JIKO (Jurnal Informatika dan Komputer) Accredited KEMDIKTISAINTEK, No.0173/C3/DT.05.00/2025
Vol. 9, No. 1, April 2026, pp. 74-84 p-ISSN: 2614-8897
DOI: 10.33387/jiko.v9i1.11546 e-ISSN: 2656-1948

ROBUSTNESS EVALUATION OF GRADIENT BOOSTING MODELS FOR GRADUATION PREDICTION UNDER COHORT-BASED DISTRIBUTION SHIFTS

Rifandito Daniswara[1], Chanifah Indah Ratnasari[2*]

[1,2] Department of Informatics, Faculty of Industrial Technology, Universitas Islam Indonesia, Sleman, 55584, Indonesia
Email: [1]rifandito.daniswara@students.uii.ac.id, [*2]chanifah.indah@uii.ac.id

(Received: 28 February 2026, Revised: 12 March 2026, Accepted: 31 March 2026)

Abstract

Student graduation rate is a critical performance indicator for higher education institutions, particularly in accreditation assessment. Early prediction of on-time graduation supports academic planning and quality assurance. Although prior studies report high predictive accuracy using conventional cross-validation, limited attention has been given to robustness under cohort-based distribution shifts. This study evaluates the robustness of three gradient boosting models—Histogram-Based Gradient Boosting (HGB), Extreme Gradient Boosting (XGBoost), and Light Gradient Boosting Machine (LightGBM)—for predicting on-time graduation using structured academic trajectory data from 370 labeled instances across three cohorts. Two validation strategies were employed: a stratified 80:20 split and Leave-One-Group-Out (LOGO) validation. Under stratified evaluation, all models achieved macro F1-scores above 0.74, with HGB obtaining the highest score (0.7568). However, LOGO evaluation revealed substantial performance degradation, with mean F1-scores below 0.51 and increased variability across cohorts, indicating sensitivity to distribution shifts. XGBoost demonstrated comparatively better stability under distribution shifts. These findings indicate that high predictive accuracy under random splits does not guarantee cross-cohort robustness. This study therefore serves as a preliminary robustness-oriented comparison of validation settings and highlights the importance of distribution-aware validation for reliable deployment in educational data mining.

Keywords: *Graduation prediction, Distribution shift, Leave-One-Group-Out, Gradient boosting, Robustness evaluation, Educational data mining*

This is an open access article under the CC BY license.

**Corresponding Author: Chanifah Indah Ratnasari*

1. INTRODUCTION

Student graduation rate is one of the key performance indicators used to evaluate higher education institutions in Indonesia. In the latest accreditation guidelines issued by the National Accreditation Institution for Higher Education (BAN-PT) for undergraduate programs [1], institutions seeking the "Unggul" accreditation status must meet specific graduation benchmarks. These include the percentage of students graduating within one curriculum period (PK1MTK) ≥ 45%, the proportion of graduates completing their study within 1.5 curriculum periods (RPK1.5MTK) ≤ 30%, and the percentage of graduates within two curriculum periods (PK2MTK) ≥ 75%. These indicators are part of the "Efektivitas Kinerja Program Studi" (Effectiveness of Study Program Performance), which directly affects institutional accreditation outcomes. Therefore, early prediction of student graduation status is practically important for academic planning and quality assurance.

Various machine learning approaches have been applied to predict student graduation outcomes. Previous studies have implemented Support Vector Machine (SVM) and Naïve Bayes classifiers and reported satisfactory performance on moderate-sized academic datasets [2], [3]. Other works explored k-Nearest Neighbor (k-NN) and Decision Tree algorithms and also achieved competitive accuracy levels [4], [5]. More recent studies adopted ensemble-based methods such as Random Forest and Extreme Gradient Boosting (XGBoost), often combined with MissForest imputation to handle missing values [6]. In some cases, missing values were not treated purely as noise but were interpreted as meaningful academic

signals, such as students not attending exams or withdrawing from courses [7]. Neural network-based models, including Multilayer Perceptron (MLP) and Convolutional Neural Networks (CNN), have also been investigated, particularly when feature extraction and hierarchical modeling were required [8], [9]. These studies demonstrate that machine learning techniques can provide strong predictive performance in graduation and dropout prediction tasks. However, most prior graduation prediction studies primarily report predictive performance with limited discussion of whether model performance remains stable across different student cohorts.

Despite these advances, most existing studies evaluate models using aggregated datasets with conventional stratified splits or k-fold cross-validation. Such evaluation strategies assume that training and testing data are drawn from similar distributions. In practice, however, academic data are often subject to temporal and cohort-based variation. Changes in curriculum structure, enrollment size, academic policies, and external socioeconomic conditions may alter both the composition and distribution of student records across different cohorts. This condition is known as distribution shift, where the statistical properties of the data vary between training and deployment phases [10], [11]. When distribution shift occurs, models that appear accurate under random splits may experience performance degradation when applied to new cohorts. Therefore, evaluation limited to conventional cross-validation may not fully reflect model reliability in real institutional settings.

Recent research in machine learning has emphasized the importance of robustness evaluation under naturally occurring distribution shifts [10], [11]. One practical strategy to simulate such conditions is group-based validation, where data are partitioned according to meaningful groups rather than random sampling. Leave-One-Group-Out (LOGO) cross-validation is one such approach, in which each group is iteratively treated as a test set while the remaining groups are used for training [12], [13]. Compared to standard leave-one-out or random k-fold validation, LOGO allows evaluation under structured heterogeneity across predefined groups, such as cohorts or academic years. This strategy provides a more realistic estimate of model generalization when inter-group imbalance exists.

Although graduation prediction has been widely studied, robustness evaluation under cohort-based distribution shifts remains limited in educational data mining. Most prior studies emphasize predictive performance without explicitly assessing stability across naturally separated cohorts. This creates a gap between reported validation results and expected deployment performance on future student cohorts. Therefore, unlike previous graduation prediction studies, the present study examines whether conclusions derived from conventional validation remain valid under cohort-separated evaluation.

