Electronic Integrated Computer Algorithm Journal
Vol. 4, No. 1, October 2026, pp 1~7

ISSN: 3031-0350, DOI: 10.62123/enigma.v4i1.158

Comparison of Machine Learning Methods for Student Graduation Prediction

Nina Padmasari [1*], Aisyah Rahmani Putri [1]

[1] *Information Systems, Faculty of Computer Science, Universitas Indo Global Mandiri*, Indonesia.

DOI : 10.62123/enigma.v4i1.158

Received : April 22, 2026
Revised : July 07, 2026
Accepted : August 08, 2026

Keywords:
Machine Learning, Student Graduation, Random Forest, Decision Tree, Naïve Bayes

ABSTRACT

Student graduation prediction has become an important topic in higher education because it supports academic quality improvement, institutional performance evaluation, and early intervention for students at risk of delayed graduation. This study aims to compare the performance of machine learning methods used for student graduation prediction through a literature review approach. The literature was collected from Google Scholar, focusing on studies published between 2018 and 2025. Fourteen relevant articles were reviewed, consisting of twelve empirical studies included in the comparative analysis and two review articles used as supporting references. The main algorithms discussed are Random Forest, Decision Tree (C4.5), and Naïve Bayes, with additional comparisons involving K-Nearest Neighbor (K-NN), Support Vector Machine (SVM), Logistic Regression, and Neural Network. The findings indicate that Random Forest generally demonstrates stable predictive performance across many studies, while Decision Tree provides better model interpretability and Naïve Bayes offers computational efficiency. However, differences in dataset characteristics, feature selection, preprocessing techniques, and evaluation strategies significantly influence the reported performance of each algorithm. Therefore, no single machine learning method can be considered universally superior for student graduation prediction. The review also identifies GPA, completed credits, attendance, and active semesters as the most influential factors in predicting student graduation.

1. INTRODUCTION

1.1 Background

On-time student graduation is a key indicator of higher education success. Low graduation rates can reflect issues with the learning process, curriculum, academic management, or other external factors affecting students. Therefore, many educational institutions are beginning to utilize machine learning and data mining to build predictive systems that can identify at-risk students early on [1][2]. In a broader context, the systematic review indicates that predicting academic performance has become a significant focus in Educational Data Mining (EDM). The review examined 17 survey papers and 74 research papers and demonstrated that the increasing volume of educational data is driving the use of machine learning methods to aid student monitoring, early intervention, and more informed academic decision-making [3].

Various studies show that student graduation predictions generally use academic attributes such as GPA, GPA, number of credits, attendance, and active semesters. Some studies have expanded this to include demographic factors, social factors, LMS data, and other external variables. However, the best algorithms are not always consistent across studies. Some studies rank Random Forest as the superior method, others find Decision Tree/C4.5 to be superior, and still others show Naïve Bayes, K-NN, or even Logistic Regression to be more competitive in certain contexts [1][2][4][5][6]. Because previous research results are still mixed, a study is needed that not only summarizes accuracy but also compares the implementation, effectiveness, advantages, and limitations of each method. Therefore, this study is structured as a literature review to compare the three most relevant and frequently occurring methods in the files you uploaded: Random Forest, Decision Tree/C4.5, and Naïve Bayes, and then place them in the context of other algorithm findings to strengthen the analysis [1][3].

1.2 Problem Formulation

1. What machine learning methods are most commonly used for student graduation prediction?
2. How are Random Forest, Decision Tree, and Naïve Bayes implemented in previous studies?
3. Which method is the most effective based on evaluation results?
4. What factors most influence student graduation?

1.3 Problem Formulation

1. Identify commonly used machine learning algorithms for graduation prediction.
2. Analyze implementations of Random Forest, Decision Tree, and Naïve Bayes.

*Corresponding Author Email: 2024210005@uigm.students.ac.id

3. Compare their effectiveness based on evaluation metrics.
4. Provide recommendations for the most suitable method.

1.4 Research Contribution

This study synthesizes findings from 14 selected articles. Twelve empirical studies are comparatively analyzed in Table 1, while two review papers are used as supporting references to strengthen the theoretical background. It emphasizes that selecting the best method should not rely solely on accuracy but also consider model stability, interpretability, and implementation needs. [1], [3]

2. LITERATURE REVIEW

2.1 Educational Data Mining and Graduation Prediction

EDM is an interdisciplinary field that combines machine learning, statistics, data mining, educational psychology, and computer science to extract patterns from educational data. The goal is not only to predict student performance but also to support curriculum development, academic monitoring, and more personalized interventions. In EDM, the general process includes problem definition, data collection, preprocessing, technique selection, model training, evaluation, and deployment [3].

