Artículo / Investigación
Article / Research

Analítica de datos: incidencia de la contaminación ambiental en la salud pública en Medellín (Colombia)

Data Analytics: incidence of air pollution on public health in Medellin, Colombia

Juan S. Parra-Sánchez, Ana I. Oviedo-Carrascal y Ferney O. Amaya-Fernández

Recibido 8 abril 2019 / Enviado para modificación 14 octubre 2020 / Aceptado 25 octubre 2020

RESUMEN

Objetivo Analizar el impacto de la contaminación del aire por material particulado $PM_{2,5}$ y su relación con el número de asistencias a entidades de salud por enfermedades respiratorias por medio de analítica de datos.
Métodos Se analizaron datos del Área Metropolitana de Medellín, Colombia, ciudad ubicada en un valle estrecho densamente poblado e industrializado y que ha presentado episodios críticos de contaminación en los últimos años. Se analizaron tres fuentes de datos: datos meteorológicos aportados por el SIATA (Sistema de Alerta Temprana de Medellín y el Valle de Aburrá); datos de contaminación por material particulado $PM_{2,5}$ aportados por SIATA; y reportes de los RIPS (Registros Individuales de Prestación de Servicios de Salud) aportados por la Secretaría de Salud.
Resultados Se evidenció la relación entre la concentración de $PM_{2,5}$ con las asistencias médicas por los diagnósticos de IRA, EPOC y asma. En un episodio crítico de contaminación por $PM_{2,5}$, se encontraron los siguientes retardos en la atención médica: entre 0 y 2 días para el IRA, 0 y 7 días para el EPOC y 0 y 5 días para el asma.
Discusión Se encontraron coeficientes de correlación que evidencian la asociación de la concentración de $PM_{2,5}$ con las asistencias por los diagnósticos de IRA, EPOC y asma. La mayor correlación entre las tres morbilidades se presentó para el asma. La variable meteorológica de mayor correlación con la variable objetivo es la temperatura del aire para el caso de EPOC y asma. En el caso de IRA, la variable con mayor correlación es la velocidad del viento. Por otro lado, el día de la semana es una variable de gran importancia a la hora de realizar un estudio de atenciones por enfermedades.

Palabras Clave: Análisis de datos; ciencia de los datos; enfermedades respiratorias; contaminación del aire (*fuente: DeCS, BIREME*).

ABSTRACT

Objective To analyze the impact of air pollution by $PM_{2,5}$ particulate matter and its relationship with the number of attendances to health entities for respiratory diseases through data analytics.
Methods Data from the Metropolitan Area of Medellín, Colombia, a city located in a densely populated and industrialized narrow valley and that has presented critical episodes of contamination in recent years, were analyzed. Three data sources were analyzed: meteorological data provided by SIATA (Early Warning System of Medellín and the Aburra Valley), $PM_{2,5}$ particulate matter contamination data provided by SIATA, and RIPS reports (Individual Registers for the Provision of Health Services) provided by the health department.
Results The relationship between the concentration of $PM_{2,5}$ and medical care for the diagnoses of ARI, COPD and asthma was evidenced. In a critical episode of $PM_{2,5}$ contamination, the following delays in medical care were found: between 0-2 days for IRA, 0-7 days for COPD, and 0-5 days for asthma.

JP: Ing. Electricista. Esp. Inteligencia de Negocios. M. Sc. Ingeniería. Universidad Pontificia Bolivariana. Medellín, Colombia.
juans.parra@upb.edu.co
AO: Ing. Sistemas. Ph. D. Ingeniería Electrónica. Universidad Pontificia Bolivariana. Medellín, Colombia.
ana.oviedo@upb.edu.co
FA: Ing. Electrónico. Ph. D. Ingeniería. Universidad Pontificia Bolivariana. Medellín, Colombia.
ferney.amaya@upb.edu.co

DOI: https://doi.org/10.15446/rsap.V22n6.78985

Discussion Correlation coefficients were found that show the association of the concentration of $PM_{2.5}$ with the attendances for the diagnoses of ARI, COPD, and asthma. The highest correlation between the three morbidities was found for asthma. The meteorological variable with the highest correlation with the objective variable is air temperature in the case of COPD and asthma. In the case of IRA, the variable with the highest correlation is wind speed. On the other hand, the day of the week is a variable of great importance when carrying out a study of care for diseases.

Key Words: Analysis of data; data science; respiratory diseases; air pollution (*source: MeSH, NLM*).

Un estudio de la Organización de las Naciones Unidas (ONU) en 2014 señala que, por primera vez en la historia, más de la mitad del planeta vive en ciudades (1). Esto implica diferentes problemáticas en materia de desigualdad social, movilidad, inseguridad y contaminación del aire. Esta última problemática es una de las más críticas teniendo en cuenta su alto impacto en la salud pública (2).

La contaminación del aire incluye la presencia de partículas de NO_2, oxidantes fotoquímicos, dióxido de azufre, monóxido de carbono y material particulado fino, las cuales afectan directamente la salud ya que aumentan el riesgo de padecer diferentes enfermedades, principalmente las cardiovasculares y las respiratorias, que causan muertes prematuras (2).

