Celal Bayar University Journal of Science

A Comparative Study of Ensemble Learning Models for Predicting Student On-Time Graduation

Ahmet Kala[1]*, Cem Özkurt[2], Hasan Yaşar[1]

[1]Sakarya University of Applied Sciences, Department of Information Technologies, 54050 Sakarya, Türkiye.
[2] Sakarya University of Applied Sciences, Department of Computer Engineering, 54050 Sakarya, Türkiye.
* ahmetkala@subu.edu.tr
* Orcid No: 0000-0002-0598-1181

Received: February 21, 2025
Accepted: June 18, 2025
DOI: 10.18466/cbayarfbe.1644296

Abstract

On-time graduation is one of the primary goals of undergraduate programs, aiming to equip students with the necessary knowledge and skills in a specific field to ensure their rapid integration into the workforce. However, delays in graduation can postpone career entry, increase financial burdens, and create psychological stress. Therefore, accurately predicting students' graduation outcomes at early stages is critically important for higher education institutions to develop effective educational strategies and provide timely interventions for at-risk individuals. This study aimed to determine the most effective approach for predicting on-time graduation by comparing ensemble learning methods with traditional machine learning models using demographic data, high school performance, and progressively accumulated university academic data over six academic periods. The models used include ensemble methods such as Random Forest, CatBoost, and XGBoost, as well as traditional models like Logistic Regression, Decision Trees, and Support Vector Machines. Across seven prediction checkpoints (T0–T6), CatBoost consistently outperformed all other models, achieving 83.0% accuracy, 84.8% F1-score, 84.4% precision, and 85.1% recall in the final prediction stage (T6). Among traditional models, Logistic Regression performed the best, with 83.5% accuracy and 84.7% F1-score at T6. Statistical analyses (paired t-test and Wilcoxon test) confirmed that ensemble models significantly outperformed traditional ones in terms of both accuracy and F1-score ($p < 0.05$).

Keywords: On-time graduation, Ensemble learning, Machine learning.

1. Introduction

Predicting student graduation performance is critically important for higher education institutions, as it contributes to enhancing academic success, ensuring student retention, and optimizing the effective use of institutional resources. On-time graduation rates are not only indicators of individual achievement but also reflect the overall performance of educational institutions. Increasing these rates improves student satisfaction, strengthens institutional reputation, and promotes more effective collaboration between students and academic staff. Additionally, it enables more efficient allocation and utilization of institutional resources.

Early identification of students at risk of academic failure allows for timely and personalized interventions, thereby improving academic outcomes. Such early intervention mechanisms help identify the challenges students face and facilitate the development of targeted support strategies. Regular monitoring and analysis of graduation rates also support the evaluation of educational policies and the development of more effective teaching methods.

In recent years, the growing accessibility of educational data has significantly expanded the application of machine learning (ML) and ensemble learning techniques in the education domain. These methods are increasingly employed to predict academic performance and design support systems tailored to students' needs. While traditional methods such as statistical analysis and decision trees have been widely used, advanced ensemble learning models—such as Random Forest (RF), Gradient Boosting, CatBoost, and XGBoost—offer greater potential for capturing complex patterns in educational data. These models integrate multiple algorithms to provide more accurate and reliable predictions. Ensemble methods synthesize multiple learning algorithms to

improve accuracy and robustness, which has made them particularly valuable in fields such as bioinformatics, medical diagnostics, and education analytics [1-3]. Although AdaBoost and Random Forest have been commonly utilized in the literature, more advanced techniques like Gradient Boosting, XGBoost, LightGBM, and CatBoost have received limited attention in graduation prediction studies targeting undergraduate students [4–12].

This study aims to address this gap by evaluating the effectiveness of advanced ensemble learning techniques—namely CatBoost, Gradient Boosting, and XGBoost—in predicting on-time graduation. The primary objective is to support institutional planning and improve student success by enabling early identification of at-risk undergraduate students. The study compares ensemble learning methods with traditional machine learning algorithms and analyzes their performance across seven academic checkpoints (T0–T6). The dataset includes demographic information, high school academic records, national university entrance exam (YKS) scores, and higher education academic performance data collected from undergraduate students at a public university.

This study provides a comparative evaluation of advanced ensemble learning methods for predicting on-time graduation, highlights the effectiveness of underexplored models such as CatBoost, introduces a time-series-based analytical framework, and offers practical contributions by demonstrating how the results can be integrated into student information systems to support early intervention strategies.

The remainder of this paper is structured as follows: Section II presents a comprehensive review of recent studies on on-time graduation prediction using machine learning models; Section III details the dataset, feature engineering, and methodological framework adopted in this study; Section IV reports and discusses the model evaluation results across multiple prediction periods; and Section V concludes the paper by summarizing the key findings, practical implications, and suggestions for future research.

2. Studies on Student Graduation Prediction in The Literature

Educational data mining (EDM) aims to analyze student performance and develop strategies to improve academic outcomes. One of the primary goals in this field is predicting students' graduation times and optimizing educational processes. Machine learning and ensemble methods are widely used for this purpose. Classification and regression algorithms are frequently employed in academic success prediction, and recent studies highlight the advantages of ensemble learning approaches in improving prediction accuracy.

Academic studies on on-time graduation prediction vary in terms of the methods used, datasets, and educational levels they target (undergraduate, master's, doctoral). A comparative summary of these studies is presented in Table 1.

Table 1. Comparative Overview of Recent Studies on On-Time Graduation Prediction Using Ensemble Methods.

Ref.	Ensemble Models	Other ML Models	Best Model	Dataset Content
[4]	RF, Bootstrap Aggregating, AdaBoost	DT, SMO, Random Tree	RF (83,74% accuracy)	Demographic attributes, CGPA, year of entry, and year of graduation
[5]	RF, AdaBoost	LR, DT, Gaussian, SVM, KNN, MLP	MLP (91.87% accuracy)	GPAs
[6]	AdaBoost, Bagging, RF, Rotation Forest	None	Rotation Forest (75.95% accuracy)	Grades
[7]	Stacking Ensemble, RF, AdaBoost	LDA, LR, CART, KNN, SVM, ANN, NB	Stacking Ensemble (LR, NB, KNN, NN) (74.82% accuracy)	Demographic attributes, first-year academic performance, financial status
[8]	AdaBoost	LR, DT, SVM	AdaBoost (82% F1-Score)	Demographic attributes, GPAs, etc.
[9]	Gradient Boosting, RF, Adaptive Boosting, XGBoost, LightGBM, Categorical Boosting	LR, LDA, NB, KNN, SVM, DT	LR with Ensemble-SMOTE (87.24% accuracy)	Student profiles, registration details, grades, and GPAs
[10]	CART with Bagging, CART with Boosting	CART with RF	Boosting with CART (87.755% accuracy)	GPAs, attendance, credit hours, length of study
[11]	Bagging, Boosting, Stacking	LR, ANN, DT, NB	Boosting with LR (86.82% accuracy)	Graduation data
[12]	RF	None	RF (91% accuracy)	GPAs, Graduation Status, etc.

