Data Visualization Paper

Olumide Aluko

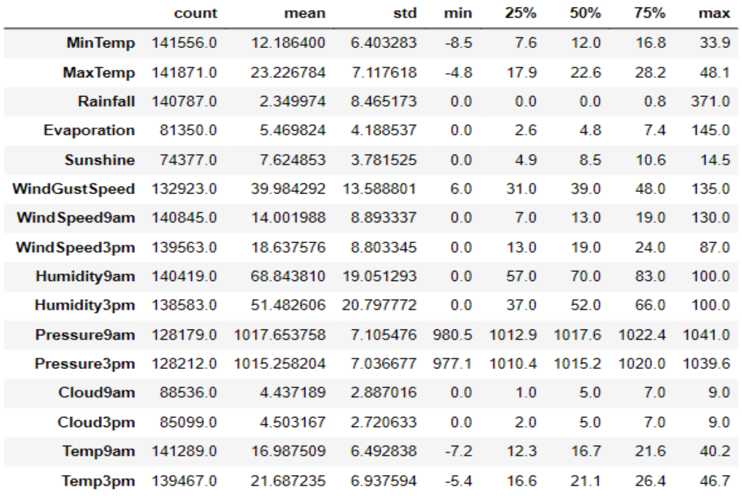
Anderson College of Business and Computing

Regis University

MSCC696: Data Science Practicum II

Prof. John Koenig

August 21, 2022



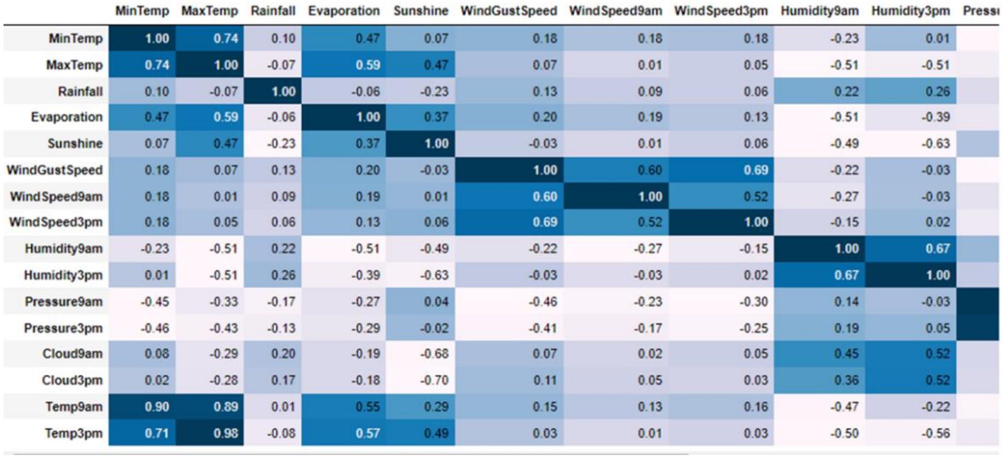
	count	mean	std	min	25%	50%	75%	max
MinTemp	141556.0	12.186400	6.403283	-8.5	7.6	12.0	16.8	33.9
MaxTemp	141871.0	23.226784	7.117618	-4.8	17.9	22.6	28.2	48.1
Rainfall	140787.0	2.349974	8.465173	0.0	0.0	0.0	0.8	371.0
Evaporation	81350.0	5.469824	4.188537	0.0	2.6	4.8	7.4	145.0
Sunshine	74377.0	7.624853	3.781525	0.0	4.9	8.5	10.6	14.5
WindGustSpeed	132923.0	39.984292	13.588801	6.0	31.0	39.0	48.0	135.0
WindSpeed9am	140845.0	14.001988	8.893337	0.0	7.0	13.0	19.0	130.0
WindSpeed3pm	139563.0	18.637576	8.803345	0.0	13.0	19.0	24.0	87.0
Humidity9am	140419.0	68.843810	19.051293	0.0	57.0	70.0	83.0	100.0
Humidity3pm	138583.0	51.482606	20.797772	0.0	37.0	52.0	66.0	100.0
Pressure9am	128179.0	1017.653758	7.105476	980.5	1012.9	1017.6	1022.4	1041.0
Pressure3pm	128212.0	1015.258204	7.036677	977.1	1010.4	1015.2	1020.0	1039.6
Cloud9am	88536.0	4.437189	2.887016	0.0	1.0	5.0	7.0	9.0
Cloud3pm	85099.0	4.503167	2.720633	0.0	2.0	5.0	7.0	9.0
Temp9am	141289.0	16.987509	6.492838	-7.2	12.3	16.7	21.6	40.2
Temp3pm	139467.0	21.687235	6.937594	-5.4	16.6	21.1	26.4	46.7

