Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871

Vol. 7, No. 2, April, Page. 1818-1833
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5651

Optimization Of Hybrid K-Means–Naïve Bayes Using Optuna for Classification of Global Plastic Waste Management Levels

Aulya Fani Madani[*1], Poningsih[2], Zulia Almaida[3], Widodo Saputra[4]

[1]Information Systems, STIKOM Tunas Bangsa, Indonesia
[2,4]Master of Informatics Study Program, STIKOM Tunas Bangsa, Indonesia
[3]Accounting Computerization, STIKOM Tunas Bangsa, Indonesia

Email: [1]aulyafaniramadani@gmail.com

Received : Jan 28, 2026; Revised : Feb 23, 2026; Accepted : Mar 12, 2026; Published : Apr 18, 2026

Abstract

The rapid growth of plastic waste has become a serious global environmental challenge, while existing waste management analysis methods often struggle to handle large and heterogeneous environmental datasets. This study aims to improve the classification of global plastic waste management performance by integrating K-Means clustering and Naïve Bayes with Optuna-based hyperparameter optimization. Using a dataset of global plastic waste indicators from multiple countries during 2020–2024, K-Means is first applied to generate waste management level clusters, which are then classified using Naïve Bayes. The hybrid model is further optimized by tuning the var_smoothing parameter using Optuna. Experimental results show that the hybrid approach improves classification performance compared to the baseline Naïve Bayes model, while the optimized model increases accuracy from 89% to 95% along with improvements in precision, recall, F1-score, and ROC-AUC. These results indicate that combining clustering-based labeling with automated hyperparameter optimization can enhance the reliability of machine learning models for large-scale environmental data analysis. Therefore, the proposed approach can support more accurate evaluation of global plastic waste management and assist data-driven environmental policy development.

Keywords : *clustering, naïve bayes, optuna, plastic waste management, classification.*

1. INTRODUCTION

Plastic waste management has become an increasingly critical global environmental issue due to the rising consumption of plastics each year. Plastic production has reached over 350 million tons per year, while recycling rates remain far from adequate, resulting in the accumulation of waste on land and in water bodies that threatens ecosystems [1]. The trend in plastic usage from 1950 to 2021 has shown a significant and continuous increase [2]. Plastic waste management requires effective recycling processes and appropriate marketing strategies to increase the value and competitiveness of recycled plastic products [3]. Moreover, meta-analysis studies indicate that plastic waste such as bags, bottles, and food packaging are the most dominant types in the environment [4]. The increase in plastic waste imports to Indonesia in 2018 has further exacerbated the domestic waste management situation [5].

As multidimensional environmental data continues to expand, data mining analysis is crucial for identifying patterns and supporting data-driven policy decisions. This approach has been applied across various fields, including healthcare, finance, education, disaster mitigation, and environmental management [6], [7], [8], [9], [10].

The K-Means algorithm is widely applied in environmental analysis due to its ability to cluster data based on characteristic similarity. Research on plastic pollution has demonstrated that K-Means can partition regions into clusters with varying levels of pollution risk [11]. This method has also been successfully employed in disaster vulnerability mapping, risk analysis, and the optimization of public services, such as waste management routing [12], [13], [14]. The effectiveness of the K-Means algorithm

Vol. 7, No. 2, April, Page. 1818-1833
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5651

can be improved by combining it with cluster validation methods, such as the Elbow, Silhouette, Davies–Bouldin, and Calinski–Harabasz methods, to determine the most appropriate number of clusters [15].

As a classification method, Naïve Bayes is a probabilistic approach widely used due to its simplicity, speed, and effectiveness with continuous data. Gaussian Naïve Bayes has been successfully applied in waste image classification, while studies on global plastic waste have found that imbalanced data distributions can reduce model performance [16], [17]. This algorithm has also been applied in air quality modeling, retail sales, and environmental pollution studies [18], [19], [20].

The hybrid combination of K-Means and Naïve Bayes offers dual advantages, namely the formation of homogeneous clusters for model stability alongside probabilistic-based classification. Studies on flood-prone areas, social assistance ranking, and disaster risk segmentation have demonstrated that the hybrid approach can improve accuracy and mitigate the instability of raw data [21], [22], [23].

Nevertheless, most previous studies have continued to rely on default parameters, resulting in suboptimal model performance. The Optuna framework has been proven to enhance model performance through more efficient hyperparameter search. In a study on stroke patient mortality risk prediction, XGBoost accuracy increased from 73% to 86% after tuning, in line with the findings of this study, which identify XGBoost as the best-performing model, achieving an accuracy of 86% in mortality prediction and 82% in length-of-stay prediction, accompanied by a significant improvement in AUC values [24], [25]. In an indigenous disease prediction system, LightGBM accuracy improved to >92% [26]. Optuna has also enhanced the performance of Random Forest in predicting contraceptive use in Ethiopia and in an OCB biomarker model, achieving a ROC-AUC of 0.902 [27], [28]. The consistency of these improvements demonstrates that hyperparameter optimization is essential for large and heterogeneous datasets.

Despite the growing use of machine learning for environmental analysis, limited studies have explored the integration of clustering-based labeling, probabilistic classification, and automated hyperparameter optimization within a unified framework for global plastic waste management analysis. Therefore, this study proposes a hybrid K-Means–Naïve Bayes model optimized using Optuna to improve classification performance on heterogeneous environmental datasets. The main contribution of this research is the development of an integrated analytical pipeline that enhances classification reliability while providing interpretable insights into global waste management patterns.

2. METHODS

This study aims to develop a more accurate and stable classification model for global plastic waste management levels through the integration of clustering techniques, classification methods, and hyperparameter optimization. The model is designed using the Knowledge Discovery in Databases (KDD) approach, which includes data cleaning, feature transformation, cluster formation, model training, hyperparameter optimization, and final evaluation stages [29]. The dataset used is obtained from Kaggle and comprises waste management indicators from ten countries over the period 2020–2024.

2.1. Data Collection

The dataset used in this study is obtained from an open-access source, namely the Global Environmental Impact dataset. This dataset consists of 3,000 data records from 10 countries over the period 2020–2024 [30].

The data include the following variables: Plastic Waste Tons, Plastic Recycled Tons, Burned Waste Tons, Waste Collected Tons, Illegal Dumping Cases, Government Interventions, Plastic Bans Enforced, Awareness Campaigns, Monitoring Stations, Country, Region, and Date. The dataset was verified to ensure completeness and consistency prior to further processing.

Vol. 7, No. 2, April, Page. 1818-1833
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5651

2.2. Data Preprocessing

The research dataset comprises 3,000 plastic waste management records from ten countries over the period 2020–2024. The dataset used in this study includes numerical, categorical, and temporal features.