Based on this background, this study evaluates the robustness of graduation prediction models under cohort-based distribution shifts using Leave-One-Group-Out (LOGO) validation. Three gradient boosting algorithms—Extreme Gradient Boosting (XGBoost) [14], Histogram-Based Gradient Boosting (HGB), and Light Gradient Boosting Machine (LightGBM) [15]—are assessed within an identical modeling pipeline. The contribution of this study is threefold. First, it compares conventional validation and cohort-based robustness validation within the same graduation prediction setting. Second, it provides empirical evidence that strong results under pooled-data validation may not remain stable when the test data are separated by cohort. Third, it examines the relative stability of gradient boosting models under cross-cohort distribution shifts, thereby supporting the adoption of robustness-aware evaluation in educational data mining.

2. RESEARCH METHOD

This study applies a supervised machine learning framework to predict student on-time graduation using structured academic data. The experimental pipeline consists of data preprocessing, feature engineering, feature selection, model training, hyperparameter tuning, performance evaluation, and robustness assessment.

Two complementary validation strategies were employed. First, a stratified 80:20 train–test split was used to evaluate predictive performance under aggregated data conditions [16]. Second, robustness was assessed using Leave-One-Group-Out (LOGO) cross-validation to simulate cohort-based distribution shifts [11], [13]. Distribution shift refers to changes in statistical properties between training and testing data that may influence model generalization [10], [11]. This dual evaluation framework enables comparison between pooled-data evaluation and cross-cohort generalization. The complete research workflow is illustrated in Figure 1.

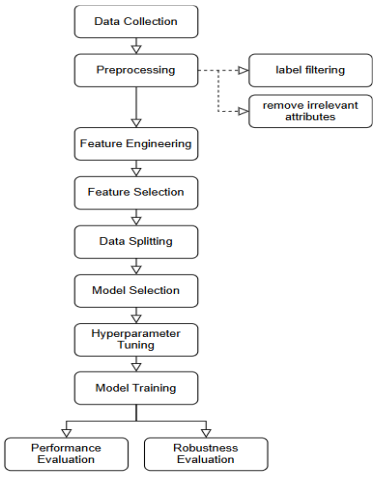

Figure 1. Overall Research Workflow

2.1 Data Collection

The dataset was obtained from the academic records of the undergraduate Informatics Program, Universitas Islam Indonesia. It consists of three student cohorts enrolled between 2019 and 2021, hereafter referred to as Group 1, Group 2, and Group 3. These cohorts serve as natural groups for robustness evaluation using Leave-One-Group-Out validation.

A total of 567 student records were collected, comprising 190 records from Group 1, 198 from Group 2, and 179 from Group 3. To prevent potential data leakage and ensure fairness in prediction, only academic information up to the sixth semester was considered in this study. All records were anonymized prior to analysis.

The original dataset contains 23 attributes, including student identification number (NIM), semester grade point average (IP1–IP6), cumulative grade point average (IPK1–IPK6), semester credits (SKS1–SKS6), attendance-related attributes (Jml_hadir, Jml_izin, and Jml_cuti), and graduation status as the target label. A simplified representation of these attributes is presented in Table 1.

Table 1. Original Dataset Attributes

Attribute	Description	Type
NIM	Student Number	Integer
IP1-IP6	Semester GPA	Float
IPK1-IPK6	Cumulative GPA	Float
SKS1-SKS6	Semester Credits	Integer
Jml_hadir	Attendance Count	Integer
Jml_izin	Excused Absence Count	Integer
Jml_cuti	Leave of Absence Count	Integer
Status	Graduation Label	Integer

The target variable "status" originally consists of five categories: not on-time graduation (0), on-time graduation (1), dropout (2), anomalous record (3), and not yet graduated (4). Since this study focuses on binary classification of graduation timeliness, the filtering and reorganization of these labels are described in Section 2.2.

2.2 Data Preprocessing

The preprocessing stage was conducted to align the dataset with the objective of binary classification for graduation timeliness. As stated in Section 2.1, the original label contains multiple categories. In this study, only records corresponding to on-time and not on-time graduation were retained, while the remaining categories were excluded to ensure consistency with the binary classification framework.

In addition to label filtering, several attributes were removed based on relevance considerations. The NIM attribute was excluded because it functions solely as an administrative identifier and does not provide predictive value. Attendance-related variables (Jml_hadir, Jml_izin, and Jml_cuti) were also excluded because they are recorded as cumulative totals rather than semester-based measurements. Consequently, these attributes do not reflect temporal academic progression, which is central to the semester-level modeling approach adopted in this study. Irrelevant variables were excluded to improve feature relevance and reduce potential noise in the modeling process [13].

2.3 Feature Engineering

To better capture academic progression dynamics beyond raw semester-level records, several statistical and temporal transformation features were constructed from semester GPA and credit variables. Temporal change was modeled using first-order differencing to represent short-term performance trends across consecutive semesters [17]. In addition, descriptive statistical measures were computed to summarize academic stability and dispersion over time. Prior studies have shown that feature engineering can improve predictive performance, particularly when raw records are transformed into more informative representations [18]. In this study, the same principle was applied to structured academic progression data through the following formulations.

Let x_t denote the semester GPA at semester t, and c_t denote the semester credit load at semester t, where $t = 1, 2, \ldots, T$ and $T = 6$ in this study. The formulas used to calculate engineered features are defined as follows:

a. Semester GPA Differences (GPA Delta)

$$\Delta x_t = x_t - x_{t-1} \quad (1)$$

b. Average Semester Credits

$$\bar{c} = \frac{1}{T}\sum_{t=1}^{T} c_t \quad (2)$$

c. GPA variance

$$Var(x) = \frac{1}{T-1}\sum_{t=1}^{T}(x_t - \bar{x})^2 \quad (3)$$

d. Minimum GPA

$$x_{min} = \min_{t \in \{1,\ldots,T\}} x_t \quad (4)$$

e. Maximum GPA

$$x_{max} = \max_{t \in \{1,\ldots,T\}} x_t \quad (5)$$

f. GPA Range

$$Range(x) = x_{max} - x_{min} \quad (6)$$

The complete list of engineered features derived from these formulations is summarized in Table 2.

