**Machine Learning project for rainfall prediction**



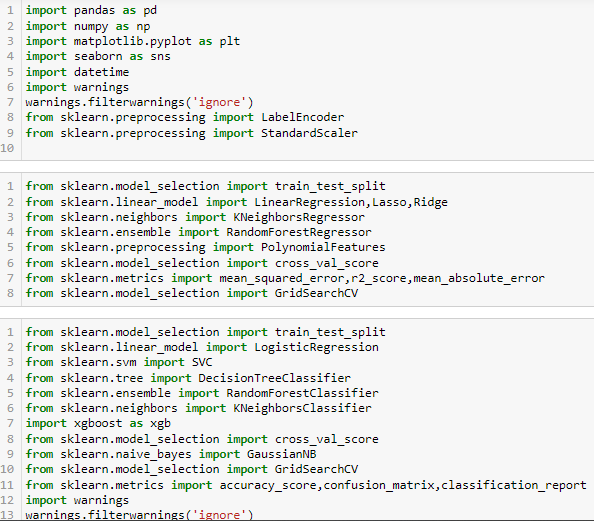
Hello there,

We all know the role of rain for sustaining life or earth. This project aims at predicting the amount of rainfall, or if there will be rainfall tomorrow or not.

**Problem statement:**

1. Design a predictive model with the use of machine learning algorithms to forecast whether or not it will rain tomorrow.
2. Design a predictive model with the use of machine learning algorithms to predict how much rainfall could be there.

**Important relevant libraries:**

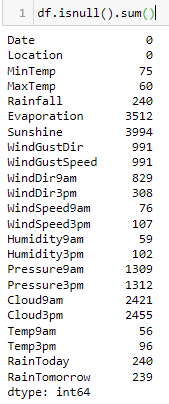
****

**Source of data:**

Data has been stored inside a csv file namely ‘weather1.csv’.

The dataset has 8245 rows and 23 columns.

**Gaining information from data:**

****

As evident from the above metadata, the dataset contains huge number of null values which needs to be dealt with. Since some columns contains very high number of null values, removing those rows would mean huge loss of data. So we have to replace the data with either mean, median or mode.

We will be trying all the above options and some default regression and classification models on them and check which dataset is performing best.

mean.PNG mean

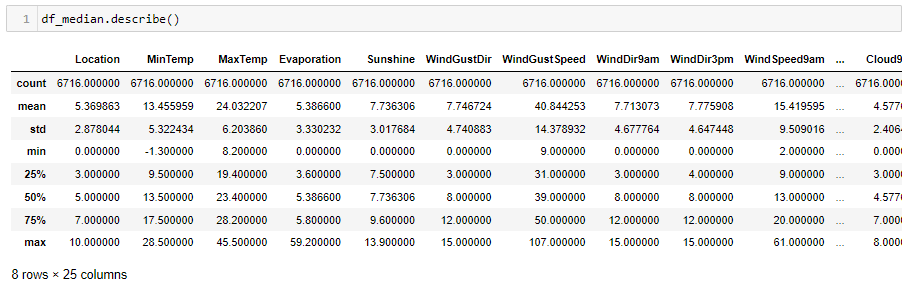
medain.PNG median

mode.PNG mode

df\_median is giving the best results for the regression problems. So for calculation rainfall amount we are using df\_median.

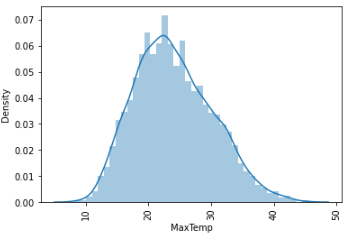
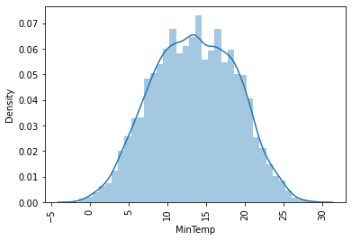
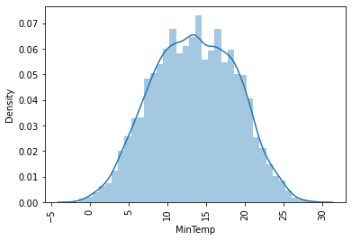
df\_median is also giving the best results for classification problems. So we are approaching futher with df\_median.

