BUAP

Reconocimiento de patrones de mortalidad en la población mexicana usando aprendizaje automático

Tesis que para obtener el título de:

Maestro en ciencias de la computación

Presenta:

Rubén Aguirre Agustín

Asesor:

Dr. Guillermo De Ita Luna

Co-asesora:

Dra. Mireya Tovar Vidal

Benemérita Universidad Autónoma de Puebla

Facultad de Ciencias de la Computación

Puebla, México

Octubre 2023

Agradecimientos

Al Consejo Nacional de Ciencia y Tecnología (CONACYT), por la beca otorgada en la totalidad de los años de estudios.
Al Dr. Guillermo de Ita Luna y a la Dra. Mireya Tovar Vidal, por su valioso apoyo en la realización de éste trabajo de tesis y por su excelente labor docente.
A mi familia y amigos, por su apoyo incondicional en los momentos de necesidad.
A todos mis profesores, pues son ellos quienes me motivaron con su ejemplo y me brindaron grandes enseñanzas.

Resumen

En este trabajo se realiza la implementación de herramientas de aprendizaje automático como son redes neuronales artificiales, máquinas de vector soporte y la aplicación del algoritmo de clasificación conocido como CART. El objetivo de aplicar estos sistemas de aprendizaje fue el reconocer los patrones existentes en los datos de la mortalidad en la población mexicana de los años 2010 al 2021. En otras palabras, se buscó relacionar variables socieconómicas con causas de defunción específicas.
En los primeros capítulos se describen las variables seleccionadas, y en general, se presentan los datos recolectados. También se explica la metodología a seguir: definiendo que variables se usarán como entrada para cada causa; ésta selección se hace a través de la correlación de Pearson y del análisis de regresión mediante mínimos cuadrados generalizados (GLS por sus siglas en Inglés).
Después se procede a aplicar algoritmos provenientes del aprendizaje automático sobre los datos recolectados. Finalmente, se muestran los resultados y las conclusiones obtenidas.
Una de las principales conclusiones obtenidas en este trabajo, fue reconocer que para este tipo de problema, el aplicar un algoritmo de clasificación tipo árbol de decisión (CART) obtuvo mejores resultados que los obtenidos al aplicar redes neuronales artificiales y máquinas de vector soporte.
Los resultados de este trabajo de tesis se presentan a modo de gráficas y tablas en la última sección de este documento de tesis.

Palabras clave: ANN, SVM, Reconocimiento de patrones, Mexico, Mortalidad, Aprendizaje automático, Ciencia de datos .

Contenido

1. Introducción

Uno de los principales problemas a nivel global es mejorar la calidad de vida de la población. Organismos como la Organización Mundial de la Salud (OMS) y la Organización de las Naciones Unidas (ONU), han impulsado de forma universal a los gobiernos de los distintos países, sobre todo los que se encuentran en desarrollo, a destinar mayores recursos en cuestiones de salud, educación y seguridad [33].

Existen diversas variables de importancia que permiten medir la calidad de vida en una población, por ejemplo, se miden indicadores en las áreas de demografía, salud, economía, empleo, seguridad, etc [23]. Dentro del área demografica, hay tres causas principales que le hacen oscilar, estas son: natalidad, mortalidad y migración.

De interés particular para este trabajo de investigación, fue realizar un análisis; primeramente del tipo estadístico y posteriormente, aplicando algoritmos de aprendizaje sobre datos recolectados acerca de la mortalidad en la población mexicana de los años 2010 al 2021.

Los valores sobre mortalidad en México los hemos correlacionado con diversas variables de la población, como son: grados de seguridad, salud y situación económico–social de la población de México, durante los años indicados [37, 43].

En México, el Instituto Nacional de Estadística y Geografía (INEGI) reporta anualmente estadísticas sobre las defunciones junto con las causas principales de muerte en la población Mexicana[18]. Sin embargo, estos estudios no aportan información sobre la correlación de la mortalidad en México con otras variables socio–económicas, como pueden ser: empleo, salud, PIB, etc.

Por lo anterior, es necesario realizar un estudio de la mortalidad en el país siguiendo un enfoque diferente, así, en este proyecto se propone utilizar el aprendizaje automático. Según [14], el aprendizaje automático se ocupa de desarrollar algoritmos capaces de aprender y, constituye junto con la estadística, el corazón del análisis inteligente de datos.

Lo primero que debemos tener, para aplicar algoritmos de aprendizaje automático, es el recolectar serie de datos sobre cuales aplicar estos algoritmos. Así que, la primer tarea en este trabajo, fue la construcción de tablas de datos sobre mortalidad, y sobre variables socio–económicas a utilizar para la correlacción, todo sobre la población Mexicana de los años 2010 al 2021.

1.1. Objetivo general

Aplicar algoritmos de aprendizaje automático a datos sobre mortalidad en México para reconocer patrones subyacentes en los datos.

1.1.1. Objetivos específicos

- Recopilar datos públicos de mortalidad.
- Recopilar datos públicos de carácter socieconómico.
- Unificar y estandarizar los datos recopilados.
- Aplicar métodos estadísticos para encontrar las primeras correlaciones significativas entre variables.
- Implementar diversos métodos del aprendizaje automático sobre los datos recolectados.
- Realizar un análisis de las correlaciones entre mortalidad y las otras variables seleccionadas.
- Analizar e interpretar los resultados obtenidos

2. Marco Teórico

En este capítulo se presentan algunos conceptos necesarios y fundamentales para el desarrollo de esta investigación; además, se presenta de forma detallada el problema a resolver.

2.1. Planteamiento del problema

Como se presentó anteriormente, la mortalidad forma parte importante de los indicadores de calidad de vida según organismos como la ONU y la OMS. En el país, el Instituto Nacional de Estadística y Geografía (INEGI) reporta anualmente la estadística de defunciones generales junto con sus causas principales, ofrece además los porcentajes de mortalidad de acuerdo al género y la edad; también se puede consultar tasas de defunciones por entidad federativa. Sin embargo, esta información no es suficiente, puesto que no evalúa la relación que existe entre cada causa de muerte y diferentes variables socieconómicas de México. Por lo anterior, es necesario realizar un estudio de la mortalidad en el país que relacione causas de muerte de la población Mexicana con variables socioeconómicas del país. En este trabajo de tesis, se propone utilizar métodos de aprendizaje automático para encontrar patrones que relacionen causas de muerte en México con variables socieconómicas del país.

2.2. Estado del arte

En México se han realizado diversos estudios tales como Mortalidad y Pobreza [43], Análisis de mortalidad por causas [8]. Sin embargo, estos son estudios en su mayoría estadísticos, por ejemplo, las tablas de vida [6] y los análisis de regresión como sucede con las tablas modelo [5]. Sin embargo, aunque los métodos estadísticos son la forma mas tradicional de análizar datos, hoy día es necesario complementar estas investigaciones con herramientas de aprendizaje automático para obtener diferentes patrones de comportamiento.
Hay otras investigaciones realizadas por instituciones publicas como el INEGI [18], en estos frecuentemente reportan las estadísticas y proyecciones en demografía, natalidad y mortalidad.
Otro tipo de investigaciones son aquellas relacionadas con una única causa de muerte, por ejemplo, [1] donde se mide la probabilidad de una muerte por Covid19 de acuerdo a algunos factores del paciente. En [7], se realiza un estudio con el objetivo de predecir el cáncer de mama utilizando tecnicas de aprendizaje automático.

En [28], se presenta un analisis sobre la mortalidad infantil en Colombia aplicando herramientas de aprendizaje automático. Sin embargo, no es común encontrar estudios sobre la relación de las causas de muerte con variables socieconómicas.

En [10] se realiza un análisis del impacto de los gastos del gobierno en la mortalidad infantil. Los autores recopilan información de la Organización Mundial de la Salud (OMS) del periodo comprendido entre los años 2013 a 2017. Algunas de las variables disponibles son: índice de pobreza, producto interno bruto, acceso a los servicios de salud, expectativa de vida escolar, población y número de muertes. Los autores utilizaron el algoritmo de Árboles de Clasificación y Regresión (CART) y un modelo Bayesiano para poder clasificar los datos. Los resultados obtenidos muestran que los gastos del gobierno relacionados a los servicios de salud pueden reducir las tasas de mortalidad neonatal(primeros 28 días) mientras que, aquellos gastos no relacionados directamente a la salud pueden reducir la mortalidad en infantes de hasta cinco años de edad.

En [30] se realiza un estudio sobre las tasas de mortalidad materna en la India; para ello los autores utilizaron la información disponible recabada por el gobierno de la India, en el cual se toman en cuenta variables como la fertilidad, planeación familiar, salud reproductiva y nutrición. Además, tratan de incluir datos de infraestructura como el número de hospitales y el número de médicos. Se utilizan el algoritmo CART, redes neuronales y máquinas de vector soporte. Los resultados obtenidos demuestran que el acceso a los servicios de salud y la infraestructura son los factores más relevantes para la mortalidad materna.

Otro estudio relevante fue publicado por Lancet Regional Journal [29], en esta investigación los autores utilizan un modelo de regresión de Poisson para predecir las causas especificas de mortalidad afectadas durante la pandemia de Covid19 basándose en los datos de mortalidad en el periodo 2015-2019. Los autores en [29] demostraron que causas como la Diabetes, las infecciones respiratorias, las enfermedades isquémicas del corazón y las relacionadas a la hipertensión se incrementaron. Por otro lado, algunas otras causas se redujeron como aquellas enfermedades relacionadas con parásitos y las muertes por accidentes no relacionadas con el tráfico.

Por otro lado, en Perú se realizó una investigación sobre la relación del Covid19 con otras causas de muerte [22]. Algunas de las conclusiones importantes del artículo fueron que no se relacionan con el Covid19 las muertes por infarto al miocardio (cuando no llega sangre a alguna parte del corazón), paro cardíaco (problema de impulso eléctrico) e insuficiencia cardíaca (cuando la presión no es suficiente). Además, esto explicó las deficiencias parciales de los servicios para atender emergencias circulatorias en la mayoría de los hospitales peruanos. Sin embargo, en este artículo no queda clara la metodología utilizada para llegar a dichas conclusiones.

En [42] se realizó un análisis sobre el impacto de variables sociales, económicas y ambientales en la salud de una provincia en China. El método utilizado es la combinación de un algoritmo Random Forest y Extreme Gradient Boosting, ambos relacionados con árboles de decisión. Los resultados demostraron que los factores económicos tienen un impacto directo en la salud

mientras que, los factores ambientales tienen un efecto retardado en la salud y por último los factores sociales tienen una relación demasiado compleja.
En [12] se hace una investigación con el objetivo de predecir la mortalidad en Dinamarca para personas mayores de 65 años. Los datos utilizados son privados y corresponden al año 2016, en ellos se incluyen variables para edad, sexo, residente o inmigrante, nivel educativo, estado civil y servicios de salud recibidos (enfermeros en casa). Para el análisis de los datos se utilizaron algoritmos XGBoost, Random Forest y regresión de Lasso. Del anterior estudio se demostró que la batería de algoritmos es capaz de clasificar bastante bien con una métrica de área bajo la curva (AUC) con valor de 0,87.
En [35] se hace un análisis de la expectativa de vida para tratar de identificar los factores más importantes que la afectan, ya sea positivamente o negativamente. Los autores utilizan la base de datos de la Organización Mundial de la Salud (OMS) del año 2010 al 2015 en las areas económicas, salud, atributos personales y atributos sociales. Los datos fueron procesados utilizando unicamente algoritmos de árbol de decisión; especificamente el algoritmo CART, Random Forest, Extratree y XGBoost. Los resultados obtenidos muestran que la seguridad social, el producto interno bruto, el índice de masa corporal y las enfermedades de transmisión sexual son los factores más importantes en general.
Como se ha explicado anteriormente, no existen estudios que relacionen variables socio-económicas con causas de muerte en el país, entonces, es necesario realizar una investigación que nos permita modelar estas relaciones aplicando métodos de aprendizaje automático.

2.3. Aprendizaje automático

El aprendizaje automático es un subcampo de las ciencias computacionales que tiene que ver con el desarrollo de algoritmos, los cuales para ser útiles, se basan en un conjunto de ejemplos de algún fenómeno. Estos ejemplos pueden ser tomados de la naturaleza o ser generados por el ser humano u algún otro algoritmo [4].
Además, el aprendizaje automático también puede ser definido como el proceso de resolver un problema práctico siguiendo el enfoque siguiente:

1. Recolectar un conjunto de datos
2. De forma algorítmica, construir un modelo basado en ese conjunto de datos
3. Seleccionar el algoritmo adecuado al problema del área de aprendizaje automático.
4. Analizar y dar significado a los resultados obtenidos.

El modelo de aprendizaje tiene el propósito de resolver el problema práctico que generó los datos del primer paso. En particular, los algoritmos de aprendizaje automático pueden ser divididos en supervisado y no supervisado.

En el aprendizaje supervisado, el conjunto de datos es una muestra de ejemplos etiquetados $\{x_i, y_i\}_i^N$, cada elemento x_i dentro de N, se conoce como un vector característica. Un vector característica, es un vector en el cual cada dimensión $j = 1...D$ contiene un valor que describe las muestras ejemplo de alguna manera. La meta de un algoritmo supervisado es usar el conjunto de datos para producir un modelo, el cual tome un vector característica x como entrada junto con las salidas de la información, tal que pueda deducir la etiqueta para éste vector característica.

En el aprendizaje no supervisado, el conjunto de datos es una colección de ejemplos no etiquetados $\{x_i\}_i^N$. De nuevo, x es un vector característica y la meta de un algoritmo no supervisado es crear un modelo que tome un vector característica x como entrada y lo transforme en otro vector o en un valor, que pueda ser usado para resolver un problema práctico.

2.4. Redes neuronales artificiales

Una red neuronal artificial (ANN), se puede definir como un sistema de cómputo formado por un número de elementos simples de procesamiento altamente interconectados, el cual procesa información mediante sus estados de respuesta dinámica a entradas externas [11].

Otra definición es la que dice que las redes neuronales son dispositivos de procesamiento (algoritmos o hardware) que están modelados siguiendo la estructura del cortex cerebral de un mamífero pero a mucho menor escala. Una red neuronal grande puede tener cientos de miles de unidades de procesamiento mientras que el de un mamífero tiene billones de neuronas con un correspondiente incremento en la magnitud de su interacción completa y su comportamiento emergente [25].

Las redes neuronales son típicamente organizadas en capas; las capas están formadas por un número de nodos interconectados los cuales contienen una función de activación. Los patrones son presentados a la red via la capa de entrada, la cual se comunica con una o mas capas ocultas, dónde el procesamiento se hace por un sistema de conexiones con pesos. Las capas ocultas entonces quedan enlazadas a las capas de salida [32]. En la Figura **2-1**, se puede observar la imagen de una red neuronal simple con sólo dos capas de neuronas ocultas.

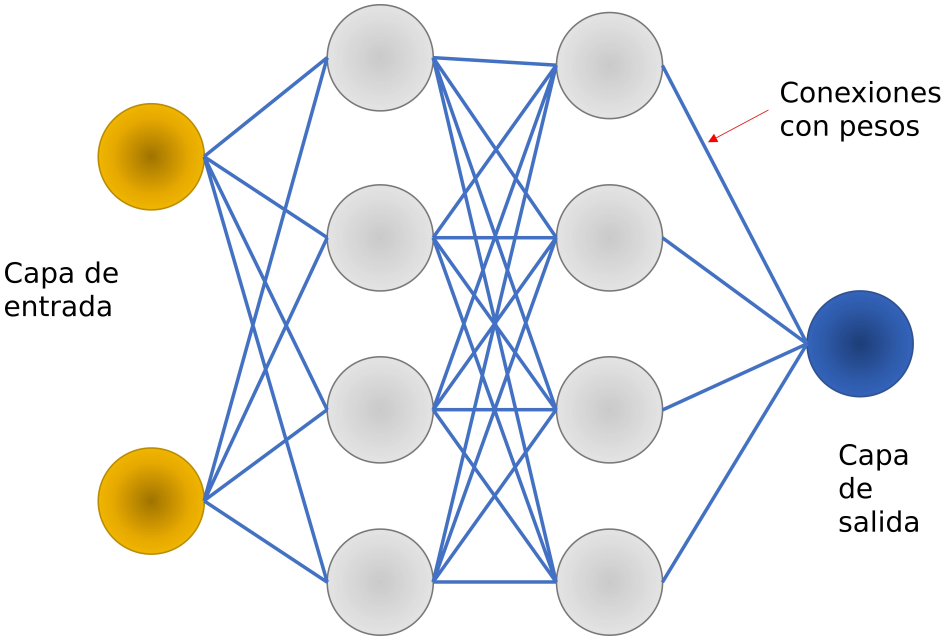

Figura 2-1.: Red Neuronal básica.

2.4.1. Redes Neuronales Recurrentes

Las redes neuronales recurrentes (RNN) son un tipo de arquitectura de red neuronal artificial diseñada para trabajar con datos secuenciales o datos que tienen una estructura temporal. A diferencia de las redes neuronales feedforward convencionales, las RNN tienen conexiones retroalimentadas, lo que les permite mantener y utilizar información sobre estados anteriores en la secuencia [17].

