Diferenciación de categorías de causas de defunción no fetal mediante aprendizaje automático e interpretación con SHAP a partir de microdatos de defunción no fetal complementados con información contextual municipal en Santander (2015–2019)

Carlos Giovanni Durán Acevedo

Director

Luis Ángel Anillo Arrieta

Universidad Nacional Abierta y a Distancia UNAD

Escuela de Ciencias Básicas Tecnología e Ingeniería ECBTI

Maestría en Ciencia de Datos y Analítica

2026

Resumen

El presente estudio analiza el potencial de los registros de defunciones no fetales y de las variables sociodemográficas del Departamento Administrativo Nacional de Estadística (DANE) para modelar e interpretar patrones asociados a distintos grupos de causas de muerte en el departamento de Santander. Se clasificaron las defunciones en tres categorías: sistema circulatorio, neoplasias y externas, integrando información individual y contextual a nivel municipal. El enfoque es analítico e interpretativo, no orientado al despliegue operativo de los modelos. Para el análisis se implementaron algoritmos supervisados basados en árboles (XGBoost y CatBoost), comparados con un modelo lineal multinomial de referencia. El desempeño se evaluó mediante métricas como accuracy, precisión, recall y F1-score, incorporando validación cruzada estratificada para examinar la estabilidad de los resultados. Posteriormente, se aplicó el método SHAP para identificar la contribución relativa de las variables en la clasificación, permitiendo identificar patrones y perfiles asociados a las categorías analizadas, así como las dificultades de diferenciación existentes entre algunas de ellas. Los hallazgos evidencian que los datos abiertos del DANE permiten construir aproximaciones estructurales útiles para el análisis territorial de la mortalidad y para apoyar la toma de decisiones en salud pública.

***Palabras clave*:** aprendizaje automático, mortalidad no fetal, shap, microdatos, salud pública.

Abstract

This study analyzes the potential of non-fetal death records and sociodemographic variables provided by the National Administrative Department of Statistics (DANE) to model and interpret patterns associated with different groups of causes of death in the department of Santander. Deaths were classified into three categories: circulatory system diseases, neoplasms, and external causes, integrating individual-level and municipal-level contextual information. The study adopted an analytical and interpretative approach rather than focusing on the operational deployment of predictive models. For the analysis, supervised tree-based algorithms (XGBoost and CatBoost) were implemented and compared with a multinomial logistic regression model used as a baseline. Model performance was evaluated using metrics such as accuracy, precision, recall, and F1-score, incorporating stratified cross-validation to assess the stability of the results. Subsequently, the SHAP method was applied to identify the relative contribution of the variables to the classification process, enabling the identification of patterns and profiles associated with the analyzed categories, as well as the difficulties in differentiating some of them. The findings show that DANE's open-access microdata provide useful structural approximations for territorial mortality analysis and can support decision-making processes in public health.

Keywords**:** machine learning, non-fetal mortality, shap, microdata, public health.

Tabla de Contenido

Lista de Tablas

Lista de Figuras

Lista de Apéndices

Introducción

La mortalidad es uno de los acontecimientos más importantes para estimar o evaluar la salud de una población, además de que permite la formulación de políticas públicas en aspectos sanitarios. Analizar las causas de este acontecimiento permite identificar perfiles de riesgo, así como identificar patrones territoriales que afectan en la distribución de causas de muerte. En Colombia, el Departamento Administrativo Nacional de Estadística (DANE, 2004) dispone de microdatos de defunciones no fetales que, junto con información sociodemográfica y económica a nivel municipal, ofrecen una base amplia y sistemática para el estudio de la mortalidad desde una perspectiva estructural.

Tradicionalmente, estos registros han sido utilizados mediante análisis descriptivos y modelos estadísticos paramétricos, centrados en tabulaciones por edad, sexo o causa básica de muerte. Si bien estos enfoques han aportado información valiosa, pueden resultar limitados para capturar interacciones complejas entre variables individuales y contextuales. En este sentido, surge la necesidad de explorar alternativas metodológicas que permitan examinar de manera multivariable y no lineal los patrones asociados a distintos grupos de causas de muerte, particularmente en contextos regionales donde el aprovechamiento analítico de los datos abiertos aún es incipiente.

En este contexto, la presente investigación tiene como propósito evaluar el potencial de los microdatos de defunciones no fetales del DANE (2004), integrados con variables sociodemográficas individuales e indicadores contextuales municipales, para sustentar modelos supervisados de clasificación multiclase en el departamento de Santander. Específicamente, se analiza la capacidad de estos datos para diferenciar registros asociados a tres grandes grupos de causas de muerte: enfermedades del sistema circulatorio, externas y neoplasias.

El estudio adopta un enfoque cuantitativo basado en técnicas de aprendizaje supervisado, comparando un modelo lineal de referencia (regresión logística multinomial) con modelos no lineales basados en algoritmos de boosting. La evaluación no se limita al desempeño predictivo, sino que incorpora un análisis interpretativo mediante el uso de herramientas de explicabilidad, con el fin de identificar la contribución relativa de las variables en la diferenciación de los grupos de causas.

El valor agregado de esta investigación radica en la evaluación del potencial analítico de los datos abiertos oficiales para el modelamiento supervisado interpretable en un contexto regional. Más que desarrollar un sistema predictivo operativo, el estudio busca determinar si los microdatos administrativos permiten capturar parcialmente patrones multivariables diferenciables entre grupos de causas de muerte y si dichos patrones pueden interpretarse de manera estructurada dentro del modelo aplicado. De este modo, se contribuye a fortalecer el vínculo entre la estadística pública, el análisis territorial y el uso metodológicamente riguroso de técnicas contemporáneas de aprendizaje automático.

Descripción del problema

Las estadísticas vitales son la principal fuente para hacer monitoreo de nacimientos y defunciones. No obstante, es necesario que sean de calidad para ser útiles, pues se pueden usar de diversas formas para la orientación de políticas y prácticas de salud pública (Mahapatra et al., 2007).

En Colombia, el Departamento Administrativo Nacional de Estadística (DANE) es la entidad encargada de consolidar, validar, procesar y difundir la información de estadísticas vitales. Actualmente, el Sistema de Estadísticas Vitales en Colombia sólo incluye registros de nacimientos y defunciones fetales y no fetales, los cuales se encuentran disponibles tanto en cuadros consolidados desagregados por sexo, edad, departamento entre otros, como en microdatos anonimizados (DANE, 2019).

Respecto a los registros de defunciones no fetales, los investigadores suelen usar los cuadros consolidados o información agregada. Aun cuando usan microdatos, estos son utilizados frecuentemente para la construcción de indicadores, dejando de lado su uso para el análisis a nivel individual. Por tal motivo, nace la necesidad de explorar si la información que contienen los microdatos, complementada con variables contextuales a nivel municipal permite diferenciar patrones sociodemográficos relacionados con diversas causas de defunción no fetal para el periodo 2015-2019 en el departamento de Santander, así como analizar el papel de las variables sociodemográficas y contextuales en dicha diferenciación.

Justificación

La presente investigación se enfocará en profundizar el análisis de las causas de defunciones no fetales a partir de la información de los microdatos de defunciones no fetales del DANE, complementados con variables contextuales a nivel municipal. Pese a que estos registros han sido utilizados para la generación de indicadores, su potencial a nivel individual no ha sido tan explorado, limitando la identificación de arquetipos asociados a las distintas causas de defunción no fetal.

Para este propósito, se utilizan modelos de aprendizaje automático, permitiendo capturar relaciones entre variables, junto con herramientas de interpretabilidad como SHAP, que facilitan el análisis de la contribución de las variables en las clasificaciones efectuadas por el modelo aplicado.

En consecuencia, el proyecto no solo busca modelar la información contenida, sino también interpretarla, aportando evidencia sobre el alcance y las limitaciones de estos registros para su uso en análisis más detallados. Por ende, se propone ampliar el uso tradicional de los microdatos más allá de enfoques descriptivos, evidenciando su potencial como fuente de información para el análisis de fenómenos poblacionales desde una perspectiva individual.

Objetivos

Objetivo General

Analizar el potencial de los microdatos de defunciones no fetales, complementados con información contextual a nivel municipal, para diferenciar las categorías de causas de defunción no fetal (enfermedades del sistema circulatorio, neoplasias y causas externas) mediante modelos de aprendizaje automático en el departamento de Santander durante el periodo 2015–2019.

Objetivos Específicos

Caracterizar las defunciones no fetales en Santander (2015–2019) según variables sociodemográficas.

Construir modelos de aprendizaje supervisado para la clasificación de las categorías de causas de defunción no fetal.

Evaluar la capacidad de los modelos para diferenciar las categorías de causas de defunción no fetal.

Interpretar los resultados del modelo mediante SHAP para analizar la contribución de las variables en la diferenciación de las categorías de causas de defunción no fetal.

Marco de Referencia

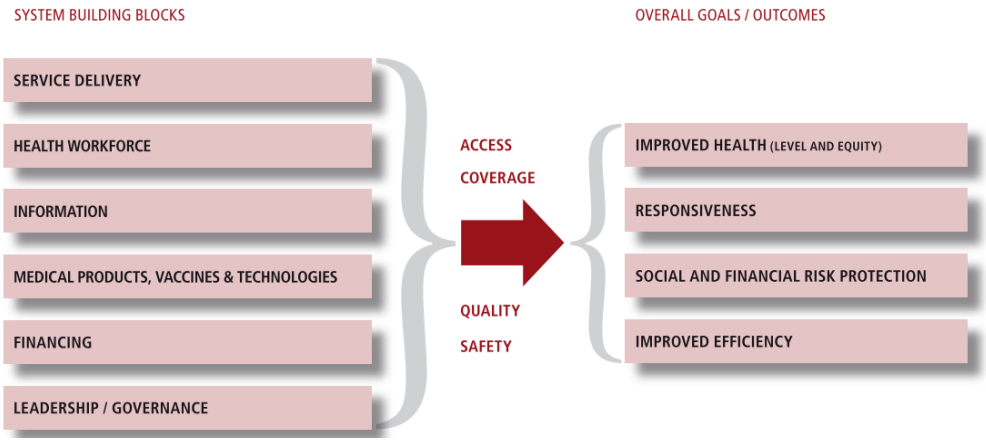
Sistemas de Información en Salud

La Organización Mundial de la Salud adopta un marco conceptual de seis bloques para entender cómo funciona un sistema de salud y cómo puede mejorarse. Esta entidad señala que un sistema de salud tiene múltiples objetivos, entre ellos lograr un mayor acceso y cobertura de intervenciones de salud efectivas, así como mejorar la equidad en salud de manera responsiva y financieramente justa (World Health Organization [WHO], 2007).

Entre estos seis bloques se destaca el de la información (Figura 1). La OMS (2007) define que un buen sistema de información en salud es aquel que garantiza la producción, el análisis, la difusión y el uso de información confiable y oportuna sobre los determinantes y el estado de la salud. Para ello, el sistema debe generar datos a nivel poblacional y de establecimientos, provenientes de censos, registros civiles y encuestas de hogares, así como contar con la capacidad de sintetizar la información y promover su disponibilidad y uso.

Figura 1

Sistema de Bloques



Nota. Tomado de Everybody's business : strengthening health systems to improve health outcomes : WHO's frmaework for action, por World Health Organization (2007).

Esta última la OMS (2007) la cataloga como prioritaria, pues enfatiza en fortalecer sistemas que analicen y utilicen información confiable proveniente de múltiples fuentes, en colaboración con organismos internacionales. Sin embargo, AbouZahr et al., 2015 han sido críticos con la OMS, dado que esta como responsable de la Clasificación Estadística Internacional de Enfermedades y Problemas de Relacionados con la Salud como estándar internacional no ha sido efectiva en movilizar los recursos organizacionales mínimos para dar apoyo operativo a los países en su implementación.

Estadísticas Vitales y Microdatos en Colombia

En el contexto colombiano, bajo el decreto 266 de 1953 (Presidencia de la Republica) y complementado con el decreto 3167 de 1968, se establece al Departamento Administrativo Nacional de Estadística- DANE como el único ente autónomo y responsable de la difusión y producción de las estadísticas oficiales. Actualmente la difusión de registros oficiales de estadísticas vitales abarca nacimientos, defunciones fetales y defunciones no fetales. Estos registros son confiables dado que están cobijados bajo el decreto 1171 de 1997 donde se describe el personal de salud autorizado para hacerlos en el país (Ministerio de Salud y Protección Social, 2024).

De igual manera, a través del Archivo Nacional de Datos (ANDA) dispone las bases de datos anonimizadas el cual contienen las variables de los certificados de nacido vivo y defunciones fetales y no fetales. Estos son conocidas como microdatos. Para este estudio se explorará específicamente los microdatos de defunciones no fetales.

Calidad y Uso de los Microdatos de Defunciones no Fetales

El Indicador global de consistencia, coherencia y completitud, es un indicador que evalúa el grado en que las variables diligenciadas en el formulario de nacimiento y defunción cumplen con los principios ya mencionados (consistencia, coherencia y completitud). Para hacer seguimiento, el DANE emplea una herramienta informática que valida de manera detallada estos principios, generando un reporte con los errores detectados (DANE, 2019).

Desde el 2008 este indicador ha venido mejorando, pasando de 3,8 % de errores a 0,3% en el 2019, es decir, una reducción del 91,6%. El informe destaca que las variables que presentan mayor error son las relacionadas con nivel educativo, localidad y seguridad social del fallecido (DANE, 2019).

Bajo este panorama, el DANE (2019) afirma que con esta información es posible conocer cómo se comporta la mortalidad en el país, además de identificar las causas que originan las muertes como las diferencias a nivel nacional, departamental y municipal. También la constituye como fuente vital para el cálculo de indicadores como las tasas de mortalidad infantil o materna entre otras.

Microdatos en la Literatura.

A lo largo del tiempo, la literatura ha usado estos datos para hacer análisis de tipo descriptivo, trabajos como los de Castañeda-Millán & Eslava-Schmalbach, 2024; Martínez Chacón, 2025; Otero, 2013 se han enfocado en analizar tasas o tendencias de causas de muerte, ya sea por cortes de edad, departamento o entornos específicos. Con ello se resalta que los aportes de los autores convergen en una misma línea de análisis ampliamente explorada y centrada en el manejo de los datos a nivel agregado.

