Assessment Task 3: Data mining in action

32130 Fundamentals of Data Analytics
Assessment 3

Vishal Poojary

Student ID: 13739863

1.Introduction

The purpose of the report is to illustrate the accuracy of classifiers which is used to classify the “RainTommorrow”. Two data set is used for this purpose “Assignment3-TrainingData.csv” and “Assignment3-UnknownData.csv”. In order to get the dataset ready pre-processing and cleaning have been done before classification. The process of classification have been tested using various method and the method which gave the highest accuracy have been selected for classification of “Rain Tommorrow” in the “Assignment3-UnknownData.csv”. Knime and Excel have been used for all the process.

2.Data Mining Problem

There are two datasets and both gives us the datapoints of weather conditions location (city) wise. The first data set which is the training dataset includes the parameter of “Rain Tomorrow” where 1 represents the rain will happen tomorrow and 0 represents that the rain won’t happen. The other data file is unknown data where the parameter of “Rain Tomorrow” is absent. I will need to build a classifier by analyzing the training dataset and predict the same parameter in the unknow dataset which is prediction of whether there will be rain the next day or not.

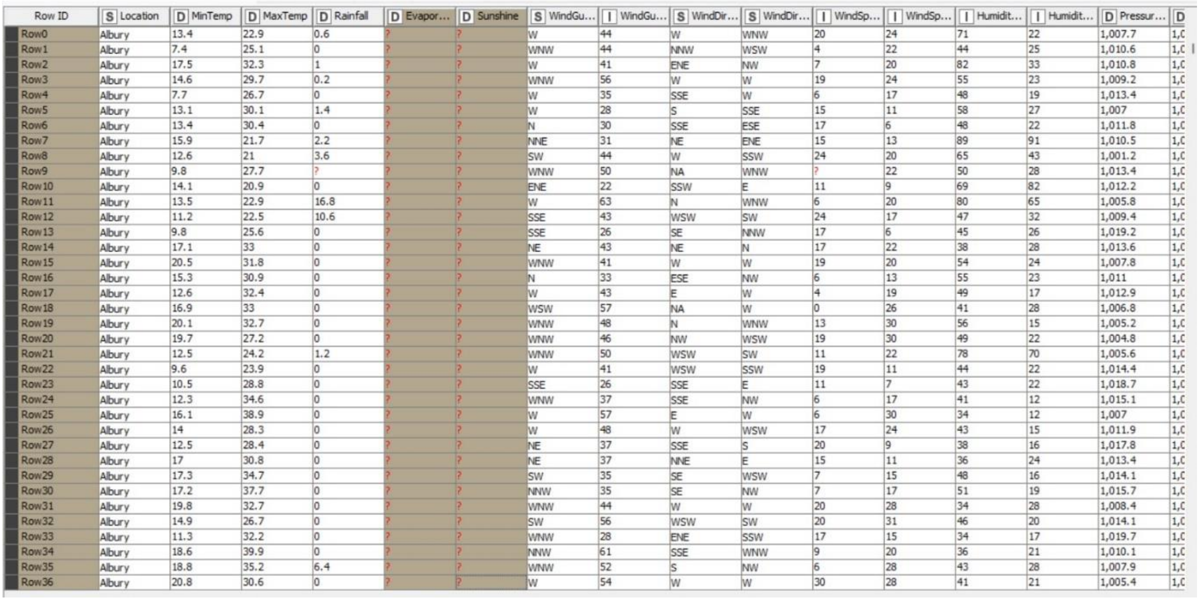
The Training data has 99516 rows where the unknown data has 42677 rows. The datasets have unwanted data, missing value and attributes which needs to be normalized.

2.Data pre-processing and transformation

Data pre-processing have been conducted to execute the model in an efficient way i.e to increase the accuracy of the classifiers.

