Análisis de Data Mining sobre el Impacto de las Condiciones de Vivienda y Factores Socioeconómicos en la Mortalidad por Enfermedades Respiratorias en el Perú en 2022

Saldaña Hidalgo Diego Antonio   
Departamento de Ingeniería  
Universidad del PacíficoLima, Perú

Vargas Collas Jhoan Leandro   
Departamento de Ingeniería  
Universidad del PacíficoLima, Perú Verde Huayney Paulo Giusepe   
Departamento de Ingeniería  
Universidad del PacíficoLima, Perú

*Resumen: Las enfermedades respiratorias representan un grave problema epidemiológico en todo el mundo, y la Organización Mundial de la Salud (2017) representa hasta 650.000 muertes por año. En Perú, este problema se agrava en las regiones montañosas, donde las bajas temperaturas contribuyen al aumento de casos, especialmente entre niños menores de 5 años y entre mujeres adultas. Por otro lado, la calidad de la vivienda, ligada a los niveles de pobreza, afecta significativamente la salud de la población, especialmente en lugares expuestos a condiciones climáticas extremas. Esta investigación busca abordar esta problemática a través de un análisis de Data Mining usando la metodología Knowledge Discovery in Databases (KDD), el cual incluyó un análisis exploratorio, la aplicación de un modelo de regresión logística para predecir riesgos y la búsqueda de reglas de asociación identificar patrones complejos. Los resultados muestran que el acceso a agua potable reduce significativamente la mortalidad por enfermedades respiratorias, mientras que las bajas temperaturas aumentan el riesgo, especialmente en regiones vulnerables del sur. Estos hallazgos subrayan la influencia de factores ambientales y de infraestructura en la problemática de estudio, quedando en evidencia la necesidad de políticas integrales que combinen mejoras estructurales, como acceso a servicios básicos, y estrategias de prevención focalizadas en las regiones más afectadas.*

# Introducción

De acuerdo con la Organización Mundial de la Salud (2017), uno de los principales problemas epidemiológicos por los que atraviesan la mayoría de los países en el mundo es la morbilidad por enfermedades respiratorias, llegando a alcanzar hasta 650,000 defunciones anualmente. Se menciona también que este fenómeno no solo representa una amenaza a la salud pública, sino que también produce un costo social y económico a los países, motivo por el cual se impulsan activamente campañas de prevención mediante vacunas y otras intervenciones sanitarias. Estas enfermedades, según las define Soriano et al. (2020), comprometen el correcto funcionamiento de las vías respiratorias y estructuras pulmonares, y son resultado de la interacción con factores de riesgo como la contaminación del aire, el tabaquismo, las bajas temperaturas y otros patógenos.

Abordando el caso de estudio en cuestión, Giraldo y Pérez (2014) que, en el Perú, las enfermedades respiratorias se agravan en los departamentos ubicados en la zona de la sierra, donde las bajas temperaturas desempeñan un papel predominante en la intensificación de estos padecimientos. De hecho, actualmente el país se enfrenta a una emergencia sanitaria; según informa la Oficina de Relaciones Institucionales del Gobierno del Perú (2024) en una nota de prensa, EsSalud registró más de 966,000 casos de infecciones respiratorias agudas en lo que va del año, siendo las principales víctimas menores de 5 años y mujeres adultas. Ante esta situación se realizan esfuerzos gubernamentales por intentar prevenir estas enfermedades.

Por otro lado, otro factor importante dentro de las causas de estas afecciones es la condición de vivienda de las personas. En base a lo que menciona Quispe (2005), la calidad de los materiales de construcción de las viviendas en un determinado departamento está directamente relacionada con los niveles de pobreza registrados; el autor destaca que, geográficamente, las regiones más expuestas a condiciones climáticas extremas, como el frío, la lluvia y el calor intenso, son aquellas que presentan un mayor número de viviendas construidas con materiales de baja calidad.

Aunque la literatura ha identificado las causas y determinantes de las enfermedades respiratorias desde una perspectiva médica, la relación entre las muertes por estas enfermedades y las condiciones de vivienda, el acceso a los servicios de salud y los factores geográficos no está clara. En este sentido, esta investigación pretende responder a la siguiente pregunta: ¿De qué manera influyen las condiciones de vivienda, los factores demográficos, socioeconómicos y geográficos en la mortalidad por enfermedades respiratorias en el contexto peruano? Y para ello se plantea la siguiente hipótesis.

* Las condiciones de vivienda precarias, junto con un acceso limitado a los servicios de salud y factores sociodemográficos desfavorables, incrementan significativamente el riesgo de mortalidad por enfermedades respiratorias.

