International Journal of Informatics and Data Science
ISSN 3026-7315 (media online)
Volume 3, No 1, December 2025, Page: 34−43
DOI: doi.org/10.64366/ijids.v3i1.571
https://journals.adaresearch.or.id/ijids/index

Comparative Analysis of Machine Learning Classification and Clustering for Predicting PIP Eligibility

Mohamad Ilyas Abas[1,*], Rozmin binti Mohamad Rasli[2], Nur'aini Mufaidhah Tangahu[1], Anisya R. Ibrahim[1], Indrayana Saleh[1], Isminarti[3], Asminar[4], Deny Wiria Nugraha[5], Arthur Gani Koto[6], Samid Saripi[7]

[1] Computer Science, Universitas Muhammadiyah Gorontalo, Gorontalo, Indonesia
[2] Faculty of Computing and Meta-Technology, Sultan Idris Education University, Perak, Malaysia
[3] Mechatronics Engineering, Politeknik Bosowa, South Sulawesi, Indonesia
[4] Department of Electrical Engineering, Universitas Halu Oleo, Southeast Sulawesi, Indonesia
[5] Magister of Informatics Technology, Universitas Tadulako, Central Sulawesi, Indonesia
[6] Geography, Universitas Muhammadiyah Gorontalo, Gorontalo, Indonesia
[7] Primary Education, Universitas Muhammadiyah Gorontalo, Gorontalo, Indonesia
Email: [1,*]ilyasabas@umgo.ac.id, [2]roznim@meta.upsi.edu.my, [3]nuraint@gmail.com, [4]anisyari@gmail.com, [5]indrayanasaleh@gmail.com, [6]isminarti@politeknikbosowa.ac.id, [7]asminar.ft@uho.ac.id, [8]deny.wiria.nugraha@untad.ac.id, [9]arthur@umgo.ac.id, [10]samid@umgo.ac.id
Correspondence Author Email: ilyasabas@umgo.ac.id
Received: 15/12/2025, Revised: 26/12/2025, Accepted: 30/12/2025, Published: 31/12/2025

Abstract–Accurate identification of students eligible for the Program Indonesia Pintar (PIP) is essential to support fair and effective distribution of educational financial assistance. Manual selection processes may be time-consuming and susceptible to inconsistent decision-making. This study comparatively evaluates five machine learning classification algorithms, Decision Tree, Random Forest, K-Nearest Neighbor (KNN), Support Vector Machine (SVM), and Logistic Regression, for predicting PIP eligibility and additionally applies K-Means clustering to explore socioeconomic patterns among students. The dataset was obtained from SMA Negeri 1 Telaga Biru and consists of 1,092 student records containing demographic and socioeconomic attributes. Data preprocessing included missing-value handling, feature selection, Ordinal Encoding, undersampling, Min-Max normalization, and a 70:30 train–test split. Classification performance was evaluated using Accuracy, Precision, Recall, and F1-Score, while clustering quality was assessed using the Silhouette Score. Decision Tree and Random Forest achieved the highest classification performance, with an Accuracy of 92.38%, Precision of 92.10%, Recall of 99.26%, and F1-Score of 95.54%. K-Means produced a Silhouette Score of 0.4431, indicating a reasonably good clustering structure. The main contribution of this research is a comparative evaluation of five complementary classification approaches on PIP eligibility data from a senior high school, combined with K-Means-based socioeconomic pattern analysis. The results show that tree-based models provide strong decision-support performance, while clustering offers complementary insights into student characteristics. This framework can support more consistent, transparent, and data-driven PIP recipient selection.

Keywords: Program Indonesia Pintar (PIP); Machine Learning; Classification; Clustering; Educational Data Mining

1. INTRODUCTION

Education is a fundamental right of every citizen and plays a vital role in improving the quality of human resources and supporting national development. However, economic inequality remains one of the primary barriers preventing many students from accessing quality education. Students from low-income families often face financial difficulties that increase the risk of school dropout and reduce their opportunities to complete formal education. To address this issue, the Government of Indonesia introduced the Program Indonesia Pintar (PIP)**,** an educational financial assistance program designed to support students from economically disadvantaged families and ensure their continued participation in the education system [1]. The Program Indonesia Pintar (PIP) was officially launched in 2014 as a national strategic program aimed at expanding access to education. The program has continued to be implemented and expanded to improve educational equity and reduce the risk of students discontinuing their education due to financial constraints [2]. Through this initiative, the government seeks to provide eligible students with greater opportunities to continue their education regardless of their socioeconomic background.

The importance of education is explicitly recognized in the Constitution of the Republic of Indonesia (UUD 1945), which places the advancement of national education among the fundamental objectives of the state. Furthermore, Article 3 of Law No. 20 of 2003 concerning the National Education System emphasizes the development of learners into individuals who are knowledgeable, creative, independent, responsible, and capable of contributing positively to society. Ensuring equal access to education is therefore essential for promoting educational equity and supporting the development of high-quality human resources for sustainable national development [3].

Although the Program Indonesia Pintar has significantly contributed to improving educational access, determining eligible beneficiaries remains a challenging task. The selection process requires evaluating multiple socioeconomic indicators, including household income, parental education, parental occupation, ownership of social assistance cards, and other supporting criteria. In many educational institutions, these assessments are still conducted manually, making the process time-consuming and potentially susceptible to inconsistencies, subjective judgment, and human error. As the number of applicants continues to increase, manual evaluation becomes increasingly inefficient and may affect the fairness and transparency of scholarship distribution.

Copyright © 2025 The Author(s). Published by *ADA Research Center*. Page 34
This work is licensed under a Creative Commons Attribution 4.0 International License.

International Journal of Informatics and Data Science
ISSN 3026-7315 (media online)
Volume 3, No 1, December 2025, Page: 34−43
DOI: doi.org/10.64366/ijids.v3i1.571
https://journals.adaresearch.or.id/ijids/index

Recent advances in information technology, particularly in the field of data mining, provide promising approaches for addressing these challenges. Data mining involves the systematic discovery of meaningful patterns, relationships, and useful knowledge from datasets through computational and analytical techniques. Its ability to transform raw data into actionable information has contributed to its application across various domains, including business, healthcare, government, and education. In educational settings, data mining has been applied to analyze student-related data, identify patterns, and support predictive and decision-making tasks, including student performance prediction and other educational outcomes [4].