The observations and the pre-processing method used are as follows:

1. Filtered column
   Evaporation and Sunshine column were excluded due data being missing for both columns. (Refer image 1)



Row ID	S Location	D MinTemp	D MaxTemp	D Rainfall	D Evapor...	D Sunshine	S WindGu...	I WindGu...	S WindDir...	S WindDir...	I WindSp...	I WindSp...	I Humidit...	I Humidit...	D Pressur...	D
Row0	Albury	13.4	22.9	0.6	?	?	W	44	W	WNW	20	24	71	22	1,007.7	1,0
Row1	Albury	7.4	25.1	0	?	?	WNW	44	NNW	WSW	4	22	44	25	1,010.6	1,0
Row2	Albury	17.5	32.3	1	?	?	W	41	ENE	NW	7	20	82	33	1,010.8	1,0
Row3	Albury	14.6	29.7	0.2	?	?	WNW	56	W	W	19	24	55	23	1,009.2	1,0
Row4	Albury	7.7	26.7	0	?	?	W	35	SSE	W	6	17	48	19	1,013.4	1,0
Row5	Albury	13.1	30.1	1.4	?	?	W	28	S	SSE	15	11	58	27	1,007	1,0
Row6	Albury	13.4	30.4	0	?	?	N	30	SSE	ESE	17	6	48	22	1,011.8	1,0
Row7	Albury	15.9	21.7	2.2	?	?	NNE	31	NE	ENE	15	13	89	91	1,010.5	1,0
Row8	Albury	12.6	21	3.6	?	?	SW	44	W	SSW	24	20	65	43	1,001.2	1,0
Row9	Albury	9.8	27.7	?	?	?	WNW	50	NA	WNW	?	22	50	28	1,013.4	1,0
Row10	Albury	14.1	20.9	0	?	?	ENE	22	SSW	E	11	9	69	82	1,012.2	1,0
Row11	Albury	13.5	22.9	16.8	?	?	W	63	N	WNW	6	20	80	65	1,005.8	1,0
Row12	Albury	11.2	22.5	10.6	?	?	SSE	43	WSW	SW	24	17	47	32	1,009.4	1,0
Row13	Albury	9.8	25.6	0	?	?	SSE	26	SE	NNW	17	6	45	26	1,019.2	1,0
Row14	Albury	17.1	33	0	?	?	NE	43	NE	N	17	22	38	28	1,013.6	1,0
Row15	Albury	20.5	31.8	0	?	?	WNW	41	W	W	19	20	54	24	1,007.8	1,0
Row16	Albury	15.3	30.9	0	?	?	N	33	ESE	NW	6	13	55	23	1,011	1,0
Row17	Albury	12.6	32.4	0	?	?	W	43	E	W	4	19	49	17	1,012.9	1,0
Row18	Albury	16.9	33	0	?	?	WSW	57	NA	W	0	26	41	28	1,006.8	1,0
Row19	Albury	20.1	32.7	0	?	?	WNW	48	N	WNW	13	30	56	15	1,005.2	1,0
Row20	Albury	19.7	27.2	0	?	?	WNW	46	NW	WSW	19	30	49	22	1,004.8	1,0
Row21	Albury	12.5	24.2	1.2	?	?	WNW	50	WSW	SW	11	22	78	70	1,005.6	1,0
Row22	Albury	9.6	23.9	0	?	?	W	41	WSW	SSW	19	11	44	22	1,014.4	1,0
Row23	Albury	10.5	28.8	0	?	?	SSE	26	SSE	E	11	7	43	22	1,018.7	1,0
Row24	Albury	12.3	34.6	0	?	?	WNW	37	SSE	NW	6	17	41	12	1,015.1	1,0
Row25	Albury	16.1	38.9	0	?	?	W	57	E	W	6	30	34	12	1,007	1,0
Row26	Albury	14	28.3	0	?	?	W	48	W	WSW	17	24	43	15	1,011.9	1,0
Row27	Albury	12.5	28.4	0	?	?	NE	37	SSE	S	20	9	38	16	1,017.8	1,0
Row28	Albury	17	30.8	0	?	?	NE	37	NNE	E	15	11	36	24	1,013.4	1,0
Row29	Albury	17.3	34.7	0	?	?	SW	35	SE	WSW	7	15	48	16	1,014.1	1,0
Row30	Albury	17.2	37.7	0	?	?	NNW	35	SE	NW	7	17	51	19	1,015.7	1,0
Row31	Albury	19.8	32.7	0	?	?	WNW	44	W	W	20	28	34	28	1,008.4	1,0
Row32	Albury	14.9	26.7	0	?	?	SW	56	WSW	SW	20	31	46	20	1,014.1	1,0
Row33	Albury	11.3	32.2	0	?	?	WNW	28	ENE	SSW	17	15	34	17	1,019.7	1,0
Row34	Albury	18.6	39.9	0	?	?	NNW	61	SSE	WNW	9	20	36	21	1,010.1	1,0
Row35	Albury	18.8	35.2	6.4	?	?	WNW	52	S	NW	6	28	43	28	1,007.9	1,0
Row36	Albury	20.8	30.6	0	?	?	W	54	W	W	30	28	41	21	1,005.4	1,0

