Journal of Applied Informatics and Computing (JAIC)
Vol.10, No.4, August 2026, pp. 3753~3760
e-ISSN: 2548-6861

Implementation of a Hybrid Model Using Principal Component Analysis, K-Means, and Naïve Bayes for Tuition Fee Category Prediction

Jessika [1]*, Munirul Ula [2]*, Nurdin [3]*
* Departement of Information Technology, Universitas Malikussaleh, Lhokseumawe, Indonesia
nurdin@unimal.ac.id [3]

Article Info

Article history:

Received 2026-06-03
Revised 2026-07-28
Accepted 2026-08-11

Keyword:

Tuition Fee Categories,
Principal Component Analysis,
K-Means Clustering,
Naïve Bayes Classifier,
Machine Learning.

ABSTRACT

The determination of Tuition Fee Categories in higher education institutions is commonly conducted through manual verification of students' socioeconomic documents, which may lead to subjectivity and inconsistencies in decision-making. This study proposes a hybrid machine learning approach that integrates Principal Component Analysis (PCA), K-Means Clustering, and Naïve Bayes Classifier within a semi-supervised learning framework for student socioeconomic classification based on pseudo-labels generated from clustering results. The dataset used in this study consists of 452 student records with 12 socioeconomic attributes obtained from the New Student Admission system of STAIN Teungku Dirundeng Meulaboh in 2025. Data preprocessing includes attribute selection, categorical transformation using One Hot Encoding, and feature standardization. PCA is applied to reduce dimensionality from 18 features to 12 principal components while retaining 95% of the total variance. The processed data are clustered using K-Means with the optimal number of clusters determined as 8 based on Elbow and Silhouette Score analysis. These clusters are used as pseudo-labels for training the Naïve Bayes classifier. Experimental results show that the proposed model achieves 98.89% training accuracy and 97.80% testing accuracy, with a weighted average F1-score of 0.98. The results indicate that the proposed hybrid approach is effective in capturing underlying socioeconomic patterns and provides a stable classification performance. However, the model is based on pseudo-labels rather than official tuition fee categories. Therefore, further validation using real labeled data is recommended to enhance generalizability and practical applicability.

I. INTRODUCTION

The advancement of information and communication technology (ICT) has significantly transformed data management and decision-making processes in higher education institutions. Universities are increasingly required to optimize data utilization to improve administrative efficiency, transparency, and accountability in academic services, including student financial management [1]. In this context, data-driven approaches and intelligent systems have become essential tools to support evidence-based decision-making in educational institutions.

One of the critical components in higher education financing systems in Indonesia is the determination of the Single Tuition Fee (Uang Kuliah Tunggal or UKT), which aims to align educational costs with students' socioeconomic conditions [2]. However, in practice, the determination process is still widely conducted through manual or semi-manual procedures involving document verification and subjective judgment. This often leads to inconsistencies in classification results, where students with similar socioeconomic backgrounds may receive different UKT categories, raising concerns regarding fairness and transparency.

In addition, the increasing number of new student admissions each year introduces additional complexity in managing socioeconomic data. Many institutions still face challenges in fully digitizing the UKT classification process,

resulting in inefficiencies and potential human errors during data processing. Therefore, an automated and data-driven approach is required to improve consistency, efficiency, and objectivity in the classification process [3].

Recent developments in machine learning have introduced various approaches for analyzing educational data and supporting classification tasks. Machine learning methods are capable of identifying hidden patterns within complex datasets, including students' socioeconomic attributes, to support predictive modelling [4]. However, most existing studies still rely on fully supervised learning approaches that require predefined labels, which are often unavailable or inconsistently defined in real-world educational datasets.

To address this limitation, this study adopts a semi-supervised learning approach, where unlabeled data is first processed using clustering techniques to generate pseudo-labels, which are then used as target variables for classification modeling. This approach allows the system to learn from inherent data structures without relying on manually assigned labels, making it more suitable for educational datasets where labeling is costly and subjective [5].

Despite the increasing adoption of machine learning in educational classification problems, most existing studies focus on single-method approaches. For instance, some studies apply Random Forest or boosting-based models for UKT classification, while clustering techniques such as K-Means are often used independently for grouping students based on socioeconomic similarity [2], [5]. However, these approaches rarely integrate dimensionality reduction, clustering, and classification into a unified hybrid framework.

The research gap in this study lies in the limited integration of Principal Component Analysis (PCA), K-Means clustering, and Naïve Bayes classification within a single semi-supervised learning framework for UKT-related classification tasks. PCA has been widely used to reduce dimensionality and eliminate redundant features, thereby improving computational efficiency and data representation quality [6]. Meanwhile, K-Means clustering has proven effective in grouping data based on similarity patterns across various educational datasets [7].

Based on these considerations, this study proposes a hybrid model integrating PCA for dimensionality reduction, K-Means for pseudo-label generation, and Naïve Bayes for classification. Naïve Bayes is selected due to its simplicity and strong performance in probabilistic classification tasks, particularly in small to medium-sized datasets [8]. The integration of these methods is expected to improve data representation, clustering quality, and classification consistency compared to single-method approaches.