As part of the government's educational intervention, the Program Indonesia Pintar (PIP) provides financial assistance to students from poor and vulnerable households to support their continued participation in education. The eligibility and beneficiary selection process involves the integration of several national education and social welfare data sources, including the Basic Education Data (Dapodik), the Integrated Social Welfare Data (DTKS), and the Indonesia Smart Card (KIP). The distribution process follows technical guidelines established by the government to support more accurate and accountable beneficiary selection [5]. In this context, knowledge discovery and data mining provide relevant analytical approaches for extracting meaningful patterns from complex datasets and supporting data-driven decision-making [6].

2. RESEARCH METHODOLOGY

2.1 Research Stages

This study aims to develop a machine learning-based approach for predicting student eligibility for the Program Indonesia Pintar (PIP) scholarship. The proposed methodology was designed to evaluate the effectiveness of several classification algorithms in identifying eligible recipients while employing clustering analysis to explore hidden patterns within the dataset. The research was conducted through a series of systematic stages, including data collection, data preprocessing, algorithm implementation, model evaluation, and results analysis. Each stage was designed to ensure that the developed models were trained using high-quality data and evaluated under a consistent experimental framework. The overall research workflow is illustrated in Figure 1.

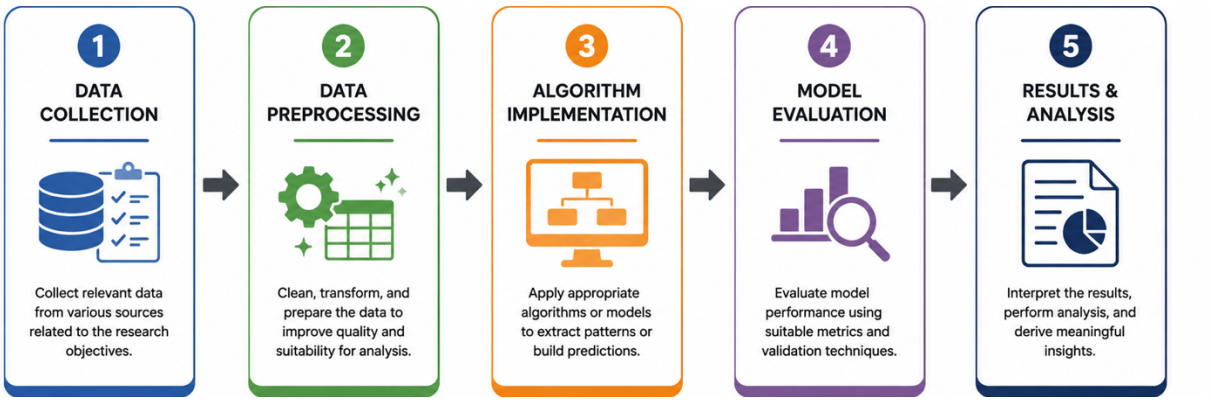


Figure 1. Research Stage

2.2 Data Collection

The dataset used in this study was obtained from SMA Negeri 1 Telaga Biru. It consists of student demographic, educational, and socioeconomic information that is relevant to determining eligibility for the Program Indonesia Pintar (PIP). The collected attributes include transportation type, Family Welfare Card (KPS) status, father's education level, father's occupation, father's income, mother's education level, mother's occupation, mother's income, Indonesia Smart Card (KIP) status, number of siblings, and the PIP eligibility recommendation provided by the school. These variables were selected because they represent the socioeconomic conditions of students and are commonly considered during the PIP selection process. The dataset was used as the primary input for both the classification and clustering models to identify eligible students and analyze similarities among student characteristics.

2.3 Data Preprocessing

Data preprocessing is a critical stage in the machine learning pipeline Effective preprocessing is essential because inconsistencies, missing values, irrelevant attributes, and inappropriate feature representations can negatively affect the reliability of machine learning models [7], [8], [9], as the quality of the input data directly influences the performance and reliability of the predictive models. Raw educational datasets often contain missing values, irrelevant attributes, categorical variables, and inconsistencies that may reduce model accuracy if left untreated. Therefore, a comprehensive preprocessing procedure was conducted to improve data quality and ensure that the dataset was suitable for both classification and clustering analyses. The preprocessing stage consisted of several sequential operations. Initially, the dataset was examined to identify missing values and inconsistencies in the recorded student

Copyright © 2025 The Author(s). Published by *ADA Research Center*.
This work is licensed under a Creative Commons Attribution 4.0 International License.

ISSN 3026-7315 (media online)
Volume 3, No 1, December 2025, Page: 34−43
DOI: doi.org/10.64366/ijids.v3i1.571
https://journals.adaresearch.or.id/ijids/index

information. Missing entries were then handled to preserve data completeness while minimizing information loss. Subsequently, feature selection was performed to remove attributes that contributed little to the prediction process, thereby reducing data dimensionality and improving computational efficiency.

Because most of the predictor variables were categorical, all categorical attributes were transformed into numerical representations using Ordinal Encoding, enabling them to be processed by machine learning algorithms. To address the imbalance between eligible and non-eligible PIP recipients, the dataset was balanced using the undersampling technique, ensuring that the classification models were trained on a more representative distribution of class labels. Furthermore, all numerical features were normalized using Min-Max Normalization, which scales each attribute to the range of 0–1. This normalization step is particularly important for distance-based algorithms, such as K-Nearest Neighbor (KNN) and Support Vector Machine (SVM), because it prevents variables with larger numerical ranges from dominating the distance calculations [10].

Finally, the preprocessed dataset was divided into training and testing subsets to facilitate objective model evaluation. The training dataset was used to construct the predictive models, whereas the testing dataset was reserved exclusively for evaluating model performance on unseen data. These preprocessing procedures ensured that the dataset was clean, consistent, balanced, and appropriately formatted, thereby providing a reliable foundation for the subsequent machine learning experiments. The detailed preprocessing workflow consisted of the following steps:

2.3.1 Missing Value Handling

The dataset was first inspected to identify incomplete records. Missing values were found in the father's education level and mother's education level attributes. To preserve the completeness of the dataset while minimizing information loss, the missing values were replaced using the mode, which represents the most frequently occurring value in each respective attribute.

