Practical Data Science: Assignment 3 – Predicting Heart Disease

Tim Kirkbride (s3650791) and Joshua Grosman (s3494389)

s3494389@student.rmit.edu.au ; s3650791@student.rmit.edu.au

RMIT University

7th of May, 2018

Abstract:

**Introduction:**

This report is concerned with developing the most appropriate classification model to predicting the absence or presence of heart disease in individuals based on a selection of descriptive features. The first stage of this report will involve data preprocessing, where all relevant features will be unpacked and their relevance to the task at hand will be determined. Relevant features will then be explored through visualization. Here univariate visualizations will be employed to highlight the distributions of the different features, while multivariate visualization will also be used for portraying any notable relationships between features.

The second stage of this report will be concerned with the classification analysis. In this stage, several different supervised machine learning techniques will be applied to the data, and the best will ultimately be determined based on its accuracy. Classification techniques which may be considered appropriate for such an analysis include information-based, similarity-based and probability-based learning algorithms (). Considering this, one algorithm from each learning type will be implemented on the data to determine the most effective type. Specifically, algorithms to be considered in the analysis will include the k-nearest neighbours, decision tree, and naïve Bayes learning algorithms. Parameters relevant to each technique will be explored as appropriate to determine the best model of each specific learning type. Ultimately, the best models of each type will be compared to each other to determine the best model overall for predicting the presence of heart disease. Comparison of models will be based primarily on prediction accuracy.

**Methodology:**

The data used in this analysis was obtained from the UCI Machine Learning Repository and can be accessed using the link below. The dataset contained 270 observations, with each one corresponding to a different individual. There were 13 descriptive features in the dataset and one target feature.

[http://archive.ics.uci.edu/ml/datasets/Statlog+%28Heart%29](http://archive.ics.uci.edu/ml/datasets/Statlog+(Heart))

*Target Feature –*

The target feature in this analysis was the “heart\_disease” feature which was comprised of two levels. These are described as follows:

* Absent: If the individual does not have heart disease
* Present: If the individual does have heart disease

In the data, the absent and present levels of the target feature were denoted by 1 and 2 respectively. Evidently, the target feature is only comprised of two levels, and as such this analysis would be a binary classification problem.

*Descriptive Features –*

The names and data types of the descriptive features is highlighted below:

* **Age:** continuous
* **Sex:** nominal (Female, Male)
* **Chest\_pain:** nominal (Typical, Atypical, Non\_Anginal, Asymptomatic)
* **Resting\_bp:** continuous
* **Cholesterol:** continuous
* **High\_blood\_sugar:** nomial (False, True)
* **Resting\_ecg:** nominal (Normal, Abnormal, LV\_Growth)
* **Max\_heart\_rate:** continuous
* **Angina:** nominal (False, True)
* **Oldpeak:** continuous
* **ST\_slope:** ordinal (Downsloping, Flat, Upsloping)
* **Major\_vessels:** continuous
* **Thalassemia:** nominal (Normal, Fixed\_Defect, Reversable\_Defect)

Many of these features are clearly self-explanatory, while some of the more convoluted features are described in more detail on the website. Resting\_ecg refers to resting electrocardiographic results, while the LV\_Growth levels within this feature represents a left ventricular mass. Angina is described as a temporary discomfort/pain in the chest following physical activity. Oldpeak is the degree of ST depression, which is where the ST segment trace in an electrocardiogram is very low (). Furthermore, ST\_slope refers to the slope of the ST depression (). Finally, thalassemia may be briefly described as an inherited blood disorder.

Initially, all levels of nominal/ordinal features were in quantified state and were represented by numbers. For example, for the sex feature, the female and male levels were represented as 0 and 1, respectively. Although this format was desirable for classification analysis, it proved inefficient for data exploration. As such, a new dataset with appropriately named levels was also created to account for this, while the original dataset was also retained.

