Modelo de predicción para identificar la gravedad de una enfermedad de respiración aguda (ERA) para las personas de Bogotá relacionados con el agente contaminante PM 2.5 y otros factores ambientales

Marco Antonio Méndez Espitia

Asesor

Jhoana Patricia Romero Leiton

Universidad Nacional Abierta y a Distancia UNAD

Escuela de Ciencias Básicas, Tecnología e Ingeniería ECBTI

Especialización en Ciencia de Datos y Analítica

Semillero de Investigación Mathphysics

2025

Dedicatoria

A Dios, quien ha sido mi guía constante a lo largo de este camino, agradezco profundamente por los buenos y malos momentos que me han forjado durante el desarrollo de esta especialización. En cada desafío y triunfo, sentí su presencia iluminando mi senda.

A mi amada esposa, gracias por tu apoyo incondicional, por ser mi refugio y sostén moral cuando decidí embarcarme en esta especialización. Tu comprensión y ánimo han sido fundamentales para alcanzar este logro.

A mi madre, por estar siempre presente, aunque a la distancia, brindando su amor y respaldo. Tu apoyo ha sido un pilar que me ha dado fuerzas en los momentos más difíciles.

A mi querida hija, gracias por sacrificar tiempo conmigo, entendiendo que este esfuerzo es para asegurar un futuro mejor para ti. Todo esto ha sido pensado en lo mejor para nuestro porvenir.

A mis jefes, por su comprensión y apoyo durante los periodos laborales más pesados, permitiéndome continuar con mis estudios sin interrupciones. Su flexibilidad y respaldo han sido esenciales para alcanzar este objetivo.

Finalmente, a mi padre, que desde el cielo sigue cuidándome y guiándome. Se´ que, aunque no estás físicamente conmigo, tu espíritu me acompaña y me impulsa a ser mejor cada día.

A todos ustedes, gracias por acampanarme en este viaje y por el amor y apoyo que han sido fundamentales para alcanzar este logro.

Agradecimientos

A todos los docentes que, con su dedicación y sabiduría, contribuyeron al desarrollo de esta especialización, les extiendo mi más profundo agradecimiento. Cada clase, cada lección y cada palabra de aliento ha dejado una marca indeleble en mi formación.

En especial, a la profesora Jhoana, gracias por sus valiosos consejos y observaciones, que no solo enriquecieron mi aprendizaje, sino que también me guiaron hacia la excelencia. Su apoyo constante fue una luz que iluminó mi camino en los momentos de duda.

Quiero extender mi gratitud a la Secretaría de Salud de Bogotá´, cuyo respaldo fue fundamental para la realización de este proyecto. La información suministrada de los RIPS fue el pilar sobre el cual se edificó´ esta investigación, y sin su colaboración, este logro no habría sido posible.

A todos ustedes, mi más profundo agradecimiento por su apoyo y por ser parte esencial de este éxito.

Resumen

La contaminación es un problema mundial que afecta tanto a naciones industrializadas como en desarrollo, y Colombia no es la excepción. En Bogotá, la calidad del aire representa un reto importante, afectando de manera especial a la infancia y la tercera edad.

Mediante la aplicación de técnicas de aprendizaje automático, se llevó a cabo la exploración de datos provenientes de cuatro fuentes distintas. Se consideró información de la Secretaría Distrital de Salud, así como datos públicos, incluyendo la Red de Monitoreo de Calidad del Aire (RMCAB), registros abiertos sobre la humedad en Bogotá y reportes de incendios proporcionados por el Cuerpo de Bomberos de la ciudad. Tras la unificación de las bases de datos y siguiendo la metodología CRISP-DM, se llevó a cabo el modelado de datos para desarrollar dos modelos de clasificación binaria capaces de evaluar la gravedad de la enfermedad.

Este proceso se fundamenta en la exploración y preprocesamiento de los datos, precedido por un procedimiento de Extracción, Transformación y Carga (ETL, por sus siglas en inglés).

Posteriormente, los modelos serán evaluados segmentando la información en conjuntos de entrenamiento, validación y prueba, lo que permitirá ajustar hiperparámetros clave del algoritmo LightGBM, tales como learning rate, num leaves y max depth.

La selección de los valores óptimo se llevará a cabo considerando métricas de rendimiento, como la sensibilidad o la precisión, según los requerimientos del modelo. El objetivo final es generar una predicción a ocho días en el futuro, basada en el período promedio de incubación de una Enfermedad Respiratoria Aguda.

Con la selección de los hiperparámetros óptimos para los modelos, se logró´ desarrollar un primer modelo que predice la gravedad de la enfermedad, clasificándola entre leve y grave,

con un rendimiento del 86 % en la métrica de sensibilidad. Este resultado es notable, ya que se buscó maximizar la detección de casos graves. Por otro lado, el segundo modelo, que distingue entre enfermedades leves y medias, alcanzó una precisión del 80 %, lo que refleja un buen rendimiento al aprovechar el equilibrio de clases. Al ensamblar estos modelos en un único algoritmo, será posible clasificar nuevos registros en alguna de las tres categorías.

Palabras claves: cobertura de predicción, contaminación atmosférica, ensamble de modelos, modelos predictivos, salud pública.

Abstract

Pollution is a global issue that affects both industrialized and developing nations, and Colombia is no exception. In Bogotá, air quality poses a significant challenge, especially impacting children and the elderly.

Using machine learning techniques, data from four different sources was explored. Information from the District Health Secretariat was considered, along with public data, including the Air Quality Monitoring Network (RMCAB), open records on humidity in Bogota´, and fire reports provided by the city's Fire Department. After unifying the databases and following the CRISP-DM methodology, data modeling was carried out to develop two binary classification models capable of assessing the severity of the disease. This process is based on data exploration, preceded by an ETL (Extract, Transform, Load) process. The models were evaluated by segmenting the data into training, testing, and validation sets, allowing for the adjustment of key LightGBM hyperparameters, such as learning rate, num leaves, and max depth. The optimal values were selected based on metrics such as sensitivity or accuracy, depending on the model's needs. The ultimate goal is to make an 8-day future prediction, based on the average incubation period of an Acute Respiratory Disease.

By selecting the optimal hyperparameters for the models, a first model was developed to predict disease severity, classifying it as mild or severe, achieving an 86 % sensitivity score. This result is significant, as the goal was to maximize the detection of severe cases. On the other hand, the second model, which differentiates between mild and moderate diseases, achieved 80 % accuracy, reflecting good performance by leveraging class balance. By assembling these models into a single algorithm, it will be possible to classify new records into one of the three categories.

Keywords: Air pollution, Model ensemble, Predictive models, Prediction coverage,

Public health.

Tabla de Contenido

Lista de Figuras

Lista de Apéndices

Introducción

El presente proyecto tiene como objetivo general predecir la gravedad de los casos positivos de Enfermedad Respiratoria Aguda (ERA) en la ciudad de Bogotá, Colombia, utilizando datos abiertos y privados recopilados entre 2019 y 2023, mediante la aplicación de un modelo de aprendizaje automático. Este enfoque se fundamenta en la necesidad de abordar una problemática de salud pública significativa, dado que las ERA representan una amenaza considerable para las poblaciones más vulnerables, especialmente los menores de 5 años y los adultos mayores de 60 años.

El marco referencial del proyecto se basa en diversas investigaciones previas que han aplicado algoritmos de aprendizaje automático para la predicción de enfermedades respiratorias. Se han revisado estudios que utilizan datos climáticos y variables ambientales, como temperatura, humedad y precipitaciones, con el objetivo de aumentar la precisión de las predicciones que se realicen. Este contexto teórico proporciona una base sólida para el desarrollo del modelo propuesto.

La metodología empleada en este proyecto sigue el enfoque CRISP-DM, que incluye etapas de entendimiento del negocio, entendimiento de los datos, preparación de los datos, modelado, evaluación y despliegue. A través de un análisis exploratorio de datos (EDA), se identificaron patrones y se prepararon los datos para el modelado, utilizando diversas técnicas de aprendizaje automático para optimizar los resultados.

Los resultados obtenidos indican que el modelo final logró un rendimiento del 86 % en la métrica de recall para la identificación de casos graves, mientras que el segundo modelo, que distingue entre enfermedades de severidad media y leve, alcanzó una precisión del 80 %. Estos resultados reflejan la efectividad del enfoque adoptado y la importancia de implementar

estrategias preventivas y de manejo que disminuyan las tasas de hospitalización y mortalidad.

En conclusión, este proyecto no solo busca generar conocimiento sobre las ERA, sino también ofrecer una solución practica para la predicción y clasificación de la gravedad de estas enfermedades, contribuyendo así a mejorar la atención médica y optimizar el uso de recursos en el sistema de salud.