This study utilizes student data obtained from the New Student Admission System (Penerimaan Mahasiswa Baru PMB) of STAIN Teungku Dirundeng Meulaboh for the year 2025. The dataset represents secondary data that has been systematically documented and structured, making it suitable for computational processing in machine learning-based analysis.

II. METHODOLOGY

A. Dataset Description

This study utilizes a student dataset obtained from the New Student Admission System (Penerimaan Mahasiswa Baru PMB) of STAIN Teungku Dirundeng Meulaboh for the year 2025. The dataset is secondary data that has been systematically documented and structured, making it suitable for computational processing in machine learning-based modeling.

The dataset consists of 452 student records with 12 attributes representing socioeconomic and academic characteristics. These attributes include student identification number, gender, age, gender, number of siblings, number of siblings currently enrolled in higher education, number of siblings still attending school, admission wave, parents' occupation, residential status, achievements, and ownership of the Indonesia Smart Card (Kartu Indonesia Pintar-KIP). These variables were selected due to their relevance in representing students' socioeconomic conditions, which are used as the basis for forming data-driven groupings in this study.

B. Research Workflow

This study adopts a semi-supervised learning framework, where unlabeled data is transformed into labeled data through clustering-based pseudo-labeling. The overall workflow consists of several structured stages: data preprocessing, dimensionality reduction, clustering, pseudo-label validation, classification, and evaluation. Initially, problem identification is conducted to understand the limitations of the current UKT classification process, which is still prone to subjectivity and inconsistency. After data collection, preprocessing is performed, including handling missing values, normalization of numerical features, and encoding categorical variables to ensure compatibility with machine learning algorithms.

Following preprocessing, Principal Component Analysis (PCA) is applied to reduce data dimensionality. In this study, PCA retains 95% of the cumulative explained variance, which is selected to ensure that most of the original information is preserved while reducing noise and computational complexity. The next stage involves K-Means clustering, which groups students based on similarity in socioeconomic characteristics. The optimal number of clusters (k) is determined using a combination of the Elbow Method and Silhouette Score analysis, ensuring both compactness and separation of clusters.

Before cluster results are used as pseudo-labels, cluster validation is performed using Silhouette Score and visual inspection of PCA-based scatter plots. This step ensures that the generated clusters are meaningful and sufficiently separable to be used as training labels. Once validated, cluster labels are assigned as pseudo-labels, which are then used as target variables in the classification stage. The Naïve Bayes algorithm is applied to learn patterns from the labeled data and perform classification.

Model evaluation is conducted using accuracy, precision, recall, F1-score, and confusion matrix analysis. Additionally, k-fold cross-validation is applied to ensure model stability and generalization performance. If performance is suboptimal, iterative tuning is performed by adjusting PCA components and the number of clusters (k). The final stage involves interpretation of results to evaluate the effectiveness of the proposed hybrid model in supporting structured and data-driven decision-making in educational financial classification systems.

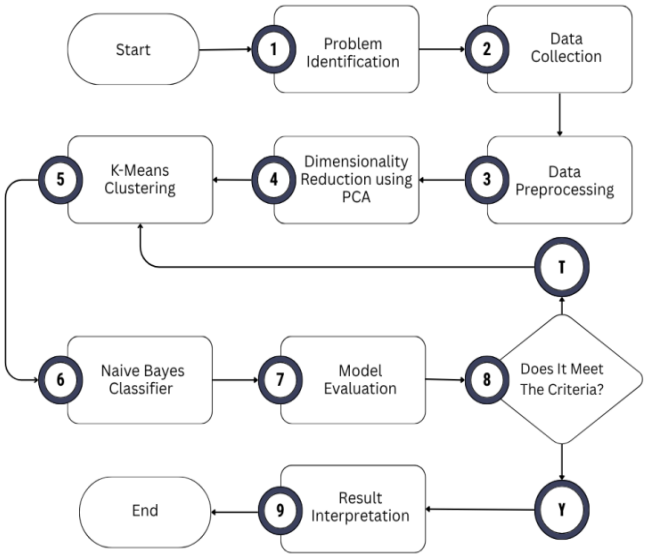

Figure 1. Research Framework

C. Tuition Fee Category

Tuition Fee Category (Uang Kuliah Tunggal or UKT) is a higher education financing policy in Indonesia aimed at promoting fairness and equal access to education. Based on the Regulation of the Minister of Education, Culture, Research, and Technology Number 2 of 2024 concerning the Standard Unit Cost of Higher Education Operational Expenses, students are charged a single tuition fee component per semester adjusted to their family's economic capability [3].

