Green Intelligent Systems and Applications

Research Article
Volume 6(2), 2026, 217–230
https://doi.org/10.53623/gisa.v6i2.1289

Analysis of the Effect of Attributes on Single Tuition Fee Grouping Using Multiple Linear Regression and Extreme Gradient Boosting

Dea Annona Prayetno Putri*, Erna Budhiarti Nababan, Mohammad Andri Budiman

Faculty of Computer Science and Information Technology, Universitas Sumatera Utara, Medan, Indonesia

*Correspondence: deanonpp@usu.ac.id

SUBMITTED: 7 July 2026; REVISED: 18 August 2026; ACCEPTED: 20 August 2026

ABSTRACT: Objectively determining the Single Tuition Fee (UKT) level was essential for aligning educational costs with students' socioeconomic conditions. This study compared Multiple Linear Regression (MLR) and Extreme Gradient Boosting (XGBoost) for predicting the final (Pleno) UKT level using data from 12,355 students, including socioeconomic, school-related, and administrative attributes. Data preprocessing involved cleaning, handling missing values and duplicates, transforming categorical variables, and splitting the dataset into training and testing sets (80:20). Models were evaluated under "operational" and "strict" scenarios using accuracy, precision, recall, F1-score, confusion matrices, and AUC. In the operational scenario, XGBoost achieved 93.42% training and 90.15% testing accuracy, outperforming MLR, which achieved 84.10% testing accuracy. Under the strict scenario, XGBoost accuracy decreased to 63.27–63.54%, compared with 49.74–49.86% for MLR. These findings demonstrated XGBoost's greater ability to capture nonlinear relationships and complex attribute interactions. Permutation importance identified Coordinator- and Verifier-assigned UKT as the dominant predictors, while SHAP analysis identified per capita income as the most influential socioeconomic factor. Overall, XGBoost showed superior predictive performance but should be used as a decision-support tool rather than as the sole determinant of student tuition fees.

KEYWORDS: Single Tuition Fee (UKT); Multiple Linear Regression; XGBoost; prediction; SHAP; decision support system.

1. Introduction

The Single Tuition Fee (UKT) represented an important component of Indonesia's higher education financing system, designed to promote fair and proportional access to education based on students' socioeconomic circumstances [1]. The determination of UKT was intended to reflect students' financial capacity so that tuition fees could be allocated according to differences in household economic conditions [2]. The UKT policy also functioned as a cross-subsidy mechanism, whereby students from families with greater financial capacity could be assigned higher tuition levels, while those from economically disadvantaged households could receive lower tuition levels [1, 3, 4].

At the University of North Sumatra, UKT was classified into eight levels, ranging from Level 1 to Level 8. The classification process considered several socioeconomic attributes, including the number of dependents, water source, household electricity capacity, ownership of movable and immovable assets, investments, possession of a poverty certificate, total family income, and total family expenditure [5–7]. The submitted information was subsequently evaluated through a verification and plenary process to assess its validity and determine the appropriate UKT level [8]. However, this process presented several challenges. Manual verification could be time-consuming and potentially susceptible to subjective judgment, while the available socioeconomic attributes might not comprehensively represent household financial capacity. In particular, family income and expenditure alone might provide an incomplete representation because they did not fully account for differences in household characteristics, assets, financial commitments, and other relevant socioeconomic factors [9].

Previous studies investigated several computational approaches for UKT determination. Artificial Neural Network-based classification was applied to support automated UKT classification [10], while K-means clustering was used to group students according to selected socioeconomic characteristics [11]. These studies demonstrated the potential of data-driven approaches to improve the efficiency and consistency of UKT determination. Nevertheless, important research gaps remained. First, previous studies generally relied on conventional household socioeconomic variables, whereas educational-background attributes received limited consideration. Second, most studies emphasized classification or clustering performance without adequately explaining how individual attributes influenced model predictions [12]. Such interpretability was particularly important because UKT classification directly affected students' financial obligations. Third, comparative evidence between conventional statistical approaches and advanced nonlinear machine-learning models remained limited, particularly regarding the trade-off between predictive performance and model interpretability.

To address these gaps, this study incorporated school of origin and previous school fees as complementary educational-background attributes alongside conventional socioeconomic variables. These variables were investigated because household income and expenditure might fluctuate and might not fully capture differences in household resources, family size, and expenditure patterns [13]. Previous school fees could provide supplementary information regarding prior household educational expenditure, while school of origin could reflect differences in educational environments and access to educational resources [14]. However, these variables were treated as complementary predictors rather than direct measures of socioeconomic status.

The novelty of this study lay in integrating expanded student attributes, comparative predictive modeling, and explainable artificial intelligence. Multiple Linear Regression (MLR) and Extreme Gradient Boosting (XGBoost) were compared to evaluate the performance of a conventional linear model against a nonlinear ensemble-learning approach capable of capturing complex relationships and interactions among predictors. Furthermore, permutation importance and SHAP analyses were employed to identify influential attributes and interpret their contributions to model predictions. This approach extended previous UKT studies by evaluating not only predictive performance but also model transparency and the contribution of individual attributes.

Therefore, this study aimed to evaluate the influence of socioeconomic, educational-background, and administrative attributes on final UKT classification and to compare the predictive performance of MLR and XGBoost under operational and strict modeling scenarios. It further aimed to identify the most influential predictors through permutation importance and SHAP analyses, thereby providing an interpretable data-driven framework for supporting UKT decision-making.

