Rainfall Prediction Using Machine Learning

Dona Jince[1], Mariyam Biju[1] and Sonia Abraham[1]

[1] *Mar Athanasius College of Engineering, Kothamangalam, Kerala, India*

donamaleri@gmail.com, mariyam.biju10@gmail.com

Abstract. Accurate rainfall forecasting is essential for agriculture, water management, and disaster preparedness. This project use three machine learning algorithms to predict whether it rains in 24 hours or not: Artificial Neural Network (ANN), Support Vector Machine (SVM), and Random Forest (RF). The analysis is based on historical weather data from the Kaggle WeatherAUS dataset. Work with preliminary data, including missing value handling, modeling, and feature selection. The design is integrated into the web interface, so users can get weather forecasts in real time. Among the tested algorithms, random forest performed the best with 90.1% accuracy when tested. This study demonstrates the power of machine learning to improve rainfall prediction and decision making in many areas.

1. INTRODUCTION

Water management, agriculture, disaster response, and public safety are some of the areas in which rain forecasting is important. Farmers have used forecasts to plan irrigation, prevent crop losses and improve planting plans. Water managers rely on forecasts to control water flows, regulate water flow, and reduce the risk of flooding or famine. Authorities depends on reliable rainfall data to issue flood and landslide warnings, protecting lives, and to minimize property damage, but they are less useful in the face of climate change, geographic variations, and seasonal changes.

More complex models are required for identifying patterns within the weather data such as temperature, humidity, wind speed, and atmospheric pressure. Machine learning is able to learn from past data and find hidden patterns in large datasets. The models built using these algorithms can provide better forecasts than the traditional ways. The advanced techniques such as Artificial Neural Networks (ANN), Support Vector Machines (SVM), and Random Forests are used to explain the relationships between these features and rain. The goal is to find the best model to predict whether it rains in 24 hours or not.Machine learning tools will be used to improve decision-making, water management, and disaster preparedness in agriculture. It will support research and development on future climate forecasts, paving the way for more accurate and up-to-date forecast models.

Section 2 contains the methodology of this project, which shows the detailed steps taken to develop and implement the rainfall prediction models. It begins with the dataset description followed by data preprocessing steps, feature selection, data standardization, machine learning models and evaluation metrics used to evaluate the models .Section 3 contains the results and discussion and the last section contains the conclusion.

2. METHODOLOGY

2.1 Dataset Description

The first step of every development of a machine learning model is to collect the necessary data for training the model. The dataset used in this project is the WeatherAUS dataset, sourced from Kaggle. It contains detailed weather observations recorded across various locations in Australia over a period of 10 years. The dataset contains a total of 22 distinct features and 145,460 observations. Among these features, 17 are continuous variables, while the remaining six are discrete variables. For instance, the target variable “to forecast is whether or not it will rain tomorrow, indicated by a binary value of ‘‘yes’’ or ‘‘no’’ . In this context, ‘‘yes’’ signifies that it will rain the following day if the rainfall for that day is recorded as 1mm or more.

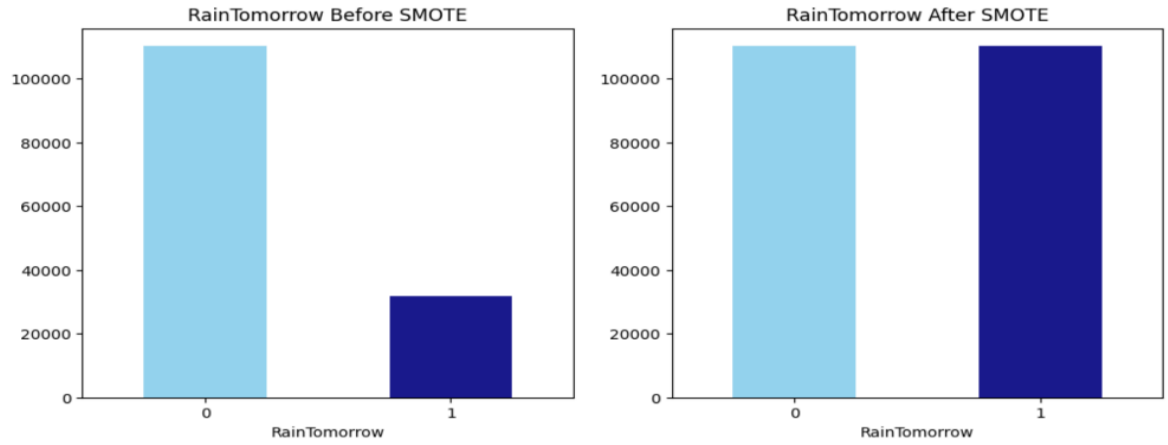
To provide a concise overview of the dataset variables, Table 1 presents a summary of their characteristics and descriptions. This information aids in understanding the nature of the dataset and the types of variables considered for rainfall prediction.