A partir de ello se plantea como objetivo general el explorar y comprender las causas de muerte relacionadas con enfermedades respiratorias, identificando patrones y asociaciones entre variables demográficas, socioeconómicas, condiciones de vivienda y acceso a servicios de salud, utilizando técnicas de Data Mining. Y como objetivos específicos los siguientes puntos:

* Analizar el impacto de factores demográficos, como la edad y el sexo, en el número de defunciones por enfermedades respiratorias.
* Evaluar el impacto de las condiciones de la vivienda, incluidos los materiales de construcción y el costo de la vivienda, sobre el riesgo de mortalidad por enfermedades respiratorias.
* Examinar el acceso a servicios básicos, como agua potable y atención médica, y su relación con el número de defunciones por enfermedades respiratorias.

# estado del arte

En la investigación titulada ‘*Factores asociados a la no utilización de los servicios formales de prestación en salud en la población peruana; análisis de la encuesta nacional de hogares 2015*’, Benites et al. (2017) hacen uso de variables demográficas como la edad, el género, el estado civil, el nivel educativo, la región y uso de los servicios médico de las personas para encontrar los determinantes de la brecha de acceso a atención médica en Perú. Los resultados indican que esta situación es particularmente grave en los departamentos más alejados de la capital y en las personas con menor nivel educativo. Este enfoque en el estudio de problemática del sistema sanitario peruano es adoptado por el presente trabajo para el análisis de las defunciones, integrando el acceso a servicios de salud con factores demográficos que influyen en la mortalidad por enfermedades respiratorias.

Por su parte Owusuaa et al. (2022) proporcionan un análisis integral de los predictores de mortalidad en pacientes con enfermedad pulmonar obstructiva crónica (EPOC), enfatizando la importancia de las variables socioeconómicas y geográficas y el acceso a los servicios de salud. Las investigaciones han identificado factores importantes que influyen en la mortalidad, como la hospitalización previa por exacerbaciones agudas y comorbilidad cardiovascular, así como variables demográficas como la edad y el sexo. Estos resultados resaltan la interrelación entre el estado de salud de los pacientes y su contexto demográfico, lo que resulta en un acceso desigual a la atención médica. La inclusión de estas variables en el análisis proporciona una mejor comprensión de cómo los entornos sociales y geográficos afectan la salud respiratoria, lo que corresponde al objetivo del presente estudio de descubrir las causas de muerte por enfermedades respiratorias en el Perú.

Finalmente, Rabbi y Mazzuco (2021) presentan el paper titulado ‘*Treinta años después: una revisión del método Lee-Carter para pronosticar la mortalidad’* donde se examinan los usos y limitaciones de uno de los modelos aceptados y usados para predecir el índice de mortalidad de un país. Los autores explican que esta metodología utiliza artificios estadísticos y matemáticos para transformar las tasas de mortalidad históricas de un país, clasificadas por grupos de edad, en la estimación del número de defunciones para los años siguientes. Sin embargo, las limitaciones del modelo incluyen la suposición de que las tendencias de mortalidad son constantes a lo largo del tiempo, lo que no resulta representativo en todos los contextos; además, el método no incorpora factores socioeconómicos o ambientales que pueden influir en las tasas de mortalidad. Dada esta posibilidad de mejora en la literatura sobre defunciones, en este estudio se plantea considerar las características que le faltan al modelo Lee-Carter, tales como el análisis de factores contextuales y el uso de técnicas de Data Mining que podrían proporcionar proyecciones más precisas y contextualizadas de la mortalidad en el futuro.

# METODOLOGÍA

A fin de organizar los procedimientos, se escogió la metodología Knowledge Discovery in Databases (KDD). A continuación, se expondrán las fases selección y limpieza y preprocesamiento. Posteriormente, en el Setup experimental se plantean 3 tareas de Data Mining. La primera consiste en un análisis exploratorio para determinar las características más importantes. La segunda tarea consiste en la aplicación de un modelo de regresión logística para predecir la probabilidad de ocurrencia de ciertos eventos asociados con enfermedades respiratorias. Finalmente, la tercera tarea implementa reglas de asociación mediante el algoritmo Apriori para identificar patrones y relaciones complejas entre las variables.

## Entendimiento del problema

De acuerdo con el análisis y e investigación acerca de los factores que incrementan el problema epidemiológico que se vive en el país, se estableció que para poder estudiar de forma efectiva y e innovadora los fallecimientos debido a enfermedades respiratorias, las variables de interés son las características demográficas, socioeconómicas, condiciones de vivienda y acceso a servicios de salud de las regiones del Perú.

## Selección y recopilación de datos

Uno de los Datasets escogidos fue extraído del Sistema Información de Fallecidos del Sistema Informático Nacional de Defunciones - SINADEF (2022) en el que se muestra atributos en relación con información de fallecidos como, perfil de la persona, centro de atención y causas de muerte.

Por otro lado, el segundo Dataset escogido proviene de á La Encuesta Nacional de Hogares (ENAHO) en el 2022. Esta detallada información particular de las personas acerca de su vivienda, salud, alimentación, ingresos, etc. Cada sección del ENAHO es un datase distinto con dimensiones distintas. El tamaño anual de la muestra 2022 es de 36 822 viviendas particulares, correspondiendo 24 206 viviendas al área urbana y 12 616 viviendas al área rural, y corresponde a un tipo de muestra probabilístico, de áreas, estratificado, multietápico e independiente en cada departamento de estudio.

