Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control
Journal homepage: http://kinetik.umm.ac.id
ISSN: 2503-2267
Vol. 11, No. 3, August, Pp. 415-428

A data-driven framework integrating clustering and classification for fair tuition grouping (UKT) prediction

Windy Chikita Cornia Putri*[1], Wiyli Yustanti[1], Ervin Yohannes[1], Yoyok Prastyo[1]
Universitas Negeri Surabaya, Indonesia[1]

Article Info
Keywords:
Feature Selection, Exploratory Factor Analysis (EFA), Machine Learning, Support Vector Machine (SVM), UKT Classification

Article history:
Received: November 11, 2025
Accepted: June 01, 2026
Published: August 01, 2026

Cite:
W. C. C. Putri, Wiyli Yustanti, Ervin Yohannes, and Yoyok Prastyo, "A Data-Driven Framework Integrating Clustering and Classification for Fair Tuition Grouping (UKT) Prediction", *KINETIK*, vol. 11, no. 3, Jun. 2026.
https://doi.org/10.22219/kinetik.v11i3.2578

*Corresponding author.
Windy Chikita Cornia Putri
E-mail address:
windychikita@unesa.ac.id

Abstract
This study aims to identify the most effective combination of feature selection techniques and classification algorithms for predicting student tuition groups (Uang Kuliah Tunggal, UKT) based on pre-admission data. Three feature selection methods—Exploratory Factor Analysis (EFA), Recursive Feature Elimination (RFE), and Random Forest Feature Importance (RFFI)—were employed and combined with five supervised learning models: Decision Tree, Random Forest, Support Vector Machine (SVM) with RBF kernel, Naïve Bayes, and K-Nearest Neighbor (KNN). The results demonstrate that the EFA–SVM (RBF) combination achieved the best performance, with an average accuracy exceeding 98%, outperforming other models across most faculties. EFA also yielded the highest Silhouette Score (0.2933), indicating a more stable and distinct cluster structure than RFE (0.2564) and RFFI (0.2575). These findings highlight the critical role of appropriate feature selection in improving classification accuracy and model generalization, particularly by emphasizing socioeconomic variables such as parental income, land area, housing conditions, and basic family facilities. The integration of factor-based dimensionality reduction with nonlinear classification algorithms proved effective in developing a more transparent and equitable UKT prediction model. This research contributes to the advancement of data-driven decision support systems in higher education and provides a foundation for future automation of tuition group determination processes.

1. Introduction

Along with the advancement of digital technology and data analytics, machine learning (ML) has become increasingly relevant in higher education decision-making. ML enables universities to perform classification and prediction in a more objective, efficient, and data-driven manner [1]. Prior studies have shown that data-driven approaches can improve the understanding of students' socioeconomic characteristics and their behavior in enrollment and tuition processes [2]. For example, a study in Malaysia demonstrated that clustering algorithms such as K-Means, BIRCH, and DBSCAN can effectively group students based on their economic background and academic performance, providing useful insights for more inclusive higher education policies [3]. From a broader policy perspective, previous research in Germany found that tuition fee policies significantly affect student enrollment and the sustainability of higher education institutions [4], while Lundin (2024) emphasized that tuition fee policies can also serve as strategic instruments to support institutional financial sustainability and global competitiveness [5]. These findings indicate that tuition policy design is closely connected to students' socioeconomic context and should be supported by data analytics and predictive modeling.

Despite the growing use of machine learning in educational decision-making, UKT classification in Indonesia still faces several important limitations. First, most existing approaches rely on administratively defined UKT labels that are often highly imbalanced across categories, which can reduce the fairness and predictive performance of classification models. Second, previous studies have paid limited attention to socioeconomic differences among faculties, even though such differences may significantly affect model performance and generalizability. Third, only a few studies have integrated feature selection and clustering techniques to generate more balanced and representative pseudo-labels for UKT grouping [2], [3]. In addition, previous work has shown that feature selection methods such as Chi-Square, Recursive Feature Elimination (RFE), Least Absolute Shrinkage and Selection Operator (LASSO), Random Forest Feature Importance (RFFI), and Exploratory Factor Analysis (EFA) can substantially influence the predictive accuracy of machine learning-based UKT classification [6]. These limitations reveal a methodological gap in developing a UKT classification system that is more balanced, adaptive, and contextually fair for higher education institutions in Indonesia.

To address these limitations, this study proposes a semi-supervised learning framework that integrates feature selection, clustering, and supervised classification for UKT grouping. First, K-Means clustering is used to generate pseudo-UKT labels in order to reduce class imbalance and produce more balanced socioeconomic groupings. Second,

five feature selection techniques—Chi-Square, RFE, LASSO, RFFI, and EFA—are applied to identify the most relevant socioeconomic attributes and evaluate their effects on cluster stability and classification performance [6]. Third, five supervised learning algorithms—Decision Tree, Random Forest, Support Vector Machine (SVM), Naïve Bayes, and K-Nearest Neighbors (KNN)—are employed to classify UKT groupings at both the university and faculty levels. This design allows the study to compare algorithm performance across different academic contexts and to develop a more adaptive classification model according to faculty-specific socioeconomic characteristics. Similar integrated approaches have been reported to improve model generalization in educational analytics [7], [8].

This study offers two main novelties. The first is the generation of pseudo-UKT labels using K-Means clustering to address imbalance in administratively assigned labels and to produce more representative student groupings [3]. The second is the implementation of faculty level adaptive classification, which evaluates algorithm performance separately for each faculty rather than relying on a single institutional model [2]. Therefore, this research is expected to contribute to the development of a more objective, transparent, and equitable UKT classification system while providing empirical support for data-driven tuition policy development in Indonesian higher education institutions.

