

**Análisis de la relación entre patrones climáticos y enfermedades respiratorias agudas en
Bogotá, Colombia: un estudio de 2025**

Roberto Julián Herrera Pérez

Asesor

Fernando Luis Carrascal

Universidad Nacional Abierta y a Distancia UNAD
Escuela de Ciencias Básicas, Tecnología e Ingeniería ECBTI
Especialización en Ciencia de Datos y Analítica
2025

Nota de Aceptación

Fernando Luis Carrascal
Director de Trabajo de Grado

Jurado

Jurado

Dedicatoria

A mi madre que me mira desde el cielo y es testigo del esfuerzo y el empeño para cumplir con este proposito.

Resumen

El cambio climático y la contaminación atmosférica han intensificado la incidencia de enfermedades respiratorias en Colombia, especialmente en grandes ciudades. Este estudio, titulado "Análisis de la relación entre patrones climáticos y enfermedades respiratorias agudas en Bogotá, Colombia: Un estudio de 2025", busca evaluar la relación entre las variaciones climáticas y las enfermedades respiratorias más comunes, utilizando modelos estadísticos como la correlación y la regresión lineal a través de Python. El análisis se centra en identificar las enfermedades respiratorias predominantes durante el 2020 y el 2024, explicando cómo las condiciones climáticas influyen en su prevalencia y determinar así cuáles son las variables climáticas que tienen un mayor impacto en estas enfermedades.

El trabajo se fundamenta en estudios previos internacionales, como el análisis del impacto del ozono en la mortalidad en Turquía y revisiones sobre factores climáticos asociados con el asma infantil. Aunque dichos estudios no están directamente vinculados al contexto colombiano, aportan una base metodológica valiosa. El trabajo de investigación pretende generar un marco teórico y analítico que facilite la predicción de escenarios de riesgo, contribuyendo a estrategias preventivas de salud pública. Así, busca no solo entender cómo el clima afecta la salud respiratoria, sino también desarrollar herramientas prácticas que puedan ser utilizadas por instituciones educativas y de salud para mitigar los impactos de este fenómeno creciente.

Palabras claves: Aprendizaje Automático, Clima, Correlación, Enfermedades respiratorias, Regresión lineal.

Abstract

Climate change and air pollution have intensified the incidence of respiratory diseases in Colombia, particularly in large cities. This study, entitled “Analysis of the Relationship Between Climate Patterns and Respiratory Diseases Acute in Bogota, Colombia: A 2025 Study,” aims to evaluate the relationship between climate variations and the most common respiratory diseases by applying statistical models such as correlation and linear regression through Python. The analysis focuses on identifying the predominant respiratory diseases during 2020 and 2024, explaining how climatic conditions influence their prevalence, and determining which climatic variables have the greatest impact on these diseases.

The research is grounded in previous international studies, such as the analysis of ozone-related mortality in Turkey and reviews on climatic factors associated with childhood asthma. Although these studies are not directly linked to the Colombian context, they provide a valuable methodological foundation. This research seeks to develop a theoretical and analytical framework that enables the prediction of risk scenarios, thereby contributing to preventive public health strategies. In doing so, it aims not only to understand how climate affects respiratory health, but also to design practical tools that educational and health institutions can use to mitigate the impacts of this growing phenomenon.

Keywords: Machine Learning, Climate, Correlation, Respiratory Diseases, Linear Regression.

Tabla de Contenido

Introducción	10
Descripción del Problema	13
Justificación	15
Objetivos	17
Objetivo General.....	17
Objetivos Específicos	17
Marco Teórico.....	18
Metodología	34
Resultados – Desarrollo de los Modelos.....	36
Datos de Salud Unificado Entre 2020 y 2024	37
Calidad del Aire entre 2020 y 2024.....	45
Unificación de los Dataframe de Salud y Calidad del Aire.....	55
Análisis de Correlación	56
Prueba Kruskal-Wallis.....	59
Regresión Lineal Múltiple	60
Causalidad de Granger.....	63
Conclusiones.....	66
Recomendaciones	70
Referencias.....	72
Apendices.....	76

Lista de Tablas

Tabla 1 <i>Bases de Datos Objeto de Análisis</i>	37
Tabla 2 <i>Registros de Salud por Año</i>	38
Tabla 3 <i>Cantidad de Casos por Municipio de Ocurrencia</i>	39
Tabla 4 <i>Cantidad de Casos por Estado</i>	39
Tabla 5 <i>Cantidad de Casos por Genero</i>	40
Tabla 6 <i>Cantidad de Casos por Municipio por Estado</i>	40
Tabla 7 <i>Cantidad de Casos por Rango de Edad en Bogotá</i>	42
Tabla 8 <i>Puntos de Corte del ICA (Índice de calidad del Aire)</i>	46
Tabla 9 <i>Puntos de Corte ICA llevados al modelo</i>	47
Tabla 10 <i>Cantidad de Mediciones por Nivel</i>	49
Tabla 11 <i>Relación de Cantidad de Medición por Categoría por Año</i>	52

Lista de Figuras

Figura 1 <i>Distribución de Casos por Estado en Bogotá</i>	42
Figura 2 <i>Distribución de Casos por Grupo de Edad y Estado en Bogotá</i>	43
Figura 3 <i>Distribución de Edades por Genero en Bogotá</i>	44
Figura 4 <i>Tendencia Mensual de Casos por Año en Bogotá</i>	45
Figura 5 <i>Mapa de Calor del ICA Promedio por Año por Estación en Bogotá</i>	51
Figura 6 <i>Distribución de Frecuencias del ICA por Año</i>	52
Figura 7 <i>Mapa de Calor de Clasificación de Categorías por Año</i>	53
Figura 8 <i>Tendencia Mensual del ICA Promedio</i>	54
Figura 9 <i>ICA Promedio por Contaminante por Año</i>	54
Figura 10 <i>Muestra del Dataframe Unido Entre Calidad del Aire y Salud</i>	55
Figura 11 <i>Diagrama de Dispersión entre ICA y los Casos de Salud</i>	56
Figura 12 <i>Correlación de Pearson por Contaminante</i>	58
Figura 13 <i>Mapa de calor correlación</i>	60
Figura 14 <i>Modelo de Regresión Lineal</i>	61
Figura 15 <i>Distribución de Residuos</i>	62
Figura 16 <i>Relación entre los Residuos y las Predicciones</i>	62

Lista de Apéndices

Apéndice A <i>Enlaces de Información</i>	76
---	----

Introducción

El cambio climático constituye uno de los desafíos más urgentes y complejos de la humanidad en el siglo XXI, no solo por sus impactos ambientales, sino también por las profundas repercusiones que tiene en la salud pública. En este escenario, las enfermedades respiratorias han adquirido una relevancia particular, debido a su estrecha relación con factores climáticos y atmosféricos. La variabilidad de la temperatura, la humedad y la precipitación, sumada a la persistente contaminación del aire en las grandes ciudades, ha intensificado la incidencia de afecciones respiratorias en diferentes regiones del mundo. Colombia no es ajena a esta situación: las principales ciudades del país han experimentado un aumento sostenido en la prevalencia de estas enfermedades, lo que evidencia la necesidad de comprender con mayor precisión la interacción entre clima y salud.

En el contexto colombiano, los patrones climáticos muestran características de alta variabilidad, influenciados tanto por situaciones locales como por dinámicas globales, tales como los fenómenos de El Niño y La Niña. Estos cambios, en combinación con los altos niveles de material particulado y contaminantes presentes en ciudades como Bogotá, Medellín y Cali, generan condiciones propicias para la aparición y el agravamiento de enfermedades respiratorias. Este panorama plantea interrogantes sobre cuáles son los factores climáticos más determinantes en la salud de la población y cómo su análisis puede aportar al diseño de estrategias preventivas y de mitigación en el ámbito de la salud pública.

El presente trabajo, titulado “Análisis de la relación entre patrones climáticos y enfermedades respiratorias en Colombia: Un estudio de 2025”, se centra en evaluar esta relación a partir de la aplicación de modelos estadísticos de correlación y regresión lineal implementados en Python. La investigación toma como referencia los datos entre el 2020 y el 2024, tanto en lo

relacionado con la incidencia de enfermedades respiratorias más comunes, como en el comportamiento de variables climáticas relevantes. De esta manera, se busca no solo identificar los patrones predominantes, sino también explicar cómo las condiciones del clima inciden en la prevalencia de estas enfermedades y cuáles son las variables que presentan un mayor impacto en el contexto colombiano.

La fundamentación de este estudio se apoya en literatura científica internacional, entre la que se destacan investigaciones sobre la mortalidad asociada al ozono en Turquía y estudios que relacionan el asma infantil con factores climáticos. Si bien estos referentes no se desarrollan específicamente en Colombia, aportan una base metodológica y conceptual sólida que orienta el análisis en este trabajo. Asimismo, la revisión de estos antecedentes permite comprender que, aunque la relación entre clima y salud ha sido explorada en diferentes países, aún existe un vacío importante en el ámbito latinoamericano y, en particular, en el caso colombiano, donde las particularidades geográficas y socioambientales requieren aproximaciones propias y contextualizadas.

La pertinencia de esta investigación se sustenta en la necesidad de generar un marco teórico y analítico que facilite la predicción de escenarios de riesgo, contribuyendo al fortalecimiento de las políticas y estrategias en salud pública. Identificar con claridad la influencia de los patrones climáticos en las enfermedades respiratorias no solo tiene valor académico, sino también práctico, al ofrecer herramientas que pueden ser utilizadas por instituciones educativas, entidades de salud y organismos gubernamentales en la formulación de programas de prevención. En un país donde las desigualdades sociales y las condiciones ambientales agudizan la vulnerabilidad de ciertos grupos poblacionales, este tipo de análisis representa un aporte significativo al bienestar colectivo.

En suma, la introducción de este trabajo plantea la necesidad de estudiar con rigurosidad la relación entre clima y enfermedades respiratorias en Colombia, entendiendo el fenómeno desde una perspectiva estadística y contextualizada. El propósito central es avanzar en la comprensión de cómo la variabilidad climática incide en la salud de la población, y al mismo tiempo, proponer un camino hacia la construcción de estrategias basadas en la evidencia que permitan mitigar los efectos adversos de un fenómeno creciente y de alcance global.

Descripción del Problema

El cambio climático ha exacerbado la variabilidad climática global, impactando directamente la salud humana. En Colombia, las variaciones bruscas de temperatura, humedad y precipitación, combinadas con niveles elevados de contaminación en las grandes ciudades, han incrementado la prevalencia de enfermedades respiratorias. Estas afecciones son especialmente notorias en la cotidianidad urbana, donde frases como "¡Están dando unas gripas!" reflejan la percepción generalizada de un problema de salud pública que no se ha abordado de manera integral.

La relación entre el clima y las enfermedades respiratorias no es nueva, pero su análisis sistemático sigue siendo limitado, especialmente en contextos locales como el colombiano. Las condiciones climáticas variantes, como lluvias extremas, olas de calor, inundaciones, fenómenos del niño o de la niña actúan como desencadenantes para afecciones como el asma, la bronquitis y otras infecciones respiratorias. Estos efectos se ven agravados por factores antropogénicos, como la contaminación vehicular y la deforestación, que aumentan la vulnerabilidad de las poblaciones expuestas. (Cediel Becerra et al., 2023)

Aunque existen estudios internacionales, como el análisis del impacto del ozono en la salud cardiovascular y respiratoria en Turquía (İbrahim & Özkan, 2024) o la revisión de factores climáticos relacionados con el asma infantil (Wang et al., 2024), el contexto colombiano presenta particularidades que requieren atención. La diversidad climática del país y sus dinámicas urbanas únicas hacen que los hallazgos de otros países sean insuficientes para comprender completamente el problema en Colombia.

En este sentido, los modelos predictivos y el análisis estadístico son herramientas clave para desentrañar la relación entre las condiciones climáticas y las enfermedades respiratorias.

Estas herramientas permiten identificar variables críticas, predecir escenarios de riesgo y proponer estrategias para mitigar el impacto de estos fenómenos en la población.

El problema afecta a diferentes sectores. Desde una perspectiva educativa, plantea un desafío para las instituciones académicas que buscan contribuir al entendimiento y la solución de problemas sociales mediante la investigación científica. Además, tiene implicaciones directas en las políticas de salud pública, donde los resultados de este estudio podrían ayudar a diseñar estrategias preventivas basadas en evidencia científica, mejorando la calidad de vida de la población colombiana. Por lo tanto, es esencial realizar estudios localizados que analicen la relación entre los patrones climáticos y las enfermedades respiratorias en Colombia, utilizando datos específicos y metodologías avanzadas, para aportar soluciones tangibles a un problema que afecta a todos los habitantes del país.

¿De qué manera las variables climáticas se relacionan en la prevalencia de las enfermedades respiratorias agudas en Bogotá, Colombia durante el año 2024?

Justificación

La relevancia de este estudio radica en la necesidad urgente de entender cómo las condiciones climáticas impactan la salud respiratoria de la población colombiana. Las enfermedades respiratorias representan una carga significativa para el sistema de salud, afectando tanto a individuos como a colectivos, especialmente en las ciudades principales como Bogotá, Medellín, Cali, Barranquilla, con altos niveles de contaminación. Este problema, amplificado por el cambio climático, demanda respuestas fundamentadas en el análisis científico y el uso de tecnologías avanzadas.

Desde el ámbito académico y científico, este estudio aporta una perspectiva integral mediante el uso de modelos estadísticos y herramientas de análisis de datos. La ciencia de datos permite identificar patrones y relaciones complejas entre las variables climáticas y las enfermedades respiratorias, lo que representa una oportunidad para transformar datos en conocimiento útil para la toma de decisiones. Además, el desarrollo de modelos predictivos basados en inteligencia artificial podría permitir la creación de sistemas de alerta temprana, capaces de identificar condiciones propicias para la propagación de enfermedades respiratorias.

En términos de relevancia global, este estudio se conecta con investigaciones internacionales que han analizado el impacto de factores climáticos en la salud humana, como el caso del ozono en Turquía y su influencia en la mortalidad (İbrahim & Özkan, 2024). Estos antecedentes destacan la importancia de evaluar cómo el clima afecta la salud respiratoria, aunque contextualizando los hallazgos a las particularidades climáticas, sociales y demográficas de Colombia.

A nivel educativo, este trabajo promueve el uso de enfoques interdisciplinarios, integrando conocimientos de climatología, epidemiología y análisis de datos. Esto no solo

enriquece la formación de estudiantes y profesores (Gilma C. et al., 2019), sino que también fortalece la capacidad de las instituciones educativas para abordar problemas reales mediante la investigación aplicada. Los resultados de este estudio pueden ser utilizados como base para desarrollar programas de prevención en escuelas, colegios y universidades, fomentando una conciencia colectiva sobre los efectos del clima en la salud.

