Análisis de la Correlación entre Variables Atmosféricas y Contaminantes mediante Técnicas de Machine Learning para la Predicción de la Calidad del Aire en el Oriente Antioqueño

Seminario de Analítica y Ciencia de Datos

# Resumen Descriptivo del Proyecto

[Este proyecto busca identificar la correlación entre variables y niveles de contaminantes atmosféricos en el oriente antioqueño, con el fin de brindar herramientas a entidades ambientales para gestionar la calidad del aire.]

Marco Teórico:

1. Revisión de Literatura.

[Se realizó una revisión de estudios publicados que encajaran con los objetivos de nuestra propuesta, resultado de la búsqueda se eligieron los siguientes:

El estudio del **año 2022**, publicado en la revista Sustainability llamado **Machine Learning for Determining Interactions between Air Pollutants and Environmental Parameters in Three Cities of Iran**, presenta un enfoque sólido basado en machine learning para analizar y predecir la calidad del aire en ciudades iraníes, Los estudios más relevantes que se mencionan incluyen investigaciones sobre la interacción entre la contaminación del aire y factores meteorológicos, como los trabajos de Zhou et al. (2017) y Yang et al. (2018), que exploraron cómo las condiciones climáticas afectan la concentración de contaminantes.

El objetivo principal era analizar las interacciones entre los contaminantes del aire (como PM2.5, PM10, NO2, SO2, O3, y CO) y parámetros ambientales (temperatura, humedad, velocidad del viento, presión atmosférica y vegetación) en tres ciudades iraníes (Teherán, Tabriz y Shiraz), entre 2015 y 2019, utilizando técnicas de machine learning.

La conclusión del estudio fue que la precisión limitada del modelo subraya la complejidad del problema, destacando que la contaminación del aire está influenciada por una combinación de factores meteorológicos, geográficos e industriales que no fueron completamente capturados en este estudio. Aunque la vegetación mostró efectos positivos, su capacidad para mitigar la contaminación resultó ser menor de lo esperado, sugiriendo la necesidad de explorar otras soluciones más integrales.

1. El estudio del **año 2022**, publicado en la revista Sustainability llamado **Long-Term Correlations and Cross-Correlations in Meteorological Variables and Air Pollution in a Coastal Urban Region**, realiza un análisis detallado y bien fundamentado de la persistencia de la contaminación del aire y su relación con factores climáticos, utilizando técnicas estadísticas avanzadas que permiten una mejor comprensión de las dinámicas a largo plazo.

El artículo aborda el problema de la contaminación atmosférica en una región costera urbana, específicamente en el área del puerto de Aratu en el estado de Bahía, Brasil. El objetivo es analizar las correlaciones a largo plazo entre las variables meteorológicas (irradiación solar, temperatura, humedad relativa y velocidad del viento) y los contaminantes atmosféricos (ozono (O3), óxidos de nitrógeno (NOx) y partículas PM10) para evaluar cómo estos factores interactúan y persisten en el tiempo. Sin embargo, una limitación del estudio es la falta de datos a largo plazo, ya que se analizaron solo datos de 2019.

Además, el estudio se centra en un área geográfica restringida, lo que podría limitar la generalización de los resultados a otras regiones con diferentes características meteorológicas o industriales. En general, el artículo contribuye significativamente al campo de la ciencia atmosférica, proporcionando un enfoque novedoso para evaluar las interacciones entre contaminación y clima en una región costera industrial. Las recomendaciones prácticas derivadas del análisis, como la necesidad de mitigar las emisiones de NOx, son particularmente útiles para la formulación de políticas ambientales más efectivas.

1. El estudio del **año 1998**, publicado en la revista Atmospheric Environment llamado **Principal Component and Canonical Correlation Analysis for Examining Air Pollution and Meteorological Data**, el artículo aborda el problema de la contaminación del aire en la ciudad de Atenas, Grecia y cómo ésta se relaciona con variables meteorológicas. Utiliza técnicas avanzadas de análisis de datos multivariados (PCA y CCA) para explorar las relaciones entre las concentraciones de contaminantes y las condiciones meteorológicas. El objetivo identificar patrones que ayuden a entender las fuentes de contaminación y su comportamiento en diferentes condiciones climáticas, en verano e invierno. Realiza un uso efectivo de técnicas multivariadas para simplificar y encontrar patrones en los datos complejos de contaminación y meteorología.