EDA – Summary of Data

- 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



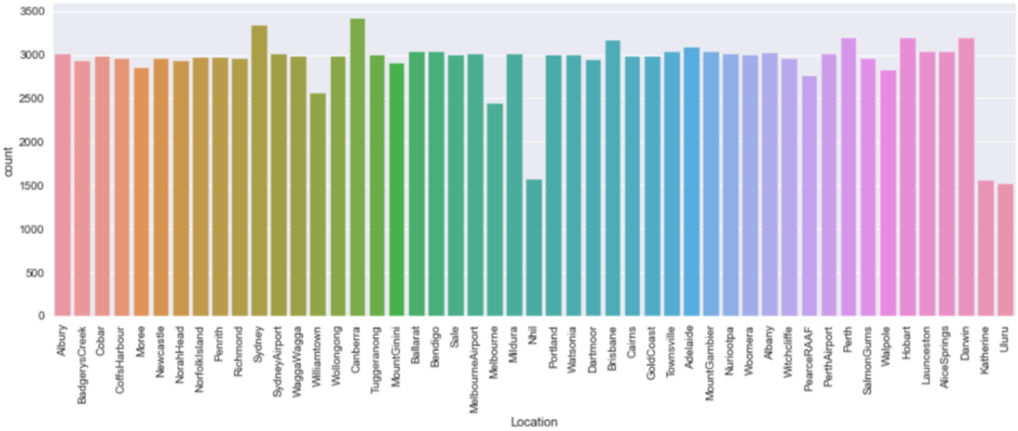
	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustSpeed	WindSpeed9am	WindSpeed3pm	Humidity9am	Humidity3pm	Pressı
MinTemp	1.00	0.74	0.10	0.47	0.07	0.18	0.18	0.18	-0.23	0.01	
MaxTemp	0.74	1.00	-0.07	0.59	0.47	0.07	0.01	0.05	-0.51	-0.51	
Rainfall	0.10	-0.07	1.00	-0.06	-0.23	0.13	0.09	0.06	0.22	0.26	
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	
Cloud3pm	0.02	-0.28	0.17	-0.18	-0.70	0.11	0.05	0.03	0.36	0.52	
Temp9am	0.90	0.89	0.01	0.55	0.29	0.15	0.13	0.16	-0.47	-0.22	
Temp3pm	0.71	0.98	-0.08	0.57	0.49	0.03	0.01	0.03	-0.50	-0.56	

EDA – Correlation of Data

- 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.

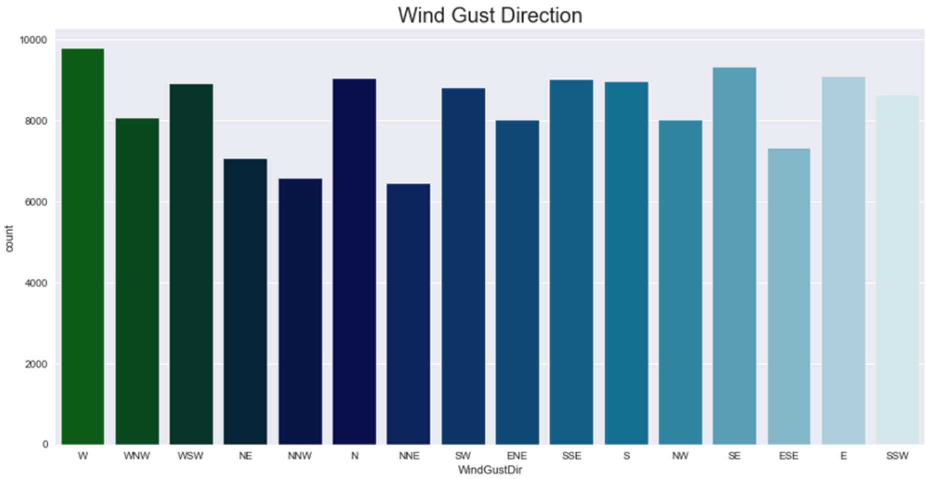


EDA - Distribution of Location

- Most occurred location is Canberra followed by Sydney.

