

Arjona Meza, A.E. 2025. Revisión de Artículos que Utilizan Machine Learning para la Predicción de Contaminantes Atmosféricos, p. 215-230. En: J.G. Cerón Bretón, R. M. Cerón Bretón y M. López Cisneros (ed.). Tópicos Selectos de Contaminación Ambiental. Universidad Autónoma de Campeche. 366 p. p. ISBN 978-607-8907-33-5. doi 10.26359/EPOMEX01202510.

Revisión de Artículos que Utilizan Machine Learning para la Predicción de Contaminantes Atmosféricos

Armando Enrique Arjona Meza

³Facultad de Química
Universidad Autónoma del Carmen

Resumen

Las revisiones van más allá de una simple recopilación de artículos, debido a que buscan aportar algo meramente importante dentro de una biblioteca, por ese motivo se requieren cumplir algunos criterios esenciales para poder ser consideradas como revisiones. Este presente trabajo tiene como objetivo principal realizar una revisión descriptiva que permita dar a conocer la utilidad y eficiencia de emplear machine learning para predecir contaminantes en el área ambiental. Hoy en día la contaminación ambiental se está saliendo de control, provocando que cada vez estén más presentes las catástrofes naturales, ésta es la principal motivación de este artículo. Dentro de los estudios revisados mostraron que el uso de técnicas de machine learning, como redes neuronales, regresión lineal y modelos de series de tiempo, ha demostrado ser eficiente en la predicción de la calidad del aire y la concentración de contaminantes. Estos modelos pueden ayudar en la toma de decisiones y la gestión ambiental, permitiendo una respuesta más rápida y precisa a los problemas de contaminación. Estos hallazgos pueden ser de gran relevancia para la comunidad científica y los responsables de la toma de decisiones, ya que pueden contribuir a la implementación de soluciones más efectivas para proteger el medio ambiente.

Palabras clave: Machine Learning, contaminación atmosférica, revisión.

Abstract

The purpose of this work is to conduct a descriptive review that demonstrates the usefulness and effectiveness of using machine learning to predict pollutants in the environment. Environmental pollution is a growing concern, leading to an increase in natural disasters. This article aims to address this issue. The reviewed studies have shown that Machine Learning techniques, such as neural networks, linear regression, and time series models, are effective in predicting air quality and pollutant concentration. These models can aid in decision-making and environmental management, enabling a faster and more accurate response to pollution issues. These findings are relevant to the scientific community and decision-makers, as they can contribute to the implementation of more effective solutions to protect the environment.

Keywords: Machine Learning, Atmospheric pollution, Review.

Introducción

El hecho de realizar una revisión de artículos puede llegarse a pensar que es un trabajo innecesario debido a que cualquier persona puede tomarse el tiempo de leer muchos artículos de un determinado tema para distintos fines, sin embargo, las revisiones van más allá de una simple recopilación, estas requieren que se cumplan algunos criterios para poder aportar algo meramente importante dentro de una biblioteca, logrando de esta manera, que tengan relevancia en las distintas áreas y pueda ser reconocida como un trabajo útil. A continuación, se mencionará los distintos puntos de vista de diversos expertos respecto a qué son las revisiones y cuáles son los criterios para cumplir.

Según Vera (2009), sostiene que el artículo de revisión se considera como un estudio minucioso, selectivo y crítico que integra la información esencial de manera coherente y completa, en una perspectiva global y unificada. Aunque no es un trabajo original, recopila la información más relevante sobre un tema específico. Su objetivo principal es examinar la bibliografía publicada y situarla en una perspectiva adecuada. El autor también destaca que la revisión puede ser vista como un estudio en sí mismo, en el cual el revisor resume y analiza la información disponible sobre un tema específico, basándose en una cuidadosa búsqueda de la literatura médica. Dado que los resultados de los estudios individuales pueden verse influenciados por varios factores, combinar los resultados de diferentes estudios sobre el mismo tema puede ser útil para llegar a conclusiones sobre la evidencia científica relacionada con la prevención, diagnóstico o tratamiento

de una enfermedad específica, u otras áreas de investigación. La diferencia fundamental entre una revisión y un estudio original o primario radica en la unidad de análisis, no en los principios científicos aplicados.

Reyes (2020) señala que “los artículos de revisión son importantes porque contribuyen a la actualización, síntesis y difusión de nuevos conocimientos. Esto es válido no solo para medicina sino para todas las ciencias y las humanidades”. Además, el autor expresa su desacuerdo con la afirmación de que los artículos de revisión sean considerados una forma de plagio encubierto, argumentando que los autores presentan una revisión crítica de la literatura y agregan su propio juicio fundamentado sobre la validez de las conclusiones y recomendaciones.

Cabe mencionar que existen distintos tipos de revisiones, el autor Salamani (2015) establece como la primera a las revisiones “Descriptivas”, que proporcionan una visión general sobre conceptos actuales en áreas en constante evolución y tienen gran utilidad en la enseñanza. La segunda son las revisiones “Exhaustivas”, que son artículos muy especializados y extensos, con trabajos bastante largos sobre un tema específico o perspectiva particular. La tercera son las “Evaluativas”, que responden preguntas muy concretas sobre aspectos etiológicos, diagnósticos, clínicos o tratamientos, utilizando evidencia empírica sólida. Se refiere a las últimas como “Casos clínicos con revisión”, que exponen un caso clínico con intervenciones terapéuticas y revisan la evidencia clínica relacionada con ese caso en particular.

