Rainfall Prediction in Australia Using Machine Learning Techniques

Manar Adel
Department of Computer Engineering
AAST
Alexandria, Egypt
manarkhedr02@gmail.com

Mayar Hesham
Department of Computer Engineering
AAST
Alexandria, Egypt
mayarfamy201@gmail.com

Ali Darwish
Department of Computer Engineering
AAST
Alexandria, Egypt
alidarwish618@gmail.com

ABSTRACT

in Australia, the unpredictability of rainfall rises substantial problems for water resource management and disaster preparedness, that directly affect both drought and flood planning. Our research addresses these challenges by using advanced Machine Learning techniques to enhance rainfall prediction accuracy. We have used a comprehensive dataset provided by the Australian Bureau of Meteorology, containing weather conditions across various locations collected from the year 2008 to the year 2017. By using machine learning algorithms, our study aims to improve the predictive accuracy of rainfall, therefore providing insight for water resource management and emergency planning. Keywords— Rainfall Prediction, Machine Learning, Weather Forecasting, Data Analysis, Predictive Modeling)

INTRODUCTION

Accurately predicting rainfall is critical for Australia, a nation with diverse climates. Traditional methods often struggle to provide timely forecasts that enable proactive decisions. Machine learning offers a promising approach to improve rainfall prediction by leveraging historical weather data. This study investigates this approach by training and testing multiple machine learning models on a comprehensive dataset encompassing various weather parameters collected across different Australian locations..

RELATED WORK

Several research efforts have explored different techniques for weather forecasting using statistical and machine learning approaches. These studies have focused on predicting various atmospheric parameters. The employed techniques range from simpler models like linear regression to more complex ones like neural networks. Notably, studies [1] demonstrate that machine learning can significantly outperform traditional statistical models in precipitation prediction.

However, there's a gap in applying these advancements specifically to Australian weather data. Australian climate is known for its high variability and unpredictable nature, which presents a unique challenge for rainfall prediction models. Our research aims to address this gap by focusing on utilizing machine learning techniques for improved rainfall prediction accuracy specifically tailored for Australian climatic conditions..

PROPOSED MODEL

Our Approach To improve rainfall prediction accuracy, we developed a model that leverages a combination of machine learning techniques. This model is optimized through a process called hyperparameter tuning, which involves adjusting various settings within the chosen machine learning algorithms to achieve the best performance.

Data Preprocessing and Feature Selection Prior to model training, we performed thorough data preprocessing as suggested by [2] to ensure the quality of our data. This included:

Dealing with Missing Values: We addressed missing values in each column using appropriate techniques. This may involve techniques like imputation with the mean, median, or mode depending on the nature of the data and the specific column.
Feature Selection: We don't blindly throw all available data at the models. Instead, we carefully select key features from the weather dataset based on two key criteria:
Statistical Correlation: We perform exploratory data analysis to identify features that exhibit a statistically significant correlation with rainfall occurrence. Features with weak or no correlation are excluded. Feature Importance: We train preliminary machine learning models and analyze their feature importance rankings. These rankings highlight features that contribute most significantly to the model's predictions. By focusing on these high-impact features, we aim to improve model efficiency and avoid overfitting.

EXPERIMENTAL WORK

Dataset: To train our rainfall prediction models, we utilized a large dataset provided by the Australian Bureau of Meteorology (ABM). This comprehensive dataset encompasses daily weather observations collected across various locations in Australia over a decade. It includes various weather features potentially influencing rainfall, such as temperature, humidity, air pressure, wind speed, and even prior day's rainfall data.

Evaluation Metrics: Once we developed different models for rain prediction, we employed several performance metrics to assess their effectiveness. These metrics are essentially tests that evaluate the models' ability

to predict rain accurately. Here's a breakdown of what each metric tells us:

Accuracy: This metric measures the overall proportion of correctly predicted rain/no-rain events. Precision: This metric indicates the proportion of predicted rain events that were actual rain occurrences. Recall: This metric captures the proportion of actual rain events that were correctly predicted.

F1-Score: This score combines precision and recall, providing a better overall picture of model performance.

ROC-AUC Score: This score represents the Area Under the Receiver Operating Characteristic Curve (ROC), which evaluates the model's ability to distinguish between rain and no-rain events. By analyzing these metrics together, we can identify the model that delivers the most accurate and reliable rainfall predictions for Australian weather patterns. Text

Results:

Logistic Regression: Our Logistic Regression model achieved a promising accuracy of 77.90%, indicating its ability to make accurate rain predictions for most cases

. This is further supported by the ROC AUC score of 0.7794, which signifies the model's good capability in distinguishing between rainy and non-rainy days.