La principal ventaja de las RNN es su capacidad para capturar dependencias a largo plazo en datos secuenciales debido a su estructura recurrente. Sin embargo, las RNN también tienen limitaciones, como el problema de desvanecimiento y explosión del gradiente, que pueden dificultar el entrenamiento de modelos efectivos para secuencias muy largas. Para abordar estas limitaciones, han surgido variantes de RNN, como las Long Short-Term Memory (LSTM) y las Gated Recurrent Unit (GRU), que han demostrado ser más efectivas en la captura de dependencias a largo plazo.

Las conexiones recurrentes pueden mejorar el rendimiento de las redes neuronales apalancando su habilidad para entender dependencias secuenciales. Sin embargo, la memoria producida de las conexiones recurrentes puede ser severamente limitada por los algoritmos de entrenamiento de las redes neuronales recurrentes. Las redes neuronales son específicamente diseñadas para evitar este problema conocido como explosión de gradiente o desvanecimiento de gradiente. Las redes LSTM tienen una celda de memoria interna que puede mantener información durante largos periodos de tiempo. Esta celda de memoria es lo que le permite capturar dependencias a largo plazo en los datos secuenciales [36].

Las LSTMs tienen tres puertas principales que regulan el flujo de información en la celda de memoria: la puerta de olvido (forget gate), la puerta de entrada (input gate) y la puerta de salida (output gate). Estas unidades reciben las señales de activación de la celda de diferentes fuentes y controlan la activación de la celda por multiplicadores. Las redes LSTM propagan los errores por mucho más tiempo que las redes neuronales recurrentes ordinarias.
En la Figura **2-2** se puede ver la representación de una celda LSTM con sus respectivas puertas de entrada, salida y olvido.

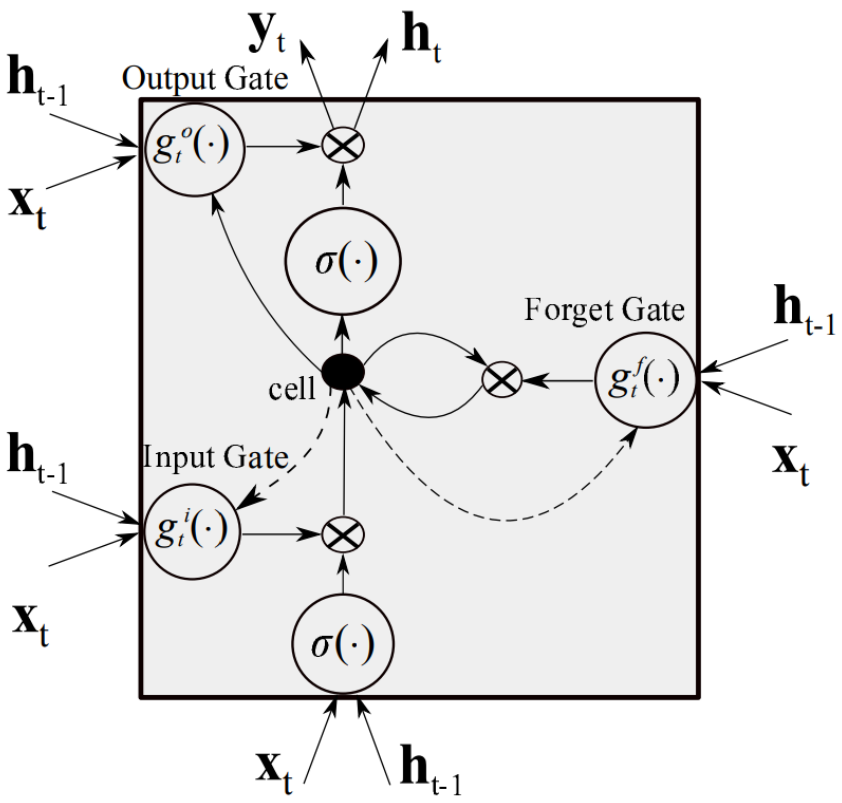

Figura 2-2.: Representación de una celda LSTM (imagen tomada de [36])

Las redes LSTM destacan en tareas como:

- Modelado de lenguaje. Por ejemplo, para predecir la probabilidad de que ocurra una palabra dada una secuencia de palabras anteriores dentro de un texto.
- Traducción automática de un idioma a otro.
- Reconocimiento de voz, es decir, convertir el habla a texto.
- Producción de series temporales. Por ejemplo, para predecir pronósticos climáticos o precios en el mercado de acciones.

2.4.2. Redes Neuronales Residuales

Las redes neuronales residuales, comúnmente conocidas como ResNets, son un tipo de arquitectura de redes neuronales profundas utilizadas en el campo del aprendizaje profundo y la visión por computadora. Fueron introducidas por Kaiming He, Xiangyu Zhang, Shaoqing Ren y Jian Sun en un artículo titulado "Deep Residual Learning for Image Recognition" [13].

A diferencia de las redes neuronales comunes (feedforward), las ResNets se caracterizan por su estructura residual o saltos de conexiones que directamente agregan cada salida de capa a la siguiente salida de capa. En una ResNet, las capas de la red incluyen conexiones directas (skip connections) que saltan una o más capas, lo que permite que la información fluya más fácilmente a través de la red. Estas conexiones directas se suman a la salida de las capas posteriores, lo que permite que la información original se mantenga sin cambios en alguna medida, evitando así el problema del "desvanecimiento de gradientes" que puede dificultar el entrenamiento de redes neuronales muy profundas [44].

En la Figura **2-3**, se puede observar el bloque básico de construcción de una red neuronal residual o ResNet. Como se observa, la primer salida de una red neuronal se suma a la siguiente capa; por esto se conoce como conexión salto o skip.

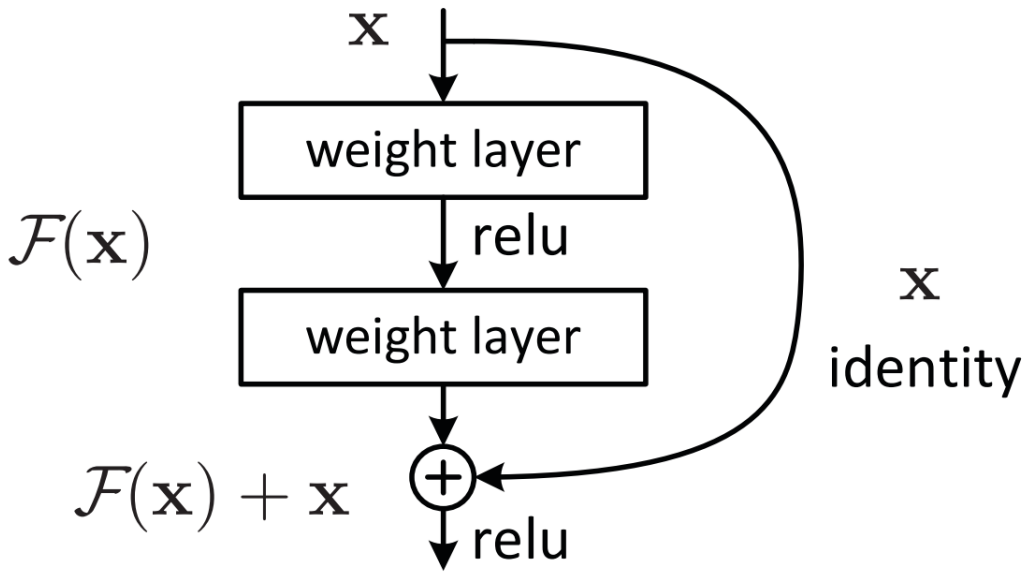

Figura 2-3.: Representación de un bloque simple en una ResNeT (imagen tomada de [13])

2.5. Máquinas de vector soporte

Las Máquinas de Vector Soporte (SVM) son modelos de aprendizaje supervisado con algoritmos que analizan datos y que son usados para el análisis de regresión y clasificación. Un modelo SVM es una representación de las muestras como puntos en el espacio, mapeado de tal manera que las muestras de diferentes categorías sean divididas por una clara brecha, la cual debe ser lo más amplia posible. Sin embargo, cuando las muestras de los datos no están etiquetadas, se intenta de forma natural, realizar agrupamientos de los subconjuntos de datos y entonces, se mapean los nuevos datos a estos grupos recién formados [25].
En la Figura **2-4** se muestra la representación esquemática de una máquina de vector soporte usada en la clasificación de dos categorías. Los puntos rojos pertenecen a la categoría A mientras que los azules a la categoría B. LA SVM determina un hyperplano que nos permite clasificar (en este caso, separar en el plano) adecuadamente las dos categorías de datos.

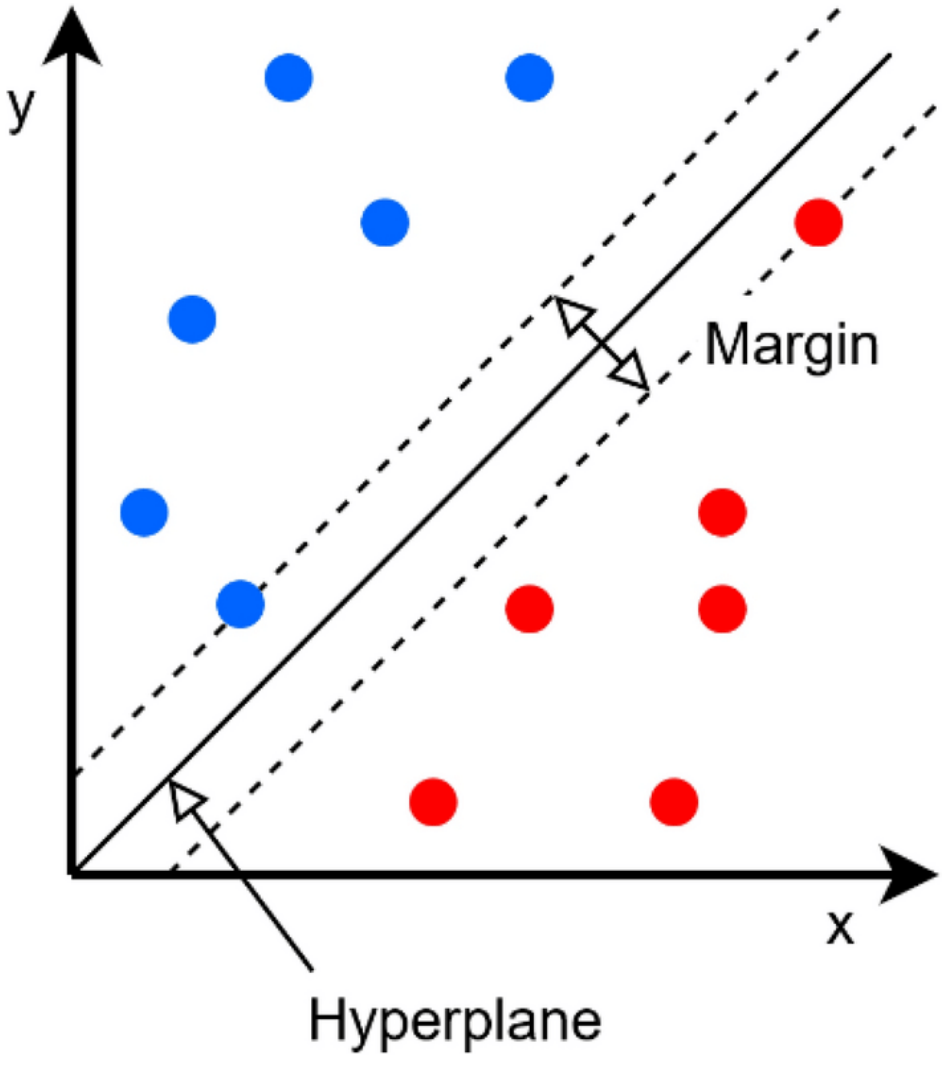

Figura 2-4.: Ejemplo de SVM lineal para dos dimensiones (imagen de [2])

2.5.1. SVM multiclase

Las Support Vector Machines (SVM) multiclase son una extensión de las SVM tradicionales diseñadas para abordar problemas de clasificación con más de dos clases. Mientras que, las SVM binarias se utilizan para separar datos en dos clases, las SVM multiclase permiten clasificar datos en una de múltiples clases [24]. Hay varias formas de implementar SVM multiclase:

Uno contra todos o uno contra resto. En este enfoque se crea una clase binaria separada para cada clase, por ejemplo:

- SVM (A vs no A)
- SVM (B vs no B)
- SVM (C vs no C)

De lo anterior, se elige la SVM que provee el mayor margen de separación entre las muestras de datos.

Uno contra uno. Se crea una SVM binaria para cada par de clases. Por ejemplo, para tres clases A,B y C se obtienen tres SVMs binarias:

- SVM (A vs B)
- SVM (A vs C)
- SVM (B vs C)

Durante la predicción cada SVM binaria vota por una de las clases y la más votada se selecciona como la predicción final.

SVM multiclase sobre características. Una solución más elegante es construir un clasificador de dos clases sobre un vector de características derivado de la tupla característica y clase. En la fase de prueba el clasificador elige el argumento máximo entre estas clases y en la fase de entrenamiento se comparan las diferencias de los márgenes entre las clases más cercanas.

2.5.2. SVM online

Una SVM online es una variante de las tradicionales SVMs que está diseñada para manejar flujos de datos(streaming) o datos en linea, donde los datos van llegando secuencialmente en el tiempo. Son particularmente útiles en situaciones en las que se necesita actualizar el modelo SVM mientras llegan nuevos datos y sin tener que reentrenar el modelo desde cero. En [21] los autores proponen un método de selección de modelo de aprendizaje incremental para máquinas de vector soporte utilizando una validación cruzada que permite el ajuste de los hiperparámetros de la SVM según se adquieren nuevas muestras de datos. En este artículo se obtuvieron buenos resultados en un conjunto de pruebas sobre un clasificador de correos electrónicos y en otro conjunto sobre clasificación de situaciones en video.

2.6. Metricas de evaluación para ANN y SVM

En esta sección se presentan dos de las métricas más utilizadas para validar los modelos de regresión, ya sean usando redes neuronales artificiales, o bien las máquinas de vector soporte.

2.6.1. Error absoluto medio (MAE)

Existen diferente métricas utilizadas para evaluar modelos de regresión y las más comunes son el error absoluto medio (MAE) y el error cuadrático medio (MSE) [20]. En el caso del error absoluto medio (MAE), el error de la regresión se calcula como un promedio de las diferencias absolutas entre los valores objetivo y las predicciones. El MAE es una puntuación lineal, lo que significa que todas las diferencias individuales se ponderan por igual en el promedio. En (2-1), se muestra su definición.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2-1}$$

donde:

- y es el valor objetivo.
- $\hat{y}$ es el valor predicho.
- i es el índice de la muestra actual.
- n es el número de predicciones.

Cabe resaltar que ésta métrica penaliza errores enormes de mejor forma que el caso de la métrica error cuadrático medio. Por lo tanto, no es tan sensible a los valores atípicos como para el error cuadrático medio [38].

2.6.2. Error cuadrático medio (MSE)

El error cuadrático medio (MSE), básicamente mide el error cuadrado promedio de las predicciones. Para cada punto, calcula la diferencia cuadrada entre las predicciones y el objetivo; luego promedia esos valores [20]. En (2-2), se muestra su definición.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{2-2}$$

donde:

- y es el valor objetivo.

- $\hat{y}$ es el valor predicho por el modelo.
- i es el índice para la muestra actual.
- n es el número de predicciones realizadas.

Cuanto mayor sea el valor de MSE, se considera que es peor el modelo. El valor MSE nunca es negativo, ya que los errores de predicción individuales se elevan al cuadrado antes de ser sumados. Aúnque un resultado de cero para MSE indicaría un modelo perfecto.
Si se hace una predicción muy errónea, la cuadratura empeorará aún más el error y puede sesgar la métrica para sobreestimar la deficiencia del modelo. Éste es un comportamiento particularmente problemático cuando se utilizan datos ruidosos [38].

