Ali Malenchik

Final Project Report

March 6, 2021

Executive Summary

Rainfall prediction is vital to many industries in strategic planning and effective resource management. The goal of this project was to perform accurate short-term binary rainfall prediction using machine learning techniques. Using historical records of climate and atmospheric parameters including temperature, rainfall, wind direction, wind speed, humidity, pressure, and cloud cover, I attempted to predict whether or not it will rain on a particular day for a particular city in Australia. Three cities were considered using data collected by Australia's Bureau of Meteorology: Darwin, Perth, and Brisbane.

Multiple algorithms were trained separately on each city's weather data, including Logistic Regression, Random Forest, and Naïve Bayes. The models were then tested using reserved unseen data to make predictions, and key metrics were compared in order to evaluate the performance and determine the best model.

Interestingly, the results of model evaluation showed that model performance differed depending on the city, and the features utilized in the model varied by city as well. For the cities of Perth and Brisbane, the best model was found to be Logistic Regression, while the city of Darwin demonstrated the best performance using Random Forest. This indicates that no singular model would necessarily perform well on all cities across the board.

Introduction

Background

Having an accurate forecast of upcoming weather conditions is important in many industries. For example, agriculture relies on precise weather predictions in order to efficiently plan when to plant, irrigate, and harvest. Other businesses such as construction, airlines, and outdoor sporting events are also dependent on weather predictions to determine when they can operate without disturbance. Using historical weather data, predictive analytics helps weather services to provide precise weather forecasts and allows businesses to make strategic decisions as a result.

Problem Statement

Can the occurrence of rain be predicted using information about the previous day's weather? In other words, can one confidently answer the question, "Will it rain tomorrow?"

Scope

This project will utilize Python to explore several binary classification algorithms using collected weather data to predict whether or not it will rain the following day. The models will be optimized and evaluated using multiple metrics in order to select the best performing classifier. It will also be determined which characteristics or features are relevant in the prediction of rain. It should be noted that the data used in this project includes weather from Australia only, and therefore the resulting predictive model may only apply to specific cities of Australia and not other climates.

Methods

Data Source

The dataset used for this analysis was obtained from Kaggle and contains approximately 10 years of daily weather observations from various locations across Australia. Climate and atmospheric data was gathered from numerous weather stations' real-time systems in an automated fashion and consolidated by Australia's Bureau of Meteorology. Observations include daily minimum and maximum temperatures, rainfall, wind, evaporation, sunshine, humidity, pressure, and other weather characteristics. The target variable for prediction is RainTomorrow, a binary variable indicating whether or not it rained the day following the observed weather features. RainTomorrow is denoted as Yes if it rained at least 1mm the day following the gathered observations.

Data Import & Cleansing

Comma-separated data was loaded into a data frame using Pandas read_csv function. The data frame consisted of 145,460 entries at time of import. Multiple cleaning and pre-processing steps were taken.

1) **Additional columns:** In order to create a more useful feature for prediction, the Date column was used to derive the Month of the observation using Pandas DatetimeIndex function. Including the month as a feature allows us to consider seasonal weather trends.
2) **Drop Missing Values:** Since the size of the dataset is extremely large, and missing data is most likely due to a lack of data collection for specific metrics rather than missing at random, it was decided to drop records containing missing values and then utilize cities with the most complete dataset. In total, 89040 rows having a missing value were dropped using the dropna() function. After removing the records containing missing

values, I obtained the record count for each city to determine the top three locations having the most data available for use in modeling:

Location	Count
Darwin	3062
Perth	3025
Brisbane	2953

3) **Encoding Target Variable:** The target variable, “RainTomorrow”, was transformed to binary values for use in modeling. Values of “Yes” were reassigned to 1, and values of “No” were reassigned to 0.

4) **Data Frame Split:** The data was split into three separate data frames, df_Darwin, df_Perth, and df_Brisbane. This allows us to consider each location independently during exploratory analysis, modeling, etc.

5) **Test-Train Split:** Each city’s data frame was then split into test and train using sklearn’s test_train_split. 75% of the data was allocated for the training of data, while 25% was reserved for later use in testing the model performance. After the split, the size of the test and train datasets were calculated:

Location	Count (x_train)	Count (x_test)
Darwin	2296	766
Perth	2268	757
Brisbane	2214	739

6) **Encoding Features:** The independent categorical features (RainToday, Location, WindGustDir, WindDir9am, WindDir3pm) were encoded using Pandas get_dummies function. Original columns were then dropped to prevent redundancy in the dataset. Note: this activity was performed after graph analysis to avoid the need for re-labeling. After this encoding, the training data frame structure increased to 66 columns.

Exploratory Analysis

Outliers

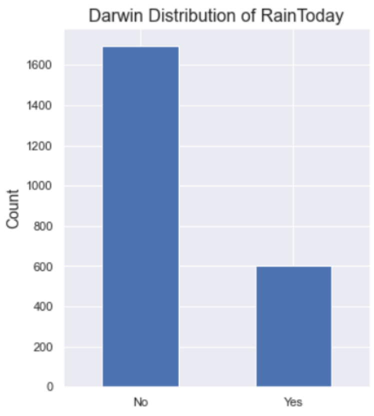
Outlier analysis was performed on numeric variables within each data frame, which revealed that the Rainfall feature contained the highest number of outliers across each city. The top 3 columns having outliers for each city are displayed below.

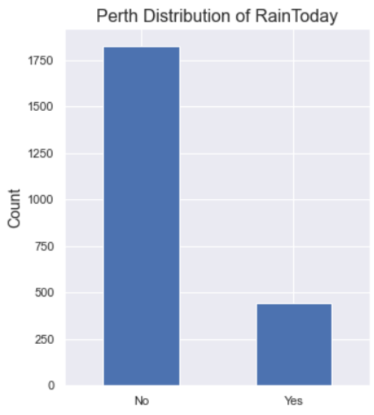
Location	Column	Number of Outliers
Darwin	Rainfall	440
	Sunshine	127
	Humidity9am	120
Perth	Rainfall	510
	WindGustSpeed	45
	Humidity3pm	31
Brisbane	Rainfall	447
	Sunshine	112
	Humidity3pm	109

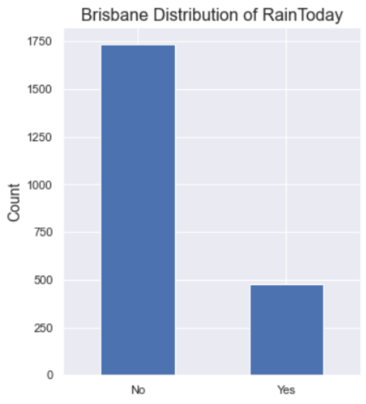
Plots

As part of exploratory data analysis, descriptive statistics were calculated, and univariate & bivariate plots were created to observe the distribution of variables. Using the plots and statistics, the following observations were made:

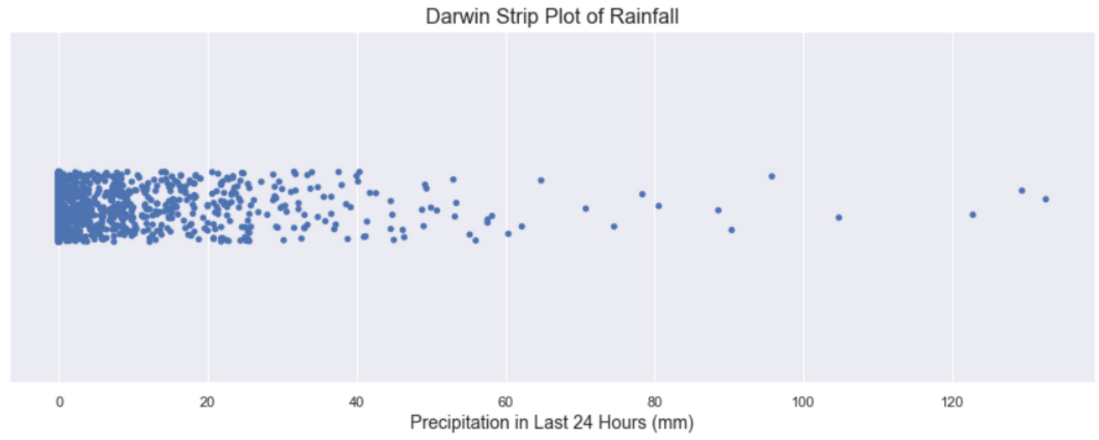
1) Using the histograms of RainToday, it can be concluded that the distribution is imbalanced for all 3 cities, with a majority of data points having a value of "No".

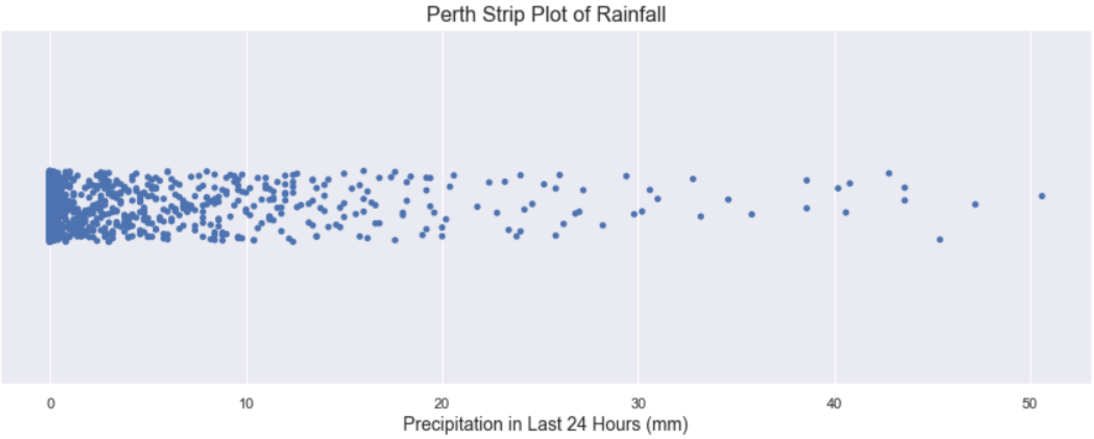


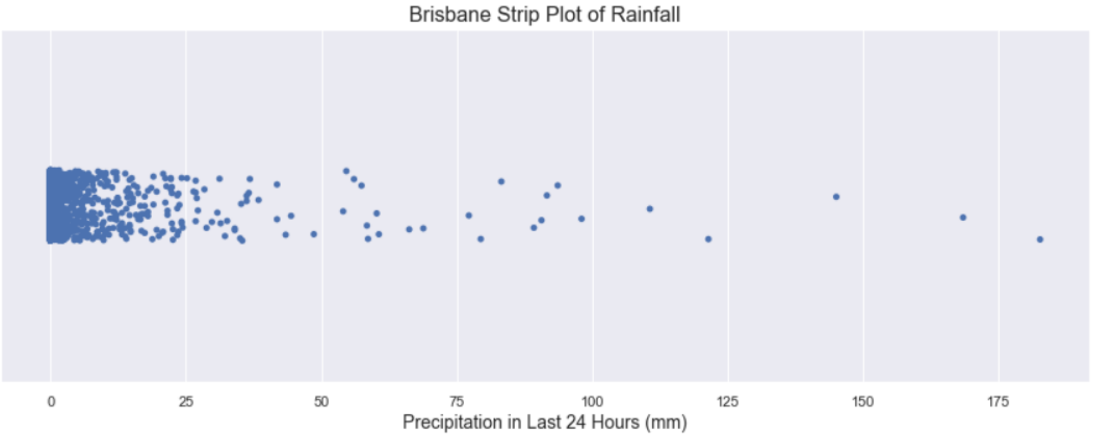




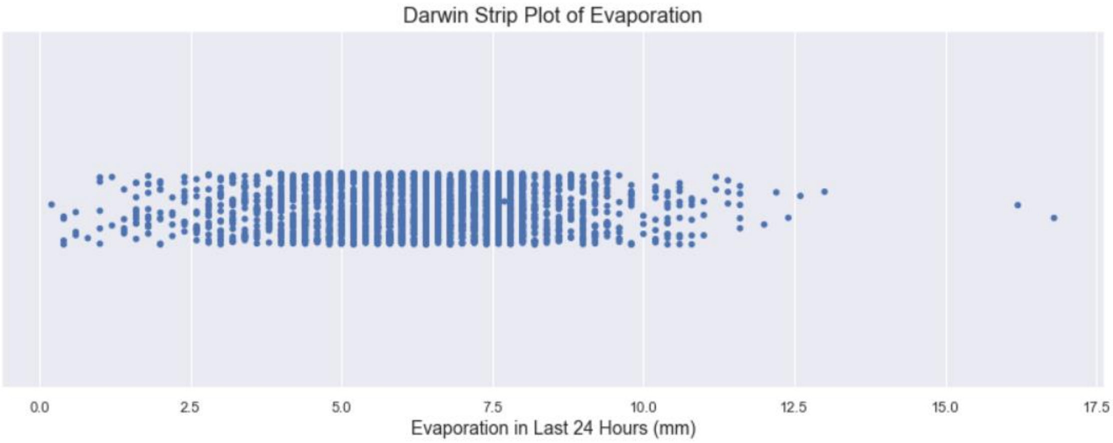
2) Using the strip plots of Rainfall, it can be concluded that all three cities generally saw minimal rainfall on most days, as the data points are densely clustered below 10 mm of precipitation in a 24 hour period. Perth saw the least amount of precipitation on a given day, with a maximum of only 50.6 mm, whereas Darwin and Brisbane had wider ranges and saw maximums of 132.6 and 182.6, respectively.