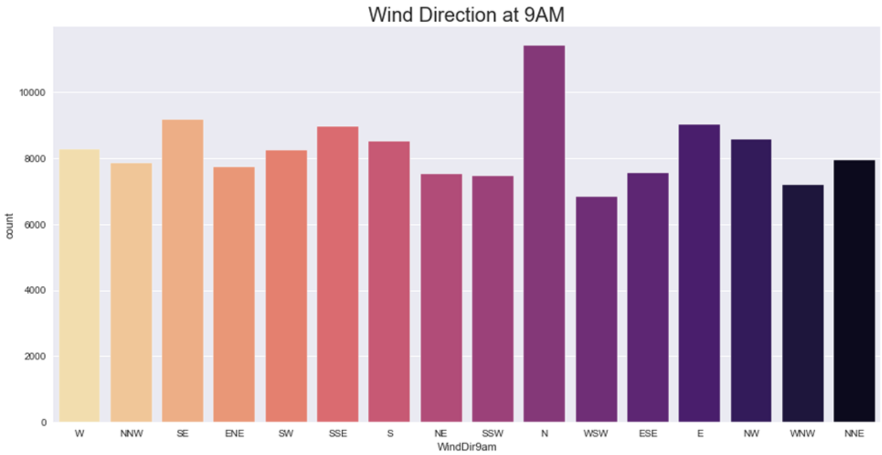
- Most of the locations have a frequency near 3000.

- Nhil, Katherine and Uluru have occurred the least.



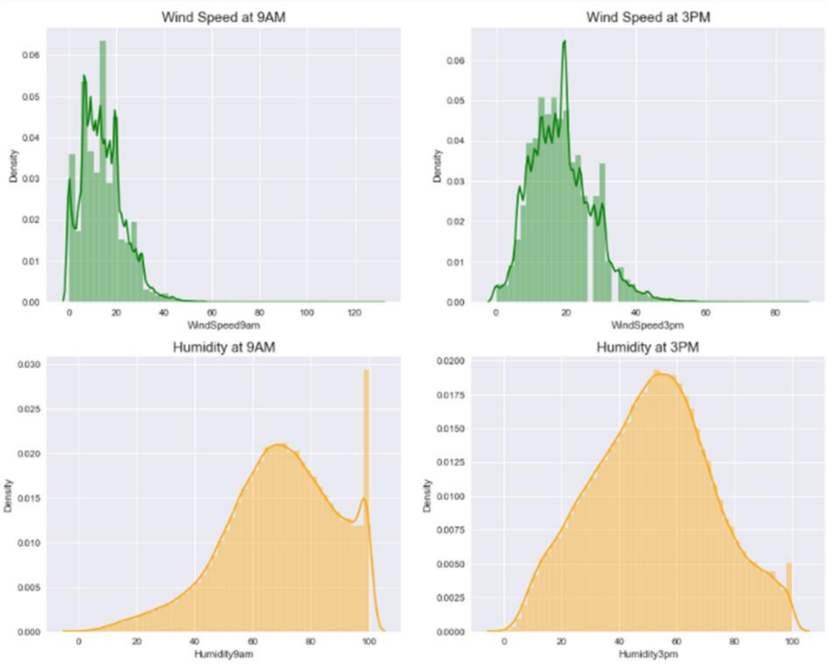
EDA – Wind Gust Direction

- Wind Gust Direction for maximum records(nearly 17,500) is West.



EDA – Wind Direction at 9AM

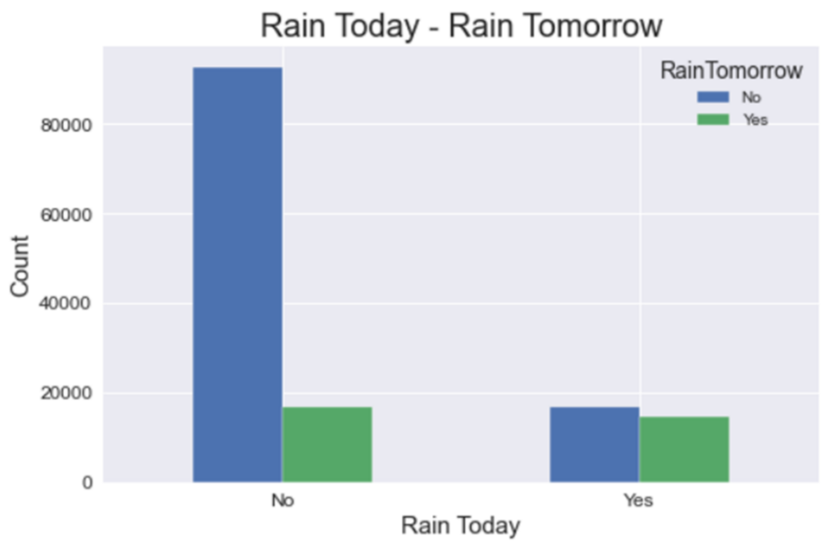
- Wind Direction at 9AM for maximum records is North followed by North-West and East.



