COMPARATIVE ANALYSIS OF RANDOM FOREST AND SUPPORT VECTOR CLASSIFIER FOR PREDICTING STUDENTS' ON-TIME GRADUATION

Nurus Sarifatul Ngaeni

Information System
STIMIK Tunas Bangsa Banjarnegara, Banjarnegara, Indonesia
www.stb.ac.id
nurussyarifatulngaeni@gmail.com
(*) Corresponding Author

Abstract—*On-time graduation is one of the key indicators of educational quality in higher education. The influencing factors range from students' internal issues and academic abilities to institutional policies. However, academic management has not yet been able to classify the data and analyze the underlying factors contributing to delayed graduation. By identifying these factors, management can formulate appropriate academic solutions or policies. The purpose of this study is to build a prediction model for on-time graduation using machine learning algorithms. This study compares the classification performance of the Random Forest algorithm and the Support Vector Classifier (SVC). The dataset, consisting of 1,298 student records, includes academic data such as study program, GPA, TOEFL score, cohort year, and study duration. Model performance was evaluated using accuracy, F1 score, and ROC-AUC metrics, followed by a confusion matrix analysis. The final evaluation revealed that the Random Forest algorithm achieved the best performance, with an accuracy of 91.86%, an F1 score of 91.86%, and a ROC-AUC of 97.39%. Meanwhile, the SVC model obtained an accuracy of 81.12% and an F1 score of 81.09%. Based on these results, it can be concluded that the Random Forest algorithm is more reliable as a prediction model in the academic domain. The main contribution of this study is the development of an early detection system for students at risk of delayed graduation. Furthermore, the findings can serve as a basis for designing more solution-oriented academic policies in accordance with the conditions at STIMIK Tunas Bangsa Banjarnegara.*

Keywords: *machine learning algorithms, on-time graduation, student data analysis, study duration.*

Abstrak—*Kelulusan tepat waktu merupakan salah satu indikator penting kualitas pendidikan di perguruan tinggi. Faktor-faktor yang memengaruhi berkisar dari masalah internal mahasiswa, kemampuan akademik, hingga kebijakan institusi. Namun, pihak manajemen akademik belum mampu melakukan klasifikasi data serta menganalisis faktor-faktor penyebab keterlambatan kelulusan. Dengan mengetahui faktor-faktor tersebut, manajemen dapat merumuskan solusi atau kebijakan akademik yang tepat. Penelitian ini bertujuan membangun model prediksi kelulusan tepat waktu menggunakan algoritma machine learning. Penelitian ini membandingkan kinerja klasifikasi algoritma Random Forest dan Support Vector Classifier (SVC). Dataset yang digunakan terdiri atas 1.298 mahasiswa, meliputi data akademik berupa program studi, IPK, skor TOEFL, angkatan, dan lama studi. Evaluasi performa dilakukan menggunakan metrik akurasi, F1 score, ROC-AUC, dan dilanjutkan dengan analisis confusion matrix. Hasil evaluasi menunjukkan bahwa algoritma Random Forest memberikan kinerja terbaik dengan akurasi 91,86%, F1 score 91,86%, dan ROC-AUC sebesar 97,39%. Sementara itu, model SVC memperoleh akurasi 81,12% dan F1 score 81,09%. Berdasarkan hasil tersebut dapat disimpulkan bahwa algoritma Random Forest lebih andal digunakan sebagai model prediksi di bidang akademik. Kontribusi utama penelitian ini adalah pengembangan sistem deteksi dini bagi mahasiswa yang berisiko tidak lulus tepat waktu. Selain itu, hasil penelitian ini dapat dijadikan dasar dalam merumuskan kebijakan akademik yang lebih solutif sesuai dengan kondisi di STIMIK Tunas Bangsa Banjarnegara.*

Kata Kunci*: algoritma machine learning, kelulusan tepat waktu, analisis data mahasiswa, lama studi,* Vol. 21, No. 2 September 2025 | DOI: 10.33480/pilar.v21i2.7048

