RESEARCH ARTICLE • OPEN ACCESS

Comparison of Random Forest, XGBoost and LightGBM Methods on the Human Development Index Classification

Yunna Mentari Indah, Rafika Aristawidya, Anwar Fitrianto, Erfiani, and L.M. Risman Dwi Jumansyah

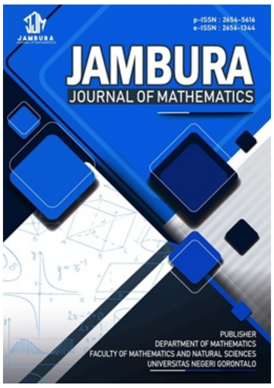
JAMBURA
JOURNAL OF MATHEMATICS

Volume 7, Issue 1, Pages 14–18, February 2025

Received 11 October 2024, Revised 12 December 2024, Accepted 25 December 2024, Published 1 February 2025

To Cite this Article : Y. M. Indah, et al.,"Comparison of Random Forest, XGBoost and LightGBM Methods on the Human Development Index Classification", *Jambura J. Math*, vol. 7, no. 1, pp. 14–18, 2025, *https://doi.org/10.37905/jjom.v7i1.28290*

JOURNAL INFO • JAMBURA JOURNAL OF MATHEMATICS


p-ISSN : 2654-5616
e-ISSN : 2656-1344
JAMBURA
JOURNAL OF MATHEMATICS
PUBLISHER
DEPARTMENT OF MATHEMATICS
FACULTY OF MATHEMATICS AND NATURAL SCIENCES
UNIVERSITAS NEGERI GORONTALO

Homepage	:	http://ejurnal.ung.ac.id/index.php/jjom/index
Journal Abbreviation	:	Jambura J. Math.
Frequency	:	Biannual (February and August)
Publication Language	:	English (preferable), Indonesia
DOI	:	https://doi.org/10.37905/jjom
Online ISSN	:	2656-1344
Editor-in-Chief	:	Hasan S. Panigoro
Publisher	:	Department of Mathematics, Universitas Negeri Gorontalo
Country	:	Indonesia
OAI Address	:	http://ejurnal.ung.ac.id/index.php/jjom/oai
Google Scholar ID	:	iWLjgaUAAAAJ
Email	:	info.jjom@ung.ac.id

Jambura Journal of Mathematics, Volume 7, Issue 1, Pages 14–18, February 2025 *https://doi.org/10.37905/jjom.v7i1.28290*

Research Article

Comparison of Random Forest, XGBoost and LightGBM Methods on the Human Development Index Classification

Yunna Mentari Indah[1,*], **Rafika Aristawidya**[1], **Anwar Fitrianto**[1], **Erfiani**[1], and **L.M. Risman Dwi Jumansyah**[1]

[1] *Study Program of Statistics and Data Science, IPB University, Bogor, Indonesia*

ARTICLE HISTORY
Received 11 October 2024
Revised 12 December 2024
Accepted 25 December 2024
Published 1 February 2025

KEYWORDS
Classification
Random Forest
XGBoost
LightGBM
Human Development Index

ABSTRACT. *Machine learning classification is an effective tool for categorizing data based on patterns, which is particularly useful in analyzing the Human Development Index (HDI) in Indonesia. HDI serves as a key indicator of regional development progress, making it crucial to classify HDI categories at the regency/city level to support targeted development planning. This study aims to compare the performance of three ensemble-based classification methods—Random Forest, XGBoost, and LightGBM—in classifying HDI categories in Indonesia. Data from the Central Bureau of Statistics (BPS) in 2023, comprising 514 observations across nine variables, was used for analysis. The study applied these algorithms to analyze the most influential variables affecting HDI. The results show that LightGBM outperformed both Random Forest and XGBoost, achieving an accuracy of 0.937 without outlier handling and 0.944 with outlier handling. Additionally, per capita expenditure was identified as the most influential factor in predicting HDI. These findings contribute to the field of statistical modeling by demonstrating how ensemble methods can improve classification accuracy and provide valuable insights for data-driven policymaking, thus enhancing regional development planning and supporting future HDI-related research.*

1. Introduction

The Human Development Index (HDI) is a measure of human development achievement for a region or area based on several essential components of quality of life formed through three main dimensions, namely long and healthy life using life expectancy at birth indicators, knowledge measured using expected years of schooling and average years of education, and a decent standard of living using purchasing power indicators through actual expenditure per capita. Indonesia's HDI in 2023 reached 74.39, an increase of 0.62 points (0.84 percent) compared to the previous year, which was 73.77 [1].