Descripción del Problema

A lo largo de 2024, la ciudad capital ha experimentado varios días sin lluvias y una incidencia inusual de incendios forestales. Según la Red de Monitoreo de Calidad del Aire de Bogotá (RMCAB), las partículas contaminantes como el PM10 destacan por exceder con frecuencia los límites establecidos por las normas de calidad del aire. Localidades como Puente Aranda, Kennedy y Fontibón presentan niveles especialmente elevados de contaminación (N. Y. Rojas, 2024), lo que incrementa el riesgo de enfermedades respiratorias agudas (ERA). La exposición continua a estos agentes contaminantes debilita los órganos respiratorios, haciéndolos más vulnerables a infecciones al reducir su capacidad funcional (Ministerio de Salud, 2017). Este proceso es exacerbado por la presencia de sustancias no biológicas, que generan reacciones de óxido-reducción, causando daños en las membranas celulares de las vías respiratorias superiores e inferiores (Olaya, 2024).

El incumplimiento de las normativas ambientales ha dificultado el control de la contaminación. La calidad del aire es esencial para la salud pública, ya que niveles bajos de contaminación contribuyen a mejorar la salud cardiovascular y respiratoria tanto a corto como a largo plazo, reduciendo la carga de morbilidad asociada a enfermedades derivadas de la contaminación (Organización Mundial de la Salud, 2016). Las ERA son un grupo de patologías relacionadas con la contaminación ambiental, principalmente causadas por agentes como el PM2,5 y PM10 (Onatra, Vargas, Páez, Rojas, y López, 2009).

Justificación

La contaminación representa un problema significativo para la salud pública, siendo un factor clave en el desarrollo de diversas enfermedades, incluidas las enfermedades respiratorias agudas (ERA). Los efectos perjudiciales del material particulado PM2.5 en los pulmones están bien documentados (Lian, Xu, H, y Xing, 2016). Este agente penetra profundamente en los pulmones, irrita y daña las paredes alveolares, afectando la función pulmonar y aumentando el riesgo de cáncer de pulmón y muerte prematura. Además, la exposición a PM2.5 genera un impacto económico significativo debido a los costos asociados con la atención médica.

A pesar de la evidencia existente sobre los efectos del PM2.5, hasta donde sabemos la investigación actual carece de modelos predictivos basados en la ubicación que puedan anticipar la gravedad de las ERA según el lugar de residencia de las personas. Los modelos existentes se centran en determinar si una persona es propensa a desarrollar este tipo de enfermedades, sin considerar la variabilidad geográfica y temporal en la exposición a contaminantes. La capacidad de predecir la gravedad de los casos de ERA basándose en las concentraciones de PM2.5 por ubicación y día, junto con otras variables ambientales como la humedad y las precipitaciones, podría permitir la implementación de medidas preventivas más efectivas, reduciendo los costos de atención médica y mejorando la salud y calidad de vida de la población.

Para abordar esta necesidad, se propone el uso de datos recopilados de la Red de Monitoreo de la Calidad del Aire de Bogotá (RMCAB), que incluye información sobre humedad, precipitaciones y otros agentes contaminantes, junto con registros de consultas médicas del sistema RIPS, correspondientes al periodo de 2019 a 2023. Aplicando técnicas de análisis y aprendizaje automático, se diseñará un modelo capaz de predecir la gravedad de las ERA. Los resultados de esta investigación podrían ser presentados a las autoridades responsables de la

calidad del aire en Bogotá o a la Secretaría de Salud, con el objetivo de desarrollar planes preventivos o focalizados para mejorar la calidad de vida de los ciudadanos afectados por la contaminación del aire.

Objetivos

Objetivo General

Desarrollar un modelo de predicción para identificar la gravedad de la Enfermedad Respiratoria Aguda (ERA) en personas de Bogotá, basado en el análisis de datos ambientales y contaminantes atmosféricos, utilizando técnicas de aprendizaje automático.

Objetivos Específicos

Analizar la relación entre la contaminación del aire y la incidencia de la Enfermedad Respiratoria Aguda en Bogotá.

Implementar un proceso de Extracción, Transformación y Carga (ETL) para integrar datos de diversas fuentes, incluyendo la Red de Monitoreo de Calidad del Aire y registros de salud.

Desarrollar modelos de clasificación basados en algoritmos de aprendizaje automático para predecir la gravedad de la ERA.

Evaluar el rendimiento de los modelos utilizando métricas como precisión, sensibilidad y especificidad.

Proponer estrategias basadas en los resultados del modelo para la mitigación del impacto de la contaminación del aire en la salud pública.

Marco Referencia

Estado del Arte

El presente estudio se enmarca en la necesidad de predecir la gravedad de las enfermedades respiratorias agudas (ERA) en Bogotá mediante algoritmos de aprendizaje automático. Se han revisado diversas investigaciones previas sobre la aplicación de estos algoritmos en la predicción de enfermedades respiratorias y otras condiciones de salud. Por ejemplo, en el trabajo propuesto en Aplicación de analítica de datos para predicción de infección respiratoria aguda en Colombia (Jiménez Manjarrés, 2019) utilizó datos climáticos como temperatura, humedad y precipitación, así como información de palabras clave relacionadas con ERA, experimentando con modelos como regresión lineal múltiple, LASSO y redes neuronales. Los resultados, con errores entre 6,44 % y 19,54 % según el MAPE, fueron prometedores, aunque el estudio tuvo un enfoque regional. En contraste, el proyecto actual se centra exclusivamente en Bogotá.

Por su parte, en el proyecto Predicción de la epidemia del virus sincitial respiratorio en Bogotá, D.C., utilizando variables climatológicas (González-Parra, Querales, y Aranda, 2016) se enfocó en predecir las semanas de brotes de infección por virus sincitial, una causa importante de ERA en niños y adultos mayores. Utilizó datos de una entidad estadounidense y encontró que variables como humedad mínima, velocidad del viento y temperatura mínima fueron cruciales para mejorar las predicciones aplicando clasificadores bayesianos. En este estudio, la variable de humedad a 2 metros, extraída de datos abiertos de Colombia, se considera esencial.

Otro enfoque relevante es el análisis de series de tiempo, como el de (Martha´ Eraso, 2014), que aplicó modelos como regresión lineal dinámica, ARIMA y regresión de Poisson dinámica.

Identificó patrones en ciertas zonas de Bogotá, destacando la ausencia de tendencias estacionales en los casos de ERA. Este enfoque mejora la gestión de la variabilidad de los casos, facilitando la identificación de casos positivos.

Finalmente, en el proyecto Minería de datos para el descubrimiento de patrones en enfermedades respiratorias en Bogotá, Colombia (Trabajo de grado) (E. Rojas y Aguilar, 2017) utilizó algoritmos de clustering, destacando la importancia de la edad en la clasificación de casos. En su estudio, los clústeres formados giraban en torno a la edad, mostrando cómo ciertas enfermedades están asociadas a edades y géneros específicos.

Comparado con estos estudios, este proyecto se enfoca en el contexto único de Bogotá, con sus características ambientales y demográficas particulares, abarcando un periodo amplio (2019-2023) y utilizando datos públicos. El objetivo es desarrollar un modelo que no solo prediga casos positivos de ERA, sino también su gravedad, permitiendo a hospitales y autoridades de salud optimizar la gestión de recursos y mejorar la planificación médica.

Marco Contextual

Las Enfermedades de Respiración Aguda (EDA), desafortunadamente es una de las causas de muerte en los niños menores a 5 años (Aristizabal, Hernández, y Medina, 2009), así mismo, por cada 100.000 menores puede morir 6.6 niños por ejemplo en Usme (Murcia Urrego, 2019), entonces este problema de salud pública es importante analizar y crear todas las herramientas posibles para su correcto manejo.

Igualmente, la Secretaría Distrital de Salud, recomienda énfasis al cuidado de los menores de 5 años pero también a las personas mayores de 60 años (Secretaría de Salud de Bogotá, 2025), ya que son poblaciones de mayor riesgo en morbilidad y mortalidad, ya que de las consultar puede llegar a ser el 32 % de la población en infantes de este rango de edad. Lo más

preocupante es que debido a este riesgo, en el 2024 el 55.4 % de los casos de hospitalización fue relacionados por ERA en estos dos grupos poblacionales (Marulanda, 2024) resaltando el autocuidado.

En conclusión, las Enfermedades de Respiración Aguda (ERA) son una amenaza significativa para las poblaciones más vulnerables, lo que exige una atención prioritaria en las políticas de salud pública. La protección de los menores de 5 años y los adultos mayores de 60 años debe ser una prioridad, implementando estrategias preventivas y de manejo que disminuyan las tasas de hospitalización y mortalidad. Fomentar el autocuidado y la prevención son medidas esenciales para enfrentar este desafío, asegurando una mayor calidad de vida y aliviando la presión sobre el sistema de salud.

Marco Conceptual

Para el desarrollo del proyecto es fundamental abordar la teoría de los conceptos relacionados con la contaminación atmosférica, ya que esta es uno de los principales causantes de las enfermedades respiratorias agudas (ERA). En segundo lugar, se describirán las variables meteorológicas y contaminantes que se utilizarán en el modelamiento, como el material particulado 2.5, la temperatura, la humedad y los incendios forestales, entre otros.