Se estima que, debido a la contaminación del aire, alrededor de tres millones de muertes prematuras se producen al año en el mundo (3). Esto es aproximadamente el doble del número de muertes producidas por accidentes de tránsito, según un estudio de la Organización Mundial de la Salud (4). Los pronósticos no son alentadores. De acuerdo con E. Van der Wall, el número de muertes prematuras en el mundo debidas a la contaminación del aire en 2050 será de 6,6 millones anuales (5).

En el caso particular de Colombia, según el Departamento Nacional de Planeación (DNP), 8241 personas murieron por causas de contaminación del aire entre 2015 y 2016 (6). Pero no solo el costo se paga en pérdidas humanas, también se producen sobrecostos en el sector salud, que atiende estas problemáticas. Los costos ascienden a 1,6 billones de pesos al año, según el DNP (7).

La revisión del estado de la cuestión revela que la contaminación del aire es un problema que se debe abordar en el contexto de las ciudades inteligentes del futuro. Esto implica diferentes desafíos, entre ellos, la capacidad de adquirir información oportuna sobre los eventos de la ciudad mediante un uso cada vez más amplio de las Tecnologías de la Información y la Comunicación (8).

Uno de los puntos de partida para solucionar problemáticas como la contaminación ambiental es convertir los datos en información fiable y usar esta última de forma eficiente para tomar decisiones de gestión y mejorar así la calidad de vida de los ciudadanos. Por esta razón, desde la perspectiva de la ciencia de datos, aquellos datos que surgen de las ciudades inteligentes dan lugar a muchos desafíos que constituyen un nuevo campo interdisciplinario de investigación (9).

Algunos de esos retos son: la complejidad y cantidad de los datos, los cuales usualmente presentan relaciones altamente no-lineales que requieren técnicas de procesamiento robustas; la preparación y limpieza de los datos, los cuales vienen de diferentes fuentes con diferentes formatos; y el procesamiento y selección de técnicas que se usarán, que dependen de la naturaleza de los datos, de los recursos computacionales disponibles y del objetivo del análisis (10).

En este trabajo de investigación se aplicó la analítica de datos para el estudio de la contaminación ambiental y su impacto en la salud de la población, específicamente con el fin de producir conocimiento que permita determinar la asociación entre la contaminación atmosférica y las asistencias hospitalarias por enfermedades respiratorias en el marco de ciudades inteligentes. Como caso de estudio se usaron los datos del Área Metropolitana de Medellín (Valle de Aburrá), la cual ha presentado episodios críticos de contaminación en los últimos años y se encuentra entre las zonas más contaminadas de Colombia, junto con Bogotá, Ráquira y Yumbo (7).

MATERIALES Y MÉTODOS

Para la experimentación mediante analítica de datos, la metodología empleada fue CRISP-DM *Cross-Industry Standard Process for Data Mining* (11), mediante la aplicación de cinco fases: comprensión de la problemática, comprensión de los datos, preparación de los datos, modelamiento, evaluación y despliegue. A continuación, se describen los materiales, métodos y técnicas usadas en las diferentes fases de la metodología.

Fase 1. Comprensión de la problemática

En este estudio fue importante precisar el concepto de área fuente. El Acuerdo Metropolitano N.º 8 del 25 de marzo de 2011 establece como área fuente "una determinada zona o región, urbana, suburbana o rural, que, por albergar múltiples fuentes de emisión, es considerada como un área especialmente generadora de sustancias contaminantes del aire" (12, p21).

A la fecha del estudio se encontraba en vigencia el Acuerdo Metropolitano 15 de 2016, en el cual se resalta que la concentración de contaminantes en la base del Valle de Aburrá está directamente influenciada por las condiciones meteorológicas y climáticas de la zona.

En el acuerdo metropolitano en mención se clasifica la cuenca del Valle de Aburrá como área fuente de contaminación por material particulado de PM_{10}. No obstante, los esfuerzos de control están destinados a prevenir los efectos, debido a la exposición al material particulado de $PM_{2.5}$, teniendo en cuenta su mayor impacto en la salud de la población (13-15).

Adicionalmente, las condiciones topográficas particulares del territorio también influyen en la dispersión y transporte de contaminantes. En este sentido, desde el acuerdo del POECA se establece que "el Valle de Aburrá cumple con las condiciones de Cuenca Atmosférica por lo que se entiende que comparte el suministro de aire en toda su extensión y prueba de ello son los valores relativamente homogéneos de las concentraciones de partículas finas $PM_{2.5}$ que se registran en las diferentes estaciones de monitoreo que indican la exposición de la población" (16, p3).

Con el objetivo de analizar la incidencia de la contaminación ambiental en la salud pública en Medellín, específicamente en las atenciones hospitalarias, se define, de acuerdo con una revisión del estado de la cuestión (17-23), las siguientes consideraciones:

- El contaminante identificado para estudiar la relación entre la contaminación ambiental y la salud es el $PM_{2.5}$
- Las enfermedades respiratorias que más relación tienen con la contaminación ambiental son IRA, EPOC y Asma.
- Se identificó que el tiempo que transcurre entre el evento de contaminación y la asistencia a la entidad de salud es de 0 a 7 días.
- El estudio de cada una de las morbilidades se realiza en forma separada, creando un modelo analítico para cada enfermedad.

