ISI Journal of Information Systems and Informatics

Published By Asosiasi Doktor Sistem Informasi Indonesia

Stacking Ensemble Learning for University Student Dropout Prediction

Aden Nia Firdaus[1], Yoannes Romando Sipayung[2]

[1,2,3] Informatics and Computer Engineering Education Program, Univeristas Ngudi Waluyo, Indonesia

Received:
November 13, 2025
Revised:
January 12, 2026
Accepted:
January 27, 2026
Published:
February 18, 2026

Corresponding Author:
Author Name*:
Aden Nia Firdaus
Email*:
adennia010304@gmail.com

DOI:
10.63158/journalisi.v8i1.1403

Abstract. Student dropout in STEM programs remains a persistent challenge for higher education institutions, reducing educational quality, weakening retention outcomes, and increasing inefficiencies in resource utilization. This study develops an interpretable Stacking Ensemble Learning approach to predict STEM student dropout risk and identify key academic and socioeconomic determinants that can support data-driven early intervention. Following the CRISP-DM framework, we analyze 3,630 student records from the UCI Machine Learning Repository containing demographic, academic, and socioeconomic attributes. The proposed stacking architecture combines Random Forest, Gradient Boosting, and XGBoost as base learners with Logistic Regression as a meta-learner, while SMOTE–Tomek Links is employed to address class imbalance and reduce boundary noise. Experimental results show that the model achieves strong predictive performance with 90.91% accuracy and ROC–AUC of 95.72%, demonstrating stable discrimination and outperforming individual base models. Feature importance analysis indicates that early academic trajectory variables—especially first- and second-semester success rates, total approved units, and average grades—are the most influential predictors of dropout risk. The proposed framework contributes a practical, interpretable early warning model by integrating stacking ensemble learning with imbalance handling and trajectory-based feature engineering, supporting actionable intervention planning in higher education.

Keywords: Stacking Ensemble Learning; Student Dropout Prediction; STEM Education; SMOTE–Tomek Links; Educational Data Mining

1. INTRODUCTION

Student dropout in STEM programs remains a persistent and high-impact challenge for higher education institutions, especially in Indonesia where national policies increasingly tie institutional accountability to graduation rates, retention, and data-driven quality assurance. When STEM students discontinue their studies, the consequences extend beyond individual academic trajectories: institutions lose tuition revenue and efficiency in resource allocation, program performance indicators weaken, and the national agenda to build digitally capable, industry-ready human capital is slowed. In this context, reducing dropout is not only an educational concern but also a strategic requirement linked to accreditation outcomes, institutional performance evaluation, and government-driven initiatives that prioritize timely completion and improved academic services.

Despite this urgency, dropout in academically demanding STEM programs is shaped by interacting factors that universities often struggle to monitor early and consistently. The simultaneous decline in enrollment and the rise in dropout rates underscore how socioeconomic constraints and academic performance can combine to disrupt students' continuity while also reducing institutional efficiency [1]. Beyond institutional metrics, dropout also affects educational quality, human resource development, financial sustainability, and graduates' competitiveness in the labor market [2]. These realities make early identification of at-risk students a practical necessity: the earlier universities can detect risk, the more feasible it becomes to deliver targeted academic support, financial guidance, and advising interventions before students disengage irreversibly.

Recent advances in Artificial Intelligence (AI) and Machine Learning (ML) have expanded the feasibility of analyzing large-scale academic records to predict dropout risk. Accordingly, many studies model dropout prediction as a binary classification task and report promising performance using various ML algorithms [3] [4]. However, much of this work emphasizes predictive accuracy as the primary objective, often leaving a critical institutional need insufficiently addressed: interpretability. For academic leaders and program managers, knowing who is at risk is not enough—effective policy and intervention design requires understanding why students are at risk, which factors are most influential, and how those factors can be translated into actionable support strategies. Without interpretable evidence, even high-performing predictive models may remain underutilized in real academic decision-making.

Ensemble learning offers a strong pathway to improve robustness and generalization by combining multiple models rather than relying on a single learner [5]. Among ensemble approaches, Stacking Ensemble Learning (SEL) integrates diverse base learners and optimizes their combined output through a meta-learner—commonly Logistic Regression—to improve accuracy and stability [6]. Although SEL has demonstrated strong performance in educational analytics, its use for STEM dropout prediction with explicit attention to interpretability and operational deployment remains limited, particularly within Indonesian higher education settings. This gap matters because institutional adoption depends not only on performance metrics but also on transparent explanations that align with advising workflows, policy targets, and quality assurance requirements.

To address this gap, this study develops an interpretable Stacking Ensemble Learning model for predicting STEM student dropout risk while simultaneously identifying key academic and socioeconomic determinants. Using the CRISP-DM framework, the proposed approach integrates Random Forest, Gradient Boosting, and XGBoost as base learners with Logistic Regression as a meta-learner, and applies SMOTE–Tomek Links to mitigate class imbalance and improve minority-class discrimination. Experimental results demonstrate strong predictive performance, achieving 90.91% accuracy and a ROC–AUC of 95.72%, while maintaining stable class separation. Beyond performance, the novelty of this work lies in pairing a high-performing SEL architecture with an interpretability-driven analysis that surfaces concrete, institutionally meaningful drivers of risk—bridging the common disconnect between model accuracy and decision usability in prior studies.

The contributions of this research are threefold. First, it implements a stacking ensemble that combines Random Forest, XGBoost, and Gradient Boosting to enhance predictive performance for STEM dropout classification in an Indonesian university context. Second, it prioritizes interpretability through feature importance analysis to clarify which academic and non-academic variables most strongly influence dropout risk, supporting explainable early-warning decisions rather than opaque predictions. Third, it translates the model's insights into strategic guidance for universities to design data-driven intervention programs aligned with retention targets and national quality assurance priorities. Feature importance results emphasize early academic trajectory indicators—particularly first- and second-semester success rates, total approved units, and average grades—as the most influential predictors of dropout risk, reinforcing the practical value of embedding the model into academic information systems as an actionable early warning tool.

2. METHODS

This study adopts the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework because it offers a well-established, industry-standard methodology for organizing machine learning and data mining work into a clear, auditable, and repeatable process [7]. CRISP-DM is particularly suitable for educational prediction tasks because it connects technical modeling choices (e.g., feature engineering, imbalance handling, algorithm selection) to institutional goals (e.g., early warning and intervention). Importantly, CRISP-DM is iterative: insights from evaluation can trigger revisions to earlier stages (such as redefining success criteria, improving data preprocessing, or refining features), supporting continuous model improvement rather than a one-pass pipeline [8]. The six interconnected phases used in this study are summarized in Figure 1 and operationalized through the activities described in the following subsections.