Table 2. Engineered Features

Feature Name	Description
ΔIP2–ΔIP6	GPA change between consecutive semesters
Avg_SKS	Average semester credits
Var_IP	GPA variance
Min_IP	Minimum GPA
Max_IP	Maximum GPA
Range_IP	GPA range

2.4 Feature Selection

Feature selection was performed to identify inter-feature associations and reduce redundancy prior to

model training. Correlation-based analysis and heatmap visualization have been widely used to support feature filtering in predictive modeling, particularly for detecting highly associated variables that may contribute overlapping information [19]. In this study, redundancy was analyzed using the Pearson correlation coefficient, which measures the strength of linear association between continuous variables [20]. The resulting correlation matrix was visualized as a heatmap to facilitate interpretation of inter-feature relationships, as shown in Figure 2.

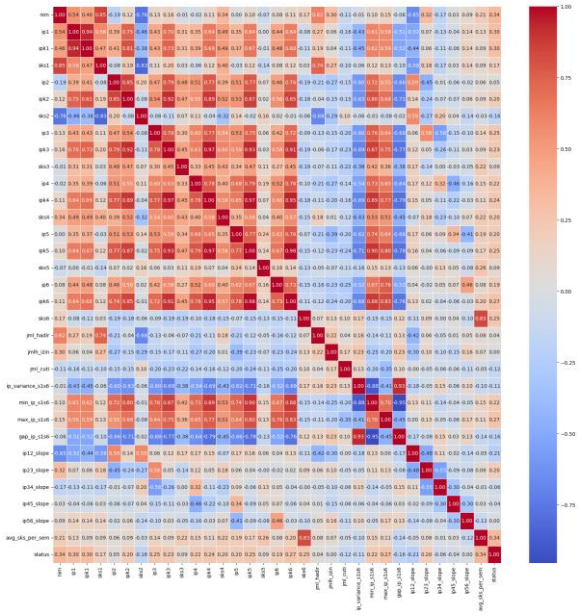

Figure 2. Pearson Correlation Matrix

As shown in Figure 2, semester GPA and cumulative GPA within the same semester exhibited strong associations, indicating substantial redundancy between the two representations. In addition, GPA range and minimum GPA showed high associations with GPA variance, suggesting overlapping information related to performance dispersion. A strong association was also observed between sixth-semester credits and average semester credits. Following Pearson-based filtering, feature pairs with coefficients above 0.80 were treated as highly collinear, and only one representative feature from each highly associated pair or group was retained [21], [22]. As a result, the final dataset consisted of 18 selected features, including semester GPA, semester credits, GPA delta, GPA variance, maximum GPA, and average semester credits.

2.5 Data Splitting

The data-splitting strategy in this study was designed to support two distinct evaluation objectives while utilizing the same 370 labeled instances. The difference between the two approaches lies solely in the partitioning mechanism. To prevent information leakage and ensure fair comparison, all model development procedures, including preprocessing within the training set, were conducted exclusively on the respective training portions.

a. Stratified 80:20 Split

For performance evaluation, a stratified 80:20 train–test split was applied to the aggregated dataset [16], [20]. Stratification preserves the class proportions in both the training and testing subsets, ensuring that they reflect the overall dataset distribution. A fixed random state value was used to maintain reproducibility.

The held-out test set was reserved exclusively for final performance evaluation. The distribution of instances after splitting is reported in Section 3.

b. Leave-One-Group-Out (LOGO)

In structured or grouped data settings, standard Leave-One-Out Cross-validation (LOOCV) may provide overly optimistic estimates of predictive performance due to correlation between training and test sets [13]. Therefore, a Leave-One-Group-Out (LOGO) validation strategy was implemented to better emulate out-of-group prediction scenarios and assess model generalization under distribution shifts [12], [13]. In this framework, cohort was used as the grouping variable. Given three cohorts in the dataset, the number of folds is:

$$K = 3$$

In each iteration, one cohort was treated as a completely unseen test set, while the remaining cohorts were combined as training data. Unlike the stratified split, class distributions in the test fold were not artificially controlled; instead, they reflected the natural distribution within each cohort. This design enables assessment of model generalization across group-based distribution variations. The configuration of the LOGO folds is illustrated in Table 3.

Table 3. LOGO Fold Configuration

Fold	Training Groups	Testing Group
Fold 1	Group 2 + Group 3	Group 1
Fold 2	Group 1 + Group 3	Group 2
Fold 3	Group 1 + Group 2	Group 3

2.6 Model Selection

The dataset contains several missing values (NaN), each of which conveys meaningful information. Considering this characteristic, this study employs three gradient boosting classifiers that are well suited for structured tabular data and capable of handling missing values natively without explicit imputation.

Three classifiers were evaluated:

a. Histogram-Based Gradient Boosting (HGB) implemented in Scikit-learn [16],
b. Extreme Gradient Boosting (XGBoost) [14], and
c. Light Gradient Boosting Machine (LightGBM) [15].

HGB uses feature binning to reduce computational complexity while maintaining competitive predictive performance [16]. XGBoost combines regularization, gradient-based optimization, and parallel processing to improve predictive accuracy on structured datasets [14]. LightGBM applies a leaf-wise tree growth strategy, enabling efficient learning on high-dimensional structured data [15]. Overall, these models were selected due to their efficiency,

robustness, and strong performance in tabular classification tasks [14], [15].

2.7 Hyperparameter Optimization

Hyperparameter tuning was performed using RandomizedSearchCV with stratified 5-fold cross-validation [16]. Instead of sampling from continuous ranges, this study employed a predefined discrete search space for each hyperparameter to balance computational efficiency and model stability.

