DIJEMSS: Dinasti International Journal of Education Management and Social Science
E-ISSN: 2686-6331
P-ISSN: 2686-6358
https://dinastipub.org/DIJEMSS dinasti.info@gmail.com +62 811 7404 455
DOI: https://doi.org/10.38035/dijemss.v6i6

Comparative Analysis of Gradient Boosting, XGBoost, and KNN on Predicting Student Graduation in Imbalance and Balance Data Schemes

Muhammad Rizki Hubu[1], Irfan Pratama[2]
[1]Universitas Mercu Buana Yogyakarta, Yogyakarta, Indonesia, rizkihubu20@gmail.com
[2]Universitas Mercu Buana Yogyakarta, Yogyakarta, Indonesia, irfanp@mercubuana-yogya.ac.id

Corresponding Author: rizkihubu20@gmail.com[1]

Abstract: The objective of this research is to compare the performance of three machine learning algorithms: Gradient Boosting, XGBoost, and K-Nearest Neighbors (KNN) in predicting student graduation using a quantitative approach and comparative experimental methods. The analysis process follows the CRISP-DM stages, which include business understanding, data understanding, data preparation, modeling, evaluation, and implementation. The dataset used consists of approximately 1,251 data points from students of the 2018–2020 cohort with an imbalanced distribution, namely 73.78% graduated on time and 6.22% did not graduate on time. The variables analyzed include academic and non-academic data, such as total credits, GPA per semester, number of repeated courses, and number of leaves. To address the data imbalance, the SMOTE-TOMEK balancing technique was applied. The results of this research indicate that XGBoost showed an improvement in performance after balancing, with accuracy, precision, recall, and F1-score reaching 1.0000. Gradient Boosting shows consistent performance with a score of 0.9992, both before and after balancing. KNN also experienced an increase in accuracy from 0.9928 to 0.9968 after the balancing process. Findings from the confusion matrix results show a significant improvement in classification. Therefore, the implementation of the SMOTE-TOMEK technique has proven effective in improving the performance of classification models on imbalanced data, and XGBoost is recommended as the main algorithm for predicting student graduation.

Keywords: Graduation Prediction, Data Imbalance, Gradient Boosting, XGBoost.

INTRODUCTION

Timely student graduation is essential to measure the quality and effectiveness of higher education. The Academic Information System (SIAKAD) has become an important tool in monitoring and evaluating student learning outcomes periodically. Data generated by SIAKAD, such as Cumulative Achievement Index (IPK), attendance, and other student activities, can be used as a basis for predicting the possibility of timely graduation. [1] However, to predict timely graduation with SIAKAD data, there are still several problems in practice. One of the main problems is data imbalance, where fewer students graduate on time

than those who do not, thus affecting the performance of the prediction model. [2] In addition, data quality is also an obstacle, such as missing data, inconsistent recording, and limited variables available in the system, which ultimately reduce the accuracy of the prediction model.

On the other hand, the integration of non-academic data such as psychological conditions, motivation, and socio-economic factors of students is still rarely utilized optimally in SIAKAD, even though these factors contribute significantly to the success of student studies. Therefore, the development of a timely graduation prediction model requires a multidimensional approach that combines academic and non-academic data more comprehensively [3]. In general, machine learning algorithms fall into three main categories: supervised learning, unsupervised learning, and reinforcement learning [4]. Jingjing Wang, Chunxiao Jiang, and Senior Member (2019). Each category has a different approach to learning patterns from data. Supervised learning is one of the most commonly used machine learning approaches [6]. In this method, the learning process is carried out in a directed or guided manner, where the computer is trained using a labeled dataset [7]. This dataset acts as a guide (training dataset), so that the algorithm can learn the relationship between input and output, and produce predictions or classifications according to the expected target [4].

The model is trained with labeled data so that it can identify patterns and produce accurate predictions when it finds new data [8]. In the context of higher education in Indonesia, student graduation prediction is becoming increasingly important as the complexity of academic data increases.\ Emphasizing the importance of predictive analytics to improve student learning experiences and institutional strategies. Algorithmic models are even able to increase prediction accuracy by up to 85% compared to conventional methods. Research conducted also shows that machine learning algorithms provide more accurate results than traditional statistical methods in predicting graduation [9].

To address the problem of data imbalance, research conducted by Luque et al., 2019 showed that the data-level approach (resampling) is more effective than the algorithm-based approach [10]. One of the techniques that is considered superior is SMOTE-TOMEK, a method that combines synthetic data generation (SMOTE) with cleaning overlapping data between classes (Luque et al., 2019). This technique obtained the best mean of ranks value (1,700) compared to other resampling methods. In addition, research conducted by Agustina et al., 2024 showed the effectiveness of SMOTE-TOMEK because it was proven to be able to significantly improve the performance of the classification model [11]. In addition, Atlantic et al., 2024 showed that the Gradient Boosting Machine (GBM) algorithm excels in handling numeric and categorical data in the classification of student graduation [12]. The results of their study showed that GBM achieved an accuracy of 71.09%, higher than CART which was only 67.97%.

