Heart Disease Data Indicator (DS-160 Final Project & Report)

By Ferdinand Yeke

DS-160-01 Intro into Data Science

December 8, 2025

Bellarmine University

Dataset Source: Indicators of Heart Disease (2022 UPDATE)

Introduction:

This is a 2020-2022 dataset describing health conditions, history, and symptoms of the respondents, which was from the CDC (Central Disease Control) annual telephone survey. This telephone survey took place in the United States of America in 2022 in all 50 states. With these interviews, the BRFSS (Behavioral Risk Factor Surveillance System, which the CDC is part of) does more than 400,00 adults interviews each year. In this dataset, about 246,000 adults were interviewed about their health.

In this dataset, not only there are 246,000 adults interview responses, but there are also 40 columns for each question which is their **Sex**, **General Health**, **Physical Health Days** (in the last 30 days if they have bad physical health), **Mental Health Days** (in the last 30 days if the had bad mental health), **Last Check Up Time**, **Physical Activities** (if they have been doing physical activities often or not), **Sleep hours**, **Removed Teeth**, **Had Heart Attack**, **Had Angina**, **Had Stroke**, **Had Skin Cancer**, **Had COPD** (Chronic Obstructive Pulmonary Disease), **Had Depressive Disorder**, **Had Kidney Disease, Had Arthritis**, **Had Diabetes**, **Deaf Or Hard Of Hearing**, **Blind Or Vision Difficulty**, **Difficulty Concentrating**, **Difficulty Walking**, **Difficulty Dressing and/or Bathing**, **Difficulty Errands**, **Smoker Status, E Cigarette Usage**, **Chest**

Scan, Race/Ethnicity Category, **Age Category**, **Height In Meters**, **Weight In Kilograms**, **BMI** (Body Mass Index), **Alcohol Drinkers**, **HIV Testing**, **FluVaxLast12** (Flu Vaccination in the last 12 years), **PneumoVaxEver**, **TatanusLast10Tdap**, **HighRiskLastYear**, and **CovidPos**.

Research Question:

-***Can my model predict potential of heart disease based on HadHeartAttack and other important variables like GeneralHealth?***

-Can my model accurately classify the potential of heart disease based on numerical and categorical columns like hasSmoked, BMI, SleepHours, and the like?

This classification problem is meaningful since it can help doctors and medical professionals on the various factors on what can have an individual being suspectable to heart disease or not, while also doing predictions as well.

Dataset Description:

- **Number of rows: 246,022**
- **Number of Columns: 40**
- **Variables:**
    - **GeneralHealth: Ranges from Excellent to Poor (Categorical)**
    - **State: Describes the data the respondent was in (Categorical)**
    - **Sex: Describes the Sex of the respondent (Categorical)**
    - **PhysicalHealthDays: Describes in the last 30 days if the respondent had any negative physical health conditions. (Categorical)**

- **MentalHealthDays:** Describes in the last 30 days if the respondent had any negative mental health conditions. (Categorical)
- **LastCheckupTime:** Describes the last time the respondent had a medical check up with a doctor (Categorical).
- **PhysicalActivites:** Describes if the respondent regularly does physical activities. (Categorical & Binary)
- **SleepHours:** Describes the respondent average amount of sleep in hours. (Numerical)
- **RemovedTeeth:** Describe the respondent's teeth removal amount (Categorical)
- **HadHeartAttack:** Describes if the respondent had a heart attack or not (Categorical & Binary)
- **HadAngina:** Describes if the respondent had Angina or not (Categorical & Binary)
- **HadStroke:** Describes if the respondent had a Stroke or not (Categorical & Binary).
- **HadAsthma:** Describes if the respondent had Asthma or not (Categorical & Binary)
- **HadSkinCancer:** Describes if the respondent had Skin Cancer or not (Categorical & Binary)
- **HadCOPD:** Describes if the respondent had Chronic Obstructive Pulmonary Disease or not (Categorical & Binary)
- **HadDepressiveDisorder:** Describes if the respondent had a depressive disorder or not (Categorical & Binary).

- **DeafOrHardOfHearing:** Describes if the respondent has deafness or struggling to hear (Categorical).
- **BlindOrVisionDifficulty:** Describes if the respondent is blind or having difficulty seeing or not. (Categorical & Binary)
- **DifficultyConcentrating:** Describes if the respondent has difficulty concentrating or not. (Categorical & Binary)
- **DifficultyWalking:** Describes if the respondent has difficulty walking or not (Categorical & Binary)
- **DifficultyDressingBathing:** Describes if the respondent has difficulty dressing or bathing or not (Categorical)
- **DifficultyErrands:** Describes if the respondent has difficulty doing errands/task or not (Categorical & Binary)
- **SmokerStatus:** Describes the respondent's smoker status (Categorical)
- **ECigaretteUsage:** Describes the respondent's E Cigarette Usage (Categorical)
- **ChestScan:** Describes if the respondent has done a chest scan or not (Categorical & Binary)
- **RaceEthnicityCategory:** Describes respondent's race/ethnicity (Categorical)
- **AgeCategory:** Describes the respondent's age in categories like from Age 18 to 24, 25 to 29, etc. (Categorical)
- **HeightInMeters:** Describes the respondent's height in meters (Numerical)
- **WeightInKilograms:** Describes the respondent's weight in kilograms (Numerical)
- **BMI:** Describes the respondent's Body Mass Index (Numerical)

- **AlcoholDrinkers:** Describes the respondent's alcohol consumption status (Categorical).
- **HIVTesting**: Describes if the respondent has done HIV Testing or not (Categorical and Binary)
- **FluVaxLast12: Describes if the respondent has taken the flu vaccine in the last 12 years** (Categorical)
- **PneumoVaxEver:** Describes if the respondent has ever taken the Pneumococcal Vaccine or not (Categorical & Binary)
- **TetanusLast10Tdap:** Describes if the respondent has ever taken the tetanus vaccines (Categorical)
- **HighRiskLastYear:** Describes if the respondent was of high risk of heart disease or not (Categorical & Binary)
- **CovidPos:** Describes if the respondent was ever marked positive having covid (Categorical).