Table 1. Description of Dataset Features

Feature Category	Features
Numerical	Plastic Waste Tons, Plastic Recycled Tons, Burned Waste Tons, Waste Collected Tons, Illegal Dumping Cases, Government Interventions, Plastic Bans Enforced, Awareness Campaigns, Monitoring Stations
Categorical	Country, Region
Temporal	Day, Month, Year

Table 1 presents the feature categories used in this study. The preprocessing stage includes checking for missing values, normalizing numerical features using StandardScaler, and applying One-Hot Encoding to Country and Region. Date features are decomposed into day, month, and year components to enhance temporal representation. To provide an overview of the raw data structure, an example of the research dataset comprising nine numerical variables is presented in Table 2.

Table 2. Sample of Raw Numeric Dataset

V1	V2	V3	V4	V5	V6	V7	V8	V9
2393.68	535.52	842.34	4	1	235.16	1	10	35
1494.46	1274.11	881.7	5	3	4292.34	0	12	28
1382.28	572.55	577.54	1	1	391.7	0	0	29
1382.3	258.1	750.74	3	4	1699.11	1	12	47
4649.71	54.06	286.07	5	2	877.99	0	7	46
2866.71	643.08	1381.93	3	9	2237.0	1	1	2
1131.08	416.08	1874.5	0	9	994.03	1	4	18
2469.81	1039.32	907.06	2	2	1866.22	1	7	1
1998.96	310.96	884.9	2	4	2418.98	0	4	2
215.91	788.1	1116.48	3	3	600.15	0	11	45
….	….	…..	….	….	….	….	….	….

Table 2 shows a sample of the raw numerical dataset used in this study. Each row represents a record of plastic waste management indicators, while the columns (V1–V9) correspond to numerical variables related to waste generation, recycling, policy implementation, and monitoring activities. This sample is provided to illustrate the structure and distribution of the data prior to preprocessing. The complete dataset undergoes normalization and transformation before being used in the clustering and classification stages.

2.3. Research Framework

The present research aims to design and optimize a classification model using a hybrid K-Means and Naïve Bayes approach. The research workflow is to assess the model's effectiveness prior to and following the optimization stage, allowing improvements in algorithm performance to be analyzed objectively.

Vol. 7, No. 2, April, Page. 1818-1833
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5651

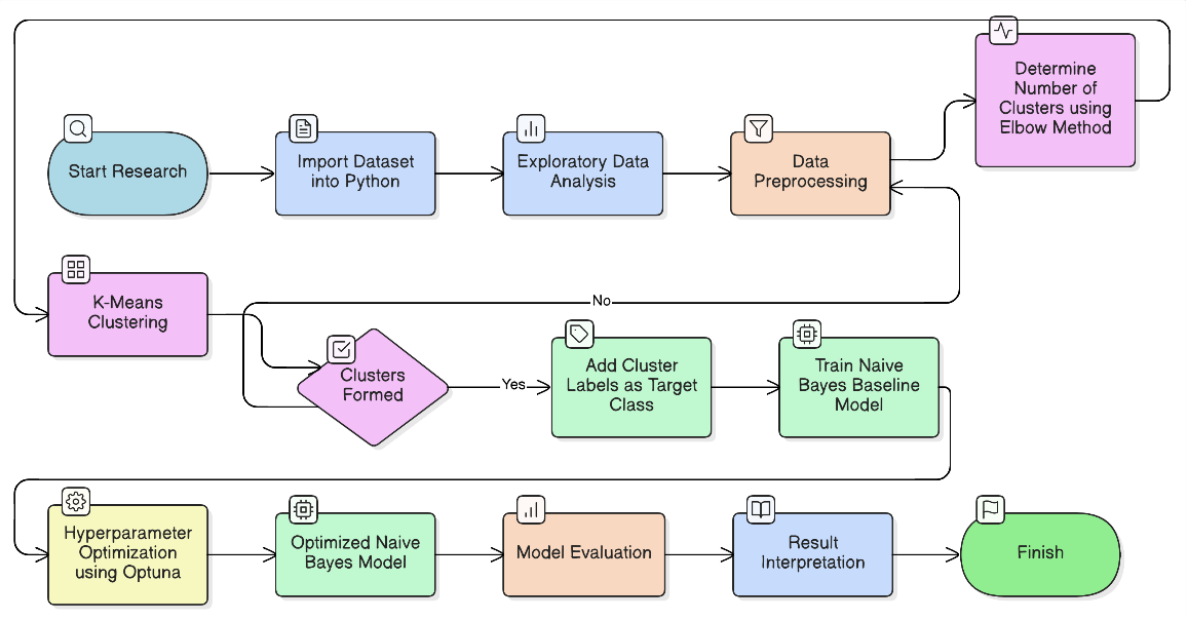

Figure 1. Research Framework of the Proposed Method for Global Plastic Waste Classification

The research framework presented in Figure 1 illustrates the progression of the study's phases. The study starts with gathering the dataset and performing exploratory data analysis, followed by preprocessing steps such as normalization and feature transformation. Afterwards, the best number of clusters is identified via the Elbow Method before executing the K-Means clustering algorithm to assign cluster labels. These labels are then used as the target class for building the Naïve Bayes classification model. After the baseline model is obtained, hyperparameter optimization is performed using Optuna to improve model performance. Finally, the optimized model is evaluated using several classification metrics and the results are interpreted to analyze the effectiveness of the proposed approach.

2.3.1. Data Collection

The research is based on global plastic waste management data from ten countries over the period 2020–2024. This data includes quantitative indicators representing the status of plastic waste management and serves as the basis for pattern formation and evaluation of classification model performance.

2.3.2. Data Preprocessing

The preprocessing stage is carried out to improve data quality prior to analysis. This process includes handling missing values, normalizing numerical data, and transforming categorical attributes. This stage is essential to ensure that classification algorithms can operate optimally, as emphasized in the foundational concepts of data mining–based classification [31], [32], [33].

2.3.3. Clustering with K-Means

The K-Means algorithm is used to cluster countries based on the similarity of their plastic waste management characteristics. The number of clusters is determined using the Elbow Method, which identifies the optimal point based on changes in the Sum of Squared Error (SSE). This method has been proven effective in various studies on environmental data clustering and waste management [34], [35], [36].

The distance between data points and cluster centroids is calculated using the Euclidean distance as follows:

Vol. 7, No. 2, April, Page. 1818-1833
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5651

$$d(x_i, c_j) = \sqrt{\sum_{k=1}^{n}(x_{ik} - c_{jk})^2} \quad (1)$$

Where:
x_i = represents the i-th data point,
c_j = denotes the centroid of cluster j, and
n = is the number of attributes.

The Elbow Method determines the optimal number of clusters by minimizing the SSE:

$$SSE = \sum_{j=1}^{k}\sum_{x_i \in C_j}|x_i - c_j|^2 \quad (2)$$

where :
k : represents the number of clusters.