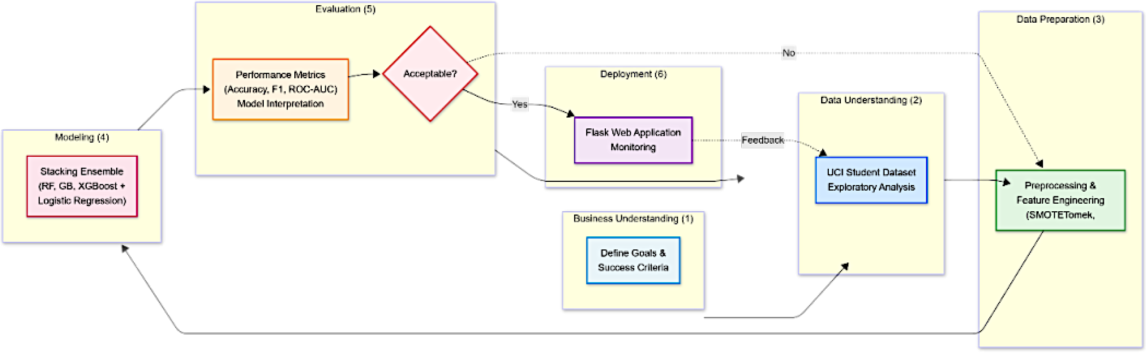

Figure 1. CRISP-DM Methodological Framework for Predicting Students' Academic and Socioeconomic Performance

CRISP-DM was selected because it is comprehensive yet flexible, enabling systematic planning while remaining adaptable to the practical issues commonly encountered in real datasets—missing values, noisy labels, class imbalance, and multicollinearity. Its phase-by-phase structure also helps ensure methodological transparency, which is essential when developing decision-support tools in higher education environments where model outputs must be interpretable and actionable. Table 1 presents the CRISP-DM phases and their core objectives, which guide the overall workflow in this research.

Table 1. CRISP-DM Process Stages

Phase	Brief Description
Business Understanding	Define business and data mining objectives and establish project success criteria.
Data Understanding	Collect, explore, and verify data quality.
Data Preparation	Select, clean, and transform data to make it suitable for modeling.
Modelling	Select modeling techniques, build models, and adjust parameters to achieve optimal results.
Evaluation	Evaluate models against business objectives and predetermined criteria.
Deployments	Implement models in the form of ready-to-use reports, systems, or applications.

The mapping between CRISP-DM phases and the structure of this paper is provided in Table 4 to maintain traceability between methodological stages and reported outcomes, ensuring that each modeling decision can be connected back to a specific phase and objective.

2.1. Business Understanding

The business understanding phase clarifies the problem definition, translates institutional needs into analytical goals, and defines measurable success criteria for the project [10]. In this study, the primary objective is to develop a predictive system that classifies student outcomes as Dropout or Graduate using a Stacking Ensemble Learning approach. The practical motivation is to support higher education institutions in early identification of at-risk students and to strengthen data-driven decision-making for improving graduation outcomes and retention strategies.

Success criteria are established in two complementary dimensions. First, the model must achieve competitive predictive performance aligned with benchmarks reported in comparable dropout prediction research (e.g., robust accuracy and discriminative power). Second, the model must support institutional usability by producing outputs that can inform intervention planning—meaning that interpretability is treated as a requirement rather than an optional add-on. This phase therefore defines a dual target: high classification performance and clear identification of key drivers that universities can act upon.

2.2. Data Understanding

This study uses the Predict Students' Dropout and Academic Success dataset from the UCI Machine Learning Repository [11]. The original dataset contains 4,424 student records and 37 variables spanning demographic, academic, admission-related, and macroeconomic attributes, along with a target label representing academic status. To formulate a clear supervised learning task with verified outcomes, records labeled "Enrolled" were excluded, producing 3,630 samples for binary classification. The retained dataset includes demographic variables (e.g., gender, age at enrollment, marital status, application mode), first- and second-semester academic variables (e.g., enrolled units, evaluations, approved units, grades), admission variables (e.g., admission grade, attendance regime), macroeconomic indicators (e.g., unemployment rate, inflation rate, GDP), and the target variable (Graduate/Dropout). The class distribution before and after filtering is reported in Table 2.

Table 2. Distribution of Target Variables Before and After Filtering

Status	Before Filtering	Percentage	After Filtering	Percentage
Graduate	2,209	49.93%	2,209	60.88%
Dropout	1,421	32.12%	1,421	39.12%
Enrolled	794	17.95%	-	-
Total	4,424	100%	3,630	100%

Exploratory Data Analysis (EDA) is conducted to understand variable distributions, detect anomalies, and identify early signals associated with dropout [12]. Descriptive statistics summarize central tendency and dispersion for numerical features and provide frequency/proportion summaries for categorical attributes. Additionally, correlation analysis (via a Pearson correlation matrix) is used to examine linear relationships among numerical variables and to flag potential multicollinearity concerns. A multicollinearity threshold is applied to identify highly correlated features—particularly among semester-based academic indicators such as approved units and grades—which are also among the most strongly associated with the target outcome [13]. These insights guide subsequent feature engineering and selection decisions in the Data Preparation phase. To improve interpretability and reduce dimensionality while preserving meaningful structure, several engineered features are constructed to summarize cumulative academic progress, early trajectory patterns, and financial condition. Table 3 lists the engineered features used, including their source variables and analytical intent.

Table 3. Summary of Engineered Features

Engineered Feature	Source Variables	Data Type	Analytical Purpose
total_units_approved	Curricular units 1st sem (approved), Curricular units 2nd sem (approved)	Numerical	Capture cumulative academic progress
avg_grade	Curricular units 1st sem (grade), Curricular units 2nd sem (grade)	Numerical	Represent overall academic achievement
success_rate_1st_sem	Curricular units 1st sem (approved), Curricular units 1st sem (enrolled)	Numerical	Measure early academic success
success_rate_2nd_sem	Curricular units 2nd sem (approved), Curricular units 2nd sem (enrolled)	Numerical	Capture continuation of academic trajectory
economic_health	Debtor, Tuition fees up to date, Scholarship holder	Numerical	Summarize students' socioeconomic condition

2.3. Data Preparation

Data preparation transforms raw data into a modeling-ready format through cleaning, encoding, transformation, and imbalance handling. Consistent with CRISP-DM, this phase includes data quality checks, selection of relevant attributes, and feature construction aligned with the research goals. Categorical variables are encoded to ensure compatibility with tree-based learners and the meta-learner, while numerical features may be normalized or standardized depending on modeling requirements and stability considerations. These preprocessing steps support robust learning and reduce the risk of distorted model behavior due to scale differences, noisy categories, or sparsity [19], [20].

