Published by
NIGERIAN SOCIETY OF PHYSICAL SCIENCES
Available online @ https://asr.nsps.org.ng/index.php/asr

African Scientific Reports 5 (2026) 541

African Scientific Reports

Daily rainfall prediction in Lagos using logistic regression, random forest, and support vector machine with web development

Oladimeji Lukman Abiola*, Oke Samuel Abayomi, Akinade Oludayo Olugbenga, Oguntola Toyin Omoyeni, Adebayo Azeez Ademola

Department of Statistics, Ladoke Akintola University of Technology, Ogbomoso, Nigeria

Abstract

Accurate rainfall prediction is critical for agricultural planning, flood-risk management, and urban infrastructure resilience, particularly in tropical coastal cities such as Lagos, Nigeria, where unpredictable rainfall events cause significant socioeconomic disruption. This study evaluates three machine learning (ML) classification models—logistic regression (LR), random forest (RF), and support vector machine (SVM)—for daily binary rainfall prediction using a 22-year meteorological dataset of 8,314 observations sourced from Visual Crossing. Two features were engineered from the raw data: daily temperature range and a seasonal indicator. To address the asymmetric cost of false negatives in a tropical rainfall context, Youden's J statistic was applied to optimise the decision threshold of each model, prioritising sensitivity over the conventional 0.5 default. RF achieved the strongest overall performance, recording an accuracy of 76.28%, sensitivity of 79.27%, F1-score of 76.44%, and area under the receiver operating characteristic curve (AUC-ROC) of 0.8454, outperforming SVM (AUC-ROC: 0.8215) and LR (AUC-ROC: 0.8001). Humidity, cloud cover, dew point, and visibility emerged as the most consistent predictors across models, while moon phase and wind speed showed negligible importance in all three classifiers. All three trained models were deployed in an interactive R Shiny web application, enabling non-technical users, including farmers, planners, and policymakers, to obtain real-time rainfall predictions from meteorological inputs. This study provides an end-to-end rainfall prediction pipeline tailored to the Lagos environment and demonstrates the practical value of coupling ML with accessible deployment frameworks for climate decision-making in developing countries.

DOI: 10.46481/asr.2026.5.3.541

Keywords: Rainfall prediction, Machine learning, Random forest, Logistic regression, Support vector machine

Article History:
Received: 12 May 2026
Received in revised form: 02 July 2026
Accepted for publication: 03 July 2026
Available online: 28 August 2026

1. Introduction

Rainfall is one of the most important natural phenomena affecting agriculture, water resources management, energy, mining, and other sectors [1]. Accurate rainfall prediction is essential for reducing the negative effects of climate variability, including floods, droughts, food insecurity, and infrastructure damage. Variations in rainfall timing and quantity have long made rainfall prediction a difficult task for meteorological experts [2]. This highlights the importance of rainfall prediction, particularly in many developing

*Corresponding Author Tel. No.: +234-803-554-0640.
Email address: laoladimeji24@lautech.edu.ng (Oladimeji Lukman Abiola)

countries where farming relies primarily on rainfall as a water source [3]. Lagos, Nigeria's most populated city and commercial centre, is especially vulnerable to heavy rainfall that often disrupts businesses, causes highway traffic congestion, floods residential areas, and leads to loss of lives and property [4]. Lagos's coastal tropical climate, characterised by distinct wet and dry seasons and convective rainfall events, makes accurate daily rainfall classification especially difficult. These risks caused by unpredictable rainfall are major reasons why rainfall prediction is needed for Lagos.

Machine learning (ML) offers a powerful alternative to the explicit physics-based forecasting used by traditional prediction models because of its ability to capture complex non-linear relationships between meteorological variables [5]. Gradient boosting, random forest (RF), and decision tree (DT) models were applied to a 40-year dataset in a study by Patel *et al*. [6]. Among the evaluated models, gradient boosting was the most effective, achieving 93% accuracy for characteristics that influence rainfall predictability, with RF and DT methods following closely. Another study, which used daily Indian precipitation data to provide climate insights, reported that RF had the best classification accuracy, whereas autoregressive integrated moving average (ARIMA) and neural-network models were the most effective for predicting meteorological characteristics [7]. In a Moroccan study that applied five ML algorithms—DT, RF, k-nearest neighbours (KNN), AdaBoost, and XGBoost—as well as a long short-term memory (LSTM) network to monthly rainfall data, XGBoost and AdaBoost were identified as meta-learners with the potential to integrate and use data from multiple sources, thereby enhancing prediction performance [8].

Ejike *et al*. [9] evaluated the influence of climate change on model performance in tropical and temperate regions through ML-based rainfall prediction. Logistic regression (LR), RF, and support vector machine (SVM) models were applied to meteorological data with variables such as temperature, humidity, and air pressure. RF and SVM outperformed LR because of their ability to detect non-linear relationships in complex data. However, LR continued to provide interpretable baseline results and was useful for comparison. That study also reported that model effectiveness depends on the weather and emphasised the need to evaluate predictive models according to climate. Collectively, these findings suggest that ensemble models such as RF and gradient boosting, and margin-based models such as SVM, are well suited to tropical rainfall classification, while LR provides an interpretable baseline. This evidence motivates the selection of the three classifiers used in the present study.

SmartForecast, an interactive R Shiny dashboard for rainfall forecasting in Subang, was developed by Bahrom R [10]. That study showed that interactive dashboards improve the usability of rainfall forecast models for non-technical users. Although the study demonstrated how Shiny dashboards can be used to forecast rainfall, it examined only one model and did not conduct a thorough comparison. More generally, although ML-based rainfall prediction is becoming increasingly popular, few studies have focused specifically on Lagos using long-term datasets, and even fewer have combined rigorous model evaluation with accessible end-user deployment. In addition, the asymmetric cost of false negatives in tropical rainfall contexts is rarely considered in current studies: farmers and city planners suffer substantially more when they anticipate no rain and rainfall occurs. This study addresses these gaps by: (i) evaluating and comparing LR, RF, and SVM for daily rainfall classification in Lagos using a 22-year meteorological dataset; (ii) applying Youden's J threshold optimisation to prioritise sensitivity in a cost-asymmetric prediction context; and (iii) deploying all three models in an interactive R Shiny application accessible to non-technical users.

2. Materials and methods

This study used a dataset obtained from Visual Crossing (https://www.visualcrossing.com/weather-data), a weather forecasting website. The dataset spans 22 years and contains 8,314 observations with 12 variables, of which 11 are independent variables.

Data preprocessing involved three steps. First, temperature values were converted from degrees Fahrenheit to degrees Celsius to align with Nigerian measurement standards. Second, missing values were handled through imputation. Three variables contained missing values: wind direction (0.5%), sea-level pressure (1.3%), and visibility (2.6%). Mean imputation was applied to approximately normally distributed variables, and median imputation was applied to skewed variables. Third, the class distribution of the target variable (Rain/No Rain) was examined. Of the 8,314 observations, 4,415 were labelled "No Rain" and 3,899 were labelled "Rain", representing a 53% to 47% split. The imbalance was mild enough not to require external intervention. The dataset was partitioned into a training set (80%, $n = 6,651$) and a test set (20%, $n = 1,640$). The split was performed in chronological order to respect the temporal structure of the data and prevent future observations from leaking into the training set.

2.1. Model selection

Three classification algorithms were selected based on evidence from the literature and their complementary properties. LR was chosen as an interpretable baseline that quantifies predictor contributions through standardised coefficients, making it suitable for identifying the relative importance of meteorological variables. RF was selected for its ensemble structure, resistance to overfitting, and native feature-importance output through Gini impurity reduction. SVM was included because of its effectiveness in high-dimensional spaces and robustness to outliers through margin maximisation [9, 11].