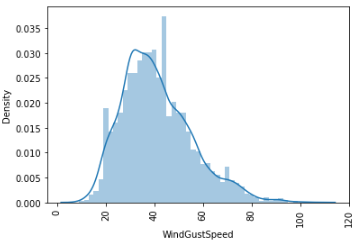
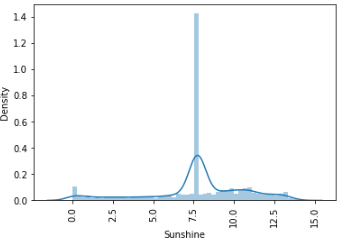
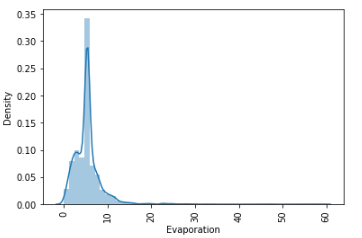
**Exploratory data analysis**

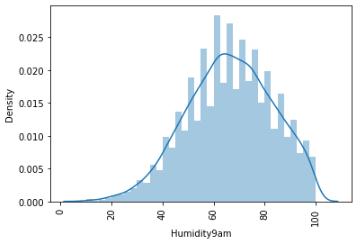
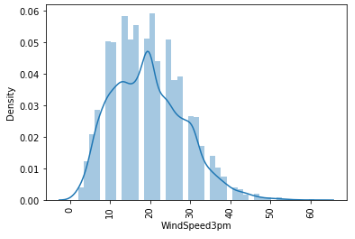
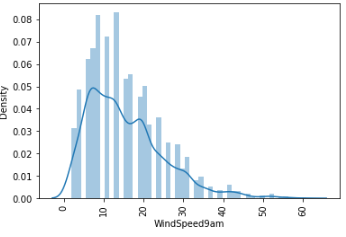


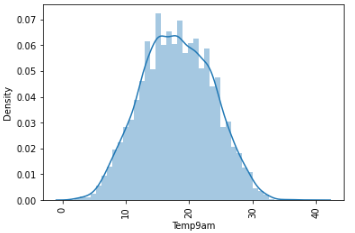
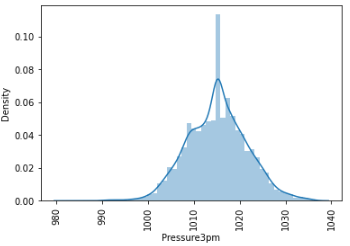
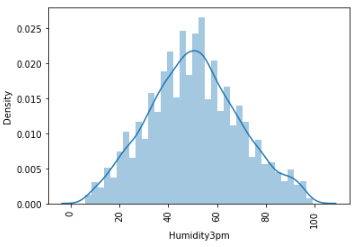
Some columns has high difference between mean and median, hence high skewness might be present. Also some columns has high difference between 75th percentile and max. Hence outliers might be present.

**Univariate analysis:**



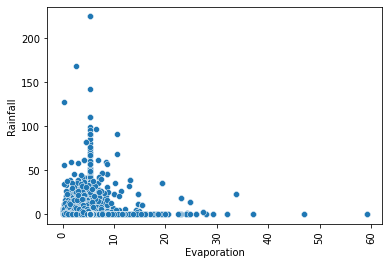
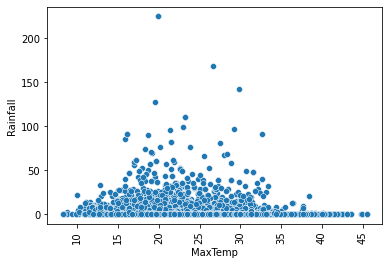
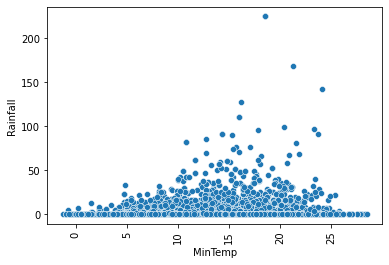