2.3.2 Feature Selection

Feature selection was conducted to remove attributes considered less relevant to predicting PIP eligibility. Three variables, namely transportation type**,** father's occupation**,** and mother's occupation, were excluded because preliminary analysis indicated that their contribution to prediction performance was relatively limited. Removing irrelevant features helps reduce model complexity, improve computational efficiency, and minimize unnecessary noise during model training.

2.3.3 Categorical Data Encoding

Most predictor variables consisted of categorical values that could not be processed directly by machine learning algorithms. Therefore, all categorical attributes were converted into numerical values using Ordinal Encoding**.** Binary variables such as "Yes" and "No" were encoded as 1 and 0, respectively, while ordinal variables such as education level and income category were encoded according to their natural hierarchical order.

2.3.4 Data Normalization

Feature scaling was performed using the Min-Max Scaler, which transforms all numerical attributes into a common range between 0 and 1. Normalization prevents variables with larger numerical values from dominating the learning process and is particularly important for distance-based algorithms such as K-Nearest Neighbor (KNN) and Support Vector Machine (SVM). Feature scaling is particularly important for distance-based learning methods because differences in feature magnitude can influence distance calculations and consequently affect classification performance [11], [12].

2.3.5 Train–Test Split

After preprocessing, the dataset was divided into 70% training data and 30% testing data. The training set consisted of 764 records, while the testing set contained 328 records. A fixed random_state = 12 was used to ensure reproducibility and consistency of the experimental results.

2.4 Applied Algorithms

This study evaluates the performance of five supervised machine learning algorithms for predicting student eligibility for the Program Indonesia Pintar (PIP). In addition, the K-Means clustering algorithm was employed to explore hidden patterns within the dataset.

2.4.1 Decision Tree

Decision Tree is one of the most widely used classification techniques in data mining. It represents decision-making processes using a hierarchical tree structure, where the root node represents the most influential attribute, internal nodes correspond to decision rules, and leaf nodes represent class labels [7]. The model recursively partitions the dataset based on attribute values until the final classification is obtained. Decision Tree is easy to interpret and closely resembles human decision-making; however, it is prone to overfitting when appropriate pruning techniques are not

Copyright © 2025 The Author(s). Published by *ADA Research Center*. Page 36
This work is licensed under a Creative Commons Attribution 4.0 International License.

ISSN 3026-7315 (media online)
Volume 3, No 1, December 2025, Page: 34−43
DOI: doi.org/10.64366/ijids.v3i1.571
https://journals.adaresearch.or.id/ijids/index

applied. Decision Tree is particularly useful when model interpretability is important because its hierarchical structure can be translated into explicit decision rules [13].

2.4.2 Random Forest

Random Forest is an ensemble learning algorithm that extends the Decision Tree model by constructing multiple decision trees using randomly selected subsets of training samples and predictor variables [8]. The final prediction is determined through majority voting among all trees, resulting in greater stability and improved predictive accuracy. Random Forest is also capable of handling missing values, reducing the influence of outliers, and performing implicit feature selection, making it suitable for complex classification problems involving large datasets.

2.4.3 K-Nearest Neighbor (KNN)

K-Nearest Neighbor (KNN) is a supervised classification algorithm that assigns a class label based on the majority class among the **k** closest training instances [9]. The underlying assumption is that observations with similar characteristics are likely to belong to the same class. Because KNN relies on distance calculations, feature normalization is essential to ensure that all variables contribute equally during classification. In this study, the number of nearest neighbors was set to **k = 5**.

2.4.4 Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised learning algorithm that identifies the optimal hyperplane for separating data into different classes while maximizing the margin between them [10]. Owing to its strong generalization capability, SVM has been widely applied to classification problems involving both linear and nonlinear decision boundaries. In this study, the Radial Basis Function (RBF) kernel was employed with the regularization parameter C = 1**.**

2.4.5 Logistic Regression

Logistic Regression is a statistical classification model used to estimate the probability of binary outcomes [11]. It models the relationship between predictor variables and the target class using the logistic function, producing probability values between zero and one. Despite its relatively simple mathematical formulation, Logistic Regression serves as a strong baseline model for binary classification problems due to its computational efficiency and interpretability. In this study, L2 regularization with C = 1.0 was applied.

2.4.6 K-Means Clustering

In addition to supervised classification, this study applies the K-Means clustering algorithm to group students based on similarities in their socioeconomic characteristics. Unlike classification, clustering does not require predefined class labels. Instead, it partitions observations into homogeneous clusters by minimizing the distance between each observation and its corresponding cluster centroid. The resulting clusters provide complementary insights into student characteristics and help identify hidden patterns within the dataset.

2.5 Model Evaluation

2.5.1 Classification Evaluation

The classification models were evaluated using four widely adopted performance metrics: Accuracy, Precision, Recall, and F1-Score. Accuracy measures the proportion of correctly classified instances, Precision evaluates the reliability of positive predictions, Recall measures the model's ability to correctly identify eligible PIP recipients, and F1-Score provides a balanced assessment by combining Precision and Recall. Machine learning-based classification has been widely applied in educational data mining to predict student-related outcomes and support data-driven educational decision-making [14]. The comparative evaluation involved Decision Tree, Random Forest, K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Logistic Regression, and Naïve Bayes. The experimental results showed that Decision Tree and Random Forest achieved the highest performance, obtaining an Accuracy of 92.38%, Precision of 92.10%, Recall of 99.26%, and an F1-Score of 95.54%. The exceptionally high Recall indicates that both algorithms successfully identified nearly all students who were eligible for the PIP program. Furthermore, the high F1-Score demonstrates an excellent balance between Precision and Recall, indicating reliable and consistent classification performance.

K-Nearest Neighbor (KNN) achieved an Accuracy of 90.55% with an F1-Score of 94.49%, while Support Vector Machine (SVM) obtained an Accuracy of 90.24% and an F1-Score of 94.37%. Logistic Regression produced an Accuracy of 88.72% with an F1-Score of 93.47%, indicating competitive but slightly lower performance than Decision Tree and Random Forest. In contrast, Naïve Bayes recorded the lowest performance, with an Accuracy of 61.28% and an F1-Score of 69.25%, suggesting that the probabilistic assumptions of the model were less suitable for the characteristics of the PIP dataset. Overall, the experimental results demonstrate that Decision Tree and Random Forest provide the most effective classification performance for predicting PIP recipient eligibility.

