Explainable Artificial Intelligence for Accurate and Interpretable Student Selection Prediction Without Aggregate Scores Using SKD Examination Data

Mohammad Alomari[1,*]

[1]Department of Information Systems, Faculty of AI and Data Science, Universitas Pelita Harapan, Indonesia

ABSTRACT

Student selection processes are often based on aggregate scores, which can obscure the contribution of individual assessment components and limit interpretability. This study proposes an explainable machine learning approach to predict student selection outcomes using SKD examination data. An XGBoost model was developed and evaluated under two configurations, including a baseline model with aggregate score and a proposed model without it. The results show that the proposed model achieves high predictive performance, with an accuracy of 95.84% and ROC-AUC of 0.9935, demonstrating that individual test components provide sufficient information for reliable prediction. To enhance transparency, SHAP was applied to provide both global and local explanations of model decisions. The analysis reveals that cognitive test components, particularly TIU and TWK, are the most influential factors, while non-academic variables such as formation and location also contribute to the prediction. These findings highlight the importance of interpretability in high-stakes decision-making and suggest that accurate and transparent student selection systems can be developed without relying on aggregate scores. The proposed approach provides a robust framework for supporting fair, interpretable, and data-driven selection policies.

Keywords Explainable Artificial Intelligence, Student Selection Prediction, XGBoost, SHAP, Educational Data Mining

Introduction

Student selection is a critical process in educational institutions, as it determines the quality of admitted candidates and directly influences academic outcomes. In many admission systems, decisions are primarily based on aggregate scores derived from multiple assessment components. While this approach simplifies evaluation, it often reduces transparency by masking the contribution of individual test elements. As a result, stakeholders may find it difficult to understand which specific factors influence the selection outcomes. In high-stakes contexts such as higher education admission, the lack of transparency can raise concerns related to fairness, accountability, and trust in the decision-making process [1].

In recent years, machine learning techniques have been increasingly adopted to support decision-making in education. Models such as decision trees, random forests, and gradient boosting methods have demonstrated strong performance in predicting academic success and admission outcomes [2]. Among these approaches, XGBoost has gained popularity due to its ability to handle structured data efficiently and produce high predictive accuracy [3]. However, most existing studies focus primarily on optimizing performance metrics, often overlooking the importance of model interpretability. As a result, many predictive

Submitted 5 May 2026
Accepted 20 June 2026
Published 2 November 2026

*Corresponding author
Mohammad Alomari,
mh.alomari@aau.edu.jo

Additional Information and Declarations can be found on page 234

DOI: 10.63913/ail.v2i4.68
© Copyright
2026 Alomari

Distributed under
Creative Commons CC-BY 4.0

How to cite this article: M. Alomari, "Explainable Artificial Intelligence for Accurate and Interpretable Student Selection Prediction Without Aggregate Scores Using SKD Examination Data," *Artif. Intell. Learn.*, vol. 2, no. 4, pp. 221-236, 2026.

models operate as black boxes, providing limited insight into how decisions are made.

To address this limitation, Explainable Artificial Intelligence has emerged as an important area of research that aims to make machine learning models more transparent and interpretable. Techniques such as SHAP provide a systematic way to quantify the contribution of each feature to model predictions, both at the global and individual levels [4]. These methods have been successfully applied in domains such as healthcare and finance, where understanding model decisions is essential. Despite these advances, the application of explainable AI in student selection remains relatively limited, particularly in studies that aim to balance predictive performance with interpretability.

Another important limitation in the current literature is the strong reliance on aggregate scores as input features. Aggregate scores are typically constructed from individual test components, which can introduce redundancy and potentially lead to information leakage. This reliance may result in models that appear highly accurate but fail to provide meaningful insights into the underlying decision process. Furthermore, the influence of non-academic factors, such as geographic location or program allocation, is often not thoroughly examined. These factors may play a role in real-world selection systems, but their impact is rarely analyzed in a transparent and systematic way [5].

Based on these limitations, this study addresses the research gap by proposing an explainable machine learning approach for student selection prediction that does not depend on aggregate scores. The proposed method uses XGBoost for prediction and SHAP for interpretation, enabling both high performance and transparency. This study aims to evaluate whether accurate predictions can be achieved using only individual test components, to identify the most influential factors affecting selection outcomes, and to provide interpretable insights that support more transparent decision-making. By integrating predictive modeling with explainable AI, this research contributes to the development of more reliable, interpretable, and policy-relevant student selection systems.

Literature Review and Related Works

The application of machine learning in education has gained significant attention in recent years, particularly in predicting student performance and supporting admission decisions. Various studies have explored classification models such as decision trees, support vector machines, and ensemble methods to improve predictive accuracy in educational datasets [6], [7]. These approaches have demonstrated the potential of data-driven techniques to enhance traditional selection processes, which are often limited by manual evaluation and aggregate scoring systems.

Among the existing machine learning methods, ensemble techniques such as random forests and gradient boosting have shown superior performance due to their ability to capture complex relationships in structured data [8], [9]. XGBoost, in particular, has emerged as a state-of-the-art algorithm because of its scalability, regularization mechanisms, and strong predictive performance across various domains [10]. In the context of education, several studies have applied XGBoost to predict academic success, dropout risk, and admission outcomes, reporting improved accuracy compared to conventional models [11], [12].

Despite these advancements, most studies primarily focus on maximizing prediction performance while treating the model as a black box. This lack of interpretability poses a significant limitation, especially in high-stakes decision-making scenarios such as student selection [13]. Decision-makers often require not only accurate predictions but also clear explanations of how and why decisions are made. Without interpretability, the adoption of machine learning models in real-world educational systems remains limited due to concerns related to transparency and trust [14].