Por último, este proyecto tiene un impacto directo en la población general, proporcionando información valiosa para los tomadores de decisiones en el ámbito de la salud pública. Identificar las variables climáticas que más afectan la incidencia de enfermedades respiratorias permitirá diseñar estrategias preventivas y de intervención más efectivas, alineadas con las necesidades locales. Además, contribuirá al desarrollo de políticas públicas informadas, dirigidas a mitigar los efectos adversos del cambio climático en la salud de la población colombiana. (Urrego-Sandoval, 2021)

Objetivos

Objetivo General

Evaluar la relación entre los patrones climáticos y las enfermedades respiratorias más comunes en Colombia durante el 2024, identificando correlaciones significativas a través del uso de modelos estadísticos.

Objetivos Específicos

Identificar las enfermedades respiratorias más comunes reportadas en Bogotá, Colombia durante el 2024, así como los datos climáticos disponibles para el mismo periodo.

Explicar la relación entre los patrones climáticos y las enfermedades respiratorias, describiendo cómo las variables climáticas influyen la incidencia de estas enfermedades.

Emplear modelos estadísticos analizando los datos climáticos y de salud, determinando las variables que tienen un mayor impacto en la prevalencia de enfermedades respiratorias en Colombia.

Marco Teórico

Este marco teórico pretende consolidar algunos conceptos que serán útiles para el desarrollo del trabajo de investigación, las herramientas asociadas a ciencia de datos utilizadas y los estudios previos que han impulsado y que apoyan el desarrollo del presente análisis.

Inicialmente se aborda el concepto de ciencia de datos con el fin de dar un esbozo a la línea de estudio de la especialización y se convierte en el marco conceptual, así mismo, el lenguaje de programación Python que es el eje principal con el que se realizan todos los modelos, cálculos y visualizaciones. Por otra parte, se abordan las descripciones de Pandas, Matplotlib, Scikit-learn, como principales librerías y herramientas que apalancan los modelos dentro de Python para el desarrollo del estudio. Se exponen los conceptos de Regresión lineal y Correlación como base estadística para el análisis de datos que permitirá aproximarnos a los resultados esperados. Por último, se complementa con algunos estudios y artículos similares que corroboran la importancia y pertinencia del desarrollo del presente estudio.

El primer concepto ampliado que utilizaremos y que es justamente la línea de estudio de la especialización y la cual será la base de todo el proyecto de investigación es la ciencia de datos que se compone de tres áreas. La primera es el Big Data, que se emplea para procesar los datos. La segunda es la minería de datos, cuya finalidad es encontrar patrones, incluso sin que estos fueran antes imaginados. Por último, la visualización de los datos, cuyo propósito es facilitar la comprensión de la información de manera clara y propiciar su socialización. cada uno de los componentes de la ciencia de datos se encuentra en continua evolución. Sin duda, este hecho influye en la forma en que se lleva a cabo la investigación científica e impacta en diferentes áreas del conocimiento (Lemus-Delgado & Pérez Navarro, 2023). Sobre este particular existen un sin

número de libros, artículos y publicaciones, lo que pretendemos es dar únicamente un concepto general de lo que será la base del estudio.

En ese sentido podemos comprender como el Big Data y la visualización de datos a partir de la información recolectada y las transformaciones aplicadas son fundamentales para un correcto entendimiento de la información, su interpretación y la posterior toma de decisiones, esto permite esquematizar que no basta con tomar datos aleatorios si no que él debe apalancarse con herramientas de visualización óptimas para los diferentes interlocutores.

El campo del Big Data y el análisis de datos ha ganado gran popularidad en los últimos años realizando importantes contribuciones tanto en el mundo académico como en el ámbito empresarial. La visualización de datos, a nivel mundial, inició con la construcción de tablas, gráficas y procesamiento de datos simples; y ha evolucionado significativamente hacia visualizaciones interactivas de mayor atracción y utilidad en términos de interpretación para los lectores con la introducción de herramientas sencillas como Canva e Infogram [1]. La visualización de datos se presenta como una forma sencilla, rápida y familiar de acceder a la información y posteriormente a la interpretación más eficiente de los resultados. El procesamiento y la manipulación de los datos se ha llevado a cabo por avances importantes en el desarrollo de sistemas computacionales. Como resultado, se ha generado una revolución en campos como las comunicaciones, la tecnología, la economía y la educación. (Grisales Aguirre et al., 2024)

Y se observa la integración entre todos los conceptos que se aplicaran y que son base para el presente estudio ya que todos se van encadenando y suministrando las herramientas necesarias y como esto puede tener un impacto social muy importante brindado soluciones reales a problemas del día a día de la sociedad.

Gracias al vertiginoso avance en las nuevas Tecnologías de la Información y la Comunicación (TIC) y al desarrollo de áreas de la ingeniería como las ciencias de la computación, la electrónica, la mecatrónica, entre otras, los dispositivos tecnológicos como laptops, teléfonos y relojes inteligentes, sensores y dispositivos wearables conectados a Internet implantando lo que hoy se conoce como IoT (Internet de las Cosas), son cada vez más accesibles a los ciudadanos, instituciones educativas, gobiernos y organizaciones del sector productivo. Se ha posibilitado, de cierta forma, algunas mejoras en la calidad de vida de la sociedad en general, permitiendo, por ejemplo: a) que el ciudadano tenga acceso inmediato a datos que contribuyen con su crecimiento personal e intelectual, b) que las instituciones educativas ofrezcan herramientas pedagógicas innovadoras para la labor de enseñanza-aprendizaje, c) que el gobierno tenga un control más eficiente de los recursos económicos, y que implemente y vigile políticas públicas que generen desarrollo, y d) finalmente, que el sector productivo, brindar servicios y productos que satisfagan las necesidades de los consumidores, además de la evaluación del rendimiento y la productividad de sus trabajadores.(Cardona et al., 2022)

Por otra parte, el lenguaje de programación utilizado para el desarrollo de todo el trabajo de investigación será Python el cual es un lenguaje de programación interpretado, interactivo y orientado a objetos. Incorpora módulos, excepciones, tipos de datos dinámicos de muy alto nivel y clases. Admite múltiples paradigmas de programación más allá de la programación orientada a objetos, como la programación procedimental y funcional. Python combina una potencia notable con una sintaxis muy clara. Tiene interfaces para muchas llamadas del sistema y bibliotecas, así como para varios sistemas de ventanas, y es extensible en C o C++. También se puede utilizar como lenguaje de extensión para aplicaciones que necesitan una interfaz programable. Por

último, Python es portátil: se ejecuta en muchas variantes de Unix, incluidas Linux y macOS, y en Windows. (Python Software Foundation, 2024)

Es por esto que se ha observado la importancia y relevancia que ha tomado este lenguaje de programación en los diferentes ámbitos y en la aplicación de la inteligencia artificial.

Python es un lenguaje de programación con muchas facilidades en el contexto de aprendizaje y mucha aplicabilidad en los entornos de infraestructura tecnológica, esto nos lleva a considerar una necesidad de verlo como un lenguaje que abrirá oportunidades laborales a los futuros egresados de la academia y a los profesionales activos en el ámbito del desarrollo de software. La flexibilidad multiplataforma de Python ha permitido crecer en soluciones de inteligencia artificial en los últimos años, y la integración de soluciones existentes, creación de nuevas soluciones para analizar datos o formas de automatizar actividades cotidianas, serán las bases de los nuevos patrones de evolución tecnológica relacionada al desarrollo de software. (Monroy Alfaro, 2022)

Así mismo, Pandas es una de las bibliotecas que se utilizarán para todo el desarrollo del trabajo de investigación, en el momento que se esté realizando la exploración y transformación de datos a través de Python. esta biblioteca es de código abierto con licencia BSD que proporciona estructuras de datos y herramientas de análisis de datos de alto rendimiento y fáciles de usar para el lenguaje de programación Python. Pandas es un paquete de Python que ofrece estructuras de datos rápidas, flexibles y expresivas diseñadas para que trabajar con datos “relacionales” o “etiquetados” sea fácil e intuitivo. Su objetivo es ser el componente básico de alto nivel para realizar análisis prácticos de datos del mundo real en Python. Además, tiene el objetivo más amplio de convertirse en la herramienta de análisis y manipulación de datos de código abierto más potente y flexible disponible en cualquier lenguaje. (NumFOCUS, 2024).

Para la generación de gráficos e imágenes se utilizará el paquete Matplotlib el cual es un paquete de gráficos 2D utilizado por Python para el desarrollo de aplicaciones, scripts interactivos y generación de imágenes con calidad de publicación en interfaces de usuario y sistemas operativos. (Juan D. Hunter, 2007)

Adicionalmente se espera utilizar Scikit-learn el cual es un módulo de Python que integra una amplia gama de algoritmos de aprendizaje automático de última generación para problemas supervisados y no supervisados de escala media. Este paquete se centra en acercar el aprendizaje automático a los no especialistas mediante un lenguaje de alto nivel de uso general. Se hace hincapié en la facilidad de uso, el rendimiento, la documentación y la coherencia de la API. Tiene dependencias mínimas y se distribuye bajo la licencia BSD simplificada, lo que fomenta su uso tanto en entornos académicos como comerciales. El código fuente, los binarios y la documentación se pueden descargar desde <http://scikit-learn.sourceforge.net> (Pedregosa et al., 2022).

Para el análisis de los datos y con el objetivo de identificar su relacionamiento se aplicarán conceptos de regresión lineal y correlación que permiten identificar el grado de influencia de una variable independiente en una variable de salida y de esta manera realizar predicciones al comportamiento de dicha variable. La correlación y regresión son conceptos fundamentales para el estudio de fenómenos en los que interesa analizar la relación entre dos o más variables (Engel y Sedlmeier, 2011). El razonamiento correlacional, la covariación y las relaciones entre datos multivariados son la base para el procesamiento de información que el individuo requiere en situaciones de incertidumbre (Gea, 2014). Por tanto, surge la necesidad de analizar datos provenientes de varias variables, situaciones que están presentes a lo largo de la

vida del ser humano, ya sea en situaciones cotidianas o profesionales. (Osorio Amaya et al., 2023)

Es así como se pueden ir consolidando las diferentes herramientas que serán útiles para el desarrollo del presente proyecto y todo esto confluye en la importancia de la inteligencia artificial en proyectos de investigación.

Como todas las tecnologías, la inteligencia artificial ha permeado y está cambiando tanto la educación como las formas de hacer investigación, fomentando la innovación y elevando la calidad y aplicación de los resultados enfocados hacia la solución de problemas sociales, ambientales, geopolíticos, entre otros. Igualmente, estas tecnologías aportan elementos importantes para la investigación científica, tales como procesamiento ágil y efectivo de grandes volúmenes de datos, identificación precisa de patrones y tendencias, predicciones mucho más exactas, establecimiento de correlaciones y tendencias que no son fácilmente identificadas a través de los métodos tradicionales de investigación; además, ha desarrollado diferentes herramientas específicas enfocadas en la investigación, por ejemplo, ChatGPT y muchas otras. (Díaz Subieta, 2024)

Y de esta manera podemos observar que los usos de la inteligencia artificial son incrementales y muy diversos, hasta este momento se ha evolucionado de gran manera y se seguirá creciendo ya que la inteligencia artificial apalancará ese mismo desarrollo, entre los usos más comunes actualmente están: búsqueda y optimización; machine learning y razonamiento probabilístico; redes neurales; visión por computadora; procesamiento de lenguaje natural y representación del conocimiento; planificación y toma de decisiones, y sistemas difusos.

Se cita el artículo “Effects of climate and environmental factors on childhood and adolescent asthma: A systematic review based on spatial and temporal analysis evidence” ya que

este artículo es uno de los más relacionados en el problema de estudio planteado, tiene todo un análisis de la relación entre las enfermedades respiratorias en niños y adolescentes generados por factores climáticos, si bien el estudio tiene una limitación y es que solo se enfocó en niños que utilicen inhaladores resultado de un diagnóstico de asma, si da una mirada y establece un marco teórico muy completo para la relación entre las enfermedades respiratorias y las variaciones del clima. Realiza una revisión y análisis de diferentes fuentes bibliográficas, teniendo en cuenta diferentes autores, diferentes regiones, periodos, condiciones climáticas, exposiciones y factores del entorno, así como el análisis espacial.

Esta revisión subraya la intrincada relación entre los factores climáticos y ambientales, como las temperaturas extremas, los espacios verdes y los contaminantes atmosféricos, con la exacerbación del asma infantil, influida por la edad y el sexo, y notablemente más grave en las estaciones frías. Los estudios anteriores se han centrado más en el análisis temporal que en el espacial, y carecen de una visualización exhaustiva de las zonas de riesgo de asma. Los estudios futuros deberían combinar análisis espaciales y temporales, sobre todo en zonas de alta prevalencia del asma, para ampliar nuestra comprensión de la variabilidad mundial del asma. Una futura dirección de investigación crítica implica comparar las incidencias del asma en diferentes zonas climáticas y los distintos efectos de las variables meteorológicas y la estacionalidad en la exposición a los desencadenantes del asma, como los aeroalérgenos, los agentes infecciosos y su efecto interactivo con el clima y los factores ambientales. La bibliografía actual carece de estudios sobre la incidencia del asma en distintas zonas climáticas, lo que constituye un área vital para futuras exploraciones. La investigación del asma en distintas zonas climáticas podría aclarar cómo contribuyen los aspectos ambientales al asma, con vistas a estrategias de prevención y tratamiento adaptadas a los climas locales. (Wang et al., 2024)

De esto lo que podemos observar es que efectivamente se requieren robustecer los estudios y análisis de las enfermedades respiratorias relacionadas a los cambios en las condiciones climáticas y de esta manera poder establecer políticas mejor alineadas y actividades que prevengan la propagación de enfermedades más críticas.

De igual forma traemos a colación el artículo “Influenza requiere un manejo bajo la perspectiva de "Una Salud" en Colombia” ya que este artículo es otro de los principales que apalancan el problema planteado, es una revisión específicamente de un tipo de enfermedad en relación con cambio climático, aporta un aspecto adicional y es como las virales que afectan a los humanos por supuesto también afectan a los animales, de una forma distinta pero también los afecta y lo que genera son mutaciones de los virus que posteriormente son transmitidas nuevamente a los humanos de formas más impactantes. En este artículo se habla nuevamente del concepto de One Health como la necesidad de entender la enfermedad de manera integral entendiendo todas las condiciones que la genera y las consecuencias que acarrea teniendo presente la relación entre la salud humana, la animal y el ambiente. Tiene una revisión exhaustiva de la influenza en Colombia la afectación a diferentes especies y la evolución que ha tenido el virus. De esta manera el artículo concluye que:

La influenza es un ejemplo claro de un problema que requiere una aproximación desde la perspectiva de Una Salud, lo cual implica tener una visión global antes que local, buscando la salud, bienestar, sostenibilidad y equilibrio entre los seres humanos, los animales y el medio ambiente. Con la participación de múltiples entidades y organizaciones a nivel mundial y la mayor aceptación del concepto de Una Salud, esta aproximación requiere involucrar de una manera más directa a expertos en áreas relacionadas no solo con la salud y el medio ambiente, sino también en ecología, ciencias sociales, biología, economía, epidemiología, política pública y

gobernanza, entre otras. En este contexto, Colombia cuenta con un Programa Nacional para la prevención, manejo y control de la infección respiratoria aguda en humanos, respaldado por el Ministerio de Salud y Protección Social; en este programa se hace énfasis particular a la influenza y los esfuerzos se dirigen primordialmente a la población colombiana. Por su parte, el Instituto Colombiano Agropecuario (ICA) encargado de la sanidad animal, desarrolla el proyecto de prevención y control de la influenza aviar, lo que ha permitido auto declarar al país como libre de influenza aviar, pero reconociendo la existencia de virus de influenza en cerdos. Se requiere entonces una articulación entre estas instituciones y los demás organismos responsables del manejo de la salud humana y animal con el fin de visualizar en un contexto real el desafío que representa la influenza en el país. (Uribe-Soto et al., 2020)

Esto que se plantea allí no solo sucede con la Influenza si no que pasa con todo tipo de enfermedades respiratorias y en definitiva se requiere un mayor análisis, un mayor despliegue y la toma de decisiones concretas que permitan mitigar la propagación de nuevas enfermedades.

