JURNAL ACCOUNTING INFORMATION SYSTEM
AIMS
Jurnal Accounting Information System (AIMS)
Volume 9 No. 1 | March 2026 : 25-39
DOI: 10.32627
http://journal.masoemuniversity.ac.id/index.php/aims
p-ISSN: 2615-7381
e-ISSN: 2621-7279

Classification of Delayed Students Graduation Risk : A Comparative Analysis of Naive Bayes, XGBoost, and Random Forest

Khafka Fadillah Wibawa N[1], Aldy Rialdy Atmadja[2], Nur Lukman[3]
[1,2,3] Informatics, Sunan Gunung Djati State Islamic University, Indonesia
khafka.fadillahww@gmail.com

Article Info

Article history:
Accepted January 2026
Revised February 2026
Approved February 2026
Published March 2026

ABSTRACT

One of the critical challenges affecting the effectiveness of higher education systems is delayed student graduation, which not only impacts institutional performance but also increases the financial and psychological burden on students. This study aims to classify the risk of delayed graduation by developing and evaluating machine learning models based on new student admission data. The dataset was obtained from the New Student Admission Center of UIN Sunan Gunung Djati Bandung and consists of students' biodata, including socioeconomic characteristics and the educational background of students and their parents. The research was conducted following the CRISP-DM framework, encompassing business understanding, data understanding, data preparation, modeling, evaluation, and deployment planning. During the data preparation stage, preprocessing techniques such as data cleaning, encoding of categorical variables, and feature selection were applied to improve data quality. Three machine learning algorithms—Naïve Bayes, Random Forest, and XGBoost—were implemented and optimized using hyperparameter tuning to achieve optimal performance. Model evaluation was carried out using accuracy, precision, recall, F1-score, and ROC-AUC metrics to ensure a comprehensive comparison.The experimental results demonstrate that the Random Forest algorithm outperformed the other models, achieving an accuracy of 0.633, precision of 0.677, recall of 0.694, F1-score of 0.685, and ROC-AUC of 0.668. These findings indicate that machine learning models based on admission data are capable of providing a reasonably effective early prediction of delayed graduation risk. Nevertheless, the model performance can be further enhanced by incorporating academic performance variables during the study period. This study is expected to support higher education institutions in formulating data-driven strategies and early intervention programs for students with a high risk of delayed graduation.

Keywords: CRISP-DM; Graduation Delay; Machine Learning; Naïve Bayes; New Student Admission Data; Random Forest; Xgboost.

p-ISSN: 2615-7381
e-ISSN: 2621-7279

INTRODUCTION

Delayed student graduation has become a major concern in higher education systems because of its multidimensional impact on both students and educational institutions. For students, prolonged study duration increases financial burdens due to additional tuition fees and living expenses, delays access to employment opportunities, and may negatively affect psychological well-being, including stress, anxiety, and decreased academic motivation[1]. From an institutional perspective, delayed graduation can reduce academic efficiency, negatively influence graduation rate indicators, disrupt academic planning, and ultimately affect institutional reputation and accreditation outcomes. Consequently, addressing delayed graduation is not only an academic issue but also a strategic concern for higher education management. Study duration is influenced by students' initial characteristics at the time of admission[2], Therefore, new student admission data can be utilized for early detection and academic intervention. These early-stage attributes often shape students' academic trajectories throughout their study period. Therefore, data collected during the new student admission process can serve as a valuable source for early detection of students who are at risk of delayed graduation, enabling institutions to design timely academic guidance and intervention programs.

