Predicción de causas de defunción en Guatemala en base a datos demográficos de la población

Irving Fabricio Morales Acosta, Oscar Ignacio Escriba Rodas y Ricardo Andrés Chuy Morales

Departamento de Ingeniería en Ciencias de la Computación y Tecnologías de la información, Universidad del Valle de Guatemala, Guatemala. 2025

RESUMEN:

Este trabajo examinó la mortalidad en Guatemala a lo largo de mes, día y año, cumpliendo los objetivos de identificar patrones temporales y etarios mediante análisis exploratorio y clustering en tres grupos (jóvenes, adultos y adultos mayores). La minería de reglas con Apriori reveló asociaciones demográficas clave —como la vinculación de la baja escolaridad con causas externas de muerte y la relación entre sexo y lugar de ocurrencia—, enriqueciendo la segmentación etaria con información para intervenciones foca21lizadas. En la fase de modelos supervisados, Random Forest obtuvo el mejor desempeño con 56.63 % de precisión, superando ampliamente la línea base aleatoria (20 %) pero sin alcanzar el umbral del 60 % planteado, lo que subraya la necesidad de incorporar variables clínicas y técnicas de balanceo avanzadas para mejorar la discriminación de causas crónicas y neoplásicas. Las limitaciones incluyen solapamiento en los clusters (baja silueta) y falta de datos médicos, mientras que las recomendaciones apuntan a enriquecer el dataset, explorar ensamblados más sofisticados y realizar análisis departamentales para diseñar alertas tempranas y políticas de salud pública más precisas.

PALABRAS CLAVE: clasificación, causas de defunción, aprendizaje automático, random forest, regresión logística multinomial, KNN, a priori, datos demográficos, salud pública, Guatemala.

Prediction of Cause of Death in Guatemala based on the population's demographic data.

ABSTRACT

This study explored mortality patterns in Guatemala across months, days, and years, achieving its goals of uncovering temporal and age-related trends through exploratory analysis and clustering into three cohorts (young, adult, and elderly) . Apriori rule mining uncovered key demographic associations—such as the link between low educational level and external causes

of death, and the relationship between sex and location of occurrence—complementing the age segmentation with insights for targeted interventions . In supervised modeling, Random Forest led with 56.63 % accuracy, well above the 20 % random baseline but below the 60 % target, highlighting the need for clinical variables and advanced balancing techniques to better distinguish chronic versus neoplastic causes . Limitations include cluster overlap (low silhouette) and the absence of medical data; recommendations involve enriching the dataset, exploring more advanced ensemble methods, and conducting department-level analyses to develop early-warning systems and more precise public health policies.

KEY WORDS: classification, causes of death, machine learning, random forest, multinomial logistic regression, KNN, demographic data, public health.

Introducción

En Guatemala, la optimización de los sistemas de salud pública y la prevención de enfermedades dependen críticamente de datos precisos y oportunos sobre las causas de muerte. Sin embargo, los análisis tradicionales se han limitado a menudo a cifras anuales agregadas. De tal forma que se omiten patrones temporales (meses, días, temporadas) y factores demográficos clave como la edad, sexo, escolaridad y estado civil. Esta limitación restringe la identificación de grupos vulnerables y períodos críticos, obstaculizando la implementación de intervenciones preventivas. A pesar de contar con registros detallados de defunciones desde 2009, la exploración detallada de factores temporales y demográficas es aún escasa, impidiendo una comprensión profunda de los patrones de mortalidad.

Por lo tanto, este estudio aborda la necesidad de un análisis exploratorio que permita investigar simultáneamente la influencia de las características demográficas (edad, sexo, escolaridad y estado civil) y los patrones temporales (anuales, mensuales y diarios) en la mortalidad general y sus principales causas en Guatemala, durante el período 2009-2022. La investigación se enfoca en descubrir patrones emergentes no evidentes en análisis convencionales, identificando característica que vinculan a las personas con un tipo de causa de muerte.

Problema científico

¿De qué manera las características demográficas (edad, sexo, escolaridad y estado civil) y los patrones temporales (anuales, mensuales y diarios) influyen en la mortalidad general y en las principales causas de muerte en Guatemala, durante el período comprendido entre 2009 y 2022?

Objetivos

El objetivo general de este artículo es analizar la relación entre los patrones temporales y las características demográficas con la mortalidad general y sus principales causas en Guatemala usando datos del año 2009 hasta el 2022. Con el fin de identificar periodos críticos y grupos poblacionales más vulnerables. Esto se logra con el desarrollo de modelos de aprendizaje automático capaces de predecir la causa macro de defunción de un individuo basándose en sus características.

Para lograrlo, se plantean los siguientes objetivos específicos:

1. Determinar la existencia de variaciones significativas en las defunciones totales y por causas específicas según mes, día o año, durante el período de estudio (2009–2022).

2. Identificar grupos de población especialmente vulnerables (según edad, sexo, escolaridad y estado civil) frente a las variaciones temporales en la mortalidad, para establecer períodos críticos específicos para estos grupos.

3. Explorar la interacción entre características demográficas y causas de defunción con la ayuda del algoritmo A Priori, destacando patrones en la población que puedan informar políticas públicas de prevención y respuesta oportuna.

4. Implementar y evaluar modelos de clasificación de aprendizaje automático (K-Nearest Neighbors, Random Forest, y Regresión Logística) para predecir la causa macro de defunción con una precisión mínima del 60%.

