Tecnolochicas México

“Análisis de Datos con Python 2024”

Proyecto final

"Análisis de Datos de Pacientes con COVID-19: Evaluación de Condiciones Preexistentes y Resultados Clínicos Globales"

TEAM 20

Integrantes:

- **De la Cruz Munguía Arely** arelydlcm07@gmail.com
- **Pérez Mendoza Leisly** leislymp@utbispuebla.edu.mx
- **Valle Núñez Gabriela** gabrielavallenunez@gmail.com

Fecha de entrega: 28/10/2024

Índice

Proyecto Tecnolochicas MÓDULO 2 Team 20

Integrantes:
- De La Cruz Munguia Arely arelydlcm07@gmail.com
- Pérez Mendoza Leisly leislymp@utbispuebla.edu.mx
- Valle Núñez Gabriela gabrielavallenunez@gmail.com

Análisis Exploratorio y Limpieza de Datos de Pacientes con COVID-19: Un Enfoque en Condiciones Preexistentes y Resultados Clínicos Globales

Introducción

El COVID-19, causado por el virus SARS-CoV-2, surgió a finales de 2019 y rápidamente se convirtió en una pandemia global. La propagación del virus, altamente contagioso, llevó a una crisis de salud pública sin precedentes, afectando a millones de personas en todo el mundo. La enfermedad se caracteriza por síntomas que varían desde leves, como fiebre y tos, hasta graves, como neumonía y fallo multiorgánico, que en muchos casos llevan a la muerte.

A nivel global, el COVID-19 ha provocado millones de defunciones, exacerbando las desigualdades en los sistemas de salud, especialmente en países con recursos limitados. La rapidez con la que se extendió el virus desbordó los sistemas de atención médica, llevando a una escasez de recursos como camas de hospital, ventiladores, y equipos de protección personal. Además, la pandemia tuvo un impacto devastador en las economías, aumentó la pobreza, y alteró profundamente las dinámicas sociales y laborales.

Las campañas de vacunación a gran escala han sido fundamentales para reducir la mortalidad y controlar la propagación del virus, aunque la aparición de nuevas variantes y la desigualdad en el acceso a las vacunas siguen siendo desafíos importantes.La pandemia de COVID-19 ha planteado desafíos significativos para los sistemas de salud en todo el mundo, afectando a millones de personas de diferentes maneras. Los pacientes con condiciones preexistentes, como diabetes, hipertensión y enfermedades respiratorias, han sido identificados como particularmente vulnerables a desarrollar complicaciones graves. Este proyecto tiene como objetivo explorar y analizar un conjunto de datos proporcionado por el Sistema Nacional de Salud de México, el cual contiene información anonimizada de más de un millón de pacientes diagnosticados con COVID-19 en el año de 2020.

A través de un análisis exploratorio y de limpieza de datos, buscamos identificar patrones relevantes y comprender mejor cómo factores demográficos y clínicos, como la edad, el sexo, las enfermedades preexistentes y el tipo de atención médica recibida, influyen en la evolución clínica de los pacientes con COVID-19. Además, este proyecto pretende ofrecer una visión detallada de la distribución de estas variables y cómo pueden afectar para considerar que el paciente tiene un bajo o alto riesgo.

El análisis se enfoca en limpiar y estructurar el dataset para facilitar su interpretación y presentación, utilizando técnicas como la imputación de datos faltantes, el mapeo de variables categóricas, el casting de variables de acuerdo a lo que se necesite. Este estudio preliminar sienta las bases para investigaciones más profundas que podrían incorporar técnicas de modelado y análisis estadístico avanzado para identificar factores de riesgo específicos y prever desenlaces clínicos.

Con este proyecto, buscamos contribuir al entendimiento de los factores que afectan la gravedad de la enfermedad en pacientes con COVID-19 y poder indicar si los pacientes presentan un menor o mayor riesgo de contraer la enfermedad de acuerdo a todos sus antecedentes.

Objetivo

El objetivo general de este proyecto es investigar cómo las condiciones de salud preexistentes influyen en la probabilidad de mortalidad y contagio en pacientes con COVID-19 y qué factores de riesgo están asociados a la hospitalización. Esto permitirá identificar patrones y factores críticos que pueden guiar las políticas de salud pública y las estrategias de intervención.

Objetivos específicos

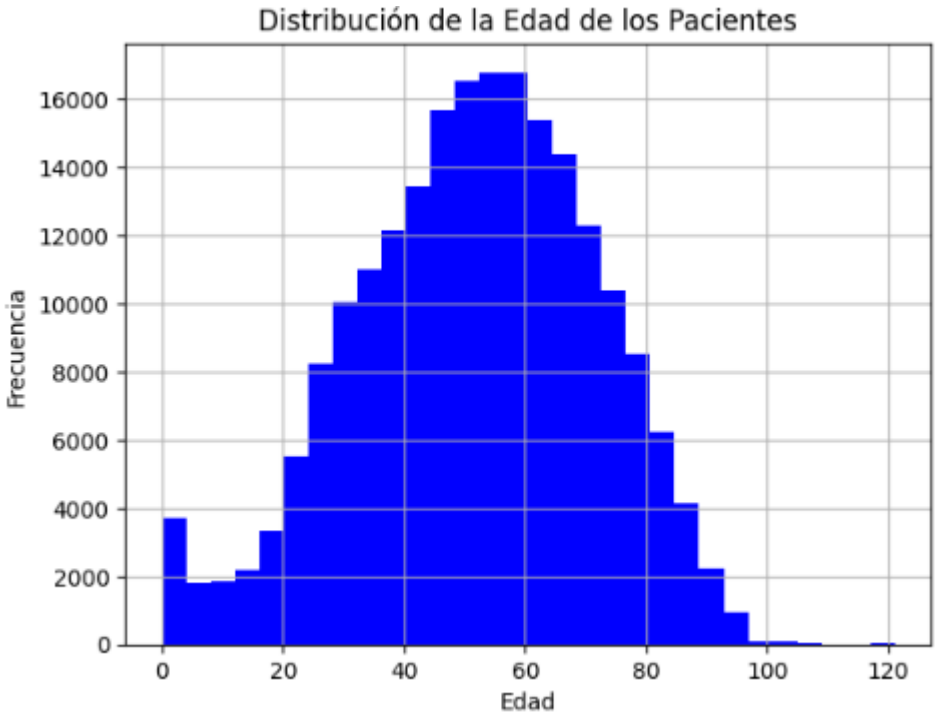
- Limpiar y organizar las columnas seleccionadas para garantizar que los datos estén listos para su análisis.
- Investigar diferencias en la tasa de mortalidad entre diferentes grupos de edad y género.
- Evaluar la relación entre las comorbilidades y el desenlace del paciente (alta o defunción) para determinar si ciertas condiciones preexistentes influyen en los resultados de los pacientes con COVID-19.

Planteamiento del problema

El problema radica en la falta de una comprensión clara y detallada sobre cómo las características demográficas y clínicas de los pacientes afectan el curso y el desenlace de la enfermedad. Además, las posibles disparidades en la atención y resultados entre diferentes regiones del país sugieren la existencia de factores adicionales que no han sido completamente explorados. Por lo tanto, este proyecto se enfoca en la limpieza, organización y análisis de un conjunto de datos de COVID-19 en México, con el objetivo de identificar cómo los padecimientos previos, como diabetes, hipertensión, obesidad, entre otros, afectan el riesgo de complicaciones graves o fallecimiento en pacientes contagiados. A través de este análisis, se busca clasificar a los pacientes en grupos de alto y bajo riesgo en función de sus comorbilidades, proporcionando así una herramienta valiosa para priorizar la atención médica y diseñar estrategias preventivas más efectivas. Esto contribuirá a una mejor comprensión del impacto de las condiciones preexistentes en la evolución del COVID-19 y apoyará la toma de decisiones en salud pública para la protección de los grupos más vulnerables.

Posibles preguntas de investigación

1. ¿Cuál es la distribución de la edad de los pacientes en el dataset?



2. ¿Cuál es la proporción de pacientes masculinos y femeninos en el dataset?
   La proporción de pacientes masculinos es de: 53.36%
   La proporción de pacientes femeninos es de: 46.64%

3. ¿Cuántos pacientes fueron diagnosticados con COVID en cada uno de los grados (1-3)?
   Los pacientes diagnosticados por grado son:
   - Positivo-Leve: 4,672
   - Positivo-Moderado: 1,787
   - Positivo-Grave:89,483

4. ¿Cuántos pacientes fueron hospitalizados en comparación con los que fueron enviados a casa?
   Los pacientes hospitalizados fueron: 116,375 y los pacientes enviados a casa (alta) fueron: 97,537

5. ¿Qué porcentaje de los pacientes diagnosticados con COVID ya tenía neumonía?
   Un 50% de los contagiados ya tenian como padecimiento neumonia

6. ¿Cuál es el porcentaje de pacientes embarazadas en el dataset?
   El porcentaje de pacientes embarazadas es de 1.45%

Descripción del dataset

El dataset tiene por nombre “COVID-19 Dataset”, este fue recuperado de la plataforma de kaggle en el siguiente link: https://www.kaggle.com/datasets/meirnizri/covid19-dataset

El dataset fue proporcionado por el gobierno mexicano. Contiene una gran cantidad de información anonimizada sobre pacientes, incluyendo condiciones preexistentes. El dataset crudo consta de 21 características únicas y 1,048,576 pacientes únicos.