In recent years, machine learning–based approaches have gained increasing attention due to their capability to extract meaningful patterns from large-scale and complex educational data[3]. Methods such as Random Forest, Naive Bayes, and XGBoost have been widely used in various predictive studies, ranging from academic performance analysis and dropout risk identification to the development of early warning systems for at-risk students [4]. Previous research conducted by Kingsley Okoye entitled 'Machine Learning Model (RG-DMML) and Ensemble Algorithm for the Prediction of Students' Retention and Graduation in Education', employing the CRISP-DM methodology, demonstrated that machine learning can effectively be used as a tool to predict the risk of delayed graduation. [5]. Fajar Riskiyono, in his study entitled 'Implementation of Random Forest Algorithm for Graduation Prediction', demonstrated that Random Forest can serve as a highly effective model for predicting student graduation [6]. In addition, Siti Nurlia, in her study entitled 'Implementation of Naive Bayes Classifier in Predicting Student Graduation', demonstrated that the Naive Bayes algorithm achieved strong evaluation results, with an F1-score of 90%, in predicting student graduation using academic data collected during the study period [7]. Wilda Imama Sabilla, in her study entitled 'Implementation of SMOTE and Under-Sampling on Imbalanced Datasets for Corporate Bankruptcy Prediction', demonstrated that SMOTE is an effective technique for handling data imbalance, achieving a recall value of 95.45% [8]. Despite the extensive application of machine learning in educational prediction tasks, the utilization of new student admission (PMB) data for predicting delayed graduation remains relatively limited. This gap highlights an opportunity to develop early-stage predictive models that focus on pre-academic variables. Such

p-ISSN: 2615-7381
e-ISSN: 2621-7279

models can support preventive academic interventions, improve institutional decision-making, and strengthen data-driven academic policies. Ultimately, the implementation of early warning systems based on admission data has the potential to sustainably enhance the quality and effectiveness of higher education services [9].

METODE

This study employs a quantitative approach by adopting the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework. The CRISP-DM framework was selected due to its systematic and flexible structure, and its proven effectiveness [10]. The stages of this framework include business understanding, data understanding, data preparation, modeling, evaluation, and deployment, as illustrated in Figure 1.

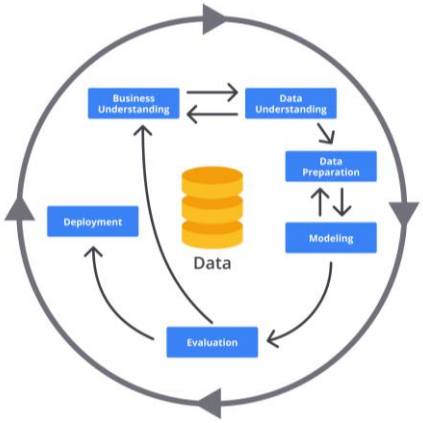

Figure 1. CRISP-DM Process Flow

Business Understanding

The rising incidence of delayed student graduation has emerged as a significant issue, affecting educational costs, access to employment opportunities, and institutional reputation. This research proposes a machine learning–based model utilizing new student admission data to accurately identify potential graduation delays and support data-driven academic policy development.

Data Understanding

This study uses a dataset containing various attributes that reflect the profiles of new students, including personal identity, school of origin, previous major, socioeconomic background, and parents' occupations, obtained from the Central Unit of New Student Admissions (PMB) at UIN Sunan Gunung Djati Bandung. In addition, a dataset from the Center for Information Technology and Data Repository of UIN Sunan Gunung Djati Bandung was used, which contains students' academic information such as first registration date, commencement of active study period, and graduation date. These two datasets were subsequently integrated through a data joining process to produce a comprehensive dataset.

p-ISSN: 2615-7381
e-ISSN: 2621-7279

Dataset Description

The integrated dataset consists of 48,068 records with 80 features representing various important aspects, including student identity, educational background, family economic conditions, and other social indicators. During the initial exploration stage, exploratory data analysis (EDA) was conducted to assess data quality and consistency, including the identification of missing values, outlier detection, duplicate value detection, and evaluation of the target variable distribution.

The target variable in this study is on-time graduation, which consists of two categories: on-time (0) and delayed (1). Initial distribution analysis indicates the presence of class imbalance, with 8,477 students classified as delayed and 6,257 students graduating on time. Students who did not continue their studies until the final semester were categorized as delayed.

Data Preparation

At this stage, the raw data were transformed into a clean dataset ready for the modeling process. The steps performed included data cleaning, feature selection and engineering, encoding, scaling, and data balancing [11].

