Introduction

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. Heart disease is a formidable global health challenge, affecting the heart and causing serious health issues. Its multifaceted nature demands complex strategies for prevention and early detection. Despite significant medical advancement, heart disease still plagues the U.S. as the leading cause of death. Its prevalence is composed of an array of risk factors, including diabetes, obesity, smoking, and sedentary lifestyles.

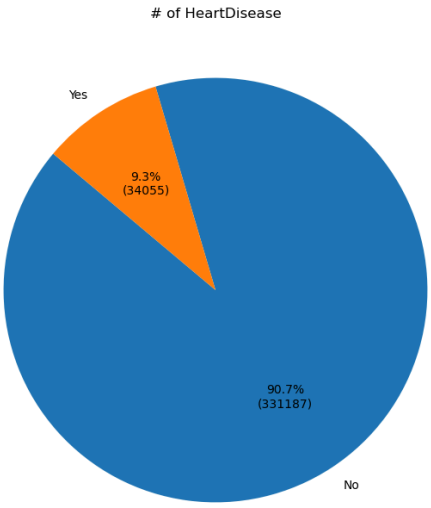
Patients , health care providers, researchers, healthcare systems and insurance would benefit from the prevention and detection of this persistent health issue. Tackling heart disease requires innovative approaches to analyze vast amounts of health data. By leveraging machine learning and ai, researchers can identify patterns and correlations that might otherwise go unnoticed. Patient and healthcare providers would benefit significantly from the use of data science to combat heart disease; Data-driven insights can lead to more accurate diagnosis and tailor treatments. Healthcare systems and insurance companies can optimize resource allocations and streamline operations by analyzing patient data and thus develop more accurate risk assessment models.

In what follows, I will build 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.The data were retrieved from CDC's 2022 Behavioral Risk Factor Surveillance System survey data (BRFSS). This survey aims to collect prevalence data among U.S. residents regarding risk behaviors and preventive health practices. The original dataset from the CDC consisted of 328 columns and approximately 450,000 records. A threshold of 80% missing data was set, leading to the removal of 213 columns where more than 80% of the data was missing. Additionally, 74 columns irrelevant to the analysis were removed based on the data dictionary, including IMONTH (month of the survey), INCOME (income bracket), and EMPLOY (employment status). Following encoding of the remaining columns. After purging most NA values, the final dataset comprised 41 columns and around 365,000 records. Data transformation involved remapping encoded values to their corresponding answers according to the data dictionary. Furthermore, formatting errors in the Height, Weight, and BMI columns were corrected.

Demographics:

To begin the analysis, it is essential to determine the prevalence of heart disease within the dataset.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.(Fig 1.) This initial step provides a foundational understanding of the dataset's composition and sets the stage for further analysis.

Fig 1.



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. One significant factor is that men are more likely to experience heart attacks throughout their lifespan, with men being roughly twice as likely as women to have a heart attack at some point. Let's investigate if this holds true within our dataset:

Among all females in the dataset (187,074 individuals), 7.32% (13,685 individuals) had heart disease. Among all males in the dataset (178,168 individuals), 11.43% (20,370 individuals) had heart disease. These figures further illustrate the higher prevalence of heart disease among men compared to women within the dataset.(Fig 2.)

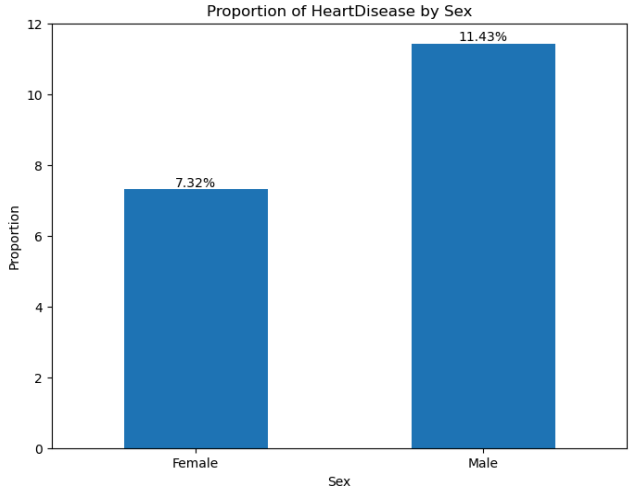
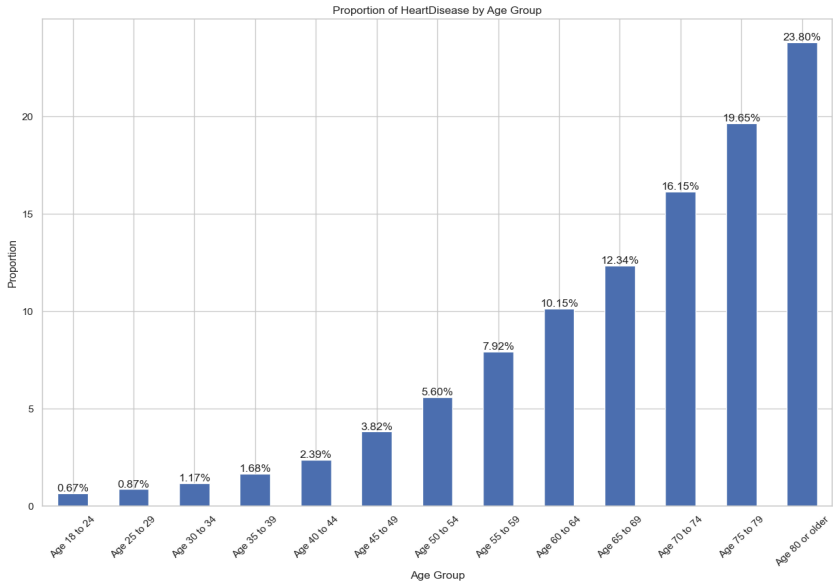


