JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)

VOL. 12. NO. 1 AUGUST 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i1.7469

HYBRID K-MEANS, FP-GROWTH, AND RANDOM FOREST FOR ACCURATE UKT PREDICTION IN WEST-SOUTHEAST ACEH

Hijrah[1*]; Jessika[2]

Information Technology and Database Technical Implementation Unit[1]
State Islamic Institute of Teungku Dirundeng Meulaboh, West Aceh, Aceh, Indonesia[1]
https://staindirundeng.ac.id[1]
hijrah@staindirundeng.ac.id[*]

Master Program in Information Technology Author Study Program[2]
Malikussaleh University, North Aceh, Aceh, Indonesia[2]
https://unimal.ac.id[2]
jesjessikaa@gmail.com

(*) Corresponding Author
(Responsible for the Quality of Paper Content)

Abstract— *Predicting the Single Tuition Fee (UKT) in economically vulnerable areas remains challenging due to diverse socioeconomic conditions and the limited capacity of existing models to capture interactions among contributing factors. This study proposes a hybrid framework that integrates segmentation, pattern discovery, and classification to improve UKT prediction accuracy in the Southwest Aceh region. Data from 452 student records were analyzed through three main stages: cluster-based socioeconomic segmentation, association pattern discovery to reveal dominant factor relationships, and the integration of these hybrid features into a Random Forest model. To ensure model stability and prevent data leakage, the framework was evaluated using a multilevel 5-fold cross-validation strategy. The developed model achieved 93.4% accuracy, demonstrating a significant improvement over baseline demographic models while identifying key socioeconomic determinants of UKT assignment. The proposed framework demonstrates how hybrid learning strategies can improve fairness and transparency in educational financial decision systems. These findings highlight the potential of the hybrid framework to support fairer and more data-driven tuition policies, offering a robust evaluation design for educational finance modeling in Indonesia.*

Keywords: *Association Analysis, Cluster Segmentation, Educational Data Mining, Hybrid Learning, UKT Prediction*

Intisari— *Prediksi besaran Uang Kuliah Tunggal (UKT) di wilayah rentan ekonomi masih menghadapi tantangan karena keragaman kondisi sosial-ekonomi serta keterbatasan model yang mampu menangkap interaksi antarfaktor. Penelitian ini mengusulkan kerangka hybrid yang mengintegrasikan pendekatan segmentasi, penemuan pola, dan klasifikasi untuk meningkatkan akurasi prediksi UKT di wilayah Barat-Selatan Aceh. Data dari 452 mahasiswa dianalisis melalui tiga tahap utama: segmentasi sosial-ekonomi berbasis klaster, identifikasi pola asosiasi untuk menyingkap keterkaitan faktor dominan, dan integrasi fitur hybrid tersebut ke dalam model Random Forest. Untuk menjamin stabilitas model dan mencegah kebocoran data (data leakage), kerangka ini divalidasi menggunakan pendekatan Stratified 5-fold Cross-Validation. Model yang dikembangkan mencapai akurasi 93,4%, menunjukkan peningkatan kinerja yang signifikan dibanding model demografi dasar serta mengungkap faktor sosial-ekonomi kunci dalam penentuan UKT. Temuan ini menunjukkan potensi kerangka hybrid dalam mendukung kebijakan UKT yang lebih adil dan berbasis data di Indonesia.*

Kata Kunci: *Analisis Asosiasi, Data Mining Pendidikan, Pembelajaran Hybrid, Prediksi UKT, Segmentasi Klaster*

INTRODUCTION

Education plays a fundamental role in national development, yet its accessibility is heavily influenced by the accuracy of financial management systems. Economic disparities significantly affect students' ability to persist in higher education, influencing performance and dropout risks, especially among those under severe financial pressure [1]. Lack of transparency in tuition fee allocation and governance inefficiencies further affect students' perceptions and highlight the need to improve educational service quality [2]. A critical challenge in this system is the improper grouping of Single Tuition Fees (UKT), where inaccuracies in socioeconomic evaluation result in low-income students being wrongly assigned to higher tuition categories. This misalignment is not merely an administrative error; it reinforces socioeconomic disparities and restricts access for underprivileged groups [3]. As a result, inaccurate UKT grouping can intensify financial stress and increase dropout risk despite tuition-assistance policies [4].