- **Data Source Credibility: A survey done in 2022 by the CDC related to hundreds of thousands respondent's health status that is a heart disease indicator or not (Indicators of Heart Disease (2022 UPDATE)).**

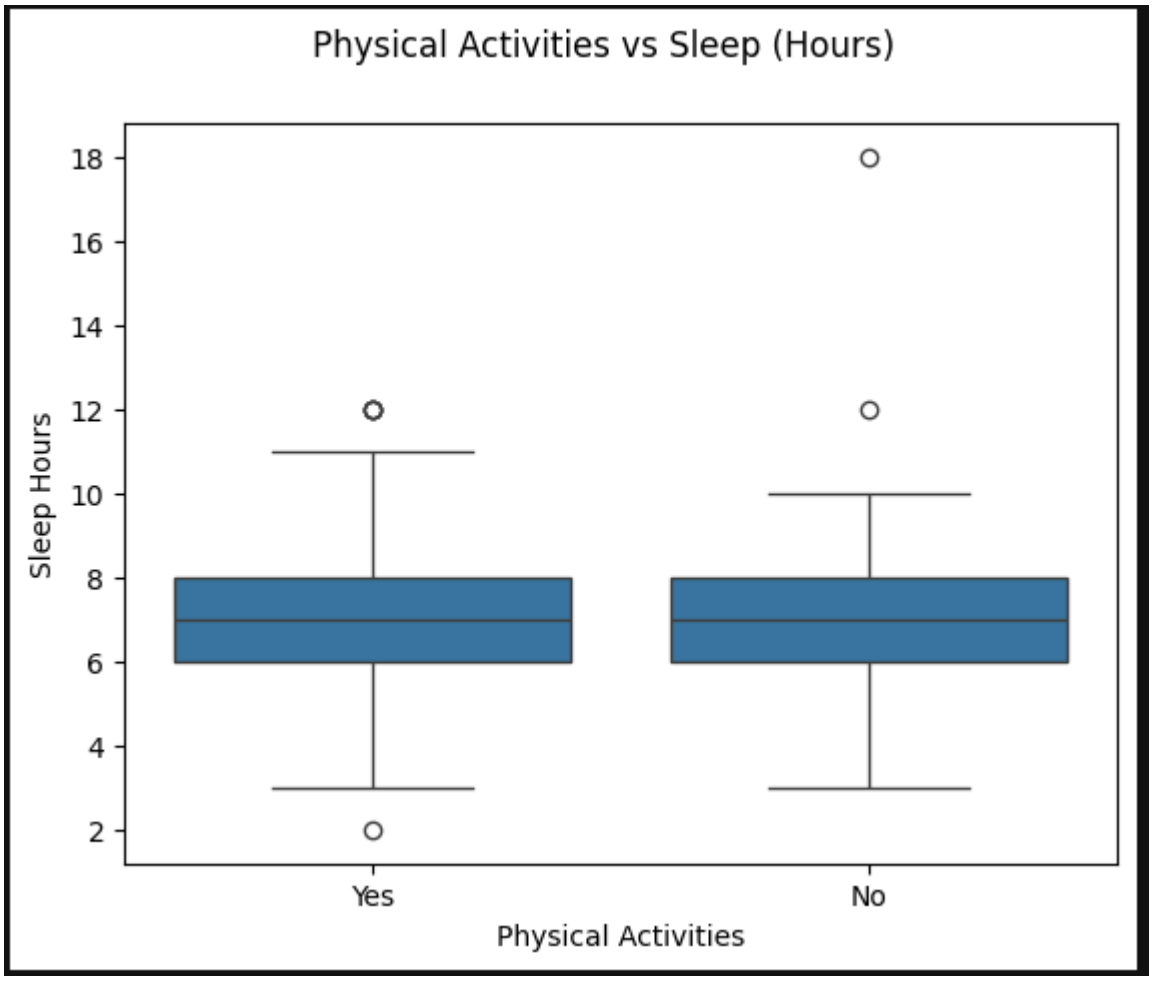
EDA Analysis:

- **Starting with Python:**
  - **Data Statistics:**

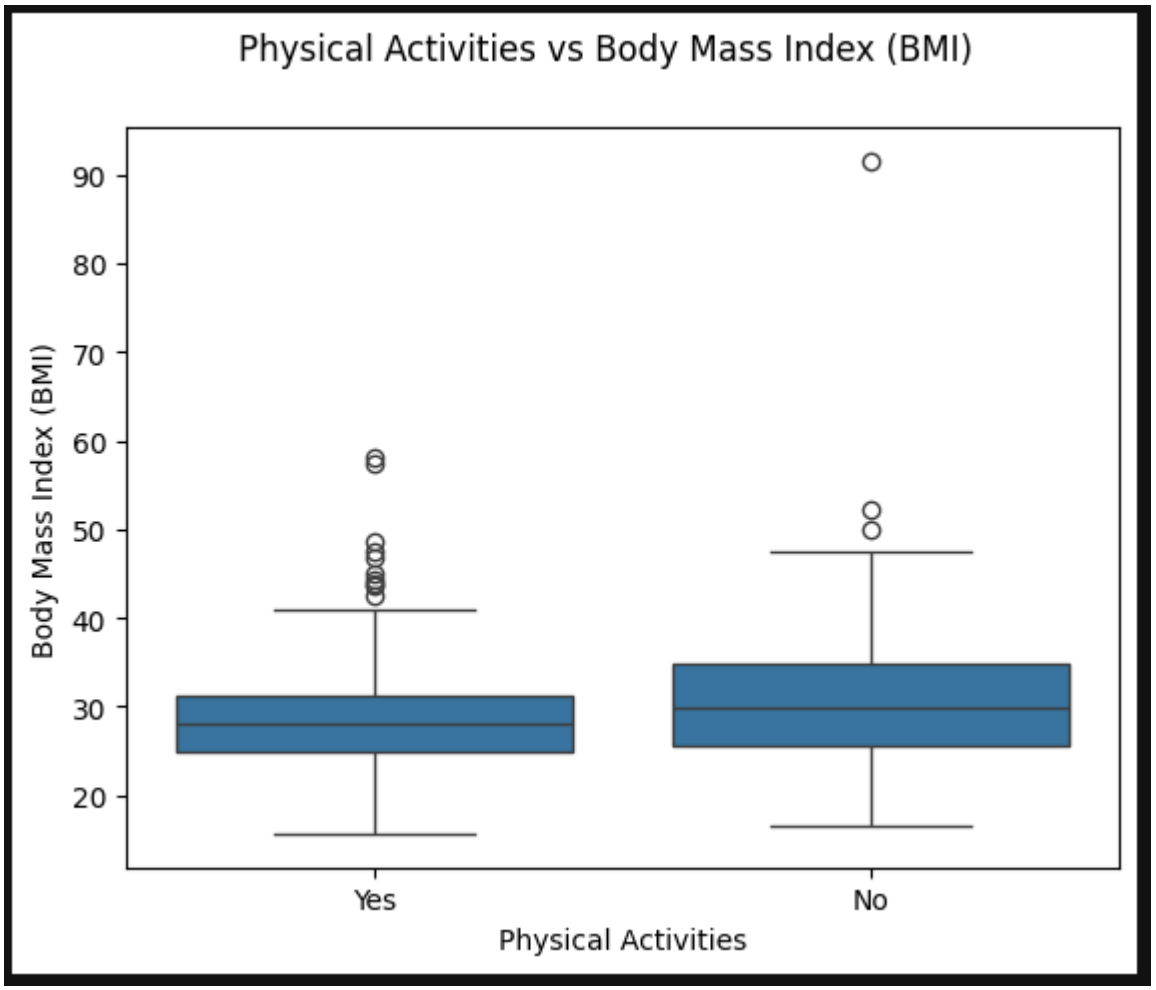
	PhysicalHealthDays	MentalHealthDays	SleepHours	HeightInMeters	WeightInKilograms	BMI
count	246022.000000	246022.000000	246022.000000	246022.000000	246022.000000	246022.000000
mean	4.119026	4.167140	7.021331	1.705150	83.615179	28.668136
std	8.405844	8.102687	1.440681	0.106654	21.323156	6.513973
min	0.000000	0.000000	1.000000	0.910000	28.120000	12.020000
25%	0.000000	0.000000	6.000000	1.630000	68.040000	24.270000
50%	0.000000	0.000000	7.000000	1.700000	81.650000	27.460000
75%	3.000000	4.000000	8.000000	1.780000	95.250000	31.890000
max	30.000000	30.000000	24.000000	2.410000	292.570000	97.650000

- In these statistics, what I noticed here is that the mean amount of physical health days (days where the respondents did not feel good) is at 4 days, like with the mental health days. The Average for sleep is at 7 hours. The max weight in kilograms here is 292.57, while the minimum is 83.61. The Standard Deviation of BMI is 6.51 as well.

- **Plots:**

- 

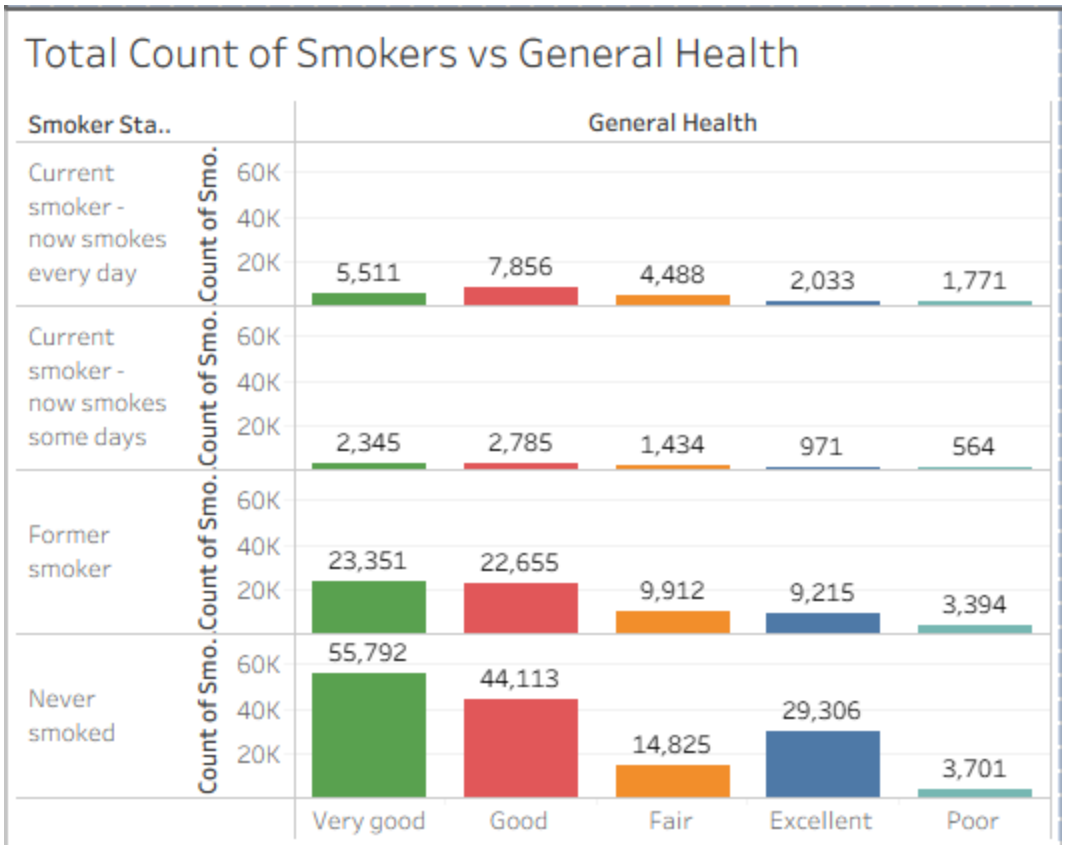
- Here, the first plot here is the comparison between Physical Activities vs Sleep (Hours). In this, it shows that with the respondents that regularly do physical activities, The average is at 7-7.5, with the max being 11 hours and min 3. There are some outliers in the Yes portion, where one outlier is **at 12 hours** while the other is at **2 hours.** The respondents that do not really do regular physical activities, the average amount of sleep that is present there is 7, just like the Yes portion. The max in the No portion is 10 hours, while the min is 3. Strangely, in the No portion, there are some extreme outliers where there were some that slept for 18 hours, and some at 12 hours.

