Heart disease is the leading cause of death in the united states and it is estimated 655,000 Americans die each year from heart disease. The aims of the research were to investigate whether patients with heart disease can be distinguished from patients without heart disease. Patients with heart disease are more likely to have a heart attack and the research will investigate if there are variables that can explain why some people get heart disease. This could assist with heart disease prevention strategies and therefore reduce the number of deaths from heart disease.

**Method**

Dataset was downloaded from Kaggle and Jupyter notebook was used: Pandas, NumPy, Matplotlib and Scikit-learn packages were used for data preparation, visualization, and machine learning. The dataset consisted of 303 data points.

**Features and target**

The features included age, sex, chest pain, cholesterol (mg/DI), maximum heart rate achieved, old peak (ST depression induced from exercise), slope of peak exercise ST segment and number of major vessels coloured by fluoroscopy. The target was heart disease coded as 0: no heart disease, 1:has heart disease. Full descriptions of each variable can be found at <https://archive.ics.uci.edu/ml/datasets/Heart+Disease>

**Exploratory data analysis**

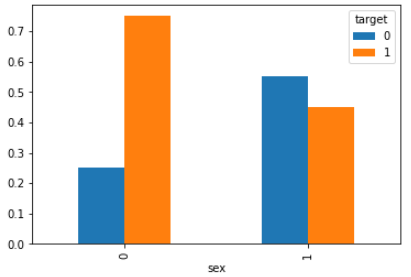
All categorical features were visualized with each on the x-axis and the target on the y-axis using bar charts. The y-axis represented the percentage of people who had heart disease and did not have heart disease for multiple categories (e.g., people who were female and had heart disease and people who were female and did not have heart disease).

Bins were created for the continuous variables to show visualisations with the target. For example, age had 5 bins: 25-40, 40-50, 50-60, 60-70 and 70-80.

**Machine learning**

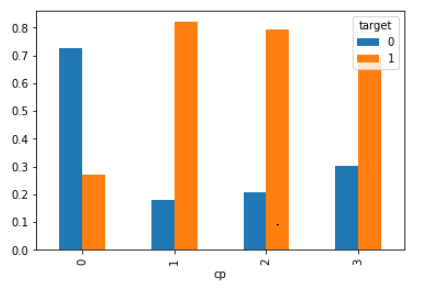
A random forest classifier was used to fit the data and 5-fold cross validation was used. The target classes were reasonably balanced thus classification accuracy was the metric recorded. A grid search was implemented to find the optimal number of trees to build in the model between 1-100 and an elbow graph was visualised to determine this. A second random forest model was also built with sex and slope removed considering these features had the least importance in the first model determined by the random forest classifier’s feature importance attribute.

**Results**



*Figure 1*. Bar chart depicting sex on the x-axis and heart disease on the y-axis

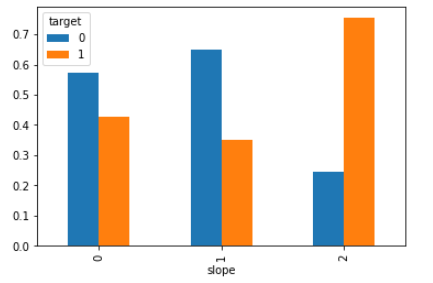
As can be seen from figure 1, a large proportion of females had heart disease whereas heart disease was just under 50% for the males.



*Figure 2*. Bar chart depicting chest pain on the x-axis and heart disease on the

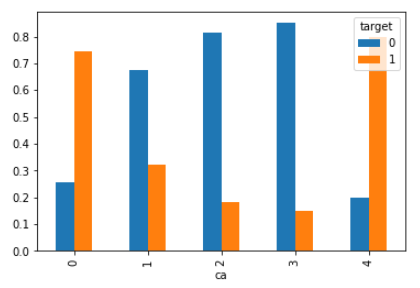
y-axis

Figure 2 shows a large proportion of people who had chest pain whether it is mild, moderate, and severe had heart disease



*Figure 3.* Bar chart depicting slope of peak exercise at ST segment on x-axis

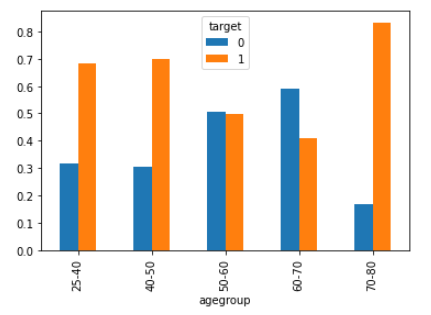
and heart disease on y-axis

 figure 3 depicts upsloping ST segment is similar between people with heart disease and people without. A larger proportion of people without heart disease have a flat ST segment Whereas most people who had heart disease also had a down sloping ST segment.

