MvLA Report 2020

Michael Haskins

28/11/2020

Introduction

The heart data set (file heart_data.csv) consists of nine variables, each measuring different physiological characteristics of patients at risk of heart disease. Five variables are continuous, and three are binary. One variable, Slope, is ordinal, but will be treated as continuous. The tenth variable in the data set is Class, and this denotes whether heart disease is absent (1) or present (2).

This report will consist of two sections;

1. Cluster Analysis:
    1. Identifying subsets present in the data
    2. Identifying subsets at higher risk of heart disease
2. Classification:
    1. Predicting the patients with heart disease

Cluster Analysis

Identifying subsets present in the data

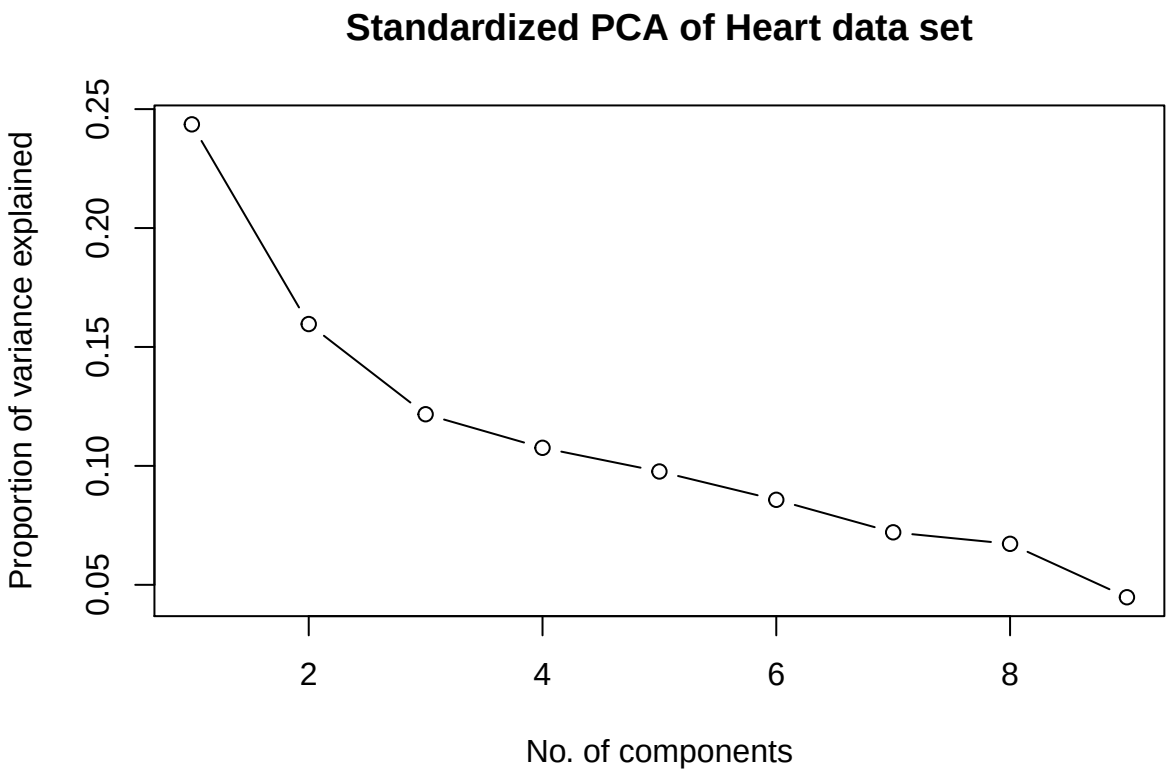
Section Aim:

- To identify subsets of patients in the data set who have similar physiological characteristics. and determine if these subsets are at higher risk of heart disease.

