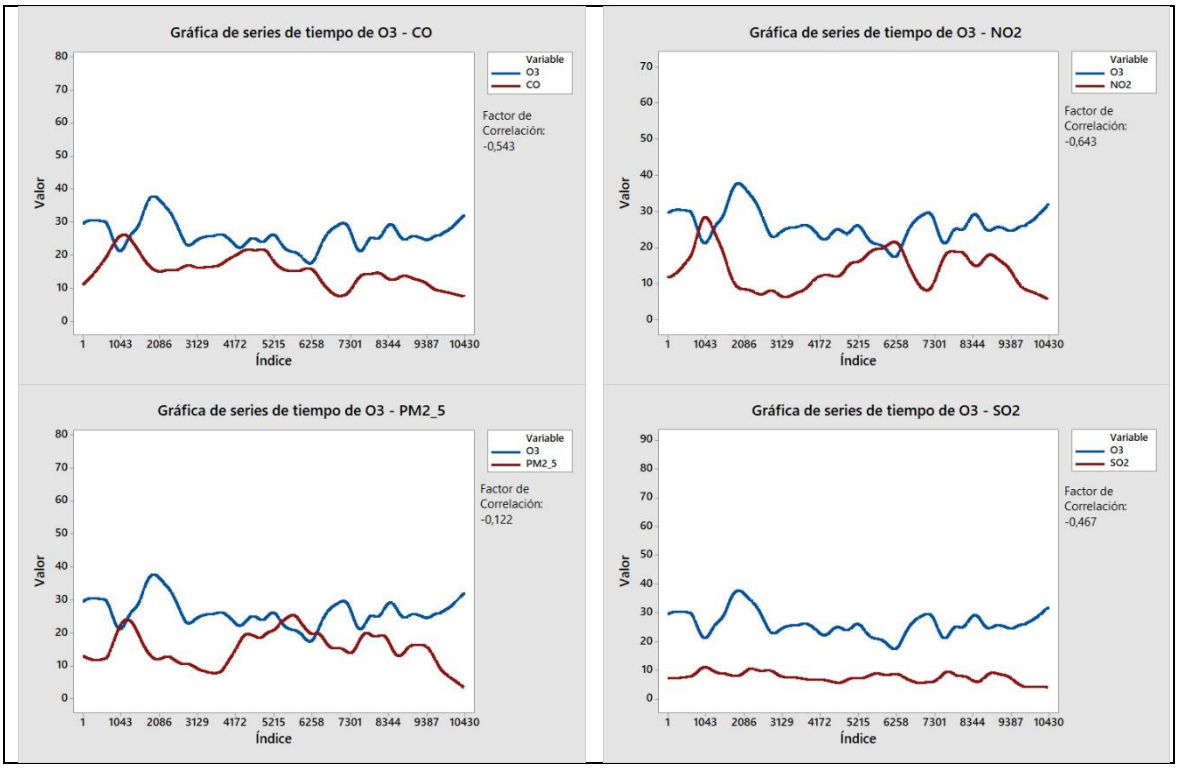


Ilustración 33: Diagramas de series de tiempo y factores de correlación de O3 obtenidos con Cobweb

En la lustración anterior se observa un alto nivel de correspondencia inversa del O3 con el CO, NO2 y SO2 que presentan factores de correlación mayores a 0.45, además, las gráficas de series de tiempo presentan curvas reflejadas. Por otro lado, el PM2_5 presenta un factor de correlación inversa muy bajo, mismo que se observa en la gráfica como en la zona central existe un comportamiento directo entre los 2 componentes.

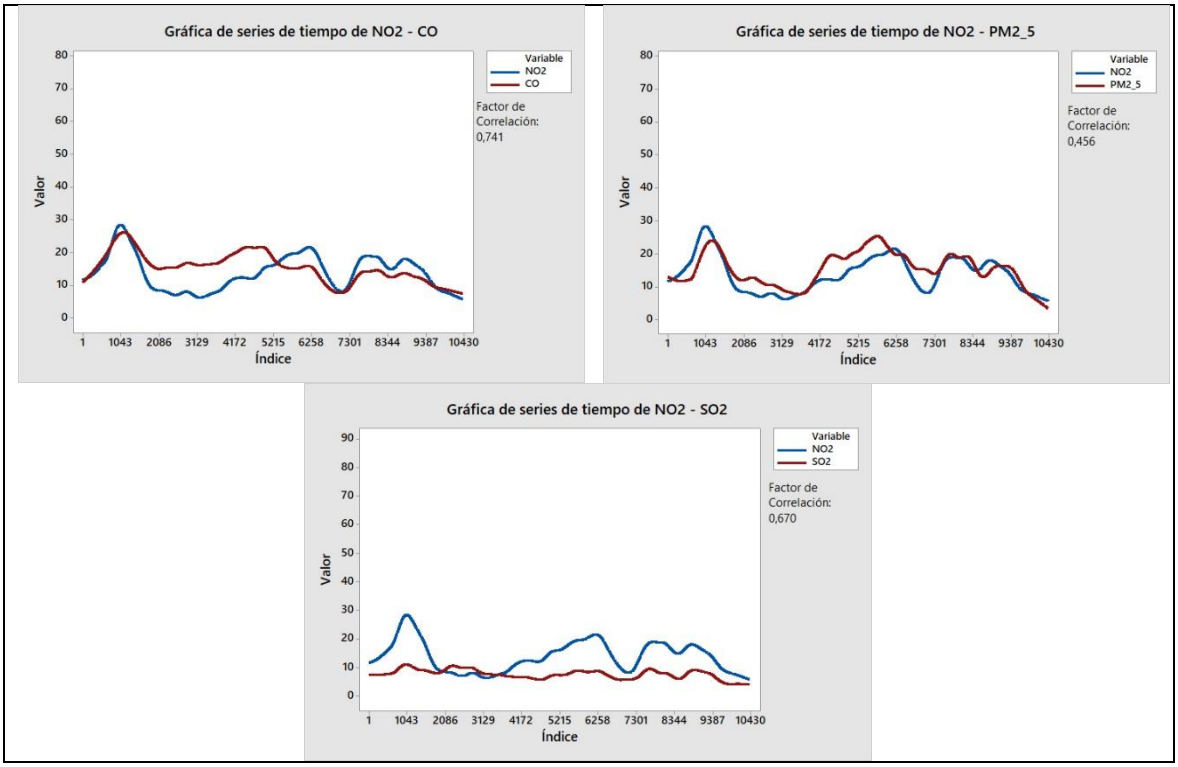


Ilustración 34: Diagramas de series de tiempo y factores de correlación de NO2 obtenidos con Cobweb.

La ilustración 34 muestra las gráficas de serie de tiempo del NO2 frente a los demás contaminantes. Se observa un alto nivel de correlación positiva en todos los casos con valores de los factores de correlación que van desde 0.45 hasta 0.75. Al ser una correlación positiva se observa como las gráficas describen comportamientos similares en toda su extensión.

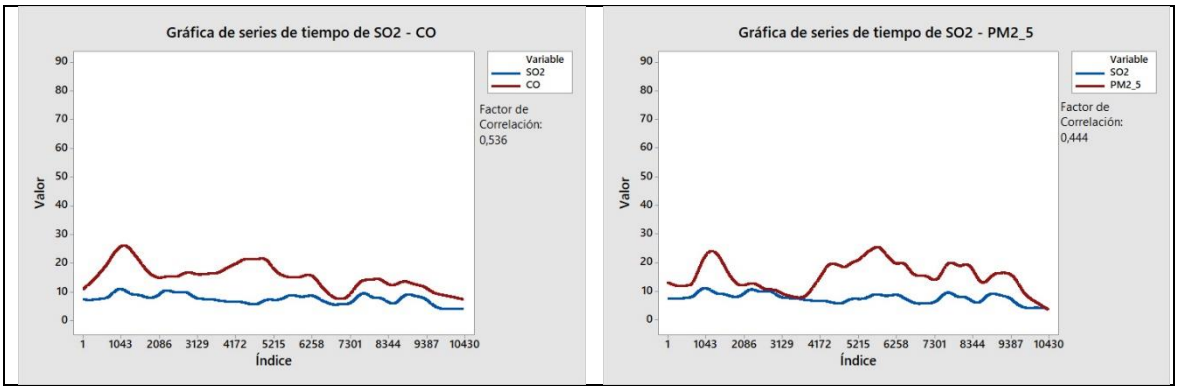


Ilustración 35: Gráficas de series de tiempo y factores de correlación de SO2 obtenidos con Cobweb

La ilustración anterior muestra los resultados obtenidos de los factores de correlación del SO2 frente al CO y PM2_5 y sus gráficas de series de tiempo. Las gráficas se evidencia la gran similitud que presenta el comportamiento de los contaminantes y que se corresponden con los valores de los factores de correlación que se encuentran por encima del 0.44.

