**EAS 509**

**Statistical Leaning II**

**Project-1**

**Team Members:**

Sai Teja Dondeti (saitejad)

Sai Teja Sankarneni (saitejas)

Nishant Varma Indukuri (nindukur)

Shanawaz Pathan (spathan)

**DATA LOADING**

We have chosen the Land Mines dataset for this project, cause detection of mines which are hidden under the ground are crucial when it comes to the life of the people and their property.

In order to procced with this data set we need to know few statistics about it. We are loading the data from the second sheet of the **Mine\_Dataset.xls** into a data frame named ‘my\_data’.

This data set contains 4 features they are ‘V’, ‘H’, ‘S’ and ‘M’ for our reference now printing the top 20 rows of our data set.

Table

Description automatically generated with medium confidence

Now printing the structure of our data set including number of entries or rows and number of features or columns and the data type of each column.

Text

Description automatically generated

Our dataset contains ‘338’ rows and ‘ 4’ columns and the datatype of each column is numeric.

Getting more details and basic statistics of our data set like mean, median, min and max etc.

Text

Description automatically generated

**PREPROCESSING**

After knowing the basic knowledge of our dataset, we now perform data preprocessing in R by extracting, normalizing, and combining feature variables with a target variable into a new data frame.

Once we are done with the preprocessing, now we print the dimensions of our normalized dataset.

Graphical user interface, text, application, email

Description automatically generated

From this dataset we have 4 features they are ‘V’, ‘H’, ‘S’ and ‘M’ were ‘V’ is voltage , ‘H’ is high, ‘S’ is soil type and ‘M’ is mine type.

Table

Description automatically generated

From the above column we have chosen ‘M’ mine type as our predictor column and now fetching the unique elements in ‘M’.

Graphical user interface, application

Description automatically generated

From the above results we came to know that we have ‘1’,’2’,’3’,’4’ and ‘5’ as unique elements in ‘M’. (5 output labels).

**VISUALIZATION**

Finding the relation between the target column and predictor features. First we find the relation between voltage and mine type from the normalized dataset.

A picture containing chart

Description automatically generated

*Voltage vs Mine Type*

From the above graph the result is gives the scatter plot between the ‘V’ and ‘M’.

Now we find the relation between high and mine type from the normalized dataset.

**A picture containing chart

Description automatically generated**

*High vs Mine Type*

From the above graph the result gives the scatter plot between the ‘H’ and ‘M’.

Now we find the relation between soil type and mine type from the normalized dataset.

Chart, scatter chart

Description automatically generated

*Soil Type vs Mine Type*

From the above graph result is gives the scatter plot between the ‘S’ and ‘M’. from all the above results it looks like the relation between soil type and high with mine type is quite similar.

**CORRELATION MATRIX**

Now we will calculate the correlation matrix of the columns in normalized\_data.

This means that for each pair of columns in normalized\_data, the function will calculate the correlation coefficient between them, which measures the strength of the linear relationship between the two variables. The resulting correlation matrix will be a square matrix where each entry (i,j) represents the correlation coefficient between the i-th and j-th columns of normalized\_data.

The values in the correlation matrix will range between -1 and 1, where a value of 1 indicates a perfect positive correlation between two variables, 0 indicates no correlation, and -1 indicates a perfect negative correlation.

Graphical user interface, text, application

Description automatically generated

**VOLTAGE TYPE DISTRIBUTION**

Now we plot the distribution of voltage over the mine type column which gives the line graph of the V and M variables in my\_data, with a violet line connecting the data points.

Chart, histogram

Description automatically generated

*Voltage Type distribution*

**BAR CHART TO VISUALIZE DIFFERENT MINE TYPES**

In this part we are going to count the number of mine types with each of their unique value of ‘M’.

Chart, bar chart

Description automatically generated

*Count of each mine type*

From the above graph ‘1’ mine type have around 70 and mine type ‘2’ have the count nearly 68 whereas the mine type ‘3’ ,’4’ and ‘5’ have the similar count of 65.

**BOXPLOT TO FIND THE OUTLIERS**

Now from the normalized dataset we find the outliers of each column.

Chart, box and whisker chart

Description automatically generated

*Boxplot of each variable*

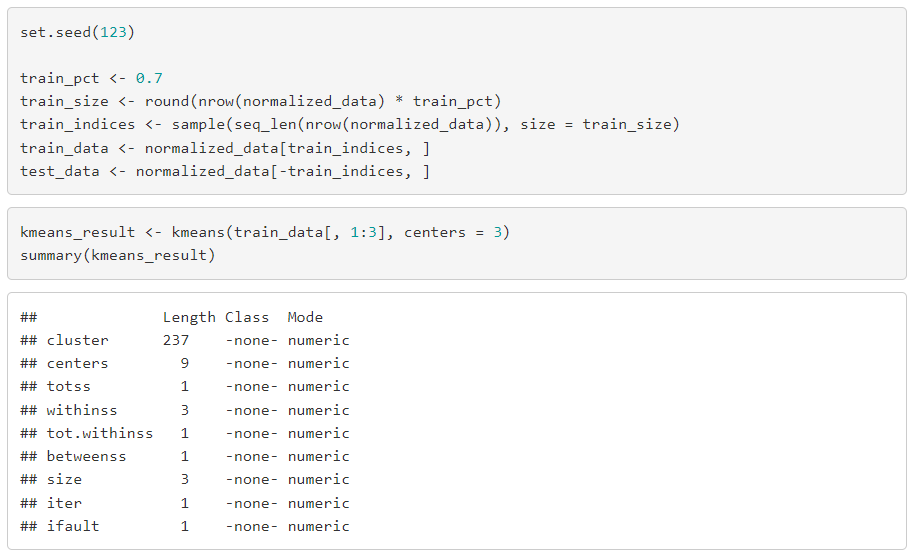
From the above graphical results, it is clear that voltage has very minimal outliers whereas the rest of the columns are clear from the outliers.

