Actividad física y salud cardiovascular: analizar cómo diferentes niveles de ejercicio influyen en indicadores de riesgo

ÍNDICE

1. Introducción y contexto..... 3
2. Descripción del dataset.....3
3. Hipótesis iniciales.....5
4. Decisiones técnicas.....5
5. Limpieza y transformaciones.....6
6. Análisis exploratorio.....13
6.1 Análisis univariante..... 13
6.2 Análisis bivariante..... 14
6.3 Análisis multivariante.....15
7. Visualizaciones.....17
8. Conclusiones.....17
9. Recomendaciones..... 17
ANEXO..... 18

1. Introducción y contexto

El objetivo de este trabajo es realizar un Análisis Exploratorio de Datos (EDA) sobre el conjunto de datos del dataset “Heart Disease 2022”, con el fin de comprender su estructura, detectar patrones relevantes, errores y así poder llegar a sacar unas conclusiones que respondan a las hipótesis que nos planteamos.

Este dataset está relacionado con la salud física y mental, puesto que queremos desarrollar el siguiente tema: “Actividad física y salud cardiovascular: analizar cómo diferentes niveles de ejercicio influyen en indicadores de riesgo (colesterol, IMC, frecuencia cardíaca en reposo, hipertensión)”.

Más concretamente, este dataset recoge información procedente de encuestas de salud realizadas en Estados Unidos a una serie de personas, donde contiene datos como variables demográficas, hábitos de vida, indicadores de salud general y antecedentes médicos. El análisis de este tipo de datos es especialmente relevante en el ámbito de la salud, ya que permite identificar factores de riesgo asociados a enfermedades cardiovasculares y apoyar la toma de decisiones preventivas.

El EDA permite detectar valores faltantes, incoherencias, variables poco informativas y posibles relaciones entre variables que pueden guiar a análisis posteriores.

2. Descripción del dataset

El conjunto de datos analizado contiene **445.132 entradas** y **40 variables** (columnas), lo que lo convierte en un dataset de gran tamaño.

Las variables incluidas pueden agruparse en las siguientes categorías:

Variables demográficas

- **State**: estado de residencia de la persona encuestada
- **Sex**: género de la persona
- **Age Category**: rango de edad de la persona
- **Race Ethnicity Category:** categoriza a la persona según su raza o grupo étnico

Variables de salud general

- **General Health**: Percepción del estado general de salud de la persona
- **Physical Health Days**: Número de días con mala salud física en el último mes
- **Mental Health Days**: Número de días con mala salud mental en el último mes
- **Sleep Hours**: Promedio de horas de sueño

Variables de hábitos de vida

- **Physical Activities**: Práctica de actividad física
- **Smoker Status**: Hábito de fumar
- **ECigarette Usage**: Uso de cigarrillos electrónicos
- **Alcohol Drinkers**: Consumo de alcohol

Variables médicas y antecedentes

- **Had Heart Attack**: indica si la persona ha sufrido un ataque al corazón
- **Had Angina**: si la persona ha sufrido una angina de pecho
- **Had Stroke**: si la persona ha sufrido un accidente cerebrovascular
- **Had Asthma**: si la persona ha sido diagnosticada de asma
- **Had Diabetes**: si la persona tiene diabetes
- **Had Kidney Disease**: si la persona tiene alguna enfermedad renal
- **Had Arthritis**: indica si la persona ha sido diagnosticada de artritis
- **Had Depressive Disorder**: Indica si la persona ha tenido algún trastorno depresivo
- **Had COPD**: indica si la persona padece EPOC (enfermedad pulmonar obstructiva crónica)
- **Had Skin Cancer**: señala si la persona ha tenido cáncer de piel
- **Removed Teeth**: indica si la persona ha perdido piezas dentales

Limitaciones físicas y cognitivas

- **Deaf Or Hard Of Hearing**: indica si la persona es sorda o tiene dificultades auditivas
- **Blind Or Vision Difficulty**: señala si la persona presenta ceguera o dificultades de visión
- **Difficulty Concentrating**: indica si la persona tiene dificultades para concentrarse
- **Difficulty Walking**: refleja dificultades para caminar
- **Difficulty Dressing Bathing**: indica si la persona tiene dificultades para vestirse o bañarse sin ayuda
- **Difficulty Errands**: refleja si la persona tiene dificultades para realizar tareas o recados, como por ejemplo hacer la compra

Medidas corporales

- **BMI**: Índice de Masa Corporal
- **Height in Meters**: Altura de la persona en metros
- **Weight in Kilograms**: Peso de la persona en kilogramos

Variables de prevención diagnóstico y seguimiento sanitario

- **Flu Vax Last 12**: indica si la persona ha recibido la vacuna contra la gripe en los últimos 12 meses
- **Pneumo Vax Ever**: señala si la persona ha recibido alguna vez la vacuna contra el neumococo
- **Tetanus Last 10 Tdap**: indica si la persona ha recibido una vacuna contra el tétanos en los últimos 10 años
- **Covid Pos**: indica si la persona ha dado positivo en COVID-19 en algún momento
- **Chest Scan**: indica si la persona se ha realizado alguna vez una TAC del tórax
- **HIV Testing**: señala si la persona se ha realizado alguna prueba del VIH
- **Last Checkup Time**: refleja el tiempo que ha pasado desde el último chequeo general
- **High Risk Last Year**: indica si la persona ha tenido riesgos altos de salud durante el último año

Cabe destacar que el dataset contiene valores nulos en varias variables, lo cual se tendrá en cuenta durante el análisis exploratorio y la toma de decisiones.

