The Task:

The purpose of this analysis is to predict the survival of passengers of the Titanic. So, to solve this problem, the nature of the relationship between of each variable with Survived must be understood along with the individual characteristic of each variable. A statistical model is then adopted to further the analysis and arrive at the results and interpretation.

The Dataset:

The dataset contains the following variables:

**PassengerId**

**Survived**

**Pclass**

**Name**

**Sex**

**Age**

**SibSp**

**Parch**

**Ticket**

**Income**

**Fare**

**Cabin**

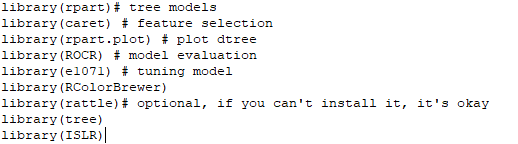
**Embarked**

The statistical model:

Classification is the method of predicting the class of a given input data point. Classification problems are common in machine learning and they fall under the Supervised learning method.

We have adopted the decision tree classification model analysis in this case. When the true goal of our data analysis is to be able to predict which of several non-overlapping groups an observation belongs to, the techniques we use are known as classification techniques. So, to predict the survival of the passengers from the given data with the help of the other variables, the decision tree classification model is best suited to the purpose. In the following pages of documentation, the approach steps have been clearly outlined.

Setting up the R model by loading the required libraries:



Data:

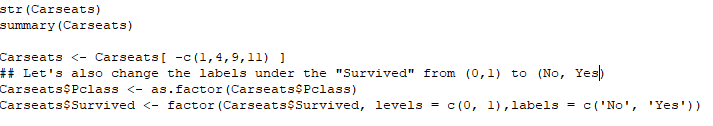
In the next step, the data is read into the R environment from the file.

setwd("C:\\Users\\ADMIN\\Desktop\\R Models\\Decision Tree")

Carseats <- read.csv("Titanic.csv")

Data selection and data type modification:

In the given dataset, the columns which are named as “PassengerId”, “Name”, “Ticket” and “Fare” contains customerid details, names of the passengers, ticket details and fare are of no relevance to our model and is thus excluded. Two of the variables, Pclass and Survived are converted to factors and Survived was in binary form so we converted 0 = “No” and 1 = “Yes”.



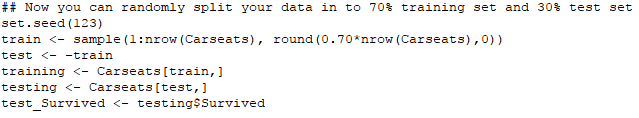
Checking for missing values:

After the data has been cleaned off all the outliers, it is then checked for any missing values in the following manner:

264 missing values were found and has been removed.

Splitting the data:

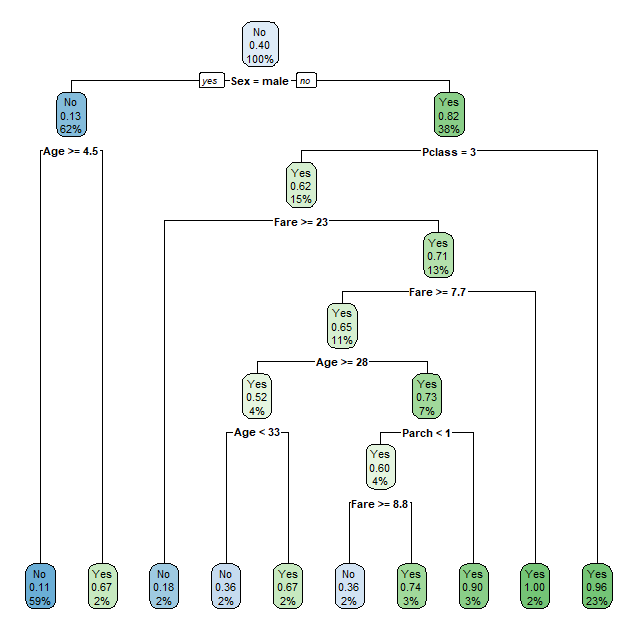
The data is then split into two parts. 70 percent of the data is split into ‘development’ and the remaining 30 percent is named ‘validation’. The development part is for training the decision tree model and the validation part is for testing the robustness of the model.