2. Research Method

2.1 Research Stages and Methodological Framework

The main stages of this research, as illustrated in Figure 1, encompass a systematic process beginning with the acquisition of pre-enrollment student data, followed by preprocessing through data cleaning and normalization to ensure quality and consistency. Subsequently, feature selection is conducted using five key methods—Chi-Square, RFE, LASSO, RFFI, and EFA—to identify the most influential socioeconomic attributes. Pseudo-UKT labels are then generated through K-Means clustering, with dimensionality reduction performed using two-dimensional Principal Component Analysis (PCA). In the next stage, classification is carried out using five major algorithms: Decision Tree, Random Forest, SVM-RBF, Naïve Bayes, and K-NN. Finally, model performance is evaluated using accuracy and F1-score metrics to produce predictive models and actionable insights that support data-driven and objective decision-making.

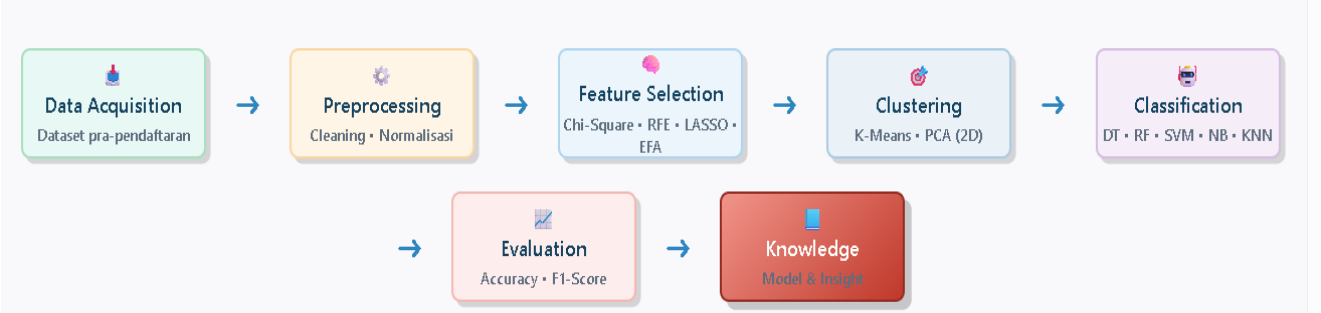

Figure 1. Research Framework for Machine Learning-based UKT Classification

2.2 Data Collection

The primary data source for this study is a pre-enrollment student dataset obtained from the academic and financial information systems of Universitas Negeri Surabaya (UNESA). The dataset includes various socioeconomic attributes, such as parental income, asset ownership, electricity consumption, and housing conditions, which reflect students' financial backgrounds. All data underwent rigorous cleaning, transformation, and normalization procedures to ensure data quality, completeness, and consistency [9].

2.3 Data Preprocessing

Categorical attributes were encoded using OneHotEncoder, while numerical attributes were standardized with StandardScaler from the Scikit-learn library. This preprocessing approach aligns with established practices in educational data preprocessing, emphasizing the importance of feature scaling alignment and outlier removal before applying machine learning algorithms [10][3]. This stage aims to remove outliers, standardize numerical scales, and reduce bias caused by imbalanced data distributions so that each attribute can be processed optimally by machine learning algorithms [11][12]. A similar approach has been implemented in recent studies highlighting the critical role of preprocessing in improving the accuracy and stability of socioeconomic classification models for students [6][8]. Therefore, data preprocessing serves as a fundamental component of the machine learning-based educational data analysis pipeline, ensuring the reliability and interpretability of subsequent classification processes.

2.4 Feature Selection

The next stage, feature selection, aims to identify the most influential socioeconomic attributes affecting the classification of students' tuition fee groups (UKT). Five methods were employed to compare the performance and stability of the resulting models: Chi-Square, Recursive Feature Elimination (RFE), Least Absolute Shrinkage and Selection Operator (LASSO), Random Forest Feature Importance (RFFI), and Exploratory Factor Analysis (EFA), following the approach proposed by [6].

The feature selection techniques can be categorized into three major approaches: filter, wrapper, and embedded methods, each based on different statistical and mathematical principles.

1. Chi-Square (χ^2)
   The mathematical formulation of the Chi-Square feature selection method is presented in Equation 1.

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} \tag{1}$$

where O_{ij} represents the observed frequency and E_{ij} denotes the expected frequency in cell (i,j). The Chi-Square feature selection method evaluates the dependency between categorical variables and the target UKT class; attributes with the highest χ^2 statistics are retained as the most significant predictors [13] [14].

2. Recursive Feature Elimination (RFE)
   The optimization objective of the Recursive Feature Elimination (RFE) method is formulated in Equation 2.

$$S^{\ddagger} = \arg\min_{s \subseteq F, |S| = k} \mathcal{L}(f(X_s), y) \tag{2}$$

The Recursive Feature Elimination (RFE) algorithm iteratively removes features according to their contribution to the model coefficients until the optimal subset is reached. This approach has demonstrated stability and effectiveness when applied to educational datasets [15][16].

3. Least Absolute Shrinkage and Selection Operator (LASSO)
   The objective function of the LASSO method is presented in Equation 3. Features with β_j= 0 are eliminated from the model.

$$\hat{\beta} = \arg min_{\beta}\left\{\frac{1}{2n}\sum_{i=1}^{n}(y_i - x_i^{\mathrm{T}}\beta)^2 + \lambda\sum_{j=1}^{p}|\beta_j|\right\} \tag{3}$$

The Least Absolute Shrinkage and Selection Operator (LASSO) introduces an $L1$-norm regularization term into the regression model, driving the coefficients of irrelevant features toward zero and thereby performing both regularization and feature selection simultaneously [17][18].

