Report: Rain in Australia Classifier

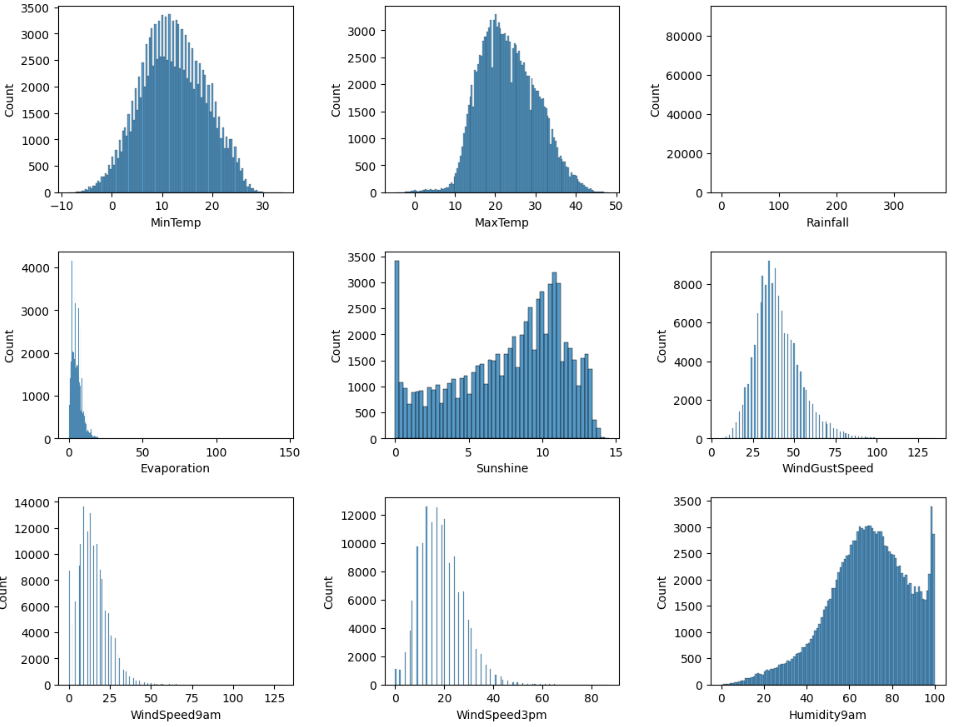
Illia Novikov SD2 ID 91651

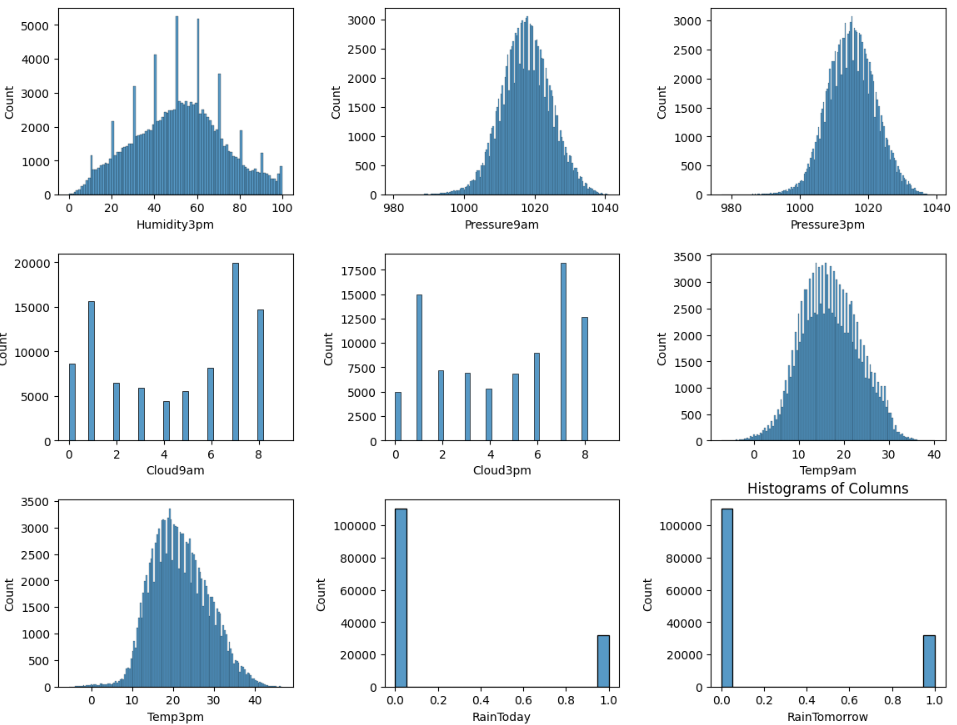
1. **Short description of the chosen dataset**: The "Rain in Australia" dataset is a collection of weather data from across Australia. It includes daily records from many weather stations, with information on rainfall, temperature, humidity, wind speed, wind directions etc. This dataset helps us understand how much it rains in different parts of Australia and how weather patterns change over time. It's useful for studying weather forecasts and learning more about Classifying Models.

