**Part 1**

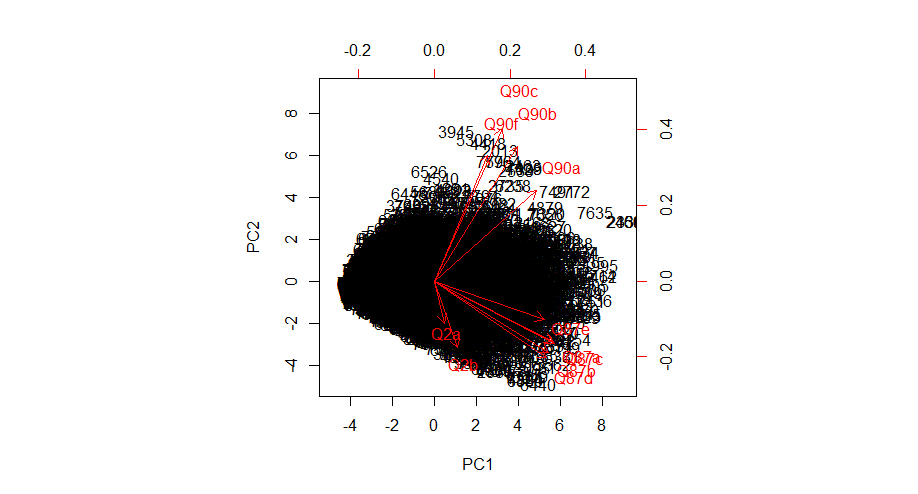
**1.1 Background**

In this part, a section of data from European Working Conditions Survey 2016 was extracted to form a dataset for analysis.

European Working Conditions Survey was a survey launched in 1990 to provide an overview of working conditions in Europe. The data set contains data on age, gender, subjective well-being and dimensions of engagement. Subjective well-being and dimensions of engagement were measured through a series of statements which respondents have to respond with a numerical value in the scale of one to six or one to five. In subjective well-being, there are five statements where respondents have to rate with numerical values between 1 to 6, with response 1 being all of the time, 2 being most of the time, 3 being more than half of the time, 4 being less than half of the time, 5 being some of the time, 6 being at no time. The statements are: 1) felt cheerful and in good spirits, 2) felt calm and relaxed, 3) felt active and vigorous, 4) woke up feeling fresh and rested, and 5) their daily life was filled with things that interest them. For Dimensions of Engagement, there were four statements with response options being numerical values between 1 to 5, with response 1 being always, 2 being most of the time, 3 being sometimes, 4 being rarely, 5 being never. The statements are 1) At my work I feel full of energy 2) I am enthusiastic about my job 3) Time flies when I am working 4) I am good at my job.

**1.2 Principal Component Analysis**

A principal component analysis could be conducted to provide a summary of data by reducing its dimensionality. A biplot can be constructed to visualise the data after principal component analysis is carried out with each observation plotted based on its principal component scores. Each number in the biplot represents an observation. From the bilpot, it can be observed that broadly there are two clusters of questions. First cluster contains Questions 87a-87e that represent questions regarding subjective well-being. Second cluster contains questions 90a-90f that represent questions regarding dimensions of engagement. The questions within subjective well-being cluster and employee engagaement cluster are located close to each other. That means questions within each cluster are correlated with each other.

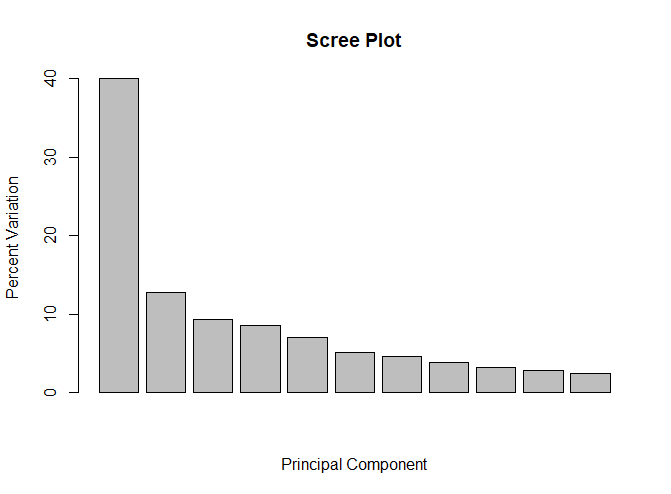


**Figure 1**: Biplot

From this biplot in Figure 1, it can be observed that the first loading vector places a heavy weight on questions in subjective well-being cluster, with much less weight on questions in engagement cluster. An observation with higher principal component one (PC1) value generally will score a high numerical value on questions in subjective well-being and high but slightly lesser numerical value on questions in subjective well-being cluster.

