Scientific Journal of Informatics
Vol. 11, No. 3, Aug 2024
p-ISSN 2407-7658 https://journal.unnes.ac.id/journals/sji/index e-ISSN 2460-0040

Analysis of Student Graduation Prediction Using Machine Learning Techniques on an Imbalanced Dataset: An Approach to Address Class Imbalance

Dedy Hermanto[1*], Desy Iba Ricoida[2], Desi Pibriana[3] , Rusbandi[4], Muhammad Rizky Pribadi[5]
[1,4,5]Informatic Department, Faculty of Computer Science and Engineering,
Universitas Multi Data Palembang, Indonesia
[2,3]Information System Department, Faculty of Computer Science and Engineering,
Universitas Multi Data Palembang, Indonesia

Abstract.
Purpose: Machine learning is a key area of artificial intelligence, applicable in various fields, including the prediction of timely graduation. One method within machine learning is supervised learning. However, the results are influenced by the distribution of data, particularly in the case of imbalanced classes, where the minority class is significantly smaller than the majority class, affecting classification performance. Timely graduation from a university is crucial for its sustainability and accreditation. This research aims to identify a suitable method to address the issue of predicting timely graduation by managing class imbalance using SMOTE (Synthetic Minority Oversampling Technique).
Methods: This study uses a five-year dataset with 26 attributes and 1328 records, including status labels. The preprocessing stages involve applying five classification algorithms: Decision Tree (DT), Naive Bayes (NB), Logistic Regression (LR), K-Nearest Neighbors (KNN), and Random Forest (RF). Each algorithm is used both with and without SMOTE to handle the class imbalance. The dataset indicates that 60.84% of the cases represent timely graduations. To mitigate the imbalance, over/under-sampling methods are employed to balance the data. The evaluation metric used is the confusion matrix, which assesses the classification performance.
Result: Without SMOTE, the accuracies were 89.12% for DT, 79.65% for NB, 89.47% for LR, 87.72% for KNN, and 90.88% for RF. With SMOTE, the accuracies were 88.89% for DT, 81.48% for NB, 91.05% for LR, 92.59% for KNN, and 89.81% for RF. The algorithms NB, LR, and KNN showed improvement with SMOTE, with KNN yielding the best results.
Novelty: Based on the comparison results, a comparison of five algorithms with and without SMOTE can reasonably classify several of the algorithms being compared.

Keywords: Classification, Machine learning, SMOTE, Timely graduation, University
Received May 2024 / **Revised** June 2024 / **Accepted** June 2024

INTRODUCTION

Universities represent the pinnacle of educational institutions, offering tertiary education across three stages: Bachelor (S1), Master (S2), and Doctoral (S3) levels. Accreditation in higher education serves as a means for universities to validate their operations and gain recognition for their contributions. A crucial aspect of accreditation is the timely graduation of students, a criterion assessed by the National Accreditation Board for Higher Education (BAN-PT) or the Institute for Independent Accreditation (LAM), the designated authorities responsible for evaluating universities. Timely graduation entails students completing their studies within a maximum period of four years [1]. Indonesia boasts 4,432 universities and 42,913 study programs [2], encompassing universities, institutes, polytechnics, and academies. Given the significance of timely graduation as an evaluation parameter impacting university accreditation assessments, it becomes imperative for these institutions to establish appropriate criteria for selecting prospective students capable of graduating within the stipulated timeframe. As previously written in the higher education accreditation guidelines, educational effectiveness and productivity are measured by the average length of study and timely graduation in a higher education institution [3]–[7].

*Corresponding author.
Email addresses: dedy@mdp.ac.id (Dedy)
DOI: 10.15294/sji.v11i3.5528

Assessing a university's success involves comparing the punctuality of graduates with the academic performance of students who fail to meet required standards. This aspect is explored in a study [8], which predicts the likelihood of student dropout by categorizing students into clusters based on their propensity to drop out. Decision-making can involve observing student lectures or developing predictions based on provided inputs. Previous research by other scholars has investigated graduation punctuality, linking forecasted outcomes to student attributes [9]. Moreover, predictions are formulated by incorporating input judgments derived from machine learning techniques [10]–[13]. Researchers have also previously examined timely graduation using an algorithm without SMOTE [14].

This research employs DecisionTree [15]–[17], Naive Bayes [18], Logistic Regression, KNN [19], [20], and Random Forest [21], [22] classification techniques to investigate the timely completion rate of academic programs, building on previous research conducted using these five methods [14]. After data cleaning, the dataset will be analyzed using these techniques, with the Synthetic Minority Over-sampling Technique (SMOTE) employed for comparison. The Decision Tree, Naive Bayes, Logistic Regression, KNN, and Random Forest algorithms are utilized to assess system accuracy, with additional metrics such as precision, recall, and F1 score considered in the evaluation.

