Pronóstico del aumento de ingresos en agrupaciones de hospitales por diagnósticos de enfermedades respiratorias basado en las condiciones metereológicas y la calidad del aire del Valle de Aburrá

Seminario de Analítica y Ciencia de Datos

Resumen Descriptivo del Proyecto

Para tratar de evitar que personas con deficiencias respiratorias puedan verse afectadas ante las condiciones de la calidad del aire, presentamos un modelo alimentado por variables climáticas e ingresos en hospitales con el fin de generar el pronóstico de zonas donde el aire se vea afectado y aumente la capacidad de hospitales por esta problemática.

1. Revisión de Literatura.

La calidad del aire ha sido un tema de abundante estudio en el Valle del Aburrá, dado que al estar rodeado por montañas genera una estrechez que impide el flujo libre de gases y partículas en el aire (Área Metropolitana Valle del Aburrá, 2019). Es por ello que desde finales de los años 90 se creó el Sistema de Alertas Tempranas SIATA como un proyecto para la gestión ambiental y de riesgos del Área Metropolitana del Valle de Aburrá, realizando monitoreos en tiempo real con modelamientos hidrológicos y meteorológicos para pronosticar ocurrencias de fenómenos naturales que puedan generar riesgos a la población (Universidad de EAFIT, 2020).

A medida que avanzaba el tiempo de creación e implementación de la herramienta se fueron innovando métodos y tecnologías para realizar el seguimiento de las diferentes variables fenomenológicas. De esta forma, observamos que desde 2019 se realizan implementaciones de modelos de aprendizaje autónomo como el uso de Redes Neuronales Artificiales (RNA) para pronosticar la concentración diaria del material particulado menor a 2.5 micras (PM2.5) en el Valle de Aburrá (Colombia). Bajo este estudio establecieron un intervalo de pronóstico de un día de anticipación, a partir de información de tres estaciones de la Red de Monitoreo de Calidad del Aire del Área Metropolitana, concluyendo que la precisión de la predicción del número de días con eventos de contaminación crítica es adecuada con un coeficiente de regresión del 86.2%; observando que los modelos de RNA funcionan mejor cuando la información disponible corresponde a las distintas variables relacionada, por ende, son importantes tanto las variables meteorológicas como los registros de los contaminantes (Salazar et al., 2019, 1-2). Bajo estos hallazgos, se puede considerar al modelo de redes neuronales como posible candidato para la modelación con datos actuales, realizando la combinación de las variables independientes según lo mencione la matriz de correlación entre ellas.

Podemos decir que a pesar de que los modelos de RNA son robustos y generan buenas métricas de precisión entre valores predictores y reales, a menudo se realizan comparaciones entre modelos para abarcar diferentes formas de mejora en selección de modelos según el comportamiento de los datos, así como lo realizaron en (Murillo et al., 2019) donde realizaron una comparación entre RNA y Regresión de Soporte Vectorial mejorado con Optimización de Enjambre de Partículas (SVR-PSO) utilizando una metodología de caracterización, pronóstico y evaluación del sistema. Adicionalmente, realizaron la división de variables en 4 grupos, Variables Meteorológicas Horarias (HMV), Variables Meteorológicas Temporales (TMV), Concentraciones Horarias de Contaminantes (HPC) y Concentraciones Temporales de Contaminantes (TPC), esto con el fin de introducir información sobre el comportamiento temporal de las fuentes de emisión y actualizar la información disponibles antes de la predicción. Sin embargo, reportan que según sus resultados y el test de sesgo fraccional (FB) se subestiman los valores predichos, lo que resulta en poca capacidad predictiva cuando hay cambios abruptos. Aun así, el modelo mantiene la capacidad de representar la tendencia del comportamiento y que la mejor combinación de variables para el pronóstico resultó con HMV + TMV + TPC como entrada frente a temporadas de bajas y altas precipitaciones. Finalmente, recomiendan que en próximos casos de estudio se estudien métodos para completar la información faltante de los registros agregando una incertidumbre mínima.

