TRABAJO DE GRADO
Opción Seminario-Diplomado.

Aplicación de la Ciencia de Datos para analizar la Calidad del Aire en el área metropolitana de Bucaramanga, Colombia

Corporación Universitaria Remington

Facultad de Ingeniería

Ingeniería de Sistemas

Jhon Fredy Ospina Montoya – Kirlian Valencia Ochoa[1]

Tutor temático seminario: Ivonne Castaño Osorio[2]

Tutor metodológico seminario: Jhon Edison Amortegui[3]

Opción de trabajo de grado: Seminario

2024

[1] Estudiantes noveno semestre de Ingeniería de Sistemas Uniremington, sede Pereira, Email: jhon.ospina.4060@miremington.edu.co, kirlian.valencia.7897@miremington.edu.co

[2] Tutor temático Seminario Big Data y ciencia de datos, Email: ivonne.castano@uniremington.edu.co

[3] Tutor metodológico Seminario Big Data y ciencia de datos, Email: jhon.amortegui@uniremington.edu.co

Dedicatoria

A mi familia

Quienes siempre apoyaron mi sueño de ser ingeniero, gracias por siempre apoyarme de manera incondicional y por creer en mi

A mis compañeros

Los cuales estuvieron conmigo en esta travesía llena de retos, mismos que nos ayudaron a crecer intelectualmente y crecer profesionalmente.

Tabla de contenidos

Resumen

En este trabajo se empleará un proceso de ciencia de datos para analizar un dataset con el fin de revisar la información allí inmersa y con ello poder identificar posibles problemas con los índices que se muestran de la medición obtenida gracias a sensores instalados en puntos estratégicos de la ciudad de Bucaramanga para medir la calidad del aire. Este análisis inicia con la recolección y posterior limpieza de los datos, para asegurar una gran calidad y precisión. Teniendo en cuenta el manejo de valores faltantes se aplicarán visualizaciones de datos para identificar posibles patrones o tendencias en la contaminación del aire. Se finalizará con la interpretación de los resultados obtenidos para poder proporcionar estrategias de mitigación para que sean implementadas por las autoridades locales, basados en los índices permitidos de contaminación. Los datos utilizados son específicos sobre la contaminación del aire en el área metropolitana de Bucaramanga (AMB), aquí se indica que en el año 2018 se estableció una medida ambiental por parte de la administración municipal, la cual consistió en implementar sensores especializados en 5 sectores estratégicos de la ciudad, estos sensores a su vez fueron instalados en 5 estaciones (EST. LAGOS, EST. LA CIUDADELA, EST. SANTA CRUZ, EST. SAN FRANCISCO y EST. LAGOS DEL CACIQUE) en cada una de ellas se pudo medir la presencia en el aire de partículas contaminantes que son perjudiciales para la salud respiratoria de la población Material particulado de 10 Microgramos y 2.5 Microgramos (PM10 y PM2.5).

Palabras clave: Ciencia de datos, Metodología, Big Data, Calidad de Aire, Salud Pública.

Pregunta orientadora de la búsqueda

Por lo planteado en la resolución 2254 de 2017 y teniendo en cuenta que el área metropolitana de Bucaramanga implementó un sistema para medir la calidad del aire como está publicado en su página web:

Este sistema de vigilancia beneficiará a la comunidad al poder contar con información continua y en tiempo real, cuyos resultados permitirán evaluar la incidencia de estos contaminantes en la salud de la población expuesta. (Área Metropolitana de Bucaramanga, 2021)

Se hizo necesario una forma de analizar los datos obtenidos, teniendo en cuenta que será una gran cantidad de datos se hace necesario analizarlos mediante Big Data como lo plantea López y Zarza (2017):

> Big data se refiere normalmente a la aplicación de un enfoque científico práctico a la resolución de problemas de datos en los cuales se necesita atacar eficientemente! a uno o más de los tres atributos principales: volumen o cantidad de datos. (p. 56)

El análisis es importante para tener en cuenta los resultados obtenidos por las estaciones, pues analizando los datos individuales es posible llegar a ciertas conclusiones o posibles patrones que se presenten en ciertos momentos, de ser así estos datos son útiles para poder plantear posibles oportunidades de mejora, en este caso se eligió una metodología descriptiva para este fin.