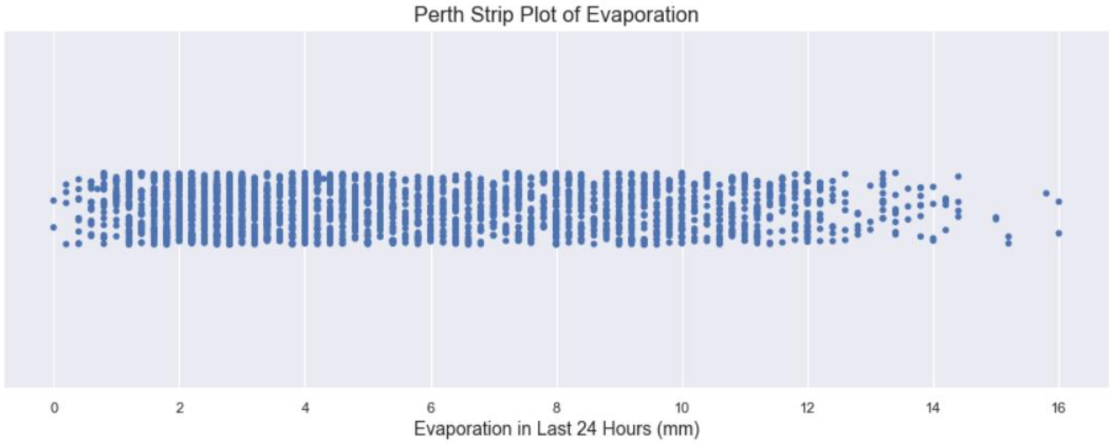


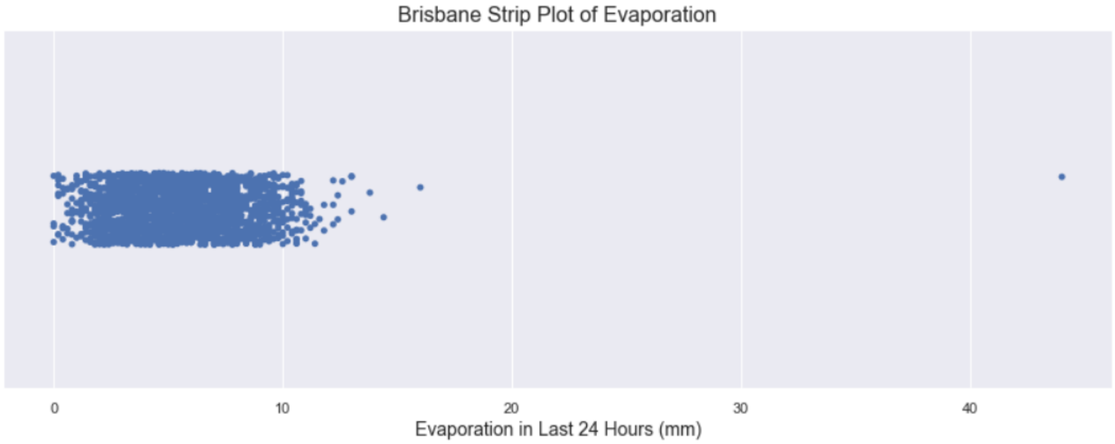




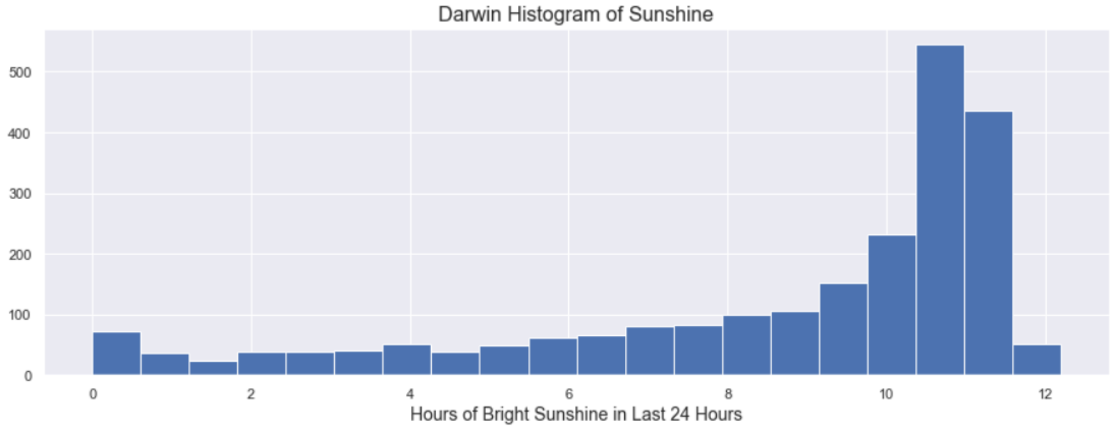
3) Using the strip plots of Evaporation, it was observed that most values were relatively evenly spread between approximately 0 and 12 mm. Darwin and Perth saw a few outliers closer to the 16 mm range, and Brisbane saw the highest value at 44 mm.