3. Hipótesis iniciales

Antes de comenzar el análisis exploratorio, se plantean las siguientes hipótesis:

- **Hipótesis 1:** Las personas que tienen peor estado de salud general (General Health) presentan una mayor prevalencia de enfermedades cardiovasculares como ataques al corazón o angina de pecho.

- **Hipótesis 2:** Existen diferencias significativas en la incidencia de problemas cardíacos en función del sexo y la edad.

- **Hipótesis 3:** Los hábitos de vida poco saludables (sedentarismo, tabaquismo, consumo de alcohol, bajo número de horas de sueño) están asociados a un mayor riesgo de enfermedades del corazón.

- **Hipótesis 4:** Variables como el IMC y la presencia de diabetes muestran relación con los antecedentes cardiovasculares.

Estas hipótesis servirán como guía durante el EDA y podrán ser confirmadas o descartadas a partir de los datos.

4. Decisiones técnicas

Tratamiento de valores nulos
El dataset contiene valores faltantes en diversas variables, como por ejemplo Last Check Up Time, Removed Teeth, Height Meters o BMI.

Se ha decidido no eliminar de forma inmediata las filas con valores nulos, ya que esto supondría una pérdida considerable de información dada la magnitud del dataset. En su lugar, se evaluará el porcentaje de valores faltantes por variable para decidir si:

- Se imputan valores (mediana, por ejemplo)
- Se descarta la variable si el número de valores nulos (NaN) es alto

Análisis de variables numéricas
Para las variables numéricas (como BMI, Sleep Hours o Physical Health Days), se prevé utilizar estadísticas descriptivas y representaciones gráficas como histogramas y boxplots.
Estas herramientas permiten identificar distribuciones asimétricas, valores extremos y posibles errores que puedan haber en los datos.

Análisis de variables categóricas
Para las variables categóricas, se prevé analizar mediante tablas de frecuencia y gráficos de barras. Esta elección facilita la comparación entre categorías y ayuda a detectar posibles desbalances, por ejemplo, entre personas con y sin antecedentes médicos.

Análisis de relaciones entre variables
Para explorar la relación entre las variables numéricas y la presencia de enfermedades cardiovasculares, se prevé analizar comparaciones entre grupos (por ejemplo, entre personas que hayan sufrido o no un ataque al corazón).

Este enfoque permite evaluar si ciertas características están asociadas a una mayor prevalencia de problemas cardíacos.

5. Limpieza y transformaciones

En esta sección se describen las decisiones de limpieza y transformación aplicadas a las variables del dataset, con el objetivo de mejorar la calidad de los datos y facilitar su análisis posterior.

State

- **Valores:**
  Nombres de estados y territorios

- **Tipo de variable:**
  Categórica nominal con múltiples categorías

- **Limpieza:**
  No es necesario realizar ninguna limpieza ya que no encontramos valores duplicados ni inconsistencias

- **Transformación:**
  Conversión al tipo category

Sex

- **Valores**:
  Female y Male

- **Tipo de variable**:
  Es una variable categórica nominal (no numérica y sin orden inherente)

- **Limpieza:**
  No es necesario limpiar nada puesto que no se detectaron errores de escritura ni valores inconsistentes

- **Transformación:**
  Conversión al tipo category en pandas. Esta transformación no modifica los valores, pero mejora la eficiencia de memoria y facilita el análisis de variables categóricas

General Health

- **Valores**:
  Poor, Fair, Good, Very good, Excellent

- **Tipo de variable:**
  Es una variable categórica ordinal (existe un orden lógico entre las categorías)

- **Limpieza**:
  No es necesario limpiar nada ya que todas las categorías son válidas

- **Transformación**:
  Consideramos transformar los valores de esta variable por valores numéricos:
  Poor = 1
  Fair = 2
  Good = 3
  Very good = 4
  Excellent = 5

  Esta transformación permite calcular promedios, correlaciones y otros análisis cuantitativos

Physical Health Days / Mental Health Days

- **Valores**:
  Número de días entre 0 y 30

- **Tipo de variable**:
  Es una variable numérica discreta

- **Limpieza**:
  No es necesario realizar ninguna limpieza. El valor 30 representa “todos los días” y no se considera un error

- **Transformación**:
  No es necesario realizar ninguna transformación

Last Checkup Time

- **Valores**:
  Intervalos de tiempo desde el último chequeo médico: <1 year, 1-2 years, >5 years, Never

- **Tipo de variable**:
  Es una variable categórica ordinal

- **Limpieza**:
  No es necesario realizar ninguna limpieza

- **Transformación**:
  Consideramos transformar los valores de esta variable por valores numéricos:

  Within past year = 1
  Within past 2 years = 2
  Within past 5 years = 3
  5 or more years ago = 4

  Esta codificación facilita el análisis de prevención y su relación con otras variables de salud

Variables binarias de salud y hábitos como:

Physical Activities, Had Heart Attack, Had Stroke, Had Asthma, Chest Scan, Alcohol Drinkers, HIV Testing, Flu Vax Last 12, entre otras.