Table 1 Description of WeatherAUS dataset attributes

Attribute	Value Type	Description
Date	string	The day on which the measurement is carried out
Location	string	Station location name Meteorological
MinTemp	float	Minimum temperature in degrees Celsius
MaxTemp	float	Maximum temperature in degrees Celsius.
Rainfall	float	Amount of rain recorded during the day in mm.
Evaporation	float	Class A pan evaporation(mm) in 24h until 9a.m
Sunshine	float	Number of hours of radiant sun during the day
WindGustDir	string	The strongest windgust direction
WindGustSpeed	float	The strongest speed of wind gust in the 24 hours (km/h)
WindDir9am	string	Wind direction at 9 a.m.
WindDir3pm	string	The wind direction at 3 p.m
WindSpeed9am	float	Average wind speed(km/h) 10 minutes before 9a.m
WindSpeed3pm	float	Average wind speed(km/h) in the10min before 3p.m
Humidity9am	float	Humidity(%) at 9a.m
Humidity3m	float	Humidity(%) at 3p.m
Pressure9am	float	Atmospheric pressure(hpa) at the level of evil, at 9a.m
Pressure3pm	float	Atmospheric pressure(hpa) at the level of evil, at 3p.m
Cloud9am	float	The proportion of the sky not covered by clouds at 9am is measured in "oktash".
Cloud3pm	float	Fraction of sky covered by clouds at 3 p.m.The unit of measurement is the same as in Cloud 9 am measurements.
Temp9am	float	Temperature at 9a.m in degrees Celsius
Temp3pm	float	Temperature at 3p.m in degrees Celsius
RainToday	string	Boolean: Yes if precipitation exceeds 1mm in the 24h to 9a.m else No.
RainTomorrow	string	Boolean: Yes if it rains in 24 hours, else No

2.2 Data Pre-processing

Data pre-processing is an important step in every machine learning process. Raw data we collected will be noisy, insignificant, and inconsistent, which can impact the performance of forecast models. In this project, the WeatherAUS dataset presented challenges such as missing values and inconsistent data, which required various preprocessing to ensure that the data was entered correctly. For categorical features such as WindGustDir and RainToday, type (i.e., maximum value) is used to fill in missing text. This approach minimizes bias by ensuring that categories are generally similar across the data. For non-critical numbers, the multiple sequential chained equations (MICE) method is used. MICE makes it more robust by taking into account the relationship between variables by remodeling each variable as a function of the other variable. For example, for the case of "no", the number of data samples exceeds the number of "yes" cases. This imbalance can lead to a biased model where the algorithm is biased towards the majority class (i.e. "no"). To solve this problem, the Synthetic Minority Oversampling Technique (SMOTE) is adopted. SMOTE creates a synthetic model (YES) for a small number of classes, thereby evaluating the dataset and helping the model learn from the two classes before processing by the machine learning algorithms. This function uses label encoding, which gives each group a unique number. This encoding converts categorical data into a format that the model can interpret without specifying the order of the categories . Figure 1 shows the class distribution before and after smote.



(a) class distribution before balancing (b) class distribution after balancing

Figure 1 Smote

2.3 Feature Selection

Feature selection is an important step in building accurate and correct machine learning models. It involves identifying the most important features that help in the prediction task while discarding irrelevant or unnecessary features. This reduces model complexity and improves performance. In this project, filtering is applied to feature selection using the f_classif function, which evaluates the statistical relationship between each feature and the target variable (RainTomorrow). This includes an ANOVA F-value that shows the importance of each feature in predicting the target. Features with higher F values are important because they have a better relationship with the target. The top 10 most important features identified by this process are Sunshine, WindGustSpeed, Humidity9am, Humidity3pm, Pressure9am, Pressure3pm, Cloud9am, Cloud3pm, Temp3pm, and RainToday. These features were selected based on their high statistical correlation with precipitation estimates. For example, sunlight and cloud cover play a significant role in weather conditions that cause rain, while humidity and pressure are direct indicators of humidity and stability of the air. By reducing the number of features to the most important features, the model training process becomes more efficient and the risk of overfitting is reduced due to the model being trained less on irrelevant information.

