Regis University MSDS696 Data Science Practicum II

Olumide Aluko

Contents

Problem or Situation

Research Question

Data

Proposed Methodology

Project Timeline

Problem or Situation

Rainfall Prediction is a complex and uncertain task that significantly impacts human society.

Well-timed and accurate forecasting can proactively help reduce human and financial loss.

This project presents a set of experiments that use standard machine learning methods to build models that can predict whether it will rain tomorrow or not based on the weather data for that day in major cities in Australia.

Research Question

Can we predict whether it will rain tomorrow or not using data?

Solution: Design a predictive classification model (*Decision Tree and Logistics Regression*) using machine learning algorithms to forecast whether or not it will rain tomorrow in Australia.

Data

- Dataset Source: https://www.kaggle.com/code/ankitjoshi97/rainfall-in-australia-eda-prediction-89-acc/data

- The dataset is taken from Kaggle and contains about 10 years of daily weather observations from many locations across Australia.

- **Dataset Description:**
- Number of columns: 23
- Number of rows: 145460
- Number of Independent Columns: 22
- Number of Dependent Column: 1

Proposed Methodology

- Download dataset
- Import data into Jupiter Notebook
- Import required libraries
- Process data
- Perform data Exploration
- Clean and remove outliers in data
- Find categorical and numerical features in dataset
- Split data into training and testing set
- Perform feature scaling
- Use Decision Tree & Logistic Regression to build a predictive model whether or not it will rain tomorrow.
- Summarize results and conclusion

3. Import Libraries

Let's import the necessary libraries.

In [1]:

```
import matplotlib.pyplot as plt
import seaborn as sns
import pandas as pd
import numpy as np
import matplotlib
%matplotlib inline
import os
from sklearn.impute import SimpleImputer
from sklearn.preprocessing import MinMaxScaler, OneHotEncoder
from sklearn.tree import DecisionTreeClassifier, plot_tree, ex
from sklearn.metrics import accuracy_score, confusion_matrix
import pyarrow as pa
from sklearn.ensemble import RandomForestClassifier
import joblib
# Warnings configuration
# ============================================================
import warnings
warnings.filterwarnings('ignore')
```

Libraries

EDA – Summary of Data

- Observations:-

  - The average minimum temperature is 12.19 and average maximum temperature is 23.22-degree Celsius.
  - The mean rainfall is 2.35 mm. -The average sunshine received is 7.62 hour.
  - The average wind gust speed is 40.00 km/hr., and the median evaporation is 4.8 mm

```
ain.describe().T
```

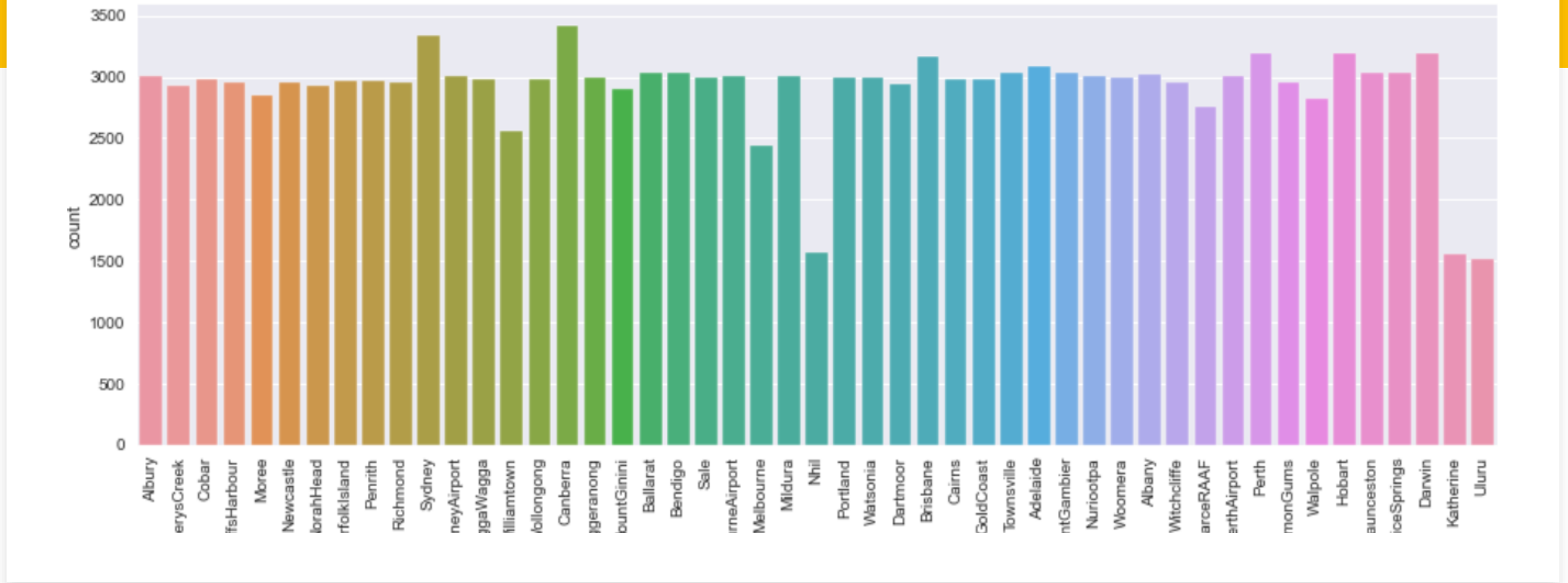
	count	mean	std	min	25%	50%	75%	ma
MinTemp	141556.0	12.186400	6.403283	-8.5	7.6	12.0	16.8	33.
MaxTemp	141871.0	23.226784	7.117618	-4.8	17.9	22.6	28.2	48.
Rainfall	140787.0	2.349974	8.465173	0.0	0.0	0.0	0.8	371.
Evaporation	81350.0	5.469824	4.188537	0.0	2.6	4.8	7.4	145.
Sunshine	74377.0	7.624853	3.781525	0.0	4.9	8.5	10.6	14.
lGustSpeed	132923.0	39.984292	13.588801	6.0	31.0	39.0	48.0	135.
dSpeed9am	140845.0	14.001988	8.893337	0.0	7.0	13.0	19.0	130.
dSpeed3pm	139563.0	18.637576	8.803345	0.0	13.0	19.0	24.0	87.
umidity9am	140419.0	68.843810	19.051293	0.0	57.0	70.0	83.0	100.
umidity3pm	138583.0	51.482606	20.797772	0.0	37.0	52.0	66.0	100.
ressure9am	128179.0	1017.653758	7.105476	980.5	1012.9	1017.6	1022.4	1041.
essure3pm	128212.0	1015.258204	7.036677	977.1	1010.4	1015.2	1020.0	1039.
Cloud9am	88536.0	4.437189	2.887016	0.0	1.0	5.0	7.0	9.
Cloud3pm	85099.0	4.503167	2.720633	0.0	2.0	5.0	7.0	9.
Temp9am	141289.0	16.987509	6.492838	-7.2	12.3	16.7	21.6	40.
Temp3pm	139467.0	21.687235	6.937594	-5.4	16.6	21.1	26.4	46.