Copyright © 2025 The Author(s). Published by *ADA Research Center*. Page 37
This work is licensed under a Creative Commons Attribution 4.0 International License.

International Journal of Informatics and Data Science
ISSN 3026-7315 (media online)
Volume 3, No 1, December 2025, Page: 34−43
DOI: doi.org/10.64366/ijids.v3i1.571
https://journals.adaresearch.or.id/ijids/index

2.5.2 Clustering Evaluation

The clustering analysis was conducted using the K-Means algorithm to group students according to similarities in their socioeconomic characteristics. Unlike classification, clustering identifies natural groupings within the data without relying on predefined class labels [15]. The quality of clustering was evaluated using the Silhouette Score, which measures both cluster cohesion and separation. The experimental results produced a Silhouette Score of 0.4431, indicating a reasonably good clustering structure. Cluster analysis revealed that Cluster 0 consisted of 370 students, all of whom were categorized as eligible PIP recipients. Meanwhile, Cluster 1 contained 676 students, including 524 eligible and 152 non-eligible students. These findings indicate that K-Means successfully grouped most students with similar characteristics into the same cluster. However, the presence of both eligible and non-eligible students within Cluster 1 suggests that the separation between clusters was not entirely distinct.

Overall, the clustering results demonstrate that K-Means is effective for exploratory data analysis and for identifying patterns among students with similar socioeconomic backgrounds. Nevertheless, compared with supervised classification models, clustering is less suitable for making definitive eligibility decisions because it does not utilize predefined class labels [15]. Therefore, K-Means should be considered a complementary analytical tool that supports data exploration, whereas classification algorithms are more appropriate for decision-support systems used to determine PIP recipients..

3. RESULT AND DISCUSSION

3.1 Classification Results

To evaluate the effectiveness of the proposed machine learning models in predicting student eligibility for the Program Indonesia Pintar (PIP), classification experiments were conducted using five supervised learning algorithms: Decision Tree, Random Forest, K-Nearest Neighbor (KNN), Support Vector Machine (SVM), and Logistic Regression. Model performance was assessed using four widely adopted evaluation metrics, namely Accuracy, Precision, Recall, and F1-Score. To provide a detailed evaluation of the best-performing classifier, the confusion matrix of the Decision Tree model is presented in Figure 2. The confusion matrix illustrates the distribution of correctly and incorrectly classified instances, allowing a comprehensive analysis of the model's predictive capability.

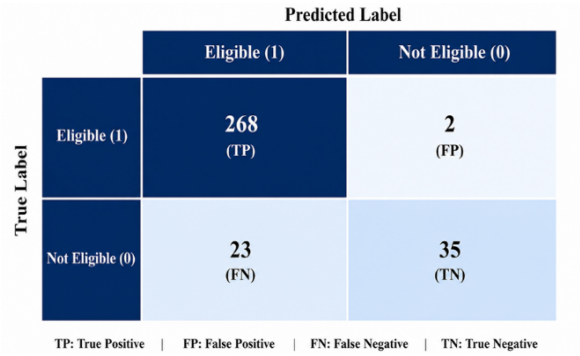


Figure 2. Confusion Matrix of the Decision Tree Classifier

As shown in Figure 2, the Decision Tree model correctly classified 303 out of 328 testing instances, while only 25 instances were misclassified. Specifically, the model correctly identified 268 eligible students (True Positives) and 35 non-eligible students (True Negatives). In contrast, only 2 students were incorrectly classified as eligible (False Positives), while 23 eligible students were incorrectly predicted as non-eligible (False Negatives). These results indicate that the Decision Tree classifier effectively distinguishes between eligible and non-eligible students while maintaining a very low prediction error.

The excellent classification capability of the Decision Tree model is further confirmed by its evaluation metrics. The model achieved an Accuracy of 92.38%**,** indicating that more than nine out of ten students were classified correctly. The Precision of 92.10% demonstrates that the majority of students predicted as eligible were indeed eligible, while the Recall of 99.26% indicates that almost all students who deserved PIP assistance were successfully identified. Furthermore, the F1-Score of 95.54% reflects an excellent balance between Precision and Recall, making the Decision Tree model highly reliable for supporting scholarship recipient selection.

3.2 Comparison of Classification Algorithms

The experimental comparison among the five classification algorithms is presented in **Table 1**.

Table 1. Performance Comparison of Classification Algorithms

Model	Accuracy	Precision	Recall	F1-Score
Decision Tree	92.38%	92.10%	99.26%	95.54%

Copyright © 2025 The Author(s). Published by *ADA Research Center*. Page 38
This work is licensed under a Creative Commons Attribution 4.0 International License.

ISSN 3026-7315 (media online)
Volume 3, No 1, December 2025, Page: 34−43
DOI: doi.org/10.64366/ijids.v3i1.571
https://journals.adaresearch.or.id/ijids/index

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	92.38%	92.10%	99.26%	95.54%
K-Nearest Neighbor	90.55%	90.79%	98.52%	94.49%
Support Vector Machine	90.24%	89.93%	99.26%	94.37%
Logistic Regression	88.72%	89.23%	98.15%	93.47%

Based on Table 1, Decision Tree and Random Forest achieved the highest performance among all evaluated models. Both algorithms produced identical evaluation metrics, with an Accuracy of 92.38%, Precision of 92.10%, Recall of 99.26%, and an F1-Score of 95.54%. These results indicate that tree-based algorithms are highly effective for predicting PIP eligibility because they can capture complex relationships among socioeconomic attributes while maintaining excellent generalization performance. The K-Nearest Neighbor (KNN) algorithm ranked third, achieving an Accuracy of 90.55% and an F1-Score of 94.49%. Although its performance remained competitive, it was slightly lower than that of the tree-based models. This difference can be attributed to the distance-based nature of KNN, which is sensitive to data distribution and local neighborhood structures. Similarly, Support Vector Machine (SVM) demonstrated strong classification performance, with an Accuracy of 90.24% and an F1-Score of 94.37%. The model also achieved a Recall of 99.26%, indicating its excellent ability to identify students eligible for financial assistance. However, its lower Precision suggests that it produced slightly more false positive predictions than Decision Tree and Random Forest.