Additionally, the Cohen's Kappa score of 0.5585 suggests that the model effectively handles the potential imbalance between rainy and non-rainy days in the dataset, performing better than random chance.

Precision, Recall, and F1-score across the two classes are detailed as shown in TABLE I :

TABLE I. EVALUATION METRICS FOR LOGISTIC REGRESSION

Classes	Evaluation Metrics		
	Precision	***Recall***	***F1 Score***
0 (No Rain)	84.41%	81.41%	82.75%
1 (Rain)	82.09%	84.74%	83.39%

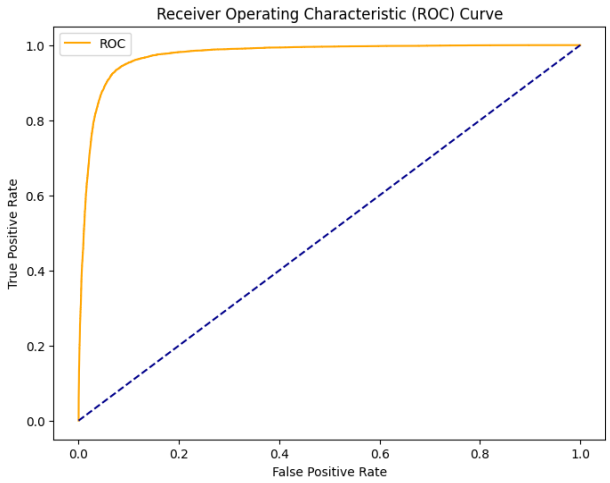


Fig. 1. Reciever Operating Characteristic (ROC) Curve for Logistic Regression

The ROC curve as shown in Fig. 1. for the Logistic Regression model showcases an effective trade-off between sensitivity and specificity. The curve remains well above the diagonal line of no-discrimination, reflecting the model's proficiency in distinguishing between the positive and negative classes.

The confusion matrix of the model provided additional insights into its classification accuracy shown in Fig. 2. :

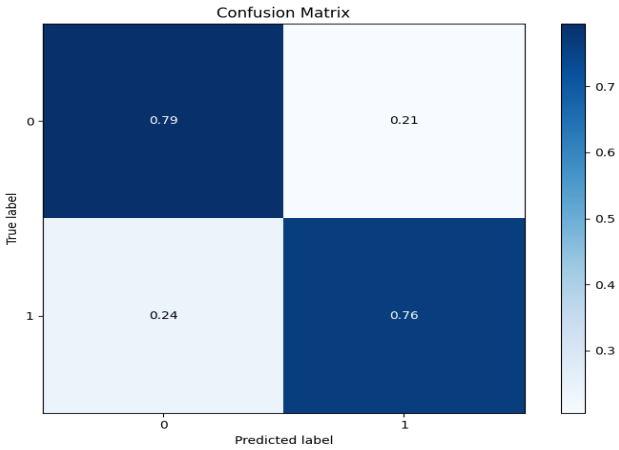


Fig. 2. Confusion Matrix for Logistic Regression

True Negatives (TN): 79% of the non-rainy days were correctly identified, indicating a strong specificity.

False Positives (FP): 21% of the days were incorrectly predicted as rainy, reflecting areas for improvement in reducing over-predictions.

False Negatives (FN): 24% of the actual rainy days were not predicted, highlighting a need for enhancing sensitivity.

True Positives (TP): 76% of rainy days were accurately forecasted.

Decision Tree: We evaluated the Decision Tree model's ability to predict rain in Australia. The Decision Tree model achieved an impressive accuracy of 83.08%, indicating a substantial improvement compared to Logistic Regression.

ROC AUC Score: The ROC AUC score reached a notable high of 0.8206. This score reflects the model's strong ability to distinguish between days with and without rain. In simpler terms, the model excels at differentiating rainy days from non-rainy days.

Cohen's Kappa Score: The Cohen's Kappa score of 0.6417 indicates substantial agreement beyond chance. This emphasizes the model's effectiveness in classifying rain events, performing significantly better than random guessing.

