CMPT 353 (Computational Data Science) Project

Predicting Heart Disease

December 04, 2023

Vraj Patel

Abhay Dhiman

Simon Fraser University

Background

As data science students, we're focusing our project on analyzing heart health data, a critical issue since heart disease is a leading cause of death globally. We aim to explore various factors, from lifestyle choices to genetic factors, to understand what contributes to heart disease. By examining a comprehensive dataset that includes health indicators and lifestyle habits, our goal is to uncover the key predictors of heart disease and apply our data skills to a real-world challenge. This project is not only crucial for our course but also a significant step as we transition into our professional data science careers.

1. Project Objective

Our project is centered on exploring the intricate link between various health indicators and heart disease risk. We aim to leverage machine learning tools to predict heart disease likelihood in individuals, analyzing factors like BMI, lifestyle habits, and existing health conditions. Through this analysis, we aspire to uncover insights that can shape healthcare approaches and personal health decisions. Through this analysis, we hope to reveal insights that can influence healthcare practices and individual health choices.

2. The Data

2.1 Data Gathering

When looking for a database, our main focus was to find one with simple variables and a lot ofrow entries. The main reason behind this was that a disease prediction based on simple daily-life factors can give better results compared to data which is limited (sources from lab research).

Our database contains 18 factors which are described below. The data was originally collected by CDC from a survey of over 400000 individuals with over 300 variables [6]. We then used the data variables that were specific to our model purpose.

2.2 Data Description

Variables	Description
HeartDisease	Indicates if the individual has heart disease (Yes/No)
BMI	Body Mass Index, a measure of body fat based on height and weight
Smoking	Indicates if the individual smokes (Yes/No)
AlcoholDrinking	Indicates of the individual drinks alcohol (Yes/No)
Stroke	Indicates if the individual has had a stroke (Yes/No)
PhysicalHealth	Days in the past 30 days with bad physical health
MentalHealth	Days in the past 30 days with bad mental health
DiffWalking	Indicates if the individual has difficulty walking (Yes/No)
Sex	Gender of the individual
AgeCategory	Age category of the individual (e.g., '55-59', '80 or older')
Race	Race of the individual (e.g., White, Black)
Diabetic	Indicates if the individual is diabetic (Yes/No/Borderline)
PhysicalActivity	Indicates if the individual engages in physical activity (Yes/No)
GenHealth	General health condition of the individual (e.g., 'Very good', 'Fair')
SleepTime	Average number of hours of sleep per night
Asthma	Indicates if the individual has asthma (Yes/No)
KidneyDisease	Indicates if the individual has kidney disease (Yes/No)
SkinCancer	Indicates if the individual has skin cancer (Yes/No)

2.3 Data Cleaning

In order to properly analyze our data, we needed to format each data entry in a uniform manner. Hence, we took the following steps:

- The first task for cleaning was to convert data from Yes/No to (0/1). Almost every column had either true or false values which were converted easily. Only issue was Diabetic column, where we had to consider a few cases of conditional diabetes i.e. borderline diabetes and diabetes during pregnancy.
- After visualizing through use of box plots to identify outliers, BMI values which are greater than 80 were removed, as in real sense, they are above normal body range and would be the result of some other serious health issue which would itself lead to heart disease, making those BMI values unreliable for heart disease.
- PhysicalHealth, MentalHealth, AgeCategory, GenHealth were distributed well, hence did not need any filtering. We analyzed this through box plots and histograms.

They all did have a cluster point, but the "outliers" in the plot were reasonable to include in prediction.

- The range of SleepTime was 1 to 24. Hence data was cleaned out by removing individuals with sleep time which was greater than 12 hours. As in real sense, that also lies above normal sleeping time for an individual. That amount of sleep hours is normal for infants and babies of age 2-3 yrs, and for them this prediction model would not be accurate due to lack of other factors. Also, values under 3 hours were also removed, as they would also point to some other disease.
- Based on initial data, we had 319795 rows of entry (individuals). After removing unrealistic data entries, we had a total of 33985 rows for our Data Analysis.

