Classification of Daily Rainfall Using XGBoost on Imbalanced Data in The Special Region of Yogyakarta

Ayyesa Azzahra Mulia Ramadani[1)]*, Danur Wijayanto[2)]

[1.2] Information Technology Study Program, Universitas `Aisyiyah Yogyakarta
* Coressponding author : 2211501032@student.unisayogya.ac.id

Abstract

Rainfall is a meteorological parameter that influences various sectors, such as agriculture, water resource management, and disaster mitigation; however, the process of classifying it still faces challenges, particularly due to imbalanced data across categories. This study aims to evaluate the performance of the XGBoost algorithm in classifying daily rainfall in the Special Region of Yogyakarta using NASA POWER data from 2000 to 2025, with input variables including air temperature, relative humidity, wind speed, and surface pressure. The evaluation was conducted using accuracy, precision, recall, and F1-score metrics to provide a more comprehensive overview of the model's performance. The results indicate that the model achieved an accuracy of 0.82 and performed well in identifying light rain, and began to identify moderate rain, although not yet optimally; however, its performance remains limited for higher-intensity rain classes. This suggests that imbalanced data distribution remains a primary factor affecting model performance, making data quality and balance critical considerations in the development of rainfall classification models.

Keywords- Rainfall, XGBoost, SMOTE, Classification

1. INTRODUCTION

The amount of water that falls on the ground is the definition of rainfall, which occurs over a specific period of time and is measured in millimeters (mm) [1]. As one of the primary meteorological elements, rainfall plays a crucial role in various aspects of life, such as agriculture, water resource management, disaster mitigation, and development planning [2]. The Special Region of Yogyakarta consists of mountainous and lowland areas, resulting in high rainfall intensity in this region and potentially causing impacts such as flooding and landslides [3]. Data from the Yogyakarta City Statistics Agency (BPS) indicates that the annual average temperature ranges from 26.3°C to 27.8°C. In 2023, the recorded rainfall total was 1,955.0 mm, increasing to 3,058.4 mm in 2024. The highest number of rainy days occurred in 2022, totaling 214 days. These changes in temperature and rainfall indicate the need for a data-driven approach to more accurately assess hydrometeorological conditions [4]. These changes are influenced not only by local conditions, such as landforms and vegetation, but also by global climate change and extreme weather events [5]. Thus, rainfall analysis is a crucial step in understanding rainfall patterns based on historical data, thereby enhancing the ability to make informed, data-driven decisions [6].

These increasingly unpredictable weather patterns highlight the need for analytical methods capable of processing large datasets more effectively. It is at this stage that the Data Mining approach becomes relevant. Data Mining is a method that leverages mathematical, statistical, and Machine Learning concepts to process various databases and transform them into useful information [7][8]. Advances in Machine Learning technology provide more adaptive solutions for data processing [9]. One of the algorithms used in data-driven modeling is Extreme Gradient Boosting (XGBoost) [10]. This algorithm has the capability to handle complex weather data, yielding good results through ensemble learning techniques [11]. XGBoost builds decision trees incrementally, where each new tree corrects the errors of the previous one through the gradient boosting process [12]. Additionally, XGBoost can automatically handle missing data, prevent overfitting through regularization, and process various data types, including both numeric and categorical data [13]. Rainfall data from NASA's Prediction of Worldwide Energy Resources (POWER) was used in this study, covering the period from 2000 to 2025. This dataset was selected because it is comprehensive, open-access, covers various weather parameters, and has minimal missing data, thereby supporting the development of a Machine Learning model to classify rainfall conditions into four categories based on historical data. The use of data over a long time span aims to capture a wider variety of rainfall patterns, allowing the model to better learn the data's characteristics.