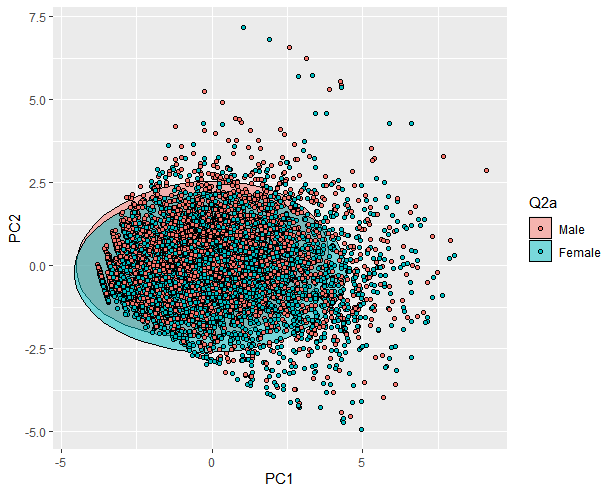
The second loading vector places most of its weight on employee engagement score and negative weight on subjective well-being score. An observation with higher principal component two (PC2) value generally will score a high value on questions in dimensions of engagement cluster and lower score on questions in subjective well-being cluster.

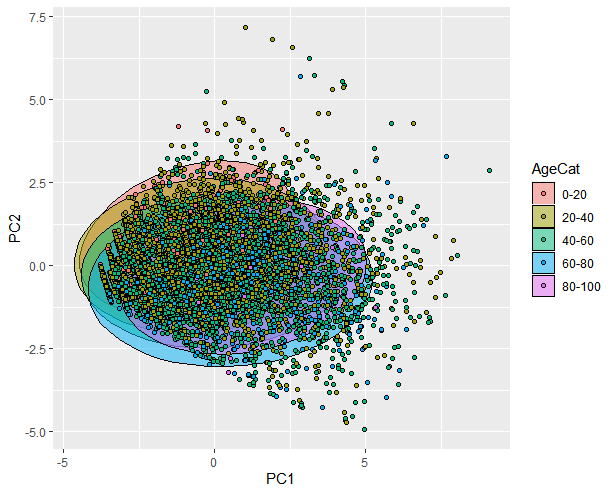
Next, a scree plot of the data is constructed to illustrate the proportion of the total variance in the data explained by each of the principal component.



**Figure 2**: Scree Plot

The first principal component explains about 40% of the variance in the data. The next principal component explains 13% of the variance. Together, the first two principal components explain more than half of the variance in the data, which means it could do a good job summarising the data.

**1.3 Data Analysis**

****

**Figure 3:** 95% confidence interval ellipse in Principal Component Analysis Score Plot

Apart from subjective well-being and dimensions of engagement, the dataset also contains gender (Q2a) and age (Q2b) which could form natural clusters for us to analyse our data. Before the analysis is conducted, gender data is converted into a categorical variable and age is separated into five distinct age groups. A plot is constructed where each observation is plotted based on its PC1 and PC2 score and a 95% confidence interval ellipse is drawn for each group of data. From the plot on the left, it can be observed that Male has slightly higher PC2 score. That implies that male as a respondent group represents a slightly more engaged workforce group with slightly lower subjective well-being compared to the female group. On the other hand, the plot on the right shows that younger workforce tends to have higher PC2 scores. That means younger workforce as a workforce group is more engaged in their work but have lesser subjective well-being compared to older workforce.

**1.4 Conclusion**

The dataset could be separated out into two principal components that could explain about 53% of the variance in the data. Based on the variation, the key observation is being younger and being male is associated with being more engaged at work and having lower subjective well-being. Further investigation should be done on the full set of data to find out more about the underlying factors.