Sin embargo, en los últimos años se han presentado investigaciones que usan estos microdatos de una manera distinta, especialmente aplicando algoritmos de machine learning. Trabajos como los de Araujo Zarate et al. (2024) y Ortiz (2020) se caracterizan por usar diversos modelos para clasificar problemas de carácter binario, como la clasificación de nacido vivo y defunción fetal. Los autores justifican el uso de estos algoritmos por su capacidad de pronosticar con precisión eventos complejos además de producir conocimiento

Concepto de Reutilización de Datos

La aplicación de estas técnicas se podría fundamentar desde el concepto de reutilización de datos. Según Van de Sandt et al. (2019) este concepto es complejo y varía según la disciplina que aplique o incluso entre individuos. Los autores realizaron una búsqueda sistemática de las definiciones establecidas y para este caso se tomó la definición de Sun & Khoo (2017). Los autores establecen que los datos que se utilizan para fines distintos a los que fueron recopilados originalmente, se consideran reutilizados. En otros terrenos se le considera “uso secundario de datos de investigación”. Esta última expresión Law (2006) la define como el uso de datos de investigación para analizar problemas distintos a aquellos para los cuales fueron originalmente recolectados. Estos pueden provenir de fuentes administrativas, del sector salud o educativo, de censos o de estudios previos.

Potencial Analítico de los Microdatos de Defunciones no Fetales.

En síntesis, se podría establecer que los microdatos de defunciones no fetales son aptos para ser usados en técnicas de aprendizaje automático especialmente el supervisado. No obstante, aún no se ha explorado el potencial que estos tienen para diferenciar categorías de causas de defunción no fetal. Por ende, este estudio pretende evaluar el potencial de los microdatos de

defunciones no fetales para identificar e interpretar patrones asociados a distintas categorías de causas de defunción no fetal, integrando variables sociodemográficas y contextuales.

Síntesis de la Fundamentación Teórica y Técnica

La tabla 1 presenta un resumen de los temas y los autores que brindan el respaldo científico y técnico a este marco de referencia.

Tabla 1

Síntesis de conceptos centrales del marco de referencia

Tema	Idea central	Autores
Sistemas de información en salud.	Los datos son fundamentales para analizar el estado de salud y apoyar decisiones.	WHO (2007); AbouZahr et al. (2015
Microdatos y estadísticas vitales.	Los registros del DANE son fuentes confiables y disponibles a nivel individual.	DANE (2019); Ministerio de Salud (2024)
Calidad de los datos	La mejora en calidad respalda su uso en análisis	DANE (2019
Uso tradicional	Se han utilizado principalmente en análisis descriptivos (tasas y tendencias).	Castañeda-Millán & Eslava-Schmalbach (2024); Otero (2013); Martínez Chacón (2025)
Uso reciente (ML)	Se han aplicado modelos para predecir eventos binarios	Ortiz (2020); Araujo Zarate et al. (2024)
Reutilización de datos	Uso de datos para analizar problemas distintos a aquellos para los que fueron originalmente recolectados	van de Sandt et al. (2019); Sun & Khoo (2017); Law (2006)

Metodología

Enfoque Metodológico

El presente estudio desarrolla un enfoque predictivo supervisado con fines analíticos e interpretativos. No obstante, no se orienta al desarrollo de un modelo para despliegue en el campo de la salud. Para cumplir con el objetivo general, se evaluarán los microdatos de defunciones no fetales comprendidos entre los años 2015 y 2019 en el departamento de Santander, agregando variables propias del municipio de residencia del fallecido como la estructura de actividad económica y la distribución de la población.

En primera instancia se realiza un análisis descriptivo de las variables sociodemográficas con el objetivo de caracterizar las defunciones no fetales en el periodo de estudio. Posteriormente se pretende identificar patrones y asociaciones multivariables relacionadas con las causas de defunción analizadas sin establecer relaciones causales. En esta etapa se emplean modelos supervisados lineales y no lineales.

Respecto al modelo lineal, se utiliza un modelo de regresión logística multinomial, con el fin de comparar su desempeño frente a modelos basados en árboles de decisión (no lineal), particularmente XGBoost y CatBoost.

Este ejercicio se realiza con el propósito de identificar qué enfoque captura mejor las relaciones de los microdatos tratados. Respecto al desempeño general de los modelos, se evaluará mediante métricas de clasificación (accuracy, precisión, recall y F1-score), incorporando validación cruzada estratificada para analizar la estabilidad de los resultados y reducir la dependencia de una única partición entrenamiento-prueba.

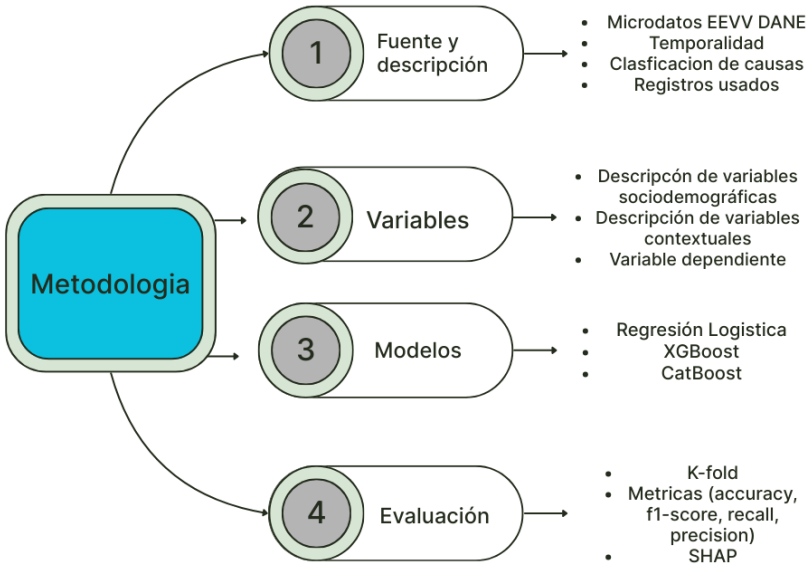
Por último, se aplica SHAP (SHapley Additive Explanations) al modelo de mejor desempeño, con el propósito de interpretar la contribución de cada variable en la clasificación de

las causas de defunción establecidas. Este tratamiento permite avanzar más allá de la predicción permitiendo identificar perfiles asociados a cada causa.

A continuación, se presenta un flujograma (ver Figura 2) el cual recopila las etapas ejecutadas, desde la preparación y recolección de los datos, hasta la interpretación de los modelos supervisados aplicados.

Figura 2

Flujograma general de la metodología del estudio



Fuente: Elaboración propia

Fuente y Descripción de los Datos

Fuente de los Datos

La base de datos utilizada en este estudio se conforma a partir de tres fuentes oficiales provenientes del Departamento Administrativo Nacional de Estadística (DANE). En primer lugar, se disponen los microdatos de defunciones no fetales para el periodo de tiempo 2015-

2019. En segundo lugar, las proyecciones de población, elaboradas por el DANE con base en los resultados del Censo Nacional Población y Vivienda- CNPV-2018. Por último, variables derivadas de las cuentas nacionales. Los datos de proyección poblacional y estructura económica de los municipios fueron unificados con los microdatos de defunciones no fetales mediante el código DIVIPOLA[1] del municipio de residencia.

Descripción del Conjunto de Datos

Unificado el dataset, se obtuvieron 54.265 registros de personas fallecidas en el intervalo de tiempo establecido, de los cuales 22 registros presentaron valores nulos. Estos valores nulos se dieron a causa de que las personas fallecidas no tenían código divipola registrado, debido a su condición de residencia registrada en el extranjero, por lo tanto, fueron excluidas del estudio.

Con lo anterior, se decidió hacer un análisis descriptivo tomando variables como grupos etarios, sexo y causas de defunción no fetal, con el propósito de entender la dinámica de defunciones no fetales en Santander.

Creación de Variables

Con el propósito de estructurar la variable de análisis, se definieron tres categorías conceptuales de causas de muerte, fundamentadas en la estructura de la Lista 6/67 para la Tabulación de la Mortalidad de la OPS/OMS:

Enfermedades del sistema circulatorio (código 0): Abarca causas crónicas no transmisibles, principalmente enfermedades cardiovasculares y cerebrovasculares.

Neoplasias (código 1): Incluye los diversos tipos de tumores malignos con alta incidencia en el departamento.

[1] El DANE define DIVIPOLA como un estándar nacional que codifica y lista las entidades territoriales a saber: departamentos, municipios, corregimientos departamentales, así como los centros poblados, tanto inspecciones de policía, como caseríos y corregimientos municipales en el área rural.

Causas externas (código 2): muertes por accidentes y eventos violentos de origen no médico.

Una vez establecidas las tres categorías principales, se identificaron las causas específicas más representativas dentro de cada grupo establecido, de acuerdo con su frecuencia acumulada en el periodo estudiado. Las causas incluidas fueron:

Sistema circulatorio: enfermedades isquémicas del corazón, enfermedades cerebrovasculares, enfermedades hipertensivas

Neoplasias: Tumor maligno del estómago, Tumor maligno del colon, de la unión rectosigmoidea, recto y ano, y Tumor maligno de la tráquea, los bronquios y el pulmón.

Externas: accidentes de transporte terrestre, lesiones autoinfligidas intencionalmente y agresiones u homicidios.

Adicionalmente se crearon cuatro variables de tipo contextual y socioeconómico. La variable de porcentaje de población rural del municipio de residencia establecida como la razón de personas en centros poblados y rural disperso respecto a la población total del municipio, el porcentaje de participación en actividades primarias y el porcentaje de participación en actividades secundarias. Por último, una variable dicotómica de desplazamiento la cual toma el valor de 1 si el individuo falleció en un municipio diferente al de su residencia, y 0 en caso contrario.

La recodificación de variables categóricas no fue necesaria, puesto que los microdatos de defunciones no fetales del DANE presentan una codificación numérica estandarizada, respaldada por un diccionario técnico oficial.

En este punto, el dataset se estableció en 22376 registros de defunciones no fetales, correspondiente a personas fallecidas con residencia en el departamento de Santander

equivalentes al 41,2% del total de defunciones no fetales registradas en el periodo de estudio. Las enfermedades del sistema circulatorio representan 15352 casos (68,61%), neoplasias 3408 casos (15,23%), y las causas externas 3616 casos (16,16%).

La selección de estas categorías respondió a una delimitación metodológica orientada a trabajar con grupos epidemiológicamente relevantes y con suficiente representatividad estadística dentro del periodo analizado. En consecuencia, el estudio no busca modelar la totalidad de causas de mortalidad registradas en Santander, sino analizar el comportamiento de modelos supervisados sobre categorías previamente definidas.

Depuración del dataset

Para el desarrollo del modelo, fue necesario hacer un proceso de depuración, en el cual se identificó que ciertas variables influyentes presentaban registros clasificados como “sin información”. En el proceso se encontraron 7.718 ocurrencias de la categoría “sin información” distribuidas entre cinco variables (Tabla 2). Estas ocurrencias corresponden a 5.398 registros únicos afectados.

Tabla 2

Registros "sin información" por variable

Código[2]	Descripción	Ocurrencias
est_civil	Estado civil del fallecido	2.954
gru_ed1	Grupo Etario	3
nivel_edu	Nivel educativo alcanzado.	4.284
seg_social	Régimen de afiliación	422
area_res	Área de residencia del fallecido	55

[2] Codificación registrada en los microdatos de defunciones no fetales del DANE

Código[2]	Descripción	Ocurrencias
Total Ocurrencias		7.718
Registros únicos afectados		5.398

Nota. Un registro puede presentar la categoría "sin información" en más de una variable. Por tal motivo, el número de ocurrencias es superior al número de registros únicos afectados.

No obstante, estos registros no corresponden estrictamente a valores faltantes generados durante el procesamiento, sino también a una opción contemplada en ciertas secciones del certificado de defunción procesado por el DANE, como el nivel educativo y el estado civil. Según el manual de crítica y codificación de certificados de nacido vivo y de defunción – EEVV esta opción puede producirse en distintos escenarios, por ejemplo, cuando la persona autorizada de diligenciar el formulario no le es posible obtener el dato o en situaciones asociadas a inconsistencias en el diligenciamiento, tales como casillas en blanco o múltiples marcaciones dentro de una misma variable. En estos casos, el procedimiento de validación establece el uso de la categoría "sin información" como mecanismo de estandarización administrativa (DANE, 2004). Esta diversidad de causas impide asumir que todos los registros clasificados bajo la categoría "sin información" respondan a un mismo mecanismo de ausencia de información.

Como se puede apreciar en la Tabla 2, el nivel educativo fue la variable de mayor ocurrencia, seguido del estado civil. En particular, el comportamiento observado para el nivel educativo coincide con lo señalado en el marco de referencia sobre el indicador global de consistencia, coherencia y completitud reportado por el DANE, donde se destaca como una de las variables de mayor error.

Ante este panorama, se planteó el objetivo de identificar si los registros clasificados como "sin información" se encontraban asociados a determinadas características observadas de los

individuos. Para ello, se realizó una comparación entre los registros afectados y no afectados mediante una prueba de independencia Chi-cuadrado. Como se puede apreciar en la Tabla 3, no se observaron diferencias estadísticamente significativas para variables como sexo, área de residencia y la causa de muerte tratada. En cambio, variables como grupo etario o desplazamiento, entendida como la situación de fallecimiento en un municipio distinto al de residencia, presentaron diferencias estadísticamente significativas entre los registros afectados y no afectados.

No obstante, en el caso del grupo etario, las diferencias porcentuales observadas entre ambos grupos fueron reducidas, no superando los dos puntos porcentuales, mientras que la mayor diferencia se presentó en la condición de desplazamiento. En general, estos resultados sugieren que los registros afectados y no afectados presentan una distribución similar en la mayoría de las variables analizadas.

Tabla 3

Prueba Chi-Cuadrado

Variable	χ^2	gl	p-valor
Sexo	0,56	1	0,456
Grupo Etario	112,03	24	< 0,001
Area Residencia	5,89	2	0,053
Se desplazó	176,76	1	< 0,001
Causa de muerte (target)	3,96	2	0,138

Ante este panorama, se buscaron alternativas para el tratamiento de estos registros, como la inclusión de la categoría “sin información” dentro del modelo, sin embargo, al ser una

categoría que no representa una característica interpretable, su inclusión podría ocasionar que el modelo diferencie los registros a partir de la ausencia administrativa de información y no con base en los patrones asociados a las causas de mortalidad analizadas.

Por otro lado, también se analizó la aplicación de métodos de imputación. No obstante, Galván & Medina (2007) reconocen que el uso de estos se debe tomar con cautela, priorizando el contexto que se está manejando. En este sentido, realizar un análisis de datos sin la elección a priori de un método de imputación resalta que el análisis exploratorio y la consistencia de los datos son los que marcan la pauta para tomar la mejor decisión. En este caso, la categoría "sin información" podía originarse tanto por ausencia del dato como por inconsistencias en el diligenciamiento del certificado, por lo que la imputación de variables categóricas como nivel educativo, estado civil y régimen de afiliación podía introducir información artificial.

Bajo este contexto, se decidió eliminar los registros con estas características. Como se observa en la Tabla 4, a pesar de la eliminación de estos registros, la estructura porcentual se mantuvo relativamente estable en variables como sexo, grupos etarios y las categorías de causas de muerte analizadas, ratificando que la depuración empleada no representó alteraciones sustanciales en la composición general del conjunto de datos.

