Análisis y predicción de la tasa de mortalidad por cáncer de pulmón, tráquea y bronquios: un estudio comparativo en tres departamentos de Colombia mediante machine learning y modelos de series de tiempo

Óscar Eduardo Vásquez Parra

Asesor

Luis Angel Anillo Arrieta

Universidad Nacional Abierta y a Distancia UNAD

Escuela de Ciencias Básicas, Tecnología e Ingeniería ECBTI

Especialización en Ciencia de Datos y Analitica

2025

_____

Nombre Director de Trabajo de Grado

_____ _____

Jurado Jurado

2025

Dedicatoria

A Dios, por ser mi guía constante, fuente de fortaleza y sabiduría en cada paso de este camino.

A mis padres, quienes con su amor incondicional, esfuerzo y sacrificio han sido el pilar fundamental en mi formación. Gracias por enseñarme el valor de la perseverancia, la responsabilidad y el compromiso, valores que han guiado mi camino hasta este logro.

A mis profesores y mentores, quienes con su dedicación y conocimiento han dejado una huella imborrable en mi formación. Sus enseñanzas han sido una fuente invaluable de inspiración.

Y, por último, a todas aquellas personas que, de una u otra manera, han contribuido a mi crecimiento académico y personal. Cada gesto de apoyo ha sido un impulso para llegar hasta aquí.

Este logro es también de ustedes. ¡Gracias!

Resumen

El cáncer de pulmón, tráquea y bronquios es una de las principales causas de mortalidad a nivel mundial, con un aumento significativo en los últimos años. Este estudio analiza y predice la tasa de mortalidad en los departamentos de Antioquia, Risaralda y Quindío mediante modelos de series de tiempo, como ARIMA (AutoRegressive Integrated Moving Average), ETS (Exponential Smoothing State Space Model) y Holt-Winters (Triple Exponential Smoothing), con el propósito de identificar el modelo más preciso. Los resultados revelan que el modelo ARIMA, con parámetros (1,1,1), ofrece la mejor precisión en la predicción de tendencias futuras. Además, se identifican variaciones en la tasa de mortalidad por cada 100.000 habitantes entre los departamentos estudiados, con algunas regiones mostrando descensos y otras un posible incremento en los próximos años. La detección de estos patrones es crucial para la toma de decisiones en salud pública y el diseño de estrategias de prevención. Este estudio resalta la importancia de los modelos predictivos en la gestión sanitaria y en la formulación de políticas que reduzcan el impacto de esta enfermedad en la población colombiana. Por ello, resulta fundamental seguir explorando nuevas metodologías basadas en inteligencia artificial y aprendizaje automático para mejorar la precisión de las predicciones.

***Palabras clave*:** Cáncer, Análisis, Modelos, Tendencias, Predicciones.

Abstract

Lung, tracheal, and bronchial cancer is one of the leading causes of mortality worldwide, with a significant increase in recent years. This study analyzes and predicts the mortality rate in the departments of Antioquia, Risaralda, and Quindío using time series models such as ARIMA (AutoRegressive Integrated Moving Average), ETS (Exponential Smoothing State Space Model), and Holt-Winters (Triple Exponential Smoothing), with the aim of identifying the most accurate model. The results reveal that the ARIMA model, with parameters (1,1,1), provides the highest precision in forecasting future trends. Additionally, variations in the mortality rate per 100,000 inhabitants are identified among the studied departments, with some regions showing declines and others a possible increase in the coming years. Detecting these patterns is crucial for public health decision-making and the design of prevention strategies. This study highlights the importance of predictive models in healthcare management and policy-making to mitigate the impact of this disease on the Colombian population. Therefore, it is essential to continue exploring new methodologies based on artificial intelligence and machine learning to improve prediction accuracy.

Keywords**:** Cancer, Analysis, Models, Trends, Predictions.

Tabla de Contenido

Lista de Tablas

Lista de Figuras

Introducción

El cáncer es una de las principales causas de mortalidad en Colombia y en el mundo (Vergara et al.,2017). En el país, esta enfermedad ha cobrado miles de vidas, convirtiéndose en un problema de salud pública que requiere atención prioritaria. Sin embargo, los avances en la medicina, especialmente en el diagnóstico temprano mediante instrumentos de medicina nuclear como la radiografía, tomografía por emisión de positrones (PET) y la gammagrafía (Llamas, 2022), han permitido mejorar las tasas de detección y tratamiento, logrando que muchos pacientes tengan mayores oportunidades de recuperación y sobrevida.

Entre los diferentes tipos de cáncer, el cáncer de pulmón, tráquea y bronquios se encuentra entre los más prevalentes y letales en Colombia. Según estudios epidemiológicos (Zhou et al., 2022), estas neoplasias han mostrado un crecimiento alarmante en las últimas décadas, lo que refleja una problemática que va más allá de factores individuales y abarca aspectos ambientales, socioeconómicos y demográficos. Según Andújar et al. (2021), entre los posibles factores de este tipo de cáncer podemos encontrar el tabaquismo, la contaminación del aire, exposición ocupacional, la urbanización e industrialización, los hábitos de vida y alimentación y por supuesto que factores genéticos en menores proporciones.

Si bien se han realizado investigaciones previas en cuanto al cáncer a nivel país, se decidió individualizar el estudio de este fenómeno para un tipo de cáncer en 3 departamentos (Antioquia, Risaralda y Quindío) dado a un análisis exploratorio previo que indicaban que eran las zonas que más contaban con este tipo de patologías, presentando no solo altas tasas de incidencia sino de de mortalidad.

Por lo tanto, más allá de querer determinar un número de casos, es crucial identificar patrones específicos dentro de los tres departamentos dada la información de los datos históricos,

para pronosticar el rumbo de esta enfermedad en los años posteriores, este estudio propone el uso de técnicas de Machine Learning (Zhang et al., 2023) como lo son los modelos de series de tiempo para predecir la tasa de mortalidad futura y mediante la evaluación de diversas métricas de desempeño, se busca determinar que modelo ofrece buenos resultados en términos de eficiencia.

Los hallazgos de esta investigación no solo permitirán conocer la evolución de la mortalidad en cada ciudad, sino que también proporcionarán información valiosa para la formulación de políticas públicas en salud. Se espera que los resultados de este estudio sirvan como una base científica para que las entidades gubernamentales y de salud implementen estrategias de prevención.

¿Cómo se comportará la tasa de mortalidad por cáncer de pulmón, tráquea y bronquios en Antioquia, Risaralda y Quindío en los próximos años según el modelo de series de tiempo más óptimo?

En este sentido, se propone la pregunta de investigación que permita tener un enfoque en una problemática que, aunque no es silenciosa afecta la vida de una población en general con el fin de tener una representación visual que facilite conocer más sobre estas afecciones en el tiempo y buscar solventar poco a poco este mal.

Justificación

El problema abordado en este proyecto se da por la creciente prevalencia de enfermedades cancerígenas a nivel nacional el cual hace que sea de mucha importancia investigar este problema debido a sus graves implicaciones para la salud pública y el bienestar de la población. Como señala Alarcón et al. (2021), en un estudio del Instituto Nacional De Cancerología, "Se deben ampliar los estudios de factores de riesgo en el cáncer de pulmón y se requieren trabajos observacionales que permitan contemplar otros factores que pueden estar impactando en la incidencia de esta enfermedad para continuar ideando estrategias de prevención" y así no tener que ver más el incremento de las tasas de mortalidad.

El cáncer de pulmón, tráquea y bronquios representa una de las principales causas de mortalidad a nivel mundial, con más de dos millones de muertes registradas en 2019 (Wang et al.,2023). La alta letalidad asociada a estos tipos de cáncer subraya la necesidad de desarrollar herramientas predictivas que permitan estimar con precisión las tasas de mortalidad para hacer estudios efectivos gracias a mediciones históricas. En este contexto, la implementación de modelos de aprendizaje supervisado también ofrece una oportunidad valiosa para analizar grandes volúmenes de datos clínicos y epidemiológicos, facilitando la identificación de patrones y factores de riesgo que influyen en la supervivencia de los pacientes (Tabio et al.,2021).

La integración de técnicas de aprendizaje supervisado en la investigación oncológica contribuye al avance del conocimiento científico. Al identificar nuevos factores pronósticos y validar biomarcadores asociados con la mortalidad, se pueden desarrollar intervenciones más efectivas y estrategias de prevención más precisas. La colaboración entre instituciones médicas y centros de investigación es esencial para recopilar y analizar datos de manera conjunta,

fortaleciendo la capacidad predictiva de los modelos y, en última instancia, mejorando los resultados en salud para los pacientes afectados por estos tipos de cáncer.