D. Educational Data Mining (EDM)

Educational Data Mining (EDM) is an interdisciplinary field that combines data mining, machine learning, and statistical analysis techniques to discover hidden patterns in educational data. The primary objective of EDM is to generate meaningful information from academic and non-academic data to support decision-making processes in educational institutions. Through this approach, institutions can understand student learning behavior, predict academic performance, and adapt learning strategies based on individual characteristics. EDM also serves as a foundation for the development of adaptive learning systems that integrate technology and analytics within the context of digital transformation in higher education [9].

E. Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a dimensionality reduction technique used to transform high-dimensional data into a lower-dimensional representation while preserving most of the original variance. Previous research demonstrates that PCA is effective in simplifying complex educational datasets while maintaining essential information structures [6].

In this study, PCA is applied to reduce redundancy and multicollinearity among attributes. The selection of 95% explained variance threshold is used to balance information retention and dimensionality reduction efficiency, ensuring that the transformed dataset still represents the original data structure effectively. The integration of PCA with clustering and classification improves data quality by reducing noise and enhancing computational efficiency, making it suitable for hybrid modeling approaches in educational data analysis [10].

F. K-Means Clustering

K-Means is one of the most widely used unsupervised learning algorithms for clustering data based on similarities among data objects. The algorithm partitions a dataset into a predefined number of clusters (k), where each data point is assigned to the nearest centroid. Its primary objective is to minimize intra-cluster variation while maximizing inter-cluster separation, thereby producing representative and meaningful data groupings. In the field of education, K-Means has been widely applied to identify hidden patterns in student-related data, such as academic performance, attendance, socioeconomic conditions, and other educational data grouping tasks [7], [11].

Mathematically, the K-Means algorithm aims to minimize an objective function known as the Within Cluster Sum of Squares (WCSS), which is formulated as follows:

$$WCSS = \sum_{i=1}^{k} \sum_{x \in C_i} \left| |x - \mu_i| \right|^2 \quad (1)$$

Where x represents a data point, μ_i denotes the centroid of cluster (C_i), and $\left| |x - \mu_i| \right|$ represents the Euclidean distance between the data point x and the centroid μ_i. A lower WCSS value indicates a higher degree of similarity among members within the same cluster.

In this study, clustering is performed after the dimensionality reduction stage using Principal Component Analysis (PCA). Applying PCA prior to K-Means helps reduce attribute redundancy, eliminate noise, and improve cluster separability. The optimal number of clusters (k) is determined using a combination of the Elbow Method and Silhouette Score analysis. The Elbow Method is employed to identify the point at which the decrease in WCSS begins to level off, while the Silhouette Score is used to evaluate cluster compactness and separation.

Once the optimal value of k is determined, the K-Means algorithm is executed through an iterative process that begins with centroid initialization. Each data point is assigned to the nearest centroid based on Euclidean distance, after which centroid positions are updated using the mean values of cluster members. This process is repeated until convergence is achieved, indicated by no significant changes in centroid positions or cluster memberships [12], [13].

The resulting clusters are not directly used in the classification stage. Instead, cluster validation is first conducted using the Silhouette Score and PCA-based cluster visualization to ensure that the generated clusters are sufficiently compact and well separated. This validation step is essential to confirm that the clusters adequately represent the underlying socioeconomic patterns of students. After validation, the cluster labels are assigned as pseudo-labels and subsequently used as target classes in the Naïve Bayes classification stage within a semi-supervised learning framework.

G. Naïve Bayes Classifier

The Naïve Bayes algorithm is a probabilistic classification method based on Bayes' Theorem, which is widely used in machine learning, particularly for classifying data with numerical and categorical attributes. The method works by calculating the probability of a data instance belonging to a particular class based on its attribute values [6], [14]. Mathematically, Naïve Bayes is formulated as follows:

$$P(H|X) = \frac{P(X|H).P(H)}{P(X)} \quad (2)$$

Where P(H|X) represents the probability that hypothesis H is true given data X (posterior probability), P(X|H) denotes the probability of observing data X if hypothesis H is true (likelihood), P(H) represents the prior probability of hypothesis H, and P(X) denotes the probability of observing data X (evidence) [15], [16].

In this study, Naïve Bayes is employed as the classification model within a semi-supervised learning framework. Unlike conventional supervised learning approaches that require manually labeled data, the classifier is trained using pseudo-labels generated from the validated K-Means clustering process. This approach enables the model to learn the underlying patterns of students' socioeconomic characteristics without relying on predefined UKT labels.

One of the main advantages of Naïve Bayes is its computational efficiency and its ability to perform well on datasets with heterogeneous attributes and relatively limited sample sizes. Furthermore, the algorithm has demonstrated competitive performance in educational data classification tasks due to its simplicity and robustness [17].