Precision, Recall, and F1-score across the two classes are detailed as shown in TABLE II :

TABLE II. EVALUATION METRICS FOR DECISION TREE

Classes	Evaluation Metrics		
	Precision	***Recall***	***F1 Score***
0 (No Rain)	84.41%	81.41%	82.75%
1 (Rain)	82.09%	84.74%	83.39%

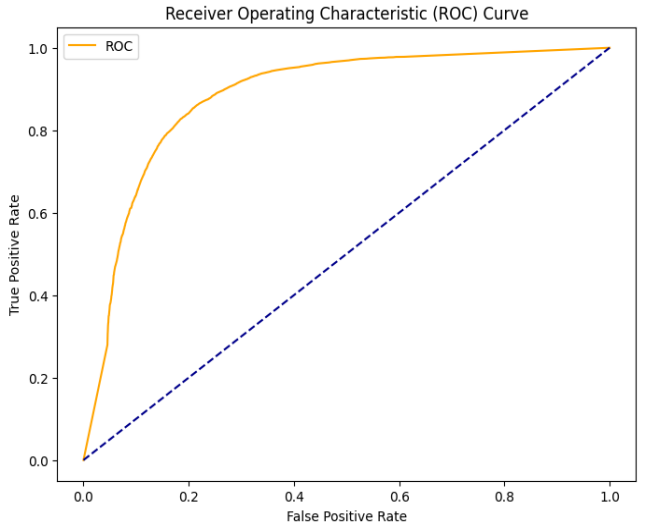


Fig. 3. Reciever Operating Characteristic (ROC) Curve for Decision Tree

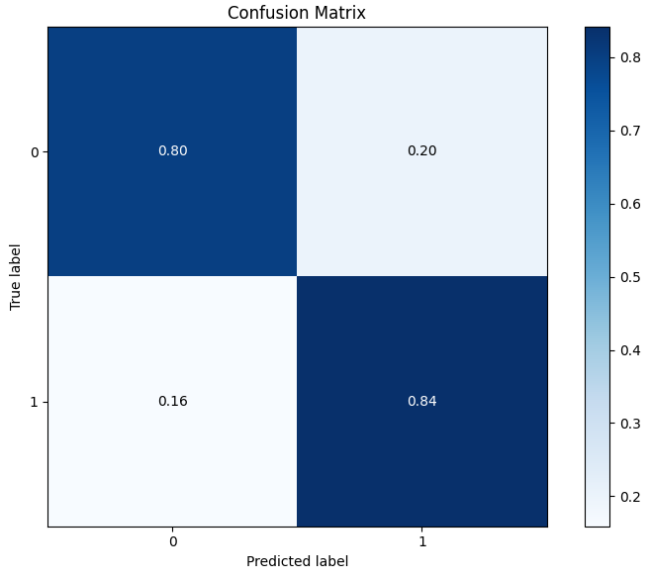


Fig. 4. Confusion Matrix for Decision Tree

Random Forest: The Random Forest model significantly outperformed the Logistic Regression and Decision Tree models, achieving an outstanding accuracy of 88.80%. This means the model can make highly accurate rain predictions for a very high percentage of cases.

The model's exceptional performance is further supported by the ROC AUC score of 0.8880. This score suggests the model excels at distinguishing between rainy days (positive class) and non-rainy days (negative class).

The Cohen's Kappa score of 0.7761 indicates a high degree of accuracy, especially considering the potential imbalance between rainy and non-rainy days in the dataset. In simpler terms, the model performs well even if the data has more of one type of day (rainy or non-rainy) compared to the other.

Precision, Recall, and F1-score across the two classes are detailed as shown in TABLE III :

TABLE III. EVALUATION METRICS FOR RANDOM FOREST

Classes	Evaluation Metrics		
	Precision	***Recall***	***F1 Score***
0 (No Rain)	90.34%	86.82%	88.55%
1 (Rain)	87.38%	90.77%	89.05%

