# ANALISIS DE FACTORES QUE INFLUYEN EN ENFERMEDADES DEL CORAZON

PROYECTO DATA SCIENCE – TRABAJO FINAL.

ERNESTO LEIMSIEDER

DOCENTE: IGNACIO RUSSO

# HOJA DE RUTA

## Abstract

El presente documento trata sobre el estudio de diferentes variables que inciden en el rubro salud. Aquí se encuentran las distintas etapas que deben cumplirse para lograr llegar a dar respuestas a las interrogantes que se hacen al inicio.
Con este fin, se tomó un Dataset para conocer la situación actual de un grupo heterogéneo de personas y con esto luego quienes hagan uso del presente trabajo, poder llegar a saber qué medidas se deberían llegar a implementar por parte de diferentes instituciones relacionadas con la salud ya sean públicas o privadas.
Esta información permitirá a los profesionales de la salud, tomar medidas orientadas a evitar la aparición de enfermedades o problemas de salud mediante el control de los factores causales o detener su avance y atenuar sus consecuencias una vez establecida.

## Definición de objetivo

Con el Dataset elegido la intención es poder obtener información que permita tener datos ciertos sobre distintos grupos de personas y como influyen los hábitos de alimentación, actividad física, etc. en distintos tipos de enfermedades, en base a esto poder tomar medidas que sirvan como forma de tener una población más saludable.

## Preguntas/hipótesis

Preguntas de interés que se pueden realizar a partir el Dataset seleccionado:

1. Qué relación existe entre el BMI (Indice de masa corporal) según el género
de las personas?
2. Se puede diferenciar la cantidad de horas de sueño según el género de las personas?

3. Qué relación existe entre la salud física y la salud mental de las personas?

4. Qué relación existe entre **la edad** de las personas y los distintos tipos de enfermedades que puede tener una persona (diabetes, problemas cardíacos, hipertensión, etc.)?

5. Qué relación existe entre **el género** de las personas y los distintos tipos de enfermedades que puede tener una persona (diabetes, problemas cardíacos, hipertensión, etc.)?

6. Qué relación existe entre la raza de las personas y la posibilidad de que tengan cáncer en la piel?

## Contexto comercial

Se ha proporcionado el archivo "heart_2020_cleaned.csv" que contiene información sobre distintos tipos de problemas relacionados con la salud de las personas (tabaquismo, ingesta de alcohol), enfermedades como diabetes, cáncer, etc. También se brinda información como raza, edad, actividad deportiva, tiempo de descanso, etc.
Se deberá extraer los datos y prepararlos para su visualización y poder llegar a conclusiones según lo obtenido.

## Exploratory Data Analysis (EDA)

1- El índice de BMI (Indice de masa corporal) en mujeres es levemente superior que en varones.

2 - La cantidad de horas de sueño es prácticamente igual tanto en mujeres como en varones.

3 - Existe una relación directa entre la salud física y mental de las personas.

4 - Un poco más de 2/3 de las personas realizan algún tipo de actividad física, además se mantiene la relación por género.

5 - El índice de BMI es mayor entre personas de 40 y 60 años.

6 - Se puede concluir que la gran mayoría de las personas no son diabéticas, una cierta cantidad lo es y los otros son insignificantes.

## Storytelling

Mediante el Dataset elegido para este proyecto en donde se obtienen datos de un grupo heterogéneo de personas (características tales como edad, raza, tiempo dedicado al descanso, etc.), distintos tipos de enfermedades (diabetes, asma, problemas cardíacos), hábitos como el consumo de tabaco y alcohol lo que se busca es llegar a conclusiones que permitan saber que grupos de personas están predispuestas a tener complicaciones en su salud, de esta forma se podrían prevenir ciertos riesgos.
Además, permitiría tomar medidas en cuanto al cuidado de la salud de las personas, lo cual redundaría en una mejor atención en los centros hospitalarios ya que se podrían gestionar de una mejor forma los recursos médicos (humanos y materiales).

---

Dataset "heart_2020_cleaned.csv"

| | HeartDisease | BMI | Smoking | AlcoholDrinking | Stroke | PhysicalHealth | MentalHealth | DiffWalking | Sex | AgeCategory | Race | Diabetic | PhysicalActivity | GenHealth | SleepTime | Asthma | KidneyDisease | SkinCancer |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | No | 16.60 | Yes | No | No | 3.0 | 30.0 | No | Female | 55-59 | White | Yes | Yes | Very good | 5.0 | Yes | No | Yes |
| 1 | No | 20.34 | No | No | Yes | 0.0 | 0.0 | No | Female | 80 or older | White | No | Yes | Very good | 7.0 | No | No | No |
| 2 | No | 26.58 | Yes | No | No | 20.0 | 30.0 | No | Male | 65-69 | White | Yes | Yes | Fair | 8.0 | Yes | No | No |
| 3 | No | 24.21 | No | No | No | 0.0 | 0.0 | No | Female | 75-79 | White | No | No | Good | 6.0 | No | No | Yes |
| 4 | No | 23.71 | No | No | No | 28.0 | 0.0 | Yes | Female | 40-44 | White | No | Yes | Very good | 8.0 | No | No | No |