Fase 2. Comprensión de los datos

Para este estudio se consideraron principalmente tres grandes conjuntos de datos, agrupados así: datos de contaminación por material particulado, datos meteorológicos y datos relacionados con las asistencias por enfermedades respiratorias.

Datos de contaminación por material particulado

Es importante precisar que las estaciones que hacen parte de la red de monitoreo de calidad del aire están ubicadas en zonas urbanizadas. Estas estaciones representan la calidad del aire en un área de influencia y no son reflejo de situaciones aisladas. Se dividen en dos tipos: estaciones de representatividad poblacional y estaciones de tráfico.

Las estaciones de representatividad poblacional son, específicamente, las que se toman de referencia para el análisis de las alertas para enfrentar episodios críticos de contaminación atmosférica a nivel metropolitano y, por ende, son las que permiten aplicar las medidas expuestas en los diferentes acuerdos de orden metropolitano. Teniendo en cuenta que los datos disponibles para estaciones de representatividad poblacional en los ámbitos de la salud son del municipio de Medellín, se tienen dos posibles estaciones, llamadas Universidad Nacional de Colombia y Tanques la Ye EPM.

Analizando la completitud de los datos meteorológicos, la estación Tanques la Ye EPM tiene disponibles datos desde enero del 2015, mientras que la estación Universidad Nacional de Colombia cuenta con datos desde el año 2011. Los datos de salud corresponden al rango de años desde el 2014 hasta el 2016. Por este motivo, para este estudio se seleccionó la estación Universidad Nacional de Colombia, núcleo El Volador. Por lo anterior, y teniendo en cuenta el concepto de área fuente, es pertinente para este estudio, realizar la elección de una de las estaciones para extraer los datos.

Datos meteorológicos

Las variables de estudio son el resultado de un procesamiento realizado a la base de datos de SIATA para la estación seleccionada. Se encuentran reportadas la velocidad del viento, temperatura del aire, presión, precipitación, humedad y radiación. Para cada una de estas variables se halló el valor máximo, mínimo y el promedio.

Datos relacionados con las asistencias por enfermedades respiratorias

Para este estudio se utilizaron los RIPS (Registros Individuales de Prestación de Servicios de Salud), proporcionados por la Secretaría de Salud del municipio de Medellín para los años 2014 a 2016. De acuerdo con el Ministerio de Salud y Protección Social, se definen los RIPS como "el conjunto de datos mínimos y básicos que el Sistema General de Seguridad Social en salud requiere para los procesos de dirección, regulación y control y como soporte de la venta de servicios" (24). Sin embargo, y pese a que los distintos generadores y usuarios de los datos y la información reconocen su valor e importancia, en la actualidad se han identificado grandes falencias tanto en la calidad del registro primario como en la utilización para la gestión de IPS (instituciones prestadoras de servicios) y EAPB (Entidades Administradoras de Planes de Beneficios), así como para la formulación de la política sobre la salud pública y el aseguramiento. Esta situación muestra la necesidad de mejorar la calidad, oportunidad y confiabilidad de los datos que se reportan

para que contribuyan a la toma de decisiones en condiciones de mayor certidumbre (25).

Los RIPS están conformados por cuatro clases de datos, que se aplican dependiendo del servicio de salud registrado: datos de identificación, datos del servicio, datos del motivo de la atención y datos del valor. Para este estudio se tomaron dos archivos: el de usuarios y el de consulta. Fue necesario realizar una fusión de los datos del archivo de usuarios con el de consulta, teniendo en cuenta que la Secretaría de Salud de Medellín protege los derechos de los pacientes y entrega los datos de forma anónima, a fin de respetar la privacidad de los usuarios.

Fase 3. Preparación de datos

El gran volumen de datos comprende alrededor de 80 158 registros de la estación seleccionada y alrededor de 1.5 millones de los RIPS. Este volumen se debe a la integridad de los datos, la cantidad de datos nulos, la calidad encontrada en los datos y su completitud.

En la base de datos entregada por SIATA, se encontraron algunos problemas de calidad de datos. Por ejemplo, de los 80 158 registros que se tienen en la estación, se encontraron problemas con respecto a la granularidad de los registros, completitud y consistencia. Los datos fueron sometidos a una depuración que consistió en eliminar los registros que no tengan datos meteorológicos asociados a una concentración o viceversa. También los datos horarios fueron convertidos a datos diarios.

Con los registros de meteorología y concentraciones de $PM_{2.5}$, se descartaron aquellos días en los que, para el cálculo del promedio, no contaban con un 75% de los datos de día calculado.

De los días comprendidos desde 2014 a 2016, que en total suman 1096, se tomaron 942, que corresponden a aproximadamente el 86 % de los registros.

Es importante considerar que, teniendo en cuenta que no se encuentran disponibles las horas de atención sino las fechas de atención de los usuarios, la granularidad para este estudio fue de día. Por ello, fue necesario conservar esta misma granularidad para los datos medioambientales y meteorológicos.