Bako (2023) addressed the limited research on predicting postgraduate students' graduation time in Nigerian universities by proposing a model that applies the ADASYN technique to handle data imbalance and uses the Random Forests ensemble method for prediction. The results showed that balancing the dataset with ADASYN improved classification performance, and the proposed RF-based model achieved the highest prediction accuracy at 83.74% [4].

Rismayati et al. (2022) emphasized that graduating on time is a key goal for both students and higher education institutions, and individual student characteristics play a significant role in the duration of study. Several algorithms were tested, including LR, DT, Gaussian, RF, AdaBoost, SVM, KNN, and MLP Classifier. The results showed that the MLP Classifier achieved the highest accuracy at 91.87% [5].

Pandey and Taruna (2014) investigated the accuracy of ensemble learning techniques for early prediction of students' academic performance, aiming to support timely guidance and counseling. Focusing on a four-year engineering undergraduate program, the study employed five ensemble methods based on AdaBoost, Bagging, RF, and Rotation Forest algorithms. Among these, Rotation Forest achieved the highest performance with a model accuracy of 75.95%, emerging as the most effective method for predicting student performance in the early stages of the program [6].

Desfiandi and Soewito (2023) compared LR, SVM, DT, and ensemble learning methods such as AdaBoost to predict on-time graduation of Computer Science students. Using student data from 2015 to 2019, the study found that the AdaBoost decision tree model achieved the highest performance with an F1-score of 0.82. The results demonstrated that ensemble learning outperformed traditional models and proved to be an effective tool for educational prediction tasks [7].

Kang (2019) compared ten different machine learning algorithms to predict whether first-generation college students would graduate within six years, using data from their first year. The study evaluated methods such as LDA (linear discriminant analysis), LR, DT, SVM, AdaBoost, and stacking. Among these, the stacking ensemble model stood out as one of the top three performers, achieving an accuracy rate of 74.82% [8].

Law et al. (2024) addressed the issue of class imbalance in on-time graduation prediction by comparing various resampling strategies and proposing Ensemble-SMOTE, a method that combines different SMOTE variants. The findings showed that logistic regression with Ensemble-SMOTE achieved the highest performance between the 6th and 10th trimesters, with an accuracy of 87.24%, recall of 92.50%, F1-score of 91.30%, and F2-score of 92.02%. Statistical analyses confirmed that Ensemble-SMOTE was the most effective class imbalance treatment method based on performance metrics [9].

Ananto (2024) emphasized the importance of predicting student graduation status for monitoring academic progress and guiding educational interventions. The study examined the effectiveness of ensemble machine learning methods at Klabat University by combining algorithms such as Random Forest, Gradient Boosting, and Bagging. Trained on historical academic data, including GPA and attendance, the ensemble models outperformed individual classifiers in terms of accuracy, precision, and recall [10].

Lagman et al. (2020) highlighted that nearly half of first-time, full-time undergraduate students fail to complete their degrees, according to the National Center for Education Statistics. To address this issue, the study emphasized the importance of early prediction and identification of students at risk of delayed graduation, allowing institutions to implement appropriate retention and remediation strategies. The study demonstrated that using ensemble methods significantly improved classification accuracy, confirming their effectiveness in student graduation prediction tasks [11].

Rachmawati et al. (2023) emphasized that increasing the rate of on-time graduation is essential for maintaining the credibility and sustainability of universities, as it positively influences departmental accreditation. To address this, the study aimed to classify the accuracy of student graduation time using the Random Forest algorithm, based on data from Computer Science students at Dian Nuswantoro University between the academic years 2008–2017. The classification model developed with the Random Forest algorithm achieved an accuracy of 93% on training data and 91% on test data [12].

In general, the literature shows that ensemble methods such as AdaBoost and Random Forest are widely used, whereas more advanced approaches like Gradient Boosting, XGBoost, LightGBM, and CatBoost have been applied in relatively few studies [2-10]. This indicates that the potential of advanced ensemble learning techniques in predicting on-time graduation remains underexplored. Additionally, they highlight the need for developing advanced ensemble techniques to address challenges such as data imbalance and feature selection.

3. Method

In this study, ensemble learning techniques have been employed to predict whether undergraduate students at a public university would graduate on time. The applied methods include ensemble models such as Random Forest (RF), AdaBoost, Gradient Boosting, XGBoost, and CatBoost. Their performances have been compared with those of traditional machine learning algorithms,

including Logistic Regression (LR), Decision Trees (DT), Artificial Neural Networks (ANN), and Support Vector Machines (SVM).

3.1 Dataset

The dataset consists of records from undergraduate students enrolled at Sakarya University of Applied Sciences, a public university in Türkiye, and was obtained with institutional approval. Since the study does not involve any direct interaction with human participants—such as surveys, interviews, observations, experiments, or psychological tests—ethical committee approval was not required. The dataset includes demographic information (gender, age, and region of residence), academic performance in higher education (cumulative GPAs (grade point averages) and department), high school records, and results from the Higher Education Institutions Exam (YKS). The dataset consisted of 724 students, with 51.2% (371 students) graduating on time and 48.8% (353 students) not graduating on time. The features included in the dataset are presented in Table 2.

Table 2. Undergraduate degree dataset

Group	Feature	Description
Demographic (D)	Gender	Student's gender (Female/Male)
	Age	Student's age
	AddressRegion	Region of residence within the province
High School and University Entrance (H)	PreferenceOrder	Order of program preference
	PlacementRank	Ranking position of the selected program
	PlacementScore	The placement score for the selected program
	HighSchoolGPA	High school grade point average (GPA)
	HighSchoolTopStudentStatus	Top student status in high school (YES/NO)
	HighSchoolStatus	Educational status of the high school
	HighSchoolGroup	High school category group
Academic Performance (P)	Department	The program where the student studied
	FirstSemester,	Cumulative GPA after semesters
	SecondSemester	Cumulative GPA after second semester
	ThirdSemester	Cumulative GPA after third semester
	FourthSemester	Cumulative GPA after fourth semester
	FifthSemester	Cumulative GPA after fifth semester
	SixthSemester	Cumulative GPA after sixth semester
Output	GraduationStatus	Student's Graduation Outcome (GRADUATED/NOT GRADUATED)

3.2 Data Preprocessing

Data preprocessing comprises three main steps: data integration, data cleaning, and data transformation. In the first step, higher education academic performance data provided by the Student Information System (SIS) have been integrated with high school and Higher Education Institutions Exam (YKS) results, which are made available to universities through the web service of the Measuring, Selection and Placement Center (ÖSYM). During the second step, data cleaning, missing, erroneous, or inconsistent data that could negatively impact the analysis and modeling process are identified and either corrected or removed.
In the final step, categorical variables in the dataset are transformed into a suitable format, as machine learning models operate on numerical data. For this purpose, categorical variables are encoded into numeric values using the LabelEncoder method, which assigns a unique number to each category. To ensure fair evaluation and weighting across features with different scales, numerical variables are standardized using the mean and standard deviation, rescaling them to have a mean of 0 and a standard deviation of 1. After standardization, the Min-Max scaling method is applied to map values into a range between 0 and 1.