En las características Booleanas:

1 significa "sí".

2 significa "no".
Los valores 97,98 y 99 representan datos faltantes.

Las variables dentro del dataset representan lo siguiente:

- **sex**: Género del paciente. 1 para femenino y 2 para masculino.
- **age**: Edad del paciente.
- **classification**: Resultados de la prueba de COVID. Los valores 1-3 indican que el paciente fue diagnosticado con COVID en diferentes grados. Un valor de 4 o superior indica que el paciente no es portador de COVID o que la prueba fue inconclusa.
- **patient type**: Tipo de atención recibida por el paciente. 1 significa que fue dado de alta a su casa, y 2 indica que fue hospitalizado.
- **pneumonia**: Indica si el paciente ya tiene inflamación de los sacos de aire.
- **pregnancy**: Indica si la paciente está embarazada o no.
- **diabetes**: Indica si el paciente tiene diabetes.
- **copd**: Indica si el paciente tiene Enfermedad Pulmonar Obstructiva Crónica (EPOC).
- **asthma**: Indica si el paciente tiene asma.
- **inmsupr**: Indica si el paciente está inmunosuprimido.
- **hypertension**: Indica si el paciente tiene hipertensión.
- **cardiovascular**: Indica si el paciente tiene una enfermedad relacionada con el corazón o los vasos sanguíneos.
- **renal chronic**: Indica si el paciente tiene una enfermedad renal crónica.
- **other disease**: Indica si el paciente tiene alguna otra enfermedad.
- **obesity**: Indica si el paciente tiene obesidad.
- **tobacco**: Indica si el paciente es usuario de tabaco.
- **usmr**: Indica si el paciente fue tratado en unidades médicas de primer, segundo o tercer nivel.
- **medical unit**: Tipo de institución del Sistema Nacional de Salud que brindó la atención.
- **intubed**: Indica si el paciente fue conectado a un ventilador.
- **icu**: Indica si el paciente fue admitido en una Unidad de Cuidados Intensivos (UCI).
- **date died**: Si el paciente falleció, indica la fecha de defunción; de lo contrario, aparece el valor 9999-99-99.

Tipos de datos.

```
Tipos de datos por columna:
USMER                  int64
MEDICAL_UNIT           int64
SEX                    int64
PATIENT_TYPE           int64
DATE_DIED             object
INTUBED                int64
PNEUMONIA              int64
AGE                    int64
PREGNANT               int64
DIABETES               int64
COPD                   int64
ASTHMA                 int64
INMSUPR                int64
HIPERTENSION           int64
OTHER_DISEASE          int64
CARDIOVASCULAR         int64
OBESITY                int64
RENAL_CHRONIC          int64
TOBACCO                int64
CLASIFFICATION_FINAL   int64
ICU                    int64
dtype: object
```

Posible solución

La solución propuesta consiste en llevar a cabo un proceso integral de limpieza y procesamiento de datos utilizando Python y la biblioteca Pandas. Este proceso incluye la eliminación de valores nulos, la transformación de fechas de ingreso a valores numéricos que reflejen la antigüedad del paciente, y un análisis exploratorio inicial para identificar relaciones entre variables clave. Estas tareas permitirán preparar los datos de manera eficiente para su posterior análisis y modelado, asegurando la calidad y coherencia de la información.

El procesamiento cuidadoso de los datos es fundamental para garantizar que estén en un formato adecuado para futuras etapas de análisis avanzado, incluyendo la implementación de modelos de Machine Learning. Este enfoque preliminar sentará las bases para desarrollar un modelo predictivo que pueda identificar el nivel de riesgo de los pacientes con mayor precisión, facilitando así la toma de decisiones informadas y la planificación estratégica en el ámbito de la salud.

Para futuras consideraciones queremos realizar lo siguiente como parte de la solución:

- Análisis Estadístico: Realizar un análisis de incidencia para identificar las condiciones preexistentes más comunes entre los pacientes fallecidos y hospitalizados. Utilizar técnicas estadísticas para correlacionar estas condiciones con los desenlaces fatales y la necesidad de hospitalización.
- Modelos Predictivos: Desarrollar modelos predictivos para anticipar la probabilidad de mortalidad y hospitalización en función de las condiciones preexistentes y otras variables relevantes. Esto puede ayudar a priorizar recursos y orientar intervenciones.

- Optimización de Recursos: Usar los hallazgos para mejorar la asignación de recursos médicos, como ventiladores y camas de UCI, y diseñar campañas de prevención específicas para los grupos de mayor riesgo.

Consideraciones futuras

1. Adaptación de Políticas: Actualizar las políticas de salud pública basadas en nuevos hallazgos para proteger mejor a las poblaciones vulnerables y mejorar la preparación para futuras pandemias.
2. Investigación Continua: Continuar investigando el impacto de nuevas variantes del virus y otros factores emergentes que puedan influir en los desenlaces de la enfermedad.
3. Impacto a largo plazo: Evaluar las consecuencias a largo plazo en la salud de los pacientes sobrevivientes y en los sistemas de salud.
4. Identificar las condiciones de salud preexistentes más comunes entre los pacientes fallecidos por COVID-19.
5. Determinar qué condición preexistente tiene la mayor correlación con la mortalidad.
6. Analizar las características comunes de los pacientes que requieren hospitalización.
7. Investigar si hay grupos de riesgo con múltiples condiciones preexistentes que presentan una mayor tasa de mortalidad.
8. Evaluar la importancia de las condiciones preexistentes en la necesidad de hospitalización.
9. Examinar la relación entre la hospitalización y variables como edad y género.

Conclusión

El análisis de cómo las condiciones de salud preexistentes afectan la mortalidad y hospitalización por COVID-19 proporcionará información valiosa para mejorar la respuesta a la pandemia y fortalecer los sistemas de salud. Al identificar y abordar los factores de riesgo más críticos, se pueden implementar estrategias más efectivas para proteger a las poblaciones vulnerables y gestionar recursos de manera más eficiente.

Proyecto Tecnolochicas MÓDULO 3 Team 20

Integrantes:
- De La Cruz Munguia Arely arelydlcm07@gmail.com
- Pérez Mendoza Leisly leislymp@utbispuebla.edu.mx
- Valle Núñez Gabriela gabrielavallenunez@gmail.com

Análisis Exploratorio y Visualización de Datos de Pacientes con COVID-19: Un Enfoque en Condiciones Preexistentes y Resultados Clínicos Globales

Introducción

En el contexto de la pandemia global de COVID-19, el acceso a grandes volúmenes de datos sobre pacientes ha abierto nuevas oportunidades para mejorar nuestra comprensión de la enfermedad, sus factores de riesgo y sus resultados clínicos. Sin embargo, para aprovechar al máximo esta información, es fundamental aplicar enfoques sólidos de ciencia de datos. La calidad de las conclusiones derivadas de los datos depende en gran medida de su correcta limpieza y análisis exploratorio. Estos pasos iniciales permiten identificar inconsistencias, valores atípicos y patrones ocultos, y son cruciales para construir un conjunto de datos confiable.

El análisis exploratorio de datos (EDA, por sus siglas en inglés) es esencial para descubrir relaciones clave entre las condiciones preexistentes de los pacientes y sus desenlaces clínicos tras una infección por COVID-19. A través de la visualización de datos, podemos transformar cifras crudas en representaciones gráficas claras, facilitando la comprensión y la comunicación de los hallazgos a un público amplio, desde investigadores hasta responsables de la toma de decisiones en salud pública.

La limpieza de datos y el EDA no solo ayudan a garantizar la integridad de los datos, sino que también sientan las bases para el desarrollo de modelos predictivos que podrían aplicarse para identificar grupos de riesgo, predecir desenlaces clínicos y optimizar los tratamientos. Además, la visualización de los resultados puede ofrecer un medio poderoso para identificar tendencias globales y apoyar el diseño de intervenciones efectivas en diferentes contextos epidemiológicos.

En este proyecto, nos proponemos realizar un análisis exhaustivo de un conjunto de datos globales de pacientes con COVID-19, con un enfoque especial en sus condiciones preexistentes y resultados clínicos. Mediante la aplicación de técnicas de limpieza de datos, análisis exploratorio y visualización, buscamos no solo comprender mejor las dinámicas de la enfermedad, sino también sentar las bases para futuras investigaciones y aplicaciones clínicas basadas en ciencia de datos.

Objetivo

El objetivo principal de este proyecto es analizar el impacto de las condiciones de salud preexistentes en los resultados clínicos de los pacientes con COVID-19 alrededor del mundo, incluyendo su influencia en la mortalidad, la probabilidad de contagio y los factores asociados a la hospitalización. A través del análisis exploratorio y la visualización de datos, se busca identificar patrones y factores críticos que puedan guiar políticas de salud pública y diseñar estrategias más efectivas.