Link: <https://www.kaggle.com/datasets/jsphyg/weather-dataset-rattle-package/data>

1. **Statistical analysis of dataset:**

**Histograms**

****

****

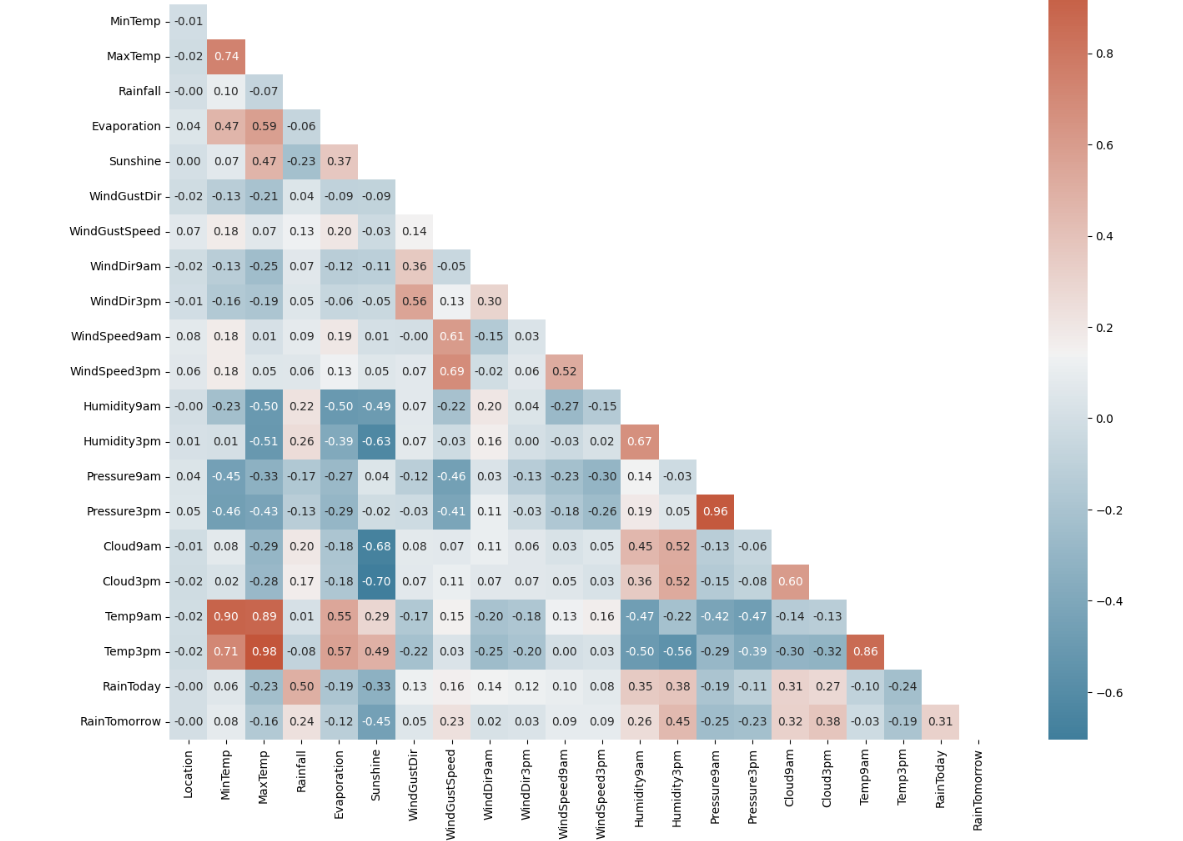
A graph with a bar graph

Description automatically generated

Observations:

* Most features are normally distributed
* The *Sunshine* distribution is interesting as the high frequency of 0 values represents days where it is overcast all day and the strong decline in frequency after around 11 hours is a reflection of the limited number of days of the year where it is the sun out for that many hours
* The Humidity9am distribution is particularly interesting due to the large spikes in frequencies
* Rainfall feature is only seen in the distribution of values between 0 and 1