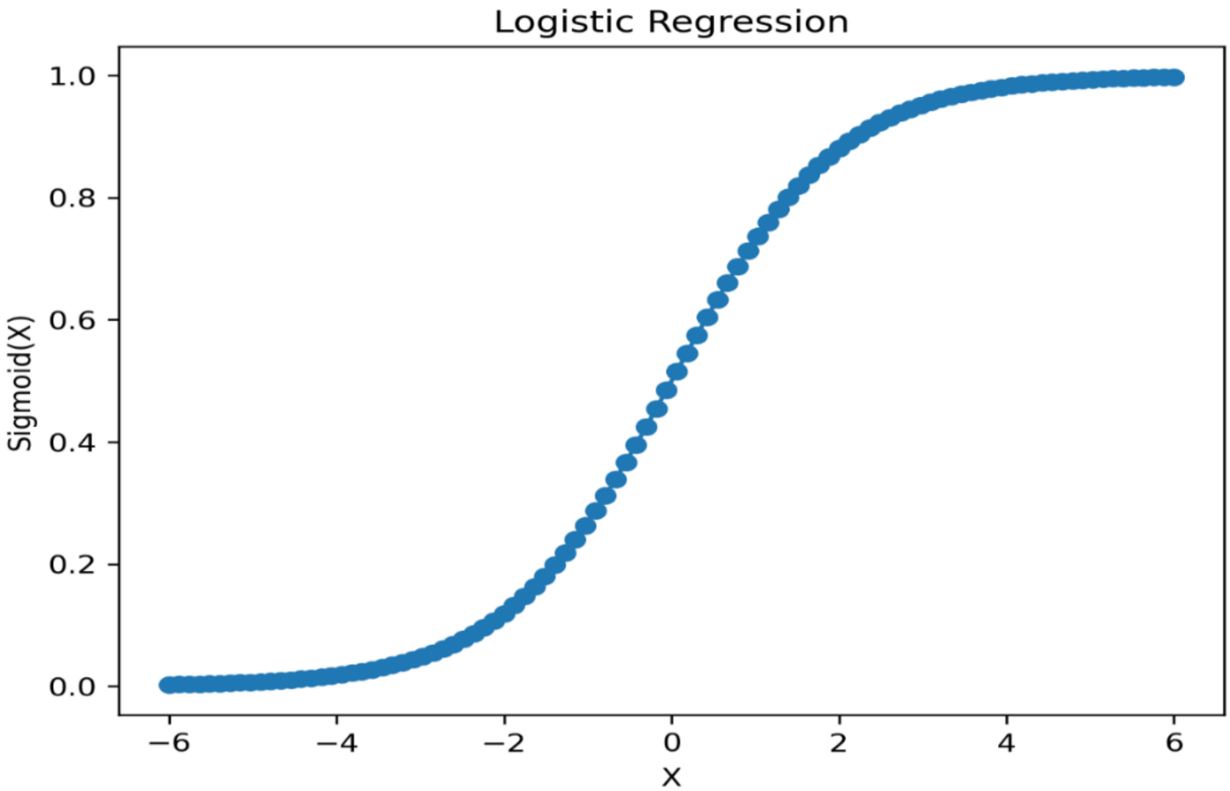

Figure 1: Logistic regression sigmoid curve.

2.2. Logistic regression

Logistic regression is an algorithm used for classification when the dependent variable is categorical [12]. It is suitable for this study because the response variable is dichotomous, categorised as Rain and No Rain. The regression model is defined in Eq. (1) as follows:

$$P(X) = \frac{1}{1 + \exp(-X\beta)}. \tag{1}$$

A linear regression line equation cannot be fitted because the dependent variable values are 0 and 1. The scatter plot can produce only an S-shaped curve, as shown in Figure 1.

Figure 1 shows the shape of a logistic regression curve. A rise in the log-odds increases the probability of an event occurring, while a decline in the log-odds decreases that probability. The general regression equation is not appropriate for fitting a logistic regression model because it can give values outside the probability range. Logistic regression instead transforms probabilities into log-odds, which can take real-number values, as shown in Eq. (2) and Eq. (3):

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right), \tag{2}$$

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki}. \tag{3}$$

2.3. Random forest classifier

Random forest is a sampling-with-replacement ML algorithm used for both regression and classification problems. It was developed by Breiman in 2001 and is based on combining many decision trees to reduce overfitting and improve predictive ability. The algorithm obtains the final prediction for classification problems through voting and takes an average of predictions for regression problems [13].

Let $D_b(X)$ denote the prediction from the b^{th} tree. The RF classifier is defined in Eq. (4):

$$\widehat{Y} = \text{majority vote}\{D_1(X), D_2(X), \ldots, D_B(X)\}. \tag{4}$$

The RF classifier structure is illustrated in Figure 2.

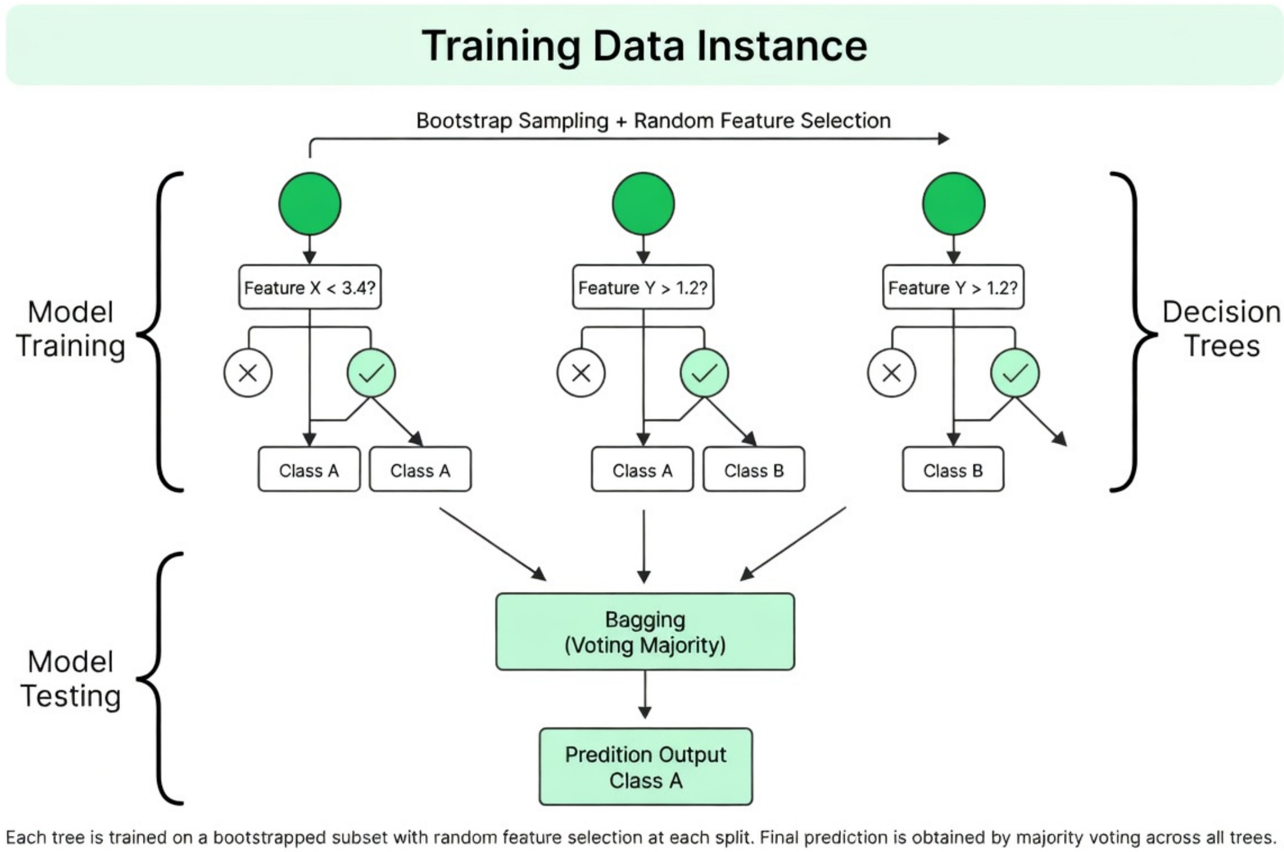

Figure 2: Random forest classifier illustration.

2.4. *Support vector machine*

Support vector machine, like LR and RF, is a supervised ML model that constructs a line called a hyperplane that maximises the distance between the different classes in a dataset [11]. SVM constructs a hyperplane that maximises the margin between classes, as shown in Eq. (5):

$$\mathbf{w}^{\mathrm{T}}\mathbf{x} + b = 0. \tag{5}$$

For non-linear classification, a radial basis function (RBF) kernel is used to map inputs into a higher-dimensional feature space, as shown in Eq. (6):

$$K(\mathbf{x}, \mathbf{y}) = \exp\left(-\gamma \|\mathbf{x} - \mathbf{y}\|^2\right). \tag{6}$$

The SVM classifier is illustrated in Figure 3.