DOI : https://doi.org/10.21063/jtif.2026.V14.1.59-67

Various Machine Learning approaches have been applied for weather prediction or classification. The K-Nearest Neighbor (KNN) algorithm yields good results with 75% accuracy, though the F1-score varies across classes [14]. Other approaches, such as the Support Vector Machine (SVM), are also used in weather forecasting, achieving an accuracy of 54.55% in a two-class classification (rain and cloudy) [15]. In decision tree-based approaches, the C4.5 Decision Tree demonstrates very high classification accuracy of up to 99.12%, particularly with a 90:10 data split using 2,048 training data points [16]. However, a single Decision Tree model has limitations compared to ensemble methods like Random Forest, which are more stable and possess better generalization capabilities [17]. Further research compared the Random Forest and Naïve Bayes algorithms in classifying daily rainfall data in Indonesia. The results showed that Random Forest achieved higher accuracy (86.55%) compared to Naïve Bayes (36.61%), and produced better precision, recall, F1-score, and AUC values [18]. As machine learning methods have evolved, the XGBoost algorithm has also been widely used in various classification studies and demonstrated strong performance. The boosting approach used allows the model to be built incrementally to improve prediction results [19]. However, to date, no study has specifically applied XGBoost for multi-class daily rainfall classification using long-term data coverage in the Special Region of Yogyakarta. This underscores the need for further research in this context.

Based on the need for data-driven analysis and advancements in Machine Learning methods, this study focuses on applying the Extreme Gradient Boosting (XGBoost) algorithm to classify daily rainfall into three categories: light rain (0–20 mm), moderate rain (21–50 mm), and heavy rain (51–100 mm). This classification approach is crucial for supporting the precise and accurate identification of weather conditions. This study aims to evaluate the performance of the XGBoost algorithm in classifying daily rainfall in the Special Region of Yogyakarta. The results of this study are expected to serve as a reference for data-driven decision-making in sectors sensitive to rainfall conditions, such as the agricultural sector in determining planting times, disaster mitigation to anticipate floods and landslides, as well as regional planning and environmental management. To achieve this objective, this study developed a daily rainfall classification model using the XGBoost algorithm, which was then evaluated using accuracy, precision, recall, and F1-score metrics.

2. METHODS AND DATA

This study employs a quantitative approach with classification for a daily rainfall prediction model using the XGBoost algorithm. The research framework consists of five main interrelated stages: (1) literature review, (2) data collection, (3) data preprocessing, (4) machine learning modeling, and (5) model evaluation. The complete research process is shown in Figure 1.

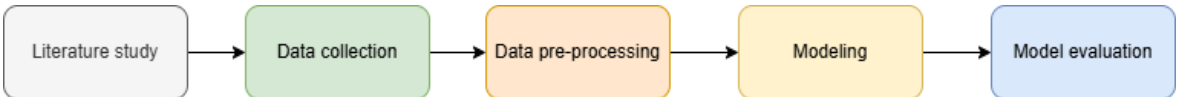

Figure 1. Research phase

2.1. Literature Study

This study began with an in-depth literature review to understand the theory [20], the development of the XGBoost algorithm, weather classification methods, and previous research trends related to rainfall prediction. This review was used to identify key concepts, findings from previous research, and existing gaps in rainfall classification research based on local data from the Special Region of Yogyakarta.

2.2. Data collection

During the data collection phase, this study utilized meteorological data sourced from NASA's Prediction of Worldwide Energy Resources (POWER), which is managed by the NASA Langley Research Center and is openly accessible via https://POWER.larc.nasa.gov/. The data used consists of

DOI : https://doi.org/10.21063/jtif.2026.V14.1.59-67

daily data covering the period 2000–2025, including atmospheric parameters such as Year (YEAR), Day of the Year (DOY), 2-meter Air Temperature (T2M, °C), 2-meter Relative Humidity (RH2M, %), 10 m Wind Speed (WS10M, m/s), Surface Pressure (PS), and Precipitation (PRECTOTCORR, mm). These parameters were selected because they are the primary factors influencing rainfall prediction [21]. The precipitation parameter serves as the main variable in rainfall condition analysis, while the other parameters are used to represent supporting atmospheric conditions [22]. The data was downloaded in CSV (Comma Separated Values) format for initial verification and preparation for the subsequent data processing stage.

2.3. Data Preprocessing

The data preprocessing stage was conducted to ensure the dataset is clean, consistent, and ready for use in the modeling process [23]. Raw data obtained from the NASA POWER platform was adjusted to be more representative and free from values that could potentially interfere with the analysis [24]. The first step involved renaming parameters, where technical terms such as T2M, RH2M, WS10M, PS, and PRECTOTCORR were changed to more informative names—namely, 2-meter Air Temperature, 2-meter Relative Humidity, 10-meter Wind Speed, Surface Pressure, and Precipitation—without altering the original data values.

