Jurnal Ilmiah Edutic : Pendidikan dan Informatika /Vol.12, No.2 : 2025

ISSN 2407-4489
pp. 130-140

Optimizing UKT Prediction Based on Socio-Economic Features: A Multimodel Evaluation with Feature Selection Srategies

Windy Chikita Cornia Putri[a], Wiyli Yustanti[b], Ervin Yohannes[c]

[a,b,c] *Engineering Faculty, Universitas Negeri Surabaya, Surabaya, Indonesia*
email: : [a] windychikita@unesa.ac.id, [b] wiyliyustanti@unesa.ac.id, [c] ervinyohannes@unesa.ac.id

Abstract

Determining the tuition fee group (UKT) for new students in Indonesian public universities represents a complex challenge that demands an equitable and data-driven decision approach. This study proposes an integrative feature-selection framework that combines five popular techniques Chi-Square, Recursive Feature Elimination (RFE), LASSO Regression, Random Forest Importance, and Exploratory Factor Analysis (EFA) to identify the most influential attributes from 53 socioeconomic variables of applicants at Universitas Negeri Surabaya. As its key novelty, the study extracts intersecting features consistently selected across all five methods and evaluates their impact on the performance of five classification algorithms: Support Vector Machine (SVM), Decision Tree, Random Forest, K-Nearest Neighbor (KNN), and Naïve Bayes. Experimental results show consistent performance improvements across all models after reducing the feature set from 53 to 16 attributes. Accuracy increases were recorded for SVM (0.7550 → 0.7810), Decision Tree (0.6334 → 0.6702), Random Forest (0.5941 → 0.7123), KNN (0.6815 → 0.6994), and Naïve Bayes (0.6430 → 0.6643), accompanied by parallel gains in precision, recall, and F1-score. These findings demonstrate that integrative feature selection not only enhances predictive accuracy but also reduces data complexity, providing a replicable methodological contribution for transparent and adaptive socioeconomic-based classification systems in higher education..

Keywords: Feature selection, UKT prediction, Socioeconomic data.

ISSN 2407-4489
pp. 130-140

INTRODUCTION

The classification of tuition groups under the Single Tuition Fee (UKT) scheme for new students in Indonesian public universities (PTN) plays a pivotal role in supporting equity-oriented higher education financing. Although the system is formally regulated through Ministry of Education and Culture Regulation No. 25/2020, its implementation still faces persistent challenges in ensuring transparency, consistency, and accuracy. Many institutions including Universitas Negeri Surabaya (UNESA) continue to rely on manual verification of documents such as salary slips, electricity bills, and property tax, which often fail to represent families' true socio-economic conditions (Siregar & Darma, 2021). These limitations highlight the need for systematic, data-driven approaches capable of processing complex socio-economic variables more objectively

Recent studies have explored machine-learning-based UKT classification using algorithms such as Decision Tree, K-Nearest Neighbor, and Support Vector Machine (Syam et al., 2024; Yustanti & Rochmawati, 2022). However, most of these works employ only a single feature selection technique, which creates risks of redundancy, unstable model performance, and the exclusion of important variables. At UNESA, the UKT classification process involves more than 50 socio-economic attributes, underscoring the urgency of identifying the most relevant and stable features. As the institution transitions toward Legal Entity State University (PTN-BH) status, the demand for fair, accountable, and analytically transparent financial governance becomes increasingly critical.

To address these gaps, this study proposes an integrative feature-selection framework combining five widely used techniques Chi-Square, Recursive Feature Elimination (RFE), LASSO Regression, Random Forest Importance, and Exploratory Factor Analysis (EFA). Unlike previous works that rely on a single method (Yustanti, 2021; 2022), this study identifies intersection features consistently selected across all techniques and evaluates their effect on five classification algorithms: Support Vector Machine (SVM), Decision Tree, Random Forest, K-Nearest Neighbor (KNN), and Naïve Bayes. By emphasizing methodological integration and stability, the study contributes a more accurate, interpretable, and replicable approach to UKT classification supporting the broader agenda of data-driven governance and AI-enabled decision-making in Indonesian higher education.

