Machine Learning in Health Care

Andy Gray
445348

Part 1: Clustering

1. Run K-means clustering on the above heart disease dataset and answer the following questions

1. **Why should the attribute "*class*" in *heart-c.csv* ("*num*") not be included for clustering?**

The class num is the ended cluster predictions, so in essence these outputs are the labels. As K-Means is an unsupervised algorithm these are not needed and are a way for us to check that the algorithm has plotted/predicted well.

2. **Run K-means algorithm by choosing different numbers of clusters, *numCluster* = 2, 3, 4,5, then observe the differences of clusters generated:**
    1. **How are the *Within Cluster Sum of Squared Errors*[1] changed for different numbers of clusters?**

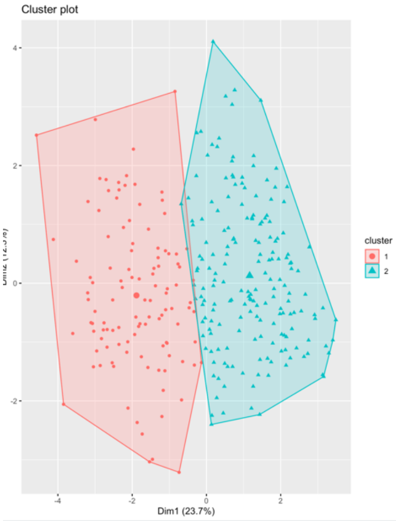
As the numbers of clusters is increased the within cluster sum of squares increases. K = 2 has the values 1306.399, 1865.346 and a between_SS/total_SS of 17.3%. K = 3 has the values 1500.4056, 505.1190, 938.7798 and a between_SS/total_SS of 23.2%. K = 4 has the values 466.9881, 838.7374, 812.6521, 597.6057 and a between_SS/total_SS of 29.2%. K = 5 has the values 458.2797, 811.3692, 451.2931, 339.3949, 529.1168 and a between_SS/total_SS of 32.5%.

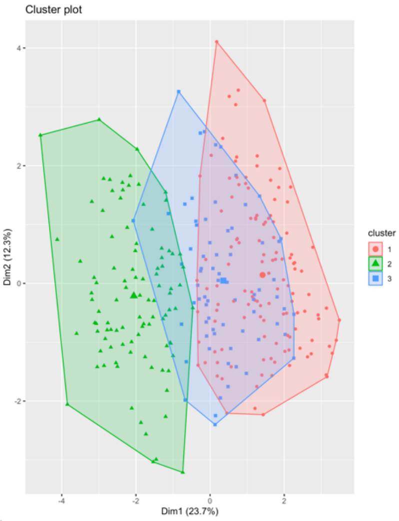
2. **What can you conclude?**

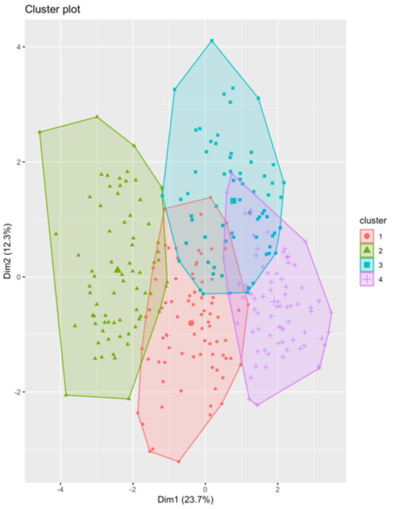
That as the cluster k numbers gets bigger the distance between the datapoints or variance per cluster is getting bigger. Showing that the datapoints are quite far apart from the centroids and not a very good k value to be using.

