UNIVERSIDAD CENTRAL DEL ECUADOR

FACULTAD DE INGENIERÍA Y CIENCIAS APLICADAS

INGENIERÍA EN COMPUTACIÓN

Análisis de Mortalidad Hospitalaria en la provincia de Guayaquil

INTEGRANTES:

- CASTILLO CONDE SERGHY OMAR
- WILLAN ALEXANDER DIAZ CORDOBA
- YANIRY MABELY FLOREZ RIVERA
- JUAN ANDRES TOSCANO LUCERO

APRENDIZAJE AUTOMÁTICO
PHD Zoila Ruiz

QUITO - ECUADOR
2023 - 2023

.

Índice general

Resumen

En la presente investigación, se propone un enfoque basado en técnicas de aprendizaje automático para predecir la aparición de enfermedades cardio respiratorias, y cardio pulmonares que se relacionan con casos de COVID-19. El estudio está enfocado en Ecuador, específicamente en Guayaquil ya que fue una de las ciudades más afectadas por la pandemia de COVID-19 durante la primera ola en 2020. Según informes de los medios de comunicación y las autoridades de salud, la ciudad experimentó una alta tasa de mortalidad y un colapso del sistema de salud debido al gran número de casos.

Para esta práctica se utilizó un conjunto de datos de defunciones generales que incluía información demográfica de las defunciones, junto con causas probables de las mismas.

El proceso de trabajo se dividió en varias etapas. En primer lugar, se realizó un análisis exploratorio de los datos para comprender la distribución de las variables y detectar posibles correlaciones. A continuación, se aplicaron técnicas de preprocesamiento, como la normalización y la codificación de variables categóricas, para preparar los datos para el modelado.

Se evaluaron varios algoritmos de aprendizaje automático, incluyendo árboles de decisión, para la selección de variables, varios algoritmos de aprendizaje Supervisado y No Supervisado para la cauterización y análisis de resultados. Además, se identificaron las variables más influyentes en las predicciones, como la edad, y parroquias donde los problemas Cardio Pulmonares fueron más frecuentes.

Palabras Clave

ÇOVID-19", .ªprendizaje Automático", .ªprendizaje Supervisado", .ªprendizaje no Supervisado", .ªrboles Rand'omicos", .ªnfermedades Cardio Pulmonares"

Abstract

In this research, an approach based on machine learning techniques is proposed to predict the occurrence of cardiorespiratory and cardiopulmonary diseases related to COVID-19 cases. The study is focused on Ecuador, specifically in Guayaquil, as it was one of the cities most affected by the COVID-19 pandemic during the first wave in 2020. According to reports from the media and health authorities, the city experienced a high mortality rate and a collapse of the health system due to the large number of cases.

For this practice, a dataset of general deaths was used, which included demographic information of the deaths, along with probable causes of death. The work process was divided into several stages.

Firstly, an exploratory analysis of the data was carried out to understand the distribution of the variables and detect possible correlations. Then, preprocessing techniques, such as normalization and coding of categorical variables, were applied to prepare the data for modeling.

Several machine learning algorithms were evaluated, including decision trees for variable selection, various supervised and unsupervised learning algorithms for clustering and analysis of results. In addition, the most influential variables in the predictions were identified, such as age and parishes where cardiopulmonary problems were more frequent.

KeyWords

ÇOVID-19", "Machine Learning", "Supervised Learning", Ünsupervised Learning", Random Forests", Çardiopulmonary Diseases"

0.1. Introducción

En el mundo las capacidades humanas con la tecnología nos han dado avances muy significativos en el campo laboral, académico y social. Esto también conlleva a muchas situaciones donde nuestra vida corre peligro, en el Ecuador una de las temáticas en alza son las muertes, donde un índice ocurre por diferentes situaciones.

Las Defunciones Generales, corresponde a los hechos vitales de defunciones ocurridos en el territorio nacional. La inscripción de estos hechos se realiza en las oficinas a nivel nacional de la Dirección General de Registro Civil Identificación y Cedulación y en la Corporación de Registro Civil de Guayaquil.(Carrera y Llumiquinga, 2020)

El comportamiento de los registros de defunciones generales desde el año 2004 registra una tendencia creciente hasta el año 2020. El número de defunciones inscritas en el último año presenta un incremento de 2.3 puntos en relación al año 2019.(Carrera y Llumiquinga, 2020)

Esto se acentúa en los países más pobres, donde la tasa de mortalidad asociada con los accidentes de tránsito alcanza las 28.5 muertes por cada 100,000 habitantes.(Congacha y cols., s.f.)

Para nuestro contexto, las muertes se dan por un registro en la Base de Datos de cada entidad pública, como lo es el Registro Civil, Ministerio de Salud, Ministerio del Interior entre otros. Para este análisis somos conscientes de lo siguiente:

Los datos estadísticos que se investigan se obtienen en coordinación con los establecimientos de salud, dónde principalmente se originan los hechos vitales y con oficinas de la Dirección General de Registro Civil, Identificación y Cedulación(Carrera y Llumiquinga, 2020). Los diferentes registros se manejan de manera sistemática por lo general se mantiene en un extenso registro la cual no ayuda a mantener las diferentes maneras de integrar el momento necesario y las causas por el cual el ciudadano o la persona involucrada ha dado su deceso a priori de toda la información que se maneja en el registro de cada entidad involucrada:

- Entidades de Salud (hospitales, clínicas)
- Sistema de Transito
- Sistema de Investigación
- Sistema de Registro Civil.

0.2. Inteligencia Artificial

La IA es la capacidad de las máquinas para usar algoritmos, aprender de los datos y utilizar lo aprendido en la toma de decisiones tal y como lo haría un ser humano. (Rouhiainen, 2018)

La IA tiene una capacidad de crecimiento exponencial en el tiempo, ya que el hecho de tener un margen de error mínimo, así como no poseer agotamiento como un humano, la hace muy eficaz en los distintos ámbitos de la vida.

Algunas de las aplicaciones de la IA son las siguientes:

- Reconocimiento de imágenes estáticas, clasificación y etiquetado.
- Mejoras del desempeño de la estrategia algorítmica comercial.
- Procesamiento eficiente y escalable de datos de pacientes.
- Mantenimiento predictivo.
- Detección y clasificación de objetos.
- Distribución de contenido en las redes sociales.
- Protección contra amenazas de ciberseguridad.

0.2.1. Big Data

Proceso de recopilación, almacenamiento, procesamiento y analítica de un volumen masivo de datos que las organizaciones pueden implementar en su actividad, de una manera tanto externa, enfocando su objetivo en el aumento de beneficios mediante la captación de clientes, como interna, enfocando su objetivo a la mejora de la eficiencia en su actividad mediante sus empleados y técnicas de trabajo. (Serrano Muñoz, 2023)

Hay que tener en cuenta que Big Data no es lo mismo que datos de gran volumen y estructurados almacenados en bases de datos. Estos datos se han ido almacenando a lo largo de décadas a diferencia que la Big Data, donde los datos se van generando a cada segundo y deben ser tratados con aplicaciones informáticas no convencionales, extrayendo así información que tienen los datos logrando obtener patrones y tendencias para poder realizar una predicción acertada.

0.2.2. Aprendizaje automático

Es una rama de la Inteligencia Artificial que se encarga de generar algoritmos que tienen la capacidad de aprender y no tener que programarlos de manera explícita. (Sandoval Serrano y cols., 2018)

Lo único de lo que se tiene que preocupar el desarrollador es de tener una base de entrenamiento lo suficientemente adecuada como para no caer en un subajuste o sobre ajuste para que el algoritmo aprenda bien y sepa dar soluciones a futuros problemas de manera autónoma.

0.2.3. Tipos de Aprendizaje Automático

Existen diferentes tipos o técnicas de aprendizaje automático entre los que podemos encontrar:

- Aprendizaje supervisado
- Aprendizaje no supervisado
- Aprendizaje semisupervisado
- Aprendizaje por refuerzo

Aprendizaje supervisado

Es cuando se entrena al algoritmo otorgándole las preguntas, denominadas características, y las respuestas, denominadas etiquetas. (Candia Oviedo, 2019) Pudiendo realizar predicciones solo conociendo las características. EL aprendizaje supervisado posee dos algoritmos:

- Regresión: Dibuja una recta sobre la que se indicará la tendencia de los datos sean continuos o discretos, si fuese el caso, utilizándose para este último regresión logística.