Finalmente, el último Dataset integrado corresponde a las Variables Meteorológicas de las Estaciones automáticas de intercambio internacional proporcionado por el Servicio Nacional de Meteorología e Hidrología del Perú – SENAMHI. De este se obtienen las variables correspondientes a la temperatura, precipitación y humedad de las distintas localidades del Perú para el año 2022.

Para mayor comprensión de la estructura de datos, la Tabla 1 muestra las dimensiones de las fuentes mencionadas.

Tabla 1. Dimensiones iniciales de las fuentes de datos

|  |  |  |
| --- | --- | --- |
| Base de Datos | Número de Filas | Número de Columnas |
| SINADEF | 21,205 | 33 |
| ENAHO Vivienda | 44,122 | 11 |
| ENAHO Salud | 114,709 | 49 |
| SENAMHI | 408,960 | 16 |

## Limpieza y preprocesamiento

3.1. *Procesamiento de los Datos de Vivienda*

Se utilizó la base de datos de vivienda de la Encuesta Nacional de Hogares (ENAHO) para analizar las condiciones habitacionales. Inicialmente, se eliminaron columnas que no aportaban información relevante o eran redundantes, tales como:

• 'CONGLOME', 'VIVIENDA', 'HOGAR': Identificadores específicos que no contribuían al análisis general.

• 'AÑO': Redundante, ya que todos los registros correspondían al mismo año.

• Preguntas específicas sobre alquiler y tipo de vivienda: No aportaban al objetivo del estudio.

Se evaluó la cantidad de valores nulos por fila y se eliminaron aquellas con más de tres valores nulos para mantener la integridad de los datos. Además, se eliminaron columnas con más del 10% de valores nulos.

Para la variable 'El hogar tiene acceso al servicio de agua todos los días de la semana', se imputaron los valores faltantes utilizando la moda, asegurando así la consistencia de los datos.

Se clasificó la calidad de la vivienda en tres categorías basadas en los materiales predominantes en paredes, pisos y techos:

• Alta: Uso de materiales duraderos y de buena calidad como concreto armado, tejas, ladrillos, losetas y parquet.

• Media: Materiales de calidad intermedia como calamina, fibra de cemento, madera, adobe y cemento.

• Baja: Materiales menos duraderos como paja, caña, quincha y tierra.

La clasificación consideró que la presencia de al menos un material de categoría baja determinaba que la vivienda fuese clasificada como de calidad baja. Si no había materiales de categoría baja pero sí de categoría media, se clasificaba como media. Solo las viviendas con todos sus materiales de categoría alta fueron clasificadas como de calidad alta.

Se procedió de manera similar para clasificar la calidad del servicio de agua, teniendo en cuenta la procedencia del agua, su potabilidad y la frecuencia de acceso:

• Alta: Agua potable proveniente de la red pública dentro de la vivienda, con acceso todos los días.

• Media: Agua potable de la red pública fuera de la vivienda, pero dentro del edificio, con acceso todos los días.

• Baja: Agua no potable, acceso no diario o procedente de fuentes públicas externas.

Finalmente, se eliminaron las columnas utilizadas para estas clasificaciones y se obtuvo un conjunto de datos depurado para el análisis.

* 1. *Procesamiento de los Datos de Salud*

Se analizó la base de datos de salud de la ENAHO, enfocándose en variables relevantes para el estudio. Se eliminaron columnas que no aportaban valor analítico, como identificadores específicos, detalles de enfermedades no pertinentes y preguntas de carácter muy específico.

Para manejar los valores nulos, se rellenaron con ceros las columnas relacionadas con consultas médicas y razones para no acudir a establecimientos de salud. Se creó una nueva columna categórica 'Lugar\_de\_consulta' que combinó las diferentes opciones de lugares donde los encuestados buscaron atención médica.

Se calcularon los tiempos totales de espera para ser atendido, unificando las unidades de medida (meses, días, horas, minutos) en una sola métrica en minutos. Asimismo, se consolidaron las razones para no acudir a centros de salud en la columna 'Motivo\_no\_atencion'.

Se estandarizaron las respuestas sobre medicamentos tomados para tratamientos específicos, facilitando su análisis posterior.

* 1. *Procesamiento de Datos Meteorológicos*

Se incorporaron datos meteorológicos correspondientes al año 2022, incluyendo temperatura (TEMP), humedad relativa (HR) y precipitación (PP). Se extrajeron las variables de año, mes y día a partir de la fecha y se eliminaron columnas no relevantes para el análisis.

Los valores nulos en las variables meteorológicas se imputaron utilizando la media de cada variable. Se agruparon los datos por mes y código ubigeo, calculando los promedios mensuales para cada variable. Se añadió la columna 'ubigeo\_departamento' para facilitar la integración con otros conjuntos de datos.