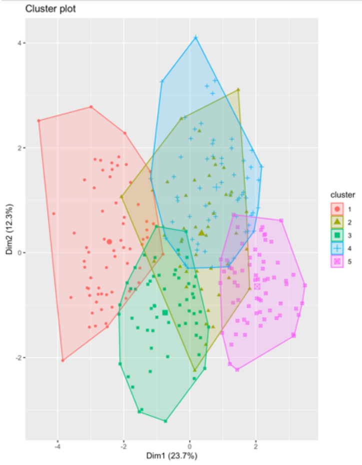
3. **How can you explain this conclusion from clustering analysis point of view?**

When we look at the visualisations of the data points, we can see there is a lot of overlapping of the data, which is expected with regards to the high dimensionality nature of the dataset. However, the data doesn't seem to cluster well for example, when we have 5 clusters compare to just the 2 clusters and with the high within cluster value getting higher as the ks get bigger indicate that the k is going in the wrong way and the fact we can only go to 2 ks minimum, it shows that for this dataset 2 ks is the ideal number. Which we know as the dataset has two labels within the "num" column, showing that there are two main clusters within the data.



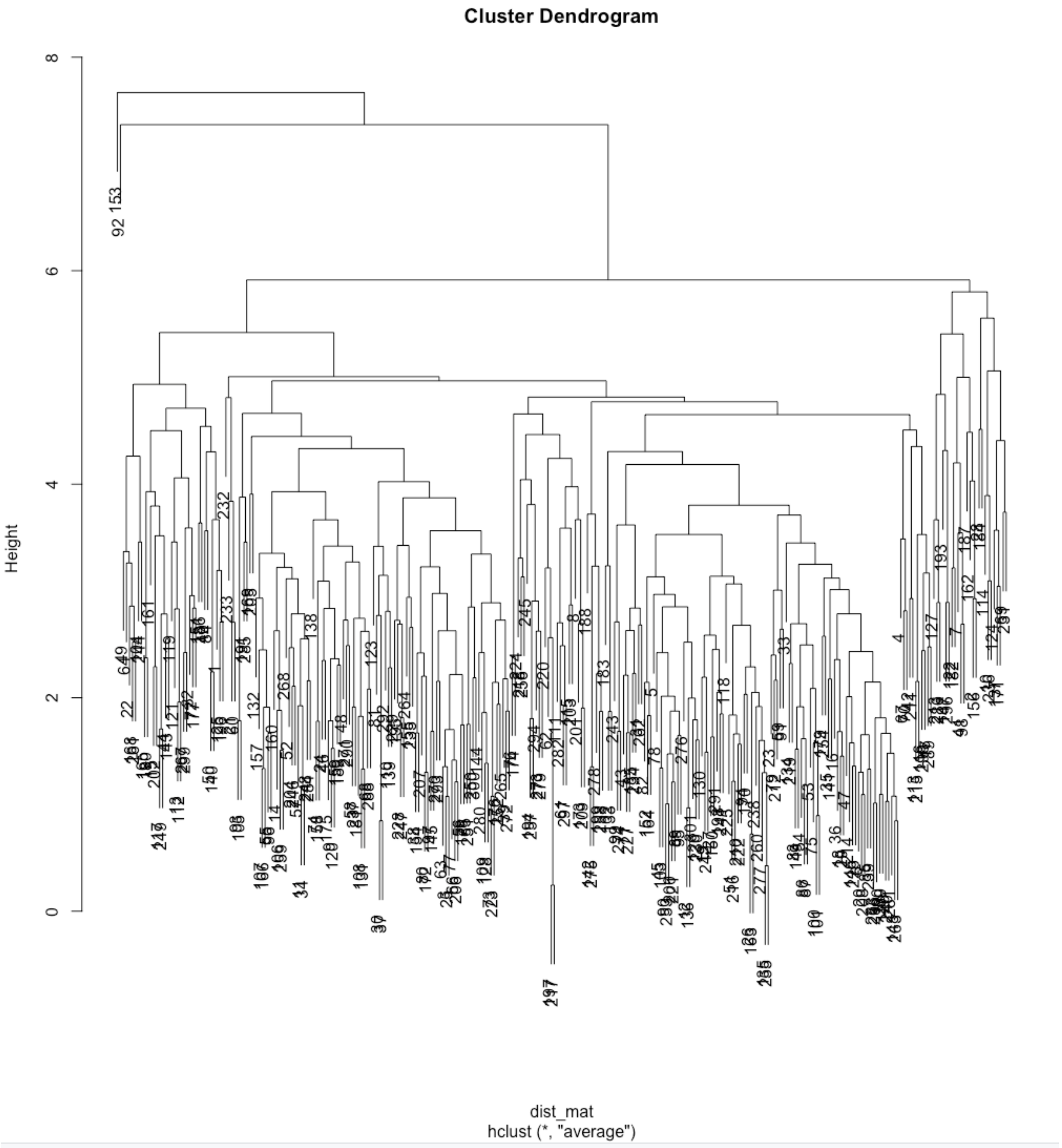






2. Run the hierarchical clustering on above heart disease dataset, and answer the following questions

1. **Show the clustering results in tree structure**



2. **Describe the link method you used**

The method used to configure the distribution is the Euclidean distance while the method used for linking the hierarchical cluster is the average method. This takes an average of the distributed Euclidean distance to then create a hierarchy.

3. **What are the strengths and limitations of this link method in hierarchical clustering?**

It shows the flow and the link of the clustering well. However, due to it being an average of the data values from the centroids, a lot of information can potentially be lost with it.