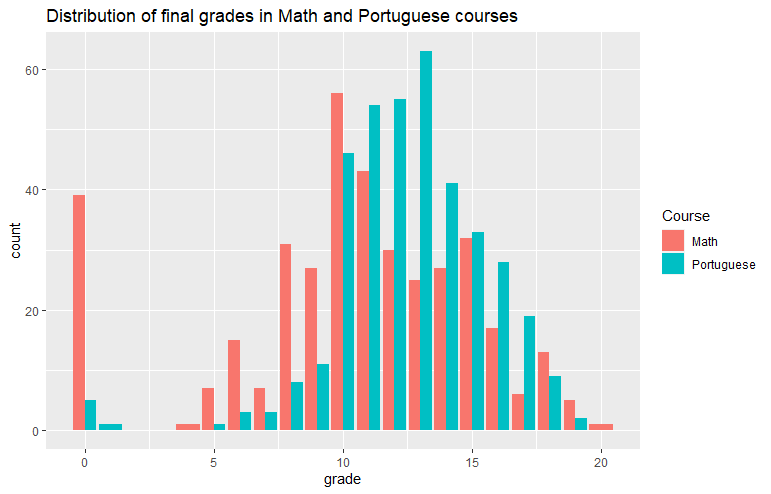
**Part 2**

**2.1 Background**

In part 2 two files containing data about student’s performance in two subjects, Mathematics and Portuguese language of two Portuguese school are given. For convenience, the two dataset will be known as Mathematics and Portugese dataset in this article. Various attributes that could affect student’s performance are included in the data, such as family size, parent's cohabitation status, mother's education, father's education and others. The objective here is to build a regression model for the variable G3 (final grade) without using the variables G1 (first period grade) and G2 (second period grade) and interpret the model and assess its predictive performance. Give the explicit instruction to exclude G2 and G3, they are removed from our analysis entirely from the start.

**2.2 Data Visualisation**

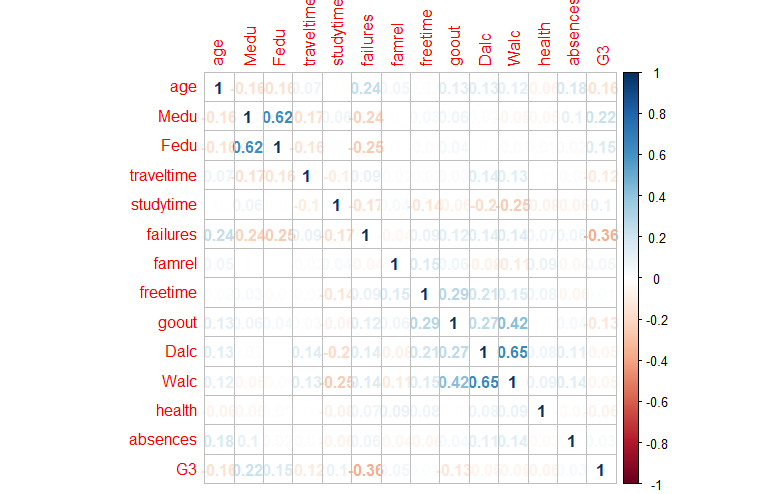
A histogram is constructed to visualise the distribution of G3 (final grade), which is the variable to be predicted under the regression models.



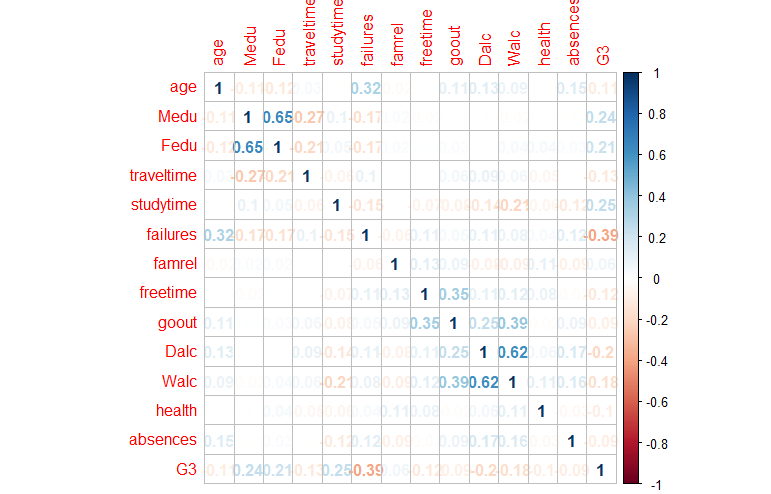
**Figure 4**: Histogram for distribution of final grades in Mathematics and Portuguese courses

The histogram displays a distribution that looks somewhat like a normal distribution. A huge majority of data falls in the middle. One interesting thing to note is that there are quite a high number of students who are getting a 0 for their final grade for mathematics.

Next, correlation plots for both datasets are constructed to provide insight into the relationships between each feature.



