# A Stacking Ensemble Model for Predicting Student High School Graduation Outcomes

Fitriyani[1], Ari Amir Alkodri[2,*], Fajar Aswin[3]

[1,2]*ISB Atma Luhur, Pangkalpinang, Kep.Bangka Belitung 33117, Indonesia*

[3]*Politeknik Manufaktur Negeri Bangka Belitung, Sungailiat Bangka, Kep. Bangka Belitung 33211, Indonesia*

(Received: June 8, 2025; Revised: August 10, 2025; Accepted: November 19, 2025; Available online: December 9, 2025)

**Abstract**

This study develops and evaluates machine learning models to predict high school graduation outcomes and identify at-risk students for early intervention. Using a quantitative approach, data from 1,017 students across three public high schools were analyzed, encompassing academic performance (average yearly scores), behavioral factors (attendance rates and extracurricular participation), and socio-economic background (proxied by parental occupation). A comparative modeling strategy was applied, beginning with a Decision Tree baseline and advancing to a Stacking Ensemble model that integrated three heterogeneous base learners—Support Vector Machine (SVM), k-Nearest Neighbors (k-NN), and Decision Tree—combined through a Logistic Regression meta-model. Both models were optimized using GridSearchCV and adjusted for class imbalance between graduates (93.4%) and at-risk students (6.6%). The results showed that academic variables, particularly third-year average scores (mean = 82.6, SD = 6.4) and attendance rate (mean = 94.3%), were the strongest predictors of graduation, while socio-economic indicators had minimal impact. The Stacking Ensemble achieved a notable improvement over the Decision Tree, reaching an accuracy of 99.6%, precision of 0.909, recall of 1.000, F1-score of 0.952, and AUC of 1.000, compared to the baseline accuracy of 94.9% (F1-score = 0.519, AUC = 0.83). These findings indicate the superior predictive capability of the ensemble model in identifying students at risk of non-graduation. The study's novelty lies in combining interpretable and high-performance models to construct a practical early-warning framework that can guide educators and policymakers in targeted academic interventions. However, the near-perfect metrics also suggest potential overfitting, emphasizing the need for validation using external datasets before broader application. Overall, this research contributes a robust, data-driven methodology for improving student retention through predictive analytics in educational settings.

*Keywords:* Student Performance Prediction, Educational Data Mining, Early-Warning System, Stacking Ensemble, Decision Tree, Logistic Regression.

## 1. Introduction

Student graduation rates are widely recognized as critical indicators of educational system effectiveness and institutional quality [1] in developing countries such as Indonesia, ensuring that students graduate on time remains a persistent challenge, particularly at the high school level where academic, behavioral, and socio-economic pressures intersect [2]. Pangkalpinang City, the capital of Bangka Belitung Islands Province, has implemented various initiatives to improve educational outcomes. However, the absence of predictive systems for identifying at-risk students continues to limit the effectiveness of proactive interventions [3].

The emergence of Educational Data Mining (EDM) and machine learning has created new opportunities to support early-warning systems in schools. A recent 10-year review of EDM highlights that these techniques allow institutions to process large amounts of student data, identify meaningful patterns, and predict outcomes with greater accuracy than traditional methods, thereby helping to increase student performance and graduation rates [4]. By leveraging predictive models, schools can move from reactive strategies to proactive ones, enabling early identification of students at risk of delayed or incomplete graduation [5].

Numerous studies have applied machine learning algorithms to predict student academic outcomes. For instance, a 2024 systematic review confirms that machine learning models, including decision trees, consistently outperform

***Corresponding author: Ari Amir Alkodri (arie_a3@atmaluhur.ac.id)**
**DOI: https://doi.org/10.47738/jads.v5i2.XXX**

traditional statistical models in forecasting student performance and identifying at-risk students due to their capacity to handle large, non-linear datasets [5]. Furthermore, more recent works have emphasized ensemble and hybrid approaches. One study demonstrated the effectiveness of the Random Forest algorithm, an ensemble method, in developing an early-warning system [6], while another confirmed that combining various prediction algorithms improves the early identification of unsuccessful students [7]. In the Indonesian context, studies by Kurniawan et al. [8] and Aprilia et al. [9] confirmed the feasibility of applying such models to secondary education, though much of the research remains concentrated in large metropolitan or university-level settings.

Despite these advances, research focusing specifically on high school students in smaller Indonesian cities such as Pangkalpinang remains scarce. Most existing studies emphasize higher education or large-scale datasets [10], [11]. This leaves a gap in understanding how predictive models perform when applied to localized, smaller-scale datasets in the Indonesian high school system.

This study focuses on implementing machine learning techniques to predict high school graduation outcomes in Pangkalpinang City, Indonesia. The research employs Decision Tree and Stacking Ensemble methods to analyze multi-year student data encompassing academic performance, attendance, extracurricular participation, and socio-economic background. By comparing these models, the study seeks to evaluate the potential of hybrid ensemble learning to enhance prediction accuracy and support data-informed interventions in schools.

To achieve these objectives, the study investigates which academic, attendance, and demographic factors most strongly influence graduation outcomes in public high schools and examines whether machine learning models, particularly the Decision Tree and Stacking Ensemble, can accurately predict students' graduation status based on those factors. The study also compares the performance of ensemble-based approaches against single classifiers to determine which offers superior predictive accuracy and generalizability for this context. We hypothesize that (1) academic performance and attendance records will be the strongest predictors of graduation, outweighing demographic factors, and (2) the Stacking Ensemble model will achieve higher accuracy, recall, and overall reliability than the single Decision Tree classifier.