	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustSpeed	WindSpeed9am	WindSpeed3pm	Humidity9am	Humidity3pm	Pressu
MinTemp	1.00	0.74	0.10	0.47	0.07	0.18	0.18	0.18	-0.23	0.01	
MaxTemp	0.74	1.00	-0.07	0.59	0.47	0.07	0.01	0.05	-0.51	-0.51	
Rainfall	0.10	-0.07	1.00	-0.06	-0.23	0.13	0.09	0.06	0.22	0.26	
Evaporation	0.47	0.59	-0.06	1.00	0.37	0.20	0.19	0.13	-0.51	-0.39	
Sunshine	0.07	0.47	-0.23	0.37	1.00	-0.03	0.01	0.06	-0.49	-0.63	
WindGustSpeed	0.18	0.07	0.13	0.20	-0.03	1.00	0.60	0.69	-0.22	-0.03	
WindSpeed9am	0.18	0.01	0.09	0.19	0.01	0.60	1.00	0.52	-0.27	-0.03	
WindSpeed3pm	0.18	0.05	0.06	0.13	0.06	0.69	0.52	1.00	-0.15	0.02	
Humidity9am	-0.23	-0.51	0.22	-0.51	-0.49	-0.22	-0.27	-0.15	1.00	0.67	
Humidity3pm	0.01	-0.51	0.26	-0.39	-0.63	-0.03	-0.03	0.02	0.67	1.00	
Pressure9am	-0.45	-0.33	-0.17	-0.27	0.04	-0.46	-0.23	-0.30	0.14	-0.03	
Pressure3pm	-0.46	-0.43	-0.13	-0.29	-0.02	-0.41	-0.17	-0.25	0.19	0.05	
Cloud9am	0.08	-0.29	0.20	-0.19	-0.68	0.07	0.02	0.05	0.45	0.52	

EDA - Correlation

- Observations:-

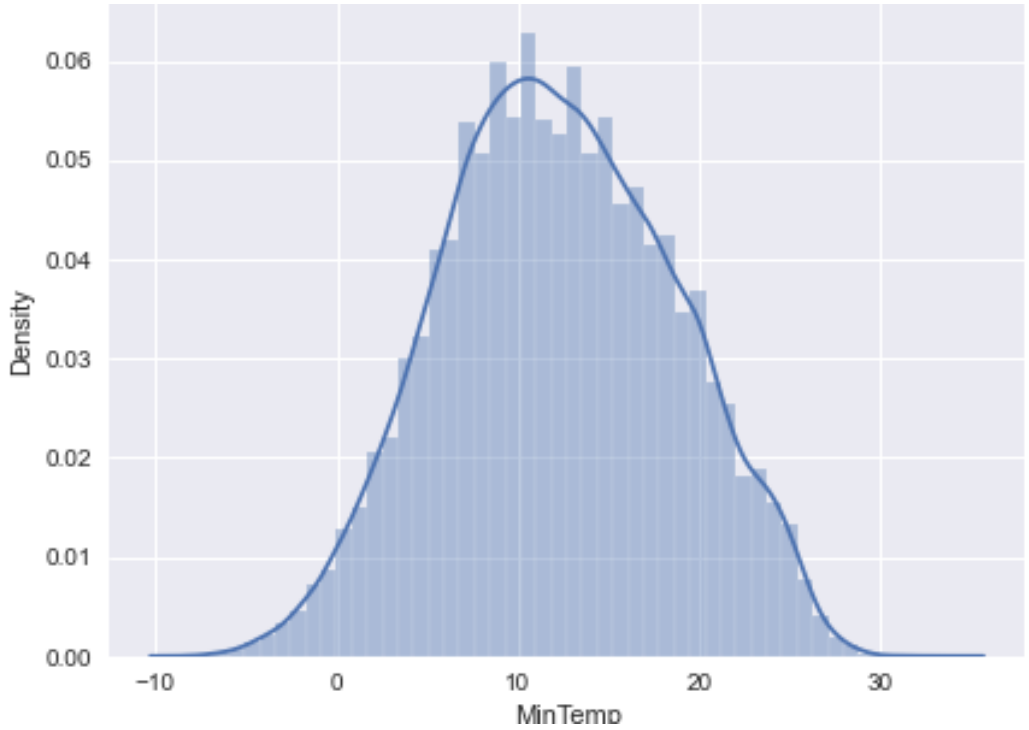
- Max Temp and Temp3pm have a strong positive correlation of 0.97.
- Pressure9am and Pressure3pm have a strong positive correlation of 0.96.
- Min Temp and Temp9am have a strong positive correlation of 0.90.
- Max Temp and Temp9am have a strong positive correlation of 0.88.

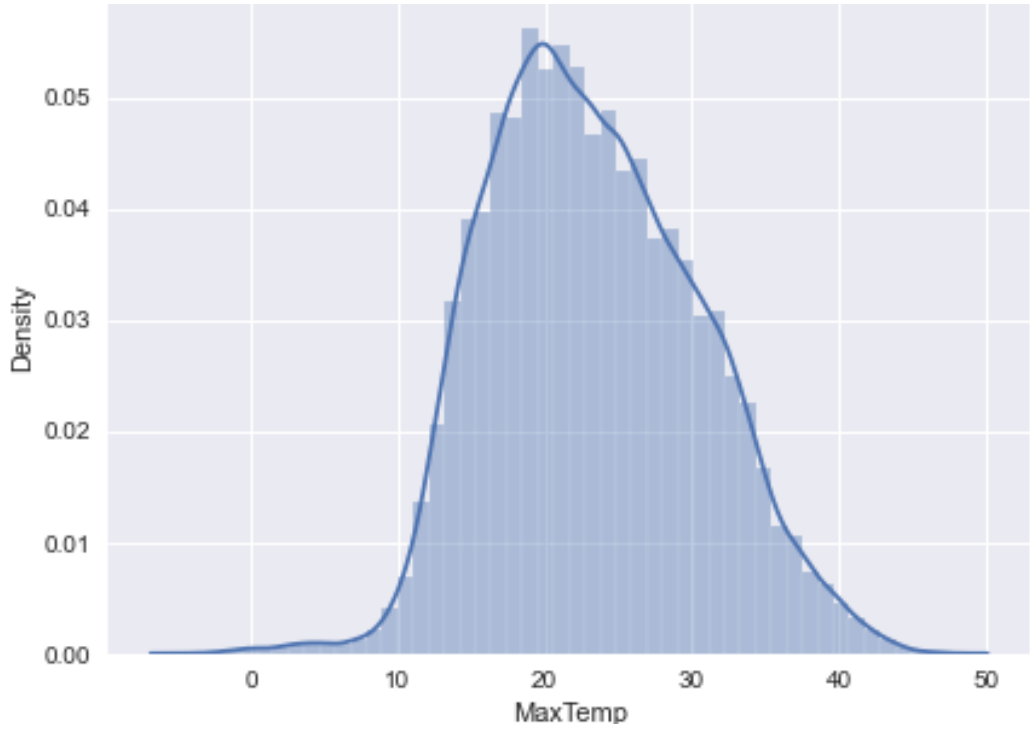


EDA - Distribution of Location

Observations:-

- Most occurred location is Canberra followed by Sydney.
- Most of the locations have a frequency near 3000.
- Nhil, Katherine and Uluru have occurred the least.