Atendiendo a la problemática de los efectos negativos de una mala calidad del aire, en 2020 una revista de salud pública presentó una relación entre las partículas contaminantes y asistencias médicas por los diagnósticos de infecciones respiratorias agudas (IRA), enfermedad pulmonar obstructiva crónica (EPOC) y asma. Entre sus resultados, reportaron que, “la mayor correlación entre las tres morbilidades se presentó para el asma. La variable meteorológica de mayor correlación con la variable objetivo es la temperatura del aire para el caso de EPOC y asma. En el caso de IRA, la variable con mayor correlación es la velocidad del viento” (Parra et al., 2020). Además, encontraron correlación entre la métrica de PM2.5 y las asistencias médicas de estas enfermedades, que en episodios críticos se encontraron retardos a atención médica de 0-2 días para IRA, 0-7 para EPOC y 0-5 para ASMA, todo esto a partir de una metodología de 4 pasos que involucran la comprensión de la problemática, comprensión de los datos, preparación de los datos y modelamiento análitico. También, es apreciable el uso que le dieron a las agrupaciones de datos en divisiones de 4-9 clusters con un óptimo de 8 subdivisiones bajo el algoritmo k-means con el fin de realizar segmentaciones respecto a los días y el comportamiento de las variables meteorológicas y medio ambientales. No obstante, bajo la revisión de sus resultados mostraron que el mejor modelo que representó el 60.85% de los registros fue el modelo de redes neuronales, lo que evidencia la complejidad del ajuste del modelo a estas variables para una predicción de las mismas.

Por otro lado, consideramos que estos temas no se han interesado únicamente en las expresiones matemáticas, modelos estadísticos y archivos programables, sino que por el contrario han aumentado las investigaciones en términos del conocimiento de las herramientas, sus aplicaciones y beneficios de su uso, de esta forma aquellas persona que no tenga un conocimiento profundo del funcionamiento de las herramientas puede de igual forma presentar propuestas de mejora basado en estas y de su aplicabilidad en los campos de trabajo. Por ejemplo, lo reportado en 2023 por Juan Pablo Vasquez Arenas, el cual realizó una recopilación de posibles atributos y limitantes que pueden presentar el uso de algoritmos de machine learning en el campo de la predicción de concentraciones de contaminantes atmosféricos para la contribución de la gestión ambiental. El autor tuvo la apreciación que hasta su investigación no logró identificar un algoritmo que tuviese el mejor desempeño para este tipo de fenómeno debido a que “se suelen usar algoritmos específicos en cada caso de estudio, dificultando su reusabilidad y repetibilidad”, aún así, expresa que los modelos más comunes del campo corresponden a redes neuronales y modelos de regresión. Lo anterior da pie a la implementación de nuevos modelos que se ajusten de mejor forma a los datos que estén presentándose al momento de investigación, usando las recomendaciones del uso de las particularidades topográficas y fenómenos meteorológicos como la inversión térmica, perturbación en la dinámica de vientos, o cambios en la temperatura superficial dentro del conjunto de datos de entrenamiento (Vasquez, 2022).

Se podría pensar que estos efectos son muy dependientes de la zona geográfica, condiciones climáticas, topografía, vientos, entre muchas otras. No obstante, es pertinente detallar estos fenómenos en otros lugares del mundo, dado que la calidad del aire impacta también en la dispersión de contaminantes por corrientes aéreas (Vidal & Pérez, 2018) y que adicionalmente se estima que la calidad del aire disminuya por efectos de la rápida urbanización, la expansión de la población y el aumento de actividades industriales. Por ello, nos atrevemos a dar una vista de referencia al artículo “Supervised Machine Learning Approaches for Predicting Key Pollutants and for the Sustainable Enhancement of Urban Air Quality: A Systematic Review” donde ofrecen información sobre avances recientes usando el método PRISMA y el análisis de datos de la calidad del aire bajo beneficios y retos únicos. En el artículo se contempla que las técnicas de aprendizaje supervisado frecuentes son LSTM (Long Short-Term Memory), RF (Random Forest), ANN (Artificial Neural Networks) y SVR (Support Vector Regression) y que las bases de datos para el desarroll fueron obtenidas por Google Scholar, MDPI, IEEE Xplore, Springer, and Elsevier. Adicionalmente, se rescata de la información recopilada una tabla de diferentes referencias y modelos de predicción usados para diferentes variables de medición de la calidad del aire tales como PM2.5, PM10, Ozono, AQI, NH3, CO2, NX, SO2, entre otros (Essamlali, 2024).

