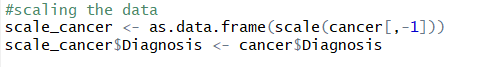
**Machine Learning for Data Science**

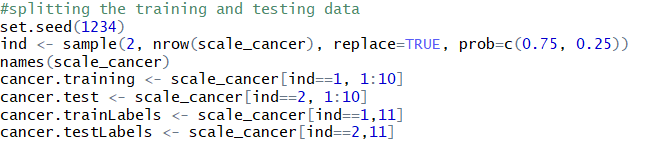
**KNN classification –**

**Steps to be followed:**

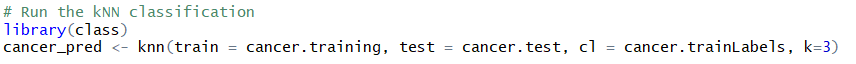
1. Loading the dataset into R studio and preprocessing it by scaling the dataset



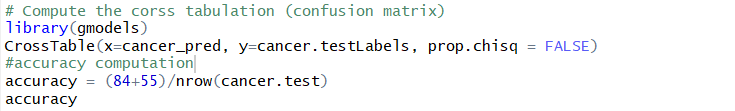
1. Split the dataset into 75% training and remaining into testing



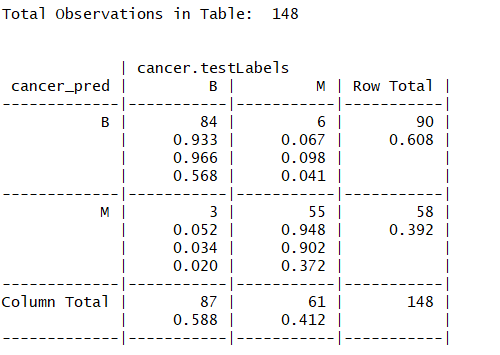
1. Fit the KNN model for classification for predicting the class of the ‘Diagnosis’ variable into Malignant or Benign with the three nearest neighborhood values(k=3)



1. Compute the confusion matrix and calculate the accuracy



1. We have the below confusion matrix and its accuracy as **93.9%**



Thus, the KNN model predicted the class of the cancer$Diagnosis to be **94% accurate**. Also, it misclassified the M as B for 6 times and B as M for 3 times. So, the error can be calculated as 1-accuracy = **0.0608(6% error).** This is overall a very good model with the given number of data.

**Question no:2**

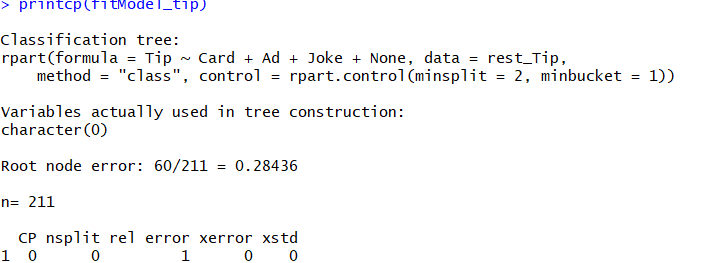
**Decision Tree-**

**Steps to be followed:**

1. Loading the dataset into R studio and visualizing the data
2. Including the rpart library for constructing the decision tree for the given data and building the model with it.



1. After fitting the model for the above decision tree, we see that the split of the tree is found to be 0, which means that the split is not happening.



1. Also, we are getting the error message as below when we try to plot the tree



1. This error message clearly indicates that the tree is not grown. It is taking all the observations as individual roots as the split is not happening which indicates that the split =0 and there are no variables getting involved in the tree construction.
2. Tried changing the cp value to be very minimal as cp=0.001, but still I don’t see the split is happening. This indicates that the rpart routine has not found enough information for splitting the tree into many leaf nodes.
3. As there is no much information obtained from the decision tree, with the given input dataset, let us make our decision rules by visualizing the data

**Decision Rules:**

Removing the Card variable from the dataset as it is just an indicator of the other variables and it will be an insignificant variable.

**R1: IF (Ad = 1) AND (Joke = 0) THEN Tip = 0**

**R2: IF (Ad = 1) AND (None = 0) THEN Tip = 0**

**R3: IF (None = 0) AND (Joke = 0) THEN Tip = 1**

**R4: IF (Ad = 0) AND (Joke = 0) THEN Tip = 1**

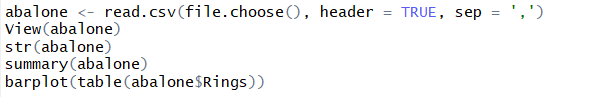
These are couple of rules which can be defined. Again, the values are very randomized, so it is not definite rules.

**Question no:3**

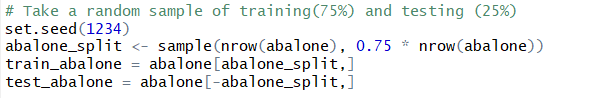
**Random Tree-**

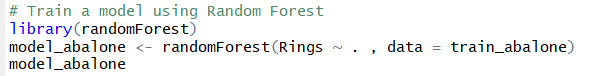
**Steps to be followed:**

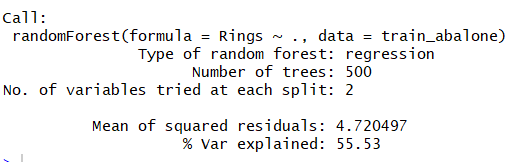
1. Load the datasets into R studio and check for the visualization and statistical summary of the dataset.



1. Split the dataset into training and testing with 3:1 ratio respectively and fit the model with random forest







1. We can see from the above output that there 500 number of tress got generated to train our model
2. Also, we can notice that the type of random forest is ‘regression’ which is obvious because the input dataset are numerical variables.
3. Now that the model is trained with the training data, let us perform the prediction with the 25% testing data



1. Because we have done regression here for this random forest model, it is ideal to use RMSE value to compute how well the random forest has predicted the test outcomes





1. We get the root mean square error of 2.085



1. Here the root mean squared error is about 21% percent of the mean.
2. The RMSE value should be as low as possible. Here like the 1st question, we have not done the scaling. May be, if we preprocess the dataset like removing the insignificant variables and scaling the data, then we might get the lowest RMSE.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*