Anshul Verma

EM-623

Final Project

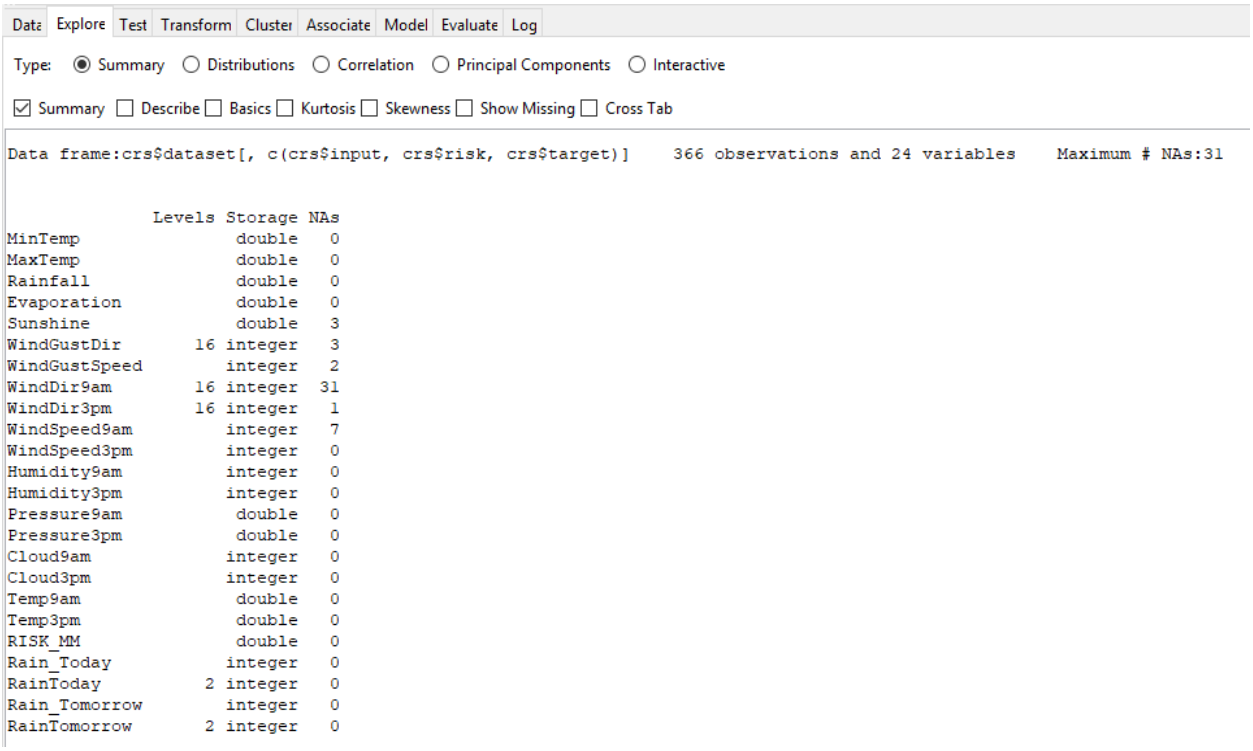
Decision Tree for Determining Rainfall Next Day

Phase 1 – Business Understanding:

The purpose of this study is to analyze the dataset 'weather.csv' and build a model to determine whether it will rain the next day. This analysis will be done using R (Library: Rattle) and Microsoft Excel. The model would be decision tree based on the results of the analysis.

Phase 2 – Data Understanding:

First, we load the dataset into Rattle. By executing the 'Summary' in the 'Explore' tab we find that the dataset has 366 observations and 24 variables. The variables mostly consist of weather conditions on given dates of specific years. According to the summary, there are 38 observations with missing values. Also, there is a maximum of 31 NAs in the dataset.



	Levels	Storage	NAs
MinTemp		double	0
MaxTemp		double	0
Rainfall		double	0
Evaporation		double	0
Sunshine		double	3
WindGustDir	16	integer	3
WindGustSpeed		integer	2
WindDir9am	16	integer	31
WindDir3pm	16	integer	1
WindSpeed9am		integer	7
WindSpeed3pm		integer	0
Humidity9am		integer	0
Humidity3pm		integer	0
Pressure9am		double	0
Pressure3pm		double	0
Cloud9am		integer	0
Cloud3pm		integer	0
Temp9am		double	0
Temp3pm		double	0
RISK_MM		double	0
Rain_Today		integer	0
RainToday	2	integer	0
Rain_Tomorrow		integer	0
RainTomorrow	2	integer	0

Therefore, the dataset would require some cleaning before we can begin the analysis.

