Rain Prediction –Weather forecasting

Problem Statement-

Rain Dataset is to predict whether or not it will rain tomorrow. The Dataset contains about 10 years of daily weather observations of different locations in Australia.

a) Design a predictive model with the use of machine learning algorithms to forecast whether or not it will rain tomorrow.

b) Design a predictive model with the use of machine learning algorithms to predict how much rainfall could be there.

Dataset Description:

Number of columns: 23

THEY ARE:

Date, Location, MinTemp, MaxTemp, Rainfall, Evaporation, Sunshine, WindGustDir, WindGustSpeed, WindDir9am, WindDir3pm, WindSpeed9am, WindSpeed3pm, Humidity9am, Humidity3pm, Pressure9am, Pressure3pm, Cloud9am Cloud3pm, Temp9am, Temp3pm, Rain Today, Rain Tomorrow

FEATURE ENGINEERING:

We had seen many null values in most of the columns, so it should be filled as for that we will use Imputer techniques to fill the data.

As, for imputers we have used **Simple imputer, KNN Imputer, Iterative Imputer.**

After filling the data, we have also observed that all columns are object data type, which is not correct as some columns are numeric but still it shows object data type.

So, we have also converted those object columns into numeric.

EXPLORATORY DATA ANALYSIS:

We have analysed columns while taking the best possible scenario,

As we analyse the columns individually, we observed

- Melbourne has the most amount of rainfall
- The strongest wind gust comes mostly from North of Australia
- The wind direction at 9am shows it at North
- We can see that at 3pm the wind direction changes and comes to Southeast and also good amount of data shows South.
- We can also see as per value count both days has majority that it will not rain.
- We have also tried to analyse with line plot
    - As we have taken Sunlight and Rainfall, we can see a constant line which shows that they don't posses any relation between them.
    - We can also see that rainfall and evaporation had a directly proportional relation relationship between them.
    - We had also seen that cloud columns also have a directed relationship with rainfall.

- We have also tried to observe through Histogram Plot
- As we know that it will help mostly to observe the numerical columns so, as we can observe that mostly the columns are normally distributed, but few columns show some skewness windspeed9am, windspeed3pm are right skewed, Humidity9am is left skewed.

- We have also observed that mean and median of the dataset is somewhat similar and show the data to be normal.

As, from the above analysis we conclude that the data the wind is blowing from north to south from morning to evening rainfall also has relation with clouds and evaporation and majority of the data shows rain will not occur.

Pre-Processing Pipeline:

As for the object columns we have encoded the columns with label encoder show that we can build a model and interpret the data.

CORRELATION:

As we can observe that Rain Tomorrow Correlation humidity3pm and least with Sunshine.

Building Machine Learning Models:

Firstly, as we know that our problem statement shows us to build two types of models.

- As for predicting it will rain tomorrow or not, we have built Classification models.
- For predicting the amount of rainfall, we have built Regression model

So, as for the Classification model we have build different types of classification model-

- **Logistic Regression Model-**

  As we can see that this model is giving us the accuracy score of 83 %. It also shows the cv at 6 with the best score.

- **KNN Classifier Model-**

  As before getting the accuracy score lets do some hyper parameter tuning, which gives the best parameters as-
  algorithm-kd tree, leafsize-10, n_neighbours=1

  As we can see that this model is giving us the accuracy score of 84%.and cross validation score 78%.

KNeighbors Classifier report:

We can see that precision, recall and f1score for both the outputs are Close.

	precision	recall	f1-score	support
0	0.89	0.91	0.90	2589
1	0.68	0.64	0.66	781
accuracy			0.85	3370
macro avg	0.79	0.78	0.78	3370
weighted avg	0.85	0.85	0.85	3370

- **Random Forest Classifier:**

As before getting the accuracy score let's do some hyper parameter tuning, which gives the best parameters as-
criterion=gini, max_features=auto

As we can see that this model is giving us the accuracy score of 88%.and cross validation score 83%.

Random Forest Classifier Report:

	precision	recall	f1-score	support
0	0.90	0.96	0.93	2589
1	0.84	0.64	0.72	781
accuracy			0.89	3370
macro avg	0.87	0.80	0.83	3370
weighted avg	0.88	0.89	0.88	3370

➢ Decision Tree Classifier:

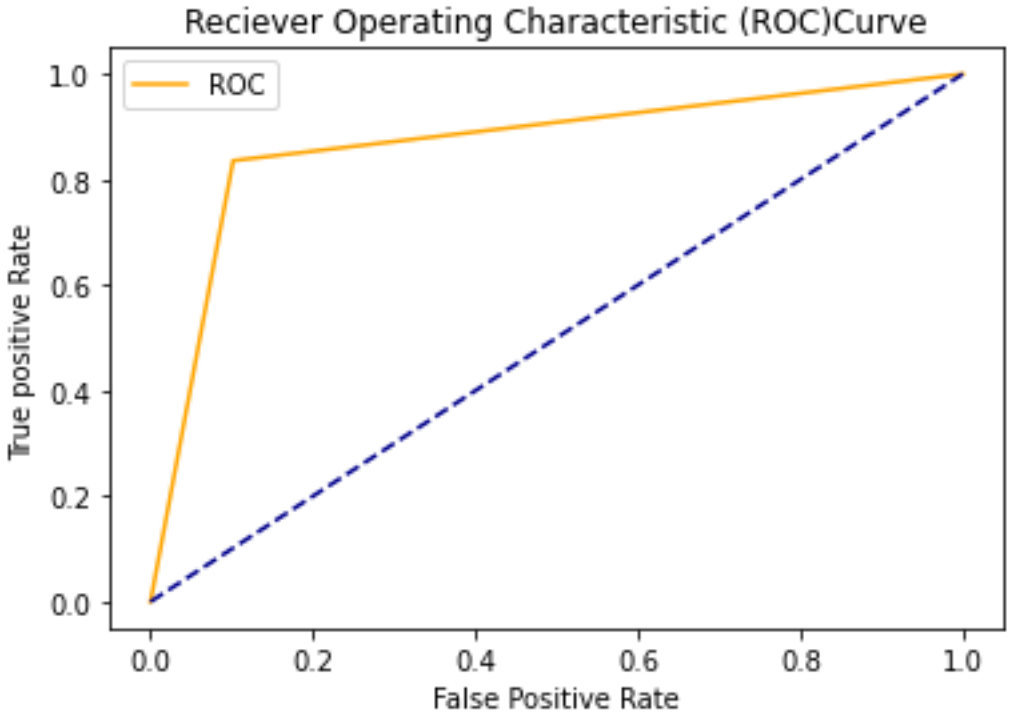
As before getting the accuracy score let's do some hyper parameter tuning, which gives the best parameters as-
criterion=entropy, max_features=auto, splitter='best'

As we can see that this model is giving us the accuracy score of 82%.and cross validation score 74%.

Decision Tree Classifier Report:

	precision	recall	f1-score	support
0	0.89	0.88	0.88	2589
1	0.61	0.64	0.63	781
accuracy			0.82	3370
macro avg	0.75	0.76	0.75	3370
weighted avg	0.83	0.82	0.82	3370

Aoc Roc as per random forest classifier:



As we conclude from model building that random forest classifier model is giving the best score so, we are going to save this model to predict.

As for the Auc Roc Curve it gives a secured result.

Now, as for the second prediction we have build a build different types of regression models-

- **Linear Regression Model-**

As we can see that this model is giving us the accuracy score of 80 %. It also shows the cv at 6 with the best score.
We will also do some hyper parameter tuning which gives best parameters for-

Lasso - alpha=0.0001, random state=17, selection='random'
Ridge - alpha=10, random state=37, solver='sag'
It also gives a cross validation score as 80%.

- **<u>Random Forest Regressor:</u>**

As before getting the accuracy score let's do some hyper parameter tuning, which gives the best parameters as-
criterion='mse', max_features='auto'

As we can see that this model is giving us the accuracy score of 87%.and cross validation score 84%.

- **K Neighbors Regressor:**

As before getting the accuracy score lets do some hyper parameter tuning, which gives the best parameters as-
algorithm-kd tree, leafsize-10, n_neighbours=5

As we can see that this model is giving us the accuracy score of 76%.and cross validation score 76%.

- **Bagging Regressor:**

As before getting the accuracy score let's do some hyper parameter tuning, which gives the best parameters as-
max_features=3, random state=2

As we can see that this model is giving us the accuracy score of 58%.and cross validation score 52%.

- **Decision Tree Regressor:**

As before getting the accuracy score let's do some hyper parameter tuning, which gives the best parameters as-
criterion="mse", splitter="best", max_features="auto"

As we can see that this model is giving us the accuracy score of 75%.and cross validation score 68%.

As we conclude from model building that Random Forest Regressor model is giving the best score so, we are going to save this model to predict.

CONCLUSION:

This above discussion has presented a supervised rainfall learning model which used machine learning algorithms to classify rainfall data. We used different machine learning algorithm to check the accuracy of rainfall prediction.

We have compared LOGISTIC REGRESSION, RANDOM FOREST, KNEIGHBOUR and DECICISION TREE classifiers.

We have also compared LINEAR REGRESSION, KNEIGHBOUR, BAGGING, RANDOM FOREST and DECISION TREE Regressor.

From the above we can conclude that Random Forest is the Machine learning algorithm which is suitable for rainfall prediction.