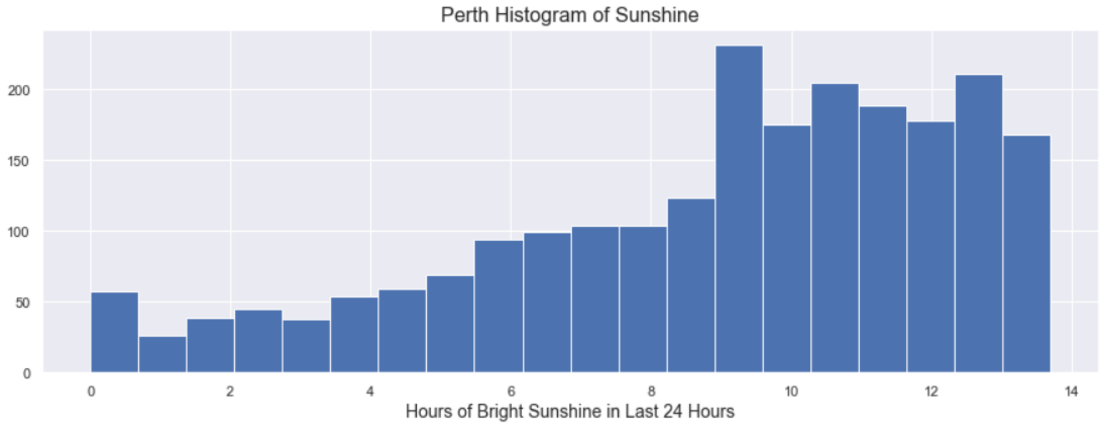


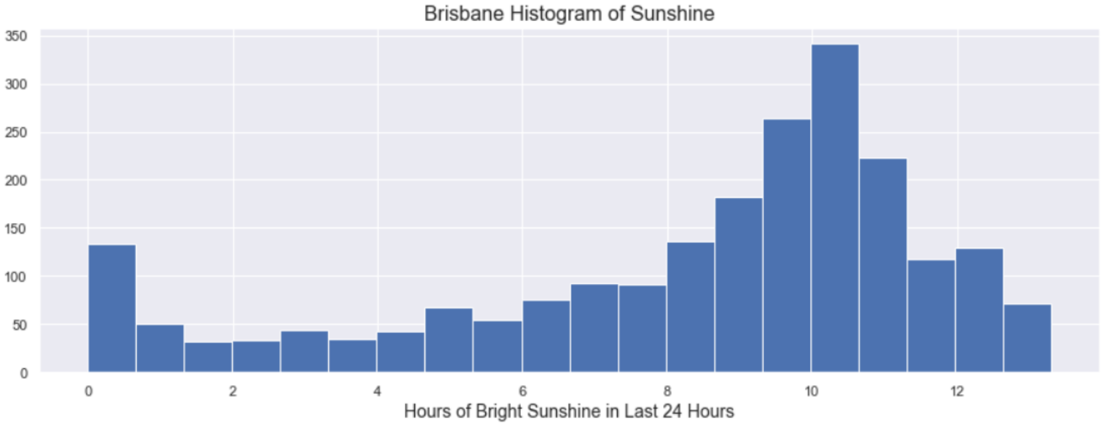




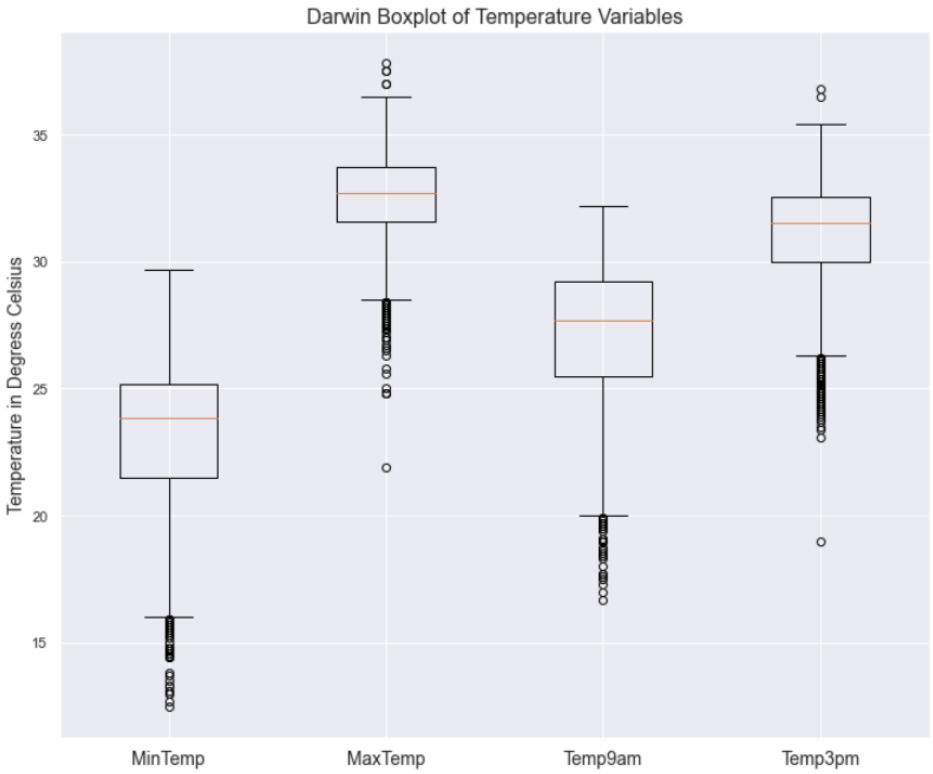
4) The histograms of Sunshine reveal that the distribution is negatively skewed for all three cities, with the average duration of bright sunshine between 8 and 9 hours per day.

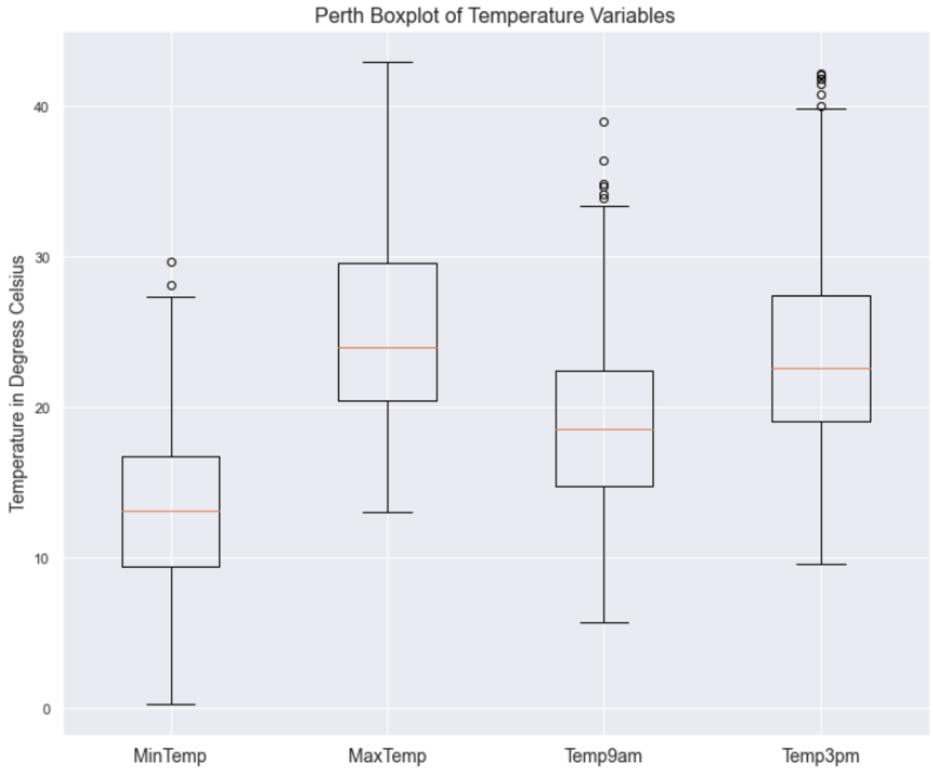


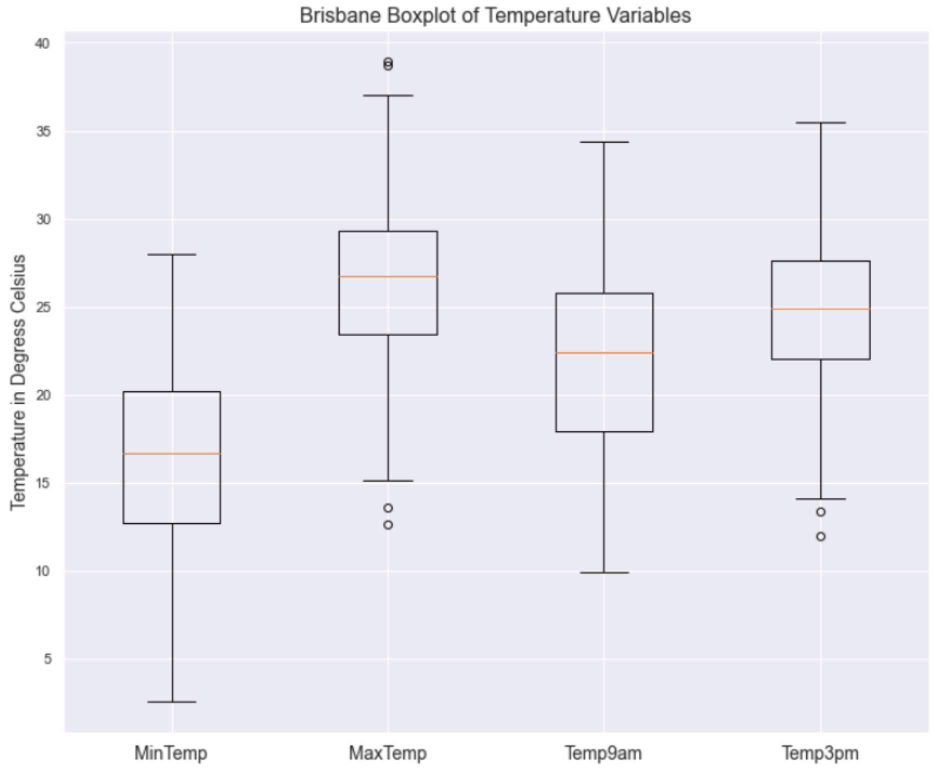




5) The box plots of temperature variables demonstrate that Darwin has a considerable number of outliers related to temperature in comparison to the other two cities. It also has a higher average temperature than the other 2 cities, with an average Temp9am and Temp3pm of 27 and 31 degrees Celsius. Brisbane was the next highest, with an average of 22 and 25 degrees Celsius. Finally, Perth had the lowest averages at 19 and 24 degrees Celsius. The lowest temperatures recorded for Darwin, Perth, and Brisbane were 12.5, .3, and 2.6 degrees Celsius, respectively. The highest temperatures recorded were 37.8, 42.9, and 38.9 degrees Celsius.