2. Materials and Methods

This study employed a quantitative comparative approach to evaluate the influence of student attributes on the determination of the final Single Tuition Fee (UKT) level and to compare the predictive performance of Multiple Linear Regression (MLR) and Extreme Gradient Boosting (XGBoost). Historical student records were used, with the final UKT level determined during the plenary process (*UKT Pleno*) serving as the response variable. Predictor variables comprised socioeconomic characteristics, educational-background information, and administrative UKT assessments.The analytical procedure consisted of five main stages: (1) data acquisition and variable selection, (2) data preprocessing and feature engineering, (3) data partitioning and validation, (4) model development using MLR and XGBoost under operational and strict scenarios, and (5) model evaluation and interpretation. Figure 1 presents the overall methodological framework, illustrating the sequence from raw-data acquisition and preprocessing to model development, performance evaluation, and interpretation.

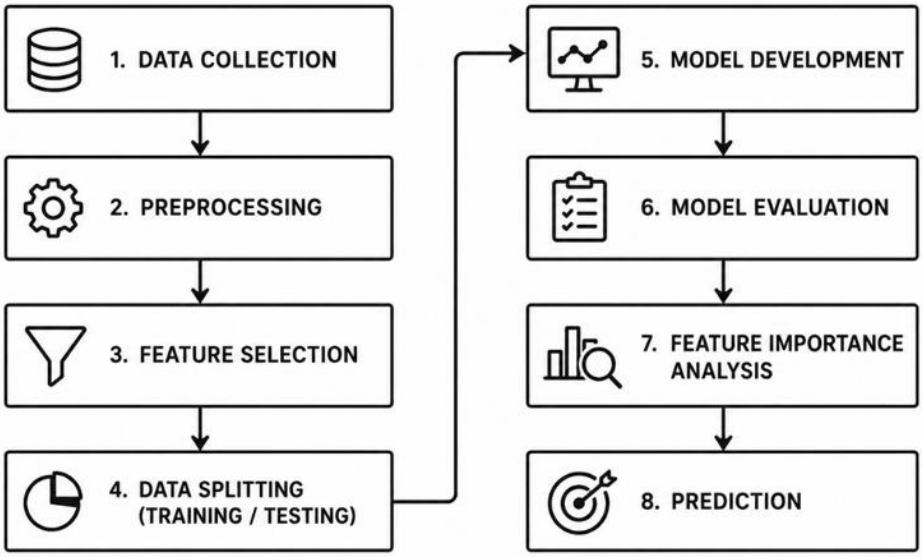

Figure 1. Research Architecture.

2.1. Data Acquisition and Predictor Variables.

The dataset consisted of historical records of 12,355 students and initially contained 109 attributes related to student characteristics, socioeconomic conditions, educational backgrounds, and UKT assessment processes. The raw attributes were screened to identify variables relevant to UKT determination and the objectives of the predictive analysis. Variables were selected based on their socioeconomic relevance, availability, data quality, and potential contribution to model development.The final predictor set included household characteristics, utility information, asset ownership, income and expenditure indicators, educational-background attributes, and administrative UKT assessments. In particular, school of origin and

previous school fees were included as complementary educational-background variables, whereas the Coordinator- and Verifier-assigned UKT levels represented administrative assessments conducted before the final plenary decision. Table 1 summarizes the predictor variables, their data types, and their operational definitions.

Table 1. Predictor Variables Used for UKT Modeling and Their Operational Definitions.

No.	Predictor variable	Type	Operational definition
1	Number of Dependents	Numerical	Number of household members financially supported by the family.
2	Water Source	Categorical	Primary source of water used by the student's household.
3	PLN Electricity Capacity	Numerical/Categorical	Installed household electricity capacity supplied by PLN.
4	Movable Property	Numerical/Categorical	Ownership or value of movable household assets.
5	Immovable Property	Numerical/Categorical	Ownership or value of land, buildings, and other immovable assets.
6	Investment Information	Numerical/Categorical	Ownership or value of investments and other financial assets.
7	Certificate of Poverty	Categorical	Possession of an official certificate indicating economic disadvantage.
8	Total Family Income	Numerical	Total household income during the specified assessment period.
9	Total Family Expenditure	Numerical	Total household expenditure during the specified assessment period.
10	Per Capita Income	Numerical	Household income adjusted for the number of household members or dependents.
11	School of Origin	Categorical	Type or category of school attended before university enrollment.
12	Previous School Fees	Numerical	Educational fees paid at the student's previous school.
13	UKT Coordinator Level	Ordinal	UKT level recommended during the coordinator assessment stage.
14	UKT Verifier Level	Ordinal	UKT level assigned during verification before the final plenary decision.

2.2. Data Preprocessing and Feature Engineering.

Data preprocessing was performed before model development to improve data completeness, consistency, and analytical reliability. The procedure included duplicate removal, missing-value treatment, categorical encoding, outlier assessment, feature scaling where required, and feature selection. Duplicate observations were removed to prevent repeated records from disproportionately influencing model estimation. Missing values in numerical variables were imputed using the median of the corresponding training variable because the median was less sensitive to skewed distributions and extreme socioeconomic values. Missing categorical observations were imputed using the most frequent category. Importantly, all imputation parameters were estimated from the training data and subsequently applied to the testing data to prevent information leakage. Nominal categorical variables, including water source and school of origin, were transformed using one-hot encoding. Ordinal variables, particularly ordered UKT categories, were encoded according to their natural ranking. Numerical variables were examined for extreme observations using the interquartile range (IQR):

$$IQR = Q_3 - Q_1 \tag{1}$$

Potential outliers were identified using the lower and upper boundaries:

$$L = Q_1 - 1.5(IQR) \tag{2}$$

$$U = Q_3 + 1.5(IQR) \tag{3}$$

where and represent the first and third quartiles, respectively. Because extreme socioeconomic values could represent genuine household conditions, observations outside these boundaries

were not automatically removed. They were examined for plausibility, and only erroneous or implausible values were corrected, excluded, or capped when justified.