The next stage involved data cleaning through the examination of missing values and the detection of anomalies. Missing values in the numerical data were handled using the mean imputation method to maintain the stability of the data distribution [25], while extreme values or outliers were addressed using the Interquartile Range (IQR) method to ensure the data remained within a reasonable range without altering its primary characteristics, thereby allowing for a more comprehensive understanding of the data's patterns and characteristics [11].

Class labeling was performed by grouping daily rainfall values (PRECTOTCORR) into four categories: 0–20 mm (light rain), 21–50 mm (moderate rain), and 51–100 mm (heavy rain). The dataset was split into training and test data using an 80:20 time-based split, where the training data came from the initial period and the test data from the subsequent period. The goal was to ensure the model had sufficient data during the training process while allowing for objective testing using the available data [26]. Next, the Synthetic Minority Over-sampling Technique (SMOTE) was applied to the training data to increase the number of samples in the minority classes. Meanwhile, the test data retained its original distribution to ensure the model evaluation remained objective and to prevent data leakage

2.4. Modeling

In the modeling stage, the XGBoost (Extreme Gradient Boosting) algorithm is used to classify rainfall. In XGBoost, decision trees are constructed sequentially to correct errors arising from the previous tree; ultimately, the resulting model achieves high precision and is capable of understanding the complex patterns present in the data [27]. The XGBoost model training process is performed by optimizing the following objective function:

$$Obj(\theta) = L(\theta) + \Omega(\theta) \quad (1)$$

Note **:**

- $L(\theta)$) is the Loss Function, which measures how far the model's predictions deviate from the actual labels.
- $\Omega(\theta)$ is the regularization term, which aims to control model complexity and prevent overfitting [28].

This objective function ensures that the model not only minimizes prediction error through the loss function but also remains stable and capable of generalizing by adding regularization to each decision

DOI : https://doi.org/10.21063/jtif.2026.V14.1.59-67

tree. This formula serves as the basis for the model training process and the evaluation of XGBoost performance in this study [29].

2.5. Model Evaluation

The classification of the model's output is evaluated using several metrics: accuracy, precision, recall, and F1-score [30]. This evaluation is based on the confusion matrix, which consists of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

$$F1\text{-}score = \frac{2\times (Precision\times Recall)}{Precision+Recall} \quad (5)$$

The accuracy calculation in Equation (2) is performed comprehensively to measure the extent to which the model can correctly predict all test data. Equation (3) serves to assess the model's precision in generating accurate positive predictions, that is, measuring positive predictions that truly correspond to the actual state. Recall in Equation (4) describes the model's ability to identify all positive cases present in the data. Meanwhile, the F1-score in Equation (5) is a harmonic mean that combines precision and recall, providing a more balanced assessment, particularly for data with class imbalance[31]. These four metrics are used to evaluate the model's quality in recognizing rainy conditions in the test data, with a particular emphasis on recall and the F1-score because these two metrics better reflect the model's performance in situations involving unbalanced daily rainfall data.

3. RESULTS AND DISCUSSION

This study utilized 9,481 daily rainfall records obtained from NASA POWER for the Special Region of Yogyakarta (DIY) during the period 2000–2025. The data includes several meteorological parameters, namely 2-meter air temperature (T2M), 2-meter relative humidity (RH2M), 10-meter wind speed (WS10M), surface pressure (PS), and precipitation (PRECTOTCORR). Meteorological parameters are used as features in the classification process, while precipitation is used as the basis for class label formation. Table 1 displays an example of the data used in the study. Subsequently, the data is processed through preprocessing and labeling stages before being used in Machine Learning modeling.

Table 1. Sample dataset

YEAR	DOY	T2M	RH2M	WS10M	PS	PRECTOTCORR
2000	1	26.9	88.13	3.97	100.15	1.33
2000	2	26.66	87.33	4.65	100.1	4.82
2000	3	25.88	89.38	2.88	100.14	19.39
2000	4	26.05	85.95	1.67	100.24	6.39
2000	5	25.9	89.55	4.01	100.32	2.93

DOI : https://doi.org/10.21063/jtif.2026.V14.1.59-67

Based on the correlation results, rainfall appears to have a moderate positive correlation with 2-meter relative humidity, as well as a weak negative correlation with 2-meter air temperature. Meanwhile, the correlations with surface pressure and wind speed are very weak.