To address this issue, Explainable Artificial Intelligence has been introduced to improve the interpretability of machine learning models. Techniques such as SHAP provide a unified framework for interpreting predictions by quantifying the contribution of each feature [15]. SHAP has been widely applied in domains such as healthcare, finance, and risk analysis to enhance model transparency and support decision-making [16], [17]. In the educational domain, however, the use of SHAP and other explainability methods is still relatively limited, particularly in the context of admission prediction [18].

Another important limitation in the literature is the heavy reliance on aggregate features, such as total scores, which are often derived from multiple components [19]. While these features may improve model performance, they can introduce redundancy and reduce interpretability by masking the contribution of individual variables. Furthermore, existing studies often overlook the role of non-academic factors, such as geographic location or program allocation, which may influence selection outcomes in real-world settings [20]. Therefore, there is a need for approaches that not only achieve high predictive performance but also provide interpretable insights without depending on aggregate features.

Methodology

This study proposes an explainable machine learning framework for predicting student selection outcomes using SKD examination data. The overall research workflow is illustrated in figure 1, which outlines the main stages including data collection, preprocessing, feature construction, model development, evaluation, and explainability analysis. The process begins with dataset preparation and cleaning, followed by the construction of two experimental scenarios to assess the impact of aggregate features. The XGBoost model is then trained and evaluated using multiple performance metrics. Finally, SHAP is applied to provide both global and local explanations of the model predictions, ensuring that the system is both accurate and interpretable.

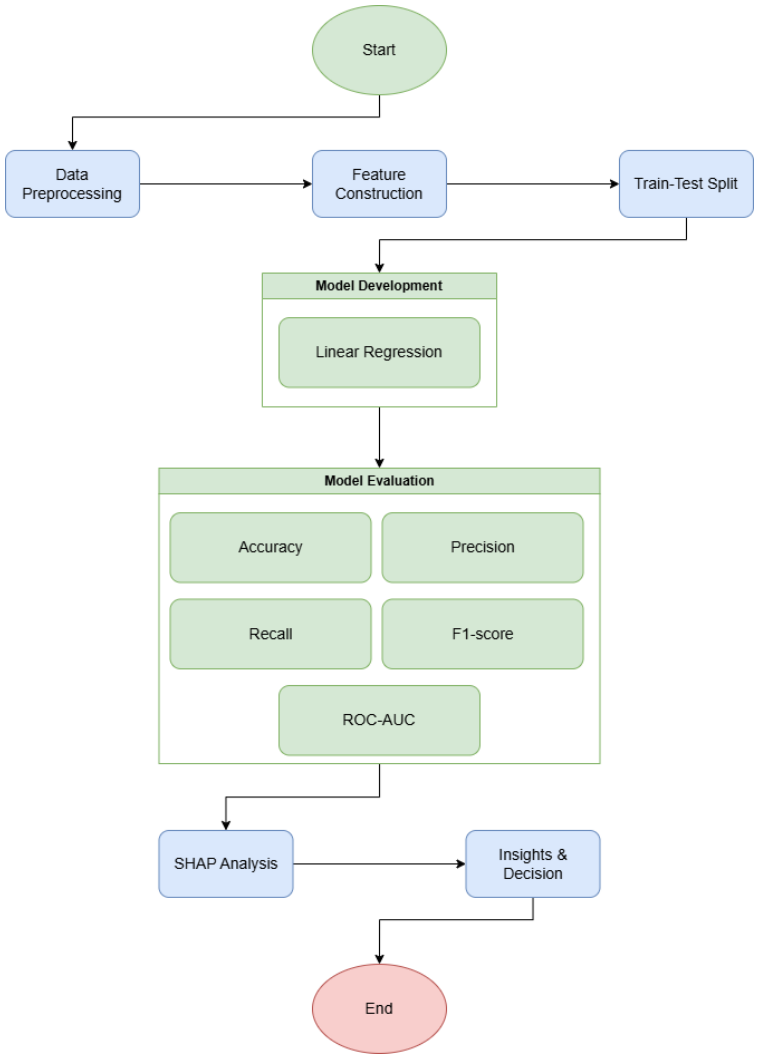


Figure 1 Research framework

Dataset Description

The dataset used in this study consists of SKD examination results from the 2024 student admission process, containing 16,106 records and multiple attributes. Each record represents an individual candidate and includes demographic information, educational background, and test scores. The main academic features include TWK, TIU, and TKP scores, while contextual features include study program, formation code, and location.

To ensure data consistency, candidates who did not attend the examination were removed from the dataset. The target variable was defined as a binary classification problem, where candidates labeled as passing were assigned a value of 1 and those not passing were assigned a value of 0. The dataset is imbalanced, with a higher proportion of non-passing candidates compared to passing candidates.

Data Preprocessing

Data preprocessing was conducted to handle missing values and transform the dataset into a format suitable for machine learning. Numerical features were

imputed using the median value to reduce the impact of outliers, while categorical features were imputed using the most frequent value.

Categorical variables were transformed into numerical representations using one-hot encoding, which converts each category into binary features. This step is necessary because the XGBoost model requires numerical input.

The dataset was then split into training and testing sets using an 80:20 ratio. Stratified sampling was applied to maintain the class distribution in both sets, which is important due to class imbalance.

