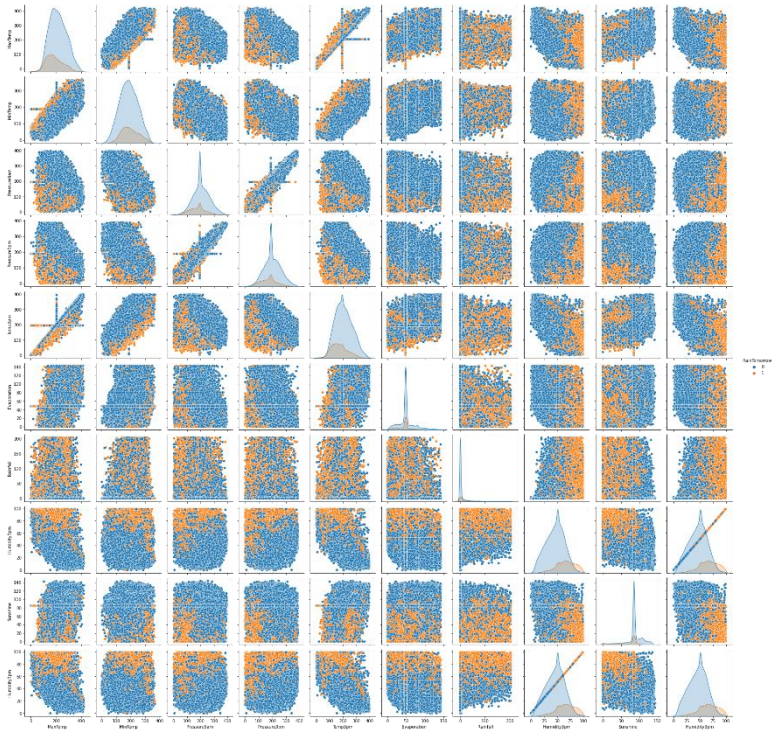
RAIN PREDICTION MODEL

Introduction

Weather is something everybody deals with, and accurate data of it like what is coming can help users to make informed decisions. With weather apps for iOS and Android, people can exactly know when to expect a change in the weather conditions. Machine learning applications are spread all over the entire workflow of weather prediction breaking that workflow down into observations, data assimilation, numerical weather forecasting, and post-processing and dissemination. Across those areas, machine learning could be used for anything from weather data monitoring to learning the underlying equations of atmospheric motions. In this assignment, we are presented with the daily weather observation data recorded by numerous Australian weather stations located all across the nation covering 17 different regions and our analysis aims to develop a classification model that could predict whether it will Rain Tomorrow or not? The analysis aims at analyzing the effect of the various parameter that could help in forecasting the probability of rain the next day.

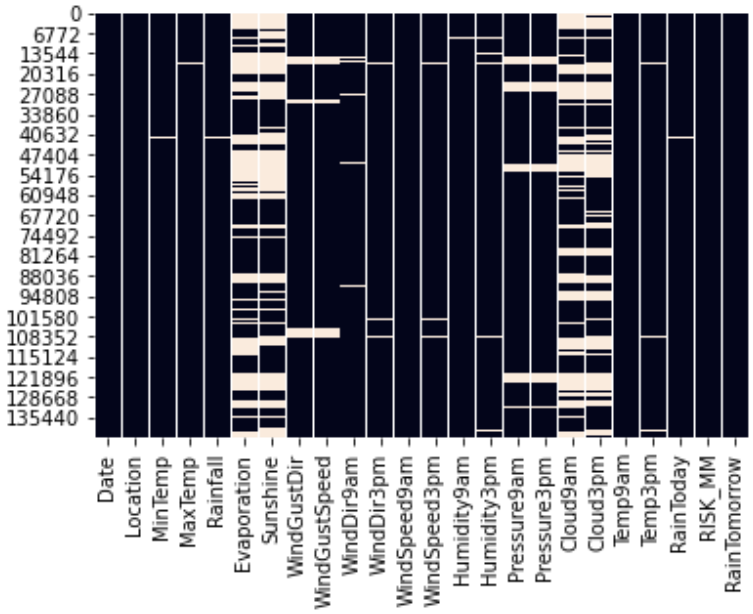
Data Understanding

The Australian weather dataset that we are given contains 24 variables recorded for 1,42,193 observations, which were observed and recorded at different weather stations located in different parts of the country.



