

Heart Disease Prediction with Emphasis on Factors Causing It

Milestone: Model Exploration and Selection

Group 63

Krishna Hemant

Yanming Liu

617-834-9804

617-959-5728

durgasharan.k@northeastern.edu

liu.yanmi@northeastern.edu

Percentage of Effort Contributed by student 1: 50%

Percentage of Effort Contributed by student 2: 50%

Signature of Student 1: Krishna Hemant

Signature of Student 2: Yanming Liu

Submission Date: 3/28/2022

Problem Setting

Cardiovascular health has been one of the most important aspects of human beings over centuries and science and medicine have been trying to predict, diagnose and solve problems as efficiently as possible. As technology and the amount of data generated have increased multifold, computer science started playing a role in helping doctors to cure heart related problems with a deeper understanding of it. Data science, especially data mining algorithms, has piqued interest in academia or healthcare research for detecting patterns, classifying or predicting values in various areas which could improve a person's health and also contribute data for further understanding of our complex body.

Problem Definition

The purpose of the project lies in predicting if the person has heart disease or not using a collection of features from patients and healthy volunteers. All the features can be obtained by the healthcare app, so as to avoid tedious medical examinations in hospitals. Furthermore, we aim to find the most influential factors contributing to people's disease using exploratory data analysis, which can help in preventing serious implications to the patients in the future. We aim to explore and compare different data mining techniques to further understand the data and come up with an accurate model. Lastly, we would try to do a correlation analysis between heart disease and covid with additional relevant datasets and will show if we need to take special measures on protecting patients with heart disease.

Data Sources

Kaggle is a website for data scientists and machine learning to learn, discuss, compete, and solve real problems. The heart disease dataset can be found in the following link:

<https://www.kaggle.com/sulianova/cardiovascular-disease-dataset>

Data Description

The heart disease dataset consists of 70000 records of patient data, 11 attributes and one target column. It can be categorised into 3 parts, objective, examination, and subjective. Objective attributes are basic information, like Age, Height, Weight, and Gender. The examination contains Systolic and Diastolic blood pressure, Cholesterol and Glucose levels. Subjective information refers to whether the person smokes, drinks, or actively exercises. The target variable is if the patient has heart disease or not. The dataset contains numeric and categorical variables, but all are encoded by numerical data types in the original dataset.

Table 1 Variables and Their Data Types

	Full Name	Feature Type	Abb. Name	Data Type
1	Age	Objective Feature	age	int (days)
2	Height	Objective Feature	height	int (cm)
3	Weight	Objective Feature	weight	float (kg)
4	Gender	Objective Feature	gender	categorical code
5	Systolic blood pressure	Examination Feature	ap_hi	int
6	Diastolic blood pressure	Examination Feature	ap_lo	int
7	Cholesterol	Examination Feature	cholesterol	1: normal, 2: above normal, 3: well above normal
8	Glucose	Examination Feature	gluc	1: normal, 2: above normal, 3: well above normal
9	Smoking	Subjective Feature	smoke	binary
10	Alcohol intake	Subjective Feature	alco	binary
11	Physical activity	Subjective Feature	active	binary
12	Presence or absence of cardiovascular disease	Target Variable	cardio	binary

Data Preprocessing

A Glimpse of Data

Show the First 5 Records

Table 2 Original Data for the 5 Rows

id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
0	18,393	2	168	62	110	80	1	1	0	0	1	0
1	20,228	1	156	85	140	90	3	1	0	0	1	1
2	18,857	1	165	64	130	70	3	1	0	0	0	1
3	17,623	2	169	82	150	100	1	1	0	0	1	1
4	17,474	1	156	56	100	60	1	1	0	0	0	0

We got basic ideas for further operations in the Data Cleaning section from this original data.