Principal Component Analysis(PCA)

Principal Component Analysis(PCA) is a dimensionality-reduction method to reduce the number of variables of a data set, while preserving as much information as possible (trade a little accuracy for simplicity). Perfoming a PCA makes it easier to identify clusters/groups present in the data. In PCA the most important component is 1 followed by 2 and so on. This is because the variance the components account for is listed in descreasing order from PC1 onwards.

I conducted a standardized PCA on this data set. The data has been standardized so that the variables with the highest variance do not dominate the PCA. If the data was not standardized the PCA would be completely dominated by the SerumCholestral varaible as this has the highest varaince (of 2671.467). Below is a graph showing the proportion of variance that can be explained by each principal component.



Note that the proportion of variance accounted for by PC1 is very small (0.24). When the fifth PC is included the cumulative proportion of the variance of the dataset accounted for is only 0.73. This indicates that this dataset is not very suited to PCA.

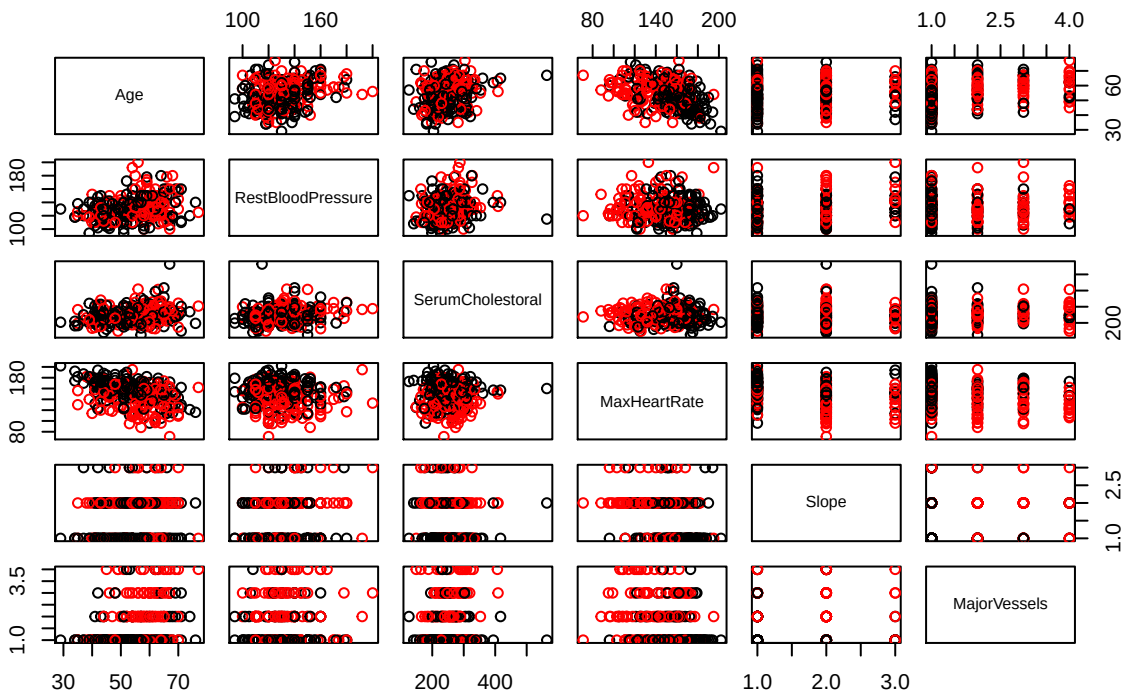
The most influential variables in the standardized PCA for PC1 were MaxHeartRate and age. This is because their coefficients or 'loadings' (See below) in the PC1 vector have the largest magnitude. As they have different signs this indiactes that PC1 contrasts people who have a high MaxHeartRate and are of a young age. No further valuable insights can be gained from PCA due to its poor performance when applied to this dataset.

```
                    PC1   PC2   PC3   PC4   PC5   PC6   PC7   PC8   PC9
Age               -0.46  0.28 -0.01  0.23 -0.26  0.23 -0.25  0.38 -0.57
Sex               -0.05 -0.52  0.46  0.12  0.38  0.32  0.23  0.46 -0.01
RestBloodPressure -0.26  0.36  0.21 -0.53  0.16  0.59 -0.06 -0.20  0.24
SerumCholestoral  -0.19  0.49 -0.26  0.07  0.58 -0.29  0.33  0.34  0.13
FastingBloodSugar -0.13  0.20  0.74 -0.15 -0.11 -0.57 -0.15  0.08  0.07
MaxHeartRate       0.49  0.26  0.23 -0.08  0.23  0.06  0.26 -0.24 -0.67
ExerciseInduced   -0.36 -0.34 -0.12 -0.16  0.52 -0.22 -0.43 -0.34 -0.30
Slope             -0.37 -0.24 -0.12 -0.48 -0.30 -0.17  0.62 -0.01 -0.23
 [ reached getOption("max.print") -- omitted 1 row ]
```

