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>

*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.

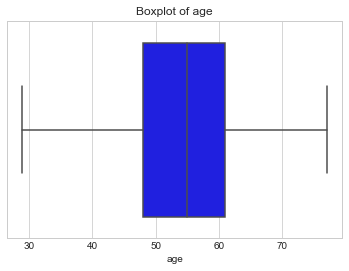
The new descriptive dataset was then explored through visualization. For continuous variables, boxplots were used to highlight differences between the levels of the target feature with respect to each feature. Bar charts were used for the categorical variables to show the differences in observation across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

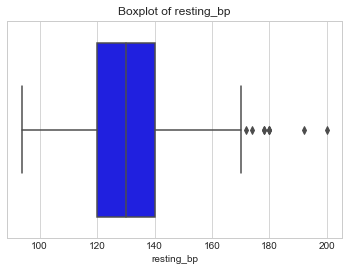
Following this, the classification analysis was conducted. Here, the original data was used. The data was split into training and testing segments using a 75:25 ratio. As such, the training data comprised 216 observations, while the test data had 54 observations. A k-nearest neighbours algorithm was 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.

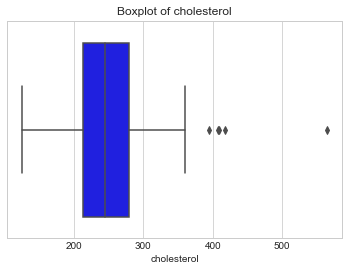
