**Rain Prediction in Australia using XGBoost Classifier**

By Mushfiqur Rahman Khan Majlish

**Methods:**

1. Necessary libraries and datasets are imported. The dataset was downloaded from Kaggle.

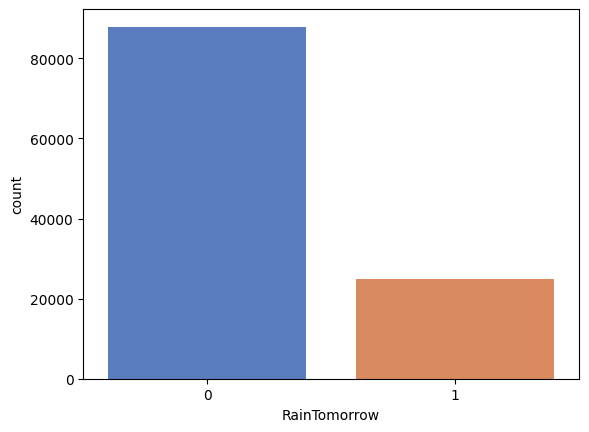
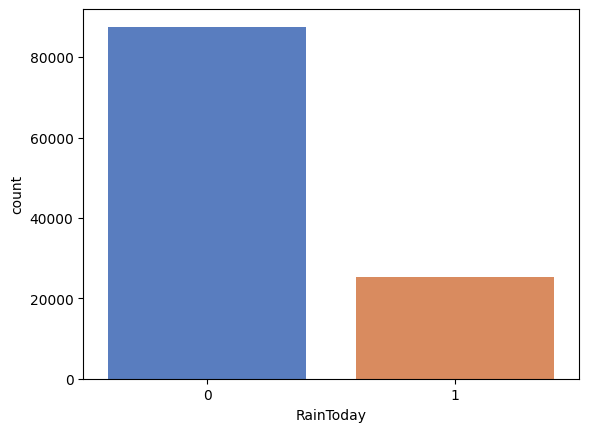
It contains 23 variables and 145460 observations. The dataset has missing values as well as several duplicated values.

1. The missing values are dropped from the dataset and duplicated values are removed. Now the dataset has 112925 observations and 17 variables. 6 variables are removed as majority of them contained null values.
2. The categorical variables are converted to numerical variables by label encoding.
3. Suitable graphical representations are used to notice and compare the pattern of variables.
4. Pearson correlation was calculated and presented in a heatmap.
5. XGBoost Classifier was applied which was used to predict the species.

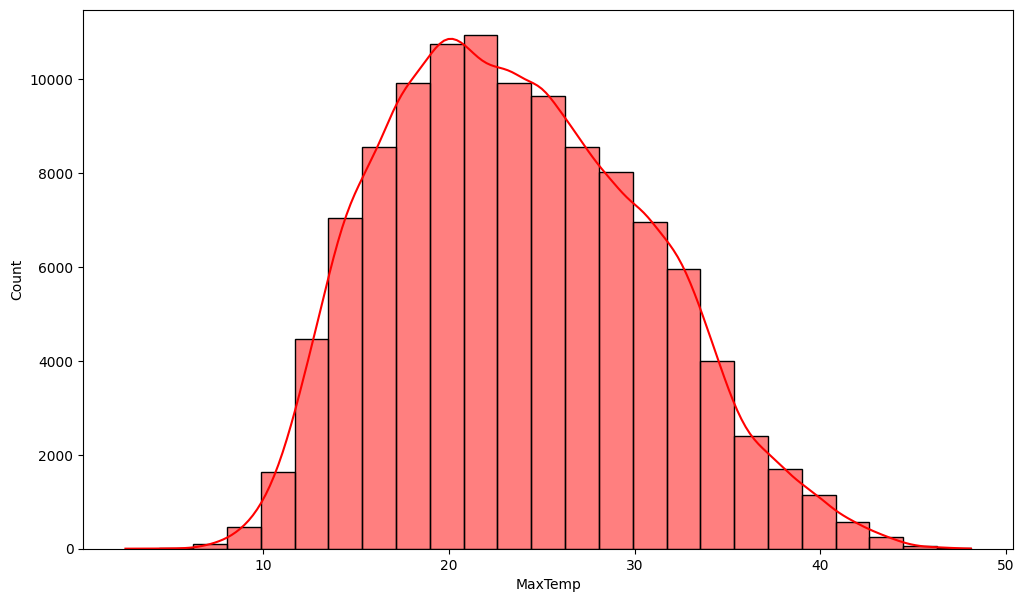
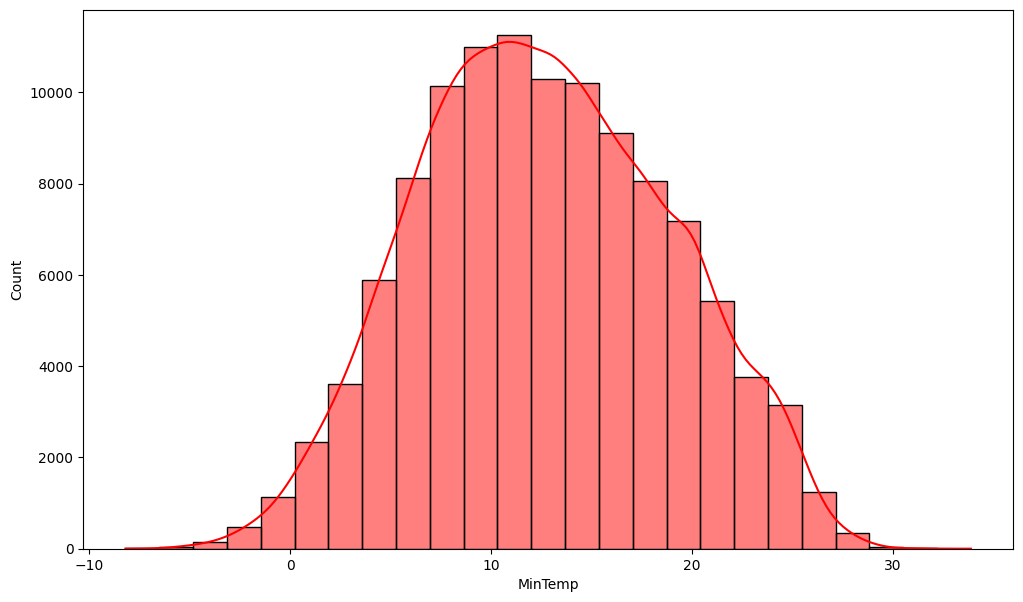
**Variable List:**

