Correlation Between Heart Activity, Certain Biomedical Traits and Heart Disease

Introduction

Heart disease is the leading cause of death in both men and women. 647,000 people die of heart disease every year. In America about 805,000 people have a heart attack every year and 605,000 of these are a first heart attack. And about 1 in 5 heart attacks are silent so the damage has been done without the person being aware of it.

Currently cardiovascular disease in the costliest disease in America with the price tag of $555 million in year 2016. On Average an employee with cardiovascular disease costs their employer $1,100 more per year due to loss in productivity. If the study can help with prediction of heart disease, preventive measures can be taken to reduce cost.

In this study we will look into correlations between heart trends, and or other biomedical traits that can help us predict the presence of heart disease. If there is a correlation between some heart activities and, or other biomedical traits this study can be used to help with heart disease prevention. As a high risk patient including men, the elderly, individuals with diabetes and persons with family members who have suffered from heart disease, early detection and prevention is important as heart disease is a silent threat and shows no symptoms before occurrence.

Prevention is paramount because in the event of a heart attack if the any of the heart muscles are damaged they cannot regrow, also if any of the valves in the heart become stiff and calcified there is no way to restore the flexibility and it must be replaced or repaired.

Data

Data Acquisition and Cleaning

The data set comes from UCI machine learning repository, the full dataset had 76 attributes but the most widely used for machine learning till date was a subset of 14 attributes from the experiments from Cleveland database. The variable that describes the presence of heart disease is called target where 0 is presence of disease and 1 is the absence of the disease. Column names used in the original data set were abbreviated making the variables unclear, hence each column was renamed to the full medical name.The data type of each columns was checked and corrected, so that all columns were numerical. The isnull and sum function was used to determine a count of the total number of missing values in each column all null and missing values were deleted. Outliers were also identified and deleted. Variables are listed below

1. age
2. sex
3. chest pain type (4 values)
   -- Value 0: asymptomatic
   -- Value 1: atypical angina
   -- Value 2: non-anginal pain
   -- Value 3: typical angina
4. resting blood pressure
5. serum cholesterol in mg/dl
6. fasting blood sugar > 120 mg/dl
7. resting electrocardiographic results (values 0,1,2)
   -- Value 0: showing probable or definite left ventricular hypertrophy by Estes' criteria
   -- Value 1: normal
   -- Value 2: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
8. maximum heart rate achieved

9. exercise induced angina
10. oldpeak = ST depression induced by exercise relative to rest
11. the slope of the peak exercise ST segment
    0: down-sloping; 1: flat; 2: up-sloping
12. number of major vessels (0-3) colored by fluoroscopy
13. thalassemia: 3 = normal; 6 = fixed defect; 7 = reversible defect
14. Target 0 = heart disease 1 = no heart disease

Data Exploration

Heart Disease is a wide spread disease and is a leading cause of death, however it is preventable if detected in the very early stages before any damage has been done to the heart vessels. In the exploration of this data we will try to find any correlation between the biomedical traits of the individuals in relation to heart disease. For clarification the terms of the data is listed below:

1. age
2. sex
3. chest pain type (0,1,2,3)
    -- Value 0: asymptomatic
    -- Value 1: atypical angina
    -- Value 2: non-anginal pain
    -- Value 3: typical angina
4. resting blood pressure
5. serum cholesterol in mg/dl
6. fasting blood sugar > 120 mg/dl (fbs > 120 mg/dl = 1, fbs < 120 mg/dl = 0)
7. resting electrocardiographic results (values 0,1,2)
8. maximum heart rate achieved
9. exercise induced angina (1 = yes; 0 = no)
10. old-peak = ST depression induced by exercise relative to rest
11. slope: the slope of the peak exercise ST segment -
    -- 0: down-sloping;
    -- 1: flat;
    -- 2: up-sloping
12. number of major vessels (0-3) colored by fluoroscopy
13. thalassemia: 1 = normal; 2 = fixed defect; 3 = reversible defect
14. target: 0 = heart disease; 1 = no heart disease

There were a total of 10 charts made from the data comparing various medical traits. the first 4 charts are scatter plots where 3 variables could be determined the x variable, y variable and the color difference represented the presence or not of heart disease. While the last 6 charts were made form categorical data where the presence of heart disease was used to classify the other variable.

Pair Plot Table

Resting blood pressure VS Cholesterol

In the first chart (Scatter plot) there is a comparison between resting blood pressure and cholesterol, there is no correlation between resting blood pressure and cholesterol also there is no correlation between these two variables as it relates to heart disease increases.

C**holesterol VS Age**

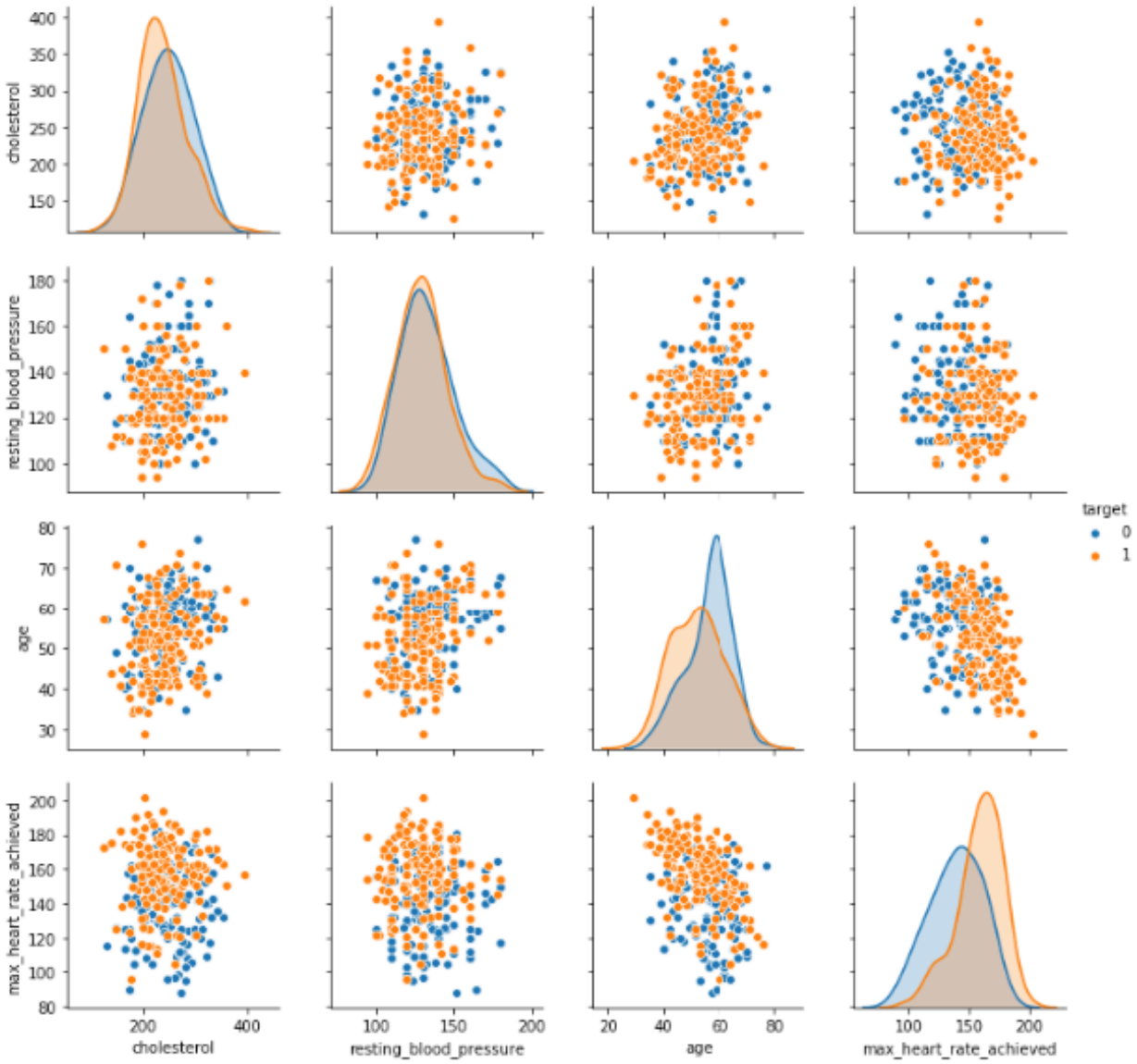
The next chart we are see a slight positive correlation between the age and the cholesterol is also visible that as the older the individuals get the more likely they were to have heart disease. Hence there is a slight increase in the cholesterol levels as the individuals age.