Visually comparing the data

Please see the below graphs comparing the continuous variables of the dataset.

Heart Data Continuous Variables (Red denotes Heart Disease present)



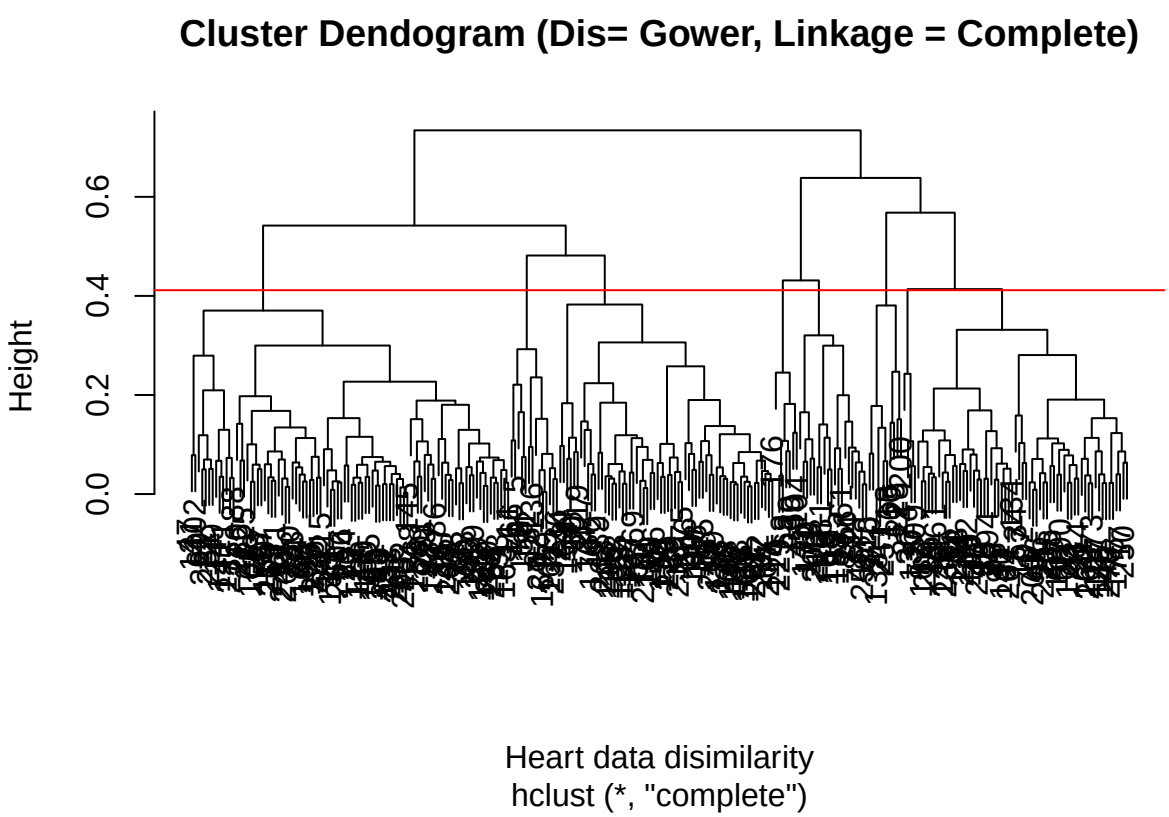
Not much can be learned from the graphs that are comparing the variables Age, RestBloodPressure and SerumCholestral to one another, but the graphs including the final 3 variables yield some interesting insights. Patients with a lower MaxHeartRate reading typically have heart disease and patients with high slope and MajorVessels readings appear to generally have heart disease. This will be discussed further later on in this report.