2.5. *Youden's J statistic for threshold optimisation*

The default decision threshold of 0.5 in binary classification is based on the notion that the costs of false positives and false negatives are the same; that is, forecasting "No Rain" on a day that actually rains has the same consequence as predicting "Rain" on a dry day. Lagos, being in a tropical region, does not follow this notion. Predicting no rainfall when it actually rains (false negative) has a significantly higher cost than predicting rainfall when it does not rain (false positive), because unexpected rainfall deprives farmers and other planners of the opportunity for early preparation or preventive measures. This study uses Youden's J statistic [14] to estimate the ideal threshold for each classifier. The statistic is defined in Eq. (7):

$$J = \text{Sensitivity} + \text{Specificity} - 1. \tag{7}$$

The Youden's J-optimised decision thresholds for LR, RF, and SVM were determined to be 0.4908593, 0.4795, and 0.4057616, respectively. These thresholds replace the conventional 0.5 default and shift the models toward higher sensitivity, reducing the risk of false negatives in the Lagos context.

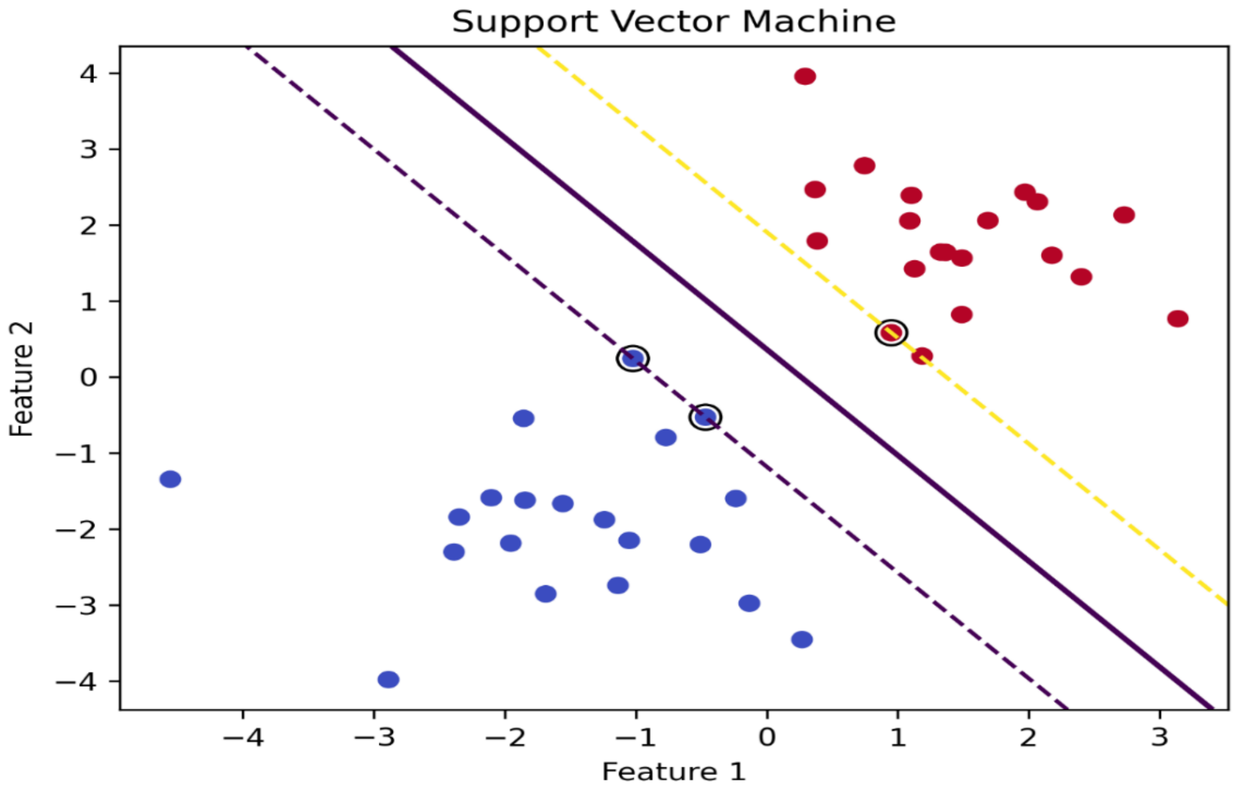

Figure 3: Support vector classifier.

Table 1: Confusion matrix.

	Actual	
Predicted	Rain	No Rain
Rain	True positives (TP)	False positives (FP)
No Rain	False negatives (FN)	True negatives (TN)

All three models underwent iterative hyperparameter tuning during training to achieve better accuracy. A confusion matrix, a performance-measurement table that compares the actual outcomes with the predicted outcomes, was used to compare how well each model performed. The generic confusion matrix is shown in Table 1. Other metrics used to evaluate model performance were accuracy, precision, sensitivity, F1 score, and AUC-ROC; their formulae are given in Eqs. (8)–(11).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{8}$$

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{9}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN}, \tag{10}$$

$$\text{F1 score} = 2 \times \frac{\text{Precision} \times \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}}. \tag{11}$$

3. Results and discussion

3.1. Correlation heatmap

Figure 4 presents the correlation heatmap of the 11 predictor variables.

Figure 4 shows strong positive correlations between temperature and dew point, which is consistent with the physical relationship between air temperature and atmospheric moisture content. Humidity and cloud cover also show a notable positive correlation,