3. Exploratory Data Analysis

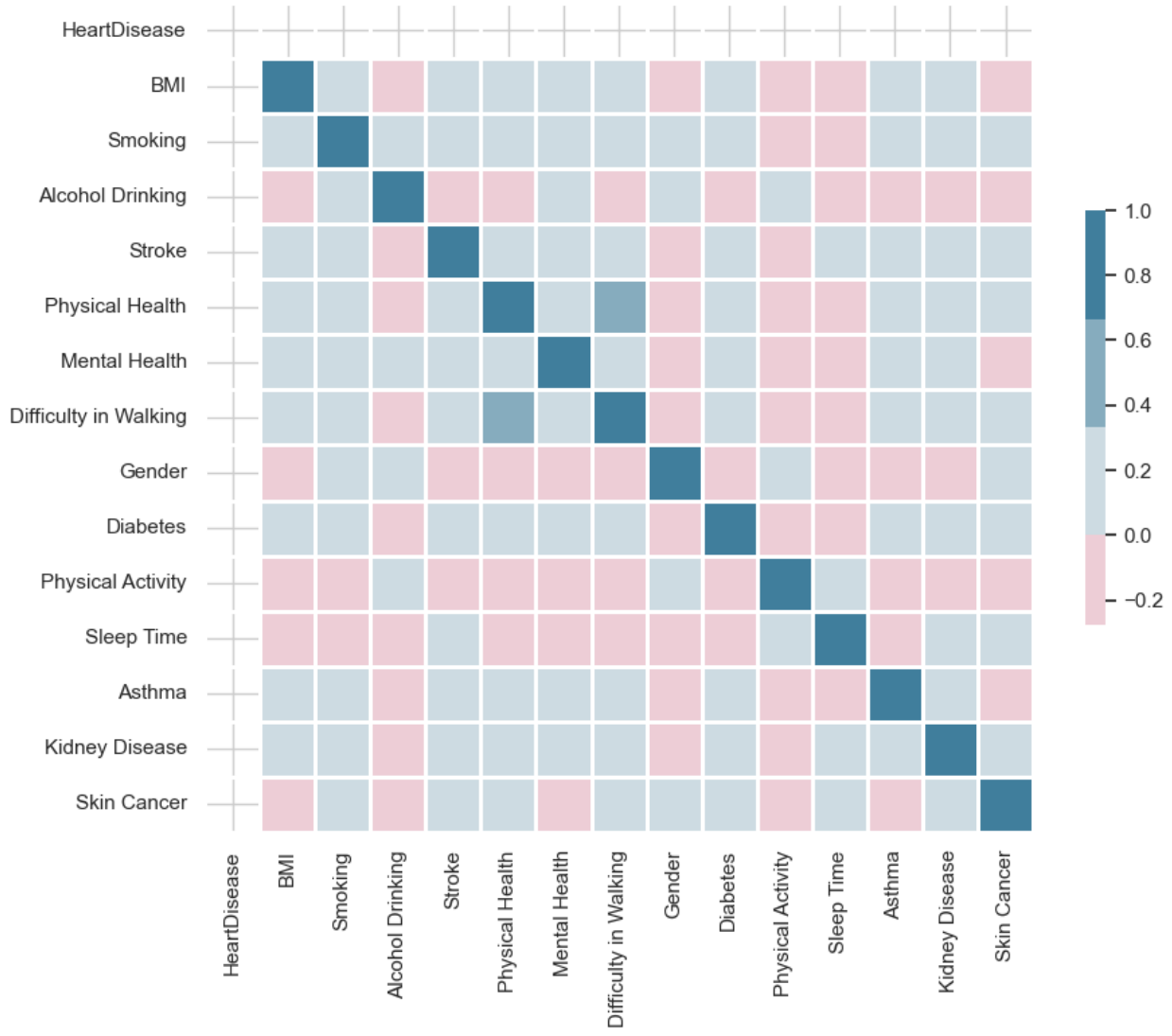

Figure 1

We produced the heat-map (Figure 1) in our analysis to observe multicollinearity among selected health variables (i.e. high correlation between multiple independent variables [3]) among selected health variables, 'Heart Disease', 'BMI', 'Physical Health', 'Mental Health', 'Sleep Time', and 'Diabetic'. We needed to test this relationship amongst our variables to help us select the variables to use for our regression model. This is because we would not be able to distinguish between the individual effects of the independent variables on the dependent variables, hence this would compromise the reliability of our regression model. From this heat-map, we found:

- There is a moderate positive correlation between the presence of heart disease and Physical Health
- There is positive correlation with Diabetic variable, indicating a potential link between heart disease and diabetes.
- There is also a moderate positive correlation between Diabetes and Physical Health. This suggests diabetes also affects overall physical health.
- There is a moderate positive correlation between Physical Health and Mental Health, suggesting a strong connection between the two.
- There is a negative correlation between Sleep Time and other variables.
- There is a positive correlation and quite weak between heart disease and BMI.