- **Valores**:
  Yes, No

- **Tipo de variable**:
  Es una variable booleana

- **Limpieza**:
  No es necesario realizar ninguna limpieza

- **Transformación**:
  Consideramos transformar los valores de esta variable por valores numéricos:
  Yes = 1
  No = 0

  Esta transformación mejora la eficiencia computacional y permite su uso en modelos estadísticos

Sleep Hours

- **Valores**:
  Horas de sueño (1-24h)

- **Tipo de variable**:
  Es una variable numérica continua

- **Limpieza**:
  No es necesario realizar ninguna limpieza. Aunque algunos valores son poco frecuentes, se consideran posibles y eliminarlos podría introducir sesgos

- **Transformación**:
  No es necesario realizar ninguna transformación

Removed Teeth

- **Valores**:
  None, 1 to 5, 6 or more (but not all), All

- **Tipo de variable**:
  Es una variable categórica ordinal

- **Limpieza**:
  No es necesario realizar ninguna limpieza

- **Transformación**:
  Consideramos transformar los valores de esta variable por valores numéricos:
  None = 0
  1 to 5 = 1
  6 or more = 2
  All = 3

Had Diabetes

- **Valores**:
  No, Yes, Pre-diabetes y Diabetes during Pregnancy

- **Tipo de variable**:
  Es una variable categórica nominal

- **Limpieza**:
  No es necesario realizar ninguna limpieza ya que todas las categorías aportan información relevante

- **Transformación**:
  No es necesario realizar ninguna transformación. De forma adicional, se ha creado una variable binaria que indica si la persona tiene o ha tenido diabetes (sí/no)

Smoker Status / E Cigarette Usage

- **Valores**:
  Never
  Former
  Current (some days)
  Current (every day)

- **Tipo de variable**:
  Es una variable categórica ordinal

- **Limpieza**:
  No es necesario realizar ninguna limpieza

- **Transformación**:
  Consideramos transformar los valores de esta variable por valores numéricos:
  Never = 0
  Former = 1
  Some days = 2
  Every day = 3

Race Ethnicity Category

- **Valores**:
  Cinco categorías

- **Tipo de variable**:
  Es una variable categórica nominal

- **Limpieza**:
  No es necesario realizar ninguna limpieza

- **Transformación**:
  No es necesario realizar ninguna transformación

Age Category

- **Valores**:
  Rangos de edad

- **Tipo de variable**:
  Es una variable categórica ordinal

- **Limpieza**:
  No es necesario realizar ninguna limpieza

- **Transformación**:
  No es necesario realizar ninguna transformación

Height In Meters

- **Valores**:
  Altura en metros

- **Tipo de variable**:
  Es una variable numérica continua

- **Limpieza**:
  Se consideraron valores fuera del rango típico adulto (1.3 m - 2.2 m) como no

plausibles y se marcaron como valores perdidos (NaN)

- **Transformación**:
  No es necesario realizar ninguna transformación

Weight In Kilograms

- **Valores**:
  Peso en kilogramos

- **Tipo de variable**:
  Es una variable numérica continua

- **Limpieza**:
  Se marcaron como valores perdidos los pesos fuera del rango 30-250 kg

- **Transformación**:
  No es necesario realizar ninguna transformación adicional

BMI

- **Valores**:
  Índice de Masa Corporal

- **Tipo de variable**:
  Es una variable numérica continua

- **Limpieza**:
  No es necesario realizar ninguna limpieza

- **Transformación**:
  Se ha mantenido la variable original y se ha creado una columna adicional con categorías:

  Underweight (< 18.5)
  Normal (18.5–24.9)
  Overweight (25–29.9)
  Obese (≥ 30)

Tetanus Last 10 Tdap

- **Valores**:
  Cuatro categorías

- **Tipo de variable**:
  Es una variable categórica nominal

- **Limpieza**:
  No es necesario realizar ninguna limpieza

- **Transformación**:
  No es necesario realizar ninguna transformación

Covid Pos

- **Valores**:
  No, Yes, Positive (home test)

- **Tipo de variable**:
  Es una variable categórica nominal

- **Limpieza**:
  Se unificaron todas las categorías positivas en un único valor “positivo”

- **Transformación**:
  No se ha transformado la variable original, sino que se ha creado una versión binaria adicional

Resultado del tratamiento de valores NaN

1. **Variables categóricas**: Sex, State, Race Ethnicity Category, Smoker Status, etc.

   Se han reemplazado los valores faltantes por la moda de la columna. Esto permite mantener la distribución general sin introducir valores arbitrarios

2. **Variables binarias**: Physical Activities, Had Heart Attack, Had Asthma, etc.

   Los valores faltantes se han reemplazado por 0 (“No”). Se considera razonable asumir la ausencia de la condición cuando no hay reporte

3. **Variables numéricas**: Sleep Hours, Weight In Kilograms, BMI, Physical Health Days, etc.

   Los valores faltantes se han imputado con la mediana de la columna. Se ha elegido la mediana porque es robusta frente a valores extremos.

6. Análisis exploratorio

En esta sección se realiza un análisis exploratorio de los datos con el objetivo de comprender la distribución de las variables, identificar patrones relevantes y detectar posibles relaciones entre ellas. Para ello, el estudio se estructura en tres niveles de análisis.

6.1 Análisis univariante

Este primer análisis se centra en el estudio individual de cada variable para conocer su distribución, valores más frecuentes y posibles anomalías. Por ello, en primer lugar realizamos una revisión general de la estructura del dataset, comprobando el número de variables que tiene.

Seguidamente, el análisis lo dividimos según la naturaleza de las variables, siendo estas categóricas y numéricas.