Squires (2015) propone los siguientes. En primer lugar, menciona la revisión exhaustiva de todo lo publicado, que consiste en un artículo de bibliografía comentada. Estos trabajos son extensos y altamente especializados, pero no ofrecen información precisa para responder a una pregunta específica de un profesional interesado. El autor también menciona la revisión descriptiva, que proporciona una actualización sobre conceptos útiles en áreas en constante evolución. Este tipo de revisión es útil tanto para la enseñanza como para aquellos interesados en mantenerse al día en campos relacionados. Asimismo, se destaca la revisión evaluativa, la cual responde a preguntas específicas sobre aspectos etiológicos, diagnósticos, clínicos o terapéuticos. Estas revisiones, conocidas actualmente como preguntas clínicas basadas en la evidencia científica, se centran en una pregunta concreta y ofrecen una evaluación crítica. Si bien, es muy importante resaltar con todas estas citas que las revisiones buscan aportar un juicio y crítica, convirtiéndose algo más allá de ser una simple recopilación; un ejemplo, cuando un experto realiza una revisión de una serie de artículos publicados en cierto año, él puede expresar su opinión respecto a cómo llevaron a cabo sus investigaciones, de igual manera, comparar las metodologías empleadas realizando al final una conclusión, así como también es válido que brinde algunas recomendaciones, todo en pro a mejorar las investigaciones de aquella área o tema.

Hoy en día se está apreciando como el calentamiento global y el efecto invernadero están ocasionando en desmedida una serie de desastres naturales como olas extremas de calor, huracanes, inundaciones y desertifica-

ción, las malas noticias son que se espera que en los años futuros todas estas circunstancias sigan en aumento, una forma metafórica de decirlo es que la tierra va directa hacia un barranco que, a medida que pasan los años, esta comienza a acelerar más hacia el final del camino. Según Grijalva, Jiménez y Ponce (2020)

El problema ambiental generalmente surge como resultado de la alteración de las interacciones entre la sociedad y el medio ambiente. Esta perturbación puede tener múltiples impactos, entre ellos la contaminación del aire, el agua y el suelo. Estos impactos ambientales a su vez tienen diversas implicaciones para la sociedad humana, incluyendo el bienestar, el valor de la naturaleza y la salud humana.

En este trabajo se realizará una revisión descriptiva de una serie de artículos que como requisito debieron ser publicados, como máximo, hace 4 años, con la finalidad de trabajar con información reciente y actualizada, además, deberán tener como tema principal la utilización de machine learning en el área de contaminación ambiental o afines, específicamente enfocándose en temas de ingeniería química. También se contarán con un pequeño análisis y criterio para poder concluir la información, buscando como principal punto el dar a conocer la importancia de utilizar machine learning en las predicciones de contaminantes ambientales, con esto de igual manera se busca denotar que tan eficiente puede ser el empleo de estos métodos para poder brindar confiabilidad y así comenzar a utilizarlos para que poco a poco comience la evolución de las investigaciones, gracias a la asistencia de inteligencia artificial que permitan un análisis más rá-

pido, preciso, confiable, práctico y exacto. En pocas palabras, usar la inteligencia artificial puede potenciar a las investigaciones y como consecuencia, acelerar las soluciones a la contaminación ambiental, por eso el interés de emplearla como herramienta para

ir frenando los cambios catastróficos que se aproximan. El objetivo del presente trabajo fue realizar una revisión descriptiva que permita dar a conocer la utilidad y eficiencia de emplear machine learning para predecir contaminantes en el área ambiental.

Marco teórico

Williams (2013, citado por Palacios y Moreno, 2022) denomina contaminación atmosférica o contaminación ambiental a la presencia de cualquier agente (físico, químico o biológico) o una combinación de varios agentes en lugares, formas y concentraciones tales que sean nocivos para la salud, para la seguridad y el bienestar de la población, o puedan ser perjudiciales para los seres vivos en general. La contaminación ambiental es también la incorporación a los cuerpos receptores de sustancias sólidas, líquidas o gaseosas o la combinación de ellas, siempre que alteren desfavorablemente las condiciones naturales de estos o puedan afectar la salud, la higiene y el bienestar de la población.

El artículo “Efectos de la calidad del aire sobre la salud” de Valls (2017, citado por Castro y Solano, 2021) analiza de manera detallada los agentes contaminantes que afectan la calidad del aire y cómo su deterioro, ya sea debido a causas antropogénicas o naturales, tiene efectos negativos en la salud humana y los ecosistemas. A nivel global, también contribuye al cambio climático. Las causas antropogénicas, es decir, aquellas generadas por la actividad humana, tienen los efectos más significativos y han aumentado en las últimas décadas. Estas fuentes emiten diversos contaminantes a la atmósfe-

ra, como el dióxido de carbono (CO_2), monóxido de carbono (CO), dióxido de azufre (SO_2), monóxido de nitrógeno (NO), dióxido de nitrógeno (NO_2), ozono (O_3) troposférico, amoníaco (NH_3), ácido sulfhídrico (H_2S) y partículas con diferentes medidas y composiciones, incluyendo metales, compuestos inorgánicos y compuestos orgánicos persistentes, así como compuestos orgánicos volátiles. Entre las partículas, su tamaño determina su grado de peligrosidad. Las partículas respirables inferiores a $10\text{ }\mu\text{g}$ son particularmente perjudiciales, ya que pueden ingresar fácilmente al sistema respiratorio. Además, las partículas de tamaño inferior a $2.5\text{ }\mu\text{g}$ pueden pasar desde los alvéolos pulmonares hacia la sangre, afectando órganos, tejidos y células de todo el cuerpo.

