**A PROJECT REPORT**

**ON**

**Rainfall Prediction**



Submitted by

ANKEET SAMBYAL

**ACKNOWLEDGMENT**

I would like to express my sincere gratitude to my mentors from DataTrained for helping me work on this project. Their timely and valuable inputs have helped me in the completion of this project successfully.

Finally, I would like to thank my family who have always been a great support to me and have been very helpful in various stages of project completion. Without their support it would not be possible to pursue my PG course and internship.

**References:**

* <https://www.kaggle.com/datasets/jsphyg/weather-dataset-rattle-package>
* Definitions adapted from

<http://www.bom.gov.au/climate/dwo/IDCJDW0000.shtml>

* Datasource:
* <http://www.bom.gov.au/climate/dwo/> and <http://www.bom.gov.au/climate/data>.

**TABLE OF CONTENTS:**

# Introduction

* + Business Problem Framing
    - Problem Statement
    - Objective
  + Conceptual Background of the Domain Problem
  + Review of literature
  + Motivation for the Problem Undertaken

# Analytical Problem Framing

* + Mathematical/ Analytical Modelling of the Problem
  + Data Sources and their formats
  + Hardware & Software Requirements & Tools Used
  + Data Analysis
  + EDA Concluding Remark
  + Pre-processing Pipeline

# Model/s Development and Evaluation

* + Building Machine Learning Models

# Conclusions

* + Key Findings
  + Concluding Remarks

**INTRODUCTION**

Predicting rainfall is an important step in generating data for Weather forecasting. Rainfall predictions are a key process for providing Weather forecasting assessments with inputs to provide information people and organizations can use to reduce weather-related losses and enhance societal benefits, including protection of life and property, public health and safety, and support of economic prosperity and quality of life. Machine Learning (ML) can be helpful in overcoming such issues; for example, ML can be used to predict rainfall and apply it to foresee crop health and yield. Predictive analysis is a subset of data mining that forecasts future probabilities and patterns. Forecasting could be applied in air traffic, severe weather alerts, marine, agriculture, utility companies, private sector and military application.

**Business Problem Framing:**

Weather forecasting is the application of science and technology to predict the conditions of the atmosphere for a given location and time. Weather forecasts are made by collecting quantitative data about the current state of the atmosphere at a given place and using meteorology to project how the atmosphere will change.

# Problem Definition: Design a predictive model with the use of machine learning algorithms to forecast whether or not it will rain tomorrow in Australia.

**Objective:** To predict two things

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.

**Conceptual Background of the Domain Problem:**

The [weather](https://goweatherforecast.com/) has a great impact on various aspects of human life. So, efforts have been made for years to improve the accuracy of weather forecasting to ensure a better life.

**What is weather forecasting?**

Weather forecasting refers to the process in which science and technology are applied to predict the conditions of the atmosphere for a given location and time. Humans have been making an unofficial weather forecasting attempt millennia ago, and official weather forecasting dates back to the nineteenth century. Weather forecasting is done by collecting data on the current state of the atmosphere and applying a scientific understanding of atmospheric processes to predict atmospheric progression.

Rainfall forecasting has been around for years using traditional methods that employ statistical techniques to assess the correlation between rainfall, geographic coordinates (such as latitude and longitude), and other atmospheric factors (like pressure, temperature, wind speed, and humidity).

**Key Importance of weather forecasting**

The ultimate goal of weather forecasting is to protect human lives and property, improve health, safety, and economic prosperity. Weather Forecasting is crucial since it helps to determine future climate changes. With the use of latitude, we can determine the probability of snow and hail reaching the surface. We are able to identify the thermal energy from the sun that is exposed to a region.

On an everyday basis, many use weather forecasts to determine what to wear on a given day. Since outdoor activities are severely curtailed by heavy rain, snow and wind chill, forecasts can be used to plan activities around these events, and to plan ahead and survive them.

**Review of literature:**.

Rainfall remains one of the most influential meteorological parameters in many aspects of our daily lives. With effects ranging from damage to infrastructure in the event of a flood to disruptions in the transport network, the socio-economic impacts of rainfall are noteworthy. Floods and similar extreme events are consequences of climate change that are expected to occur more frequently and have catastrophic effects in years to come. More interestingly, recent studies have highlighted that weather conditions can potentially increase air pollution (another major topic of discourse alongside climate change in recent times) in winter and summer periods. It is pertinent to reiterate that increased air pollution results in health conditions such as asthma and similar problems related to the lungs. Therefore, as a mitigation approach, many studies have investigated and proposed rainfall forecasting techniques in preparation for any eventuality. However, in order to enhance human mobility activities and enhance agriculture and industrial development, these approaches must provide efficient and timely predictions.

Rainfall Prediction is the application area of data science and machine learning to predict the state of the atmosphere. It is important to predict the rainfall intensity for effective use of water resources and crop production to reduce mortality due to flood and any disease caused by rain.

Rainfall forecasting has gained utmost research relevance in recent times due to its complexities and persistent applications such as flood forecasting and monitoring of pollutant concentration levels, among others. Existing models use complex statistical models that are often too costly, both computationally and budgetary, or are not applied to downstream applications. Therefore, approaches that use [Machine Learning algorithms](https://www.sciencedirect.com/topics/computer-science/machine-learning-algorithm) in conjunction with time-series data are being explored as an alternative to overcome these drawbacks..

# Motivation for the Problem Undertaken:

# Artificial intelligence helps in understanding past weather models and this can make decision-making faster. The predictions of extreme weather phenomenon can be of tremendous value in minimizing the damage caused, and plan ahead to prevent casualties. The AI system can make more accurate short-term predictions, including for critical storms and floods. Climate change is making it harder to anticipate adverse weather conditions, as the frequency and severity of heavy rain increases, which researchers believe will lead to both significant material damage and death. To this end, this study presents a [comparative analysis](https://www.sciencedirect.com/topics/computer-science/comparative-analysis) using simplified rainfall estimation models based on conventional Machine Learning algorithms to forecast whether or not it will rain tomorrow and to predict how much rainfall could be there.

**ANALYTICAL PROBLEM FRAMING**

**Mathematical/ Analytical Modelling of the Problem:**

The main thing that I found in problem statement is, the data given for us is supervised data. The problem statement contains both continuous and categorical data. I have performed both univariate and bivariate analysis to analyze these values using different plots like pie plot, count plot, distribution plot, factor plot etc. These plots give better pattern for analyzing the data. In this project I have done various mathematical and statistical analysis such as describing the statistical summary of the columns in which I found that the count is not same for all the columns which means null values present. Since the dataset contains object data type, I used label encoding method to convert the object data into numerical data. Checked for correlation between the features and visualized it using heat map for both the target variables separately and created models individually for each of the target variables.

**Data Sources and their formats:**

* The Dataset contains about 10 years of daily weather observations of different locations in Australia. The dataset is provided by Datatrained in the project portal which is in the CSV format.
* I have used this CSV for the processing.
* The dataset contains 8425 rows and 23 columns which is comprised of categorical columns. The datset has most of the features in float data type and some in the object data type.
* While describing the data I found skewness and outliers present in some of the columns. Since some of these columns are categorical they need not be removed.

# Hardware & Software Requirements & Tools Used:

### **Hardware required:**

* Processor: core i5 or above
* RAM: 8 GB or above
* ROM/SSD: 250 GB or above

### **Software required:**

### Anaconda 3- language used

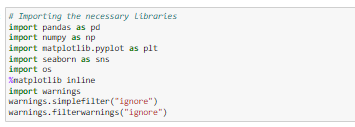
### Python 3

# Data Analysis

* Firstly, I have uploaded the dataset in the Jupyter notebook.
* Then I have imported the necessary libraries and dataset.

### **Libraries:**

The important libraries that I have used for this project are below.



### **import numpy as np:**

It is defined as a Python package used for performing the various numerical computations and processing of the multidimensional and single dimensional array elements. The calculations using Numpy arrays are faster than the normal Python array.

### **import pandas as pd:**

Pandas is a Python library that is used for faster data analysis, data cleaning and data pre-processing. The data-frame term is coming from Pandas only.

### **import matplotlib.pyplot as plt and import seaborn as sns:**

Matplotlib and Seaborn acts as the backbone of data visualization through Python.

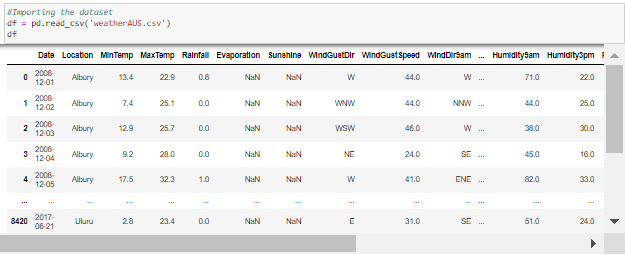
**Matplotlib**: It is a Python library used for plotting graphs with the help of other libraries like Numpy and Pandas. It is a powerful tool for visualizing data in Python. It is used for creating statistical interferences and plotting 2D graphs of arrays.

**Seaborn**: It is also a Python library used for plotting graphs with the help of Matplotlib, Pandas, and Numpy. It is built on the roof of Matplotlib and is considered as a superset of the Matplotlib library. It helps in visualizing univariate and bivariate data.

* **import warnings**

The warnings module was introduced in PEP 230 as a way to warn programmers about changes in language or library features in anticipation of backwards incompatible changes coming with Python 3.0. Since warnings are not fatal, a program may encounter the same warn-able situation many times in the course of running. Using “ignore” never displays warnings which match

* I have imported the given dataset from jupyter notebook as follows



* Checked the dimension of the dataset.