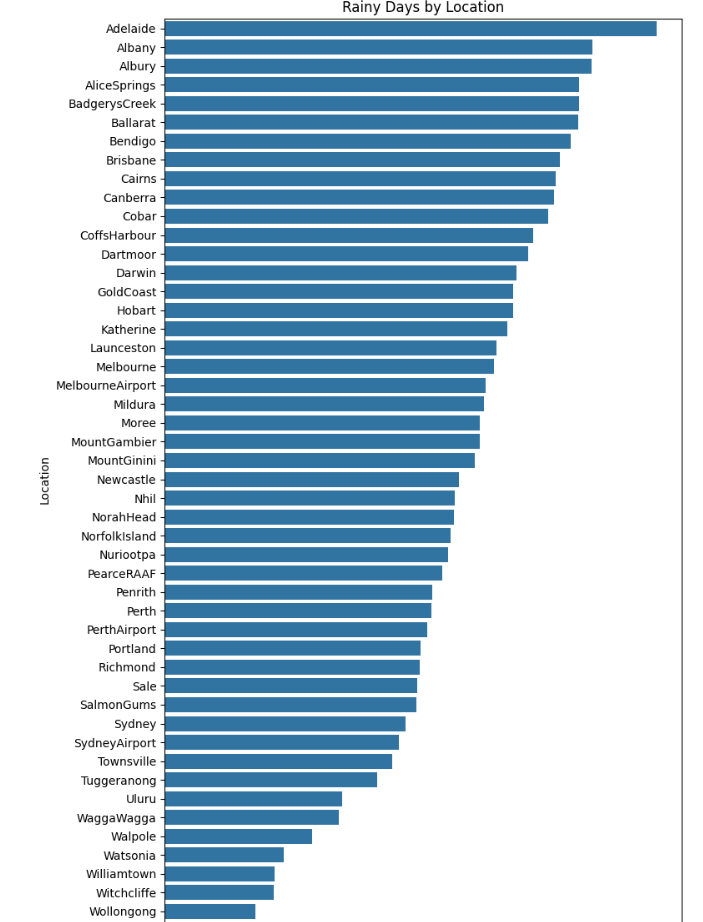
**Correlation**

****

Observations:

* Nothing surprising; all correlation as expected

**Distribution (Rain by Location)**

****

Observations:

* The above chart is useful for a quick check on the differences between locations regarding the number of rainy days

A lot of work was done for data preprocessing and cleaning because much data was missing or was in inappropriate format. All workflow can be seen in Jupyter Lab file.

1. **Models Applied and Why:** Predicting if it rains tomorrow in Australia is a classification problem (rain or no rain). I used *neural networks*, *logistic regression*, and *random forests*. Each of these models can provide different insights into the problem.

