

# Heart Disease Prediction with Emphasis on Factors Impacting it

Group 63

Krishna Hemant

Yanming Liu

617-834-9804

617-959-5728

[durgasharan.k@northeastern.edu](mailto:durgasharan.k@northeastern.edu)

[liu.yanmi@northeastern.edu](mailto: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: 4/24/2022

<b>Problem Setting</b>	<b>3</b>
<b>Problem Definition</b>	<b>3</b>
<b>Data Sources</b>	<b>3</b>
<b>Data Description</b>	<b>4</b>
Table 1 Variables and Their Data Types	4
<b>Data Preprocessing</b>	<b>4</b>
Table 2 Original Data for the 5 Rows	5
Table 3 Summary Statistics	5
<b>Data Cleaning</b>	<b>5</b>
Fig. 1 Pair plot among numerical variables	6
Table 4 Summary Statistics of Filtered Data	7
<b>Data Transformation and Preparation</b>	<b>7</b>
<b>Data Exploration and Analysis</b>	<b>7</b>
Fig. 2 Numeric Features Distribution among Subjects who have heart disease or not	8
Fig. 3 Categorical Features Frequency among Subjects who have heart disease or not	9
Fig. 3 Influential Numerical Factors Correlations Analysis	10
Fig. 4 Chi-Square Test Results for Categorical Data	11
<b>1) K Nearest Neighbours (KNN)</b>	<b>11</b>
Advantages:	11
Disadvantages:	12
Procedure and result:	12
Fig. 5 Optimal Choice on K and Cut off Value according to Accuracy and Sensitivity	12
<b>2) TREES</b>	<b>12</b>
<b>2.1) DECISION TREE</b>	<b>13</b>
Advantages	13
Disadvantages	13
Table 5 Confusion Matrix and Classification Report for Best Decision Tree	14
<b>2.2) BOOSTED TREE</b>	<b>15</b>
<b>2.3) RANDOM FOREST</b>	<b>15</b>
<b>3) Logistic Regression</b>	<b>15</b>
<b>4) Neural Network</b>	<b>17</b>
Fig. 10 Classification Summary for Neural Network	17
<b>Final Model Performance Report</b>	<b>18</b>
<b>Conclusion</b>	<b>18</b>
<b>References</b>	<b>18</b>

**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. Numerous people have lost their lives because of careless examination in the beginning of the heart disease stage, we plan to make this a more positive experience and aid them.

**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 accurate models 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**