INTRODUCTION

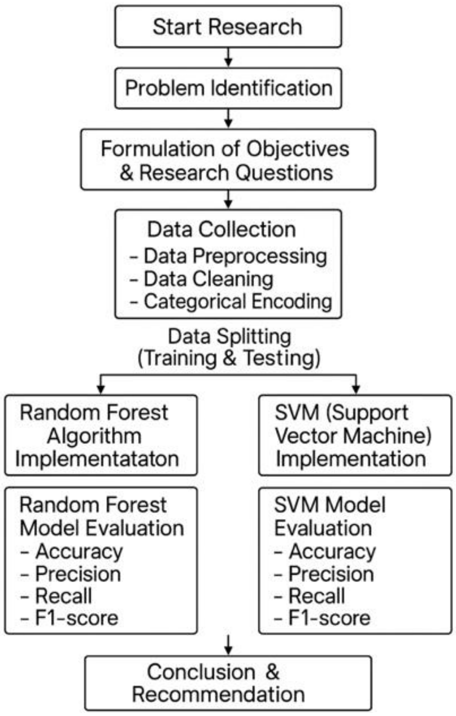
On-time graduation is a key indicator used as a benchmark for the quality of higher education. The determining factors analyzed for their impact on student graduation are academic achievement, attendance and involvement in academic activities. (Puspa et al., 2025). This is evidenced by the percentage of on-time student graduations at universities during the accreditation process for study programs and universities (Hairani, 2021). Data from the Directorate General of Higher Education (DGHE) in 2021 showed that on-time graduation rates for undergraduate students only reached 30–40% (Rohadhatul Aisy and Pramono 2023). This means that the number of Indonesian undergraduate students who fail to complete their eight semesters on time remains quite high (Purba, 2025). Utilizing technology, particularly machine learning, is expected to provide a gateway to more effective and efficient solutions (Wicaksono et al., 2023). This challenge, of course, not only causes personal and institutional losses, but also impacts graduates' competitiveness in the workforce (Ngaeni et al., 2024). To address this, several analyses using machine learning models are needed (Satrio Junaidi et al., 2024). One of these is identifying the factors causing the high rate of late graduation (Darmawan et al., 2023). Several factors contributing to delays in study include GPA, repeating courses, delays in the preparation of theses, and several technical factors within the student academic realm (Oon Wira Yuda et al., 2022). Therefore, researchers need a model-based approach to identify, classify, and analyze the patterns that cause delays in study. The goal of this analysis is for universities to formulate preventative policies tailored to the challenges or problems that contribute to overstudy.

Machine learning approaches have been widely applied in various sectors, from education to industry (Mu'tashim & Zaidiah, 2023). Predictive models are developed to address various challenges, including delays in student graduation. Previous research has demonstrated a reliable Support Vector Machine (SVM) algorithm for classification. Meanwhile, the Random Forest model has been proven to produce significant results Ermamtita & Hafyz, 2025). Each model is reliable in handling even large-scale data. Furthermore, these models can be relied upon to handle non-linear data and produce more accurate results.

By developing a student graduation prediction system, STIMIK Tunas Bangsa's internal management can evaluate and provide early intervention for students at risk of not graduating on time. Decisions and policies can be formulated based on the results of student data analysis to address the problems or challenges, thereby improving the quality of higher education more efficiently (Kartini et al., 2023).

MATERIALS AND METHODS

The method used in this study is to compare the Random Forest and Support Vector Machine (SVM) algorithms. The research design can be seen in the research flowchart presented in Figure 1.



Source: (Research Results, 2025)

Figure 1. Research Flowchart

Subjects of Research

The research subjects used were graduate data obtained from the internal STIMIK Tunas Bangsa Campus, amounting to 3368 data. However, after going through the data cleaning process, the number of suitable and representative data used was 1,298 data.

Objects of Research

The research objects used are the Random Forest and Support Vector Classifier (SVC) models which are used in the process of analyzing student accuracy in completing studies.

Tools and Materials

Some of the tools and materials used during the research process. The tools used in this research consist of two types, namely hardware and software, the details are as follows: It consists of two types: hardware and software. The details are as follows:

1. Hardware
   The hardware specifications used in this study consisted of macOS Catalina, equipped with a 2.5 GHz Dual-Core Intel Core i5 processor and 16 GB of RAM.
2. Software
   The software environment included Windows 11 Home Single Language as the operating system, Python as the programming language, and Google Colaboratory as the primary platform for cloud-based modeling and data analysis.