The performance of the classification model is evaluated using accuracy, precision, recall, and F1-score. These metrics provide a comprehensive assessment of classification performance across different aspects of prediction quality. The formulations of these evaluation metrics are defined as follows:

The algorithm also performs well with large numbers of attributes and imbalanced class distributions when evaluated using metrics such as precision, recall, and F1-score [17]. The formulations of these evaluation metrics are as follows:

$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})} \quad (3)$$

$$\text{Precision} = \frac{\text{TP}}{(\text{TP} + \text{FP})} \quad (4)$$

$$\text{Recall} = \frac{\text{TP}}{(\text{TP} + \text{FN})} \quad (5)$$

$$\text{F1} - \text{Score} = \frac{2 \times (\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})} \quad (6)$$

Where TP (True Positive) represents the number of positive instances correctly classified, TN (True Negative) represents the number of negative instances correctly classified, FP (False Positive) denotes the number of negative instances incorrectly classified as positive, and FN (False Negative) represents the number of positive instances incorrectly classified as negative [5].

To further evaluate model robustness, this study employs k-fold cross-validation, which measures the stability and generalization capability of the classifier across multiple data partitions. In addition, a confusion matrix is analyzed to identify classification errors and examine the distribution of correctly and incorrectly predicted classes. The combination of these evaluation techniques provides a more comprehensive assessment of the effectiveness of the proposed hybrid model.

H. Hybrid Model

The proposed model integrates Principal Component Analysis (PCA), K-Means clustering, and Naïve Bayes classification within a semi-supervised learning framework. This hybrid approach combines dimensionality reduction, unsupervised learning, and supervised learning techniques to improve data representation, clustering quality, and classification performance.

The process begins with PCA, which transforms the original dataset into a lower-dimensional feature space while retaining the majority of the information contained in the original variables. By reducing redundancy, multicollinearity, and noise, PCA produces a more compact and representative dataset that facilitates subsequent analysis [18], [19]

The transformed data are then processed using the K-Means algorithm to identify groups of students with similar socioeconomic characteristics. Unlike conventional supervised learning approaches that require predefined class labels, K-Means generates clusters directly from the data structure. The resulting cluster labels are subsequently validated and assigned as pseudo-labels, representing latent socioeconomic categories derived from the dataset.

After the pseudo-labeling process, Naïve Bayes is employed as the classification model to learn the patterns associated with each cluster and predict the corresponding class of new observations. Through this mechanism, the proposed framework combines the strengths of unsupervised and supervised learning, allowing classification to be performed even when manually labeled data are unavailable.

The overall workflow of the proposed hybrid model consists of six stages: data preprocessing, dimensionality reduction using PCA, clustering using K-Means, cluster validation, pseudo-label generation, and classification using Naïve Bayes. The model is then evaluated using accuracy, precision, recall, F1-score, confusion matrix analysis, and k-fold cross-validation to assess both predictive performance and model stability.

The integration of PCA, K-Means, and Naïve Bayes is expected to provide several advantages over single-method approaches. PCA improves data quality by reducing irrelevant information, K-Means uncovers hidden structures within the data and generates meaningful pseudo-labels, while Naïve Bayes performs efficient probabilistic classification. Consequently, the proposed hybrid model is expected to produce more consistent and robust classification results for UKT-related socioeconomic grouping tasks [20], [21].

III. Result And Discussion

A. Results of Problem Identification

The determination of Tuition Fee Categories (UKT) at STAIN Teungku Dirundeng Meulaboh is currently conducted through manual document verification based on students' socioeconomic information, such as parents' occupation, family dependents, residential status, and ownership of social assistance indicators. This process relies heavily on manual assessment, which may lead to inconsistencies and variations in decisions among students with similar socioeconomic conditions.

Furthermore, the increasing number of student admissions each year has made the evaluation process more complex and time-consuming. The absence of a structured data-driven mechanism may reduce efficiency and increase the risk of subjective judgment during the classification process. To address these challenges, this study proposes a hybrid semi-supervised learning approach that integrates Principal Component Analysis (PCA), K-Means Clustering, and Naïve Bayes. PCA is employed to reduce data dimensionality, K-Means is used to identify socioeconomic patterns and generate pseudo-labels, and Naïve Bayes is applied to classify the generated labels. This approach is expected to support a more objective, consistent, and data-driven student socioeconomic classification process.

B. Results of Data Collection

The dataset used in this study consists of 452 student records with 12 attributes. Each record represents an individual student, while each attribute describes socioeconomic and demographic characteristics related to the student and their family. The variables include region, gender, age, parents' occupation, number of siblings, number of siblings currently enrolled in higher education, number of siblings still attending school, admission wave, residential status, achievements, ownership of the Indonesia Smart Card (KIP), and other relevant attributes used in the UKT classification process.

The collected dataset represents the raw data obtained from the New Student Admission System before preprocessing. During the preprocessing stage, attributes that do not contribute to the modeling process are removed, while the remaining attributes are transformed and prepared for further analysis. The initial appearance of the dataset is shown in Figure 2, which illustrates the data structure before preprocessing and feature transformation are performed.