Con esta investigación se espera obtener un análisis que permita observar mediante el modelo futuras predicciones y ciertas características de dicha problemática de manera más sencilla y así cerrar las brechas de conocimiento identificadas y proporcionar una base sólida para futuras investigaciones y acciones de intervención en el que se pueda extender dicho modelo o análisis a otras ciudades y no solo en las implicadas para carácter comparativo dada su población. Como señala Woodcock et al. (2009), la implementación de modelos predictivo basado en datos epidemiológicos podría proporcionar información valiosa para la toma de decisiones en política de salud pública.

La realización de este proyecto aplicado cobra una relevancia aún mayor en la actualidad, especialmente en un contexto donde la salud pública se ha vuelto más vulnerable debido a diversos factores, sobre todo después del Covid-19 (Esendağli et al., 2021), que ha dejado secuelas significativas en la vida de la población y ha resaltado la importancia de abordar esta problemática de las enfermedades de tipo cancerígenas de manera integral. Así, mediante la tasa de mortalidad y el dominio de otras variables se busca radicar más atención a casos puntuales con el objetivo de abordar esto de manera efectiva en la que los beneficiarios directos incluyen a los individuos afectados por dichas enfermedades, así como a los profesionales de la salud y los responsables de formular políticas públicas.

Objetivos

Objetivo General

Determinar el modelo de series de tiempo óptimo para predecir la tasa de mortalidad de cáncer de pulmón, tráquea y bronquios en los departamentos de Antioquia, Risaralda y Quindio.

Objetivos Específicos

Analizar los datos históricos mediante visualizaciones para identificar patrones y comparar las tendencias observadas entre los departamentos seleccionados para ofrecer recomendaciones de salud pública.

Identificar los parámetros del modelo ARIMA mediante las características usuales en series de tiempo (estacionariedad y autocorrelación).

Evaluar las métricas de rendimiento en los modelos empleados.

Marco Teórico

El cáncer tiene un impacto significativo en la sociedad y las estadísticas relacionadas con esta enfermedad permiten analizar su evolución a lo largo del tiempo y comprender su impacto en grandes grupos de personas. Estos datos reflejan cuántas personas son diagnosticadas y fallecen a causa del cáncer cada año, cuántas viven con la enfermedad en la actualidad, la edad promedio al momento del diagnóstico y la cantidad de pacientes que siguen con vida después de recibirlo. Además, permiten identificar diferencias entre distintos grupos según edad, género, etnia, ubicación geográfica y múltiples factores.

El análisis de las tasas de cáncer a lo largo del tiempo ha ayudado a detectar tendencias y evaluar cambios en el riesgo de desarrollar distintos tipos de cáncer o de fallecer por esta causa y con la llegada de la inteligencia artificial y puntualmente el Machine learning a nuestras vidas, no está de más implementar modelos que ayuden a encontrar nuevas formas de batallar con esta calamidad pública mediante los distintos datos y así permitir la evaluación de estas técnicas con sus respectivas métricas.

Aunque las estadísticas no se aplican de manera individual a cada paciente, son fundamentales para gobiernos, profesionales de la salud y responsables de políticas públicas pues permiten diseñar estrategias efectivas para abordar los desafíos que representa el cáncer y evaluar el impacto de las iniciativas dirigidas a su control y tratamiento.

Según el Ministerio de Salud (MinSalud, 2023) el diagnóstico de cáncer implica una serie de procedimientos que permiten detectar la presencia de células anormales en el cuerpo. Los métodos varían según el tipo de cáncer sospechado en los cuales incluyen radiografías, tomografía computarizada (TC), resonancia magnética (RM), ecografía, tomografía por emisión de positrones (PET) entre otras, dado esto mediante bases de datos abiertos se encontró que uno

de los procedimientos más requeridos corresponde a radiografias puntualmente de Tórax el cual es la principal para la detección de tipos de cáncer como pulmón, tráquea y bronquios.

En Colombia, los 5 tipos de cáncer más comunes son: Mama, próstata, colon y recto, estómago y pulmón, tráquea y bronquios en el que estos últimos de manera individual se definen como:

Cáncer de Pulmón

Enfermedad caracterizada por el crecimiento descontrolado de células malignas del tracto respiratorio, en particular del tejido pulmonar, y uno de los tipos de cáncer más frecuentes a nivel mundial (Rodriguez et al., 2018).

Cáncer de Tráquea

Es una neoplasia maligna poco frecuente que se desarrolla en la tráquea, la estructura tubular que conecta la laringe con los bronquios principales (Hetnał et al., 2010). Este tipo de cáncer puede manifestarse con síntomas como tos persistente, dificultad para respirar y, en casos avanzados, hemoptisis (expectoración de sangre).

Cáncer de Bronquios

Se refiere a tumores malignos que se originan en los bronquios, las principales vías aéreas que se ramifican desde la tráquea hacia los pulmones (Fuentes & Corona, 2002). Este tipo de cáncer puede provocar obstrucción de las vías respiratorias, síntomas como tos crónica, infecciones pulmonares recurrentes y disnea (dificultad para respirar). La exposición a factores de riesgo como el tabaquismo y la inhalación de sustancias carcinógenas aumenta la probabilidad de desarrollar cáncer bronquial.

Cualquiera de estas afecciones hace parte del grupo de las enfermedades respiratorias que en conexión con una región sus causas pueden ser multifacéticas y abarcar una serie de factores

interrelacionados que impactan la salud de la población, en nuestro caso se puede estimar el análisis en tres ciudades capitales de Antioquia, Risaralda y Quindío el cual aunque en tamaño poblacional tiene sus grandes diferencias se tendrá que usar tasas ajustadas por habitantes.

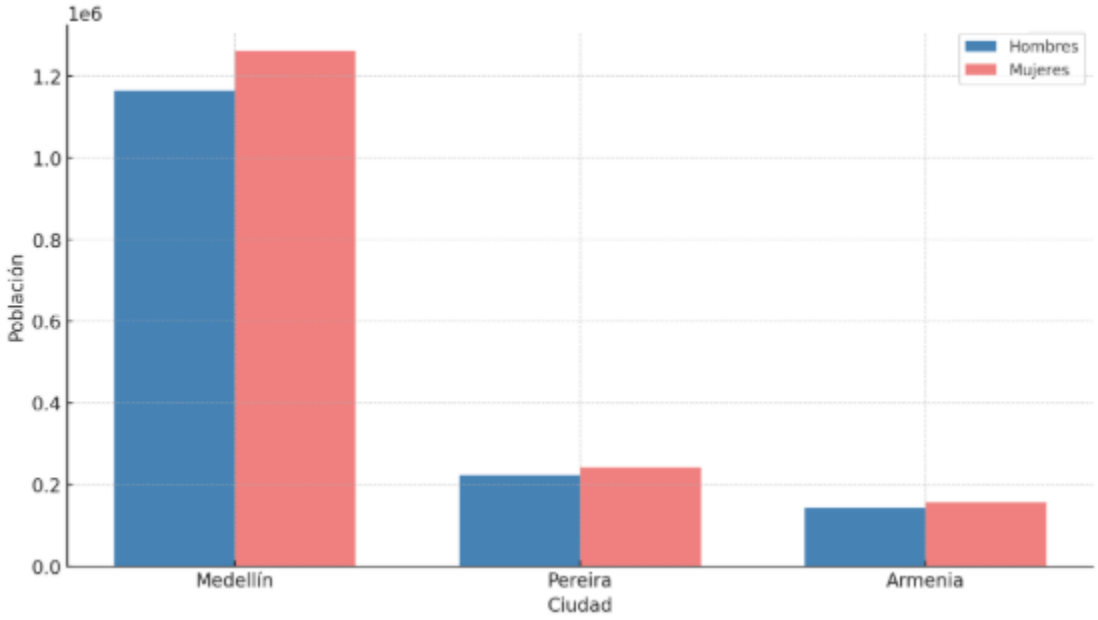
Según el Departamento Administrativo Nacional de Estadística (DANE, 2018) en el Censo Nacional de Población y Vivienda 2018 se obtuvieron las siguientes cifras:

- Medellín 2.427.129 habitantes.
- Pereira 467.269 habitantes.
- Armenia 301.226 habitantes.

En el que por sexo se visualiza en el siguiente histograma que la mayoría de población corresponde al genero femenino.

Figura 1

Población por Sexo en las Capitales de los 3 Departamentos Seleccionados



*Nota.*Tomada de DANE, 2018.

Teniendo en cuenta esto se filtra el cáncer en cuestión a estudiar con información de la tasa de mortalidad.