Tabla 4

Comparación estructural del conjunto original y el conjunto depurado

Variable	Sin depuración (%)	Con depuración (%)
Causa de muerte		
Sistema Circulatorio	68,61	68,32
Neoplasias	15,23	15,25
Externas	16,16	16,42

Variable	Sin depuración (%)	Con depuración (%)
Sexo		
Hombre	57,93	57,78
Mujer	42.07	42,21
Grupo Etario		
0-14 años	0,49	0,34
15-29 años	9,10	9,29
30-44 años	6,74	6,83
45-59 años	8,89	8,82
60-74 años	24,11	23,70
75 años y más	50,66	51,02

Estos resultados permitieron establecer un dataset consistente para el entrenamiento de los modelos supervisados aplicados. Finalmente, el dataset quedó conformado por 16.798 registros.

A pesar de ello, es pertinente reconocer que cualquier método de tratamiento de estos registros implica limitaciones metodológicas, Mientras que conservar la categoría “sin información” podía generar un aprendizaje de patrones no interpretables dentro del modelo, también aplicar métodos de imputación podía introducir información artificial o sesgos que se derivan del procedimiento utilizado. En este sentido, se consideró que la eliminación de dichos registros representaba la opción metodológica más adecuada en cuanto a los objetivos planteados y las características del dataset analizado. En consecuencia, los resultados del estudio deben interpretarse considerando esta limitación metodológica.

Descripción de las Variables

Para efectos de mayor claridad al lector, las variables utilizadas en el estudio se estructuraron en tres grupos: sociodemográficas, contextuales y dependiente. Las sociodemográficas como su nombre lo indica describen las características individuales del decesado; las contextuales, son las que caracterizan el municipio de residencia, por último, la dependiente corresponde a la causa de muerte definida previamente como variable objetivo en los modelos de aprendizaje supervisado.

A continuación, se presenta la descripción de las variables incluidas en el análisis. La Tabla 5 resume las variables sociodemográficas, mientras que la Tabla 6 muestra las variables contextuales y la variable dependiente.

Tabla 5

Variables sociodemográficas utilizadas en el estudio

Variable	Descripción	Tipo
sexo	Sexo del fallecido (masculino=1 / femenino=0)	Dicotómica
est_civil	Estado civil del fallecido	Categórica Nominal
gru_ed1	Grupo Etario	Categórica Ordinal
nivel_edu	Nivel educativo alcanzado.	Categórica Ordinal
seg_social	Régimen de afiliación	Categórica Nominal
area_res	Área de residencia del fallecido	Categórica Nominal
Se_desplazó	Indica si el fallecimiento ocurrió en un lugar distinto al de residencia habitual del individuo	Dicotómica

.

Tabla 6

Variables contextuales y variable dependiente

Variable	Descripción	Tipo
pobr_res_percent	Población en centros poblados y rural disperso (%) del municipio de residencia.	Núm. (%)
var_prim_percent	Valor agregado bruto del sector primario (%) en el municipio de residencia.	Núm. (%)
var_terc_percent	Valor agregado bruto del sector terciario (%) en el municipio de residencia	Núm. (%)
target	Categoría de causa de muerte: 0 = sistema circulatorio, 1 = neoplasias, 2 = externas.	Categórica

Nota. Núm. (%) = variable numérica expresada en porcentaje.

Como se puede apreciar, estas son las variables definidas para el análisis mediante modelos de aprendizaje supervisado, orientado a identificar las variables que influyen en la clasificación de las causas establecidas.

Modelos y Procedimientos Analíticos

Preparación de los Datos

El dataset final se dividió en un 70 % para entrenamiento y 30 % para prueba, utilizando partición estratificada con el propósito de mantener la proporción original entre las tres categorías de causa de muerte (sistema circulatorio, neoplasias y externas).

Dependiendo del modelo aplicado, las variables categóricas se trataron acorde a los requerimientos de cada algoritmo. Para la regresión logística multinomial, las variables categóricas nominales se transformaron en dicotómica excluyendo una categoría por variable para evitar problemas de colinealidad, permitiendo una correcta estimación de los efectos. Por otro lado, las variables categóricas ordinales se procesaron como numéricas aprovechando la codificación original del DANE y el orden propio de sus categorías.

En la Tabla 7 se expone un ejemplo de cómo queda transformada una variable categórica nominal, en este caso, el estado civil.

Tabla 7

Ejemplo categorización de variables

id	Estaba separado(a), divorciado(a)	Estaba viudo(a)	Estaba soltero(a)	Estaba casado(a)	llevaba menos de dos años viviendo con su pareja
1	1	0	0	0	0
2	0	1	0	0	0
3	0	0	0	0	1
4	0	0	1	0	0

La caracterización completa de las variables categóricas nominales usadas en el estudio se encuentra en el Anexo A. De igual manera, los Anexos B y C detallan las variables categóricas ordinales. Estos anexos sirven como referencia para la interpretación de los resultados en los gráficos tipo beeswarm y waterfall plot, donde se analizan los efectos de dichas variables.

Para los modelos basados en árboles (XGBoost), la presencia de múltiples variables dicotómica no representa un problema estructural, ya que estos algoritmos no dependen de supuestos de independencia lineal entre predictores. En el caso de CatBoost, no fue necesaria la transformación a variables dicotómicas, ya que este algoritmo incorpora un mecanismo interno para el tratamiento estadístico de variables categóricas, permitiendo su utilización directa en el proceso de entrenamiento.

La ejecución y el análisis se realizó mediante lenguaje Python en entorno vs code. Allí se emplearon librerías especiales como pandas, sckit-learn y matplotlib entre otras.

Modelos Supervisados Aplicados

Como se puede observar en la Tabla 4, la variable objetivo (target) presentó un desbalance moderado, donde las causas relacionadas con el sistema circulatorio fueron las de

mayor frecuencia dentro del conjunto analizado en Santander para el periodo 2015-2019. Esta característica podría afectar el aprendizaje del modelo, ocasionando el favorecimiento de la clase mayoritaria respecto a las minoritarias.

En la literatura existen diversos enfoques para tratar este desequilibrio como lo es el sobremuestreo (oversampling) de la clase minoritaria, el submuestreo de la clase mayoritaria o la modificación de la función de perdida para dar a la clase minoritaria fallida un costo más alto. El sobremuestreo, especialmente la técnica de sobremuestreo sintético de la clase minoritaria (SMOTE) se ha destacado como uno de los métodos más implementados en las últimas dos décadas, sin embargo, no implica que sea necesariamente el más beneficioso, para todos los contextos.

Autores como Tarawneh et al. (2022) evaluaron más de 70 métodos de sobremuestreo incluyendo SMOTE, utilizando nueve conjuntos de datos reales desbalanceados. Los autores argumentaron que gran parte de la literatura valida estos métodos a partir del rendimiento del clasificador empleado para clasificar los conjuntos de datos sobremuestrados mediante medidas como Precisión, Exactitud, Exhaustividad, Medida F, Media geométrica, Especificidad, entre otras. Sin embargo, señalaron que estas evaluaciones asumen que las observaciones sintéticas generadas pertenecen efectivamente a la clase minoritaria, sin verificar explícitamente la validez de dicha suposición. Por ende, los autores aplicaron un método de validación específico, dando como resultado que todos los métodos de sobremuestreo estudiados generaron muestras sintéticas con alta probabilidad de pertenecer a la clase mayoritaria, cuestionando la fiabilidad de estas técnicas para el aprendizaje a partir de datos con clases desbalanceadas.

No obstante, la literatura no presenta consenso respecto a una estrategia universal para el tratamiento del desbalance de clases. Weiss et al. (2007). encontraron que el desempeño relativo

de técnicas como el sobremuestreo, el submuestreo y el aprendizaje sensible al costo depende en gran medida de las características del conjunto de datos analizado. Por ejemplo, los autores observaron que en conjuntos de datos con más de 10.000 observaciones los enfoques sensibles al costo tendieron a superar a los métodos de muestreo, mientras que en conjuntos de menor tamaño el sobremuestreo mostró resultados competitivos.

De igual manera Bakirarap & Elhan (2023) distinguen el sobremuestreo y submuestreo como las estrategias más utilizadas para afrontar el desbalance de clases, sin embargo, también señalan que pueden introducir limitaciones. Por ejemplo, el uso de sobremuestreo podría generar observaciones repetidas favoreciendo el sobreajuste del modelo, Mientras que el submuestreo implica la eliminación de información, lo que podría ocasionar una pérdida de información relevante limitando el aprendizaje del modelo.

En consecuencia, los autores examinaron distintos esquemas de ponderación de clases aplicados a modelos Random Forest y SVM, donde se observaron mejoras en las categorías minoritarias y en métricas de evaluación.

Bajo este contexto, se decidió emplear una estrategia de aprendizaje sensible al costo (Cost-Sensitive Learning), ya que permite abordar el desbalance de clases mediante la asignación de pesos diferenciados durante el entrenamiento, sin modificar la distribución original de los datos. Con esto se buscó mantener la estructura de los registros analizados, concediendo mayor importancia a las categorías minoritarias y reduciendo el sesgo de los modelos hacia la clase mayoritaria, sin necesidad de generar observaciones sintéticas ni eliminar registros del conjunto de datos**.**

A continuación, se describen los modelos supervisados aplicados y las estrategias empleadas para el tratamiento del desbalance de clases.

Regresión Logística Multinomial**.** Es un modelo supervisado lineal que estima la probabilidad de pertenencia a cada clase mediante la aplicación de la función logística sobre una combinación lineal de las variables explicativas, permitiendo modelar directamente la relación entre características predictoras y una variable dependiente categórica (Beysolow, 2017).

Para tratar el tema de desbalance, se aplicó el algoritmo de class_weight=" balanced". Según la documentación de scikit-learn, este parámetro asigna automáticamente pesos inversamente proporcionales a la frecuencia relativa de cada clase dentro del conjunto de entrenamiento, otorgando mayor importancia a las categorías minoritarias durante el proceso de ajuste del modelo.

En este estudio se incluyó como modelo de referencia con la intención de comparar el desempeño de un enfoque lineal frente a modelos no lineales basados en árboles de decisión.

XGBoost es un algoritmo de ensamble basado en gradient boosting que construye árboles secuenciales con el objetivo de minimizar el error residual y optimizar una función de pérdida diferenciable (Chen & Guestrin, 2016). Debido al desbalance en la variable dependiente, se aplicó una ponderación diferencial mediante sample_weight en la función de pérdida utilizando pesos balanceados inversamente proporcionales a la frecuencia relativa de cada clase dentro del conjunto de entrenamiento. Con esta estrategia, el algoritmo penaliza con mayor dureza los errores asociados a las clases minoritarias.

CatBoost es un algoritmo de gradient boosting que incorpora ordered boosting y un mecanismo estadístico específico para el tratamiento de variables categóricas Prokhorenkova et al. (2019).Para tratar el desbalance de clases, se aplicó una ponderación manual de pesos en la función de pérdida para compensar el desbalance de la variable dependiente.

Evaluación e Interpretabilidad del Modelo

Métricas Utilizadas

Para evaluar el desempeño de los modelos supervisados implementados en este estudio, se utilizaron métricas de clasificación el cual permiten identificar que tan bien está diferenciando las causas de muerte tratada. Según Hat et al. (2012), la evaluación de un clasificador debe realizarse sobre observaciones no utilizadas durante el entrenamiento, con el pretexto de evitar estimaciones optimistas del rendimiento y aproximar su capacidad de generalización.

Las métricas empleadas incluyen la exactitud (*accuracy*), la precisión (*precision*), la sensibilidad o *recall* y el índice F_1. Estas medidas cuantifican la proporción de aciertos y el equilibrio entre la detección de verdaderos positivos y la reducción de errores de clasificación, ofreciendo una acercamiento del rendimiento del modelo.

La evaluación de este estudio se realizó en dos niveles. En primera nivel, en la fase de entrenamiento, se implementó validación cruzada estratificada con el objetivo de estimar el desempeño promedio de cada modelo y su estabilidad a través de múltiples particiones del conjunto de datos.

A continuación, la Tabla 8 presenta un resumen de los parámetros bases aplicados para la validación cruzada y las particularidades de cada modelo.

Tabla 8

Resumen de modelos supervisados implementados

Modelo	Principio teórico	Parámetros	Ventajas
Logístico Multinomial	Modelo supervisado lineal para clasificación multiclase basado en la función logística aplicada a una combinación lineal de variables	solver='lbfgs', max_iter=500, random_state=42, class_weight= 'Balanced'	Modelo de referencia lineal, interpretable, eficiente computacionalmente

Modelo	Principio teórico	Parámetros	Ventajas
XGBoost	Boosting secuencial basado en gradiente; combina árboles débiles para minimizar error residual(Chen & Guestrin, 2016)	objective='multi:softpro', num_class=3, n_estimators=500, max_depth=5, learning_rate=0.05	Alta precisión, regularización L1/L2, eficiente en conjuntos desbalanceados.
CatBoost	Boosting ordenado que corrige *target leakage* y mejora el tratamiento de variables categóricas(Prokhorenkova et al., 2019)	iterations=500, depth=6, learning_rate=0.1, loss_function='MultiClass'	Reduce sesgo de predicción, estable con datos heterogéneos, robusto a sobreajuste.

Una vez seleccionado el modelo de mejor desempeño en validación cruzada, el segundo nivel consistió, primero en ajustar los hiperparámetros mediante técnicas como Randomized SearchCV o GridSearchCV y segundo, evaluar sobre el conjunto de prueba (30 % de los datos). En esta fase se analizaron tanto las métricas globales como las métricas por causa, además se empleó la matriz de confusión, con el propósito de examinar el comportamiento del modelo en cada causa de defunción.

En la Tabla 9 se presentan las expresiones matemáticas y la interpretación de las métricas utilizadas en el análisis.

Tabla 9

Definición y descripción de las métricas de evaluación

Métrica	Formula	Descripción
Precision	$\frac{TP}{TP + FP}$	Proporción de predicciones positivas que fueron correctas
Recall	$\frac{TP}{TP + FN}$	Capacidad del modelo para identificar correctamente los casos positivos.

Métrica	Formula	Descripción
F_1-score	$2 \times \frac{\text{Precision x Recall}}{\text{Precision+Recall}}$	Media armónica entre precisión y recall. Evalúa el equilibrio entre ambas métricas.
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$	Proporción total de observaciones correctamente clasificadas.
Macro avg	Promedio aritmético de las métricas por clase.	Evalúa el desempeño medio del modelo sin ponderar por el tamaño de las clases.
Weighted avg	Promedio ponderado por el número de instancias en cada clase.	Refleja mejor el rendimiento global en presencia de clases desbalanceadas.