Objetivos específicos

- **Limpieza y organización de datos**: Realizar la limpieza y organización de un nuevo conjunto de datos sobre casos de COVID-19 a nivel mundial, seleccionando las columnas relevantes para asegurar que los datos estén preparados para el análisis.
- **Integración de datos**: Relacionar el nuevo conjunto de datos con el previamente utilizado, garantizando la coherencia y completitud de la información, y preparando los datos para el análisis conjunto.
- **Análisis de comorbilidades y resultados clínicos**: Evaluar la relación entre comorbilidades y el desenlace clínico de los pacientes (alta o defunción), para determinar la influencia de ciertas condiciones preexistentes en los resultados de pacientes con COVID-19 a nivel global.
- **Exploración de variables**: Identificar y clasificar las variables presentes en el conjunto de datos (categóricas y numéricas), aplicando normalización cuando sea necesario para estandarizar los datos.
- **Visualización y análisis estadístico**: Visualizar los datos mediante gráficos y realizar un análisis estadístico de las distribuciones y correlaciones entre variables, para identificar patrones clave que puedan guiar futuras investigaciones.

Planteamiento del problema

El problema central radica en la falta de una comprensión profunda sobre cómo las características demográficas y clínicas de los pacientes, como las condiciones preexistentes, influyen en el curso y desenlace de la enfermedad por COVID-19. Además, la existencia de disparidades en los resultados clínicos entre distintas regiones del mundo sugiere la presencia de factores adicionales que aún no han sido completamente explorados. Este proyecto se enfoca en la limpieza, organización y análisis de un conjunto global de datos de pacientes con COVID-19, con énfasis en cómo condiciones preexistentes como diabetes, hipertensión y obesidad impactan el riesgo de complicaciones graves o fallecimiento.

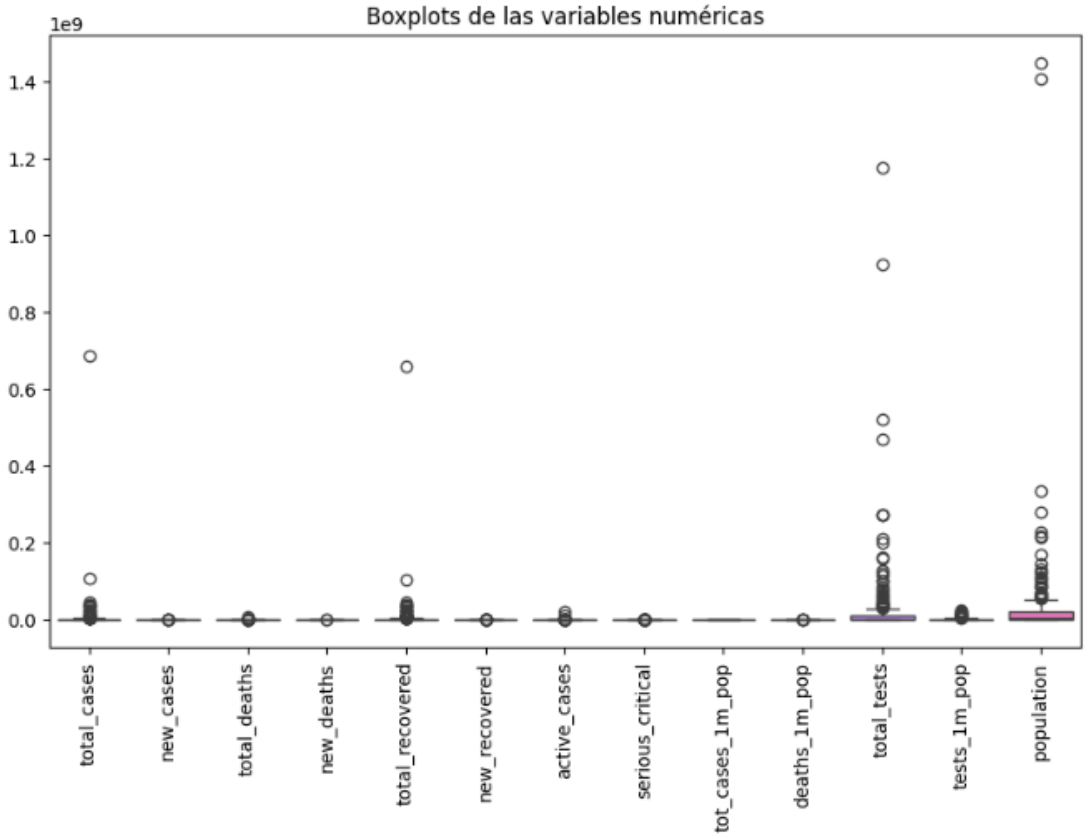
A través del análisis exploratorio y visualización de datos, se busca identificar patrones clave y clasificar a los pacientes en grupos de alto y bajo riesgo en función de sus comorbilidades. Este enfoque no solo ayudará a priorizar la atención médica, sino que también proporcionará una herramienta valiosa para diseñar estrategias preventivas más efectivas, adaptadas a las necesidades de cada grupo de riesgo. De esta manera, el proyecto contribuirá a una mejor comprensión del impacto de las condiciones preexistentes en la evolución del COVID-19, apoyando la toma de decisiones en salud pública para proteger a los grupos más vulnerables y optimizar los recursos médicos.

Posibles preguntas de investigación

1. **¿Cuáles son las estadísticas descriptivas del total de casos y de nuevos casos de COVID-19?**

```
Estadísticas descriptivas para la columna: total_cases
           Estadística         Valor
0             Promedio  5.922601e+06
1              Mediana  2.080335e+05
2  Media Truncada (5%)  1.340789e+06
3  Desviación Estándar  4.593526e+07
4                Rango  6.870217e+08
5         Percentil 25  2.517500e+04
6         Percentil 75  1.324580e+06
7   Rango Intercuartil  1.299404e+06

================================================

Estadísticas descriptivas para la columna: new_cases
           Estadística         Valor
0             Promedio    125.094828
1              Mediana      0.000000
2  Media Truncada (5%)      0.000000
3  Desviación Estándar   1339.736929
4                Rango  14511.000000
5         Percentil 25      0.000000
6         Percentil 75      0.000000
7   Rango Intercuartil      0.000000
```

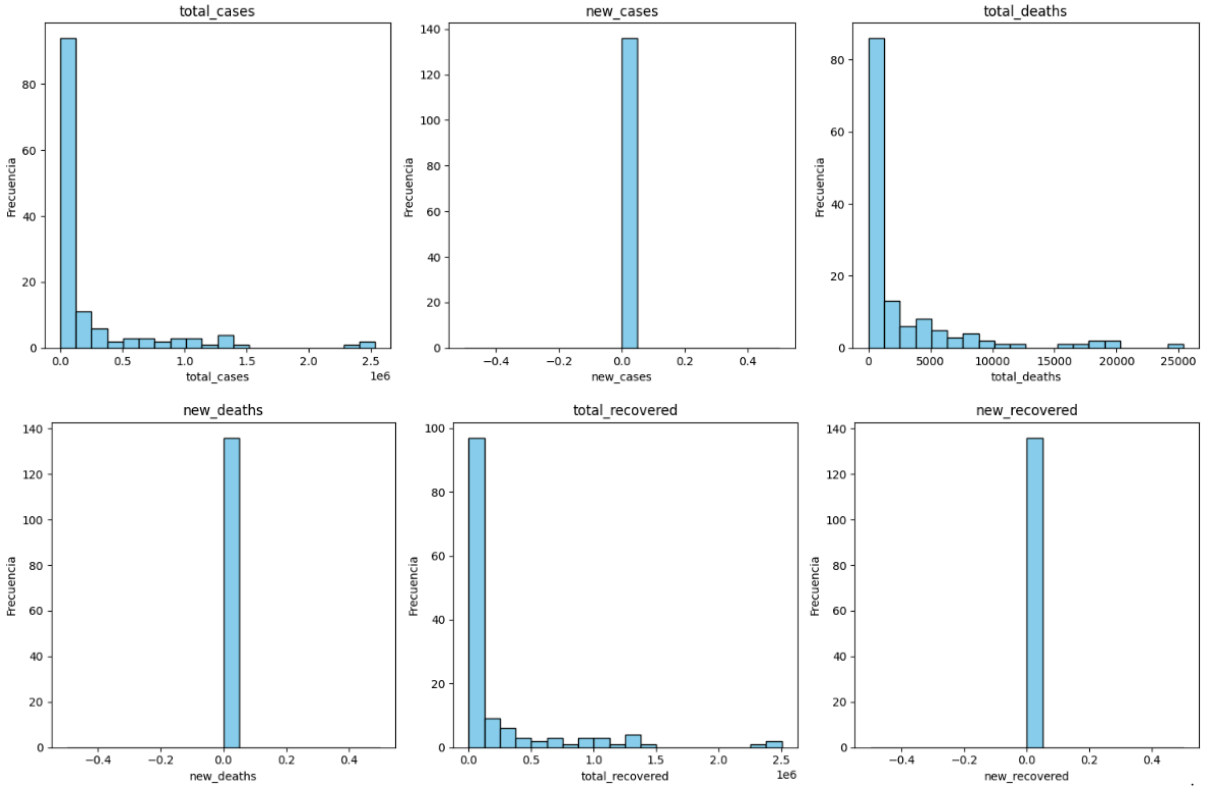
2. **Visualización gráfica mediante 'boxplots' de las variables numéricas en el data set**

