Diseño de un Modelo Predictivo de Machine Learning para la Estimación de Concentraciones de PM2.5 en Áreas Urbanas

Valeria Carvajal Galleguillos

Profesor Guía: Werner Kristjanpoller

Universidad Técnica Federico Santa María
Mayo 2024

1 Introducción

La calidad del aire es un determinante crítico de la salud pública y el bienestar ambiental. Entre los contaminantes atmosféricos, las partículas finas, en particular las PM2.5 (partículas con un diámetro aerodinámico menor a 2.5 micrómetros), son de gran interés y relevancia de estudio, debido a que tienen la capacidad de ingresar profundamente en los pulmones y entrar en el torrente sanguíneo, afectando no solo el sistema respiratorio, sino también el sistema cardiovascular.

Estudios epidemiológicos han demostrado que la exposición prolongada a altas concentraciones de PM2.5 puede aumentar el riesgo de enfermedades respiratorias, cardiovasculares y cáncer de pulmón. El incremento en las tasas de morbilidad y mortalidad genera la necesidad urgente de un sistema de monitoreo y de herramientas de predicción eficaces de estas partículas (Smith, 2019; Liu, 2020). Por otra parte, a nivel global, las concentraciones de PM2.5 varían significativamente, influenciadas por factores tanto naturales como antropogénicos, tales como la meteorología y las actividades industriales y vehiculares. Este dinamismo en las concentraciones hace que la predicción precisa de PM2.5 sea un desafío considerable y de gran relevancia para la implementación de políticas de salud y medioambientales (Chang, 2021). Ambas problemáticas subrayan la importancia de desarrollar modelos predictivos precisos y robustos que puedan anticipar las concentraciones de PM2.5 con suficiente antelación para implementar medidas de mitigación efectivas. Este enfoque no solo mejora la calidad del aire y la salud pública, sino que también proporciona una herramienta valiosa para la formulación de políticas ambientales basadas en datos.

El uso de técnicas de machine learning en la predicción de PM2.5 ha emergido como una solución prometedora, al contar con la capacidad de manejar grandes volúmenes de datos y modelar complejidades no lineales (Zhou, 2018). Sin embargo, la efectividad de diferentes algoritmos de machine learning y su aplicación en distintos contextos geográficos y temporales aún requiere una exploración detallada. La variabilidad en los factores que afectan las concentraciones de PM2.5, como las condiciones meteorológicas, las fuentes de emisión y las características geográficas, puede influir significativamente en el rendimiento de los modelos predictivos. Por ejemplo, estudios han mostrado que algoritmos como los bosques aleatorios y las redes neuronales profundas pueden capturar mejor las relaciones no lineales y las interacciones entre múltiples variables predictoras en comparación con los modelos lineales tradicionales (Liang, 2020).

Tomando en cuenta lo anterior, este estudio se enfoca en diseñar y validar un modelo predictivo basado en machine learning para estimar las concentraciones de PM2.5, utilizando un conjunto de datos que incluye variables meteorológicas y datos sobre actividades humanas. La investigación no solo contribuirá al cuerpo académico existente mediante la comparación de varios algoritmos de aprendizaje automático, sino que también proporcionará insights prácticos

para mejorar los sistemas de alerta temprana y las estrategias de mitigación de la contaminación del aire.

2 Objetivos

2.1 Objetivo General

Diseñar un modelo predictivo de machine learning que estime de manera precisa las concentraciones diarias de PM2.5 en áreas urbanas, proporcionando un instrumento predictivo que mejore la gestión de la calidad del aire.

2.2 Objetivos Específicos

1. Recopilar y procesar datos históricos de concentraciones de PM2.5, variables meteorológicas y datos sobre actividades humanas relacionadas con la emisión de partículas.
2. Realizar un análisis exploratorio para identificar las relaciones y patrones entre las concentraciones de PM2.5 y las variables predictoras.
3. Implementar y evaluar diferentes algoritmos de machine learning, incluyendo regresión lineal, bosques aleatorios y redes neuronales, para determinar el modelo más preciso y eficiente.
4. Validar el modelo predictivo utilizando un conjunto de datos de prueba y evaluar su rendimiento mediante métricas estadísticas.

3 Metodología

Recopilación de Datos

Para desarrollar un modelo predictivo robusto, es fundamental comenzar con una recopilación exhaustiva de datos de diversas fuentes relevantes. En el caso de las concentraciones de PM2.5, los datos se obtendrán de estaciones de monitoreo ambiental proporcionados por organismos gubernamentales como el Ministerio del Medio Ambiente de Chile, así como por instituciones académicas y de investigación. Estos datos deben abarcar un período de al menos 10 años para capturar adecuadamente las variaciones estacionales y de largo plazo. Además, se utilizarán datos con una frecuencia diaria para captar las fluctuaciones diarias en las concentraciones, lo que es crucial para un modelo predictivo preciso.