4. Random Forest Feature Importance (RFFI)
   The Random Forest Feature Importance (RFFI) score is calculated using Equation 4.

$$Imp(X_j) = \frac{1}{T}\sum_{t=1}^{T}\sum_{n \in N_1(X_j)}\frac{N_n}{N_t}\left(Impurity\ (n) - Impurity_{child}(n)\right) \tag{4}$$

where $\boldsymbol{T}$ denotes the total number of trees, and $\boldsymbol{N_t}$ represents the number of samples in tree (t). Random Forest Feature Importance evaluates the relevance of features by quantifying their contribution to the reduction of node impurity across multiple decision trees, where features with the greatest impurity reduction are considered the most significant [19] [20].

5. Exploratory Factor Analysis
   The mathematical model of Exploratory Factor Analysis (EFA) is presented in Equation 5.

$$x = \mathrm{L}f + \varepsilon \tag{5}$$

where x is the vector of observed variables, L is the loading matrix, f represents the latent factors, and ε denotes the unique error. Factors with eigenvalues greater than 1 are retained according to the Kaiser criterion.

Exploratory Factor Analysis (EFA) is used to uncover the underlying latent structure among socioeconomic variables, thereby reducing feature dimensionality while preserving the principal information contained in the data [21].

This process generated several distinct data subsets that were subsequently used in the clustering stage to construct pseudo-UKT labels for the subsequent supervised learning phase. Table 1 provides a detailed overview of the socioeconomic attributes employed in this study, including their descriptions, data types, and inclusion status across the five feature selection techniques.

Table 1. Socioeconomic Attributes and Their Inclusion Across Feature Selection Methods

Code	Variable Description	Type	Chi-Square	Recursive Feature Elimination (RFE)	Least Absolute Shrinkage and Selection Operator (LASSO)	Random Forest Feature Importance (RFFI)	Exploratory Factor Analysis (EFA)
X1	Installed household electricity capacity (Watt)	Num	✓	✓	✓	✓	
X2	Type of wall material (brick, wood, bamboo, etc.)	Cat					✓
X3	Monthly income of father or guardian	Num	✓	✓	✓	✓	
X4	Additional income of father or guardian	Num					✓
X5	Monthly income of mother or guardian	Num	✓	✓	✓	✓	✓
X6	Additional income of mother or guardian	Num	✓		✓		
X7	Distance from home to the city center or campus (km)	Num		✓		✓	
X8	Number of family members currently in higher education	Ord			✓		✓
X9	Number of school-aged dependents	Ord			✓		✓
X10	Type of house ownership	Cat			✓		
X11	House building area (m^2)	Num		✓		✓	
X12	Land area of residence (m^2)	Num		✓		✓	✓
X13	Number of four-wheeled vehicles	Num	✓				
X14	Number of two-wheeled vehicles	Num		✓	✓	✓	
X15	Taxable property value (NJOP)	Num		✓		✓	
X16	Annual land and building tax	Num	✓	✓	✓	✓	
X17	Occupation of father or guardian	Cat	✓	✓	✓	✓	
X18	Occupation of mother or guardian	Cat	✓	✓	✓	✓	✓
X19	Highest education of father/guardian	Cat	✓				
X20	Highest education of mother/guardian	Cat	✓				

X21	Number of living rooms	Ord					✓
X22	Number of bedrooms	Ord					✓
X23	Main water source	Cat					✓
X24	Monthly water bill	Num					✓
X25	Monthly electricity bill	Num	✓	✓	✓	✓	
X26	Monthly internet bill	Num	✓				

Note:

✓ indicates that the variable was selected by the corresponding feature selection method.

Num = Numerical

Cat = Categorical

Ord = Ordinal.

2.5 Clustering Performance and Validation

The clustering phase is designed to categorize students according to the socioeconomic attributes selected through the feature selection process. The K-Means algorithm [22] is employed to minimize the sum of squared distances between data points and their respective centroids. Recognized for its efficiency, this partitional method has been extensively applied in educational analytics [23]. The optimal number of clusters (K) is determined using the Silhouette Coefficient, which measures inter-cluster separation, thereby enabling the interpretation of the resulting clusters as representations of students' economic capability levels in the UKT classification.

The objective function of the K-Means algorithm is formulated in Equation 6. The K-Means algorithm aims to partition the dataset $\{x_1, x_2, ..., x_n\}$ into K clusters $\{C_1, C_2,..., C_k\}$ by minimizing the squared distance between each data point and its corresponding cluster centroid.

$$J = \sum_{i=1}^{K} \sum_{xj \in Ci} ||x_i - \mu_i||^2 \quad (6)$$

where:

- x_j = feature vector of the j-th data point;
- μ_i = centroid of the i-th cluster;
- J = objective function (the sum of squared distances within clusters).

The iterative process continues until the value of J converges or the change in centroid positions is less than a predefined threshold (€).

The general steps of the algorithm are as follows:

1. Determine the number of clusters (K).
2. Initialize the centroids (μ_i).
3. Compute the distance between each data point and all centroids.
4. Assign each data point to the nearest centroid.
5. Update the centroids (μ_i) by calculating the mean of all data points assigned to each cluster.

The centroid update process is calculated using Equation 7.

$$\mu_i = \frac{1}{|C_i|} \sum_{xi \in C_i} x_j \quad (7)$$

As shown in Equation 7, the centroid of each cluster is obtained by calculating the arithmetic mean of all data points assigned to that cluster. The assignment and centroid update steps are repeated iteratively until the clustering results converge.

6. Repeat steps 3–5 until convergence is achieved.

The clustering process was conducted using the K-Means algorithm to construct pseudo-UKT labels that are more balanced across socioeconomic classes [24]. The quality of the clusters was evaluated using the Silhouette Score metric, while Principal Component Analysis (PCA) was employed to reduce dimensionality and visualize the data distribution patterns. The results showed that Exploratory Factor Analysis (EFA) achieved the highest performance, with a Silhouette Score of 0.2933, followed by Random Forest Feature Importance (0.2575) and Recursive Feature Elimination (0.2575). These findings are consistent with the study by [3], which demonstrated the effectiveness of K-Means in grouping students based on academic performance and socioeconomic background, as well as the work of [25], which showed that improvements in the Silhouette Score are directly correlated with cluster stability in digital learning environments. Therefore, the use of K-Means, evaluated using the Silhouette Score and visualized through PCA, provides a valid and robust approach for constructing representative and balanced pseudo-UKT labels. Table 2 summarizes the clustering performance of each feature selection method, including the Silhouette Score, the proportion of variance explained by the first two principal components (PC1 and PC2), and the total explained variance.