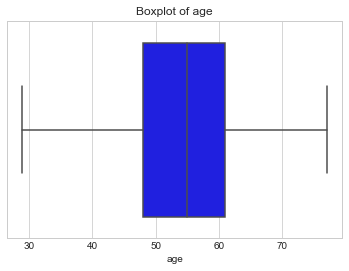
The new descriptive dataset was then explored through visualization. For continuous variables, boxplots were used to highlight differences between the levels of the target feature with respect to each feature. Bar charts were used for the categorical variables to show the differences in observation across the levels of each feature. Subsequently, multivariate visualizations were generated for different pairings of the descriptive features, as well as the target feature. These visualizations included multivariate boxplots, scatter plots and proportional bar charts.

Following this, the classification analysis was conducted. Here, the original data was used. The data was split into training and testing segments using a 75:25 ratio. As such, the training data comprised 216 observations, while the test data had 54 observations. A k-nearest neighbours algorithm was then applied to the segmented data. A variety of models were constructed using various combinations of the algorithm’s parameters in order to identify the most ideal model. A decision tree algorithm was also used and again models were constructed using different parameters. Finally, models using the naïve Bayes algorithm were constructed with different combinations of parameters. The best model within each learning type was identified based on the relevant models’ confusion matrices, classification error rates, precision, recall and F1-scores. Ultimately, these models were then compared to each other using the same metrics in order to identify the best model overall.

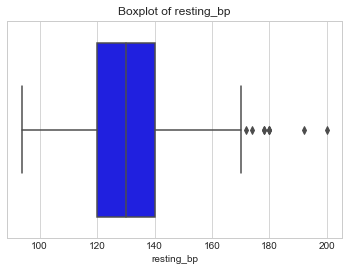
**Results:**

*Data Visualization –*

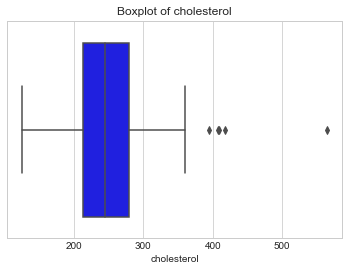
* Univariate



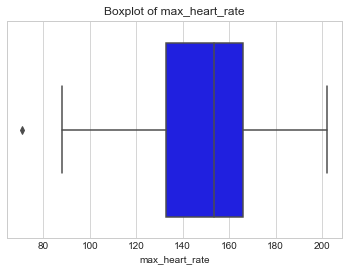
As we can see here the dataset that we are studying is made up of mostly older people with an average age of around 55. The width of the box plot shows that the majority of observed cases will feature people aged in their late 40s to their early 60s with some extreme cases of people in their 30s to people in their 80s.



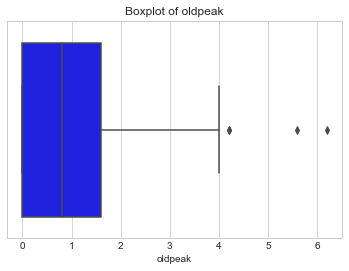
With this boxplot we can see that the majority of people in this data set have a resting blood pressure between 120-140. According to the heart foundation, this is the expected range for most people, so anything outside of this range should be considered unhealthy and will be tested for its relationship with heart disease.



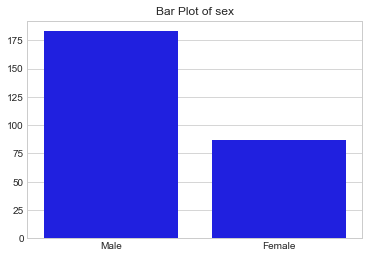
This boxplot shows the distribution of cholesterol levels of people in the dataset. The majority of observations are between 200-300 with some exceptions outside this range. A cholesterol level of less than 200 is considered healthy, so the majority of people in this study can be said to have high cholesterol. This is a major cause of heart disease, heart attack, and stroke so in our predictions, attention must be paid to the patients cholesterol level.



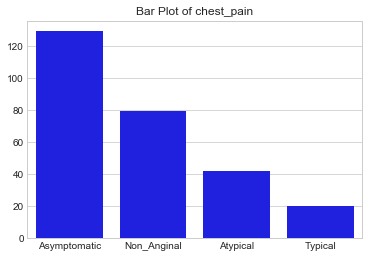
We can see in the visualisation above that the majority of patients maximum heart rates were centered around the 140-160 range with a large variation between 90-200.