En esta línea de artículos tenemos un estudio realizado en California, Estados Unidos, el cual utiliza un conjunto de datos recopilado pasivamente por sensores digitales que monitorean el uso de medicamentos inhalados y aplica análisis estadísticos utilizando modelos lineales mixtos y Machine Learning a partir de Random Forest. Este estudio identificó las asociaciones de tres contaminantes del aire, derivados de superficies de contaminación del aire dióxido de nitrógeno (NO₂), material particulado fino (PM 2.5) y ozono (O₃), en relación con síntomas de enfermedades respiratorias.

En primera medida este estudio nos permite corroborar la viabilidad del proyecto planteado para analizar el problema de estudio, ya que, si bien utiliza unos datos a partir de una medición en particular, si permite evidenciar la relación entre las condiciones ambientales y la

enfermedad respiratoria. Adicionalmente corrobora una variable adicional vista en uno de los anteriores artículos y es el aporte de la contaminación causada por los vehículos de motor a combustión a partir del uso de combustibles fósiles, porque con el aumento de los vehículos en las calles esta condición se hace más crítica.

Los resultados de este estudio son muy importantes y son la base para continuar sobre la línea planteada ya que concluye que:

En resumen, nuestro estudio confirmó que las exposiciones a corto plazo a la contaminación del aire están significativamente asociadas con síntomas respiratorios adversos, medidos por sensores de inhaladores digitales. Combinando superficies de alta exposición espaciotemporal, espacio de actividad durante todo el año y lugares de uso de medicación de rescate, pudimos identificar con precisión las exposiciones a la contaminación del aire espaciotemporal y, por lo tanto, estimar con mayor precisión el impacto de la contaminación del aire en los síntomas respiratorios. Además, este estudio pudo integrar tres tipos de contaminantes del aire en un único modelo de resultados de salud, identificando efectivamente los efectos marginales de cada contaminante en el uso de medicación de rescate. Creemos que los resultados de nuestro estudio fortalecen la evidencia de que incluso pequeñas fluctuaciones diarias en la contaminación del aire están asociadas con síntomas respiratorios diarios.

En definitiva, este artículo da grandes luces en como plantear y abordar el problema planteado entendiendo de forma ampliada las diferentes variables que afectan el modelo y la variable de salida, pero observando los beneficios de hacer este tipo de estudios para la salud pública que a la final es responsabilidad de todos. (Su et al., 2024)

Este es otro muy buen ejemplo del tipo de estudios de variables del clima que afectan la salud humana, en este caso es realizado en Corea del Sur y aborda la necesidad de realizar

pronósticos del oxígeno en la atmósfera y la calidad del aire para mitigar los riesgos de enfermedades respiratorias.

Este estudio utiliza redes neuronales para el análisis de los datos obtenidos. Se centra también en establecer que son necesarias acciones concretas que permitan robustecer el análisis del clima, las condiciones del aire y su afectación en la salud humana en pro de disminuir la enfermedad respiratoria y por ende las muertes asociadas.

En parte del planteamiento del estudio se estableció que: Para evaluar el impacto de varios contaminantes del aire en la salud, se utilizan múltiples enfoques y modelos dentro de las estadísticas epidemiológicas. Estas metodologías combinan datos de calidad del aire con factores epidemiológicos como el riesgo relativo, el riesgo de referencia y la fracción atribuible a diferentes niveles de concentraciones de contaminación, que generalmente culminan en análisis de la tasa de mortalidad (Oliveri Conti et al., 2017; Krzyzanowski et al., 2002). Entre las herramientas disponibles, el modelo AirQ+ es reconocido como una herramienta confiable para evaluar los impactos de los contaminantes del aire en la salud (Asl et al., 2018).

Este aspecto es fundamental por que corrobora la necesidad de evaluar las condiciones climáticas para identificar relaciones con enfermedades respiratorias comunes y que permita adicionalmente predecir a partir de unos insumos cuando podría estarse gestando una nueva afección viral, de esta manera podrían mitigarse muchas condiciones, aun cuando vimos en artículos anteriores que se requiere de un mayor involucramiento de otros actores por que se deben tomar decisiones de otra índole que no necesariamente dependen de la población en general.

Por lo tanto, el artículo es concluyente en que la contaminación del aire es un problema crítico de salud pública que afecta significativamente la salud y el bienestar. (Shams et al., 2024)

Este artículo permite complementar el anterior en una condición más cercana, ya que se centra en el análisis medioambiental de una ciudad de Colombia, la cual puede representar condiciones similares ya que comparte problemáticas que se presentan a lo largo del país. En este caso en particular no es un estudio técnico, con recopilación y análisis de datos, pero brinda un buen marco teórico para corroborar que las emisiones de dióxido de carbono generadas por el exceso de vehículos en las calles aumentan la contaminación que se ve retribuido en una mala condición del aire, por una parte, afecta el cambio climático y por otra es el cultivo perfecto para la generación de nuevas enfermedades respiratorias. De esto lo que podemos observar es la relación que veíamos anteriormente del ambiente, la salud humana y la animal, la necesidad de ver el problema de forma integral, analizar las causas que afectan al clima, en su mayoría por efectos generados por los seres humanos y las consecuencias que esto acarrea. Esto es lo que finalmente concluye el artículo.

Existe un ineficiente sistema de vigilancia de calidad de aire, excesiva dependencia de combustibles fósiles, las mayores concentraciones de contaminantes atmosféricos como los Óxidos de Nitrógeno (NOx) y Material Particulado (PM10 y PM2,5) se atribuyen al transporte público y son responsables de efectos negativos en la salud y el medioambiente, dentro de las partículas (PM2,5) aparece el Carbono Negro o “Black Carbon” (BC), siendo el principal responsable de enfermedades respiratorias y cardiovasculares, así como el aumento de temperaturas ya que es el segundo precursor del cambio climático y con un tercio del forzamiento radiactivo del Dióxido de Carbono (CO2), absorbe más radiación de la que se refleja, es bioacumulable y persistente en la atmósfera e interactúa con otros contaminantes primarios y secundarios produciendo una degradación de la calidad de aire.(Carbonell Chams, 2023)

Este artículo permite vislumbrar la importancia que tiene la revisión de los aspectos relacionados al cambio climático en relación con la salud humana, establece una revisión de que sucede en caso de que la temperatura suba 1° es decir, que sucede si la temperatura aumente 2.5°C en vez de 1.5°C y el impacto que esto tendría en todo el planeta y por supuesto directamente en la salud humana.

Se aborda el problema que eso implicaría a nivel de descongelamiento de los polos y el aumento en los niveles del agua, el incremento en la temperatura de los océanos y cómo afectaría los arrecifes de coral, la generación de nuevos y más potentes huracanes, tifones y tormentas, el incremento de los incendios forestales y para los humanos y animales causan agotamiento por calor, insolación, hipertermia y deshidratación que podría conllevar la muerte.

Se establece también que las consecuencias derivadas del cambio climático afectarían de forma desproporcionada afectando en mayor medida a mujeres embarazadas, niños, ancianos, personas en condición de discapacidad, minorías y personas en condición de pobreza.

De esta manera este artículo se relaciona directamente con el problema en cuestión ya que si bien no es un estudio directamente de las enfermedades respiratorias si es un primer esbozo de la problemática que genera el cambio climático y todos los impactos que puede tener, es por esto que el autor concluye el artículo de reflexión con el siguiente párrafo:

Luego de varios meses de detectado el primer caso, aún persisten más preguntas que respuestas. Si bien el COVID-19 ha cambiado la vida de millones de personas, no va a terminar con la humanidad, pero se debe reaccionar y tomar como un llamado de alarma y atención que incentive a reflexionar sobre cómo nuestras acciones afectan el ecosistema en el que vivimos. Seguramente esta no será la última pandemia y vendrán otras enfermedades, pero si no se aprende la lección y se continúa teniendo el mismo impacto negativo en la naturaleza, tal vez en

el futuro el planeta no sea tan benevolente con nuestra especie como lo ha sido hasta ahora.

(Petrone & Petrone, 2020)

Adicionalmente se contempla el artículo “Determination of Ozone induced mortality in Aydın province, Turkey in 2022 using AIRQ+ Modeling” en este artículo, si bien no es en Colombia, sirve de guía para el análisis de variables relacionadas al clima, incluye el análisis de unas condiciones medioambientales de una población en Turquía y su relación con la mortalidad, este tipo de estudios permiten evaluar la forma de esquematizar análisis de condiciones climáticas en relación a la afectación humana, y este artículo tiene algo en particular adicional de gran valor y es la implementación de modelos de recolección y análisis de datos, ya que utiliza instrumentos de medición de las condiciones de oxígeno para luego trabajarlas con herramientas estadísticas para evaluar su impacto en una variable de salida, en este caso la tasa de mortalidad.

El estudio concluye que:

Cuando se examinan los resultados de los datos obtenidos en nuestro estudio y se tienen en cuenta los estudios anteriores sobre contaminantes atmosféricos, se observa un paralelismo entre los resultados. Los niveles de ozono a nivel del suelo provocan problemas cardiovasculares y respiratorios y aumentan los ingresos hospitalarios. El efecto a largo plazo del ozono sobre la mortalidad es innegable. Además, se observa que el nivel de ozono en el aire varía a lo largo del día, según las estaciones, los días laborables y los fines de semana. En nuestro estudio, sólo se ha estudiado el efecto del ozono sobre la mortalidad, y el número de estudios realizados hasta ahora en este ámbito es muy bajo. Creemos que nuestro estudio inspirará otros estudios a este respecto.

(İbrahim & Özkan, 2024)

Este artículo da un muy buen punto de partida para continuar realizando análisis de variables medioambientales que afectan a la salud humana, si bien las condiciones del clima en

la zona del estudio serán diferentes a las de Colombia, no se puede desconocer el resultado de cómo afecta el clima en la salud en cualquier parte del mundo, en algunos lugares en mayor o menor proporción, pero las afectaciones son indudables.

En Colombia, la evaluación de la calidad del aire se encuentra regulada por la Resolución 2254 del 1 de noviembre de 2017, expedida por el Ministerio de Ambiente y Desarrollo Sostenible, la cual establece los estándares de calidad del aire y los valores máximos permisibles para contaminantes como PM₁₀, PM_{2.5}, O₃, NO₂, SO₂ y CO. Esta normativa define los métodos oficiales de medición, los promedios temporales exigidos y los criterios para la operación de los Sistemas de Vigilancia de Calidad del Aire (SVCA), constituyéndose en el marco técnico que guía la medición y reporte de la contaminación atmosférica en el país (Ministerio de Ambiente y Desarrollo Sostenible, 2017). Su incorporación en este marco teórico asegura coherencia metodológica y el uso de criterios oficiales para el cálculo del Índice de Calidad del Aire (ICA).

Asimismo, representa un pilar normativo para interpretar la exposición de la población a contaminantes atmosféricos y sus posibles efectos sobre la salud, al establecer valores guía alineados con estándares internacionales y con evidencia científica consolidada. Esta normativa permite contextualizar los niveles de riesgo asociados a cada categoría del ICA y facilita relacionar los niveles de contaminación con estudios epidemiológicos sobre enfermedades respiratorias, cardiovasculares u otros impactos en la salud (Ministerio de Ambiente y Desarrollo Sostenible, 2017). Por ello, su integración en el marco teórico fortalece la comprensión del vínculo entre calidad del aire y salud pública desde una perspectiva regulatoria y técnica.

Por último, se tuvo en cuenta los conceptos que abarcan La causalidad de Granger ya que este constituye uno de los enfoques más reconocidos para evaluar relaciones dinámicas entre variables temporales y determinar si la información pasada de una serie contribuye

significativamente a mejorar la predicción de otra. Este método parte del principio de que una variable X “Granger-causa” a una variable Y cuando los valores rezagados de X contienen información estadísticamente relevante para anticipar el comportamiento futuro de Y, por encima de lo que podría lograrse únicamente con los valores históricos de Y. El fundamento teórico de este enfoque fue establecido por Clive W. J. Granger en su artículo seminal *Investigating Causal Relations by Econometric Models and Cross-spectral Methods*, en el cual demuestra que la causalidad, entendida en un sentido predictivo, puede evaluarse mediante modelos econométricos que comparan la capacidad explicativa de distintas especificaciones con y sin los rezagos de la variable potencialmente causal (Granger, 1969). Este concepto ha sido ampliamente adoptado en estudios de econometría, ciencias sociales y funciona para el análisis ambiental debido a su solidez metodológica y a su aplicabilidad en contextos donde las relaciones no pueden observarse de forma experimental.

En el contexto del presente trabajo, la inclusión de la prueba de causalidad de Granger permite evaluar si las variaciones del Índice de Calidad del Aire (ICA) anteceden y explican de manera significativa los cambios en los casos de enfermedades respiratorias reportados en Colombia. Dado que el estudio analiza series temporales mensuales, este modelo resulta especialmente adecuado para identificar patrones predictivos o dependencias dinámicas entre las condiciones ambientales y los indicadores de salud pública. Su aplicación contribuye a fortalecer la rigurosidad del análisis, al permitir no solo observar correlaciones, sino también explorar la direccionalidad temporal de la relación. En consecuencia, este enfoque complementa los análisis descriptivos y correlacionales realizados, proporcionando un soporte metodológico robusto para evaluar si la calidad del aire puede considerarse un determinante anticipado de las afectaciones en salud respiratoria.

Metodología

El estudio será de tipo descriptivo-correlacional, orientado a identificar patrones y relaciones entre variables climáticas y la incidencia de enfermedades respiratorias en Colombia durante el año 2024. Este enfoque permite explorar la asociación entre las condiciones climáticas y las enfermedades sin intervenir directamente en las variables. Para esta identificación y entendimiento el método utilizado será KDD (Knowledge Discovery in Databases), que permite obtener información a partir de la exploración de las bases de datos, su relacionamiento y comprensión del problema.