3.3 Data Splitting

To train and test the models, the dataset has been split into two subsets:

- 70% of the data is allocated to the training set, allowing the model to learn patterns and relationships within the data.
- 30% is designated as the test set, enabling performance evaluation on previously unseen data.

Since the distribution between graduates and non-graduates is balanced (51.2% and 48.8%, respectively), no additional resampling techniques such as SMOTE or undersampling were applied.

This approach is essential for assessing the model's ability to generalize to new data and for identifying potential issues such as overfitting or underfitting.

3.4 Prediction Timelines

Table 3 outlines the temporal availability of features used for on-time graduation prediction across seven distinct time points (T0 to T6). At all stages, static features—such as demographic data (D), high school academic performance, higher education entrance exam results (H), and department information (P)—are consistently available and serve as foundational predictors. Starting from T1 (end of the first semester), cumulative GPA scores are incrementally added as features, reflecting students' academic progress over time.

Table 3. Prediction timelines

Feature	T0	T1	T2	T3	T4	T5	T6
Demographic (D), high school academic performance and higher education entrance exam results (H) and Department (P)	✓	✓	✓	✓	✓	✓	✓
Cumulative GPA after 1st Semester		✓	✓	✓	✓	✓	✓
Cumulative GPA after 2nd Semester			✓	✓	✓	✓	✓
Cumulative GPA after 3rd Semester				✓	✓	✓	✓
Cumulative GPA after 4th Semester					✓	✓	✓
Cumulative GPA after 5th Semester						✓	✓
Cumulative GPA after 6th Semester							✓

3.5 Models

Both ensemble learning and traditional machine learning methods have been used for the binary classification task of predicting graduation status.

Ensemble Learning Methods:

- AdaBoost (Adaptive Boosting): AdaBoost is an ensemble learning technique based on the principle of combining weak learners to create a stronger classifier. It is widely used in classification problems and belongs to the family of boosting algorithms [13].

- CatBoost: CatBoost is a boosting algorithm specifically designed for categorical data. It can handle categorical features automatically and is commonly applied in classification, regression, and time-series tasks due to its efficiency and accuracy [14].

- Gradient Boosting: This method builds a model iteratively by training weak learners sequentially and minimizing errors through loss function optimization. It is effective for both classification and regression problems [15].

- XGBoost (Extreme Gradient Boosting): XGBoost is an enhanced version of gradient boosting, offering superior computational speed and scalability. It is widely utilized in classification, regression, and ranking applications, particularly with large datasets [16].

- Random Forest (RF): Random Forest is a robust ensemble technique composed of multiple decision trees. It improves generalization and reduces overfitting by training on different subsets of the data, making it suitable for both classification and regression tasks [17].

Traditional Machine Learning Methods:

- Logistic Regression (LR): Logistic Regression is a supervised learning algorithm commonly used in binary classification. It applies the logistic function to estimate the probability of a binary outcome and is widely used for modeling event likelihood [18].

- Artificial Neural Networks (ANN): ANN are multi-layered computational networks inspired by biological neural systems. They perform well on large and complex datasets, making them suitable for both classification and regression [19].

- Support Vector Machines (SVM): SVM is a supervised learning model used for both classification and regression tasks. Its core

objective is to find the optimal hyperplane that separates classes. SVM is particularly effective in high-dimensional feature spaces [20].

- Decision Trees (DT): DT is a tree-based model that recursively splits the data for decision-making. It is simple to interpret and performs well on small datasets, although pruning may be necessary to prevent overfitting [21].

The parameters of the models used are presented in the Table 4.

Table 4. The parameters of the models

Model	Important Parameters
AdaBoost	learning_rate=0.1, n_estimators=1000, algorithm='SAMME.R', random_state=42
CatBoost	learning_rate=0.1, iterations=100, depth=3, random_seed=42, verbose=0
Gradient Boost	learning_rate=0.1, n_estimators=100, max_depth=3, random_state=42
XGBoost	learning_rate=0.1, n_estimators=100, max_depth=3, random_state=42
RF	criterion='gini', n_estimators=100, random_state=42
LR	C=1.0, penalty='l2', solver='lbfgs', random_state=42
ANN	hidden_layer_sizes=(100,), learning_rate_init=0.001, activation='relu', batch_size='auto', solver='adam', random_state=42
SVM	C=1.0, kernel='rbf', gamma='scale', degree=3, random_state=42
DT	criterion='gini', splitter='best', random_state=42

The ensemble and traditional machine learning models have been trained and evaluated on a system featuring a 13th Gen Intel® Core™ i5-13500H processor (2.60 GHz), 16 GB of RAM, an NVIDIA GeForce RTX 4050 GPU, and a 64-bit architecture, running Windows 11 IoT Enterprise LTSC. The models have been implemented in Python within the Anaconda Navigator environment, utilizing libraries such as Pandas, Keras, and NumPy. The training phase has been conducted using a designated training dataset, and model performance has subsequently been assessed on a separate test dataset. Only CPU resources have been employed throughout the process.

3.6 Model Evaluation

To assess the performance of the models, various classification metrics are used:

- Accuracy: This metric determines the proportion of correctly classified instances out of all instances in the dataset. It is calculated as:
  Accuracy = (TP + TN) / (TP + TN + FP + FN)

- Precision: Precision measures how many of the instances predicted as positive are actually positive. It is defined as:
  Precision = TP / (TP + FP)

- Recall (True Positive Rate or Sensitivity): Recall indicates the percentage of actual positive cases that the model correctly identifies. It is computed as:
  Recall = True Positive Rate (TPR) = TP / (TP + FN)

- F1-Score: The F1-score is the harmonic mean of precision and recall, balancing both metrics. It is given by the formula:
  F1-Score = (2 * Recall * Precision) / (Recall + Precision)

4. Results and Discussion

In the study, the effectiveness of various machine learning models, including traditional methods and ensemble learning techniques, was evaluated using binary classification for on-time graduation prediction. The dataset consisted of 724 students, with 51% (371 students) graduating on time and 49% (353 students) not graduating on time. Model performance was assessed using common classification metrics such as accuracy, recall, precision, and F1 score. Additionally, the analyses were conducted across seven different time points (T0-T6), reflecting the incremental accumulation of academic data over time.