* 1. *Procesamiento de Datos de Mortalidad*

Se utilizó la base de datos del Sistema Informático Nacional de Defunciones (SINADEF) para analizar las causas de muerte relacionadas con enfermedades respiratorias en 2022. Se eliminaron columnas irrelevantes y se unieron los datos con la tabla de ubigeos para obtener información geográfica detallada.

Se extrajo el mes de ocurrencia de la defunción y se filtraron las causas de muerte para enfocarse en las más relevantes para el estudio, seleccionando las cinco causas respiratorias más frecuentes: insuficiencia respiratoria aguda, neumonía, enfermedad pulmonar obstructiva crónica, bronconeumonía y fibrosis pulmonar.

Se eliminó cualquier registro con valores nulos en las variables clave para asegurar la calidad del análisis.

* 1. *Integración de los Conjuntos de Datos*

Se realizó la combinación de los conjuntos de datos de SINADEF y variables meteorológicas utilizando como claves el mes y el código de departamento. Esto permitió analizar la incidencia de causas de muerte respiratorias en relación con variables climáticas.

Se asignaron las estaciones del año según el mes para evaluar posibles patrones estacionales en las causas de muerte. Posteriormente, se integraron los datos de salud y vivienda de la ENAHO, alineando las claves geográficas y temporales.

Para manejar los valores nulos resultantes de la integración, se imputaron utilizando la moda dentro de cada región específica, asegurando así la coherencia de los datos.

Esta integración proporcionó un conjunto de datos completo que combina información sobre condiciones de vivienda, salud, clima y mortalidad, facilitando un análisis multidimensional de los factores que influyen en las enfermedades respiratorias.

# SETUP EXPERIMENTAL

* 1. *Tarea 1. Análisis exploratorio*

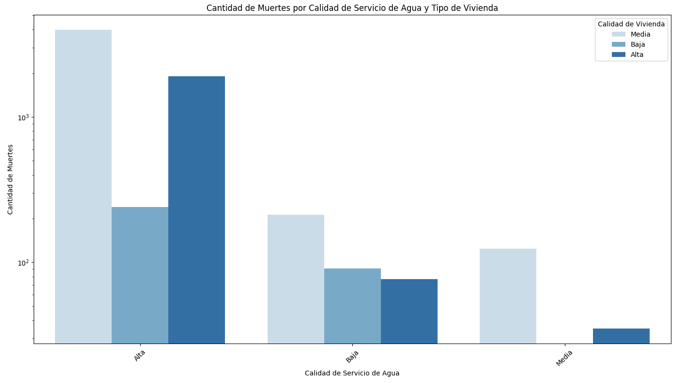
Figura 1. *Correlación entre las condiciones Climáticas, socioeconómicas y edad (One-Hot Encoding)*

Gráfico, Gráfico en cascada

Descripción generada automáticamente

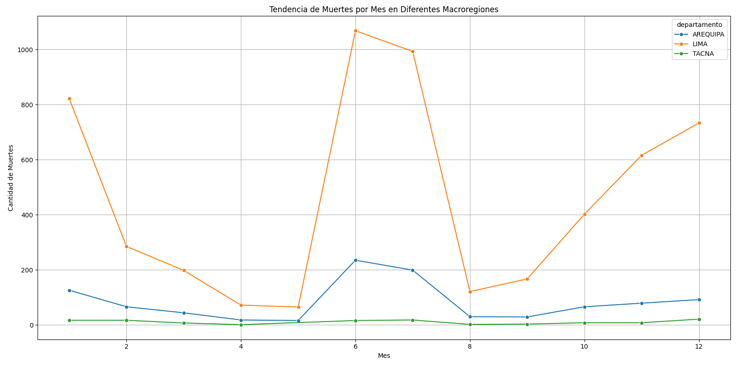
La Figura 1 muestra la matriz de correlación entre las variables climáticas, socioeconómicas y edad, utilizando un codificado one-hot para variables categóricas como la calidad de la vivienda y el servicio de agua. Un hallazgo importante es la alta correlación positiva entre Precipitación (PP) y Humedad Relativa (HR) (0.97), lo cual es esperado ya que mayor precipitación generalmente resulta en mayor humedad. También hay correlaciones significativas y negativas entre la calidad del agua y la vivienda en sus niveles "media" y "alta". Esto sugiere que una mejor calidad de vivienda o agua está relacionada negativamente con algunos factores climáticos, lo que podría influir en la mortalidad y las condiciones de vida en las regiones evaluadas. La correlación entre Edad y las demás variables es baja, lo cual indica que, en este contexto, la edad no parece tener una relación significativa con los otros factores estudiados.