Fig 2

To further understand the dataset, visualizing the distribution of heart disease across different age groups is necessary.

Fig. 3



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.
For example, among respondents aged 80 and older, 23.80% report having heart disease, whereas among those aged 18-24, only 0.67% report the same.(Fig. 3)

In addition, it is valuable to explore the potential connections between heart disease and race/ethnicity(Fig. 4). Visualizing this relationship can provide insights into how heart disease prevalence varies across different racial and ethnic groups within the dataset. Understanding these patterns can help identify at-risk populations and inform targeted prevention and treatment strategies.

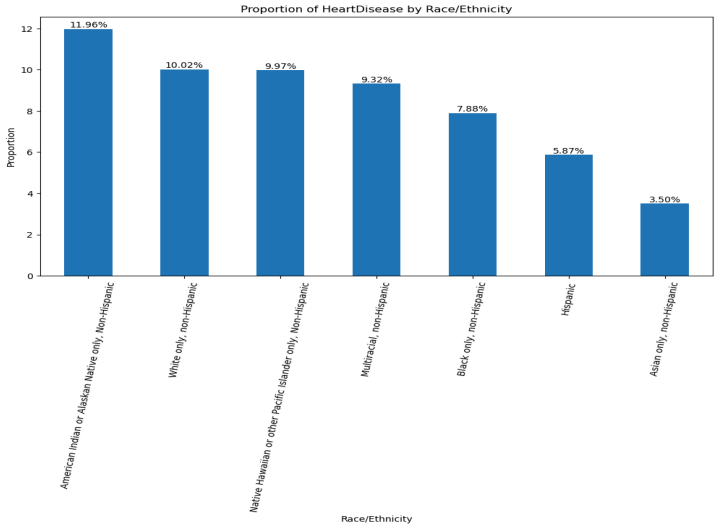


Fig 4

Upon visualizing the relationship between heart disease and race/ethnicity, 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. Following that, I sought to investigate personal habits and behaviors that may be prevalent among individuals with heart disease. This insight indicates that other factors, such as lifestyle or genetic predisposition, might play more prominent roles in influencing heart disease prevalence.

Smoking:

One such lifestyle factor is 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% (Fig. 5). From this observation, it can be inferred that smoking is associated with a higher risk of developing heart disease compared to non-smokers.

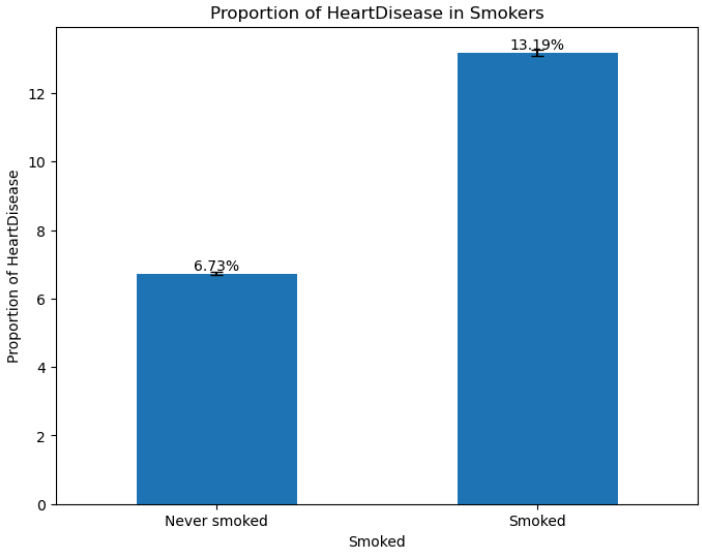


Fig 5.

There is a significant association between smoking and heart disease ($p < 0.05$).

