STEVENS
INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®

CS 513 A
Knowledge Discovery and Data Mining

Rainfall in Australia

Group Members

Chetan Goyal
20005334

Muzaffar Turak
10476985

Vishal Mandala
10474183

Contents

1. Dataset Description
2. Objective
3. EDA - Exploratory Data Analysis
4. Classification algorithms used
   - a. KNN Methodology
   - b. Adaboost
   - c. Gaussian Naive Bayes
   - d. Artificial Neural Network
   - e. Decision Trees (CART)
   - f. Extra Trees
   - g. Random Forest
   - h. SVM Methodology
5. Conclusion

Dataset description

Rain in Australia dataset-
https://www.kaggle.com/datasets/jsphyg/weather-dataset-rattle-package

- Raw data -
  - Number of columns - 23
  - Number of rows - 145460
  - Categorical rows - 7
  - Numerical rows - 16
  - Target Feature - ‘RainTomorrow’

- Data after EDA -
  - Number of columns - 22
  - Number of rows - 142193
  - Categorical rows - 7
  - Numerical rows - 15
  - Column dropped - ‘Temp3pm’

- Data distribution of test and training -
  - 30% for test
  - 70% for training

Numerical columns -

- MinTemp
- MaxTemp
- RainFall
- Evaporation
- Sunshine
- WindGustSpeed
- WindSpeed9am
- WindSpeed3pm
- Humidity9am
- Humidity3pm
- Pressure9am
- Pressure3pm
- Cloud9am
- Cloud3pm
- Temp9am
- Temp3pm

Categorical columns -

- Date
- Location
- WindGustDir
- WindDir9am
- WindDir3pm
- RainToday

Target Column -

- RainTomorrow

Objective

Problem Statement -

This dataset contains about 10 years of daily weather observations from many locations across Australia.

RainTomorrow is the target variable to predict. It means -- did it rain the next day, Yes or No? This column is Yes if the rain for that day was 1mm or more.

Objective -

Predict next-day rain by training classification models on the target variable RainTomorrow.

To achieve this by first performing EDA (Exploratory Data Analysis) which results into data preprocessing and feature engineering. Then to prepare few ML models to help us make the prediction if it will rain next day - or not.

Exploratory Data Analysis

Data preprocessing

1. Removed all the rows that contained null values in our target variable Rain tomorrow. This resulted in a reduction of 255 rows. Compared to over 14000 rows in total that we have, this is an acceptable reduction

2. Checked for any column that contained over 50% null values in their cells. None such column existed so no column was removed here. All columns had less than 50% null values

3. Checked for cardinality in the columns-
   a. First discovered how many categorical columns exist - 7 such columns
   b. Of all the categorical columns- date had 3436 unique values which results in high cardinality and requires feature engineering on date
   c.

4. In the numerical columns - replaced all the null values with the mean of each column respectively

5. In the categorical columns - replaced all the null values with mode of each column respectively

Exploratory Data Analysis

Feature Engineering

1. High Cardinality in Date

   Solution - Separated day and month from date column, then appended day and month columns and dropped date from the dataset

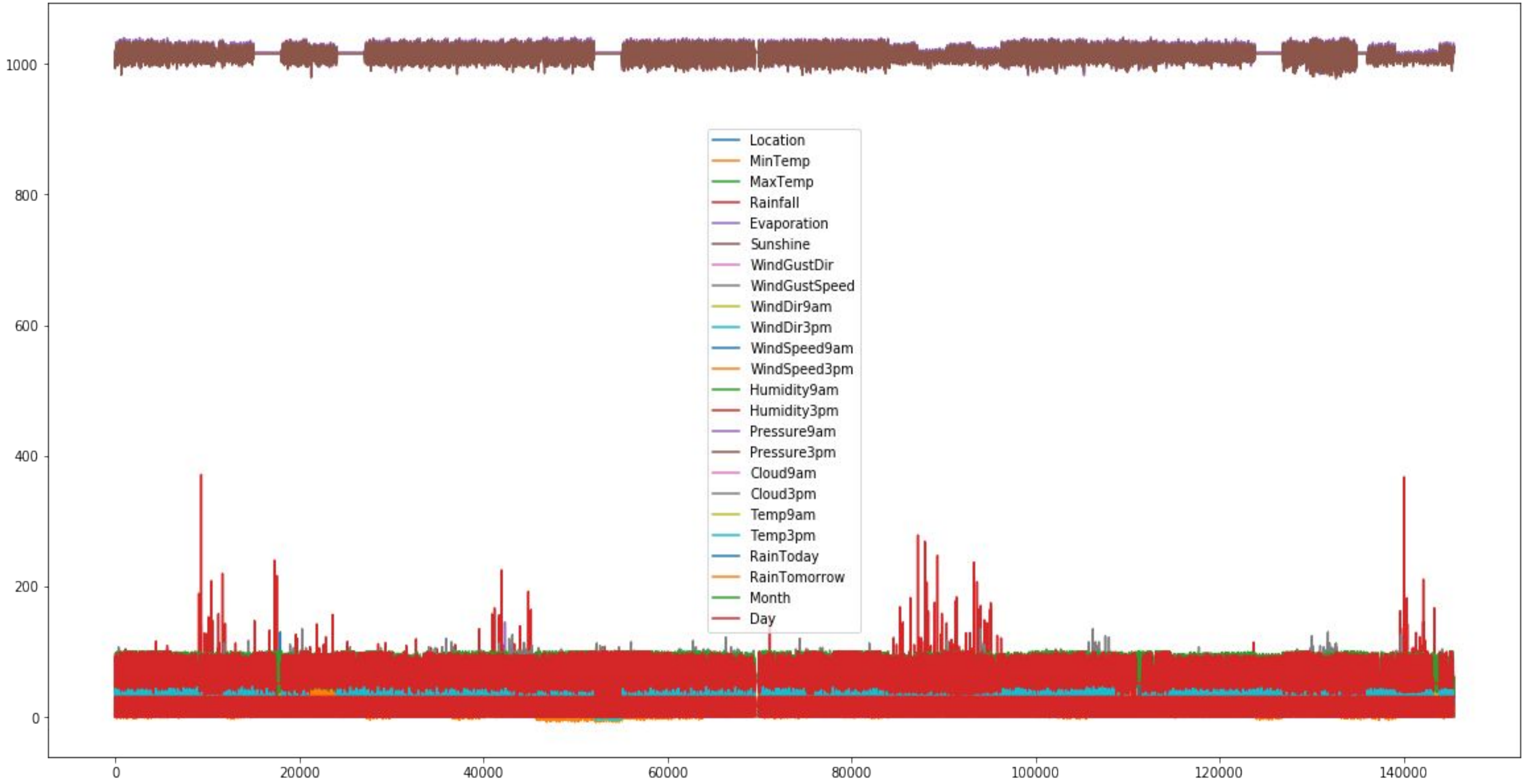
   This reduces cardinality as only 12 unique values on month exist, and 31 unique values for day exist compared to 3436 unique values of date alone

   Note - On experimental trial we also observed that if we included both day and month compared to only one of them, our algorithm models' accuracy improved, marginally though by 0.01 approx