Summary Statistics

Table 3 Summary Statistics

	age	height	weight	systolic_blood_pressure	diastolic_blood_pressure
count	70,000	70,000	70,000	70,000	70,000
mean	0	164	74	129	97
std	0	8	14	154	188
min	0	55	10	-150	-70
25%	0	159	65	120	80
50%	0	165	72	120	80
75%	0	170	82	140	90
max	0	250	200	16,020	11,000

It shows some outliers, which will be removed in the section, Data Cleaning, with the help of pair plots and heart disease-related domain knowledge.

Data Cleaning

The data was imported to a data frame in pandas and the first step was to rename the columns of the data frame to understandable and more logical terms. Then the data type of columns was rightly changed to categorical and numerical variables. The 'Age' Variable seemed to have been set in days and we converted it into years. We dropped the 'id' column as it's of no use on further analysis. The data was checked for null values before proceeding with the next steps, and data entry consistency for categorical variables was ensured.

A pair plot was plotted between the numerical variables(Height, Weight, Systolic and Diastolic blood pressure) to get an overall outlook at the outliers in the data. Upon closer analysis and by getting a statistical summary of the numerical variables, limits were set to remove outliers from the data frame.

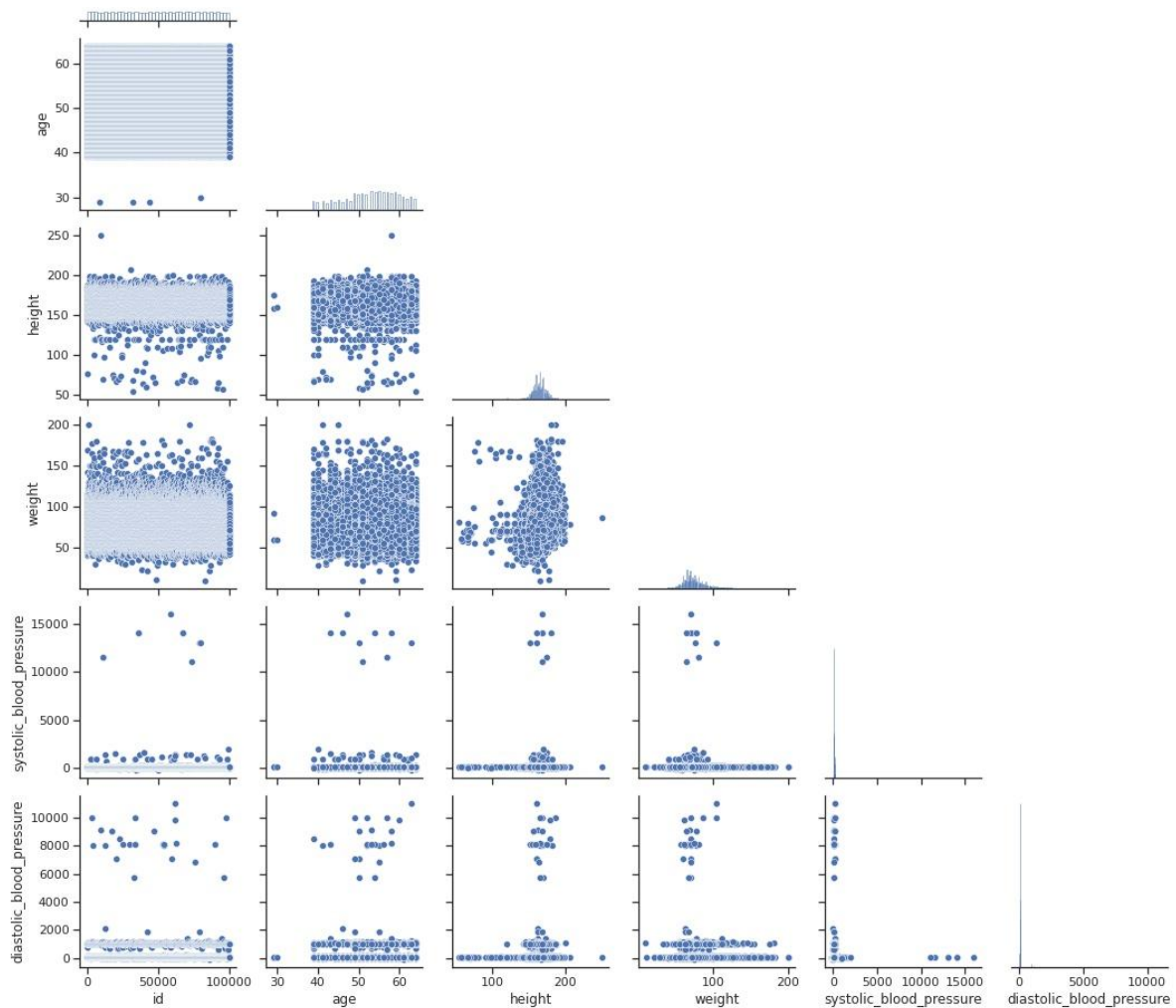


Fig. 1 Pair plot among numerical variables

Here are the limits set based on domain knowledge from online resources, and you can find them in reference:

1. The heights for men and women were analysed and by taking 3 standard deviations and adding/subtracting a bit, the range was set to $144 \text{ cm} < \text{Height} < 214 \text{ cm}$
2. For Blood pressures, we added 3 filters,
 - a. Systolic Blood Pressure $>$ Diastolic Blood Pressure