Based on the reviewed literature, graduation prediction is usually formulated as a binary classification problem: on-time or late. Setiono's research, for example, categorizes graduations into these two categories using attributes such as GPA, credits, attendance, and active semesters [2].

2.2 Random Forest

Random Forest is an ensemble algorithm based on a collection of decision trees designed to improve prediction accuracy and reduce overfitting. In a study by Afandi et al., Random Forest achieved 97.50% accuracy, 98% precision, 97% recall, and 97% F1-score, outperforming Decision Tree and Naïve Bayes on academic data from Informatics students [1]. In other studies, Random Forest has also demonstrated strong performance in various educational contexts. A study by Andrianof et al. reported 87.5% accuracy, 86.3% precision, and 85.9% recall for predicting student graduation based on academic data, with GPA and attendance as the most important factors [7]. Meanwhile, a study on student graduation potential from baseline secondary education data showed Random Forest achieved 89% accuracy and an AUC of 0.94, and was deemed effective for classifying graduation potential based on baseline data [8].

Furthermore, a study by Zeniarja et al. showed that in feature selection scenarios and comparisons of several algorithms, Random Forest still produced the highest accuracy of 77.35% compared to Naïve Bayes, Decision Tree, K-NN, and SVM. Interestingly, the study also found that feature selection did not always improve accuracy, as the combination of complete attributes actually produced the best results [9].

2.3 Decision Tree (C4.5)

Decision Trees are a popular classification method because their models are easily interpreted in the form of decision rules. This advantage is crucial in predicting student graduation rates, as academics often require models that are explainable, not simply accurate. In Setiono's research, Decision Tree/C4.5 achieved 80% accuracy with 10-Fold Cross-Validation and was successfully implemented as an interactive web application for an early warning system. The research also showed that attendance was the most significant factor, followed by GPA and credit progress [2].

In another study, the C4.5 algorithm was reported to achieve 99.64% accuracy and 99.5% AUC for predicting student graduation rates [10]. Furthermore, Kurniawati's study comparing C4.5, K-NN, ID3, Logistic Regression, and Naïve Bayes found that C4.5 produced the highest accuracy of 94.33%, while K-NN achieved the lowest at 50.93% [4]. These highly variable results indicate that Decision Tree/C4.5 can perform very well, but is also very sensitive to data quality, feature characteristics, and evaluation strategy.

2.4 Naïve Bayes

Naïve Bayes is a simple and fast probabilistic algorithm that assumes independence between features. Its main advantages are computational efficiency and ease of implementation. In a study by Afandi et al., Naïve Bayes achieved an accuracy of 86.25%, lower than Decision Tree and Random Forest, but still exhibiting a fairly high precision of 93% [1].

On the other hand, some studies have shown Naïve Bayes to be a very powerful method. A study by Meiriza et al. reported an accuracy of 97.6378% for predicting student graduation using the Naïve Bayes Classifier on data from Sriwijaya University Information Systems students [11]. Widaningsih's research also showed that Naïve Bayes outperformed C4.5, kNN, and SVM for graduation prediction, with an accuracy of 76.79% and an AUC of 0.850. These findings indicate that Naïve Bayes can be very competitive, but the results are highly dependent on the data structure and the fulfillment of model assumptions [12].

2.5 Naïve Bayes

Although the primary focus of this study is the three algorithms mentioned above, the file you uploaded also demonstrates the role of other algorithms. A study by Aulia Putri et al. showed that K-NN outperformed Naïve Bayes and SVM with 87.8% accuracy, 87.8% precision, and 84% recall on data from 379 final-year students [5]. Meanwhile, a UNISA study also showed that K-NN achieved 91.82% accuracy and outperformed Naïve Bayes and Decision Tree on a specific dataset [13].

