BUKTI KORESPONDENSI
ARTIKEL JURNAL INTERNASIONAL BEREPUTASI

Judul Artikel : Improving Student Graduation Timeliness Prediction Using SMOTE and Ensemble Learning with Stacking and GridSearchCV Optimization

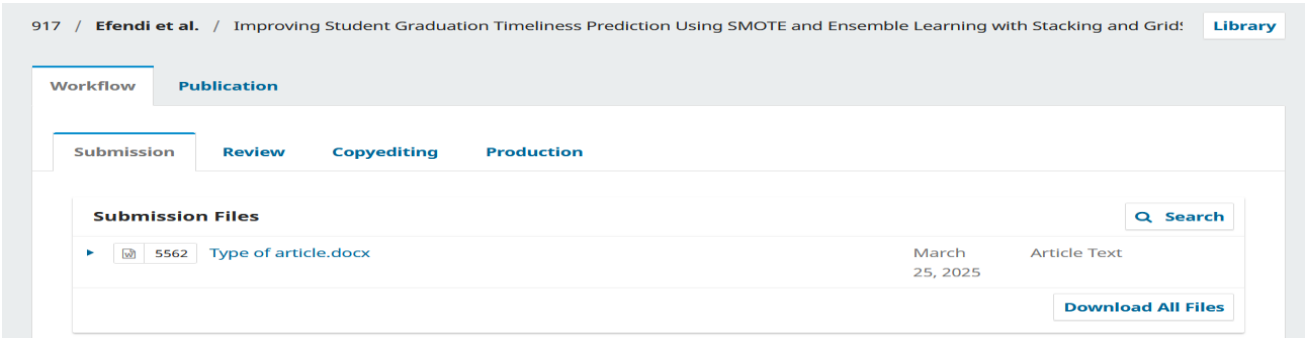
Jurnal : Data and Metadata, 2025, volume 4, no 917, halaman 1-10

Penulis : 1. Akmar Efendi
2. Iskandar Fitri
3. Gunadi Widi Nurcahyo

No.	Perihal	Tanggal
1.	Submit Artikel di Portal jurnal	25 Maret 2025
2.	Konfirmasi Editor Tehadap Tinjauan Artikel	26 Maret 2025
3.	Review Artikel	30 Maret 2025
4.	Upload Perbaikan Artikel	11 April 2025
5.	Accept Artikel	23 April 2025
6.	Publish Artikel di Jurnal Data and Metadata	25 April 2025

1. Submit Artikel di Portal Jurnal (25 Maret 2025)

- Screenshoot Submit Artikel di Portal Jurnal



- Screenshoot Artikel Submit

Type of article: Original

Improving Student Graduation Timeliness Prediction Using SMOTE and Ensemble Learning with Stacking and GridSearchCV Optimization

Mejora de la Predicción de la Oportunidad de Graduación Estudiantil Utilizando SMOTE y Aprendizaje por Ensamblado con Stacking y Optimización mediante GridSearchCV

Akmar Efendi [1], ORCID (https://orcid.org/0009-0008-4787-537X)
Iskandar Fitri [2], ORCID (https://orcid.org/0009-0005-7665-1074)
Gunadi Widi Nurcahyo [3], ORCID (https://orcid.org/0000-0003-0714-0244)

[1] Universitas Islam Riau, Department of Informatics Engineering. Pekanbaru, Indonesia.
[2,3] Universitas Putra Indonesia YPTK Padang, Department of Information Technology. Padang, Indonesia.

Corresponding author: Akmar Efendi, akmarefendi@eng.uir.ac.id

ABSTRACT
Introduction: Timely graduation is a key indicator of higher education success. Predicting student graduation time remains a challenge due to the complex interplay of academic and non-academic factors. This study aims to enhance graduation prediction accuracy using machine learning and ensemble techniques.
Methods: The study collected data via questionnaires from students and alumni of Universitas Islam Riau, covering academic performance, attendance, study habits, social support, stress levels, and extracurricular participation. After preprocessing and label encoding, SMOTE was applied to balance the class distribution between on-time and delayed graduates. Various machine learning models including Gaussian Naïve Bayes, SVM, Decision Tree, and KNN were tested. Ensemble methods, particularly voting and stacking, were implemented to boost performance. GridSearchCV was employed for hyperparameter tuning.
Results: The stacking ensemble model optimized with GridSearch achieved the highest accuracy of 99.37%, outperforming individual models such as SVM (96%) and Naïve Bayes (95%). The classification report showed high precision, recall, and F1-score (all above 0.99), and the confusion matrix indicated minimal misclassification.
Conclusions: Integrating SMOTE with ensemble learning and hyperparameter optimization significantly improved the model's ability to predict student graduation timeliness. The proposed method outperformed previous studies and provided a reliable framework for academic risk prediction in higher education.

Keywords: Student Graduation, Ensemble Learning, SMOTE, Stacking, GridSearchCV, Machine Learning.

RESUMEN
Introducción: La graduación oportuna es un indicador clave del éxito en la educación superior. Predecir el tiempo de graduación de los estudiantes sigue siendo un desafío debido a la compleja interacción de factores académicos y no académicos. Este estudio tiene como objetivo mejorar la precisión en la predicción de la graduación utilizando técnicas de aprendizaje automático y métodos de ensamblado.
Métodos: El estudio recopiló datos mediante cuestionarios dirigidos a estudiantes y egresados de la Universitas Islam Riau, abarcando el rendimiento académico, la asistencia, los hábitos de estudio, el apoyo social, los niveles de estrés y la participación extracurricular. Tras el preprocesamiento y la codificación de etiquetas, se aplicó SMOTE para equilibrar la distribución entre las clases de estudiantes que se gradúan a tiempo y los que se retrasan. Se evaluaron varios modelos de aprendizaje automático, incluidos Gaussian Naïve Bayes, SVM, Árbol de Decisión y KNN. Se implementaron métodos de ensamblado, particularmente voting y stacking, para mejorar el rendimiento. GridSearchCV se utilizó para la optimización de hiperparámetros.
Resultados: El modelo de ensamblado con stacking optimizado mediante GridSearch logró la mayor precisión con un 99.37%, superando a modelos individuales como SVM (96%) y Naïve Bayes (95%). El informe de clasificación mostró altos valores de precisión, recall y F1-score (todos superiores a 0.99), y la matriz de confusión indicó una clasificación errónea mínima.

Conclusiones: La integración de SMOTE con aprendizaje por ensamblado y la optimización de hiperparámetros mejoró significativamente la capacidad del modelo para predecir la puntualidad en la graduación de los estudiantes. El método propuesto superó a estudios anteriores y proporcionó un marco confiable para la predicción del riesgo académico en la educación superior.

Palabras clave: Graduación Estudiantil, Aprendizaje Por Ensamblado, SMOTE, Stacking, GridSearchCV, Aprendizaje Automático.

INTRODUCTION

Timely graduation rates are one of the key indicators of success for higher education institutions.[1] Universitas Islam Riau (UIR), as one of the leading private universities in Indonesia, also faces challenges in improving its on-time graduation rates. Various academic and non-academic factors may influence student graduation, such as teaching quality, availability of learning facilities, social support, and students' psychological conditions. However, identifying and understanding these factors is often complex and requires an in-depth analytical approach.[2]

With the advancement of technology, the application of machine learning in educational data analysis has opened up new opportunities for discovering patterns that influence academic success.[3] This approach can handle large datasets and detect hidden patterns that are difficult to capture using traditional methods.[4,5] Therefore, this study aims to explore the influence of academic and non-academic factors on student graduation at UIR using various machine learning algorithms.

Previous research has explored different approaches to predicting student graduation using machine learning and deep learning methods. One study using the Naïve Bayes algorithm achieved an accuracy of 85%, demonstrating that probabilistic methods can effectively handle data, although they have limitations in dealing with unstructured data of high complexity.[6] Meanwhile, the Deep Neural Network (DNN) approach achieved a higher accuracy of 87%, utilizing its architecture to better capture complex data patterns.[7] Another study implemented ensemble learning using AdaBoost, which enhanced the performance of the Decision Tree and reached an F1-score of 82%, confirming that boosting techniques can improve model generalization.[8]

A more complex approach was proposed by combining Support Vector Machine (SVM) with SMOTEN-ENN, Spearman's correlation, and randomized search, resulting in an accuracy of 73%, although it still fell short compared to other methods.[9] The K-Nearest Neighbors (KNN) algorithm was also evaluated in another study and obtained the highest accuracy of 89%, demonstrating the effectiveness of distance-based methods in handling graduation prediction tasks.[10]

This study employs a variety of machine learning algorithms to predict student graduation by leveraging the strengths of each method in handling different data characteristics. Multinomial Naïve Bayes (MNB) was chosen for its ability to manage text-based and probabilistic data, especially in processing questionnaire data that has undergone word-weighting stages.[11] MNB is effective in estimating the probability of categories based on word frequency, making it suitable for text-based analysis. Linear Support Vector Machine (SVM) was used due to its capability to find optimal hyperplanes for separating classes in high-dimensional data.[12] and its stable performance—especially when combined with data balancing techniques like SMOTE—helps improve model accuracy when dealing with imbalanced data.[13]

In addition, Decision Tree (DT) was applied for its ability to handle both categorical and numerical data, as well as for its interpretability. DT works by building rule-based decision trees that allow the identification of key factors influencing student graduation.[14] Meanwhile, K-Nearest Neighbors (KNN) was used as a comparative model because it operates based on similarity between data samples, making it effective for relatively small sample sizes. KNN also has advantages in capturing non-linear relationships in data, providing additional insights into the factors affecting graduation.[15]

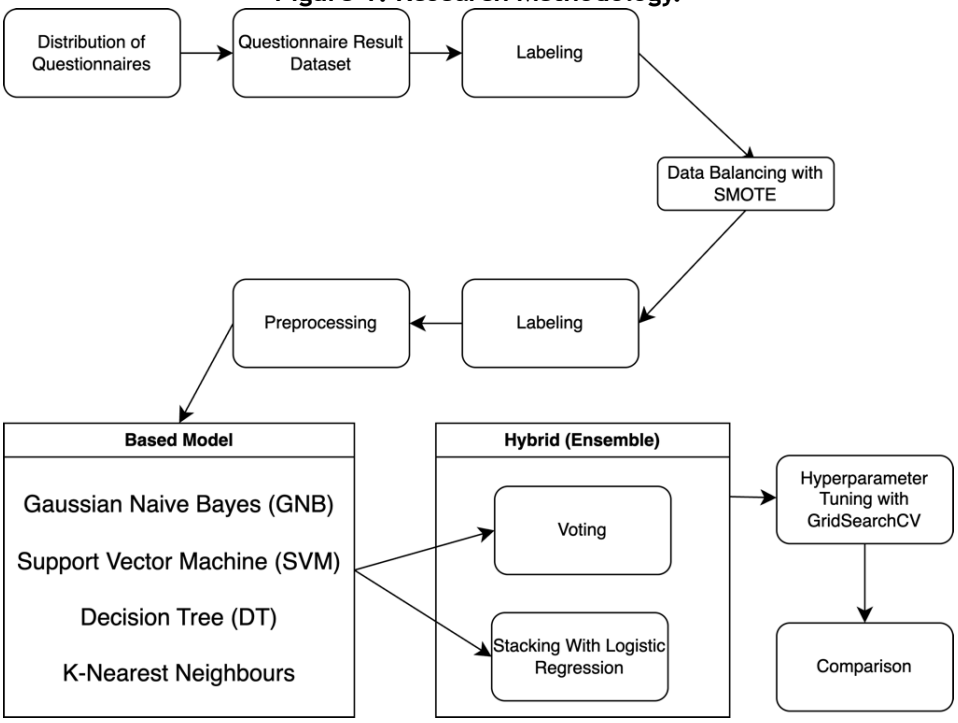
To improve prediction performance, this study applied Voting and Stacking ensemble methods with Logistic Regression. The Voting method acts as an ensemble technique that combines predictions from several models to enhance classification accuracy.[16] Stacking with Logistic Regression is used as a meta-classifier to combine predictions from base classifiers and produce a more accurate final decision.[17] Logistic Regression was selected as the meta-classifier due

to its good generalization ability and its effectiveness in both binary and multi-class classification tasks.[18]
o address data imbalance, this study adopted SMOTE (Synthetic Minority Over-sampling Technique), which generates synthetic samples for the minority class, resulting in a more balanced dataset. Data imbalance often causes models to be biased toward the majority class, so applying SMOTE allows the model to learn more effectively from the underrepresented category.[19] urthermore, this study used GridSearchCV for hyperparameter tuning, aimed at finding the best parameter combinations for each model. This tuning process is crucial to avoid underfitting or overfitting and to ensure that the model performs optimally.[20]
Compared to previous studies, this research integrates several best-practice approaches, including the use of SMOTE to address class imbalance, the implementation of ensemble methods such as voting and stacking to boost model performance, and the use of GridSearchCV for model parameter optimization. Therefore, this study presents a more comprehensive approach to predicting student graduation, with the goal of producing more accurate results than the individual methods used in earlier research.

METHODS

Figure 1 shows the methodology used in this research.

Figure 1. Research Methodology.



This study began with data collection, in which questionnaires were distributed to students and alumni of Universitas Islam Riau. The questionnaire was designed to gather both academic and non-academic information that may influence student graduation. The academic aspects included academic performance, Grade Point Average (GPA), total credits taken, and attendance rate, while the non-academic aspects covered psychological factors, social support, and learning motivation. Table 1 presents the questionnaire distributed to students of Universitas Islam Riau.

Table 1. A List of Questions.

Academic Factors	
1	What is your average GPA so far?
2	How often do you attend lectures?
3	Are you actively involved in academic activities (e.g., class discussions, group assignments)?

4	On average, how much time do you spend studying each day?
	Non Academic
1	How much social support do you receive from your family?
2	Do you currently have a part-time job?
3	What is your stress level during your college years?
4	Do you have adequate access to learning facilities (e.g., library, internet)?
5	Are you involved in any organizational or extracurricular activities on campus?

After the questionnaires were collected and compiled into a questionnaire-based dataset, a preprocessing stage was carried out to ensure the quality of the data before further analysis. This stage involved the use of label encoding. Label encoding is a data preprocessing technique used to convert categorical text data into numerical values. [21] This is done by assigning a specific number to each unique category within a feature. This technique is important because most machine learning algorithms cannot process textual data directly. By converting categories into numerical representations, models can interpret and process the information mathematically. Label encoding is especially useful when categorical data is ordinal or has a limited number of categories. The use of this technique improves processing efficiency and allows algorithms to learn from the data more effectively.[22]

In addition, one of the challenges in predictive analysis is the imbalance in the number of samples between classes. Therefore, SMOTE (Synthetic Minority Over-sampling Technique) was applied to balance the number of samples between the "On-Time Graduation" and "Not On-Time Graduation" classes. This technique aims to generate synthetic data to strengthen the minority class, reducing model bias toward the majority class and improving predictive accuracy. Once the data was balanced, the next step was the modeling process.[23]

In the modeling and classification phase, several machine learning algorithms were used to build the predictive model. The models tested included Gaussian Naïve Bayes (GNB), Linear Support Vector Machine (SVM), Decision Tree, and K-Nearest Neighbors (KNN). These four algorithms were selected due to their individual strengths in handling various types of data. Subsequently, ensemble learning techniques were applied to improve model performance. Two main ensemble learning approaches used were voting and stacking. In the voting technique, individual models are combined using majority voting to enhance prediction accuracy .[24] In contrast, stacking combines the results of several base models and further analyzes them using Logistic Regression as a meta-learner to produce a more accurate final prediction.[25] Table 2 presents the models used in this study.

Table 2. Model Testing.

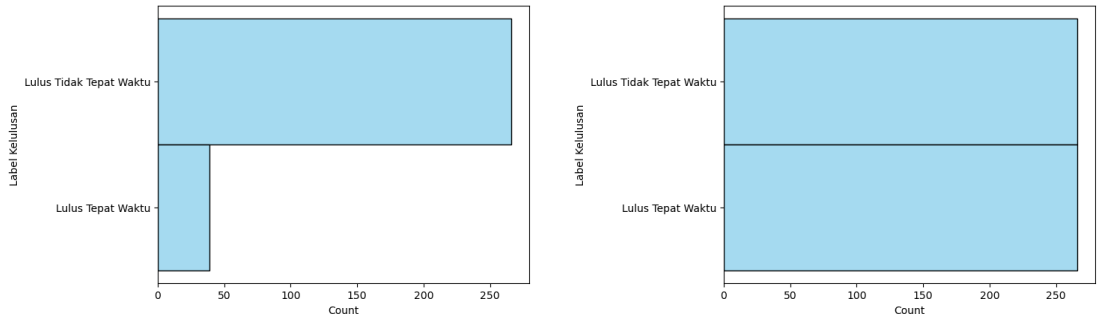
NO	Model
1	Gaussian Naive Bayes
2	Support Vector Machine
3	Decision Tree
4	K-Nearest Neighbours
5	Ensemble With Voting
6	Ensemble With Stacking
7	Ensemble With Voting + GridSearch
8	Ensemble With Stacking + GridSearch

After the model was built, hyperparameter optimization was performed using GridSearchCV. This technique is used to find the best parameter combinations for each algorithm applied, with the goal of improving model performance.[26] Once the optimization process was completed, the models were evaluated and compared to determine the best-performing algorithm for predicting student graduation based on the collected dataset. The final results of this process were used for model comparison, in which the model with the best performance was recommended as the optimal approach for analyzing the factors that influence whether students graduate on time or not.

