Introduction

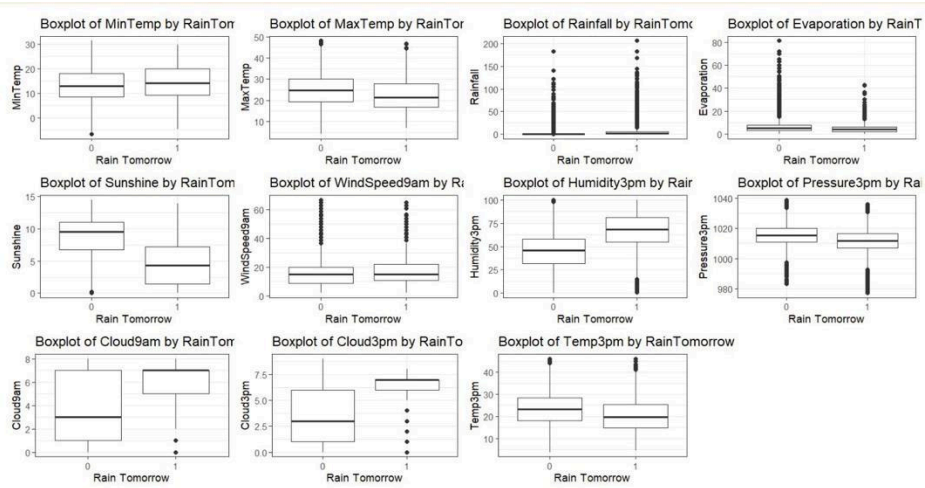
The significance of accurately predicting rainfall cannot be overstated, especially in a continent like Australia, known for its diverse climatic zones ranging from deserts to rainforests. This study leverages an extensive dataset comprising daily weather observations from multiple Australian locales, containing variables such as temperature, humidity, wind speed, and atmospheric pressure. The primary objective is to predict the occurrence of rain the following day, with the binary target variable "RainTomorrow" indicating the presence of at least 1mm of rainfall.

Exploratory Data Analysis

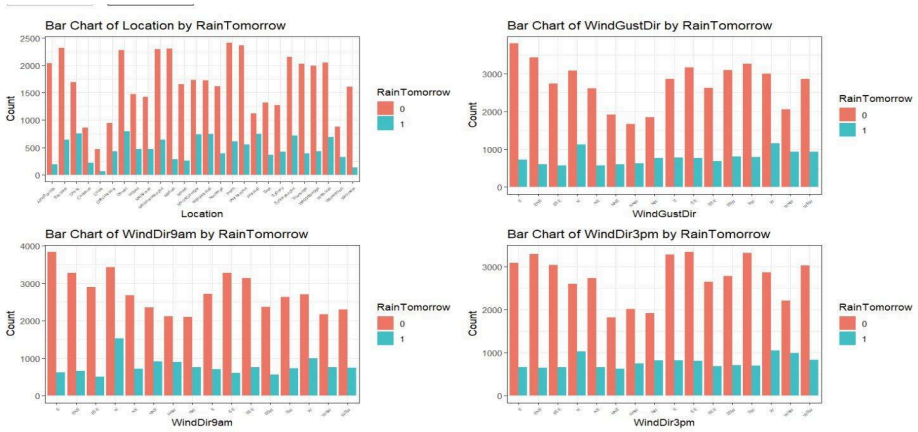
The initial phase involved an exploratory data analysis of the dataset, revealing missing values in several columns, necessitating cleansing. We first attempted imputation but came to realize the process was out of the scope of our capabilities and were directed to remove rows with missing data. With this, we went from 150,000 rows, to just over 50,000 rows in our dataset.

Categorical variables such as" Location", "WindGustDir,", "WindDir9am", and "WindDir3pm" were encoded for model compatibility. We also needed to encode our target variable, "RainTomorrow" from "Yes" or "No" to 1 or 0, respectively. These actions enabled us to perform more cohesive analysis with our data.

Looking at how our variables are distributed using boxplots, we can see the variables "Cloud3pm", "Cloud9am" and "Sunshine" have a large difference in means distributions, possibly indicating they are good predictors of "RainTomorrow".



Continuing in our EDA, we used bar charts to also assess variables. “WindGustDir”. “Location”, “WindDir9am”, and “WindDir3pm” also show a large difference between whether there will be rain tomorrow, also potentially indicating a strong predictor variable.



Our final dataset was imbalanced with about 20% of the dataset being classified as 1, meaning it did “RainTomorrow”, which is important to note as we enter analysis. Finally, the dataset was partitioned into training (45,136 records) and testing (11,284 records) sets, laying the groundwork for model training and evaluation.

Model Implementation and Evaluation

Logistic Regression

The logistic regression model identified variables such as "Sunshine," "Rainfall," "Perth," and "Perth Airport" as significant predictors of rain on the following day. This is in line with our

EDA. The logistic regression model performed quite will and achieved an accuracy of 85.85% and a balanced accuracy of 74.88%, with an AIC (Akaike Information Criterion) score of 29,853.

Confusion Matrix

	0	1
0	8321	1099
1	498	1366

K-Nearest Neighbors (KNN)

The K-Nearest Neighbors algorithm is a simple, yet effective non-parametric method used for classification and regression. In the context of predicting rainfall in Australia, KNN considers the features of a given day (such as temperature, humidity, wind speed, and pressure) and identifies the 'k' closest days in the historical dataset based on these features. The algorithm then predicts rain for the next day based on the majority vote or average of these 'k' nearest neighbors. Tuning the model with a K-value of 9, the KNN model slightly trailed the logistic regression with an accuracy of 85.21% and a balanced accuracy of 73.29%.

Confusion Matrix

	0	1
0	8330	1180
1	489	1285

Bagged Tree

Bagged Trees is an ensemble learning technique. It involves training multiple decision trees on different subsets of the original dataset and then aggregating their predictions. By averaging several trees, the Bagged Tree method reduces the variance part of the error. The Bagged Tree model outperformed the aforementioned models, registering an accuracy of 86.02%

and a balanced accuracy of 75.22%, with a confidence interval (CI) ranging from 85.3% to 86.52%.

Confusion Matrix

	0	1
0	8304	1063
1	515	1402

Support Vector Machine (SVM)

The SVM model lagged in performance, achieving an accuracy of 78.23% and a notably lower balanced accuracy of 50.18%, indicating a potential model bias or overfitting to a particular the accuracy is no better than a coin flip.

Confusion Matrix

	0	1
0	8819	2456
1	0	9

From this confusion matrix, we can determine the model is predicting mostly that it will not rain tomorrow and has a large false negative rate . Thus, incorrectly predicting approximately 2,500 variables. This could be because our dataset is imbalance. The SVM model is sensitive to imbalanced data, and, in hindsight, it would've been more efficient to use a model not sensitive to imbalanced datasets such as Gradient Boosting Machine (GBM).

Conclusion

The comparative analysis of the models reveals that the Bagged Tree approach holds a slight edge over Logistic Regression and KNN in predicting rainfall in Australia, with the SVM model falling behind in terms of both accuracy and balanced accuracy. These findings the complexity of meteorological predictions and the need for tailored models that can accommodate the intricate

patterns of weather variables. Future research could explore more sophisticated algorithms, feature engineering techniques, and the integration of external data sources to enhance predictive accuracy.