BAREKENG: Journal of Mathematics and Its Applications
December 2025 Volume 19 Issue 4 Page 2765–2776
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol19no4pp2765-2776

# COMPARATIVE ANALYSIS OF MACHINE LEARNING MODELS FOR RAINFALL CLASSIFICATION IN YOGYAKARTA

**Dina Tri Utari[1*], Ghalang Rambu Putera Palage[2], Faiz Fadhlirobby[3], Artheta Bimo Nuswantoro[4]**

[1,2,3,4]*Department of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Islam Indonesia*
*Jln. Kaliurang Km 14.5, Sleman, Yogyakarta, 55584, Indonesia*

*Corresponding author's e-mail: * **dina.t.utari@uii.ac.id***

***Article History:***

*Received: 24th February 2025*
*Revised: 15th April 2025*
*Accepted: 19th May 2025*
*Available online: 1st September 2025*

***Keywords:***

*Decision Tree;*
*KNN;*
*Logistic Regression;*
*Rainfall Classification;*
*Weather Variables;*
*Yogyakarta.*

***ABSTRACT***

*Precise rainfall classification is most important for meteorological forecasting and disaster risk mitigation, particularly in regions such as Yogyakarta, which are vulnerable to extreme weather events. Although previous studies have examined rainfall classification through the lens of meteorological variables, a notable lack of research has systematically evaluated the effectiveness of diverse machine learning algorithms for categorizing rainfall types within this specific locale. This study aims to rectify this gap by incorporating essential weather variables, specifically temperature, humidity, atmospheric pressure, and precipitation, into predictive models that utilize K-Nearest Neighbors (KNN), decision trees, and logistic regression techniques. Among the evaluated models, the decision tree demonstrated the highest degree of accuracy across both training and testing datasets. An examination of feature significance indicated that precipitation emerged as the most pivotal variable, aligning with the fundamental physical mechanisms associated with rainfall. This study contributes significantly to the evolving field of weather informatics by illustrating the utility of machine learning approaches in classifying regional rainfall. However, the parameters of this research are limited to specific meteorological variables and do not account for spatial or temporal variations, which could potentially influence the model's broader applicability. Future research endeavors could augment this framework by integrating remote sensing data and methodologies for spatiotemporal modeling.*



***How to cite this article:***

D. T. Utari, G. R. P. Palage, F. Fadhlirobby, and A. B. Nuswantoro, "COMPARATIVE ANALYSIS OF MACHINE LEARNING MODELS FOR RAINFALL CLASSIFICATION IN YOGYAKARTA," *BAREKENG: J. Math. & App.*, vol. 19, iss. 4, pp. 2765-2776, December, 2025.


Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*

**Research Article • Open Access**

## 1. INTRODUCTION

Weather classification, an essential meteorological component, entails systematically categorizing atmospheric phenomena based on empirical evidence rather than predictive models of forthcoming conditions. Conventional classification techniques were predominantly dependent on observational heuristics, such as the interpretation of a red sunset as an indication of forthcoming clear weather, which, despite their intuitive attractiveness, were deficient in scientific rigor and methodological consistency. This circumstance prompted the development of data-centric approaches bolstered by technological innovations and systematic observational practices [1]. Contemporary meteorological sciences utilize a comprehensive temperature, humidity, dew point, wind velocity and orientation, solar irradiance, and precipitation levels gathered from terrestrial observation stations, satellite systems, and numerical weather prediction (NWP) frameworks [2]. While conventional forecasting predominantly seeks to project forthcoming weather conditions, the process of weather classification is crucial for the identification of anomalies, the establishment of early warning mechanisms, and the ongoing assessment of climatic changes. The incorporation of machine learning (ML) methodologies, particularly classification algorithms, has facilitated the identification of intricate patterns within multivariate datasets, thereby enhancing both the precision and promptness of decision-making in sectors significantly influenced by climatic variables.

Numerous investigations have elucidated the efficacy of machine learning methodologies, including logistic regression, decision trees, and K-Nearest Neighbors (KNN), in meteorological classification tasks. These analytical models leverage historical meteorological datasets to forecast future weather classifications predicated upon a defined array of input features, with their predictive accuracy typically enhancing as the quantity of variables increases [3]. Classification methodologies rooted in machine learning have exhibited notable potential in differentiating various types of precipitation or rainfall, thereby facilitating enhanced resource allocation and disaster management strategies [4], [5], [6].

In Indonesia, specifically in Yogyakarta, rainfall patterns are integral to various sectors, including agriculture, water resource management, and disaster preparedness. The region is characterized by two primary types of rainfall, each exhibiting distinct microphysical properties and resultant implications. The precise categorization of these rainfall types is paramount for enhancing flood forecasting, optimizing irrigation practices, and advancing crop resilience strategies.

Combining weather variables, including temperature, humidity, and atmospheric pressure, with machine learning methodologies can yield a more intricate comprehension of rainfall dynamics in Yogyakarta. By utilizing historical meteorological data alongside real-time observations, this study seeks to establish robust predictive models that effectively categorize types of rainfall. Such classifications are imperative for enhancing quantitative precipitation estimations, which are important for agricultural strategizing and flood risk mitigation [7], [8]. Another investigation in the climatic field analyzed precipitation and thermal patterns in North Sumatra Province. The findings indicated a progressive augmentation in precipitation levels, with Medan exhibiting higher mean monthly values and a positive thermal trend aligning with the broader weather change phenomenon. Nonetheless, the relationship between precipitation and temperature exhibited a lack of robustness. These findings are instrumental in formulating effective adaptation strategies to mitigate the impacts of climate change within the region [9].