In the West-South Aceh region specifically, these challenges are compounded by unique socioeconomic vulnerabilities. The local economy is heavily reliant on informal sectors such as agriculture, plantations, and fisheries, leading to highly fluctuating and undocumented household incomes. Consequently, standard UKT categorization indicators which often assume steady monthly wages frequently fail to capture the true financial capacity of students' families in this area. This specific regional challenge makes accurate tuition fee prediction both highly complex and urgently needed. The application of Educational Data Mining (EDM) has grown significantly in recent years, proving its reliability not only in evaluating student academic performance [5], [6], but also in solving complex administrative challenges. Recent studies have demonstrated the urgency of utilizing machine learning algorithms, such as K-Means clustering and Random Forest, to establish data-driven financial policies and transparent tuition fee (UKT) classifications [7], [8], [9].

Recent studies emphasize that the integration of hybrid machine learning models and systematic predictive frameworks is crucial for addressing complex challenges in educational data mining, particularly in ensuring accuracy and decision-making transparency [10]. Advancements in machine learning allow more precise analyses of socio-economic and educational data to support equitable fee categorization [11]. K-Means clustering has been effective in grouping students based on socioeconomic similarities for data-driven UKT decisions [12].

Furthermore, recent advancements in 2025 have shown that combining unsupervised learning like K-Means with supervised classification can significantly enhance the precision of student performance and resource allocation models [13] and is capable of handling large datasets in education [14]. However, most studies apply clustering as a standalone technique, limiting interpretability of socio-economic disparities [15]. FP-Growth can discover hidden patterns, but rules generated without segmentation tend to overlook subgroup characteristics [16].

Random Forest is robust in educational prediction tasks [17]. For instance, a recent study by Alkhaidar [11], successfully applied the Random Forest method to classify Single Tuition Fees (UKT) at Lhokseumawe State Polytechnic, Aceh. While their model achieved a high overall accuracy of 96%, it exhibited significant limitations in predicting minority classes due to imbalanced socio-economic data, with the recall metric dropping to as low as 0.39 for certain UKT groups. This highlights that standalone models often rely on raw features without integrating clustering or association rule insights, potentially reducing accuracy and interpretability for specific vulnerable groups [18].

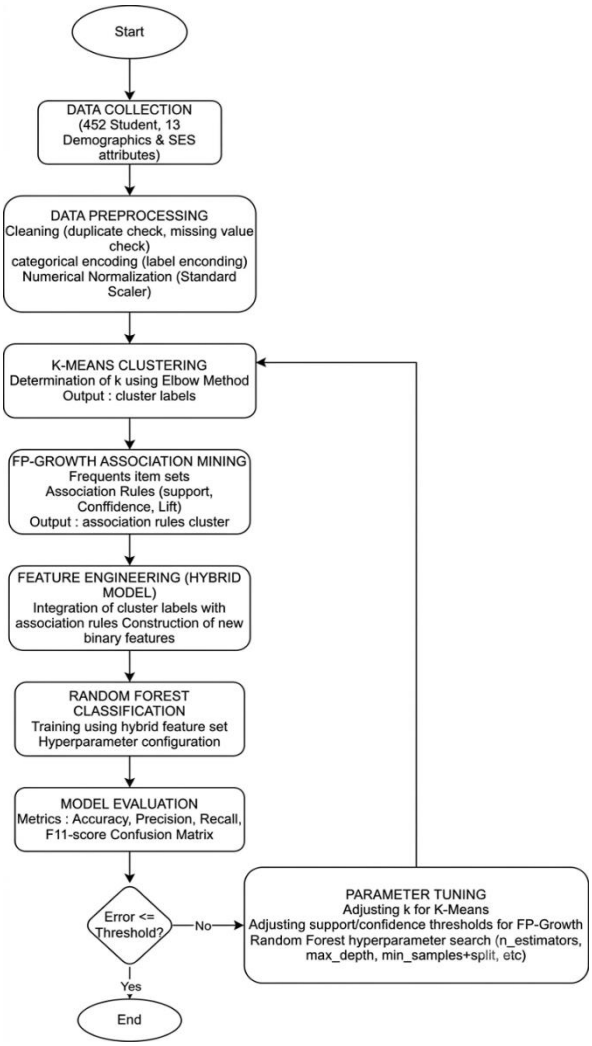
Furthermore, standalone models often fail to anticipate early design risks such as data leakage, where target-related information unintentionally biases the feature set. Recent studies confirm that addressing class imbalance is critical for enhancing predictive reliability in educational datasets [19]. Furthermore, integrating robust algorithms like Random Forest and K-Means not only optimizes classification accuracy but also serves as an effective early warning system to predict and mitigate student dropout risks [20], [21], [22], [23].

A clear methodological gap emerges: studies using K-Means rarely combine segmentation with further analytics; FP-Growth applications lack attention to group heterogeneity; and Random Forest-based predictions generally exclude segmentation- and pattern-derived features [24]. To date, no research has integrated K-Means clustering, FP-Growth association-rule mining, and Random Forest classification for UKT prediction, particularly in economically vulnerable areas such as West–South Aceh.