4.1 Accuracy Analysis

Table 5 and Figure 1 present the accuracy results of different models across prediction periods T0 to T6. In the early prediction periods, ensemble methods generally achieved higher accuracy rates. Specifically, Random Forest (RF) and CatBoost showed accuracies of 60.1% and 58.7%, respectively, in T0. The success of ensemble methods in early predictions can be attributed to their strong feature selection and ability to generalize well over data.

The low accuracy observed at T0 is attributed to the unavailability of academic performance data from higher education at that stage. In subsequent periods (T1–T6), a significant improvement in prediction performance was observed as cumulative semester GPA scores were incrementally incorporated into the model. As the prediction periods progressed (T1-T6), an observable increase in accuracy was seen across all models. Notably, CatBoost, XGBoost, and Random Forest reached high accuracy values of 83.0%, 82.1%, and 81.7%, respectively, in T6. This demonstrates that ensemble methods tend to perform better in the long run,

showcasing their higher capacity for generalization. On the other hand, classical machine learning methods also showed improvement over time. Models like LR and SVM started with low accuracy in T0 but reached 83.5% and 81.2% accuracy in T6, respectively. However, this improvement was more limited compared to ensemble methods.

Table 5. Accuracy results

Models	T0	T1	T2	T3	T4	T5	T6
RF	**60.1**	70.2	73.4	76.1	78.9	79.8	81.7
CatBoost	58.7	72.0	72.0	74.3	79.4	79.8	83.0
XGBoost	58.3	68.8	73.4	77.5	75.7	**80.3**	82.1
AdaBoost	56.0	**73.4**	**74.3**	**78.4**	**80.3**	79.4	79.8
GradientBoost	56.0	69.7	**74.3**	72.9	75.7	74.8	81.2
Ensemble (Avg)	**57.8**	**70.8**	**73.5**	**75.8**	**78.0**	**78.8**	**81.6**
DT	56.0	61.5	67.0	65.6	69.7	72.5	74.3
LR	54.1	66.5	72.5	74.3	76.1	**80.3**	**83.5**
SVM	53.7	65.1	67.0	71.1	75.7	78.0	81.2
ANN	51.4	60.6	65.1	72.0	71.1	72.0	78.9
Traditional (Avg)	**53.8**	**63.4**	**67.9**	**70.8**	**73.2**	**75.7**	**79.5**

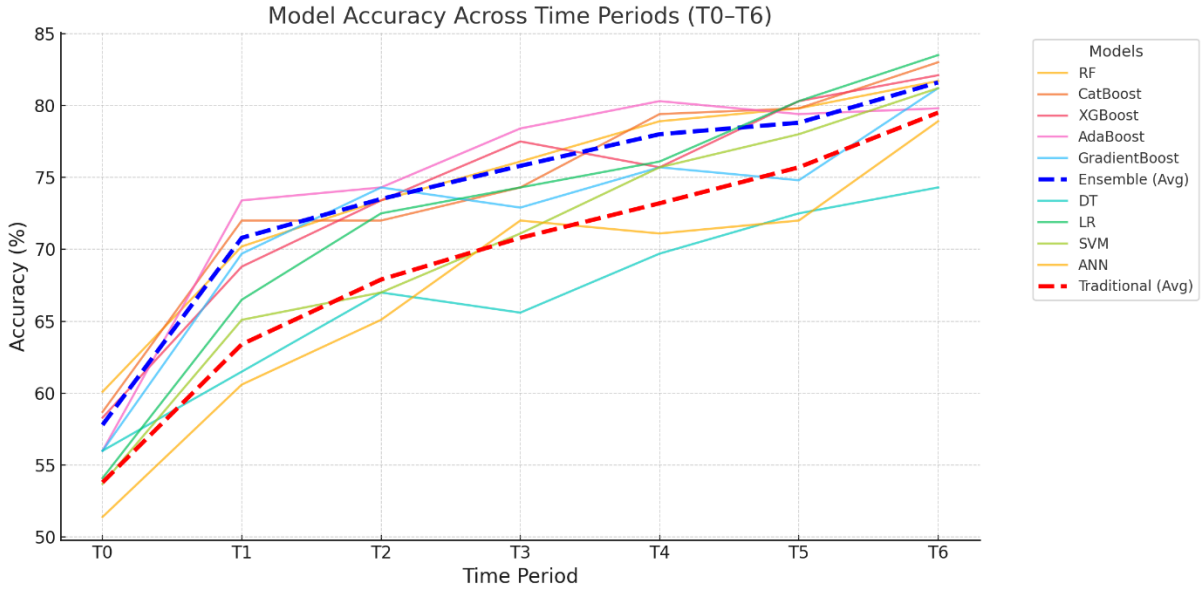

Figure 1. Accuracy results

4.2 F1-Score Comparison

Table 6 and Figure 2 present the F1-Score results across different models and prediction periods from T0 to T6. In the early prediction periods, ensemble methods again show superior performance compared to classical machine learning methods. For instance, Random Forest (RF) and CatBoost exhibit F1-Scores of 62.0% and 61.2%, respectively, in T0.

As the prediction periods progress, all models show a general increase in F1-Score. Notably, the highest F1-Scores are reached by the ensemble methods. CatBoost achieves the highest score of 84.8% in T6, closely followed by XGBoost at 84.2% and Random Forest at 83.2%. These results highlight that ensemble models continue to outperform their classical counterparts throughout the prediction periods, particularly as the models have more data and time to refine their predictions. On the other hand, classical models such as LR and SVM also show improvements in F1-Score over time, with LR achieving 84.7% in T6 and SVM reaching 83.0%.