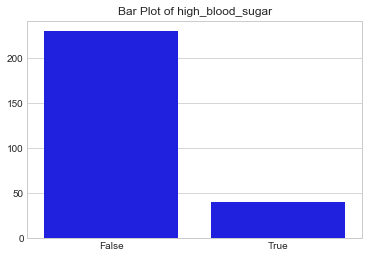
The oldpeak in the above box plot refers to the extent of the depression in an ECG when the patient has been exercising. It can be seen that mostly this depression does not exist (0) or is minor (<2) with some larger outliers.



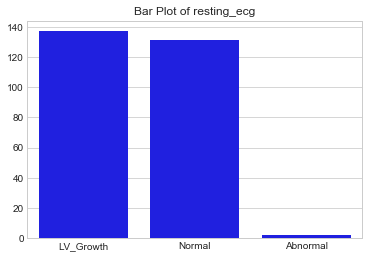
This bar plot shows that the vast majority of people in the study were male. Almost 2/3 of the cases featured males with a 1/3 featuring females.



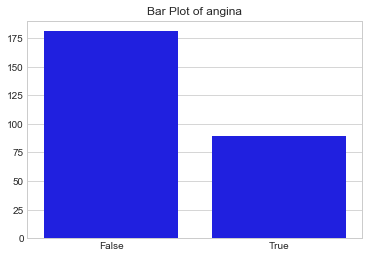
This bar chart shows the distribution of chest pain Asymptomatic Angina, Atypical Angina, Typical Angina and Non-Aginal pain. Here we can see that the most common type of pain is Asymptomatic angina.



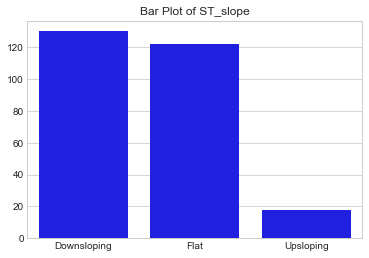
The above bar chart shows that the vast majority of patients did not have high blood sugar. High blood sugar is an indicator of heart disease as there is a positive relationship between blood sugar and heart disease (i.e. as blood sugar increases so does the risk of heart disease).



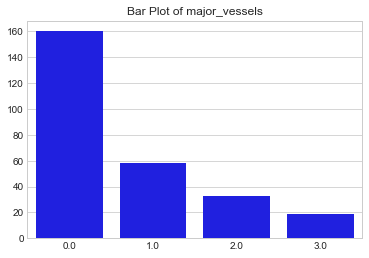
The resting ecg is a Electrical Cardio Graph result of a heart at rest. Here we can see that the main results of this test is either LV\_Growth, which indicates left ventricular growth (or a thickening of the heart muscles which is related to heart diseases and strokes), or a normal result (i.e. no ventricular growth).



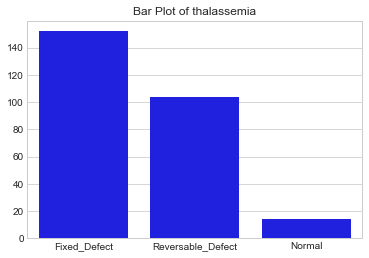
This bar chart shows whether or not the patient has Angina which is pain or discomfort caused by a lack of blood and oxygen for the heart. There are multiple symptoms of angina and it can be an indicator of heart attack and disease. Here we can see that the majority of patients included in the dataset did not have angina.



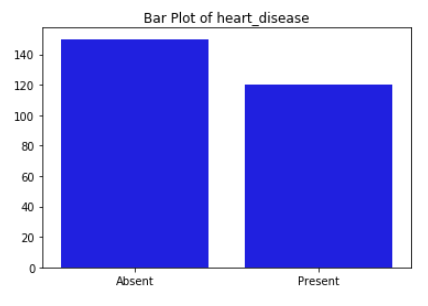
The ST\_Slope shown here is the heart rates reaction to stress testing (i.e. exercise). The most common reactions to this stress test is a slight downward slope of an ecg test or this graph remaining flat.