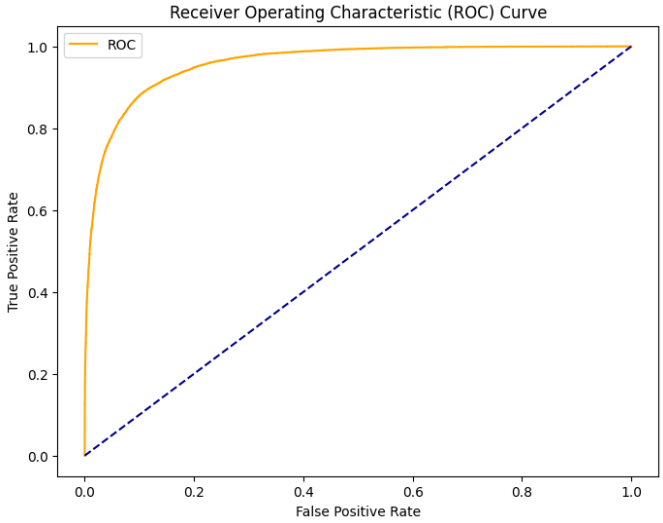


Fig. 5. Reciever Operating Characteristic (ROC) Curve for Random Forest

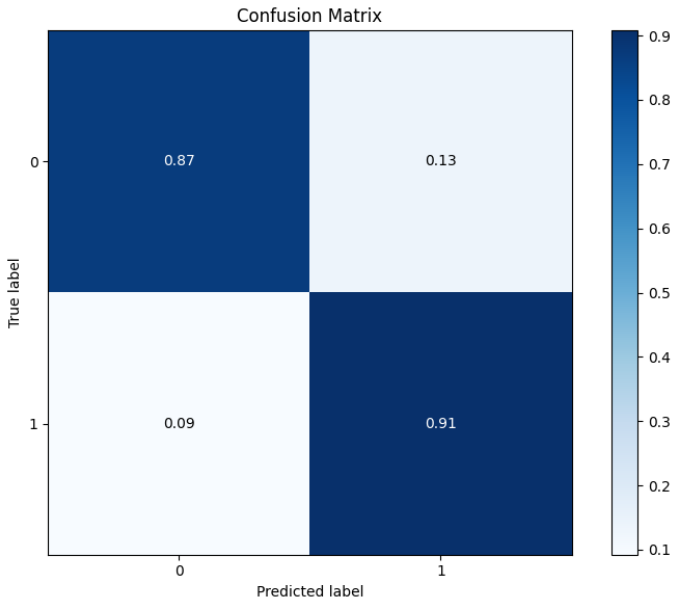


Fig. 6. Confusion Matrix for Random Forest

XGBoost: The XGBoost model surpassed all previous models, achieving a stellar accuracy of 92.87%. This represents a significant leap forward in rain prediction accuracy compared to the other models we tested. In simpler terms, the model can make highly accurate rain predictions for a very high percentage of cases in the data.
The model's exceptional performance is further reinforced by the ROC AUC score of 0.9287.

This score reflects the XGBoost model's superior ability to distinguish between rainy days and non-rainy days. Finally, Cohen's Kappa score of 0.8574 indicates an excellent level of agreement beyond chance. This emphasizes the model's precision in rain prediction, meaning it makes very few mistakes when classifying rainy days.

Precision, Recall, and F1-score across the two classes are detailed as shown in TABLE IV :

TABLE IV. EVALUATION METRICS FOR XGBOOST

Classes	**Evaluation Metrics**		
	Precision	***Recall***	***F1 Score***
0 (No Rain)	94.39%	91.11%	92.72%
1 (Rain)	91.45%	94.62%	93.01%

Fig. 7. Reciever Operating Characteristic (ROC) curve for XGBOOST

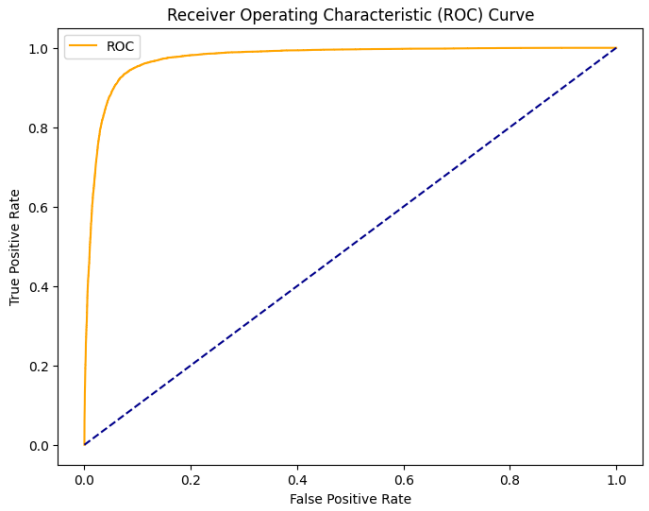
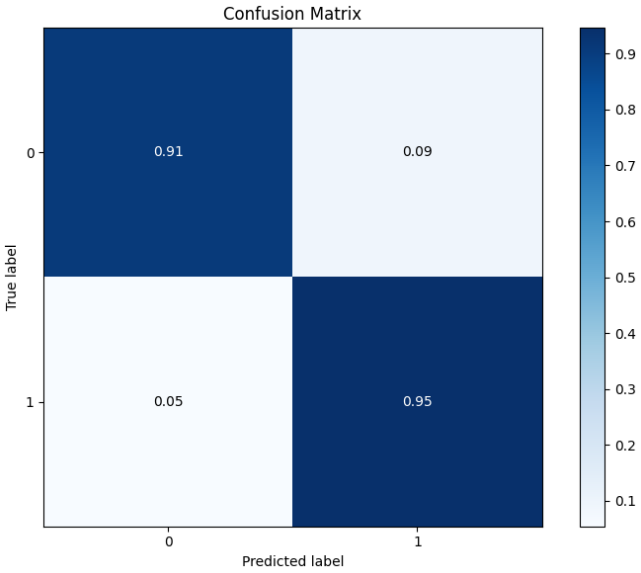