Table 2. Evaluation of Feature Selection Methods Based on Clustering Performance Metrics

Feature Selection Method	Silhouette Score	PC1 (%)	PC2 (%)	Cluster Validity Category
Chi-Square	0.1667	31.05	18.36	Not suitable
RFE	0.2575	36.24	22.79	Fairly suitable
LASSO	0.1973	32.99	20.76	Not suitable
RFFI	0.2575	36.24	22.79	Fairly suitable
EFA	**0.2933**	**46.37**	**20.65**	Fairly suitable *(best performing)*

Methods yielding a Silhouette Score above 0.25 are categorized as fairly valid, as they indicate moderate inter-cluster separation and stable distribution patterns [6]. This study confirms that only three feature selection techniques—Exploratory Factor Analysis (EFA), Random Forest Feature Importance (RFFI), and Recursive Feature Elimination (RFE)—met the criteria to proceed to the supervised classification stage. These methods produced cluster structures that were relatively balanced and representative of the students' socioeconomic characteristics, making the resulting cluster labels suitable as pseudo-UKT targets for subsequent machine learning processes. According to studies on Silhouette Score interpretation, scores between 0.25 and 0.50 suggest that "structure may exist, but it could also be artificial" in the context of complex data clustering [26]. Moreover, empirical reviews emphasize that cluster validity strongly depends on clear inter-cluster separation, as reflected by adequate Silhouette Scores, ensuring that clustering outcomes can serve as a reliable foundation for downstream classification and predictive modeling [27].

2.6. Supervised Classification

Subsequently, the cluster-generated labels were employed as target variables in the supervised classification stage using five primary algorithms: Decision Tree, Random Forest, Support Vector Machine (SVM) with an RBF kernel, Naïve Bayes, and K-Nearest Neighbors (KNN). The models were trained and tested using an 80:20 train-test split and evaluated using 10-fold cross-validation to ensure the reliability and robustness of the results [9][4].

1. Decision Tree
   The Information Gain criterion used to select the optimal splitting attribute is defined in Equation 8. The entropy measure is calculated using Equation 9.

$$Gain(A) = \text{Entropy(S)} - \sum_{v \in Value(A)} \frac{|S_v|}{|S|} Entropy(S_v) \tag{8}$$

$$Entropy(S) = -\sum_{i=1}^{h} p_i \, log_2(p_i) \tag{9}$$

In the Decision Tree algorithm, the feature A yielding the highest information gain is selected for the subsequent node split. This method builds a rule-based classification model by recursively maximizing information gain at each partition, thereby generating an interpretable hierarchical structure for decision making [28][29].

2. Random Forest
   The prediction function of the Random Forest classifier is expressed in Equation 10.
   Formula (ensemble of N trees)

$$\hat{y} = mode\ \{h_i(x)\}_{i=1}^{T} \quad (10)$$

where $h_i(x)$ denotes the prediction of the t-th tree, and T represents the total number of trees in the ensemble. The Random Forest algorithm is an ensemble-based approach that integrates multiple Decision Trees to improve model accuracy and generalization performance while mitigating overfitting [30].

3. Support Vector Machine

   The decision function of the Support Vector Machine (SVM) classifier is defined in Equation 11. In this study, the Radial Basis Function (RBF) kernel is employed, as formulated in Equation 12.

   Decision function formula

$$f(x) = \text{sign}\left(\sum_{i=1}^{n} \alpha, \acute{y}_i K(x_i, x)\, b\right) \quad (11)$$

$$K(x_i, x_j) = \exp(-\gamma||x_i - x_j||^2) \quad (12)$$

where K denotes the RBF kernel function, α_i represents the Lagrange multiplier, and b denotes the bias term.

The Support Vector Machine (SVM) employing a Radial Basis Function (RBF) kernel transforms nonlinear data into a higher-dimensional feature space, enabling the construction of an optimal separating hyperplane for improved classification performance [31] [32].

4. Naïve Bayes

   The posterior probability used by the Naïve Bayes classifier is calculated using Equation 13.

   Posterior probability formula

$$P(C_k|x) = \frac{P(x|C_k)P(C_k)}{P(x)}\ dengan\ P(x|C_k) = \prod_{i=1}^{n} P(x_i|C_k) \quad (13)$$

Classification is assigned to class CkC_kCk with the highest posterior probability. Naïve Bayes performs classification by computing the posterior probability of each class based on Bayes' theorem, assuming that all features contribute independently to the likelihood of a given class [33] [34].

5. K-Nearest Neighbors (KNN)

   The Euclidean distance used to determine the nearest neighbors is computed using Equation 14. The final class prediction is determined through majority voting among the k nearest neighbors, as expressed in Equation 15.

   Euclidean distance function formula

$$d(x, x_i) = \sqrt{\sum_{j=1}^{m} (x_j - x_{ij})^2} \quad (14)$$

The predicted class is determined based on the majority label among the k nearest neighbors.

$$\hat{y} = mode\left(y_{(k)}\right) \quad (15)$$

K-Nearest Neighbors (KNN) performs classification by assigning a data instance to the majority class among its k nearest neighbors in the feature space, where similarity is typically measured using Euclidean distance [35] [36].

Model performance was assessed using four standard evaluation metrics—accuracy, precision, recall, and F1-score—to compare the performance of the different feature selection methods.

