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 presences 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.

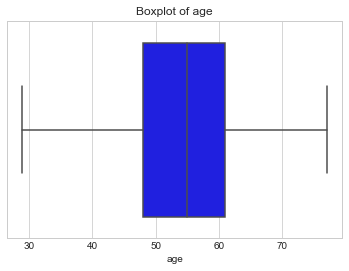
The new descriptive dataset was then explored through visualization. For continuous variables, boxplots were used to highlight differences between the levels of the target feature with respect to each feature. Bar charts were used for the categorical variables to show the differences in observation across 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.

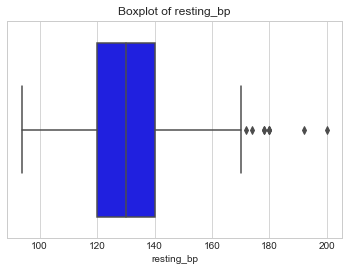
**Results:**

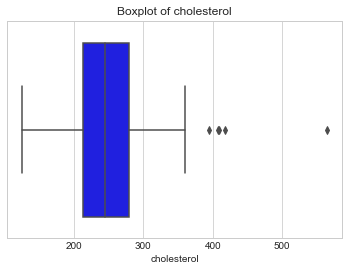
*Data Visualization –*

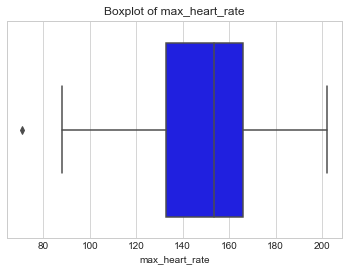
* Univariate

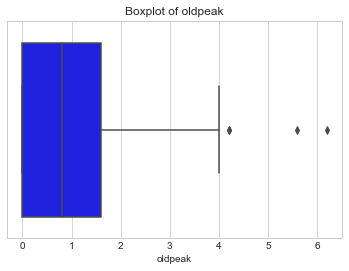


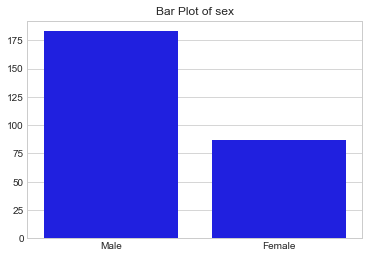
As we can see here the dataset that we are studying is made up of mostly older people with an average age of around 55. The width of the box plot shows that the majority of observed cases will feature people aged in their late 40s to their early 60s with some extreme cases of people in their 30s to people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



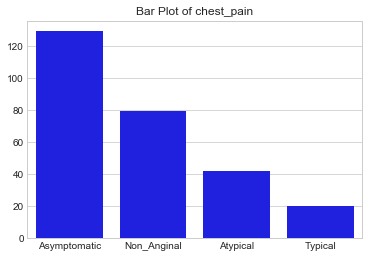


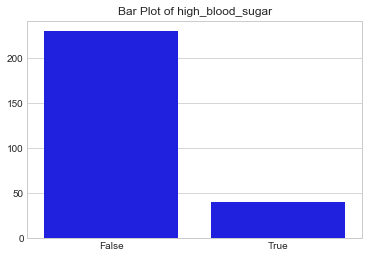


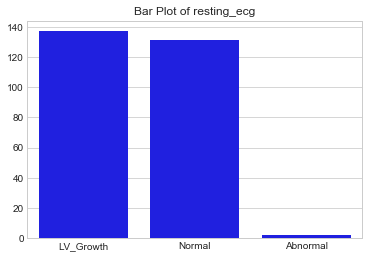


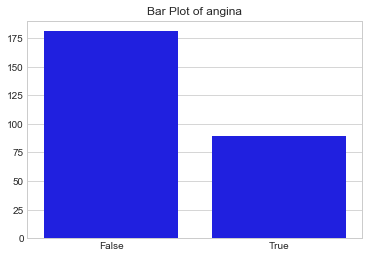


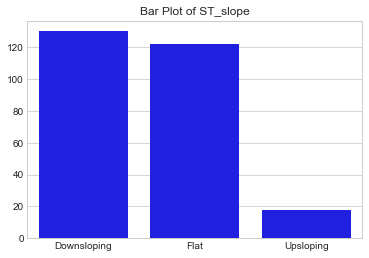
This bar plot shows that the vast majority of people in the study were male. Almost 2/3rds of the cases featured males with a 1/3 featuring females.