Incorporating ML algorithms into meteorological forecasting, particularly in rainfall prediction, represents a notable progression in the field. The investigation introduced a service-oriented architecture for meteorological forecasting employing Artificial Neural Networks (ANN) alongside Decision Tree methodologies. The findings indicated an enhancement in the precision of average weather variable classifications [10]. Research conducted by Adhane demonstrated that the amalgamation of ANN with fuzzy logic yields superior results compared to conventional techniques in meteorological forecasting, utilizing parameters such as temperature, humidity, pressure, and wind velocity [11]. Furthermore, another study applied six meteorological parameters in conjunction with Learning Vector Quantization (LVQ) to classify rainfall. Despite the successful implementation of the LVQ methodology, the classification outcomes were deemed inadequate [12]. This presents a clear research gap in leveraging more robust ML algorithms to classify rainfall types with higher accuracy and relevance to local weather conditions.

This study addresses this gap by constructing a rainfall classification model predicated on historical meteorological data from Yogyakarta, incorporating essential variables such as temperature, humidity, and atmospheric pressure. The principal objective is to assess the efficacy of ML algorithms in precisely differentiating between types of rainfall phenomena. This classification is a fundamental precursor to

enhancing quantitative precipitation estimation and bolstering early warning systems in regions susceptible to flooding. For example, discerning whether an impending rain event is convective or stratiform can assist farmers in preparing for the consequent effects on soil moisture and agricultural health [13]. Furthermore, this research can potentially improve flood forecasting and management initiatives. By categorizing rainfall occurrences according to their characteristics, governing bodies can more effectively predict flooding hazards linked to various types of precipitation. Convective rainfall, for instance, is frequently correlated with intense, short-duration incidents that may result in flash floods, whereas stratiform rainfall generally exhibits a more extended duration and reduced intensity [4], [14]. Enhanced classification facilitates timely interventions and the judicious allocation of resources during extreme meteorological events.

## 2. RESEARCH METHODS

This study uses meteorological variables from the Yogyakarta area to classify rainfall types such as cloudy, light rain, moderate rain, heavy rain, and very heavy rain. The chosen input parameters encompass precipitation, temperature, humidity, duration of sunshine, wind speed, and wind direction, which serve as critical indicators of atmospheric conditions pertinent to rainfall classification. These variables were chosen owing to their significant correlation with precipitation attributes in tropical climatic conditions. The preprocessing procedures encompassed normalization, imputation of missing values, and the categorical transformation of wind direction to improve interpretability. The classification of rainfall types and the selection of influencing variables were based on predefined threshold parameters provided by the Yogyakarta Meteorology, Climatology, and Geophysics Agency (BMKG).

This study focuses strictly on classification methodologies rather than predictive forecasting, thereby allowing for determining rainfall types for the same day contingent upon extant meteorological conditions, without extrapolating future climatic trends. Temporal granularity is established daily, and the principal objective is to facilitate immediate decision-making processes rather than engage in prolonged climatic forecasting endeavors. Several constraints have been recognized: the predictive models, particularly the K-Nearest Neighbors (KNN) algorithm, exhibit a heightened sensitivity to noise, decision trees are susceptible to overfitting unless adequately regularized, and incomplete or inconsistent records from meteorological stations may compromise the integrity of the data. Furthermore, given that the model is exclusively trained on datasets from Yogyakarta, its generalizability may be constrained to other geographical areas characterized by divergent climatic conditions.

### 2.1 K-Nearest Neighbor

K-Nearest Neighbor (KNN) is an instance-based learning framework prevalently employed for classification and regression. This algorithm functions by allocating data points into distinct categories based on their proximity to preexisting labeled instances within the training dataset. The foundational tenet of KNN posits that analogous data points typically reside nearby within the feature space, thereby facilitating classification determinations based on the predominant class of neighboring instances [15].

One of the foremost benefits of the KNN algorithm is its straightforward implementation and high degree of interpretability. It necessitates minimal training duration, as it merely retains the training dataset and executes computations at the point of prediction [16]. Furthermore, KNN demonstrates a proficient capability to address multi-class classification challenges, rendering it adaptable for various applications, such as meteorological forecasting, particularly in classification tasks.

The KNN methodology ascertains the classification of a specific test instance by examining its $k$ nearest neighbors, wherein $k$ is a user-specified parameter that signifies the number of the closest training examples to be considered. The classification procedure is executed utilizing various distance metrics, with the Euclidean distance being one of the most extensively utilized. This metric quantifies the straight-line distance between two points in Euclidean space, thereby measuring resemblance among data points.

From a mathematical perspective, the Euclidean distance between two points $P(x_1, y_1)$ and $Q(x_2, y_2)$ within a two-dimensional space can be expressed as:

$$d(P, Q) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \quad (1)$$

In the context of higher-dimensional feature spaces, the Euclidean distance extends to:

$$d(P,Q) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (2)$$

where $x_i$ and $y_i$ denote the feature values of the two data points across $n$ dimensions.

The following steps outline the process of implementing KNN for classification:

1. Determine the optimal number of neighbors $k$ for classification. A small value of $k$ can make the model sensitive to noise, while a large value may oversmooth the decision boundary. Cross-validation techniques can be used to find the best $k$ that minimizes classification error **[17]**.
2. For each instance in the dataset, the KNN algorithm calculates the distance between the new data point (the instance to be classified) and all other points in the training dataset.
3. After calculating the distances, the algorithm identifies the $k$ nearest neighbors to the input data point based on the calculated distances. These neighbors will be used to determine the classification of the new instance **[18]**.
4. Once the nearest neighbors are identified, the algorithm employs a voting mechanism to classify the new instance. Each neighbor votes for its class, and the class with the majority votes is assigned to the new data point. In the case of a tie, various strategies can be employed, such as choosing the class of the nearest neighbor or using weighted voting based on distance **[19]**.

