# Heart Disease Prediction

DANNY HUANG

# Introduction – Heart Disease

According to the 2024 Heart Disease and Stroke Statistics: A Report of U.S. and Global Data From the American Heart Association, heart disease has been the leading cause of death in the U.S. for 100 years.

18.6 million people died from cardiovascular disease globally, representing about 32% of all global deaths.

# Problem

Heart disease remains a leading cause of mortality, and early detection is crucial for effective treatment. However, identifying at-risk individuals using health data is challenging. This project aims to address this issue by developing a machine learning model that predicts heart disease with at least 85% accuracy, using data from the CDC.

# Stake Holders

Patients and Healthcare providers

Scientist and Researchers

# Objective: Developing a Model

By leveraging survey data and machine learning, we aim to identify individuals at higher risk. By building a predictive model of heart disease incidence using related comorbidities and associated risk factors as predictors, aiming to enhance early detection and preventative interventions through analysis of survey data.

# Data Collection

The data we are using are survey data record from 2022 Behavioral Risk Factor Surveillance System survey data (BRFSS)

This survey collects prevalence data among U.S. residents regarding risk behaviors, preventative health practices, chronic health conditions and health care access.

- The original dataset retrieved from the CDC had 328 columns and ~450,000 records.

The survey is conducted via telephone interviews, and also includes state-specific questions. Participants are 18 years and older.

# Data cleaning/wrangling

The original dataset consisted of 328 columns and over 450,000 records.

We set a threshold of 80% of the total records, and dropped all columns that were missing more than 80% of the dataset, 213 columns were dropped. Leaving us with 115 columns.

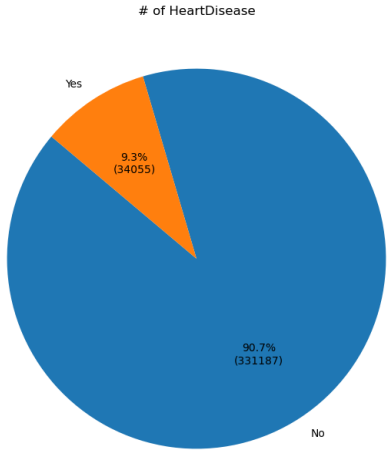
Then went through the data dictionary and removed 74 columns that were not in scope of answering our problem.

- Some examples of the removed columns were: IMONTH (month which survey was taken), INCOME (income bracket), EMPLOY(employment status)

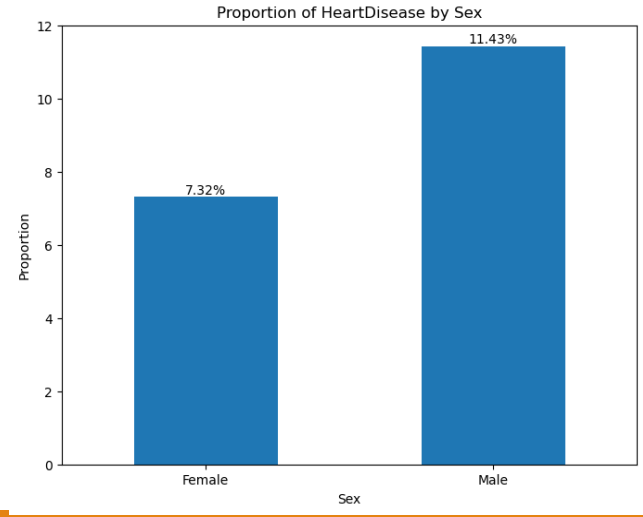
All columns were encoded so we renamed all 41 columns that were left. And after removing most of the NAs values, we were left with a dataset of 41 columns, and ~365,000 records.