Heirarchical Clustering

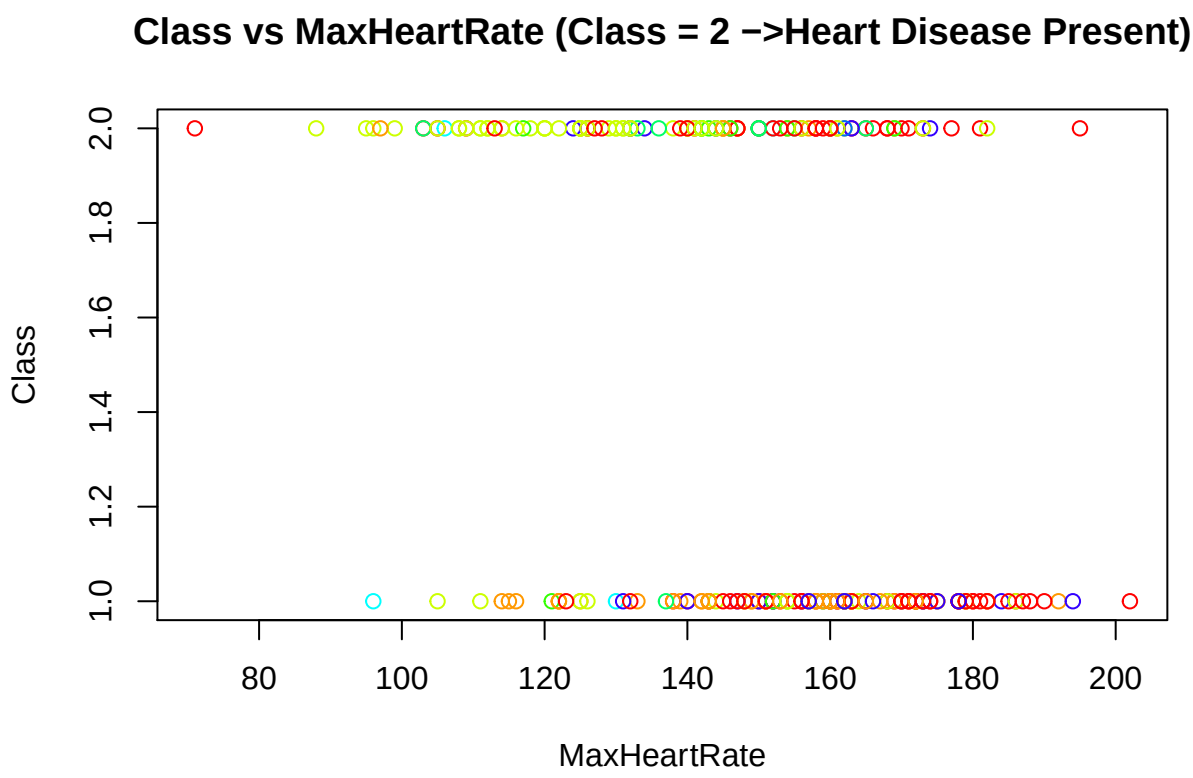
Hierarchical Clustering (HC) analysis groups similar observations into groups called clusters. This reveals group structure in the data.