For MLR, continuous predictors were standardized where required to facilitate comparison among variables measured on different scales. Tree-based XGBoost modeling did not require feature scaling because its decision-tree structure was invariant to monotonic changes in predictor scale. The same preprocessing transformations fitted to the training data were applied unchanged to the testing data.

2.3. Modeling Scenarios.

The models were developed under two predictor scenarios to evaluate both predictive performance and the influence of administrative information. The operational scenario included socioeconomic, educational-background, and administrative UKT attributes, including the Coordinator- and Verifier-assigned UKT levels. This scenario represented prediction within the existing administrative workflow. The strict scenario excluded attributes generated during the UKT assessment process, particularly variables that could contain information closely related to the final plenary decision. The strict scenario therefore evaluated the extent to which the final UKT level could be predicted using student socioeconomic and educational-background characteristics alone. Comparing these scenarios also enabled assessment of the contribution of administrative variables and reduced the risk of overstating model performance due to target-proximal predictors. This addition is particularly important because your results showed a major performance decline from approximately 90% to 63% for XGBoost when moving from operational to strict mode. The Methods must therefore define exactly which variables were included/excluded.

2.4. Data Partitioning and Model Validation.

The complete dataset was partitioned into training and independent testing subsets using an 80:20 ratio. The partitioning procedure maintained a comparable distribution of UKT categories between the two subsets to minimize distributional differences, particularly for less frequently represented UKT levels. The training subset was used for model development, feature preprocessing, hyperparameter optimization, and cross-validation, whereas the testing subset was reserved exclusively for final model evaluation. Model selection and hyperparameter optimization were conducted using stratified -fold cross-validation within the training dataset. In each iteration, folds were used for model fitting and the remaining fold for validation. The procedure was repeated until each fold had served as the validation subset. The independent testing dataset was not used during model development or hyperparameter selection, thereby reducing information leakage and providing a more objective estimate of model generalizability.

2.5. Multiple Linear Regression.

Multiple Linear Regression was employed as the conventional statistical benchmark for examining relationships between the predictor variables and final UKT level. The general model was expressed as:

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} + \varepsilon_i \quad (4)$$

where represents the predicted final UKT level for student; is the intercept; represents the regression coefficient associated with predictor; represents the value of predictor for student; is the total number of predictors; and represents the residual error.

Before interpreting the regression results, the assumptions of linearity, independence of residuals, homoscedasticity, approximate normality of residuals, and absence of severe multicollinearity were evaluated. Multicollinearity was assessed using the variance inflation factor (VIF):

$$VIF_j = \frac{1}{1 - R_j^2} \quad (5)$$

where represents the coefficient of determination obtained by regressing predictor on the remaining predictors.

For XGBoost, the main hyperparameters included n_estimators = [value], max_depth = [value], learning_rate = [value], min_child_weight = [value], subsample = [value], colsample_bytree = [value], reg_alpha = [value], and reg_lambda = [value], which were selected based on validation performance while controlling model complexity and reducing overfitting. Model validation was conducted using an 80:20 training-testing split, while k-fold cross-validation (k = [5/10]) was applied to the training data during model selection and hyperparameter evaluation. Where necessary, stratified cross-validation was used to maintain a comparable distribution of UKT categories across folds. The analysis was implemented in Python [version], using pandas and NumPy for data processing, scikit-learn [version] for preprocessing, regression, cross-validation, and performance evaluation, xgboost [version] for XGBoost modeling, and SHAP [version] for model interpretation. The independent testing set was used only for final evaluation and was not involved in model fitting or hyperparameter selection to minimize information leakage and provide a more objective estimate of model generalization.

3. Results and Discussion

This study develops a UKT (Single Tuition Fee) prediction system that classifies new student profiles into one of nine Plenary ("Pleno") UKT tiers using Multiple Linear Regression (MLR). MLR was selected for its capacity to process multiple independent variables simultaneously, enabling the identification of patterns linking student data to UKT classification outcomes. The Plenary UKT tier serves as the target variable, while student attributes function as predictor variables that support the classification process. The features used in the model are grouped into two categories: socioeconomic features and operational features. Socioeconomic features describe a student's financial status, family background, and prior educational history before the administrative verification process begins. These include total family income, paternal and maternal salaries, number of dependents, household electricity capacity (PLN), previous school tuition fees, parental education and occupation, admission pathway, KIP-Kuliah scholarship status, faculty, and study program. Because these attributes are unaffected by verifier or

coordinator decisions, they serve as more objective initial indicators of a student's family economic capacity.

Operational features, by contrast, are derived from the administrative stages of UKT determination, namely the Verifier UKT Tier, Coordinator UKT Tier, New System UKT Tier, and Old System UKT Tier. As preliminary assessments made prior to the final plenary decision, these features are closely correlated with the Plenary UKT outcome and can substantially improve model accuracy. However, their use warrants careful scrutiny, as heavy reliance on variables already closely aligned with the final decision risks target leakage. This categorization allows the model to be evaluated from two complementary perspectives: operational prediction, which reflects performance when all available administrative information is used, and initial objectivity, which reflects the intrinsic influence of socioeconomic and school-related attributes when prior administrative decisions are excluded. This distinction ensures that the results demonstrate not only predictive accuracy but also academically and ethically sound interpretability within the context of UKT decision-making.