Data Collection

Data was collected from the public relations and administration departments of the STIMIK Tunas Bangsa Banjarnegara campus. The initial data was in Excel format, but for ease of processing, the file was converted to CSV format for analysis using Google Collaboration.

Preprocessing Data

The available dataset then enters the preparation process, which aims to obtain valid data. In the data collection stage, researchers undertake at least three stages: data preprocessing, cleaning, and category encoding. For cleaning, researchers select duplicate data, remove blanks, and remove unnecessary attributes. Meanwhile, category encoding converts categorical data into numeric data so it can be read and processed by the machine learning model.

Data Splitting

Splitting the test data is necessary to measure how well the model generalizes to previously learned datasets. This data division aims to avoid prediction errors during the model evaluation phase (Hermanto et al., 2024). The training data serves as training data and to build the analysis model, while the testing data is used to analyze the accuracy of the algorithm model that has been run. If the model's performance is stable, it indicates that the model has the reliability to make predictions using other datasets (Lestari et al., 2024). Researchers separated the test and training data with an 80:20 ratio. This ratio has been proven in the literature to provide stable performance even when handling large-scale data. This 80:20 test and training data ratio allows the model to see and learn the data more comprehensively and avoid overfitting.

Random Forest Model Evaluation

After processing the training and test data, the next step is to evaluate the performance of each model (Imran et al., 2022). The Random Forest model will then process the data classification using the required features (Purba, 2025). The purpose of this model evaluation process is to measure how well the model predicts student data through accuracy, precision, recall, and F1 score (Dewi et al., 2024).

Hyperparameters are used in the implementation of the Random Forest and SVM algorithms to achieve a balance between accuracy and data generalization. The main parameters used in the Random Forest algorithm aim to maintain prediction stability, prevent overfitting, and eliminate dependence on any one variable. 100 decision trees and a depth level of 10 are used. Internal nodes will be split if there are at least 2 samples. Random state is also used to ensure consistent replica results.

Implementation of the SVC (Support Vector Classifier) Algorithm

Another algorithm model that will be compared in this study is the Support Vector Machine Algorithm. Parameter C is used in the implementation of the SVC algorithm to overcome classification errors and the model sensitivity remains stable even on non-linear data. The Radial Basis Function (RBF) kernel is used to more effectively measure data. The parameter used is C = 1.0 to maintain a balanced decision boundary and accuracy level for the training data. The kernel coefficients are adjusted based on the data distribution (gamma = scale).

SVM Model Evaluation

The SVM (Support Vector Machine) model will be tested using a confusion matrix, which includes evaluation metrics such as accuracy, precision, recall, and F1 score. For the confusion matrix, there are at least four main components used as the basis for evaluating the performance or results of the SVM model. The following are the four main components of the confusion matrix:

1. True Negative (TN), which means the number of students who do not graduate on time is actually students who do not graduate on time.
2. False Positive (FN), which means students who graduate on time actually do not graduate on time.
3. True Positive (TP), which means students who are predicted to graduate on time actually do graduate on time.
4. False Negative (FN), which means students who do not graduate on time actually do graduate on time.

Model evaluation is a crucial step for both Random Forest and SVC (Support Vector Classifier) models (Sabita & Trisnawati, 2023). Through evaluation, researchers can determine how well the model classifies students who have the potential to graduate on time (Law et al., 2024). This process also serves as an early detection tool for students at risk of delaying their studies(Diantika et al., 2024). Of the four main components of the confusion matrix used, here are some evaluation metrics:

1. Accuracy

The accuracy value is obtained from the sum of the correct predictions (TP + TN). Overall, it can be seen how often the model or algorithm correctly classifies all review data.

2. Precision

Precision is the ratio of the number of review data correctly classified as positive (TP) compared to the total number of data classified as positive (TP + FP). Precision measures how accurately the model classifies review data labeled positive.

3. Recall (Sensitivity)

Recall measures the number of review data labeled positive and correctly classified by the model. This process allows the model to evaluate its performance in classifying positive data.

4. F1-Score

The F1-score balances precision and recall, especially when the data is imbalanced.

RESULTS AND DISCUSSION

Data Collection

The data used were student and graduate data obtained from the academic department of the STIMIK Tunas Bangsa Banjarnegara campus. The data was collected in Excel format and then sorted according to the required research attributes. The dataset was then converted to CSV format to facilitate analysis using Google Collaboration. An excerpt of the dataset structure is presented in Table 1.