Marco teórico (Materiales y métodos)

El presente estudio busca profundizar en la relación entre las características demográficas, como la edad, el sexo, la escolaridad, el estado civil y los patrones temporales (estacionales, mensuales y diarios) con la mortalidad general y las principales causas de muerte en Guatemala, además de tener en cuenta también el tipo de atención recibida y lugar de ocurrencia del fallecido, entre otros. Esto por medio de la realización de modelos de aprendizaje automático con la capacidad de clasificar a personas que han fallecido en su respectivo tipo de causa de defunción.

Lenguajes de Programación y Herramientas

Para la manipulación, el análisis y la modelización de los datos, se emplearon los lenguajes de programación R y Python. En el caso de R, este se utilizó para la descarga inicial y la unificación de los archivos .sav, así como para el análisis exploratorio de datos (EDA) y el entrenamiento y visualización de resultados de los modelos de clasificación realizados. El

lenguaje es una herramienta mundialmente conocida por su utilidad en el análisis, minería de datos y aprendizaje automático. Esta herramienta fue fundamental en todas las etapas del proyecto y específicamente utilizamos RStudio para poder agilizar el desarrollo y trabajo colaborativo mediante el controlador de versiones de git y github.

El uso de Python fue un poco más limitado, pero de igual manera bastante útil y práctico. Principalmente se usó para el preprocesamiento de datos valores vacíos en las variables y también el manejo de valores NA dentro del set de datos. Utilizando principalmente la librería de pandas.

Se utilizaron datos del Instituto Nacional de Estadística (INE) para el período comprendido entre 2009 y 2022. Estos registros se descargaron inicialmente en formato .sav. Dado que la uniformidad de las variables no era completa a lo largo de todos los años, se realizó un proceso exhaustivo de análisis para identificar las variables presentes de manera consistente en todo el período. Finalmente, se consolidó un conjunto de datos único en formato csv, llegando a ser un set de datos con aproximadamente 1,000,000 de observaciones y 19 variables.

Transformaciones en los datos

Una vez obtenido el set de datos completo, aun fueron necesarias algunas transformaciones para poder realizar predicciones utilizando distintas técnicas. Dentro de cada observación del set datos, se encuentra una variable llamada “Caudef”, la cual represente la causa de defunción del apersona utilizando el código CIE-10, el cuál representa la codificación de diagnóstico de morbilidad y mortalidad (INE, 2017).

El hecho es que hay cientos de causas de defunción con dicho código. Por lo que se decidió agrupar las causas de defunción en 5 clases distintas en base propio código, agrupando por la primera letra (la cuál justamente representa al tipo/grupo) del código CIE-10. En el proceso de clasificación de los códigos de defunción, se establecieron 5 categorías macro: Crónicas, Infecciosas, Causas Externas y Neoplasias) y “Síntomas y causas mal definidas”, el resto de letras con baja frecuencia se descartó. La categoría “Crónicas” comprendía la mayor cantidad de casos (52.27%), seguida de “Causas externas” (16.67%), “Neoplásicas” (12.60%), “Síntomas y causas mal definidas” (12.45%) e “Infecciosas” (6.02%).

Posteriormente a dichas categorías se les hizo un downsampling, una técnica de balanceo de datos que consiste en reducir la cantidad de observaciones de las clases desbalanceadas a la clase con menor cantidad de registros. Dejando a todos las clases con 41,000 observaciones aproximadamente, para que cada clase represente cerca de un 20% de los datos cada una. Sobre

este nuevo set, se dividió en 2 sets de datos. Uno que contiene el 70% de los datos, que será utilizado para poder entrenar los modelos y el otro 30% para probar el modelo y sus capacidades predictoras (siempre manteniendo el balance entre las clases en ambas pariciones).

Métricas clave:

Las principales métricas son el accuracy o precisión el cuál es un porcentaje que indica que del 1 al 100 o bien del 0 al 1, cual fue el porcentaje de aciertos del modelo. Junto con esto, se estará mostrando una "Matriz de confusión" de los resultados. Una tabla que representa visualmente las predicciones de los modelos, que es muy común en problemas de clasificación, donde se puede ver claramente en donde tuvo aciertos y donde se equivocó más. Además de esto en los modelos se evaluará también el sobreajuste. Con sobreajuste/overfitting, nos referimos al fenómeno donde un modelo es muy bueno prediciendo con datos similares a los del entrenamiento, pero falla considerablemente más con datos nuevos.

Modelos

Para abordar este proyecto, se usaron diversos modelos de aprendizaje automático, incluyendo A Priori, K-Nearest Neighbors (KNN), Random Forest y Regresión Logística. Aprendizaje automático involucra el proceso mediante el cual se usan modelos matemáticos de datos para ayudar a un equipo a aprender sin instrucciones directas (Microsoft Axure, 2025). Estos algoritmos permitirán predecir el tipo de defunción en función de las características demográficas y demás características mencionadas anteriormente relacionados con el fallecimiento de la persona, utilizando los sets de datos de manera automática. Con esto se buscó encontrar los factores específicos que influyen en las decisiones de los modelos para clasificar la causa de muerte.

Explicación teórica de los modelos utilizados:

La Regresión Logística es un modelo estadístico fundamentalmente utilizado para problemas de clasificación. Pertenece a la familia de los modelos lineales generalizados y se basa en la idea de estimar la probabilidad de que una observación/registro pertenezca a una clase particular (James et al., 2013). En el caso de clasificación multiclase, como el nuestro con 5 causas de defunción, se suelen emplear extensiones como la regresión logística multinomial.

K-Nearest Neighbors (KNN) es un algoritmo de aprendizaje supervisado no paramétrico y basado en instancias. Esto significa que no construye un modelo explícito durante la fase de entrenamiento, sino que simplemente almacena todo el conjunto de datos de entrenamiento. Cuando se presenta un nuevo punto de datos para clasificar, KNN busca los "K" vecinos más

cercanos en el espacio de características. La cercanía se determina utilizando una métrica de distancia que es comúnmente la distancia euclidiana (James et al., 2013).