2. Label encoded all the categorical values - from string to numbers

Exploratory Data Analysis

Feature Engineering - Plotting the graph for all the values

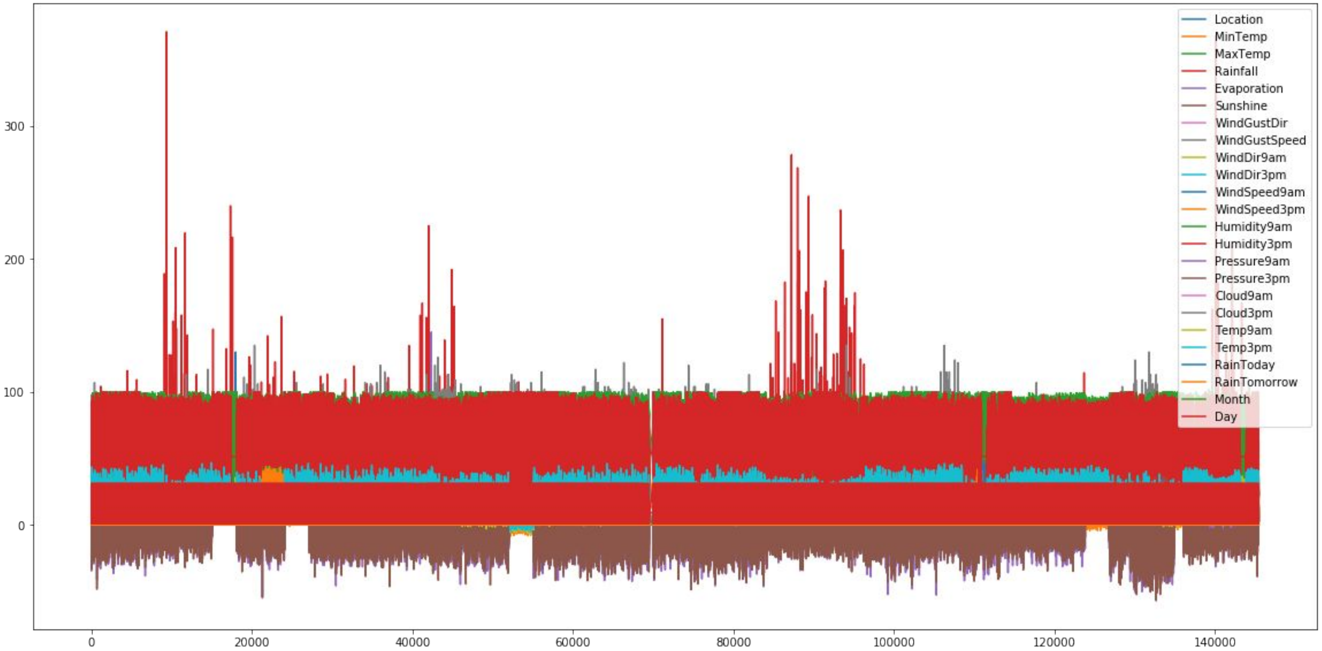


Exploratory Data Analysis

Feature Engineering - Normalisation

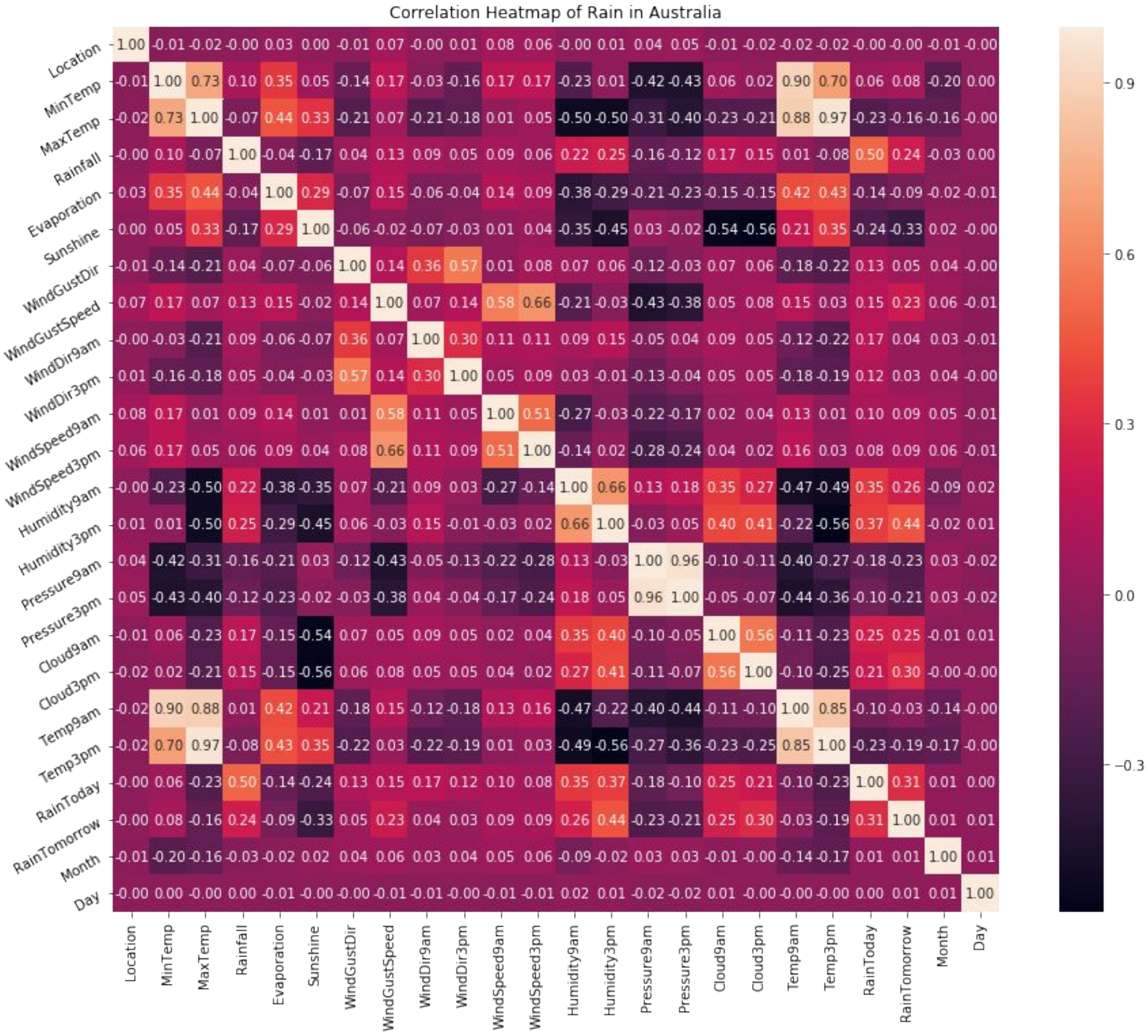
We noticed that while all other columns have values that are in close range, the values of Pressure9am and Pressure3pm are very much higher.

So these columns required to be normalised. We used **z-score** to normalise these columns and then multiplied it by 10. The new range observed was closely - (-55, 35) which brings it quite close to other columns' values.