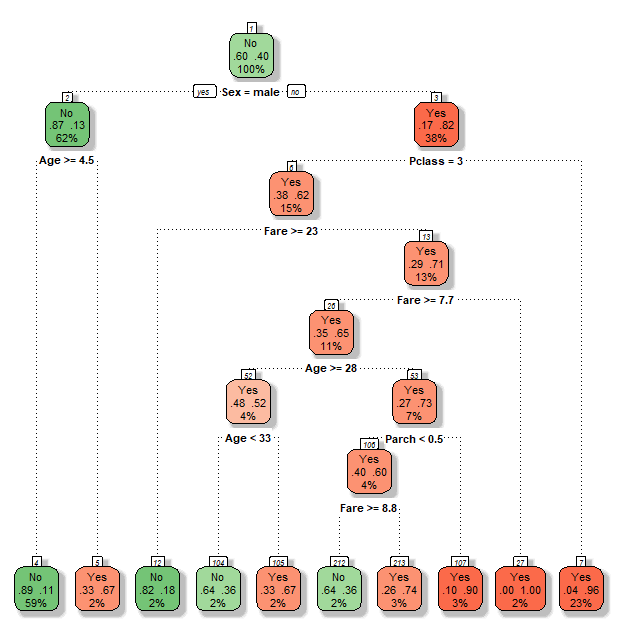
**Running the Decision Tree model:**

Once the data has been divided into two parts, a decision tree has been performed with the Survived as the dependent variable. Here, we are taking the value of cp in such a way that we could get a full grown tree.

C:\Users\ADMIN\Desktop\New folder (2)\dtree model.PNG

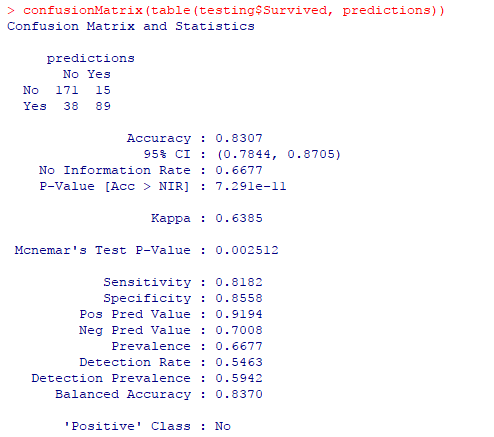


For a better view fancy Rpart plot has been performed where the number of passengers died mentioned in green and those who survived mentioned in red.



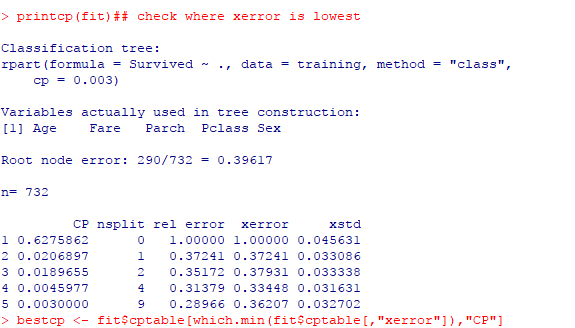
**Validation** of the model:

Checking the confusion matrix to understand how good the model is with a full grown tree.