Se tomaron las variables meteorológicas con el fin de revisar la correlación entre ellas y eliminar las que presentan una correlación mayor al 0,7. Sin embargo, llama la atención la alta correlación encontrada entre la temperatura y la humedad, en la cual se tiene que entre la máxima temperatura del aire y la mínima humedad la correlación es del -0,85; entre el promedio de la temperatura del aire y la mínima humedad la correlación es del -0,74. Debido a la alta correlación existente entre la temperatura y la humedad, se trabajó con el dato de la temperatura y se eliminó la humedad, teniendo en cuenta la facilidad en la medición para futuros escenarios.

Con respecto a la elección del valor máximo, mínimo o del promedio entre las variables meteorológicas se realizó un procedimiento en el que se utilizó un árbol de decisión para elegir las variables de mayor correlación con la variable objetivo, considerando que para la misma magnitud se tiene el valor mínimo, máximo y el promedio. En cuanto a la precipitación y la radiación, hay que tener en cuenta que el valor mínimo es cero, por lo que analíticamente es más adecuado escoger el valor máximo y no el promedio. Para la temperatura y la velocidad del viento se tomó el promedio diario y para la presión, precipitación y radiación el valor máximo.

Cabe resaltar que las atenciones están divididas por los tres grupos de enfermedades mencionados (IRA, EPOC y asma) con el correspondiente retardo entre la medición y la atención en días. Es decir, el registro N.º 1 (del día 1) contendrá las variables descritas y las atenciones por la enfermedad del día posterior y de los siete posteriores.

Fase 4. Modelamiento analítico

La minería de datos permite descubrir estructuras y relaciones interesantes, inesperadas o valiosas en grandes conjuntos de datos. El modelamiento es la capacidad de aplicar una técnica a un conjunto de datos para predecir una variable objetivo o encontrar un patrón desconocido (17). Hay varias tareas en la minería de datos basadas en algunos principios matemáticos: clasificación, regresión y agrupamiento. Todas ellas están basadas en el paradigma del aprendizaje inductivo.

El presente artículo es el resultado de un estudio que, con base en la minería de datos, se propuso realizar dos tareas específicas: la primera, establecer un modelo para predecir un rango de asistencias por enfermedades respiratorias a entidades de salud (y, a la par, analizar las correlaciones de las variables meteorológicas y de contaminación por $PM_{2.5}$); y, segundo, establecer un modelo de *clustering* (o agrupamiento) con el fin de determinar cuántas clases (o rangos) se van a utilizar para la predicción. A continuación, se describen dichos modelos.

Clustering o agrupamiento

Para determinar cuántas clases deberán crearse, se utilizaron clústeres que evidencian los posibles escenarios de días de acuerdo con las variables meteorológicas y medioambientales. Esto se traduce en que el número de grupos (clases) de días deberían ser consistentes con el comportamiento en las atenciones por las enfermedades respiratorias. Se debe tener en cuenta que el hecho de tener muy pocas clases hace que el resultado no sea tan significativo, pues una predicción en un rango muy amplio perderá precisión. Por otro lado, el tener muchas clases hace que las características de cada grupo sean repetitivas y que el aprendizaje del modelo sea más complejo. Para la creación de los clústeres, se utilizó

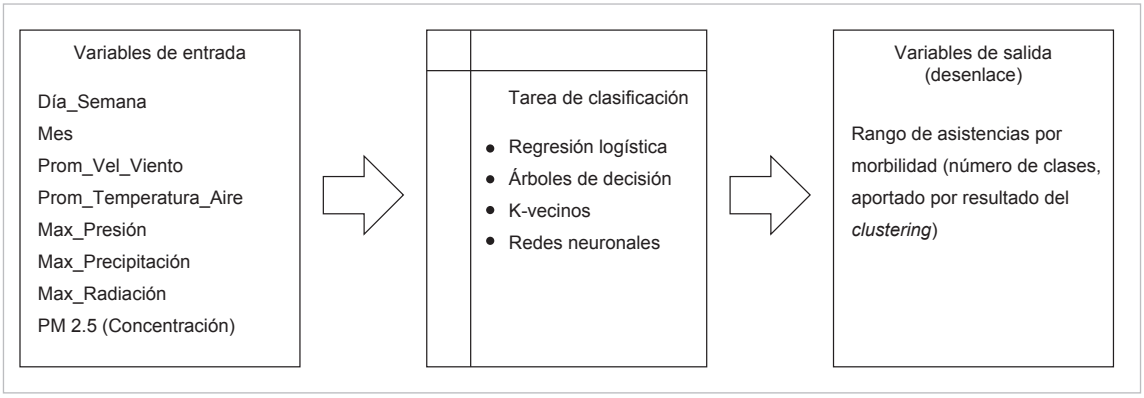
el algoritmo *k-means*. Este algoritmo divide el conjunto de datos en un número predefinido de grupos. *k-means* es el método más comúnmente utilizado para agrupamiento y consiste en definir *k* centroides, uno por clúster. Los datos se asocian al centroide más cercano.