EDA – Wind Speed & Humidity

- 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%.

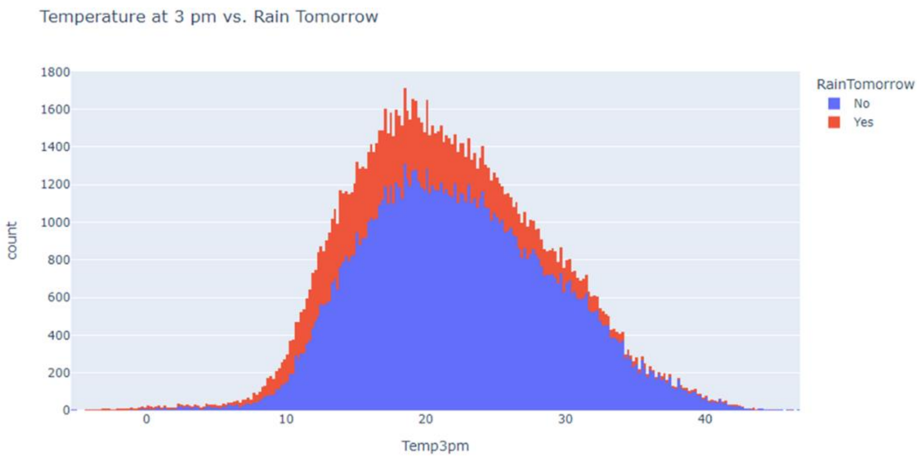


EDA – Rain Today & Rain Tomorrow

- 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.

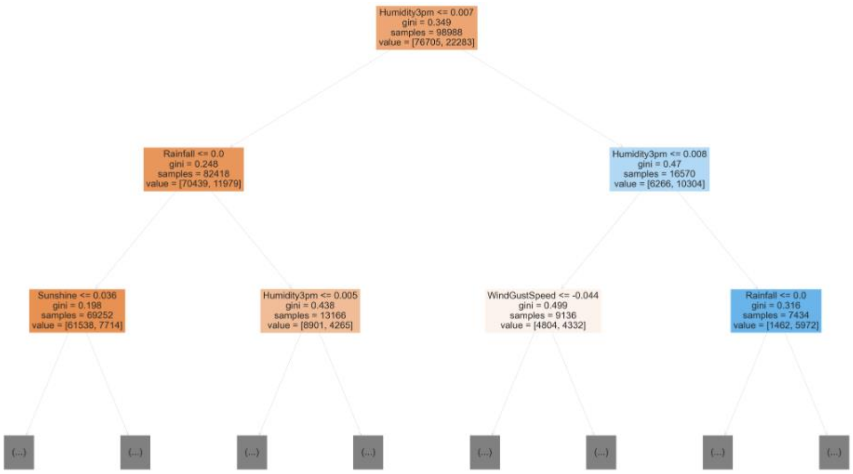
- For nearly 20,000 records it rained today but didn't rain tomorrow.



EDA – Temperature at 3 PM VS. Rain

- 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.