Maximum heart rate achieved VS Age

There is a negative correlation between maximum heart rate and age, so as people age we can see that the heart rate slows down significantly, we can also see by that the number or yellow colored markers, heart disease is more prevalent in individuals as they age. And hence lower heart rate seems to be more prevalent with individuals with heart disease than not.

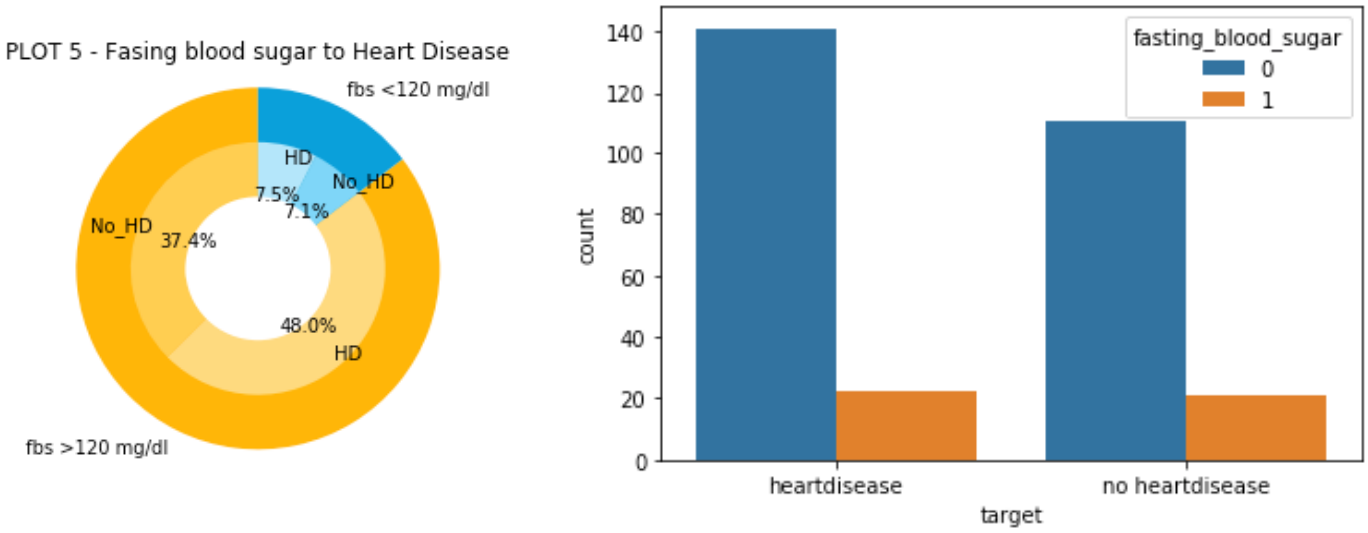
Maximum heart rate achieved VS Resting blood pressure

There is a no correlation between maximum heart rate and resting blood pressure, but as 0 signifies heart disease we see that at all levels of resting blood pressure there are cases of heart disease without any trend however as maximum heart rate increases we see a gentle decrease in heart disease cases. So we can again say that lower heart rate can be associated to heart disease.



Pie chart Fasting blood sugar Percentage levels

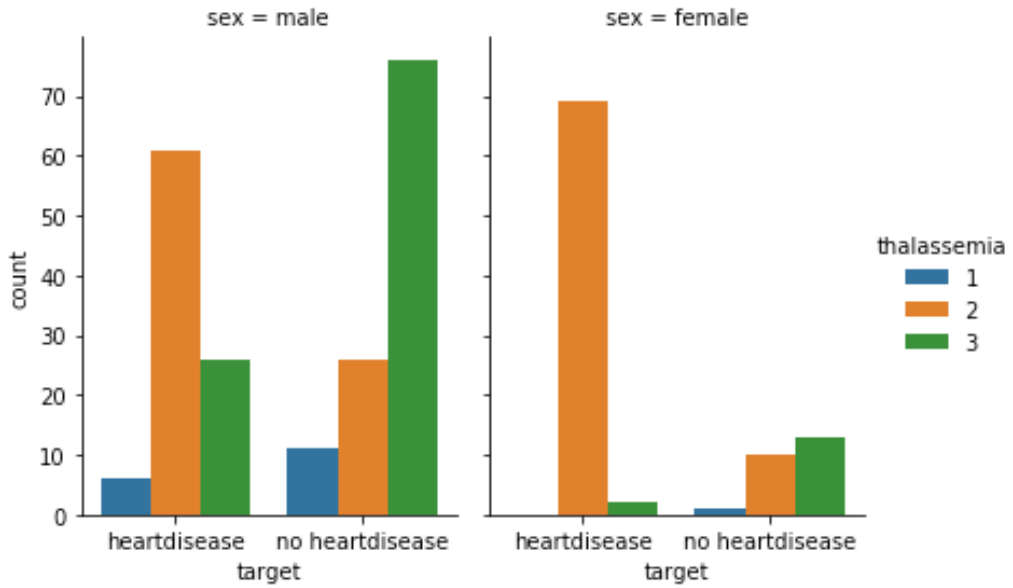
There was a significant number of people had their fasting blood sugar below 120 mg/dl and out of those there was about 56.2% had fasting blood sugar below 120 mg/dl had heart disease, and about 50% of the people with blood sugar above 120 mg/dl had heart disease so we can see about the same percentage of people had heart disease in these two cases so from this chart there doesn't seem to be any correlation between fasting blood sugar and heart disease.