2.4 Data Normalization

MinMaxScaler was used to scale the features, ensuring all inputs were standardized between 0 and 1 which helps models like ANN and SVM that are sensitive to feature magnitude.

2.5 Machine Learning Models

In this project, we use three different machine learning algorithms to predict whether it will rain in the next 24 hours: Artificial Neural Network (ANN), Support Vector Machine (SVM), and Random Forest (RF). Each of these models approaches the problem in a unique way and has its own unique strengths when it comes to handling complex weather data. ANN is seen as a simplified version of how the human brain works. They consist of a network of "neurons" that work together to identify patterns in data. In this project, we use a multi-layer ANN called ReLU (rectified linear unit) to do its job. The ReLU function helps the network focus on important patterns in weather data, such as temperature changes, which are important for precipitation prediction. Neural networks are particularly good at identifying nonlinear relationships between different climate features, making them a powerful tool for making predictions without knowing the weather. > The goal of SVM is to draw a clear line (or hyperplane) that best separates the data into two groups (in our case, rainy days and rainless days). Think of it like drawing a straight line on a graph dividing two sets of points. The model tries to find the “best” line that makes the biggest difference between the two groups (rainy and rainless). For this project, we use the output of SVM and adjust a parameter called “C” to strike the balance between getting the right results and over-tightening the classification. SVMs are particularly useful for datasets with many features because they can handle the data well. Let’s work with trees, and each tree makes a prediction. The model then combines the predictions from each tree to make a final decision. Imagine asking 100 weather experts and each expert gives you a prediction; the random forest listens to everyone and chooses the most common answer. This makes it more powerful and less likely to be over-trained, meaning it can not only remember training information but also amplify new, unseen information well. In this project, we fixed the main problem of the trees and the depth of each tree growth to make sure we get the best prediction. Random forests are particularly powerful when dealing with complex data like weather, where many variables can affect the results. We are improving prediction accuracy and decision making in real-world applications like agriculture and water management. Table 2 shows the hyperparameters used in each models.

Table 2 Hypermeters setting for machine learning models

Model	Hyperparameters
RF	max_depth: 17, min_samples_leaf: 1, min_samples_split: 2, n_estimators: 100, random_state: 12345
ANN	random_state :42, verbose :1, hidden_layer_sizes :(100, 50), activation: relu , solver : adam
SVM	random_state : 42, C :1, kernel :linear

2.6 Evaluation of the Model

In machine learning, performance metrics express how well the algorithm performs based on various factors such as accuracy, precision, recall, and F1 score.

Accuracy

The percentage of correct test data predictions is referred to as accuracy. It is easy to calculate by dividing the number of forecasts by the number of correct guesses.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

Precision

The precision score is used to assess the model’s correctly counting genuine positives among all positive predictions.

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall

The recall is used to assess the model's performance in terms of accurately counting true positives among all actual positive values.

$$\text{Recall} = \frac{TP}{TP + FN}$$

F1 Score

The F1 score is a compromise between the true and recall scores and is used to judge when choosing the correct one or you may end up with a model with too many false positives or false positives.

$$\text{F1 Score} = \frac{2\,(\text{Precision} * \text{Recall})}{\text{Precision} + \text{Recall}}$$

3. RESULTS AND DISCUSSIONS

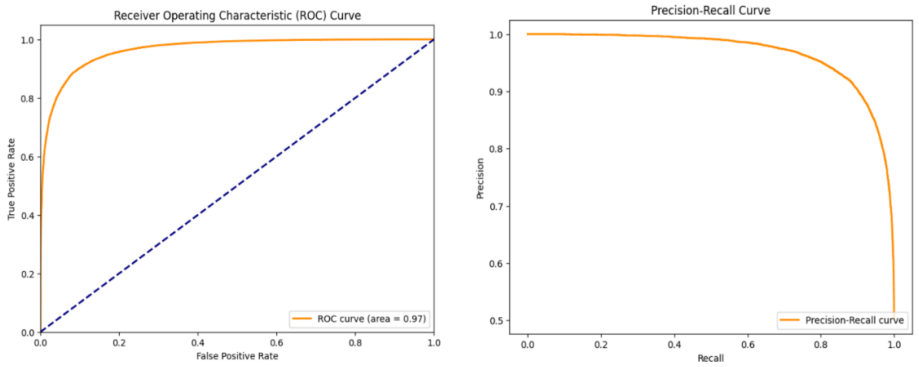
To measure the performance of RF, ANN and SVM models various evaluation metrics are used. Random Forest model outperforms all other models with 90.2% precision, 90.1% recall, and 90.2% F1-score followed by ANN model. Table 3 and 4 shows the results of the evaluations. Figure 2 shows the performance of the random forest model.

