Machine Learning Based Early Warning Model for Delayed Student Graduation: A Sixth-Semester Prediction Approach

Benny Daniawan[1*], Suwitno[2], Andri Wijaya[3], Ardiane Rossi Kurniawan Maranto[4], Junaedi[5]

[1,2,3,4,5]*Faculty of Science and Technology, Universitas Buddhi Dharma, Tangerang, Indonesia*

*corr-author: b3n2y.miracle@gmail.com

Abstract - Delayed student graduation is a critical issue in higher education because it affects academic planning, student support, and institutional performance evaluation. This study develops a leakage controlled machine learning framework for early identification of students at risk of delayed graduation, using academic records available through the sixth semester. A dataset of 564 students was used, with graduation status defined as on-time for students graduating in the eighth semester or earlier and delayed for those graduating after the eighth semester. To prevent temporal data leakage, post-outcome variables were excluded from the predictor set. Five supervised learning algorithms were evaluated: Decision Tree, Support Vector Machine, Random Forest, Naïve Bayes, and K-Nearest Neighbor. Preprocessing was performed using one-hot encoding and standardization within a pipeline, and model performance was assessed using stratified five-fold cross-validation. The tuned Random Forest achieved the most balanced performance, with 0.956 accuracy, 0.861 delayed-class precision, 0.805 delayed-class recall, 0.832 delayed-class F1-score, and 0.977 ROC-AUC. The tuned SVM with a threshold of 0.30 achieved higher delayed-class recall (0.857) and ROC-AUC (0.980). Feature-importance analysis indicated that fourth and fifth semester GPAs were the strongest predictors. These findings show that machine learning can support early academic intervention and data driven decision making in higher education.

Keywords: delayed graduation, education data mining, stratified five-fold cross-validation, tuned random forest, tuned support vector.

I. INTRODUCTION

In education, categorizing student graduation is critical for decision making and program design. Many educational institutions have high student dropout rates due to factors such as demographic and personal characteristics, including age, gender, academic background, financial capacity, and degree options [1]. By developing graduation predictions, educational institutions can identify students at risk of not graduating on time and take preventive actions to support their academic progress [2, 3]. One approach to predicting graduation is to use Machine Learning techniques.

Machine Learning is a method of developing artificial intelligence that allows computer systems to learn from data and generate predictions or actions without being explicitly programmed [4, 5]. In predicting student graduation, Machine Learning can analyze patterns in historical data from students who graduated or did not graduate on time to predict the graduation of active students [6-8]. This approach enables large scale processing of academic data and the identification of complex relationships among attributes that are difficult to analyze conventionally.

Research by Alhakami demonstrates the effectiveness of machine learning in analyzing student data patterns and predicting graduation probabilities [9]. Other studies were also conducted to predict student graduation using Machine Learning with several techniques such as Logistic Regression (LR), Naive Bayes (NB), Neural Network (NN), Support Vector Machine (SVM), Random Forest (RF), Artificial Neural Network (ANN), C4.5, and so on. These techniques allow the system to build prediction models from patterns identified in the training data and then apply them to new data to predict student graduation [10-13].

The reason for using Machine Learning to predict student graduation is its ability to process complex data and uncover patterns invisible to humans. Using Machine Learning algorithms, the system may identify factors that influence student graduation, such as academic grades, participation in learning activities, and sociodemographic characteristics [14, 15]. This technique is theoretically consistent with educational data mining, which analyzes historical data to support evidence based decision making.

Several studies have been conducted in predicting student drop out. One study conducted by Osorio and Santacoloma created an early warning system to identify first-semester students at high risk of dropping out. This

was based on a machine learning model trained on historical data of first-semester students. The results predicted "at-risk" students with a sensitivity of 61.97% [16]. Two machine learning approaches, namely Logistic Regression and Decision Tree, were used to predict student dropouts at the Karlsruhe Institute of Technology (KIT), where Decision Tree performed better, achieving an accuracy of 95% after 3 semesters [17]. The Decision Tree method has been used in various recommendation systems in the educational domain, such as major selection recommendations and academic decision making, due to its ability to generate easy-to-interpret decision rules [18]. Research on student academic performance prediction shows that decision tree based algorithms perform well in educational classification tasks. Previous studies comparing J48 and Random Forest have shown that ensemble methods such as Random Forest tend to achieve higher predictive performance than single decision trees, especially on complex academic data [19].

Previous research has shown that several machine learning methods are helpful in predicting student graduation and academic success. However, most research still focuses on predicting college dropout or general student performance without conducting a thorough comparative analysis of several algorithms in the context of late graduation, and few have examined differences in dominant factors across groups or faculties. In addition, studies that specifically identify key academic attributes that influence untimely graduation are limited.