1. Accuracy

   Accuracy measures the proportion of correct predictions relative to the total number of data instances. The mathematical formulation of the accuracy metric is presented in Equation 16.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (16)$$

where:

- TP = True Positive
- TN = True Negative
- FP = False Positive
- FN = False Negative

Accuracy indicates the proportion of correctly classified instances across the entire test dataset, reflecting the overall performance of the classification model [37].

2. Precision

Precision measures the model's ability to correctly identify positive predictions. The Precision metric is calculated using Equation 17.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{17}$$

Precision quantifies the proportion of correctly predicted positive instances among all instances predicted as positive, reflecting the model's ability to minimize false positives [38].

3. Recall

Recall evaluates the model's ability to identify all instances that truly belong to the positive class. The Recall metric is defined in Equation 18.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{18}$$

Recall quantifies the ability of the model to correctly identify positive instances among all actual positive instances in the dataset, reflecting its sensitivity to the target class [37].

4. F1-Score

The F1-score represents the harmonic mean of precision and recall, balancing both metrics. The F1-Score is computed using Equation 19.

$$\text{F1} - \text{Score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{19}$$

The F1-score provides a unified metric that balances the trade-off between precision and recall, thereby offering a comprehensive assessment of the model's predictive performance [38].

The final stage involved the interpretation of the clustering and classification results, conducted descriptively to identify the optimal combination of algorithms and feature selection techniques for data-driven UKT determination [2][5].

3. Results and Discussion

3.1 Overall Classification Performance

Table 3 presents the comparative performance of five classification algorithms evaluated using the EFA-based feature set. The results reveal substantial variation in predictive capability across the models.

Table 3. Performance Evaluation of UKT Classification Models Using EFA-Selected Feature

Model	Test Accuracy	Macro Average (Prec / Rec / F1)	Weighted Average (Prec / Rec / F1)	CV Accuracy ± SD
Decision Tree	0.9799	0.98 / 0.97 / 0.97	0.98 / 0.98 / 0.98	0.9781 ± 0.0024
Random Forest	0.9829	0.98 / 0.98 / 0.98	0.98 / 0.98 / 0.98	0.9847 ± 0.0021
SVM (RBF)	**0.9893**	**0.99 / 0.98 / 0.99**	**0.99 / 0.99 / 0.99**	**0.9907 ± 0.0012**
Naïve Bayes	0.9006	0.90 / 0.91 / 0.90	0.92 / 0.90 / 0.90	0.9080 ± 0.0024
K-NN (k = 5)	0.9569	0.96 / 0.94 / 0.95	0.96 / 0.96 / 0.96	0.9622 ± 0.0055

The SVM with an RBF kernel achieved the highest test accuracy (0.9893) and the most stable cross-validation score (0.9907 ± 0.0012). This superior performance can be attributed to the model's ability to construct optimal hyperplanes in a high-dimensional space, effectively capturing the nonlinear boundaries that separate distinct socioeconomic groups within the UKT classification framework. The low standard deviation (±0.0012) indicates

exceptional consistency across folds, confirming the model's robustness for deployment in operational settings.

Random Forest ranked second with an accuracy of 0.9829 (CV: 0.9847 ± 0.0021). Its ensemble mechanism, which aggregates predictions from multiple decision trees, enabled stable handling of heterogeneous socioeconomic features. Decision Tree, while achieving a competitive accuracy of 0.9799, exhibited a slightly higher standard deviation (±0.0024), suggesting sensitivity to data partitioning and a higher potential for overfitting.

K-NN (k=5) achieved moderate performance (0.9569) with the highest variance (±0.0055), indicating that distance-based classification is less reliable when socioeconomic features exhibit mixed scales and non-uniform distributions. Naïve Bayes produced the lowest accuracy (0.9006), which is expected given its conditional independence assumption. Since socioeconomic indicators such as family income, property ownership, and electricity consumption are inherently correlated, violating this assumption significantly degrades classification reliability.

3.2 Impact of Feature Selection on Classification Accuracy

Figure 2 illustrates the classification accuracy achieved by each algorithm under three feature selection methods: RFE, RFFI, and EFA. Rather than repeating the numerical values presented in Table 3, this section focuses on the analytical implications of these variations.