- 

- In the second plot, Physical Activities is being compared to Body Mass Index. In the side where respondents said that they regularly do physical activities, the average BMI is around 27 to 28. The Max for this is at around 41, while the Minimum is at around 12. In this, there are a lot of outliers here, ranging from 42 to 59 BMI. In the side where respondents said that they do not regularly do physical activities, the average BMI is at around 30, which is somewhat slightly higher than the average of respondents that do physical activities. The max in the respondents that said no is 48 BMI, while the min is 12 to 13. In the No side, there are far less outliers, but there are some extreme outliers where some of the respondents have a BMI of around 90.

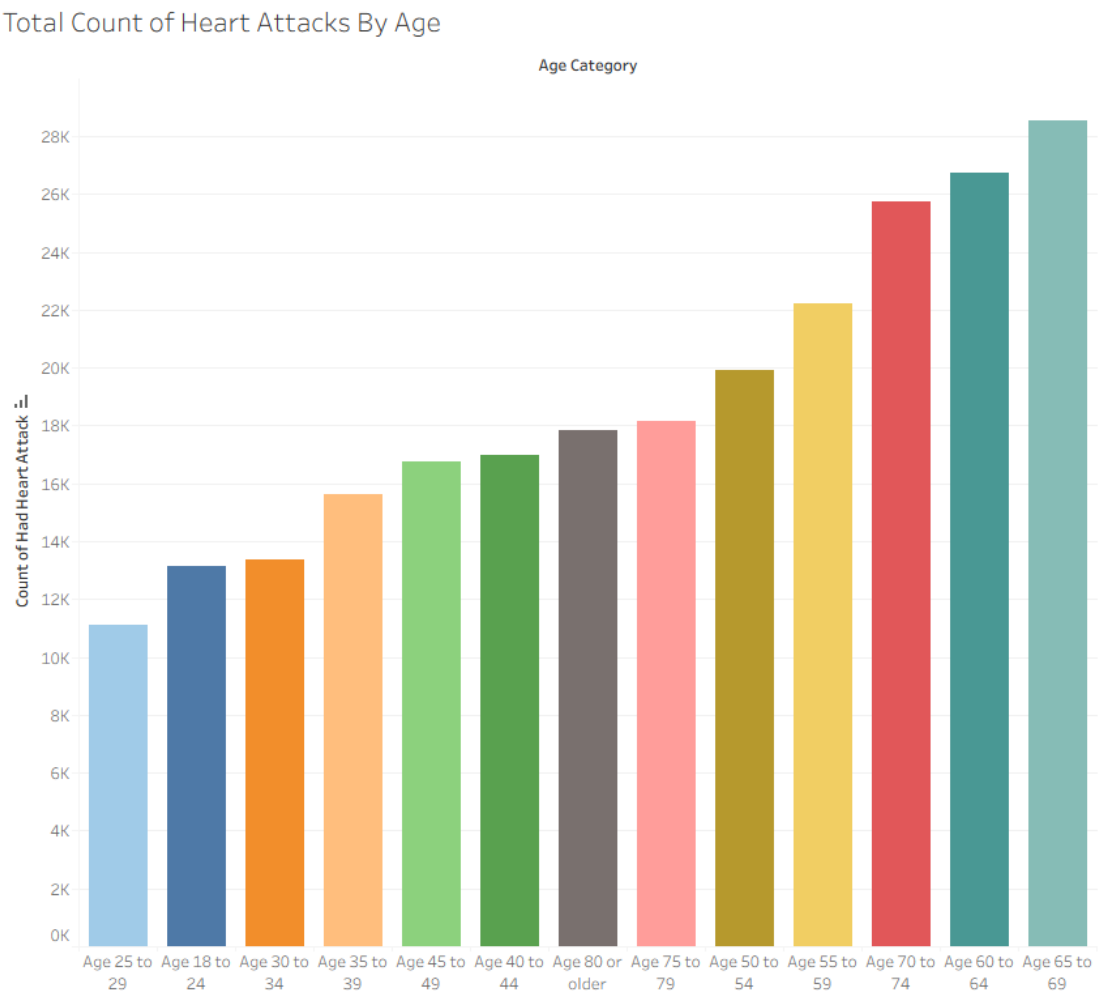
- **Tableau:**
  - 

  - In the first Tableau graph visualization, this goes over the general health of the respondents that do drink and respondents that don't drink. In the "Very Good" sector, the respondents that responded with having very good health while drinking alcohol is 54,266. For much of this, they might have very good control of their habits. With the respondents that responded with having very good health while not drinking alcohol is 32,733. In the "Good" sector, 37,379 responded with not drinking alcohol, while 40,030 said they do. In this sector, it is actually the closest/even between the yes and no respondents. In the "Excellent" sector, 15,423 responded with not drinking alcohol, while 26,102 said that they do. In the "Fair" sector, 18,492 responded with not drinking alcohol, while 12,167 responded that they do. In the "Poor" sector, 6,688 respondents said they do not drink alcohol, while 2,742 said that they do.