 - b. $50 < \text{Systolic Blood Pressure} < 350 \text{ (mmHg)}$
 - c. $0 < \text{Diastolic Blood Pressure} < 300 \text{ (mmHg)}$

Here are the statistics of the Numerical variables after the data cleaning,

Table 4 Summary Statistics of Filtered Data

	age	height	weight	systolic_blood_pressure	diastolic_blood_pressure
count	68,414	68,414	68,414	68,414	68,414
mean	53	165	74	127	81
std	7	8	14	17	10
min	29	144	11	60	1
25%	48	159	65	120	80
50%	53	165	72	120	80
75%	58	170	82	140	90
max	64	207	200	240	182

Data Transformation and Preparation

We sampled 5000 records from the data frame of 70,000 records for ease of analysis and plotting and also for the future purposes of running machine learning models. One hot encoding was performed on the Gender column having values (Male, Female) to create two columns with (1,0) as categorical values. The target variable and predictor variables were separated into two sets for the future machine learning models. For each of them, training and validation sets were obtained with a ratio of 70/30. The predictor or feature variables were normalised using the StandardScaler() approach for a uniform distribution of data in all columns and one doesn't outweigh the other.

Data Exploration and Analysis

We conduct an overall analysis on quantifying numeric variable distribution or class frequency of categorical features for subjects who have heart disease compared to those who do not. Plots A, B, C from figure 2 illustrate that - high age, low height, high weight are characteristics of a heart disease patient. For normal people, the systolic blood pressure versus diastolic blood pressure is 120/80 which meets our standards in medical settings. However, those two pressures increase if the subject has heart disease, like Fig. 2 D, E.