**ALGORITHMS PERFORMED:**

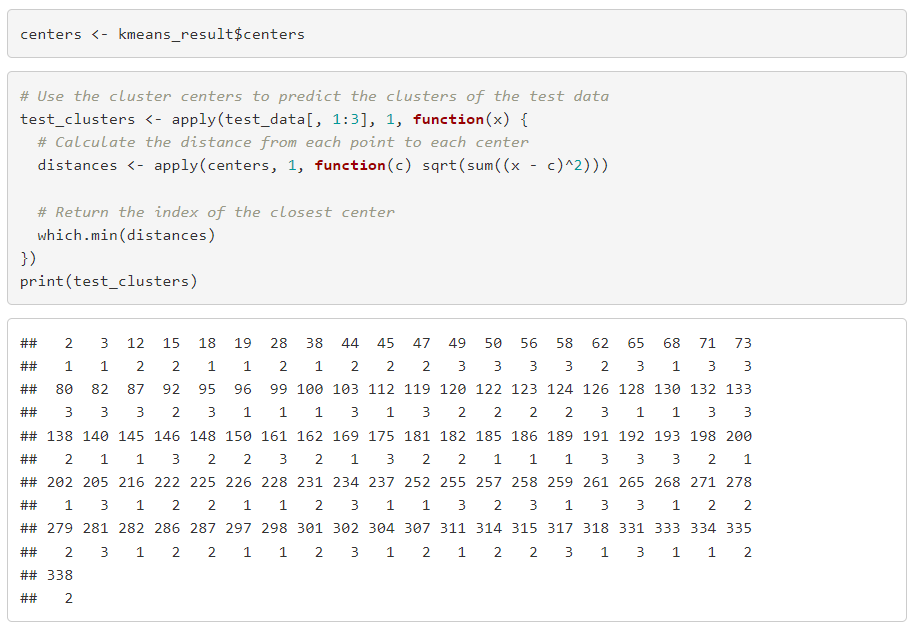
1. **K – MEANS CLUSTERING**
2. **HIREARCHIAL CLUSTERING**
3. **DECISION TREE ALGORITHM**
4. **NEURAL NETWORKS**
5. **K – NEAREST NEIGHBOURS (KNN)**
6. **MULTINOMIAL LOGISTIC REGRESSION**
7. **NAÏVE BAYES**
8. **K- MEANS CLUSTERING**

It is a popular unsupervised machine learning algorithm which works on the basic idea of grouping similar data points together and discovering underlying patterns for which we choose a fixed number of clusters in a dataset.

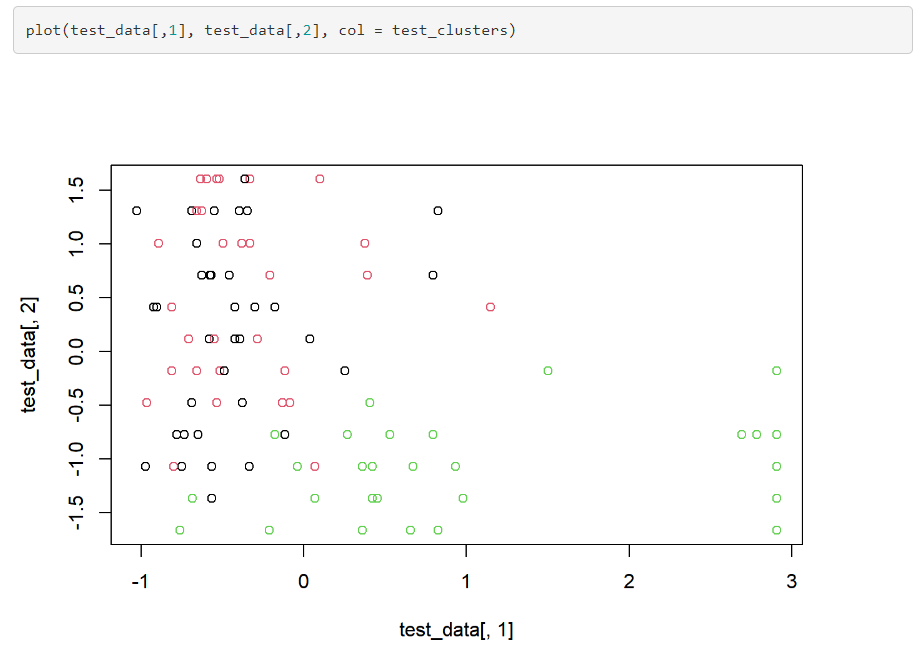
We first choose a target number of centroids needed in a dataset. Then every data point is allocated to each of the clusters through reducing in the cluster sum of squares. It then comes to a halt when there is no change in their values or the defined number of iterations have been achieved. After final reassignment we name the cluster as the final cluster.



* After splitting the data into test and train data we, perform K-Means clustering on it.