## 2. Literature Review

### 2.1. Quantitative Approaches in Educational Research

Machine learning has increasingly been applied to education as a quantitative approach for predicting student outcomes. Oppong [12] reviewed machine learning applications in educational data mining and reported that 87% of studies relied on supervised learning, confirming the central role of classification algorithms. Ahmed et al. [13] further demonstrated that methods such as Logistic Regression, Decision Trees, Random Forest, and XGBoost can achieve high accuracy in predicting academic success, while also identifying key social and demographic factors. Similarly, Tamir et al. [14] compared decision trees, logistic regression, and neural networks, showing that decision trees outperformed other models in accuracy and interpretability, underscoring their value as baseline classifiers.

In addition to algorithm selection, the predictive power of machine learning models depends on the inclusion of appropriate features. Pandey [15] confirmed that academic performance indicators, such as grades, are strong predictors of student outcomes when applied in decision tree models. Swiderski [16] demonstrated that behavioral indicators like attendance rates significantly influence high school graduation, highlighting their role as early-warning signals. Muhammad et al. [11] expanded this scope by incorporating extracurricular and employability-related features, showing that multi-dimensional datasets can enhance predictive power and inform student support strategies. Together, these studies emphasize that both methodological choices and feature selection are critical for building robust predictive frameworks in education.

### 2.2. Data Preprocessing, Feature Engineering, and Class Imbalance

Preprocessing ensures the integrity of educational datasets before modeling. Khairy et al. [17] emphasized that normalization, encoding, and handling missing values are essential steps for improving model reliability. Bum et al. [10] showed that careful data preparation enhanced the accuracy of regression-based predictions of student academic

performance. Tamir et al. [14] also highlighted the impact of structured preprocessing pipelines in improving model performance, particularly in large-scale predictive frameworks.

A related challenge in educational prediction is class imbalance, where successful students typically outnumber those at risk. Khairy et al. [17] recommended balancing strategies such as re-weighting and resampling to improve fairness in predictions. Oppong [12] reported that many studies adopted feature selection and balancing techniques to avoid bias toward the majority class. These findings confirm that preprocessing and imbalance handling are not only technical necessities but also crucial for ensuring equitable and accurate identification of at-risk students.

### 2.3. Model Development in Educational Prediction

Decision trees remain a popular choice for their interpretability, as shown in studies by Pandey [15] and Swiderski [16]. However, logistic regression continues to provide stable and probabilistic insights into student outcomes, making it useful for practical applications [11]. Almalawi et al. [5] compared multiple models and concluded that ensemble methods, particularly when logistic regression is used as a meta-model, outperform individual classifiers, offering both accuracy and generalizability.

### 2.4. Evaluation Metrics

The evaluation of predictive models in education requires metrics beyond accuracy. A systematic review of predictive models in education confirms that machine learning algorithms are superior for handling complex educational data, making the use of metrics like precision and recall critical for a meaningful assessment of model performance [5]. Tamir et al. [14] supported this by applying confusion matrix metrics such as precision, recall, and F1-score, which better capture the balance between identifying at-risk students and minimizing false alarms. Oppong [12] added that AUC and cross-validation are increasingly employed to validate models and ensure their generalizability across different datasets.

## 3. Methodology

### 3.1. Research Design

This study employs a quantitative research approach using supervised machine learning techniques to develop predictive models for high school graduation rates in Pangkalpinang City. The research follows a comparative methodology, implementing and evaluating two distinct algorithms: Logistic Regression and Decision Tree, to determine the most effective approach for the specific context of local high school education. Figure 1 illustrates the overall development of the prediction model, starting from data collection and preprocessing to feature selection, model training, and evaluation using GridSearchCV cross-validation. This visualization clarifies how each stage contributes to building a robust predictive framework.

### 3.2. Data Collection

Data collection was carried out through collaboration with three public high schools in Pangkalpinang City, chosen to represent geographical spread, diversity of student populations, and institutional readiness to participate in the study. The dataset consists of 1017 student records covering three academic years (2022–2024), providing a sufficient sample size for reliable model training and evaluation.

The collected dataset comprises four main categories of variables. Academic variables capture students' average academic scores for each school year, reflecting overall performance in the first, second, and third years. Behavioral variables include student attendance, measured as the number of school days attended annually, and participation in extracurricular activities, expressed as the number of activities joined per year. Socio-economic variables consist of parents' occupations, which serve as proxies for family income levels and educational background. Finally, the outcome variable records graduation status, where "LULUS" indicates Pass, while other values represent as Fail.

In this study, parental occupation was utilized as a proxy for socioeconomic status (SES) due to the limited availability of direct income or educational background data within the school's administrative system. The choice of this proxy reflects both practical and contextual considerations. In the Indonesian education setting, parental occupation often serves as a general indicator of household income potential and social position, which may indirectly influence

 ISSN 2723-6471


students' access to learning support, technology, and educational resources. Although this approach does not capture the full complexity of socioeconomic factors, it provides a feasible and contextually relevant approximation given the constraints of available data.