During optimization, model selection was guided by the macro-averaged F1-score (F1_macro), which assigns equal weight to each class regardless of class frequency. The macro F1-score is computed as the average of class-wise F1-scores, as defined in Equation (7). This metric was chosen to ensure balanced optimization under class imbalance conditions [23].

$$F1_{macro} = \frac{1}{2}(F1_0 + F1_1) \quad (7)$$

where $F1_0$ and $F1_1$ denote the F1-scores of each class. Hyperparameter optimization was conducted exclusively on the training data to prevent information leakage. Within this training portion, stratified 5-fold cross-validation preserved class proportions across folds. A fixed number of randomized search iterations (*n_iter*=40) was applied to explore the predefined search space efficiently. The best-performing configuration obtained from cross-validation was then retrained on the full training set prior to final evaluation. The predefined hyperparameter ranges for each model are summarized in Table 4.

Table 4. Hyperparameter Search Space

Model	Parameter	Search Space
HGB	learning_rate	{0.01, 0.03, 0.05, 0.1, 0.2}
	max_iter	{100, 200, 300, 400, 500}
	max_depth	{None, 3, 5, 7, 9, 11}
	min_samples_leaf	{5, 10, 15, 20, 25}
	L2_regularization	{0.0, 0.001, 0.01, 0.1, 1, 5, 10, 50}
	max_bins	{64, 128, 255}
XGB	learning_rate	{0.01, 0.03, 0.05, 0.1, 0.2}
	n_estimators	{100, 200, 300, 400, 500}
	max_depth	{2, 3, 4, 5, 6, 7, 9, 11}
	min_child_weight	{5, 10, 15, 20, 25}
	reg_lambda	{0.0, 0.001, 0.01, 0.1, 1, 5, 10, 50}
	subsample	{0.6, 0.8, 1.0}
	colsample_bytree	{0.6, 0.8, 1.0}
LGBM	learning_rate	{0.01, 0.03, 0.05, 0.1, 0.2}
	n_estimators	{100, 200, 300, 400, 500}
	max_depth	{-1, 3, 5, 7, 9, 11}
	num_leaves	{7, 15, 31, 63, 127}
	min_child_samples	{5, 10, 15, 20, 25}
	reg_lambda	{0.0, 0.001, 0.01, 0.1, 1, 5, 10, 50}
	reg_alpha	{0.0, 0.001, 0.01, 0.1, 1}
	subsample	{0.6, 0.8, 1.0}
	subsample_frequency	{0, 1, 5}
	colsample_bytree	{0.6, 0.8, 1.0}

2.8 Model Training and Threshold Optimization

All models were trained and validated in a cloud-based Google Colab environment with Python 3 and CPU resources. The environment provided 12.7 GB of RAM and 107.7 GB of disk storage. Supporting libraries included Pandas and NumPy for data processing, Matplotlib for visualization, and Scikit-learn for model implementation.

Each classifier was trained using the optimized hyperparameters obtained from the randomized search procedure described in Section 2.7. Training was conducted exclusively on the designated training subset to prevent information leakage. After fitting, probabilistic outputs were generated for subsequent evaluation. The final evaluation results are presented in Section 3.

2.9 Performance Evaluation

Following model training, performance evaluation was conducted on the held-out test set. Model performance was assessed using accuracy, precision, recall, and macro-averaged F1-score (F1_macro) [23]. Due to class imbalance, F1_macro was designated as the primary evaluation metric because it assigns equal importance to both classes and provides a balanced assessment of predictive performance. The macro F1-score was computed as previously defined in Section 2.7. When multiple thresholds yield identical macro F1-scores, threshold selection prioritizes the configuration with the smallest number of false positives.

Let True Positives (TP), False Positives (FP), False Negatives (FN), and True Negatives (TN) denote the entries of the confusion matrix. The evaluation metrics are defined as follows:

a. Accuracy

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (8)$$

b. Precision

$$Precision = \frac{TP}{TP+FP} \quad (9)$$

c. Recall

$$Recall = \frac{TP}{TP+FN} \quad (10)$$

d. Class-specific F1-score:

$$F1_c = 2\frac{Precision_c \cdot Recall_c}{Precision_c + Recall_c} \quad (11)$$

To further enhance evaluation robustness, decision threshold optimization was conducted instead of using the default threshold of 0.5. The optimal threshold was selected using out-of-fold (OOF) predicted probabilities obtained from stratified 5-fold cross-validation [20].

The OOF probability for each instance *i* is defined in Equation (12):

$$\hat{p}_i = Pr(\hat{y}_i = 1 | x_i; \theta_{-k(i)}) \quad (12)$$

where $\theta_{-k(i)}$ represents the model trained without the fold containing instance i, ensuring that predictions are generated without information leakage.

Predicted class labels under threshold t are obtained as:

$$\hat{y}_i(t) = 1[\hat{p}_i \geq t] \quad (13)$$

For each threshold $t \in T = \{0.00, 0.01, \ldots, 0.99\}$, the macro-averaged F1-score was computed. The optimal threshold was determined by maximizing macro F1, as defined in Equation (14):

$$t^* = arg\max_{t \in T} F1_{macro}\left(y, \hat{y}(t)\right) \quad (14)$$

where t^* denotes the optimal decision threshold, T represents the candidate threshold set, y is the true label, and $\hat{y}^{(t)}$ is the predicted label obtained using threshold t. This formulation aligns with theoretical analyses of F1-optimal decision rules under imbalanced classification settings [24]. Threshold optimization was conducted exclusively on the training data, and the selected threshold was subsequently applied to the held-out test set for final performance evaluation, ensuring independence between model calibration and final assessment.

2.10 Robustness Assessment

In this study, robustness refers to the ability of a model to maintain stable predictive performance under group-based distribution shifts [10]. Robustness under natural distribution shifts has been widely discussed in machine learning literature, particularly in domain generalization and out-of-distribution (OOD) evaluation settings [11], [25].

Robustness evaluation was conducted under the Leave-One-Group-Out (LOGO) framework described in Section 2.5. In each iteration, one cohort was treated as an unseen test domain, while the remaining cohorts were used for training. This setting simulates temporal distribution shifts across cohorts and enables evaluation under naturally varying class distributions.