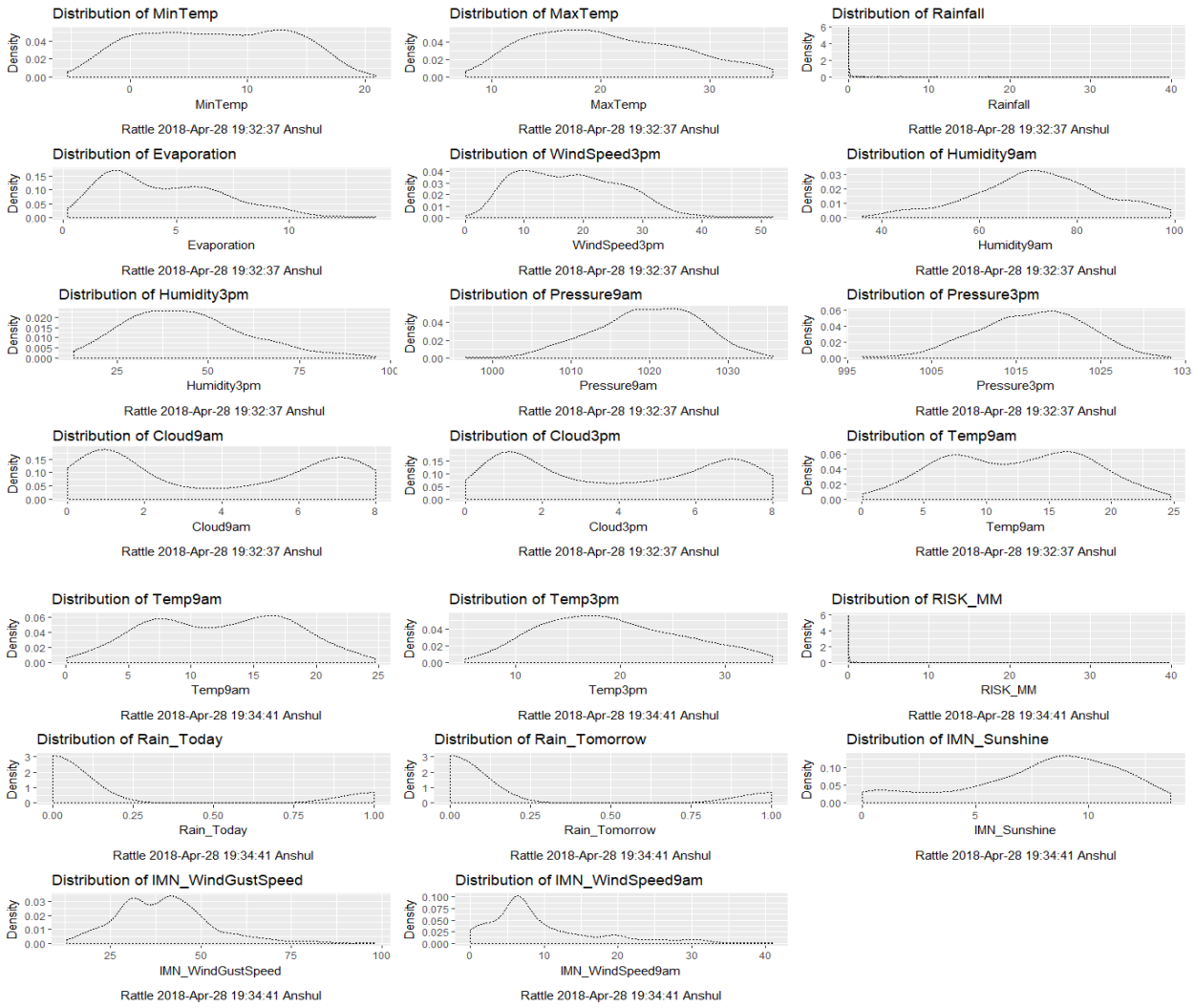
Phase 3 – Data Preparation:

First, we convert the binary categorical variables 'RainToday' and 'RainTomorrow' into binary numeric variables using Excel. For this, we use the 'if' statement and set the condition as: **If RainTomorrow= Yes, then 1, otherwise 0**. We then remove the original variables.