Tasa de Mortalidad

La tasa de mortalidad cruda por tipo de cáncer, por 100.000 personas-año, es un indicador epidemiológico que mide la cantidad de muertes causadas por un tipo específico de cáncer en una población determinada durante un período de tiempo (Marcus, 2019). Sin ajustar por factores como la edad o el sexo, se determina mediante la ecuación:

$$\boldsymbol{TM_x = \left(\frac{F_x}{P_X}\right) * 10^n}$$

Donde,

$TM_x = Tasa\ de\ mortalidad\ dentro\ de\ un\ grupo\ X.$

$F_x = Número\ de\ fallecimientos\ dentro\ de\ un\ conjunto\ X\ en\ el\ período\ de\ tiempo\ considerado.$

$P_X = Número\ total\ de\ personas\ en\ el\ conjunto\ X.$

$n = Potencia\ del\ factor\ conversión\ para\ expresar\ en\ unidades\ en\ nuestro\ caso\ n = 5.$

Para nuestro caso *X*= Tipo de Cáncer, por lo que expresará cuántas personas fallecen por un tipo específico de cáncer en una población de 100,000 habitantes en un año y permitiendo comparar la carga de mortalidad entre diferentes tipos de cáncer o regiones.

Según estudios epidemiológicos, estas neoplasias han mostrado un crecimiento alarmante en las últimas décadas, lo que refleja una problemática que va más allá de factores individuales y abarca aspectos ambientales, socioeconómicos y demográficos. Entre los posibles factores de este tipo de cáncer podemos encontrar el tabaquismo, la contaminación del aire, exposición ocupacional, la urbanización e industrialización, los hábitos de vida y alimentación y por supuesto que factores genéticos en menores proporciones.

Algunos de los factores ambientales juegan un papel crucial, ya que las principales ciudades están expuestas a la contaminación del aire, alérgenos y condiciones climáticas que pueden exacerbar las enfermedades respiratorias, estudios hechos en Ámsterdam enuncian que las personas que vivían en ese momento cerca de carreteras o zonas de mucho tránsito tenían un 35 % de probabilidad de tener más exposición producto del hollín, que los que habitaban en un lugares apartados o rurales (Van Roosbroeck et al., 2006).

Otros aspectos a tener en cuenta de vital importancia están relacionados a los estilos de vida y comportamientos de salud, como el tabaquismo, la actividad física y la dieta, a tal punto que son objetivos claves para los programas de prevención, en el caso puntual del tabaco según el Ministerio de Salud de Colombia (MinSalud, 2023), cada año mueren más de 34.800 personas por enfermedades atribuibles al tabaquismo, lo que es un mal endémico que no se regula fuertemente dado a ser una sustancia psicoactiva legal.

En el proceso nos encontramos con otras variables que poseen cierta importancia a la hora de hacer un análisis exploratorio detallado e incluso uno predictivo de clasificación para predecir la supervivencia del paciente (estado vital), cuya información de entrenamiento además de la edad y el sexo encontramos la topografía (Indica el sitio o ubicación del tumor), esta localización es codificada siguiendo la tercera edición de la clasificación internacional de enfermedades oncológicas (CIE-O-3), Base de diagnóstico que representa la fuente de donde fue diagnosticado el tumor (Diagnóstico por Certificado de Defunción, Clínico, verificación microscópica,etc), Comportamiento (que representa si el tumor el maligno, benigno, incierto, entre otros) y el CIE-10 que es un sistema de codificación desarrollado por la Organización Mundial de la Salud (OMS) donde en nuestro caso se tienen en cuenta C33 – Tumor maligno de la tráquea y C34 – Tumor maligno de los bronquios y del pulmón.

Ahora bien, desde la ciencia de datos ya se ha venido registrando importantes estudios asociados a aportar en el campo de la salud mediante algoritmos o modelos que brinden al profesional nuevas destrezas en sus labores cotidianas, entre algunos informes relacionados a las variables de estudio se tiene en el campo del Big Data iniciativas que permitirán que se mejore el diagnóstico, tratamiento y sobre todo pronósticos de las patologías que se enfrentan los pacientes, esto con modelos como el de Ram et al. (2015), que dice que es capaz de predecir en un tiempo casi que real la cantidad de pacientes con asma que podrían acudir a urgencias con un 70% de precisión, rigiéndose en gran parte por las búsquedas en diversos medios como Google, Twitter, distintas redes sociales y sensores ambientales, siendo desde su momento de creación empleado para validar si dicho modelo puede alcanzar una gran efectividad.

En lo que respecta a temas de contaminación del aire se ha empleado el Machine Learning y el modelo de Redes Neuronales Artificiales para predecir el comportamiento que toma la variable PM2.5 y así generar medidas de sean entre otras de tipo preventivo para las enfermedades respiratorias, en el que se diagnostiquen a partir de valores que presenten altos índices de contaminación (Garcia, 2021) y similarmente para casos de consumo de ciertas sustancias se ha visto estudios predictivos, cuyo entre tantos enfoques caben destacar el planteado por Montaño et al. (2014), que desde diversas variables psicosociales y de personalidad plantea determinar el consumo o detección de nicotina en cierta población con el uso de técnicas de clasificación que usualmente proceden del Data Mining.

En el presente trabajo se buscará por medio de series temporales predecir la tasa de mortalidad del Cáncer en los próximos años mediante diversos modelos definiendo por medio de las métricas el más efectivo y así analizar posibles tendencias con resultados que documenta el rendimiento y visualizaciones que demuestren lo optimo que es el modelo.

Las series de tiempo son secuencias de datos registrados en intervalos de tiempo específicos y ordenados cronológicamente, estas series permiten analizar patrones, tendencias y estacionalidad en diversas áreas, como economía, ventas, meteorología, y salud (Pavlyshenko, 2019). Según Coutin (2007) estas técnicas de aprendizaje supervisado del ML "ha sido empleada mundialmente en salud pública con frecuencia creciente, debido a sus bondades para modelar adecuadamente el comportamiento de los eventos médicos" y por esto la hace útil e importante para la obtención de pronósticos en cuanto al comportamiento o diagnóstico de enfermedades. Los modelos a tener en cuenta en este proyecto son:

Modelo ARIMA (Autoregressive Integrated Moving Average)

El modelo ARIMA fue desarrollado por George Box y Gwilym Jenkins en la década de 1970 con el propósito de obtener mejores pronósticos en el control de la contaminación (Box y Jenkins,1994), combina tres componentes principales:

AR (AutoRegresivo). Usa valores pasados de la serie para predecir futuros.

I (Integrado). Diferencia la serie para hacerla estacionaria.

MA (Media Móvil). Usa errores pasados en la predicción.

Estos componentes se asocian con 3 parámetros (p,d,q) que se determinan con algunos métodos que incluyen matemáticas avanzadas y que mediante la programación por medio de gráficos nos puede brindar una mejor comprensión:

Autocorrelación Parcial (PACF). Indica cuántos valores pasados se usan para predecir el futuro.

Dickey-Fuller Aumentada (ADF). Indica cuántas veces se necesita diferenciar la serie para hacerla estacionaria (es decir, que su media y varianza sean constantes en el tiempo) y uno último denominado.

Autocorrelación (ACF). Representa el número de términos de media móvil (errores pasados) usados en la predicción.

Se estima que una de las primeras utilizaciones de los modelos ARIMA en el ámbito sanitario se debe a Keewan Choi y Stephen Tacker en Atlanta, EE.UU., haciendo parte del grupo Epidemiología del Centro para el Control de Enfermedades (CDC) y la Universidad de Emory respectivamente, empleando este método para pronosticar el número de muertes esperadas por influenza y neumonía (Choi y Thacker, 1981).

Modelo ETS (Exponential Smoothing State Space Model)

El modelo ETS fue desarrollado en 2002 por Rob J. Hyndman, Anne B. Koehler, J. Keith Ord y Ralph D. Snyder para mejorar las técnicas de suavizamiento exponencial mediante un enfoque basado en modelos de espacio de estados, es una de las técnicas más utilizadas en la predicción de series temporales (Shi, 2021), se basa en el método de suavizamiento exponencial, que asigna diferentes pesos a los valores pasados de una serie de datos, dando mayor importancia a los más recientes, como indica Hyndman et al. (2008) en cuanto el término “suavizado exponencial” refleja el hecho de que los pesos disminuyen exponencialmente a medida que las observaciones se hacen más antiguas.

Modelo Holt-Winters (Triple Exponential Smoothing)

El modelo Holt-Winters fue desarrollado por Charles C. Holt en 1957 y extendido por Peter R. Winters en 1960 para pronosticar series de tiempo con tendencia y estacionalidad. Su enfoque se basa en tres componentes: nivel, que representa el valor base de la serie; tendencia, que captura cambios a lo largo del tiempo; y estacionalidad, que modela patrones repetitivos en períodos fijos (Educohack Press 2025).