Let M_k denote the evaluation metric (macro F1-score or precision) obtained on the held-out group in fold k, and let K represent the total number of cohorts. Robustness was quantified using the mean and standard deviation of M_k across folds, defined as:

$$\mu_M = \frac{1}{K}\sum_{k=1}^{K} M_k \quad (15)$$

$$\sigma_M = \sqrt{\frac{1}{K}\sum_{k=1}^{K}(M_k - \mu_M)^2} \quad (16)$$

A model is considered more robust if it achieves both high mean performance and low variability across folds. This statistical characterization is consistent with domain generalization evaluation practices, where cross-domain stability is regarded as a key indicator of reliable out-of-distribution performance [25].

3. RESULT AND DISCUSSION

This section presents the experimental results and provides analytical discussion aligned with the research objectives. The presentation begins with dataset characteristics after preprocessing, followed by predictive performance under aggregated evaluation. Subsequently, robustness performance under cohort-based distribution shifts is analyzed.

3.1 Dataset Characteristics After Preprocessing

After applying the preprocessing procedure described in Section 2.2, the dataset was reduced from 567 records to 370 labeled instances consisting only of on-time (1) and not on-time (0) graduation categories. The distribution of labels across cohorts is summarized in Table 5.

Table 5. Label Distribution After Preprocessing

Group	Total Instances	Not On-Time (0)	On-Time (1)
Group 1	143	86	57
Group 2	134	87	47
Group 3	93	12	81

As shown in Table 5, Group 1 and Group 2 exhibit relatively balanced distributions, although the majority class remains "not on-time." In contrast, Group 3 demonstrates a substantial shift, where the majority of students graduated on time (81 out of 93 instances). This difference indicates the presence of cohort-level distribution variation.

Such variation reflects the type of natural distribution shift discussed in Section 1 and Section 2.5. When the class proportion differs substantially across cohorts, a model trained on aggregated data may not generalize consistently to unseen groups. Therefore, the use of macro-averaged metrics becomes important to ensure balanced evaluation across classes.

3.2 Predictive Performance Under Stratified 80:20 Split

Predictive performance under aggregated data conditions was evaluated using a stratified 80:20 train–test split. The stratification procedure preserved the class proportions between training and testing subsets. The resulting label distribution after splitting is illustrated in Figure 3, confirming that both subsets reflect the overall dataset composition.

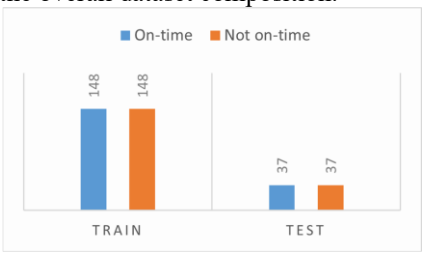

Figure 3. Label Distribution After Stratified 80:20 Split

Hyperparameter optimization produced distinct configurations for each model. The best macro F1-score obtained during 5-fold cross-validation on the training portion is summarized in Table 6.

Table 6. Optimal Hyperparameters and Cross-Validation Macro F1-Score

Model	Selected Configuration (Key Parameters)	Macro F1 (CV)
HGB	learning_rate=0.01, max_iter=300, max_depth=3, min_samples_leaf=25	0.7323
XGB	learning_rate=0.1, n_estimators=100, max_depth=6, min_child_weight=5, reg_lambda=10	0.7084
LGBM	learning_rate=0.03, n_estimators=400, max_depth=11, num_leaves=7, reg_lambda=50	0.7255

Based on cross-validation results, HGB achieved the highest macro F1-score (0.7323), followed by LGBM (0.7255) and XGB (0.7084). Although the differences are moderate, this indicates that HGB provided slightly better balance between precision and recall during internal validation. Each model was subsequently retrained on the full training set using its respective optimized configuration.

Before final testing, decision threshold calibration was applied. The selected optimal thresholds are reported in Table 7.

Table 7. Optimal Decision Threshold

Model	Optimal Threshold
HGB	0.53
XGB	0.51
LGBM	0.52

The optimal thresholds remain close to the default value of 0.50, suggesting that only minor probability calibration was required to maximize macro F1-score and minimize false positives. The final performance on the held-out test set is presented in Table 8.

Table 8. Final Performance on Held-Out Test Set

Model	Accuracy	Precision	Recall	Macro F1
HGB	0.7568	0.7568	0.7568	0.7568
XGB	0.7568	0.7714	0.7297	0.7566
LGBM	0.7432	0.7500	0.7297	0.7432

As shown in Table 8, HGB achieved the highest macro F1-score (0.7568), although the difference compared to XGB (0.7566) is marginal. LGBM produced slightly lower performance (0.7432). Overall, the three gradient boosting models demonstrate comparable predictive capability under aggregated evaluation.

XGB obtained the highest precision (0.7714), indicating fewer false positive predictions. However, its recall (0.7297) is slightly lower than that of HGB. This reflects a trade-off between precision and recall. Since the macro F1-score balances both metrics equally, HGB achieves the most balanced overall performance. To further interpret the aggregated test result, the confusion matrix of XGB is presented in Figure 4.

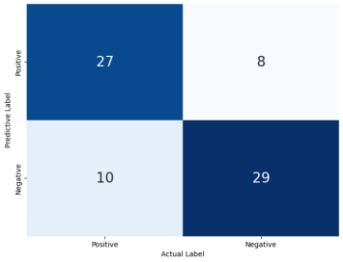

Figure 4. Confusion Matrix of XGB on the Hold-Out Test Set

As shown in Figure 4, correct predictions are obtained for both classes, with relatively limited false positives and false negatives. This indicates that, under stratified aggregated evaluation, XGB maintains a reasonably balanced classification behavior between on-time and not-on-time graduation classes. Thus, the macro F1-score obtained in the hold-out test is supported not only by overall accuracy but also by a relatively stable trade-off between precision and recall.

