Comparative Analysis of Machine Learning Models for Rainfall Prediction in Australia

Sreelekshmy Rengith srengith@umd.edu

***Abstract*—This paper presents a comparative analysis of various machine learning models to predict rainfall in Australia using the Rain in Australia dataset. The aim of this study is to evaluate the performance of five different machine learning algorithms: Logistic Regression, Decision Tree, Naive Bayes, Linear Discriminant Model, and K-Nearest Neighbors (KNN). Additionally, cross-validation is employed to assess the models' generalization capabilities. We also generate Receiver Operating Characteristic (ROC) curves for each model and compare their Area Under the Curve (AUC) metrics. Finally, based on the evaluation results, we discuss which model is the most suitable for rain prediction in Australia.**

I. Introduction

Accurate rainfall prediction is of paramount importance in Australia, with far-reaching implications across various sectors such as agriculture, water resource management, public safety, and event planning. The essence of this study lies in addressing the critical challenge of rain prediction and the subsequent need to develop robust machine learning models for this purpose. In this pursuit, we employ a diverse ensemble of models, including Logistic Regression, Decision Tree, Naive Bayes, Linear Discriminant Model, and K-nearest neighbors (KNN). To rigorously evaluate their performance and reliability, this study incorporates cross-validation techniques and generates ROC curves to compare the models based on their Area Under the Curve (AUC) metrics.

The significance of rainfall prediction cannot be overstated; it serves as a pivotal tool for informed decision-making. Accurate forecasts empower farmers to plan irrigation, enable reservoir managers to prudently manage water levels, and allow emergency services to proactively prepare for potential flood events. Thus, this study assumes the mantle of contributing to the advancement of dependable rain prediction models. By enhancing the precision of our forecasts, we aspire to facilitate better-informed decisions across various sectors.

The dataset underpinning this study is the ”Rain in Australia” dataset, a repository of ten years' worth of historical weather observations. This dataset encompasses a rich array of parameters, including temperature, wind direction, and rainfall. Further details regarding the dataset's characteristics and sourcing will be elaborated upon in subsequent sections.

While our study builds upon the solid foundation of prior research in weather prediction and machine learning, it is essential to underscore the contribution of Comparative Analysis of Rainfall Prediction Models Using Machine Learning in Islands with Complex Orography by Ricardo Aguasca-Colomo, Dagoberto Castellanos-Nieves, Máximo Méndez. This seminal work has been particularly instrumental in informing and guiding our research towards addressing the unique challenges presented by islands with complex orography.

II. Methodology

Data Collection: The study utilizes the ”Rain in Australia” dataset, which comprises ten years of historical weather observations. The dataset includes various meteorological parameters such as temperature, wind direction, rainfall, and more. It is sourced from the ”weatherAUS.csv” file.

Exploratory Data Analysis (EDA):

Before proceeding with model development, Exploratory Data Analysis (EDA) is conducted to gain insights into the dataset's characteristics and relationships between variables. Several visualizations are created to explore correlations and patterns within the data:

- A histogram showing ”Location” vs. rainy days, with a breakdown by ”RainToday.” (Fig.1.)
- A histogram illustrating ”Temperature at 3 pm” vs. ”Rain Tomorrow,” categorized by ”RainTomorrow.” (Fig.2.)
- A histogram depicting the relationship between ”Rain Tomorrow” and ”Rain Today.” (Fig.3.)
- A scatter plot showcasing ”Min Temp” vs. ”Max Temp,” with points colored by ”Rain Today.” (Fig.4.)
- Another scatter plot visualizing ”Temp (3 pm)” vs. ”Humidity (3 pm),” with points colored by ”Rain Tomorrow.” (Fig.5.)

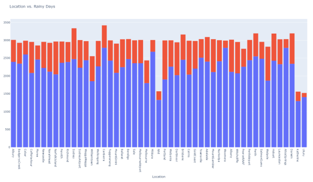

Fig. 1. Location vs. rainy days

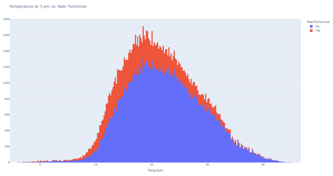

Fig. 2. ”Temperature at 3 pm” vs. ”Rain Tomorrow,”

Data Preprocessing: Data cleaning is an essential step to ensure the dataset's quality and consistency. Missing values in

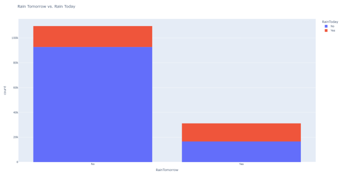

Fig. 3. ”Rain Tomorrow” and ”Rain Today.”

