

Reporte de análisis exploratorio de datos: enfermedades cardiovasculares

Fecha de elaboración: 17 de junio de 2025

Caso: Predicción de Enfermedades Cardiovasculares

Nombre: Bernardo Corona Domínguez

1. Objetivos del reporte

El presente documento tiene como propósito exponer, de manera clara y accesible, los principales hallazgos del análisis exploratorio realizado sobre el conjunto de datos heart_2020_cleaned.csv. El objetivo es identificar patrones, distribuciones y relaciones relevantes para la construcción futura de un modelo predictivo que permita anticipar la presencia de enfermedades cardiovasculares en la población.

2. Descripción del conjunto de datos y sus variables

El conjunto de datos analizado contiene 319,795 registros y 18 variables, entre las que se incluyen datos demográficos (edad, sexo, raza), factores de riesgo (tabaquismo, consumo de alcohol, historial de diabetes, asma, enfermedad renal, cáncer de piel), indicadores de salud física y mental, horas de sueño y la variable objetivo: HeartDisease.

Variables numéricas:

- BMI (Índice de Masa Corporal)
- PhysicalHealth (Días de mala salud física en el mes)
- MentalHealth (Días de mala salud mental en el mes)
- SleepTime (Horas promedio de sueño por noche)

Variables categóricas:

- HeartDisease
- Smoking
- AlcoholDrinking
- Stroke
- DiffWalking
- Sex
- AgeCategory
- Race
- Diabetic
- PhysicalActivity

- GenHealth
- Asthma
- KidneyDisease
- SkinCancer

3. Principales hallazgos

3.1 Análisis univariado

- La mayoría de los registros corresponde a personas adultas mayores de 50 años, con un predominio en el grupo de 60-64 años.
- La variable objetivo HeartDisease presenta un desbalance: aproximadamente 8% de la población estudiada reporta enfermedad cardiovascular.
- El índice de masa corporal (BMI) tiende a concentrarse entre 25 y 35, reflejando una alta prevalencia de sobrepeso y obesidad.
- La distribución de días de mala salud física y mental muestra asimetría positiva, con la mayoría de la población reportando menos de 10 días en el mes.

3.2 Análisis bivariado

- Se observa una mayor prevalencia de enfermedad cardiovascular en personas con antecedentes de diabetes, hipertensión, y aquellas que presentan dificultades para caminar.
- El riesgo de enfermedad incrementa con la edad, especialmente a partir de los 55 años.
- El hábito de fumar y el consumo de alcohol están correlacionados con una mayor proporción de casos de enfermedad cardiovascular, aunque el efecto es moderado.
- Existen diferencias notables entre sexos y razas respecto a la incidencia de la enfermedad, siendo los hombres y ciertas razas más afectados.

3.3 Análisis multivariado

- El análisis de correlación entre variables numéricas revela una asociación positiva moderada entre BMI y mala salud física, pero no hay correlaciones fuertes entre las variables numéricas.
- La acumulación de factores de riesgo (edad avanzada, diabetes, inactividad física, mala salud general) coincide con una mayor prevalencia de enfermedades del corazón.
- El análisis cruzado entre categorías de edad y sexo muestra que la incidencia de enfermedad cardiovascular es más alta en hombres mayores de 60 años.

4. Conclusiones

El análisis exploratorio permite afirmar que la presencia de enfermedades cardiovasculares está fuertemente asociada con la edad, el estado de salud general y la coexistencia de factores de riesgo como diabetes, obesidad y hábitos poco saludables. La distribución y correlación de variables justifican el enfoque multivariado para el modelado predictivo y subrayan la importancia de campañas de prevención focalizadas en grupos vulnerables.

5. Visualizaciones y tablas de apoyo

Incorpora aquí las gráficas y tablas generadas en Python que consideres más ilustrativas para respaldar los hallazgos (por ejemplo: histogramas de edad y BMI, boxplots de BMI vs. HeartDisease, heatmap de correlaciones, gráficos de barras de prevalencia por sexo y edad, etc.).