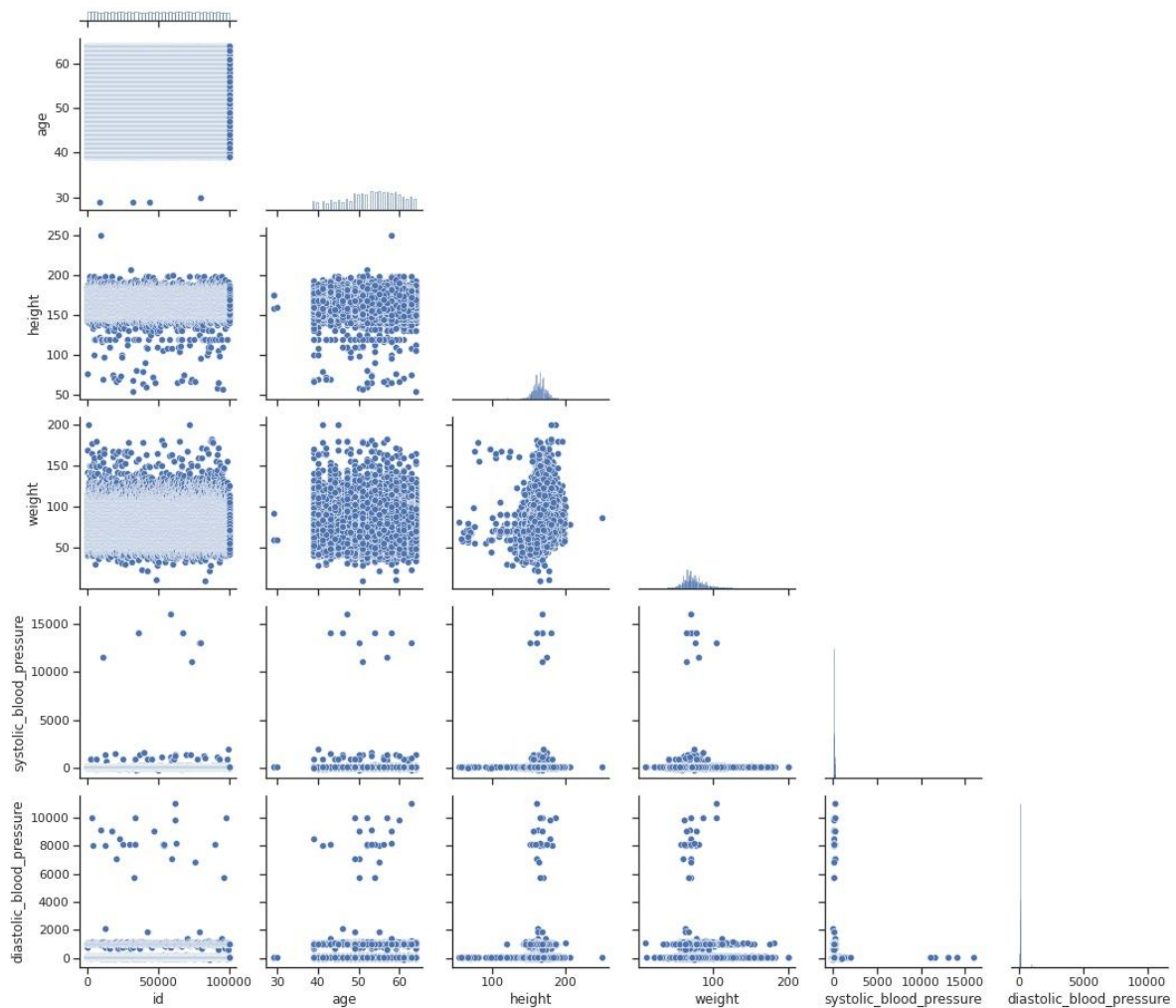
	age	height	weight	systolic_blood_pressure	diastolic_blood_pressure
count	70,000	70,000	70,000	70,000	70,000
mean	0	164	74	129	97
std	0	8	14	154	188
min	0	55	10	-150	-70
25%	0	159	65	120	80
50%	0	165	72	120	80
75%	0	170	82	140	90
max	0	250	200	16,020	11,000

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.  $\text{Systolic Blood Pressure} > \text{Diastolic Blood Pressure}$
  - 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**