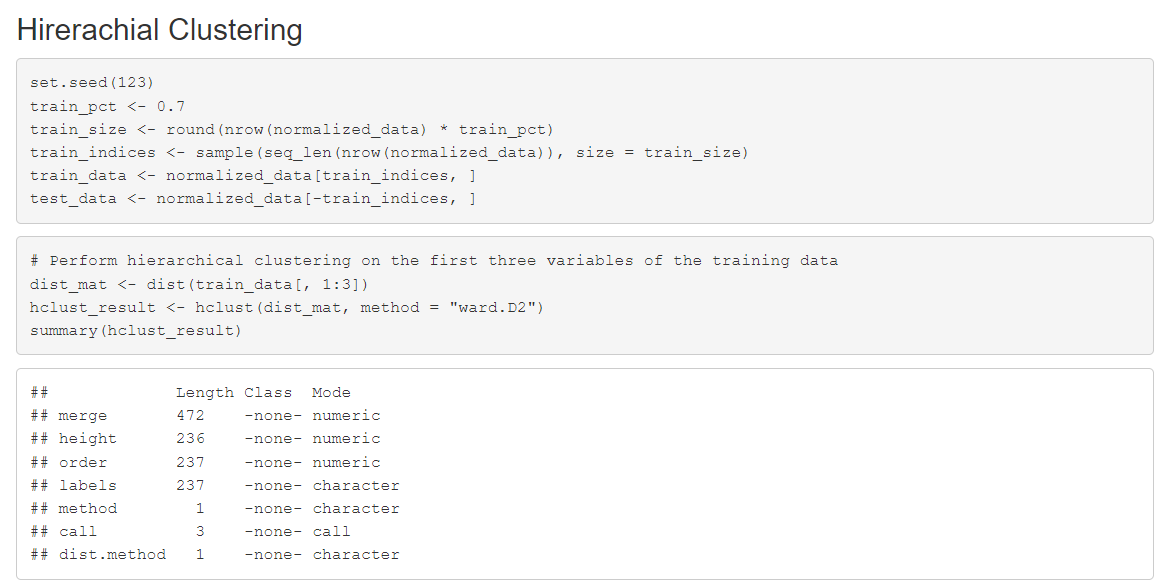
* After the data has been divided into clusters of similar types, we, print the clusters from the test data as the final clusters.
* We then plot a graph showing the division of data into clusters of test data 1 and test data 2



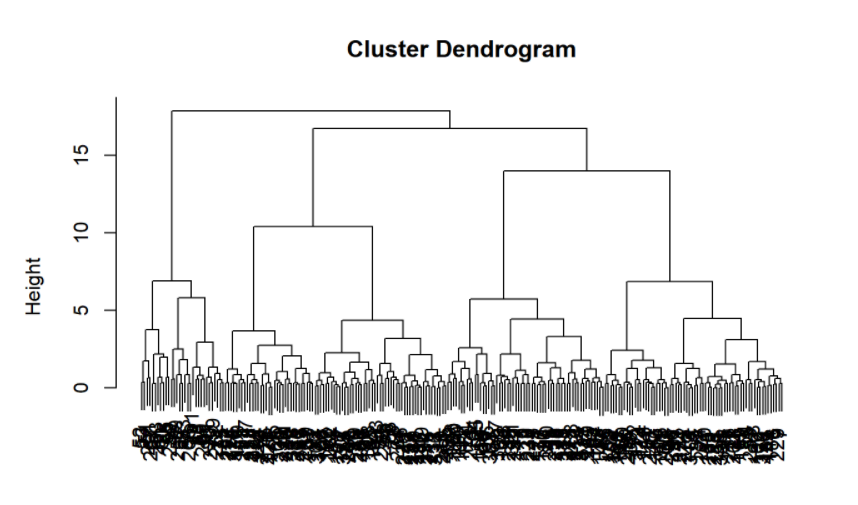
1. **HIREARCHIAL CLUSTERING**

A clustering algorithm known as "hierarchical clustering" creates a hierarchy of clusters by repeatedly breaking up a collection of objects into smaller subsets until distinct clusters are produced. It can be either agglomerative, in which case each object starts forming its own cluster and then all of the clusters are combined, or divisive, in which case all of the objects begin in the same cluster and are then separated repeatedly. Dendrograms, which show the hierarchical links between the clusters, are one way to visualize the results of hierarchical clustering.

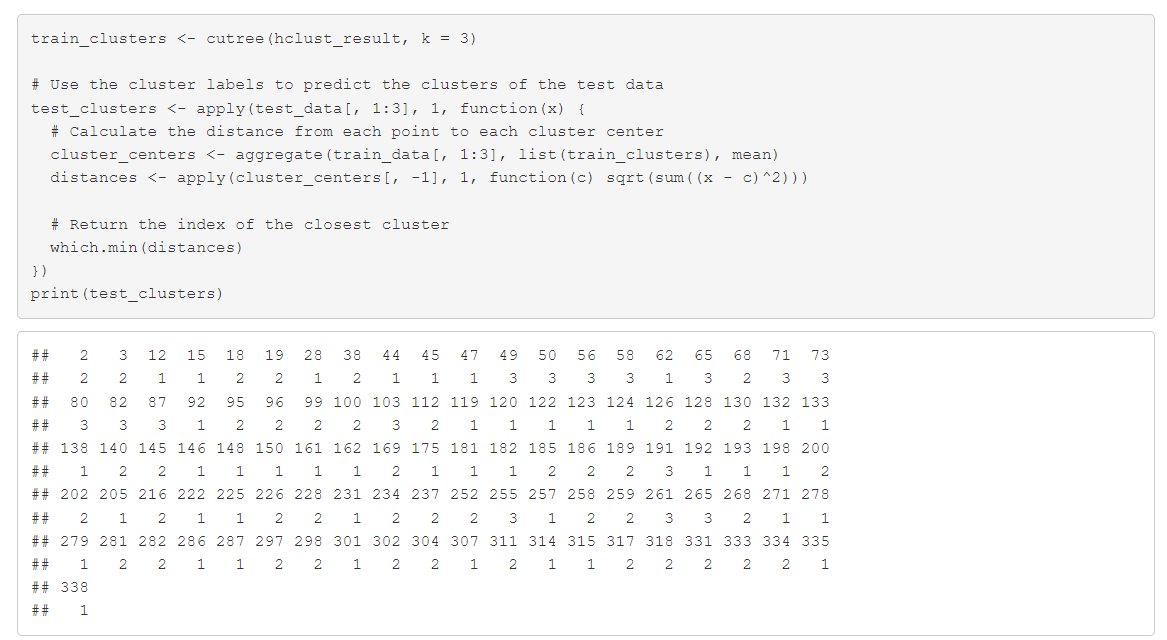
On the other hand, K-means clustering is a partitioning technique that seeks to group a set of objects into K clusters while minimizing the sum of squares within each cluster. It operates by updating the cluster centers after iteratively allocating each object to its nearest cluster center.