Among the evaluated models, Logistic Regression obtained the lowest classification performance, with an Accuracy of 88.72% and an F1-Score of 93.47%**.** Despite its lower performance, Logistic Regression still produced satisfactory prediction results and may serve as a computationally efficient baseline model for binary classification problems. Overall, the experimental results demonstrate that Decision Tree and Random Forest are the most suitable algorithms for predicting PIP recipient eligibility. Their superior Accuracy and F1-Score indicate that they provide more reliable and consistent predictions than the other evaluated classification models.

3.3 Clustering Results

In addition to supervised classification, this study applied the K-Means clustering algorithm to explore hidden patterns within the student dataset. Unlike classification, clustering groups observations based solely on similarities in their characteristics without relying on predefined class labels. The optimal number of clusters was determined using the Elbow Method**,** resulting in two clusters**.**

The quality of clustering was evaluated using the Silhouette Score**,** which measures both cluster cohesion and separation. The experiment produced **a** Silhouette Score of 0.4431, indicating a reasonably good clustering structure. This value suggests that students within the same cluster share similar socioeconomic characteristics while maintaining moderate separation from students in other clusters. The distribution of students across clusters is presented in Table 2**.**

Table 2. Distribution of Students in Each Cluster

Cluster	Eligible	Not Eligible	Total
Cluster 0	370	0	370
Cluster 1	524	152	676

As shown in Table 2**,** Cluster 0 consisted of 370 students, all of whom were classified as eligible PIP recipients. This finding indicates that the cluster exhibits highly homogeneous characteristics associated with financial assistance eligibility. Conversely, Cluster 1 contained 676 students, comprising 524 eligible and 152 non-eligible students. The presence of both classes within the same cluster indicates that, although K-Means successfully grouped students with similar characteristics, complete separation between eligible and non-eligible students was not achieved.

These findings suggest that K-Means is effective for exploratory data analysis, particularly for identifying hidden patterns among students with similar socioeconomic conditions. However, compared with supervised classification models, clustering remains less appropriate for making final eligibility decisions because it does not utilize predefined class labels during model development.

Overall, the experimental results indicate that classification algorithms**,** particularly Decision Tree and Random Forest, are more suitable for supporting decision-making in the PIP recipient selection process, whereas **K-**Means clustering provides complementary insights that can assist educational institutions in understanding the distribution and characteristics of prospective beneficiaries. The combination of supervised and unsupervised learning therefore offers a more comprehensive analytical framework for evaluating student eligibility for educational financial assistance.

3.4 Discussion

The experimental results demonstrate that Decision Tree and Random Forest consistently outperformed the other evaluated machine learning algorithms in predicting student eligibility for the Program Indonesia Pintar (PIP). Both models achieved the highest Accuracy (92.38%) and F1-Score (95.54%), indicating their strong capability to classify eligible and non-eligible students based on socioeconomic and educational attributes. The comparable performance

Copyright © 2025 The Author(s). Published by *ADA Research Center*. Page 39
This work is licensed under a Creative Commons Attribution 4.0 International License.

ISSN 3026-7315 (media online)
Volume 3, No 1, December 2025, Page: 34−43
DOI: doi.org/10.64366/ijids.v3i1.571
https://journals.adaresearch.or.id/ijids/index

of these two algorithms suggests that tree-based learning methods are particularly effective for educational datasets containing a mixture of categorical and ordinal variables. This observation is consistent with recent comparative evidence showing that tree-based algorithms can demonstrate strong performance on structured tabular datasets. Their hierarchical decision-making mechanism enables them to capture complex relationships among predictor variables while maintaining high interpretability, making them suitable for decision-support applications in educational institutions [13].

In contrast, K-Nearest Neighbor (KNN) and Support Vector Machine (SVM) also produced competitive performance, although both achieved slightly lower Accuracy than the tree-based models. The lower performance of KNN may be attributed to its dependence on distance calculations, where classification results are highly influenced by the distribution of neighboring observations [16]. Although Min-Max normalization was applied to reduce scale differences among variables, the presence of overlapping socioeconomic characteristics among students may have limited the algorithm's ability to distinguish between eligible and non-eligible recipients. Similarly, SVM achieved a high Recall (99.26%), demonstrating its effectiveness in identifying eligible students, but its lower Precision indicates that it generated more false-positive predictions than Decision Tree and Random Forest.

Among the evaluated classification models, Logistic Regression obtained the lowest Accuracy (88.72%). This outcome may be explained by the assumption of a linear relationship between predictor variables and the target class. Student eligibility for the PIP program is determined by multiple socioeconomic factors that often interact in a nonlinear manner. Consequently, a linear decision boundary is less capable of capturing these complex relationships than tree-based algorithms, resulting in slightly lower predictive performance. The confusion matrix further supports these findings. The Decision Tree classifier correctly classified 303 out of 328 testing instances, with only 25 misclassified observations. In particular, the model correctly identified 268 eligible students while producing only two false-positive predictions. This result is particularly important because minimizing the number of incorrectly classified students contributes to a fairer and more reliable scholarship selection process. The high Recall achieved by the model indicates that nearly all students who deserved financial assistance were successfully identified, thereby reducing the risk of excluding eligible beneficiaries.

The clustering experiment complements the classification results by providing an alternative perspective on the dataset. Unlike supervised classification, K-Means groups students according to similarities in their characteristics without using predefined class labels [17]. The obtained Silhouette Score of 0.4431 indicates a reasonably good clustering structure, suggesting that students within the same cluster share similar socioeconomic profiles. However, the presence of both eligible and non-eligible students in the same cluster demonstrates that clustering alone is insufficient for determining scholarship eligibility. Instead, clustering should be regarded as an exploratory analytical tool that helps educational institutions understand the underlying characteristics of prospective recipients rather than as a standalone decision-making approach [17].