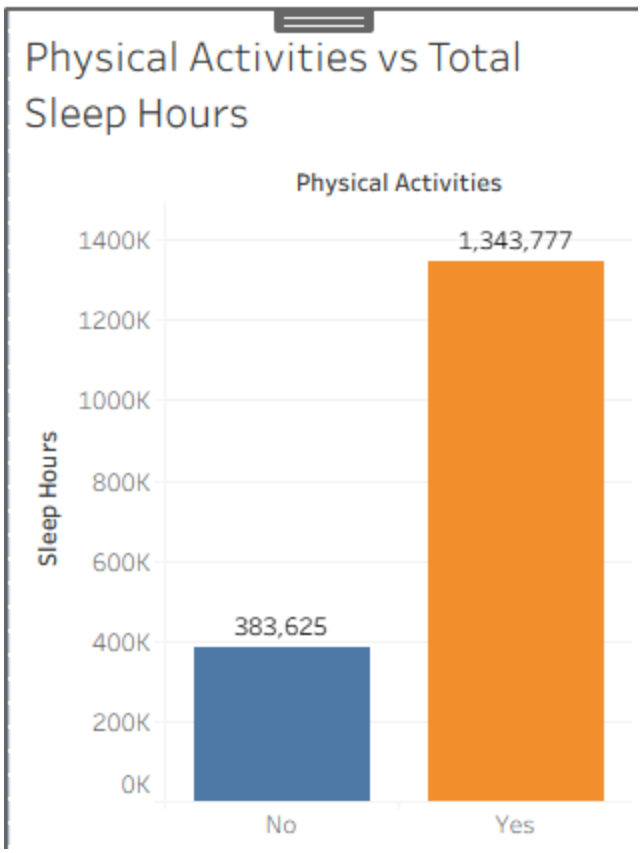
- 

- In the second Tableau visualization, this is the comparison of the Total Smoker status of respondent's vs the General Health.
  - In the "Never Smoked" Sector:
    - **55,792 respondents reported** having "**Very good**" health while not ever smoking.
    - **44,113 respondents** reported having "**Good**" health while never smoking.
    - **14,825 respondents reported** having "**Fair**" health while never smoking.
    - **29,306 respondents reported** having "**Excellent**" health while never smoking.

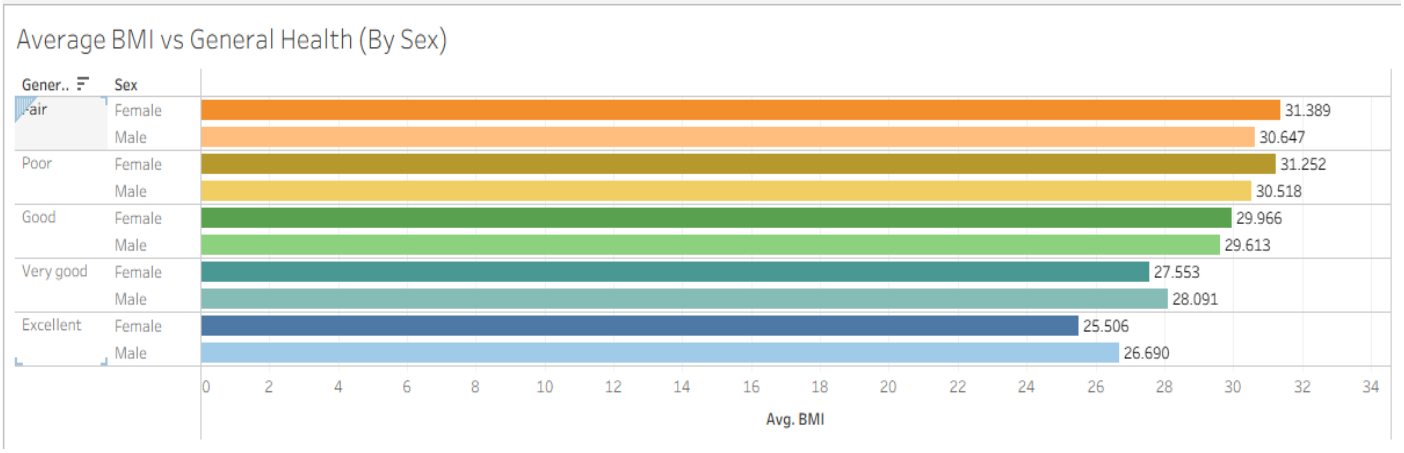
        - **3,701 respondents reported** having "**poor**" health while never smoking.
    - In the "Former Smoker" Sector:
        - **23,351 respondents reporte**d having "**Very good**" health while being a former smoker.
        - **22,655 respondents** reported having "**Good**" health while being a former smoker.
        - **9,912 respondents** reported having "Fair" Health while being a former smoker.
        - **9,215 respondents** reported having "Excellent" Health while being a former smoker.
        - **3,394 respondents** reported having "Poor" health while being a former smoker.
- Those are just some of the smoker types. The interesting thing about this that seems obvious is that those that never smoker had the highest count of Very Good to fair health in this dataset. Something that also seems interesting about this is that those that are currently smoking and are former smokers have a lower count of poor health than on the sector with poor health that never smoked before.

- 

- In the Third Tableau Visualization, this is the Total Count of Heart Attacks by Age. The Ages between 65 to 69 seems to have the most amount of heart attacks in the entire dataset, while Age 70 to 74 and 60 to 64 follow suit. The concerning thing that I see in this dataset is seeing younger adults from age 18 to 24 and 25 to 29 having a count of heart attacks at all. Even when they have the least amount of heart attacks in the dataset, it is still concerning since having heart attacks at a younger age as an adult is not normal.