Because dropout prediction datasets often exhibit class imbalance—where one class (typically graduates) is more frequent—this study applies SMOTE–Tomek Links as a hybrid sampling strategy to both oversample the minority class and remove ambiguous borderline instances. This approach helps improve minority-class sensitivity while simultaneously reducing overlap and noise between classes, which can otherwise inflate accuracy while masking poor dropout detection. The use of SMOTE–Tomek Links is aligned with best practices for imbalanced classification in educational data mining and improves the model's ability to discriminate dropout cases reliably [21].

Feature engineering (Table 2) is applied during this phase to reduce redundancy among strongly correlated semester-level variables and to capture more policy-relevant constructs such as early success rates and cumulative progress. These engineered indicators are designed to be more interpretable for academic stakeholders, enabling direct translation into intervention logic (e.g., first-year performance monitoring, credit accumulation thresholds, financial risk screening). The resulting dataset is then finalized for modeling.

2.4. Modeling

In the modeling phase, multiple algorithms are trained and combined using Stacking Ensemble Learning, which aims to improve predictive accuracy and robustness by integrating complementary strengths from different learners [22]. This study employs Random Forest, Gradient Boosting, and XGBoost as base models because they perform strongly on structured tabular data, capture nonlinear relationships, and handle complex interactions among academic, demographic, and socioeconomic variables. The outputs of these base learners are then combined by a Logistic Regression meta-learner, selected for its stability and interpretability when learning optimal weights from the base models' predictions.

To ensure a valid stacking procedure and reduce overfitting, the stacking pipeline uses cross-validation to generate out-of-fold predictions from the base learners for training the meta-learner. This approach prevents information leakage because the meta-learner is trained on predictions produced from validation folds rather than on in-sample predictions. In this study, the Stacking Classifier is configured with 5-fold cross-validation (cv = 5) and uses the auto stack method, enabling the framework to select an appropriate stacking strategy based on the prediction outputs. This design strengthens generalization and improves the reliability of the final model when applied to unseen student records.

Hyperparameters for each component model are specified to balance predictive performance and stability while maintaining reproducibility. All models use a consistent random_state = 42 to ensure that results can be replicated across runs. The Random Forest is configured with 100 trees (n_estimators = 100) and a maximum depth of 10 (max_depth = 10) to control model complexity while capturing meaningful feature interactions. Gradient Boosting uses 100 estimators, a learning rate of 0.1, and maximum depth of 5 to enable incremental error correction without excessive variance. XGBoost is similarly set to 100 estimators, a learning rate of 0.1, and maximum depth of 5, with log loss as the evaluation metric to optimize probabilistic classification quality. The Logistic Regression meta-learner is configured with maximum iterations of 1000 to ensure convergence and applies L2 regularization (default) to reduce overfitting and stabilize learned coefficients. The full hyperparameter configuration is summarized in Table 5.

Table 5. Hyperparameter Configuration of the Stacking Ensemble Model

Model	Hyperparameter	Value
Random Forest	Number of trees (n_estimators)	100
	Maximum depth (max_depth)	10
	Random state	42
Gradient Boosting	Number of estimators	100
	Learning rate	0.1
	Maximum depth	5
	Random state	42
XGBoost	Number of estimators	100
	Learning rate	0.1
	Maximum depth	5

Model	Hyperparameter	Value
	Evaluation metric	Log loss
	Random state	42
Logistic Regression (Meta-learner)	Maximum iterations	1000
	Regularization	L2 (default)
	Random state	42
Stacking Classifier	Cross-validation folds (cv)	5
	Stack method	Auto

2.5. Evaluation

Model evaluation assesses whether the developed solution meets both technical performance targets and institutional decision-support needs. Performance is evaluated using standard classification metrics, emphasizing not only overall accuracy but also discriminative power through ROC–AUC to reflect the model's ability to separate dropout and graduate outcomes across decision thresholds [23], [24]. The stacking ensemble's results are also compared to individual base learners to quantify performance gains attributable to ensembling and to verify that improvements are consistent rather than incidental. In addition to predictive metrics, evaluation also considers interpretability requirements. Feature influence is examined using feature importance analysis to identify the most impactful academic and socioeconomic predictors. This interpretability-oriented evaluation supports institutional actionability by revealing which student attributes contribute most strongly to risk classification, enabling interventions that are targeted, transparent, and aligned with academic advising practices.

3. RESULTS AND DISCUSSION

3.1. Dataset Characteristics

The final dataset comprises 3,630 student records representing demographic, academic, and economic attributes. Most students are unmarried and enrolled in daytime classes, with an average previous qualification grade of 132.92, indicating generally strong prior academic performance. Academic indicators show mean values of 273.75 and 256.58 for first- and second-semester grades, respectively, with an overall average grade of 265.17. Success rates average 0.70 in the first semester and 0.66 in the second semester, suggesting a moderate decline in academic performance over time. From a socioeconomic perspective, the economic_health variable has a mean value of −12.87, reflecting heterogeneous financial conditions among students. The target distribution indicates that 60.9% of students graduate, while 39.1% drop out, confirming the presence of class imbalance and justifying the application of resampling techniques during preprocessing. Overall, the diversity and variability of features support the use of an ensemble learning approach to capture complex, nonlinear relationships in dropout prediction. Figure 2 illustrates the class distribution, showing a higher proportion of graduate students compared to dropouts. This imbalance highlights the need for appropriate class-balancing strategies to prevent biased model learning.