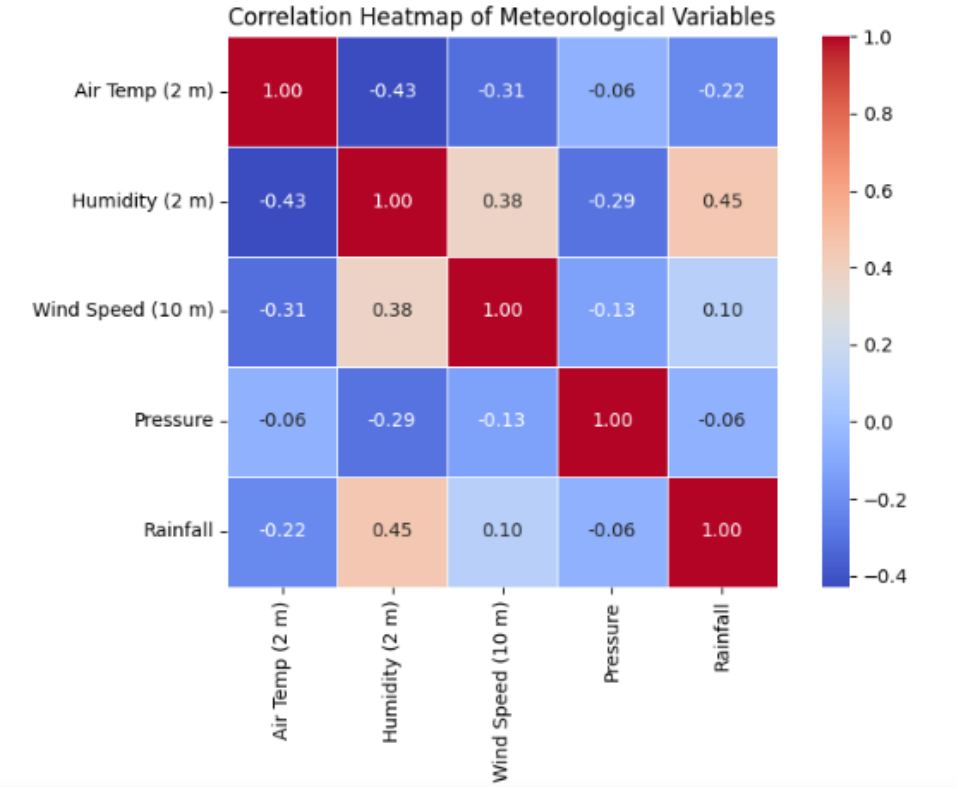

Figure 2. Correlation heatmap

Table 2 shows the distribution of data counts across each rainfall category. The data is dominated by the light rain category with 8,666 days, moderate rain with 430 days, and heavy rain with 39 days. This indicates class imbalance in the dataset used.

Table 2. Number of data categories

Label	Rainfall	Category	Number of Days
0	0–20 mm	Light rain	8666
1	21–50 mm	Moderate rain	430
2	51–100 mm	Heavy rain	39

To address data imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) was applied to the training data. This method works by generating new synthetic data for the minority class, thereby making the data distribution across classes more balanced. After data balancing using SMOTE, the model was then trained using the XGBoost algorithm and evaluated using the test data. The model used was configured with several parameters, namely n_estimators set to 200, max_depth set to 4, learning_rate set to 0.1, subsample set to 0.8, and colsample_bytree set to 0.8. Additionally, to reduce bias toward the majority class, class weights were applied during the model training process. The model evaluation results are shown in Table 3.

DOI : https://doi.org/10.21063/jtif.2026.V14.1.59-67

Table 3. Evaluation table

Class	Precision	Recall	F1-score
0	0.95	0.86	0.91
1	0.16	0.33	0.22
2	0.04	0.17	0.07
Macro avg	0.39	0.45	0.40
Accuracy		0.82	

The evaluation results show that the model is able to accurately identify the light rain class, but still struggles to identify the moderate and heavy rain classes. This is influenced by an unbalanced data distribution, in which the light rain class is far more dominant. The use of class weights helps improve the model's ability to identify the moderate rain class, although its performance for the heavy rain class remains unstable. This is influenced by the unbalanced data distribution and the very limited amount of data in certain classes. Although data balancing was performed using SMOTE, the generated data still originates from existing patterns, so its variation remains limited, resulting in the model not yet being able to capture the differences between classes effectively.