Thus, the purpose of this study is to analyze and compare the performance of three machine learning algorithms gradient boosting, XGBoost, and K-nearest neighbors (KNN) in predicting student graduation, both in imbalanced and balanced data conditions.

METHOD

Research Stages

To achieve the objectives of this study, the researcher compiled systematic steps starting from the data collection process to the evaluation and analysis stages of the results. This stage is designed to compare the performance of the Gradient Boosting, XGBoost, and K-Nearest Neighbors (KNN) algorithms in predicting student graduation in conditions of unbalanced data and balanced data. The overall research stages can be seen in Figure 1 below:

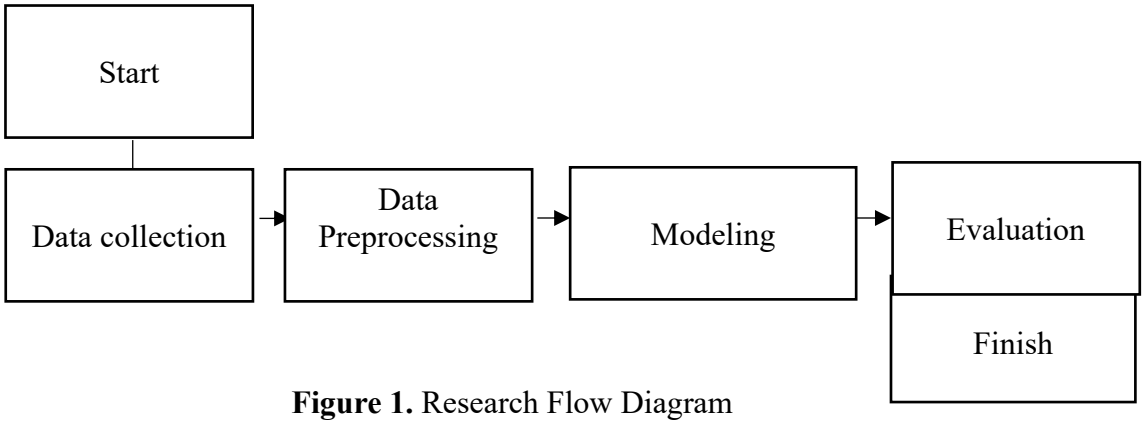

Figure 1. Research Flow Diagram

Data Collection

The data used in this study were obtained from the university's academic system which includes student information, such as demographic data, academic achievement, and graduation status. The data were collected in tabular form and included a number of relevant features such as GPA, number of credits taken, study period, and activity per semester [13]. After the acquisition process, the data was cleaned to overcome missing values and anomalous data. For the analysis needs of the imbalance and balance data schemes, the original data that tends to be imbalanced is then processed using resampling techniques, such as SMOTE-TOMEK, to create balanced data conditions. The entire data collection and pre-processing process is carried out while maintaining the confidentiality and integrity of student data.

Data Pre-Processing

The data pre-processing stage is carried out to ensure the quality of the data used in modeling student graduation predictions. The initial steps include data cleaning, namely overcoming missing values, duplicate data, and inconsistencies in attributes. Furthermore, the data is converted into a numeric format through an encoding technique for categorical variables. Normalization is done so that the scale between features is uniform, especially for the KNN algorithm which is sensitive to distance. Given that the focus of this study is also on imbalance and balance data schemes, balancing techniques such as SMOTE are applied to produce a balanced class distribution. For the model training and evaluation process, the processed data is divided into test data and training data [14].

Modeling

In this study, the modeling process is carried out by comparing three classification algorithms, namely Gradient Boosting, XGBoost, and K-Nearest Neighbors (KNN). Modeling is carried out in two data schemes, namely imbalanced data (without class balancing) and balanced data (using the SMOTE oversampling technique) [15]. Each model is trained using data that has gone through a preprocessing and normalization process. Performance evaluation is carried out using accuracy, precision, recall, and F1 score metrics. The goal is to find the most effective algorithm for predicting student graduation, as well as to determine the effect of data balancing on the performance of each model.