Random Forest es un algoritmo de aprendizaje conjunto. Su funcionamiento se basa en la construcción de un gran número de árboles de decisión independientes durante la fase de entrenamiento. Cada árbol se entrena con una submuestra aleatoria del conjunto de datos original y además, en cada nodo de decisión, solo se considera un subconjunto aleatorio de variables. Este algoritmo tiene alta precisión, robustez frente a datos faltantes y valores atípicos, y maneja un gran número de características sin necesidad de escalado. También proporciona una medida de la importancia de las características, lo que es útil para entender cuáles variables son más influyentes en la predicción (James et al., 2013).

El algoritmo Apriori es un algoritmo clásico de minería de reglas de asociación, no es un clasificador en el sentido estricto como los anteriores, pero es fundamental para descubrir patrones frecuentes en grandes conjuntos de datos. Su objetivo es identificar conjuntos de ítems que aparecen juntos con una frecuencia mínima (soporte) y luego derivar reglas de asociación de la forma "si {antecedente} entonces {consecuente}" con una confianza mínima (James et al., 2013).

Descripción de los datos y análisis exploratorio

El análisis exploratorio se realizó sobre el set de datos antes de las transformaciones. Es decir, cuando los registros con causas de defunciones poco comunes no habían sido fiiltrados y las observaciones eran alrededor de 1,100,000. Este conjunto comprende 19 variables, categorizadas principalmente en demográficas, temporales y de ubicación, además de la variable objetivo.

Las variables numéricas son un tanto reducidas. Las variables son el Año_registro (Añoreg), el Mes_registro (Mesreg), el Año_defuncion (year), el Dia_ocurrencia (Diaocu) y el Mes_ocurrencia (Mesocu), que capturan la fecha del evento de defunción o bien el registro de la misma. También se tiene la Edad_difunto (Edadif) proporciona la edad precisa del individuo al momento de fallecer.

Por otro lado, las variables categóricas, que en su mayoría son códigos o identificadores, ofrecen información demográfica y contextual detallada. Se tiene: Asistencia_recibida (Asist) sobre el tipo de atención médica; Causa_defuncion (Caudef), el código específico de la causa de muerte; Departamento_ocurrencia (Depocu), Departamento_registro (Depreg), Departamento_nacimiento (Dnadif), Departamento_residencia (Dredif), Municipio_residencia

(Mredif), y Municipio_nacimiento_difunto (Mnadif), que detallan las ubicaciones geográficas relevantes.

Además, se incluye el Estado_civil (Ecidif), la Escolaridad_difunto (Escodif), el Periodo_edad_difunto (Perdif,), el Sitio_ocurrencia (Ocur), y el Sexo_difunto (Sexo, donde 1 es hombre y 2 es mujer). Como parte del preprocesamiento, se derivó la variable letra_inicial del código de defunción, y la variable objetivo, causa_macro, que agrupa las causas de defunción en 5 clases principales para fines de clasificación.

A continuación, se muestran algunos de los descubrimientos más interesantes y relevantes del análisis exploratorio sobre las variables antes mencionadas.

Clustering

Para el clustering se eligieron las variables numéricas y se hizo un breve análisis para verificar el comportamiento:

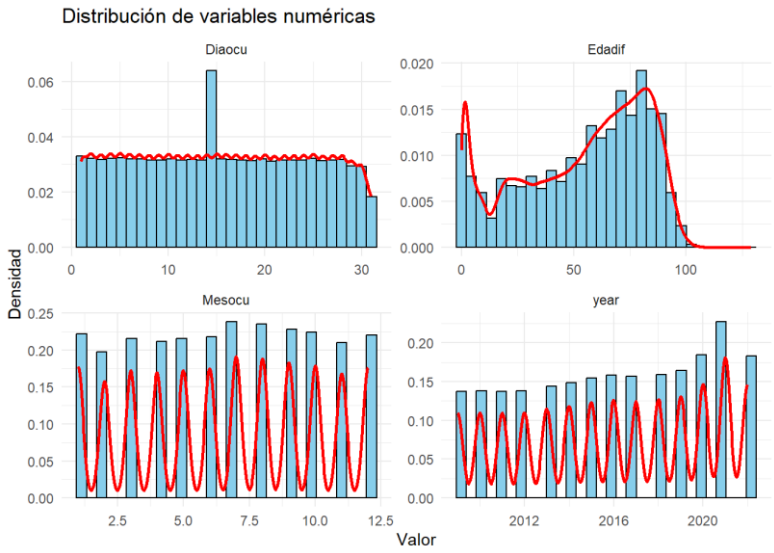

Figura 1. Distribuciones de frecuencia en variables numéricas utilizadas en el algoritmo.

En el caso de los meses realmente no se ve alguna tendencia en relación con los meses del año en relación a las defunciones. En el caso de los años se puede ver que con el tiempo han ingresado las defunciones, especialmente en los años de 2020 - 2022. Las defunciones parecen ser bastante mayores en personas alrededor de los 60-70 años y en niños también. Como último insight de este breve análisis se puede ver que el día en el cual hay más defunciones significativamente es el día 14 del mes.

Agrupamiento con K-means:

Utilizando 3 grupos y el algoritmo de K-means se obtiene el siguiente diagrama de clusters.