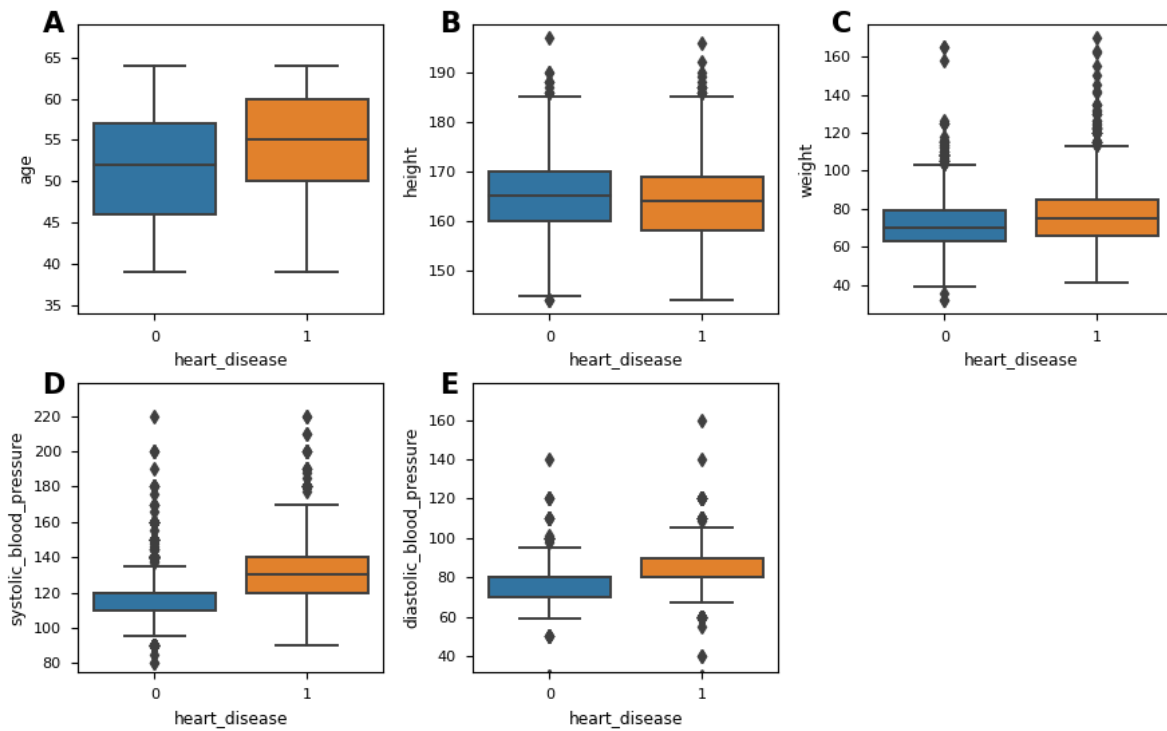
	age	height	weight	systolic_blood_pressure	diastolic_blood_pressure
count	68,414	68,414	68,414	68,414	68,414
mean	53	165	74	127	81
std	7	8	14	17	10
min	29	144	11	60	1
25%	48	159	65	120	80
50%	53	165	72	120	80
75%	58	170	82	140	90
max	64	207	200	240	182

**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**

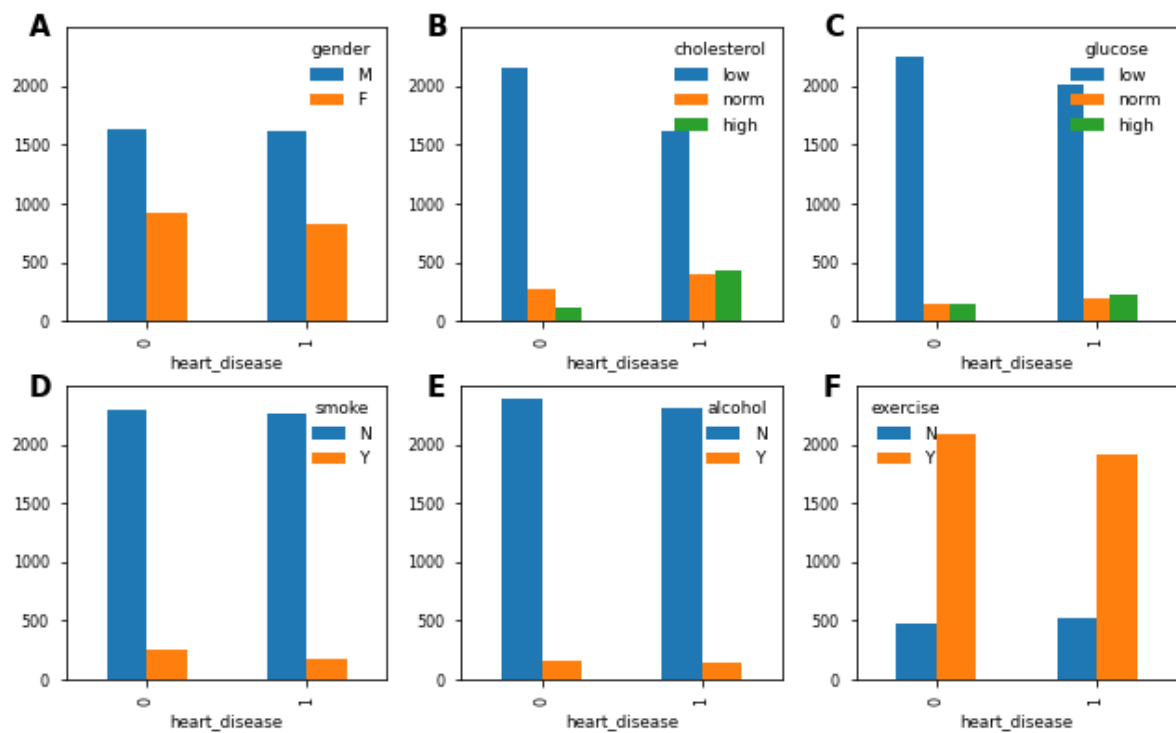
We conduct an overall analysis on quantifying numeric variable distribution or class frequency of categorical features for subjects who have heart disease compared to those who do not. Plots A, B, and C from figure 2 illustrate that - high age, low height, and high weight are characteristics of a heart 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**