This data now serves as an input for our rain prediction classification model that could help us identify the relationship between independent and dependent variables. Since, the model aims

at predicting the probability of the 'Rain Tomorrow', which is a categorical variable, we need to identify and transform all variables into desired input for our consideration. Based on the observations, we identified that only Seven categorical variables i.e., Date, Location, WindGustDir, WindDir9am, WindDir3pm, RainToday, and RainTomorrow are present in our data, and the rest all remaining 17 variables are continuous variables. Moreover, before training the machine learning model, we need to ensure that the data doesn't contain any missing / NA values, which in our case is present.

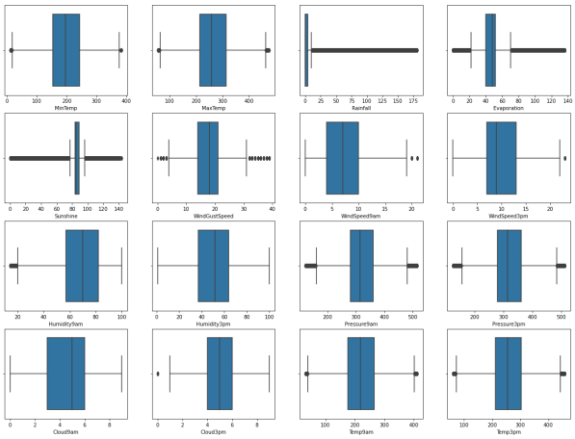


Data Exploration and Transformation

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.

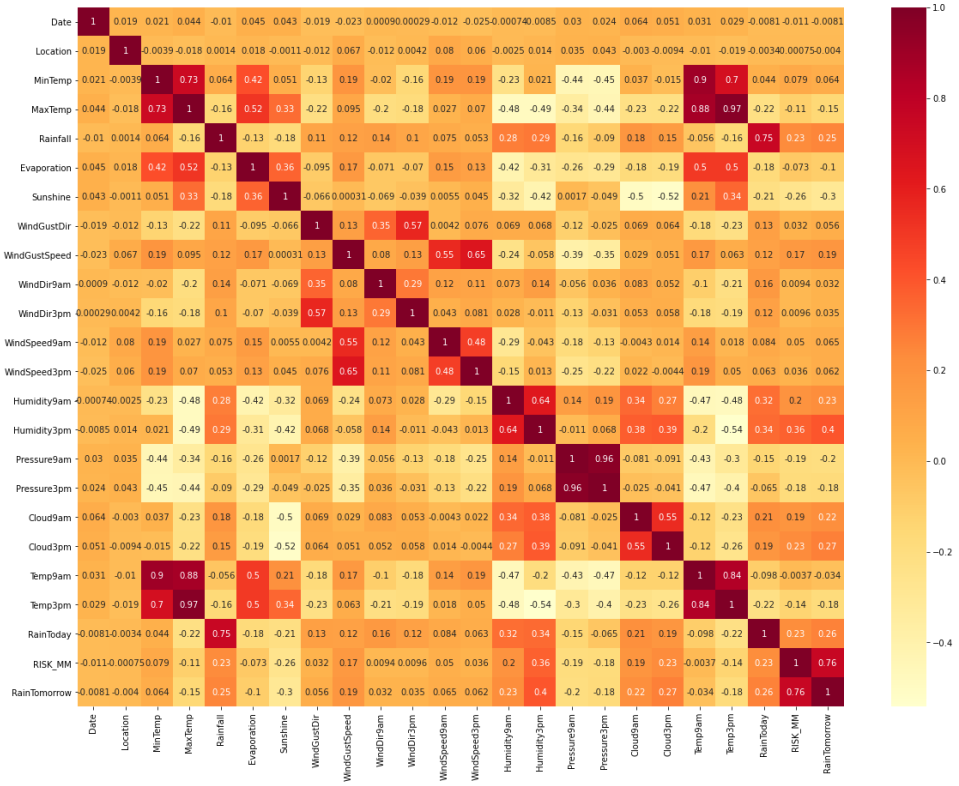
As we need to scale the data it is very important to deal with outliers. Moreover, in order to make sure that the data fed into the model is normally distributed and doesn't contain any extreme input values, the data has been treated by identifying and removing outliers present in the given data. A total 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.



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.