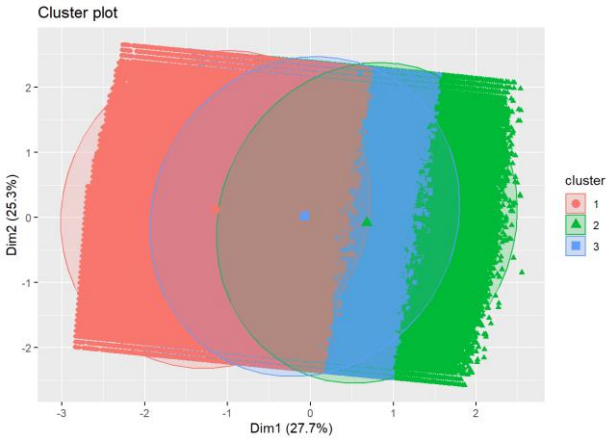

Figura 2. Resultados de agrupación con k-means con valor k = 3.

Con estos resultados podemos ver que realmente el algoritmo ha tenido dificultades para encontrar grupos realmente separados entre sí. El agrupamiento fue realizado con 4 variables al final: el año, mes, día de ocurrencia y la edad. Realmente estas primeras 3 no ayudaron mucho en el agrupamiento de los datos, sino que realmente la variable principal que agrupó los datos fueron los rangos de edad.

Interpretación de los grupos obtenidos:

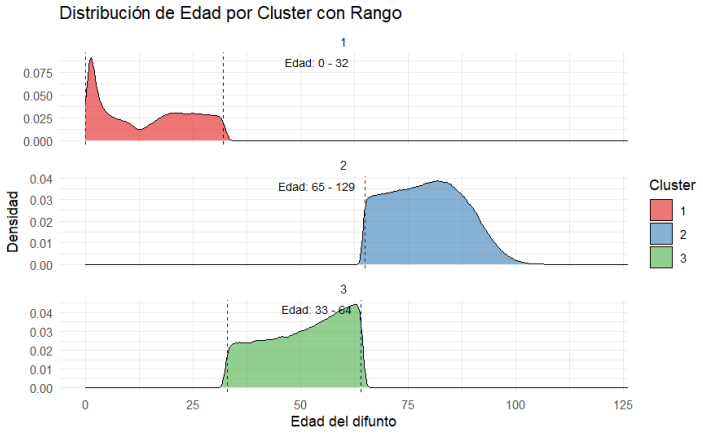

Figura 3. Distribución de edad por los clusters encontrados

Los grupos se dividen principalmente en 3. Podríamos llamarlos jóvenes (cluster 1), adultos (cluster 3) y adultos mayores (cluster 2) con rangos de edad aproximados de 0-30, 30-60 y 60+ respectivamente. En base a estos grupos de edad se analizaron las defunciones con respecto a distintos datos demográficos y temporales

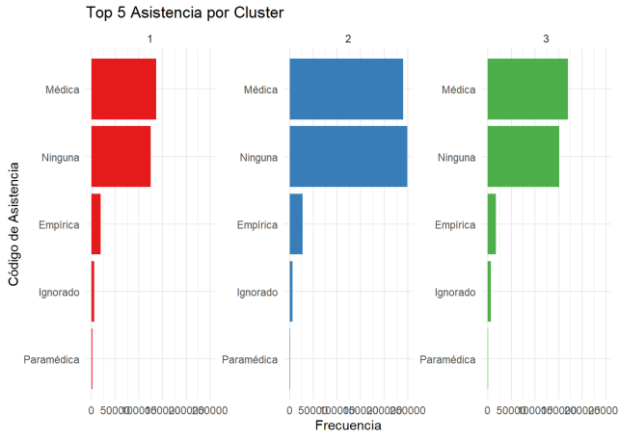

Figura 4. Tipos de asistencia médica por los grupos de edad determinados por el clustering.

Con respecto a la atención médica, en el gráfico se puede ver que, si hay intervención considerable del equipo médico, pero muchas personas independientemente de la edad no recibieron asistencia y la intervención de equipo paramédico parece ser casi nula. Se ve claramente que hay un gran problema con la asistencia a personas en riesgo por enfermedad o accidentes. Y encima de esto, se puede ver que hay falta de intervención médica si analizamos el lugar de ocurrencia por cada uno de los grupos. Dentro de la investigación, se encontrpo también que la mayoría de las defunciones ocurrierón en domicilios y no en hospitales sin importar la edad de las personas. Lo cuál debe ser una alerta en el sistema de salud.

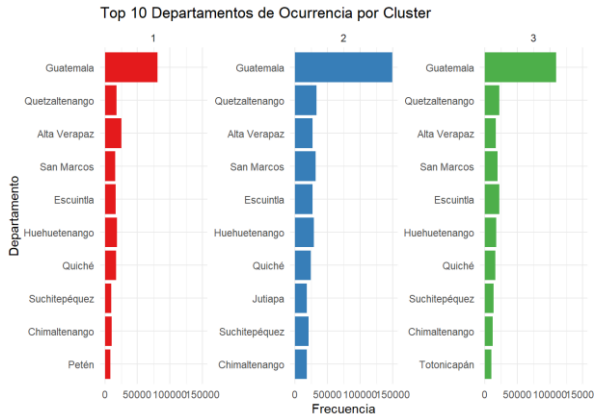

Figura 5. Diagrama de frecuencia para las defuncions por departamento y los grupos de edad del clustering.

El lugar con más defunciones es el área más urbanizada. Seguramente porque la densidad de población en la capital es mayor, hay muchas más personas en la capital Ciudad de Guatemala que en los demás departamentos. Aparte de la gran cantidad de defunciones en la capital del país, las defunciones en los departamentos son menores en comparación. A pesar de ser el área con mayor población y desarrollo, se identifica una población vulnerable y una gran cantidad de fallecidos en dicha ubicación.