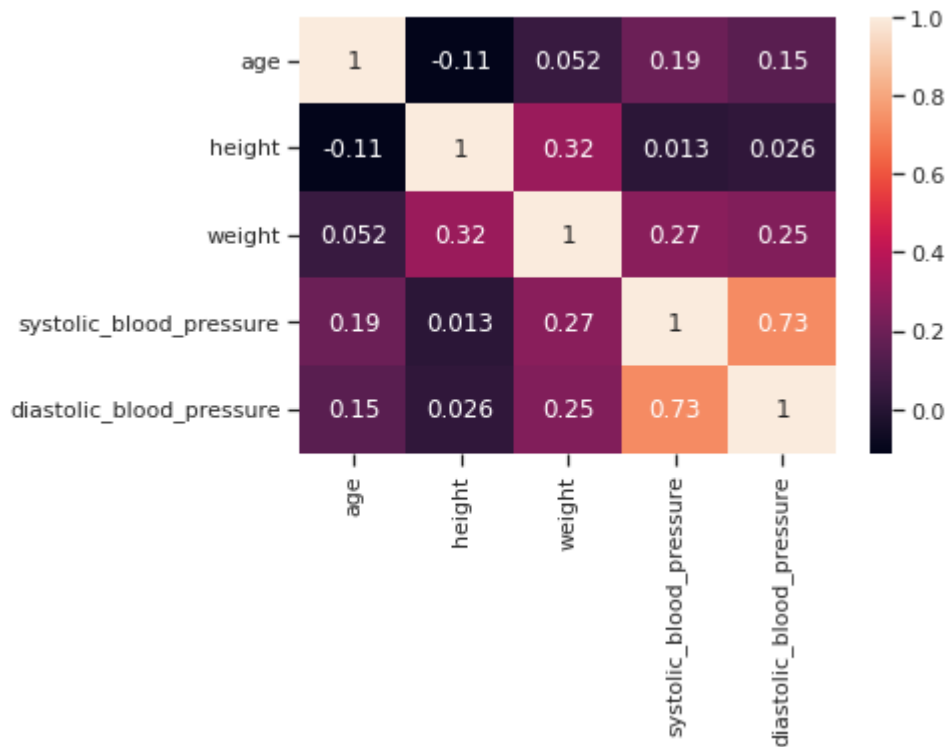
Our data is approximately balanced, as Fig. 3 A(below) shows, which enacts a superior learning performance for the further method. It can be noticed that the percentage of patients with high cholesterol and high glucose significantly increases in the case where the person has heart disease (Fig 3 B,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





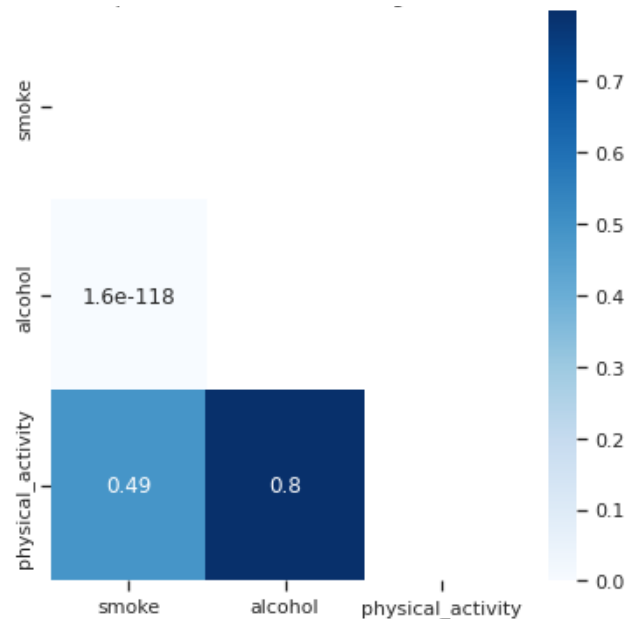
**Fig. 3 Categorical Features Frequency among Subjects who have heart disease or not**