Table 1. Sample Data of Student Collection from Excel Files

No	tplhrtrlsm	tglhrtrlsm	nmpstmspst	nlipktrlsm	toefltrlsm	skriptrlsm	tahuntrlsm	jalur	tepat_waktu
1	Semarang	16/02/1994	Informatika	03.24	400	Pengaruh Variasi Algoritma Terhadap Akurasi Prediksi Data Mahasiswa	2012	4	TRUE
2	Tegal	25/04/1993	Informatika	03.48	400	Analisis Keanekaragaman Pola Akses dalam Jaringan Komputer	2011	2	FALSE
3	Kota	05/10/91	Informatika	03.05	400	Pelabelan Sentimen Otomatis Pada Media Sosial	2009	0	TRUE
4	Banjar	05/06/90	Informatika	03.06	420	Identifikasi Sinyal Suara Menggunakan Algoritma Deep Learning	2008	0	FALSE
5	Semarang	09/09/89	Informatika	03.06	400	Aplikasi Pembayaran SPP Berbasis Web	2008	0	FALSE

Source: (Research Results, 2025)

The data used in this study were 1,298 students from 2012 to 2025 at the STIMIK Tunas Bangsa Banjarnegara campus. The number of students per study program was 697 in informatics and 601 in information systems.

Data Collection Result

The dataset, in CSV file format, consists of 1,298 student and graduate data. During the data collection process, only a few items related to the research topic were collected. Some of the data or information used in this study are:

1. nmpstmspst = Name of Study Program (Informatics/Information Systems)
2. nlipktrlsm = Student GPA
3. toefltrlsm = Student TOEFL Score
4. tahuntrlsm = Student Class Year
5. skriptrlsm = Thesis Title
6. tepat_waktu = Study Duration Accurate (<=4 Years) or Not (>4 Years)

Data Preprocessing

The raw data in Excel format, processed into CSV format, then underwent a cleaning process. Preprocessing was performed in Google Collaboration using several Python programming language libraries, including pandas, numpy, seaborn, and others. The following are several steps in the data cleaning process.

Dividing Training and Testing Data

Before entering the analysis process by the classification model, the data is divided into two parts, training data and testing data. The data is divided into two proportions: 20% training data and 80% testing data. The total dataset used in this study was 1,298, with 1,038 training data and 260 test data.

Data Remapping

The remapping stage is used to group the data contained in the place of birth column, where the initial data obtained is presented as city or district names. This resulted in an imbalance in the data distribution during the analysis process. To improve the analysis process, researchers grouped places of birth by province. For example, several cities such as Banjarnegara, Salatiga, Wonosobo and others will be included in the "Central Java" province category.

The next step is to simplify the student birth date data into birth year. The purpose of this regrouping is to enable the analysis process to capture data on student age cohorts. This student birth year data is considered sufficient to represent age cohorts, encompassing generational groups or student age categories. For example, to compare data on students born between 1995 and 2000 and 2001 and 2005.

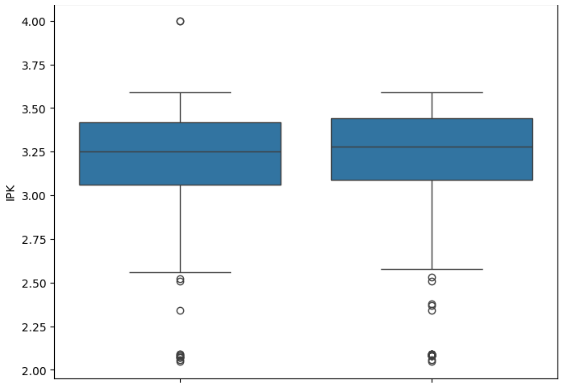
IPK Between Study Programs

To compare the distribution of GPA (Indeks Prestasi Kumulatif / IPK) across different cohorts and study programs, the researchers utilized a set of command codes as illustrated in Figure 2.

```
# Urutkan DataFrame berdasarkan IPK
df_sorted = df.sort_values(by='nlipktrlsm', ascending=False)

# Buat boxplot dengan Seaborn
plt.figure(figsize=(8, 6))
sns.boxplot(x='nmpstmspst', y='nlipktrlsm', data=df_sorted)

# Tambahkan label dan judul
plt.xlabel('Angkatan')
plt.ylabel('IPK')
plt.title('Boxplot IPK per Program Studi')

# Tampilkan plot
plt.show()
```

Source: (Research Results, 2025)
Figure 2. IPK Distribution Command Code

The execution of the command code produced a boxplot visualization, as presented in Figure 3.