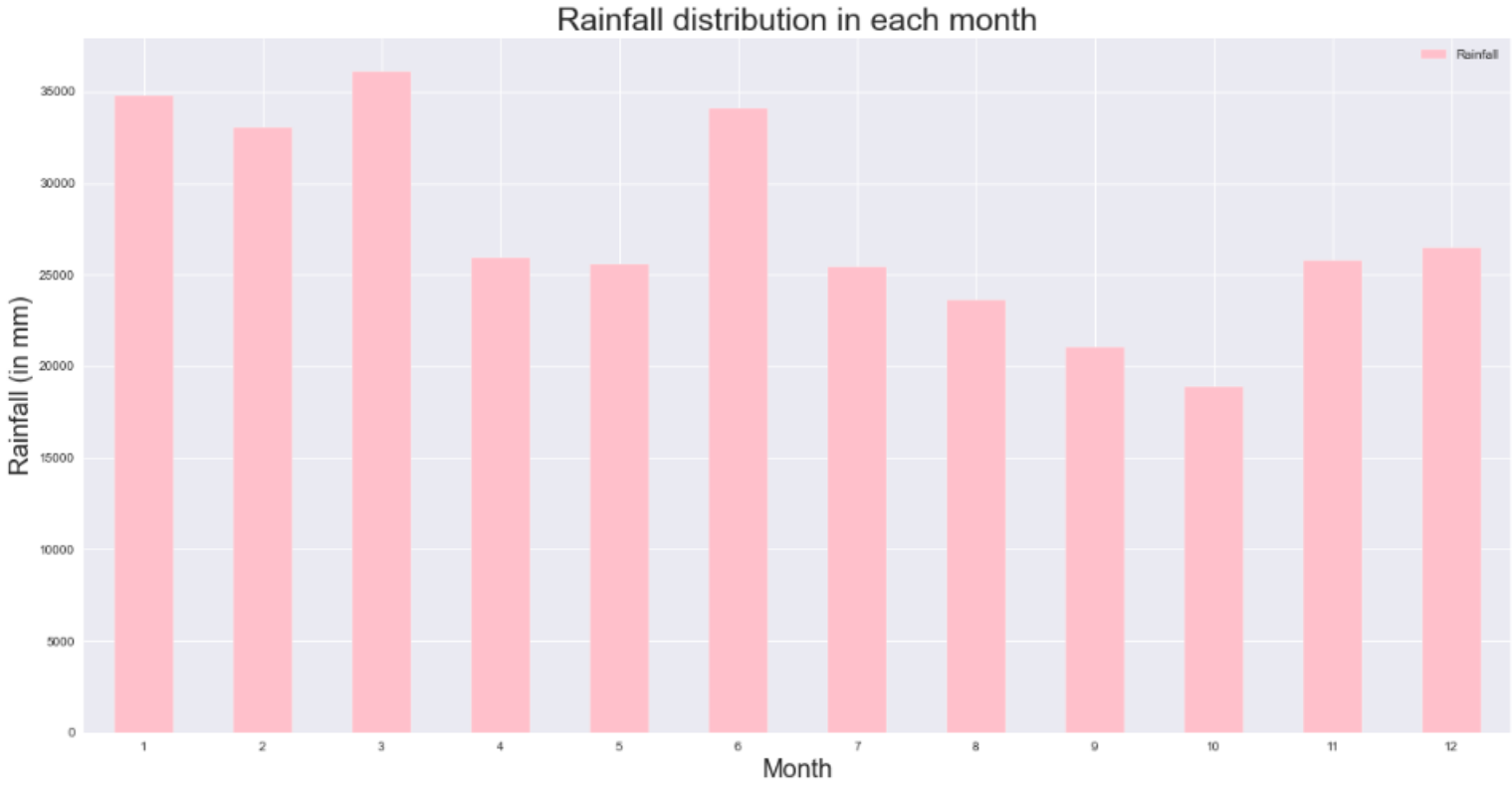




EDA – Min/Max Temp

- Observations:-

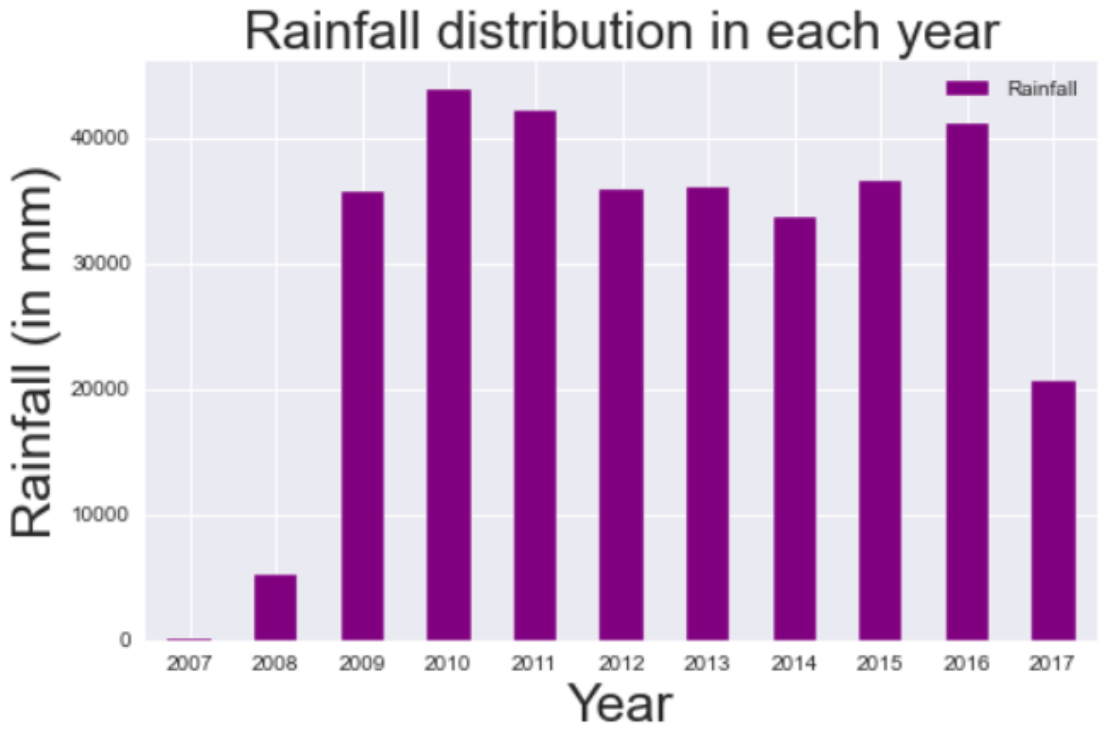
- The Highest concentration of points for minimum temperature is between 10-to-12-degree Celsius.
- The Highest concentration of points for maximum temperature is between 18to 22-degree Celsius.



EDA – Rainfall Distribution by Month

- Observations:-

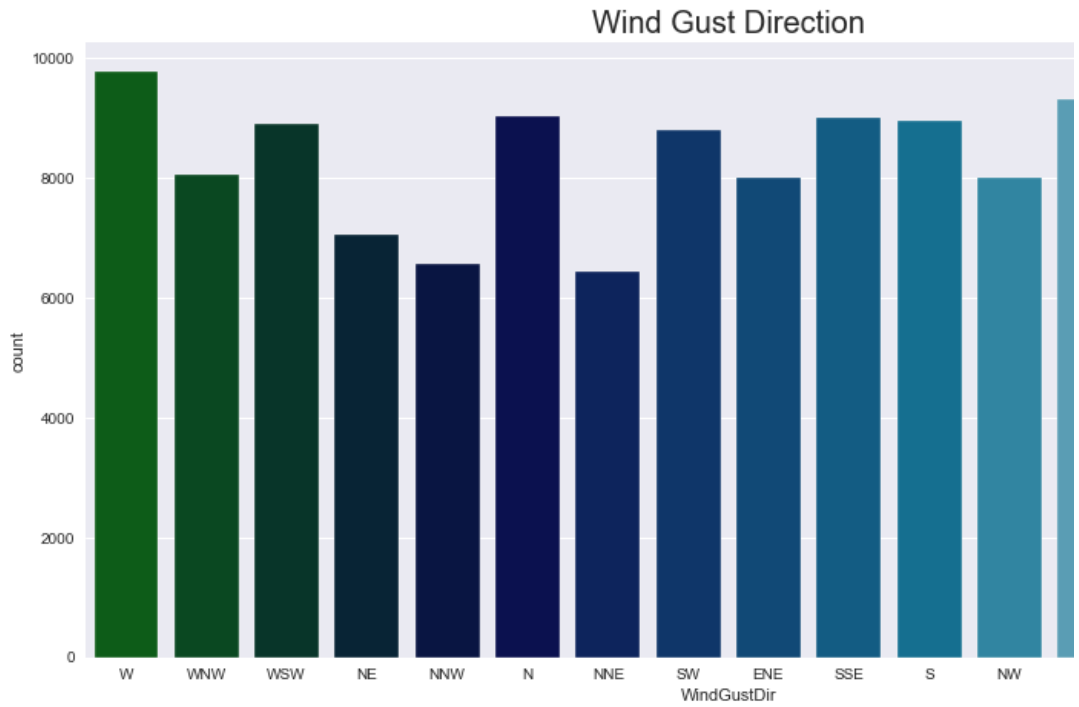
- Maximum rainfall (greater than 35,000 mm) occurs in March.
- January and June also experience high rainfall(nearly 35,000 mm) followed by February.
- Minimum rainfall occurs in October followed by September.