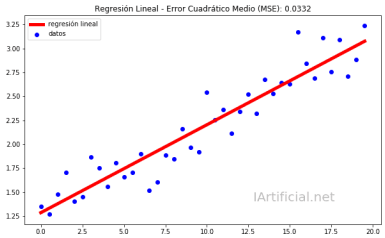

Figura 1: Algoritmo de Regresión, Fuente: IArtificial.net

- Clasificación: Clasifica los datos de acuerdo a patrones, teniendo asi diferentes grupos con datos que presentan similitud entre sí pudiendo predecir la agrupación de datos con características similares. Los valores deben ser discretos.

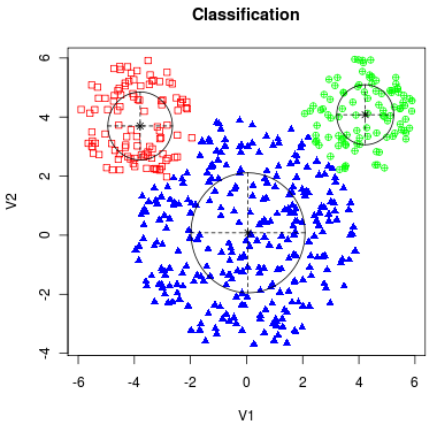

Figura 2: Algoritmo de clasificación, Fuente: Fsancho

0.2.4. Aprendizaje no supervisado

El aprendizaje no supervisado consiste en que la red descubra por si misma características, regularidades, correlaciones o categorías en los datos de entrada y se obtengan de forma codificada en la salida. (Candia Oviedo, 2019)

El algoritmo debe ser capaz de agrupar de acuerdo a las características de los datos, es decir similitud y correlación que se encuentran presentes en estos, para ello hace uso del clustering.

- Clustering: consiste en agrupar un conjunto de objetos (no etiquetados) en subconjuntos de objetos llamados Clusters. Cada Cluster está formado por una colección de objetos que son similares entre sí, pero que son distintos respecto a los objetos de otros Clusters. (Franco Martín y cols., 2021)

0.2.5. Aprendizaje semisupervisado

Es una técnica de aprendizaje automático en que se hace uso de datos etiquetados y no etiquetados, busca aprovechar la combinación de métodos supervisados y no supervisados para mejorar el comportamiento de aprendizaje. (Brusil Cruz, 2020) Se divide en dos tipos:

- Aprendizaje transductivo: Predice los valores de la variable objetivo del conjunto test. El conjunto se divide en entrenamiento y test, en entrenamiento la variable objetivo es conocida, mientras que en test no está dada.
- Aprendizaje inductivo: Obtiene una función de predicción usando los datos de las variables objetivo conocidas y las que no.

0.2.6. Aprendizaje por refuerzo

En este tipo de aprendizaje un agente trata de aprender un comportamiento mediante interacciones de prueba y error en un ambiente dinámico e incierto. (Morales y González, 2012)

0.2.7. Validación cruzada

Murillo et al (2019), mencionan que la estrategia de Validación Cruzada consiste en la división de un conjunto de muestras que se pueden analizar en dos conjuntos disjuntos de datos. Uno de estos conjuntos entrenará las muestras que contiene, y los resultados obtenidos se aplicarán al otro conjunto que será utilizado para la clasificación de muestras.(Murillo y cols., 2019)

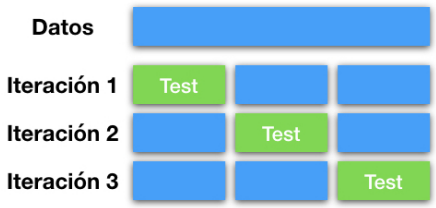

Figura 3: Validación cruzada, Fuente: Analytics Lane

Esto se repite con varias iteraciones, en donde, se calculará la media aritmética de los errores generados en las diferentes medidas de evaluación diferentes.

Se pueden encontrar diferentes tipos de validaciones cruzadas como: k iteraciones, aleatoria, dejando uno fuera.

- **K iteraciones:** También conocido como k-fold cross-validation, consiste en dividir los datos de muestra en k subconjuntos . Uno de los subconjuntos se utiliza como datos de prueba y el resto (K-1) como datos de entrenamiento. El proceso de validación cruzada se repite durante K iteraciones, con cada uno de los posibles subconjuntos de datos de prueba. Finalmente, se realiza la media aritmética de los resultados de cada iteración para obtener un único resultado.
- **Aleatoria:** O rando cross-validation, divide aleatoriamente los datos de muestra en conjuntos de entrenamiento y prueba. Este proceso se repite varias veces y se calcula la media aritmética de los resultados para obtener un único resultado.
- **Dejando uno fuera:** leave-one-out cross-validation, utiliza un solo punto de datos como conjunto de prueba y el resto como conjunto de entrenamiento.

Sin embargo, también es evidente que todo este proceso tiene desventajas presentes, dentro de las cuales se encuentran las siguientes:

- Computacional mente costosa.
- Inestable si el conjunto de datos es pequeño.
- Sesgada si el conjunto de datos no es representativo.

Por otro lado, presenta las siguientes ventajas:

- Proporciona una estimación más precisa del rendimiento del modelo.
- Utiliza todos los datos disponibles para el entrenamiento y la prueba.
- Útil para conjuntos de datos pequeños.

0.2.8. Análisis de datos: variables y tipos

El análisis de datos es la etapa o proceso en el que se limpia el dataset, se transforman los datos para poder obtener información útil que ayuden con las conclusiones y toma de decisiones. Es aquí donde se encuentran las variables y sus tipos, estas deben ser elegidas de una manera adecuada para tener un buen entrenamiento.

Una variable, no es más que una característica o atributo que medir y describe al fenómeno que se esta estudiando,es utilizada en el entrenamiento y construcción de un modelo.

Se puede encontrar distintos tipos de variables, de acuerdo a la función de su naturaleza o forma de medición.

- **Independientes:** Se utilizan como entradas para el modelo, es decir son las etiquetas o ejemplos sobre las cuales el modelo tendrá que predecir datos nuevos.
- **Dependiente:** variable que el modelo intenta predecir o clasificar.
- **Categóricas:** representan categorías o etiquetas discretas, siendo de tipo nominal (sin orden) u ordinales (ordenadas).

- **Numéricas:** pueden ser continuas, cualquier valor presente dentro de un rango,o discretas, valores separados.
- **Textuales:** características en forma de textos.
- **Temporales:** aquellas variables que se encuentran relacionadas con el tiempo.
- **Imagen o audio:** en este caso, se hace referencia a la representación numérica de estas formas de datos, es decir, su forma binaria.
- **Derivadas o transformadas:** variables que se crean a partir de las originales cuando se realizan cálculos matemáticos o transformaciones que quieren capturar patrones, presentes pero complejos, en los datos.

Otro aspecto a tomar en cuenta es la escala de las variables, refiriéndose al rango de valores que esta toma en un conjunto de datos. Tiene importancia, ya que aporta con un impacto relevante en el rendimiento e interpretación de los modelos de machine learning. Esos impactos son los siguientes:

- **Normalización y estandarización:** Es recomendable realizar este proceso para que todas las variables presenten la misma escala, permitiendo a los algoritmos que las variables que presentan escalas grandes no dominen las predicciones y todo sea equilibrado.
- **Interpretación de coeficientes:** las escalas de las variables afecta la interpretación de los coeficientes que no son sencillamente comparables.
- **Visualización:** estas escalas también influyen en la apariencia e interpretación de gráficos.
- **Eficiencia computacional:** la escala adecuada de las variables mejora el proceso de entrenamiento del modelo, permitiendo la convergencia mas rápida del algoritmo.

