Nota Científica

Un caso de estudio para la predicción de las partículas contaminantes PM2.5 mediante redes bayesianas

Resumen

La contaminación del aire afecta a la salud pública e individual debido a que incrementa el riesgo de morbilidad y mortalidad. Las partículas suspendidas (PM) son contaminantes del aire que se clasifican en partículas menores a 2.5 micras (PM2.5) y partículas menores a 10 micras (PM10). La exposición a largo plazo de PM2.5 y PM10 provoca reacciones inflamatorias, disminuyendo la inmunidad e incrementando la predisposición a infecciones y al desarrollo de enfermedades crónicas. Varios estudios han enfatizado que la exposición crónica a contaminantes del aire retrasa y/o complica la recuperación de personas infectadas por COVID-19. En este contexto, el objetivo de este estudio consistió en aplicar redes bayesianas para la predicción de los límites máximos establecidos por la Organización Mundial de la Salud (OMS) de las concentraciones promedio de PM2.5 (25 $\mu g/m^3$) en periodos de exposición de 24 horas de la ciudad de Xalapa, Veracruz, México. La predicción del modelo obtenido fue de 82.52%. Los resultados obtenidos, demuestran que la aplicación de técnicas de aprendizaje automático podría coadyuvar al desarrollo de estrategias de control y medidas oportunas orientadas a una efectiva gestión ambiental para enfrentar con éxito futuras epidemias.

Palabras clave: *Partículas suspendidas PM2.5, redes bayesianas, aprendizaje automático, contaminación del aire.*

Introducción

Los contaminantes atmosféricos tienen alto impacto en la salud humana y en el medio ambiente. Las personas que viven en zonas con contaminación del aire son más propensas a padecer enfermedades como asma, enfermedad pulmonar obstructiva crónica o alteraciones del sistema inmunológico. Las partículas finas (líquidos o sólidos suspendidos en el aire), el ozono y otros contaminantes del aire pueden ocasionar procesos de estrés oxidante e inflamación de las vías respiratorias y los pulmones. Las partículas suspendidas (PM) son emitidas por fuentes primarias como los vehículos o la industria, pero también se forman en la atmósfera como resultado de interacciones químicas entre diferentes contaminantes y se catalogan en PM2.5 (menores a 2.5 micras) y PM10 (menores a 10 micras).

De acuerdo con la Organización Mundial de la Salud (OMS), en 2018 las PM estuvieron relacionadas con la muerte prematura de 4.2 millones de personas y específicamente, las PM2.5 son uno de los principales factores de riesgo asociados con los índices de mortalidad debido a que incrementan la inflamación y favorecen la morbilidad causada por virus.

La exposición crónica a partículas PM2.5 o PM10, con la posible unión de diferentes virus o bacterias, tiene un significativo impacto negativo en el sistema inmune humano. Diversos estudios respaldan que la exposición crónica a contaminantes del aire retrasa y/o complica la recuperación de pacientes de COVID-19 y conduce a formas más severas y letales de esta enfermedad (Domingo y Rovira 2020; Jiang et al., 2020; Fattorini y Regoli, 2020).

En el año 2018, las ciudades de Xalapa y Minatitlán sobrepasaron los límites de concentración de PM2.5, PM10, ozono y dióxido de nitrógeno, establecidos por la NOM de calidad del aire (Programa de gestión para mejorar la calidad del aire en el estado de Veracruz de Ignacio de la Llave, 2018). En este sentido, motivados por generar indicadores predictivos que contribuyan al desarrollo de políticas ambientales y sostenibles basadas en la reducción de los niveles de contaminación del aire encauzadas para poder hacer frente a futuras epidemias, se propone la aplicación

de técnicas de aprendizaje automático para la predicción de niveles de concentración de PM 2.5.

El aprendizaje automático (AA) es una subárea de la inteligencia artificial orientada a la generación de modelos computacionales que sean capaces de inducir conocimiento a partir de un conjunto de datos para la predicción, toma de decisiones o descubrimiento de patrones en grandes conjuntos de datos. Algunos de los algoritmos de aprendizaje automático más utilizados son las redes neuronales, redes bayesianas y árboles de decisión.