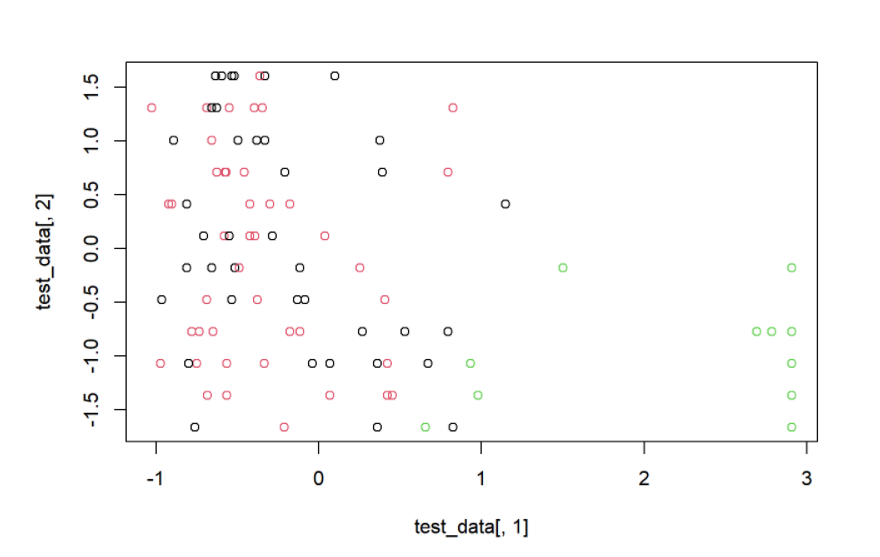


* First, we have randomly split the data into training and testing datasets.
* Then we performed hierarchical clustering on the first three variables of the training data using **Ward’s** method.



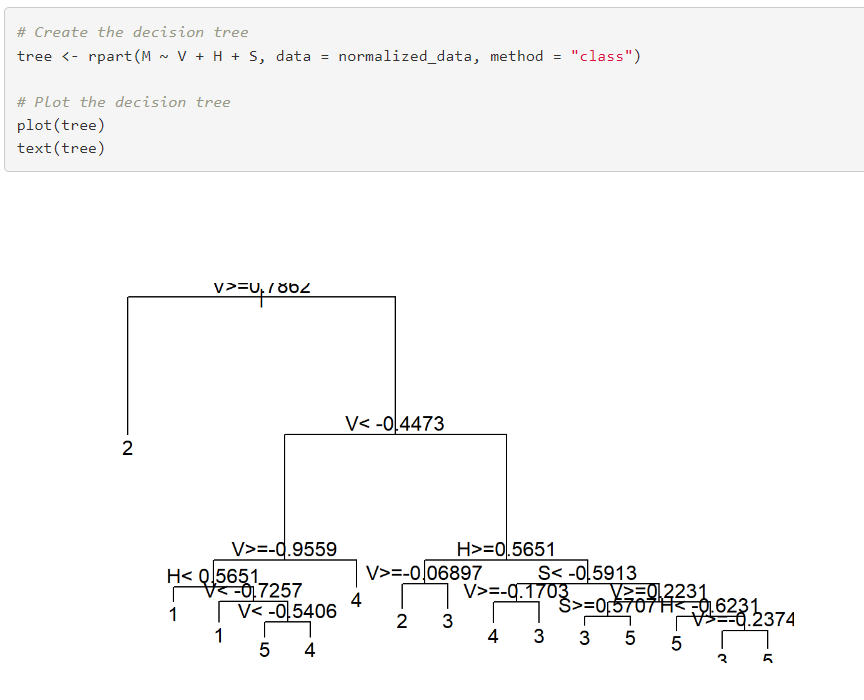
* We plotted the resulting dendrogram to visualize the cluster structure.
* Then we cut the dendrogram into 3 clusters and extract the cluster labels for the training data.

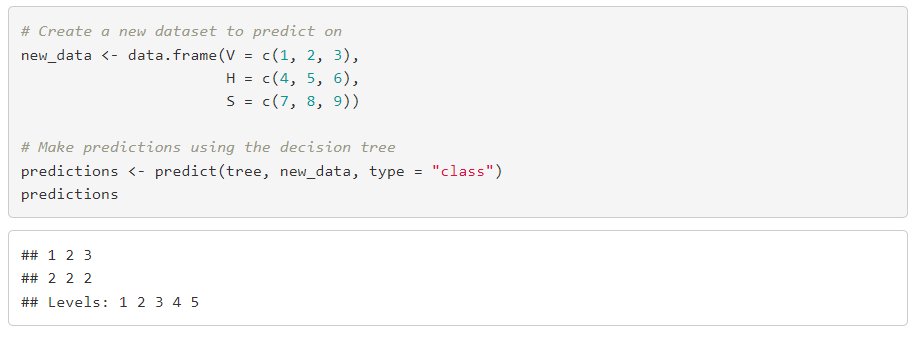




1. **DECISION TREE ALGORITHM**

Decision trees have the ability to perform classification tasks. They are characterized by nodes and branches where the tests on each attribute are represented at the nodes and the outcome is represented at the branches and class labels are represented at the leaf nodes. Hence it uses a tree-like model based on various decisions that are used to compute their probable outcomes.