Next, we remove some categorical variables, which also have many missing values, such as – WindGustDir, WindDir9am, WindDir3pm, Date. This is done by setting the variable type to 'Ignore' under the 'Data' tab, and using 'Cleanup' option under the 'Transform' tab.

Next, we fill the missing values of variables 'Sunshine', 'WindGustSpeed', and 'WindSpeed9am' with their mean. This is done by using 'Impute (Mean)' option under the 'Transform' tab.

Next, we study the distributions of the variables to determine outliers, and also to see how data is distributed. For this, we generate histograms of the variables, by selecting the 'Distributions' option under the 'Explore' tab. The histograms are shown below:

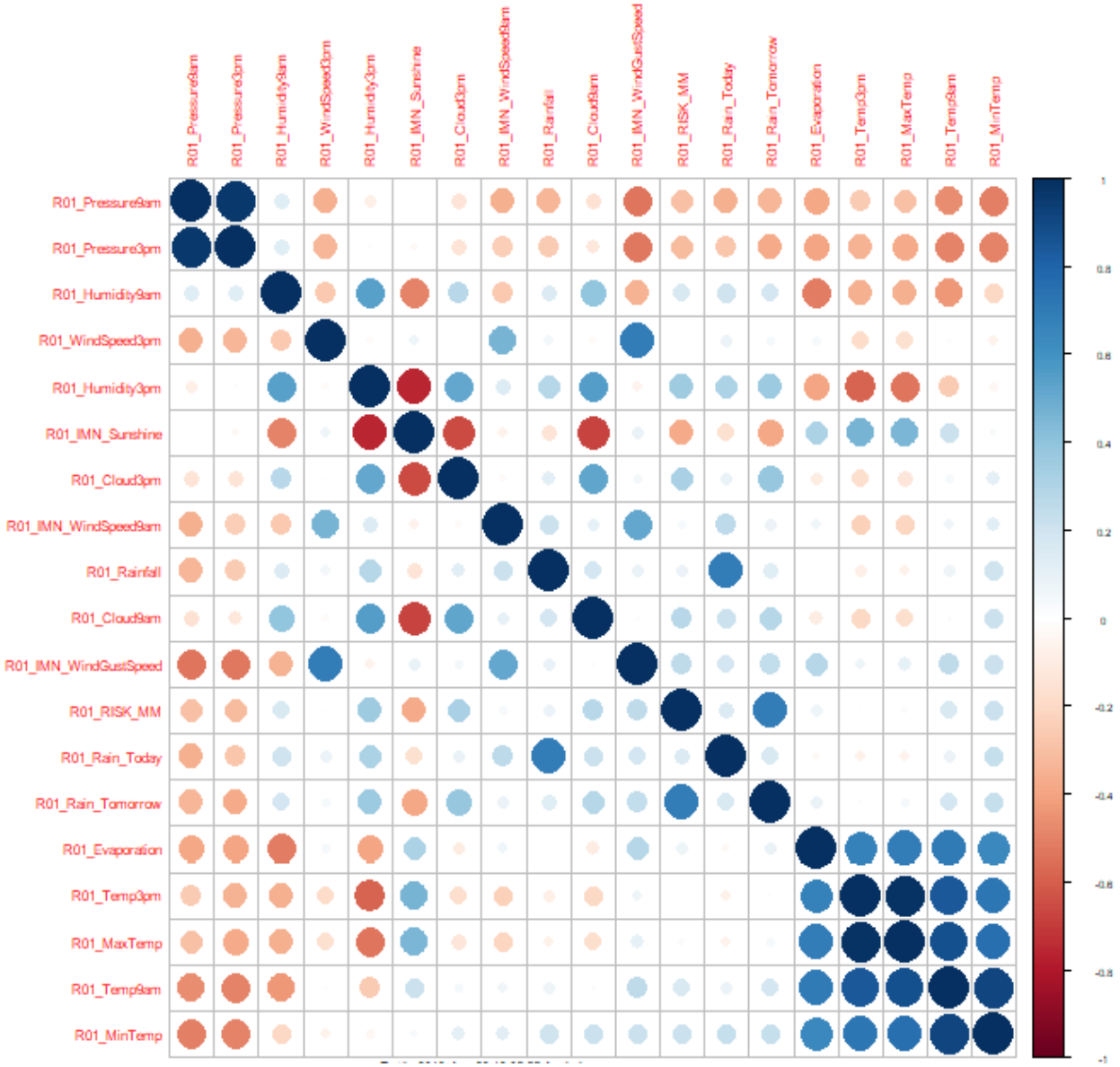


We see that there are no outliers in the variables.