As a result, this study contributes a leakage controlled and early warning framework for identifying delayed student graduation using academic records available up to the sixth semester. The study evaluates whether academic predictors available before graduation can produce reliable predictions [17, s20]. Then this study can fills a gap by comparing multiple Machine Learning algorithms to detect individuals experiencing graduation delays and determining the most relevant academic elements using historical university data. The goal of this study is to improve academic efficiency at one of Tangerang's XYZ universities by identifying the key characteristics that contribute to on-time graduation, enabling the institution to better manage resources by focusing on supporting and coaching students who may not graduate on time. This study examines the effectiveness of several algorithms, including Decision Tree, Support Vector Machine, Random Forest, Naïve Bayes, and K-Nearest Neighbor, in assessing students' graduation status and identifying academic issues that delay graduation.

II. METHOD

This section describes the methodological procedures used to develop and evaluate a machine learning based early warning model to identify students at risk of delayed graduation. The study adopted the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework, consisting of business understanding, data understanding, data preparation, modeling, evaluation, and knowledge deployment. Fig. 1 depicts the entire flow of this research as a framework that shows the interaction between the stages of the data analysis process, ending in the categorization and interpretation of the most important elements influencing student graduation. This technique tries to construct an accurate prediction model while also giving a full knowledge of the academic qualities that influence graduation delays.

A. Business Understanding

This research starts from Business Understanding, which is the stage where problems that occur in business processes are defined simply and precisely, at this stage requires an understanding of business processes including regulations that regulate how to manage [21, 22]. The data for this study were gathered at the University and focused on students who graduated on time and those who graduated late. The purpose is to discover the causal causes and determine the potential of students in the following semester who are unlikely to graduate on time.

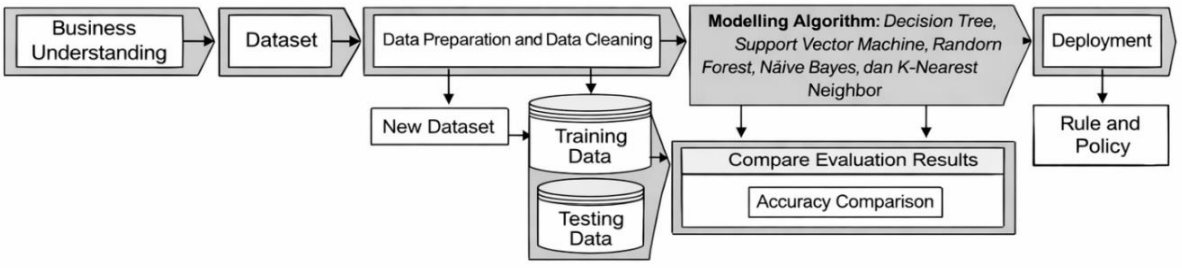

Fig. 1 Proposed thinking framework model

B. Data Understanding

The next stage of Data Understanding is a process that brings together what data is owned and what data is needed. The dataset is studied to identify it, and the type of data will greatly determine the type of algorithm used for machine learning [21, 22]. At this stage, the data obtained comprised 564 students from three faculties: the Faculty of Business. The Faculty of Social Affairs and Humanities, and the Faculty of Science and Technology. The attributes used are shown in Table I. Attributes such as Nim are removed, the seventh through fourteenth semester achievement index is removed, and replaced with a on time or delayed label. Civil status and employment status were also removed because the data showed that all students were unmarried and not employed. Gender data and Graduation_Status labels that were originally in text form were converted to numerical values using label encoders, as machine learning methods cannot process raw text [23]. This is very important because raw data should be encoded into numerical vectors [24].

C. Data Preparation and Data Cleaning

The next stage of Data Preparation is the process of data treatment towards a useful quality model. A good, accurate model starts with well prepared data. Some of the things commonly done at this stage are to double check the correctness of the data, manage data outliers, and address missing data and data inconsistencies [21, 22]. In the dataset to be used, data validation is used to eliminate incomplete records, such as students who register only in the first semester but do not continue their studies in the next semester, the data will be deleted to ensure the accuracy of processing. Categorical variables, namely major and gender, were transformed using one-hot encoding. Numerical variables, including age, semester GPA from the first to sixth semester, and zero GPA count, were standardized using StandardScaler. To avoid preprocessing leakage, encoding and scaling were embedded in a machine learning pipeline and applied only to the training folds during cross-validation. Meanwhile, for students who have reached sixth semester but do not have a semester achievement index, the missing value will remain zero because the student is active but does not complete academics at that time.

