MAJOR PROJECT DEC -ML12B1

Name: Aman Kumar

What is the average rainfall in Cairns?

=>

5.68

Which place has the rainfall above 200 cm?

=>

Location	Rainfall
Cairns	206.2
Townsville	206.8
CoffsHarbour	208.5
Darwin	210.6
CoffsHarbour	219.6
Williamtown	225.0
Townsville	236.8
Cairns	247.2
Cairns	268.6
Cairns	278.4
Darwin	367.6

Summary of the project:

The dataset used was about weather in Australia.

Independent variables:

'Location', 'MinTemp', 'MaxTemp', 'Rainfall', 'Evaporation', 'Sunshine','WindGustDir', 'WindGustSpeed', 'WindDir9am', 'WindDir3pm', 'WindSpeed9am', 'WindSpeed3pm', 'Humidity9am', 'Humidity3pm', 'Pressure9am', 'Pressure3pm', 'Cloud9am', 'Cloud3pm', 'Temp9am','Temp3pm', 'RainToday'

Target Variable:

'RainTomorrow'

The model should be designed such that it takes all the necessary values as input and predict whether it would rain tomorrow or not. The target variable is of binary classification.

The three algorithms used were:

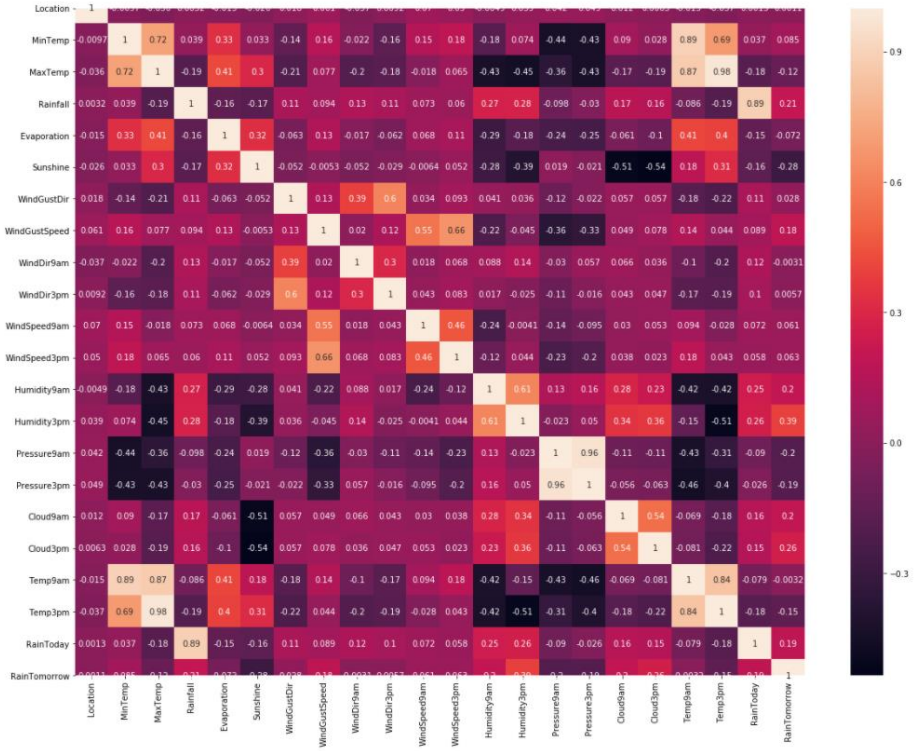
1. Logistic Regression
2. KNN (K-Nearest Neighbours)
3. SVM (Support Vector Machine)

After implementing them here are the accuracy score:

1. Logistic regression : `0.8603611657428929`
2. KNN: `0.8607634543178974`
3. SVM: `0.8237976041480422`

After comparing them it can be said the KNN marginally has higher accuracy than Logistic regression.

Since SVM doesn't perform well with large datasets it took much time to execute than Logistic and KNN.



For the above problem KNN would be best even though it takes more time to train the model because the dataset consists of few not fully ‘independent’ variables (see the heatmap figure).

Few variables have dependency greater than 0.4 even 0.5 thus it wouldn’t be suitable for logistic regression.

Best suitable Algorithm for this dataset:

KNN (K-Nearest Neighbours)