0.3. Selección de variables

La selección de variables es un proceso fundamental en el aprendizaje automático, ya que permite identificar las características más relevantes de un conjunto de datos para construir modelos predictivos precisos y eficientes.(Solís-Salazar y Madrigal-Sanabria, 2022)

En este proceso, se busca identificar las variables que tienen una mayor influencia en la variable objetivo, eliminando aquellas que no aportan información relevante o que pueden generar ruido en el modelo. Existen diferentes técnicas para la selección de variables, que se pueden clasificar en dos categorías principales: métodos basados en filtros y métodos basados en wrappers.(Martín-Rodríguez y cols., 2020)

Los métodos basados en filtros se basan en medidas estadísticas para evaluar la relevancia de las variables, como la correlación o la información mutua.

Estos métodos son rápidos y eficientes, pero no tienen en cuenta la interacción entre las variables y el modelo predictivo.

Por otro lado, los métodos basados en wrappers utilizan un modelo predictivo para evaluar la relevancia de las variables, seleccionando aquellas que mejoran

el rendimiento del modelo. Estos métodos son más precisos, pero también más costosos computacionalmente.

Entre las técnicas más comunes de selección de variables se encuentran:

- **Análisis de componentes principales (PCA):** Esta técnica se utiliza para reducir la dimensionalidad de los datos, identificando las variables que explican la mayor parte de la variabilidad en los datos. PCA es útil cuando se trabaja con conjuntos de datos con muchas variables, ya que permite reducir el número de variables sin perder información relevante.
- **Regresión Lasso:** Esta técnica utiliza una función de penalización para reducir el número de variables en un modelo de regresión. Lasso es útil cuando se trabaja con conjuntos de datos con muchas variables y se busca construir un modelo de regresión parsimonioso, es decir, con pocas variables explicativa.
- **Árboles de decisión:** Esta técnica se utiliza para identificar las variables más importantes en un conjunto de datos, dividiendo los datos en subconjuntos cada vez más homogéneos. Los árboles de decisión son útiles cuando se trabaja con conjuntos de datos con variables categóricas o discretas.
- **Random Forest:** Esta técnica es una extensión de los árboles de decisión, que utiliza múltiples árboles para mejorar la precisión del modelo. Random Forest es útil cuando se trabaja con conjuntos de datos con muchas variables y se busca construir un modelo preciso y robusto.

Selección de variables por regresión lineal

La selección de variables por regresión lineal es una técnica utilizada en el aprendizaje automático para identificar las variables más relevantes en un conjunto de datos y construir modelos predictivos precisos y eficientes.(Ruiz y Irais, 2018)

En este proceso, se utiliza la regresión lineal para evaluar la relación entre cada variable independiente y la variable dependiente, y se seleccionan aquellas variables que tienen una mayor influencia en la variable objetivo.

Para llevar a cabo la selección de variables por regresión lineal, se pueden utilizar diferentes técnicas, como la regresión lineal simple, la regresión lineal múltiple y la regresión logística. En la regresión lineal simple, se evalúa la relación entre una variable independiente y la variable dependiente, mientras que en la regresión lineal múltiple se evalúa la relación entre varias variables independientes y la variable dependiente. Por otro lado, la regresión logística se utiliza cuando la variable dependiente es categórica.

Entre las ventajas de la selección de variables por regresión lineal se encuentran su simplicidad y eficiencia computacional, así como su capacidad para identificar las variables más relevantes en un conjunto de datos. Sin embargo, esta técnica puede presentar limitaciones cuando las variables independientes están altamente correlacionadas, lo que puede generar problemas de multicolinealidad.

Selección de variables por Arboles Randómicos

La selección de variables por árbol random forest es una técnica utilizada en el aprendizaje automático para identificar las variables más relevantes en un conjunto de datos y construir modelos predictivos precisos y eficientes.(Guerrero-Muguerza, 2020)

Esta técnica se basa en la construcción de múltiples árboles de decisión, que se combinan para mejorar la precisión del modelo. Para llevar a cabo la selección de variables por árbol random forest, se sigue el siguiente proceso:

- Se divide el conjunto de datos en un conjunto de entrenamiento y un conjunto de prueba.
- Se construyen múltiples árboles de decisión utilizando diferentes subconjuntos de variables y observaciones del conjunto de entrenamiento.
- Se evalúa la precisión de cada árbol utilizando el conjunto de prueba.
- Se combinan los árboles para mejorar la precisión del modelo final.

Entre las ventajas de la selección de variables por árbol random forest se encuentran su capacidad para manejar conjuntos de datos con muchas variables y su capacidad para identificar interacciones no lineales entre las variables. Además, esta técnica es resistente al ruido y a los valores atípicos en los datos.

Entre las limitaciones de la selección de variables por árbol random forest se encuentran su costo computacional y su dificultad para interpretar los resultados. Además, esta técnica puede presentar problemas de sobre ajuste si se utiliza un número excesivo de árboles.

0.4. Imputación de datos

Imputación de datos en machine learning, o data imputation en inglés, es el proceso de reemplazar valores faltantes en un conjunto de datos con valores estimados. Este proceso es importante en machine learning porque muchos algoritmos no pueden manejar valores faltantes y pueden producir resultados incorrectos si se les da un conjunto de datos incompleto.

. La calidad de la imputación de datos puede afectar significativamente la precisión de los modelos de machine learning, por lo que es importante elegir el método de imputación adecuado para el conjunto de datos en cuestión. análisis de la varianza por las medidas de dispersión centraletc etc nosotros analizamos las variables con valores ausentes para hacer la imputación antes de utilizar en el método de Boruta

0.4.1. medidas de dispersión

Las medidas de dispersión proporcionan detalles acerca de la variabilidad de una variable y buscan condensar en un solo valor la extensión de la dispersión que se presenta en un conjunto de datos. Las medidas de dispersión más comunes incluyen el rango de variación, la varianza, la desviación estándar y el coeficiente de variación.Ricardi (2011)

0.4.2. medidas de tendencia central

En estadística, se utilizan medidas de tendencia central para resumir un conjunto de valores en un solo valor que represente un centro en torno al cual se encuentran ubicados los datos. Las medidas más comunes son la media, mediana y moda. Por otro lado, existen medidas de dispersión que miden el grado de variación de los valores de la variable, es decir, evalúan en qué medida los datos difieren entre sí. En conjunto, estas medidas permiten describir la posición de un conjunto de datos(Quevedo, 2011).

En nuestro trabajo consideramos lo siguientes:

- Los indicadores pueden afectar en la relación
- Algunas de las variable en estudio, por lo que no causaría la afectación en ello
- Podemos detallar algunos de los algoritmos, puede variar en el desarrollo
- Las variables pueden dar varios factores como lo son en valores atípicos. Así no se usaría la varibale prua, sino un estimado reducido de la variable.

0.5. Mortalidad hospitalaria

La mortalidad hospitalaria es un tema importante en la atención médica, ya que se refiere a la cantidad de pacientes que fallecen durante su estancia en el hospital. Existen varios estudios que se han enfocado en analizar la mortalidad en diferentes patologías.(Rosales y Adoración, 2019)

Por ejemplo, un estudio se enfocó en analizar la mortalidad en pacientes con hemorragia digestiva alta, encontrando que presentar un episodio de hemorragia digestiva alta supone un exceso de mortalidad que va más allá del ingreso hospitalario. Otro estudio se enfocó en analizar los factores de riesgo asociados a la mortalidad neonatal en un hospital de Nicaragua, encontrando que la edad gestacional, el peso del recién nacido y la presencia de patologías en el recién nacido son factores de riesgo para la mortalidad neonatal.

La sepsis es otra patología que ha sido objeto de estudio en relación a la mortalidad hospitalaria. Un estudio se enfocó en evaluar el impacto del asesoramiento dietético en pacientes que han tenido un ingreso hospitalario y que presentan desnutrición, encontrando que el grupo intervención que recibió asesoramiento dietético aumentó de peso, mientras que los controles perdieron peso. Es importante mencionar que la mortalidad hospitalaria no solo se refiere a la cantidad de pacientes que fallecen durante su estancia en el hospital, sino también a la mortalidad diferida, que se refiere a la cantidad de pacientes que fallecen después de haber sido dados de alta.