METHODS

This research was motivated by the impact of timely graduation on university accreditation. Universities strive to ensure timely graduation to meet accreditation standards, prompting them to seek the best prospective students to achieve this goal. However, private universities face challenges in attracting top prospective students, as state universities are the preferred choice for most students in Indonesia. To address this issue, one approach is to predict the attributes that influence a student's timely graduation. The first step is to develop a tool to identify the most suitable student candidates or provide guidance to help students graduate on time.

The proposed framework in this study consists of five phases: data collection, data labeling, preprocessing, modeling, and machine learning (ML) categorization. This approach involves dividing the dataset into two subsets: 80% for training data and 20% for testing data. The outcomes of this study encompass assessment findings in terms of accuracy, precision, recall, and F1-Score. Figure 1 depicts the sequential steps of the proposed framework utilized in this research.

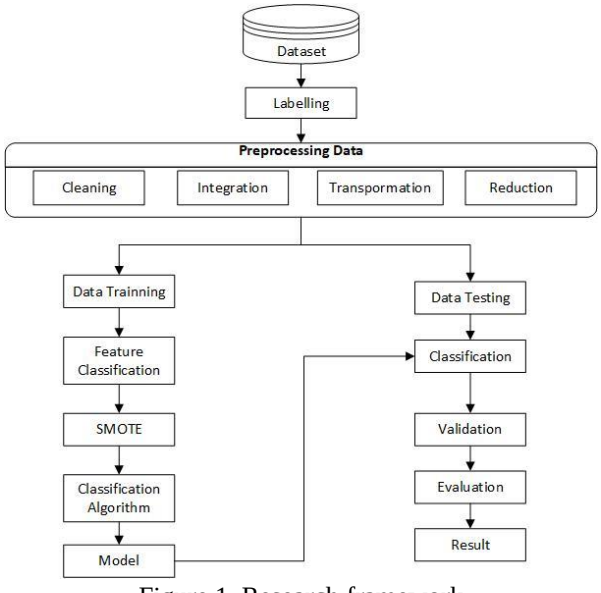

Figure 1. Research framework

Dataset

This research utilizes data from graduate students at the University of Palembang spanning the years 2015 to 2019 across five undergraduate study programs. The characteristics examined in this study include the duration of attendance at the university, gender, place of education, study plan, Semester Achievement Index (IPS), SKS for social studies, and credits taken in lecture sessions. Previous research on timely graduation has explored various attributes for predictive purposes.

In a study by [23], variables such as experience, gender, language proficiency, and skills were employed for predictions in e-learning system user profiling. Another study [11] investigated attributes including gender, study program, school type, primary education institution, region, grade, and graduation year for predictive purposes. Similarly, research by [24] focused on predicting course completion, GPA, gender, place of origin, and study program, utilizing comparable attributes. Consequently, this research incorporates six main attributes: university tenure, gender, school of origin, study program, social studies credits, and total credits.

The IPS and SKS attributes are further subdivided into ten attributes each (IPS1-IPS10 and SKS1-SKS10). This study focuses on data from a single university in Palembang, categorized by study program, which will be integrated into the model under development. The data collection process and the features utilized are summarized in Table 1, following established conventions in prior research.

Table 1. Initial input variables

No.	Variable	Description
1	Enter_year (Ey)	The year in which students entered the University
2	Time (tm)	Distinction between morning (M) and afternoon (N) lecture hours
3	Gender (gd)	Gender categorization: male (M) or female (F)
4	School_type (schT)	Classification of students' school location: private (Pr) or public (Pu)
5	Major (Mj)	Student's chosen department or major, including IF (Informatics), SI (Information System), TE (Electronic Engineering), MJ (Management), and AK (Accounting)
6	IPS1 – 10	Student achievement index scores for the 1st through 10th semesters
7	SKS1 – 10	Course credits received by students for each semester (1 – 24)
8	Status (st)	Graduation status based on GPA and study period, categorized as Graduated (0) and Not Graduated (1)

Preprocessing

The dataset collection process yielded a total of 1,307 rows of data. The preprocessing stage begins with (i) data collection, which involves extracting data from the academic domain through the Learning Management System (LMS) application owned by MDP University, specifically the lecture information system, to obtain initial data. This dataset comprises 26 variables utilized in this research, as detailed in Table 1.

The initial step involves gathering academic data from graduate students spanning 2015 to 2019, resulting in a total of 1,328 rows used in this study, with 808 instances of non-timely graduation (60.84%) and 520 instances of timely graduation (39.16%), indicating an imbalance in the dataset [25]–[27]. (ii) Data cleaning follows, which involves identifying and handling inappropriate or missing data. Unsuitable data is removed from the academic data repository, and missing values are imputed using an average value approach. Specifically, missing values are replaced with the calculated average value. Table 2 presents information on average data in row 1, with rows 2 to 10 indicating instances lacking necessary fields or being empty, often due to students dropping out after the first semester due to poor performance, necessitating the removal of such outlier data. (iii) Data transformation is the next stage, involving compilation of data and assignment of pass or fail labels as per machine learning requirements. Lastly, (iv) data reduction involves partitioning the dataset into training data (80%) and test data (20%), utilizing the entire dataset to develop a training data model.