The dataset has 8425 rows and 23 columns

* Checked the columns in the dataset



**Dataset Description:**

**`Date`** - The date of observation

**`Location`** -The common name of the location of the weather station

**`MinTemp**` -The minimum temperature in degrees celsius

**`MaxTemp`** -The maximum temperature in degrees celsius

**`Rainfall`** -The amount of rainfall recorded for the day in mm

**`Evaporation`** -The so-called Class A pan evaporation (mm) in the 24 hours to 9am

**`Sunshine`** -The number of hours of bright sunshine in the day.

**`WindGustDir**`- The direction of the strongest wind gust in the 24 hours to midnight

**`WindGustSpeed`** -The speed (km/h) of the strongest wind gust in the 24 hours to midnight

**`WindDir9am`** -Direction of the wind at 9am

**`WindDir3pm`** -Direction of the wind at 3pm

**`WindSpeed9am`** -Wind speed (km/hr) averaged over 10 minutes prior to 9am

**`WindSpeed3pm`** -Wind speed (km/hr) averaged over 10 minutes prior to 3pm

**`Humidity9am`** -Humidity (percent) at 9am

**`Humidity3pm`** -Humidity (percent) at 3pm

**`Pressure9am`** -Atmospheric pressure (hpa) reduced to mean sea level at 9am

**`Pressure3pm`** -Atmospheric pressure (hpa) reduced to mean sea level at 3pm

**`Cloud9am` -** Fraction of sky obscured by cloud at 9am.

**`Cloud3pm`** -Fraction of sky obscured by cloud

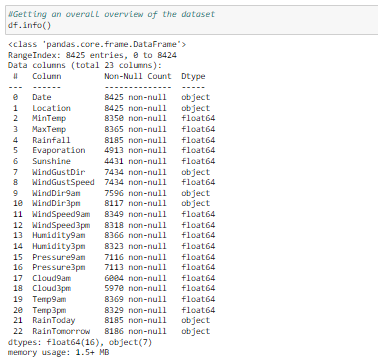
**`Temp9am`** -Temperature (degrees C) at 9am

**`Temp3pm`** -Temperature (degrees C) at 3pm

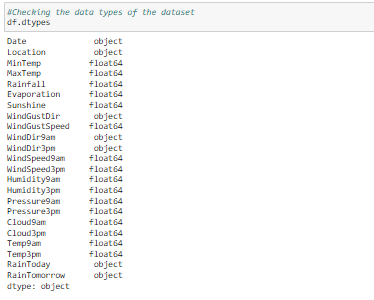
**`RainToday`** -Boolean: 1 if precipitation (mm) in the 24 hours to 9am exceeds 1mm, otherwise 0

**`RainTomorrow`** -The amount of next day rain in mm. Used to create response variable . A kind of measure of the "risk".

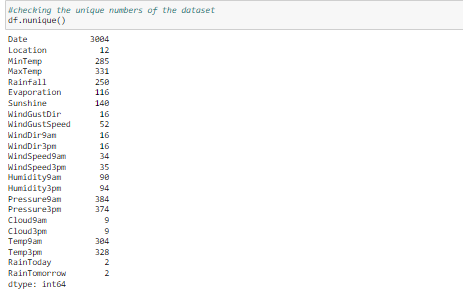
* Checked the information of the data frame using info()



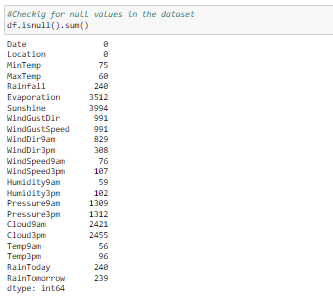
* Checked the data types of the features.



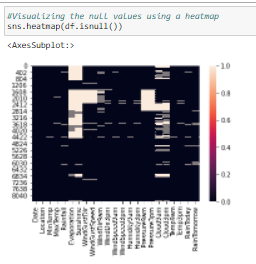
* Checked the number of unique values present in the dataset.



* I have checked for null values and found null values present in the dataset



* I have visualized it using heat map as below



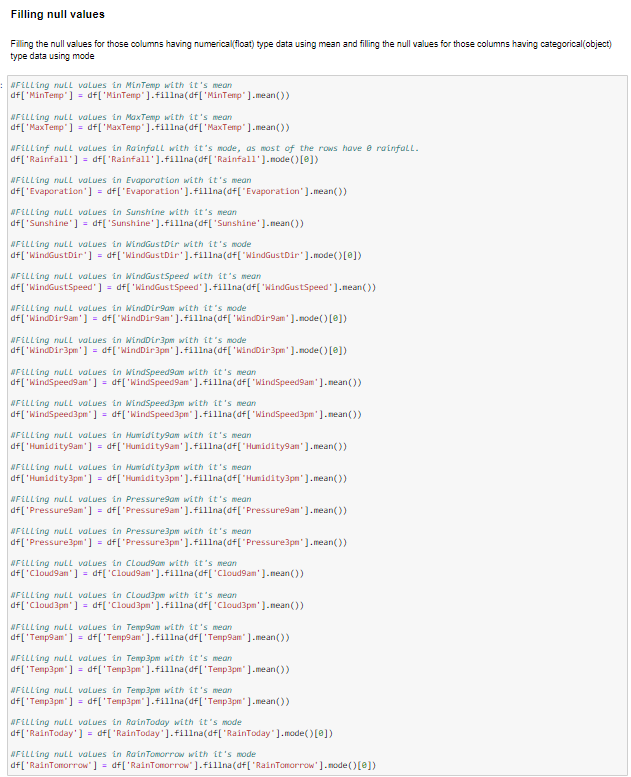
# EDA Concluding Remark

## Observations

* We can see that our dataset comprises of 8425 rows and 23 columns.
* By observing the dataset we can say that we have two target variables.
* The values present in our target column 'RainTomorrow' have data in categorical data type. So we can use Classification model in this case. While the other target variable 'Rainfall' is in continuous data making it a regression problem.
* EDA and preprocessing is same for prediction of both the targets.
* In the above output we can see that there are missing values in many columns and few even have approximately 50% of the data empty. Such columns need to be removed since it provides no insights.
* Though we have the following columns with nearly 50% of missing values... they are important features. Hence we cannot drop them
  + Sunshine has 3994 null values which says 47.4% of the data is missing
  + Evaporation has 3512 null values which says 41.6% of the data is missing
* There are null values in the target columns too.
* All the null values need to be treated using respective imputation methods

# Pre-processing Pipeline

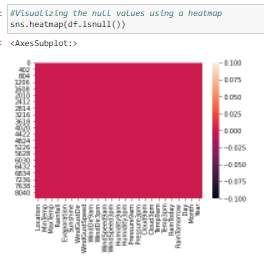
* I have used Imputation methods to fill those null values.



* Next I have converted the date feature to datetime format.
* Extracted the date, month and year from the date feature to individual columns.
* Dropped the Date feature as it is redundant.

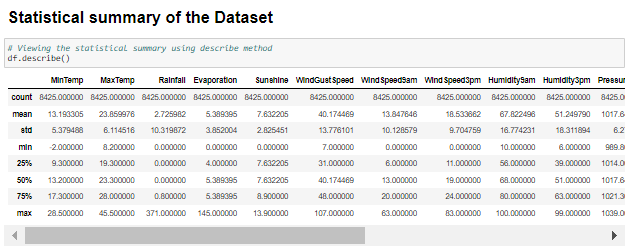


* Checked the null values and plotted the same using a heatmap to confirm there are no null values in the dataset.



This shows there are no more null values in the dataset.

* Described the data using describe() to view the statistical summary of the dataset.



**Observations :**

The describe method shows statistical summary of the numerical columns only, it doesnot displays the categorical data. From the above table we can say that

- The count of all the columns is uniform which says there are no null values in the dataset.

- The Rainfall columns has mean value higher than the median (50% - 2nd quantile), so the distribution is skewed to right.

- Other columns have mean almost equal to median, which says the distribution of curve is normal

- The following columns have huge difference between the 75% (3rd quantile) and the max values, which shows the chance for presence of outliers

- MinTemp

- MaxTemp

- Rainfall

- Evaporation

- Sunshine

- WindGustSpeed

- WindSpeed9am

- WindSpeed3pm

- Humidity9am

- Humidity3pm

- Temp9am

- Temp3pm

- The following columns have no much difference between the 75% (3rd quantile) and the max values, which shows there are less chance for presence of outliers

- Pressure9am

- Pressure3pm

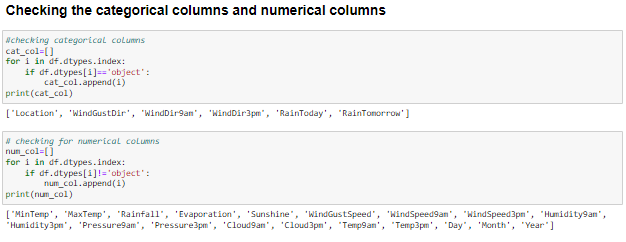
- Cloud9am

- Cloud3pm

* Checked the value count of each column. By using for loop value count function.



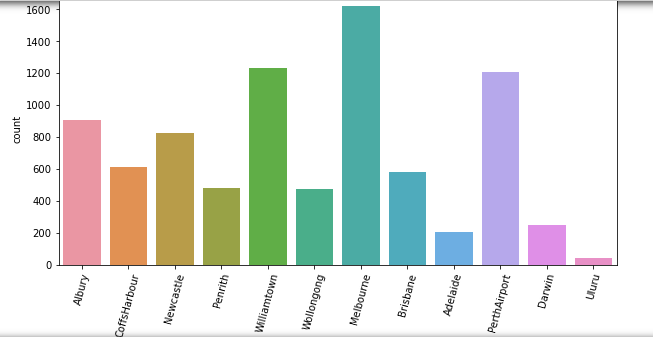
* Checked for the features with object type data and float type data using for loop and stored them in individual lists for better and easy plotting of graphs.