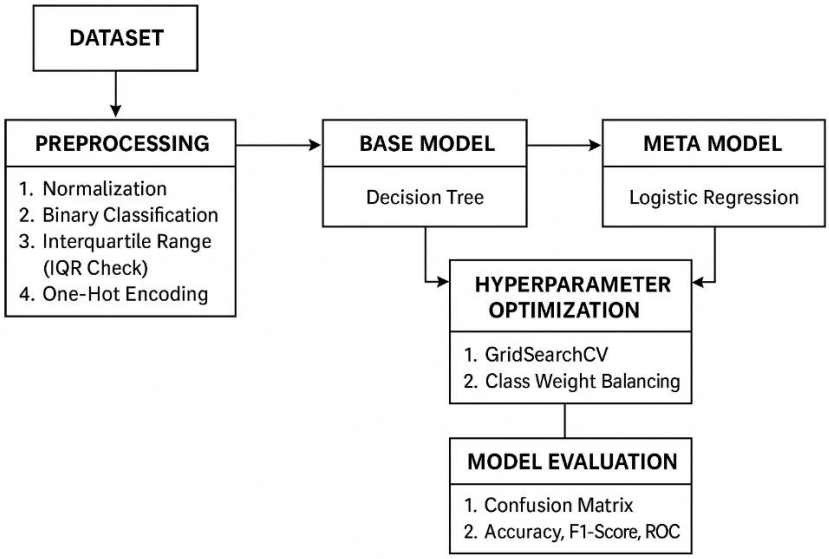


**Figure 1**. Predictive modeling workflow, from data preprocessing to model evaluation

## 3.3. Data Preprocessing

Prior to modeling, a structured preprocessing pipeline was applied to prepare the dataset for analysis. Variable names were first harmonized into English for consistency, and categorical values such as gender and parental occupation were standardized through string normalization and label harmonization. Ambiguous or missing categorical entries (e.g., unknown or unemployed parental occupations) were inspected manually and treated systematically. Specifically, missing categorical values were imputed using the most frequent category within each school, while anomalous or ambiguous labels were consolidated under a unified "Other/Unknown" category. This approach ensured that categorical features remained valid for encoding without introducing artificial bias or loss of representational integrity. The target variable, graduation status, was transformed into a binary classification label (1 = PASS, 0 = FAIL). Data validation was conducted using missing value inspection and outlier detection (via descriptive statistics and interquartile range (IQR) checks) to ensure data integrity.

Feature engineering was limited to the preparation of existing variables for modeling. Categorical variables such as gender, father's occupation, and mother's occupation were prepared using One-Hot Encoding (OHE) for low-cardinality variables. Numerical variables (scores, attendance, extracurricular counts) were standardized through z-score normalization (StandardScaler) to ensure comparability of features with different value ranges. No additional derived features, such as aggregate statistics or trend indicators, were created beyond the original variables collected in the dataset.

Finally, the dataset was partitioned into training and testing subsets using an 75/25 stratified train–test split, ensuring proportional representation of the graduation outcome variable. To avoid data leakage and guarantee reproducibility, all transformations were implemented within a scikit-learn Pipeline and ColumnTransformer, enabling consistent application of preprocessing during cross-validation and model evaluation.

## 3.4. Model Development

To construct the predictive framework, a two-stage modeling strategy was employed. A Decision Tree classifier was first implemented as a baseline model, given its interpretability and ability to capture non-linear relationships between features. Equations (1) and (2) show that the splitting criterion of the Decision Tree is based on Information Gain (IG) derived from Entropy (H), defined as follows:

 ISSN 2723-6471


$$H(S) = -\sum_{i=1}^{C} p_i \cdot \log_2(p_i) \quad (1)$$

$$IG(S,A) = H(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} H(S_v) \quad (2)$$

$p_i$ represents the probability of class $i$, $S$ is the dataset, and $S_v$ is the subset of $S_v$ after splitting by attribute $A$. A minimum information gain threshold was applied to avoid overfitting by eliminating splits that did not significantly reduce impurity.

Subsequently, a stacking ensemble model was developed to enhance predictive performance. This ensemble integrates predictions from multiple base learners—Support Vector Machine (SVM), k-Nearest Neighbors (k-NN), and Decision Tree—which serve as inputs to a Logistic Regression meta-model. The meta-model learns an optimal combination of base predictions to produce the final output (3), computed as:

$$\hat{y} = \sigma(\omega_0 + \omega_1 h_1 + \omega_2 h_2 + w_3 h_3) \quad (3)$$

Where $y$ is the final output, $h$ are the outputs of base learners, $\omega$ denotes the learned weights, and $\sigma(z)$denotes the logistic sigmoid activation function (4):

$$\sigma(z) = \frac{1}{1+e^{-z}} \quad (4)$$

To optimize the performance of each model, GridSearchCV was applied for hyperparameter tuning across multiple parameter combinations with stratified k-fold cross-validation, ensuring generalizability and robustness. The algorithmic procedures for model construction and optimization are summarized in figure 2. An analysis of the target variable revealed significant class imbalance, with 950 students (93.4%) in the 'PASS' category compared to only 67 students (6.6%) in the 'FAIL' category, yielding a ratio of approximately 14.2:1. a class-weight balancing (5) strategy was implemented to penalize misclassifications of the minority class more heavily. The adjusted weight $w_j$ for each class $j$ was computed as:

$$w_j = \frac{N}{n_j \times K} \quad (5)$$

$N$ is the total number of samples, $n_j$ is the number of samples in class $j$, and $K$ is the total number of classes. This weighting mechanism enhances model sensitivity to at-risk students while maintaining robust overall performance, thereby enabling more balanced and equitable predictions of graduation outcomes.

**Algorithm 1** Decision Tree Construction with Information Gain Threshold

**Require:** Training dataset $D$, minimum information gain threshold $\tau_{min}$
**Ensure:** Decision tree model
Initialize root node with complete dataset $D$
BuildTree$node, data$
**if** stopping criterion met (min samples leaf, max depth, etc.) **then**
  $node.label \leftarrow$ majority class in $data$
  **return**
**end if**
$best\_gain \leftarrow -\infty$
$best\_split \leftarrow$ None
**for** each feature $f$ in $data.features$ **do**
  **for** each possible split value $v$ for $f$ **do**
    $left\_data, right\_data \leftarrow$ split$(data, f, v)$
    $current\_gain \leftarrow IG(data, left\_data, right\_data)$
    **if** $current\_gain > best\_gain$ **then**
      $best\_gain \leftarrow current\_gain$
      $best\_split \leftarrow (f, v)$
    **end if**
  **end for**
**end for**
**if** $best\_gain < \tau_{min}$ **then**
  $node.label \leftarrow$ majority class in $data$
  **return**
**end if**
$node.split\_feature \leftarrow best\_split.feature$
$node.split\_value \leftarrow best\_split.value$
$left\_data, right\_data \leftarrow$ split$(data, node.split_feature, node.split_value)$
$node.left \leftarrow$ BuildTree$(new_node(), left_data)$
$node.right \leftarrow$ BuildTree$(new_node(), right_data)$