Respecto al análisis de datos de monitoreo de la calidad del aire se han utilizado diversas técnicas de aprendizaje automático, destacando el uso de redes neuronales (Feng et al., 2015; Franceschi et al, 2018; Cabaneros et al., 2019) y en menor medida, redes bayesianas (Vitolo et al., 2018; Zhou et al., 2020) y árboles de decisión (Deleawe et al., 2010; Desarkar y Das, 2018).

Particularmente, el objetivo de nuestro estudio consistió en aplicar redes bayesianas para la predicción de los límites máximos de las concentraciones promedio de PM2.5 establecidos por la Organización Mundial de la Salud (OMS). Se utilizaron períodos de exposición de 24 horas en la ciudad de Xalapa, de enero a diciembre de 2019, debido a que los efectos de los contaminantes del aire, en la salud humana, dependen primordialmente de la cronicidad a la exposición de concentraciones de PM.

Materiales y métodos

Redes Bayesianas

Una red bayesiana (RB) es un grafo acíclico dirigido donde los nodos (círculos) representan a las variables y los arcos (flechas) representan relaciones de dependencia directa entre las variables. Para cada nodo existe una distribución de probabilidad local que depende del estado de sus padres. Una RB consta de un modelo estructural (cualitativo) que brinda una representación visual de las interacciones entre las variables y de un conjunto de distribuciones locales de probabilidad (cuantitativo), que permite efectuar inferencia probabilística y mide el impacto de una variable o conjuntos de variables sobre otras. Ambas partes determinan una distribución de probabilidad conjunta de las variables de un determinado problema que se puede expresar de manera compacta mediante el uso extensivo de la independencia condicional (Pearl, 1998; Friedman et al., 1997), como se muestra en la ecuación (1).

$$P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | P_a(X_i)) \qquad (1)$$

Donde $P_a(X_i)$ representa al conjunto de padres de X_i (nodos cuyos arcos apuntan a X_i). Dicha ecuación también muestra como recuperar la probabilidad conjunta del producto de distribuciones locales de probabilidad condicional.

Los clasificadores bayesianos son un caso especial de una red bayesiana en la cual existe una variable especial que es la clase y las demás variables son los atributos.

Las redes bayesianas trabajan eficientemente con datos discretos. La discretización es el proceso que consiste en transformar los atributos continuos en un número finito de intervalos, donde cada valor de los atibutos continuos se asigna a un intervalo que incluya a dicho valor.

Selección de la base de datos

El estado de Veracruz cuenta con 3 estaciones automáticas para el monitoreo de la calidad del aire (SEDEMA) ubicadas en Xalapa, Minatitlán y Poza Rica. En el presente estudio, hicimos uso de datos de los contaminantes atmosféricos: monóxido de carbono (CO), óxidos de nitrógeno (NOx), ozono (O3), PM10, PM2.5, dióxido de azufre (SO2) y datos meteorológicos: presión barométrica (PB), precipitación pluvial (PP), temperatura (TMP), direccción del viento (DV) y velocidad del viento (VV) obtenidos del Sistema de Monitoreo de la Calidad del Aire (SINAICA) de la ciudad de Xalapa para el período de enero a diciembre de 2019 (SINAICA, 2020).

Preprocesamiento de datos

Para el análisis de datos, usamos la eliminación de datos para el tratamiento de datos faltantes. Las clases se asignaron de acuerdo con los límites máximos permisibles de las concentraciones promedio de PM2.5 en períodos de exposición de 24 horas (25 $\mu g/m^3$) dictaminados por la Organización Mundial de la Salud (OMS, 2018). Si el promedio de las concentraciones de PM25 en períodos de 24 horas sobrepasan los límites permisibles se clasifica como “mayor”, en caso contrario se clasifica como “menor”. Cuando el número

de muestras de una de las clases (la clase mayoritaria) sobrepasa el número de muestras de la otra (la clase minoritaria), se presenta un desbalanceo de clases que puede producir un deterioro importante en la efectividad del clasificador, en particular con los patrones de las clases menos representadas (Sun, et al., 2009). Para el tratamiento de dicho problema, usamos la técnica de submuestro aleatorio *SpreadSubSample*.