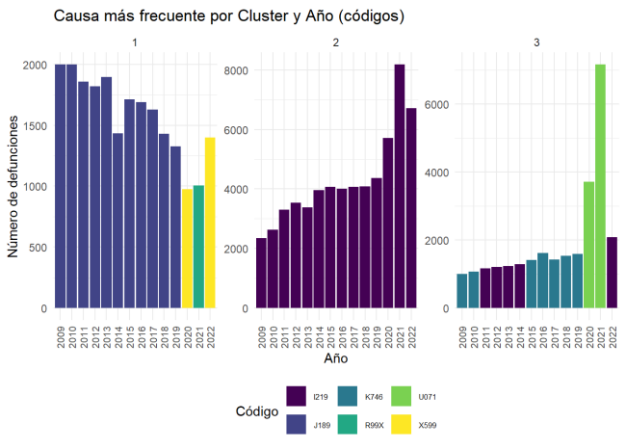

Figura 6. Gráfico que describe la causa de defunción más común por grupo de edad a lo largo del tiempo.

En el caso de las personas más jóvenes, la razón de defunción más común había sido la Neumonía(J189) hasta el 2020, de este año en adelante predominan otras causas que "no están bien identificadas" (códigos X599 y R99X). Por otra parte, en el caso de las personas del grupo 3 (personas de edad intermedia) la principal razón de defunción alterna entre Infartos (I219) y Cirrosis (K746), a excepción de los años de pandemia Covid-19. En el caso de las personas de mayor edad (grupo 2) la principal causa en todos los años fue infartos, incluidos los años de pandemia. Algo la pandemia fue letal no solo por el propio virus sino por el debilitamiento del cuerpo, exponiéndolo a otras enfermedades y empeorando la salud en general.

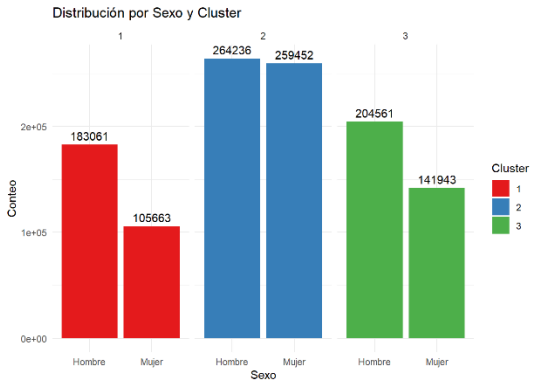

Figura 7. Gráfico que muestra la cantidad de muertes por sexo y grupo de edad sobre todos los datos.

En todos los casos, hay más defunciones de hombres que mujeres, independientemente del rango de edad. La única diferencia es que a medida que el grupo de edad es mayor la diferencia entre la cantidad de difuntos por sexo disminuye. En el grupo 2 (las personas de mayor edad), la cantidad de difuntos casi no varía dependiendo del sexo. A diferencia de las personas menores a 60 años (grupos 1 y 3) done se ve claramente que los hombres son más propensos a fallecer o al menos hay más defunciones en el sexo masculino.

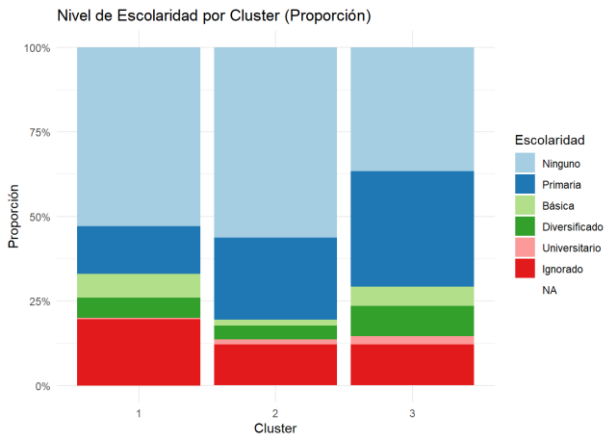

Figura 8. Proporción del nivel de escolaridad de los difuntos por grupo de edad.

Por último, tenemos el nivel de escolaridad de las personas. De todas las observaciones independientemente del rango de edad, el nivel de escolaridad es bastante bajo. Las personas fallecen sin haber alcanzado algún nivel de escolaridad o bien habiendo alcanzado nivel primaria. Es difícil correlacionar la escolaridad con causas de defunción específicas con únicamente este grafico o implicar que la escolaridad influye en la cantidad de fallecidos con solo esta información. Pero aun así, este gráfico levanta otra alerta y muestra cono la población sin importar la edad, maneja un estado de educación pobre. Lo cuál podría ser otra vulnerabilidad que atente contra la vida de las personas. Una vez realizado el análisis exploratorio, nos dispusimos a elegir los algoritmos de aprendizaje para predecir la causa de defunción.

Explicación de elección de algoritmos a utilizar

Esta elección se basa en la capacidad de los modelos previamente explicados para manejar tanto datos categóricos como numéricos para hacer predicciones multiclase. Además, estos algoritmos han sido previamente empleados y validados en investigaciones relacionadas con la salud y el análisis de defunciones, tanto en Guatemala como en Latinoamérica. Lo que respalda su inclusión para el presente estudio.

KNN se eligió por su eficacia en problemas de clasificación, demostrada en análisis previos donde este método mostró un alto porcentaje de precisión. Este modelo es relativamente simple, al tener un único parámetro K. Su naturaleza basada en la similitud en vecinos de las observaciones permite estudiar la influencia de las variables demográficas en la causa de muerte.