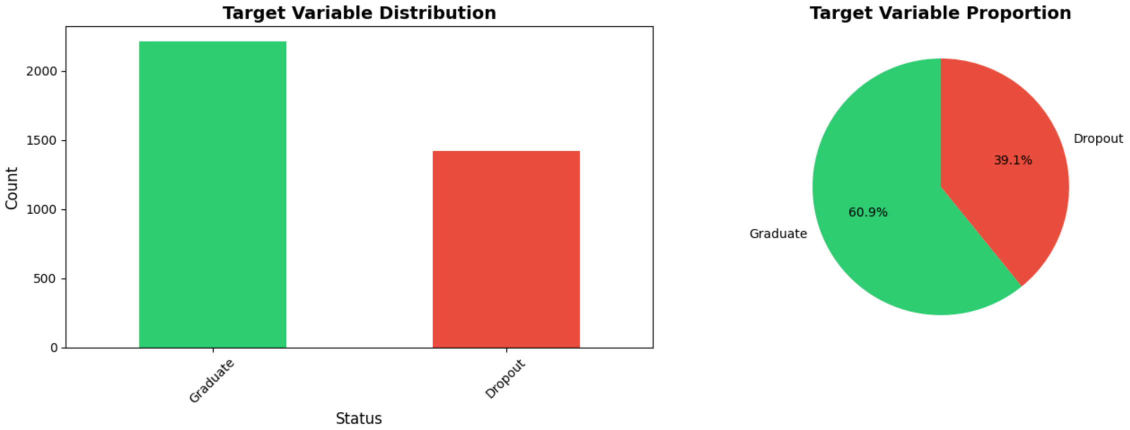

Figure 2. Proportion of Student Status

The visualization in Figure 2 presents the distribution and proportion of student status based on their final academic outcomes. The bar chart on the left shows that 2,209 students graduated, while 1,421 students dropped out, indicating a class imbalance between the two categories. The pie chart on the right further clarifies this proportion, with 60.9% of students graduating and 39.1% dropping out. This imbalance in distribution is an important consideration in the data preprocessing stage, particularly in handling class imbalance prior to the model training process.

Figure 3 presents the correlation matrix for the most influential features. Early academic indicators—success_rate_1st_sem, success_rate_2nd_sem, total_units_approved, and avg_grade—exhibit the strongest positive correlations with the target variable (r = 0.55–0.74). In contrast, demographic variables such as age at enrollment, gender, and debtor status display weak negative correlations, indicating limited predictive contribution relative to academic performance. The dataset exhibits clear class imbalance and strong variability across academic, demographic, and socioeconomic features. Early academic performance indicators dominate the correlation structure, while demographic variables contribute marginally. These characteristics justify both the application of class-balancing techniques and the use of an ensemble modeling strategy to capture complex relationships underlying student dropout risk.

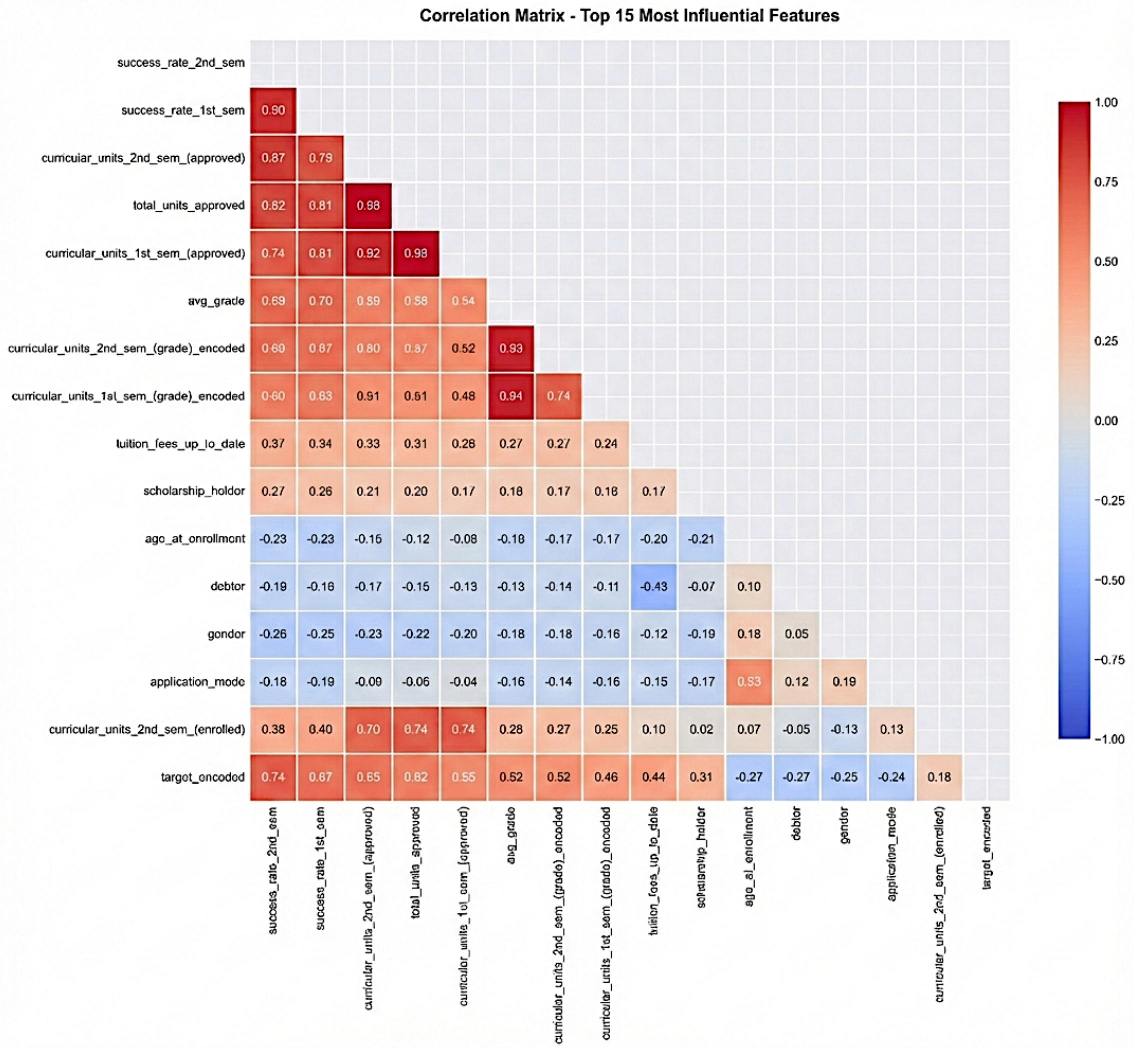

Figure 3. Correlation Matrix - Top 15 most influential features

3.2. Feature Engineering

Figure 4 presents the outcomes of the feature engineering stage, where five composite variables were created to summarize students' academic progression and socioeconomic conditions in a compact, interpretable form. This step reduces redundancy among semester-level indicators while keeping the signals that matter most for dropout prediction—particularly early academic momentum and financial vulnerability. The engineered features were designed to be both model-friendly and institution-friendly, meaning they can support accurate prediction while remaining easy to translate into early-warning rules and intervention triggers.