Posteriormente, se contextualizarán las enfermedades respiratorias agudas, que son el objetivo principal del proyecto, con el objetivo de estimar la gravedad de la enfermedad en caso de que una persona dé positivo. Finalmente, se desarrollarán los conceptos relacionados con el aprendizaje automático enfocado en modelos de clasificación, dada la naturaleza del problema.

Contaminación Atmosférica

La presencia de sustancias que puedan causar molestias o representar un riesgo para la vida en el planeta se puede catalogar como contaminación atmosférica. Aunque estas sustancias

pueden tener un origen natural, como resultado de eventos naturales, la actividad humana ha dejado una huella significativa en la contaminación actual. Las emisiones dañinas provenientes de procesos industriales y la quema de combustibles fósiles son las principales fuentes de esta contaminación (Martínez Ataz y Díaz de Mera Morales, 2004).

Se debe tener en cuenta que la concentración y los diferentes contaminantes en la atmósfera es variable, ya que dependerá de la ubicación y su distribución de aquellas fuentes contaminantes, así mismo de la ubicación y características climáticas y topográficas de la ciudad o pueblo (Mora-Barrantes, Sibaja-Brenes, y Borbón-Alpizar, 2021). Así mismo, es importante que, según la Organización Mundial de la Salud, el 23 % de las muertes se relacionan con el ambiente, un porcentaje relevante, ya que estas muertes se pueden vincular a un exposoma, donde la exposición a diferentes contaminantes desde la preconcepción de una persona tendrá consecuencias a largo plazo (Emjen, 2020).

Dentro la clasificación de los contaminantes se puede dar por:

Su forma física se clasifica en gases y aerosoles (Romero Placeres, Olite, y A´ lvarez Toste, 2006).

Por su origen, dónde se clasifica en:

Primarios: Relacionado con las partículas sólidas y líquidas suspendidas en el aire, gases o vapores (Romero Placeres y cols., 2006).

Secundarios: Relacionado con el ácido sulfúrico y sulfatos, ozono, otros contaminantes fotoquímicos (Romero Placeres y cols., 2006).

Para el proyecto, se desarrollará bajo la clasificación de la clasificación por su origen, ya que los atributos del modelo serán con base a medidas de materiales particulados como el PM2.5 y PM10, como concentraciones de diferentes contaminantes secundarios. Con base a lo anterior:

Primario – Partículas (Material Particulado): Es un conjunto de partículas sólidas y líquidas que son emitidas al aire, como es el hollín de diésel, polvo de vías o de procesos productivos (Arciniénagas Suárez, 2012).

Igualmente, al ser una mezcla con diferentes orígenes, tamaños, forma y composición química tendrá un impacto sobre la salud humana, pues desde la Organización Mundial de la Salud, ya estableció una conexión entre la exposición a estas partículas y los niveles de mortalidad y morbilidad (Egas, Naulin, y Préndez, 2018). Este tipo de contaminante, se clasifican en tres categorías:

Partículas gruesas PM10: Tienen un diámetro aerodinámico entre 2.5 y 10 µm (Olaya-Ochoa, Ovalle, y Urbano, 2017).

Partículas finas PM2.5: Tienen un diámetro aerodinámico menor a 10 µm (Olaya-Ochoa y cols., 2017).

Partículas ultrafinas PM1: Tienen un diámetro aerodinámico menor a 1 µm (Olaya-Ochoa y cols., 2017).

Gases de efecto invernadero: El efecto invernadero es un fenómeno que realiza la captura la energía solar reteniéndola en la atmósfera. Los gases que causan este efecto se clasifican en dos:

Naturales: Son aquellos gases emitidos por las acciones propias de la

naturaleza, como es la evaporación para mantener un equilibrio en la temperatura de la tierra Palacios (2018).

Antropogénico: Son los gases emitidos por las acciones del ser humano, como es la quema de agentes fósiles en la industria o la ganadería masiva Palacios (2018).

De estos gases de efecto invernadero, existen 6 de mayor importancia por sus niveles de

contaminación, de acuerdo con el protocolo de Kioto (Actual y perspectivas, S., 2024), estableciendo un control de los límites:

Dióxido de Carbono CO2: Es un gas proveniente de la extracción y quema de combustibles fósiles como principal fuente (Dióxido de carbono, 2024).

Metano CH4: Como componente principal del gas natural, el metano es un absorbente de calor mucho más potente que el dióxido de carbono, siendo responsable del 25 % del calentamiento global UNEP (2022).

Óxidos nitrosos NOx: Estos óxidos son los responsables de formar ozono fotoquímico o smog, el cual puede causar lluvia acida (Comisión Europea, 2024).

Hidrofluocarbonados HFC: Son gases sintéticos que se utilizan para enfriar y refrigerar, pero es uno de los contaminantes más potente, aunque solamente represente el 2 % de los gases invernadero (Coalición Clima y Aire Limpio (CCAC), 2024).

Perfluorcarburo PFC: Son compuestos artificiales que se utilizan en procesos de fabricación industrial (Parlamento Europeo, 2024).

Hexafluoruro de azufre SF6: Es un gas sintético, incoloro y sin olor se utiliza como aislante eléctrico por su alta eficiencia, pero es 23.500 más potente que el dióxido de carbono (BBC News Mundo, 2019).

Variables Meteorológicas y Contaminantes

Para tener un contexto completo de las variables que se utilizarán, se definen a continuación:

Incendio Forestal: Es un fuego iniciado que se propaga sin control, quemando toda vegetación rural o urbana, afectando la fauna y personas que estén cerca (CNE, 2024).

Humedad: Se hace relación al vapor de agua que hay en el aire, que dependerá de las

formaciones de nubes como de las precipitaciones, ya que la humedad dependerá de la cantidad de agua existente en el momento (Humedad, 2014).

Precipitación: Es cualquier tipo de precipitación que se origina en la atmósfera y alcanza la superficie terrestre. En esta definición incluye la lluvia o llovizna, granizo o aguanieve (de Cambio Climático de las Américas, 2024).

Óxido nítrico NO: Contaminante que ha contribuido a la destrucción de la capa de ozono, a pesar de que el 80 % de su emisión proviene de fuentes naturales, su producción ha aumentado a causa de la quema de combustibles fósiles o la fertilización intensiva de suelos agrícolas (Acosta González y cols., 2022).

Dióxido de Nitrógeno NO2: Las fuentes principales de emisión son los vehículos con combustión interna, teniendo en cuenta que este componente es uno de los causantes de las sibilancias y el uso de medicamentos para los niños que sufren asma, así mismo aumenta que una persona pueda sufrir alergias al polen en las zonas contaminadas, siendo un componente que compromete la salud de la población (González-Díaz, Lira-Quezada, Villarreal-González, y Canseco-Villarreal, 2022).

Óxidos de Nitrógeno NOX: Estos componentes contribuyen a la formación de lluvia acida afectando la calidad del agua. Igualmente, estos componentes son emitidos por los vehículos de combustión interna como principal fuente (Darquea, 2018).

Dióxido de Azufre SO2: Es un gas inodoro cuando se encuentra en bajas concentraciones. Su fuente principal de contaminación es por la combustión de carbón, siendo responsable del smog que se presenta en el aire contaminado. Su impacto es la salud es importante, ya que puede se relaciona con el ácido sulfúrico que puede causar daños permanentes en los pulmones (Apaza Cabrera, 2018).

Monóxido de Carbono CO: Es un gas altamente tóxico porque envenena la sangre, ya que evita que se transporte oxigeno de manera correcta. Normalmente se origina en la combustión incompleta de la gasolina en los vehículos de combustión interna (Apaza Cabrera, 2018).

Dióxido de Carbono CO2: A pesar de desempeñar un papel importante en el ciclo del carbono debido a su efecto invernadero natural, el aumento de este gas ha desequilibrado ese efecto, causando que la captura de calor incremente de manera no natural (Apaza Cabrera, 2018).

Ozono O3: Es un gas que se forma por la interacción natural entre compuestos orgánicos y óxidos de nitrógeno NOx, pero también por fuente que generen grandes cantidades de energía como rayos (Apaza Cabrera, 2018).

Velocidad del viento: Es la relación entre la distancia recorrida por el aire respecto al tiempo en recorrerla (DANE, 2024).

Dirección del viento: Hace referencia de donde proviene el viento, esta dirección se calcula con un instrumento llamado veleta (del viento, 2023).

Temperatura: Es la manifestación de energía cinética presente en el aire en un punto y momento específico. Existe varias escalas como Celsius, Fahrenheit y Kelvin (es la temperatura y cómo se mide? — Clima.com, 2023).

Radiación Solar: Es la energía liberada por el Sol, que se transmite a través del espacio mediante ondas electromagnéticas (IDEAM, 2024).

Presión Barométrica: Es la fuerza por unidad que superficie que ejerce la atmósfera en un punto especifico (atmosférica, 2022).