Modelo predictivo para los rangos de asistencias por morbilidad

Con el fin de evidenciar la correlación de las variables predictoras (especialmente el $PM_{2.5}$) con respecto al número de asistencias hospitalarias y a su vez analizar cómo influyen los retardos en las atenciones en el resultado del modelo con una mayor precisión y pertinencia, la clasificación permite minimizar las diferencias entre los registros de una misma clase (varianza intraclase) y maximizar las diferencias entre las clases (varianza interclase). Cuando se trata de predecir una variable, se utilizan métodos supervisados, es decir, métodos donde se conocen las variables de entrada (predictoras) con su respectiva variable de salida (objetivo). En esta etapa en la cual se crean modelos, se articulan las variables de entrada (predictoras) y se evalúa su relación con la variable de salida (objetivo). Por lo tanto, es posible evaluar qué tanto podrían explicar las variables de entrada el comportamiento de una determinada salida.

Las técnicas que serán utilizadas para esta fueron: redes neuronales, regresión logística, árboles de decisión (DT) y k-NN (algoritmo de vecinos más cercanos). De los registros que se tienen para el modelo, se utilizó un 75% para entrenamiento y un 25% para la evaluación. La Figura 1 muestra un esquema simplificado del modelo analítico.

Figura 1. Esquema simplificado del modelo analítico



Fase 5. Evaluación

Evaluación del *clustering* (agrupamiento). En este estudio se utilizó la evaluación de la cohesión del clúster mediante la métrica SSW (*sum of squared within*) y que se calcula de acuerdo con la ecuación:

$$SSW = \sum_{i=1}^{k} \sum_{x \in C_i} dist^2 (m_i, x)$$

Donde *k* corresponde al número de clústeres; x, a un punto del clúster C_i; y m_i, al centroide del clúster C_i.

Evaluación de la predicción

La evaluación de los modelos de clasificación compara los valores predichos con los valores reales de las instancias en el conjunto de datos. El modelo de clasificación tiene diferentes maneras de ser evaluado, una de ellas es el *error absoluto medio* (MAE). Esta es la métrica más simple y directa para evaluar el grado de divergencia entre dos conjuntos de valores, y se define mediante la siguiente ecuación:

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y(i) - x(i)|$$

donde n es el conjunto de instancias, x es la función del modelo e y es la función de destino con las etiquetas correctas de las instancias. En este caso, cada residuo tiene la misma contribución en el MAE final.

Otra forma de calcular el grado de divergencia es el error cuadrático medio (MSE), que es probablemente el más utilizado para evaluar los modelos analíticos. Se calcula utilizando la igualdad:

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y(i) - x(i))^2$$

Esta métrica penaliza los grandes residuos, lo que significa que, si el modelo se aproxima más correctamente a las instancias del conjunto de datos, pero tiene errores significativos en algunos de ellos, la penalización es mayor que la de la métrica MAE. El RMSE se define como la raíz cuadrada del MSE. Comparte características con las nombradas, pero tiene una ventaja adicional; la escala del error coincide con la escala de los datos originales. El RMSE se calcula utilizando la siguiente ecuación:

$$RMSE = \sqrt{MSE}$$

Se espera que, entre mejor sea un modelo, los resultados de MAE y RMSE sean cercanos a 0.

RESULTADOS

Clustering

A continuación, se muestra el resultado para el número de clústeres (categorías) de asistencias por morbilidad (Tabla 1).

Tabla 1. Selección de número de clases

Número de clústeres	SSW
4	566,1
5	447,1
6	423,9
7	411,3
8	399,4
9	496,7

De acuerdo con la Tabla 1, el número de grupos que se pueden realizar con la mejor cohesión es ocho. Los clústeres con sus centroides se muestran en la Tabla 2.

Tabla 2. Clústeres calculados con las variables meteorológicas

Atributo	Todos los datos (914)	Cluster 1 (113)	Cluster 2 (93)	Cluster 3 (146)	Cluster 4 (170)	Cluster 5 (99)	Cluster 6 (95)	Cluster 7 (93)	Cluster 8 (105)
Mes	7	9	4	7	7	9	3	3	10
Dia_semana	Saturday	Sunday	Saturday	Monday	Tuesday	Thursday	Sunday	Thursday	Saturday
Prom_Veloc_Viento	1,63	1,46	1,89	1,63	1,68	1,41	1,86	2,07	1,10
Prom_Temperatura	23,20	23,69	23,83	23,01	23,93	23,10	22,26	23,73	21,66
Max_Presion	641,92	641,83	641,81	641,97	641,82	641,98	642,11	641,73	642,14
Max_Precipitacion	1,92	1,80	0,72	1,89	1,18	2,00	2,82	1,50	3,83
Max_Radiación	814,10	872,50	822,96	815,72	824,64	859,32	712,61	826,68	762,16
$PM_{2,5}$	31,99	26,01	32,84	30,51	31,71	29,31	33,57	41,61	32,76

Modelo predictivo

En la Tabla 3 se muestran los coeficientes de correlación de la concentración del $PM_{2,5}$ modelo con las asistencias por enfermedad pulmonar obstructiva crónica (EPOC), de acuerdo con el retardo (lag) en las atenciones. La concentración del $PM_{2,5}$ se encuentra entre las variables con mayor correlación con un coeficiente del 0,097, aproximadamente.