Figura 2. *Cantidad de Muertes por Calidad de Servicio de Agua y Tipo de Vivienda.*



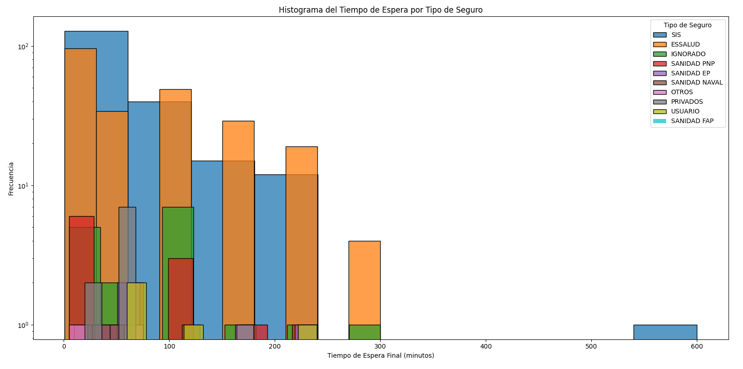
La figura 2 muestra la cantidad de muertes según la calidad del servicio de agua y el tipo de vivienda destaca cómo estas variables interactúan en la mortalidad por enfermedades respiratorias. Se observa que la mayoría de las muertes se concentran en viviendas de calidad alta y media con servicios de agua de buena calidad, lo que podría estar relacionado con una mayor capacidad de registro y acceso a sistemas de salud en estas áreas. En contraste, las viviendas de menor calidad y con servicios de agua deficientes presentan un número significativamente menor de muertes, lo que podría indicar un subregistro de casos en comunidades con menos recursos y acceso limitado a servicios de salud.

Figura 3. *Tendencia de Muerte por Mes en Diferentes Macroregiones.*



La figura 3 muestra la tendencia de muertes por mes en diferentes departamentos destaca las variaciones estacionales y regionales en la mortalidad por enfermedades respiratorias. Se observa que el departamento de Lima presenta picos significativos de mortalidad al inicio y al final del año, lo cual podría estar asociado a cambios estacionales y condiciones climáticas específicas que afectan la salud respiratoria. Por otro lado, Arequipa muestra un patrón más moderado, con fluctuaciones a mitad de año, mientras que Tacna mantiene un nivel de mortalidad relativamente bajo y constante a lo largo de los meses. Estos hallazgos refuerzan la importancia de considerar las diferencias regionales en el diseño de políticas de salud pública, ya que cada región puede enfrentar desafíos particulares en términos de condiciones ambientales y acceso a servicios de salud.

Figura 4. *Histograma del Tiempo de Espera por Tipo de Seguro*



La figura 4 revela una distribución desigual en los tiempos de atención, destacando al SIS como el seguro con mayores frecuencias en tiempos prolongados, lo que indica una carga significativa en su gestión. ESSALUD también presenta alta frecuencia, aunque en un rango más moderado. Los seguros privados y otros como SANIDAD EP y PRIVADOS muestran tiempos de espera más cortos, sugiriendo un acceso más eficiente. La mayoría de los datos se concentra entre 0 y 100 minutos, con una disminución notable en frecuencias a medida que los tiempos aumentan, lo que resalta la necesidad de políticas que optimicen los tiempos de espera, especialmente en los seguros públicos, para mejorar la equidad en el acceso a la atención sanitaria.

* 1. *Tarea 2. Regresión Logística*
     1. *Pre-procesamiento de Datos*

Durante el pre-procesamiento, se llevó a cabo una eliminación manual de variables redundantes y aquellas con alta cardinalidad, como provincia y distrito, que no aportaban valor significativo al modelo. Las variables categóricas se codificaron mediante Label Encoding para convertirlas en un formato numérico adecuado. Además, se normalizaron las características numéricas usando StandardScaler para asegurar una escala uniforme, evitando así sesgos en el peso de las variables y mejorando la estabilidad del entrenamiento del modelo. La variable objetivo (target) seleccionada para el modelo fue **causa de muerte**, categorizada en cinco clases distintas, representando diferentes causas de fallecimiento relacionadas con enfermedades respiratorias.

* + 1. *Selección de Atributos*

Para identificar las variables más relevantes, se aplicó el método de Recursive Feature Elimination (RFE) utilizando un modelo de regresión logística como evaluador. Este proceso seleccionó un total de 10 atributos clave, incluyendo edad, tipo de seguro, temperatura, calidad del servicio de agua, PP, estación, dominio, estrato, administración de oxígeno, y departamento. La elección de estos atributos se basó en su alta capacidad predictiva para la clasificación de las causas de muerte, destacando aquellos con un mayor peso en la determinación de la variable objetivo. Esta selección contribuyó a un análisis más eficiente al reducir la complejidad del modelo y enfocar el entrenamiento en las variables con mayor impacto.

* + 1. *Configuración del Modelo*

El modelo de regresión logística multinomial fue configurado con el solver *lbfgs* y un ajuste de max\_iter para asegurar la convergencia, considerando el tamaño del dataset y la tarea de clasificación multiclase. La variable objetivo fue causa de muerte, lo que implica un desafío adicional debido a la naturaleza multiclase del problema. Los datos fueron divididos en un 80% para entrenamiento y un 20% para pruebas, garantizando así una evaluación objetiva del rendimiento del modelo. Se calcularon los *odds ratios* para interpretar la influencia de cada atributo en las probabilidades de pertenecer a una clase específica.