Detalles de implementación

La RB fue generada mediante el uso de WEKA -*Waikato Environment for Knowledge Analysis*- (Waikato Environment for Knowledge Analysis, 2014). Empleamos el método de discretización CAIM (Kurgan y Cios, 2004), *Attribute Selection* y el algoritmo de búsqueda: *RepeatedHillClimber* de *Bayes Net*, con validación cruzada de 10 hojas. El rendimiento de nuestro clasificador está basado en la precisión (número de clasificaciones correctas dividido entre el tamaño del conjunto de prueba).

Resultados

En la Figura 1, se muestra la red bayesiana para la predicción de los límites de las concentraciones promedio de PM2.5 cada 24 horas con una precisión de 82.52%. Como se puede observar, la variable SO2 es el variable padre de la clase PM2.5, adicionalmente existe una relación de dependencia de la clase PM2.5 con las variables: PB, CO, TMP, PM10 y PP.

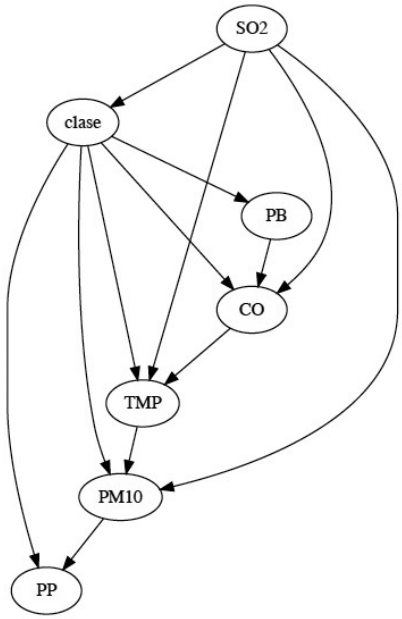

Figura 1. Red bayesiana para la predicción del límite de las concentraciones promedio de PM2.5 cada 24 horas. Los nodos representan las variables y los arcos relaciones de dependencia.

En el modelo obtenido, se clasificaron correctamente 3,050 (82.52%) de 3,696 ejemplos. Del grupo "mayor" al límite de concentraciones promedio de PM.25 cada 24 horas (25 $\mu g/m^3$) se clasificaron correctamente 1,510 ejemplos y 338 ejemplos fueron clasificados erróneamente. En el grupo "menor" (niveles de concentración menores o iguales a 25 $\mu g/m^3$) se clasificaron acertadamente 1540 ejemplares y 308 ejemplos fueron clasificados erróneamente.

Discusión

Gran parte de las partículas comprendidas entre 2.5 y 10 μm son formadas en la atmósfera mediante un proceso químico a partir de gases precursores como dióxido de azufre (SO2), óxidos de nitrógeno (NOx), compuestos orgánicos volátiles (COVs) y amoníaco (NH3). Compuestos como el sulfato y nitrato amónico ($(NH_4)_2SO_4$ y NH_4NO_3) llegan a constituir hasta el 40% de las partículas PM2.5 y se originan por la oxidación en la atmósfera de SO_2 y NO_2 y su interacción con amoníaco (NH_3). Esta reacción es favorecida en condiciones de alta temperatura y humedad y elevada insolación. Los COVs reaccionan en la atmósfera con el ozono (O_3), NOx y otros componentes y generan compuestos carbonosos sólidos y/o líquidos que constituyen alrededor del 25-30% del PM2.5 y PM10 (Environment Canada and Health Canada, 2000; European Commission, 2004).

Las partículas ultrafinas (<0.1 μm) y finas (<= 2.5 μm) se pueden formar en la atmósfera por nucleación de gases como ácido sulfúrico, amoníaco y agua y también por condensación de gases, coagulación de partículas pequeñas, evaporación de neblina y gotas de agua en las que los gases se han disuelto y reaccionado. La formación de estas partículas depende de las condiciones de temperatura, presión y humedad. Así, por ejemplo, el proceso de nucleación se ve favorecido por disminución en la temperatura o aumento en la humedad relativa (Eastern y Peter, 1994; Environmental Protection Agency, 2009).

