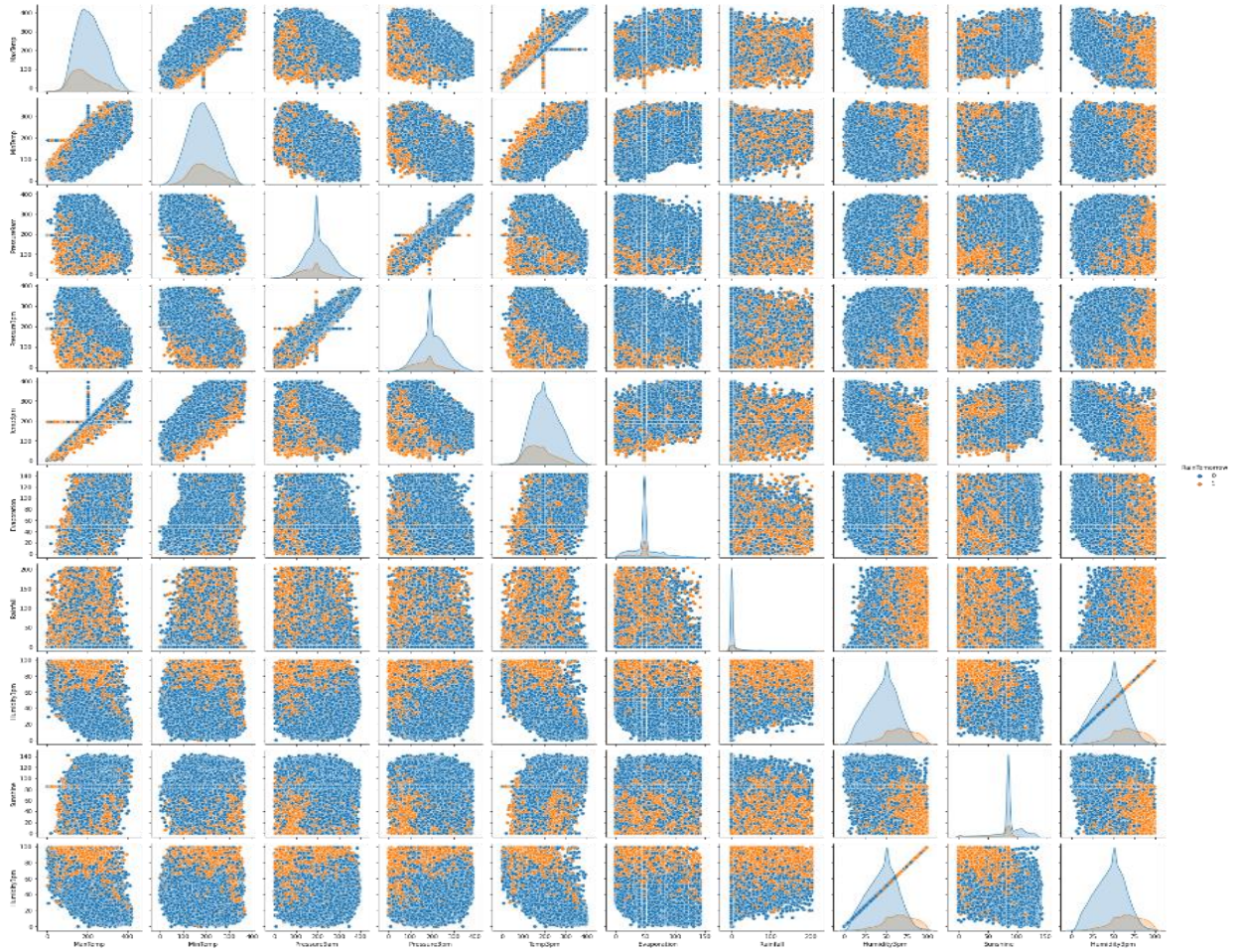
RAIN PREDICTION IN AUSTRALIA

Introduction

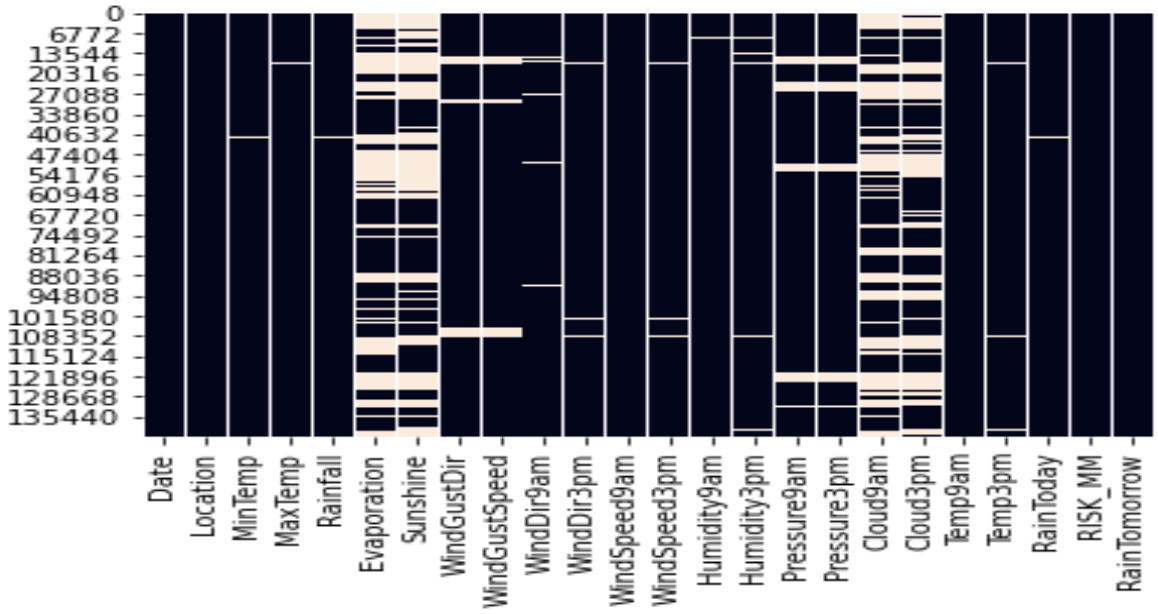
Weather is something that everyone has to deal with, and accurate data about it, such as what is on the way, can help users make informed decisions. People can use weather apps for iOS and Android to predict when the weather will change. Machine learning applications are used throughout the entire weather prediction workflow, which is divided into observations, data assimilation, numerical weather forecasting, and post-processing and dissemination. Machine learning could be used for everything from weather data monitoring to learning the underlying equations of atmospheric motions. In this assignment, we are given daily weather observation data from numerous Australian weather stations located across the country, covering 17 different regions, and our study aims to create a classification model that can predict whether or not it will rain tomorrow. The analysis aims to examine the impact of various parameters that could aid in forecasting the likelihood of rain the next day.

Data Understanding

The Australian weather dataset that we have been given contains 24 variables recorded for 1,42,193 observations that were observed and recorded at various weather stations located throughout the country.



This data is now being used as an input for our rain prediction classification model, which could assist us in determining the relationship between independent and dependent variables. Because the model's goal is to predict the probability of 'Rain Tomorrow,' which is a categorical variable, we must identify and transform all variables into the desired input for our consideration. Based on our observations, we found that our data contains only seven categorical variables, namely Date, Location, WindGustDir, WindDir9am, WindDir3pm, RainToday, and RainTomorrow, with the remaining 17 variables being continuous variables. Furthermore, prior to training the machine learning model, we must ensure that the data is free of missing / NA values, which are present in our case.