Source: (Research Results, 2025)
Figure 3. Boxplot of GPA between Study Programs

As shown in Figure 3, the boxplot provides a comparison of GPA scores between students in the Information Systems and Informatics study programs. From this visualization, several conclusions and analytical results can be drawn, highlighting differences in GPA distribution across the two study programs.

1. Median GPA Scores across Study Programs Are Equivalent

The median GPA score between students in the Information Systems and Informatics study programs is around 3.3. This proves that the median GPA levels of students in both study programs are considered equivalent.

2. Distribution of GPA Scores is Almost Equal

The data representation shows that the distribution of scores across study programs is almost equal. This is evidenced by the consistent main GPA range, which includes the lower and upper limits of student scores.

3. Lower Scores Are Present in the Information Systems Study Program

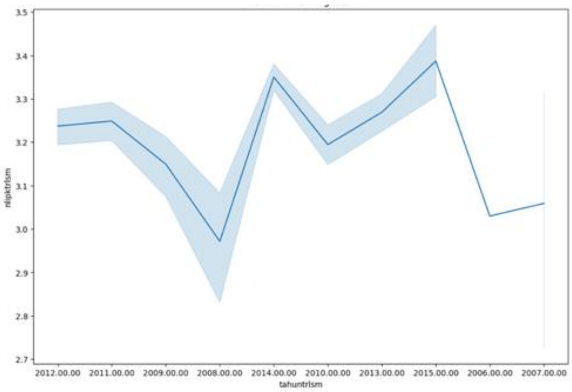
Based on the majority scores, there are several extreme scores that are more prevalent in the Information Systems study program. These scores can be said to be lower than the average score of the majority.

4. IPK Ranges Across Study Programs Are Almost Equal

The score ranges for students in the Informatics and Information Systems study programs are similar, ranging from a IPK of 2.0 to a maximum of 4.0. This data demonstrates consistency and similarity in grade ranges across study programs. However, some Information Systems students have lower-than-average IPK.

PILAR Nusa Mandiri: Journal of Computing and Information System
Vol. 21, No. 2 September 2025 | DOI: 10.33480/pilar.v21i2.7048

IPK Trends per Student Class



Source: (Research Results, 2025)
Figure 4. Student Generation GPA Trend Graph

As presented in Figure 4, the GPA trend graph illustrates the academic performance of each student cohort over the years. The figure shows that GPA values fluctuate annually without following a consistent upward or downward pattern, indicating that each cohort possesses unique characteristics in terms of academic achievement.

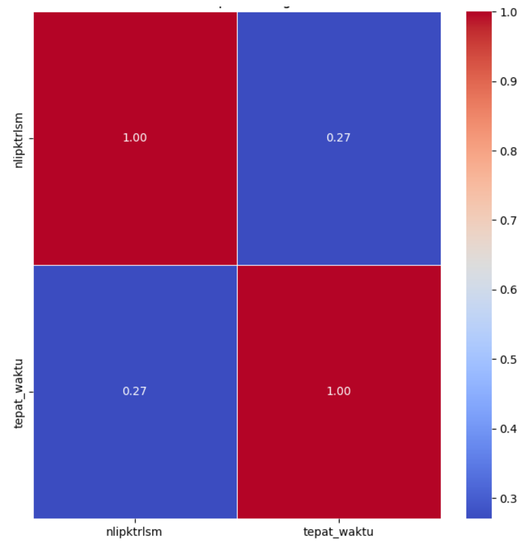
For example, the data for the 2008 cohort showed a significant decline, falling below 3.0, the lowest among all cohorts. This data should be of particular interest to the 2008 cohort for further analysis, perhaps to identify the underlying factors behind the low GPAs or the student population.

The analysis reveals that the most significant increase in GPAs occurred in the 2014 cohort, reaching between 3.35 and 3.4. However, this trend was not stable for long. The 2016 cohort saw another decline, reaching around 3.05. Subsequent data can be considered more stable within this range.