Also, we see that the variables need to be normalized and rescaled.

To normalize the variables, we use the Min-Max Transformation. This is done by selecting the 'Rescale (Scale [0-1])' option under the 'Transform' tab. We then remove the original variables and keep the normalized variables.

Next, we analyze the correlation between the variables, to determine if any two variables have strong correlation with each other or with the target variable 'RainTomorrow'. This is done by selecting the 'Correlation' option under the 'Explore' tab. We get the following results:

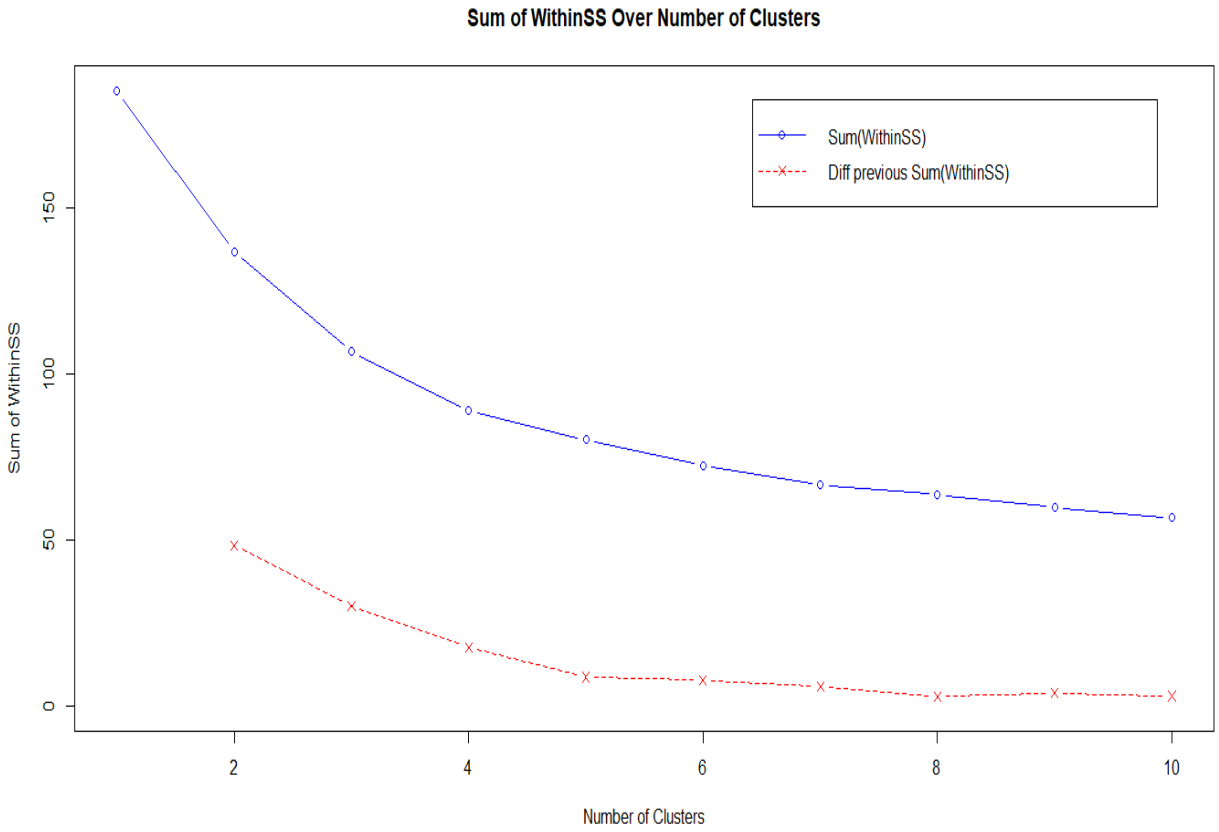


We see that there is strong correlation between variables ‘Pressure9am’ and ‘Pressure3pm’. So, we remove variable ‘Pressure3pm’. Also, we see the variables ‘MaxTemp’ and ‘Temp3pm’ have strong correlation. So, we remove the variable ‘Temp3pm’. Then, we see that variables ‘MinTemp’ and ‘Temp9am’ have strong correlation. So, we remove variable ‘Temp9am’. Then we see that variable ‘WindSpeed3pm’ and ‘WindGustSpeed’ have strong correlation. So, we remove variable ‘WindSpeed3pm’. Then we see that variable ‘Humidity3pm’ and ‘Sunshine’ have strong correlation. So, we remove variable ‘Humidity3pm’. Also, we see that variable ‘Rain_Today’ and “Rainfall’ have strong correlation. So, we remove the variable ‘Rainfall’.

We also see that the variable ‘RISK_MM’ has strong correlation with the target variable ‘RainTomorrow’. So, we remove the variable ‘RISK_MM’.

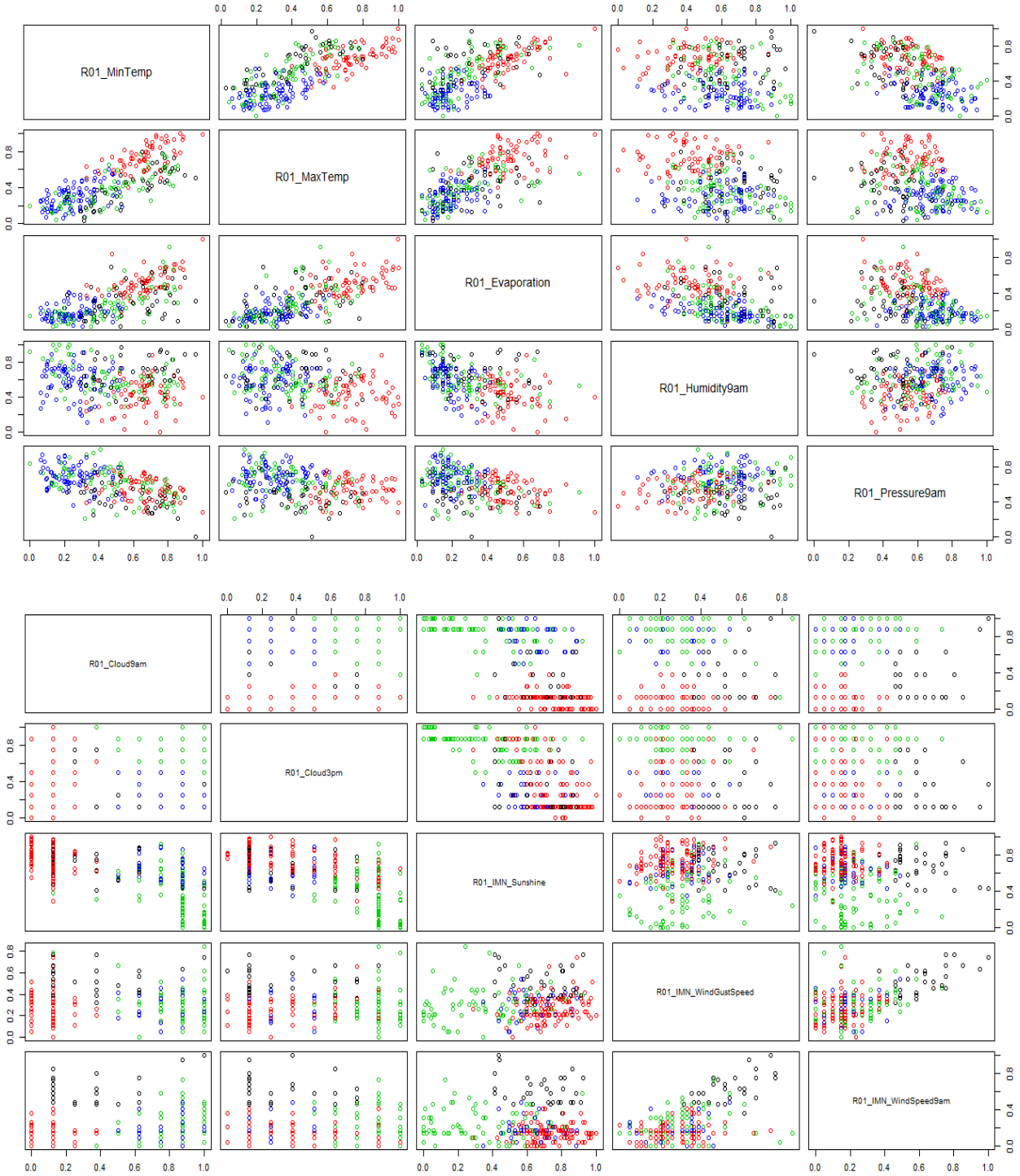
This leaves us with 12 variables.