Fig. 8. Confusion Matrix for XGBoost



Belolw are the comparsion of accuracies between the models and their time taken as show in TABLE V and Fig. 9.

TABLE V. ACCURACIES COMPARSION TABLE

	Model			
	Logistic Regression	***Decision Tree***	***Random Forest***	***XGB BOOST***
Accuracy	77.90%	83.08%	88.80%	92.87%

Accuracy vs Time Taken for execution

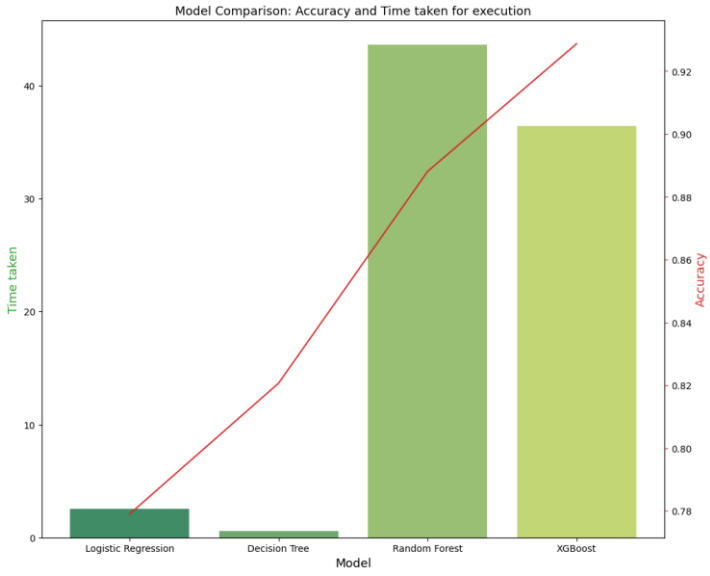


Fig. 9. Accuracy vs time taken for each model

CONCLUSION

In this study, we compared various machine learning models for their effectiveness in predicting rainfall in Australia. Among the models tested, the XGBoost model emerged as

the most accurate, achieving a remarkable accuracy rate of 92.87%. This high level of performance, coupled with an impressive ROC Area under Curve (AUC) of 0.9287 and a Cohen's Kappa score of 0.8574, underscores XGBoost's ability to handle complex predictive tasks with superior precision. However, it's important to consider the computational efficiency alongside accuracy. The XGBoost model required approximately 34.27 seconds to execute, which, while relatively efficient, was the second-highest time taken among the models evaluated. This presents a trade-off between accuracy and speed that must be considered, especially for real-time applications where rapid prediction might be crucial. The Decision Tree and Random Forest models also performed commendably but did not reach the accuracy level of the XGBoost model. On the other hand, these models required less time for execution, highlighting a potential area where XGBoost could be optimized further.

In conclusion, while the XGBoost model stands out for its high accuracy in rainfall prediction, its computational demand suggests an area for improvement. Future work could explore optimizing the XGBoost model to reduce its runtime without compromising its predictive accuracy. Additionally, integrating real-time data and testing the models in different climatic conditions could further enhance their applicability and robustness for operational use in weather forecasting and disaster preparedness.

REFERENCES

[1] Xiang, et al., "Precipitation Forecasting in Northern Bangladesh Using a Hybrid Machine Learning Model," SVR and ANN Algorithms, Rebus University online repository, 2017. [Online]. Available: https://rebus.us.edu.pl/bitstream/123456789/9999/1/article.pdfY.

[2] Aakash Parmar, Kinjal Mistree and Mithila Sompura, "Machine Learning Techniques for Rainfall Prediction: A Review", International Conference on Innovations in Information Embedded and Communication Systems (ICIIECS), March 2017..