3.1. Dataset Description.

The dataset comprises 12,355 records with 17 attributes, covering student identity, academic information, socioeconomic conditions, and UKT tier assignments across multiple assessment stages. It forms the basis for analyzing how socioeconomic attributes and school-related data influence the Plenary UKT tier, using MLR combined with an Explainable AI (XAI) approach. The Previous School Tuition Fee (SPP) attribute is included as an additional indicator of a family's financial capacity to fund education prior to the student's enrollment in higher education. The dataset also contains several attributes representing UKT assessment outcomes at different stages: Old System UKT Tier, New System UKT Tier, Verifier UKT Tier, Coordinator UKT Tier, and Plenary UKT Tier.

3.2. Multiple Linear Regression Modeling Results.

MLR modeling was conducted to predict the Plenary UKT tier based on socioeconomic, school, and operational attributes available in the dataset. The model incorporates regularization to improve prediction stability and reduce sensitivity to data complexity. Data were split into 80% training and 20% testing (holdout) subsets, with the training set used to build the model and the test set used to evaluate its predictive performance on unseen data. The model was evaluated under two scenarios: operational mode, which incorporates administrative features such as the Verifier UKT Tier, Coordinator UKT Tier, and System UKT Tier, and strict mode, which relies primarily on socioeconomic and school-related attributes that are less influenced by administrative decisions. Figure 2 presents the accuracy results of the MLR model under both scenarios. The operational mode consistently outperformed the strict mode. With the SPP attribute included, the operational mode achieved a training accuracy of 84.81% and a test accuracy of 84.10%. Excluding the SPP attribute produced identical results (84.81% training, 84.10% test), indicating that the inclusion of SPP did not materially affect operational-mode performance, a result consistent with administrative features already dominating the model's predictive capability. In strict mode, the model with the SPP attribute achieved a training accuracy of 52.12% and a test accuracy of 49.74%, while the model without SPP achieved 52.24% training accuracy and 49.86% test accuracy. These substantially lower

results reflect the absence of operational features closely tied to the final Plenary UKT decision, leaving the model to rely solely on socioeconomic and school-related attributes.

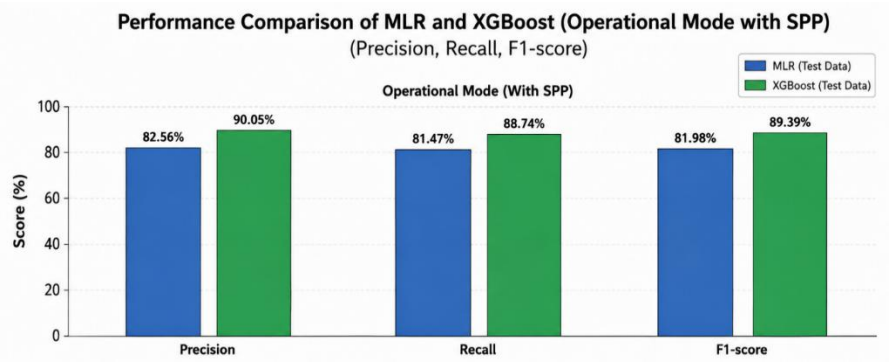

Figure 2. MLR model accuracy.

Figure 3 presents the marginal association between each feature and the target class, measured using mutual information (MI) on the training data, and displays the fifteen features with the strongest association. The four operational features dominate the ranking, led by K.UKT Verifikator and Level UKT (MI ≈ 1.22 and 1.20), followed by K.UKT Sistem Lama and K.UKT Koordinator (MI ≈ 0.87 and 0.85). This pattern aligns with the accuracy results in Figure 2 and reinforces the target leakage concern, as these features carry nearly twice the informational association with the Plenary UKT tier compared to the strongest socioeconomic predictors. Among socioeconomic features, income-related variables form the next cluster, with Total Penghasilan and its transformed variants (pln_kva_per_log1p_income, sqrt_penghasilan, log_penghasilan, cbrt_penghasilan) showing similar MI values of approximately 0.50–0.54, suggesting that raw income magnitude, rather than its transformation, drives relevance. Income-per-dependent features show slightly lower MI (≈0.36–0.37), while more granular salary components (log_gaji_reg, gaji_orangtua, max_gaji_ortu, gaji_ayah_share) rank lowest (≈0.25–0.31). Overall, Figure 3 confirms that operational features hold disproportionately strong marginal association with the Plenary UKT tier relative to socioeconomic attributes, highlighting both their predictive value and their leakage risk..