2.3.4. Classification Using Naïve Bayes

Once the clusters are formed, the Naïve Bayes algorithm is employed to build the classification model. This method is chosen for its probabilistic nature and its efficiency in handling high-dimensional data, despite the assumption of feature independence [37]. The model serves as an initial reference before optimization is performed.

The Naïve Bayes classifier uses the following formula :

$$P(C|X) = \frac{P(X|C)P(C)}{P(X)} \quad (3)$$

Where:
$P(C \mid X)$ = represents the probability of class Cgiven predictor X.

Since the dataset contains continuous variables, Gaussian Naïve Bayes is applied:

$$P(x_i \mid C_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{(x_i - \mu_k)^2}{2\sigma_k^2}\right) \quad (4)$$

Where:
μ = Mean
σ = Standard deviation

2.3.5. Hyperparameter Optimization Using Optuna

To enhance the performance of the classification model, hyperparameter optimization is conducted using Optuna. The optimization process aims to automatically identify the best parameter configuration through a series of trials, enabling the model to achieve optimal performance [38]. This stage leverages a flexible and reproducible Python-based modeling approach and is executed within an interactive notebook environment to facilitate evaluation of experimental results [39].

The optimization process aims to identify the optimal parameter configuration by maximizing the objective function:

$$\theta^* = \arg\max_{\theta \in \Theta} f(\theta) \quad (7)$$

Where:
θ = represents the model parameters
$f(\theta)$ = denotes the evaluation metric such as accuracy or ROC-AUC.

Vol. 7, No. 2, April, Page. 1818-1833
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5651

2.3.6. Model Evaluation

The model's performance is evaluated using classification metrics such as accuracy, precision, recall, F1-score, and the confusion matrix. Evaluation is conducted on models both before and after optimization to assess the impact of hyperparameter tuning on performance improvement.

Accuracy :

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (6)$$

Where :
TP = True Positive
TN = True Negative
FP = False Positive
FN = False Negative

2.3.7. Model Comparison and Analysis

The final stage, the study compares the performance of the baseline Naïve Bayes model with the Optuna-optimized model. The analysis focuses on improvements in accuracy and prediction stability, thereby assessing the effectiveness of the optimization process in the context of global plastic waste management classification.

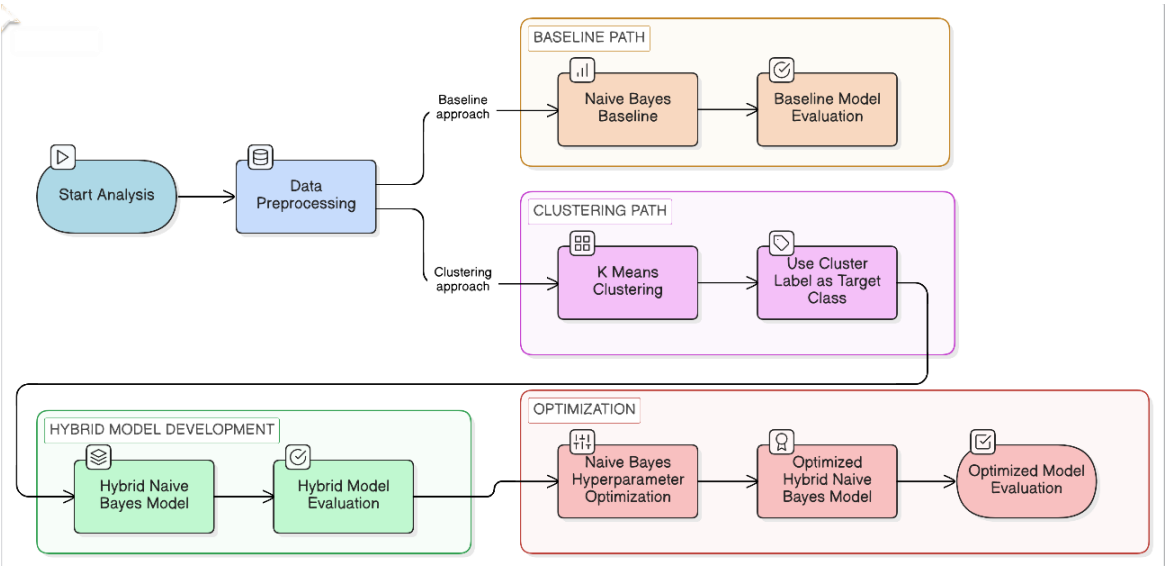

Figure 2. Comparison Framework of Naïve Bayes, Hybrid K-Means–Naïve Bayes, and Optimized Model

Figure 2 illustrates the comparison framework used in this study. The process begins with data preprocessing to prepare the dataset for modeling. A baseline classification model using Naïve Bayes is first trained and evaluated. In parallel, K-Means clustering is performed to generate cluster labels representing waste management levels. These labels are then used as target classes to build a Hybrid K-Means–Naïve Bayes model. After evaluating the hybrid model, hyperparameter optimization using Optuna is conducted to obtain the optimized model. The final stage evaluates the optimized model to analyze performance improvements compared to previous approaches.

3. RESULTS

This section presents and discusses the experimental results of classifying global plastic waste management levels using Naïve Bayes, Hybrid K-Means–Naïve Bayes, and Hybrid K-Means–Naïve

Vol. 7, No. 2, April, Page. 1818-1833
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5651

Bayes with Optuna optimization. The model's performance is assessed through metrics including accuracy, precision, recall, F1-score, confusion matrix, and ROC-AUC curves to assess its discrimination ability and stability. Furthermore, the performance of models before and after hyperparameter optimization is compared to identify the improvements achieved through the hybrid approach and optimization process.

3.1. Data Preprocessing Results

The preprocessing stage is conducted to ensure the quality and readiness of the data prior to modeling. This process includes checking and handling missing values, standardizing numerical features, and transforming categorical attributes using one-hot encoding. Additionally, temporal attributes are decomposed into day, month, and year components to enrich the temporal representation of the data.

Following preprocessing, the dataset is split into training and testing sets. This procedure aims to ensure that the data used for model training and testing phases are balanced and representative. The preprocessing results indicate that the data have been standardized and are ready for use with the K-Means and Naïve Bayes algorithms.

3.2. Clustering Results Using K-Means

The K-Means algorithm was applied to cluster the data based on similarities in plastic waste management characteristics. The number of clusters was determined using the Elbow Method, which identifies the optimal cluster count. Elbow Method visualization was employed to detect the "elbow" point, representing the condition where increasing the number of clusters no longer results in a significant reduction in inertia.