Image 1

2. Number to String
   "RainTomorrow" attribute was converted from whole number to string using Number to String node in order to predict through classifiers. Note: Classifiers do not accept target attribute to be in whole number.

3. Missing Values
   "Wind Gus Speed", "Wind Speed 9am", "Cloud 9am", "Cloud 3pm" all have missing values hence missing value node was used and all the missing values were replaced by the mean of the values of their specific column.

4. Normalization
   In "Rainfall", "Humidity 9am" and "Humidity 3pm" columns the data were unbalanced in which lowest being 0 and highest being 371. Please refer image 2 & 3

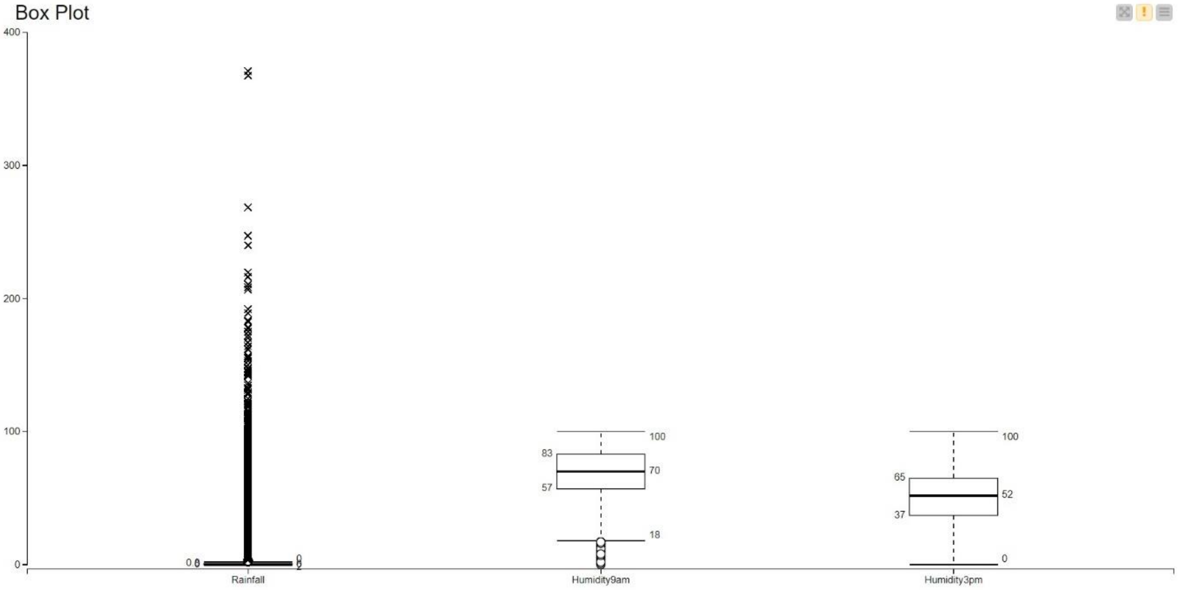


Image 2

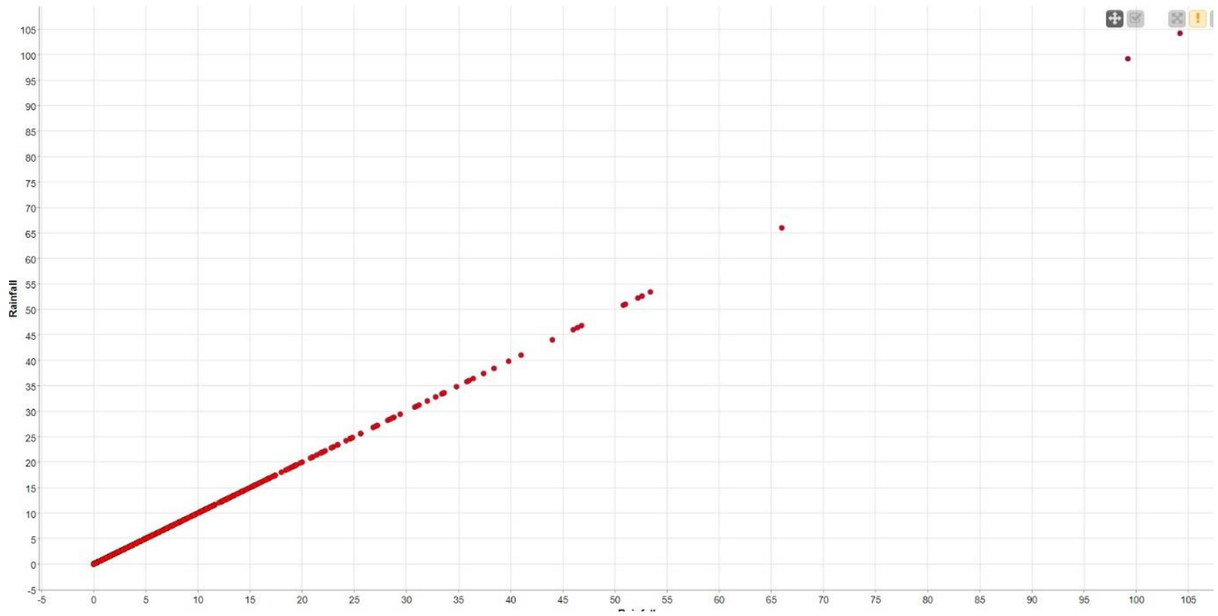


Image 3

As you can see from image 2 and 3 that the distribution is highly uneven and therefore have been normalized using min-max normalization.

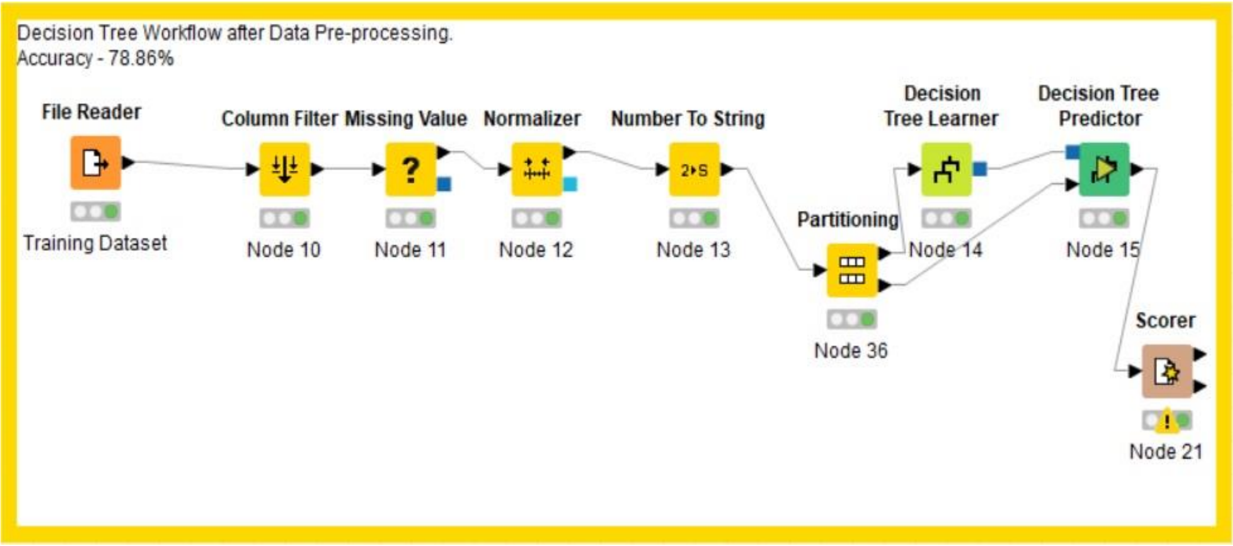
3.Problem solving process

After following the above-mentioned pre-processing steps. Different model was developed. Following classifiers were used – Decision Tree, K-Nearest Neighbour, Random Forest, Tree Ensemble.