### 2.2 Decision Tree

The algorithm that incorporates condition control statements is called a decision tree. It assists in discerning a strategy or methodology most likely to achieve the desired objective. A decision tree (DT) structure contains internal nodes indicating that the structure is a test on a particular attribute. At the same time, the branches depict the experiment's outcomes, and each leaf node signifies a class label. The pathway from the roots to the leaves represents the classification rules. Three types of nodes exist: decision nodes, chance nodes, and end nodes **[20]**. The Algorithm functions in the following steps:

1. When all scenarios are categorized under the same class, the resulting tree is designated as a leaf, and this leaf is subsequently assigned the Class label once more.
2. For each value, one must compute the necessary metrics derived from the testing parameters and subsequently ascertain the gain of information garnered from examining these parameters.
3. Utilizing a selection criterion, one must identify the most suitable parameter to utilize on the branch.

The calculation of entropy can be performed as follows:

$$Entropy(S) = \sum_{i=1}^{c} -p_i \log(p_i) \quad (3)$$

where:
$S$ is the dataset
$c$ is the number of classes
$p_i$ is the probability of class $i$ in the dataset.

Information Gain quantifies the decrease in entropy after the partitioning of the dataset. It is represented by:

$$IG(S,A) = Entropy(S) - \sum_{v \in V} \frac{|S_v|}{|S|} Entropy(S_v) \quad (4)$$

where:
$S$ is the original dataset,
$A$ is the attribute used for splitting,
$V$ is the set of possible values for the attribute $A$,
$S_v$ is the subset of $S$ corresponding to the attribute value $v$,
$|S_v|/|S|$ represents the proportion of $S_v$ in $S$.

### 2.3 Logistic Regression

Logistic regression is a special case of the Generalized Linear Model (GLM) specifically designed for binary classification problems, where the dependent variable takes only two possible values, typically 0 or 1. The model predicts the probability that a given observation belongs to one of these two classes. While multinomial logistic regression is used when the dependent variable consists of more than two unordered categories [21].

Suppose that categorical variables $Y$ with more than two possible levels, namely $\{\{1, 2, \dots, C\}$. Given the predictors $X_1, X_2, \dots, X_p$ , multinomial logistic regression models the probability of each level $c$ of $Y$ by [22],

$$p_c(\mathbf{x}) \coloneqq \mathbb{P}\left[Y = c | X_1 = x_1, \dots, X_p = x_p\right] = \frac{e^{\beta_{0c} + \beta_{1c}X_1 + \cdots + \beta_{pc}X_p}}{1 + \sum_{l=1}^{C-1} e^{\beta_{0l} + \beta_{1l}X_1 + \cdots + \beta_{pl}X_p}} \tag{5}$$

for $c = 1, 2, \dots, C-1$ and (for the last level $C$)

$$p_C(\mathbf{x}) \coloneqq \mathbb{P}\left[Y = C | X_1 = x_1, \dots, X_p = x_p\right] = \frac{1}{1 + \sum_{l=1}^{C-1} e^{\beta_{0l} + \beta_{1l}X_1 + \cdots + \beta_{pl}X_p}} \tag{6}$$

The multinomial logistic model has an interesting interpretation in terms of logistic regressions. Taking the quotient between **Equation (5)** and **Equation (6)** gives

$$\frac{p_c(\mathbf{x})}{p_C(\mathbf{x})} = e^{\beta_{0c} + \beta_{1c}X_1 + \cdots + \beta_{pc}X_p} \tag{7}$$

for $c = 1, 2, \dots, C-1$. Therefore, applying a logarithm to both sides, we have:

$$\log\left(\frac{p_c(\mathbf{x})}{p_C(\mathbf{x})}\right) = \beta_{0c} + \beta_{1c}X_1 + \cdots + \beta_{pc}X_p \tag{8}$$

## 3. RESULTS AND DISCUSSION

The dataset employed in this study comprises secondary daily data sourced from the Yogyakarta Meteorology, Climatology, and Geophysics Agency (BMKG). The sample utilized in this research encompasses climate condition data, including variables such as precipitation, temperature, wind, humidity, sunshine duration, wind speed, and wind direction, covering the period from 2021 to 2023. Subsequently, the comprehensive dataset is partitioned into sub-datasets, specifically 80% for training purposes and the remaining for testing, to optimize the model's performance that yields the most favorable outcomes.

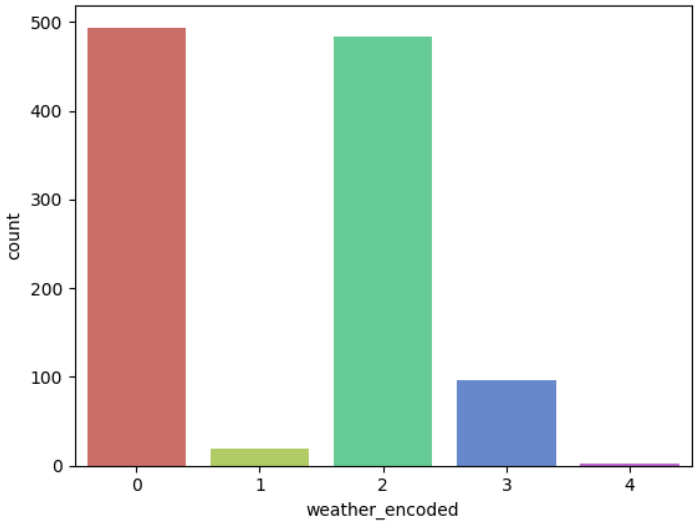



**Figure 1. The Variables Distribution**

From the analysis presented in **Figure 1**, it is evident that the dataset predominantly comprises instances of cloudy (code: 0) and light rain (code: 2) weather conditions, encompassing approximately