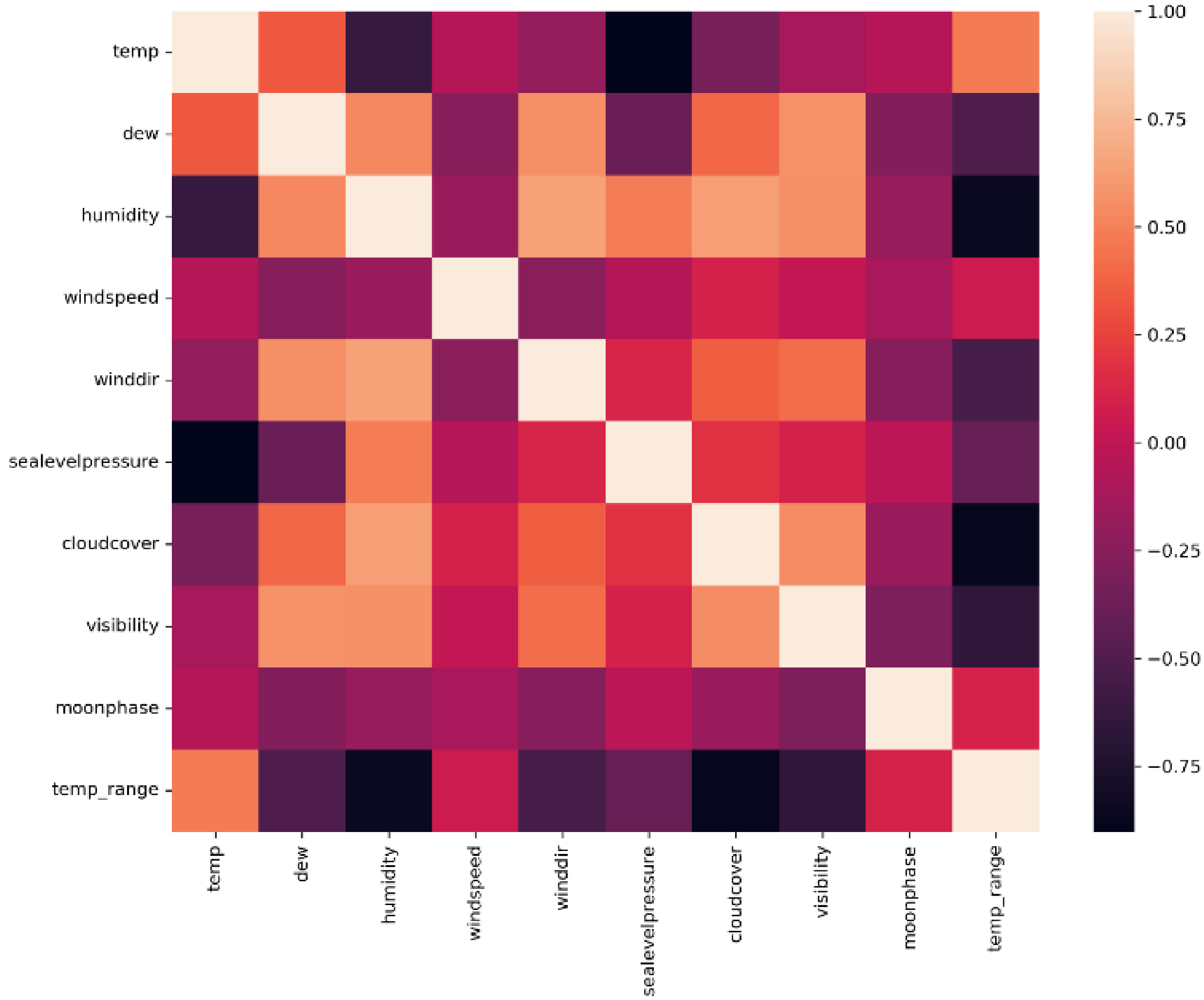

Figure 4: Correlation heatmap of the features.

Table 2: Logistic regression confusion matrix.

		Actual	
		No Rain	Rain
Predicted	No Rain	575	176
	Rain	269	620

suggesting that they capture overlapping meteorological information. These intercorrelations are worth noting because they may influence feature-importance rankings across models, particularly where collinear variables compete for explanatory weight.

3.2. Models

3.2.1. Logistic regression

The LR confusion matrix is presented in Table 2. The corresponding feature-importance plot is shown in Figure 5.

Figure 5 presents the feature importance of the LR model, quantified by absolute standardised regression coefficients (absolute Z-scores), which measure each predictor's contribution to the probability of rainfall after accounting for scale differences. Visibility, temperature, dew point, and the engineered temperature range emerged as the most dominant predictors. Visibility's prominence in LR may reflect its sensitivity to atmospheric moisture conditions; reduced visibility is commonly associated with fog and pre-rainfall humidity accumulation in coastal tropical cities such as Lagos. The strong contribution of dew point aligns with the physical principle that dew-point temperature is a direct measure of atmospheric moisture, and rainfall becomes increasingly likely as air temperature

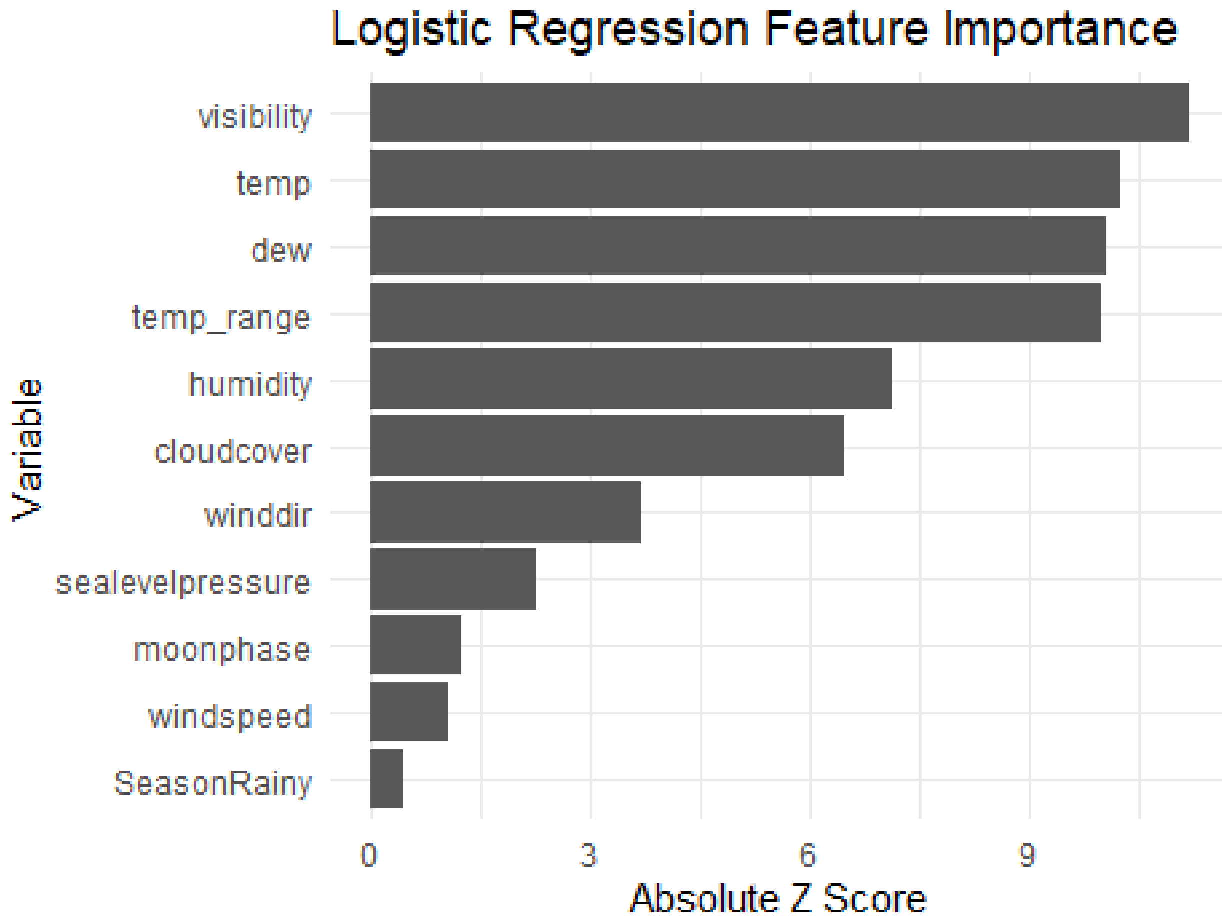

Figure 5: Logistic regression feature importance.

Table 3: Random forest confusion matrix.

		Actual	
		No Rain	Rain
Predicted	No Rain	620	165
	Rain	224	631

approaches the dew point. The significance suggests that days with compressed diurnal temperature ranges are more associated with rainfall events.

Wind speed, moon phase, and season recorded *p*-values greater than 0.05, indicating no statistically significant contribution to the model's predictive power. The negligible importance of moon phase is unsurprising because the gravitational influence of the moon on atmospheric moisture is negligible at the local scale of daily rainfall prediction. The near-zero importance of the season variable, despite Lagos having clearly defined wet and dry seasons, is notable. A likely explanation is that the other predictors, particularly humidity, dew point, and cloud cover, already implicitly encode seasonal variation. In other words, the season variable becomes redundant because the model can infer the season from the atmospheric conditions themselves.

3.2.2. Random forest

The RF confusion matrix is presented in Table 3. The corresponding feature-importance chart is shown in Figure 6.