Un estudio analizó la mortalidad diferida a 6 meses en pacientes con hemorragia digestiva alta, encontrando que la mortalidad secundaria a hemorragia digestiva puede ocurrir hasta 6 meses después del alta hospitalaria. La mortalidad hospitalaria es un tema que preocupa a los profesionales de la salud, ya que se busca brindar una atención médica de calidad que permita salvar la vida de los pacientes. Es por ello que se han llevado a cabo diversos estudios para identificar los factores de riesgo asociados a la mortalidad en diferentes patologías,

con el objetivo de identificar aquellos pacientes susceptibles de beneficiarse de un tratamiento y seguimiento más estrecho.

Es importante mencionar que la mortalidad hospitalaria no solo se refiere a la calidad de atención médica, sino también a factores sociales y económicos que pueden influir en la salud de los pacientes. Por ejemplo, un estudio analizó la influencia de la consulta previa con empíricos sobre la morbilidad y mortalidad de niños internados por infección respiratoria baja o diarrea aguda, encontrando que la gravedad al ingreso fue mayor en pacientes que consultaron con el curandero y con el farmacéutico.

0.5.1. Mortalidad hospitalaria en el Ecuador

A continuación, se presentan algunos hallazgos relevantes:

- Las enfermedades transmitidas por el agua siguen siendo un problema de salud pública importante en Ecuador, con poblaciones indígenas teniendo una mayor probabilidad de enfermarse y morir debido a estas enfermedades que otros grupos étnicos.
- La pandemia de COVID-19 ha causado una mortalidad excesiva en Ecuador, con la tasa de mortalidad superando hasta el 300 por ciento de las muertes oficiales por COVID-19.
- Un estudio sobre la mortalidad por sepsis en una unidad de cuidados intensivos (UCI) en Quito-Ecuador encontró que la disfunción de más de dos órganos era un factor de riesgo para la mortalidad por sepsis, mientras que la resucitación temprana exitosa y el tratamiento antibiótico adecuado eran factores protectores.
- La leptospirosis es una enfermedad descuidada en Ecuador, con la morbilidad y mortalidad siendo subestimadas debido a la financiación insuficiente del sector de la salud, la incompetencia política, la falta de liderazgo y una larga crisis económica.
- La mortalidad debido a lesiones por accidentes de tráfico en adultos mayores (60 años o más) es mayor que en aquellos menores de 60 años, con una tendencia creciente en años de vida potencial perdidos (AVPP).
- Un estudio sobre la mortalidad en Ecuador relacionada con factores dietéticos encontró que el grupo con el mayor número de muertes correspondía a enfermedades cardiovasculares seguidas de enfermedades cerebro vasculares, y las provincias amazónicas mostraron menos muertes en relación con otras provincias en Ecuador.

En general, la mortalidad hospitalaria en Ecuador está influenciada por varios factores, como enfermedades transmitidas por el agua, COVID-19, sepsis, leptospirosis, lesiones por accidentes de tráfico y factores dietéticos. Estos hallazgos destacan la necesidad de intervenciones y políticas de salud pública efectivas para abordar estos problemas y mejorar los resultados de salud en Ecuador.

0.5.2. Mortalidad hospitalaria y machine learning

El uso de machine learning en la mortalidad hospitalaria es un tema de gran interés en la industria de la salud. El machine learning es una técnica de inteligencia artificial que permite a las computadoras aprender de los datos y mejorar su rendimiento en tareas específicas. En el contexto de la mortalidad hospitalaria, el machine learning se utiliza para predecir la probabilidad de que un paciente muera durante su estancia en el hospital.

El machine learning se basa en el análisis de grandes cantidades de datos de pacientes, incluyendo su historial médico, resultados de pruebas y otros factores relevantes. A partir de estos datos, los algoritmos de machine learning pueden identificar patrones y correlaciones que pueden ser utilizados para predecir la probabilidad de mortalidad de un paciente.

La predicción de la mortalidad hospitalaria puede ser útil para los médicos y el personal del hospital, ya que les permite identificar a los pacientes que tienen un mayor riesgo de morir y tomar medidas para prevenirlo. Por ejemplo, los médicos pueden ajustar el tratamiento de un paciente o proporcionar una atención más intensiva para reducir el riesgo de mortalidad.

Sin embargo, es importante tener en cuenta que el machine learning no es una solución perfecta para predecir la mortalidad hospitalaria. Los algoritmos de machine learning pueden ser influenciados por factores que no están relacionados con la salud del paciente, como la edad, el género o la raza. Además, los algoritmos pueden no ser capaces de tener en cuenta factores que son difíciles de medir, como la calidad de la atención médica o la experiencia del personal del hospital.

0.6. Defunciones en el Ecuador

En los reportes de cada registro de mortalidad en nuestro país los años 2019 - 2020 podemos verificar con los siguientes datos :

- En año 2019, se registraron 73.431 defunciones generales lo cual nos da un indicio de por cada 3 a 4 por cada 1000 habitantes mueren diariamente

Estos registros al ser proporcionados la magnitud de la demografía nacional en la cual las persona tienen decesos. En este caso los hombres y mujeres, nacionalidades, entre otras características que la conforman

Entre los casos mas comunes y generales que se han dado defunciones a nivel nacional han sido las siguientes:

- Accidentes de Transito
- Asesinato
- Desnutrición
- Defunción Natural
- Enfermedades Crónicas
- Varias

En el año de 2019 a principios de 2020 el mundo vivió una incertidumbre a nivel mundial la cual afecto a muchas personas.

El COVID 2019 fue una de las causas mas comunes de muerte que creció en gran masa en la historia entre 2019 al 2020, muchos de los países se encontraron en zona de riesgo para ello Ecuador se encontró en los países con un mayor numero de muertes.

Para ello en muchas de las investigaciones que se dieron en el desarrollo de informes y expedientes de cada persona fallecida en el Ecuador tiene relevancia con el contexto social.

Una de las problemáticas actuales en Ecuador es la falta de un sistema de datos abiertos sobre la COVID-19 porque la única información a la que puede acceder la ciudadanía y los medios de comunicación son las cifras emitidas diaria-mente por el COE-N y los reportes de defunciones inscritas diarias del Registro Civil(Parra y Carrera, 2021). Esto en conocimiento hacia los ciudadano se dan a entender que la implicación a niveles de mortalidad se han generado varios puntos y picos en donde el país se conmociono y esto hizo que los diferentes casos de mortalidad quedaran atrás ante esta crisis mundial.

Tomando en cuenta todo lo antes mencionado podemos rescatar en este ámbito investigativo, que los estados o los diferentes casos de mortalidad en el Ecuador puede ser un gran impacto en nuestro desarrollo. Para lo que empezamos a generar las siguiente interrogantes:

- Que causas de muerte son mas frecuentes
- El COVID 19 es un punto de inflexión en el estado de defunción en estos años
- Los accidentes de transito alrededor del país toma un gran impacto alrededor del COVID 19

Estas interrogantes las podemos revisar en nuestro desarrollo de investigación en nuestro modelo de predicción

0.7. Aplicación de modelo

A continuación se describe todo lo que se utilizo en el estudio. Para ello usamos la herramienta de Software RStudio, la cual nos ayudara en la predicción y desarrollo de nuestros modelos matemáticos, así también con algoritmos de Aprendizaje Automático.

Dividimos para el desarrollo de nuestro código en los diferentes algoritmos de desarrollo:

- Selección de Variables
    - Imputación de Variables

    - Aplicación de Algoritmo de selección de Variables (Arboles Randómicos)
- Pre - Procesamiento de datos
    - Selección de la población
    - Manejo de datos faltantes (Outliders)
- Modelo Clustering
    - Modelo Som
    - Resultados de la Clusterizacion
- Análisis de Clusters

estas fueron las etapas que se utilizaron para la investigación y manejo de datos.

Librerías utilizadas

Boruta

El algoritmo de Boruta es un contenedor construido alrededor del algoritmo de clasificación de bosques aleatorios. Intenta capturar todas las características importantes e interesantes que podría tener en su conjunto de datos con respecto a una variable de resultado.

En una secuencia mas especifica de su proceso podemos detallarlo de la siguiente manera:

- Primero, duplica el conjunto de datos y mezcla los valores en cada columna. Estos valores se denominan características de sombra.
- Luego, entrena un clasificador, como un clasificador de bosque aleatorio, en el conjunto de datos. Al hacer esto, se asegura de tener una idea de la importancia, a través de la precisión de disminución media o la impureza de disminución media, para cada una de las características de su conjunto de datos.
- Cuanto mayor sea la puntuación, mejor o más importante.