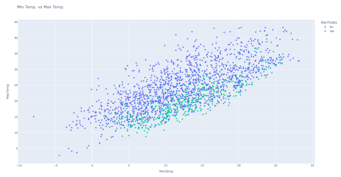

Fig. 4. ”Min Temp” vs. ”Max Temp,”

the ”RainToday” and ”RainTomorrow” columns are addressed by removing rows with missing data. For numeric columns with missing values, the SimpleImputer is employed with a mean strategy to fill in the gaps. Categorical columns with missing values are imputed using the ”most_frequent” strategy.

Data Encoding: Categorical data is encoded using label encoding techniques. This transformation ensures that categorical variables are in a numerical format, making them compatible with machine learning algorithms.

Feature Selection: The dataset is divided into feature variables (inputs) and the target variable. Feature selection is based on columns that are considered relevant for predicting rainfall.

Data Splitting: The dataset is split into training and testing sets to facilitate model evaluation. A 75-25% split is used, where 75% of the data is allocated for training, and 25% for testing. This ensures that the models are trained on a substantial portion of the data while reserving a separate portion for evaluating their performance. Model Selection:

Five different machine learning models are selected for this study: Logistic Regression, Decision Tree, Naive Bayes, Linear Discriminant Model, K-nearest neighbors (KNN)

Cross-Validation:

To evaluate the performance and robustness of each model, k-fold cross-validation with k=10 is employed. This technique partitions the training dataset into ten subsets, with each model trained and tested ten times, using different combinations of training and testing sets.

Model Evaluation: For each model, the following evaluation metrics are computed: Accuracy: The ratio of correct predictions to the total number of predictions. Confusion Matrix: A matrix representing the true positive, true negative, false positive, and false negative values. Precision, Recall, and F1-Score: Metrics that provide insights into the model’s performance with respect to positive and negative classes. Receiver Operating Characteristic (ROC) Curve:

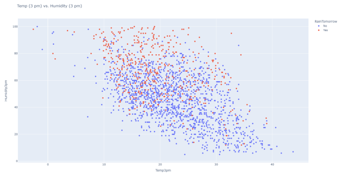

Fig. 5. ”Temp (3 pm)” vs. ”Humidity (3 pm),”

For models supporting predict_proba, ROC curves are generated. These curves illustrate the model’s ability to distinguish between positive and negative classes across various threshold values. The Area Under the Curve (AUC) is calculated for each ROC curve, providing a quantitative measure of model performance.

Model Comparison: The study concludes with a comparative analysis of all five models, considering their performance metrics, ROC curves, and AUC scores.

Selection of Optimal Model:

The choice of the optimal model is based on a holistic assessment of accuracy, precision, recall, F1-score, and the AUC metric. The model that exhibits the highest predictive performance will be selected.

Distance formula

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

d = distance
(x_1, y_1) = coordinates of the first point
(x_2, y_2) = coordinates of the second point

Fig. 6. Distance formula employed by KNN Classifier model

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

Fig. 7. Sigmoid Function of the Logistic Regression Model

III. Results

The Rain in Australia dataset was used to generate machine learning models for predicting whether it will rain tomorrow. Five different models were implemented and evaluated using cross-validation, and their performance was compared. Additionally, Receiver Operating Characteristic (ROC) curves were generated for each model to compare their Area Under the Curve (AUC) metrics. Finally, the optimal model for rainfall prediction was selected based on its overall performance.

Logistic Regression: Logistic Regression demonstrated strong performance in predicting rainfall. With a mean accuracy of 0.84 during cross-validation, it achieved a good balance between precision and recall. The model’s confusion matrix shows that it correctly predicted a substantial number of rainy and non-rainy days. Its accuracy of 0.84 indicates that it can reliably classify days as rainy or not. Furthermore, the model’s ROC curve exhibits an AUC of 0.88, suggesting that it can effectively discriminate between positive and negative instances. Logistic Regression is also relatively interpretable, making it a practical choice when model interpretability is important.