Predicting HDI accurately is an essential challenge for the government in determining development programs that are right on target and by regional priorities [2]. In addition, the level of achievement of HDI values in each district/city in Indonesia is influenced by the development programs implemented by the local government. The selected development program must be by the priorities and right on target based on the HDI category owned by each region. Therefore, a decision system is needed to accurately determine the classification of HDI categories in each district/city in Indonesia [3]. In machine learning, classification is a supervised learning technique employed to examine a given dataset and develop a model that divides the data into specific and distinct categories [4].

Decision trees are widely regarded as simple and intuitive tools for predicting outcomes, as they separate "high" and "low" values of a predictor about the target variable. However, despite their advantages, decision tree methodologies often need more accuracy when applied to complex datasets, such as those involving large volumes of data or intricate interactions between variables [5].

In addition, the method can produce less stable trees where small changes in learning data can cause significant changes in the trees formed and tend to overfit. Hence, to increase stability and avoid overfitting, the ensemble method is applied. Ensemble classification is considered more resistant to noise and can minimize bias and variance compared to single learning [6]. Ensemble classification is a method that combines several classification algorithms to increase model power and improve classification performance [7].

Several comparative studies of classification methods have been conducted, such as research conducted by Airlangga [8], which compared ensemble technique classification methods such as Extra Trees Classifier and LightGBM with traditional machine learning algorithms. The analysis showed that ensemble classification methods such as Extra Trees Classifier and LightGBM perform better than non-ensemble classification methods. In addition, research in the classification of dry beans using a comparison of the Gradient Boosting Machine, Random Forest, and Light GBM methods [9] showed that LightGBM is the best classification method. Therefore, this study aims to compare classification methods, namely Random Forest, XGBoost, and LightGBM, on the 2023 Human Development Index data. The results of this

[*] Corresponding Author.

Email : *yunnamentari@apps.ipb.ac.id* (Y. M. Indah)

Homepage : http://ejurnal.ung.ac.id/index.php/jjom/index / E-ISSN : 2656-1344

study can provide information related to the classification of HDI, the level of classification accuracy, and variables that affect the HDI.

2. Methods

2.1. Data and Data Source

The data used in this study is secondary data derived from the Central Statistics Agency (BPS) website in 2023. The data has 514 observations from all regencies/cities in Indonesia with the Human Development Index (Y) as the response variable, which is divided into four categories and consists of eight explanatory variables, namely Life Expectancy (X_1), Average Years of Schooling (X_2), Per Capita Expenditure (X_3), Percentage of Poor Population (X_4), Expected Years of Schooling (X_5), Open Unemployment Rate (X_6), Labor Force Participation Rate (X_7), and Percentage of Households with Access to Adequate Drinking Water Sources (X_8), According to [1], human development achievements in a region at a particular time can be grouped into four groups. This grouping aims to organize regions into groups similar to human development achievements.

1. Very high category : HDI $\geq$ 80
2. High category: 70 $\leq$ HDI $<$ 80
3. Medium category : 60 $\leq$ HDI $<$ 70
4. Low category: HDI $<$ 60

An explanation of the data types of the nine variables used in this study is presented in Table 1.

Table 1. Description of response variables and explanatory variables

Variables	Description	Data Type
Y	Human Development Index (HDI)	Categorical: 1 = low 2 = medium 3 = high 4 = very high
X_1	Life Expectancy	Numerical
X_2	Average Years of Schooling	Numerical
X_3	Per capita expenditure	Numerical
X_4	Percentage of Poor Population	Numerical
X_5	Expected Years of Schooling	Numerical
X_6	Open Unemployment Rate (TPT)	Numerical
X_7	Labor Force Participation Rate (TPAK)	Numerical
X_8	Percentage of Households with Access to Adequate Drinking Water Sources	Numerical

2.2. Data Analysis Steps

The classification analysis steps using HDI data are as follows:

1. Perform data pre-processing.
2. Divide the data with 80% training data and 20% testing data.
3. Perform classification using Random Forest, XGBoost, and LightGBM methods without handling outlier data.
4. Perform classification using the Random Forest, XGBoost, and LightGBM methods with outlier data handling using the Interquartile Range (IQR) method, replacing outlier values with the median or values closer to the IQR limit.
5. Comparing the three methods based on the best accuracy value.
6. Determine the explanatory variables that have the most influence on HDI data.