BMI

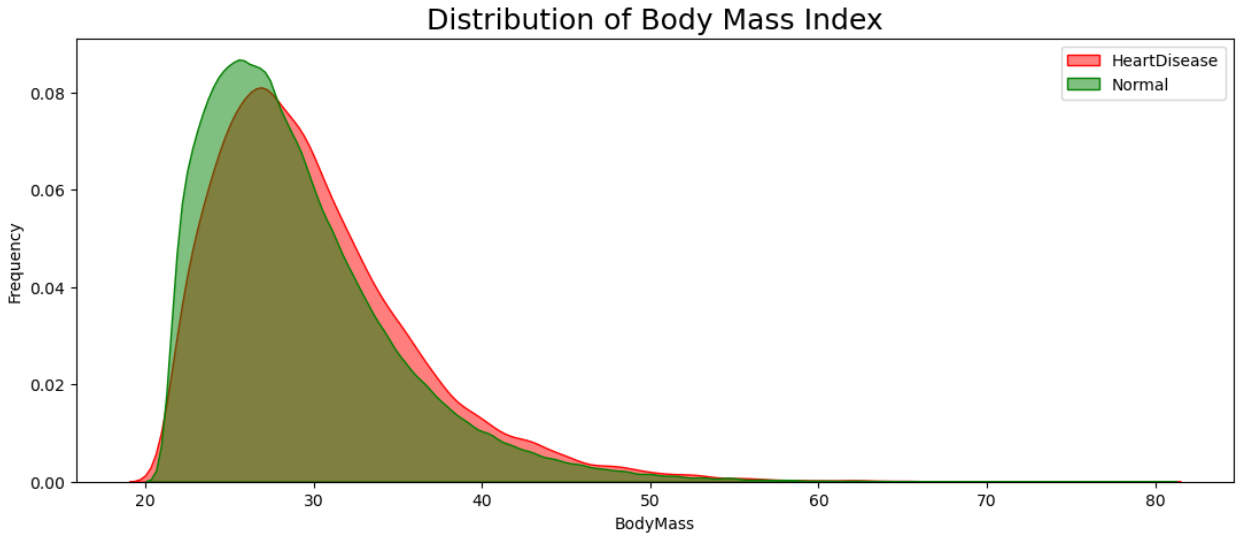

Figure 2

Figure 2 illustrates the distribution of Body Mass Index (BMI) among individuals categorized by their heart disease status. From this plot, we observe that both groups – those with heart disease (shown in red) and without heart disease (shown in green) – have similar BMI distributions. However, the group with heart disease shows a slightly higher frequency in the higher BMI range, indicating a potential association between higher BMI and the prevalence of heart disease. The overlap in the distributions suggests that while BMI might be a factor in heart disease risk, it is not the sole determinant.

Smoking

Figure 3

The bar plot in Figure 3 displays the distribution of heart disease among smokers and non-smokers, highlighting a higher prevalence of heart disease in smokers. This correlation is significant for us because we'd like to see if the smoking status of an individual has an effect on heart disease. This can help us in improving our predictive model.

Alcohol Drinkers

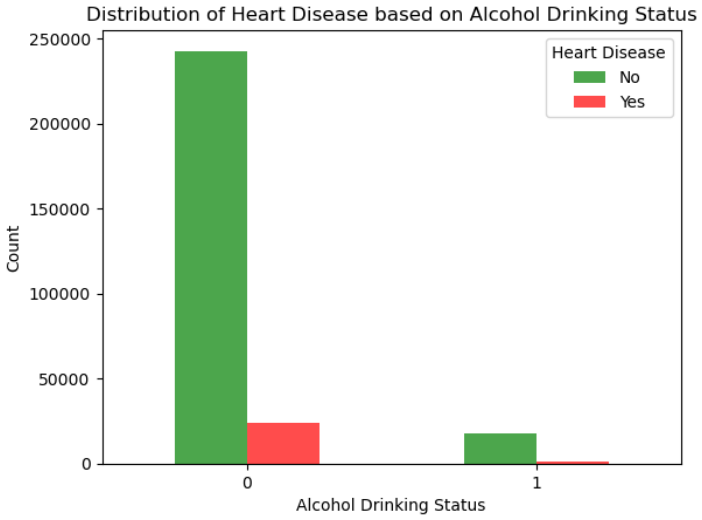

Figure 4

Figure 4 displays the relationship between alcohol drinking status and heart disease prevalence. It reveals a notable distribution pattern, with different counts of heart

disease cases among alcohol drinkers and non-drinkers. This indicates that on a dataset with more than 250000 individuals, drinking alcohol has a huge impact on heart disease as compared to those who do not drink which is almost close to 0.