---

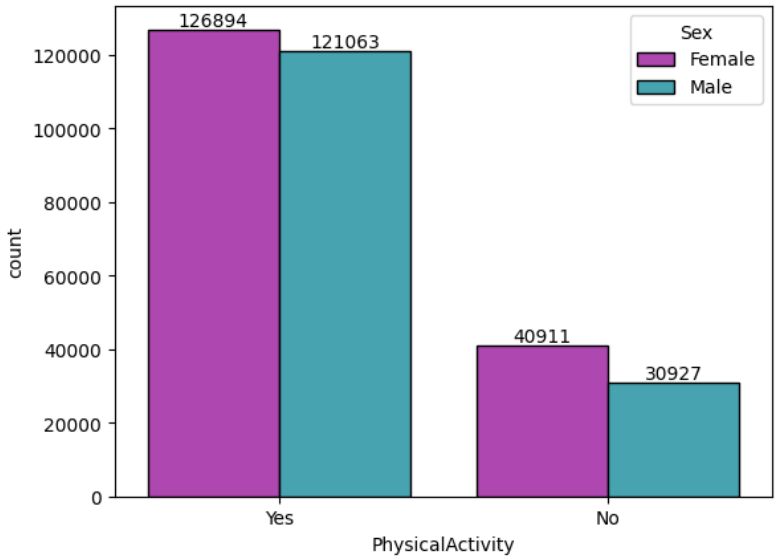
Teniendo en cuenta dicho Dataset y en base al interés mencionado anteriormente, se pudieron obtener los siguientes gráficos que indican ciertas características y que permitieron saber el comportamiento de las personas que fueron motivo de este estudio.

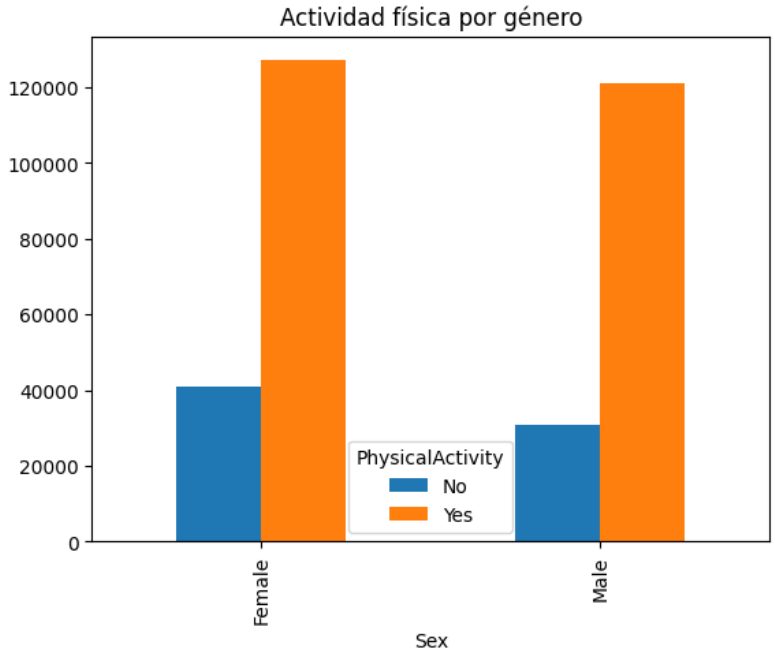
1. Actividad física según el género.

Un poco más de dos tercios de las personas realizan algún tipo de actividad física.

Si tomamos la información que brinda el Dataset, además se mantiene la relación por género.

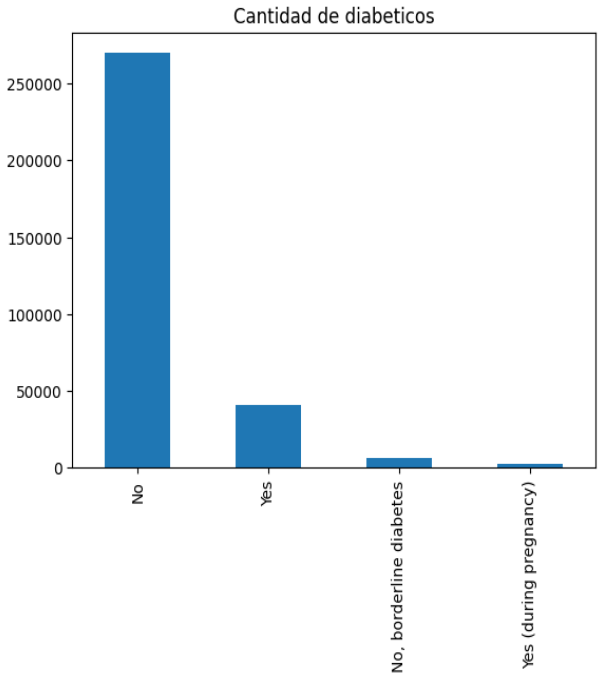
Los siguientes dos gráficos muestran estos datos de distinta forma.



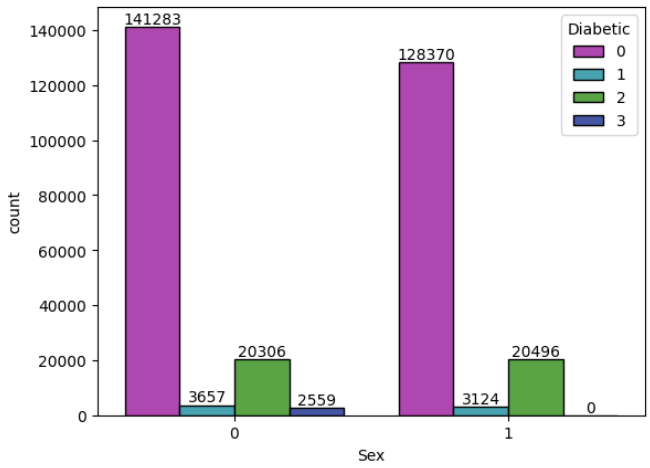
2. Cantidad de diabéticos.