Since this data set contains continuous and binary data the dissimilarity matrix I will be using is gower's general dissimilarity measure. This is the recommended dissimilarity measure to use when dealing with mixed data. The linkage method I used was complete linkage, I believe this linkage method is the best for this dataset as it will join the final clusters at a much larger measure of dissimilarity ensuring that there is good internal similarity in the clusters.

A dendrogram, like the one seen below, is used to visually display the hierarchical relationships between observations. The observations that have been joined at the very bottom are most similar to one another, and then the further up in the graph the more disimilar the clusters are that are being joined.



The above dendogram describes the group structure present in the data. The height of the dendogram describes the dissimilarity (with respect to linkage method) between groups as they were clustered. The recommeneded cut off height, which is used for determining how many clusters/groups are present in the data set is h + 3sh, where h is the mean height at which groups are joined, and sh is the standard deviation of such heights. The recommended cut off height for this data set is 0.4115239, this is the organge line in the above plot. From this it can be determined that there are 8 clusters/groups in the dataset.



Above is a visualization of max heart rate vs class. Each of the different coulours signify clusters which were obtained from the heirarchical clustering that was performed earlier in this report. Visually not much can be deduced from this, except that the cluster being represented by the green dot appears to generally have heart disease.

```
      true_label
groups  1  2
     1 52 40
     2 54  8
     3 14 47
     4  7  7
     5  2  8
     6  5  2
     7  0  4
     8 16  4
```

In the above table we can see how many members of each of the 8 clusters have heart disease in the true_label second column and don't have heart disease by looking in the true_label first column. From this we can see that the members of clusters 2 and 8 mainly do not have heart disease and the members of groups 3 and 7 have heart disease. It is not possible to gain any other useful insights from this table.

The below calculated Rand Index will more precisely compute how accurately these clusters reflect the true group structure of the data.

```
       RI
0.5408784
```

```
       ARI
0.08607804
```

The Rand index can have a value between 0 and 1, with 0 indicating that the two data clusterings do not agree on any pair of points and 1 indicating that the data clusterings are exactly the same.

The heirarchical data clustering result (obtained from using Glover's dissimilairity measure, complete linkage and having 8 clusters) when compared to the true clusters classification of the data has a Rand Index score of 0.5408784. This Rand Index score was not calculated with respect to agreement by chance, so is therefore an unadjusted Rand Index.

The adjusted Rand Index, which is calculated with respect to agreement by chance, is 0.08607804. This is a very low score and suggests there is little agreement between the clusters and the real data in terms of classifying the patients with heart disease. Therefore it is not clear from heirarchical data clustering what subsets of patients are at higher risk of heart disease.