Next, we try to find out clusters in the dataset. For this, we perform K-Means Clustering on the dataset. We use ‘Elbow Method’ for determining the optimum number of clusters. We get the following graph:



From the graph, we see that the optimum number of clusters is 4.

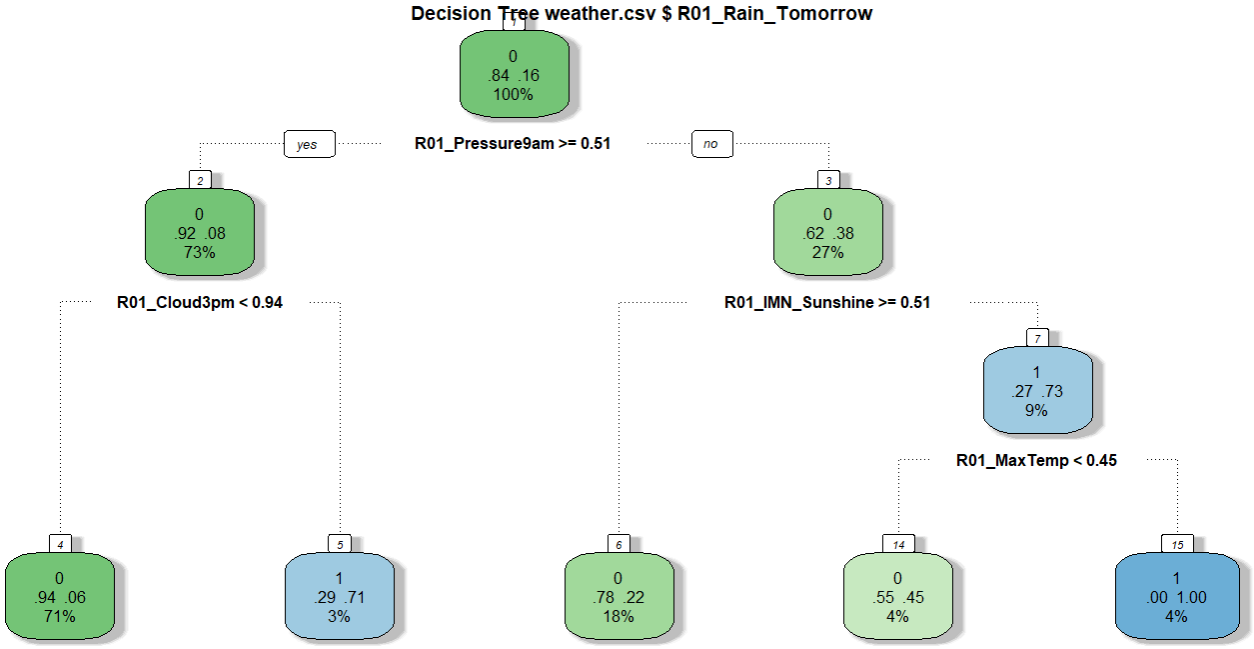
Then, we perform clustering by selecting ‘KMeans (clusters=4)’ option under the ‘Cluster’ tab. We get the following results:



Phase 4 – Modelling:

For this phase, we create a partition in the dataset for training and testing the model. We set partition as 70 (training) / 30 (testing). Then, we set the variable 'RainTomorrow' as target variable.