This plot shows the distribution of the number of heart vessels that were shown as “coloured” in a fluoroscopy. A fluoroscopy is a procedure that a patient will undergo similar to an X-ray to detect heart disease. Here we can see that just over half of the observations returned zero major vessels as coloured.

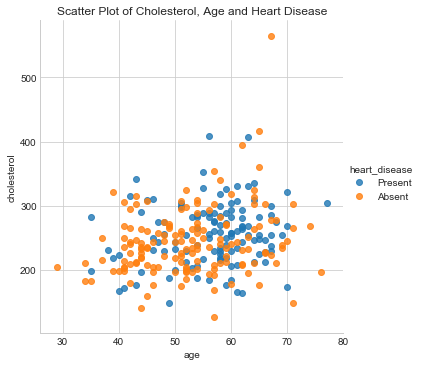


In the above bar chart we can see that only a small minority of patients had no thalassemia with most having the disorder treated (fixed defect) and a smaller amount having the disorder but not having treatment yet (reversable defect).

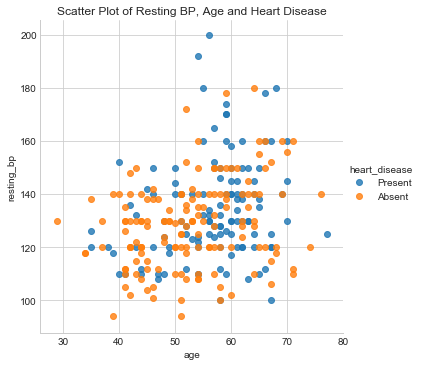


The bar plot of the target feature is shown below. From this it can be seen that the two levels are almost balanced, however there are slightly more people within the absent level compared to the present level.

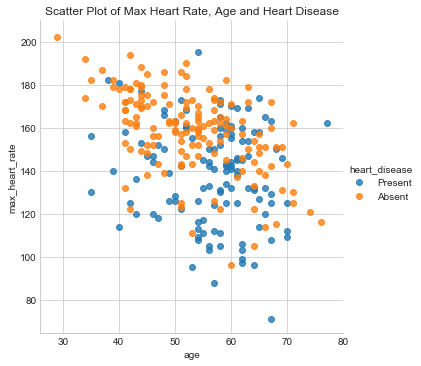
* Multivariate

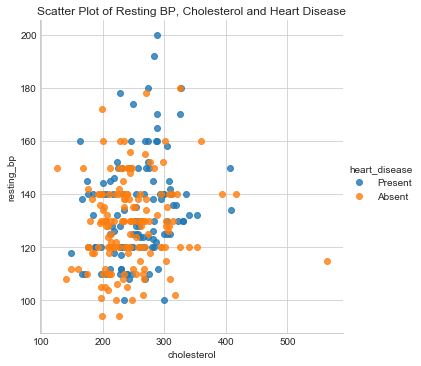


This plot of heart disease by cholesterol level and age shows that while the level of cholesterol doesn’t seem to have a major influence on the rate of heart disease there, as age increases the likelihood of heart disease also increases. Although this isn’t a completely clear correlation, after the age of 55 it becomes much more likely that heart disease will be present.

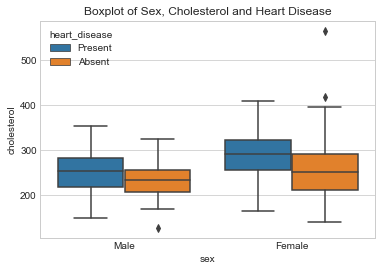


This increase in the likelihood of heart disease after the age of 55 is also seen in this plot. Unlike the previous plot however, there appears to be a link between an increase in the blood pressure indicates an increased chance of heart disease. The strength of this correlation should not be overstated however, as there is less recorded instances of resting blood pressures over 180 and the cause of this increase in heart disease could be linked to the increase in age (as discussed previously). Later in our report we will discuss which features we believe are the best for predicting the presence of heart disease.