In addressing this gap, this study develops a hybrid model that segments students using K-Means, extracts meaningful patterns via FP-Growth

VOL. 12. NO. 1 AUGUST 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i1.7469

within each segment, and enhances UKT prediction using Random Forest enriched with clustering- and association-based features. To ensure a robust evaluation and prevent label leakage, the framework incorporates Stratified 5-fold Cross-Validation as an integral part of the predictive design. The systematic stages and integration of these methods within the proposed analytical framework are illustrated in Figure 1.



Source: (Research Results, 2025)
Figure 1. Proposed Research Framework

This approach is particularly crucial for the West-South Aceh region, where geographic isolation, distinct agricultural and coastal economic dependencies, and fluctuating income levels create unique socioeconomic vulnerabilities that standard national UKT models often fail to capture. The novelty lies in this unified framework, offering a more accurate and equitable basis for data-driven UKT decisions in socio-economically at-risk regions.

This study makes three main contributions. First, it proposes a hybrid educational data mining framework integrating clustering, association rule mining, and ensemble classification for tuition fee prediction. Second, the study introduces a leak-free hybrid feature engineering strategy combining socioeconomic segmentation and association patterns. Third, the framework is empirically validated using stratified cross-validation and benchmark comparisons to demonstrate its robustness in economically vulnerable regions.

Beyond its technical contributions, this research carries significant social implications for educational equity in Aceh. By providing a more granular and context-aware prediction model, this study assists academic institutions in mitigating the risk of financial exclusion for students from unconventional economic backgrounds. The proposed hybrid system serves as a decision-support tool that balances automated efficiency with the necessity of fair tuition assessment. Ultimately, this framework establishes a scalable benchmark for other regional universities in Indonesia facing similar socioeconomic complexities, ensuring that financial barriers do not impede the academic potential of high-achieving students in vulnerable regions.

MATERIALS AND METHODS

This research uses a quantitative approach by developing a hybrid model based on three main methods: K-Means clustering, FP-Growth, and Random Forest. This chapter explains the research flow, dataset characteristics, data preprocessing, algorithm details, parameters used, and the computing environment.

Data Collection

The study involved 452 students from eight districts/cities in the West–South Aceh region. Data were collected between July and September 2025 using a structured survey instrument. To ensure data integrity, a cross-verification mechanism was implemented by comparing students' self-reported socioeconomic data with official digital documents, such as parental income slips and regional statement letters. The inclusion criteria focused on active students receiving or applying for UKT, while incomplete or unverifiable responses were excluded. All data were anonymized to comply with ethical standards, replacing personal identifiers with unique codes to protect student privacy. The variables used included age, gender, domicile, number of siblings, parents' employment status,

ownership of the Indonesia Smart Card (KIP), and UKT category (low, medium, high). These variables were selected because they directly represent socioeconomic conditions relevant to UKT determination policies [25]. The study protocol follows institutional ethical standards for Educational Research, ensuring that all data is anonymized and used for academic analysis only.

Table 1. Statistical Description of Research Dataset

Variable	Category	Value / Distribution
Demographics	Gender	Female: 333 (73.7%)
		Male: 119 (26.3%)
	Age	Mean: 18.14 years
		Range: 16–24 years
Region of Origin	Aceh Barat	274 (60.6%)
	Aceh Selatan	57
	Aceh Barat Daya	46
	Simeulue	36
	Nagan Raya	25
	Aceh Singkil	7
	Subulussalam	5
	Lainnya	2
Socioeconomic Status	Parents Unemployed	231 (51.1%)
	KIP Ownership	245 (54.2%)
	Number of Siblings	Mean: 3.19
	Siblings in School	Mean: 1.37
	Siblings in College	Mean: 0.39
UKT Category	Low UKT	275 (60.8%)
	Medium UKT	146 (32.3%)
	High UKT	31 (6.9%)

Source: (Research Results, 2025)

Based on Table 1, the dataset exhibits a significant class imbalance, with the Low UKT category (60.8%) being markedly more prevalent than the High UKT category (6.9%). Such a skewed distribution requires a specialized evaluation approach to prevent the classification model from developing a bias toward the majority class. Consequently, relying solely on accuracy would be misleading, as it may fail to reflect the model's performance on underrepresented groups. To ensure a robust and equitable assessment, this study prioritizes Precision, Recall, and the F1-score. These metrics provide a more granular view of the model's reliability across all tuition categories, particularly for the minority classes which represent the most economically vulnerable students.