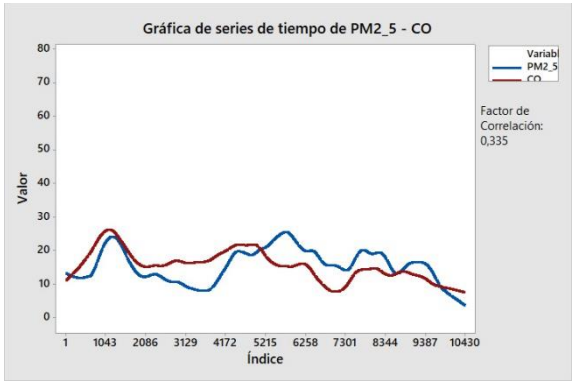


Ilustración 36: Gráfica de series de tiempo y factor de correlación entre PM2_5 y CO obtenido con Cobweb

Como resultado final de este algoritmo se tiene el factor de correlación entre el PM2_5 y el CO junto con su respectiva gráfica de serie de tiempo (ilustración 36). La correlación es positiva y el factor de correlación supera el 0.3, lo que significa un comportamiento similar entre estos contaminantes y que se refleja en la gráfica de serie de tiempo.

7.2.5. **DBSCAN**

Es un algoritmo que agrupa los datos basado en la densidad del clúster, misma que está definida por el número de vecinos de un punto dentro de un área establecida, y dependiendo del número de vecinos se obtiene los *corepoints* y los *borderpoints*. Para que un punto sea considerado un *corepoint* debe tener un mínimo de puntos dentro de su área mientras que los *borderpoints* serán aquellos puntos que si bien se encuentran dentro del vecindario no cumplen con el mínimo de vecinos cercanos y serán los que delimitan el clúster (Neijman, 2011).

Resultados

El tiempo de ejecución de este algoritmo fue de 50974 ms. Al igual que el caso anterior, con DBSCAN se obtiene únicamente el número de clústeres y el número de elementos por cada clúster. En donde el clúster más grande es el clúster 1 que contiene el 99% de los datos como se observa en la tabla 10 y en la ilustración 37.

	Clúster 0	Clúster 1	Clúster 2	Clúster 3
Ítems	22	12839	63	8
%	0,17%	99%	0,49%	0,06%

Tabla 10: Número de elementos por clúster obtenidos con DBSCAN

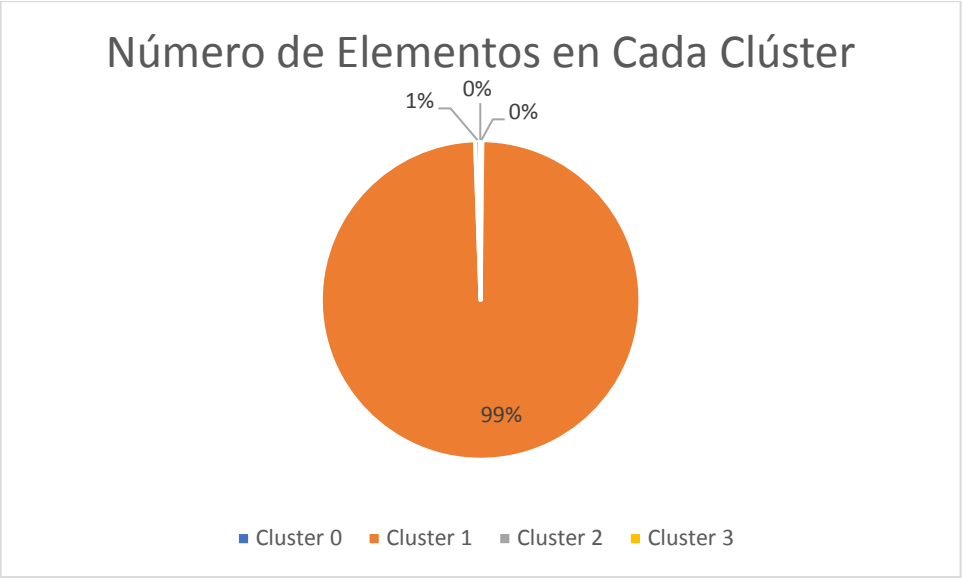


Ilustración 37: Número de elementos por cada clúster obtenido con DBSCAN

Se calcula los factores de correlación entre cada uno de los contaminantes utilizando los datos contenidos en el clúster 1. A continuación se muestra una serie de ilustraciones con los valores de los factores junto a sus respectivas gráficas de series de tiempo.

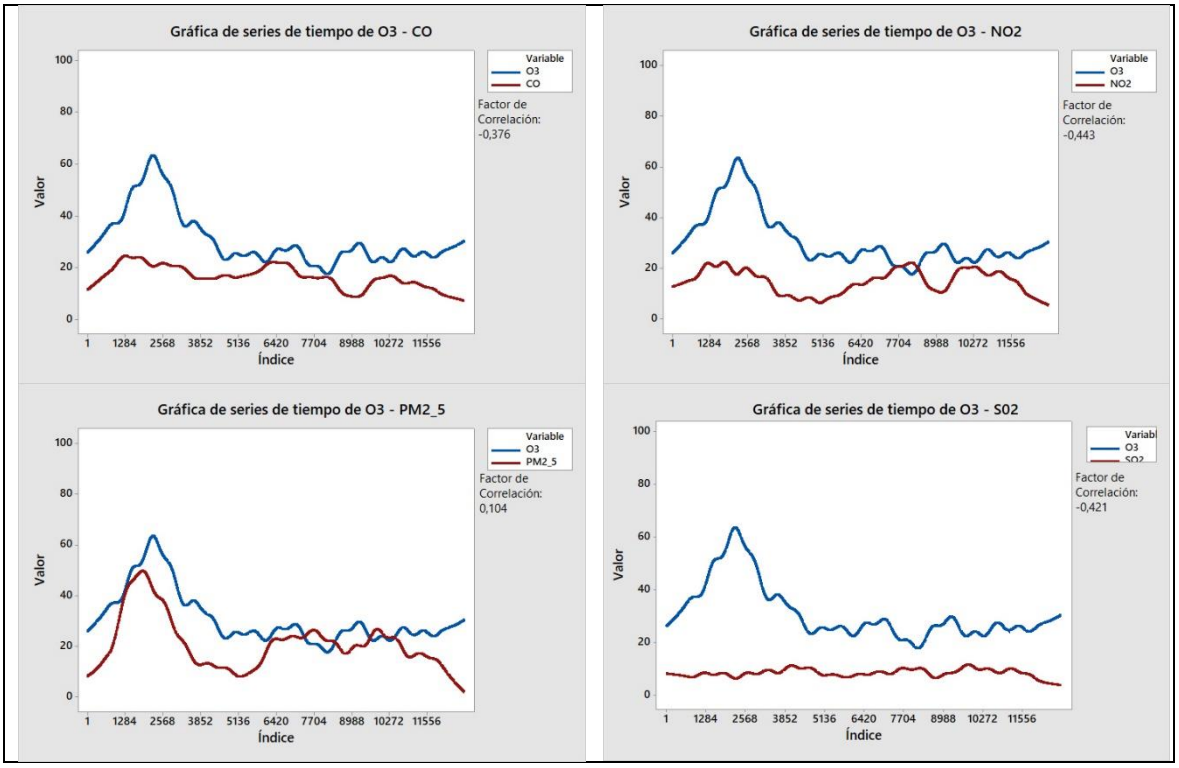


Ilustración 38: Gráficas de series de tiempo y factores de correlación de O3 obtenidos con DBSCAN

En la imagen anterior se muestra los factores de correlación de O3 frente a lo demás contaminantes obtenidos con DBSCAN. Los factores de correlación superan el 0.3 excepto con el PM2_5 el cual tiene un valor de 0.1 mismo que se refleja en su gráfica de series de tiempo al observarse como en la primera mitad presenta un comportamiento prácticamente idéntico mientras que para la segunda mitad se rompe completamente dicho comportamiento.