Zero values in semester GPA variables were not treated as ordinary missing values because they represented academically meaningful conditions, such as inactive study status, leave, or failure to complete academic activities in a given semester. Therefore, these zero values were retained as valid observations. To capture the cumulative effect of academic inactivity, an additional feature was constructed by counting the number of semesters with zero GPA within each prediction horizon. To ensure temporal validity and prevent data leakage, this study defined the prediction point at the end of the sixth semester. This prediction horizon was selected because the earliest graduation in the dataset occurred in the seventh semester therefore, academic information up to the sixth semester was available before any graduation outcome was observed. The target variable was constructed from the graduation semester, where students who graduated in the eighth semester or earlier were classified as on-time, while students who graduated after the eighth semester were classified as delayed. After constructing the target label, post-outcome variables such as graduation semester, final GPA, total credits at graduation, and semester GPA after the sixth semester were excluded from the predictor set.

TABLE I
RESEARCH DATA ATTRIBUTES

No	Attributes	Data Type	Information
1	Major	Nominal	Student major
2	Gender	Binary	Male or Female
3	Age	Numerical	Student's Age
4	IPS 1	Numerical	Semester 1 GPA
5	IPS 2	Numerical	Semester 2 GPA
6	IPS 3	Numerical	Semester 3 GPA
7	IPS 4	Numerical	Semester 4 GPA
8	IPS 5	Numerical	Semester 5 GPA
9	IPS 6	Numerical	Semester 6 GPA
10	Zero_IPS_Count_s6	Numerical	Zero GPA count up to Semester 6
11	Graduation_Status	Binary	0 = on time 1 = delayed

D. Model Development

The modeling stage was conducted after the dataset had been validated and confirmed to contain no missing or inconsistent values. At this phase, the classification process was performed using five supervised learning algorithms: Decision Tree (DT), which constructs hierarchical decision rules by recursively partitioning the dataset based on selected attributes and is valued for its interpretability and ability to handle both categorical and numerical variables [25]. Support Vector Machine (SVM), which identifies an optimal separating hyperplane in high dimensional space and can model nonlinear relationships through kernel functions while maintaining robustness to outliers [26, 27]. Random Forest (RF) is an ensemble based technique that combines multiple decision trees built on randomized subsets of the data to reduce variance and improve generalization performance [28, 29]. Naïve Bayes (NB), a probabilistic classifier grounded in Bayes' theorem that estimates posterior class probabilities under conditional independence assumptions and is computationally efficient for large scale datasets [26, 30], and K-Nearest Neighbor (K-NN), a distance-based method that assigns class labels according to similarity measures by considering the closest training instances within the feature space [26]. Each algorithm was evaluated using stratified 5-fold cross-validation, which was used to preserve the proportions of on-time and delayed students in each fold. The folds were shuffled using a fixed random seed to ensure reproducibility. This validation strategy was selected because the dataset had an imbalanced class distribution, with fewer delayed students than on-time students, as shown in Table II.
In order for model performance to increase significantly, it is necessary to tune the hyperparameters correctly [31]. This research uses hyperparameter tuning, which is shown in Table III.

In addition to hyperparameter tuning, probability-threshold analysis was also performed for the five tuned algorithms. This analysis examined thresholds from 0.10 to 0.90 to evaluate the trade-off between precision and recall for the delayed graduation class. The threshold was selected by considering the F1-score of the delayed class and the practical need to maintain high recall for early warning intervention.

E. Model Evaluation

Model performance was evaluated using accuracy, precision, recall, F1-score, confusion matrix, and ROC-AUC [32]. Because the main objective was to identify students at risk of delayed graduation, the delayed class was treated as the positive class. Therefore, recall and F1-score for the delayed class were emphasized in model selection.

III. RESULT AND DISCUSSION

A. Model Performance Evaluation

This section presents the analytical results obtained from the application of several machine learning techniques, namely Decision Tree, Support Vector Machine, Random Forest, Naïve Bayes, and K-Nearest Neighbors. The data analysis was conducted using Google Colab as the computational environment. The results of a comparison of the five algorithms using 5-Fold cross validation are shown in Table IV.

Based on the performance evaluation shown in Table IV, the validation results indicate that SVM achieved the most balanced performance for identifying students at risk of delayed graduation. SVM achieved an accuracy of 0.938, a recall of 0.882, an F1-score of 0.798, and an ROC-AUC of 0.980. Although Random Forest achieved the highest accuracy 0.950 and ROC-AUC 0.982, its lower recall 0.714 indicates that it missed more students who were actually late than SVM. Furthermore, the KNN and Decision Tree models showed moderate performance but indicated symptoms of overfitting. KNN achieved an accuracy of 0.943, a recall of 0.674, an F1-Score of 0.761, and an ROC-AUC of 0.925. Decision Tree experienced a significant performance decline on the test data with an F1-score of 0.734, indicating that the model was too sensitive to variations in the training data.