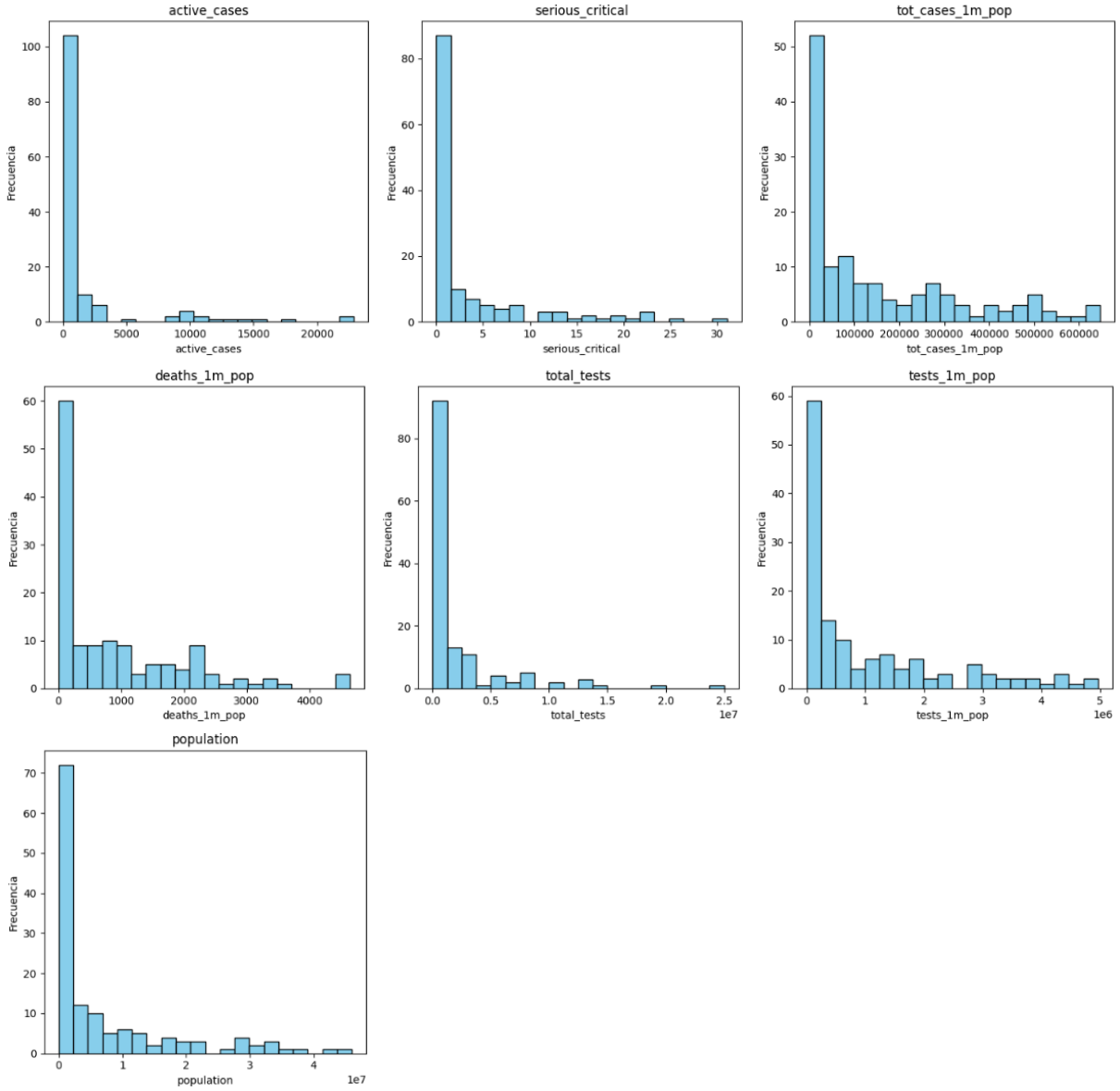


3. **¿Cuáles son las estadísticas de las variables numéricas del data set considerando 'outliers'? ¿Cuáles son sin considerarlos?**

```
Estadísticas antes de filtrar:
                          mean     median           std
total_cases       5.922601e+06   208033.5  4.593526e+07
new_cases         1.250948e+02        0.0  1.339737e+03
total_deaths      5.916825e+04     1971.0  4.609176e+05
new_deaths        2.586207e-02        0.0  2.779394e-01
total_recovered   5.604029e+06   133659.5  4.411465e+07
new_recovered     1.794914e+02        0.0  1.602637e+03
active_cases      1.699781e+05      705.5  1.561362e+06
serious_critical  2.064138e+02        1.0  2.580700e+03
tot_cases_1m_pop  1.950318e+05   123760.5  1.983633e+05
deaths_1m_pop     1.211015e+03      779.5  1.283547e+03
total_tests       3.007449e+07  1671065.5  1.143474e+08
tests_1m_pop      1.961049e+06   717353.5  3.531832e+06
population        3.424541e+07  5533165.0  1.377903e+08

Estadísticas después de filtrar:
                          mean     median           std
total_cases       2.538560e+05    43133.0  4.804614e+05
new_cases         0.000000e+00        0.0  0.000000e+00
total_deaths      2.613632e+03      402.5  4.718002e+03
new_deaths        0.000000e+00        0.0  0.000000e+00
total_recovered   2.382020e+05    32831.5  4.760592e+05
new_recovered     0.000000e+00        0.0  0.000000e+00
active_cases      1.837022e+03      116.0  4.270799e+03
serious_critical  3.544118e+00        0.0  6.394289e+00
tot_cases_1m_pop  1.548342e+05    80118.5  1.790661e+05
deaths_1m_pop     8.888897e+02      414.5  1.064556e+03
total_tests       2.073955e+06   362186.0  3.941998e+06
tests_1m_pop      1.036060e+06   393308.0  1.315977e+06
population        7.150891e+06  1956102.0  1.057144e+07
```

4. **Visualización gráfica mediante 'histogramas' de las variables numéricas en el data set. Sin normalizar los datos**