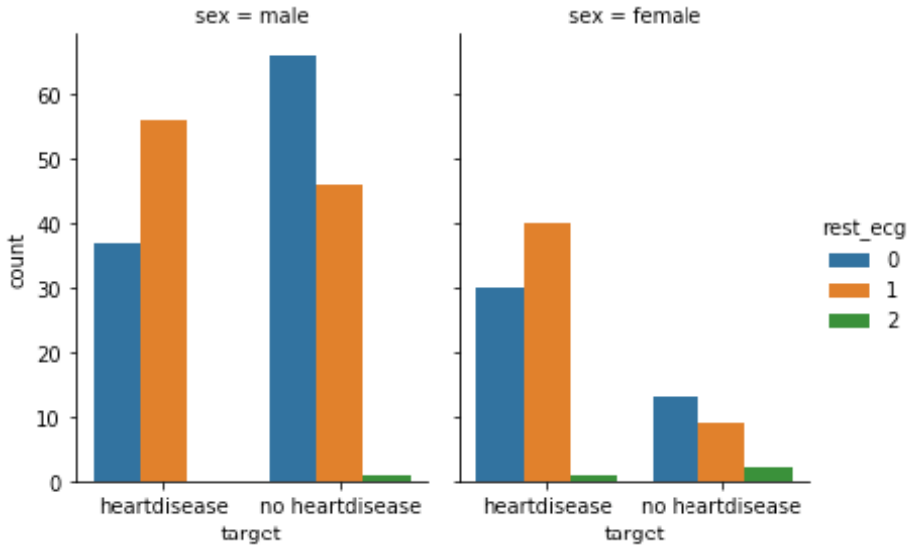
Bar chart

comparing the types of thalassemia with presence of heart disease

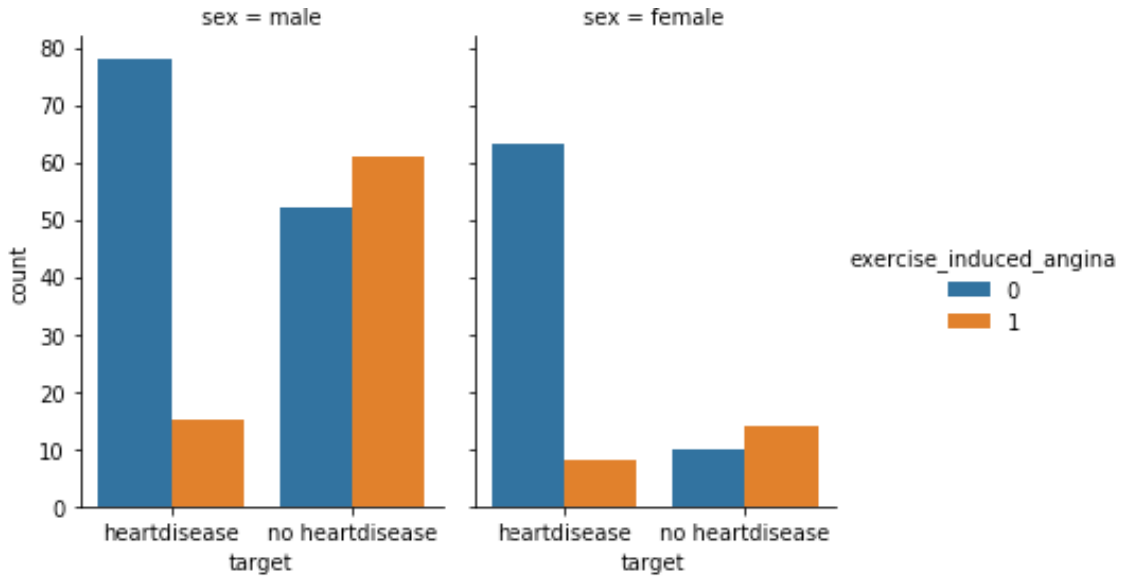
There are 2 types of thalassemia which are either reversible which can be treated and corrected and the fixed type that comes from gene mutation and is not reversible. From the charts we can see that most people that do not have heart disease also do not thalassemia while approximately 64% of the people with heart disease had the reversible form of thalassemia which is a large percentage of people and from this chart we can suggest a correlation between reversible thalassemia and heart disease.



Bar chart comparing Resting electrocardiographic results with presence of heart disease
There is no significant difference in results for the resting electrocardiograph so we can say there is no significant correlation directly between resting electrocardiograph results and heart disease.



Bar chart Comparing Exercise induced agina Percentage levels with presence of heart disease
So we can see that out of tested patients the number of people with heart disease that have exercise induced agina is approximately 77% and the number of patients that have heart disease without exercise induced agina is approximately 30% there fore we can say that from these charts we can assume that there is a higher chance that someone with exercise induced agina to develop or have heart disease.