Alcohol:

Another lifestyle factor to consider is alcohol consumption. 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%(Fig. 6). This observation suggests that alcohol consumption may not be a significant contributing factor to the prevalence of heart disease in this dataset.

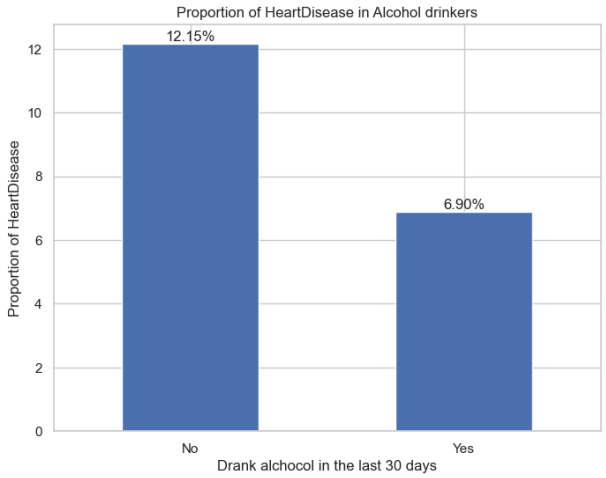


Fig. 6

Exercise:

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% (Fig. 7). This suggests that regular exercise may indeed be linked to a reduced likelihood of heart disease among the general population

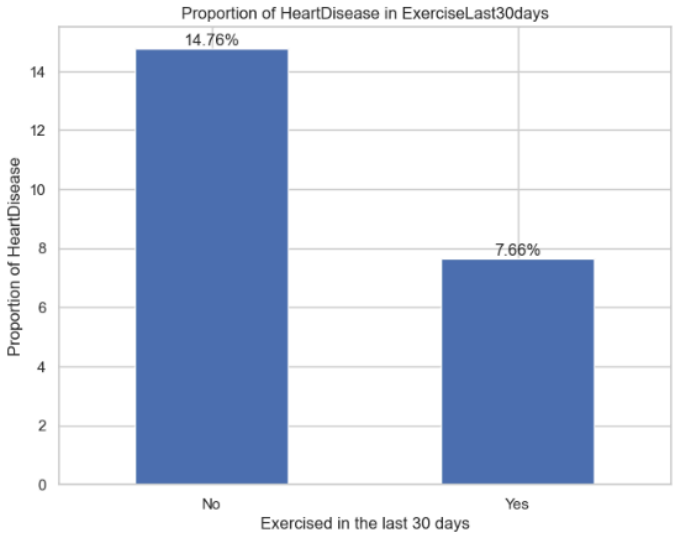


Fig. 7.

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 (Fig. 8). 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. Those who partake in a more active lifestyle remain at risk of developing heart disease due to its multifaceted nature, as mentioned before where a combination of genetic, environmental, and lifestyle factors play a role in its onset and development.

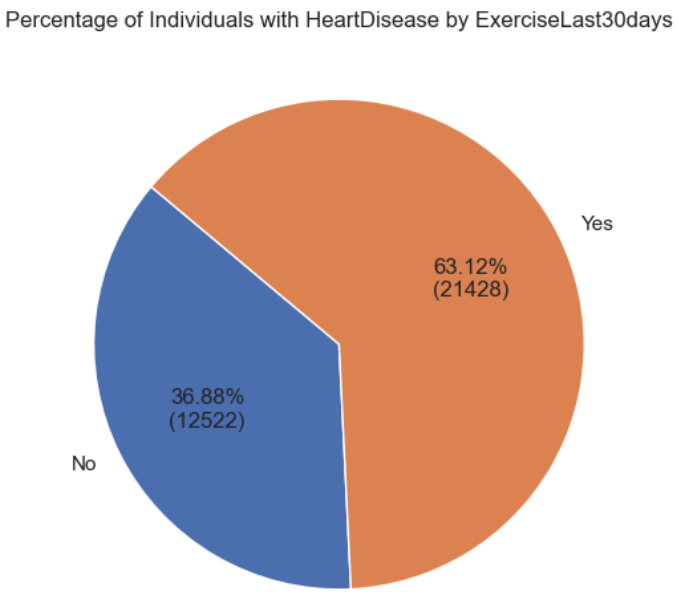


Fig. 8

Sleep:

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 (Fig. 9). This suggests that sleep duration may not be a prominent factor influencing the prevalence of heart disease within this dataset.