Table 2. Missing value

No.	*Ey*	*Tm*	*Gd*	*schT*	*Mj*	*IPS1*	*IPS2*	***.....***	*IPS9*	*IPS10*	*SKS1*	*SKS2*	***.....***	*SKS9*	*SKS10*	*st*
1	2015	M	M	Pu	SI	2.58	3	***.....***	1.28	0	19	21	***.....***	12	0	0
2	2015	M	M	Pr	SI	0.42	0	***.....***	0	0	19	0	***.....***	0	0	1
3	2015	N	M	Pr	SI	0.95	0	***.....***	0	0	19	0	***.....***	0	0	1
4	2015	N	M	Pr	SI	2.05	0	***.....***	0	0	19	0	***.....***	0	0	1
5	2016	M	M	Pu	IF	2.35	0	***.....***	0	0	20	0	***.....***	0	0	1
6	2016	M	M	Pu	IF	2.34	3.09	***.....***	0	0	20	18	***.....***	0	0	1
7	2016	M	M	Pr	IF	1.66	0	***.....***	0	0	20	0	***.....***	0	0	1
8	2017	N	F	Pr	SI	3.24	0	***.....***	0	0	19	0	***.....***	0	0	1
9	2017	N	M	Pr	SI	1.96	0	***.....***	0	0	19	0	***.....***	0	0	1
10	2017	M	F	Pu	SI	1.38	0	***.....***	0	0	19	0	***.....***	0	0	1

SMOTE (Synthetic Minority Over-sampling Technique)

This approach is employed to enhance the accuracy of study findings [28], [29]. The primary objective of the SMOTE method is to mitigate the class imbalance issue observed in research. The SMOTE technique involves generating synthetic data by duplicating minority data through data resampling. There are two processes [30]: (i) undersampling and (ii) oversampling. Here is an elucidation of SMOTE, a technique used in machine learning: (i) Calculate the difference between a sample and its closest neighboring sample, (ii) multiply this difference by a random integer ranging from 0 to 1, (iii) add this adjusted difference to the original sample to create a new synthetic example in the feature space, (iv) repeat this process with subsequent nearest neighbors until the desired number is reached, as specified by the user. The most commonly used approach in SMOTE is oversampling, which generates synthetic samples to balance the dataset. An example illustrating the process of SMOTE is presented in Figure 2.

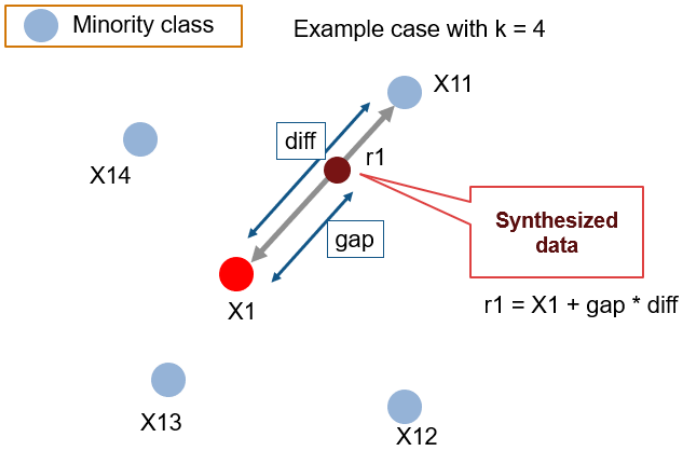

Figure 2. SMOTE working procedure

SMOTE is utilized to address imbalanced datasets in various research domains [31]–[33]. The current research employs the SMOTE algorithm to enhance the quality of test data for model generation. The process involved in the SMOTE algorithm for handling imbalanced data is depicted in Figure 3. The procedure begins with utilizing the available data, detailed in the preprocessing section, comprising a total of 1,328 entries with 60.84% representing graduates not meeting timelines and 39.16% representing timely graduates. This imbalanced dataset is then utilized in the research employing the SMOTE technique for oversampling. Figure 4 illustrates the dataset before and after applying SMOTE.