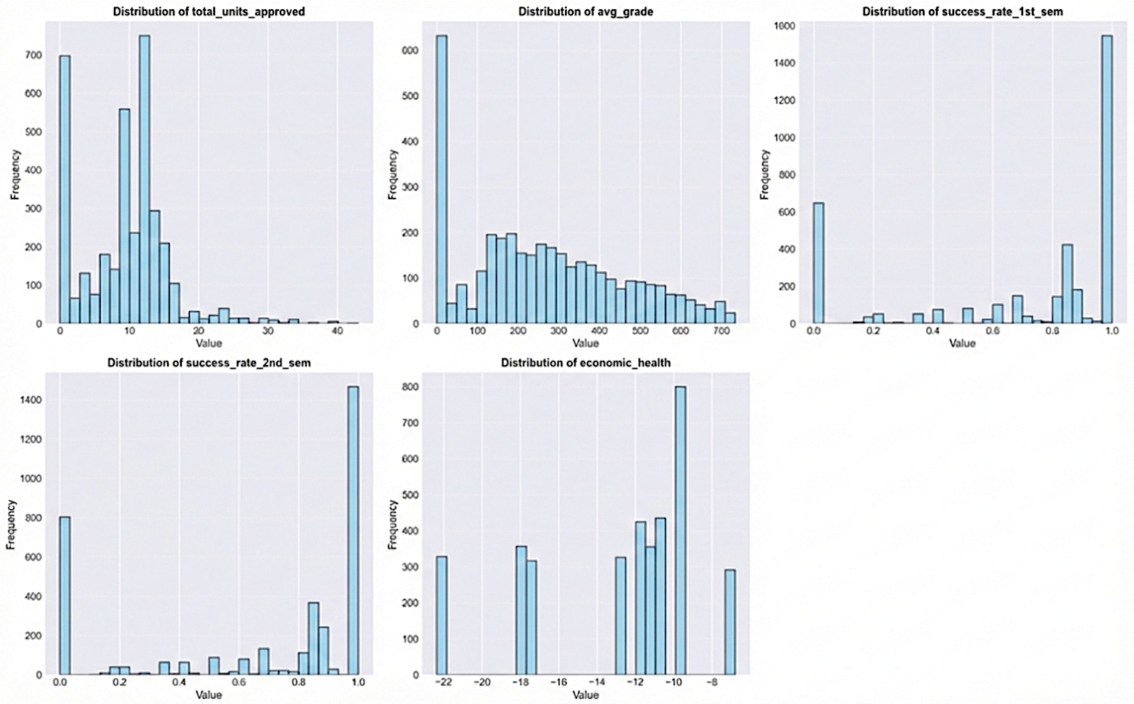

Figure 4. Feature Engineering

Two features capture cumulative academic performance. total_units_approved represents overall workload completion across the first two semesters, while avg_grade summarizes achievement across the same period. In Figure 4, both show distributions approaching normality, indicating stable central tendencies and making them reliable summary indicators of progress and performance. In contrast, success_rate_1st_sem and success_rate_2nd_sem—the proportions of approved units relative to enrolled units—show clear bimodal patterns clustered near 0 and 1. This polarization suggests strong discriminative value: students tend to fall into “consistently progressing” versus “struggling/unstable” trajectories, which is precisely the behavioral split an early warning system needs to detect.

The socioeconomic indicator economic_health, constructed from tuition payment status, scholarship status, and debtor status, displays a multimodal distribution in Figure 4, reflecting distinct financial profiles within the student population. Overall, these engineered variables condense high-dimensional academic and financial information into transparent indicators that highlight early trajectory disruption and financial strain as primary dropout signals, strengthening both predictive relevance and interpretability.

3.3. Handling Imbalanced Data

Figure 5 shows the effect of applying SMOTE–Tomek to address the original class imbalance in the dataset (60.8% Graduate vs. 39.2% Dropout). Because standard classifiers tend to favor the majority class, this imbalance can lead to a model that appears accurate overall but performs poorly in detecting dropout cases—the group that matters most for an early warning system. To reduce this bias, SMOTE–Tomek was used as a hybrid resampling strategy: SMOTE generates synthetic examples of the minority class (Dropout), while Tomek Links removes ambiguous samples near overlapping class boundaries, effectively cleaning borderline noise and sharpening separation between classes.

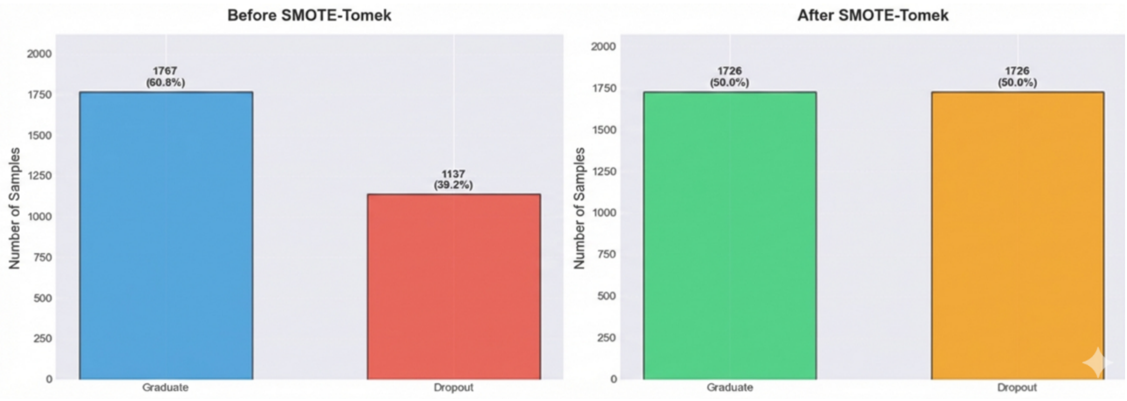

Figure 5. Handling Imbalanced Data using Smote-tomek

As illustrated in Figure 5, the resampling process produces a balanced class ratio of 50%:50%, ensuring that dropout patterns are represented as strongly as graduate patterns during training. This balanced representation improves the model's ability to learn minority-class characteristics and typically strengthens performance on dropout-focused metrics such as recall and F1-score, which are critical for institutional deployment. In practice, higher recall means fewer at-risk students are missed, while a stronger F1-score indicates a better balance between correctly identifying dropout cases and avoiding excessive false alarms.

3.4. Model Evaluation

The proposed Stacking Ensemble Learning model was evaluated to determine whether it can function as a reliable early warning mechanism for identifying students at risk of dropping out while keeping false alarms at a manageable level. Evaluation focuses on three complementary views of performance: (1) error structure through the confusion matrix (how the model succeeds or fails for each class), (2) class-wise effectiveness using precision, recall, and F1-score, and (3) threshold-independent discrimination using ROC–AUC. In addition, the Matthews Correlation Coefficient (MCC) is reported to provide a balanced single-number summary that accounts for all outcomes in the confusion matrix.