En la tesis “Prototipo de uso de machine learning aplicada en el pronóstico de la contaminación por $\text{PM}_{2.5}$ en la localidad de Kennedy de la ciudad de Bogotá” los autores Castro y Solano (2021) presentan una lista breve de los principales contaminantes atmosféricos y sus afecciones a la salud: El dióxido de azufre (SO_2) puede afectar la función pulmonar, empeorar enfermedades respiratorias preexistentes, especialmente la bronquitis, y reducir la capacidad de los pulmones para eliminar partículas extrañas.

El dióxido de nitrógeno (NO_2) puede causar problemas respiratorios, especialmente en personas asmáticas y niños, a diferencia del monóxido de nitrógeno (NO), que es relativamente inofensivo. El ozono (O_3) producido naturalmente en la estratosfera es beneficioso, ya que protege a la Tierra y a sus habitantes de la radiación ultravioleta dañina del sol. El material particulado (MP) consiste en partículas sólidas o líquidas suspendidas en el aire, con una composición química diversa y un tamaño que varía de 0.005 a 100 μm de diámetro aerodinámico. Se originan principalmente por la quema incompleta de combustibles en motores diésel y combustibles sólidos como madera y carbón. También pueden formarse por la condensación de vapores ácidos y compuestos orgánicos semi-volátiles, a través de complejas reacciones del NO_2 y SO_2 en la atmósfera que finalmente generan sales de nitratos y sulfatos. Las partículas más perjudiciales para la salud son aquellas con un diámetro aerodinámico menor a 10 μm (PM_{10}) y, especialmente, aquellas con un diámetro aerodinámico menor a 2.5 μm ($\text{PM}_{2.5}$).

En el estudio de “Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the Global Burden of Diseases Study 2015.” de Cohen, *et al.* (2015, citado por Iskandaryan, Ramos & Trilles, 2020) indica que las partículas $\text{PM}_{2.5}$ fueron el quinto factor de riesgo de mortalidad en 2015. Por lo tanto, prevenir o reducir las consecuencias causadas por la contaminación del aire es un problema crucial. Contar con información sobre la calidad del aire nos inducirá a tomar medidas de protección, lo que puede llevar a la población a realizar sus

actividades diarias en lugares menos contaminados (evitando áreas altamente contaminadas). Sin embargo, el análisis de los datos y la búsqueda de soluciones inteligentes siguen siendo tareas desafiantes. Por lo tanto, es fundamental aplicar métodos y técnicas productivas para analizar de manera más efectiva y eficiente los grandes volúmenes de datos, convirtiendo lo invisible en visible y extrayendo la información oculta detrás de los datos.

De los problemas y definiciones mencionados anteriormente, podemos observar que la calidad del aire se considera un componente esencial en el concepto de ciudad inteligente. La calidad del aire se ha convertido en un problema masivo en muchas áreas. Según la Organización Mundial de la Salud (OMS), cada año más de siete millones de personas mueren a causa de este problema, y más del 80% de la población de áreas urbanas vive en lugares donde la calidad del aire supera los límites establecidos por la OMS (Apte *et al.*, 2018).

La ciencia de datos se considera el cuarto paradigma científico, sucediendo al experimento, la teoría y las ciencias computacionales. Se refiere al análisis de datos como una ciencia empírica, donde se aprende directamente de los propios datos. Esto puede tomar la forma de recopilación de datos seguida de un análisis abierto sin preconcepciones (conocido como descubrimiento o exploración de datos). El segundo método empírico implica formular una hipótesis, recopilar datos nuevos o existentes para abordar la hipótesis, y analizar y confirmar o refutar la hipótesis (o determinar la necesidad de información o estudio adicional). En ambos métodos, las conclusiones se

basan en los datos. En muchos proyectos de ciencia de datos, se examinan los datos originales primero, lo que ayuda a formular una hipótesis que luego se investiga. Como en cualquier ciencia experimental, el resultado puede llevar a la reformulación de la hipótesis original. El concepto clave es que la ciencia de datos es una ciencia empírica que realiza el proceso científico directamente sobre los datos. Es importante tener en cuenta que la hipótesis puede surgir de una necesidad o ser una reafirmación de una necesidad en términos de una hipótesis técnica. La ciencia de datos implica extraer conocimientos útiles directamente de los datos mediante el descubrimiento o la formulación y prueba de hipótesis (Castro & Solano, 2021).

Según Hinestroza y Cárdenas (2018), en sectores como el informático, salud, corporativo y transporte, se ha encontrado una solución a la tarea de aprendizaje y predicción mediante el uso del *Machine Learning*, una disciplina de la inteligencia artificial. Esta herramienta busca mejorar el análisis de datos para lograr predicciones futuras, ya sea a través de la implementación de nuevos sistemas o la mejora de los existentes, utilizando algoritmos basados en información antigua o reciente para optimizar el rendimiento del sistema. Por otro lado, *Machine Learning*, promete ser útil en campos como la medicina, robótica y mecánica. Sin embargo, existe preocupación acerca de los posibles daños que esta herramienta pueda tener. La creación de organismos mecánicos más inteligentes y la reducción de la necesidad de intervención humana, plantea la incertidumbre de volverse mentalmente obsoletos debido a una dependencia excesiva de estos sistemas.