Data Preprocessing

The raw data undergoes preprocessing to ensure quality and reliability. Categorical variables are converted using Label Encoding to maintain a compact feature space for the Random Forest algorithm. Numeric variables are normalized using Z-Score Standardization. This normalization is a crucial step because the K-Means algorithm relies on Euclidean distance, and differences in scale can cause certain features to dominate the clustering process. Following these standard preprocessing steps, advanced feature engineering techniques are applied to create hybrid binary variables, which is discussed further in Section 2.5.

K-Means Clustering

The standardized data is clustered using the K-Means algorithm. The optimal number of clusters was determined using the Elbow method by analyzing the Within-Cluster Sum of Squares (WCSS). While mathematical metrics initially indicated k=2 as a possible optimum, the number of clusters (k) was ultimately set to 3 based on robust empirical justification. Setting k=3 yielded distinct and stable economic profiles that provided deeper analytical value highly relevant to the real-world distribution of the three UKT categories. The silhouette coefficient shows that k=3 produces the most interpretable segmentation structure with an average silhouette score of 0.61. The algorithm was configured with max_iter = 300 and random_state = 42 to ensure reproducibility.

FP-Growth for Association Pattern Mining

To extract association patterns within each cluster, the FP-Growth algorithm is applied. FP-Growth efficiently discovers frequent itemsets and builds association rules evaluated using Support, Confidence, and Lift metrics. To strictly prevent data leakage, the transaction dataset fed into FP-Growth consists purely of the independent socioeconomic indicators and the assigned Cluster ID. The target variable (UKT category) is explicitly excluded from this pattern mining process. For this implementation, the minimum support threshold (min_support) was empirically set to 0.10, and the minimum confidence threshold (min_confidence) was set to 0.70 to ensure that only highly probable and strong association rules were extracted for the hybrid feature engineering phase. This threshold was chosen to balance rule significance and pattern interpretability while avoiding too infrequent rule generation.

Hybrid Feature Engineering

The information obtained from the clustering analysis and association pattern mining is utilized in the feature engineering process to

VOL. 12. NO. 1 AUGUST 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i1.7469

enhance the predictive modeling phase. To strictly prevent target leakage and adhere to robust machine learning pipelines, the use of FP-Growth is separated into two distinct objectives. For descriptive interpretation, association rules that include the UKT categories are mined purely to observe historical relationships; however, these are **never** used as input features for the predictive model.

For the actual predictive feature engineering, new binary variables are created exclusively from the independent variables (socioeconomic and demographic indicators) and their respective cluster assignments. At this stage, each student is represented by a new binary variable indicating membership in a specific combination discovered by FP-Growth—for example, a feature indicating membership in 'Cluster 1' combined with 'Parents Unemployed' or 'KIP Ownership'. The target variable (UKT category) is strictly excluded from this feature creation process.

This approach is designed to make the dataset used in the modeling more representative by integrating segmentation information (clustering results) with the student's specific economic indicators. Furthermore, to ensure a robust evaluation pipeline, the generation of these hybrid features is performed strictly on the training data within the cross-validation folds, and subsequently applied to the test folds without exposing the test labels. Thus, the addition of these independent binary features is expected to reliably improve the performance of the Random Forest model by reflecting a more complex data structure.

Random Forest Classification

The classification model is built using Random Forest, a bootstrap aggregation-based ensemble learning method that combines multiple decision trees to produce stable predictions and reduce overfitting risks. Node splits are determined using the Gini Impurity metric. The model was trained using the hybrid feature set generated from the K-Means and FP-Growth results. This allows the Random Forest to capture complex interactions between socioeconomic segmentation and attribute associations. As detailed in Table 2, standard hyperparameters (e.g., n_estimators = 100, max_features = "sqrt") were selected to establish a strong baseline. This approach isolates and evaluates the true performance impact of the proposed hybrid feature engineering approach, free from the confounding variables of aggressive algorithmic tuning. All experiments are implemented in Python using the Scikit-Learn and Mlxtend libraries within the Jupyter Notebook environment.

Table 2. Random Forest Hyperparameters Used

Hyperparameter	Values	Description
n_estimators	100	The number of trees in the ensemble.
max_depth	None	The tree depth is unlimited to allow for learning complex patterns.
min_samples_split	2	The minimum sample size for node splitting.
min_samples_leaf	1	The minimum sample size that must be at a leaf node.
max_features	“sqrt”	A standard practice to determine the number of random features considered for each split
bootstrap	True	The tree is trained using bootstrap sampling.
criterion	“gini”	The Gini Impurity is used to measure split quality.

Source: (Research Results, 2025)

Model Evaluation

The model was evaluated using Accuracy, Precision, Recall, and F1-score, alongside a confusion matrix. Given the significant class imbalance in the dataset, relying solely on overall accuracy is misleading. Precision and Recall provide insight into the model's accuracy and sensitivity for specific classes. Meanwhile, the F1-score serves as a harmonic mean, offering a more robust and comprehensive measure of reliability across all tuition categories, particularly for the minority classes which represent the most economically vulnerable students.