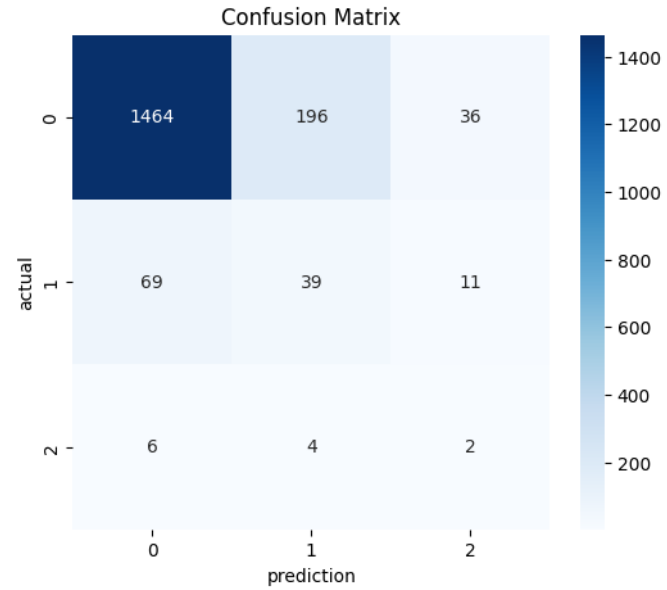

Figure 3. Confusion matrix

Based on the confusion matrix, classification errors frequently occur in the moderate and heavy rain classes, which are often predicted as light rain. This indicates that the model still struggles to distinguish between classes, primarily because the amount of data in those classes is very limited. Although SMOTE was used, the generated data still comes from existing data, so the resulting variations remain limited. Consequently, the model is not yet able to recognize the patterns of each class effectively, as shown in Figure 3.

DOI : https://doi.org/10.21063/jtif.2026.V14.1.59-67

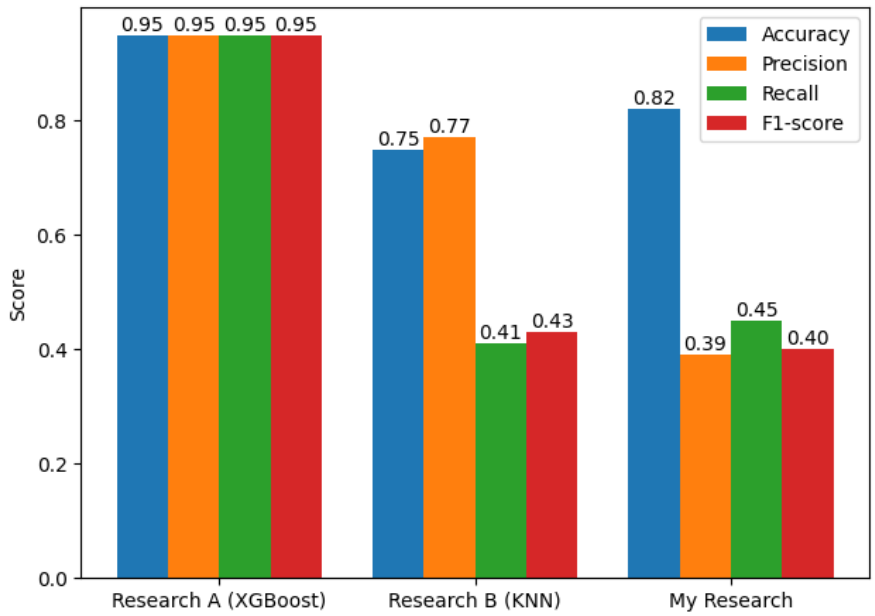

Figure 4. Comparative study

Compared to previous studies, the results obtained in this study are still lower than those of studies that also used XGBoost, which achieved accuracies of up to 95%. This difference is due to the fact that previous studies used simpler binary classification, whereas this study employed the more complex multiclass classification. Additionally, the type of data also plays a role, as the daily data used exhibits higher variability compared to hourly data, making the data patterns more difficult for the model to recognize. When compared to the K-Nearest Neighbor (KNN) method, the results differ only slightly in terms of accuracy, with KNN achieving around 75% while this study's results are slightly higher. However, when viewed in terms of precision, recall, and F1-score, KNN indicates that the model's ability to recognize each class is not yet consistent. In this study, XGBoost shows a performance improvement, albeit not very significant, particularly in terms of stability in recognizing several classes. This indicates that a boosting-based approach can help the model learn more complex data patterns. Overall, the differences in the results obtained are still within a reasonable range and are more influenced by the method, the number of classes, and the unbalanced data conditions. As shown in Figure 4.

