**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.

1. 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 Errors1* changed for different numbers of clusters?
   2. What can you conclude?
   3. How can you explain this conclusion from clustering analysis point of view?

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.

The data frame is then spilt into a train and test set while a 75-25% split to the data.

1. Run the classifier with default parameters.
   1. How accurately can the classifier predict those that develop heart disease? What is in the output that signifies this?
   2. How many people are misclassified as developing heart disease? Where is this answer found in the output?
2. Plot and submit the ROC curve for the class that develops heart disease. What is another measure of accuracy commonly used? Please provide this.

Chart

Description automatically generated

Random Forrest:

1. Did you undertake any prepossessing? If so, why?
2. Run the classifier with default parameters.
   1. How accurately can the classifier predict those that develop heart disease? What is in the output that signifies this?
   2. How many people are misclassified as developing heart disease? Where is this answer found in the output?
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.

Chart

Description automatically generated

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. 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 but changing parameters.

1. Briefly explain how this classifier works from a theoretical point of view.
2. 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.
   1. Were there any features that could be removed? Please print the output that helped you make this decision.
   2. Did changing the way data is sampled during training/testing affect the accuracy?
   3. What about some of the internal parameters specific to the classifier? Please explain how one of these parameters can affect accuracy.
3. 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?