Decision Tree: The Decision Tree model showed moderate performance with a mean accuracy of 0.75 during cross-validation. While it exhibited reasonably high recall for both

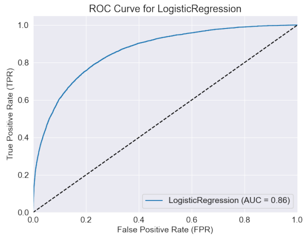

Fig. 8. ROC curve for Logistic Regression model

rainy and non-rainy days, it had a lower precision, resulting in a trade-off between precision and recall. The confusion matrix shows that it correctly classified a considerable number of non-rainy days but had more difficulty with rainy days. The accuracy of 0.79 indicates decent overall performance. However, the ROC curve displayed an AUC of 0.70, suggesting that it might not be as effective at distinguishing between classes compared to other models.

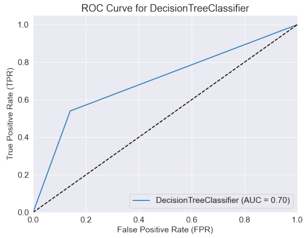

Fig. 9. ROC curve for Decision Tree model

K-Nearest Neighbors (KNN): K-Nearest Neighbors (KNN) performed well, with a mean accuracy of 0.81 during cross-validation. It showed high accuracy, especially in correctly classifying non-rainy days. However, it had a lower precision for rainy days, resulting in a slightly lower F1-score. The confusion matrix indicates a good balance between true positives and true negatives. With an accuracy of 0.84, KNN is a reliable model for rainfall prediction. The ROC curve displayed an AUC of 0.87, indicating strong discriminatory power. KNN's performance suggests that it's a robust choice for this classification task.

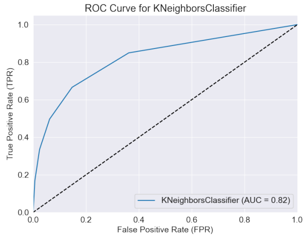

Fig. 10. ROC curve for K-Nearest Neighbors model

Naive Bayes (GaussianNB): Gaussian Naive Bayes (GaussianNB) demonstrated respectable performance with a mean accuracy of 0.80 during cross-validation. It achieved a good balance between precision and recall for both rainy and non-rainy days, as indicated by the confusion matrix. The model's accuracy of 0.81 suggests that it reliably predicts rainfall. The ROC curve displayed an AUC of 0.85, indicating its ability to distinguish between classes effectively. GaussianNB is known for its simplicity and speed, making it a practical choice for certain applications.

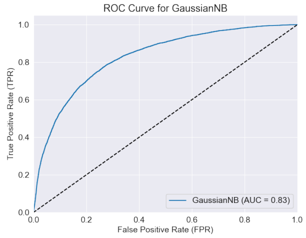

Fig. 11. ROC curve for Naive Bayes model

Linear Discriminant Analysis (LDA): Linear Discriminant Analysis (LDA) exhibited strong performance, with a mean accuracy of 0.84 during cross-validation. It displayed a good balance between precision and recall for both rainy and non-rainy days, as seen in the confusion matrix. With an accuracy of 0.84, LDA is a reliable model for predicting rainfall. The ROC curve displayed an AUC of 0.88, indicating its excellent discriminatory power. LDA also offers the advantage of being interpretable, making it a suitable choice when model transparency is desired. Its high accuracy and AUC make it a compelling option for rainfall prediction.

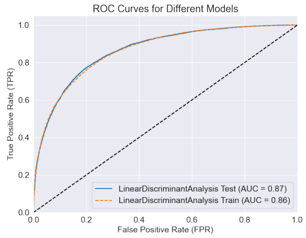

Fig. 12. ROC curve for Linear Discriminant Analysis model

Comparison and Decision:

- Among the models, Logistic Regression, K-Nearest Neighbors (KNN), and Linear Discriminant Analysis (LDA) performed the best in terms of accuracy and AUC.
- Logistic Regression and LDA achieved the highest mean accuracy during cross-validation, both at 0.84, and also had the highest AUC of 0.88. These models displayed a good balance between precision and recall.
- K-Nearest Neighbors (KNN) achieved a slightly lower mean accuracy of 0.81 during cross-validation but performed well on the test set with an accuracy of 0.84 and an AUC of 0.87.
- Naive Bayes (GaussianNB) and Decision Tree had lower mean accuracies during cross-validation (0.80 and 0.75, respectively) and also showed lower AUC values.