Tabla 1

Aspectos Clave del Análisis del Cáncer desde la Epidemiología y la Ciencia de Datos

Dimensión	Concepto / Descripción	Fuente(s)
Impacto social y estadístico	El cáncer afecta a millones de personas cada año. Las estadísticas permiten entender su evolución y distribución.	MinSalud (2023)
Métodos de diagnóstico	Incluyen radiografías, TC, RM, PET y otros. La radiografía de tórax es clave en cánceres respiratorios.	MinSalud (2023)
Cáncer respiratorio	Incluye cáncer de pulmón, tráquea y bronquios, asociados con tabaquismo y exposición ambiental.	Rodríguez et al. (2018); Hetnał et al. (2010); Fuentes & Corona (2002)
Indicador epidemiológico clave	Tasa de mortalidad cruda por tipo de cáncer por 100.000 personas-año.	Marcus (2019)
Factores de riesgo	Contaminación del aire, tabaquismo, dieta, ocupación, genética, urbanización.	Van Roosbroeck et al. (2006); MinSalud (2023)
Análisis regional	Evaluación en Medellín, Pereira y Armenia con tasas ajustadas por población.	DANE (2018)
Aplicaciones de IA y ML	Modelos predictivos con ARIMA, ETS, Holt-Winters. Permiten anticipar tendencias de mortalidad.	Pavlyshenko (2019); Hyndman et al. (2008); Choi & Thacker (1981)
Casos de aplicación en salud	ML para predecir consultas por asma, contaminación y consumo de nicotina.	Ram et al. (2015); García (2021); Montaño et al. (2014)

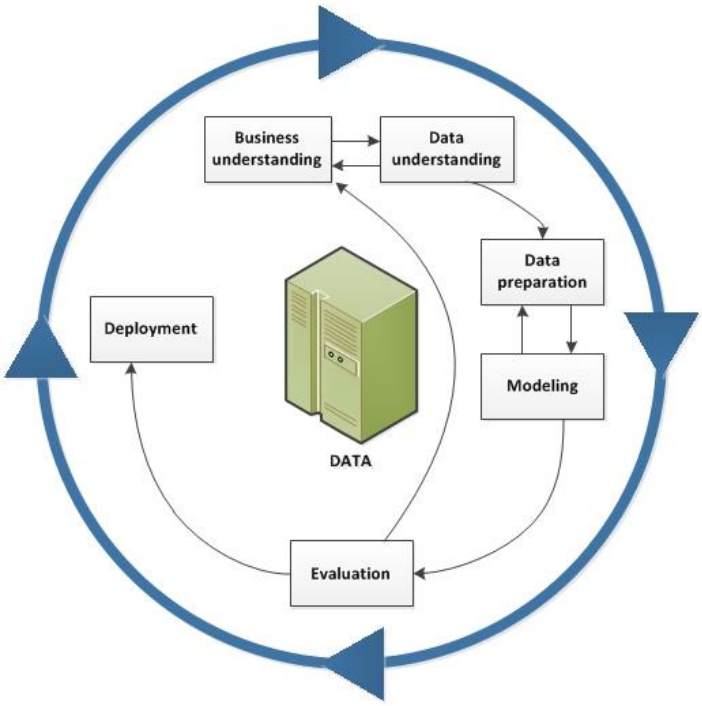
Nota. La tabla resume los principales ejes de análisis del cáncer a partir de estadísticas, estudios epidemiológicos y técnicas de predicción mediante ciencia de datos.

Metodologia

Para el desarrollo de nuestro proyecto hemos decidido utilizar la metodología CRISP-DM (Proceso de Estándar Cruzado para Minería de Datos). Es una metodología que se utiliza con amplitud en la minería de datos y en el análisis de los mismos. Fue desarrollada inicialmente para proyectos de minería de datos, pero debido a su flexibilidad se puede aplicar un amplio campo de proyectos de análisis de datos y en generar a las ciencias de datos.

Figura 2

Métodologia CRISP-DM



Nota. Tomada de IBM, 2021.

Esta metodología consta de varios pasos que se pueden aplicar a un proyecto, los cuales vamos a identificar a continuación, pero que estarán enfocados en nuestro proyecto aplicado que estamos estudiando.

Comprensión del Negocio

El objetivo de nuestro proyecto es determinar un modelo que permita predecir la tasa de mortalidad del cáncer de pulmón, tráquea y bronquios con ayuda de registros históricos del país, pero con un enfoque en tres departamentos y analizar su tendencia.

Comprensión de Datos

Realizamos una recopilación de datos asociados a procedimientos de diagnóstico de cáncer, tasa de mortalidad por tipo de cáncer en todo el país, por medio del Instituto Nacional de Cancerologia de Colombia y los datos abiertos del Ministerio de Salud.

Preparación de los Datos

Con los diferentes datasets encontrados para llevar a cabo la investigación del proyecto procedimos a realizar la respectiva limpieza y transformación de datos, donde realizaremos la eliminación de datos incompletos o incorrectos y la integración de los diferentes conjuntos de datos, los cuales se implementaran en un solo dataset, donde se integraran los datos que tengan la información más relevante, en el que nos permitan hacer una investigación más minuciosa del proyecto y así poder implementarlo en nuestra predicción. También es importante realizar tareas como imputación de datos faltantes, codificación de variables categóricas y normalización.

Modelado

En esta fase ya tenemos nuestros datos limpios y listos para realizar un análisis exploratorio de los datos (EDA), aplicando el modelo de serie de tiempo ARIMA (Autoregressive Integrated Moving Average) en donde se requiere de 3 parámetros para usar valores pasados de la serie, predecir valores futuros, hacer que la serie sea estacionaria (sin tendencias o cambios drásticos) y usar el error de predicciones pasadas para mejorar la

predicción futura, entre otras técnicas estadísticas y otros modelos como el ETS (Exponential Smoothing State Space Model) y el Holt-Winters (Triple Exponential Smoothing).

Evaluación

Para esta fase, evaluamos el modelo y los resultados obtenidos en el análisis. Esto implica validar la precisión del modelo final, mediante distintas métricas que nos indique la calidad del modelo y también realizar ajustes de los hiperparámetros según sea necesario.

Despliegue

El fin de esta fase es comunicar los resultados del análisis de manera efectiva a todas las partes interesadas, como los evaluadores del proyecto, profesores y comunidad estudiantil en general, lo que podría implicar también la elaboración de informes técnicos, exposiciones, presentación de resultados en conferencias o seminarios y posterior publicación de dichos resultados.

Documentación

Documentación de todo el proceso desde su etapa inicial hasta su implementación, donde se pueda evidenciar como se realizó cada proceso del proyecto.

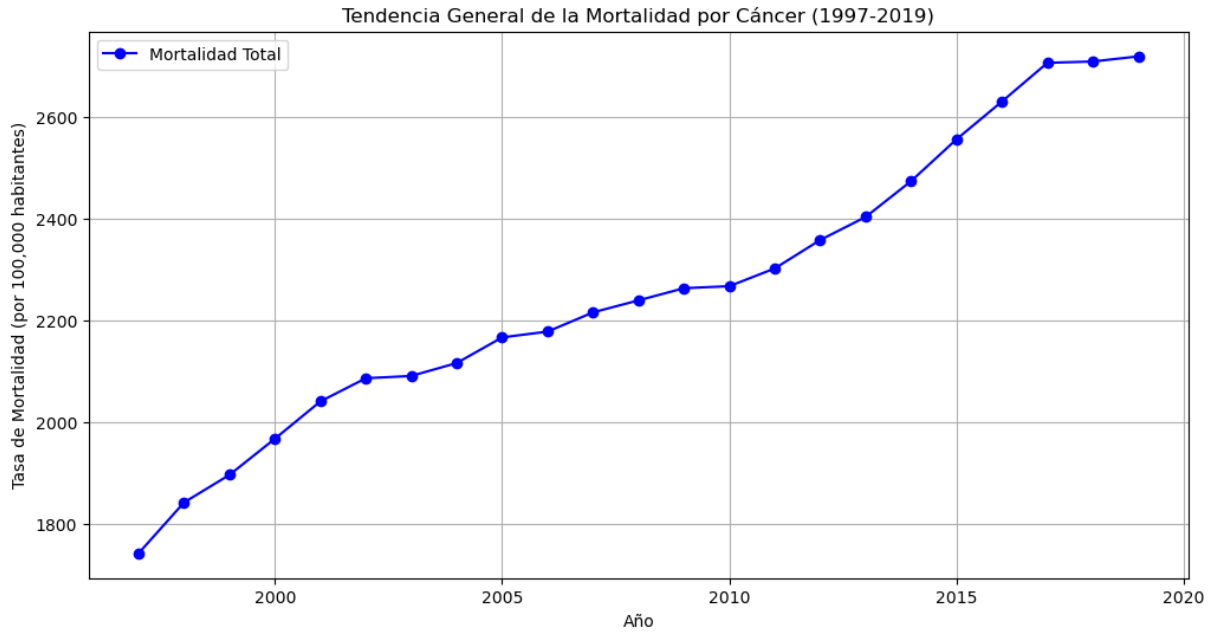
Es importante mencionar que, dado el carácter sensible de los datos de salud y la necesidad de proteger la privacidad de las personas, se debe cumplir a cabalidad con todas las regulaciones y estándares éticos relevantes en la recopilación, comunicación y análisis de los datos epidemiológicos. Si seguimos estos pasos para la implementación de la metodología CRISP-DM, podremos llevar a cabo un análisis predictor efectivo del cáncer en cualquier ciudad del país, y así aportar al entendimiento y abordaje de este importante problema de salud pública.