3.4.1. Confusion Matrix

The confusion matrix provides the most operationally meaningful view of model behavior because it shows how often the system correctly flags dropout cases and how often it either misses at-risk students or raises unnecessary alerts. The model achieves 90.91% accuracy, correctly classifying 660 out of 726 instances. Its errors are asymmetric in a way that is generally favorable for institutional deployment: false positives are low, meaning the system does not overwhelm staff with unnecessary interventions, while recall for dropout remains strong enough to support proactive outreach.

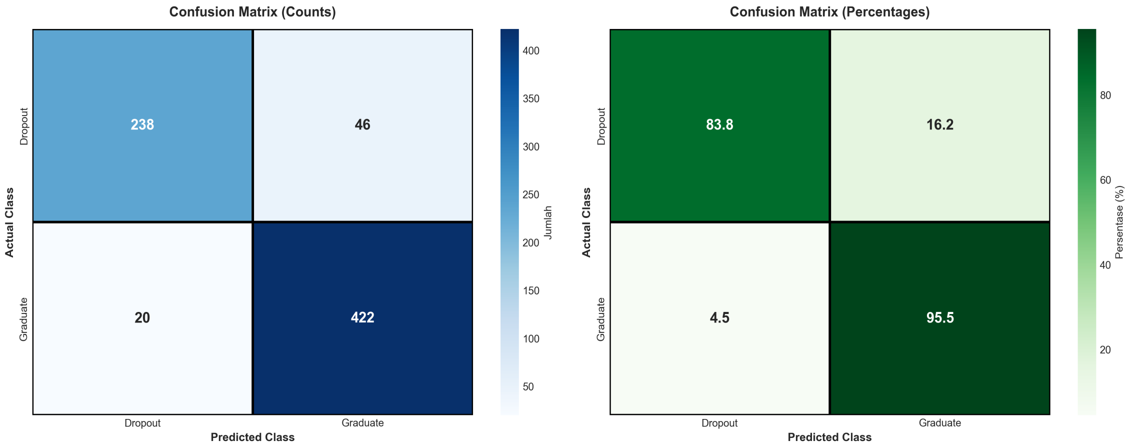

Figure 6. Confusion Matrix Analysis

Figure 6 shows that the model correctly identifies 238 dropout students (true positives) and misses 46 dropout students (false negatives). In percentage terms, 83.8% of dropout cases are detected, while 16.2% are not flagged. For graduates, the model correctly classifies 422 students (true negatives) and incorrectly flags 20 graduates as dropout (false positives), which corresponds to 95.5% correct graduate identification and only 4.5% false alarms. This profile indicates a practical balance: the model detects most at-risk students while keeping unnecessary intervention workload relatively low.

3.4.2. Classification Reports

While the confusion matrix highlights the pattern of errors, the classification report quantifies performance in a standardized way that is easier to compare across studies and models. Precision reflects how trustworthy the "at-risk" flag is, recall reflects how many at-risk students are actually captured, and F1-score balances both. In early warning contexts, recall for the Dropout class is particularly important because missed cases represent students who may not receive timely support.

Table 6 shows that the Dropout class achieves precision = 0.9225, recall = 0.8380, and F1-score = 0.8782. This means that when the model predicts dropout, it is correct most of the time, and it successfully captures the majority of actual dropout cases. The Graduate class achieves precision = 0.9017, recall = 0.9548, and F1-score = 0.9275, indicating strong stability in identifying students who complete their studies. Macro and weighted averages confirm that performance is balanced across classes rather than being driven by one class only.

Tabel 6. Classification Report

Class	Precision	Recall	F1-Score	Support
Dropout	0.9225	0.8380	0.8782	284
Graduate	0.9017	0.9548	0.9275	442
Accuracy	**0.9091**	**0.9091**	**0.9091**	**0.9091**
Macro Avg	0.9121	0.8964	0.9029	726
Weighted Avg	0.9098	0.9091	0.9082	726

3.4.3. ROC-AUC

Because institutions may adjust the decision threshold depending on available resources (e.g., how many students can be supported each semester), it is important to verify that the model discriminates well across thresholds—not only at one fixed cut-off. ROC–AUC evaluates this threshold-independent separation between Dropout and Graduate outcomes. Figure 7 shows an AUC of 0.9572, which indicates excellent discriminative capability. The curve remains near the upper-left area, reflecting a strong true positive rate with a low false positive rate across a wide range of thresholds. The selected operating threshold of 0.464 represents a balanced point: lowering the threshold would capture more at-risk students (higher recall) but increase false positives and staff workload, while raising it would reduce false positives but risk missing more dropout cases. The chosen threshold supports preventive intervention by maintaining strong detection while keeping false alarms manageable.