2.7. Algoritmo ID3 (Iterative Dichotomiser 3)

El algoritmo ID3 es un algoritmo inventado por Ross Quinlan que sirve para generar un árbol de decisión a partir de un conjunto de datos. El algoritmo ID3 aprende árboles de decisión construyendo de forma top-down comenzando por la pregunta ¿cuál atributo debería ser probado en la raíz del árbol? Para responder a esta pregunta, cada atributo de instancia se evalúa utilizando una prueba estadística para determinar que tan bien clasifica las muestras del conjunto de entrenamiento. El mejor atributo es seleccionado y usado como la prueba para el nodo raíz del árbol. Entonces, se continúa creando un descendiente del nodo raíz para cada posible valor de sus atributos, y las muestras de entrenamiento son ordenadas respecto al nodo descendente apropiado (bajo la rama correspondiente al valor de prueba de este atributo). El proceso completo se repite usando las muestras de prueba asociadas con cada nodo descendiente para seleccionar el mejor atributo en ese punto del árbol (nodo).
Esta es una forma de búsqueda ávida, aceptable para un árbol de decisión en el cual el algoritmo nunca hace retrocesos (backtracking) para considerar mejores opciones [26].
El algoritmo ID3 utiliza la ganancia de información para seleccionar el mejor atributo en cada paso del creciente árbol de decisión. En la ecuación (2-3) se muestra el cálculo de la entropia, mientras que en la ecuación (2-4) la de la ganancia de información.

$$entropia(T) = \Sigma_{i=1}^{c} - p_i \log_2 p_i \tag{2-3}$$

donde:

- T es una colección de objetos
- i son las posibles respuestas de los objetos
- c es el número de clases

- p_i es la probabilidad de los posibles valores

$$ganancia(T, X) = entropia(T) - \Sigma_{i=1}^{n} \frac{|T_i|}{|T|} entropia(T_i) \quad (2\text{-}4)$$

donde:

- T es una colección de objetos
- T_i es tamaño del subconjunto de datos despues de dividirlo por el atributo i
- X es el dato de prueba actual
- n número de subconjuntos en el que se divide T

Este modelo es un algoritmo de aprendizaje inductivo, es decir, busca establecer leyes o principios generales sobre la base de la observación de varios o todos los componentes de un conjunto o clase.
El algoritmo ID3 sigue una política en la cual selecciona una hipótesis sobre otra basándose en elegir el primer árbol aceptable que encuentre usando su búsqueda de simple a complejo dentro del espacio de árboles posibles. En otras palabras, la estrategia de búsqueda del algoritmo ID3 se puede descomponer en dos partes:

- Tiene preferencia por los árboles más cortos.
- Selecciona los árboles con mayor ganancia de información cercanos a la raíz.

Algunas características del algoritmo ID3 se describen a continuación:

- ID3 solo funciona con atributos categóricos.
- Puede generar problemas cuando hay valores faltantes en el conjunto de datos.
- Puede tener sesgo hacia atributos con un mayor número de valores variables.
- Puede sobreajustarse a los datos de entrenamiento (overfitting).

2.8. Algoritmo C4.5

El algoritmo C4.5 fue propuesto por Ross Quinlan en 1993 como una mejora del algoritmo ID3. La idea principal detrás de C4.5 es construir un árbol de decisión que divida el conjunto de datos en subconjuntos más pequeños y homogéneos con respecto a la variable objetivo (la variable que se desea predecir). El algoritmo utiliza medidas de impureza, como la ganancia

de información, para determinar qué atributo y valor de atributo son los mejores para realizar una división en cada nodo del árbol. Luego, se repite este proceso de manera recursiva para construir el árbol completo [31].
El algoritmo C4.5 se basa en la utilización del criterio ratio de ganancia como se ve en la ecuación (2-6). Se trata de un cociente entre la ganancia de información, previamente definida en la ecuación (2-4), y la cantidad de información definida en la ecuación (2-5).

$$cantidad\ informacion(T, X) = -\Sigma_{i=1}^{n} \frac{|T_i|}{|T|} \log_2 \frac{|Ti|}{|T|} \tag{2-5}$$

donde:

- T_i es el tamaño del subconjunto de datos después de dividirlo por el atributo i
- T es una colección de objetos
- X es el dato de prueba actual
- n número de subconjuntos en el que se divide T

$$ratio\ ganancia(T, X) = \frac{ganancia(T, X)}{cantidad\ informacion(T, X)} \tag{2-6}$$

Además, el algoritmo incorpora una poda de árbol de clasificación. La poda está basada en la aplicación de una prueba de hipótesis que trata de responder la interrogante de expandir o no una rama en el árbol de decisión [27].
Algunas características del algoritmo C4.5 son descritas a continuación:

- C4.5 puede utilizar atributos categóricos y continuos.
- Puede tratar con valores faltantes a la hora de calcular la ganancia de información.
- Utiliza la ganancia de información normalizada conocida como ratio de ganancia. Esto corrige el sesgo de ID3 hacia atributos con más valores posibles.
- Al realizar una poda posterior reduce el sobreajuste (overfitting).

2.9. Algoritmo de árbol de decisión CART

Los árboles de decisión o de clasificación son un modelo surgido en el ámbito del aprendizaje automático (Machine Learning) y de la inteligencia artificial que, partiendo de una base de datos, crea diagramas de construcciones lógicas que nos ayudan a resolver problemas. A

esta técnica también se la denomina segmentación jerárquica. Es una técnica explicativa y descomposicional que utiliza un proceso de división secuencial, iterativo y descendente que partiendo de una variable dependiente, forma grupos homogéneos definidos específicamente mediante combinaciones de variables independientes en las que se incluyen la totalidad de los casos recogidos en la muestra [34].
El algoritmo CART es el acrónimo de Árboles de Clasificación y Regresión diseñado por Breiman et. al. en 1984. Con este algoritmo se generan árboles de decisión binarios u multinarios, es decir, los nodos de un árbol CART pueden tener más de dos hijos, lo que permite modelar relaciones más complejas en los datos. Este modelo admite variables de entrada y salida nominales, ordinales y continuas, por lo que se pueden resolver tanto problemas de clasificación como de regresión.
CART es muy similar al algoritmo C4.5 con la diferencia de que soporta variables objetivo con valores numéricos (regresión). CART construye árboles binarios usando la característica y umbral que aporta la mayor ganancia de información en cada nodo [39]. El algoritmo utiliza el índice de Gini para calcular la medida de impureza, su ecuación se muestra en (2-7).

$$G(A_i) = \Sigma_{j=i}^{M_i} p(a_{ij}) G(C/A_{ij}) \tag{2-7}$$

Siendo:

$$G(A_{ij}) = -\Sigma_{j=i}^{M_i} p(C_k/A_{ij})(1 - p(C_k/A_{ij})) \tag{2-8}$$

donde:

- A_{ij} es el atributo empleado para ramificar el árbol.
- j es el número de clases.
- M_i es el conjunto de valores distintos que tiene el atributo A_i.
- $p(A_{ij})$ constituye la probabilidad de que A_i tome su j-ésimo valor.
- $p(C_k/A_{ij})$ representa la probabilidad de que un ejemplo sea de la clase C_k cuando su atributo A_i toma su j-ésimo valor.

Loas algoritmos de clasificación mediante árboles de decisión han resultado ser métodos robustos para correlacionar variables - atributos con valores de clase, reportando además una jerarquía de significancia sobre los valores atributos para determinar los valores de clase.

2.10. Coeficiente de Pearson

El coeficiente de correlación de Pearson se define como la medida de la fuerza de la relación entre dos variables y su asociación con cada una de ellas. En otras palabras, el coeficiente de correlación de Pearson calcula el efecto del cambio en una varible cuando la otra cambia [16].
El coeficiente de correlación de Pearson tiene una gran importancia estadística; busca trazar una línea a través de los datos de dos variables para mostrar su relación. Tal relación lineal puede ser positiva o negativa.
La fórmula para calcular este coeficiente se muestra en (2-9):

$$r_{xy} = \frac{\Sigma_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\Sigma_{i=1}^{n}(y_i - \bar{y})^2}} \tag{2-9}$$

donde:

- n es el tamaño de la muestra.
- x_i, y_i son las muestras de los puntos indexados con i.
- $\bar{x}$ es la media aritmética simple definida como $\bar{x} = (1/n)\Sigma_{i=1}^{n} x_i$, analogamente para $\bar{y}$.

El coeficiente obtenido estará dentro del intervalo $[-1, 1]$, con las siguientes interpretaciones:

- Si $r = 1$, se dice que hay una perfecta correlación positiva.
- Si $0 < r < 1$, existe una correlación positiva.
- Si $r = 0$, no existe ninguna correlación.
- Si $-1 < r < 0$, existe una correlación negativa.
- Si $r = -1$, se dice que hay una perfecta correlación negativa. A veces, es nombrada como relación inversamente proporcional.

2.11. Análisis de regresión

El análisis de regresión es una técnica estadística que permite comprobar la hipótesis de que una variable depende de una u otras variables. Además, el análisis de regresión brinda una estimación de la magnitud del impacto de un cambio en una variable sobre otra. Por supuesto, esta última característica es de vital importancia para predecir los valores futuros [15].

2.11.1. Cuadrados mínimos generalizados (GLS)

Es una técnica que tiene como objetivo modelar la relación entre variables dependientes (resultado) y las variables independientes (variables predictoras). El análisis por cuadrados mínimos generalizados (GLS), es un tipo de análisis de regresión que toma en cuenta la heterocedasticidad (variaciones desuniformes en los datos) y las correlaciones de los mismos [40].

En el análisis de regresión los valores p se usan para evaluar la significancia estadística de las relaciones entre variables independientes (predictoras) y dependientes (resultado). Si el valor p es pequeño el modelo sugiere que hay suficiente evidencia para concluir que la variable independiente tiene una relación estadística significativa con la variable dependiente. Por otro lado, cuando el valor es muy grande no existe relación significativa entre las variables [19].

3. Metodología

En este capítulo, primero se presentan los datos obtenidos de las diferentes fuentes disponibles y después se presentan las variables seleccionadas para cada herramienta de aprendizaje automático.

3.1. Conjunto de datos

Para comenzar con la investigación, fue necesaria la búsqueda y recolección de los datos disponibles en los repositorios abiertos del INEGI(Instituto Nacional de Estadística y Geografía), el CONEVAL (Consejo Nacional de Evaluación de la Política de Desarrollo Social) y el CONAPO (Consejo Nacional de Población) [1]. Sin embargo, esta tarea no fue fácil, ya que en México se registran muy pocas variables socioeconómicas y las que se registran son por periodos muy cortos. A pesar de esto, se consiguió generar un conjunto de datos con 15,233,229 registros útiles y 103,367 registros incompletos. Los datos obtenidos permitieron considerar para este estudio, las siguientes variables:

- Empleo. El número de personas con ocupación, obtenido del INEGI (encuesta nacional de ocupación y empleo).
- Seguridad. La incidencia delictiva, obtenido del INEGI (encuesta nacional de victimización y percepción sobre seguridad pública).
- Pobreza. El número de personas en situación de pobreza de acuerdo al CONEVAL.
- Internet. El número de personas con acceso a internet, resultado de los censos y proyecciones. Obtenido del INEGI (encuesta nacional sobre disponibilidad y uso de tecnologías de información).
- PIB. El Producto Interno Bruto. Obtenido de INEGI (Cuentas de bienes y servicios).
- Acceso-Salud. El número de personas que cuentan con una afiliación a los servicios de atención médica. Obtenido de INEGI (Sistema de cuentas nacionales de México, SALUD).

[1] INEGI:www.inegi.org.mx CONEVAL: www.coneval.org.mx CONAPO: www.gob.mx/conapo

- Población. Registro de la población del país a lo largo del periodo 2010-2021. Obtenido de CONAPO (Población y proyecciones).

Para este estudio también fue necesario obtener los registros de mortalidad general en el periodo disponible (2010-2021). De estos registros, los datos se clasificaron de acuerdo a la lista mexicana de enfermedades, con el fin de obtener las causas de muertes principales dentro del mismo periodo. De esta clasificación se obtienen las siguientes causas:

- Corazón. Enfermedades del corazón.
- Diabetes. Diabetes Mellitus.
- Tumores. Aquellos tumores evaluados como malignos.
- Influenza. Relacionadas con la influenza y neumonía.
- Hígado. Enfermedades relacionadas al hígado.
- Cerebrov. Enfermedades cerebrovasculares.
- Agresión. Defunciones derivadas de una agresión.
- Accidente. Defunciones derivadas de un accidente.
- Pulmonar. Enfermedades pulmonares obstructivas crónicas.
- I.Renal. Enfermedades relacionadas con la insuficiencia renal.
- Perinatal. Relacionada con ciertas afecciones en el periodo perinatal.
- Suicidio. Derivado de lesiones autoinfligidas intencionalmente.

En algunos casos, como los datos de la pobreza, fue necesario aplicar un polinomio de interpolación, ya que no se contaba con información anual disponible. Esto es debido a que las instituciones no realizan publicaciones de datos de esa manera. Para esta tarea se utilizó regresión polinomial. En la Tabla **3-1**, se observan en la segunda columna, los datos originales mientras que en la tercera los resultados obtenidos. El polinomio se muestra en la ecuación (3-1).

$$y = -0,9816x^5 + 14{,}074x^4 - 68{,}855x^3 + 130{,}05x^2 - 68{,}914x + 528{,}13 \tag{3-1}$$

Para trabajar con los datos se utilizó una normalización con respecto de la población total anual; esto se hizo con el fin de trabajar dentro de un mismo espacio muestral. Por ejemplo, en la ecuación (3-2), se muestra el cálculo para la variable empleo en el año 2010. Esta normalización se realizó para las variables de empleo, seguridad, pobreza, internet, PIB y acceso a salud. Esto se puede observar en la Tabla **3-2**.

Tabla 3-1.: Resultados obtenidos al realizar la interpolación

AÑO	POBREZA(100K) ORIGINAL	POBREZA (100K) eq. (3-1)
2010	528.13	528.13
2011	-	531.05
2012	533.50	533.50
2013	-	548.78
2014	553.42	553.42
2015	-	546.70
2016	534.18	534.18
2017	-	524.31
2018	524.26	524.26
2019	-	537.00
2020	556.54	556.54

$$\frac{PoblacionOcupadaEn2010(100k)}{PoblacionTotalEn2010(100k)} * 100 = EmpleoNormalizadoEn2010 \tag{3-2}$$

Tabla 3-2.: Variables consideradas para el análisis (por 100k hab.)

AÑO	EMPLEO	SEGURIDAD	POBREZA	INTERNET	PIB	SALUD	POBLACIÓN
2010	40.55	26.84	46.43	28.84	13141.03	60.29	1137.49
2011	40.86	25.31	46.03	32.61	13431.34	61.84	1153.67
2012	41.79	30.05	45.62	34.99	13733.77	63.32	1169.36
2013	41.56	35.09	46.33	38.86	1374.36	64.74	1184.54
2014	41.20	34.73	46.14	39.56	13958.32	66.08	1199.36
2015	41.71	29.25	45.05	50.56	14250.22	67.37	1213.48
2016	41.45	30.16	43.53	52.45	14462.16	68.60	1227.15
2017	41.66	31.74	42.27	56.67	14609.78	69.77	1240.42
2018	42.42	30.17	41.84	58.36	14777.26	70.89	1253.28
2019	43.15	26.59	42.43	62.80	14602.21	71.95	1265.78
2020	41.47	23.95	43.56	64.93	13298.12	72.96	1277.92
2021	42.77	23.87	43.86	68.67	13980.21	73.92	1289.72

Para identificar algunas relaciones entre las diferentes variables y causas de defunción se siguieron dos perspectivas. La primera es utilizar el coeficiente de correlación de Pearson y la segunda utilizar un método de análisis de regresión. Ambos métodos se usan con la finalidad de obtener conjuntos variables-causas para ser alimentados a máquinas de vector soporte (SVM) y a redes neuronales artificiales. Por otro lado, para el método de CART, se siguió otra perspectiva que se detalla en la sección 3.1.3.

3.1.1. Método de Pearson

Para correlacionar variables se utilizó el coeficiente de Pearson, el cual se implementó en Python versión 3.9 utilizando la librería statsmodels 0.14.0 [41]. De este análisis se obtuvo la Tabla **3-3**.

Tabla 3-3.: Coeficientes de Pearson

CAUSA\VAR	EMPLEO	SEGURIDAD	POBREZA	INTERNET	PIB	ACCESO-SALUD
CORAZON	0.51	-0.58	-0.51	0.85	0.03	0.85
DIABETES	0.33	-0.51	-0.44	0.79	-0.05	0.79
TUMOR	0.68	-0.36	-0.82	0.97	0.46	0.97
INFLUENZA	0.45	-0.60	-0.44	0.79	-0.07	0.79
HIGADO	0.65	-0.39	-0.82	0.95	0.44	0.94
CEREBROV.	0.18	-0.40	-0.47	0.70	0.03	0.67
AGRESION	0.55	-0.62	-0.71	0.63	0.16	0.60
ACCIDENTE	-0.71	0.34	0.74	-0.92	-0.38	-0.95
PULMONAR	0.09	0.29	-0.54	0.20	0.56	0.18
I.RENAL	0.41	-0.51	-0.55	0.74	0.10	0.73
PERINATAL	-0.68	0.32	0.68	-0.95	-0.35	-0.97
SUICIDIO	.64	-0.46	-0.55	0.90	0.22	0.91

Tabla 3-4.: Correlación (Pearson) de variables seleccionadas y sus causas

CAUSA\VAR	EMPLEO	SEGURIDAD	POBREZA	INTERNET	PIB	ACCESO-SALUD
CORAZON	0	1	0	1	0	1
DIABETES	0	1	0	1	0	1
TUMOR	1	0	0	1	0	1
INFLUENZA	0	1	0	1	0	1
HIGADO	0	0	1	1	0	1
CEREBROV.	0	0	1	1	0	1
AGRESION	0	1	1	1	0	0
ACCIDENTE	0	0	1	1	0	1
PULMONAR	0	0	1	0	1	0
I.RENAL	0	1	0	1	1	1
PERINATAL	1	0	0	1	0	1
SUICIDIO	0	1	1	0	1	0

De estos valores, se seleccionaron aquellos con un coeficiente absoluto mayor a 0.5, así se obtienen las variables de mayor relevancia. Finalmente, éstas variables son las que serán utilizadas para alimentar los algoritmos de redes neuronales y de las máquinas de vector soporte.