Feature Construction and Experimental Design

To evaluate the role of aggregate features in the prediction task, two feature configurations were constructed. The first configuration, referred to as the baseline model, includes all available features, including the aggregate score (skd.total). The second configuration, referred to as the proposed model, excludes the aggregate score and relies only on individual test components and other attributes. The exclusion of skd.total is motivated by the fact that it is derived directly from individual test scores, which may introduce redundancy and potentially lead to information leakage in the modeling process. By removing this feature, the model is encouraged to learn from primary attributes, resulting in more transparent and interpretable predictions. This experimental design enables a direct comparison between a performance-oriented model that utilizes aggregate information and a more interpretable model that emphasizes individual feature contributions.

Model Development

The prediction model is developed using the XGBoost classifier, an ensemble learning method based on gradient boosting of decision trees. The objective of XGBoost is to minimize a regularized loss function:

$$\mathcal{L}(\phi) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{1}$$

$l(y_i, \hat{y}_i)$is the loss function and $\Omega(f_k)$is the regularization term defined as:

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \parallel w \parallel^2 \tag{2}$$

$T$is the number of leaves and $w$represents leaf weights.

The prediction is constructed as:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i) \tag{3}$$

At each boosting iteration, a new function is added using second-order optimization:

$$\mathcal{L}^{(t)} \approx \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \tag{4}$$

$g_i$and $h_i$are the first and second derivatives of the loss function.

The model is configured with:

Number of estimators: 300

Learning rate: 0.05

Maximum depth: 4

Subsample: 0.9

Colsample_bytree: 0.9

To handle class imbalance, the scale_pos_weight parameter is computed as:

$$\text{scale_pos_weight} = \frac{\text{number of negative samples}}{\text{number of positive samples}} \tag{5}$$

Model Evaluation

The model performance is evaluated using multiple classification metrics to provide a comprehensive assessment of its predictive capability. Accuracy is used to measure the overall correctness of the model by calculating the proportion of correctly classified instances. Precision evaluates the correctness of positive predictions by measuring how many predicted positive cases are actually true positives. Recall assesses the model's ability to identify positive cases by calculating the proportion of actual positives that are correctly detected. The F1-score is used as a balanced metric that combines precision and recall through their harmonic mean, providing a more robust evaluation in the presence of class imbalance. In addition, the ROC-AUC metric is employed to measure the model's ability to distinguish between classes across different classification thresholds. To further analyze prediction performance, a confusion matrix is used to provide a detailed breakdown of true positives, true negatives, false positives, and false negatives. This evaluation framework is applied to both the baseline model and the proposed model to examine the impact of excluding the aggregate score on overall performance.

Explainability Analysis

To improve interpretability, SHAP is used to explain the model predictions. SHAP is based on Shapley values from cooperative game theory and assigns each feature a contribution value.

The SHAP value for feature iis defined as:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N|-|S|-1)!}{|N|!} [f(S \cup \{i\}) - f(S)] \tag{6}$$

$S$is a subset of features

$N$is the set of all features

Global interpretability is obtained using:

SHAP summary plot

SHAP feature importance

Local interpretability is obtained using:

SHAP waterfall plot

This approach allows the model to provide both overall insights and instance-level explanations, making it suitable for high-stakes decision-making scenarios such as student selection.

Algorithm 1: Explainable Student Selection Prediction

Input: dataset D, feature set X, target y

Output: prediction model M, evaluation metrics, SHAP explanations

Process:

Start

Define target variable:

$$y_i = \begin{cases} 1, & \text{if candidate passes} \\ 0, & \text{otherwise} \end{cases}$$

Define feature configurations:

$$X_{total} = X \cup \{\text{skd.total}\}$$
$$X_{no_total} = X \setminus \{\text{skd.total}\}$$

For each feature set $X_j \in \{X_{total}, X_{no_total}\}$:

Preprocessing:

Fill missing numerical values:

$$x_{num} \leftarrow \text{median}(x_{num})$$

Fill missing categorical values:

$$x_{cat} \leftarrow \text{mode}(x_{cat})$$

Apply one-hot encoding to categorical features

Split dataset:

$$D_{train}, D_{test} = \text{split}(D, 80{:}20)$$

Compute class weight:

$$w = \frac{N_{negative}}{N_{positive}}$$

Train XGBoost model:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i)$$

Minimize objective function:

$$\mathcal{L}(\phi) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

Predict output:

$$\hat{y}_i = M(x_i)$$

Evaluate model:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$
$$Precision = \frac{TP}{TP + FP}$$
$$Recall = \frac{TP}{TP + FN}$$
$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

End For

Select final model:

$$M_{final} = M_{no_total}$$

Compute SHAP values:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f(S \cup \{i\}) - f(S)]$$

Generate explanations:

Global: SHAP summary plot

Local: SHAP waterfall plot

End

Results

Model Performance

The performance of the proposed XGBoost model was evaluated using two configurations, namely a baseline model that includes the aggregate score (skd.total) and a proposed model that excludes this feature. The comparison between these two models is presented in table 1. The baseline model achieved higher performance across all evaluation metrics, with an accuracy of 97.59%, F1-score of 0.906, and ROC-AUC of 0.9976. In contrast, the proposed model without the aggregate score achieved an accuracy of 95.84%, F1-score of 0.844, and ROC-AUC of 0.9935. Although there is a measurable decrease in performance, the difference remains relatively small considering the removal of a highly informative feature.