Table 6. F1-Score results

Models	T0	T1	T2	T3	T4	T5	T6
RF	**62.0**	73.0	75.8	78.2	80.3	82.0	83.2
CatBoost	61.2	75.1	74.5	76.1	81.5	81.8	**84.8**
XGBoost	61.3	71.7	75.8	79.1	78.0	**82.4**	84.2
AdaBoost	59.7	76.8	**76.7**	**81.3**	**82.4**	81.2	81.8
GradientBoost	60.3	**73.2**	76.1	74.2	77.1	77.2	82.7
Ensemble (Avg)	**60.9**	**74.0**	**75.8**	**77.8**	**79.9**	**80.9**	**83.3**
DT	57.9	64.7	68.1	67.2	69.7	73.7	76.1
LR	56.1	69.2	74.6	75.9	77.6	82.2	84.7
SVM	53.9	67.5	69.2	73.6	78.0	80.2	83.0
ANN	53.5	61.9	66.4	73.8	71.7	73.4	80.8
Traditional (Avg)	**55.4**	**65.8**	**69.6**	**72.6**	**74.3**	**77.4**	**81.2**

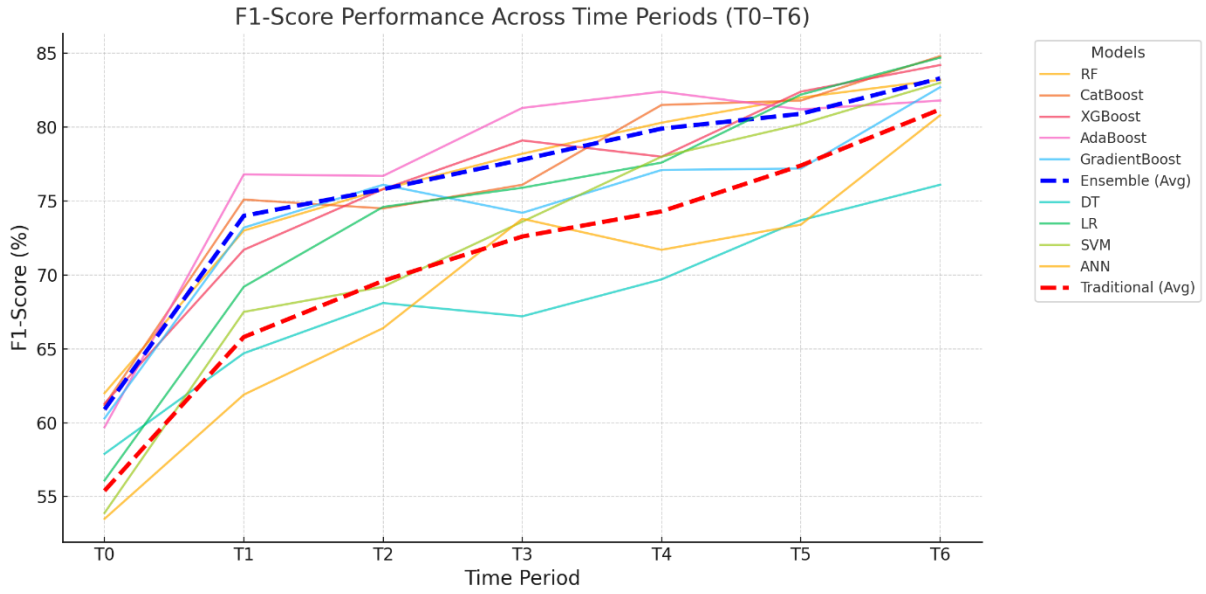

Figure 2. F1-Score results

4.3 Precision Performance

Table 7 and Figure 3 present the Precision results for the different models across prediction periods T0 to T6. As observed in the previous metrics, ensemble methods tend to perform better than classical machine learning models, especially in the early prediction periods. For example, Random Forest (RF) and CatBoost achieve Precision scores of 65.7% and 64.0%, respectively, in T0.

As the prediction periods progress, the Precision scores for all models generally increase, with ensemble methods consistently leading the way. In T6, Logistic Regression (LR) achieves the highest Precision score of 87.0%, surpassing all other models. Other ensemble methods such as RF and CatBoost also show strong performances, with Precision scores of 84.6% and 84.4%, respectively. Gradient Boosting also demonstrates strong performance, reaching 84.5% in T6, highlighting the efficacy of these models in maintaining high Precision. In contrast, classical methods like DT, SVM, and ANN show slower improvements and still fall behind ensemble methods by the end of the evaluation period.

Table 7. Precision results

Models	T0	T1	T2	T3	T4	T5	T6
RF	**65.7**	73.3	76.5	79.5	**83.2**	81.3	84.6
CatBoost	64.0	74.2	75.4	78.8	81.1	81.8	84.4
XGBoost	63.2	72.3	76.5	**81.6**	78.3	81.5	82.5
AdaBoost	60.7	**74.4**	77.3	78.5	81.5	82.2	81.8
GradientBoost	60.3	72.0	**78.8**	78.7	80.9	77.5	84.5
Ensemble (Avg)	**62.8**	**73.2**	**76.9**	**79.4**	**81.0**	**80.9**	**83.6**
DT	61.7	65.8	73.3	71.3	78.4	78.5	78.8
LR	59.8	70.7	76.5	79.3	81.1	**82.5**	**87.0**
SVM	60.2	69.9	71.7	74.6	78.3	80.2	83.3
ANN	57.0	66.7	71.4	76.8	78.4	77.8	81.5
Traditional (Avg)	**59.7**	**68.3**	**73.2**	**75.5**	**79.1**	**79.8**	**82.7**