- 

- In the Fourth Tableau Visualization, this is a Total Count of Sleep hours VS the Physical Activities by respondents. In this, the respondents that does regular activities have a total amount of sleep hours more than the total amount of respondents that do not do regular physical activity.



- In the Fifth and last Tableau Visualization, it is the Average BMI (Body Mass Index) vs the General Health of the respondents. With most of the respondents saying that they have “Excellent” Health, the average BMI is 26.690 for the Male respondents, while 25.506 is the average for the Female Respondents. With most of the respondents saying that they have “Very Good” Health, the average BMI is 28.091 for the Male respondents, and 27.553 for the Female respondents. With most of the respondents saying that they have “Fair” Health, the average BMI is 31.389 for the Female respondents, and 30.647 for the Male respondents. With most of the respondents saying that they have ”Poor” Health, the average BMI is 30.518 for the Male respondents, and 31.252 for the Female Respondents.

Preprocessing:

With this dataset, there came three versions of this, the **heart_2020_cleaned.csv**, **heart_2022_with_nans.csv,** and **heart_2022_no_nans.csv.** The name is self-explanatory, and the heart_2022_no_nans.csv was used, due to it being more recent as well as it being already cleaned. Since the data I imported was cleaned, I did not need to do any manual cleaning for the dataset. The columns are named right to where there are no spaces, the names themselves look good following the “Camel-Case” as well. The rows are also all filled, where I did not have to do any mean imputation or any imputation in the dataset.

With the feature encoding, I decided keep columns such as **Sex, GeneralHealth, PhysicalActivities, SleepHours, HadHeartAttack, HadAngina, HadStroke, HadCOPD, HadKidneyDisease, HadDiabetes, DifficultyWalking, SmokerStatus, AgeCategory, BMI, AlcoholDrinkers, HIVTesting, HighRiskLastYear.** I then dropped all the rest. After dropping the columns that were not the ones kept, I then made dummy variables for ALL the Categorical

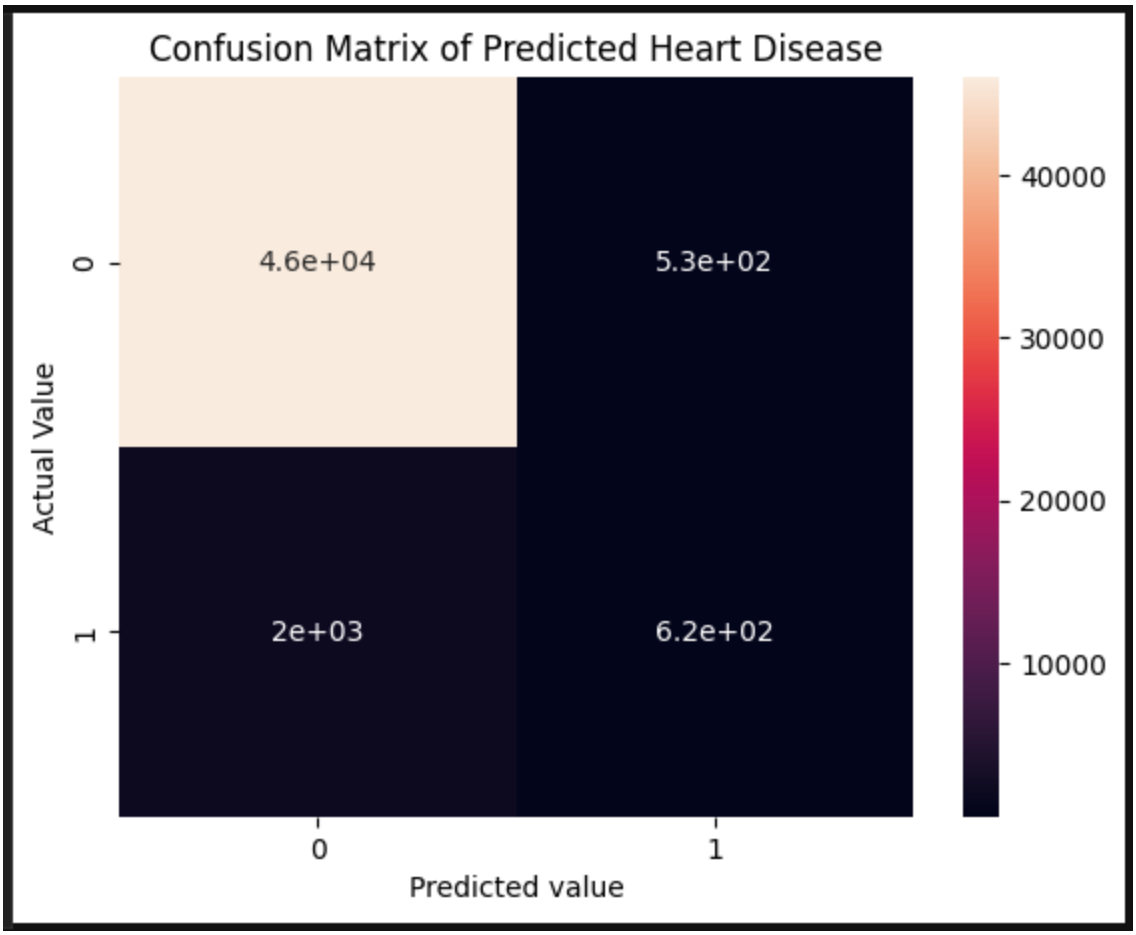
columns. However, **HadHeartAttack will have a drop_first = True** applied to it, since this column will be fully binary as well as it **being the target variable.** HadKidneyDisease will also have a drop_first = True to reduce column size.

With Train/Test split, I have Variable "X" set equal to the new dataset, while dropping the HadHeartAttack_Yes column. Then, I have dependent variable "y" being set equal to the HadHeartAttack_Yes column. Again, HadHeartAttack_Yes column will be the target variable.