RESULTS,

The first stage after the data is obtained is to carry out the data labeling process. Figure 2 is a distribution based on labels.

Figure 2. Label Distribution Before and After Oversampling.



2A. Before 2B. After

Based on Figure 2, there are two visualizations in the form of horizontal bar charts showing the number of students according to graduation labels: *Lulus Tepat Waktu* (On-Time Graduation) and *Lulus Tidak Tepat Waktu* (Not On-Time Graduation). In chart 2A, it is evident that the majority of students fall into the "Not On-Time Graduation" category, while only a small portion successfully graduate on time. This imbalance indicates that the dataset has an uneven class distribution, which can cause machine learning models to become biased toward the majority class.

To address this issue, an oversampling technique using SMOTE (Synthetic Minority Over-sampling Technique) was applied. SMOTE works by synthesizing new data points for the minority class based on existing characteristics, thereby creating realistic additional data. The results of this process are shown in chart 2B, where the number of data points in both categories becomes balanced. The application of SMOTE is crucial to ensure that the trained model can learn patterns from both classes proportionally, thereby improving the model's predictive ability and generalization to new data. After the labeling process, the next step is model development. Figure 3 presents the confusion matrix from the ensemble model using the stacking technique with GridSearchCV.

Figure 3. Confusion Matrix

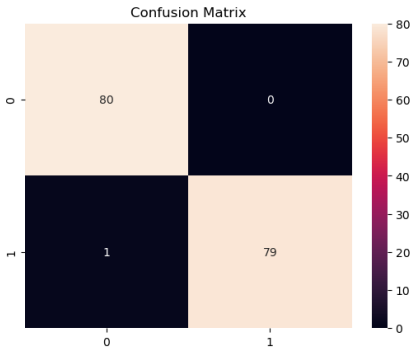


Figure 3 is a visualization of the confusion matrix used to evaluate the performance of the classification model for two classes: “On-Time Graduation” (label 1) and “Not On-Time Graduation” (label 0). The matrix consists of four main components: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN).

Based on the figure, the model successfully classified 80 data points correctly as "Not On-Time Graduation" (TN) and 79 data points correctly as "On-Time Graduation" (TP). There was only one misclassification, where a data point that should have been labeled as "On-Time Graduation" was incorrectly predicted as "Not On-Time Graduation" (FN). No data points were incorrectly predicted as "On-Time Graduation" when they actually belonged to the "Not On-Time Graduation" class (FP = 0).

Overall, these results indicate that the model has very high accuracy, with an extremely low error rate. This suggests that the model is highly effective in distinguishing between the two classes, especially after applying data balancing using the SMOTE technique. The next section, presented in Figure 4, is the classification report.

Figure 4. Classification Report

```
                          precision    recall  f1-score   support

      Lulus Tepat Waktu       0.99      1.00      0.99        80
Lulus Tidak Tepat Waktu       1.00      0.99      0.99        80

               accuracy                           0.99       160
              macro avg       0.99      0.99      0.99       160
           weighted avg       0.99      0.99      0.99       160

Accuracy : 0.99375
-----------------------------------------
```

Figure 4 displays the performance evaluation results of the classification model using evaluation metrics such as precision, recall, and F1-score for the two classes: Lulus Tepat Waktu (On-Time Graduation) and Lulus Tidak Tepat Waktu (Not On-Time Graduation). The model demonstrated very high performance, with precision and recall values ranging from 0.99 to 1.00 for both classes. This indicates that the model is highly accurate in identifying students who graduate on time as well as those who do not, with a very low error rate.
The F1-score, which also reached 0.99 for both classes, reflects the model's excellent balance between precision and recall. Overall, the model achieved an accuracy of 99.375%, meaning that out of 160 test data points, only one was misclassified. The macro average and weighted average values also show high consistency across all metrics, affirming that the model is not biased toward either class. These results indicate that the classification model developed in this study is highly reliable and effective in predicting students' on-time graduation, especially after the dataset was balanced using the SMOTE oversampling technique. The overall model testing results are presented in Table 3.

Table 3. Overall Test Results.

NO	Model	Accuracy	Precision	Recall	F1-Score
1	Gaussian Naive Bayes	95%	95%	95%	95%
2	Support Vector Machine	96%	95%	96%	95%
3	Decision Tree	93%	93%	93%	93%
4	K-Nearest Neighbours	94%	94%	94%	94%
5	Ensemble With Voting	98%	98%	98%	98%
6	Ensemble With Stacking	98%	98%	98%	98%
7	Ensemble With Voting + GridSearch	99%	99%	99%	99%
8	Ensemble With Stacking + GridSearch	99%	99%	99%	99%

Table 3 presents the performance evaluation results of various classification models applied to student graduation data, based on metrics such as accuracy, precision, recall, and F1-score. In general, all models demonstrated relatively high performance, with accuracy levels above 90%. The Gaussian Naive Bayes model achieved an accuracy of 95%, followed by Support Vector Machine (SVM) with 96%, while the Decision Tree performed slightly lower at 93%. The K-Nearest Neighbours (KNN) model recorded an accuracy of 94%.
Ensemble models showed superior performance. Both the Voting and Stacking ensemble methods achieved an accuracy of 98%, with consistently high values for precision, recall, and F1-score. Further improvement was obtained through the combination of ensemble learning and hyperparameter tuning using GridSearch, specifically in the Ensemble Voting + GridSearch and Ensemble Stacking + GridSearch models, both of which reached the highest accuracy of 99% across all evaluation metrics. These results indicate that the ensemble approach, particularly when optimized with GridSearch, provides the most reliable and accurate classification performance in predicting on-time student graduation compared to other models.

DISCUSSION

Figure 1 presents the research methodology, which consists of several key stages starting from questionnaire distribution, data labeling, to modeling and evaluation. The study began with data collection through the distribution of questionnaires designed to measure academic and non-academic factors that influence student graduation. The questionnaire data were then processed through a preprocessing stage and labeled based on graduation criteria. One of the challenges encountered was class imbalance in the data, where the number of students who

did not graduate on time was significantly higher. To address this issue, an oversampling technique using SMOTE (Synthetic Minority Over-sampling Technique) was applied, aiming to balance the data and avoid bias in the prediction model.
The next step involved building classification models using algorithms such as Gaussian Naïve Bayes, Support Vector Machine, Decision Tree, and K-Nearest Neighbours. In addition to these base models, this study also applied ensemble learning approaches, namely voting and stacking, to enhance model accuracy and generalization. Furthermore, hyperparameter optimization was carried out using GridSearchCV to achieve the best performance from each algorithm. Through these stages, the evaluation results showed that the ensemble model combined with GridSearch delivered the best performance, achieving an accuracy of 99%. This indicates that the methodological approach used in this study successfully produced a highly reliable predictive model for classifying students based on their on-time graduation status. These results also outperform previous studies, as shown in Table 4.

Table 4. Comparison with Previous Research.

NO	Researcher	Model	Accuracy
1	Rahmaddeni, et al. 2022[27]	XGBoost + SVM	79%
2	Ulfah, et al. 2022[28]	SVM + GridSearchCV	95%
3	Rahmiati, et al. 2023[29]	Naive Bayes	95%
4	Herianto, et al. 2024[25]	SMLOS (Stacking Machine Learning Optuna SMOTE)	95%
5	Van FC, et al. 2025[17]	Stacking + Adaboost + Hyperparameter Tuning + SMOTE	97%
6	Suandi, et al. 2024[24]	Majority Voting + SMOTE	97%
7	Putra, et al. 2025[30]	Stacking + SMOTE	88%
8	Anam, et al. 2025[16]	Voting Hard + SMOTE	94%
9	This Study	SMOTE + Ensemble With Stacking + GridSearch	99%

Table 4 presents a comparison of accuracy results from various previous studies that employed different approaches in data classification, particularly in the context of predictive analysis using machine learning. The table shows that most studies have applied a combination of classification methods and performance enhancement techniques such as SMOTE, ensemble learning, and hyperparameter tuning to address data imbalance and improve model accuracy.
The study by Rahmaddeni et al. (2022), which used a combination of XGBoost and SVM, achieved an accuracy of only 79%,[27] whereas Ulfah et al. (2022) and Rahmiati et al. (2023), who applied SVM with GridSearchCV and Naïve Bayes respectively, each attained an accuracy of 95%.[28] [29] More recent studies, such as Herianto et al. (2024) with the SMLOS method, as well as Van FC et al. (2025) and Suandi et al. (2024) using stacking and voting approaches combined with SMOTE and hyperparameter tuning, successfully achieved higher accuracy, reaching 97%.[25] [17] [24]

In comparison, the present study demonstrates the most superior performance with an accuracy of 99%, achieved through the integration of SMOTE, ensemble learning using stacking techniques, and hyperparameter optimization via GridSearch. These results confirm that the integrative approach adopted in this study provides more optimal outcomes than those reported in previous research. This success highlights the effectiveness of ensemble techniques and data balancing in enhancing classification performance, as well as the significant potential for their application in real-world data analysis and natural language processing tasks.

CONCLUSIONS

This study successfully developed a predictive model for on-time student graduation by integrating various machine learning algorithms and relevant supporting techniques. By employing base algorithms such as Gaussian Naïve Bayes, Support Vector Machine (SVM), Decision Tree, and K-Nearest Neighbours, along with ensemble learning approaches like voting and stacking, the research demonstrated that combining methods can yield more optimal prediction outcomes. One of the key strengths of this study lies in the application of SMOTE to handle class imbalance in the dataset and the use of GridSearchCV for hyperparameter optimization. The results showed that the ensemble model using the stacking approach,

combined with SMOTE and hyperparameter tuning, achieved the highest performance with an accuracy of 99%, outperforming models from previous studies.
Model evaluation using the confusion matrix and classification report indicated that the developed model not only excelled in accuracy but also provided balanced and reliable predictions with minimal misclassification. This confirms that the integrative approach employed is highly effective in identifying and classifying the factors influencing student graduation. These findings are expected to support decision-making in higher education institutions, particularly in efforts to improve academic performance and student development.
Moving forward, this research can be expanded in two main directions. First, by broadening the dataset to include data from various universities across Indonesia, which would enhance the generalizability of the predictive model. Second, by integrating the model into academic information systems in real-time, enabling it to serve as a decision-support tool for early detection and intervention for students at risk of delayed graduation. With these two focuses, the predictive model becomes not only analytical but also practical, offering direct contributions to the improvement of higher education management.

REFERENCES

1. Bakri R, Astuti NP, Ahmar AS. Machine Learning Algorithms with Parameter Tuning to Predict Students' Graduation-on-time: A Case Study in Higher Education. J Appl Sci Eng Technol Educ. 2022 Dec 30;4(2):259–65.

2. Casanova VS, Pullido ML. Factors Of Graduate Students' Attrition And Retention In Occidental Mindoro State College Graduate School. IJERSC. 3(2):826–31.

3. López-Meneses E, López-Catalán L, Pelícano-Piris N, Mellado-Moreno PC. Artificial Intelligence in Educational Data Mining and Human-in-the-Loop Machine Learning and Machine Teaching: Analysis of Scientific Knowledge. Applied Sciences. 2025 Jan 14;15(2):1–21.

4. Ersozlu Z, Taheri S, Koch I. A review of machine learning methods used for educational data. Educ Inf Technol. 2024 Nov;29(16):22125–45.

5. Taye MM. Understanding of Machine Learning with Deep Learning: Architectures, Workflow, Applications and Future Directions. Computers. 2023 Apr 25;12(5):1–26.

6. Mehta S. Playing Smart with Numbers: Predicting Student Graduation Using the Magic of Naive Bayes. ITALIC. 2023 Nov 23;2(1):60–75.

7. Darenoh NV, Bachtiar FA, Perdana RS. Prediction of On-time Student Graduation with Deep Learning Method: -. J ICT Res Appl. 2024 Jun 27;18(1):1–20.

8. Desfiandi A, Soewito B. Student Graduation Time Prediction Using Logistic Regression, Decision Tree, Support Vector, and Adaboost Ensemble Learning. International Journal of Information System and Computer Science. 7(3):195–9.

9. Haikal MF, Palupi I. Predicting Employability of University Graduates Using Support Vector Machine Classification. Building of Informatics, Technology and Science. 2024;6(2).

10. Rismayati R, Ismarmiaty I, Hidayat S. Esemble Implementation for Predicting Student Graduation withClassification Algorithm. IJECSA. 1(1):35–42.

11. Anam MK, Putra PP, Malik RA, Putra TA, Elva Y, Mahessya RA, et al. Enhancing the Performance of Machine Learning Algorithm for Intent Sentiment Analysis on Village Fund Topic. Journal of Applied Data Sciences. 2025;6(2):1102–15.

12. Sharma H, Pangaonkar S, Gunjan R, Rokade P. Sentimental Analysis of Movie Reviews Using Machine Learning. Shah H, Patel R, Patel N, Buyya R, Chatterjee I, editors. ITM Web Conf. 2023;53:02006.

13. Anam MK, Firdaus MB, Suandi F, Lathifah, Nasution T, Fadly S. Performance Improvement of Machine Learning Algorithm Using Ensemble Method on Text Mining. In: 2024

International Conference on Future Technologies for Smart Society (ICFTSS) [Internet]. Kuala Lumpur, Malaysia: IEEE; 2024 [cited 2025 Mar 16]. p. 90–5. Available from: https://ieeexplore.ieee.org/document/10691363/

14. Putra PP, Anam MK, Defit S, Yunianta A. Enhancing the Decision Tree Algorithm to Improve Performance Across Various Datasets. intensif. 2024 Aug 1;8(2):200–12.

15. Danny M, Muhidin A, Jamal A. Application of the K-Nearest Neighbor Machine Learning Algorithm to Preduct Sales of Best-Selling Products. Brilliance. 2024 Jun 28;4(1):255–64.

16. Anam MK, Lestari TP, Yenni H, Nasution T, Firdaus MB. Enhancement of Machine Learning Algorithm in Fine-grained Sentiment Analysis Using the Ensemble. ECTI-CIT Transactions. 2025 Mar 8;19(2):159–67.

17. Van Fc LL, Anam MK, Bukhori S, Mahamad AK, Saon S, Nyoto RLV. The Development of Stacking Techniques in Machine Learning for Breast Cancer Detection. J Appl Data Sci. 2024 Jan 1;6(1):71–85.

18. Munthe IR, Rambe BH, Hanum F, Amanda AT, Hutagaol ASR, Harianto R. Implementation of Stacking Technique Combining Machine Learning and Deep Learning Algorithms Using SMOTE to Improve Stock Market Prediction Accuracy. J Appl Data Sci. 2024 Dec 1;5(4):2079–91.

19. Anam MK, Van Fc LL, Hamdani H, Rahmaddeni R, Junadhi J, Firdaus MB, et al. Sara Detection on Social Media Using Deep Learning Algorithm Development. JAETS. 2024 Dec 15;6(1):225–37.

20. Alemerien K, Alsarayreh S, Altarawneh E. Diagnosing Cardiovascular Diseases using Optimized Machine Learning Algorithms with GridSearchCV. J Appl Data Sci. 2024 Dec 1;5(4):1539–52.

21. Bolikulov F, Nasimov R, Rashidov A, Akhmedov F, Cho YI. Effective Methods of Categorical Data Encoding for Artificial Intelligence Algorithms. Mathematics. 2024 Aug 18;12(16):1–21.

22. Hien DTT, Thi C, Kim T, The D, Nguyen C. Optimize the Combination of Categorical Variable Encoding and Deep Learning Technique for the Problem of Prediction of Vietnamese Student Academic Performance. IJACSA. 2020;11(11):274–80.

23. Anam MK, Munawir M, Efrizoni L, Fadillah N, Agustin W, Syahputra I, et al. Improved Performance of Hybrid GRU-BiLSTM for Detection Emotion on Twitter Dataset. J Appl Data Sci. 2024 Jan 1;6(1):354–65.

24. Suandi F, Anam MK, Firdaus MB, Fadli S, Lathifah L, Yumami E, et al. Enhancing Sentiment Analysis Performance Using SMOTE and Majority Voting in Machine Learning Algorithms. In: Lumombo L, Rahmi A, Suwarno S, Ardi N, Kurniawan DE, editors. Proceedings of the 7th International Conference on Applied Engineering (ICAE 2024) [Internet]. Dordrecht: Atlantis Press International BV; 2024 [cited 2025 Mar 25]. p. 126–38. (Advances in Engineering Research; vol. 251). Available from: https://www.atlantis-press.com/doi/10.2991/978-94-6463-620-8_10

25. Herianto H, Kurniawan B, Hartomi ZH, Irawan Y, Anam MK. Machine Learning Algorithm Optimization using Stacking Technique for Graduation Prediction. J Appl Data Sci. 2024 Sep 1;5(3):1272–85.

26. Jumanto J, Rofik R, Sugiharti E, Alamsyah A, Arifudin R, Prasetiyo B, et al. Optimizing Support Vector Machine Performance for Parkinson's Disease Diagnosis Using GridSearchCV and PCA-Based Feature Extraction. J Inf Syst Eng Bus Intell. 2024 Feb 28;10(1):38–50.