Part 2: Classification

Naïve Bayes:

1) **Did you undertake any prepossessing? If so, why?**

The data frame first had the variables X and Patient_ID removed due to these values not having any significant values. X Is just the column number originally and Patient_ID is the auto id number given to the patient when they are first entered into a database system.

The data frame then has any NA values removed and the class variable is converted to a factor value. This was done in order to make sure that all the variables had values and that any observations that had missing values would be removed completely from the dataset. The class changed to a factor was done to allow the algorithms to know that the class is are a category data type that is needed to classify the data predictions.

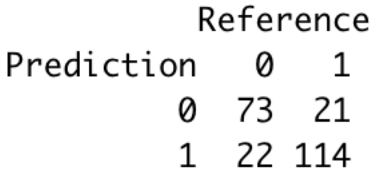
The data frame is then spilt into a train and test set while a 75-25% split to the data. This was done to allow us to see how well the models would predict on an unseen dataset. The 75% of the data was used to train the model while the 25% withheld was used to test the model's predictions.

2) **Run the classifier with default parameters.**
   a. **How accurately can the classifier predict those that develop heart disease? What is in the output that signifies this?**

There are two ways we can determine how well this model can predict. We could use its accuracy percentage while it was training, which was 77.83%, and we can also use its confusion matrix output to show how well it predicted correctly or incorrectly predicted.