Los tiempos de residencia de las partículas en la atmósfera puede ser de horas o de varias semanas dependiendo del tamaño de las partículas, pero también de las condiciones de temperatura, presión y humedad. Lo anterior implica que, si hay cambios en la temperatura

y la humedad en la atmósfera, también habrá cambios en las concentraciones de partículas (European Commission, 2004; Environmental Protection Agency, 2009). Como se puede notar, la presencia, formación y tiempo de residencia de partículas como las PM2.5 se ven influenciadas por varios factores como la fuente de origen, los gases precursores, la temperatura, la presión atmosférica, la hora del día y eventualmente por la velocidad del viento. Los resultados obtenidos con las redes bayesianas que se muestran en la Figura 1, son congruentes con lo descrito anteriormente ya que permiten establecer que las variables que impactan directamente sobre la concentración de partículas son la temperatura (TMP), la presión (PB), la lluvia (PP) y las concentraciones de SO2, CO y PM10.

Conclusiones

El modelo cualitativo de nuestra red bayesiana permitió modelar la interacción de las concentraciones de contaminantes del aire y variables meteorológicas para la predicción de los límites de las concentraciones promedio cada 24 horas (25 μg/m3) de PM2.5 decretados por la OMS, donde las variables SO2, TMP, CO, PM10, PB y PP están directamente relacionadas con la clase.

La principal ventaja del uso de redes bayesianas, consistió en modelar las relaciones entre los diferentes factores contaminantes y meteorológicos de las concentraciones de PM2.5. Los resultados obtenidos pueden alentar a los organismos reguladores a promover cambios en las políticas ambientales, ya que el control de la fuente de contaminación lograría la reducción de los efectos nocivos de los contaminantes ambientales para afrontar futuras epidemias.

Bibliografía

Cabaneros, S. M. S., Calautit, J. K., y Hughes, B. R. (2019). A review of artificial neural network models for ambient air pollution prediction. *Environmental Modelling & Software*. 119, 285-304.

Deleawe, S., Kusznir, J., Lamb, B., y Cook, D. J. (2010). Predicting air quality in smart environments. *Journal of ambient intelligence and smart environments*, 2(2), 145-154.

Desarkar, A., y Das, A. (2018). Implementing decision tree in air pollution reduction framework. In Smart Computing and Informatics (pp. 105-113). Springer, Singapore.

Domingo, J. L., y Rovira, J. (2020). Effects of air pollutants on the transmission and severity of respiratory viral infections. *Environmental Research*, 109650.

Eastern, R.C., Peter, L.K. (1994). Binary homogeneous nucleation: temperature and relative humidity fluctuations, nonlinearity, and aspects of new particles production in the atmosphere. *Journal of Applied Meteorology*, 775-784.

Environment Canada and Health Canada. (2000). Priority Substances List Assessment Report, Respirable Particulate Matter Less Than or Equal to 10 Microns. *Canadian Environmental Protection Act,* 1999.

Environmental Protection Agency (2009). Integrated Science Assessment for Particulate Matter. EPA/600/R-08/139F.

European Commission (2004), Second Position Paper on Particulate Matter. CAFE Working Group on Particulate Matter.

Fattorini, D., y Regoli, F. (2020). Role of the chronic air pollution levels in the Covid-19 outbreak risk in Italy. *Environmental Pollution*, 114732.

Feng, X., Li, Q., Zhu, Y., Hou, J., Jin, L., & Wang, J. (2015). Artificial neural networks forecasting of PM2. 5 pollution using air mass trajectory based geographic model and wavelet transformation. *Atmospheric Environment*, 107, 118-128.

Franceschi, F., Cobo, M., y Figueredo, M. (2018). Discovering relationships and forecasting PM10 and PM2. 5 concentrations in Bogotá, Colombia, using artificial neural networks, principal component analysis, and k-means clustering. *Atmospheric Pollution Research*, 9(5), 912-922.

Friedman, N., Geiger D. y Goldszmidt M. (1997). Bayesian networks classifiers. *Machine Learning,* 29,131–163.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I. (2009). The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter*, 11(1).

Jiang, Y., Wu, X. J., & Guan, Y. J. (2020). Effect of ambient air pollutants and meteorological variables on COVID-19 incidence. *Infection Control & Hospital Epidemiology*, 1-11.