Nota. TP = verdaderos positivos; TN = verdaderos negativos; FP = falsos positivos; FN = falsos negativos. Las métricas "macro avg" y "weighted avg" se obtienen promediando los resultados por clase, la primera sin ponderación y la segunda ponderando por el tamaño de cada grupo.

Fuente*:* Adaptado de de Han et al. (2012).

Interpretabilidad del Modelo

Como se definió previamente, el método SHAP (SHapley Additive Explanations), propuesto por Lundberg & Lee (2017), constituye un marco para explicar las decisiones de modelos supervisados, asignando a cada variable un valor que cuantifica su contribución relativa en la decisión que toma los modelos para clasificar las causas.

Los autores señalan que, aunque los modelos modernos pueden capturar relaciones no lineales e interacciones complejas, su funcionamiento interno suele ser menos transparente que el de los modelos lineales tradicionales. En el presente estudio, Los valores SHAP se aplicaron al modelo que mostró el mejor desempeño y estabilidad en validación cruzada y cuya capacidad de generalización fue confirmada en el conjunto de prueba independiente. Este análisis no persigue explicar relaciones causales, sino examinar la contribución relativa de las variables sociodemográficas y contextuales en la diferenciación de las categorías consideradas.

Resultados

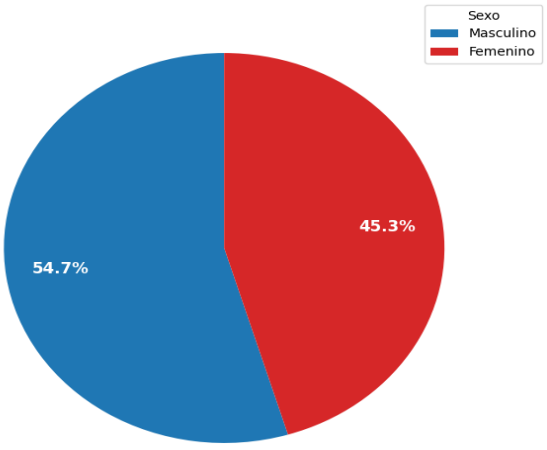
Análisis Descriptivo

Las defunciones no fetales en Santander presentaron una amplia diferencia entre el género masculino y femenino (Figura 3). En promedio, el 54.7% de las muertes correspondieron al género masculino, mientras que el 45,3% a mujeres, es decir se presentó una diferencia de 9.4 puntos porcentuales. Cabe destacar que un pequeño porcentaje (0,01%) se registró como indeterminado, sin embargo, este último fue excluido en el análisis.

Esta dinámica en el departamento sigue la tendencia global. Estudios como el de Wu et al. (2021) , demostraron que los hombres tienen un 60% más de riesgo de mortalidad que las mujeres, incluso si se ajusta por edad. Además, recalca que factores como el tabaquismo y las enfermedades cardiovasculares impactan en esta brecha, pese a que no la explican del todo, lo que sugiere que también intervienen variables contextuales y socioculturales.

Figura 3

Distribución promedio de defunciones no fetales por sexo en Santander (2015-2019)



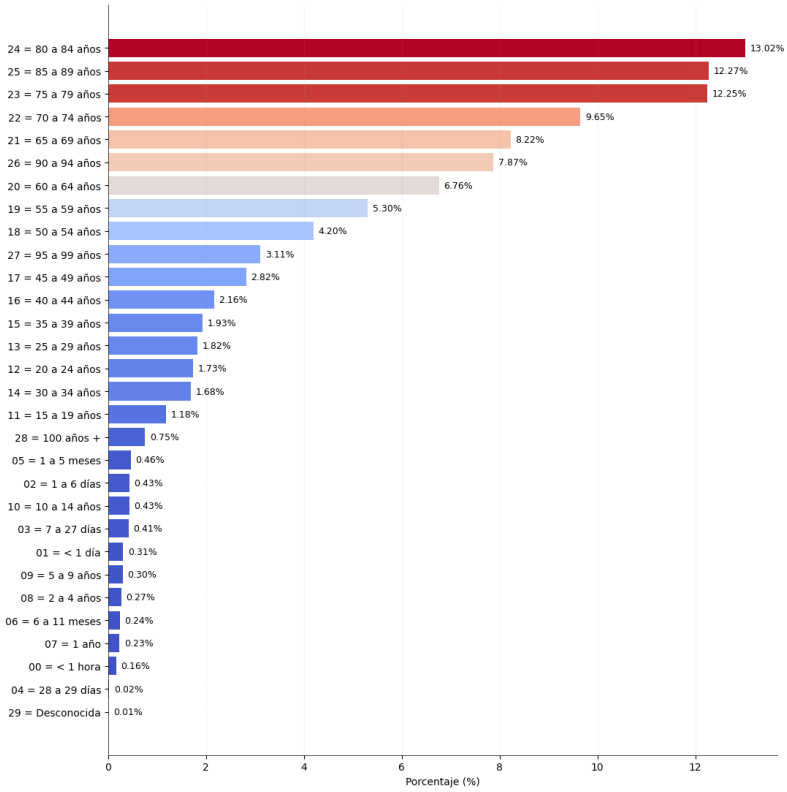
Fuente: Elaboración propia

Distribución por Grupos Etarios

Los resultados por grupos etarios (Figura 4), muestran que la mortalidad se concentra en adultos mayores. Especialmente el grupo de 80 a 84 años. Este grupo etario representa el 13,02 % del total de decesos, seguido por el de 85 a 89 años con 12,27 %, y el de 75 a 79 años con 12,25 %. Estos grupos representan más de un tercio de todos los decesos registrados.

Figura 4

Distribución por grupos etarios (2015-2019)



Fuente: Elaboración propia

Las edades tempranas representan una menor proporción, por ejemplo, los menores de un año tan solo abarcan el 2 % de las defunciones, mientras que los adolescentes y adultos jóvenes de 15 a 29 años suman alrededor del 4,7 %.

En síntesis, la mortalidad en Santander se concentra principalmente en edades avanzadas, especialmente a partir de los 65 años, lo que refleja el impacto de las enfermedades del sistema circulatorio como las principales causas de muerte en la región.

Evolución de las Principales Causas de Muerte en Santander

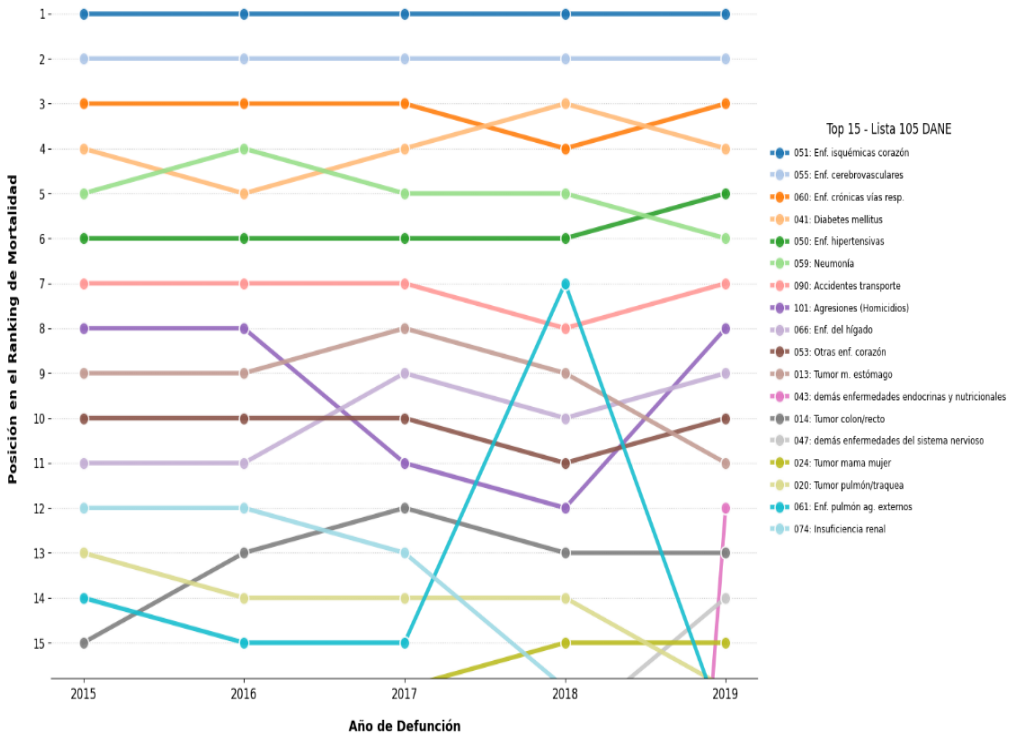
La evolución de la mortalidad en el departamento de Santander durante el periodo 2015-2019 (Figura 5) refleja una transición epidemiológica donde se hacen presente las enfermedades tradicionales, las causas externas con un impacto creciente y las enfermedades oncológicas. El análisis realizado permitió destacar que las enfermedades isquémicas del corazón se mantienen como la causa principal de defunción durante el periodo analizado. Este comportamiento, sumado a la persistencia de las enfermedades cerebrovasculares en la segunda posición establecen a las enfermedades del sistema cardiovascular como la principal causa de mortalidad en el departamento.

Respecto a las causas externas, los accidentes de transporte terrestre mantienen una presencia crítica y constante, oscilando entre los puestos siete y ocho del ranking general, lo que podría indicar una problemática de seguridad vial. Por otro lado, pero no menos importante, las agresiones y homicidios (101) presentan una volatilidad marcada, con fluctuaciones que los llevan a entrar y salir del grupo de las diez principales causas de muerte. Esto indica que el entorno y la violencia hacen que la esperanza de vida sea más variable que por causas médicas.

Desde lo social, es preocupante que las causas externas aún resalten y especialmente en Santander, pese a que en Colombia desde el año 2002, se ha disminuido estas causas, especialmente en homicidios (Dávila y Pardo, 2019).

Figura 5

Evolución de causas de muerte en Santander (2015-2019)



Fuente: Elaboración propia

Por último, el comportamiento de las causas neoplásicas en Santander se mantuvo estable, incluso disminuyendo su frecuencia en algunos años del periodo analizado. El tumor maligno de estómago y el tumor maligno de colon, recto y ano se mantuvieron dentro del top 15.

De igual manera, la presencia sostenida del tumor maligno de la tráquea, los bronquios y el pulmón en el registro histórico resalta la importancia de considerar factores de exposición ambiental y hábitos de vida.

Finalmente, se observa que mientras causas clínicas como la diabetes mellitus (041) ascienden puestos de manera progresiva, las enfermedades del sistema respiratorio muestran una mayor variabilidad anual. De esta forma, la caracterización permite discriminar con mayor

claridad los perfiles sociodemográficos que definen la exposición de la población santandereana a riesgos crónicos, eventos violentos o patologías oncológicas.

Resultados Modelo con Validación Cruzada

La validación cruzada estratificada permitió estimar el desempeño promedio y la estabilidad de los modelos supervisados evaluados sobre el conjunto de entrenamiento (70 % de los datos). En la Tabla 10 se presentan las métricas promedio y sus respectivas desviaciones estándar para cada uno de los modelos analizados.

Tabla 10

Desempeño promedio en validación cruzada (5-fold)

Modelo	Accuracy (±SD)	F1-macro(±SD)	Recall-macro (±SD)	Precision(±SD)
Regresión logística	0.6732 ± 0.0090	0.6225 ± 0.0085	0.6751 ± 0.0081	0.6087 ± 0.0086
XGBoost	0.6753 ± 0.0077	0.6288 ± 0.0066	0.6707 ± 0.0046	0.6206 ± 0.0072
CatBoost	0.6457 ± 0.0139	0.6140 ± 0.0108	0.6730 ± 0.0087	0.6108 ± 0.0094

Los resultados evidencian que tanto los modelos lineales como no lineales bajo condiciones similares alcanzan desempeños competitivos en problemas de clasificación multiclase. La regresión logística multinomial se caracterizó por alcanzar, con un margen mínimo, el mejor recall-macro promedio con una estimación de 0.6751 (± 0.0081). Esto indica que el modelo fue capaz de identificar correctamente, en promedio, el 67,51% de las observaciones pertenecientes a cada causa.

Respecto a los modelos basados en boosting, el XGBoost alcanzó un mayor desempeño, en métricas como el accuracy (0,6753 ± 0,0077), el f1 score (0.6288 ± 0.0066) y especialmente

en la precisión con un valor de 0.6206((± 0.0072), es decir, en promedio el 62,06 % de las observaciones clasificadas por el modelo en cada causa correspondían realmente a dicha causa.

CatBoost, de igual manera presentó métricas con buen desempeño, sin embargo, fueron inferiores por un margen mínimo frente a los otros modelos tratados. Al compararlo con XGBoost, como parte de la familia Boosting, el modelo tuvo un recall-macro moderadamente superior que el XGBoost, pues, solo fue por una diferencia de 0.23 puntos porcentuales.

En conclusión, la validación cruzada estratificada permitió determinar el comportamiento de los modelos bajo condiciones similares en términos de estabilidad y desempeño. Los resultados sugieren que no existen diferencias significativas entre los modelos evaluados, pues cada uno presentó ventajas como desventajas, por lo cual no se determinó ningún modelo ampliamente superior. Sin embargo, para efectos de este estudio, se decidió explorar más a fondo el modelo XGBoost mediante técnicas de tuning, considerando su F1-macro promedio moderadamente más alto y su mejor desempeño global en métricas como accuracy.

Optimización de hiperparámetros del modelo XGBoost

Ya seleccionado el modelo XGBoost, se procedió a hacer un proceso de optimización de hiperparámetros por medio de GridSearchCV. Este es una herramienta que combina búsqueda en cuadricula y validación cruzada para determinar los mejores parámetros de los modelos de aprendizaje automático. El proceso calcula todas las combinaciones posibles de parámetros dentro de un rango particular, escogiendo el mejor según una métrica definida (Zhao et al., 2024). En la Tabla 11 se presentan los hiperparámetros que se aplicaron para optimizar el modelo XGBoost.