Exploratory Data Analysis

Feature Engineering - Correlation HeatMap



Exploratory Data Analysis

Insights from the Correlation Heatmap

- MaxTemp and Temp3pm have an extremely high correlation - 0.97
- MinTemp and Temp9am have a high correlation - 0.90
- MaxTemp and Temp9am have a moderately high correlation - 0.88
- Location and RainTomorrow had no correlation - 0.0

Due to the first point of MaxTemp and Temp3pm having an extremely high correlation- Temp3pm was dropped and no more considered. On experimenting with this - a small improvement in our ML model was observed

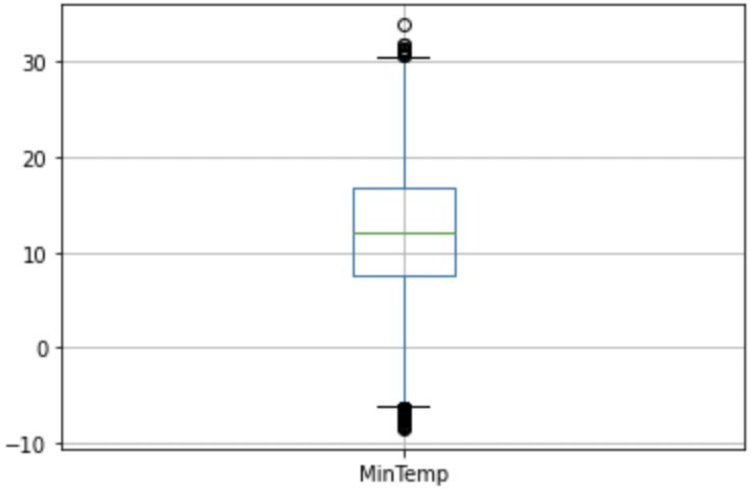
We also noticed that location and RainTomorrow had no correlation, but our accuracy on ML model was decreased when location was dropped

Exploratory Data Analysis

Outliers using the Boxplot

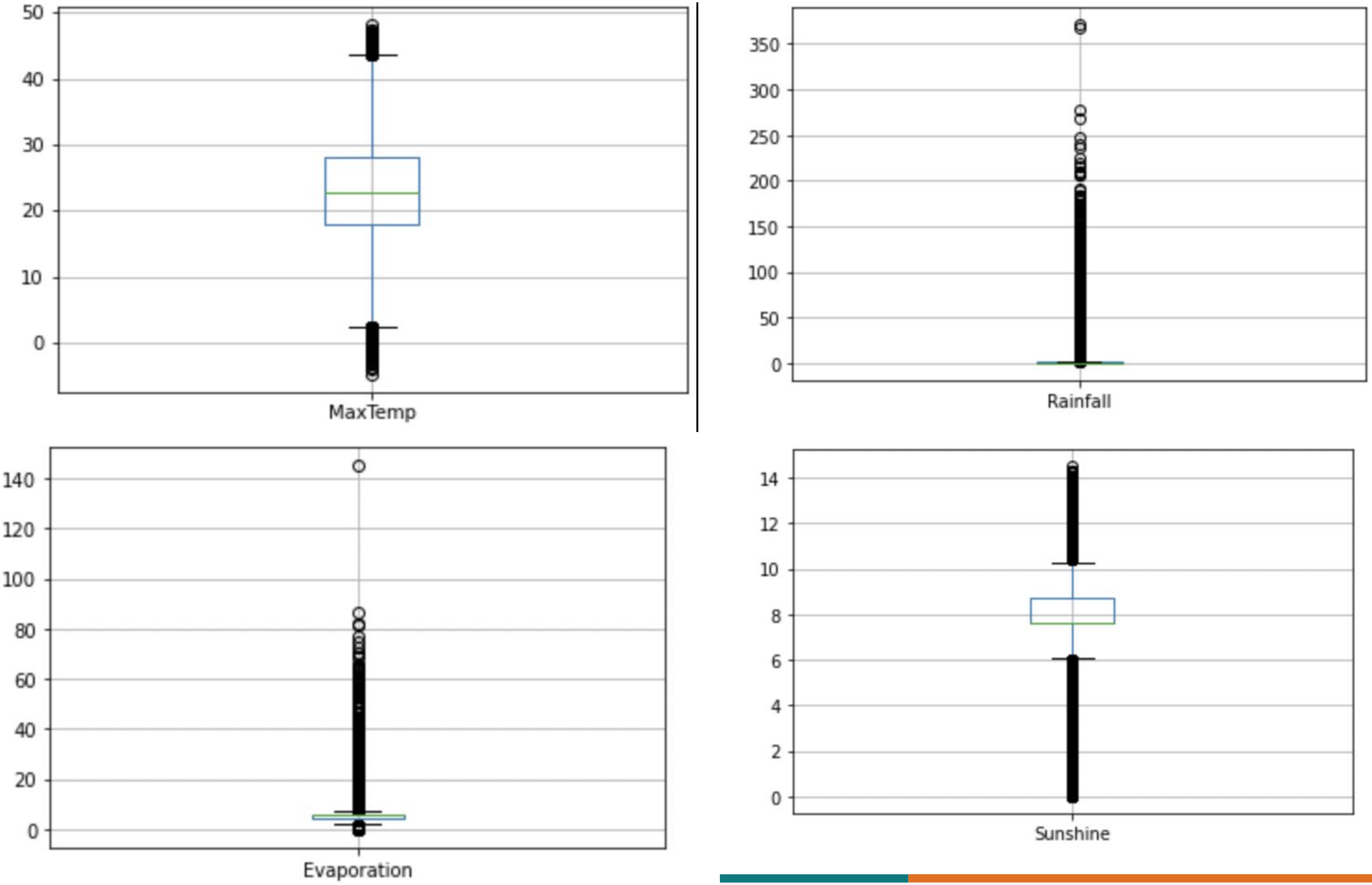
- MinTemp varies from -8.50 to a maximum value of 33.90 and the range e
- MaxTemp varies from -4.80 to a maximum of 48.10
- Rainfall shows a lot of variation in its value because sometimes it may rain very less of 0 and sometimes a bit on the extreme side upto 371
- Since Evaporation is the cause of rainfall it also varies a lot starting from 0 to a high of 145
- Sunshine varies from 0 to 14.5 with a mean value of 7.62

Following are the boxplot of all the variables taken into account:



Exploratory Data Analysis

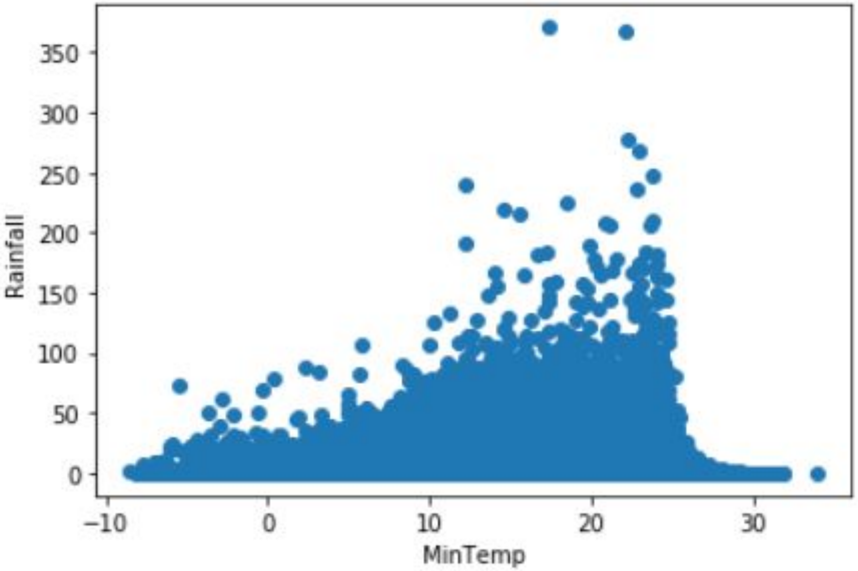
Box Plot for the required variables



Exploratory Data Analysis

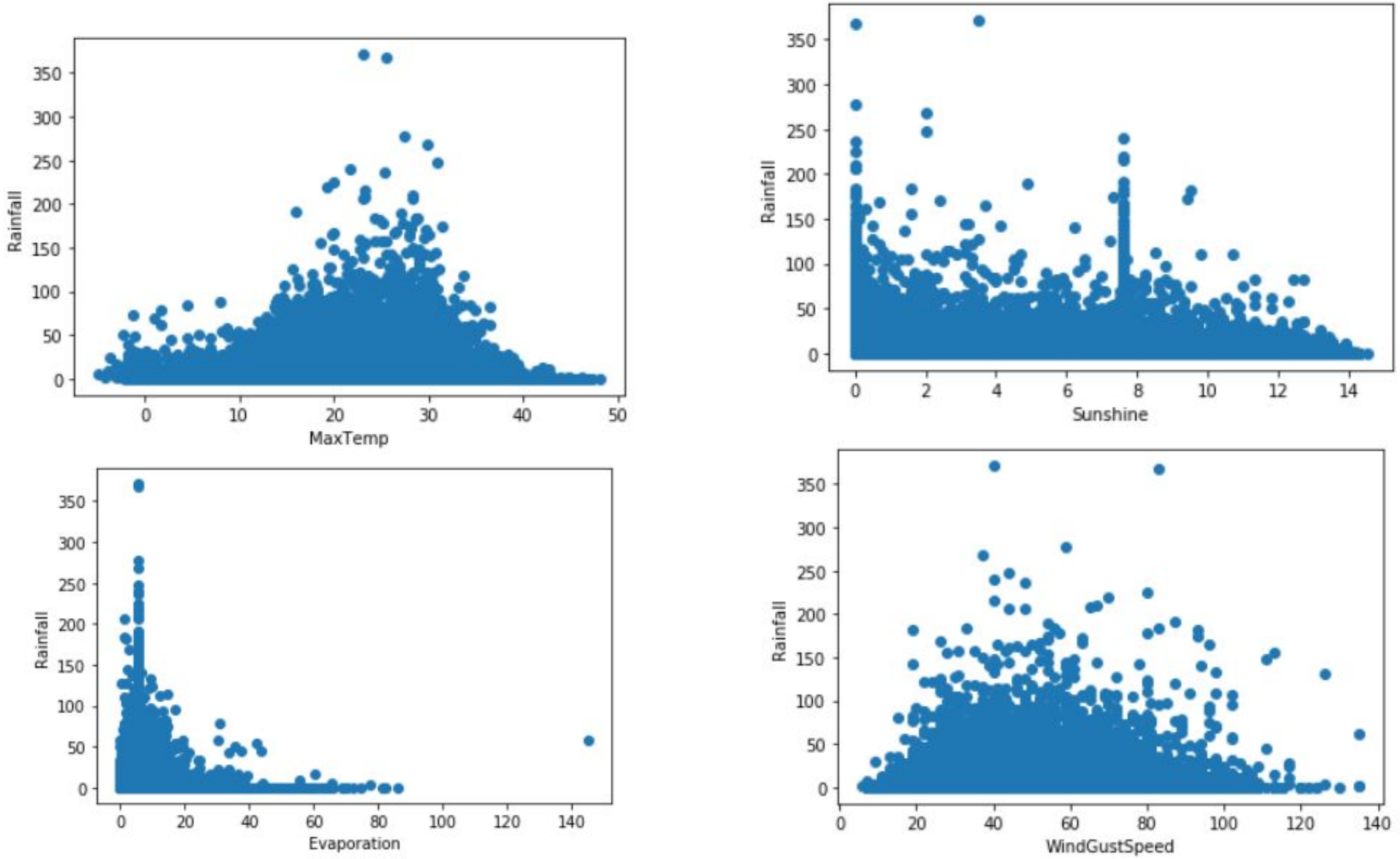
Scatter Plot for the required variables

- The diagram graphs pairs of numerical data, with one numerical data, with one variable on each axis, to look for a relationship between them.
- If the variables are correlated, the points will fall along a line or curve.
- The better the correlation, the tighter the points will hug the line.



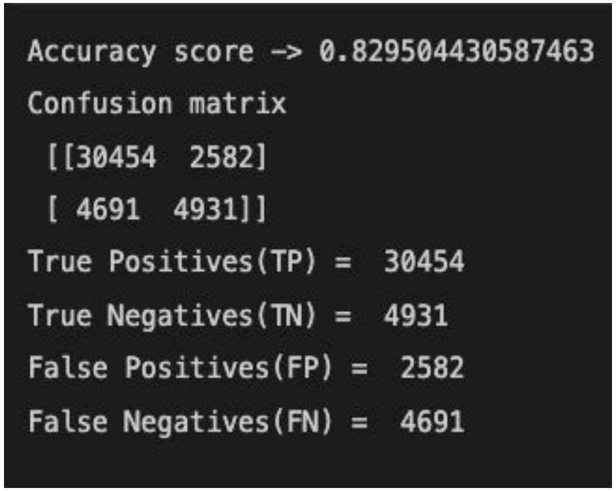
Exploratory Data Analysis

Scatter Plot for the required variables



KNN Methodology

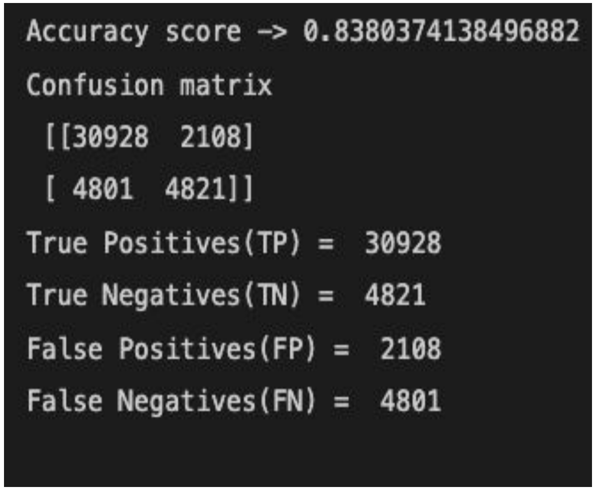
Accuracy (K = 3) ~ 82.95%
Accuracy (K = 5) ~ 83.80
Accuracy (K = 10) ~ 84.28