Figure 6 shows the feature-importance chart for the RF model, scaled with respect to the most significant variable and calculated using the mean reduction in Gini impurity. Humidity stands out with a scaled significance score of 100 as the most important predictor. Cloud cover is the second most important predictor, with a score of almost 68, closely followed by visibility, temperature, wind direction, temperature range, and dew point. Season again shows little to no significance, consistent with the LR model. Humidity's dominant position in the RF model has a clear physical basis: relative humidity directly quantifies the amount of water

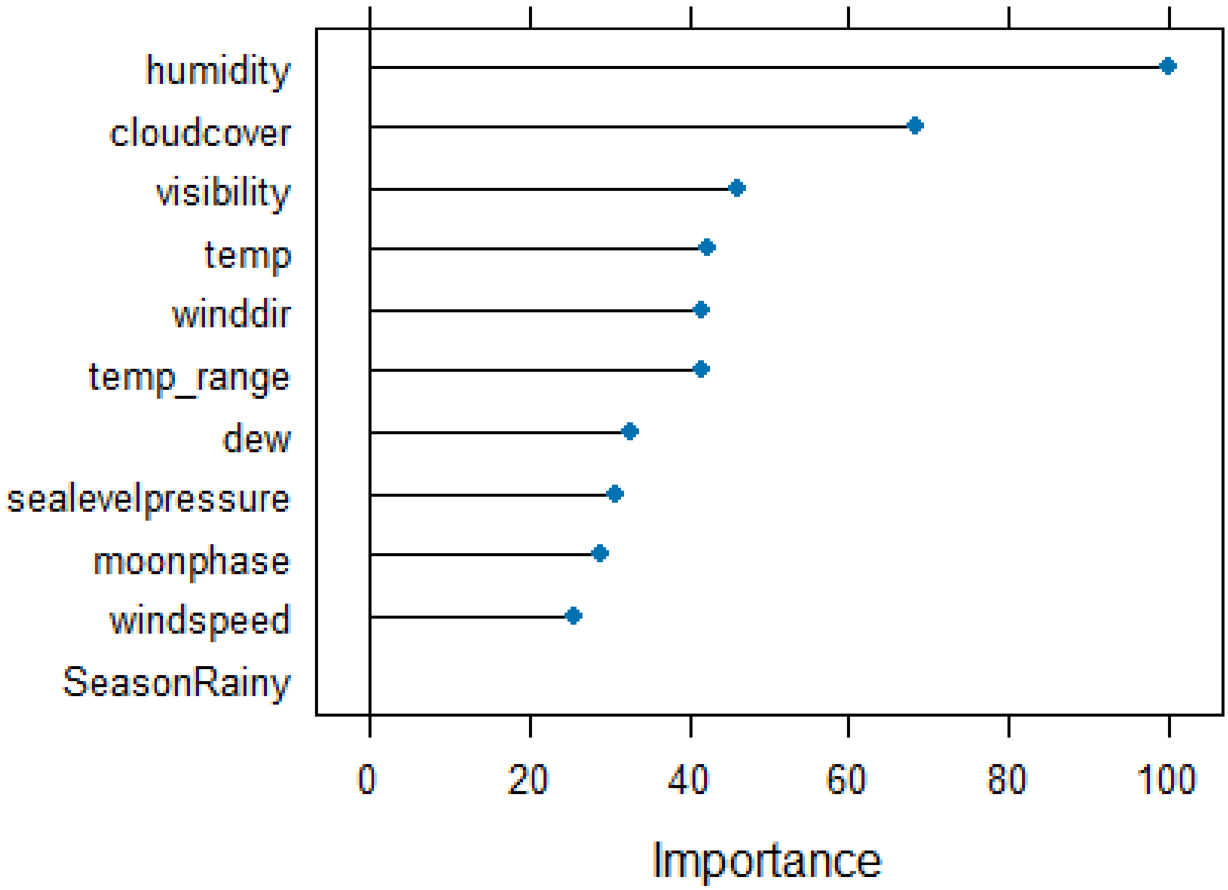

Figure 6: Random forest feature importance.

vapour in the atmosphere relative to the saturation point. As humidity approaches 100%, the atmosphere's capacity to hold water vapour is exhausted, making precipitation increasingly likely. This makes humidity the most direct atmospheric precursor to rainfall and explains why the RF model, which captures non-linear feature interactions through ensemble decision trees, ranks it highest. Cloud cover's second-place ranking similarly reflects a physical relationship: cloud formation is a prerequisite for most rainfall events, particularly the convective rainfall common in Lagos's tropical climate.

The discrepancy between LR and RF feature rankings is also worth noting. Visibility ranks highly in LR but is less prominent in RF, while humidity dominates RF but not LR. This divergence reflects a fundamental difference in how the two models process information. LR captures linear marginal effects, making visibility, which has a strong linear inverse relationship with moisture, influential. RF, however, captures interaction effects through tree splits, allowing it to use the more complex, non-linear relationship between humidity and rainfall that a linear model cannot fully exploit.

3.2.3. Support vector machine

The SVM confusion matrix is presented in Table 4. The corresponding feature-importance chart is shown in Figure 7.

Figure 7 presents the feature importance of the tuned SVM model trained with an RBF kernel ($C = 2$, $\sigma = 0.009674655$). Temperature and dew point show the most predictive power for SVM, reflecting their strong co-variation. Wind direction, wind speed, and moon phase have the lowest relevance ratings. Moon phase shows no significance for the third time in a row, indicating that it adds little predictive information to any of the three models. The prominence of temperature and dew point in the SVM model is consistent with their strong co-variation: dew point is mathematically bounded by air temperature, and the difference between them, the dew-point depression, is a classical meteorological indicator of atmospheric saturation. The SVM model's margin-maximising mechanism likely identified this paired relationship as a strong separating boundary between rain and no-rain classes. The relatively

Table 4: Support vector machine confusion matrix.

		Actual	
		No Rain	Rain
Predicted	No Rain	653	229
	Rain	191	567

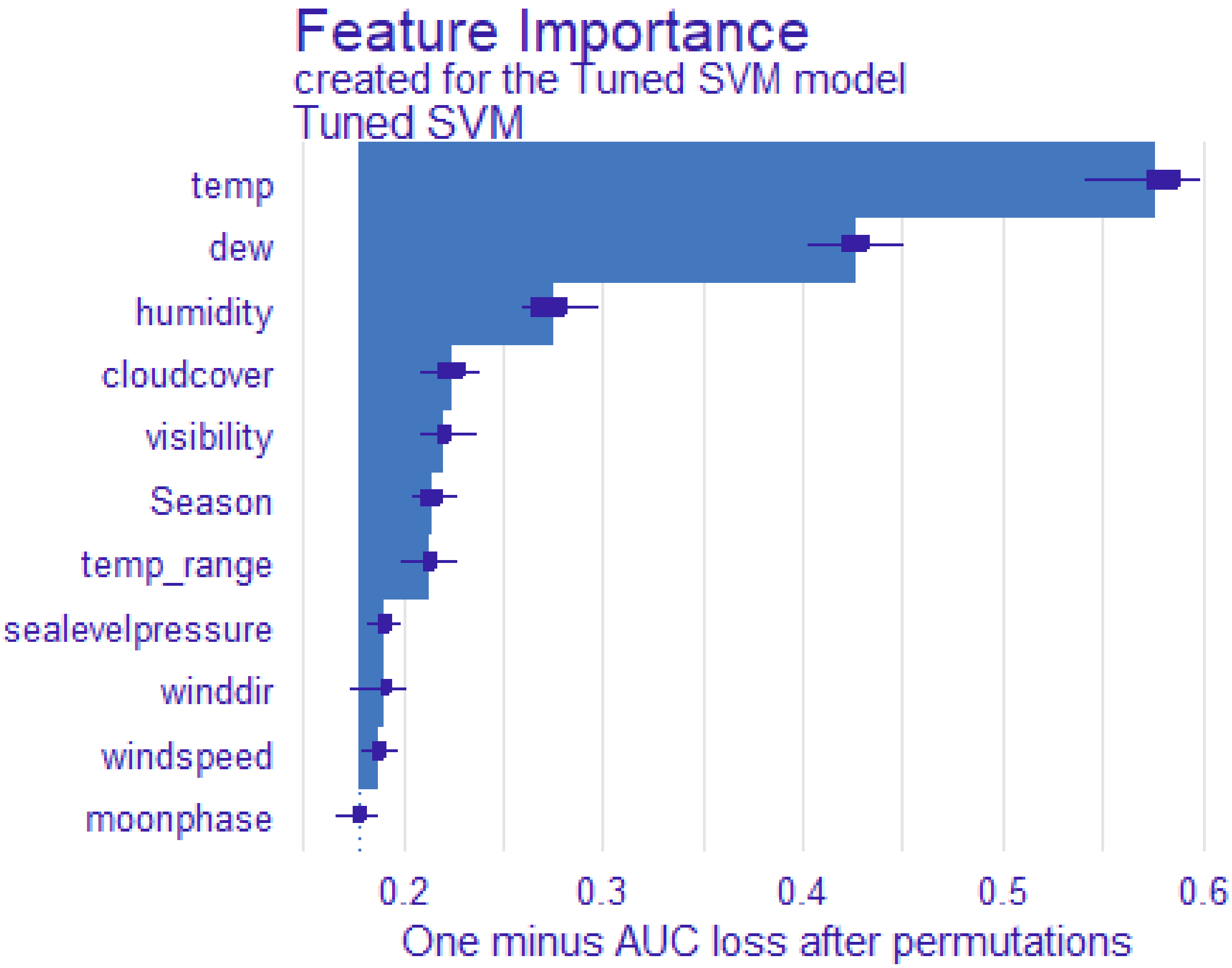

Figure 7: Support vector machine feature importance.

lower prominence of humidity in SVM compared with RF may reflect the sensitivity of the RBF kernel to feature scaling and intercorrelation, where humidity and dew point overlap in information content; the SVM may distribute weight differently from RF's tree-based splitting.