Los autores Castro y Solano (2021) mencionan que el machine learning abarca diversas aplicaciones en áreas como motores de búsqueda, diagnósticos médicos, detección de fraude en tarjetas de crédito, análisis del mercado de valores, clasificación de secuencias de ADN, reconocimiento del habla y lenguaje escrito, juegos y robótica. Para abordar estos temas, es fundamental comprender los diferentes tipos de problemas de machine learning. En el aprendizaje supervisado, se entrena al algoritmo utilizando datos etiquetados con respuestas correctas. Después del entrenamiento, el algoritmo utiliza la experiencia adquirida para predecir resultados cuando se le presentan nuevos datos sin etiquetas. En el aprendizaje no supervisado, el algoritmo se entrena utilizando datos sin etiquetas, con el objetivo de encontrar patrones que ayuden a comprender el conjunto de datos. En el aprendizaje por refuerzo, el algoritmo aprende observando el entorno y recibiendo retroalimentación basada en sus acciones, además de que lo hace a través de ensayo y error, seleccionando y perfeccionando estrategias que llevan al éxito. El sobreajuste es un fenómeno en el que los modelos de machine learning se ajustan demasiado a los datos de entrenamiento, sin generalizar correctamente para nuevos datos, esto puede llevar a resultados imprecisos y falta de capacidad para predecir. Para evitar el sobreajuste, se pueden aplicar estrategias como la retención de datos y la validación cruzada, que ayudan a reducir este fenómeno y mejorar la generalización del modelo.

Dentro de los algoritmos supervisados, destacan aquellos utilizados en investigaciones que buscan realizar predicciones basadas en datos. Entre ellos se encuentran el

Support Vector Machine (svm), las Redes Neuronales, los Árboles de Decisión y los Bosques Aleatorios, entre otros. Una revisión reciente mostró que el svm y las Redes

Neuronales están entre las técnicas más populares en el área de la salud (Jiang 2017, citado por Pedrero *et al.* (2021)).

Métodología

La revisión de artículos es de tipo descriptiva, que consiste en realizar una síntesis de los artículos seleccionados que irá acompañado de una crítica que tiene como finalidad resaltar la eficiencia e importancia de cómo se empleó el machine learning para predecir las contaminaciones ambientales.

La selección de artículos siguió los siguientes lineamientos:

Se realizó una búsqueda exhaustiva de artículos científicos utilizando términos clave relacionados con machine learning, predicción de contaminantes ambientales y sus diferentes compuestos. Estos términos incluirán palabras clave como “*Machine Learning*”, “*contaminantes ambientales*”, “predicción”, “modelos predictivos”, entre otros.

Se priorizaron los repositorios de acceso abierto y reconocidos en el área científica, como bases de datos académicas y revistas científicas de renombre. Algunos ejemplos pueden incluir Springer, Redalyc, Scielo, ScienceDirect y Google Scholar.

Se establecieron criterios claros para la selección de artículos. Se incluirán aquellos que estén relacionados con la aplicación de machine learning para predecir datos de

contaminantes ambientales, sin discriminar ningún método específico de machine learning. Se excluyeron aquellos artículos que no cumplieron con los criterios establecidos, que no son accesibles en los repositorios seleccionados o que tienen una antigüedad mayor a 4 años.

Se llevó a cabo una revisión sistemática de los artículos encontrados, utilizando los términos clave y los criterios de inclusión y exclusión establecidos. Se extrajo la información relevante de cada artículo, como los métodos de machine learning utilizados, los compuestos contaminantes abordados, los resultados obtenidos y cualquier otro aspecto pertinente.

Se realizó un análisis crítico de los artículos seleccionados, identificando las tendencias, hallazgos y limitaciones comunes. Se sintetizó la información recopilada en forma de resumen, tablas comparativas o cualquier otro formato adecuado.

Los resultados de la revisión literaria se presentaron de manera clara y concisa, destacando las principales conclusiones y recomendaciones encontradas en los artículos revisados.

Resultados y discusión

El primer artículo para revisar fue el de “Satellite-based estimation of NO_2 concentrations using a machine-learning model: A case study on Rio Grande do Sul, Brazil” de Becerra, *et al.* (2023), los autores plantean que, en muchas áreas de Brasil, las mediciones de NO_2 a nivel del suelo enfrentan dificultades debido a la escasez y distribución desigual de estaciones de monitoreo. Por lo tanto, sugieren que las observaciones satelitales combinadas con modelos de aprendizaje automático pueden ayudar a superar esta falta de datos. Este estudio investiga la capacidad de un enfoque de aprendizaje automático (un algoritmo estadístico no lineal de Bosques Aleatorios, RF) para reconstruir la distribución espaciotemporal a largo plazo de NO_2 a nivel del suelo desde 2006 hasta 2019. Los resultados muestran que el modelo RF predice el NO_2 con precisión, con un coeficiente de determinación $R^2 = 0.68$, utilizando validación cruzada de 10 pliegues. También se observa una tendencia decreciente en la concentración total de NO_2 en la región estudiada, con el año 2006 presentando las concentraciones más altas y 2017 las más bajas. Este estudio resalta que los algoritmos estadísticos no lineales como RF pueden ser herramientas complementarias a las observaciones in situ y satelitales para el mapeo de NO_2 .

El siguiente artículo fue “Prototipo de red neuronal artificial para el pronóstico de eventos críticos por partículas $\text{PM}_{2.5}$ en el centro de la ciudad de Manizales.” de García (2021), aquí se presenta un prototipo de red neuronal artificial con el objetivo de prevenir enfermedades respiratorias a partir de los

altos índices de contaminación atmosférica. Para ello, se utilizaron ocho variables para el pronóstico de calidad de aire, y se subdividió el conjunto de datos en train y en test. Se corrieron diferentes algoritmos para distintas configuraciones y conjunto de variables predictoras, y se determinó que, a pesar de haber utilizado modelos más complejos de redes neuronales recurrentes, como los de capas apiladas LSTM o CRU, se obtuvo un mejor resultado con modelos más simples basados en una única capa con neuronal LSTM.