Data Cleaning

The initial dataset consisted of 48,068 records and 80 features. Several attributes were removed because they did not provide significant contributions to the modeling process, including student names, full addresses, identification numbers, and administrative data such as information update dates and system account status [12]. Missing data imputation was performed selectively. For categorical attributes such as living arrangement, school major, father's education, mother's education, school type, father's occupation, and mother's occupation, the category 'Others' was used to fill missing values. For the father's income and mother's income attributes, missing values were imputed using the category 'IDR 0–400,000'. For the admission pathway and on-time graduation attributes, imputation was conducted using the mode of each feature, as it best represents the general tendency of the data population [13].

The KPS recipient attribute was imputed with a value of 0, while the number of dependents attribute was filled with the value '2 persons', based on the median of the distribution. For the transportation mode attribute, missing values were imputed using the 'Does not own' category, and the funding type attribute was filled with the 'Self-funded' category[13]. This data cleaning process ensures that the dataset used for modeling is free from noise and statistical anomalies [14]. The dataset that has passed this stage is ready to be further processed in the feature engineering and numerical transformation stages to build a robust model [15].

In this study, feature engineering was conducted by systematically creating several new features to enhance the model's ability to capture influential patterns. The first step involved transforming income categories using numerical mapping,

p-ISSN: 2615-7381
e-ISSN: 2621-7279

where each income category was converted into a numerical data type. Subsequently, a new feature, econ_score, was constructed to represent the family's economic condition by considering total income and the number of dependents. The formula used is presented as follows.

$$\text{Econ\_index\_raw} = log(1 + Income\ Total) - Dependent\ Count + 0.5$$
$$\text{econ\_score} = \frac{(\text{Econ\_Index\_Raw} - \min(\text{Econ\_Index\_Raw}))}{(\max(\text{Econ\_Index\_Raw}) - \min(\text{Econ\_Index\_Raw}))} \quad (1)$$

The formula combines a logarithmic transformation (log1p) to stabilize the distribution of income data with a penalty based on the number of household dependents. All resulting values were then normalized using a Min–Max Scaler to ensure they fall within the range [0, 1]. The next step involved constructing the parent_socio_score feature, which represents the parents' socioeconomic status by combining the education level and occupation type of both the father and the mother. Each category was converted into an ordinal numerical scale and mapped into new features, namely edu_map for education and job_map for occupation. The average of these four variables was then calculated. To maintain scale homogeneity across features, the parent_socio_score values were normalized using Min–Max Scaling, as defined in the following formula.

$$\text{Parent\_Socio\_Raw} = \frac{\text{Edu}_{\text{father}} + \text{Edu}_{\text{mother}} + \text{Job}_{\text{father}} + \text{Job}_{\text{mother}}}{4}$$
$$\text{parent\_socio\_score} = \frac{(\text{Parent\_Socio\_Raw} - \min(\text{Parent\_Socio\_Raw}))}{(\max(\text{Parent\_Socio\_Raw}) - \min(\text{Parent\_Socio\_Raw}))} \quad (2)$$

In addition to these two main indices, several other derived features were also developed, including:

1. Fam_dependency_rate (FDR), which represents the ratio between family dependency burden and household financial capacity.
$$\text{fdr\_Raw} = \frac{\text{Dependent Count}}{\log(1 + \text{Income Total}) + 10^{-6}}$$
$$\text{fdr} = \frac{\text{fdr\_Raw} - \min(\text{fdr\_Raw})}{\max(\text{fdr\_Raw}) - \min(\text{fdr\_Raw})} \quad (3)$$
2. Socioecon_composite_score, which is a composite feature combining parent_socio_score and econ_score to provide a holistic view of the family's socioeconomic condition.
$$\text{socioecon\_composite\_score} = \text{parent\_socio\_score} + \text{econ\_score} \quad (4)$$
3. Econ_load_ratio, which measures the family's economic capacity after accounting for dependency burden.
$$\text{econ\_load\_ratio} = \frac{\text{econ\_score}}{\text{fdr}+1} \quad (5)$$
4. Avg_parent_edu and max_parent_edu, which represent the average and highest education levels of both parents, respectively, and are assumed to influence the child's educational values.