The comparison between supervised and unsupervised learning highlights their complementary roles in educational data analysis. Classification algorithms are more appropriate for predicting eligibility because they learn directly from historical labeled data, whereas clustering provides additional insight into hidden patterns that may not be immediately observable [15], [17]. Integrating these two approaches enables decision-makers to obtain both accurate predictions and a deeper understanding of the socioeconomic distribution of students. The findings of this study are consistent with previous research showing that machine learning and tree-based approaches can provide effective prediction and classification performance for structured educational datasets [14]. Their superior performance can be attributed to their ability to model nonlinear decision boundaries, automatically select informative features, and maintain robustness against noisy data. Furthermore, the results support previous studies demonstrating the use of clustering techniques for exploring patterns and grouping students based on similarities in educational or socioeconomic characteristics [16], [17]. From a practical perspective, the proposed approach can assist schools in improving the transparency, consistency, and efficiency of the PIP recipient selection process. By utilizing machine learning models, educational institutions can reduce manual evaluation, minimize subjective bias, and accelerate decision-making while ensuring that financial assistance is allocated to students who meet the established eligibility criteria.

Despite these promising results, several limitations should be acknowledged. First, the dataset was collected from a single educational institution, which may limit the generalizability of the developed models. Second, the study considered only socioeconomic and demographic attributes available in the school database. The selection of relevant attributes is important because feature selection can influence the predictive performance of machine learning models in student-related prediction tasks [18]. Incorporating additional variables, such as academic performance, attendance records, or household dependency ratios, may further improve prediction accuracy. Future studies may also compare different class-imbalance handling techniques, including alternative resampling strategies, to determine their effects on model performance [19] - [20]. Finally, this research focused on conventional machine learning algorithms. Future studies may investigate ensemble learning strategies, gradient boosting methods, or deep learning models and evaluate their performance using larger and more diverse datasets collected from multiple schools to improve the robustness and applicability of the proposed framework.

Recent studies have increasingly applied machine learning to the classification of PIP beneficiaries, but most have focused on a single algorithm or a limited comparison of methods. Darussalam et al. [21] applied Naive Bayes to PIP recipient classification at SDN 2 Purwawinangun and reported an accuracy of 96.47%, using socioeconomic

Copyright © 2025 The Author(s). Published by *ADA Research Center*. Page 40
This work is licensed under a Creative Commons Attribution 4.0 International License.

ISSN 3026-7315 (media online)
Volume 3, No 1, December 2025, Page: 34−43
DOI: doi.org/10.64366/ijids.v3i1.571
https://journals.adaresearch.or.id/ijids/index

and academic attributes. Putra et al. [22] compared K-Nearest Neighbor (KNN), Naive Bayes Classifier (NBC), and C4.5 for PIP recipient classification at SMKN 4 Pekanbaru; KNN achieved the best accuracy of 84.20%, while C4.5 provided a simpler decision model. Suciko et al. [23] also used Naive Bayes to classify prospective PIP recipients using parental income, number of siblings, KIP, and KPS status, obtaining an accuracy of 85.9% and an AUC of 0.973. These studies demonstrate the feasibility of machine learning for PIP selection, but they remain centered on individual algorithms or small algorithm sets.

Other recent studies reinforce this limitation. Hidayat and Kurniawan [24] implemented Naive Bayes for PIP eligibility prediction at MTs Al-Hasanah and reported 100% accuracy, precision, recall, and F1-score; however, the authors acknowledged the risk of overfitting because the dataset contained only 105 records. Santuni et al. [25] applied C4.5/Decision Tree to 299 students at SD Negeri 01 Balai Ahad and obtained 84.27% accuracy, while Syamsudin et al. [27] developed an SVM model for PIP eligibility at a public vocational school in Cirebon and achieved 85.19% test accuracy. Although these studies provide valuable evidence, they evaluate isolated algorithm families and are conducted on different school contexts and dataset characteristics, making direct comparison across a broader set of classifiers difficult.

Based on these studies, a research gap remains in the systematic comparative evaluation of diverse classification algorithms for PIP eligibility within the same dataset and experimental setting. Existing PIP studies predominantly investigate Naive Bayes, C4.5/Decision Tree, KNN, or SVM separately, with limited evidence comparing tree-based, distance-based, margin-based, and probabilistic approaches under identical preprocessing and evaluation conditions. In addition, previous studies generally emphasize classification performance and provide limited complementary analysis of socioeconomic structure through unsupervised clustering. Therefore, this study addresses the gap by evaluating Decision Tree, Random Forest, KNN, SVM, and Logistic Regression using the same PIP dataset and preprocessing pipeline, while additionally applying K-Means clustering to identify socioeconomic patterns. The specific contributions are: (1) a controlled comparison of five classification algorithms for PIP eligibility prediction using Accuracy, Precision, Recall, and F1-Score; (2) identification of the most effective model for the studied school context; and (3) complementary K-Means analysis to reveal socioeconomic groupings that cannot be obtained from supervised classification alone.

The contribution of this research is therefore not merely the application of machine learning to PIP data, but the provision of a unified experimental framework that enables the relative strengths of different classification paradigms to be assessed under consistent conditions and complemented by socioeconomic clustering analysis. This approach provides empirical evidence for selecting an appropriate predictive model and offers additional analytical information that can support more transparent and consistent PIP recipient selection.

4. CONCLUSION

This study comparatively evaluated five machine learning classification algorithms, Decision Tree, Random Forest, K-Nearest Neighbor (KNN), Support Vector Machine (SVM), and Logistic Regression, for predicting student eligibility for the Program Indonesia Pintar (PIP), and complemented the supervised analysis with K-Means clustering. The comparative design directly addresses the research gap identified from recent PIP studies, which have generally focused on individual algorithms or limited algorithm combinations. Under the same preprocessing and experimental conditions, Decision Tree and Random Forest achieved the best overall performance, each obtaining an Accuracy of 92.38%, Precision of 92.10%, Recall of 99.26%, and F1-Score of 95.54%. These results indicate that tree-based models are particularly effective for the socioeconomic and categorical characteristics represented in the SMA Negeri 1 Telaga Biru dataset. The K-Means analysis generated two clusters with a Silhouette Score of 0.4431, providing complementary information about socioeconomic similarities among students. The main contribution of this research is a unified comparative framework for evaluating diverse classification paradigms for PIP eligibility while simultaneously using clustering to explore socioeconomic patterns. The findings provide empirical evidence that Decision Tree and Random Forest are the most effective models among the five evaluated methods for this dataset and can therefore serve as decision-support tools for more consistent and transparent PIP selection. K-Means should be used as a complementary exploratory method rather than as the primary eligibility decision model. In practical terms, the proposed framework can help schools reduce reliance on manual assessment and improve the consistency of beneficiary identification. Nevertheless, the study is limited by the use of data from a single school and a restricted set of socioeconomic and demographic attributes. Future research should validate the framework on larger, multi-school datasets and compare additional ensemble, boosting, and deep learning methods, as well as alternative class-imbalance strategies.