**Algorithm 2** Stacking Ensemble Training and Prediction

**Require:** Training data $D_{train}$, base models $M_1, M_2, ..., M_k$, meta-model $M_{meta}$
**Ensure:** Trained stacking ensemble
**Step 1: Training Phase**
**for** each base model $M_i$ in $\{M_1, M_2, ..., M_k\}$ **do**
  $M_i \leftarrow$ train$(M_i, D_{train})$
**end for**
$D_{meta} \leftarrow \emptyset$
**for** each instance $(x_j, y_j)$ in $D_{train}$ **do**
  $predictions \leftarrow [M_1.predict(x_j), M_2.predict(x_j),$
  ..., $M_k.predict(x_j)]$
  $D_{meta} \leftarrow D_{meta} \cup \{(predictions, y_j)\}$
**end for**
$M_{meta} \leftarrow$ train$(M_{meta}, D_{meta})$
**Step 2: Prediction Phase** Predict$x$
$base\_preds \leftarrow [M_1.predict(x), M_2.predict(x),$
..., $M_k.predict(x)]$
$final\_prediction \leftarrow M_{meta}.predict(base\_preds)$
**return** $final\_prediction$

**Algorithm 3** GridSearchCV with Cross-Validation

**Require:** Model $M$, parameter grid $G$, dataset $D$, number of folds $k$
**Ensure:** Optimal parameters $P^*$
$best\_score \leftarrow -\infty$
$best\_params \leftarrow$ None
**for** each parameter combination $P$ in grid $G$ **do**
  $scores \leftarrow \emptyset$
  **for** $fold = 1$ to $k$ **do**
    $(D_{train}, D_{val}) \leftarrow$ k_fold_split$(D, fold)$
    $M_{temp} \leftarrow M$.set_parameters$(P)$
    $M_{temp}$.fit$(D_{train})$
    $score \leftarrow M_{temp}$.evaluate$(D_{val})$
    $scores \leftarrow scores \cup \{score\}$
  **end for**
  $avg\_score \leftarrow$ mean$(scores)$
  **if** $avg\_score > best\_score$ **then**
    $best\_score \leftarrow avg\_score$
    $best\_params \leftarrow P$
  **end if**
**end for**
$M_{final} \leftarrow M$.set_parameters$(best_params)$
$M_{final}$.fit$(D)$
**return** $M_{final}, best\_params$

**Figure 2.** Algorithms for model construction and evaluation: Decision Tree building, Stacking Ensemble training, and GridSearchCV optimization

### 3.5. Model Evaluation

The evaluation of both the Decision Tree baseline and the Logistic Regression meta-model was conducted using a confusion matrix and several derived performance metrics. The confusion matrix organizes model predictions into four categories: True Positives (TP), students correctly identified as failing to graduate on time; False Positives (FP), students incorrectly predicted as failing; True Negatives (TN), students correctly identified as graduating on time; and False Negatives (FN), students incorrectly predicted as graduating.

From the confusion matrix, several performance metrics were derived. Precision was calculated to measure the proportion of correctly identified at-risk students among those predicted as failing, while recall (sensitivity) quantified the model's ability to capture all actual failing students. The F1-score, as the harmonic-mean of precision and recall, was used to provide a balanced measure particularly suited to imbalanced datasets. In addition, the Area Under the ROC Curve (AUC) was employed to evaluate the model's overall discriminative ability across different threshold values.

To ensure robust and unbiased evaluation, all metrics were computed under a cross-validation procedure implemented within the GridSearchCV framework, where stratified folds were applied to preserve the distribution of the target classes during training and validation.

## 4. Result and Discussion

The results presented below are based on the analysis of the collected dataset, which profiled 1,017 students across academic, behavioral, and demographic dimensions. This section first examines the underlying relationships between these variables before evaluating the predictive models trained on them.

### 4.1. Correlation Analysis

Figure 3 illustrates the correlation heatmap representing relationships among students' academic performance, attendance, and socio-economic variables. Darker shades indicate stronger positive correlations, revealing that early-year academic performance has a significant relationship with subsequent achievements. In our dataset, we observed a strong correlation between the first-year and third-year scores ($r = 0.72$), indicating that students who performed well academically early on tended to sustain high achievement in later years. Similarly, the first-year attendance percentage (1st_Attend) was positively correlated with attendance in subsequent years, with a correlation of $r = 0.68$ between first and third-year attendance.

The strength of these correlations in our Pangkalpinang dataset not only confirms the persistence of academic and behavioral patterns but also provides a local quantitative benchmark. This finding aligns with recent international evidence. For example, [18] analyzed over 4,700 Chinese university students and found that synchronous class attendance was significantly correlated with course outcomes, reporting correlation coefficients in the range of $r = 0.32–0.45$. Similarly, [19] reported that early continuous assessment scores showed correlations as high as $r = 0.61$ with final exam performance. The stronger correlations observed in our study may be attributed to the structured, traditional classroom environment of Pangkalpinang's public high schools, where attendance and performance are more directly and consistently linked.

Attendance has also been emphasized as a persistent driver of academic outcomes across multiple post-pandemic cohorts. A recent U.S. study on chronic absenteeism found that consistent attendance strongly predicted achievement levels over a three-year period, highlighting long-term consequences of early absenteeism [16], [20]. These findings align with the present study, where first-year attendance strongly correlated with later academic outcomes, reinforcing the idea that interventions to improve attendance should be prioritized early in high school.

At the same time, recent research warns that the strength of these correlations can vary by learning context. [21] showed that in fully online environments, simple engagement indicators such as login frequency or page views often correlated weakly with final performance ($r < 0.20$). This suggests that while attendance and performance are robust predictors in traditional classroom settings such as Pangkalpinang's public high schools, predictive strength may be weaker in digital-only contexts.