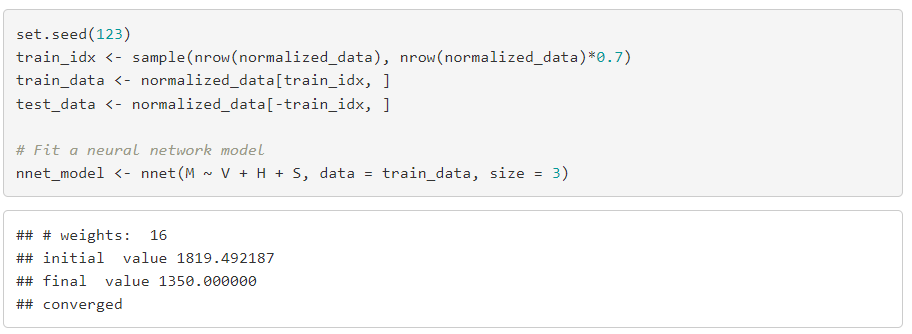




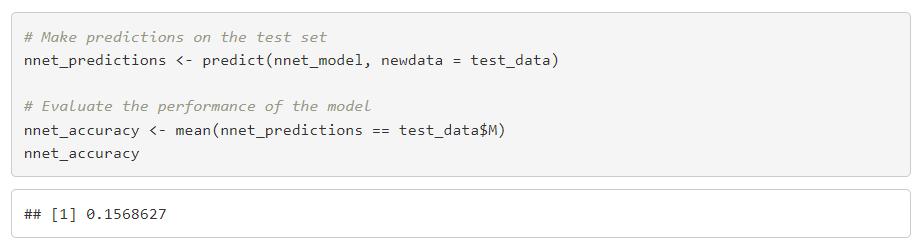
1. **NEURAL NETWORKS**

A Neural network is an information processing machine which is similar to human nervous system containing interconnected information processing units. They are used due to their power to parallel process the data.

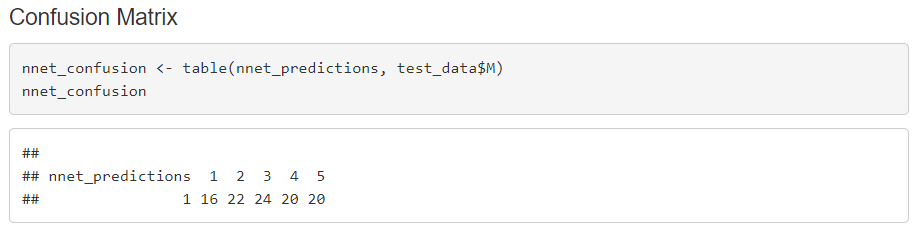
* The data is first split into training and testing data after normalizing it

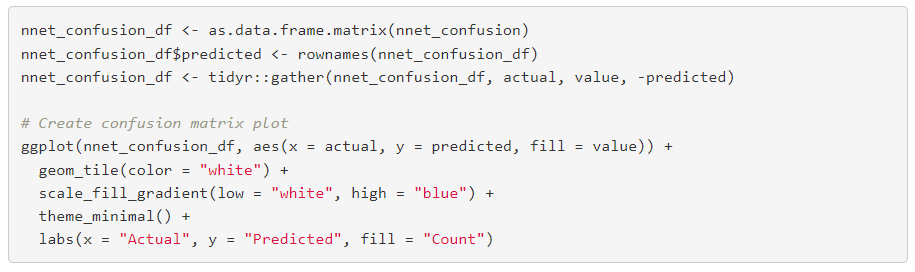


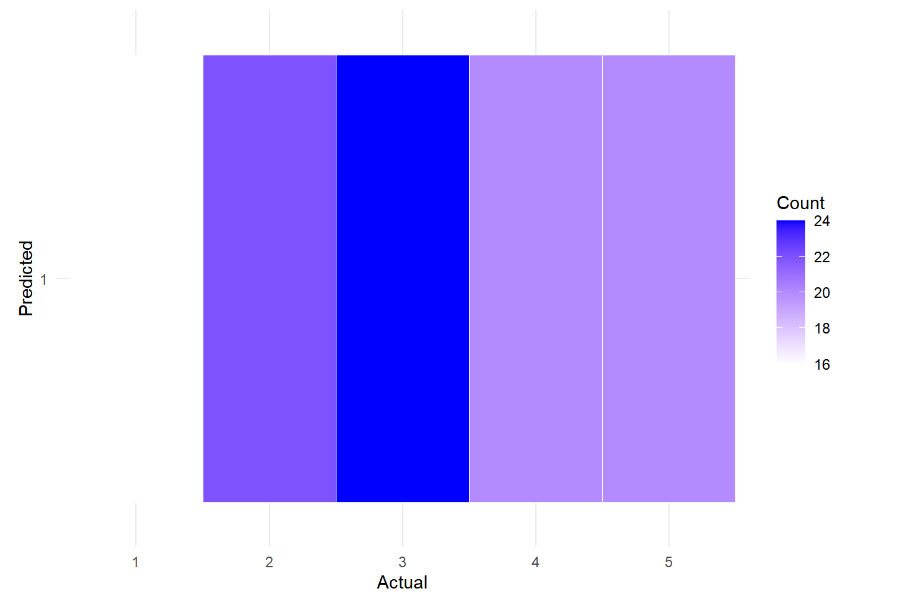
* After making predictions we have taken the accuracy of the predictions made by this model on the testing data and have got the following



* We have achieved an accuracy of 0.15686
* Through this we have deduced that neural network model isn’t optimal for this data set.
* Drawing a confusion matrix to analyze the outcomes.







1. **K-NEAREST NEIGHBOUR [KNN]**

It is a supervised machine learning algorithm which classifies a new data point into a target class based on the features of the neighboring data points. Which means that KNN checks how similar a data point is to neighbor or any other classes of data points, identifies which it is most similar to and classifies it into it.