27. Rahmaddeni R, Anam MK, Irawan Y, Susanti S, Jamaris M. Comparison of Support Vector Machine and XGBSVM in Analyzing Public Opinion on Covid-19 Vaccination. Ilk J Ilm. 2022 Apr 30;14(1):32–8.

28. Ulfah AN, Anam MK, Sidratul Munti NY, Yaakub S, Firdaus MB. Sentiment Analysis of the Convict Assimilation Program on Handling Covid-19. JUITA. 2022 Nov 14;10(2):209–16.

29. Rahmiati R, Anam MK, Paradila D, Mardainis M, Machdalena M. Application of Naïve Bayes Algorithm for Non-Cash Food Assistance Recipients in Kampar Regency. SinkrOn. 2023 Jan 4;8(1):433–41.

30. Putra PP, Anam MK, Chan AS, Hadi A, Hendri N, Masnur A. Optimizing Sentiment Analysis on Imbalanced Hotel Review Data Using SMOTE and Ensemble Machine Learning Techniques. J Appl Data Sci. 2025 May 1;6(2):921–35.

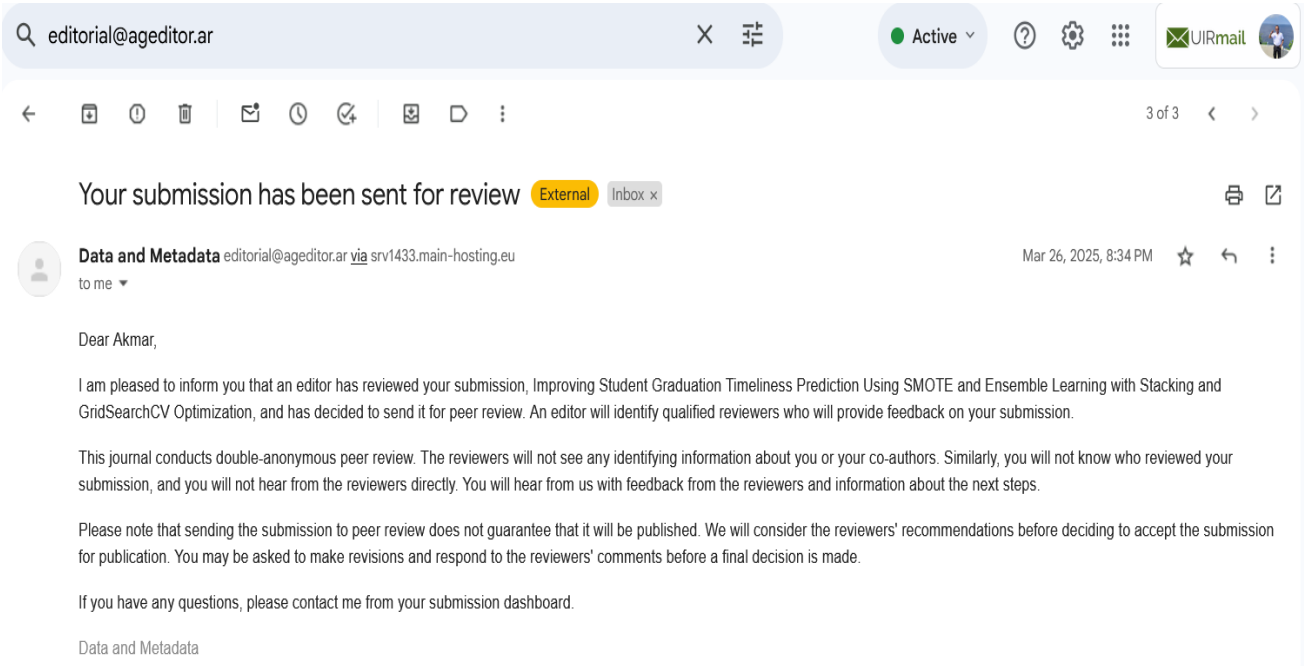
FINANCING
No financing

CONFLICT OF INTEREST
The authors declare that there is no conflict of interest".

AUTHORSHIP CONTRIBUTION:
Conceptualization: Akmar Efendi
Data curation: Akmar Efendi
Formal analysis: Akmar Efendi
Research: Akmar Efendi, Gunadi Widi Nurcahyo, and Iskandar Fitri
Methodology: Akmar Efendi, Gunadi Widi Nurcahyo, and Iskandar Fitri
Project management: Akmar Efendi
Resources: Akmar Efendi
Software: Akmar Efendi
Supervision: Iskandar Fitri
Validation: Iskandar Fitri
Display: Iskandar Fitri
Drafting - original draft: Akmar Efendi
Writing - proofreading and editing: Gunadi Widi Nurcahyo

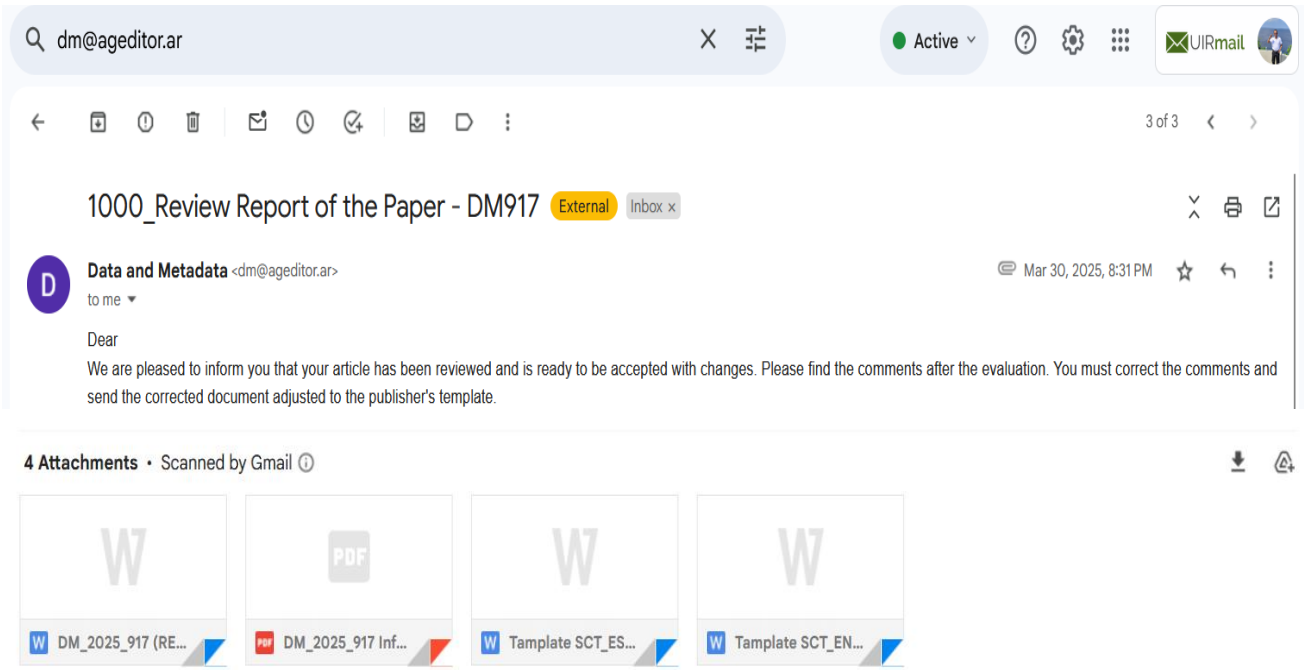
2. Konfirmasi Editor Tehadap Tinjauan Artikel (26 Maret 2025)

- **Bukti Screenshoot Tinjauan Artikel**



3. Review Artikel (30 Maret 2025)

- **Screenshoot E-mail pemberitahuan Revisi**



- **Screenshoot Komentar Revisi Artikel**

Type of article: Original

Improving Student Graduation Timeliness Prediction Using SMOTE and Ensemble Learning with Stacking and GridSearchCV Optimization

Mejora de la Predicción de la Oportunidad de Graduación Estudiantil Utilizando SMOTE y Aprendizaje por Ensamblado con Stacking y Optimización mediante GridSearchCV

Akmar Efendi [1], ORCID (https://orcid.org/0009-0008-4787-537X)
Iskandar Fitri [2], ORCID (https://orcid.org/0009-0005-7665-1074)
Gunadi Widi Nurcahyo [3], ORCID (https://orcid.org/0000-0003-0714-0244)

[1] Universitas Islam Riau, Department of Informatics Engineering. Pekanbaru, Indonesia.
[2,3] Universitas Putra Indonesia YPTK Padang, Department of Information Technology. Padang, Indonesia.

Corresponding author: Akmar Efendi, akmarefendi@eng.uir.ac.id

REVISTA
Dear authors:

Thank you for choosing this journal to publish your results; below are a few notes

REVISTA
In general: the results of the work are not clearly defined. These should be presented independently

RESUMEN
Introducción: La graduación oportuna es un indicador clave del rendimiento en la educación superior. Este estudio tiene como objetivo mejorar la precisión en la predicción de la oportunidad de graduación estudiantil mediante técnicas de aprendizaje automático por ensamblado combinadas con SMOTE y optimización de hiperparámetros.
Métodos: Este es un estudio cuantitativo predictivo. La población incluye a estudiantes y egresados de la Universitas Islam Riau. Se obtuvo una muestra de 160 participantes mediante muestreo intencional. Los datos se recopilaron utilizando cuestionarios estructurados que abarcan variables académicas (por ejemplo, promedio académico, créditos, asistencia) y no académicas (por ejemplo, estrés, apoyo social, participación extracurricular). Después del preprocesamiento y la codificación de etiquetas, se aplicó SMOTE para equilibrar la distribución de clases. Se probaron varios clasificadores (Naïve Bayes, SVM, Árbol de Decisión, KNN), y se implementó aprendizaje por ensamblado (voting y stacking) optimizado mediante GridSearchCV.
Resultados: El modelo de ensamblado tipo stacking optimizado con GridSearchCV alcanzó el mejor desempeño, con una precisión del 99.37%, valores de precisión y recall superiores a 0.99, y una tasa mínima de errores de clasificación. Este modelo superó a los modelos individuales y a enfoques previos en la literatura.
Conclusiones: La integración de SMOTE, métodos de ensamblado y GridSearchCV mejora significativamente la precisión predictiva para la graduación oportuna de los estudiantes. El

REVISTA
Adjust according to the modifications to the previous summary

DISCUSSION
The findings of this study reveal that the integration of SMOTE, ensemble learning, and GridSearchCV optimization yields a highly accurate predictive model for student graduation timeliness. The stacking ensemble model optimized with GridSearchCV achieved an accuracy of 99.37%, outperforming all baseline and ensemble-only models. The performance across all evaluation metrics—precision, recall, and F1-score—exceeded 0.99, indicating near-perfect classification.

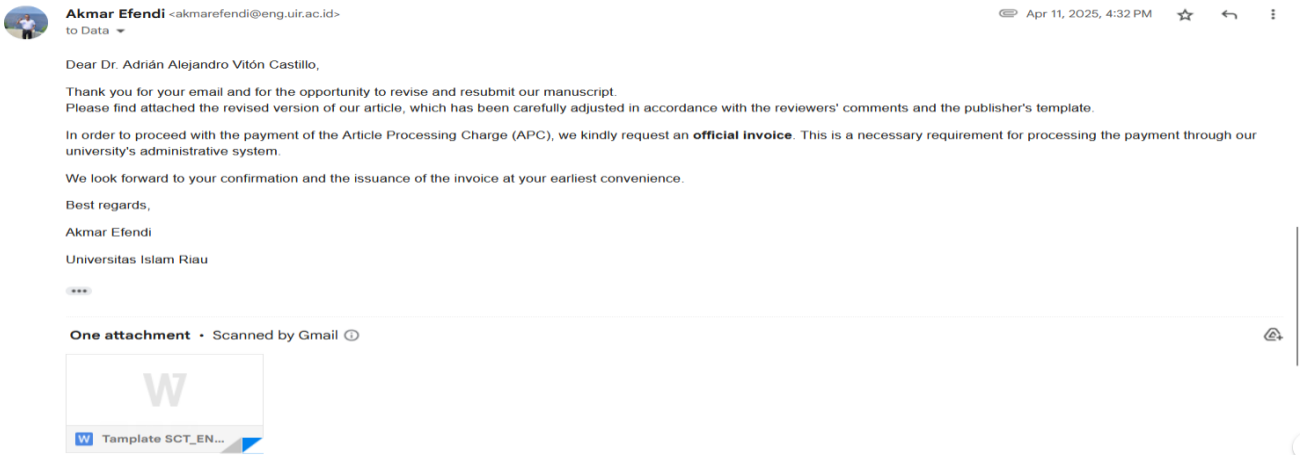
REVISTA
The discussion should be based on the results found. The authors should assess them, providing their own opinions, while also comparing them with previous similar studies
March 28, 2025, 10:45 AM

CONCLUSIONS
This study aimed to develop a predictive model for student graduation timeliness by integrating SMOTE, ensemble learning techniques (voting and stacking), and hyperparameter tuning using GridSearchCV. The findings confirm that this integrative approach is highly effective, as evidenced by the stacking ensemble model achieving a predictive accuracy of 99.37%, with precision, recall, and F1-scores above 0.99.
By addressing class imbalance through SMOTE and refining model performance using GridSearchCV, the study succeeded in enhancing classification quality beyond the capabilities of individual machine learning models. Compared to previous related studies, the proposed model demonstrates superior performance and robustness, proving its potential as a reliable framework for academic risk detection.

REVISTA
The conclusions are generalizing of the content and must respond to the objective.

4. Upload Perbaikan Artikel (11 April 2025)

- Screenshoot Upload Perbaikan Artikel



Akmar Efendi <akmarefendi@eng.uir.ac.id>
to Data

Apr 11, 2025, 4:32 PM

Dear Dr. Adrián Alejandro Vitón Castillo,

Thank you for your email and for the opportunity to revise and resubmit our manuscript.
Please find attached the revised version of our article, which has been carefully adjusted in accordance with the reviewers' comments and the publisher's template.

In order to proceed with the payment of the Article Processing Charge (APC), we kindly request an **official invoice**. This is a necessary requirement for processing the payment through our university's administrative system.

We look forward to your confirmation and the issuance of the invoice at your earliest convenience.

Best regards,

Akmar Efendi

Universitas Islam Riau

One attachment • Scanned by Gmail

Tamplate SCT_EN...

- Screenshoot Artikel Yang Diupload Hasil Perbaikan

Improving Student Graduation Timeliness Prediction Using SMOTE and Ensemble Learning with Stacking and GridSearchCV Optimization

Mejora de la Predicción de la Oportunidad de Graduación Estudiantil Utilizando SMOTE y Aprendizaje por Ensamblado con Stacking y Optimización mediante GridSearchCV

Akmar Efendi [1], ORCID (https://orcid.org/0009-0008-4787-537X)
Iskandar Fitri [2], ORCID (https://orcid.org/0009-0005-7665-1074)
Gunadi Widi Nurcahyo [3], ORCID (https://orcid.org/0000-0003-0714-0244)

[1] Universitas Islam Riau, Department of Informatics Engineering. Pekanbaru, Indonesia.
[2,3] Universitas Putra Indonesia YPTK Padang, Department of Information Technology. Padang, Indonesia.

Corresponding author: Akmar Efendi, akmarefendi@eng.uir.ac.id

ABSTRACT
Introduction: Timely graduation is a key performance indicator in higher education. This study aims to improve the accuracy of predicting student graduation timeliness using ensemble machine learning techniques combined with SMOTE and hyperparameter optimization.
Methods: This is a quantitative predictive study. The population includes students and alumni of Universitas Islam Riau. A sample of 160 respondents was obtained via purposive sampling. Data were collected using structured questionnaires encompassing academic variables (e.g., GPA, credits, attendance) and non-academic variables (e.g., stress, social support, extracurricular activity). After preprocessing and label encoding, SMOTE was applied to balance class distribution. Several classifiers (Naïve Bayes, SVM, Decision Tree, KNN) were tested, with ensemble learning (voting and stacking) implemented and optimized using GridSearchCV.
Results: The stacking ensemble model optimized with GridSearchCV achieved the highest performance with an accuracy of 99.37%, precision and recall above 0.99, and minimal misclassification. This outperformed individual models and previous approaches in the literature.
Conclusions: The integration of SMOTE, ensemble methods, and GridSearchCV significantly enhances predictive accuracy for student graduation timeliness. The resulting model provides a robust framework for academic risk detection and early intervention.

Keywords: Student Graduation, Ensemble Learning, SMOTE, Stacking, GridSearchCV, Machine Learning.

RESUMEN
Introducción: La graduación oportuna es un indicador clave del rendimiento en la educación superior. Este estudio tiene como objetivo mejorar la precisión en la predicción de la oportunidad de graduación estudiantil mediante técnicas de aprendizaje automático por ensamblado combinadas con SMOTE y optimización de hiperparámetros.
Métodos: Este es un estudio cuantitativo predictivo. La población incluye a estudiantes y egresados de la Universitas Islam Riau. Se obtuvo una muestra de 160 participantes mediante muestreo intencional. Los datos se recopilaron utilizando cuestionarios estructurados que abarcan variables académicas (por ejemplo, promedio académico, créditos, asistencia) y no académicas (por ejemplo, estrés, apoyo social, participación extracurricular). Después del preprocesamiento y la codificación de etiquetas, se aplicó SMOTE para equilibrar la distribución de clases. Se probaron varios clasificadores (Naïve Bayes, SVM, Árbol de Decisión, KNN), y se implementó aprendizaje por ensamblado (voting y stacking) optimizado mediante GridSearchCV.
Resultados: El modelo de ensamblado tipo stacking optimizado con GridSearchCV alcanzó el mejor desempeño, con una precisión del 99.37%, valores de precisión y recall superiores a 0.99, y una tasa mínima de errores de clasificación. Este modelo superó a los modelos individuales y a enfoques previos en la literatura.
Conclusiones: La integración de SMOTE, métodos de ensamblado y GridSearchCV mejora significativamente la precisión predictiva para la graduación oportuna de los estudiantes. El

modelo resultante proporciona un marco sólido para la detección de riesgos académicos y la intervención temprana.