Resultados

La tasa de mortalidad en el país en cuanto al cáncer ha ido en aumento cada año en la Figura 3, se puede apreciar como desde el 1997 se ha presentado un aumento significativo por cada 100.000 habitantes hasta el 2019.

Figura 3

Tendencia de la Tasa de Mortalidad del Cáncer por Año



Teniendo en cuenta esta creciente se decidió buscar un tipo de cáncer que estuviera asociado a instrumentos que emplearan radiación para su diagnostico en el cual en un top 10 con valores porcentuales como se ve en la Figura 4, vemos que a pesar de que el electrocardiograma es uno de los más utilizados al no emplear radiación elegimos el siguiente de la lista que corresponde a Radiografia de Torax con un 15.1% cuyo principal diagnostico afecta al cáncer de pulmón, traquea y bronquios siendo el suficiente motivo para elegir y profundizar en este tipo de patología.

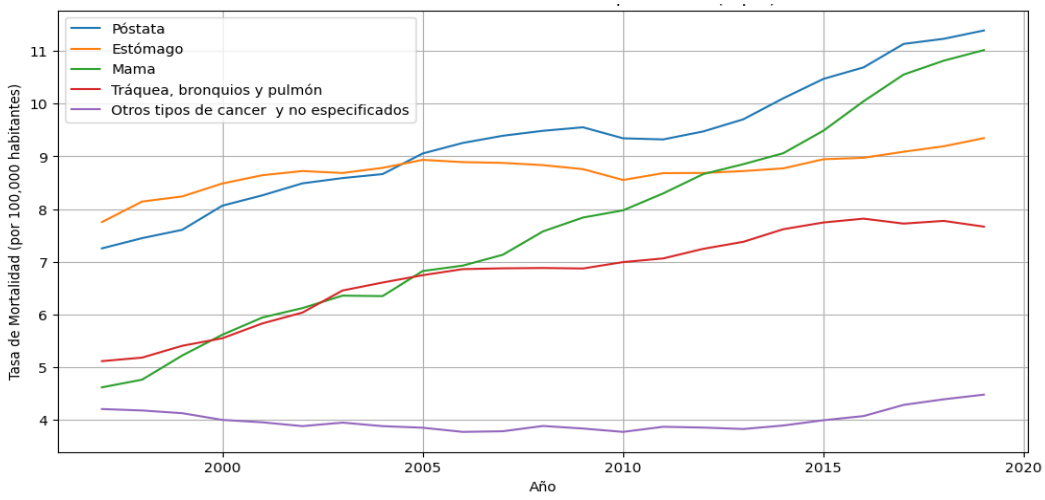
Figura 4

Procedimientos de Imagen Diagnostica más Aplicados

Una vez dada esta elección se visualizó la evolución de manera más puntual por tipo de cáncer a nivel nacional en cuanto a la tasa de mortalidad cada año, notando que más allá de que el procedimiento de radiografia de torax sea uno de los más empleados, este en cuanto a enfermedades cancerígenas está en un cuarto lugar detrás de otros tipos como cáncer de Próstata, Estómago y Mama.

Figura 5

Evolución de la Tasa de Mortalidad por tipo de Cáncer

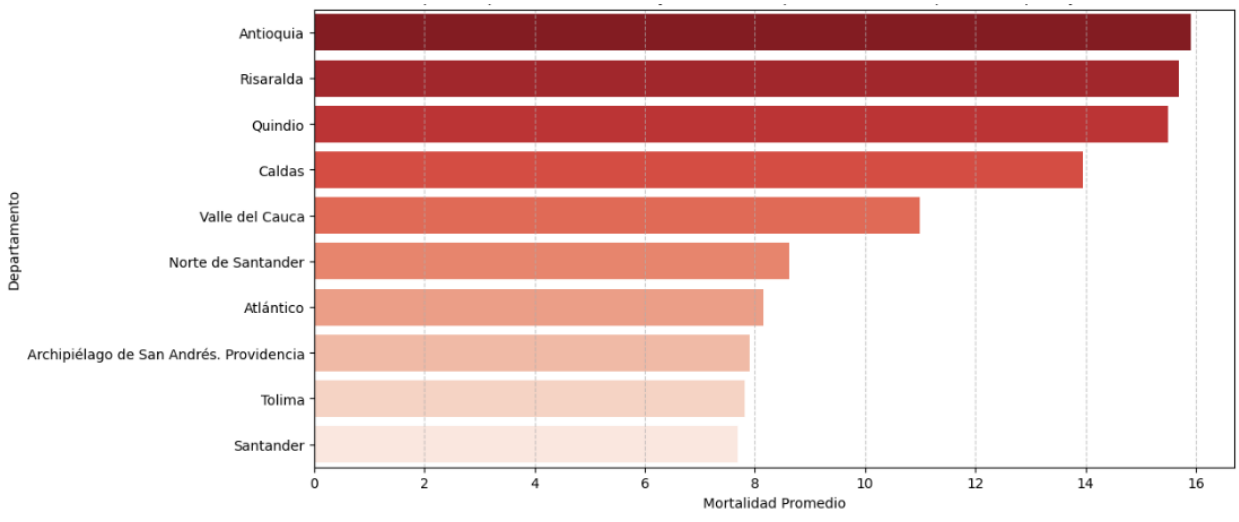


Vemos que en lo que respecta la evolución de la tasa de mortalidad por cáncer de tráquea, bronquios y pulmón (línea roja) que a lo largo del tiempo se observa una tendencia claramente ascendente, lo que indica un incremento sostenido en el número de fallecimientos asociados a esta enfermedad, por otro lado a partir del 2015 vemos un patrón en el que la curva muestra signos de estabilización con un ligero decrecimiento en los últimos años del análisis, lo que podría sugerir efectos de intervenciones médicas, campañas de prevención o mejoras en los tratamientos pero que no termina de sentenciar el posible rumbo de las futuras tendencias como se podria decir de los otros 4 casos.

Segmentando la base de datos para el tipo de cáncer elegido se hizo un diagramas de barra horizontal con el top 10 de los departamentos que en promedio más se han visto afectados por esta enfermedad, siendo Antioquia, Risaralda y Quindío con una tasa de mortalidad superior a las 14 unidades por 100.000 habitantes

Figura 6

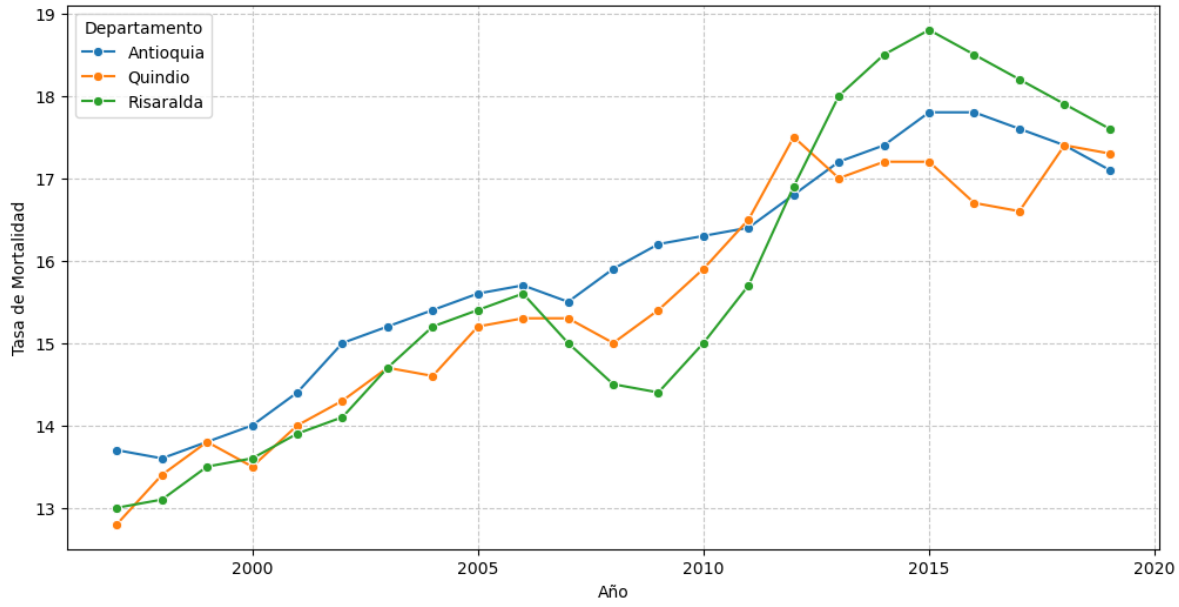
Top 10 de Departamentos con Mayor Tasa de Mortalidad



De acuerdo a lo anterior se representaron mediante diagramas de lineas los 3 departamentos en el que cada línea permite analizar su comportamiento a lo largo del tiempo y compararlos entre sí.