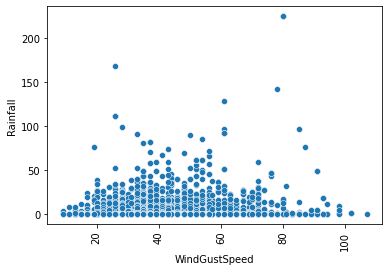
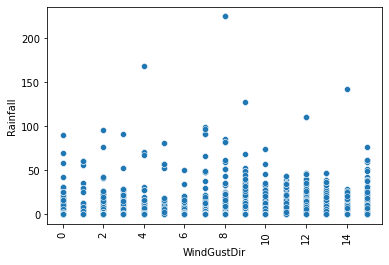
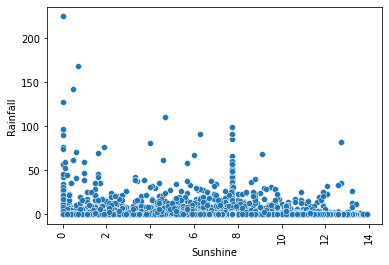


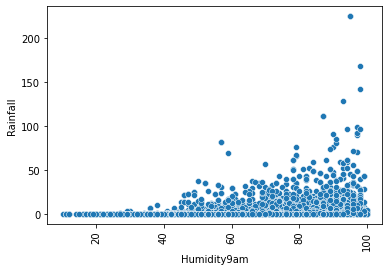
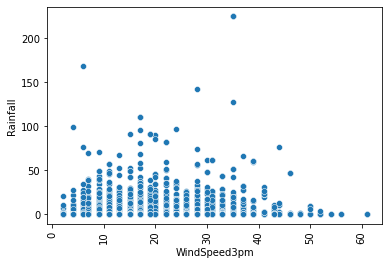
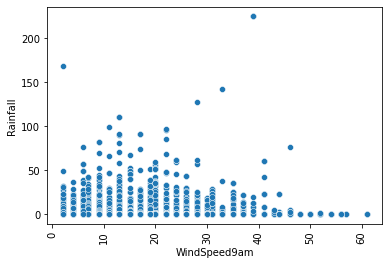


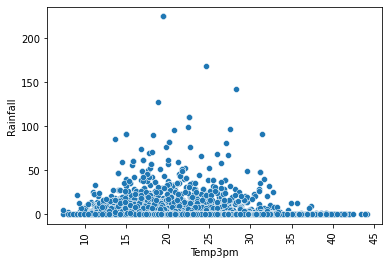
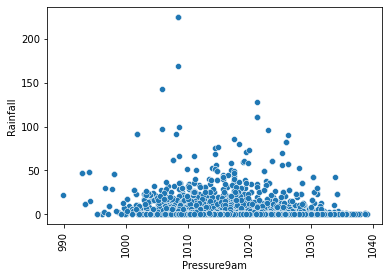
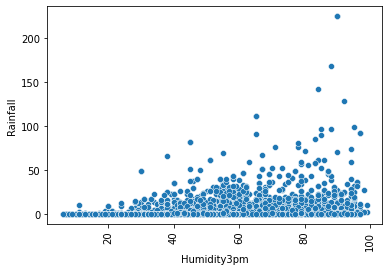
It is evident from univariate analysis that skewness is present in many columns which needs to be removed.