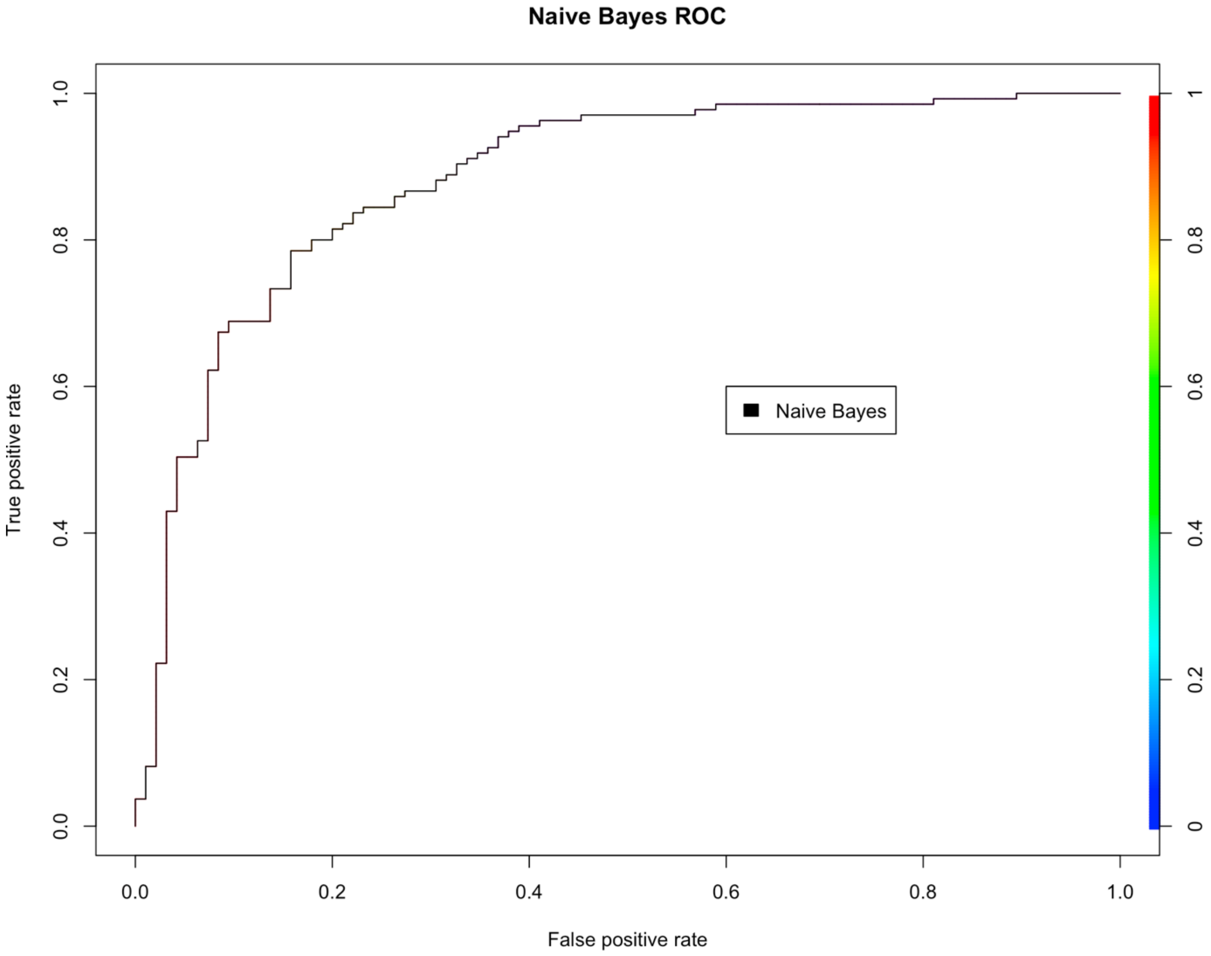
   b. **How many people are misclassified as developing heart disease? Where is this answer found in the output?**

21 people have been misclassified as developing heart disease but actually did not have it. This output was found when predicting using the testing dataset and the output was showing in the confusion matrix.



```
          Reference
Prediction  0   1
         0 73  21
         1 22 114
```

3) **Plot and submit the ROC curve for the class that develops heart disease. What is another measure of accuracy commonly used? Please provide this.**



The Area Under Curve (AUC) is another accuracy predictor. The AUC for the naïve bayes model was 0.8821053, as well as a confusion matrix can be used to determine accuracy.

Random Forest:

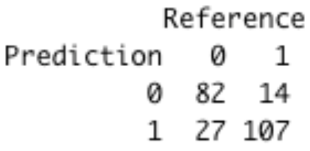
1) Did you undertake any prepossessing? If so, why?

The same dataset and train test split data was used as the naïve bayes. So all the pre-processing was the same as before. No additional processing was done.

2) Run the classifier with default parameters.

a. How accurately can the classifier predict those that develop heart disease? What is in the output that signifies this?

The random forest had an accuracy percentage of 84.35% with the training data.