Enfermedades de Respiración Aguda (ERA)

Las Enfermedades Respiratorias Agudas (ERA) constituyen un grupo de patologías que afectan el aparato respiratorio. Son causadas por diversos microorganismos, como virus y

bacterias, y se caracterizan por su aparición repentina y su duración generalmente inferior a dos semanas (de Salud, 2017). Este grupo de enfermedades representa una significativa causa de morbilidad y mortalidad en todas las edades, constituyendo un serio problema de salud pública en Colombia. Las ERA pueden clasificarse de dos maneras: según su etiología (origen de la enfermedad) o según su localización (Macedo y Mateos, 2006).

Según su etiología, las enfermedades pueden tener origen bacteriano, viral, parasitario, o ser especificas e inespecíficas. Según su localización, se clasifican en altas y bajas. La clasificación de las ERA a partir de su etiología se basa en la ocurrencia de la afección en más de un sitio del cuerpo (de la Salud, 1992). Esta clasificación cuenta con 10 grupos. En la clasificación según la localización, el sistema respiratorio se divide en dos secciones: altas y bajas. También es importante considerar las estaciones del año, ya que los factores ambientales y meteorológicos influyen significativamente en la propagación de las bacterias y virus responsables.

Modelos de Aprendizaje Automático de Clasificación

Para hablar de modelos de aprendizaje automático de clasificación, se debe hablar del concepto mismo de aprendizaje automático. El machine learning, o aprendizaje automático, va más allá de ser una simple tendencia; es una herramienta clave que emplean las grandes compañias, desde las tecnológicas hasta las bancarias, e incluso los propios gobiernos, ven como una gran oportunidad para diversas aplicaciones (Raschka y Mirjalili, 2019) No se trata únicamente de una definición teórica, sino de la implementación de ciertas técnicas y métodos que, con el enorme volumen de datos generados, desde las compras en un supermercado hasta las transacciones diarias de una entidad financiera, es posible transformar esos datos en información valiosa y, mediante los algoritmos adecuados, convertir esa información en conocimiento

(Raschka y Mirjalili, 2019).

Dada la relación entre la IA y el aprendizaje automático, desde un enfoque técnico, se define como un algoritmo que aprende a partir de una experiencia, en función de una tarea específica, evaluado mediante métricas y que mejora conforme dicha experiencia evoluciona (Díaz, 2021). Sin embargo, este algoritmo con modelos matemáticos se distingue de la programación clásica, ya que, en esta, la combinación de datos y reglas produce una salida, mientras que en el aprendizaje automático, los datos combinados con las salidas generan nuevas reglas. Al entender esta diferencia entre ambos enfoques, el aprendizaje automático, como una rama de la IA, aprende de un conjunto de datos mediante un proceso de entrenamiento para identificar patrones utilizando modelos matemáticos (Pineda, 2022).

Dentro de la clasificación, existen dos ramas principales, que se relacionan a continuación:

Modelos Supervisados: Dentro de los modelos que se clasifican dentro del aprendizaje automático (AA) supervisado, es importante que el conjunto de datos cuente con la etiqueta (variable a predecir) (Pineda, 2022), también conocida como variable objetivo. Según el tipo de variable objetivo, se pueden identificar dos tipos de modelos supervisados: los modelos de regresión, que predicen un valor continuo, y los modelos de clasificación, que predicen dos o más categorías (Pineda, 2022). Algunas aplicaciones para los modelos supervisados incluyen la predicción del comportamiento de la TRM del dólar (regresión) o la categorización de un correo electrónico como spam o no spam (clasificación). Además, es importante tener en cuenta que un modelo de clasificación puede ser binario (dos categorías) o multiclase (más de dos categorías).

Modelos no Supervisados: En este tipo de aprendizaje, se diferencia por la ausencia de la variable objetivo, utilizando únicamente las variables descriptivas (Pineda, 2022). El modelo

buscara´ patrones o relaciones matemáticas entre estas variables a partir de características comunes (Pineda, 2022). Dentro de las aplicaciones o alcances de los modelos no supervisados, se incluyen la segmentación de una base de clientes en marketing y el apoyo para reducir la dimensionalidad de un conjunto de datos con una cantidad considerable de variables (Pineda, 2022).

Para el desarrollo de este proyecto, se realizó con modelos supervisados para clasificar un target, en este caso multiclase. A continuación, se relaciona algunos algoritmos de clasificación que serán importantes en el desarrollo:

1. Máquinas de Soporte Vectorial (SVM)
2. Árboles de Decisión
3. K-Nearest Neighbors (K-NN)
4. Gradient Boosting Machines (GBM)
5. LightGBM
6. XGBoost

Así mismo, se desarrollarán estrategias en función de mejorar el rendimiento del modelo, aprovechando métodos de ensamble. Para esto, métodos como el Voting (Hard y Soft) y el Stacking son estrategias que pueden ayudar a mejorar la precisión, reducir o gestionar el sobreajuste, proporcionar robustez al combinar diferentes modelos, y una de las ventajas es su flexibilidad para ser aplicados a varios problemas de clasificación o regresión.

El método de ensamble con clasificador de Voting, en términos simples, actúa como un supermodelo al ensamblar diferentes modelos. Se explicará el alcance de cada uno:

Hard: Para esta técnica, similar a un proceso democrático donde el ganador se determina por la mayoría de los votos, en aprendizaje automático, la expectativa es obtener predicciones de

múltiples modelos y considerar la predicción más frecuente como la predicción final. Así, en este caso, la predicción final se realiza por voto mayoritario, con el agregador seleccionando las predicciones de los modelos subyacentes (Manga' y cols., 2023).

Soft: Para esta técnica, no se basa en la predicción directa de clases, sino en la probabilidad promedio de cada clase. De esta manera, al calcular la probabilidad promedio para cada clase y luego compararlas, se selecciona la probabilidad más alta como la predicción final (Lasotte, Garba, Malgwi, y Buhari, 2022).

Stacking: Para aprovechar las predicciones de diferentes modelos "base", estas predicciones se combinan en un "supermodelo", ya que la filosofía de esta estrategia es aprender de lo que se ha aprendido, aprovechando así los sesgos de los diferentes modelos. Los pasos generales para aplicar esta estrategia incluyen seleccionar los diferentes modelos a entrenar, entrenar el metamodelo o "supermodelo" utilizando las predicciones de cada uno en el conjunto de validación, seguir los pasos habituales al entrenar un modelo, y finalmente evaluarlo (Dzˇeroski y Zˇ enko, 2004).

Metodología

Método

Para el desarrollo de este proyecto, se aplicará la metodología CRISP-DM, ya que es una metodología que permite el manejo de proyectos donde el objetivo es extraer valor a los datos:

Entendimiento de Negocio

Se realizará las investigaciones y estudios pertinentes, para comprender que variables pueden afectar en que una persona pueda adquirir alguna enfermedad de respiración aguda ERA. Así mismo, que entidades pueden tener información relacionada con dicho problema de salud. De esta manera, se buscarán las bases en las fuentes de extracción requeridas para la exploración, limpieza y selección de las variables. En este caso serán las siguientes fuentes requeridas:

1. Información a la Secretaría Distrital de Bogotá de Salud para acceder a la información de RIPS.
2. Información de la red de RMCAB para obtener la información de calidad del aire como meteorológica de la ciudad de Bogotá.
3. Información de humedad de fuentes de datos abiertos u otra entidad ambiental.
4. Información de incendios forestales de fuentes de datos abiertos u otra entidad ambiental.

Entendimiento de los Datos

Con las bases obtenidas para el periodo de 2019 a 2023, de las diferentes fuentes requeridas para el desarrollo del modelado, se realizará el entendimiento de las variables como su significado, de esta manera, tener un contexto global y particular de cada una de las fuentes y como puede afectar en nuestro modelamiento.

Preparación de los Datos

Con los diccionarios de datos elaborados y la contextualización de las variables, se procederá a realizar el análisis exploratorio de datos (EDA). Este análisis incluirá el uso de gráficos y técnicas estadísticas para examinar y analizar cada una de las variables. Aplicando esta metodología EDA, se garantizarán las transformaciones necesarias, las imputaciones requeridas y la creación de nuevas variables. Esto permitirá obtener atributos de alta calidad y en cantidad suficiente para asegurar un buen rendimiento del modelo.

Modelado

Con la selección de los atributos y con el objetivo de crear un algoritmo con dos modelos de clasificación binaria, se emplearán diversos algoritmos de aprendizaje automático para obtener los mejores resultados. Además, se aplicarán diferentes técnicas para optimizar las métricas, tales como el balanceo de clases, el ensamble de modelos y el ajuste de hiperparámetros. Si es necesario, se ajustarán los pasos aplicados en la preparación de los datos, ya que estos pasos son interactivos y se adaptan a las necesidades del proceso.

Evaluación

La evaluación se basará en las métricas de la precisión y sensibilidad, para determinar si el resultado del modelo cumple con los estándares de un modelo confiable.

Despliegue

Asegurando unos modelos confiables basado en las métricas, se desarrolló un notebook llamado “Apéndice F 2 Producción - Predicción Gravedad.ipynb” que realizara´ las transformaciones y aplicara´ los modelos entrenados a los nuevos registros. En esta fase, se elaboró un archivo con las especificaciones, permitiendo la lectura de nuevos archivos y la generación de predicciones sobre estos nuevos datos.