Shape: (452, 12)

	Id	Gender	Usia	Wilayah	Jlh_Saudara	Jlh_Saudara_Kuliah	Jlh_Saudara_Sekolah	Gel_Masuk	Pekerjaan_Ortu	Status_TempatTinggal	Sertifikat_Prestasi	KIP
0	1761	F	18	Aceh Singkil	6	0	1	prestasi	tidak tetap	rumah sendiri	0	ada
1	1762	F	18	Aceh Barat	1	0	0	prestasi	tidak tetap	rumah sendiri	0	ada
2	1768	F	18	Nagan Raya	2	0	1	prestasi	tidak tetap	rumah sendiri	0	ada
3	1770	F	17	Aceh Barat	2	0	1	prestasi	tidak tetap	rumah sendiri	1	ada
4	1773	F	17	Nagan Raya	2	0	1	prestasi	tidak tetap	rumah sendiri	1	ada

Figure 2. Initial Dataset Structure

C. Results of Preprocessing Data

Preprocessing was conducted to prepare the dataset for machine learning analysis and to ensure data quality, consistency, and suitability for subsequent modeling stages. An initial data inspection was performed to identify missing values and attribute relevance. The results showed that no missing values were found; therefore, no imputation was required. The regional attribute was removed due to its limited relevance to students' socioeconomic conditions, while the student ID was temporarily excluded from the modeling process as it functions only as an identifier and not a predictive feature.

Categorical variables were then transformed into numerical form using One Hot Encoding (OHE) to enable processing by machine learning algorithms while avoiding ordinal interpretation of categories. This transformation increased the number of features from 10 to 18. An example of the results is shown in Figure 3.

Usia	Jlh_Saudara	Jlh_Saudara_Kuliah	Jlh_Saudara_Sekolah	Sertifikat_Prestasi	Gender_F	Gender_M	Gel_Masuk_mandiri	Gel_Masuk_prestasi	Gel_Masuk_ujian bersama
18	6	0	1	0	1	0	0	1	0
18	1	0	0	0	1	0	0	1	0
18	2	0	1	0	1	0	0	1	0
17	2	0	1	1	1	0	0	1	0
17	2	0	1	1	1	0	0	1	0

Figure 3. OHE Results

Finally, feature standardization was applied using StandardScaler to ensure all variables were on a comparable scale. This step is essential because PCA and K-Means are sensitive to differences in feature magnitude. The standardized data are shown in Figure 4.

Usia	Jlh_Saudara	Jlh_Saudara_Kuliah	Jlh_Saudara_Sekolah	Sertifikat_Prestasi	Gender_F	Gender_M	Gel_Masuk_mandiri	Gel_Masuk_prestasi	Gel_Masuk_ujian bersama
-0.120520	2.000538	-0.422407	-0.024645	-0.743818	0.597794	-0.597794	-0.398716	0.754615	-0.539841
-0.120520	-1.556674	-0.422407	-1.037348	-0.743818	0.597794	-0.597794	-0.398716	0.754615	-0.539841
-0.120520	-0.845231	-0.422407	-0.024645	-0.743818	0.597794	-0.597794	-0.398716	0.754615	-0.539841
-0.999153	-0.845231	-0.422407	-0.024645	1.344416	0.597794	-0.597794	-0.398716	0.754615	-0.539841
-0.999153	-0.845231	-0.422407	-0.024645	1.344416	0.597794	-0.597794	-0.398716	0.754615	-0.539841

Figure 4. Data Standardization Results Using StandarScaler

D. Result of Dimensionality Reduction using PCA

In this study, Principal Component Analysis (PCA) was applied after the One Hot Encoding and standardization processes to reduce feature dimensionality and minimize redundancy among variables. PCA was configured to retain 95% of the cumulative explained variance to ensure that most of the information in the dataset was preserved while reducing computational complexity.

Based on the results, the original 18 features were transformed into 12 principal components. This indicates that several original variables contained correlated information that could be effectively represented in a lower-dimensional space without significant information loss. The explained variance ratio illustrates the contribution of each principal component to the total variance in the dataset, as shown in Figure 5.

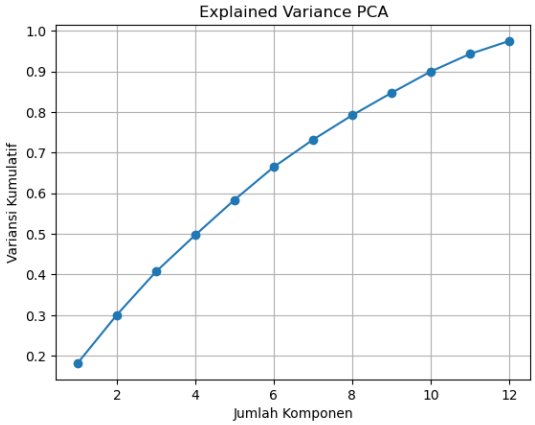

Figure 5. Explained Variance Ratio

The results show that the first principal component (PC1) contributes approximately 18.19% of the total variance, making it the most influential component in the transformed dataset. PC1 primarily represents the dominant variation in students' socioeconomic characteristics, indicating that a significant portion of the data structure can be explained by a small number of components, while the remaining components capture more specific variations with lower contribution.