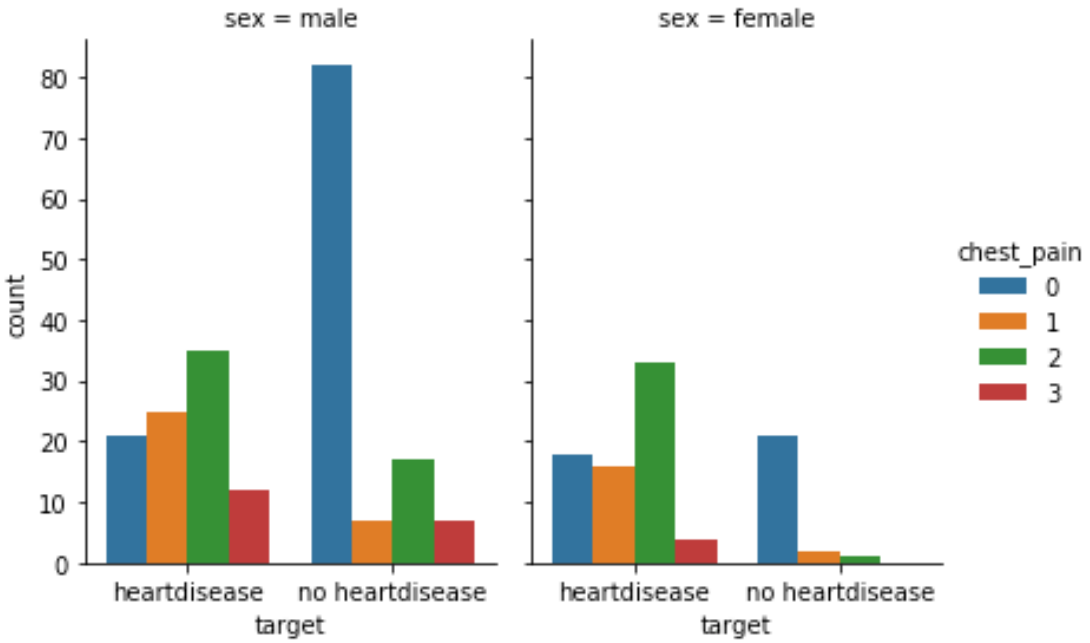


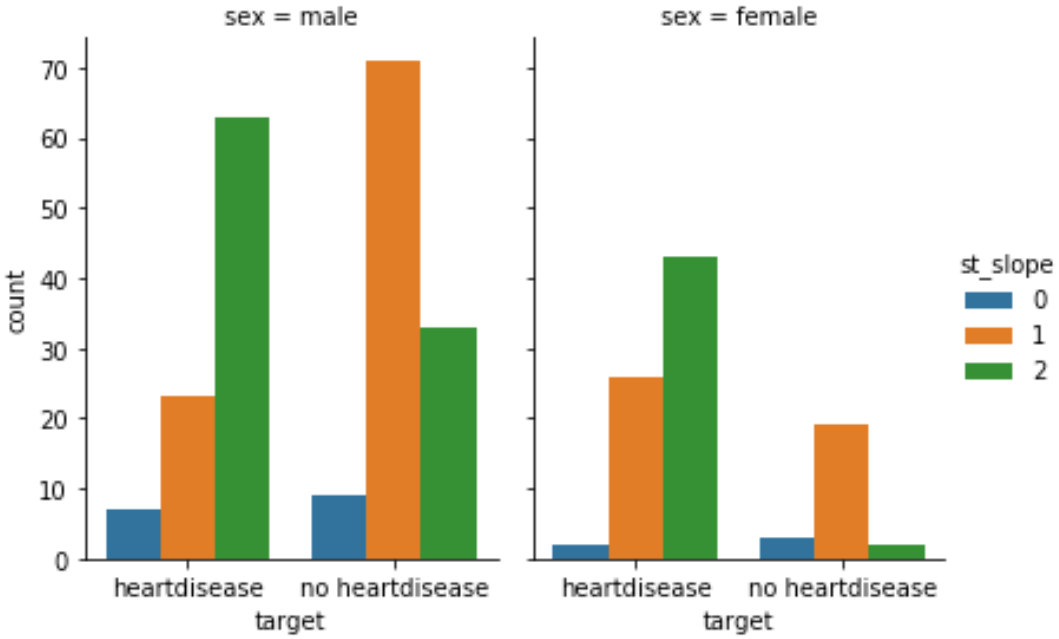
Bar chart comparing chest pain and heart disease

From the results of this graph the most prevalent type of chest pain in the individuals with heart disease was asymptomatic chest pain, this was found in high number, though there was also a sizable number of people without heart disease that also have that type of heart pain so having asymptomatic chest pain may be an indication of heart disease.

Bar chart comparing st_slope and heart disease

The predictions from the chart suggest that people without heart disease have up-sloping ST segment while those with heart disease have a high number of flat ST segment, suggesting a correlation between flat ST segment and heart disease.





Summary

Out of the 14 attributes selected we are trying to determine which of these attribute exhibit any behaviors that can help us predict the presence of heart disease. By using simple charts we will try to find any surface level correlation to heart disease and from our first four analysis. Out of these four characteristics: resting blood pressure, cholesterol, maximum heart rate and age, In plot 4 Comparing maximum heart rate and resting blood pressure, heart disease patients are found spread out through all ranges of resting blood pressure but on maximum heart rate axis they concentrated on the lower region where the maximum heart rates are lower, similarly looking at plot 3 where maximum heart rate is compared to age, as a person gets older we see that their maximum heart rate decreases also we see that heart disease patients increase, showing again correlation between heart disease and maximum heart rate. In plot 1 resting blood pressure to cholesterol there is no correlation in this scatter plot because the patients with the heart disease is evenly mixed in with those without heart disease on both axes. Hence for the individual data from our surface level graph observation a low maximum heart rate is a strong variable in heart disease detection.

As we look into the categorical data from plot 6 to plot 9, there are six attributes that we make observations of. From the graphs we see that there seems to be high correlation between heart disease and the following variables: exercise induced agina (plot 6), reversible form of thalassemia in (plot 7), asymptomatic chest pain (plot 9) and a flat st_slope (plot 10). While fasting blood sugar (plot 5) and resting electrocardiograph (plot 8) does not show strong correlation to heart disease. So in conclusion from the results of this surface level overview of the data there are five attributes that if are found in a patient that predicts a high probability of heart disease which are low maximum heart rate, exercise induced agina, reversible form of thalassemia, asymptomatic chest pain and a flat st_slope.

STATISTICAL TESTING

In relation with the aim of this project we are trying to find out some of the biological traits of individuals who have heart disease. This data set consists of 14 attributes, however the attributes that were tested are in this research are ten as stated:

15. chest pain type (4 values)
16. resting blood pressure
17. serum cholesterol in mg/dl
18. fasting blood sugar > 120 mg/dl
19. resting electrocardiographic results (values 0,1,2)
20. maximum heart rate achieved
21. exercise induced angina
22. oldpeak = ST depression induced by exercise relative to rest
23. the slope of the peak exercise ST segment
24. thal: 3 = normal; 6 = fixed defect; 7 = reversible defect

These attributes consist of continuous variable and categorical variables. The continuous variable were analyzed using the two sample t test for equal means these test have some assumptions that need to be verified to ensure that the test yields correct results. The t test assumptions are:

1. Independent observations
2. Normal distribution
3. Equal variances

Hypothesis Test Procedures and Assumption Checks for Continuous Variables
The procedures used to test the continuous variables are as follows:

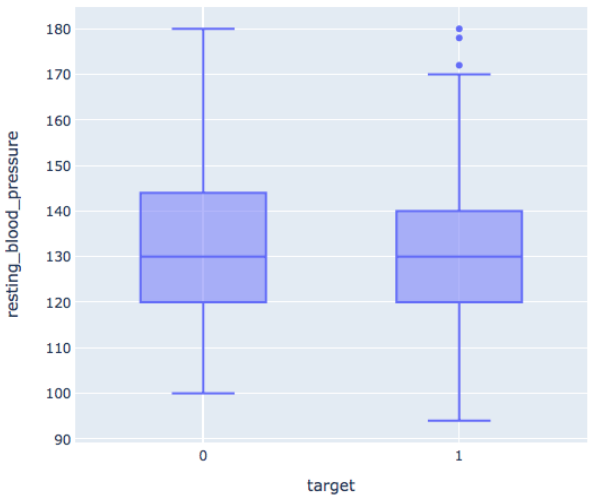
4. The Shapiro-wilk test is run to check for normality of the distributions. However the t test normal distribution assumption is satisfied for a sample size above 30 because of the CLT central limit theorem the sample sizes in this study satisfy this criterion. Where the null hypothesis (Ho) is that the distribution is normally distributed while the alternative hypothesis (Ha) is that the distribution is not normally distributed.