Tabla 11

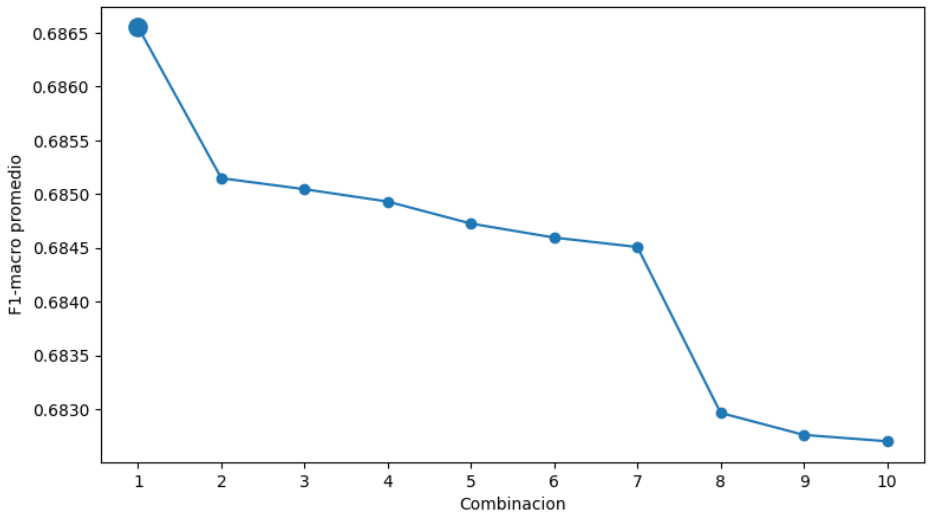
Hiperparámetros Evaluados

Hiperparámetros	Valores
learning_rate	0.01, 0.05, 0.1
max_depth	3, 5, 7
n_estimators	100, 300, 500
subsample	0.8, 1
colsample_bytree	0.8, 1

En total, el modelo se entrenó 540 veces, evaluando todas las combinaciones posibles de los hiperparámetros con una validación cruzada de 5 folds. La Figura 6, muestra que las 10 mejores combinaciones se establecieron en un f1-macro promedio de 0.68.

Figura 6

Top 10 combinaciones GridSearchCV



Fuente: Elaboración propia

La mejor combinación identificada se estableció en learning_rate = 0.05, max_depth = 3, n_estimators = 300, subsample = 0.8 y colsample_bytree = 0.8, por ende, esta configuración fue empleada para evaluar el modelo frente al conjunto de prueba y analizar su desempeño en la clasificación de cada defunción no fetal tratada.

Evaluación del Modelo XGBoost en el Conjunto de Prueba

El modelo XGBoost fue evaluado sobre el conjunto de prueba independiente (30 % de los datos), con el propósito de analizar su capacidad de generalización y examinar su comportamiento en cada uno de los grupos de causas de muerte. Esta evaluación permite verificar el desempeño del modelo en observaciones no utilizadas durante el entrenamiento y profundizar en la diferenciación estructural entre categorías.

En el conjunto de prueba, el modelo alcanzó una exactitud (accuracy) del 66 %, con un F1-macro de 0.63 y un recall-macro de 0.69. Estos resultados son consistentes con los obtenidos en la validación cruzada, lo que sugiere un desempeño estable y la ausencia de evidencia de sobreajuste significativo. En cuanto al comportamiento del modelo en cada categoría de causa de muerte, en la Tabla 12 se presentan las métricas desagregadas por clase, junto con el número de observaciones correspondientes.

Tabla 12

Resultado Métricas XGBoost

Clase	Precisión	Recall	F1-Score	Support
0	0.89	0.64	0.75	3480
1	0.29	0.59	0.39	777
2	0.70	0.83	0.76	837
Accuracy			0.66	5094

Clase	Precisión	Recall	F1-Score	Support
Macro avg	0,63	0.69	0.63	5094
Weighted avg	0,77	0,66	0.69	5094

Nota. Los valores se calculan sobre el conjunto de prueba (30 % de los registros).

El análisis por categoría evidencia un comportamiento diferenciado del modelo. Las enfermedades del sistema circulatorio (clase 0) presentan el mayor nivel de identificación, con un recall de 0.64 y un F1-score de 0.75, lo que indica que este grupo posee características estructurales relativamente bien diferenciadas dentro de las variables consideradas, aunque no de forma completamente exhaustiva.

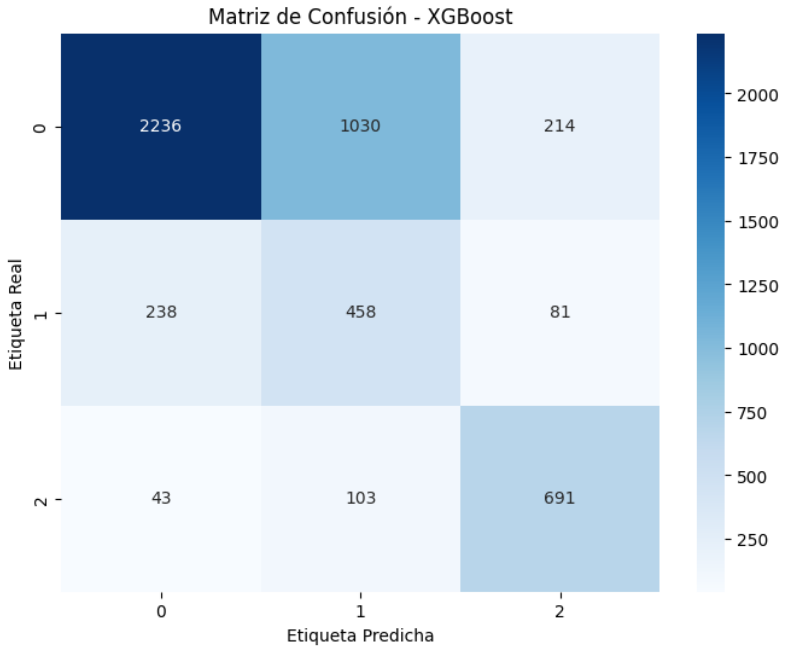
En contraste, las neoplasias (clase 1) muestran un desempeño considerablemente menor, con un recall de 0.59 y un F1-score de 0.39, evidenciando mayores dificultades para su correcta clasificación. Este resultado sugiere una posible superposición estructural entre este grupo y las demás categorías, así como una menor presencia de variables altamente discriminantes en la base de datos utilizada, lo que conduce a una alta tasa de confusión.

Por último, las causas externas (clase 2) presentan un desempeño relativamente sólido, con un recall de 0.83 y un F1-score de 0.76, lo que indica una adecuada capacidad de diferenciación por parte del modelo, cercana a la observada en las enfermedades del sistema circulatorio, aunque con un mejor balance entre sensibilidad (recall) y precisión (precision).

Bajo este panorama, se aplicó la matriz de confusión como complemento para desarrollar un análisis más certero. Esta técnica permite visualizar la distribución de aciertos y errores entre las categorías evaluadas.

Figura 7

Matriz de Confusión - XGBoost



Fuente: Elaboración propia

La matriz (ver Figura 7) refleja que gran parte de la proporción de errores se concentran es en la diferenciación de causas del sistema circulatorio (clase 0) y neoplasias (clase 1), clasificando 1030 casos de causas del sistema circulatorio como neoplasias, mientras que 238 casos de neoplasias son clasificados como enfermedades del sistema circulatorio, evidenciando una confusión bidireccional destacable entre estas dos causas. Por otro lado, la confusión de ambas clases con las causas externas (clase 2) es considerablemente menor, sugiriendo una mejor capacidad del modelo para distinguir este último grupo.

Este comportamiento indica la existencia de una superposición entre las variables que están asociadas a causas de enfermedades del sistema circulatorio y neoplasias. Es probable que ambas categorías compartan características similares en términos de edad, sexo u otros

determinantes, lo que dificulta su separación mediante modelos de clasificación basados exclusivamente en este tipo de información.

Por último, las causas externas presentan un comportamiento diferenciado, con un alto número de clasificaciones correctas (691 casos) y niveles relativamente bajos de confusión hacia las otras categorías, lo que refuerza la idea de que este grupo posee patrones más claramente identificables dentro de los datos.

En conjunto, estos resultados sugieren que el desempeño global del modelo está condicionado principalmente por la dificultad para distinguir entre enfermedades del sistema circulatorio y neoplasias, más que por una incapacidad general del modelo, lo que resalta la importancia de considerar la heterogeneidad entre clases en la interpretación de las métricas agregadas.

Análisis SHAP

Fundamento

Con el propósito de interpretar el comportamiento del modelo XGBoost seleccionado, se aplicó el método SHAP (SHapley Additive exPlanations), basado en la teoría de valores de Shapley de la teoría de juegos cooperativos. Este enfoque permite descomponer la predicción de cada observación en contribuciones marginales atribuibles a cada variable, facilitando tanto el análisis de importancia global como la interpretación local de casos individuales.

El análisis se realizó exclusivamente sobre el conjunto de prueba, de manera que los resultados reproducen el comportamiento del modelo ante datos no utilizados durante el entrenamiento, fortaleciendo así la validez interpretativa.

Dado que el modelo mostró un comportamiento diferenciado entre las tres categorías de causas de defunción no fetal, el análisis interpretativo se orienta a examinar la contribución de

las variables (magnitud) en cada causa tratada a través de graficas waterfall para mostrar explicaciones individuales y graficas beeswarm, las cuales muestra un resumen denso en información sobre cómo las características principales de un conjunto de datos impactan la salida del modelo. Cada instancia de la explicación dada está representada por un solo punto en cada fila de características. La posición x del punto está determinada por el valor SHAP (Lundberg, 2018).

Clase 0 – Enfermedades del sistema circulatorio

El beeswarm plot correspondiente a la clase de enfermedades del sistema circulatorio permite identificar las variables con mayor contribución a la predicción del modelo en el conjunto de prueba. Las variables se presentan ordenadas según su contribución promedio absoluta (mean |SHAP|), lo que establece una jerarquía de influencia estructural dentro de esta categoría. La figura 8 no solo muestra la magnitud del efecto, sino también su dirección y la relación entre valores altos y bajos de cada variable.

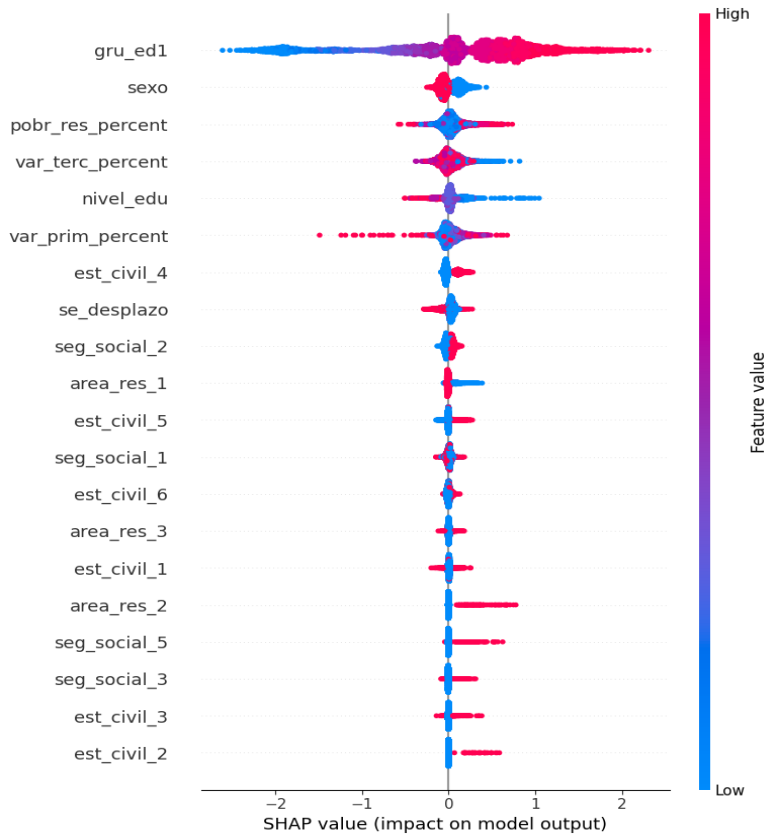
El análisis evidencia que el grupo etario (gru_ed1) es la característica más importante en promedio para clasificar o predecir esta causa tratada. A medida que aumentan los grupos etarios, la contribución de esta variable se vuelve progresivamente positiva, favoreciendo la clasificación de una observación dentro de esta categoría.

En cuanto al sexo, se presenta como la segunda de mayor promedio, donde se identifica que el sexo femenino presenta contribuciones positivas, mientras que el sexo masculino muestra contribuciones negativas en la probabilidad de pertenecer a esta categoría. No obstante, este resultado debe interpretarse en términos relativos dentro del modelo, y no como una relación causal directa. En particular, la contribución negativa asociada al sexo masculino sugiere una

mayor asociación de este grupo con otras causas de muerte, como las causas externas, más que una menor propensión intrínseca a enfermedades del sistema circulatorio.

Figura 8

Beeswarm Plot Clase 0 (Enfermedades del sistema circulatorio)



Fuente: Elaboración propia

Respecto al nivel educativo, se observa que niveles más bajos de escolaridad están asociados con un mayor aporte positivo hacia la clasificación de esta causa. Sin embargo, este efecto se encuentra estrechamente relacionado con la variable edad, dado que los grupos etarios más avanzados corresponden a cohortes que históricamente han tenido menor acceso a la educación formal. En este sentido, se evidencia una correlación estructural entre edad y nivel educativo, lo que sugiere que parte del efecto atribuido a la escolaridad podría estar capturando diferencias generacionales.

Por su parte, las variables relacionadas con la actividad económica no presentan un patrón claro en su contribución al modelo. Tanto las actividades del sector primario como del sector terciario muestran efectos heterogéneos, lo que indica un bajo poder discriminante de estas variables en la identificación de enfermedades del sistema circulatorio dentro del conjunto de datos analizado.

Por último, se observa que variables como la condición de no aseguramiento en salud (seg_social_5) y la residencia en centros poblados y áreas rurales(área_res_2) presentan contribuciones positivas.

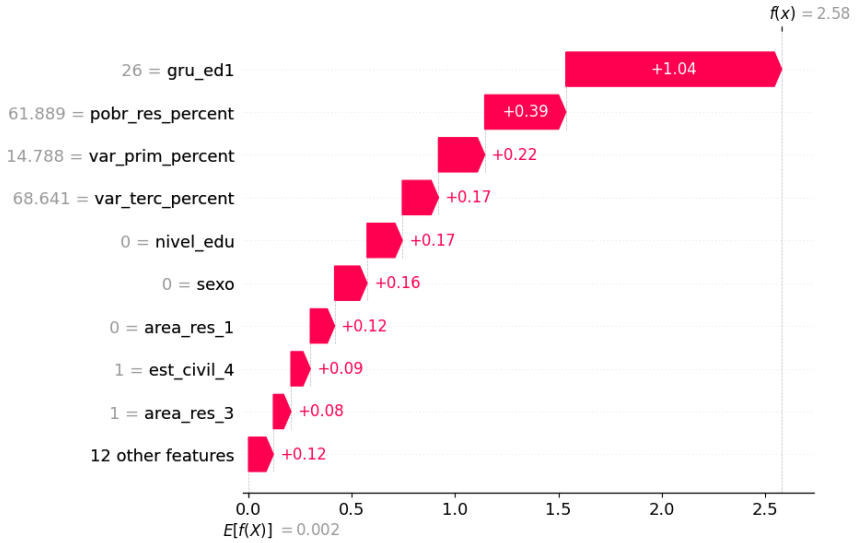
No obstante, estas variables se ubican entre las de menor importancia relativa dentro del modelo, lo que indica que, si bien aportan información, su capacidad discriminante es limitada en comparación con variables como la edad (grupo etario).