Finalmente, en la búsqueda de información más reciente de modelos aplicados en el área con el Valle de Aburrá nos topamos con el artículo (Pérez et al., 2023) donde se desarrolla un marco para pronosticar concentraciones de material particulado (PM2.5) promediadas durante 24 horas con 4 días de anticipación en estaciones terrestres sobre el área metropolitana del Valle de Aburrá, Colombia. en esta ocasión utilizaron datos de diversas fuentes incluidas incluidas observaciones in situ de PM2.5 en tiempo real, pronósticos meteorológicos del Sistema de pronóstico global (GFS), pronósticos de profundidad óptica de aerosoles (AOD) del Servicio de monitoreo atmosférico europeo Copernicus (CAMS) y productos de incendios activos del Espectrorradiómetro de imágenes de resolución moderada (MODIS). Bajo estas variables, realizaron la comparación del desempeño entre modelos basados ​​en árboles, bosques aleatorios (RF) y potenciación de gradiente (GB), con regresión lineal como base para métricas de error, usando la optimización con hiper parámetros para la mejora de ajuste del modelo. Sin embargo, la estrategia GB-MO presentó el mejor desempeño, con la capacidad de predecir el episodio con la peor calidad del aire durante el período de prueba con más de un día de anticipación. sin embargo, los autores reportan que una de las desventajas de los modelos basados ​​en árboles es su incapacidad para hacer predicciones hábiles fuera del dominio, es decir, una mala estimación en la extrapolación de los predictores. Algo adicional y apreciable, es que para mitigar la desviación de los valores al ajuste usaron una función de pérdida personalizada que considera la distribución de probabilidad de los valores objetivo.

2. Modelos y Métodos ya existentes.

La temática de la calidad del aire en el Valle de Aburrá ha presentado diferentes enfoques llevando a cabo técnicas de Machine Learning, es por ello que iremos detallando diferentes aplicaciones que se han desarrollado y que son de utilidad para el proyecto que se pretende ejecutar.

Regresión de vectores de soporte mejorada con PSO: En el estudio *Forecasting concentrations of air pollutants using support vector regression improved with particle swarm optimization* (Murillo et al., 2019)menciona que la combinación de las dos técnicas como los es el SVR y el PSO, les dió como resultado un mejor RMSE siendo este una métrica que mide la diferencia entre los valores predecidos y los valores reales observados con un valor de 9.76 μg/m³ , por otro lado, en este desarrollo ejecutaron pruebas como Shapiro que dio como resultado un valor p de 0.041 indicando que los datos no siguen una distribución normal y un resultado significativo mediante la prueba de Friedman. La metodología usada frente a la manipulación de los datos meteorológicos del SIATA mencionan que no eliminaron aquellos registros que tenían valores nulos con un porcentaje promedio de 37.6% por lo tanto puede ser un factor a tener en cuenta y otro aspecto es que usaron datos correspondientes solo a un periodo de tiempo siendo este el año 2013. Sin embargo, este estudio es de utilidad para nuestro trabajo ya que el modelo mencionado presenta robustez frente a la falta de información, además, con el análisis Post Hoc podríamos tenerlo en cuenta para encontrar la selección de combinación de variables más significativas para el modelo que escojamos.

Modelos basados en árboles Gradient Tree Boosting: La investigación con el nombre *Forecasting 24 h averaged PM2.5 concentration in the Aburrá Valley using tree-based machine learning models, global forecasts, and satellite information* (Pérez et al., 2023)indican que usaron en total unas 269 variables tomadas de diferentes bases de datos como lo son del SIATA, datos de la Copernicus Atmosphere Monitoring Service (CAMS) sobre el Aerosol Optical Depth (AOD), que mide la cantidad de aerosol en la atmósfera. También se incluyeron el índice Infrared Fire Radiative Power (IFRP) y otros datos meteorológicos como la altura de la capa límite (PBL), tasa de precipitación (PRATE), y la cobertura de nubes (TCC). Es un estudio robusto que brinda un pronóstico del contaminante PM2.5 con un horizonte de predicción de cuatro días, dando como resultado dio un RMSE aproximado de 7μg/m³ en esas 96 horas, siendo un modelo capaz de anticipar los eventos de mala calidad del aire, sin embargo, después de 24 horas de predicción, el modelo mostró un sesgo positivo, subestimando las concentraciones de PM2.5 en eventos extremos y también aumentando el error en la precisión del modelo. Para nuestra propuesta es de vital importancia este estudio ya que nos da un indicio en cómo podemos manipular los datos nulos ya que hicieron un método de interpolación para rellenar esos datos.