The relatively small performance gap among models suggests that trajectory-based academic features provide strong predictive signals, regardless of the specific gradient boosting variant used. Under conventional stratified evaluation, all models demonstrate stable and satisfactory classification performance.

Nevertheless, aggregated evaluation assumes similar training and testing distributions. As discussed earlier, such conditions may not reflect real cohort-based variations. Therefore, robustness analysis under LOGO validation is necessary to examine cross-cohort generalization performance.

3.3 Robustness Under Cohort-Based Distribution Shifts (LOGO)

To further examine model behavior under cohort heterogeneity, robustness testing was conducted using the Leave-One-Group-Out (LOGO) setting, where each cohort was alternately treated as an unseen test set. This evaluation focuses on how performance changes when the class distribution differs from the training data.

Figures 5–7 present the detailed performance patterns for HGB, XGB, and LGBM across LOGO folds. A consistent trend emerges: in certain groups—particularly Group 2—recall increases substantially, while accuracy, precision, and F1-score decline. This indicates that the models tend to predict the positive class (on-time graduation) more frequently in these cohorts. As a result, more true positives are captured (high recall), but false positives also increase, lowering precision and consequently reducing the F1-score. In contrast, when evaluated on Group 3, precision reaches its highest value while recall drops, suggesting a more conservative prediction pattern.

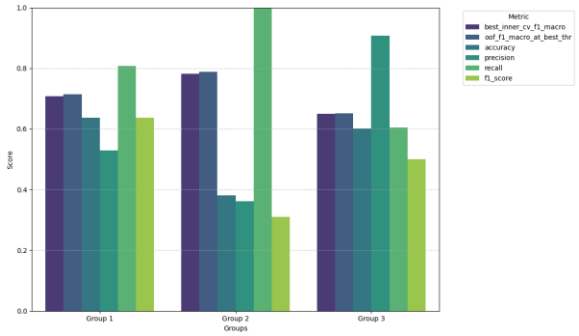

Figure 5. HGB performance across LOGO folds, showing changes in accuracy, precision, recall, and macro F1 under cohort-based distribution shifts

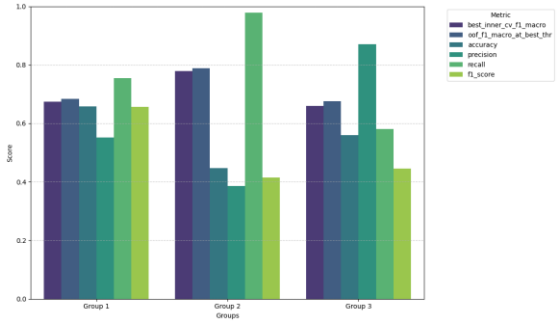

Figure 6. XGB performance across LOGO folds, showing changes in accuracy, precision, recall, and macro F1 under cohort-based distribution shifts

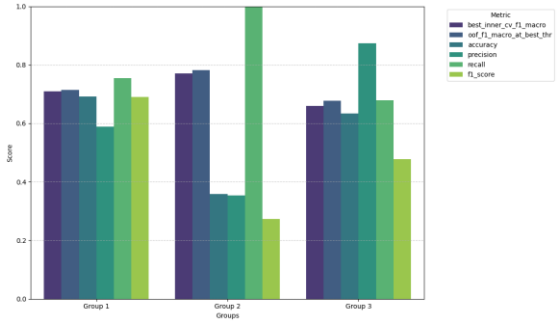

Figure 7. LGBM performance across LOGO folds, showing changes in accuracy, precision, recall, and macro F1 under cohort-based distribution shifts

The aggregated mean performance across LOGO folds is summarized in Table 9.

Table 9. Mean Performance Across LOGO Folds

Model	Accuracy	Precision	Recall	Macro F1
HGB	0.5397	0.5992	0.8039	0.4820
XGB	0.5547	0.6027	0.7711	0.5056
LGBM	0.5616	0.6051	0.8111	0.4798

All models maintain relatively high recall (above 0.77), confirming their sensitivity in identifying students who graduate on time. However, precision remains around 0.60, reflecting a tendency to overpredict the positive class under cohort shifts. Consequently, F1-scores remain below 0.51. Compared to the stratified evaluation (F1 $\approx$ 0.75), LOGO evaluation reduced the mean F1-score by approximately 33%, highlighting the sensitivity of the models to cohort-based distribution shifts. Among the three models, XGB achieves the highest mean F1-score (0.5056), indicating comparatively better cross-cohort performance. To evaluate stability, Table 10 reports the standard deviation of metrics across LOGO folds.

Table 10. Standard Deviation Across LOGO Folds

Model	Accuracy	Precision	Recall	Macro F1
HGB	0.1388	0.2796	0.1975	0.1641
XGB	0.1048	0.2459	0.1997	0.1314
LGBM	0.1785	0.2601	0.1678	0.2086

LGBM exhibits the highest variability in macro F1-score (0.2086), suggesting lower stability across cohorts. In contrast, XGB shows the lowest macro F1 standard deviation (0.1314), indicating more consistent behavior under distribution shifts.

The substantial degradation compared to the stratified evaluation can be linked to differences in label proportions across cohorts. As shown previously, Group 3 presents a markedly imbalanced distribution (12 vs. 81). When such a cohort is used as the test set, the mismatch between training and testing class proportions alters the effective decision boundary, leading to performance fluctuations. To further examine how this instability appears at the class-prediction level, the confusion matrices of the XGB model across LOGO folds are presented in Figure 8.