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 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**

Three types of 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). Firstly, we briefly introduce the models. Then, the reason and attention for each choice are illustrated. Finally, the implementation details and predictive performance for each model are shown.

#### **1) K Nearest Neighbours (KNN)**

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

#### **Advantages:**

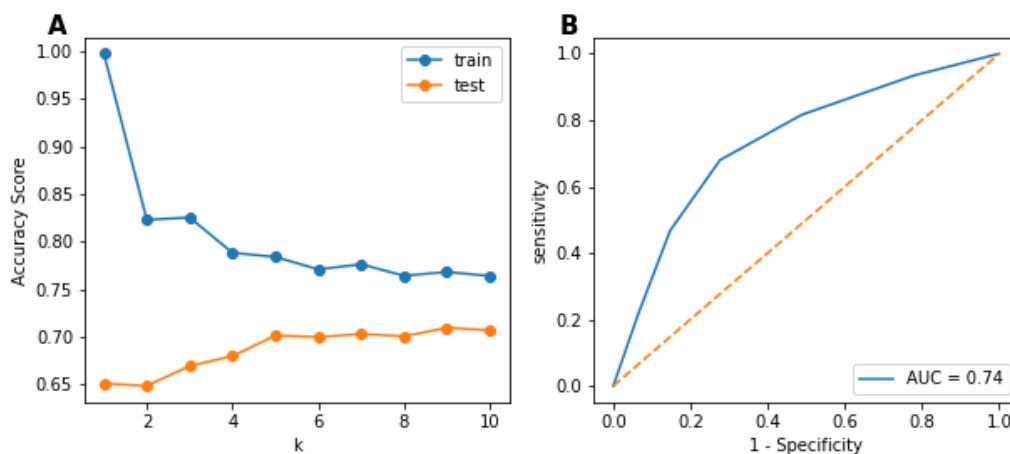
The reason why we choose the KNN method lies in its flexible tasks adaption, whether the response variable is numeric or categorical. No parameter assumption is required for predicting whether a subject has heart disease or not, and the performance is exceptionally good with large records with multiple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

**Procedure and result:**

We encoded the categorical variables with a dummy one and then split them into train data and validation data. Numeric data is normalized based on the 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 the ROC curve is drawn to determine the optimal cut off value at which the model can achieve high sensitivity. We implemented the KNN model on our dataset and we got an accuracy of 70.2% when k equals 5. The reason why we chose k=5 is to avoid overfitting as shown in Fig.5 A, the accuracy score of the model on test data decreases slightly when k is greater than 5. Fig. 5 B shows that the optimal cutoff value is 0.6 for the classifier for it has the biggest sensitivity difference between the model and a random classifier. The area under the curve is 0.74 which tells that it isn't a bad model.



**Fig. 5 Optimal Choice on K and Cut off Value according to Accuracy and Sensitivity**

**2) TREES**

A full Decision Tree is fitted on the whole dataset with 68,414 rows after removing the outliers. As the tree is unstable, cross-validation is introduced for the purpose of removing the effect of the occasional split conditions. Grid search on the parameters combinations of the

depth of the tree, minimum number of samples to split, and minimum impurity decrease are conducted for the best parameter choice for improving the predictive accuracy. Boosted tree as an alternative is implemented. Ensemble learning methods like Random forest are also tested.

### **2.1) DECISION TREE**

The 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.