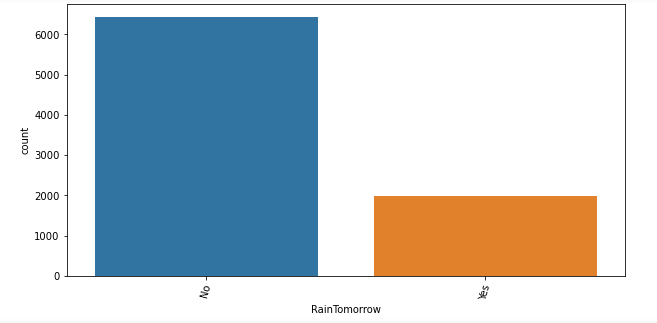
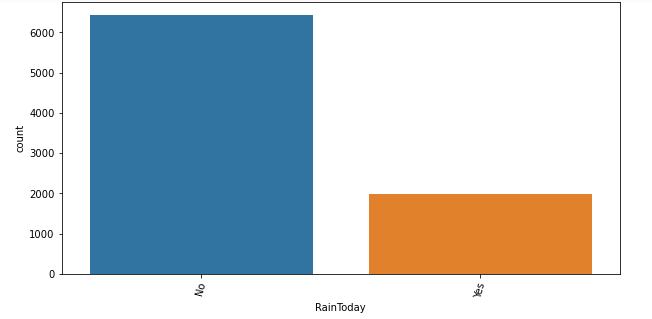
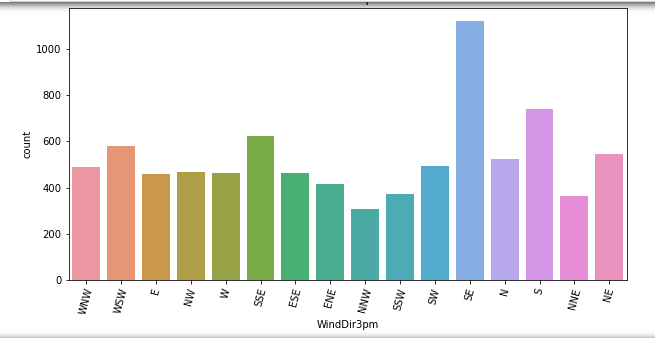
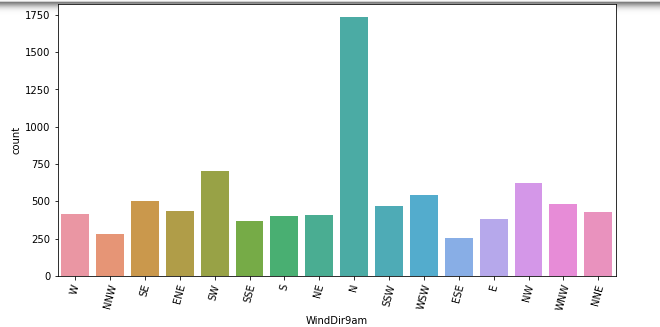
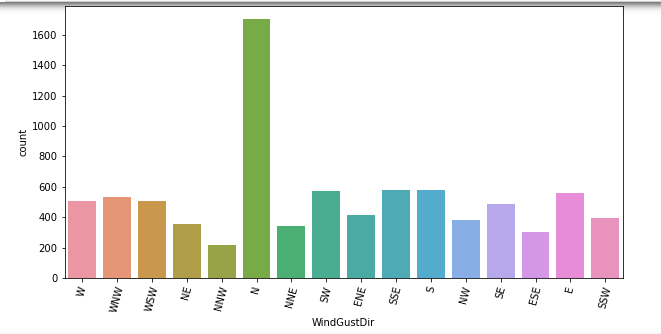
* Visualized each feature using seaborn and matplotlib libraries by plotting count plot, distribution plot, Scatterplot, bar plot and factor plot.
* Univariate Analysis

### **Visualizing the various categorical columns present in the dataset**

* **Plotting count plots of the categorical columns**



Location

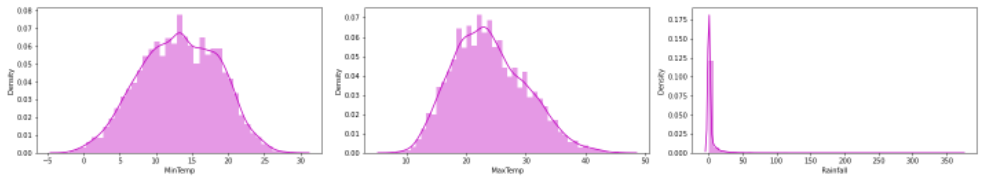


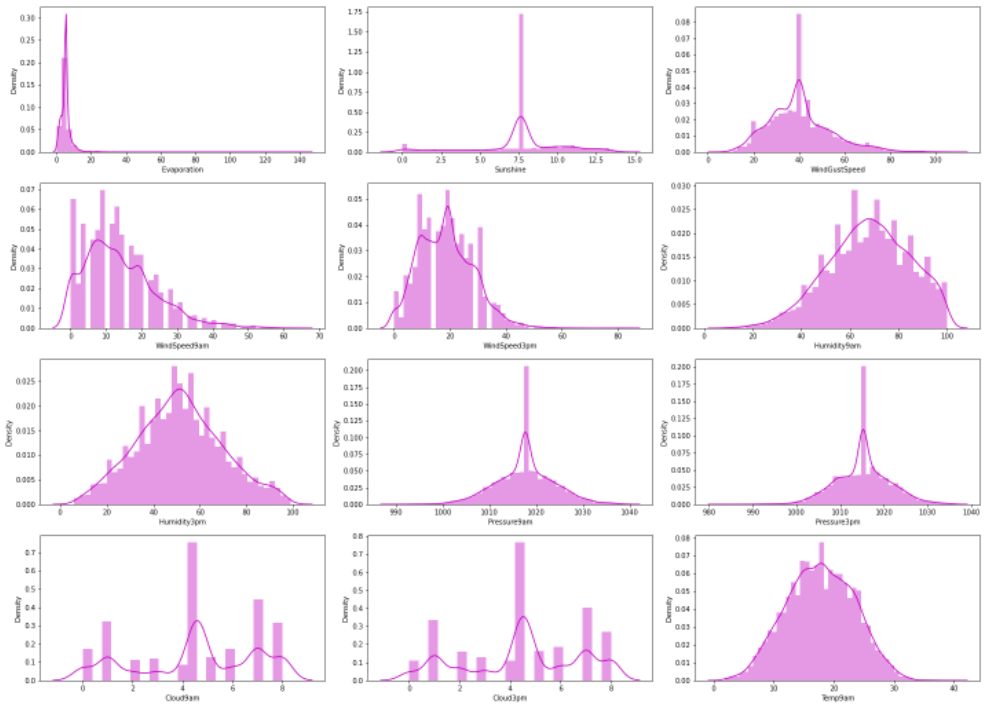
# Observations

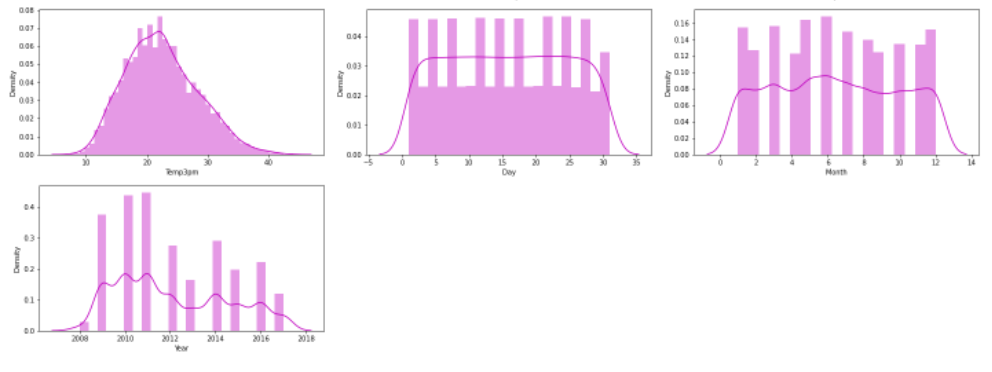
* Melbourne has highest count in the Location column, where as Uluru has the least count
* N has the highest count in the WindGustDir, where as the NNW has the least count
* N has the highest count in the WindDir9am, where as NNW has the least count
* SE has the highest count in the WindDir3pm, where as the NNW has the least count
* No RainToday has high count
* No RainTomorrow has high count

## **Visualizing the distribution of the numerical columns**

## **Plotting distplot for all the numerical columns**







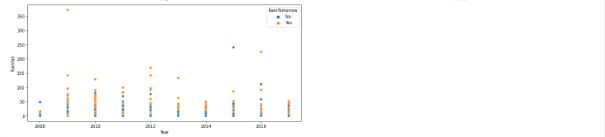
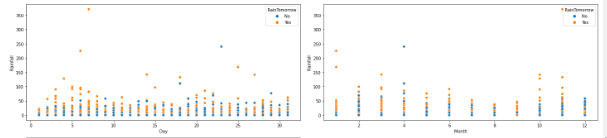
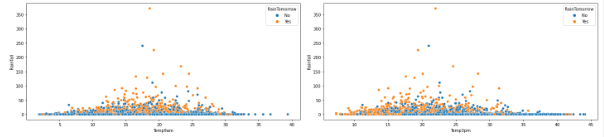
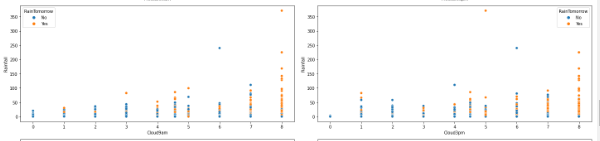
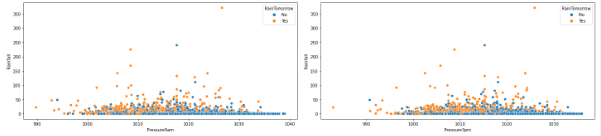
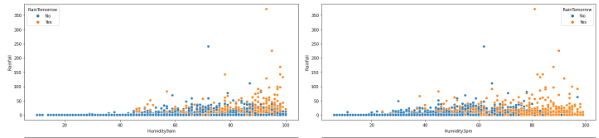
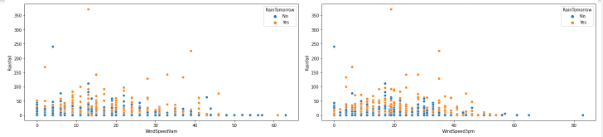
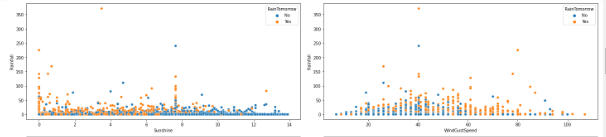
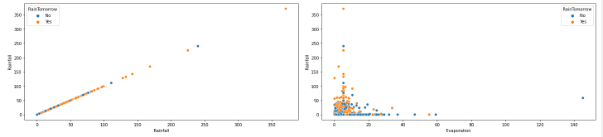
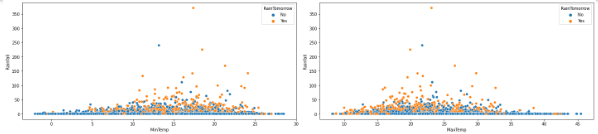
# Observations

* We can see most of the features have normal distribution
* Some of the features are skewed to the right and left as expected , this confirms the presence of the skewness.