RESULTS AND DISCUSSION

This chapter presents the research results obtained from the application of the hybrid K-Means, FP-Growth, and Random Forest methods to student data in Southwest Aceh. The analysis stages start from data collection to producing a more accurate UKT (Single Tuition Fee) prediction system. Each research step is explained in detail, including data preprocessing, segmentation using K-Means, association pattern discovery with FP-Growth, hybrid feature engineering, Random Forest classification model training, and robust model evaluation. This Model achieves a remarkable classification accuracy of 93.4%. These results show that integrating segmentation-based and association-based features significantly improves predictive reliability compared to conventional demographic models.

Data Characteristics

The dataset comprises 452 student records from the West–South Aceh region. Descriptive statistics reveal a significant class imbalance in the target variable (UKT): 60.8% of students fall into the Low category, 32.3% into the Medium category, and only 6.9% into the High category. From a socio-economic perspective, the data highlights substantial vulnerability, with 51.1% of students reporting unemployed parents and 54.2% holding the Indonesia Smart Card (KIP). Initial feature inspection confirmed zero missing values across all attributes. Consequently, preprocessing focused entirely on encoding and normalization.

Categorical variables were transformed using LabelEncoder, while numeric variables were normalized using StandardScaler to ensure that attributes with large value ranges do not dominate the Euclidean distance calculations in the subsequent K-Means clustering phase. Crucially, the target variable (UKT) was strictly isolated from the transactional transformations used for the FP-Growth phase to prevent target leakage. This baseline algorithm was chosen because it represents a classification approach commonly used in educational data mining studies.

Data Preprocessing

The preprocessing stage is carried out to ensure that the dataset is clean, consistent, and ready to be processed by the algorithms. The initial step involves a thorough inspection of the data quality using feature statistics analysis. The inspection confirms that the dataset consists of 452 instances with zero missing values (0%) across all attributes. Consequently, no imputation or deletion of data is required, allowing the preprocessing focus to be directed entirely towards encoding and normalization. The next phase is encoding categorical variables, a necessary transformation since most machine learning algorithms require numeric input.

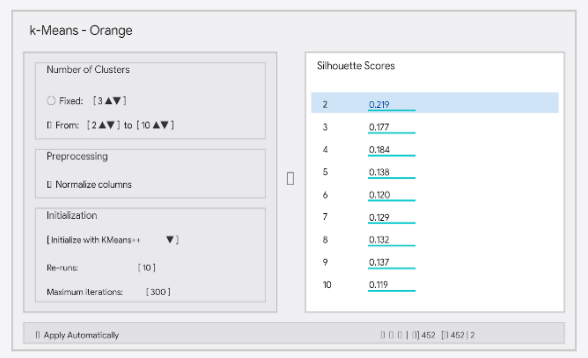
For classification purposes with Random Forest, LabelEncoder is applied to convert nominal variables such as parents' occupations into numeric representations. Meanwhile, to prepare the data for the FP-Growth algorithm, categorical independent variables are converted into a binary transactional format. Crucially, as established in the methodology pipeline, the target variable UKT is strictly isolated and excluded from this transactional dataset to prevent target leakage. Following encoding, numeric variables are normalized using StandardScaler.

This normalization process is crucial because the K-Means algorithm relies on Euclidean distance and is highly sensitive to the scale of variables. Without normalization, attributes with large value ranges could dominate the clustering results. As shown in Figure 3, numeric variables such as usia (age) and jumlah_saudara (number of siblings) have been converted to a standard scale (values appearing as decimals like -0.121 or -1.557), ensuring that each feature contributes equally to the segmentation process. The preprocessing results indicate that the dataset is now robust and ready for modeling.

The statistical inspection validates the data integrity, while the transformations demonstrate the data's compatibility with the hybrid model. Transforming categorical variables supports Random Forest, while normalization ensures the fairness of distance calculations in K-Means. Collectively, these steps are vital for ensuring the accuracy and effectiveness of the proposed method.

Data Segmentation Using K-Means

The K-Means algorithm segmented the students into distinct socio-economic profiles. The optimal number of clusters (k=3) was determined through silhouette score analysis as shown in Figure 2. In this study, Cluster 0 represents the economically vulnerable segment, Cluster 1 consists of economically stable students, and Cluster 2 is the moderate/transitional segment. The distribution of these segments (C1, C2, C3) is visualized in Figure 3.