Performance of DecisionTreeClassifier

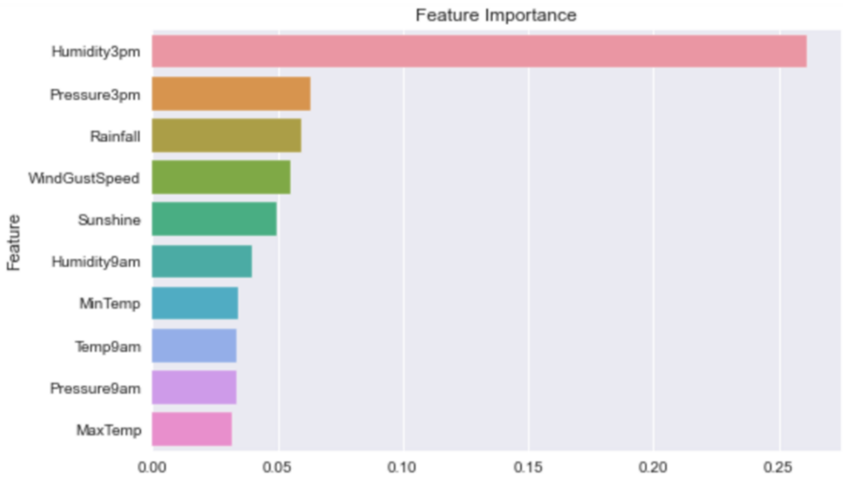
- RainTomorrow shows 78.8% 'No' and 21% 'Yes' in validation data.

- 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

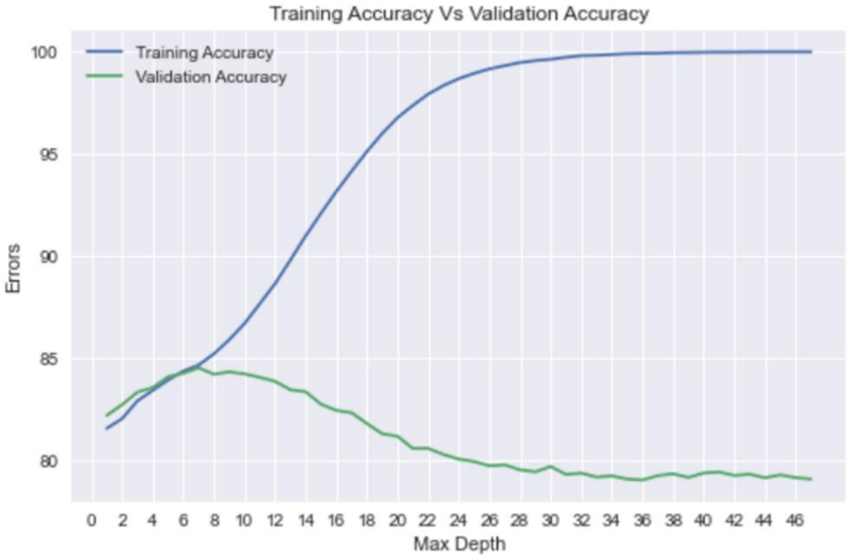


Evaluation – Feauture Importance

- 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.



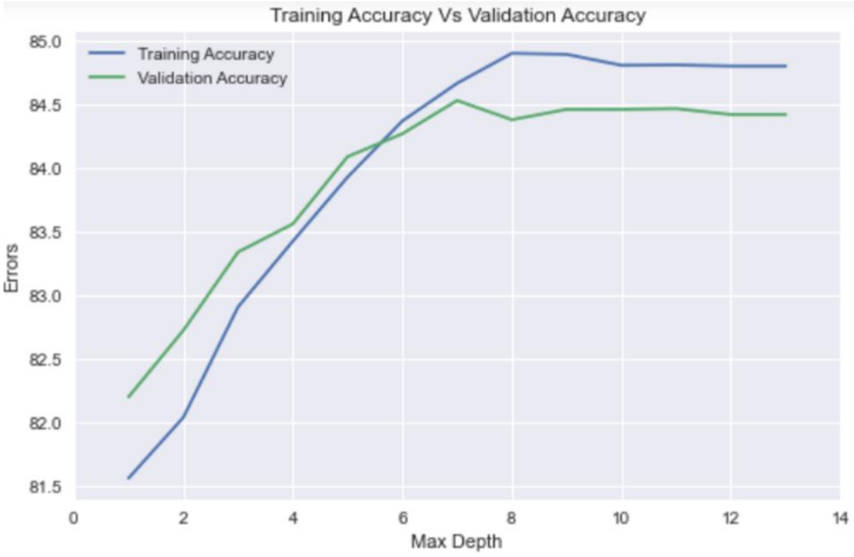
Hyperparameter Tuning – Tuning Graph 1

- 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



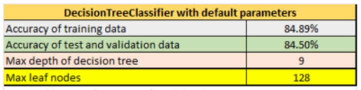
Hyperparameter Tuning – Tuning Graph 2

- 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.



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