2.3. Random Forest

Breiman introduced the Random Forest method in 2001. Random forest has two functions, namely classification and prediction [4]. Random forest is one of the classification algorithms included in *ensemble learning* [10]. Random forest performs classification by adopting an *ensemble* approach from various trees through majority emergence to reach the final decision [11]. The following is the process in random forest classification presented in Figure 1 [12].

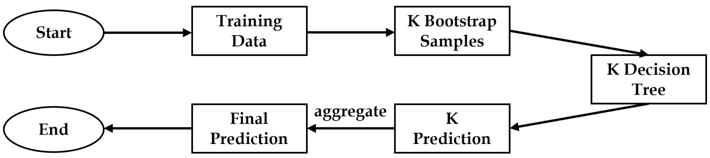


Figure 1. Random forest process

2.4. XGBoost

XGBoost stands for eXtreme Gradient Boosting, is an *ensemble* learning algorithm with a boosting method developed by Tianqi Chen in 2014. The XGBoost algorithm is decision tree-based [13] and is based on gradient boosting, which centers on examples misclassified by previous classifiers [14]. XGBoost is designed to prevent overfitting and optimize computational capability by simplifying the objective function that combines predictive and regularization terms, which controls model complexity and prevents overfitting while maintaining optimal computational speed [15].

In a tree-based algorithm, the inner nodes represent the values for the testing attributes and the leaf nodes with scores represent the decisions. The prediction result is the total score predicted by the K tree as eq. (1) [16].

$$\hat{y}_1 = \sum_{k}^{k} f_k\left(x_i\right), f_k \in F, \tag{1}$$

where each paragraph can be composed of multiple subparagraphs, the $\sum_{i-1}^{n} l(y_i, \hat{y}_1)$ is a loss function that can be differentiated to measure whether the model is suitable for the training dataset and the $\sum_{i-1}^{n} \Omega(f_k)$ is an item that determines the complexity of the model which can be seen in eq. (2).

$$\text{obj}(\theta) = \sum_{i-1}^{n} l(y_i, \hat{y}_1) + \sum_{i-1}^{n} \Omega(f_k). \tag{2}$$

2.5. LightGBM

LightGBM stands for Light Gradient Boosting Machine, which is one of the developments of gradient boosting that uses a decision tree-based learning algorithm developed to have higher speed [17]. Basic classifiers with decision trees are generated during training process, and weight parameters are calculated for each classifier in iterations [12].

$$f_m\left(X\right) = \partial_1 f_1\left(X\right) + \partial_2 f_2\left(X\right) + \cdots + \partial_m f_m(X). \tag{3}$$

All the base classifiers and their weights are then integrated to create the final model as eq. (3). From eq. (3), $f_m(X)$ means the base classifier and ∂m represent the weight parameter of each classifier.

3. Results and Discussion

3.1. Data Pre-Processing

Before classification, the data must be pre-processed to ensure the dataset is clean and suitable for analysis. This includes checking for missing values in the data, with the results showing no missing values. Next, this process involves checking for multicollinearity and identifying outliers. From 514 observations, the proportion of HDI response variables is shown in Figure 2.

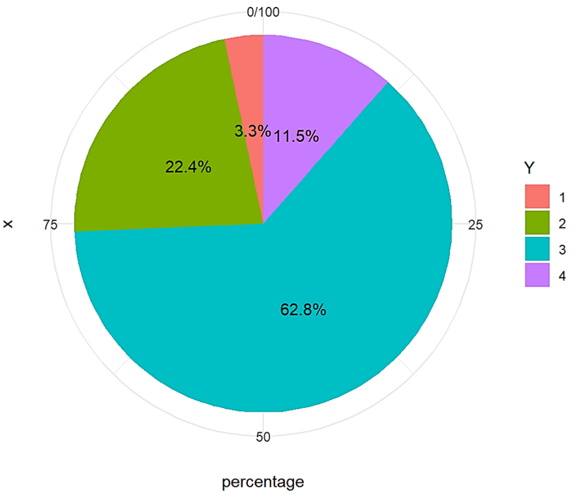


Figure 2. Proportion of HDI categories

Based on Figure 2, the proportion of low categories is 3.30% in as many as 17 observations, medium categories are 22.40% in as many as 115 observations, high categories are 62.80% in as many as 323 observations, and very high categories are 11.50%, as many as 59. Next, make a correlation between variables, here is a display of the correlation between variables in Figure 3.