Evaluation

This journal presents a comparative analysis of three machine learning algorithms Gradient Boosting, XGBoost, and K-Nearest Neighbors (KNN) in predicting student graduation, both in imbalanced and balanced data conditions. This research is relevant and important considering the issue of data imbalance often arises in the education domain. Methodologically, the journal has compiled clear steps, starting from preprocessing, data

balancing, to model evaluation with appropriate metrics [2]. However, the evaluation can be strengthened with statistical analysis to show the significance of differences in performance between models [16]. In addition, the explanation regarding parameter selection and model tuning still needs to be clarified so that research replication becomes more open. Overall, this journal makes a good contribution to the study of the application of prediction algorithms in the field of higher education.

RESULT AND DISCUSSION

Data preprocessing

Class distribution before and after balancing

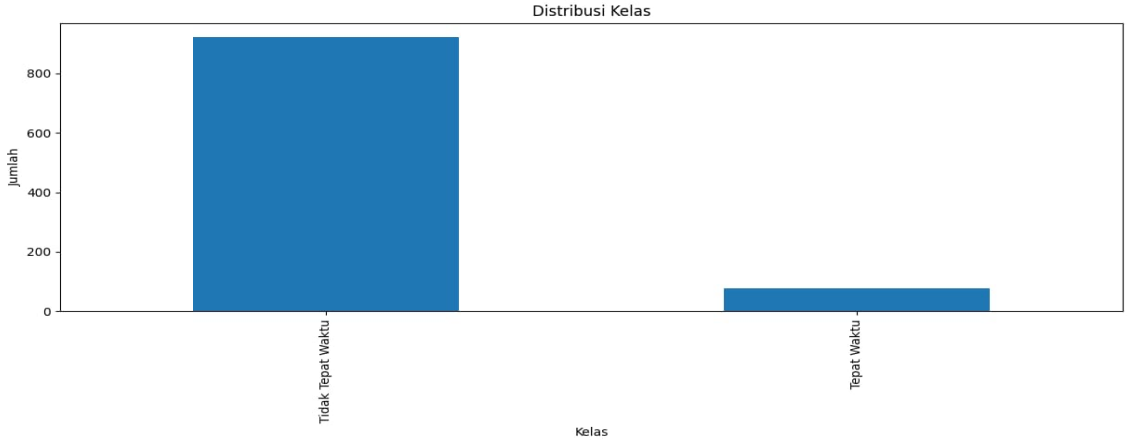

Figure 2. Class distribution before balancing

The results in Figure 2 illustrate the class distribution in the data before balancing. There is a disparity in the number of non-punctual classes which is much more dominant compared to the punctual class. This class imbalance can cause bias in the classification model: the model tends to be more accurate in predicting the majority class but less efficient in predicting the minority class.

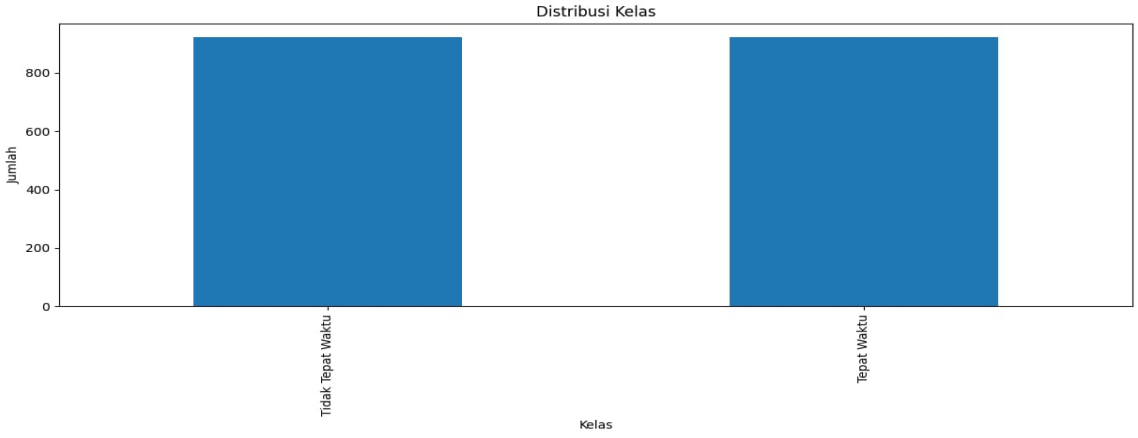

Figure 3. Class distribution after balancing

The results in Figure 3 illustrate the class distribution in the data before balancing. There is a disparity in the number of non-punctual classes which is much more dominant compared to the punctual class. This class imbalance can cause bias in the classification model: the model tends to be more accurate in predicting the majority class but less efficient in predicting the minority class.