TABLE II
CLASS DISTRIBUTION

Class	Definition	Frequency
On Time	Graduation Semester ≤8	487
Delayed	Graduation Semester >8	77

TABLE III
HYPERPARAMETER TUNING

Model	Final Hyperparameters
SVM	kernel = RBF, C = 1, gamma = 0.05, class_weight = balanced
Random Forest	n_estimators = 100, max_depth = 6, max_features = sqrt, min_samples_split = 5, min_samples_leaf = 1, class_weight = balanced_subsample
Decision Tree	random_state = 42, class_weight = balanced
KNN	n_neighbors = 5
Naïve Bayes	GaussianNB

TABLE IV
EVALUATION RESULT USING 5-FOLD CROSS VALIDATION

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
SVM	0.938	0.733	0.882	0.798	0.980
Random Forest	0.950	0.912	0.714	0.795	0.982
KNN	0.943	0.878	0.674	0.761	0.925
Decision Tree	0.929	0.764	0.715	0.734	0.839
Naive Bayes	0.229	0.145	0.949	0.251	0.907

Meanwhile, Naive Bayes produced the lowest accuracy 0.229, despite a very high recall value 0.949. This phenomenon in Naive Bayes represents a skewed prediction condition, where the model experiences bias and classifies almost all samples into the late student class. This occurs because the assumption of feature independence in Naive Bayes is violated by the characteristics of academic data (such as Semester GPA), which exhibit strong inter semester correlation. The next step is to generate a confusion matrix and classification report, which are presented in Fig. 2 and Table V. This comparison presents the results of hypertuning the best parameters using threshold testing for the SVM and Random Forest methods, which performed best.

Fig. 2 presents a visualization of the confusion matrix of the two best optimized models: Tuned Random Forest (A) and Tuned SVM (B). This matrix visualizes the distribution of object predictions by actual class (True Label) and predicted class (Predicted Label) to determine the specific error rate for each algorithm.

In Fig. 2(A), the Tuned Random Forest model demonstrates very tight performance on the majority class (on-time), generating only 10 false positive errors (on-time students predicted to be delayed). Furthermore, this model failed to identify 15 students who were actually at risk of graduating late or delayed (False Negatives), successfully capturing only 62 of 77 at risk students. Conversely, Figure 2(B) confirms the impact of lowering the decision threshold to 0.30 in the Tuned SVM model. This model proved significantly more sensitive in detecting academic risk, reducing the number of false negatives to only 9 students. This means that, out of 77 students who experienced actual delays in graduation, the Tuned SVM correctly identified 68 of them. However, this high sensitivity resulted in a spike in the False Positive rate to 24 students. In a real world scenario, this means that 24 students were predicted to be late when they actually graduated on time.

The shift in absolute values in these two confusion matrices indicates a clear trade-off between minimizing false alarms in Random Forest and maximizing early detection coverage in SVM. To see how this absolute value distribution affects the comparative standardization of algorithm performance scores, a detailed metric evaluation is presented in Table V.

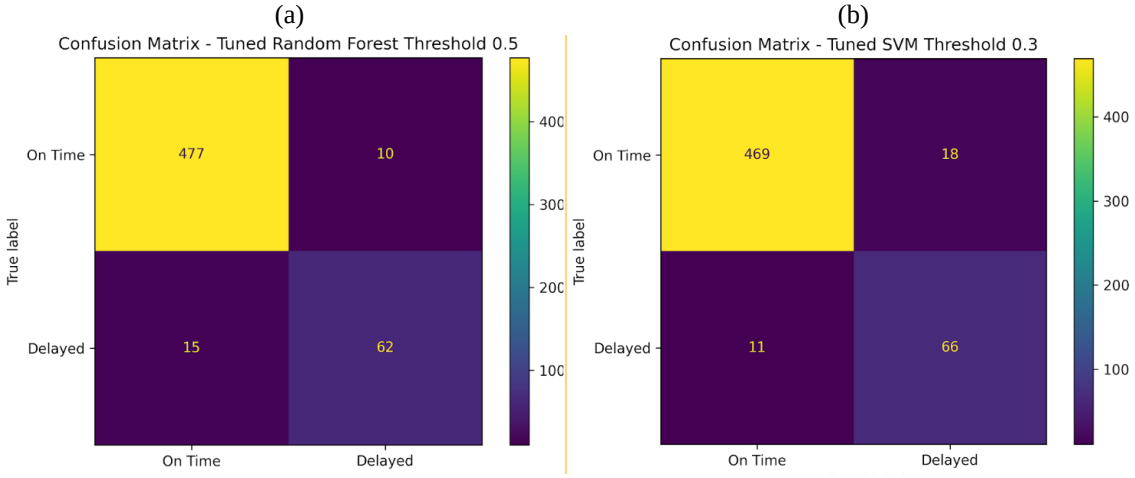

Fig. 2 Confusion matrix (a) tuned random forest, (b) tuned SVM

TABLE V
RANDOM FOREST AND SVM PERFORMANCE

Model	Threshold	Accuracy	Precision Delayed	Recall Delayed	F1 Delayed	ROC-AUC
Tuned Random Forest	0.500	0.956	0.861	0.805	0.832	0.977
Tuned SVM	0.300	0.949	0.786	0.857	0.820	0.980