Se puede llegar a la conclusión de que la gran mayoría de las personas no son diabéticas, una pequeña parte si lo es y los otros dos grupos son casi insignificantes en cuanto al total de personas.



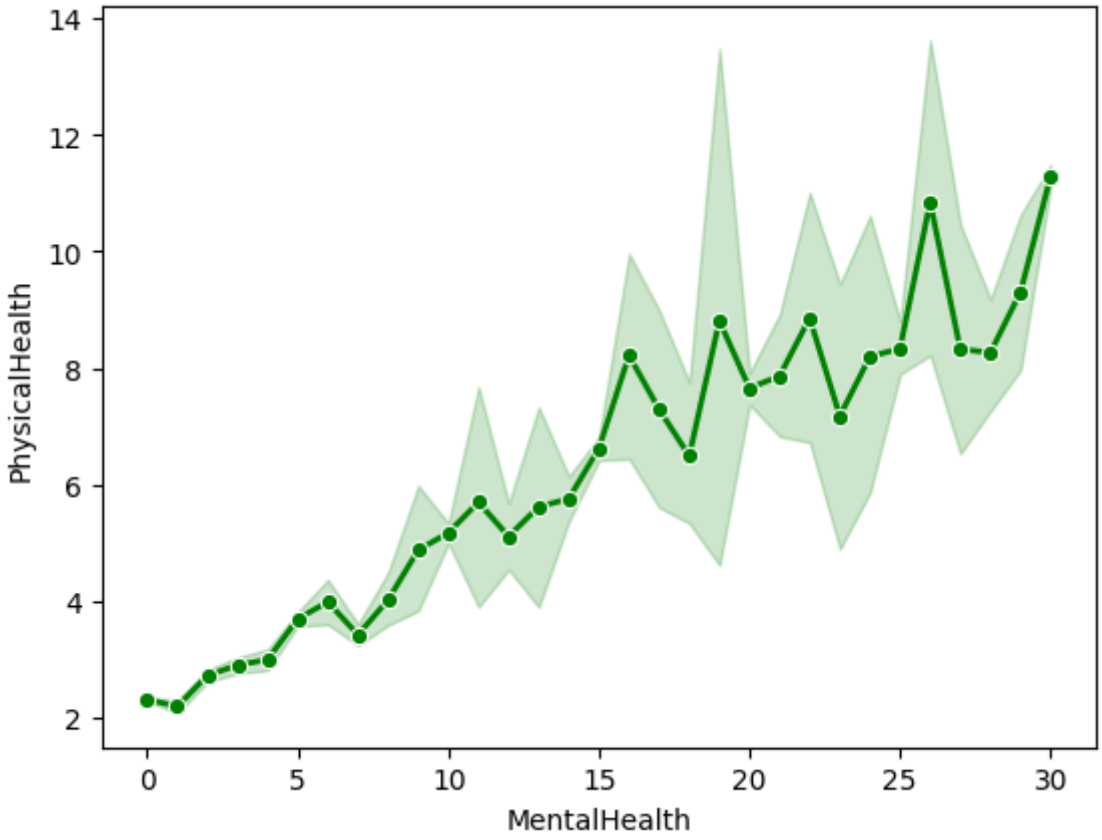

3. Cantidad de diabéticos por género.

Se puede llegar a la conclusión de que la relación Diabetes-Género es igual tanto para hombres y mujeres.



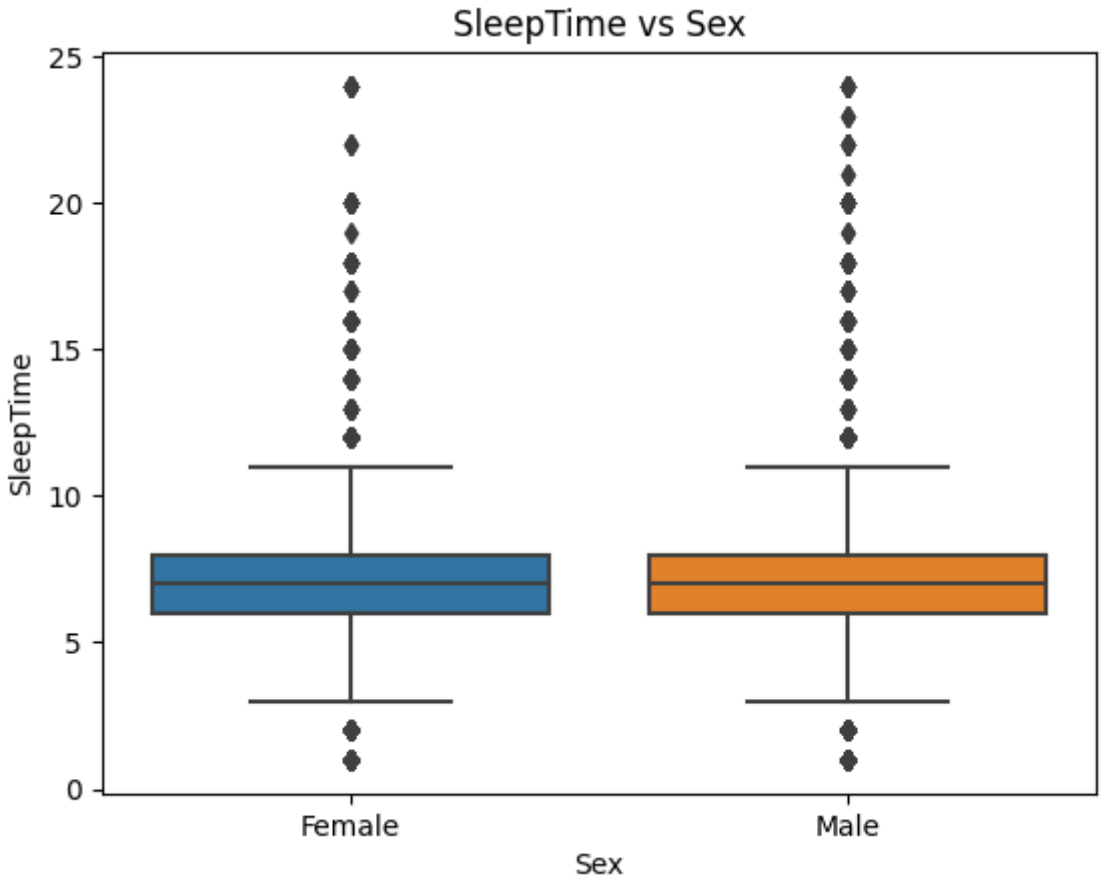
4. Salud mental y física.