**Bivariate analysis:**

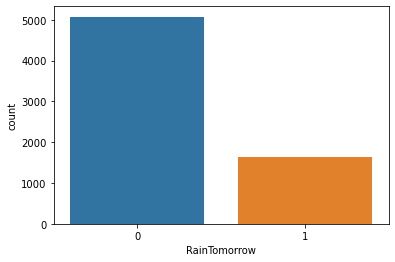








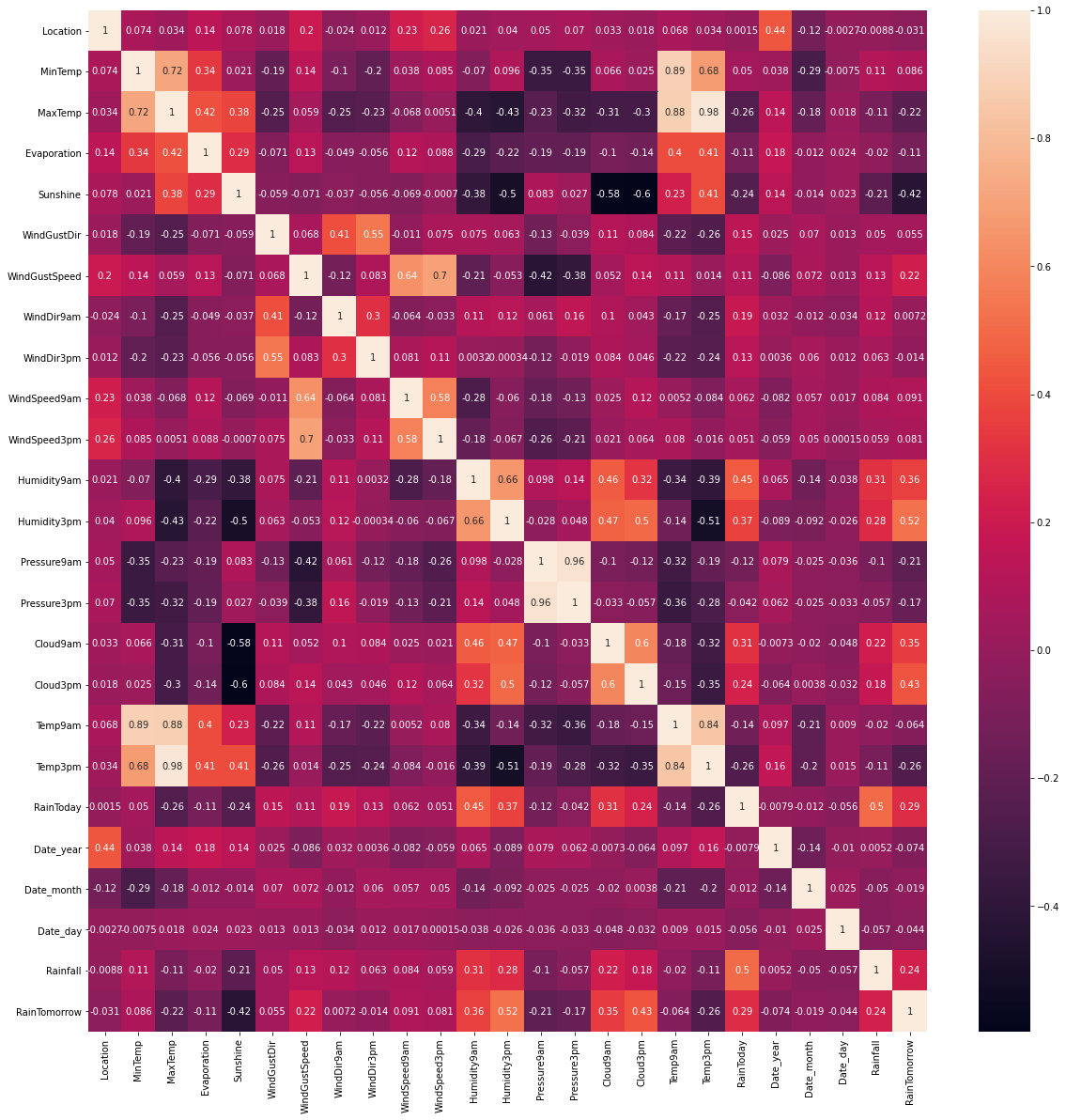
**Data preprocessing:**

Imbalance is present in the labels. So we are applying Synthetic Minority Oversampling Technique(SMOTE) is used here to balance both the labels.