En la Tabla **3-4**, se muestran las variables seleccionadas con sus causas, un 1 significa que la variable es significativa para esa causa y un 0 que no.

3.1.2. Método de cuadrados mínimos generalizados

En el análisis de regresión se puede empezar por una hipótesis nula, es decir, que no hay relación entre una variable dependiente e independiente. El valor p nos dice la probabilidad de observar que en los datos la hipótesis nula es verdadera. Si la probabilidad de que la hipótesis nula sea verdadera es muy pequeña, se puede rechazar y por lo tanto, concluir que la variable es significativa.
Tomando los mismos datos que para la correlación con el coeficiente de Pearson, se realizó un análisis de regresión con método de mínimos cuadrados generalizados utilizando Python version 3.9 y la librería statsmodels 0.14.0. De éste análisis se seleccionaron aquellos con valor $p < 0.05$. Los valores p se muestran en la Tabla **3-5**. En este modelo se observa que hay variables que tienen valores cercanos a 1 por lo que estas no son significativas; por otro lado, hay variables con valores muy cercanos a 0, en estos casos podemos rechazar la hipótesis nula y concluir que se trata de variables significativas en el modelo de regresión.

Tabla 3-5.: Variables y sus causas seleccionadas (GLS)

CAUSA	EMPLEO	SEGURIDAD	POBREZA	PIB	INTERNET	ACCESO-SALUD
CORAZON	0.7420	0.9170	0.8930	0.0190	0.0307	0.0330
DIABETES	0.0370	0.8570	0.2490	0.0440	0.4380	0.0138
TUMORES	0.8130	0.0057	0.7200	0.9800	0.0069	0.0340
INFLUENZA	0.8800	0.5940	0.0055	0.0050	0.7180	0.0390
HIGADO	0.9060	0.9350	0.0053	0.0064	0.0068	0.0731
CEREBROV.	0.0170	0.9250	0.0388	0.8140	0.0298	0.7440
AGRESION	0.0040	0.2900	0.0001	0.001	0.5770	0.8440
ACCIDENTE	0.2370	0.6110	0.0085	0.0110	0.2710	0.0071
PULMONAR	0.0069	0.9360	0.0069	0.0480	0.8460	0.8640
I.RENAL	0.7110	0.0256	0.7230	0.7020	0.0348	0.0201
PERINATAL	0.7630	0.0334	0.6110	0.0016	0.3770	0.0307
SUICIDIO	0.9500	0.0498	0.0451	0.6960	0.8430	0.0243

3.1.3. Método árboles de decisiones CART

Para el método de árboles de decisión se utilizaron todas las variables para cada causa de muerte, esto es porque el algoritmo permite discernir, por el mismo, la sigificancia de cada una de las variables consideradas en las tablas, basándose en el coeficiente de Gini. Este coeficiente se explicó en la ecuación (2-7). Además, se agregaron datos relacionados a la edad y el género con cada causa de muerte. Así, los nuevos atributos agregados son:

- H. Género hombre.
- M. Género mujer.

- L15. Porcentaje de personas con causa de muerte seleccionada dentro del rango de edad menor a 15.
- 15-24. Porcentaje de personas con causa de muerte seleccionada dentro del rango de edad 15 a 24 años.
- 25-34. Porcentaje de personas con causa de muerte seleccionada dentro del rango de edad 25 a 34 años.
- 35-44. Porcentaje de personas con causa de muerte seleccionada dentro del rango de edad 35 a 44 años.
- 45-64. Porcentaje de personas con causa de muerte seleccionada dentro del rango de edad 45 a 64 años.
- GEQ65. Porcentaje de personas con causa de muerte seleccionada dentro del rango de edad de 65 años o más.

En la Figura **3-1**, se muestra como quedaría, por ejemplo, el encabezado de la tabla para las enfermedades del corazón.

AÑO	VARIABLES						EDAD						GENERO		CORAZON
	SALUD	EMPLEO	SEGURIDAD	POBREZA	INTERNET	PIB	L15	15-24	25-34	35-44	45-64	GEQ65	H	M	

Figura 3-1.: Encabezado para las enfermedades del corazón.

En el siguiente capítulo se verán las implementaciones y resultados obtenidos de la aplicación del algoritmo CART.

4. Implementación y resultados

En este capítulo se describen tanto las implementaciones como resultados obtenidos al aplicar redes neuronales, máquinas de vector soporte y los árboles de decisión sobre las tablas construidas en base a los datos sobre la mortalidad en México y diferentes variables socioeconómicas, también de México y durante los años de 2010 a 2021.

4.1. Método de Pearson y mínimos cuadrados generalizados (GLS)

En la sección anterior se comentó que mediante el coeficiente de Pearson y el análisis de regresión por cuadrados mínimos generalizados se logró reconocer la mayor significancia entre el conjunto de variables y las causas de muerte.
El conjunto conteniendo las variables más significativas será utilizado como entrada de alimentación a una red neuronal, así también alimentará la entrada a un algoritmo que implementa una máquina de vector soporte.
En la sección 4.1.1 se muestran los detalles de implementación de las redes neuronales, mientras que en la sección 4.1.2 se muestran los resultados obtenidos de la aplicación de la máquina de vector soporte.

4.1.1. Redes neuronales artificiales

Para la implementación de las redes neuronales artificiales se utilizó el lenguaje de programación Python versión 3.9. Además, se utilizó la librería Sci–kit Learn [3]. En la Tabla **4-1**, se muestra la estructura de la red neuronal utilizada.
De la estructura de la red neuronal artificial es importante resaltar que la capa de normalización toma los valores de entrada y normaliza todos estos datos, lo cual permite una pequeña mejora en la tarea de regresión. Además, para las funciones de evaluación se usaron MAE y MSE.
En la Figura **4-1** se muestran los resultados del entrenamiento para algunas causas de muerte con variables socioeconómicas seleccionadas mediante método de Pearson. El orden de las causas de muerte es el siguiente: en la parte superior izquierda, enfermedades del corazón; en la parte superior derecha, diabetes; en la parte inferior izquierda, tumores; y en la parte inferior derecha influenza. Aquí se puede observar que la red se comporta como es esperado,

Tabla 4-1.: Modelo de la red neuronal artificial implementada

LAYER(TYPE)	OUTPUT SHAPE	PARAM #
NORMALIZATION	(NONE,4)	9
DENSE	(NONE,128)	640
DENSE	(NONE,128)	16512
DENSE	(NONE,128)	16512
DENSE	(NONE,1)	129
TOTAL PARAMS : 33802		
TRAINABLE PARAMS: 33793		
NON-TRAINABLE PARAMS: 9		

es decir, mientras más iteraciones más se reduce el error hasta que se estabiliza en los valores cercanos a cero.

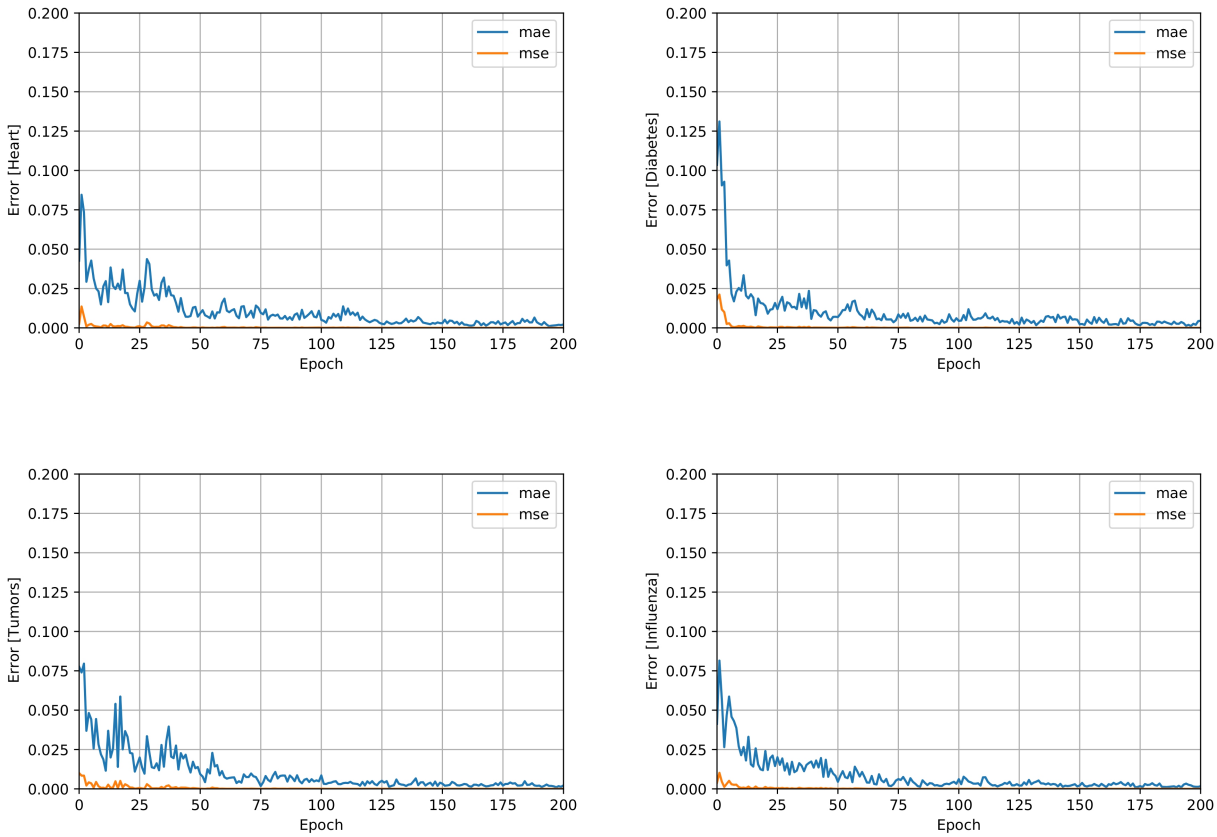

Figura 4-1.: Resultados del entrenamiento ANN para algunas causas de muerte

Por otro lado, en la Figura **4-2** se muestran los resultados obtenidos para algunas causas de

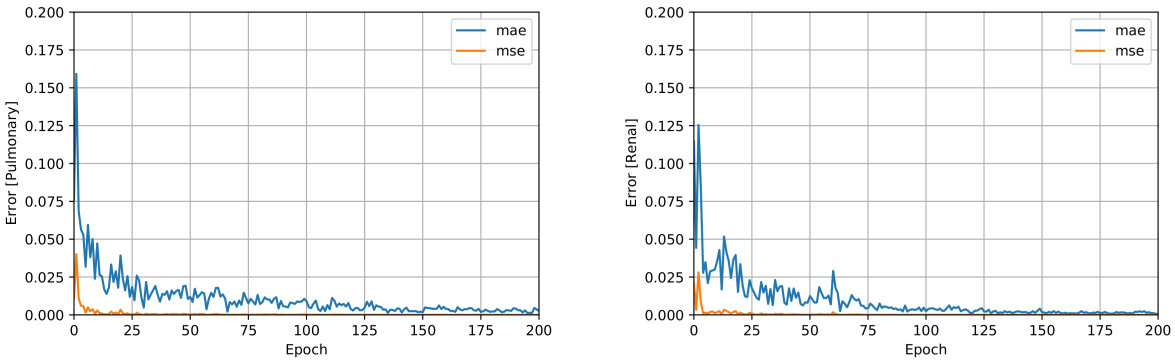

Figura 4-2.: Entrenamiento para enfermedades pulmonares e insuficiencia renal, aplicando GLS como función de evaluación.

muerte, dónde las variables socieconómicas son seleccionadas utilizando análisis GLS.

4.1.2. Máquinas de vector soporte

Las máquinas de vector soporte también se implementaron usando la librería Sci–Kit Learn, se utilizó un kernel polinomial de grado 4, como se muestra en la Figura **4-3**.

```
from matplotlib.units import ma
from sklearn import svm
from sklearn.pipeline import make_pipeline
from sklearn.preprocessing import StandardScaler
RegModel=svm.SVR(kernel='poly',C=1.0,gamma='auto',degree=4,epsilon=0.001)
print(RegModel)
```

```
SVR(degree=4, epsilon=0.001, gamma='auto', kernel='poly')
```

Figura 4-3.: Modelo SVM, grado 4 para el kernel polinomial

Para su entrenamiento, se utilizaron las métricas MAE y MSE. Del entrenamiento se obtuvieron los resultados mostrados en la Tabla **4-2**; en éste caso, la selección de variables corresponde al coeficiente de Pearson.

Tabla 4-2.: Métricas para conjunto de pruebas SVM(Pearson).

CAUSA	MAE	MSE
CORAZON	1.02E-03	1.17E-06
DIABETES	1.15E-03	1.44E-06
TUMORES	6.82E-04	5.17E-07
INFLUENZA	4.39E-04	2.62E-07
HIGADO	3.79E-04	1.83E-07
CEREBROV.	3.23E-04	1.41E-07
AGRESION	9.86E-04	1.22E-06
ACCIDENTE	6.69E-04	6.18E-07
PULMONAR	9.42E-04	1.20E-06
I.RENAL	3.14E-04	1.41E-07
PERINATAL	3.37E-04	1.52E-07
SUICIDIO	4.76E-04	2.67E-07

En contraste, se muestran en la Tabla **4-3**, los resultados obtenidos para las máquinas de vector soporte en las cuales las variables seleccionadas corresponden a las obtenidas mediante el análisis de regresión GLS.

Tabla 4-3.: Metricas para conjunto de pruebas SVM (GLS).

CAUSA	MAE	MSE
CORAZON	2.54E-02	1.13E-05
DIABETES	1.70E-03	4.13E-06
TUMORES	2.34E-03	7.13E-06
INFLUENZA	1.92E-03	4.31E-06
HIGADO	1.71E-03	5.56E-06
CEREBROV.	1.41-03	2.821E-05
AGRESION	1.16E-03	3.21E-05
ACCIDENTE	1.00E-03	5.26E-05
PULMONAR	4.70E-03	2.81E-05
I.RENAL	5.62E-04	2.22E-07
PERINATAL	1.23E-03	4.03E-05
SUICIDIO	3.22E-04	2.56E-05

De los resultados obtenidos se puede ver que los valores son bastantes cercanos a cero, siendo los de la métrica MSE mucho mas pequeños que aquellos pertenecientes a la métrica MAE. Esto se debe a que, para el conjunto de datos analizado, nuestro modelo presenta errores bastante pequeños y no hay valores atípicos en los datos [38]. Dicho de otra manera, la

métrica MSE es grande cuando los errores(diferencia entre valor predicho y valor deseado), varían grandemente entre sí como consecuencia de elevar al cuadrado los errores, tal y como se expresó en la ecuación (2-2).

4.2. Método árboles de decisiones CART

En el caso del algoritmo de árbol de decisión, se vuelven a considerar todas las variables recolectadas para cada causa de muerte. Por ejemplo, en la Tabla **4-4**, se muestran los datos utilizados para las enfermedades del corazón; se puede ver que incluye todas las variables socieconómicas, la edad y el género. Se recomienda regresar a la sección 3.1.3 para ver el significado de las variables que forman el encabezado de la tabla.
La implementación se realiza con la librería Sci–Kit Learn y la librería Graphviz [9] para poder renderizar el grafo.

Tabla 4-4.: Datos utilizados para la causa enfermedades del corazón con el algoritmo CART.