## **Bivariate Analysis**

## **Visualizing the relation between the features and the two target variables**

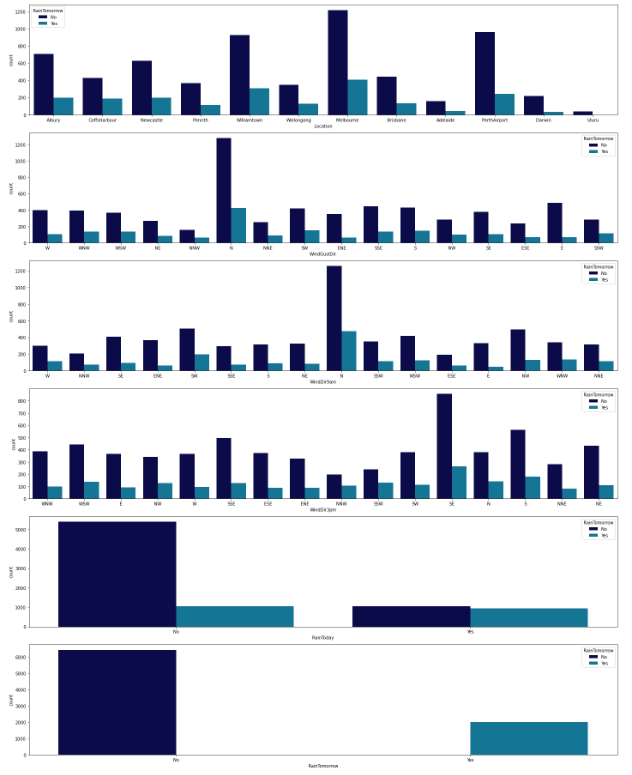
## **Plotting Scatterplot for all the numerical columns with X as the feature , Y as ‘Rainfall’ and hue as ‘RainTomorrow’**



# Observations

From the above plots we can say that

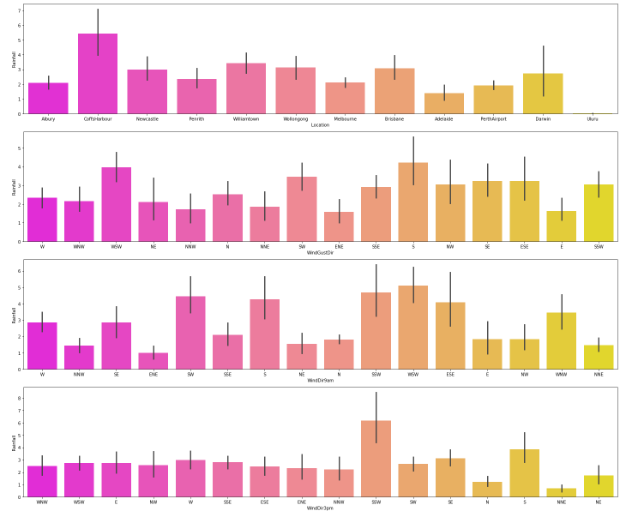
* As the MinTemp is between 10 and 20 it shows there is lower level of rainfall with high chance of no rainfall tomorrow
* As the MaxTemp is between 10 and 35 there is a lower level of rainfall with high chance of rainfall tomorrow
* Most of the rainfall is lower level and is densely populated at the lower level of evaporation. The more the evaporation, more the chance of rain tomorrow
* As the sunshine is lower there is more chance of rainfall and rain tomorrow
* When the wind speed is between 20 and 60 there is a chance of lower level rainfall with equal chance of rain tomorrow
* When the windspeed9am is between 5 and 30 there is a rainfall of lower level with high chance of rain tomorrow
* When the windspeed3pm is between 5 and 45 there is a rainfall of higher level with high chance of rain tomorrow
* When humidity9am is between 45 and 100 there is rainfall with higher chance of rain tomorrow increases as the humidity increases
* When humidity3pm is between 30 and 100 there is rainfall with higher chance of rain tomorrow increases as the humidity increases
* When the Pressure9am is between 1000 and 1030 there is rainfall with higher chance of no rain tomorrow as the pressure increases
* When the Pressure3pm is between 1000 and 1020 there is rainfall with higher chance of no rain tomorrow as the pressure increases
* If Cloud9am is higher there is higher rain fall and higher chance of rain tomorrow
* If Cloud3pm is higher there is higher rain fall and higher chance of rain tomorrow
* When Temp9am is between 10 and 30 there is a rainfall and higher chance of no rain tomorrow
* When Temp3pm is between 10 and 30 there is a rainfall and higher chance of no rain tomorrow
* There is more rainfall in the first 10 days of the month and there is high chance of rain tomorrow compared to other days
* There is high rain fall in the months of 2,3,4, and 10, 11 with high chance of rain tomorrow compared to other months
* The year 2012 has higher rainfall with higher chances of rain tomorrow than other years
* **Plotting Countplot for all the categorical columns with X as the feature and hue as ‘RainTomorrow’**

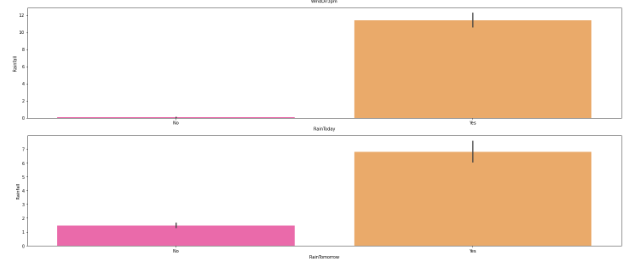


# Observations

From the above plots we can say that

* Melbourne has high rainfall with lower chance of rain tomorrow
* Uluru has lower rainfall and there is no chance for rain tomorrow
* N has high rainfall in WindGustDir with higher chance of no rain tomorrow
* In WindGustDir NNW has lower rainfall with lowest chance of rain tomorrow
* N has high rain fall in WindDir9am with higher chance of no rain tomorrow
* SE has high rain fall in WindDir3pm with higher chance of no rain tomorrow
* There is high chance for no rain today with no rain today and no rain tomorrow
* **Plotting barplot for all the categorical columns with X as the feature and Y as ‘Rainfall’**

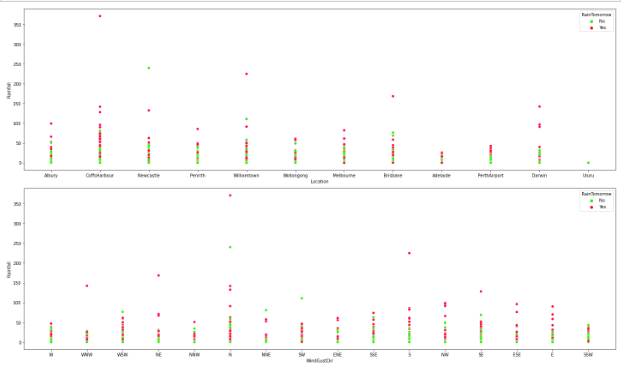


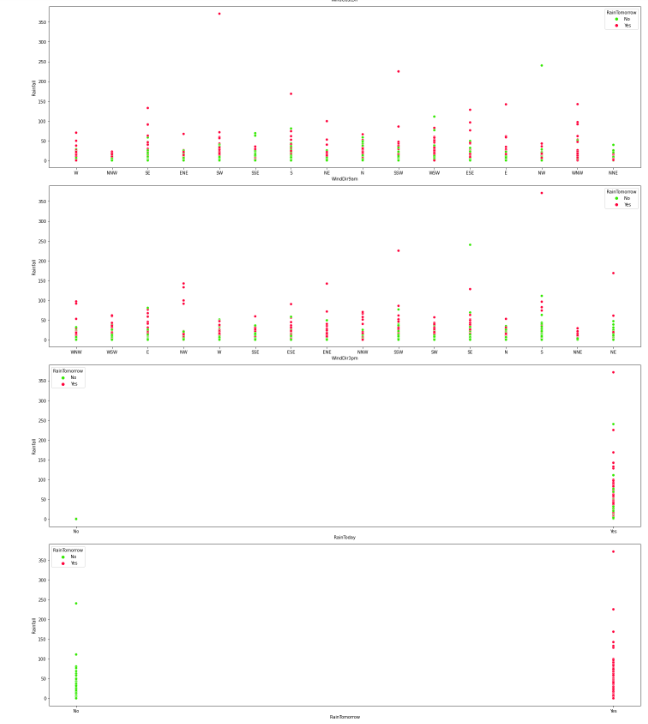
****

# Observations

From the above plots we can say that

* CoffsHarbour has higher rainfall and uluru has lowest rainfall compared to others
* S has higher rainfall and E has lowestrainfall in WindGustDir
* Rain fall is highest in the WSW of WindDir9am and ENE has lowest rainfall
* Rain fall is highest in the SSW of WindDir3pm and NNE has lowest rainfall
* RainToday has higher chance of rainfall
* Rain tomorrow has higher chance of rainfall
* **Plotting Scatterplot for all the categorical columns with X as the feature and Y as ‘Rainfall’ and hue as ‘RainTomorrow’**