Para la recolección de datos se contempla:

- Datos climáticos: Registros climáticos proporcionados por el Instituto de Hidrología, Meteorología y Estudios Ambientales (IDEAM), incluyendo variables como temperatura, humedad, precipitación y calidad del aire. Estos datos serán obtenidos por regiones del país.
- Datos de salud: Registros del Ministerio de Salud y Protección Social, y de bases de datos hospitalarias (información pública a las que sea posible acceder), que incluirán información sobre la prevalencia e incidencia de enfermedades respiratorias, tales como infecciones respiratorias agudas (IRA), asma y neumonía.
- Período de análisis: Los datos corresponderán al año 2024.

El muestreo será de tipo no probabilístico por conveniencia, considerando regiones representativas de diferentes climas en Colombia (Bogotá, Medellín, Cali, Barranquilla). Este enfoque permitirá capturar la variabilidad geográfica del país.

Para el procesamiento de datos se realizará una depuración de las bases de datos para corregir valores atípicos, datos faltantes y posibles inconsistencias. Los datos serán

transformados, si es necesario, para asegurar la comparabilidad entre variables con diferentes unidades de medida.

Se utilizarán estadísticos descriptivos y visualizaciones (e.g., diagramas de dispersión, mapas de calor) para identificar tendencias preliminares entre variables. Se emplearán coeficientes de correlación de Pearson para determinar la fuerza y dirección de las relaciones entre las variables climáticas y la incidencia de enfermedades respiratorias. Se aplicará un modelo de regresión lineal para evaluar el impacto de las variables climáticas (e.g., temperatura, humedad, precipitación) en las enfermedades respiratorias, identificando aquellas variables con mayor significancia estadística.

Todos los análisis se realizarán utilizando Python. Se emplearán bibliotecas específicas de ciencia de datos (Pandas, Scikit-learn, Matplotlib) para procesar, analizar y modelar los datos. Para garantizar la robustez del análisis, se realizarán pruebas de diagnóstico sobre los modelos estadísticos, como la colinealidad entre variables, normalidad de los residuos y ajuste.

El trabajo de investigación podría enfrentarse a limitaciones como la disponibilidad incompleta de datos en algunas regiones o la precisión de los registros climáticos y de salud. Estas limitaciones serán abordadas con técnicas de imputación de datos para evaluar la estabilidad de los resultados.

Resultados – Desarrollo de los Modelos

Para el presente estudio que integra datos ambientales y de salud, provenientes de múltiples fuentes institucionales (IDEAM, SIVIGILA), la heterogeneidad en los formatos de registro por ejemplo, diferencias en codificación, nombres de municipios o unidades de medida puede generar sesgos analíticos y errores de interpretación (Pérez & González, 2023).

Por ello, la limpieza y estandarización constituyen una etapa crítica en el proceso de preparación de datos para la analítica avanzada, conocida como *data wrangling* (Wickham, 2014).

Por lo tanto, se abordaron diferentes etapas:

1. Normalización de nombres y formatos: homologar campos como “departamento”, “municipio” y “año” para permitir su integración.
2. Conversión de tipos de datos: asegurar que fechas y valores numéricos estén en formato interpretable por los algoritmos estadísticos.
3. Eliminación de duplicados y valores faltantes: manteniendo el principio de representatividad sin distorsionar las series originales.
4. Exploración de datos: Se aplica estadística descriptiva para el entendimiento de los conjuntos de datos, variables críticas y entendimiento de la información.
5. Análisis de correlación
6. Regresión lineal múltiple
7. Causalidad de Granger

Tabla 1*Bases de Datos Objeto de Análisis*

Base de datos	Fuente	Propósito dentro del modelo analítico
Calidad del aire	IDEAM / SINA	Contiene indicadores anuales de concentración de contaminantes (PM10, PM2.5, NO ₂ , SO ₂ , CO, O ₃), necesarios para evaluar variaciones en el índice de la calidad del aire
Datos salud 2020 - 2024	SIVIGILA	Casos confirmados de Infección Respiratoria Aguda Grave (IRAG) entre 2020 y 2024 con variables demográficas, clínicas y geográficas. Permite análisis recientes y comparaciones con promedios históricos.

Nota. Fuentes utilizadas para el desarrollo del estudio

Se realizó la exploración de datos de las bases de datos objeto de estudio de forma independiente y posteriormente se realizó un análisis al data frame unificado, de la siguiente manera:

Datos de Salud Unificado Entre 2020 y 2024

Se extrajeron las variables críticas para el estudio y se identificó la distribución de datos para cada variable crítica.

VARIABLES CRÍTICAS:

- Año: Representa el año donde se presentó el evento de salud
- Departamento ocurrencia: Describe el departamento donde se presentó el evento de salud.
- Municipio ocurrencia: Describe el municipio donde se presentó el evento de salud.

- Nom_est_f_caso: Describe si el caso fue Descartado, Confirmado por laboratorio o Probable
 - Sexo: Se refiere al género de la persona, clasificado como M: Masculino y F: Femenino
 - Edad: Se refiere a la edad del paciente que notifica el evento de salud
 - Fec Not: Describe la fecha de notificación del evento de salud.
 - Confirmados: Es una variable categórica que toma valor 1 para los casos Confirmado por laboratorio y 0 para los casos en estado Descartado o Probable.
- a. Distribución de los casos por año.

Tabla 2*Registros de Salud por Año*

Año	Cantidad de Registros
2022	107.290
2023	72.558
2020	21.904
2024	1.083
2021	718

Nota. Datos obtenidos de la Base de Salud 2020 – 2024 Sivigila

Observamos que el año con mayor cantidad de registros fue el 2022 con 107.290 que representa el 52.7% del total de las observaciones y presenta una gran diferencia con los demás años.

- b. Distribución de casos por municipio:

Tabla 3*Cantidad de Casos por Municipio de Ocurrencia*

Municipio	Cantidad de Registros
Bogotá	71.088
Cali	13.950
Medellín	13.931
Barranquilla	7.048
Soacha	4.749
Villavicencio	3.706
Arauca	3.376
Manizales	2.290
Ocaña	2.235
Santa Marta	2.230

Nota. Datos obtenidos de la Base de Salud 2020 – 2024 Sivigila

El municipio donde más se presentan casos es en Bogotá con 71.088 registros que representa el 57.1% de los municipios más representativos de los casos presentados durante los años de estudio.

c. Casos confirmados por estado

Tabla 4*Cantidad de Casos por Estado*

Estado Caso	Cantidad de Registros
Descartado	137.589
Confirmado por laboratorio	51.078
Probable	14.886

Nota. Datos obtenidos de la Base de Salud 2020 – 2024 Sivigila

Se evidencia que los casos confirmados por laboratorio que podríamos decir son positivos son 51.078 que representan un 25.1% de la totalidad de los casos registrados, lo cual es un porcentaje alto e indica que se invierten muchos esfuerzos en identificar realmente los casos positivos.

d. Total, de casos por genero distribuidos entre Masculino y Femenino

Tabla 5

Cantidad de Casos por Genero

Sexo	Cantidad de Registros
Masculino	107.212
Femenino	96.341

Nota. Datos obtenidos de la Base de Salud 2020 – 2024 Sivigila

De los casos reportados se observa que el 52,6% son reportados por el género masculino el 47.3% restante es reportado por el género femenino.

e. Casos por municipio y por estado

Tabla 6

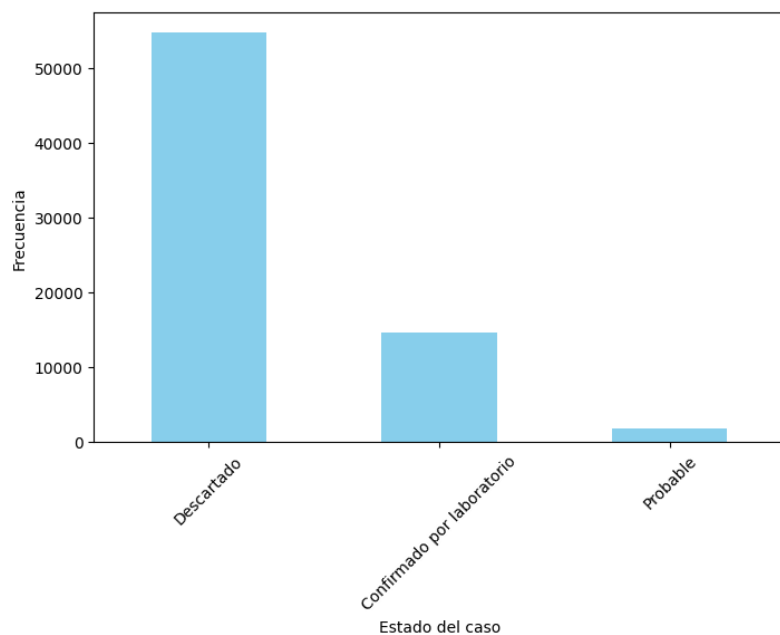
Cantidad de Casos por Municipio por Estado

Estado	Confirmado por Laboratorio	Descartado	Probable
Departamento Ocurrencia			
Bogotá	14644	54651	1793
Antioquia	6337	15948	1747
Valle	5480	11377	2173

Estado	Confirmado por Laboratorio	Descartado	Probable
Departamento Ocurrencia			
Atlántico	3707	6812	976
Cundinamarca	2684	9634	661
Cesar	1611	2194	637
Córdoba	1396	588	288
Santander	1265	354	225
Norte Santander	1145	5222	386
Cauca	1120	2198	400

Nota: Datos obtenidos de la Base de Salud 2020 – 2024 Sivigila

Observamos la relación entre el municipio de ocurrencia y el estado de los casos, donde identificamos por ejemplo que, en Bogotá, de los 71.088 casos reportados el 20,5% son confirmados. De igual manera

Figura 1*Distribución de Casos por Estado en Bogotá*

f. Casos por estado por rango de edades

Se realizó una agrupación por rango de edades cada 10 años filtrando el municipio de ocurrencia para Bogotá.

Tabla 7*Cantidad de Casos por Rango de Edad en Bogotá*

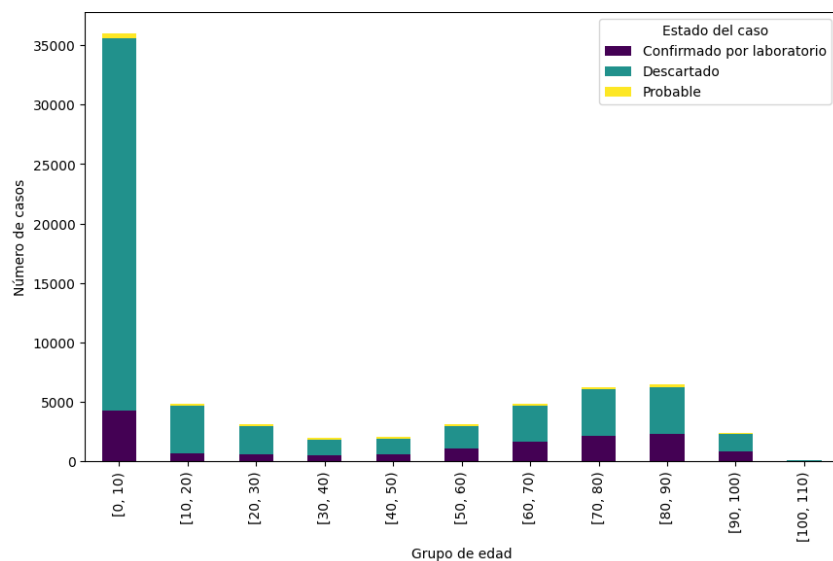
Estado	Confirmado por Laboratorio	Descartado	Probable
Grupo Edad			
[0, 10)	4259	31352	367
[10, 20)	639	4052	111
[20, 30)	600	2363	143

Estado	Confirmado por Laboratorio	Descartado	Probable
Grupo Edad			
[30, 40)	529	1281	144
[40, 50)	607	1319	128
[50, 60)	1044	1881	203
[60, 70)	1654	3013	207
[70, 80)	2139	3891	198
[80, 90)	2301	3960	236
[90, 100)	841	1495	53
[100, 110)	31	44	3

Nota. Datos obtenidos de la Base de Salud 2020 – 2024 Sivigila

Figura 2

Distribución de Casos por Grupo de Edad y Estado en Bogotá

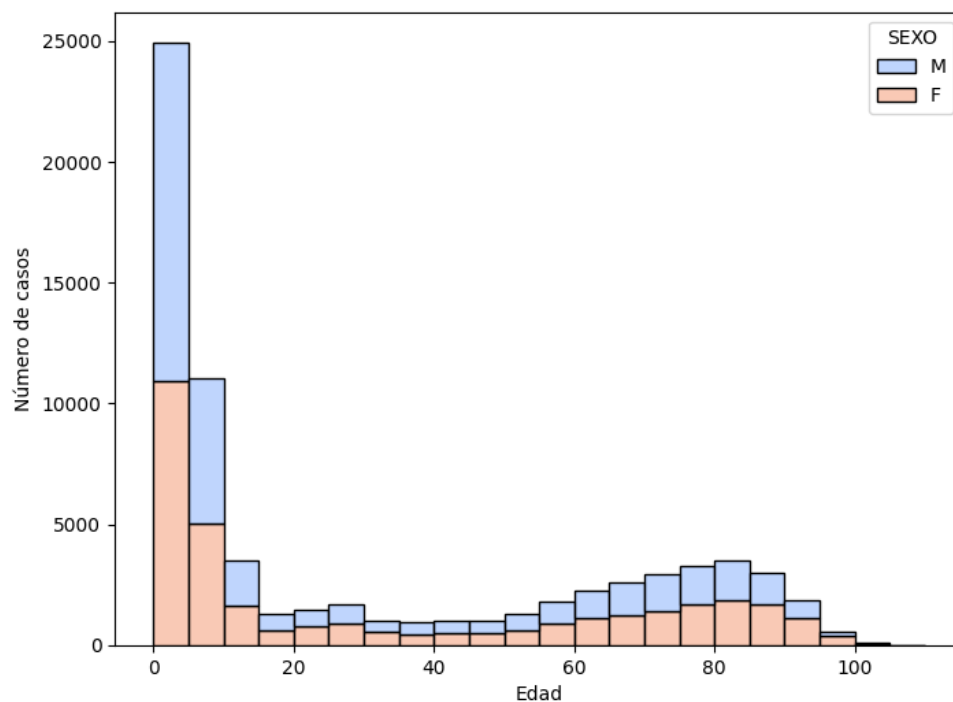


De esto podemos observar que niños entre 0 y 10 años representan el 29.1% de los casos confirmados, pero así mismo representan el 57.3% de los casos descartados, lo que indica que se presentan muchos niños con síntomas que requieren exámenes médicos y que finalmente se descartan de tener infección respiratoria. Adicionalmente permite corroborar que son los niños los principales afectados en este tipo de afecciones. De igual manera, se observa un aumento en los casos confirmados en los grupos entre los 60 y los 90 que también representan un grupo vulnerable de personas para la confirmación de casos de infección respiratoria.

Así mismo podemos observar la distribución de casos en Bogotá por rango de edad y por género, donde se evidencia la distribución entre Hombres y Mujeres.