AÑO	SALUD	EMPLEO	SEGURIDAD	POBREZA	INTERNET	PIB	L15	15-24	25-34	35-44	45-64	GEQ65	H	M	CORAZON
10	60.29	40.55	26.84	46.43	28.84	13141.03	0.688	0.755	1.449	2.968	18.42	75.721	52.225	47.775	0.092669
11	61.84	40.86	25.31	46.03	32.61	13431.34	0.006	0.711	1.433	3.056	18.62	75.579	52.477	47.523	0.092009
12	63.32	41.79	30.05	45.62	34.99	13733.77	0.006	0.689	1.338	3.014	18.702	75.614	52.189	47.811	0.094002
13	64.74	41.56	35.09	46.33	38.86	13741.36	0.007	0.676	1.314	2.898	18.597	75.823	52.529	47.471	0.09892
14	66.08	41.2	34.73	46.14	39.56	13958.32	0.004	0.717	1.348	3.00	18.454	76.053	52.738	47.262	0.102058
15	67.37	41.71	29.25	45.05	50.56	14250.22	0.004	0.707	1.391	2.933	18.333	76.243	52.868	47.132	0.107022
16	68.6	41.45	30.16	43.53	52.45	14462.16	0.003	0.659	1.434	3.111	18.64	75.844	53.048	46.952	0.112154
17	69.77	41.66	31.74	42.27	56.67	14609.78	0.003	0.677	1.546	3.061	18.655	75.749	53.15	46.85	0.114228
18	70.89	42.42	30.17	41.84	58.36	14777.26	0.003	0.613	1.473	3.058	18.807	75.771	53.562	46.438	0.119316
19	71.95	43.15	26.59	42.43	62.8	14602.21	0.003	0.591	1.463	3.035	18.864	75.785	53.362	46.638	0.124646
20	72.96	41.47	23.95	43.56	64.93	13298.12	0.002	0.454	1.298	2.928	19.77	75.385	55.584	44.416	0.174225
21	73.92	42.77	23.87	43.86	68.67	13980.21	0.002	0.477	1.302	2.874	19.174	76.003	54.699	45.301	0.174805

En la Figura **4-4** se muestra el árbol de decisión obtenido después de ejecutar el algoritmo CART para la causa de muerte enfermedad de corazón. Como se puede observar, retorna solo las variables con mayor relevancia para ésta causa de muerte.
En cada nodo del árbol a excepción de las hojas, se muestran en la primera línea los criterios de selección. Por ejemplo, para el nodo raíz el criterio es que el valor de la variable H (porcentaje de población de hombres) sea menor o igual a 54,131. A partir de este criterio se crean las siguientes ramas.
Para todos los nodos se presentan los valores para la métrica MSE aunque con el nombre squared_error, debido a que la librería Sci-Kit Learn es la encargada de generar el árbol. En la siguiente línea de cada nodo del árbol, se observa la cantidad de muestras que se utilizaron para derivar los criterios, en el caso del nodo raíz fueron 9 instancias que casaron con el criterio del nodo del árbol.
Finalmente, se muestran los valores que asignará el predictor tomando en cuenta cada criterio. Por ejemplo, para una muestra con valor $H <= 54{,}131$ y $EMPLEO <= 42{,}12$ el valor del modelo será 0,174; un valor cercano al que se tiene para los años 2020 y 2021 en enfermedades del corazón.

Por otro lado, en la Figura **4-5** se observa que la raíz corresponde al criterio $POBREZA <= 44{,}74$. Un caso interesante es el seguido por una muestra que cumpla con los criterios $POBREZA <= 44.74$ y $POBREZA > 43{,}545$ para los cuales el modelo retorna el valor de $0,08$; éste valor de pobreza está dentro del periodo 2016 a 2021 y es el más pequeño que puede retornar el modelo. En la Figura **A-2** del anexo, en la parte inferior izquierda, se observa que la mortalidad en el periodo perinatal tiene una tendencia a la baja a partir del 2016.

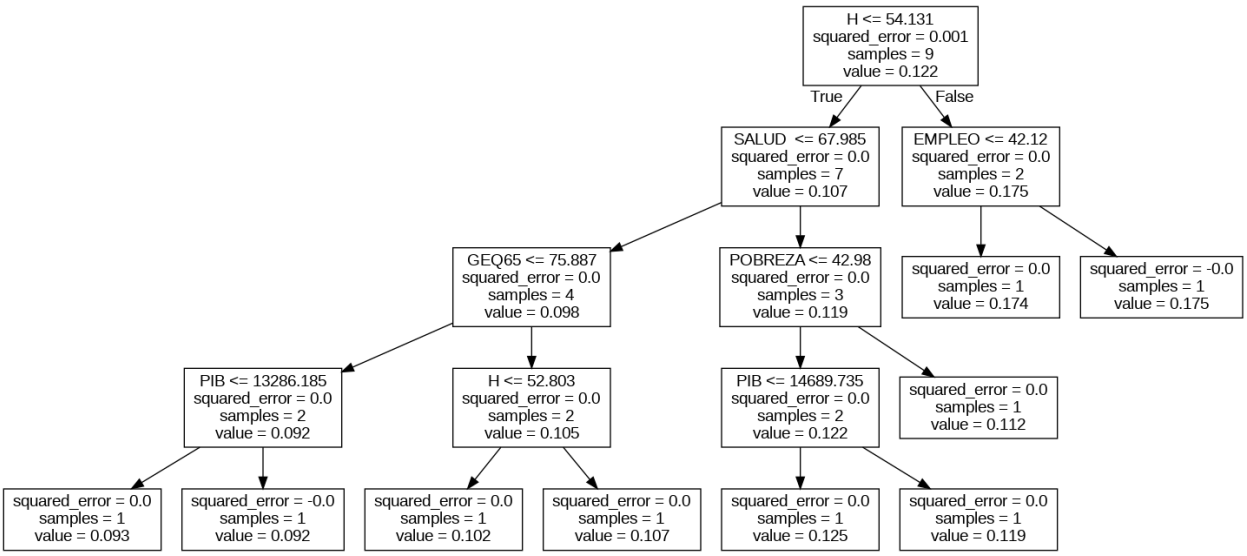

Figura 4-4.: Árbol de decisión para enfermedades del corazón.

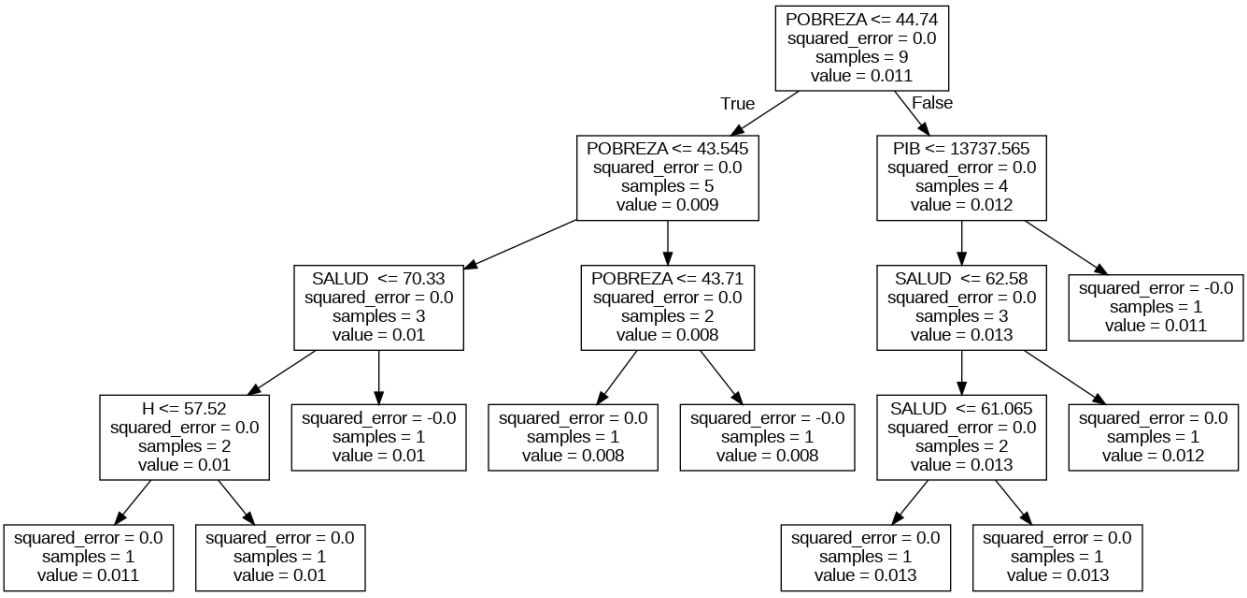

Figura 4-5.: Árbol de decisión para las muertes relacionadas con el periodo perinatal.

De igual forma en la Figura **4-6** se muestra el árbol para la causa de muerte suicidio, ésta vez se inicia tomando en cuenta la edad, específicamente el rango que comprende desde los 15 hasta los 24 años. El porcentaje de la población total del país que fallece por esta causa es bastante regular por lo que los valores, en casi todos los nodos, es el mismo. Sin embargo, observando la jerarquía del árbol se pude concluir que la edad, el empleo y el internet son las variables más relevantes para esta causa de muerte.

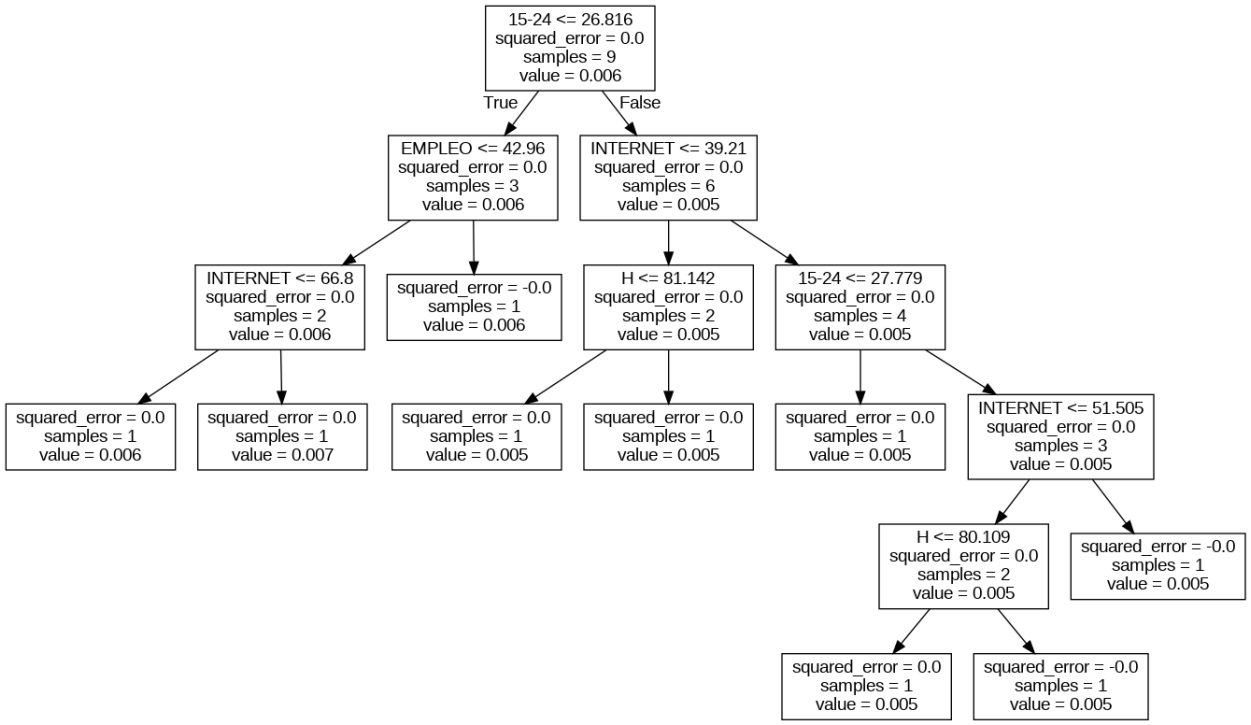

Figura 4-6.: Árbol de decisión para muertes por suicidio.

A partir de los resultados del algoritmo de árbol de decisión se pueden obtener valores relacionados con la significancia de cada variable; estos resultados para las variables relacionadas con la causa de muerte enfermedad del corazón se muestran en la Figura **4-7** y los de la causa de muerte por afecciones en el periodo perinatal en la Figura **4-8**. Por otro lado, en la sección 4.3.3 se muestran los resultados conseguidos a partir de la relevancia de las variables de manera gráfica.

4.2.1. Reglas derivadas del árbol de decisión

A partir del árbol de decisión se pueden generar reglas derivadas de acuerdo a las variables seleccionadas. Además, la librería permite obtener las probabilidades para cada rama del árbol en nuestro conjunto de datos. En la Figura **4-9** se observan las reglas obtenidas para las enfermedades del corazón, en la Figura **4-10** las reglas para las afecciones en el periodo

```
importancia=model.feature_importances_

SALUD  : 0.08791332093046365
EMPLEO : 2.0968278076253623e-05
SEGURIDAD : 0.0
POBREZA : 0.008039336516907759
PIB : 0.001801115538489012
45-64 : 0.0
GEQ65 : 0.018588208636451255
H : 0.883637050099612
M : 0.0
```

Figura 4-7.: Relevancia de cada variable, enfermedad del corazón

```
importancia=model.feature_importances_
SALUD  : 0.021244824275239542
POBREZA : 0.9085186460850487
PIB : 0.06989246314892898
L15 : 0.0
H : 0.0003440664907828561
```

Figura 4-8.: Relevancia de cada variable, periodo perinatal

perinatal y en la Figura **4-11** las reglas para las muertes por suicidio. En el anexo A se pueden observar las reglas obtenidas para el resto de causas de muerte.

```
Si H<=54.131 and SALUD<=67.98 and GEQ65<=75.88 and PIB<=13286:
VALOR=0.92 (PROBABILIDAD 0.222)
Si H<=54.131 and SALUD<=67.98 and GEQ65<=75.88 and PIB>13286:
VALOR=0.93 (PROBABILIDAD 0.222)
Si H<=54.131 and SALUD<=67.985 and GEQ65<=75.88 and H<=52.80:
VALOR=0.102 (PROBABILIDAD 0.111)
Si H<=54.131 and SALUD<=67.985 and GEQ65<=75.88 and H>52.80:
VALOR=0.107 (PROBABILIDAD 0.055)
Si H<=54.13 and SALUD<=67.98 and POBREZA<=42.98 and PIB<=14689.735:
VALOR=0.125 (PROBABILIDAD 0.055)
Si H<=54.13 and SALUD<=67.98 and POBREZA<=42.98 and PIB>14689.735:
VALOR=0.119 (PROBABILIDAD 0.055)
Si H<=54.13 and SALUD<=67.98 and POBREZA>42.98:
VALOR=0.112 (PROBABILIDAD 0.055)
Si H<=54.13 and EMPLEO<=42.12:
VALOR=0.174 (PROBABILIDAD 0.111)
Si H>54.13 and EMPLEO>42.12:
VALOR=0.175 (PROBABILIDAD 0.111)
```

Figura 4-9.: Reglas obtenidas mediante el árbol de decisión para enfermedades del corazón

```
Si POBREZA <=43.545 and SALUD <=70.33 and H <=57.52:
    VALOR=0.011 (PROBABILIDAD 0.111)
Si POBREZA <=43.545 and SALUD <=70.33 and H >57.52:
   VALOR=0.010 (PROBABILIDAD 0.0555)
Si POBREZA <=43.545 and SALUD >70.33:
   VALOR=0.010 (PROBABILIDAD 0.0555)
Si POBREZA >43.545 and POBREZA <=43.71:
   VALOR=0.008 (PROBABILIDAD 0.555)
Si POBREZA <=44.74 and POBREZA >43.71:
   VALOR=0.008 (PROBABILIDAD 0.0555)
Si POBREZA >44.74 and PIB <=13737.565 and SALUD <=61.065:
   VALOR=0.013 (PROBABILIDAD 0.222)
Si POBREZA >44.74 and PIB <=13737.565 and SALUD <= 62.58 and SALUD
   >61.065:
   VALOR=0.013 (PROBABILIDAD 0.222)
Si POBREZA >44.74 and PIB <=13737.565 and SALUD >62.58:
   VALOR=0.012 (PROBABILIDAD 0.111)
Si POBREZA >44.74 and PIB >13737.565:
   VALOR=0.011 (PROBABILIDAD 0.111)
```

Figura 4-10.: Reglas obtenidas mediante el árbol de decisión para afecciones en el periodo perinatal

```
Si 15-24 <=26.816 and EMPLEO <=42.96 and INTERNET <=66.8:
   VALOR=0.006 (PROBABILIDAD 0.222)
Si 15-24 <=26.816 and EMPLEO <=42.96 and INTERNET >66.8:
   VALOR=0.007 (PROBABILIDAD 0.0555)
Si 15-24 <=26.816 and EMPLEO >42.96:
   VALOR=0.006 (PROBABILIDAD 0.222)
Si 15-24 >26.816 and INTERNET <=39.21 and H <=81.142:
   VALOR=0.005 (PROBABILIDAD 0.0555)
Si 15-24 >26.816 and INTERNET <=39.21 and H >81.142:
   VALOR=0.005 (PROBABILIDAD 0.0555)
Si 15-24 >26.816 and INTERNET >39.21 and 15-24<=27.779:
   VALOR=0.005 (PROBABILIDAD 0.222)
Si INTERNET >39.21 and 15-24 >27.779 and INTERNET <=51.505 and H <=80.109:
   VALOR=0.005 (PROBABILIDAD 0.0555)
Si INTERNET >39.21 and 15-24 >27.779 and INTERNET <=51.505 and H >80.109:
   VALOR=0.005 (PROBABILIDAD 0.0555)
Si 15-24 >27.779 and INTERNET >51.505:
   VALOR=0.005 (PROBABILIDAD 0.0555)
```

Figura 4-11.: Reglas obtenidas mediante el árbol de decisión para muertes por suicidio

4.3. Gráficas obtenidas

A partir de las implementaciones, se puede representar la información de forma gráfica. Sin embargo, para el caso de las redes neuronales artificiales y para las máquinas de vector soporte, por su naturaleza de "blackbox" (caja negra), no se pudieron obtener patrones explicables aunque ambas herramientas son capaces de predecir los valores para el conjunto de pruebas. Por lo anterior, únicamente se muestran de forma gráfica los resultados de la ponderación de los métodos estadísticos: Pearson y GLS.

Finalmente, se muestran las gráficas que representan los resultados obtenidos mediante los árboles de decisión del algoritmo CART.