Existe una relación directa entre la salud física y mental de las personas, quienes tienen algún tipo de actividad física están mejor en cuanto a salud mental que quienes no realizan actividad física.



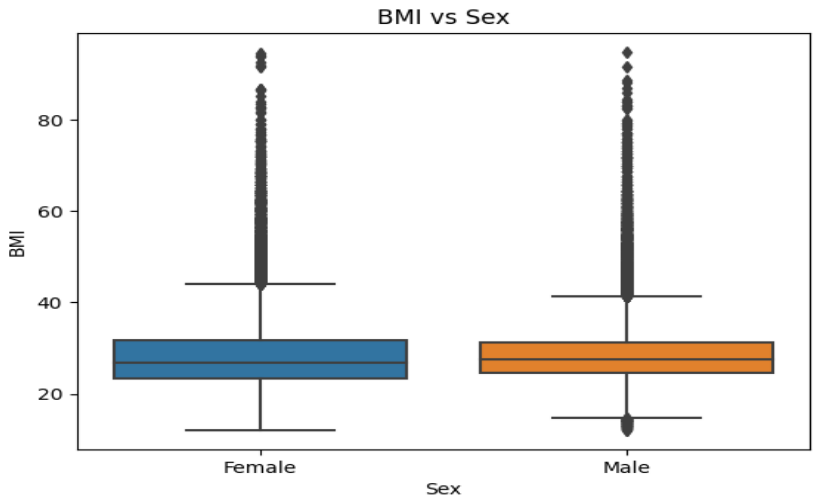
5. Tiempo de descanso por género.

La cantidad de horas de sueño es prácticamente igual tanto en hombres como en mujeres, no habiendo diferencias significantes en este sentido.



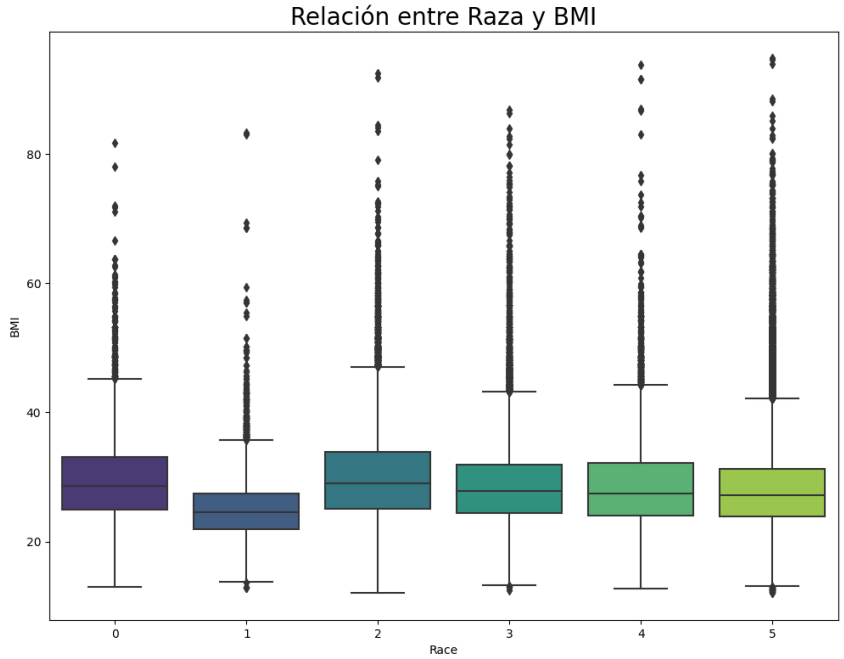
6. BMI (Índice de Masa Corporal) según el género de las personas.

El Índice de Masa Corporal en mujeres es levemente superior que en varones.




7. BMI (Índice de Masa Corporal) según la raza de las personas.

1) Las razas 0 y 2 tienen un valor más alto de BMI que las otras.
2) Los datos de la raza 2 se encuentran más dispersos que el resto.
3) Los valores de la raza 1 están por debajo de las demás razas.
4) La raza 5 tiene más outliers que las otras.