Based on the results of hyperparameter optimization and classification threshold adjustments, summarized in Table V, two models with different performance characteristics were selected for implementation in the early warning system. The Tuned Random Forest model with a standard threshold of 0.50 demonstrated the best global classification performance. This model achieved the highest accuracy of 0.956 and an F1-score of 0.832 for the delayed class. Furthermore, the Tuned Random Forest's main advantage lies in its high delayed precision value of 0.861. This indicates that the model has a low error rate in predicting late students, thereby minimizing the risk of misclassifying students who actually graduate on time as late (false positives).

On the other hand, the Tuned SVM with a threshold of 0.30 offers more operationally sensitive performance. Using this threshold successfully boosted the recall value to 0.857, higher than the Tuned Random Forest's 0.805. Regarding academic intervention policies, the high recall value indicates that the Tuned SVM is more accurate at identifying students who are truly at risk of a graduation delay. However, this increased sensitivity results in a Precision Delay of 0.786, indicating the system will produce a slightly higher false alarm rate than the Random Forest model.

By comparing these two tuned models, this study provides implementation flexibility: the Tuned Random Forest model is preferred if efficient prediction accuracy is a top priority. In contrast, the Tuned SVM is recommended if the institution prioritizes minimizing the risk of undetected at risk students (False Negatives).

A comparative evaluation of the models was conducted through Receiver Operating Characteristic (ROC) curve analysis, as presented in Fig.3. The ROC curve visualizes the dynamic trade-off between the True Positive Rate (Sensitivity) on the Y-axis and the False Positive Rate (Specificity) on the X-axis across the entire spectrum of decision thresholds (classification thresholds). Based on the graph plot in Figure 3, both curves from the Tuned Random Forest and Tuned SVM models exhibit ideal curvature characteristics, as they move toward the upper left corner of the graph. This visual phenomenon indicates that both models have a strong ability to distinguish between students with the potential to graduate on time and those at risk of being delayed.

Quantitatively, the Area Under the Curve (AUC) values for both algorithms were above 0.970, with the Tuned SVM slightly outperforming, with an AUC of 0.980, compared to the Tuned Random Forest model's 0.977. The marginal advantage of the Tuned SVM indicates that, in terms of conditional probability, the SVM has slightly greater consistency in decision space stability when the threshold parameter is shifted across alternative values (in this study, it was optimized at a threshold of 0.30). The high AUC values in both models confirm that the hyperparameter optimization (tuning) process has successfully eliminated classification bias, making the model highly reliable for application as an early warning academic decision support system.

The findings of this study make important contributions to the development of decision support systems at the university level. This predictive model was built using academic information available up to the sixth semester. This aligns with research by Berens et al. [20], which shows that early prediction of student dropout using demographic and academic data from first to fourth semesters can reveal progressive semester-over-semester dynamics and significantly improve prediction accuracy. Furthermore, research by Amjed et al. [33], shows that semester grades are among the most influential factors in determining student performance. The use of historical data attributes prior to the graduation period in this study successfully mitigated the risk of temporal data leakage. This approach mirrors a real life operational early warning system, where academic stakeholders can identify at risk students before the expected graduation period.

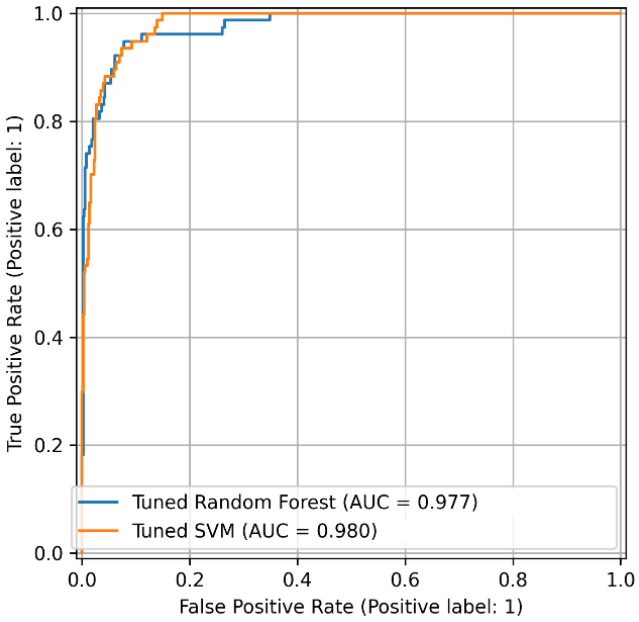

Fig. 3 ROC curve of tuned models

The final evaluation results showed that the Tuned Random Forest model achieved the most balanced performance. These characteristics demonstrate that Random Forest is highly suitable for implementation when institutions aim to balance prediction reliability and intervention efficiency. On the other hand, the SVM model produced more false positives than Random Forest, highlighting the practical trade-off between precision and recall. The SVM model can be an alternative if the institution prioritizes maximum detection coverage for students at risk of tardiness. At the same time, Random Forest is more appropriate if the institution prioritizes balanced prediction results with a minimal false alarm rate for management efficiency.