1. Dependent Variable: Species (species of Iris flower)
2. Independent Variable: MinTemp, MaxTemp, Rainfall, WindGustDir, WindGustSpeed, WindDir9am, WindDir3pm, WindSpeed9am, WindSpeed3pm, Humidity9am, Humidity3pm, Pressure9am, Pressure3pm, Temp9am, Temp3pm, RainToday

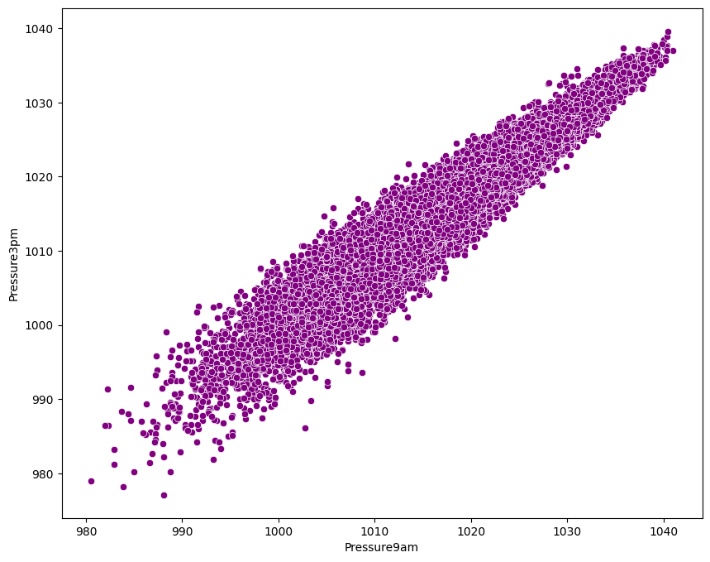
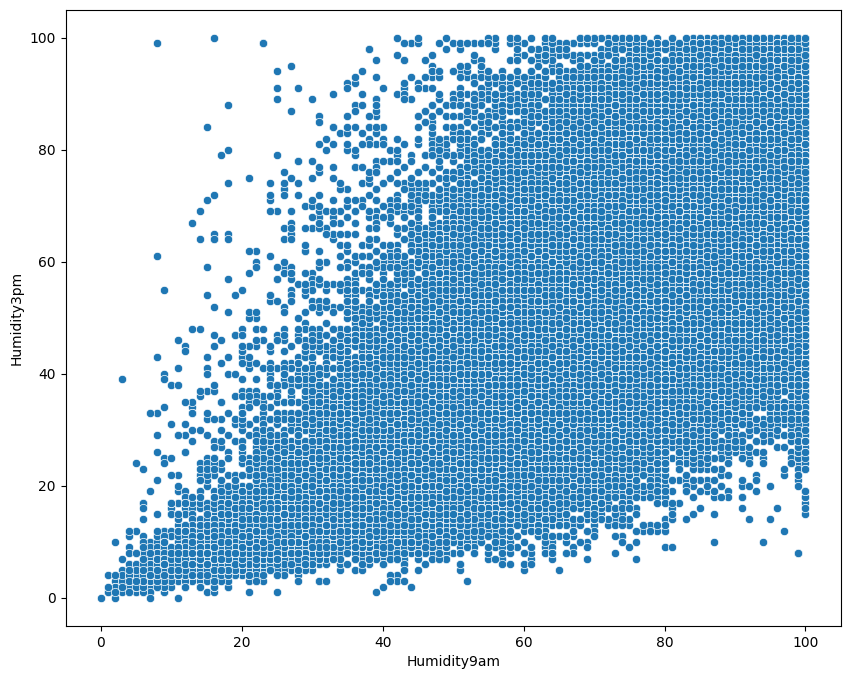
**Data Visualization:**