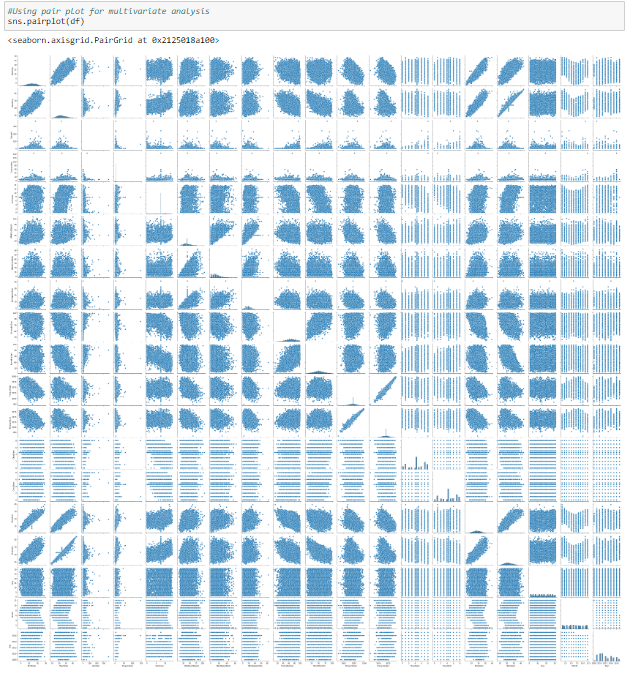
****

****

# Observations

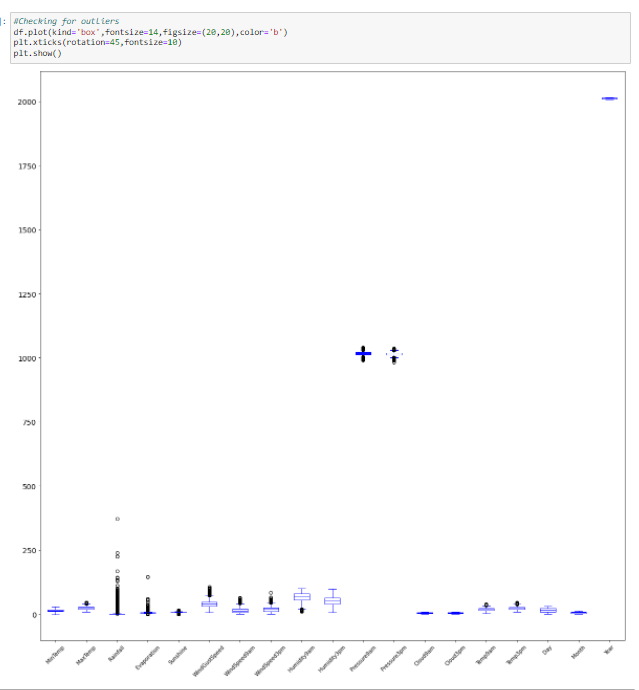
From the above plots we can say that

* CoffsHarbour has higher rainfall with higher chance of rain tomorrow and uluru has lowest rainfall with high chance of rain tomorrow compared to others
* N has higher rainfall with higher chance of higher chance of rain tomorrow and E has lowestrainfall in WindGustDir
* Rain fall is highest in the WSW in WindDir9am with higher chance of rain tomorrow and ENE has lowest rainfall
* Rain fall is highest in the SSW in WindDir3pm with higher chance of rain tomorrow and NNE has lowest rainfall with high chance of rain tomorrow
* RainToday has higher chance of rainfall
* Rain tomorrow has higher chance of rainfall
* **Performed Multivariate analysis using pairplot method in seaborn.**



# Observations

* Most of the pair plots are linear and outliers are observed in some of the features.
* Performed Data Cleaning for removing outliers.



* Plotted Box plot of all the features to check for the outliers present in the dataset. I found out that there were outliers in all the columns except for the following columns

- MinTemp,

- Humidity3pm,

- Cloud9am,

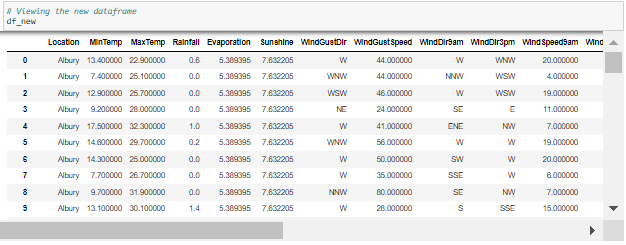
- Cloud3pm,

- Day,

- Month,

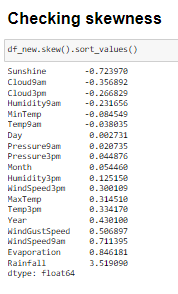
- Year

* After identifying the outliers I have added all the columns to a list and removed the outliers from them using ZScore method and IQR method.
* The ZScore method has a dataloss of 5.19% which is less than 10% and IQR method has a dataloss of 51%, which is too high. The ZScore method has showed less data loss than the IQR method so carried on with the dataframe created after removal of the outliers using ZScore method.



This is the new dataframe after removing the outliers using ZScore

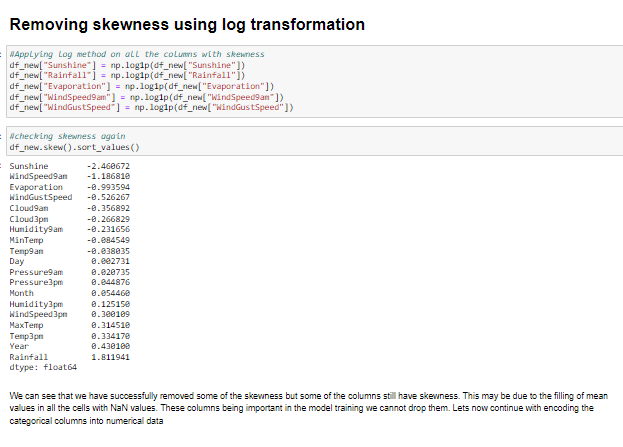
* Checked the skewness using skew() option.



# Observations

We can see that the following features have skewness

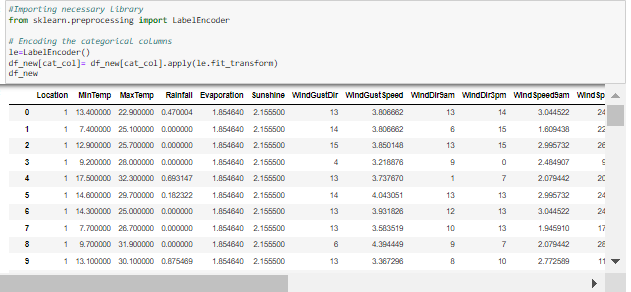
* Sunshine
* Rainfall
* Evaporation
* Windspeed9am
* Windgustspeed .
* I have removed skewness using log transformation method as follows and Checked for the skewness again



* Some of the columns having high skewness could not be dropped due its importance in the dataset. This skewness may also be a result of a major chunk of the values in the columns being Null values and filling all of them using the same mean median values.

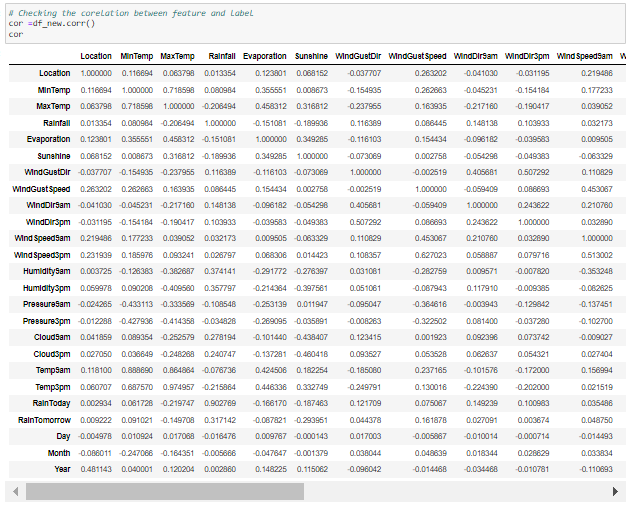
# Label Encoding

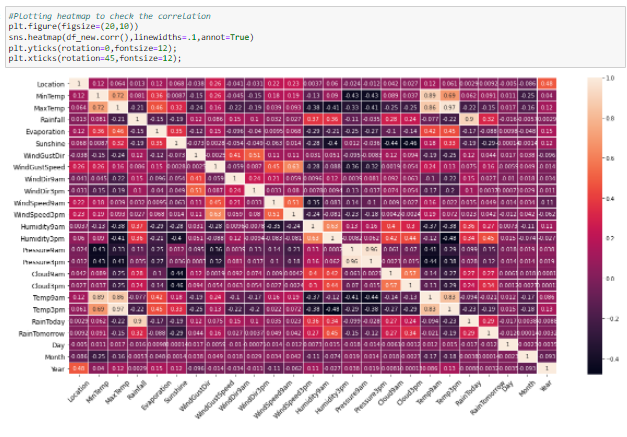
* Performed label encoding to convert the categorical data into the numerical form for the better convenience in model training.



# Correlation

* Checked the correlation between the features and visualized it using heat map.





**Observations**

- MaxTemp and MinTemp are higly correlated with Temp9am and Temp3pm

- MaxTemp and MinTemp are higly correlated with each other.

- MaxTemp and MinTemp are higly negatively correlated with Pressure9am, Pressure3pm

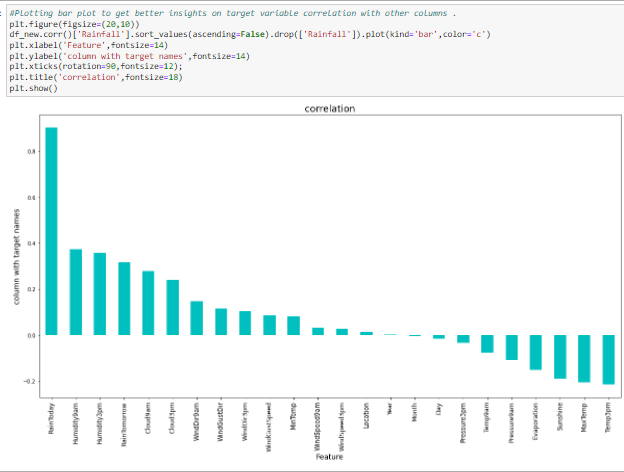
- MaxTemp is also higly negatively correlated with Humidity9am, Humidity3pm

- WindDir3pm is highly negatively correlated with the Humidity9am

- WindSpeed3pm is highly negatively correlated with Temp9am and Temp3pm

This may lead to multicolinearity issue which needs to be addressed in the later part

* Plotted bar plot to get better insights on correlation of both target variables with other columns individually.



**Observations**

- The target 'Rainfall' is highly positively correlated with the feature 'RainToday'

- The target 'Rainfall' is positively well correlated with the following features

- Humidity9am

- Humidity3pm

- RainTomorrow

- Cloud9am

- Cloud3pm

- The target 'Rainfall' is least positively correlated with the following features

- WindDir9am

- WindGustDir

- WindDir3pm

- WindGustspeed

- MinTemp

- WindSpeed9am

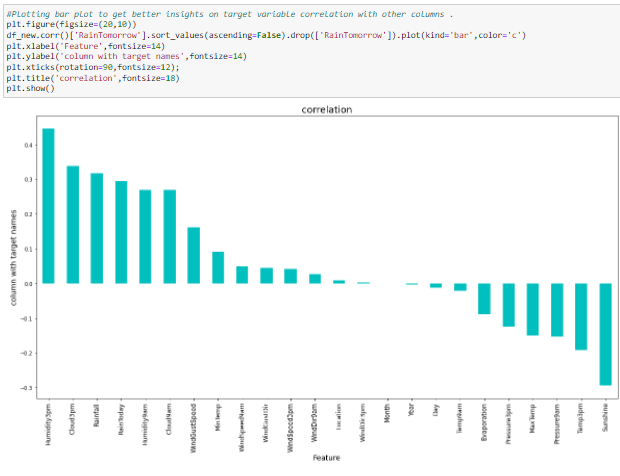
- WindSpeed3pm

- Location

- The features 'Year','Month' has least correlation with the target 'Rainfall', which says 'Year' and 'Month' has least impact on 'Rainfall'. Hence they can be dropped

- The target 'Rainfall' is most negatively correlated with the feature 'Temp3pm'

- The target 'Rainfall' is negatively correlated with the remaining features



**Observations**

- The target 'RainTomorrow' is positively well correlated with the following features

- Humidity3pm

- Cloud3pm

- Rainfall

- RainToday

- Humidity9am

- Cloud9am

- The target 'RainTomorrow' is least positively correlated with the following features

- WindGustspeed

- MinTemp

- WindSpeed9am

- WindGustDir

- WindSpeed3pm

- WindDir9am

- Location

- WindDir3pm

- The feature 'Month' has least correlation with the target 'RainTomorrow', which says 'Month' has least impact on 'RainTomorrow'. Hence it can be dropped.