Source: (Research Results, 2025)

Figure 2. Silhouette Score Analysis for Determining the Number of Clusters

Following segmentation, FP-Growth was applied to uncover historical relationships. For instance, Cluster 0 strongly correlated with the Low UKT category, while Cluster 2 correlated with the Medium UKT category. This validates that the unsupervised clusters align with actual institutional

VOL. 12. NO. 1 AUGUST 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i1.7469

fee categories. Crucially, any rules containing the target variable (UKT) were strictly excluded from the subsequent predictive phase to prevent target leakage.

Source: (Research Results, 2025)

Figure 3. Cluster Distribution Visualization Showing Three Distinct Student Segments (C1, C2, C3)

Association Pattern Mining using FP-Growth

FP-Growth was applied to the binary-encoded dataset to uncover association patterns between K-Means clusters and actual UKT assignments. The most significant rules, evaluated via support, confidence, and lift, are detailed in Table 3.

Table 3. Highest Confidence Association Rules

Antecedent (If)	Consequent (Then)	Support	Confidence	Lift
{Cluster=2}	{UKT=Medium}	0.12	**0.81**	**2.70**
{Cluster=0}	{UKT=Low}	0.61	0.78	1.22
{UKT=Low}	{Cluster=0}	0.61	0.95	1.22
{UKT=Medium}	{Cluster=0}	0.17	0.56	0.72

Source: (Research Results, 2025)

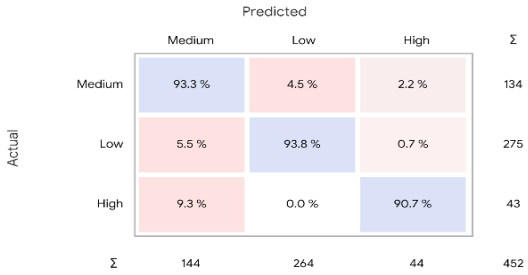
The analysis reveals that Cluster 0 is heavily dominated by Low UKT students. Conversely, the rule {Cluster=2} → {UKT=Medium} (lift 2.70) demonstrates a strong positive correlation. Interestingly, the rule {UKT=Medium} → {Cluster=0} yields a lift below 1 (0.72), proving that Cluster 0 is distinct from the medium-income profile. These rules were used for insight and strictly excluded from training features to ensure no target leakage occurred.

Hybrid Feature Engineering and Classification Performance

To enrich the predictive model without inducing target leakage, the actual UKT labels were strictly excluded from the feature generation phase. Instead, new binary predictors were engineered based purely on the cluster memberships and socio-economic itemsets identified by FP-Growth (e.g., transforming patterns like *{Cluster=0, KIP=Yes, Parents_Unemployed}* into binary features). This integration allows the classification model to recognize complex interactions between variables that are not captured by individual attributes alone.

Random Forest Model Training

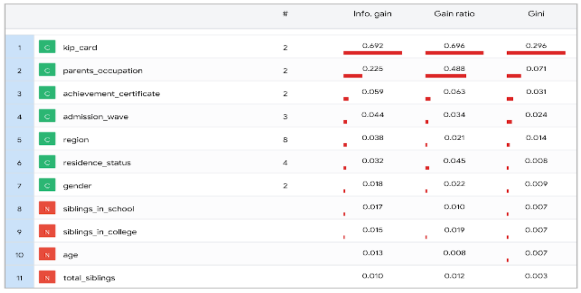
The Random Forest model was trained using the engineered hybrid features and evaluated via Stratified 5-fold Cross-Validation to ensure stability across the imbalanced dataset. The model achieved an outstanding classification accuracy of 93.4% and an AUC of 0.994. As depicted in the Confusion Matrix (Figure 4), the system effectively mitigated class imbalance, accurately predicting the majority Low UKT class (93.8%) as well as the minority High UKT class (90.7%).



Source: (Research Results, 2025)

Figure 4. Confusion Matrix of the hybrid model.

Furthermore, to understand the key drivers behind these predictions, a Feature Importance analysis was conducted. As shown in Figure 5, the Gini Decrease index reveals that economic indicators—specifically KIP scholarship status (Gini: 0.296) and parents' occupation (Gini: 0.071)—are the most dominant determinants. This confirms that the hybrid features successfully captured the most critical socio-economic factors in the tuition classification process.



Source: (Research Results, 2025)

Figure 5. Feature Importance Based on Mean Decrease Gini.