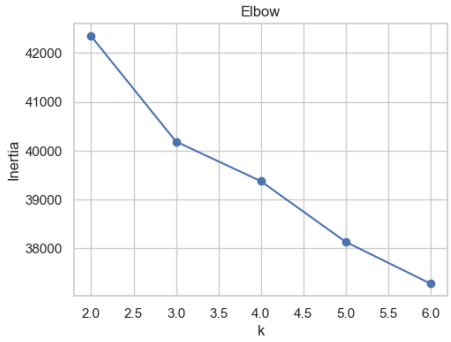

Figure 3. Elbow Method Curve for K-Means

Figure 3 illustrates the Elbow Method used to determine the optimal number of clusters in the dataset. The curve shows the relationship between the number of clusters and the Sum of Squared Errors (SSE). The visible elbow point appears at $k = 3$, indicating that three clusters (High, Medium, and Low) provide the best balance between model simplicity and clustering quality.

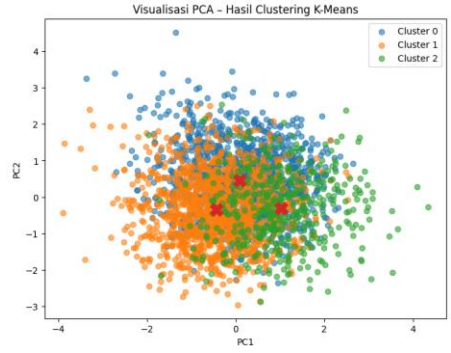

Figure 4. K-Means Cluster Distribution Visualized via PCA

Vol. 7, No. 2, April, Page. 1818-1833
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5651

Figure 4 presents the visualization of cluster distribution after applying K-Means using Principal Component Analysis (PCA). The visualization reduces high-dimensional data into two dimensions, allowing clearer observation of how the data points are grouped into clusters.

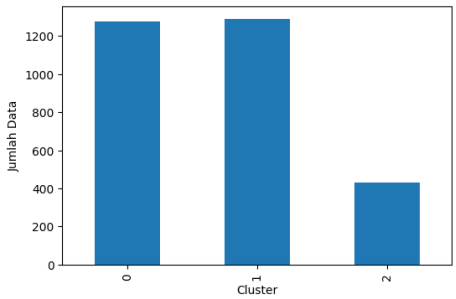

Figure 5. Number of Data Points per Cluster

Figure 5 shows the number of observations contained in each cluster generated by the K-Means algorithm. This visualization helps illustrate the distribution of data across clusters and indicates whether the clustering results are balanced.

Table 3. Sample of Plastic Waste Management Dataset with K-Means Cluster Labels

No	Country	Region	Plastic Waste Tons	Plastic Recycled Tons	Burned WasteTons	…	C
1	USA	Mountain	2393.68	535.52	842.34	…	0
2	India	Rural	1494.46	1274.11	881.70	…	1
3	China	Urban	1382.28	572.55	577.54	…	1
4	Russia	Urban	1382.30	258.10	750.74	…	0
5	Philippines	Forest	4649.71	54.06	286.07	…	1
6	Philippines	Suburban	2866.71	643.08	1381.93	…	0
7	India	Mountain	1131.08	416.08	1874.50	…	0
8	Brazil	Suburban	2469.81	1039.32	907.06	…	0
9	USA	Forest	1998.96	310.96	884.90	…	1
10	Philippines	Rural	215.91	788.10	1116.48	…	1
…	….	….	….	….	….	…	….

Table 3 presents a sample of the dataset after the clustering process. The table includes several waste management indicators along with the cluster labels produced by the K-Means algorithm. These cluster labels are later used as the target classes in the classification stage.

3.3. Naïve Bayes Classification Results

The Naïve Bayes model was employed as the baseline model in this study to evaluate the initial classification performance of global plastic waste management levels. The evaluation was conducted prior to the application of clustering and hyperparameter optimization methods. Model performance was analyzed using the confusion matrix, ROC curve, and classification metrics including accuracy, precision, recall, and F1-score. These evaluation results serve as a reference for comparing improvements in the hybrid K-Means–Naïve Bayes model and the Optuna-optimized model.

Figure 6 shows the confusion matrix of the Naïve Bayes baseline model. The matrix illustrates the distribution of correct and incorrect predictions across the three classes, providing insight into how well the model performs in distinguishing each category.

Vol. 7, No. 2, April, Page. 1818-1833
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5651

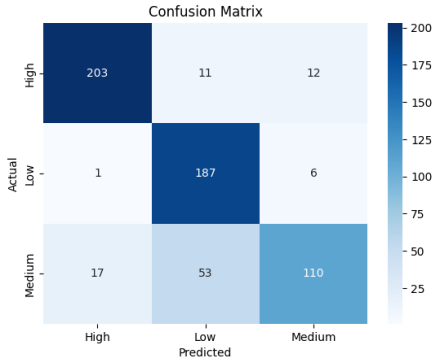

Figure 6. Confusion Matrix of Naïve Bayes Classification

Figure 7 displays the ROC curve of the Naïve Bayes model. The Naïve Bayes model achieved an accuracy of 83%, with precision and recall values of approximately 0.84 and 0.83, respectively. However, a more detailed evaluation using class-wise ROC-AUC revealed very low values, namely 0.09 for the Low class, 0.48 for the Medium class, and 0.31 for the High class. This indicates that although the model appears reasonably accurate overall, its ability to discriminate between individual classes remains poor. A likely cause is class imbalance, which leads the model to predominantly predict the majority class correctly without effectively learning patterns associated with the minority classes.

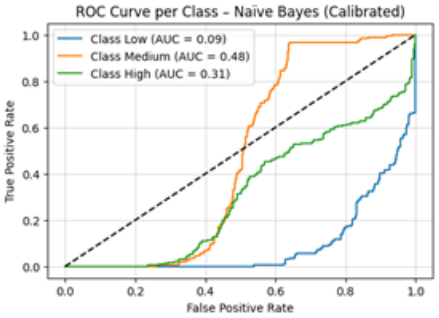

Figure 7. ROC Curve of Naïve Bayes Model

Figure 8 summarizes the main performance metrics of the Naïve Bayes model, including precision, recall, and F1-score. These metrics provide a quantitative evaluation of the baseline model before applying clustering and optimization techniques.

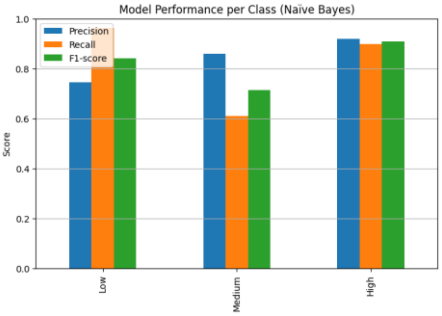

Figure 8. Performance Metrics of Naïve Bayes Model

3.4. Hybrid K-Means–Naïve Bayes Classification Results

To enhance classification performance, the hybrid K-Means–Naïve Bayes approach was applied by utilizing the K-Means clustering results as class labels in the classification stage. This strategy aims to form more homogeneous data groups prior to classification. Evaluation results indicate that the hybrid

Vol. 7, No. 2, April, Page. 1818-1833
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5651

model outperforms the baseline Naïve Bayes model, as evidenced by a reduction in classification errors in the confusion matrix, particularly for the medium class. These findings suggest that integrating clustering results assists Naïve Bayes in learning data patterns more systematically, thereby improving classification performance.