Para las **variables categóricas**, se calculan las tablas de frecuencias y porcentajes con el fin de conocer la distribución relativa de cada categoría. Este enfoque permite identificar desbalances, así como obtener una visión general de las características más comunes de la población analizada. En este caso, se observa una ligera mayor representación de mujeres frente a hombres. En cuanto a la edad, los grupos de mayor edad son los que presentan una mayor frecuencia, destacando especialmente el rango de 65-69 años. Respecto al estado de salud cardiovascular, la mayoría de individuos no ha sufrido un ataque al corazón, lo que indica que esta condición es minoritaria dentro del conjunto de datos.

El análisis del IMC muestra que una parte considerable de la población se sitúa en rangos de sobrepeso, lo que resulta relevante dado su posible vínculo con enfermedades cardiovasculares y otros problemas de salud.

En el caso de las **variables numéricas**, se analizan medidas descriptivas y se emplean representaciones gráficas como diagramas de caja (boxplots). Este tipo de gráfico permite evaluar la dispersión de los datos, la presencia de asimetrías y la existencia de outliers que podrían influir en análisis posteriores.

En este caso, se observa que muchas de estas variables presentan distribuciones asimétricas y discretas, lo que justifica el uso de medidas importantes como la mediana para su análisis. La variable de salud general (General Health) muestra una concentración elevada en categorías intermedias, mientras que los valores extremos son menos frecuentes.

Respecto a los hábitos de vida, el uso de cigarrillos electrónicos es claramente minoritario, con una distribución fuertemente desbalanceada hacia el no consumo. Por el contrario, el consumo de alcohol presenta una distribución más equilibrada entre consumidores y no consumidores.

Por último, en cuanto a la diabetes, la mayoría de la población no presenta este diagnóstico, siendo los casos positivos una minoría claramente diferenciada.
En conjunto, este análisis univariante permite identificar desbalances en varias variables, así como características predominantes de la población, sentando las bases para el posterior análisis bivariante, en el que se explorarán posibles relaciones entre estas variables y la presencia de enfermedades cardiovasculares.

6.2 Análisis bivariante

El siguiente análisis es el bivariante, cuyo objetivo es explorar las relaciones entre dos variables y evaluar posibles asociaciones de interés, concretamente HadHeartAttack y HadAngina. Para ello, se combinan herramientas de visualización gráfica con pruebas estadísticas, lo que permite obtener una visión más completa y robusta de las asociaciones observadas. A continuación, se presentan los resultados organizados en función de las hipótesis planteadas.

Hipótesis 1: Relación entre salud general y enfermedades cardiovasculares

A partir de las representaciones gráficas, se observa que el porcentaje de individuos que han sufrido un ataque al corazón o angina aumenta progresivamente a medida que empeora la percepción de la salud general (GeneralHealth). Esta tendencia es consistente para ambas variables cardiovasculares, mostrando una relación clara y positiva.

Conclusión:
La hipótesis 1 queda confirmada visualmente. Las personas con peor salud general presentan un mayor riesgo de padecer enfermedades cardiovasculares.

Hipótesis 2: Influencia del sexo y la edad en el riesgo cardiovascular

En cuanto al sexo, las comparaciones entre hombres y mujeres muestran diferencias leves en la prevalencia de ataque al corazón y angina. Sin embargo, estas diferencias no resultan estadísticamente significativas, por lo que no se puede afirmar una asociación clara entre el sexo y estas patologías.

Por el contrario, la edad muestra un patrón muy definido. El porcentaje de casos positivos, tanto de ataque al corazón como de angina, aumenta de forma clara conforme se incrementa la edad, lo que indica una relación positiva entre envejecimiento y riesgo cardiovascular.

Conclusión:
Las personas de mayor edad presentan un mayor riesgo de enfermedades cardiovasculares. El sexo, aunque muestra ligeras diferencias, no parece ser un factor determinante en este conjunto de datos.

Hipótesis 3: Hábitos de vida y su relación con las enfermedades cardiovasculares

El análisis de la actividad física revela que los individuos que no realizan PhysicalActivities presentan un mayor porcentaje de casos positivos, tanto de ataque al corazón como de angina, lo que sugiere un efecto protector del ejercicio regular.

Respecto al consumo de tabaco (SmokerStatus), se observa que el porcentaje de casos positivos aumenta con el consumo, destacando que los exfumadores presentan un riesgo similar o incluso superior al de quienes fuman algunos días, lo que podría estar relacionado con la duración o intensidad del consumo previo.

En el caso del uso de cigarrillos electrónicos (ECigaretteUsage), no se identifica una relación positiva con las enfermedades cardiovasculares. De hecho, el porcentaje de casos positivos disminuye entre los usuarios frecuentes, lo que sugiere la ausencia de una asociación clara en este conjunto de datos.

Por último, el análisis de las horas de sueño (SleepHours) muestra resultados mixtos. Visualmente, los diagramas de caja presentan medianas y rangos similares entre los grupos con y sin enfermedad cardiovascular, así como la presencia de outliers en ambos casos. No obstante, las pruebas estadísticas aportan matices importantes:

- Para HadHeartAttack, el test de Mann-Whitney U no muestra diferencias estadísticamente significativas (p-value > 0.05).

- Para HadAngina, el mismo test sí detecta una diferencia significativa (p-value < 0.05), a pesar de que esta no sea evidente a simple vista en los gráficos.

Conclusión:
La actividad física y el consumo de tabaco muestran una relación positiva con el riesgo de enfermedades cardiovasculares. El uso de cigarrillos electrónicos no parece incrementar dicho riesgo. En cuanto al sueño, los resultados ponen de manifiesto la importancia de complementar el análisis visual con pruebas estadísticas, ya que se detecta una relación significativa con la angina, pero no con el ataque al corazón.