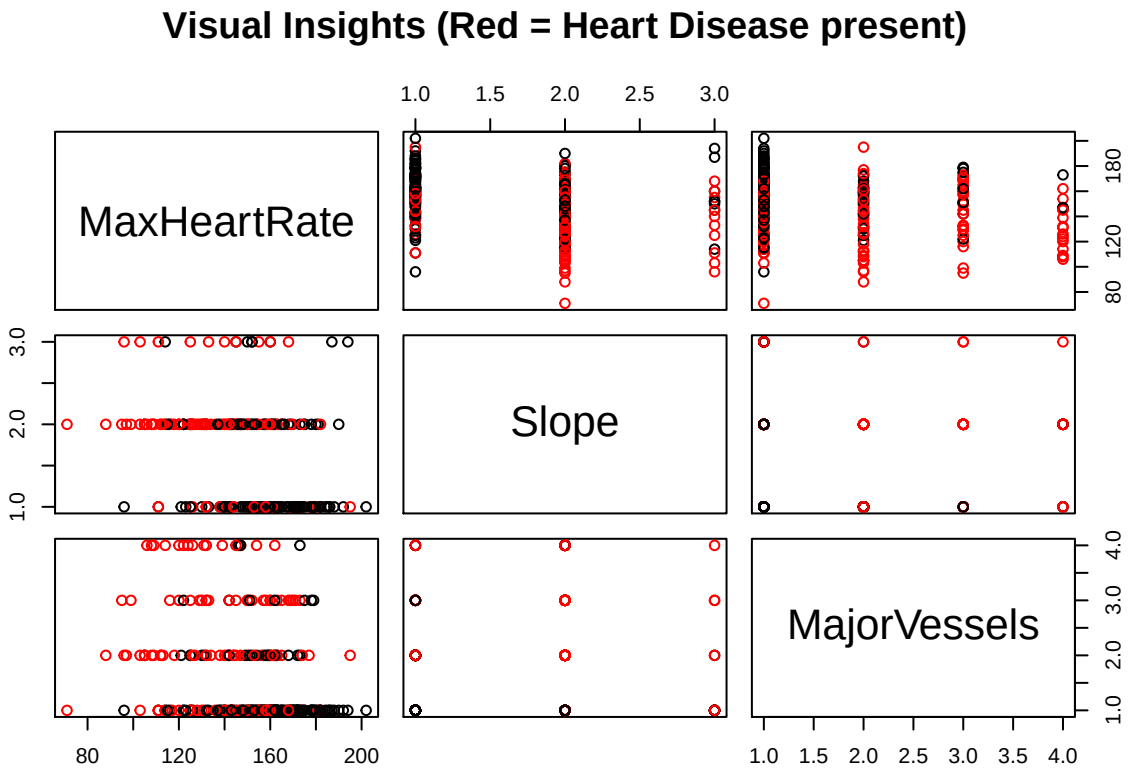
Identifying subsets at higher risk of heart disease

Section Aim:

- To determine if any of the identified subsets are at higher risk of heart disease.

Due to the poor performance of the PCA it is not possible to identify any subsets and thus impossible to accurately determine if any of the identified subsets in PCA are at higher risk of heart disease.

The low Rand index scores which were calculated in Heirarchel Clustering analysis mean that there is little aggreement between the actual group structure of the data (i.e. group1 have heart disease & group2 don't) and the calculated eight clusters. It was shown above that the members of clusters 2 and 8 mainly do not have heart disease and the members of groups 3 and 7 have heart disease. The remaining clusters contain a mixture of patients who have and don't have heart disease, making it impossible to identify any common characteristics of patients who have heart disease from this analysis.



In this graph it is possible to see that patients with a MajorVessels reading of greater than or equal to two are likely to have heart disease, it is even more likely patients have heart disease if they have also recorded a reading under 140 for MaxHeartRate and a reading greater than 2 for Slope.

Classification

The purpose of classification in this data set is to use the known data to determine if new samples have heart disease.

I will not be standardizing the data for any of the following classification methods in order to ensure that the output is easily interpreted by a wide ranging audience. For example, the group means from an unstandardized LDA are far more informative that those of a standardized LDA and easier to interpret for someone who doesn't have extensive knowledge of standardization (If this doesn't make sense now it will later on in this report).

Predicting the patients with heart disease

Section Aim:

- To provide accurate predictions of patients with heart disease.

K-Nearest Neighbours (KNN)