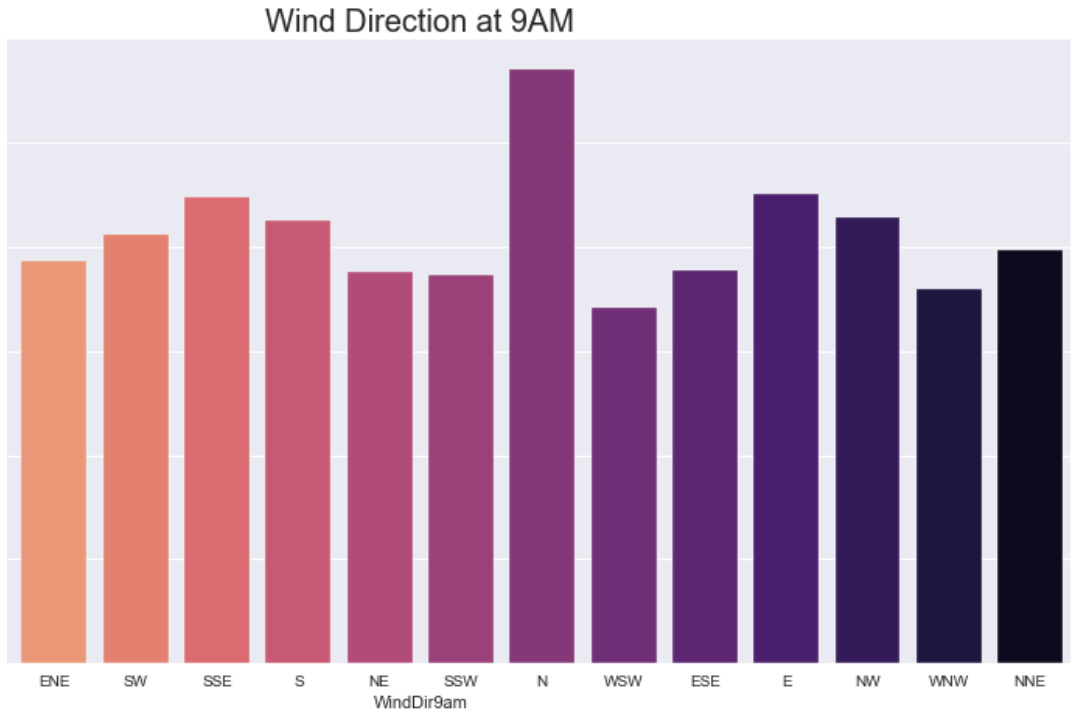


EDA - Rainfall Distribution by Year

- Observations:-

- Maximum rainfall(greater than 40,000 mm) occurred in 2010 followed by 2011 and 2016.
- 2009, 2012, 2013, 2014 and 2015 experienced rainfall between 30,000-40,000 mm.
- Least rainfall(less than 200 mm) occurred in 2007 followed by 2008 and 2017 (greater than 20,000 mm).