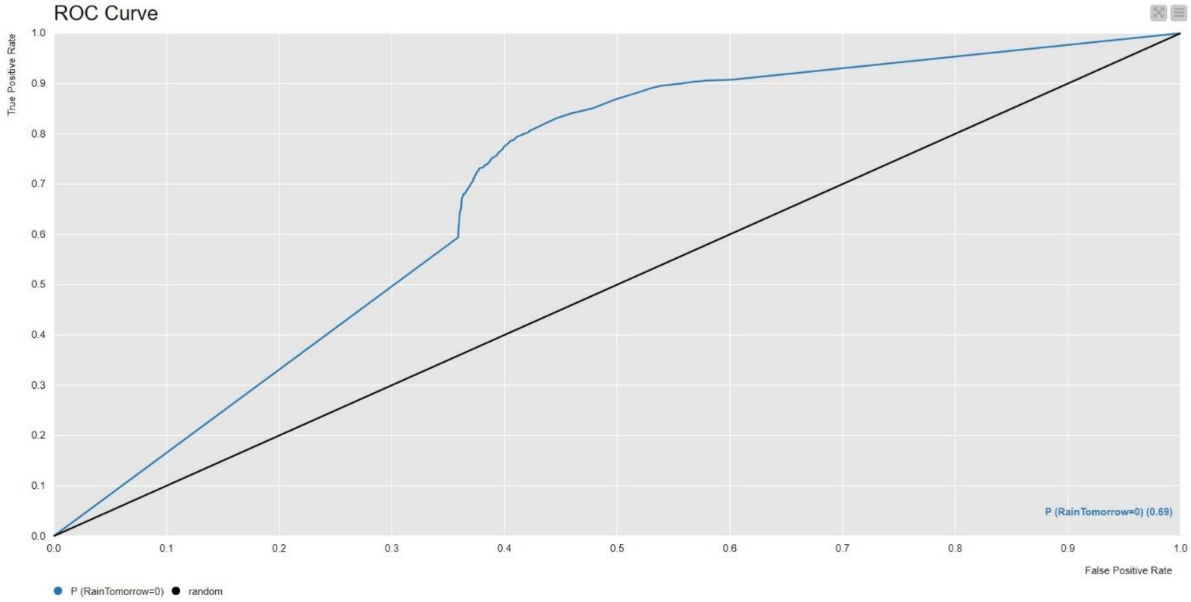
For all the classifiers data processing was done using Column filter, Missing value, Normalizer, Number to String and Partitioning nodes. Partitioning was done 70-30 – 70% data was used as training data and 30% as test data – so the classifiers learn from the 70% and predicts the “Rain Tomorrow” class in 30% of the data.

4.Classification techniques used

1) Decision Tree
   Default settings were used for this classifier except “Gain ratio” was used instead of “Gini index” as a quality measure in the decision tree learner as that gave more accuracy when the classification was attached to the scorer to check the accuracy.

