**ARTICLE: Rainfall Weather Forecasting**

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**.

DATA ANALYSIS :

1. Number of columns:  There are total **23** columns in the dataset.
2. Date  - The date of observation
3. Location  -The common name of the location of the weather station
4. MinTemp  -The minimum temperature in degrees celsius
5. MaxTemp -The maximum temperature in degrees celsius
6. Rainfall  -The amount of rainfall recorded for the day in mm
7. Evaporation  -The so-called Class A pan evaporation (mm) in the 24 hours to 9am
8. Sunshine  -The number of hours of bright sunshine in the day.
9. WindGustDi r- The direction of the strongest wind gust in the 24 hours to midnight
10. WindGustSpeed -The speed (km/h) of the strongest wind gust in the 24 hours to midnight
11. WindDir9am -Direction of the wind at 9am
12. WindDir3pm -Direction of the wind at 3pm
13. WindSpeed9am -Wind speed (km/hr) averaged over 10 minutes prior to 9am
14. WindSpeed3pm -Wind speed (km/hr) averaged over 10 minutes prior to 3pm
15. Humidity9am -Humidity (percent) at 9am
16. Humidity3pm -Humidity (percent) at 3pm
17. Pressure9am -Atmospheric pressure (hpa) reduced to mean sea level at 9am
18. Pressure3pm -Atmospheric pressure (hpa) reduced to mean sea level at 3pm
19. Cloud9am - Fraction of sky obscured by cloud at 9am.
20. Cloud3pm -Fraction of sky obscured by cloud
21. Temp9am-Temperature (degrees C) at 9am
22. Temp3pm -Temperature (degrees C) at 3pm
23. RainToday -Boolean: 1 if precipitation (mm) in the 24 hours to 9am exceeds 1mm, otherwise 0
24. RainTomorrow -The amount of next day rain in mm. Used to create response variable . A kind of measure of the "risk".

**EXPLORATORY DATA ANALYSIS (EDA):**

**A) Univariate Data Analysis :**

1. High Rainfall is predicted for 'Mellbourne'location then for 'Williamtown' and 'PerthAirport' and lowest is predicted for 'Uluru' then 'Darwin'.

2.'MinTemp' column shows normal curve having range 5 to 20.

3.'MaxTemp' column also shows normal curve having range 20 to 30.

4.'Rainfall' column shows most of the data points are concentrated near zero that means the column contains skewness.

5.'Evaporation' column also shows that most of the values lies near zero and it shows skewness.

6.'Sunshine' column shows a small normal curve with median at 7.5.

7.'WindGustSpeed' columns shows somewhat normal trend but little bit skewness with range 20-60.

8.'WindDir9am' column shows wind speed is mostly in North direction and less in NW and ESE at 9 am.

9.'WindDir3pm' column shows wind direction is mostly SE and less in NNW.

10.'Humidity9am' column shows somewhat skewness with range 40-80.

11.'Humidity3pm' column shows normal curve with most humidity in range 40-80 at 3pm.

12.'Pressure9am' column shows somewhat normal curve.

13.'Pressure3pm' column shows same normal curve as 'Pressure3pm'.

14.Both 'Cloud9am' and 'Cloud3pm' shows almost same curve.

15.Both 'Temp9am' and 'Temp3pm' shows perfect normal curve.

16.'RainToday' column shows mostly rainfall.

17.'RainTomorrow' column shows less rainfall in majority.

*NOTE : We will combine 'WindSpeed9am' and 'WindSpeed3pm','Humidity9am' and 'Humidity3pm', 'Cloud9am' and 'Cloud3pm', 'Temp9am' and 'Temp3pm', 'Pressure9am' and 'Pressure3pm' columns into average of them as they have equal values.*

**B) Bi-Variate Data Analysis:**

1. Avg\_WindSpeed has maximum rainfall for Tomorrow.

2. Avg\_Humidity shows maximum rainfall for RainTomorrow.

3. Avg\_cloud shows maximum rainfall for RainTomorrow.

4. Avg\_Temp shows less rainfall for RainTomorrow.

5. Avg\_Pressure shows same Rainfall for RainTomorrow.

**Pre-processing:**

**Co-relation Matrix:**

* Avg\_Temp and MinTemp shows Highest correlation of 0.82, Avg\_Temp and MaxTemp shows highest correlation of 0.96, also WindGustSpeed and Avg\_WindSpeed shows highest correlation of 0.72.

**Skewness :**

* The columns 'Evaporation','Rainfall','RainToday','WindGustSpeed','Avg\_WindSpeed' have highest skewness and removed it by using powertransformation.

**Outliers :**

* ‘MaxTemp','Evaporation','Sunshine','WindGustSpeed','RainToday','Avg\_WindSpeed','Avg\_Humidity','Avg\_Temp','Avg\_Pressure' columns shows the outliers and removed it by using IQR method.
* After successfully removing the outliers from the dataset, we got the percentage loss of almost 25%.

**MODEL BUILDING & EVALUATION :**

* I have applied Logistic Regression , RandomForest Classifier , Decision Tree Classifier , SVC , XGB Classifier and, Gradient Boosting Classifier for model building and evaluation.
* After doing Cross Validation check for each model ;

1. We choose the model on basis of lowest difference between model accuracy score and cross validation score of that model,we observe that we got less difference and good accuracy for SVC,and performed hyper parameter tunning for it.

**HYPER PARAMETER TUNNING :**

* Best parameter ; {'C': 10, 'gamma': 0.01, 'kernel': 'rbf'}
* GSV.best\_score\_ : 84.91
* Accuracy : 86.77
* AUC score is 100%

**Conclusion :**

* SVC is the best for model building we got the accuracy score as 86% and the AUC score is 100% which is very well good.
* 'RainTomorrow' column shows less rainfall in majority.
* Avg\_Pressure shows same Rainfall for RainTomorrow.