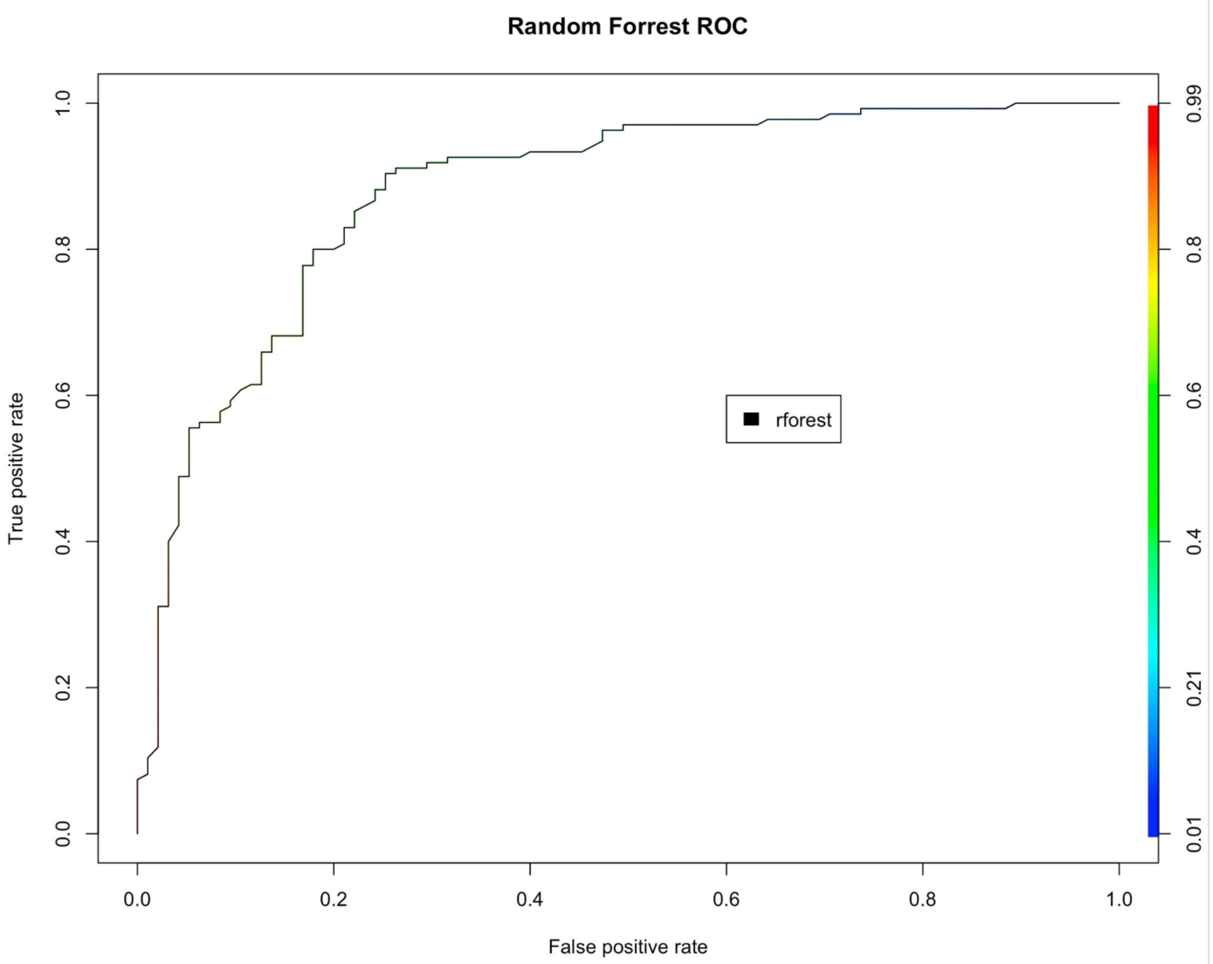


```
          Reference
Prediction  0   1
         0 82  14
         1 27 107
```

b. **How many people are misclassified as developing heart disease? Where is this answer found in the output?**

14 people were misclassified as developing heart disease. This answer was found in the top right of the confusion matrix, or the false positive section of the matrix.