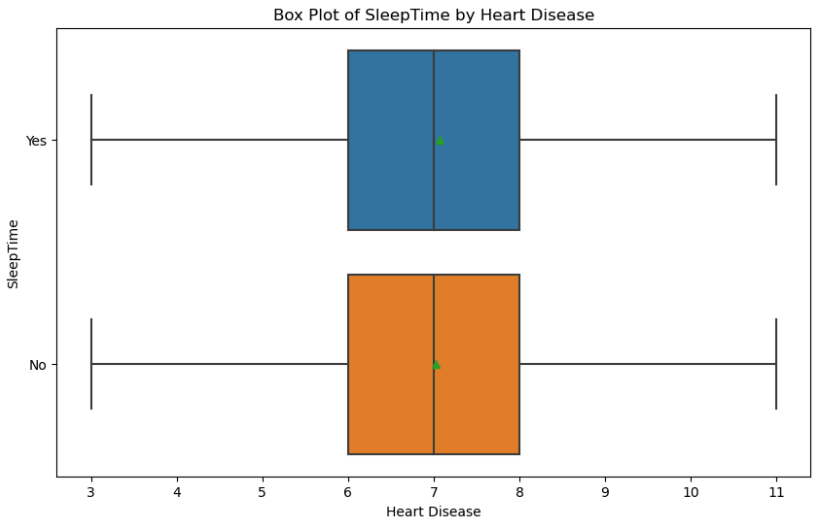
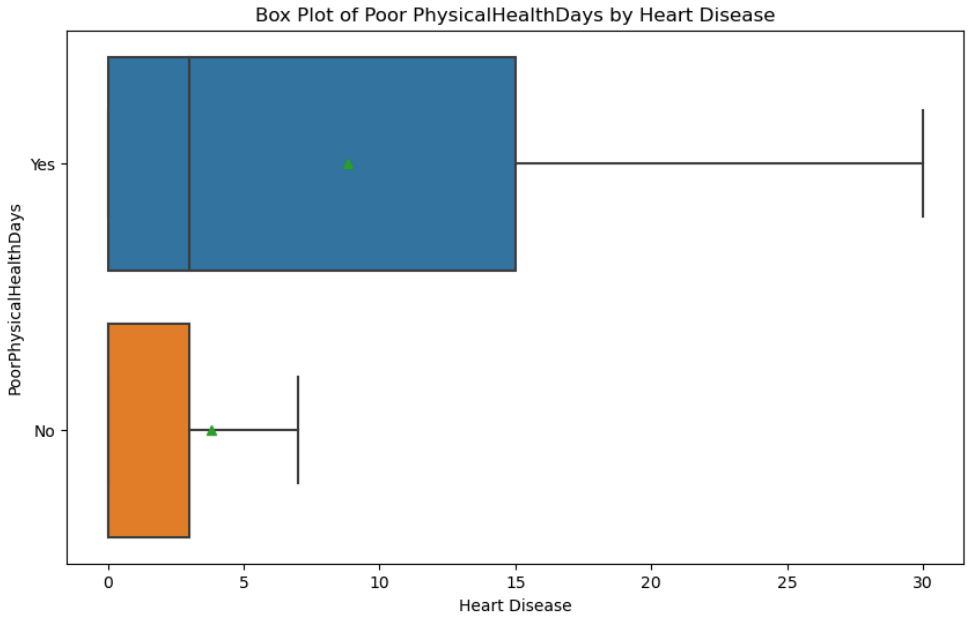


Fig. 9

Poor Physical Health Days:

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.

Additionally, the variability in poor physical health days is greater among individuals with heart disease, which is seen with the larger standard deviation of 11.67 days compared to 8.03 days for those without heart disease. The quartile values, including the 25th, 50th, and 75th percentiles, further reveal that individuals with heart disease consistently experience more poor physical health days across. This suggests that poor physical health is more prevalent and variable among individuals with heart disease, highlighting the impact of the condition on overall physical well-being. Fig. 10



BMI

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). The range of BMI values are similar for both groups as well as their standard deviations. Furthermore, BMI values are consistently higher at the 25th, 50th, and 75th percentiles for individuals with heart disease. These observations suggest a potential relationship between higher BMI and the presence of heart disease.