Darwin Boxplot of Temperature Variables
Temperature in Degress Celsius
35
30
25
20
15
MinTemp
MaxTemp
Temp9am
Temp3pm


Perth Boxplot of Temperature Variables
Temperature in Degress Celsius
40
30
20
10
0
MinTemp
MaxTemp
Temp9am
Temp3pm

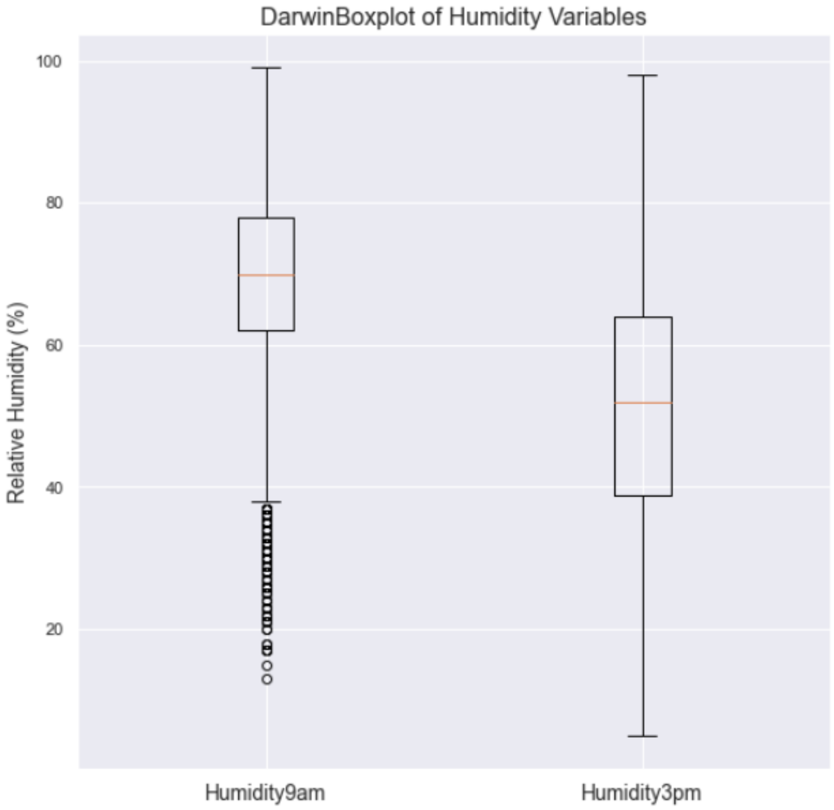


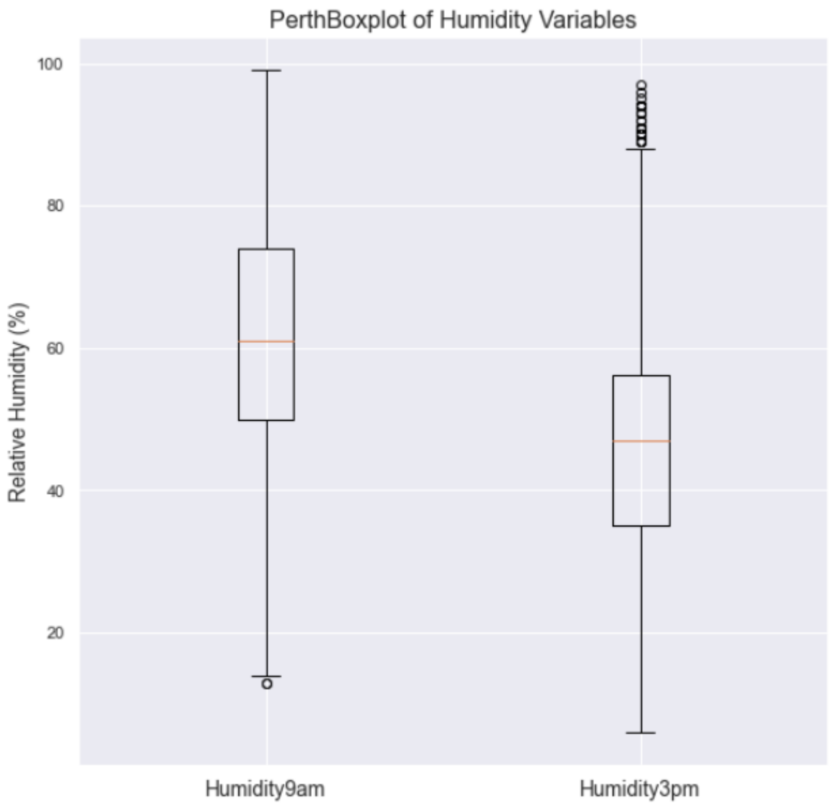
6) The box plots and descriptive statistics associated to wind speed variables display relatively similar distributions across all three cities, with Darwin seeing the highest wind gust speed at 102 kilometers per hour, and having a higher average speed at 9 and 3pm by around 20 kpm when compared to Perth and Brisbane.

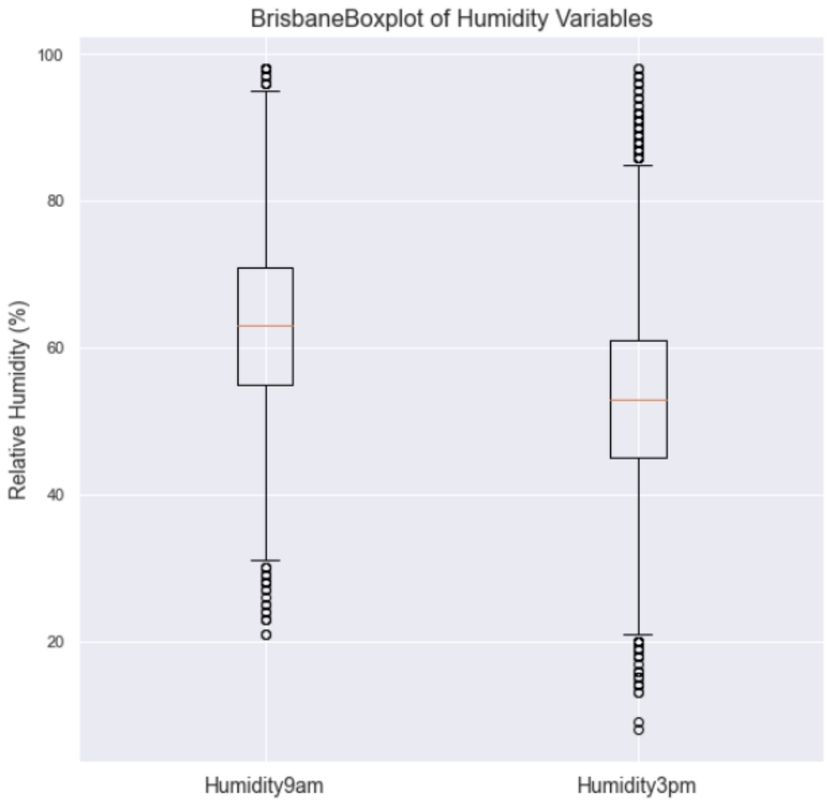
Location	Variable	Minimum	Mean	Maximum
Darwin	WindGustSpeed	15	40.5	102
	WindSpeed9am	2	15.2	50
	WindSpeed3pm	4	20.9	52
Perth	WindGustSpeed	13	35.1	83
	WindSpeed9am	2	11.2	30
	WindSpeed3pm	2	14.7	31
Brisbane	WindGustSpeed	11	28.3	93