45.11% and 44.20% of the total dataset, respectively. In contrast, weather phenomena such as moderate rain (code: 3), heavy rain (code: 1), and very heavy rain (code: 4) constitute less than 10% of the overall dataset. The scarcity of data about moderate rain, heavy rain, and very heavy rain may significantly undermine the model's predictive accuracy when classifying these specific weather conditions due to insufficient training data. This imbalance poses a significant challenge, as it can potentially introduce bias into the learning process and diminish the model's efficacy, particularly in classifying minority classes. In the current investigation, no methodologies, such as Synthetic Minority Over-sampling Technique (SMOTE), random oversampling, or class weighting, were employed to mitigate the imbalance. Consequently, the model may demonstrate commendable accuracy for the majority classes, while significantly underperforming in minority classes, which is an inherent limitation that warrants careful consideration when interpreting the findings. Subsequent research endeavors are strongly suggested to integrate strategies for addressing imbalance to enhance model generalization and predictive accuracy across all climatic conditions, particularly for infrequent yet consequential phenomena such as heavy and very heavy precipitation.

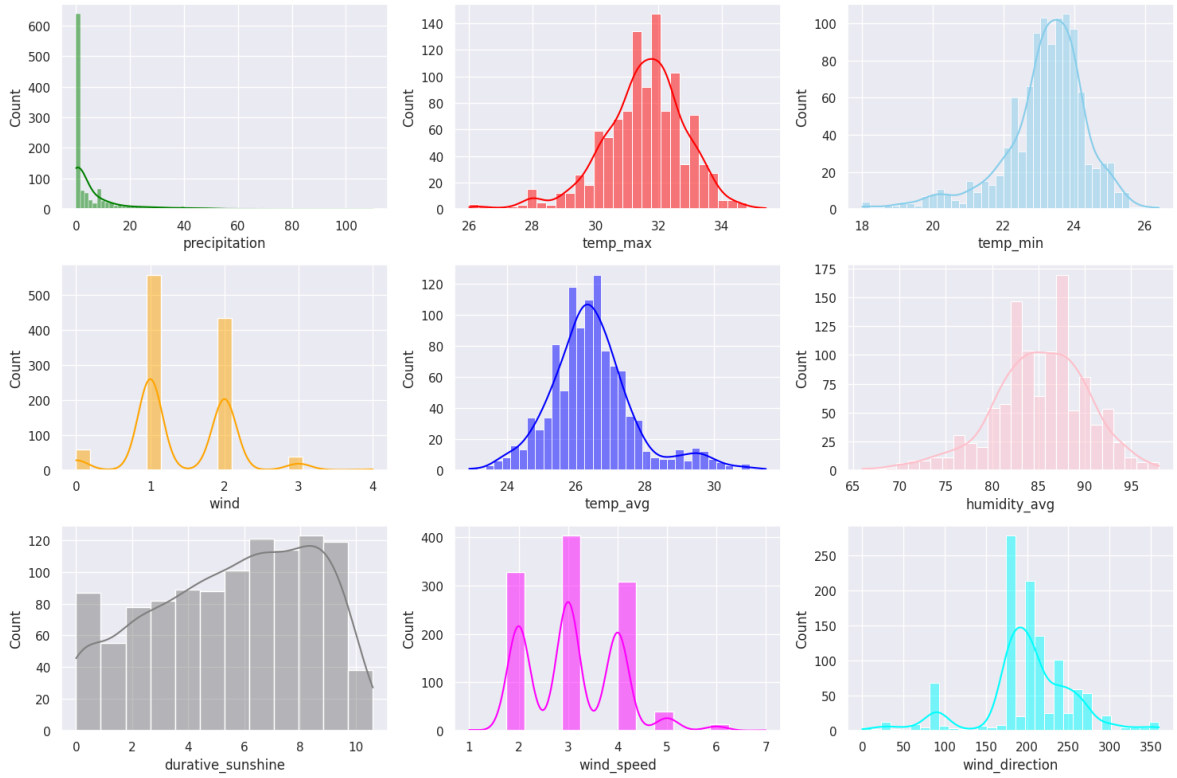



**Figure 2. The Variables Distribution**

**Figure 2** illustrates the comprises a matrix of histograms accompanied by Kernel Density Estimation (KDE) curves, which delineate the distributions of multiple meteorological variables. In the upper-left quadrant, the precipitation variable (green) demonstrates a markedly right-skewed distribution, signifying that most days experience minimal or no precipitation. At the same time, significant rainfall occurrences characterize a minority of days. The upper-middle quadrant illustrates the maximum temperature (red), which adheres to an approximately normal distribution centered around 31–32°C, indicative of a standard daily temperature range. The upper-right quadrant presents the minimum temperature (light blue), which similarly appears to be normally distributed, with values aggregating around 22–24°C but exhibiting a slight leftward skew.

Transitioning to the middle row, the left quadrant exemplifies the wind variable (orange), which reveals a bimodal distribution, thereby indicating the presence of two predominant wind conditions that manifest with notable frequency. The middle quadrant depicts the average temperature (dark blue), conforming to a near-normal distribution centered around 26–28°C, which suggests a prevalent temperature range. The right quadrant visualizes the average humidity (pink), characterized by a right-skewed distribution, with most values concentrated around 85–95%, implying a generally humid climate.

The bottom-left quadrant illustrates the duration of sunshine (gray), where the distribution appears relatively uniform with a slight right skew, indicating a diversity of sunshine durations. However, extended sunshine periods appear more prevalent. The bottom-middle quadrant presents wind speed (magenta) and

exhibits a bimodal distribution, suggesting that specific wind speed values occur with greater frequency. Lastly, the bottom-right quadrant denotes wind direction (cyan), where peaks around 200–250 degrees imply a prevailing wind direction within that range, while other directional occurrences are less frequent.