4.3.1. Pearson

Es necesario recordar que para estas gráficas no se tomaron en cuenta los atributos de edad y género. Por otro lado, los resultados difieren del análisis de regresión porque el coeficiente de Pearson toma en cuenta la distancia absoluta mientras que, en el análisis de regresión se considera la distancia cuadrática media. En la Figura **4-12** se pueden ver los porcentajes correspondientes a las causas de muerte: corazón, diabetes, tumores, influenza, hígado y cerebrovasculares. De igual forma, en la Figura **4-13** se muestran las gráficas de pastel para las causas de muerte: agresiones, accidentes, enfermedades pulmonares, insuficiencia renal, afecciones en el periodo perinatal y suicidios.

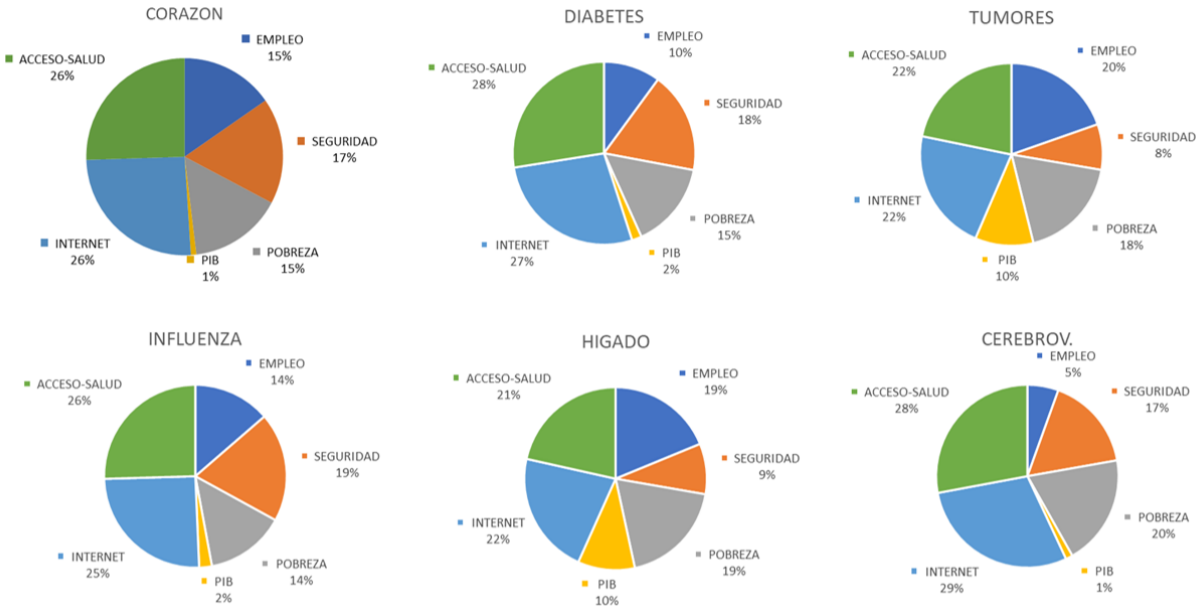

Figura 4-12.: Gráficas obtenidas con coeficiente de Pearson

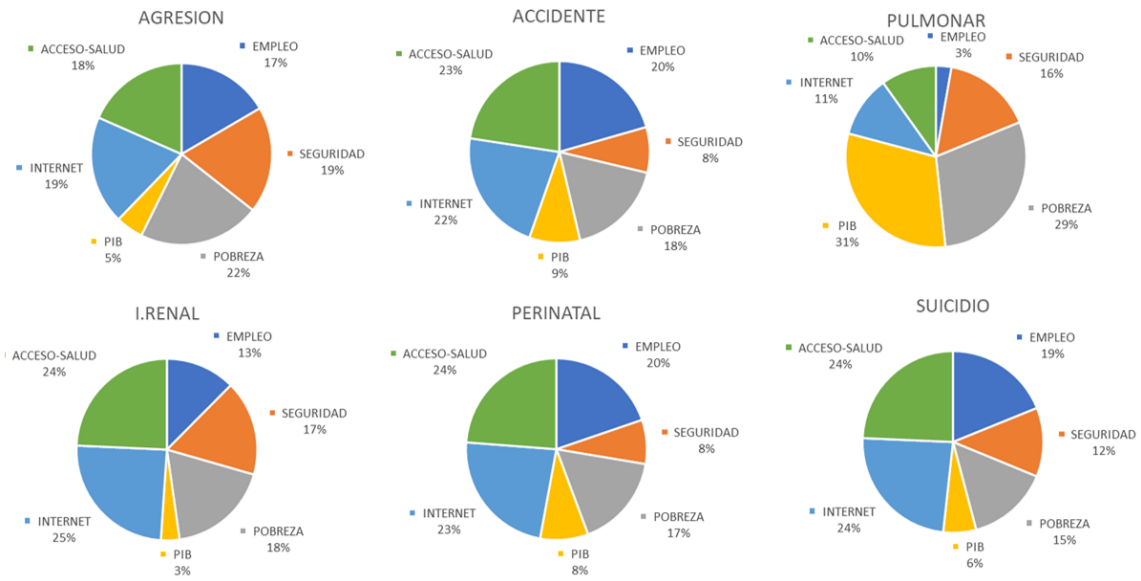

Figura 4-13.: Gráficas obtenidas con coeficiente de Pearson

4.3.2. Mínimos cuadrados generalizados GLS

Por la naturaleza del análisis de regresión únicamente se puede saber si una variable está o no contribuyendo al modelo a partir de los valores p. Por ello, para el caso de las gráficas obtenidas, solo se muestran las variables significativas para cada causa de muerte, en la Figura **4-14** parte izquierda. Además, en la Figura **4-14** parte derecha, se pueden observar la importancia de las variables para todas las causas en una gráfica de pastel; se puede ver que el acceso a la salud y la pobreza son los factores más relevantes para las causas de muerte en general.

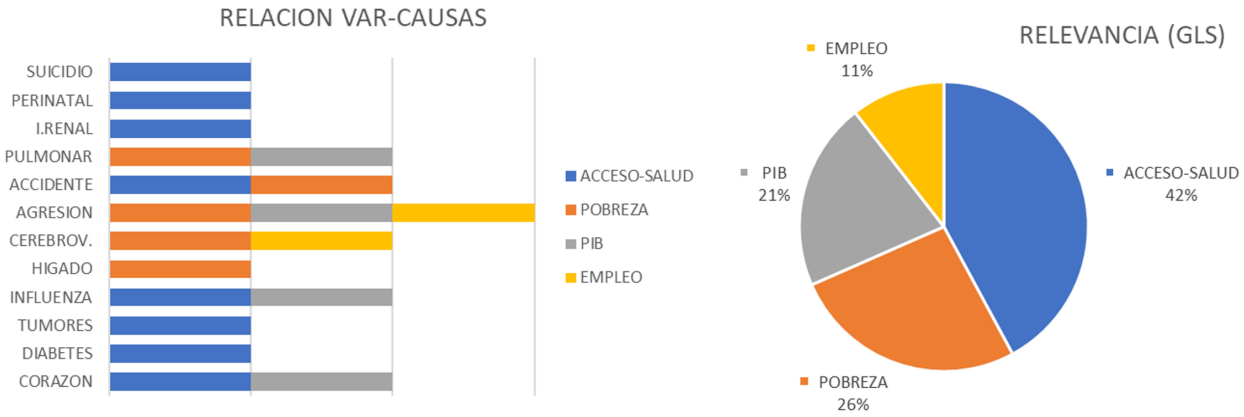

Figura 4-14.: Gráfica correspondiente a la regresión GLS

4.3.3. Árbol de decisión CART

De la interpretación de los resultados obtenidos mediante los árboles de decisión es posible generar gráficas que ponderen la relevancia de las variables socieconómicas para cada causa de muerte. En la Figura **4-15** se muestran las variables con mayor relevancia para las siguientes causas de muerte:

- Enfermedades del corazón
- Diabetes
- Tumores
- Influenza
- Enfermedades del hígado
- Enfermedades cerebrovasculares

En la Figura **4-16** se muestran los resultados obtenidos para el resto de causas de muertes descritas a continuación:

- Agresiones
- Accidentes
- Enfermades pulmonares
- Insuficiencia renal
- Afecciones en el periodo perinatal
- Suicidios

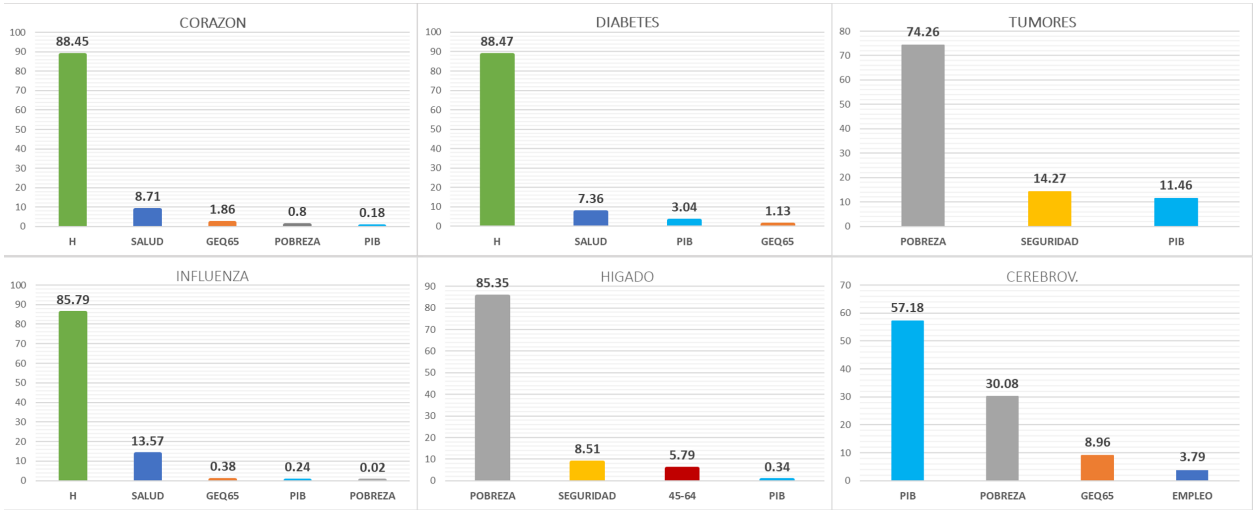

Figura 4-15.: Gráficas obtenidas con algoritmo CART (1)

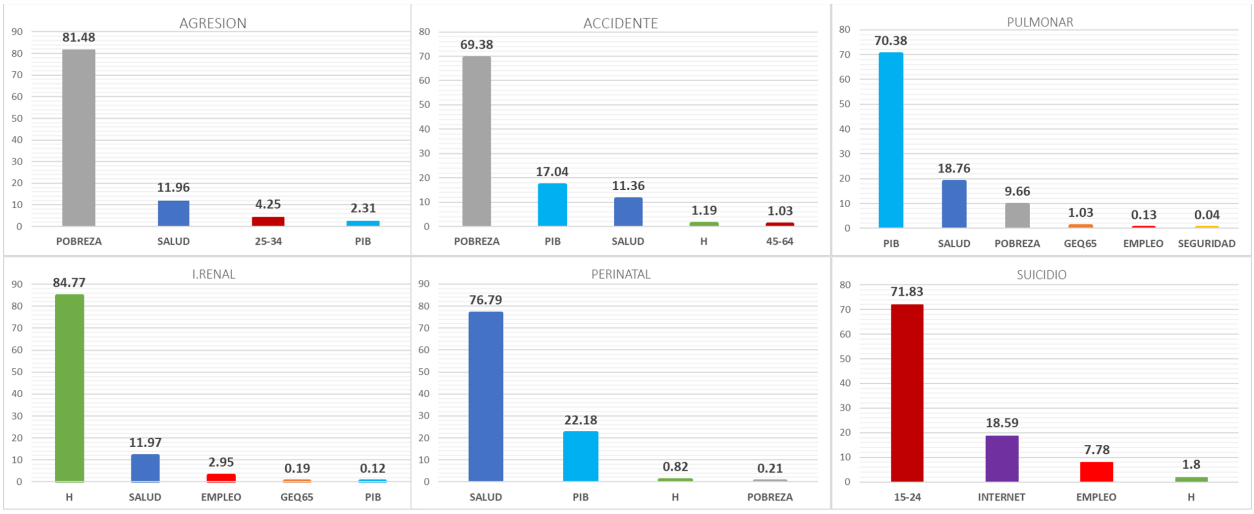

Figura 4-16.: Gráficas obtenidas con algoritmo CART (2)

Por último, en la Figura **4-17** se presenta la relevancia, dado como porcentaje, de las variables consideradas en este análisis. Esta gráfica es de suma importancia, debido a que mediante esta se puede observar la jerarquía de los factores que afectan al país para las 12 causas de muerte principales. Los factores que más afectan son: la pobreza con el 22,73 %, el género con el 20,45 % y el acceso a los servicios de salud con 19,32 %.

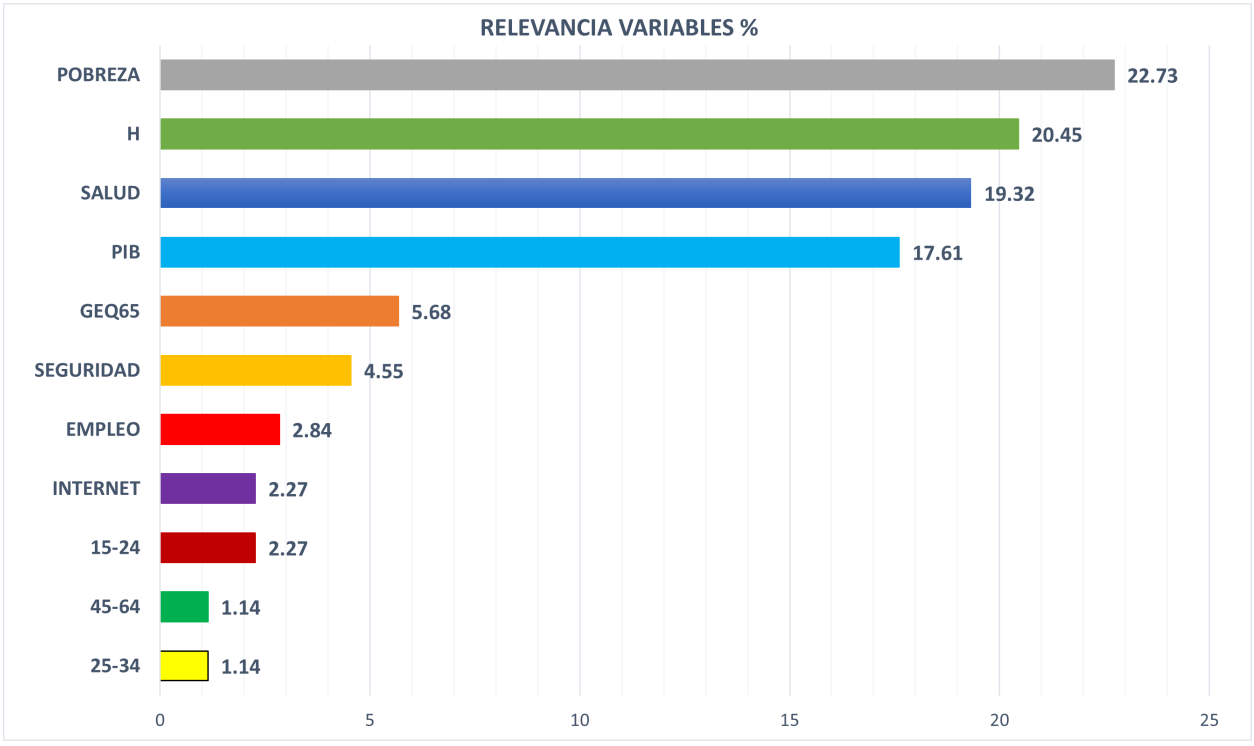

Figura 4-17.: Relevancia de las variables por porcentaje (CART)

5. Conclusiones

De acuerdo a los resultados obtenidos, se pueden obtener varias conclusiones. Sin embargo, es importante recordar la metodología que se siguió:

- Primero se recolectaron los datos de las causas de muerte y de las variables socioeconómicas disponibles.
- Posteriormente, se realizó la correlación de las variables con causas de muerte. Para esto se utilizó el coeficiente de Pearson.
- Después, se realizó un análisis de regresión por cuadrados mínimos generalizados para tener otra forma de seleccionar las variables para cada una de las causas de muerte.
- A continuación, se aplicaron los dos primeros modelos de aprendizaje automático: redes neuronales artificiales (3 capas densas con 128 neuronas) y máquinas de vector soporte (kernel polinomial).
- Finalmente, se procesaron todas las variables para cada causa de muerte con el algoritmo de árbol de decisión CART.