Figura 3

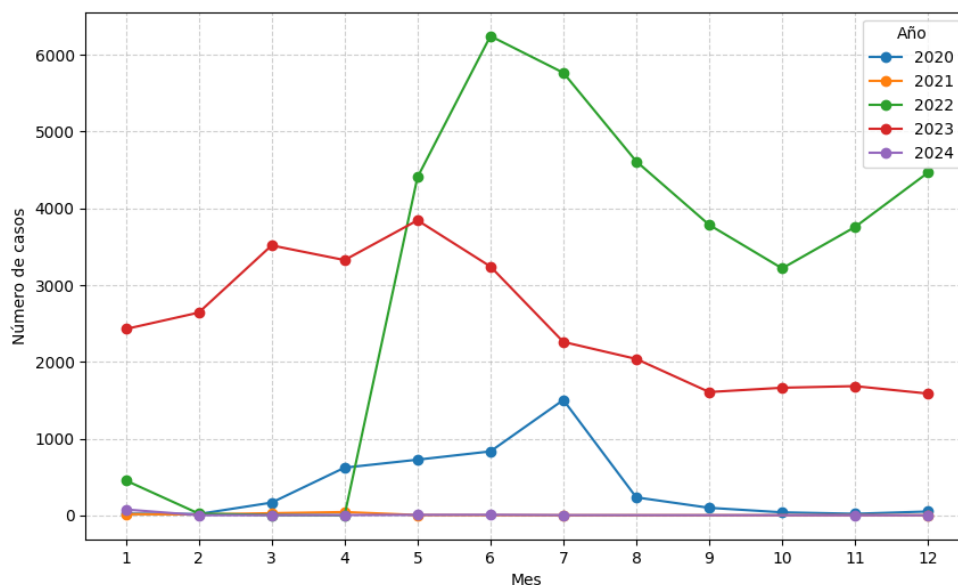
Distribución de Edades por Genero en Bogotá



g. Tendencia de reporte de casos por año en Bogotá

Figura 4

Tendencia Mensual de Casos por Año en Bogotá



Con este grafico podemos observar el aumento de los casos que hubo en 2022 después del mes de mayo y la tendencia descendiente que hubo en 2023, los años 2021 y 2024 no son representativos en comparación con la tendencia de los demás años.

Calidad del Aire entre 2020 y 2024

Este dataset contiene la medición de todas las variables críticas en las diferentes estaciones por departamento y municipio en diferentes años. Para realizar la exploración de estos datos se estableció un comparativo entre los puntos de corte para cada variable establecidos por el Ministerio de Ambiente y Protección social en donde se establece cuando una variable está en un determinado nivel, para esto se creó una matriz con los puntos de corte para cada variable en cada nivel. Adicionalmente se realizó la medición del Índice de Calidad del Aire con la fórmula que tiene establecido el Ministerio de Ambiente, esto permitió realizar los diferentes análisis y comparativos.

Tabla 8*Puntos de Corte del ICA (Índice de calidad del Aire)*

Índice de Calidad del Aire			Puntos de corte del ICA						
ICA	Color	Categoria	PM ₁₀	PM _{2.5}	CO	SO ₂	NO ₂	O ₃	O ₃
			μg/m ³ 24 Horas	μg/m ³ 24 Horas	μg/m ³ 8 Horas	μg/m ³ 1 Hora	μg/m ³ 1 Hora	μg/m ³ 8 Horas	μg/m ³ 1 Hora
0-50	Verde	Buena	0 – 54	0 – 12	0 – 5094	0 – 93	0 – 100	0 – 106
51- 100	Amarillo	Aceptable	55 - 154	13 – 37	5095 – 10819	94 – 197	101 – 189	107 – 138
101- 150	Naranja	Dañina ala salud de Grupos Sensibles	155 - 254	38 – 55	10820 – 14254	198 – 486	190 – 677	139 – 167	245 – 323
151- 200	Rojo	Dañina a la salud	255 - 354	56 – 150	14255 – 17688	487 – 797	678 – 1221	168 – 207	324 – 401
201- 300	Púrpura	Muy dañina a la salud	355 – 424	151 – 250	17689 – 34862	798 – 1583	1222 – 2349	208 - 393	402 – 794
301- 500	Marrón	Peligrosa	425 - 604	251 - 500	34863 - 57703	1584 - 2629	2350 - 3853	394	795 - 1185

Nota. Información extraída de la Resolución 2254 del 01 de Noviembre de 2017 del Ministerio de Ambiente y Desarrollo Sostenible

h. Matriz de puntos de corte llevada al modelo para realizar el comparativo contra las mediciones reales relacionadas en el dataset.

Tabla 9

Puntos de Corte ICA llevados al modelo

	Variable	Exposure_hours	Conc_low	Conc_high	Index_low	Index_high	Category
0	PM2.5	24	0.0	12	0	50	Verde
1	PM2.5	24	13.0	37	51	100	Amarillo
2	PM2.5	24	38.0	55	101	150	Naranja
3	PM2.5	24	56.0	150	151	200	Rojo
4	PM2.5	24	151.0	250	201	300	Purpura
5	PM2.5	24	251.0	500	301	500	Marron
6	PM10	24	0.0	54	0	50	Verde
7	PM10	24	55.0	154	51	100	Amarillo
8	PM10	24	155.0	254	101	150	Naranja
9	PM10	24	255.0	354	151	200	Rojo
10	PM10	24	355.0	424	201	300	Purpura
11	PM10	24	425.0	604	301	500	Marron
12	O3	8	0.0	106	0	50	Verde

	Variable	Exposure_hours	Conc_low	Conc_high	Index_low	Index_high	Category
13	O3	8	107.0	138	51	100	Amarillo
14	O3	8	139.0	167	101	150	Naranja
15	O3	8	168.0	207	151	200	Rojo
16	O3	8	208.0	393	201	300	Purpura
17	NO2	1	0.0	100	0	50	Verde
18	NO2	1	101.0	189	51	100	Amarillo
19	NO2	1	190.0	677	101	150	Naranja
20	NO2	1	678.0	1221	151	200	Rojo
21	NO2	1	1222.0	2349	201	300	Purpura
22	NO2	1	2350.0	3853	301	500	Marron
23	CO	8	0.0	5094	0	50	Verde
24	CO	8	5095.0	10819	51	100	Amarillo
25	CO	8	10820.0	14254	101	150	Naranja
26	CO	8	14255.0	17688	151	200	Rojo
27	CO	8	17689.0	34862	201	300	Purpura
28	CO	8	34863.0	57703	301	500	Marron

	Variable	Exposure_hours	Conc_low	Conc_high	Index_low	Index_high	Category
29	SO2	1	0.0	93	0	50	Verde
30	SO2	1	94.0	197	51	100	Amarillo
31	SO2	1	198.0	486	101	150	Naranja
32	SO2	1	487.0	797	151	200	Rojo
33	SO2	1	798.0	1583	201	300	Purpura
34	SO2	1	1584.0	2629	301	500	Marron

Nota. Datos obtenidos de la Resolución 2254 del 01 de Noviembre de 2017 del Ministerio de Ambiente y Desarrollo Sostenible

Se estableció una función dentro del modelo para realizar la validación de en qué nivel se encontraba cada medición de acuerdo a la matriz de puntos de corte, posteriormente se realizó una interpolación por estación y por municipio. El resultado de toda la transformación fue llevada a un nuevo archivo para posteriormente este ser tratado y analizado. Como resultado de ese cálculo se obtuvo:

- i. Cantidad de mediciones por nivel

Tabla 10

Cantidad de Mediciones por Nivel

Clasificación	Cantidad de Registros
None	14934

Clasificación	Cantidad de Registros
Purpura	3546
Rojo	3188
Marron	2059
Naranja	1932
Verde	1633

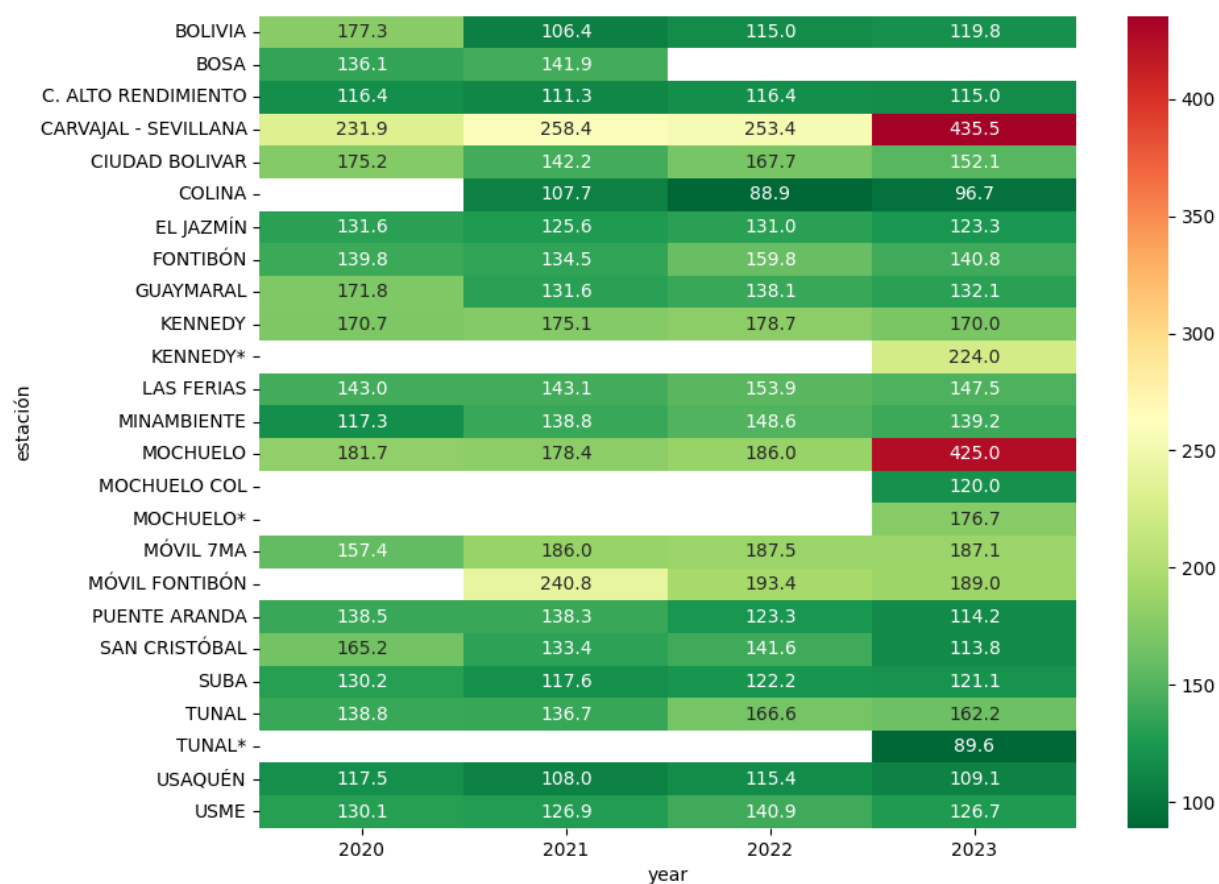
Nota. Datos obtenidos de la Base de Calidad del Aire, IDEAM

De lo anterior podemos observar que de 28.732 registros el 51.9% no puede ser categorizado y esto se debe principalmente es a que son variables que son medidas pero que no son tenidas en cuenta en el cálculo del Índice de Calidad del Aire, y que por lo tanto no están en la matriz de Puntos de Corte y son excluidas. Por otra parte, el 30.6% de los casos están en los niveles más altos entre Purpura, Marron y Rojo.

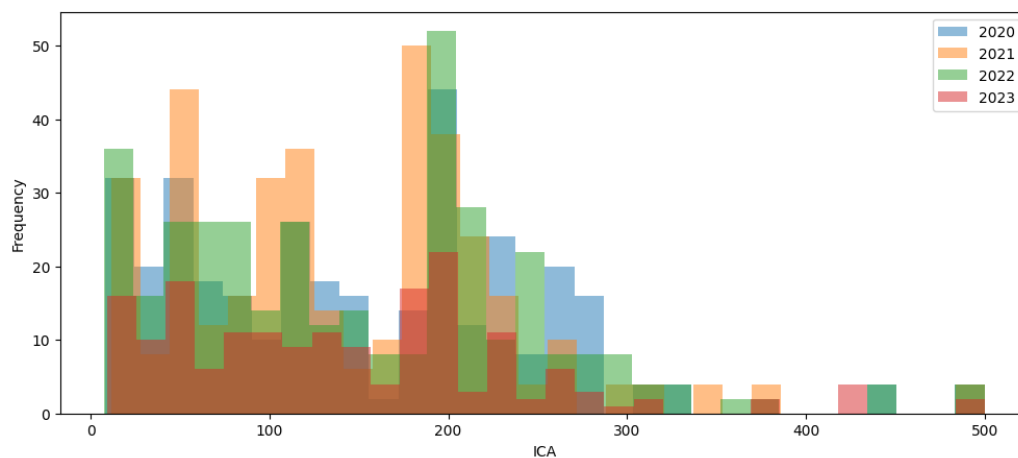
j. Al dataset original se realizó un filtro para identificar solo el municipio de Bogotá y se realizo una distribución de los registros de medición en las diferentes estaciones, teniendo en cuenta el valor ICA promedio.

Figura 5

Mapa de Calor del ICA Promedio por Año por Estación en Bogotá



k. Frecuencia de medición del ICA por año

Figura 6*Distribución de Frecuencias del ICA por Año*

Podemos observar que hay unos picos en 2021 y 2022 que se situaron cercanos a 200 pero que en todos los años se observan algunos momentos donde el Índice de Calidad del aire sube por encima de 300 si bien no son la frecuencia más alta se presentan picos altos como sucedió en 2022 y 2023.