METHOD

This study adopts the Knowledge Discovery in Databases (KDD) approach, as its systematic pipeline is well-suited for extracting knowledge from large and complex datasets, particularly those involving students' socio-economic information. Contemporary research in the fields of educational data mining and learning analytics has explicitly emphasized that the KDD process which encompasses the stages of Selection, Preprocessing, Transformation, Data Mining, and Interpretation/Evaluation provides a highly structured framework for developing classification models that are both transparent and reliable (Romero & Ventura, 2024; Lin et al., 2023).

ISSN 2407-4489
pp. 130-140

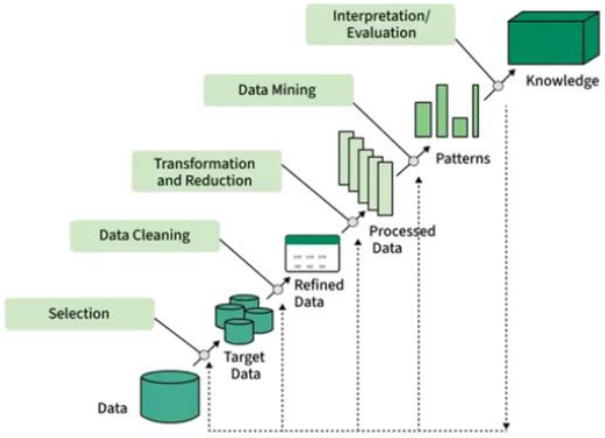


Fig 1. Knowledge Discovery in Databases (KDD) Process (source: GeeksforGeeks, 2025)

2.1 Selection (Data Selection)

In this stage, the dataset is sourced from prospective student admissions at Universitas Negeri Surabaya, consisting of 53 socio-economic attributes. These features include information on family dependents, asset ownership, household expenditures, housing facilities, and parental income. The target label for classification is the UKT (Single Tuition Fee) group, categorized from K1 to K8. Attribute selection is based on completeness, contextual economic relevance, and its relationship to the dependent variable.

2.2 Preprocessing (Data Preprocessing)

Preprocessing represents a critical early step in the Knowledge Discovery in Databases (KDD) pipeline, aimed at ensuring data quality prior to modeling or classification. In this study, several preprocessing procedures were applied systematically. First, missing values were addressed using a combination of deletion and imputation strategies: records with incomplete key socioeconomic variables such as parental income or home ownership were removed to avoid bias, while non-critical attributes with low missingness were imputed using the mode for categorical variables. Second, numerical features such as income, electricity capacity, and land/building size were normalized using min–max scaling to ensure that differences in magnitude did not distort algorithmic learning, particularly for distance-based models like KNN and SVM. Third, categorical variables such as parental occupation, home ownership, and vehicle type were encoded using one-hot encoding to convert them into machine-readable binary representations without imposing ordinal relationships. The following figure summarizes the preprocessing stages applied in this study.