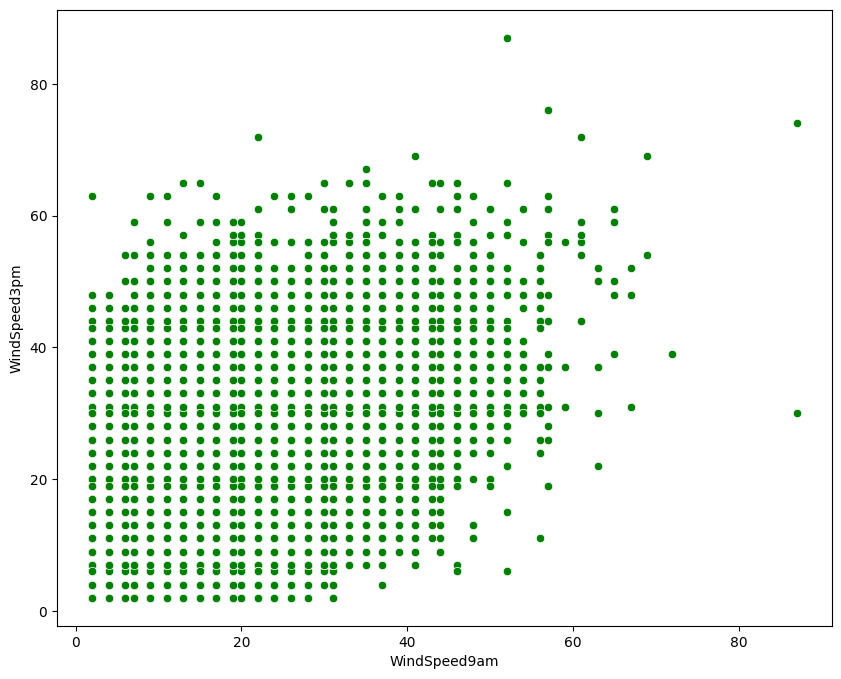
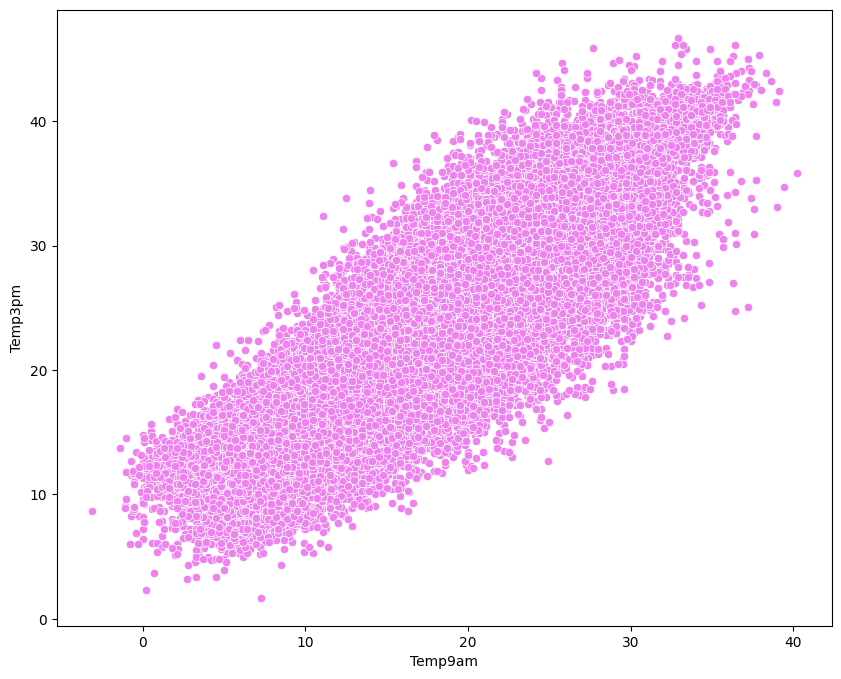
It can be observed from both graphs that, the count of rain today and rain tomorrow follow a similar pattern, it will not rain today and tomorrow (0) being highest in both cases.