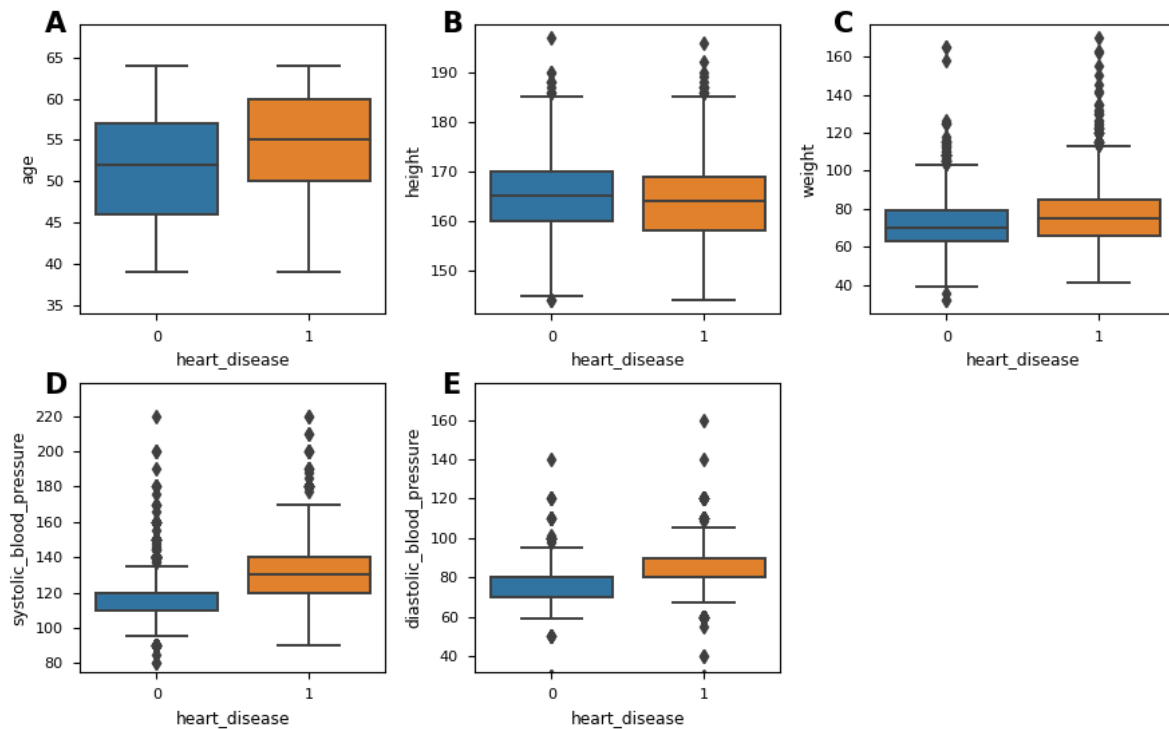


Fig. 2 Numeric Features Distribution among Subjects who have heart disease or not

Our data is approximately balanced, as Fig. 3 A(below) shows, which enacts a superior learning performance for the further method. It can be noticed that the percentage of patients with high cholesterol and high glucose significantly increases in the case where the person has heart disease (Fig 3 B, C). This also deepens our understanding of heart disease. Due to poor body conditions which lead to having heart disease, patients tend to smoke less, have alcohol intake and work out less, like Fig. 3 D, E, F.