Moon phase registers no predictive importance in SVM, consistent with its performance in LR and RF. The persistent absence of moon phase importance across all three models, which use fundamentally different learning mechanisms, provides strong convergent evidence that lunar cycles carry no meaningful signal for daily rainfall prediction in Lagos at the daily timescale studied here.

3.3. Model comparison

Table 5 and Figure 8 present a comparative summary of all three models across five evaluation metrics.

Random forest achieved the highest accuracy (76.28%), recall (79.27%), F1-Score (76.44%), and AUC-ROC (84.54%), making it the strongest overall performer. Its high recall is particularly important in the Lagos context, where false negatives—failing to predict actual rain—carry greater real-world consequences than false alarms and its high F1-score, reinforces its position as the best-performing model when precision and recall are balanced jointly. SVM achieved the highest precision (74.80%), indicating

Table 5: Model comparison across evaluation metrics.

Model	Accuracy	Precision	Recall	F1 score	AUC-ROC
Logistic regression	0.7299	0.6981	0.7814	0.7366	0.8001
Random forest	0.7628	0.7380	0.7927	0.7644	0.8454
SVM	0.7439	0.7480	0.7123	0.7297	0.8215

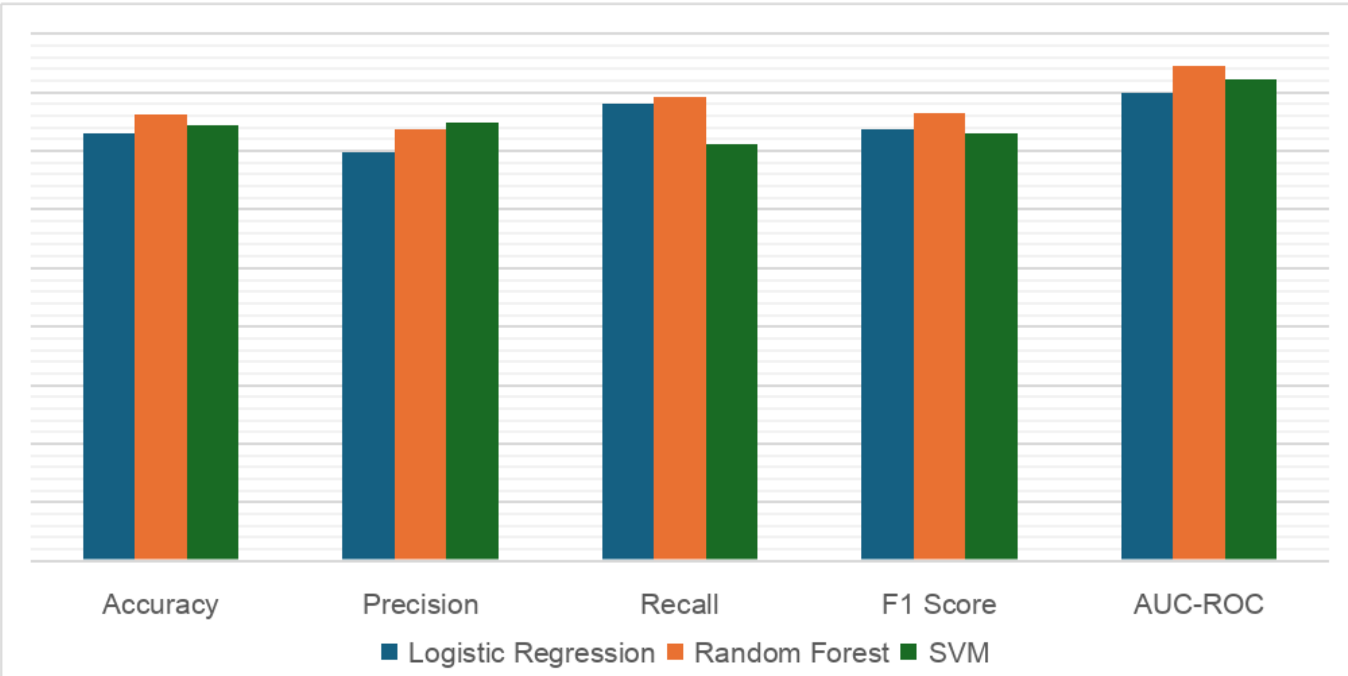

Figure 8: Model comparison visualisation.

that when SVM predicts rain, it is the most reliable of the three models, although it misses more actual rain days than RF. Logistic regression, despite being the simplest model, achieved a recall of 78.14%, nearly matching RF, while maintaining a competitive AUC-ROC (80.01%). This demonstrates that a linear model retains meaningful predictive power for this dataset and remains a viable choice where interpretability is prioritised over marginal performance gains.

The AUC-ROC values above 0.80 for all three models indicate good discriminative ability; the models meaningfully distinguish between rain and no-rain days rather than relying on class frequency alone. These results are consistent with Ejike *et al*. [9], who similarly found RF and SVM to outperform LR on tropical meteorological data, although the margin of difference here is modest, reinforcing LR's value as an interpretable baseline.

3.4. Implementation of the Shiny app for rainfall prediction

For easy access to the prediction models by non-technical users, a Shiny app was used to deploy the three trained models. The app provides an interactive dashboard where hypothetical or real-time weather data can be entered for real-time weather prediction. The prediction result is displayed as "Rain" or "No Rain" to support user understanding, alongside the evaluation metrics for each model. On a separate tab, a visual representation of all five evaluation metrics across the three models is shown. The LR, RF, SVM, and model-comparison tabs are shown in Figures 9–12, respectively.

The app can be accessed at http://hardeybaz.shinyapps.io/Rainfall_Prediction_App.

4. Conclusion and future research

This study demonstrated that ML classification models can effectively predict daily rainfall occurrence in Lagos, Nigeria, using routinely collected meteorological data. Among the three classifiers evaluated, RF emerged as the strongest overall performer, achieving the highest accuracy (76.28%), recall (79.27%), F1-score (76.44%), and AUC-ROC (84.54%). Importantly, all three models achieved AUC-ROC values exceeding 0.80, indicating that each model possesses meaningful discriminative ability beyond chance classification. The application of Youden's J statistic for threshold optimisation improved model sensitivity by prioritising the