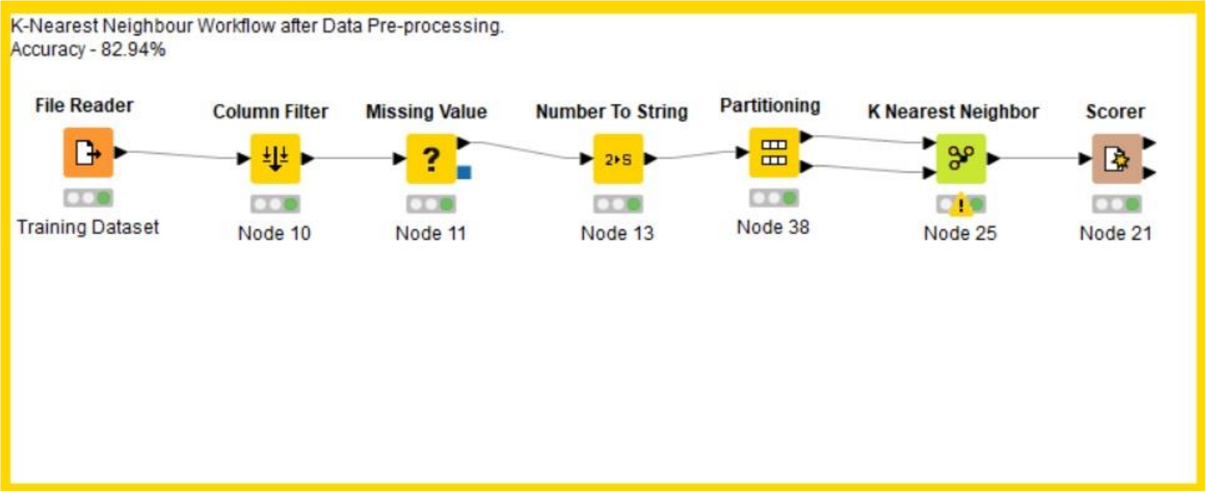


Decision Tree Workflow after Data Pre-processing.
Accuracy - 78.86%

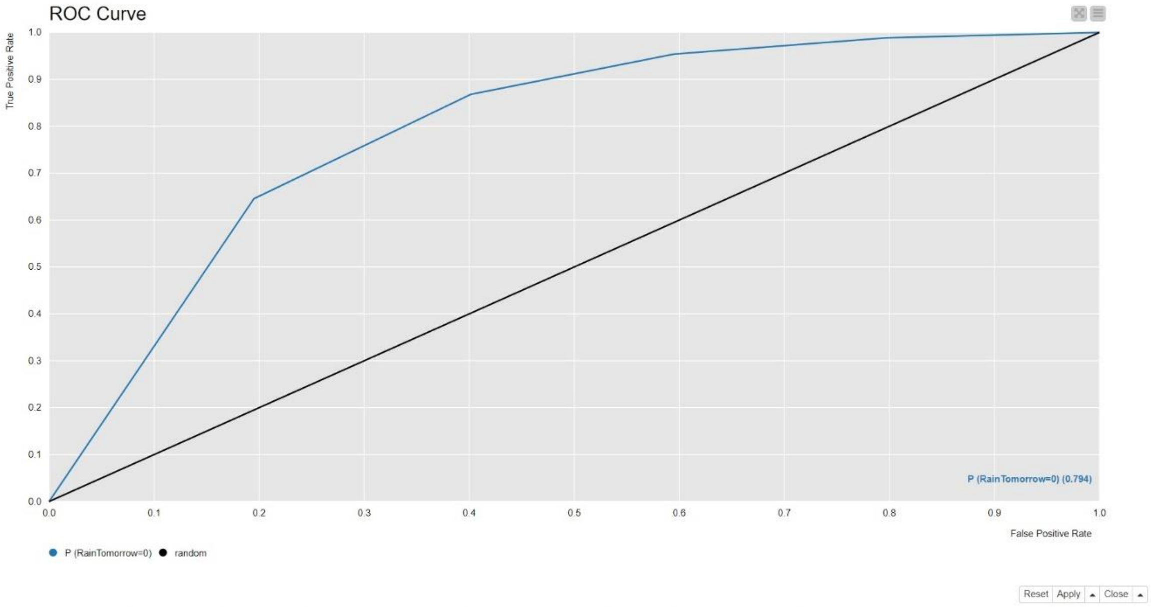


2. K – Nearest Neighbour

I changed the value of K in the K-nearest neighour node to 4 as all other attributes were providing me with less accuracy when connected to scorer node.