5. DECLARATIONS

5.1 Author Contributions

Conceptualization: M.I.A., R.B.M.R., N.M.T., A.R.I., I.S., I., A., D.W.N., A.G.K., and S.S.;
Methodology: M.I.A., R.B.M.R., N.M.T., A.R.I., I.S., and D.W.N.;

Copyright © 2025 The Author(s). Published by *ADA Research Center*. Page 41
This work is licensed under a Creative Commons Attribution 4.0 International License.

ISSN 3026-7315 (media online)
Volume 3, No 1, December 2025, Page: 34−43
DOI: doi.org/10.64366/ijids.v3i1.571
https://journals.adaresearch.or.id/ijids/index

Software: A.R.I., I.S., and D.W.N.;
Validation: M.I.A., R.B.M.R., N.M.T., and D.W.N.;
Formal Analysis: M.I.A., N.M.T., A.R.I., D.W.N., A.G.K., and S.S.;
Investigation: N.M.T., A.R.I., I., and A.;
Resources: N.M.T., A.R.I., I., A., A.G.K., and S.S.;
Data Curation: N.M.T., A.R.I., I.S., and M.I.A.;
Writing – Original Draft Preparation: M.I.A., N.M.T., A.R.I., D.W.N., A.G.K., and S.S.;
Writing – Review and Editing: M.I.A., R.B.M.R., N.M.T., A.R.I., I.S., I., A., D.W.N., A.G.K., and S.S.;
Visualization: A.R.I., I.S., and D.W.N.;
Supervision: M.I.A. and R.B.M.R.;
Project Administration: M.I.A.;
All authors have read and agreed to the published version of the manuscript..

5.2 Data Availability

The data presented in this study are available on reasonable request from the corresponding author. The dataset was obtained from SMA Negeri 1 Telaga Biru and contains student demographic, educational, and socioeconomic information used for the analysis. Due to the confidential nature of student-related information, the dataset is not publicly available. Access to the data may be provided upon reasonable request and subject to applicable institutional and privacy requirements.

5.3 Funding

This research was supported by the Department of Computer Science, Universitas Muhammadiyah Gorontalo, Indonesia. The funding provided institutional support for the implementation of the research and the completion of this study.

5.4 Institutional Review Board Statement

This study did not involve clinical trials, medical interventions, or experiments involving human subjects. The research used educational and socioeconomic data obtained from SMA Negeri 1 Telaga Biru for analytical purposes. Therefore, formal Institutional Review Board approval was not applicable to this study.

5.5 Informed Consent Statement

Informed consent was not applicable because this study involved the secondary analysis of existing educational data and did not involve direct interaction or intervention with participants. The analysis was conducted using the available dataset for research purposes while maintaining the confidentiality of student-related information.

5.6 Declaration of Competing Interest

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

REFERENCES

[1] N. T. H. Trinh, “Higher Education and Its Role for National Development. A Research Agenda with Bibliometric Analysis,” *Interchange*, vol. 54, no. 2, pp. 125–143, 2023, doi: 10.1007/s10780-023-09493-9.
[2] Roudoh, W. S. Salsabilah, A. Wahyudin, and S. Hidayati, “Rethinking the Indonesia Smart Program (PIP): A Policy Review and Solutions for Better Student Outcomes,” *Nidhomul Haq: Jurnal Manajemen Pendidikan Islam*, vol. 9, no. 3, pp. 769–780, 2025, doi: 10.31538/ndhq.v9i3.43.
[3] P. Ninghardjanti, W. Murtini, A. Hindrayani, and K. B. Sangka, “Evaluation of the Smart Indonesia Program as a Policy to Improve Equality in Education,” *Sustainability*, vol. 15, no. 6, p. 5114, 2023, doi: 10.3390/su15065114.
[4] A. Zickafoose, O. Ilesanmi, M. Diaz-Manrique, A. E. Adeyemi, B. Walumbe, R. Strong, G. Wingenbach, M. T. Rodriguez, and K. Dooley, “Barriers and Challenges Affecting Quality Education (Sustainable Development Goal #4) in Sub-Saharan Africa by 2030,” *Sustainability*, vol. 16, no. 7, p. 2657, 2024, doi: 10.3390/su16072657.
[5] C. Liu, E. Fakharizadi, T. Xu, and P. S. Yu, “Recent Advances in Domain-Driven Data Mining,” *International Journal of Data Science and Analytics*, vol. 15, no. 1, pp. 1–7, 2023, doi: 10.1007/s41060-022-00378-1.
[6] X. Shu and Y. Ye, “Knowledge Discovery: Methods from Data Mining and Machine Learning,” *Social Science Research*, vol. 110, p. 102817, 2023, doi: 10.1016/j.ssresearch.2022.102817.
[7] M. Bilal, G. Ali, M. W. Iqbal, M. Anwar, M. S. A. Malik, and R. Abdul Kadir, “Auto-Prep: Efficient and Automated Data Preprocessing Pipeline,” *IEEE Access*, vol. 10, pp. 107764–107784, 2022, doi: 10.1109/ACCESS.2022.3198662.
[8] B. L. Ortiz, V. Gupta, R. Kumar, A. Jalin, X. Cao, C. Ziegenbein, A. Singhal, S. W. Choi, and M. Tewari, “Data Preprocessing Techniques for AI and Machine Learning Readiness: Scoping Review of Wearable Sensor Data in Cancer Care,” *JMIR mHealth and uHealth*, vol. 12, p. e59587, 2024, doi: 10.2196/59587.

Copyright © 2025 The Author(s). Published by *ADA Research Center*.
This work is licensed under a Creative Commons Attribution 4.0 International License.