Some columns has very low correlation with the target. Those columns are increases the errors in Linear and Logistic regression problems. So here we will be creating two datasets and test their accuracy on machine learning algorithms:-

1. df🡪 columns having low correlation with the target have been removed.
2. df\_irev🡪 some irrelevant columns will still be present.

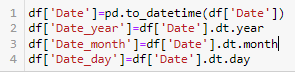


We have used this heatmap to check the correlation of attributes with target.

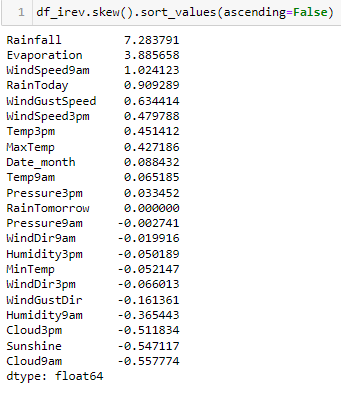
In df dataset 'WindDir9am', 'Location', 'WindGustDir', 'WindDir3pm', 'Date\_year', 'Date\_month', 'Date\_day' columns have been removed.

In df\_irev dataset, only 'Location','Date\_year','Date\_day' columns have been removed.

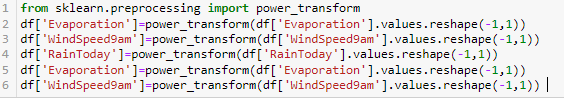
**Note:** Previously we have converted the dates from object to DateTime format , extracted the day,month and year and stored them in columns in the dataset.



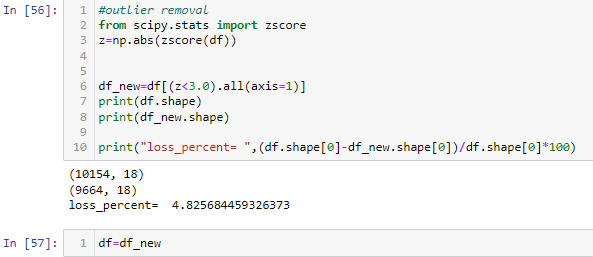
The columns were then checked for skewness.



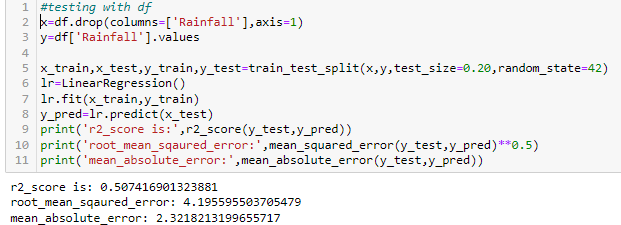
Skewness removal was done using power\_transform only on columns having high skewness.

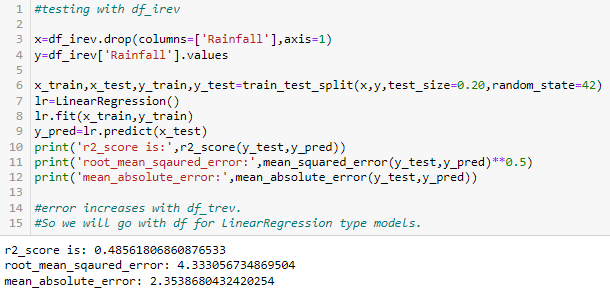


Outliers were removed after skewness removal.



**Rainfall prediction model training**

 results from df

 results from df\_irev.