4. CONCLUSION

This study shows that the XGBoost algorithm is quite capable of classifying daily rainfall with an accuracy of 0.82, particularly for the dominant light rain class. However, in a multi-class setting with an imbalanced data distribution, the model's performance is not yet consistent. The model begins to recognize moderate rain but still struggles with heavy rain due to the very limited amount of data. Although SMOTE was used, the resulting data variation remains limited, causing the model to tend to follow the pattern of the majority class. The limitations of this study lie in the data imbalance and the very small sample size in certain classes. Practically, the model can be used as an initial guide to identify light rain and begin distinguishing moderate rain. Further research is recommended to test other balancing methods, such as ADASYN or SMOTE-ENN, and to use a more balanced dataset so that the model's performance can be improved.

ACKNOWLEDGMENTS

The author would like to express gratitude to the Information Technology Program at Aisyiyah University of Yogyakarta for the academic support and facilities provided during the conduct of this research. The author also extends his deepest appreciation to his academic advisor for the guidance, supervision, and feedback provided throughout the research process and the preparation of this article.

DOI : https://doi.org/10.21063/jtif.2026.V14.1.59-67

Furthermore, the author would like to thank his parents for their prayers, support, and encouragement, which enabled this research to be successfully completed.

REFERENCES

[1] D. A. H. Panggabean, F. M. Sihombing, and N. M. Aruan, "Prediksi Tinggi Curah Hujan dan Kecepatan Angin Berdasarkan Data Cuaca dengan Penerapan Algoritma Artificial Neural Network (ANN)," *SEMINASTIKA*, vol. 3, no. 1, pp. 1–7, Nov. 2021, doi: 10.47002/seminastika.v3i1.237.

[2] H. Sitepu, D. Harisuseno, and J. S. Fidari, "Evaluasi Data Curah Hujan Satelit ERA-5 pada Berbagai Periode Data Hujan di Sub DAS Bodor Evaluation of ERA5 Satellite Rainfall Data at Various Rainfall Data Periods in Bodor Sub Watershed," *Jurnal Teknologi dan Rekayasa Sumber Daya Air*, vol. 03, no. 02, pp. 626–636, 2023, doi: 10.21776/ub.jtresda.003.vol.no02.053.

[3] M. Sulistiyono, B. Satria, A. Sidauruk, and R. Wardhana, "Rainfall Prediction Using Multiple Linear Regression Algorithm," *JITK (Jurnal Ilmu Pengetahuan dan Teknologi Komputer)*, vol. 9, no. 1, pp. 17–22, Aug. 2023, doi: 10.33480/jitk.v9i1.4203.

[4] Badan Pusat Statistik DIY., *Statistik Lingkungan Hidup Daerah Istimewa Yogyakarta*. Badan Pusat Statistik Provinsi Daerah Istimewa Yogyakarta, 2025.

[5] I Gusti Ngurah Putu Dharmayasa, Cathleen Ariella Simatupang, and Doni Marisi Sinaga, "NASA Power's: an Alternative Rainfall Data Resources for Hydrology Research and Planning Activities in Bali Island, Indonesia," *Journal of Infrastructure Planning and Engineering (JIPE)*, vol. 1, no. 1, pp. 1–7, Apr. 2022, doi: 10.22225/jipe.1.1.2022.1-7.

[6] D. Sangaji and T. Sutabri, "Analisis XGBoost dan Random Forest untuk Prediksi Curah Hujan dalam Mendukung Mitigasi Karhutla," *Jurnal Pustaka AI (Pusat Akses Kajian Teknologi Artificial Intelligence)*, vol. 5, no. 1, pp. 13–18, Apr. 2025, doi: 10.55382/jurnalpustakaai.v5i1.905.