**Results:**

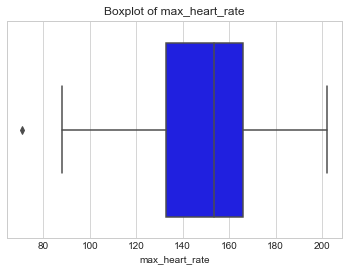
*Data Visualization –*

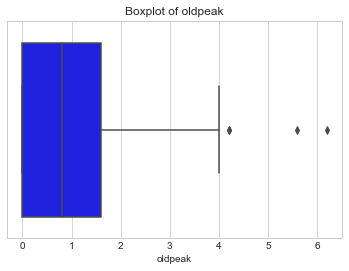
* Univariate

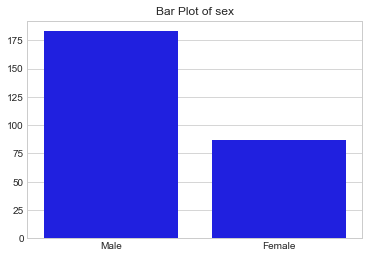


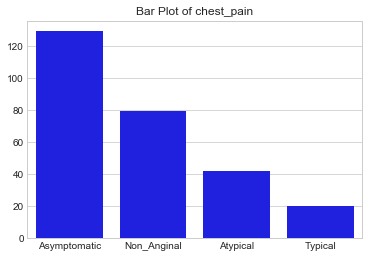


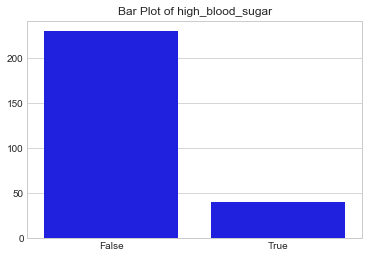


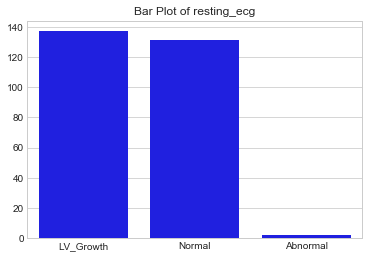


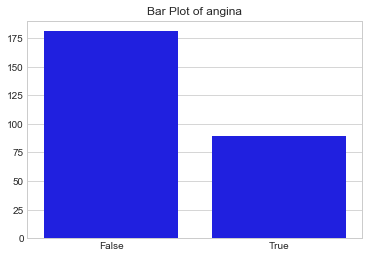


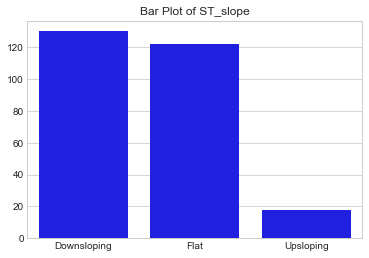


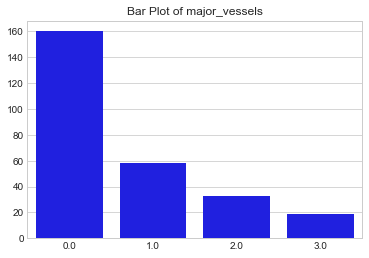


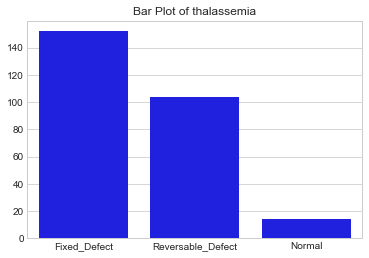




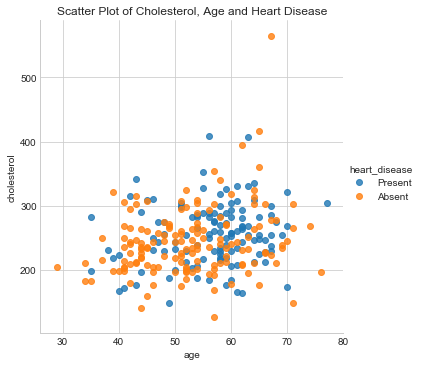


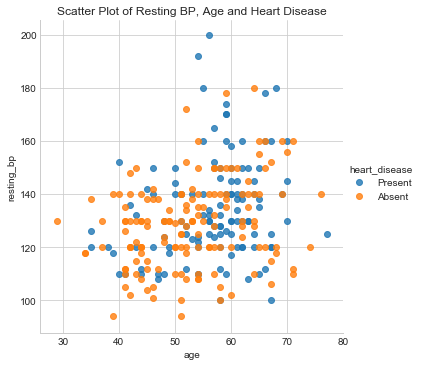


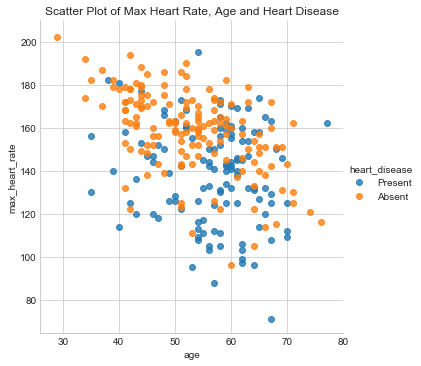


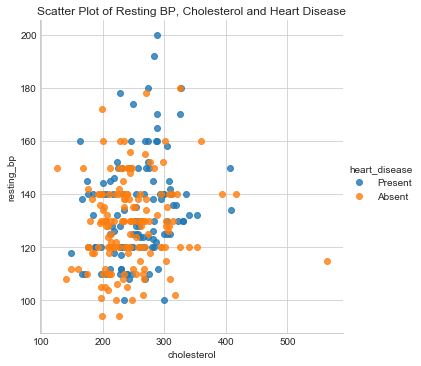


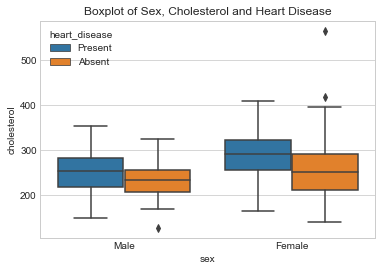
* Multivariate

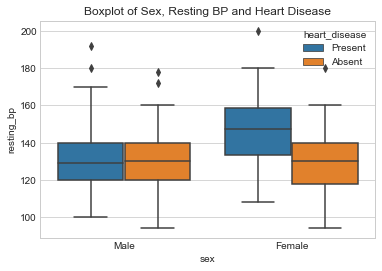


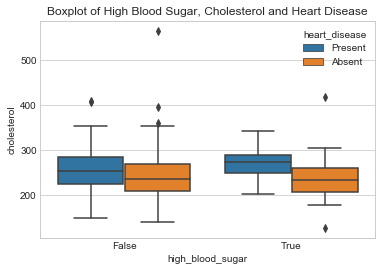


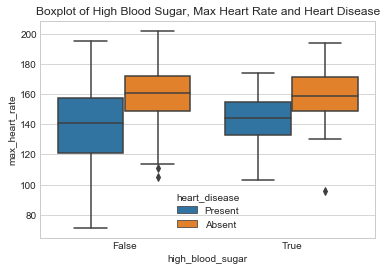


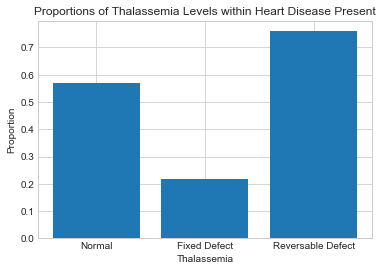


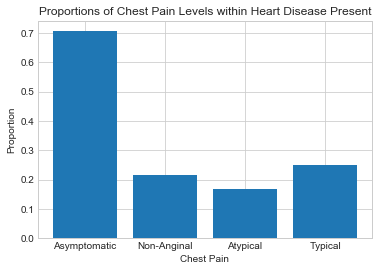












*Modeling –*

**Discussion:**

**Conclusion:**