We generate a Decision Tree for the data. This is done by selecting 'Tree' option under the 'Model' tab. We set the model parameters as: Min Split = 20, Min Bucket = 7, Max Depth = 3, Complexity = 0.0200. We get the following model:



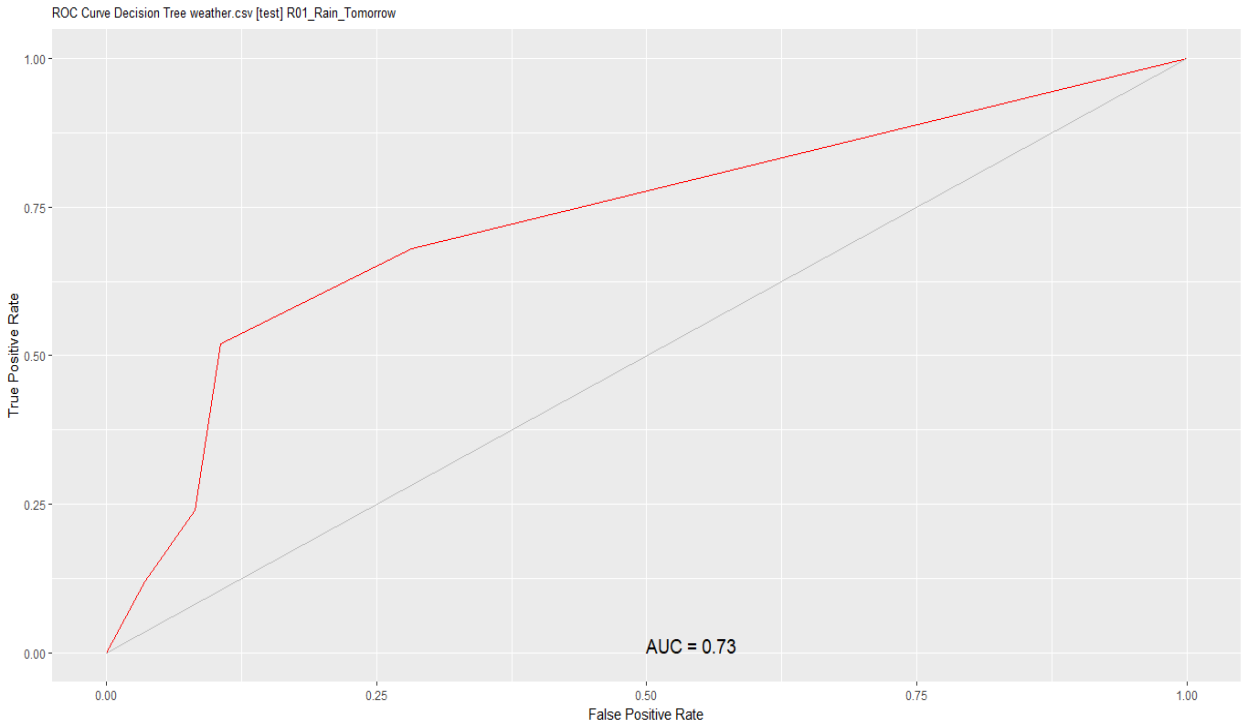
Phase 5 - Evaluation:

To evaluate the model, we generate the Error Matrix. This is done by selecting 'Error Matrix' option under the 'Evaluate' tab. We get the following results:

```
Error matrix for the Decision Tree model on weather.csv [test] (counts):

       Predicted
Actual   0 1 Error
     0 78 7   8.2
     1 19 6  76.0

Error matrix for the Decision Tree model on weather.csv [test] (proportions):

       Predicted
Actual    0   1 Error
     0 70.9 6.4   8.2
     1 17.3 5.5  76.0

Overall error: 23.6%, Averaged class error: 42.1%
```

We see that Overall Error = 23.6%. This means that the model can be improved further.

We also plot the ROC curve for the model. This is done by selecting 'ROC' option under the 'Evaluate' tab. We get the following plot:



We see that the Area Under Curve (AUC) for the plot = 0.73. This means that the model is good, but it can be further improved.

Phase 6 – Deployment:

We can use this model for determining if it will rain the next day. The model gives good results. It can still be improved further.

Conclusion:

We have created a Decision tree model for determining if it will rain the next day. From the model we see that there is a high chance of rain the next day, if air pressure at 9am is high, and clouds at 3pm are more. The model has good results, but we can improve it further.