Una vez que se había realizado este proceso, se analizaron las capacidades de las herramientas de aprendizaje automático. De este análisis destaca que, tanto las redes neuronales artificiales como las máquinas de vector soporte tienen un buen desempeño en las tareas de regresión; esto se puede ver de forma clara, por ejemplo, en la Figura **4-1** y en la Tabla **4-2**. Sin embargo, los patrones encontrados forman parte de los modelos de SVM y ANN, es decir, pueden relacionar los datos, aunque por su naturaleza de "blackbox" (caja negra) no se puede obtener información adicional que, de forma explícita, pondere la relevancia de las diferentes variables consideradas.
Aunque de los modelos de ANN y SVM no se pudo conseguir información adicional, se realizaron gráficas con los métodos estadísticos que se utilizaron para identificar la relevancia de las variables respecto de las causas de muerte. De estas gráficas se desprende información útil y además, se puede ver que difieren entre ellos. Esto se puede explicar analizando su comportamiento, ambos se basan en regresión lineal aunque uno de ellos utiliza la distancia cuadrática media como parámetro.
Por otro lado, el algoritmo de árbol de decisión si fue capaz de obtener patrones que relacionan la importancia de las variables contra cada una de las causas de muerte. En el árbol de

la Figura **4-4**, por ejemplo, se observa que el género hombre se selecciona como la raíz pues tiene relevancia en un mayor número de muestras del conjunto de datos.
Por lo anterior, el algoritmo de árbol de decisión es el método más útil cuando se trata de relacionar datos, y sus resultados serán los que se tomen como conclusión de esta investigación. De igual manera, los resultados obtenidos con el análisis de regresión GLS consiguió resultados muy similares; esto es cuando se observa la frecuencia de las variables (Figura **4-14** y Figura **4-17**).
Del algoritmo de árbol de decisión utilizado, algoritmo CART, se puede identificar (Figura **4-16**) cuales atributos (variables socioeconómicas) tienen mayor influencia para cada una de las principales causas de muerte en la población Mexicana.
Por ejemplo, el suicidio resultó ser la única causa relacionada con el número de usuarios con acceso a internet y sobretodo, que afecta fuertemente a los individuos dentro del rango de edad 15-24. Para la causa de muerte por agresión, la pobreza destaca sobre el resto de factores. Es decir, confirma una relación entre pobreza y violencia. Para las causas de Diabetes, enfermedades del corazón, insuficiencia renal e influenza, el atributo – género masculino es el más suceptible. De los resultados anteriores se infiere que las personas con edades mayores o iguales a los 65 años del género masculino deberían de tener prioridad en los servicios de salud, dado que son el grupo más afectado por las primeras causas de muerte en México.
Finalmente, en la Figura **4-17**, se puede observar que alconsiderar todas las causas de muerte de forma general, entonces los atributos de pobreza y el acceso a los servicios de salud serán las variables más relevantes (con mayor correlación) para la mortalidad de la población Mexicana. Además, en contraste, los factores como la seguridad, el empleo y el internet, realmente repercuten débilmente sobre el total de defunciones, aúnque considerando de forma específica cada causa de muerte, hay algunas pocas causas de muerte donde los atributos anteriores si son relevantes.

A. Anexo: información complementaria

A continuación se muestran las gráficas correspondientes a la evolución, durante el periodo 2010-2021, de las diferentes variables socieconómicas consideradas en este estudio (ver Figura **A-1**). Los valores se presentan por cada 100 mil habitantes de la población Mexicana.

Figura A-1.: Variables socioeconómicas durante el periodo 2010-2021

También se incluyen las principales causas de muerte para la población Mexicana, igualmente, durante el periodo 2010-2021 (ver Figura **A-2**).
Además, se presenta la evolución de los valores de la población Mexicana, normalizada por cada 100 mil habitantes (ver Figura **A-3**).
En la Tabla **A-1** se muestran las diez principales causas de muerte de la población Mexicana durante el periodo 2010 al 2021.
Finalmente, se muestran las reglas obtenidas para el resto de causas de muerte las cuales describen los patrone encontrados en los datos.

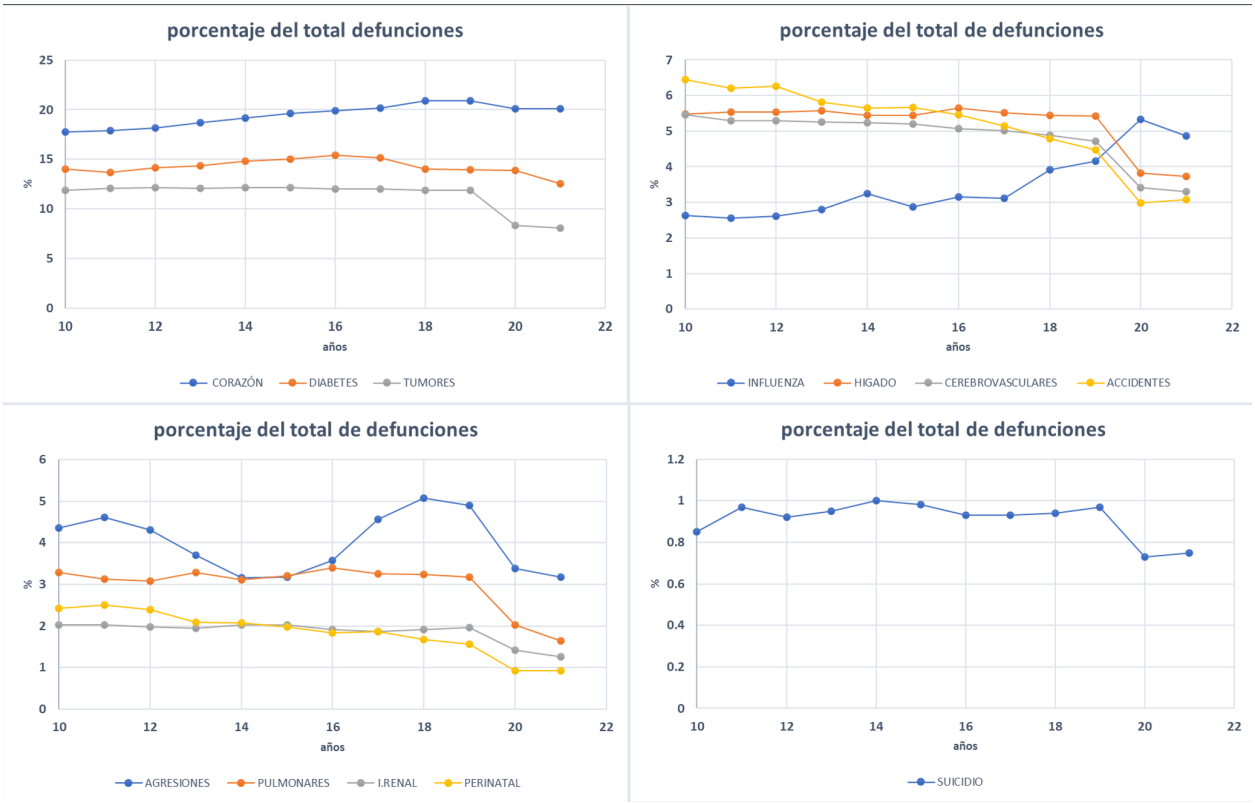

Figura A-2.: Porcentaje del total de defunciones correspondiente a cada causa de muerte durante el periodo 2010-2021

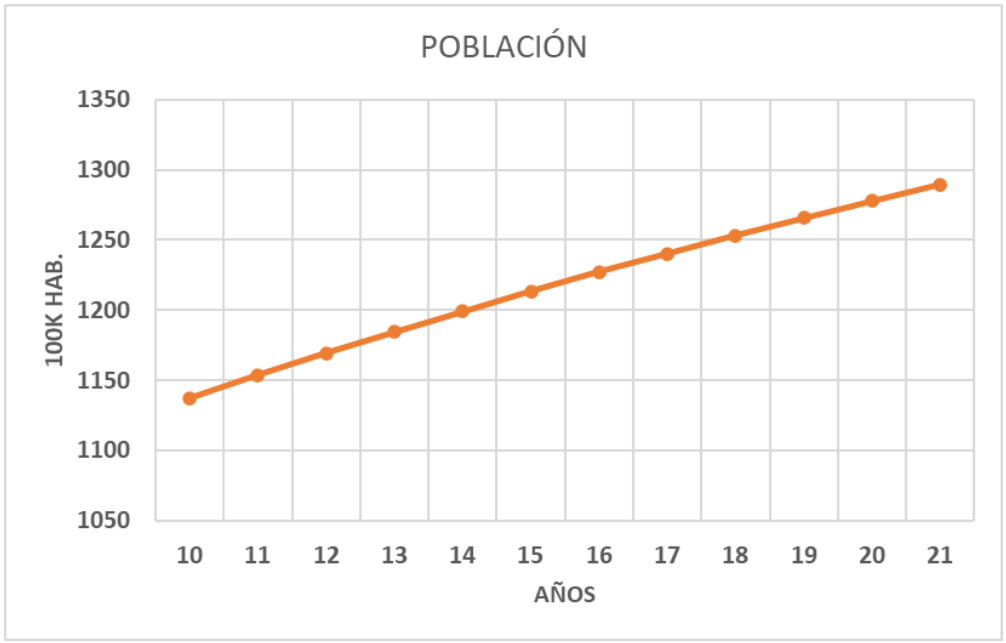

Figura A-3.: Población por 100k hab. en el periodo 2010-2021

Tabla A-1.: Principales causas de muerte en el periodo 2010-2021

	2010-2013	**2014**	**2015**
1	CORAZÓN	CORAZÓN	CORAZÓN
2	DIABETES	DIABETES	DIABETES
3	TUMORES	TUMORES	TUMORES
4	ACCIDENTES	ACCIDENTES	ACCIDENTES
5	HIGADO	HIGADO	HIGADO
6	CEREBROVASCULARES	CEREBROVASCULARES	CEREBROVASCULARES
7	AGRESIONES	INFLUENZA	PULMONARES
8	PULMONARES	AGRESIONES	AGRESIONES
9	INFLUENZA Y NEUMONIA	PULMONARES	INFLUENZA
10	PERINATAL	PERINATAL	INSUFICIENCIA RENAL
	2016-2017	**2018-2019**	**2020**
1	CORAZÓN	CORAZÓN	CORAZÓN
2	DIABETES	DIABETES	COVID19
3	TUMORES	TUMORES	DIABETES MELLITUS
4	HIGADO	HIGADO	TUMORES MALIGNOS
5	ACCIDENTES	AGRESIONES	INFLUENZA
6	CEREBROVASCULARES	CEREBROVASCULARES	HIGADO
7	AGRESIONES	ACCIDENTES	CEREBROVASCULARES
8	PULMONARES	INFLUENZA	AGRESIONES
9	INFLUENZA	PULMONARES	ACCIDENTES
10	INSUFICIENCIA RENAL	INSUFICIENCIA RENAL	PULMONARES
	2021		
1	COVID19		
2	CORAZÓN		
3	DIABETES		
4	TUMORES		
5	INFLUENZA		
6	HIGADO		
7	CEREBROVASCULARES		
8	AGRESIONES		
9	ACCIDENTES		
10	PULMONARES		

```
Si H<= 50.12 and SALUD <=66.725 and GEQ65 <= 61.591:
   VALOR=0.079 (PROBABILIDAD .0555)
Si H<= 50.12 and SALUD <=66.725 and GEQ65 > 61.591 and SEGURIDAD <=28.445:
   VALOR=0.073 (PROBABILIDAD 0.111)
Si H<= 50.12 and SALUD <=66.725 and GEQ65 > 61.591 and SEGURIDAD >28.445:
   VALOR=0.073 (PROBABILIDAD 0.111)
Si H<= 50.12 and SALUD <=66.725 and SEGURIDAD <=29.705 and L15 <=0.039:
   VALOR=0.082 (PROBABILIDAD 0.222)
Si H<= 50.12 and SALUD <=66.725 and SEGURIDAD <=29.705 and L15 >0.039:
   VALOR=0.083 (PROBABILIDAD 0.222)
Si H<= 50.12 and SALUD <=66.725 and SEGURIDAD >29.705 and SEGURIDAD
   <=30.95:
   VALOR=0.087 (PROBAILIDAD 0.0555)
Si H<= 50.12 and SALUD <=66.725 and SEGURIDAD >29.705 and SEGURIDAD
   >30.95:
   VALOR=0.086 (PROBABILIDAD 0.0555)
Si H>50.12 and PIB <=13639.165:
   VALOR=0.120 (PROBABILIDAD 0.0555)
Si H>50.12 and PIB >13639.165:
   VALOR=0.109 (PROBABILIDAD 0.0555)
```

Figura A-4.: Reglas obtenidas mediante el árbol de decisión para muertes por diabetes

```
Si POBREZA <=44.305 and SEGURIDAD <=27.06:
   VALOR=0.072 (PROBABILIDAD 0.222)
Si POBREZA <=44.305 and SEGURIDAD <=27.06 and GEQ65 <=54.514:
   VALOR=0.069 (PROBABILIDAD 0.111)
Si POBREZA <= 44.305 and SEGURIDAD >27.06 and GEQ65 >54.514:
   VALOR=0.068 (PROBABILIDAD 0.0555)
Si POBREZA >44.305 and PIB <=13737.565 and SEGURIDAD <=28.445 and 25-34
   <=3.009:
   VALOR=0.062 (PROBABILIDAD 0.111)
Si POBREZA >44.305 and PIB <=13737.565 and SEGURIDAD <=28.445 and 25-34
   >3.009:
   VALOR=0.062 (PROBABILIDAD 0.111)
Si POBREZA >44.305 and PIB <=13737.565 and SEGURIDAD >28.445:
   VALOR=0.063 (PROBABILIDAD 0.222)
Si POBREZA >44.305 and PIB >13737.565 and SEGURIDAD <=31.99:
   VALOR=0.066 (PROBABILIDAD 0.0555)
Si POBREZA >44.305 and PIB >13737.565 and SEGURIDAD >31.99 and PIB
   <=13849.84:
   VALOR=0.064 (PROBABILIDAD 0.064)
Si POBREZA >44.305 and PIB >13737.565 and SEGURIDAD >31.99 and PIB
   >13849.84:
   VALOR=0.065 (PROBABILIDAD 0.0555)
```

Figura A-5.: Reglas obtenidas mediante el árbol de decisión para muertes por tumores

```
Si H <=58.461 and PIB <=13737.565 and SALUD <=62.58:
   VALOR=0.013 (PROBABILIDAD 0.0555)
Si H <=58.461 and SALUD <=65.41 and PIB <=13737.565 and SALUD >62.58:
   VALOR=0.014 (PROBABILIDAD 0.0555)
Si H <=58.461 and SALUD <=65.41 and PIB >13737.565:
   VALOR=0.015 (PROBABILIDAD .111)
Si SALUD <=70.86 and SALUD >65.41 and GEQ65 <=67.03 and POBREZA <=44.835
   and H <=55.677:
   VALOR=0.018 (PROBABILIDAD 0.0555)
Si H <=58.461 and SALUD <=70.86 and SALUD >65.41 and GEQ65 <=67.03 and
   POBREZA <=44.835 and H>55.677:
   VALOR=0.018 (PROBABILIDAD 0.0555)
Si H<=58.461 and SALUD <=70.86 and SALUD>65.41 and GEQ65<=67.03 and
   POBREZA >44.835:
   VALOR=0.017 (PROBABILIDAD 0.111)
Si H <=58.641 and SALUD<=70.86 and SALUD>65.41 and GEQ65 >67.03:
   VALOR=0.016 (PROBABILIDAD 0.222)
Si H<=58.461 and SALUD >70.86:
   VALOR=0.025 (PROBABILIDAD 0.222)
Si H> 58.461:
   VALOR=0.042 (PROBABILIDAD 0.111)
```

Figura A-6.: Reglas obtenidas mediante el árbol de decisión para muertes por influenza

```
Si POBREZA <= 44.305 and SEGURIDAD <= 27.055:
   VALOR=0.033 (PROBABILIDAD 0.222)
Si POBREZA <= 44.305 and SEGURIDAD <= 27.055 and 45-64 <=44.945:
   VALOR=0.032 (PROBABILIDAD 0.222)
Si POBREZA <= 44.305 and SEGURIDAD <= 27.055 and 45-64 <=44.945 and L15
   <=0.45:
   VALOR=0.031 (PROBABILIDAD 0.111)
Si POBREZA <= 44.305 and SEGURIDAD <= 27.055 and 45-64 <=44.945 and L15
   >0.45:
   VALOR=0.031 (PROBABILIDAD 0.111)
Si POBREZA > 44.305 and 45-64 <=44.501 and PIB <=13846.045 and L15
   <=0.551:
   VALOR=0.029 (PROBABILIDAD 0.0555)
Si POBREZA > 44.305 and 45-64 <=44.501 and PIB <=13846.045 and L15 >0.551:
   VALOR=0.029 (PROBABILIDAD 0.0555)
Si POBREZA > 44.305 and 45-64 <=44.501 and PIB >13846.045:
   VALOR= 0.029 (PROBABILIDAD 0.111)
Si POBREZA > 44.305 and 45-64 >44.501 and SEGURIDAD <=32.17:
   VALOR=0.030 (PROBABILIDAD 0.0555)
Si POBREZA > 44.305 and 45-64 >44.501 and SEGURIDAD >32.17:
   VALOR= 0.031(PROBABILIDAD 0.0555)
```