#### **Advantages**

- A 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 training size of 75% on both sample (5000 rows) and full data set (68,414 rows). The predictive accuracies are 64.1% and 63.5% respectively. The relatively small performance illustrates that thousands of records are good enough to detect potential patterns under the data. We improve the accuracy based on the model trained on the full data set, and full data will be used in the following models. The average accuracy using 5-fold cross-validation on the full tree is 62.7%, and we assume this metric as a baseline. This tends to be a more possible one with the consideration of split randomness for the similar highest impurity decrease. We observed a ~10% increase in accuracy, from 62.7% to 73.2%, after we applied the best parameter using the grid search method. Here are the parameters:

```
criterion = "gini",
random_state = 1,
max_depth = 6,
```

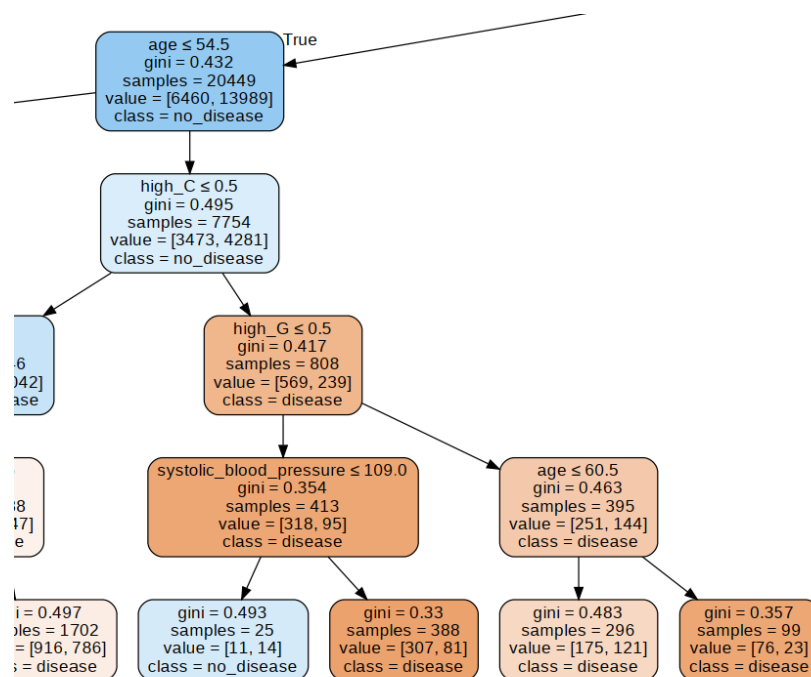
min\_impurity\_decrease = 0.0004,

min\_samples\_split = 10

**Table 5 Confusion Matrix and Classification Report for Best Decision Tree**

Confusion Matrix (Accuracy 0.7318)			precision	recall	f1-score	
Prediction						
Actual	0	1	no-disease	0.713111	0.789152	0.749207
			disease	0.755671	0.672565	0.711700
	0	6853 1831				
	1	2757 5663	accuracy	0.731759	0.731759	0.731759

We used the package graphviz to get a clear output of our model in its tree form, as is shown in Fig. 6, with the parameters above and we have included a snippet here since the original is too large to fit in.