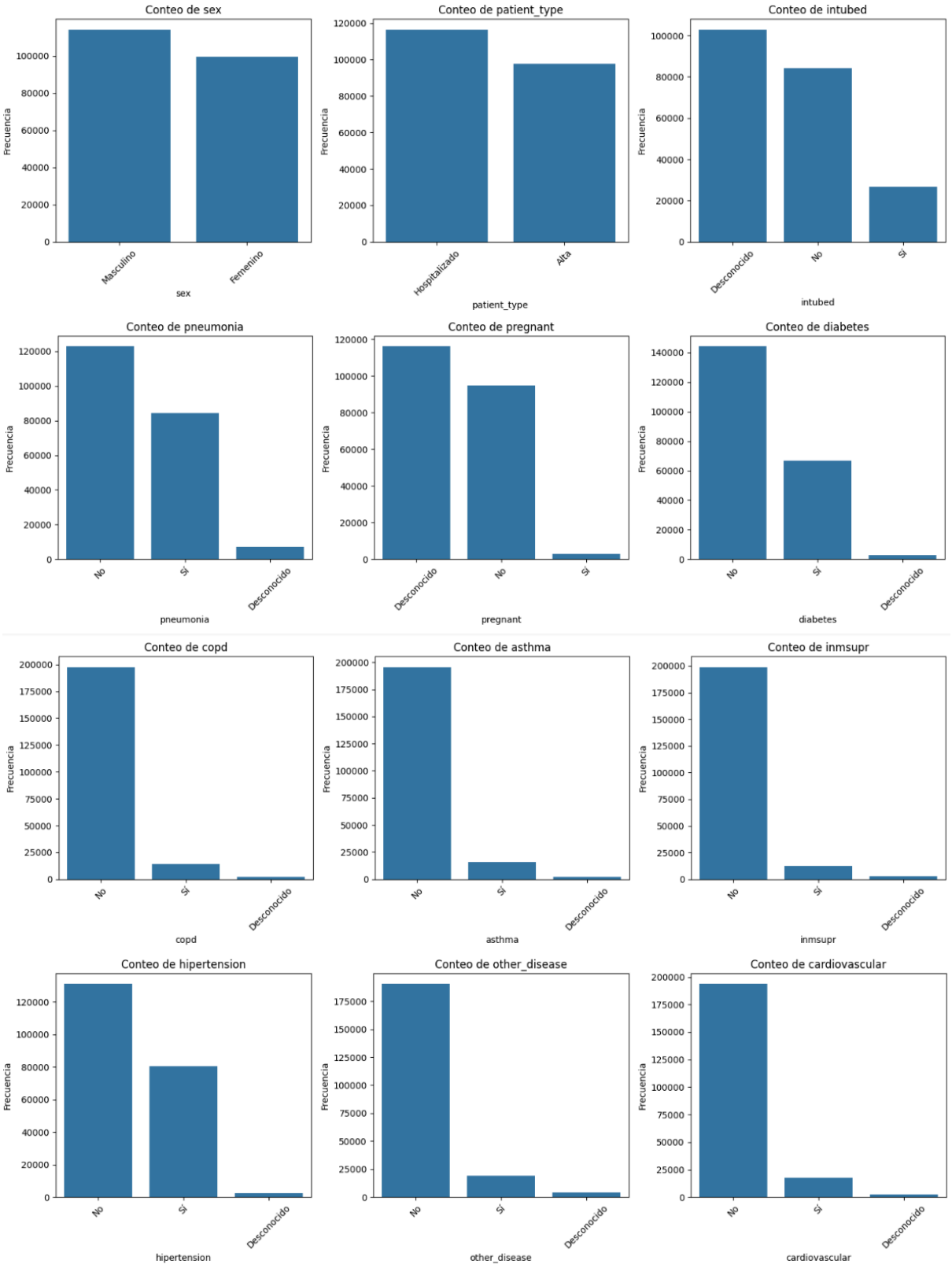


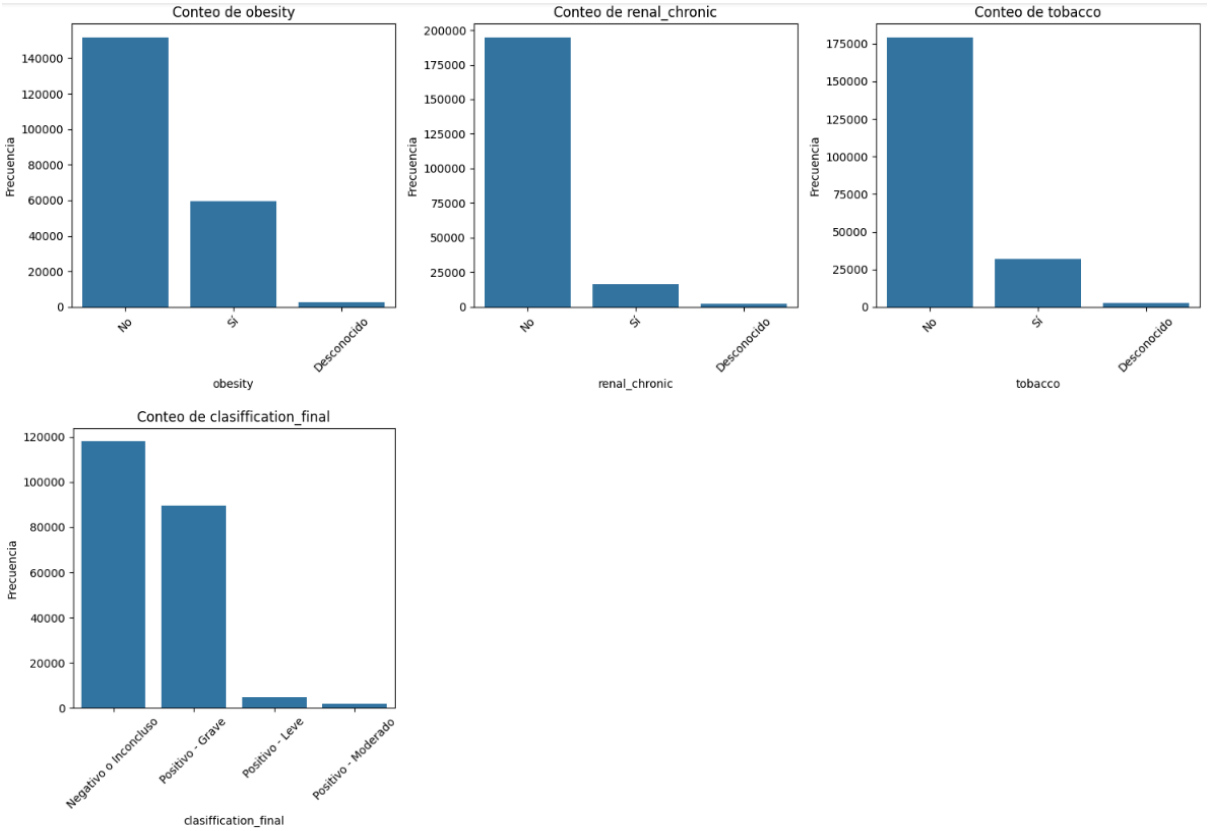


5. **¿Cuál es la asimetría y la curtosis de la variable 'total_deaths' ?**

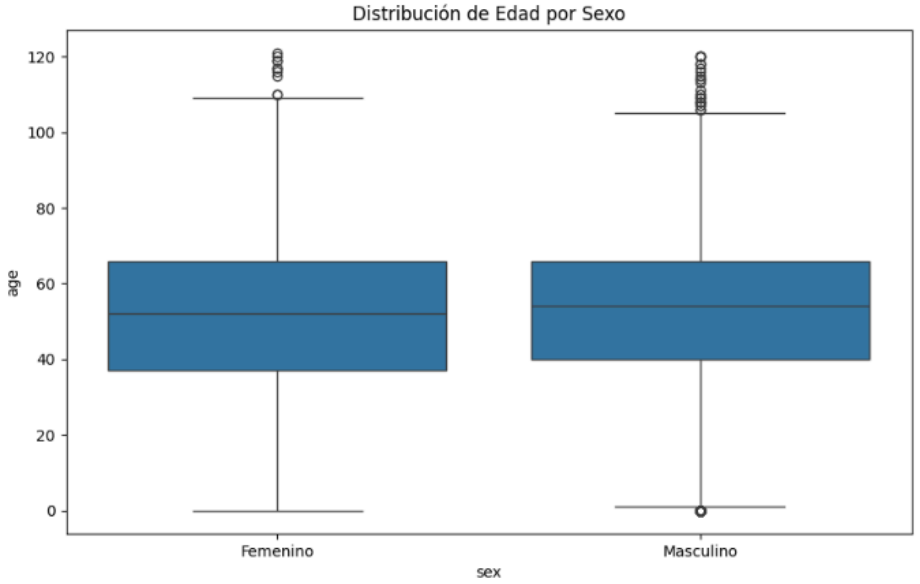
```
Asimetría para total_cases: 2.795689146580793
Curtosis para total_cases: 8.531498558351307


Asimetría para new_cases: 0.0
Curtosis para new_cases: 0.0


Asimetría para total_deaths: 2.666721346785501
Curtosis para total_deaths: 7.393748923084674


Asimetría para new_deaths: 0.0
Curtosis para new_deaths: 0.0
```

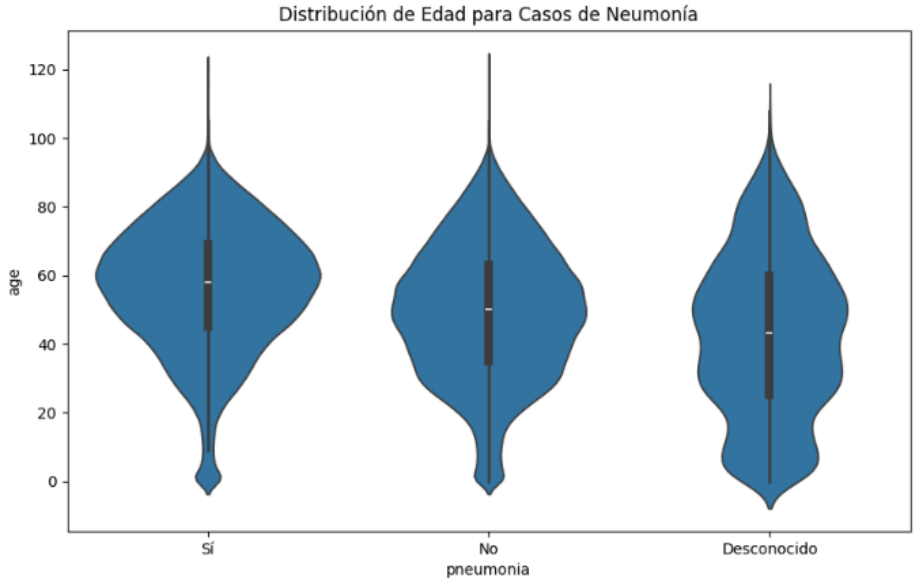
6. **Visualización con gráficos de barra de las variables categóricas**