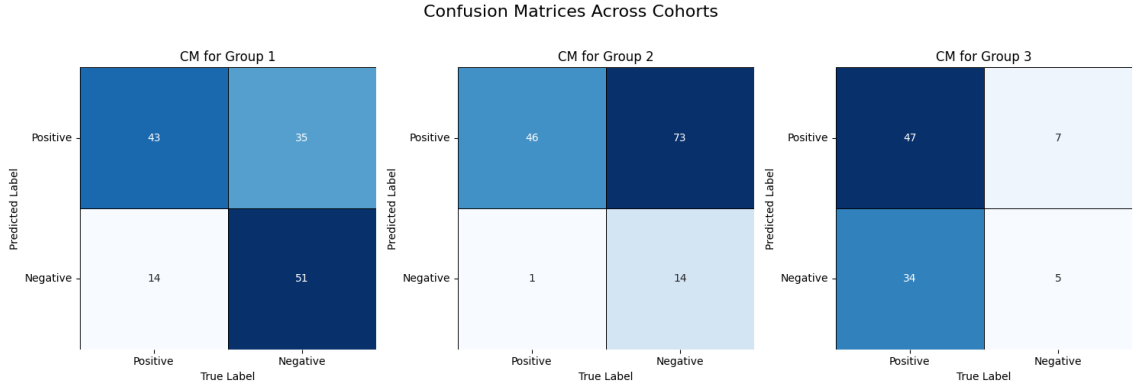

Figure 8. Confusion matrices of the XGB model across LOGO folds, showing how cohort-based distribution shifts alter the balance of false positives and false negatives

Figure 8 shows that the decline in LOGO performance is not caused by a uniform failure pattern across all cohorts. Instead, the dominant error type changes depending on the held-out group. In Group 1, the model still produces a moderate balance between correct and incorrect predictions, although misclassifications remain noticeably higher than in the aggregated hold-out setting. In Group 2, false positives increase substantially (73), indicating that the model becomes overly permissive in predicting the positive class. This explains why recall remains high while precision declines in this fold. In contrast, when Group 3 is used as the test cohort, false negatives become more prominent (34), suggesting a more conservative prediction behavior toward the positive class. As a result, precision increases while recall decreases. These findings indicate that the reduction in macro F1-score under LOGO is mainly driven by unstable error balance across cohorts rather than by a complete inability to classify one class.

Overall, the LOGO results demonstrate that cohort-based distribution shifts significantly influence predictive behavior by altering the balance between false positives and false negatives across folds, revealing performance instability that is not observable under random stratified splits.

3.4 Discussion

The findings of this study answer the research question regarding robustness under cohort-based distribution shifts. Under aggregated stratified evaluation, all three gradient boosting models demonstrate competitive performance, with macro F1-scores above 0.74. This suggests that trajectory-based academic features are informative for predicting graduation timeliness. However, robustness evaluation reveals a different perspective. When exposed to unseen cohorts through LOGO validation, performance decreases substantially, and variability across folds increases. This indicates that model reliability depends strongly on the similarity between training and testing distributions.

Among the evaluated models, XGB achieves the best balance between mean performance and variability under LOGO. This suggests that its regularization mechanism may contribute to improved stability across cohorts, although performance differences remain moderate. Overall, these results emphasize the importance of robustness-aware evaluation in educational data mining. Evaluation based solely on random splits may overestimate real-world performance when cohort-level variation exists. Incorporating group-based validation provides a more realistic estimate of model generalization in institutional deployment scenarios. The study therefore supports the argument that predictive accuracy should be complemented with stability analysis to ensure reliable academic decision support systems.

4. CONCLUSION

This study evaluated the robustness of student graduation prediction models under cohort-based distribution shifts. Using academic trajectory features derived from semester GPA and credit records, three gradient boosting algorithms—HGB, XGBoost, and LightGBM—were assessed under both stratified random splitting and Leave-One-Group-Out (LOGO) validation. Under aggregated stratified evaluation, all models demonstrated strong predictive performance, achieving macro F1-scores above 0.74, indicating that trajectory-based academic features provide meaningful signals for predicting graduation timeliness. However, when evaluated under cohort-based LOGO validation, performance declined substantially, with mean macro F1-scores below 0.51 and increased variability across cohorts. These results confirm that models trained on aggregated data may not generalize consistently when cohort-level distribution differences exist. Among the evaluated models, XGBoost showed comparatively better stability, achieving the highest mean F1-score and lowest variability under LOGO evaluation. The findings highlight the importance of robustness-aware validation strategies in educational data mining. For practical deployment in academic decision support systems, model evaluation should therefore incorporate group-based validation to obtain realistic generalization estimates.

Despite these contributions, the findings should be interpreted in light of several limitations. First, robustness evaluation was conducted on only three cohorts, which limits the granularity of cross-cohort analysis. Second, the number of samples retained after preprocessing was relatively modest, which may affect the stability of model estimation under LOGO evaluation. Third, the feature set was derived mainly from structured academic trajectory variables and did not incorporate broader contextual factors that may vary across cohorts. Therefore, the present results should be regarded as an initial robustness-oriented evaluation rather than a final claim of cross-cohort generalization. Future research may extend this work by using larger longitudinal datasets, incorporating richer contextual features, and exploring domain adaptation techniques, temporal modeling, or recalibration strategies to mitigate performance degradation under cohort shifts and improve long-term institutional reliability.

5. REFERENCE

[1] Badan Akreditasi Nasional Perguruan Tinggi, "IAPS 5.0," 2025. Available: https://www.banpt.or.id/wp-content/uploads/2025/06/Peraturan-BAN-PT-Nomor-13-Tahun-2025-ttg-IAPS-5.0.pdf

[2] E. Haryatmi and S. Pramita Hervianti, "Penerapan Algoritma Support Vector Machine Untuk Model Prediksi Kelulusan Mahasiswa Tepat Waktu," *Jurnal RESTI (Rekayasa Sistem dan Teknologi*

Informasi), vol. 5, no. 2, pp. 386–392, Apr. 2021, doi: 10.29207/resti.v5i2.3007.