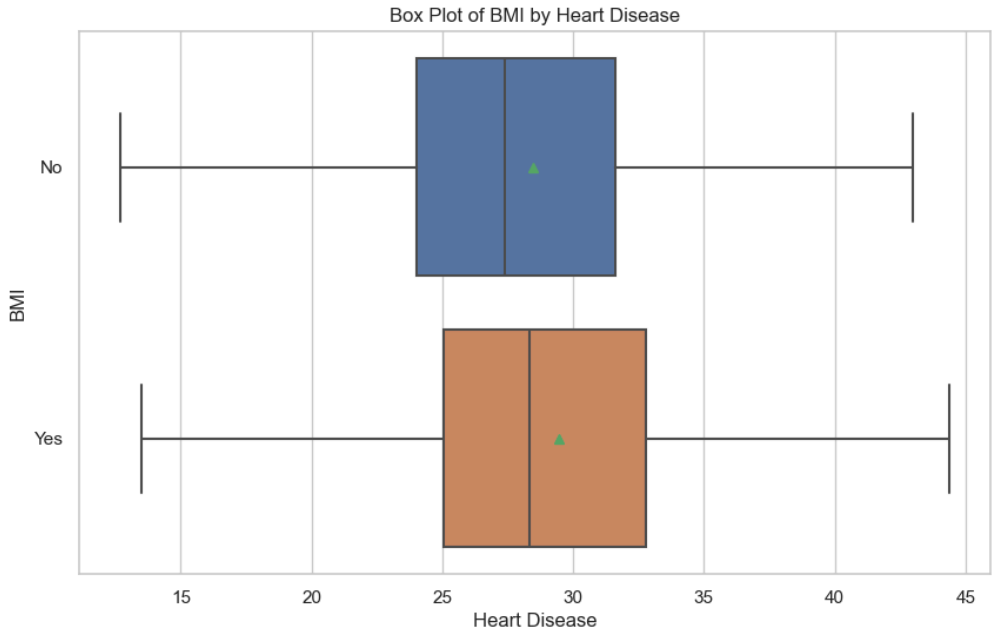
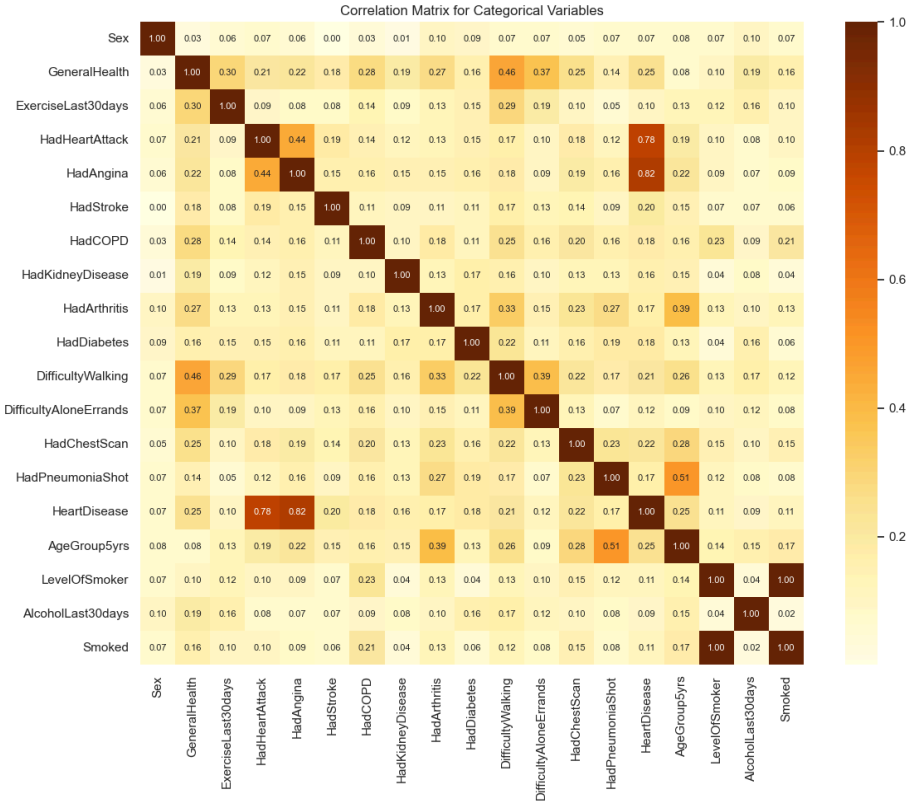


Fig. 11

Correlation matrix:

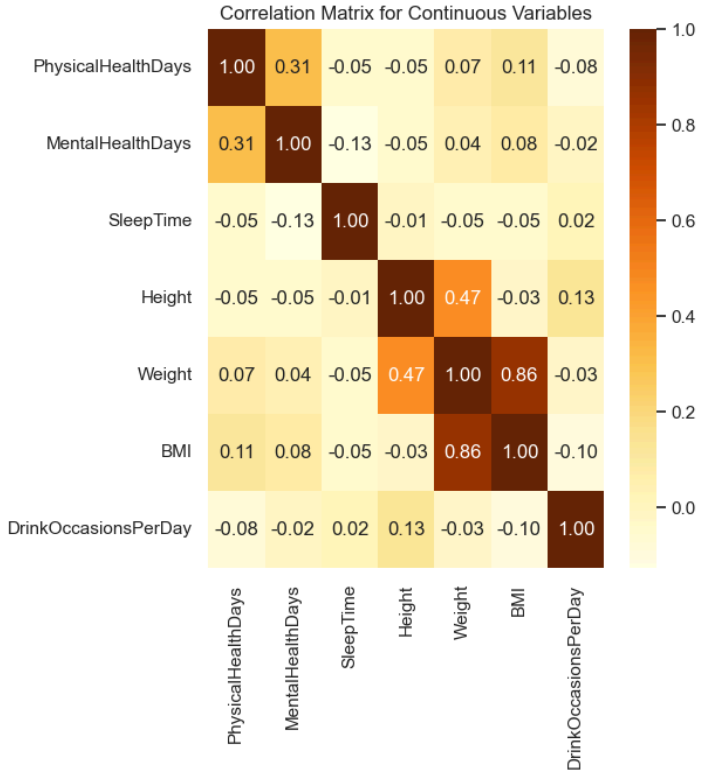
Cramer's V correlation was used to evaluate the strength of associations in categorical values. To improve clarity, columns exhibiting minimal correlation with heart disease were excluded from the analysis.(Fig. 12)



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.