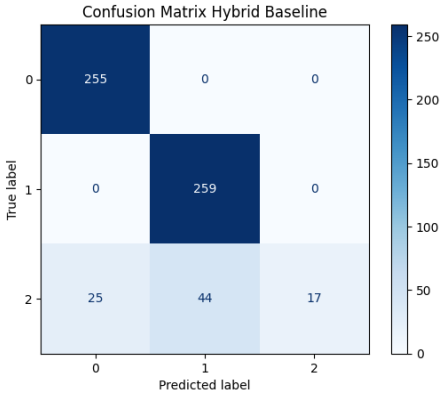

Figure 9. Confusion Matrix of Hybrid K-Means–Naïve Bayes Model

Figure 9 presents the confusion matrix of the Hybrid K-Means–Naïve Bayes model. Compared with the baseline model, the hybrid approach shows improved classification results with fewer misclassifications across the classes.

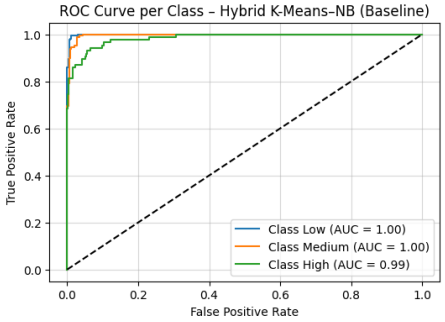

Figure 10. ROC Curve of Hybrid K-Means–Naïve Bayes Model

Figure 10 illustrates the ROC curve of the Hybrid K-Means–Naïve Bayes model. The application of K-Means clustering prior to Naïve Bayes significantly improved model performance. Accuracy increased to 89%, with precision and recall reaching 0.90 and 0.89, respectively, and class-wise ROC-AUC values approaching perfection (0.99–1.00). These results indicate that cluster formation helps the model separate the data more effectively, thereby substantially enhancing classification performance for each class. The proposed approach achieves superior overall accuracy and exhibits stronger class discrimination performance than the standalone Naïve Bayes model.

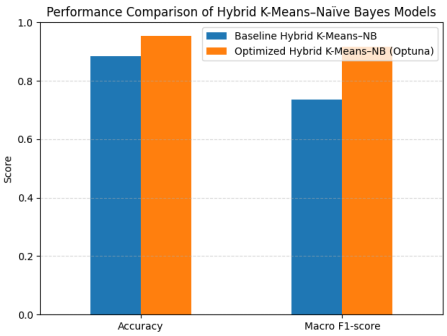

Figure 11. Performance Metrics of Hybrid K-Means–Naïve Bayes Model

Vol. 7, No. 2, April, Page. 1818-1833
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5651

Figure 11 presents the performance metrics obtained from the Hybrid K-Means–Naïve Bayes model. The results demonstrate improvements in accuracy and F1-score compared with the optimized model.

3.5. Hyperparameter Optimization Results Using Optuna

The next stage involved hyperparameter optimization of the Naïve Bayes model using Optuna. The optimization process focused on tuning the *var_smoothing* parameter through an automated trial-based search with a Stratified K-Fold cross-validation scheme.

The optimization results demonstrate improved model accuracy and stability compared to the hybrid model without optimization. The optimization history graph shows that accuracy values consistently increased with the number of trials, while the parameter importance analysis indicates that *var_smoothing* has a significant impact on model performance.

Figure 12 shows the confusion matrix of the optimized hybrid model after applying Optuna. The results indicate fewer classification errors and more consistent predictions across the classes.

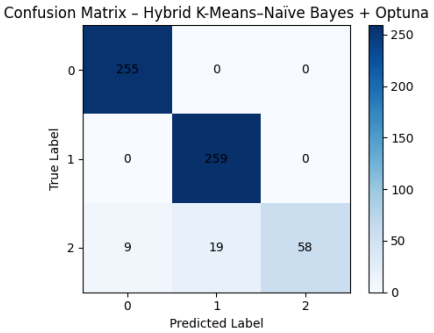

Figure 12. Confusion Matrix of Optimized Hybrid Model

Figure 13 presents the ROC curve of the optimized hybrid model. Optuna-based hyperparameter tuning enhanced performance, with accuracy reaching 95%, the F1-score achieved 0.95, and the class-wise ROC-AUC values attained 1.00, indicating optimal class discrimination capability. Through the combination of clustering and parameter tuning, the model is able to predict all classes with very high accuracy, with well-calibrated prediction probabilities and no observable trade-offs among classes. These results demonstrate that hyperparameter optimization can further refine model performance both overall and at the class level.

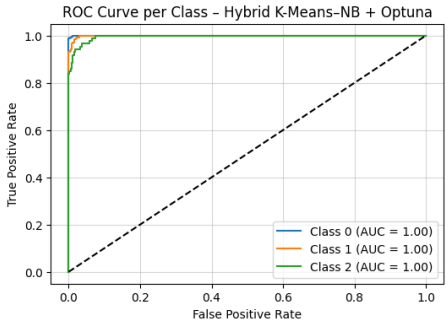

Figure 13. ROC Curve of Optimized Hybrid Model

Figure 14 illustrates the performance results of the optimized hybrid model for each class. The precision, recall, and F1-score values across the Low, Medium, and High categories indicate consistently strong classification performance. Although the High class records a slightly lower recall, the overall results demonstrate stable and well-balanced model effectiveness.

Vol. 7, No. 2, April, Page. 1818-1833
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5651

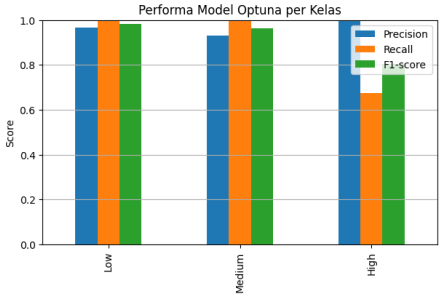

Figure 14. Performance Metrics of Optimized Hybrid Model

3.6. Model Comparisson

This section presents a performance comparison among three classification approaches: Naïve Bayes as the baseline model, Hybrid K-Means–Naïve Bayes, and the hyperparameter-optimized model using Optuna. The comparison is conducted to evaluate the impact of cluster formation and the optimization process on improving the classification performance of global plastic waste management levels. The evaluation focuses on metrics including accuracy, precision, recall, and F1-score.

