Modelo predictivo para la mortalidad en urgencias por trauma: integración de variables clínicas y socioespaciales en Cali 2012-2013

Facultad de ingeniería, diseño y ciencias aplicadas

Trabajo de grado II

Trabajo de grado para optar al título de Magíster en Ciencia de Datos

---

Alumnos:
Nicolas Orozco Echeverri
Andrea Valencia Orozco

Tutores:
Alberto Federico García Marín, MD, MSc.
Santiago Ortíz Arias, PhD (c).

4 de diciembre de 2024

Índice

1. Introducción

El trauma, definido como cualquier lesión física resultante de una fuerza externa, representa un desafío significativo para la salud pública a nivel mundial. Esta patología representa un desafío significativo para la salud pública a nivel mundial, siendo una de las principales causas de mortalidad en adultos jóvenes menores de 45 años, con un estimado de 1.9 millones de muertes en todo el mundo [1]. En el contexto colombiano, el trauma también emerge como una de las principales causas de mortalidad, según datos del Departamento Administrativo Nacional de Estadística (DANE) para el año 2022 [2]. Específicamente, en la ciudad de Cali, para el año 2021 se reportó una tasa de mortalidad por lesiones de causa externa de 75.16 fallecimientos por 100000 habitantes, y específicamente por homicidios, se presentó una tasa de 50.12 muertes por 100000 habitantes [3]. Es crucial destacar que la génesis de eventos traumáticos y el pronóstico a largo plazo de los pacientes están influenciados por diversos determinantes sociales de la salud. Estos incluyen la situación socioeconómica, el nivel educativo, el entorno físico y social del vecindario, el empleo, las redes de apoyo social y el acceso a la atención sanitaria de forma oportuna [4].

La influencia de los determinantes sociales de la salud en la generación de eventos traumáticos, especialmente aquellos relacionados con la violencia, ha sido ampliamente documentada [4], lo que subraya la importancia de abordar estos factores en el contexto de la salud pública. En Colombia, un país marcado por desafíos significativos en términos de inequidad y afecciones de los determinantes sociales de la salud, las estadísticas del DANE para el año 2022 revelan cifras preocupantes. Con un índice GINI nacional de 0.556 y aproximadamente el 36.6 % de la población en situación de pobreza monetaria y el 13.8 % en pobreza extrema [5], estas cifras ilustran la magnitud del problema y destacan la urgencia de abordar estos desafíos en el contexto de la prevención y el direccionamiento adecuado para la atención del trauma en el país, con el fin de reducir la mortalidad de los pacientes [6].

Este proyecto tiene como objetivo proporcionar una comprensión integral de los factores que inciden en la mortalidad de los pacientes traumatizados en Cali, empleando variables clínicas, sociales y geoespaciales del lugar del trauma. No solo se busca desarrollar un modelo predictivo más preciso para evaluar los resultados de mortalidad, sino también entender las complejas dinámicas que influyen en la atención de emergencia y la salud pública en la región. Al emplear técnicas de análisis de datos y explorar la interacción entre variables clínicas, sociales y geoespaciales, este proyecto contribuye al campo de la epidemiología y la atención médica de emergencia. Al profundizar en la comprensión de los determinantes del trauma y la mortalidad, se espera que este estudio proporcione una base sólida para el diseño de intervenciones más efectivas y políticas de salud pública destinadas a mejorar el cuidado de los pacientes traumatizados, no solo en Cali, sino también en otras comunidades similares.

2. Contexto y Antecedentes

El trauma constituye un importante obstáculo para la salud pública a escala global, siendo una de las principales razones detrás de la mortalidad en adultos jóvenes menores de 45 años, con aproximadamente 1.9 millones de fallecimientos en todo el mundo, como se ha señalado en un estudio de referencia [1]. En el contexto colombiano, el trauma también se destaca como una causa relevante de mortalidad, como lo demuestran las estadísticas del DANE para el año 2022 [2].

Según el Instituto Nacional de Medicina Legal y Ciencias Forenses, en 2021 la tasa de mortalidad por lesiones externas en Bogotá fue de 31.97 por 100000 habitantes, con un total de 2480 fallecimientos, de los cuales cerca del 50 % (1134 casos) fueron por homicidios. En Me-

dellín, la tasa de mortalidad fue de 41.26 por 100000 habitantes, con un total de 1050 casos, de los cuales menos del 40 % (404 casos) fueron por homicidios. En la ciudad de Cali, la tasa de mortalidad fue el doble que en Bogotá, con 75.16 casos por 100000 habitantes, totalizando 1695 fallecimientos, de los cuales aproximadamente el 70 % (1135 casos) fueron por homicidios[3]. Considerando la epidemiología local de la ciudad de Cali y las estadísticas de desigualdad nacional y local, se enfrentan retos particulares debido a la alta tasa de eventos traumáticos y una infraestructura insuficiente e inefectiva para la atención de estos casos[7]-[9].

En terminós del ámbito clínico, en múltiples estudios se ha demostrado que el tiempo de la atención prehospitalaria es un factor crítico que impacta en la mortalidad de los pacientes, aumentando el riesgo de morir hasta en un 8 %−10 % por cada 10 minutos de demora en la atención [10]-[12]. Desde 1976, se considera la "hora dorada" del trauma, el cual identifica el período crítico de la atención de los pacientes, requiriendo una intervention rápida y eficaz para mejorar la supervivencia [6], [13]. Sin embargo, en las regiones en desarrollo como Cali, la logística y la calidad de la atención prehospitalaria para atender y transportar a los pacientes a centros de trauma es considerablemente deficiente.

Además de los factores clínicos y logísticos, los determinantes sociales y geoespaciales [4], [14], [15], también juegan un papel crucial en los resultados del trauma. La violencia urbana, la pobreza y la desigualdad en el acceso a servicios de salud son condiciones que pueden exacerbar la severidad y las consecuencias de las lesiones. Estos elementos sociodemográficos y geográficos complejos requieren ser analizados conjuntamente para entender completamente su impacto en la mortalidad por trauma.

Metodológicamente, el avance en las técnicas de aprendizaje automático ha abierto nuevas posibilidades para el análisis de grandes conjuntos de datos en salud, en especial en la atención de pacientes con trauma [16]-[18]. El aprendizaje supervisado, mediante algoritmos como regresiones logísticas, árboles de decisión, bosques aleatorios y redes neuronales, facilita la modelación y predicción de los resultados de salud con alta precisión. Estos modelos son capaces de integrar una amplia variedad de variables clínicas, temporales y socioambientales, proporcionando predicciones detalladas sobre la supervivencia de los pacientes y apoyando la toma de decisiones de tratamiento en tiempo real.

La combinación de datos de salud pública con métodos avanzados de análisis de datos representa una convergencia poderosa que puede transformar la atención del trauma. Investigaciones en este campo no solo mejoran nuestro entendimiento de los determinantes de la supervivencia en trauma sino que también promueven el desarrollo de políticas más efectivas y sistemas de atención médica más responsivos y equitativos.

3. Planteamiento del Problema y Justificación

En Cali, al igual que en muchas otras partes del mundo, el trauma se destaca como una de las principales causas de mortalidad, especialmente entre personas menores de 45 años. Esta preocupante estadística coloca al trauma no solo como una crisis de salud pública, sino también como una carga significativa para el desarrollo socioeconómico. En el contexto local de Cali, los eventos traumáticos derivados de múltiples causas —accidentes de tránsito, violencia urbana, y más— no solo terminan en pérdidas de vidas humanas, sino también en numerosos casos de discapacidad y complicaciones de salud prolongadas, lo que repercute en toda la estructura social y económica de la ciudad.

La realización de este proyecto es crucial debido a las profundas implicaciones que el trau-

ma tiene sobre la juventud y, por extensión, sobre la economía y la cultura local. El trauma es la principal causa de mortalidad en los adultos jóvenes en Colombia, lo que resulta en una pérdida significativa de años potencialmente productivos y saludables. Además, las consecuencias de la discapacidad y las complicaciones a largo plazo tras eventos traumáticos imponen una pesada carga financiera sobre el sistema de salud, reduciendo la capacidad económica de los individuos afectados y sus familias, y limitando el crecimiento económico general. La necesidad de investigar y desarrollar un modelo predictivo que pueda mejorar la comprensión y manejo del trauma es evidente, no solo para optimizar la respuesta médica inmediata y los resultados a largo plazo de los pacientes, sino también para ayudar en la formulación de políticas públicas y estrategias de prevención más efectivas. Este estudio apunta a proporcionar datos empíricos y recomendaciones que puedan guiar a los entes gubernamentales y sistemas de salud en la implementación de soluciones holísticas y sostenibles, promoviendo así un entorno más seguro y una sociedad más resiliente.

La integración de las características clínicas, sociales y geoespaciales del lugar del evento traumático puede afectar significativamente los resultados de supervivencia de los pacientes traumatizados cuando se analiza utilizando modelos de aprendizaje supervisado. Al incorporar una amplia gama de puntos de datos, incluida información clínica, factores sociales y datos geoespaciales, los algoritmos de aprendizaje supervisado pueden proporcionar un análisis integral que mejore la predicción y la comprensión de los resultados de los pacientes en la atención traumatológica. Este enfoque holístico permite una evaluación más matizada de los factores que influyen en las tasas de supervivencia, lo que lleva a predicciones más precisas e intervenciones personalizadas para pacientes traumatizados.

Los algoritmos de aprendizaje supervisado, como las regresiones logísticas, los árboles de decisión, los bosques aleatorios y las redes neuronales artificiales, pueden procesar y analizar este conjunto diverso de datos para identificar patrones y relaciones que contribuyen a los resultados de los pacientes. Estos algoritmos, con sus distintos niveles de complejidad y precisión, pueden aprovechar eficazmente la integración de características clínicas, sociales y geoespaciales para predecir los resultados de supervivencia en pacientes con trauma. Además, el uso de técnicas de aprendizaje por refuerzo, puede mejorar aún más la adaptabilidad y las capacidades predictivas de estos modelos al aprender de las interacciones con el entorno.

4. Objetivos

Objetivo General

Analizar el impacto en la sobrevida de los pacientes con trauma basado en las características clínicas, sociales y geoespaciales del sitio del evento, con el uso de modelos de aprendizaje supervisado, para optimizar la atención del trauma y las redes de direccionamiento desde el punto de vista de salud pública.

Objetivos Específicos

- Analizar las características sociodemográficas, geoespaciales y clínicas del ingreso de los pacientes con trauma atendidos en tres centros médicos de la ciudad.
- Determinar el impacto de las características sociales del sitio del trauma y la distancia al centro de atención en la sobrevida de los pacientes
- Predecir la mortalidad de los pacientes con trauma con el uso de modelos de aprendizaje supervisado.

5. Marco Teórico

Se presenta a continuación la construcción del marco teórico, el cual consiste en identificar información pertinente y específica para establecer una síntesis por medio de conceptos y definiciones sobre el trauma como un problema significativo de salud pública en Colombia, destacando su manejo y las implicaciones sociales y medicas asociadas. En consonancia con el objetivo de predecir la mortalidad de los pacientes con trauma, se detallarán algunos de los modelos de aprendizaje supervisado más utilizados en la práctica, ofreciendo una perspectiva integral de estas herramientas analíticas avanzadas.

5.1. El trauma en Colombia: Un problema de salud pública

El trauma por cualquier causa es una enfermedad de interés en salud pública y es considerada una epidemia global, de especial interés en países de medianos y bajos ingresos y con mayor impacto en zonas de guerra y regiones discriminadas por la violencia multifactorial [19]. Según los datos del libro *"The Economicos of crime"*, América Latina y la región Caribe son las regiones del mundo con mayores tasas de mortalidad por violencia, y además, se reporta a Colombia como uno de los países con mayores costos del sistema asociados a la violencia urbana como en relación a los conflictos sociales internos [20].

En Colombia, la institución encargada de la evaluación de las muertes por trauma de cualquier tipo es el Instituto de Medicina Legal y Ciencias Forenses, este anualmente publica una revista con el balance de los eventos ocurridos a lo largo del año, la cual se denomina *Forensis, datos para la vida*. Para el año 2014, se recibieron 25225 casos por muertes violentas, con una tasa de 52.92 casos por cada 100000 habitantes; de estos el homicidio fue la primera causa de muerte con un 50.07 % de los casos y se reporta un aproximado de 900878 años de vida perdidos para la economía del país [21]. Adicionalmente, para constatar que este problema continua e inclusive en los últimos años se ha evidenciado un incremento de la violencia en el país, en la última versión de la revista para el año 2022, se reportaron 30437 necropsias médico legales por muertes violentas, con una tasa de 59.34 casos por 100000 habitantes y los años de vida perdidos ascienden hasta los 1191720, siendo una cifra alarmante para el país en comparación con años pasados, denotando la importancia de abordar la violencia y el trauma desde una perspectiva holística en las diferentes regiones del país [22].

5.2. La atención prehospitalaria, un determinante en la supervivencia

La atención prehospitalaria ha sido un tema de gran interés en la investigación médica, con el fin de generar estrategias de mejor atención a los pacientes de forma más oportuna y especializada desde momentos muy tempranos de ocurridos los eventos de trauma. Lastimosamente, dada la heterogeneidad de las atenciones de países desarrollados y países en vía de desarrollo, los resultados de múltiples revisiones sistemáticas de la literatura y metaanálisis no han dado conclusiones certeras en relación al tema.

Una de las primeras revisiones sistemáticas del tema fue realizada por Hill et al. [23] en esta revisión de 36 estudios identificaron que no había diferencias significativas entre el estado de remisión y la mortalidad hospitalaria, pero con una heterogeneidad muy alta entre los estudios, con lo cual esta conclusión es discutible. Posteriormente, Henry et al. [6] realizaron una revisión sistemática de 14 estudios, en los cuales identificaron que los sistemas prehospitalarios de atención del trauma, específicamente en países en desarrollo reducen la mortalidad en los pacientes jóvenes. Adicionalmente, Harmsen et al. [24] realizaron una revisión sistemática de la literatura identificando en 20 artículos que el transporte rápido a un centro de trauma en pacientes que presentan un trauma craneoencefálico y los pacientes hemo-dinámicamente ines-

tables con lesiones penetrantes mejora los desenlaces clínicos. Además de estos estudios, existen varias revisiones de la literatura inconclusas dada la heterogeneidad de los estudios, pero se evidencia que la atención prehospitalaria y los sistemas de trauma son claves para mejorar los desenlaces clínicos de los pacientes [25]-[27].

5.3. El impacto de la distancia en la mortalidad

Se han realizado varios estudios para investigar la relación entre la distancia al centro de trauma y la mortalidad en pacientes con trauma. En un análisis retrospectivo, Díaz et al. [28] encontraron que un aumento en la distancia se asoció con un mayor riesgo de mortalidad, incluso después de ajustar por la gravedad de la enfermedad y la edad del paciente. Este hallazgo sugiere que la distancia juega un papel crucial en los resultados de los pacientes traumatizados, afectando significativamente su pronóstico.

Crandall et al. [29] examinaron específicamente a pacientes con heridas por arma de fuego en áreas urbanas y encontraron resultados similares. Descubrieron que aquellos que sufrieron heridas a una mayor distancia de un centro de trauma tuvieron tiempos de transporte más largos y una mayor mortalidad. Estos hallazgos resaltan la importancia del acceso rápido a la atención médica especializada, especialmente en casos de trauma grave donde cada minuto cuenta.

Jarman et al. [30] ampliaron esta investigación al analizar la relación entre la distancia al centro de trauma y la mortalidad en adultos con lesiones traumáticas en Maryland. Encontraron una asociación significativa entre un aumento en la distancia y un mayor riesgo de muerte, especialmente en pacientes trasladados a centros de trauma de nivel 3 o públicos. Estos resultados subrayan la necesidad de políticas y sistemas de atención médica que aseguren una distribución equitativa de los recursos y una accesibilidad adecuada a la atención de trauma.

Además, la revisión sistemática realizada por Chambers et al. [31] confirmaron la importancia de la accesibilidad a la atención médica de emergencia. Encontraron que los aumentos en la distancia o el tiempo de viaje a los centros de atención de urgencia y emergencia pueden aumentar el riesgo de mortalidad, especialmente en casos de infarto agudo de miocardio. Este estudio refuerza la idea de que la distancia a los centros de atención de trauma no solo impacta la mortalidad de los pacientes traumatizados, sino también de otros grupos de pacientes que requieren atención médica urgente. En conjunto, estos hallazgos resaltan la necesidad de políticas y sistemas de salud que garanticen un acceso rápido y equitativo a la atención médica especializada, especialmente en áreas donde la distancia puede ser un factor limitante en la supervivencia de los pacientes traumatizados.

5.4. Centralizar la referencia del trauma: La necesidad de regionalizar la atención

La centralización de recursos y experiencia en sistemas formales de atención traumatológica ha demostrado reducir la tasa de mortalidad por lesiones graves y mejorar los resultados funcionales a nivel internacional. Mullins y Mann [32] realizaron una revisión sistemática que evidenció la efectividad de los centros de trauma/sistemas de trauma en la sobrevida hospitalaria de los pacientes. Aunque la mayoría de los estudios revisados tenían limitaciones metodológicas, los resultados sugieren una reducción significativa en la mortalidad prevenible y una disminución del riesgo de muerte después de la implementación de estos sistemas.

Los hallazgos de los estudios basados en la evaluación por panel de expertos, comparaciones con registros nacionales de trauma y estudios poblacionales respaldan la eficacia de los centros/-

sistemas de trauma. Se observó una reducción del número de muertes prevenibles, así como una disminución del riesgo de muerte en un 15-20 % después de la implementación de estos sistemas. Estos resultados sugieren consistentemente que la centralización de la atención traumatológica puede mejorar los resultados en pacientes con trauma grave [33].

Sin embargo, la mayoría de estos estudios se han realizado en contextos geográficos y políticos diferentes, con sistemas de emergencia estructurados y centros de trauma definidos. Por lo tanto, es necesario realizar investigaciones en entornos como la ciudad de Cali, donde se pueda evaluar la aplicabilidad y los beneficios de los sistemas de trauma en contextos similares en Latinoamérica.

5.5. Aprendizaje Supervisado

El aprendizaje supervisado es una rama del aprendizaje automático en la cual una metodología seleccionada se entrena para predecir un objetivo basado en datos de entrada etiquetados. En este proceso, se proporciona al algoritmo un conjunto de datos de entrenamiento con etiquetas correspondientes. A partir de estos datos, el algoritmo aprende una regla que posteriormente aplica para predecir las etiquetas de nuevas observaciones. Los algoritmos de aprendizaje supervisado se dividen en dos categorías principales: algoritmos de regresión y algoritmos de clasificación. Los algoritmos de regresión buscan predecir y modelar variables cuantitativas basándose en las variables de entrada, mientras que los algoritmos de clasificación determinan una categoría de una variable cualitativa, en función de un conjunto de variables de entrada [34].