The results indicate that the proposed model is still able to maintain strong predictive capability even without relying on the aggregate score. This suggests that the individual test components, including TWK, TIU, and TKP, provide sufficient information for accurate prediction of selection outcomes. The high ROC-AUC value further confirms that the model can effectively distinguish between passing and non-passing candidates. These findings support the idea that aggregate scores may not be strictly necessary for building accurate predictive models, and that more interpretable models can be developed by focusing on individual feature contributions.

Table 1 Performance comparison of XGBoost models

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
XGBoost With skd.total	0.9759	0.8383	0.9854	0.9059	0.9976
XGBoost Without skd.total	0.9584	0.7540	0.9591	0.8443	0.9935

Confusion Matrix Analysis

The confusion matrix of the proposed model is presented in figure 2, illustrating the distribution of correct and incorrect predictions across both classes. The model correctly classified 2,460 candidates as not passing and 328 candidates as passing, demonstrating strong overall classification performance. Notably, the number of false negatives is very low, with only 14 cases where successful candidates were incorrectly predicted as unsuccessful. This indicates that the model is highly effective in identifying candidates who should pass the selection process. On the other hand, there are 107 false positives, meaning that some candidates who did not pass were predicted as passing. This trade-off is reflected in the high recall value of 0.96, which shows that the model prioritizes capturing as many successful candidates as possible, even at the cost of slightly increasing the number of false positives.

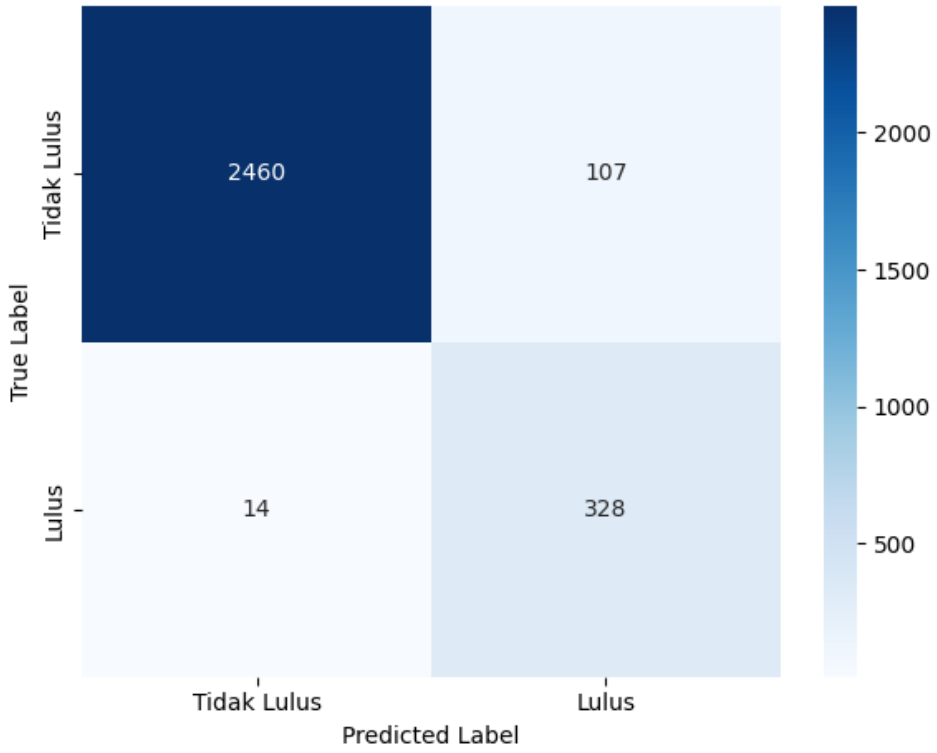


Figure 2 **Confusion matrix of XGBoost model without aggregate score**

Feature Importance Analysis

The feature importance derived from the XGBoost model is presented in figure 3, highlighting the relative contribution of each feature to the prediction task. The results show that skd.tiu is the most influential feature, followed by skd.twk, kode.formasi, and skd.tkp, indicating that cognitive ability measures play a central role in determining selection outcomes. The prominence of TIU and TWK suggests that analytical and general knowledge competencies are key factors in distinguishing successful candidates. In addition to these academic variables, non-academic features such as kode.formasi and lokasi.formasi also exhibit meaningful contributions to the model. This indicates that contextual or structural factors related to program placement and geographic allocation may influence the selection results. The presence of these variables among the top features suggests that the decision process is not solely driven by test performance, but also shaped by institutional or regional considerations.