# EDA - Demographics

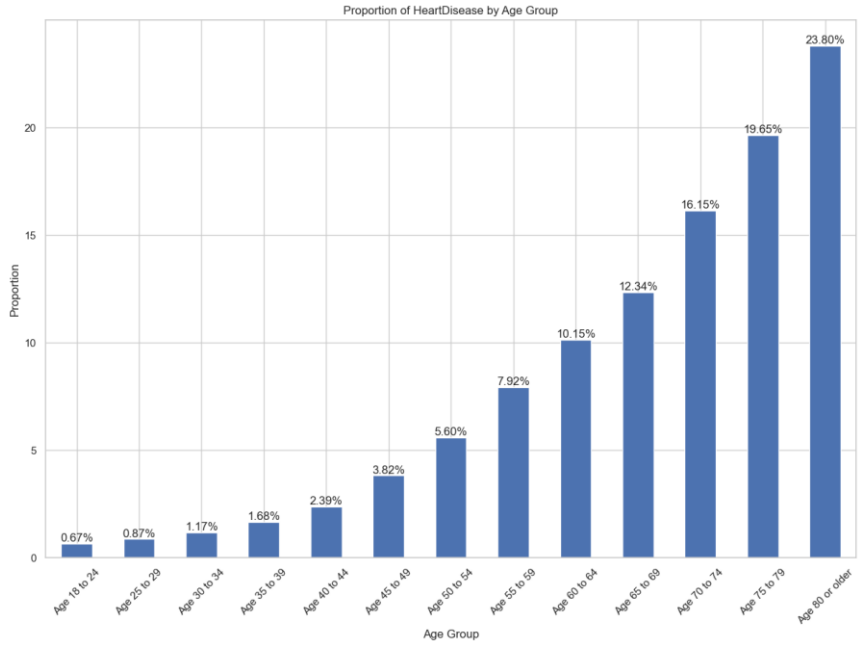
The data reveals that 9.3% of respondents (34,055 individuals) reported having heart disease, while 90.7% of respondents (331,187 individuals) did not report heart disease.

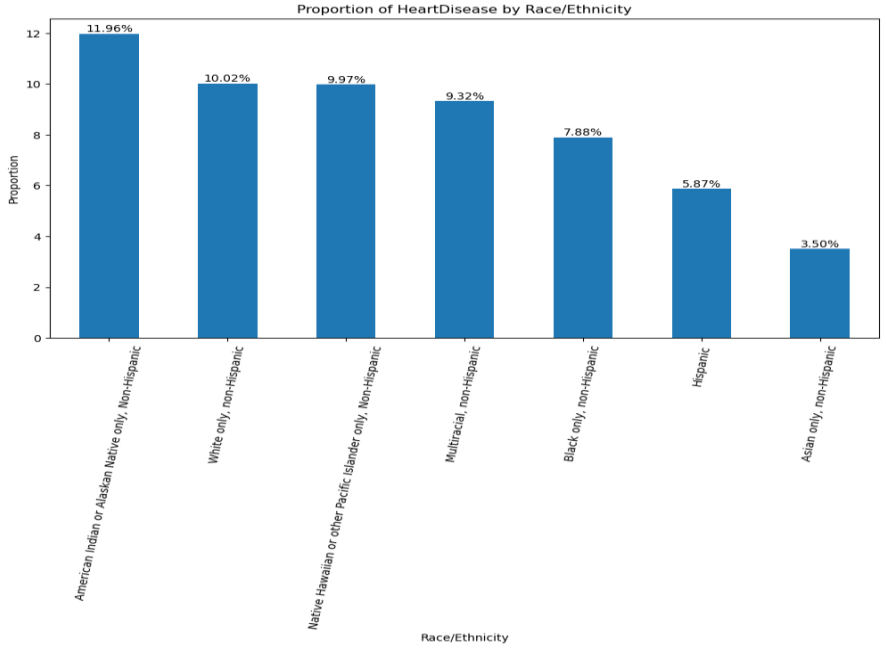



- According to the American Heart Association, It is well-documented that men are generally more prone to developing heart disease compared to women. Research shows that men often develop heart disease about 10 years earlier than women.