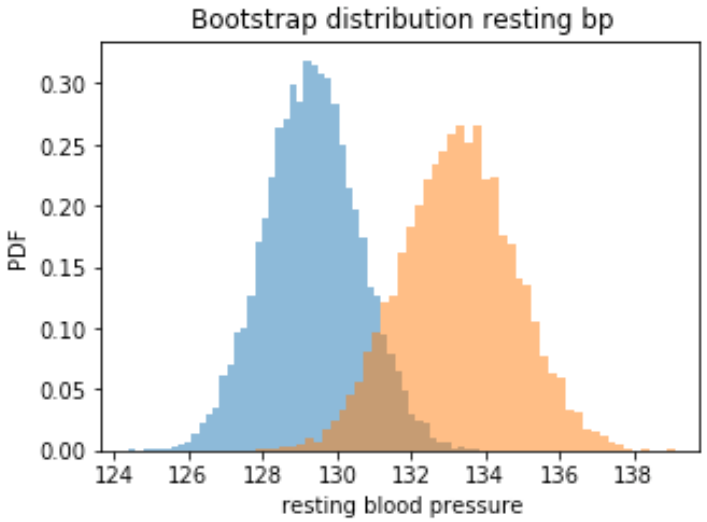
5. Bootstrap method is used to plot the distribution of means from random samples to visualize the central limit theorem that when sample size is above 30 we can assume normal distribution.

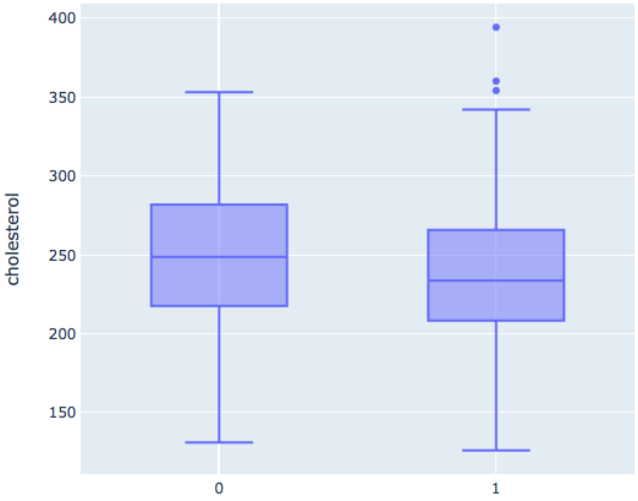
6. Levene's test for equal variances is used to satisfy the third criterion for linear regression. Where the null hypothesis (Ho) is that the distribution has equal variance while the alternative hypothesis (Ha) is that the distribution does not have equal variance. Hence if the p-value for this test is below 0.05 then the null hypothesis is rejected and the distribution is not equal in which case a non parametric test must be run. Here we the non parametric gets used is the Kruskall-Wallis test.

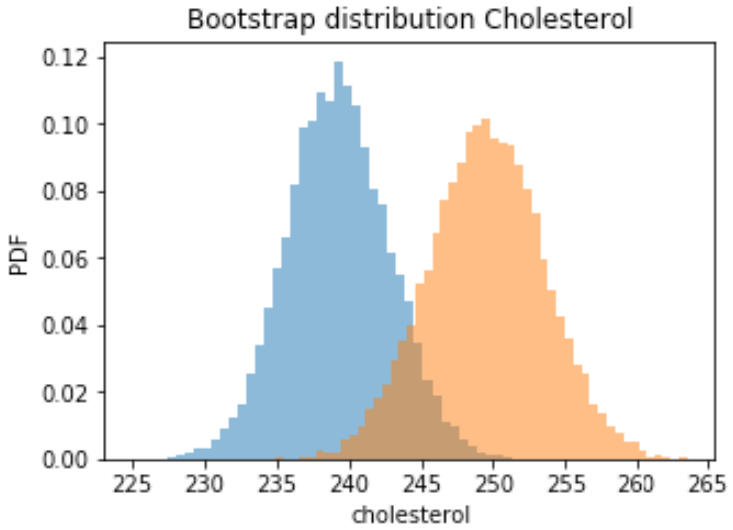
7. The two sample t test for independence is run. With the null hypothesis stating there is no difference between the means of the two samples. The p value is compared to the alpha of 0.05, if less then null hypothesis is rejected.
8. Kuskal-Wallis test for non-parametric distribution With the null hypothesis stating there is no difference between the median of the two samples. The p value is compared to the alpha of 0.05, if less then null hypothesis is rejected. The Kruskall Wallis test is run when the distribution does not have equal variance and the Leven's meaning the p-value for the Levene's test was below 0.05, and the null was rejected.

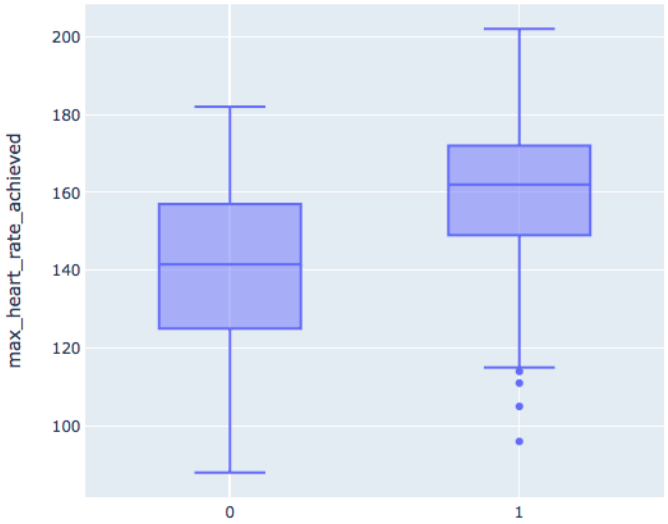
Also in order to see how these variables relate to one another the Pearsons test for correlation was run on the continuous variables to see how the continuous variable interact with one another