El mejor modelo hallado utilizando redes neuronales completamente conectadas, y poco profundas, no logró capturar más del 40% de la varianza del modelo real. Por lo tanto, para futuros modelos comentaron los autores que es deseable aplicar nuevas arquitecturas de redes neuronales, como las redes neuronales recurrentes (RNN) o las redes neuronales convolucionales que han demostrado lograr mejores resultados en este tipo de problemáticas. Como conclusión los autores expresaron que el prototipo de red neuronal aplicada para predecir el contaminante $\text{PM}_{2.5}$ demostró las posibilidades sobre las que se puede trabajar al aplicar métodos de machine learning en los distintos aspectos que nos afectan como seres humanos. A través de la creación de una red neuronal artificial, se logra predecir con éxito el comportamiento de la variable $\text{PM}_{2.5}$, lo que puede ser de gran ayuda para la comunidad en la prevención de enfermedades respiratorias. Este artículo ofrece una revisión bibliográfica sobre la contaminación del aire y sus efectos en la salud humana, lo que permite contextualizar el problema y la importancia

de la investigación. Además, se destaca la importancia de la subdivisión del conjunto de datos en train y en test, lo que permite evaluar la eficiencia de los modelos, algo muy bueno a su favor, sin embargo, podría haber incluido una discusión más amplia sobre las limitaciones y posibles mejoras de los modelos utilizados.

El artículo “Modelo de Predicción Material Particulado ($PM_{2.5}$) en Bogotá” de Salas (2022) presenta un modelo que se basa en la regresión lineal y las series de tiempo para predecir la concentración de $PM_{2.5}$ en el aire. Los resultados muestran que el modelo logra una predicción precisa de la calidad del aire, con un error medio del 5.6. Además, se destaca la importancia de considerar variables como el tráfico, la lluvia, la radiación uv y la temperatura para reducir aún más el error en la predicción. En general, el estudio representa una alternativa prometedora para la predicción del material particulado, lo que permitiría establecer rutas de acción para mitigar el daño causado por estos contaminantes en el medio ambiente. Como parte de la crítica del artículo, el uso de las técnicas de regresión lineal y series de tiempo es adecuado para el análisis de datos de series temporales y la predicción de valores futuros. Además, se destaca que en este artículo mencionaron las variables que fueron clave para que el modelo tenga una buena predicción, así como dicen la importancia de considerar variables adicionales para reducir aún más el error de esta. Sin embargo, es importante tener en cuenta que el modelo presentado en el artículo se basa en datos históricos y no tiene en cuenta factores externos que pueden afectar la calidad del aire, como cambios en la política ambiental

o eventos climáticos extremos. Por lo tanto, es necesario tener precaución al utilizar este modelo como la única fuente de información para la toma de decisiones.

El artículo “Predicción de la contaminación del aire por partículas (PM_{10}) mediante machine learning k-nearest neighbors” de Botero y Sánchez (2024) analiza el uso del aprendizaje automático para predecir la contaminación del aire, específicamente las partículas PM_{10} , en el área de Kennedy de Bogotá, Colombia. El estudio empleó el algoritmo de los k-vecinos más cercanos (KNN) para analizar datos históricos obtenidos de una estación de monitoreo de calidad del aire. El algoritmo KNN es un método de aprendizaje supervisado que clasifica nuevas muestras en función de su similitud con los puntos de datos existentes. En este caso, se utilizó para predecir los niveles de PM_{10} en el área de Kennedy encontrando los “k” puntos de datos más cercanos en el conjunto de datos históricos y utilizando sus valores para estimar el nuevo nivel de PM_{10} . Los resultados mostraron que el algoritmo KNN pudo predecir con precisión los niveles de PM_{10} con un alto grado de confianza. Los autores señalaron que la precisión de las predicciones dependía de la calidad de los datos y la selección del valor apropiado para “k”. En general, el estudio demostró el potencial del uso de algoritmos de aprendizaje automático, como KNN, para predecir los niveles de contaminación del aire. La precisión de las predicciones puede proporcionar información valiosa para la toma de decisiones y el desarrollo de políticas relacionadas con la gestión de la calidad del aire. Sin embargo, los autores también señalaron la importancia de considerar las limitaciones del modelo

y la necesidad de un monitoreo continuo y análisis de datos de calidad del aire. En términos de eficiencia, los algoritmos de aprendizaje automático pueden procesar grandes cantidades de datos de manera rápida y precisa, lo que los convierte en una herramienta útil para predecir los niveles de contaminación del aire en tiempo real. Además, el uso del aprendizaje automático puede ayudar a identificar patrones y tendencias en los datos de contaminación del aire que pueden no ser evidentes de inmediato mediante el análisis estadístico tradicional.

El posterior artículo se titula “*Air pollution prediction with machine learning: a case study of Indian cities*” de Kumar y Pande (2022), aquí se presenta un estudio sobre la predicción de la calidad del aire en 23 ciudades de la India utilizando técnicas de aprendizaje automático. Los investigadores utilizaron seis años de datos de contaminación del aire y aplicaron técnicas de preprocesamiento y selección de características para identificar los contaminantes que afectan directamente el índice de calidad del aire. Luego, se emplearon cinco modelos de aprendizaje automático para predecir la calidad del aire, y se compararon los resultados utilizando métricas estándar. El modelo de Naive Bayes Gaussiano obtuvo la mayor precisión, mientras que el modelo de Máquina de Vectores de Soporte tuvo la menor precisión.