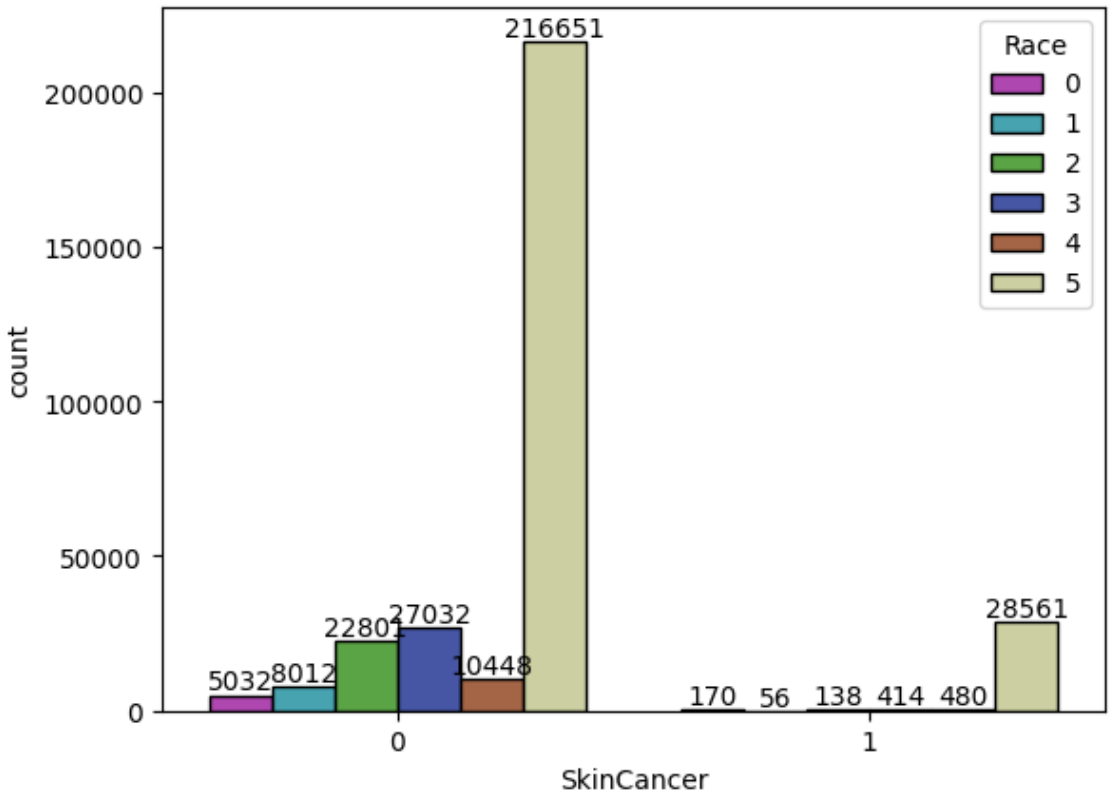
8. Cáncer de piel según raza.

Se puede ver en el siguiente gráfico que no hay incidencia en la raza que tienen las personas en cuanto al cáncer de piel.

Las personas de raza blanca son más que el resto pero porque la muestra es mayor en esa raza.

SkinCancer (0 -> NO / 1 -> SI)

Race (5 -> White)