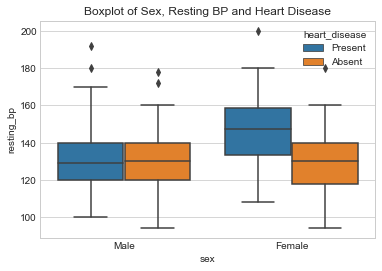
  
With this plot we can see that as the age of the patients increases, their max heart rate will slightly decrease. With this relationship you can see that as the age increases and the max heart rate decreases, the rate of heart disease will increase. Although there is a clear relationship in the plot, the extent to which feature is more influential in the cause of heart disease will be investigated in the modeling stage.



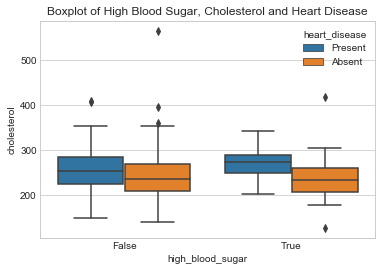
This plot which shows the relationship between resting blood pressure and cholesterol does not show as clear of a link between the features and the presence of heart disease as the previous plots which featured age. This means that it is likely that age is a major factor in the prediction of heart disease. For this plot we can see a slight increase in heart disease as cholesterol and blood pressure increase however, it must be seen as a minor relationship.



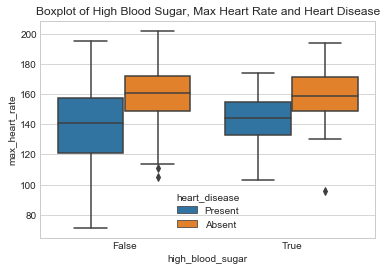
These faceted boxplots show that a higher cholesterol measurement is more likely to be associated with heart disease. Also we can see that the distribution of cholesterol levels is similar across the two sexes, however it should be noted that females slightly tended to have an overall higher cholesterol level.



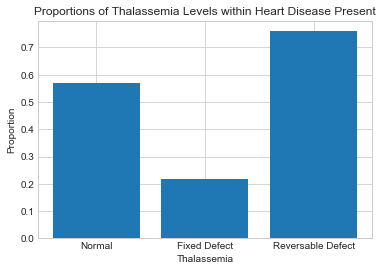
In this faceted boxplot we can see that the resting blood pressure for men is essentially the same regardless of the presence or absence of heart disease (with the extreme values being slightly different) however for women, there is a clear link between a high blood pressure level and the presence of heart disease. This finding is interesting especially when viewed with the previous scatted plot which showed a slight correlation between an increase in heart disease with high resting blood pressure.



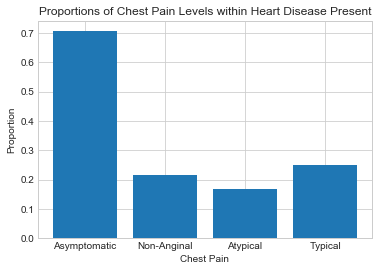
The relationship between the High Blood Sugar, Cholesterol and Heart Disease variables has been visualised above. We can see in the box plot that the cholesterol level of patients without heart disease is somewhat constant regardless of whether or not the person has high blood sugar (with the main difference being the level of variation). When heart disease is present however, we can see that there is overall a higher cholesterol level and that this difference is increased with high blood sugar.



The faceted box plots above show the relationship between High Blood Sugar, Max Heart Rate and the target variable, Heart Disease. The greater width of box plots when high blood sugar is not present (“False”) shows that there is more variation in the heart rates of patients when they do have high blood sugar. With regards to the target variable, there appears to be a significant link between a lower maximum heart rate and heart disease regardless of whether or not the patient has high blood sugar levels.



The relationship between Thalassemia and Heart Disease is visualised in this Bar Chart. For cases in which the patient did not have Thalassemia (a blood disorder) the rate of heart disease was around 55%. Interestingly, whether or not Thalassemia was fixed seems to be a good predictor of heart disease with only 20% of cases in which the patient had been treated for the disorder having heart disease, while over 70% who had the disorder but had not been treated also had heart disease.



This chart shows the proportion of heart disease occurrence with each type of chest pain, so around 25% of patients reporting “Typical” chest pain will have heart disease, while the other 75% of patients with typical chest pain will not have heart disease. Across all possible types of chest pain, “Asymptomatic” pain is most likely the best predictor with over 70% of cases associated with heart disease, this is more than twice the rate of any other type of pain.