	WindSpeed9am	2	7.1	37
	WindSpeed3pm	2	11.1	28

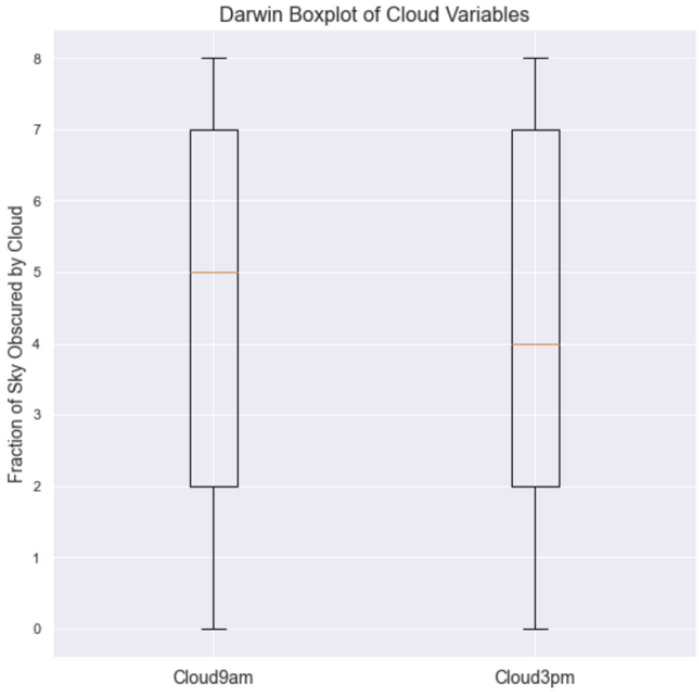
7) Using the box plots of Humidity variables, we can see that averages are not dissimilar for the cities, with Darwin having an average relative humidity of 68.1% at 9am and 51.3% at 3pm, Perth having an average relative humidity of 61.7% at 9am and 46.5% at 3pm, and Brisbane having an average relative humidity of 63.4% at 9am and 53.1 % at 3pm. However, the IQRs were observed to be different, with Perth having the largest IQR for Humidity9am, and Darwin having the largest IQR for Humidity3pm.

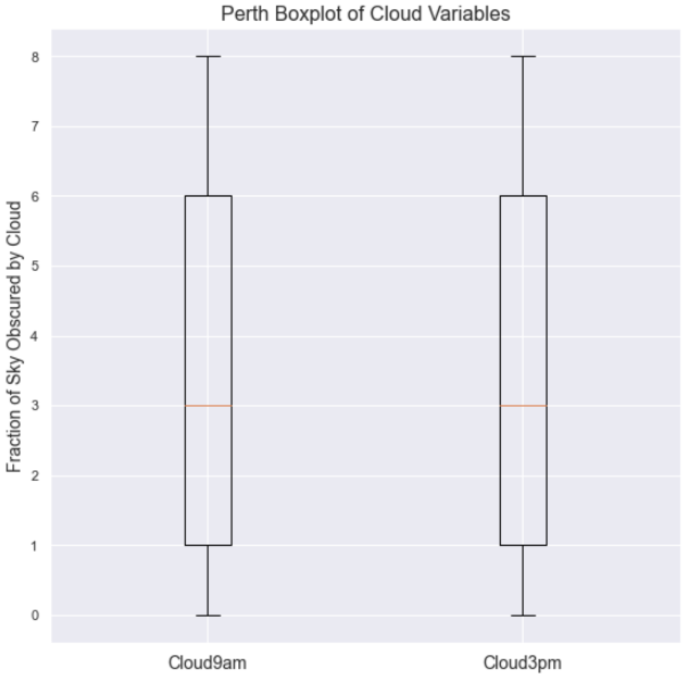


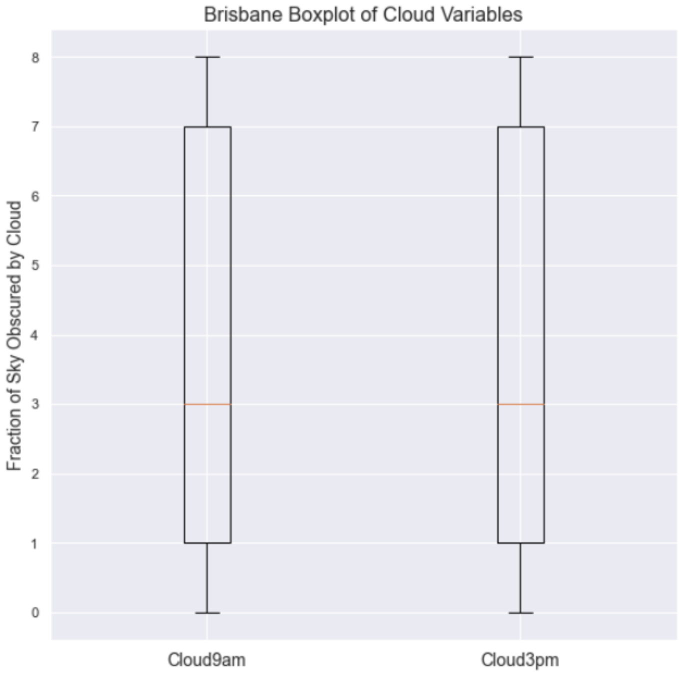

PerthBoxplot of Humidity Variables
Relative Humidity (%)
100
80
60
40
20
Humidity9am
Humidity3pm


BrisbaneBoxplot of Humidity Variables
Relative Humidity (%)
100
80
60
40
20
Humidity9am
Humidity3pm

8) The box plots of Cloud variables showed that Darwin was on average “cloudier”, with a mean fraction of sky obscured by clouds at 9am and 3pm of 5 and 4, whereas Perth and Brisbane had means of 3 and 3.