Teniendo en cuenta la metodología descriptiva y aplicando las fases de CRISP-DM tal como lo muestra Nores (2004):

> Existen varias metodologías para orientar el proceso de minería de datos; ellas pretenden facilitar la realización de nuevos proyectos con características similares, optimizar la

> planificación y dirección de éstos, reducir su complejidad y permitir hacer un mejor seguimiento a ellos. (p. 18)

Basado en los resultados obtenidos de micropartículas PM2,5 (2.5 microgramos) y PM10 (10 microgramos) se tratará de dar respuesta a la siguiente pregunta problema ¿Haciendo uso del análisis de la ciencia de datos es posible determinar si existe relación entre la humedad y la cantidad de micropartículas PM2,5 y PM10 en el aire del área metropolitana de Bucaramanga y cómo afecta la calidad del aire de la ciudad?

Metodología de búsqueda de la información

Se utilizó la metodología CRISP-DM de IBM debido a que permite contextualizar muy bien la problemática que se va a analizar, así mismo permite seguir 10 pasos para entender y aclarar cada aspecto del tema elegido:

Paso 1. Comprensión del negocio:

En este paso se trata de comprender la problemática demostrada inicialmente en la pregunta orientadora y con base en los datos existentes poder determinar cuáles datos son útiles para responder a la problemática planteada. Teniendo en cuenta la información existente se identificará qué factores influyen negativamente en la calidad del aire en la ciudad de Bucaramanga.

Dando respuesta a la Resolución 2254 (2017) emitida por el ministerio de Medio Ambiente, donde según Ministerio de medio ambiente y desarrollo sostenible es necesario:

> Establecer la norma de calidad del aire o nivel de inmisión y adoptar disposiciones para la gestión del recurso aire en el territorio nacional, con el objeto de garantizar un ambiente sano y minimizar riesgos sobre la salud humana que pueda ser causado por la exposición a contaminantes en la atmósfera. (p. 1)

Se implementó un sistema de vigilancia de calidad del aire (SVCA) tipo 3, el cual consta de 5 estaciones en puntos estratégicos de la ciudad, cada estación tiene la capacidad de medir micropartículas suspendidas en el aire como lo son PM2,5 (2.5 microgramos) y PM10 (10 microgramos), asimismo humedad y radiación solar, entre otras variables climáticas.

La función de este SVCA sería mantener en tiempo real información específica de la calidad del aire en el área metropolitana de Bucaramanga y de esa manera permitir evaluar los contaminantes presentes para tomar medidas que vayan en pro de la salud pública.

Paso 2. Enfoque analítico: una vez se establece una problemática se define el enfoque analítico sobre el cual se va a trabajar, se debe basar en lo que es necesario abordar. Se escogió el enfoque descriptivo ya que con este tipo de enfoque se podrá revisar el tipo de relaciones que hay entre las mediciones de las diferentes estaciones. Según como lo indica Maldonado, (2022):

> Un correcto análisis descriptivo de los datos permite obtener insights de manera sencilla, lo cual facilita a las juntas directivas, y altas gerencias de las empresas, la toma de decisiones. Adicional- mente, nos sirve para detectar errores en los datos. Sin embargo, su simpleza no permite identificar relaciones más complejas entre variables, por lo tanto, su aporte es limitado y está muy centrado en la capacidad del analista para tomar decisiones en base a los resultados del análisis. (p. 108)

Así de una forma acertada, se analizará como se ve afectado el PM2,5 (2.5 microgramos) y PM10 (10 microgramos) dependiendo de otro tipo de factores.

Paso 3. Requisitos de datos: se utilizará una base de datos brindada por la página principal de la Alcaldía de Bucaramanga que arroja los resultados de un estudio que inició a principios del 2018 hasta mediados del 2020 donde varias estaciones de medición arrojaban estos resultados tales como: CIUDADELA, EST. SANTA CRUZ, EST. SAN FRANCISCO y EST. LAGOS DEL CACIQUE. Dichas estaciones nos mostraban las partículas en el aire y así se analizará cómo esto puede afectar la calidad del aire.