- The target 'RainTomorrow' is highly negatively correlated with the feature 'Sunshine

- The target 'RainTomorrow' is negatively correlated with the remaining features

# Model/s Development and Evaluation

# Building Machine Learning Models

# 1. Model training for target 'RainTomorrow'

# I have dropped the column ‘Month’ from the dataframe as it is least correlated with the target ‘RainTomorrow’.

# 

# Then I split the features and label into X and Y

# 

# SCALING - I have scaled the data using standard scaler

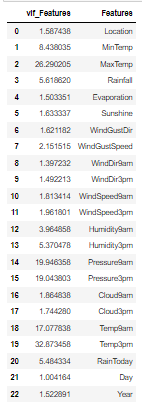
# 

# The data has been scaled using Standard Scaler

# Checking VIF (Variance Inflation Factor)

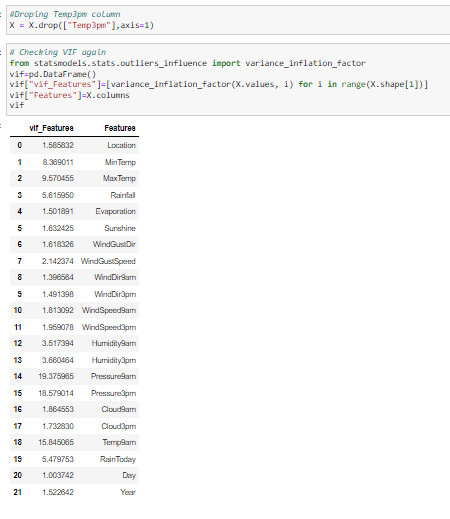
# I have checked the VIF for all the features to know about the multicollinearity.



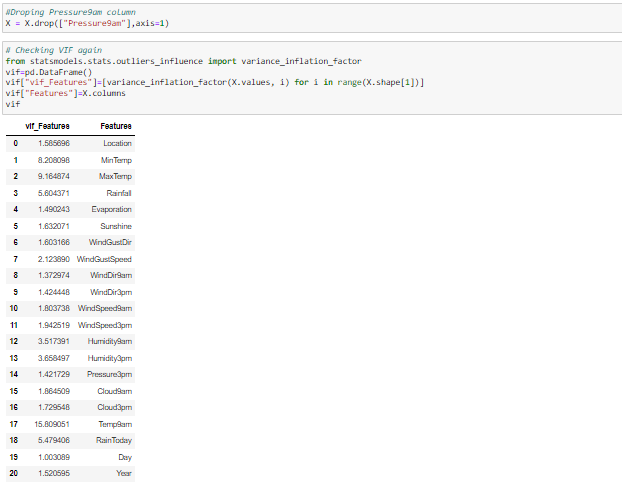


Since there can be seen high VIF for Total volume, the column shall be dropped and check for VIF again.

* I have dropped the ‘Temp3pm’ feature and checked the VIF again

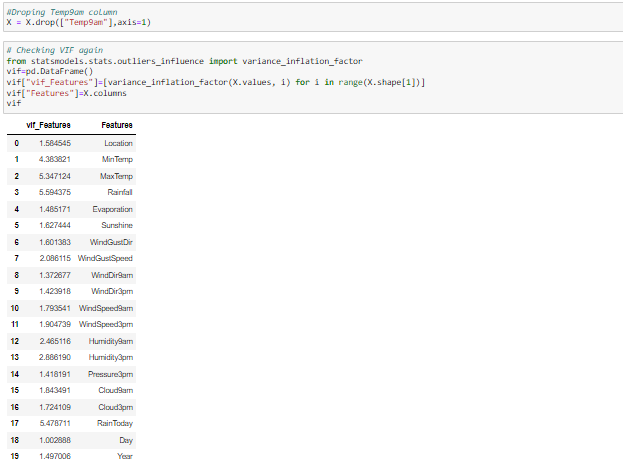


* After dropping ‘Temp3pm’ ‘Pressure9am’ can be seen having high VIF, hence dropping that column and checked the VIF again.



After dropping Pressure9am, Temp9am can be seen having high VIF,hence dropping that column.

* Dropped the column ‘Temp9am’ and check the VIF again

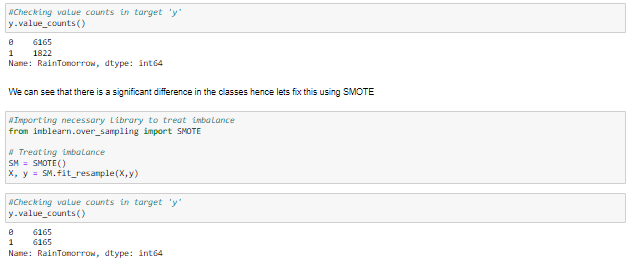


After dropping Temp9am, Rainfall can be seen having high VIF. But the Rainfall is well positively correlated with target 'RainTomorrow'. So we shall not drop this column

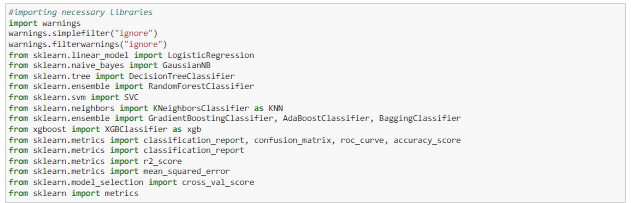
* The multicolinearity has been treated to some extent, But some features still have VIF values above 4. But dropping too many columns have a negative impact on the model. So letting them be for now and moving forward with next step.

# Oversampling

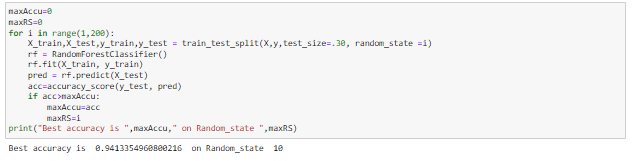
* I have checked for the value counts in the target ‘y’ to fix the imbalance in the data using SMOTE Oversampling technique and checked for the value counts again.



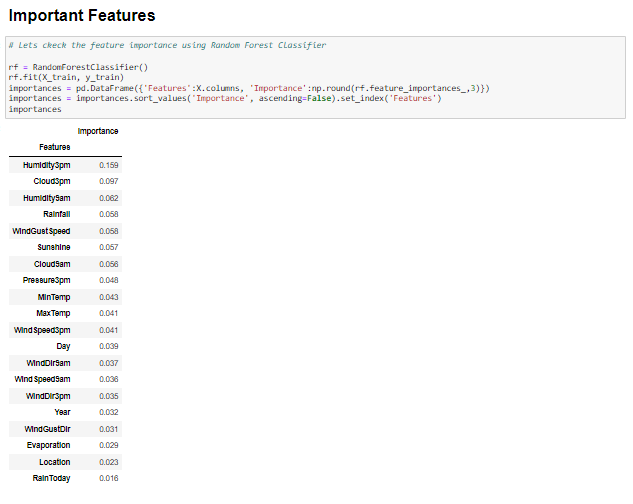
* Now we can see that the imbalance issue has been fixed. Proceeding with model development
* Here the target has classes as output data. Hence this is a classification problem. We have to use the classification algorithms in machine learning for training and testing.
* Importing the necessary libraries and machine learning algorithms for model training and testing



* I have used necessary steps for finding the random state and Accuracy as follows



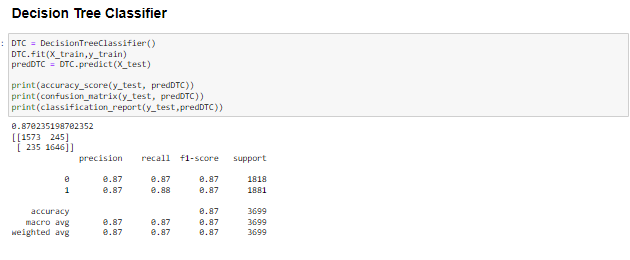
* I have checked for the feature importance using random forest classifier as follows



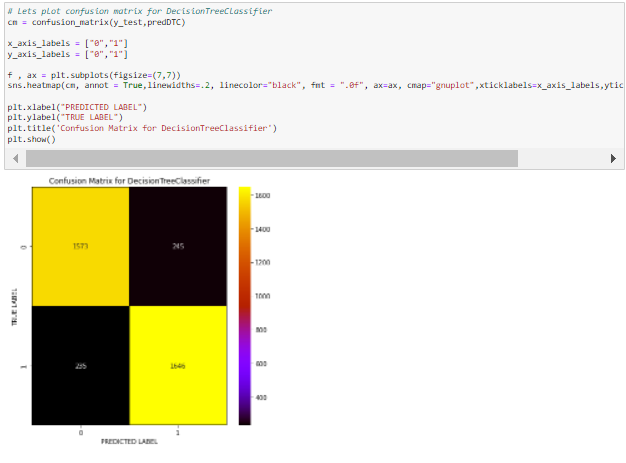
* Then I have created the train-test split as follows taking test size as 30% , which means 70% of the data is taken as the train data . The random state used here is the random state we have got the best accuracy which we previously calculated



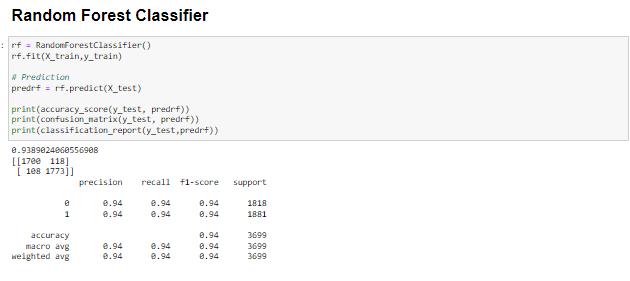
* Then I have checked the accuracy of the model using Decision Tree classifier