Junto con los datos, se recopilarán variables meteorológicas que afectan la dispersión y concentración de las partículas. Estas variables incluirán la temperatura máxima y mínima diaria, la humedad relativa, la velocidad y dirección del viento, la presión atmosférica y las precipitaciones diarias. Estos datos se obtendrán de servicios meteorológicos nacionales y regionales. La incorporación de estas variables meteorológicas es esencial, ya que factores como la temperatura y la humedad pueden influir significativamente en los niveles de particulado fino.

Asimismo, es importante incluir datos sobre actividades humanas que contribuyen a la emisión de partículas. Estos datos abarcarán el tráfico vehicular diario en las principales vías urbanas, obtenidos de registros de tráfico y conteos automáticos, así como datos sobre producción y emisiones industriales diarias provenientes de reportes de plantas industriales. Otros datos relevantes pueden incluir información sobre actividades de construcción y quemas agrícolas. Estos datos complementarios permitirán al modelo capturar una imagen más completa de las fuentes de contaminación.

Preprocesamiento de Datos

El preprocesamiento de datos es una etapa crítica para asegurar que los datos recopilados sean adecuados para el análisis y la construcción del modelo. Este proceso comienza con la limpieza de datos, que implica la identificación y eliminación de valores atípicos mediante técnicas estadísticas. Además, se manejarán los datos faltantes utilizando métodos como la imputación con la media o el *K-NN imputing*, asegurando así que el conjunto de datos sea completo y representativo.

Una vez limpiados los datos, se procederá a la transformación de los mismos. Las variables se normalizarán utilizando técnicas como *Min-Max Scaling o Z-score*, lo que garantiza que todas las variables contribuyan equitativamente al modelo. La técnica de *Min-Max Scaling* se define como:

$$X' = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}} \tag{1}$$

donde X' es el valor escalado, X es el valor original, X_{min} es el valor mínimo de la variable y X_{max} es el valor máximo de la variable. Por otro lado, la normalización *Z-score* se define como:

$$X' = \frac{X - \mu}{\sigma} \tag{2}$$

donde X' es el valor normalizado, X es el valor original, μ es la media de la variable y σ es la desviación estándar de la variable. Adicionalmente, se aplicarán transformaciones logarítmicas a las variables sesgadas para normalizar su distribución. Estas transformaciones son fundamentales para mejorar la performance del modelo y evitar problemas de escala y distribución.

El *feature engineering* también juega un papel importante en el preprocesamiento de datos. Se generarán nuevas características basadas en combinaciones de variables existentes, como el índice de calidad del aire promedio de los días anteriores. Además, se crearán variables derivadas, como la diferencia de temperatura entre el día y la noche o el índice de ventilación, que es el producto de la velocidad del viento y la altura de la capa de mezcla. Estas nuevas variables pueden mejorar significativamente la capacidad predictiva del modelo.

Análisis Exploratorio de Datos (EDA)

El análisis exploratorio de datos (EDA) es una etapa crucial para entender los datos y descubrir patrones y relaciones significativas. Este análisis comenzará con la visualización de datos mediante histogramas, boxplots y scatter plots para visualizar la distribución de las variables y detectar valores atípicos. Estas visualizaciones ayudan a identificar la naturaleza de los datos y las posibles anomalías que podrían afectar el modelo.

Además de la visualización, se realizará un análisis correlacional para identificar las relaciones lineales entre las variables independientes y las concentraciones de PM2.5. Una matriz de correlación permitirá observar qué variables tienen una relación más fuerte, lo que puede guiar la selección de características para el modelo.

El análisis de tendencias también es una parte importante del EDA. Se analizarán las tendencias temporales y estacionales en las concentraciones de PM2.5 utilizando series temporales. Además, se realizará un análisis de componentes principales (PCA) para reducir la dimensionalidad de los datos y identificar patrones subyacentes. Este análisis permite simplificar el modelo sin perder información relevante, mejorando así su eficiencia y precisión.

Implementación de Modelos de Machine Learning

Para la implementación del modelo predictivo, se seleccionarán varios algoritmos de machine learning que han demostrado ser efectivos en problemas similares. Estos algoritmos incluyen la regresión lineal, los bosques aleatorios

y las redes neuronales profundas. La regresión lineal se utilizará como línea base, ya que es un modelo simple que proporciona una referencia para comparar otros modelos más complejos. Los bosques aleatorios se seleccionan por su capacidad para capturar relaciones no lineales y manejar la variabilidad en los datos, mientras que las redes neuronales profundas son adecuadas para modelar complejidades y relaciones no lineales avanzadas.