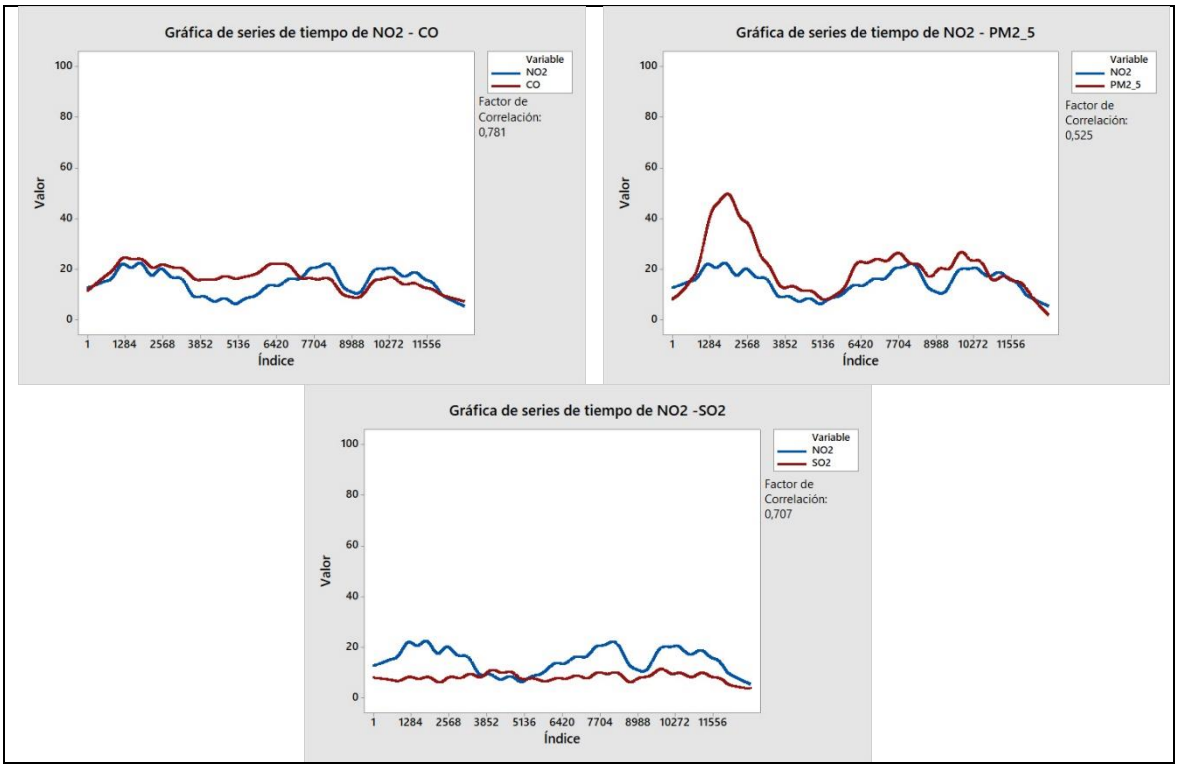


Ilustración 39: Gráficas de series de tiempo y factores de correlación de NO2 obtenidos con DBSCAN

En la ilustración 39 se muestra los factores de correlación de NO2 y sus respectivas gráficas de series de tiempo. En los 3 casos las gráficas describen un comportamiento muy similar a lo largo de todo su trayecto lo que resulta en una correlación directa que se corresponde con los valores de los factores, los cuales superan el 0.5.

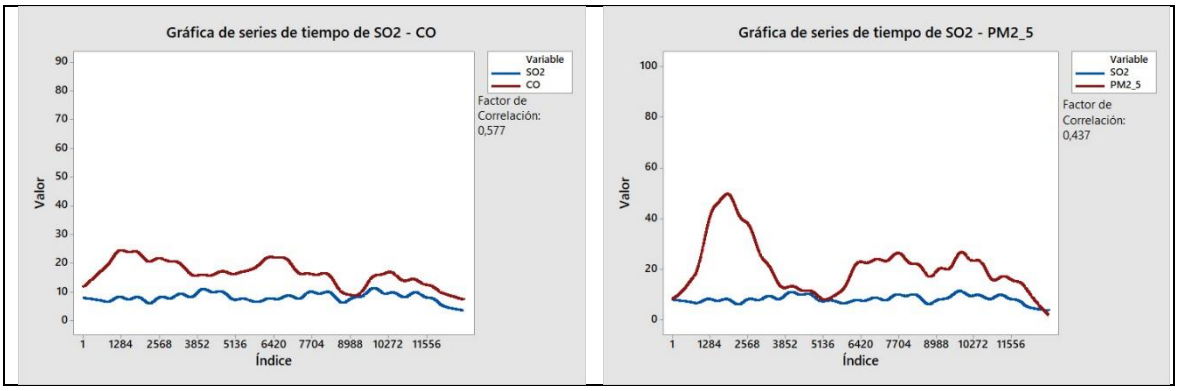


Ilustración 40: Gráficas de series de tiempo y factores de correlación de SO2 obtenidos con DBSCAN

La ilustración 40 muestra los factores de correlación de SO2 junto a sus gráficas de series de tiempo. Se observa como el factor de correlación con CO es superior al que se obtiene con PM2_5, esto se evidencia en las gráficas, ya que con el PM2_5 en la primera cuarta parte del grafico se observa como el nivel de PM2_5 crece mucho más que el CO, y en el resto presentan valores que, si bien no son los mismos, se manejan en valores de entre diez y treinta.

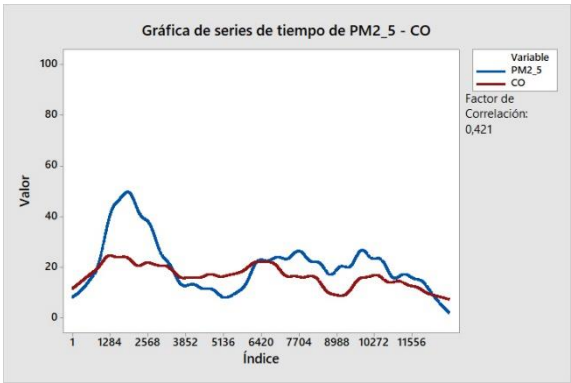


Ilustración 41: Gráfica de series de tiempo y factor de correlación entre PM2_5 y CO obtenido con DBSCAN

Por último, en la ilustración 41 se observa los resultados obtenidos entre el PM2_5 y el CO. Las gráficas de series de tiempo presentan un comportamiento similar excepto en

un pequeño tramo en la zona inicial, en donde se observa como el PM2_5 crece independientemente de los valores de CO.

7.3. Conclusión

Luego de evaluar con éxito todos los algoritmos propuestos, se ha obtenido el tiempo de ejecución del algoritmo y los factores de correlación entre los contaminantes, esta información será analizada decidir cuál es el mejor algoritmo para variables de contaminación del aire.

8. Análisis de Resultados

En esta etapa, se analiza los resultados obtenidos en base a los criterios establecidos en la fase anterior para seleccionar el mejor algoritmo para análisis de variables de contaminación del aire.

La siguiente tabla muestra los factores de correlación entre cada uno de contaminantes, además del tiempo de ejecución de cada algoritmo. Se observa que los factores de correlación son diferentes en todos los algoritmos, y, debido a que no se cuenta con información que valide los resultados no es posible determinar la exactitud o el error de cada algoritmo.

Contaminante	Contaminante	Factor de Correlación				
		K-Means	X-Means	EM	CobWeb	DBSCAN
O3	NO2	-0,30	-0,58	-0,66	-0,64	-0,44
O3	SO2	-0,10	-0,27	0,00	-0,47	-0,42
O3	PM2_5	0,05	0,07	0,14	-0,12	0,10
O3	CO	-0,12	-0,24	-0,38	-0,54	-0,38
NO2	SO2	0,26	0,39	0,10	0,67	0,71
NO2	PM2_5	0,30	0,16	0,21	0,46	0,53
NO2	CO	0,38	0,43	0,47	0,74	0,78
SO2	PM2_5	0,33	0,23	0,11	0,44	0,44
SO2	CO	0,19	0,22	0,44	0,54	0,58
PM2_5	CO	0,05	0,04	0,01	0,33	0,42
Tiempo de Ejecución (seg)		0,5	1,8	112,7	6	51

Tabla 11: Factores de correlación entre contaminantes y tiempos de ejecución obtenidos por cada algoritmo

Por otro lado, considerando el tiempo de ejecución, k-means es el algoritmo con el mejor resultado, obteniendo 500 ms frente a los demás que crecen considerablemente hasta llegar a los 2 minutos con apenas trece mil registros aproximadamente. Por lo tanto, k-means es el mejor algoritmo para variables de contaminación del aire.

8.1. Conclusión

k-means se presenta como el mejor algoritmo debido a que obtiene tiempos de ejecución bastante cortos frente a los demás algoritmos, esto es debido a que trabaja con un número específico de clústeres y se basa únicamente en calcular la distancia de los centroides, mientras que los demás algoritmos aparte de calcular el número de clústeres óptimo realizan procesos más complejos como una validación cruzada como es el caso de EM o determinar la forma del clúster y la densidad como lo hace DBSCAN y Cobweb. Estas operaciones no agregan valor que justifique los tiempos altos de ejecución.

9. Implementación

Luego de obtener el mejor algoritmo para variables de contaminación del aire, se procede a realizar un análisis de los datos utilizando dicho algoritmo en periodos de tiempo más cortos para obtener factores de correlación más precisos en distintas etapas del día.