Hipótesis 4: IMC, diabetes y riesgo cardiovascular

El análisis del IMC categorizado muestra que el porcentaje de casos positivos aumenta conforme se incrementa el peso corporal, tanto para el ataque al corazón como para la angina. De forma interesante, también se observa un aumento del riesgo en personas con bajo peso, lo que sugiere la influencia de otros factores de riesgo subyacentes.

Asimismo, la presencia de diabetes (Diabetes_binary) se asocia claramente con un mayor porcentaje de enfermedades cardiovasculares, evidenciando una relación positiva para ambas variables analizadas.

Conclusión:
La hipótesis 4 queda confirmada visualmente. Tanto un IMC elevado como la presencia de diabetes se asocian con un mayor riesgo de padecer enfermedades cardiovasculares.

En conjunto, el análisis bivariante permite identificar relaciones relevantes entre factores sociodemográficos, hábitos de vida y condiciones de salud con la presencia de enfermedades cardiovasculares, reforzando los hallazgos del análisis univariante y proporcionando una base sólida para análisis multivariantes posteriores.

6.3 Análisis multivariante

Por último, se realiza el análisis multivariante, el cual permite analizar de forma conjunta varias variables y obtener una visión más global del comportamiento de los datos.

Hipótesis 1: Relación entre salud general y enfermedades cardiovasculares

A través de los gráficos establecidos en el análisis multivariante, es preciso observar como la salud general de los individuos es clave a la hora de tener en cuenta una posibilidad mayor o menor de este tipo de enfermedades, dejando en claro también que el tener un horario considerado “sano” por la medicina es una variable destacable a la par, puesto que esta clase de enfermedades se ven agravadas por la falta de sueño o la recurrencia de este.

Conclusión:

La hipótesis 1 queda de vuelta confirmada, ya que incluso con variables extras que podemos añadir al gráfico, como lo son las horas de sueño, sigue siendo un problema evidente que recae sobre aquellas personas cuya salud general es más deficiente.

Hipótesis 2: Influencia del sexo y la edad en el riesgo cardiovascular

En este caso, se puede hacer un gráfico simple que represente esta hipótesis. En todos los rangos de edad podemos vislumbrar con evidencias como los hombres, en cualquier caso, tienen más probabilidades, más que las mujeres, de mostrar este tipo de tendencias referidas a un riesgo cardiovascular. Además también es evidente como la tendencia a este tipo de riesgos se ve agravada por la edad, a medida que esta avanza.

Conclusión:

La hipótesis 2 queda confirmada, puesto que con un solo gráfico vemos claramente esa diferenciación que planteábamos, dependiendo del sexo y la edad de los individuos. Con esto, hemos podido observar que las enfermedades cardiovasculares afectan más a un hombre adulto que a una mujer joven.

Hipótesis 3: Hábitos de vida y su relación con las enfermedades cardiovasculares

Con respecto al análisis multivariante, podemos observar el gráfico sobre el impacto del tabaquismo y el cigarrillo electrónico en este tipo de enfermedades. Visualizamos como el uso de ambos con frecuencia es claramente un problema muy grande en referencia a todo este tipo de problemas médicos, a la vez que el consumo nulo de ambos de ellos es lo mejor para la salud.

Conclusión:

La hipótesis 3 queda confirmada también, puesto que hemos visto que este tipo de malos hábitos para la salud, agravan la probabilidad de presentar este tipo de enfermedades, peligrosas para la salud de las personas y que puede poner vidas en riesgo.

Hipótesis 4: IMC, diabetes y riesgo cardiovascular

En este caso, esta hipótesis se ve confirmada por gráficos previos y reforzada con el *heatmap*. Es claramente visible como el IMC, y sobre todo la diabetes, tienen una relación clara con las enfermedades cardiovasculares, por lo que, la gente con un IMC insano, o bien personas que padezcan de diabetes, son claramente más propensas a sufrir de este tipo de riesgos médicos.

Conclusión:

Esta hipótesis se ha vuelto a confirmar con el análisis multivariante, después de haberlo visto con el análisis bivariante. Por ello, es importante intentar llevar una mayor vida sana, más si la persona padece de estas patologías comentadas anteriormente.

7. Visualizaciones

Los gráficos utilizados para este análisis se encuentran en el archivo main en el repositorio EDA_Actvidades_Enfermedades. No obstante, hemos incluido los gráficos más relevantes en el anexo de esta memoria. Para visualizarlos, ir a dicho apartado.

8. Conclusiones

Con respecto a todo lo que vimos previamente, entendemos los siguientes puntos:

- La salud general de las personas afecta a la probabilidad de padecer una enfermedad cardiovascular; esto sumado a que otro tipo de variables a lo largo del dataset nos hacen reforzar esta misma idea, da lugar a pensar que esto es una realidad y debemos tener cuidado con nuestra salud general.

- Habíamos dejado claro en el análisis bivariante que la edad tenía una clara relación con la proporción de enfermedades cardiovasculares, y tras el análisis multivariante, podemos sumar a esto que los hombres, son a su vez, más propensos a padecer esta clase de patologías en cualquier rango de edad.

- Hemos observado como, hábitos poco saludables como el tabaquismo, un mal horario de sueño, o no hacer ejercicio, dan resultado a una mayor probabilidad de enfermedades cardiovasculares. Lo que sí sorprende es que las personas que utilizan un cigarrillo electrónico no se ven significativamente tan afectadas como podríamos esperar.