ISSN 2407-4489
pp. 130-140

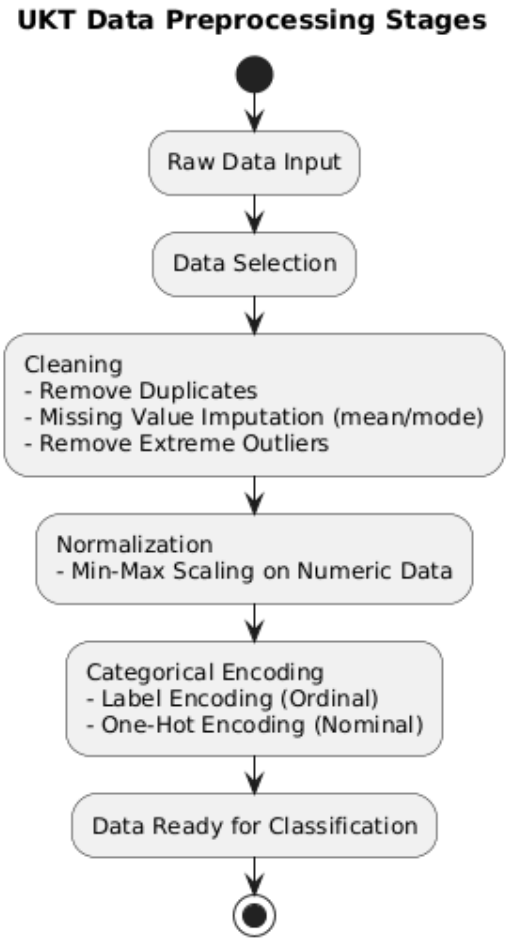


Fig 2. UKT Data Preprocessing Stages

2.3 Feature Transformation

In the feature transformation stage, five feature selection techniques were employed to extract the most relevant socioeconomic attributes for UKT classification. The Chi-Square method measures the association between categorical features and the target class, making it particularly suitable for discrete socioeconomic variables (Verma et al., 2025). Recursive Feature Elimination (RFE) iteratively removes the least contributing features using an estimator such as SVM and has been proven effective, although it requires higher computational time (Bulut, 2025). LASSO Regression offers automatic feature selection based on L1 regularization, promoting model simplicity and reducing the risk of overfitting (Ahmad et al., 2025).

These three approaches collectively represent the filter, wrapper, and embedded methods in feature selection, and complement one another in constructing an optimal subset of features that enhances classification performance.

ISSN 2407-4489
pp. 130-140

Table 1. Feature Selection Techniques

No	Feature Selection Technique	Process Flow	Formula / Basis
1	Chi-Square (χ^2)	Tests the independence between categorical features and the target label, then ranks features based on the highest χ^2 statistics.	$\chi^2 = \Sigma ((O_{ij} - E_{ij})^2 / E_{ij})$, where O_{ij} is the observed frequency and E_{ij} is the expected frequency.
2	Recursive Feature Elimination (RFE)	Uses an estimator (e.g., SVM) to iteratively eliminate the least contributing features until the optimal number of features is reached.	No fixed formula; relies on model coefficients (e.g., SVM weights). In each iteration, the feature with the smallest weight is removed.
3	LASSO Regression (L1)	Applies an L1 penalty to the regression model, forcing the coefficients of irrelevant features to become zero, effectively eliminating them.	$\min w \{\sum(y_i - X_i w)^2 + \lambda \sum$
4	Random Forest Importance	Trains a Random Forest model and computes the Gini Importance of each feature based on its contribution to reducing impurity at tree decision nodes.	$GI(f) = \Sigma t \Sigma n \in N_t\ p(n) \cdot \Delta i(n)$, where $p(n)$ is the proportion of samples at node *n*, and $\Delta i(n)$ is the impurity decrease at that node.
5	Exploratory Factor Analysis (EFA)	Extracts latent factors from feature correlations, then selects representative features from the main factors (eigenvalue > 1).	Based on Eigen decomposition of the correlation matrix: $R = \Lambda\Lambda^T + \Psi$

2.4 Data Mining (Classification)

The classification stage represents the core of the Knowledge Discovery in Databases (KDD) process, as it is at this point that hidden patterns within the data are uncovered through the application of machine learning algorithms. In the context of this study, classification is employed to predict the Tuition Fee Group (UKT) for incoming university students based on previously processed and selected socio-economic attributes. Five supervised classification algorithms are utilized: Support Vector Machine (SVM), Decision Tree, Random Forest, K-Nearest Neighbor (KNN), and Naïve Bayes. These algorithms were chosen due to their widespread representation in literature and the diversity of their underlying characteristics ranging from linear to non-linear approaches, interpretability, ensemble learning, and probabilistic modeling.

Performance evaluation is conducted using an 80:20 train-test split, along with stratified sampling to preserve the proportional distribution of UKT class labels. Furthermore, k-fold cross-validation (with k = 5) is employed to ensure empirical generalization and model stability, in line with best practices in modern classification research (Hooda & Kalra, 2023; Ibrahim et al., 2022).