De acuerdo a las gráficas de series de tiempo obtenidas en la etapa de pruebas se observa que no existe un patrón de comportamiento general de los contaminantes, es decir, pueden existir diferentes patrones de comportamiento entre los mismos contaminantes a lo largo de la muestra, por lo tanto, es necesario realizar un análisis por horas para determinar cuáles son los patrones de comportamiento a lo largo de día.

Las ilustraciones de la 42 a la 45 muestran las gráficas de dispersión de cada uno de los contaminantes en relación al resto, además de sus curvas suavizadas que reflejan el comportamiento general de los datos. En el eje vertical se muestra el valor de los contaminantes y el eje horizontal marca la hora en la que fueron medidos los contaminantes.

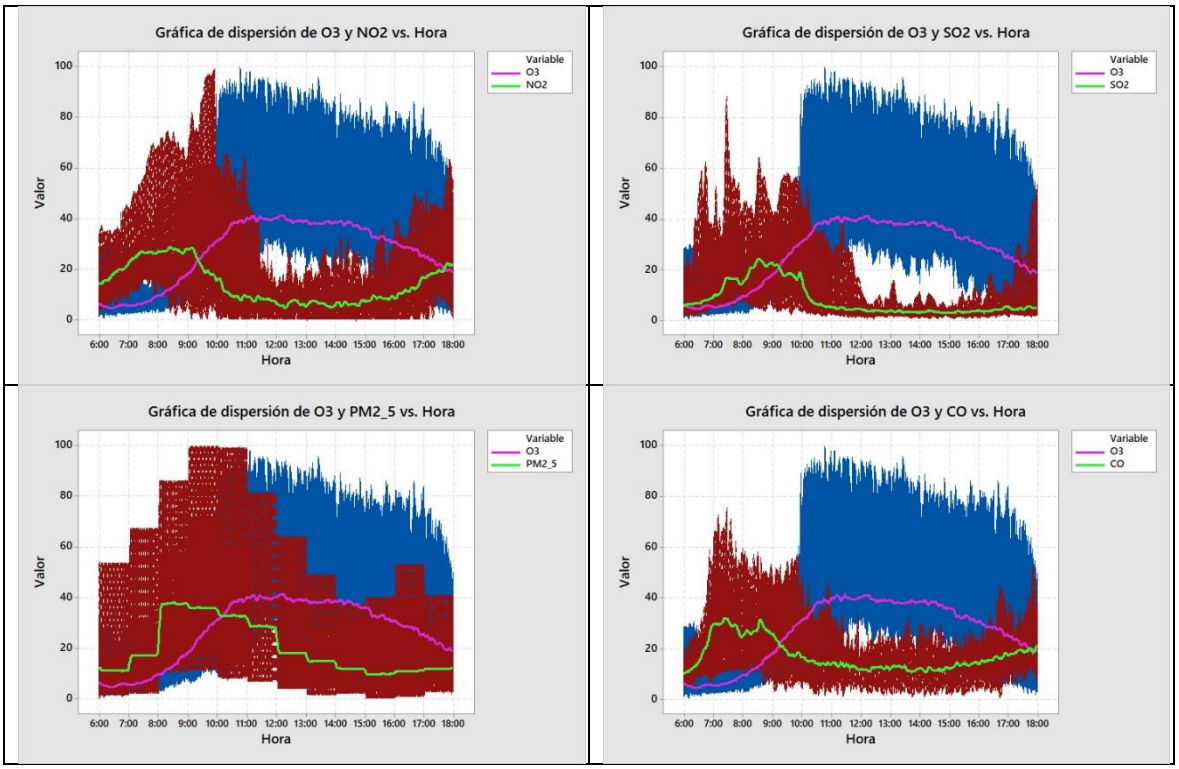


Ilustración 42: Gráficas de dispersión de O3 y curvas suavizadas

La ilustración 42 muestra las gráficas de dispersión de O3 frente a los demás contaminantes. La zona azul representa los valores medidos del contaminante principal y la zona roja representa al contaminante respectivo con el cual es comparado en cada gráfica. Suavizando las curvas se observa como en la mañana los contaminantes varían considerablemente, luego, al medio día sus niveles se mantienen o varían muy poco, y al atardecer, nuevamente se produce un cambio brusco en su comportamiento. Tomando como referencia estas gráficas, se ha establecido tres rangos horarios dentro de los cuales se obtienen factores de correlación más precisos, esto es: de 8 a 10 de la mañana, de 11 a 1 de la tarde y de 4 a 6 de la tarde. La siguiente tabla muestra los valores obtenidos en cada rango. Existe una fuerte correlación inversa con el NO2 y CO tanto en la mañana como en la tarde ya que los factores superan 0.4. Con el SO2 únicamente existe una correspondencia en la mañana con un factor de correlación del 0.35.

Contaminante	Contaminante	Factores de correlación		
		8-10 AM	11-1 PM	4-6 PM
O3	NO2	-0,58	-0,18	-0,60
O3	SO2	-0,36	-0,04	-0,10
O3	PM2_5	0,17	0,27	-0,01
O3	CO	-0,59	0,13	-0,43

Tabla 12: Factores de Correlación de O3 obtenidos con k-means

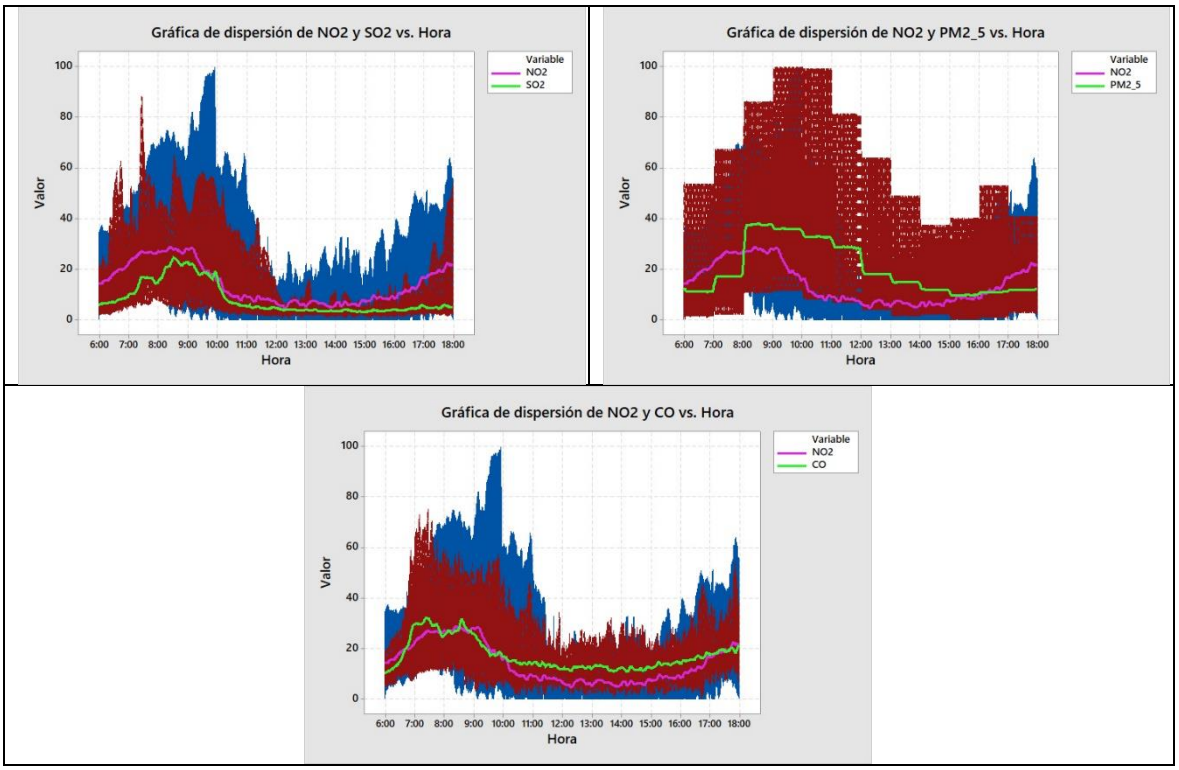


Ilustración 43: Gráficas de dispersión de NO2 y curvas suavizadas

La ilustración 43 muestra las gráficas de dispersión de NO2 frente a los demás contaminantes y sus respectivas curvas suavizadas. Al igual que en el caso anterior, los contaminantes presentan tres escenarios distintos en la mañana, medio día y al atardecer. Utilizando los datos de las mismas horas que en el caso anterior se obtiene los factores de correlación respectivos que se encuentran en la tabla 13. Los contaminantes están

fuertemente correlacionados ya que únicamente se observa valores menores a 0.15 con el PM2_5 en la mañana y tarde, con el SO2 al medio día. En las demás etapas los factores van desde 0.3 hasta 0.7 como es el caso del CO y SO2 en la mañana, lo cual indica una fuerte correlación.