**Figure 5**: Correlation matrix of student performance in Mathematics dataset



**Figure 6**: Correlation matrix of student performance in Portuguese dataset

The correlation matrix shows that the correlations between each feature are quite similar across the two different datasets. From the matrix, it can be observed that the number of past class failures has the highest correlation with final grade G3 (-0.36 for Mathematics and -0.39 for Portuguese). Given that the correlation is negative, it can be suggested that people with a higher number of past class failures would be less likely to get a higher score on the final examination.

**2.3 Multiple Linear Regression Model**

Multiple linear regression is a linear approach for predicting a quantitative response on the basis of a number of predictor variables. For both Mathematics and in Portuguese language dataset, the data is split into training set and test set. In this case, 70% of the data is allocated as training set and 30% of the data is allocated as test set. A multiple linear regression model would then be fitted through the training set of each dataset. The output and summary of the models will be described in the following sections.

**2.3.1 Mathematics dataset output**

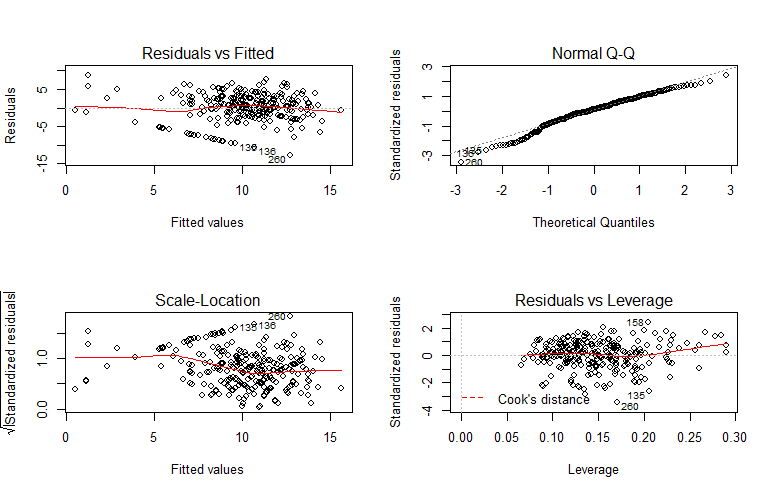
Number of past class failures (Coefficient: -1.73), which is a highly significant predictor, is significant at 0.001 level. The coefficient of -1.73 is quite highly negative, suggesting a strong negative relationship between number of past class failures and final grade. It means for every increase in past class failures, a final grade decrease of -1.72 under this model would be predicted. Two variables, romantic relationship and number of school absences are significant at 0.05 level. The coefficient is quite highly negative for romantic relationship but weakly positive for number of school absences. In simple terms, having a romantic relationship or having a high number of past class failures are predictive of lower final grades for Mathematics based on the model. The F-statistic is significant for this model, which indicates that there is evidence that there is a relationship between the predictors and the response.

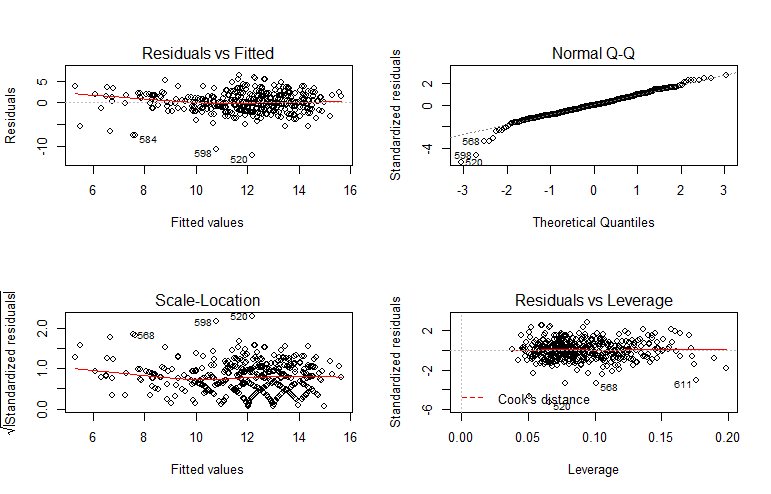
**2.3.2 Portuguese dataset output**

For Portuguese dataset, three variables which are number of past class failures (Coefficient: -1.45), wanting to take higher education (Coefficient:2.33) and school type (Coefficient: -1.04) are highly significant at 0.001 level. Three variables, age (Coefficient:2.33), number of school absences (Coefficient: -1.04) and extra educational support (Coefficient: -1.12) are significant at 0.01 level. Two variables, which are being a male (Coefficient: -0.61) and extra paid classes (Coefficient: -1.07) are significant at 0.05 level. Broadly, it can be inferred that having a higher number of past class failures, coming from Mousinho da Silveira school, having a higher number of school absences, needing extra paid classes and educational support, being a male are predictive of lower final Portuguese grade in the model. Whereas wanting to take higher education and being slightly older are predictive of higher final Portuguese grade in the model. The F-statistic is significant for this model, which indicates that there is evidence that there is a relationship between the predictors and the response.