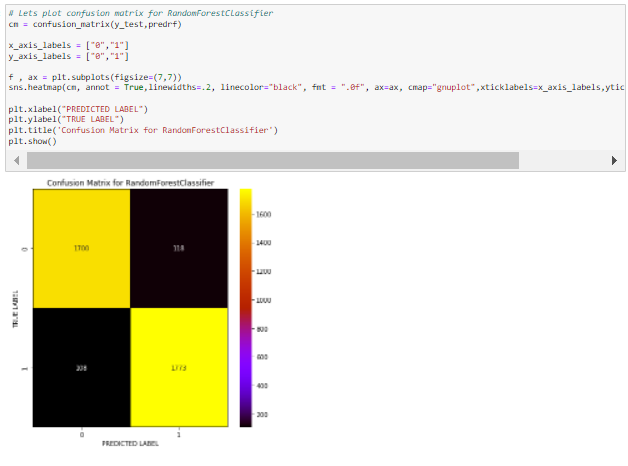
* We can observe that DTC has an accuracy of 88%
* Then plotted the confusion matrix for the same as follows



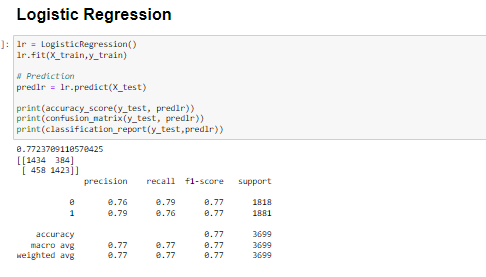
* Then I have checked the accuracy of the model using Random forest classifier



* We can observe that RFC has an accuracy of 95%
* Then plotted the confusion matrix for the same as follows



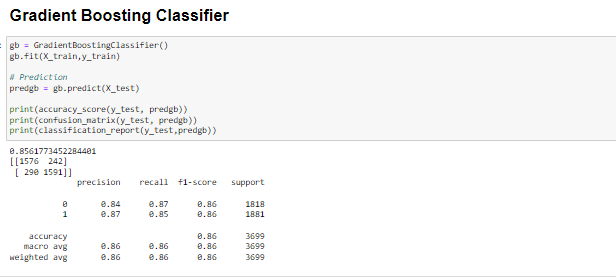
* Then I have checked the accuracy of the model using Logistic Regression



* We can observe that the above model has an accuracy of 77%
* Then plotted the confusion matrix for the same as follows



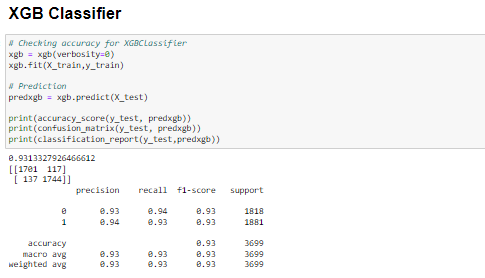
* Then I have checked the accuracy of the model using Gradient boosting classifier



* We can observe that GBC has an accuracy of 85.6%
* Then plotted the confusion matrix for the same as follows



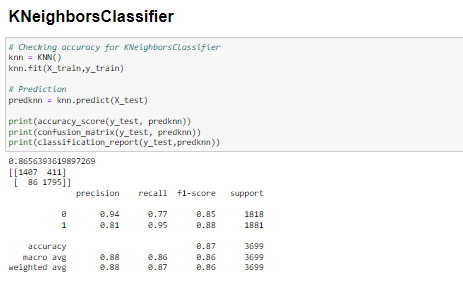
* Then I have checked the accuracy of the model using XGradient boosting classifier



* We can observe that XGB has an accuracy of 93.1%
* Then plotted the confusion matrix for the same as follows



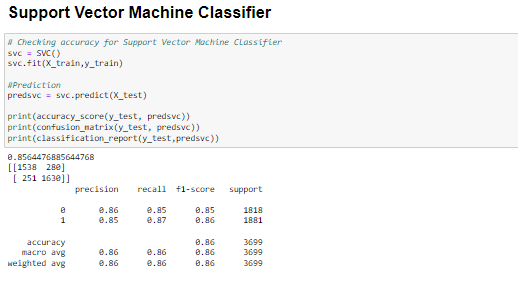
* Then I have checked the accuracy of the model using K Nearest Neighbors classifier



* We can observe that KNN has an accuracy of 87%
* Then plotted the confusion matrix for the same as follows



* Then I have checked the accuracy of the model using Support Vector Machine classifier



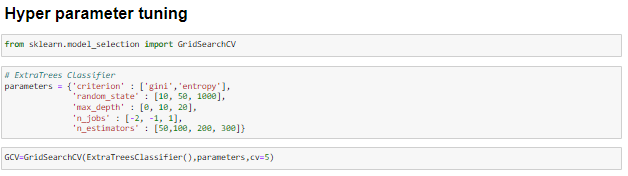
* We can observe that SVC has an accuracy of 87%
* Then plotted the confusion matrix for the same as follows



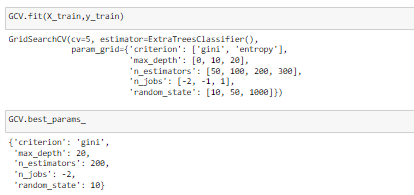
* Then I have checked for the Cross validation scores for all the above models as follows



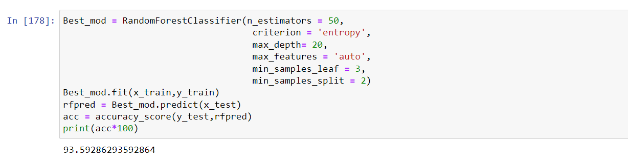
* From the above we can observe the difference between the accuracy score and cross validation score is least in SVC and Gradient Boosting classifier but Random Forest classifier has higher accuracy than other classifiers. Hence we can predict that RFC is the best model.
* Then I have tuned the selected best model using Hyper parameter tuning for better performance as follows



* I have tuned the Random Forest classifier and fit the current model for better performance as follows



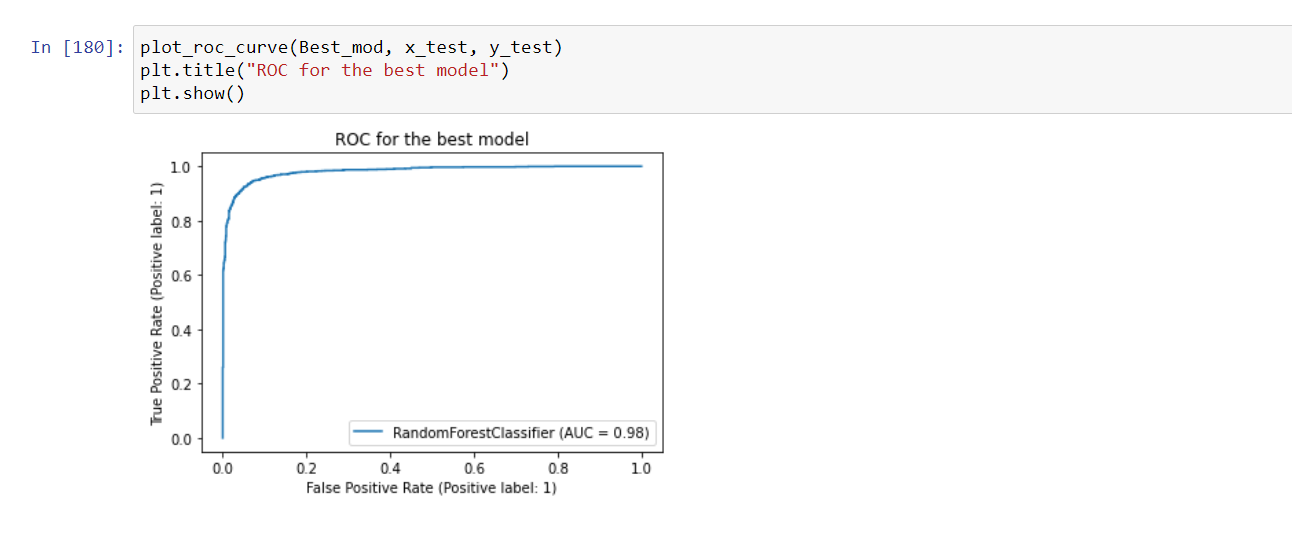
* This has resulted in the improvement of accuracy as 93.59%, which is an expected level of accuracy



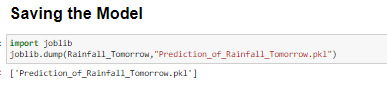
* Then plotted the confusion matrix for the same as follows



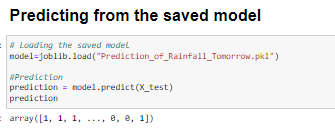
* Then I have plotted the AUC ROC curve for the selected best model as follows



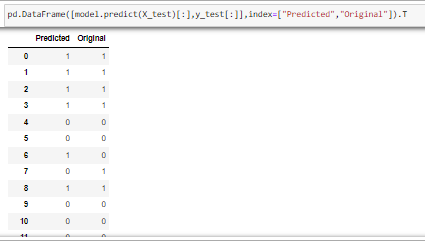
* The predicted model has 98% AUC which is approximately ideal
* Then I have saved the model importing and using joblib as follows



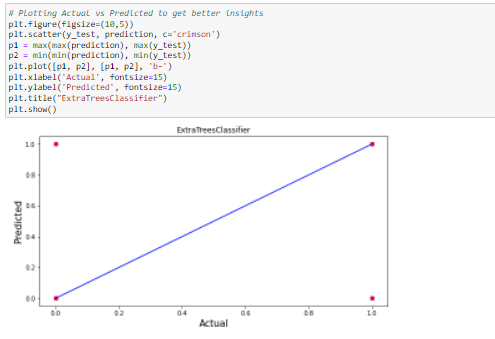
* Predicted using the saved model as follows



* I have created a dataframe of the actual vs predicted values for better view and comparison of outputs from the predicted model



* Then I have plotted Actual vs Predicted to get better insights as follows



* This shows that all the actual values look similar to predicted which is the expected outcome of the model training using machine learning

# 2. Model training for target 'Rainfall'

# I have dropped the column ‘Month’ and ‘Year’ from the dataframe as they are least correlated with the target ‘Rainfall’.