El entrenamiento de los modelos implicará dividir los datos en conjuntos de entrenamiento (70%) y prueba (30%). Esta división asegura que el modelo se entrene en una parte de los datos y se evalúe en otra, lo que ayuda a evitar el sobreajuste y proporciona una evaluación más precisa de su rendimiento. Los modelos se entrenarán utilizando el conjunto de entrenamiento y se ajustarán los hiperparámetros mediante técnicas de validación cruzada, como la validación cruzada *k-fold*.

Las redes neuronales profundas se implementarán utilizando capas completamente conectadas (*Fully Connected Layers*) y funciones de activación no lineales como la función ReLU (*Rectified Linear Unit*). La salida de una neurona j en una capa oculta se calcula como:

$$a_j = f\left(\sum_{i=1}^{n} w_{ij}x_i + b_j\right) \tag{3}$$

donde f es la función de activación (ReLU en este caso), w_{ij} son los pesos sinápticos, x_i son las entradas de la neurona, y b_j es el sesgo (bias). La función de activación ReLU se define como:

$$f(x) = \max(0, x) \tag{4}$$

El error cuadrático medio (MSE) se utilizará como la función de pérdida para entrenar las redes neuronales, definida como:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{5}$$

Validación y Optimización

La validación y optimización del modelo son etapas cruciales para asegurar su robustez y generalizabilidad. Se aplicará la validación cruzada k-fold para dividir los datos en múltiples subconjuntos y entrenar y evaluar el modelo en diferentes combinaciones de estos subconjuntos. Esta técnica ayuda a evitar el sobreajuste y proporciona una evaluación más confiable del rendimiento del modelo.

La optimización de los hiperparámetros se realizará utilizando técnicas como *Grid Search o Random Search*. Estas técnicas permiten encontrar la mejor combinación de hiperparámetros que maximicen el rendimiento del modelo. La optimización adecuada de los hiperparámetros es esencial para mejorar la precisión y eficiencia del modelo predictivo.

Comparación y Selección del Modelo

Después de entrenar y evaluar varios modelos, se compararán sus resultados para seleccionar el más preciso y eficiente. La comparación se basará en las métricas de evaluación y en la capacidad del modelo para generalizar a nuevos datos. El modelo seleccionado debe no solo ser preciso, sino también eficiente en términos de tiempo de cómputo y recursos necesarios.

Una vez seleccionado el modelo, se procederá a su implementación en un entorno de producción. Este proceso incluirá el despliegue del modelo en un sistema que permita realizar predicciones diarias de concentraciones de PM2.5. Se

desarrollará una interfaz de usuario que facilite la visualización de las predicciones y la emisión de alertas tempranas cuando se prevean niveles peligrosos de PM2.5.

Además, se integrará el modelo predictivo con sistemas de alerta temprana para informar a la población y a las autoridades sobre posibles niveles peligrosos de contaminación. Esta integración es crucial para que las predicciones del modelo se utilicen de manera efectiva en la toma de decisiones y en la implementación de medidas de mitigación.

Validación Final y Documentación

Finalmente, se realizará una validación independiente del modelo utilizando un conjunto de datos no utilizado durante el proceso de entrenamiento. Esta validación final asegurará que el modelo implementado sea robusto y generalizable a diferentes condiciones y contextos.

Se documentará todo el proceso metodológico, desde la recopilación de datos hasta la implementación y validación del modelo. Se elaborará un informe detallado que incluya los resultados y conclusiones del estudio, así como recomendaciones para futuras investigaciones y aplicaciones prácticas. La documentación adecuada es esencial para garantizar la reproducibilidad del estudio y para proporcionar una guía clara para otros investigadores o profesionales que deseen utilizar o mejorar el modelo.

References

Chang, H., L. C. . W. J. (2021). The impact of meteorological factors on pm2.5 concentrations: A case study in beijing, china. *Atmospheric Environment*, 246:118096.

Liang, Y., Z. L. C. X. . Z. Y. (2020). Assessing the performance of different machine learning algorithms for predicting pm2.5 concentration. *Environmental Modelling & Software*, 124:104600.

Liu, Y., C. R. . Z. Y. (2020). Long-term exposure to pm2.5 and incidence of cardiovascular diseases: A cohort study in china. *Environmental Health Perspectives*, 128(7):077007.

Smith, K. R., J. M. . A. H. R. (2019). Health effects of fine particulate air pollution: Lines that connect. *Journal of the Air & Waste Management Association*, 59(6):674–690.

Zhou, J., W. X. . G. H. (2018). Application of machine learning models for forecasting pm2.5 concentration in beijing, china. *Science of the Total Environment*, 636:1021–1029.