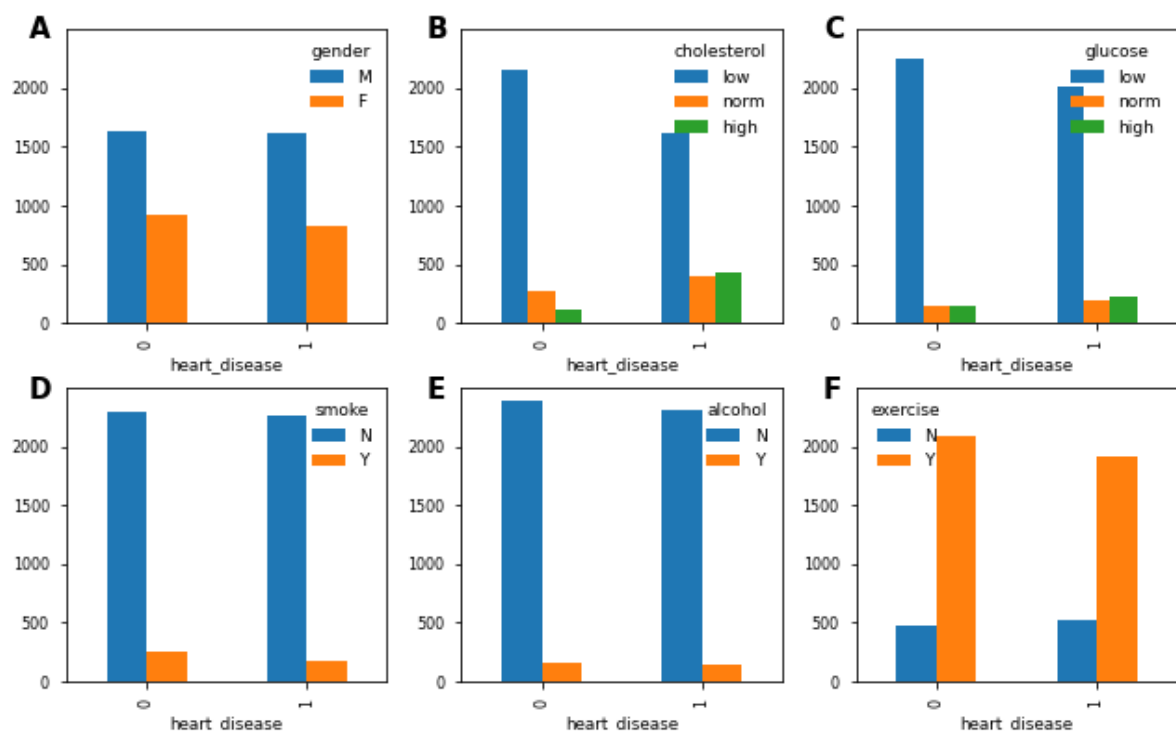


Fig. 3 Categorical Features Frequency among Subjects who have heart disease or not

Correlation analysis was performed on the numerical variables and the blood pressures have a high correlation value. As seen from the plot, age and weight have very little correlation as a person starts maintaining weight in an interval after a certain age. Height and weight are correlated as they directly affect each other. Blood pressure seems to have the most correlation with weight than any other variable.