 ISSN 2723-6471


Taken together, these findings show that the correlations identified in this study are consistent with recent global evidence: early academic scores and attendance are strong predictors of long-term performance, especially in conventional classroom environments. These correlations justify their inclusion as central features in predictive modeling aimed at identifying at-risk students in high school education.

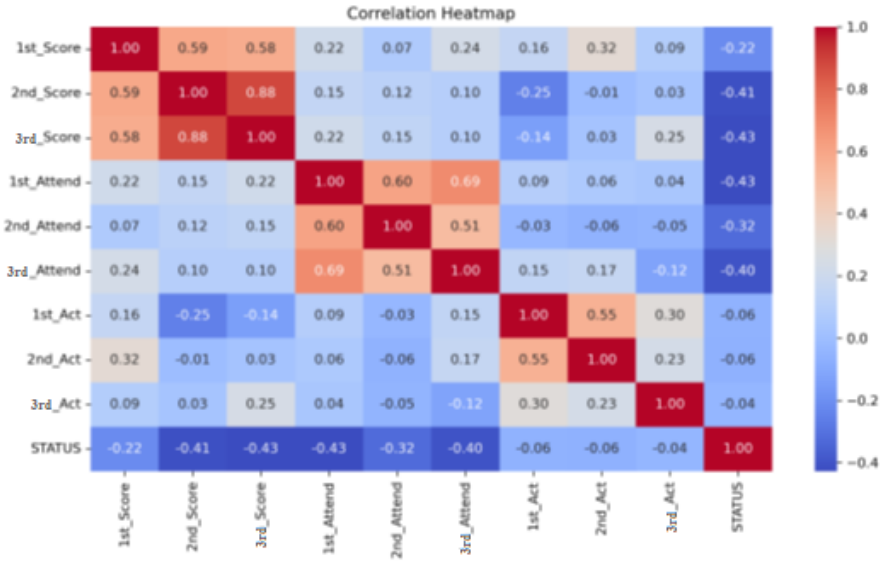


**Figure 3**. Correlation heatmap of academic, attendance, and socio-economic variables.

## 4.2. Model Performance

Two machine learning approaches were compared: a baseline decision tree classifier and a stacking ensemble model. The baseline decision tree, tuned with hyperparameters via GridSearchCV, achieved an accuracy of 0.949, recall of 0.700, and precision of 0.412, yielding an F1-score of 0.519. The area under the ROC curve (AUC = 0.830) indicates good discriminative ability. However, the relatively low precision means that while the tree was effective at capturing a majority of at-risk students (recall), it also frequently misclassified students who were not at risk as at-risk. For instance, in practice, this could translate into a school counselor being alerted for many students who may not actually need interventions, creating inefficiencies in resource allocation.

By contrast, the stacking ensemble, which combined multiple base learners with a logistic regression meta-model (also tuned with GridSearchCV), delivered markedly superior performance. It achieved an accuracy of 0.996 and an AUC of 1.000. Precision (0.909), recall (1.000), and F1-score (0.952) all indicate that the model not only captured all at-risk students (perfect recall) but also minimized false alarms (high precision). Figure 4 presents the Receiver Operating Characteristic (ROC) curves comparing the Decision Tree and Stacking Ensemble models. The Stacking Ensemble curve is closer to the upper-left corner, demonstrating higher classification performance (AUC = 1.000) compared to the Decision Tree model (AUC = 0.830. In practical terms, this suggests that the ensemble could serve as a highly reliable decision-support tool for schools in Pangkalpinang, allowing educators to focus interventions precisely where they are most needed.

Nevertheless, the ostensibly perfect discriminative ability (AUC = 1.000) must be interpreted with caution, as it raises important questions regarding overfitting and generalizability. While internal validation procedures were rigorously followed, such exemplary performance on a relatively small and localized dataset (n=1,017 from three schools in one city) could indicate that the model has learned patterns highly specific to this cohort and context. The risk is that the model may not maintain this level of performance when applied to data from new academic years, other schools in Pangkalpinang, or different regions of Indonesia. This potential limitation underscores that our results represent a powerful proof-of-concept for the local context rather than a guarantee of universal performance. The critical next step is external validation on unseen, independent datasets to confirm the model's robustness and practical utility beyond the initial training data.

 ISSN 2723-6471


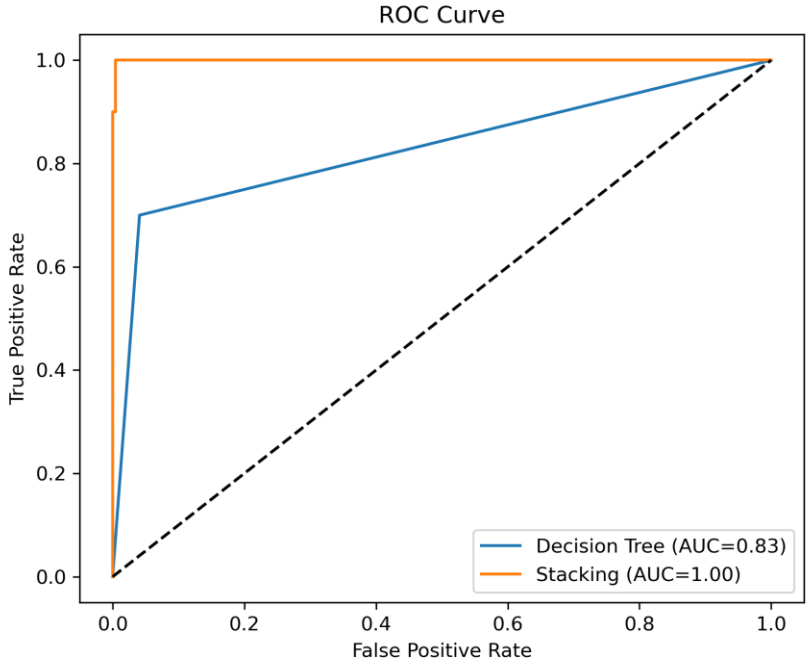


**Figure 4.** ROC curves comparing the Decision Tree and Stacking Ensemble models

Comparison of Model Performance with Recent Studies as shown in table 1. These results are consistent with recent findings in the field. [22] reported that stacking ensembles achieved an AUC of approximately 0.95 when predicting university student performance, outperforming individual classifiers such as decision trees (AUC ≈ 0.85). In another study, [23] found that ensemble models reduced false positives by nearly 20% compared to single classifiers in predicting dropout among secondary school students. Similarly, [24] highlighted logistic regression as an effective meta-model, showing F1-scores above 0.90 in their stacking framework for predicting engineering student outcomes.