Table 4. Performance Comparison of Classification Models

Model	Accuracy	Precision	Recall	F1-Score
Naïve Bayes	0.83	0.84	0.83	0.83
Hybrid K-Means–Naïve Bayes	0.89	0.90	0.89	0.85
Hybrid K-Means–Naïve Bayes + Optuna	**0.95**	**0.96**	**0.95**	**0.95**

Table 4 compares the comparative results of the three classification models examined in this research. The results show a gradual improvement from the baseline Naïve Bayes model to the hybrid model and finally to the Optuna-optimized model, indicating that clustering and hyperparameter optimization contribute positively to classification performance.

4. DISCUSSION

The experimental results demonstrate that the proposed hybrid framework integrating K-Means clustering, Naïve Bayes classification, and Optuna-based hyperparameter optimization yields noticeable improvements in the analysis of global plastic waste management data. As shown in Table 4, the optimized hybrid model achieves higher classification accuracy compared to the baseline Naïve Bayes classifier. These findings indicate that combining clustering-based data organization with automated parameter tuning can strengthen the predictive performance of probabilistic classification models when applied to environmental datasets.

The consistent enhancement observed across evaluation metrics suggests that the proposed framework not only increases overall prediction accuracy but also improves the stability of classification outcomes across varying data patterns. From a methodological standpoint, the application of K-Means clustering assists in grouping data instances that share similar characteristics prior to the classification stage. This step enables the classifier to learn patterns from more homogeneous data groups. As a result, the Naïve Bayes algorithm can estimate class probabilities more reliably because the feature distribution becomes more structured.

Furthermore, the incorporation of Optuna for hyperparameter optimization significantly contributes to improving the performance of the classification model. Rather than depending on manually selected parameter settings, the optimization procedure systematically explores the parameter space to determine configurations that produce the best model performance. Through this automated

Vol. 7, No. 2, April, Page. 1818-1833
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5651

search process, the model gains better generalization ability when analyzing environmental indicators such as recycling rates, waste generation levels, and waste treatment patterns.

Table 5. Performance Comparison with Related Classification Studies

Reference	Method	Accuracy
2025 [16]	Naïve Bayes for Waste Classification	79.50%
2025 [40]	Naïve Bayes for Air Quality Classification	90.00%
2022 [41]	Naïve Bayes + K-Means Clustering	80.80%
2025 [42]	DenseNet121 Waste Classification	91.00%
Proposed Method	**Hybrid K-Means + Naïve Bayes + Optuna**	**95.00%**

To further strengthen the evaluation of the proposed method, Table 5 presents a comparative analysis with several prior studies in the fields of environmental data classification and related machine learning applications. Previous research applying Naïve Bayes for waste classification reported an accuracy of 79.50%, while studies on air quality classification achieved approximately 90%. A recent study integrating Gaussian Naïve Bayes with K-Means clustering for air quality and pollution assessment reported an accuracy of 93%, demonstrating that the combination of supervised and unsupervised learning approaches can enhance classification performance. In addition, deep learning-based waste classification methods such as DenseNet121 achieved an accuracy of around 91%. Compared with these studies, the proposed hybrid K-Means–Naïve Bayes framework optimized using Optuna achieved the highest accuracy of 95%. This comparison indicates that integrating clustering-based data structuring with automated hyperparameter optimization provides measurable performance improvements across different environmental and analytical domains

Compared with these studies, the proposed hybrid K-Means–Naïve Bayes model optimized with Optuna demonstrates competitive performance. The higher accuracy obtained in this study suggests that combining clustering with probabilistic classification and automated optimization can produce a more effective analytical framework for environmental datasets. However, differences in dataset characteristics, preprocessing strategies, feature composition, and evaluation procedures should be considered when interpreting the comparison results. Therefore, the comparison presented in this study should be viewed as contextual evidence rather than a strict experimental benchmark.

The performance of the proposed hybrid model should also be interpreted in relation to the characteristics of the dataset and the experimental design used in this study. The dataset used in this research consists of global plastic waste indicators collected from multiple countries and environmental monitoring sources. The preprocessing stage includes feature transformation, categorical encoding, and exploratory data analysis to ensure data quality and consistency. Nevertheless, conducting further research with larger environmental datasets, more regions, and extended time periods would help assess how robust and scalable the proposed approach is.

Moreover, although the integration of clustering and hyperparameter optimization improves predictive performance, it also introduces additional computational processes during the model development stage. The clustering stage must be performed before classification, and the optimization process requires multiple training iterations. However, these additional processes occur only during model training and do not significantly affect the efficiency of the final prediction stage. Once the optimal parameters are identified, the deployed model remains relatively lightweight and computationally efficient, making it suitable for practical environmental data analysis scenarios.

5. CONCLUSION

This study evaluated a hybrid K-Means–Naïve Bayes classification framework optimized using Optuna for classifying global plastic waste management levels. The baseline Naïve Bayes model showed moderate performance and limited class discrimination due to heterogeneous data. The integration of

Vol. 7, No. 2, April, Page. 1818-1833
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5651

K-Means clustering improved classification by forming more homogeneous class labels, resulting in higher accuracy and better ROC-AUC performance.

Optuna-based hyperparameter optimization further enhanced model stability and increased accuracy from 89% to 95%. These results demonstrate that combining clustering-based labeling with automated optimization can significantly improve probabilistic classification on complex environmental datasets. From an informatics perspective, this study provides a data-driven analytical approach that can support environmental monitoring and decision-making related to global plastic waste management. Future research may explore other algorithms and feature selection methods to further improve model performance.

CONFLICT OF INTEREST

The authors confirm that there are no financial, professional, or personal conflicts that could have influenced the results or interpretation of this study. The research was conducted independently without any conflict related to the research object or involved parties.

ACKNOWLEDGEMENT

The authors sincerely thank STIKOM Tunas Bangsa for the support and facilities provided during the completion of this research. The authors also appreciate the providers of the dataset that made this study possible.

REFERENCE

[1] World Bank, "Tackling the plastics pollution crisis by channeling private capital to projects that reduce plastic waste.," World Bank. Accessed: Nov. 13, 2025. [Online]. Available: https://www.worldbank.org/en/news/feature/2024/01/25/tackling-the-plastics-pollution-crisis-by-channeling-private-capital-to-projects-that-reduce-plastic-waste?utm_source=chatgpt.com

[2] OECD, *Global Plastics Outlook*. 2022. doi: 10.1787/de747aef-en.

[3] A. Yuliarsono, "Analisis Strategi Pemasaran dan Pengolahan Daur Ulang Limbah Plastik," *remik*, vol. 9, no. 3, pp. 780–790, Aug. 2025, doi: 10.33395/remik.v9i3.14871.

[4] M. R. Kelly, M. R. Cordova, S. Jobling, and R. C. Thompson, "Meta-analysis of the spatial distribution and composition of plastic macro-debris in Indonesia," *Reg. Stud. Mar. Sci.*, vol. 90, no. June, p. 104460, 2025, doi: 10.1016/j.rsma.2025.104460.