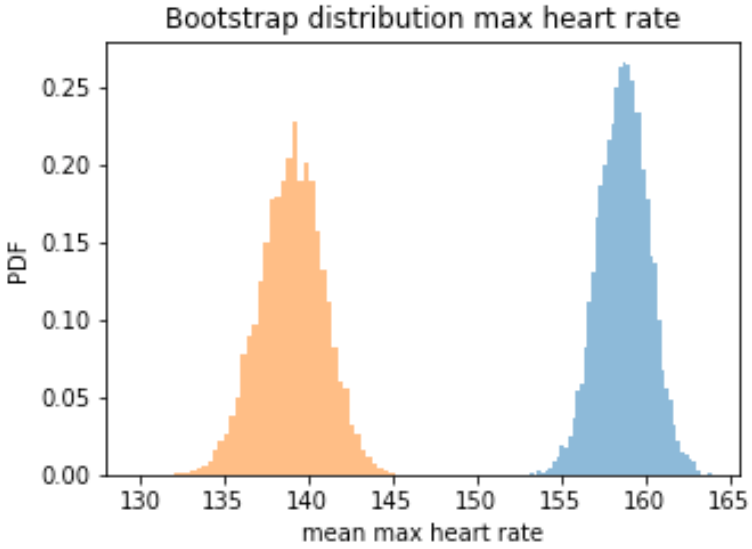


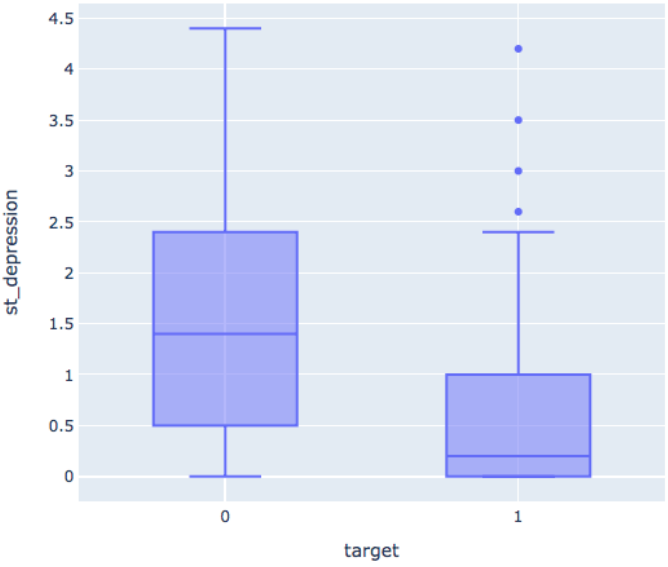


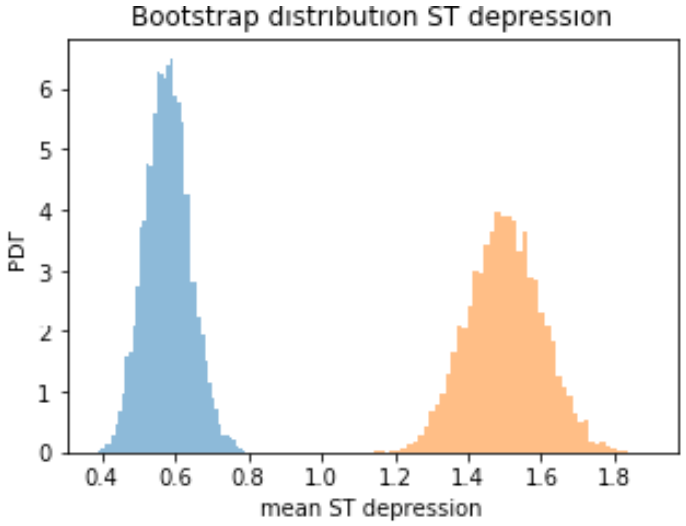












Test for Categorical Variables

The Chi-Square test of independence is used to determine if there is a significant relationship between two nominal (categorical) variables. In this study we are checking for the relationship between each individual categorical variable and heart disease to

see which variables may have have an effect on heart disease. With the null hypothesis stating there is no independence between the variable in question and heart disease.

Results

There were 4 continuous variables : cholesterol, resting blood pressure, maximum heart rate, ST depression and 6 categorical attributes tested : fasting blood sugar, exercise induced aging, thalassemia, resting electrocardiograph, chest pain, and the slope of the peak exercise ST segment.

All three continuous variables cholesterol, resting blood pressure, maximum heart rate and ST depression are found to have non normal distributions hence the bootstrap method was used to visualize the central limit theorem, which were normally distributed allowing us to use the two sample t test.

Using Levene's test we determined that the distribution of people with heart disease and those without heart disease cholesterol levels had equal variance this was the same for resting blood pressure this is significant because the two sample t test and ANOVA test assumptions will be violated and a different test would need to be run on these variables if the variance are unequal. While the distribution of people with heart disease and those that do not have it, for maximum heart rate and ST depression do not have equal variance therefore another test would have to be run to check if they are individually significant related to heart disease. In this study the Kruskal-Wallis test for non-parametric distribution is run to see if there is significant difference in the medians of the distributions.

From the two sample t test we determine that for cholesterol pvalue of 0.107 is greater than the standard alpha of 0.05 so we fail to reject the null and we can say there is not enough evidence for significant difference between the cholesterol distribution of the heart disease sample and non-heart disease sample. We confirm this by doing a ANOVA test which uses the same assumptions and get a p value 0.058. The t test on resting blood pressure has a pvalue 0.1 which is higher than the average alpha of 0.05 gain we fail to reject the null meaning there is not enough evidence to suggest that the means of the sample with or without heart disease has no significant difference. For the variables maximum heart rate and ST depression the Kruskal wall test was run and the pvalue for them respectively were 0.00 and 0.00 so they were both below the standard alpha of 0.05 hence we would can state that there is evidence to suggest a difference in median values for the sample with and with out heart disease for maximum heart rate and ST depression. In other words the above tests means that individually the cholesterol and resting blood pressure are not significant in checking for heart disease by themselves, while maximum heart rate and ST depression from this study suggests that they may be significant in identifying heart disease individually by themselves.

There was also correlation matrix between these four variable and age there was no real strong correlation between these variables as the strongest correlation was between age and maximum heart rate which was negative correlation of 0.4.

The chi square test for independence is carried on all six categorical variables fasting blood sugar, exercise induced agina, thalassemia, resting electrocardiograph, chest pain, and the slope of the peak exercise ST segment, as they are compared to heart disease to see if there is any relationship between them.

Fasting blood sugar is the only variable that when alone has no evidence to suggest there is a relationship between itself and heart disease with a pvalue of 0.78, which is greater than the alpha which is set to 0.01 hence we fail to reject the null hypothesis which means we cannot prove there any association or relationship between those three variable and heart disease.

While ST segment, thalassemia, exercise induced angina, chest pain variables have pvalues of 0.00 while resting electrocardiograph has a pvalue of 0.005 and are lower than the set alpha of 0.01 so there is a high probability that there is association between these variables and heart disease. Disclaimer these variables are being assessed according to their individual effects on heart disease, however the variable that were found insignificant may become significant when being tested in the presence of other variables for its effect on heart disease.

