Predicting Student Achievement Status Using Decision Tree and Random Forest with SHAP in Vocational School Admissions

Irwan Syahputra, Solly Aryza, Muhammad Iqbal

Abstract

This study investigates the ability of Decision Tree and Random Forest models to predict student achievement status using a synthetic vocational-school admission dataset. The dataset contains 1,000 records with Achievement as the target, consisting of 215 achievement and 785 non-achievement records. Predictors include academic year, school origin, vocational program, report-card score, Mathematics, Indonesian Language, English, selection-test score, and home distance. The data were divided using a stratified 80:20 train-test split, and both models were evaluated using accuracy, balanced accuracy, precision, recall, specificity, F1-score, ROC-AUC, MCC, and confusion matrices. Random Forest achieved 77.00% accuracy but only 51.58% balanced accuracy and 45.28% ROC-AUC, whereas Decision Tree obtained 57.00% accuracy, 46.44% balanced accuracy, and 46.05% ROC-AUC. SHAP analysis of Random Forest identified vocational program, Mathematics score, English score, and selection-test score as the strongest global contributors. The results demonstrate that high accuracy on imbalanced data does not necessarily indicate strong discriminative performance. The synthetic dataset does not contain sufficiently strong predictive signals for reliable achievement classification, highlighting the need for validation using real student-achievement data.

Keywords: student achievement; Decision Tree; Random Forest; SHAP; educational data mining.

Irwan Syahputra
Magister Teknologi Informasi, Universitas Pembangunan Panca Budi, Indonesia
e-mail: isyah6845@gmail.com

Solly Aryza, Muhammad Iqbal
e-mail: sollyaryzalubis@pancabudi.ac.id , wakbalpb@yahoo.co.id
2nd International Conference on the Epicentrum of Economic Global Framework (ICEEGLOF)
Theme: Navigating The Future: Business and Social Paradigms in a Transformative Era.
https://proceeding.pancabudi.ac.id/index.php/ICEEGLOF

Introduction

Educational institutions increasingly generate structured data during student admission and learning activities. These data can support educational data mining and predictive analytics, including the identification of students with different achievement profiles. Machine-learning models are useful for discovering nonlinear relationships that are difficult to identify using conventional descriptive analysis.

Tree-based algorithms are frequently applied to educational prediction because they can handle mixed data types and nonlinear interactions. Decision Tree provides a relatively transparent rule structure, while Random Forest aggregates multiple trees to improve stability and reduce variance. Previous admission studies have shown that Random Forest can be competitive with other classification models in predicting prospective-student outcomes [1].

However, predictive performance should not be evaluated using accuracy alone, particularly when the target class is imbalanced. Educational data-mining studies have shown that imbalance can produce misleadingly high accuracy while minority-class recognition remains poor [2]. Therefore, metrics such as balanced accuracy, recall, specificity, ROC-AUC, and MCC are important for a more complete assessment.

Interpretability is also essential in educational applications. SHapley Additive exPlanations (SHAP) provides a unified framework for explaining how individual features contribute to model predictions [3]. In admission-related applications, transparency and fairness are increasingly considered alongside predictive performance [4].

This study focuses on student achievement status rather than admission status. It examines whether academic and admission-time characteristics in a synthetic vocational-school dataset provide sufficient information to predict whether a student has an achievement record. The study also uses SHAP to identify which features most strongly influence the Random Forest model. The research question is: can admission-time academic characteristics reliably predict achievement status, and which factors contribute most strongly to the model output?

Literature Review

Brianorman and Sucipto compared Decision Tree, Random Forest, and Naive Bayes for prospective-student prediction and found that Random Forest slightly outperformed the other models in accuracy [1]. Their work supports the relevance of ensemble tree methods for admission analytics, although the prediction target differs from the achievement-status target examined in this paper.

Class imbalance is a recurring issue in educational prediction. Wongvorachan et al. compared several resampling techniques for educational data mining and emphasized that model evaluation should consider the imbalance ratio and minority-class performance [2]. This consideration is directly relevant to the present dataset, where only 21.5% of records belong to the achievement class.