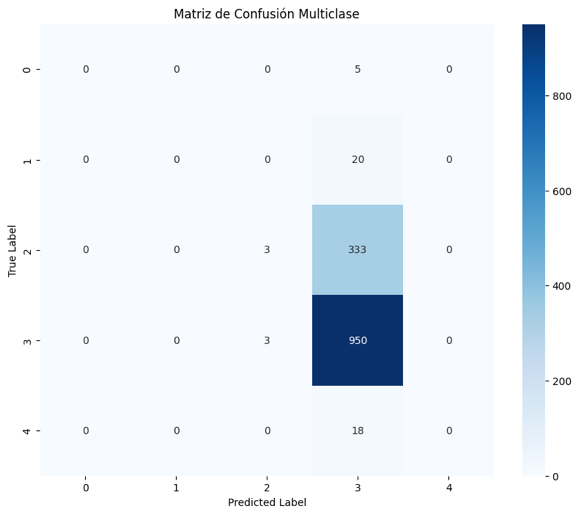
* + 1. *Resultados*

*Tabla 2. Odd ratios*

|  |  |
| --- | --- |
| Atributo | Odd Ratio |
| TIPO SEGURO | 0.902 |
| EDAD | 1.470 |
| TEMP | 0.714 |
| PP | 0.996 |
| Estación | 0.732 |
| DOMINIO\_x | 1.710 |
| ESTRATO\_x | 1.554 |
| Administración de oxígeno | 0.823 |
| Calidad\_servicio\_agua | 0.655 |
| departamento | 1.275 |

Los odds ratios indican cómo las variables predictoras influyen en la probabilidad de pertenecer a una categoría de causa de muerte. Por ejemplo, un odds ratio de 1.470 para "EDAD" sugiere un aumento del 47% en la probabilidad de pertenecer a una causa específica por cada año adicional de edad. En contraste, "TEMP" con un odds ratio de 0.714 indica una reducción del 28.6% en esa probabilidad al aumentar la temperatura, y "CALIDAD\_SERVICIO\_AGUA" con 0.655 implica que un mejor acceso al agua reduce en un 34.5% la probabilidad de ciertas causas de muerte.

*Imagen 1. Matriz de confusión*

**

La matriz de confusión revela un fuerte sesgo hacia la clase 3, que es la más representada, mientras que las otras clases no son correctamente identificadas. Esto demuestra que el modelo tiene dificultades para clasificar correctamente las clases minoritarias, enfocándose solo en la clase predominante.

*Tabla 3. Reporte de Clasificación Multiclase*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Accuracy | 0.72 |  |  |  |
| Clase | Precision | Recall | F1-score | Support |
| 0 | 0.00 | 0.00 | 0.00 | 5 |
| 1 | 0.00 | 0.00 | 0.00 | 20 |
| 2 | 0.50 | 0.01 | 0.02 | 336 |
| 3 | 0.72 | 1.00 | 0.83 | 953 |
| 4 | 0.00 | 0.00 | 0.00 | 18 |

El modelo mostró un **accuracy global del 72%**, indicando un buen desempeño en la clase más representada (clase 3) con un **F1-score de 0.83** y un recall del 100%. Las otras clases, sin embargo, presentaron menor precisión y recall, especialmente la clase 2. Esto sugiere la necesidad de ajustes para mejorar el balance y la predicción de todas las categorías. En general, el modelo es efectivo para la clase principal y se recomienda optimización para fortalecer la predicción en clases menos representadas.

* 1. *Tarea 3. Reglas de asociación.*
     1. *Selección de atributos*

En esta sección se utilizó el dataframe preprocesado, aplicándose las siguientes modificaciones. En primer lugar, se eliminaron las columnas que no posean valor interpretativo como 'ubigeo\_inei', 'Dominio\_x', y 'Estrato\_x'. Del mismo modo, se suprimieron aquellas que después de la experimencación presentaban resultados conflictivos y poco interpretables como 'HR', 'PP', y 'Lugar\_de\_consulta' Como resultado, las dimensiones iniciales de 6660 filas y 26 columnas se redujeron a 13 columnas.

Seguidamente, las variables categóricas restantes resultaron ser de baja cardinalidad y se aplicó el one-hot encoding. Por otro lado, la variable edad fue agrupada en rangos de 10, obteniéndose grupos desde [0,10[ hasta [100,110[. Por su parte la variable temperatura fue separada en 3 usando los percentiles 33 y 66 quedando una separación de temperatura en alta, baja y media. En esta sección el número de columnas asciende a 65

Adicionalmente se buscó reducir las columnas que no aporten información relevante, para ello se obtuvo la correlación de las variables y se eliminaron aquellas que superen el 90% de similitud. El método usado fue la correlación de Jaccard el cual es especialmente útil para medir similitudes entre variables binarizadas ya que evalúa las coincidencias entre 2 variables en relación con el total de elementos únicos. El cálculo se observa en la ecuación 1.

