PREDICTING RAIN IN AUSTRALIA

CPTS 437 MACHINE LEARNING

AYLEAH HILL

HYPOTHESIS

Given a dataset that has 10 years of weather forecasts in Australia, what is the effect/results of different prediction models that are trained on a series of days on predicting if it will rain in Australia the following day?

The Random Forest Classifier is hypothesized to perform the best.

OBJECTIVES

- Determine what kinds of data cleaning must be done to have a suitable data set to work with
- Determine what prediction models best predict rain in Australia the next day
- Determine which prediction model is the best based on its runtime and results

DATA

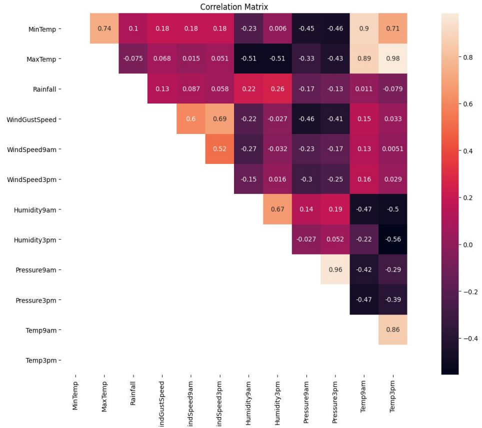
The dataset has 10 years of weather observations from cities across Australia. The data has 23 columns: 1 column with the date, 21 columns with weather data and 1 column that indicates whether it rains tomorrow. Links for data are below.

METHODS

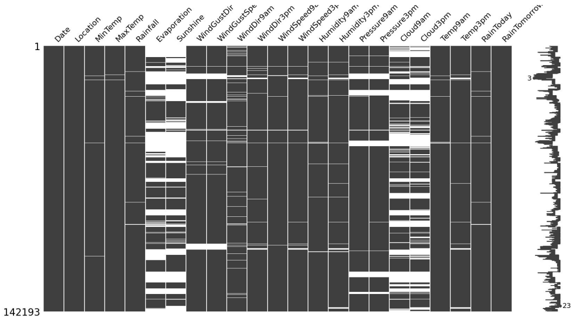
Exploratory Data Analysis was used initially to clean and understand the data. The data was cleaned of duplicates, missing data/labels, highly correlated features, and outliers. The data also needed to be rebalanced (via oversampling) to improve accuracy since most of the data is categorized as “No rain tomorrow” and only a few instances say “Yes rain tomorrow”.

Ran a series of models on the cleaned data: **AdaBoost, Decision Tree, Logistic Regression, Multilayer Perceptron, Naive Bayes Classifier, Random Forest Classifier, Support Vector Machine, Voting Classifier**.

Following each classifier, performance measures were made with ROC curves,Confusion Matrices, and F1 score, which were then compared.



Correlation Matrix in preprocessing where features that are heavily correlated are brighter. Features that were heavily correlated were removed.



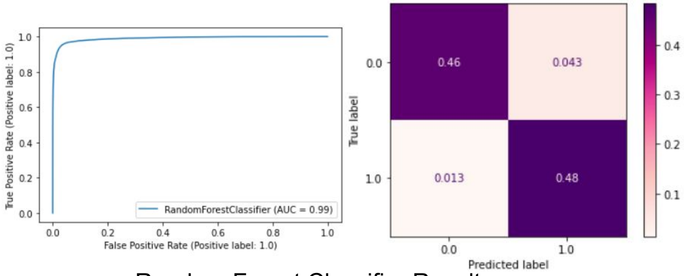
Missing Data chart where white space indicates missing data.

RESULTS

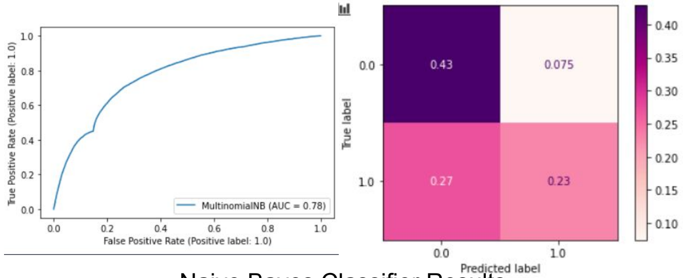
Results showed varying performances between the classifiers with the lowest performance being from the Naive Bayes Classifier with an F1 score of 0.64 while the Random Forest Classifier performed the best with an F1 score of 0.94 and ran the fastest of all classifiers. The Voting Classifier had an F1 score of 0.91.

CONCLUSION

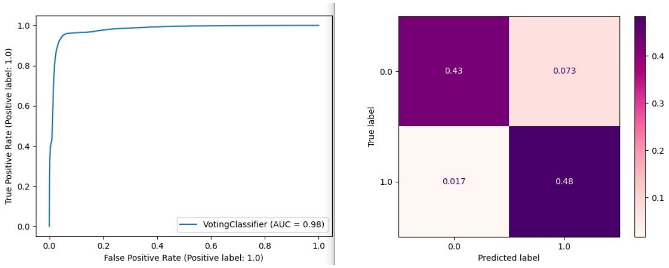
Some models work better than others at being able to predict if it will rain in Australia the following day based on the weather forecast of the current day. Using an ensemble method such as a Random Forest Classifier proves to be the best method for making predictions in this context.



Random Forest Classifier Results



Naive Bayes Classifier Results



Voting Classifier Results

The dataset was provided on Kaggle and the data was sourced by the Australian Government’s Bureau of Meteorology:
Kaggle: https://www.kaggle.com/jsphyg/weather-dataset-rattle-package Data source: http://www.bom.gov.au/climate/dwo/ http://www.bom.gov.au/climate/data