Figura 7

Comparativa de la Tendencia en los 3 Departamentos Seleccionados a lo Largo de los Años



El estudio permite ver que desde finales de los años 90 hasta 2015, la tasa de mortalidad en los tres departamentos muestra una tendencia ascendente, reafirmando que este tipo de cáncer ha sido un problema creciente en la población. En términos absolutos, la tasa de mortalidad pasó de aproximadamente 13 a más de 18 por 100,000 habitantes en algunos casos, lo que representa un incremento significativo.

Se observó un pico máximo en la tasa de mortalidad alrededor del año 2015, particularmente en Risaralda, donde la mortalidad alcanzó un valor superior a 18 casos, esto más allá que su población sea muy inferior a departamentos como el de Antioquia, a partir de este punto, la tendencia comienza a disminuir ligeramente, lo que podría estar relacionado con

mejoras en los tratamientos, estrategias de prevención o reducción de factores de riesgo como el tabaquismo entre otros. En cuanto a Quindío su evolución es similar a la de Antioquia, aunque con ligeras oscilaciones.

Con el objetivo de encontrar el modelo con mejor precisión, se implementaron tres enfoques de series temporales: ARIMA, ETS y Holt-Winters, evaluando su desempeño en cada departamento de interés, por lo que se extraen los datos históricos de mortalidad y los años correspondientes, esto permitiendo trabajar con información específica de cada región y analizar tendencias individuales en la tasa de mortalidad. Posteriormente, los datos se dividen en conjuntos de entrenamiento y prueba, utilizando un 75% de los datos para entrenar los modelos y el 25% restante para evaluar su precisión. Este procedimiento garantiza que los modelos sean evaluados en datos no vistos, reflejando mejor su capacidad de generalización.

El primer modelo implementado es ARIMA (AutoRegressive Integrated Moving Average), que es ampliamente utilizado para series temporales debido a su capacidad de capturar patrones estacionales y tendencias, utilizando inicialmente un orden de (2,1,2), lo que significa que el modelo emplea dos rezagos autorregresivos, una diferenciación para hacer la serie estacionaria y dos rezagos en el promedio móvil.

El segundo modelo implementado es ETS (Exponential Smoothing State Space Model), una técnica basada en suavizamiento exponencial que permite modelar tendencias y estacionalidad en los datos, estableciendo una estructura con tendencia y estacionalidad aditiva, lo que significa que el modelo asume que las variaciones en la serie son consistentes a lo largo del tiempo.

Y por último el modelo Holt-Winters que es una variante de suavizamiento exponencial que incorpora tendencia y estacionalidad multiplicativa, asumiendo que la estacionalidad tiene

un comportamiento multiplicativo, lo que implica que las variaciones pueden crecer o disminuir en proporción al nivel de la serie.

Una vez esto se generaron las métricas para la evaluación del mejor modelo como se observa en la Tabla 2.

Tabla 2

Métricas de Error para los Modelos de Predicción en Cada Departamento

Departamento	Modelo	MAE	MSE	RMSE	MAPE
Antioquia	ARIMA	0.4616	0.4326	0.6577	2.6662
Antioquia	ETS	0.5444	0.5779	0.7602	3.1442
Antioquia	Holt-Winters	0.5455	0.5809	0.7621	3.1505
Risaralda	ARIMA	0.3978	0.2120	0.4604	2.2070
Risaralda	ETS	3.6394	18.4807	4.2989	20.2125
Risaralda	Holt-Winters	3.6469	18.5879	4.3114	20.2550
Quindío	ARIMA	0.7063	0.6952	0.8338	4.1655
Quindío	ETS	0.7237	0.7247	0.8513	4.2654
Quindío	Holt-Winters	0.7294	0.7225	0.8500	4.2972

Nota. MAE = Error Absoluto Medio, MSE = Error Cuadrático Medio, RMSE = Raíz del Error Cuadrático Medio, MAPE = Error Porcentual Absoluto Medio.

En cuanto al desempeño de los modelos por departamentos tenemos que:

Para Antioquia el modelo ARIMA obtiene los menores valores de error, lo que indica que es el modelo más preciso para este departamento mientras que los modelos ETS y Holt-Winters

presentan errores más elevados y valores similares, lo que indican que no son tan efectivos para capturar la tendencia de la serie temporal.

Por parte de Risaralda el modelo ARIMA vuelve a ser el modelo con mejor desempeño ya que los demás presentan errores significativamente altos, con MSE superiores a 18 y RMSE alrededor de 4.3, indicando una menor capacidad para modelar correctamente la serie de tiempo, motivo que pueden estar sobreajustando o no capturando adecuadamente la estructura de la serie.

Y por último se termina de consolidar para el Quindío el modelo ARIMA por sus valores más bajos de error aunque los otros modelos tienen errores ligeramente más altos. Aunque la diferencia entre los modelos no es tan marcada como en Risaralda, ARIMA sigue destacando por su mayor precisión.

Por lo tanto, para futuras predicciones de la mortalidad por cáncer de tráquea, bronquios y pulmón, se eligió ARIMA como el modelo más confiable, procediendo a ajustar correctamente sus parámetros.

En la Tabla 3 vemos que el estadístico ADF obtenido es de 1.8073, lo que indica que la serie no muestra evidencia suficiente para rechazar la hipótesis nula (H0), que postula la presencia de una raíz unitaria, haciendo que la serie no sea estacionaria. Así mismo el p-valor es de 0.9984, significativamente mayor al umbral de 0.05, lo que refuerza que la serie tiene una tendencia y no es estacionaria y entonces se procedió a hacer una diferenciación haciendo que el parámetro d=1. Los valores críticos sugieren que para rechazar H0, el valor estadístico ADF debería ser menor que -3.1271 (umbral del 5%) y como el valor calculado es positivo se deduce que estamos lejos de cumplir esta condición.

Tabla 3

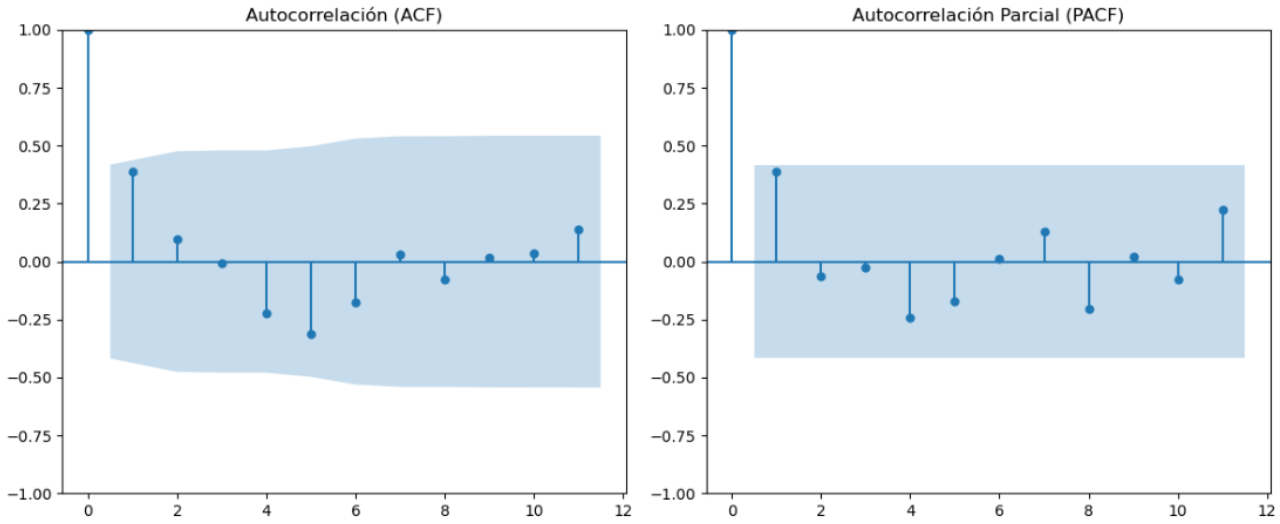
Resultados de la Prueba de Dickey-Fuller Aumentada (ADF)

Parámetro	Valor
Estadístico ADF	1.8073
p-valor	0.9984
Lags usados	9
Observaciones usadas	13
Valor crítico (1%)	-4.0689
Valor crítico (5%)	-3.1271
Valor crítico (10%)	-2.7017

Una vez dado el parámetro “d” con una sola diferenciación se obtuvieron las siguientes gráficas de autocorrelación.