*Figure 4.* Bar chart depicting number of major vessels coloured by

fluoroscopy on the x-axis and heart disease on the y-axis

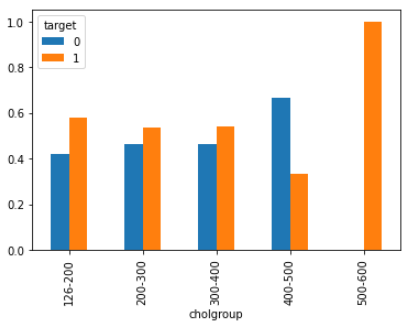
As can be seen from figure 4, people who have no major vessels coloured by fluoroscopy tended to be people who had heart disease and people who had 4 major vessels coloured by fluoroscopy also tended to be people with heart disease. Between 1-3 categories was mainly people who did not have heart disease.



*Figure 5.* Bar chart depicting age groups on x-axis and heart disease on y-

axis

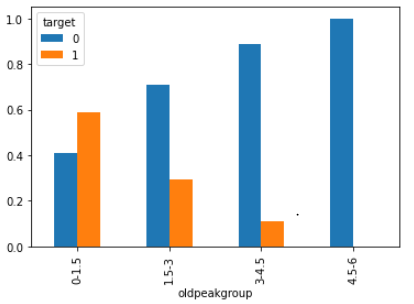
Figure 5 shows people between 25-50 were around 70% more likely to have heart disease while people between 50-60 were around the same proportion. People who were between 60-70 were less likely to have heart disease while most people aged between 70-80 had heart disease.



*Figure 6.* Bar chart depicting cholesterol on x-axis and heart

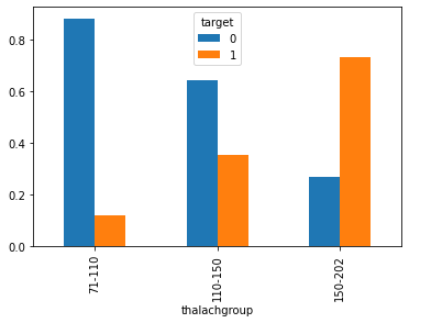
disease on y-axis

Figure 6 shows when cholesterol is between 126-400, the proportion of heart disease and no heart disease is similar. Majority of people who had cholesterol between 400-500 did not have heart disease. Every person who had cholesterol greater than 500 had heart disease.



*Figure 7.* Bar chart depicting old peak group on x-axis and heart disease on y-

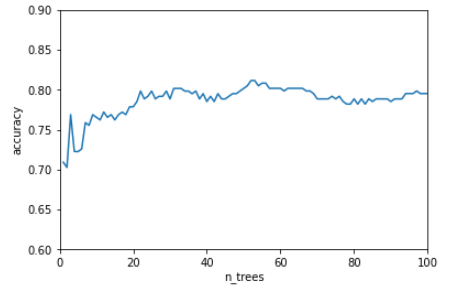
axis

 As can be seen from figure 7, when old peak is between 0-1.5, majority of people have heart disease while 1.5-3 the majority is people without heart disease. Between 3.4-5, 90% of people do not have heart disease while 4.5 or more only consists of people without heart disease.

*Figure 8.* Bar chart depicting maximum heart rate achieved on x-axis and

heart disease on y-axis

Figure 8 shows most people who had a maximum heart rate between 71-110 did not have heart disease. Around 60% of people with maximum heart rate between 110-150 did not have heart disease. Majority of people who had maximum heart rate greater than 150 had heart disease.



*Figure 9.* Graph shows random forest classification accuracy on the y-axis

and the x-axis represents number of trees used to build the model in the range

of 1-100.

Figure 9 shows when the n\_trees parameter is approximately 25, it levels off and classification accuracy is around 80%.

|  |  |  |
| --- | --- | --- |
| Model | N\_tree | Classification accuracy |
| Random forest (8 features) | 25 | 80% |
| Random forest (6 features) | 25 | 84% |

*Table 1.* Comparison of two random forest models

Table 1 shows the random forest model with 6 features (excluded slope and sex) had higher classification accuracy (84%) compared to the random forest model that had 8 features (80%).

Discussion

The aims of the research were to explore if a set of features could predict heart disease and this was confirmed. Combining age, sex, cholesterol, chest pain, slope of peak exercise at ST segment, maximum heart rate achieved, old peak and number of major vessels coloured by fluoroscopy resulted in high classification accuracy at predicting heart disease. Further simplifying the model by removing sex and slope from the feature set increased classification accuracy from 80 to 84%. This makes sense considering sex and slope had the least importance in the first model determined by the random forest classifier’s feature importance attribute. One limitation in the study was only accuracy metric was used to evaluate the random forest models. The classes were reasonably balanced but precision and recall metrics may have added more insight into how good the models were. In conclusion, the combination of age, cholesterol, chest pain, maximum heart rate achieved, old peak and number of major vessels coloured by fluoroscopy predicted heart disease with high classification accuracy. Therefore, close inspection of these variables could assist with heart disease prevention strategies and assist treatment.