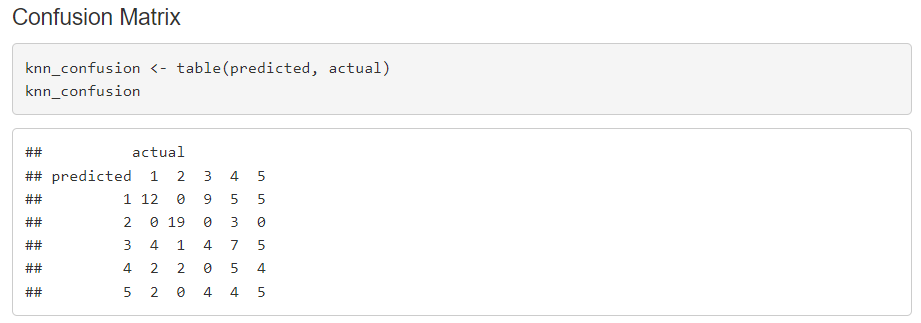
* First, we split the data into training and testing data and select the value of k which is number of neighbors to consider

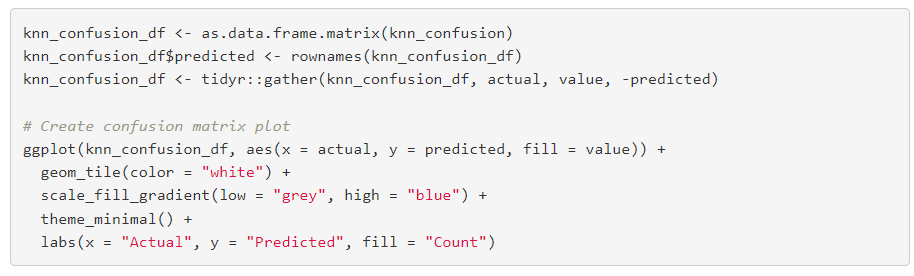


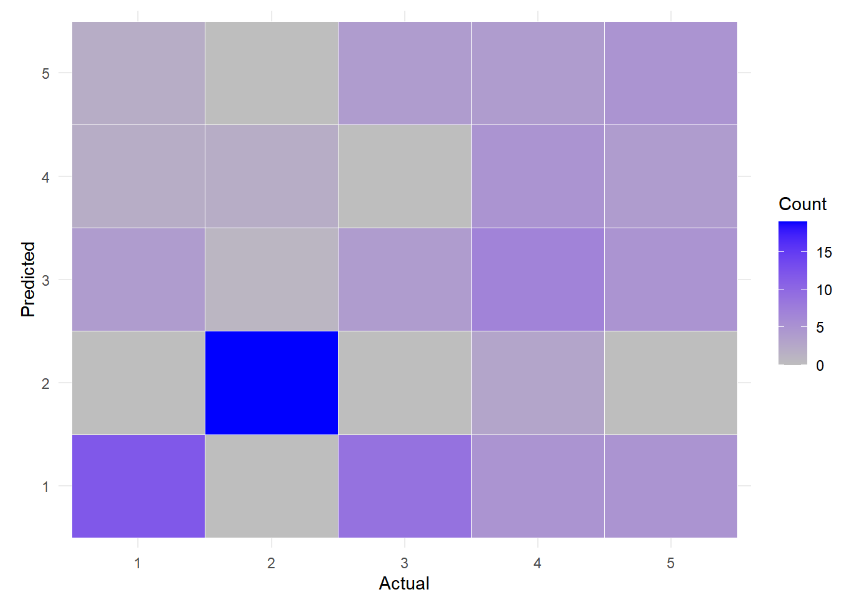
* Then after we have trained the model, we use the testing data to calculate the accuracy of the model



* We got an accuracy of 44% using KNN Model.
* Drawing a confusion matrix to analyze the outcomes.