En el marco de este proyecto, se evaluarán los modelos de clasificación más relevantes. Estos modelos desempeñan un papel esencial en la predicción del riesgo de incidencia de enfermedades, la identificación de biomarcadores para monitorear la evolución de patologías o la respuesta a tratamientos, y la optimización de decisiones terapéuticas personalizadas, contribuyendo así a elevar la calidad de la atención médica, con lo cual se describen los siguientes modelos:

5.5.1. Regresión Logística

La regresión logística es uno de las metodologías más utilizadas para la clasificación. El modelo de regresión logística surge del deseo de modelar las probabilidades de las clases de salida a partir de una función que es lineal en $(x_1,\ldots,x_n)$, asegurando al mismo tiempo que las probabilidades de salida sumen uno y se mantengan entre cero y uno, como cabría esperar de las probabilidades. Si se entrena un modelo de regresión lineal en varios ejemplos donde $Y = 0$ o 1, se podría terminar prediciendo algunas probabilidades que son menores que cero o mayores que uno, lo que no tiene sentido. En su lugar, se utiliza un modelo de regresión logística (o modelo logit), es un modelo lineal generalizado, el cual, a partir de una función de enlace se modela la probabilidad de pertenencia o no a una categoría [34].

La Ecuación 1 representa el modelo de regresión logística. Al igual que en la regresión lineal, esta ecuación combina los valores de entrada x de manera lineal mediante el uso de pesos o coeficientes para predecir un valor de salida y. Sin embargo, a diferencia de la regresión lineal, el resultado obtenido es una probabilidad que, posteriormente, se convierte en un valor binario (0 o 1) para generar la predicción del modelo:

$$y = \frac{\exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_j x_j)}{1 + \exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_j x_j)} \tag{1}$$

Donde β_1, β_2, β_3, …, β_j son los coeficientes de regresión del modelo. La expresion del lado derecho de la ecuación se conoce con el nombre de función logística multivariada [34].

Para optimizar la efectividad de este modelo en escenarios con un alto número de variables, se pueden aplicar técnicas de regularización como la regresión logística LASSO, Ridge, o una combinación de ambas llamada Elastic-net. La regresión logística Ridge (regularización ℓ_2) penaliza la suma de los cuadrados de los coeficientes, limitando su tamaño pero sin anularlos completamente, lo cual es útil cuando muchas variables tienen importancia moderada. Por otro lado, la regresión logística LASSO (regularización ℓ_1) penaliza la suma de los valores absolutos de los coeficientes, lo que puede llevar a que algunos de ellos se reduzcan a cero, facilitando así la selección de variables al eliminar predictores menos significativos. Elastic-net combina las penalizaciones de ℓ_1 y ℓ_2, ofreciendo un balance entre la selección de variables y la reducción de coeficientes, lo que puede ser especialmente útil en situaciones donde las variables están altamente correlacionadas. Estas técnicas de regularización no solo previenen el sobreajuste, sino que también contribuyen a mejorar la interpretabilidad del modelo al simplificar la estructura de los coeficientes [34].

5.5.2. Análisis Discriminante Lineal y Cuadrático

El análisis discriminante lineal (LDA), también conocido como análisis discriminante normal (NDA) o análisis de función discriminante (DFA), sigue un marco de modelo generativo. Esto significa que la metodología LDA modela la distribución de datos para cada clase y utilizan el teorema de Bayes para clasificar nuevos puntos de datos. Bayes calcula probabilidades condicionales: la probabilidad de un evento dado que otro evento ha ocurrido. Los algoritmos de LDA hacen predicciones utilizando Bayes para calcular la probabilidad de si un conjunto de datos de entrada pertenecerá a una salida particular. El LDA funciona identificando una combinación lineal de características que separa o caracteriza dos o más clases de objetos o eventos. El LDA hace esto proyectando datos con dos o más dimensiones en una dimensión para que puedan clasificarse más fácilmente. Por lo tanto, la técnica es a veces referida como reducción de dimensionalidad. Esta versatilidad asegura que el LDA puede ser utilizado para problemas de clasificación de datos multiclase. Así, el LDA es a menudo aplicado para mejorar el funcionamiento de otros algoritmos de clasificación de aprendizaje como árboles de decisión, bosques aleatorios o máquinas de vectores de soporte [35].

Por otro lado, el análisis discriminante cuadrático (QDA) es una extensión del LDA. Esto hace que QDA sea más flexible que LDA, ya que no asume que todas las clases comparten la misma varianza general. En QDA, se modela la probabilidad condicional de cada clase como una distribución normal multivariante, con su propia media y matriz de covarianza. Esto permite que QDA capture relaciones más complejas dentro de los datos, especialmente cuando las diferencias entre grupos no son linealmente separables. Como tal, QDA puede ser más adecuado para conjuntos de datos donde las clases presentan variaciones significativas en sus características internas, pero a costa de requerir más datos para estimar de manera efectiva los parámetros y siendo potencialmente más susceptible a problemas de sobreajuste si el tamaño de la muestra es pequeño [35].

5.5.3. Máquinas de Vectores de Soporte

Las Máquinas de Vectores de Soporte (SVM, por sus siglas en inglés) son métodos de clasificación y regresión reconocidos por su capacidad para generar predicciones altamente precisas, evitando el sobreajuste en los datos de entrenamiento. Estas herramientas son especialmente valiosas para analizar conjuntos de datos con muchas variables predictoras y se utilizan en una amplia variedad de campos. Su mecanismo subyacente implica mapear los datos a un espacio de características de alta dimensión donde los conjuntos de datos pueden ser clasificados, aún

en situaciones donde no es posible una separación lineal. Se establece un hiperplano divisorio entre las categorías, y los datos se manipulan de tal manera que este plano se convierte en un hiperplano. Posteriormente, se utilizan las características de nuevos datos para predecir la categoría a la que pertenecen [35].

Para ilustrar visualmente los principios de las SVM, se presentan en las Figuras 1, 2 y 3 la capacidad de este método para clasificar datos complejos, las gráfican fueron tomadas de [35].

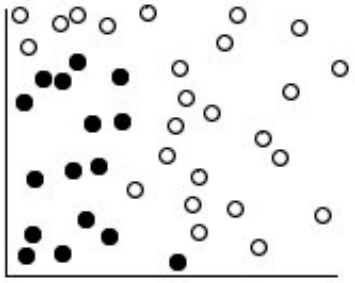


Figura 1: Conjunto de datos original.

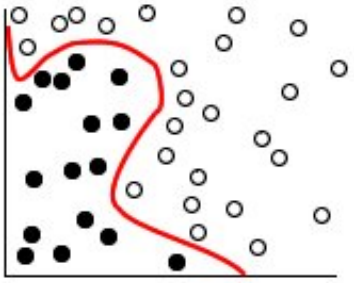


Figura 2: Datos con un separador añadido.

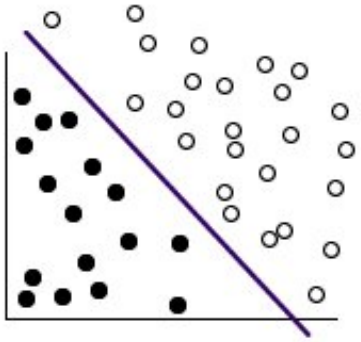


Figura 3: Datos transformados.

La eficacia de estas herramientas depende en gran medida de la elección adecuada de la función kernel. Entre las más comunes se encuentran las funciones Lineal, Polinómica, de Base Radial (RBF) y Sigmoide. Una función kernel lineal es adecuada cuando los datos pueden separarse linealmente de manera sencilla. Para datos más complejos, se recomienda explorar las otras funciones disponibles. Es fundamental experimentar con estas diferentes funciones para identificar la que mejor se adapte a cada caso, dado que cada una utiliza algoritmos y parámetros distintos [35].

5.5.4. Árboles de Decisión

El árbol de decisión es un algoritmo de aprendizaje supervisado no paramétrico ampliamente utilizado para la clasificación y la regresión. Este algoritmo presenta una estructura de árbol jerárquica, con un nodo raíz desde donde se origina, ramas que representan decisiones, y nodos hoja que corresponden a los resultados. Funciona mediante una estrategia de búsqueda codiciosa que identifica los puntos de división óptimos, un proceso que se repite hasta clasificar efectivamente los datos. Aunque un árbol sencillo puede resultar en nodos hoja con datos uniformes, un árbol demasiado complejo corre el riesgo de sobreajustarse a los datos de entrenamiento. Para evitar esto, se utiliza la poda del árbol, que simplifica el modelo conforme a la Navaja de Occam, y se emplea la validación cruzada para asegurar que el modelo generalice bien a nuevos datos [34].

Cómo elegir el mejor atributo en cada nodo

Si bien hay varias formas de seleccionar el mejor atributo en cada nodo, dos métodos, la ganancia de información y la impureza de Gini, actúan como criterio de división popular para los modelos de árboles de decisión. Ayudan a evaluar la calidad de cada condición de prueba y qué tan bien podrá clasificar las muestras en una clase [34].

Entropía y ganancia de información

La entropía es un concepto que se deriva de la teoría de la información, que mide la impureza

de los valores de la muestra. Se define con la siguiente fórmula [34]:

$$\text{Entropía}(S) = -\sum_{c \in C} p(c) \log_2 p(c)$$

donde:

- S representa el conjunto de datos.
- C representa las clases en el conjunto S.
- $p(c)$ representa la proporción de puntos de datos que pertenecen a la clase c respecto al número total de puntos de datos en el conjunto S.

Los valores de entropía pueden estar entre 0 y 1:

- Si todas las muestras en S pertenecen a una sola clase, la entropía será 0.
- Si las muestras están uniformemente distribuidas entre dos clases, la entropía alcanza su máximo de 1.

Para encontrar la mejor característica sobre la cual dividir y formar un árbol de decisión óptimo, se busca el atributo con la menor entropía. La *ganancia de información* se define como la diferencia en la entropía antes y después de dividir un conjunto según un atributo específico a y se calcula con la fórmula [34]:

$$\text{Ganancia de Información}(S, a) = \text{Entropía}(S) - \sum_{v \in \text{valores}(a)} \frac{|S_v|}{|S|} \text{Entropía}(S_v)$$

donde:

- $|S_v|/|S|$ representa la proporción de los valores en S_v al número de valores en el conjunto de datos S.
- Entropía(S_v) es la entropía del subconjunto S_v que resulta de dividir S por el valor v del atributo a.

El atributo con la mayor ganancia de información es considerado como el mejor para dividir los datos, ya que proporciona la clasificación más clara según la variable objetivo [34].

Índice de Gini

El índice de Gini cuantifica la pureza de un nodo, reflejando la probabilidad de que dos registros seleccionados al azar del mismo nodo pertenezcan a clases diferentes. Un índice de Gini alto indica una menor pureza; por ende, para mejorar la clasificación, se da preferencia a la variable que presente un valor más bajo de Gini ponderado. El índice de Gini se define como:

$$1 - \sum_{i=1}^{n} (P_i)^2 \tag{2}$$

Donde P_i es la probabilidad de que un ejemplo sea de la clase i.

Si se seleccionan dos elementos de una población al azar, entonces deben ser de la misma clase y la probabilidad de esto es 1 si la población es pura:

1. Funciona con la variable objetivo categórica “éxito” o “fracaso”.
2. Realiza solo divisiones binarias.
3. Cuanto mayor sea el valor de Gini, mayor será la homogeneidad.
4. CART (árbol de clasificación y regresión) utiliza el método Gini para crear divisiones binarias.

5.5.5. Algoritmos de Ensamble

Las técnicas de ensamble combinan modelos individuales para mejorar la estabilidad y el desempeño de la predicción del modelo. Se dice que este método permite una mejora en el rendimiento de la predicción combinando múltiples modelos de aprendizaje, usualmente denotados como clasificadores débiles en un solo modelo, donde cada una de las predicciones se combinan para obtener una única predicción [35].

Dentro de los enfoques de ensable, el **bagging y el boosting** son los más destacados. El bagging entrena distintos modelos en paralelo, cada uno con un subconjunto aleatorio de los datos. El boosting, en cambio, entrena modelos de forma secuencial, donde cada modelo subsiguiente se enfoca en corregir los errores de su predecesor, aprendiendo de manera iterativa y continuada. Al agrupar estos modelos, los ensamblajes tienden a ser más flexibles y estables ante variaciones en los datos, lo que reduce tanto el sesgo como la varianza [35].

5.5.6. Bosque Aleatorio

El bosque aleatorio (Random Forest, RF por sus siglas en inglés) es una versión mejorada de los árboles de decisión ensamblados mediante bagging. Para entender mejor el algoritmo de bosque aleatorio, es útil primero comprender el algoritmo de bagging. Suponiendo que se tiene un conjunto de datos de mil instancias, los pasos para el bagging son [35]:

1. Crear muchas (e.g., cien) submuestras aleatorias del conjunto de datos.
2. Entrenar un modelo CART en cada muestra.
3. Dado un nuevo conjunto de datos, calcular la predicción promedio de cada modelo y agregar la predicción de cada árbol para asignar la etiqueta final por voto mayoritario.

Los bosques aleatorios diversifican aún más este proceso: en lugar de permitir que cada árbol revise todas las variables y sus valores para encontrar el mejor punto de división, cada árbol en un bosque aleatorio solo considera una muestra aleatoria de características para seleccionar los puntos de división. Esto reduce la correlación entre los árboles y aumenta la generalización del modelo. El número de características a considerar en cada división (m) es un parámetro clave del algoritmo. Además, durante la construcción del bosque, es posible evaluar la importancia de cada característica observando el impacto de cada una en la reducción del error, como podría ser el índice Gini en problemas de clasificación. El promedio de estas disminuciones de error a lo largo de todos los árboles proporciona una medida de la importancia de las características [35].

5.5.7. Extreme Gradient Boosting

El concepto de aumento de gradiente (XGBoost) fue introducido inicialmente por Friedman [36] y posteriormente adaptado a los árboles de decisión por Chen y Guestrin [37], quienes desarrollaron una versión avanzada conocida como XGBoost, abreviatura de 'Extreme Gradient Boosting'. Este algoritmo está especialmente diseñado para optimizar velocidad y rendimiento, y es un enfoque de aprendizaje supervisado que utiliza el ensamblaje de árboles para mejorar la capacidad de generalización del modelo de aprendizaje automático.

XGBoost construye múltiples árboles de decisión secuencialmente, donde cada nuevo árbol se enfoca en corregir los errores cometidos por los árboles anteriores, mejorando así la precisión de las predicciones paso a paso. Es reconocido por su eficaz manejo de datos dispersos y por su habilidad para manejar valores faltantes de manera adecuada. Aunque XGBoost puede requerir

un tiempo de entrenamiento mayor debido a su proceso iterativo detallado, ofrece un rendimiento sobresaliente y resultados precisos en la predicción, en gran parte gracias a su eficiente optimización de uso en el ordenador [37].

5.5.8. Redes Neuronales

Las redes neuronales artificiales (ANN, por sus siglas en inglés) son aproximaciones no lineales a la forma en que funciona el cerebro. Se definen como sistemas de mapeos no lineales cuya estructura se basa en principios observados en los sistemas nerviosos humanos y animales; constan de un gran número de procesadores simples ligados por conexiones con "pesos". Las unidades de procesamiento se llaman neuronas; cada unidad recibe entradas de otros nodos y genera una salida simple escalar que depende de la información local disponible, guardada internamente o que llega a través de las conexiones con pesos.

Existen diferentes modelos ANN, pero una característica común que tienen la mayoría de estas es que tienen el siguiente proceso de operación bajo la conceptualización de un sistema (ver Figura 4) [38]:

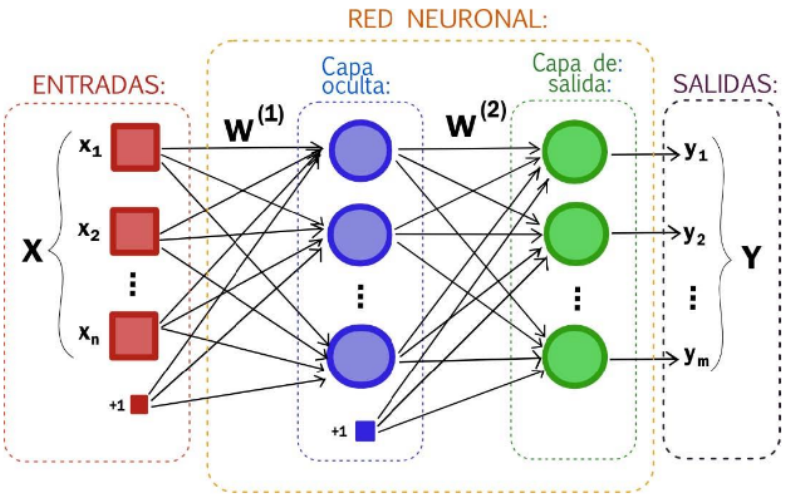


Figura 4: Representación esquemática de una red neuronal. Tomado de [38].

Las entradas constituyen el estímulo que recibe la neurona artificial del ambiente externo, y la salida representa la reacción frente a dicho estímulo. La capacidad de la neurona para ajustarse y aprender del entorno se manifiesta a través de la modificación de los pesos sinápticos. Estos pesos son denominados parámetros libres del modelo, permitiendo su ajuste para cumplir con una tarea específica.

Funciones de activación

La función de activación tiene el rol de determinar el estado de actividad de una neurona, es decir, dicta el valor de la salida de la neurona en función de su entrada, indicando si la neurona se encuentra inactiva (con una salida de 0 o -1) o activa (con una salida de 1). A continuación, se exponen varias de las funciones de activación más habituales [39]:

Función sigmoide

$$f(x) = \frac{1}{1 + e^{-x}} \tag{3}$$

Función tangente hiperbólica

$$f(x) = \frac{1 - e^{-2x}}{1 + e^{-2x}} \tag{4}$$

Función ReLU

$$f(x) = \text{máx}(0, x) \tag{5}$$

5.6. Medidas de desempeño del modelo

El desempeño del modelo se evaluó mediante la matriz de confusión, detallada en la Tabla 1. Esta herramienta permite calcular métricas críticas para valorar el rendimiento del algoritmo, incluyendo la exactitud, precisión, sensibilidad (recall), tasa de error y el área bajo la curva ROC (ver Tabla 2).

Tabla 1: Matriz de confusión.

Predicción	**Positivo**	**Negativo**
Positivo	Verdadero Positivo (VP)	Falso Negativo (FN)
Negativo	Falso Positivo (FP)	Verdadero Negativo (VN)

Las columnas de la matriz corresponden al número de predicciones para cada clase pronosticada, y las filas reflejan las instancias en su clase real, proporcionando una visión integral del comportamiento predictivo del modelo [40].

Tabla 2: Métricas de evaluación de modelos de clasificación.