HEART DISEASE CLASSIFICATION ANALYSIS

The purpose of this study is to use machine learning to help identify some of the factors that can help us identify the presence of heart disease in a patient. There are several machine learning classifications that can be use to help us detect out of the variable we are testing which more accurately predicts the presence of heart disease. In this study we look at the following classification methods:

1. Knn Classifier
2. Naive bayes classifier
3. Logistic Regression classifier
4. SVM classifier
5. Decision tree classifier
6. Random forest classifier
7. Gradient boost classifier
8. Bagging classifier

Analysis result terms

Accuracy

The accuracy of a machine learning classification algorithm is one way to measure how often the algorithm classifies a data point correctly. Accuracy is the number of correctly predicted data points out of all the data points. More formally, it is defined as the number of true positives and true negatives divided by the number of true positives, true negatives, false positives, and false negatives.

$$Accuracy = \frac{TrueNegatives + TruePositive}{TruePositive + FalsePositive + TrueNegative + FalseNegative}$$

Precision
Precision evaluates how precise a model is in predicting positive labels. Precision answers the question, out of the number of times a model predicted positive, how often was it correct?
Also precision = true positive/actual results

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive}$$

Recall
Recall calculates the percentage of actual positives a model correctly identified (True Positive). When the cost of a false negative is high, you should use recall.
Also recall = true positive/predicted results

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative}$$

Suport
Support may be defined as the number of samples of the true response that lies in each class of target values.

F1-Score
F1-score, which takes both precision and recall into account to ultimately measure the accuracy of the model. The F1 score gives more weight to false negatives and false positives while not letting large numbers of true negatives influence your score.

$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$

Knn Classifier
K-nearest neighbors (KNN) algorithm uses 'feature similarity' to predict the values of new datapoints which further means that the new data point will be assigned a value based on how closely it matches the points in the training set.

	precision	recall	f1-score	support
0	0.91	0.83	0.87	35
1	0.79	0.88	0.84	26
accuracy			0.85	61
macro avg	0.85	0.86	0.85	61
weighted avg	0.86	0.85	0.85	61

Accuracy = 0.85246

Naive bayes classifier

Using Bayes theorem, we can find the probability of A happening, given that B has occurred. Here, B is the evidence and A is the hypothesis. The assumption made here is that the predictors/features are independent. That is presence of one particular feature does not affect the other. Hence it is called naive.

	precision	recall	f1-score	support
0	0.82	0.80	0.81	35
1	0.74	0.77	0.75	26
accuracy			0.79	61
macro avg	0.78	0.78	0.78	61
weighted avg	0.79	0.79	0.79	61

Accuracy = 0.78689

Logistic Regression Classifier

Logistic regression is a classification algorithm. It is used to predict a binary outcome based on a set of independent variables.

	precision	recall	f1-score	support
0	0.87	0.74	0.80	35
1	0.71	0.85	0.77	26
accuracy			0.79	61
macro avg	0.79	0.79	0.79	61
weighted avg	0.80	0.79	0.79	61

Accuracy = 0.78689

SVM classifier

The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space(N — the number of features) that distinctly classifies the data points.

	precision	recall	f1-score	support
0	0.83	0.71	0.77	35
1	0.68	0.81	0.74	26
accuracy			0.75	61
macro avg	0.76	0.76	0.75	61
weighted avg	0.77	0.75	0.76	61

Accuracy = 0.7541

Decision Tree Classifier

The goal of using a Decision Tree is to create a training model that can use to predict the class or value of the target variable by learning simple decision rules inferred from prior data(training data)

	precision	recall	f1-score	support
0	0.88	0.63	0.73	35
1	0.64	0.88	0.74	26
accuracy			0.74	61
macro avg	0.76	0.76	0.74	61
weighted avg	0.78	0.74	0.74	61

Accuracy = 0.7377

Random Forest Classifier

It is an ensemble tree-based learning algorithm. The Random Forest Classifier is a set of decision trees from randomly selected subset of training set. It aggregates the votes from different decision trees to decide the final class of the test object.

	precision	recall	f1-score	support
0	0.87	0.77	0.82	35
1	0.73	0.85	0.79	26
accuracy			0.80	61
macro avg	0.80	0.81	0.80	61
weighted avg	0.81	0.80	0.80	61

Accuracy = 0.80328

Gradient Boost

Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function.

	precision	recall	f1-score	support
0	0.84	0.77	0.81	35
1	0.72	0.81	0.76	26
accuracy			0.79	61
macro avg	0.78	0.79	0.78	61
weighted avg	0.79	0.79	0.79	61