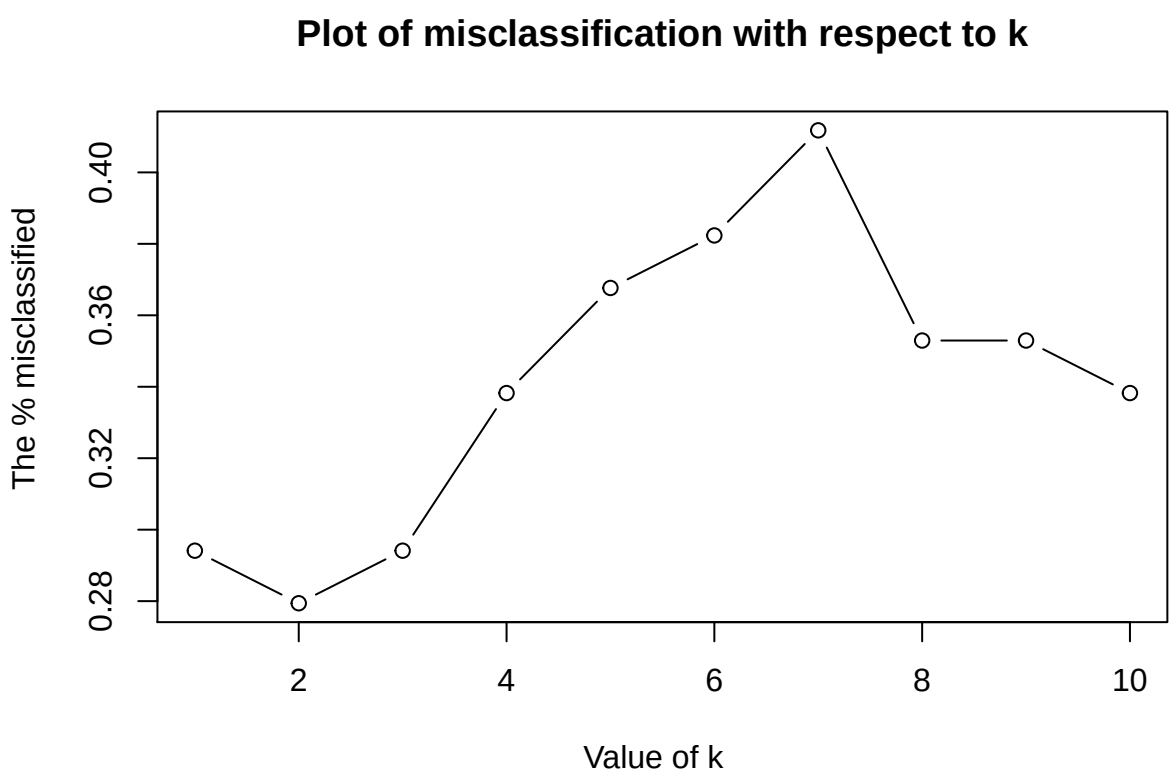
KNN classification is a non-parametric method of assigning group membership. It makes no assumption of the spread of the data within each class. The consequence of this is that it is not possible to measure uncertainty concerning any assignments.

This method works by comparing an observation to the k nearest datapoints and assigning that observation to the cluster who contains the most of the k datapoints it is closest to.

Choosing a value for k is very important to this method. To decide on a value for K it is important to split the known data into a training, test and validation set. This will be done in the ratio of (50%, 25%, 25%). This is a common split, because if the training set is too large the results can look better than they actually are due to a lack of samples in the validation and testing sets. The training set is used to classify the 'unlabelled' points of the test set. The test set is used find the value of k that is best for classification and the validation set is used to estimate the classification error of the best k identified by the test set.

I performed a KNN classification using only datapoints containing the variables MaxHeartRate, Slope and MajorVessels. I did this because earlier in this report I found these by visual inspection to be the most influential on whether a patient has heart disease. The best value of k, in terms of minimizing misclassification, was 2 which resulted in a miclassification rate of 27.94%. Although this is quite high, it means it is possible to make classification predictions for new datapoints with a misclassification rate of 27.94%.

```
      Value of k       The % misclassified
 [1,]          1                 0.2941176
 [2,]          2                 0.2794118
 [3,]          3                 0.2941176
 [4,]          4                 0.3382353
 [5,]          5                 0.3676471
 [6,]          6                 0.3823529
 [7,]          7                 0.4117647
 [8,]          8                 0.3529412
 [9,]          9                 0.3529412
[10,]         10                 0.3382353
```



If all the variables were to be included in the KNN classification it would result in a best value of k (in terms of minimizing misclassification) of 5 with a misclassification rate of 32.35%. Therefore it is possible to conclude that the three variables MaxHeartRate, Slope and MajorVessels are better at classifying new points than the whole dataset, as the lowest misclassification from that KNN classifcation was 27.94%.