MlBench

MLBench es un marco para el aprendizaje automático distribuido. Su propósito es mejorar la transparencia, la reproducibilidad, la solidez y proporcionar medidas de rendimiento justas, así como implementaciones de referencia, ayudando a la adopción de métodos de aprendizaje automático distribuido

En estos aspectos MlBench tiene dos objetivos principales:

- Ser una suite de evaluación comparativa justa y fácil de usar para algoritmos y sistemas (marcos de software y hardware).
- Para proporcionar implementaciones de referencia confiables y reutilizables de algoritmos de ML distribuidos.

Características

- Para la reproducibilidad y la simplicidad, actualmente nos enfocamos en el aprendizaje automático supervisado estándar , incluidas las tareas estándar de aprendizaje profundo, así como los modelos clásicos de aprendizaje automático lineal.
- Proporcionamos implementaciones de referencia para cada algoritmo y tarea, para facilitar la migración.
- Proporcionan conjuntos de datos y tareas definidas con precisión para tener una comparación justa y precisa de todos los algoritmos, marcos y hardware.
- Fácil de ejecutar en nubes públicas .

Caret

Incluye una serie de funciones que facilitan el uso de decenas de métodos complejos de clasificación y regresión. Utilizar este paquete en lugar de las funciones originales de los métodos presenta dos ventajas:

- Permite utilizar un código unificado para aplicar reglas de clasificación muy distintas, implementadas en diferentes paquetes.
- Es más fácil poner en práctica algunos procedimientos usuales en problemas de clasificación. Por ejemplo, hay funciones específicas para dividir la muestra en datos de entrenamiento y datos de test o para ajustar parámetros mediante validación cruzada.

RandomForest

Hay leyes que exigen que las decisiones tomadas por modelos utilizados en la emisión de préstamos o seguros sean explicables. Esto último se conoce como interpretación del modelo y es una de las razones por las que vemos que los modelos de bosques aleatorios se utilizan sobre otros modelos como las redes neuronales.

El algoritmo de bosque aleatorio funciona mediante la agregación de las predicciones realizadas por múltiples árboles de decisión de profundidad variable. Cada árbol de decisión en el bosque se entrena en un subconjunto del conjunto de datos llamado conjunto de datos de arranque.

La parte de las muestras que quedaron fuera durante la construcción de cada árbol de decisión en el bosque se denomina conjunto de datos fuera de la bolsa

Recordemos cómo a la hora de decidir los criterios con los que dividir un árbol de decisión, medimos la impureza que produce cada característica.

Hmisc

Contiene muchas funciones útiles para el análisis de datos, gráficos de alto nivel, operaciones de servicios públicos, funciones para calcular el tamaño y la potencia de la muestra, simulación, importación y anotación de conjuntos de datos, imputación de valores faltantes, elaboración avanzada de tablas, agrupamiento de variables, manipulación de cadenas de caracteres, conversión de objetos R. a código LaTeX y html, recodificación de variables, almacenamiento en caché, computación paralela simplificada, estimación estadística general de ventana móvil y asistencia en la interpretación del análisis de componentes principales.

Stringi

Le proporciona muchas funciones relacionadas con la limpieza de datos, la extracción de información y el procesamiento del lenguaje natural:

- Concatenación de cadenas, relleno, envoltura y extracción de subcadenas.
- Búsqueda de patrones (p. ej., con expresiones regulares similares a Java de ICU).
- Recopilación, clasificación y clasificación.
- Generación de cadenas aleatorias.
- Transliteración de cadenas, mapeo y plegado de casos.
- Normalización Unicode.
- Formato y análisis de fecha y hora.

Dplyr

Es un versión optimizada de su paquete plyr. El paquete dplyr no proporciona ninguna nueva funcionalidad a R, en el sentido que todo aquello que podemos hacer con dplyr lo podríamos hacer con la sintaxis básica de R.

Una importante contribución del paquete dplyr es que proporciona una "gramática"(particularmente verbos) para la manipulación y operaciones con data frames. Con esta gramática podemos comunicar mediante nuestro código que es lo que estamos haciendo en los data frames a otras personas (asumiendo que conozcan la gramática). Esto es muy útil, ya que proporciona una abstracción que anteriormente no existía.

Modeest

R no dispone de una función en su paquete base que nos permita calcular la moda. La función mode devuelve el tipo o modo de almacenamiento de un objeto. Hay múltiples formas de calcular la moda haciendo uso de otras funciones de R. Sin embargo, ahora optamos por cargar el paquete modeest y usar la función mlv que devuelve el valor de un vector numérico.

Moments

Funciones para calcular:

- Momentos.
- Curtosis de Pearson.
- Curtosis de Geary y asimetría
- Ensayos relacionados con ellos (Anscombe-Glynn, D'Agostino, Bonett-Seier).

Dummy

Crear variables dummy implica transformar datos de un formato "alto", en el que cada columna contiene la información de una variable, a datos con un formato "ancho", en los que múltiples columnas contienen la información de las dos variables, codificada de manera binaria, esto es, con 0 y 1.

Kohonen

Mapas autoorganizados para mapear espectros o patrones de alta dimensión en 2D; Se utiliza la distancia euclidiana. Modelado a partir de la función SOM.

TicToc

Encargado de mantener una secuencia organizada para los procesos llevadoa a cabo.

0.7.1. Ggplot2

ggplot2 es un sistema para crear gráficos declarativamente, basado en The Grammar of Graphics . Usted proporciona los datos, le dice a ggplot2 cómo asignar variables a la estética, qué primitivas gráficas usar y se ocupa de los detalles.

Catools

Contiene varias funciones de utilidades básicas que incluyen: funciones de estadísticas de ventana en movimiento (desplazamiento, ejecución), lectura/escritura de archivos binarios GIF y ENVI, cálculo rápido de AUC, clasificador LogitBoost, codificador/decodificador base64, redondeo de suma sin errores y suma acumulada, etc.

Foreign

Leer y escribir datos almacenados por algunas versiones de 'Epi Info', 'Minitab', 'S', 'SAS', 'SPSS', 'Stata', 'Systat', 'Weka', y para leer y escribir algunos 'dBase 'archivos.

MiscFuncs

Una función para generar plantillas de roxygen para funciones genéricas y métodos asociados.

Selección de Variables

Para nuestro proceso tomamos un dataset que nos provee información sobre defunciones los cuales se dan por las siguientes variables:

- Provincia de inscripción
- Cantón de inscripción
- Parroquia de inscripción
- Año de inscripción
- Mes de inscripción
- Día de inscripción
- Fecha de inscripción
- Nacionalidad
- Código del país
- Sexo
- Fecha de nacimiento
- Año de nacimiento
- Mes de nacimiento
- Día de nacimiento
- Fecha de fallecimiento
- Año de fallecimiento
- Mes de fallecimiento
- Día de fallecimiento
- Código de edad al fallecer
- Edad al fallecer
- Provincia de residencia habitual del fallecido
- Área de residencia habitual del fallecido (a)
- Estado civil y/o conyugal
- Nivel de instrucción alcanzado

- Sabe leer y escribir
- Etnia
- Lugar de ocurrencia del fallecimiento
- Provincia de fallecimiento
- Cantón de fallecimiento
- Parroquia de fallecimiento
- Área de fallecimiento del fallecido (a)
- Certificado por
- Tipo presuntivo de la muerte accidental o violenta
- Lugar donde ocurrió el hecho muerte accidental o violenta
- Causa básica de defunción (categorías - 4 caracteres)
- Se realizó autopsia/necropsia?
- Causa básica de defunción (categorías - 3 caracteres)
- Lista corta de causas de defunción
- Lista condensada (103 causas)
- Lista de tabulación 2 para la mortalidad subcategorías (80 causas)
- Lista condensada (67 causas)A
- Lista condensada desagregada (67 causas)B
- Residente de el País.