E. Result of K-Means Clustering

At this stage, K-Means Clustering was applied to group students based on similarities in their socioeconomic characteristics. The clustering process was conducted on the PCA-transformed dataset to improve data separability and reduce feature redundancy, ensuring that clustering is performed on the most informative components.

The optimal number of clusters was determined using the Elbow Method and Silhouette Score analysis. The results indicate that k = 8 provides the best balance between cluster cohesion and separation, supported by a significant reduction in WCSS and the highest Silhouette Score obtained at this value. The evaluation results are presented in Figure 6.

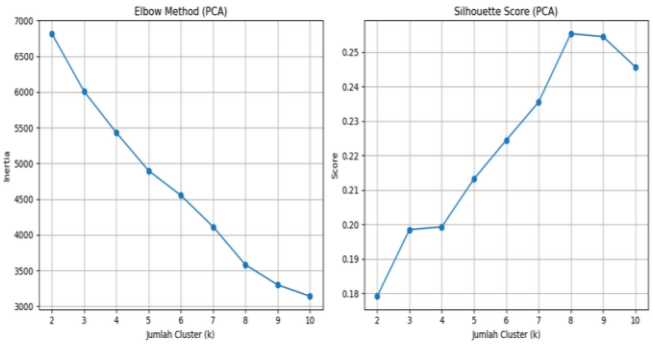

Figure 6. Elbow Method and Silhouette Score Graphs

The K-Means algorithm was then applied using the optimal cluster number, resulting in eight distinct student groups. The distribution of these clusters is illustrated in Figure 7.

Distribusi Kelas:

	Jumlah	Persentase (%)
0	76	16.81
1	159	35.18
2	31	6.86
3	49	10.84
4	19	4.20
5	38	8.41
6	77	17.04
7	3	0.66

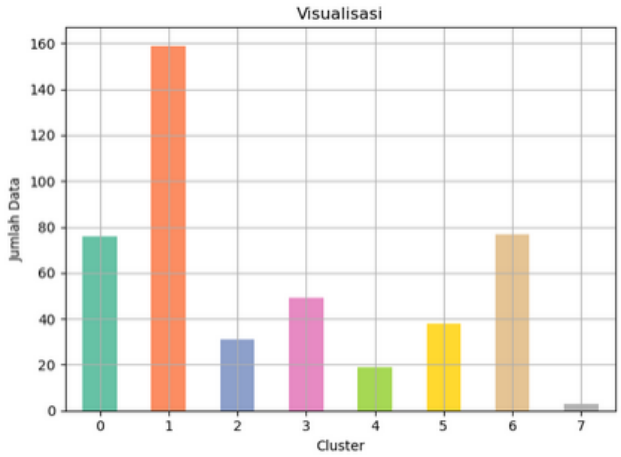

Figure 7. Visualization of K-Means Data Distribution

Based on the visualization results, the data points are distributed across all clusters with varying densities. Some clusters contain a relatively higher number of students, indicating dominant socioeconomic patterns within the dataset, while others are more sparsely populated. In addition, partial overlap between clusters can be observed, suggesting similarities in certain socioeconomic characteristics among different student groups. These clusters are subsequently used as pseudo-labels in the classification stage, where each cluster represents a latent socioeconomic category derived from the underlying data structure.

F. Result of Naïve Bayes Classifier

The Naïve Bayes classifier was employed to predict student groups based on features obtained from Principal Component Analysis (PCA) and pseudo-labels generated through K-Means clustering. This approach follows a semi-supervised learning framework, where the model is trained using labels derived from unsupervised clustering rather than actual ground truth data.

The dataset was split into training and testing sets using an 80:20 ratio to evaluate the model's generalization performance on unseen data. This partitioning ensures that the evaluation reflects the model's ability to classify new instances based on learned patterns from pseudo-labeled data. Model performance was assessed using accuracy, precision, recall, and F1-score metrics to evaluate classification effectiveness. In addition, a confusion matrix was utilized to analyze misclassification patterns across clusters and to identify areas where class overlap occurs. These evaluation results also serve as an indication of potential overfitting and model stability within the proposed hybrid framework.

G. Result of Model Evaluation

The model evaluation was conducted within a semi-supervised learning framework, where classification models were trained using pseudo-labels generated from K-Means clustering after dimensionality reduction using PCA. The evaluation was performed to assess model performance in terms of accuracy, precision, recall, F1-score, generalization ability, and stability across different data splits. The comparison of Naïve Bayes and Logistic Regression performance is presented in Figure 8.