# EDA - Demographics



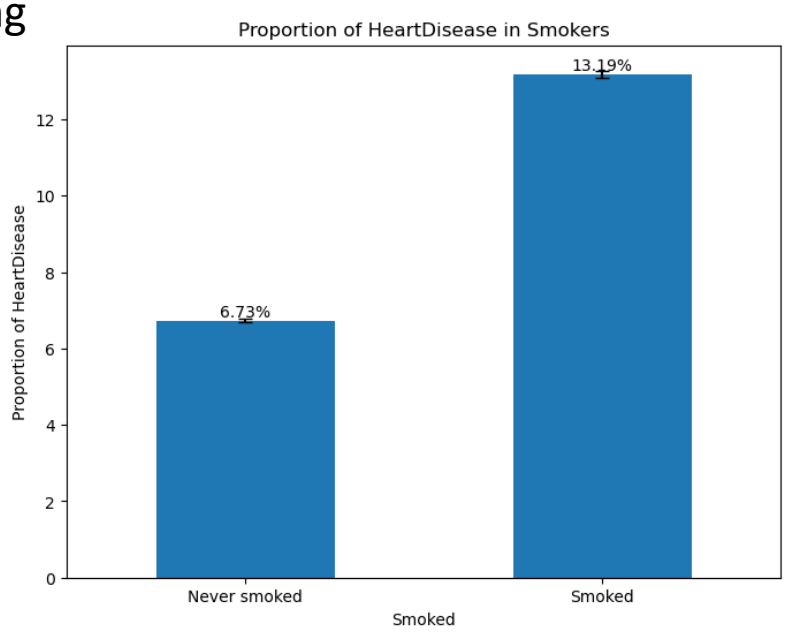




It is evident that as we age, the incidence of heart disease increases. It goes to say that our bodies become more susceptible to sickness and other factors that can cause heart problems.

No outstanding patterns or significant disparities were observed. The analysis suggests that heart disease prevalence does not vary markedly across different racial and ethnic groups within this dataset.

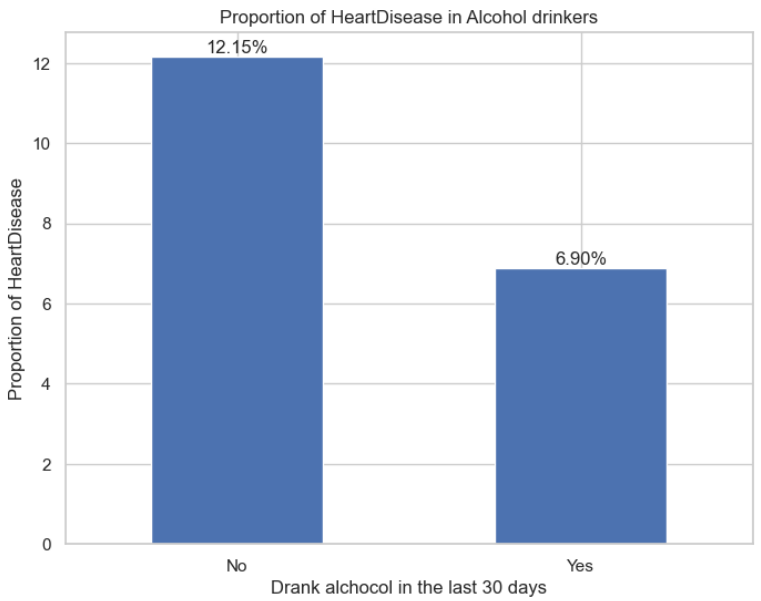
# EDA – Risk factors

Smoking




Among individuals who smoked, 13.19% were found to have heart disease, while among those who never smoked, the prevalence of heart disease was 6.73%

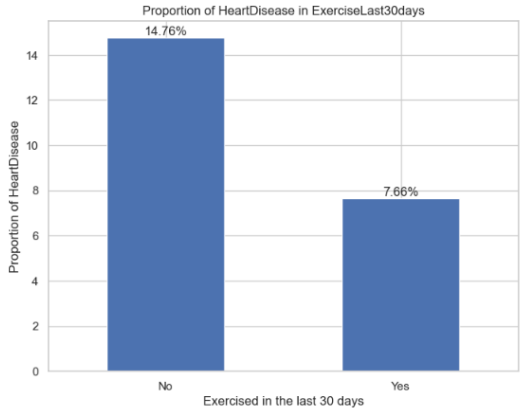
Alcohol




Among individuals who drank in the last 30 days, 6.90% were found to have heart disease, while among those who did not drink in the last 30 days, the prevalence of heart disease was 12.15%