Palabras clave: plantear entre 3 a 6 palabras claves.

INTRODUCTION

Timely graduation rates are one of the key indicators of success for higher education institutions.[1] Universitas Islam Riau (UIR), as one of the leading private universities in Indonesia, also faces challenges in improving its on-time graduation rates. Various academic and non-academic factors may influence student graduation, such as teaching quality, availability of learning facilities, social support, and students' psychological conditions. However, identifying and understanding these factors is often complex and requires an in-depth analytical approach.[2]

With the advancement of technology, the application of machine learning in educational data analysis has opened up new opportunities for discovering patterns that influence academic success.[3] This approach can handle large datasets and detect hidden patterns that are difficult to capture using traditional methods.[4,5] Therefore, this study aims to explore the influence of academic and non-academic factors on student graduation at UIR using various machine learning algorithms.

Previous research has explored different approaches to predicting student graduation using machine learning and deep learning methods. One study using the Naïve Bayes algorithm achieved an accuracy of 85%, demonstrating that probabilistic methods can effectively handle data, although they have limitations in dealing with unstructured data of high complexity.[6] Meanwhile, the Deep Neural Network (DNN) approach achieved a higher accuracy of 87%, utilizing its architecture to better capture complex data patterns.[7] Another study implemented ensemble learning using AdaBoost, which enhanced the performance of the Decision Tree and reached an F1-score of 82%, confirming that boosting techniques can improve model generalization.[8]

A more complex approach was proposed by combining Support Vector Machine (SVM) with SMOTEN-ENN, Spearman's correlation, and randomized search, resulting in an accuracy of 73%, although it still fell short compared to other methods.[9] The K-Nearest Neighbors (KNN) algorithm was also evaluated in another study and obtained the highest accuracy of 89%, demonstrating the effectiveness of distance-based methods in handling graduation prediction tasks.[10]

This study employs a variety of machine learning algorithms to predict student graduation by leveraging the strengths of each method in handling different data characteristics. Multinomial Naïve Bayes (MNB) was chosen for its ability to manage text-based and probabilistic data, especially in processing questionnaire data that has undergone word-weighting stages.[11] MNB is effective in estimating the probability of categories based on word frequency, making it suitable for text-based analysis. Linear Support Vector Machine (SVM) was used due to its capability to find optimal hyperplanes for separating classes in high-dimensional data.[12] and its stable performance—especially when combined with data balancing techniques like SMOTE—helps improve model accuracy when dealing with imbalanced data.[13]

In addition, Decision Tree (DT) was applied for its ability to handle both categorical and numerical data, as well as for its interpretability. DT works by building rule-based decision trees that allow the identification of key factors influencing student graduation.[14] Meanwhile, K-Nearest Neighbors (KNN) was used as a comparative model because it operates based on similarity between data samples, making it effective for relatively small sample sizes. KNN also has advantages in capturing non-linear relationships in data, providing additional insights into the factors affecting graduation.[15]

To improve prediction performance, this study applied Voting and Stacking ensemble methods with Logistic Regression. The Voting method acts as an ensemble technique that combines predictions from several models to enhance classification accuracy.[16] Stacking with Logistic Regression is used as a meta-classifier to combine predictions from base classifiers and produce a more accurate final decision.[17] Logistic Regression was selected as the meta-classifier due to its good generalization ability and its effectiveness in both binary and multi-class classification tasks.[18]

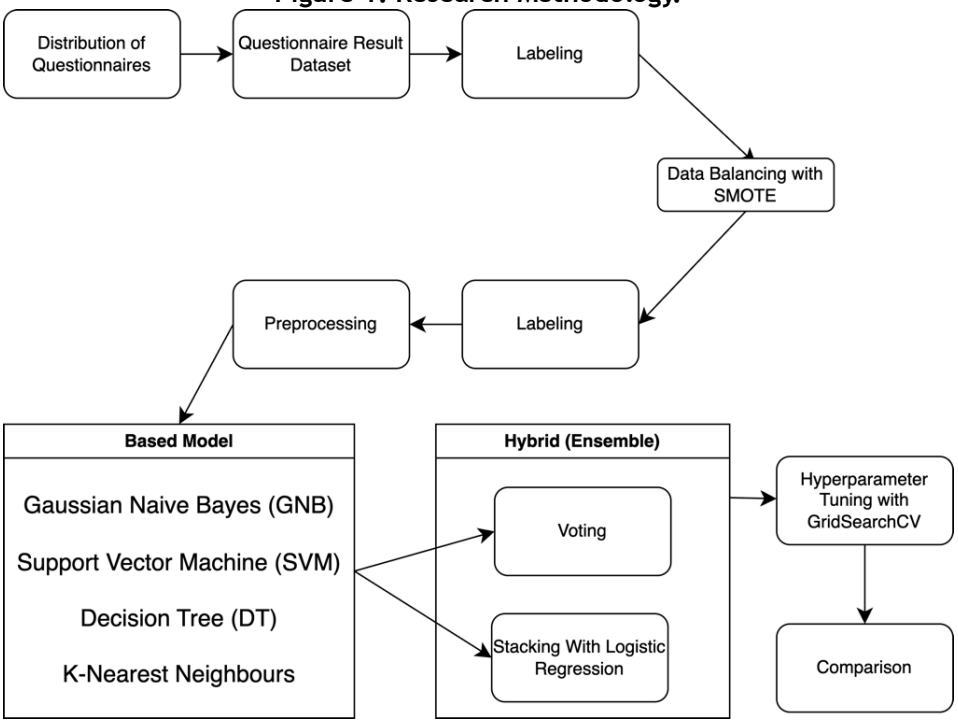
o address data imbalance, this study adopted SMOTE (Synthetic Minority Over-sampling Technique), which generates synthetic samples for the minority class, resulting in a more balanced dataset. Data imbalance often causes models to be biased toward the majority class, so applying SMOTE allows the model to learn more effectively from the underrepresented category.[19] urthermore, this study used GridSearchCV for hyperparameter tuning, aimed at finding the best parameter combinations for each model. This tuning process is crucial to avoid underfitting or overfitting and to ensure that the model performs optimally.[20]

Compared to previous studies, this research integrates several best-practice approaches, including the use of SMOTE to address class imbalance, the implementation of ensemble methods such as voting and stacking to boost model performance, and the use of GridSearchCV for model parameter optimization. Therefore, this study presents a more comprehensive approach to predicting student graduation, with the goal of producing more accurate results than the individual methods used in earlier research. Therefore, the objective of this study is to develop a highly accurate predictive model for student graduation timeliness by applying ensemble learning techniques integrated with SMOTE and GridSearchCV optimization.

METHODS

The research process was structured into several systematic stages to ensure the development of an accurate and reliable predictive model. These stages are visually summarized in Figure 1, which outlines the flow from data collection to model evaluation.

Figure 1. Research Methodology.



This research applied a quantitative predictive approach, designed to develop a robust machine learning model for predicting student graduation timeliness. As a predictive study, it focused on building and evaluating algorithms that can learn patterns from historical data and generalize to future cases. The choice of this method is grounded in its suitability for educational data analysis, especially where classification tasks such as on-time or delayed graduation are involved.

Sample

The population of this study consisted of students and alumni from Universitas Islam Riau, specifically those who had completed at least one academic year or graduated between 2019

and 2024. From this population, a total of 160 respondents were selected using purposive sampling, with the inclusion criteria being the completeness of academic records and willingness to participate in the survey.
The questionnaire was designed to gather both academic and non-academic information that may influence student graduation. The academic aspects included academic performance, Grade Point Average (GPA), total credits taken, and attendance rate, while the non-academic aspects covered psychological factors, social support, and learning motivation. Table 1 presents the questionnaire distributed to students of Universitas Islam Riau.

Table 1. A List of Questions.

Academic Factors	
1	What is your average GPA so far?
2	How often do you attend lectures?
3	Are you actively involved in academic activities (e.g., class discussions, group assignments)?
4	On average, how much time do you spend studying each day?
Non Academic	
1	How much social support do you receive from your family?
2	Do you currently have a part-time job?
3	What is your stress level during your college years?
4	Do you have adequate access to learning facilities (e.g., library, internet)?
5	Are you involved in any organizational or extracurricular activities on campus?

Data Labeling
After data preprocessing, a labeling process was conducted to classify students into two categories:
- "On-Time Graduation" (label = 1): students who completed their studies within the standard duration as defined by the university curriculum (typically 4 years or less for undergraduate programs).
- "Not On-Time Graduation" (label = 0): students who exceeded the standard study duration.

This classification was based on the graduation year and the year of enrollment recorded in the academic database. By calculating the difference between these two timestamps, each student was automatically assigned a binary label representing their graduation timeliness. This labeling served as the target variable for the machine learning models.

Preprocessing
After the questionnaire is labeled, a preprocessing stage is carried out to ensure data quality before further analysis. This stage involves the use of label coding. Label coding is a data preprocessing technique used to convert categorical text data into numeric values.[21] This is done by assigning a specific number to each unique category within a feature. This technique is important because most machine learning algorithms cannot process textual data directly. By converting categories into numerical representations, models can interpret and process the information mathematically. Label encoding is especially useful when categorical data is ordinal or has a limited number of categories. The use of this technique improves processing efficiency and allows algorithms to learn from the data more effectively.[22]
In addition, one of the challenges in predictive analysis is the imbalance in the number of samples between classes. Therefore, SMOTE (Synthetic Minority Over-sampling Technique) was applied to balance the number of samples between the "On-Time Graduation" and "Not On-Time Graduation" classes. This technique aims to generate synthetic data to strengthen the minority class, reducing model bias toward the majority class and improving predictive accuracy. Once the data was balanced, the next step was the modeling process.[23]

Model Development
In the modeling and classification phase, several machine learning algorithms were used to build the predictive model. The models tested included Gaussian Naïve Bayes (GNB), Linear Support Vector Machine (SVM), Decision Tree, and K-Nearest Neighbors (KNN). These four algorithms were selected due to their individual strengths in handling various types of data.

Subsequently, ensemble learning techniques were applied to improve model performance. Two main ensemble learning approaches used were voting and stacking. In the voting technique, individual models are combined using majority voting to enhance prediction accuracy .[24] In contrast, stacking combines the results of several base models and further analyzes them using Logistic Regression as a meta-learner to produce a more accurate final prediction.[25] Table 2 presents the models used in this study.

Table 2. Model Testing.

NO	Model
1	Gaussian Naive Bayes
2	Support Vector Machine
3	Decision Tree
4	K-Nearest Neighbours
5	Ensemble With Voting
6	Ensemble With Stacking
7	Ensemble With Voting + GridSearch
8	Ensemble With Stacking + GridSearch

After the model was built, hyperparameter optimization was performed using GridSearchCV. This technique is used to find the best parameter combinations for each algorithm applied, with the goal of improving model performance.[26] Once the optimization process was completed, the models were evaluated and compared to determine the best-performing algorithm for predicting student graduation based on the collected dataset. The final results of this process were used for model comparison, in which the model with the best performance was recommended as the optimal approach for analyzing the factors that influence whether students graduate on time or not.

Model Evaluation

The performance of each predictive model was evaluated using several standard classification metrics, namely accuracy, precision, recall, and F1-score. These metrics were chosen to provide a comprehensive assessment of the model's ability to correctly classify students based on their graduation timeliness.

- Accuracy measures the overall correctness of the model, calculated as the ratio of correctly predicted instances to the total number of predictions.
- Precision evaluates the proportion of true positive predictions among all predicted positives, highlighting the model's ability to avoid false alarms.
- Recall assesses the model's sensitivity by measuring how many actual positive cases were correctly predicted.
- F1-score, as the harmonic mean of precision and recall, provides a balanced view of the model's effectiveness, especially when dealing with class imbalance.

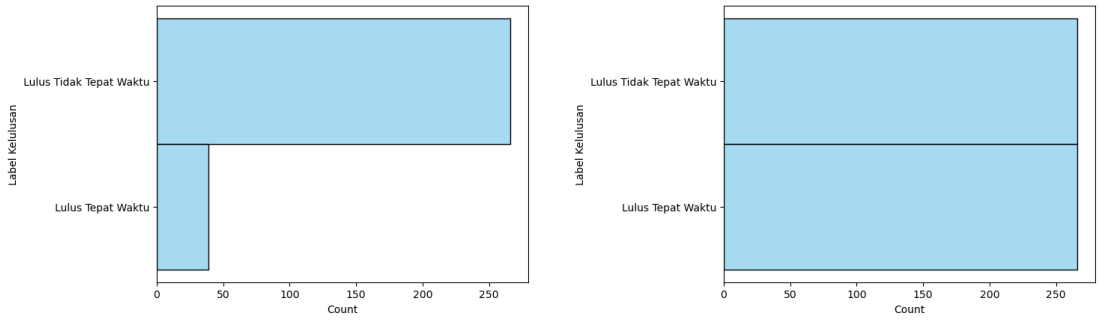
In addition, a confusion matrix was generated to visualize the performance in terms of true positives, true negatives, false positives, and false negatives. This matrix helps to identify the types of classification errors made by the model. Evaluation was conducted on a test dataset comprising 30% of the total samples, which was separated during the initial data splitting process to ensure objective performance assessment.

All evaluations were conducted using the Scikit-learn library in Python, ensuring standardized computation and reproducibility. These metrics were used to compare the baseline models with ensemble approaches, particularly those optimized using GridSearchCV, to determine the most effective model in predicting graduation timeliness.

RESULTS

The first stage after the data is obtained is to carry out the data labeling process. Figure 2 is a distribution based on labels.

Figure 2. Label Distribution Before and After Oversampling.



2A. Before 2B. After

Based on Figure 2, there are two visualizations in the form of horizontal bar charts showing the number of students according to graduation labels: *Lulus Tepat Waktu* (On-Time Graduation) and *Lulus Tidak Tepat Waktu* (Not On-Time Graduation). In chart 2A, it is evident that the majority of students fall into the "Not On-Time Graduation" category, while only a small portion successfully graduate on time. This imbalance indicates that the dataset has an uneven class distribution, which can cause machine learning models to become biased toward the majority class.

To address this issue, an oversampling technique using SMOTE (Synthetic Minority Over-sampling Technique) was applied. SMOTE works by synthesizing new data points for the minority class based on existing characteristics, thereby creating realistic additional data. The results of this process are shown in chart 2B, where the number of data points in both categories becomes balanced. The application of SMOTE is crucial to ensure that the trained model can learn patterns from both classes proportionally, thereby improving the model's predictive ability and generalization to new data. After the labeling process, the next step is model development. Figure 3 presents the confusion matrix from the ensemble model using the stacking technique with GridSearchCV.

Figure 3. Confusion Matrix

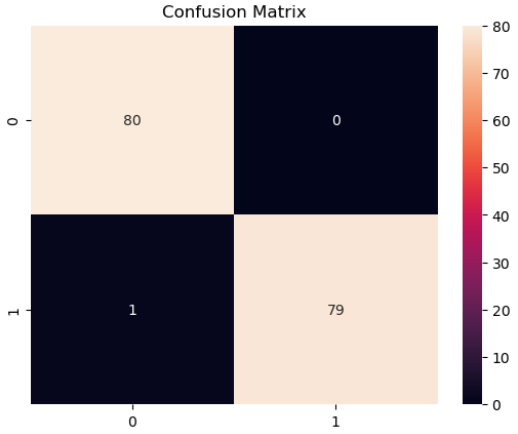


Figure 3 is a visualization of the confusion matrix used to evaluate the performance of the classification model for two classes: “On-Time Graduation” (label 1) and “Not On-Time Graduation” (label 0). The matrix consists of four main components: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN).

Based on the figure, the model successfully classified 80 data points correctly as "Not On-Time Graduation" (TN) and 79 data points correctly as "On-Time Graduation" (TP). There was only one misclassification, where a data point that should have been labeled as "On-Time Graduation" was incorrectly predicted as "Not On-Time Graduation" (FN). No data points were incorrectly predicted as "On-Time Graduation" when they actually belonged to the "Not On-Time Graduation" class (FP = 0).

Overall, these results indicate that the model has very high accuracy, with an extremely low error rate. This suggests that the model is highly effective in distinguishing between the two classes, especially after applying data balancing using the SMOTE technique. The next section, presented in Figure 4, is the classification report.