Table 1 provides a numerical comparison of model performance metrics between this study and several recent works in educational prediction. While our stacking ensemble shows favorable metrics, it is crucial to interpret these comparisons with caution due to significant differences in contextual factors. The studies listed in table 1 differ from the present work in several key aspects that influence performance benchmarks, including the educational level, with studies such as Zhang et al. [25] and Vaarma et al. [23] conducted in university settings, which involve different student populations and attrition patterns than our high school context. Furthermore, the scale and nature of the datasets vary considerably; for instance, Mustofa et al. [24] and Villar & Andrade [26] utilized larger, potentially more heterogeneous national or multi-institutional datasets, whereas our study focuses on a localized, homogeneous dataset from one city. The predictive features also differed significantly across studies, with some incorporating employability-related attributes or online learning metrics, while our model relied primarily on academic and behavioral data available within the school system. Finally, the institutional and national contexts, spanning countries like China, Finland, and Pakistan with distinct educational systems and grading cultures, directly affect baseline performance and the difficulty of the prediction task.

**Table 1**. Comparison of Model Performance with Recent Studies

| Study | Accuracy | Precision | Recall | F1-Score | AUC | Context / Notes |
|---|---|---|---|---|---|---|
| **This Study (Stacking Ensemble)** | 0.996 | 0.909 | 1.0 | 0.952 | 1.0 | Local Indonesian HS; Academic/Attendance features |
| **Zhang et al.** [25] | 0.940 | 0.720 | 0.800 | 0.758 | 0.950 | Chinese University; Multi-source data |
| **Mustofa et al.** [24] | 0.930 | 0.880 | 0.920 | 0.900 | 0.930 | Multi-institutional College data |
| **Varma et al.** [23] | 0.920 | 0.750 | 0.780 | 0.764 | 0.900 | Finnish University; Demographic & engagement data |
| **Villar et al.** [26] | 0.900 | 0.600 | 0.700 | 0.650 | 0.850 | Pakistani University; Online learning platform |

Therefore, the performance advantage observed in table 1 should not be interpreted as a direct indicator of a superior model, but rather as evidence that our stacking ensemble performed exceptionally well within its specific localized context. The high performance may be attributable to the consistency of the local data and the model's tight alignment with Pangkalpinang's educational environment. This finding underscores the potential of tailored models for specific institutional contexts, even if their absolute performance metrics are not directly comparable to those from vastly different educational settings.

The consistency of our findings with the broader literature lies not in the absolute performance values, but in the relative effectiveness of ensemble methods. The fact that stacking improved upon a single Decision Tree classifier aligns with findings from [22] and [24] reinforcing that hybrid approaches can be beneficial for educational prediction across diverse contexts.

At the same time, decision trees continue to serve as useful baseline models due to their interpretability. [26] showed that while tree-based methods alone were less accurate than ensembles, they remained valuable for explaining which features (e.g., scores or attendance) most strongly influenced predictions, something educators often demand in practice. Thus, while the stacking model in this study offered exceptional predictive power, its black-box nature raises questions about transparency.

Taken together, the results reinforce two key insights: (1) tuning with GridSearchCV improved the fairness and performance of both baseline and meta-models, and (2) stacking ensembles with logistic regression as a meta-model can achieve near-perfect classification in educational datasets, though at the cost of reduced interpretability. These findings support the view that predictive analytics in education benefits most from hybrid strategies that combine the accuracy of ensembles with the interpretability of simpler models.

## 4.3. Feature Importances

The feature importance analysis, detailed in table 2, quantitatively confirms that academic scores and attendance were the most influential predictors of student outcomes. In contrast, demographic variables such as gender and parental occupation contributed minimally. The analysis, based on the mean decrease in impurity (Gini importance) from the Decision Tree model, showed that the 3rd Year Score (3rd_Score) was the most critical feature with an importance score of 0.512, followed by the 1st Year Score (1st_Score) at 0.217. The importance scores for attendance features, such as 1st_Year Attendance (1st_Attend), were also substantial at 0.092.

**Table 2**. Feature Importance

| Rank | Feature | Feature Importance Score | Contribution to Prediction |
|---|---|---|---|
| 1 | 3rd Year Score (3rd_Score) | 0.512 | Highest impact on final status; strongest single predictor. |
| 2 | 1st Year Score (1st_Score) | 0.217 | Strong early predictor of academic trajectory. |
| 3 | 2nd Year Score (2nd_Score) | 0.103 | Reinforces the pattern of academic consistency. |
| 4 | 1st Year Attendance (1st_Attend) | 0.092 | Key early indicator of student engagement. |
| 5 | 2nd Year Attendance (2nd_Attend) | 0.045 | Signal of sustained participation. |
| ... | ... | ... | ... |
| 12 | Father's Occupation | 0.008 | Minimal predictive influence. |
| 13 | Mother's Occupation | 0.006 | Minimal predictive influence. |
| 14 | Gender | 0.003 | Negligible predictive influence. |

Conversely, the importance scores for demographic variables were orders of magnitude lower. The combined importance for all encoded categories of Father's Occupation was 0.008, for Mother's Occupation was 0.006, and for Gender was 0.003. These negligible values provide clear numerical evidence that static demographic attributes were far less informative than dynamic academic and behavioral data for predicting graduation in this context. This finding resonates with the nature of the dataset itself: the three years of high school performance and attendance represent consistent, quantifiable records of student engagement, whereas demographic attributes remained static.