Tipo de Estudio

El tipo de estudio de este proyecto es un estudio aplicado, ya que su objetivo es abordar una problemática concreta mediante la integración y procesamiento de datos reales. En este caso, el proyecto se basa en la recopilación de información de cuatro fuentes diferentes de archivos, que contienen datos relevantes sobre enfermedades respiratorias agudas. Estos archivos serán cargados, procesados y cruzados para generar un conjunto único de datos.

Una vez que se haya consolidado este set de datos, se procederá a una fase de exploración y limpieza, utilizando técnicas de análisis exploratorio de datos (EDA) para identificar patrones, detectar inconsistencias y preparar los datos para su posterior modelado. El objetivo final es desarrollar un modelo predictivo que clasifique la gravedad de las enfermedades respiratorias agudas, lo que permitirá una mejor comprensión de los casos y una toma de decisiones más eficiente en el ámbito de la salud.

Este enfoque aplicado no solo busca generar conocimiento, sino también ofrecer una solución práctica para la predicción y clasificación de la gravedad de las enfermedades respiratorias, lo cual tiene implicaciones directas para mejorar la atención médica y optimizar el uso de los recursos en el sistema de salud.

Recolección de Datos

Para el desarrollo del modelo, se recopilaron datos de cuatro (4) fuentes distintas, provenientes de diversas instituciones, incluyendo tanto fuentes de acceso libre como entidades públicas. Las fuentes utilizadas fueron las siguientes:

1. Reporte de eventos forestales entre el 2009 y 2024 en la ciudad de Bogotá.
2. Reporte de humedad para la ciudad de Bogotá.
3. Reportes Red de Monitoreo de Calidad del Aire de Bogotá RMCAB.

4. Reportes Registro Individuales de Prestación de Servicios RIPS para la ciudad de Bogotá.

Con la descarga de la información, se realizó una exploración inicial con el lenguaje de programación de Python, obteniendo las siguientes observaciones:

Eventos Forestales

Esta información es de carácter público, desde la página de Datos Abiertos de Bogotá. La información se descargó de la página https://datosabiertos.bogota.gov.co/dataset/areas-afectadas-por-incendios-forestales-bogota-d-c, donde la descripción de la información desde la página web indica es información recopilada por el Cuerpo Especial de Bomberos de la ciudad. Esta´ información se actualiza de manera anual. En la carga inicial desde un archivo .shp, se tiene 287 registros con 11 columnas.

Reporte de Humedad

El reporte es de uso público, ya que se puede descargar de Datos Abiertos desde la página https://www.datos.gov.co/Ambiente-y-Desarrollo-Sostenible/Humedad-del-Aire-2-metros/uext-mhny/about data, información que es recopilada desde los sensores de las estaciones de monitoreo que se encuentran en los convenios con el IDEAM, ya que esta información es validada por dicha entidad. El set disponible al 09 de enero del 2025 tiene 77.4M de registros y 12 columnas, pero el set descargado el 16 de junio del 2024, contaba con 9.4M de registros con la misma cantidad de columnas.

Reportes RMCAB

Es un sistema de monitoreo ambiental continúo implementado en la ciudad de Bogotá, que cuenta con 16 estaciones distribuidas en toda la ciudad, siendo una móvil. Estas estaciones monitorean:

1. Partículas como PM10, PM2.5, PST.
2. Gases como CO, SO2, NOx, O3, CH4, NMCH4.
3. Variables meteorológicas como precipitación, velocidad y dirección del viento, temperatura, humedad relativa, presión atmosférica, radiación.

Los reportes descargados manualmente, fueron entre los años 2019 al 2023 desde la página http://rmcab.ambientebogota.gov.co/Report/stationreport . El total de registros después de la consolidación de cada uno de los reportes fue de 32.452 registros y 39 columnas.

Reporte RIPS

Este conjunto de datos, aunque proviene de una entidad pública como la Secretaría de Salud de Bogotá, al set el registro uno a uno de los pacientes que son atendidos en la red pública de salud, contiene información sensible y confidencial, por este motivo, la solicitud realizada con el SDQS 2261052024 – Solicitud datos RIPS, la generación de las columnas estuvo sujeto a los campos no confidenciales a criterios de la Secretaría de Salud de Bogotá para el periodo entre el año 2019 y 2023. El total de filas fue 1.046.590 casos de pacientes atendidos por alguna enfermedad de respiración aguda ERA con 17 columnas.

Resultados

De acuerdo con la metodología seleccionada CRISP-DM para el desarrollo del proyecto, se obtuvieron los siguientes resultados:

Comprensión del Negocio

Con base al descargue de la información, se consolidaron las bases obtenidas de fuentes públicas y de entes gubernamentales.

Análisis de los Datos

Dentro el notebook Apéndice C 1 Lectura y Unificación.ipynb, se realizó el cargue de cada una de las bases como la unificación de los 4 sets para generar el set de datos de entrenamiento.

Donde se realizó un análisis de datos comenzando con la carga de las librerías necesarias y la definición de las rutas para los archivos a utilizar. Se cargó el archivo de eventos forestales, filtrando los registros entre 2019 y 2023 y excluyendo localidades específicas, lo que resultó en 100 registros. Se realizó un conteo de los valores en la columna LOCALIDAD y se obtuvieron los valores únicos, asegurando la consistencia de los nombres mediante transformaciones. Posteriormente, se creó una nueva columna, Clasificación Incendio, y se generó un pivot para calcular el total del área afectada y la cantidad de eventos forestales.

A continuación, se cargó el set de humedad, filtrando por fechas y obteniendo más de 4 millones de registros. Se creó un pivot para calcular promedios, mínimos y máximos por fecha. Para los reportes RMCAB, se implementaron funciones de limpieza y se consolidaron los datos en un solo conjunto. Se homologaron los nombres de localidades en diferentes conjuntos de datos y se realizaron cruces entre ellos para asegurar la integridad de la información. Finalmente, se exportó el conjunto de datos final como ”Base Train.csv”, listo para su análisis posterior.

El código relacionado con los pasos anteriores puede ser visto en el archivo “Apéndice C 1 Lectura y Unificación.ipynb”.

Procesamiento de los Datos

Ver notebooks “Apéndice C 1 Lectura y Unificación.ipynb, ”Apéndice D 2 Modelamiento LightGBM- leve grave Final 1 50.ipynb” y “Apéndice E 2 Modelamiento LightGBM- leve media Final 50.ipynb”.

Modelamiento

Dentro del proceso del modelamiento, se realizó la exploración y limpieza de la información, para los notebooks “Apéndice D 2 Modelamiento LightGBM- leve grave Final 1 50.ipynb” y “Apéndice E 2 Modelamiento LightGBM- leve media Final 50.ipynb”, se realizó un análisis de datos comenzó con la carga del archivo ”Base Train.csv”, donde se convirtió el campo fecha atención al formato de fecha y se seleccionó aleatoriamente una muestra del 50 % del total de registros. Se identificaron las variables con más del 30 % de valores nulos, eliminando aquellas que no aportaban información relevante, como’CO2’ y ’Vel Viento 10M’. Para las variables críticas como ’AREA AFECT’ e ’Incendio Forestal’, se imputaron los valores nulos con 0, mientras que otros datos como ’Precipitacion’ y ’PM2.5’ se imputaron con el promedio por localidad. Además, se eliminaron variables identificadoras para evitar sesgos en el modelo.

Para definir la variable objetivo, se utilizó información del Excel . Apéndice J Tabla Enfermedades Gravedad.xlsx”, que clasifica la gravedad según la variable grupoCIE10. Durante el proceso de exploración y transformación, se realizaron diversas modificaciones: se codificó la variable sexo nombre, se extrajeron mes y día de fecha atención, y se eliminaron variables altamente correlacionadas, como NOX. Se generaron graficas de violín e histogramas para cada

variable, y se identificaron registros anómalos mediante el cálculo de valores mínimos, máximos y cuartiles, complementados con visualizaciones de cajas y bigotes. Finalmente, se establecieron rangos para las variables PM10 y Dir Viento, y se realizó el escalamiento de las variables, excluyendo la variable objetivo y el año.

Los pasos anteriores se pueden ver en los notebooks “Apéndice D 2 Modelamiento LightGBM- leve grave Final 1 50.ipynb” y “Apéndice E 2 Modelamiento LightGBM- leve media Final 50.ipynb”.