p-ISSN: 2615-7381
e-ISSN: 2621-7279

$$\text{avg\_parent\_edu} = \frac{\text{Edu}_{\text{father}} + \text{Edu}_{\text{mother}}}{2}$$
$$\text{max\_parent\_edu} = \max(\text{Edu}_{\text{father}}, \text{Edu}_{\text{mother}}) \quad (6)$$

Encoding and Scaling

The encoding process was applied to transform categorical features into numerical representations so that they can be processed by machine learning algorithms. The approach employed consisted of three techniques:

1. One-Hot Encoding (OHE) was applied to features with a limited number of categories, such as admission_pathway, transportation_mode, funding_type, living_arrangement, and school_major [16].
2. Target Encoding (TE) was used for categorical features with a large number of categories, such as father_education, father_occupation, mother_education, mother_occupation, school_type, and major. [17].
3. Label Encoding (LE) was applied specifically to features with only two categories, such as the target feature on_time_graduation [18].

Following feature encoding, standardization was applied using StandardScaler to balance feature scales and mitigate the dominance of features with larger magnitudes in the modeling process.

Data Balancing

The distribution of the target variable on_time_graduation indicates an imbalance, with 57.5% of students experiencing delayed graduation and 42.5% graduating on time. This imbalance has the potential to reduce the model's ability to effectively learn patterns associated with the minority class [19]. Therefore, the Synthetic Minority Oversampling Technique (SMOTE) was applied to mitigate class imbalance by creating synthetic minority class samples through interpolation between closely related data points in the feature space [8].

Final Dataset and Data Splitting

The processed dataset comprises 14,734 records and 23 features, including seven engineered and scaled features, along with a binary target variable, on_time_graduation, where 0 indicates on-time (tepat waktu) graduation and 1 indicates delayed.

After completing all data cleaning, transformation, and feature engineering processes, the dataset was divided into two main subsets: the training set and the testing set [20]. The data was split with a proportion of 80% for the training set and 20% for the testing set.

Modeling

The modeling stage employed three machine learning algorithms, Random Forest (RF), Naïve Bayes (NB), and Extreme Gradient Boosting (XGBoost). These algorithms were selected to represent different classification characteristics, where Naïve Bayes offers a simple probabilistic baseline model, Random Forest provides

p-ISSN: 2615-7381
e-ISSN: 2621-7279

robustness through ensemble bagging, and XGBoost is capable of capturing complex non-linear relationships through boosting techniques. This combination enables a balanced and comprehensive comparison of model performance. Hyperparameter tuning was performed using RandomizedSearchCV with three-fold cross-validation (cv = 3)[21]. which was chosen to efficiently explore the hyperparameter space while reducing computational cost. The tuning process focused on optimizing key parameters of each algorithm to improve generalization performance. To further assess the impact of data imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) was applied. Each model was evaluated under four scenarios: without tuning and without SMOTE, without tuning with SMOTE, with tuning without SMOTE, and with both tuning and SMOTE. The baseline models used default parameter settings without data balancing or tuning, serving as reference benchmarks.

Evaluation and Deployment

1. The model performance was evaluated using several metrics: Accuracy to represent overall correctness.
2. Precision to indicate the model's ability to correctly identify students with delayed graduation.
3. Recall to measure the model's capability in capturing all students who graduate late.
4. F1-score to reflect the balance between precision and recall.
5. ROC-AUC to assess the quality of separation between on-time and delayed graduation classes.

The final stage of the study focused on deploying the best-performing model into a web-based system using Streamlit. The web application was designed with a simple user interface, allowing users to input new student data to obtain estimated probabilities of delayed graduation.

RESULT AND DISCUSSION

Random Forest Model Result

The Random Forest model was trained under four different scenarios, all of which outperformed the baseline Random Forest model. The baseline model achieved an accuracy of 0.612, precision of 0.630, recall of 0.655, F1-score of 0.642, and ROC-AUC of 0.628. The evaluation results of the four scenarios are presented in Table 1.