En consecuencia, estos factores pueden interpretarse como determinantes secundarios, cuya influencia, aunque consistente, es considerablemente menor frente a los principales determinantes demográficos.

A continuación, mediante el grafico waterfall (Figura 9) se presenta un caso particular exitoso con el objetivo de identificar que registros o condiciones de las variables contribuyeron al modelo para clasificarlo como causa por enfermedad del sistema circulatorio.

Figura 9

Waterfall-Enfermedad Sistema circulatorio



Fuente: Elaboración propia.

Como se puede observar, la explicación del modelo refleja que la condición correspondiente al grupo etario 26 (90 a 94 años) fue la de mayor contribución positiva, complementada con características como residencia en un municipio con predominio rural (61.9 %) y ausencia de nivel educativo. Adicionalmente, se aprecia que variables como sexo femenino y estado civil de viudez también presentaron aportes positivos dentro de la predicción realizada por el modelo. Conjuntamente, estas variables representaron las condiciones de mayor contribución para clasificar este registro como una defunción asociada a enfermedades del sistema circulatorio. En este sentido, este tipo de graficas SHAP posibilita identificar factores que contribuyen dentro del modelo, sirviendo como referencia para la elaboración de hipótesis en torno a posibles perfiles asociados a esta causa de defunción.

Clase 1- Causas Neoplasias.

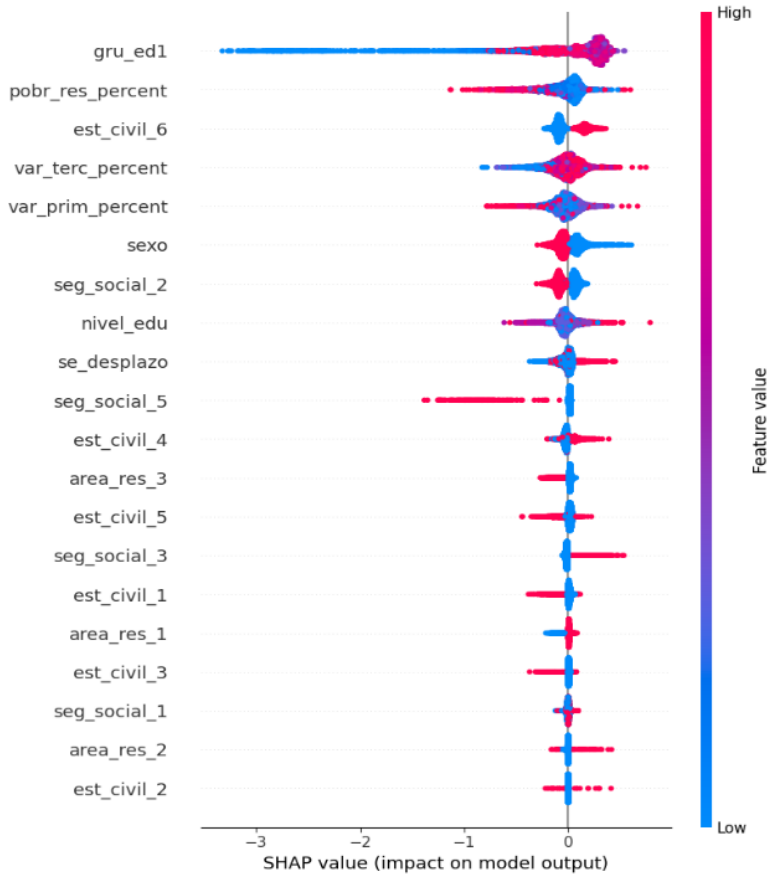
En el caso de las neoplasias (Figura 10), el análisis de los valores SHAP permite identificar los patrones que el modelo utiliza para realizar sus predicciones, incluso en una categoría donde el desempeño predictivo fue más limitado. A diferencia de lo observado en las enfermedades del sistema circulatorio, la variable edad mantiene un efecto positivo en los grupos etarios más altos, pero con un patrón menos definido, evidenciando una mayor dispersión en su contribución. En particular, se observa que no solo los grupos de mayor edad, sino también algunos grupos intermedios, presentan aportes positivos, lo que sugiere una menor capacidad de esta variable para discriminar claramente esta categoría frente a otras.

Asimismo, variables como el estado civil casado (est_civil_6) y el sexo femenino presentan contribuciones positivas hacia la clasificación de neoplasias. Por otro lado, la categoría seg_social_5 muestra un efecto negativo, lo cual es consistente con los resultados observados en enfermedades del sistema circulatorio, donde esta condición se asocia más fuertemente con dicha categoría. En contraste, la pertenencia a regímenes especiales de seguridad social, como el correspondiente a seguridad social de excepción (seg_social_3), presenta contribuciones positivas hacia la clasificación como neoplasias.

No obstante, es importante destacar que estas contribuciones no deben interpretarse como relaciones causales ni como determinantes reales de las neoplasias, sino como los criterios internos que el modelo utiliza para realizar sus predicciones. En este sentido, la presencia de patrones menos definidos y, en algunos casos, inconsistentes, es coherente con el bajo desempeño observado en esta clase, reflejando la dificultad del modelo para identificar señales claras que permitan diferenciar las neoplasias de otras causas, particularmente de las enfermedades del sistema circulatorio.

Figura 10

Beeswarm Plot Clase 1 (Neoplasias)



Fuente: Elaboración propia

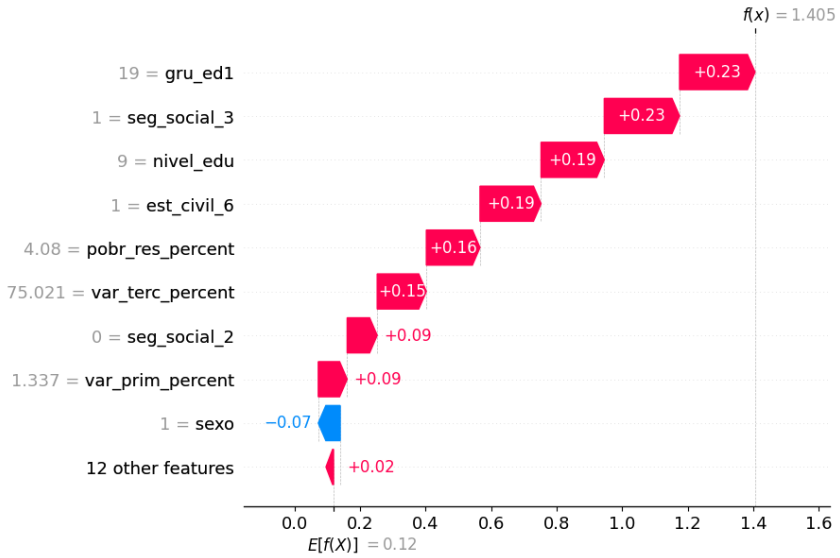
Con el fin de seguir con la dinámica, se presenta un caso específico donde el modelo logró identificar “exitosamente” la causa tratada. Aunque en este punto, las clasificaciones se deben más a la semejanza de clases que a un perfil diferenciable. La Figura 11 muestra qué variables tomó el modelo para decidirse por la causa tratada.

Como se puede contemplar, la persona fallecida pertenecía al grupo etario 19 (55 a 59 años) y presentaba régimen de seguridad de excepción. Estas dos variables fueron las que presentaron mayor impacto en igual magnitud, seguido del nivel educativo con condición de profesional y estado civil casado. En conjunto, estas 4 condiciones fueron as que presentaron

mayor contribución dentro de la predicción del modelo. Una posible respuesta a esta decisión es que las causas relacionadas con el sistema circulatorio se presentan más en personas de mayor edad y con menor nivel educativo, por lo cual, la clasificación observada podría responder más a un proceso de descarte entre categorías que a la identificación de un perfil claramente diferenciable asociado a las neoplasias.

Figura 11

Waterfall- Causa Neoplasia



Fuente: Elaboración propia.

Clase 2 – Causas Externas

El análisis de los valores SHAP para las causas externas (Figura 12) evidenció que el grupo etario influye en gran proporción para la decisión del modelo. A diferencia de las causas del sistema circulatorio, el beeswarm plot muestra que grupos etarios bajos presentan mayores contribuciones positivas. En cuanto al sexo, se observa que el sexo masculino presenta contribuciones positivas relevantes, lo que sugiere una mayor contribución de esta variable dentro de la clasificación de causas externas realizadas por el modelo.

En el nivel educativo, se identifica que mayores niveles de escolaridad tienden a generar contribuciones positivas en la clasificación de causas externas. Sin embargo, este resultado debe interpretarse con cautela, ya que puede estar influenciado por la estructura etaria de la población. En particular, los grupos más jóvenes que presentan mayor probabilidad de causas externas también han tenido históricamente mayor acceso a la educación, lo que sugiere la presencia de una relación indirecta mediada por la edad.

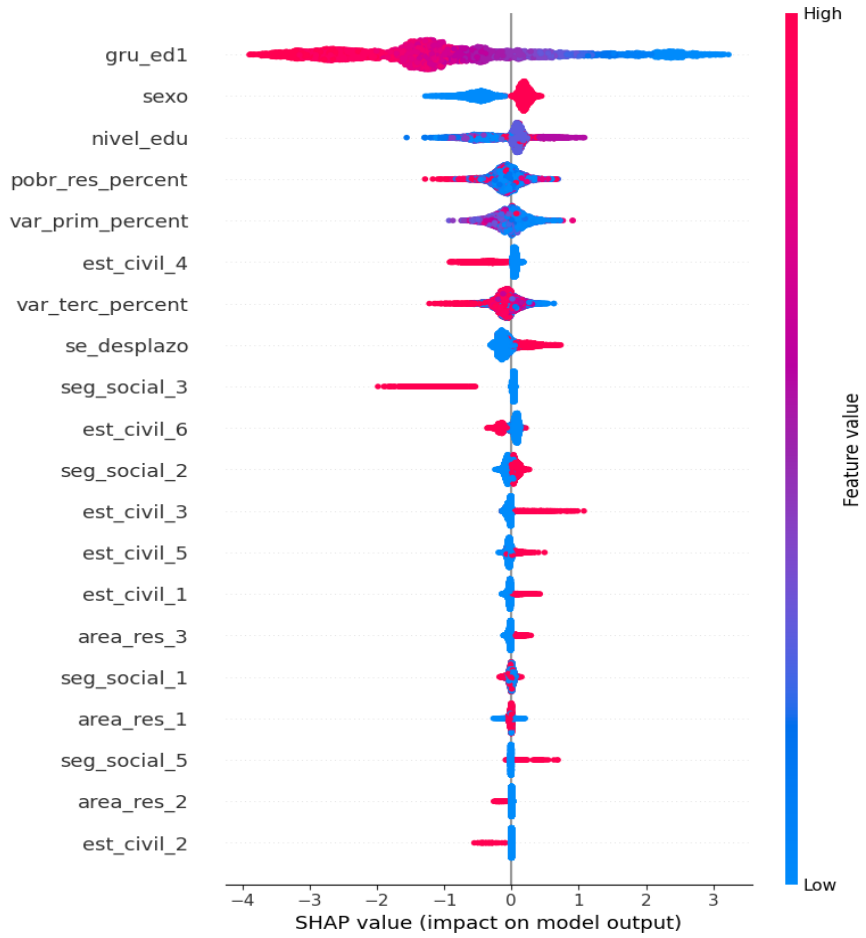
Por su parte, las variables económicas muestran patrones menos consistentes. En particular, se observa que un mayor porcentaje de participación en el sector terciario está asociado con contribuciones negativas, lo que indicaría una menor probabilidad de clasificación en causas externas. Sin embargo, este efecto no es tan marcado como el de las variables demográficas, lo que sugiere un menor poder discriminante.

Finalmente, la variable dicotómica asociada a la condición de desplazamiento entendida como el hecho de que la defunción ocurra en un municipio distinto al de residencia presenta contribuciones positivas hacia la clasificación en causas externas. Este resultado podría estar capturando dinámicas asociadas a movilidad, accidentes o eventos violentos fuera del lugar habitual de residencia.

En conjunto, estos hallazgos muestran que la clasificación o predicción de causas externas está fuertemente determinada por factores demográficos como la edad y el sexo, mientras que otras variables reflejan efectos indirectos o contextuales, en línea con la naturaleza más circunstancial de este tipo de causas de muerte.

Figura 12

Beeswarm Plot Clase 2 (Externas)



Fuente: Elaboración propia

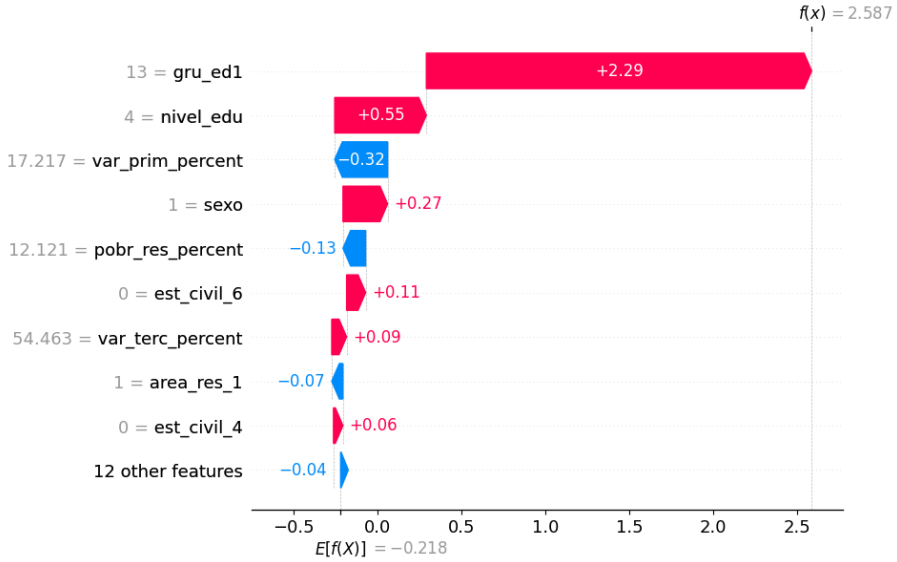
Para profundizar el perfil referenciado, se analiza un caso individual correctamente clasificado (Figura 13). El modelo tomó como principal variable el grupo etario, que en este caso correspondía a una persona de 25 a 29 años (grupo etario 13).

Seguidamente, se complementa con la variable de nivel de educación igual 4 (media académica o clásica) y sexo masculino. Por último, se puede observar que, para este perfil, el modelo dio un f(x)= 2.587, el cual, al transformarse en probabilidad es más del 90%, es decir,

indica un alto nivel de confianza del modelo al clasificar esta observación dentro de la categoría de causas externas.