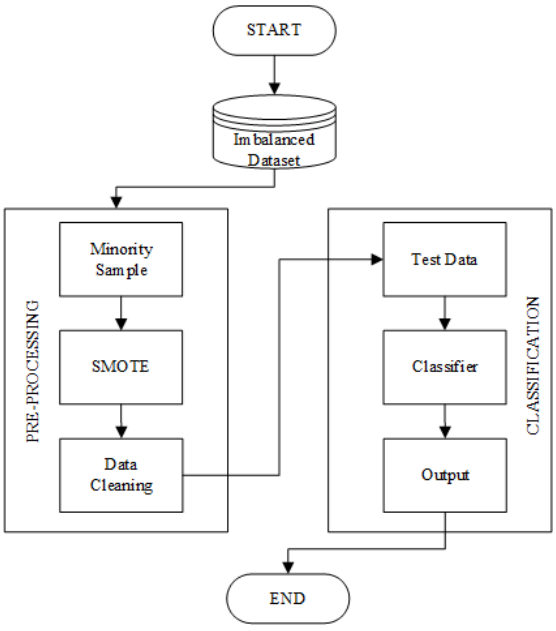

Figure 3. SMOTE step process to classification

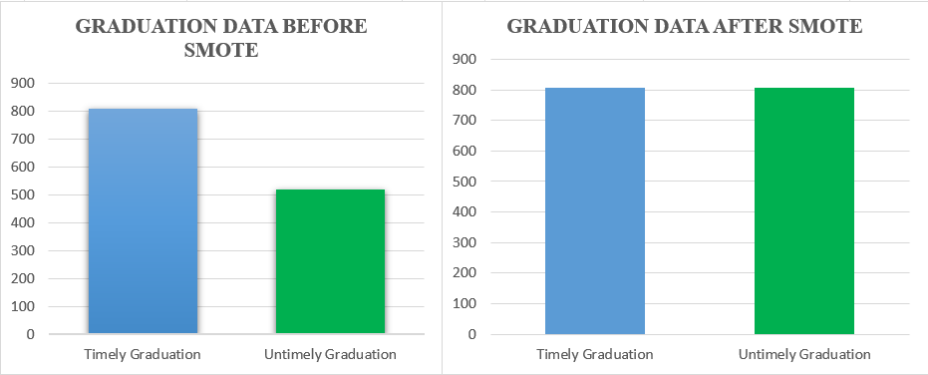

Figure 4. Total data before and after SMOTE

Feature selection

Feature selection is conducted to determine the most relevant features for this research. Researchers utilized the Chi-Square method, a widely recognized technique in machine learning research [34]–[37]. Initially, 26 features were considered for evaluation, and their suitability was assessed using the Chi-Square method. The obtained results for this weighting involve calculations between observed frequency and expected frequency, which are elements of the Chi-Square method, as presented in equation (1):

$$X_c^2 = \sum \frac{(Oi-Ei)^2}{Ei} \quad (1)$$

Description:

X = chi-square statistic O = observed frequency E = expected frequency

Nineteen influential features were identified through the Chi-Square method, refining the initial set of twenty-six. These selected features significantly impact classification tasks when using machine learning techniques, particularly with carefully chosen classification algorithms such as Naive Bayes or Random Forest. The weight values derived from the Chi-Square technique for this study are detailed in Table 3.

Table 3. Attribute weighting results

No.	Variable	Weight
1	Enter_year	52.51
2	Time	22.70
3	Gender	53.68
4	School_type	69.31
5	Major	49.64
6	IPS1	369.82
7	IPS2	385.07
8	IPS3	448,24
9	IPS4	411,50
10	IPS5	411,79
11	IPS6	406,41
12	IPS7	257.03
13	IPS8	284.19
14	IPS9	258.64
15	IPS10	108.85
16	SKS1	26.51
17	SKS2	334.23
18	SKS3	403.28
19	SKS4	415.56
20	SKS5	369.59
21	SKS6	187.11
22	SKS7	222.37
23	SKS8	426.00
24	SKS9	256.19
25	SKS10	160.47
26	Status	0

Evaluation

This step incorporated an evaluation technique employed in this study, specifically utilizing the Confusion Matrix methodology. The Confusion Matrix serves to evaluate the efficacy of categorization [38], [39]. This method produces several computed metrics, including precision, recall, accuracy, and F1 Score. These computations are essential due to the interdependencies among the requisite data for each step. The results of this evaluation are illustrated through equations (2), (3), (4), and (5). Metrics such as precision, recall, accuracy, and F1 Score are utilized to assess the efficacy of data testing throughout the study. Accuracy assesses the alignment of research predictions with actual outcomes, while recall evaluates the similarity between datasets. The F1 Score signifies the balance between Precision and Recall values and is employed in studies to gauge the consistency between these metrics. The Confusion Matrix methodology scrutinizes the performance of classification, encompassing positive and negative classifications, resulting in true positive, false positive, true negative, and false negative outcomes. These outcomes delineate the classification of data. The Confusion Matrix results from data processing are presented in Table 4, encompassing four distinct components.