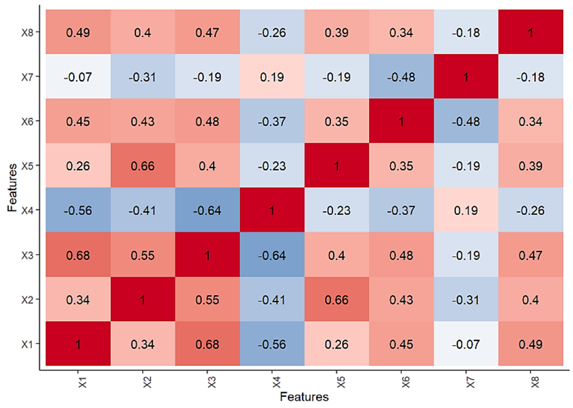


Figure 3. Heatmap correlation between variables

Based on the correlation matrix in Figure 3, it can be seen that the education dimension, specifically the correlation results, shows that Life Expectancy (X_1) has a strong positive relationship with Per Capita Expenditure (X_3) and Average Years of Schooling (X_2). In contrast, the Percentage of Poor Population (X_4) negatively correlates with X_1 and X_3. A strong relationship between X_2 and Expected Years of Schooling (X_5) is also observed. Other variables, like the Labor Force Participation Rate (X_7) and the Open Unemployment Rate (X_6), show weaker correlations. These correlations are crucial to consider for multicollinearity in the classification analysis.

Then, multicollinearity and outlier data are checked, presented in Table 2 and Figure 4. In checking no multicollinearity, it uses the VIF value or the value of the correlation matrix of all variables. The following are the results of multicollinearity testing.

Table 2. Multicollinearity in explanatory variables

X_1	X_2	X_3	X_4
2.714896	3.631882	2.749658	2.443979
X_5	X_6	X_7	X_8
2.641420	1.616458	1.416050	1.474205

In Table 2, shows that the VIF value given between HDI variables has a VIF value of less than 10 (VIF $<$ 10). So, there is no multicollinearity between all the variables.

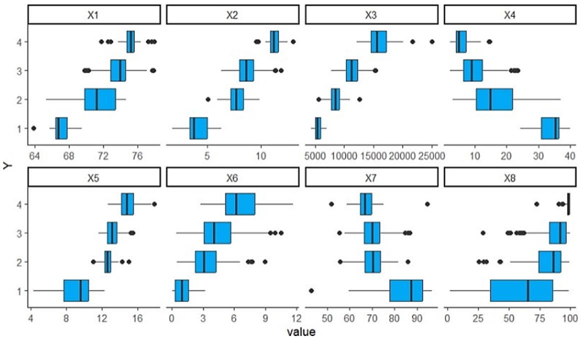


Figure 4. Outlier data checking

This boxplot shows that some variables have extreme values that deviate from the main data distribution. The black dots outside the whiskers on the boxplot represent values that are far outside the interquartile range. Some variables, such as life expectancy (X_1) and access to safe drinking water (X_8) have few outliers, while variables such as expenditure per capita (X_3) and percentage of poor people (X_4) show more outliers, especially in the medium and high HDI categories. X_2 shows low and medium HDI outliers, reflecting significant educational variations. X_5 has outliers in high and very high HDI, indicating differences in the predicted length of education. X_6 has outliers in low and very high HDI, reflecting significant variations in unemployment rates. X_7 has outliers in low and medium HDI, reflecting differences in labor force participation. This suggests a large variation in the data for certain variables that may need to be considered in further analysis.

3.2. Classification Analysis

After pre-processing the data, the next step is classification analysis, which categorizes data into predefined classes or groups based on specific features or attributes. The classification using Random Forest, XGBoost, and LightGBM methods. The classification results with and without outlier data handling are shown in Table 3.

Table 3. Classification accuracy results

	Method	Accuracy
Without Outlier Handling	Random Forest	0.921
	XGBoost	0.911
	LightGBM	**0.937**
Outlier Handling	Random Forest	0.911
	XGBoost	0.931
	LightGBM	**0.944**

Based on these results in Table 3, it is found that the LightGBM method is the best in determining the classification of Human Development Index data. In addition, the accuracy obtained from the three methods is similar when handling outlier data or not handling outlier data. The following displays of the accuracy comparison between the LightGBM, Random Forest, and XGBoost methods with and without outlier handling. Outlier handling is marked in green, while without outlier handling is marked in blue, as shown in Figure 5.