Paso 4. Recopilación de datos: se realiza la recopilación de datos inicial, la cual es analizada para identificar si existen vacíos en la información o hay información sobrante no necesaria. El ingeniero de datos puede utilizar métodos estadísticos para evaluar, sustituir o completar de la forma más eficiente la información que tenga este tipo de falencias, para que al momento de que el experto haga la revisión tener la mayor certeza posible sobre los datos que trabajará, estos se

catalogan como datos estáticos, por estar alojados en la página web, según como lo define Roma (2019): Los datos estáticos se encuentran almacenados de forma perdurable, es decir, a largo plazo. Por lo tanto, la captura de estos datos es similar al proceso de captura que se ha venido haciendo en los procesos de minería. (p. 128)

Se analizó la base de datos extraída de la página web y se pudo identificar que habían muchos datos en 0, datos que no servirán debido a que en los histogramas no generarían nada y sería imposible analizar dichos gráficos por esto se optó por eliminar dichos datos y así tener mayor agilidad en dicho análisis.

Paso 5. Comprensión de datos: en esta fase se debe tener claro que los datos recolectados si apuntan a resolver el tema general, se debe ubicar variables y revisar la relación entre ellas para posteriormente ser usadas en el modelo, se pueden utilizar herramientas estadísticas e histogramas para evaluar la calidad de los datos como lo presenta López (2009):

> DQGuard, de Knowledge Integrity Incorporated, captura las reglas de calidad de los datos y reglas de negocios en un lenguaje formal. Presenta un marco para definir las reglas de calidad de datos, manejando estas como contenido y genera aplicaciones para medir y reportar los resultados de probar estas reglas. (p. 32)

En caso de que se determine que hay poca calidad en los datos se debe retornar a la fase anterior.

Paso 6. Preparación de datos: este es un paso que toma mucho tiempo sabiendo que es fundamental se deben transformar los datos de la manera que sea más fácil trabajarlos, deben tener un formato generalizado que ayude a eliminar duplicidades y datos faltantes. Se transformaron los datos de tal forma que todas las columnas estuvieran en formato decimal y así poder tener mayor

certeza en la calidad de los datos también se encontró varias similitudes en las columnas existentes por esto se optó por tomar los datos de las columnas que tuvieran similitudes.

Paso 7. Modelado: partiendo de la elección del enfoque a utilizar que en este caso será el enfoque descriptivo se debe buscar que la implementación de modelos mediante comparaciones que hará el científico de datos apoyado en diferentes algoritmos, con datos ya conocidos, se ajusta para garantizar que las variables que se usen sean las adecuadas para tratar el tema, como lo indica Minguillón (2017): el objetivo es asegurar que los modelos construidos a partir de los datos disponibles funcionan correctamente para nuevos datos que haya que pro- cesar (clasificar, agrupar, etc.) (p. 70)

Paso 8. Evaluación: la evaluación y el modelado se hacen casi que a la par, nos permite calificar la calidad del modelo y si nos funciona para resolver la problemática del tema inicial haciendo que el modelo desarrollado sea preciso y confiable, así es precisado por Minguillón (2017):

> En fases anteriores nos hemos preocupado de asegurar la fiabilidad y plausibilidad del modelo; en cambio, en esta fase nos centraremos en evaluar el grado de acercamiento a los objetivos de negocio y en la búsqueda, si las hay, de razones de negocio por las cuales el modelo es ineficiente. (p. 32).

Paso 9. Implementación: se debe realizar una retroalimentación total del modelo de datos y se pone a prueba involucrando a todas las partes interesadas, dependiendo del propósito es posible probar en grupos pequeños, si se trata de escenarios empresariales.

Paso 10. Retroalimentación: posterior a la implementación, se tendrá que tener muy en cuenta los comentarios de los implicados para pulir el modelo y verificar su impacto, se deben

hacer mejoras y establecer un punto en el cual el experto confía en la efectividad y se emplea en el caso real.