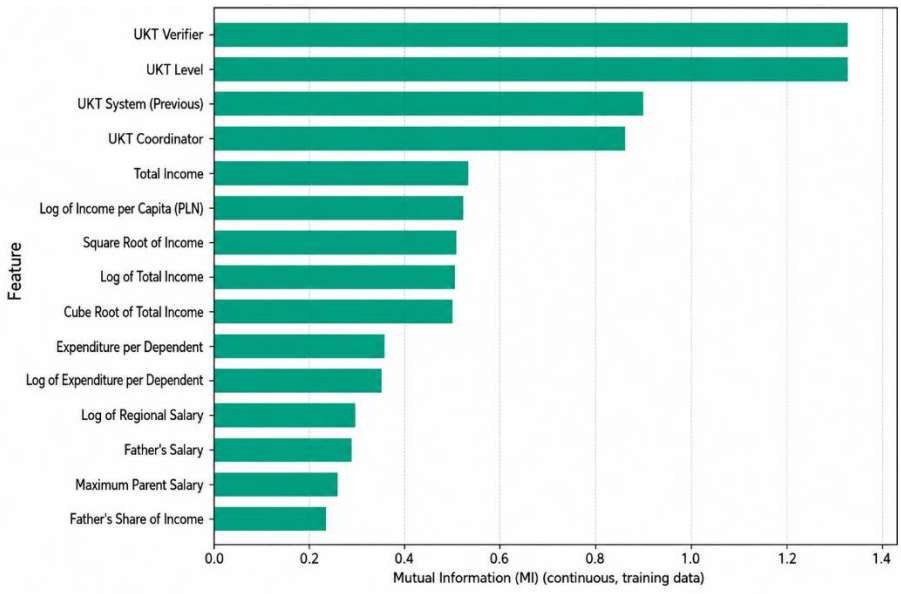

Figure 3. Feature-Marginal Association.

3.3. Confusion Matrix.

The confusion matrix for the UKT (Single Tuition Fee) determination model using Multiple Linear Regression shows an accuracy rate of 84.10% on the holdout test data. These results indicate that the model classifies UKT groups reasonably well, particularly for groups with large sample sizes, such as K3, K5, K6, K7, K8, and K9. Correct predictions are observed along the main diagonal: 25 instances for K2, 190 for K3, 119 for K4, 180 for K5, 193 for K6, 185 for K7, 525 for K8, and 661 for K9. Groups K8 and K9 demonstrated the strongest predictive performance, yielding the highest number of correct predictions. Prediction errors generally occurred between adjacent UKT groups; for instance, some actual K8 instances were predicted as K6 or K7, while some actual K9 instances were predicted as K8. This suggests that the model still struggles to distinguish between UKT groups with relatively similar input characteristics, particularly among numerically adjacent classes. Meanwhile, group K1 could not be predicted effectively due to its very limited sample size, resulting in no correctly classified K1 instances. The confusion matrix results are presented in Figure 4.

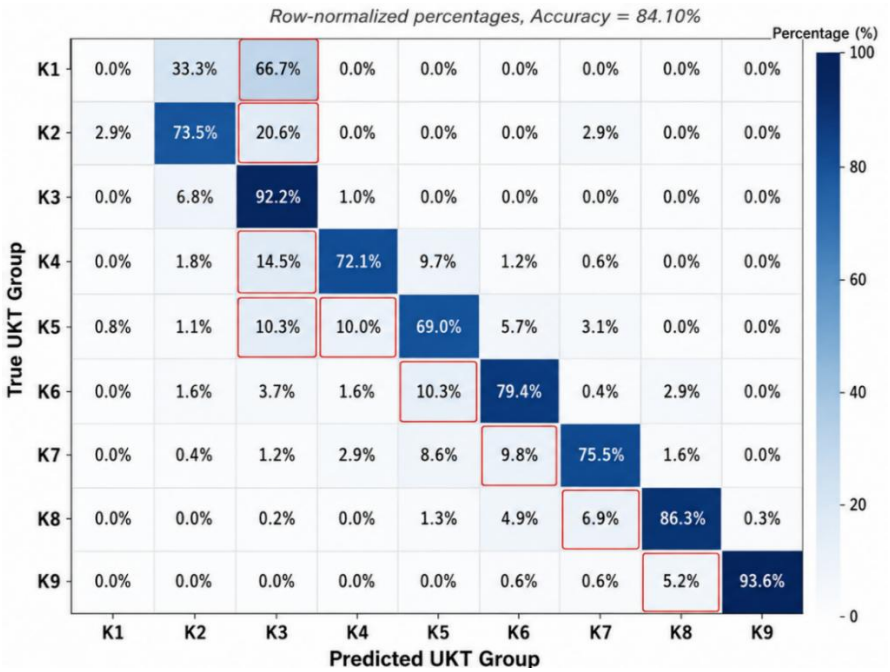

Figure 4. MLR Confusion Matrix.

3.4. XGBoost Modeling Results.

In this study, XGBoost modeling was employed to predict the Plenary UKT group based on socioeconomic attributes, school data, and operational attributes available in the dataset. The XGBoost model was selected for its ability to capture complex, non-linear relationships between variables. Unlike linear models, XGBoost employs a decision tree-based ensemble approach, constructing multiple trees sequentially to correct the prediction errors of preceding trees. Through this mechanism, XGBoost is expected to yield more stable and accurate UKT group predictions, particularly for data characterized by significant attribute variation. As with the MLR model, the data were split into 80% training and 20% testing (holdout) subsets, with the training data used to build the model and the test data used to evaluate its predictive performance on unseen data. The model was evaluated under the same two scenarios described

previously: operational mode, which utilizes administrative attributes associated with the UKT determination process, and strict mode, which relies primarily on socioeconomic attributes and school data less influenced by administrative decisions. Figure 5 presents the accuracy results of the XGBoost model.