Table 1. RF Model Performance

Model	SMOTE	Tuning	Accuracy	Precision	Recall	F1-Score	ROC-AUC
RF	✗	✗	0.633	0.680	0.684	0.682	0.669

p-ISSN: 2615-7381
e-ISSN: 2621-7279

RF	✓	✗	0.634	0.676	0.700	0.687		0.667
RF	✗	✓	0.633	0.677	0.694	0.685		0.668
RF	✓	✓	0.628	0.669	0.700	0.684		0.653

Overall, Random Forest demonstrated the most stable performance among the three algorithms. When hyperparameter tuning was applied using RandomizedSearchCV, the model's performance remained consistent, with accuracy ranging from approximately 0.628 to 0.634 and ROC-AUC from 0.653 to 0.668. This indicates that Random Forest exhibits strong robustness against variations in data and balancing techniques. Although performance improved slightly after tuning, the increase was not substantial.

Naïve Bayes Model Result

Naïve Bayes exhibited stable performance, although it tended to underperform compared to Random Forest and XGBoost in most scenarios. The baseline Naïve Bayes model recorded an accuracy of 0.598, precision of 0.615, recall of 0.641, F1-score of 0.627, and ROC-AUC of 0.617. Training under the four experimental scenarios led to improvements in overall performance, summarized in Table 2.

Table 2. NB Model Performance

Model	SMOTE	Tuning	Accuracy	Precision	Recall	F1-Score	ROC-AUC
NB	✗	✗	0.612	0.662	0.665	0.663	0.631
NB	✓	✗	0.612	0.662	0.665	0.663	0.628
NB	✗	✓	0.612	0.662	0.665	0.663	0.631

p-ISSN: 2615-7381
e-ISSN: 2621-7279

NB	✓	✓	0.592	0.663	0.592	0.625	0.628

Across all scenarios, accuracy ranged from approximately 0.592 to 0.612, with precision and recall nearly identical at around 0.662–0.665, and ROC-AUC approximately 0.628–0.631. This indicates that the model tends to produce balanced results between the two classes, but it lacks high sensitivity for identifying students at risk of delayed graduation.

XGBoost Model Result

The baseline XGBoost model yielded relatively low performance, with an accuracy of 0.601, precision of 0.620, recall of 0.667, F1-score of 0.643, and ROC-AUC of 0.624. After training under the four defined scenarios, the XGBoost model showed improved performance, as presented in Table 3. All four experimental scenarios significantly boosted the performance of the XGBoost model, especially with regard to Recall and F1-Score, which showed marked increases.

Table 3. XGBoost Model Performance

Model	SMOTE	Tuning	Accuracy	Precision	Recall	F1-Score	ROC-AUC
XGB	✗	✗	0.588	0.631	0.684	0.657	0.608
XGB	✓	✗	0.588	0.631	0.684	0.657	0.608
XGB	✗	✓	0.606	0.600	0.946	0.734	0.670

p-ISSN: 2615-7381
e-ISSN: 2621-7279

XGB	✓	✓	0.615	0.610	0.917	0.733	0.646

Discussion

This study aimed to develop a predictive model for identifying the risk of delayed graduation using new student admission (PMB) data. Baseline models were initially developed to evaluate the fundamental performance of Random Forest (RF), Naïve Bayes (NB), and XGBoost without feature engineering, data balancing, or hyperparameter tuning.

Table 4. Base Line Model

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
RF	0.612	0.630	0.655	0.642	0.628
NB	0.598	0.615	0.641	0.627	0.617
XGB	0.601	0.620	0.667	0.643	0.624

As shown in Table 4, all baseline models demonstrated moderate predictive performance, with Random Forest slightly outperforming the other algorithms. These baseline results provide a valid benchmark for assessing the impact of subsequent optimization strategies. This baseline serves as a benchmark for evaluating the effects of feature engineering, hyperparameter optimization, and SMOTE on overall model performance. Following feature engineering, hyperparameter tuning, and the application of SMOTE, model performance showed noticeable changes, as summarized in Table 5.