Figure 4. Classification Report

```
                          precision    recall  f1-score   support

      Lulus Tepat Waktu       0.99      1.00      0.99        80
Lulus Tidak Tepat Waktu       1.00      0.99      0.99        80

               accuracy                           0.99       160
              macro avg       0.99      0.99      0.99       160
           weighted avg       0.99      0.99      0.99       160

Accuracy : 0.99375
-----------------------------------------
```

Figure 4 displays the performance evaluation results of the classification model using evaluation metrics such as precision, recall, and F1-score for the two classes: Lulus Tepat Waktu (On-Time Graduation) and Lulus Tidak Tepat Waktu (Not On-Time Graduation). The model demonstrated very high performance, with precision and recall values ranging from 0.99 to 1.00 for both classes. This indicates that the model is highly accurate in identifying students who graduate on time as well as those who do not, with a very low error rate.
The F1-score, which also reached 0.99 for both classes, reflects the model's excellent balance between precision and recall. Overall, the model achieved an accuracy of 99.375%, meaning that out of 160 test data points, only one was misclassified. The macro average and weighted average values also show high consistency across all metrics, affirming that the model is not biased toward either class. These results indicate that the classification model developed in this study is highly reliable and effective in predicting students' on-time graduation, especially after the dataset was balanced using the SMOTE oversampling technique. The overall model testing results are presented in Table 3.

Table 3. Overall Test Results.

NO	Model	Accuracy	Precision	Recall	F1-Score
1	Gaussian Naive Bayes	95%	95%	95%	95%
2	Support Vector Machine	96%	95%	96%	95%
3	Decision Tree	93%	93%	93%	93%
4	K-Nearest Neighbours	94%	94%	94%	94%
5	Ensemble With Voting	98%	98%	98%	98%
6	Ensemble With Stacking	98%	98%	98%	98%
7	Ensemble With Voting + GridSearch	99%	99%	99%	99%
8	Ensemble With Stacking + GridSearch	99%	99%	99%	99%

Table 3 presents the performance evaluation results of various classification models applied to student graduation data, based on metrics such as accuracy, precision, recall, and F1-score. In general, all models demonstrated relatively high performance, with accuracy levels above 90%. The Gaussian Naive Bayes model achieved an accuracy of 95%, followed by Support Vector Machine (SVM) with 96%, while the Decision Tree performed slightly lower at 93%. The K-Nearest Neighbours (KNN) model recorded an accuracy of 94%.
Ensemble models showed superior performance. Both the Voting and Stacking ensemble methods achieved an accuracy of 98%, with consistently high values for precision, recall, and F1-score. Further improvement was obtained through the combination of ensemble learning and hyperparameter tuning using GridSearch, specifically in the Ensemble Voting + GridSearch and Ensemble Stacking + GridSearch models, both of which reached the highest accuracy of 99% across all evaluation metrics. These results indicate that the ensemble approach, particularly when optimized with GridSearch, provides the most reliable and accurate classification performance in predicting on-time student graduation compared to other models.

DISCUSSION

The findings of this study reveal that the integration of SMOTE, ensemble learning, and GridSearchCV optimization yields a highly accurate predictive model for student graduation timeliness. The stacking ensemble model optimized with GridSearchCV achieved an accuracy of 99.37%, outperforming all baseline and ensemble-only models. The performance across all

evaluation metrics—precision, recall, and F1-score—exceeded 0.99, indicating near-perfect classification.

This high performance is attributable to two key factors. First, the use of ensemble learning, particularly stacking, effectively combined the strengths of several base classifiers (Naïve Bayes, SVM, Decision Tree, and KNN), resulting in a model that generalized well across both majority and minority classes. Second, the application of SMOTE helped address the imbalanced dataset by generating synthetic samples for the underrepresented class (on-time graduates), which significantly improved model sensitivity and fairness.

A comparative analysis with previous studies is presented in Table 4, which summarizes the accuracy achieved by different research approaches in predicting student graduation:

Table 4. Comparison with Previous Research.

NO	Researcher	Model	Accuracy
1	Rahmaddeni, et al. 2022[27]	XGBoost + SVM	79%
2	Ulfah, et al. 2022[28]	SVM + GridSearchCV	95%
3	Rahmiati, et al. 2023[29]	Naive Bayes	95%
4	Herianto, et al. 2024[25]	SMLOS (Stacking Machine Learning Optuna SMOTE)	95%
5	Van FC, et al. 2025[17]	Stacking + Adaboost + Hyperparameter Tuning + SMOTE	97%
6	Suandi, et al. 2024[24]	Majority Voting + SMOTE	97%
7	Putra, et al. 2025[30]	Stacking + SMOTE	88%
8	Anam, et al. 2025[16]	Voting Hard + SMOTE	94%
9	This Study	SMOTE + Ensemble With Stacking + GridSearch	99%

Compared to these studies, our approach achieves the highest accuracy. Notably, while Van FC et al. reached 97% with a combination of stacking, Adaboost, and SMOTE, our integration of GridSearchCV further refined model performance by tuning the hyperparameters, which likely contributed to the observed performance gain.[17] Similarly, Herianto et al., using SMLOS, reported 95%, yet lacked the layered optimization pipeline applied in our research. [25]

These results also reaffirm findings in other domains (e.g., Alemerien et al.) where model tuning via GridSearchCV significantly enhanced performance, especially in datasets with complexity and imbalance. [20] The consistency across metrics and near-zero misclassification rate demonstrates the reliability and stability of the proposed model for practical implementation.

From an academic management perspective, this model can serve as a decision-support system for identifying students at risk of delayed graduation, enabling universities to design personalized interventions and advising systems. By integrating such a predictive model into academic dashboards, institutions can shift toward more proactive and data-driven academic policies.

CONCLUSIONS

This study aimed to develop a predictive model for student graduation timeliness by integrating SMOTE, ensemble learning techniques (voting and stacking), and hyperparameter tuning using GridSearchCV. The findings confirm that this integrative approach is highly effective, as evidenced by the stacking ensemble model achieving a predictive accuracy of 99.37%, with precision, recall, and F1-scores above 0.99.

By addressing class imbalance through SMOTE and refining model performance using GridSearchCV, the study succeeded in enhancing classification quality beyond the capabilities of individual machine learning models. Compared to previous related studies, the proposed model demonstrates superior performance and robustness, proving its potential as a reliable framework for academic risk detection.

Thus, the objective of this study—to produce an accurate, optimized model for predicting on-time graduation—has been achieved. The resulting model can be integrated into academic decision-support systems to assist universities in monitoring students' progress and initiating early interventions for those at risk of delayed graduation.

REFERENCES

1. Bakri R, Astuti NP, Ahmar AS. Machine Learning Algorithms with Parameter Tuning to Predict Students' Graduation-on-time: A Case Study in Higher Education. J Appl Sci Eng Technol Educ. 2022 Dec 30;4(2):259–65.

2. Casanova VS, Pullido ML. Factors Of Graduate Students' Attrition And Retention In Occidental Mindoro State College Graduate School. IJERSC. 3(2):826–31.

3. López-Meneses E, López-Catalán L, Pelícano-Piris N, Mellado-Moreno PC. Artificial Intelligence in Educational Data Mining and Human-in-the-Loop Machine Learning and Machine Teaching: Analysis of Scientific Knowledge. Applied Sciences. 2025 Jan 14;15(2):1–21.

4. Ersozlu Z, Taheri S, Koch I. A review of machine learning methods used for educational data. Educ Inf Technol. 2024 Nov;29(16):22125–45.

5. Taye MM. Understanding of Machine Learning with Deep Learning: Architectures, Workflow, Applications and Future Directions. Computers. 2023 Apr 25;12(5):1–26.

6. Mehta S. Playing Smart with Numbers: Predicting Student Graduation Using the Magic of Naive Bayes. ITALIC. 2023 Nov 23;2(1):60–75.

7. Darenoh NV, Bachtiar FA, Perdana RS. Prediction of On-time Student Graduation with Deep Learning Method: -. J ICT Res Appl. 2024 Jun 27;18(1):1–20.

8. Desfiandi A, Soewito B. Student Graduation Time Prediction Using Logistic Regression, Decision Tree, Support Vector, and Adaboost Ensemble Learning. International Journal of Information System and Computer Science. 7(3):195–9.

9. Haikal MF, Palupi I. Predicting Employability of University Graduates Using Support Vector Machine Classification. Building of Informatics, Technology and Science. 2024;6(2).

10. Rismayati R, Ismarmiaty I, Hidayat S. Esemble Implementation for Predicting Student Graduation withClassification Algorithm. IJECSA. 1(1):35–42.

11. Anam MK, Putra PP, Malik RA, Putra TA, Elva Y, Mahessya RA, et al. Enhancing the Performance of Machine Learning Algorithm for Intent Sentiment Analysis on Village Fund Topic. Journal of Applied Data Sciences. 2025;6(2):1102–15.

12. Sharma H, Pangaonkar S, Gunjan R, Rokade P. Sentimental Analysis of Movie Reviews Using Machine Learning. Shah H, Patel R, Patel N, Buyya R, Chatterjee I, editors. ITM Web Conf. 2023;53:02006.

13. Anam MK, Firdaus MB, Suandi F, Lathifah, Nasution T, Fadly S. Performance Improvement of Machine Learning Algorithm Using Ensemble Method on Text Mining. In: 2024 International Conference on Future Technologies for Smart Society (ICFTSS) [Internet]. Kuala Lumpur, Malaysia: IEEE; 2024 [cited 2025 Mar 16]. p. 90–5. Available from: https://ieeexplore.ieee.org/document/10691363/

14. Putra PP, Anam MK, Defit S, Yunianta A. Enhancing the Decision Tree Algorithm to Improve Performance Across Various Datasets. intensif. 2024 Aug 1;8(2):200–12.

15. Danny M, Muhidin A, Jamal A. Application of the K-Nearest Neighbor Machine Learning Algorithm to Preduct Sales of Best-Selling Products. Brilliance. 2024 Jun 28;4(1):255–64.

16. Anam MK, Lestari TP, Yenni H, Nasution T, Firdaus MB. Enhancement of Machine Learning Algorithm in Fine-grained Sentiment Analysis Using the Ensemble. ECTI-CIT Transactions. 2025 Mar 8;19(2):159–67.

17. Van Fc LL, Anam MK, Bukhori S, Mahamad AK, Saon S, Nyoto RLV. The Development of Stacking Techniques in Machine Learning for Breast Cancer Detection. J Appl Data Sci. 2024 Jan 1;6(1):71–85.

18. Munthe IR, Rambe BH, Hanum F, Amanda AT, Hutagaol ASR, Harianto R. Implementation of Stacking Technique Combining Machine Learning and Deep Learning Algorithms Using SMOTE to Improve Stock Market Prediction Accuracy. J Appl Data Sci. 2024 Dec 1;5(4):2079–91.

19. Anam MK, Van Fc LL, Hamdani H, Rahmaddeni R, Junadhi J, Firdaus MB, et al. Sara Detection on Social Media Using Deep Learning Algorithm Development. JAETS. 2024 Dec 15;6(1):225–37.

20. Alemerien K, Alsarayreh S, Altarawneh E. Diagnosing Cardiovascular Diseases using Optimized Machine Learning Algorithms with GridSearchCV. J Appl Data Sci. 2024 Dec 1;5(4):1539–52.

21. Bolikulov F, Nasimov R, Rashidov A, Akhmedov F, Cho YI. Effective Methods of Categorical Data Encoding for Artificial Intelligence Algorithms. Mathematics. 2024 Aug 18;12(16):1–21.

22. Hien DTT, Thi C, Kim T, The D, Nguyen C. Optimize the Combination of Categorical Variable Encoding and Deep Learning Technique for the Problem of Prediction of Vietnamese Student Academic Performance. IJACSA. 2020;11(11):274–80.

23. Anam MK, Munawir M, Efrizoni L, Fadillah N, Agustin W, Syahputra I, et al. Improved Performance of Hybrid GRU-BiLSTM for Detection Emotion on Twitter Dataset. J Appl Data Sci. 2024 Jan 1;6(1):354–65.

24. Suandi F, Anam MK, Firdaus MB, Fadli S, Lathifah L, Yumami E, et al. Enhancing Sentiment Analysis Performance Using SMOTE and Majority Voting in Machine Learning Algorithms. In: Lumombo L, Rahmi A, Suwarno S, Ardi N, Kurniawan DE, editors. Proceedings of the 7th International Conference on Applied Engineering (ICAE 2024) [Internet]. Dordrecht: Atlantis Press International BV; 2024 [cited 2025 Mar 25]. p. 126–38. (Advances in Engineering Research; vol. 251). Available from: https://www.atlantis-press.com/doi/10.2991/978-94-6463-620-8_10

25. Herianto H, Kurniawan B, Hartomi ZH, Irawan Y, Anam MK. Machine Learning Algorithm Optimization using Stacking Technique for Graduation Prediction. J Appl Data Sci. 2024 Sep 1;5(3):1272–85.

26. Jumanto J, Rofik R, Sugiharti E, Alamsyah A, Arifudin R, Prasetiyo B, et al. Optimizing Support Vector Machine Performance for Parkinson's Disease Diagnosis Using GridSearchCV and PCA-Based Feature Extraction. J Inf Syst Eng Bus Intell. 2024 Feb 28;10(1):38–50.

27. Rahmaddeni R, Anam MK, Irawan Y, Susanti S, Jamaris M. Comparison of Support Vector Machine and XGBSVM in Analyzing Public Opinion on Covid-19 Vaccination. Ilk J Ilm. 2022 Apr 30;14(1):32–8.

28. Ulfah AN, Anam MK, Sidratul Munti NY, Yaakub S, Firdaus MB. Sentiment Analysis of the Convict Assimilation Program on Handling Covid-19. JUITA. 2022 Nov 14;10(2):209–16.

29. Rahmiati R, Anam MK, Paradila D, Mardainis M, Machdalena M. Application of Naïve Bayes Algorithm for Non-Cash Food Assistance Recipients in Kampar Regency. SinkrOn. 2023 Jan 4;8(1):433–41.

30. Putra PP, Anam MK, Chan AS, Hadi A, Hendri N, Masnur A. Optimizing Sentiment Analysis on Imbalanced Hotel Review Data Using SMOTE and Ensemble Machine Learning Techniques. J Appl Data Sci. 2025 May 1;6(2):921–35.

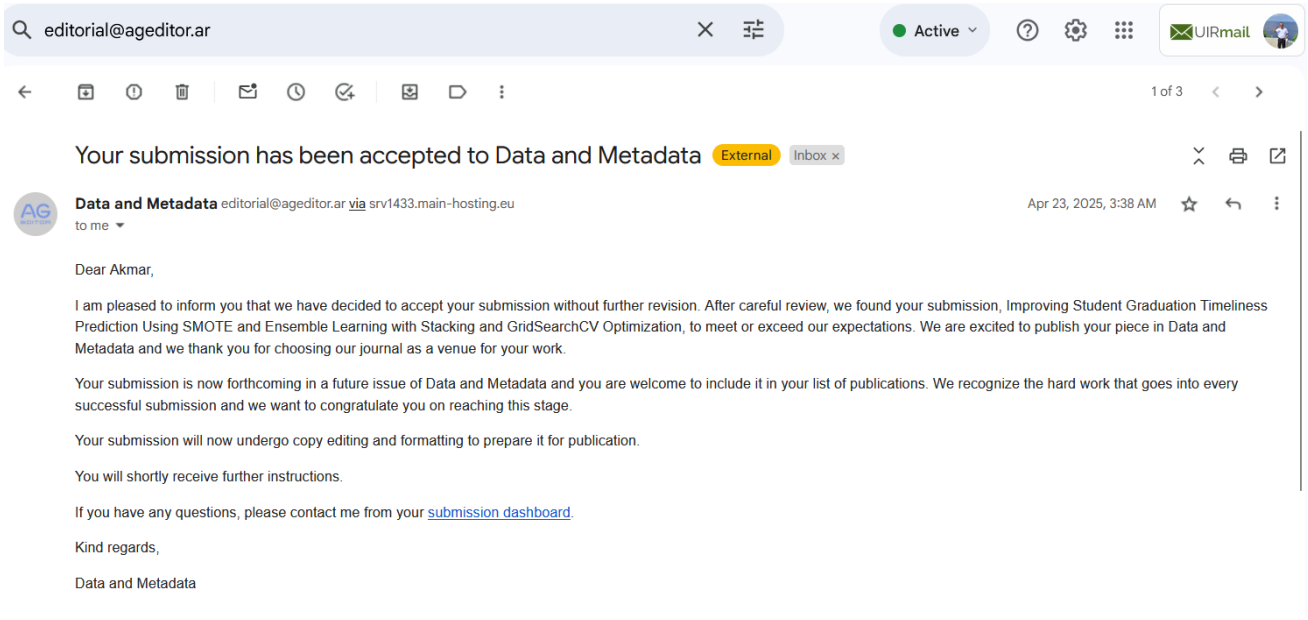
FINANCING
The authors would like to express their sincere gratitude to Universitas Islam Riau (UIR) for the financial support provided for this research. This support was essential for the smooth execution and successful completion of the study.

CONFLICT OF INTEREST
None

AUTHORSHIP CONTRIBUTION:
Conceptualization: Akmar Efendi
Data curation: Akmar Efendi
Formal analysis: Akmar Efendi
Research: Akmar Efendi, Gunadi Widi Nurcahyo, and Iskandar Fitri
Methodology: Akmar Efendi, Gunadi Widi Nurcahyo, and Iskandar Fitri
Project management: Akmar Efendi
Resources: Akmar Efendi
Software: Akmar Efendi
Supervision: Iskandar Fitri
Validation: Iskandar Fitri
Display: Iskandar Fitri
Drafting - original draft: Akmar Efendi
Writing - proofreading and editing: Gunadi Widi Nurcahyo

5. Accept Artikel (23 April 2025)

- **Sreenshoot Bukti Accepted**



editorial@ageditor.ar

Active

UIRmail

1 of 3

Your submission has been accepted to Data and Metadata External Inbox ×

Data and Metadata editorial@ageditor.ar via srv1433.main-hosting.eu
to me

Apr 23, 2025, 3:38 AM

Dear Akmar,

I am pleased to inform you that we have decided to accept your submission without further revision. After careful review, we found your submission, Improving Student Graduation Timeliness Prediction Using SMOTE and Ensemble Learning with Stacking and GridSearchCV Optimization, to meet or exceed our expectations. We are excited to publish your piece in Data and Metadata and we thank you for choosing our journal as a venue for your work.