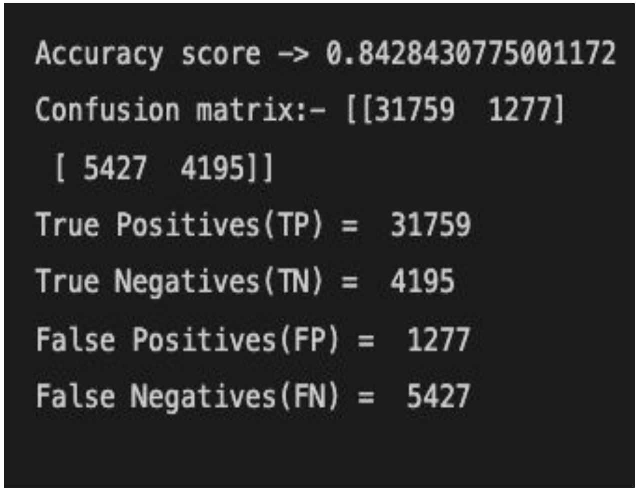
```
Accuracy score -> 0.8295044305874631
Confusion matrix
 [[30454  2582]
 [ 4691  4931]]
True Positives(TP) =  30454
True Negatives(TN) =  4931
False Positives(FP) =  2582
False Negatives(FN) =  4691
```

K = 3

K = 5



```
Accuracy score -> 0.8380374138496882
Confusion matrix
 [[30928  2108]
 [ 4801  4821]]
True Positives(TP) =  30928
True Negatives(TN) =  4821
False Positives(FP) =  2108
False Negatives(FN) =  4801
```

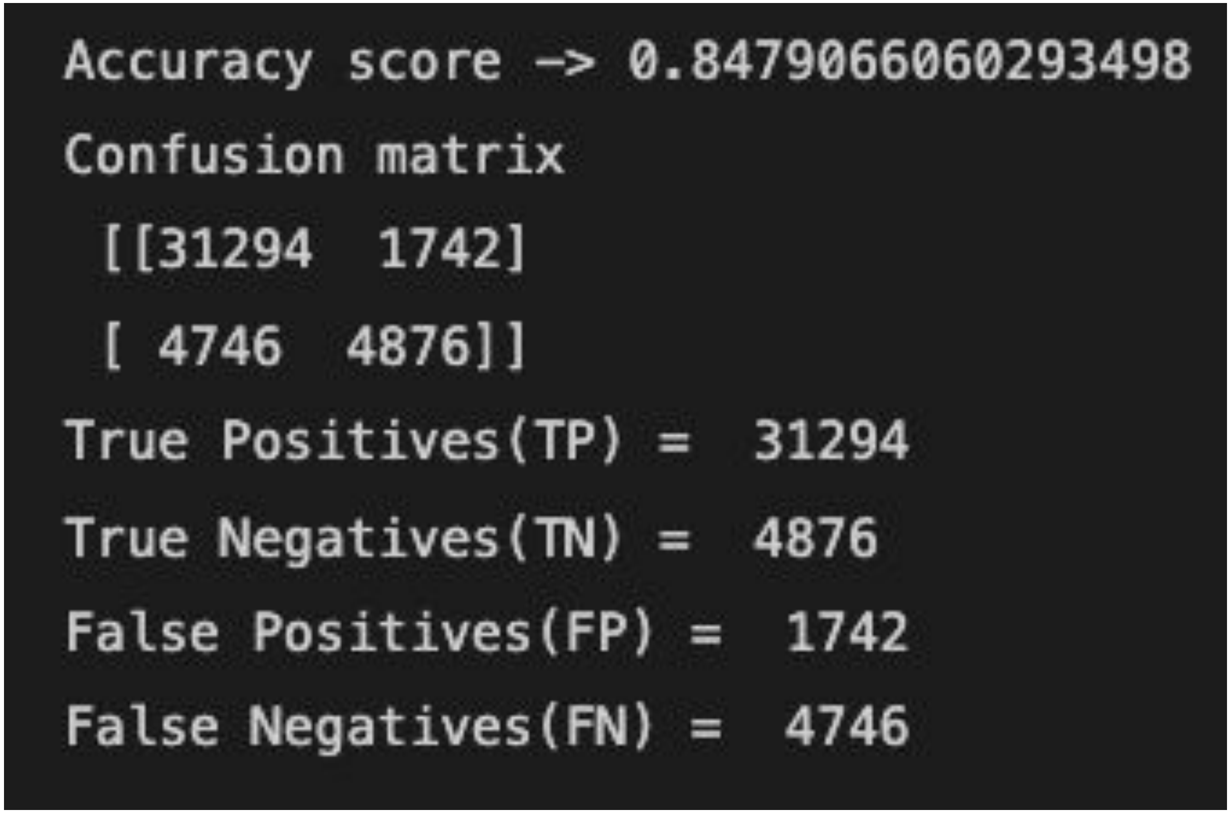


```
Accuracy score -> 0.8428430775001172
Confusion matrix:- [[31759  1277]
 [ 5427  4195]]
True Positives(TP) =  31759
True Negatives(TN) =  4195
False Positives(FP) =  1277
False Negatives(FN) =  5427
```

K = 10

Adaboost

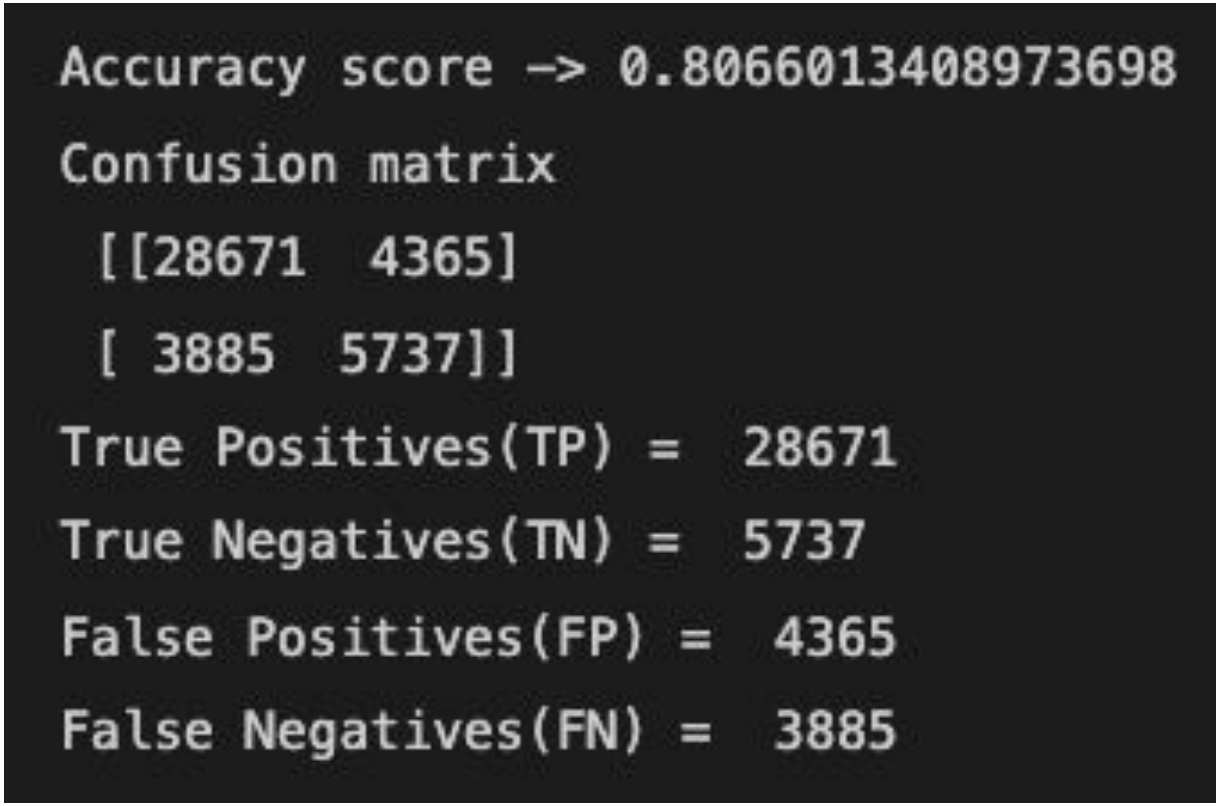
Accuracy ~ 84.79%



```
Accuracy score -> 0.8479066060293498
Confusion matrix
 [[31294  1742]
 [ 4746  4876]]
True Positives(TP) =  31294
True Negatives(TN) =  4876
False Positives(FP) =  1742
False Negatives(FN) =  4746
```