Table 5. Model Performance

Model	SMOTE	Tuning	Accuracy	Precision	Recall	F1-Score	ROC-AUC
RF	✗	✗	0.6	0.6	0.6	0.682	0.669

p-ISSN: 2615-7381
e-ISSN: 2621-7279

			3 3	8 0	8 4		
RF	✓	✗	0. 6 3 4	0. 6 7 6	0. 7 0 0	0.687	0.667
RF	✗	✓	0. 6 3 3	0. 6 7 7	0. 6 9 4	0.685	0.668
RF	✓	✓	0. 6 2 8	0. 6 6 9	0. 7 0 0	0.684	0.653
NB	✗	✗	0. 6 1 2	0. 6 6 2	0. 6 6 5	0.663	0.631
NB	✓	✗	0. 6 1 2	0. 6 6 2	0. 6 6 5	0.663	0.628
NB	✗	✓	0. 6 1 2	0. 6 6 2	0. 6 6 5	0.663	0.631
NB	✓	✓	0. 5 9 2	0. 6 6 3	0. 5 9 2	0.625	0.628
XGB	✗	✗	0. 5 8 8	0. 6 3 1	0. 6 8 4	0.657	0.608
XGB	✓	✗	0. 5 8 8	0. 6 3 1	0. 6 8 4	0.657	0.608
XGB	✗	✓	0. 6 0 6	0. 6 0 0	0. 9 4 6	0.734	0.670
XGB	✓	✓	0. 6 1 5	0. 6 1 0	0. 9 1 7	0.733	0.646

p-ISSN: 2615-7381
e-ISSN: 2621-7279

Feature engineering contributed to performance improvements by enabling the models to better capture interactions among input variables. Random Forest demonstrated stable and consistent performance across different scenarios, indicating strong robustness to data transformations. The best-performing configuration was Random Forest with SMOTE but without hyperparameter tuning, achieving an accuracy of 0.634, precision of 0.676, recall of 0.700, F1-score of 0.687, and ROC-AUC of 0.667. This result suggests that handling class imbalance had a more significant impact than parameter optimization for this dataset. The evaluation process employed multiple metrics—accuracy, precision, recall, F1-score, and ROC-AUC—to ensure result validity and avoid reliance on a single performance indicator, particularly under imbalanced class conditions.

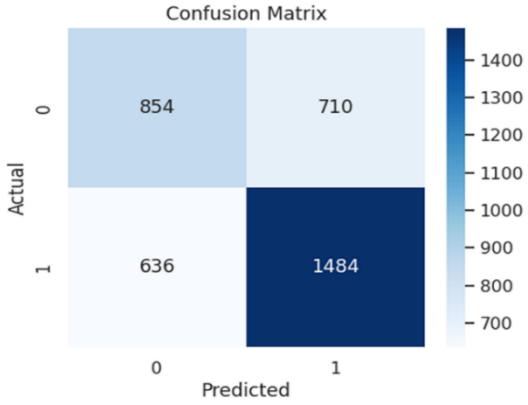

Figure 2. Confusion Matrix Best RF

The confusion matrix in Figure 2 shows a relatively balanced classification between on-time and delayed graduation classes, with a reduced number of false negatives compared to the baseline model. This is particularly important for early warning systems, as minimizing false negatives helps ensure that at-risk students are not overlooked. This indicates that SMOTE effectively addressed class imbalance in the training data.