Conteo de sex
Frecuencia
Masculino
Femenino
sex
Conteo de patient_type
Hospitalizado
Alta
patient_type
Conteo de intubed
Desconocido
No
Si
intubed
Conteo de pneumonia
pneumonia
Conteo de pregnant
pregnant
Conteo de diabetes
diabetes
Conteo de copd
copd
Conteo de asthma
asthma
Conteo de inmsupr
inmsupr
Conteo de hipertension
hipertension
Conteo de other_disease
other_disease
Conteo de cardiovascular
cardiovascular

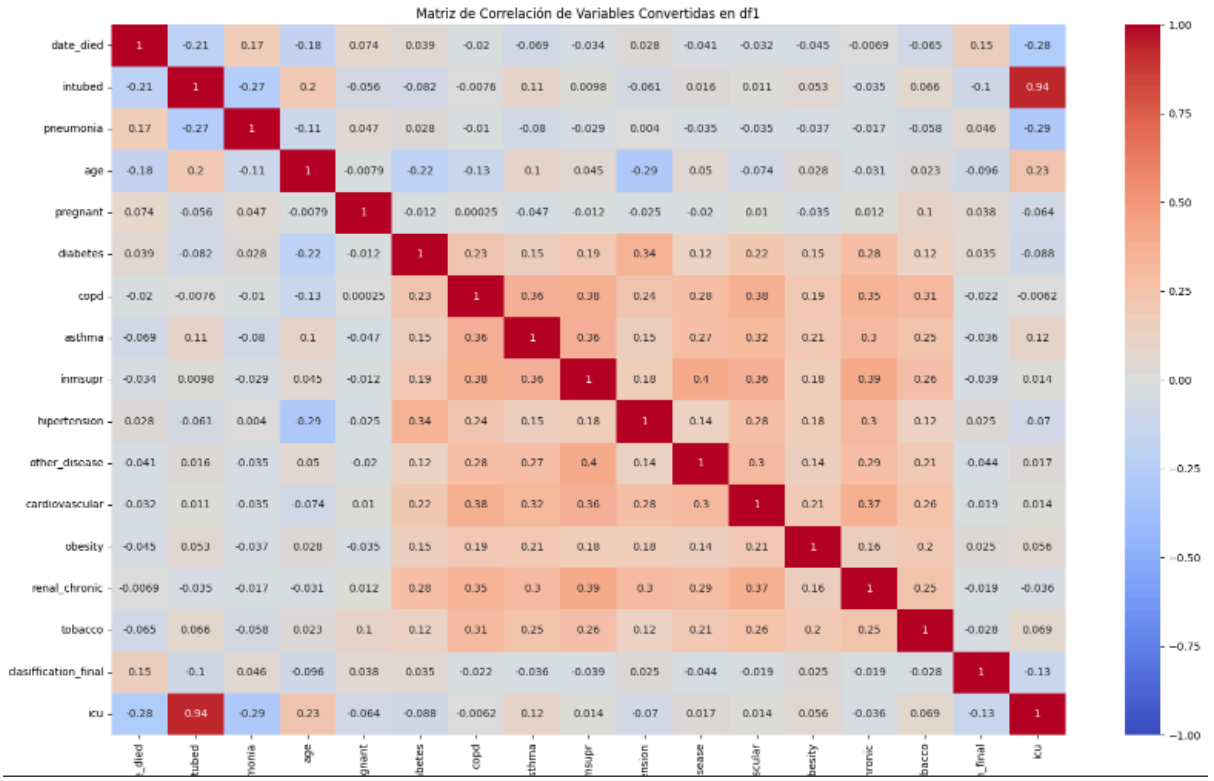


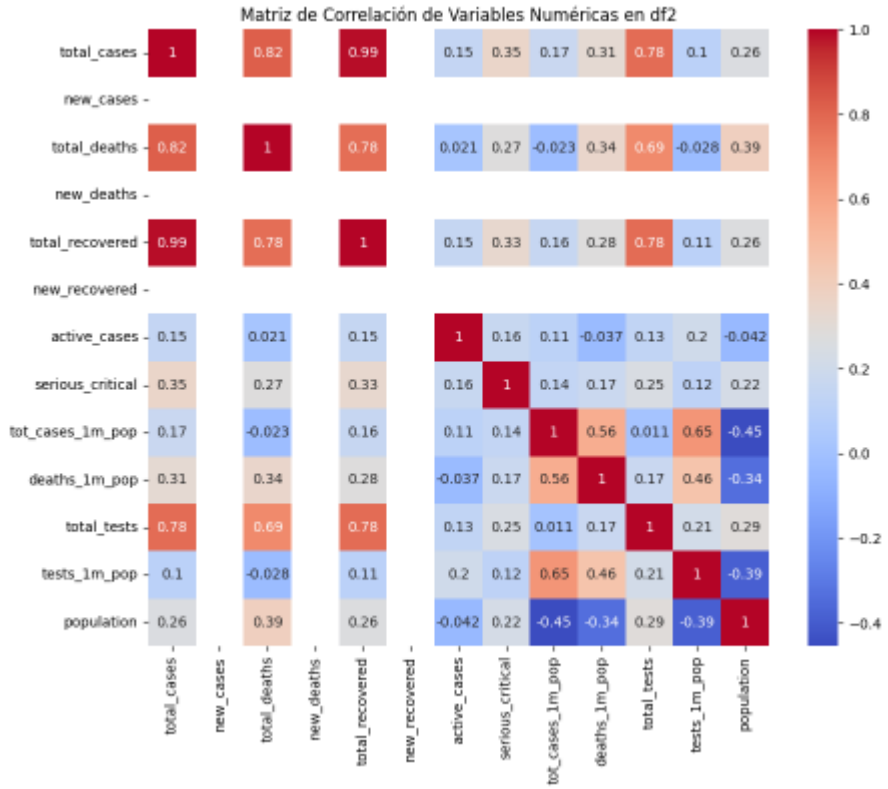
7. **Visualización gráfica con “barplots” y “violinplots” de la distribución entre las variables “age” - “sex” y “age” - “pneumonia”**