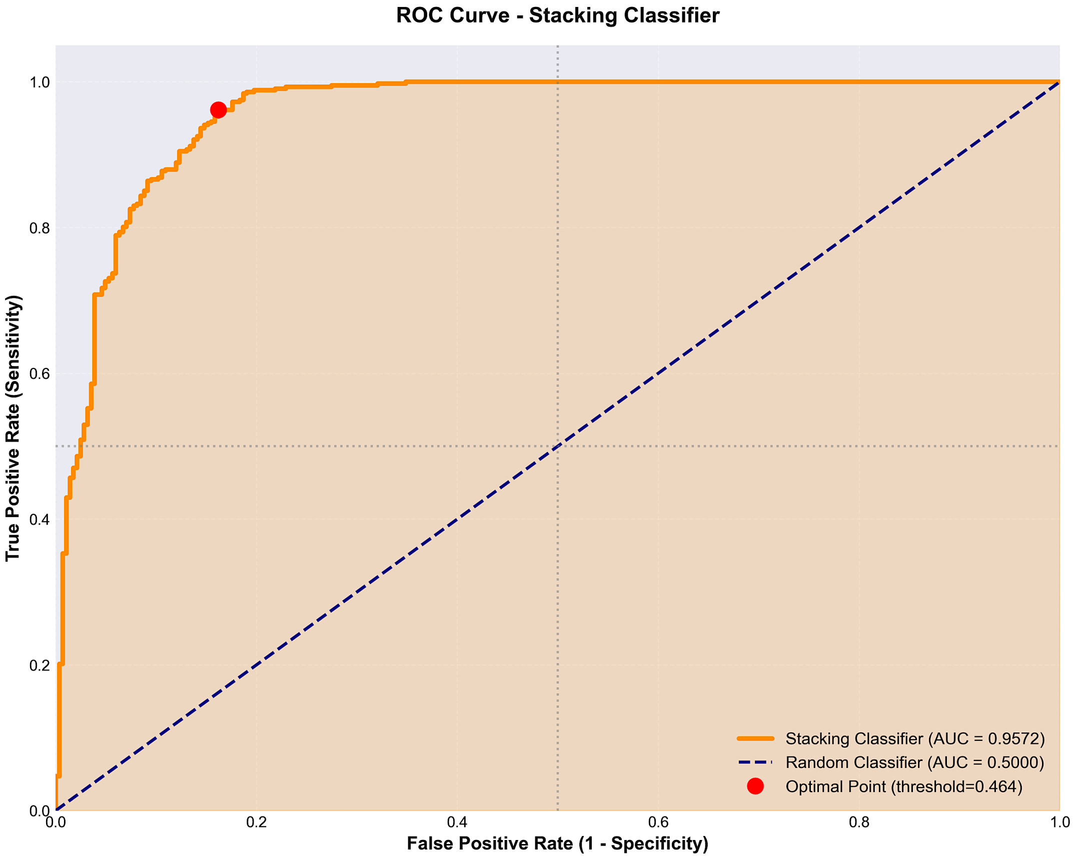

Figure 7. ROC-AUC analysis of Stacking Classifier

3.4.4. Matthews Correlation Coefficient (MCC)

o complements the above metrics, the MCC is reported as a balanced measure that considers true positives, true negatives, false positives, and false negatives simultaneously. The model achieves an MCC of 0.82, indicating strong agreement between predicted and actual labels. This reinforces that performance is not inflated by class distribution effects and remains stable across both outcome groups, strengthening the model's suitability for early warning deployment where misclassification costs are not equal.

3.5. Discussion

The findings of this study confirm that a stacking-based ensemble architecture can deliver strong predictive performance while still producing outputs that are meaningful for institutional decision-making. The proposed Stacking Ensemble Learning model achieves 90.91% accuracy, an ROC–AUC of 0.9572, and an MCC of 0.82, indicating both high discrimination and stable agreement between predicted and actual outcomes. Importantly, the error structure is operationally acceptable for early warning use: the model keeps false alarms relatively low (20 false positives) while still detecting the majority of at-risk students (238 true positives, with 46 false negatives). This balance matters in practice because universities must identify enough at-risk students to justify intervention programs without overwhelming academic support units with excessive alerts.

From an interpretability standpoint, the feature importance results reinforce a consistent message in recent dropout prediction research: early academic trajectory dominates risk explanation. Variables such as first- and second-semester success rates, total approved units, and average grades emerge as the most influential predictors, showing that dropout risk is strongly shaped by whether students build early academic momentum. This is aligned with [1], who emphasize that explainable AI is most valuable when it highlights factors that institutions can act on for personalized intervention. The dominance of cumulative and early-semester performance indicators also supports the observations reported by [2] and [3], where academic progress measures consistently outperform demographic and macroeconomic attributes in classifying at-risk students. In practical terms, this study strengthens the argument that early warning systems should prioritize first-year trajectory monitoring—particularly credit completion consistency and early course success—because these signals offer both predictive power and clear intervention pathways (e.g., tutoring, course redesign, academic advising, or structured study support).

Methodologically, the results provide evidence that stacking improves reliability beyond what single learners typically achieve by combining complementary strengths across different algorithms. This supports the argument by [20] that two-layer ensemble strategies can reduce generalization error by leveraging diverse inductive biases—here represented by Random Forest's robustness, Gradient Boosting's iterative correction, and XGBoost's strong optimization—then integrating them through a Logistic Regression meta-learner. The high ROC–AUC value further aligns with [25], who report that stacking frameworks often perform particularly well in complex classification tasks with heterogeneous feature types. In this study, the practical benefit of stacking is visible in the balanced precision–recall tradeoff: dropout predictions maintain high precision while preserving meaningful recall, which is essential when the cost of missing at-risk students (false negatives) can be higher than the cost of issuing limited false alerts.

The improvement in minority-class detection is also closely linked to the preprocessing strategy. The use of SMOTE–Tomek Links addresses not only the class ratio problem but also the quality of the decision boundary by removing borderline overlap. Similar to the trends reported by [18], balancing the dataset supports stronger dropout recall without driving false positives to an impractical level. This matters for deployment because institutional early warning systems must reliably detect dropout risk patterns that may otherwise be underrepresented during training, especially when the graduate class is dominant. By reducing boundary noise, the model becomes more stable when exposed to new cohorts, making predictions less sensitive to small shifts in student profiles.

This study contributes to the literature by demonstrating a practical bridge between performance and usability: stacking ensemble learning provides strong predictive capability, while feature engineering and importance analysis translate model behavior into actionable risk signals that align with how universities design interventions. At the same time, several considerations remain important for interpretation and future work. The binary formulation excludes the "Enrolled" group, which improves label clarity but may reduce exposure to transitional trajectories; future studies could explore time-to-event or multi-class approaches to capture uncertainty during ongoing study status. In addition, expanding validation across institutional contexts—particularly within Indonesian STEM programs—would further strengthen external validity and support implementation as a scalable, policy-relevant early warning tool.

4. CONCLUSION

This study shows that Stacking Ensemble Learning can predict STEM student dropout accurately and provide usable insights for early intervention. The proposed model achieves 90.91% accuracy and a ROC–AUC of 0.9572, indicating strong class separation. Combining Random Forest, Gradient Boosting, and XGBoost with a Logistic Regression meta-learner captures complementary patterns, while SMOTE–Tomek Links helps address class imbalance and improves detection of dropout cases. Feature importance results consistently point to early academic trajectory signals—especially first- and second-semester success rates, total approved units, and average grades—as the most influential predictors, making the model suitable as an interpretable early warning component in academic information systems.

Key limitations include reliance on a single-source dataset, exclusion of qualitative/behavioral factors (e.g., motivation, engagement), and limited individual-level transparency despite global feature importance; the binary setup also omits "Enrolled" cases, which may contain transitional patterns. Future work should validate the approach on multi-institutional/longitudinal data, integrate multimodal indicators (e.g., LMS activity), and apply SHAP or counterfactual explanations to strengthen case-level interpretability.