Table 4. Confusion matrix

		Predicted Values	
		Positive	Negative
Actual Values	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Negative (FN)	True Negative (TN)

The parameters from Table 4 have the following meanings: True Positive (TP) is a value that is correctly identified as positive. False Positive (FP) is a value that is not true but is identified as a positive. False Negative (FN) is a value that is true but is considered a negative. True Negative (TN) is a negative value and is correctly considered negative.

$$Precesion = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$Accuracy = \frac{TP+TN}{TP++TN+FP+FN} \quad (4)$$

$$F1\ Score = 2\ X\ \frac{Precision\ x\ Recall}{Precision+Recall} \quad (5)$$

RESULTS AND DISCUSSIONS

This research utilizes a dataset from academic records spanning active students up to the graduating class of 2015-2019. Machine learning algorithms are employed as a common approach to predict timely graduation among students [40] [23], [41], [42]. Among various machine learning methods, classification stands out. In this study, classification is implemented using Naive Bayes and Random Forest algorithms. What sets this research apart is the integration of these algorithms, aiming to enhance data processing performance using the SMOTE algorithm. The machine learning methods employed herein include Decision Tree (DT), Naive Bayes (NB), Logistic Regression (LR), K-Nearest Neighbors (KNN), and Random Forest (RF). The results of this methodological fusion are presented in Table 5, showcasing the evaluation results of data testing without leveraging SMOTE and illustrating the outcomes of data testing with SMOTE.

Table 1. Comparison of results with and without leveraging SMOTE

Classifier	SMOTE	Accuracy	Precision	Recall	AUC	F1
Decision Tree	Without	89.12	80.26	99.19	95.1	88.72
	With	88.89	83.16	97.53	90.5	89.78
Naive Bayes	Without	79.65	91.14	58.54	92.7	71.28
	With	81.48	98.11	64.2	94.6	77.6
Logistic Regression	Without	89.47	89.08	86.18	95.1	87.61
	With	91.05	91.3	90.74	97.0	91.02
KNN	Without	87.72	86.07	85.73	95.7	85.89
	With	92.59	89.20	96.91	97.1	92.89
Random Forest	Without	90.88	85.93	94.31	97.2	89.92
	With	89.81	83.77	98.77	98.5	90.71

The results presented in this research were also discussed in prior studies [14] without the use of SMOTE, employing five existing algorithms. The current research extends these findings by applying SMOTE to all previously selected algorithms. The confusion matrix testing conducted in this research, detailed in Table 5, illustrates the outcomes of data testing both without leveraging SMOTE and with leveraging SMOTE. These results demonstrate improved accuracy when SMOTE is integrated, with KNN achieving the highest improvement among the algorithms tested. Specifically, KNN achieved an accuracy of 87.72% without leveraging SMOTE and 92.59% with leveraging SMOTE, resulting in a 4.87% increase in accuracy upon integrating the SMOTE algorithm with KNN. An example illustrating these accuracy improvements for KNN is presented in Figure 5.

accuracy: 87.72%

	true 0	true 1	class precision
pred. 0	145	18	88.96%
pred. 1	17	105	86.07%
class recall	89.51%	85.37%	

accuracy: 92.59%

	true 0	true 1	class precision
pred. 0	143	5	96.62%
pred. 1	19	157	89.20%
class recall	88.27%	96.91%	

Figure 5. KNN classifier accuracy without and with leveraging SMOTE

Figure 6 displays the Receiver Operating Characteristic (ROC) curve and the Area Under Curve (AUC) using the KNN Classifier algorithm with SMOTE integration. The AUC obtained from these results is 0.971, corresponding to 97.1%.