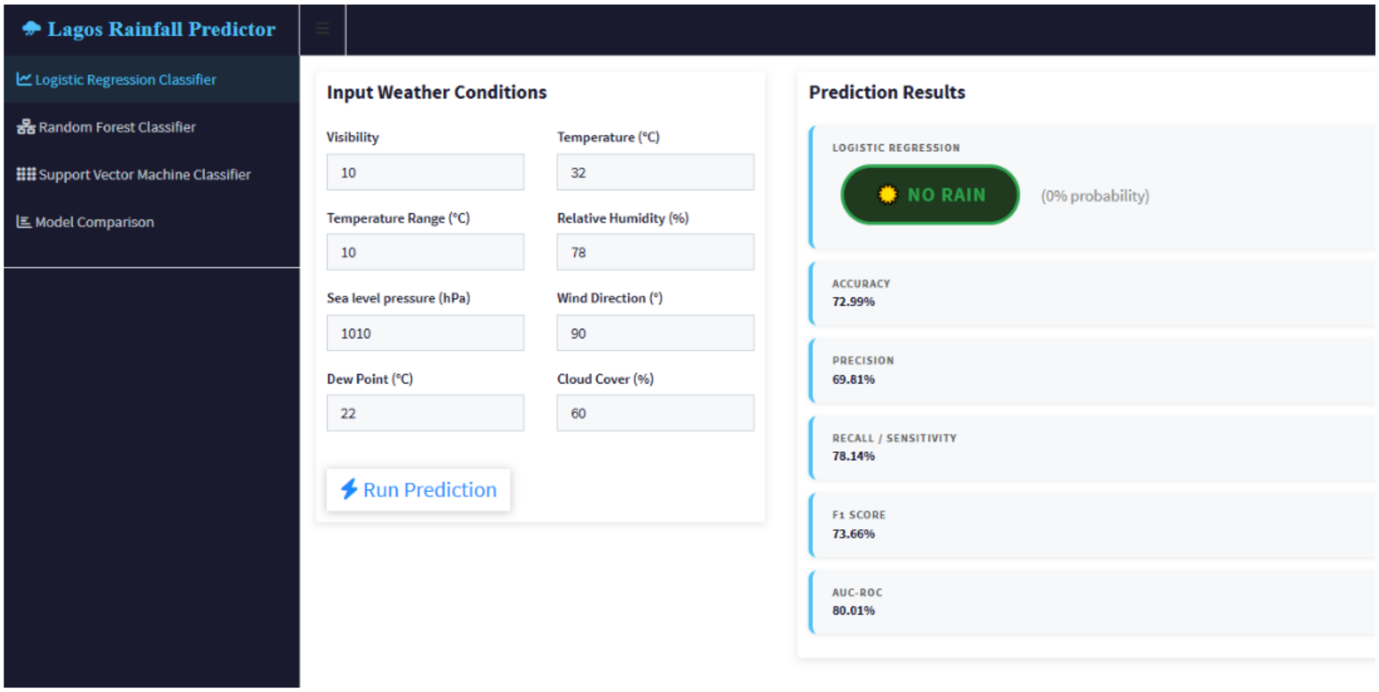

Figure 9: Logistic regression tab.

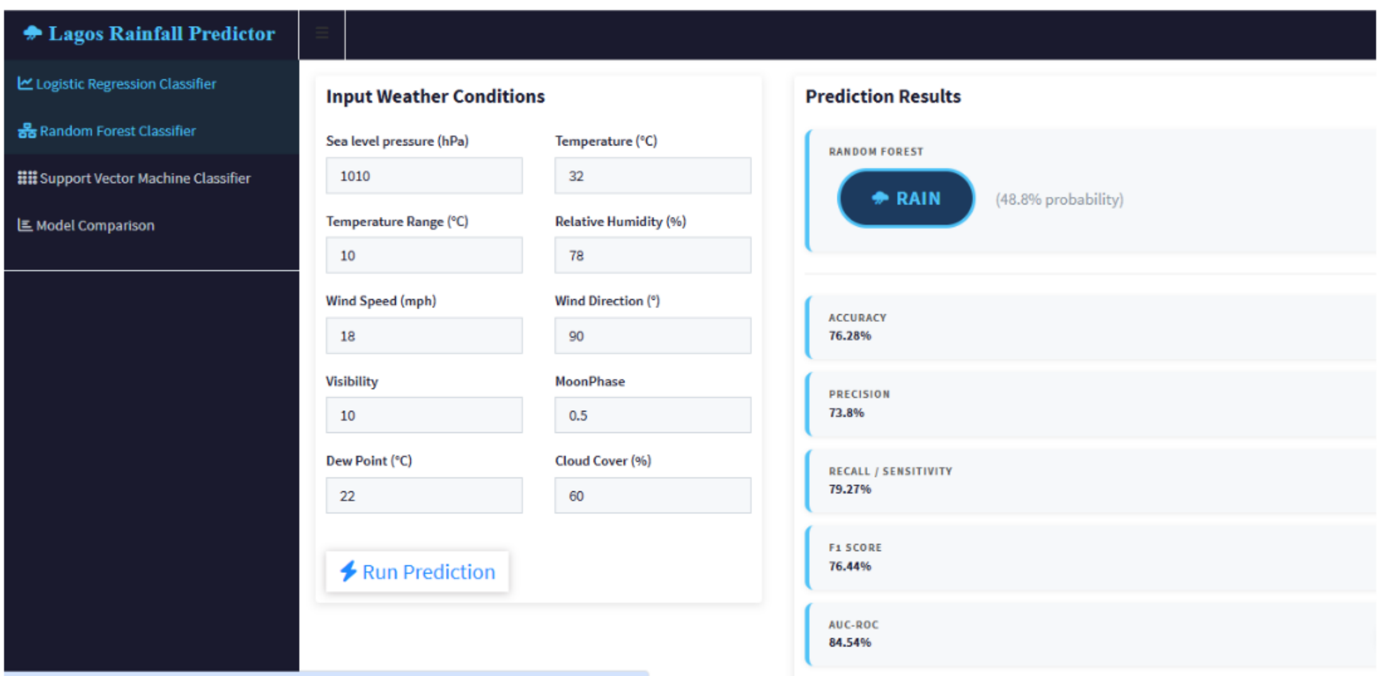

Figure 10: Random forest classifier tab.

detection of actual rain days, which is critical in a city where undetected rainfall events have documented consequences for flooding, agriculture, and urban infrastructure.

The findings carry practical implications for meteorological forecasting and urban planning in Lagos and comparable tropical cities. The dominance of humidity, cloud cover, dew point, and visibility as predictors across all three models suggests that monitoring these variables provides the strongest early signals for imminent rainfall. The near-zero importance of the engineered season variable, despite Lagos's wet and dry seasons, indicates that atmospheric variables already capture seasonal information implicitly,

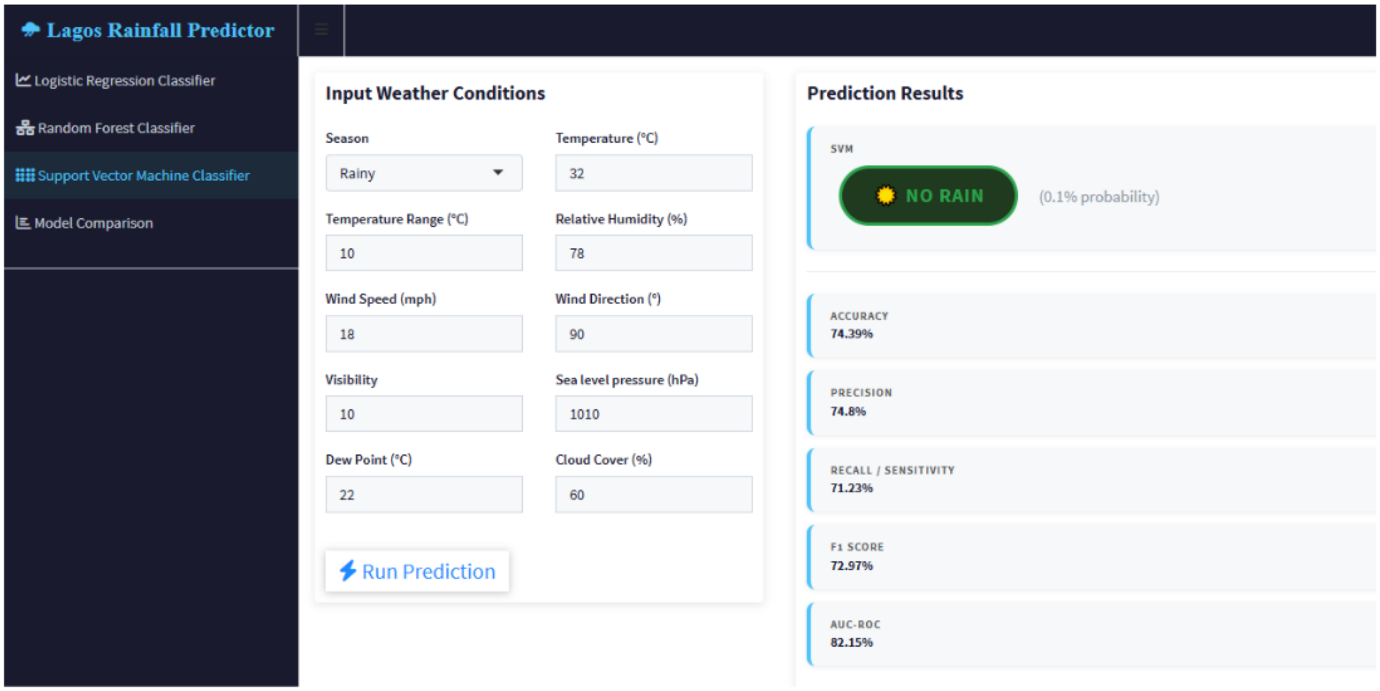

Figure 11: Support vector machine classifier tab.