A square root transformation was employed on the dataset to mitigate skewness and address potential outliers. This transformation is frequently utilized to diminish right skewness and stabilize variance while preserving the relative relationships among the data points. Significant alterations can be discerned in the graphical representation upon executing the transformation, encompassing a more symmetric distribution and a probable attenuation of extreme values.

In the KNN algorithm, the parameter representing the number of neighbors ($k$) is instrumental in ascertaining the resultant classification or regression output. In this instance, $k$ is designated as 5, indicating that the algorithm evaluates the five closest neighbors for any specified data point to generate a prediction. This selection balances bias and variance, ensuring that the model does not exhibit excessive sensitivity to noise (as observed with minimal $k$ values) while simultaneously capturing localized patterns within the data. A specification of $k = 5$ facilitates the smoothing of predictions by averaging across multiple neighbors, thereby diminishing the probability of misclassification attributable to outliers or slight variations in the dataset.

The decision tree constitutes a supervised learning algorithm employed for classification endeavors, wherein data is partitioned into branches predicated on feature values to facilitate predictive modeling. In this instance, the classifier is parameterized with *max_depth = range (1, 8)*, indicating that the model will evaluate various tree depths from 1 to 7 to ascertain the optimal depth that reconciles complexity with performance. Furthermore, *max_leaf_nodes = 15* constrains the maximum quantity of terminal nodes, averting excessive branching and mitigating the potential for overfitting. The parameter *random_state = 0* guarantees reproducibility by regulating the stochasticity inherent in tree partitioning. The model aspires to attain a harmonious balance between accuracy and generalization by calibrating these hyperparameters, rendering it apt for classification undertakings involving structured data.

The logistic regression model, characterized by the designated parameters, constitutes a prevalent classification methodology that employs L2 regularization (ridge penalty) to mitigate the propensity for overfitting. The tolerance (*tol = 0.0001*) specifies the cessation criteria for the optimization procedure, thereby ensuring convergence when the loss function variation diminishes beneath this stipulated threshold. The regularization strength (*C = 1.0*) modulates the balance between attaining minimal error and preserving model complexity, where reduced values necessitate more rigorous regularization. The intercept scaling (*intercept_scaling = 1*) holds significance in scenarios where the model is fitted without prior feature scaling, consequently influencing the magnitude of the intercept coefficient. The *lbfgs*' solver (Limited-memory Broyden–Fletcher–Goldfarb–Shanno) represents an efficient optimization technique tailored for small to medium-sized datasets, adept at accommodating L2 regularization. Ultimately, the maximum number of iterations (*max_iter = 100*) guarantees that the optimization endeavor concludes after 100 iterations should convergence remain unachieved. These configurations yield a methodical framework for training a logistic regression model, integrating regularization with numerical stability.

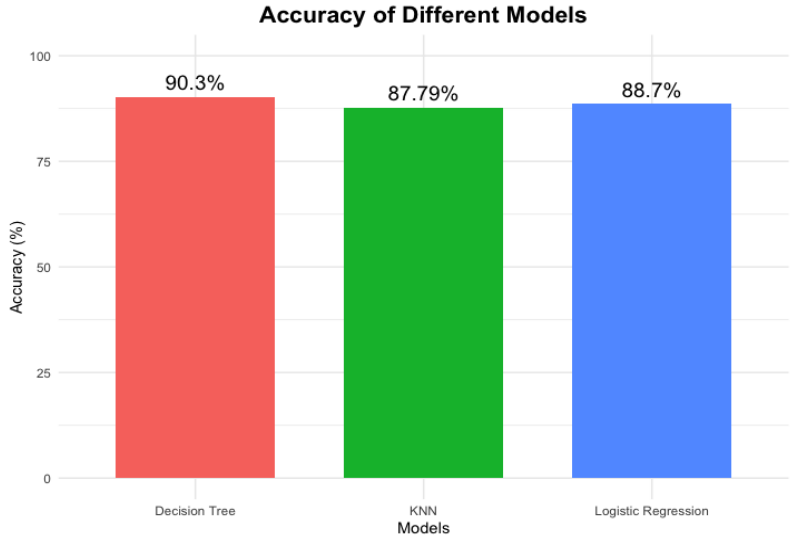
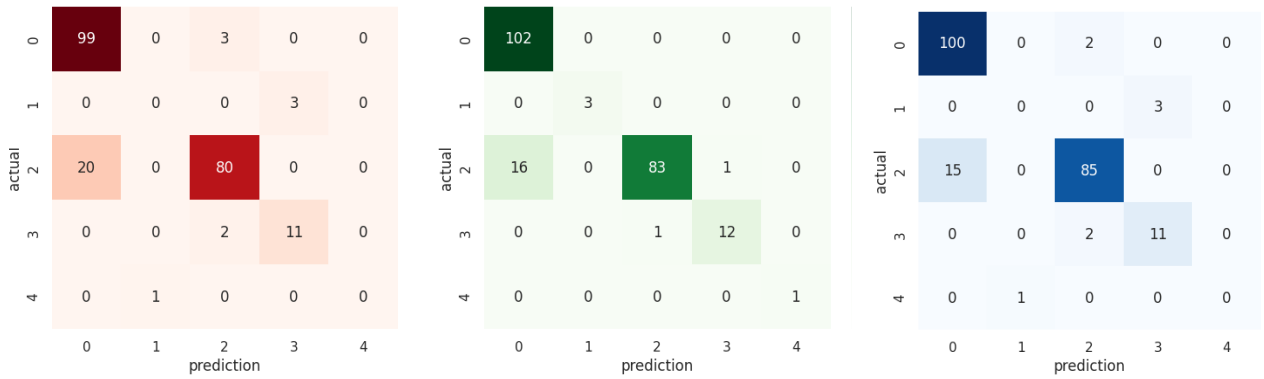



**Figure 3. The Comparison of the Accuracy of Training Data**

The bar chart in **Figure 3** shows the training accuracy associated with three machine learning algorithms: KNN, decision tree, and logistic regression. The decision tree model attains the apex of training accuracy at 90.30%, succeeded by logistic regression at 88.70%, and KNN at 87.79%.