Métrica	**Fórmula**
Exactitud	$\frac{\text{VP} + \text{VN}}{N}$
Error de Clasificación	$\frac{\text{FP} + \text{FN}}{N}$
Recall (Exhaustividad)	$\frac{\text{VP}}{\text{VP} + \text{FN}}$
Precisión	$\frac{\text{VP}}{\text{VP} + \text{FP}}$
Medida-F	$\frac{2 \cdot \text{Precisión} \cdot \text{Recall}}{\text{Precisión} + \text{Recall}}$
Especificidad	$\frac{\text{VN}}{\text{FP} + \text{VN}}$
Sensibilidad	$\frac{\text{VP}}{\text{VP} + \text{FN}}$
FNR (Tasa de Falsos Negativos)	$\frac{\text{FN}}{\text{VP} + \text{FN}}$
FPR (Tasa de Falsos Positivos)	$\frac{\text{FP}}{\text{VN} + \text{FP}}$

Kappa de Cohen:
Esta métrica calcula una puntuación que expresa el nivel de acuerdo entre dos anotadores en

un problema de clasificación.

Se define como:

$$k = \frac{P_o - P_e}{1 - P_e} \tag{18}$$

Donde P_o es la probabilidad empírica de acuerdo en la etiqueta asignada a cualquier muestra (la relación de acuerdo observada), y P_e es el acuerdo esperado cuando ambos anotadores asignan etiquetas al azar. P_e se estima utilizando un previo empírico por anotador sobre las etiquetas de clase [41].

En la Figura 5 se muestra una representación gráfica de una curva ROC.

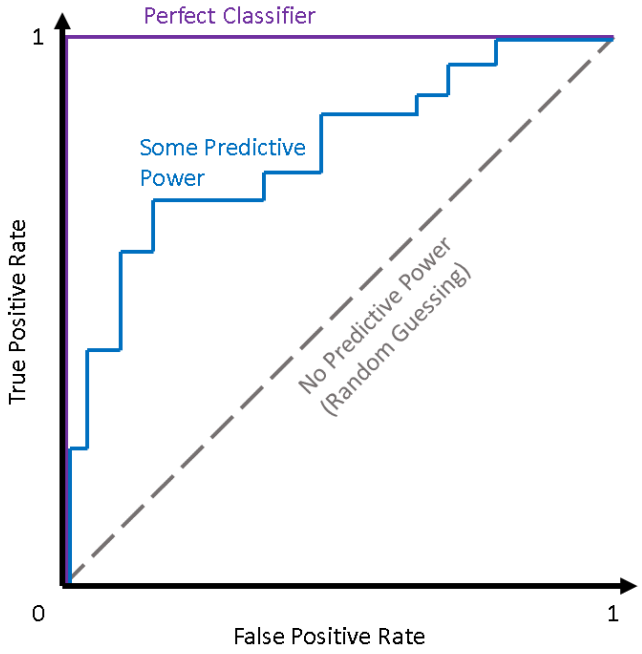


Figura 5: Curva ROC y representación área bajo la curva (AUC). Tomado de [42].

Tabla 3: Discriminación de la curva ROC.

Condición	Interpretación
Si ROC < 0.5	Sugiere No Discriminación
Si $0.5 \leq$ ROC < 0.7	Discriminación regular
Si $0.7 \leq$ ROC < 0.8	Discriminación aceptable
Si $0.8 \leq$ ROC < 0.9	Discriminación buena
Si ROC ≥ 0.9	Discriminación excelente

5.7. Optimización Bayesiana

El método de Optimización Bayesiana descansa sobre un enfoque probabilístico de búsqueda global, el cual utiliza el teorema de Bayes para ajustar continuamente la distribución de probabilidad de los hiperparámetros basándose en la evidencia acumulada. La Optimización Bayesiana procede a través de evaluaciones iterativas de una función de costo en múltiples configuraciones

de parámetros. La principal ventaja reside en su capacidad para identificar eficientemente el óptimo global —ya sea máximo o mínimo— de la función objetivo, utilizando el menor número de pruebas posibles. Este proceso se beneficia de un modelo estadístico sofisticado que anticipa los resultados en configuraciones que todavía no se han explorado. Este modelo estadístico dirige de manera inteligente la selección del siguiente conjunto de hiperparámetros a evaluar, optimizando el equilibrio entre la exploración de nuevas configuraciones y la explotación de aquellas que han demostrado ser efectivas anteriormente. Este enfoque asegura que el proceso de optimización no solo es eficiente sino también exhaustivo, maximizando las posibilidades de mejorar el desempeño del modelo al refinar los hiperparámetros basándose en un conocimiento progresivamente más profundo y preciso del espacio de búsqueda [43].

5.8. Problema del desbalance de clases

En el ámbito del aprendizaje automático, uno de los desafíos más comunes en las tareas de clasificación es el desbalance de clases. Este problema ocurre cuando una clase en un conjunto de datos tiene significativamente más instancias que otra. La clase con más instancias se denomina clase mayoritaria, mientras que la clase con menos instancias se conoce como clase minoritaria. Este desbalance se cuantifica mediante el ratio de desequilibrio (*Imbalance Ratio, IR*), que es la proporción entre el número de muestras en la clase mayoritaria y la minoritaria. Un IR elevado indica un mayor desequilibrio, lo que puede sesgar los resultados del modelo en favor de la clase mayoritaria, especialmente si se utiliza la precisión como métrica de evaluación [44].

Este problema es crítico en aplicaciones como el diagnóstico médico y la detección de fraudes, donde las clases de mayor interés suelen ser minoritarias. Un modelo que clasifica incorrectamente la clase minoritaria puede tener consecuencias graves, como fallar en la detección de una transacción fraudulenta [44].

Además del desbalance de clases, otros desafíos asociados incluyen el tamaño reducido de las muestras, el solapamiento de clases y el desbalance dentro de las clases, todos los cuales pueden degradar el rendimiento de los modelos supervisados [44].

Métodos para abordar el desbalance de clases

Para enfrentar el desbalance de clases, se han propuesto dos enfoques principales: el basado en la manipulación de datos y el basado en algoritmos [44].

1. **Enfoque basado en la manipulación de datos:**

   Este enfoque se centra en ajustar la distribución de las clases en el conjunto de datos antes de entrenar el clasificador, generalmente durante la fase de preprocesamiento. Es ampliamente utilizado en problemas de clasificación con clases desbalanceadas, ya que permite lograr un ratio relativamente balanceado. Una técnica particularmente popular dentro de este enfoque es la Técnica de Sobremuestreo de Minorías Sintéticas (*SMOTE*). SMOTE genera datos sintéticos para la clase minoritaria utilizando el algoritmo de k-vecinos más cercanos, equilibrando así el ratio de las clases. Esta técnica es preferible a simplemente duplicar las muestras existentes, ya que ayuda a prevenir el sobreajuste [44].

   Aunque SMOTE ofrece varias ventajas, como la prevención del sobreajuste, también presenta limitaciones, como el sobremuestreo de muestras ruidosas o aquellas ubicadas en los bordes de las clases. Para mitigar estos problemas, se han desarrollado variantes y estrategias adicionales, tales como *SMOTE-IPF*, *SMOTE-LOF*, *Borderline-SMOTE*, y

SVM-SMOTE. De estas, *Borderline-SMOTE* se destaca al enfocarse en sobremuestrear las muestras minoritarias que se encuentran en los límites entre clases, mejorando así la capacidad discriminativa del modelo [44].

En situaciones donde los conjuntos de datos contienen tanto variables categóricas como numéricas, se emplea una variante de SMOTE conocida como *SMOTE-NC* (Synthetic Minority Over-sampling Technique for Nominal and Continuous). Este método extiende SMOTE para trabajar con datos mixtos, permitiendo el sobremuestreo de la clase minoritaria mientras se preservan las relaciones entre las variables categóricas y numéricas. SMOTE-NC crea muestras sintéticas en el espacio de las variables numéricas, al tiempo que maneja las variables categóricas de manera adecuada, lo que lo convierte en una opción ideal cuando se enfrenta a conjuntos de datos mixtos [45].

2. **Enfoque basado en algoritmos:**

   Este enfoque ajusta los algoritmos de aprendizaje para que presten más atención a las clases minoritarias, disminuyendo así el impacto del desbalance sin necesidad de alterar los datos originales. Estos métodos pueden integrarse con cualquier algoritmo de aprendizaje, lo que los hace versátiles y menos complejos de configurar [44]..

   Recientemente, han surgido enfoques avanzados como las redes generativas adversarias (*GAN*) para la generación de datos. Las GAN permiten crear nuevas muestras de datos sintéticos que imitan la distribución de los datos reales, mejorando la representación de las clases minoritarias [44].

6. Estado del Arte

Para la construcción del estado del arte de este proyecto, se llevó a cabo una búsqueda de literatura académica, seleccionando documentos y estudios publicados de diversos año. Esta revisión reveló que, en el ámbito de la predicción de mortalidad u otro eventos de interés para pacientes que han sufrido traumas, se han implementado modelos estadísticos variados, incluyendo técnicas de estadística espacial y/o aprendizaje automático supervisado.

A continuación, se presenta el resumen de algunas investigaciones y artículos que se relacionan con el problema de investigación expuesto:

Yu et al. (2024) [46] en su artículo “Predicting the complexity and mortality of polytrauma patients with machine learning models”, exploraron la viabilidad de utilizar modelos de aprendizaje automático para predecir la complejidad y mortalidad en pacientes con politraumatismos. Este estudio se realizó en el Centro de Trauma Hospitalario de Pueblo de Pizhou, Jiangsu, China, analizando a 756 pacientes ingresados entre 2020 y 2022. Los investigadores emplearon varios algoritmos de aprendizaje automático, incluyendo SVM, Random Forest, ANN y XGBoost, encontrando que el modelo XGBoost superó significativamente a los demás con una precisión del 90 % y un AUC del 82 % para la predicción de mortalidad. Adicionalmente, se identificaron varios predictores clínicos significativos, como el hematoma intracraneal, que contribuyen a la complejidad y mortalidad en estos pacientes. Este descubrimiento no solo subraya la eficacia del XGBoost sobre los sistemas de evaluación de trauma existentes como el Injury Severity Score (ISS) y el Glasgow Coma Scale (GCS), sino que también propone una metodología avanzada para mejorar la gestión clínica de los pacientes politraumatizados. Los resultados de este estudio proporcionan una base sólida para futuras investigaciones sobre la aplicación de modelos de

aprendizaje automático en el ámbito del trauma y sugieren una posible mejora en las estrategias de atención médica a través de la personalización del tratamiento basado en análisis predictivos.

Hunter et al. [16] en su publicación "Science fiction or clinical reality: a review of the applications of artificial intelligence along the continuum of trauma care", realizaron una revisión comprensiva que explora el emergente campo de aplicación de la inteligencia artificial (IA) en la atención del trauma. Los autores se enfocaron en cómo la IA, con su avanzado aprendizaje automático, se está integrando en la predicción de lesiones, la gestión del volumen de emergencias, la evaluación y pronóstico de pacientes en situaciones de trauma. Se destacaron los usos actuales de algoritmos de IA para predecir la severidad de las lesiones en accidentes vehiculares y cómo estos pueden optimizar las respuestas de emergencia en el lugar del incidente. La revisión señaló que la IA puede asistir a los servicios de emergencia en el terreno y mejorar la eficiencia en el traslado de pacientes críticos, impactando directamente en la toma de decisiones clínicas y los resultados de los pacientes. Sin embargo, se enfatiza que, a pesar de su potencial, la aplicación de la IA en la cirugía de trauma aún está en sus etapas iniciales, y se necesita una investigación adicional, incluyendo ensayos clínicos prospectivos y la validación de algoritmos, para comprender completamente su capacidad para predecir trayectorias de pacientes y mejorar la atención de trauma.

Vaz et al. [47] en su publicación "Open data and injuries in urban areas—A spatial analytical framework of Toronto using machine learning and spatial regressions", investigaron la problemática de las lesiones, un tema de salud pública muchas veces no reconocido adecuadamente. El estudio utilizó datos abiertos para examinar las lesiones urbanas como una de las principales causas de años potenciales de vida perdidos antes de los 65 años en Canadá, sugiriendo la necesidad de políticas dirigidas geográficamente para su prevención y control. Utilizando datos del National Ambulatory Care Reporting System, se analizaron las lesiones en Toronto entre 2004 y 2010 a través de un marco analítico espacial que utiliza el aprendizaje automático y regresiones espaciales. Este enfoque permitió identificar patrones geográficos y temporales en la ocurrencia de lesiones y proporcionó una visión detallada de los factores de riesgo y su distribución por toda la ciudad. El análisis se complementó con datos del proyecto Wellbeing Toronto, y se realizaron autocorrelaciones espaciales a escalas global y local para evaluar más de 1.7 millones de lesiones. Los hallazgos revelaron comunidades específicas con perfiles de lesiones significativamente altos, y la comparación entre tres modelos de regresión espacial demostró que estas herramientas estadísticas son robustas y deberían usarse para informar políticas de prevención de lesiones. La investigación refuerza la importancia de un enfoque basado en datos para mejorar la respuesta a las lesiones urbanas, lo que podría llevar a una mejor asignación de recursos y a un aumento de la participación comunitaria y democrática en la toma de decisiones médicas relacionadas con la prevención de lesiones en áreas urbanas.

Nehemiah et al. [18] en su manuscrito "Machine Learning for Predicting Outcomes in Trauma", se concentraron en estudiar la incertidumbre que rodea a los modelos de predicción basados en aprendizaje automático (ML) para el triaje y la evaluación de pacientes con trauma. Con el objetivo de revisar y comparar cómo los diferentes modelos de ML pronostican resultados en este contexto, los autores realizaron una revisión sistemática de estudios que involucraban ML para predecir diversos resultados relacionados con el trauma, con la hipótesis de que los modelos que predicen resultados similares comparten características comunes. Analizaron datos de 2.433.180 pacientes de estudios observacionales que se enfocaban en medidas de resultado como la supervivencia/mortalidad, necesidad de hospitalización o shock/hemorragia, entre otros. A pesar de la variabilidad en el desempeño de los diferentes modelos de ML con un AUC que

oscilaban entre 0.035 a 0.927, se destacó que los estudios más robustos demostraron una capacidad significativa de predicción utilizando redes neuronales. Sin embargo, se reconoció que la implementación de ML en la práctica clínica requiere de ensayos observacionales prospectivos y estudios que confirmen su valor práctico, sugiriendo que es necesario establecer una evidencia sólida y de alta calidad sobre los impactos clínicos y económicos antes de que el ML pueda ser adoptado ampliamente.

Ghandour et al. [48] en el artículo "Analyzing Factors Associated with Fatal Road Crashes: A Machine Learning Approach", abordaron la tarea de comprender los factores de riesgo que contribuyen a las fatalidades en accidentes de tráfico. La investigación se fundamentó en la importancia de analizar estos factores para reducir el número de lesiones fatales en carreteras a nivel mundial. Utilizando un enfoque de aprendizaje automático, los autores desarrollaron un modelo híbrido que combinó optimización secuencial mínima y árboles de decisión para identificar los factores que conducen a lesiones mortales en accidentes de carretera. Este modelo se construyó, probó y validó usando datos del Lebanese Road Accidents Platform (LRAP), que incluyó 8482 casos de accidentes de tráfico con lesiones fatales como variable de resultado. De las nueve variables independientes examinadas en el estudio, siete resultaron ser significativamente asociadas con la ocurrencia de fatalidades. Los hallazgos del estudio tuvieron la intención de ser utilizados por legisladores y partes interesadas para desarrollar políticas más informadas relacionadas con la seguridad vial y para fomentar la creación de programas de seguridad que pudieran mitigar los factores de riesgo de accidentes fatales en carreteras. El estudio destacó el valor de los análisis de datos basados en aprendizaje automático para ofrecer una mejor comprensión y enfoque en la prevención de choques mortales.

Tsiklidis et al. [48] en su artículo "Using the National Trauma Data Bank (NTDB) and machine learning to predict trauma patient mortality at admission", presentaron un estudio en el cual se entrenó un clasificador de boosting de 400 estimadores para predecir probabilidades de supervivencia de pacientes de trauma. NTDB proporcionó datos de 799.323 pacientes, detallando su sobrevida o muerte y registrando 32 características. Enfocándose en 8 de esas características: presión arterial sistólica, frecuencia cardiaca, tasa respiratoria, temperatura, saturación de oxígeno, género, edad y la escala de coma de Glasgow. El estudio usó un conjunto de entrenamiento equilibrado para predecir la probabilidad de supervivencia al momento del ingreso hospitalario, logrando distinguir entre pacientes fallecidos y supervivientes en un 92.4 % de los casos. A través del análisis de la importancia de las variables y las curvas de dependencia parcial, se encontró que la escala de coma de Glasgow y la presión arterial sistólica fueran los predictores más significativos, mientras que otros factores tuvieron influencias más sutiles. Los valores de Shapley, que miden la contribución relativa de cada característica, se calcularon para varios pacientes, proporcionando una forma cuantitativa de identificar señales de advertencia específicas del paciente. Este estudio subrayó el potencial del aprendizaje automático aplicado a grandes conjuntos de datos para mejorar la comprensión de los riesgos a nivel del paciente y facilitar la toma de decisiones en el cuidado de emergencias de trauma.

En el estudio "Hospital mortality prediction in traumatic injuries patients: comparing different SMOTE-based machine learning algorithms", Hassanzadeh et al. [49] desarrollaron una investigación enfocada en predecir la mortalidad hospitalaria en pacientes con lesiones traumáticas mediante el uso de técnicas avanzadas de aprendizaje automático. Utilizando un conjunto de datos desequilibrado de 126 pacientes de trauma ingresados en la UCI en Hamadan, Irán, de marzo de 2020 a marzo de 2021, el estudio aplicó técnicas de SMOTE para equilibrar los datos

antes de procesarlos con una variedad de algoritmos, incluyendo Árboles de Decisión, Random Forest, Naive Bayes, NNA, SVM y XGBoost. La evaluación de estos modelos se llevó a cabo a través de un compendio de métricas incluyendo sensibilidad, especificidad, valores predictivos positivo y negativo, precisión, área bajo la curva ROC y el promedio geométrico de los puntajes F1. Los resultados mostraron una mejora significativa en la predicción tras el balanceo de datos, con Random Forest y las Redes Neuronales basadas en SMOTE-NC destacándose por su alto rendimiento en todos los criterios evaluados, lo que indicó el potencial de estas técnicas para asistir en la toma de decisiones clínicas en la UCI, mejorando potencialmente las políticas de atención de trauma.