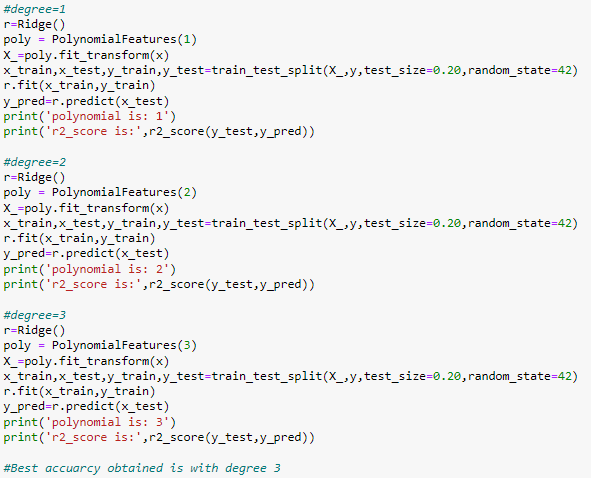
From the above two snaps, we have concluded that, LinearRegression model is performing better with dataset df. So we will approach futher with df in the LinearRegression problems.

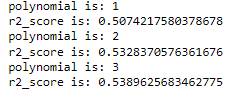
Similarly we will check , which dataset is working better with which algorithm and do hyper parameter tuning and training of that algorithm with the better dataset among df and df\_irev.

The regression algorithms used for rainfall prediction are:-

1. Linear Regression
2. Ridge Regerssion
3. Lasso Regression
4. KNN regression
5. RandomForest regression

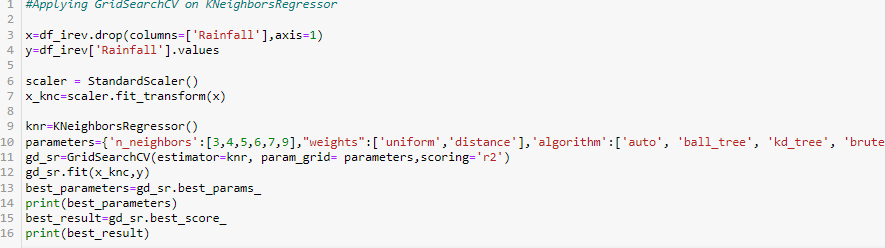
In Linear,Ridge,Lasso regression we had to find the best polynomial. The code for determining the best polynomial is shown below.





Best accuracy is obtained with degree=3. Hence will approach further with degree=3

Of all the models tried, we have obtained best results from KNeighbors Regressor. So we are discussing the hyperparameter tuning of KNeighbors regressor only.



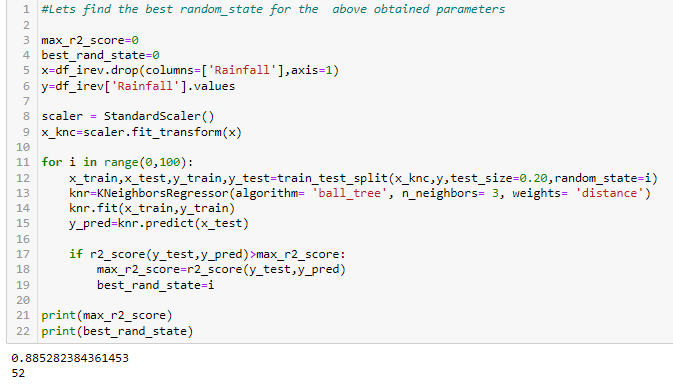
Before applying KNeighbors regression, we have applied Standard scaler because KNeighbors regressor is a distance based algorithm.

The best parameters obtained from hyperparameter tuning are:-

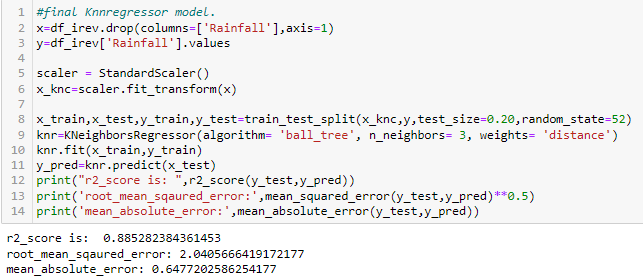
'algorithm': 'ball\_tree', 'n\_neighbors': 3, 'weights': 'distance'.