- Asimismo, visualizamos como aquellas variables como el IMC y padecer de diabetes son también muy relevantes, en especial esta última, puesto que aquellas personas que se ven afectadas por la dicha, ven abierta una mayor probabilidad a verse expuestos a este tipo de riesgos.

9. Recomendaciones

Finalmente, nuestras recomendaciones con respecto a la información adquirida después de nuestros análisis es que se mantenga una salud general que sea la mejor posible, a pesar de que la mayoría de la población se conforme con la salud habitual. También, es importante que se tenga en cuenta ciertos malos hábitos como el tabaquismo, que se pueda ostentar un horario de sueño correcto, y tener un IMC y una actividad física adecuada y saludable para correr el menor riesgo posible de padecer este tipo de patologías.

ANEXO

Visualizaciones

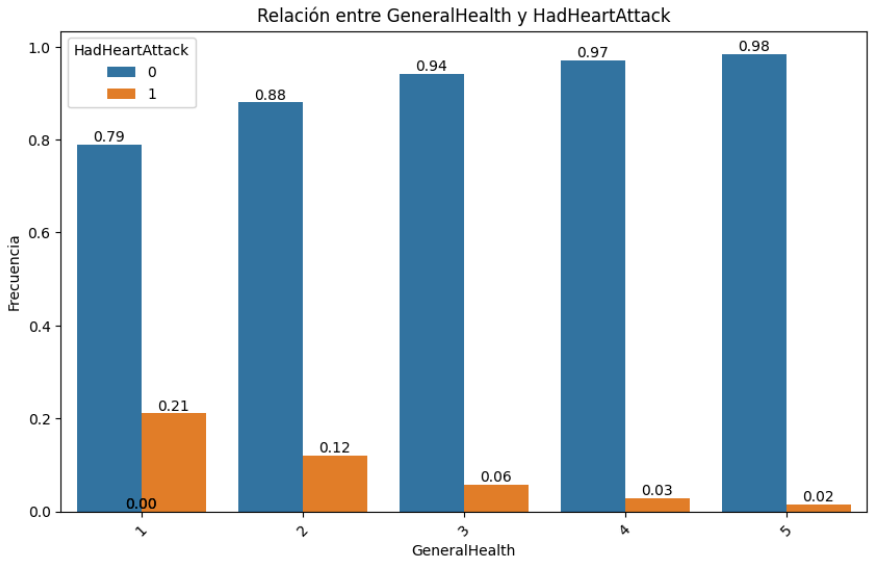

Gráfico 1. Distribución de GeneralHealth y HadHeartAttack. El porcentaje de Yes en HadHeartAttack aumenta conforme GeneralHealth empeora.

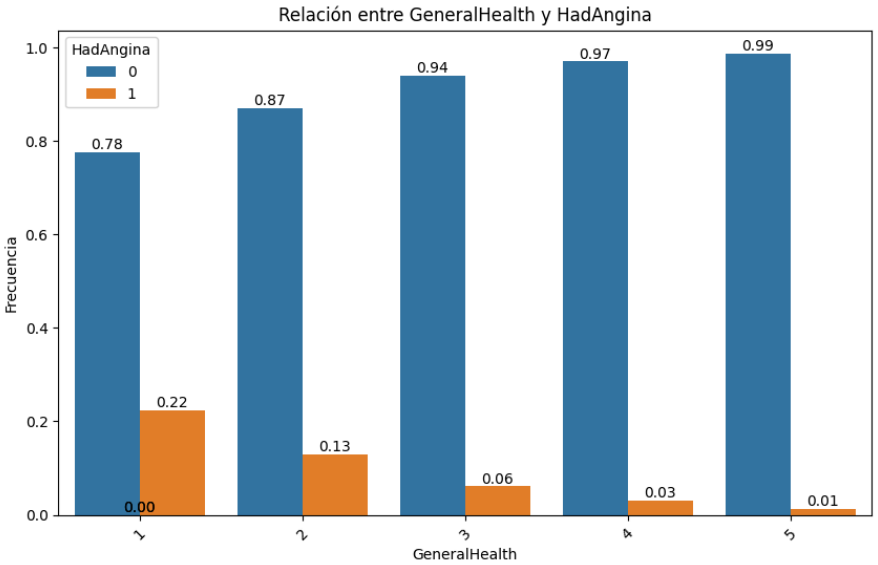

Gráfico 2. Distribución de GeneralHealth y HadAngina. El porcentaje de Yes en HadAngina aumenta conforme GeneralHealth empeora.