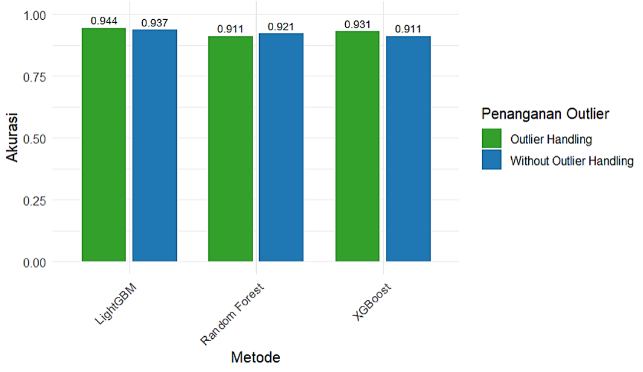


Figure 5. Bar chart of classification model accuracy

Basen on Figure 5, LightGBM shows higher accuracy with outlier handling, which is 0.944 compared to 0.937 without outlier handling. In Random Forest, the accuracy without outlier handling was 0.921, but after outlier handling, the accuracy slightly decreased to 0.911. In XGBoost, the accuracy without handling is 0.911 and increases with outlier handling to 0.931. Overall, the LightGBM method showed higher accuracy than the other methods. Furthermore, determining the most influential explanatory variables on HDI using the LightGBM method as shown in Figure 6.

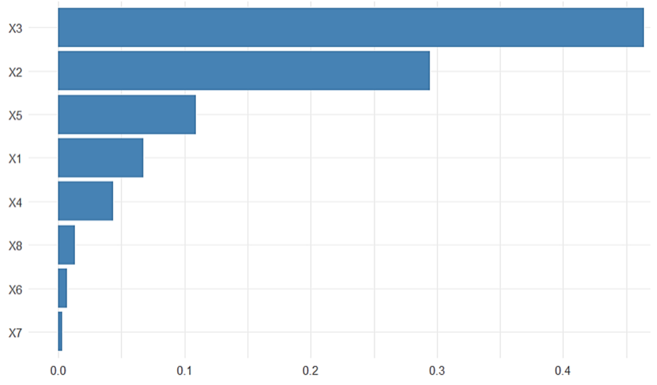


Figure 6. Most important explanatory variables

Based on Figure 6, it was found that the explanatory variable of per capita expenditure was the most influential, followed by the explanatory variables of average and expected years of schooling.

4. Conclusion

The performance of the three machine learning methods, Random Forest, XGBoost, and LightGBM, are compared based on the data used. LightGBM is the best-performing method in accuracy, both with and without outlier handling. LightGBM achieves the highest accuracy, 0.937 without outlier handling and 0.944 with outlier handling, demonstrating its ability to handle data, including data cleaned of extreme values effectively.

However, XGBoost increases accuracy after outlier handling, from 0.911 to 0.931, and is resilient to extreme values; LightGBM remains more consistent and efficient. On the other hand, although achieving a good accuracy of 0.921 without outlier handling, Random Forest experiences a decrease in performance (0.911) once outliers are addressed, indicating that this model is more sensitive to changes in the data distribution.

It is also important to note that the variable per capita expenditure has been identified as the most influential factor in the predictions. Understanding the impact of this variable has significant implications, particularly in the context of the Human Development Index (HDI) classification. This variable can provide valuable insights into socio-economic conditions and contribute to more accurate HDI assessments.

Author Contributions. **Yunna Mentari Indah:** Conceptualization, methodology, software, methodology, formal analysis, data curation, writing—original draft preparation, writing—review and editing, visualization. **Rafika Aristawidya:** Formal analysis, writing—original draft preparation, writing—review and editing. **Anwar Fitrianto:** Conceptualization, supervision. **Erfiani:** Conceptualization, supervision. **L.M. Risman Dwi Jumansyah:** Conceptualization, supervision. The final draft was approved by each author, and the study was validated and approved by each author upon their confirmation.

Acknowledgement. Thanks to the lecturers and assistants of the visualization data course of the Department of Statistics and Data Science, IPB University who have guided and provided advice and solutions when there were obstacles in the process of preparing this article and especially thankful for the editors and reviewers who have supported us in improving this manuscript.

Funding. This research received no external funding.

Conflict of interest. The authors declare no conflict of interest related to this article.