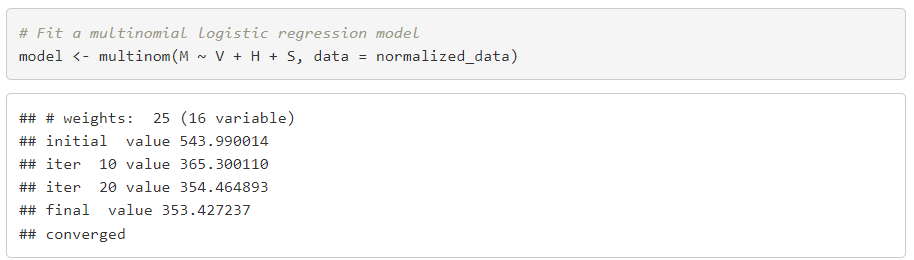
****

****

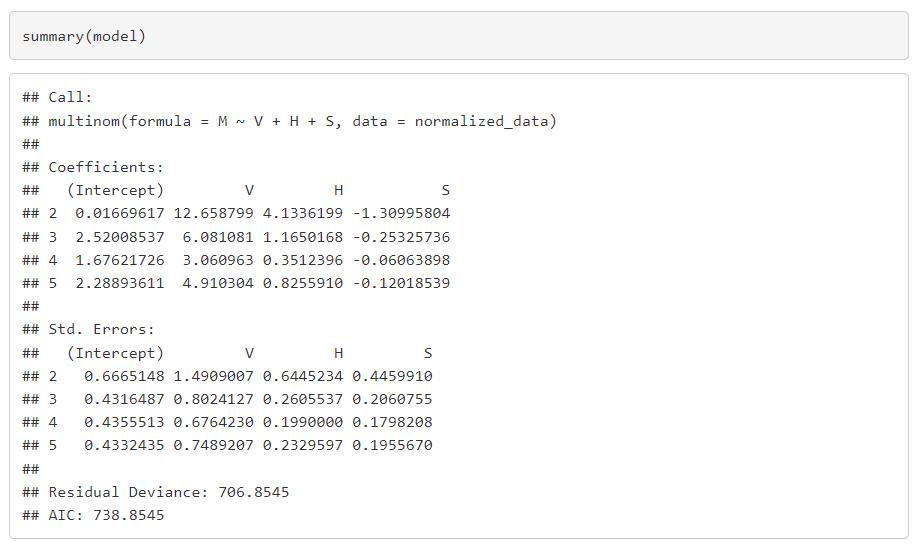
****

1. **MULTINOMIAL LOGISTIC REGRESSION**

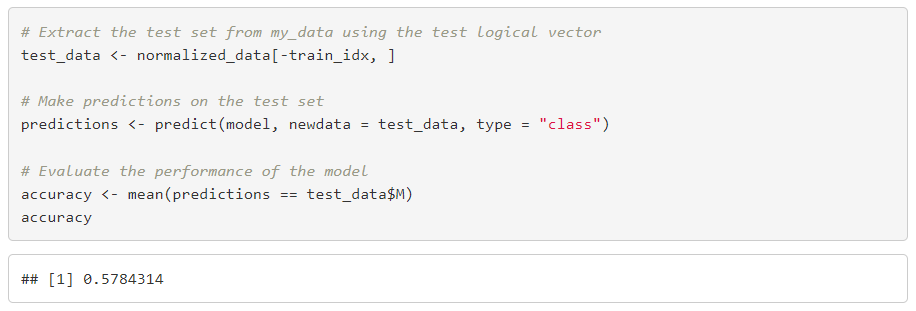
It is used to predict nominal target variable. Generally, this algorithm is used when the dependent variable is nominal of more than two levels. Used to describe data and explain relation of one nominal variable and a minimum of one continuous independent variable.



* Here is the summary of the model

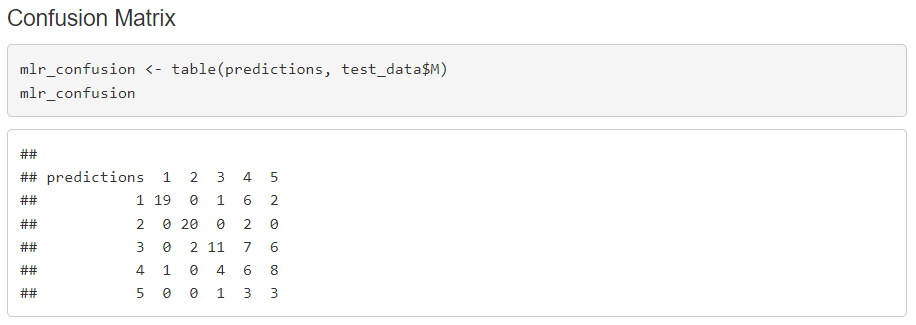


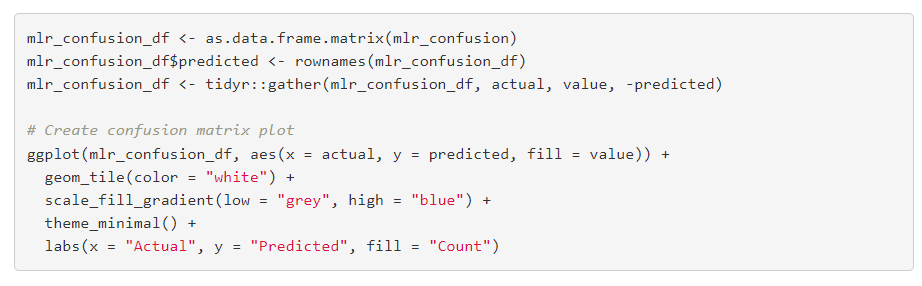
* Evaluating the performance of the model on the test data and taking the accuracy.

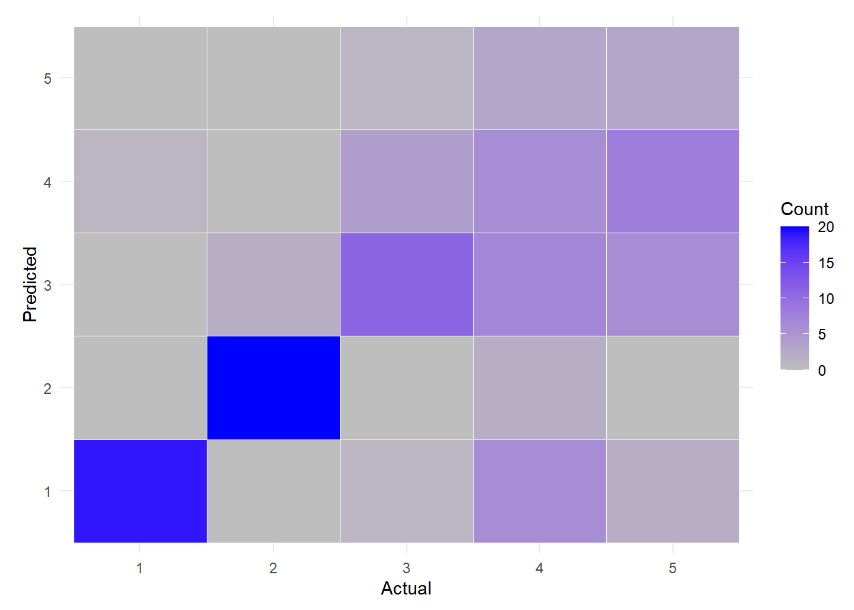


* As we can see the accuracy, we got is 57.8%

* Drawing a confusion matrix to analyze the outcomes.