SHAP, introduced by Lundberg and Lee, assigns feature-level contributions to individual predictions based on Shapley values [3]. Tree-based SHAP methods later enabled efficient explanations of models such as Decision Tree and Random Forest [5]. Explainability is especially important in education because prediction systems should support human interpretation rather than provide opaque automated judgments.

Raftopoulos et al. examined fairness and transparency in student admission prediction and demonstrated that performance metrics alone are insufficient for evaluating educational decision models [4]. Accordingly, the present study treats SHAP as an analytical and auditing mechanism rather than as proof of causal relationships between predictors and student achievement.

Research Methodology

The study uses the uploaded synthetic dataset dataset_sintetis_ppdb_dwitunggal2_1000.xlsx containing 1,000 records and 15 columns. The target variable is Prestasi (Achievement), encoded as 1 for 'Ada' and 0 for 'Tidak ada'. The target distribution is 215 achievement records (21.5%) and 785 non-

achievement records (78.5%). ID_Calon and Status_Penerimaan were excluded to avoid identifiers and downstream information. Jenis_Kelamin, Kategori_Ekonomi, and Sumber_Informasi were also excluded from the core model to reduce dependence on non-academic or potentially sensitive attributes.

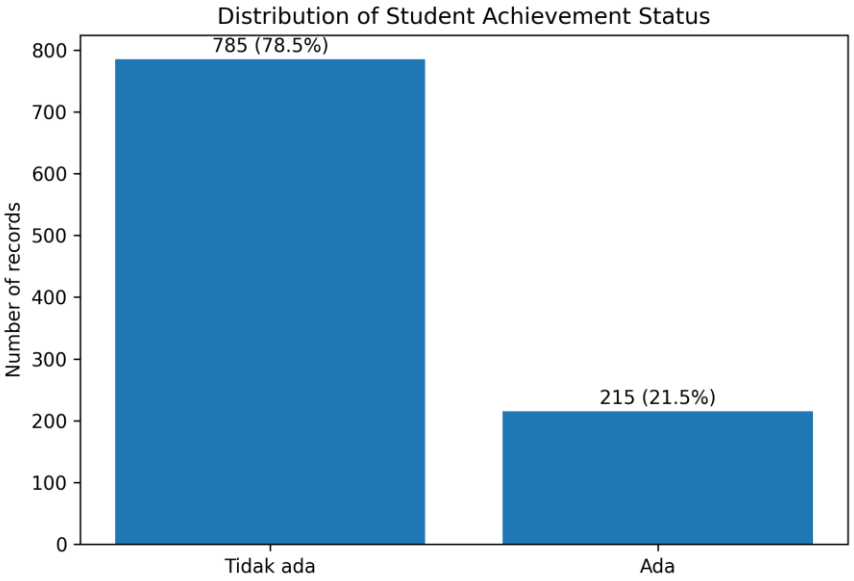


Figure 1. Distribution of student achievement status.

The predictor set consists of Tahun_Ajaran, Asal_Sekolah, Program_Keahlian, Nilai_Rapor, Nilai_Matematika, Nilai_Bahasa_Indonesia, Nilai_Bahasa_Inggris, Nilai_Tes_Seleksi, and Jarak_Rumah_km. Categorical variables were transformed using one-hot encoding, while numeric variables were passed directly to the tree-based models.

The dataset was divided using a stratified 80:20 train-test split with random_state=42. Decision Tree was configured with max_depth=5, min_samples_leaf=5, and balanced class weights. Random Forest used 250 trees, max_depth=8, min_samples_leaf=3, sqrt feature sampling, and balanced class weights. These settings were selected for a reproducible preliminary experiment rather than claimed as globally optimal hyperparameters.

Model performance was assessed using accuracy, balanced accuracy, precision, recall, specificity, F1-score, ROC-AUC, MCC, and confusion matrices. Because the target was imbalanced, balanced accuracy and ROC-AUC were emphasized alongside accuracy. SHAP TreeExplainer was then applied to the Random Forest model. Mean absolute SHAP values were aggregated to original variables to obtain global feature importance.

Results

Metric	Decision Tree	Random Forest
Accuracy	57.00%	77.00%
Balanced Accuracy	46.44%	51.58%
Precision	17.91%	33.33%
Recall	27.91%	6.98%
Specificity	64.97%	96.18%
F1-score	21.82%	11.54%
ROC-AUC	46.05%	45.28%
MCC	-0.0620	0.0625

Table 1. Performance comparison of Decision Tree and Random Forest.