EDA – Wind Gust Direction & Wind Direction at 9AM

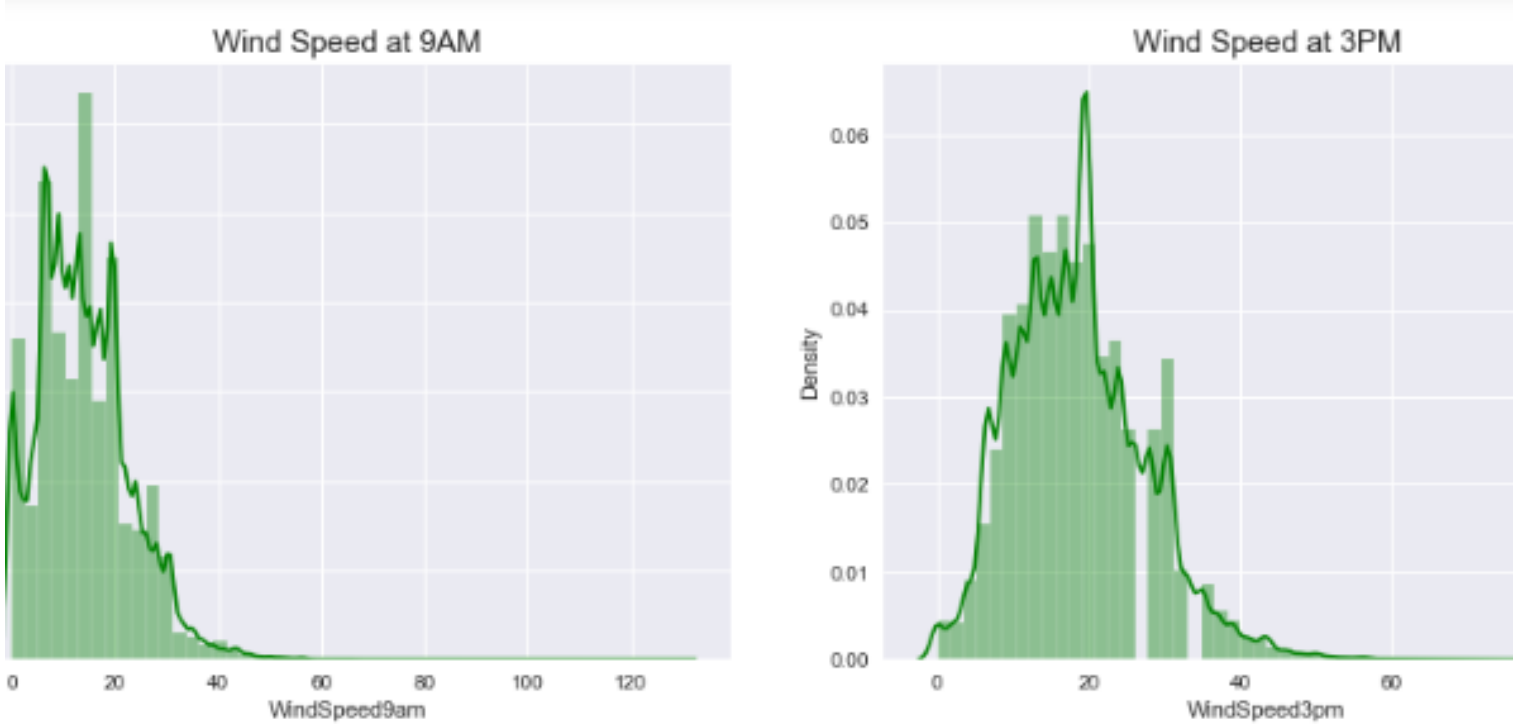
- Observations:-

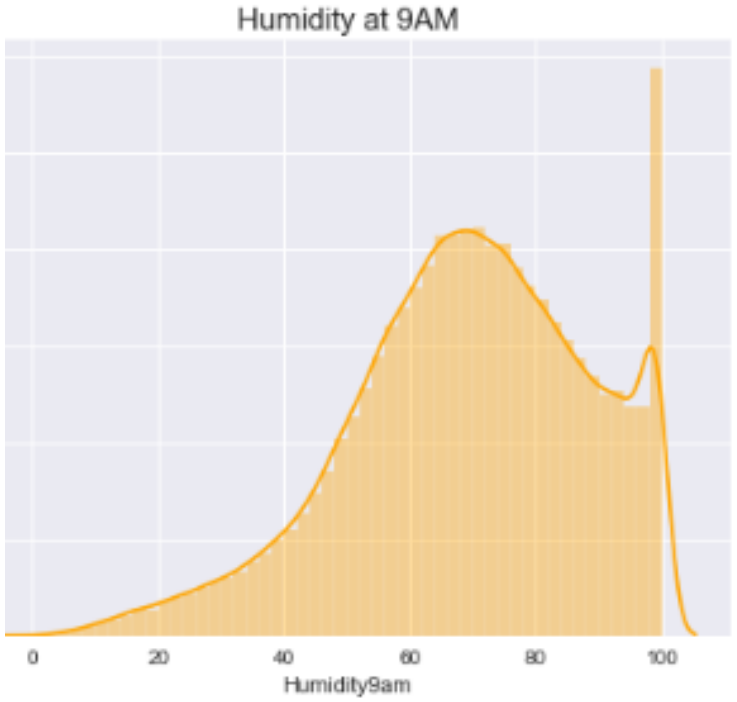
- Wind Gust Direction for maximum records(nearly 17,500) is West.
- Wind Direction at 9AM for maximum records is North followed by North-West and East.

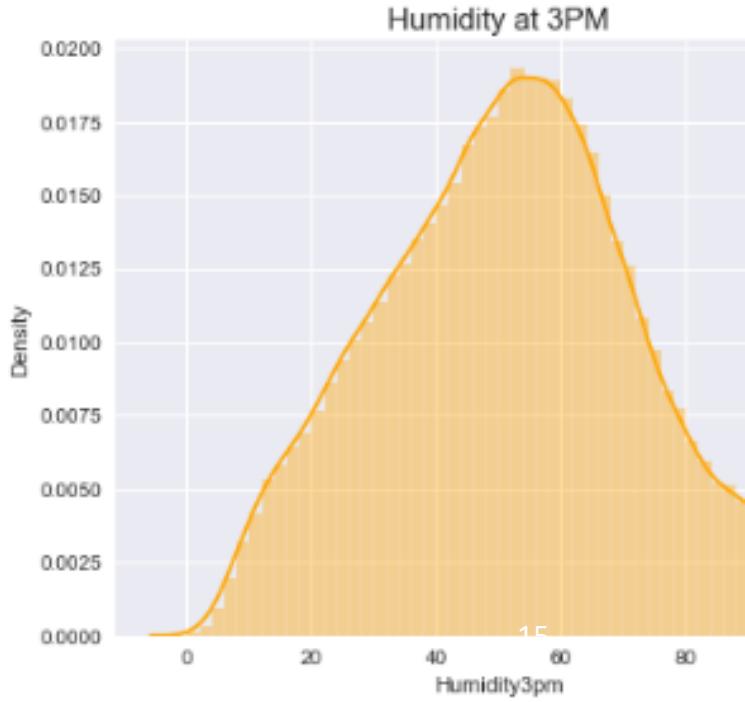
EDA – Wind Speed & Humidity

- Observations:-

  - Maximum wind speed at 9AM ranges from 10 to 20 km/hr. whereas at 3PM it ranges from 15 to 22 km/hr.
  - Highest concentration of points for humidity at 9AM is between 60-80% whereas at 3PM it's 40-70%.