Gaussian Naive Bayes

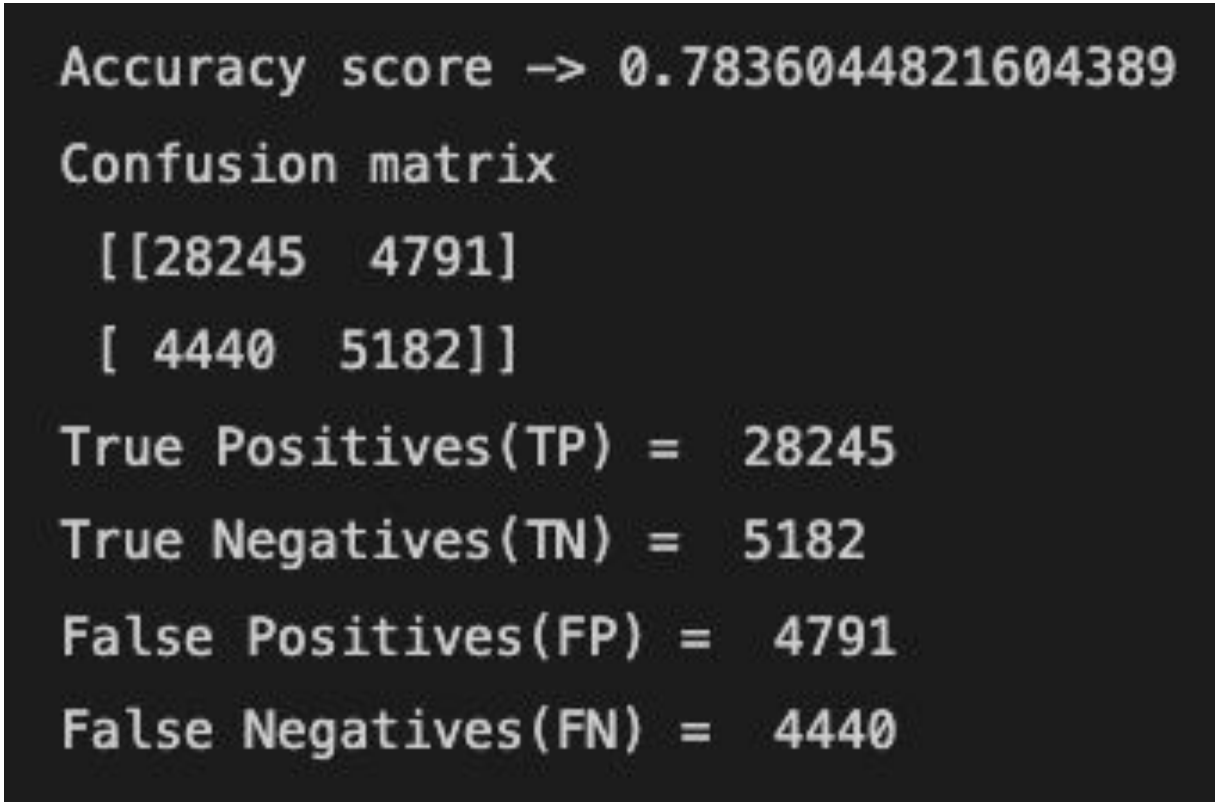
Accuracy ~ 80.66%



```
Accuracy score -> 0.8066013408973698
Confusion matrix
 [[28671  4365]
 [ 3885  5737]]
True Positives(TP) =  28671
True Negatives(TN) =  5737
False Positives(FP) =  4365
False Negatives(FN) =  3885
```

Decision Trees (CART)

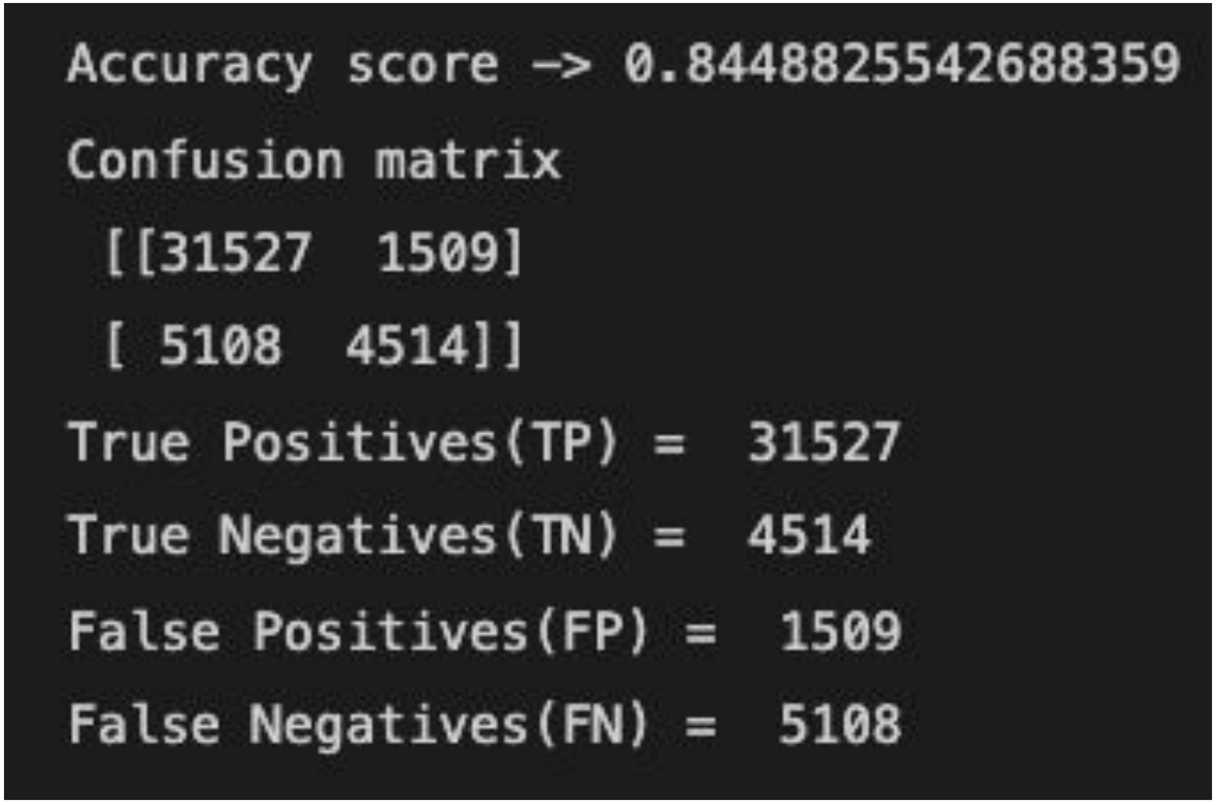
Accuracy ~ 78.36%



```
Accuracy score -> 0.7836044821604389
Confusion matrix
 [[28245  4791]
 [ 4440  5182]]
True Positives(TP) =  28245
True Negatives(TN) =  5182
False Positives(FP) =  4791
False Negatives(FN) =  4440
```