B. Feature Importance and Academic Interpretation

In addition to evaluating prediction accuracy, this study conducted a predictor variable analysis to identify the factors most strongly influencing the risk of delayed graduation. Feature contributions were calculated using the Mean Decrease in Impurity (Gini Importance) metric from the Tuned Random Forest, the algorithm with the best performance. The results of extracting the 5 most influential features are presented in Fig. 4.

Fig. 4 shows a significant phenomenon where the fourth semester (ips4) and fifth Semester (ips5) Achievement Index variables dominate the graph, with the highest contribution scores. This finding indicates that students' academic performance midway through their studies (late second year to early third year) is the most critical indicator in determining on-time graduation. Academically, this phase is generally filled with advanced courses with a high theoretical and practical weight, as well as the initial planning of their final project research topic. Failure to maintain stable IPS grades during this crucial phase directly affects the accumulated credit load that must be repeated, thereby increasing the likelihood that students will be categorized as delayed.

Interestingly, age ranked third as the most dominant non academic predictor, surpassing academic record in the early semesters (ips3, ips2, ips6, and ips1). This represents a correlation between students' maturity or personal readiness during their studies and their consistent completion of their academic workload. Conversely, demographic features such as gender and study program ranked lowest, with very marginal contributions. The low contribution values of these features provide important justification for the fairness of the prediction model and the absence of gender or study program bias. The risk of delayed graduation is driven more by students' actual academic performance than by demographic factors.

Based on the feature importance analysis, the fourth and fifth semester GPA variables were the most influential predictors of delayed students. This finding confirms that the middle phase of study is a critical juncture for identifying at risk students. Although data were used through the sixth semester, performance in fourth and fifth served as the decision making anchor because this phase represents a crucial transition period toward deepening the subject matter of the skill set and preparing for the final assignment. Age also contributed significantly to the prediction, although its influence was lower than that of the semester GPA development variable. Meanwhile, gender and major variables showed relatively marginal contributions. This phenomenon indicates that academic development variables are a far more important determinant of graduation risk than demographic background or student study program factors.

The results of this study provide data driven insights for proactive academic decision making. Students who consistently show a declining trend in semester GPA through the sixth semester can be prioritized for intensive academic advising and administrative support. This proposed model can serve as a decision support tool to help academic advisors identify students who need a qualitative assessment and timely intervention.

This study has limitations due to its reliance on structured academic records sourced from a single institution. Therefore, future research needs to validate this model using multi institutional datasets and integrate additional, more dynamic variables, such as attendance rates, financial constraints, activity on the Learning Management System (LMS), and academic advising track records. Furthermore, integrating explainable artificial intelligence (XAI) methods, such as SHAP or LIME, is highly recommended for future research to provide more transparent, accountable, and actionable intervention recommendations for university stakeholders.

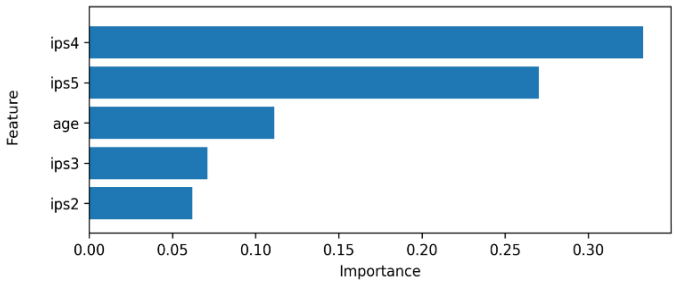

Fig. 4 Top 5 feature importance tuned random forest

IV. CONCLUSION

This study proposed a machine learning based prediction framework to identify students at risk of delayed graduation. By using only predictors available before the graduation outcome and excluding post outcome attributes, the model reduced the risk of temporal data leakage and provided a more realistic early warning design. The tuned Random Forest model achieved the most balanced performance, with an accuracy of 0.956, a delayed class F1-score of 0.832, and an ROC-AUC of 0.977. The tuned SVM model with a threshold of 0.30 achieved higher delayed class recall, making it a relevant alternative when early detection is prioritized. Feature importance analysis showed that fourth and fifth semester GPAs were the dominant predictors, indicating that the middle phase of study is critical for monitoring graduation risk.