Figura A-7.: Reglas obtenidas mediante el árbol de decisión para muertes por enfermedades del hígado

```
Si PIB <=13860.785 and POBREZA <=46.38 and GEQ65 <=76.259 and EMPLEO
    <=41.325:
   VALOR=0.027 (PROBABILIDAD 0.0555)
Si PIB <=13860.785 and POBREZA <=46.38 and GEQ65 <=76.259 and EMPLEO
    >41.325:
   VALOR=0.027 (PROBABILIDAD 0.0555)
Si PIB <=13860.785 and POBREZA <=46.38 and GEQ65 >76.259:
   VALOR=0.028 (PROBABILIDAD 0.111)
Si PIB <=13860.785 and POBREZA >46.38:
   VALOR=0.028 (PROBABILIDAD 0.222)
Si PIB >13860.785 and PIB <=14115.215:
   VALOR=0.029 (PROBABILIDAD 0.222)
Si PIB >14115.215 and EMPLEO <=41.555:
  VALOR=0.029 (PROBABILIDAD 0.111)
Si PIB >14115.215 and EMPLEO >41.555 and GEQ65 <=74.179:
   VALOR=0.028 (PROBABILIDAD 0.111)
Si PIB >14115.215 and EMPLEO >41.555 and GEQ65 >74.179 and POBREZA
    <=43.66:
  VALOR=0.028 (PROBABILIDAD 0.0555)
Si PIB >14115.215 and EMPLEO >41.555 and GEQ65 >74.179 and POBREZA >43.66:
  VALOR=0.028 (PROBABILIDAD 0.0555)
```

Figura A-8.: Reglas obtenidas mediante el árbol de decisión para muertes por enfermedades cerebrovasculares

```
Si POBREZA <=44.74 and 25-34 <=30.947:
   VALOR=0.026 (PROBABILIDAD 0.111)
Si POBREZA <=44.74 and 25-34 >30.947 and SALUD <= 73.44 and 25-34
    <=31.359:
   VALOR=0.029 (PROBABILIDAD 0.0555)
Si POBREZA <=44.74 and SALUD <=72.455 and 25-34 >31.359:
   VALOR=0.029 (PROBABILIDAD 0.0555)
Si POBREZA <=44.74 and SALUD <=73.44 and 25-34 >31.359 and SALUD>72.455:
  VALOR=0.029 (PROBABILIDAD 0.0555)
Si POBREZA <=44.74 and 25-34>30.947 and SALUD >73.44:
   VALOR=0.028 (PROBABILIDAD 0.0555)
Si POBREZA > 44.74 and SALUD <=64.03 and PIB <=13437.4:
   VALOR=0.023 (PROBABILIDAD 0.111)
Si POBREZA >44.74 and SALUD <=64.03 and PIB >13437.4:
   VALOR=0.022 (PROBABILIDAD 0.111)
Si POBREZA > 44.74 and SALUD >64.03 and PIB <=13849.84:
   VALOR=0.020 (PROBABILIDAD 0.222)
Si POBREZA > 44.74 and SALUD >64.03 and PIB <=13849.84:
   VALOR=0.017 (PROBABILIDAD 0.222)
```

Figura A-9.: Reglas obtenidas mediante el árbol de decisión para muertes por agresiones

```
Si SALUD <=69.185 and 25-34 <=15.727 and PIB <=14356.19:
   VALOR=0.031 (PROBABILIDAD 0.222)
Si SALUD <=69.185 and 25-34 <=15.727 and PIB >14356.19:
   VALOR=0.031 (PROBABILIDAD 0.111)
Si SALUD <=69.185 and 25-34 <=16.026 and 25-34 >15.727:
   VALOR=0.030 (PROBABILIDAD 0.0555)
Si SALUD <=61.805 and 25-34 >16.026:
   VALOR=0.034 (PROBABILIDAD 0.111)
Si SALUD <=69.185 and 25-34 >16.026 and SALUD >61.805:
   VALOR=0.032 (PROBABILIDAD 0.111)
Si SALUD >69.185 and SEGURIDAD <=30.955 and POBREZA <=42.995 and H
   <=76.844:
   VALOR=0.028 (PROBABILIDAD 0.0555)
Si SALUD >69.185 and SEGURIDAD <=30.955 and POBREZA <=42.995 and H
   >76.844:
   VALOR=0.027 (PROBABILIDAD 0.0555)
Si SALUD >69.185 and SEGURIDAD <=30.955 and POBREZA >42.995:
   VALOR=0.026 (PROBABILIDAD 0.0555)
Si SALUD >69.185 and SEGURIDAD >30.955:
   VALOR=0.029 (PROBABILIDAD 0.0555)
```

Figura A-10.: Reglas obtenidas mediante el árbol de decisión para muertes por accidentes

```
Si PIB <=14115.215 and SALUD <=73.44 and GEQ65 <=90.318 and EMPLEO
   <=41.01:
   VALOR=0.017 (PROBABILIDAD 0.093)
Si PIB <=14115.215 and SALUD <=73.44 and GEQ65 <=90.318 and EMPLEO >41.01:
   VALOR=0.17 (PROBABILIDAD 0.093)
Si PIB <=14115.215 and SALUD <=73.44 and GEQ65 >90.318 and PIB <=13846.045
    and EMPLEO <=41.325:
   VALOR=0.016 (PROBABILIDAD 0.0555)
Si PIB <=14115.215 and SALUD <=73.44 and GEQ65 >90.318 and PIB <=13846.045
    and EMPLEO >41.325:
   VALOR=0.016 (PROBABILIDAD 0.0555)
Si PIB <=14115.215 and SALUD <=73.44 and GEQ65 >90.318 and PIB >13846.045:
   VALOR=0.017 (PROBABILIDAD 0.093)
Si PIB <=14115.215 and SALUD >73.44:
   VALOR=0.014 (PROBABILIDAD 0.277)
Si PIB >14115.215 and GEQ65 <=90.675 and POBREZA <=42.35:
   VALOR=0.019 (PROBABILIDAD 0.111)
Si PIB >14115.215 and GEQ65 <=90.675 and POBREZA >42.35:
   VALOR=0.019 (PROBABILIDAD 0.111)
Si PIB >14115.215 and GEQ65 >90.675:
   VALOR=0.018 (PROBABILIDAD 0.111)
```

Figura A-11.: Reglas obtenidas mediante el árbol de decisión para muertes por enfermedades pulmonares

```
Si H <=56.683 and SALUD <=65.41 and GEQ65 <=60.456:
   VALOR=0.010 (PROBABILIDAD 0.0555)
Si H <=56.683 and SALUD <=65.41 and GEQ65 >60.456:
   VALOR=0.010 (PROBABILIDAD 0.0555)
Si H <=56.683 and SALUD >65.41 and EMPLEO <=41.685 and SALUD <=67.34:
   VALOR=0.011 (PROBABILIDAD 0.0835)
Si H <=56.683 and EMPLEO <=41.685 and SALUD <=69.185 and SALUD >67.34:
   VALOR=0.011 (PROBABILIDAD 0.0835)
Si H <=56.683 and EMPLEO <=41.685 and SALUD >69.185:
   VALOR=0.011 (PROBABILIDAD 0.222)
Si H <=56.683 and SALUD >65.41 and EMPLEO >41.685 and PIB <=14115.215:
   VALOR=0.011 (PROBABILIDAD 0.0835)
Si H <=56.683 and SALUD >65.41 and EMPLEO >41.685 and PIB >14115.215:
   VALOR=0.011 (PROBABILIDAD 0.0835)
Si H >56.683 and H <=57.047:
   VALOR=0.012 (PROBABILIDAD 0.222)
Si H >57.047:
   VALOR=0.012 (PROBABILIDAD 0.111)
```

Figura A-12.: Reglas obtenidas mediante el árbol de decisión para muertes por insuficiencia renal

Bibliografía

[1] A. Becerra-Sánchez, A. Rodarte-Rodríguez, N. I. Escalante-García, J. E. Olvera-González, J. I. De la Rosa-Vargas, G. Zepeda-Valles, and E. d. J. Velásquez-Martínez. Mortality analysis of patients with covid-19 in mexico based on risk factors applying machine learning techniques. *Diagnostics*, 12(6):1396, 2022.

[2] A. Bhandari, M. Ibrahim, C. Sharma, R. Liong, S. Gustafson, and M. Prior. Ct-based radiomics for differentiating renal tumours: a systematic review. *Abdominal Radiology*, 46:1–12, 05 2021. https://doi.org/10.1007/s00261-020-02832-9 doi:10.1007/s00261-020-02832-9.

[3] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, and G. Varoquaux. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122, 2013.

[4] A. Burkov. *The hundred-page machine learning book*, volume 1. Andriy Burkov Quebec City, QC, Canada, 1 edition, 2019.

[5] D. J. Cambría. Mortalidad como indicador económico y social. argentina y américa latina. *Revista de Salud Pública*, 16(2):57–66, 2012.

[6] C. A. D. Cervantes. Análisis de la mortalidad por causas en méxico: Tendencias y proyecciones al 2015. Master's thesis, El Colegio de México, 2008.

[7] C. Collaborative et al. Machine learning risk prediction of mortality for patients undergoing surgery with perioperative sars-cov-2: the covidsurg mortality score. *The British journal of surgery*, 108(11):1274, 2021.

[8] S. Eligio. La mortalidad por causas en mexico y las ganancias en las esperanzas de vida. Master's thesis, Universidad Nacional Autónoma de México, 2002.

[9] E. R. Gansner and S. C. North. An open graph visualization system and its applications to software engineering. *Software: practice and experience*, 30(11):1203–1233, 2000.

[10] L. P. Garcia, I. J. C. Schneider, C. de Oliveira, E. Traebert, and J. Traebert. What is the impact of national public expenditure and its allocation on neonatal and child mortality? a machine learning analysis, Apr 2023. URL: http://dx.doi.org/10.1186/s12889-023-15683-y, https://doi.org/10.1186/s12889-023-15683-y doi:10.1186/s12889-023-15683-y.

[11] GeeksForGeeks. Backpropagation in data mining. https://www.geeksforgeeks.org//backpropagation-in-data-mining/. Accessed: 2023-02-08.

[12] A. V. Hansen, L. H. Mortensen, C. T. Ekstrøm, S. Trompet, and R. Westendorp. Predicting mortality and visualizing health care spending by predicted mortality in danes over age 65, Jan 2023. URL: http://dx.doi.org/10.1038/s41598-023-28102-4, https://doi.org/10.1038/s41598-023-28102-4 doi:10.1038/s41598-023-28102-4.

[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[14] J. Hernández Orallo et al. *Introducción a la Minería de Datos*. Pearson Educación, 1 edition, 2004.

[15] A. Holmes, B. Illowsky, and S. Dean. *Introductory business statistics*. Rice University, 1 edition, 2017.

[16] G.-H. Huang. Measure of association. In *International Encyclopedia of Education*, pages 260–263. Elsevier Ltd, 2010.

[17] IBM. Qué son las redes neuronales recurrentes? Accessed: 2023-09-29. URL: https://www.ibm.com/mx-es/topics/recurrent-neural-networks.

[18] INEGI. *Comunicado de prensa 378: Estadísticas de defunciones registradas*. INEGI, 1 edition, 2022.

[19] K. G. Jöreskog and A. S. Goldberger. Factor analysis by generalized least squares. *Psychometrika*, 37(3):243–260, 1972.

[20] H. Kinsley and D. Kukieła. *Neural Networks from Scratch in Python: Building Neural Networks in Raw Python*. Harrison Kinsley, 1 edition, 2020.

[21] I. Lawal and S. Abdulkarim. Adaptive svm for data stream classification. *South African Computer Journal*, 29, 07 2017. https://doi.org/10.18489/sacj.v29i1.414 doi:10.18489/sacj.v29i1.414.

[22] Y. Liu, S. Zhao, L. Yang, L. Aliaga-Linares, and D. He. All-cause mortality during the covid-19 pandemic in peru. *IJID Regions*, 5:177–179, 2022. URL: https://www.sciencedirect.com/science/article/pii/S277270762200131X, https://doi.org/10.1016/j.ijregi.2022.10.005 doi:10.1016/j.ijregi.2022.10.005.

[23] I. Madrid. *Metodología de indicadores de calidad de vida 2021*. Instituto Nacional de Estadística, Madrid, 1 edition, 2021.

[24] C. D. Manning. *An introduction to information retrieval*. Cambridge university press, 2009.

[25] L. Marvin. *MACHINE LEARNING: Neural Networks, Decision Trees and Support Vector Machine with IBM SPSS Modeler*. Scientific Books, 1 edition, 2022.

[26] T. Mitchell. *Machine Learning*. McGraw-Hill international editions - computer science series. McGraw-Hill Education, 1997. URL: https://books.google.com.mx/books?id=xOGAngEACAAJ.

[27] A. Moujahid, I. Inza, and P. Larranaga. Tema 8: Árboles de clasificación, 2008.

[28] V. Ortiz Velásquez. Aplicación de técnicas de aprendizaje automático para la segmentación y clasificación de características sociodemográficas asociadas a tasas de mortalidad infantil utilizando datos reportados por el dane colombia entre los años 2008 al 2017. *Universidad de Bogotá*, 1, 2020.

[29] L. S. Palacio-Mejía, J. E. Hernández-Ávila, M. Hernández-Ávila, D. Dyer-Leal, A. Barranco, A. D. Quezada-Sánchez, M. Alvarez-Aceves, R. Cortés-Alcalá, J. L. Fernández- Wheatley, I. O. nez Hernández, E. Vielma-Orozco, M. de la Cruz Muradás-Troitiño, O. Muro-Orozco, E. Navarro-Luévano, K. Rodriguez-González, J. M. Gabastou, R. López-Ridaura, and H. López-Gatell. Leading causes of excess mortality in mexico during the covid-19 pandemic 2020–2021: A death certificates study in a middle-income country. *The Lancet Regional Health - Americas*, 13:100303, 2022. URL: https://www.sciencedirect.com/science/article/pii/S2667193X2200120X, https://doi.org/10.1016/j.lana.2022.100303 doi:10.1016/j.lana.2022.100303.

[30] S. S. Patel. Explainable machine learning models to analyse maternal health. *Data Knowledge Engineering*, 146:102198, 2023. URL: https://www.sciencedirect.com/science/article/pii/S0169023X23000587, https://doi.org/10.1016/j.datak.2023.102198 doi:10.1016/j.datak.2023.102198.

[31] J. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann series in machine learning. Elsevier Science, 1993. URL: https://books.google.com.mx/books?id=HExncpjbYroC.

[32] G. Rinu. Artificial neural networks for machine learning–structure and layers. Accessed: 2023-02-9. URL: https://medium.com/javarevisited/artificial-neural-network-for-machine-learning-structure-layers-a031fcb279d7.

[33] J.-M. Robine. Indicadores de la esperanza de salud. *Boletín de la Organización Mundial de la salud: Revista internacional de salud pública 1999*, 202:106–110, 1999.

[34] F. J. P. Rodríguez. *Estadística y Machine Learning con R: Ejercicios resueltos con R*. Editorial académica española, 1 edition, 2017.

[35] A. E. Ronmi, R. Prasad, and B. A. Raphael. How can artificial intelligence and data science algorithms predict life expectancy - an empirical investigation spanning 193 countries. *International Journal of Information Management Data Insights*, 3(1):100168, 2023. URL: https://www.sciencedirect.com/science/article/pii/S2667096823000150, https://doi.org/10.1016/j.jjimei.2023.100168 doi:10.1016/j.jjimei.2023.100168.

[36] H. Salehinejad, S. Sankar, J. Barfett, E. Colak, and S. Valaee. Recent advances in recurrent neural networks. *arXiv preprint arXiv:1801.01078*, 2017.

[37] A. Sen. La vida y la muerte como indicadores económicos. *Investigación y ciencia*, 202:6–13, 1993.

[38] SitioBigData. Aprendizaje automático y las métricas de regresión. Accessed: 2023-02-9. URL: https://sitiobigdata.com/2018/08/27/machine-learning-metricas-regresion-mse/.

[39] SKLearn. Sklearn decision trees. Accessed: 2023-09-29. URL: https://scikit-learn.org/stable/modules/tree.html#tree-algorithms-id3-c4-5-c5-0-and-cart.

[40] StatsModels. Linear regression model. Accessed: 2023-02-10. URL: https://www.statsmodels.org/stable/generated/statsmodels.regression.linear_model.GLS.html.

[41] StatsModels. Statsmodels api. Accessed: 2023-02-10. URL: https://www.statsmodels.org/stable/api.html.

[42] L. Wen, W. Pan, S. Liao, W. Pan, H. Xu, and C. Hu. A combination-based machine learning algorithm estimating impacts of social, economic, and environmental on resident health—on china's provincial panel data. *Engineering Applications of Artificial Intelligence*, 123:106135, 2023. URL: https://www.sciencedirect.com/science/article/pii/S0952197623003196, https://doi.org/10.1016/j.engappai.2023.106135 doi:10.1016/j.engappai.2023.106135.

[43] D. I. R. Zavala. *Mortalidad y pobreza: México, 1990.* El Colegio de México, 1 edition, 2000.

[44] B. Zeng. *Towards understanding residual neural networks.* PhD thesis, Massachusetts Institute of Technology, 2019.