**PROBLEM 1**

First, we import the necessary libraries in order to proceed with the tasks



We read the Employee dataset



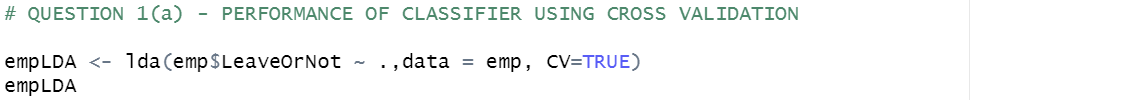
We need to convert all the string variables which are categorical into numerical variables

Text

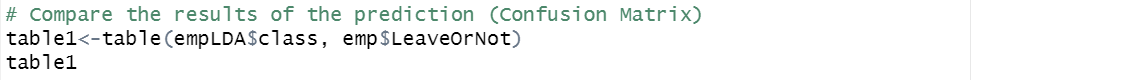
Description automatically generated

**QUESTION 1(a) - What is the performance of the classifier using cross-validation?**

Now we create the LDA model for LeaveOrNot variable. This LDA is based on Cross Validation. Hence, CV attribute of LDA function is set to True.



We compare the results by confusion matrix which is in a form of a table



Below is the output of the above code that clearly shows confusion matrix

Text

Description automatically generated

We must calculate the accuracy of the model. For that we use the below snippet of code



Below is the output for the accuracy of the model



The accuracy comes out to be 68.51% which is not that great. There are ways through which we can improve the accuracy and bring it to the acceptable range of 75% and above.

**QUESTION 1(b) - What is the performance of the classifier using training and testing?**

We create a LDA classifier but this time we use the training and testing approach.

We set the seed so that the train test split is not predetermined and is more random in nature.



Now we create a training and testing dataset from the main dataset but subsetting and splitting the dataset in 70:30 ration.

Text

Description automatically generated

Now we create the actual model using the training dataset



We predict the target variable using the above model on the testing dataset. We store the outcome statistics as a confusion matrix in a table.



The outcome confusion matrix can be seen in the below output snippet

Text

Description automatically generated with medium confidence

To validate this classifier, we need both the accuracy and the precision score.

Background pattern

Description automatically generated with low confidence

The outcome statistics that is the accuracy and the precision of the classifier is 70.4% and 61.66% respectively which can be seen below

Background pattern

Description automatically generated with low confidence

**QUESTION 1(c) - Would certain misclassification errors be worse than others? If so, how would you suggest measuring this?**

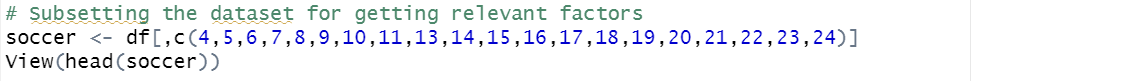
Misclassification is always an issue when it comes to classifiers. Wrong classification always leads to unfair and biased actions. In this case there are more 400 instances where the outcome was wrongly classified. In this particular case, lets assume if an employee is predicted to leave, then a company might unjustly fire that employee just because the model predicted it. It will surely lead to a case of wrongful termination. On the other hand, if an employee wants to leave but the model predicted that the employee should not leave, then the company might employ an employee who wishes to leave. This might result in poor performance from the employee and may affect the overall performance of the company. In conclusions, misclassification errors are certainly worse and may lead to serious negative consequences.

**PROBLEM 2**

First, we read the dataset



Then we subset the dataset to get the relevant factors

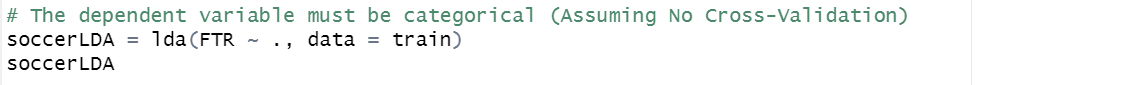


We then set the seed to ensure that the training and testing split is random. Then we split the dataset in 70:30 split ratio.

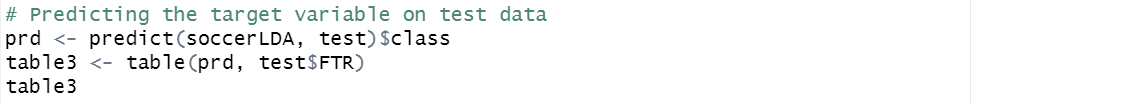
Text

Description automatically generated

We finally create a model using the training dataset



Now we predict the target variable on test data. We then save the performance and validation metrics in a table as a confusion matrix. We can clearly see the number of correctly classified outcomes in the confusion matrix below. The accuracy of the classifier is 78.07%.



Background pattern

Description automatically generated with low confidence

**PROBLEM 3**

The journal article titled ‘A cluster analysis of climate change mitigation behaviors among SMTEs (Small and Medium Tourist Enterprises’ aims to study the discrepancy between the proposal of climate mitigation methodologies, implementation and execution of climate mitigation methodologies, and innovation of climate mitigation methodologies among 400 and more accommodation providers from South-West England which is a major tourist attraction. Coles, Zschiegner, and Dinan (2013) also conclude that “One size fit for all” approach is inappropriate and counterproductive in mitigating climate change. The research used cluster analysis to measure the knowledge gap between proposal, implementation and execution, and innovation of climate mitigation strategies. For this methodology, the research relied on a survey that included a questionnaire to gather details of relationship between the Small and Medium Tourist Enterprises and environment, climate change, environmental practices, and business operation characteristics.

Three clusters were mainly formed and analyzed over the course of the research. The three distinctive clusters of Small and Medium Tourism Enterprise (SMTE) were identified based on how they innovated to mitigate climate change. The smallest (12%) had implemented a variety of procedural and management changes and was the most proactive and forward-thinking. A second cluster (23%), on the other hand, had implemented several process improvements, but its approach to management innovations was both incomplete and ambiguous. The largest cluster (65%) had mostly implemented simple process improvements but had failed to implement management innovations to assess, monitor, and act on their environmental performance. These findings show that the contribution of lodging providers to emissions reduction objectives has been at best negligible.

Cluster analysis aims at identifying heterogenous groups consisting of homogenous elements which allowed researchers to identify existence of distinctive groups of business exhibiting similar intra-group mitigation behaviors. The research followed a two-step approach mainly utilizing the Statistics Package for Social Sciences (SPSS) version 15. The first step employed a hierarchical approach using Ward method to determine the number of clusters. Later, in the second step K-means method was used to determine the cluster membership of 416 businesses that took part in the research. The research concluded and presented a very interesting finding that there were significant differences among clusters regarding the way in which constituents businesses managed energy and water. Businesses in cluster one was seriously concerned about mitigating climate change and were committed in monitoring and adopting to energy and water conservation measures. Businesses constituting cluster two were somewhat similar in adoption of measures in mitigating climate change as cluster one but did little in terms of innovation and following up after implementation of measures. On the contrary, businesses forming cluster three were disinterested in mitigating climate change. A very few businesses in cluster three introduced a low climate mitigating steps but it amounted to nothing. Overall, there was a significant gap, and this study conclusively proves that. Hence, giving a solid cause to introduce strict climate mitigation laws and policies.