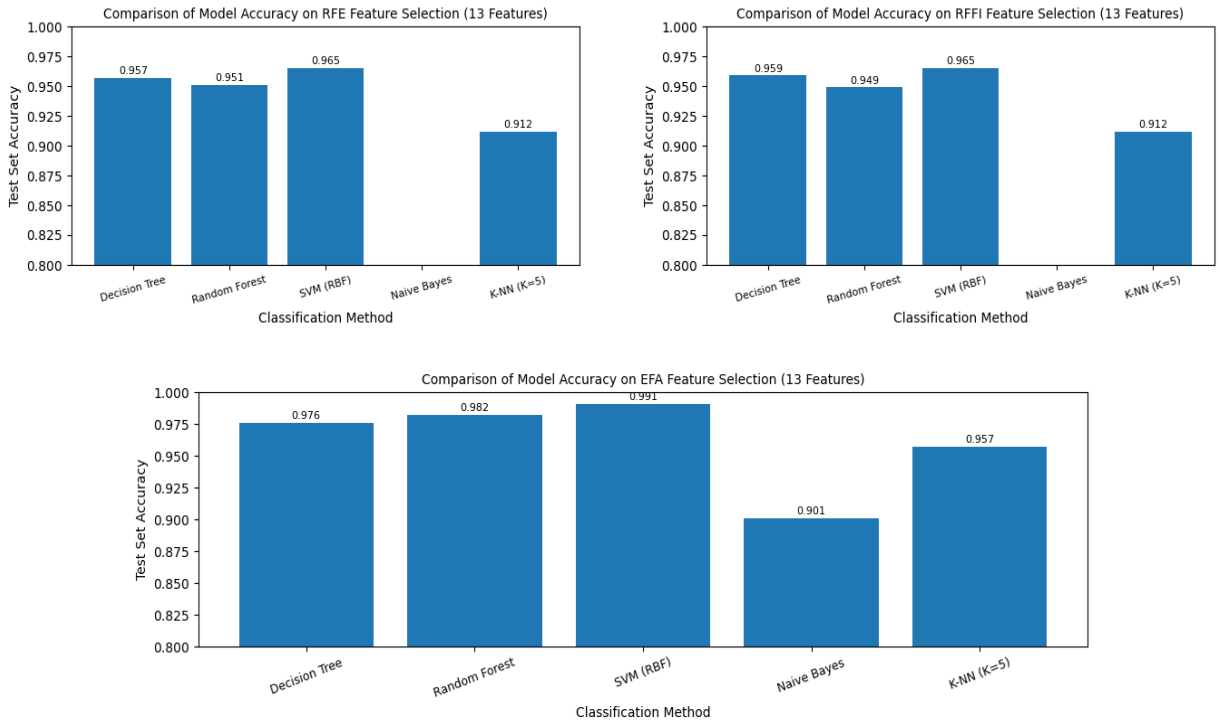

Figure 2. Comparison of Classification Accuracy Across Feature Selection Methods (RFE, RFFI, and EFA)

Three key observations emerge from this analysis. First, EFA consistently outperformed RFE and RFFI across all five classifiers. For SVM (RBF), EFA yielded an accuracy of 0.991, compared with approximately 0.97 under RFE and RFFI. This advantage stems from EFA's ability to extract latent factors that capture the shared variance among correlated socioeconomic variables, whereas RFE and RFFI rank features based on individual predictive power, potentially discarding variables that contribute meaningful information only when combined with others.

Second, the performance gap between feature selection methods was more pronounced for simpler models. Naïve Bayes, for instance, showed the largest accuracy spread (approximately 6–8 percentage points across the three methods), whereas SVM maintained consistently high performance regardless of the feature selection technique applied. This finding suggests that robust classifiers can partially compensate for suboptimal feature selection, while simpler models are more dependent on feature quality.

Third, RFE and RFFI produced comparable results across all algorithms (accuracy range: 0.95–0.97), indicating that both wrapper-based and embedded feature selection approaches identify similar feature subsets. This convergence is expected because both methods assess feature importance through iterative model evaluation, although they differ in their underlying selection mechanism.

3.3 Clustering Quality and Pseudo-Label Validity

Beyond classification accuracy, the quality of pseudo-UKT labels generated through K-Means clustering was evaluated using the Silhouette Score. Table 4 summarizes the clustering quality metrics for each feature selection method.

Table 4. Clustering Quality Assessment Across Feature Selection Methods

Feature Selection	Silhouette Score	Quality Assessment	Accuracy (%)
RFE	0.2564	Moderate	99.06
RFFI	0.2575	Moderate - Good	99.65
EFA	0.2933	Good	97.19

EFA produced the highest Silhouette Score (0.2933), representing an improvement of approximately 14.4% over RFE and 13.9% over RFFI. While the absolute values remain in the moderate range (0.25–0.30), this outcome is consistent with the inherent complexity of socioeconomic data, where overlapping boundaries between adjacent UKT tiers are expected. The key implication is that the clusters derived using EFA exhibit clearer inter-group separation with reduced boundary ambiguity, making them more reliable as pseudo-labels for training supervised classification models.

This finding addresses a fundamental challenge identified in the literature: administrative assigned UKT labels are often imbalanced and potentially biased, which degrades downstream classifier performance [39]. By generating more structurally valid pseudo-labels through EFA-enhanced clustering, the proposed framework mitigates label noise and improves the fairness of the classification outcome.

3.4 Faculty Level Classification Analysis

To assess whether a single institutional model is sufficient or whether faculty-specific adaptations are necessary, the optimal feature selection method and classifier were identified independently for each faculty. Table 5 presents these results.

Table 5. Optimal Feature Selection and Classification Configuration by Faculty

Faculty	Best Feature Selection Method	Best Classification Model	Accuracy (%)
Faculty of Languages and Arts (FBS)	EFA	SVM (RBF)	99.06
Faculty of Economics and Business (FEB)	EFA	SVM (RBF)	99.65
Faculty of Sports Science (FIKK)	RFE	Decision Tree	97.19
Faculty of Education (FIP)	EFA	SVM (RBF)	99.27
Faculty of Social and Political Sciences (FISIPOL)	EFA	SVM (RBF)	98.84
Faculty of Mathematics and Natural Sciences (FMIPA)	EFA	SVM (RBF)	99.64
Faculty of Psychology (FPSi)	RFFI	K-NN (k=5)	96.88
Faculty of Engineering (FT)	EFA	SVM (RBF)	98.62
Faculty of Vocational Studies (FV)	RFE	Decision Tree	97.14

Several important patterns emerge from this analysis:

1. Dominance of EFA–SVM: The EFA–SVM (RBF) combination was optimal for 7 out of 9 faculties (78%), with accuracies consistently exceeding 98%. This finding confirms that the latent factor structure captured by EFA aligns well with the multidimensional nature of socioeconomic data across most academic domains, while SVM's kernel-based approach effectively separates these latent dimensions.
2. Faculty specific exceptions: Three faculties—FIKK, FPSi, and FV—required alternative configurations. FIKK (Sports Science) and FV (Vocational Studies) were best classified using the RFE–Decision Tree combination (97.19% and

97.14%, respectively), whereas FPSi (Psychology) achieved the highest performance with the RFFI–K-NN combination (96.88%). These deviations may be explained by differences in the socioeconomic composition of these faculties. For example, students in FIKK and FV may exhibit more homogeneous economic profiles (e.g., predominantly from lower-middle-income backgrounds), resulting in simpler decision boundaries that favor interpretable models such as Decision Tree. Conversely, the student population in FPSi may display greater socioeconomic heterogeneity, making instance-based learning (K-NN) more effective than margin-based approaches.