El uso del Análisis de Componentes Principales (PCA) y Análisis de Correlación Canónica (CCA) permite descubrir relaciones ocultas que no serían evidentes con análisis simples. Sin embargo, la correlación baja entre los conjuntos de datos de contaminación y meteorología (0.1–0.4) sugiere que otros factores no considerados podrían influir en la contaminación atmosférica en la región. Además, aunque los resultados identifican patrones importantes, la resolución espacial y temporal del estudio se limita a una única estación de monitoreo en Atenas, lo que podría no capturar variaciones en otras partes de la ciudad.

En resumen, el artículo proporciona una visión valiosa sobre la dinámica de la contaminación del aire en Atenas, pero futuras investigaciones deberían incorporar más estaciones de monitoreo y otras variables, como las fuentes móviles y estacionarias de emisiones, para obtener una imagen más completa del problema.

1. En el estudio del **año 2017** llamado **Estimación de la calidad del aire en la cuidad de Talca utilizando algoritmos de aprendizaje automático, se enfocan en un problema estacional (Otoño-invierno)** **de contaminación del arie** **por material particulado**, asociado a las características meteorológicas y altas emisiones de calefactores domiciliarios a leña. El trabajo se enfoca en generar un modelo de predicción de la calidad el aire en la cuidad de Talca, localizada en la zona central de Chile, prediciendo con un día de anticipación el material particulado de diámetro igual o menor a 2.5 micrómetros, utilizando tres modelos de aprendizaje Perceptrón Multicapa (MLP), redes neuronales supervisadas basadas en Self-Organizing (SOM) y Random Forest.

Los modelos utilizan los registros históricos de calidad del aire de los meses de abril y agosto de los años 2014 y 2015, además de registros de la velocidad del viento y temperatura, recolectados por tres estaciones de medición en la cuidad. El autor compara los resultados de los modelos de regresión de los tres algoritmos de aprendizaje automático, analizando la capacidad de predecir episodios de emergencia ambiental, en donde se evidencia que el modelos MLP y el algoritmos TTOSOM (basado en Self-Organizing) tienen resultados similares donde las dos primeras estaciones dan resultados confiables pero para la tercera estación presentan mediciones peak más pronunciados, lo que aumenta el rango entre las mediciones de concentración de material particulado.

En cambio, con los algoritmos de Random Forest los resultados para las estaciones de medición fueron significativamente superiores a los obtenidos por los dos algoritmos neurales artificiales. Por tal motivo se concluyó que el algoritmo de Randon Forest supero los problemas señalados en el estudio, reduciendo el porcentaje de errores de subestimación.

1. El estudio del año publicado en el año **2022** en la revista **Advances in Meteorology llamado**​ **The Influence of Data Length on the Performance of Artificial Intelligence Models in Predicting Air Pollution** se tiene como objetivo evaluar cómo la longitud de los datos afecta la precisión de modelos de IA para predecir la concentración de PM2.5 en Dorset, Ontario, Canadá. Dado que la contaminación del aire es un problema crítico de salud pública, la predicción precisa de PM2.5 permite tomar medidas tempranas para mitigar sus efectos. El proyecto se centra en la comparación de tres modelos de IA: Máquina de Aprendizaje Extremo (ELM), Red Neuronal del Método de Manejo de Datos (GMDHNN) y Árbol de Regresión de Potenciación de Gradiente (GBR).

Para analizar el impacto de la proporción de datos, los investigadores utilizaron nueve divisiones diferentes de entrenamiento/prueba, variando entre 50/50 y 90/10. Los resultados mostraron que ELM se destacó en precisión y adaptabilidad, especialmente con una proporción de 60/40, gracias a su capacidad para manejar variaciones en los datos de entrenamiento y prueba. Este hallazgo sugiere que una división óptima y el uso de modelos IA robustos son esenciales para predicciones precisas en escenarios ambientales, con posibles aplicaciones en la gestión de calidad del aire y salud pública

1. En el estudio de Vásquez Arenas, J. P. (2022) publicado en el **CESET** se realizó un análisis en el Valle de Aburrá, en Colombia, centrado en los beneficios y limitaciones del machine learning para predecir concentraciones de PM2.5. El trabajo revisó múltiples casos de éxito publicados en revistas indexadas, en donde se evaluaron diferentes algoritmos incluyendo redes neuronales y modelos de ensamble como AdaBoost y XGBoost, indicando que estos métodos son efectivos en contextos urbanos donde las emisiones varían considerablemente, contemplando las características topográficas, meteorológicas y de emisión del valle de aburra. Determinando que no se tiene un algoritmo especifico tuviese un mejor desempeño para este tipo de aplicación.]
2. Modelos y Métodos ya existentes.