Stroke

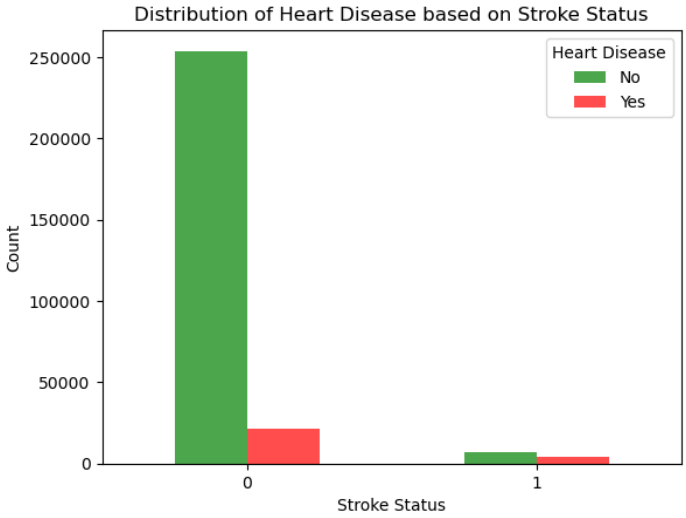

Figure 5

Figure 5 depicts the correlation between stroke history and heart disease occurrence. It shows distinct counts of heart disease cases among individuals with and without a history of stroke, indicating a potential link between stroke history and an increased risk of heart disease.

Physical Health

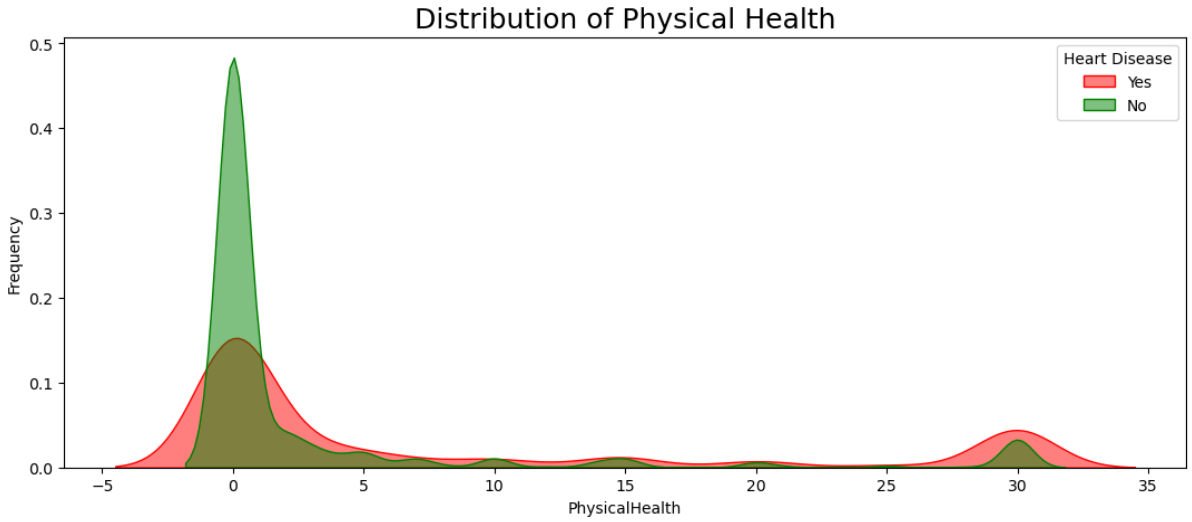

Figure 6

In Figure 6, the red curve (heart disease) demonstrates a wider spread, suggesting a more varied range of physical health conditions among those with heart disease. There is a noticeable overlap between the two distributions, especially in the lower range of 'Physical Health' issues, indicating that not all individuals with heart disease have high levels of physical health problems.