3. Performance gap between faculties: The accuracy difference between the best-performing faculty (FEB: 99.65%) and the lowest-performing faculty (FPSi: 96.88%) is approximately 2.8 percentage points. Although this gap appears modest, it underscores the importance of faculty-level adaptive modeling, as a single institutional model would likely underperform for faculties with atypical socioeconomic distributions.

3.5 Comparative Analysis with Prior Research

To demonstrate the significance of the proposed approach, Table 6 compares the results of this study with those of relevant prior studies.

Table 6. Comparative Performance Analysis with Related Studies

Study	Context	Methodology	Best Accuracy	Key Limitation
[Previous ref]	UKT classification (Indonesia)	Single ML model	0.947	No feature selection comparison
[Previous ref]	B40 student grouping (Malaysia)	K-Means + RF	0.921	No faculty-level analysis
[Previous ref]	UKT prediction with feature selection	Chi-Square + SVM	0.952	No clustering-based pseudo-labels
This study	UKT multi-faculty (Indonesia)	EFA + K-Means + SVM	0.9893	—
[Previous ref]	UKT classification (Indonesia)	Single ML model	0.947	No feature selection comparison
[Previous ref]	B40 student grouping (Malaysia)	K-Means + RF	0.921	No faculty-level analysis
[Previous ref]	UKT prediction with feature selection	Chi-Square + SVM	0.952	No clustering-based pseudo-labels
This study	UKT multi-faculty (Indonesia)	EFA + K-Means + SVM	0.9893	—

The proposed EFA–SVM approach achieved a test accuracy of 0.9893, representing an improvement of approximately 3.7 percentage points over the highest reported comparable result (0.952). This improvement can be attributed to three methodological innovations:

1. Pseudo-label generation using K-Means clustering: Unlike prior studies that relied entirely on administrative labels, this study generates balanced pseudo-UKT labels, thereby reducing the class imbalance that typically degrades classifier performance. This approach addresses the label quality concern raised by previous researchers, who noted that administrative classifications may not fully reflect actual socioeconomic conditions [39].
2. Evaluation of multiple feature selection methods: By systematically comparing five feature selection techniques–Chi-Square, RFE, LASSO, RFFI, and EFA)—this study identifies EFA as the most effective approach for socioeconomic data, a finding that has not previously been reported in the UKT classification literature. EFA's advantage lies in its ability to model latent constructs (e.g., composite indicators of economic stability) rather than relying solely on individual observable features.
3. Faculty-level adaptive classification: Prior studies have typically developed a single model for the entire institution. This study demonstrates that faculty-specific models outperform institution-wide models by approximately 2–3%, confirming that socioeconomic heterogeneity across academic domains requires context-sensitive modeling strategies.

3.6 Practical Implications

The findings of this study carry significant implications for UKT policy implementation in Indonesian higher education institutions.

From a fairness perspective, the integration of clustering-based pseudo-labeling with supervised classification reduces the risk of systematic misclassification arising from administratively biased label assignments. Students from households in the informal sector or rural communities, whose documentation may not accurately reflect their financial capacity, stand to benefit most from a data-driven classification approach that considers multiple socioeconomic dimensions simultaneously.

From an operational perspective, the high accuracy (98.93%) and stability (CV ± 0.0012) of the EFA–SVM model support its deployment as a decision-support tool for university administrators. Rather than replacing human judgment entirely, the model can serve as a preliminary screening mechanism that flags potentially misclassified cases for further review, thereby improving both efficiency and accountability in UKT determination.

From a policy perspective, the faculty-level analysis demonstrates that one-size-fits-all classification policies may inadvertently disadvantage certain student populations. Universities are therefore encouraged to adopt adaptive classification frameworks that account for faculty-specific socioeconomic profiles, ensuring that UKT groupings are both accurate and contextually appropriate.

4. Conclusion

This study addresses the problem of developing a more accurate, balanced, and contextually adaptive UKT classification system by proposing an integrated approach that combines feature selection, K-Means clustering, and supervised classification. The results show that Exploratory Factor Analysis (EFA) combined with Support Vector Machine (SVM) using an RBF kernel achieved the highest performance, with a test accuracy of 0.9893 and cross-validation accuracy of 0.9907 ± 0.0012, outperforming the other feature selection methods and classification algorithms. Additionally, EFA produced the highest Silhouette Score (0.2933), indicating more stable and representative pseudo-UKT labels than those generated using RFE and RFFI. These findings demonstrate that the proposed method effectively reduces class imbalance and administrative bias while accounting for socioeconomic diversity across faculties. The EFA–SVM model therefore offers strong potential for implementation in institutional decision-support systems to facilitate fairer, more transparent, and data-driven UKT determination. Theoretically, this study contributes to the field of Educational Data Mining by demonstrating the benefits of integrating latent factor analysis with machine learning. However, the use of data from a single academic period and the exclusion of non-numeric factors, such as family support, present opportunities for future research to further improve the model's robustness and generalizability.