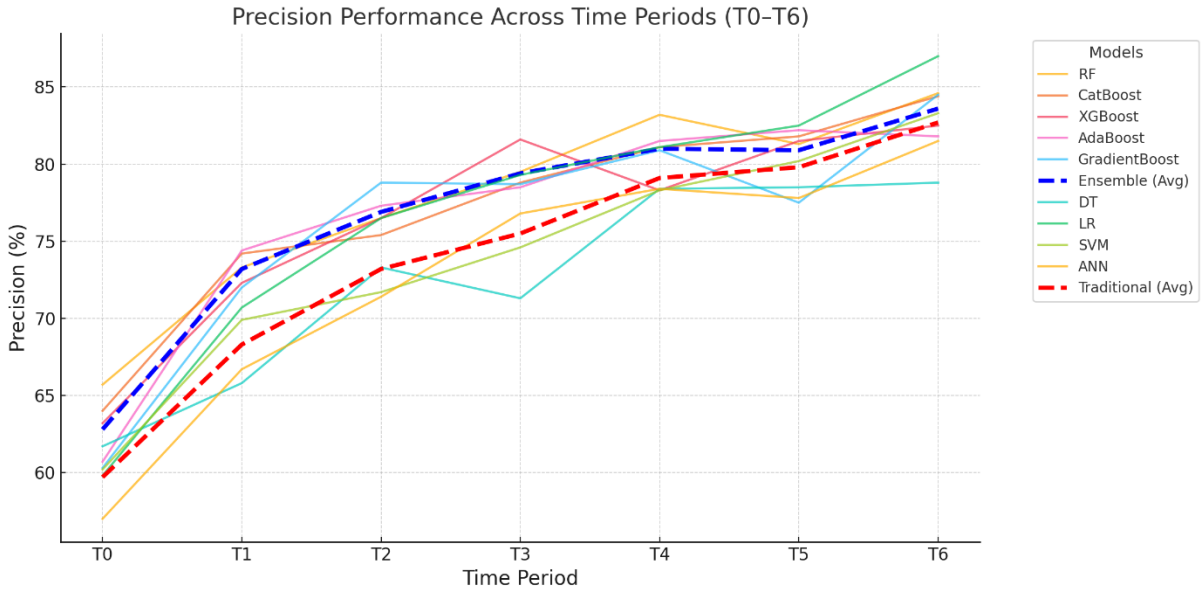

Figure 3. Precision results

4.4 Recall Analysis

Table 8 and Figure 4 present the Recall results across different models and prediction periods from T0 to T6. In the early prediction periods, GradientBoosting, XGBoost, and AdaBoost exhibit relatively strong performance. For instance, GradientBoosting achieves a Recall of 60.3% in T0, while XGBoost and AdaBoost have Recall values of 59.5% and 58.7%, respectively.

As the prediction periods progress, XGBoost shows the most notable improvement, achieving the highest Recall value of 86.0% in T6. CatBoost also demonstrates strong performance, reaching 85.1% in T6, while AdaBoost and GradientBoosting follow closely with Recall values of 81.8% and 81.0%, respectively. In contrast, classical methods such as LR and SVM show steady improvements but do not match the performance of the top ensemble models. LR reaches 82.6% in T6, while SVM achieves 82.6%, demonstrating solid but slower progress over time.

Table 8. Recall results

Models	T0	T1	T2	T3	T4	T5	T6
RF	58.7	72.7	75.2	76.9	77.7	82.6	81.8
CatBoost	58.7	76.0	73.6	73.6	81.8	81.8	85.1
XGBoost	**59.5**	71.1	75.2	76.9	77.7	**83.5**	**86.0**
AdaBoost	58.7	**79.3**	**76.0**	**84.3**	**83.5**	80.2	81.8
GradientBoost	60.3	74.4	73.6	70.2	73.6	76.9	81.0
Ensemble (Avg)	**59.2**	**74.7**	**74.7**	**76.4**	**78.9**	**81.0**	**83.1**
DT	54.5	63.6	63.6	63.6	62.8	69.4	73.6
LR	52.9	67.8	72.7	72.7	74.4	81.8	82.6
SVM	48.8	65.3	66.9	72.7	77.7	80.2	82.6
ANN	50.4	57.9	62.0	71.1	66.1	69.4	80.2
Traditional (Avg)	**51.7**	**63.7**	**66.3**	**70.0**	**70.3**	**75.2**	**79.8**

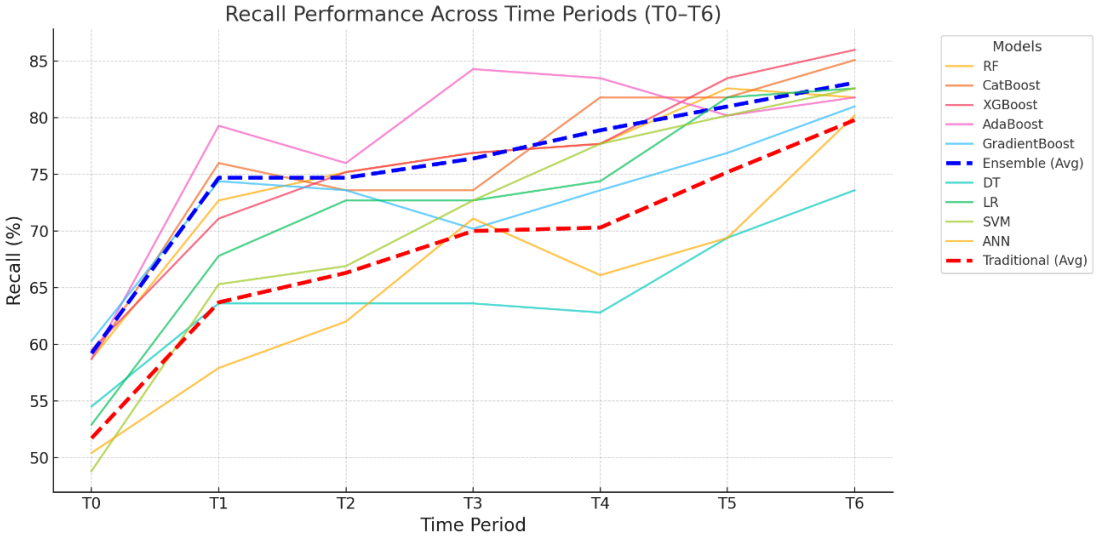

Figure 4. Recall results

4.5 Feature Importance Analysis

Feature importance scores were computed separately for each of the four ensemble models used in this study: Random Forest, XGBoost, AdaBoost, and Gradient Boosting. The built-in feature_importances_ attribute of each model was utilized to extract variable contributions, and the top-ranked features were compared across models. To highlight the general pattern of agreement, the average importance score across models was also calculated and visualized (see Figure 5).

According to the results, the cumulative GPA after the 6th semester consistently emerged as the most influential variable in all models. This finding indicates that students' academic performance in the later stages of their undergraduate education plays a critical role in predicting on-time graduation. The 4th and 3rd semester GPA values also demonstrated high importance, reinforcing the conclusion that academic progress in advanced semesters strongly correlates with successful and timely degree completion. Although model-specific variations exist, the "Department" variable exhibited relatively high importance, particularly in the XGBoost model. In contrast, pre-university variables such as high school GPA, placement score, and preference order showed lower importance scores across all models. This suggests that while these factors may provide some insight during early prediction stages, their influence diminishes in the presence of university-level academic performance data.

In Figure 5, the average importance score is marked by a black line, emphasizing the features with consistently high contribution across models. The results confirm that dynamic in-university academic indicators, especially cumulative GPA, are more informative and reliable for predicting on-time graduation. These findings offer valuable implications for the development of early warning systems and data-driven academic advising tools in higher education.