Figura 8

Autocorrelación ACF y PACF para los Párametros p y q del Modelo ARIMA



Para seleccionar los parámetros “p” y “q” en el modelo ARIMA, observamos las gráficas donde en la gráfica de ACF (izquierda), observamos que hay una caída rápida a valores no significativos después del primer o segundo rezago, lo que permite determinar que la serie tiene un componente de promedio móvil (MA) de bajo orden, es decir que “q” podría ser 1 o 2.

Por otra parte en la gráfica de PACF (derecha), hay un corte claro después del primer rezago, lo que indica que la serie tiene un componente autorregresivo (AR) con un orden de p=1. Dado el anterior análisis se decidió que:

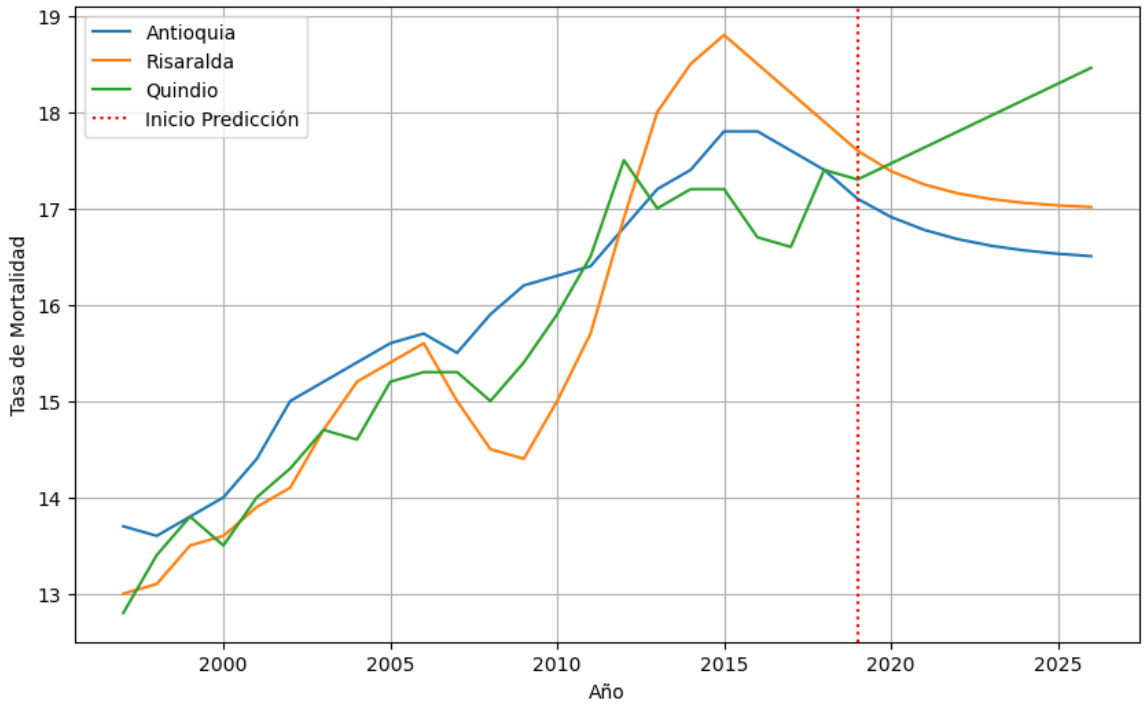
- p=1 - Porque la PACF muestra un corte claro después del primer rezago.
- q=1 - Porque la ACF muestra una caída rápida.
- d=1 – Porque se verificó con una prueba de raíz unitaria si la serie ya es estacionaria o si necesita diferenciación y como se necesitó de una diferenciación dado a que tenia un valor de p-valor superior a 0.05 se tuvo que hacer el respectivo procedimiento para que fuera una serie estacionaria.

De tal modo que los parámetros fueron (1,1,1).

Por lo que se predicen valores futuros y se visualizan en un gráfico con el cambio en la serie a partir de 2020 hasta 2026, como se aprecia en la Figura 9 donde se permite comparar la evolución de la mortalidad en los tres departamentos y analizar tendencias.

Figura 9

Predicción mediante el modelo ARIMA de la tasa de mortalidad por Cáncer de Tráquea, Bronquios y Pulmón hasta 2026.



Para el departamento de Antioquia vemos que antes de 2019, la tasa de mortalidad muestra una tendencia creciente hasta aproximadamente 2015, seguida de una leve disminución, luego la predicción proporciona una ligera caída o estabilización en los próximos años, lo que podría indicar un posible control en la mortalidad o limitaciones del modelo para capturar nuevas dinámicas.

Risaralda hasta 2019 mostró una tasa de mortalidad con muchas variaciones, con un pico alrededor de 2015 y luego un descenso en el que la proyección muestra una tendencia descendente, similar a lo ocurrido con Antioquia pero con una tasa que llega a un valor casi constante de 17 personas.

Por último para Quindío se observa una gran cantidad de fluctuaciones antes de 2019, con un incremento fuerte entre 2010 y 2015, seguido de una caída que no determinó una disminución continua ya que a diferencia de los otros departamentos, la predicción indica una tendencia creciente, lo que podría ser preocupante si se materializa, es posible que el modelo capture un patrón de recuperación después de la caída previa, pero habría que evaluar si hay factores externos que expliquen esta tendencia al alza con una pendiente de crecimiento exponencial para los siguientes años.

Los valores númericos de la predicción se pueden verificar en la Tabla 4.

Tabla 4

Predicción de la Tasa de Mortalidad por Cáncer de Tráquea, Bronquios y Pulmón

Año	Antioquia	Risaralda	Quindío
2019	16.91	17.39	17.47
2020	16.78	17.25	17.63
2021	16.68	17.16	17.80
2022	16.61	17.10	17.96
2023	16.56	17.06	18.13
2024	16.53	17.03	18.29
2025	16.50	17.01	18.46
2026	16.49	17.00	18.63

Nota. Los valores representan la predicción de la tasa de mortalidad por cáncer de tráquea, bronquios y pulmón en tres departamentos de Colombia para el período 2019-2026.

Conclusiones

El cáncer de pulmón, tráquea y bronquios representa un porcentaje significativo dentro de las tasas de mortalidad por enfermedades oncológicas en Colombia. Se empleó tres modelos estadísticos ampliamente utilizados en la predicción de series temporales siendo evaluados con base en métricas de error para determinar que el modelo ARIMA fue el más adecuado, proporcionando estimaciones más precisas en comparación con los otros métodos analizados con valores de sus parámetros de (1,1,1) permitiendo así que cada predicción se base en el último valor observado del conjunto de datos, su tendencia y el error previo .

Los hallazgos revelaron que, en promedio, Antioquia es el departamento con la tasa de mortalidad más alta para este tipo de cáncer. Sin embargo, se evidenció que Risaralda ha alcanzado picos de hasta 19 muertes por cada 100.000 habitantes, aunque a partir del año 2020 ha mostrado una tendencia descendente. En contraste, Quindío presenta una proyección de incremento en la mortalidad, lo que sugiere la necesidad de medidas preventivas y estrategias de intervención efectivas para mitigar el impacto de la enfermedad en esta región.

Este análisis comparativo demostró la utilidad de los modelos de series de tiempo en la predicción de la tasa de mortalidad por cáncer de pulmón, tráquea y bronquios pero indicando que a través del uso del Machine Learning y técnicas como ARIMA, ETS y Holt-Winters, se puedo identificar la evolución de las tasas de mortalidad de cualquier enfermedad que se desee analizar en nuestro ente territorial, brindando información clave para la formulación de políticas públicas en salud. La evidencia sugiere que, si bien algunos departamentos han mostrado descensos en las tasas de mortalidad, otros presentan tendencias al alza, lo que resalta la necesidad de implementar estrategias de intervención focalizadas.

Recomendaciones

El estudio destaca la importancia de la predicción de la mortalidad por cáncer como un insumo clave para la planificación de políticas en el sector salud. La implementación de modelos de series de tiempo permite anticipar tendencias y facilitar la toma de decisiones fundamentadas en datos. Con base en los resultados obtenidos, se recomienda a las autoridades de salud fortalecer los programas de prevención enfocadas en los factores de riesgo, especialmente el consumo de tabaco y la exposición a contaminantes, con búsqueda de obtener más inversión en la salud pública, que permita dirigir recursos a los departamentos con mayor riesgo proyectado y mejorando el acceso a diagnóstico temprano, tratamientos oncológicos y programas de seguimiento. Asimismo, es fundamental continuar explorando nuevas metodologías basadas en inteligencia artificial y aprendizaje automático para mejorar la precisión de los modelos predictivos.