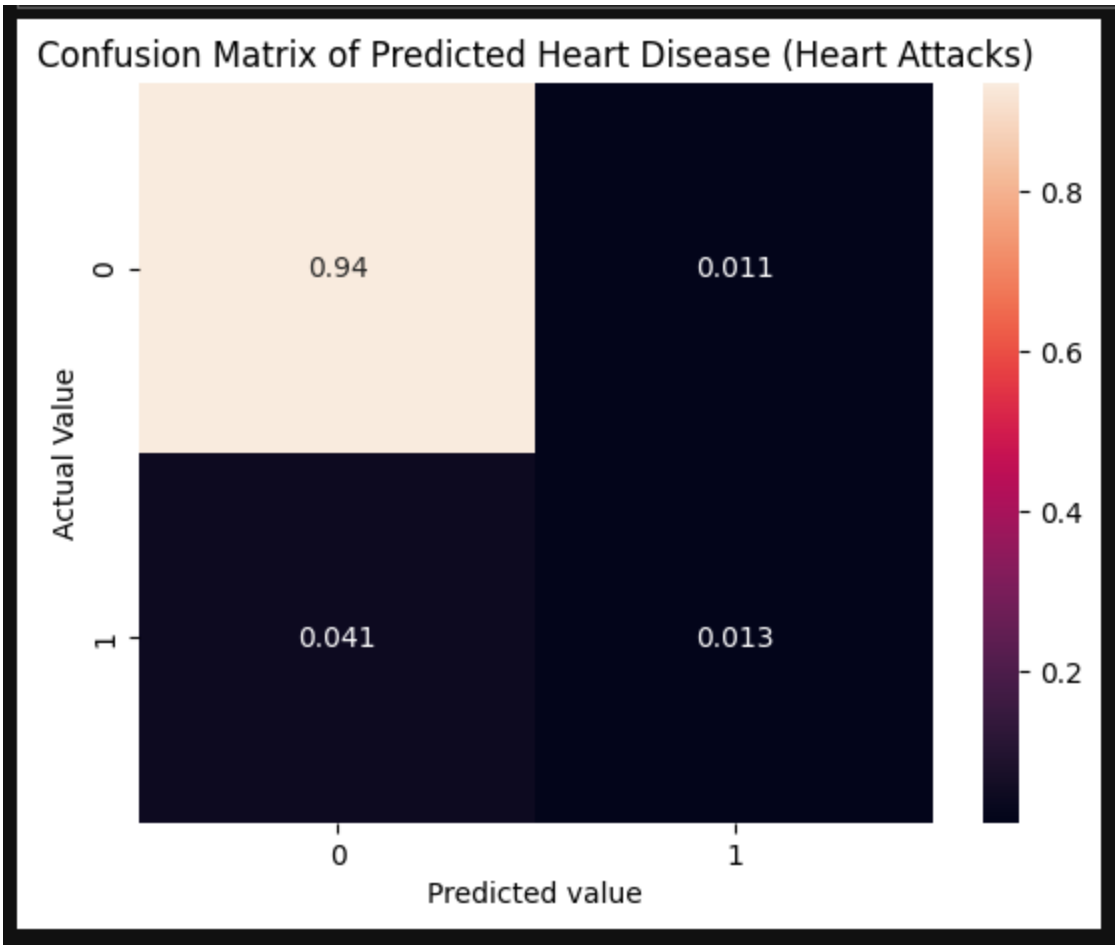
Logistic Regression Model

The model configuration is having variable "model" being set to the LogisticRegression() function, where the parameters for that is max_iter being set to 1000. The model was then fitted with the X_train.values, and y_train.values. The variable y_pred is the model's prediction from all the X_test.values.

The confusion matrix of the Model:

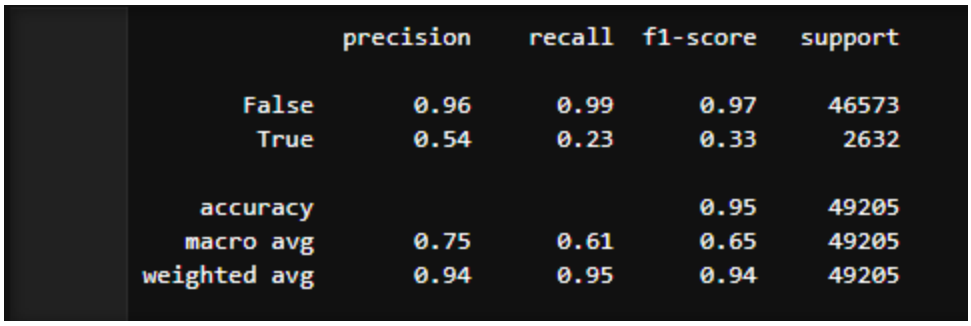

Confusion Matrix of Predicted Heart Disease
4.6e+04
5.3e+02
2e+03
6.2e+02
Actual Value
0
1
Predicted value
0
1
40000
30000
20000
10000

(Normalized):



- Starting with the normalized version of the confusion matrix (Which will be more of a clearer explanation). Here, this shows the model's prediction in a normalized format. In the true negatives, it predicted that 94% haven't had heart disease, and it was right. In the False Positives, it predicted that 0.01% had heart disease, and it is wrong. In the False Negatives, it predicted that 0.04% haven't had heart disease, and it is wrong. In the True Positive, it predicted that 0.013% had disease, and it is right.

The Precision, Recall, and F1-Score of the Model:

- 

	precision	recall	f1-score	support
False	0.96	0.99	0.97	46573
True	0.54	0.23	0.33	2632
accuracy			0.95	49205
macro avg	0.75	0.61	0.65	49205
weighted avg	0.94	0.95	0.94	49205

- First, the precision score of this model is at 96% at false, and 54% true. The Recall score of this model is at 99% for false, and 23% true. The F1-Score of this model is at 97% false and 33% true, **giving total accuracy at 95%**, which for a model on medical data, seems good.