**Fig. 6 A section of the pruned tree with depth = 5**

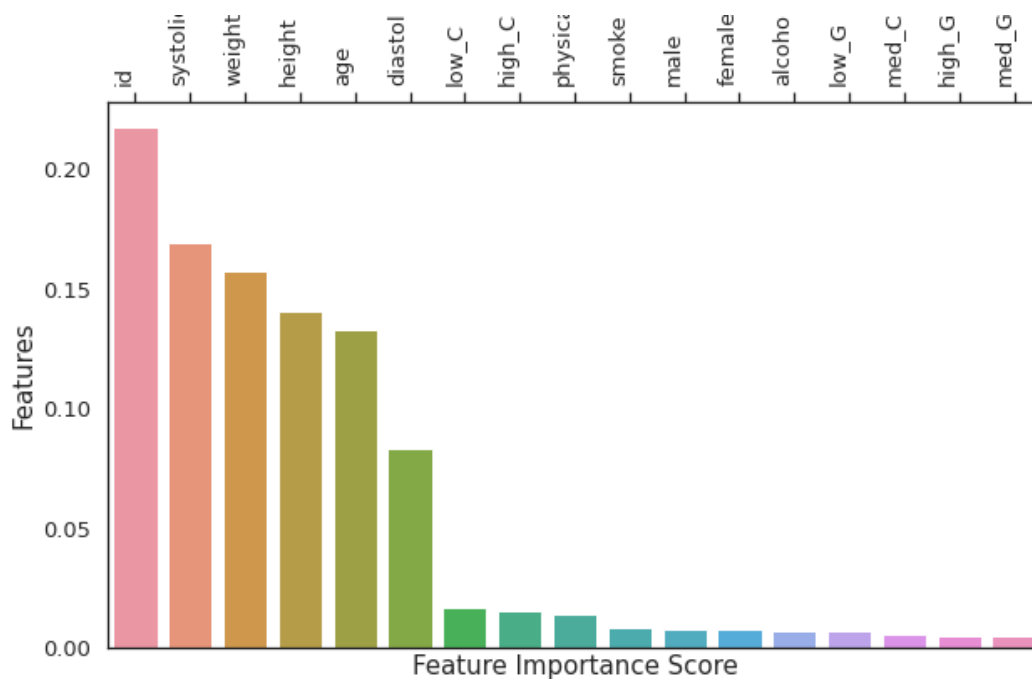
## **2.2)BOOSTED TREE**

Gradient Boosting builds an additive model in a forward stage-wise fashion; it allows for the optimization of arbitrary differentiable loss functions. The predictive accuracy that we got from the Boosted Tree method with 200 estimators is 72.4%.

## **2.3)RANDOM FOREST**

Random Forest is a robust algorithm to combine various decision trees with subsets of high-weighted features. An ensemble learner always manages plenty of techniques to get average overall performance. Random forests can also be an effective method to determine important predictors from the overall set.

The predictive accuracy for the Random Forest method with 200 estimators is 72.4%. We listed the feature importance for selecting subsets of predictors to achieve better predictive performance, as shown in Fig. 7.



**Fig. 7 Feature Importance of Using Random Forest**

## **3) Logistic Regression**

Logistic Regression aims to find the probability or propensity of a new record to belong to one class. This entails the purpose of binary classification. Similar to the linear regression method, it assumes that the predictors are not related and that there is no multicollinearity

among predictor variables. We use dummy coding to represent the categorical variables and drop the first column for each dummy variable to avoid multicollinearity.

The predictive accuracy of Logistic Regression is 70.2% without any penalty. The accuracy improves with the lasso method and it performs slightly better than the ridge method with higher accuracy of 72.97% than 72.95% from the ridge method. Both metrics under feature selection exceed the baseline of 70.2%. The benefits of penalising the predictors can be seen in these methods.

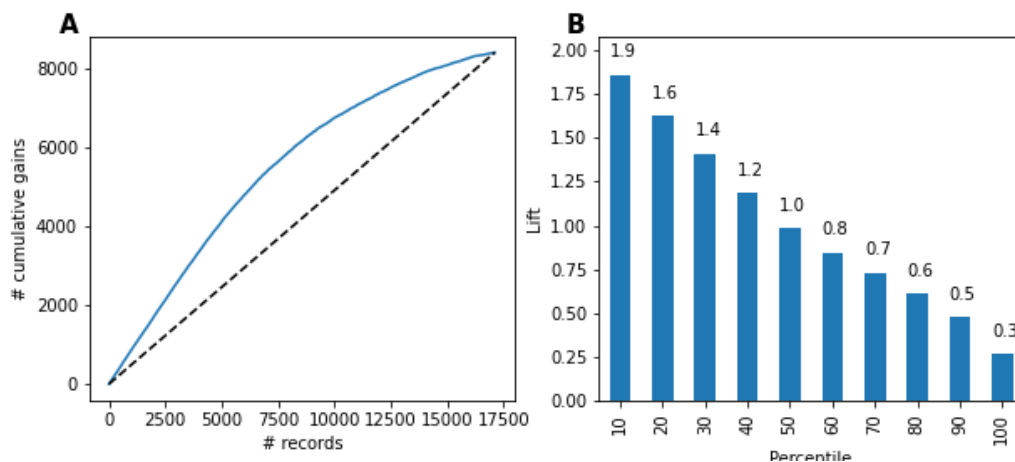
We have adopted the model with the best accuracy to get the coefficient of predictors and response, as is shown in Fig. 8. For continuous predictors, the coefficients for age, weight, and blood pressures, show a positive value denoting that all else being equal, a change of 1 unit in the predictors will increase the likelihood of the importance of the predictors by the magnitude of the coefficient. On the contrary, a negative one, like weight, indicates a higher of these predictors coincides with a lower possibility of heart disease. This is exactly what concludes in Fig.2 of the Data Exploration and Analysis section. For categorical variables, a male with high medium/high cholesterol, and middle glucose tends to have heart disease. However, exercise, smoking, and drinking behaviour are associated with a lower possibility of heart disease. This is because of oversampled data among these features. We can see a huge number of smoking, drinking, and exercise subjects than those who do not. The model may regard having heart disease as a numerous result of this behaviour.