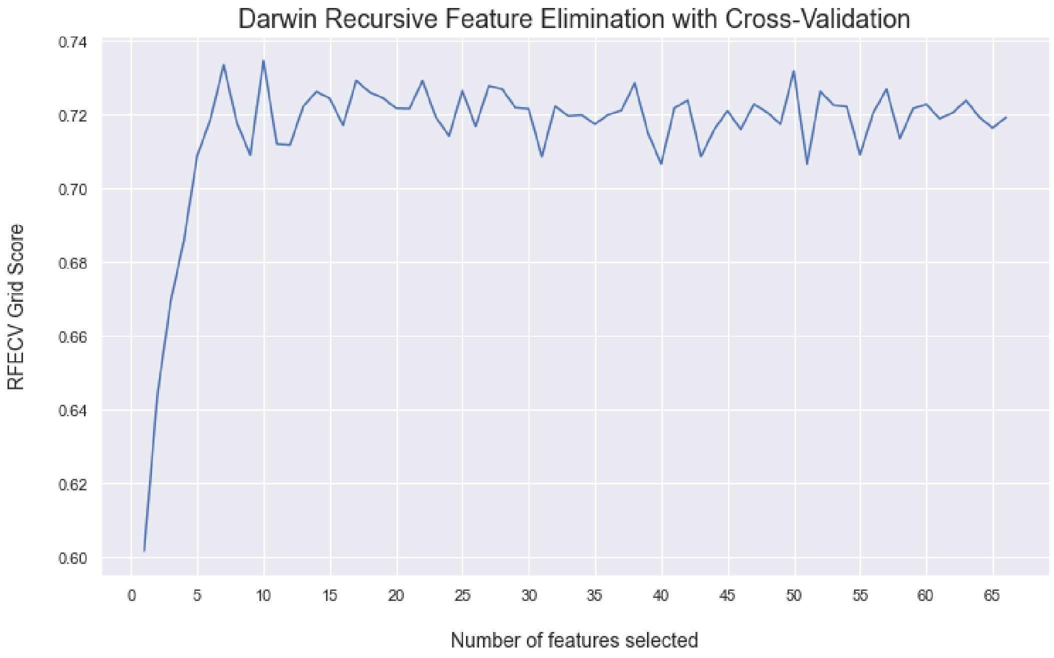




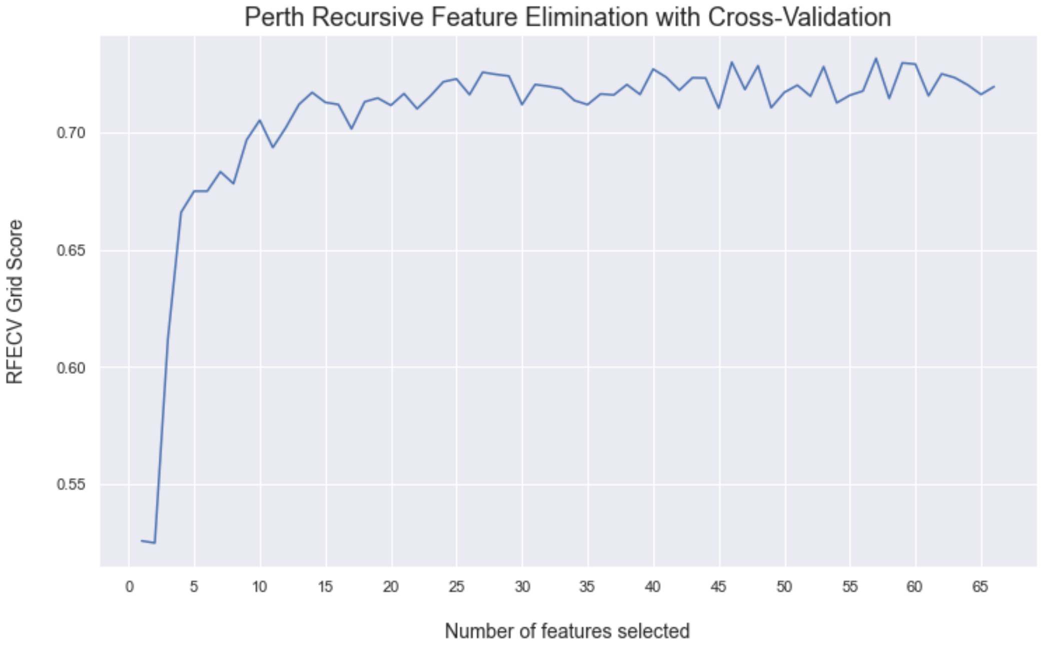
Feature Selection

Feature selection was performed on each training dataset using sklearn’s RFECV (Recursive Feature Elimination using Cross Validation). F1 score was optimized in a random forest classifier. F1 score was chosen as the metric for optimization in order to consider both precision and recall. This is ideal for target variables having imbalanced classes.

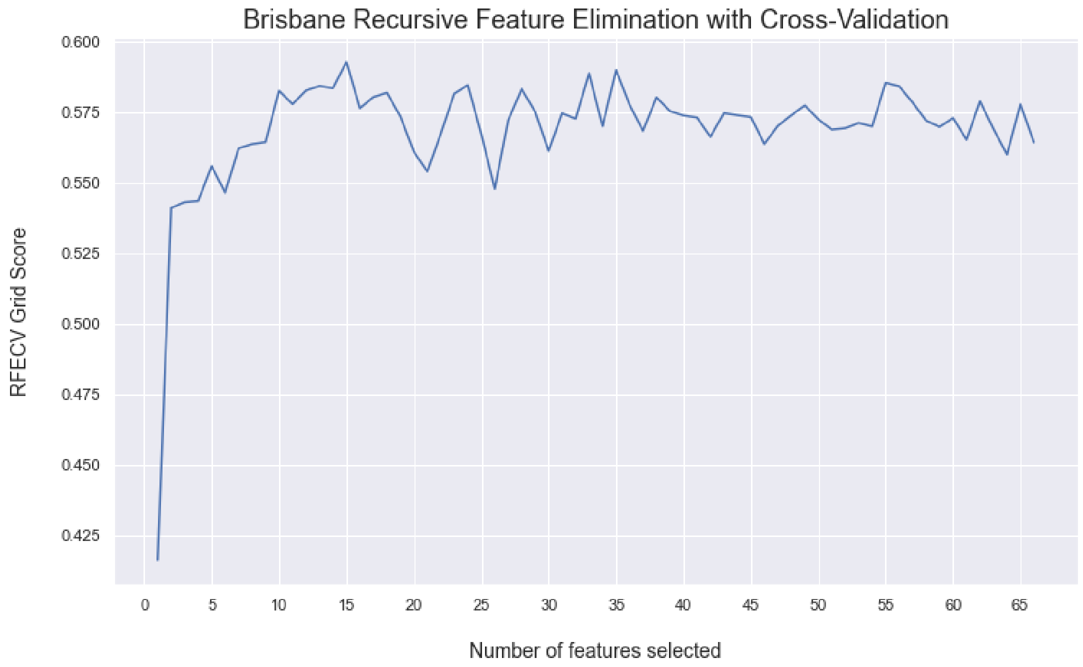
In feature selection for Darwin, the optimal number of features was determined to be 10, with the highest RFECV grid score (around .735). The 10 most important features were Humidity3pm, Sunshine, Pressure9am, WindGustSpeed, Pressure3pm, Rainfall, MinTemp, Temp3pm, MaxTemp, and Temp9am.



Similarly, RFECV was performed on Perth and the optimal number of features was found to be 57, a drastic difference from Darwin. The 10 most important features were Sunshine, Humidity3pm, Pressure3pm, Pressure9am, Temp3pm, MaxTemp, Rainfall, Humidity9am, WindGustSpeed, and Cloud3pm.



Finally, RFECV was performed on Brisbane and the optimal number of features was determined to be 15. The 10 most important features were Humidity3pm, Sunshine, Cloud3pm, Humidity9am, MinTemp, Temp3pm, Pressure3pm, Pressure9am, Temp9am, and Cloud9am.



Based on the feature elimination process, it seemed that Pressure, Humidity, Sunshine, Cloud, Wind Gust Speed, and Temperature variables were most important across all cities. Though Month, RainToday, and WindGustDir features were selected for modeling for Perth, the level of importance was relatively low.

Model Deployment

For this binary classification problem, three models were trained on each data frame and hyperparameters were tuned to optimize the performance. First, logistic regression was performed using sklearn's LogisticRegression. Hyperparameters chosen for tuning were k, C, and penalty. Next, Random Forest training was performed using sklearn's RandomForestClassifier. Hyperparameters selected for tuning were k, n_estimators, class_weight, max_depth, and max_features. Finally, Naïve Bayes training was performed using sklearn’s GaussianNB. The hyperparameter selected for tuning was var_smoothing.

The table below displays the optimal hyperparameter values that were calculated during grid search with 5-fold cross validation.