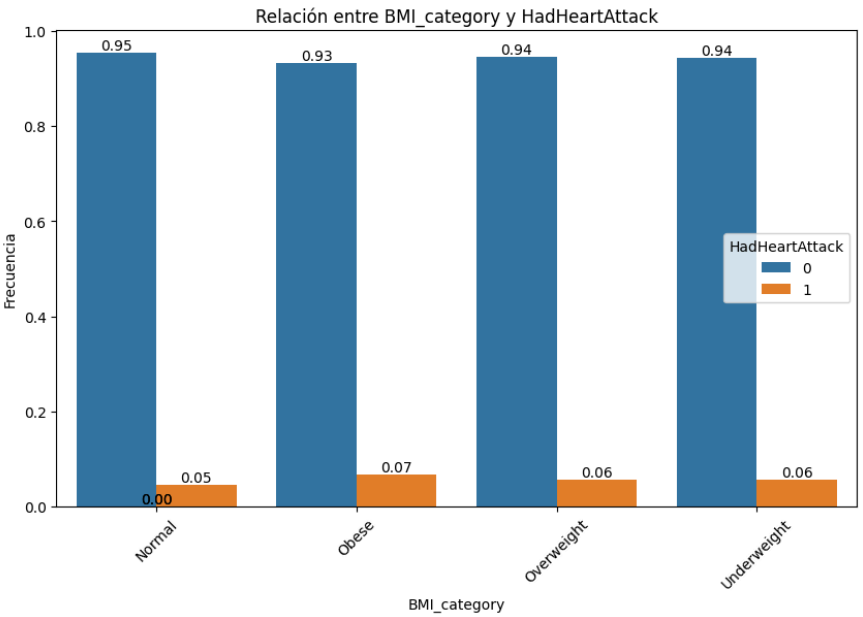

Gráfico 3. Distribución de BMI_Category y HadHeartAttack. El porcentaje de Yes aumenta con el peso para ambas variables. Incluso, aumenta en personas que están por debajo de su peso, indicando otros factores de riesgo en ambas variables.

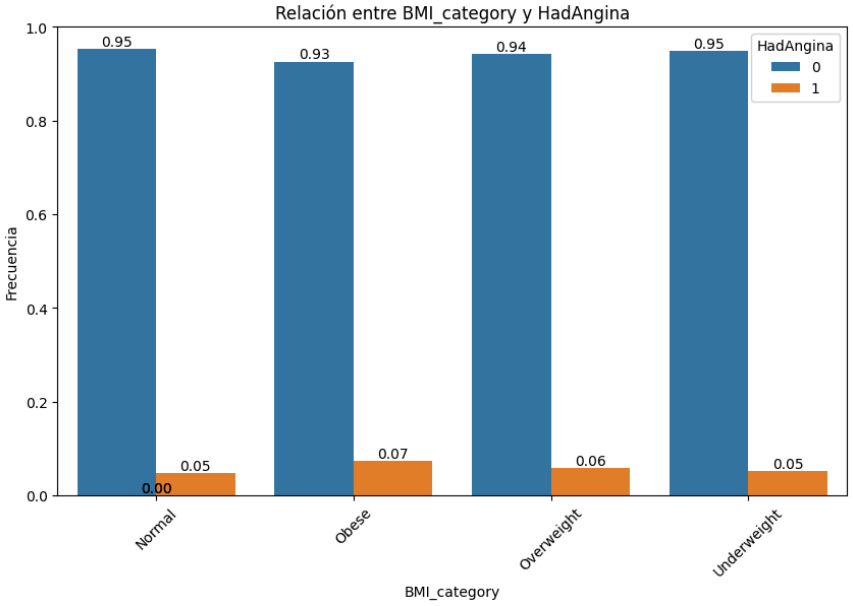

Gráfico 4. Relación de BMI_Category y HadAngina. El porcentaje de Yes aumenta con el peso para ambas variables. Incluso, aumenta en personas que están por debajo de su peso, indicando otros factores de riesgo en ambas variables.

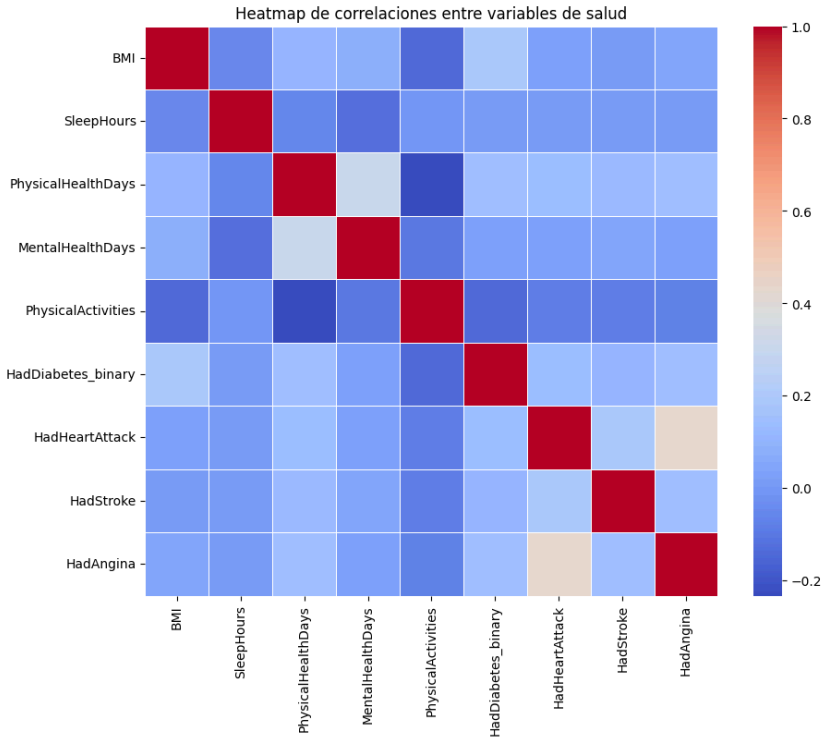

Gráfico 5. Heatmap de correlaciones entre variables de salud. Este gráfico nos permite relacionar las variables más importantes, observando las que se ven más relacionadas claramente entre sí.