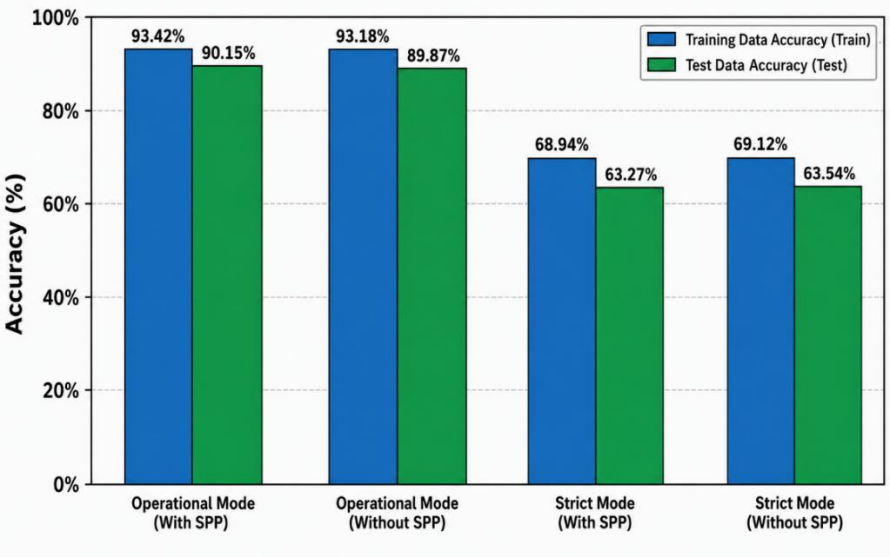

Figure 5. XGBoost model accuracy.

As shown in Figure 5, the XGBoost model demonstrates strong performance in the operational scenario. With the SPP attribute included, the model achieved a training accuracy of 93.42% and a test accuracy of 90.15%. Without the SPP attribute, the model achieved a training accuracy of 93.18% and a test accuracy of 89.87%. These results indicate that the XGBoost model effectively learns underlying data patterns while maintaining high performance on unseen data. The relatively small gap between training and test accuracies further suggests that the model does not suffer from excessive overfitting in the operational scenario. In the strict mode scenario, accuracy declined relative to the operational mode: with the SPP attribute, the model achieved a training accuracy of 68.94% and a test accuracy of 63.27%, while without SPP, it achieved a training accuracy of 69.12% and a test accuracy of 63.54%. This decline indicates that, when administrative attributes are excluded or restricted, the model's ability to predict the Plenary UKT group diminishes, likely because socioeconomic and school-related attributes alone are insufficient to fully capture the patterns underlying the final UKT assignment decisions.

3.5. Confusion Matrix.

The confusion matrix for the XGBoost model achieved an accuracy rate of 90.15% on the holdout test data under the operational scenario with the SPP attribute included. These results indicate that the XGBoost model classifies UKT groups highly effectively, particularly for groups with large sample sizes, such as K3, K5, K6, K7, K8, and K9. Correct predictions are observed along the main diagonal: K1 (2 instances), K2 (30), K3 (196), K4 (135), K5 (220), K6 (216), K7 (211), K8 (548), and K9 (670). As with the MLR model, groups K8 and K9 demonstrated the strongest predictive performance, recording the highest number of correct predictions among all groups. Prediction errors again occurred predominantly between adjacent UKT groups: of the actual K8 instances, 24 were predicted as K7, 15 as K6, and 14 as K9,

while of the actual K9 instances, 28 were predicted as K8, 5 as K7, and 3 as K6. This error pattern indicates that the model still struggles to distinguish between UKT groups with relatively similar input characteristics, particularly among numerically adjacent classes. The confusion matrix results, presented in Figure 6, demonstrate that XGBoost possesses robust classification capability, as the majority of predictions are concentrated along the main diagonal, indicating that correct predictions substantially outnumber errors. Notably, even group K1, despite its very limited sample size, achieved two correct predictions, although its performance remained weaker than that of classes with larger sample sizes. Overall, the XGBoost model under the operational scenario with SPP demonstrates strong predictive performance for the Plenary UKT group, characterized by high accuracy and error patterns that remain largely confined to adjacent UKT groups.

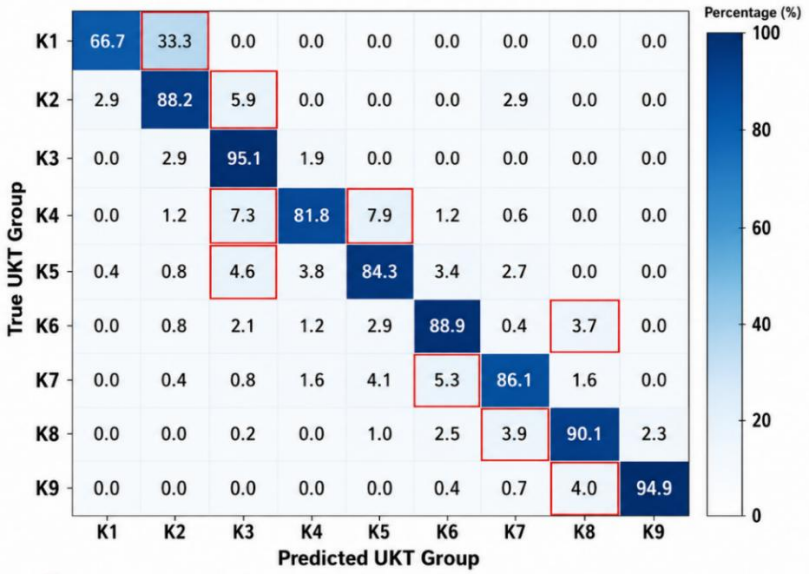

Figure 6. Result Confusion Matriks XGBoost.