[En el análisis crítico que se realizó a los estudios anteriormente mencionados se prestó un mayor interés en cuales modelos y métodos se utilizaron para la solución de su problema.

**Long-Term Correlations and Cross-Correlations in Meteorological Variables and Air Pollution in a Coastal Urban Region** :

**Métodos utilizados:** Este estudio emplea el Análisis de Fluctuación Detrendizado (DFA) para identificar dependencias a largo plazo en las concentraciones de contaminantes (O₃, NOx, PM₁₀) y el Análisis de Correlación Cruzada Detrendizado (ρDCCA) para estudiar correlaciones entre esas concentraciones y variables meteorológicas como la radiación solar, la temperatura, la humedad y la velocidad del viento. Estos métodos permiten descomponer las series temporales y detectar correlaciones no estacionarias.

**Datos utilizados**

Los datos utilizados en el estudio provienen de tres estaciones de monitoreo de calidad del aire ubicadas en ciudades de Brasil (Botelho, Gamboa y Malembá). Se recopilaron datos horarios sobre concentraciones de contaminantes y variables meteorológicas durante todo el año 2019. Estos datos incluyen:

*Contaminantes:* O3, NOx y PM10.

*Variables meteorológicas:* Humedad relativa, irradiación solar, temperatura del aire y velocidad del viento.

El estudio utiliza dos técnicas estadísticas avanzadas, como se mencionó anteriormente, estas se utilizaron de la siguiente manera:

La metodología incluye dividir los perfiles de las series temporales en ventanas de tiempo, aplicar el DFA para detectar la persistencia y usar el ρDCCA para cuantificar las correlaciones cruzadas.

El estudio propone que las agencias ambientales deberían enfocarse en reducir las emisiones de NOx, un precursor del ozono, especialmente en áreas industriales como el puerto de Aratu. Las correlaciones a largo plazo entre las variables meteorológicas y la contaminación permiten prever episodios de mala calidad del aire, lo que podría guiar futuras políticas de mitigación.

**Útil o descartable para nuestra solución propuesta:** Dado que en nuestro estudio los datos incluyen mediciones horarias, el Análisis de Fluctuación Detrended (DFA) puede ser útil para identificar tendencias persistentes en los niveles de contaminación y cómo evolucionan con el tiempo, especialmente para las series de tiempo no estacionarias, como los datos de calidad del aire que vamos a utilizar. En el caso de métodos detallados, como el DCCA multivariante que es usado en conjuntos de datos climatológicos o financieros más complejos, podrían no aportar un valor significativo y se podrían dejar de lado.

**Machine Learning for Determining Interactions between Air Pollutants and Environmental Parameters in Three Cities of Iran**

**Datos utilizados**

*Datos de contaminación del aire:* Registrados por el sistema de monitoreo del Departamento de Medio Ambiente de Irán, que recopila información sobre los niveles de CO, O3, NO2, SO2, PM10, PM2.5, y el Índice de Calidad del Aire (AQI).

*Datos meteorológicos:* Incluyen variables como la temperatura, humedad relativa, velocidad del viento y presión atmosférica, extraídos de estaciones meteorológicas automatizadas.

*Datos de vegetación:* Se utilizó el índice de vegetación de diferencia normalizada (NDVI) para cuantificar el estado de la vegetación y su posible impacto en la calidad del aire.

**Metodología de solución: El** estudio emplea el modelo XGBoost, un algoritmo de machine learning basado en la técnica de "boosting", que es una mejora de los árboles de decisión. XGBoost fue elegido por su capacidad para manejar grandes conjuntos de datos y ofrecer predicciones más precisas en comparación con otros algoritmos. La metodología incluyó:

*Análisis estadístico:* Se utilizó el coeficiente de correlación de Pearson para evaluar las relaciones entre los contaminantes del aire y los parámetros ambientales.