The elevated accuracy of the decision tree model implies a proficient fitting of the training data; however, this may concurrently signify a potential risk of overfitting, particularly in instances where the model exhibits excessive complexity. The logistic regression model demonstrates marginally superior performance compared to KNN, likely attributable to its efficacy in identifying a linear decision boundary. Conversely, the slightly diminished accuracy of KNN may be influenced by hyperparameter considerations, including the number of neighbors ($k$) and the selection of the distance metric. While the decision tree model showcases the highest accuracy, validating these models on a test dataset is imperative to evaluate their generalization capabilities.




**(a)** **(b)** **(c)**

**Figure 4. The Confusion Matrix of Different Models**

**Figure 4** illustrates the confusion matrices corresponding to three distinct classification algorithms: **Figure 4** (a) KNN depicted in red, **Figure 4** (b) decision tree illustrated in green, and **Figure 4** (c) logistic regression represented in blue. The matrices provide a comparative analysis of actual versus predicted class labels within a five-class classification challenge. KNN exhibits difficulties with light rain (code: 2), erroneously categorizing 20 instances as belonging to cloudy (code: 0) while accurately predicting 80 instances. The decision tree algorithm demonstrates enhanced efficacy, accurately classifying 83 instances of light rain (code: 2) and reducing overall misclassifications. Logistic regression emerges as the most effective approach, achieving 85 correct predictions for light rain (code: 2) and displaying fewer inaccuracies across all class categories. Both decision tree and logistic regression surpass the performance of KNN.

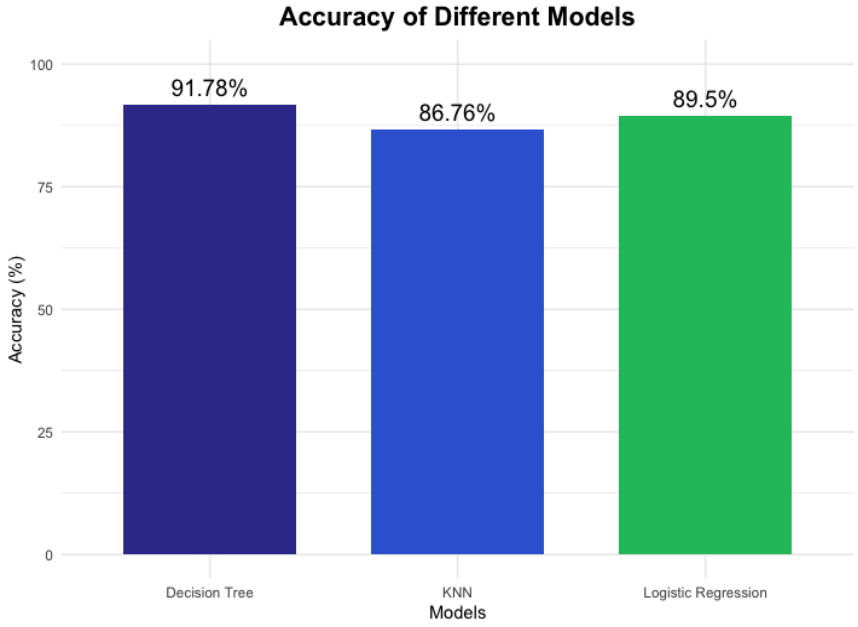



**Figure 5. The Comparison of the Accuracy of Testing Data**

Based on **Figure 5**, the decision tree model exhibits the highest classification accuracy, achieving a remarkable 91.78%, and is succeeded by logistic regression with an accuracy of 89.50% and KNN with an accuracy of 86.76%. This finding suggests that the decision tree model demonstrates superior performance in classification accuracy, resulting in a lower frequency of overall misclassifications. Logistic regression also demonstrates commendable performance, marginally exceeding KNN yet falling short of the decision tree's efficacy. KNN is characterized by the lowest accuracy among the three models. This accuracy is corroborated by the confusion matrix analysis, which indicates its pronounced challenges with misclassifications, especially concerning class light rain (code: 2).

Furthermore, **Table 1** presents the contribution of nine key variables in influencing the type of rainfall classification. These variables likely include meteorological factors such as precipitation, minimum temperature, maximum temperature, average temperature, wind, average humidity, sunshine duration, wind speed, and wind direction, which are crucial in determining rainfall patterns. The decision tree model's high accuracy suggests it effectively captures and leverages these influential factors, making it a reliable choice for rainfall classification. The importance of these variables in the model helps provide deeper insights into how specific climatic conditions contribute to rainfall, enabling better decision-making for rainfall classifications and climate-related analyses.

**Table 1. Features Importance**

| Feature | Importance |
|---|---|
| precipitation | 0.925798 |
| minimum temperature | 0.015525 |
| maximum temperature | 0.005102 |
| average temperature | 0.000000 |
| wind | 0.000000 |
| average humidity | 0.019682 |
| sunshine duration | 0.025015 |
| wind speed | 0.000000 |
| wind direction | 0.008877 |

**Table 1** presents the feature importance scores from the decision tree model for classifying type of rainfall in Yogyakarta. The most influential factor is precipitation, which is expected since rainfall directly determines rainy conditions. Other weather-related factors, such as humidity average and durative sunshine, play a minor role, as humidity influences cloud formation, while sunlight duration affects evaporation and atmospheric moisture. Temperature variables, including maximum and minimum temperatures, have minimal impact, suggesting that temperature fluctuations alone are not strong variables of rain. Interestingly, wind, wind speed, and average temperature have no contribution, indicating they do not significantly influence the type of rainfall classification in this dataset. Wind direction shows slight importance, possibly due to its effect on moving moisture-laden air masses. Overall, the model heavily relies on precipitation as the primary factor, with humidity and sunshine duration providing additional but limited support in classifying rainfall.