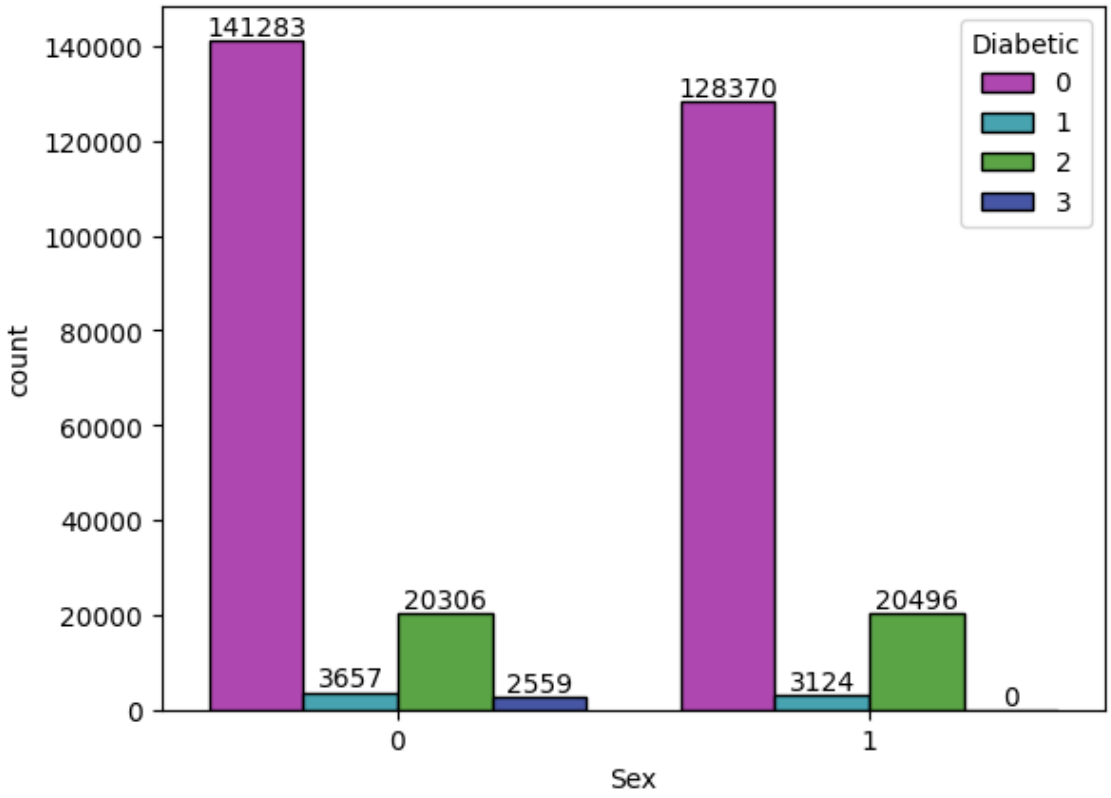


9. Diabetes según el género.

Se puede ver que la relación que existe entre diabéticos (por si o por no) es igual tanto en hombres como mujeres.

Diabetic (0 -> NO / 2 -> SI)

Sex (0 -> Mujer / 1 -> Hombre)

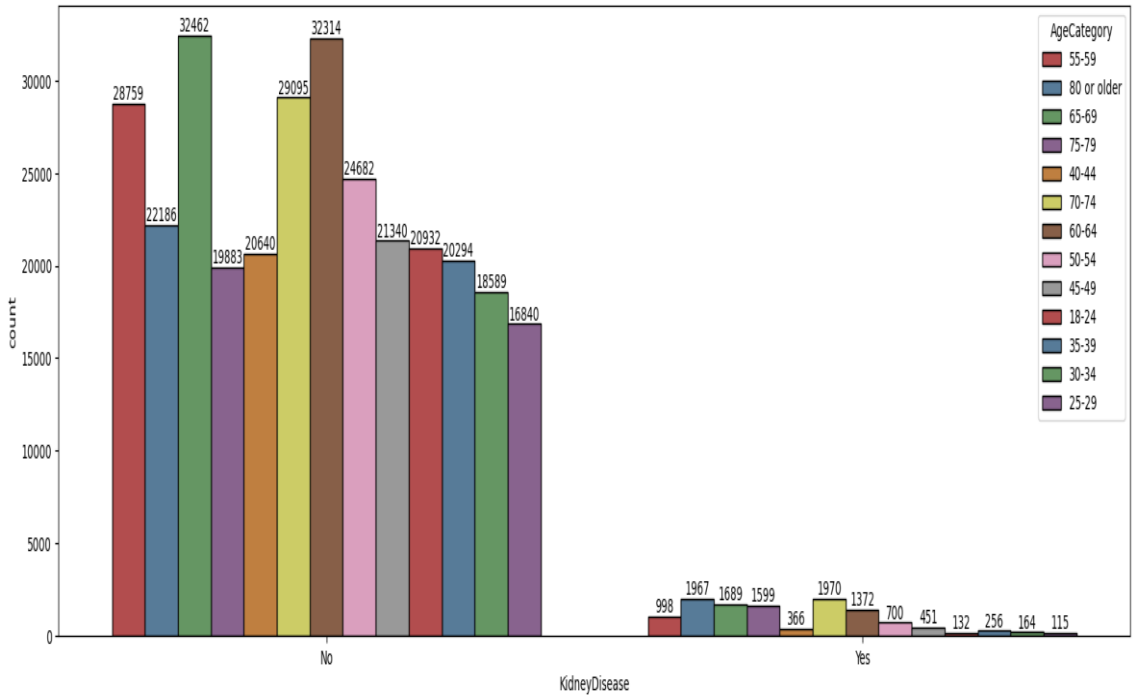


10. KidneyDisease según rango de edades.

Podemos ver en el gráfico que la gran mayoría de las personas no tienen trastornos renales y que los más jóvenes están en los valores más bajos de quienes si los tienen.