[7] A. S. Agung, A. A. Fauzi, A. A. Nur Risal, and F. Adiba, "Implementasi Teknik Data Mining terhadap Klasifikasi Data Prediksi Curah Hujan BMKG Di Sulawesi Selatan," *Jurnal Tekno Insentif*, vol. 17, no. 1, pp. 22–23, Apr. 2023, doi: 10.36787/jti.v17i1.955.

[8] T. Hardiani and R. N. Putri, "Implementasi Metode Naïve Bayes Classifier Untuk Klasifikasi Stunting Pada Balita," *Digital Transformation Technology*, vol. 4, no. 1, pp. 621–627, Aug. 2024, doi: 10.47709/digitech.v4i1.4481.

[9] W. Puji, "Penggunaan Aplikasi Machine Learning (Ml) dalam Kurikulum Perubahan Iklim," *Journal of Education Research*, vol. 5, no. 4, 2024.

[10] G. Almuzadid and R. Subhiyakto, "Stroke Risk Classification Using the Ensemble Learning Method of XGBoost and Random Forest," *Journal of Applied Informatics and Computing (JAIC)*, vol. 9, no. 3, p. 828, 2025, [Online]. Available: http://jurnal.polibatam.ac.id/index.php/JAIC

[11] A. Syahreza, N. K. Ningrum, and M. A. Syahrazy, "Perbandingan Kinerja Model Prediksi Cuaca: Random Forest, Support Vector Regression, dan XGBoost," *Edumatic: Jurnal Pendidikan Informatika*, vol. 8, no. 2, pp. 526–534, Dec. 2024, doi: 10.29408/edumatic.v8i2.27640.

[12] C. Valentino *et al.*, "Analisis Kinerja XGBoost Menggunakan Bayesian Optimization dalam Prediksi Harga Ethereum," *JNATIA: Jurnal Nasional Teknologi Informasi dan Aplikasinya*, vol. 3, no. 4, Aug. 2025.

[13] A. Khairunnisa, "Perbandingan Model Random Forest dan XGBoost Untuk Prediksi Kejahatan Kesusilaan di Provinsi Jawa Barat," *JIKO (Jurnal Informatika dan Komputer)*, vol. 7, no. 2, p. 202, Sep. 2023, doi: 10.26798/jiko.v7i2.799.

[14] A. D. P. Putri, M. Al Haris, F. Fauzi, and S. Amri, "K-Nearest Neighbor (KNN) Method for Weather Data Prediction," *Journal Of Data Insights*, vol. 3, no. 1, pp. 56–64, Jun. 2025, doi: 10.26714/jodi.v3i1.214.

[15] I. Fau, "Penerapan Data Mining Dengan Metode Support Vector Machine Untuk Prediksi Cuaca," *Bulletin of Data Science*, vol. 4, no. 1, Oct. 2024, [Online]. Available: https://ejurnal.seminar-id.com/index.php/bulletinds

[16] N. Akhir dari Penulis Pertama *et al.*, "Penerapan Algoritma Decision Tree C4.5 untuk Prediksi Cuaca di Kota Semarang," *INDEXIA: Informatic and Computational Intelligent Journal*, vol. 07, no. 01, pp. 45–52, 2025.

[17] P. Ayu Firnanda *et al.*, "Analisis Perbandingan Decision Tree dan Random Forest dalam Klasifikasi Penjualan Produk pada Supermarket," *Emerging Statistics and Data Science Journal*, vol. 3, no. 1, 2025.