Mental Health

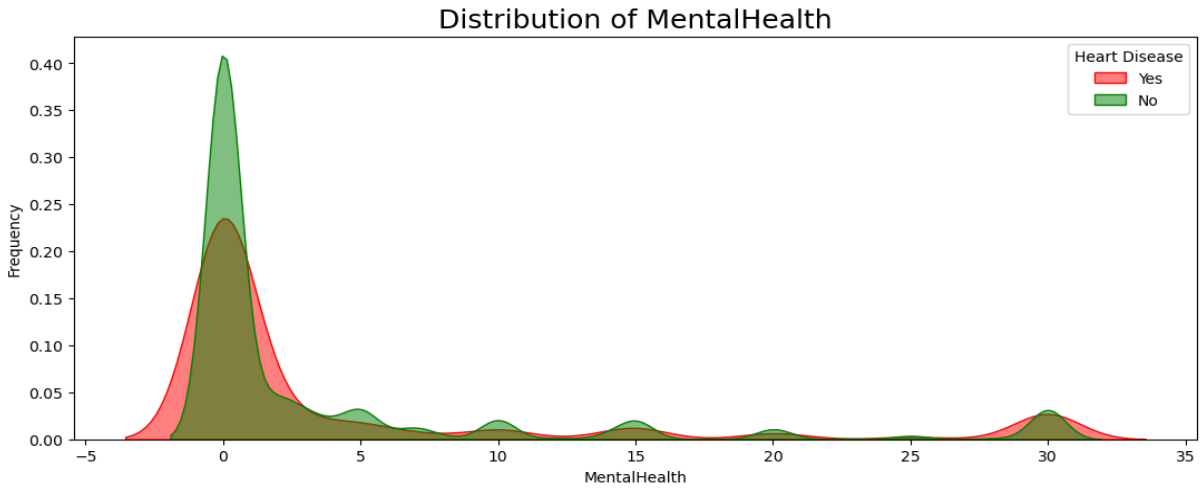

Figure 7

In Figure 7, both distributions seem to be centered around a value between 0 and 5, suggesting that the median or mean of "MentalHealth" scores for both groups is within this range. The tails of graph, for individuals without heart disease, are longer showing a wide range of "MentalHealth" values. This also might suggest that there is more variation amongst individuals without heart disease compared to those with it.

Difficulty in Walking

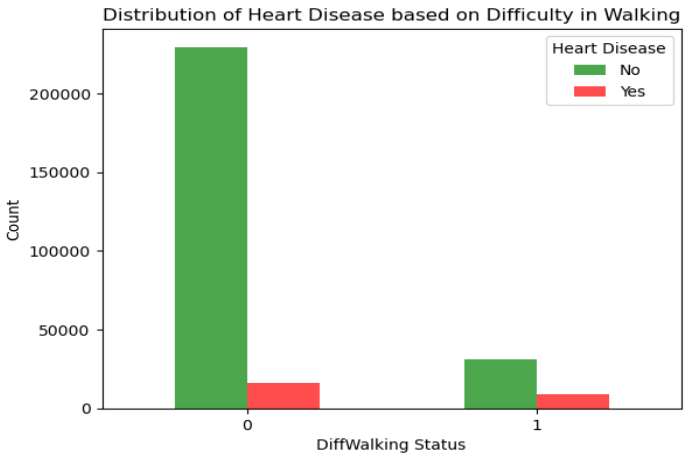

Figure 8

Figure 8 shows that a large number of individuals who do not have difficulty walking also do not have a heart disease. A much smaller number of individuals who do not have difficulty walking have heart disease. It is evident that the number of individuals without heart disease is much higher than those within both walking difficulty categories.

Gender

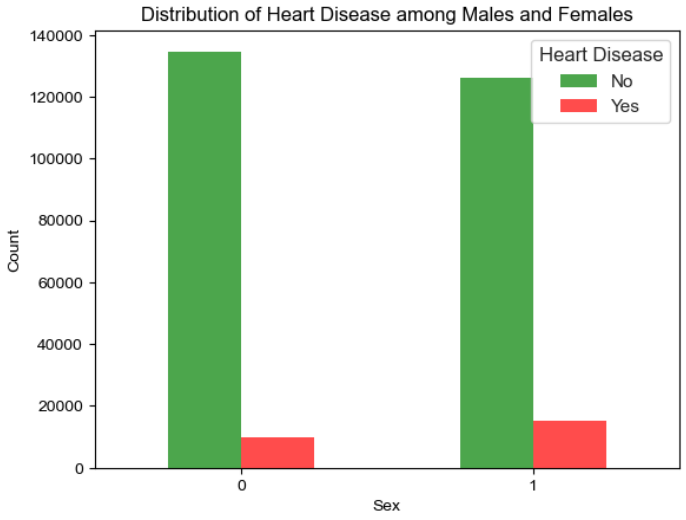

Figure 9

In Figure 9, the bar graph illustrates the distribution of heart disease among genders, with '1' indicating males and '0' indicating females. Both genders have a much higher

count of individuals without heart disease compared to those with it, and the numbers are relatively similar across the genders. However, the count is higher for males.