ISSN 2407-4489
pp. 130-140

Table 2. Classification Methods

No	Classification Technique	Approach	Formula
1	Support Vector Machine (SVM)	Margin-based classifier with Radial Basis Function (RBF) kernel, suitable for non-linear data.	$\min_{w,b,\xi} \ \frac{1}{2} \|w\|^2 + C \sum \xi_i$ subject to: $y_i(w^T\phi(x_i)+b) \geq 1-\xi_i,\ \xi_i \geq 0$
2	Decision Tree (CART)	Tree-based learning that splits data based on impurity (e.g., Gini or Entropy).	Gini: $1 - \sum p_i^2$ Entropy: $- \sum p_i \log_2(p_i)$
3	Random Forest	Ensemble learning based on multiple decision trees built from random subsets of features/data.	$\hat{f}(x) = \frac{1}{T} \sum_{t=1}^{T} f_t(x)$
4	K-Nearest Neighbor (KNN)	Lazy learner; classification based on majority vote from nearest neighbors in feature space.	$d(x, x_i) = \sqrt{\sum (x_j - x_{ij})^2}$
5	Naïve Bayes	Probabilistic model assuming feature independence.	$P(C_k

2.5 Interpretation and Evaluation

To assess the effectiveness of the classification models in predicting the *Uang Kuliah Tunggal* (UKT) group, performance evaluation was conducted using four widely adopted metrics in multiclass classification settings. These metrics not only provide an overview of overall accuracy but also evaluate the balance between precision, recall (sensitivity), and their harmonic mean through the F1-score.

The macro-averaging approach was employed to ensure that each UKT class receives equal weight in the evaluation process, regardless of the distribution of data across the classes. This strategy is particularly important in education-related datasets where class imbalance may exist due to socioeconomic disparities.

The following is a description of the evaluation metrics used to measure classification performance:

a. Accuracy: The ratio of correctly predicted instances to the total number of instances.
b. Precision: The proportion of true positive predictions among all positive predictions for each class.
c. Recall: The proportion of true positive predictions among all actual positive instances for each class.
d. F1-Score: The harmonic mean of precision and recall, providing a single score that balances both.

These metrics collectively offer a comprehensive view of model performance, especially when combined with stratified sampling and k-fold cross-validation to ensure robust generalization.

ISSN 2407-4489
pp. 130-140

Table 3. Evaluation Metrics for Classification Model Performance

No.	Evaluation Metric	Formula	Description
1.	Accuracy	Accuracy = (TP + TN) / (TP + TN + FP + FN)	Measures the proportion of correct predictions over the total number of instances. Suitable when class distribution is balanced.
2.	Precision(Macro Average)	Precision = (1/N) Σ (TP_i / (TP_i + FP_i))	Assesses how precise the model is for each class by minimizing false positives. Useful when the cost of false positives is high.
3.	Recall(Macro Average)	Recall = (1/N) Σ (TP_i / (TP_i + FN_i))	Measures the model's ability to capture all true postives for each class. Important when false negatives carry significant consequences.
4.	F1-Score(Macro Average)	F1 = 2 × (Precision × Recall) / (Precision + Recall)	The harmonic mean of precision and recall. Ideal when a balance between precision and recall is crucial.

RESULTS AND DISCUSSION

A total of 16 selected features were identified as the intersection of the five feature selection methods employed in this study. These features represent the most consistently significant predictors across Chi-Square, Recursive Feature Elimination (RFE), LASSO Regression, Random Forest Importance, and Exploratory Factor Analysis (EFA). Selecting intersection features strengthens the statistical validity of the model by ensuring that the retained attributes are not merely relevant under a single criterion but demonstrate cross-method stability, robustness, and multivariate significance. This approach reduces the risk of model overfitting caused by method-specific biases such as Chi-Square's sensitivity to categorical distributions or LASSO's tendency to shrink correlated variables while preserving features that repeatedly show strong discriminative power.

From a socioeconomic standpoint, the retained features align well with theoretical expectations and empirical findings in poverty and welfare assessment research. Father's and mother's income emerged as primary indicators of household welfare, while their occupations provide additional insight into long-term financial security and income regularity. The number of family dependents captures the household's economic burden, and residential characteristics such as home ownership, flooring type, and electricity capacity function as reliable proxies for living standards. Indicators of asset ownership including vehicles, electronic appliances, and land or building area further reflect purchasing power and accumulated wealth. The recurrence of these attributes across all five selection techniques reinforces their statistical and conceptual relevance.