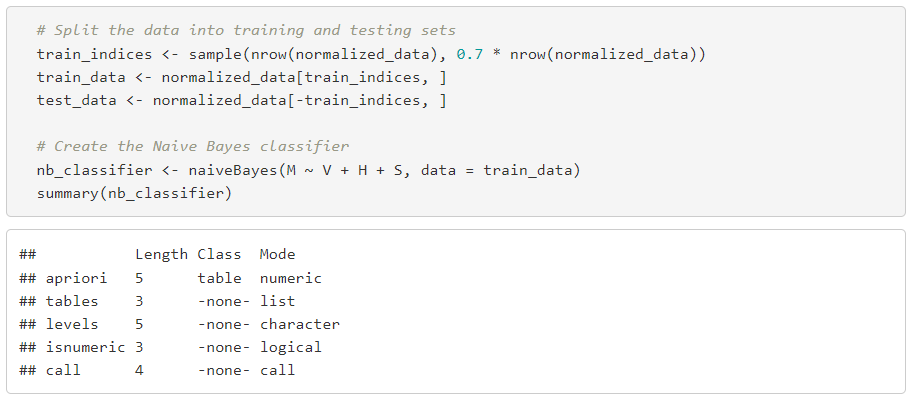




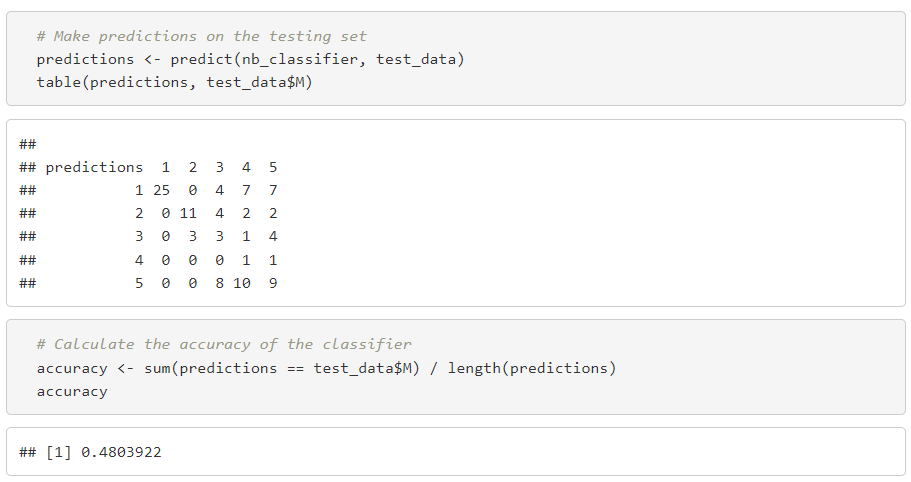
1. **NAÏVE BAYES**

Naive bayes is a supervised classification algorithm based on Bayes theorem which gives us the probability of an event A given that an event B occurred.

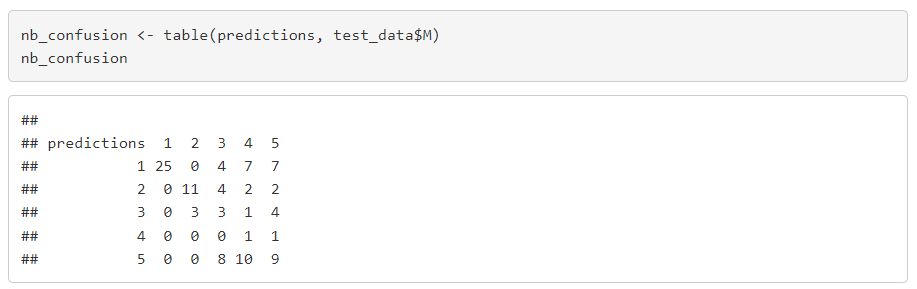
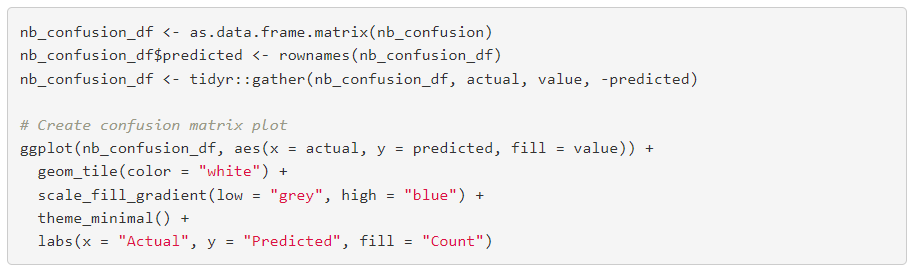
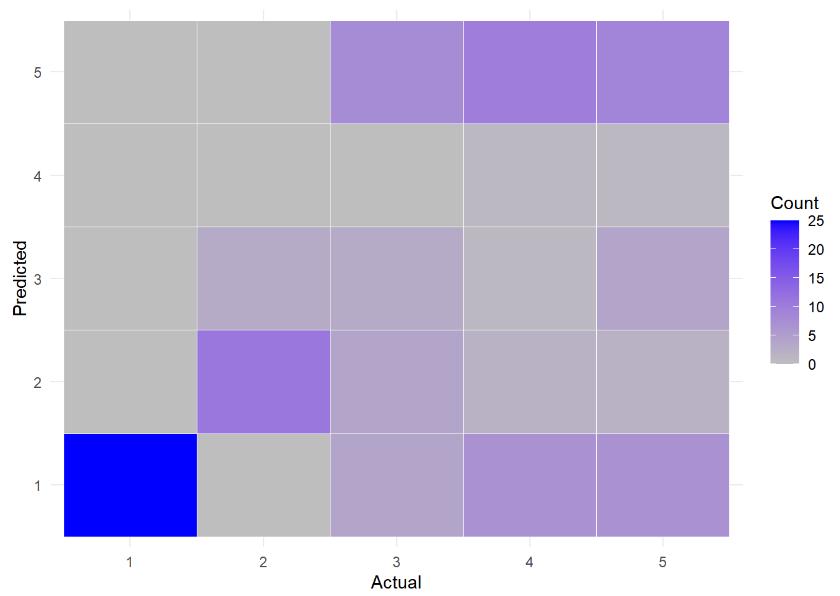
It is called Naïve as it Assumes that the occurrence of a certain feature is independent of the occurrence of other features.



* Evaluating the performance of the model on the test data and taking the accuracy.



* The accuracy we got is 48.03%
* Drawing a confusion matrix to analyze the outcomes.

**Conclusion:**

From the correlation matrix of the dataset, we can clearly observe that that output variable is not very much dependent on all the input features therefore, we are unable to attain great accuracies using the given “mine\_dataset”. We obtained best accuracy while performing Multinomial Logistic Regression.