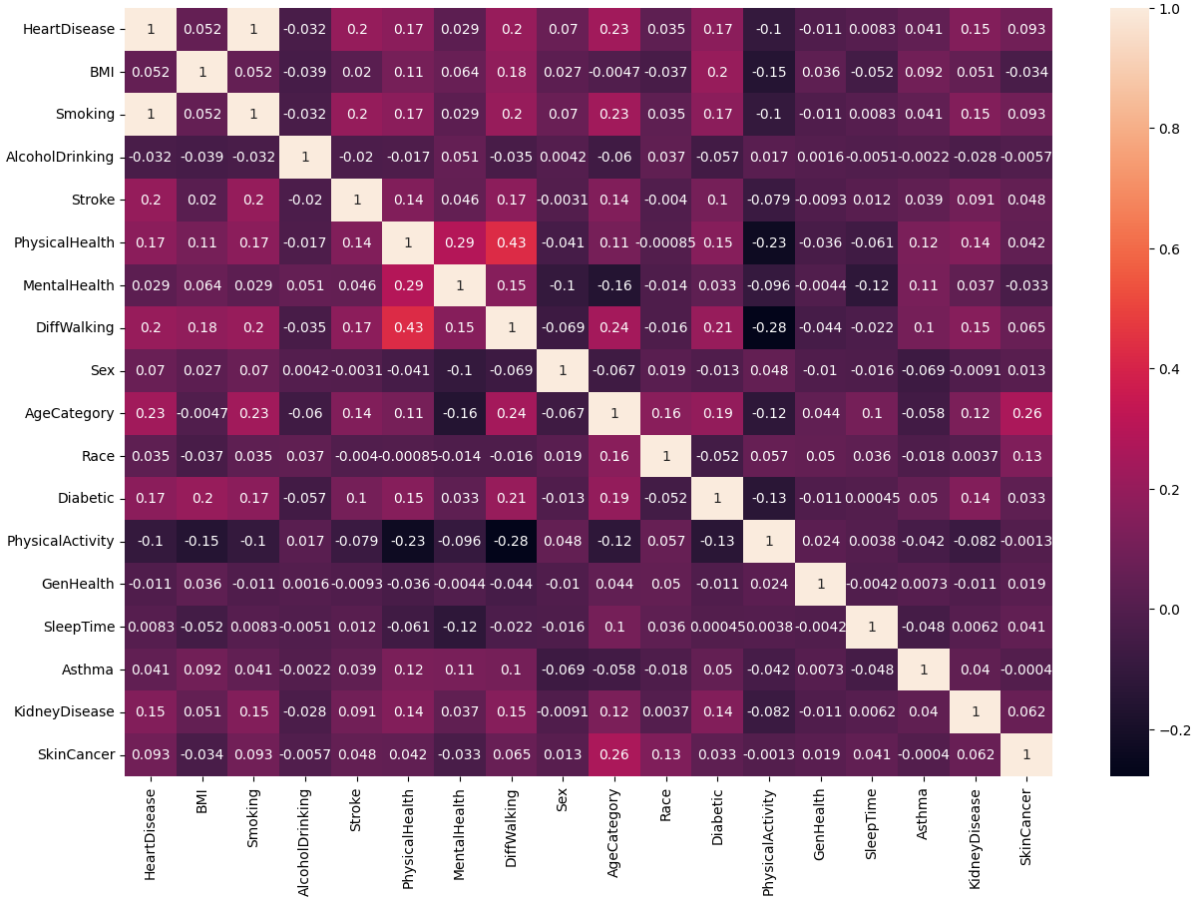
KidneyDisease = Nefropatía.

Nefropatía -> Es un trastorno renal en el cual anticuerpos (llamados IgA) se acumulan en el tejido del riñón. Nefropatía se refiere a un daño, enfermedad u otras anomalías del riñón.



Observando el mapa de calor (heatmap) además de ver la relación mencionada anteriormente, también podemos obtener la relación que existe (en menor medida) entre las siguientes variables:

1) Dificultad para caminar (DiffWalking) - Estado físico (PhysicalHealth).
2) Salud mental (MentalHealth) - Estado físico (PhysicalHealth).
3) Dificultad para caminar (DiffWalking) - Edad (AgeCategory).
4) Dificultad para caminar (DiffWalking) - Diabetes (Diabetic).
5) Cancer de piel (SkinCancer) - Edad (AgeCategory).



La tabla que devuelve el describe().T del Dataset me permite saber que variables son numéricas y cuales son categóricas.
Si tomo cada variable por separado lo que estoy realizando es un análisis univariado y para cada variable categórica puedo decir que no tienen sentido los percentiles ya que los valores son 1 y 0.

| | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| **HeartDisease** | 319795.0 | 0.085595 | 0.279766 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 |
| **BMI** | 319795.0 | 28.325.399 | 6.356.100 | 12.02 | 24.03 | 27.34 | 31.42 | 94.85 |
| **Smoking** | 319795.0 | 0.085595 | 0.279766 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 |
| **AlcoholDrinking** | 319795.0 | 0.068097 | 0.251912 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 |
| **Stroke** | 319795.0 | 0.037740 | 0.190567 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 |
| **PhysicalHealth** | 319795.0 | 3.371.710 | 7.950.850 | 0.00 | 0.00 | 0.00 | 2.00 | 30.00 |
| **MentalHealth** | 319795.0 | 3.898.366 | 7.955.235 | 0.00 | 0.00 | 0.00 | 3.00 | 30.00 |
| **DiffWalking** | 319795.0 | 0.138870 | 0.345812 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 |
| **Sex** | 319795.0 | 0.475273 | 0.499389 | 0.00 | 0.00 | 0.00 | 1.00 | 1.00 |
| **AgeCategory** | 319795.0 | 6.514.536 | 3.564.759 | 0.00 | 4.00 | 7.00 | 9.00 | 12.00 |
| **Race** | 319795.0 | 4.396.742 | 1.212.208 | 0.00 | 5.00 | 5.00 | 5.00 | 5.00 |
| **Diabetic** | 319795.0 | 0.300386 | 0.716480 | 0.00 | 0.00 | 0.00 | 0.00 | 3.00 |
| **PhysicalActivity** | 319795.0 | 0.775362 | 0.417344 | 0.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| **GenHealth** | 319795.0 | 2.220.904 | 1.534.647 | 0.00 | 1.00 | 2.00 | 4.00 | 4.00 |
| **SleepTime** | 319795.0 | 7.097.075 | 1.436.007 | 1.00 | 6.00 | 7.00 | 8.00 | 24.00 |
| **Asthma** | 319795.0 | 0.134061 | 0.340718 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 |
| **KidneyDisease** | 319795.0 | 0.036833 | 0.188352 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 |
| **SkinCancer** | 319795.0 | 0.093244 | 0.290775 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 |

Modelo de clasificación.