This result is consistent with recent empirical findings by [15] and [16] who also showed that academic engagement and attendance carry greater predictive weight than static demographic factors. The quantitative feature importance underscores that socioeconomic descriptors alone are poor proxies for the dynamic processes of learning and persistence observed in this student population.

From a policy perspective, these findings suggest that schools should prioritize monitoring first-year scores and attendance records. Early detection of declining attendance or poor performance can provide actionable signals for timely interventions. For example, a student with low 1st_Attend but average 1st_Score may be academically capable yet disengaged a candidate for counseling or mentoring programs. Conversely, students with persistently low scores may benefit from remedial support or differentiated instruction strategies.

Importantly, the minimal contribution of demographic factors implies that interventions should focus less on static characteristics such as parental occupation, and more on real-time student behavior and performance. This focus ensures that resource allocation is equitable and data-driven, targeting students based on demonstrated need rather than background assumptions. particularly valuable, as it identified that low academic performance combined with poor attendance created exponentially higher risk profiles.

### 4.4. Practical Implications

The findings offer a clear blueprint for improving student retention by prioritizing early-warning systems based on first-year academic scores and attendance. This enables a shift from reactive to proactive support, allowing schools to identify and assist at-risk students before academic difficulties become irreversible. For equitable and efficient policy, interventions should be triggered by these dynamic behavioral indicators rather than static socio-economic assumptions. This data-driven approach ensures help is allocated based on demonstrated need, preventing bias. Successful implementation requires pairing this technology with robust teacher training, equipping educators with the data literacy to interpret alerts and the skills to deliver tailored responses, such as counseling or remedial tutoring.

Finally, the deployment of predictive models must be guided by ethical oversight. Model outputs should serve as decision-support tools, not automated judgments, with teachers and counselors providing essential human context. Periodic model retraining and fairness audits are necessary to maintain equity and transparency, ensuring these systems enhance support without compromising ethical integrity.

### 4.5. Limitations and Future Research

This study has several limitations. First, its focus on high schools in Pangkalpinang may limit the generalizability of the findings to other regions of Indonesia. Second, the three-year data collection period could be extended in future longitudinal studies to better capture long-term trends. A key limitation involves the measurement of socioeconomic status (SES). Using only parental occupation as a proxy fails to capture the full complexity of SES. Incorporating additional indicators like household income and parental education would improve the model's accuracy.

Future research should prioritize external validation by testing the model in different Indonesian contexts to assess its true transferability. Internally, bootstrapping techniques could be used to better estimate the model's performance stability. Furthermore, exploring more advanced analytical methods and incorporating additional data sources (e.g., teacher assessments or psychological factors) could enhance predictive power and insight. Techniques to address class imbalance, such as SMOTE, should also be investigated to improve model resilience.

## 5. Conclusion

This study demonstrates the successful application of a stacking ensemble model to predict high school graduation in Pangkalpinang, Indonesia, achieving high accuracy by leveraging routinely collected academic and attendance data. The findings offer three key contributions with broader implications: first, providing a pragmatic blueprint for early-warning systems in resource-constrained environments by proving that effective predictive analytics can rely on simple, available data; second, advancing a theory of "dynamic equity" by showing that behavioral data are dominant predictors while demographic proxies are minimal, thereby shifting the intervention focus from static background to dynamic engagement for more equitable support; and third, advancing the methodological discourse on transparent educational AI by demonstrating that a hybrid approach balancing an interpretable Decision Tree with a powerful stacking

ensemble can integrate accuracy and explainability to build trust among educators. In summary, this research provides a validated, context-sensitive framework arguing that the strategic use of simple data can significantly improve educational equity and outcomes. Future work should focus on external validation across diverse regions and the integration of this predictive framework into real-world intervention protocols.

## 6. Declarations

### 6.1. Author Contributions

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

### 6.2. Data Availability Statement

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

### 6.3. Funding

Acknowledgments are addressed to Direktorat Riset, Teknologi dan Pengabdian kepada Masyarakat Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi Indonesia for providing funding in the Fundamental Research Scheme - Regular Research Program 2025 (Grant No. 0419/C3/DT.05.00/2025).

### 6.4. Institutional Review Board Statement

An ethical review was waived for this study as it used exclusively anonymized, secondary data provided by schools under formal collaboration. The research involved no direct student interaction or identifiable personal data. All procedures adhered to the ISB Atma Luhur Research Ethics Policy and the national guidelines from Indonesia's Ministry of Education, Culture, Research, and Technology.

### 6.5. Informed Consent Statement

Individual consent was waived as the study used pre-existing, anonymized administrative data. Institutional approval was granted by the participating school principals and the Pangkalpinang regional education authority. The research adhered to ethical guidelines for data protection, ensuring confidentiality throughout.

### 6.6. 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] J. M. Aiken, R. de Bin, M. Hjorth-Jensen, and M. D. Caballero, "Predicting time to graduation at a large enrollment American university," *PLoS One*, vol. 15, no. 11 November, pp. 1-12, Nov. 2020,

[2] L. Huang, L. R. Roche, E. Kennedy, and M. B. Brocato, "Using an Integrated Persistence Model to Predict College Graduation," *International Journal of Higher Education*, vol. 6, no. 3, pp. 40-52, May 2017,

[3] V. Zhang, B. Jeffries, and I. Koprinska, "A Machine Learning Approach for Predicting Student Progress in Online Programming Education," *Int J Artif Intell Educ*, vol. 2025, no. 1, pp. 1-12, 2025,

[4] E. Kalita, S. S. Oyelere, S. Gaftandzhieva., "Educational data mining: A 10-year review," *Discover Computing*, vol. 28, no. 81, pp. 1-21, 2025, doi: 10.1007/s10791-025-09589-z.