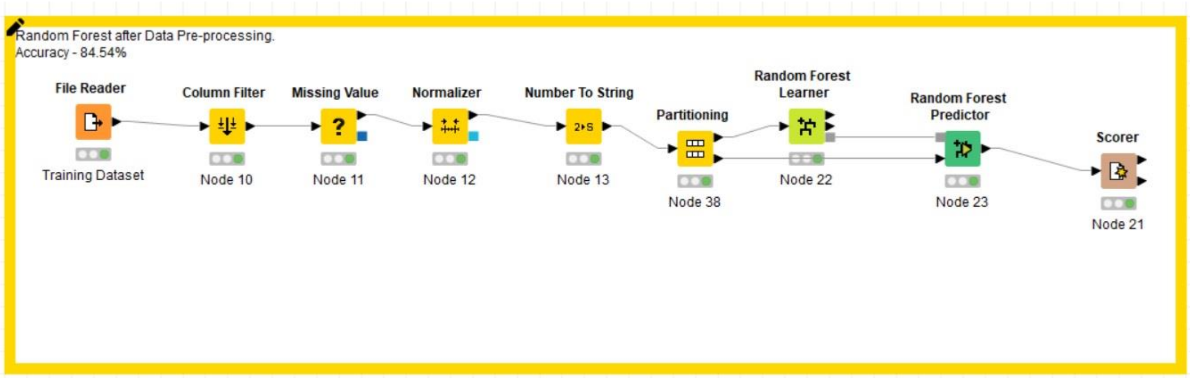


K-Nearest Neighbour Workflow after Data Pre-processing.
Accuracy - 82.94%

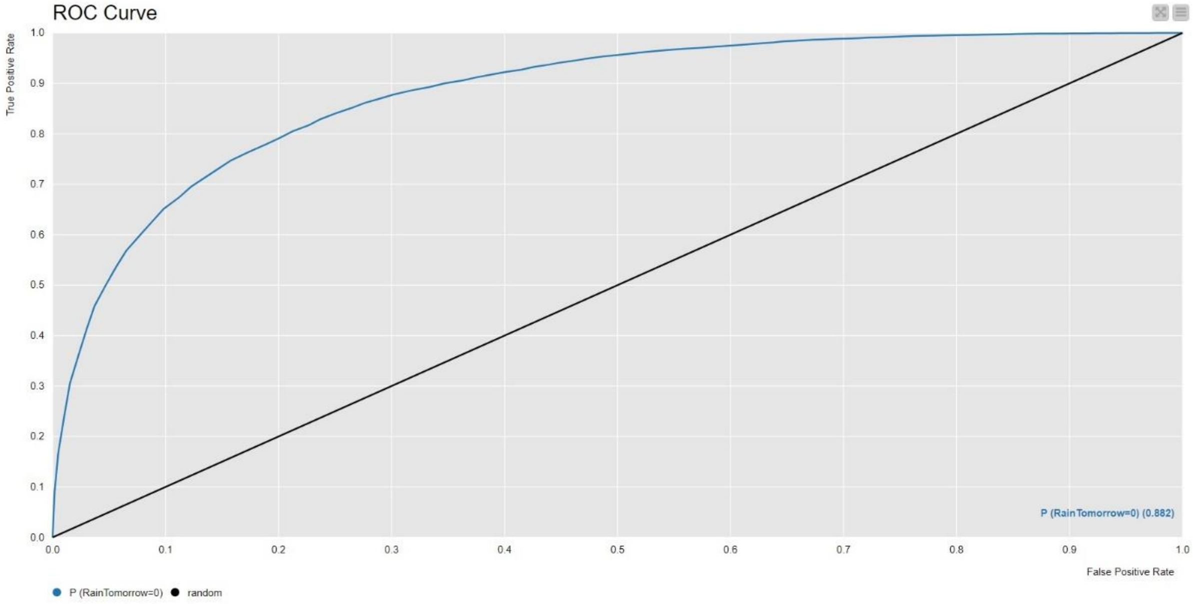


3. Random Forest

Default settings were used for Random Forest classifiers along with pre-processing.