Neural Networks: Captures complex relationships between weather variables.

Automatically learns important featurerepresentations**.** The parameters were chosen randomly referring to the size of the dataset and numbers of features.

Logistic Regression: Fast and effective with large datasets. Effective starting point for binary classification. The default parameters without hyper tuning.

Random Forest: Handles large feature sets and reduces overfitting. Shows complex interactions between features. The default parameters without hyper tuning.

**A screenshot of a computer

Description automatically generated**

1. **The Results:**

**Neural Networks Performance**

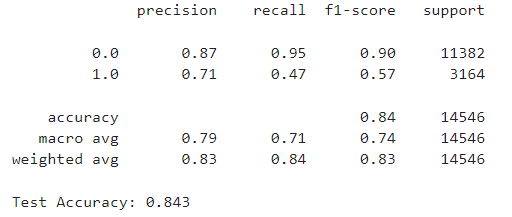
**A screenshot of a computer

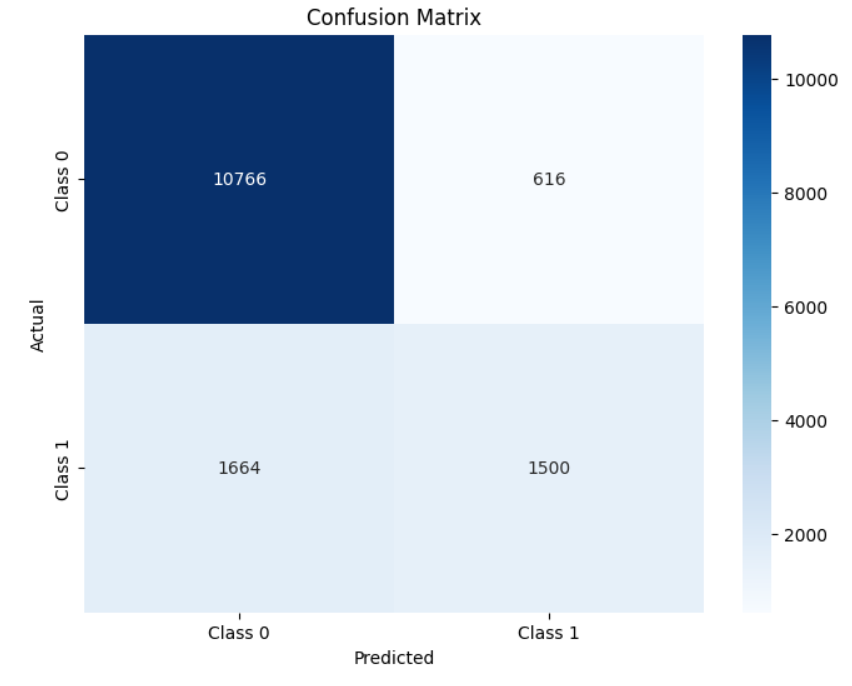
Description automatically generated**

**A blue squares with white text

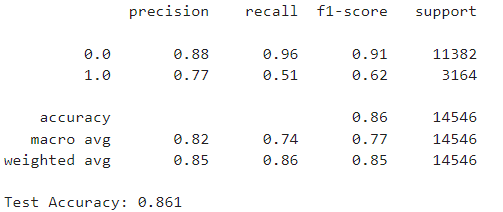
Description automatically generated**

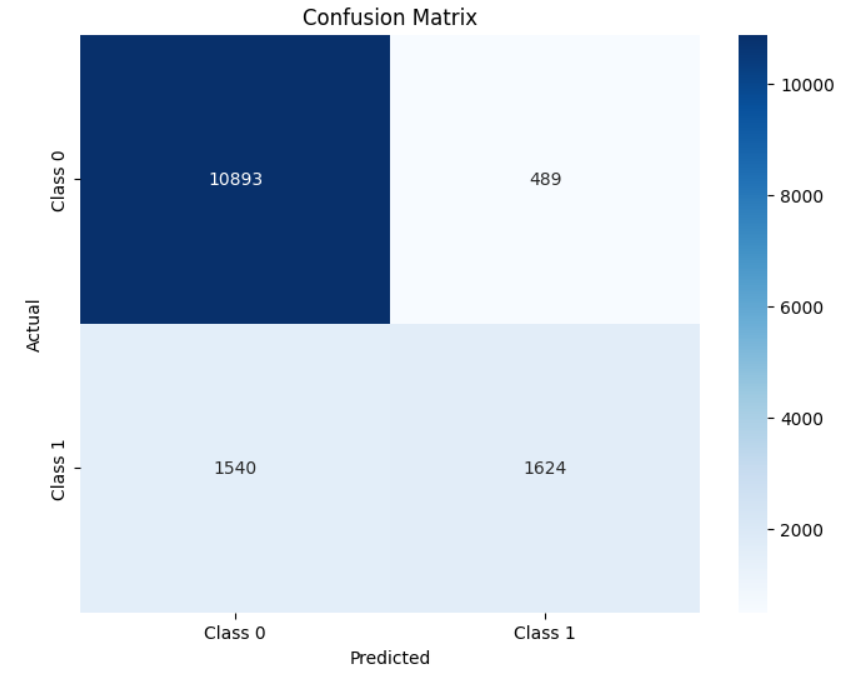
**Logistic Regression Performance**

****

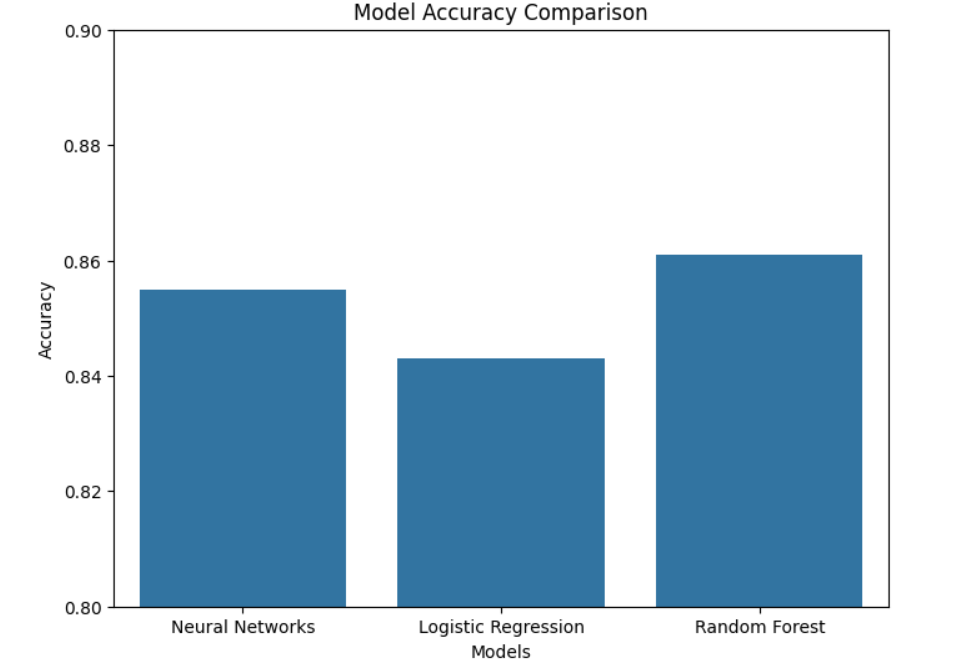
****

**Random Forest Performance**

****

****

**Models Comparing**

****

1. **Final Conclusion:** The models showed all similar performance around 85% of accuracy. My approach was decent, but many things can be improved. I had good results. To get better further results I can pay more attention to hyper tuning, better data preprocessing and more conscious approach to models’ parameters.