[5] A. Almalawi, B. Soh, A. Li, and H. Samra, "Predictive models for educational purposes: A systematic review," *Big Data and Cognitive Computing*, vol. 8, no. 12, pp. 1-17, 2024, doi: 10.3390/bdcc8120187.

[6] G. Akçapınar, M. N. Hasnine, R. Majumdar, B. Flanagan, and H. Ogata, "Developing an early-warning system for spotting at-risk students by using eBook interaction logs," *Smart Learning Environments*, vol. 6, no. 1, pp. 1-12, Dec. 2019,

[7] G. Akçapınar, A. Altun, and P. Aşkar, "Using learning analytics to develop early-warning system for at-risk students," *International Journal of Educational Technology in Higher Education*, vol. 16, no. 1, pp. 1-13, Dec. 2019,

[8] M. Kurniawan and S. Muhamad Isa, "Application of Data Mining for Prediction of High School Student Graduation Rates," *Jurnal Indonesia Sosial Teknologi*, vol. 5, no. 11, pp. 1-20, 2024,

[9] F. Aprilia, R. A. Anggraini, and Y. D. Putri, "Student Graduation Prediction Using Machine Learning Algorithms: Application of Linear and Logistic Regression on Educational Factors," *ROUTERS: Journal of Information Systems and Technology,*,vol. 3, no. 1, pp. 55–64, Feb. 2025

[10] S. Bum, I. B. Iorliam, E. O. Okube, and A. Iorliam, "Prediction of Student's Academic Performance Using Linear Regression," *Nigerian Annals Of Pure And Applied Sciences*, vol. 1, no. Dec., pp. 259–264, Dec. 2019,

[11] Muhammad Hadiza Baffa, Muhammad Abubakar Miyim, and Abdullahi Sani Dauda, "Machine Learning for Predicting Students' Employability," *UMYU Scientifica*, vol. 2, no. 1, pp. 001–009, Feb. 2023,

[12] S. O. Oppong, "Predicting Students' Performance Using Machine Learning Algorithms: A Review," *Asian Journal of Research in Computer Science*, vol. 16, no. 3, pp. 128–148, 2023,

[13] W. Ahmed, M. A. Wani, P. Plawiak, S. Meshoul, A. Mahmoud, and M. Hammad, "Machine learning-based academic performance prediction with explainability for enhanced decision-making in educational institutions," *Sci Rep*, vol. 15, no. 1, pp. 1-12, Dec. 2025,

[14] T. S. Tamir, "Traffic Congestion Prediction using Decision Tree, Logistic Regression and Neural Networks," *IFAC-PapersOnLine*, vol. 53, no. 5, pp. 512–517, Jan. 2020

[15] M. Pandey and V. K. Sharma, "A Decision Tree Algorithm Pertaining to the Student Performance Analysis and Prediction General Terms Data mining," *International Journal of Computer Applications*, vol. 61, no. 13, pp. 1-5, 2013.

[16] T. Swiderski, S. C. Fuller, and K. C. Bastian, "Student-Level Attendance Patterns Across Three Post-Pandemic Years," *Educ Eval Policy Anal*, vol. 2025, no. 1, pp. 1-12, 2025

[17] D. Khairy, N. Alharbi, M. A. Amasha, M. F. Areed, S. Alkhalaf, and R. A. Abougalala, "Prediction of student exam performance using data mining classification algorithms," *Educ Inf Technol (Dordr)*, vol. 29, no. 16, pp. 21621–21645, Nov. 2024,

[18] W. Ha, L. Ma, Y. Cao, Q. Feng, and S. Bu, "The effects of class attendance on academic performance: Evidence from synchronous courses during Covid-19 at a Chinese research university," *Int J Educ Dev*, vol. 104, no. Jan., pp. 1-12, Jan. 2024,

[19] O. Ojajuni et al., "Predicting Student Academic Performance Using Machine Learning," in *Computational Science and Its Applications – ICCSA 2021, Springer Science and Business Media Deutschland GmbH*, vol. 2021, no. 1, pp. 481–491, 2021.

[20] G. Keppens, "School absenteeism and academic achievement: Does the timing of the absence matter?," *Learn Instr*, vol. 86, no. Aug., pp. 1-19, Aug. 2023,

[21] M. I. Martínez-Serna, J. S. Baixauli-Soler, M. Belda-Ruiz, and J. Yagüe, "The effect of online class attendance on academic performance in finance education," *The International Journal of Management Education*, vol. 22, no. 3, pp. 1-23, Nov. 2024

[22] A. M. Rabelo and L. E. Zárate, "A model for predicting dropout of higher education students," *Data Science and Management*, vol. 8, no. 1, pp. 72–85, Mar. 2025

[23] M. Vaarma and H. Li, "Predicting student dropouts with machine learning: An empirical study in Finnish higher education," *Technol Soc*, vol. 76, no. Mar., pp. 1-24, Mar. 2024,

[24] S. Mustofa, Y. R. Emon, S. Bin Mamun, S. A. Akhy, and M. T. Ahad, "A novel AI-driven model for student dropout risk analysis with explainable AI insights," *Computers and Education: Artificial Intelligence*, vol. 8, no. jun., pp. 1-22, Jun. 2025,

[25] V. Zhang, B. Jeffries, and I. Koprinska, "A Machine Learning Approach for Predicting Student Progress in Online Programming Education," *Int J Artif Intell Educ*, vol. 2025, no. 1, pp. 1-12, 2025,

[26] A. Villar and C. R. V. de Andrade, "Supervised machine learning algorithms for predicting student dropout and academic success: a comparative study," *Discover Artificial Intelligence*, vol. 4, no. 1, pp. 1-12, Dec. 2024,