Celal Bayar University Journal of Science
Volume 21, Issue 4, 2025, p 79-91
Doi: 10.18466/cbayarfbe.1644296

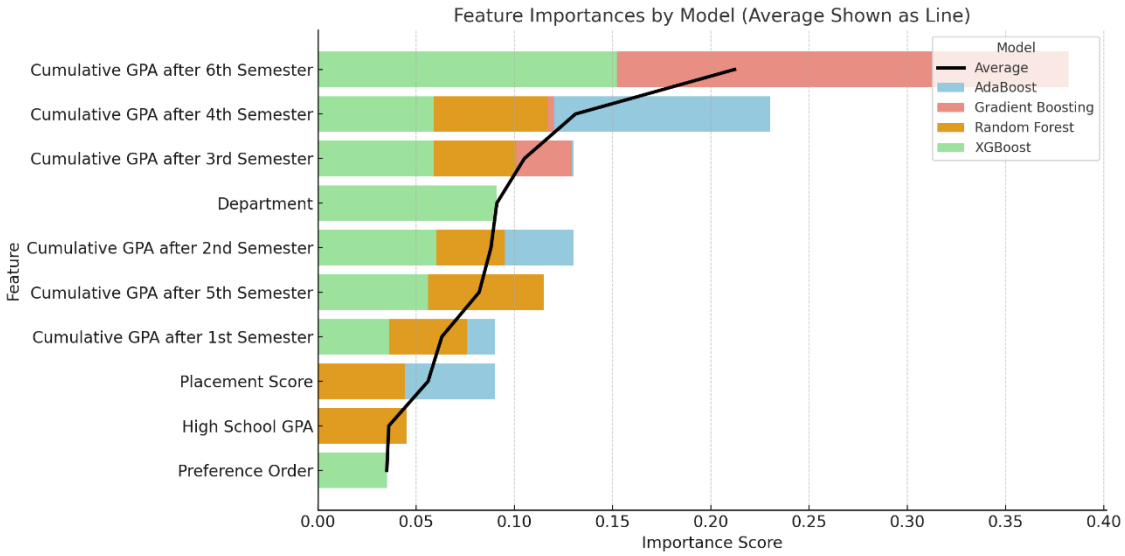

Figure 5. Feature importance of models

4.6 Comparison of Traditional and Ensemble Methods

The experimental results highlight significant differences between traditional machine learning methods and ensemble learning techniques in predicting on-time graduation. The performance of both categories of models was evaluated across multiple prediction periods (T0-T6) using accuracy, F1-score, precision, and recall metrics. The findings provide key insights into the advantages and limitations of each approach.

4.6.1 Early Prediction Performance

In the early prediction period (T0), ensemble methods consistently outperformed traditional machine learning models. Specifically, Random Forest (RF) and CatBoost achieved higher accuracy (60.1% and 58.7%, respectively) compared to traditional models such as Logistic Regression (54.1%) and Support Vector Machine (SVM) (53.7%). A similar trend was observed in other evaluation metrics, with ensemble methods demonstrating superior F1-score, precision, and recall in the initial stages. This suggests that ensemble models, leveraging multiple decision trees and boosting mechanisms, are better suited for early-stage predictions where limited data is available (see Figure 6).

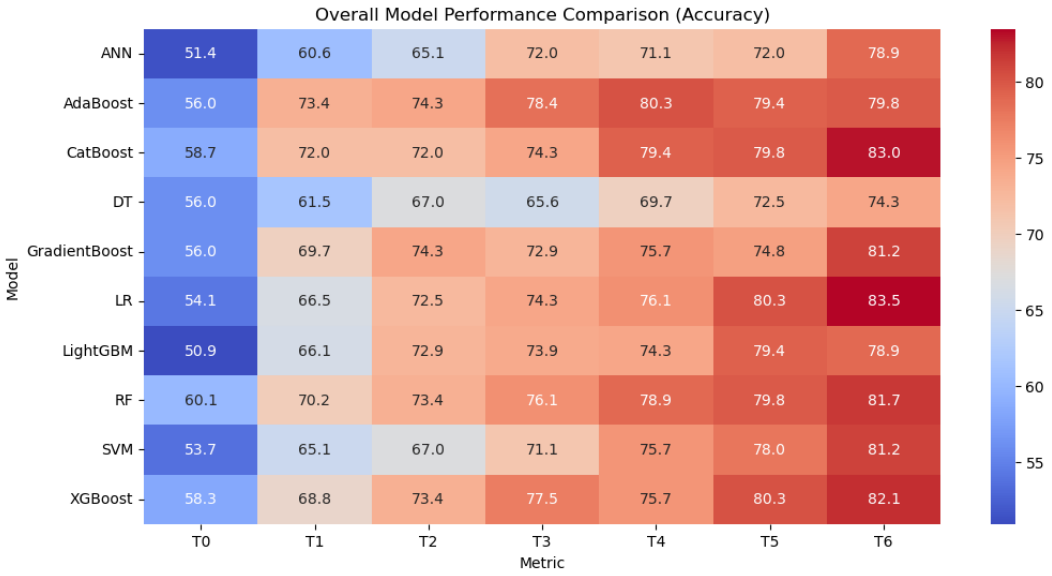

Figure 6. Overall model performance comparison (Accuracy)

4.6.2 Performance Progression Over Time

As the prediction periods progressed (T1-T6), all models showed improved performance, reflecting the benefit of increased data accumulation. However, ensemble methods maintained their superiority in most cases. For instance, by T6, CatBoost, XGBoost, and Random Forest achieved accuracy rates of 83.0%, 82.1%, and 81.7%, respectively, while traditional models such as Logistic Regression and SVM attained 83.5% and 81.2%. Although the gap between traditional and ensemble methods narrowed over time, ensemble techniques consistently demonstrated higher generalization ability, particularly in handling complex patterns in the dataset.

4.6.3 Stability Across Metrics

The comparison of F1-score, precision, and recall further reinforced the advantages of ensemble methods. By T6, CatBoost and XGBoost achieved the highest F1-scores (84.8% and 84.2%, respectively), surpassing traditional models like Logistic Regression (84.7%) and SVM (83.0%). Precision followed a similar trend, with Logistic Regression (87.0%) being the only traditional model that rivaled ensemble techniques such as Random Forest (84.6%) and CatBoost (84.4%). In terms of recall, XGBoost achieved the highest score of 86.0%, while CatBoost and Logistic Regression followed closely at 85.1% and 82.6%, respectively. These results indicate that ensemble methods not only provide robust predictive capabilities but also maintain a stable balance between precision and recall, ensuring better classification of both positive and negative cases.

4.6.4 Statistical Comparison of Ensemble and Traditional Models

To determine whether the performance differences between ensemble methods and traditional machine learning models are statistically significant, paired t-tests and Wilcoxon signed-rank tests were conducted on the average accuracy and F1-score values across the prediction periods T0 to T6.

The results are as follows:

- For accuracy, the p-value from the t-test was 0.0004, and from the Wilcoxon test was 0.0156.
- For F1-score, the p-value from the t-test was 0.0004, and from the Wilcoxon test was 0.0156.