Adicionalmente, se buscó identificar y eliminar columnas cuyos valores estuvieran desbalanceados, es decir, donde el valor predominante (generalmente 0) representaba el 80% o más de las filas. Este criterio permitió reducir la cantidad de columnas originales, optimizando el análisis al centrarse en variables con mayor diversidad en sus valores. El número final de columnas relevantes fue 18.

* + 1. *Aplicación del modelo de reglas de asociación*

En consonancia con el objetivo de la investigación y debido a la naturaleza de los datos usados, buscar reglas de asociación entre características demográficas, climáticas, condiciones de vivienda y salud permite reconocer interacciones complejas entre las variables. Para ello, se aplicó el algoritmo Apriori, utilizando las métricas de soporte, confianza y lift, con el objetivo de evaluar la fuerza y relevancia de las asociaciones. En el análisis se integró un filtro de antecedentes, el cual permite interpretar las reglas generadas en función de categorías específicas de las variables.

* + 1. *Resultados*

*Tabla 4. Reglas de asociación de Enfermedades.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Antecedentes | Consecuentes | Soporte | Confianza | Lift |
| Causa\_insuficiencia respiratoria | Estacion\_primavera | 0.09 | 0.36 | 1.2 |
| Causa\_insuficiencia respiratoria | Edad\_group\_[70, 80) | 0.07 | 0.25 | 1.12 |
| Causa\_insuficiencia respiratoria | Calidad\_vivienda\_alta | 0.08 | 0.32 | 1.05 |
| Causa\_insuficiencia respiratoria | Edad\_group\_[80, 90) | 0.08 | 0.31 | 1.02 |

En los resultados de la Tabla 4, la regla más prominente muestra que la insuficiencia respiratoria tiene una asociación moderada con la estación de primavera (lift=1.2), indicando un incremento del 20% en la probabilidad de ocurrencia respecto a la frecuencia esperada por azar. Adicionalmente, se observaron asociaciones sutiles con grupos etarios de 70-80 y 80-90 años, y una leve correlación con viviendas de alta calidad, sugiriendo factores multidimensionales en la manifestación de esta condición respiratoria.

*Tabla 5. Reglas de asociación de Tipo de vivienda.*

*|*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Antecedentes | Consecuentes | Soporte | Confianza | Lift |
| Calidad\_vivienda Alta | EDAD\_GROUP [90, 100) | 0.05 | 0.17 | 1.08 |
| Calidad\_vivienda Alta | EDAD\_GROUP [80, 90) | 0.09 | 0.31 | 1.03 |
| Calidad\_vivienda Media | EDAD\_GROUP [60, 70) | 0.09 | 0.14 | 1.01 |

La Tabla 5 muestra resultados que sugieren que los individuos de viviendas de alta calidad presentan una probabilidad ligeramente superior de fallecimiento en grupos etarios avanzados, con un lift de 1.08 para el rango [90, 100) y 1.03 para [80, 90), indicando una asociación marginal pero estadísticamente relevante. Adicionalmente una asociación entre los fallecidos con calidad de vivienda media y el grupo el grupo de [60, 70) años.

*Tabla 6. Reglas de asociación de Macro región*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Antecedentes | Consecuentes | Soporte | Confianza | Lift |
| Macro región SUR | Temperatura Baja, No suministro de oxígeno. | 0.17 | 1 | 3.1 |

La Tabla 6 presenta un solo resultado el cual muestra importantes hallazgos en la relación geográfica y condiciones de salud. Con un lift de 3.1 y una confianza de 1, el análisis indica que, en la totalidad de los registros de muerte en esta macro región, estas dos condiciones están intrínsecamente relacionadas, sugiriendo una vulnerabilidad sistémica en el acceso a recursos sanitarios en zonas de bajas temperaturas.

# CONCLUSIONES

Los resultados del análisis de minería de datos sobre la mortalidad por enfermedades respiratorias en Perú durante 2022 revelan patrones complejos e interrelaciones significativas:

Factores Determinantes y su Interpretación:

## La edad como predictora dominante (OR=1.470) revela una vulnerabilidad progresiva en la población adulta mayor, especialmente notable cuando se combina con condiciones socioeconómicas específicas.

## El efecto protector del acceso a agua potable (OR=0.655) demuestra que la infraestructura sanitaria básica actúa como barrera preventiva contra la mortalidad respiratoria, sugiriendo que las intervenciones en servicios básicos podrían tener un impacto más significativo que algunas intervenciones médicas directas.

## La relación inversa con la temperatura (OR=0.714) no solo indica un riesgo climático directo, sino que revela vulnerabilidades sistémicas en la capacidad de respuesta del sistema de salud en regiones frías.

Patrones Geográficos y Temporales:

## La distribución heterogénea de la mortalidad entre regiones refleja disparidades estructurales en el sistema de salud, más allá de simples diferencias demográficas.

## La estacionalidad marcada en la mortalidad, particularmente en primavera (lift=1.2), sugiere la existencia de factores ambientales específicos que actúan como catalizadores de la morbilidad respiratoria.