Dirago et al. [50] en su publicación "Geospatial Analysis of Social Vulnerability, Race, and Firearm Violence in Chicago", examinaron la asociación entre la vulnerabilidad social y la violencia por armas de fuego en Chicago, enfocándose en cómo las inequidades estructurales y raciales, junto con disparidades socioeconómicas, se manifestaron en el índice de vulnerabilidad social (SVI) y su correlación con la incidencia de la violencia por armas de fuego (UFV) a través del tiempo. El equipo de investigación analizó datos de asaltos con armas de fuego de 2001 a 2019, que fueron geocodificados utilizando ArcGIS y comparados con los SVI de 2018 proporcionados por el CDC. Aplicaron regresión de Poisson con errores robustos y el índice Local Moran's I para evaluar la tasa de incidentes de UFV por cápita y la autocorrelación espacial, respectivamente. Los hallazgos revelaron que los tractos censales con SVI bajos experimentaron tasas significativamente más altas de violencia por armas de fuego comparados con aquellos con SVI más altos, destacando que los tractos censales con muy bajo SVI tenían un riesgo 1.7 veces mayor de incidentes de UFV. Los resultados de este estudio histórico reafirmaron la relación significativa entre la vulnerabilidad social y la violencia por armas de fuego, poniendo de relieve la necesidad de políticas enfocadas en la vulnerabilidad social para combatir efectivamente la violencia interpersonal y mejorar la salud pública en ciudades con problemas similares a los de Chicago.

Jay [51] en su artículo "Alcohol outlets and firearm violence: a place-based case-control study using satellite imagery and machine learning", investigó la relación entre los puntos de venta de alcohol y la violencia por armas de fuego en Filadelfia, Pensilvania. El autor desarrolló un método innovador que utiliza imágenes satelitales de alta resolución y aprendizaje automático para emparejar visualmente lugares con características similares del entorno construido. El estudio se centró en tiendas de cerveza y bares/restaurantes utilizando un marco de estudio de casos y controles para comparar manzanas de la ciudad con incidentes de disparos en 2017-2018 frente a manzanas sin incidentes, basando el análisis en una red neuronal convolucional y un modelo de incrustación estocástica distribuida. La regresión logística estimó las probabilidades relativas (OR) de un tiroteo en la misma manzana que un punto de venta de alcohol y dentro de distancias de una o dos manzanas, considerando factores adicionales como el uso del suelo, la composición demográfica y la actividad de drogas ilegales. Los resultados indicaron que los lugares con tiendas de cerveza dentro de una manzana tenían un OR=1.5 de tener un tiroteo, y los lugares con bares/restaurantes en la misma manzana un OR=1.6, lo que muestra una mayor probabilidad de violencia por armas de fuego en comparación con lugares sin estos establecimientos. Estos hallazgos son consistentes con estudios anteriores y proporcionan una evidencia más sólida del efecto causal de los puntos de venta de alcohol en la violencia por armas de fuego cercana. Este método de emparejamiento basado en la similitud visual ofrece un enfoque innovador que podría mejorar los estudios observacionales que abordan los riesgos basados en el lugar.

Bedard et al. [52] en su artículo "A scoping review of worldwide studies evaluating the effects of prehospital time on trauma outcomes", realizaron un examen detallado de cómo el tiempo de atención prehospitalaria influyó en los resultados de pacientes con trauma. Se llevó a cabo una revisión de alcance de la literatura publicada en MEDLINE desde 2009 hasta 2020, donde se identificaron y revisaron 808 artículos, de los cuales 96 cumplieron con los criterios de inclusión. Estos estudios, predominantemente de alta calidad, a menudo se basaron en datos de registros de trauma y resaltaron la falta de investigaciones en países de ingresos bajos y medios, con una notable ausencia de estudios centrados en poblaciones africanas. La mortalidad se utilizó como medida de resultado en el 93 % de los artículos, comúnmente definida como "mortalidad hospitalaria", en lugar de en un marco temporal específico post-trauma. El tiempo prehospitalario fue evaluado más frecuentemente desde el momento del despacho de los servicios de emergencias médicas hasta la llegada al centro de trauma terciario. Pocos estudios analizaron resultados fisiológicos como la insuficiencia de múltiples órganos. Los investigadores concluyeron que la literatura revisada representa desproporcionadamente a entornos de altos ingresos y se enfoca principalmente en la mortalidad en momentos específicos del intervalo prehospitalario. Loa autores sugieren que los registros de trauma están infravalorados como herramienta para profundizar en el impacto del tiempo prehospitalario sobre la morbimortalidad. Además, recomiendan investigaciones futuras que expandan el enfoque hacia la mejora de la reanimación y el transporte prehospitalarios, especialmente en África y en países de ingresos bajos y medios.

Rojas et al. [53] en su artículo "Optimization Model for the Location of Prehospital Care Ambulances in the city of Cali Colombia"presentaron un modelo de localización destinado a mejorar la eficiencia de los servicios de ambulancias en Cali, Colombia. Este trabajo fue motivado por la preocupante estadística de que en 2015, la ciudad experimentó más de 309 muertes en accidentes de tráfico, con un 70 % involucrando motocicletas. Los autores desarrollaron un modelo basado en el problema del Máximo Cubrimiento Esperado (MEXCLP), con el objetivo de maximizar la demanda satisfecha ajustada por la disponibilidad del servicio, calculada a través del promedio global de ocupación de ambulancias y su distribución histórica de demanda. Este enfoque buscó optimizar la cobertura geográfica y temporal de las ambulancias utilizando datos históricos, empleando lenguajes de programación matemática y la plataforma NEOS Server for Optimization para la resolución del modelo. El estudio fue una contribución significativa para la posible mejora en la prestación de servicios de emergencia en la ciudad.

En resumen, recientes investigaciones han explorado la aplicación de técnicas analíticas y de inteligencia artificial para la evaluación de riesgos y la gestión de emergencias, demostrando su eficacia en diferentes áreas. Yu et al. (2024) y Hassanzadeh et al. (2023) emplearon modelos de aprendizaje automático, como XGBoost y Random Forest, para predecir la severidad de lesiones y la mortalidad en pacientes con trauma, mostrando altos niveles de precisión en contextos clínicos específicos. Por otro lado, estudios como los de Vaz et al. (2021) y Dirago et al. (2024) utilizaron análisis geoespaciales para evaluar el impacto urbano de lesiones y la violencia, identificando patrones de riesgo en entornos urbanos. Además, trabajos como los de Ghandour et al. (2020) y Rojas et al. (2017) aplicaron modelos híbridos y técnicas avanzadas para identificar riesgos en accidentes de tráfico y mejorar los servicios prehospitalarios, mientras que Hunter et al. (2023) y Nehemiah et al. (2017) exploraron el uso de la inteligencia artificial para mejorar la capacidad predictiva en la atención clínica de traumas.

A pesar de estos avances, cada estudio tiende a enfocarse en un solo tipo de datos o en un ámbito específico, lo que limita una comprensión más integral de los factores que influyen en

la sobrevida de los pacientes con trauma. Nuestro estudio aborda estas limitaciones al integrar datos clínicos, sociales y geoespaciales para evaluar de manera más completa la sobrevida de los pacientes. Este enfoque permite considerar múltiples factores de manera conjunta, ofreciendo una visión más amplia y detallada de los determinantes que afectan los resultados de salud. Al combinar diferentes tipos de datos, nuestro trabajo busca no solo mejorar la comprensión de los factores que afectan la mortalidad, sino también optimizar la atención del trauma y la planificación de recursos desde una perspectiva de salud pública. Así, nuestro estudio complementa la literatura existente al subrayar la importancia de una aproximación multidimensional en la gestión de emergencias médicas.

7. Metodología

En esta sección se describen los procedimientos implementados para abordar el problema planteado. Este enfoque estructurado abarca todas las fases clave del proyecto, desde la comprensión del problema y la preparación de los datos, hasta el modelado y la evaluación de los resultados obtenidos. El objetivo final es la elaboración y somentimiento de un artículo científico, en el que se expondrán los hallazgos, las implicaciones de los modelos generados y su contribución al campo de estudio.

7.1. Datos de estudio

Se llevó a cabo un estudio analítico de una base de datos perteneciente a una cohorte observacional, diseñada específicamente para valorar la eficacia del TRISS como herramienta pronóstica de la mortalidad. Dicho estudio incorporó a pacientes que sufrieron traumas de gravedad moderada y severa ($ISS > 8$), y que fueron atendidos en cuatro centros hospitalarios de Cali en el periodo comprendido entre diciembre de 2012 y junio de 2013.

Criterios de inclusión:

- Ingreso al hospital para la atención de lesiones traumáticas.
- Edad mayor de 17 años.
- Transferencia directa del sitio de trauma o, en caso de remisión, no haber recibido intervenciones previas.

Criterios de exclusión:

- Intervalo entre el trauma y el ingreso superior a 6 horas.
- Intervención quirúrgica o transfusión en la institución hospitalaria que remite.
- Trauma menor (ISS menor a 8).
- Quemaduras.
- Lesiones traumáticas originadas por un proceso no traumático (ejemplo: caída con lesiones menores, originada en un accidente cerebrovascular, que puede causar daño neurológico mayor).
- Embarazo detectado en la evaluación clínica.
- Individuos declarados muertos a la admisión al servicio de urgencias.

Esta base de datos pertenece al Dr. Alberto Federico García Marín, quien actuó como director del estudio observacional y facilitó el acceso a la misma para el uso de las variables clínicas de los pacientes de la cohorte en el desarrollo de este proyecto de grado.

7.2. Geografía del sitio de estudio

El municipio de Santiago de Cali se encuentra localizado en el sur occidente de Colombia. El área urbana de la ciudad se divide en 22 comunas y una zona de expansión rural. El municipio tiene una población aproximada de 2.0 millones de habitantes, 53.2 % mujeres y 46.8 % hombres, de los cuales la mayoría de los habitantes son adultos jovénes entre los 20 y 35 años de edad, abarcando aproximadamente el 30 % de la población de la ciudad.

Previamente en el momento de la recolección de los datos y hasta la actualidad en la ciudad de Cali no existe un sistema de referenciación de hospitales para la atención de los pacientes traumatizados, pues en la ciudad a pesar de la frecuencia de estos eventos, no hay un sistema de la atención del trauma ni centros de atención certificados para la atención del trauma. Por lo cual, esta cohorte solo se limitó a la recolección de pacientes traumatizados en cuatro instituciones de la ciudad que contaran con el personal experimentado dedicado al cuidado integral de las urgencias traumáticas las 24 horas del día. Estos centros son los siguientes:

- Hospital Universitario del Valle (HUV)
- Fundación Valle del Lili
- Clínica Nuestra
- Hospital San Juan de Dios

7.3. Descripción de las variables

7.3.1. Variables del estudio inicial

En la base de datos original se registraron variables demográficas, el mecanismo del trauma, las constantes fisiológicas al ingreso y las características de las lesiones requeridas para el cálculo del TRISS (Tabla 4). Se registró el estado del paciente al egreso, vivo o muerto. Cuando existió duda acerca del origen traumático del fallecimiento, se desconoció completamente la descripción anatómica de las lesiones, o se ignoró el desenlace, se constató esta información en el Instituto Nacional de Medicina Legal y Ciencias Forenses.

Tabla 4: **Variables iniciales de la base de datos**

Variable	Tipo de Variable	Valores posibles	Dominio*
Sexo	Categórica Nominal	0. Masculino 1. Femenino	1
Edad	Cuantitativa Discreta	{>= 18}	1
Hospital	Categórica Nominal	1. Clínica Nuestra 2. Hospital Universitario del Valle (HUV) 3. Fundación Valle del Lili 4. Hospital San Juan de Dios	2
Sitio del Trauma	Categórica Nominal	Dirección del evento	2

Continuación Tabla: Variables iniciales de la base de datos

Variable	Tipo de Variable	Valores posibles	Dominio*
Mecanismo de Trauma	Categórica Nominal	0. Herida por arma de fuego 1. Herida arma cortopunzante 2. Trauma contundente 3. Lesiones de tránsito 4. Caídas 5. Golpes 6. Sin dato	1
Trauma penetrante	Categórica Nominal	0. No 1. Si	1
Tipo de Aseguramiento	Categórica Nominal	0. Sin aseguramiento 1. Régimen contributivo 2. Régimen subsidiado 3. Régimen especial 4. Otro	3
Tipo de Traslado	Categórica Nominal	0. Vehículo particular 1. Fuerza pública 2. Ambulancia básica 3. Ambulancia medicalizada 4. Otro 5. Sin dato	2
Escala de Coma de Glasgow	Cuantitativa discreta	$\{3, 4, 5, ..., 15\}$	1
Presión Arterial Sistólica al ingreso	Cuantitativa Discreta	$(0, \infty)$	1
Frecuencia Respiratoria al ingreso	Cuantitativa Discreta	$(0, \infty)$	1
Clasificación de lesiones anatómicas en cabeza	Cuantitativa Discreta	$\{0 - 6\}$	1
Clasificación de lesiones anatómicas en cara	Cuantitativa Discreta	$\{0 - 6\}$	1

Continuación Tabla: Variables iniciales de la base de datos

Variable	Tipo de Variable	Valores posibles	Dominio*
Clasificación de lesiones anatómicas en tórax	Cuantitativa Discreta	$\{0-6\}$	1
Clasificación de lesiones anatómicas en abdomen/pelvis	Cuantitativa Discreta	$\{0-6\}$	1
Clasificación de lesiones anatómicas en extremidades	Cuantitativa Discreta	$\{0-6\}$	1
Clasificación de lesiones anatómicas en externo	Cuantitativa Discreta	$\{0-6\}$	1
Clasificación de lesiones anatómicas en cabeza	Cuantitativa Discreta	$\{0-6\}$	1
Injury Severity Score (ISS)	Cuantitativa Discreta	$\{0, 1, 2, ..., 75\}$	1
Revised Trauma Score Scale (RTSS)	Cuantitativa Continua	$(0, \infty)$	4
Probabilidad de sobrevida	Cuantitativa Continua	$[0, 1]$	4
Probabilidad de muerte	Cuantitativa Continua	$[0, 1]$	4
Estado al Egreso	Categórica Nominal	0. Vivo 1. Muerto	5
Días de Hospitalización	Categórica Discreta	$(0, \infty)$	5

**Dominio: 1. Características del paciente; 2. Determinante sociodemográfico; 3. Determinante social; 4. Variables calculadas a partir de las iniciales; 5. Desenlaces clínicos.*

7.3.2. Variables de georeferenciación

A partir de las variables de la base de datos, en especial de la variable "Sitio del traumaçon el uso del API de OpenRouteService [54], se obtuvieron variables importantes para tener la capacidad de georeferenciar los eventos de trauma en la ciudad y calcular la ruta de las distancias más cortas del lugar de los eventos al centro de atención y a los otros centros, con el fin de identificar el centro más cercano. Es importante recalcar que, aquellos registros que tuvieron algún tipo de error en la geolocalización inicial con el API, fueron recuperados de forma manual con el uso de Google Maps para identificar la longitud y la latitud del evento y poder obtener las otras variables. La tabla de variables generada por este proceso se presenta en la Tabla 5:

Tabla 5: **Variables de georeferenciación**

Variable	Tipo de Variable	Valores posibles	Dominio*
Longitud	Cuantitativa Continua	$(-\infty, \infty)$	2
Latitud	Cuantitativa Continua	$(-\infty, \infty)$	2
Longitud Centro de atención	Cuantitativa Continua	$(-\infty, \infty)$	2
Latitud Centro de atención	Cuantitativa Continua	$(-\infty, \infty)$	2
Comuna	Cuantitativa Discreta	Comuna del evento	2
Distancia ruta más corta del evento al centro de atención	Cuantitativa Continua	$(0, \infty)$	2
Distancia ruta más corta del evento al HUV	Cuantitativa Continua	$(0, \infty)$	2
Distancia ruta más corta del evento a Clínica Nuestra	Cuantitativa Continua	$(0, \infty)$	2
Distancia ruta más corta del evento a FVL	Cuantitativa Continua	$(0, \infty)$	2
Distancia de retraso	Cuantitativa Continua	$(0, \infty)$	2
Atendido en el más cercano	Categórica Nominal	0. Si 1. No	2

**Dominio: 1. Características del paciente; 2. Determinante sociodemográfico; 3. Determinante social; 4. Variables calculadas a partir de las iniciales; 5. Desenlaces clínicos.*

7.3.3. Variables sociodemográficas

Posterior a tener la geolocalización de los eventos de trauma en la ciudad, se consultaron las bases de datos del DANE conjunto con solicitudes formales para obtener los archivos con las variables sociodemográficas de la ciudad de Cali para el año 2018 en el cual se realizó el Censo Nacional de Población y Vivienda (CNPV2018), con el fin de obtener las variables sociales, culturales y económicas en relación al sitio del evento del trauma. Esto fue realizado utilizando herramientas de mapeo avanzadas, incluyendo el software QGIS [55], así como varias librerías de R, específicamente *fs* [56], *sf* [57], *terra* [58], *glue* [59] y *distanceto* [60]; las cuales permitieron poder recuperar y organizar dichas variables referenciando el sitio del evento a la manzana perteneciente.

A continuación se muestran las variables obtenidas:

Tabla 6: **Variables Sociodemográficas**

Variable	Tipo de Variable	Valores posibles	Dominio*
Escala de Vulnerabilidad Social	Categórica Nominal	0. Baja 1. Media-Baja 2. Media 3. Media-Alta 4. Alta	3
Indice de pobreza Multidimensional	Cuantitativa Continua	$(0, 100)\,\%$	3
Riesgo de Embarazo Adolescente	Categórica Nominal	0. Baja 1. Media-Baja 2. Media 3. Media-Alta 4. Alta	3
Elegibilidad Jóvenes para Empleo	Categórica Nominal	0. Baja 1. Media-Baja 2. Media 3. Media-Alta 4. Alta	3
Viviendas Totales en la manzana	Cuantitativa Continua	$(0, \infty)$	3
Hogares Totales en la manzana	Cuantitativa Continua	$(0, \infty)$	3
Habitantes	Cuantitativa Continua	$(0, \infty)$	3
Déficit Habitacional (Cuantitativo)	Cuantitativa Continua	$(0, \infty)$	3

Continuación Tabla: Variables de georeferenciación

Variable	Tipo de Variable	Valores posibles	Dominio*
Déficit Habitacional (Cualitativo)	Cuantitativa Continua	$(0, \infty)$	3
Porcentaje de déficit de vivienda cuantitativo	Cuantitativa Continua	$(0, 100)$	3
Porcentaje de déficit de vivienda cualitativo	Cuantitativa Continua	$(0, 100)$	3
Porcentaje de déficit Medido	Cuantitativa Continua	$(0, 100)$	3

**Dominio: 1. Características del paciente; 2. Determinante sociodemográfico; 3. Determinante social; 4. Variables calculadas a partir de las iniciales; 5. Desenlaces clínicos.*

7.4. Proceso de preparación, exploración y modelación

En este proyecto, la limpieza y preparación de datos fueron fundamentales para el análisis. Se eliminaron entradas duplicadas y se gestionaron datos incompletos mediante imputación o eliminación, asegurando un conjunto de datos representativo. Se corrigieron errores, valores atípicos e inconsistencias, y se estandarizaron los formatos de datos. Además, se resolvieron discrepancias para mantener la coherencia y exactitud. La manipulación de datos implicó transformar y reorganizar los datos en bruto en un formato estructurado, integrando datos de múltiples fuentes, convirtiendo tipos de datos, seleccionando variables relevantes y realizando ingeniería de características. Estas acciones prepararon los datos para análisis posteriores, garantizando resultados robustos y confiables.