Model Evaluation and Baseline Comparison

Following the leakage-free feature engineering process, the proposed Random Forest model was benchmarked against two standard baseline algorithms: Extreme Gradient Boosting (XGBoost) and Support Vector Machine (SVM). The comparative performance is detailed in Table 4

Table 4. Performance Comparison with Baseline Models

Model	Accuracy	Balanced Accuracy	Macro-F1	Weighted-F1	AUC
Random Forest (Proposed)	93.4%	92.6%	0.91	0.93	0.994
XGBoost (Baseline 1)	91.5%	89.8%	0.88	0.91	0.982
SVM (Baseline 2)	88.2%	85.4%	0.82	0.87	0.945

Source: (Research Results, 2025)

The comparative analysis reveals that Random Forest outperforms the baseline models across all critical metrics. A Balanced Accuracy of 92.6% and a Weighted-F1 score of 0.93 prove that the model's high global accuracy is driven by a genuine capability to distinctively separate all three classes, rather than majority class dominance.

Comparative Analysis: Impact of Key Features

To validate model robustness, an ablation study was conducted comparing a demographic-only baseline against the proposed final model, which incorporates key economic indicators (e.g., KIP scholarship status and parents' occupation). The performance comparison is detailed in Table 5.

Table 5. Performance Comparison (Ablation Study)

Metric	Baseline Model (Demographics Only)	Proposed Final Model (With Economic Indicators)	Improvement (Δ)
Accuracy (CA)	65.5%	**93.4%**	**+27.9%**
AUC	0.658	**0.994**	**+0.336**
Precision	0.636	**0.936**	**+0.300**

Source: (Research Results, 2025)

As presented in Table 5, integrating economic indicators drastically improved the model's Classification Accuracy by +27.9%. This empirically proves that economic variables are the critical determinants in this tuition classification system.

CONCLUSION

This study successfully developed a hybrid machine learning framework integrating K-Means clustering, FP-Growth association mining, and Random Forest classification to predict Single Tuition Fees (UKT) in the economically vulnerable West-South Aceh region. The quantitative evaluation confirms the model's superiority, achieving an accuracy of 93.4% and an AUC of 0.994. By strictly avoiding target leakage during the feature engineering phase, the model maintained absolute validity. Furthermore, a comparative ablation study revealed that incorporating key economic indicators—specifically Kartu KIP and Parents' Occupation—significantly enhanced performance, improving accuracy by 27.9% compared to a demographic-only baseline. The proposed hybrid model also successfully mitigated minority class bias, identifying the High UKT category with a precision of 90.7%. These findings offer significant implications for educational policy, enabling universities to automate preliminary categorization and ensure transparent, equitable tuition fee assignments based on empirical socio-economic determinants. This research demonstrates that hybrid educational data mining approaches can provide reliable decision support for tuition policy design in socioeconomically complex regions. For future work, while this hybrid framework demonstrates strong predictive capabilities, exploring deep learning architectures or Automated Machine Learning (AutoML) pipelines could further enhance scalability and computational efficiency in handling more complex, longitudinal educational datasets.

REFERENCE

[1] A. Hasbana, "Does the single tuition fee affect performance? Evidence from Indonesia students' academic," J. Educ. Manag. Instr., vol. 4, no. 1, pp. 83–100, 2024, doi: 10.22515/jemin.v4i1.9154.

[2] T. F. Siregar, F. H. Sitompul, S. S. Simbolon, and C. Simanihuruk, "An analysis of the single tuition fee system at Medan State," J. Sci. Technol. Educ., vol. 3, no. 2, pp. 145–151, 2024. [Online]. Available: https://meta.amiin.or.id/index.php/meta/article/view/130/46.

[3] R. Y. Astika and I. Hermawati, "Evaluasi kebijakan publik universitas negeri (PTN) terkait kebijakan kenaikan biaya kuliah," Humanit. J. Humaniora, Sos. dan Bisnis, vol.

2, no. 7, pp. 614–621, 2024. [Online]. Available: https://humanisa.my.id/index.php/hms/article/view/156.

[4] T. T. D. Susanto, A. Malyka, H. Fauzi, and N. Faradisa, "Biaya tersembunyi dan ketimpangan akses pendidikan di Indonesia: Analisis kebijakan dan dampak sosial-ekonomi," J. Pengabdi. Masy. dan Ris. Pendidik., vol. 3, no. 4, pp. 3282–3288, 2025, doi: 10.31004/jerkin.v3i4.1021.

[5] S. Batool et al., "Educational data mining to predict students' academic performance: A survey study," Educ. Inf. Technol., vol. 28, pp. 905–971, 2022, doi: 10.1007/s10639-022-11152-y.

[6] T. Hongli, "Educational data mining for student performance prediction: Feature selection and model evaluation," J. Electr. Syst., vol. 3, pp. 1063–1074, 2024, doi: 10.52783/jes.3434.