Your submission is now forthcoming in a future issue of Data and Metadata and you are welcome to include it in your list of publications. We recognize the hard work that goes into every successful submission and we want to congratulate you on reaching this stage.

Your submission will now undergo copy editing and formatting to prepare it for publication.

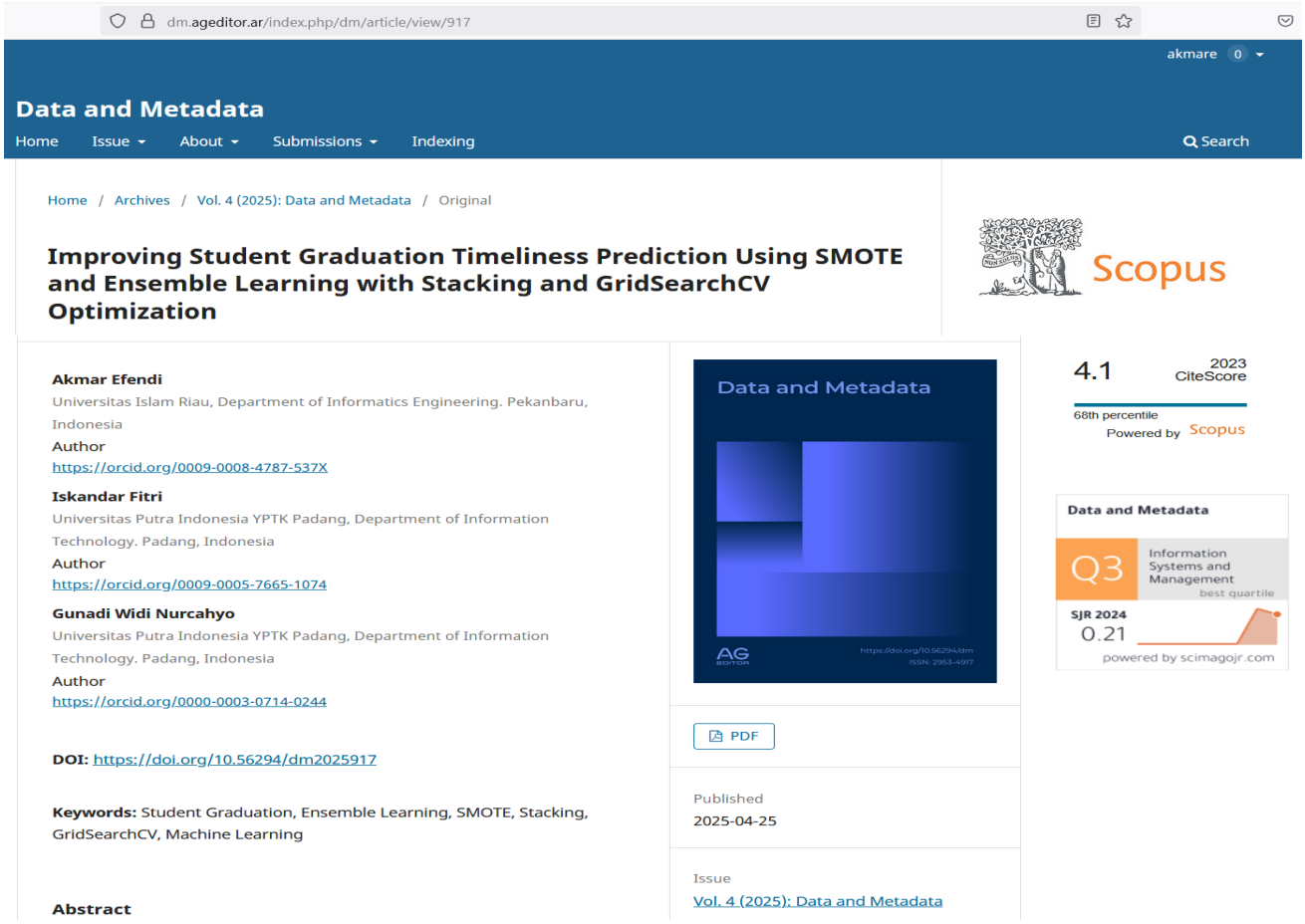
You will shortly receive further instructions.

If you have any questions, please contact me from your submission dashboard.

Kind regards,

Data and Metadata

6. Publish Artikel di Jurnal Data and Metadata (25 April 2025)



dm.ageditor.ar/index.php/dm/article/view/917

akmare 0

Data and Metadata

Home Issue About Submissions Indexing Search

Home / Archives / Vol. 4 (2025): Data and Metadata / Original

Improving Student Graduation Timeliness Prediction Using SMOTE and Ensemble Learning with Stacking and GridSearchCV Optimization

Akmar Efendi
Universitas Islam Riau, Department of Informatics Engineering. Pekanbaru, Indonesia
Author
https://orcid.org/0009-0008-4787-537X

Iskandar Fitri
Universitas Putra Indonesia YPTK Padang, Department of Information Technology. Padang, Indonesia
Author
https://orcid.org/0009-0005-7665-1074

Gunadi Widi Nurcahyo
Universitas Putra Indonesia YPTK Padang, Department of Information Technology. Padang, Indonesia
Author
https://orcid.org/0000-0003-0714-0244

DOI: https://doi.org/10.56294/dm2025917

Keywords: Student Graduation, Ensemble Learning, SMOTE, Stacking, GridSearchCV, Machine Learning

Abstract

Data and Metadata

PDF

Published
2025-04-25

Issue
Vol. 4 (2025): Data and Metadata

Scopus

4.1 2023 CiteScore
68th percentile
Powered by Scopus

Data and Metadata
Q3 Information Systems and Management
best quartile
SJR 2024
0.21
powered by scimagojr.com

Introduction: Timely graduation is a key performance indicator in higher education. This study aims to improve the accuracy of predicting student graduation timeliness using ensemble machine learning techniques combined with SMOTE and hyperparameter optimization.
Methods: This is a quantitative predictive study. The population includes students and alumni of Universitas Islam Riau. A sample of 160 respondents was obtained via purposive sampling. Data were collected using structured questionnaires encompassing academic variables (e.g., GPA, credits, attendance) and non-academic variables (e.g., stress, social support, extracurricular activity). After preprocessing and label encoding, SMOTE was applied to balance class distribution. Several classifiers (Naïve Bayes, SVM, Decision Tree, KNN) were tested, with ensemble learning (voting and stacking) implemented and optimized using GridSearchCV.
Results: The stacking ensemble model optimized with GridSearchCV achieved the highest performance with an accuracy of 99.37%, precision and recall above 0.99, and minimal misclassification. This outperformed individual models and previous approaches in the literature.
Conclusions: The integration of SMOTE, ensemble methods, and GridSearchCV significantly enhances predictive accuracy for student graduation timeliness. The resulting model provides a robust framework for academic risk detection and early intervention.

Section
Original

How to Cite
Efendi A, Fitri I, Nurcahyo GW. Improving Student Graduation Timeliness Prediction Using SMOTE and Ensemble Learning with Stacking and GridSearchCV Optimization. Data and Metadata [Internet]. 2025 Apr. 25 [cited 2025 May 16];4:917. Available

References

1. Bakri R, Astuti NP, Ahmar AS. Machine Learning Algorithms with Parameter Tuning to Predict Students' Graduation-on-time: A Case Study in Higher Education. J Appl Sci Eng Technol Educ. 2022 Dec 30;4(2):259–65.

2. Casanova VS, Pullido ML. Factors Of Graduate Students' Attrition And Retention In Occidental Mindoro State College Graduate School. IJERSC. 3(2):826–31.

3. López-Meneses E, López-Catalán L, Pelícano-Piris N, Mellado-Moreno PC. Artificial Intelligence in Educational Data Mining and Human-in-the-Loop Machine Learning and Machine Teaching: Analysis of Scientific Knowledge. Applied Sciences. 2025 Jan 14;15(2):1–21.

4. Ersozlu Z, Taheri S, Koch I. A review of machine learning methods used for educational data. Educ Inf Technol. 2024 Nov;29(16):22125–45.

5. Taye MM. Understanding of Machine Learning with Deep Learning: Architectures, Workflow, Applications and Future Directions. Computers. 2023 Apr 25;12(5):1–26.

6. Mehta S. Playing Smart with Numbers: Predicting Student Graduation Using the Magic of Naive Bayes. ITALIC. 2023 Nov 23;2(1):60–75.

7. Darenoh NV, Bachtiar FA, Perdana RS. Prediction of On-time Student Graduation with Deep Learning Method: -. J ICT Res Appl. 2024 Jun 27;18(1):1–20.

8. Desfiandi A, Soewito B. Student Graduation Time Prediction Using Logistic Regression, Decision Tree, Support Vector, and Adaboost Ensemble Learning. International Journal of Information System and Computer Science. 7(3):195–9.

9. Haikal MF, Palupi I. Predicting Employability of University Graduates Using Support Vector Machine Classification. Building of Informatics, Technology and Science. 2024;6(2).

10. Rismayati R, Ismarmiaty I, Hidayat S. Esemble Implementation for Predicting Student Graduation withClassification Algorithm. IJECSA. 1(1):35–42.

11. Anam MK, Putra PP, Malik RA, Putra TA, Elva Y, Mahessya RA, et al. Enhancing the Performance of Machine Learning Algorithm for Intent Sentiment Analysis on Village Fund Topic. Journal of Applied Data Sciences. 2025;6(2):1102–15.

12. Sharma H, Pangaonkar S, Gunjan R, Rokade P. Sentimental Analysis of Movie Reviews Using Machine Learning. Shah H, Patel R, Patel N, Buyya R, Chatterjee I, editors. ITM Web Conf. 2023;53:02006.

13. Anam MK, Firdaus MB, Suandi F, Lathifah, Nasution T, Fadly S. Performance Improvement of Machine Learning Algorithm Using Ensemble Method on Text Mining. In: 2024 International Conference on Future Technologies for Smart Society (ICFTSS) [Internet]. Kuala Lumpur, Malaysia: IEEE; 2024 [cited 2025 Mar 16]. p. 90–5. Available from: https://ieeexplore.ieee.org/document/10691363/

14. Putra PP, Anam MK, Defit S, Yunianta A. Enhancing the Decision Tree Algorithm to Improve Performance Across Various Datasets. intensif. 2024 Aug 1;8(2):200–12.

15. Danny M, Muhidin A, Jamal A. Application of the K-Nearest Neighbor Machine Learning Algorithm to Preduct Sales of Best-Selling Products. Brilliance. 2024 Jun 28;4(1):255–64.

16. Anam MK, Lestari TP, Yenni H, Nasution T, Firdaus MB. Enhancement of Machine Learning Algorithm in Fine-grained Sentiment Analysis Using the Ensemble. ECTI-CIT Transactions. 2025 Mar 8;19(2):159–67.

17. Van Fc LL, Anam MK, Bukhori S, Mahamad AK, Saon S, Nyoto RLV. The Development of Stacking Techniques in Machine Learning for Breast Cancer Detection. J Appl Data Sci. 2024 Jan 1;6(1):71–85.

18. Munthe IR, Rambe BH, Hanum F, Amanda AT, Hutagaol ASR, Harianto R. Implementation of Stacking Technique Combining Machine Learning and Deep Learning Algorithms Using SMOTE to Improve Stock Market Prediction Accuracy. J Appl Data Sci. 2024 Dec 1;5(4):2079–91.

19. Anam MK, Van Fc LL, Hamdani H, Rahmaddeni R, Junadhi J, Firdaus MB, et al. Sara Detection on Social Media Using Deep Learning Algorithm Development. JAETS. 2024 Dec 15;6(1):225–37.

20. Alemerien K, Alsarayreh S, Altarawneh E. Diagnosing Cardiovascular Diseases using Optimized Machine Learning Algorithms with GridSearchCV. J Appl Data Sci. 2024 Dec 1;5(4):1539–52.

21. Bolikulov F, Nasimov R, Rashidov A, Akhmedov F, Cho YI. Effective Methods of Categorical Data Encoding for Artificial Intelligence Algorithms. Mathematics. 2024 Aug 18;12(16):1–21.

22. Hien DTT, Thi C, Kim T, The D, Nguyen C. Optimize the Combination of Categorical Variable Encoding and Deep Learning Technique for the Problem of Prediction of Vietnamese Student Academic Performance. IJACSA. 2020;11(11):274–80.

23. Anam MK, Munawir M, Efrizoni L, Fadillah N, Agustin W, Syahputra I, et al. Improved Performance of Hybrid GRU-BiLSTM for Detection Emotion on Twitter Dataset. J Appl Data Sci. 2024 Jan 1;6(1):354–65.

24. Suandi F, Anam MK, Firdaus MB, Fadli S, Lathifah L, Yumami E, et al. Enhancing Sentiment Analysis Performance Using SMOTE and Majority Voting in Machine Learning Algorithms. In: Lumombo L, Rahmi A, Suwarno S, Ardi N, Kurniawan DE, editors. Proceedings of the 7th International Conference on Applied Engineering (ICAE 2024) [Internet]. Dordrecht: Atlantis Press International BV; 2024 [cited 2025 Mar 25]. p. 126–38. (Advances in Engineering Research; vol. 251). Available from: https://www.atlantis-press.com/doi/10.2991/978-94-6463-620-8_10

25. Herianto H, Kurniawan B, Hartomi ZH, Irawan Y, Anam MK. Machine Learning Algorithm Optimization using Stacking Technique for Graduation Prediction. J Appl Data Sci. 2024 Sep 1;5(3):1272–85.

26. Jumanto J, Rofik R, Sugiharti E, Alamsyah A, Arifudin R, Prasetiyo B, et al. Optimizing Support Vector Machine Performance for Parkinson's Disease Diagnosis Using GridSearchCV and PCA-Based Feature Extraction. J Inf Syst Eng Bus Intell. 2024 Feb 28;10(1):38–50.

27. Rahmaddeni R, Anam MK, Irawan Y, Susanti S, Jamaris M. Comparison of Support Vector Machine and XGBSVM in Analyzing Public Opinion on Covid-19 Vaccination. Ilk J Ilm. 2022 Apr 30;14(1):32–8.

28. Ulfah AN, Anam MK, Sidratul Munti NY, Yaakub S, Firdaus MB. Sentiment Analysis of the Convict Assimilation Program on Handling Covid-19. JUITA. 2022 Nov 14;10(2):209–16.

Data and Metadata. 2025; 4:917
doi: 10.56294/dm2025917

ORIGINAL

Improving Student Graduation Timeliness Prediction Using SMOTE and Ensemble Learning with Stacking and GridSearchCV Optimization

Mejora de la Predicción de la Oportunidad de Graduación Estudiantil Utilizando SMOTE y Aprendizaje por Ensamblado con Stacking y Optimización mediante GridSearchCV

Akmar Efendi[1], Iskandar Fitri[2], Gunadi Widi Nurcahyo[2]

[1]Universitas Islam Riau, Department of Informatics Engineering. Pekanbaru, Indonesia.
[2]Universitas Putra Indonesia YPTK Padang, Department of Information Technology. Padang, Indonesia.

Cite as: Efendi A, Fitri I, Nurcahyo GW. Improving Student Graduation Timeliness Prediction Using SMOTE and Ensemble Learning with Stacking and GridSearchCV Optimization. Data and Metadata. 2025; 4:917. https://doi.org/10.56294/dm2025917

Submitted: 03-08-2024 **Revised:** 12-12-2024 **Accepted:** 01-05-2025 **Published:** 02-05-2025

Editor: Dr. Adrián Alejandro Vitón Castillo

Cooresponding author: Akmar Efendi

ABSTRACT

Introduction: timely graduation is a key performance indicator in higher education. This study aims to improve the accuracy of predicting student graduation timeliness using ensemble machine learning techniques combined with SMOTE and hyperparameter optimization.
Method: this is a quantitative predictive study. The population includes students and alumni of Universitas Islam Riau. A sample of 160 respondents was obtained via purposive sampling. Data were collected using structured questionnaires encompassing academic variables (e.g., GPA, credits, attendance) and non-academic variables (e.g., stress, social support, extracurricular activity). After preprocessing and label encoding, SMOTE was applied to balance class distribution. Several classifiers (Naïve Bayes, SVM, Decision Tree, KNN) were tested, with ensemble learning (voting and stacking) implemented and optimized using GridSearchCV.
Results: the stacking ensemble model optimized with GridSearchCV achieved the highest performance with an accuracy of 99,37 %, precision and recall above 0,99, and minimal misclassification. This outperformed individual models and previous approaches in the literature.
Conclusions: the integration of SMOTE, ensemble methods, and GridSearchCV significantly enhances predictive accuracy for student graduation timeliness. The resulting model provides a robust framework for academic risk detection and early intervention.