[5] M. A. Septiono, "Indonesia Waste Trade Updates: Focusing on Plastic and Paper Waste in Indonesia Grid-Arendal," no. November 2022, p. 33, 2022, doi: 10.13140/RG.2.2.12149.45280.

[6] R. Aspiah and Taghfirul Azhima Yoga Siswa, "Implementasi Correlation Based Feature Selection (Cfs) Untuk Peningkatan Akurasi Algoritma C4.5 Dalam Prediksi Performa Akademik Mahasiswa Berbasis Learning Management System," *J. Ilm. Betrik*, vol. 13, no. 2, pp. 199–207, Aug. 2022, doi: 10.36050/betrik.v13i2.523.

[7] M. Gibril and R. Selamat, "Sistem Deteksi Fraud Menggunakan Data Mining, Data Warehouse, dan OLAP di Bank of India Indonesia," *J. Compr. Sci.*, vol. 4, no. 8, pp. 2570–2580, 2025, doi: 10.59188/jcs.v4i8.3543.

[8] R. Y. Hayuningtyas and R. Sari, "Implementasi Data Mining Dengan Algoritma Multiple Linear Regression Untuk Memprediksi Penyakit Diabetes," *J. Tek. Komput.*, vol. 8, no. 1, pp. 40–44, Jan. 2022, doi: 10.31294/jtk.v8i1.11552.

[9] A. Prasetio, M. M. Effendi, and M. N. Dwi M, "Analisis Gempa Bumi Di Indonesia Dengan Metode Clustering," *Bull. Inf. Technol.*, vol. 4, no. 3, pp. 338–343, Sep. 2023, doi: 10.47065/bit.v4i3.820.

[10] R. Nugraha, N. Suarna, I. Ali, and D. Rohman, "Optimasi Pengelolaan Sampah Melalui Model Pengelompokan Dengan Algoritma K-Means," *J. Inform. dan Tek. Elektro Terap.*, vol. 13, no. 1, pp. 646–652, 2025, doi: 10.23960/jitet.v13i1.5694.

[11] C. Darmawan, Y. Setiyawan, R. A. Prasetyo, and S. K. Qurrota'Ayyun, "Penerapan Algoritma K-means dan Metode Elbow Untuk Clustering Tingkat Pencemaran Sampah Plastik pada Kabupaten/Kota di Seluruh Indonesia," *G-Tech J. Teknol. Terap.*, vol. 8, no. 1, pp. 349–358, Jan.

Vol. 7, No. 2, April, Page. 1818-1833
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5651

2024, doi: 10.33379/gtech.v8i1.3637.

[12] Isni Rinjani, Saeful Anwar, and Ruli Herdiana, "PENGELOMPOKAN DAERAH BENCANA ALAM MENGGUNAKAN ALGORITMA K-MEANS CLUSTERING," *J. Ilm. Sist. Inf. dan Ilmu Komput.*, vol. 3, no. 1, pp. 35–51, Mar. 2023, doi: 10.55606/juisik.v3i1.417.

[13] A. N. B. Prasetyo, M. Maimunah, and P. Sukmasetya, "K-Means Clustering Method for Determining Waste Transportation Routes to Landfill," *J. Ris. Inform.*, vol. 5, no. 3, pp. 277–284, 2023, doi: 10.34288/jri.v5i3.219.

[14] M. Hanafi, B. Warsito, and R. Gernowo, "Sistem Informasi Manajemen Pengumpulan dan Pengangkutan Sampah Padat dengan Efisiensi Rute Menggunakan K-Means Clustering dan Travelling Salesman Problem," *J. Sist. Inf. Bisnis*, vol. 12, no. 2, pp. 106–115, 2022, doi: 10.21456/vol12iss2pp106-115.

[15] I. F. Ashari, E. Dwi Nugroho, R. Baraku, I. Novri Yanda, and R. Liwardana, "Analysis of Elbow, Silhouette, Davies-Bouldin, Calinski-Harabasz, and Rand-Index Evaluation on K-Means Algorithm for Classifying Flood-Affected Areas in Jakarta," *J. Appl. Informatics Comput.*, vol. 7, no. 1, pp. 89–97, Jul. 2023, doi: 10.30871/jaic.v7i1.4947.

[16] A. A. Alimun, H. Harlinda, and H. Azis, "Klasifikas Sampah Menggunakan Metode Naive Bayes," *LINIER Lit. Inform. dan Komput.*, vol. 2, no. 3, pp. 459–466, Oct. 2025, doi: 10.33096/linier.v2i3.3155.

[17] Muhammad Satria Nugraha, Imiel Ardhanenggar Tallane, Nabila Nur Fadhilah, Putri Citra Arrahma, Rifa Abdussalam, and Anna Dina Kalifia, "Analisis Data Sampah Plastik Dunia Pada Tahun 2023 Dengan Metode Naive Bayes," *J. Teknol. Komput. dan Inf.*, vol. 12, no. 2, pp. 152–157, 2024, doi: 10.52072/jutekinf.v12i2.1165.

[18] Sugeng Dwi Budi Priantoro, M Ghofar Rohman, and Moh Rosidi Zamroni, "Klasifikasi Kualitas Udara Dengan Metode Naive Bayes Berbasis Web," *Rabit J. Teknol. dan Sist. Inf. Univrab*, vol. 10, no. 2, pp. 1024–1035, 2025, doi: 10.36341/rabit.v10i2.6447.

[19] N. W. Wardani, P. G. S. C. Nugraha, and G. S. Mahendra, "Implementasi Naïve Bayes Pada Data Mining Untuk Mengklasifikasikan Penjualan Barang Terlaris Pada Perusahaan Ritel," *JST (Jurnal Sains dan Teknol.*, vol. 12, no. 3, pp. 656–668, 2024, doi: 10.23887/jstundiksha.v12i3.38605.

[20] D. D. Purwanto and E. S. Honggara, "Klasifikasi Kategori Hasil Perhitungan Indeks Standar Pencemaran Udara dengan Gausian Naïve Bayes (Studi Kasus: ISPU DKI Jakarta 2020)," *J. Intell. Syst. Comput.*, vol. 4, no. 2, pp. 102–108, 2022, doi: 10.52985/insyst.v4i2.259.

[21] I. M. Sinatrya, A. B. Pohan, Y. Yunita, H. Amalia, and A. F. Lestari, "Penerapan Integrasi Algoritma K-Means Dan Naïve Bayes Untuk Klasifikasi Wilayah Rawan Banjir Di Jakarta," *Comput. Sci.*, vol. 5, no. 2, pp. 67–76, 2025, doi: 10.31294/coscience.v5i2.6900.

[22] F. M. Sarimole and L. Nurmayanti, "Sistem Data Mining Penentuan Prioritas terhadap Penerima Bantuan Bencana Banjir dengan Metode Naive Bayes dan Klusterisasi K-Means (Studi Kasus: Wilayah Cengkareng 2025)," *J. Pengabdi. Nas. Indones.*, vol. 6, no. 3, pp. 685–697, 2025, doi: 10.63447/jpni.v6i3.1609.