## 4. CONCLUSION

Among the three machine learning models, KNN, decision tree, and logistic regression, the decision tree model demonstrates superior accuracy in type of rainfall classification, attaining 90.30% on the training dataset and 91.78% on the test dataset. Among the input variables, precipitation emerges as the predominant influencing factor, which is anticipated given its direct correlation with rainy conditions. Other meteorological factors, such as average humidity and duration of sunlight, assume a secondary significance as humidity plays a pivotal role in cloud genesis. In contrast, sunlight duration impacts evaporation rates and atmospheric moisture content. Temperature variables, encompassing maximum and minimum temperatures, exert a negligible influence, suggesting that variations in temperature alone do not serve as robust indicators of precipitation.

Notwithstanding these encouraging outcomes, the investigation is not devoid of limitations. Firstly, the dataset is confined to a particular geographic locale (Yogyakarta) and may lack generalizability to other tropical or non-tropical regions characterized by divergent climatic conditions. Secondly, the analysis incorporated only a limited selection of meteorological variables; additional potentially pertinent factors, such as wind direction, atmospheric pressure, or cloud cover, were excluded due to data availability constraints. Thirdly, the research utilized solely conventional machine learning models. Implementing more sophisticated methodologies, such as ensemble learning or deep neural networks, may enhance classification accuracy and improve generalizability. Prospective inquiries could rectify these limitations by integrating various climatic variables, broadening the geographic focus, and employing more advanced modeling methodologies. Moreover, incorporating real-time data and external environmental determinants, such as topography or land utilization, could significantly augment model robustness and applicability within operational forecasting frameworks.

## AUTHOR CONTRIBUTIONS

Dina Tri Utari: Conceptualization, Formal Analysis, Investigation, Methodology, Supervision, Validation, Writing - Original Draft, Writing - Review and Editing. Ghalang Rambu Putera Palage: Data Curation, Formal Analysis, Project Administration. Faiz Fadhlirobby: Data Curation, Formal Analysis, Software. Artheta Bimo Nuswantoro: Data Curation, Formal Analysis, Visualization. All authors discussed the results and contributed to the final manuscript.

## FUNDING STATEMENT

This research has been funded by the Department of Statistics, Universitas Islam Indonesia, with contract number: 74/Kajur.Stat/10/Jur.Stat/XII/2024.

## ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to the Department of Statistics, Universitas Islam Indonesia, for the continuous support, guidance, and resources provided during this work.

## CONFLICT OF INTEREST

The authors declare that there is no conflict of interest related to this work.

## REFERENCES

[1] S. Scher, “TOWARD DATA-DRIVEN WEATHER AND CLIMATE FORECASTING: APPROXIMATING A SIMPLE GENERAL CIRCULATION MODEL WITH DEEP LEARNING,” *Geophys Res Lett*, vol. 45, no. 22, Nov. 2018, doi: https://doi.org/10.1029/2018GL080704.

[2] S. Adebayo, F. O. Aweda, I. A. Ojedokun, and J. A. Agbolade, “METEOROLOGICAL DATA PREDICTION OVER SELECTED STATIONS IN SUB-SAHARA AFRICA: LEVERAGING ON MACHINE LEARNING ALGORITHM,” *Ruhuna Journal of Science*, vol. 13, no. 2, pp. 129–140, Dec. 2022, doi: https://doi.org/10.4038/rjs.v13i2.120.

[3] K. Naren Athreyas, E. Gunawan, and B. K. Tay, “ESTIMATION OF VERTICAL STRUCTURE OF LATENT HEAT GENERATED IN THUNDERSTORMS USING CLOUDSAT RADAR,” *Meteorological Applications*, vol. 27, no. 2, Mar. 2020, doi: https://doi.org/10.1002/met.1902.

[4] W. Ghada *et al.*, “STRATIFORM AND CONVECTIVE RAIN CLASSIFICATION USING MACHINE LEARNING MODELS AND MICRO RAIN RADAR,” *Remote Sens (Basel)*, vol. 14, no. 18, p. 4563, Sep. 2022, doi: https://doi.org/10.3390/rs14184563.

[5] W. Ghada, N. Estrella, and A. Menzel, “MACHINE LEARNING APPROACH TO CLASSIFY RAIN TYPE BASED ON THIES DISDROMETERS AND CLOUD OBSERVATIONS,” *Atmosphere (Basel)*, vol. 10, no. 5, p. 251, May 2019, doi: https://doi.org/10.3390/atmos10050251.

[6] H. Jain, R. Dhupper, A. Shrivastava, D. Kumar, and M. Kumari, "LEVERAGING MACHINE LEARNING ALGORITHMS FOR IMPROVED DISASTER PREPAREDNESS AND RESPONSE THROUGH ACCURATE WEATHER PATTERN AND NATURAL DISASTER PREDICTION," *Front Environ Sci*, vol. 11, Nov. 2023, doi: https://doi.org/10.3389/fenvs.2023.1194918.

[7] W. Ghada, J. Bech, N. Estrella, A. Hamann, and A. Menzel, "WEATHER TYPES AFFECT RAIN MICROSTRUCTURE: IMPLICATIONS FOR ESTIMATING RAIN RATE," *Remote Sens (Basel)*, vol. 12, no. 21, p. 3572, Oct. 2020, doi: https://doi.org/10.3390/rs12213572.

[8] P. Asha, A. Jesudoss, S. P. Mary, K. V. S. Sandeep, and K. H. Vardhan, "AN EFFICIENT HYBRID MACHINE LEARNING CLASSIFIER FOR RAINFALL PREDICTION," *J Phys Conf Ser*, vol. 1770, no. 1, p. 012012, Mar. 2021, doi: https://doi.org/10.1088/1742-6596/1770/1/012012.