Following the feature transformation and selection process, model performance was evaluated under two conditions: using the full set of 53 socioeconomic features and using only the 16 intersection features. This two-stage evaluation aims to quantify the effect of feature reduction on efficiency, interpretability, and predictive accuracy. Five supervised classification algorithms Support Vector Machine (SVM), Decision Tree, Random Forest, K-Nearest Neighbor (KNN), and Naïve Bayes were tested under both scenarios using macro-averaged performance metrics (accuracy, precision, recall, and F1-score). The comparative analysis enables an assessment of whether the statistically grounded feature reduction enhances model generalization while minimizing complexity.

ISSN 2407-4489
pp. 130-140

Table 4. Classification Model Performance Results

No	Algorithm	Number of Features	Accuracy	Precision	Recall	F1-Score
1	SVM	53 (all)	0.7550	0.7412	0.7261	0.7234
		16 (selected)	0.7810	0.7635	0.7511	0.7542
2	Decision Tree	53 (all)	0.6334	0.6422	0.6230	0.6251
		16 (selected)	0.6702	0.6594	0.6463	0.6501
3	Random Forest	53 (all)	0.5941	0.5920	0.5795	0.5810
		16 (selected)	0.7123	0.7001	0.6935	0.6952
4	K-Nearest Neighbor (KNN)	53 (all)	0.6815	0.6670	0.6591	0.6612
		16 (selected)	0.6994	0.6851	0.6778	0.6812
5	Naïve Bayes	53 (all)	0.6430	0.6521	0.6404	0.6451
		16 (selected)	0.6643	0.6610	0.6512	0.6537

The findings from the classification results demonstrate that the integrative feature selection approach significantly improves model performance, particularly for the Support Vector Machine (SVM) and Random Forest algorithms. The increase in SVM accuracy from 0.7550 to 0.7810, and in Random Forest from 0.5941 to 0.7123, reflects the successful identification of features that are truly relevant to the classification of new students' tuition categories (UKT). This approach addresses the issues of feature redundancy and irrelevance in high-dimensional data, which often lead to overfitting or reduced model performance.

Moreover, reducing the number of features from 53 to 16 substantially shortens training time, making the model more efficient and practical for real-world applications, such as automated UKT classification systems in universities. These results align with previous studies such as Wiyli Yustanti et al. (2022), which emphasized the importance of feature optimization in improving the accuracy of socioeconomic-based classification systems within higher education institutions.

Compared to other studies that employed single-method feature selection, this study offers a novel contribution by integrating five feature selection techniques and identifying their intersection as a stable and consistent set of attributes. This methodological innovation enhances both the robustness and interpretability of the classification model.

ISSN 2407-4489
pp. 130-140

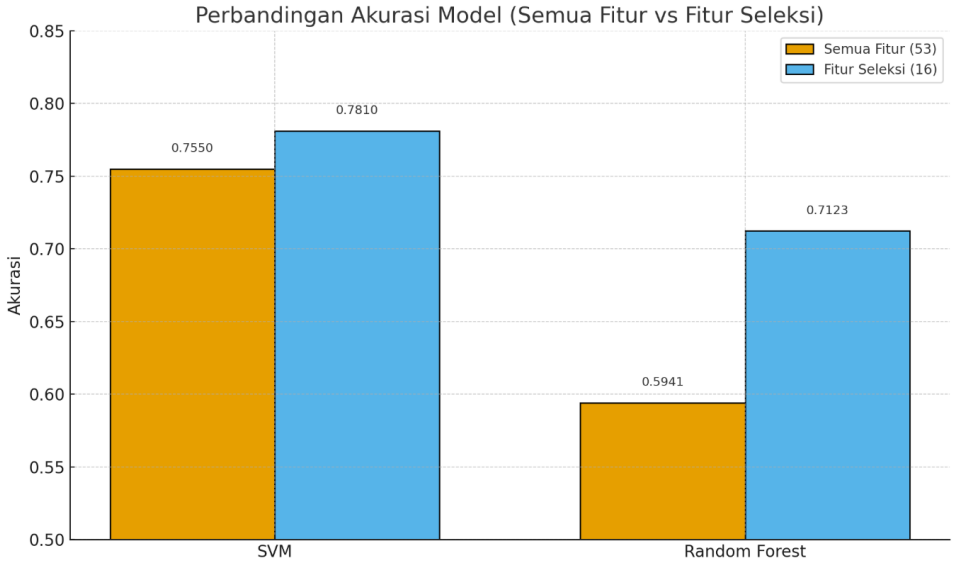


Fig 3. Comparison of Model Accuracy (All Features vs. Selected Features)