Vemos en nuestro detallado de características que posee el dataset, este contiene información variada, pero asi mismo no toda es de utilidad. Además, podemos apreciar que varios campos están como datos vacíos (cuando la pregunta no aplica)

0.7.2. Imputación de Variables

Antes de proceder a la selección de variables, es fundamental llevar a cabo un exhaustivo análisis y descripción de las variables presentes en el conjunto de datos. Este proceso implica examinar en detalle cada variable, comprendiendo su naturaleza, rango de valores, distribución y posibles relaciones con otras variables.

Sin embargo, es común encontrarse con la presencia de valores ausentes o faltantes en algunas de las variables. Estos valores pueden deberse a diversas razones, como errores de medición, problemas en la recopilación de datos o simplemente la falta de información en ciertos casos. La presencia de valores ausentes puede afectar negativamente el rendimiento y la validez de los modelos de aprendizaje automático.

Es necesario realizar una imputación de los valores faltantes. La imputación es un proceso mediante el cual se estima o se reemplazan los valores ausentes por valores plausibles basados en la información disponible en el conjunto de datos. El objetivo principal de la imputación es preservar la integridad y la representatividad de la variable en cuestión, evitando distorsiones o sesgos en el análisis posterior.

Existen varias técnicas de imputación que se pueden utilizar, como la imputación basada en la media, la imputación basada en modelos o la imputación múltiple. Cada técnica tiene sus propias ventajas y consideraciones, y la elección de la técnica adecuada dependerá del contexto y las características del conjunto de datos.

Es importante destacar que la imputación de valores faltantes debe realizarse con cuidado y siguiendo buenas prácticas. Es fundamental tener en cuenta el impacto potencial de la imputación en los resultados del análisis y la interpretación de los datos. Además, es recomendable realizar una evaluación posterior de la calidad de la imputación y considerar la incertidumbre asociada a los valores imputados.

para los datos actuales se se creo algunas funciones que ejecutan un análisis de varianza, claro primero cambiando todos los datos ausentes o faltantes por NA.

```
48 medidas_variabilidad <- function(x, na.rm = FALSE) {
49   lista <- c(format(coef_var(x, na.rm=na.rm), scientific = F),
50              format(var(x, na.rm=na.rm), scientific = F),
51              format(sd(x, na.rm=na.rm), scientific = F),
52              format(range(x,na.rm = na.rm), scientific = F))
53           #-------coeficiente de varianza----varianza --------desviacion estandar------rango
54   lista
55 }
56
57 verificacion <- function(x, y, na.rm = FALSE, dispersion = mean) {
58   y <- impute(x, dispersion)
59   lista <- c(medidas_variabilidad(x, na.rm=T),
60              medidas_variabilidad(y, na.rm=T))
61   matriz <- matrix(lista, nrow = 5, ncol = 2)
62   matriz
63 }
```

Figura 4: Funciones, Fuente: Equipo de Investigación

en estas funciones se verifica el valor de la varianza al imputar con cualquier medida de dispersión.

```r
colSums(is.na(spss))
# segun los valores ausentes se necesita hacer la imputacion de las columnas
#anio_insc   dia_insc edad   cod_edad  nac_fall  etnia
#est_civil   niv_inst sabe_leer   autopsia     lugar_ocur    lug_viol
colSums(is.na(imputados))
#------------------------revision de medidas de variabilidad--------------
#anio_insc
verificacion(spss$anio_insc, imputados$anio_insc, T, mean)

#dia_insc

verificacion(spss$dia_insc, imputados$dia_insc, T, mean)

#edad

verificacion(spss$edad, imputados$edad, T, mean)

#cod_edad

verificacion(spss$cod_edad, imputados$cod_edad, T, mode)

#nac_fall
```

Figura 5: Verificación, Fuente: Equipo de Investigación

una vez definidas podemos utilizarlas para verificar cada columna que contenga datos faltantes, una vez verificadas se comprobó que la mayoría de variables no presentan un incremento en su varianza al realizar la imputación. es decir, que se puede utilizar cualquier método de imputación en este caso, por la media o la moda dependiendo de la variable.

```r
#------------------- IMPUTACION DE DATOS ---------------------------------
imputados$anio_insc <- impute(spss$anio_insc, mean)
imputados$dia_insc <- impute(spss$dia_insc, mean)
imputados$edad <- impute(spss$edad, mean)
imputados$cod_edad <- impute(spss$cod_edad, mode)
imputados$etnia <- impute(spss$etnia, mode)
imputados$est_civil <- impute(spss$est_civil, mode)
imputados$niv_inst <- impute(spss$niv_inst, mode)
imputados$sabe_leer <- impute(spss$sabe_leer, mode)
imputados$autopsia <- impute(spss$autopsia, mode)
imputados$lugar_ocur <- impute(spss$lugar_ocur, mode)
imputados$lug_viol <- impute(spss$lug_viol, mode)
imputados$nac_fall <- impute(spss$nac_fall, mode)

colSums(is.na(imputados))

```

Figura 6: Imputación, Fuente: Equipo de Investigación

al final se imputaron las variables y se guardaron en un CSV para su utilización.

0.7.3. Aplicación de Algoritmo de selección de Variables (Arboles Randómicos)

Se ha optado por emplear el método de random forest (boruta) para llevar a cabo la selección de variables. Este enfoque nos brinda la capacidad de identificar

y elegir las variables más relevantes para nuestro análisis. A continuación, se describe el funcionamiento de este método:

El método de random forest (boruta) es una técnica de selección de variables que se basa en la construcción de múltiples árboles de decisión. En este proceso, se generan árboles aleatorios utilizando diferentes subconjuntos de variables del conjunto de datos original.

Luego, se compara el desempeño de las variables originales con el de las variables aleatorias generadas. Si una variable original tiene un rendimiento significativamente mejor que las variables aleatorias en términos de su importancia en la predicción o clasificación, se considera como una variable relevante.

El método boruta también tiene en cuenta las variables que no alcanzan un rendimiento significativo en comparación con las variables aleatorias. Estas variables se denominan "tentativasz se someten a un proceso de prueba adicional para determinar si son realmente relevantes o no.

```
> bor <- TentativeRoughFix(boruta)
> print(bor)
Boruta performed 14 iterations in 3.350887 hours.
Tentatives roughfixed over the last 14 iterations.
 30 attributes confirmed important: anio_fall, anio_nac, area_fall, area_res, autopsia and 25 more;
 3 attributes confirmed unimportant: anio_insc, mes_nac, prov_insc;
```

Figura 7: Variables, Fuente: Equipo de Investigación

una vez terminado la implementación se selecciono 30 variables como las mas importantes

0.8. Pre - Procesamiento de datos

0.8.1. Selección de la población

Para llevar a cabo la selección de la población de interés, se procedió a cargar el archivo CSV que contenía las variables obtenidas mediante el método de árboles random forest (boruta). A continuación, se aplicó un filtro al conjunto de datos original en función de la variable de provincia de fallecimiento, que es la variable que nos interesa analizar.

Este proceso de filtrado nos permitió seleccionar únicamente los registros del conjunto de datos que corresponden a la provincia de Guayas, lo que nos permitió enfocarnos en la población que nos interesa estudiar. De esta manera, pudimos reducir el tamaño del conjunto de datos y enfocarnos en los registros que son relevantes para nuestro análisis.

```r
#Variable a consideracion
variables <- "variables_seleccionadas.csv"
spss <- read.csv2(spssPoblacion, header = TRUE, sep = ";")
names(spss) <- tolower(names(spss))

var <- read.csv2(variables, header = TRUE, sep = ";")

var$x[26] <- "causa"


#Filtrado - sbudataset
spss_poblacion <- spss[spss$prov_fall == provincia, unlist(
write.csv2(spss_poblacion, salida, row.names = FALSE)
```

Figura 8: Población, Fuente: Equipo de Investigación

Es importante destacar que el proceso de selección de la población debe realizarse cuidadosamente y siguiendo buenas prácticas. Es fundamental tener en cuenta el impacto potencial de la selección de la población en los resultados del análisis y la interpretación de los datos. Además, es recomendable realizar una evaluación posterior de la calidad de la selección de la población y considerar la incertidumbre asociada a los resultados obtenidos.