Table 1 shows that Random Forest achieved substantially higher overall accuracy than Decision Tree (77.00% versus 57.00%). However, this difference should not be interpreted as strong evidence of superior achievement prediction. The majority class represents 78.5% of the dataset, meaning that a model can achieve high accuracy by predominantly predicting the non-achievement class.

The balanced accuracy of Random Forest was only 51.58%, close to random classification, while ROC-AUC was 45.28%. Decision Tree also produced weak discrimination, with 46.44% balanced accuracy and 46.05% ROC-AUC. These results indicate that the available predictors do not provide a strong separable signal for the Prestasi label in this synthetic dataset.

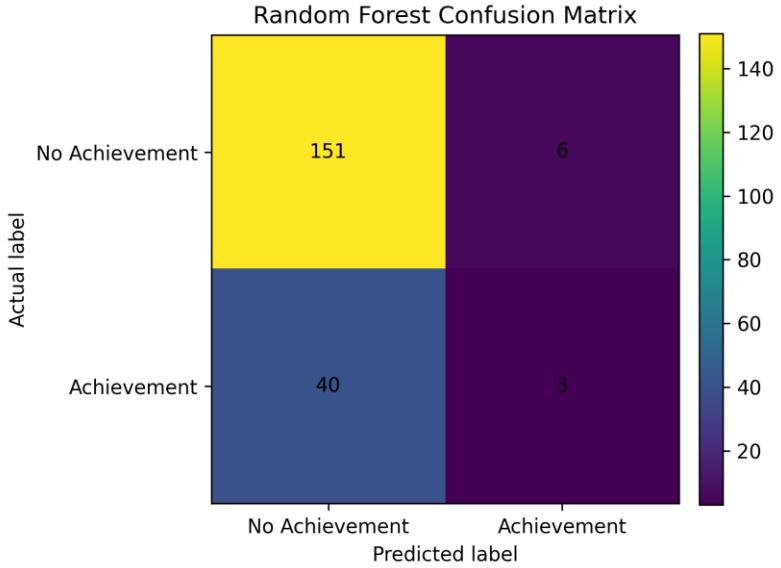


Figure 2. Confusion matrix of the Random Forest achievement-status model.

Figure 2 illustrates the limitation more clearly. Random Forest correctly classified 151 non-achievement records but identified only 3 of 43 achievement records in the test set. Forty achievement records were classified as non-achievement. Consequently, the model reached high specificity (96.18%) but very low recall for the achievement class (6.98%). This demonstrates why accuracy alone is inadequate for this problem.

Rank	Variable	Mean |SHAP|
1	Program_Keahlian	0.03324
2	Nilai_Matematika	0.02654
3	Nilai_Bahasa_Inggris	0.02500
4	Nilai_Tes_Seleksi	0.02297
5	Tahun_Ajaran	0.02179
6	Jarak_Rumah_km	0.02145
7	Nilai_Bahasa_Indonesia	0.01970
8	Nilai_Rapor	0.01737

Table 2. Global SHAP importance of the Random Forest model.