Contaminante	Contaminante	Factores de correlación		
		8-10 AM	11-1 PM	4-6 PM
NO2	SO2	0,63	0,09	0,31
NO2	PM2_5	-0,13	0,47	0,12
NO2	CO	0,72	0,35	0,38

Tabla 13: Factores de correlación de NO2 en distintos periodos del día

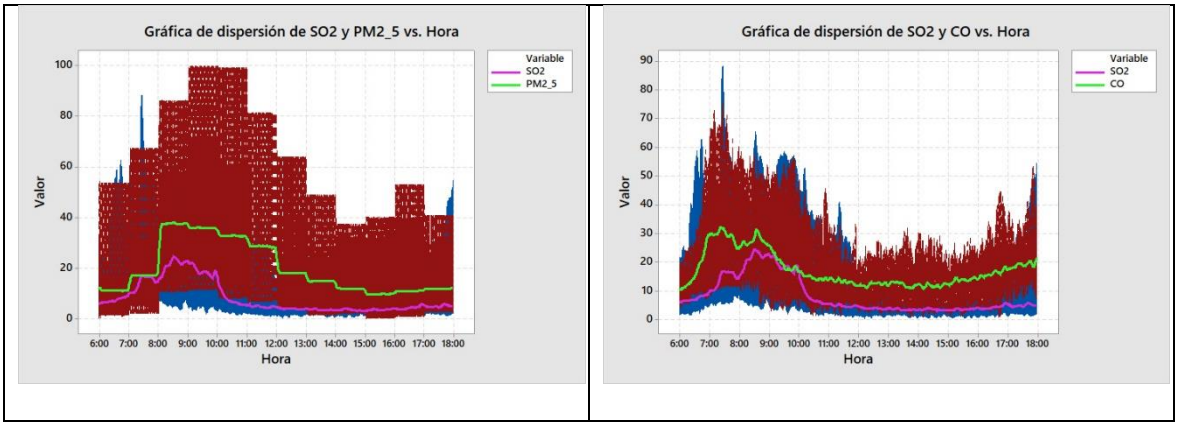


Ilustración 44: Gráficas de dispersión de SO2 y curvas suavizadas

La ilustración 44 muestra las gráficas de dispersión de SO2 frente al PM2_5 y CO. Se observa distintos escenarios en las mismas fracciones de tiempo que en los casos anteriores, por lo tanto, se obtiene los factores de correlación en las mismas horas como se muestra en la tabla 14. Existe únicamente una correlación positiva con CO en la mañana y al medio día con factores de correlación de 0.5 y 0.4 respectivamente.

Contaminante	Contaminante	Factores de correlación		
		8-10 AM	11-1 PM	4-6 PM
SO2	PM2_5	-0,08	-0,17	-0,20
SO2	CO	0,49	0,38	0,07

Tabla 14: Factores de correlación de SO2 en distintas etapas del día

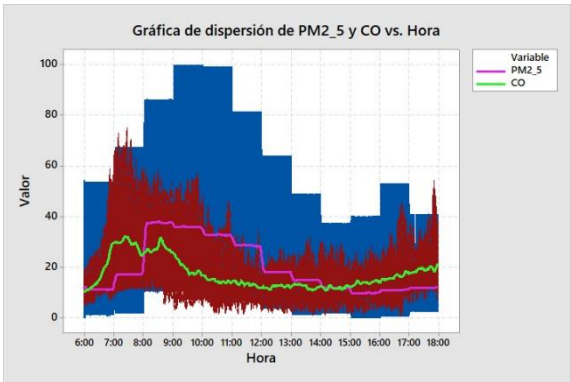


Ilustración 45: Gráfica de dispersión de PM2_5 y CO con sus curvas suavizadas

Por último, la ilustración 45 muestra la gráfica de dispersión de PM2_5 frente a CO, la cual presenta los mismos escenarios que el caso anterior. A continuación, la tabla 15 muestra los factores de correlación obtenidos en cada periodo. Se observa que los valores de los factores son bajos al medio día y en la tarde, por lo tanto, no existe una correlación entre dichos contaminantes. Por otro lado, si bien entre las ocho y diez de la mañana el factor rodea el 0.2 su correlación sigue siendo leve.

Contaminante	Contaminante	Factores de correlación		
		8-10 AM	11-1 PM	4-6 PM
PM2_5	CO	-0,21	0,09	0,07

Tabla 15: Factores de correlación de PM2_5 y CO en distintas etapas del día

Con los resultados obtenidos se genera la tabla 16, en donde se muestra a modo de resumen todos los factores de correlación que existen entre los distintos contaminantes en tres epatas del día que representan: la mañana, medio día y en la tarde.

Factores de correlación obtenidos por k-means				
Contaminante	Contaminante	Factor de Correlación		
		8-10 AM	11-1 PM	4-6 PM
O3	NO2	-0,55	-0,18	-0,60
O3	SO2	-0,36	-0,04	-0,10
O3	PM2_5	0,17	0,27	-0,01
O3	CO	-0,59	0,13	-0,43
NO2	SO2	0,63	0,09	0,31
NO2	PM2_5	-0,13	0,47	0,12
NO2	CO	0,72	0,35	0,38
SO2	PM2_5	-0,08	-0,17	-0,20
SO2	CO	0,49	0,38	0,07
PM2_5	CO	-0,21	0,09	0,07

Tabla 16: Factores de correlación de todos los contaminantes en distintas etapas del día

Al final, haciendo uso del operador "*Select by Weights*" es posible determinar el grado de incidencia que tiene cada contaminante sobre los demás elementos. La tabla 17 muestra dichos valores en un rango de 0 y 1, donde a mayor valor, mayor será el grado de incidencia de un elemento. Entonces, O3 es el contaminante que influye en gran medida sobre los demás contaminantes en los 3 escenarios del día; el NO2 Influye únicamente sobre el CO al medio día; el SO2 influye en menor medida sobre el NO2, PM2 y CO al medio día; El PM2_5 influye sobre el SO2, CO y NO2 en la mañana, sobre el NO2 y CO al medio día y sobre el NO2, SO2 y CO en la tarde. Por último, se observa que el CO mantiene un nivel de influencia muy bajo en todo el día.

Contaminante	Peso		
	8-10 AM	11-1 PM	4-6 PM
O3	1	1	1
NO2	0	0,294	0
SO2	0,001	0,888	0,102
PM2_5	0,460	0,381	0,172
CO	0,134	0	0,088

Tabla 17: Pesos de los contaminantes en distintas etapas del día

10. Conclusiones y Trabajos futuros

Fueron comparados cinco algoritmos de clusterización en el estudio: K-means, X-means, Cobweb, Maximization Expectation y DBSCAN; con un *dataset* de 13000 registros aproximadamente, donde, luego de realizar todas las pruebas planteadas se ha determinado que el mejor algoritmo de minería de datos para variables de contaminación del aire es k-means, ya que con este algoritmo obtiene resultados en un tiempo de ejecución muy corto.

Utilizando gráficas de series de tiempo se identificó que existen distintos patrones de comportamiento entre los contaminantes, por lo tanto, no es posible establecer un factor de correlación que refleje el comportamiento de los datos en toda la muestra.

Utilizando gráficas de dispersión se identificó tres escenarios idóneos para obtener valores más precisos: en la mañana de ocho a diez, al medio día de once a una y en la tarde de cuatro a seis. Estos son periodos de tiempo en los cuales los datos cambian su comportamiento y por lo tanto cambia la correlación entre las variables.

En la mañana, la mayoría de los contaminantes presentan una correlación alta excepto con el material particulado el cual tiene factores de correlación igual o menores a dos. Al medio día existe una correlación en NO2 - PM2_5, SO2 – CO y NO2 - CO ya que sus

factores de correlación se encuentran cercanos al 0.5. En la tarde existe una fuerte correlación entre O3 - NO2, NO2 - SO2 - CO con factores superiores a 0.2. Además, se logró establecer que contaminantes influyen sobre los demás, siendo el O3 el que tiene mayor grado de incidencia en el comportamiento de los demás contaminantes.

Se observa que existe un cambio considerable en todas las variables tanto en la mañana como en la tarde, lo que indica que existen factores externos que afectan el comportamiento de los contaminantes por lo tanto sería necesario involucrar variables meteorológicas como por ejemplo el viento y la temperatura para obtener mejores resultados.

Bibliografía

Anu, J., Divyadharshini, M., & Science, C. (2017). Air Pollution Analysis using Clustering Algorithms, 110–113.

Athanasiadis, I. N., & Kaburlasos, V. G. (2006). Air Quality Assessment Using Fuzzy Lattice Reasoning (FLR). *Fuzzy Systems, 2006 IEEE International Conference on*, (4), 29–34. https://doi.org/10.1109/fuzzy.2006.1681690

Crowe, C. (2011). Entornos invisibles. *Ministerio de Educación. Instituto Nacional de Educación Tecnológica.*, 3–45. Retrieved from http://www.inet.edu.ar/wp-content/uploads/2012/11/C1_Parque_de_diversionesR.pdf

Dan Pelleg, A. M. (2015). X-means. *CEUR Workshop Proceedings*, *1542*, 33–36. https://doi.org/10.1017/CBO9781107415324.004

Du, X. (2016). Mining PM2 . 5 and Traffic Conditions for Air Quality, 33–38.