**2.4 Diagnostics Plots**

Diagnostics test on how well our model is fitting the data could be tested by running diagnostics plots.

 **Figure 7**: Diagnostics Plots of Mathematics dataset



**Figure 8**: Diagnostics Plots of Portuguese dataset

There are four diagnostics plots for each regression on each dataset. The plots are quite similar for both datasets. Thus, the following descriptions are applicable for both datasets. From the first plot (Residuals vs Fitted), the data appear to be well modelled by a linear relationship. From the second plot (Normal Q-Q), it looks like the residuals are normally distributed. In the third plot (Scale-Location), the residuals appear randomly spread, which means that the assumption of equal variance (homoscedasticity) seems to hold. In the fourth plot (Residuals vs Leverage), the cook’s distance lines could hardly be seen, which means there is no influential case of data on the regression.

**2.5 Predictive Performance**

To assess the predictive performance of the models, the trained models are used to predict the values of G3 final grade on the test set. For each dataset, the full model is first used for prediction. Then, a regression with only the significant variables is used for prediction instead, as identified in the regression output. The model will be called partial model for convenience. The partial model of Mathematics dataset model contains number of past class failures, romantic relationship and number of school absences. The partial model of Portuguese dataset model contains school type, number of past class failures, extra educational support and wanting to take higher education. We use Mean Squared Error (MSE) to evaluate the predictive performance of the models. Lastly, a prediction is made on the other dataset using both the full and partial models. For example, mathematics dataset model is used to predict the final grade of Portuguese. The purpose is to assess how well the model generalise to other datasets.

**2.3.1 Mathematics dataset model**

The MSE obtained for prediction on test set using full model is 19.53. When the full model is used to predict the final grade on Portuguese dataset, an MSE of 16.84 is obtained

Next, the MSE obtained for prediction on test set using partial model is 20.35. When the partial model is used to predict the final grade on Portuguese dataset, an MSE of 12.06 is obtained.

**2.3.1 Portuguese dataset model**

The MSE obtained for prediction on test set using full model is 10.94. When the full model is used to predict the final grade on Mathematics dataset, an MSE of 20.82 is obtained.

Next, the MSE obtained for prediction on test set using partial model is 11.19. When the partial model is used to predict the final grade on Mathematics dataset, an MSE of 21.49 is obtained.

**2.6 Conclusion**

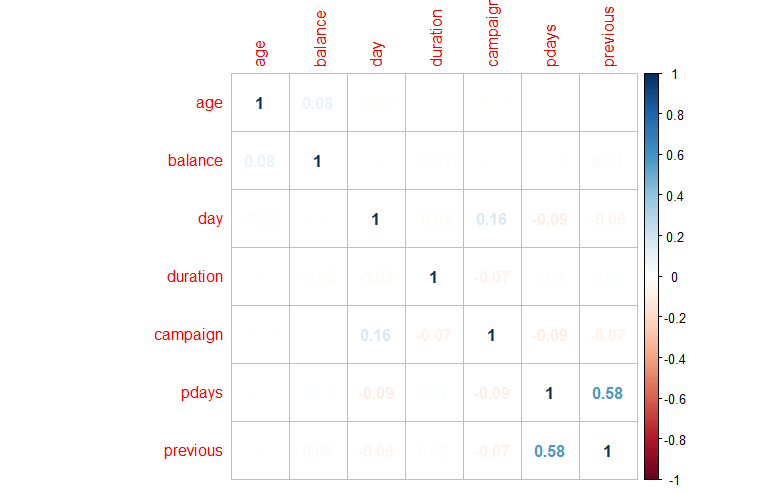
There are a few observations that could be made from the predictive performance results. In both Mathematics and Portuguese datasets, the MSE increases only slightly when variables that are not statistically significant are dropped. This means that a simpler partial model with only the statistically significant variables included can still provide comparative performance. When the multiple linear regression model fitted on Mathematics dataset is used to predict the final grade on Portuguese dataset, a slightly better results are obtained. But worse results are obtained when the multiple linear regression model fitted on Portuguese dataset is used to predict the final grade on Mathematics dataset. It seems the multiple linear regression models from these datasets don’t generalise to other datasets very well.