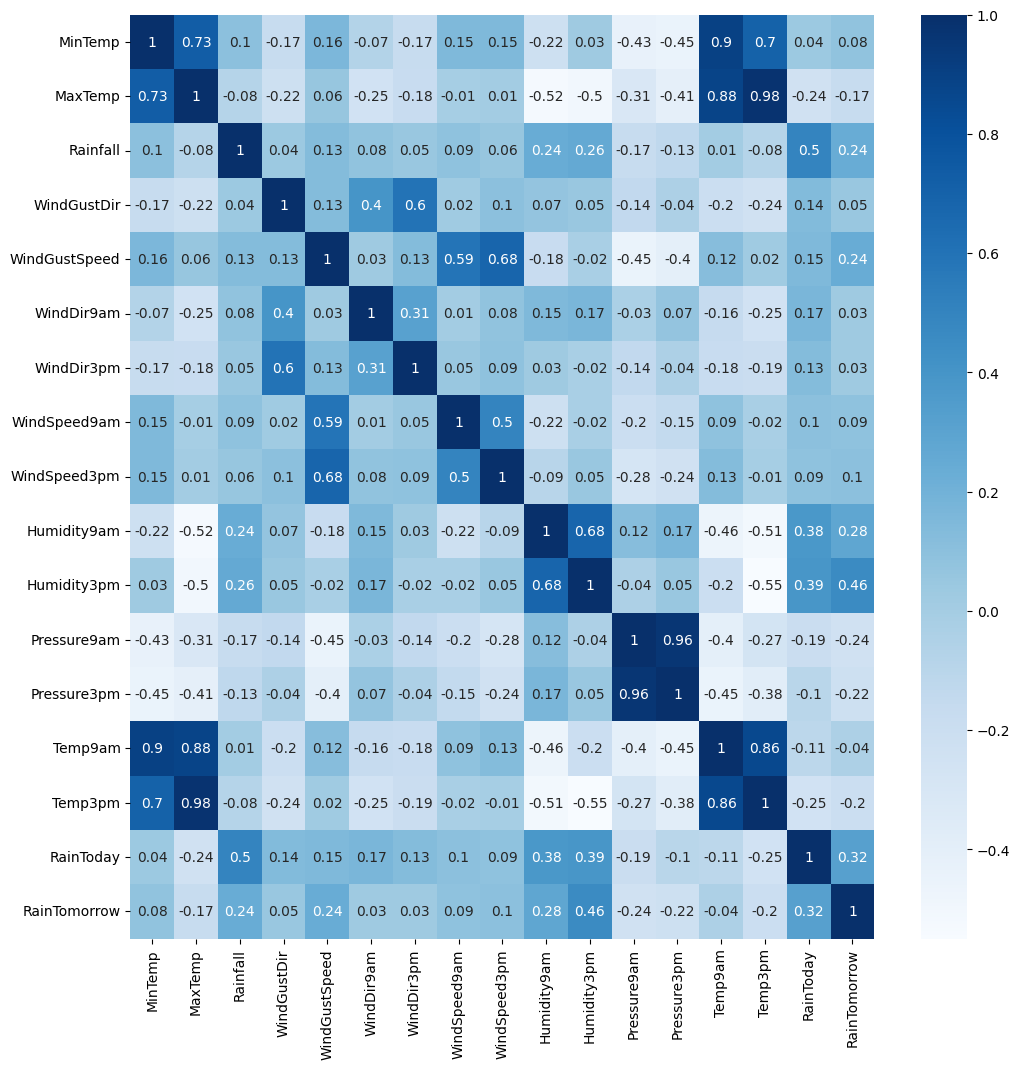
The histogram showcases similar pattern for minimum and maximum temperature. The highest minimum temperature is 8-15 whereas the highest maximum temperature 19-25.



The scatterplot showcases that there is positive relationship between humidity of two different times (9am and 3pm). This is also similar for pressure.



The scatterplot showcases that there is positive relationship between temperature of two different times (9am and 3pm). This is also similar for wind speed.



This is the heatmap which showcases the correlation of the numerical variables in the dataset. We can see from the heatmap that there are several strong and weak relationships (positive and negative) among the variables.

**Result:**

An XGBoost Classifier was employed to predict whether there will be rain tomorrow or not. 1 refers to the value of raining and 0 refers means there will be no rain tomorrow. Rain tomorrow is the dependent variable and rest are the independent variables. The model predicted with 85.8% of accuracy which means it a good fit.