Keywords: Student Graduation; Ensemble Learning; SMOTE; Stacking; GridSearchCV; Machine Learning.

RESUMEN

Introducción: la graduación oportuna es un indicador clave del rendimiento en la educación superior. Este estudio tiene como objetivo mejorar la precisión en la predicción de la oportunidad de graduación estudiantil mediante técnicas de aprendizaje automático por ensamblado combinadas con SMOTE y optimización de hiperparámetros.
Método: este es un estudio cuantitativo predictivo. La población incluye a estudiantes y egresados de la Universitas Islam Riau. Se obtuvo una muestra de 160 participantes mediante muestreo intencional. Los datos se recopilaron utilizando cuestionarios estructurados que abarcan variables académicas (por ejemplo, promedio académico, créditos, asistencia) y no académicas (por ejemplo, estrés, apoyo social, participación extracurricular). Después del preprocesamiento y la codificación de etiquetas, se aplicó SMOTE para equilibrar la distribución de clases. Se probaron varios clasificadores (Naïve Bayes, SVM, Árbol de Decisión, KNN), y se

implementó aprendizaje por ensamblado (voting y stacking) optimizado mediante GridSearchCV.
Resultados: el modelo de ensamblado tipo stacking optimizado con GridSearchCV alcanzó el mejor desempeño, con una precisión del 99,37 %, valores de precisión y recall superiores a 0,99, y una tasa mínima de errores de clasificación. Este modelo superó a los modelos individuales y a enfoques previos en la literatura.
Conclusiones: la integración de SMOTE, métodos de ensamblado y GridSearchCV mejora significativamente la precisión predictiva para la graduación oportuna de los estudiantes. El modelo resultante proporciona un marco sólido para la detección de riesgos académicos y la intervención temprana.

Palabras clave: Graduación de estudiantes; Aprendizaje Conjunto; SMOTE; Apilamiento; GridSearchCV; Aprendizaje Automático.

INTRODUCTION

Timely graduation rates are one of the key indicators of success for higher education institutions.[1] Universitas Islam Riau (UIR), as one of the leading private universities in Indonesia, also faces challenges in improving its on-time graduation rates. Various academic and non-academic factors may influence student graduation, such as teaching quality, availability of learning facilities, social support, and students' psychological conditions. However, identifying and understanding these factors is often complex and requires an in-depth analytical approach.[2]

With the advancement of technology, the application of machine learning in educational data analysis has opened up new opportunities for discovering patterns that influence academic success.[3] This approach can handle large datasets and detect hidden patterns that are difficult to capture using traditional methods.[4,5] Therefore, this study aims to explore the influence of academic and non-academic factors on student graduation at UIR using various machine learning algorithms.

Previous research has explored different approaches to predicting student graduation using machine learning and deep learning methods. One study using the Naïve Bayes algorithm achieved an accuracy of 85 %, demonstrating that probabilistic methods can effectively handle data, although they have limitations in dealing with unstructured data of high complexity.[6] Meanwhile, the Deep Neural Network (DNN) approach achieved a higher accuracy of 87 %, utilizing its architecture to better capture complex data patterns.[7] Another study implemented ensemble learning using AdaBoost, which enhanced the performance of the Decision Tree and reached an F1-score of 82 %, confirming that boosting techniques can improve model generalization.[8]

A more complex approach was proposed by combining Support Vector Machine (SVM) with SMOTEN-ENN, Spearman's correlation, and randomized search, resulting in an accuracy of 73 %, although it still fell short compared to other methods.[9] The K-Nearest Neighbors (KNN) algorithm was also evaluated in another study and obtained the highest accuracy of 89 %, demonstrating the effectiveness of distance-based methods in handling graduation prediction tasks.[10]

This study employs a variety of machine learning algorithms to predict student graduation by leveraging the strengths of each method in handling different data characteristics. Multinomial Naïve Bayes (MNB) was chosen for its ability to manage text-based and probabilistic data, especially in processing questionnaire data that has undergone word-weighting stages.[11] MNB is effective in estimating the probability of categories based on word frequency, making it suitable for text-based analysis. Linear Support Vector Machine (SVM) was used due to its capability to find optimal hyperplanes for separating classes in high-dimensional data.[12] and its stable performance—especially when combined with data balancing techniques like SMOTE—helps improve model accuracy when dealing with imbalanced data.[13]

In addition, Decision Tree (DT) was applied for its ability to handle both categorical and numerical data, as well as for its interpretability. DT works by building rule-based decision trees that allow the identification of key factors influencing student graduation.[14] Meanwhile, K-Nearest Neighbors (KNN) was used as a comparative model because it operates based on similarity between data samples, making it effective for relatively small sample sizes. KNN also has advantages in capturing non-linear relationships in data, providing additional insights into the factors affecting graduation.[15]

To improve prediction performance, this study applied Voting and Stacking ensemble methods with Logistic Regression. The Voting method acts as an ensemble technique that combines predictions from several models to enhance classification accuracy.[16] Stacking with Logistic Regression is used as a meta-classifier to combine predictions from base classifiers and produce a more accurate final decision.[17] Logistic Regression was selected as the meta-classifier due to its good generalization ability and its effectiveness in both binary and multi-class classification tasks.[18]

To address data imbalance, this study adopted SMOTE (Synthetic Minority Over-sampling Technique), which generates synthetic samples for the minority class, resulting in a more balanced dataset. Data imbalance

often causes models to be biased toward the majority class, so applying SMOTE allows the model to learn more effectively from the underrepresented category.[19] Furthermore, this study used GridSearchCV for hyperparameter tuning, aimed at finding the best parameter combinations for each model. This tuning process is crucial to avoid underfitting or overfitting and to ensure that the model performs optimally.[20]

Compared to previous studies, this research integrates several best-practice approaches, including the use of SMOTE to address class imbalance, the implementation of ensemble methods such as voting and stacking to boost model performance, and the use of GridSearchCV for model parameter optimization. Therefore, this study presents a more comprehensive approach to predicting student graduation, with the goal of producing more accurate results than the individual methods used in earlier research. Therefore, the objective of this study is to develop a highly accurate predictive model for student graduation timeliness by applying ensemble learning techniques integrated with SMOTE and GridSearchCV optimization.

METHOD

The research process was structured into several systematic stages to ensure the development of an accurate and reliable predictive model. These stages are visually summarized in figure 1, which outlines the flow from data collection to model evaluation.

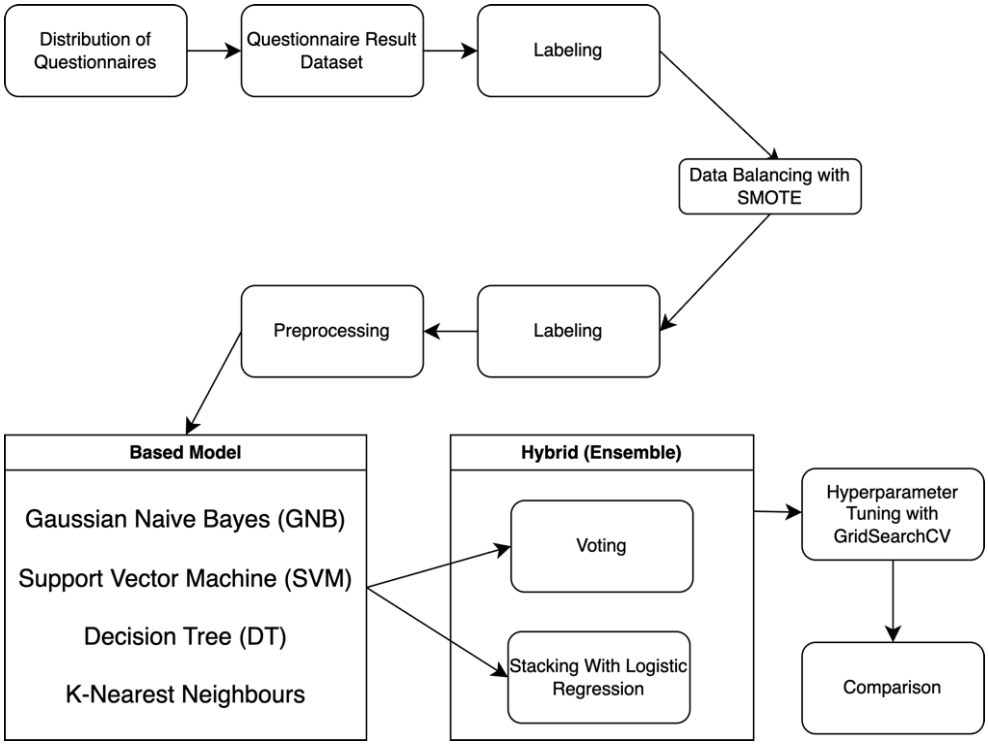


Figure 1. Research Methodology

This research applied a quantitative predictive approach, designed to develop a robust machine learning model for predicting student graduation timeliness. As a predictive study, it focused on building and evaluating algorithms that can learn patterns from historical data and generalize to future cases. The choice of this method is grounded in its suitability for educational data analysis, especially where classification tasks such as on-time or delayed graduation are involved.

Sample

The population of this study consisted of students and alumni from Universitas Islam Riau, specifically those who had completed at least one academic year or graduated between 2019 and 2024. From this population, a total of 160 respondents were selected using purposive sampling, with the inclusion criteria being the completeness of academic records and willingness to participate in the survey.

The questionnaire was designed to gather both academic and non-academic information that may influence student graduation. The academic aspects included academic performance, Grade Point Average (GPA), total credits taken, and attendance rate, while the non-academic aspects covered psychological factors, social support, and learning motivation. Table 1 presents the questionnaire distributed to students of Universitas Islam Riau.

Table 1. A List of Questions	
Academic Factors	
1	What is your average GPA so far?
2	How often do you attend lectures?
3	Are you actively involved in academic activities (e.g., class discussions, group assignments)?
4	On average, how much time do you spend studying each day?
Non Academic	
1	How much social support do you receive from your family?
2	Do you currently have a part-time job?
3	What is your stress level during your college years?
4	Do you have adequate access to learning facilities (e.g., library, internet)?
5	Are you involved in any organizational or extracurricular activities on campus?

Data Labeling

After data preprocessing, a labeling process was conducted to classify students into two categories:

- "On-Time Graduation" (label = 1): students who completed their studies within the standard duration as defined by the university curriculum (typically 4 years or less for undergraduate programs).
- "Not On-Time Graduation" (label = 0): students who exceeded the standard study duration.

This classification was based on the graduation year and the year of enrollment recorded in the academic database. By calculating the difference between these two timestamps, each student was automatically assigned a binary label representing their graduation timeliness. This labeling served as the target variable for the machine learning models.

Preprocessing

After the questionnaire is labeled, a preprocessing stage is carried out to ensure data quality before further analysis. This stage involves the use of label coding. Label coding is a data preprocessing technique used to convert categorical text data into numeric values.[21] This is done by assigning a specific number to each unique category within a feature. This technique is important because most machine learning algorithms cannot process textual data directly. By converting categories into numerical representations, models can interpret and process the information mathematically. Label encoding is especially useful when categorical data is ordinal or has a limited number of categories. The use of this technique improves processing efficiency and allows algorithms to learn from the data more effectively.[22]

In addition, one of the challenges in predictive analysis is the imbalance in the number of samples between classes. Therefore, SMOTE (Synthetic Minority Over-sampling Technique) was applied to balance the number of samples between the "On-Time Graduation" and "Not On-Time Graduation" classes. This technique aims to generate synthetic data to strengthen the minority class, reducing model bias toward the majority class and improving predictive accuracy. Once the data was balanced, the next step was the modeling process.[23]

Model Development

In the modeling and classification phase, several machine learning algorithms were used to build the predictive model. The models tested included Gaussian Naïve Bayes (GNB), Linear Support Vector Machine (SVM), Decision Tree, and K-Nearest Neighbors (KNN). These four algorithms were selected due to their individual strengths in handling various types of data. Subsequently, ensemble learning techniques were applied to improve model performance. Two main ensemble learning approaches used were voting and stacking. In the voting technique, individual models are combined using majority voting to enhance prediction accuracy.[24] In contrast, stacking combines the results of several base models and further analyzes them using Logistic Regression as a meta-learner to produce a more accurate final prediction.[25] Table 2 presents the models used in this study.

Table 2. Model Testing	
N	**Model**
1	Gaussian Naive Bayes
2	Support Vector Machine
3	Decision Tree
4	K-Nearest Neighbours
5	Ensemble With Voting
6	Ensemble With Stacking

https://doi.org/10.56294/dm2025917

7	Ensemble With Voting + GridSearch
8	Ensemble With Stacking + GridSearch

After the model was built, hyperparameter optimization was performed using GridSearchCV. This technique is used to find the best parameter combinations for each algorithm applied, with the goal of improving model performance.[26] Once the optimization process was completed, the models were evaluated and compared to determine the best-performing algorithm for predicting student graduation based on the collected dataset. The final results of this process were used for model comparison, in which the model with the best performance was recommended as the optimal approach for analyzing the factors that influence whether students graduate on time or not.

Model Evaluation

The performance of each predictive model was evaluated using several standard classification metrics, namely accuracy, precision, recall, and F1-score. These metrics were chosen to provide a comprehensive assessment of the model's ability to correctly classify students based on their graduation timeliness:

- Accuracy measures the overall correctness of the model, calculated as the ratio of correctly predicted instances to the total number of predictions.
- Precision evaluates the proportion of true positive predictions among all predicted positives, highlighting the model's ability to avoid false alarms.
- Recall assesses the model's sensitivity by measuring how many actual positive cases were correctly predicted.
- F1-score, as the harmonic mean of precision and recall, provides a balanced view of the model's effectiveness, especially when dealing with class imbalance.

In addition, a confusion matrix was generated to visualize the performance in terms of true positives, true negatives, false positives, and false negatives. This matrix helps to identify the types of classification errors made by the model. Evaluation was conducted on a test dataset comprising 30 % of the total samples, which was separated during the initial data splitting process to ensure objective performance assessment.

All evaluations were conducted using the Scikit-learn library in Python, ensuring standardized computation and reproducibility. These metrics were used to compare the baseline models with ensemble approaches, particularly those optimized using GridSearchCV, to determine the most effective model in predicting graduation timeliness.

RESULTS

The first stage after the data is obtained is to carry out the data labeling process. Figure 2 is a distribution based on labels.

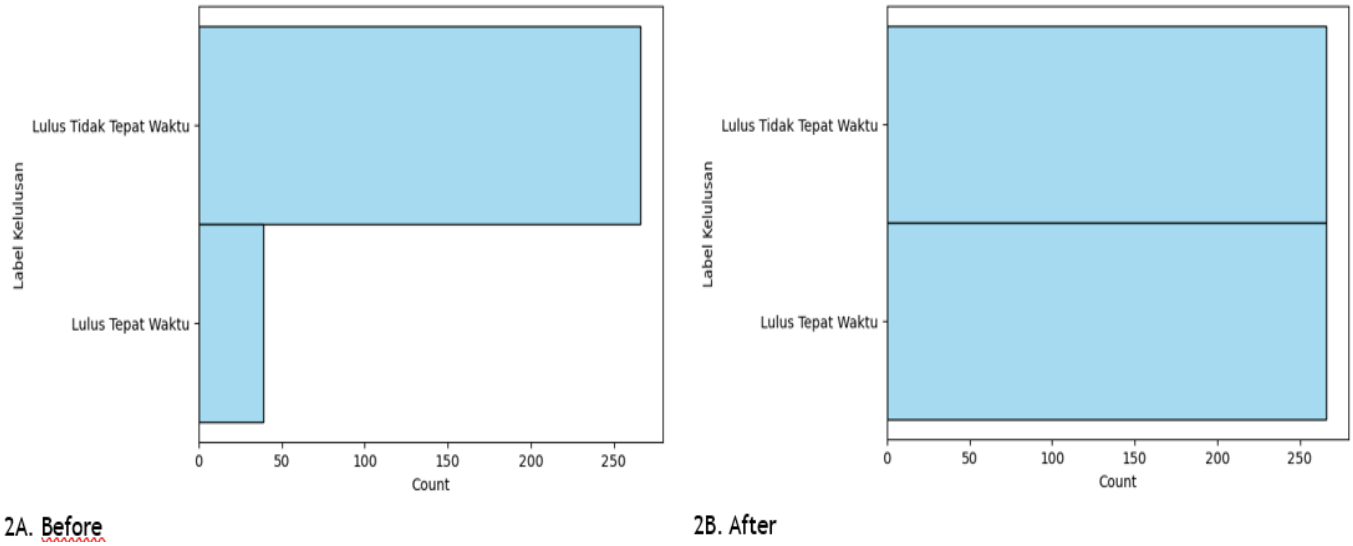


Figure 2. Label Distribution Before and After Oversampling

Based on figure 2, there are two visualizations in the form of horizontal bar charts showing the number of students according to graduation labels: *Lulus Tepat Waktu* (On-Time Graduation) and *Lulus Tidak Tepat Waktu* (Not On-Time Graduation). In chart 2A, it is evident that the majority of students fall into the "Not On-Time Graduation" category, while only a small portion successfully graduate on time. This imbalance indicates that

the dataset has an uneven class distribution, which can cause machine learning models to become biased toward the majority class.