1. Relación de la cantidad de mediciones en cada categoría por año.

Tabla 11*Relación de Cantidad de Medición por Categoría por Año*

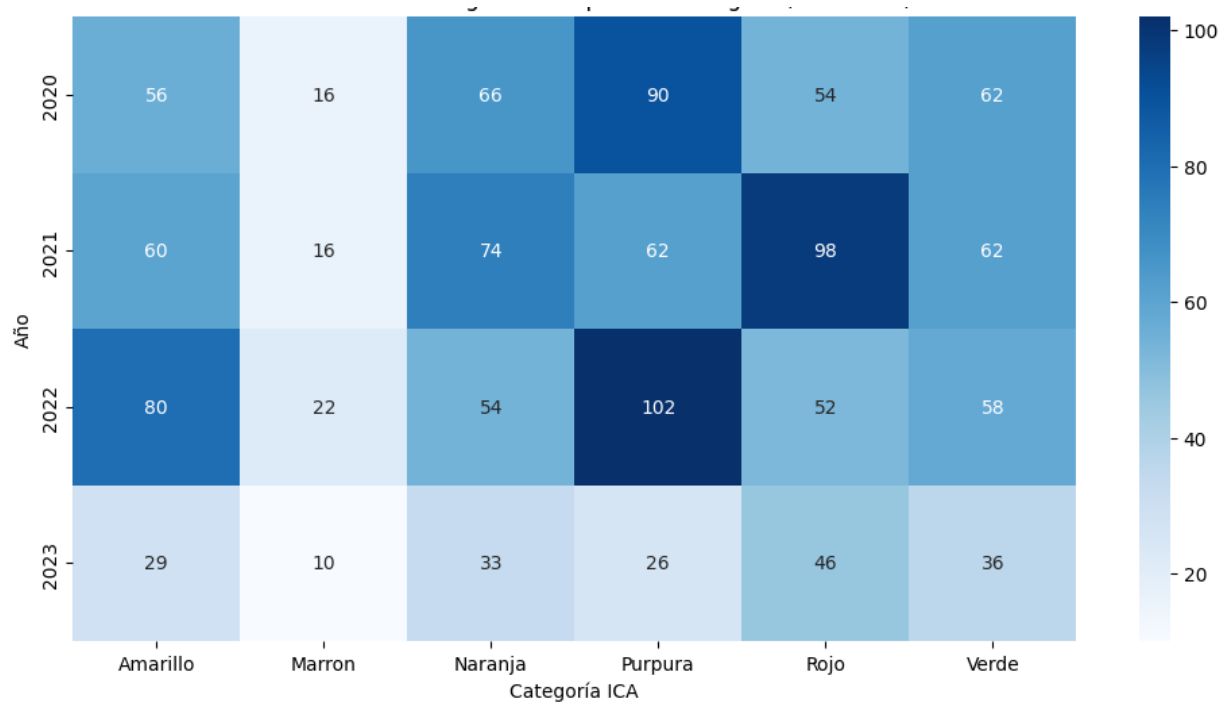
ICA_category	Amarillo	Marron	Naranja	Purpura	Rojo	Verde
year						
2020	56	16	66	90	54	62
2021	60	16	74	62	98	62
2022	80	22	54	102	52	58

ICA_category	Amarillo	Marron	Naranja	Purpura	Rojo	Verde
year						
2023	29	10	33	26	46	36

Nota. Datos obtenidos de la Base de Calidad del Aire, IDEAM

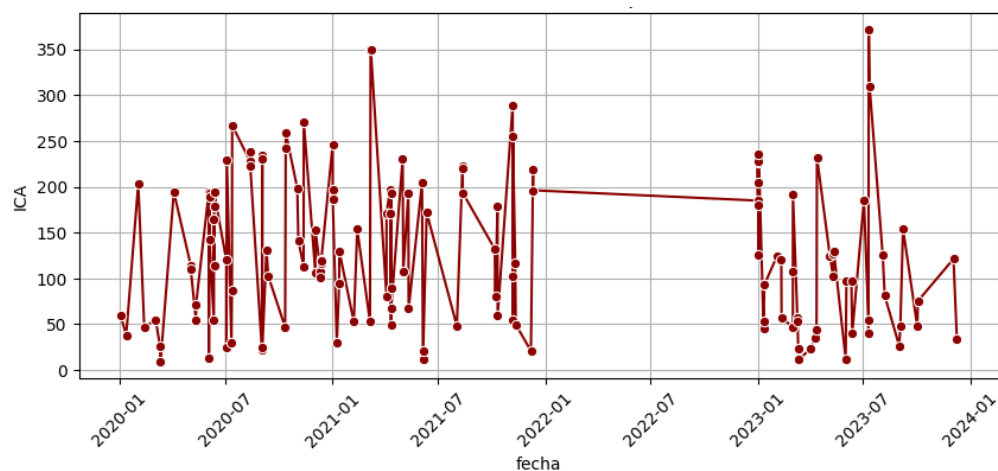
Figura 7

Mapa de Calor de Clasificación de Categorías por Año



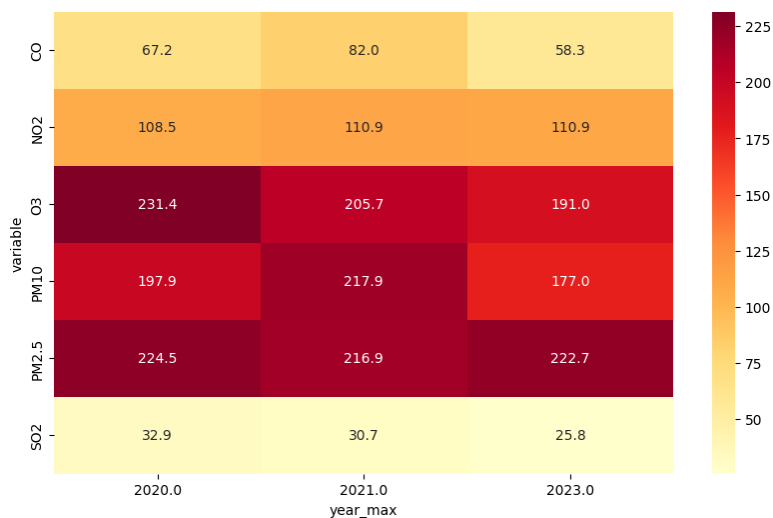
Podemos corroborar que en el año 2021 y 2022 es donde se presenta la mayor cantidad de mediciones en categoría Purpura y Rojo, esto coincide parcialmente con el pico de casos que habíamos observado en el Dataset de salud, sin constituirse como una conclusión definitiva.

m. Tendencia mensual del ICA promedio

Figura 8*Tendencia Mensual del ICA Promedio*

Con esta tendencia podemos observar la gran variabilidad que hay en las mediciones del ICA, no se observa una tendencia específica en ninguno de los años, se observan unos picos importantes en 2021 y 2023 pero no una tendencia clara.

n. ICA promedio por variable contaminante por año:

Figura 9*ICA Promedio por Contaminante por Año*

Se observa que al calcular el ICA promedio a partir de la formula establecida por el ministerio para cada una de las variables el PM2.5, el PM1.0 y el O3 son los contaminantes con los mayores valores siendo estas las variables más representativas en el Índice Final en los 3 años.

Unificación de los Dataframe de Salud y Calidad del Aire

Teniendo en cuenta la temporalidad de las mediciones tanto en salud como en calidad del aire, se construyó una variable llamada periodo y se realizó el merge entre ambas bases de datos, a partir de eso se realizó el análisis de Correlación, Regresión Lineal y Causalidad de Granger. Con este nuevo dataframe se pudo unir los Casos de Salud vs las variables y el índice de calidad del aire para Bogotá y se retiraron los registros que no tuvieran medición del ICA. De esto en total quedaron 517 registros para los diferentes modelos que se van a evaluar. Se garantizo que este dataframe quedara limpio de valores nulos, vacíos y que cada variable tuviera los tipos requeridos.

Figura 10

Muestra del Dataframe Unido Entre Calidad del Aire y Salud

	variable	year	ICA	ICA_matched	ICA_category	fechas/horas del máximo	municipio	fecha	year_max	month_max	Periodo	Casos_Salud	Fecha
0	CO	2020	54.252970	True	Amarillo	2020-06-11 01:00:00	BOGOTA	2020-06-11 01:00:00	2020.0	6.0	2020-06	833.0	2020-06-01
1	SO2	2020	30.107527	True	Verde	2020-03-10 12:00:00	BOGOTA	2020-03-10 12:00:00	2020.0	3.0	2020-03	166.0	2020-03-01
2	SO2	2020	30.107527	True	Verde	2020-07-11 04:00:00	BOGOTA	2020-07-11 04:00:00	2020.0	7.0	2020-07	1503.0	2020-07-01
3	PM2.5	2020	183.840426	True	Rojo	2020-06-11 11:00:00	BOGOTA	2020-06-11 11:00:00	2020.0	6.0	2020-06	833.0	2020-06-01
4	SO2	2020	9.139785	True	Verde	2020-03-11 12:00:00	BOGOTA	2020-03-11 12:00:00	2020.0	3.0	2020-03	166.0	2020-03-01
...
512	PM10	2023	125.252525	True	Naranja	2023-01-02 10:00:00	BOGOTA	2023-01-02 10:00:00	2023.0	1.0	2023-01	2700.0	2023-01-01
513	PM10	2023	125.747475	True	Naranja	2023-01-02 00:00:00	BOGOTA	2023-01-02 00:00:00	2023.0	1.0	2023-01	2700.0	2023-01-01
514	SO2	2023	11.827957	True	Verde	2023-03-12 08:00:00	BOGOTA	2023-03-12 08:00:00	2023.0	3.0	2023-03	3520.0	2023-03-01
515	SO2	2023	11.827957	True	Verde	2023-03-12 00:00:00	BOGOTA	2023-03-12 00:00:00	2023.0	3.0	2023-03	3520.0	2023-03-01
516	NO2	2023	124.845996	True	Naranja	2023-05-04 20:00:00	BOGOTA	2023-05-04 20:00:00	2023.0	5.0	2023-05	3843.0	2023-05-01

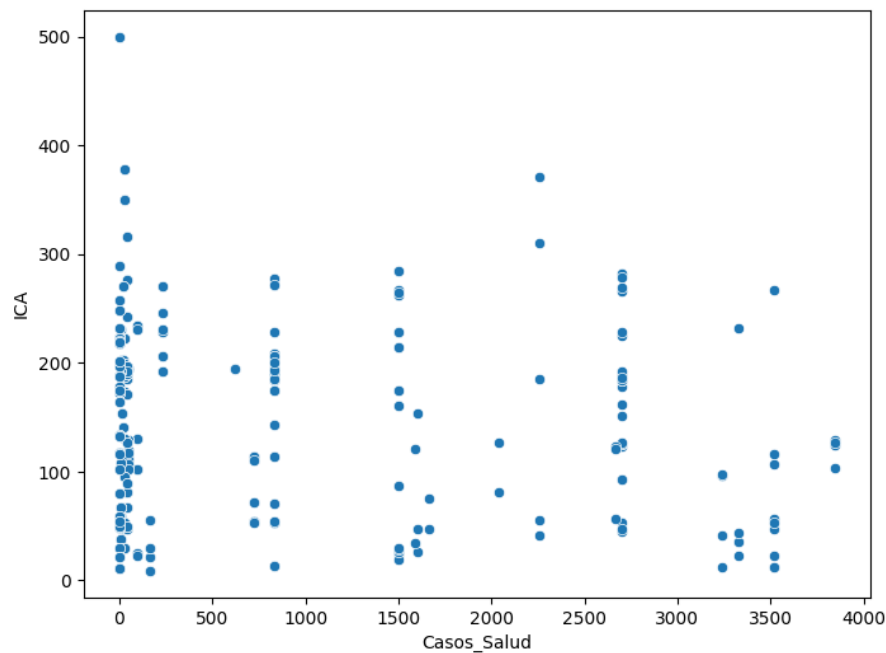
517 rows x 13 columns

Análisis de Correlación

Con el nuevo dataframe se hizo un análisis de correlación de diferentes formas para identificar el impacto que pueda existir en la generación de los casos de salud.

Figura 11

Diagrama de Dispersión entre ICA y los Casos de Salud



o. Se realizó el cálculo de la correlación de Pearson, de Spearman y de Kendal, para identificar

Correlaciones ICA ↔ Casos de Salud

Pearson: $r = -0.1299$ | $p = 0.0143$

Spearman: $\rho = -0.1187$ | $p = 0.0253$

Kendall: $\tau = -0.0826$ | $p = 0.0248$

Los resultados de correlación obtenidos indican que la relación entre el Índice de Calidad del Aire (ICA) y los casos reportados de enfermedades respiratorias es negativa y

estadísticamente significativa, aunque de magnitud débil en los tres coeficientes calculados (Pearson, Spearman y Kendall). Esto implica que, de manera directa e inmediata, los cambios en el ICA no explican de forma sustancial la variación en los casos de salud. La naturaleza negativa de la correlación sugiere que, en los datos analizados, una mejoría del ICA podría coincidir con un leve incremento en los casos de salud o viceversa, lo cual resulta contraintuitivo y evidencia que la relación entre calidad del aire y afectaciones respiratorias no se manifiesta de forma lineal en este análisis inicial.

Este comportamiento débil y aparentemente inverso refuerza la necesidad de considerar efectos más complejos en la interacción clima–salud. Factores como la estacionalidad, los rezagos temporales entre la exposición a contaminantes y la aparición de síntomas, así como la influencia de otros contaminantes específicos, pueden estar afectando la relación observada. Por ello, estos resultados justifican la aplicación de modelos más avanzados para capturar patrones más profundos y evaluar si el impacto real de la calidad del aire se manifiesta bajo configuraciones temporales o no lineales que no son visibles en una correlación directa

p. Se esquematizo la correlación por cada contaminante que conforma el Índice de Calidad del aire obteniendo los siguientes resultados:

Correlación por contaminante

CO -0.251820

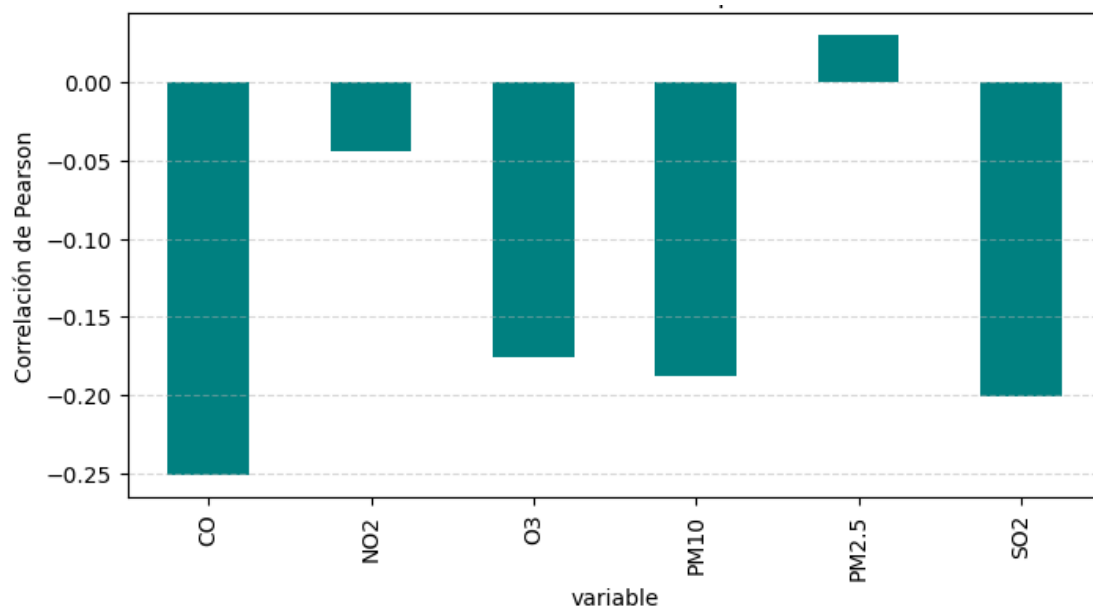
NO2 -0.044385

O3 -0.176289

PM10 -0.187985

PM2.5 0.030066

SO2 -0.200843

Figura 12*Correlación de Pearson por Contaminante*

El análisis de correlación desagregado por tipo de contaminante muestra diferencias importantes en la forma como cada sustancia se asocia con los casos de enfermedades respiratorias. En general, se observa que la mayoría de los contaminantes presentan correlaciones negativas, aunque de magnitud baja, destacándose el monóxido de carbono (CO) con el coeficiente más pronunciado (-0.2518). Esto indica que, para CO, O₃, PM10, SO₂ y NO₂, las fluctuaciones del ICA asociado a cada contaminante no se relacionan de forma directamente proporcional con los cambios en los casos de salud. La leve correlación positiva del PM2.5 (0.0300), aunque prácticamente nula, refleja que este contaminante no presenta una relación lineal clara con los casos reportados, lo cual puede deberse a que el impacto del material particulado fino es más notorio en ventanas temporales rezagadas —algo que no captura la correlación simple por simultaneidad.