González, F. (2013). Aplicación de técnicas de Minería de Datos a datos obtenidos por el Centro Andaluz de Medio Ambiente (CEAMA), 1–68.

Lagares, P., & Puerto, J. (2001). Población y muestra. Técnicas de muestreos. *Management Mathematics for European Schools*, 1–19. https://doi.org/10.4067/S0071-17132000003500023

Neijman, N. (2011). DBSCAN A Density-Based Spatial Clustering of Application with. *Linköpings Universitet - ITN 2011-11-30*, 1–8.

Raipure, S., & Mehetre, D. (2015). Wireless sensor network based pollution monitoring system in metropolitan cities. *2015 International Conference on Communications and Signal Processing (ICCSP)*, 1835–1838. https://doi.org/10.1109/ICCSP.2015.7322841

Red de Monitoreo de Calidad del Aire de Cuenca de la EMOV EP. (2015). Informe de la Calidad del Aire de Cuenca, 124. https://doi.org/10.1017/CBO9781107415324.004

Riquelme, J. C., Ruiz, R., & Gilbert, K. (2006). Miner??a de datos: Conceptos y tendencias. *Inteligencia Artificial*, *10*(29), 11–18.

Rodríguez, D., Cuadrado, J., & Sicilia, M. (2007). Comparación de diferentes algoritmos de clustering en la estimación de coste en el desarrollo de software. *Ingeniería Del Software*, *3*(1), 6–22. Retrieved from http://en.scientificcommons.org/44226406

Rodríguez, O. (n.d.). Metodología para el Desarrollo de Proyectos en Minería de Datos CRISP-DM, 12. Retrieved from http://www.oldemarrodriguez.com/yahoo_site_admin/assets/docs/Documento_CRISP-DM.2385037.pdf

Sahafizadeh, E., & Ahmadi, E. (2009). Prediction of Air Pollution of Boushehr City Using Data

Mining. *2009 Second International Conference on Environmental and Computer Science*, 33–36. https://doi.org/10.1109/ICECS.2009.18

Sellers, C. A. (2013). Publicación de los contaminantes atmosféricos de la estación de monitoreo en tiempo real de la ciudad de Cuenca, utilizando servicios estándares OCG. *Journal of Chemical Information and Modeling*, *53*(9), 1689–1699. https://doi.org/10.1017/CBO9781107415324.004

Taneja, S., Sharma, N., Oberoi, K., & Navoria, Y. (2016). Predicting Trends in Air Pollution in Delhi using Data Mining, 1–6.

Bifet, A. (2013). Data mining in real time. *Informatica37*.

Cabrera, A. (2017). *Instituto de Estudios de Régimen Seccional del Ecuador.* Cuenca: Sistemas de Infomación Geofráfica.

Duque, N. D. (2011). MIneria de Datos para el analisis de datos metereologicos. *Departamento de Informatica y Computacion, Universidad Nacional de Colombia Sede Manizales*.

EMOV-EP, R. d. (2015). Informe de la Calidad del Aire de Cuenca.

Fernandez-Camacho, R., Brito Cabeza, I., Aroba, J., Gomez-Bravo, F., Rodriguez, S., & de la Rosa, J. (2015). Assessment of ultrafine particles and noise measurements using fuzzy. *Science of the Total Enviroment*.

García, F. (2013). Aplicación de técnicas de minería de datos a datos obtenidos por el centro de Andaluz de medio ambiente. *Universidad de Granada*.

Gorbea, P. (2013). Tendencias transdiciplinarias en los estudios métricos de la información y su relación con la gestión de la información y del conocimento. *Perspectivas em Gestao & Conhecimento*.

Lima, F. (2017). Universidad del Azuay: LIDI.

Medina, F., & Gómez, C. (2014). Funcionalidades de la minería de datos. *Revista Ingeniería y Región*.

Office-of-air-and-radiation. (2003). Air Quality Index A guide to air quality and your health. *United States Environmental Protection Agency*.

Office-of-air-and-radiation. (2011). Air Quality Guide for Nitrogen Dioxide. *United Estates Evironmental Protection Agency*.

Oficina-de-Calidad-del-Aire-y-Radiación. (2015). Guía de la calidad del aire sobre el ozono. *Agencia de los EEUU para la protección del medio ambiente*.

Oficina-de-calidad-del-aire-y-radiacion. (2015). Guía de la calidad del aire sobre la contaminación por partículas. *Agenmcia de los EEUU para la protección del medio ambiente*.

Red_de_Monitoreo_del_Calidad_del_Aire_de_Cuenca, E. (2015). *Informe de la Calidad del Aire de Cuenca.* Cuenca.

Roiger, R. J. (2017). *Data Mining: A Tutorial-Based Primer.* Boca Raton - Florida: CRC Press.

Sellers Walden, C. A. (2012). Publlicación de Contaminantes Atmosféricos de la Estación de Monitoreo de la Ciudad de Cuenca, Utlizando Servicios Estándares OGC. *Universidad de Azuay*.

Sellers, C. A. (2013). Publicacion de los contaminantes atmosfericos de la estacion de monitoreo en tiempo real de la ciudad de Cuenca, utilizando servicios estandares OGC. *Universidad del Azuay*.

Snee, R. D. (2015). A Practical Approach to Data Mining: I Have All This Data: What Should I Do? *Quality Engineering*.

Snehal, S., Manoj, C., & Shweta, G. (2014). Implementation of WSN Based Air Pollution Monitoring Sysem using Data Mining Technique. *International Journal of Computer Scince and Mobile Computing*.

Tulas. (Libro VI). Normas de Calidad del Aire Ambiente. *Anexo 4*.

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: **"IMPACTO DE LA APLICACIÓN DE ALGORITMOS DE MINERÍA DE DATOS EN VARIABLES DE CONTAMINACIÓN DE AIRE"**, presentado por el estudiante **John Javier Ortega Guamán**, previa a la obtención del grado de **Ingeniero en Sistemas y Telemática**, para el día **LUNES 15 DE MAYO DE 2017 A LAS 07h20,** La sustentación se realizará en el **laboratorio del IERSE.**

Cuenca, 10 de mayo de 2017

Dra. Jenny Ríos Coello
Secretaria de la Facultad

Ing. Marcos Orellana Cordero .....

Ing. Chester Sellers Walden .....

Dr. Francisco Salgado Arteaga .....

mjmr/

Comunicado OK
12-05-2017.

Oficio Nro. 060-2017-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 "Impacto de la aplicación de algoritmos de minería de datos en variables de contaminación de aire", presentado por John Javier Ortega Guamán estudiante de la Escuela de Ingeniería de Sistemas y Telemática, y revisado por el Ing. Marcos Orellana, 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. Marcos Orellana y como miembros del Tribunal a Francisco Salgado Ph.D. e Ing. Chester Sellers.

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: John Javier Ortega Guamán**
1.2 **Director sugerido**: Ing. Marcos Orellana Cordero
1.3 **Codirector (opcional)**: _____
1.4 **Tribunal:** Ing. Chester Sellers Walden/ Dr. Francisco Salgado Arteaga
1.5 **Título propuesto: "IMPACTO DE LA APLICACIÓN DE ALGORITMOS DE MINERÍA DE DATOS EN VARIABLES DE CONTAMINACIÓN DE AIRE"**
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. Marcos Orellana Cordero

1.6.4 No aceptado
- Justificación:

_____
_____
_____

Tribunal

Ing. Marcos Orellana Cordero

Ing. Chester Sellers Walden

Dr. Francisco Salgado Arteaga

Sr. John Javier Ortega Guamán

Dra. Jenny Ríos Coello
Secretario de Facultad

Fecha de sustentación: día **LUNES 15 DE MAYO DE 2017 A LAS 07h20**

UNIVERSIDAD DEL AZUAY

RÚBRICA PARA LA EVALUACIÓN DEL PROTOCOLO DE TRABAJO DE TITULACIÓN

1.1 Nombre del estudiante: John Javier Ortega Guamán
1.2 Director sugerido: Ing. Marcos Orellana Cordero
1.3 Codirector (opcional):
1.4 Título propuesto: "IMPACTO DE LA APLICACIÓN DE ALGORITMOS DE MINERÍA DE DATOS EN VARIABLES DE CONTAMINACIÓN DE AIRE"
1.5 Revisores (tribunal): Ing. Chester Sellers Walden/ 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. Marcos Orellana Cordero

Ing. Chester Sellers Walden

Dr. Francisco Salgado Arteaga

UNIVERSIDAD DEL AZUAY

Escuela Sistemas y Telemática

Oficio Estudiante: Solicitud aprobación de Protocolo de Trabajo de Titulación

IST-RE-EST-02
Versión 01
04/04/2017
Página 1 de 1

Lugar de Almacenamiento	Retención	Disposición Final
F: Archivo Secretaría de la Facultad	5 años	Almacenar en archivo pasivo de la Facultad

Cuenca, 11 de Mayo del 2017

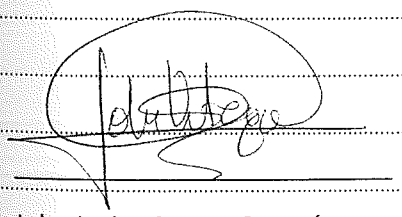
Ingeniero,
Oswaldo Merchán Manzano
DECANO DE LA FACULTAD DE CIENCIAS DE LA ADMINISTRACIÓN
UNIVERSIDAD DEL AZUAY

De mi consideración,

Estimado Señor Decano, yo **John Javier Ortega Guamán** con C.I. **0301689170**, código estudiantil **69259**, 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 **"Impacto de la aplicación de algoritmos de minería de datos en variables de contaminación del aire"** previo a la obtención del título de Ingeniero en Sistemas y Telemática para lo cual adjunto la documentación respectiva.