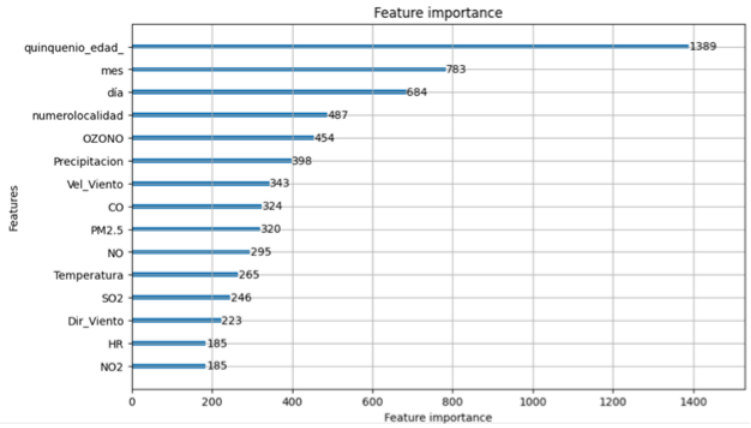
Valoración

Para el modelo que clasifica las enfermedades entre y leve y grave, las enfermedades leves y medias como clase 0, y las graves como clase 1, se identificó que las variables con mayor correlación con la variable objetivo son “Quinquenio de la edad”, “Radiación Solar”, “Ozono”, “Año”, “El género de la persona”, “Los incendios pequeños, catalogados dentro de la variable Otros Incendios”.

Así mismo, con una tasa natural del 1.4 % en la clase minoritaria, se aplicó un balanceo de clases con la estrategia SMOTE. Por otro lado, entre las 15 variables más importantes e influyentes en el modelo, la edad de la persona destaca como un factor clave. Asimismo, las variables relacionadas con la fecha, como el mes y el día, muestran una relevancia significativa, reflejando coherencia con los picos habituales observados en la ciudad de ERA.

Figura 1

Importancia de las Variables en el Modelo Leves-Graves

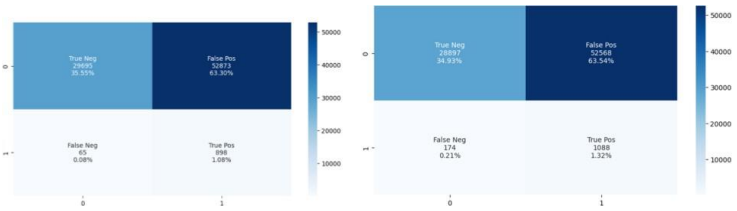


Nota. La Figura Muestra la Importancia de las Variables Dentro del Modelo Entrenado.

Al entrenar los modelos aplicando el algorimo LightGBM, con tunning de hiperparámetros, se obtuvo un recall del 93 % para el set de test, mientras con el set de evaluación fue del 86 %, pesar de esta reducción, el rendimiento del modelo sigue siendo sólido, indicando un desempeño satisfactorio.

Figura 2

Matrices de Confusión (Error) (Test/Validación) - Modelo Leves Graves



Nota. La Figura Muestra las Matrices de Confusión entre los Resultados de Test y Validación, para Calcular las Métricas.

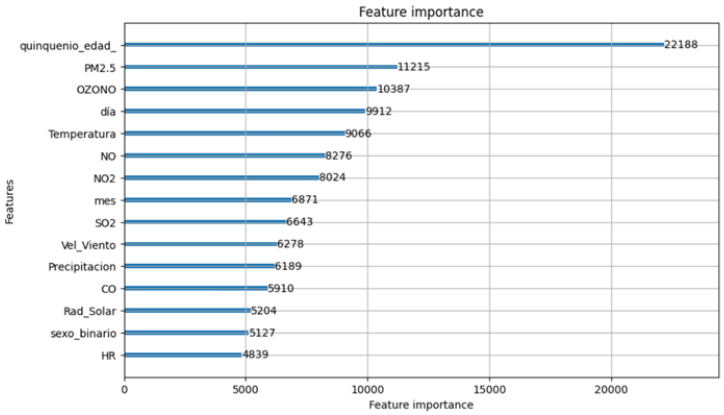
Para el modelo que clasifica las enfermedades entre y leve y media, donde el primer

modelo clasifica las enfermedades leves y medias como clase 0, y las graves como clase 1, se identificó que las variables con mayor correlación con la variable objetivo son “Localidad”, “HR”, “Presión Barométrica”, “PM2.5” y. “Precipitación”.

Así mismo, con una tasa natural del 44.4 % en la clase minoritaria, se aplicó un balanceo de clases con la estrategia SMOTE. Por otro lado, entre las 15 variables más importantes e influyentes en el modelo, la edad de la persona destaca como un factor clave. En este modelo el material particulado 2.5 como el contaminante OZONO, influenciaron dentro del modelo.

Figura 3

Importancia de Variables en el Modelo Leves Media

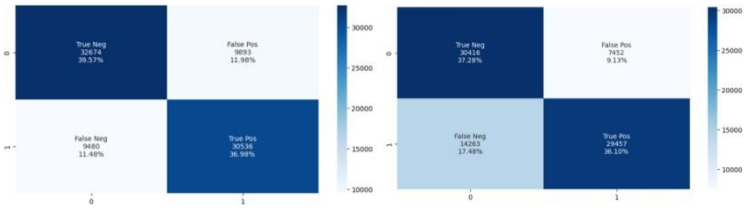


Nota. La Figura Muestra la Importancia de las Variables Dentro del Modelo Entrenado.

Al entrenar los modelos aplicando el algoritmo LightGBM, con tunning de hiperparámetros, se obtuvo una precisión del 76 % para el set de test, mientras con el set de evaluación fue del 80 %, mejorando el rendimiento del modelo, indicando un desempeño satisfactorio.

Figura 4

Matrices de Confusión (Error) (Test/Validación) - Modelo Leves Media



Nota. La Figura Muestra las Matrices de Confusión entre los Resultados de Test y Validación, para Calcular las Métricas.

Implementación del Modelo

Entonces con la selección de las mejores métricas para los modelos entrenados, en el primer modelo (“leve - grave”), que predice entre una enfermedad grave y no grave con una incidencia natural del 1.4 %, se priorizó alcanzar un alto recall para identificar el mayor número posible de casos graves. El modelo final logró un rendimiento del 86 % en esta métrica.

Por otro lado, el segundo modelo (“leve - media”) se enfocó en distinguir entre una enfermedad de severidad media y leve, con una incidencia natural del 44.4 %. El objetivo en este caso fue optimizar el área bajo la curva ROC (ROC AUC) para equilibrar la precisión y la sensibilidad. El modelo obtuvo una precisión del 80 % y una sensibilidad del 67 % para la clase Media (1). Con los modelos entrenados y exportados, se desarrolló el notebook “Apéndice F 2 Producción - Predicción Gravedad.ipynb”, donde realiza el cargue, preparación, limpieza de los datos, así mismo, la predicción de los modelos entrenados para los nuevos registros de entrada. Los archivos de los modelos entrenados son:

1. Modelo leve grave “’Apéndice H LGBMfinal 50.pkl”
2. Modelo leve Media “Apéndice G LGBMfinal 50 leve media.pkl”

Conclusiones

Este trabajo examina la influencia de contaminantes ambientales, como los materiales particulados PM2.5 y PM10, así como de variables ambientales como la humedad y los reportes de incendios. Para ello, se analizaron datos recopilados en archivos planos y Excel provenientes de cuatro fuentes diferentes utilizadas a lo largo del proyecto. Después de un análisis exhaustivo de 44 variables obtenidas de estas diversas entidades, se llevó a cabo una minuciosa exploración y limpieza de los datos, siguiendo la metodología de Análisis Exploratorio de Datos (EDA).

Como resultado de este proceso, el modelado final se realizó utilizando 22 variables seleccionadas por su relevancia y calidad.

Con el modelamiento, se tuvo en cuenta que la distribución natural de los casos reveló una tasa de gravedad del 1.4 %, en comparación con el 54.8 % de los casos leves y el 43.8 % de los casos moderados. Dado este desequilibrio, se adoptó una estrategia de modelado en dos etapas. En la primera etapa, se unificaron los casos leves y moderados para entrenar un modelo enfocado en predecir la clase minoritaria de casos graves. En la segunda etapa, sobre los casos clasificados como" leve-moderado "por el primer modelo, se implementó un segundo modelo para diferenciar entre casos leves y moderados.

Esta estrategia de modelado permitió mejorar la precisión en la clasificación de la clase minoritaria, a pesar de trabajar solo con el 50 % de la información disponible. Las métricas obtenidas reflejaron un buen desempeño del enfoque adoptado, destacando la efectividad de dividir el problema en dos fases para manejar el desequilibrio de clases y optimizar la predicción de casos graves.

En el primer modelo, se priorizó la optimización de la métrica de recall, ya que, debido a la naturaleza crítica de los casos graves, es fundamental reducir al mínimo los falsos negativos.

Un alto número de falsos negativos implica que muchos casos graves no son detectados, lo que puede resultar en la falta de tratamiento o atención oportuna, con potenciales consecuencias graves para los pacientes, incluida la muerte. Por ello, es crucial que el modelo detecte la mayor cantidad posible de casos graves.

Cuando se trata de enfermedades graves, la detección temprana es esencial. Identificar una enfermedad respiratoria aguda (ERA) grave en una etapa temprana permite iniciar el tratamiento adecuado antes de que la condición del paciente se deteriore. Esto puede mejorar significativamente las posibilidades de recuperación, reducir la duración de la enfermedad y prevenir complicaciones severas. La capacidad de un modelo para detectar tempranamente estos casos puede marcar la diferencia entre una recuperación exitosa y un desenlace fatal.