0.8.2. Manejo de datos faltantes (Outliders)

En el proceso de manejo de outliers, se llevó a cabo la eliminación de varias columnas que se consideraron no relevantes para el análisis. Entre estas columnas se encontraban las fechas, ya que la información de las fechas ya estaba presente en variables separadas, lo que evitaba la duplicación de datos. Además, se decidió eliminar la variable de autopsia debido a que presentaba un alto porcentaje de datos faltantes, superando el 70 por ciento.

La eliminación de estas columnas no importantes permitió simplificar y limpiar el conjunto de datos, eliminando información redundante o con una alta proporción de valores ausentes. Al reducir la cantidad de variables, se facilita el análisis y se evita la introducción de ruido o sesgos en los modelos de aprendizaje automático.

```r
dataset <-replace_with_na_all(dataset, ~.x %in% c("N/A", "missing", "na", " "))

#eliminar valores ausentes
dataset$autopsia <- NULL
dataset$dia_insc <- NULL
dataset$fecha_insc <-NULL
dataset$dia_fall <- NULL
dataset$fecha_fall <- NULL
dataset$dia_nac <- NULL
dataset$fecha_nac <- NULL


row.has.na <- apply(dataset, 1, function(x){
  any(is.na(x))
})

dataset <- dataset[!row.has.na,]
```

Figura 9: Outliers, Fuente: Equipo de Investigación

Es importante tener en cuenta que la eliminación de columnas debe realizarse con precaución y basarse en una evaluación cuidadosa de la relevancia de cada variable para el análisis en cuestión. Es posible que en otros contextos o análisis, las fechas o la variable de autopsia sean consideradas importantes y no se eliminen.

También se eliminaron las filas de datos faltantes que no son relevantes en el estudio mediante una función que recorre las variables buscando NA´s

0.9. Modelo Clustering

0.9.1. Modelo Som

En la investigación se lleva a cabo una técnica de análisis exploratorio para identificar patrones o grupos dentro del conjunto de datos. La técnica de mapas auto-organizativos (SOM) es una técnica de aprendizaje no supervisado que permite visualizar y analizar la estructura de los datos en un espacio de menor dimensión. En este caso, se está utilizando SOM para agrupar las observaciones en clusters basados en sus características y similitudes.

```r
dataset<- dataset_total

dataset <- dummy(dataset)

dataset.sc<- data.matrix(dataset)
set.seed(0)

som_grid <- somgrid(xdim=5,ydim = 5, topo= "rectangular")
tic("Time to run SOM: ") # Inicia el tiempo
som_model <- som(X=dataset.sc, grid=som_grid, rlen = 100, alpha = c(0.05, 0.01), keep.data=TRUE
toc() # Finaliza el tiempo


#generar clusters
clusters <- som_model$unit.classif
conteo <- table(clusters)
conteo_frame <- as.data.frame(conteo)
```

Figura 10: Modelo SOM Código, Fuente: Equipo de Investigación

La conversión de variables categóricas en variables ficticias es un paso importante en el análisis de datos, ya que permite que las variables categóricas sean tratadas como variables numéricas en el modelo. Esto es necesario para que el modelo SOM pueda trabajar con estas variables y agrupar las observaciones en clusters.

La creación de una cuadrícula SOM de 5x5 es una elección específica que se hace en este caso, debido a que una cuadricula mayor no tendría un impacto real en la investigación. La elección del tamaño de la cuadrícula puede afectar el rendimiento del modelo y la interpretación de los resultados. Es importante tener en cuenta que la elección del tamaño de la cuadrícula debe basarse en una evaluación cuidadosa de los datos y los objetivos del análisis.

se procede a graficar el modelo SOM con una paleta de colores especifica.

```
#graficar
#preparar la paleta de colores
colBlueHotRed <- function(n, alpha = 1){
  rainbow(n, end = , alpha = alpha)[n:1]
}
pretty_palette <- c("#1f77b4",'#ff7f0e', '#2ca02c', '#d62728', '#9467bd', '#8c564b', '#e377c2')

colors <- function(n, alpha=1){
  rev(heat.colors(n,alpha))
}
```

Figura 11: Paleta de Color, Fuente: Equipo de Investigación

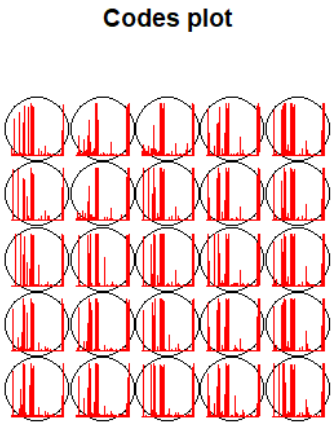

Figura 12: Modelo SOM, Fuente: Equipo de Investigación

Además, es posible visualizar el modelo SOM generado de varias maneras para obtener información relevante para la investigación. La visualización del modelo SOM puede ayudar a identificar patrones y relaciones entre las variables, así como a comprender la estructura de los clusters generados.

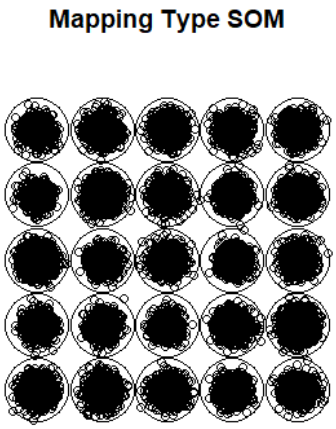

Figura 13: Modelo SOM Distribución, Fuente: Equipo de Investigación

Existen varias técnicas de visualización que se pueden utilizar para representar el modelo SOM, como mapas de calor, gráficos de burbujas y gráficos de densidad. Cada técnica tiene sus propias ventajas y consideraciones, y la elección de la técnica adecuada dependerá del contexto y las características del conjunto de datos.

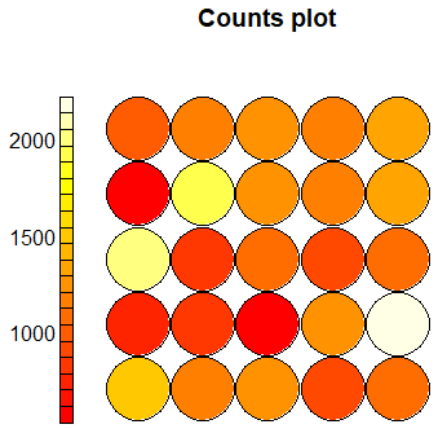

Figura 14: Modelo SOM calor, Fuente: Equipo de Investigación

La visualización del modelo SOM también puede ayudar a identificar los clusters más importantes o relevantes para la investigación. Por ejemplo, se pueden utilizar diferentes colores o etiquetas para resaltar los clusters que contienen un mayor número de observaciones o que presentan características específicas de interés.

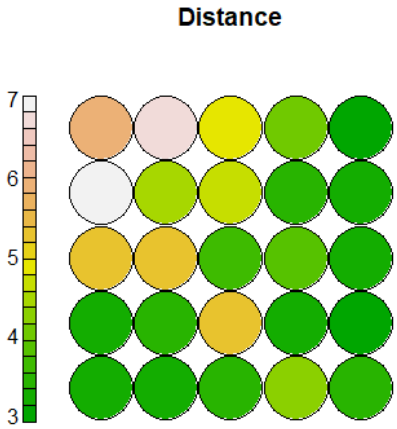

Figura 15: Modelo Distancia, Fuente: Equipo de Investigación

0.9.2. Resultados de la Clusterización

Para obtener los resultados del análisis de agrupamiento, se llevó a cabo un análisis del codo para determinar la cantidad óptima de clusters finales que se encuentran relacionados. El análisis del codo es una técnica común utilizada en el análisis de agrupamiento para determinar el número óptimo de clusters que deben utilizarse en el análisis.