Respecto al análisis exploratorio de datos (EDA), este permitió una comprensión profunda del conjunto de datos mediante el uso de técnicas de visualización y estadísticas descriptivas. Estas herramientas ayudaron a identificar patrones, anomalías y posibles hipótesis, orientando así la selección de técnicas estadísticas o de modelado y asegurando resultados sólidos y fiables.

El conjunto de datos se dividió en dos partes: el conjunto de entrenamiento, que representó el 70 % del total, correspondiente a 387 instancias; y el conjunto de validación, que abarcó el 30 % restante, con 166 instancias. La base de datos está categorizada en dos etiquetas relacionadas con el desenlace: *muerto* y *vivo*, donde *muerto* se codificó como 1 y *vivo* como 0.

Se evaluaron distintos modelos de clasificación, cada uno con sus propias particularidades, entre los cuales se incluyeron:

- **Regresión logística**
  - Regresión logística
  - Con regularización LASSO
  - Con regularización Ridge
  - Con regularización Elastic-net

- **Máquina de vectores de soporte**
  - Kernel sigmoide
  - Kernel lineal
  - Kernel RBF
  - Kernel polinomial
- **Métodos de ensamblaje**
  - Random Forest
  - XGBoost
- **Redes neuronales**
  - Red neuronal artificial feedforward

7.4.1. Creación del Pipeline

Para la preparación y modelado de los datos, se desarrolló un pipeline estructurado en varias etapas clave. Este pipeline fue diseñado para asegurar una limpieza, transformación y selección óptimas de los datos durante el uso de los modelos de clasificación.

Etapas del Pipeline

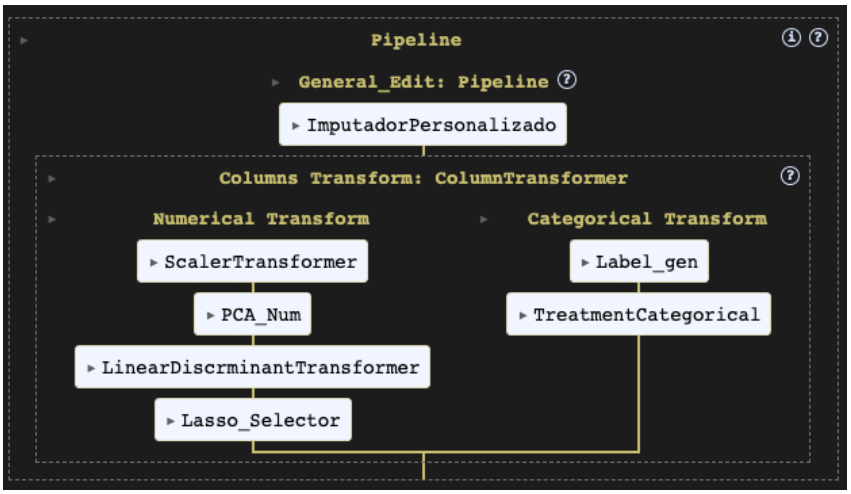


Figura 6: Esquema del pipeline de procesamiento de datos

1. **Imputación personalizada**: Se utilizaron técnicas de imputación para completar los valores faltantes en el dataset. Se diseñó un imputador específico para adaptarse a las particularidades de nuestros datos. Dado que los valores faltantes solo se presentaban en variables numéricas, se optó por la imputación simple con la mediana o el método de vecinos más cercanos (KNN), seleccionados mediante optimización bayesiana para ajustar los mejores hiperparámetros.
2. **Transformaciones numéricas**:
   - **Escalado**: Se aplicó un escalador para normalizar las características numéricas y asegurar que todas operen en la misma escala.

- **Análisis de componentes principales (PCA)**: Se utilizó PCA para reducir la dimensionalidad del conjunto de datos mientras se preserva la mayor cantidad de variabilidad posible. El número de componentes principales fue determinado también por optimización bayesiana.
- **LDA**: Se aplicó LDA para identificar combinaciones lineales de características que mejor diferencian las clases (vivo/muerto). Esta información se añadió como una nueva variable al conjunto de datos.
- **Selección de características con LASSO**: Se utilizó LASSO para identificar las características más relevantes. Este método aplica una penalización que reduce algunos coeficientes a cero. El parámetro de regularización también se ajustó mediante optimización bayesiana.

3. **Transformaciones categóricas**:

   - **Codificación de etiquetas**: Para las variables categóricas nominales se utilizó el método de codificación One-Hot Encoder, y para las ordinales, el Ordinal Encoder.

7.4.2. Optimización Bayesiana de hiperparámetros

Los rangos de valores para el proceso de optimización bayesiana abarcaron aspectos como el tipo de imputación, la estrategia de imputación, el número de vecinos más cercanos, la cantidad de componentes principales y el valor de alpha para el método LASSO. Se definieron los siguientes rangos de valores para los hiperparámetros en el proceso de optimización bayesiana:

1. Tipo de imputación (*type_imputer*):
   - Rango: (0, 1)
   - Los valores entre 0 y 0.5 indican el uso de KNN para la imputación, mientras que los valores mayores a 0.5 indican el uso de imputación simple usando una medida de resumen.
2. Estrategia de imputación (*strategy*):
   - Rango: (0, 1)
   - Los valores entre 0 y 0.5 corresponden a la estrategia de imputación por media, y los valores mayores a 0.5 corresponden a la imputación por mediana.
3. Número de vecinos más cercanos (*knn_k*):
   - Rango: (3, 10)
   - Define el número de vecinos a considerar en la imputación KNN.
4. Cantidad de componentes principales (*num_comp*):
   - Rango: (2, 30)
   - Define el número de componentes principales a considerar en el ACP.
5. Valor de alpha para LASSO (*alpha*):
   - Rango: (0.01, 0.7)
   - Define el valor de penalización en la regresión *LASSO*, con valores entre 0.01 y 0.7.

7.4.3. Protocolo de entrenamiento y evaluación de modelos

En el desarrollo del presente estudio, se procedió a la búsqueda de hiperparámetros óptimos empleando el conjunto de entrenamiento, que constaba de 387 instancias. Para esta tarea, se utilizó la técnica de validación cruzada, específicamente mediante el método de Stratified K-Fold cross-validator. La elección de Stratified K-Fold como método de validación cruzada se fundamenta en su capacidad para preservar la proporcionalidad de las clases en cada uno de los segmentos del conjunto de datos. Esta característica es de suma importancia, especialmente en contextos donde existe un desequilibrio significativo entre las clases. Métodos más convencionales, como el K-Fold tradicional, podrían resultar en particiones que no reflejan adecuadamente la composición original del conjunto, introduciendo así sesgos potenciales en la evaluación del rendimiento del modelo. La metodología de Stratified K-Fold garantiza que cada pliegue conserve una distribución de clases proporcional, reflejando fielmente la distribución global del conjunto. Este enfoque es esencial para modelos de clasificación en los cuales la precisión en la representación de las clases es crítica para la validación de su eficacia y fiabilidad.[61]

En la Figura 7 se presenta el esquema ilustrativo del proceso de validación cruzada.

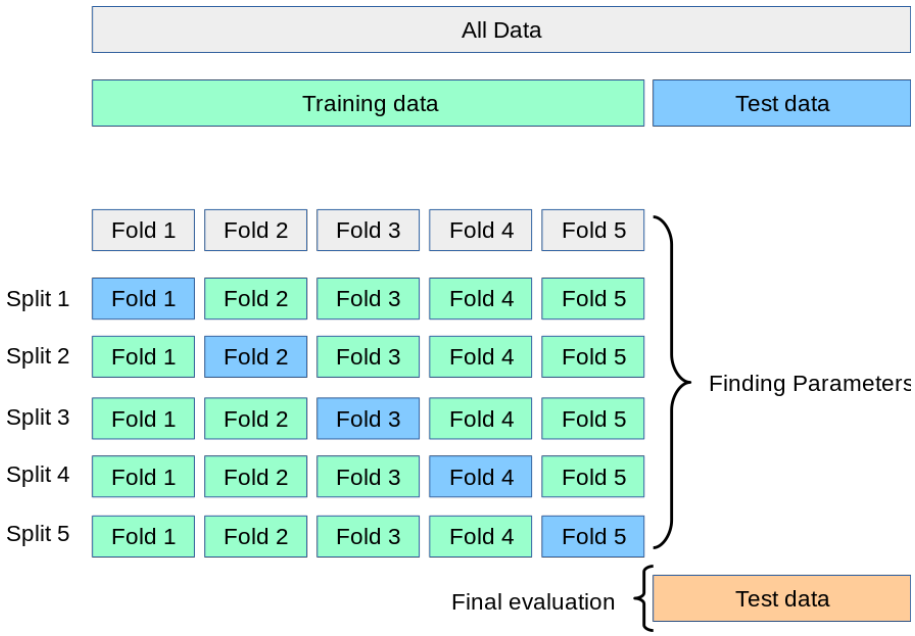


Figura 7: Representación esquemática validación cruzada. Tomado de **https://scikit-learn.org/stable/modules/cross˙validation.html**.

En el proceso de selección de los valores óptimos de los hiperparámetros, se decidió emplear el método de Optimización Bayesiana. Este enfoque se aplicó sistemáticamente a cada uno de los modelos entrenados en los K-Folds generados por la técnica de validación cruzada, donde K=2.

Para encontrar la combinación óptima de hiperparámetros utilizando este enfoque, se realizaron 50 pruebas por cada modelo. El espacio de búsqueda de hiperparámetros se construyó de manera dinámica durante este proceso.

La validación final del modelo se llevó a cabo aplicándolo al conjunto de prueba de 166 instancias. Este paso fue crucial para evaluar críticamente la capacidad de generalización del modelo en un contexto que simula condiciones de aplicación real. La evaluación permitió iden-

tificar el comportamiento del modelo frente a nuevos datos, fuera del entorno controlado de entrenamiento. Posteriormente, se realizó una estimación detallada y un análisis de las métricas de desempeño sobre el conjunto de prueba. El objetivo de este análisis fue verificar la estabilidad y confiabilidad del modelo bajo condiciones operativas no completamente controladas. A través de este proceso, se buscó asegurar que el modelo no solo fuera preciso, sino también robusto y fiable en su implementación práctica.
.

7.5. Muestras balanceadas

Para abordar el problema de desbalance de clases presente en los datos originales, donde la variable objetivo estaba desigualmente distribuida (muertos: 119 (21.5 %) y vivos: 434 (78.5 %)), se implementó un proceso de validación con el objetivo de evaluar la estabilidad de las métricas de los modelos previamente entrenados. Este proceso consistió en la extracción de tres muestras aleatorias balanceadas, cada una representando el 30 % del total de los datos. Las muestras fueron diseñadas para equilibrar las clases objetivo (vivo y muerto), permitiendo así una evaluación más justa de las métricas clave, como la F1, sensibilidad, especificidad y AUC, utilizando los modelos más efectivos previamente identificados con los datos desbalanceados.

Para complementar este enfoque de balanceo y mejorar la representatividad del conjunto de datos de entrenamiento (n=387), se aplicó la técnica SMOTE-NC. Esta técnica se utilizó para generar instancias sintéticas adicionales mediante la interpolación entre puntos de las clases minoritarias. En el caso de las variables continuas, SMOTE-NC emplea una interpolación lineal, mientras que para las variables nominales, selecciona la categoría más frecuente entre los k vecinos más cercanos. Este método no solo equilibra eficazmente las clases, sino que también asegura la coherencia y la integridad estructural de los datos, evitando la creación de muestras sintéticas que puedan distorsionar la representación del fenómeno estudiado.

8. Resultados

Esta sección inicia con la presentación de los hallazgos más relevantes obtenidos del análisis exploratorio de datos. A continuación, se resaltan conclusiones clave que fueron cruciales para la selección de variables importantes en el modelado predictivo para clasificación, utilizando varios algoritmos dentro de un flujo de trabajo tipo Pipeline. Posteriormente, se exponen los resultados obtenidos de los modelos de clasificación, evaluados mediante diversas métricas de desempeño. Por último, se presentan mapas que describen la distribución espacial, facilitando la visualización simultánea de variables clínicas y sociodemográficas en relación con la localización de comunas y estratos socioeconómicos en la ciudad de Cali.

8.1. Análisis exploratorio de datos

El conjunto de datos original constaba de 553 observaciones (pacientes) y 67 variables. En cuanto a la variable objetivo, la distribución de la muestra reveló que el 21.3 % de los pacientes fallecieron, lo que equivale a 119 individuos. Durante el proceso de análisis y enriquecimiento de datos, el número de variables aumentó a 79. Las nuevas columnas se derivaron de las variables existentes con el objetivo de mejorar o recuperar información relevante. Aunque no se encontraron registros duplicados, se observó la presencia de datos faltantes, los cuales se detallan en los Anexos, en la Tabla 9. Este análisis reveló un patrón significativo de ausencias informativas; por ejemplo, la columna ‘soat’ presentó 330 datos faltantes, equivalentes al 59.67 % del total, y la columna ‘seguridad social’ mostró 159 faltantes (28.75 %). Las columnas ‘hora egreso’ y ‘fecha egreso’ registraron 23 ausencias cada una (4.16 %). Aunque las tasas de datos faltantes en otras variables críticas fueron menores, algunas columnas presentaron solo un dato faltante,

representando apenas un 0.18 %.

En cuanto a los valores atípicos, se decidió no emplear el método del rango intercuartílico para su identificación debido al contexto clínico específico. Los casos extremos pueden representar variaciones significativas en la severidad o tipo de trauma, proporcionando perspectivas importantes sobre condiciones clínicas particulares o manifestaciones atípicas. En su lugar, se adoptó una metodología alternativa que implica la revisión de 20 casos por variable, identificando aquellos valores que se sitúan por debajo o por encima de los umbrales predeterminados para evaluar su relevancia clínica. Con base en el juicio médico o experto, se optó por mantener los datos sin modificaciones en lo que respecta a los valores atípicos, garantizando así la integridad y la relevancia clínica del análisis.

Caracterización sociodemográfica y clínica

Para comprender mejor el conjunto de datos, se realizó un análisis exploratorio de datos que incluyó tanto análisis univariado como bivariado. En el análisis univariado, se examinaron las distribuciones y características de cada variable individualmente mediante el uso de histogramas, diagramas de cajas y gráficos de barras. El análisis bivariado se centró en explorar las relaciones entre las variables independientes y la variable dependiente (clases vivo/muerto) utilizando gráficos de dispersión, diagramas de cajas y bigotes, y gráficos de barras agrupadas. Además, se construyó una tabla (ver Anexo, Tabla 9.) que presenta los estadísticos descriptivos y los p–value para cada variable en relación con las clases vivo/muerto, identificando diferencias significativas. En este documento, se presenta la tabla mencionada del análisis bivariado, mientras que los gráficos y otros resultados del análisis univariado y bivariado se encuentran disponibles en el notebook de Jupyter para una revisión más detallada.

En la tabla mencionada se describieron las características clínicas y demográficas de los pacientes con trauma, clasificadas por el resultado de vivo/muerte. Se encontró una mortalidad significativamente mayor en el Hospital Universitario del Valle (68.9 %) en comparación con otras instituciones (p–value < 0.001). La mayoría de los afectados eran hombres (85.9 %), quienes también representaron el 92.4 % de los fallecidos (p–value $= 0.030$). Las lesiones por arma de fuego tuvieron una alta letalidad, afectando al 70.1 % de los fallecidos frente al 38.5 % de los supervivientes (p–value < 0.001). Los traumas penetrantes se presentaron en el 75.4 % de los casos fatales (p–value < 0.001). Diferencias significativas en la presión arterial sistólica y la frecuencia respiratoria indicaron condiciones críticas en los fallecidos (p–value < 0.001). La escala de Glasgow fue significativamente más baja en los fallecidos (5.0) comparado con los no fallecidos (15.0) (p–value < 0.001).

Además, se encontró una asociación significativa entre la cobertura de seguridad social y la mortalidad (p–value < 0.001). El 56.3 % de los fallecidos fueron atendidos en un centro más lejano en comparación con el 45.4 % de los no fallecidos (p–value $= 0.045$), sugiriendo que la distancia al centro de atención puede influir en la mortalidad. La vulnerabilidad social no tuvo un impacto significativo en la mortalidad (p–value $= 0.5$), y el riesgo de embarazo adolescente tampoco mostró diferencias significativas (p–value $= 0.14$). La elegibilidad para empleo juvenil no fue relevante para la mortalidad (p–value $= 0.7$).

Estas variables se incluyeron en el análisis predictivo porque, aunque individualmente no mostraron diferencias significativas, en combinación con otras variables podrían proporcionar información valiosa sobre los resultados de los pacientes y ayudar a construir modelos predictivos más robustos y comprensivos.

Riesgos relativos univariados

Se utilizó la regresión de Poisson con errores estándar robustos para llevar a cabo un análisis univariado, asociando cada variable independiente con el desenlace de interés: la mortalidad. Esto permitió calcular los riesgos relativos (RR). La elección de este método frente a la regresión logística se fundamentó en varias ventajas: ofrece una interpretación directa de los RR, más intuitiva en estudios epidemiológicos; es adecuada para modelar tasas de incidencia y eventos, proporcionando estimaciones más estables y precisas; y los errores robustos corrigen la sobre-dispersión, mejorando la confiabilidad de las estimaciones. Los detalles completos de los resultados se incluyeron en la sección de Anexos, Figura 10.

Los hallazgos más notables incluyeron RRs elevados en lesiones de abdomen y pelvis (RR aproximado de 10) y en extremidades, ambos con significativa asociación con el desenlace de interés. En contraste, la Escala de Glasgow y lesiones en el tórax mostraron RRs bajos, indicando una menor probabilidad del desenlace, con el color azul denotando significancia estadística. Se observó variabilidad en los intervalos de confianza para lesiones en la cabeza y se notó el impacto moderado de factores socioeconómicos como el nivel de vulnerabilidad y la seguridad social. El transporte por ambulancia medicalizada se asoció con mejores resultados, mientras que mecanismos específicos de lesión, como los impactos por arma de fuego, fueron factores críticos para la probabilidad del desenlace.

Variables seleccionadas

A partir del análisis exploratorio y el cálculo de los p–value para evaluar la asociación con el desenlace, así como la estimación de RRs, y con la supervisión del criterio experto del médico especialista, se compiló una lista de variables seleccionadas para los análisis de modelación subsiguientes. Esta lista se detalla en los Anexos, Tabla 11.

8.2. Evaluación de modelos y análisis de resultados

Se evaluaron diversos modelos de clasificación como parte de un pipeline optimizado (Figura 6), que incluyó la búsqueda exhaustiva de los mejores hiperparámetros para la preparación de datos y algoritmos de clasificación, asegurando un rendimiento óptimo. Debido a la naturaleza desbalanceada de las clases, se utilizó el F1-score como métrica principal. También se construyó la matriz de confusión y se calcularon otras métricas para una visión más completa del rendimiento del modelo.