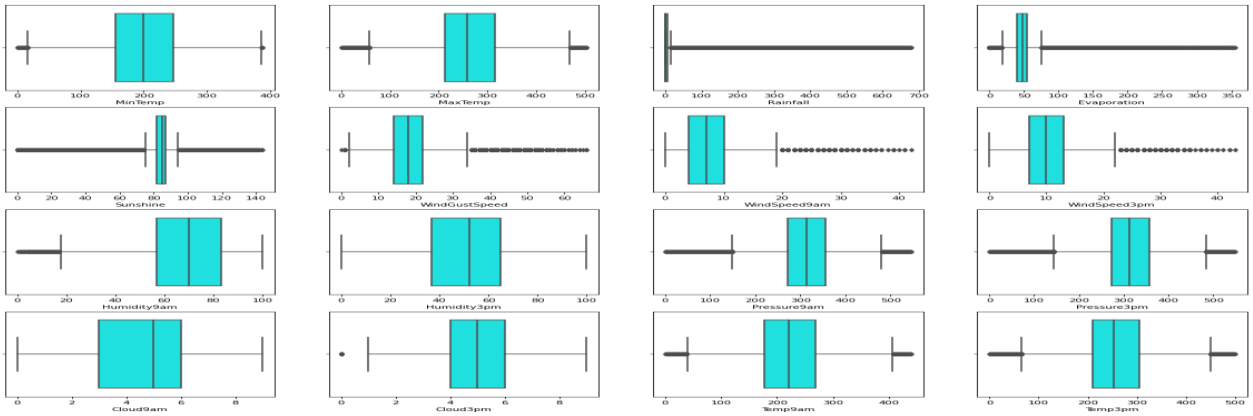


Data Exploration and Transformation

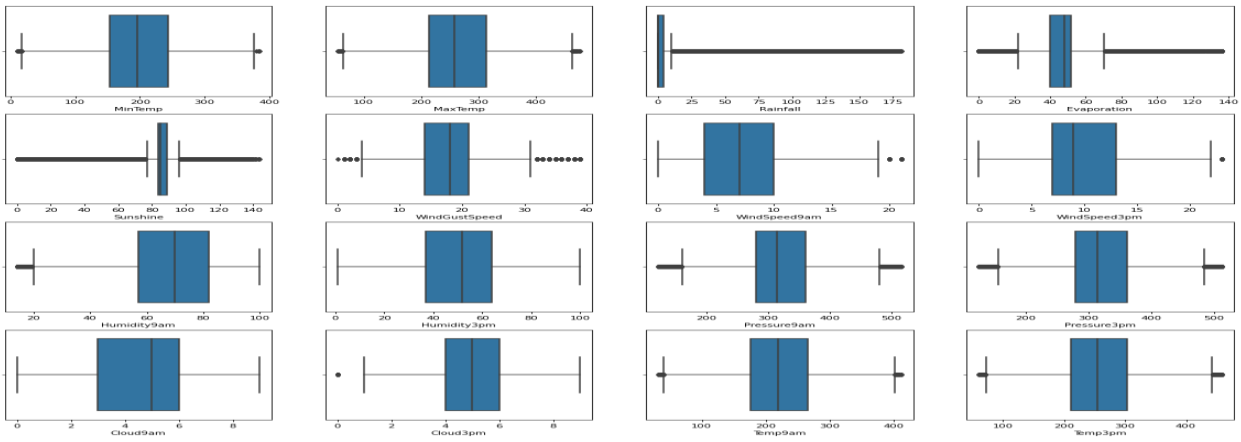
In order to make sure that the data that is being fed to the machine learning model needs to be processed to ensure the reliability of the model. Therefore, to deal with the data imputation problem, I have replaced the numerical imputation values with median values while categorical imputation is replaced with mode, the values that have been repeated most.

As we need to scale the data it is very important to deal with outliers. Moreover, in order to make sure that the data fed into the model is normally distributed and doesn't contain any extreme input values, the data has been treated by identifying and removing outliers present in the given data. A total of 13,740 values were identified as outliers and were removed during data preprocessing. Rather than using a percentile method, I performed outlier detection with Standard Deviation. All values under 3σ are considered for training the model.

In order to understand the impact of the given variables with respect to 'RainTomorrow', we tried to implement the sklearn Feature selection model to identify the most impactful variables. Based on the univariate selection method, we observed that 'Rainfall' is the most impactful variable while 'MinTemp' has the lowest impact in the determination of the Rainfall tomorrow.