Referencias Bibliográficas

Adeyinka, D. A., & Muhajarine, N. (2020). Time series prediction of under-five mortality rates for Nigeria: Comparative analysis of artificial neural networks, Holt-Winters exponential smoothing and autoregressive integrated moving average models. *BMC Medical Research Methodology*, 20, 292. https://doi.org/10.1186/s12874-020-01159-9

Altuhaifa, F. (2023). Time Series Prediction of Lung Cancer Death Rates on the Basis of SEER Data. *JCO clinical cancer informatics*, 7, e2300011. https://doi.org/10.1200/CCI.23.00011

Alarcón, M.L., Brugés, R., Carvajal, C., Vallejo, C., & Beltrán, R. (2021). Características de los pacientes con cáncer de pulmón de célula no pequeña en el Instituto Nacional de Cancerología de Bogotá. *Revista Colombiana de Cancerología*, 25. https://doi.org/10.35509/01239015.706

Andújar-Vera, F., García-Fontana, C., González-Salvatierra, S., Martínez-Heredia, L., Muñoz-Torres, M., García-Fontana, B. (2021). Relación genética entre las enfermedades pulmonares de origen ambiental u ocupacional y la osteoporosis: un enfoque bioinformático. *Revista de Osteoporosis y Metabolismo Mineral*, 13(4), 130-136. https://dx.doi.org/10.4321/s1889-836x2021000400005

Bellinger, C., Mohomed-Jabbar, M., Zafane, O., Osornio-Vargas, A. (2017). A systematic review of data mining and machine learning for air pollution epidemiology. *BMC Public Health*. https://doi.org/10.1186/s12889-017-4914-3

Box G, Jenkins G, Reinsel G. Time Series Analysis: Forecasting and Control. 3th ed. Canada: Prentiuce Hall Canada;1994.

Choi, K., & Thacker, S. B. (1981). An evaluation of influenza mortality surveillance, 1962-1979. I. Time series forecasts of expected pneumonia and influenza deaths. *American journal of epidemiology*, 113(3), 215–226. https://doi.org/10.1093/oxfordjournals.aje.a113090

Coutin, M. G., (2007). Use of ARIMA models for communicable disease surveillance. *Revista Cubana de Salud Pública*.

Departamento Administrativo Nacional de Estadística (2018). *Estadísticas Demografía y Población.* Recuperado de https://www.dane.gov.co/index.php/estadisticas-por-tema/demografia-y-poblacion/proyecciones-de-poblacion

Esendağli, D., Yilmaz, A., Akçay, Ş., y Özlü, T. (2021). Post-COVID syndrome: pulmonary complications. *Turkish journal of medical sciences*, 51. https://doi.org/10.3906/sag-2106-238.

Fuentes-Valdés, E., y Corona-Mancebo, S. B. (2002). Tumores primarios malignos de tráquea y bronquios principales. *Revista Cubana de Cirugía*, 41(3), 176-184

Garcia, D. (2021) Artificial neural network prototype for the forecast of critical events due to PM 2.5 particles in the center of the city of Manizales. *Fundación Universitaria Los Libertadores*. https://repository.libertadores.edu.co/server/api/core/bitstreams/ad9ae5e0-3056-49b9-887a-55c8cfd44886/content

Hetnał, M., Kielaszek-Ćmiel, A., Wolanin, M., Korzeniowski, S., Brandys, P., Małecki, K., ... & Kokoszka, A. (2010). Tracheal cancer: Role of radiation therapy. *Reports of Practical Oncology and Radiotherapy*, 15(5), 113-118.

Hyndman, R. J., & Khandakar, Y. (2008). Automatic time series forecasting: The forecast package for R. *Journal of Statistical Software*, 27(3), 1–22. https://doi.org/10.18637/jss.v027.i03

Introduction to Time Series Analysis. (2025). (n.p.): Educohack Press.

Kaplan, A., Cao, H., FitzGerald, J. M., Iannotti, N., Yang, E., Kocks, J. W. H., Kostikas, K., Price, D., Reddel, H. K., Tsiligianni, I., Vogelmeier, C. F., Pfister, P., Mastoridis, P. (2021). Artificial Intelligence/Machine Learning in Respiratory Medicine and Potential Role in Asthma and COPD Diagnosis. *The journal of allergy and clinical immunology*. 9(6), 2255–2261. https://doi.org/10.1016/j.jaip.2021.02.014

Llamas O. A., (2022). Reflexiones históricas sobre la calidad en medicina nuclear en Colombia: regulación normativa y dinámica del mercado. *Revista Colombiana de Cancerología*. 26, 2, 124–126.

López-Campos, J. L., Tan, W., & Soriano, J. B. (2016). Global burden of COPD. *Respirology (Carlton, Vic.), 21(1),* 14–23. https://doi.org/10.1111/resp.12660

Ministerio de Salud y Protección Social. (2023). Estado de la salud pública en Colombia: Informe anual. Recuperado de https://www.minsalud.gov.co/Paginas/Cada-anno-mas-de-34-800-muertes-en-Colombia-estan-relacionadas-con-el-consumo-de-productos-de-tabaco.aspx

Montaño, J., Gervilla, E., Cajal, B., & Palmer, A. (2014). Técnicas de clasificación de data mining: una aplicación al consumo de tabaco en adolescentes. *Anales de Psicología / Annals of Psychology*, 30(2), 633–641. https://doi.org/10.6018/analesps.30.2.160881

Pavlyshenko, B. M. (2019). Machine-learning models for sales time series forecasting. Data, 4(1), 15.

Ram S., Zhang W., Williams M., & Pengetnze Y. (2015) Predicting asthma-related emergency department visits using big data. *IEEE J Biomed Heal Informatics*. 1216–1223. https://doi.org/10.1109/jbhi.2015.2404829

Rodriguez-Serret, J. E., García-Gómez, O., Salcedo-Quintero, S., Rosell Nicieza, I., & Pons Porrata, L. (2018). Caracterización clínica, tomográfica e histopatológica de pacientes con cáncer de pulmón. *MediSan*, 22(9), 1138-1147.

Shi, Y. (2021). Forecasting mortality rates with the penalized exponential smoothing state space model. *Journal of the Operational Research Society*, 73(5), 955–968. https://doi.org/10.1080/01605682.2021.1892465

Tabio-Lage, A., Collado-Otero, J., Gómez-Trueba, G., & Ropero-Toirac, R. (2021). Supervivencia global de pacientes con carcinoma del pulmón de células no pequeñas. *Revista Cubana de Oncología*, 19(2). Recuperado de https://revoncologia.sld.cu/index.php/onc/article/view/96

Van Roosbroeck, S., Wichmann J., Janssen N., Hoek G., van Wijnen J., Lebret E., et al. (2006) Long-term personal exposure to traffic-related air pollution among school children, a validation study. *Sci Total Environ*. 565-73. http://dx.doi.org/10.1016/j.scitotenv.2006.03.034

Vergara-Dagobeth, E., Suárez-Causado, A., y Gómez-Arias, R. (2017). Plan Control del cáncer en Colombia 2012-2021. Un análisis formal. *Revista Gerencia y Políticas de Salud*, 16(33), 6-18. https://doi.org/10.11144/javeriana.rgps16-33.pccc

Wang, C., Wu, Z., Xu, Y., Zheng, Y., Luo, Z., Cao, W., Wang, F., Dong, X., Qin, C., Zhao, L., Xia, C., Tan, F., Chen, W., Li, N., & He, J. (2023). Disparities in the global burden of tracheal, bronchus, and lung cancer from 1990 to 2019. *Chinese medical journal pulmonary and critical care medicine*, 1(1), 36–45. https://doi.org/10.1016/j.pccm.2023.02.001

Wiesner, C. (2018). Salud pública y epidemiología del cáncer en Colombia. *Colombia Médica*, 49(1), 13-15.

Woodcock, J., Edwards, P., Tonne, C., Armstrong, B. G., Ashiru, O., Banister, D., ... y Roberts, I. (2009). Public health benefits of strategies to reduce greenhouse-gas emissions: urban land transport. *The Lancet*. https://doi.org/10.1016/s0140-6736(09)61714-1

Zhang, B., Shi, H., & Wang, H. (2023). Machine Learning and AI in Cancer Prognosis, Prediction, and Treatment Selection: A Critical Approach. *Journal of multidisciplinary healthcare*, 16, 1779–1791. https://doi.org/10.2147/JMDH.S410301

Zhou, B., Bie, F., Zang, R., Zhang, M., Song, P., Liu, L., Peng, Y., Bai, G., Huai, Q., Li, Y., Zhao, L., y Gao, S. (2022). Worldwide burden and epidemiological trends of tracheal, bronchus, and lung cancer: A population-based study. *eBioMedicine*, 78, 103951. https://doi.org/10.1016/j.ebiom.2022.103951

Vásquez, P. O. (2025, mayo 25). *Análisis y Predicción de la tasa de Mortalidad por Cáncer de Pulmón, Tráquea y Bronquios: Un Estudio Comparativo en tres departamentos de Colombia mediante Machine Learning y Modelos de Series de Tiempo*. [Video]. YouTube. https://youtu.be/YsN_hUkHL7g