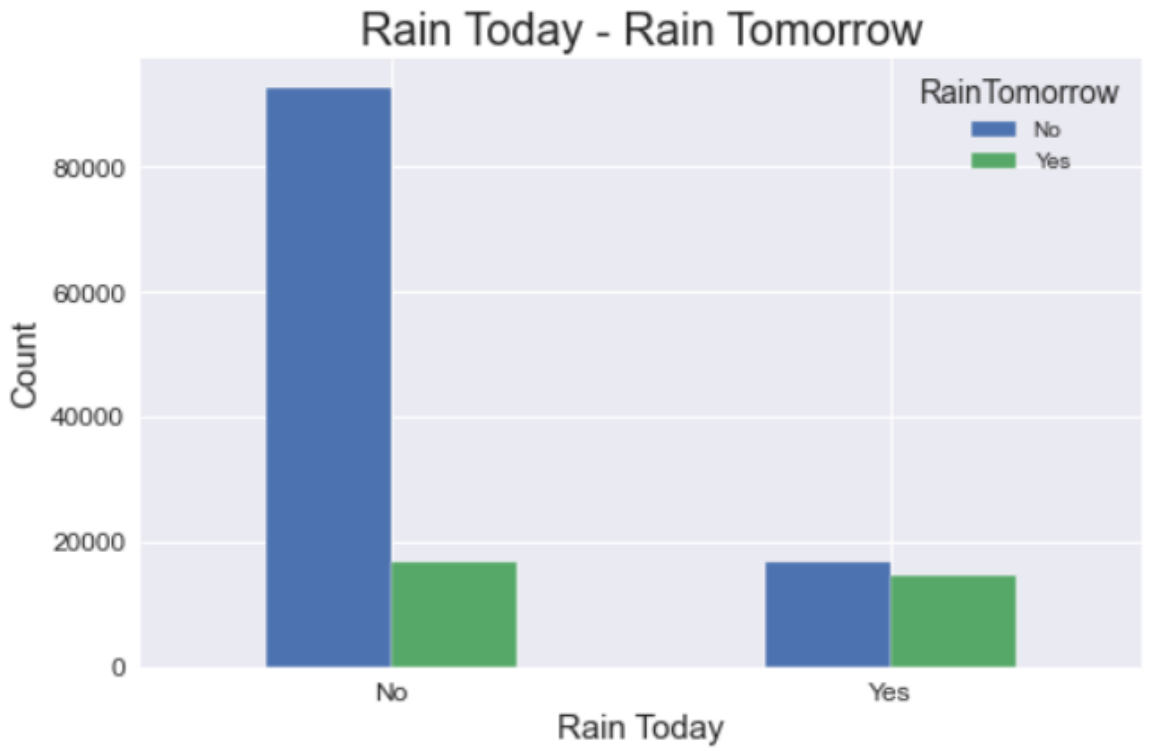


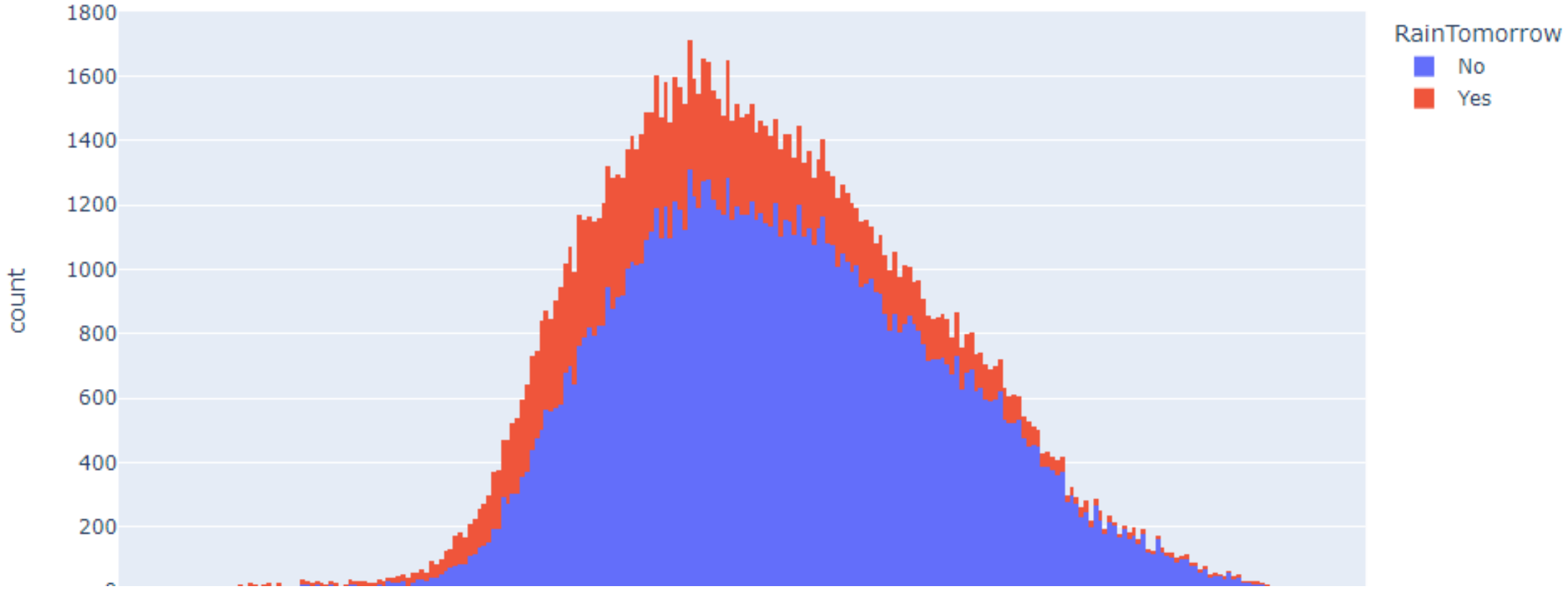


EDA – Rain Today & Rain Tomorrow

- Observations:-

  - For maximum records it didn't rain for both days.
  - For nearly 20,000 records it didn't rain today but rained tomorrow and rained for both days.
  - For nearly 20,000 records it rained today but didn't rain tomorrow.





EDA – Temperature at 3 PM VS. Rain Tomorrow

- Observations:-
  - Raintomorrow with "No" has the highest count of 1311 when Temp3pm is between (18.4 – 18.5) Celsius.
  - Raintomorrow with “YES" has the highest count of 401 when Temp3pm is between (18.4 – 18.5) Celsius.

DecisionTree Model Training - Evaluation

```
In [49]: Y_val.value_counts() / len(Y_val)

Out[49]: No     0.788289
         Yes    0.211711
         Name: RainTomorrow, dtype: float64
```

```
In [46]: X_train_pred = model.predict(X_train)
         pd.value_counts(X_train_pred)

Out[46]: No     76707
         Yes    22281
         dtype: int64
```

```
In [47]: train_probs = model.predict_proba(X_train)
         print('Training Accuracy :',accuracy_score(X_train_pred,Y_train)*100)

         Training Accuracy : 99.99797955307714
```

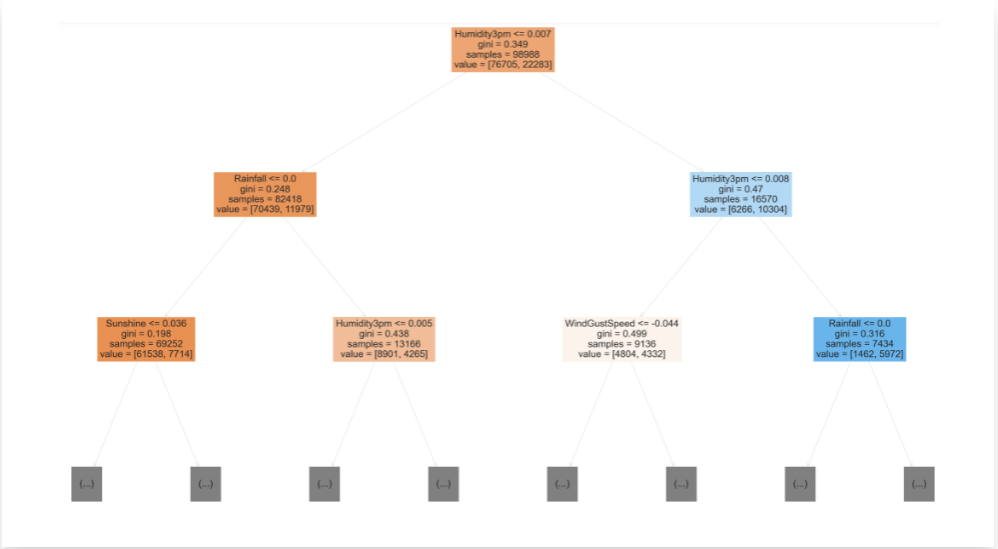
```
In [48]: print('Validation Acuracy :',model.score(X_val,Y_val)*100)

         Validation Acuracy : 79.28152747954267
```