The comparison graph reveals that the intersection-based feature selection significantly enhances model accuracy, particularly for Random Forest, which improved from 0.5941 to 0.7123, and for SVM, which increased from 0.7550 to 0.7810. This improvement indicates that feature simplification not only reduces model complexity but also strengthens learning efficiency by eliminating noisy, redundant, or weakly informative variables. More importantly, the findings highlight why certain features such as parental income, parental occupation, home ownership, electricity capacity, and asset ownership exert a substantial influence on model performance. These variables represent stable socioeconomic indicators that directly reflect a household's financial capacity, making them strong discriminators across UKT categories. For instance, parental income and occupation provide consistent long-term economic information, while housing conditions and asset ownership serve as validated proxies for wealth and living standards in household welfare research. Their consistent selection across all five feature selection methods further supports their statistical relevance and robustness.

In comparison to previous studies, the present findings offer a methodological advantage. Prior works on UKT classification, such as Syam et al. (2024) and Yustanti & Rochmawati (2022), commonly relied on a single feature selection approach or expert judgment, which limited the stability of their selected attributes and led to model performances highly dependent on one algorithm or dataset. By contrast, the integrative intersection-based approach used in this study ensures cross-method agreement, resulting in a more reliable and generalizable feature subset. The consistent performance improvements across multiple algorithms demonstrate that the selected features not only enhance accuracy but also address the shortcomings identified in earlier studies particularly issues of redundancy, unstable feature relevance, and limited interpretability. Thus, the results provide stronger empirical evidence that multi-method feature selection offers a more rigorous foundation for developing transparent and adaptive UKT classification systems.

CONCLUSIONS

Based on the findings of this study, the integrative feature selection approach has been shown to significantly enhance the efficiency and predictive performance of UKT classification models for new students at Universitas Negeri Surabaya. By combining five complementary feature selection techniques Chi-Square, Recursive Feature Elimination (RFE), LASSO Regression, Random Forest Importance, and

Exploratory Factor Analysis (EFA) the study successfully identified 16 intersection features that consistently contributed to model performance. This reduction from 53 variables not only simplified the data structure but also improved the accuracy of several algorithms, particularly Support Vector Machine and Random Forest, demonstrating the value of selecting features that exhibit both statistical stability and socioeconomic relevance.

More broadly, the study highlights the potential of machine learning to support transparent, data-driven financial governance in higher education. The integrative methodology presented here provides a structured foundation for developing UKT classification systems that are more objective, interpretable, and adaptable to policy needs.

Nonetheless, several limitations must be acknowledged. First, the dataset originates from a single university context, which introduces the possibility of institutional or demographic bias and limits the generalizability of the findings to other PTNs with different socioeconomic landscapes. Second, the analysis primarily focused on conventional performance metrics accuracy, precision, recall, and F1-score without incorporating fairness assessments that are increasingly critical in socioeconomic classification tasks. Third, the study did not evaluate the temporal stability of the selected features across multiple admission periods, leaving open questions about the long-term consistency of the model. These limitations suggest that caution is needed before applying the model directly to broader institutional settings.

Future research should prioritize validating this approach using multi-campus or nationally representative datasets to assess generalizability and potential bias. Incorporating fairness-oriented evaluation metrics will also be crucial to ensuring that classification outcomes are equitable across diverse student groups. Further exploration of explainable AI (XAI) and recommendation-based decision-support systems may enhance the transparency and policy relevance of UKT determination processes, supporting more accountable and student-centered financial governance.

ACKNOWLEDGMENT

With profound respect and gratitude, the author extends sincere thanks to Universitas Negeri Surabaya (UNESA) for the invaluable support and research facilities provided throughout the course of this study. The author also wishes to express appreciation to the Directorate of Information and Communication Technology (DTIK) UNESA for the provision of data and system access, which greatly facilitated the data collection and analysis processes.