Random Forest es elegido por ser rubusto y preciso en problemas de clasificación. Es una opción prometedora a pesar de su costo computacional en grandes conjuntos de datos. Además,

este modelo reduce el sobreajuste y mejora la generalización. Adicionalmente, este algoritmo tiene la capacidad para generar métricas de importancia de las variables. Esto permite identificar cuáles características del set de datos son las más influyentes en la predicción de las causas de defunción. De hecho, estudios previos en Guatemala, como el "Análisis de condiciones que provocan decesos en Guatemala" (Orozco, 2024), ya han utilizado Random Forest como una herramienta clave para analizar causas de defunción en el país de manera exitosa.

La Regresión Logística Multiclase también es fundamental para este estudio, ya que permite estimar la probabilidad de que una persona pertenezca a una categoría específica de causa de defunción y predecir sobre dicha probabilidad. La relevancia de este algoritmo se refuerza al considerar estudios como "Lugar de muerte y factores asociados en 12 países de América Latina" (Getzzg, 2022), donde la regresión logística multivariable fue empleada con éxito para examinar factores sociodemográficos asociados a los lugares de fallecimiento.

Finalmente, el algoritmo A Priori Aunque no es un clasificador directo, es una herramienta valiosa para la minería de reglas de asociación. Ayuda a identificar patrones y correlaciones implícitas entre las características demográficas y las causas de defunción. En investigaciones previas sobre mortalidad, como la "Desmitificación de las causas de mortalidad por COVID-19 mediante análisis de datos interpretables" (Zhang & Qian, 2024).

Discusión de resultados

El análisis exploratorio reveló que, a pesar de la variación evidente en la densidad de defunciones durante el día, con un punto máximo en el día 14 de cada mes, las variables temporales no mostraron patrones estacionales distintos y de alguna manera predecibles que impacten sustancialmente la mortalidad general. Después de la aplicación del algoritmo K-medias, dividiendo sutilmente en 3 clusters por razones prácticas, se encontró que los principales factores separadores de estos eran grupos de edad: jóvenes, adulto y adulto joven; que sin embargo, se superponen mucho según el valor de la silueta promedio $\simeq$ -0.005, lo que indica la falta de la fuerte estructura de clústering. Este solapamiento sugiere que las variables demográficas usadas (edad, fecha) por sí solas no bastan para segmentar de forma nítida, pero sí permitió identificar que las causas de defunción varían según la edad: neumonía en jóvenes, accidentes e infartos en adultos, y enfermedades crónicas en adultos mayores. Además, la baja escolaridad se mantiene constante en todos los grupos, reflejando potenciales vulnerabilidades en la población.

El algoritmo Apriori se ejecutó con tres configuraciones de soporte y confianza, generando respectivamente 564, 21 y 120 reglas. Entre las asociaciones más destacadas se encuentran:

Modelo 1:

- `{Sexo=2, Depreg=1, causa_macro=Infecciosas} ⇒ {Caudef=A419}` con lift > 10, lo que evidencia una relación muy fuerte entre el sexo femenino en cierto departamento y el código específico A419 (neumonía)

Modelo 2:

- `{Caudef=R54X} ⇒ {causa_macro=Síntomas y causas mal definidas}` con confianza = 1.00 y lift = 5.00, mostrando que ese código específico se agrupa de forma invariable bajo su macro-causa

Modelo 3:

- `{Caudef=E149} ⇒ {causa_macro=Crónicas}` con lift = 5.0, indicando que el código E149 (enfermedad renal crónica) aparece frecuentemente en esa macro-categoría

Estas asociaciones complementan el clustering etario al revelar patrones demográficos más finos: permiten, por ejemplo, diseñar campañas de prevención de accidentes en poblaciones con escolaridad limitada o enfocar recursos sanitarios en determinados departamentos según sexo y tipo de causa de muerte.

Al ajustar el algoritmo Apriori con diferentes umbrales de soporte y confianza, se identificaron patrones de conexión significativos entre los datos demográficos y las distintas clasificaciones de motivos de fallecimiento. Los indicadores de soporte, confianza y lift sirvieron para confirmar la importancia de estos vínculos; un caso claro es la estrecha relación observada entre un bajo nivel de estudios y las muertes por causas no naturales, o la conexión entre el género y el lugar donde ocurre el deceso. Estas reglas añaden valor a la agrupación de datos al revelar relaciones ocultas que podrían ser útiles para crear estrategias de intervención específicas.

Modelos predictivos supervisados

Se evaluaron tres familias de clasificadores supervisados: K-Nearest Neighbors (KNN), Regresión Logística Multiclase y Random Forest.

- **KNN**: El mejor modelo (k=25) alcanzó un 47 % de accuracy, reduciendo el sobreajuste frente a valores mayores de k, pero mostrando aún una brecha notable entre entrenamiento y prueba.
- **Regresión Logística Multiclase**: Con un accuracy de 53.3 % y mínima diferencia train-test (≈ 1.1 %), ofreció un buen equilibrio entre generalización y capacidad predictiva, aunque con dificultades para capturar interacciones no lineales en categorías complejas como "Crónicas" versus "Neoplasias".
- **Random Forest**: Lideró el desempeño con un 56.63 % de accuracy, gracias a su robustez ante ruido y habilidad para modelar relaciones no lineales. Además, identificó a la variable temporal de ocurrencia ("Ocur"), seguida de "Edadif" y "Sexo", como las más influyentes en la predicción.