Observations:-

Evaluation – Performance of DecisionTreeClassifier

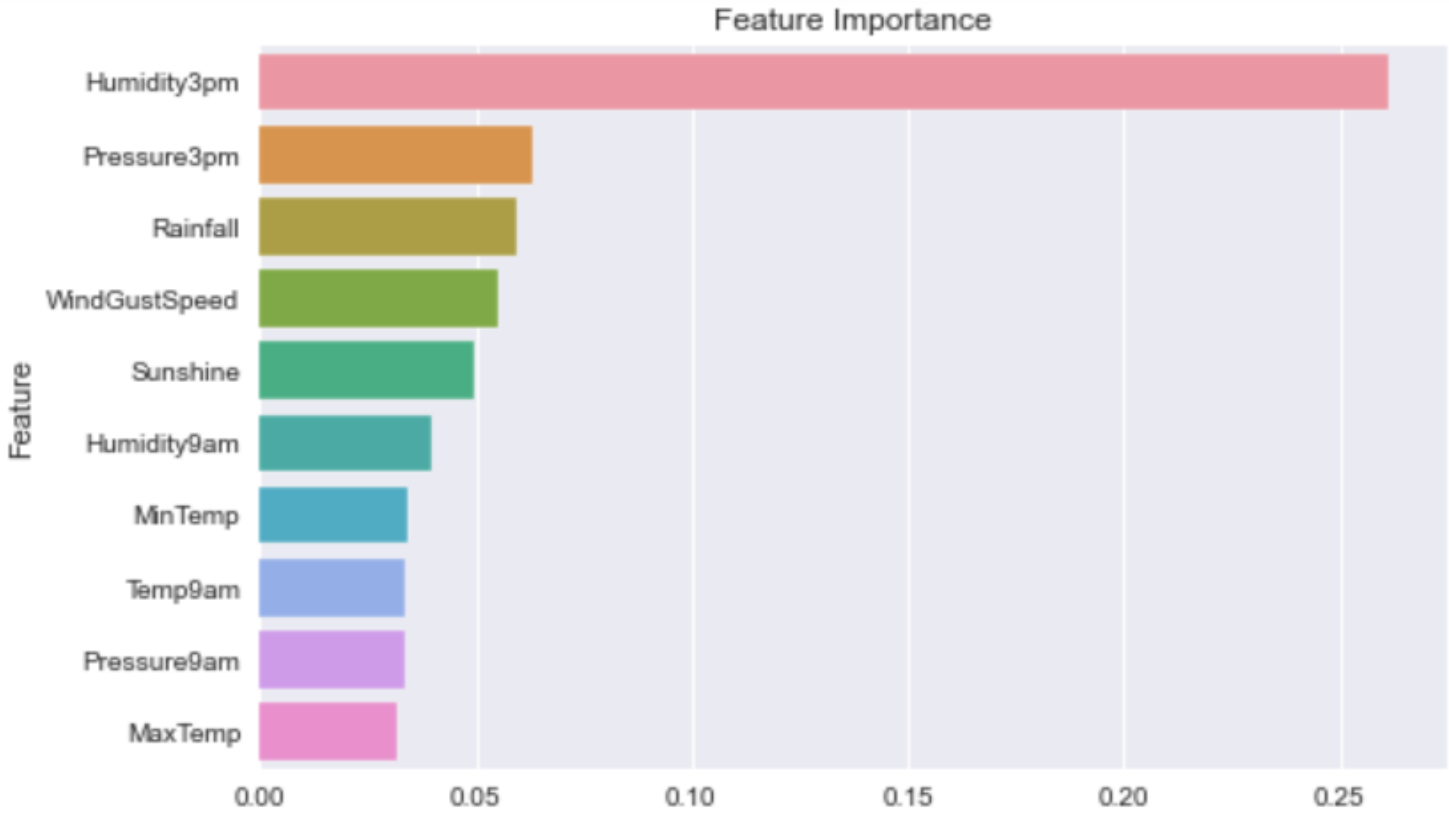
- Observations:-



- RainTomorrow shows 78.8% 'No' and 21% 'Yes' in validation data.
- The validation accuracy is 79.28
- The training set accuracy is 99.99
- The above case was an overfitting case as tree used the (max_depth = 48) and memorized the values
- And it failed to predict with low accuracy of 79.28% for test and validation dataset

DecisionTreeClassifier with default parameters	
Accuracy of training data	99.99%
Accuracy of test and validation data	79.28%
Max depth of decision tree	48

Table 1 - Performance of DecisionTreeClassifier with default parameters



Evaluation – Feature Importance

Observations:-

- Humidity3pm has the highest feature importance of 0.27
- Pressure3pm and Rainfall has the second highest feature importance of less than 0.10.
- MaxTemp has the lowest feature importance of less than 0.05.

Hyperparameter Tuning – Confusion Matrix for Training & Validation Data

Observations:-

- The training accuracy is just 83% which means the model is not memorizing and overfitting the values. See table 1.
- The validation accuracy is just 84%. See table 2.
- We now have a better performance on training and test dataset.
- DecisionTreeClassifier with a (max_depth=4) was utilized.

```
[[72602  4103]
 [12300  9983]]
```

	precision	recall	f1-score	support
No	0.86	0.95	0.90	76705
Yes	0.71	0.45	0.55	22283
accuracy			0.83	98988
macro avg	0.78	0.70	0.72	98988
weighted avg	0.82	0.83	0.82	98988

Table 1: Confusion matrix for training data

```
[[12855   728]
 [ 2104  1544]]
```

	precision	recall	f1-score	support
No	0.86	0.95	0.90	13583
Yes	0.68	0.42	0.52	3648
accuracy			0.84	17231
macro avg	0.77	0.68	0.71	17231
weighted avg	0.82	0.84	0.82	17231

Table 2: Confusion matrix for validation data

Hyperparameter Tuning – Tuning max_depth

Observations:-

- As the max_depth value without manual constraint for which our model overfitted is 48. And the max_depth value can't be 0 or lesser.
- Let's find what the best value of max_depth method for which the errors of train and validation dataset is optimal.
- From table 1, the training accuracy increases with increase in max_depth.
- From Table 1, the validation accuracy first increases and then decreases.

Out[61]:

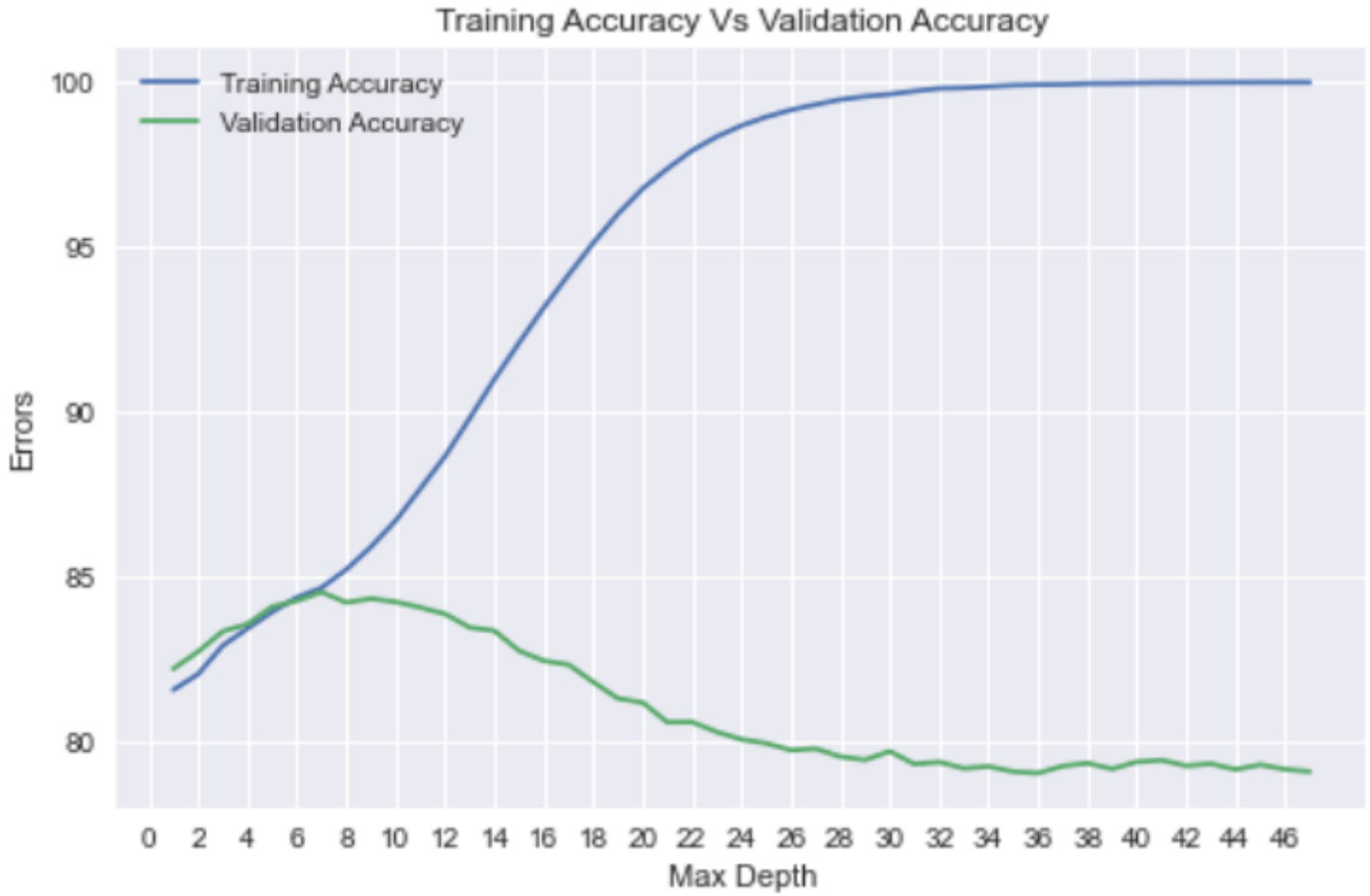
	Max_Depth	Training_Accuracy	Validation_Accuracy
0	1	81.568473	82.206488
1	2	82.045298	82.728803
2	3	82.913080	83.343973
3	4	83.429305	83.564506
4	5	83.932396	84.092624
5	6	84.372853	84.272532
6	7	84.668849	84.533689
7	8	85.219421	84.220301
8	9	85.908393	84.336370
9	10	86.703439	84.237711
10	11	87.675274	84.069410
11	12	88.655191	83.872091
12	13	89.813917	83.460043
13	14	90.997899	83.361384

