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

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

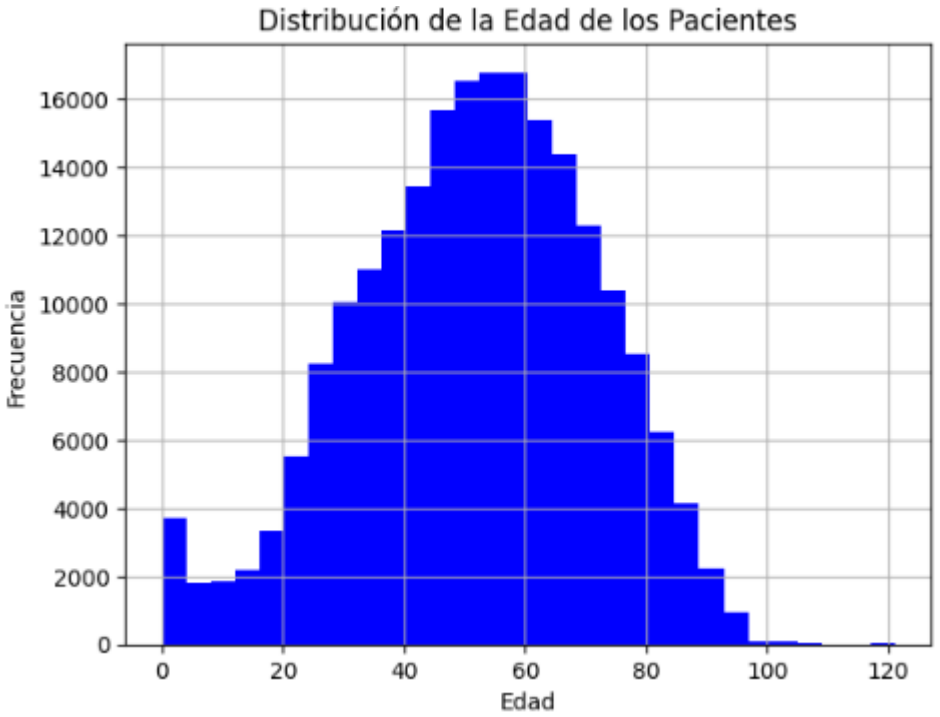
- Limpiar y organizar las columnas seleccionadas para garantizar que los datos estén listos para su análisis.
- Investigar diferencias en la tasa de mortalidad entre diferentes grupos de edad y género.
- Evaluar la relación entre las comorbilidades y el desenlace del paciente (alta o defunción) para 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.