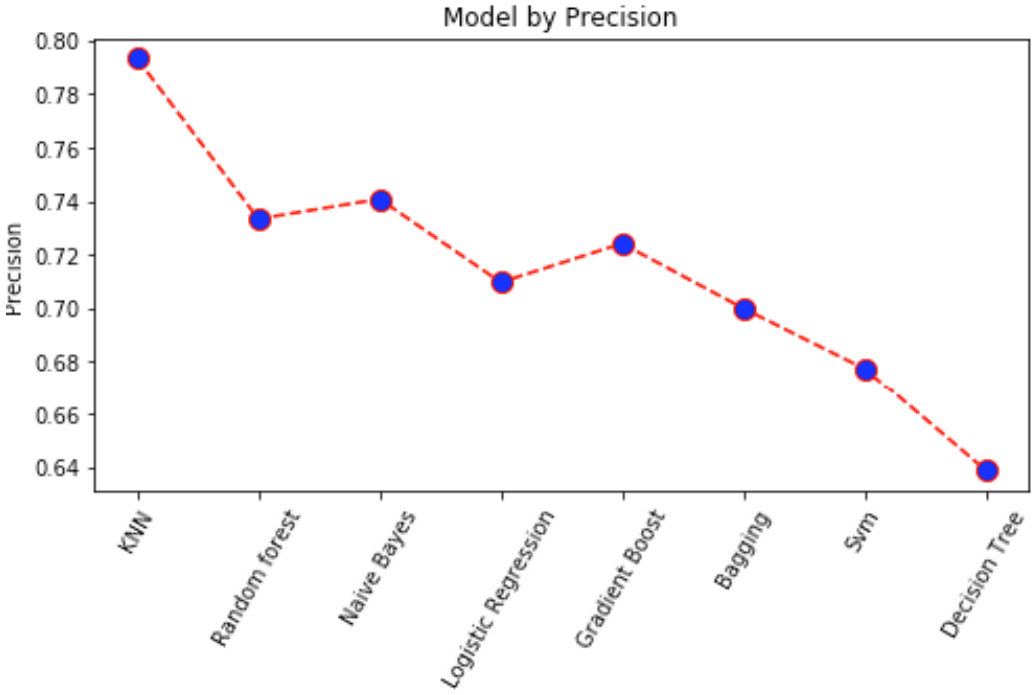
Accuracy = 0.78689

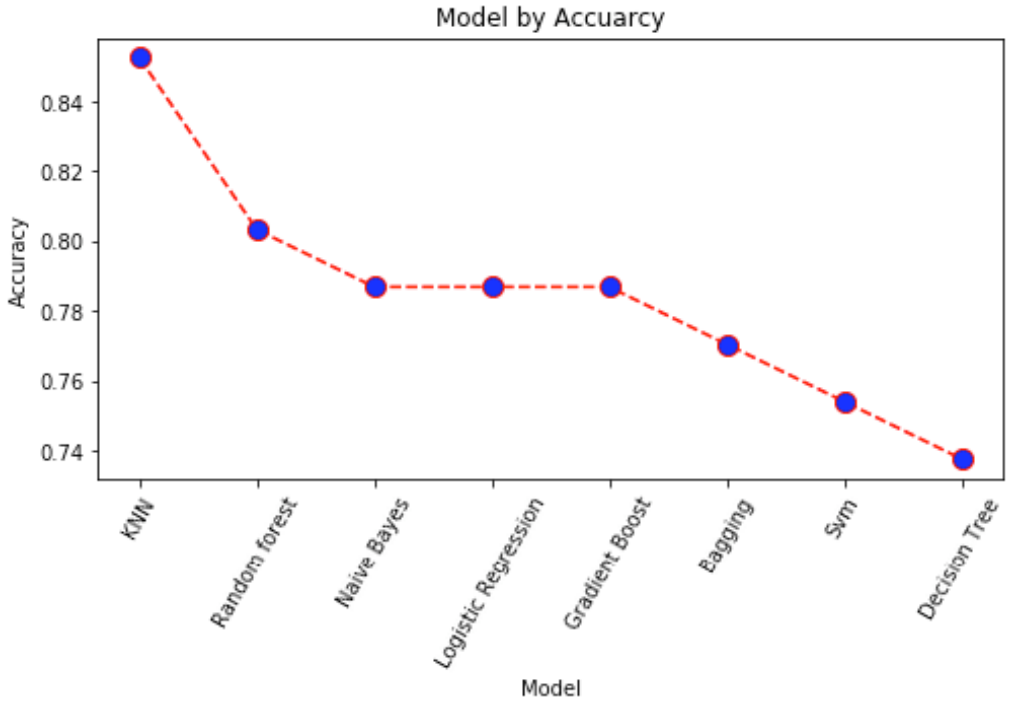
Bagging Classification

Bagging is a simple ensembling technique in which we build many *independent* predictors/ models/learners and combine them using some model averaging techniques. (e.g. weighted average, majority vote or normal average)

	precision	recall	f1-score	support
0	0.84	0.74	0.79	35
1	0.70	0.81	0.75	26
accuracy			0.77	61
macro avg	0.77	0.78	0.77	61
weighted avg	0.78	0.77	0.77	61

Accuracy = 0.77049





target: 0 = heart disease; 1 = no heart disease

The above diagrams show the precision and accuracy of the of each the classification techniques. For the diagrams we can see that knn and random forest were the model with the highest values of accuracy and precision, with knn accuracy at 0.85 and its precision for guessing the presence of heart disease at 0.84, while random forest accuracy is at 0.80 and the precision for guessing the presence of heart disease at 0.87.

Permutation Importance

In this study we use the Permutation importance to tell us which variables in the model are most significant in deterring the heart disease.
The permutation importance feature is the decrease in a model score when a single feature value is randomly shuffled. This procedure breaks the relationship between the feature and the target, thus the drop in the model score is indicative of how much the model depends on the feature.

Permutation importance knn

Weight	Feature
0.0689 ± 0.0700	num_major_vessels
0.0656 ± 0.0508	thalassemia
0.0590 ± 0.0334	st_depression
0.0557 ± 0.0445	sex_1
0.0557 ± 0.0393	chest_pain
0.0459 ± 0.0382	max_heart_rate_achieved
0.0393 ± 0.0491	age
0.0295 ± 0.0382	rest_ecg
0.0262 ± 0.0161	cholesterol
0.0230 ± 0.0334	resting_blood_pressure
0.0131 ± 0.0321	st_slope
0.0066 ± 0.0642	exercise_induced_angina
0.0033 ± 0.0131	fasting_blood_sugar

Permutation importance Random Forest

Weight	Feature
0.0885 ± 0.0572	st_depression
0.0361 ± 0.0382	exercise_induced_angina
0.0164 ± 0.0359	chest_pain
0 ± 0.0000	thalassemia
0 ± 0.0000	st_slope
0 ± 0.0000	max_heart_rate_achieved
0 ± 0.0000	cholesterol
0 ± 0.0000	resting_blood_pressure
0 ± 0.0000	age
-0.0033 ± 0.0382	sex_1
-0.0066 ± 0.0262	num_major_vessels
-0.0131 ± 0.0131	fasting_blood_sugar
-0.0361 ± 0.0131	rest_ecg

From the above diagrams we can see that in the knn has number of major blood vessels, thalassemia, St depression, Sex, chest pain all have values above 0.05 weight significance. While for random forest has St depression, exercise induced angina and chest pain have the most significant factors determining heart disease.

Forward selection

Forward selection is another method used in this study to determine the most effective variables. Forward selection is a type of stepwise regression which begins with an empty model and adds in variables one by one. In each forward step, you add the one variable that gives the single best improvement to your model. We would use forward selection in selecting the best variable from the knn and random forest classifiers. The knn results were 9 variables giving an accuracy of 0.89 as accessed with the forward selection, while the random forest classifier also had a selection of 9 variable with accuracy of 0.83. The accuracy values were confirmed by putting the selected variables back into the classifier to see if they produced better results than the full model. And the final variable did perform better when using the forward selection to predict as the model with 9 significant variables for knn had an final accuracy of 0.88 while the model with 13 variable had an accuracy of 0.85 and the random

forest classification with 9 significant variables had an accuracy of 0.83 while the model with 13 variable had an accuracy of 0.80. Also the correlation coefficients of these variable are checked ignorer to ensure they are not highly correlated because, it becomes difficult for the model to estimate the relationship between each independent variable and the dependent variable independently because the independent variables tend to change in unison. From the heat maps and correlation tables we see that non of the final selected variables are highly correlated. New would consider any value above 0.5 highly correlated all variable in Both models are below those values.

Results

From these findings we the most effective determinants of heart disease from knn classifier sex, ,thalassemia, number of major vessels ,St_slope ,St_depression, exercise induced angina , resting electrocardiographic, fasting blood sugar, chest pain, and the variables that were most significant for random forest are sex, cholesterol, chest pain, fasting_ blood sugar, St slope, number of major blood vessels, thalassemia, resting electrocardiographic, St depression. All the variable are the same except that knn classifier has cholesterol while the random forest classifier uses exercise induced angina.

References

1. https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_with_python_knn_algorithm_finding_nearest_neighbors.htm
2. https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c
3. https://careerfoundry.com/en/blog/data-analytics/what-is-logistic-regression/
4. https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47
5. https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html
6. https://towardsdatascience.com/random-forest-classification-and-its-implementation-d5d840dbead0
7. https://medium.com/mlreview/gradient-boosting-from-scratch-1e317ae4587d
8. https://en.wikipedia.org/wiki/Gradient_boosting
9. https://deepai.org/machine-learning-glossary-and-terms/accuracy-error-rate
10. https://medium.com/@erika.dauria/accuracy-recall-precision-80a5b6cbd28d
11. https://towardsdatascience.com/whats-the-deal-with-accuracy-precision-recall-and-f1-f5d8b4db1021
12. https://scikit-learn.org/stable/modules/permutation_importance.html