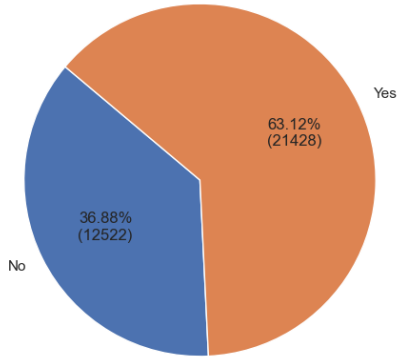
# EDA - Lifestyle

Another important lifestyle factor to consider is exercise. Among individuals who did not exercise in the last 30 days, 14.76% were found to have heart disease. In contrast, among those who did exercise in the last 30 days, the prevalence of heart disease was 7.66%

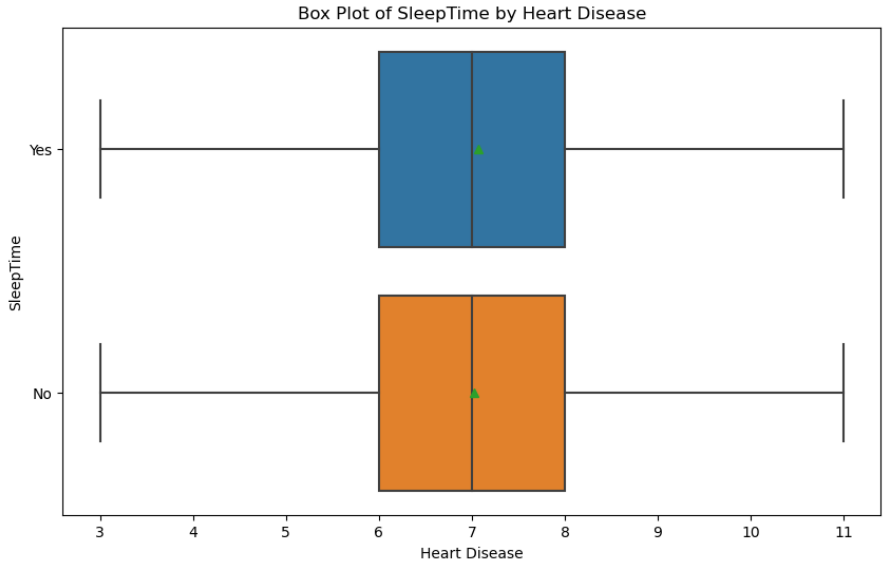



Percentage of Individuals with HeartDisease by ExerciseLast30days

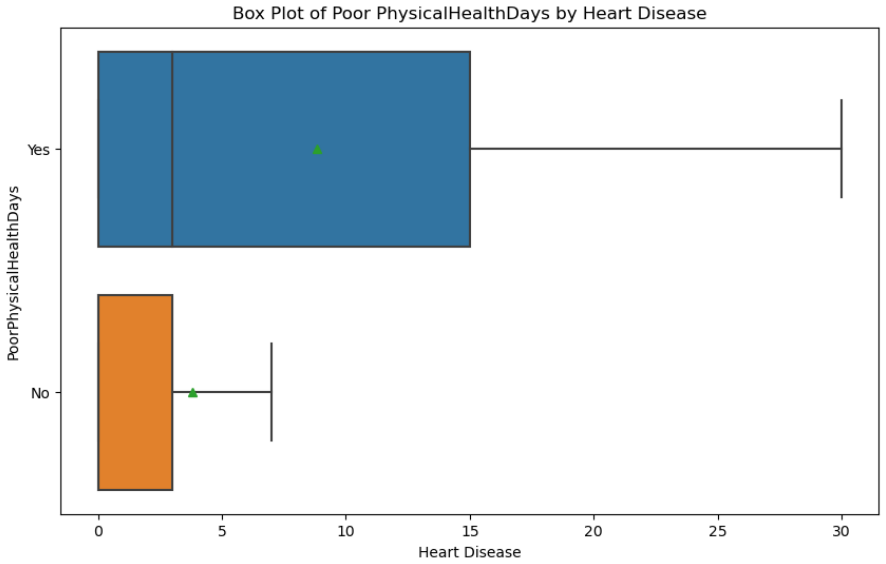
However, when focusing only on individuals with heart disease, a significant proportion, 63.12%, reported exercising in the last 30 days, while 36.88% did not. This means that while exercising may have a beneficial impact on health, other variables and factors may also contribute to the development of heart disease