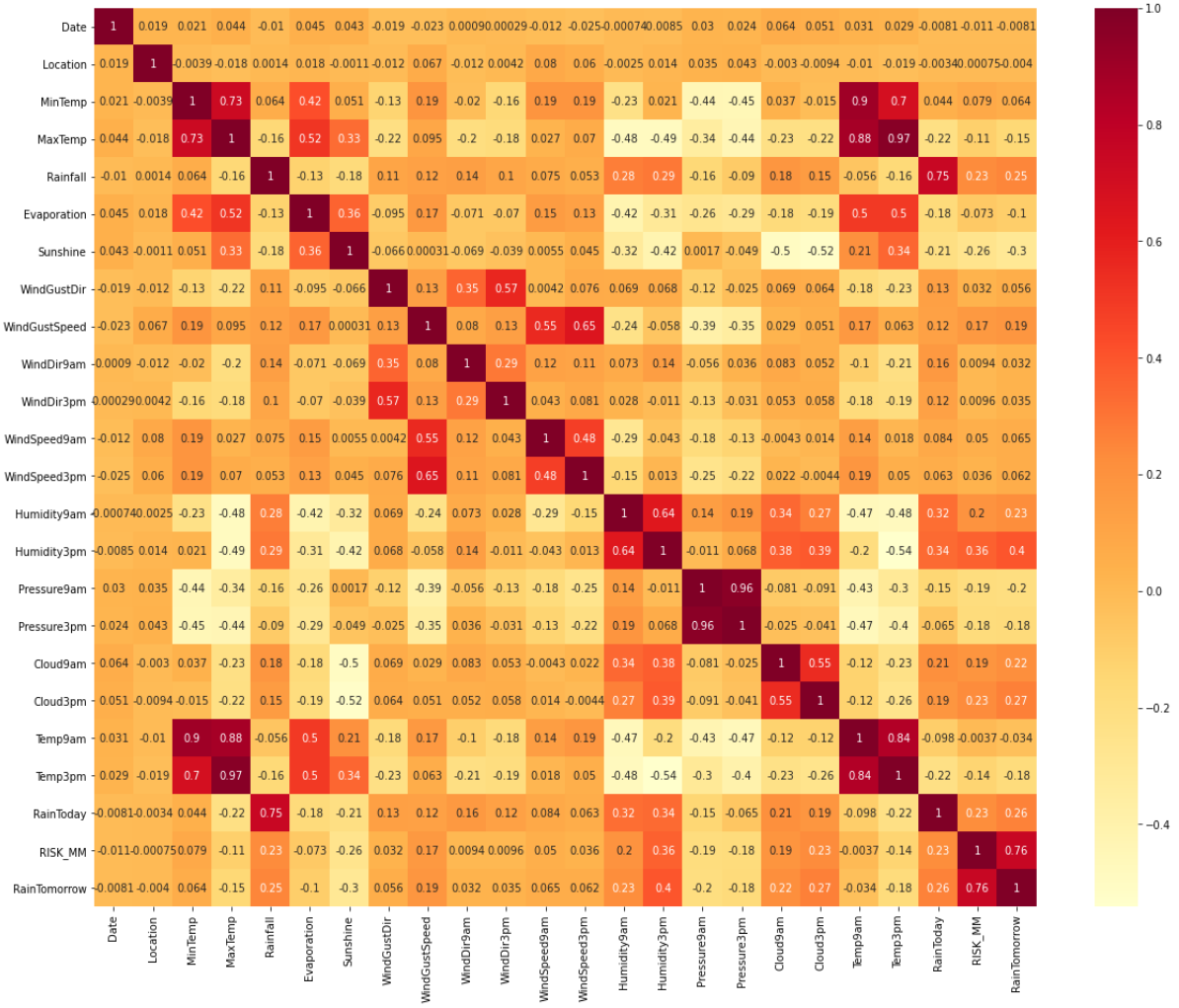


Before Outlier Removal



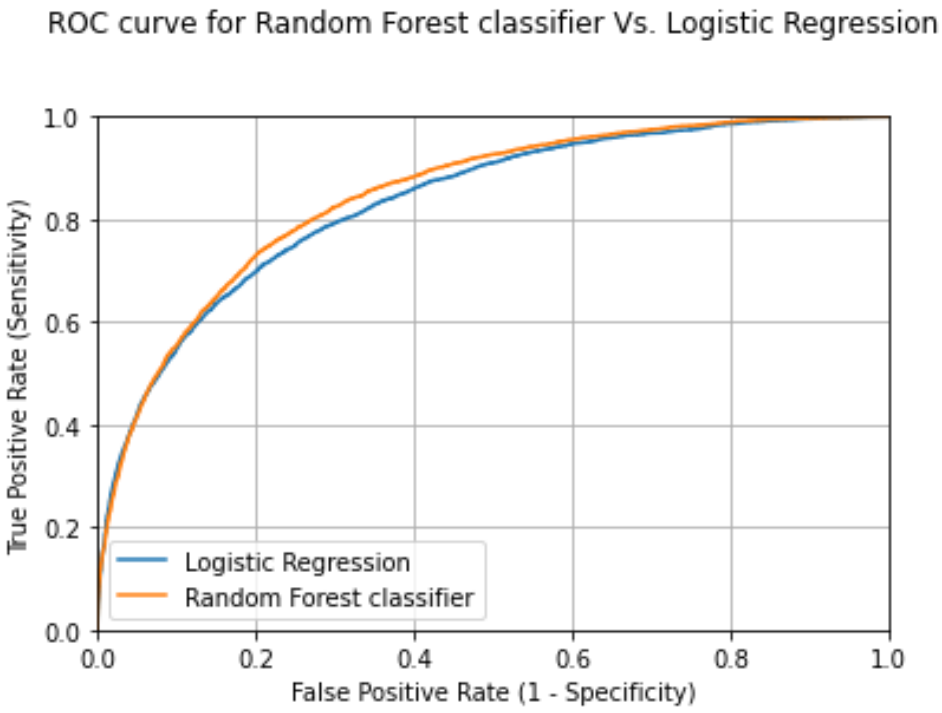
After Outlier Removal

However, Results from Correlation Matrix were not in sync with the findings from the previous method and identified 'Humidity3pm' (since we need to exclude RISK_MM from our model) as strongly correlated to the 'RainTomorrow' while 'WindSpeed9am' has the least correlation among all. Lastly, data has been rescaled using Standard Scaler from the sklearn preprocessing Library.



Modeling

In order to build an optimal classification model for the Australian weather rainfall prediction model, we used a Random Forest Classifier. The model was trained on 80% of data and 14 explanatory variables including 'Rainfall', 'Humidity3pm', 'Pressure9am', 'Pressure3pm', 'Sunshine', 'Temp3pm', 'MaxTemp', 'MinTemp', 'WindGustSpeed', 'RainToday', 'Cloud3pm', 'Humidity9am','Cloud9am' were used to predict the response of the dependent variable – 'RainTomorrow'. The model consistently had an accuracy above 85.79% which was validated using a K-Fold cross-validation technique. Apart from the Random Forest Classifier, the model's accuracy was also compared against other models such as Logistic Regression and Decision-Tree Classifier. Logistic Regression had lower accuracy than that of the Random Forest Classifier, which is 84.90%, however, Decision-Tree Classifier just turned out to be least accurate in this given situation. Lastly, Hyperparameter Optimization was conducted using 'GridSearchCv' for the chosen model in order to fine-tune the performance of the model.



Evaluation

Three different supervised learning models namely the Random Forest Classifier, Logistic Regression, and Decision-Tree Classifier have been used to make rain predictions from the given weather dataset. A number of performance metrics including Confusion Matrix, Accuracy Score, AUC ROC Curve, Precision Score, Recall Score, Specificity, and False-Positive Rate have been used to compare and evaluate the performance of the classification model. Based