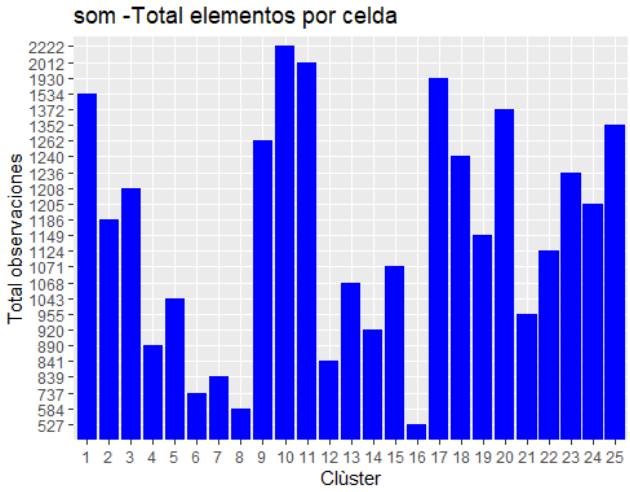

Figura 16: elementos, Fuente: Equipo de Investigación

En este caso, se realizó un análisis del codo para determinar la cantidad de clusters finales que se deben utilizar en el análisis de agrupamiento. El análisis del codo implica graficar la suma de las distancias cuadradas intra-cluster en función del número de clusters. El punto de inflexión en la curva se utiliza como

una indicación del número óptimo de clusters.

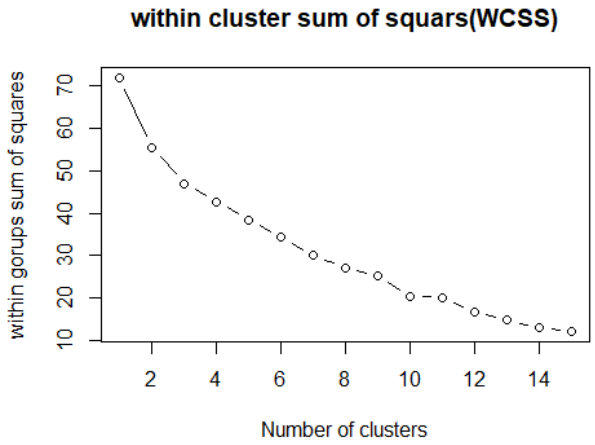

Figura 17: Análisis del CODO, Fuente: Equipo de Investigación

Gracias a este análisis, se obtuvieron 10 clusters finales que se pueden utilizar en estudios más profundos. La elección de 10 clusters se basó en la identificación del punto de inflexión en la curva del análisis del codo. Este número de clusters puede proporcionar una buena representación de la estructura de los datos y permitir la identificación de patrones y relaciones entre las variables.

podemos analizar los clusters en otras herramientas como Rapid Miner para determinar mas relaciones.

0.10. Resultados

0.10.1. Conclusiones

- En conclusión, del análisis de los resultados obtenidos del entrenamiento del modelo desarrollado, se pudo observar que la causa principal de muerte en los hospitales durante el año 2020 fue el COVID-19. Esta enfermedad afectó principalmente a personas de entre 50 y 65 años, lo que evidencia la gravedad de la situación para esta población. Además, se observó que la capacidad hospitalaria en la provincia de Guayas colapsó, lo que llevó a que muchas personas fallecieran en sus hogares o en subcentros de salud.

- En el momento del análisis, tener una predicción para el problema no era útil debido a que el virus era nuevo y la emergencia sanitaria a nivel mundial era desconocida. Sin embargo, se evidenció la falta de preparación del Ministerio de Salud y del sector hospitalario para enfrentar una emergencia de esta magnitud. En la actualidad, las predicciones pueden ser utilizadas para disminuir la mortalidad hospitalaria en función del caso que se esté analizando.

0.10.2. Recomendaciones

- Como recomendaciones, se destaca la importancia del manejo adecuado de datos faltantes y outliers en el desarrollo de modelos de aprendizaje automático. Un buen manejo y tratamiento de estos datos puede mejorar significativamente el rendimiento de los modelos, al reducir el ruido y la complejidad en el entrenamiento.

- Además, es fundamental elegir los algoritmos adecuados para el modelo, teniendo en cuenta las necesidades, requerimientos y el problema que se esté abordando. La elección de algoritmos precisos y adecuados puede mejorar la precisión de los resultados, lo que a su vez puede ayudar en la toma de decisiones y en la realización de predicciones precisas.

- Se recomienda identificar los patrones que se presenten en los resultados y actuar sobre ellos para mejorar la calidad de atención médica hospitalaria. Es importante evaluar si las soluciones implementadas respondieron adecuadamente a la problemática planteada y realizar ajustes si es necesario.

- Por último, se destaca la importancia de utilizar algoritmos de aprendizaje supervisado para mejorar el análisis de los resultados obtenidos de los algoritmos de aprendizaje no supervisado. Los algoritmos de aprendizaje supervisado pueden proporcionar una mayor precisión y control en la identificación de patrones y relaciones entre las variables, lo que puede mejorar la calidad y la interpretación de los resultados.

Referencias

Brusil Cruz, C. A. (2020). *Análisis comparativo entre aprendizaje supervisado y aprendizaje semi-supervisado para la clasificación de señales sísmicas vulcanológicas del volcán cotopaxi* (B.S. thesis). Quito, 2020.

Candia Oviedo, D. I. (2019). Predicción del rendimiento académico de los estudiantes de la unsaac a partir de sus datos de ingreso utilizando algoritmos de aprendizaje automático.

Carrera, S., y Llumiquinga, R. (2020). Registro estadístico de defunciones generales. *Quito: Instituto Nacional de Estadística y Censos (INEC). Recuperado el Enero de*.

Congacha, A. E., BARBA BRITO, J., PALACIOS PACHECO, L., y Delgado, J. (s.f.). *Caracterización de los siniestros viales en el ecuador. novasinergia [online]. 2019, vol. 2, n. 2.* Epub.

Franco Martín, P., y cols. (2021). Métodos de aprendizaje automático aplicados al desarrollo de la bioinformática.

Guerrero-Muguerza, A. M. (2020). Comparación entre regresión logística y random forest para determinación de factores de violencia de pareja en el perú. *Innovando la educación en tecnología. Actas del II Congreso Internacional de Ingeniería de Sistemas*. Descargado de `https://api.semanticscholar.org/CorpusID:247825357`

Martín-Rodríguez, F., Sanz-García, A., Moreno, L. P. M., del Pozo Vegas, C., Martín-Conty, J. L., Villamor, M. A. C., ... Rabbione, G. J. O. (2020). Modelo de riesgo de mortalidad precoz en pacientes ancianos con enfermedad aguda atendidos por servicios de emergencias prehospitalarias. *Emergencias*, *32*, 177-184. Descargado de `https://api.semanticscholar.org/CorpusID:219511745`

Morales, E., y González, J. (2012). Aprendizaje por refuerzo. *Presentacion En Linea en: https://ccc. inaoep. mx/˜ emorales/Cursos/Aprendizaje2/Acetatos/refuerzo. pdf*.

Murillo, R., Arcedalia, N., y cols. (2019). Análisis de validación cruzada bajo diferentes condiciones de ruido.

Parra, M., y Carrera, E. (2021). Evolución de la covid-19 en ecuador. *Investigación y Desarrollo*, *13*(1), 27–40.

Quevedo, F. (2011). Medidas de tendencia central y dispersión. *Medwave*, *11*(03).

Ricardi, F. Q. (2011). Medidas de tendencia central y dispersión. *Revista Biomédica Revisada Por Pares*, 1–8.

Rosales, J. C., y Adoración, R. (2019). Mortalidad intrahospitalaria y diferida en hemorragia digestiva alta. análisis de factores pronósticos en una serie prospectiva.. Descargado de `https://api.semanticscholar.org/CorpusID:155765950`

Rouhiainen, L. (2018). Inteligencia artificial. *Madrid: Alienta Editorial*.

Ruiz, L., y Irais, C. (2018). Selección de híbridos en chiles jalapeños, anchos y serranos para el bajío mexicano auxiliados con indices de selección.. Descargado de `https://api.semanticscholar.org/CorpusID:165688196`

Sandoval Serrano, L. J., y cols. (2018). Algoritmos de aprendizaje automático para análisis y predicción de datos. *Revista Tecnológica; no. 11*.

Serrano Muñoz, M. (2023). Big data y los efectos de su implementación en las empresas. estudio de caso: Vicrila industrias de vidrio slu.

Solís-Salazar, M., y Madrigal-Sanabria, J. (2022). A machine learning proposal to predict poverty.. Descargado de `https://api.semanticscholar.org/CorpusID:252680869`