*Interpolación espacial y mapeo:* Se utilizó el método de interpolación Kriging para zonificar el AQI en las ciudades estudiadas, y se emplearon imágenes de satélite para analizar la densidad de vegetación mediante NDVI.

*Entrenamiento y optimización del modelo:* Se usó el método de Grid Search para optimizar los hiperparámetros del modelo XGBoost, con el fin de mejorar su precisión en la predicción de la calidad del aire.

**Útil o descartable para nuestra solución propuesta:** Para nuestro objetivo de estudio, los enfoques de Machine Learning como XGBoost pueden llegar a ser útiles para el análisis, especialmente dado los volúmenes de datos por hora que vamos a utilizar y las relaciones complejas entre las variables. Sin embargo, se puede utilizar el análisis estadístico de Pearson como primera instancia para obtener una idea general de las correlaciones y generar una visual de cómo se comportan las variables.

**Principal Component and Canonical Correlation Analysis for Examining Air Pollution and Meteorological Data:**

**Datos utilizados**

Los datos utilizados en el estudio provienen de un período de cinco años (1988-1992) y se recolectaron en una estación de monitoreo de Atenas, cerca de una zona industrial. Incluyen:

*Contaminantes atmosféricos:* CO, NO, NO2, O3, humo y SO2, con mediciones cada hora (y una medición diaria para el humo).

*Variables meteorológicas:* Humedad relativa, temperatura, duración del sol, velocidad del viento y dirección del viento, también medidas cada hora. Las estaciones meteorológicas y de contaminación se ubicaban a 1.7 m una de la otra.

**Metodología de solución:** El análisis se realizó por separado para los datos de verano (abril a octubre) y de invierno (noviembre a marzo), utilizando dos técnicas clave:

*Análisis de Componentes Principales (PCA):* Se empleó para reducir la dimensionalidad de los datos y encontrar combinaciones lineales de las variables originales que expliquen la mayor varianza en los datos. Estas combinaciones permitieron identificar las principales fuentes de contaminación y las condiciones meteorológicas predominantes.

*Análisis de Correlación Canónica (CCA):* Se utilizó para analizar la relación entre las variables meteorológicas y las concentraciones de contaminantes, determinando qué patrones meteorológicos coinciden con niveles altos de contaminación.

*Se utilizó un software llamado PONTOS*, desarrollado internamente, para realizar estos análisis.

**Útil o descartable para nuestra solución propuesta**: La PCA puede ser útil porqué permitirá simplificar los datos meteorológicos y de contaminantes, identificando variables clave que explican la mayor parte de la variabilidad en nuestro estudio, es aplicable para el manejo de grandes cantidades de datos.

En el caso de a CCA es útil debido a que puede revelar las interrelaciones más fuertes entre las condiciones meteorológicas y las concentraciones de contaminantes, lo cual es directamente relevante con nuestro objetivo de correlación.

**Estimación de la calidad del aire en la cuidad de Talca utilizando algoritmos de aprendizaje automático,** tenemos:

**Datos utilizados:**

*Datos de contaminación del aire:* Datos registrados por el Sistema de Información Nacional del Aire (SINCA) recopilada desde abril de 2014 a agosto de 2015 de las estaciones La Florida, U.C Maule y Universidad de Talca, con variables como MP2.5.

*Datos meteorológicos* como velocidad de viento y temperatura bajo las mismas condiciones mencionadas anteriormente, tanto para estaciones como para rango de tiempo.

**Metodología Utilizada:**

* Se realizó primero una normalización de las variables de concentración de partículas y registros meteorológicos dado que presentaban una distribución log-normal, luego fueron llevados a un rango de 0-1, restando a cada valor el mínimo del parámetro y dividiéndolo por la resta entre el valor máximo y mínimo, para obtener una mejor distribución de los datos.
* Se reemplazan los datos faltantes por la media muestral o moda, e igualmente ajustando las variables dummmy. Para luego proceder a eliminar los valores atípicos.
* Análisis de Medidas de precisión para modelos de regresión: Se utilizan cuatro métricas R2 ajustada, MAPE, RSME, RSE, para tener la diferencia entre el valor predicho y el valor medio para las etapas de entrenamiento y validación de los algoritmos escogidos a estudiar.
* Modelación de Algoritmo TTOSOM derivado de (SOM): Se representan los utilizando una topología de árbol. Utilizan una búsqueda de grilla de parámetros para cada base de datos buscando el conjunto de parámetros con mejores resultados según las medidas de precisión.
* Modelación con algoritmo Perceptrón Multicapa (MLP): Dado que es un método supervisado no permite datos faltantes, utilizan el método de descartar todos los registros faltantes, igualmente que en el de TTOSOM se utiliza el método de grilla para establecer los mejores registros para hacer la predicción.
* Modelación de Algoritmos Random Forest (RF): Se realizan pruebas con 100, 500 y 1000 arboles con 2 y 3 variables. en las tres estaciones de monitoreo, se utiliza la base de datos con datos faltantes.