Overall, variables related to past cardiovascular events (e.g., heart attack, angina) show strong correlations with heart disease, reflecting their significant role as risk factors. Lifestyle factors such as exercise and alcohol consumption show weaker correlations, indicating that while they may impact heart disease, other factors may play a more dominant role.(Fig. 12)

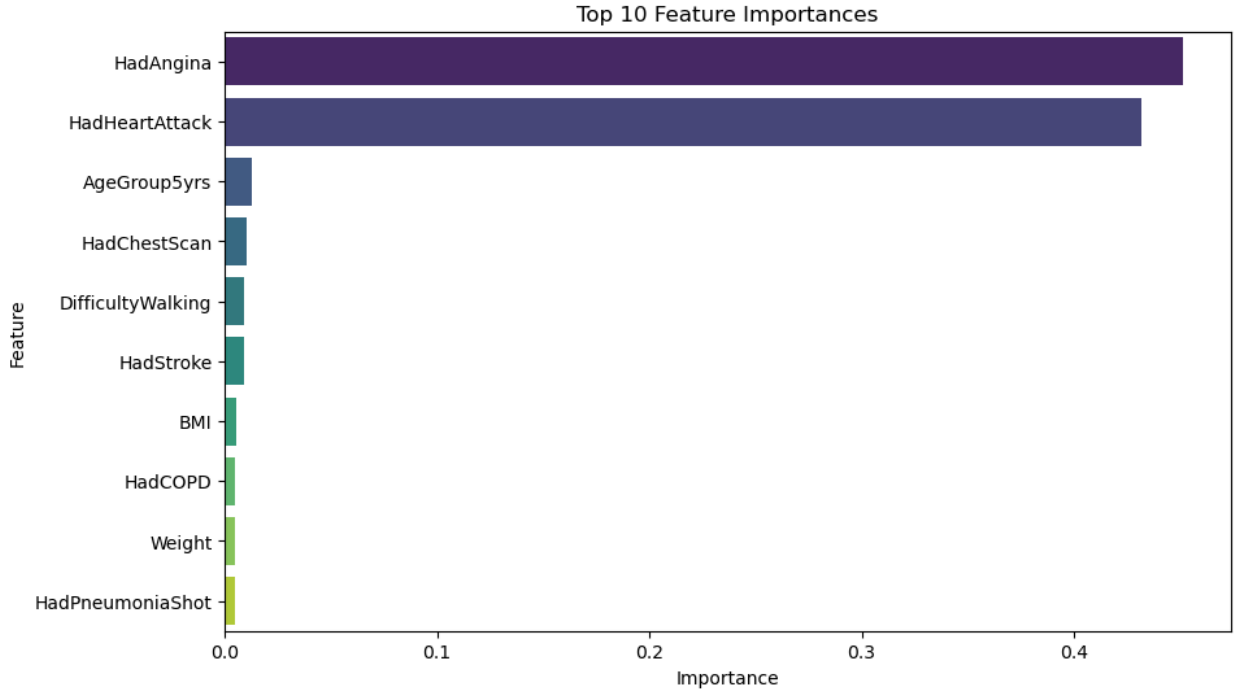
Fig.13



In summary, 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.(Fig 13)

Feature importance:

Fig. 14



Feature selection:

HadAngina and HadHeartAttack were both strong predictors of heart disease. To refine the model and gain more meaningful insights, it was decided to use only the HadAngina feature and exclude HadHeartAttack. This choice was made because both Angina and Heart Attack are prevalent indicators of heart disease, and including both in the model led to overly accurate predictions. By excluding HadHeartAttack, the model aims to provide more valuable and insightful analysis.Additionally, the ten least important features were excluded, and Principal Component Analysis (PCA) was subsequently applied to identify the component that retained 95% of the variance, and the remaining features were dropped.