Age

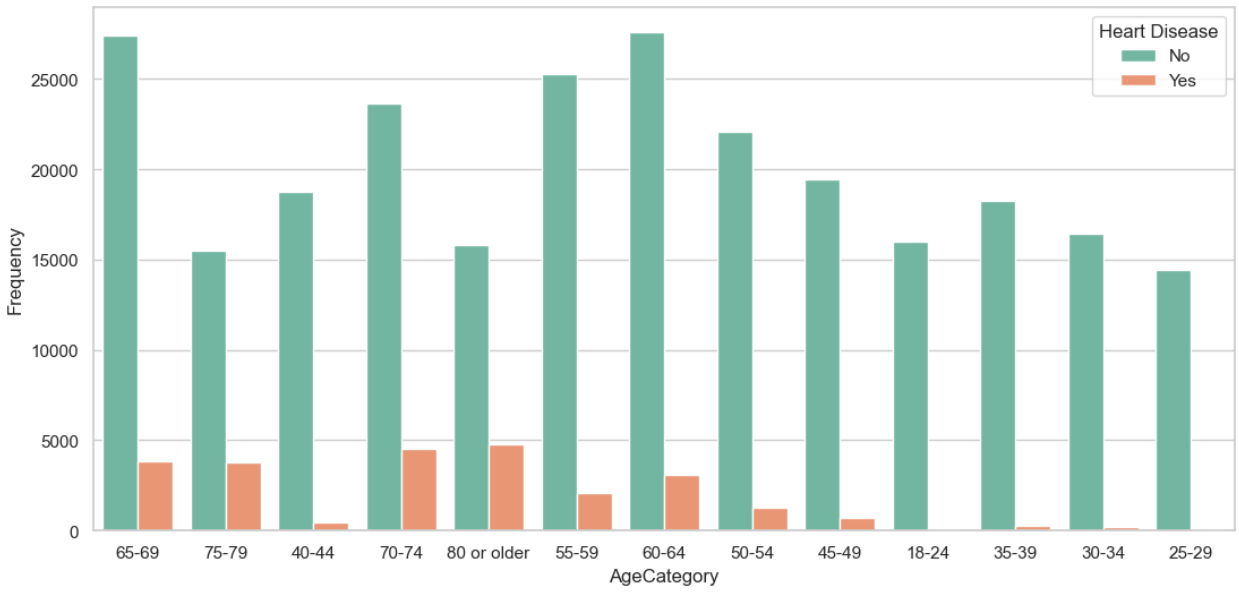

Figure 10

In Figure 10, the bar graph shows the frequency of heart disease occurrence across different age categories, indicating an overall higher prevalence of heart disease in older age groups.

Race

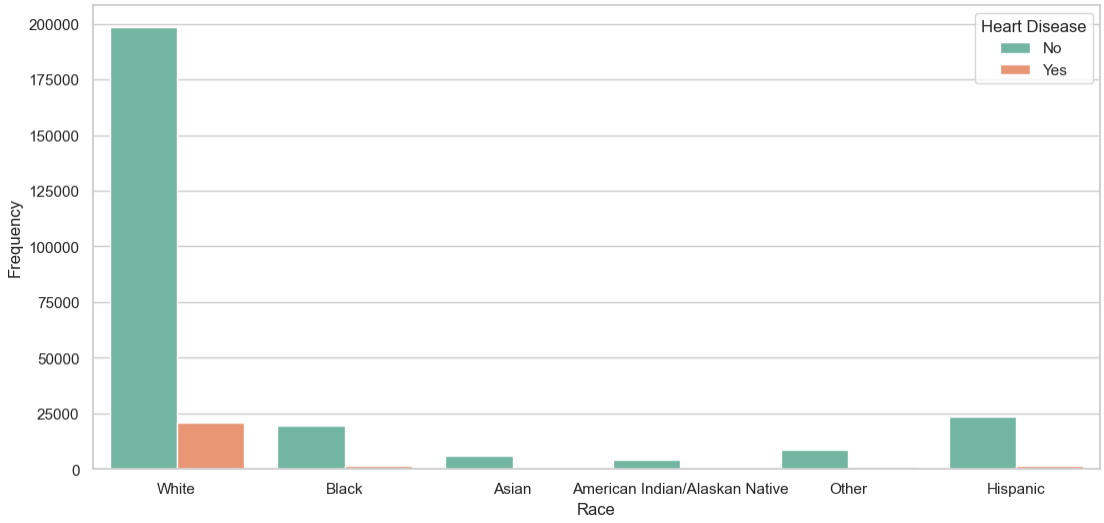

Figure 11

In Figure 11, the bar plot shows that Heart Disease has a significant difference amongst different races. Amongst White, Black, Asian, American Indian, Hispanic and others, White seem to have the highest frequency, followed by Hispanic and Black. This suggests that their lifestyle factors such as Physical Health and Mental Health might play a huge role.

Diabetes

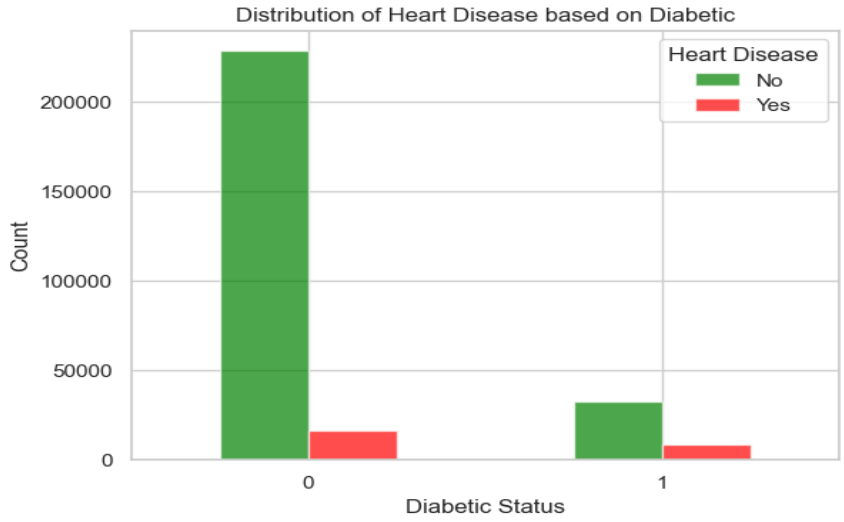

Figure 12

In Figure 12, It shows that a larger number of non-diabetic individuals do not have heart disease compared to those with diabetes, and the count of individuals with heart disease is higher among diabetics than non-diabetics.