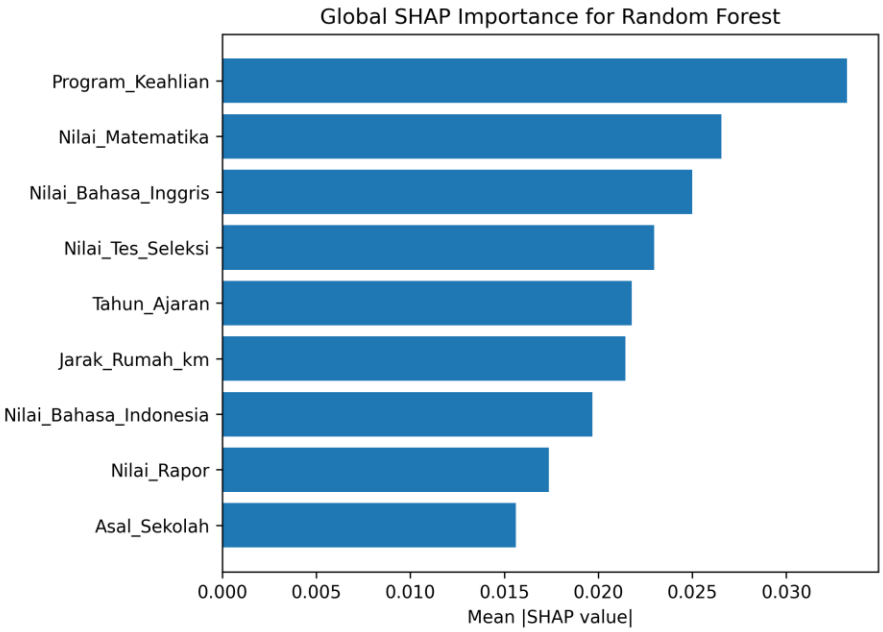


Figure 3. Global SHAP importance for Random Forest.

The SHAP analysis indicates that Program_Keahlian has the largest aggregated contribution, followed by Nilai_Matematika, Nilai_Bahasa_Inggris, Nilai_Tes_Seleksi, Tahun_Ajaran, and Jarak_Rumah_km. However, the low ROC-AUC of the underlying model means that these SHAP rankings must be interpreted cautiously. SHAP explains how the model uses the available features; it does not establish that these variables have strong or causal relationships with student achievement.

The results suggest that the synthetic data-generation process does not encode sufficiently strong relationships between academic admission variables and the Prestasi label. For a publication-oriented empirical study, this limitation can be addressed by using anonymized real achievement data, defining achievement more precisely, including longitudinal academic outcomes, or redesigning the synthetic-data generator so that the target reflects realistic educational relationships. The present findings are therefore most appropriately interpreted as a methodological simulation rather than evidence about actual students.

Conclusion

This study evaluated Decision Tree and Random Forest models for predicting student achievement status in a synthetic vocational-school admission dataset and used SHAP to explain the Random Forest model. Random Forest produced higher accuracy (77.00%) than Decision Tree (57.00%), but its balanced accuracy (51.58%) and ROC-AUC (45.28%) showed that the model had limited discriminative ability. The confusion matrix confirmed that Random Forest strongly favored the majority non-achievement class and detected only a small proportion of achievement cases. SHAP ranked vocational program, Mathematics score, English score, and selection-test score among the most influential features, but these explanations should be interpreted only as model-behavior descriptions because the predictive signal itself was weak. The main implication is that evaluating imbalanced educational data requires metrics beyond accuracy and that explainability cannot compensate for an uninformative target-data relationship. Future research should validate the approach using real, anonymized student-achievement data and longitudinal outcomes.

References

[1] Y. Brianorman and S. Sucipto, "Prediction of Prospective New Students Using Decision Tree, Random Forest, and Naive Bayes," BAREKENG: J. Math. & App., vol. 18, no. 3, pp. 1433-1446, 2024, doi: 10.30598/barekengvol18iss3pp1433-1446.

[2] T. Wongvorachan, S. He, and O. Bulut, "A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining," Information, vol. 14, no. 1, art. 54, 2023, doi: 10.3390/info14010054.

[3] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in Advances in Neural Information Processing Systems, vol. 30, 2017.

[4] G. Raftopoulos, G. Davrazos, and S. Kotsiantis, "Fair and Transparent Student Admission Prediction Using Machine Learning Models," Algorithms, vol. 17, no. 12, art. 572, 2024, doi:

[5] S. M. Lundberg et al., "From Local Explanations to Global Understanding with Explainable AI for Trees," Nature Machine Intelligence, vol. 2, pp. 56-67, 2020, doi: 10.1038/s42256-019-0138-9.

[6] L. Breiman, "Random Forests," Machine Learning, vol. 45, pp. 5-32, 2001, doi: 10.1023/A:1010933404324.

[7] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 1135-1144, doi: 10.1145/2939672.2939778.

[8] T. Fawcett, "An Introduction to ROC Analysis," Pattern Recognition Letters, vol. 27, no. 8, pp. 861-874, 2006, doi: 10.1016/j.patrec.2005.10.010.