El modelo XGBoost tuvo el mejor rendimiento en comparación con los otros modelos y mostró la mayor linealidad entre los datos predichos y los datos reales. De manera general, el estudio demuestra que las técnicas de aprendizaje automático son herramientas eficientes para predecir la calidad del aire y pueden ser útiles para monitorear

y prevenir la contaminación del aire en las ciudades. Hablando críticamente, el estudio presenta una metodología clara y detallada sobre cómo se preprocesaron los datos de contaminación del aire y se seleccionaron las características relevantes para la predicción de la calidad del aire. Además, se utilizaron varios modelos de aprendizaje automático para predecir la calidad del aire, y se compararon los resultados utilizando métricas estándar. Sin embargo, como se menciona en el artículo, el estudio tiene algunas limitaciones, como por ejemplo, que se centró solo en ciudades de la India, por lo que los resultados pueden no ser generalizables a otras regiones o países. Además, el estudio no aborda la cuestión de cómo se pueden utilizar los resultados de la predicción para mejorar la calidad del aire en las ciudades. Ya para finalizar, se considera que este artículo es una buena introducción al uso de técnicas de aprendizaje automático para predecir la calidad del aire en ciudades de la India. Sin embargo, se necesitan más investigaciones para abordar las limitaciones y explorar cómo se pueden utilizar los resultados de la predicción para mejorar la calidad del aire en las ciudades. Además, como mirándolo desde el área de ingeniería química, sería interesante explorar cómo se pueden utilizar los resultados de la predicción para diseñar y optimizar sistemas de control de la contaminación del aire en las ciudades.

El artículo: “*Prediction of Air Quality Index Using Machine Learning Techniques: A Comparative Analysis*” de Gupta *et al.* (2023) presenta un análisis comparativo de diferentes modelos de machine learning para predecir el índice de calidad del aire (AQI) en ciudades de la India. Los autores utilizaron

tres modelos de minería de datos: SVR, RFR y CR, y aplicaron la técnica de sobremuestreo de minorías sintéticas (SMOTE) para equilibrar los datos de las clases y obtener resultados más precisos y consistentes. Además, compararon los resultados de los conjuntos de datos equilibrados e inequilibrados utilizando métodos estadísticos como RMSE, MAE, MSE y R-SQUARE para confirmar la precisión de los resultados. Los resultados mostraron que el modelo SVR fue el más preciso en la predicción del AQI en las ciudades estudiadas, seguido por RFR y CR. Además, los autores encontraron que la técnica SMOTE mejoró significativamente la precisión de los modelos. Por último, el estudio destaca la importancia de equilibrar los datos de las clases y utilizar técnicas de machine learning para predecir el AQI con mayor precisión. Los resultados pueden ser útiles para mejorar la monitorización de la calidad del aire y la salud pública en las ciudades de la India.

Esta es una investigación interesante y relevante para la mejora de la calidad del aire en las ciudades de la India, en cuanto a los modelos de machine learning utilizados, SVR, RFR y CR son técnicas comunes en la minería de datos y han sido aplicadas en diversos campos. Sin embargo, es importante destacar que la elección de los modelos depende de la naturaleza de los datos y el problema a resolver. En este caso, los autores justifican su elección y presentan resultados comparativos que respaldan la eficacia del modelo SVR en la predicción del AQI. La técnica SMOTE utilizada para equilibrar los datos de las clases es una estrategia útil para mejorar la precisión de los modelos, especialmente cuando se trata de conjuntos de datos desequilibrados. Sin embargo, es importante

tener en cuenta que esta técnica puede generar datos sintéticos que no reflejan la realidad y pueden afectar la interpretación de los resultados. Se puede destacar que este artículo es una contribución valiosa para la investigación en el campo de la calidad del aire y el machine learning. A pesar de que los resultados se basan en datos específicos de las ciudades de la India y pueden no ser generalizables a otras regiones o países.

El siguiente artículo "Machine Learning-Based Prediction of Air Quality" de Liang, *et al.* (2020) se enfoca en el desarrollo de modelos de predicción del índice de calidad del aire (AQI) para eventos de contaminación aguda del aire con anticipación de 1, 8 y 24 horas. Se investigaron cinco algoritmos de aprendizaje automático (ML): bosques aleatorios, Adaptive Boosting (AdaBoost), máquina de vectores de soporte, red neuronal artificial y métodos de ensamblaje de apilamiento para entrenar modelos. Además, se observó cómo el rendimiento de la predicción disminuye con el tiempo y se midió la precisión con tres índices de error dependientes de la escala comúnmente utilizados: error absoluto medio (MAE), error cuadrático medio (RMSE) y R-cuadrado (R^2). Los autores encontraron que el modelo de ensamblaje de apilamiento fue el más eficiente para predecir AQI con anticipación de 1 hora, mientras que el modelo AdaBoost fue el más eficiente para predecir AQI con anticipación de 8 y 24 horas. Además, se encontró que el rendimiento de la predicción disminuyó con el tiempo, y que los modelos de predicción con pasos de tiempo más largos (F8-AQI y F24-AQI) se beneficiaron de la optimización de hiperparámetros y el uso de ensamblajes de apilamiento, AdaBoost y

bosques aleatorios. Los resultados sugieren que los modelos de aprendizaje automático pueden ser una herramienta efectiva para predecir AQI con anticipación de 1, 8 y 24 horas, lo que puede ser útil para la toma de decisiones en la gestión de la calidad del aire. Como crítica constructiva, hay que señalar que el artículo no profundiza en la selección de variables y la calidad de los datos utilizados para entrenar los modelos.

