**FDS ACTIVITY 2**

**TOPIC: WEATHER PREDICTION USING LOGISTIC REGRESSION**

**INTRODUCTION**

Weather conditions around the world change rapidly and continuously. Correct forecasts are essential in today’s daily life. From agriculture to industry, from traveling to daily commuting, we are dependent on weather forecasts heavily. As the entire world is suffering from the continuous climate change and its side effects, it is very important to predict the weather without any error to ensure easy and seamless mobility, as well as safe day to day operations

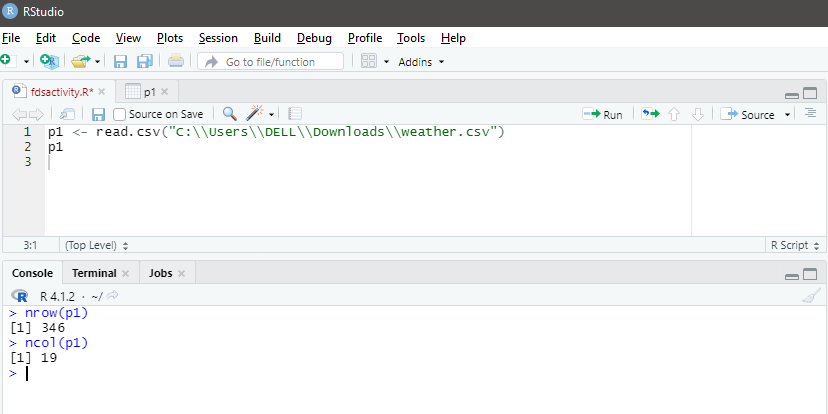
**APPROACHES**

**Logistic regression**

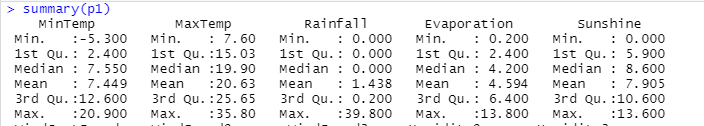
* **Logistic regression** is a process of modeling the probability of a discrete outcome given an input variable. The most common [logistic regression models](https://www.sciencedirect.com/topics/computer-science/logistic-regression-model) a binary outcome; something that can take two values such as true/false, yes/no.
* Logistic regression is a useful analysis method for classification problems, where you are trying to determine if a new sample fits best into a category

1.Collection of dataset:

The dataset was collected from www.kaggle.com , which consists of 346 tuples and 19 columns. The dataset was imported as a csv file into Rstudio

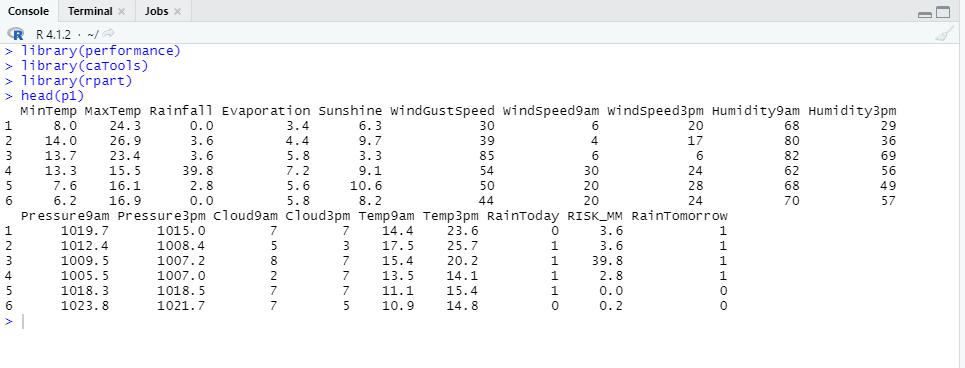


Then we use summary(p1) to determine the statistic for each attribute

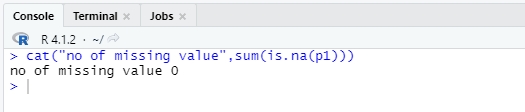


->Then we load the required libraries to execute the code head function will print

First 6 tuples



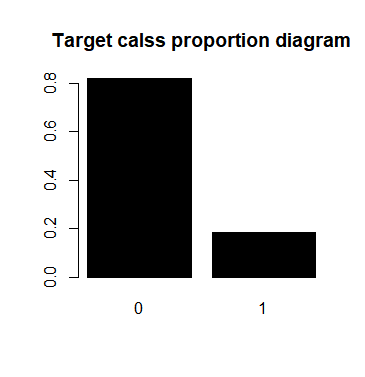
Then we clean the dataset. We check for NAs in the whole dataset



For class distribution

x <- prop.table(table(p1$RainTomorrow))

b <- barplot(x,col="black", main = "Target calss proportion diagram")



Then we split our data for training and testing.We use 70% for training purpose and 30% for the testing purpose.