To address this issue, an oversampling technique using SMOTE (Synthetic Minority Over-sampling Technique) was applied. SMOTE works by synthesizing new data points for the minority class based on existing characteristics, thereby creating realistic additional data. The results of this process are shown in chart 2B, where the number of data points in both categories becomes balanced. The application of SMOTE is crucial to ensure that the trained model can learn patterns from both classes proportionally, thereby improving the model's predictive ability and generalization to new data. After the labeling process, the next step is model development. Figure 3 presents the confusion matrix from the ensemble model using the stacking technique with GridSearchCV.

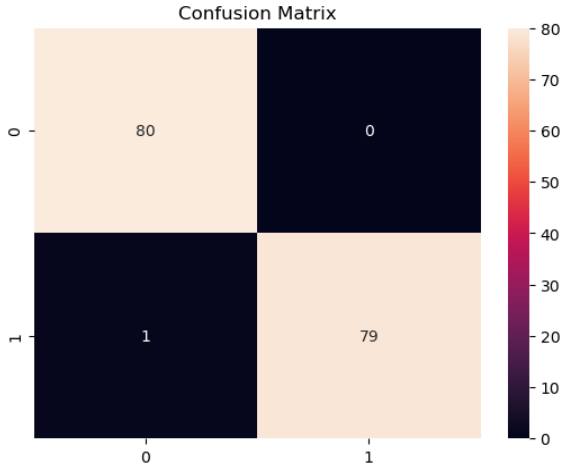


Figure 3. Confusion Matrix

Figure 3 is a visualization of the confusion matrix used to evaluate the performance of the classification model for two classes: "On-Time Graduation" (label 1) and "Not On-Time Graduation" (label 0). The matrix consists of four main components: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN).

Based on the figure, the model successfully classified 80 data points correctly as "Not On-Time Graduation" (TN) and 79 data points correctly as "On-Time Graduation" (TP). There was only one misclassification, where a data point that should have been labeled as "On-Time Graduation" was incorrectly predicted as "Not On-Time Graduation" (FN). No data points were incorrectly predicted as "On-Time Graduation" when they actually belonged to the "Not On-Time Graduation" class (FP = 0).

Overall, these results indicate that the model has very high accuracy, with an extremely low error rate. This suggests that the model is highly effective in distinguishing between the two classes, especially after applying data balancing using the SMOTE technique. The next section, presented in figure 4, is the classification report.

	precision	recall	f1-score	support
Lulus Tepat Waktu	0.99	1.00	0.99	80
Lulus Tidak Tepat Waktu	1.00	0.99	0.99	80
accuracy			0.99	160
macro avg	0.99	0.99	0.99	160
weighted avg	0.99	0.99	0.99	160

Accuracy : 0.99375

Figure 4. Classification Report

Figure 4 displays the performance evaluation results of the classification model using evaluation metrics such as precision, recall, and F1-score for the two classes: Lulus Tepat Waktu (On-Time Graduation) and Lulus Tidak Tepat Waktu (Not On-Time Graduation). The model demonstrated very high performance, with precision and recall values ranging from 0,99 to 1,00 for both classes. This indicates that the model is highly accurate in identifying students who graduate on time as well as those who do not, with a very low error rate.

The F1-score, which also reached 0,99 for both classes, reflects the model's excellent balance between precision and recall. Overall, the model achieved an accuracy of 99,375 %, meaning that out of 160 test data

points, only one was misclassified. The macro average and weighted average values also show high consistency across all metrics, affirming that the model is not biased toward either class. These results indicate that the classification model developed in this study is highly reliable and effective in predicting students' on-time graduation, especially after the dataset was balanced using the SMOTE oversampling technique. The overall model testing results are presented in table 3.

Table 3. Overall Test Results

N	Model	Accuracy	Precision	Recall	F1-Score
1	Gaussian Naive Bayes	95 %	95 %	95 %	95 %
2	Support Vector Machine	96 %	95 %	96 %	95 %
3	Decision Tree	93 %	93 %	93 %	93 %
4	K-Nearest Neighbours	94 %	94 %	94 %	94 %
5	Ensemble With Voting	98 %	98 %	98 %	98 %
6	Ensemble With Stacking	98 %	98 %	98 %	98 %
7	Ensemble With Voting + GridSearch	99 %	99 %	99 %	99 %
8	Ensemble With Stacking + GridSearch	99 %	99 %	99 %	99 %

Table 3 presents the performance evaluation results of various classification models applied to student graduation data, based on metrics such as accuracy, precision, recall, and F1-score. In general, all models demonstrated relatively high performance, with accuracy levels above 90 %. The Gaussian Naive Bayes model achieved an accuracy of 95 %, followed by Support Vector Machine (SVM) with 96 %, while the Decision Tree performed slightly lower at 93 %. The K-Nearest Neighbours (KNN) model recorded an accuracy of 94 %.

Ensemble models showed superior performance. Both the Voting and Stacking ensemble methods achieved an accuracy of 98 %, with consistently high values for precision, recall, and F1-score. Further improvement was obtained through the combination of ensemble learning and hyperparameter tuning using GridSearch, specifically in the Ensemble Voting + GridSearch and Ensemble Stacking + GridSearch models, both of which reached the highest accuracy of 99 % across all evaluation metrics. These results indicate that the ensemble approach, particularly when optimized with GridSearch, provides the most reliable and accurate classification performance in predicting on-time student graduation compared to other models.

DISCUSSION

The findings of this study reveal that the integration of SMOTE, ensemble learning, and GridSearchCV optimization yields a highly accurate predictive model for student graduation timeliness. The stacking ensemble model optimized with GridSearchCV achieved an accuracy of 99,37 %, outperforming all baseline and ensemble-only models. The performance across all evaluation metrics—precision, recall, and F1-score—exceeded 0,99, indicating near-perfect classification.

This high performance is attributable to two key factors. First, the use of ensemble learning, particularly stacking, effectively combined the strengths of several base classifiers (Naïve Bayes, SVM, Decision Tree, and KNN), resulting in a model that generalized well across both majority and minority classes. Second, the application of SMOTE helped address the imbalanced dataset by generating synthetic samples for the underrepresented class (on-time graduates), which significantly improved model sensitivity and fairness.

A comparative analysis with previous studies is presented in table 4, which summarizes the accuracy achieved by different research approaches in predicting student graduation:

Table 4. Comparison with Previous Research

N	Researcher	Model	Accuracy
1	Rahmaddeni, et al.[27]	XGBoost + SVM	79 %
2	Ulfah, et al.[28]	SVM + GridSearchCV	95 %
3	Rahmiati, et al.[29]	Naive Bayes	95 %
4	Herianto, et al.[25]	SMLOS (Stacking Machine Learning Optuna SMOTE)	95 %
5	Van FC, et al.[17]	Stacking + Adaboost + Hyperparameter Tuning + SMOTE	97 %
6	Suandi, et al.[24]	Majority Voting + SMOTE	97 %
7	Putra, et al.[30]	Stacking + SMOTE	88 %
8	Anam, et al.[16]	Voting Hard + SMOTE	94 %
9	This Study	SMOTE + Ensemble With Stacking + GridSearch	99 %

Compared to these studies, our approach achieves the highest accuracy. Notably, while Van FC et al. reached

97 % with a combination of stacking, Adaboost, and SMOTE, our integration of GridSearchCV further refined model performance by tuning the hyperparameters, which likely contributed to the observed performance gain.[17] Similarly, Herianto et al., using SMLOS, reported 95 %, yet lacked the layered optimization pipeline applied in our research.[25]

These results also reaffirm findings in other domains (e.g., Alemerien et al.) where model tuning via GridSearchCV significantly enhanced performance, especially in datasets with complexity and imbalance.[20] The consistency across metrics and near-zero misclassification rate demonstrates the reliability and stability of the proposed model for practical implementation.

From an academic management perspective, this model can serve as a decision-support system for identifying students at risk of delayed graduation, enabling universities to design personalized interventions and advising systems. By integrating such a predictive model into academic dashboards, institutions can shift toward more proactive and data-driven academic policies.

CONCLUSIONS

This study aimed to develop a predictive model for student graduation timeliness by integrating SMOTE, ensemble learning techniques (voting and stacking), and hyperparameter tuning using GridSearchCV. The findings confirm that this integrative approach is highly effective, as evidenced by the stacking ensemble model achieving a predictive accuracy of 99,37 %, with precision, recall, and F1-scores above 0.99.

By addressing class imbalance through SMOTE and refining model performance using GridSearchCV, the study succeeded in enhancing classification quality beyond the capabilities of individual machine learning models. Compared to previous related studies, the proposed model demonstrates superior performance and robustness, proving its potential as a reliable framework for academic risk detection.

Thus, the objective of this study—to produce an accurate, optimized model for predicting on-time graduation—has been achieved. The resulting model can be integrated into academic decision-support systems to assist universities in monitoring students' progress and initiating early interventions for those at risk of delayed graduation.

REFERENCES

1. Bakri R, Astuti NP, Ahmar AS. Machine Learning Algorithms with Parameter Tuning to Predict Students' Graduation-on-time: A Case Study in Higher Education. J Appl Sci Eng Technol Educ. 2022 Dec 30;4(2):259–65.

2. Casanova VS, Pullido ML. Factors Of Graduate Students' Attrition And Retention In Occidental Mindoro State College Graduate School. IJERSC. 3(2):826–31.

3. López-Meneses E, López-Catalán L, Pelícano-Piris N, Mellado-Moreno PC. Artificial Intelligence in Educational Data Mining and Human-in-the-Loop Machine Learning and Machine Teaching: Analysis of Scientific Knowledge. Applied Sciences. 2025 Jan 14;15(2):1–21.

4. Ersozlu Z, Taheri S, Koch I. A review of machine learning methods used for educational data. Educ Inf Technol. 2024 Nov;29(16):22125–45.

5. Taye MM. Understanding of Machine Learning with Deep Learning: Architectures, Workflow, Applications and Future Directions. Computers. 2023 Apr 25;12(5):1–26.

6. Mehta S. Playing Smart with Numbers: Predicting Student Graduation Using the Magic of Naive Bayes. ITALIC. 2023 Nov 23;2(1):60–75.

7. Darenoh NV, Bachtiar FA, Perdana RS. Prediction of On-time Student Graduation with Deep Learning Method: -. J ICT Res Appl. 2024 Jun 27;18(1):1–20.

8. Desfiandi A, Soewito B. Student Graduation Time Prediction Using Logistic Regression, Decision Tree, Support Vector, and Adaboost Ensemble Learning. International Journal of Information System and Computer Science. 7(3):195–9.

9. Haikal MF, Palupi I. Predicting Employability of University Graduates Using Support Vector Machine Classification. Building of Informatics, Technology and Science. 2024;6(2).

10. Rismayati R, Ismarmiaty I, Hidayat S. Esemble Implementation for Predicting Student Graduation withClassification Algorithm. IJECSA. 1(1):35–42.

11. Anam MK, Putra PP, Malik RA, Putra TA, Elva Y, Mahessya RA, et al. Enhancing the Performance of Machine Learning Algorithm for Intent Sentiment Analysis on Village Fund Topic. Journal of Applied Data Sciences. 2025;6(2):1102–15.

12. Sharma H, Pangaonkar S, Gunjan R, Rokade P. Sentimental Analysis of Movie Reviews Using Machine Learning. Shah H, Patel R, Patel N, Buyya R, Chatterjee I, editors. ITM Web Conf. 2023;53:02006.

13. Anam MK, Firdaus MB, Suandi F, Lathifah, Nasution T, Fadly S. Performance Improvement of Machine Learning Algorithm Using Ensemble Method on Text Mining. In: 2024 International Conference on Future Technologies for Smart Society (ICFTSS) [Internet]. Kuala Lumpur, Malaysia: IEEE; 2024 [cited 2025 Mar 16]. p. 90–5. Available from: https://ieeexplore.ieee.org/document/10691363/

14. Putra PP, Anam MK, Defit S, Yunianta A. Enhancing the Decision Tree Algorithm to Improve Performance Across Various Datasets. intensif. 2024 Aug 1;8(2):200–12.

15. Danny M, Muhidin A, Jamal A. Application of the K-Nearest Neighbor Machine Learning Algorithm to Preduct Sales of Best-Selling Products. Brilliance. 2024 Jun 28;4(1):255–64.

16. Anam MK, Lestari TP, Yenni H, Nasution T, Firdaus MB. Enhancement of Machine Learning Algorithm in Fine-grained Sentiment Analysis Using the Ensemble. ECTI-CIT Transactions. 2025 Mar 8;19(2):159–67.

17. Van Fc LL, Anam MK, Bukhori S, Mahamad AK, Saon S, Nyoto RLV. The Development of Stacking Techniques in Machine Learning for Breast Cancer Detection. J Appl Data Sci. 2024 Jan 1;6(1):71–85.

18. Munthe IR, Rambe BH, Hanum F, Amanda AT, Hutagaol ASR, Harianto R. Implementation of Stacking Technique Combining Machine Learning and Deep Learning Algorithms Using SMOTE to Improve Stock Market Prediction Accuracy. J Appl Data Sci. 2024 Dec 1;5(4):2079–91.

19. Anam MK, Van Fc LL, Hamdani H, Rahmaddeni R, Junadhi J, Firdaus MB, et al. Sara Detection on Social Media Using Deep Learning Algorithm Development. JAETS. 2024 Dec 15;6(1):225–37.

20. Alemerien K, Alsarayreh S, Altarawneh E. Diagnosing Cardiovascular Diseases using Optimized Machine Learning Algorithms with GridSearchCV. J Appl Data Sci. 2024 Dec 1;5(4):1539–52.

21. Bolikulov F, Nasimov R, Rashidov A, Akhmedov F, Cho YI. Effective Methods of Categorical Data Encoding for Artificial Intelligence Algorithms. Mathematics. 2024 Aug 18;12(16):1–21.

22. Hien DTT, Thi C, Kim T, The D, Nguyen C. Optimize the Combination of Categorical Variable Encoding and Deep Learning Technique for the Problem of Prediction of Vietnamese Student Academic Performance. IJACSA. 2020;11(11):274–80.

23. Anam MK, Munawir M, Efrizoni L, Fadillah N, Agustin W, Syahputra I, et al. Improved Performance of Hybrid GRU-BiLSTM for Detection Emotion on Twitter Dataset. J Appl Data Sci. 2024 Jan 1;6(1):354–65.

24. Suandi F, Anam MK, Firdaus MB, Fadli S, Lathifah L, Yumami E, et al. Enhancing Sentiment Analysis Performance Using SMOTE and Majority Voting in Machine Learning Algorithms. In: Lumombo L, Rahmi A, Suwarno S, Ardi N, Kurniawan DE, editors. Proceedings of the 7th International Conference on Applied Engineering (ICAE 2024) [Internet]. Dordrecht: Atlantis Press International BV; 2024 [cited 2025 Mar 25]. p. 126–38. (Advances in Engineering Research; vol. 251). Available from: https://www.atlantis-press.com/doi/10.2991/978-94-6463-620-8_10

25. Herianto H, Kurniawan B, Hartomi ZH, Irawan Y, Anam MK. Machine Learning Algorithm Optimization using Stacking Technique for Graduation Prediction. J Appl Data Sci. 2024 Sep 1;5(3):1272–85.

26. Jumanto J, Rofik R, Sugiharti E, Alamsyah A, Arifudin R, Prasetiyo B, et al. Optimizing Support Vector Machine Performance for Parkinson's Disease Diagnosis Using GridSearchCV and PCA-Based Feature Extraction. J Inf Syst Eng Bus Intell. 2024 Feb 28;10(1):38–50.

27. Rahmaddeni R, Anam MK, Irawan Y, Susanti S, Jamaris M. Comparison of Support Vector Machine and

XGBSVM in Analyzing Public Opinion on Covid-19 Vaccination. Ilk J Ilm. 2022 Apr 30;14(1):32–8.

28. Ulfah AN, Anam MK, Sidratul Munti NY, Yaakub S, Firdaus MB. Sentiment Analysis of the Convict Assimilation Program on Handling Covid-19. JUITA. 2022 Nov 14;10(2):209–16.

29. Rahmiati R, Anam MK, Paradila D, Mardainis M, Machdalena M. Application of Naïve Bayes Algorithm for Non-Cash Food Assistance Recipients in Kampar Regency. SinkrOn. 2023 Jan 4;8(1):433–41.

30. Putra PP, Anam MK, Chan AS, Hadi A, Hendri N, Masnur A. Optimizing Sentiment Analysis on Imbalanced Hotel Review Data Using SMOTE and Ensemble Machine Learning Techniques. J Appl Data Sci. 2025 May 1;6(2):921–35.

FINANCING

The authors would like to express their sincere gratitude to Universitas Islam Riau (UIR) for the financial support provided for this research. This support was essential for the smooth execution and successful completion of the study.

CONFLICT OF INTEREST

None

AUTHORSHIP CONTRIBUTION

Conceptualization: Akmar Efendi.
Data curation: Akmar Efendi.
Formal analysis: Akmar Efendi.
Research: Akmar Efendi, Gunadi Widi Nurcahyo, and Iskandar Fitri.
Methodology: Akmar Efendi, Gunadi Widi Nurcahyo, and Iskandar Fitri.
Project management: Akmar Efendi.
Resources: Akmar Efendi.
Software: Akmar Efendi.
Supervision: Iskandar Fitri.
Validation: Iskandar Fitri.
Display: Iskandar Fitri.
Drafting - original draft: Akmar Efendi.
Writing - proofreading and editing: Gunadi Widi Nurcahyo.