Además, no se menciona si se realizaron pruebas de validación cruzada para evaluar la capacidad de generalización de los modelos. Estos son aspectos importantes para considerar en futuras investigaciones para garantizar la robustez y la precisión de los modelos de predicción.

El empleo de técnicas de machine learning para la predicción de contaminantes ambientales ha cobrado relevancia en los últimos años debido a la necesidad de monitorear y controlar la calidad del aire en las ciudades y zonas industriales. Los estudios revisados abordan esta problemática desde diferentes enfoques y algoritmos, resaltando las ventajas y limitaciones de cada uno.

El artículo de Becerra *et al.* (2023) destaca el potencial de utilizar observaciones satelitales combinadas con modelos de aprendizaje automático para superar las limitaciones en la medición de contaminantes como el NO_2 a nivel del suelo. La precisión del modelo de Bosques Aleatorios (RF) para predecir las concentraciones de NO_2 demuestra la viabilidad de esta aproximación en la reconstrucción espaciotemporal de datos de contaminación atmosférica.

Por otro lado, el estudio de García (2021) sobre la predicción de eventos críticos de partículas $\text{PM}_{2.5}$ resalta la importancia de

utilizar modelos de redes neuronales artificiales. Aunque se observa que modelos más complejos no necesariamente mejoran la precisión de la predicción, este artículo subraya la capacidad de las técnicas de machine learning para prevenir enfermedades respiratorias al predecir la concentración de contaminantes.

El artículo de Salas (2022) presenta un enfoque diferente al utilizar modelos de regresión lineal y series de tiempo para predecir la concentración de $\text{PM}_{2.5}$. Aunque muestra una predicción precisa, su limitación radica en la falta de consideración de factores externos que podrían afectar la calidad del aire. Esto destaca la necesidad de integrar más variables y considerar el contexto ambiental más amplio en los modelos de predicción.

El estudio de Botero y Sánchez (2024) sobre la predicción de partículas PM_{10} en Bogotá demuestra la eficacia del algoritmo K-Nearest Neighbors (KNN) en la predicción precisa de niveles de contaminación del aire. Sin embargo, se reconoce la importancia de considerar las limitaciones del modelo y la necesidad de un monitoreo continuo para mejorar la calidad del aire.

El artículo de Kumar y Pande (2022) presenta un análisis exhaustivo sobre la predicción de la calidad del aire en ciudades de la India, utilizando diversas técnicas de aprendizaje automático. Aunque se destaca la precisión de ciertos modelos, se reconoce la limitación de la generalización de los resultados a otras regiones o países.

Finalmente, el estudio de Liang *et al.* (2020) sobre la predicción del AQI con anticipación de 1, 8 y 24 horas destaca la eficacia de los modelos de ensamblaje de apilamiento y AdaBoost. Sin embargo, se señala

la falta de profundización en la selección de variables y la necesidad de realizar pruebas de validación cruzada para garantizar la robustez de los modelos.

De acuerdo con estos artículos, se puede hacer un análisis más a detalle, en el que se muestra que los algoritmos no lineales como los bosques aleatorios (RF), las redes neuronales artificiales (ANN) y los modelos de ensamblaje han demostrado ser efectivos en la predicción de contaminantes como el dióxido de nitrógeno (NO_2), las partículas $\text{PM}_{2.5}$ y PM_{10} , y el índice de calidad del aire (AQI). Específicamente, los modelos RF y ANN lograron coeficientes de determinación (R^2) de hasta 0.68 y 0.4, respectivamente, en la predicción de estos contaminantes. Sin embargo, la precisión de los modelos depende en gran medida de la calidad y cantidad de datos disponibles, así como de la selección adecuada de variables predictoras. Algunos estudios destacan la importancia de incluir variables meteorológicas, de tráfico y de radiación solar, entre otras, para mejorar el rendimiento de las predicciones.

Otro aspecto relevante es el equilibrio de los conjuntos de datos, donde técnicas como

SMOTE (Synthetic Minority Over-sampling Technique) han demostrado mejorar la precisión de los modelos al abordar problemas de clases desbalanceadas. No obstante, también se evidencian limitaciones en cuanto a la generalización de los modelos a diferentes regiones geográficas y la falta de discusión sobre cómo utilizar los resultados de las predicciones para implementar acciones concretas en la gestión de la calidad del aire.

En general, los estudios revisados demuestran el potencial del machine learning como una herramienta complementaria a las mediciones in situ y satelitales para el mapeo y predicción de contaminantes ambientales. Sin embargo, es fundamental abordar las limitaciones mencionadas y continuar investigando nuevas arquitecturas y técnicas que permitan obtener modelos más precisos y generalizables.

Además, es necesario considerar la interpretabilidad de los modelos y su integración con otros enfoques, como la simulación de calidad del aire y la evaluación de impacto ambiental, para lograr una gestión integral de la contaminación atmosférica y sus efectos en la salud pública y el medio ambiente.