Figura 13

Waterfall- Causa Externa



Fuente: Elaboración propia.

Análisis comparativo de patrones estructurales

El análisis comparativo de las tres categorías evidencia que el modelo aprende estructuras diferenciadas según el tipo de causa de muerte, aunque con distintos niveles de claridad en cada caso. En las enfermedades del sistema circulatorio, la dimensión etaria constituye el eje dominante de clasificación: los grupos de mayor edad incrementan la contribución positiva hacia la clasificación de esta categoría, mientras que las edades más jóvenes actúan como factores de exclusión. Esta segmentación relativamente clara, reforzada por variables como el sexo y el nivel educativo (en interacción con la edad), explica el mejor desempeño observado en esta clase.

En las causas externas, la estructura predictiva también se organiza en torno a variables demográficas, aunque con un patrón inverso. En este caso, la combinación de edad joven y sexo

masculino define el perfil predominante, mientras que las edades avanzadas y características asociadas a etapas posteriores del ciclo de vida funcionan como elementos de descarte. Adicionalmente, variables como la condición de desplazamiento presentan contribuciones positivas, sugiriendo la relevancia de factores contextuales asociados a movilidad o eventos fuera del lugar de residencia. En esta categoría, las variables económicas y territoriales muestran una influencia más limitada.

Por el contrario, las neoplasias presentan una estructura menos definida, caracterizada por patrones más difusos y una mayor superposición con las enfermedades del sistema circulatorio. Aunque la edad mantiene un efecto positivo en grupos etarios altos, este no es tan claramente diferenciador, ya que también se observan contribuciones positivas en grupos intermedios como el caso específico analizado previamente. De igual manera, variables como el sexo femenino, el estado civil y ciertas categorías de afiliación al sistema de salud presentan efectos heterogéneos.

En conjunto, estos resultados indican que el modelo logra diferenciar con mayor claridad las categorías que presentan perfiles sociodemográficos bien definidos, como las enfermedades del sistema circulatorio y las causas externas, mientras que su capacidad disminuye cuando la categoría presenta mayor heterogeneidad o comparte características estructurales con otras clases, como ocurre en el caso de las neoplasias. La aplicación de los valores SHAP permitió identificar estos ejes diferenciadores y comprender cómo el modelo organiza la clasificación en cada caso, evidenciando tanto sus fortalezas como sus limitaciones.

Discusión

Mediante la aplicación del modelo de aprendizaje automático (XGBoost) y su respectiva interpretación mediante técnicas como SHAP, se evidenció que las categorías de causas de defunción no fetal, especialmente las enfermedades del sistema circulatorio y causas externas presentaron un perfil sociodemográfico característico en el departamento de Santander para el periodo 2015-2019.

El desempeño general estuvo marcado por un accuracy global del 66%, f1-score de 63%, recall de 69% y una precisión del 63%; estas métricas reflejan principalmente la mayor capacidad del modelo para discriminar las enfermedades del sistema circulatorio y las causas externas. No obstante, es pertinente considerar que la categoría de enfermedades del sistema circulatorio fue la de mayor representación en el dataset estudiado. Como se señaló anteriormente, las enfermedades isquémicas del corazón y cerebrovasculares se mantuvieron durante el periodo estudiado como las dos principales causas de defunción no fetal, situación que pudo favorecer el aprendizaje de patrones asociados a esta categoría y contribuir a su elevada precisión (89%) y f1-score (75%).

Sin embargo, la utilidad de este estudio no solo radica en determinar el potencial de los microdatos para identificar la categoria mayoritaria, sino también en evaluar el desempeño en las clases minoritarias tratadas como lo son las causas externas y neoplasias. Por este motivo, se implementaron estrategias de aprendizaje sensible al costo (cost-sensitive learning), con el fin de reducir el sesgo hacia la clase de mayor representación y asignar una mayor penalización a los errores cometidos sobre las demás categorías durante el proceso de entrenamiento.

Bajo este panorama, el modelo aplicado presentó un desempeño favorable para la identificación de causas externas. Esto se puede evidenciar en las métricas obtenidas sobre el

conjunto de prueba, donde alcanzó una precisión del 70% y un recall del 83%. Esta última métrica resulta especialmente relevante, ya que indica que el modelo logró identificar correctamente una elevada proporción de los registros que realmente correspondían a esta categoría. En síntesis, estos resultados sugieren que las variables sociodemográficas y contextuales complementadas con métodos de balanceo de clases, presentan potencial para diferenciar las causas externas frente a las enfermedades del sistema circulatorio y neoplasias.

No obstante, las neoplasias no presentaron un desempeño favorable, lo que evidencia una de las principales limitaciones del conjunto de variables disponibles en los microdatos de defunción no fetal. Pese a que el modelo obtuvo un recall del 59%, es decir, consiguió identificar una proporción mesurada de los casos reales de esta categoría, su baja precisión del 29%, indica una elevada presencia de falsos positivos. Lo expuesto sugiere que la capacidad del modelo aplicado no dependió exclusivamente de la frecuencia de las clases, puesto que, como se evidenció anteriormente, el modelo alcanzó métricas superiores en la clasificación de las causas externas, a pesar de que esta categoría presentaba un tamaño de muestra similar al de las neoplasias.

Esto se puede apreciar en la matriz de confusión (Figura 7) donde se evidencia una gran superposición de esta respecto a las enfermedades del sistema circulatorio. En términos absolutos el modelo clasificó 238 registros como enfermedades circulatorias, 81 como causas externas y 458 como neoplasias, es decir, una proporción considerable de los casos mal clasificados de neoplasias tiende a concentrarse en la clase mayoritaria del conjunto de datos.

Por otro lado, la matriz de confusión también evidencia que esta superposición se presenta desde las enfermedades del sistema circulatorio hacia las neoplasias, lo que sugiere la

existencia de características sociodemográficas y contextuales compartidas entre ambas categorías y una posible dificultad del modelo para distinguirlas entre sí.

Con estos resultados se podría concluir que los modelos no deben evaluarse exclusivamente desde su desempeño global cuando se tratan de problemas multiclase, pues gran parte de sus métricas pueden estar sesgadas por el buen desempeño de la clase mayoritaria. En este sentido, resulta necesario complementar el análisis con métricas desagregadas por clase, con el fin de obtener un diagnóstico más robusto del comportamiento del modelo frente a cada categoría.

A partir de estos hallazgos se procedió a la interpretación del modelo mediante técnicas de explicabilidad como SHAP, con el fin de identificar las variables que contribuyen a la clasificación de cada una de las categorías analizadas. Respecto a la clase relacionada con las enfermedades del sistema circulatorio, se evidenció que estuvo principalmente influenciada por variables como grupos etarios elevados, niveles educativos bajos y sexo femenino. No obstante, dichas contribuciones no deben interpretarse como relaciones causales, sino como una explicación del comportamiento predictivo del modelo.

En este sentido, los resultados obtenidos presentan coherencia con la evidencia epidemiológica existente. Por ejemplo. Acosta & Romero (2014), determinaron que las enfermedades del sistema circulatorio, como las enfermedades isquémicas del corazón y cerebrovasculares aumentan con la edad y presentan diferencias según el sexo. Los autores encontraron que, una vez se alcanzan los 65 años, la probabilidad de morir por enfermedades del sistema circulatorio es de 43,6% para las mujeres y 41,3% para los hombres, es decir, una diferencia de 2,3 puntos porcentuales.

Considerando esto, los resultados de este estudio son congruentes con el entorno colombiano, el cual se caracteriza por un aumento progresivo de la esperanza de vida y el envejecimiento de la población, factores que, según los mismos autores, han generado cambios en el perfil epidemiológico del país.

Respecto a la clasificación de causas externas, el modelo estuvo especialmente influenciado por variables como grupos etarios bajos y sexo masculino Este patrón coincide con lo señalado por Otero (2013), quien, con base en la literatura sobre violencia y actividades criminales (Bonilla, 2010; Yunez, 1999; OMS, 2010), reportó que los homicidios, como parte de las causas externas, son más frecuentes en los hombres que en las mujeres, sobre todo en jóvenes entre los 15 y 44 años.

De igual manera, Martínez (2025) documentó que Colombia presenta una sobremortalidad masculina por causas externas, sobre todo en edades de 10 a 14 años, donde por cada 100 mujeres fallecidas hay 182 hombres. En este sentido, los patrones identificados por el modelo resultan consistentes con la evidencia reportada en la literatura, lo que sugiere una mayor exposición de la población masculina a factores asociados con este tipo de causas de muerte.

Por último, sobre las causas neoplásicas, en el gráfico beeswarm se identificó que los grupos etarios presentaron un patrón más disperso, al igual que variables como el sexo y los niveles educativos, evidenciando una superposición con las demás causas tratadas, especialmente con las enfermedades del sistema circulatorio. Este resultado es consistente con lo planteado por Acosta & Romero (2014), quienes señalaron que la probabilidad de morir por esta causa afecta de manera similar tanto a hombres como mujeres, manteniendo variaciones relativamente estables a lo largo de distintos grupos de edad. Por ende, la menor capacidad del modelo para

discriminar esta categoría podría estar asociada a la ausencia de un patrón sociodemográfico claramente diferenciado.

En este sentido, el uso de SHAP permite interpretar los criterios empleados por el modelo en sus predicciones, además de evidenciar patrones que podrían servir como punto de referencia para futuras hipótesis y análisis de salud pública. No obstante, es importante señalar que estas interpretaciones dependen del desempeño del modelo, por lo que su validez explicativa está condicionada a la calidad de las métricas y al poder predictivo alcanzado previamente.

Limitaciones

Para interpretar los resultados obtenidos adecuadamente se deben considerar las limitaciones que presenta el estudio.

En primera instancia, este estudio empleó datos de la región de Santander para el periodo 2015-2019, por lo tanto, los resultados solo pueden interpretarse bajo dicho contexto geográfico y temporal, lo cual limita el alcance del estudio a nivel nacional, aunque puede utilizarse como guía en estudios que se hagan a futuro en otras regiones del país. En este aspecto, se restringe la generalización de los hallazgos a otros contextos diferentes al analizado**.**

De igual manera, el análisis se centró en las categorías de causas de defunción no fetal más representativas (enfermedades del sistema circulatorio, neoplasias y causas externas) conforme a la agrupación de la lista 6/67 OMS/OPS. Por tanto, los resultados obtenidos se deben interpretar en virtud de estas categorías, dado que no se evaluó el desempeño del modelo frente a la totalidad de las categorías de causas establecidas en la mencionada lista.

En segunda instancia, los microdatos de defunciones no fetales son registros producidos a partir de la información que provienen de los certificados de defunción. Estos certificados se caracterizan por contener información sociodemográfica, geográfica y administrativa. Por lo tanto, el alcance de este estudio se limita al tipo de información disponible en dichos registros. Como se expuso en el marco de referencia, la investigación aplicó el concepto de reutilización de datos (data reuse), entendido como el uso de información para fines distintos a aquellos para los que fue originalmente recopilada. Por tanto, el análisis estuvo condicionado a las variables contenidas en la fuente de información, lo cual impidió incorporar variables adicionales que pudiesen contribuir al problema de clasificación analizado. Además, el DANE, como entidad oficial para la difusión de información estadística, somete estos registros a un proceso de

anonimización, dejando variables como la edad agrupadas en grupos etarios quinquenales, lo que limita parcialmente la precisión de esta variable.

Por otro lado, se evidenció que un gran porcentaje de los registros relacionados con nivel educativo y estado civil se encontraban dentro de la categoría “Sin información”. Independientemente de si se eliminan esos registros o se aplican técnicas de imputación, el tratamiento puede influir en el aprendizaje y desempeño del modelo de aprendizaje automático. Lo anterior puede introducir sesgos en la distribución de las variables, asociados a la calidad de los registros administrativos.

Bajo este contexto, el modelo se entrenó con los tipos de variables descritas anteriormente, dando como resultado un desempeño aceptable para clasificar causas de defunción no fetal relacionadas con el sistema circulatorio y externas. Pese a tener un recall de 59%, la precisión para las neoplasias se estableció en 29%, muy por debajo respecto a las otras causas analizadas. Una de las posibles razones es que las variables contenidas en los microdatos no fueron suficientes para diferenciar esta causa frente a otras, especialmente las enfermedades del sistema circulatorio, pues como se describió anteriormente, los falsos positivos tienden a relacionarse con la clase mayoritaria o más frecuente, lo que evidencia una superposición entre ambas categorías en el espacio de variables.

El método SHAP es una herramienta utilizada para explicar la contribución de las variables en las predicciones del modelo respecto a cada causa tratada. Cuando el desempeño del modelo es aceptable, la interpretación mediante SHAP es más consistente. De lo contrario, se observarán patrones dispersos, limitando la capacidad de explicar por qué el modelo tomó esa decisión. En este sentido, su capacidad explicativa depende directamente del desempeño del modelo y de la calidad de las variables de entrada.

En conclusión, los modelos de aprendizaje automático, como el XGBoost, explicados mediante SHAP, permiten marcar puntos de referencia para generar hipótesis sobre perfiles asociados a cada causa, sin embargo, los resultados están condicionados a la disponibilidad, calidad y nivel de detalle que estos representan. Por lo tanto, los resultados deben interpretarse dentro del alcance y limitaciones metodológicas del presente estudio. En consecuencia, no deben asumirse como relaciones causales ni generalizables a otros contextos.

Conclusiones

Con base en el trabajo realizado y los resultados obtenidos se llegó a las siguientes conclusiones: Tener un buen sistema de información en salud garantiza la producción, el análisis, la difusión y la confiabilidad de los datos. El DANE como ente oficial para la difusión y producción de las estadísticas oficiales como lo son las estadísticas vitales en Colombia, ha ido mejorando a lo largo del tiempo, incluso alcanzando en la última década una tasa mínima de error cercana al 0,3%. Este marco de calidad de datos que dispone el DANE permitió dar cumplimiento al primer objetivo propuesto de manera confiable.

Allí se constató que, en Santander, para el periodo 2015-2019, más de la mitad de los decesos registrados correspondían al género masculino, específicamente el 54.7%, frente a un 45,3% del femenino. Respecto a la distribución etaria se identificó que aproximadamente el 38% concernían a edades entre los 75 y 84 años. Este comportamiento va en línea con la esperanza de vida en Colombia, la cual se ubicó en el 2020 por encima de los 77 años (Rueda, 2022).

Por último, las principales causas de defunción no fetal en este periodo de tiempo fueron las relacionadas con enfermedades del sistema circulatorio como las isquémicas del corazón y las cerebrovasculares, consideradas como crónicas por su condición a causa del envejecimiento de la población. Sin embargo, se destaca la posición de las causas externas como los accidentes de transporte y homicidio, las cuales se establecieron dentro del top 10, incluso por encima, en algunos periodos, de causas como las enfermedades del hígado y los tumores.