*Modeling –*

**K-Nearest Neighbours (KNN) Modeling:**

Initially, a KNN model with K=1 using the Euclidean distance metric was trained. The associated confusion matrix is shown below.

**Matrix 1. KNN: Euclidean, K = 1**

|  |  |  |  |
| --- | --- | --- | --- |
|  | | **Predicted** | |
| Present | Absent |
| **Observed** | Present | 22 | 14 |
| Absent | 13 | 19 |

Different variations of the KNN model were then trained in order to fine tune the parameters of the learner.

The confusion matrix for the KNN with K=1 using the Manhattan distance metric is shown below.

**Matrix 2. KNN: Manhattan, K = 1**

|  |  |  |  |
| --- | --- | --- | --- |
|  | | **Predicted** | |
| Present | Absent |
| **Observed** | Present | 22 | 14 |
| Absent | 14 | 18 |

Evidently, the Euclidean model performed slightly better than the Manhattan one. As such, all further parameter tuning employed the Euclidean metric.

The K parameter was tuned next. The confusion matrix for the KNN model with K=3 and K=5 are shown below in Matrices 3 and 4, respectively.

**Matrix 3. KNN: Euclidean, K = 3**

|  |  |  |  |
| --- | --- | --- | --- |
|  | | **Predicted** | |
| Present | Absent |
| **Observed** | Present | 27 | 9 |
| Absent | 11 | 21 |

**Matrix 4. KNN: Euclidean, K = 4**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | **Predicted** | | | |
| Present | | Absent | |
| **Observed** | Present | | 27 | | 9 |
| Absent | | 11 | | 21 |

Comparison of these two matrices emphasizes that they both perform better than the K=1 model, with fewer False Positive and False Negative counts. Furthermore, both models can be seen to perform at exactly the same level.

A weighted KNN model was also trained with K = n train. The associated confusion matrix is shown below.

**Matrix 5. KNN: Euclidean, Weighted, K = n train**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | **Predicted** | | | |
| Present | | Absent | |
| **Observed** | Present | | 35 | | 1 |
| Absent | | 27 | | 5 |

Here it can be seen that this model performs fairly poorly in that it over-predicts the “Present” level, resulting in a large amount of False Positive predictions.

As a result of this tuning, the K=3 and K=5 models were found to be the best performing.

The classification report for the K=5 model is shown below (this report would be identical to the K=3 model’s report)

**Table 1. KNN: Euclidean, K = 5**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Precision** | **Recall** | **F-1** | **Support** |
| **Present** | 0.71 | 0.75 | 0.73 | 36 |
| **Absent** | 0.7 | 0.66 | 0.68 | 32 |
| **Ave./Total** | 0.71 | 0.71 | 0.71 | 68 |

This shows that the model was 71% accurate in its predictions.

Decision Tree Modeling:

The confusion matrix for a basic decision tree model is shown below.

**Matrix 6. Decision Tree**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | **Predicted** | | | |
| Present | | Absent | |
| **Observed** | Present | | 29 | | 7 |
| Absent | | 8 | | 24 |

This shows that the model is already performing fairly well.

Some of the learner’s parameters were then altered to see if the model accuracy could be improved. Firstly, the maximum depth of the decision tree was set to 3. The associated confusion matrix is shown below.

**Matrix 7. Decision Tree: Max Depth = 3**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | **Predicted** | | | |
| Present | | Absent | |
| **Observed** | Present | | 35 | | 1 |
| Absent | | 12 | | 20 |

This model has a very good True Positive Rate, however it is not very good at predicting absence of heart disease. The max depth parameter was increased to 4, and the model’s confusion matrix is shown below.

**Matrix 8. Decision Tree: Max Depth = 4**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | **Predicted** | | | |
| Present | | Absent | |
| **Observed** | Present | | 33 | | 4 |
| Absent | | 10 | | 22 |

This model’s accuracy is very similar to the previous one, however its incorrect predictions are more balanced. The max depth parameter was increased once more to 5 and the confusion matrix is below.