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

Atentamente:



John Javier Ortega Guamán

Estudiante de la Carrera de Sistemas y Telemática

Edición autorizada de 10.000 ejemplares
Del 788.501 al 798.500
Nº 0793814

UNIVERSIDAD DEL AZUAY

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

C E R T I F I C A:

Que, el Señor **Jhon Javier Ortega Guamán** registrado con código **69259** estudiante de la Escuela de Ingeniería de Sistemas y Telemática tiene aprobado más del 80% de su Pensum de estudios.

Que, el Señor **Jhon Javier Ortega Guamán** le falta aprobar las siguientes asignaturas para finalizar sus estudios:

METODOLOGIA DE LA INVESTIGACION

INGENIERIA DE SOFTWARE II

DESARROLLO Y GESTION DE PROYECTOS INFORMATICOS

PROYECTOS TELEMATICOS

SISTEMA DE INFORMACION GERENCIAL

PRODUCCION II

IDIOMA EXTRANJERO (15 CREDITOS)

Cuenca, 11 de Mayo de 2017

FACULTAD DE ADMINISTRACION SECRETARIA

Derecho 69529
vcf.-

Edición autorizada de 10.000 ejemplares
Del 788.501 al 798.500
Nº 0796022

Lugar de Almacenamiento	Retención	Disposición Final
F: Archivo Secretaría de la Facultad	5 años	Almacenar en archivo pasivo de la Facultad

Cuenca, 11 de Mayo del 2017

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

De mi consideración,

Yo, **Marcos Patricio Orellana Cordero** informo que he revisado el protocolo de trabajo de titulación previo a la obtención del título de Ingeniero en Sistemas y Telemática, denominado **"Impacto de la aplicación de algoritmos de minería de datos en variables de contaminación del aire"**, realizado por el estudiante **John Javier Ortega Guamán**, con código estudiantil 69259, 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

Ing. Marcos Orellana Cordero

UNIVERSIDAD DEL AZUAY

UNIVERSIDAD DEL AZUAY
INGENIERIA EN SISTEMAS Y TELEMATICA
DISEÑO DE TESIS

1. DATOS GENERALES

1.1 Nombre del estudiante: John Javier Ortega Guamán

1.1.1 Código: 69259

1.1.2 Contacto:

Teléfono convencional: 4098458

Celular: 0995954510

Correo electrónico: johnortega2501@gmail.com

1.3 Director sugerido: Orellana Cordero, Marcos Ing.

1.2.1 Contacto:

Teléfono convencional: 4128640

Celular: 0999955611

Correo electrónico: marore@uazuay.edu.ec

1.3 Co-director sugerido:

1.4 Asesor metodológico:

1.5 Tribunal designado:

1.6 Aprobación: Junta Académica:

Consejo de Facultad:

1.7 Línea de Investigación de la carrera:

1.7.1 Código UNESCO: 1203 Informática de computadores

1203.17 Informática

Edición autorizada de 10.000 ejemplares
Del 788.501 al 798.500 Nº 0797227

1.7.2 Tipo de trabajo: Investigación.

1.8 Área de estudio: Minería de Datos.

1.9 Título propuesto: Impacto de la aplicación de algoritmos de minería de datos en variables de contaminación del aire.

1.10 Subtítulo:

1.11 Estado del proyecto: El proyecto consiste en un análisis de tecnologías existentes de minería de datos, para que en función de los resultados obtenidos, determinar cuáles son los mejores algoritmos para identificar relaciones entre contaminantes atmosféricos.

2. CONTENIDO

2.1 Motivación de la investigación:

La contaminación atmosférica representa un peligro para la salud de las personas y la ciudadanía se expone diariamente a este riesgo, por lo que es necesario analizar los niveles de contaminación de las grandes ciudades para tomar medidas para evitar o reducir los niveles de contaminación y, de esta forma garantizar a la ciudadanía un aire limpio y seguro.

2.2 Problemática:

En la actualidad existe un proyecto en la ciudad el cual se encarga de medir y publicar los niveles de contaminación a través de una página web. El presente trabajo tiene como propósito mejorar dicho proyecto implementando algoritmos de *data mining,* ya que la estación de monitoreo envía los datos medidos del aire cada segundo, lo que genera grandes volúmenes de información que deben ser procesados en el menor tiempo posible. Para esto, es necesario analizar un conjunto de algoritmos de *data mining* y determinar cuál es el algoritmo idóneo para el trabajo.

UNIVERSIDAD DEL AZUAY

2.3 Pregunta de investigación:

¿Cuáles son los algoritmos más relevantes de minería de datos para determinar la relación entre variables de contaminación atmosférica?

2.4 Resumen:

Este trabajo se orientará al descubrimiento de los mejores algoritmos para el análisis sobre los datos recopilados de manera sistemática por una estación de monitoreo de la calidad del aire. Se evaluará el comportamiento de los algoritmos para el análisis de 5 variables ambientales recogidas de la ciudad por un sistema de monitoreo. 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). Los algoritmos descubrirán patrones de comportamiento que luego establecerán correlaciones entre las variables de estudio; será necesario experimentar con varios algoritmos de DM (*Data mining*), lo que resultará en conocimiento para contrastar hipótesis o elaborar indicadores de gestión y predicción.

2.5 Indagación Exploratoria:

La contaminación del aire ha sido calificada por la OMS (Organización Mundial de la Salud) como uno de los grandes causantes de cáncer en los seres humanos, además, en varios estudios se ha demostrado que la contaminación ambiental es la principal causa de muertes por problemas respiratorios y cardiovasculares (Fernandez-Camacho, y otros, 2015).

El constante crecimiento del tráfico vehicular y la industrialización son unas de las principales fuentes de contaminación de las ciudades. Todos los vehículos de combustión interna

Edición autorizada de 10.000 ejemplares
Del 788.501 al 798.500 Nº 0797226

generan gases contaminantes y a esto se le suma los gases que son despedidos por las industrias que se encuentran en las cercanías de las zonas pobladas. Además, la falta o el incumpliendo de la normativa para regular y controlar la composición de los combustibles incrementa considerablemente el impacto que tienen sobre el medio ambiente. (Fernandez-Camacho, y otros, 2015).

Para realizar cualquier análisis del aire es necesario recopilar datos del medio ambiente a través de una estación de monitoreo que mide varios tipos de contaminantes.

La estación de monitoreo utilizada en la ciudad de cuenca mide 5 tipos de contaminantes: Ozono (O3), Monóxido de Carbono (CO), Dióxido de Azufre (SO2), Dióxido de Nitrógeno (NO2) y Material Particulado (PM2.5). Estos son los 5 principales elementos que aportan a la contaminación de la atmósfera. La estación tiene un rango de cobertura de 4 km de radio y se encuentra ubicada en los altos del edificio de la Empresa de Movilidad (EMOV-EP), cuyas coordenadas son 721895.66 E, 9679548.85 N (Sellers Walden, 2012). La estación también registra la precipitación, la radiación solar global, la velocidad y dirección del viento. Toda esta información es almacenada cada segundo, lo que proporciona un gran volumen de datos (Red_de_Monitoreo_del_Calidad_del_Aire_de_Cuenca, 2015).

En la actualidad la cantidad de información que se genera cada día supera los 5 exabytes y deben ser almacenados para estudiarlos, es lo que se llama Big Data. Estos lotes no pueden ser tratados por los métodos de procesamiento de información tradicionales, con lo que se desarrolló nuevos métodos capaces procesar grandes lotes de datos y obtener información, de esta manera surge lo que hoy se conoce como minería de datos (Bifet, 2013).