References

[1] BPS, "Indeks Pembangunan Manusia 2023," Jakarta: BPS, 2023.
[2] G. Alfian *et al.*, "Improving efficiency of RFID-based traceability system for perishable food by utilizing IoT sensors and machine learning model," *Food Control*, vol. 110, p. 107016, 2020, doi: 10.1016/j.foodcont.2019.107016.
[3] M. J. Paput, K. Suryowati, and M. T. Jatipaningrum, "Perbandingan Metode Random Forest dan Adaptive Boosting pada Klasifikasi Indeks Pembangunan Manusia di Indonesia," *Jurnal Statistika Industri Dan Komputasi*, vol. 8, no. 2, pp. 73–83, 2023, doi: 10.34151/statistika.v8i2.4458.
[4] P. N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*, Addison-Wesley Longman Publishing Co., Inc., 2005.

[5] J. L. Speiser, M. E. Miller, J. Tooze, and E. Ip, ”A comparison of random forest variable selection methods for classification prediction modeling,” *Expert Systems with Applications*, vol. 134, pp. 93–101, 2019, doi: 10.1016/j.eswa.2019.05.028.

[6] S. Mahmuda, D. A. Nohe, and A. M. Leonardo, ”Classification of the human development index in Kalimantan using random forest method,” in *Proceeding International Seminar of Science and Technology*, pp. 231–239, 2024, doi: 10.33830/isst.v3i1.2283.

[7] I. Syarif, E. Zaluska, A. Prugel-Bennett, and G. Wills, ”Application of bagging, boosting and stacking to intrusion detection,” in *Machine Learning and Data Mining in Pattern Recognition: 8th International Conference, MLDM 2012, Berlin, Germany, July 13-20, 2012. Proceedings 8*, Springer, 2012, pp. 593–602, doi: 10.1007/978-3-642-31537-4_46.

[8] G. Airlangga, ”Comparative Analysis of Machine Learning Models for Predicting Diabetes: Unveiling the Superiority of Advanced Ensemble Methods,” *G-Tech: Jurnal Teknologi Terapan*, vol. 8, no. 2, pp. 1272–1280, 2024, doi: 10.33379/gtech.v8i2.4246.

[9] I. Wardhana, M. Ariawijaya, V. A. Isnaini, and R. P. Wirman, ”Gradient Boosting Machine, Random Forest dan Light GBM untuk Klasifikasi Kacang Kering,” *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 6, no. 1, pp. 92–99, 2022, doi: 10.29207/resti.v6i1.3682.

[10] L. Breiman, ”Random forests,” *Machine Learning*, vol. 45, pp. 5–32, 2001.

[11] C. Yoo, D. Han, J. Im, and B. Bechtel, ”Comparison between convolutional neural networks and random forest for local climate zone classification in mega urban areas using Landsat images,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 157, pp. 155–170, 2019, doi: 10.1016/j.isprsjprs.2019.09.009.

[12] B. S. Wardani, S. Sa’adah, and D. Nurjanah, ”Measuring and Mitigating Bias in Bank Customers Data with XGBoost, LightGBM, and Random Forest Algorithm,” *Jurnal Ilmiah Teknik Elektro Komputer dan Informatika (JITEKI)*, vol. 9, no. 1, pp. 142–155, 2023, doi: 10.26555/jiteki.v9i1.25768.

[13] Y. Jiang, G. Tong, H. Yin, and N. Xiong, ”A pedestrian detection method based on genetic algorithm for optimize XGBoost training parameters,” *IEEE Access*, vol. 7, pp. 118310–118321, 2019, doi: 10.1109/ACCESS.2019.2936454.

[14] T. Chen and C. Guestrin, ”Xgboost: A scalable tree boosting system,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016, doi: 10.1145/2939672.2939785.

[15] J. Fan *et al.*, ”Comparison of Support Vector Machine and Extreme Gradient Boosting for predicting daily global solar radiation using temperature and precipitation in humid subtropical climates: A case study in China,” *Energy Conversion and Management*, vol. 164, pp. 102–111, 2018, doi: 10.1016/j.enconman.2018.02.087.

[16] S. Liang, ”Comparative Analysis of SVM, XGBoost and Neural Network on Hate Speech Classification,” *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, vol. 5, no. 5, pp. 896–903, 2021, doi: 10.29207/resti.v5i5.3506.

[17] R. Latifah and G. Erda, ”Application Of The Lightgbm Algorithm In The Classification Of Greenhouse Gas Emissions,” *Parameter: Journal of Statistics*, vol. 4, no. 1, pp. 9–15, 2024, doi: 10.22487/27765660.2024.v4.i1.17055.