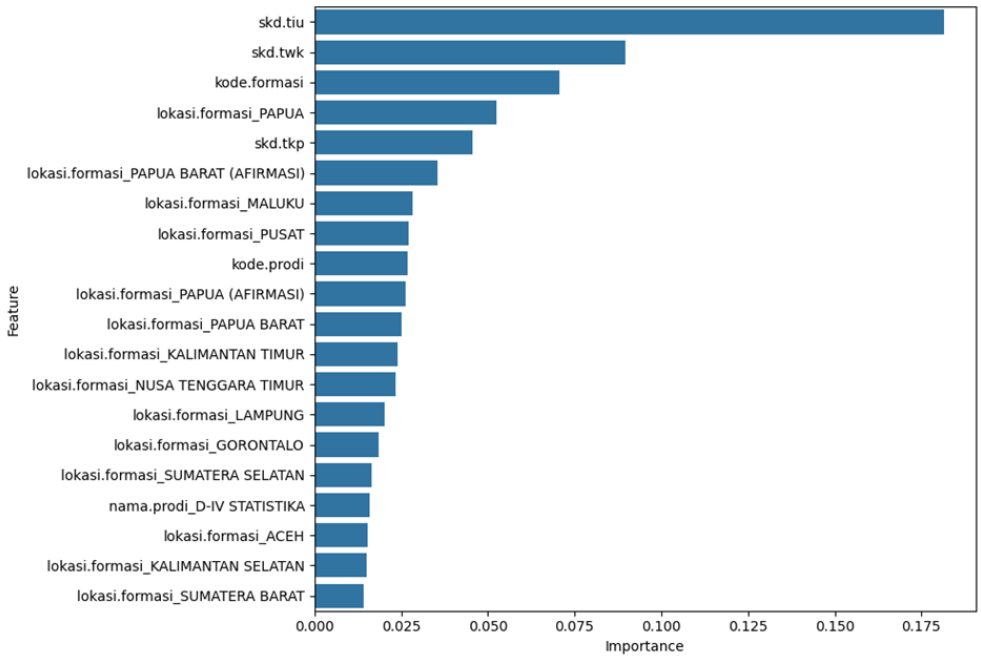


Figure 3 Top 20 feature importance from XGBoost model

SHAP Global Interpretation

The SHAP summary plot is presented in figure 4, providing a global interpretation of how each feature contributes to the model predictions across all samples. The plot shows that higher values of skd.tiu, skd.twk, and skd.tkp consistently increase the probability of passing, as indicated by the concentration of positive SHAP values associated with higher feature magnitudes. In contrast, lower values of these features tend to produce negative SHAP values, pushing predictions toward the "not passing" class. The distribution of SHAP values also reveals the relative importance and impact range of each feature, where skd.tiu exhibits the widest spread, indicating a strong and varying influence on the model output. Overall, the visualization demonstrates a stable and interpretable relationship between feature values and prediction outcomes, confirming that the model behaves consistently and aligns with expected patterns in student performance data.