Machine learning can assist academic stakeholders in prioritizing students for advising and intervention. Future research that integrates explainable artificial intelligence (XAI) methods, such as SHAP or LIME, is highly recommended to provide more transparent, accountable, and actionable intervention recommendations for university stakeholders.

ACKNOWLEDGEMENT

Thank you to the Institute for Publication, Research, and Community Service (LPPPM) at Buddhi Dharma University for the funding and opportunity provided.

REFERENCES

[1] S. Shilbayeh and A. Abonamah, "Predicting student enrolments and attrition patterns in higher educational institutions using machine learning," *Int. Arab J. Inf. Technol.*, vol. 18, no. 4, pp. 562–567, 2021, doi: 10.34028/18/4/8.

[2] K. Fahd, Kiran; Miah, Shah Jahan; Ahmed, "Predicting student performance in a blended learning environment using learning management system interaction data," *Appl. Comput. Informatics*, 2021, doi: 10.1108/ACI-06-2021-0150.

[3] C. M. Profiroiu, P. S. Angheluță, P. C. Vasilache, and C. Dima, "The Level of Education of Graduates in Romania in the Context of Globalization," *SHS Web Conf.*, vol. 92, p. 07055, 2021, doi: 10.1051/shsconf/20219207055.

[4] I. H. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions," *SN Comput. Sci.*, vol. 2, no. 3, pp. 1–21, 2021, doi: 10.1007/s42979-021-00592-x.

[5] J. P. Bharadiya, "The role of machine learning in transforming business intelligence," *Int. J. Comput. Artif. Intell.*, vol. 4, no. 1, pp. 16–24, 2023, doi: 10.33545/27076571.2023.v4.i1a.60.

[6] N. A. M. Samsudin, S. M. Shaharudin, N. A. F. Sulaiman, S. Ismail, N. S. Mohamed, and N. H. M. Husin, "Prediction of Student's Academic Performance during Online Learning Based on Regression in Support Vector Machine," *Int. J. Inf. Educ. Technol.*, vol. 12, no. 12, pp. 1431–1435, 2022, doi: 10.18178/ijiet.2022.12.12.1768.

[7] T. Handhayani and L. Hiryanto, "Predicting and Analyzing the Students' Length of Study-Time Using Support Vector Machine," vol. 8, no. 2, pp. 107–114, 2017, doi: 10.21512/comtech.v8i2.3756.

[8] M. R. P. Putra and E. Utami, "A Blending Ensemble Approach to Predicting Student Dropout in Massive Open Online Courses (MOOCs)," *JUITA J. Inform.*, vol. 13, no. 1, pp. 11–18, 2025, doi: 10.30595/juita.v13i1.24061.

[9] H. Alhakami, T. Alsubait, and A. Aljarallah, "Data mining for student advising," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 3, pp. 526–532, 2020, doi: 10.14569/ijacsa.2020.0110367.

[10] N. S. Sani, A. F. M. Nafuri, Z. A. Othman, M. Z. A. Nazri, and K. N. Mohamad, "Drop-Out Prediction in Higher Education Among B40 Students," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 11, pp. 550–559, 2020, doi: 10.14569/IJACSA.2020.0111169.

[11] A. S. Hoffait and M. Schyns, "Early detection of university students with potential difficulties," *Decis. Support Syst.*, vol. 101, pp. 1–11, 2017, doi: 10.1016/j.dss.2017.05.003.

[12] A. Ritesh, A. Jadhav, and S. Dukhan, "Forecasting Learner Attrition for Student Success at a South African University," *ACM Int. Conf. Proceeding Ser.*, pp. 19–28, 2020, doi: 10.1145/3410886.3410973.

[13] A. F. Gkontzis, "A predictive analytics framework as a countermeasure for attrition of students," *Interact. Learn. Environ.*, vol. 30, no. 6, pp. 1028–1043, 2022, doi: 10.1080/10494820.2019.1709209.

[14] S. Huang and J. Wei, "Student Performance Prediction in Mathematics Course Based on the Random Forest and Simulated Annealing," *Sci. Program.*, vol. 2022, 2022, doi: 10.1155/2022/9340434.

[15] J. Alvarado-Uribe, "Student Dataset from Tecnologico de Monterrey in Mexico to Predict Dropout in Higher Education," *Data*, vol. 7, no. 9, 2022, doi: 10.3390/data7090119.

[16] J. K. Hoyos and G. Daza, "Predictive Model to Identify College Students with High Dropout Rates," *Rev. Electron. Investig. Educ.*, vol. 25, no. 13, pp. 1–10, 2023, doi: 10.24320/redie.2023.25.e13.5398.

[17] L. Kemper, G. Vorhoff, and B. U. Wigger, "Predicting student dropout: A machine learning approach," *Eur. J. High. Educ.*, vol. 10, no. 1, pp. 28–47, 2020, doi: 10.1080/21568235.2020.1718520.