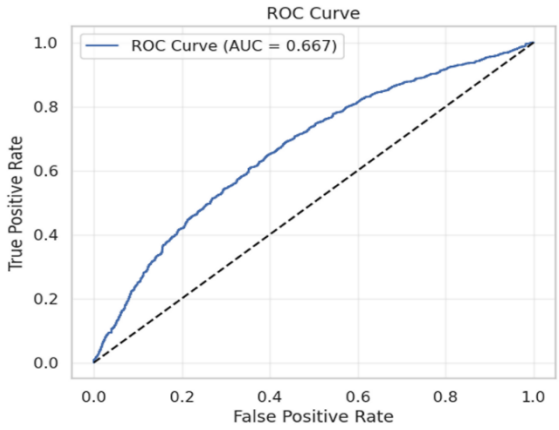

Figure 3. ROC Curve Best RF

p-ISSN: 2615-7381
e-ISSN: 2621-7279

From Figure 3, the best ROC-AUC reached 0.667, an improvement from the baseline value of 0.628, with the curve consistently above the random line, indicating moderate capability in distinguishing between the two classes. Although XGBoost achieved a higher recall and F1-score after tuning, its lower precision indicates a higher false-positive rate, which may reduce its suitability for practical academic interventions. In contrast, Random Forest provided a more balanced trade-off between precision and recall, making it the most reliable and interpretable model for implementation. From a practical perspective, the proposed model can support academic management by enabling early identification of students at risk of delayed graduation based on admission data. This allows institutions to design targeted interventions such as academic mentoring, financial assistance, or counseling programs at an early stage of study. By integrating such predictive models into academic monitoring systems, universities can improve graduation rates, optimize resource allocation, and strengthen data-driven decision-making in academic management.

CONCLUSION

Based on the results of this study, the structured application of feature engineering has been shown to significantly contribute to improving model performance. The baseline Random Forest model achieved an accuracy of 0.634, precision of 0.676, recall of 0.700, and ROC-AUC of 0.667. After adding engineered features, the model scenario using SMOTE without hyperparameter tuning demonstrated the most optimal performance. These results indicate that handling class imbalance plays an important role in improving predictive performance, particularly for early identification of students at risk of delayed graduation. In practice, this model can be used as an early warning mechanism to support academic management in designing targeted intervention strategies, such as academic mentoring, financial assistance, or counseling programs.

These findings confirm that students' socioeconomic background information in the PMB data has at least some influence on the timeliness of graduation. This suggests that admission-stage data can serve as a valuable foundation for developing predictive models that support data-driven academic policies.Future research is recommended to integrate the predictive model into campus academic systems and to include students' academic grades in the modeling stage. Further studies may also consider filtering out students who do not continue until the end of their studies and experimenting with alternative algorithms to further enhance model robustness and practical applicability.

BIBLIOGRAPHY

[1] A. Fadli, T. Limbong, R. Priskila, and V. Handrianus Pranatawijaya, "Penggunaan Algoritma Naive Bayes Untuk Memprediksi Kelulusan Mahasiswa," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 8, no. 3, pp. 3773–3779, 2024, doi: 10.36040/jati.v8i3.9791.

[2] V. Amalia, "Faktor-faktor yang Mempengaruhi Lama Studi Mahasiswa Menggunakan Analisis Survival (Studi Kasus pada Mahasiswa Program Sarjana FMIPA Universitas Brawijaya)," 2018.

[3] E. Haryatmi and S. Pramita Hervianti, "Penerapan Algoritma Support Vector Machine Untuk Model Prediksi Kelulusan Mahasiswa Tepat Waktu," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 2, pp. 386–392, 2021, doi: 10.29207/resti.v5i2.3007.

[4] A. A. Permana, R. Taufiq, R. Destriana, and A. Nur'aini, "Implementasi Algortima Naïve Bayes Untuk Prediksi Kelulusan Mahasiswa," *J. Tek.*, vol. 13, no. 1, pp. 65–70, 2024, [Online]. Available: https://jurnal.umt.ac.id/index.php/jt/article/view/10996

[5] K. Okoye, J. T. Nganji, J. Escamilla, and S. Hosseini, "Machine learning model (RG-DMML) and ensemble algorithm for prediction of students' retention and graduation in education," *Comput. Educ. Artif. Intell.*, vol. 6, no. January, p. 100205, 2024, doi: 10.1016/j.caeai.2024.100205.

[6] F. Riskiyono and D. Mahdiana, "Implementation of Random Forest Algorithm for Graduation Prediction," *Sinkron*, vol. 8, no. 3, pp. 1662–1670, 2024, doi: 10.33395/sinkron.v8i3.13750.