Además, en contextos médicos, la detección temprana de condiciones graves no solo mejora los resultados para los pacientes, sino que también ayuda a optimizar los recursos sanitarios. Al identificar rápidamente los casos que requieren atención urgente, los sistemas de salud pueden priorizar la asignación de recursos, garantizando que los pacientes más críticos reciban la atención que necesitan. Por estas razones, en el desarrollo de modelos predictivos para enfermedades graves, el recall es una métrica esencial, ya que asegura que los casos críticos no pasen desapercibidos, contribuyendo a una respuesta médica más efectiva y oportuna.

En el segundo modelo, el enfoque se centró en lograr un equilibrio entre precisión y sensibilidad, utilizando el ROC AUC como métrica de evaluación. Dado que las clases de enfermedades leves y medias están relativamente balanceadas, el modelo puede distinguir mejor entre ellas.

Reducir los falsos positivos en los casos de enfermedades medias es fundamental para evitar ansiedad innecesaria en los pacientes y prevenir tratamientos médicos inapropiados. Por

otro lado, minimizar los falsos negativos es crucial, ya que clasificar erróneamente una enfermedad media como leve puede ser peligroso, permitiendo que la condición del paciente empeore y progrese a una enfermedad grave.

Por lo tanto, el uso del ROC AUC como métrica es la mejor estrategia en este contexto, ya que permite aprovechar el balance entre las clases para lograr un diagnóstico más preciso y seguro, asegurando que los pacientes reciban la atención adecuada según la gravedad de su condición. Al integrar estos modelos en un sistema capaz de consumir las variables provenientes de las fuentes trabajadas en este proyecto, se busca maximizar el valor de la información aportada por cada variable. Este enfoque permitirá no solo combinar y aprovechar eficientemente los datos disponibles, sino también orientar los resultados hacia la predicción de la gravedad de las enfermedades. De esta manera, se contribuirá a la reducción de los costos tanto de vida como de tratamiento. Además, la herramienta podría potenciarse aún más mediante la inclusión de un mayor número de casos o la incorporación de variables adicionales que los expertos puedan recomendar, mejorando así su precisión y alcance.

Igualmente dado que el rendimiento del computador limitó el trabajo sobre la base completa de aproximadamente 1.1 millones de registros de RIPS, se recomienda considerar el uso de equipos con mayor capacidad de procesamiento o servicios de computación en la nube. Esto permitirá trabajar con la totalidad de los datos, mejorando la precisión y representatividad de los análisis.

Alternativamente, si no es posible, continuar utilizando muestras aleatorias bien estratificadas puede ser una opción válida, asegurando la diversidad y la representatividad de los datos.

Recomendaciones

Con el desarrollo del proyecto y la experiencia con la información disponible, desde el punto del modelamiento para potenciar las métricas de los modelos, sería ideal incorporar información adicional sensible, como enfermedades preexistentes, antecedentes familiares, ubicación exacta de vivienda, características físicas como el peso, entre otras. Estas variables pueden proporcionar un contexto más rico y mejorar significativamente la precisión de los modelos predictivos. No obstante, es crucial garantizar la seguridad y privacidad de los datos, cumpliendo con las normativas legales y éticas correspondientes para trabajos futuros.

Desde el cambio de la estructura de los datos iniciales utilizados, el procesamiento de información descargada desde RMCAB puede variar con nuevas mediciones o la supresión de algunas existentes, es fundamental implementar un sistema de monitoreo continuo de la estructura de datos. Esto permitirá detectar y adaptarse rápidamente a cualquier cambio, minimizando el riesgo de interrupciones en los análisis y manteniendo la integridad de los modelos.

Pero una recomendación importante, para mejorar el rendimiento de los modelos, es recomendable colaborar con expertos en las enfermedades estudiadas. Su conocimiento puede ayudar a identificar variables adicionales relevantes, interpretar mejor los resultados, y ajustar los modelos para reflejar de manera más precisa la realidad médica. Así mismo, se debe buscar informes y bases de datos adicionales que ofrezcan información más detallada sobre los casos de incendios forestales. Esto incluiría datos específicos sobre la localización, causas, impacto, y medidas preventivas. Una base de datos más robusta permitirá realizar análisis más precisos y elaborar estrategias de prevención y respuesta más eficaces.

Este aporte es igualmente significativo para la sociedad, ya que proporcionara´ una

herramienta que integra información sobre la ciudad con datos de la población general. Esto permitirá establecer relaciones entre variables de contaminación y la ubicación de los individuos, ofreciendo una perspectiva valiosa. Además, servirá como un recurso para el sistema de salud del distrito, facilitando la focalización de los recursos necesarios para reducir la morbilidad y mortalidad, especialmente entre los niños menores de 5 años y los adultos mayores.

Referencias

Acosta González, O. M., y cols. (2022). *Óxido nítrico: reactividad, propiedades y biodisponibilidad*. https://riull.ull.es/xmlui/handle/915/29023

Actual y perspectivas, S. (2024). *Protocolo de Kioto*. Ceida.org. https://www.ceida.org/prestige/Documentacion/Protocolo%20Kioto.pdf (Recuperado el 26 de junio de 2024)

Apaza Cabrera, R. (2018). *Impacto de la contaminación ambiental en la salud de la población de Arequipa Metropolitana en el periodo* 2013-2017.

Arciniegas Suarez, C. A. (2012). *Diagnóstico y control de material particulado: Partículas suspendidas totales y fracción respirable PM10*. Luna Azul, (34), 195-213. http://www.scielo.org.co/scielo.php?script=sci˙arttext&pid=S1909-24742012000100012&lng=en&tlng=es (Recuperado el 24 de junio de 2024, de http://www.scielo.org.co/scielo.php?script=sci$arttext\ pid$ =S1909 − 24742012000100012lng = $entlng$ = es)

Aristizábal, G., Hernández, L., y Medina, K. (2009). Asociación entre la contaminación del aire y la morbilidad por enfermedad respiratoria aguda en menores de 5 años en tres localidades de Bogota´. *Revista de Salud Pública*, 12(), https://pdf.sciencedirectassets.com/313022/1-s2.0-S0120491212X52005/1-s2.0-S0120491215300112/main.pdf

atmosférica, P. (2022, 24 de february*). Presión atmosférica*. Eltiempo.es. https://www.eltiempo.es/noticias/meteopedia/presion-atmosferica

BBC News Mundo. (2019). *El gas con efecto invernadero 23.500 veces más potente que el dióxido de carbono y del que muchos jamás han oído hablar*. BBC.

https://www.bbc.com/mundo/noticias-49717228 (Recuperado el 16 de septiembre de 2019) CNE. (2024). Recomendaciones y Consejos para la Prevención de Incendios Forestales. (Recuperado el 28 de junio de 2024, de Cne.go.cr)

Coalición Clima y Aire Limpio (CCAC). (2024). *Hidrofluorocarbonos* (HFC). Sitio web de la Coalición Clima y Aire Limpio. https://www.ccacoalition.org/es/short-lived-climate-pollutants/hydrofluorocarbons-hfcs (Recuperado el 26 de junio de 2024)

Comisión Europea. (2024). *Óxido de nitrógeno* (NOx). Sitio web de la Comisión Europea. https://ec.europa.eu/health/scientific˙committees/opinions˙layman/es/contaminacion-aire-interior/glosario/mno/oxidos-nitrogeno-nox-oxido-nitrico-no-dioxido-nitrogeno-no2.html (Recuperado el 26 de junio de 2024)

DANE. (2024). *Velocidad del Viento*. Gov.co. https://www.dane.gov.co/files/investigaciones/pib/ambientales/Sima/88HM-Velocidad-del- viento-4.pdf (Recuperado el 28 de junio de 2024)

Darquea, D. G. P. (2018). Estudio de emisiones contaminantes utilizando combustibles locales. INNOVA *Research Journal*, 3(3), 23–34. de Cambio Climático de las Américas, A. A. (2024). ACCA: Atlas de Cambio Climático de las Américas. Aragon.es. https://idearagon.aragon.es/lib/IDEAragon/examples/ACCA/precipitacion.html (Recuperado el 28 de junio de 2024)

Organización Panamericana de la Salud, O. P. (1992*). Clasificación Estadística Internacional de Enfermedades y Problemas Relacionados con la Salud* (10.a ed.) (10.a ed.). O. P. de la Salud: Autor.

del viento, D. (2023, 13 de july). *Dirección del viento*. Eltiempo.es. https://www.clima.com/meteopedia/direccion-viento de Salud, M. (2017). Infecciones

Respiratorias Agudas (IRA). https://www.minsalud.gov.co/salud/Paginas/InfeccionesRespiratorias-Agudas-(IRA) (Recuperado el 17 de marzo de 2017)

Dióxido de carbono. (2024). *Dióxido de carbono*. Sitio web de la NASA. https://climate.nasa.gov/en-espanol/signos-vitales/dioxido-de-carbono/?intent=111 (Recuperado el 26 de junio de 2024)

Díaz, J. (2021). Aprendizaje Automático y Aprendizaje Profundo. Ingeniare. *Revista chilena de ingeniería*, 29(2), 180–181. doi: 10.4067/S0718-33052021000200180

Dzˇeroski, S., y Zˇ enko, B. (2004). *Is combining classifiers with stacking better than selecting the best one? Machine Learning*, 54(), 255–273.