Sleep Time

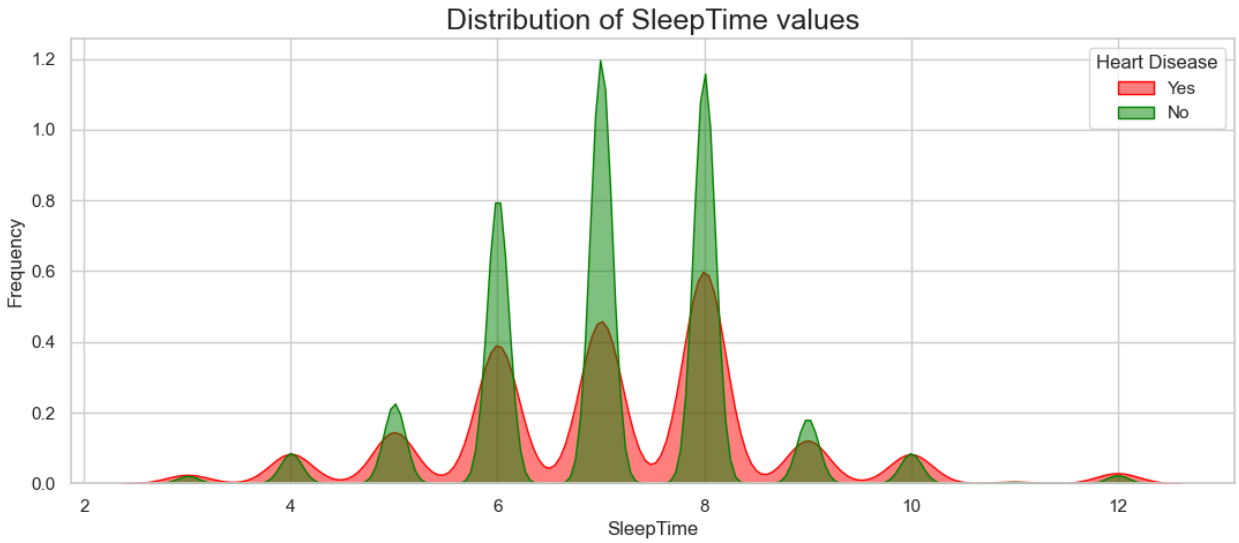

Figure 13

In Figure 13, the graph suggests that individuals with heart disease tend to have a different distribution of sleep times compared to those without, with peaks at different sleep durations. However, it still follows a similar pattern. Highest peaks are at 7 and 8 hours of sleep time.

4. Modeling

4.1 Train/Test Data

The variables we used in our training and validation data sets to predict Heart Disease are BMI, Smoking, AlcoholDrinking, Stroke, PhysicalHealth, MentalHealth, Sex, Diabetic, PhysicalActivity, Asthma, KidneyDisease, and SkinCancer. We used an 80/20 split for training and test/validation set sizes. The following models are measured for accuracy in terms of Accuracy, Precision and Recall.

4.2 Linear Regression

Looking at the models we had learned throughout this course, we wanted to start with a simple model that we could use as a baseline approach to solving our problem of predicting heart disease. From our data exploration, we identified a correlation between multiple variables and heart disease of an individual. Since linear regression assumes that there is linear relationship between the input and variables, hence we decided to use this model to see the accuracy that could be achieved. We were able to achieve an accuracy of 91.4%, with a recall of 7.9% and a precision of 51.5%. From this results, we used the accuracy levels as a baseline level to predict our other models.

4.3 Random Forest Regression

Our second model was random forest regression because it utilizes an ensemble of decision trees, compared to a single decision tree. This would help to reduce over-fitting in our data. We used hyperparameters: max_depth and n_estimators parameters to optimize our model. The model achieves a high accuracy of 91.2%, indicating that it makes correct predictions for most cases. However, there's a critical trade-off in its performance with a recall of only 3.6%.

4.4 K-Nearest Neighbors

Our next approach was K-Nearest Neighbors since it tries to find similarities between predictors and values that are within the dataset. Since kNN makes predictions based on the similarity between a data point and its k-nearest neighbors, we wanted to explore this approach to see the results compared to our other models. Amongst all the models, kNN had the lowest accuracy level. However, a significantly higher recall value of 9.6% compared to models.