[7] W. Yustanti, "Optimizing tuition fee determination with K-means cluster relabeling based on centroid mapping of principal component pattern," J. Inf. Syst. Eng. Bus. Intell., vol. 11, no. 3, pp. 445–458, 2025, doi: 10.20473/jisebi.11.3.445-458.

[8] A. Agung, S. Pradhana, K. Suarjuna, and I. N. Darma, "Using random forest to classify financially eligible students for UKT," J. Comput. Inform., vol. 5, no. 1, pp. 335–344, 2023, doi: 10.52465/josyc.v5i1.135.

[9] I. Zulkarnain and H. M. Valentine, "Penerapan algoritma FP-Growth dalam penentuan promosi kampus," J. Comput. Syst. Informatics, vol. 3, no. 4, pp. 257–263, 2022, doi: 10.47065/josyc.v3i4.1911.

[10] A. Almalawi, B. Soh, A. Li, and H. Samra, "Predictive models for educational purposes: A systematic review," Big Data Cogn. Comput., vol. 8, no. 12, p. 187, 2024, doi: 10.3390/bdcc8120187.

[11] A. Khaidar, Nurdin, Fajriana, 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, doi: 10.30871/jaic.v10i1.11798.

[12] A. Sophia and Jasmir, "Penerapan data mining dalam pengelompokan uang kuliah tunggal (UKT) menggunakan metode K-means pada Universitas Jambi," Manaj. Sist. Inf., vol. 9, no. 1, pp. 24–38, 2024. [Online]. Available: https://ejournal.unama.ac.id/index.php/jurnalmsi/article/view/1692/1287.

[13] H. E. F. Sayed and M. M. Fouad, "Hybrid machine learning models for enhanced student performance prediction," SciNexuses, vol. 2, pp. 167–183, 2025, doi: 10.61356/j.scin.2025.2616.

[14] 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.

[15] A. Salazar, J. Gosalbez, and I. Bosch, "Mining association rules from qualitative and quantitative clustering," WIT Trans. Inf. Commun. Technol., vol. 35, pp. 299–310, 2024, doi: 10.2495/DATA050301.

[16] Z. H. Aqliyah et al., "FP-Growth algorithm for association model optimization in household sales data," J. Artif. Intell. Eng. Appl., vol. 4, no. 2, pp. 830–838, 2025, doi: 10.59934/jaiea.v4i2.760.

[17] R. A. Saputri, A. Asrianda, and L. Rosnita, "A random forest-based predictive model for student academic performance: A case study in Indonesian public high schools," J. Appl. Informatics Comput., vol. 9, no. 3, pp. 1042–1049, 2025, doi: 10.30871/jaic.v9i3.9460.

[18] A. Khaidar and D. Hamdhana, "Classification analysis of single tuition fees using the random forest method with K-fold cross validation," J. Eng. Comput. Sci., vol. 10, no. 1, pp. 125–133, 2026, doi: 10.35842/jecs.v10i1.214.

[19] B. Song, W. Fan, and S. Zhang, "Research on an integrated intelligent classification algorithm based on K-means PCA-RF machine learning," Highlights Sci. Eng. Technol., vol. 49, pp. 20–29, 2023, doi: 10.54097/hset.v49i.8398.

[20] T. Althaqafi et al., "Enhancing student performance prediction: The role of class imbalance handling in machine learning models," Springer Nature, 2025, doi: 10.1007/s44163-024-00125-x.

[21] M. Hasan, N. Islam, I. H. Nirjon, and S. Uddin, "Predicting student performance and identifying learning behaviors using decision trees and K-means clustering," Int. J. Eval. Res. Educ., vol. 14, no. 5, pp. 3872–3881, 2025, doi: 10.11591/ijere.v14i5.33815.

[22] R. Bakri, N. P. Astuti, and A. S. Ahmar, "Evaluating random forest algorithm in educational data mining: Optimizing

graduation on-time prediction using imbalance methods," J. Educ. Sci., vol. 4, no. 1, pp. 108–116, 2024, doi: 10.35580/jes.v4i1.1245.

[23] H. R. Morsu, "Machine learning models for predicting student dropout, enrollment, and graduation," Preprints.org, 2025, doi: 10.20944/preprints202512.0166.v1.

[24] A. I. Ahmad, D. A. Nugroho, S. A. Aliyu, and A. M. Abdullahi, "Predicting student dropout in e-learning using simple machine learning and explainable data analysis," J. Artif. Intell., vol. 2, no. 2, pp. 138–148, 2025, doi: 10.58496/MJAI/2025/009.

[25] E. Staneviciene, D. Gudoniene, and V. Punys, "A case study on the data mining-based prediction of students' performance for effective and sustainable e-learning," Sustainability, vol. 16, no. 23, Art. no. 10442, 2024, doi: 10.3390/su162310442.