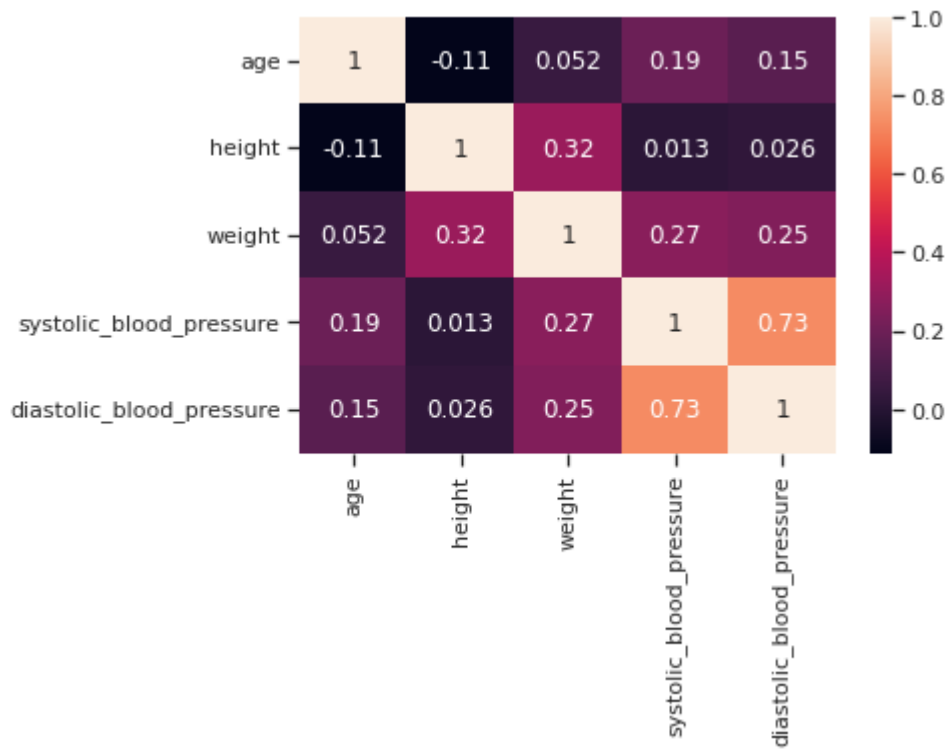


Fig. 3 Influential Numerical Factors Correlations Analysis

The Pearson Chi-squared test was done for the categorical variables using and the results show that smoking and alcohol consumption were dependent variables with a p-value less than an alpha value of 0.05 (rejects null hypothesis). Similarly, the physical activity and smoking were independent variables and failed to reject the null hypothesis with a p-value > 0.05 and is not a significant result.

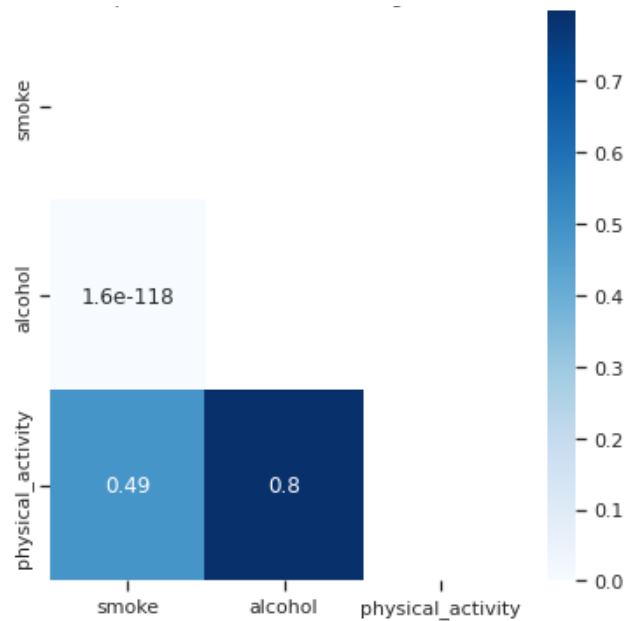


Fig. 4 Chi-Square Test Results for Categorical Data

Data Mining Models

Six models will be fitted on train data and tested on validation data for the binary classification task (If the patient is prone to heart disease or not). First, we briefly introduce the models. Then, the reason and attention for each choice are illustrated, and the implementation and results for each model are shown.. Here are the models we have tried and implemented for now.

K Nearest Neighbours (KNN)

KNN method predicts whether subjects have heart disease or not by assigning the majority class of its kth nearest neighbour in the training data, which was measured by Euclidean distance, for the project.

Advantages:

The reason why we choose KNN method lies in its flexible tasks adaption, whether the response variable is numeric or categorical. No parameter assumption is required for predicting whether a subject has heart disease or not, and the performance is exceptionally good with large records with multiple combinations of predictors to mark the class.

Disadvantages:

However, it is time consuming to calculate the distance between each record and thousands of existing training data and finding the kth nearest neighbours. Furthermore, high dimensional space will separate data points far away compared to 2 or 3 predictors, thus increasing the poor accuracy unless a huge record exists.

Steps taken and results:

We encoded the categorical variables with a dummy one, then split them into train data and validation data. Numeric data is normalized based on mean and variance of train data. The best k number is obtained by looping it from 1 to 10 to check the highest accuracy score. The prediction performance summary is conducted at the best K number, and ROC curve is drawn to determine the optimal cut off value at which the model can achieve a high sensitivity(will implement it later).