Special thanks are extended to Mrs. Wiyli Yustanti and Mr. Ervin Yohanes for their academic guidance and invaluable scientific contributions during the preparation of this article. The author is also grateful to all reviewers, fellow researchers, and other parties who have provided constructive feedback, technical support, and encouragement throughout the manuscript writing process.

It is the author's sincere hope that this research will contribute meaningfully to the advancement of data-driven policy development in higher education.

REFERENCES

Ahmad, S., Bashir, S., & Khan, M. A. (2025). Classification-based explainable AI approach for cancer diagnosis using LASSO feature selection. arXiv preprint arXiv:2503.16577. https://arxiv.org/abs/2503.16577

Bulut, M. (2025). Hybrid students' academic performance and dropout prediction models using recursive feature elimination. Information, 16(6), 476. https://www.mdpi.com/2078-2489/16/6/476

GeeksforGeeks. (2025, January 28). KDD process [Gambar]. Dalam GeeksforGeeks. https://media.geeksforgeeks.org/wp-content/uploads/20250128111402276995/kdd_process_660.webp

(Diakses 27 September 2025)

Hooda, R., & Kalra, S. (2023). Application of machine learning classification algorithms for student dropout prediction. International Journal of Educational Technology in Higher Education, 20(1), 15. https://doi.org/10.1186/s41239-023-00389-w

Ibrahim, A., Hassan, R., & Ali, N. (2022). A comparative study of classification algorithms for student academic performance prediction. Education and Information Technologies, 27, 12345–12368. https://doi.org/10.1007/s10639-021-10745-4

Kemendikbud RI. (2020). Peraturan Menteri Pendidikan dan Kebudayaan Republik Indonesia Nomor 25 Tahun 2020 tentang Standar Satuan Biaya Operasional Pendidikan Tinggi. Kementerian Pendidikan dan Kebudayaan.

Lin, Y., Chen, H., Xia, W., Lin, F., Wang, Z., & Liu, Y. (2023). A comprehensive survey on deep learning. arXiv preprint arXiv:2303.01234.

Nafi'a, M. Z. I., Kuswandi, D., & Wedi, A. (2021, December). Development of Tringo based multiliteracy learning model design as an effort to improve student writing skills. Dalam International Conference on Information Technology and Education (ICITE 2021) (pp. 172–176). Atlantis Press. https://doi.org/10.2991/assehr.k.211210.029

Romero, C., & Ventura, S. (2024). Educational data mining and learning analytics: An updated survey. arXiv preprint arXiv:2402.07956.

Saputro, S. D., Tamam, B., Rohmah, N., Saputro, A. K., Amil, A. J. U., & Salimi, M. (2025). Improving students' cognitive abilities and motivation in kinematics material through Egamerasi media. Salud, Ciencia y Tecnología. https://doi.org/10.56294/saludcyt20251429

Siregar, H. R., & Darma, D. A. (2021). Transparansi penetapan UKT di PTN: Tinjauan sistemik. Jurnal Kebijakan Publik Indonesia, 4(2), 111–120.

Syam, F. A., Wulandari, S. R., & Putri, D. T. T. (2024). Penerapan SMOTE pada klasifikasi UKT menggunakan Random Forest. Jurnal Teknologi Informasi dan Komputer, 9(2), 102–111.

Verma, A., Sharma, R., & Bhatia, P. K. (2025). Comparative analysis of feature selection techniques for student performance prediction. SN Computer Science, 6, 123. https://doi.org/10.1007/s42979-025-04035-9

Yustanti, W. (2021). Pengembangan model pengelompokan UKT berbasis algoritma unsupervised learning. Laporan Penelitian LPPM Universitas Negeri Surabaya. https://repository.unesa.ac.id

Yustanti, W., & Rochmawati, N. (2022). Analisis algoritma klasifikasi untuk memprediksi karakteristik mahasiswa pada pembelajaran daring. Jurnal Edukasi dan Penelitian Informatika (JEPIN), 8(1), 57–65. https://doi.org/10.26418/jp.v8i1.49971