Tabla 3. Correlación de la concentración de $PM_{2,5}$ con el EPOC

	EPOC	EPOC_L1	EPOC_L2	EPOC_L3	EPOC_L4	EPOC_L5	EPOC_L6	EPOC_L7
$PM_{2,5}$	0,097	0,054	0,059	0,061	0,048	0,073	0,092	0,096

Se presenta con un retardo de cero días (l=0) y con un retardo de 6 y 7 días (l=6,7). La temperatura del aire también tiene una alta correlación con el EPOC, independientemente del retardo. La mayor correlación se presenta con las variables "Día de la semana" y "Mes", lo cual era un resultado esperado, dado que los fines de semana las consultas son mucho menores que en días laborales, lo mismo ocurre con el mes.

Aunque no se tiene un estudio en otra ciudad que compare exactamente estas variables predictoras con el EPOC, en estudios en la ciudad de Albacete (26) se encontró un coeficiente de 0.190 entre material particulado y mortalidad por enfermedades respiratorias; en Madrid, el $PM_{2,5}$ aparece relacionado con los ingresos hospitalarios en la modelización con un riesgo atribuible del 2,7% en infantes (27); en Ciudad de México, las correlaciones encontradas con enfermedades respiratorias (específicamente por IRA debidas al Ozono) son del 0,157, con un retardo de 7 días, y con el bióxido de nitrógeno del 0,163 en un periodo de latencia (o retardo) de tres días (28). Si bien estos estudios no contemplan la misma metodología descrita en este trabajo, los resultados obtenidos tienen un alto grado de similitud en cuanto a la dimensión y los coeficientes encontrados en el nivel de asociación medioambiente-salud.

En cuanto a la predicción, se escogieron los dos retardos para los cuales se encontró una mayor correlación entre $PM_{2,5}$ y asistencias por EPOC. Estos fueron en el día cero y en el día 7. Para cada uno de estos días, se realizó el modelo predictivo utilizando las métricas y las técnicas de modelamiento anteriormente descritas. Los resultados se muestran en la Tabla 4. El mejor resultado se obtuvo usando redes neuronales, en el que se clasificaron correctamente

el 60,85% de los registros. En un estudio similar (17), que evalúa variables meteorológicas en relación con ingresos hospitalarios por enfermedades respiratorias en la ciudad de Curitiba (Brasil), se encontró que el modelo respondía mejor con un retardo de 6 días (similar a este estudio) y con una evaluación de clasificación de la técnica por redes neuronales del 65,86%.

Tabla 4. Resultados del modelamiento predictivo para asistencias por EPOC

Variable objetivo	EPOC_L0	EPOC_L7
Métrica/ Técnica	Redes Neuronales	
% Instancias clasificadas correctamente	60,85	58,47
MAE	0,11	0,12
RMSE	0,27	0,27

Para el caso de asistencias por asma, se encontró la mayor correlación entre $PM_{2.5}$ y asistencias hospitalarias entre las tres morbilidades analizadas. En este caso, la mayor correlación se da para los casos en que el retardo es cero (0,117) y para el día 5 (0,109), como se observa en la Tabla 5. En contraste con las asistencias por EPOC, la variable meteorológica más influyente es el promedio de la velocidad del viento y no la temperatura del aire.

Tabla 5. Correlación del $PM_{2.5}$ con las asistencias por asma

$PM_{2.5}$	ASMA	ASMA_L1	ASMA_L2	ASMA_L3	ASMA_L4	ASMA_L5	ASMA_L6	ASMA_L7
	0,117	0,071	0,076	0,080	0,094	0,109	0,106	0,075

Sin embargo, al realizar la clasificación, las métricas fueron ligeramente inferiores a las obtenidas con EPOC. De acuerdo con la Tabla 6, la técnica con mejor rendimiento sigue siendo redes neuronales (perceptrón multicapa) con un porcentaje de instancias correctamente clasificadas del 54,60% y un retardo de 5 días. Esto se debe a que el rango de asistencias por dicha enfermedad es mucho más bajo que el de EPOC y aún más que el de IRA, lo que aumenta la complejidad de la clasificación.

Tabla 6. Resultados del modelamiento predictivo para asistencias por asma

Variable objetivo	ASMA	ASMA_L5
Métrica/ Técnica	Redes Neuronales	
% Instancias clasificadas correctamente	50,72	54,60
MAE	0,14	0,13
RMSE	0,29	0,29

Finalmente, se realizó el modelamiento para las asistencias por infecciones respiratorias agudas (IRA), las cuales contienen la mayor cantidad de asistencias de las tres morbilidades analizadas (1 203 295 atenciones). En este caso se encontró un coeficiente de correlación entre el $PM_{2.5}$ y las asistencias de 0,087 para el día cero (sin retardo) y del 0,084 para el día dos (Tabla 7). En contraste con EPOC y el asma, el mes, aunque tiene una correlación importante con la variable objetivo, en pocos casos supera la correlación del $PM_{2.5}$, que se ubica como el segundo atributo más importante. La variable meteorológica de mayor correlación es la velocidad del viento, comportamiento similar a las asistencias por asma.