 C:\Users\ADMIN\Desktop\New folder (2)\mean of full grown tree.PNG

**Pruning the tree for better result:**

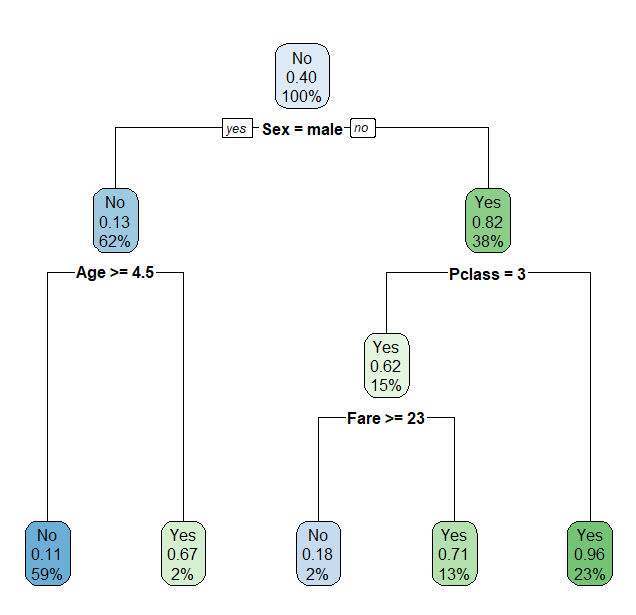
* For pruning the tree, Misclassification Error is used. The Complexity Parameter Table will help in evaluating the fitted decision tree model. The CP table will help in selecting the decision tree that minimizes the misclassification error. CP table lists down all the trees nested within the fitted tree. The best nested sub-tree can then be extracted by selecting the corresponding value for cp.

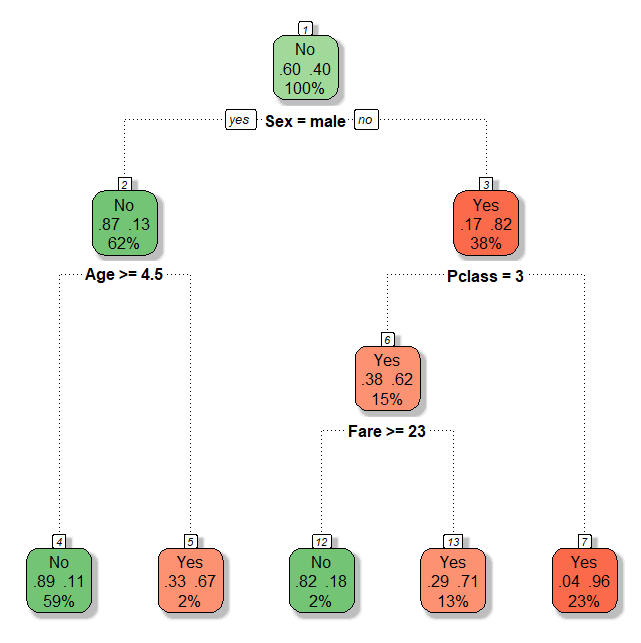


* An optimal tree size is chosen adaptively from the training data. The recommended approach is to build a fully-grown decision tree and then extract a nested sub-tree (prune it) in a way that you are left with a tree that has the minimal node impurities.

pruned <- prune(fit, cp = bestcp)

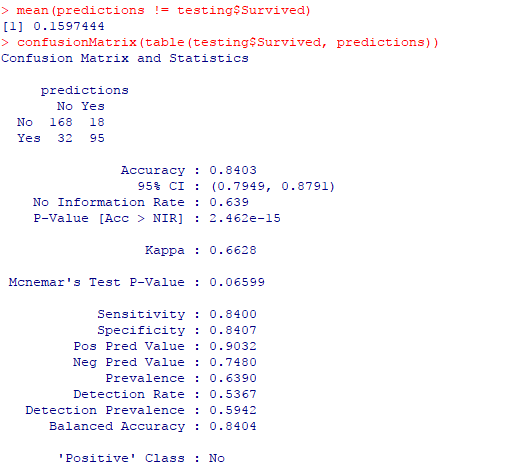
rpart.plot(pruned, extra = 106)





**Validation of the model:**

Checking the confusion matrix to understand how good the model is with a pruned tree.



The model accuracy have increased from 83% to 84%.

Plotting advanced tree for better understanding:

The following image shows that only **2%** of male with **age** of more than 4 years and 5 months and **23%** of female from 3rd **pclass** and **13%** of female who paid **fare** more than 23 had **survived**.