Conclusiones

En conclusión, la revisión descriptiva realizada sobre la utilidad y eficiencia de emplear machine learning para predecir contaminantes en el área ambiental ha proporcionado una visión amplia y actualizada sobre este tema. A través de la síntesis y análisis de varios artículos científicos, se han identificado diferentes métodos y algoritmos de machine learning utilizados en la predicción de con-

taminantes ambientales. Se ha observado que el uso de técnicas de machine learning, como redes neuronales, regresión lineal, modelos de series de tiempo y ensamblajes de modelos, ha demostrado ser eficiente en la predicción de la calidad del aire y la concentración de contaminantes en diferentes contextos y regiones. Los estudios revisados destacan la importancia de considerar varia-

bles relevantes, realizar un preprocesamiento adecuado de los datos y utilizar técnicas de selección de características para mejorar la precisión de los modelos de predicción. Además, se ha evidenciado que la aplicación de machine learning puede ayudar en la toma de decisiones y la gestión ambiental, permitiendo una respuesta más rápida y precisa a los problemas de contaminación. Sin embargo, se ha identificado la necesidad de profundizar en aspectos como la interpretación de los resultados, la escalabilidad de los modelos y la consideración de factores externos que puedan influir en la calidad del aire. Asimismo, se ha señalado la importancia de utilizar repositorios confiables y reconocidos

en el área científica para la selección de los artículos.

En pocas palabras, la revisión literaria ha proporcionado evidencia sólida sobre la utilidad y eficiencia de emplear machine learning en la predicción de contaminantes ambientales. Estos hallazgos pueden ser de gran relevancia para la comunidad científica, los responsables de la toma de decisiones y todos aquellos involucrados en la gestión y prevención de la contaminación ambiental. Se espera que esta revisión contribuya a la evolución de las investigaciones en este campo y a la implementación de soluciones más efectivas para proteger nuestro medio ambiente.

Literatura citada

- Apte, J. S., Brauer, M., Cohen, A. J., Ezzati, M., & Pope, C. A., III. (2018). Ambient PM_{2.5} reduces global and regional life expectancy. *Environmental Science & Technology Letters*, 5, 546–551.
- Becerra-Rondón, A., Ducati, J., & Haag, R. (2023). Satellite-based estimation of NO₂ concentrations using a machine-learning model: A case study on Rio Grande do Sul, Brazil. *Atmosfera*, 37(1), 111-125.
- Botero Mármol, V., & Sánchez Martín, I. D. (2024). Predicción de la contaminación del aire por partículas (PM₁₀) mediante machine learning K-Nearest Neighbors. Universidad Libre - Facultad de Ingeniería, 13(23), 57-64.
- Castro, E. & Solano, C. (2021). Prototipo de uso de machine learning aplicada en el pronóstico de la contaminación por pm_{2.5} en la localidad de Kennedy de la ciudad de Bogotá. (Tesis). Universidad Distrital Francisco José de Caldas. Bogotá, Colombia.
- García, D. (2021). Prototipo de red neuronal aplicada para predecir el contaminante PM_{2.5} en el centro de la ciudad de Manizales. *Revista Científica de Ingeniería*, 2(2).
- Grijalva, A. de las M., Jiménez, M. E., & Ponce, H. X. (2020). Contaminación del agua y aire por agentes químicos. *RECIMUNDO*, 4(4), 79-93.
- Guirao Goris, S. J. A. (2015). Utilidad y tipos de revisión de literatura [Usefulness and types of literature review]. *Ene.*, 9(2). Recuperado de <https://dx.doi.org/10.4321/S1988-348X2015000200002>
- Gupta, N., Mohta, Y., Heda, K., Armaan, R., Valarmathl, B. & Arulkumaran, G. (2023). Prediction of Air Quality Index Using Machine Learning Techniques: A Comparative Analysis. *Journal of Environmental and Public Health*.
- Hinestroza, D. & Cárdenas, J. (2018). El machine learning a través de los tiempos, y los aportes a la humanidad. (Tesis). Universidad Libre Seccional Pereira. Bogotá, Colombia.

- Iskandaryan, D., Ramos, F., & Trilles, S. (2020). Air Quality Prediction in Smart Cities Using Machine Learning Technologies Based on Sensor Data: A Review. *Applied Sciences*, 10(7).
- Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., et al. (2017). Artificial intelligence in healthcare: past, present and future. *Stroke and Vascular Neurology*, 2(4), 230-243.
- Kumar, K., & Pande, B. (2022). Air quality prediction using machine learning techniques: A case study of Indian cities. *International Journal of Environmental Science and Technology*, 20(12), 5333-5348.
- Liang, Y., Maimury, Y., Hsiang-Ling, A. & Cuevas, J. (2020). Machine learning-based prediction of air quality. *International Journal of Environmental Research and Public Health*, 17(7).
- Martí Valls, J. (2017). Efectos de la calidad del aire sobre la salud. *FMC - Formación Médica Continuada en Atención Primaria*, 24(9), 511-514.
- Palacios, I & Moreno, D. (2022). Contaminación ambiental. *Revista Científica Mundo de la Investigación y el Conocimiento*. 903-10. <https://recimundo.com/index.php/es/article/view/1545/1979>
- Pedrero, V., Reynaldós, K., Ureta, J., & Cortes, E. (2021). Generalidades del Machine Learning y su aplicación en la gestión sanitaria en Servicios de Urgencia. *Revista Médica de Chile*. 149(2).
- Reyes, H. (2020). Artículos de revisión. *Revista médica de Chile*. 148(1).
- Salas, L. (2022). Predicción de la concentración de material particulado en Bogotá mediante técnicas de machine learning. *Avances: Investigación en Ingeniería*, 17(2), 57-62.
- Vera, O. (2009). Cómo escribir artículos de revisión. *Revista Médica La Paz*. 15. http://www.scielo.org.bo/scielo.php?script=sci_arttext&pid=S1726-89582009000100010