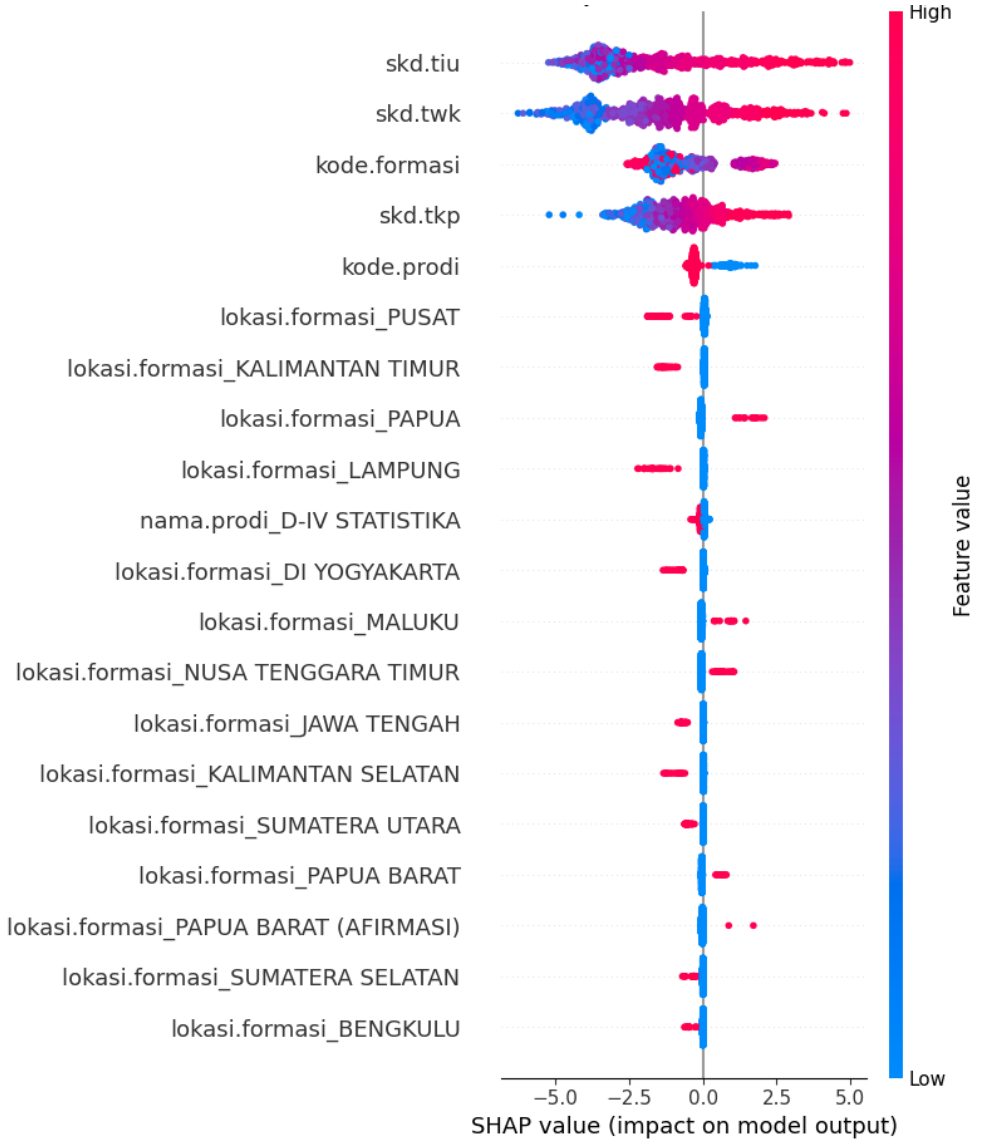


Figure 4 SHAP summary plot showing global feature impact

SHAP Feature Importance

The SHAP-based feature importance is illustrated in figure 5, providing a global ranking of features based on their average contribution to the model predictions. The results confirm that skd.tiu and skd.twk are the most influential features, followed by kode.formasi and skd.tkp, which aligns with the findings from the model-based feature importance. The dominance of TIU and TWK highlights the critical role of cognitive and reasoning abilities in determining selection outcomes, while the presence of kode.formasi indicates that structural or program-related factors also play a meaningful role. Unlike traditional importance measures, SHAP evaluates feature contributions across all instances, making the interpretation more consistent and reliable. The agreement between SHAP-based importance and XGBoost feature importance suggests that the model captures stable and meaningful patterns in the data, thereby strengthening the validity and robustness of the overall findings.

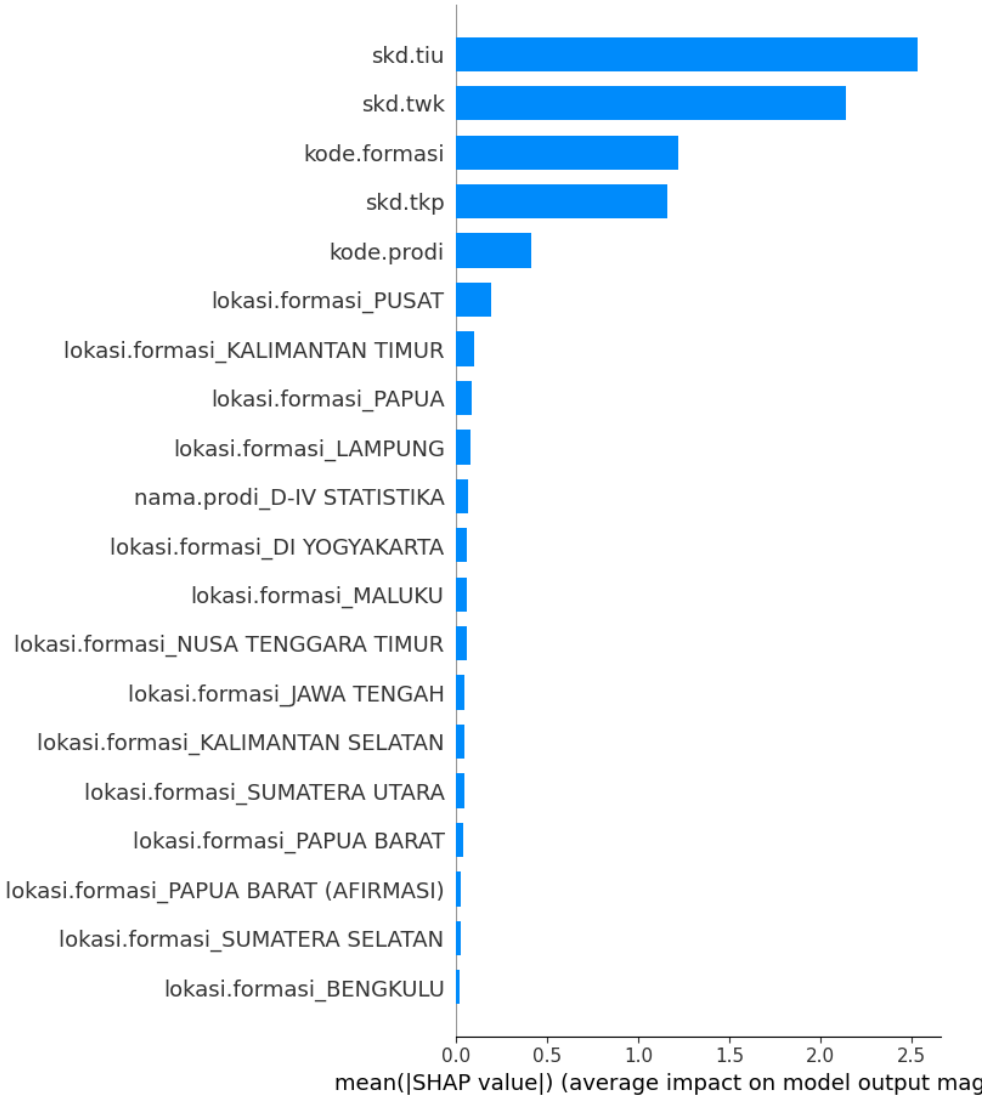


Figure 5 SHAP feature importance ranking

Local Explanation of Model Prediction

An example of local explanation using SHAP is shown in figure 6, illustrating how individual features contribute to a single prediction made by the model. In this case, the model predicts a candidate as "not passing" with a very low probability of passing, equal to 0.00018, indicating strong confidence in the decision. The SHAP waterfall plot reveals that low values of skd.twk and skd.tiu are the primary factors driving the prediction toward the negative class, with substantial negative contributions that significantly reduce the predicted probability. Although there are positive contributions from other features such as skd.tkp and program-related variables, their influence is relatively small and insufficient to counterbalance the strong negative impact of the cognitive test scores. The step-by-step accumulation of SHAP values clearly shows how each feature shifts the prediction from the base value toward the final output. This example demonstrates the model's ability to provide transparent, instance-level explanations, allowing stakeholders to understand not only what the prediction is, but also why it was made.