Según los autores Medina y Gómez los algoritmos en la minería de datos se clasifican de la siguiente manera:

- Exploración: Utilizada para encontrar relaciones sistemáticas entre las variables cuando se tiene poco o nada de conocimiento sobre los resultados próximos. Este método solo funciona como una primera etapa para la predicción.
  - De Asociación
  - Agrupamiento
- Estadísticos: Utilizados cuando se trata de encontrar una función matemática que mejor relacione los datos. Busca corregir datos faltantes en una muestra, clasificar o realizar procesos de predicción.
  - Regresión Lineal
  - Regresión Logística
  - Redes Bayesianas
  - Bayes Naive
- Clasificación de Datos: Basado en el aprendizaje inductivo.
  - Árboles de Decisión
  - Reglas de decisión
- Aprendizaje Automático: Combina la inteligencia artificial con la estadística para establecer una noción general de inferencia.
  - Lógica de Inferencia Difusa
  - Redes Neuronales

Edición autorizada de 10.000 ejemplares
Del 788.501 al 798.500 Nº 0797225

- Serie Temporal: Sirve para generar predicciones de tiempo en valores continuos, basándose únicamente en los datos que se utilizaron para crear el modelo. (Medina & Gómez, 2014)

Antes de implementar cualquier modelo es necesario organizar la información: antes, durante y después del análisis. La creación o selección de un modelo depende de los datos, por lo que datos desorganizados podrían ser motivo de la selección de un modelo erróneo. Los datos deben ser revisados constantemente para identificar cualquier anomalía durante el proceso y corregirlas a tiempo. Evaluar los resultados obtenidos y asegurarse de que las conclusiones posean cierta concordancia con el problema planteado (Snee, 2015).

Dentro de la minería de datos se realizan tareas descriptivas y predictivas para identificar patrones de relación de datos y para estimar valores futuros (García, 2013). La minería de datos se ha definido como "la extracción no trivial de información implícita, previamente desconocida y potencialmente útil de los datos" (Snehal, Manoj, & Shweta, 2014). Emplea varias técnicas informáticas tales como técnicas supervisadas o no supervisadas del algoritmo de aprendizaje, para buscar automáticamente datos grandes y derivar los patrones que se pueden utilizar para generar predicciones o tareas descriptivas como la agrupación, la asociación, etc. Las aplicaciones de minería de datos pueden utilizar una variedad de parámetros para examinar los datos, pueden incluir varios patrones de asociación donde un evento está conectado a otro evento, como comprar un cepillo de dientes y un hilo dental, secuencia o análisis de trayectoria, es decir, patrones donde un evento conduce a otro evento, como la llegada de sesiones festivas y la compra de paños. (Snehal, Manoj, & Shweta, 2014).

Para extraer y determinar patrones de relación entre los datos se hace uso de una serie de métodos inteligentes a través del uso de algoritmos como: regresión lineal, regresión logística, de asociación, lógica difusa, árboles de decisión y redes neuronales, entre otros. Esto permite obtener conocimiento histórico y predictivo que apoyan en la toma de decisiones (Gorbea, 2013) .

2.6 Objetivo general:

Identificar los algoritmos relevantes que determinen las relaciones entre las variables de contaminación atmosférica.

2.7 Objetivos específicos:

1. 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.
2. Elaborar un marco teórico sobre las técnicas de minería de datos aplicables al problema.
3. Identificar los algoritmos relevantes que determinen las relaciones entre las variables de contaminación atmosférica.
4. Analizar los resultados obtenidos de la aplicación de los algoritmos seleccionados.

2.8 Metodología:

Se estudiará los fundamentos teóricos del tema apoyado en una búsqueda bibliográfica en diferentes fuentes: eventos relacionados con la temática, revistas, conferencias y repositorios de

universidades. Se seleccionarán las fuentes bibliográficas relevantes y se procederá a crear un documento que refleje el estado del arte y un estudio comparativo sobre algoritmos de minería de datos.

Sera necesario revisar los registros generados por la estación de monitoreo para luego pasar a la experimentación sobre un mismo grupo de datos con distintos algoritmos. Todos los experimentos serán registrados para obtener un informe técnico de los resultados obtenidos.

2.9 Alcances y resultados esperados:

Culminado este proyecto se pretende determinar cuál o cuáles son los algoritmos óptimos para el análisis de contaminantes en el aire, e identificar los patrones de relación de las variables.

Se debe obtener un informe técnico, matrices y documentos de evaluación resultantes de la evaluación de algoritmos.

2.10 Supuestos y riesgos:

Riesgos	Probabilidad	Alternativas de solución
Capacidad de procesamiento baja que distorsionaría los resultados.	Media	Adquirir equipos.
Insuficiente variabilidad de los datos recolectados que respalden los resultados.	Media	Buscar nuevas fuentes de datos.
Tiempo insuficiente para terminar el trabajo.	Bajo	Solicitar una prórroga para la entrega final del proyecto.

UNIVERSIDAD DEL AZUAY

2.11 Esquema tentativo:

Capítulo 1: Introducción.

Capítulo 2: Estado del Arte.

Capítulo 3: Análisis de Algoritmos.

Capítulo 4: Análisis de Resultados.

Conclusiones y trabajos futuros.

Bibliografía.

2.12 Cronograma:

Objetivo Específico	Actividad	Semanas																				
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
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 documentación y conceptos a utilizarse dentro del desarrollo del proyecto.																					
Elaborar un marco teórico sobre las técnicas de minería de datos aplicables al problema.	Revisión de bibliografía sobre la minería de datos y sus técnicas de aplicación.																					
Identificar los algoritmos relevantes que determinen las relaciones entre las variables de contaminación atmosférica.	Experimentación de algoritmos en un grupo de datos																					
Analizar los resultados obtenidos de la aplicación de los algoritmos seleccionados	Aplicación de los algoritmos de minería de datos seleccionados en la data sintetizada de la estación de monitoreo																					

Edición autorizada de 10.000 ejemplares
Del 788.501 al 798.500 Nº 0797223

2.13 Referencias:

Bifet, A. (2013). Data mining in real time. *Informatica37*.

Fernandez-Camacho, R., Brito Cabeza, I., Aroba, J., Gomez-Bravo, F., Rodriguez, S., & de la Rosa, J. (2015). Assessment of ultrafine particles and noise measurements using fuzzy. *Science of the Total Enviroment*.

García, F. (2013). Aplicación de técnicas de minería de datos a datos obtenidos por el centro de Andaluz de medio ambiente. *Universidad de Granada*.

Gorbea, P. (2013). Tendencias transdiciplinarias en los estudios métricos de la información y su relación con la gestión de la información y del conocimento. *Perspectivas em Gestao & Conhecimento*.

Medina, F., & Gómez, C. (2014). Funcionalidades de la minería de datos. *Revista Ingeniería y Región*.

Red_de_Monitoreo_del_Calidad_del_Aire_de_Cuenca, E. (2015). *Informe de la Calidad del Aire de Cuenca*. Cuenca.

Sellers Walden, C. A. (2012). Publlicación de Contaminantes Atmosféricos de la Estación de Monitoreo de la Ciudad de Cuenca, Utlizando Servicios Estándares OGC. *Universidad de Azuay*.

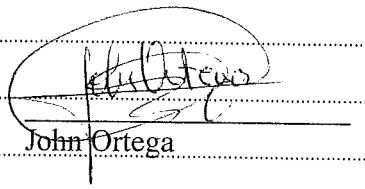
Snee, R. D. (2015). A Practical Approach to Data Mining: I Have All This Data: What Should I Do? *Quality Engineering*.

Snehal, S., Manoj, C., & Shweta, G. (2014). Implementation of WSN Based Air Pollution Monitoring Sysem using Data Mining Technique. *International Journal of Computer Scince and Mobile Computing*.

UNIVERSIDAD DEL AZUAY

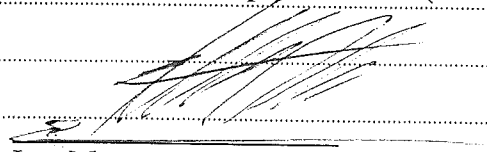
2.14 Anexos: para casos en los que se requiera respaldar el proyecto.

2.15 Firma de responsabilidad (estudiante)



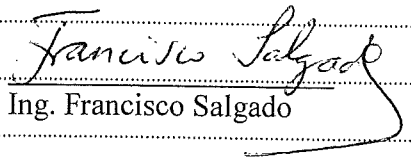
John Ortega

2.16 Firma de responsabilidad (director sugerido)



Ing. Marcos Orellana

2.17 Firma de responsabilidad (Asesor Metodológico)



Ing. Francisco Salgado

2.18 Fecha de entrega: 15 de Mayo de 2017

Edición autorizada de 10.000 ejemplares
Del 788.501 al 798.500 Nº 0797221