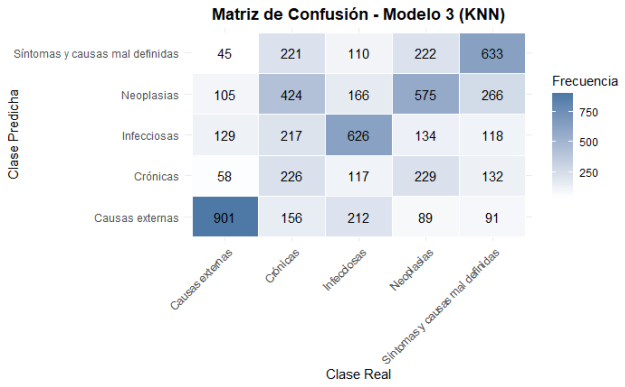

Figura 9. Matriz de confusión del mejor modelo con el algoritmo KNN.

```
Confusion Matrix and Statistics

                          Reference
Prediction                 Causas.externas Crónicas Infecciosas Neoplasias Síntomas.y.causas.mal.definidas
  Causas.externas                    10893      659         864        375                             573
  Crónicas                             401     2781        1266       2158                            1336
  Infecciosas                          707     2534        7701       1125                            1176
  Neoplasias                           219     4238        1392       6869                            2439
  Síntomas.y.causas.mal.definidas      161     2223        1087       1965                            6874
```

Figura 10. Matriz de confusión del mejor modelo con el algoritmo Random Forest

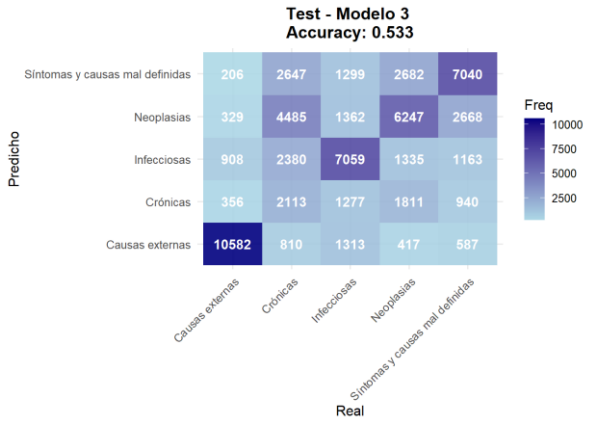

Figura 11. Matriz de confusión del mejor modelo con el algoritmo Regresión logística multiclase.

Un hallazgo consistente en todos los modelos fue la dificultad para diferenciar defunciones por enfermedades crónicas de neoplasias, sugiriendo que las variables demográficas disponibles no discriminan adecuadamente estas categorías clínicas. Esto apunta a la necesidad de enriquecer el dataset con información clínica adicional para mejorar la capacidad predictiva en estas clases específicas.

Importancia de variables detectadas con random forest:

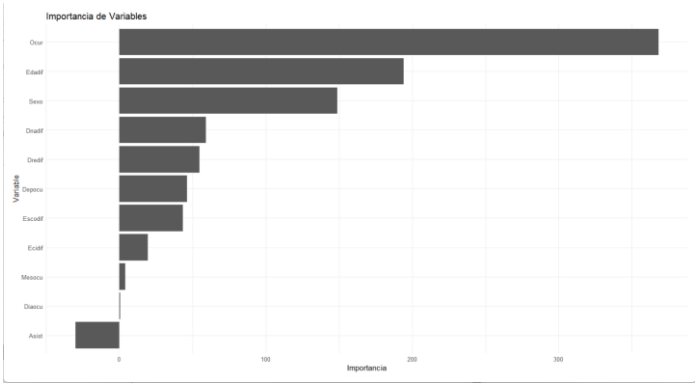

Figura 12. Gráfico que muestra la importancia de variables para clasificación usando random forest.

Según el algoritmo de Random Forest, las variables que mayor influencia tienen en el modelo son Ocur, Edadif y Sexo. Estas destacan significativamente por su contribución a la precisión de las predicciones, indicando que tienen un peso considerable en la toma de decisiones del modelo. En contraste, variables como Asist, Diaocu y Mesocu presentan una importancia mucho menor, lo que sugiere que su aporte al desempeño del modelo es limitado.

Implicaciones para salud pública

Si bien las precisiones obtenidas —aproximadamente entre el 47 % y el 56. 6 %— tal vez no parezcan muy altas, están bastante por encima del azar (20 %) y muestran tendencias que facilitan las decisiones. La agrupación junto a las reglas de vínculo y al estudio de factores esenciales propone focos de actuación: reforzar la asistencia sanitaria en colectivos sensibles (jóvenes y ancianos), optimizar el archivo y calidad de datos de decesos, y crear iniciativas formativas en sectores con poca formación. Igualmente, el modelo de Bosques Aleatorios podría funcionar como aviso previo ante variaciones en los caminos de la causa de defunción a nivel regional o demográfico.

Conclusiones

La exploración inicial reveló que, efectivamente, hay cambios notables en los fallecimientos dependiendo del mes, el día y el año, lo cual satisface por completo nuestra primera meta. Al dividir a las personas en tres grupos de edad —jóvenes, adultos y mayores— pudimos ver sin problemas cuáles son los momentos más delicados para cada grupo, así que también logramos el segundo objetivo.

Al poner en práctica el algoritmo Apriori, salieron a la luz conexiones significativas, por ejemplo, la marcada unión entre un bajo nivel de estudios y fallecimientos por factores ajenos, además del nexo entre el género y el sitio del suceso, cumpliendo así con el tercer propósito de enfatizar tendencias demográficas que sirvan de base para crear acciones específicas.