We use the following code

split <- sample.split(p1, SplitRatio = 0.7)

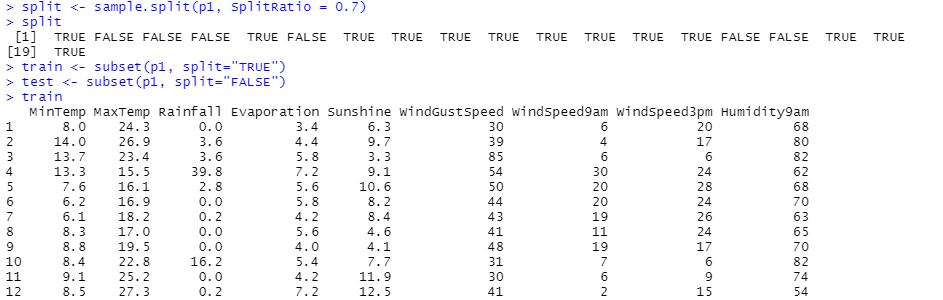
split

train <- subset(p1, split="TRUE")

test <- subset(p1, split="FALSE")

train

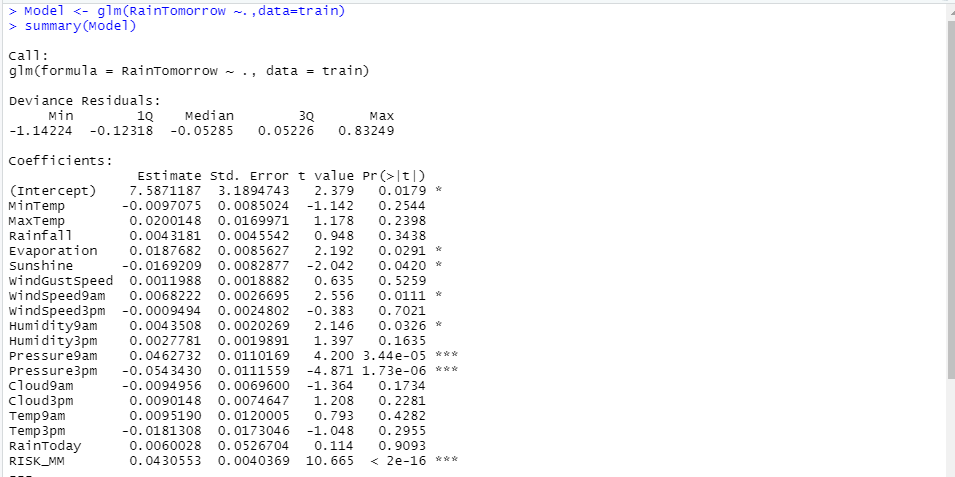
test



Then we train our model using logistic regression

Model <- glm(RainTomorrow ~.,data=train)

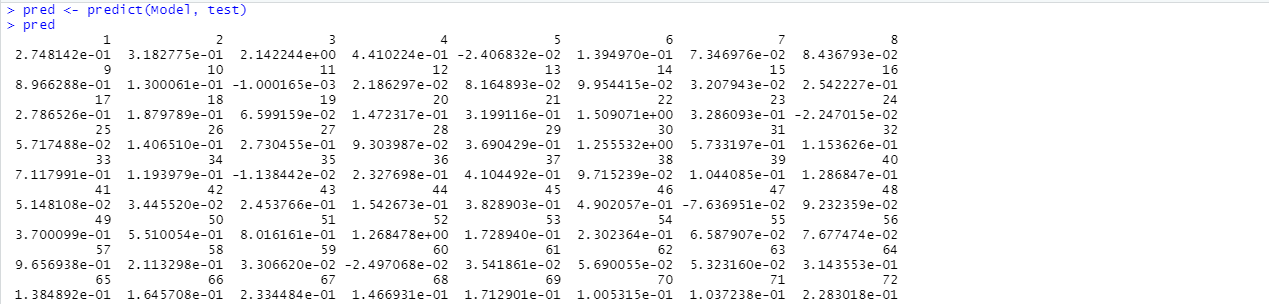
summary(Model)