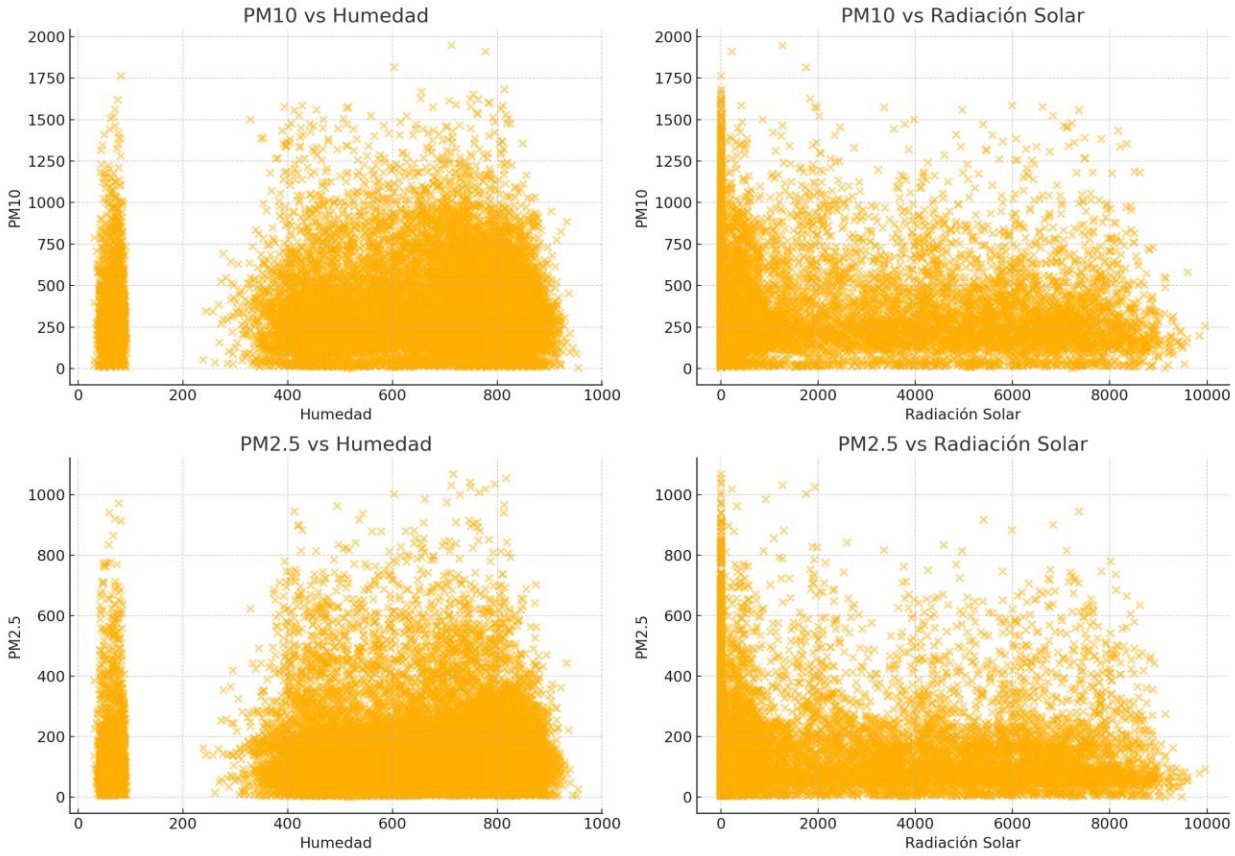
Sustentación teórica de la pregunta

Para tratar de hallar la respuesta a la pregunta orientadora una búsqueda de fuentes como la OMS donde se establece los valores máximos permitidos en cuanto a la presencia de PM10 y PM2.5, así mismo se consultó en la página gubernamental del área metropolitana de Bucaramanga donde se indica cual es el objetivo y en que consiste la implementación del sistema para medir la calidad del aire. Teniendo esto se hizo uso de la herramienta Excel con su función análisis de datos el cual realiza un proceso con los datos seleccionados, como muestra Alcalde (2015): un proceso es un conjunto de actividades mutuamen- te relacionadas que al interactuar, transforman elemen- tos de entrada y los convierten en resultados. (p. 55)

El cual permitió graficar los resultados y las figuras insertadas a continuación:

Análisis de correlación:

La correlación muestra los coeficientes entre las diferentes variables, en el caso de este documento Partículas PM10, Partículas PM2.5, Humedad y Radiación Solar, todas medidas por una misma estación. Un coeficiente de correlación cercano a 1 ó -1 indica una relación lineal, mientras que un valor cercano a 0 indica poca relación o que definitivamente no tiene relación.