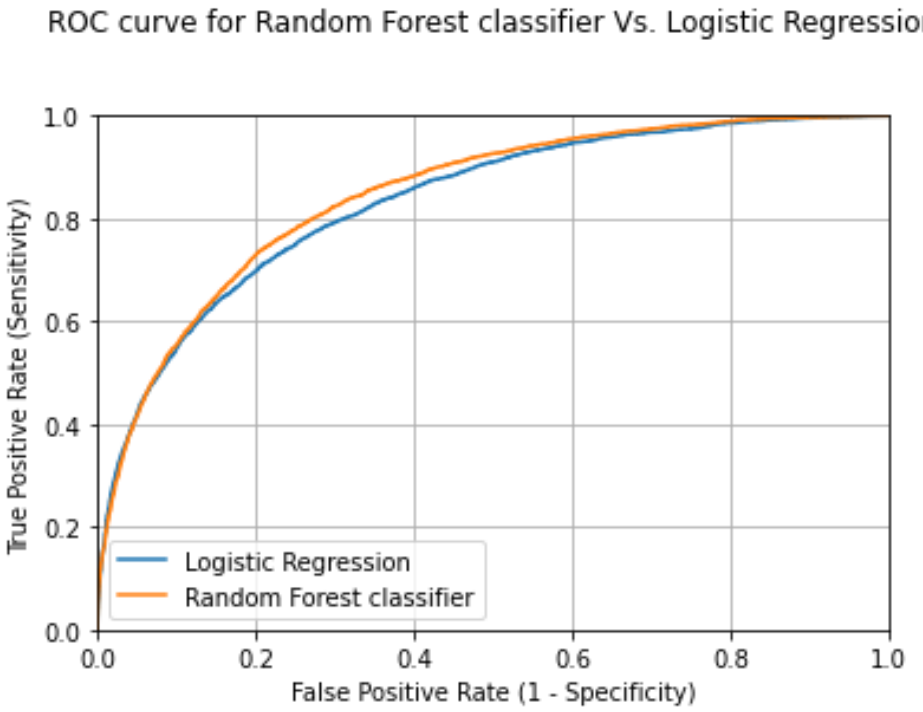
Modeling

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',

Model	Accuracy	Balanced Accuracy	ROC AUC	F1 Score	Time Taken
NearestCentroid	0.76	0.74	0.74	0.78	0.13
GaussianNB	0.80	0.71	0.71	0.80	0.14
QuadraticDiscriminantAnalysis	0.82	0.71	0.71	0.82	0.19
XGBClassifier	0.86	0.70	0.70	0.84	3.06
LGBMClassifier	0.86	0.69	0.69	0.84	0.61
RandomForestClassifier	**0.86**	0.69	0.69	0.84	16.96
BernoulliNB	0.79	0.69	0.69	0.80	0.14
LinearDiscriminantAnalysis	0.85	0.68	0.68	0.83	0.27
ExtraTreesClassifier	0.86	0.68	0.68	0.84	10.34
KNeighborsClassifier	0.84	0.68	0.68	0.83	24.00
AdaBoostClassifier	0.85	0.68	0.68	0.83	3.74
LogisticRegression	**0.85**	0.67	0.67	0.83	0.35
BaggingClassifier	0.85	0.67	0.67	0.83	5.88
CalibratedClassifierCV	0.85	0.67	0.67	0.83	77.00
DecisionTreeClassifier	0.79	0.67	0.67	0.79	0.97
SVC	0.86	0.66	0.66	0.83	357.93
LinearSVC	0.85	0.66	0.66	0.83	21.45
ExtraTreeClassifier	0.78	0.65	0.65	0.78	0.19
RidgeClassifier	0.85	0.64	0.64	0.82	0.17
RidgeClassifierCV	0.85	0.64	0.64	0.82	0.25
SGDClassifier	0.84	0.60	0.60	0.80	0.42
Perceptron	0.81	0.58	0.58	0.78	0.26
PassiveAggressiveClassifier	0.76	0.57	0.57	0.75	0.27
DummyClassifier	0.69	0.50	0.50	0.69	0.10

The outcome of a given dataset that has been trained on 29 distinct classification models

'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 construct a supervised machine model that leverages a Random Forest Classifier model to predict the likelihood of a Rain Tomorrow. The model is also able to identify and verify the existing relationship that exists between the response variable 'RainTomorrow' and other variables such as Rainfall, Humidity, Sunshine, Temperature, and more. Although the accuracy of the model can still be improved by the proper treatment of the outliers, missing values and other parameters can also be hyper-tuned to improve performance. However, once this issue is rectified, this model could assist in effectively predicting the Rainfall in Australia.