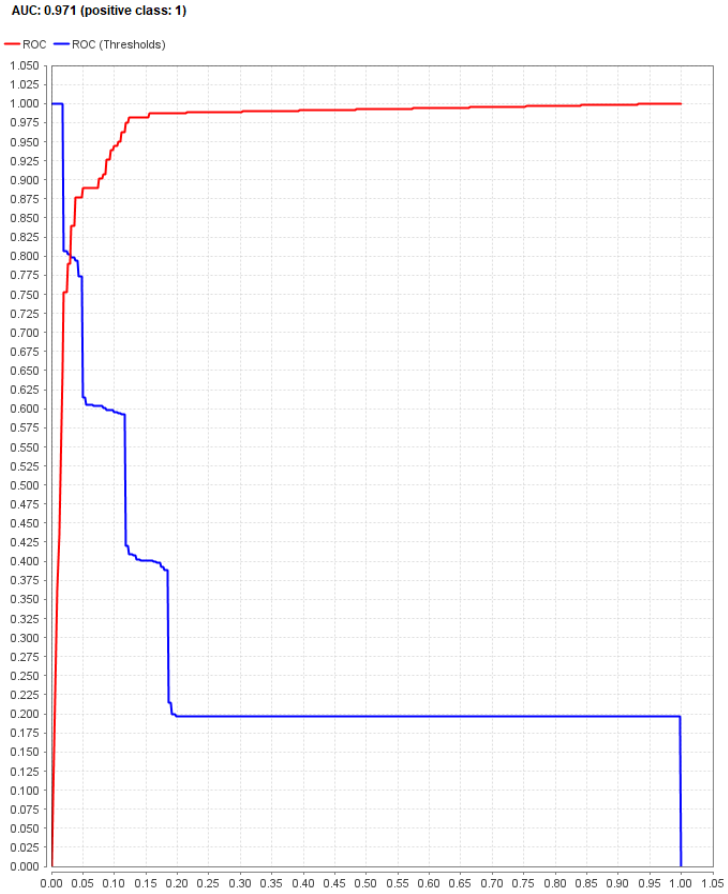

Figure 6. KNN classifier AUC leveraging SMOTE

Calculate the F1 Score using the example results from the KNN Classifier without SMOTE (result 6) and with SMOTE (result 7).

$$F1\ Score = 2\ X\ \frac{Precision\ x\ Recall}{Precision + Recall}$$

$$F1\ Score = 2\ X\ \frac{0.8607\ x\ 0.8573}{0.8607\ + 0.8573}$$

$$F1\ Score = 2\ X\ \frac{0.7378}{1.718}$$

$$F1\ Score = 0.8589 \text{ or } 85{,}89\% \quad (6)$$

$$F1\ Score = 2\ X\ \frac{Precision\ x\ Recall}{Precision + Recall}$$

$$F1\ Score = 2\ X\ \frac{0.8920\ x\ 0.9691}{0.8920\ + 0.9691}$$

$$F1\ Score = 2\ X\ \frac{0.8644}{1.8611}$$

$$F1\ Score = 0.9289 \text{ or } 92.89\% \quad (7)$$

CONCLUSION

This research employed machine learning methods to predict timely graduation from a university. The stages of data collection and model building, both with and without leveraging SMOTE, yielded varying results. Without leveraging SMOTE, the accuracy rates were as follows: Decision Tree (DT) 89.12%, Naive Bayes (NB) 79.65%, Logistic Regression (LR) 89.47%, K-Nearest Neighbors (KNN) 87.72%, and Random Forest (RF) 90.88%. When SMOTE was leveraged, the accuracies were: DT 88.89%, NB 81.48%, LR 91.05%, KNN 92.59%, and RF 89.81%.

Notably, leveraging SMOTE led to increased accuracy in several classifiers: NB improved from 79.65% to 81.48%, LR from 89.47% to 91.05%, and KNN achieved the highest improvement, rising from 87.72% to 92.59%. The KNN classifier showed the highest accuracy when SMOTE was applied.

REFERENCES

[1] BAN-PT, "National Accreditation Board for Higher Education," 2022.

[2] KEMDIKBUD, "Server PDDIKTI - Online Database," 2022.

[3] A. P. Studi *et al.*, *Criteria And Procedure For Program Study Accreditation Bachelor Of Pharmacy*. 2019.

[4] A. P. Studi *et al.*, *Performance Document Assessment Guidelines And Matrix And Self-Evaluation Report On Accreditation Of The Bachelor Of Pharmacy Study Program*. 2019.

[5] LAM-INFOKOM, *Preparation Guide Self-Evaluation Report*, no. November. 2020.

[6] LAM-TEKNIK, *Guidelines For Preparing Self-Evaluation Reports Accreditation Of Academic And Vocational Study Programs*, no. 5. 2021.

[7] LAMEMBA, *LAMEMBA Study Program Accreditation Instrument*. 2021.

[8] W. Purba, S. Tamba, and J. Saragih, "The effect of mining data k-means clustering toward students profile model drop out potential," *J. Phys. Conf. Ser.*, vol. 1007, no. 1, 2018, doi: 10.1088/1742-6596/1007/1/012049.

[9] F. N. Al Amin, Indahwati, and Y. Anggraini, "Analysis of Timeliness of Graduation Based on Characteristics of FEM and Faperta Students Using the Chart Method," *Xplore*, vol. 2, no. 1, pp. 1–8, 2013.

[10] F. Ouatik, M. Erritali, F. Ouatik, and M. Jourhmane, "Predicting Student Success Using Big Data and Machine Learning Algorithms," *Int. J. Emerg. Technol. Learn.*, vol. 17, no. 12, pp. 236–251, 2022, doi: 10.3991/ijet.v17i12.30259.