3) **Plot and submit the ROC curve for the class that develops heart disease. What is another measure of accuracy commonly used? Please provide this.**



Another metric that can be used to measure accuracy is a confusion matrix as well as AUC (Area Under Curve). A CM has been shown above and the AUC for the random forest was 0.8786355.

Random Forrest (Model) Optimised:

1. **Why did you choose this classifier over the other?**

The random forest performed better than the naïve bayes model when compared on the ROC graph, AUC value and the accuracy percentage with training. The random forest was getting a higher true positive rate to a lowers false positive value compared to the naïve bayes model.

Due to the random forest doing better, it was decided to try and improve this model by changing some of the parameters.

2. **Briefly explain how this classifier works from a theoretical point of view.**

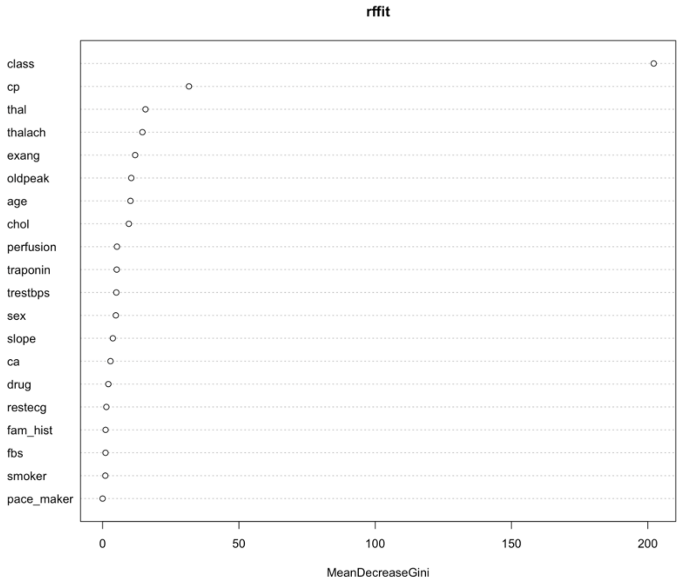
A random forest combines multiple decision trees (depending on the amount specified for training). The random forest then trains each of the decision trees on a slightly different set of the observations, splitting nodes in each tree considering a limited number of the features. The final prediction of the random forest is made by averaging the predictions of each of the individual trees.

Ref: https://towardsdatascience.com/an-implementation-and-explanation-of-the-random-forest-in-python-77bf308a9b76#:~:text=The%20random%20forest%20combines%20hundreds,predictions%20of%20each%20individual%20tree.

Ref: Hands on data science book

3. **Try to optimize the classifier to achieve a higher accuracy (no matter how small) than first found. Remember that we have a particular focus on predicting those that develop heart disease.**
    a. **Were there any features that could be removed? Please print the output that helped you make this decision.**

The value pacemaker overall had no importance on the predicting factor of the output. Therefore, this feature could be dropped with not having any impact on the model's performance.