Redes neuronales artificiales: Como tercer caso queremos recalcar el proyecto de analitica *Red neuronal artificial aplicado para el pronóstico de eventos críticos de PM2.5 en el Valle de Aburrá* (Salazar et al., 2019)que nos demuestra que el uso de redes neuronales artificiales (RNA) son eficaces para prever niveles de contaminación de la partícula PM2.5 pero enfatizan que para que el modelo tenga un mejor resultado es necesario con contar con otras variables para disminuir los errores en las predicciones, los datos meteorológicos usados son del SIATA y contaron solo con tres estaciones de monitoreo. Por otro lado el porcentaje de eficiencia del modelo fue del 70% esto nos denota la capacidad de predecir los días en que la partícula supera los umbrales de concentración críticos, para cada estación cuenta con un margen de error de predicción entre el 12 al 18,52%, encontramos utilidad en este proyecto para lo que podría ser tener en cuenta la variable de la altura de la capa de mezcla que no se tenía en el radar está puede ser un parametro clave ya que la capa de mezcla determina el aire disponible para la dispersión de todos los componentes en la atmósfera.

Otro de los estudios que tomamos como referencia es *Forecasting PM10 and PM2.5 in the Aburrá Valley (Medellín, Colombia) via EnKF based data assimilation* (López et al., 2020), tomaron el modelo LOTOS-EUROS que se usa específicamente para la simulación o predicción de contaminantes en la atmósfera, en este caso usaron la asimilación de datos por medio del Filtro de Kalman Ensamblado (EnKF) tomando la fuente de datos lo que es el SIATA, dio como resultado unas métricas como la raíz del error medio al cuadrado o comúnmente llamado RMSE en la investigación no nos indica cuál fue ese valor sin embargo indican que el valor de la métrica tuvo un buen valor en el radio de 30km frente a radio de 5 o 10 km para estimaciones de las partículas PM10 y PM2.5 y otra métrica fue el MFB (Mean Factorial BIAS) mencionando que la disminución del sesgo también fue óptimo en el radio de 30km. La utilidad que le vemos a este proyecto para nuestro estudio es la utilización del modelo LOTOS-EUROS ya que nos amplía el conocimiento sobre la interacción entre las emisiones de contaminantes, la meteorología y cómo estas afectan la calidad del aire.

Por último trabajo de referenciación, lo tomamos del CESET (Centro de Extensión Académica de la Universidad de Antioquia) que se titula *Sistema de Alerta temprana de Medellín - calidad del aire* (Gómez, 2022) que su finalidad es notificar y gestionar a tiempo los errores en los sensores de las estaciones de monitoreo del SIATA, este trabajo nos da indicio de que la entidad ha hecho procesos pertinentes para detectar estas fallas y facilitandoles la gestión y la toma de decisiones de estas. Por lo tanto, genera un valor de confiabilidad y precisión sobre la base de datos de las variables meteorológicas que usaremos en nuestro proyecto, otro aspectos a recalcar del trabajo es que el equipo de calidad del aire del SIATA establece la relevancia de monitorear las variables más críticas, entre esas están las partículas contaminantes como lo son PM2.5, PM10,y otros componentes como lo son NO,NO2,CO.

3. Diferenciación del Proyecto

Recopilando toda la información anterior en este espacio podemos resumidamente mencionar que el factor común entre ellas se basa en aprecio del modelo de Redes Neuronales Artificiales y de Regresión por mostrar buenos resultados de predicción para las manipulaciones de datos realizados, sin embargo, en la mayoría de casos usan diferentes variables para el entrenamiento de los modelos dado que el efecto de la calidad del aire se ve causado por diferentes motivos como en las combustiones, aerosoles, industrias, medio ambiente, entre otras, por ende, cada autor toma en base bajo qué causas o efectos quiere estudiar el fenómeno.

Adicionalmente, debemos tener en cuenta que dichas predicciones consideran datos históricos bajo las condiciones tomadas, es decir, que si ocurre algún tipo de situación atípica como un hecho que aumente significativamente las partículas dispersas en el aire, los modelos no tendrán este conocimiento en los tiempos alimentados. Por ello, es necesario actualizar estos modelos con información actual, así que la primera diferenciación se basará en la actualización de datos de alimentación desde las variables meteorológicas, mediciones de la calidad del aire y los registros de ingresos de pacientes con diagnósticos asociados a enfermedades respiratorias con histórico de datos del año 2020 al 2024, no obstante, también trataremos de usar algunas técnicas para la estimación de valores nulos o vacíos dado que se ha observado que las fuentes de datos presentan con frecuencia y que se reporta en la revisión bibliográfica.