Una vez hecho el análisis descriptivo, se procedió a hacer un proceso de depuración para aplicar los modelos de aprendizaje automático y dar cumplimiento al objetivo número dos. Para este se aplicaron 3 modelos de aprendizaje automático, uno con enfoque lineal como lo es la regresión logística Multinomial y dos con enfoque no lineal, como el XGBoost y CatBoost

pertenecientes a la familia boosting. Para cada modelo se aplicaron las trasformaciones necesarias para ser ejecutado correctamente, como lo fue el caso de la Regresión logística y el XGBoost, requiriendo transformar las variables categóricas nominales en dicotómicas.

La validación cruzada con 5 fold permitió evaluar el modelo de mejor desempeño y estabilidad promedio respecto a métricas como el accuracy, f1-macro, recall macro y precisión para la clasificación de las causas analizadas. El XGBoost fue el de mejor desempeño global con un margen mínimo respecto a los otros modelos, presentando un accuracy del 67,53% con una desviación estándar de 0,0077, un recall de 67,07% con desviación estándar de 0,0046 y una precisión del 62,06% con desviación estándar de 0,0072.

Haciendo énfasis en cada categoría analizada, el modelo después de pasar por un proceso de optimización de hiperparámetros, presentó un mejor desempeño para las categorías de enfermedades del sistema circulatorio y causas externas en el departamento de Santander, con una precisión del 89% y 70%, respectivamente, y un recall del 64% y 83%. En cambio, la categoría neoplásica reflejó un desempeño inferior: el modelo solo logró obtener una precisión del 29% y un recall de 59%, resultados que se atribuyen a la existencia de características sociodemográficas y contextuales compartidas especialmente con las enfermedades del sistema circulatorio, generando una superposición entre ambas categorías que dificultó su diferenciación por parte del modelo. Con lo anterior, se da por cumplido el objetivo número tres.

Como objetivo final, se utilizó el método SHAP con graficas beeswarm y waterfall para explicar cómo influyen las variables en las decisiones que toma el modelo para clasificar las categorías analizadas. A través de este, se evidenció que, para clasificar la categoría de enfermedades del sistema circulatorio, el modelo asigna mayore relevancia a grupos etarios altos y niveles educativos bajos. Estas variables están estrechamente relacionadas, pues la población

con edad avanzada en el periodo de tiempo establecido se caracterizó por tener bajos niveles de escolarización. Otra variable que influyó fue el sexo femenino; no obstante, es pertinente aclarar que estas relaciones no son causales. Es decir, el modelo no pretende afirmar que, por ser mujer, se fallece por esta causa, sino que asocia con mayor frecuencia este género frente a las otras causas tratadas, como las neoplásicas y externas. De igual manera, sucede con el porcentaje de población rural en el municipio de residencia de la persona fallecida; el modelo relaciona mayores porcentajes de ruralidad con enfermedades del sistema circulatorio frente a las otras; nuevamente, esto no quiere decir que en municipios altamente rurales la población solo fallezca por esta causa.

En cuanto a las causas externas, en las gráficas se identificó una relación inversa a las causas relacionadas con el sistema circulatorio, donde grupos etarios bajos y niveles educativos medios-altos se relacionaban más con esta causa. A diferencia de lo que se explicó anteriormente, las nuevas generaciones tienen mejor nivel de escolaridad, lo que podría explicar este comportamiento. Por otro lado, el género masculino también representa un papel importante, al relacionarse más con esta causa frente a las otras. Esto no necesariamente significa que ser hombre implique fallecer por una causa externa.

Por último, las gráficas para la categoría neoplásica mostraron una influencia dispersa, sin poderse detectar un perfil diferenciador, coincidiendo con el bajo desempeño presentado por el modelo para diferenciar esta categoría frente a las causas externas y, especialmente, frente a enfermedades del sistema circulatorio.

En conclusión, el ejercicio permitió identificar el alcance de los microdatos de defunciones no fetales complementados con variables contextuales municipales, dando como una alternativa metodológica el modelo XGBoost para clasificar las categorías analizadas, donde

especialmente se identificaron perfiles asociados a las causas por enfermedades del sistema circulatorio y externas. De igual modo el uso de técnicas de interpretabilidad como SHAP permitió comprender el aporte de las variables en las decisiones del modelo, facilitando una lectura más transparente de los resultados.

En este sentido, los resultados evidencian el potencial de los enfoques de aprendizaje automático y de interpretabilidad como herramientas complementarias para el análisis de la mortalidad, aportando evidencia útil para la toma de decisiones en salud pública en el departamento de Santander.

Recomendaciones

Con base en los hallazgos del estudio, se recomienda seguir fortaleciendo la calidad de información. Pese a que Colombia tiene un sistema de información con estándares de calidad, aun se presencian variables con un gran porcentaje de información faltante, las cuales pueden servir en proyectos futuros para un entrenamiento más robusto de modelos de aprendizaje automático.

De igual manera, se sugiere explorar otros métodos que permitan incorporar este tipo de registros con información incompleta en los modelos, ya sea mediante técnicas de imputación o enfoques que integren la ausencia de información como otra característica, y así reducir la pérdida de información condicionalmente útil para el análisis

Por otro lado, se recomienda replicar el mismo estudio, pero con otros periodos de tiempo y contextos territoriales, con el objetivo de evaluar la consistencia de los resultados en diferentes años y zonas, y el grado en que los patrones dependen de condiciones específicas del entorno o del periodo analizado.

Por último, se recomienda manejar distintos niveles de desagregación de las causas de defunción, llevando a los modelos a aprender patrones de enfermedades o causas más específicas, dependiendo de la necesidad que requiera el departamento o municipio analizado.

Referencias

AbouZahr, C., De Savigny, D., Mikkelsen, L., Setel, P. W., Lozano, R., & Lopez, A. D. (2015). Towards universal civil registration and vital statistics systems: The time is now. *The Lancet, 386* (10001), . 1407–1418.https://doi.org/10.1016/S0140-6736(15)60170-2

Acosta, K., y Romero, J. (2014). *Cambio recientes en las principales causas de mortalidad en Colombia*. https://repositorio.banrep.gov.co/server/api/core/bitstreams/e4e8a454-9003-497e-aa9b-8956c70d9825/content

Araujo, P., Martínez, D., y Contreras, J. (2024). Modelos predictivos en la clasificación de nacidos vivos y mortinatos: un estudio comparativo entre técnicas de machine learning y regresión logística en función de variables sociodemográficas y clínicas. *Revista Ciencias Biomédicas*, *13*(4), 175–189. https://doi.org/10.32997/rcb-2024-4940

Bakirarar, B. y Elhan, A. (2023). Class Weighting Technique to Deal with Imbalanced Class Problem in Machine Learning: Methodological Research. *Turkiye Klinikleri Journal of Biostatistics*, *15*(1), 19–29. https://doi.org/10.5336/biostatic.2022-93961

Beysolow, T. (2017). Introduction to deep learning using R: A step-by-step guide to learning and implementing deep learning models using R. In *Introduction to Deep Learning Using R: A Step-by-Step Guide to Learning and Implementing Deep Learning Models Using R*. Apress Media LLC. https://doi.org/10.1007/978-1-4842-2734-3

Castañeda, G., & Eslava, J. (2024). Tendencias en la mortalidad por accidentes de tránsito en motocicleta en Colombia, 2008-2021. *Revista Panamericana de Salud Pública*, *48*, 1. https://doi.org/10.26633/RPSP.2024.44

Chen, T., & Guestrin, C. (2016). XGBoost. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Dávila, C., y Pardo, A. (2019). Violencia y accidentes mortales: análisis de la mortalidad por causas externas en Colombia y México, 1998-2015. Papeles de población, 99, 249-273. http://dx.doi.org/10.22185/24487147.2019.99.10

Departamento Administrativo Nacional de Estadística. (2004). *Manual de crítica y codificación de certificados de nacido vivo y de defunción – EEVV*. https://microdatos.dane.gov.co/index.php/catalog/366/download/5594

Departamento Administrativo Nacional de Estadística. (2019). *Anuario Nacional de Estadísticas Vitales*. https://www.dane.gov.co/files/investigaciones/poblacion/anuario-EEVV-2019/anuario-nacional-de-estadisticas-vitales-colombia-2019.pdf

Galván, M., y Medina, F. (2007). *Imputación de datos: teoría y práctica*. Cepal.

Han, J., Kamber, M., & Pei, J. (2012). Classification. In *Data Mining* (pp. 327–391). Elsevier. https://doi.org/10.1016/B978-0-12-381479-1.00008-3

Law, M. (2006). Reduce, Reuse, Recycle: Issues in the Secondary Use of Research Data. *IASSIST Quarterly*, *29*(1), 5. https://doi.org/10.29173/iq599

Lundberg, S. (2018). *SHAP documentation*. https://shap.readthedocs.io/en/latest/index.html

Lundberg, S., & Lee, S.-I. (2017). *A Unified Approach to Interpreting Model Predictions* [ponencia]. Conference on Neural Information Processing Systems. DOI:10.48550/arXiv.1705.07874

Mahapatra, P., Shibuya, K., Lopez, A. D., Coullare, F., Notzon, F. C., Rao, C., & Szreter, S. (2007). Civil registration systems and vital statistics: successes and missed opportunities. *The Lancet*, *370*(9599), 1653–1663. https://doi.org/10.1016/S0140-6736(07)61308-7

Martínez, L. C. (2025). *Mortalidad por causas externas en menores de 15 años en Colombia: análisis de tendencias, magnitud y diferencias regionales (2015-2020)*. https://bdigital.uexternado.edu.co/handle/001/27350

Ministerio de Salud y Protección Social. (2024). *Manual de Principios y Procedimientos-Sistema de Registro Civil y Estadísticas Vitales*. https://www.minsalud.gov.co/sites/rid/Lists/BibliotecaDigital/RIDE/VS/ED/GCFI/manual-pyp-sistema-registro-civil-estadisticas-vitales.pdf

Ortiz, V. (2020). *Aplicación de técnicas de aprendizaje automático para la segmentación y clasificación de características sociodemográficas asociadas a tasas de mortalidad infantil utilizando datos reportados por el DANE Colombia entre los años 2008 al 2017*. http://hdl.handle.net/20.500.12010/19171

Otero, A. (2013). *Diferencias departamentales en las causas de mortalidad en Colombia.* Banco de la República de Colombia. https://repositorio.banrep.gov.co/server/api/core/bitstreams/e105a763-ebd0-4191-85cd-8ea1d8eb204c/content

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2019). *CatBoost: unbiased boosting with categorical features*. https://doi.org/10.48550/arXiv.1706.09516

Rueda Hernández, L. V. (2022). Esperanza de vida e inicio de la etapa de adulto mayor. *Salud Uninorte*, *38*(01), 5–20. https://doi.org/10.14482/sun.38.1.613.041

Sun, G., & Khoo, C. S. G. (2017). Social science research data curation. *Libellarium: Časopis Za Istraživanja u Području Informacijskih i Srodnih Znanosti*, *9*(2), 59–80. https://doi.org/10.15291/libellarium.v9i2.291

Tarawneh, A. S., Hassanat, A. B., Altarawneh, G. A., & Almuhaimeed, A. (2022). Stop Oversampling for Class Imbalance Learning: A Review. *IEEE Access*, *10*, 47643–47660. https://doi.org/10.1109/ACCESS.2022.3169512

van de Sandt, S., Dallmeier-Tiessen, S., Lavasa, A., & Petras, V. (2019). The Definition of Reuse. *Data Science Journal*, *18*. https://doi.org/10.5334/dsj-2019-022

Weiss, G. M., McCarthy, K., & Zabar, B. (2007). Cost-sensitive learning vs. sampling: Which is best for handling unbalanced classes with unequal error costs?. *Dmin*, *7*(35–41), 24.

World Health Organization. (2007). *Everybody's business : strengthening health systems to improve health outcomes : WHO's frmaework for action*. https://www.who.int/publications/i/item/everybody-s-business----strengthening-health-systems-to-improve-health-outcomes

Wu, Y.-T., Niubo, A. S., Daskalopoulou, C., Moreno-Agostino, D., Stefler, D., Bobak, M., Oram, S., Prince, M., & Prina, M. (2021). Sex differences in mortality: results from a population-based study of 12 longitudinal cohorts. *Canadian Medical Association Journal*, *193*(11), E361–E370. https://doi.org/10.1503/cmaj.200484

Zhao, Y., Zhang, W., & Liu, X. (2024). Grid search with a weighted error function: Hyper-parameter optimization for financial time series forecasting. *Applied Soft Computing*, *154*, 111362. https://doi.org/10.1016/j.asoc.2024.111362

Apéndices

Apéndice A *Tabla variables categóricas nominales transformadas a dicotómicas*

Código	Descripción
est_civil_1	No estaba casado(a) y llevaba dos o más años viviendo con su pareja
est_civil_2	No estaba casado(a) y llevaba menos de dos años viviendo con su pareja
est_civil_3	Estaba separado(a), divorciado(a)
est_civil_4	Estaba viudo(a)
est_civil_5	Estaba soltero(a)
est_civil_6	Estaba casado(a)
seg_social_1	Contributivo
seg_social_2	Subsidiado
seg_social_3	Excepción
seg_social_4	Especial
seg_social_5	No asegurado
area_res_1	Cabecera municipal
area_res_2	Centro poblado (Inspección, corregimiento o caserío)
area_res_3	Rural disperso

Apéndice B *Tabla Codificación Grupo Etario*

Código Grupo etario	Descripción
0	Menor de una hora
1	Menor de un día
2	De 1 a 6 días
3	De 7 a 27 días
4	De 28 a 29 días
5	De 1 a 5 meses
6	De 6 a 11 meses
7	De 1 año
8	De 2 a 4 años
9	De 5 a 9 años
10	De 10 a 14 años
11	De 15 a 19 años
12	De 20 a 24 años
13	De 25 a 29 años
14	De 30 a 34 años
15	De 35 a 39 años
16	De 40 a 44 años
17	De 45 a 49 años
18	De 50 a 54 años
19	De 55 a 59 años

Código Grupo etario	Descripción
20	De 60 a 64 años
21	De 65 a 69 años
22	De 70 a 74 años
23	De 75 a 79 años
24	De 80 a 84 años
25	De 85 a 89 años
26	De 90 a 94 años
27	De 95 a 99 años
28	De 100 años y más

Apéndice C *Tabla Codificación Nivel Educativo*

Código Nivel Educativo	Descripción
0	Ninguno
1	Preescolar
2	Básica primaria
3	Básica secundaria
4	Media académica o clásica
5	Media técnica
6	Normalista
7	Técnica profesional
8	Tecnológica
9	Profesional
10	Especialización
11	Maestría
12	Doctorado