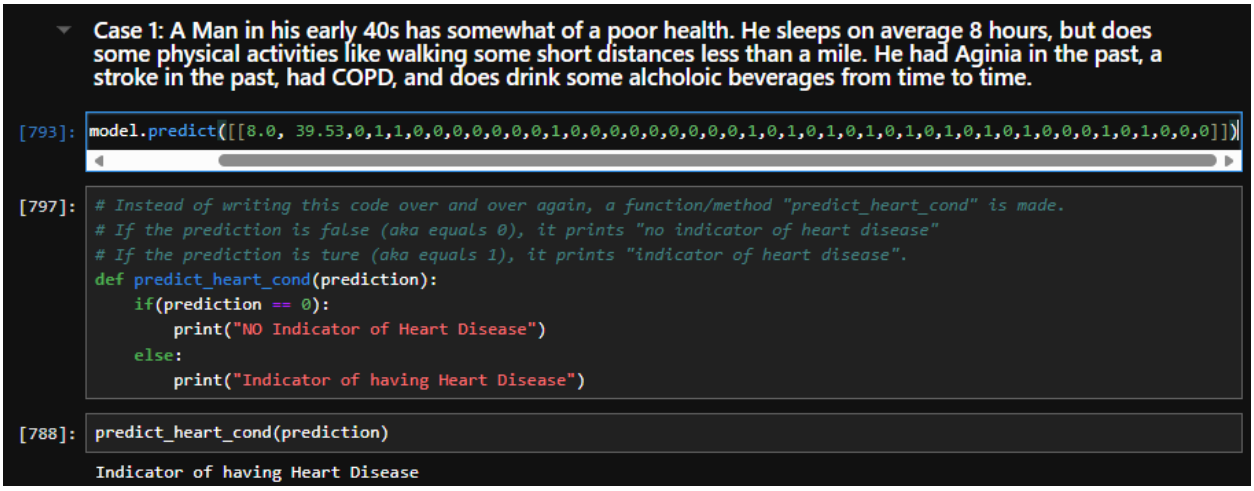
Interpretation of Results:

The model seems to perform well with good accuracy score and being great with the predictions to being right ratio.

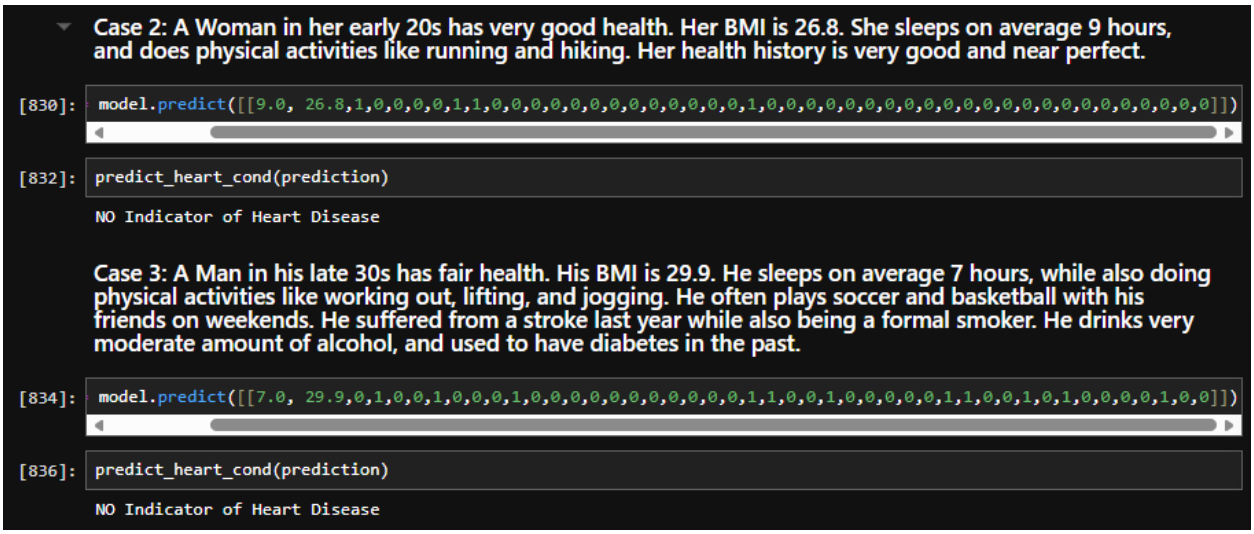
Findings & Interpretation:

When doing predictions like with example cases of a person of poor health, it seems that factors like high BMI, lack of physical activities, having Angina, a stroke, COPD, while smoking

and drinking can influence raise the indication of a person having heart disease.



But if a person has lack of problematic health problems, it decreases the indication of heart disease.



Again, the model is very strong, sitting at an accuracy score of 95%. The parts where it performed poorly is with the “True” of the HadHeartAttack_True, where many of them were low, while the highest score was at 0.53 for true.

Conclusion:

In summary of the results, the model performed considerably well with the predictions and accuracy score. The precision score for false is 0.96, while for true 0.54. The recall score for false is 0.99, while for true 0.23. The F1-Score for false is 0.97, while for truth it is 0.33. The total accuracy score for this is 0.95, which is very good and extremely promising possibly for medical use for predictions and knowing the factors that can lead to heart disease for an individual.

There are not really limitations for this dataset, basically everything from the respondent's background to health and state history is responded with.

What could be improved more with this dataset and model is to just have slightly more data for respondents that do have cases of health background that may raise an indicator of heart disease. But with this dataset, it seems already enough with 246,022 respondents, but more information, the better figuring out the many causes that can lead to heart disease and helping not just the individual, but everyone as well.

References:

Indicators of Heart Disease (2022 UPDATE)

The world's best hospital - Mayo Clinic