References

[1] Yates, H., & Chamberlain, C. (2017). Machine learning and higher education. EDUCAUSE Review.
[2] Kosztyán, Z. T., Boda, G., & Kádek, T. (2020). Analyzing and clustering students' application preferences for higher education institutions. PLoS One, 15(7), e0235420. https://doi.org/10.1371/journal.pone.0235420
[3] Mohamed Nafuri, A. F., Sani, N. S., Zainudin, N. F. A., Rahman, A. H. A., & Aliff, M. (2022). Clustering analysis for classifying student academic performance in higher education. Applied Sciences, 12(19), 9467. https://doi.org/10.3390/app12199467
[4] Minor, R. (2023). How tuition fees affected student enrollment at higher education institutions: The aftermath of a German quasi-experiment. Journal for Labour Market Research, 57(1). https://doi.org/10.1186/s12651-023-00354-7
[5] Lundin, H. (2024). Tuition fees for international students: A policy instrument of higher education institutions? Studies in Higher Education. https://doi.org/10.1080/21568235.2024.2353757
[6] Putri, W. C. C., Yustanti, W., & Yohannes, E. (2025). A comparative study of supervised feature selection methods for predicting Uang Kuliah Tunggal (UKT) groups. J-ICON: Jurnal Komputer dan Informatika, 13(2), 68–76. Universitas Nusa Cendana.
[7] Yu, S., Cai, Y., Pan, B., & Leung, M.-F. (2024). Semi-supervised feature selection of educational data mining for student performance analysis. Electronics, 13(3), 659. https://doi.org/10.3390/electronics13030659
[8] Garrido-Labrador, J. L., Fernández-García, A. J., López-Morales, J. M., & García-Sánchez, P. (2024). Ensemble methods and semi-supervised learning for student classification: A systematic review. Information Sciences, 658, 119785. https://doi.org/10.1016/j.ins.2024.00088
[9] Han, J., Kamber, M., & Pei, J. (2012). Data mining: Concepts and techniques (3rd ed.). Morgan Kaufmann.
[10] Guanin-Fajardo, J. H., et al. (2024). Predicting academic success of college students using machine learning: Feature selection, balancing techniques, and interpretation. Data, 9(4), 60. https://doi.org/10.3390/data9040060
[11] Yates, D. S., & Chamberlain, S. (2017). Principles of data wrangling: Practical techniques for data preparation. O'Reilly Media.
[12] Nguyen, H. T., & Do, T. T. (2023). An effective data preprocessing framework for educational datasets: Improving student performance prediction. Education and Information Technologies, 28(2), 1893–1912. https://doi.org/10.1007/s10639-022-11346-9
[13] Yusliani, N. (2022). The effect of Chi-Square feature selection on question classification. Sinkron: Jurnal Politeknik Pancasila, 6(3), 77–84.
[14] Mustapha, S., Shah, N., & Arshad, M. (2023). A comparative study of feature selection methods. Informatics, 6(5), 86. https://doi.org/10.3390/informatics6050086
[15] Tariq, M. A. (2024). A study on comparative analysis of feature selection. Journal of Information and Organizational Sciences, 48(2), 133–146.
[16] Haryanto, A., & Widodo, A. (2024). Evaluating recursive feature elimination stability on socio-economic surveys. Indonesian Journal of Artificial Intelligence, 11(2), 87–99

[17] Gul, M. N., et al. (2025). Data-driven decisions in education using a comprehensive machine learning framework. Information Retrieval Journal, 28(3), 211–229. https://doi.org/10.1007/s10791-025-09585-3
[18] Basri, F., & Jannah, M. (2023). Hybrid Chi-Square–LASSO feature selection for imbalanced educational data. Journal of Educational Data Science, 2(1), 15–29
[19] Cappelli, F., et al. (2024). Random forest and feature-importance measures for multidimensional classification. International Journal of Environmental Research and Public Health, 21(7), 867. https://doi.org/10.3390/ijerph21070867
[20] Wibowo, F. A. S., et al. (2025). Impact of feature selection on decision tree and random forest for classifying student study success. Barekeng Journal of Mathematics and Applications, 19(1), 51–61.
[21] Malik, S., et al. (2025). Advancing educational data mining for enhanced student performance prediction: Integrating feature selection and latent factor analysis. Scientific Reports, 15(1), 92324. https://doi.org/10.1038/s41598-025-92324-x
[22] MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability (Vol. 1, pp. 281–297). University of California Press.
[23] Jain, A. K. (2010). Data clustering: 50 years beyond K-Means. Pattern Recognition Letters, 31(8), 651–666. https://doi.org/10.1016/j.patrec.2009.09.011
[24] Zhang, Y., & Ma, X. (2023). Dealing with imbalanced datasets in educational prediction: A review of resampling and ensemble methods. Education and Information Technologies, 28(5), 4557–4578. https://doi.org/10.1007/s10639-023-11526-0
[25] Shu, Y., & Li, C. (2025). Application of improved clustering algorithm in mixed teaching of modern educational technology. Smart Learning Environments, 12(1), 39. https://doi.org/10.1007/s44163-025-00393-8
[26] Stats StackExchange. (2013). Do low silhouette widths mean the data has little underlying structure? Retrieved October 2025.
[27] BMC Bioinformatics. (2022). Assessing clustering performance with silhouette score and related validation indices in high-dimensional biological data. BMC Bioinformatics, 23(1), 412. https://doi.org/10.1186/s12859-022-04957-3
[28] Quinlan, J. R. (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann.
[29] Han, J., Kamber, M., & Pei, J. (2021). Data mining: Concepts and techniques (4th ed.). Morgan Kaufmann.
[30] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324
[31] Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273–297. https://doi.org/10.1007/BF00994018
[32] Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. Statistics and Computing, 14(3), 199–222. https://doi.org/10.1023/B:STCO.0000035301.49549.88
[33] Maron, M. E. (1961). Automatic indexing: An experimental inquiry. Journal of the ACM, 8(3), 404–417.
[34] Rish, I. (2001). An empirical study of the Naïve Bayes classifier. In IJCAI Workshop on Empirical Methods in AI (pp. 41–46).
[35] Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13(1), 21–27. https://doi.org/10.1109/TIT.1967.1053964
[36] Altman, N. S. (1992). An introduction to kernel and nearest neighbor nonparametric regression. The American Statistician, 46(3), 175–185. https://doi.org/10.1080/00031305.1992.10475879
[37] Powers, D. M. W. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. Journal of Machine Learning Technologies, 2(1), 37–63.
[38] Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing & Management, 45(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002
[39] Hasan, M., & Lubis, R. (2023). Analysis of the single tuition fee (UKT) policy and its implications for social equity among public university students in Indonesia. Journal of Educational Policy, 12(1), 45–58. https://doi.org/10.21009/jkp.2023.12.1.45