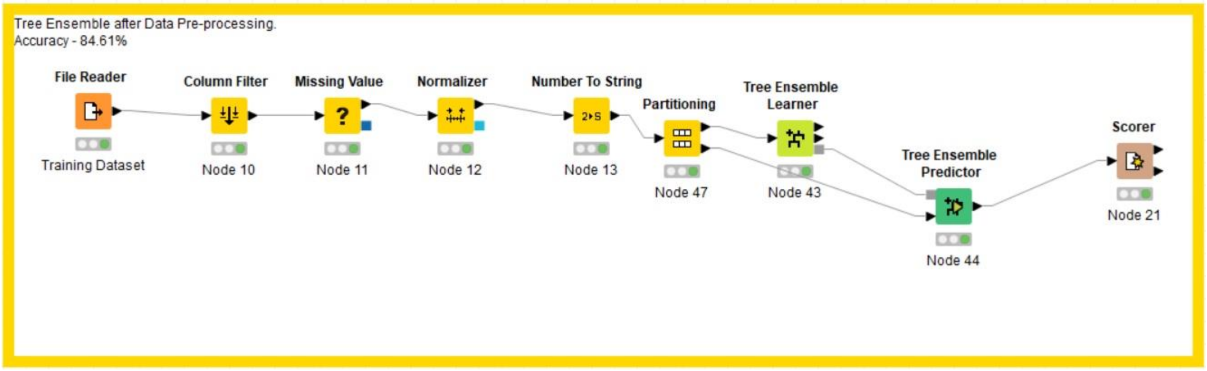


Random Forest after Data Pre-processing.
Accuracy - 84.54%

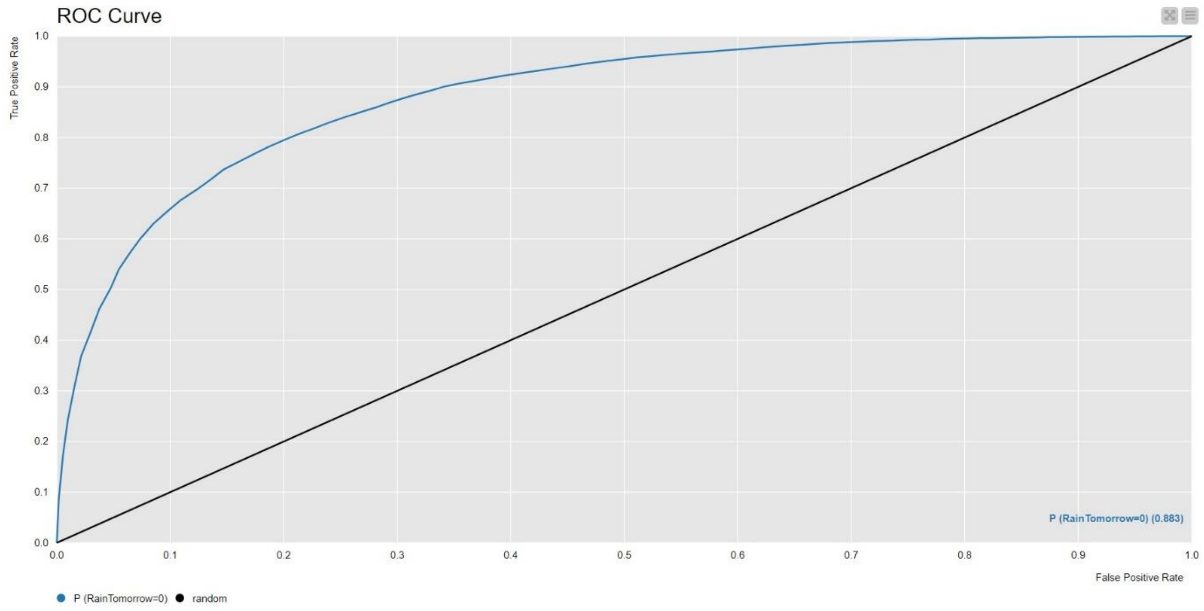


4.Tree Ensemble

Default settings were used for Random Forest classifiers along with pre processing



Tree Ensemble after Data Pre-processing.
Accuracy - 84.61%



5.Best Classifier

From the above analysis, it can be concluded that Tree Ensemble classifier gave us the highest accuracy of 84.61% compared to other classifier used. Second best is the Random Forest classifier with the accuracy of 84.54%. Both the mentioned classifier can be used to predict the target class on the real data.

6.Reflection

Through this assignment I got a thorough understanding of how data pre-processing, classification is done through Knime. Mainly I learnt about what predictive analytics is. How the AI algorithm in different classifiers learn from the training data and based on that learning it predicts in the test data. We predicted if the rain is going to happen in future based on past data. I can apply the same logic in my current work environment with customer data. Can predict if the customer is likely to churn soon. I also understood that it is very important to clean the data before pre-processing as many says garbage in is garbage out. All the time spent cleaning and pre-processing the data is well worth as it affects the final output.