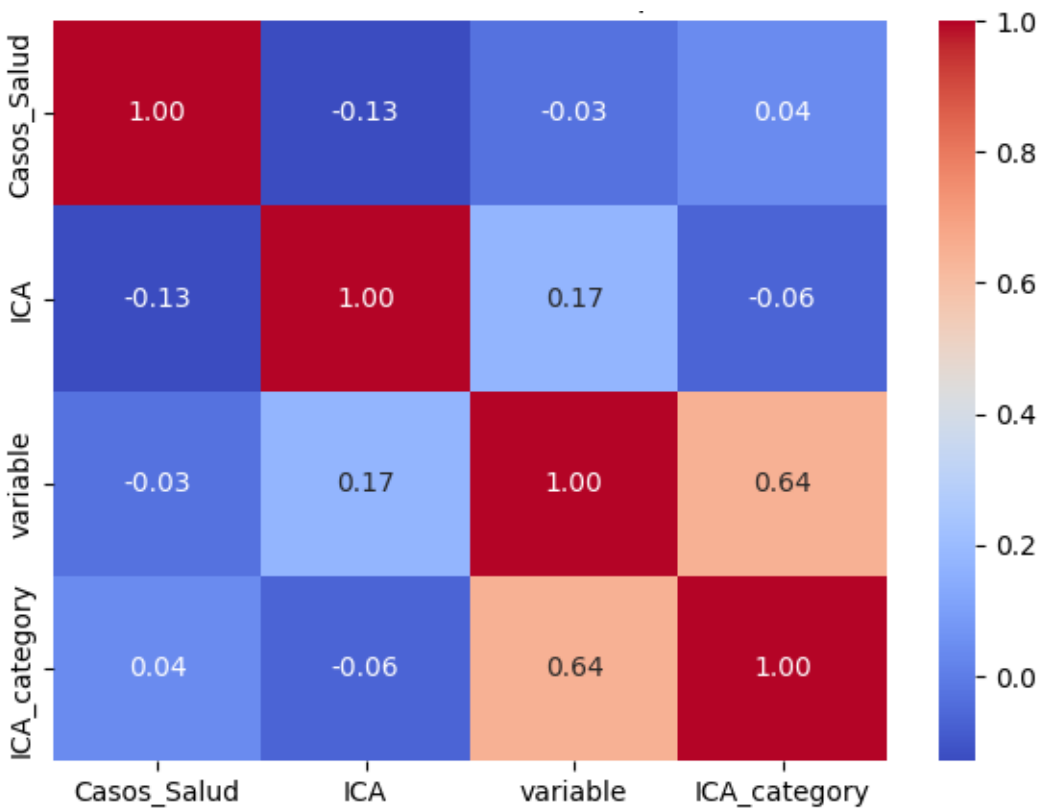
Estas diferencias sugieren que la relación entre contaminación atmosférica y salud no es homogénea entre contaminantes, lo cual coincide con la literatura que indica que cada sustancia tiene dinámicas y tiempos de exposición particulares. La baja magnitud de todas las correlaciones refuerza que los efectos de la calidad del aire sobre la salud no se manifiestan de forma lineal instantánea, sino que probablemente actúan bajo esquemas multivariados, acumulativos o con rezagos temporales. Esto justifica la posterior aplicación de modelos más complejos, como la regresión múltiple, los modelos de rezagos y la causalidad de Granger, con el fin de capturar relaciones que una simple correlación a nivel contemporáneo no puede identificar

Prueba Kruskal-Wallis

Se realizó la prueba de Kruskal – Wallis para identificar diferencias significativas que pueda haber entre las diferentes categorías del ICA

$$H = 7.1562, p = 0.2093$$

Los resultados de la prueba no paramétrica de Kruskal–Wallis indican que no existen diferencias estadísticamente significativas en el número de casos de salud entre las distintas categorías del ICA ($H = 7.1562, p = 0.2093$). Esto significa que, al comparar los grupos de valores de casos asociados a categorías como Verde, Amarillo, Naranja, Rojo, entre otras, no se identifica una variación consistente que permita afirmar que el nivel cualitativo de la calidad del aire esté asociado a aumentos o disminuciones significativas en los casos reportados. Este resultado refuerza la idea de que la relación entre contaminación y salud no se manifiesta de forma inmediata ni categórica, y que las categorías del ICA, al ser umbrales amplios y agrupados, podrían no capturar la variabilidad necesaria para evidenciar efectos diferenciales en el corto plazo, lo que justifica el uso de análisis complementarios como modelos con rezagos o enfoques multivariados para explorar relaciones más complejas y acumulativas.

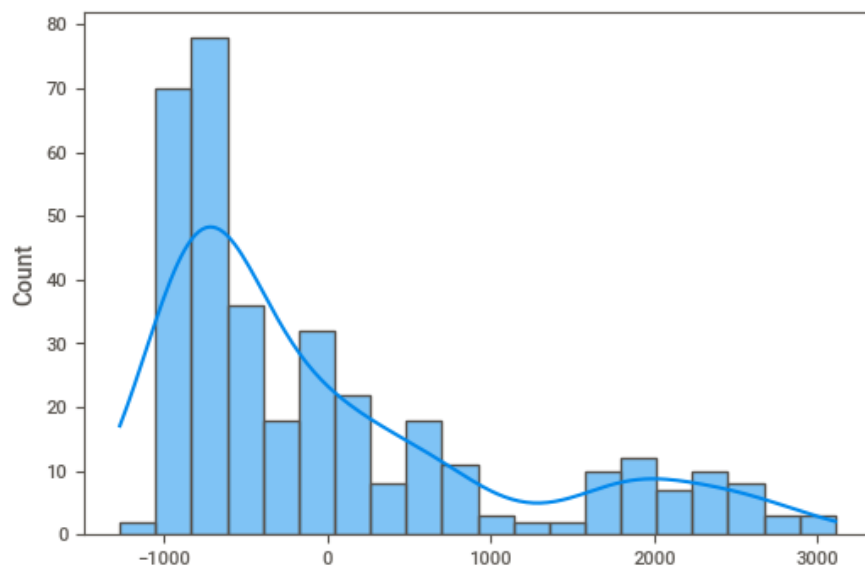
Figura 13*Mapa de calor correlación***Regresión Lineal Múltiple**

Se realizó el modelo de regresión lineal con el propósito de corroborar los datos vistos en la correlación y a pesar que se tienen indicios de que los resultados no son favorables se corrió el modelo para dejar todo el soporte analítico del caso.

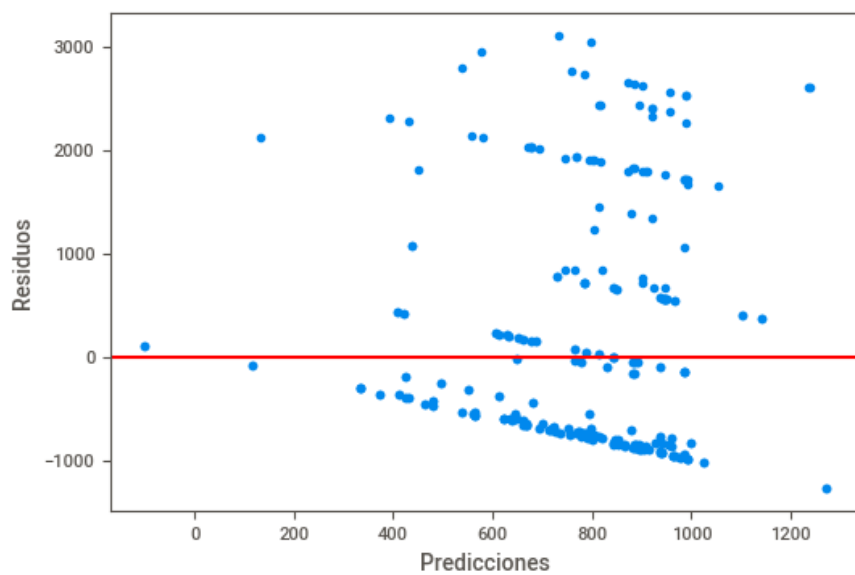
Figura 14*Modelo de Regresión Lineal*

OLS Regression Results						
Dep. Variable:	Casos_Salud	R-squared:	0.032			
Model:	OLS	Adj. R-squared:	0.016			
Method:	Least Squares	F-statistic:	1.931			
Date:	Mon, 24 Nov 2025	Prob (F-statistic):	0.0750			
Time:	22:34:18	Log-Likelihood:	-2978.5			
No. Observations:	355	AIC:	5971.			
Df Residuals:	348	BIC:	5998.			
Df Model:	6					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
const	1038.8911	160.325	6.480	0.000	723.563	1354.219
ICA	-2.9031	1.213	-2.393	0.017	-5.289	-0.517
variable_NO2	57.7336	218.459	0.264	0.792	-371.932	487.399
variable_O3	571.1217	286.479	1.994	0.047	7.673	1134.571
variable_PM10	312.3782	243.445	1.283	0.200	-166.430	791.186
variable_PM2.5	172.7121	252.637	0.684	0.495	-324.175	669.599
variable_SO2	-14.8142	204.909	-0.072	0.942	-417.831	388.202
Omnibus:	64.999	Durbin-Watson:	0.078			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	97.230			
Skew:	1.262	Prob(JB):	7.70e-22			
Kurtosis:	3.445	Cond. No.	1.29e+03			

Se visualiza la distribución de los residuos posterior a correr el modelo de regresión lineal.

Figura 15*Distribución de Residuos*

Se realiza la visualización de los residuos vs las predicciones y el comportamiento de homocedasticidad entre las variables

Figura 16*Relación entre los Residuos y las Predicciones*

El modelo de regresión lineal múltiple logró ajustarse correctamente, pero su capacidad explicativa fue limitada ($R^2 = 0.032$). La inspección de los residuos evidenció autocorrelación significativa (Durbin–Watson = 0.08), no normalidad de los errores y multicolinealidad entre las variables explicativas, lo que indica que el modelo OLS no es adecuado para analizar la relación entre contaminación del aire y enfermedades respiratorias en series temporales. Se recomienda avanzar hacia modelos con rezagos como VAR, DLM y pruebas de causalidad de Granger, los cuales permiten capturar los efectos temporales diferidos característicos de este tipo de fenómenos.

Causalidad de Granger

Teniendo en cuenta los resultados de Correlación se realizó el modelo de Causalidad de Granger para identificar si el ICA tienen un impacto en la generación de los casos de salud y se obtuvieron los siguientes resultados.

Granger Causality

number of lags (no zero) 1

ssr based F test: $F=1.1365$, $p=0.2871$, $df_denom=351$, $df_num=1$

ssr based chi2 test: $chi2=1.1462$, $p=0.2843$, $df=1$

likelihood ratio test: $chi2=1.1444$, $p=0.2847$, $df=1$

parameter F test: $F=1.1365$, $p=0.2871$, $df_denom=351$, $df_num=1$

Granger Causality

number of lags (no zero) 2

ssr based F test: $F=0.7965$, $p=0.4517$, $df_denom=348$, $df_num=2$

ssr based chi2 test: $chi2=1.6159$, $p=0.4458$, $df=2$

likelihood ratio test: $chi2=1.6122$, $p=0.4466$, $df=2$

parameter F test: $F=0.7965$, $p=0.4517$, $df_denom=348$, $df_num=2$

Granger Causality

number of lags (no zero) 3

ssr based F test: $F=0.7024$, $p=0.5512$, $df_denom=345$, $df_num=3$

ssr based chi2 test: $chi2=2.1499$, $p=0.5419$, $df=3$

likelihood ratio test: $chi2=2.1433$, $p=0.5432$, $df=3$

parameter F test: $F=0.7024$, $p=0.5512$, $df_denom=345$, $df_num=3$

Los resultados de la prueba de causalidad de Granger indican que no existe evidencia estadística de que la variación del ICA en los meses previos prediga el comportamiento de los casos de salud en Bogotá. Para los rezagos evaluados (1, 2 y 3 meses), los valores p fueron superiores al umbral de significancia del 5% ($lag1 = 0.2871$; $lag2 = 0.4517$; $lag3 = 0.5512$), lo que sugiere que los cambios en la calidad del aire no Granger-causan cambios posteriores en la incidencia de enfermedades respiratorias dentro de la ventana temporal analizada. En otras palabras, aunque existe una asociación leve entre ambas series, esta relación no presenta un componente temporal predictivo, lo que podría indicar efectos más complejos, rezagos más largos o la intervención de factores externos no incluidos en el modelo.

Las pruebas de estacionariedad ADF complementan este hallazgo al mostrar que las dos series se comportan de manera diferente en términos temporales. Mientras que la serie del ICA resultó ser claramente estacionaria ($ADF = -10.71$; $p < 0.001$), la serie de casos de salud no lo es ($ADF = -2.31$; $p = 0.1676$). Esto implica que el ICA presenta un comportamiento estable a lo largo del tiempo, sin tendencias marcadas, mientras que los casos de salud mantienen tendencias o patrones persistentes, posiblemente influenciados por ciclos estacionales, dinámicas epidemiológicas u otros determinantes no ambientales. Esta diferencia en la estructura temporal

de las dos series afecta directamente las posibilidades de establecer relaciones de causalidad en el corto plazo. En términos metodológicos, una relación Granger significativa es más difícil de detectar cuando las series no comparten el mismo nivel de integración, como ocurre en este caso, lo cual constituye un aporte importante para la interpretación analítica global del proyecto.

Conclusiones

El desarrollo integral de este proyecto permitió abordar de manera sistemática y basada en evidencia la relación entre los patrones de calidad del aire y la incidencia de enfermedades respiratorias en la ciudad de Bogotá durante el periodo 2020–2024. A través de la consolidación, depuración y unificación de múltiples fuentes de datos ambientales y sanitarios, se lograron construir series temporales robustas que sirvieron como insumo para los análisis exploratorios, correlacionales y predictivos planteados en los objetivos. El proceso de integración de las bases, sumado a la estandarización del Índice de Calidad del Aire (ICA) mediante los umbrales oficiales para Colombia, permitió disponer de un indicador ambiental comparable entre años, contaminantes y momentos específicos, lo que dotó al análisis de mayor consistencia técnica.