[11] C. Wirawan, E. Khudzaeva, T. H. Hasibuan, Karjono, and Y. H. K. Lubis, "Application of Data

mining to Prediction of Timeliness Graduation of Students (A Case Study)," *2019 7th Int. Conf. Cyber IT Serv. Manag. CITSM 2019*, pp. 18–21, 2019, doi: 10.1109/CITSM47753.2019.8965425.
[12] M. Anwar, "Prediction of the graduation rate of engineering education students using Artificial Neural Network Algorithms," *Int. J. Res. Couns. Educ.*, vol. 5, no. 1, p. 15, 2021, doi: 10.24036/00411za0002.
[13] M. Ridwan, I. Sembiring, A. Setiawan, and I. Setyawan, "Analysis of Attack Detection on Log Access Servers Using Machine Learning Classification: Integrating Expert Labeling and Optimal Model Selection," *Sci. J. Informatics*, vol. 11, no. 1, pp. 119–126, 2024, doi: 10.15294/sji.v11i1.49424.
[14] D. Iba Ricoida, D. Hermanto, D. Pibriana, and M. Rizky Pribadi, "Comparison of the Accuracy of Data Mining Algorithms in Predicting Graduation on Time," vol. 7, no. 2, pp. 69–76, 2024.
[15] A. Rifan Rudiyanto and M. Arief Soeleman, "Performance of the Decision Tree Algorithm in the Classification of Edible and Poisonous Mushrooms with Information Gain Optimization," *Sci. J. Informatics*, vol. 10, no. 4, pp. 549–558, 2023, doi: 10.15294/sji.v10i4.47864.
[16] U. Salamah, A. A. Ningrum, and M. Yunianto, "Classification of Early-Stage Lung Cancer Based on First and Second Order Statistical Variations Using the Gaussian Smoothing Filter and Decision Tree Method," vol. 10, no. 4, pp. 559–570, 2023, doi: 10.15294/sji.v10i4.48026.
[17] L. Alfaris, R. C. Siagian, A. C. Muhammad, and U. I. Nyuswantoro, "Classification of Spiral and Non-Spiral Galaxies using Decision Tree Analysis and Random Forest Model : An Investigation of the Zoo Galaxy Dataset," vol. 10, no. 2, pp. 139–150, 2023, doi: 10.15294/sji.v10i2.44027.
[18] R. Wibowo, M. A. Soeleman, and A. Affandy, "Hybrid Top-K Feature Selection to Improve High-Dimensional Data Classification Using Naïve Bayes Algorithm," *Sci. J. Informatics*, vol. 10, no. 2, pp. 113–120, 2023, doi: 10.15294/sji.v10i2.42818.
[19] N. A. Saputra, K. Aeni, and N. M. Saraswati, "Indonesian Hate Speech Text Classification Using Improved K-Nearest Neighbor with TF-IDF- ICSρF," *Sci. J. Informatics*, vol. 11, no. 1, pp. 21–30, 2024, doi: 10.15294/sji.v11i1.48085.
[20] S. Lonang, A. Yudhana, and M. K. Biddinika, "Performance Analysis for Classification of Malnourished Toddlers Using K-Nearest Neighbor," *Sci. J. Informatics*, vol. 10, no. 3, pp. 313–322, 2023, doi: 10.15294/sji.v10i3.45196.
[21] A. Khairunnisa, K. A. Notodiputro, and B. Sartono, "A Comparative Study of Random Forest and Double Random Forest Models from View Points of Their Interpretability," *Sci. J. Informatics*, vol. 11, no. 1, pp. 207–218, 2024, doi: 10.15294/sji.v11i1.48721.
[22] A. D. Goenawan and S. Hartati, "The Comparison of K-Nearest Neighbors and Random Forest Algorithm to Recognize Indonesian Sign Language in a Real-Time," *Sci. J. Informatics*, vol. 11, no. 1, pp. 237–244, 2024, doi: 10.15294/sji.v11i1.48475.
[23] A. Maulana, "Prediction of student graduation accuracy using decision tree with application of genetic algorithms," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1073, no. 1, p. 012055, 2021, doi: 10.1088/1757-899x/1073/1/012055.
[24] A. Anwarudin, W. Andriyani, B. P. DP, and D. Kristomo, "The Prediction on the Students' Graduation Timeliness Using Naive Bayes Classification and K-Nearest Neighbor," *J. Intell. Softw. Syst.*, vol. 1, no. 1, p. 75, 2022, doi: 10.26798/jiss.v1i1.597.
[25] C. Magnolia, A. Nurhopipah, and B. A. Kusuma, "Handling Imbalanced Dataset for Classifying Independent Campus Program Comments on the Twitter Application," *Edu Komputika J.*, vol. 9, no. 2, pp. 105–113, 2023, doi: 10.15294/edukomputika.v9i2.61854.
[26] W. Ustyannie and S. Suprapto, "Oversampling Method To Handling Imbalanced Datasets Problem In Binary Logistic Regression Algorithm," *IJCCS (Indonesian J. Comput. Cybern. Syst.*, vol. 14, no. 1, p. 1, 2020, doi: 10.22146/ijccs.37415.
[27] C. Kaope and Y. Pristyanto, "The Effect of Class Imbalance Handling on Datasets Toward Classification Algorithm Performance," *MATRIK J. Manajemen, Tek. Inform. dan Rekayasa Komput.*, vol. 22, no. 2, pp. 227–238, 2023, doi: 10.30812/matrik.v22i2.2515.
[28] R. Siringoringo, "Unbalanced data classification using the SMOTE algorithm and k-nearest neighbors," *J. ISD*, vol. 3, no. 1, pp. 44–49, 2018.
[29] M. R. Pribadi, D. Manongga, H. D. Purnomo, I. Setyawan, and Hendry, "Sentiment Analysis of the PeduliLindungi on Google Play using the Random Forest Algorithm with SMOTE," *2022 Int. Semin. Intell. Technol. Its Appl. Adv. Innov. Electr. Syst. Humanit. ISITIA 2022 - Proceeding*, no. July, pp. 115–119, 2022, doi: 10.1109/ISITIA56226.2022.9855372.
[30] R. Muzayanah, A. D. Lestari, J. Jumanto, B. Prasetiyo, D. A. A. Pertiwi, and M. A. Muslim, "Comparative Study of Imbalanced Data Oversampling Techniques for Peer-to-Peer Landing Loan