En la Tabla 7, se resume el desempeño de los modelos en los datos de validación y entrenamiento.

La Figura 8, se presentan los resultados de la matriz de confusión para cuatro de los mejores modelos, así como las curvas ROC para todos los modelos evaluados, utilizando el conjunto de prueba.

En el análisis realizado sobre diversos modelos predictivos de aprendizaje supervisado para predecir la mortalidad en pacientes con trauma, se observaron resultados destacados en varios algoritmos, excluyendo el SVM con kernel RBF por su posible subajuste, evidenciado por la discrepancia en el F1-score entre los datos de entrenamiento (0.35) y los de prueba (0.97).

Los modelos de Random Forest y XGBoost mostraron un desempeño robusto, con AUCs de 0.90 y 0.92 respectivamente, y registraron F1-scores sólidos tanto en el conjunto de entrenamiento (0.68 y 0.62) como en el de prueba (0.70 y 0.71), reflejando un equilibrio entre precisión

Tabla 7: Comparación del rendimiento de modelos de clasificación en los conjuntos de entrenamiento y prueba

Modelos	F1		Sensitivity		Specificity		AUC	
	Train	Test	Train	Test	Train	Test	Train	Test
Logistic regression	0.62	0.66	0.66	0.81	0.92	0.82	0.90	0.88
Lasso-Logistic regression	0.61	0.67	0.82	0.83	0.78	0.82	0.88	0.87
Ridge-Logistic regression	0.61	0.66	0.60	0.81	0.95	0.82	0.90	0.88
Elastic-net Logistic regression	0.61	0.65	0.67	0.78	0.93	0.83	0.91	0.87
SVM-Sigmoid kernel	0.58	0.61	0.60	0.47	0.88	0.98	0.76	0.73
SVM-RBF kernel	**0.35**	**0.97**	**1.00**	**0.94**	**1.00**	**1.00**	**1.00**	**1.00**
SVM-Linear kernel	**0.64**	**0.70**	**0.64**	**0.75**	**0.96**	**0.89**	**0.90**	**0.89**
SVM-Polynomial kernel	**0.65**	**0.69**	**0.57**	**0.67**	**0.96**	**0.92**	**0.84**	**0.86**
Random Forest	**0.68**	**0.70**	**0.59**	**0.72**	**0.98**	**0.91**	**0.89**	**0.90**
XGBoost	**0.62**	**0.71**	**0.80**	**0.72**	**0.96**	**0.92**	**0.94**	**0.92**
Neural Network	**0.66**	**0.73**	**0.66**	**0.83**	**0.95**	**0.88**	**0.89**	**0.90**

y recall en sus predicciones. Estas métricas, junto con las matrices de confusión y curvas ROC, subrayan la capacidad de estos modelos para clasificar los casos de mortalidad.

La red neuronal, con un AUC de 0.90 y un F1-score más bajo (0.66 en entrenamiento y 0.73 en prueba), sigue siendo una alternativa viable debido a su capacidad para manejar complejidades no lineales en datos de alta dimensión. Aunque su desempeño no es tan alto como el de Random Forest y XGBoost en términos de F1-score, su capacidad para capturar relaciones complejas en los datos la hace relevante, especialmente en contextos donde la dimensionalidad y las interacciones no lineales son significativas.

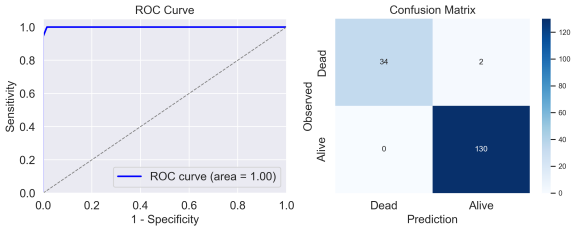
En particular, los modelos de SVM con kernel lineal y polinomial también mostraron un rendimiento notable, con AUCs de 0.89 y 0.86 respectivamente. El SVM con kernel lineal alcanzó un F1-score de 0.64 en entrenamiento y 0.70 en prueba, mientras que el SVM con kernel polinomial obtuvo un F1-score de 0.65 en entrenamiento y 0.69 en prueba. Estos resultados sugieren que los SVM también pueden ser efectivos para este tipo de problema, especialmente cuando se optimizan correctamente.

Estos resultados sugieren que tanto Random Forest como XGBoost son opciones prometedoras para aplicaciones clínicas, ofreciendo un rendimiento adecuado y capacidad de generalización, mientras que las redes neuronales podrían ser apropiadas para contextos donde las interacciones entre variables son complejas y el tamaño de la muestra es grande, reduciendo el riesgo de sobreajuste o subajuste. Los resultados de los hiperparámetros de los mejores modelos se presentan en la Tabla 12 de Anexos.

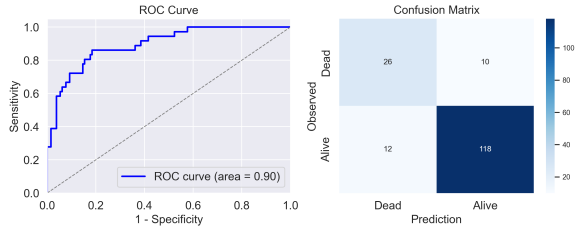
8.3. Efecto de las muestras balanceadas en los modelos destacados

Al comparar los resultados de las tres muestras (ver Anexo, Tabla 13) con los obtenidos del conjunto de prueba (Tabla 8), se observó que XGBoost y Random Forest mantuvieron estabilidad en F1 y AUC, aunque con un incremento en la sensibilidad y una disminución en la especificidad en la muestra 3, sugiriendo un mayor número de falsos positivos. RNN mostró una AUC constante, aunque con una ligera reducción en la especificidad en la muestra 3. SVM-RBF kernel, por su parte, alcanzó una sensibilidad perfecta en todas las muestras, pero con una especificidad baja, lo que sugiere un posible sobreajuste a la clase positiva.

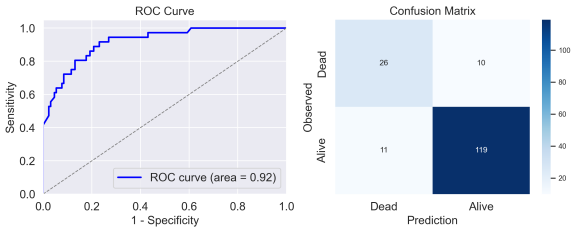
El balanceo de clases mediante SMOTE-NC tuvo un impacto diverso en el rendimiento de los modelos. En general, mejoró la sensibilidad en la mayoría de los casos, como en SVM con Kernel Polinómico y Random Forest, lo que sugiere una mejor detección de la clase minoritaria.



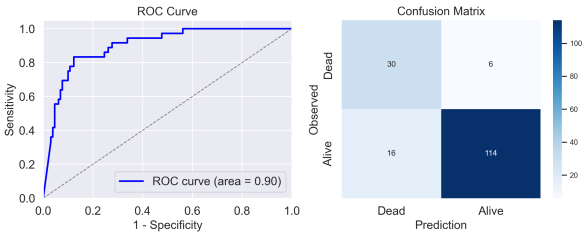
SVM-RBF Kernel



Random Forest



XGBoost



Neural Network

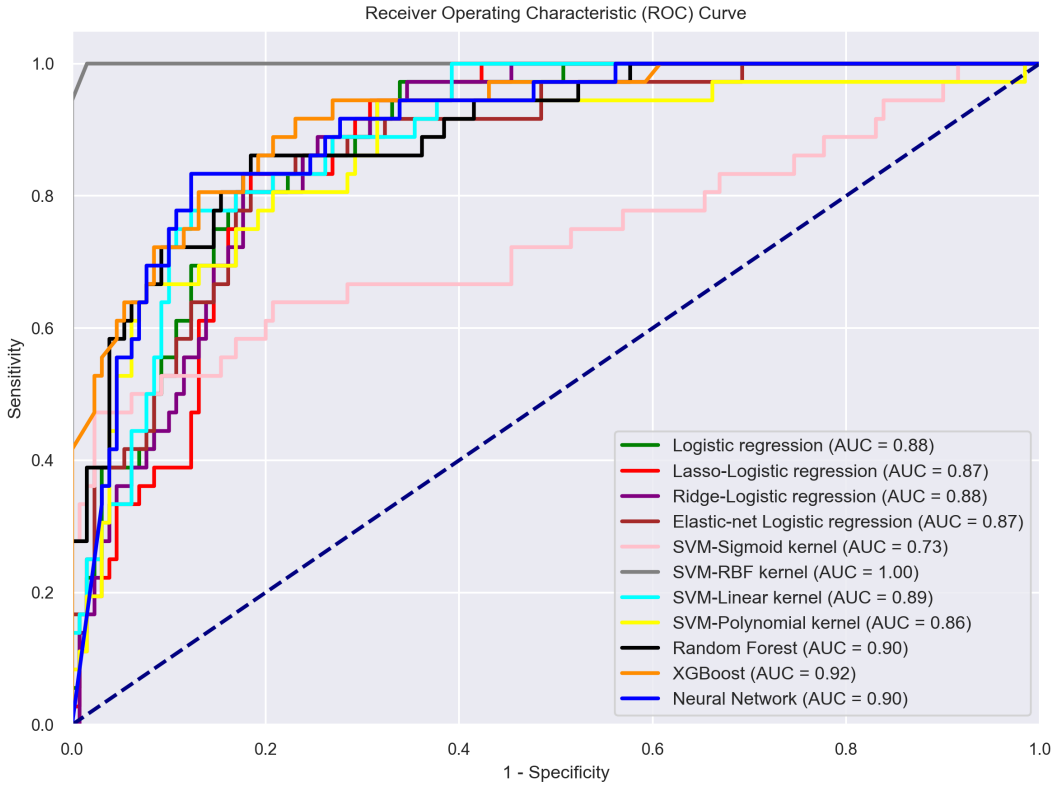


Figura 8: Curvas ROC y matrices de confusión de distintos modelos de predicción de mortalidad (regresiones logísticas, SVM, Random Forest, XGBoost y Redes Neuronales). Las curvas ROC muestran la relación entre sensibilidad y especificidad, mientras que las matrices de confusión ilustran el desempeño del modelo en términos de predicciones correctas e incorrectas.

Sin embargo, también redujo la F1 y la especificidad en modelos como Regresión Logística y XGBoost, indicando más falsos positivos. Algunos modelos, como SVM con Kernel RBF, mantuvieron un rendimiento sobreajustado, mientras que Elastic-net Logistic Regression mejoró en F1 y especificidad, pero redujo la sensibilidad.

Tabla 8: Comparación del rendimiento de modelos de clasificación en datos equilibrados con SMOTE-NC

Modelos	F1		Sensitivity		Specificity		AUC	
	Train	Test	Train	Test	Train	Test	Train	Test
Logistic regression	0.80	0.64	0.86	0.72	0.88	0.85	0.94	0.88
Lasso-Logistic regression	0.83	0.67	0.86	0.81	0.77	0.83	0.89	0.87
Ridge-Logistic regression	0.88	0.68	0.84	0.78	0.89	0.86	0.93	0.89
Elastic-net Logistic regression	0.88	0.69	0.90	0.72	0.87	0.90	0.95	0.88
SVM-Sigmoid kernel	0.77	0.54	0.90	0.81	0.52	0.67	0.82	0.74
SVM-RBF kernel	0.76	1.00	0.97	1.00	0.99	1.00	0.99	1.00
SVM-Linear kernel	0.84	0.67	0.88	0.78	0.82	0.85	0.92	0.88
SVM-Polynomial kernel	0.82	0.97	0.96	0.97	0.91	0.99	0.98	1.00
Random Forest	0.89	0.74	0.95	0.67	0.80	0.96	0.95	0.92
XGBoost	0.86	0.67	0.87	0.72	0.81	0.88	0.92	0.89
Neural Network	0.87	0.70	0.85	0.81	0.88	0.86	0.92	0.86

8.4. Visualización espacial: Mapas

Para el análisis de variables asociadas con información geográfica, como coordenadas, latitud y longitud, se generaron diversos mapas detallados de la ciudad de Cali y áreas adyacentes utilizando el software QGIS. Esta herramienta facilitó una representación precisa y detallada de las ubicaciones geográficas de interés.

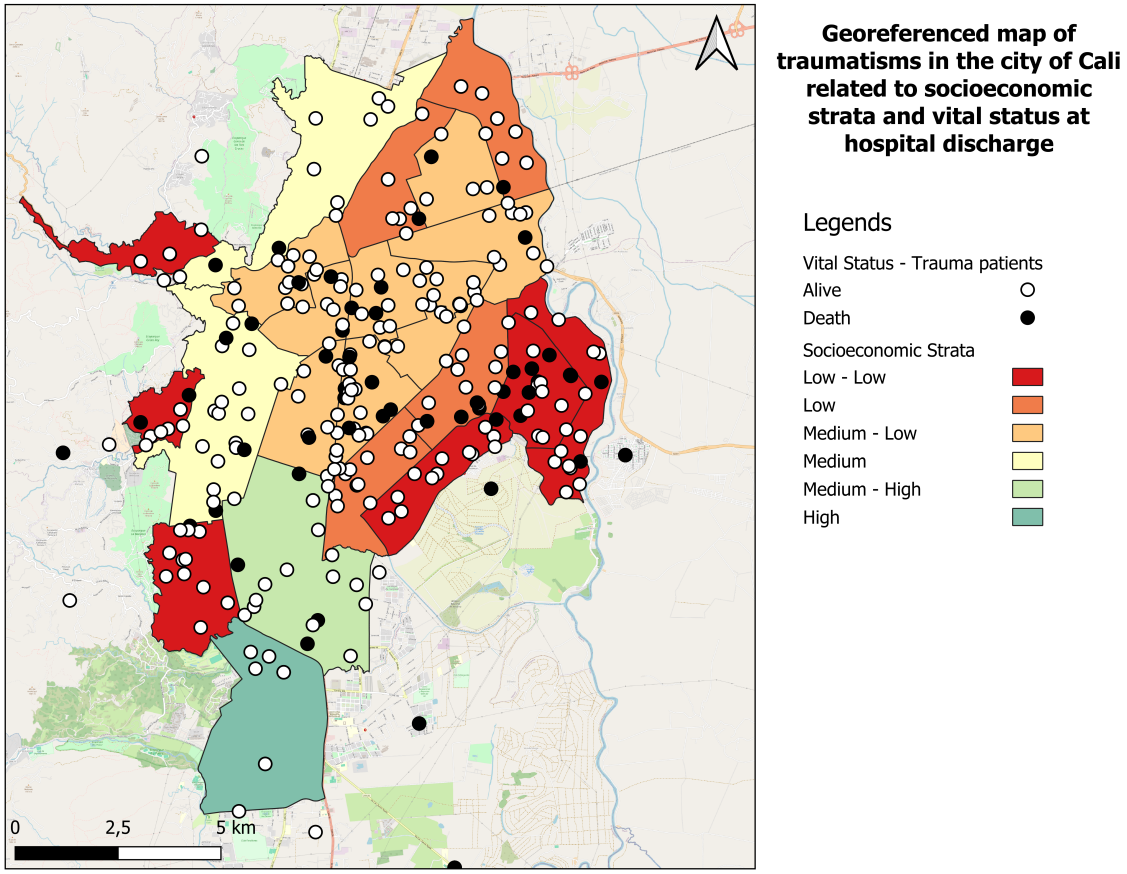


Figura 9: Geolocalización del trauma en Cali clasificado por estado vital y estrato socioecónomico del sitio del evento

El mapa de la Figura 9 es una descripción espacial del comportamiento de los eventos de trauma de la cohorte obtenida, los casos estan representados como los puntos blancos o negros, donde los puntos negros representan los pacientes fallecidos al final de la hospitalización y el mapa esta estratificado por estrato socio-ecónomico de la comuna donde ocurrió el evento. Este análisis a pesar de ser descriptivo, se puede evidenciar claramente como en las zonas más pobres de la ciudad ocurren más eventos y mueren mayor cantidad de pacientes. En la carpeta "Maps" de GitHub se pueden encontrar mapas adicionales para la visualización espacial.

9. Conclusiones y Limitaciones

A continuación, se destacan los principales resultados y aportes obtenidos a lo largo del proceso de investigación, así como los aprendizajes y recomendaciones que surgen de este trabajo:

Conclusiones

- Los modelos predictivos desarrollados en este estudio proporcionan una herramienta útil para la predicción temprana de la mortalidad en pacientes politraumatizados, lo cual es de gran valor para optimizar el tratamiento clínico y la toma de decisiones en situaciones de urgencia.
- Las variables sociodemográficas desempeñan un papel clínicamente relevante como determinantes de la salud en los resultados de los pacientes politraumatizados, al reflejar las condiciones del entorno en el que ocurre el evento traumático.
- Los modelos de ensamble, particularmente Random Forest y XGBoost, sobresalieron por su capacidad para equilibrar varianza y sesgo, además de su alta generalización y rendimiento en métricas clave como el AUC. Estos modelos demostraron ser una opción preferida para tareas de clasificación en comparación con modelos individuales, debido a su desempeño robusto.
- La técnica de balanceo de clases mediante SMOTE-NC mejoró la sensibilidad en la mayoría de los modelos, facilitando la detección de la clase minoritaria; sin embargo, a menudo redujo la F1 y la especificidad, incrementando los falsos positivos. Lo anterior sugiere que el balanceo de clases es útil para detectar la clase minoritaria, pero su aplicación debe ajustarse al contexto específico y al equilibrio deseado entre sensibilidad y precisión del modelo.

Limitaciones

- Durante el desarrollo del proyecto, se enfrentaron varias limitaciones importantes. El acceso a los datos del DANE presentó desafíos logísticos y de permisos, provocando retrasos. Además, el censo de 2018 no reflejaba los cambios más recientes, lo cual podría haber afectado la precisión de los análisis sociodemográficos. La conversión de datos a formatos específicos como Shapefile para el uso en QGIS añadió complejidad técnica. También, la integración de variables sociodemográficas con datos de eventos traumáticos implicó retos metodológicos, y la calidad y completitud de los datos era variable. Estas limitaciones deben tenerse en cuenta en estudios futuros.
- Es fundamental considerar las limitaciones de nuestro modelo al aplicarlo a conjuntos de datos más amplios o heterogéneos. Para mejorar su universalidad y estabilidad, será necesario ampliar la muestra en términos de escala y diversidad de los pacientes. Asimismo, resulta relevante considerar otros factores, como los componentes genéticos, inmunológicos y psicológicos, y explorar modelos alternativos de aprendizaje automático para mejorar la capacidad predictiva.