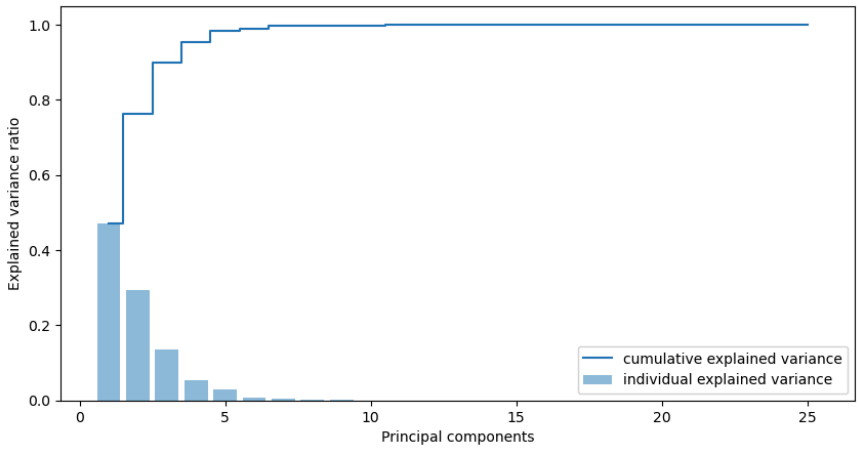


Fig. 15

Leaving us with: State, Sex, GeneralHealth, PhysicalHealthDays, MentalHealthDays, LastCheckup, ExerciseLast30days, SleepTime, HadAngina, HadStroke, HadCOPD, HadKidneyDisease, HadArthritis, HadDiabetes, Deaf, DifficultyWalking, EcigUsage, HadCovid, HeartDisease, RaceEthnicityGroup, AgeGroup5yrs, Height, Weight, BMI, DrinkOccasionsPerDay, Smoked.

Preprocessing:

Categorical data was encoded, and the specified columns were removed. Notably, some null values remained in the categorical data, which were then imputed using the most frequent value. Additionally, Synthetic Minority Over-sampling Technique (SMOTE) was utilized to address class imbalance within the dataset.

In addition, by running models with different scalers we found which scaler is best to use for each model.
MaxAbsScaler for RandomForest (Fig. 16)

Scaler	RF_accuracy	RF_recall	RF_f1_score	RF_beta_2_score	RF_precision	RF_confusion_matrix	RF_ROC AUC Score
StandardScaler	0.968554	0.836423	0.801418	0.719516	0.989059	[[64474, 50], [2190, 4520]]	0.928531
MinMaxScaler	0.968386	0.836397	0.800602	0.719355	0.986257	[[64461, 63], [2189, 4521]]	0.928738
RobustScaler	0.968919	0.836157	0.803025	0.719318	0.996247	[[64507, 17], [2197, 4513]]	0.927075
MaxAbsScaler	0.968414	0.836746	0.80092	0.719967	0.985627	[[64458, 66], [2184, 4526]]	0.929166

MinMaxScaler for Logistic Regression and XGBoost (Fig. 17)

Scaler	LogReg_accuracy	LogReg_recall	LogReg_f1_score	LogReg_beta_2_score	LogReg_precision	LogReg_confusion_matrix	LogReg_ROC AUC Score
StandardScaler	0.912486	0.859421	0.630906	0.719611	0.523379	[[59672, 4852], [1382, 5328]]	0.937138
MinMaxScaler	0.914563	0.859566	0.635831	0.721052	0.531194	[[59835, 4689], [1397, 5313]]	0.936927
RobustScaler	0.935129	0.855896	0.687707	0.728376	0.629158	[[61525, 2999], [1622, 5088]]	0.935501
MaxAbsScaler	0.914128	0.859059	0.634479	0.72006	0.529576	[[59808, 4716], [1401, 5309]]	0.936945

Scaler	XGBoost_accuracy	XGBoost_recall	XGBoost_f1_score	XGBoost_beta_2_score	XGBoost_precision	XGBoost_confusion_matrix	XGBoost_ROC AUC Score
StandardScaler	0.968793	0.836488	0.802593	0.719792	0.992969	[[64492, 32], [2191, 4519]]	0.936987
MinMaxScaler	0.968849	0.836786	0.803018	0.720338	0.992975	[[64492, 32], [2187, 4523]]	0.937343
RobustScaler	0.968975	0.836455	0.803451	0.719864	0.996251	[[64507, 17], [2193, 4517]]	0.936823
MaxAbsScaler	0.968849	0.836519	0.802878	0.719884	0.993842	[[64496, 28], [2191, 4519]]	0.937685

In the context of predicting health conditions, recall often takes precedence over accuracy. Recall emphasizes the identification of true positive cases (patients who actually have the condition being tested for). Ensuring high recall helps identify all potential candidates for further testing or treatment, ensuring that no promising treatments or necessary interventions are overlooked. Missing a true positive case (a false negative) can have significant consequences, such as patients being denied potentially life-saving treatments or experiencing worsened health due to delayed intervention. . While accuracy provides a general measure of correctness, it does not differentiate between the types of errors that could happen. Therefore we use the Beta-2 score, which weights recall more heavily than precision, to better capture the importance of finding every true positive case in our evaluations.