Figure 3 shows the class distribution in the data after the balancing process. It can be seen that the number of students with the "On Time" and non-on time types is relatively balanced, each number is almost the same. This balancing process aims to overcome the problem of class imbalance which can affect the performance of the classification model, especially in detecting minority classes. With a balanced class distribution, it is expected that the model has better and fairer prediction capabilities in both classes.

Correlation Between Attributes

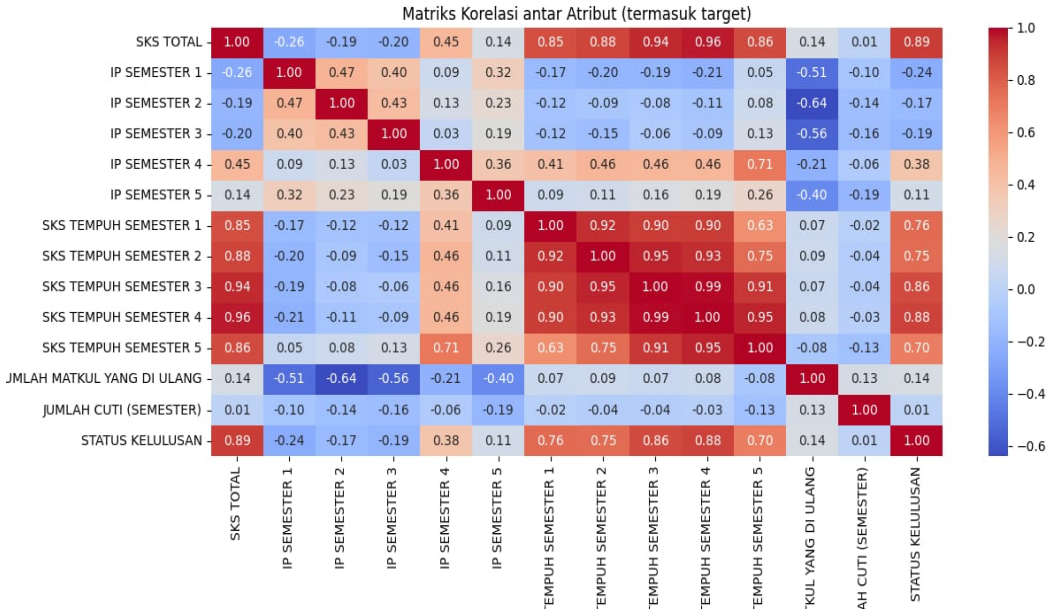

Figure 4. Matrix Corelation

Figure 4 correlation matrix between attributes shows a linear relationship between each input variable and the target student graduation status. The correlation is shown in the range of values between -1 and 1, where higher values indicate a very strong relationship with positive and negative. Based on the analysis results, the variable that has the highest correlation with graduation status is Total Credits (r = 0.89), followed by Credits Taken in Semester 4 (r = 0.88), Credits Taken in Semester 3 (r = 0.86), and Credits Taken in Semester 1 (r = 0.76). This finding shows that the more credits a student successfully completes, the more likely the student is to complete their studies on time. On the other hand, the variable of achievement index (IP) for each semester shows a lower correlation with graduation status, such as IP Semester 4 (r = 0.38) and IP Semester 1 (r = -0.24), so the influence of academic achievement per semester on graduation looks varied. In addition, the variable number of courses repeated has a negative correlation with graduation status (r = -0.14), indicating that the more courses repeated by students, the lower the chance of graduation. Meanwhile, the variable number of leave (semester) shows almost no significant effect on graduation status (r = 0.01). On the other hand, the relationship between input variables, especially between credits taken between semesters, shows quite high multicollinearity (r > 0.90), which reflects a relatively consistent pattern of credit taking between semesters by students. Overall, the results of this correlation analysis provide an initial picture of the relevant and potential variables to be further developed in building a student graduation prediction model.

Modeling and Evaluation

XGBoost

The XGBoost model is implemented using XGBClassifier with more comprehensive hyperparameter optimization. The optimized parameters include:

Table 1. XGBoost Parameters

Parameter	Value
n_estimators	200
learning_rate	0.1
max_depth	5
min_child_weight	3
subsample	0.8