[18] A. R. Anwari and S. Sukirman, "Recommendation

system to Select a Major of Vocational School Using Decision Tree," *J. Tek. Inform.*, vol. 5, no. 2, pp. 589–598, Apr. 2024, doi: 10.52436/1.jutif.2024.5.2.1327.

[19] M. Kumar, V. Bhardwaj, D. Thakral, A. Rashid, and M. T. Ben Othman, "Ensemble Learning Based Model for Student's Academic Performance Prediction Using Algorithms," *Ing. des Syst. d'Information*, vol. 29, no. 5, pp. 1925–1935, 2024, doi: 10.18280/isi.290524.

[20] J. Berens, K. Schneider, S. Gortz, S. Oster, and J. Burgoff, "Early Detection of Students at Risk - Predicting Student Dropouts Using Administrative Student Data from German Universities and Machine Learning Methods," *J. Educ. Data Min.*, vol. 11, no. 3, pp. 1–41, 2019, doi: 10.5281/zenodo.3594771.

[21] V. Plotnikova, M. Dumas, and F. P. Milani, "Applying the CRISP-DM data mining process in the financial services industry: Elicitation of adaptation requirements," *Data Knowl. Eng.*, vol. 139, p. 102013, 2022, doi: https://doi.org/10.1016/j.datak.2022.102013.

[22] C. Schröer, F. Kruse, and J. M. Gómez, "A systematic literature review on applying CRISP-DM process model," *Procedia Comput. Sci.*, vol. 181, no. 2019, pp. 526–534, 2021, doi: 10.1016/j.procs.2021.01.199.

[23] M. K. Dahouda and I. Joe, "A Deep-Learned Embedding Technique for Categorical Features Encoding," *IEEE Access*, vol. 9, pp. 114381–114391, 2021, doi: 10.1109/ACCESS.2021.3104357.

[24] T. Jo, *Machine Learning Foundations: Supervised, Unsupervised, and Advanced Learning*, 1st ed. Springer Cham, 2021. doi: 10.1007/978-3-030-65900-4.

[25] V. Matzavela and E. Alepis, "Decision tree learning through a Predictive Model for Student Academic Performance in Intelligent M-Learning environments," *Comput. Educ. Artif. Intell.*, vol. 2, p. 100035, 2021, doi: 10.1016/j.caeai.2021.100035.

[26] F. Agrusti, G. Bonavolontà, and M. Mezzini, "University dropout prediction through educational data mining techniques: A systematic review," *J. E-Learning Knowl. Soc.*, vol. 15, no. 3, pp. 161–182, 2019, doi: 10.20368/1971-8829/1135017.

[27] D. A. Otchere, T. O. Arbi Ganat, R. Gholami, and S. Ridha, "Application of supervised machine learning paradigms in the prediction of petroleum reservoir properties: Comparative analysis of ANN and SVM models," *J. Pet. Sci. Eng.*, vol. 200, no. November 2020, p. 108182, 2021, doi: 10.1016/j.petrol.2020.108182.

[28] M. Nachouki, E. A. Mohamed, R. Mehdi, and M. Abou Naaj, "Student course grade prediction using the random forest algorithm: Analysis of predictors' importance," *Trends Neurosci. Educ.*, vol. 33, p. 100214, 2023, doi: https://doi.org/10.1016/j.tine.2023.100214.

[29] T. T. Huynh-Cam, L. S. Chen, and H. Le, "Using decision trees and random forest algorithms to predict and determine factors contributing to first-year university students' learning performance," *Algorithms*, vol. 14, no. 11, 2021, doi: 10.3390/a14110318.

[30] I. B. Mahendra, I. M. G. Sunarya, and I. M. A. Wirawan, "Comparison of Multinomial, Bernoulli, and Gaussian Naïve Bayes for Complaint Classification in Pro Denpasar Application," *JUITA J. Inform.*, vol. 13, no. 1, pp. 77–86, 2025, doi: 10.30595/juita.v13i1.24828.

[31] J. A Ilemobayo, "Hyperparameter Tuning in Machine Learning: A Comprehensive Review," *J. Eng. Res. Reports*, vol. 26, no. 6 SE-Review Article, pp. 388–395, Jun. 2024, doi: 10.9734/jerr/2024/v26i61188.

[32] J. Li, "Area under the ROC Curve has the most consistent evaluation for binary classification," *PLoS One*, vol. 19, no. 12 December, 2024, doi: 10.1371/journal.pone.0316019.

[33] A. Abu Saa, M. Al-Emran, and K. Shaalan, *Factors Affecting Students' Performance in Higher Education: A Systematic Review of Predictive Data Mining Techniques*, vol. 24, no. 4. Springer Netherlands, 2019. doi: 10.1007/s10758-019-09408-7.