También, tomando en cuenta que se han reportado complicaciones en los pronósticos generando precisiones moderadas, esperamos usar escalamiento, series de tiempo y/o clusterización de los datos junto a comparaciones entre modelos y combinaciones en ensambles para tratar de mejorar el comportamiento de predicción de la variable, de tal forma que esto permita tener un mayor rango de tiempo de pronóstico para alertar con más tiempo de anticipación a las entidades que prestan el servicio de salud, dándoles una herramienta que ayude a la gestión oportuna de ingresos de pacientes por enfermedades respiratorias.

4. Referencias.

Área Metropolitana Valle del Aburrá. (2019). *Calidad del aire*. Área Metropolitana. Retrieved September 29, 2024, from https://www.metropol.gov.co/ambiental/calidad-del-aire

Essamlali, I. (2024, Enero). *Supervised Machine Learning Approaches for Predicting Key Pollutants and for the Sustainable Enhancement of Urban Air Quality: A Systematic Review*. MDPI. Retrieved October 3, 2024, from https://www.mdpi.com/2071-1050/16/3/976

Gómez, J. A. (2022). *Sistema de Alerta temprana de Medellín - calidad del aire.* Sistema de bibliotecas universidad de Antioquia. Retrieved 10 07, 2024, from https://bibliotecadigital.udea.edu.co/handle/10495/33716

López, S., Yarce, A., Pinel, N., & Heemink, A. W. (2020, Julio 01). *Forecasting PM10 and PM2.5 in the Aburrá Valley (Medellín, Colombia) via EnKF based data assimilation*. Sciencedirect. Retrieved 10 07, 2024, from https://www.sciencedirect.com/science/article/abs/pii/S1352231020302442

Murillo, J., Sepulveda, J. P., Correa, M. A., & Orrego, D. (2019, September). *Forecasting concentrations of air pollutants using support vector regression improved with particle swarm optimization: Case study in Aburrá Valley, Colombia*. Science Direct. Retrieved Octubre 02, 2024, from https://www.sciencedirect.com/science/article/abs/pii/S2212095518302438

Parra, J., Oviedo, A., & Amaya, F. (2020, Octubre 25). *Analítica de datos: incidencia de la contaminación ambiental en la salud pública en Medellín (Colombia)*. SciELO Colombia. Retrieved October 2, 2024, from http://www.scielo.org.co/scielo.php?script=sci\_arttext&pid=S0124-00642020000600609

Pérez, J., Montoya, P., Sanchez, J., & Hernández, S. (2023, Diciembre). *Forecasting 24 h averaged PM2.5 concentration in the Aburrá Valley using tree-based machine learning models, global forecasts, and satellite information*. Advances in Statistical Climatology, Meteorology and Oceanography. Retrieved Octubre 03, 2024, from https://ascmo.copernicus.org/articles/9/121/2023/ascmo-9-121-2023.html

Salazar, D., Jiménez, J., Zapata, C., & Cardona, Á. (2019, Junio). Red neuronal artificial aplicado para el pronóstico de eventos críticos de PM2.5 en el Valle de Aburrá. *Scielo*, *86*, 1-2. https://doi.org/10.15446/dyna.v86n209.63228

Universidad de EAFIT. (2020). *Sistema de Alertas Tempranas SIATA - Innovación EAFIT / Proyectos*. Universidad EAFIT. Retrieved September 29, 2024, from https://www.eafit.edu.co/innovacion/consultoria-proyectos/Paginas/siata.aspx

Vasquez, J. (2022). *Beneficios de los modelos basados en algoritmos de machine learning para la predicción de concentraciones de PM2.5 en el Valle de Aburrá*. Repositorio institucional Universidad de Antioquia. Retrieved Octubre 03, 2022, from chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://bibliotecadigital.udea.edu.co/bitstream/10495/29252/1/VasquezJuan\_2022\_BeneficiosMachinelearningCalidaddelaire.pdf

Vidal, O., & Pérez, A. (2018, Abril). *Estimación de la Dispersión de Contaminantes Atmosféricos Emitidos por una Industria Papelera Mediante el Modelo AERMOD*. SciELO Colombia. Retrieved October 3, 2024, from http://www.scielo.org.co/scielo.php?script=sci\_arttext&pid=S0121-750X2018000100031