**Matrix 9. Decision Tree: Max Depth = 5**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | **Predicted** | | | |
| Present | | Absent | |
| **Observed** | Present | | 29 | | 7 |
| Absent | | 9 | | 23 |

Evidently, the True Positive Rate has decreased.

Based on these results, the two best performing decision tree models were those with max depths of 3 and 4. As such, the classification reports for both models were constructed. They are shown below in respective order.

**Table 2. Decision Tree: Max Depth = 3**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Precision** | **Recall** | **F-1** | **Support** |
| **Present** | 0.74 | 0.97 | 0.84 | 36 |
| **Absent** | 0.95 | 0.62 | 0.75 | 32 |
| **Ave./Total** | 0.84 | 0.81 | 0.80 | 68 |

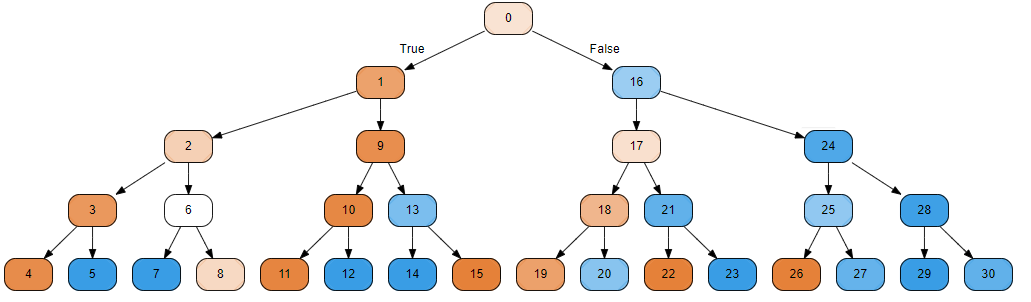
**Table 3. Decision Tree: Max Depth = 4**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Precision** | **Recall** | **F-1** | **Support** |
| **Present** | 0.77 | 0.92 | 0.84 | 36 |
| **Absent** | 0.88 | 0.69 | 0.68 | 32 |
| **Ave./Total** | 0.82 | 0.81 | 0.81 | 68 |

Based on these reports, it is clear that although the model with max depth of 3 has a better precision score, the F-1 score is slightly better for the max depth 4 model. As such, this model can be taken as the best.

The visualization of the decision tree produced by this model is shown below as well.

**Figure 1. Decision Tree for Max Depth = 4 Model**



**Naïve Bayes Modeling:**

The Naïve Bayes algorithm was applied to the training set and the confusion matrix is shown below:

**Matrix 10: Naïve Bayes**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | **Predicted** | | | |
| Present | | Absent | |
| **Observed** | Present | | 30 | | 6 |
| Absent | | 7 | | 25 |

Clearly, this model is already performing well. The associated classification report for this model is shown below.

**Table 4. Naïve Bayes**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Precision** | **Recall** | **F-1** | **Support** |
| **Present** | 0.81 | 0.83 | 0.82 | 36 |
| **Absent** | 0.81 | 0.78 | 0.79 | 32 |
| **Ave./Total** | 0.81 | 0.81 | 0.81 | 68 |

**Discussion:**

Through the use of a variety of supervised machine learning models, we can attempt to create an effective classifier that could accurately predict the presence of heart disease given a set of descriptive features. With this aim, K-Nearest Neighbour, Decision Tree and Naive Bayesian models were all investigated and with the above results we can say that a decision tree with a max depth of four nodes was the most accurate model with the Naive Bayes the second most accurate. This judgment has been made through the use of a F1 Score which is a combination of the precision of a model (the fraction of ‘positive’ predictions that correctly identified heart disease) and the recall (the proportion of heart disease cases that were identified correctly).

The selection of the parameters for these algorithms was a process of …

As the central dataset of this assignment was already a subset of a larger group of descriptive features, we did not attempt any feature engineering, however we believe that current feature selection techniques could be used in further study to test whether or not this subset was the most efficient combination of features to be used in a classification model. This feature engineering could be done through … or ... Likewise, other classification methods such as Random Forests or Neural Networks could be used to potentially improve the accuracy or efficiency of this process and could be a basis for further study.

**Conclusion:**

Brief restatement of aims

Summary of findings