**Útil o descartable para nuestra solución propuesta:** Según los resultados del estudio por la falta de confiabilidad en las metricas de R2 y MAPE se podrían descartar utilizar en nuestra solución los modelos de TTOSOM y MLP, ajustándonos a una mejor solución el Randon Forest por sus mejores resultados en el periodo de tiempo a escenarios de predicción de alertas.

Para la monografía **The Influence of Data Length on the Performance of Artificial Intelligence Models in Predicting Air Pollution** se tiene:

**Datos utilizados:**

Para este estudio se recopilaron datos horarios de la concentración de PM2.5 de la estación de monitoreo Dorset, situada en Ontario, entre el 1 de enero de 2011 y el 31 de diciembre de 2020. La estación, de tipo rural y a 318 metros de altura, registró PM2.5 en microgramos por metro cúbico cada hora.

**Metodología utilizada:**

Se implementaron técnicas de limpieza para datos perdidos (1.78% de los registros) y valores anómalos, aplicando interpolación lineal por ser el método más preciso para este conjunto de datos. Se emplea algoritmos de machine learning, incluyendo ELM, GMDHNN y GBR. Cada modelo fue entrenado y evaluado usando distintas proporciones de datos de entrenamiento/prueba, optimizando la precisión mediante el ajuste de hiperparámetros en fases de entrenamiento. La normalización de los datos fue esencial para garantizar que los modelos mantuvieran estabilidad y precisión durante el proceso de predicción. Las evaluaciones se realizaron mediante métricas como el error medio absoluto (MAE), el error cuadrático medio (RMSE) y el índice de eficiencia de Nash-Sutcliffe (NSE). Los resultados indicaron que ELM ofrece predicciones de mayor precisión, especialmente en divisiones de datos de 60/40, y se adaptó mejor a cambios en la longitud del conjunto de datos que GMDHNN y GBR.

En conclusión, estos datos representan un importante punto de análisis al reflejar una variabilidad que abarca todas las estaciones del año, permitiendo así una evaluación robusta de las predicciones a lo largo del tiempo y bajo diferentes condiciones meteorológicas

**Útil o descartable para nuestra solución propuesta:** Según las metodologías utilizadas se podría utilizar la interpolación lineal para la limpieza de datos y el modelo de ELM por su mayor precisión, descartando los demás modelos.

6. **Beneficios de los modelos basados en algoritmos de machine learning para la predicción de concentraciones de PM2.5 en el Valle de Aburrá**

**Datos utilizados:**

*Recolección de Estudios:* Basados en estudios de los años de 2020 a 2023 se escogieron los estudios más sobresalientes en la predicción de la calidad del aire en sitios con características físicas similares a las del Valle de Aburra

*Datos*: Tipo de Monitoreo, resolución temporal, fuente de datos, periodo de entrenamiento, testeo, tipo de predicciones, alerta de episodios de mala calidad.

**Metodología Utilizada:**

El estudio se desarrolló en dos fases principales. La primera fase fue una revisión exhaustiva de artículos recientes (2019-2021) sobre el uso de machine learning para la predicción de PM2.5, analizando características físicas y de emisión en el Valle de Aburrá. La segunda fase consistió en el análisis de casos con condiciones similares, enfocándose en datos de entrada (monitoreo de calidad del aire y variables meteorológicas), tipos de algoritmos utilizados (como redes neuronales, regresiones, y modelos de ensamble), y métricas de desempeño, incluyendo el error cuadrático medio (RMSE) y el coeficiente de eficiencia. Estos pasos permitieron identificar los modelos de mayor precisión y evaluar los desafíos en la reusabilidad y adaptación de estos modelos para condiciones locales