**Part 3**

**3.1 Background**

In part 3, a dataset regarding direct marketing campaigns via phone calls of a Portuguese banking institution is given. The goal in this part is to predict whether clients under the direct marketing campaign will subscribe (yes/no) to a term deposit (variable y). There are 16 variables in total. Some are quantitative variables like number of contacts performed during this campaign and number of days that passed by since last contacted from a previous campaign. Some are categorical variables like whether the client has a housing loan, marital status and contact communication type.

**3.2 Data Visualisation**

First, correlation matrix for the Bank Marketing dataset is constructed to generate a visualisation of relationships between each quantitative feature.

**Figure 9**: Correlation matrix of numerical variables of Bank Marketing dataset

From the correlation matrix, all the variables seem to be not strongly correlated apart from correlation between number of days that passed by after the client was last contacted (pdays) and number of contacts performed before this campaign(previous).

**3.3 Classification models**

Several classification techniques will be used to generate a prediction on whether the client would subscribe to a term deposit. They are logistic regression, linear discriminant analysis (LDA) classifier, quadratic discriminant analysis (QDA) classifier, K-nearest neighbours (KNN) classifier and classification trees. Last contact duration is removed from the data before any of the classification is performed to ensure that our classification model is realistic. This is because duration is not known before a call is performed and thus it highly affects the output target. If duration equals 0, then there will never be any subscription (y=0).

**3.4 Description and Predictive Performance of each model**

In the following section, a brief description of each model will be given and followed by its predictive performance on the Bank Marketing Dataset.

**3.4.1 Logistic Regression Model**

Logistic regression uses a logistic function to model a binary dependent variable. First, a logistic regression models is generated using the whole training set. From the summary, it can be observed that that three variables, being married, having a successful previous marketing campaign and contacted during month of October are highly significant at 0.001 level. Running through the prediction on the whole data, we obtain a training accuracy of 89.32% or training error rate of 11%. Next, we split 70% of the data as training set, and 30% as test set. The logistic regression model is fitted on the new training set. From the summary, we observed that four variables, which are contact communication type, having a successful previous marketing campaign, contacted during month of October and number of contacts performed during this campaign are highly significant at 0.001 level. This model is then used to make prediction on the variable y on test set. Here, the accuracy of 89.10% was obtained, slightly lesser than the training accuracy as expected. Another logistic regression model is fitted with only variables that are the most significant, which include contact communication type, outcome of the previous marketing campaign, last contact month of year and number of contacts performed during this campaign. Here, accuracy level of 88.75% is obtained. It implies that a simpler model with four predictors can do a better job than a model which contains all the predictors. Therefore, this partial model with four predictors will be used for the next few classifiers.

**3.4.1 Linear Discriminant Analysis (LDA) classifier and Quadratic Discriminant Analysis (LDA) classifier**

Both LDA classifier and QDA classifier result from assuming that the observations from each class are drawn from a Gaussian distribution, and plugging estimates for the parameters into Bayes’ theorem in order to perform prediction. The difference is that QDA assumes that each class has its own covariance matrix. For this dataset, linear discriminant analysis classifier results in an accuracy of 88.11% and quadratic discriminant analysis results in an accuracy of 86.98%. From this we could infer that linear discriminant analysis seems to be capturing the true relationship more accurately.

**3.4.2 K-Nearest Neighbours (LDA) classifier**

K-nearest neighbours (KNN) classifier classifies data based on the closeness of the data. K-Nearest Neighbours classifier with three different decision boundaries (K=1,10 and 100) are used to perform classification on the dataset. The results for K=1,10,100 is 88.04%, 88.75% and 88.53%. We can infer that the decision boundary of K = 10 gives us the best performance.

**3.4.3 Classification Tree**

Classification tree classify data based on the most commonly occurring class. 88.96% accuracy was obtained when classification tree classification is tried on the dataset. After pruning the tree, the accuracy remained the same.

**3.5 Conclusion**

Among all the classifiers, Classification Tree gives the best prediction in this bank marketing dataset. Its performance is closely followed by logistic regression, linear discriminant analysis (LDA) and K-Nearest Neighbours (LDA) classifier. Quadratic Discriminant Analysis (QDA) classifier yields the worst predictive performance in this dataset.