En términos descriptivos, los resultados evidenciaron variaciones claras tanto en la calidad del aire como en los casos notificados de enfermedades respiratorias, mostrando patrones mensuales y anuales que sugieren la influencia de factores estacionales y dinámicas propias de la movilidad urbana, el clima y el comportamiento poblacional. La distribución de categorías del ICA reveló la presencia de episodios críticos especialmente asociados a contaminantes como PM_{2.5}, PM₁₀ y O₃, lo que coincide con lo reportado por instituciones ambientales sobre la principal carga contaminante en la ciudad. Sin embargo, la tendencia de los casos de salud mostró un comportamiento más irregular y dependiente de ciclos internos del sistema de vigilancia epidemiológica.

Las correlaciones realizadas permitieron identificar una relación negativa, débil pero estadísticamente significativa, entre el ICA y los casos respiratorios. Tanto las pruebas de Pearson como Spearman y Kendall coincidieron en que la relación existe, pero es moderada y no suficientemente fuerte como para afirmar dependencia lineal o monotónica clara entre las

variables. Este hallazgo sugiere que, aunque la calidad del aire incide sobre la salud respiratoria, su efecto no se manifiesta de forma directa ni uniforme en las series temporales trabajadas, posiblemente debido al peso de otros factores concurrentes como las condiciones climáticas, la circulación de virus respiratorios, la variabilidad intrínseca del sistema de salud y riesgos individuales no observados.

El análisis por tipo de contaminante permitió profundizar en esta relación, mostrando que algunos agentes presentan una conexión más marcada con los casos de salud, como es el caso del CO, SO₂, PM10 y O₃, que exhibieron correlaciones negativas de mayor magnitud. Estos resultados coinciden con la literatura internacional y aportan evidencia contextualizada a Bogotá, aunque nuevamente la fuerza de asociación no fue alta, indicando que estos contaminantes podrían contribuir al riesgo respiratorio, pero no de forma aislada.

La prueba de Kruskal–Wallis, diseñada para evaluar si los casos respiratorios varían significativamente según las categorías del ICA, no mostró diferencias estadísticamente significativas entre grupos. Esto confirma que los incrementos o disminuciones del ICA, aunque relevantes desde el punto de vista ambiental, no se reflejan de manera inmediata ni claramente diferenciada en los registros de salud. De igual forma, los análisis de causalidad de Granger fueron consistentes al indicar que la calidad del aire no presenta un efecto causal predictivo sobre los casos respiratorios en rezagos de 1 a 3 meses. Este resultado se interpreta como evidencia de que la respuesta sanitaria a la contaminación atmosférica podría ser más compleja, no lineal o influenciada por rezagos más largos, manifestaciones crónicas o interacciones con otros determinantes epidemiológicos.

Un elemento clave para interpretar estos resultados es la estacionariedad de las series. Mientras que la serie del ICA resultó ser estacionaria, la de salud no lo fue, lo cual complica de

manera estructural la posibilidad de establecer relaciones temporales robustas. Esta diferencia en comportamiento temporal confirma que los casos respiratorios están sujetos a tendencias, ciclos y factores que no dependen exclusivamente de la calidad del aire, lo que restringe la capacidad predictiva de los modelos lineales y de los métodos de causalidad clásicos. En consecuencia, se reafirma que la calidad del aire contribuye a la explicación del fenómeno, pero no es el principal ni único determinante.

La modelación estadística, incluyendo la regresión lineal múltiple y el Random Forest, aportó elementos adicionales de interpretación. Aunque el ICA mostró un efecto significativo en la regresión, su contribución explicativa fue baja, reafirmando que la variabilidad de los casos de salud no depende principalmente de los contaminantes atmosféricos en los niveles medidos. El Random Forest permitió evaluar la importancia relativa de las variables ambientales y confirmó que ningún contaminante por sí solo domina el comportamiento de los casos de salud, evidenciando que el fenómeno es multicausal y requiere un enfoque más amplio para construir modelos predictivos más precisos.

En conjunto, los resultados obtenidos permiten concluir que sí existe una relación entre la calidad del aire y los problemas respiratorios, pero que esta es débil, compleja y no se manifiesta de manera lineal ni causal en periodos cortos de tiempo. Los contaminantes atmosféricos analizados influyen, pero su impacto no es suficiente para anticipar ni explicar por completo el comportamiento de la morbilidad respiratoria en Bogotá. Esto sugiere la necesidad de integrar variables adicionales como temperatura, humedad, presión atmosférica, circulación de patógenos estacionales, factores socioeconómicos y condiciones crónicas de salud, así como de explorar modelos no lineales o basados en series temporales multivariadas más complejas.

Finalmente, este trabajo aporta un análisis sólido y metodológicamente consistente que demuestra las limitaciones y los alcances reales de los datos disponibles en Colombia para estudiar la relación entre ambiente y salud. Los resultados invitan a fortalecer la integración de datos climatológicos, epidemiológicos y ambientales, así como a avanzar hacia sistemas de vigilancia más robustos que permitan construir modelos explicativos y predictivos con mayor precisión. Con ello, se abren oportunidades para mejorar la toma de decisiones en salud pública, la gestión ambiental y la formulación de políticas orientadas a reducir el impacto de la contaminación atmosférica sobre la salud de la población

Recomendaciones

A partir de los resultados obtenidos y de las limitaciones identificadas durante el proceso investigativo, se recomienda fortalecer los mecanismos de articulación institucional entre el Ministerio de Ambiente y Desarrollo Sostenible y el Ministerio de Salud y Protección Social, de manera que se consolide un sistema integrado de datos ambientales y sanitarios. Esta cooperación permitiría estandarizar las metodologías de captura, almacenamiento y publicación de información, reduciendo inconsistencias entre fuentes y facilitando futuros estudios que busquen comprender el impacto de la calidad del aire en la salud pública. Una infraestructura robusta y unificada no solo favorecería la trazabilidad de los indicadores, sino que también contribuiría a una toma de decisiones más ágil, fundamentada en evidencia y con alcance intersectorial.

En esa misma línea, se sugiere la conformación de grupos de trabajo interdisciplinarios, integrados por especialistas en salud pública, epidemiología, climatología, calidad del aire, estadística y ciencia de datos. Este tipo de equipos permitiría analizar los fenómenos desde una perspectiva integral, incorporando tanto los determinantes ambientales como las dinámicas epidemiológicas y sociales. Lograr una lectura multidimensional facilitaría la identificación de patrones tempranos, la construcción de modelos predictivos más precisos y la formulación de intervenciones oportunas para la prevención de enfermedades respiratorias asociadas a factores ambientales.

Asimismo, se recomienda avanzar en la implementación de sistemas de monitoreo ambiental y sanitario con mayor granularidad temporal y espacial. La periodicidad horaria de los datos de calidad del aire y los registros clínicos consolidados en escalas mensuales generan restricciones importantes para la aplicación de modelos con mayor detalle. Contar con datos

diarios o incluso subdiarios permitiría emplear métodos más sofisticados de series temporales y machine learning, mejorando la capacidad de anticipación de eventos críticos. De igual manera, disponer de información georreferenciada facilitaría el análisis de desigualdades territoriales en la exposición a contaminantes y sus efectos en salud.

Otra recomendación clave es fomentar el desarrollo de modelos predictivos integrados con variables climáticas, tales como temperatura, humedad, velocidad del viento y precipitación, ya que estos factores modulan la dispersión y concentración de contaminantes atmosféricos. Incorporar estas variables podría enriquecer la explicación del comportamiento de los casos respiratorios y mejorar la capacidad de respuesta ante posibles eventos de contaminación extrema o incrementos abruptos en la morbilidad respiratoria.

Finalmente, se sugiere que los estudios futuros consideren la inclusión de variables socioeconómicas y demográficas, la estacionalidad de agentes virales respiratorios, así como condiciones previas de salud de la población. Estos elementos permiten entender mejor la heterogeneidad del riesgo y las vulnerabilidades diferenciales entre grupos poblacionales, lo cual es esencial para diseñar políticas más equitativas y estrategias de mitigación focalizadas. La integración de estas variables, junto con la mejora de los sistemas de datos y la articulación institucional, constituye un camino prometedor para fortalecer el análisis entre ambiente y salud en Colombia y potenciar la toma de decisiones basadas en evidencia.

Referencias

- Carbonell Chams, A. J. (2023). Transporte público y problemas medioambientales en Colombia. *Observatorio Medioambiental*, 26, 23–44. <https://doi.org/10.5209/obmd.93019>
- Cardona, L. A. L., Ocampo, D. F. F., & Esponda, R. D. E. (2022). Aplicaciones de la Datificación y Big Data en América Latina entre el 2015 y 2019. *Revista Logos Ciencia & Tecnología*, 14(2), 125–143. <https://doi.org/10.22335/rlct.v14i2.1594>
- Cediel Becerra, N. M., Machado, D. F., Pineda, J., Cartín-Rojas, A., Aguirre, L. F., Vargas, R., Sánchez, M. P., Vega, S., & Morais, M. (2023). Crisis climática y Una Salud en Iberoamérica. *Revista de Medicina Veterinaria*. <https://doi.org/10.19052/mv.voll.iss46.1>
- Díaz Subieta, L. B. (2024). El uso de la inteligencia artificial en la investigación científica. *Journal History of Latin American Education / Revista Historia de la Educación Latinoamericana*, 26(43), 253–272. <https://doi.org/10.19053/upte.01227238.18014>
- Gilma C., Mantilla C., & Christina Li. (2019, diciembre 15). *Enseñanza de cambio climático y salud en facultades de medicina en Colombia* [Video recording].
- Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3), 424–438. <https://doi.org/10.2307/1912791>
- Grisales Aguirre, A. M., Espinosa Villa, L. M., & Alzate, P. (2024). La evolución de la visualización de datos: Campos de aplicación y herramientas de Business Intelligence. *Revistade Ingenierías Interfaces*, 7(1), 1–12. <https://research.ebsco.com/linkprocessor/plink?id=b0f6b0a2-2bde-3f71-acb6-8207bcee2716>

İbrahim, D., & Özkan, A. (2024). Determination of Ozone-induced mortality in Aydın province, Turkey in 2022 using AIRQ+ Modeling. *EQA*, 63, 52–57.

<https://doi.org/10.6092/issn.2281-4485/20078>

Instituto de Hidrología, Meteorología y Estudios Ambientales – IDEAM. (s.f.). *Calidad del Aire en Colombia (Promedio Anual)*. Datos Abiertos Colombia.

https://www.datos.gov.co/Ambiente-y-Desarrollo-Sostenible/Calidad-Del-Aire-En-Colombia-Promedio-Anual-/kekd-7v7h/about_data

Instituto Nacional de Salud – INS. (s.f.). *Datos de Vigilancia en Salud Pública de Colombia (SIVIGILA)*. Datos Abiertos Colombia. https://www.datos.gov.co/Salud-y-Proteccion-Social/Datos-de-Vigilancia-en-Salud-P-blica-de-Colombia/4hyg-wa9d/about_data

Juan D. Hunter. (2007). *Matplotlib: A 2D Graphics Environment* (3a ed., Vol. 9). Computing in Science & Engineering.

Lemus-Delgado, D., & Pérez Navarro, R. (2023). Ciencia de datos y estudios globales: aportaciones y desafíos metodológicos. *OpenAIRE*.

<http://journals.openedition.org/colombiaint/3824>

Ministerio de Ambiente y Desarrollo Sostenible. (2017). *Resolución 2254 del 1 de noviembre de 2017: Por la cual se establecen los estándares de calidad del aire*. República de Colombia.

Monroy Alfaro, C. R. (2022). “El lenguaje python y su potencial en el desarrollo de software de inteligencia artificial”. *Masferrer Investiga: Revista Científica de la Universidad Salvadoreña Alberto Masferrer*, 12(1), 18–41.

<https://research.ebsco.com/linkprocessor/plink?id=5c0a0338-f537-3ec9-98ba-900bcb163e74>

NumFOCUS, Inc. (2024). *Pandas Documentation*. Versión 2.2, OVHcloud .

https://pandas.pydata.org/docs/getting_started/overview.html

Osorio Amaya, E. A., Inzunza Cazares, S., & Ward Bringas, S. E. (2023). Modelización estadística para el aprendizaje de la correlación y regresión lineal : Statistical Modeling for Learning Correlation and Linear Regression. En *PNA: Revista de investigación en didáctica de la matemática* (Vol. 17, Número . 3, pp. 295–321).

<https://doi.org/10.30827/pna.v17i3.23937>

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2022). *Scikit-learn: Machine Learning in Python* (JMLR, Ed.; Vol. 23). <http://scikit-learn.sourceforge.net>

Petrone, P., & Petrone, P. (2020). Cambio climático y su impacto sobre la salud humana. *Revista Colombiana de Cirugía*, 35(3), 347–350. <https://doi.org/10.30944/20117582.723>

Python Software Foundation. (2024, noviembre). *¿Qué es Python?* © Copyright 2001-2024.

Shams, S. R., Choi, Y., Singh, D., Ghahremanloo, M., Momeni, M., & Park, J. (2024).

Innovative approaches for accurate ozone prediction and health risk analysis in South Korea: The combined effectiveness of deep learning and AirQ+. *Science of The Total Environment*, 946, 174158. <https://doi.org/10.1016/J.SCITOTENV.2024.174158>

Su, J. G., Vuong, V., Shahriary, E., Aslebagh, S., Yakutis, E., Sage, E., Haile, R., Balmes, J., & Barrett, M. (2024). Health effects of air pollution on respiratory symptoms: A longitudinal study using digital health sensors. *Environment International*, 189.

<https://doi.org/10.1016/j.envint.2024.108810>

Uribe-Soto, M., Gomez, A. P., & Ramirez-Nieto, G. (2020). Influenza needs an approach as a “one health” problem in Colombia. *Acta Biologica Colombiana*, 25(3), 421–430.

<https://doi.org/10.15446/abc.v25n3.79364>

Urrego-Sandoval, C. (2021). Ahora o nunca. Gobernanza, coproducción y bioeconomía contra el cambio climático. *OASIS*, 35, 361–365. <https://doi.org/10.18601/16577558.n35.18>

Wang, J., Cortes-Ramirez, J., Gan, T., Davies, J. M., & Hu, W. (2024). Effects of climate and environmental factors on childhood and adolescent asthma: A systematic review based on spatial and temporal analysis evidence. *Science of The Total Environment*, 951, 175863.

<https://doi.org/10.1016/J.SCITOTENV.2024.175863>

Apendices

Apéndice A

Enlaces de Información

Base de datos de la calidad del aire en Colombia

https://drive.google.com/drive/folders/1BZUXY_mCwD8PST4VR7KmlK_WdgE8fw3J?usp=sharing

Base de datos de salud 2020 – 2024

https://drive.google.com/drive/folders/1BZUXY_mCwD8PST4VR7KmlK_WdgE8fw3J?usp=sharing

Archivo en Phyton con la consolidación de los modelos empleados

<https://colab.research.google.com/drive/1T5f9sERmRIVBEFWbPyXigFhdKcefdxaL?usp=sharing>