4.5 XGboost Regression

Instead of using the algorithms covered in the course, we decided to experiment with a different approach called XGBoost (eXtreme Gradient Boosting). XGBoost is a gradient boosting framework that employs a collection of decision trees. We selected this algorithm because it shares similarities with the random forest algorithm, both being decision tree-based ensembles, but it incorporates gradient boosting [1].

We trained the model using a loss function (i.e. used to measure the difference between the predicted and actual outcomes) [2]. We also used Scikitlearn's GridSearchCV to conduct hyperparameter tuning on the max_depth, learning_rate, n_estimators, and gamma parameters to optimize our model. We achieved an accuracy of 91.2% with a recall of 7.79% and precision of 5.54%.

5. Results

Models	Accuracy	Recall	Precision
Linear Regression	0.914	0.079	0.515
Random Forest Regression	0.912	0.036	0.557
K-Nearest Neighbors	0.907	0.096	0.350
XGboost Regression	0.912	0.058	0.554

Throughout the project, we were able to investigate and observe the different aspects that may be a cause of heart disease. After the exploration of the data, we compared how each different variable effected the frequency of other variables. After the visualizations created, we were able to able to pick the most relevant and important variables that we used to build our models. Out of the 19 variables we analyzed, we selected 13 of them to base our predictions on.

We were able to explore various methods to utilize different techniques to achieve high levels of accuracy from our predictions. After building the models, we decided to use the four models to get a comparison. Linear regression, random forest regression and kNN were models that we had learned in this course. We choose linear regression for its simplicitiy and used it as a baseline to compare. We then used random forest and kNN that gave us a significantly better prediction. We then decided to choose XGboost as it builds on the random forest model.

Given the models we built, here are our observations:

- **Linear Regression:** In this case, the accuracy of linear regression is 0.914, which means that it correctly predicts the target variable 91.4% of the time. However, the recall of linear regression is only 0.079, which means that it only correctly identifies 7.9% of the positive cases. This suggests that linear regression may be missing some important information from the input features.
- **Random Forest Regression**: In this case, the accuracy of random forest regression is 0.912, which is slightly lower than that of linear regression. However, the recall of random forest regression is much higher, at 0.036. This suggests that random forest regression is better at identifying positive cases than linear regression.
- **K-Nearest Neighbors**: In this case, the accuracy of k-nearest neighbors is 0.907, which is slightly lower than that of linear regression and random forest regression. The recall of k-nearest neighbors is also lower, at 0.096. This suggests that k-nearest neighbors may not be as effective at capturing complex relationships between the input features and the target variable as linear regression or random forest regression.
- **XGboost Regression**: In this case, the accuracy of XGBoost regression is 0.912, which is the same as that of random forest regression. The recall of XGBoost regression is also higher, at 0.058. This suggests that XGBoost regression is better at identifying positive cases than random forest regression.

Random forest regression and XGBoost regression have the highest recall, while linear regression has the highest precision.

References

[1]https://www.simplilearn.com/what-is-xgboost-algorithm-in-machine-learning-article#:~:text=cost%20to%20accuracy!-,What%20is%20XGBoost%20Algorithm%3F,optimize%20their%20machine%2Dlearning%20models

[2]https://www.analyticsvidhya.com/blog/2022/06/understanding-loss-function-in-deep-learning/#:~:text=a%20loss%20function%3F-,A.,well%20it%20fits%20the%20data

[3]https://www.analyticsvidhya.com/blog/2020/03/what-is-multicollinearity/#:text=Multicollinearity

[4] https://xgboost.readthedocs.io/en/stable/

[5]https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html

[6]https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease

6. Project Experience Summary

6.1 Vraj

- Conducted exploratory data analysis by generating insightful visualizations, including heat maps and various plots, utilizing Python libraries such as matplotlib and seaborn.
- Designed and implemented four distinct predictive models using machine learning techniques by exploring a range of algorithms, including linear regression, random tree regression, K-Nearest Neighbors (KNN), and XGBoost, to assess their effectiveness in predicting heart disease.
- Conducted a comparative analysis of the performance of the four models.
- Evaluated the efficacy of each model by assessing their predictive accuracy and potentially identifying which model provided the most reliable predictions for heart disease.
- Collected insights and transformed into a report.

6.2 Abhay

- Finding a good database is hard in reality. One has to look through many, and even use them only to reject in order to find a good one.
- Getting meaning out of that data is what we learned in this project. Using sources beyond the project to clean data, learning that visualizing data at every important step is a great idea.
- Comparing data and prediction models is only reliable if we have trust in input data, and if it is cleaned properly to match our goals.