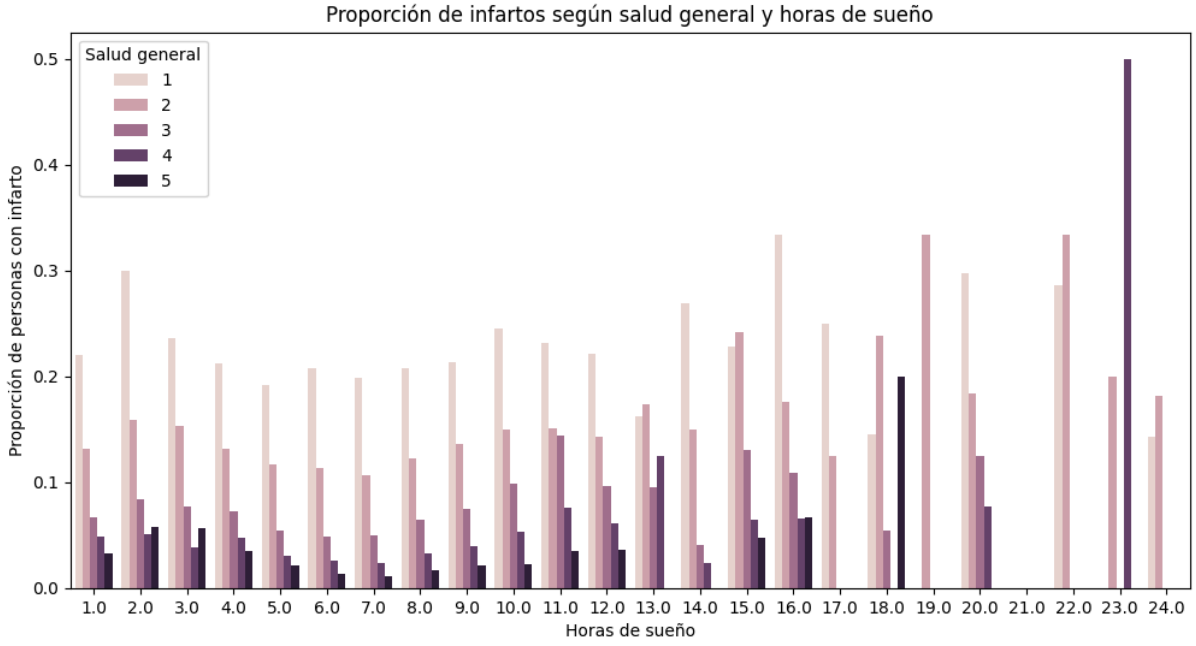

Gráfico 6. Gráfico de barras de la proporción de infartos según la salud general y las horas de sueño. El gráfico refuerza que una buena salud general y un patrón de sueño moderado se asocia con menor proporción de infartos, mientras que la mala salud y los extremos de sueño aumentan el riesgo.

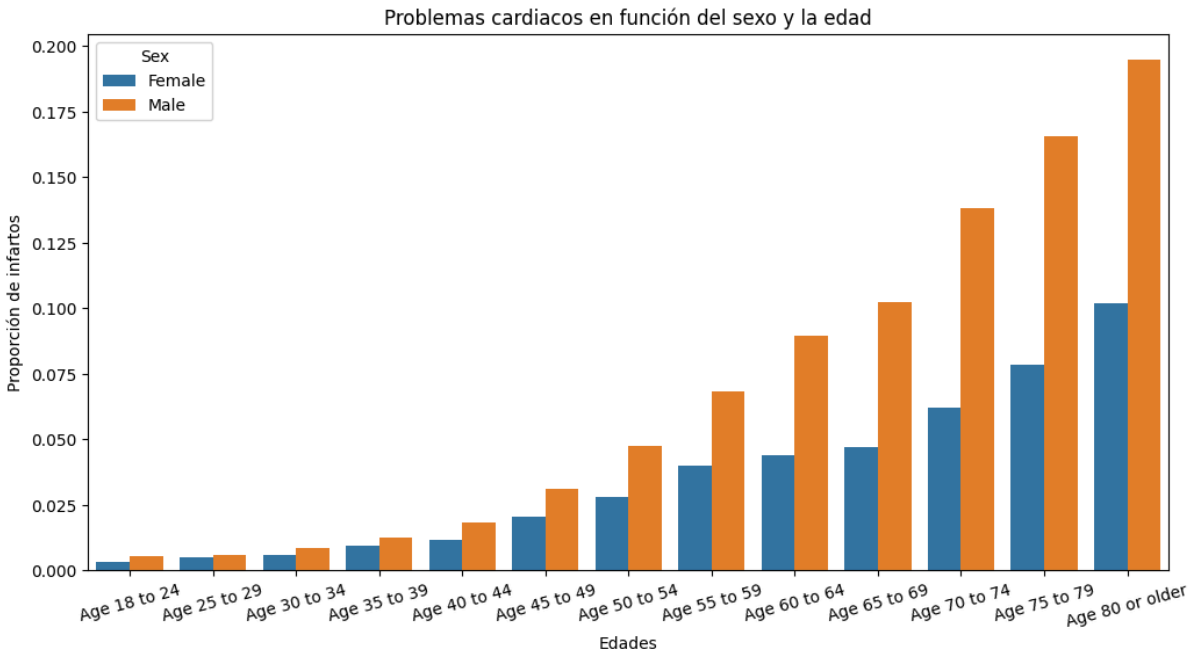

Gráfico 7. Gráfico de barras de los problemas cardíacos en función del sexo y la edad. El gráfico indica que la edad es un factor de riesgo clave para los infartos, y que el sexo masculino presenta un riesgo sistemáticamente mayor, especialmente en edades avanzadas.