Referencias

[1] R. Lozano, M. Naghavi, K. Foreman et al., «Global and Regional Mortality from 235 Causes of Death for 20 Age Groups in 1990 and 2010: A Systematic Analysis for the Global Burden of Disease Study 2010,» *Lancet*, vol. 380, n.º 9859, págs. 2095-2128, 2012.

[2] Departamento Administrativo Nacional de Estadística (DANE). «DANE - Defunciones no Fetales 2022.» Accessed 2024-02-01. (2022), dirección: https://www.dane.gov.co/index.php/estadisticas-por-tema/salud/nacimientos-y-defunciones/defunciones-no-fetales/defunciones-no-fetales-2022.

[3] I. N. de Medicina Legal y Ciencias Forenses, *Forensis 2021, Datos para la Vida*. Instituto Nacional de Medicina Legal y Ciencias Forenses, 2021. dirección: https://www.medicinalegal.gov.co/documents/20143/878249/Forensis_2021.pdf.

[4] H. Lam, N. Harshaw, K. Bresz, C. T. Brown y L. L. Perea, «Social Determinants of Health in Trauma,» *The American surgeon*, vol. 89, n.º 8, págs. 3597-3599, 2023.

[5] Departamento Administrativo Nacional de Estadística (DANE). «DANE - Desigualdad Nacional 2022.» Accessed 2024-04-16. (2022), dirección: https://www.dane.gov.co/files/operaciones/PM/cp-PM-2022.pdf.

[6] J. A. Henry y A. L. Reingold, «Prehospital trauma systems reduce mortality in developing countries: a systematic review and meta-analysis,» *The journal of trauma and acute care surgery*, vol. 73, n.º 1, págs. 261-268, 2012.

[7] A. Uribe, M. Badiel, J. W. Tejada, J. H. Loaiza, L. F. Pino y M. Aboutanos, «Tendencia del Trauma en dos Hospitales Nivel IV en Cali, Colombia. Reporte Preliminar en la Plataforma del Registro de la Sociedad Panamericana de Trauma (SPT/RT),» *Panamerican Journal of Trauma, Critical Care & Emergency Surgery*, vol. 1, n.º 3, págs. 175-181, dic. de 2012.

[8] M. Badiel, J. W. Tejada, M. Aboutanos et al., «Epidemiología Del Trauma en Dos Hospitales de Primer Nivel de Atención Del Suroccidente de Colombia. Reporte Preliminar Del Registro Internacional de Trauma de la Sociedad Panamericana de Trauma,» *Panamerican Journal of Trauma, Critical Care & Emergency Surgery*, vol. 3, n.º 1, págs. 11-15, abr. de 2014.

[9] C. Ordóñez, W. Botache, L. Pino et al., «Experiencia en dos hospitales de tercer nivel de atención del suroccidente de Colombia en la aplicación del Registro Internacional de Trauma de la Sociedad Panamericana de Trauma,» *Revista Colombiana de Cirugia*, vol. 28, págs. 39-47, 2013.

[10] T. Gauss, F. X. Ageron, M. L. Devaud et al., «Association of Prehospital Time to In-Hospital Trauma Mortality in a Physician-Staffed Emergency Medicine System,» *JAMA surgery*, vol. 154, n.º 12, págs. 1117-1124, 2019.

[11] J. F. Waalwijk, R. van der Sluijs, R. D. Lokerman et al., «The impact of prehospital time intervals on mortality in moderately and severely injured patients,» *The journal of trauma and acute care surgery*, vol. 92, n.º 3, págs. 520-527, 2022.

[12] A. A. H. Nasser, C. Nederpelt, M. El Hechi et al., «Every minute counts: The impact of pre-hospital response time and scene time on mortality of penetrating trauma patients,» *American journal of surgery*, vol. 220, n.º 1, págs. 240-244, 2020.

[13] R. A. Cowley, «The resuscitation and stabilization of major multiple trauma patients in a trauma center environment,» *Clin Med (Northfield IL)*, vol. 83, págs. 16-22, 1976.

[14] L. Fox, M. L. Serre, S. J. Lippmann et al., «Spatiotemporal approaches to analyzing pedestrian fatalities: the case of Cali, Colombia,» *Traffic injury prevention*, vol. 16, n.º 6, págs. 571-577, 2015.

[15] C. H. Lasecki, F. C. Mujica, S. Stutsman et al., «Geospatial mapping can be used to identify geographic areas and social factors associated with intentional injury as targets for prevention efforts distinct to a given community,» *The journal of trauma and acute care surgery*, vol. 84, n.º 1, págs. 70-74, 2018.

[16] O. F. Hunter, F. Perry, M. Salehi y et al., «Science fiction or clinical reality: a review of the applications of artificial intelligence along the continuum of trauma care,» *World J Emerg Surg*, vol. 18, n.º 1, pág. 16, 2023.

[17] E. J. Tsiklidis, C. Sims, T. Sinno y S. L. Diamond, «Using the National Trauma Data Bank (NTDB) and machine learning to predict trauma patient mortality at admission,» *PloS one*, vol. 15, n.º 11, e0242166, 2020.

[18] N. T. Liu y J. Salinas, «Machine Learning for Predicting Outcomes in Trauma,» *Shock (Augusta, Ga.)*, vol. 48, n.º 5, págs. 504-510, 2017.

[19] M. B. Aboutanos, F. Mora, E. Rodas et al., «Ratification of IATSIC/WHO's guidelines for essential trauma care assessment in the South American region,» *World journal of surgery*, vol. 34, n.º 11, págs. 2735-2744, 2010.

[20] R. Soares y J. Naritomi, *The Economics of Crime: Lessons for and from Latin America*. University of Chicago Press, 2010, pág. 472.

[21] «FORENSIS 2014. DATOS PARA LA VIDA,» Instituto Nacional de Medicina Legal y Ciencias Forenses. (2014), dirección: https://www.medicinalegal.gov.co/documents/20143/49520/Forensis+2014+Datos+para+la+vida.pdf.

[22] «FORENSIS 2022. DATOS PARA LA VIDA,» Instituto Nacional de Medicina Legal y Ciencias Forenses. (2022), dirección: https://www.medicinalegal.gov.co/documents/20143/989825/Forensis_2022.pdf.

[23] A. D. Hill, R. A. Fowler y A. B. Nathens, «Impact of interhospital transfer on outcomes for trauma patients: a systematic review,» *The Journal of trauma*, vol. 71, n.º 6, págs. 1885-1901, 2011.

[24] A. M. Harmsen, G. F. Giannakopoulos, P. R. Moerbeek, E. P. Jansma, H. Bonjer y F. Bloemers, «The influence of prehospital time on trauma patients outcome: a systematic review,» *Injury*, vol. 46, n.º 4, págs. 602-609, 2015.

[25] M. K. Björklund, M. Cruickshank, R. A. Lendrum y K. Gillies, «Randomised controlled trials in prehospital trauma: a systematic mapping review,» *Scand J Trauma Resusc Emerg Med*, vol. 29, n.º 1, dic. de 2021.

[26] T. Williams, J. Finn, D. Fatovich e I. Jacobs, «Outcomes of different health care contexts for direct transport to a trauma center versus initial secondary center care: A systematic review and meta-analysis,» *Prehospital Emergency Care*, vol. 17, n.º 4, págs. 442-457, oct. de 2013.

[27] A. Pickering, K. Cooper, S. Harnan, A. Sutton, S. Mason y J. Nicholl, «Impact of prehospital transfer strategies in major trauma and head injury: Systematic review, meta-analysis, and recommendations for study design,» *Journal of Trauma and Acute Care Surgery*, vol. 78, págs. 164-177, 2015.

[28] J. Diaz, P. Norris, O. Gunter, B. Collier, W. Riordan y J. Morris, «Triaging to a regional acute care surgery center: Distance is critical,» *Journal of Trauma - Injury, Infection, and Critical Care*, vol. 70, n.º 1, págs. 116-119, ene. de 2011.

[29] M. Crandall, D. Sharp, E. Unger et al., «Trauma deserts: distance from a trauma center, transport times, and mortality from gunshot wounds in Chicago,» *American Journal of Public Health*, vol. 103, n.º 6, págs. 1103-1109, 2013.

[30] M. P. Jarman, F. C. Curriero, E. R. Haut, K. P. Porter y R. C. Castillo, «Associations of distance to trauma care, community income, and neighborhood median age with rates of injury mortality,» *JAMA Surgery*, vol. 153, n.º 6, págs. 535-543, jun. de 2018.

[31] D. Chambers, A. Cantrell, S. Baxter, J. Turner y A. Booth, *Effects of increased distance to urgent and emergency care facilities resulting from health services reconfiguration: a systematic review*. Southampton, UK: NIHR Journals Library, 2020.

[32] R. J. Mullins y N. C. Mann, «Population-based research assessing the effectiveness of trauma systems,» *Journal of Trauma and Acute Care Surgery*, vol. 47, n.º 3, S59-S66, 1999.

[33] S. Baker, B. O'Neill y R. Karpf, *The Injury Fact Book*. 1984.

[34] H. Tatsat, S. Puri y B. Lookabaugh, *Machine Learning and Data Science Blueprints for Finance*. O'Reilly media, 2020.

[35] International Business Machines Corporation, *Temas*, 2019. dirección: https://www.ibm.com/es-es/topics?topic=all&page=1.

[36] J. H. Friedman, «Greedy function approximation: A gradient boosting machine,» *The Annals of Statistics*, vol. 29, n.º 5, págs. 1189-1232, 2001.

[37] T. Chen y C. Guestrin, «XGBoost: A Scalable Tree Boosting System,» en *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, ago. de 2016.

[38] J. L. Reyes-Cruz, M. G. Sánchez-Trujillo y R. Mejía-Ramírez, «The influence of higher education on entrepreneurial attitudes,» 2021.

[39] M. G. Pose, *Introducción a las redes de neuronas artificiales*, Departamento de Tecnologías de la Información y las Comunicaciones, Universidad da Coruña, 2009.

[40] E. Alpaydin, *Introduction to Machine Learning*. Cambridge, MA: The MIT Press, 2004.

[41] J. Cohen, «A coefficient of agreement for nominal scales,» *Educational and Psychological Measurement*, vol. 20, n.º 1, págs. 37-46, 1960.

[42] P. Martí Sanahuja, *Entendiendo la curva ROC y el AUC: Dos medidas del rendimiento de un clasificador binario que van de la mano*, Último acceso: 20 de abril de 2024, 2023. dirección: https://polmartisanahuja.com/entendiendo-la-curva-roc-y-el-auc-dos-medidas-del-rendimiento-de-un-clasificador-binario-que-van-de-la-mano/.

[43] B. Bischl, M. Binder, M. Lang et al., «Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges,» *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 13, n.º 2, e1484, 2023.

[44] F. Alharbi, «Sampling Strategies for Tackling Imbalanced Data in Human Activity Recognition,» PhD Thesis, Goldsmiths, University of London, 2021. dirección: https://research.gold.ac.uk/id/eprint/30647/1/COM_thesis_AlharbiF_2021.pdf.

[45] N. V. Chawla, K. W. Bowyer, L. O. Hall y W. P. Kegelmeyer, «SMOTE: Synthetic Minority Over-sampling Technique,» *Journal of Artificial Intelligence Research*, vol. 16, págs. 321-357, jun. de 2002, Submitted 09/01; published 06/02. dirección: https://www.jair.org/index.php/jair/article/view/10302.

[46] M. Yu, S. Wang, K. He et al., «Predicting the complexity and mortality of polytrauma patients with machine learning models,» *Scientific reports*, vol. 14, n.º 1, pág. 8302, 2024.

[47] E. Vaz, M. D. Cusimano, F. Bação, B. Damásio y E. Penfound, «Open data and injuries in urban areas—A spatial analytical framework of Toronto using machine learning and spatial regressions,» *PloS one*, vol. 16, n.º 3, e0248285, 2021.

[48] A. J. Ghandour, H. Hammoud y S. Al-Hajj, «Analyzing factors associated with fatal road crashes: a machine learning approach,» *International Journal of Environmental Research and Public Health*, vol. 17, n.º 11, pág. 4111, 2020.

[49] R. Hassanzadeh, M. Farhadian y H. Rafieemehr, «Hospital mortality prediction in traumatic injuries patients: comparing different SMOTE-based machine learning algorithms,» *BMC Medical Research Methodology*, vol. 23, n.º 1, pág. 101, 2023.

[50] C. Dirago, M. Poulson, J. Hatchimonji, J. Byrne y D. Scantling, «Geospatial analysis of social vulnerability, race, and firearm violence in Chicago,» *Journal of Surgical Research*, vol. 294, págs. 66-72, 2024.

[51] J. Jay, «Alcohol outlets and firearm violence: a place-based case–control study using satellite imagery and machine learning,» *Injury Prevention*, vol. 26, n.º 1, págs. 61-66, 2020.

[52] A. F. Bedard, L. V. Mata, C. Dymond et al., «A scoping review of worldwide studies evaluating the effects of prehospital time on trauma outcomes,» *International Journal of Emergency Medicine*, vol. 13, págs. 1-19, 2020.

[53] C. A. Rojas-Trejos, J. González-Velasco y M. A. López-Ramírez, «Optimization Model for the Location of Prehospital Care Ambulances in the city of Cali, Colombia,» *International Journal of Combinatorial Optimization Problems and Informatics*, vol. 8, n.º 3, págs. 64-70, 2017.

[54] A. Oleś, *openrouteservice: Openrouteservice API Client*, Python package version 2.3.3, 2023. dirección: https://openrouteservice.org/.

[55] QGIS Development Team, *QGIS Geographic Information System*, QGIS Association, 2024. dirección: https://www.qgis.org.

[56] J. Hester y H. Wickham, *fs: Cross-Platform File System Operations Based on 'libuv'*, R package version 1.5.0, 2021. dirección: https://CRAN.R-project.org/package=fs.

[57] E. Pebesma, *sf: Simple Features for R*, R package version 1.0-2, 2021. dirección: https://CRAN.R-project.org/package=sf.

[58] R. J. Hijmans, *terra: Spatial Data Analysis*, R package version 1.3-4, 2021. dirección: https://CRAN.R-project.org/package=terra.

[59] J. Hester, *glue: Interpreted String Literals*, R package version 1.4.2, 2021. dirección: https://CRAN.R-project.org/package=glue.

[60] A. N. Specified, *distanceto: Distance Matrix Computation*, R package version not specified, Year Not Specified. dirección: https://CRAN.R-project.org/package=distanceto.

[61] L. Copete, *Cómo implementar un K-Fold Cross Validation a modelos de inteligencia artificial con scikit-learn*, https://medium.com/@luiscopete/como-implementar-un-k-fold-cross-validation-a-modelos-de-inteligencia-artificial-con-scikit-learn-eb0726c5ba55, Último acceso: 2024-04-25, 2019.

A. Anexos: Análisis exploratorio de datos

Tabla 9: Resumen de datos faltantes en el conjunto de datos

Variable	Faltantes	Total	Porcentaje (%)
soat	330	553	59.67
seg_social	159	553	28.75
transporte_llevado	51	553	9.22
muerte_asociada_trauma	38	553	6.87
fecha_egreso	23	553	4.16
hora_egreso	23	553	4.16
trauma_penetrante	22	553	3.98
fecha_nacimiento	11	553	1.99
hora_trauma	8	553	1.45
mecanismo_lesion	6	553	1.08
prob_sobrev_ing	4	553	0.72
b_init	4	553	0.72
prob_muerte_ing	4	553	0.72
edad_triss	3	553	0.54
fecha_trauma	2	553	0.36
score_severidad_urgencias	1	553	0.18
pcg_triss	1	553	0.18
escala_glasgow	1	553	0.18
ggm_int	1	553	0.18
score_sev_urg_anatomico	1	553	0.18

Tabla 10: Caracterización sociodemográfica y clínica

Variable	N	Overall, N = 553	Vivo, N = 434	Muerto, N = 119	p-value
Centro de atención	553				<0.001
Clínica Nuestra		152 (27.5 %)	144 (33.2 %)	8 (6.7 %)	
Fundación Valle del Lili		146 (26.4 %)	117 (38.2 %)	29 (24.4 %)	
Universitario del Valle		255 (46.1 %)	173 (7.4 %)	82 (68.9 %)	
Edad	549	29.0 (22.0, 40.0)	29.0 (22.0, 42.0)	27.0 (23.0, 35.0)	0.12
(Missing)		4	3	1	
Presión arterial sistolica	553	118.0 (99.0, 126.0)	120.0 (103.3, 126.8)	98.0 (69.5, 120.0)	<0.001
Frecuencia respiratoria	553	20.0 (18.0, 24.0)	20.0 (18.0, 24.0)	19.0 (12.0, 24.5)	<0.001
Escala glasgow	552	15.0 (9.8, 15.0)	15.0 (14.0, 15.0)	5.0 (3.0, 13.0)	<0.001
(Missing)		1	1	0	

Continúa en la próxima página

Variable	**N**	**Overall**, N = 553	**Vivo**, N = 434	**Muerto**, N = 119	**p-value**
Score severidad lesiones	553	17.0 (10.0, 26.0)	16.0 (10.0, 24.0)	29.0 (25.0, 41.0)	<0.001
Score severidad lesiones	553				0.5
No		179 (32.4 %)	137 (31.6 %)	42 (35.3 %)	
Sí		374 (67.6 %)	297 (68.4 %))	77 (64.7 %)	
Score severidad urgencias	552	12.0 (10.0, 12.0)	12.0 (11.0, 12.0)	8.0 (7.0, 10.5)	<0.001
(Missing)		1	1	0	
Score severidad urgencias anatómico	552	7.8 (6.2, 7.8)	7.8 (7.1, 7.8)	4.7 (4.1, 6.9)	<0.001
(Missing)		1	1	0	
Distance openrout tx center	553	7534.0 (4666.0, 10692.0)	7338.0 (4586.5, 10584.3)	8210.0 (5111.0, 12376.0)	0.040
Distancia retardo min	553	0.0 (0.0, 1032.0)	0.0 (0.0, 1030.3)	653.0 (0.0, 1129.5)	0.065
Índice pobreza multidimensional	553	7.4 (0.0, 18.6)	7.0 (0.0, 17.4)	10.8 (0.0, 20.5)	0.2
Viviendas totales manzana	553	12.0 (4.0, 19.0)	12.0 (4.0, 20.0)	12.0 (4.0, 18.0)	¿0.9
Hogares totales manzana	553	33.0 (2.0, 53.0)	33.0 (2.0, 41.0)	33.0 (3.0, 53.0)	0.3
Habitaciones habitantes	553	16.0 (1.0, 33.0)	7.0 (1.0, 33.0)	16.0 (3.0, 33.0)	0.2
Déficit habitacional cuantitativo	553	0.0 (0.0, 2.0)	0.0 (0.0, 2.0)	2.0 (0.0, 2.0)	0.025
Déficit habitacional cualitativo	553	10.0 (1.0, 31.0)	7.0 (1.0, 31.0)	11.0 (3.0, 31.0)	0.2
Porcentaje deficit vivienda	553	100.0 (30.2, 100.0)	100.0 (9.1, 100.0)	100.0 (30.2, 100.0)	0.2
Porcentaje deficit vivienda cuantitativo	553	0.0 (0.0, 6.1)	0.0 (0.0, 6.1)	6.1 (0.0, 6.1)	0.007
Porcentaje deficit vivienda cualitativo	553	93.9 (18.9, 95.7)	93.9 (9.1, 95.7)	93.9 (18.9, 95.7)	0.3
Sexo	553				0.030