Random Forest

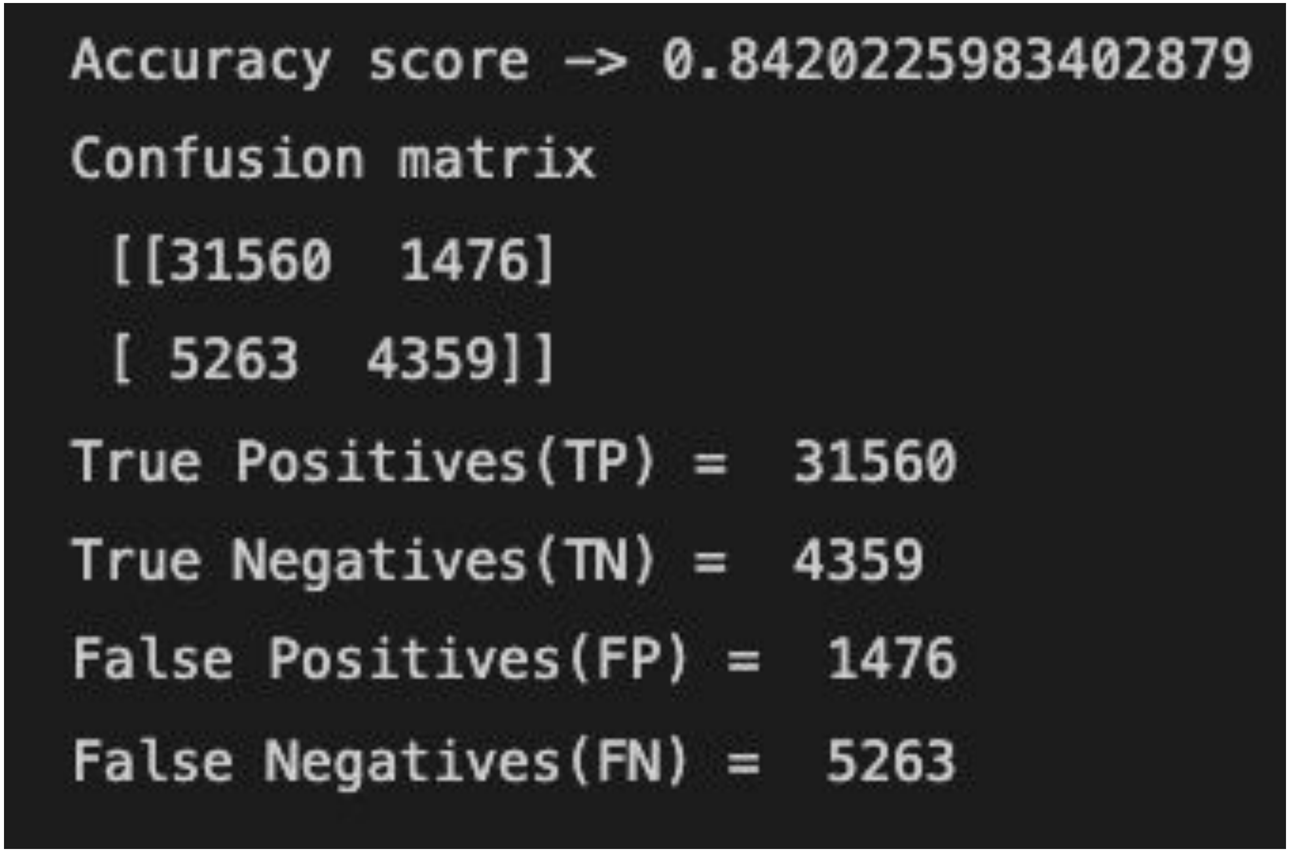
Accuracy ~ 84.45%



```
Accuracy score -> 0.8448825542688359
Confusion matrix
 [[31527  1509]
 [ 5108  4514]]
True Positives(TP) =  31527
True Negatives(TN) =  4514
False Positives(FP) =  1509
False Negatives(FN) =  5108
```