REFERENCES

[1] M. Nagy and R. Molontay, "Interpretable Dropout Prediction: Towards XAI-Based Personalized Intervention," *Int. J. Artif. Intell. Educ.*, vol. 34, no. 2, pp. 274–300, Jun. 2024, doi: 10.1007/s40593-023-00331-8.

[2] S. Kim, E. Choi, Y. K. Jun, and S. Lee, "Student Dropout Prediction for University with High Precision and Recall," *Appl. Sci.*, vol. 13, no. 10, Art. no. 6275, May 2023, doi: 10.3390/app13106275.

[3] C. H. Cho, Y. W. Yu, and H. G. Kim, "A Study on Dropout Prediction for University Students Using Machine Learning," *Appl. Sci.*, vol. 13, no. 21, Art. no. 12004, Nov. 2023, doi: 10.3390/app132112004.

[4] T. Yoon and D. Kang, "Multi-Modal Stacking Ensemble for the Diagnosis of Cardiovascular Diseases," *J. Pers. Med.*, vol. 13, no. 2, Art. no. 373, Feb. 2023, doi: 10.3390/jpm13020373.

[5] M. Nascimento, A. C. C. Nascimento, C. F. Azevedo, A. C. B. de Oliveira, E. T. Caixeta, and D. Jarquin, "Enhancing genomic prediction with stacking ensemble learning in Arabica coffee," *Front. Plant Sci.*, vol. 15, Art. no. 1373318, 2024, doi: 10.3389/fpls.2024.1373318.

[6] J. Zheng, M. Wang, T. Yao, Y. Tang, and H. Liu, "Dynamic mechanical strength prediction of BFRC based on stacking ensemble learning and genetic algorithm optimization," *Buildings*, vol. 13, no. 5, Art. no. 1155, May 2023, doi: 10.3390/buildings13051155.

[7] N. Doede, P. Merkel, M. Kriwall, M. Stonis, and B. A. Behrens, "Implementation of an intelligent process monitoring system for screw presses using the CRISP-DM standard," *Prod. Eng.*, 2024, doi: 10.1007/s11740-024-01298-8.

[8] A. M. Shimaoka, R. C. Ferreira, and A. Goldman, "The evolution of CRISP-DM for data science: Methods, processes and frameworks," *SBC Rev. Comput. Sci.*, vol. 4, no. 1, pp. 28–43, Oct. 2024, doi: 10.5753/reviews.2024.3757.

[9] C. Schröer, F. Kruse, and J. M. Gómez, "A systematic literature review on applying CRISP-DM process model," *Procedia Comput. Sci.*, vol. 181, pp. 526–534, 2021, doi: 10.1016/j.procs.2021.01.199.

[10] E. Hakim and A. Muklason, "Analysis of employee work stress using CRISP-DM to reduce work stress on reasons for employee resignation," *Data Sci. J. Comput. Appl. Inform.*, vol. 8, no. 2, pp. 75–87, 2024, doi: 10.32734/jocai.v8.i2.

[11] V. Realinho, J. Machado, L. Baptista, and M. V. Martins, "Predicting student dropout and academic success," *Data*, vol. 7, no. 11, Art. no. 146, 2022, doi: 10.3390/data7110146.

[12] A. Y. Wang, W. Epperson, R. A. Deline, and S. M. Drucker, "Diff in the Loop: Supporting Data Comparison in Exploratory Data Analysis," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, New Orleans, LA, USA, Apr. 2022, doi: 10.1145/3491102.3502123.

[13] M. B. Courtney, "Exploratory data analysis in schools: A logic model to guide implementation," *Int. J. Educ. Policy Leadersh.*, vol. 17, no. 4, May 2021, doi: 10.22230/ijepl.2021v17n4a1041.

[14] S. Marlia *et al.*, "Analysis of music features and song popularity trends on Spotify using K-Means and CRISP-DM," *Sistemasi*, 2024.

[15] M. Mujahid *et al.*, "Data oversampling and imbalanced datasets: An investigation of performance for machine learning and feature engineering," *J. Big Data*, vol. 11, no. 1, 2024, doi: 10.1186/s40537-024-00943-4.

[16] R. Joeres, D. B. Blumenthal, and O. V. Kalinina, “DataSAIL: Data splitting against information leakage,” *bioRxiv*, Nov. 17, 2023, doi: 10.1101/2023.11.15.566305.

[17] Q. H. Nguyen *et al.*, “Influence of data splitting on performance of machine learning models in prediction of shear strength of soil,” *Math. Probl. Eng.*, vol. 2021, Art. no. 4832864, 2021, doi: 10.1155/2021/4832864.

[18] Y. Zhang, L. Deng, and B. Wei, “Imbalanced data classification based on improved Random-SMOTE and feature standard deviation,” *Mathematics*, vol. 12, no. 11, Art. no. 1709, Jun. 2024, doi: 10.3390/math12111709.

[19] H. Hairani, A. Anggrawan, and D. Priyanto, “Improvement performance of the random forest method on unbalanced diabetes data classification using SMOTE-Tomek link,” *Int. J. Inform. Vis.*, 2023.

[20] J. Niyogisubizo, L. Liao, E. Nziyumva, E. Murwanashyaka, and P. C. Nshimyumukiza, “Predicting student’s dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization,” *Comput. Educ. Artif. Intell.*, vol. 3, Art. no. 100066, 2022, doi: 10.1016/j.caeai.2022.100066.

[21] M. Nascimento *et al.*, “Enhancing genomic prediction with stacking ensemble learning in Arabica coffee,” *Front. Plant Sci.*, vol. 15, Art. no. 1373318, 2024, doi: 10.3389/fpls.2024.1373318.

[22] S. Sathyanarayanan, “Confusion matrix-based performance evaluation metrics,” *Afr. J. Biomed. Res.*, vol. 27, no. 4S, pp. 4023–4031, Nov. 2024, doi: 10.53555/ajbr.v27i4s.4345.

[23] E. K. Anku and H. O. Duah, “Predicting and identifying factors associated with undernutrition among children under five years in Ghana using machine learning algorithms,” *PLoS One*, vol. 19, no. 2, Feb. 2024, doi: 10.1371/journal.pone.0296625.

[24] D. Chicco and G. Jurman, “The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification,” *BioData Min.*, vol. 16, no. 1, 2023, doi: 10.1186/s13040-023-00322-4.

[25] A. Gupta, V. Jain, and A. Singh, “Stacking ensemble-based intelligent machine learning model for predicting post-COVID-19 complications,” *New Gener. Comput.*, vol. 40, no. 4, pp. 987–1007, Dec. 2022, doi: 10.1007/s00354-021-00144-0.