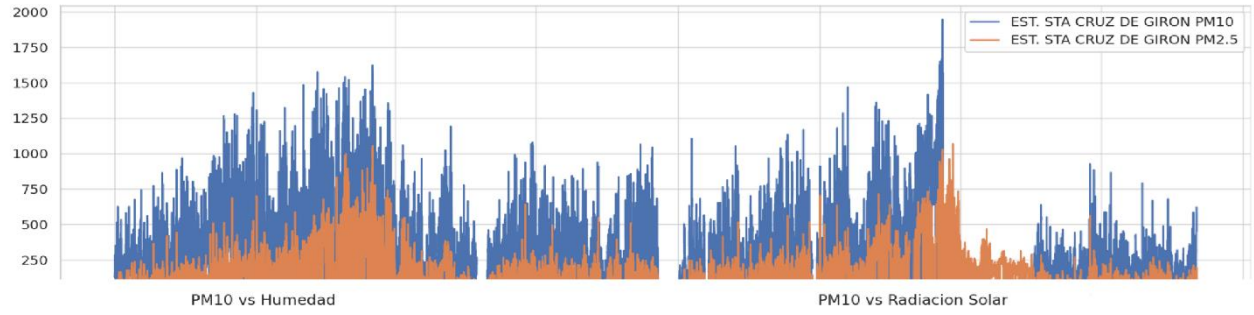
Gráfica 1. Correlación Estación Santa Cruz de Girón.

Utilizando los gráficos de dispersión se analizó la correlación de todas las variables presentes en la estación Santa Cruz de Girón. En la parte superior izquierda se muestra las variables PM10 vs Humedad, en la parte superior derecha se muestra PM10 vs Radiación solar.

En la parte inferior se muestran las variables PM2.5 vs Humedad, en la parte superior derecha se muestra PM2.5 vs Radiación solar. De esta manera se realiza la respectiva comparativa para hacer más fácil su análisis (Ver gráfico 1)

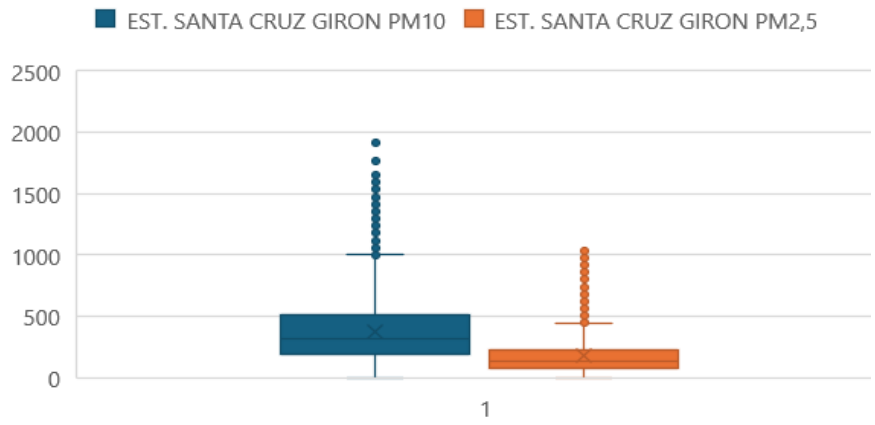
Tabla 1. Análisis de Correlación Estación Santa Cruz de Girón, se realiza la comparativa de las variables PM10, PM2.5, Humedad y Radiación solar.

	EST. STA CRUZ DE GIRON PM10	EST. STA CRUZ DE GIRON PM2.5	Humedad	Radiacion Solar
EST. STA CRUZ DE GIRON PM10	1	0.707340162	0.025027468	-0.106514753
EST. STA CRUZ DE GIRON PM2.5	0.707340162	1	0.003888261	-0.082477144
Humedad	0.025027468	0.003888261	1	-0.457558339
Radiacion Solar	-0.106514753	-0.082477144	-0.457558339	1



Grafica 2. Gráficos de barras Estación Santa Cruz de Girón

Se realiza gráfico de barras para graficar más visualmente la presencia de PM10 y PM2.5 frente a las variables de humedad y radiación solar. Al revisar los gráficos y los coeficientes de correlación, se puede determinar que si existe una relación significativa entre PM10 y PM2.5, por lo tanto, cuando existe la presencia de una también está la otra en una medida similar, por otro lado la radiación solar y la humedad con las concentraciones de partículas PM10 y PM2.5, no presentan una relación, dado que el promedio es muy bajo, cercano a 0 (cero).