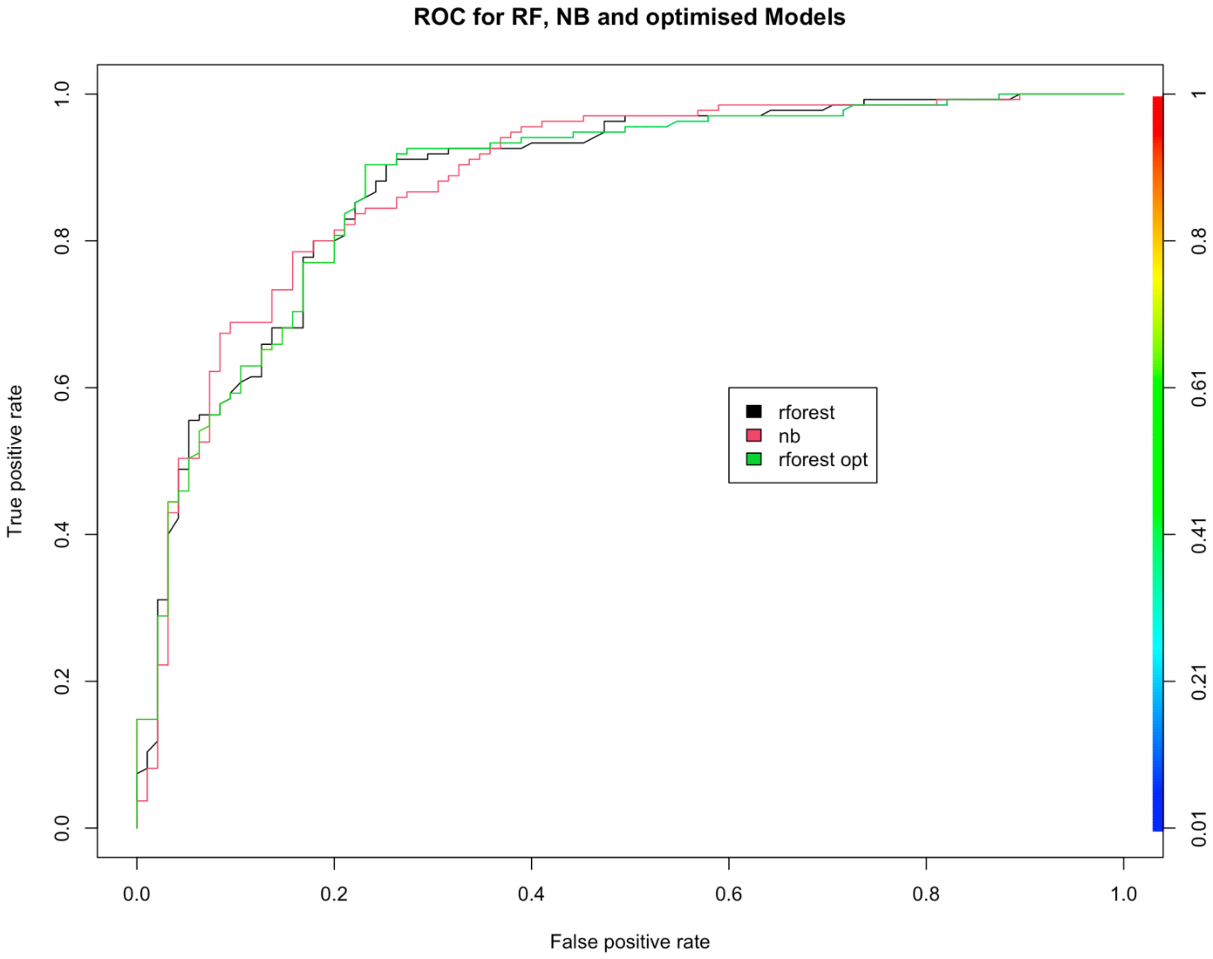
b. **Did changing the way data is sampled during training/testing affect the accuracy?**

Yes, the data can have an impact on the overall results. For example, the Random Forest produced a accuracy percentage of 82.17% but when the dataset's train and test data was reshuffled and reallocated this created a score of 84.35%. Which is not a huge difference of significance but a difference none the less.

c. What about some of the internal parameters specific to the classifier? Please explain how one of these parameters can affect accuracy.

Changing the parameters would have either a negative or positive effect on the model's predictions, as these parameters change the way in which the algorithms will work. For example, changing the number of ntree will change how many trees to grow in the forest, with the default being 5 trees.

In this case, when the random forest's parameters ntree, mtry and max nodes were changed to 800, 4 and 24 this created a slightly higher accuracy. The original random forest had an accuracy of 84.35%, error rate of 16.96% and an AUC value of 0.8786355. While the optimised model had an accuracy of 84.78%, an error rate of 15.94% and a AUC value of 0.8781676.



4. In general, a classifier is only as good as the data it is trained on. Please comment on what is needed from training data to train a good classifier. How can utilizing classifiers help feed back into healthcare settings with regards to data collection?