# 

# Then I split the features and label into X and Y

# 

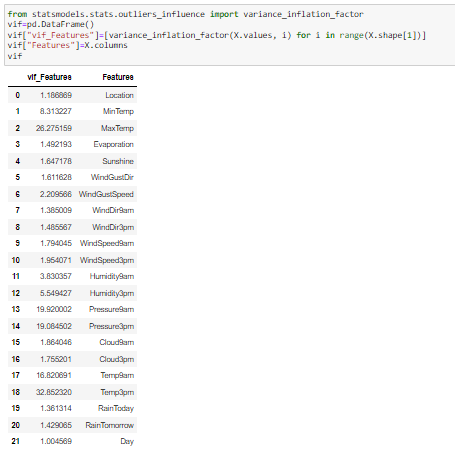
# SCALING - I have scaled the data using standard scaler

# 

# The data has been scaled using Standard Scaler

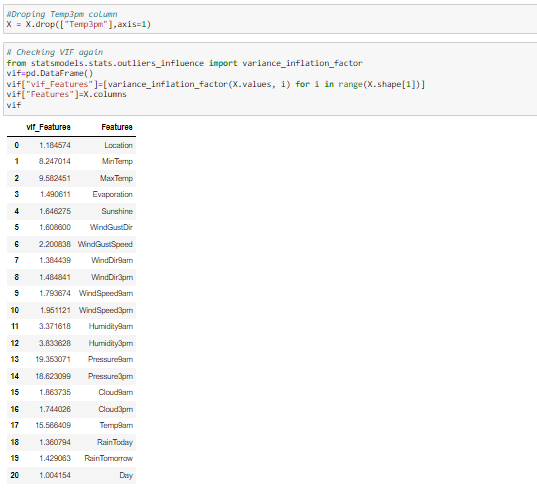
# Checking VIF (Variance Inflation Factor)

# I have checked the VIF for all the features to know about the multicollinearity.

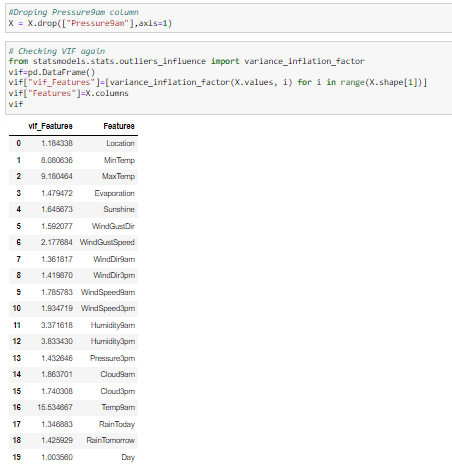
****

Since there can be seen high VIF for Total volume, the column shall be dropped and check for VIF again.

* I have dropped the ‘Temp3pm’ feature and checked the VIF again

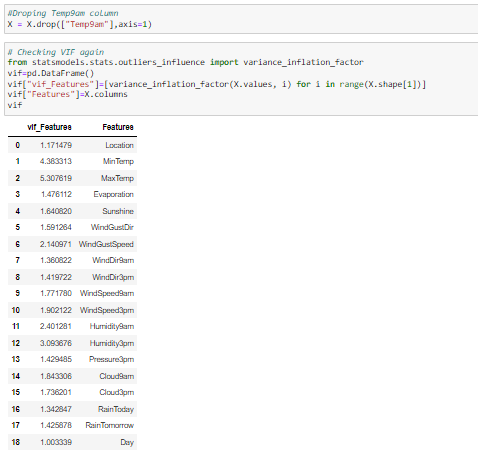


* After dropping ‘Temp3pm’ ‘Pressure9am’ can be seen having high VIF, hence dropping that column and checked the VIF again.



After dropping Pressure9am, Temp9am can be seen having high VIF,hence dropping that column.

* Dropped the column ‘Temp9am’ and check the VIF again



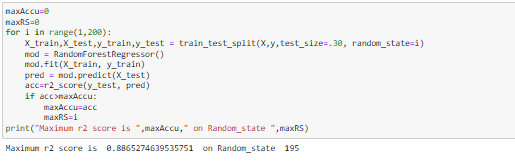
After dropping Temp9am, Rainfall can be seen having high VIF. But the Rainfall is well positively correlated with target 'RainTomorrow'. So we shall not drop this column

The multicolinearity has been treated to some extent, But some features still have VIF values above 4. But dropping too many columns have a negative impact on the model. So letting them be for now and moving forward with next step.

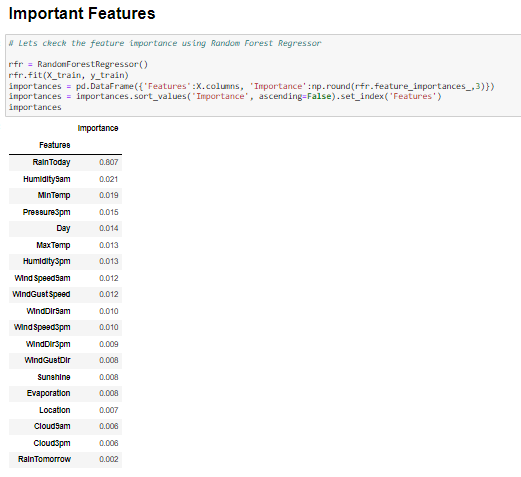
* Here the target has continuous data. Hence this is a Regression problem. We have to use the regression algorithms in machine learning for training and testing.
* Importing the necessary libraries and machine learning algorithms for model training and testing



* I have used necessary steps for finding the random state and Accuracy as follows



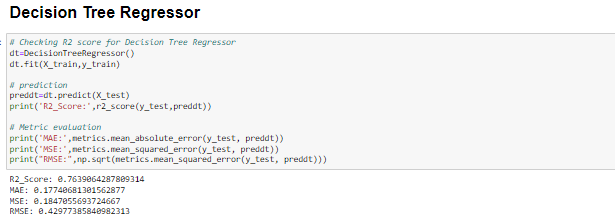
* I have checked for the feature importance using random forest classifier as follows



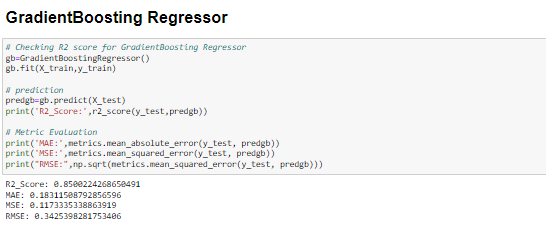
* Then I have checked the accuracy of the model using Random Forest Regressor



* We can observe that RFR has an accuracy of 88.64%
* Then I have checked the accuracy of the model using Decision Tree Regressor



* We can observe that DTR has an accuracy of 77.42%
* Then I have checked the accuracy of the model using Gradient Boosting Regressor



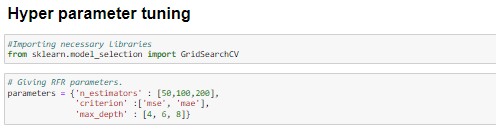
* We can observe that GBR has an accuracy of 85%
* Then I have checked the accuracy of the model using XGB Regressor



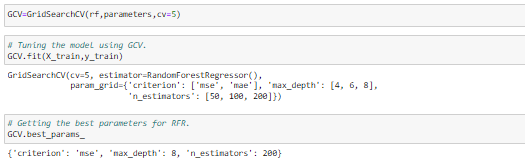
* We can observe that XGB has an accuracy of 86%
* Then I have checked for the Cross validation scores for all the above models as follows



* From the above we can observe the difference between the accuracy score and cross validation score is least in Random forest regressor. Hence we can predict that Random forest regressor is the best model.
* Then I have tuned the selected best model using Hyper parameter tuning for better performance as follows



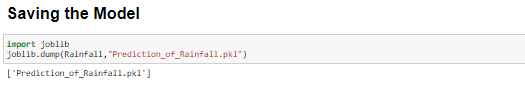
* I have tuned the Random Forest classifier and fit the current model for better performance as follows



* This has resulted in the accuracy as 88.64%, which is less than an expected level of accuracy.

But due to lots of missing data and multi collinearity rising from filling the missing data using imputation has resulted in lesser accuracy.

* Then I have saved the model importing and using joblib as follows



* Predicted using the saved model as follows



* I have created a dataframe of the actual vs predicted values for better view and comparison of outputs from the predicted model



* We can see that the Actual Values and Predicted Values look similar with few exceptions.

# Conclusions

**Key Findings and Conclusions of the Study:**

## **Findings:**

* In this project we have investigated the Rainfall and developed new knowledge to understand the most important factors influencing Rainfall prediction for weather forecasting.
* This project aimed to enhance prior understanding of how Rainfall prediction is dependent on temperature, pressure, and other factors for forecasting.
* The factors like ‘Rain Today’ ,’Huidity’,’Temperature’ had a positive impact on the Rainfall and Rain Tomorrow
* Thus one needs to pay attention to these factors more specifically and seek breakthroughs that can improve its Predictability and help in weather forecasting.

## **CONCLUDING REMARKS**

* The endeavor of this study is to identify the factors influencing rainfall tomorrow and the intensity of rainfall tomorrow in mm.
* In this project, I have done some feature engineering by using imputation methods to fill the null values in the data. Visualized the data using count plot, factor plot, Scatter plot, bar plot and distribution plot, also encoded the object data into numerical using label encoding method. Checked the statistical summary of the dataset and checked for skewness, outliers and correlation between the features.
* From the analysis it was found that rainfall prediction for tomorrow is dependent on various factors. All these factors influence the rainfall tomorrow. Majorly the Rainfall today and the Humidity, Pressure, Temperature and the clouds had a great influence on the rainfall tomorrow.
* Also the windspeed and wind direction has an impact on the rainfall tomorrow. Whereas the Day, Month, Year and Minimum temperature has least impact on the rainfall tomorrow. Hence some of these columns were dropped.
* Though the sunshine has a great importance in the prediction of rainfall tomorrow, it is seen that nearly 50% of its data is missing, we had to fill those values using imputation and this had a negative impact on the modelling as the data is more skewed due to the same values being repeated in 50% of the rows.
* After visualizing the data, I found that CoffsHarbour has higher rainfall with higher chance of rain tomorrow and Uluru has lowest rainfall with high chance of rain tomorrow compared to others and Rain today has higher chance of impacting rainfall tomorrow.
* Also, Melbourne has high rainfall with lower chance of rain tomorrow and Uluru has lower rainfall and there is no chance for rain tomorrow.
* Most of the rainfall is lower level and is densely populated at the lower level of evaporation. The more the evaporation, more the chance of rain tomorrow. As the sunshine is lower there is more chance of rainfall and rain tomorrow.

# THANK YOU