# EDA - Lifestyle



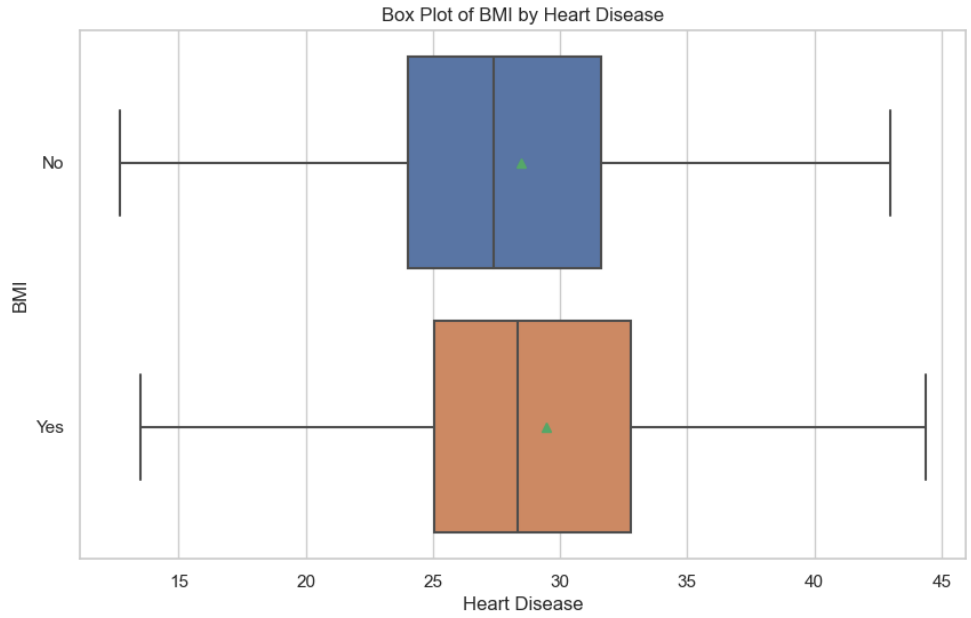

When observing the impact of sleep duration on heart disease, the data shows that there isn't much difference between the duration of sleep between individuals with and without heart disease




The analysis of poor physical health days, which refers to the number of days individuals feel physically weaker than usual, shows a significant difference between those with heart disease and those without. Individuals with heart disease have a higher average of these poor physical health days, with a mean of about 8.92 days, compared to approximately 3.87 days for those without heart disease.

# EDA - Lifestyle

The analysis of BMI data shows differences between people with and without heart disease. On average, people with heart disease have a higher BMI (29.15) compared to those without heart disease (28.22)