We have not applied cross validation separately because GridSearchCV takes cv=5 by default.

After finding the best parameters, we have used the obtained parameters to find the best random state for KNeighbors Regression model.



We have used the best random\_state obtained, i.e 52 and the best parameters obtained from hyperparameter tuning to train the final Linear Regression model.



**Model to predict tomorrow will be rainfall or not?**

The algorithms used to build the classification model are as follows:-

1. Logistic Regression

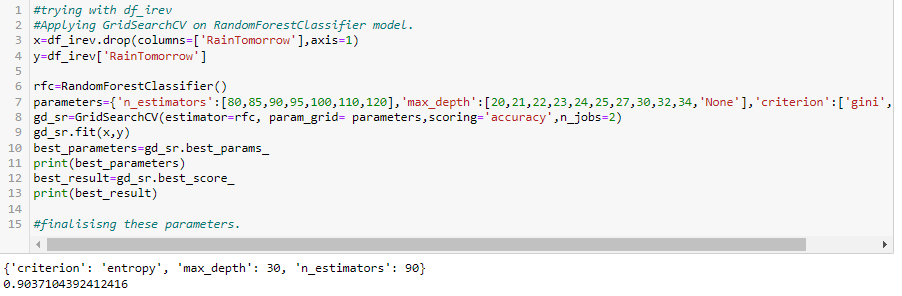
2. KNeighbors Classifier

3. Gaussian NaiveBayes algorithm

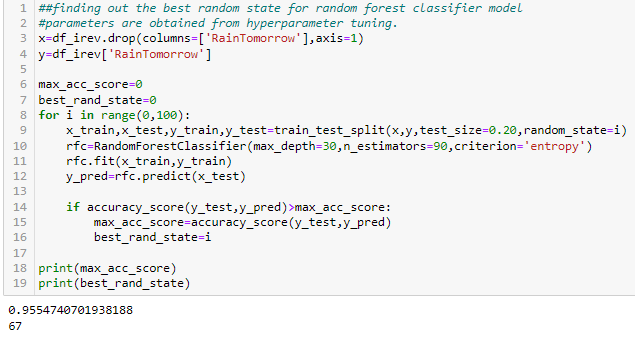
4. RandomForest Classifier model.

The best model is the Random Forest Classifier model.

We are discussing only the Random Forest Classifier model here.

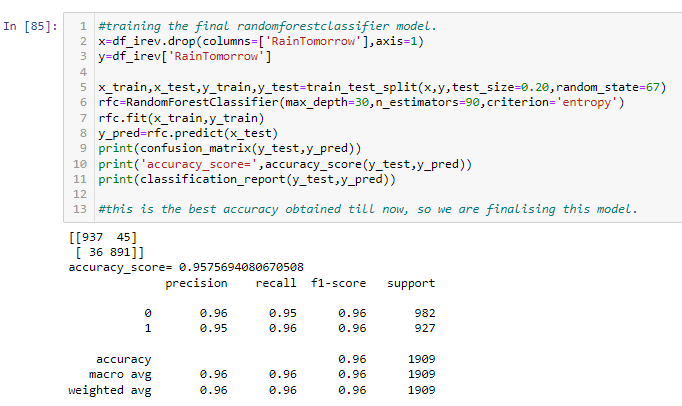
The best parameters for training the model are obtained from hyperparameter tuning.

The best random state for training the model are obtained using the best parameters obtained.



We train the final model using the best parameters ('criterion': 'entropy', 'max\_depth': 30, 'n\_estimators': 90) and the best random\_state, i.e 67

The training of the final model is shown below.



The AUC-ROC curve for the final model is shown below.