3.6. Results of XAI Implementation.

The results of applying Explainable Artificial Intelligence (XAI) using the SHAP (Shapley Additive Explanations) method demonstrate that the model can explain the factors influencing its prediction of UKT Group 5. As shown in the SHAP summary plot in Figure 7, the sqrt_Pendapatan Per Kapita (square root of per capita income) variable emerges as the most dominant factor, with SHAP values ranging from approximately −7 to +0.4. A negative SHAP value indicates that higher per capita income steers the model's prediction away from UKT Group 5 and toward a higher UKT group instead. The horizontal axis represents SHAP values, indicating the magnitude of each variable's contribution to the model's output: values to the right of zero drive the prediction toward UKT Group 5, while values to the left steer it away from that group. The color of each point represents the corresponding feature value, with red indicating high values and blue indicating low values. As shown in the figure, sqrt_Per_Capita_Income is the most dominant variable, exhibiting the widest spread of SHAP values, from approximately −7 to +0.4, indicating that per capita income exerts the greatest influence on the model's predictions. High per capita income values tend to cluster on the negative side, indicating that students with higher per capita income are less likely to be assigned to UKT Group 5 and more likely to be placed in a higher UKT group. Conversely,

lower per capita income values tend to cluster closer to the positive side, supporting a higher likelihood of placement in a mid-range UKT group.

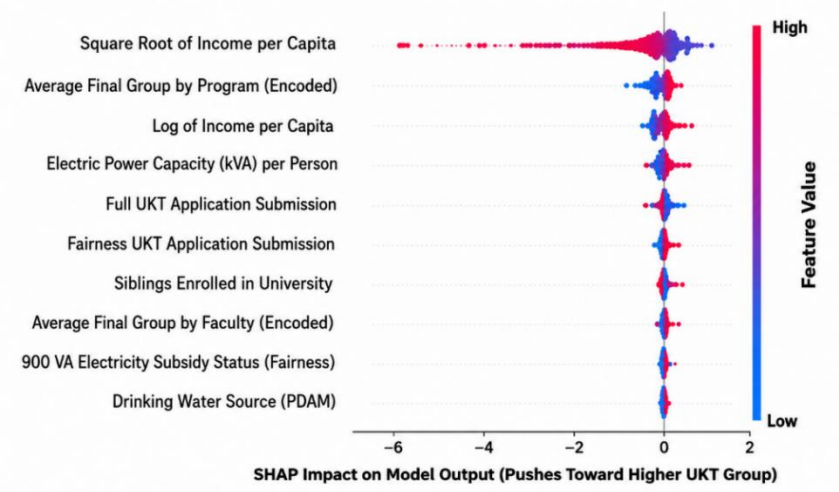

Figure 7. Figure X. SHAP summary plot of key variables influencing UKT group classification.

3.7. Discussions.

The evaluation results showed that the operational mode consistently outperformed the strict mode. For Multiple Linear Regression (MLR), the test accuracy reached 84.10% in the operational mode but decreased to approximately 49.74–49.86% in the strict mode. Similarly, XGBoost achieved its highest test accuracy of 90.15% in the operational mode with the SPP attribute, whereas the strict mode achieved only 63.27–63.54%. This finding indicated that XGBoost captured complex and nonlinear relationships among the predictor variables more effectively than MLR. However, the inclusion of the SPP attribute produced only a marginal improvement, suggesting that SPP was not a dominant predictor of UKT classification. The substantial performance difference between the operational and strict modes also demonstrated the strong contribution of administrative variables, particularly the Coordinator UKT Group and Verifier UKT Group. Permutation importance analysis supported this finding, as shuffling these variables reduced model performance by approximately 29% and 24%, respectively. Therefore, the high predictive accuracy of the operational model should be interpreted cautiously because it partly reflected patterns embedded in previous administrative decisions rather than students' socioeconomic characteristics alone. This distinction is important because a prediction system intended to support equitable UKT determination should not simply reproduce historical administrative decisions. The confusion matrix further showed that both models performed better for UKT groups with larger numbers of observations, particularly K8 and K9, whereas misclassification occurred more frequently between adjacent groups. This pattern suggested that neighboring UKT categories shared similar socioeconomic characteristics, making their boundaries more difficult for the models to distinguish. In addition, the uneven distribution of observations across UKT groups may have contributed to stronger predictive performance for the more frequently represented categories.

SHAP analysis provided further insight into the socioeconomic factors underlying the predictions. Income per capita emerged as an important explanatory variable, while household electricity capacity per person, water source, and the number of siblings enrolled in higher education also contributed to the model output. These variables represented different

dimensions of household economic capacity and financial burden, indicating that the model retained meaningful socioeconomic information even when administrative variables were excluded. Nevertheless, their lower predictive contribution relative to administrative variables explained the reduction in accuracy observed under the strict mode. Overall, XGBoost provided the strongest predictive performance and was more suitable than MLR for capturing the complex relationships associated with UKT classification. However, its dependence on administrative variables in the operational mode indicated that high predictive accuracy did not necessarily imply greater socioeconomic fairness or generalizability. Accordingly, the model should be used as a decision-support tool to complement socioeconomic assessment and administrative verification rather than as an autonomous mechanism for determining students' final UKT groups.

4. Conclusions

The results demonstrated that XGBoost outperformed Multiple Linear Regression (MLR) in predicting the Plenary UKT (Single Tuition Fee) group. In the operational mode, XGBoost achieved the highest test accuracy of 90.15%, compared with 84.10% for MLR. However, the performance of both models declined substantially in the strict mode, in which only socioeconomic and school-related attributes were considered. This decline highlighted the strong influence of administrative variables, particularly the Verifier UKT Group and Coordinator UKT Group, on the final UKT classification. In contrast, the inclusion of the tuition fee (SPP) attribute provided only marginal improvement, while socioeconomic factors, particularly per capita income, remained relevant to the prediction outcomes. Overall, XGBoost was identified as the most effective model for UKT prediction because of its superior predictive performance and ability to capture complex relationships among variables. Nevertheless, the model should be implemented as a decision-support tool that complements socioeconomic assessment and administrative verification rather than as the sole mechanism for determining final UKT groups.