Now we predict the values for the test data(30% data)

pred <- predict(Model, test)

pred



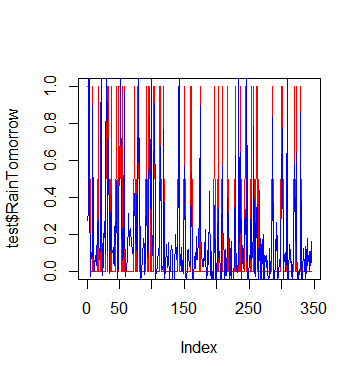
Comparison Graph ( Test vs Train)

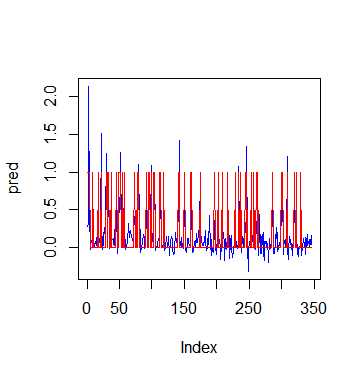
plot(test$RainTomorrow,type = "l",lty=1.8,col="red")

lines(pred,type = "l",col="blue")

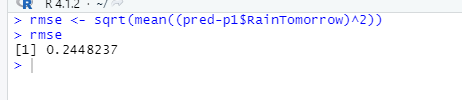
plot(pred,type = "l",lty=1.8,col="blue")

lines(test$RainTomorrow,type = "l",lty=1.8,col="red")





Now we have to calculate the RMSE value. RMSE is is the square root of the mean of the square of all of the error. RMSE is considered an excellent general-purpose error metric for numerical predictions.

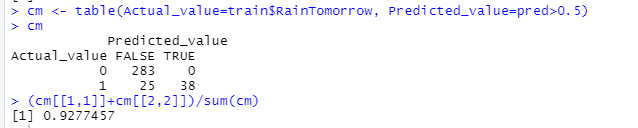


We found the confusion matrix using

cm <- table(Actual\_value=train$RainTomorrow, Predicted\_value=pred>0.5)

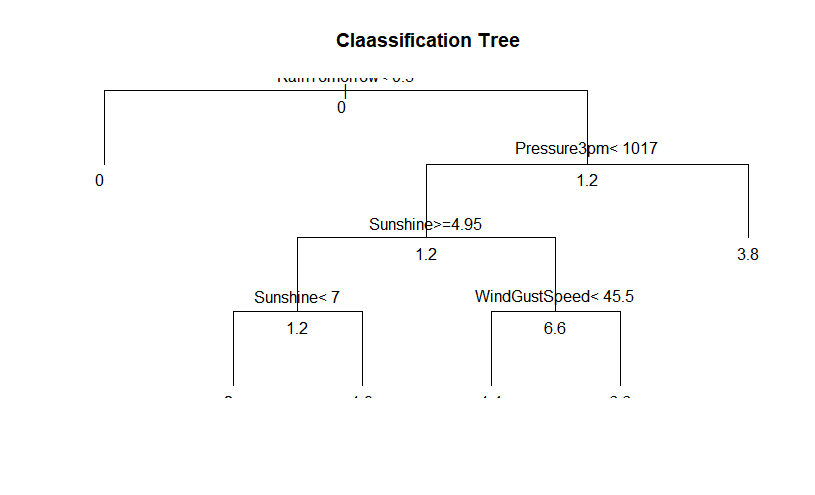
cm

(cm[[1,1]]+cm[[2,2]])/sum(cm)



Accuracy of our model is – 92%

we create the decision tree as to what attributes accurately determine the target value.



**Conclusion**

In this Activity we analyzed dataset by knowing attributes in data, mean, median and class distribution of the dataset. We identified Outcome as the target variable. We also made data visualization. And implemented two algorithm for prediction**.**