	Model	Accuracy Train	Accuracy Test
0	Naïve Bayes	0.98892	0.978022
1	Logistic Regression	1.00000	1.000000

Figure 8. Comparison of Model Accuracy

Based on the experimental results, Naïve Bayes achieved an accuracy of 98.89% on the training set and 97.80% on the testing set. Logistic Regression achieved perfect performance with 100% accuracy on both training and testing sets. The small difference between training and testing performance in Naïve Bayes indicates strong generalization capability. Meanwhile, Logistic Regression shows consistently perfect performance, which may indicate a high degree of separability in the transformed feature space resulting from PCA, although such results should still be interpreted with caution in terms of potential overfitting or data structure simplicity.

To further assess model robustness, a 5-fold cross-validation was conducted, as shown in Figure 9.

	Model	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean CV
0	Naïve Bayes	0.978022	1.000000	0.966667	0.966667	0.944444	0.971160
1	Logistic Regression	1.000000	1.000000	0.988889	0.977778	0.988889	0.991111

Figure 9. Cross Validation Results

The cross-validation results show that Logistic Regression achieved the highest average accuracy of 99.11%, followed by Naïve Bayes with 97.12%. Although Logistic Regression demonstrates slightly better average performance, Naïve Bayes exhibits more stable performance across different folds, indicating better consistency under varying data splits. The classification report further shows that Naïve Bayes achieves balanced performance across all classes, with precision, recall, and F1-score values ranging from 0.88 to 1.00. Logistic Regression achieves perfect scores across all evaluation metrics, indicating strong classification capability on the pseudo-labeled dataset.

Overall, Naïve Bayes provides a more balanced trade-off between accuracy and stability, while Logistic Regression offers higher but less conservative performance. Considering both stability and interpretability, Naïve Bayes is selected as the most reliable model within the proposed semi-supervised UKT classification framework.

IV. CONCLUSION

This study proposed a hybrid machine learning framework that integrates Principal Component Analysis (PCA), K-Means Clustering, and Naïve Bayes Classification within a semi-supervised learning approach for student socioeconomic classification based on pseudo-labels derived from clustering results. The preprocessing stage successfully prepared the dataset through cleaning, encoding, and standardization. PCA reduced the original 18 features into 12 principal components while preserving 95% of the total variance, thereby reducing redundancy and improving computational efficiency.

The K-Means clustering process identified eight optimal clusters based on socioeconomic similarities among students. These clusters were subsequently used as pseudo-labels for the classification stage. The Naïve Bayes model achieved a training accuracy of 98.89% and a testing accuracy of 97.80%, indicating strong predictive performance and good generalization capability.

Despite the high performance results, this study has several limitations. The dataset is limited to 452 student records from a single institution, which may restrict the generalizability of the model to other contexts. In addition, the classification is based on pseudo-labels generated from clustering rather than actual ground truth UKT categories.

Future research is recommended to validate the proposed model using larger and more diverse datasets from multiple institutions and to explore additional hybrid or ensemble approaches to further improve classification robustness and interpretability.

REFERENCES

[1] G. Al-Tameemi, J. Xue, I. H. Ali, and S. Ajit, "A Hybrid Machine Learning Approach for Predicting Student Performance Using Multi-class Educational Datasets," *Procedia Comput. Sci.*, vol. 238, pp. 888–895, 2024, doi: 10.1016/j.procs.2024.06.108.

[2] A. Khaidar, N. Nurdin, F. Fajriana, T. Taufiq, and D. Hamdhana, "Classification Analysis of Single Tuition Fees Using the Random Forest Method with K-Fold Cross Validation," J. Appl. Informatics Comput., vol. 10, no. 1, pp. 125–133, 2026.

[3] Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi, "Peraturan Menteri Pendidikan, Kebudayaan, Riset, dan Teknologi Nomor 2 Tahun 2024 tentang Standar Satuan Biaya Operasional Pendidikan Tinggi (SSBOPT)." 2024.

[4] I. Papadogiannis, M. Wallace, and G. Karountzou, "Educational Data Mining: A Foundational Overview," *Mach. Learn. Knowl. Extr.*, vol. 6, no. 4, pp. 1644–1664, 2024, doi: https://doi.org/10.3390/encyclopedia4040108.

[5] A. F. M. Nafuri, N. S. Sani, N. F. A. Zainudin, A. H. A. Rahman, and M. Aliff, "Clustering Analysis for Classifying Student Academic Performance in Higher Education," *Appl. Sci.*, vol. 12, no. 13, pp. 1–22, 2022, doi: https://doi.org/10.3390/app12199467.

[6] D. Wulandari, T. Prahasto, and V. Gunawan, "Penerapan Principal Component Analysis (PCA) Untuk Mereduksi Dimensi Data Penerapan Teknologi Informasi dan Komunikasi untuk Pendidikan di Sekolah," *J. Sist. Inf. Bisnis*, vol. 02, pp. 91–96, 2016, doi: 10.21456/vol6iss2pp91-96.

[7] A. Zaki, Irwan, and I. A. Sembe, "Penerapan K-Means Clustering dalam Pengelompokan Data (Studi Kasus Profil Mahasiswa Matematika FMIPA UNM)," *J. Math. Comput. Stat.*, vol. 5, no. 2, pp. 163–176, 2022.