**Útil o descartable para nuestra solución propuesta** El estudio puede ser valioso en la sección de metodología, dado que se describe las aplicaciones de machine learning para identificar relaciones entre variables meteorológicas y contaminantes. Sin embargo, no comparte características entre el Valle de Aburra y el Oriente antioqueños se tendría que ajustar para las dinámicas de este último. Nos tendríamos que ajustas las técnicas descritas en preprocesamiento y análisis para que se ajusten en las particularidades topográficas y meteorológicas del oriente. ]

1. Diferenciación del Proyecto

[El objetivo de nuestro proyecto busca correlacionar las variables meteorológicas con las concentraciones de contaminantes atmosféricos en el Oriente antioqueño, representa un esfuerzo innovador y necesario para la región.

El Oriente antioqueño, una región de creciente desarrollo urbano e industrial, carece actualmente de estudios detallados que aborden la relación entre los factores meteorológicos y la contaminación atmosférica. A diferencia de otras áreas en Colombia y el mundo, donde se han llevado a cabo estudios extensivos de calidad del aire, esta zona ha quedado rezagada en términos de análisis científico. Esto significa que no se tiene un conocimiento profundo sobre cómo las condiciones meteorológicas, como la temperatura, la humedad y la velocidad del viento, influyen en la dispersión y concentración de contaminantes.

Este proyecto aborda ese vacío, ofreciendo no solo una evaluación exhaustiva de la calidad del aire, sino también una comprensión detallada de los factores que influyen en las fluctuaciones de contaminantes como el NOx, el O3 y las partículas PM10, entre otros. Al ofrecer datos y análisis basados en mediciones horarias de estaciones de calidad del aire, el proyecto genera información relevante que será clave para futuras políticas públicas y estrategias de control ambiental en la región.

Este estudio no solo llena un vacío en la investigación local, sino que también crea oportunidades para avanzar en la comprensión de las interacciones entre las variables meteorológicas y los contaminantes atmosféricos en zonas de alta altitud y características geográficas particulares, como el Oriente antioqueño. Los resultados de este proyecto pueden servir de base para comparaciones con otras regiones y para futuras investigaciones que busquen identificar patrones similares en contextos geográficos y climáticos diferentes.

Al aplicar metodologías avanzadas y centrarse en un área geográfica previamente desatendida, este trabajo no solo contribuirá al avance científico, sino que también tendrá un impacto directo en la gestión ambiental y la salud pública de la región.]

1. Referencias.

[Palmeira, A., Pereira, É., Ferreira, P., Diele-Viegas, L. M., & Moreira, D. M. (2022). Long-Term Correlations and Cross-Correlations in Meteorological Variables and Air Pollution in a Coastal Urban Region. *Sustainability*, *14*(21), 14470.

Statheropoulos, M., Vassiliadis, N., & Pappa, A. (1998). Principal component and canonical correlation analysis for examining air pollution and meteorological data. *Atmospheric environment*, *32*(6), 1087-1095.

Rad, A. K., Shamshiri, R. R., Naghipour, A., Razmi, S. O., Shariati, M., Golkar, F., & Balasundram, S. K. (2022). Machine learning for determining interactions between air pollutants and environmental parameters in three cities of Iran. Sustainability, 14(13), 8027.

Zapata González, E. A. (2017). Estimación de la calidad del aire en la ciudad de Talca utilizando algoritmos de aprendizaje automático (Tesis de maestría, Universidad de Talca). Facultad de Ingeniería, Universidad de Talca

AlOmar, M. K., Khaleel, F., AlSaadi, A. A., Hameed, M. M., AlSaadi, M. A., & Al-Ansari, N. (2022). *The Influence of Data Length on the Performance of Artificial Intelligence Models in Predicting Air Pollution*. Advances in Meteorology, 2022, Article ID 5346647. <https://doi.org/10.1155/2022/5346647>

Vásquez Arenas, J. P. (2022). Beneficios de los modelos basados en algoritmos de machine learning para la predicción de concentraciones de PM2.5 en el Valle de Aburrá [Trabajo de grado especialización, Universidad de Antioquia]. Universidad de Antioquia, Medellín, Colombia. ]