Author Contributions

All authors contributed equally to the conceptualization, methodology, data analysis, interpretation of the results, and preparation of the manuscript. All authors reviewed and approved the final version of the manuscript.

Data Availability

The data supporting the findings of this study are available from the corresponding author upon reasonable request.

Competing Interests

The authors declare that they have no known financial or personal competing interests that could have influenced the work reported in this study.

References

[1] Raiya, M.S.; Khamil, M.R.P.; Fadillah, N.; Saputra, R.A. (2025). Implementasi Algoritma Long Short-Term Memory pada Isu Kenaikan Uang Kuliah Tunggal terhadap Minat Kuliah Mahasiswa. *Jurnal Informatika Terpadu*, *11*(1), 37–43. https://doi.org/10.54914/jit.v11i1.1656.

[2] Susetyoko, R.; Yuwono, W.; Purwantini, E.; Ramadijanti, N. (2022). Perbandingan Metode Random Forest, Regresi Logistik, Naïve Bayes, dan Multilayer Perceptron pada Klasifikasi Uang Kuliah Tunggal (UKT). *Jurnal Infomedia: Teknik Informatika, Multimedia & Jaringan*, *7*(1), 8–16. https://doi.org/10.30811/jim.v7i1.2916.

[3] Wardani, A.N.; Olfah, S.T.; Santoso, K.; Wahyudi, A. (2025). Uang Kuliah Tunggal (UKT) sebagai Penerimaan Negara Bukan Pajak (PNBP): Persepsi Mahasiswa pada PTN dan PTKIN BLU. *Jurnal Info Artha*, *9*(1), 49–77. https://doi.org/10.31092/jia.v9i1.3343.

[4] Sukamto, S.; Adriyani, Y.; Aulia, R. (2020). Prediksi Kelompok UKT Mahasiswa Menggunakan Algoritma K-Nearest Neighbor. *JUITA: Jurnal Informatika*, *8*(1), 121–130. https://doi.org/10.30595/juita.v8i1.6267.

[5] Triase, T.; Samsudin, S. (2020). Implementasi Data Mining dalam Mengklasifikasikan UKT (Uang Kuliah Tunggal) pada UIN Sumatera Utara Medan. *Jurnal Teknologi Informasi*, *4*(2), 370–376. https://doi.org/10.36294/jurti.v4i2.1711.

[6] Cheng, X.; Guo, Y.; Gao, J.; Chen, Y.; Pan, G. (2023). Coordinated Passive Maneuvering Target Tracking by Multiple Underwater Vehicles Based on Asynchronous Sequential Filtering. International Conference on Intelligent Robotics and Applications; Springer, pp. 246–255.

[7] Rohman, M.G.; Abdullah, Z.; Kasim, S.; Albab, M.U. (2026). Machine Learning to Identify Eligibility of Students Receiving Single Tuition Relief. *JITK (Jurnal Ilmu Pengetahuan dan Teknologi Komputer)*, *11*(3), 613–626. https://doi.org/10.33480/jitk.v11i3.7294.

[8] Siddique, A.; Jan, A.; Majeed, F.; Qahmash, A.I.; Quadri, N.N.; Wahab, M.O.A. (2021). Predicting Academic Performance Using an Efficient Model Based on Fusion of Classifiers. *Applied Sciences*, *11*, 11845. https://doi.org/10.3390/app112411845.

[9] Li, W.; Yin, Y.; Quan, X.; Zhang, H. (2019). Gene Expression Value Prediction Based on XGBoost Algorithm. *Frontiers in Genetics*, *10,* 1077. https://doi.org/10.3389/fgene.2019.01077.

[10] Wang, S.; He, J. (2025). Evaluating and Forecasting Undergraduate Dropouts Using Machine Learning for Domestic and International Students. *Technologies*, *13*, 480. https://doi.org/10.3390/technologies13110480.

[11] Wahyuani, A.S. (2018). Pemilihan Parameter Terbaik Menggunakan Metode ID3 untuk Penentuan Cluster Terbaik pada Metode K-Means Clustering: Studi Kasus Penentuan Golongan Uang Kuliah Tunggal Universitas Islam Negeri Maulana Malik Ibrahim. Thesis, Universitas Islam Negeri Maulana Malik Ibrahim, Malang, Indonesia.

[12] Sari, S.D. (2025). Analisis Faktor Penentuan Golongan UKT di UIN Maulana Malik Ibrahim Malang Menggunakan Regresi Probit Ordinal. Thesis, Universitas Islam Negeri Maulana Malik Ibrahim, Malang, Indonesia.

[13] Sheng, C.; Yu, H. (2022). An Optimized Prediction Algorithm Based on XGBoost. In *2022 International Conference on Networking and Network Applications (NaNA)*; IEEE, pp. 1–6. https://doi.org/10.1109/NaNA56854.2022.00082.

[14] Gajdzik, B.; Wolniak, R.; Sączewska-Piotrowska, A.; Grebski, W.W. (2025). Polish Steel Production under Conditions of Decarbonization—Steel Volume Forecasts Using Time Series and Multiple Linear Regression. *Energies*, 18(7), 1627. https://doi.org/10.3390/en18071627.