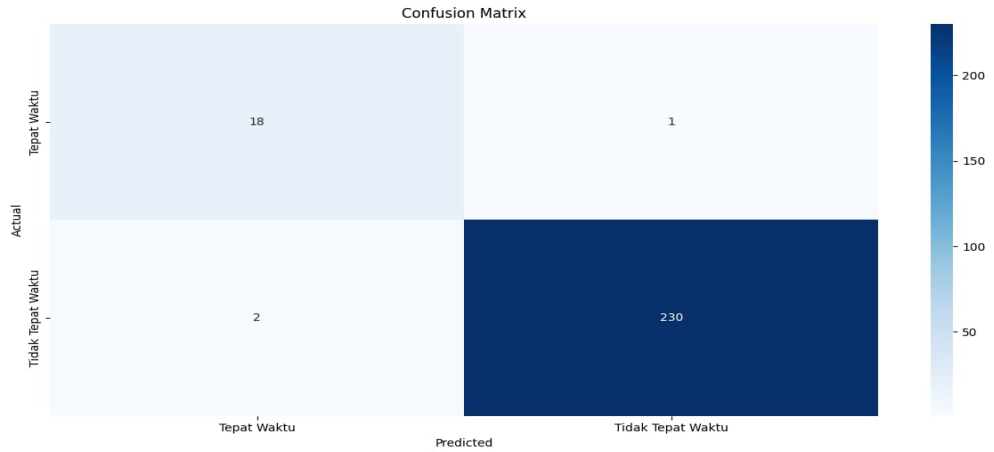

Figure 5. XGBoost Before Balancing

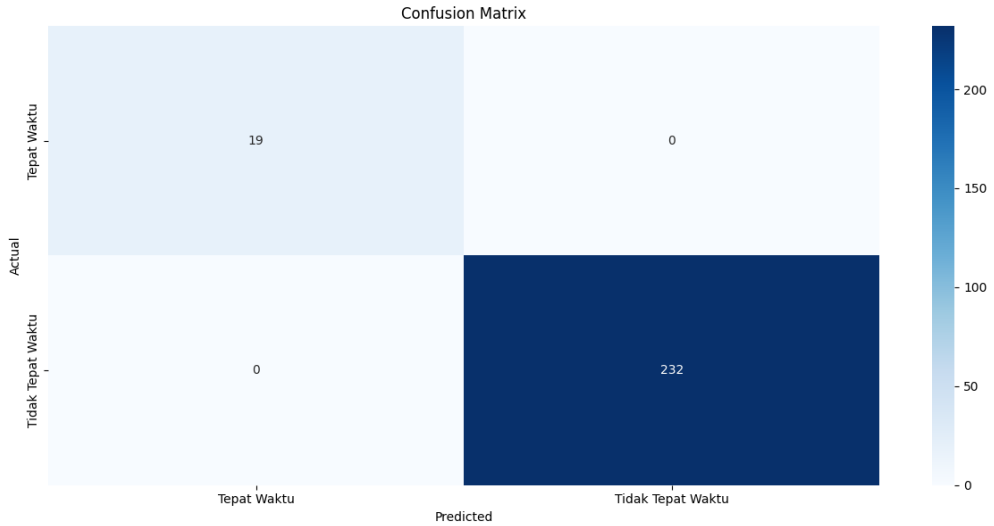

Figure 6. XGBoost After Balancing

Before balancing, the XGBoost model produced 18 True Positive (TP), 230 True Negative (TN), 1 False Positive (FP), and 2 False Negative (FN). After implementing the SMOTE-TOMEK technique, the model performance increased by producing perfect classification, namely 19 TP and 232 TN, and no classification errors (0 FP and 0 FN). These results indicate full accuracy in distinguishing between students who passed and failed.

Gradient Boosting

The Gradient Boosting model is implemented using the GradientBoostingClassifier from scikit-learn. This model uses a boosting approach by combining several weak learners into a strong learner. Based on the code implementation, the parameters used include:

Table 2. Gradient Boosting Parameters

Parameter	Value
n_estimators	100
learning_rate	0.1
max_depth	3
random_state	42

Figure 7. Gradient Boosting Before Balancing

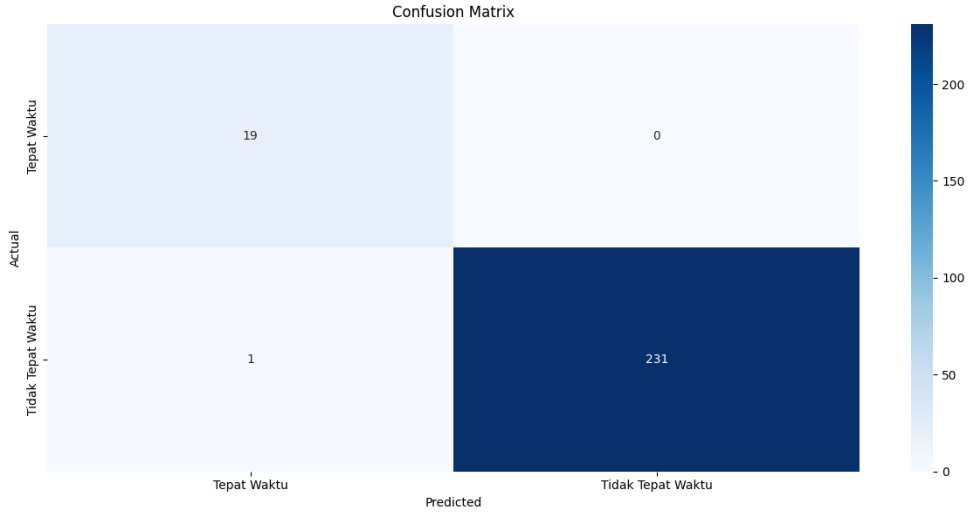

Figure 8. Gradient Boosting After Balancing

Gradient Boosting shows very good performance consistency, both before and after data balancing. The confusion matrix shows identical results, namely 19 TP, 231 TN, 0 FP, and only 1 FN. This indicates that the model has robustness to changes in data distribution, and is able to maintain prediction quality even though the data is unbalanced.

KNN

The KNN model is implemented using KNeighborsClassifier with optimized parameters:

Table 3. KNN Parameters

Parameter	Value
n_neighbors	5

weights	'uniform'
metric	'minkowski

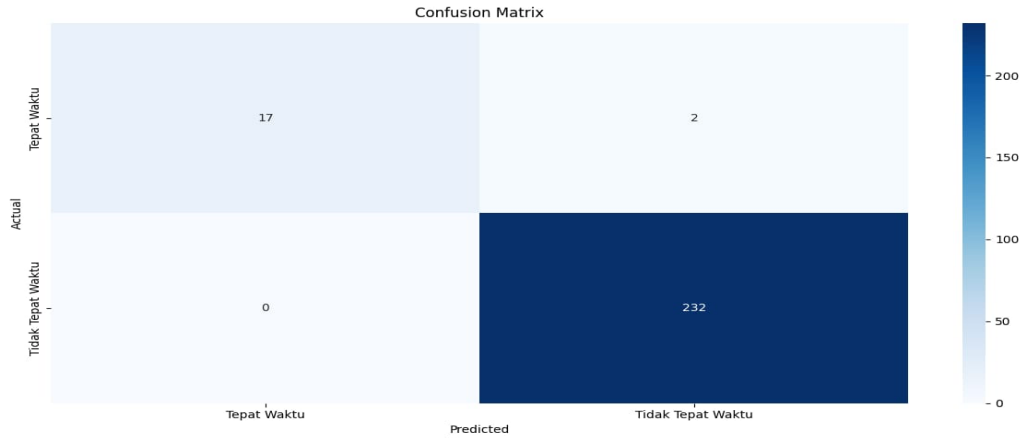

Figure 9. KNN Before Balancing

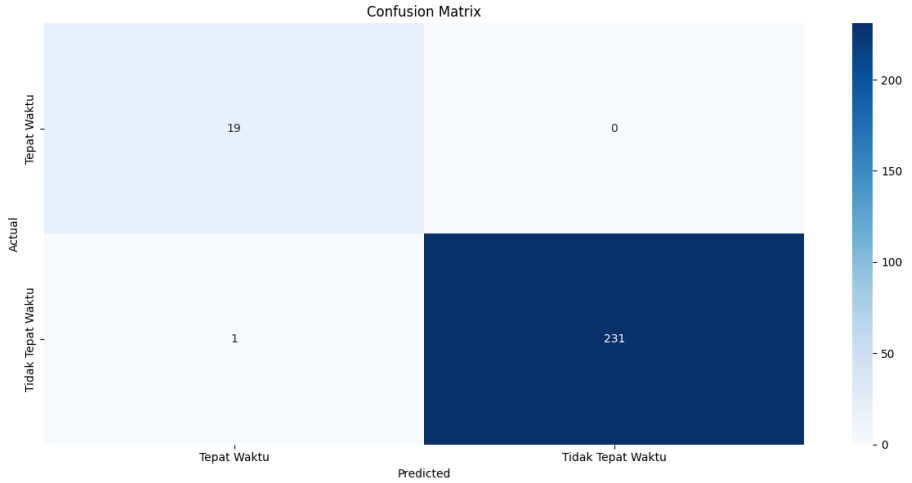

Figure 10. KNN After Balancing

Before balancing, KNN produced 17 TP, 232 TN, 2 FP, and 0 FN. After balancing, the model performance increased with a value of 19 TP and 231 TN, as well as 0 FP and 1 FN. This change shows that KNN responds positively to the balanced data distribution, marked by reduced classification errors and increased model ability to recognize graduating students more accurately.