[9] R. S. Lubis, Y. Suzana, F. Syarah, F. Fajriana, F. Rozi, and B. C. Nusantara, "DYNAMICS OF RAINFALL AND TEMPERATURE IN NORTH SUMATRA PROVINCE: COMPREHENSIVE ANALYSIS OF TEMPORAL TRENDS," *BAREKENG: Jurnal Ilmu Matematika dan Terapan*, vol. 19, no. 1, pp. 215–226, Jan. 2025, doi: https://doi.org/10.30598/barekengvol19iss1pp215-226.

[10] Z. Rustam, R. P. Yuda, H. Alatas, and C. Aroef, "PULMONARY RONTGEN CLASSIFICATION TO DETECT PNEUMONIA DISEASE USING CONVOLUTIONAL NEURAL NETWORKS," *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 18, no. 3, p. 1522, Jun. 2020, doi: https://doi.org/10.12928/telkomnika.v18i3.14839.

[11] G. Adhane, M. M. Dehshibi, and D. Masip, "ON THE USE OF UNCERTAINTY IN CLASSIFYING *AEDES ALBOPICTUS* MOSQUITOES," *IEEE J Sel Top Signal Process*, vol. 16, no. 2, pp. 224–233, Feb. 2022, doi: https://doi.org/10.1109/JSTSP.2021.3122886.

[12] S. Kasus *et al.*, "KLASIFIKASI CURAH HUJAN HARIAN MENGGUNAKAN LEARNING VECTOR QUANTIZATION," *Jurnal Ilmu Komputer Indonesia (JIK)*, vol. 7, no. 2, 2022, [Online]. Available: http://www.ogimet.com,

[13] D. Naidu, B. Majhi, and S. K. Chandniha, "DEVELOPMENT OF RAINFALL PREDICTION MODELS USING MACHINE LEARNING APPROACHES FOR DIFFERENT AGRO-CLIMATIC ZONES," 2021, pp. 72–94. doi: https://doi.org/10.4018/978-1-7998-6659-6.ch005.

[14] Z. Yang, P. Liu, and Y. Yang, "CONVECTIVE/STRATIFORM PRECIPITATION CLASSIFICATION USING GROUND-BASED DOPPLER RADAR DATA BASED ON THE K-NEAREST NEIGHBOR ALGORITHM," *Remote Sens (Basel)*, vol. 11, no. 19, p. 2277, Sep. 2019, doi: https://doi.org/10.3390/rs11192277.

[15] T. Herlambang, V. Asy'ari, R. P. Rahayu, A. A. Firdaus, and N. Juniarta, "COMPARISON OF NAÏVE BAYES AND K-NEAREST NEIGHBOR MODELS FOR IDENTIFYING THE HIGHEST PREVALENCE OF STUNTING CASES IN EAST JAVA," *BAREKENG: Jurnal Ilmu Matematika dan Terapan*, vol. 18, no. 4, pp. 2153–2164, Oct. 2024, doi: https://doi.org/10.30598/barekengvol18iss4pp2153-2164.

[16] J. Gou, H. Ma, W. Ou, S. Zeng, Y. Rao, and H. Yang, "A GENERALIZED MEAN DISTANCE-BASED K-NEAREST NEIGHBOR CLASSIFIER," *Expert Syst Appl*, vol. 115, pp. 356–372, Jan. 2019, doi: https://doi.org/10.1016/j.eswa.2018.08.021.

[17] D. M. Ali and H. A.Sadeq, "ROAD POTHOLE DETECTION USING UNMANNED AERIAL VEHICLE IMAGERY AND DEEP LEARNING TECHNIQUE," *Zanco J Pure Appl Sci*, vol. 34, no. 6, Dec. 2022, doi: https://doi.org/10.21271/ZJPAS.34.6.12.

[18] G. He, W. Wang, B. Shi, S. Liu, H. Xiang, and X. Wang, "AN IMPROVED YOLO V4 ALGORITHM-BASED OBJECT DETECTION METHOD FOR MARITIME VESSELS," *International Journal of Science and Engineering Applications*, vol. 11, no. 04, pp. 50–55, Apr. 2022, doi: https://doi.org/10.7753/IJSEA1104.1001.

[19] M. Mahasin and I. A. Dewi, "COMPARISON OF CSPDARKNET53, CSPRESNEXT-50, AND EFFICIENTNET-B0 BACKBONES ON YOLO V4 AS OBJECT DETECTOR," *International Journal of Engineering, Science and Information Technology*, vol. 2, no. 3, pp. 64–72, Sep. 2022, doi: https://doi.org/10.52088/ijesty.v2i3.291.

[20] R. Bhardwaj and V. Duhoon, "WEATHER FORECASTING USING DECISION TREE," *Journal of Engineering Research*, Nov. 2021, doi: https://doi.org/10.1155/2021/6758557.

[21] A. Purwanto, M. A. Suprayogi, E. Setiawan, J. F. R. B. Loly, G. A. Rahman, and A. Kurnia, "MULTINOMIAL LOGISTIC REGRESSION MODEL USING MAXIMUM LIKELIHOOD APPROACH AND BAYES METHOD ON INDONESIA'S ECONOMIC GROWTH PRE TO POST COVID-19 PANDEMIC," *BAREKENG: Jurnal Ilmu Matematika dan Terapan*, vol. 19, no. 1, pp. 51–62, Jan. 2025, doi: https://doi.org/10.30598/barekengvol19iss1pp51-62.

[22] E. García-Portugués, *NOTES FOR PREDICTIVE MODELING*, Version 5.10.1. https://bookdown.org/egarpor/PM-UC3M/, 2025.