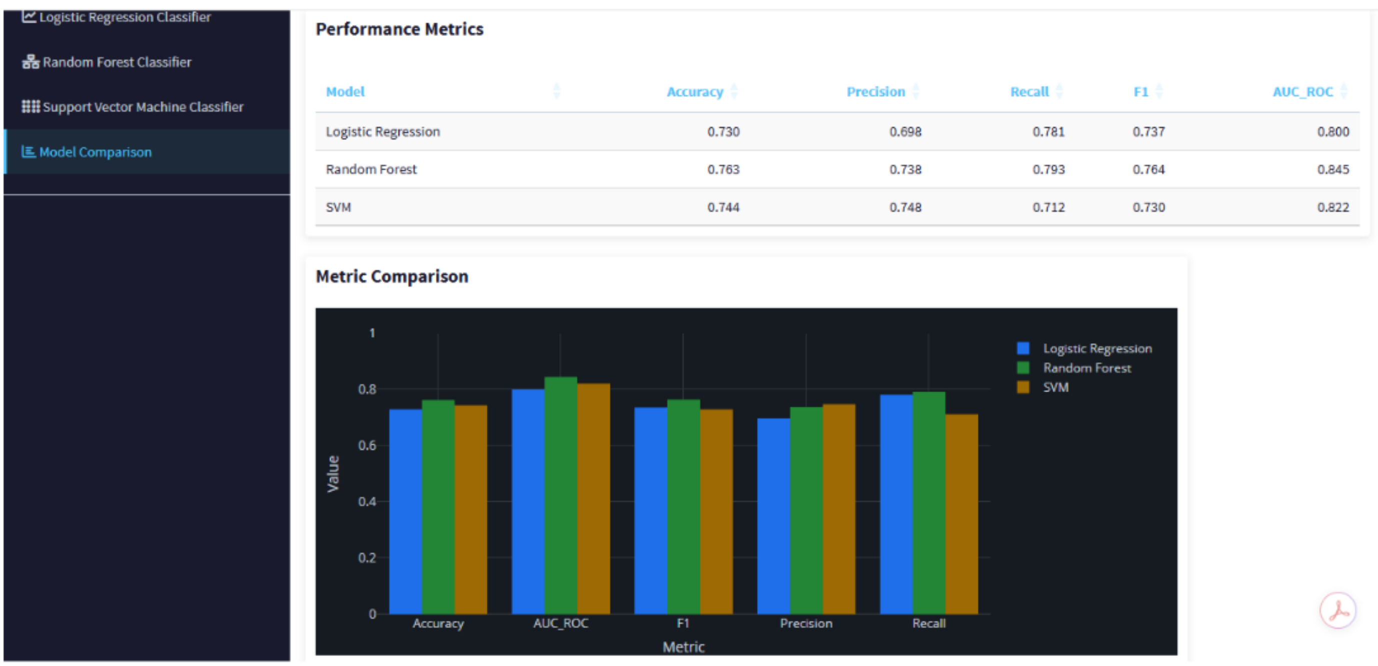

Figure 12: Model comparison tab.

which is a useful insight for future feature selection in similar studies. The consistent absence of moon phase importance across all three models provides strong evidence that lunar cycles carry no predictive signal for rainfall in Lagos. The deployment of all three models in an interactive R Shiny application bridges the gap between technical model development and practical usability, enabling agricultural planners, emergency responders, and policymakers to access real-time predictions without statistical expertise.

This study has several limitations that should be acknowledged. First, the dataset was sourced from a single provider, Visual Crossing, and cross-validation against Nigeria Meteorological Agency (NiMet) station data was not performed. Second, although the 80/20 data split is standard in ML classification, its chronological application to time-series meteorological data may introduce temporal considerations that a walk-forward cross-validation approach would better address. Finally, the models were trained and evaluated on Lagos data alone, and their generalisability to other Nigerian cities with different climatic profiles has not been tested.

Future research should explore deep-learning architectures, such as LSTM networks, which are better suited to capturing long-range temporal dependencies in daily weather sequences. Integrating real-time data feeds from NiMet weather stations into the Shiny application would enhance its operational value for practitioners. Extending the framework to other Nigerian cities with distinct climatic profiles, such as Kano or Port Harcourt, would test the generalisability of these findings. In addition, incorporating ensemble stacking of the three classifiers, rather than selecting a single best model, may yield further performance improvements.

Data availability

The data used in this study are available at https://www.visualcrossing.com/weather-query-builder/lagos/?v=wizard.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this manuscript.

Funding

The authors received no external funding for this study.

References

[1] S. Q. Dotse, I. Larbi, A. M. Limantol & L. C. De Silva, "A review of the application of hybrid machine learning models to improve rainfall prediction", Modeling Earth Systems and Environment **10** (2024) 19. https://doi.org/10.1007/s40808-023-01835-x.

[2] A. Parmar, K. Mistree & M. Sompura, "Machine learning techniques for rainfall prediction: A review", 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), Coimbatore, India, 2017, pp. 425–430. Available online: https://www.researchgate.net/publication/319503839_Machine_Learning_Techniques_For_Rainfall_Prediction_A_Review.

[3] S. D. Latif, N. A. B. Hazrin, C. H. Koo, J. L. Ng, B. Chaplot, Y. F. Huang, A. El-Shafie & A. N. Ahmed, "Assessing rainfall prediction models: Exploring the advantages of machine learning and remote sensing approaches", Alexandria Engineering Journal **82** (2023) 16. https://doi.org/10.1016/j.aej.2023.09.060.

[4] U. C. Nkwunonwo, M. Whitworth & B. Baily, "A review and critical analysis of the efforts towards urban flood risk management in the Lagos region of Nigeria", Natural Hazards and Earth System Sciences **16** (2016) 349. https://doi.org/10.5194/nhess-16-349-2016.

[5] M. I. Jordan & T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects", Science **349** (2015) 255. https://doi.org/10.1126/science.aaa8415.

[6] A. Patel, N. Keriwala, N. Soni, U. Goel, R. Bhoj, Y. Adhyaru & S. M. Yadav, "Rainfall prediction using machine learning techniques for the Sabarmati River Basin, Gujarat, India", Journal of Engineering Science and Technology Review **16** (2023) 101. https://doi.org/10.25103/jestr.161.13.

[7] U. Shah, S. Garg, N. Sisodiya, N. Dube & S. Sharma, "Rainfall prediction: Accuracy enhancement using machine learning and forecasting techniques", 2018 Fifth International Conference on Parallel, Distributed and Grid Computing (PDGC), Solan, India, 2018, pp. 776–782. https://doi.org/10.1109/PDGC.2018.8745763.

[8] M. El Hafyani, K. El Himdi & S.-E. El Adlouni, "Improving monthly precipitation prediction accuracy using machine learning models: A multi-view stacking learning technique", Frontiers in Water **6** (2024) 1378598. https://doi.org/10.3389/frwa.2024.1378598.

[9] O. Ejike, D. Ndzi & M. Z. Shakir, "Comparative study of machine learning-based rainfall prediction in tropical and temperate climates", Climate **13** (2025) 167. https://doi.org/10.3390/cli13080167.

[10] S. Bahrom, "SmartForecast: An interactive R Shiny dashboard for rainfall forecasting in Subang", APS Proceedings **24** (2025) 8. https://doi.org/10.5281/zenodo.16788019.

[11] C. Cortes & V. Vapnik, "Support-vector networks", Machine Learning **20** (1995) 273. https://doi.org/10.1007/BF00994018.

[12] D. W. Hosmer, S. Lemeshow & R. X. Sturdivant, *Applied Logistic Regression*, 3rd ed., Wiley, Hoboken, USA, 2013. https://doi.org/10.1002/9781118548387.

[13] L. Breiman, "Random forests", Machine Learning **45** (2001) 5. https://doi.org/10.1023/A:1010933404324.

[14] W. J. Youden, "Index for rating diagnostic tests", Cancer **3** (1950) 32. https://doi.org/10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3.