Model Performance Comparison

The evaluation of model performance in this study was carried out using four main metrics, namely accuracy, precision, and recall, as well as F1-score. The following is a detailed explanation for each evaluation metric:

Accuracy is a metric that measures how accurate the model is in predicting overall. The equation for calculating accuracy is:

$$Accuracy = \frac{TP + TN + FP + FN}{TP + TN}$$

Where TP (True Positive) and TN (True Negative) represent correct predictions, however, FP (False Positive) and FN (False Negative) represent incorrect predictions.

Precision measures how accurate the positive predictions produced by the model are. In the context of this study, precision indicates how accurately the model predicts students who graduate. The equation for calculating precision is:

$$Precision = \frac{TP}{TP + FP}$$

Recall evaluates the ability of the model to find all positive cases. In the context of this study, recall indicates the ability of the model to identify students who actually graduated. The equation for calculating recall is:

$$Recall = \frac{TP}{TP + FN}$$

F1-score is the harmonic mean of precision and recall that provides a balance between the two metrics. The equation for calculating F1-score is:

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

Table 4 shows the performance comparison of the three models (Gradient Boosting, XGBoost, and KNN) in the conditions before and after balancing based on the four evaluation metrics.

Table 4. Performance comparison

Model	***Accuracy***	***Precision***	***Recall***	***F1-Score***
***XGBoost* Before Balancing**	**0.9976**	**0.9976**	**0.9976**	**0.9976**
XGBoost* After *Balancing	**1.0000**	**1.0000**	**1.0000**	**1.0000**
Gradient Boosting* Before *Balancing	**0.9992**	**0.9992**	**0.9992**	**0.9992**
Gradient Boosting* After *Balancing	**0.9992**	**0.9992**	**0.9992**	**0.9992**
KNN Before *Balancing*	0.9928	0.9928	0.9928	0.9927
KNN After *Balancing*	0.9968	0.9969	0.9968	0.9968

The implementation of data balancing techniques has been proven to have a significant impact on improving the performance of prediction models. In previous studies, for example, the Random Forest algorithm experienced an increase in accuracy from 83.6% to 85.6% after the application of the data imbalance handling method. This increase confirms that balanced data distribution has a significant impact on the effectiveness of the model in classifying, especially in the context of academic predictions such as student graduation. When the data is imbalanced, the model tends to be biased towards the majority class, namely students who graduate late or do not graduate, thus ignoring the minority class which is often the main focus in predictive research, namely students who successfully graduate on time.

As a result, the use of data balancing procedures, such as the Synthetic Minority Over-sampling Technique – Tomek Links (SMOTE-TOMEK), is an important step that increases the model's capacity to generalize to new data while increasing accuracy. Data balancing also helps reduce prediction errors that often affect minority groups, which are smaller in number but important in real-world situations. These results support recent studies showing that the use of data balancing significantly improves the performance of techniques such as Extreme Gradient Boosting (XGBoost) and K-Nearest Neighbor (KNN). Therefore, addressing data imbalance is one of the important elements that cannot be ignored when building machine learning-based prediction systems, especially for use in higher education-related contexts.

CONCLUSION

From the findings of this study, it is concluded that the SMOTE-TOMEK technique is effective in dealing with data imbalance problems and significantly improves the performance of the student graduation prediction model. Among the three algorithms tested, XGBoost showed the best performance after the balancing process, with accuracy, precision, recall, and F1-score scores reaching 1.0000. Gradient Boosting also gave very good results, with a score of 0.9992 which was relatively stable in both imbalanced and balanced data conditions. Meanwhile, K-Nearest Neighbors (KNN) experienced a significant increase in performance after balancing, with accuracy increasing from 0.9928 to 0.9968. These results are supported by the confusion matrix analysis, which shows that XGBoost is able to classify all data without error, Gradient Boosting has a very low error rate, and KNN shows improvement in reducing classification errors. With these findings, XGBoost combined with the SMOTE-TOMEK technique is recommended as the main model in predicting student graduation, while Gradient Boosting can be a reliable alternative if the balancing process cannot be applied. However, this study has several limitations. First, the data used only includes academic variables from the academic information system (SIAKAD), so it has not considered non-academic factors such as learning motivation, socio-economic conditions, or involvement in student organizations that can also affect graduation. Second, the effectiveness of the SMOTE-TOMEK technique is highly dependent on parameter selection, so that in more complex data conditions or with extreme levels of imbalance, model performance may be different. Third, the scope of the algorithms tested is still limited to three supervised learning methods, without comparing them with other approaches such as deep learning. Fourth, this study does not apply feature selection techniques, which can have an impact on model complexity and the possibility of overfitting. The absence of feature selection can also reduce model interpretability and make the model more difficult to implement on a larger scale. For future development, this study can be expanded by enriching predictor variables, including relevant non-academic data. In addition, it would be very useful to compare the performance with other algorithms such as Random Forest, Support Vector Machine (SVM), or deep learning-based models such as Artificial Neural Network (ANN), especially in the context of large datasets. Implementation of feature selection techniques such as Recursive Feature Elimination (RFE), Principal Component Analysis (PCA), or importance score-based methods can also be considered to improve the efficiency and interpretability of the model. Finally, the use of data from various higher education institutions can be an important step to test the generalization of the model and ensure its reliability in various implementation contexts.