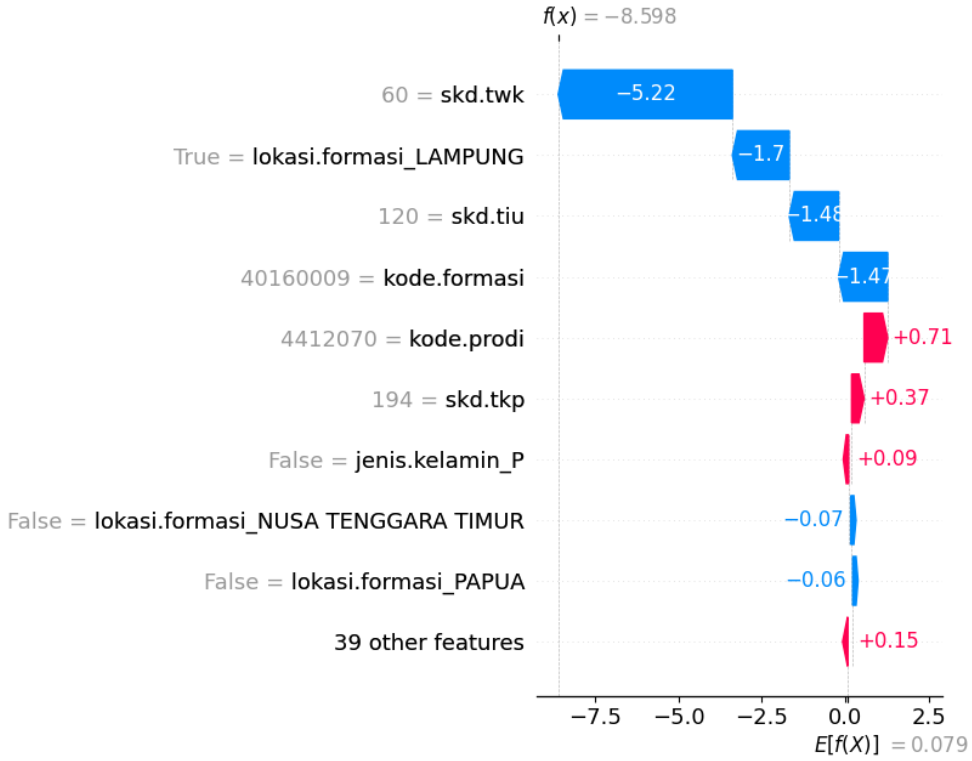


Figure 6 SHAP waterfall plot for individual prediction

Discussion

The results of this study demonstrate that accurate prediction of student selection outcomes can be achieved without relying on aggregate scores such as skd.total. Although the inclusion of this feature improves overall performance, the proposed model without it still achieves high accuracy and strong discriminative ability, indicating that the individual test components already contain sufficient predictive information. This finding is important because aggregate scores are derived from the same underlying variables and may introduce redundancy while reducing interpretability. By removing the aggregate score, the model is forced to learn from more granular features, which allows for clearer insight into how each component contributes to the final decision. The consistent performance across evaluation metrics suggests that the model generalizes well and is not overly dependent on a single dominant feature.

Further analysis using feature importance and SHAP reveals that cognitive test components, particularly skd.tiu and skd.twk, are the most influential factors in determining selection outcomes, while skd.tkp also contributes meaningfully. In addition, non-academic variables such as kode.formasi and lokasi.formasi appear among the important features, indicating that contextual or structural factors may influence the selection process. This raises important considerations regarding fairness, as decisions may not be solely based on academic performance. The confusion matrix results show that the model achieves high recall, meaning it effectively identifies candidates who should pass, with very few false negatives. At the same time, the presence of some false positives indicates a tendency to classify certain unsuccessful candidates as passing, reflecting a trade-off that may align with policy goals depending on

the context. The use of SHAP further strengthens the study by providing both global and local interpretability, enabling a clear understanding of how predictions are made at both the dataset and individual levels, which is essential for transparency and trust in decision-making systems.

Conclusion

This study demonstrates that an explainable machine learning approach can effectively predict student selection outcomes using SKD examination data while maintaining both high performance and interpretability. The experimental results show that although the inclusion of the aggregate score slightly improves model performance, the proposed model without skd.total still achieves strong accuracy and a high ROC-AUC, indicating that individual test components such as TWK, TIU, and TKP provide sufficient predictive information. The use of SHAP further enhances the model by enabling transparent interpretation at both global and individual levels, revealing that cognitive abilities, particularly TIU and TWK, are the most influential factors, while non-academic variables such as formation and location also contribute to the prediction. These findings suggest that student selection decisions are influenced not only by academic performance but also by contextual factors, which has important implications for fairness and policy evaluation. Overall, the integration of machine learning and explainable AI provides a robust and transparent framework that can support more informed, accountable, and trustworthy decision-making in student selection systems.

Declarations

Author Contributions

Conceptualization: M.A.; Methodology: M.A.; Software: M.A.; Validation: M.A.; Formal Analysis: M.A.; Investigation: M.A.; Resources: M.A.; Data Curation: M.A.; Writing Original Draft Preparation: M.A.; Writing Review and Editing: M.A.; Visualization: M.A.; All authors have read and agreed to the published version of the manuscript.

Data Availability Statement

The data presented in this study are available on request from the corresponding author.

Funding

The authors received no financial support for the research, authorship, andor publication of this article.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] S. B. Kotsiantis, “Use of machine learning techniques for educational purposes: A decision support system for forecasting students’ grades,” *Artificial Intelligence Review*, vol. 37, no. 4, pp. 331–344, 2012, doi: 10.1007/s10462-011-9234-x.