Table 3 Training and Testing Accuracies

Algorithm	Training Accuracy	Testing Accuracy
Random Forest Classifier	97.8	90.1
Artificial Neural Network	88.0	87.6
Support Vector Machine	81.2	80.8

Table 4 Precision, Recall, and F1 Score

Algorithm	Precision	Recall	F1 Score
Random Forest Classifier	90.2	90.1	90.2
Artificial Neural Network	85.9	89.9	87.9
Support Vector Machine	80.7	80.9	80.8



(a) ROC Curve (b) Precision-Recall

Figure 2 Curves of Random Forest model

In summary, Random Forest is the best-performing model for rainfall prediction due to its high accuracy.

4. CONCLUSION

Various machine learning classification techniques are investigated and evaluated at different stages of the research for predicting rainfall using the WeatherAUS dataset. After preprocessing, feature selection, and standardization, RF achieved the highest accuracy (90.1%) and balanced precision recall, making it the most reliable model. ANN showed strong performance but was slightly less accurate, while SVM struggled with the data's complexity. The integration of models into a web interface enhances accessibility, providing real-time predictions for end-users. Overall, Random Forest proves highly effective, demonstrating machine learning's potential to improve rainfall forecasting and decision-making in various sectors.

By incorporating real-time data integration could further improve prediction accuracy by enabling the model to process and analyze updated weather information continuously. This would provide more accurate forecasts, making the system more dynamic and reliable for decision-making in critical sectors like agriculture and disaster management.

REFERENCES

[1] Sarasa-Cabezuelo, A. Prediction of Rainfall in Australia Using Machine Learning. *Information* 2022, *13*, 163. https://doi.org/10.3390/info13040163

[2] M. M. Hassan et al., "Machine Learning-Based Rainfall Prediction: Unveiling Insights and Forecasting for Improved Preparedness," in IEEE Access, vol. 11, pp. 132196-132222, 2023, doi: 10.1109/ACCESS.2023.3333876.

[3] Ghosh, S., Gourisaria, M.K., Sahoo, B. *et al.* A pragmatic ensemble learning approach for rainfall prediction. *Discov Internet Things* **3**, 13 (2023). https://doi.org/10.1007/s43926-023-00044-3

[4] Appiah-Badu, N.K.A., Missah, Y.M., Amekudzi, L.K., Ussiph, N., Frimpong, T. and Ahene, E., 2021. Rainfall prediction using machine learning algorithms for the various ecological zones of Ghana. *IEEE Access*, *10*, pp.5069-5082.

[5] Kaggle, WeatherAUS dataset: https://www.kaggle.com/datasets/trisha2094/weatheraus

[6] Z. He, ‘‘Rain prediction in Australia with active learning algo rithm,’’ in Proc. Int. Conf. Comput. Autom. (CompAuto), Sep. 2021, pp. 14–18.

[7] Hussein, E. A., Ghaziasgar, M., Thron, C., Vaccari, M., & Jafta, Y. (2022). Rainfall prediction using machine learning models: literature survey. *Artificial Intelligence for Data Science in Theory and Practice*, 75-108.

[8] Mohammed, M., Kolapalli, R., Golla, N. and Maturi, S.S., 2020. Prediction of rainfall using machine learning techniques. *International Journal of Scientific and Technology Research*, *9*(1), pp.3236-3240.

[9] Parmar, A., Mistree, K. and Sompura, M., 2017, March. Machine learning techniques for rainfall prediction: A review. In *International conference on innovations in information embedded and communication systems* (Vol. 3).

[10] Basha, C.Z., Bhavana, N., Bhavya, P. and Sowmya, V., 2020, July. Rainfall prediction using machine learning & deep learning techniques. In *2020 international conference on electronics and sustainable communication systems (ICESC)* (pp. 92-97). IEEE.