[3] N. Yustira, D. Witarsyah, and E. Sutoyo, "Implementasi Algoritma NaÏve Bayes Classification Untuk Klasifikasi Kelulusan Mahasiswa Tepat Waktu," *eProceedings of Engineering*, vol. 8, no. 5, Oct. 2021. Available: https://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/view/16721

[4] E. Novianto, A. Hermawan, and D. Avianto, "Klasifikasi Algoritma K-Nearest Neighbor, Naive Bayes, dan Decision Tree untuk Prediksi Status Kelulusan Mahasiswa S1," *Rabit : Jurnal Teknologi dan Sistem Informasi Univrab*, vol. 8, no. 2, pp. 146–154, Jul. 2023, doi: 10.36341/rabit.v8i2.3434.

[5] G. A. J. Satvika, I. N. Sukajaya, and I. G. A. Gunadi, "Improving k-nearest neighbor performance using permutation feature importance to predict student success in study," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 35, no. 3, pp. 1835–1844, Sep. 2024, doi: 10.11591/ijeecs.v35.i3.pp1835-1844.

[6] I. Nirmala, H. Wijayanto, and K. A. Notodiputro, "Prediction of Undergraduate Student's Study Completion Status Using MissForest Imputation in Random Forest and XGBoost Models," *ComTech: Computer, Mathematics and Engineering Applications*, vol. 13, no. 1, pp. 53–62, Feb. 2022, doi: 10.21512/comtech.v13i1.7388.

[7] A. J. Fernandez-Garcia, J. C. Preciado, F. Melchor, R. Rodriguez-Echeverria, J. M. Conejero, and F. Sanchez-Figueroa, "A real-life machine learning experience for predicting university dropout at different stages using academic data," *IEEE Access*, vol. 9, pp. 133076–133090, 2021, doi: 10.1109/ACCESS.2021.3115851.

[8] M. Windarti and P. T. Prasetyaninrum, "Prediction Analysis Student Graduate Using Multilayer Perceptron," in *Proceedings of the International Conference on Online and Blended Learning 2019 (ICOBL 2019)*, Paris, France: Atlantis Press, May 2020, pp. 53–57. doi: 10.2991/assehr.k.200521.011.

[9] A. Salam and J. Zeniarja, "Classification of deep learning convolutional neural network feature extraction for student graduation prediction," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 32, no. 1, pp. 335–341, Oct. 2023, doi: 10.11591/IJEECS.V32.I1.PP335-341.

[10] P. W. Koh *et al.*, "WILDS: A Benchmark of in-the-Wild Distribution Shifts," Jul. 2021, doi: 10.48550/arXiv.2012.07421.

[11] Z. Shi *et al.*, "Effective robustness against natural distribution shifts for models with different training data," in *Proceedings of the 37th International Conference on Neural Information Processing Systems*, in NIPS '23. Red Hook, NY, USA: Curran Associates Inc., Dec. 2023. doi: 10.5555/3666122.3669339.

[12] Z. Liu, J. Van Niekerk, and H. Rue, "Leave-group-out cross-validation for latent Gaussian models," Jul. 2025, doi: 10.57645/20.8080.02.25.

[13] A. Adin, E. T. Krainski, A. Lenzi, Z. Liu, J. Martínez-Minaya, and H. Rue, "Automatic cross-validation in structured models: Is it time to leave out leave-one-out?," *Spat. Stat.*, vol. 62, Aug. 2024, doi: 10.1016/j.spasta.2024.100843.

[14] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, in KDD '16. New York, NY, USA: Association for Computing Machinery, 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[15] G. Ke *et al.*, "LightGBM: a highly efficient gradient boosting decision tree," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, in NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, pp. 3149–3157. doi: 10.5555/3294996.3295074.

[16] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," Jun. 2018, doi: 10.48550/arXiv.1201.0490.

[17] Y. Ben Baccar, "Comparative Study on Time Series Forecasting," Master of Science (Data Science), Institut Polytechnique de Paris, Telecom Paris, Paris, 2019. doi: 10.13140/RG.2.2.32241.02408.

[18] K. Hubbard and S. Amponsah, "Feature Engineering on LMS Data to Optimize Student Performance Prediction," in *New Frontiers in Data Science*, J. Han Henry and Stamey, Ed., Springer Nature Switzerland, 2025, pp. 125–142. doi: 10.1007/978-3-031-99879-9_8.

[19] Amiruddin and R. Ishak, "Implementasi Seleksi Fitur Klasifikasi Waktu Kelulusan Mahasiswa Menggunakan Correlation Matrix with Heatmap," *Jambura Journal of Electrical and Electronics Engineering*, vol. 4, no. 2, pp. 169–174, Jul. 2022, doi: 10.37905/jjeee.v4i2.14403.

[20] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics)*, 2nd ed. NY: Springer New York, 2009. doi: 10.1007/978-0-387-84858-7.

[21] N. Tri, W. Kurniawan, M. Ariyadi, and B. Efendi, "A Hybrid Approach of Pearson Correlation and PCA in Feature Selection for Opinion Mining," *IJID (International Journal on Informatics for Development)*, vol. 14, pp. 601–615, Feb. 2025, doi: 10.14421/ijid.2025.5195.

[22] J. B. Diekuu, M. S. Mekala, U. S. Abonie, J. Isaacs, and E. Elyan, "Predicting student next-term performance in degree programs using AI-

based approach: a case study from Ghana," *Cogent Education*, vol. 12, no. 1, 2025, doi: 10.1080/2331186X.2025.2481000.

[23] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf. Process. Manag.*, vol. 45, pp. 427–437, Mar. 2009, doi: 10.1016/j.ipm.2009.03.002.

[24] Z. C. Lipton, C. Elkan, and B. Naryanaswamy, "Optimal Thresholding of Classifiers to Maximize F1 Measure," in *Machine Learning and Knowledge Discovery in Databases*, F. and H. E. and M. R. Calders Toon and Esposito, Ed., Berlin, Heidelberg: Springer Berlin Heidelberg, 2014, pp. 225–239. doi: 10.1007/978-3-662-44851-9_15.

[25] I. Gulrajani and D. Lopez-Paz, "In Search of Lost Domain Generalization," *International Conference on Learning Representations*, Jul. 2020, doi: 10.48550/arXiv.2007.01434.