Internationally, a comparative study of models for student success in online programming courses showed that Logistic Regression achieved an AUC-ROC of 0.9354, while Random Forest, SVM, and Neural Network also performed well. This study confirmed that the combination of LMS data and external factors can improve prediction accuracy. Another study also found that Random Forest and Decision Tree were the best classifiers for predicting student exam performance, with only three misclassifications out of 253 test data sets [14].

3. RESEARH METHODS

3.1 Research Type

This study employed a literature review approach to compare the performance of machine learning algorithms used for student graduation prediction. The review focused on identifying, evaluating, and synthesizing findings from previous studies in order to provide a comprehensive comparison of commonly used prediction methods in Educational Data Mining (EDM) [3].

3.2 Inclusion and Exclusion Criteria

Inclusion criteria:

1. Discusses prediction of student graduation or academic performance.
2. Uses machine learning or data mining algorithms.
3. Presents evaluation metrics such as accuracy, precision, recall, F1-score, AUC, or ROC.
4. Describes data attributes or model implementation.

Exclusion criteria:

1. Duplicate publications.
2. Studies without experimental evaluation.
3. Papers unrelated to student graduation prediction.
4. Articles lacking sufficient methodological information.

3.3 Literature Search Strategy

The literature search was conducted using Google Scholar as the primary database. The search focused on articles published between 2018 and 2025. Several keywords and keyword combinations were used, including:

1. Student graduation prediction
2. Academic performance prediction
3. Machine learning
4. Educational data mining
5. Random Forest
6. Decision Tree
7. C4.5
8. Naïve Bayes

After the initial search, duplicate and irrelevant publications were removed. The remaining articles were screened based on their titles, abstracts, and full-text content before being included in the review.

3.4 Article Selection Process

The article selection process consisted of four stages. First, relevant studies were identified through Google Scholar using predefined search keywords. Second, duplicate and irrelevant publications were removed. Third, the remaining articles were screened according to the inclusion and exclusion criteria. Finally, fourteen articles were selected for this review. Twelve empirical studies were included in the comparative analysis presented in Table 1, while two review papers were used as supporting references to strengthen the theoretical background.

3.5 Data Extraction and Analysis

Data extracted from each selected study included the machine learning algorithm used, dataset characteristics, evaluation metrics, performance results, and important predictive attributes. The extracted information was then grouped according to algorithm type and comparatively analyzed to identify similarities, differences, advantages, limitations, and implementation contexts of each method. Rather than comparing accuracy values directly, this study also considered differences in dataset characteristics, feature selection, preprocessing techniques, and evaluation strategies to provide a more balanced interpretation of the reported findings.

3.6 Research Framework

In Figure 1, the research framework consisted of six stages:

1. Identifying the research problem,
2. Searching the literature through Google Scholar,
3. Screening articles based on inclusion and exclusion criteria,
4. Extracting relevant information,

5. Conducting comparative analysis, and
6. Synthesizing the findings to formulate conclusions and recommendations.

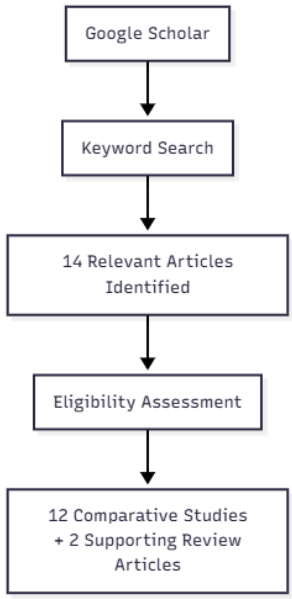


Figure 1. Research Framework

4. RESULT AND DISCUSSIONS

4.1 Method Comparison

The comparison among studies should be interpreted carefully because each study employed different datasets, feature sets, sample sizes, and evaluation methods. Consequently, differences in reported accuracy do not necessarily indicate that one algorithm is universally superior. Instead, the findings in table 1 suggest that algorithm performance is highly dependent on the characteristics of the educational dataset and the experimental setting.