Modeling:

Developing a predictive model to address the need for early detection and prevention of heart disease. By leveraging survey data and machine learning, the aims to identify individuals at higher risk. This approach seeks to reduce the incidence of heart disease and improve overall health outcomes by providing actionable insights derived from comprehensive survey data.

I've decided to use Random Forest (RF), Logistic regression and XGBoost.

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. Other things to note about this model is that it is good against overfitting as well as providing a reliable measure of feature importance. Which helps in identifying comorbidities or risk factors are most influential in predicting heart disease.

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. The issue with class imbalance will be addressed by combining the model with SMOTE, an oversampling technique.

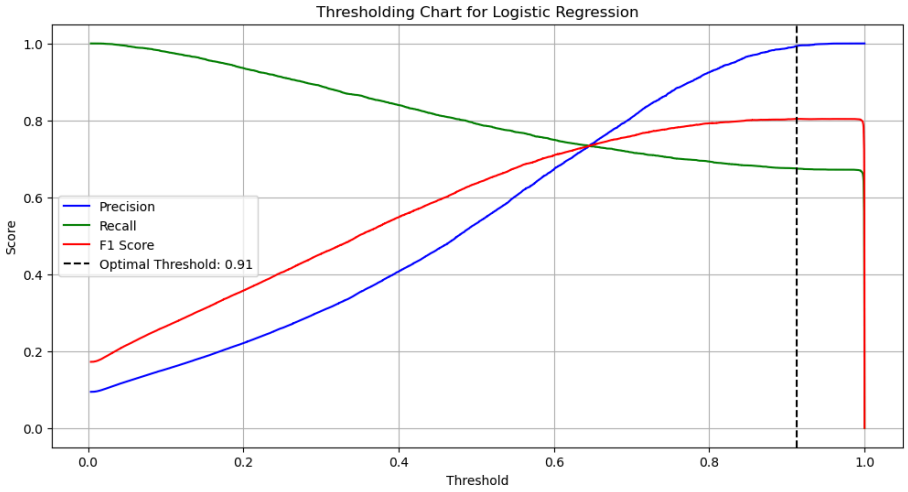
XGBoost: A gradient boosting algorithm and model, that combines several weak prediction models to produce a strong overall model.Its ability to capture complex patterns and interactions in data makes it particularly valuable when dealing with something as complex and multifaceted as heart disease.

HyperParameter and tuning:

For RandomForest and Logistic Regression, we ran grid searchs to find the best parameters. And RandomSearchCV for XGBoost. (Fig.18)

	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

Logistic Regression had the best recall and AUC, indicating it might be better for identifying more true positives but with lower precision, which may result in more false positives. Overall all three models have failed to pass the recall of .85. (Fig. 19)



Looking at the threshold chart, the trade off for recall being .85 is too severe to make this model worthwhile.

Conclusion

The major findings from the exploratory data analysis and predictive modeling of heart disease incidence allows us to further enhance our knowledge of this area. Firstly, certain comorbidities and risk factors, such as poor physical health days, higher BMI, and the presence of conditions like angina, are strongly associated with heart disease. The data also revealed that individuals with heart disease exhibit greater variability in poor physical health days and consistently higher BMI across different percentiles. Interestingly, a notable portion of individuals with heart disease reported maintaining an active lifestyle, suggesting that heart disease is influenced by a complex interplay of genetic, environmental, and lifestyle factors.

Also found that addressing class imbalance is also an important part of the process for effectiveness and reliability of a predictive model. When the dataset has a significant imbalance between patients with and without heart disease, the model can become biased towards the majority class, leading to poor performance in identifying the minority class. This is especially problematic in medical settings, where high recall is essential to ensure that no patients with heart disease are missed (false positives). The handling of missing values and class imbalance was addressed through imputation with the most frequent values and the application of SMOTE. One other thing is the impact of different scalers on model performances. Scalers such as StandardScaler, MinMaxScaler, RobustScaler, and MaxAbsScaler were tested, revealing that the choice of scaler can significantly influence model outcomes.

And finally, In evaluating predictive model performances, Logistic Regression was the most effective model. In the context of heart disease prediction, minimizing false negatives are crucial. Recall score is more sought after, although we did not hit our target score of 0.85, logistic regression still had the highest recall of 0.79. Although Random Forest showed superior performance in terms of accuracy (0.97), precision (0.97), and F1 score (0.80), its overall effectiveness is tempered by its lower recall (0.67) compared to Logistic Regression.

These findings align with the project's objective to enhance early detection and preventative interventions for heart disease through the analysis of survey data. By identifying key features and understanding their relationships with heart disease, healthcare professionals can better target at-risk populations and design more effective prevention strategies. The insights gained from this project highlight the importance of considering multiple factors when addressing heart disease.

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.