# Correlation Matrix

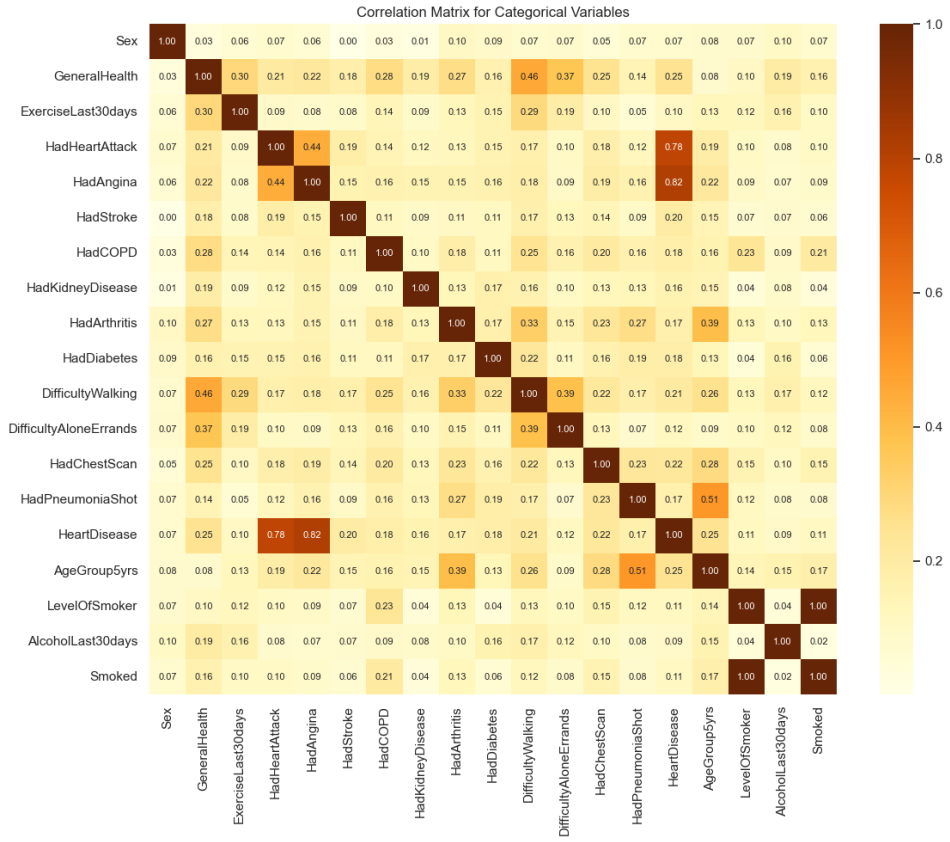
**HadAngina**: The correlation between having angina and heart disease is even stronger (0.82). This suggests that angina is highly associated with heart disease.

**HadHeartAttack**: There is a strong positive correlation between having had a heart attack and heart disease (0.78). This indicates a strong association where individuals who have experienced a heart attack are more likely to have heart disease.

**GeneralHealth**: This variable shows a moderate positive correlation with heart disease (0.25). Individuals with poorer general health are more likely to have heart disease.

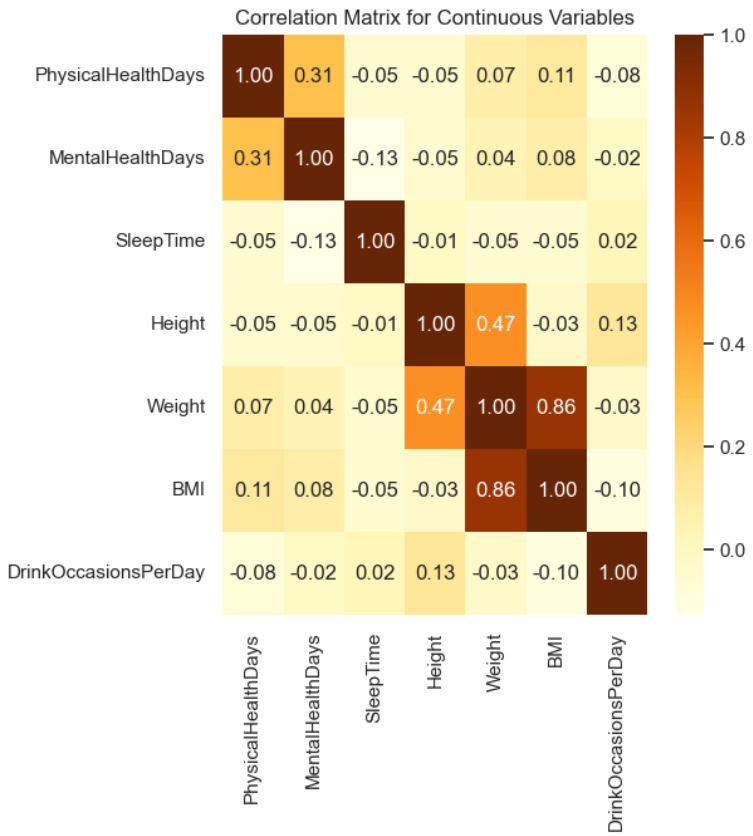
**ExerciseLast30days**: The correlation with heart disease is low (0.10), indicating that exercise has a weaker association with heart disease in this dataset. However, exercise is generally a known factor for reducing heart disease risk as mentioned before.