Table 1. Literature Extraction Results

No	Researcher	Method	Dataset	Results
1.	Andrianof et al.	Random Forest	GPA, attendance, core course grades	Accuracy 87,5%, precision 86,3%, recall 85,9%[7]
2.	Setiono	Decision Tree(C4.5)	GPA, credits, attendance, active semester	Accuracy 80%, most dominant presence [2]
3.	Suriani	C4.5	Social Studies, GPA	Accuracy 99,64%, AUC 99,5%[10]
4.	Widianingsih	C4.5, NB, kNN, SVM	Social Studies semesters 3–6, gender	Best Naïve Bayes: 76,79%, AUC 0,850[12]
5.	Kurniawati	C4.5, KNN, ID3, LR, NB	Age, marital status, Social Studies, GPA	Best C4.5: 94,33%[4]
6.	Zeniarja et al.	NB, RF, DT, KNN, SVM	12 academic attributes	Best RF: 77,35%[9]
7.	Afandi et al.	NB, DT, RF	Social Studies, credits semesters 1–6	Best RF: 97,50%, F1 97%[1]
8.	Gunawan & Paputungan	Random Forest	Grade, attendance, study time, social factors	Accuracy 89%, AUC 0,94[8]
9.	Putri et al.	K-NN, NB, SVM	Public data: 379 students	Best K-NN: 87,8%[5]
10.	Hardiani	NB, K-NN, DT	UNISA student data	Best K-NN: 91,82%[13]
11.	Meiriza et al.	Naïve Bayes	SI student data 2015–2016	Accuracy 97,6378%[11]
12.	Arévalo-Cordovilla & Peña	LR, RF, SVM, MLP	LMS + external factors	LR AUC-ROC 0,9354[6]

4.2 Analysis of Method Implementation

The comparative analysis indicates that the reported performance of machine learning algorithms varies considerably across the reviewed studies. These differences are influenced not only by the algorithms themselves but also by dataset characteristics, feature selection, preprocessing techniques, sample size, class distribution, and evaluation strategies. Therefore, the reported accuracy values should not be interpreted as direct evidence that one algorithm is universally superior.

a) Random Forest

Random Forest consistently demonstrated strong predictive performance in many of the reviewed studies. Its ensemble learning mechanism enables the model to reduce overfitting while handling complex relationships among academic variables. Studies reporting high performance generally used multiple academic attributes, including GPA, completed credits, attendance, and study duration. These characteristics allow Random Forest to capture nonlinear relationships more effectively than single-tree classifiers [1][8]. However, the superior performance of Random Forest should be interpreted carefully. Differences in dataset size, attribute composition, feature engineering, and validation techniques may contribute to higher accuracy in some studies. Consequently, its reported performance cannot be generalized to every educational dataset.

b) Decision Tree / C4.5

Decision Tree remains one of the most widely applied algorithms because of its high interpretability. Unlike ensemble models, Decision Tree produces classification rules that are easy for lecturers and university administrators to understand. This characteristic makes it particularly suitable for early warning systems and academic decision support [2]. Nevertheless, the reviewed studies reported considerable variation in Decision Tree performance. Some studies achieved relatively moderate accuracy, while others reported very high results. These variations suggest that Decision Tree is highly sensitive to data quality, feature characteristics, preprocessing procedures, and class distribution. Therefore, good predictive performance depends largely on the suitability of the dataset rather than on the algorithm alone [2][4][10].

c) Naïve Bayes

Naïve Bayes offers a computationally efficient approach with relatively simple implementation. Several studies reported competitive predictive performance despite the simplicity of the algorithm. Its efficiency makes it suitable for educational institutions with limited computational resources. However, Naïve Bayes assumes statistical independence among predictor variables. In educational datasets, important variables such as GPA, completed credits, attendance, and study duration are often correlated. When this assumption is violated, prediction performance may decrease compared with more flexible algorithms such as Random Forest. Overall, the reviewed studies demonstrate that each algorithm possesses different strengths and limitations. Rather than identifying a universally superior method, the findings suggest that algorithm selection should consider the objectives of the prediction system, data characteristics, model interpretability, computational requirements, and institutional needs [1][11].

4.3 Most Influential Factors

Across the reviewed studies, several academic attributes consistently appeared as important predictors of student graduation. Grade Point Average (GPA), completed credits, attendance, and active semesters were the most frequently reported variables influencing prediction performance. These variables directly reflect students' academic progress and learning consistency [2]. Several studies also highlighted the contribution of non-academic variables, including demographic information, learning management system (LMS) activities, motivation, study behavior, and social factors. Although these variables were not used consistently across all studies, they often improved prediction performance when combined with academic attributes [7]. The variation in selected features among the reviewed studies further explains why different machine learning algorithms produced different performance levels. Therefore, feature selection should be considered an important factor alongside algorithm selection when developing student graduation prediction models [3][6].