Kurgan, L. A. y Cios, K. J. (2004). CAIM Discretization Algorithm. *IEEE Transactions on knowledge and data engineering*, 16, 145-153.

Organización Mundial de la Salud. (2018). Calidad del aire y salud. Recuperado el 2 de marzo de 2020, de: http://www.who.int/es/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health.

Pearl, J. (1998). Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann series in representation and reasoning. Morgan Kaufmann Publishers.

Programa de gestión para mejorar la calidad del aire en el estado de Veracruz de Ignacio de la Llave (2018). Recuperado el 10 de enero de 2020, de https://www.gob.mx/cms/uploads/attachment/file/418382/31_ProAire_Veracruz.pdf.

Sistema Nacional de Información de la Calidad del Aire (SINAICA). (2020). Recuperado el 2 de enero de 2020, de https://sinaica.inecc.gob.mx/.

Sun, Y., Wong, A. K., y Kamel, M. S. (2009). Classification of imbalanced data: A review. *International journal of pattern recognition and artificial intelligence*, 23(04), 687-719.

Vitolo, C., Scutari, M., Ghalaieny, M., Tucker, A., & Russell, A. (2018). Modeling air pollution, climate, and health data using Bayesian Networks: A case study of the English regions. *Earth and Space Science*, 5(4), 76-88.

Zhou, Y., Chang, L. C., & Chang, F. J. (2020). Explore a Multivariate Bayesian Uncertainty Processor driven by artificial neural networks for probabilistic PM2. 5 forecasting. *Science of The Total Environment*, 711, 134792.

Autores:

Sonia Lilia Mestizo Gutiérrez[1*]
Nicandro Cruz Ramírez[2]
Epifanio Morales Zárate[3]
Carlos Roberto Leines Vite[4]

[1, 3, 4] Facultad de Ciencias Químicas
Universidad Veracruzana región Xalapa

[2]Instituto de Investigaciones en Inteligencia Artificial
Universidad Veracruzana región Xalapa

Correspondencia:
*smestizo@uv.mx

Recibido: 18-08-2020 Aceptado:11-12-2020
(Artículo Arbitrado)

Universidad Tecnológica de la Mixteca
UTM

Infraestructura

104 Ha. de dimensión
113 Edificios
48 Laboratorios
9 Talleres
Parque Tecnológico
Parque Solar Fotovoltaico
Agavetum

9 Institutos de Investigación

Instituto de Agroindustrias
Instituto de Computación
Instituto de Ciencias Sociales y Humanidades
Instituto de Diseño
Instituto de Electrónica y Mecatrónica
Instituto de Física y Matemáticas
Instituto de Hidrología
Instituto de Minería
Instituto de Ingeniería Industrial y Mecánica Automotriz

Oferta Educativa

Licenciaturas

Ingeniería en Electrónica
Ingeniería en Computación
Ingeniería en Diseño
Ingeniería en Alimentos
Ingeniería Industrial
Ingeniería en Mecatrónica
Ingeniería en Física Aplicada
Ingeniería en Mecánica Automotriz
Ingeniería Civil
Licenciatura en Ciencias Empresariales
Licenciatura en Matemáticas Aplicadas
Licenciatura en Estudios Mexicanos (modalidad virtual)

Posgrado

Doctorado en Robótica
Doctorado en Modelación Matemática
Doctorado en Tecnologías de Cómputo Aplicado
Doctorado en Electrónica con especialidad en Sistemas Inteligentes Aplicados
Maestría en Robótica
Maestría en Medios Interactivos
Maestría en Administración de Negocios
Maestría en Tecnologías de Cómputo Aplicado
Maestría en Tecnología Avanzada de Manufactura
Maestría en Ciencias: Productos Naturales y Alimentos
Maestría en Modelación Matemática
Maestría en diseño de Muebles
Maestría en diseño de Modas
Maestría en Ciencias de Materiales
Maestría en Electrónica con opción en Sistemas Inteligentes Aplicados
Maestría en Computación con especialidad en Sistemas Distribuidos (modalidad virtual)

INFORMES
Consulta las bases y requisitos en
www.utm.mx