[23] V. R. Prasetyo, G. Erlangga, and D. A. Prima, "Analisis Sentimen untuk Identifikasi Bantuan Korban Bencana Alam berdasarkan Data di Twitter Menggunakan Metode K-Means dan Naive Bayes," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 10, no. 5, pp. 1055–1062, 2023, doi: 10.25126/jtiik.2023107077.

[24] I. Tahyudin, A. Tikaningsih, P. Lestari, E. Winarto, and N. Hassa, "Optimizing Stroke Mortality Prediction: A Comprehensive Study on Risk Factors Analysis and Hyperparameter Tuning Techniques," *TEM J.*, vol. 13, no. 1, pp. 705–717, 2024, doi: 10.18421/TEM131-74.

[25] A. Tikaningsih, P. Lestari, A. Nurhopipah, I. Tahyudin, E. Winarto, and N. Hassa, "Optuna Based Hyperparameter Tuning for Improving the Performance Prediction Mortality and Hospital Length of Stay for Stroke Patients," *Telematika*, vol. 17, no. 1, pp. 1–16, Feb. 2024, doi: 10.35671/telematika.v17i1.2816.

[26] L.-H. Lai *et al.*, "The Use of Machine Learning Models with Optuna in Disease Prediction," *Electronics*, vol. 13, no. 23, p. 4775, Dec. 2024, doi: 10.3390/electronics13234775.

[27] J. B. Adem *et al.*, "Explainable machine learning algorithms to identify predictors of intention to use family planning among women of reproductive-age in Ethiopia: Evidence from the

Vol. 7, No. 2, April, Page. 1818-1833
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5651

Performance Monitoring and Accountability (PMA) 2021 survey data set," *BMJ Public Heal.*, vol. 3, no. 1, p. e000962, 2025, doi: 10.1136/bmjph-2024-000962.

[28] H. Gözgöz, O. Orhan, B. Akan Konuk, and P. Akan, "A machine learning model for predicting oligoclonal band positivity using routine cerebrospinal fluid and serum biochemical markers," *Am. J. Clin. Pathol.*, vol. 164, no. 6, pp. 933–945, 2025, doi: 10.1093/ajcp/aqaf119.

[29] D. Papakyriakou and I. S. Barbounakis, "Data Mining Methods: A Review," *Int. J. Comput. Appl.*, vol. 183, no. 48, pp. 5–19, 2022, doi: 10.5120/ijca2022921884.

[30] K. Yadav, "Global Environmental Impact," Kaggle. Accessed: Nov. 14, 2025. [Online]. Available: https://www.kaggle.com/datasets/khushikyad001/global-environmental-impact?resource=download

[31] Noviyanto, M. Wahyudi, and S. Sumanto, "Comparison of Supervised Learning Classification Methods on Accreditation Data of Private Higher Education Institutions," *Paradig. - J. Komput. dan Inform.*, vol. 26, no. 1, pp. 24–29, 2024, doi: 10.31294/p.v26i1.3306.

[32] M. F. M. Khalik and F. Arifin, "Klasifikasi Indeks Kedalaman Kemiskinan Provinsi Sulawesi Selatan Berbasis Decision Tree, K-Nearest Neighbor, Naive Bayes, Neural Network, dan Random Forest," *J. Edukasi dan Penelit. Inform.*, vol. 9, no. 2, p. 282, 2023, doi: 10.26418/jp.v9i2.67492.

[33] D. B. A and N. Mangla, "A Novel Network Intrusion Detection System Based on Semi-Supervised Approach for IoT," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 4, pp. 207–216, 2023, doi: 10.14569/IJACSA.2023.0140424.

[34] A. N. Azmi, A. L. O. Siregar, F. I. Lesmana, A. A. Nasir, and F. Kartiasih, "Implementasi machine learning dalam pengelompokan provinsi di Indonesia berdasarkan data pencemaran lingkungan hidup," *e-Jurnal Sumberd. dan Lingkung.*, vol. 14, no. 2, pp. 113–128, 2025, [Online]. Available: https://doi.org/10.22437/jesl.v14i2.37366

[35] M. Nurrohman, M. Maimunah, and P. Sukmasetya, "Sistem Klasterisasi Volume Sampah Organik di Kota Magelang menggunakan K-Means," *TEMATIK*, vol. 10, no. 1, pp. 146–153, Jun. 2023, doi: 10.38204/tematik.v10i1.1338.

[36] F. Salsabila, T. Ridwan, and H. H, "Analisa Volume Penyebaran Sampah Di Karawang Menggunakan Algoritma K-Means Clustering," *J. Inform. dan Tek. Elektro Terap.*, vol. 12, no. 2, 2024, doi: 10.23960/jitet.v12i2.4226.

[37] Z. Zulkipli, K. Kusrini, and S. Sudarmawan, "Prediksi Tingkat Kesehatan Lingkungan Masyarakat Dalam Program Sustainable Development Goals Menggunakan Algoritma Naive Bayes," *Infotek J. Inform. dan Teknol.*, vol. 6, no. 2, pp. 431–442, Jul. 2023, doi: 10.29408/jit.v6i2.18776.

[38] A. Efendi, I. Fitri, and G. W. Nurcahyo, "Development of a machine learning model with optuna and ensemble learning to improve performance on multiple datasets," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 41, no. 1, p. 375, Jan. 2026, doi: 10.11591/ijeecs.v41.i1.pp375-386.

[39] S. Samuel and D. Mietchen, "Computational reproducibility of Jupyter notebooks from biomedical publications," *Gigascience*, vol. 13, pp. 1–23, 2024, doi: 10.1093/gigascience/giad113.

[40] A. F. Fadhilah, A. R. Juwita, Y. E. Wicaksana, and T. Al Mudzakir, "Air Quality Classification Using Naive Bayes Algorithm With SMOTE Technique Based on ISPU Data," *JISA(Jurnal Inform. dan Sains)*, vol. 8, no. 1, pp. 16–22, Jun. 2025, doi: 10.31326/jisa.v8i1.2181.

[41] D. Barber, "Penilaian Kualitas Udara Dan Analisis Polusi Berbasis Algoritma Naive Bayesdan Klustrerisasi Data Dengan K-Means," *Bayesian Reason. Mach. Learn.*, vol. 13, no. 3, pp. 243–255, 2012, doi: 10.1017/cbo9780511804779.014.

[42] J. S. Mboli and O. A. Ogungbemi, "AI-Enabled Waste Classification as a Data-Driven Decision Support Tool for Circular Economy and Urban Sustainability," in *2025 IEEE International Smart Cities Conference (ISC2)*, IEEE, Oct. 2025, pp. 1–6. doi: 10.1109/ISC266238.2025.11293327.