4.4 Method Effectiveness

Based on the reviewed literature, Random Forest appears to provide consistently strong performance across many studies. Nevertheless, no single algorithm can be considered universally superior because prediction performance depends on dataset characteristics, feature selection, preprocessing techniques, and evaluation strategies. Therefore, algorithm selection should be aligned with the objectives and context of the prediction task. Decision Tree provides an important advantage in terms of model transparency and interpretability, allowing academic staff to understand the reasoning behind predictions and translate them into institutional policies. Meanwhile, Naïve Bayes offers fast computation and relatively simple implementation, making it appropriate for datasets that satisfy its underlying assumptions.

Therefore, selecting an appropriate machine learning algorithm should not rely solely on prediction accuracy. Universities should also consider data characteristics, implementation objectives, computational resources, and the need for model interpretability when choosing an algorithm for graduation prediction systems. In the context of predicting student graduation for higher education, Random Forest is more appropriately recommended as the primary model because:

1. It performs better in many studies.
2. It is capable of handling diverse attributes.
3. It is more resistant to overfitting.
4. It is suitable for academic decision support systems [1][8][9].

5. CONCLUSSION

This literature review compared the application of machine learning algorithms for student graduation prediction based on fourteen selected articles published between 2018 and 2025. The reviewed studies indicate that Random Forest, Decision Tree (C4.5), and Naïve Bayes are the most frequently applied algorithms in this research area, each offering different advantages depending on the prediction context.

Among the reviewed methods, Random Forest generally demonstrated stable predictive performance across various studies due to its ability to handle complex relationships among multiple academic attributes and reduce overfitting. However, this finding should not be interpreted as evidence that Random Forest is universally superior. The reviewed studies employed different datasets, feature sets, preprocessing techniques, sample sizes, and evaluation methods, which may significantly influence the reported performance of each algorithm. Decision Tree remains a valuable alternative when model transparency and interpretability are essential, particularly for academic decision support and early warning systems. Meanwhile, Naïve Bayes provides an efficient and computationally lightweight solution that performs well in datasets that satisfy its underlying assumptions.

The review also found that Grade Point Average (GPA), completed credits, attendance, and active semesters are the academic variables most frequently associated with successful graduation prediction. In addition, several studies reported that integrating non-academic information, such as demographic characteristics, learning management system (LMS) activities, and behavioral factors, may further improve prediction performance. Overall, this review suggests that selecting an appropriate machine learning algorithm should consider not only predictive accuracy but also dataset characteristics, feature selection, interpretability, computational requirements, and the intended application within higher education institutions.

6. LIMITATIONS OF THE STUDY

This study has several limitations. First, the literature search was conducted only through google scholar, which may not include all relevant publications available in other academic databases. Second, the reviewed studies employed different datasets, feature sets, preprocessing techniques, and evaluation methods, making direct comparison of algorithm performance difficult. Third, this study relied on a qualitative literature review without performing a quantitative meta-analysis. Therefore, the findings should be interpreted as a comparative synthesis rather than definitive evidence of the superiority of any particular machine learning algorithm.

Future research is encouraged to adopt a systematic literature review (slr) or meta-analysis approach using multiple academic databases, such as scopus, web of science, and ieee xplore, to provide a more comprehensive and rigorous comparison of machine learning methods for student graduation prediction.

REFERENCES

[1] M. C. Afandi, U. Indahyanti, H. Setiawan, and I. A. Kautsar, "V4_PERBANDINGAN+ALGORITMA+MACHINE+LEARNING+DALAM+MEMPREDIKSI+KELULUSAN+MAHASISWA+_11," *semanTIK*, vol. 11, no. 2, 2025.

[2] S. A. Setiono, E. Purwanto, U. Duta, and B. Surakarta, "Prediksi Kelulusan Mahasiswa Menggunakan Algoritma Decision Tree," *Seminar Nasional Teknologi Informasi dan Bisnis (SENATIB)*, p. 2025, 2025.

[3] M. Salim, N. Al-Din, and H. A. Al Abdulqader, "Students' Academic Performance Prediction Using Educational Data Mining and Machine Learning: A Systematic Review," *International Journal of Research and Innovation in Social Science*, 2024.