Continúa en la próxima página

Variable	**N**	**Overall**, N = 553	**Vivo**, N = 434	**Muerto**, N = 119	**p-value**
Femenino		78 (14.1 %)	69 (15.9 %)	9 (7.6 %)	
Masculino		475 (85.9 %)	365 (84.1 %)	110 (92.4	
Mecanismo lesion	553				<0.001
Arma cortopunzante		45 (8.1 %)	38 (8.8 %)	7 (5.9 %)	
Arma de fuego		248 (44.8 %)	166 (38.2 %)	82 (68.9 %)	
Caídas		36 (6.5 %)	32 (7.4 %)	4 (3.4 %)	
Golpes		5 (0.9 %)	4 (0.9 %)	1 (0.8 %)	
Lesiones de tránsito		210 (38.0 %)	185 (42.6 %)	25 (21.0 %)	
Trauma contundente		9 (1.6 %)	9 (2.1 %)	0 (0.0 %)	
Trauma penetrante	553				<0.001
No		253 (45.8 %)	222 (51.2 %)	31 (26.1 %)	
Sí		300 (54.2 %)	212 (48.8 %)	88 (73.9 %)	
Transporte	553				0.2
Ambulancia básica		413 (74.7 %)	332 (76.5 %)	81 (68.1 %)	
Ambulancia medicalizada		70 (12.7 %)	51 (11.8 %)	19 (16.0 %)	
Fuerza pública		24 (4.3 %)	19 (4.4 %)	5 (4.2 %)	
Vehículo particular		46 (8.3 %)	32 (7.4 %)	14 (11.8 %)	
Seguridad social	553				<0.001
Contributivo		145 (26.2 %)	113 (26.0 %)	32 (26.9 %)	
Especial		13 (2.4 %)	10 (2.3 %)	3 (2.5 %)	
Otro		11 (2.0 %)	8 (1.8 %)	3 (2.5 %)	
Sin aseguramiento		88 (15.9 %)	59 (13.6 %)	29 (24.4 %)	
Soat		139 (25.1 %)	131 (30.2 %)	8 (6.7 %)	
Subsidiado		157 (28.4 %)	113 (26.0 %)	44 (37.0 %)	
Lesión cabeza	553				<0.001
0		278 (50.3 %)	241 (55.5 %)	37 (31.1 %)	
1		23 (4.2 %)	21 (4.8 %)	2 (1.7 %)	
2		38 (6.9 %)	37 (8.5 %)	1 (0.8 %)	
3		51 (9.2 %)	46 (10.6 %)	5 (4.2 %)	
4		78 (14.1 %)	71 (16.4 %)	7 (5.9 %)	
5		74 (13.4 %)	17 (3.9 %)	57 (47.9 %)	
6		11 (2.0 %)	1 (0.2 %)	10 (8.4 %)	

Continúa en la próxima página

Variable	**N**	**Overall**, N = 553	**Vivo**, N = 434	**Muerto**, N = 119	**p-value**
Lesión cabeza	553				<0.001
No		278 (50.3 %)	241 (55.5 %)	37 (31.1 %)	
Si		275 (49.7 %)	193 (44.5 %)	82 (68.9 %)	
Lesión cara	553				0.4
0		451 (81.6 %)	355 (81.8 %)	96 (80.7 %)	
1		24 (4.3 %)	20 (4.6 %)	4 (3.4 %)	
2		47 (8.5 %)	38 (8.8 %)	9 (7.6 %)	
3		26 (4.7 %)	18 (4.1 %)	8 (6.7 %)	
4		4 (0.7 %)	3 (0.7 %)	1 (0.8 %)	
5		1 (0.2 %)	0 (0.0 %)	1 (0.8 %)	
Lesión cara	553				0.9
No		451 (81.6 %)	355 (81.8 %)	96 (80.7 %)	
Si		102 (18.4 %)	79 (18.2 %)	23 (19.3 %)	
Lesión torax	553				<0.001
0		320 (57.9 %)	249 (57.4 %)	71 (59.7 %)	
1		11 (2.0 %)	8 (1.8 %)	3 (2.5 %)	
2		5 (0.9 %)	5 (1.2 %)	0 (0.0 %)	
3		112 (20.3 %)	101 (23.3 %)	11 (9.2 %)	
4		81 (14.6 %)	61 (14.1 %)	20 (16.8 %)	
5		23 (4.2 %)	10 (2.3 %)	13 (10.9 %)	
6		1 (0.2 %)	0 (0.0 %)	1 (0.8 %)	
Lesión torax	553				0.7
No		320 (57.9 %)	249 (57.4 %)	71 (59.7 %)	
Si		233 (42.1 %)	185 (1.8 %)	48 (40.3 %)	
Lesión abdomen pelvis	553				<0.001
0		390 (70.5 %)	312 (71.9 %)	78 (65.5 %)	
1		15 (2.7 %)	15 (3.5 %)	0 (0.0 %)	
2		20 (3.6 %)	15 (3.5 %)	5 (4.2 %)	
3		39 (7.1 %)	34 (7.8 %)	5 (4.2 %)	
4		50 (9.0 %)	40 (9.2 %)	10 (8.4 %)	
5		39 (7.1 %)	18 (4.1 %)	21 (17.6 %)	
Lesión abdomen pelvis	553				0.2
No		390 (70.5 %)	312 (71.9 %)	78 (65.5 %)	
Si		163 (29.5 %)	122 (28.1 %)	41 (34.5 %)	
Lesión extremidades	553				<0.001
0		325 (58.8 %)	229 (52.8 %)	96 (80.7 %)	

Continúa en la próxima página

Variable	**N**	**Overall**, N = 553	**Vivo**, N = 434	**Muerto**, N = 119	**p-value**
1		34 (6.1 %)	30 (6.9 %)	4 (3.4 %)	
2		74 (13.4 %)	66 (15.2 %)	8 (6.7 %)	
3		100 (18.1 %)	95 (21.9 %)	5 (4.2 %)	
4		19 (3.4 %)	14 (3.2 %)	5 (4.2 %)	
5		1 (0.2 %)	0 (0.0 %)	1 (0.8 %)	
Lesión extremidades	553				<0.001
No		325 (58.8 %)	229 (52.8 %)	96 (80.7 %)	
Si		228 (41.2 %)	205 (47.2 %)	23 (19.3 %)	
Lesión externo	553				<0.001
0		452 (81.7 %)	342 (78.8 %)	110 (92.4 %)	
1		88 (15.9 %)	84 (19.4 %)	4 (3.4 %)	
2		7 (1.3 %)	6 (1.4 %)	1 (0.8 %)	
3		1 (0.2 %)	0 (0.0 %)	1 (0.8 %)	
4		2 (0.4 %)	1 (0.2 %)	1 (0.8 %)	
5		3 (0.5 %)	1 (0.2 %)	2 (1.7 %)	
Lesión externo	553				0.001
No		452 (81.7 %)	342 (78.8 %)	110 (92.4 %)	
Si		101 (18.3 %)	92 (21.2 %)	9 (7.6 %)	
Estrato evento trauma	553				0.3
1		129 (23.3 %)	96 (22.1 %)	33 (27.7 %)	
2		181 (32.7 %)	138 (31.8 %)	43 (36.1 %)	
3		170 (30.7 %)	140 (32.3 %)	30 (25.2 %)	
4		27 (4.9 %)	21 (4.8 %)	6 (5.0 %)	
5		38 (6.9 %)	31 (7.1 %)	7 (5.9 %)	
6		8 (1.4 %)	8 (1.8 %)	0 (0.0 %)	
Atención centro más lejano	553				0.045
No		289 (52.3 %)	237 (54.6 %)	52 (43.7 %)	
Sí		264 (47.7 %)	197 (45.4 %)	67 (56.3 %)	
Nivel vulnerabilidad	553				0.5
Baja		257 (46.5 %)	194 (44.7 %)	63 (52.9 %)	
Alta		4 (0.7 %)	4 (0.9 %)	0 (0.0 %)	
Media		176 (31.8 %)	141 (32.5 %)	35 (29.4 %)	
Media-Alta		99 (17.9 %)	81 (18.7 %)	18 (15.1 %)	
Media-Baja		17 (3.1 %)	14 (3.2 %)	3 (2.5 %)	

Continúa en la próxima página

Variable	**N**	**Overall**, N = 553	**Vivo**, N = 434	**Muerto**, N = 119	**p-value**
Riesgo embarazo adolescente	553				0.14
Baja		345 (62.4 %)	276 (63.6 %)	69 (58.0 %)	
Alta		6 (1.1 %)	6 (1.4 %)	0 (0.0 %)	
Media		15 (2.7 %)	14 (3.2 %)	1 (0.8 %)	
Media-Alta		52 (9.4 %)	40 (9.2 %)	12 (10.1 %)	
Media-Baja		135 (24.4 %)	98 (22.6 %)	37 (31.1 %)	
Elegibilidad jovenes empleo	553				0.7
Baja		12 (2.2 %)	9 (2.1 %)	3 (2.5 %)	
Alta		1 (0.2 %)	1 (0.2 %)	0 (0.0 %)	
Media		249 (45.0 %)	189 (43.5 %)	60 (50.4 %)	
Media-Alta		29 (5.2 %)	23 (5.3 %)	6 (5.0 %)	
Media-Baja		262 (47.4 %)	212 (48.8 %)	50 (42.0 %)	

*Para características categóricas, la tabla muestra el número (porcentaje) de pacientes en la clase; para características continuas, la tabla muestra el valor mediana(Q3-Q1) de esta característica de todos los pacientes de la clase.

B. Anexos: Optimización de hiperparámetros

Se configuraron los hiperparámetros de cuatro modelos de aprendizaje automático (SVM con núcleo RBF, Random Forest, XGBoost y Redes Neuronales Recurrentes) para predecir el riesgo de mortalidad en pacientes con trauma (ver Tabla 12).

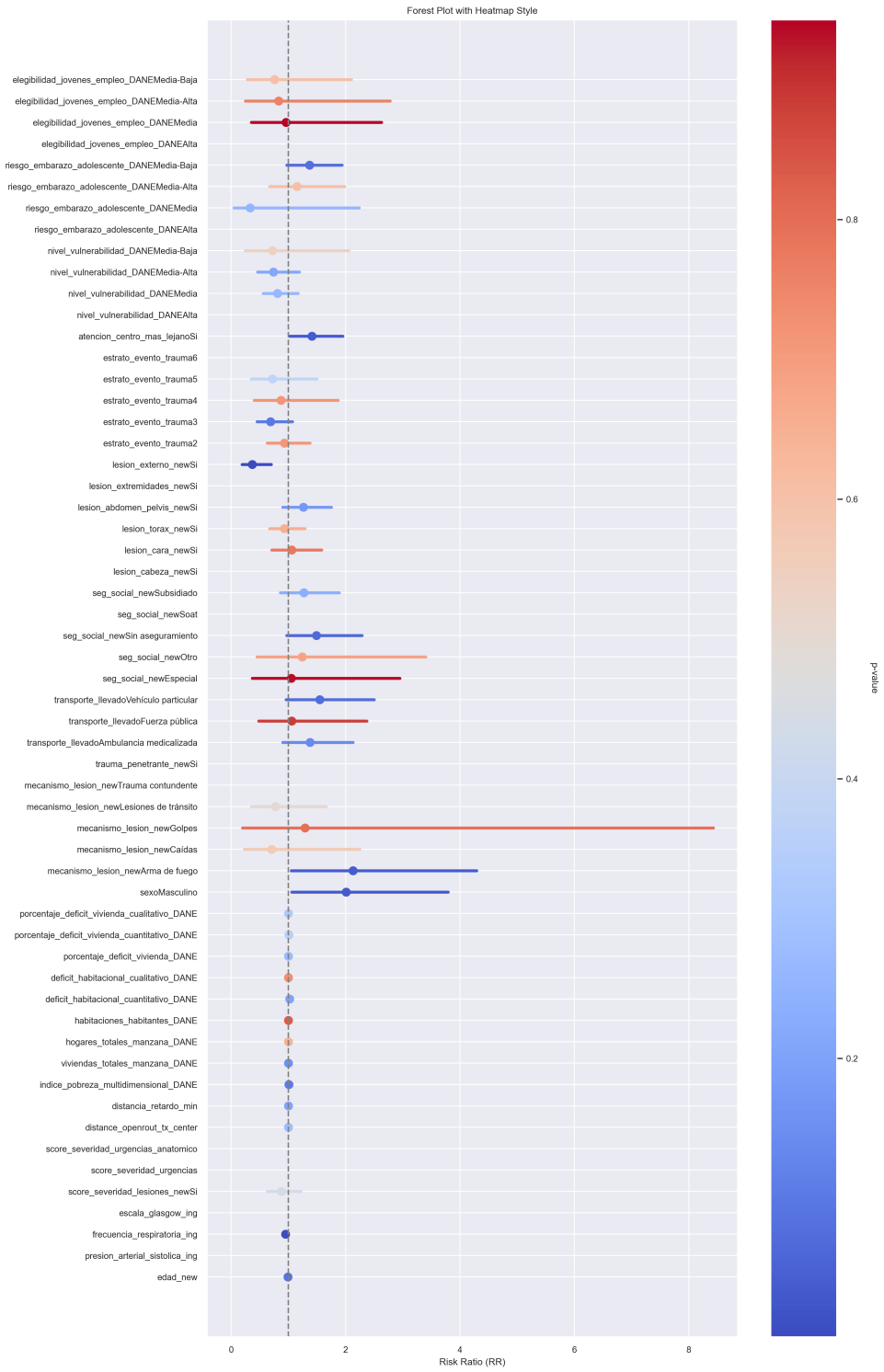


Figura 10: Análisis univariado del modelo Poisson con errores estándar robustos

Tabla 11: Lista de variables seleccionadas para el análisis de modelado

Variable	**Descripción**
edad_new	Edad ajustada
presion_arterial_sistolica_ing	Presión arterial sistólica
frecuencia_respiratoria_ing	Frecuencia respiratoria
escala_glasgow_ing	Escala de Glasgow
score_severidad_lesiones	Score de severidad de lesiones
score_severidad_urgencias	Score de urgencias general
score_severidad_urgencias_anatomico	Score anatómico de urgencias
distance_openrout_tx_center(m)	Distancia al centro de tratamiento
distancia_retardo_min	Tiempo de retardo en minutos
indice_pobreza_multidimensional_DANE	Índice de pobreza multidimensional
viviendas_totales_manzana_DANE	Viviendas totales por manzana
hogares_totales_manzana_DANE	Hogares totales por manzana
habitaciones_habitantes_DANE	Habitaciones por habitante
deficit_habitacional_cuantitativo_DANE	Déficit habitacional cuantitativo
deficit_habitacional_cualitativo_DANE	Déficit habitacional cualitativo
porcentaje_deficit_vivienda_DANE	Porcentaje de déficit de vivienda
porcentaje_deficit_vivienda_cuantitativo_DANE	Porcentaje de déficit de vivienda
porcentaje_deficit_vivienda_cualitativo_DANE	Porcentaje de déficit de vivienda
muerte	**Estado de muerte**
sexo	Género
mecanismo_lesion_new	Mecanismo de lesión
trauma_penetrante_new	Trauma penetrante
transporte_llevado	Tipo de transporte
seg_social_new	Cobertura de seguridad social
lesion_cabeza	Lesión en la cabeza
lesion_cara	Lesión en la cara
lesion_torax	Lesión en el tórax
lesion_abdomen_pelvis	Lesión en abdomen y pelvis
lesion_extremidades	Lesión en extremidades
lesion_externo	Lesión externa
estrato_evento_trauma	Estrato del evento de trauma
atencion_centro_mas_lejano	Atención en centro más lejano
nivel_vulnerabilidad_DANE	Nivel de vulnerabilidad
riesgo_embarazo_adolescente_DANE	Riesgo de embarazo adolescente
elegibilidad_jovenes_empleo_DANE	Elegibilidad para empleo joven

Tabla 12: Hiperparámetros para modelos de aprendizaje automático seleccionados

Modelo	**Ajuste de hiperparámetros**
SVM (RBF)	SVM_C=3.968277974832469, alpha=0.38178354646231627, knn_k=5.934361600823063, num_comp=21.186146011109265, strategy=0.20445224973151743, svm_rbf_gamma=4.390587181954727, type_imputer=0.027387593197926163, weight=10.716077632319228
Random Forest	alpha=0.2433968083321761, knn_k=7.486003971091739, max_depth=1.8383286989834844, min_samples_leaf=1.7423000890358622, min_samples_split=2.5216850489524094, n_estimators=128.73108977084684, num_comp=7.236005560485336, strategy=0.2857078727288883, type_imputer=0.3322628963048947
XGBoost	alpha=0.11774321288381494, gamma=3.7596621814411986, knn_k=8.048116969376334, learning_rate=0.12913140306590823, max_depth=6.229437266463972, n_estimators=642.3701731715681, num_comp=13.505866910747033, strategy=0.8933175601597346, type_imputer=0.24988436102278289
Red Neuronal	alpha=0.08752578559856267, batch_size=6.49455170558776, beta=-0.0844379205990704, hidden_layer_size_exp=7.974734746092953, knn_k=5.420276709712419, lr_init=-0.6758844978167464, max_iter=295.71912702177883, num_comp=16.58273872957414, solver=0.22976557408536602, strategy=0.9830086980623106, type_imputer=0.5590428472290561

C. Anexos: Evaluación de modelos con muestras balanceadas

Tabla 13: Comparación de métricas por modelo con muestras balanceadas

	muestra 1	**muestra 2**	**muestra 3**
XGBoost - F1	0.76	0.77	0.79
XGBoost - Sensitivity	0.88	0.88	0.93
XGBoost - Specificity	0.84	0.82	0.67
XGBoost - AUC	0.80	0.82	0.83
Random Forest - F1	0.77	0.77	0.76
Random Forest - Sensitivity	0.73	0.74	0.82
Random Forest - Specificity	0.84	0.84	0.67
Random Forest - AUC	0.80	0.82	0.83
Neural Network - F1	0.76	0.77	0.76
Neural Network - Sensitivity	0.70	0.72	0.80
Neural Network - Specificity	0.87	0.84	0.70
Neural Network - AUC	0.80	0.82	0.83
SVM-RBF kernel - F1	0.77	0.72	0.75
SVM-RBF kernel - Sensitivity	1.00	1.00	1.00
SVM-RBF kernel - Specificity	0.40	0.22	0.32
SVM-RBF kernel - AUC	0.92	0.85	0.88