on results from various comparison markers Random Forest Classifier Model has better results for the given dataset, with a mean model accuracy score of 85.79% and has a higher ROC curve which indicates the relationship between sensitivity and specificity. Furthermore, the model has a classification error of 15.22% and a recall score of 27.55%, indicating that the model can correctly predict positive observations. Since the evaluation factors of the models are relatively precise, the above classification model may adopt to predict Rain tomorrow using Random Forest Classifier.

Interpretation of Result

- Accuracy Score: This is the ratio of correctly predicted observation to the total observation. [Calculated: 85.79%]
- ROC AUC Curve: ROC is a probability curve and AUC represents the degree or measure of separability. It tells how much the model is capable of distinguishing between classes. [Selected Model has a greater area under the curve]
- Recall Score: The recall is intuitively the ability of the classifier to find all the positive samples [Calculated: 27.55%]
- Specificity: Specificity is the metric that evaluates a model's ability to predict the true negatives of each available category [Calculated: 98.12%]

- Precision Score: The precision is intuitively the ability of the classifier not to label as positive a sample that is negative. [Calculated: 77.44%]

Conclusion

After analyzing the Australian Weather dataset, we were able to build a supervised machine model that uses a Random Forest Classifier model to predict the likelihood of rain tomorrow. The model can also identify and validate the existing relationship between the response variable 'RainTomorrow' and other variables such as Rainfall, Humidity, Sunshine, Temperature, and more. Although the model's accuracy can still be improved by properly treating outliers, missing values and other parameters can also be hyper-tuned to improve performance. However, once this issue is resolved, this model may be useful in accurately predicting rainfall in Australia..