Algorithm	**Hyperparameter**	**Darwin** **Optimal Hyperparameter Value**	**Perth** **Optimal Hyperparameter Value**	**Brisbane** **Optimal Hyperparameter Value**
Logistic Regression	**k**	8	45	14
	C	1	100	0.01
	penalty	l2	l2	l2
Random Forest	**k**	10	10	8
	n_estimators	50	1000	20
	class_weight	None	balanced	None
	max_depth	8	8	6
	max_features	sqrt	sqrt	log2
Naitve Bayes	**var_smoothing**	.028	.01	.066

The best model for each city was selected by comparing the highest mean AUC ROC scores from the grid search. The table below displays the highest mean scores observed for each city.

Algorithm	**Darwin** **Highest Mean Test Score**	**Perth** **Highest Mean Test Score**	**Brisbane** **Highest Mean Test Score**
Logistic Regression	.924	.932	.888
Random Forest	.927	.908	.879

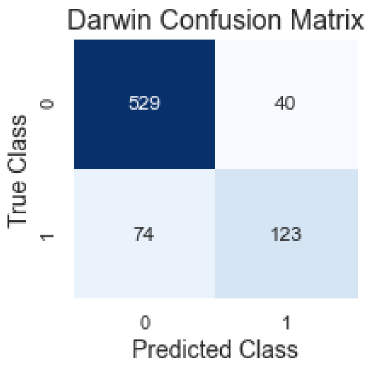
Naïve Bayes	.915	.878	.869

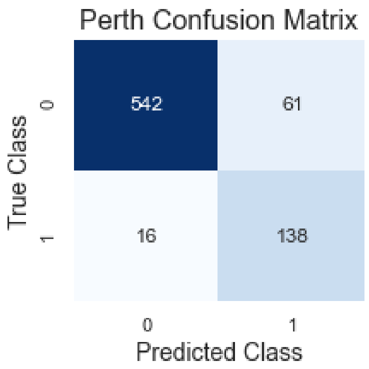
Based on the above statistics, the best models were selected for each city: for Darwin, Random Forest was the best performing model, while Perth and Brisbane both had the best performance using Logistic Regression. Though it was not considerably worse than the other algorithms, Naïve Bayes was not determined to be the best performing model for any of the three cities.

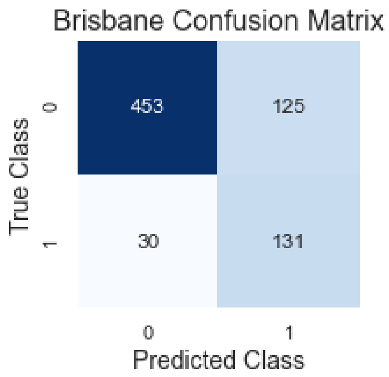
Results

Confusion Matrix

The three "best" models were tested using the reserved dataset so that predictions could be compared against actual recorded target class. Confusion matrices were created using sklearn's confusion_matrix function.







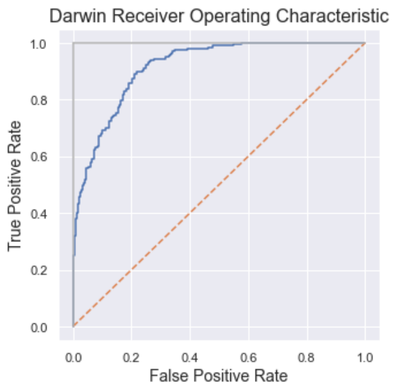
Classification Report

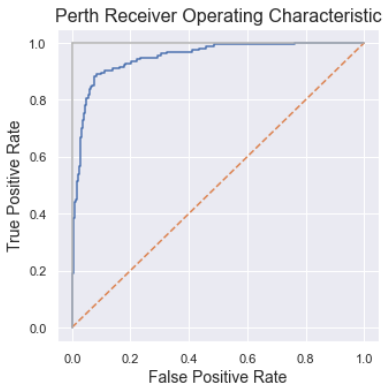
Next, evaluation metric reports were created using sklearn's classification_report. The classification reports showed that the f1-score, or the harmonic mean between precision & recall, is considerably lower for the positive class, implying that the models are more effective at determining the negative class (it will not rain tomorrow). It also showed a moderately high accuracy for each model, and that precision is relatively low for the positive classes, especially for Brisbane.

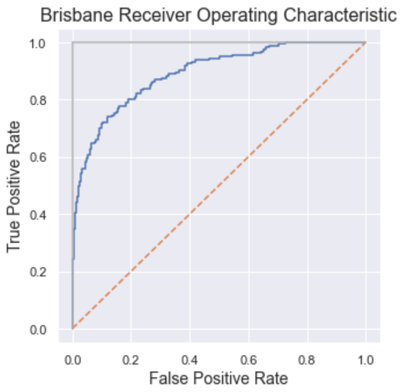
```
===========Classification Report - Darwin============
              precision    recall  f1-score   support

           0       0.88      0.93      0.90       569
           1       0.75      0.62      0.68       197

    accuracy                           0.85       766
   macro avg       0.82      0.78      0.79       766
weighted avg       0.85      0.85      0.85       766

===========Classification Report - Perth============
              precision    recall  f1-score   support

           0       0.97      0.90      0.93       603
           1       0.69      0.90      0.78       154

    accuracy                           0.90       757
   macro avg       0.83      0.90      0.86       757
weighted avg       0.91      0.90      0.90       757

===========Classification Report - Brisbane============
              precision    recall  f1-score   support

           0       0.94      0.78      0.85       578
           1       0.51      0.81      0.63       161

    accuracy                           0.79       739
   macro avg       0.72      0.80      0.74       739
weighted avg       0.85      0.79      0.80       739
```

ROC Curve

In addition to evaluation metrics, the Receiver Operating Characteristic (ROC) curve was plotted for each of the “best” models. Perth comes closest to the ideal clinical discriminator, while Darwin and Brisbane show fairly strong performance. Since ROC does not depend on the class distribution, and since we are dealing with unbalanced classes, this is an indication of good performance for the three models.







Discussion

Conclusion

The goal of this project was to predict whether it would rain tomorrow based on the available historical weather data. Based on the activities performed, it seems that the model architecture and performance varies by city, and may not be applicable on a wider scale. However, it does seem that models can be developed for individual cities with relatively strong performance. The results of the evaluation methods proved strongest when using metrics favored by imbalanced classes.

Opportunities for Improvement

During analysis and evaluation I found multiple opportunities for improvement.

1) **Imbalanced Classes** – For this project I chose to work with imbalanced target classes. As an enhancement the impact of over- or under-sampling the data on the models' performance could be explored.
2) **Missing Data** – Due to time limitations I chose to simply use the cities with the most complete datasets available. However, as can be seen in the missing data analysis, many cities have sparse or uncomprehensive data which will need to be addressed prior to

modeling. If given more time, I would perform regression to impute the data using the other available information.

3) **Correlation Analysis** – due to the high dimensionality of the data, I ran into issues performing correlation analysis using typical methods. Representing the correlations as a heatmap proved illegible due to the large number of features. After determining that the large number of WindGustDir features were not important to the models during feature selection, I would have dropped these during initial pre-processing, which could have helped with correlation analysis on the other features.

Acknowledgements

I would like to thank my professor for the challenging goals in this course, as well as the developers of sklearn as it has been vital to many of my course projects ☺.

References

Notes to accompany Daily Weather Observations. (n.d.). Australian Government Bureau of Meteorology. http://www.bom.gov.au/climate/dwo/IDCJDW0000.shtml