Considering the high mean accuracy, good AUC, and balanced precision-recall scores, both Logistic Regression and Linear Discriminant Analysis (LDA) are strong candidates for predicting rainfall. The choice between them may depend on factors such as model interpretability and ease of implementation.

In conclusion, based on the provided results, the preferred model for predicting rainfall would be either Logistic Regression or Linear Discriminant Analysis (LDA), with a slight preference for LDA due to its interpretability. However, it's essential to consider the specific requirements and constraints of the problem and perform further evaluation if necessary to make the final model selection.

IV. Discussions

The results of our machine learning models on the Rain in Australia dataset provide valuable insights into their performance and help us make informed decisions. Here, we discuss the key findings, the importance of this work, and potential future improvements or extensions:

Model Performance:

- Logistic Regression: Logistic regression achieved a mean accuracy of around 84%. It performed reasonably well in distinguishing between rainy and non-rainy days, with an AUC of 0.88.
- Decision Tree: The decision tree model had a mean accuracy of 75%. It showed potential for further optimization but had a lower AUC of 0.70.
- K-Nearest Neighbors (KNN): KNN exhibited a mean accuracy of approximately 81% and an AUC of 0.82, indicating decent performance.
- Naive Bayes: The Gaussian Naive Bayes model had a mean accuracy of 80% and an AUC of 0.85, indicating good predictive capability.
- Linear Discriminant Analysis (LDA): LDA achieved a mean accuracy of 84% and an AUC of 0.88, demonstrating competitive performance.

Expectations and Observations:

Logistic regression, LDA, and Naive Bayes performed well, as expected, given their suitability for binary classification tasks. Decision trees had lower accuracy, which might be due to overfitting; further hyperparameter tuning could enhance their performance. KNN's performance was decent, but it could benefit from feature scaling and hyperparameter optimization.

Importance:

This work is important for several reasons: Weather Prediction: Accurate weather predictions are essential for various sectors, including agriculture, transportation, and emergency services. Model Comparison: Comparing multiple machine learning models helps identify the most suitable one for a specific task, providing insights for decision-makers. Data Preprocessing: Data cleaning and preprocessing are crucial steps in building reliable machine learning models. This work showcases these essential data preparation steps.

Future Directions:

- Hyperparameter Tuning: Fine-tuning model hyperparameters can potentially improve performance for decision trees and KNN.
- Feature Engineering: Creating new features or transforming existing ones may enhance model accuracy.
- Ensemble Methods: Combining multiple models (e.g., ensemble methods like Random Forests or Gradient Boosting) could yield better results.
- More Data: Collecting more recent and diverse weather data can lead to more accurate predictions.
- Interpretable Models: Explaining model predictions to end-users, such as providing reasons for a rainy day prediction, can be crucial for real-world applications.

In conclusion, this work demonstrates the importance of evaluating various machine learning models for weather prediction. While Logistic Regression, Naive Bayes, and Linear Discriminant Analysis performed well, there is room for improvement in Decision Trees and KNN. Future work should focus on optimizing models, exploring new features, and enhancing the interpretability of predictions to make more accurate and useful weather forecasts.

References

[1] F. Provost and T. Fawcett, "Data Science for Business," O'Reilly, Dec 2013.

[2] A. Gelman and J. Hill, ”Logistic Regression: Why We Cannot Do What We Think We Can Do, and What We Can Do About It”.

[3] S. Roweis, ”k-Nearest Neighbors for Handwritten Digit Recognition”

[4] https:/umaryland.on.worldcat.orgatoztitleslink?sid=google&auinit=R&aulast=Aguasca-Colomo&atitle=Comparative+analysis+of+rainfall+prediction+models+using+machine+