[9] S. Z. Golazad, A. Mohammadi, A. Rashidi, and M. Ilbeigi, "From Raw to Refined: Data Preprocessing for Construction Machine Learning (ML), Deep Learning (DL), and Reinforcement Learning (RL) Models," *Automation in Construction*, vol. 168, p. 105844, 2024, doi: 10.1016/j.autcon.2024.105844.
[10] P. Koukaras and C. Tjortjis, "Data Preprocessing and Feature Engineering for Data Mining: Techniques, Tools, and Best Practices," *AI*, vol. 6, no. 10, p. 257, 2025, doi: 10.3390/ai6100257.
[11] R. K. Halder, M. N. Uddin, M. A. Uddin, S. Aryal, and A. Khraisat, "Enhancing K-Nearest Neighbor Algorithm: A Comprehensive Review and Performance Analysis of Modifications," *Journal of Big Data*, vol. 11, no. 1, p. 113, 2024, doi: 10.1186/s40537-024-00973-y.
[12] R. Guido, S. Ferrisi, D. Lofaro, and D. Conforti, "An Overview on the Advancements of Support Vector Machine Models in Healthcare Applications: A Review," *Information*, vol. 15, no. 4, p. 235, 2024, doi: 10.3390/info15040235.
[13] S. Uddin and H. Lu, "Confirming the Statistically Significant Superiority of Tree-Based Machine Learning Algorithms over Their Counterparts for Tabular Data," *PLOS ONE*, vol. 19, no. 4, p. e0301541, 2024, doi: 10.1371/journal.pone.0301541.
[14] M. Yağcı, "Educational Data Mining: Prediction of Students' Academic Performance Using Machine Learning Algorithms," *Smart Learning Environments*, vol. 9, no. 1, p. 11, 2022, doi: 10.1186/s40561-022-00192-z.
[15] S. M. Dol and P. M. Jawandhiya, "Classification technique and its combination with clustering and association rule mining in educational data mining—A survey," Engineering Applications of Artificial Intelligence, vol. 122, p. 106071, 2023, doi: 10.1016/j.engappai.2023.106071.
[16] R. Wandri, Y. Arta, A. Hanafiah, and R. Oktaviaani, "Prediction of Student Scholarship Recipients Using the K-Means Algorithm and C4.5," *The Indonesian Journal of Computer Science*, vol. 12, no. 1, 2023, doi: 10.33022/ijcs.v12i1.3145.
[17] S. Andre, N. Suciati, H. Fabroyir, and E. Pardede, "Educational Data Mining Clustering Approach: Case Study of Undergraduate Student Thesis Topic," *IEEE Access*, vol. 11, pp. 130072–130088, 2023, doi: 10.1109/ACCESS.2023.3332818.
[18] S. Rezvani and X. Wang, "A Broad Review on Class Imbalance Learning Techniques," *Applied Soft Computing*, vol. 143, p. 110415, 2023, doi: 10.1016/j.asoc.2023.110415.
[19] M. Imani, A. Beikmohammadi, and H. R. Arabnia, "Comprehensive Analysis of Random Forest and XGBoost Performance with SMOTE, ADASYN, and GNUS Under Varying Imbalance Levels," *Technologies*, vol. 13, no. 3, p. 88, 2025, doi: 10.3390/technologies13030088.
[20] A. Laakel Hemdanou, M. Lamarti Sefian, Y. Achtoun, and I. Tahiri, "Comparative Analysis of Feature Selection and Extraction Methods for Student Performance Prediction Across Different Machine Learning Models," *Computers and Education: Artificial Intelligence*, vol. 7, p. 100301, 2024, doi: 10.1016/j.caeai.2024.100301.
[21] E. Ahmed, "Student Performance Prediction Using Machine Learning Algorithms," *Applied Computational Intelligence and Soft Computing*, 2024, p. 4067721, 2024, doi: 10.1155/2024/406772.
[22] L. N. Darussalam, R. Kurniawan, Y. A. Wijaya, and T. Suprapti, "Algoritma Naive Bayes untuk Meningkatkan Model Klasifikasi Penerima Program Indonesia Pintar di SDN 2 Purwawinangun," Jurnal Informatika dan Teknik Elektro Terapan, vol. 13, no. 1, pp. 1229–1237, 2025, doi: 10.23960/jitet.v13i1.5882.
[23] T. A. Putra, I. Permana, Zarnelly, and Megawati, "Klasifikasi Penerima Bantuan Program Indonesia Pintar (PIP) Pada Siswa SMK Menggunakan Algoritma KNN, NBC dan C4.5," Building of Informatics, Technology and Science (BITS), vol. 6, no. 4, pp. 2131–2138, 2025, doi: 10.47065/bits.v6i4.6395.
[24] A. Suciko, Y. Hendriyani, K. Budayawan, and Syafrijon, "Penerapan Data Mining Untuk Klasifikasi Calon Siswa Penerima Program Indonesia Pintar (PIP) Menggunakan Algoritma Naive Bayes," Jurnal Pendidikan Tambusai, vol. 9, no. 2, pp. 15266–15274, 2025, doi: 10.31004/jptam.v9i2.27930.
[25] Z. Hidayat and R. Kurniawan R., "Penerapan Naive Bayes Classifier untuk Prediksi Kelayakan Penerima Bantuan Program Indonesia Pintar (PIP)," EXPERT: Jurnal Manajemen Sistem Informasi dan Teknologi, vol. 15, no. 1, pp. 22–31, 2025, doi: 10.36448/expert.v15i1.4326.
[26] N. Santuni, D. Irfan, Y. Hendriyani, and Y. Huda, "Klasifikasi Penerima Beasiswa Program Indonesia Pintar (PIP) Menggunakan Algoritma Decision Tree (Studi Kasus: SD Negeri 01 Balai Ahad)," Jurnal Pendidikan Tambusai, vol. 9, no. 1, pp. 2657–2669, 2025, doi: 10.31004/jptam.v9i1.25009.
[27] M. S. Syamsudin, I. Maulana, and Suratno, "Pemodelan Prediktif Kelayakan Beasiswa PIP Menggunakan Support Vector Machine: Studi Kasus pada Sekolah Menengah Kejuruan Negeri di Cirebon," Jurnal Digit: Digital of Information Technology, vol. 15, no. 2, pp. 46–54, 2025, doi: 10.51920/jd.v15i2.444.

Copyright © 2025 The Author(s). Published by *ADA Research Center*. Page 43
This work is licensed under a Creative Commons Attribution 4.0 International License.