Tabla 7. Correlación de las variables predictoras con las asistencias por IRA

$PM_{2.5}$	IRA	IRA_L1	IRA_L2	IRA_L3	IRA_L4	IRA_L5	IRA_L6	IRA_L7
	0,087	0,078	0,084	0,068	0,059	0,054	0,064	0,071

Con respecto a la clasificación, la técnica de redes neuronales clasificó correctamente el 54,5% de las instancias con un retardo de dos días, similar al caso del asma. Sin embargo, la regresión logística presentó un mejor rendimiento en el modelamiento (56,3%) (Tabla 8).

Tabla 8. Resultados del modelamiento predictivo para asistencias por IRA

Variable objetivo	IRA_L0	IRA_L2
Métrica/técnica	Regresión logística	
% Instancias clasificadas correctamente	50,30	56,30
MAE	0,15	0,15
RMSE	0,28	0,27

DISCUSIÓN

Se encontraron coeficientes de correlación que evidencian la asociación de la concentración de $PM_{2.5}$ con las asistencias por los diagnósticos de IRA, EPOC y asma. La mayor correlación entre las tres morbilidades se presentó para el asma.

La mayor correlación de concentración de $PM_{2.5}$ y número de atenciones se encontró a los cero y dos días de retardo para el IRA, a los cero y siete días para el EPOC y a los cero y cinco días para el asma, tal y como lo mencionaban los estudios realizados en otras ciudades del mundo.

La variable meteorológica de mayor correlación con la variable objetivo es la temperatura del aire para el caso de EPOC y asma. En el caso de IRA, la variable con mayor correlación es la velocidad del viento.

Se evidencia que el comportamiento es similar en el tiempo para las atenciones por las tres morbilidades estudiadas, con valores máximos en el mes de marzo, en el que se encuentran además los valores mayores de concentración por $PM_{2.5}$.

Por otro lado, el día de la semana es una variable de gran importancia a la hora de realizar un estudio de atenciones por enfermedades. En este caso hay una clara disminución de las atenciones el fin de semana, por lo que es una variable que tiene alta correlación, tal y como se evidenció en el modelo predictivo.

Recomendaciones

A la luz de la información recolectada y en vista de que no hay un estudio aplicado en estos términos en el país, es importante destacar que estos resultados son satisfactorios, teniendo en cuenta que dichas asistencias se están explicando por medio de 8 variables que tienen limitaciones. Para complementar este estudio y aumentar el valor de la clasificación sería necesario contar con datos como cuántos de los pacientes que asistieron son fumadores, cuáles son inactivos físicamente y cuántos tienen sobrepeso. Desde el ámbito de ciudad, cuáles son los inventarios de fuentes móviles y fijas, cómo es el comportamiento diario de dichas fuentes, si el día del análisis hay algún evento (como el día sin carro), entre otras ♦

Conflictos de intereses: Ninguno.