**Smoked**: The correlation between smoking and heart disease is lower (0.11), suggesting that smoking has a modest association with heart disease in this dataset, though smoking is a well-kjnown risk factor.

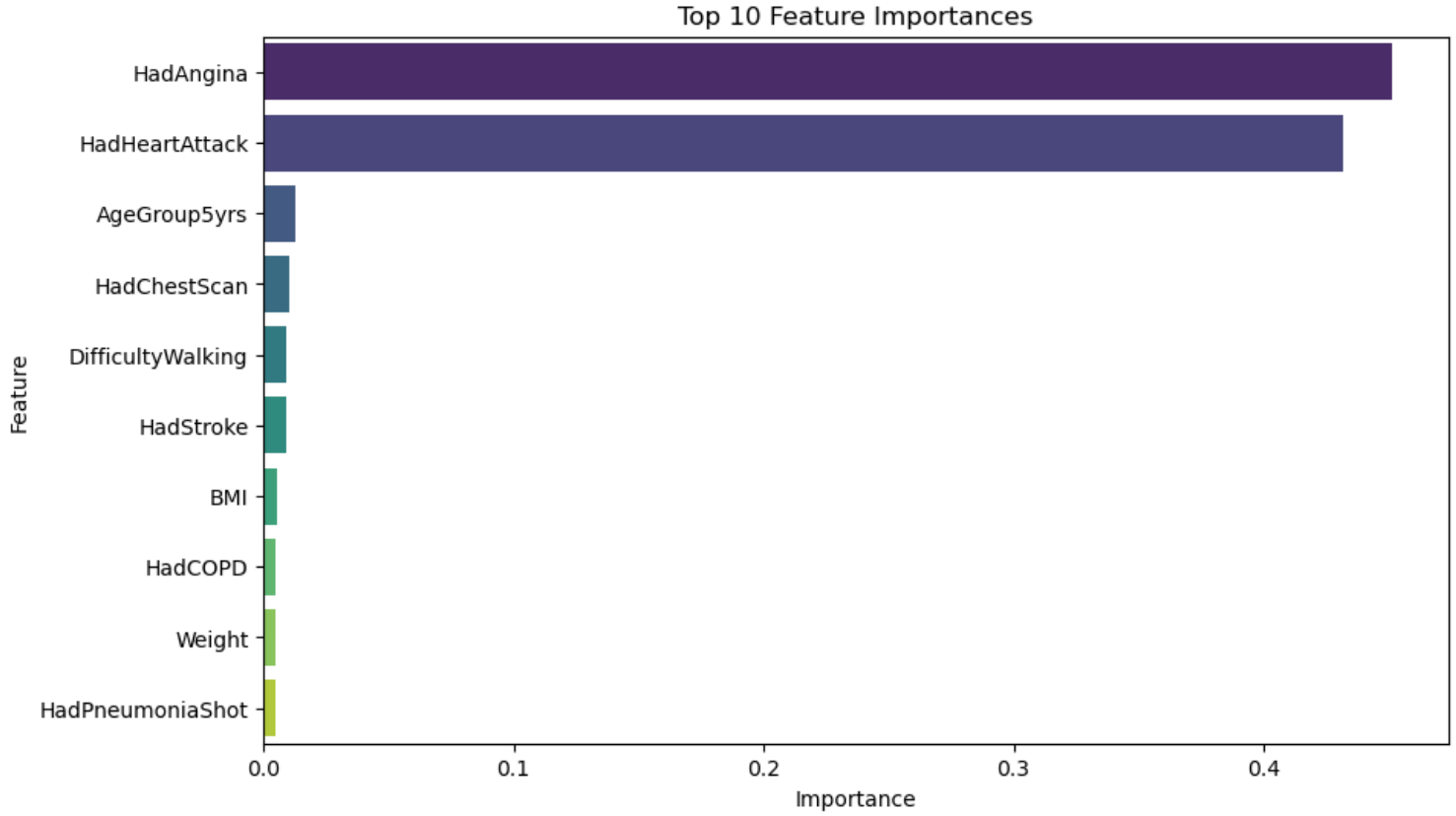


# Correlation Matrix

the continuous variable matrix indicates that while there are notable correlations between certain health-related variables (e.g., physical and mental health days), many of the associations involving sleep time are weak. The strong correlation between weight and BMI is consistent with their mathematical relationship.