Prediction," *Sci. J. Informatics*, vol. 11, no. 1, pp. 245–254, 2024, doi: 10.15294/sji.v11i1.50274.
[31] A. Fernández, S. Garcia, F. Herrera, and N. V Chawla, "SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary," *J. Artif. Intell. Res.*, vol. 61, pp. 863–905, 2018.
[32] C. Zhang, Y. Chen, X. Liu, and X. Zhao, "Abstention-SMOTE: An over-sampling approach for imbalanced data classification," *ACM Int. Conf. Proceeding Ser.*, pp. 17–21, 2017, doi: 10.1145/3176653.3176676.
[33] K. Savetratanakaree, K. Sookhanaphibarn, S. Intakosum, and R. Thawonmas, "Borderline over-sampling in feature space for learning algorithms in imbalanced data environments," *IAENG Int. J. Comput. Sci.*, vol. 43, no. 3, pp. 363–373, 2016.
[34] A. Sikri, N. P. Singh, and S. Dalal, "INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING Chi-Square Method of Feature Selection : Impact of Pre- Processing of Data," *Intell. Syst. Appl. Issn2147-6799*, vol. 11, pp. 241–248, 2023.
[35] K. M. A. Uddin *et al.*, "Machine Learning-Based Screening Solution for COVID-19 Cases Investigation: Socio-Demographic and Behavioral Factors Analysis and COVID-19 Detection," *Human-Centric Intell. Syst.*, no. October, 2023, doi: 10.1007/s44230-023-00049-9.
[36] O. S. Bachri, Kusnadi, M. Hatta, and O. D. Nurhayati, "Feature selection based on CHI square in artificial neural network to predict the accuracy of student study period," *Int. J. Civ. Eng. Technol.*, vol. 8, no. 8, pp. 731–739, 2017.
[37] M. Lezoche, J. E. Hernandez, M. del M. E. Alemany Díaz, H. Panetto, and J. Kacprzyk, "Agri-food 4.0: A survey of the supply chains and technologies for the future agriculture," *Comput. Ind.*, vol. 117, p. 103187, May 2020, doi: 10.1016/j.compind.2020.103187.
[38] Y. Yohannes, M. R. Pribadi, and L. Chandra, "Klasifikasi Jenis Buah dan Sayuran Menggunakan SVM Dengan Fitur Saliency-HOG dan Color Moments," *Elkha*, vol. 12, no. 2, p. 125, 2020, doi: 10.26418/elkha.v12i2.42160.
[39] A. Veluchamy, H. Nguyen, M. L. Diop, and R. Iqbal, "Comparative Study of Sentiment Analysis with Product Reviews Using Machine Learning and Lexicon-Based Approaches," *SMU Data Sci. Rev.*, vol. 1, no. 4, pp. 1–22, 2018.
[40] N. Mohammad Suhaimi, S. Abdul-Rahman, S. Mutalib, N. H. Abdul Hamid, and A. Md Ab Malik, *Predictive Model of Graduate-On-Time Using Machine Learning Algorithms*, vol. 1100, no. September. Springer Singapore, 2019. doi: 10.1007/978-981-15-0399-3_11.
[41] A. Syahputra, R. M. A. K. Rasyid, I. Istiningsih, A. Surwiyanta, H. Q. Djojodibroto, and E. Anan, "Prediction Of Graduation Time For Students Faculty Of Computer Science Amikom University Yogyakarta Using Naïve Bayes Method," *Conf. Ser.*, vol. 4, no. January, pp. 142–151, 2022, doi: 10.34306/conferenceseries.v4i1.709.
[42] J. Letkiewicz, H. Lim, S. Heckman, S. Bartholomae, J. Fox, and C. Montalto, "The path to graduation: Factors predicting on-time graduation rates," *J. Coll. Student Retent. Res. Theory Pract.*, vol. 16, no. 3, pp. 351–371, 2014, doi: 10.2190/CS.16.3.c.