Since the value of k is chosen specifically for the test set it is important to check how well it performs when applied to another set (the validation set). The validation set can offer a more accurate estimate of the correct classifciation rate. The misclassifcation rate for the validation set was 26.87% (See output above). It is therefore possible to conclude that when examining only the three variables MaxHeartRate, Slope and MajorVessels we can correctly predict the classifcation (if they have heart disease or not) of new patients circa 74% of the time using KNN.

Linear Discriminant Analysis (LDA)

LDA and Quadratic Discriminant Analysis (QDA) are used if it is known or assumed that there is group structure in the data. In this data set there are two groups, those who have heart disease and those who don't.

This type of analysis reveals structure in the data which then allows for the classification of future observations. This is done by using the characteristics of labelled data in order to classify the group membership of unlabelled data.

Unlike KNN, both LDA & QDA assume the use of a distribution over the data. This then enables us to quantify the uncertainty over the structure of the data and we can consider the probability of group assignment (I.e. Is an observation highly likely to be part of the group it was assigned to or not).

As there is a large number of variables in this dataset (9) this means the data set has a high dimension. In QDA the higher the dimension the more parameters that have to be estimated (While in LDA the number

of parameters is fixed). This can lead to high variance and so it is for this reason that I belive LDA is more suited to this data set.

I split 75% of the data into a training set which will be subject to LDA. We can use the result to find the prior probabilities of an observation belonging to a group and the group means for each variable. The remaining 25% of the data has been split into a test set which will be used to determine how accurate classifications made using LDA were.

Below are the prior probailities that an observation belongs to one of the groups, this is followed by the group means for each variable. From the group means, for example, we can determine that on average a person with heart disease has a MaxHeartRate recording of 138. This can be quite useful information.

Note: Group 2 has heart disease and Group 1 does not.

```
        1         2
0.5517241 0.4482759
```

```
       Age      Sex RestBloodPressure SerumCholestoral FastingBloodSugar
1 53.28571 1.517857          127.6429         243.3839          1.151786
2 56.42857 1.835165          134.4286         254.3407          1.153846
  MaxHeartRate ExerciseInduced    Slope MajorVessels
1     156.5357        1.178571 1.410714     1.312500
2     138.0330        1.549451 1.802198     2.197802
```

The miclassification percentage for LDA was 20.9% (See below), this means that we can determine the correct classification (Whether they have heart disease or not) of a new patient cicra 80% of the time using LDA. This is an improvement on KNN's rate of circa 74%.

```
[1] 0.2089552
```

Conclusion

The principal component analysis performed in this report was not very informative when applied to this dataset due to a large number of Principal components being required to account for an acceptable level of variance. In the Heirarchical clustering analysis performed it was shown that the members of clusters 2 and 8 mainly do not have heart disease and the members of groups 3 and 7 have heart disease. The remaining clusters contain a mixture of patients who have and don't have heart disease, making it impossible to identify any common characteristics of patients who have heart disease from this analysis. In my opinion it would not be very useful to study the characteristics of cluster 2, for example, as it only accounts for a small fraction of the observations present in this dataset.

It was shown that patients with a MajorVessels reading of greater than or equal to two are likely to have heart disease, it is even more likely patients have heart disease if they have also recorded a reading under 140 for MaxHeartRate and a reading greater than 2 for Slope. It was for this reason that these three variables were used in the KNN method in this report, which yielded a lower misclassification percentage than that of the KNN method which used all of the variables in the data set. By using this KNN method it is possible to correctly classify a patient circa 74% of the time.

LDA was able to better this with a correct classification percentage of circa 80% (all of the variables were used in the performed LDA).

Note: I did not standardize the data for KNN or LDA in order to ensure that the output is easily interpreted by a wide ranging audience. For example, the group means from an unstandardized LDA are far more informative that those of a standardized LDA and easier to interpret for someone who doesn't have extensive knowledge of standardization.