# Preprocessing



# Learning models

**Random forest**: An ensemble learning model that is used for its ability to handle large datasets with high dimensionality. Operates by constructing multiple decision trees, hence being called a ‘Forest’, during training and producing mode or mean predictions of each tree.

**Logistic Regression:** This learning model can be very effective for binary classification tasks, in this case whether one has heart disease or not. Because of its simplicity it serves as a reliable baseline model to compare against other models. It does not do that well with large datasets and class imbalance, so parameters such as ridge or lasso will be implemented.

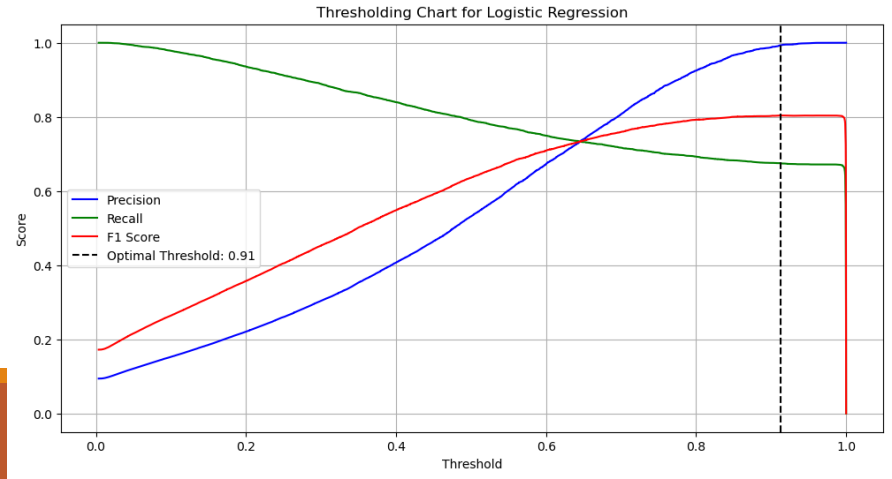
**XGBoost**: A gradient boosting algorithm and model, that combines several weak prediction models to produce a strong overall model. Its ability to capture complex patterns and interactions in data makes it particularly valuable when dealing with something as complex and multifaceted as heart disease.

# Logistic Regression

Ending up providing us with the largest recall score of 0.79

| | Metric | RF | Logistic Regression | XGBoost |
|---|---|---|---|---|
| 0 | Accuracy | 0.967740 | 0.914662 | 0.943300 |
| 1 | Precision | 0.970965 | 0.531597 | 0.687281 |
| 2 | Recall | 0.677794 | 0.791058 | 0.730402 |
| 3 | F1 Score | 0.798315 | 0.635879 | 0.708186 |
| 4 | F-Beta (Beta=2) | 0.721355 | 0.720706 | 0.721351 |
| 5 | AUC | 0.936861 | 0.936976 | 0.932811 |

But for our best model, the target of .85 recall seems impossible when looking at the optimal threshold of the precision-recall curve



# Conclusions

Moving forward, the highest recall score achieved was 0.79, falling short of our target of 0.85. To improve the model's effectiveness, one major issue was the size of the dataset, which made feature selection challenging.

With over 300 columns, the evaluation of feature importance was constrained by lack of resources. There could have been features that were removed haphazardly. This was the same issue with hyperparameters, running grid searches took a toll on my computer.

Exploring alternative methods, such as Bayesian optimization, could have alleviated some of these problems and potentially improved model performance.. Additionally, exploring alternative machine learning algorithms or ensemble methods could offer better performance and help achieve the desired recall score.