[7] S. Nuralia, H. Harliana, and T. Prabowo, "Implementasi Naïve Bayes Classifier Dalam Memprediksi Kelulusan Mahasiswa," *J. Autom. Comput. Inf. Syst.*, vol. 3, no. 1, pp. 63–72, 2023, doi: 10.47134/jacis.v3i1.57.

[8] W. I. Sabilla and C. Bella Vista, "Implementasi SMOTE dan Under Sampling pada Imbalanced Dataset untuk Prediksi Kebangkrutan Perusahaan," *J. Komput. Terap.*, vol. 7, no. 2, pp. 329–339, 2021, doi: 10.35143/jkt.v7i2.5027.

[9] R. Sepriansyah and S. D. Purnamasari, "Prediction of Student Graduation Using Naïve Bayes," *Budapest Int. Res. Critics Institute-Journal*, vol. 5, no. 3, pp. 24255–24268, 2022, [Online]. Available: https://doi.org/10.33258/birci.v5i3.6447

[10] C. Schröer, F. Kruse, J. Marx, F. Kruse, and J. Marx, "ScienceDirect ScienceDirect A Systematic Literature Review A Systematic Literature Review on Applying Process Model on Applying CRISP-DM Process Model," *Procedia Comput. Sci.*, vol. 181, no. 2019, pp. 526–534, 2021, doi: 10.1016/j.procs.2021.01.199.

[11] P. Martins, F. Cardoso, P. Váz, J. Silva, and M. Abbasi, "Performance and Scalability of Data Cleaning and Preprocessing Tools: A Benchmark on Large Real-World Datasets," *Data*, vol. 10, no. 5, pp. 1–22, 2025, doi: 10.3390/data10050068.

[12] S. Mehta, "Playing Smart with Numbers: Predicting Student Graduation Using the Magic of Naive Bayes," *Int. Trans. Artif. Intell.*, vol. 2, no. 1, pp. 60–75, 2023, doi: 10.33050/italic.v2i1.405.

[13] Subha, "Handling missing values in dataset — 9 methods that you need to know," 2024. Accessed: Jun. 12, 2025. [Online]. Available: https://medium.com/@pingsubhak/handling-missing-values-in-dataset-7-

p-ISSN: 2615-7381
e-ISSN: 2621-7279

methods-that-you-need-to-know-5067d4e32b62
[14] R. R. Deshmukh, "Data Cleaning: Current Approaches and Issues," no. June, 2015.
[15] T. Rawat, "Feature Engineering (FE) Tools and Techniques for Better Classification Performance," no. May, 2019, doi: 10.21172/ijiet.82.024.
[16] F. Bolikulov, R. Nasimov, A. Rashidov, and F. Akhmedov, "Effective Methods of Categorical Data Encoding for Artificial Intelligence Algorithms," 2024.
[17] K. Potdar, "A Comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers," no. October, pp. 10–13, 2017, doi: 10.5120/ijca2017915495.
[18] M. Guntara, "Komparasi Kinerja Label-Encoding dengan One-Hot-Encoding pada Algoritma K-Nearest Neighbor menggunakan Himpunan Data Campuran Performance Comparison of Label-Encoding with One-Hot-Encoding Methode on K-Nearest Neighbor Algorithm with Mixed Type Data Set," no. 2, pp. 352–360, 2025, doi: 10.26798/jiko.v9i2.1605.
[19] T. Wongvorachan, S. He, and O. Bulut, "A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining," *Inf.*, vol. 14, no. 1, 2023, doi: 10.3390/info14010054.
[20] M. A. Jassim and S. N. Abdulwahid, "Data Mining preparation: Process, Techniques and Major Issues in Data Analysis," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1090, no. 1, p. 012053, 2021, doi: 10.1088/1757-899x/1090/1/012053.
[21] K. Okoye, J. T. Nganji, J. Escamilla, and S. Hosseini, "Machine learning model (RG-DMML) and ensemble algorithm for prediction of students' retention and graduation in education," *Comput. Educ. Artif. Intell.*, vol. 6, no. January, 2024, doi: 10.1016/j.caeai.2024.100205.