En la etapa de creación de modelos supervisados, se logró dejar atrás el punto de partida aleatorio (20 %). Random Forest llegó a un 56. 63 % de exactitud. Sin embargo, no se alcanzó el mínimo esperado del 60 % que se había fijado como meta en el cuarto objetivo. El problema más grande que notamos fue lo difícil que resultó distinguir las muertes por enfermedades crónicas de las causadas por tumores. Esto deja claro que es necesario añadir más datos clínicos al conjunto y buscar formas más complejas de equilibrar los datos para que el modelo funcione mejor.

En conjunto, estos hallazgos integran un enfoque robusto que combina análisis temporal, reglas de asociación y modelos predictivos, sentando las bases para el desarrollo de alertas tempranas y políticas de salud pública más focalizadas en Guatemala.

Recomendaciones

Este proyecto llevo una gran serie de pasos y fue un proceso extenso. La primera recomendación es en el enfoque del proyecto, específicamente no centrarlo en la búsqueda de un accuracy muy alto. Como se vió en este artículo, aun cuando el porcentaje de accuracy no es muy alto, es posible realizar descubrimientos y encontrar patrones relevantes en los datos si se eligen los métodos adecuados.

Este estudio se realizó sobre información a lo largo de todo el país, pero enfocarse en algún departamento o población específica podría traer resultados más precisos y detallados. Por lo que la reducción de las clases podría ser una buena opción o bien centrarla en un grupo específico

Por último, dado los desafíos computacionales y técnicos que enfrentamos al procesar más de 200,000 registros, recomendamos el uso de herramientas en la nube como Jupyter Notebook en Google Colab o Amazon SageMaker para optimizar el análisis. Estas plataformas permiten ejecutar modelos pesados sin depender del hardware local y puede seguirse realizando de manera colaborativa sin mayores dificultades.

Agradecimientos

Los autores desean expresar su más profundo agradecimiento a la Ing. Lynette García por su valiosa orientación, rigurosidad académica y constante apoyo, los cuales fueron determinantes para el desarrollo y la calidad de este trabajo. Asimismo, reconocen el compromiso y la dedicación del equipo de proyecto, cuyo esfuerzo colaborativo, profesionalismo y espíritu crítico contribuyeron de manera significativa al cumplimiento de los objetivos planteados.

Material Complementario:

Descripción de los códigos de defunción:

Código	Descripción Completa
1	
A099	Gastroenteritis y colitis de origen no especificado
A09X	Código no encontrado
J180	Bronconeumonía, no especificada
J189	Neumonía, no especificada
P220	Síndrome de dificultad respiratoria del recién nacido
P369	Sepsis bacteriana del recién nacido, no especificada
V899	Persona lesionada en accidente de vehículo no especificado
X599	Exposición a factores no especificados que causan otras lesiones y las no especificadas
X954	Agresión con disparo de otras armas de fuego, y las no especificadas, calles y carreteras
X959	Agresión con disparo de otras armas de fuego, y las no especificadas, lugar no especificado

2	
C169	Tumor maligno del estómago, parte no especificada
E149	Diabetes mellitus no especificada, sin mención de complicación
I219	Infarto agudo del miocardio, sin otra especificación
I64X	Accidente vascular encefálico agudo, no especificado como hemorrágico o isquémico
J189	Neumonía, no especificada
K746	Otras cirrosis del hígado y las no especificadas
R54X	Senilidad
R98X	Muerte sin asistencia
R99X	Otras causas mal definidas y las no especificadas de mortalidad
U071	COVID-19, virus identificado

3	
E149	Diabetes mellitus no especificada, sin mención de complicación
I219	Infarto agudo del miocardio, sin otra especificación
J189	Neumonía, no especificada
K746	Otras cirrosis del hígado y las no especificadas
N189	Enfermedad renal crónica, no especificada
R98X	Muerte sin asistencia
R99X	Otras causas mal definidas y las no especificadas de mortalidad
U071	COVID-19, virus identificado
X599	Exposición a factores no especificados que causan otras lesiones y las no especificadas
X959	Agresión con disparo de otras armas de fuego, y las no especificadas, lugar no especificado

Figuras 13-15. Códigos de defunción más comunes por grupo de edad. 1 es edades de 0-32; 2 es edades de 32-64 y 3 edades de 65 en adelante.

Enlace al repositorio: https://github.com/chuy-zip/PROYECTO_3

Enlace a archivos de las distintas etapas de desarrollo (incluyendo archivos csv): Minería proyecto 3

Referencias y bibliografía

Instituto Nacional de Estadística (INE). (s. f.). *Estadísticas vitales.* https://www.ine.gob.gt/vitales/

Microsoft Azure. (s. f.). https://azure.microsoft.com/es-es/resources/cloud-computing-dictionary/what-is-machine-learning-platform

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: with applications in R*. Springer, New York, EE.UU.

Seitz, K., Cohen, J., Deliens, L., Cartin, A., Castañeda de la Lanza, C., Cardozo, E. A., Marcucci, F. C. I., Viana, L., Rodrigues, L. F., Colorado, M., Samayoa, V. R., Tripodoro, V. A., Pozo, X. & Pastrana, T. (2022). Place of death and associated factors in 12 Latin American countries: A total population study using death certificate data. Journal of Global Health, *12*(1), 04031. https://jogh.org/2022/jogh-12-04031

Orozco, D. F. (2024). Analysis of conditions leading to deaths in Guatemala. ResearchGate. https://doi.org/10.13140/RG.2.2.36194.85442

Qian, X., Zuo, Z., Xu, D., He, S., Zhou, C., Wang, Z., Xie, S., Zhang, Y., Wu, F., Lyu, F., Zhang, L. & Qian, Z. (2024). Demystifying COVID-19 mortality causes with interpretable data mining. *Scientific Reports*, *14*, 10076. https://doi.org/10.1038/s41598-024-60841-w