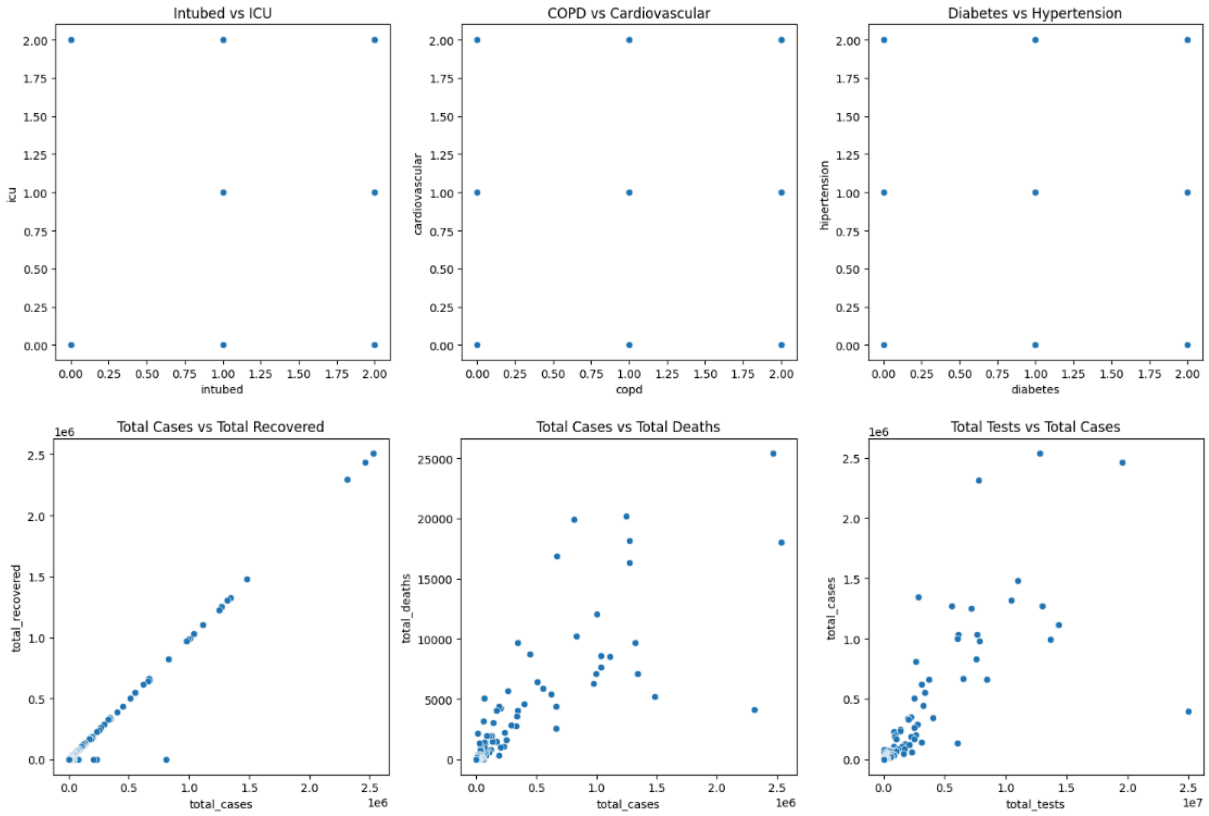


8. **¿Cuál es la correlación entre las variables en ambos data sets?**



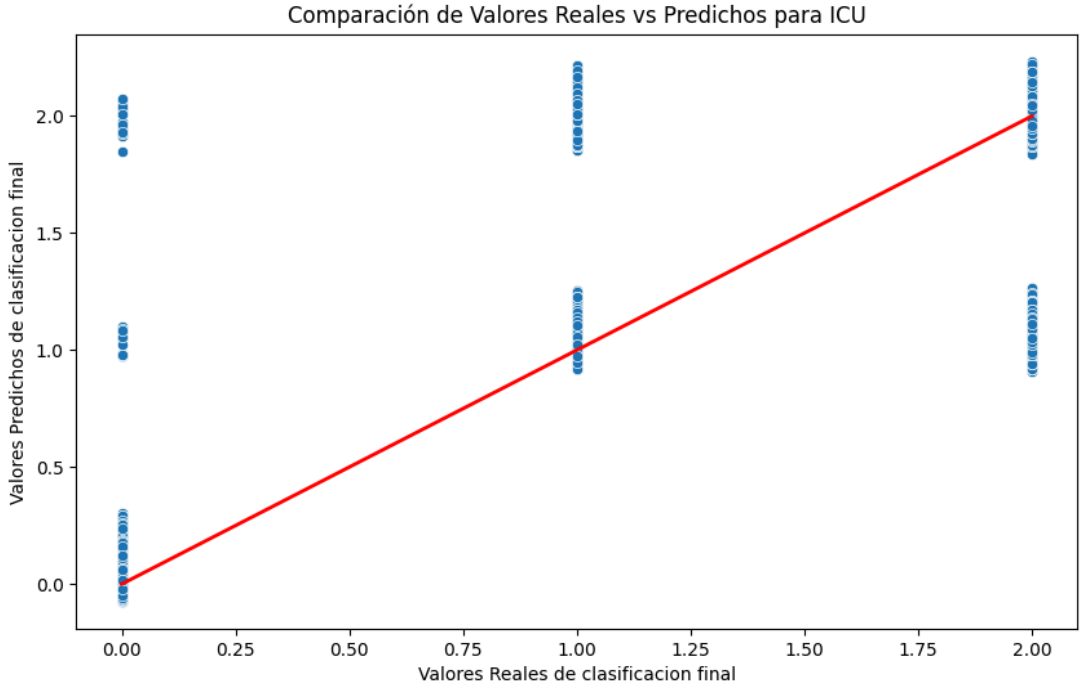


9. Gráficos de dispersión



10. Regresión lineal

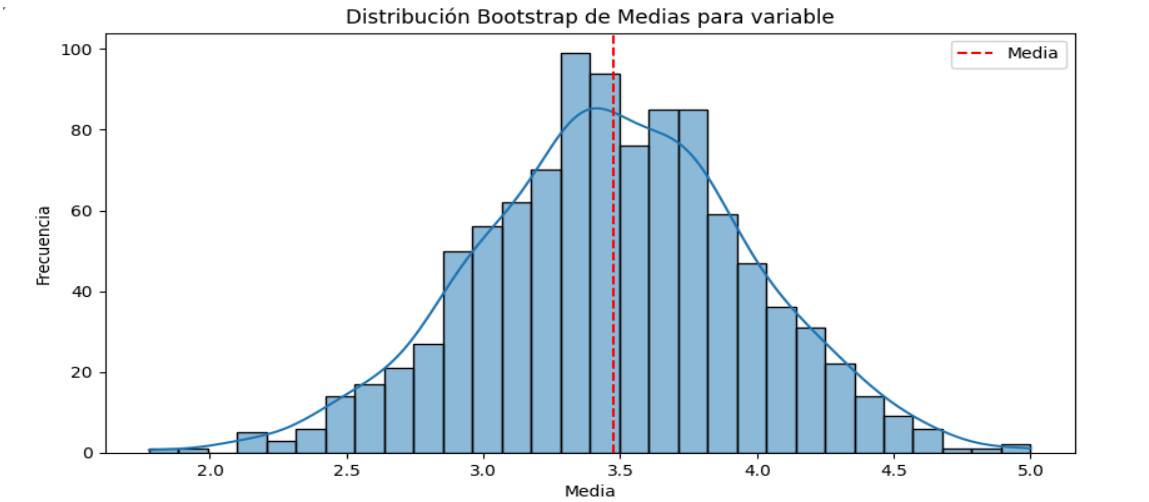
Dados los datos que tenemos pudimos notar que la regresión lineal no era un modelo que nos sirviera para nuestro análisis.



11. Distribución muestral mediante técnicas de 'Bootstrap'

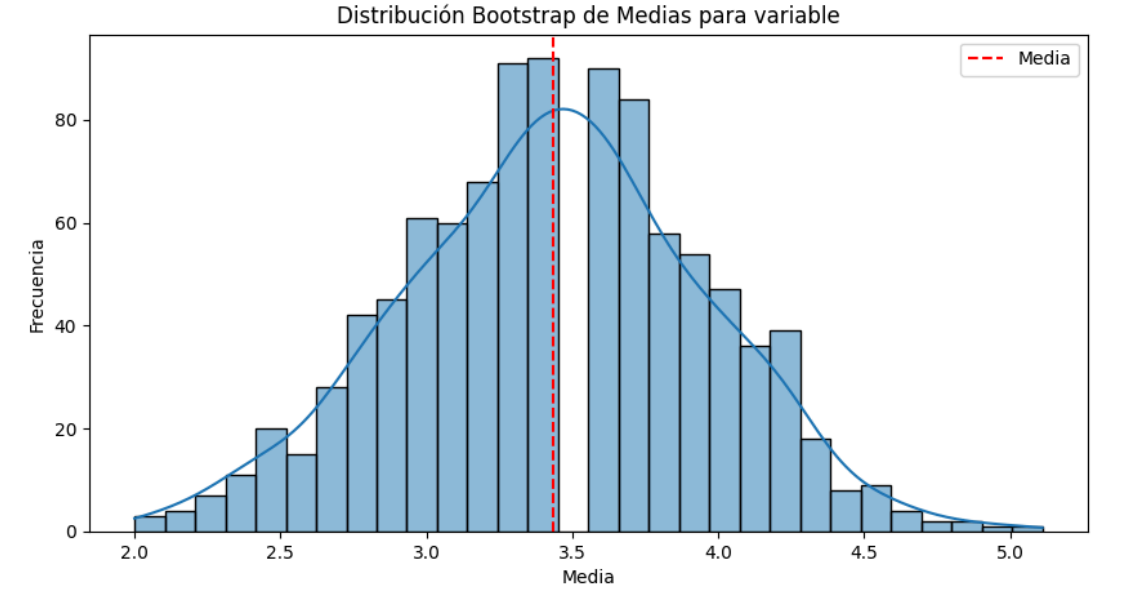
a) La distribución, su asimetría y curtosis

```
Resultados de Bootstrap para variable:
Asimetría: -0.07, Curtosis: 0.04
```



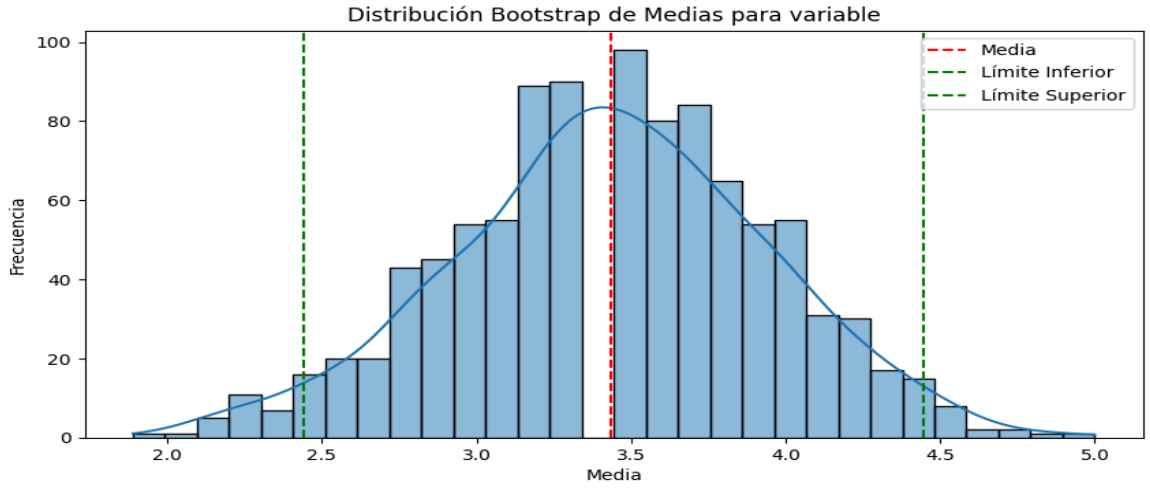
b) El error estándar

```
Resultados de Bootstrap para variable:
Error Estándar: 0.52
```