DOI : https://doi.org/10.21063/jtif.2026.V14.1.59-67

[18] N. A. Prakoso Indaryono, "Analisa Perbandingan Algoritma Random Forest dan Naive Bayes untuk Klasifikasi Curah Hujan Berdasarkan Iklim di Indonesia," *JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, vol. 9, no. 1, pp. 158–167, Feb. 2024, doi: 10.29100/jipi.v9i1.4421.
[19] J. Zhen *et al.*, "Performance of XGBoost Ensemble Learning Algorithm for Mangrove Species Classification with Multisource Spaceborne Remote Sensing Data," *Journal of Remote Sensing (United States)*, vol. 4, Jan. 2024, doi: 10.34133/remotesensing.0146.
[20] E. P. Cendana1*, "Visualization of COVID-19 Data in Yogyakarta City Using Data Studio," 2022.
[21] A. Luthfiarta, A. Febriyanto, H. Lestiawan, and W. Wicaksono, "Analisa Prakiraan Cuaca dengan Parameter Suhu, Kelembaban, Tekanan Udara, dan Kecepatan Angin Menggunakan Regresi Linear Berganda," *JOINS (Journal of Information System)*, vol. 5, no. 1, pp. 10–17, May 2020, doi: 10.33633/joins.v5i1.2760.
[22] I Dewa Gede Loka Maheswara and Ahmad Hanif Al'aziz, "Perbandingan Model Machine Learning pada Klasifikasi Curah Hujan di Bogor," *INTI Nusa Mandiri*, vol. 19, no. 2, pp. 202–210, Feb. 2025, doi: 10.33480/inti.v19i2.6296.
[23] S. Sandiwarno, "Penerapan Machine Learning untuk Prediksi Bencana Banjir," *Jurnal Sistem Informasi Bisnis*, vol. 14, no. 1, pp. 62–76, Jan. 2024, doi: 10.21456/vol14iss1pp62-76.
[24] A. Aprilia, A. B. Wahidin, and A. F. Abdurrahman, "Integration of Machine Learning and NASA POWER Dataset for Predicting Coffee Production in Lampung," *Jurnal Fisika Flux: Jurnal Ilmiah Fisika FMIPA Universitas Lambung Mangkurat*, vol. 22, no. 1, p. 44, Mar. 2025, doi: 10.20527/flux.v22i1.20980.
[25] F. Yulian Pamuji, A. Rofiqul Muslikh, R. Muhammad Arief, and D. Muti, "Komparasi Metode Mean dan KNN Imputation Dalam Mengatasi Missing Value pada Dataset Kecil," *JIP (Jurnal Informatika Polinema)*, vol. 10, no. 2, Feb. 2024, [Online]. Available: https://archive.ics.uci.edu/datasets.
[26] P. A. Saputra, R. Rahmaddeni, S. S. Irawan, R. Prianto, and D. Delfi, "Analisis Faktor Dominan Minat Beli Generasi Z di Shopee Menggunakan Algoritma Naïve Bayes," *sudo Jurnal Teknik Informatika*, vol. 4, no. 3, pp. 247–256, Nov. 2025, doi: 10.56211/sudo.v4i3.1137.
[27] C. Emilia Sukmawati *et al.*, "Efektivitas Algoritma AdaBoost dan XGBoost pada Dataset Obesitas Populasi Dewasa," *Jambura Journal of Informatics*, vol. 6, no. 2, pp. 101–111, 2024, doi: 10.37905/jji.
[28] R. Winurputra and D. E. Ratnawati, "Peramalan Penjualan Produk Menggunakan Extreme Gradient Boosting (XGBoost) dan Kerangka Kerja CRISP-DM untuk Pengoptimalan Manajemen Persediaan (Studi Kasus: UB Mart)," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 12, no. 2, pp. 417–428, Apr. 2025, doi: 10.25126/jtiik.2025129451.
[29] I. Muslim Karo Karo, "Implementasi Metode XGBoost dan Feature Importance untuk Klasifikasi pada Kebakaran Hutan dan Lahan," *Journal of Software Engineering, Information and Communication Technology*, vol. 1, no. 1, pp. 11–18, 2020, doi: Vol.1No.1,November2020pp.11-18.
[30] S. N. S. Muslim, F. Nurdiyansyah, and A. Y. Rahman, "Perbandingan Algoritma Naive Bayes dan KNN Dalam Analisis Sentimen Ulasan Pengguna Aplikasi Capcut," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 12, no. 3S1, Oct. 2024, doi: 10.23960/jitet.v12i3s1.5156.
[31] D. D. N. Cahyo and A. Sunyoto, "Analisis Perbandingan Klasifikasi dalam Data Mining pada Prediksi Hujan dengan menggunakan Algoritma LSTM dan GRU," *Jurnal Sains dan Informatika*, vol. 11, no. 1, pp. 40–49, Jun. 2025, doi: 10.34128/jsi.v11i1.1212.

DOI : https://doi.org/10.21063/jtif.2026.V14.1.59-67