REFERENCES

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.

D. Hermanto, D. I. Ricoida, D. Pibriana, and M. R. Pribadi, “Analysis of Student Graduation Prediction Using Machine Learning Techniques on an Imbalanced Dataset : An Approach to Address Class Imbalance,” vol. 11, no. 3, pp. 559–568, 2024, doi: 10.15294/sji.v11i3.5528.

R. Al-Ali, K. Alhumaid, M. Khalifa, S. A. Salloum, R. Shishakly, and M. A. Almaiah, “Analyzing Socio-Academic Factors and Predictive Modeling of Student Performance Using Machine Learning Techniques,” *Emerg. Sci. J.*, vol. 8, no. 4, pp. 1304–1319, 2024, doi: 10.28991/ESJ-2024-08-04-05.

R. Primartha, *algortima machine learning*. Banding: Informatika Bandung, 2021.

J. Wang, C. Jiang, and S. Member, “Thirty Years of Machine Learning : The Road to Pareto-Optimal Wireless Networks Line of Sight”.

E. Retnoningsih and R. Pramudita, “Mengenal Machine Learning Dengan Teknik Supervised dan Unsupervised Learning Menggunakan Python,” vol. 7, no. 2, pp. 156–165, 2020.

R. S. Nurhalizah and R. Ardianto, “Analisis Supervised dan Unsupervised Learning pada Machine Learning : Systematic Literature Review,” vol. 4, no. 1, pp. 61–72, 2024.

H. Eyke, “Towards Analogy-Based Explanations,” 2020.

A. Almalawi, B. Soh, A. Li, and H. Samra, “Predictive Models for Educational Purposes: A Systematic Review,” *Big Data Cogn. Comput.*, vol. 8, no. 12, pp. 1–42, 2024, doi: 10.3390/bdcc8120187.

A. Luque, A. Carrasco, A. Martín, and A. de las Heras, “The impact of class imbalance in classification performance metrics based on the binary confusion matrix,” *Pattern Recognit.*, vol. 91, pp. 216–231, 2019, doi: 10.1016/j.patcog.2019.02.023.

T. Agustina, A. Fitrianto, and Indahwati, “Comparison of SARIMA, Bagging Exponential Smoothing with STL Decomposition and Robust STL Decomposition for Forecasting Red Chili Production,” *Int. J. Sci. Res. Sci. Eng. Technol.*, vol. 11, no. 2, pp. 64–73, 2024, doi: 10.32628/ijsrset2411146.

V. Atlantic, E. Sulistianingsih, and H. Perdana, “Gradient Boosting Machine Pada Klasifikasi Kelulusan Mahasiswa,” *Bul. Ilm. Math. Stat. dan Ter.*, vol. 13, no. 2, pp. 165–174, 2024.

D. Kurniadi, F. Nuraeni, and S. M. Lestari, “Implementasi Algoritma Naïve Bayes Menggunakan Feature Forward Selection dan SMOTE Untuk Memprediksi Ketepatan Masa Studi Mahasiswa Sarjana,” *J. Sist. Cerdas*, vol. 5, no. 2, pp. 63–82, 2022, doi: 10.37396/jsc.v5i2.215.

D. L. Wibisono and Z. Abidin, “Prediction of Student Graduation Predicts using Hybrid 2D Convolutional Neural Network and Synthetic Minority Over-Sampling Technique,” *Recursive J. Informatics*, vol. 1, no. 1, pp. 27–34, 2023, doi: 10.15294/rji.v1i1.65646.

A. Bisri and R. Rachmatika, “Integrasi Gradient Boosted Trees dengan SMOTE dan Bagging untuk Deteksi Kelulusan Mahasiswa,” *J. Nas. Tek. Elektro dan Teknol. Inf.*, vol. 8, no. 4, p. 309, 2019, doi: 10.22146/jnteti.v8i4.529.

M. W. Dwinanda, N. Satyahadewi, and W. Andani, “Classification of Student Graduation Status Using Xgboost Algorithm,” *BAREKENG J. Ilmu Mat. dan Terap.*, vol. 17, no. 3, pp. 1785–1794, 2023, doi: 10.30598/barekengvol17iss3pp1785-1794.