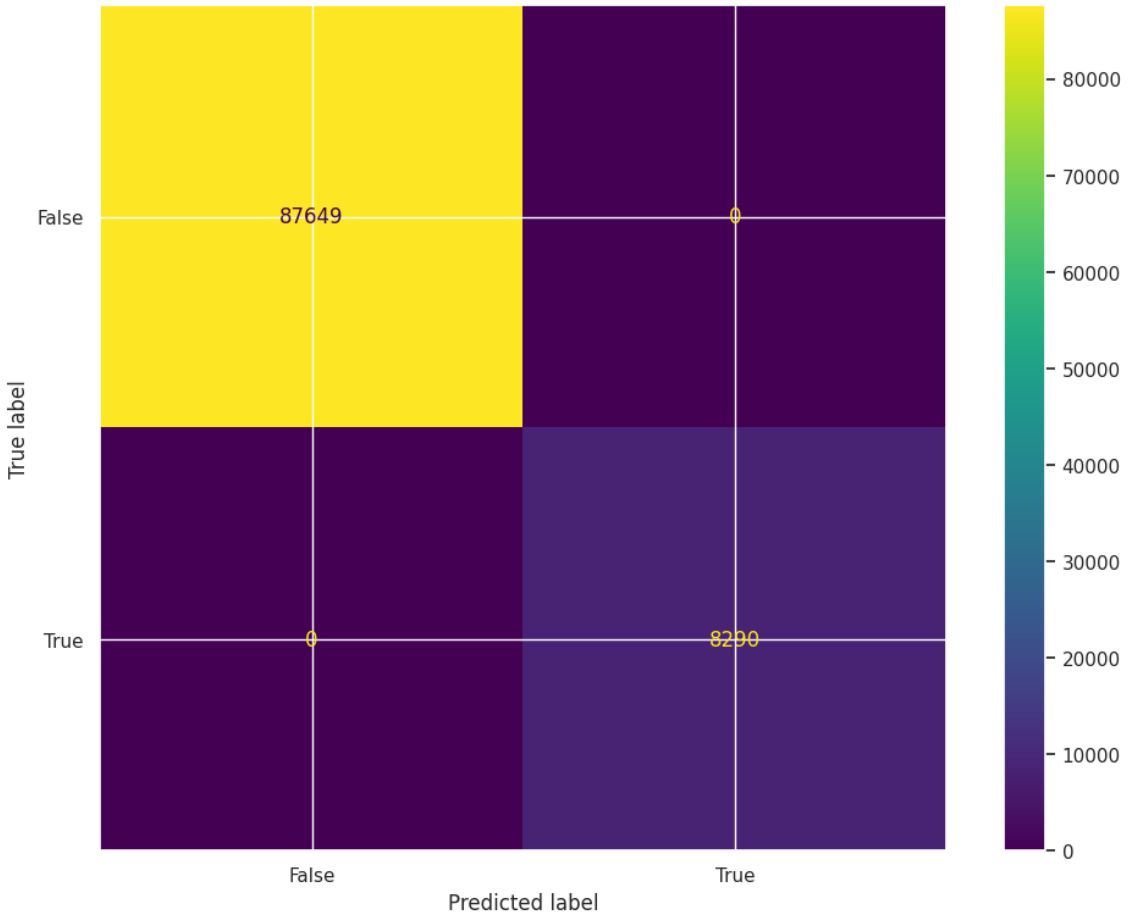
Elección del algoritmo.

De acuerdo con el tipo de variables que tenemos en nuestro Dataset y por sr una variable de clasificación la que tenemos que predecir (HeartDisease), elegimos el algoritmo de clasificación DecisionTreeClasifier.

Para evaluar la performace de nuestro modelo aplicado aplicaremos la matriz de confusión, en la misma se indican:

- Verdaderos positivos, resultados en el que el modelo predice correctamente la clase positiva.
- Falsos positivos, resultados en el que el modelo predice incorrectamente la clase positiva cuando en realidad es negativa
- Verdaderos negativos, resultado donde el modelo predice correctamente la clase negativa.
- Falsos negativos, resultado en el que el modelo predice incorrectamente la clase negativa cuando en realidad es positiva.

Tanto falsos positivos como negativos son indeseados, conocidos como error de tipo 1 y error de tipos 2.




Podemos verificar en la matriz de confusión que no existen falsos positivos y tampoco falsos negativos.
Se observa que 87649 y 8290 individuos quedaron bien clasificados, mientras que no hubo malas clasificaciones por lo expresado anteriormente.

## Insights obtenidos

* El Dataset no tiene datos nulos ni repetidos.
* Las variables ya fueron transformadas a numérico para poder hacer los cálculos que correspondan.
* Todas las variables las preciso para poder responder las preguntas de interés/hipótesis.
* No parece ser necesario agregar datos calculados para responder dichas preguntas.
* Existe una relación directa entre la salud física y mental de las personas.
* Un poco más de 2/3 de las personas realizan algún tipo de actividad física, además se mantiene la relación por género.

Habiendo sido aplicado Data Profiling al Dataset elegido para este proyecto (heart_2020_cleaned.csv) podemos obtener los siguientes insights:

* La variable que expresa enfermedad cardíaca (HeartDisease) está altamente correlacionada en general con el tabaquismo (Smoking), por lo tanto, podemos afirmar que HeartDisease es dependiente de Smoking.

Pudimos comprobar que existen variables que no se distribuyen normalmente las cuales habrá que modificar, seguramente realice raíz cuadrada para que las
distribuciones estén más cerca a la normalidad.
Las variables son HeartDisease, Smoking, AlcoholDrinking, Stroke, Diabetic,KidneyDisease, SkinCancer siendo todas ellas categóricas.