Extra Trees

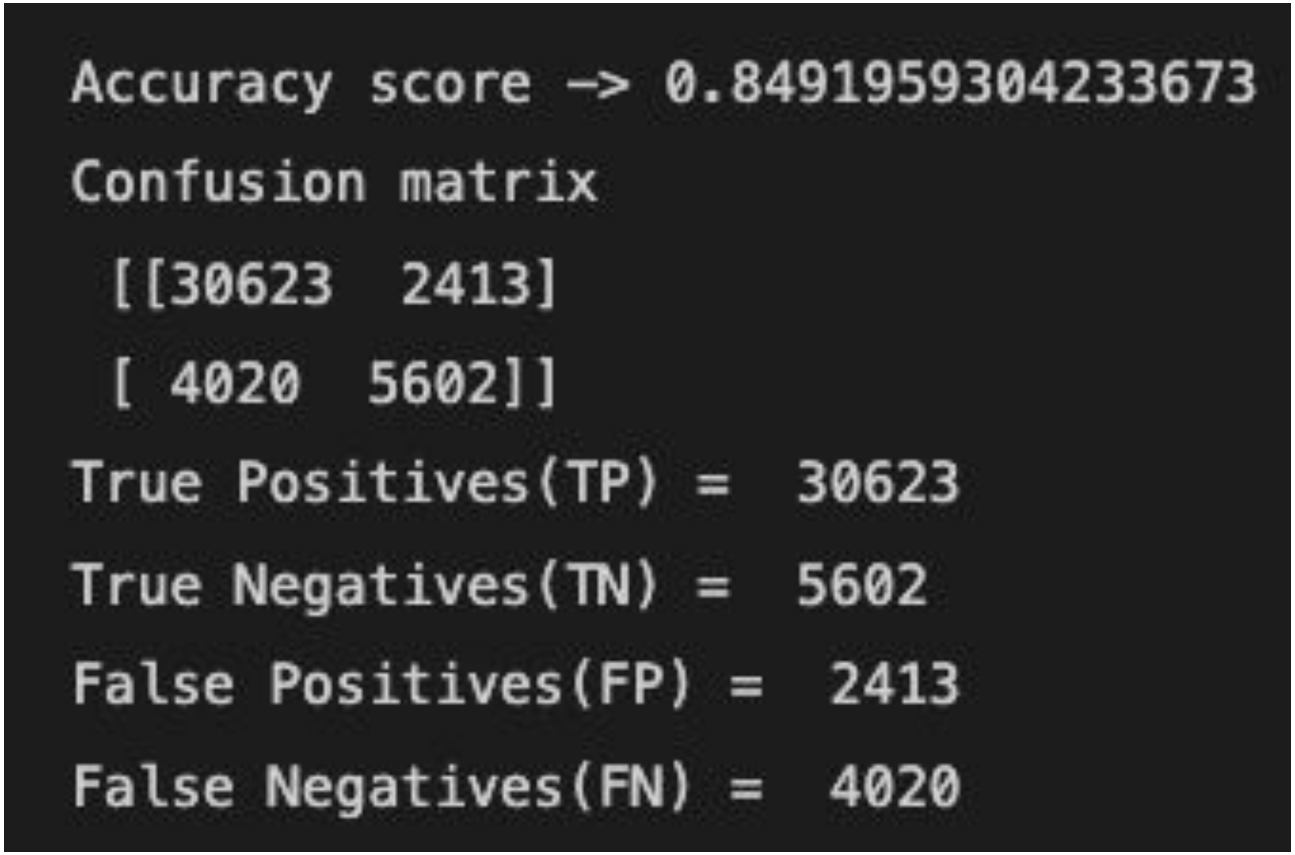
Accuracy ~ 84.20%



```
Accuracy score -> 0.8420225983402879
Confusion matrix
 [[31560  1476]
 [ 5263  4359]]
True Positives(TP) =  31560
True Negatives(TN) =  4359
False Positives(FP) =  1476
False Negatives(FN) =  5263
```

Artificial Neural Network (ANN)

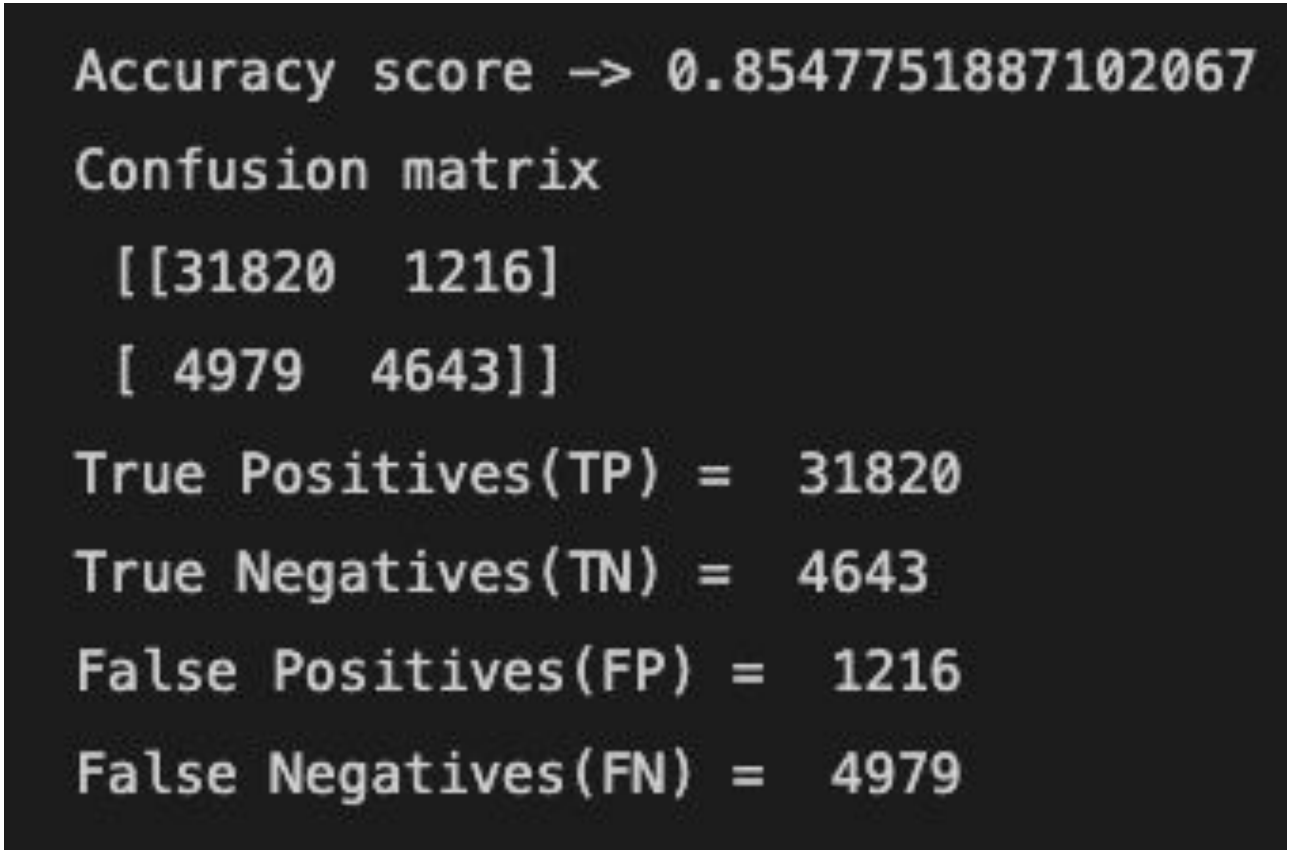
Accuracy ~ 84.92%



```
Accuracy score -> 0.8491959304233673
Confusion matrix
 [[30623  2413]
 [ 4020  5602]]
True Positives(TP) =  30623
True Negatives(TN) =  5602
False Positives(FP) =  2413
False Negatives(FN) =  4020
```

SVM Methodology

Accuracy ~ 85.48%



```
Accuracy score -> 0.8547751887102067
Confusion matrix
 [[31820  1216]
 [ 4979  4643]]
True Positives(TP) =  31820
True Negatives(TN) =  4643
False Positives(FP) =  1216
False Negatives(FN) =  4979
```

Conclusion and Insights

Predicting rain is one of the most difficult thing and to date with advanced science and precision tools, rain prediction stays a difficult job and the probability is always varying.

Despite that, based on climate and history - our algorithms are able to give an accuracy of over 85 per cent after all the preprocessing and engineering steps are taken into account. This efficiency can further be improved if outliers are worked upon.

This data is limited to the locations of Australia and is only for last 10 years. Expansion of data can help with better results and this methodology can be used over other countries and locations as well. It can help predict storms and heavy rainfall if worked upon correctly. Now wouldn’t that be great?

1870

STEVENS
INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®

stevens.edu

THE END