Egas, C., Naulin, P. I., y Pre´ndez, M. (2018). *Urban Pollution by Particulate Matter and its Effect on Morpho-Anatomical Characteristics of Four Tree Species in Santiago*, Chile. Información Tecnológica, 29(4), 111-118. https://dx.doi.org/10.4067/S0718-07642018000400111

Emjen, C. R. (2020). *Contaminación atmosférica y medioambiental y patología respiratoria. EMC-Tratado de Medicina*, 24(3), 1-9.

Clima.com, (2023, 14 de june). *¿Que´ es la temperatura y cómo se mide?* Eltiempo.es. https://www.clima.com/meteopedia/temperatura

González-Díaz, S. N., Lira-Quezada, C. E. D., Villarreal-González, R. V., y Canseco-Villarreal, J. I. (2022). Contaminación ambiental y alergia. *Revista Alergia México*, 69(), 24–30.

González-Parra, G., Querales, J. F., y Aranda, D. (2016). *Predicción de la epidemia del virus sincitial respiratorio en Bogotá, D.C., utilizando variables climatológicas*. Biomédica, 36(3), 378–389. https://doi.org/10.7705/biomedica.v36i3.2763 doi:

10.7705/biomedica.v36i3.2763 Humedad. (2014, 18 de february). Humedad. Eltiempo Previsto, S.L.U.

IDEAM, R. S. (2024). RADIACIÓN SOLAR. Gov.co. https://www.eltiempo.es/noticias/meteopedia/humedad

RADIACIÓN SOLAR - IDEAM. (s. f.). http://www.ideam.gov.co/web/tiempo-y-clima/radiacion-solar-ultravioleta

Jiménez Manjarrés, M. (2019). Aplicación de analítica de datos para predicción de infección respiratoria aguda en Colombia. http://hdl.handle.net/1992/43984

Lasotte, Y. B., Garba, E. J., Malgwi, Y. M., y Buhari, M. A. (2022). An Ensemble Machine **Learning Approach for Fake News Detection and Classification Using a Soft Voting Classifier. European Journal of Electrical Engineering and Computer Science**, 6(2), 1–7. doi: 10.24018/ejece.2022.6.2.409

Lian, Y. X., Xu, Y. H., H, S. M., y Xing, Y. F. (2016*). The impact of PM2.5 on the human respiratory system. Journal of Thoracic Disease*, 8(1), E69–E74. https://doi.org/10.3978/j.issn.2072-1439.2016.01.19)

Macedo, M., y Mateos, S. (2006). *Temas de Bacteriología y Virología Médica* (2o Edición Corregida) (2.a ed.). U. de la República: Universidad de la República.

Manga', A. R., Handayani, A. N., Herwanto, H. W., Asmara, R. A., Sulistya, Y. I., y Kasmira, K. (2023). *Analysis of the ensemble method classifier's performance on handwritten Arabic characters dataset*. ILKOM Jurnal Ilmiah, 15(1), 186–192. doi: 10.33096/ilkom.v15i1.1357.186-192

Martha´ Eraso, J. (2014). *Impacto de la variabilidad climática en las enfermedades respiratorias agudas en Bogotá - una aproximación por modelos de regresión dinámica y análisis de*

series de tiempo. http://hdl.handle.net/1992/11955)

Martínez Ataz, E., y Díaz de Mera Morales, Y. (2004). *Contaminación atmosférica*. https://dialnet.unirioja.es/servlet/libro?codigo=6152

Marulanda, M. M. (2024, mayo 18). *Incremento en casos de infección respiratoria aguda en Bogotá durante primer pico del año, ya se registran 668.282 casos*. Infobae. https://www.infobae.com/colombia/2024/05/18/incremento-en-casos-de-infeccion-respiratoria-aguda-en-bogota-durante-primer-pico-del-ano-ya-se-registran-668282-casos/

Ministerio de Salud. (2017). *Infecciones Respiratorias Agudas* (IRA). https://www.minsalud.gov.co/salud/Paginas/InfeccionesRespiratorias-Agudas-(IRA))

Mora-Barrantes, J. C., Sibaja-Brenes, J. P., y Borbón-Alpizar, H. (2021). Fuentes antropogénicas y naturales de contaminación atmosférica: estado del arte de su impacto en la calidad fisicoquímica del agua de lluvia y de niebla. *Revista Tecnología en Marcha*, 34(1), 92-103. https://dx.doi.org/10.18845/tm.v34i1.4806

Murcia Urrego, L. (2019). *Determinantes sociales de la morbimortalidad infantil por Infección Respiratoria Aguda en Bogotá 2015-2016* (Tesis de Master, Universidad de los Andes). http://hdl.handle.net/1992/44267

Olaya, C. (2024, marzo 8). *¿Que´ esta´ pasando con la calidad del aire en Bogotá? El Nuevo Siglo*, (https://www.elnuevosiglo.com.co/nacion/que-esta-pasando-con-la-calidad-del-aire)

Olaya Ochoa, J., Ovalle, D. P., y Urbano, C. L. (2017). Acerca de la estimación de la fracción PM 2.5 /PM 10. DYNA, 84(203), 343-348. https://doi.org/10.15446/dyna.v84n203.65228

Onatra, W., Vargas, S., Páez, E., Rojas, D., y López, A. (2009). Correlación entre la enfermedad respiratoria aguda (ERA) en mujeres embarazadas y la calidad del aire. *Revista U.D.C.A*

Actualidad & Divulgación Científica, 12(2), 27–37. doi: 10.31910/rudca.v12.n2.2009.689

Organización Mundial de la Salud. (2016). Calidad del aire ambiente (exterior) y salud. Informe. http://www.who.int/mediacentre/factsheets/fs313/es/

Parlamento Europeo. (2024). *Cambio climático: gases de efecto invernadero que causan el calentamiento global. Temas — Parlamento Europeo.* https://www.europarl.europa.eu/topics/es/article/20230316STO77629/cambio-climatico-gases-de-efecto-invernadero-que-causan-el-calentamiento-global

Pineda, C. (2022). *Aprendizaje automático y profundo en Python*. RA-MA.

Raschka, S., y Mirjalili, V. (2019). *Python machine learning: aprendizaje automático y aprendizaje profundo con Python, scikit-learn y TensorFlow. Marcombo.*

Rojas, E., y Aguilar, J. (2017). *Minería de datos para el descubrimiento de patrones en enfermedades respiratorias en Bogotá, Colombia* (Trabajo de grado). https://repository.ucatolica.edu.co/server/api/core/bitstreams/53df941b-73ad-494c-8553-fc724f9c3df4/content

Rojas, N. Y. (2024). Aire y problemas ambientales de Bogotá. Universidad Nacional de Colombia. https://bogota.gov.co/sites/default/files/inline- files/aire*y problemasambientalesd ebogota.pdf*

Romero Placeres, M., Olite, D. F., y A´ lvarez Toste, M. (2006). La contaminación del aire: su repercusión como problema de salud. *Revista Cubana de Higiene y Epidemiologia*, 44(2), . http://scielo.sld.cu/scielo.php?script=sci˙arttext&pid=S1561-30032006000200008&lng=es&tlng=es (Recuperado en 24 de junio de 2024, de http://scielo.sld.cu/scielo.php?script=sci*arttext pid* = *S*1561 − 30032006000200008*lng*

=$estlng = es$)

Secretaría de Salud de Bogotá. (2025). *Secretaría de Salud de Bogotá recomienda reforzar medidas de prevención contra enfermedades respiratorias*. Gov.co. https://www.minsalud.gov.co/Regiones/Paginas/Secretar%C3%ADa-de-Salud-de-Bogot%C3%A1-recomienda-reforzar-medidas-de-prevenci%C3%B3n-contra-enfermedades-respiratorias.aspx

Apéndices

Apéndice A

Lista Embebidos Enlace de Notebooks

https://github.com/rlmendez/Trabajo-Grado-UNAD---Especializacion/issues/1

1. Carta Secretaría Distrital de Salud Bogotá 2024-EE-61831.pdf
2. Diccionario de Datos.xlsx
3. Lectura y Unificación.ipynb
4. 2 modelamiento LightGBM - leve grave Final 1 50.ipynb
5. 2 modelamiento LightGBM - leve media Final 50.ipynb
6. 2 producción - Predicción Gravedad.ipynb
7. LGBMfinal 50 leve media.pkl
8. GBMfinal 50.pkl
9. Especificaciones Técnicas de Notebooks.docx
10. Tabla Enfermedades Gravedad.xlsx