## La asociación crítica en la región sur entre bajas temperaturas y falta de oxígeno (lift=3.1) evidencia una falla sistémica en la planificación de recursos médicos esenciales.

Interacciones Socioeconómicas y Sanitarias:

## La asociación entre viviendas de alta calidad y mortalidad en edades avanzadas (lift=1.08) sugiere un posible sesgo de registro que enmascara la realidad en zonas más vulnerables, indicando la necesidad de mejorar los sistemas de vigilancia epidemiológica.

## El modelo logístico (accuracy 72%) revela que la predicción de mortalidad respiratoria es más precisa cuando se consideran factores socioeconómicos junto con variables clínicas tradicionales, sugiriendo la necesidad de un enfoque más holístico en la prevención.

En base a estos hallazgos, se recomiendan las siguientes acciones:

A) Intervenciones Estructurales:

## Desarrollar un sistema integrado de vigilancia que combine datos climáticos, epidemiológicos y socioeconómicos para identificar poblaciones en riesgo.

## Implementar programas de mejora de infraestructura sanitaria básica priorizando zonas con alta vulnerabilidad climática.

## Establecer protocolos de distribución de recursos médicos que consideren patrones estacionales y geográficos identificados.

B) Políticas de Salud Pública:

## Reformular los programas de prevención considerando la multifactorialidad evidenciada en el análisis.

## Fortalecer los sistemas de registro epidemiológico en zonas vulnerables para superar el sesgo de subregistro.

## Desarrollar intervenciones específicas para grupos de alta vulnerabilidad, considerando la interacción entre edad y condiciones socioeconómicas.

Esta investigación demuestra que la mortalidad por enfermedades respiratorias en Perú es resultado de una compleja interacción entre factores ambientales, socioeconómicos y sanitarios, requiriendo un enfoque multidimensional para su abordaje efectivo. Los patrones identificados proporcionan una base para la toma de decisiones en política sanitaria y la asignación eficiente de recursos.

# Referencias

Benites-Zapata, V. A., Lozada-Urbano, M., Urrunaga-Pastor, D., Márquez-Bobadilla, E., Moncada-Mapelli, E., & Mezones-Holguín, E. (2017). Factores asociados a la no utilización de los servicios formales de prestación en salud en la población peruana: análisis de la encuesta nacional de hogares (enaho) 2015. *Revista Peruana de Medicina Experimental y Salud Pública*, *34*(3), 478–484. https://www.redalyc.org/articulo.oa?id=36353391015

Gregory Härtl. (2017, December 13). *Cada año, hasta 650.000 personas mueren por enfermedades respiratorias relacionadas con la gripe estacional*. OMS.

Oficina de Relaciones Institucionales Gobierno del Perú. (2024, June 4). *EsSalud registra más de 966 mil casos de infecciones respiratorias agudas a nivel nacional y refuerza campañas informativas*. Gob.Pe.

Owusuaa, C., Dijkland, S. A., Nieboer, D., van der Rijt, C. C. D., & van der Heide, A. (2022). Predictors of mortality in chronic obstructive pulmonary disease: a systematic review and meta-analysis. *BMC Pulmonary Medicine*, *22*(1), 125. https://doi.org/10.1186/s12890-022-01911-5

Quispe Romero, J. (2005). El Problema de la Vivienda en el Perú, Retos y Perspectivas. *Revista INVI*, *20*(53), 20–44. https://www.redalyc.org/articulo.oa?id=25805303

Rabbi, A. M. F., & Mazzuco, S. (2021). Mortality Forecasting with the Lee–Carter Method: Adjusting for Smoothing and Lifespan Disparity. *European Journal of Population*, *37*(1), 97–120. https://doi.org/10.1007/s10680-020-09559-9

Soriano, J. B., Kendrick, P. J., Paulson, K. R., Gupta, V., Abrams, E. M., Adedoyin, R. A., Adhikari, T. B., Advani, S. M., Agrawal, A., Ahmadian, E., Alahdab, F., Aljunid, S. M., Altirkawi, K. A., Alvis-Guzman, N., Anber, N. H., Andrei, C. L., Anjomshoa, M., Ansari, F., Antó, J. M., … Vos, T. (2020). Prevalence and attributable health burden of chronic respiratory diseases, 1990, 2013;2017: a systematic analysis for the Global Burden of Disease Study 2017. *The Lancet Respiratory Medicine*, *8*(6), 585–596. https://doi.org/10.1016/S2213-2600(20)30105-3

Zeña Giraldo, S., & Barceló Pérez, C. (2014). Clima e incidencia de infecciones respiratorias agudas en Ancash, Perú (2005-2013). *Revista Cubana de Higiene y Epidemiología*, *52*, 301–313. https://www.gob.pe/institucion/essalud/noticias/983311-essalud-registra-mas-de-966-mil-casos-de-infecciones-respiratorias-agudas-a-nivel-nacional-y-refuerza-campanas-informativas