[4] I. Kurniawati, "Komparasi Metode Machine Learning Pada Prediksi Kelulusan Mahasiswa A Comparative Method of Machine Learning On Predicting Students Graduation," 2018.

[5] A. Putri, C. Syaficha Hardiana, E. Novfuja, F. Try Puspa Siregar, Y. Fatma, and R. Wahyuni, "Comparison of K-NN, Naive Bayes and SVM Algorithms for Final-Year Student Graduation Prediction Komparasi Algoritma K-NN, Naive Bayes dan SVM untuk Prediksi Kelulusan Mahasiswa Tingkat Akhir," *Institut Riset dan Publikasi Indonesia (IRPI) MALCOM: Indonesian Journal of Machine Learning and Computer Science Journal Homepage*, vol. 3, no. 1, pp. 20–26, 2023.

[6] F. E. Arévalo-Cordovilla and M. Peña, "Comparative Analysis of Machine Learning Models for Predicting Student Success in Online Programming Courses: A Study Based on LMS Data and External Factors," *Mathematics*, vol. 12, no. 20, Oct. 2024, doi: 10.3390/math12203272.

[7] H. Andrianof, A. P. Gusman, and O. A. Putra, "Implementasi Algoritma Random Forest untuk Prediksi Kelulusan Mahasiswa Berdasarkan Data Akademik: Studi Kasus di Perguruan Tinggi Indonesia," *Jurnal Sains Informatika Terapan (JSIT) E-ISSN*, vol. 4, no. 1, pp. 24–28, 2025.

[8] P. H. Gunawan and I. V. Paputungan, "Sistemasi: Jurnal Sistem Informasi Deteksi Tingkat Potensi Kelulusan Calon Mahasiswa menggunakan Algoritma Random Forest Detection of Graduation Potential in Prospective Students using the Random Forest Algorithm," *Sistemasi: Jurnal Sistem Informasi*, vol. 14, no. 5, pp. 2055–2065, 2025, [Online]. Available: http://sistemasi.ftik.unisi.ac.id

[9] J. Zeniarja, A. Salam, and F. A. Ma'ruf, "Seleksi Fitur dan Perbandingan Algoritma Klasifikasi untuk Prediksi Kelulusan Mahasiswa," *Jurnal Rekayasa Elektrika*, vol. 18, no. 2, pp. 102–108, Jun. 2022, doi: 10.17529/jre.v18i2.24047.

[10] P. Studi Teknologi Rekayasa Multimedia and P. Darussalam, "Penerapan Data Mining untuk Memprediksi Tingkat Kelulusan Mahasiswa Menggunakan Algoritma Decision Tree C4.5 Uci Suriani," *Journal of Computer and Information Systems Ampera*, vol. 3, no. 2, 2023, doi: 10.51519/journalcisa.v4i2.393.

[11] A. Meiriza, E. Lestari, P. Putra, A. Monaputri, and D. A. Lestari, "Advances in Intelligent Systems Research," in *Advances in Intelligent Systems Research*, Atlantis Press, 2020.

[12] S. Widaningsih, "PERBANDINGAN METODE DATA MINING UNTUK PREDIKSI NILAI DAN WAKTU KELULUSAN MAHASISWA PRODI TEKNIK INFORMATIKA DENGAN ALGORITMA C4,5, NAÏVE BAYES, KNN DAN SVM," *Jurnal Tekno Insentif*, vol. 13, no. 1, pp. 16–25, Apr. 2019, doi: 10.36787/jti.v13i1.78.

[13] T. Hardiani, "Comparison of Naive Bayes Method, K-NN (K-Nearest Neighbor) and Decision Tree for Predicting the Graduation of â€˜Aisyiyah University Students of Yogyakarta," *International Journal of Health Science and Technology*, vol. 2, no. 1, pp. 75–85, Jan. 2021, doi: 10.31101/ijhst.v2i1.1829.

[14] D. Khairy, N. Alharbi, M. A. Amasha, M. F. Areed, S. Alkhalaf, and R. A. Abougalala, "Prediction of student exam performance using data mining classification algorithms," *Educ. Inf. Technol. (Dordr).*, vol. 29, no. 16, pp. 21621–21645, Nov. 2024, doi: 10.1007/s10639-024-12619-w.