	coefficient
age	0.050700
height	-0.003012
weight	0.010756
systolic_blood_pressure	0.056000
diastolic_blood_pressure	0.011251
smoke	-0.131320
alcohol	-0.235414
physical_activity	-0.216164
male	-0.010612
med_C	0.369231
high_C	1.084887
med_G	0.052190
high_G	-0.351904

**Fig. 8 Coefficient Between Predictors and Subjects Who Have Heart Disease**



The lift above the reference line in Fig. 9 A indicates the superior performance of the logistics regression model to a random classifier. In this way, we can rank those people who have heart disease with the highest probabilities, thus providing essential protection, preventive and curative measures for the predicted patients. Fig. 9 B illustrates that by choosing 10% of the records from the dataset, the model classifies the risk of heart disease 1.9 times better than a random classifier.



**Fig. 9 Cumulative Gains Chart and Decile-wise Lift Chart for Validation Data of Heart Disease for Logistic regression model**

#### **4) Neural Network**


Neural Network is known for high predictive performance in classification and regression tasks by imitating biological properties of the brain. We preprocessed the categorical variables into m-1 dummies. We apply the MLPClassifier from sklearn package by assigning two hidden layers with five nodes each and other default hyperparameters to achieve 73.47% overall accuracy., with the previous accuracy achieved being 73.35% from one hidden layer and 5 nodes Other hyper parameters combination test for a superior predictive performance will be tested in future under the balance of overfitting and underfitting.

Confusion Matrix (Accuracy 0.7347)		
Actual	Prediction	
	0	1
0	6675	2009
1	2528	5892

**Fig. 10 Classification Summary for Neural Network**

### **Final Model Performance Report**

Compared to the KNN, Tree-based method and Logistic Regression, the optimal accuracy of 73.47% is obtained when training a Neural Network under the best parameters, where the parameters used were 2 hidden layers with 5 nodes each and activation function being logistic and solver being lbfgs. The second highest was the Decision tree with an accuracy of 73.2% and the least accuracy obtained using 5-fold cross-validation on the full tree for the whole dataset with a size of 68,414 rows is 62.7%. By trying different parameters using grid search methods, there was an accuracy increase of 10.5% for the Tree. Overall, Neural Networks prove to be the best among all living up to its fame.



Model	KNN	Decision Tree	Boosted Tree	Random Forest	Logistic Regression	Neural Network
Accuracy	70.20%	73.20%	72.40%	72.40%	72.97%	73.47%

*Table 6 Accuracy Summary among Models*

### **Conclusion**

From all the models that we have tested we have decided to go forward with the Neural Network(MLP Classifier) as our model to be implemented for the healthcare application which can be used by the general public as a preventive measure to avoid serious cardiovascular disease.

### **References**

- Shmueli, Galit, et al. Data mining for business analytics: concepts, techniques, and applications in R. John Wiley & Sons, 2017.
- <https://wisdomplexus.com/blogs/data-mining-algorithms-classification>