REFERENCIAS

1. Naciones Unidas. World Urbanization Prospects: The 2014 Revision [Internet]. New York: ONU; 2014 [cited 2018 Nov 9]. https://bit.ly/2TPNHMq.
2. Desarkar A, Das A. A Smart Air Pollution Analytics Framework. Information and Communication Technology. 2018; 625:197-205. DOI:10.1007/978-981-10-5508-9_19.
3. The Economist. Airborne particles cause more than 3m early deaths a year [Internet]. The Economist; 2017 [cited 2018 Nov 10]. https://econ.st/3v8JUqd.
4. Organización Mundial de la Salud. Lesiones causadas por el tránsito [Internet]. Geneva: OMS; 2017 [cited 2018 Nov 15]. https://bit.ly/350C4Vm.
5. Van der Wall EE. Air pollution: 6.6 million premature deaths in 2050! Neth Heart J. 2015; 23(12):557-8. DOI:10.1007/s12471-015-0763-9.
6. El Mundo. Valle de Aburrá registra el mayor número de muertes por contaminación. Diario El Mundo [Internet]. 2017 [cited 2018 Nov 15]. https://bit.ly/3zcDgCH.
7. Universidad del Valle. Atención de males por calidad del aire cuesta $ 1,6 billones al año [Internet]. Universidad del Valle, Facultad de Salud. 2017 [cited 2018 Nov 15]. https://bit.ly/3x58725.
8. Souza A, Figueredo M, Cacho N, Araújo D, Prolo CA. Using big data and real-time analytics to support smart city initiatives. IFAC-PapersOnLine. 2016; 49(30):257-62. DOI:10.1016/j.ifacol.2016.11.121.
9. Andrienko G, Gunopulos D, Ioannidis Y, Kalogeraki V, Katakis I, Morik K, et al. Mining Urban Data (Part C). Information Systems. 2017;64:219-220. DOI:10.1016/j.is.2016.09.003.
10. Bellinger C, Mohomed Jabbar MS, Zaïane O, Osornio-Vargas A. A systematic review of data mining and machine learning for air pollution epidemiology. BMC Public Health. 2017; 17(1):907. DOI:10.1186/s12889-017-4914-3.
11. Chapman P, Clinton J, Keber R, Khabaza T, Reinartz T, Shearer C, Wirth R. CRISP-DM 1.0 Step by step BIguide. Edited by SPSS; 2000. https://bit.ly/2SdlDCi.
12. Área Metropolitana del Valle de Aburrá. Acuerdo Metropolitano N.° 8 de 25 de marzo de 2011. Gaceta Oficial [Internet]. 2011 [cited 2018 Nov 20]. https://bit.ly/2RFkbs4.
13. García Ávila PA, Rojas NY. Análisis del origen de PM10 y PM2,5 en Bogotá gráficos polares. Rev Mutis. 2016; 6(2):47-58. DOI:10.21789/22561498.1150.
14. Alessandrini ER, Stafoggia M, Faustini A, Berti G, Canova C, De Togni A, et al. Association between short-term exposure to PM2.5 and PM10 and mortality in susceptible subgroups: A multisite case-crossover analysis of individual effect modifiers. Am J Epidemiol. 2016; 184(10):744-54. DOI:10.1093/aje/kww078.
15. Área Metropolitana del Valle de Aburrá. Acuerdo Metropolitano N.° 15 de 2016. Plan Operacional para Enfrentar Episodios Críticos de Contaminación. Gaceta Oficial [Internet] 2016 [cited 2018 Nov 20]. https://bit.ly/3g53y2e.
16. Gironés JC, Minguillón J, Caihuelas JR. Minería de datos: modelos y algoritmos. Monarka; 2017.
17. Polezer G, Tadano YS, Siqueira HV, Godoi AFL, Yamamoto CI, de André PA, et al. Assessing the impact of PM2.5 on respiratory disease using artificial neural networks. Environ Pollut. 2018; 235:394-403. DOI:10.1016/j.envpol.2017.12.111.
18. Mantovani KCC, Nascimento LFC, Moreira DS, Vieira LCPF da S, Vargas NP. Poluentes do ar e internações devido a doenças cardiovasculares em São José do Rio Preto, Brasil. Cien Saude Colet. 2016; 21(2):509-16. DOI:10.1590/1413-81232015212.16102014.
19. Liu H, Tian Y, Cao Y, Song J, Huang C, Xiang X, et al. Fine particulate air pollution and hospital admissions and readmissions for acute myocardial infarction in 26 Chinese cities. Chemosphere. 2018; 192:282-8. DOI:10.1016/j.chemosphere.2017.10.123.
20. Jo YS, Lim MN, Han YJ, Kim WJ. Epidemiological study of PM2.5 and risk of COPD-related hospital visits in association with particle constituents in Chuncheon, Korea. Int J Chron Obstruct Pulmon Dis. 2018; 13:299-307. DOI: 10.2147/COPD.S149469.
21. Ignotti E, Hacon S de S, Junger WL, Mourão D, Longo K, Freitas S, et al. Air pollution and hospital admissions for respiratory diseases in the subequatorial Amazon: a time series approach. Cad Saude Publica. 2010; 26(4):747-61. DOI:10.1590/S0102-311X2010000400017.
22. Miri M, Derakhshan Z, Allahabadi A, Ahmadi E, Oliveri Conti G, Ferrante M, et al. Mortality and morbidity due to exposure to outdoor air pollution in Mashhad metropolis, Iran. The AirQ model approach. Environ Res. 2016; 151:451-7. DOI:10.1016/j.envres.2016.07.039.
23. Pannullo F, Lee D, Neal L, Dalvi M, Agnew P, O'Connor FM, et al. Quantifying the impact of current and future concentrations of air pollutants on respiratory disease risk in England. Environmental Health. 2017; 16(1):29. DOI:10.1186/s12940-017-0237-1.

24. Ministerio de la protección Social. República de Colombia. Lineamientos técnicos para el registro de los datos del registro individual de la prestación de servicios de salud [Internet]. 2009 [cited 2019 Jan 20]. https://bit.ly/3pxM7up.
25. Informe Quincenal Epidemiológico Nacional. Utilidad de los Registros Individuales de Prestación de Servicios (RIPS) para la vigilancia en salud pública [Internet] 2011; 18(16):164 -175. https://bit.ly/3w9vibu.
26. López, MJ. Contaminación Atmosférica, Morbilidad y Mortalidad en la ciudad de Albacete. Rev Clín Med Fam. 2009; 2(8):392-399. https://bit.ly/2TcvAje.
27. Linares C, Díaz J. Efecto de las partículas de diámetro inferior a 2,5 micras (PM2,5) sobre los ingresos hospitalarios en niños menores de 10 años en Madrid. Gaceta Sanitaria. 2009; 23(3):192-197. DOI:10.1016/j.gaceta.2008.04.006.
28. Téllez M, Romieu I, Polo M, Ruiz S, Meneses F, Hernández M. Efecto de la contaminación ambiental sobre las consultas por infecciones respiratorias en niños de la Ciudad de México. Salud Pública de México. 1997; 39(6):513-22. https://bit.ly/3v29JIB.