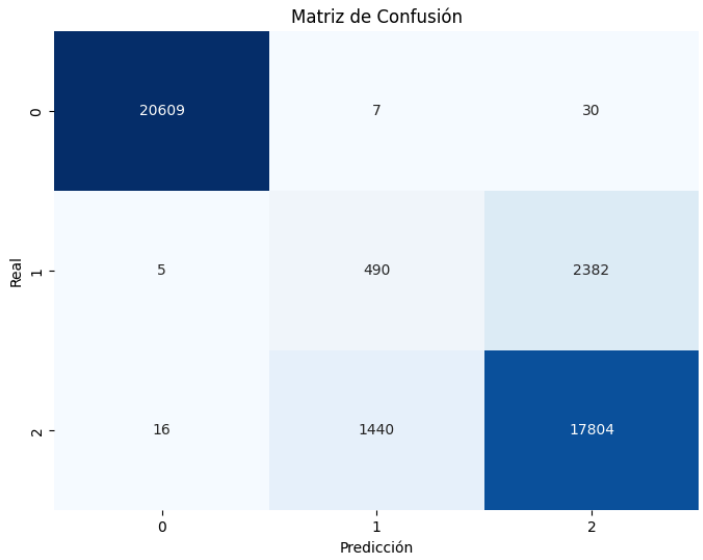
c) El intervalo de confianza que te parezca más apropiado

```
Resultados de Bootstrap para variable:
Intervalo de Confianza (95%): (2.441666666666667, 4.444444444444445)
```



12. Modelos

Árbol de decisión



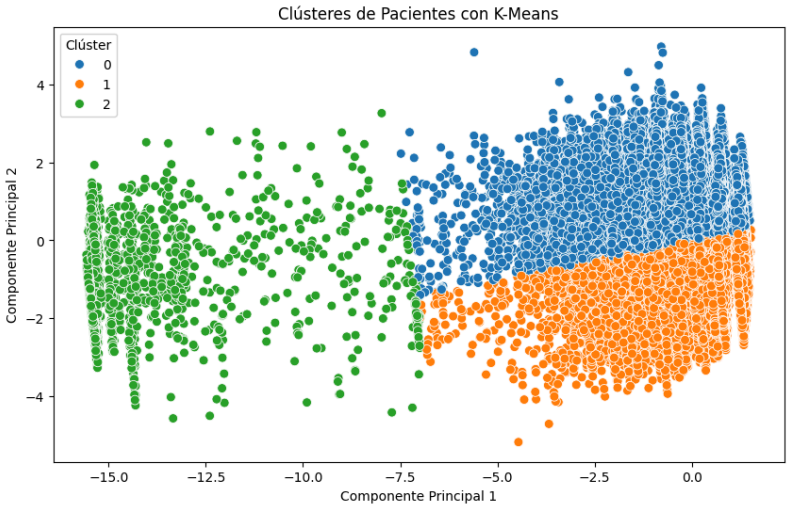
Reporte de Clasificación:

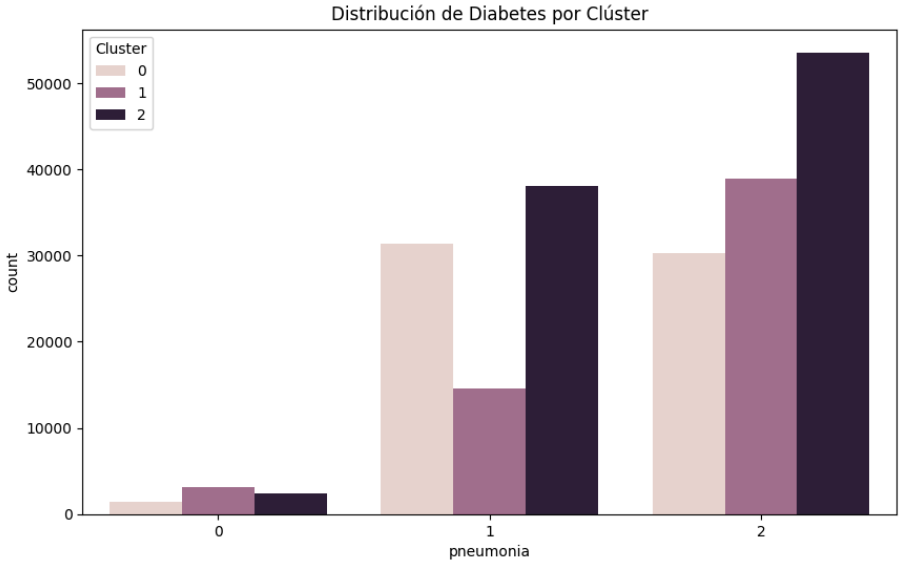
	precision	recall	f1-score	support
0.0	0.998982	0.998208	0.998595	20646.00000
1.0	0.252969	0.170316	0.203573	2877.00000
2.0	0.880689	0.924403	0.902016	19260.00000
accuracy	0.909310	0.909310	0.909310	0.90931
macro avg	0.710880	0.697642	0.701395	42783.00000
weighted avg	0.895562	0.909310	0.901655	42783.00000

Regresión logística

```
Exactitud de la Regresión Logística: 0.9321926933595119
              precision    recall  f1-score   support

         0.0       1.00      1.00      1.00     20646
         1.0       0.60      0.01      0.02      2877
         2.0       0.87      1.00      0.93     19260

    accuracy                           0.93     42783
   macro avg       0.82      0.67      0.65     42783
weighted avg       0.91      0.93      0.90     42783
```

K-means





De los 3 modelos empleados solo regresión logística tuvo un mejor desempeño prediciendo los valores de 1 como si al ingreso de cuidados intensivos. Esto se debe al tipo de variables que estamos usando y a la forma en que mapeamos todo.

Descripción del data set

- Se utilizaron 2 data sets, el del módulo 2 “COVID-19 Dataset” cuya descripción se encuentra en el apartado anterior, y uno nuevo “corona_virus”, ambos obtenidos del sitio web “Kaggle”.
    - Data set 1: https://www.kaggle.com/datasets/meirnizri/covid19-dataset

- Data set 2: https://www.kaggle.com/code/shushilshah/notebook396826ec64/input

Tipos de datos (Data set 2)

```
Tipos de datos por columna:
country_other        object
total_cases         float64
new_cases           float64
total_deaths        float64
new_deaths          float64
total_recovered     float64
new_recovered       float64
active_cases        float64
serious_critical    float64
tot_cases_1m_pop    float64
deaths_1m_pop       float64
total_tests         float64
tests_1m_pop        float64
population          float64
```

Posible solución

Para abordar la falta de comprensión sobre el impacto de las características demográficas y clínicas en los resultados de los pacientes con COVID-19, este proyecto propone un enfoque basado en la ciencia de datos, centrado en tres fases clave: la limpieza y organización de datos, el análisis exploratorio y la visualización, y como perspectiva el desarrollo de modelos predictivos, por ejemplo:

1. Modelos predictivos y clasificación de riesgo: A partir del análisis exploratorio, se desarrollarán modelos predictivos de clasificación que permitan identificar grupos de pacientes en alto y bajo riesgo en función de sus comorbilidades. Estos modelos pueden basarse en algoritmos de machine learning como árboles de decisión, random forest o regresión logística, ajustados para predecir el riesgo de complicaciones graves o muerte en los pacientes.

2. Aplicación de los resultados: Los resultados obtenidos serán útiles para la toma de decisiones en salud pública. Al clasificar a los pacientes según su nivel de riesgo, se podrá priorizar la atención médica de manera más eficiente, asignando recursos a los grupos más vulnerables. Además, los hallazgos contribuirán al diseño de estrategias preventivas y políticas de intervención más precisas para mitigar los efectos de futuras oleadas de COVID-19 o pandemias similares.

Consideraciones futuras

1. Adaptación de Políticas: Actualizar las políticas de salud pública basadas en nuevos hallazgos para proteger mejor a las poblaciones vulnerables y mejorar la preparación para futuras pandemias.
2. Investigación Continua: Continuar investigando el impacto de nuevas variantes del virus y otros factores emergentes que puedan influir en los desenlaces de la enfermedad.
3. Impacto a largo plazo: Evaluar las consecuencias a largo plazo en la salud de los pacientes sobrevivientes y en los sistemas de salud.
4. Identificar las condiciones de salud preexistentes más comunes entre los pacientes fallecidos por COVID-19.

Conclusión

El análisis de cómo las condiciones de salud preexistentes impactan la mortalidad y hospitalización por COVID-19 aporta información clave para fortalecer la respuesta global ante la pandemia y mejorar la resiliencia de los sistemas de salud. Al identificar y abordar los factores de riesgo más significativos, es posible implementar estrategias más efectivas para proteger a las poblaciones vulnerables y optimizar el uso de recursos a nivel mundial. Esta comprensión no solo ayudará a reducir el impacto de COVID-19, sino que también fortalecerá la preparación y respuesta ante futuras emergencias de salud pública.