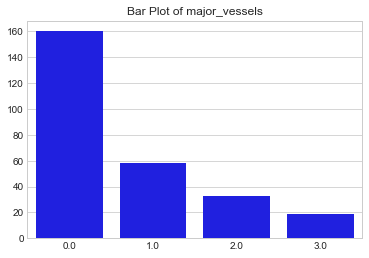


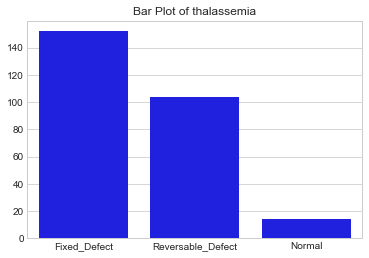




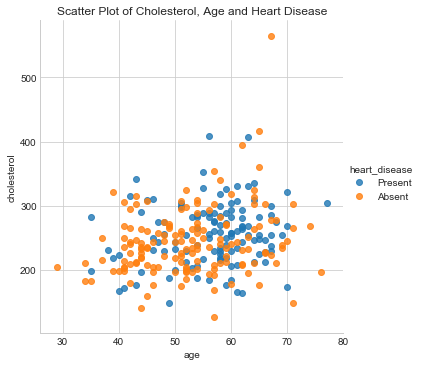




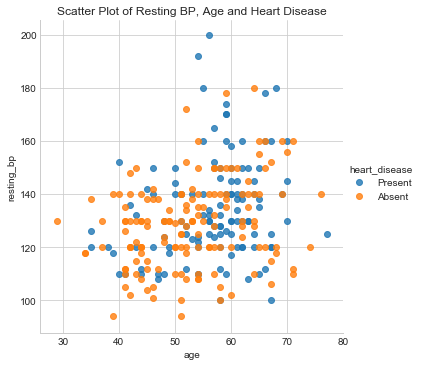




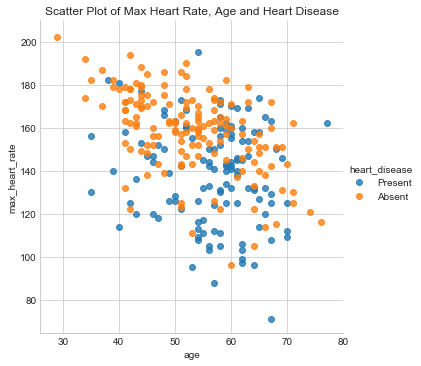
* Multivariate

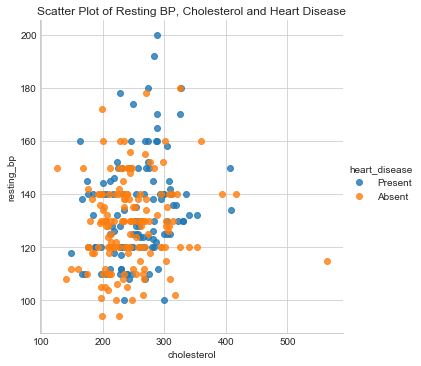


This plot of heart disease by cholesterol level and age shows that while the level of cholesterol doesn’t seem to have a major influence on the rate of heart disease there, as age increases the likelihood of heart disease also increases. Although this isnt a completely clear correlation, after the age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

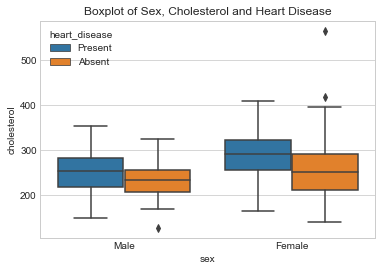


This increase in the likelihood of heart disease after the age of 55 is also seen in this plot. Unlike the previous plot however, there appears to be a link between an increase in the blood pressure indicates an increased chance of heart disease. The strength of this correlation should not be overstated 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.

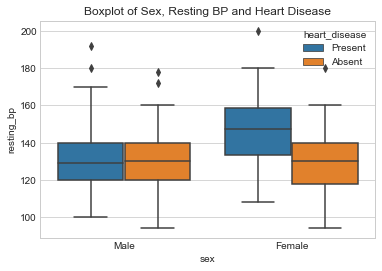
  
With this plot we can see that as the age of the patients increases, their max heart rate will slightly decrease. With this relationship you can see that as the age increases and the max heart rate decreases, the rate of heart disease will increase. Although there is a clear relationship in the plot,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



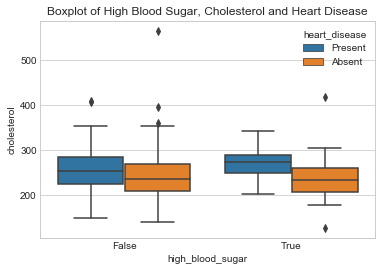
This plot which shows the relationship between resting blood pressure and cholesterol does not show as clear of a link between the features and the presence of heart disease as the previous plots which featured age. This means that it is likely that age is a major factor in the prediction of heart disease. 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.

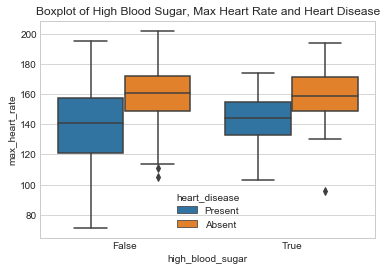


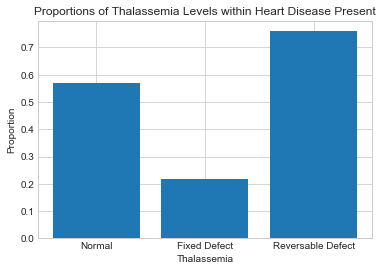
These faceted boxplots show that a higher cholesterol measurement is more likely to be associated with heart disease. Also we can see that the distribution of cholesterol levels is similar across the two sexes, however it should be noted that females tended to have an overall higher cholesterol level.

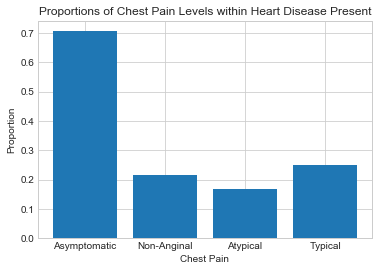


In this faceted boxplot we can see that the resting blood pressure for men is essentially the same regardless of the presence or absence of heart disease (with the extreme values being slightly different) however for women, there is a clear link between a high blood pressure level and the presence of 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.









*Modeling –*

**Discussion:**

**Conclusion:**