[2] C. Romero and S. Ventura, “Educational data mining: A review of the state of the art,” *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 40, no. 6, pp. 601–618, Nov. 2010, doi: 10.1109/TSMCC.2010.2053532.

[3] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD ’16)*, San Francisco, CA, USA, Aug. 2016, vol. 2016, no. August, pp. 785–794, doi: 10.1145/2939672.2939785.

[4] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in *Advances in Neural Information Processing Systems*, vol. 30, no. May, pp. 1-10, 2017, pp. 4765–4774.

[5] Z. C. Lipton, “The mythos of model interpretability,” *Communications of the ACM*, vol. 61, no. 10, pp. 36–43, Oct. 2018, doi: 10.1145/3233231.

[6] J. Luan, “Data mining and its applications in higher education,” *New Directions for Institutional Research*, vol. 2002, no. 113, pp. 17–36, Spring 2002, doi: 10.1002/ir.35.

[7] E. Osmanbegović and M. Suljić, “Data mining approach for predicting student performance,” *Economic Review: Journal of Economics and Business*, vol. 10, no. 1, pp. 3–12, 2012.

[8] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[9] J. H. Friedman, “Greedy function approximation: A gradient boosting machine,” *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, Oct. 2001, doi: 10.1214/aos/1013203451.

[10] M. Imran, S. Latif, D. Mehmood, and M. S. Shah, “Student academic performance prediction using supervised learning techniques,” *International Journal of Emerging Technologies in Learning*, vol. 14, no. 14, pp. 92–104, Jul. 2019, doi: 10.3991/ijet.v14i14.10310.

[11] S. K. Mohamad and Z. Tasir, “Educational data mining: A review,” *Procedia - Social and Behavioral Sciences*, vol. 97, no. November, pp. 320–324, Nov. 2013, doi: 10.1016/j.sbspro.2013.10.240.

[12] B. D. Mittelstadt, P. Allo, M. Taddeo, S. Wachter, and L. Floridi, “The ethics of algorithms: Mapping the debate,” *Big Data & Society*, vol. 3, no. 2, pp. 1–21, Nov. 2016, doi: 10.1177/2053951716679679.

[13] D. Gunning and D. W. Aha, “DARPA’s explainable artificial intelligence (XAI) program,” *AI Magazine*, vol. 40, no. 2, pp. 44–58, 2019, doi: 10.1609/aimag.v40i2.2850.

[14] S. M. Lundberg et al., “From local explanations to global understanding with explainable AI for trees,” *Nature Machine Intelligence*, vol. 2, no. January, pp. 56–67, Jan. 2020, doi: 10.1038/s42256-019-0138-9.

[15] M. T. Ribeiro, S. Singh, and C. Guestrin, “‘Why should I trust you?’: Explaining the

predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*, San Francisco, CA, USA, 2016, vol. 2016, no. August, pp. 1135–1144, doi: 10.1145/2939672.2939778.

[16] A. Adadi and M. Berrada, "Peeking inside the black box: A survey on explainable artificial intelligence (XAI)," *IEEE Access*, vol. 6, no. September, pp. 52138–52160, 2018, doi: 10.1109/ACCESS.2018.2870052.

[17] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM Computing Surveys*, vol. 51, no. 5, Art. no. 93, pp. 1–42, Aug. 2018, doi: 10.1145/3236009.

[18] D. J. Hand, "Classifier technology and the illusion of progress," *Statistical Science*, vol. 21, no. 1, pp. 1–14, 2006, doi: 10.1214/088342306000000060.

[19] L. De Pin, "S. Barocas, M. Hardt and A. Narayanan. Fairness and Machine Learning: Limitations and Opportunities. MIT Press, 2023", *tr*, vol. 20, no. 1, Feb. 2026, doi: 10.54103/1757-0522/29851.

[20] H. Sahlaoui, E. A. Abdellaoui Alaoui, A. Nayyar, S. Agoujil, and M. M. Jaber, "Predicting and interpreting student performance using ensemble models and Shapley additive explanations," *IEEE Access*, vol. 9, no. October, pp. 152688–152703, 2021, doi: 10.1109/ACCESS.2021.3124270.

[21] H. Mastour, T. Dehghani, E. Moradi, and S. Eslami, "Early prediction of medical students' performance in high-stakes examinations using machine learning approaches," *Heliyon*, vol. 9, no. 7, Art. no. e18248, Jul. 2023, doi: 10.1016/j.heliyon.2023.e18248.

[22] S. Gunasekara and M. Saarela, "Explainable AI in education: Techniques and qualitative assessment," *Applied Sciences*, vol. 15, no. 3, Art. no. 1239, Jan. 2025, doi: 10.3390/app15031239.

[23] A. Zanellati, S. P. Zingaro, and M. Gabbrielli, "Balancing performance and explainability in academic dropout prediction," *IEEE Transactions on Learning Technologies*, vol. 17, no. July, pp. 2086–2099, 2024, doi: 10.1109/TLT.2024.3425959.

[24] G. Raftopoulos, G. Davrazos, and S. Kotsiantis, "Fair and transparent student admission prediction using machine learning models," *Algorithms*, vol. 17, no. 12, Art. no. 572, Dec. 2024, doi: 10.3390/a17120572.

[25] B. Ujkani, D. Minkovska, and N. Hinov, "Course success prediction and early identification of at-risk students using explainable artificial intelligence," *Electronics*, vol. 13, no. 21, Art. no. 4157, Oct. 2024, doi: 10.3390/electronics13214157.