[8] S. Hartati, N. A. Ramdhan, and H. A. SAN, "Prediksi Kelulusan Mahasiswa dengan Naïve Bayes dan Feature Selection Information Gain," *J. Ilm. Intech Inf. Technol. J. UMUS*, vol. 4, no. 02, pp. 223–235, 2022.

[9] D. E. Prayogo and Kusrini, "Application of K-Means and Naïve Bayes Algorithms for Prediction Model of Student Interest Concentration (Case Study: Amikom University Yogyakarta)," *G-Tech J. Teknol. Terap.*, vol. 9, no. 1, pp. 511–519, 2025, doi: https://doi.org/10.70609/gtech.v9i1.6523.

[10] M. R. Gusmansyah, Rahmaddeni, Rohid, S. Daulay, and A. Rivaldi, "Perbandingan Metode Machine Learning untuk Klastering Penerima Bantuan Pendidikan Siswa," *J. Pustaka AI*, vol. 5, no. 2, pp. 193–203, 2025, doi: https://doi.org/10.55382/jurnalpustakaai.v5i2.1139.

[11] A. Sapitri, N. Nurdin, and Y. Afrilia, "Implementation of Clustering Method Using K-Means Algorithm for Grouping BPJS Health Patient Medical Record Data," *J. Appl. Informatics Comput.*, vol. 9, no. 5, pp. 2391–2398, 2025, [Online]. Available: http://jurnal.polibatam.ac.id/index.php/JAIC

[12] A. Alfitra, Nurdin, and R. Meiyanti, "Comparison of K-Means and K-Medoids Methods in Clustering High Population Density Areas in Bireuen Regency," *JITE (Journal Informatics Telecommun. Eng.*, vol. 9, no. 1, pp. 292–302, 2025, doi: 10.31289/jite.v9i1.15602.

[13] B. A. Fauzan, M. Jamaris, Junadhi, and H. Asnal, "Implementation of K-Means Clustering Algorithm for Grouping Traffic Violation Levels in Siak," *J. Teknol. dan Open Source*, vol. 5, no. 1, pp. 81–88, 2022, doi: 10.36378/jtos.v5i1.2427.

[14] S. N. Hariono, N. Nurdin, and L. Rosnita, "Comparison of K-Nearest Neighbors Method and Naïve Bayes Method in Classifying the Quality of Oil Palm Seed Varieties," *J. Inov. Teknol. dan Rekayasa*, vol. 10, no. 2, pp. 413–425, 2025, doi: 10.31572/inotera.Vol10.Iss2.2025.ID539.

[15] M. S. Samosir and L. Wati, "Penerapan Naive Bayes Untuk Memprediksi Kelulusan Mahasiswa Rekayasa Perangkat Lunak Politeknik Negeri Bengkalis," *Remik Ris. dan E-Jurnal Manaj. Inform. Komput.*, vol. 8, no. 3, pp. 838–848, 2024, doi: http://doi.org/10.33395/remik.v8i3.13964.

[16] N. Nurdin, M. Suhendri, and Y. Afrilia, "Klasifikasi Karya Ilmiah (Tugas Akhir) Mahasiswa Menggunakan Metode Naive Bayes Classifier (Nbc)," *Sist. J. Sist. Inf.*, vol. 10, pp. 268–279, 2021, doi: 10.32520/stmsi.v10i2.1193.

[17] Y. Safrina and Nurdin, "Analisis Perbandingan Metode Naïve Bayes dengan Forward Chaining Pada Sistem Pakar Diagnosa Kanker Servik," *J. Sist. Inf. Kaputama*, vol. 9, no. 2, pp. 139–145, 2025.

[18] N. Amalia, N. Nurdin, and F. Fajriana, "Z-Score Based Initialization for K-Medoids Clustering : Application on QSAR Toxicity Data," *J. Appl. Informatics Comput.*, vol. 9, no. 5, pp. 2410–2417, 2025, doi: https://doi.org/10.30871/jaic.v9i5.10448.

[19] A. A. Alharbi and J. Allohibi, "A New Hybrid Classification Algorithm for Predicting Student Performance," *AIMS Math.*, vol. 9, no. 7, pp. 18308–18323, 2024, doi: 10.3934/math.2024893.

[20] A. Khaidar, N. Nurdin, and F. Fajriana, "Single Tuition Fee Classification Using Light Gradient Boosting Machine with Confusion Matrix Analysis," Journal of Artificial Intelligence and Software Engineering, vol. 5, no. 4, pp. 1444–1454, 2025, doi: 10.30811/jaise.v5i4.847.

[21] V. Popovych and M. Drlik, "Identification of Students with Similar Performances in Micro-Learning Programming Courses with Automatically Evaluated Student Assignments," *Appl. Sci.*, vol. 14, no. 3615, pp. 1–26, 2024, doi: https://doi.org/10.3390/app14093615.