Since all p-values are below the 0.05 significance threshold, it can be concluded that ensemble methods demonstrate statistically significant superiority over traditional models in terms of both accuracy and F1-score.

5. Conclusion

The findings of this study demonstrate that ensemble learning methods provide a significant performance advantage over traditional machine learning algorithms in predicting on-time graduation. Across all evaluation metrics—accuracy, F1-score, precision, and recall—ensemble models consistently outperformed classical approaches, particularly during the early stages of prediction (T0). Algorithms such as Random Forest (RF), CatBoost, and XGBoost achieved strong predictive performance, reflecting their enhanced generalization capabilities and effective feature selection mechanisms.

In particular, RF and CatBoost yielded high accuracy scores even in the initial prediction periods and maintained this superiority as additional academic data accumulated. Likewise, these models exhibited consistently high precision and F1-scores, indicating a strong balance between sensitivity and specificity. XGBoost and CatBoost also stood out in terms of recall, suggesting their effectiveness in identifying students likely to graduate on time—a critical factor for early intervention strategies.

Although traditional methods such as Logistic Regression (LR) and Support Vector Machines (SVM) showed improvements over time, especially in later periods, they generally failed to match the predictive performance of ensemble models. While LR performed reasonably well in precision and F1-score, it lagged behind ensemble techniques in recall, limiting its ability to capture positive graduation cases effectively.

Overall, this study reaffirms the value of ensemble models for graduation prediction tasks. These models are capable of producing reliable forecasts both in early and later academic stages. Moreover, feature importance analysis revealed that cumulative academic performance throughout university education is the most decisive factor influencing graduation outcomes. This insight suggests that advanced ensemble learning frameworks can be effectively leveraged to develop early warning systems and data-driven academic advising infrastructures in higher education institutions.

Future research could focus on integrating deep learning architectures or hybrid ensemble approaches to further enhance predictive performance. Additionally, evaluating the generalizability of ensemble models across different institutional contexts would offer valuable insights into their scalability and robustness.

The models developed in this study can be operationalized as service-based tools or API components within university Student Information Systems (SIS). Once deployed, they can autonomously analyze student data at predefined academic checkpoints, identify individuals at risk of delayed graduation, and

provide actionable outputs for academic advisors and administrative staff. Such integration enhances the practical utility of the models by supporting timely, data-informed decision-making processes in academic planning and student support.

Acknowledgement

There are no acknowledgements for this manuscript.

Author's Contributions

The authors contributed equally to the study.

Ethics

There are no ethical issues after the publication of this manuscript.

References

[1]. Mienye, I. D. and Sun, Y. (2022). A survey of ensemble learning: concepts, algorithms, applications, and prospects. *IEEE Access*; 10, 99129-99149. https://doi.org/10.1109/access.2022.3207287

[2]. Sagi, O. and Rokach, L. (2018). Ensemble learning: a survey. *WIREs Data Mining and Knowledge Discovery;* 8(4). https://doi.org/10.1002/widm.1249

[3]. Ataş, P.K. (2024). Evaluate student achievement by classifying brain structure and its functionality with novel hybrid method. *Neural Computing and Applications;* 36, 3357–3368. https://doi.org/10.1007/s00521-023-09031-9

[4]. Bako, H. 2023. Predicting timely graduation of postgraduate students using random forests ensemble method. *Fudma Journal of Sciences*; 7(3): 177-185. https://doi.org/10.33003/fjs-2023-0703-1773

[5]. Rismayati, R., Ismarmiaty, I., Hidayat, S. (2022). Ensemble Implementation for Predicting Student Graduation with Classification Algorithm. *International Journal of Engineering and Computer Science Applications (IJECSA);* 1(1), 35–42. https://doi.org/10.30812/ijecsa.v1i1.1805

[6]. Pandey, M. and Taruna, S. 2014. A comparative study of ensemble methods for students' performance modeling. *International Journal of Computer Applications*; 103(8): 26-32. https://doi.org/10.5120/18095-9151

[7]. Kang, Z. 2019. Using machine learning algorithms to predict first-generation college students' six-year graduation: a case study. *International Journal of Information Technology and Computer Science*;11(9): 1-8. https://doi.org/10.5815/ijitcs.2019.09.01

[8]. Desfiandi, A. and Soewito, B. 2023. Student graduation time prediction using logistic regression, decision tree, support vector machine, and adaboost ensemble learning. *IJISCS (International Journal of Information System and Computer Science);* 7(3): 195. https://doi.org/10.56327/ijiscs.v7i2.1579

[9]. Law, T.-J., Ting, C.-Y., Ng, H., Goh, H.-N., Quek, A. (2024). Ensemble-SMOTE: Mitigating Class Imbalance in Graduate on Time Detection. *Journal of Informatics and Web Engineering;* 3(2), 229–250. https://doi.org/10.33093/jiwe.2024.3.2.17

[10]. Ananto, N. (2024). Leveraging ensemble learning for predicting student graduation: a data mining approach. *Prosiding Seminar & Conference FMI*; 2: 353-365. https://doi.org/10.47747/snfmi.v2i1.2320

[11]. Lagman, A., Alfonso, L., Goh, M., Lalata, J., Magcuyao, J., & Vicente, H. 2020. Classification algorithm accuracy improvement for student graduation prediction using ensemble model. *International Journal of Information and Education Technology*; 10(10): 723-727. https://doi.org/10.18178/ijiet.2020.10.10.1449

[12]. Rachmawati, D. A., Ibadurrahman, N. A., Zeniarja, J., Hendriyanto, N. (2023). Implementation of the Random Forest Algorithm in Classifying the Accuracy of Graduation Time for Computer Engineering Students at Dian Nuswantoro University. *Jurnal Teknik Informatika (Jutif);* 4(3): 565–572. https://doi.org/10.52436/1.jutif.2023.4.3.920

[13]. Freund, Y., Schapire, R. E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*; 55(1), 119–139.

[14]. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., Gulin, A. (2017). CatBoost: unbiased boosting with categorical features. https://doi.org/10.48550/ARXIV.1706.09516

[15]. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. The Annals of Statistics, 29(5), 1189–1232.

[16]. Chen, T., Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD *International Conference on Knowledge Discovery and Data Mining. ACM;* San Francisco California USA, pp 785–794

[17]. Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32.

[18]. Hosmer, D. W., Lemeshow, S. (2000). Applied Logistic Regression, 1st ed, Wiley, New York

[19]. Goodfellow, I., Bengio, Y., Courville, A. (2016). Deep learning, The MIT press, Cambridge

[20]. Cortes, C., Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273–297.

[21]. Breiman, L., Friedman, J. H., Olshen, R. A., Stone, C. J. (2017). Classification And Regression Trees, 1st ed, Routledge, New York