Gráfica 3. Gráfico de cajas y bigotes Estación Santa Cruz de Girón

Haciendo uso del gráfico de cajas y bigotes se puede establecer el rango medio de ambas micropartículas (PM10 y PM2.5), en el grafico izquierdo los valores que son superiores a 1000 son muy pocos, podrían ser un error en la lectura de los sensores, en el caso del gráfico de la derecha los valores por encima de 500 son muy pocos y podrían deberse a errores de lectura.

En la parte izquierda de color azul, se presenta el gráfico correspondiente a los datos obtenidos de la micropartícula PM10, se puede evidenciar que el rango mayormente esta entre 200 y 500 aproximadamente, quiere decir que entre estos 2 valores se concentra la mayor parte de la presencia de esta partícula en el aire.

En la parte derecha, de color naranja se presenta el gráfico de la micropartícula PM2.5, se puede evidenciar que el rango donde se concentra mayormente, según las lecturas es entre 100 y 230 aproximadamente, teniendo cierta consistencia y menor rango en su medición, entre menor y mayor.

Conclusiones.

Según lo analizado en los gráficos obtenidos se puede determinar que no hay una relación directa entre la presencia de micropartículas PM10 y PM2.5 con respecto a la humedad o radiación solar, por lo tanto no es posible demostrar que la contaminación presente por parte de estas partículas esté relacionada con las condiciones climáticas tenidas en cuenta en este análisis que se realiza basado en los datos obtenidos por parte de SVCA instalado en el Área Metropolitana de Bucaramanga.

La ciencia de datos demuestra la gran importancia que tiene para la ingeniería de sistemas, al combinar herramientas y métodos, necesarios para llevar a cabo el análisis de grandes cantidades de datos recolectados en este caso por los sensores instalados en el área metropolitana de Bucaramanga, con la utilización de esta se permite tener una idea general de lo que ocurre en cierto periodo.

Lista de referencias

Alcalde, I. (2015). Visualización de la información: de los datos al conocimiento: (ed.). Barcelona, Spain: Editorial UOC. Recuperado de https://elibro.net/es/ereader/remington/57832?page=55.

(area metropolitana de bucaramanga, 2018)

Casas Roma, J. Nin Guerrero, J. y Julbe López, F. (2019). Big data: análisis de datos en entornos masivos: (ed.). Barcelona, Editorial UOC. Recuperado de https://elibro.net/es/ereader/remington/117744?page=128.

Gondar Nores, J.-E., Metodologías para la Realización de Proyectos de Data Mining [Electronic Version]., 2004. from http://www.estadistico.com/arts.html?20040426

López Murphy, J. J. y Zarza, G. (2017). La ingeniería del big data: cómo trabajar con datos: (ed.). Barcelona, Spain: Editorial UOC. Recuperado de https://elibro.net/es/ereader/remington/59093?page=56.

López Porrero, B. E. (2009). Limpieza de datos: (ed.). Santa Clara, Cuba: Editorial Feijóo. Recuperado de https://elibro.net/es/ereader/remington/71744?page=32.

Maldonado, S. (2022). Analytics y Big Data: ciencia de los Datos aplicada al mundo de los negocios: (1 ed.). Santiago de Chile, RIL editores. Recuperado de https://elibro.net/es/ereader/remington/225562?page=108.

Medio Ambiente, donde según Ministerio de medio ambiente y desarrollo sostenible (Noviembre 1 de 2017) Resolución 2254 de 2017

https://www.minambiente.gov.co/documento-entidad/resolucion-2254-de-2017/

Minguillón, J. Casas, J. y Minguillón, J. (2017). Minería de datos: modelos y algoritmos: (ed.). Barcelona, Spain: Editorial UOC. Recuperado de https://elibro.net/es/ereader/remington/58656?page=32.

Minguillón, J. Casas, J. y Minguillón, J. (2017). Minería de datos: modelos y algoritmos: (ed.). Barcelona, Spain: Editorial UOC. Recuperado de https://elibro.net/es/ereader/remington/58656?page=70..