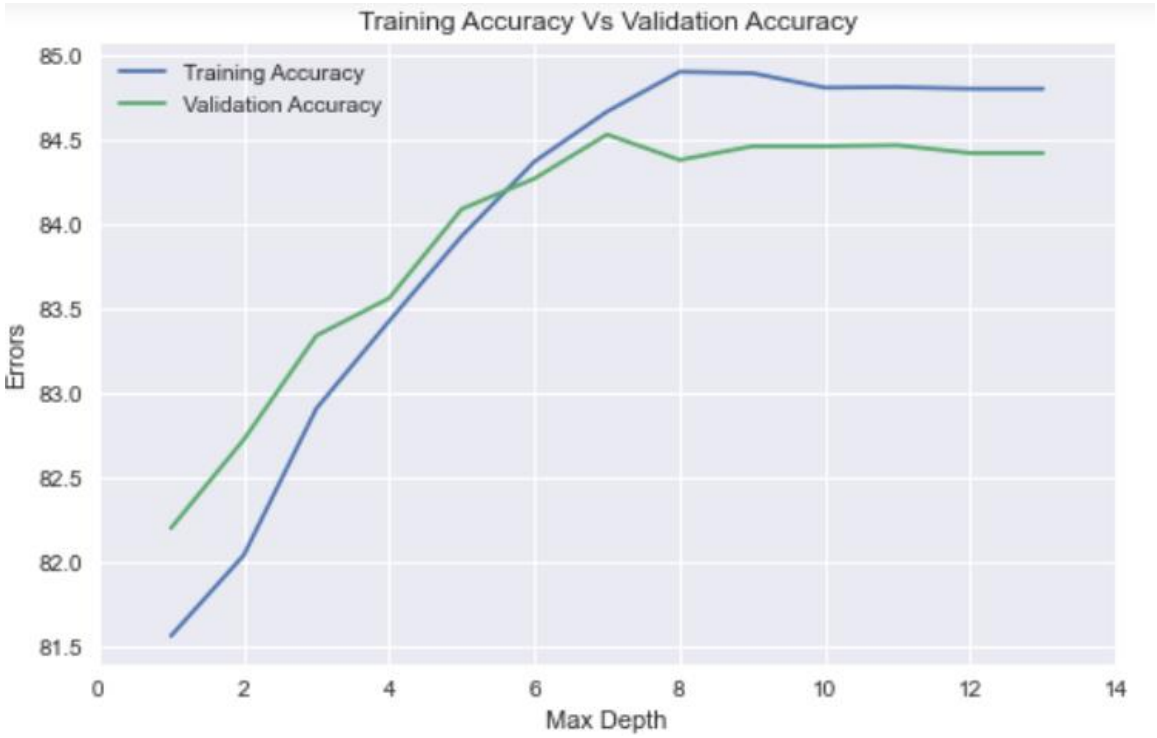
Table 1: Data frame of Training & Validation Accuracy

Hyperparameter Tuning – Tuning Graph at (max_depth = 7)

- Observations:-

- The graph shows that training accuracy increases with increase in max_depth while validation accuracy first increases (till max_depth = 7).
- And then decreases. Hence, optimal max_depth is 7.
- Build Decision Tree with (max_depth = 7), the training accuracy was 84.66
- Build Decision Tree with (max_depth = 7), the validation accuracy was 84.53





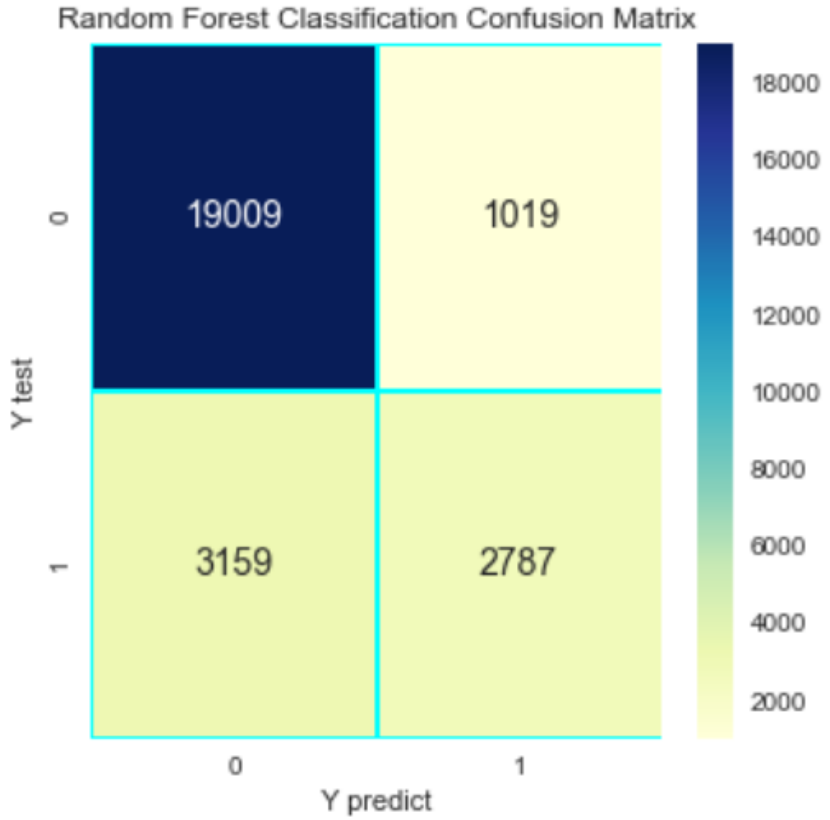
DecisionTreeClassifier with default parameters	
Accuracy of training data	84.89%
Accuracy of test and validation data	84.50%
Max depth of decision tree	9
Max leaf nodes	128

Table 2 - Performance of model with Hyperparameter tuning

Hyperparameter Adjustment – Tuning Graph at (max_depth = 9)

Observations:-

- From the graph, it seems (max_depth = 9) and (max_leaf_nodes = 128) is the optimal hyperparameters.
- Build Decision Tree with (max_depth = 9), the training accuracy was 84.89
- Build Decision Tree with (max_depth = 9), the validation accuracy was 84.46
- From the graph, performance is improved for new predictions as accuracy of training data, test data and validation data is almost the same.




```
In [76]: print("Training accuracy = ", rfc.score(X_train, Y_train) * 100, "%")
         Training accuracy =  99.99595910615429 %

In [77]: print("Validation accuracy = ", rfc.score(X_val, Y_val) * 100, "%")
         Validation accuracy =  85.58412164122802 %

In [78]: print("Test accuracy = ", rfc.score(X_test, Y_test) * 100, "%")
         Test accuracy =  84.00323400323401 %
```

Random Forest Algorithm Training – Model Score

Observations:-

- From the output, the RandomForest training accuracy is 99.99%.
- The RandomForest validation accuracy is 85.58% & the test accuracy is 84.0%
- From the above confusion matrix, there are 19,009 true negative values, 3,159 false negative values, 1,019 false positive values, and 2,787 true positive values.
- This proves that the RandomForest is the best model.

Conclusion

- For the decision tree model, the training accuracy is 99.99%, validation accuracy is 79.28% and the percentage of 'No' in validation data is 78.8%. Hence, our model is only marginally better than always predicting "No". This occurs because the training data from which our model learned remains skewed towards 'No’ Decision tree overfit.
- After an Hyperparameter tuning was applied to make some changes in the parameters of the model training to avoid overfitting. We were able to predict with a training accuracy of 84.89% and validation accuracy of 84.46% using DecisionTree.
- The RandomForest has a training accuracy of 99.99% and a validation accuracy of 85.58%. From the performance of the two models, Random Forest is greater than Decision Tree.
- Finally, one can establish that the Random Forest model is better in the sense it yields higher accuracy than other models.