We implemented the KNN model on our dataset and we got an accuracy of Accuracy = 0.7016, when k equals to 5. Can be graphically seen from the plot below,

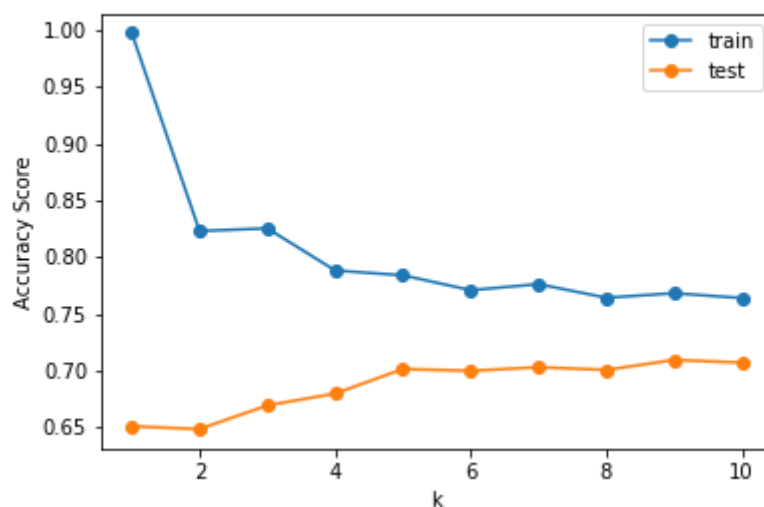


Fig. 5 Accuracy for Train and Test Data

DECISION TREES

Decision tree is an intuitive tree-like method to help with decisions. Following the conditional judgement, it will keep splitting based on conditions until all the nodes are pure.

The length of the tree and number of nodes can be controlled, also known as pruning, which can help with the accuracy and the fit of the model

Advantages

- Simple and effective method, requires very less pre-processing
- Doesn't require normalisation or scaling of data
- Very intuitive and good to understand alternate decisions
- Missing data or null values does not affect the model

Disadvantages

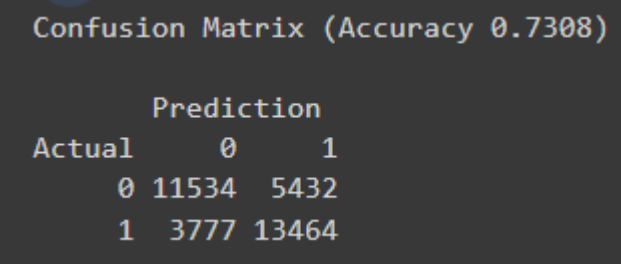
- Can easily lead to overfitting if not controlled
- A single new point can make a huge difference in the overall model
- Sensitive to noise

We fitted our heart disease dataset onto the decision tree model, with a test and train size of 50% and we got an overall accuracy of 73%. We used the entire dataset(68,414 rows) to train the model which showed a higher accuracy than the sampled dataset of 5000 rows due to more training data.. The initial accuracy was 63% and we observed a 10% increase after we Changed the following parameters,

```
criterion = "gini",  
random_state = 10,  
max_depth=5,  
max_leaf_nodes= 30
```

Results of the Implemented Model:

Classification Summary:



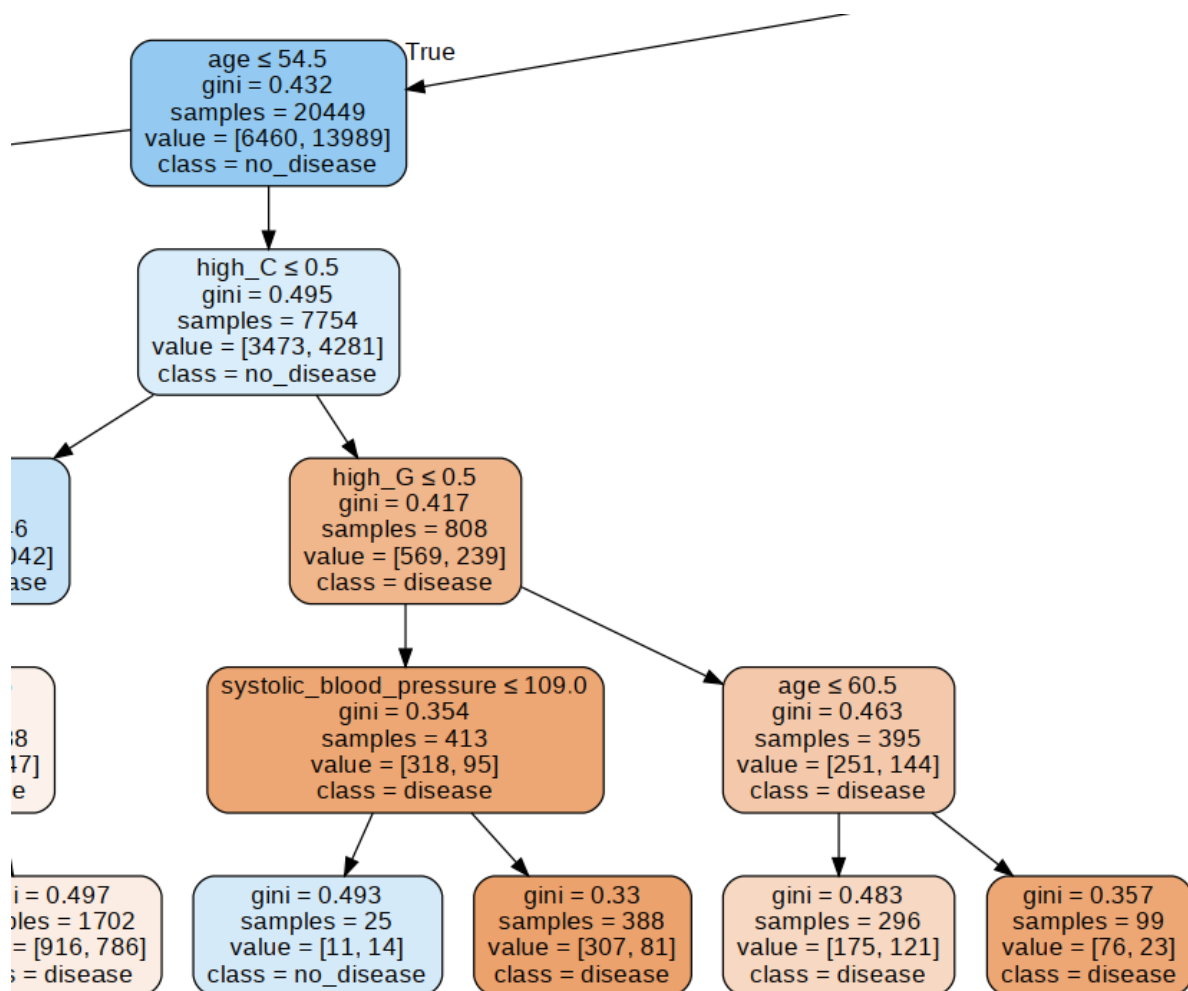
```
Confusion Matrix (Accuracy 0.7308)  
  
      Prediction  
Actual    0     1  
    0 11534  5432  
    1  3777 13464
```

Classification report:

	precision	recall	f1-score
disease	0.753315	0.679830	0.714688
no_disease	0.712532	0.780929	0.745164
accuracy	0.730786	0.730786	0.730786
macro avg	0.732923	0.730380	0.729926
weighted avg	0.732759	0.730786	0.730049

We used the package graphviz to get a clear output of our model in its tree form with the parameters above and we have included a snippet here since the original is too large to fit in,

A section of our pruned tree with depth = 5:



There are multiple ways to try to gradually improve the accuracy of the decision tree model, and we are constantly working on it to make it better.

Future models we are planning to implement once we learn it:

- Naive Bayes Classification
- Random Forest
- Logistic Regression
- Support Vector Machine

References:

[1] Shmueli, Galit, et al. Data mining for business analytics: concepts, techniques, and applications in R. John Wiley & Sons, 2017.

[2] <https://wisdomplexus.com/blogs/data-mining-algorithms-classification/>