In addition, Figure 4 also depicts a confidence interval visualized as a blue shaded area surrounding the GPA trend line. The blue shading between these lines represents the degree of variation in GPA scores for each cohort. Therefore, the wider the blue shading, the greater the variation in GPA scores for each student. Conversely, if the blue shading is thinner or closer to the line, it indicates that the scores of students from each cohort are close together.

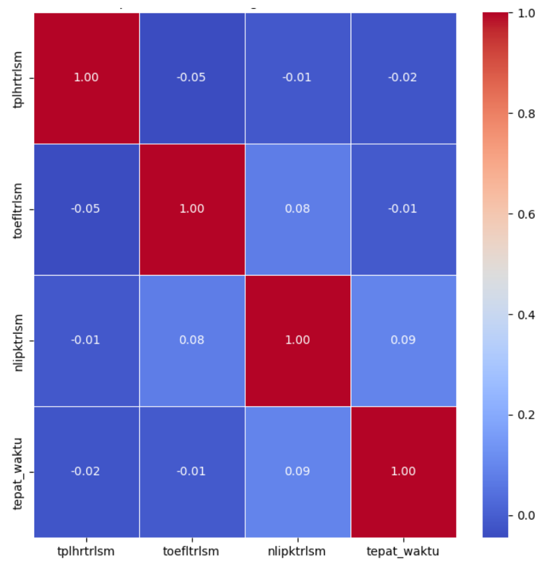
The Effect of GPA on Length of Study

The correlation heatmap in Figure 5 shows that there is still a relationship, influence, or connection between GPA and student length of study. The correlation value reaches 0.27, indicating a weak positive relationship. This means that students with high GPAs are more likely to graduate on time. However, other determinants also influence length of study, including academic policies and internal student factors such as time management, motivation or willingness, and health conditions.



Source: (Research Results, 2025)
Figure 5. Correlation Heatmap of GPA and Student Length of Study

Correlation of IPK, TOEFL, and Student Length of Study



Source: (Research Results, 2025)
Figure 6. Correlation Heatmap of Region of Origin with TOEFL, GPA, and Student Study Period

The correlation heatmap in the Figure 6 represents the relationship between each entity, starting from region of origin, GPA, graduation status, and TOEFL score. This data shows that region of origin is not a determinant of academic achievement, specifically graduation status. This fact presents a positive opportunity for students from all regions to receive their education on time without being affected by geographic location.

The correlation between GPA and on-time graduation status reached 0.09. Although relatively weak, this score demonstrates that there is still a correlation between academic achievement and students' success in graduating on time. Academics can create policies that can improve academic achievement (GPA) to boost the percentage of students graduating on time. Furthermore, the correlation between TOEFL scores and on-time graduation is relatively low, indicating that limited foreign language (English) proficiency is not a contributing factor to delayed graduation.

Classification Model Evaluation Results

1. Random Forest

The performance of the Random Forest model demonstrated excellent results, achieving both accuracy and F1-score values exceeding 91% (Table 2). In comparison, a previous study reported an accuracy of up to 98% for the Random Forest algorithm (Puspa et al., 2025). The difference lies in the methodological approach, as this study employed a comparative analysis between Random Forest and Support Vector Classifier (SVC) using a more limited dataset.

Furthermore, the evaluation results are reinforced by a near-perfect ROC-AUC score of 0.9739, as presented in Table 2, indicating that the Random Forest model is both stable and reliable in distinguishing between classes. The confusion matrix shown in Table 3 provides additional evidence, demonstrating a high proportion of correctly classified instances across both negative and positive categories, thereby underscoring the consistency of the model's predictive capability.

Table 2. Evaluation Metrics Results of Random Forest

Evaluation Matrix	Score
Average Accuracy	0.9186
Average F1 Score	0.9186
Average ROC-AUC	0.9739

Source: (Research Results, 2025)

Table 3. Results of Random Forest Confusion Matrix

	Negative (Pred)	Positive (Pred)
Negative (Actual)	342 (True Negative)	26 (False Positive)
Positive (Actual)	30 (False Negative)	275 (True Positive)

Source: (Research Results, 2025)

2. SVC (Support Vector Classifier)

The Support Vector Classifier (SVC) model achieved satisfactory performance, with an accuracy and F1-score of approximately 81% (Table 4). In addition, the model demonstrated a reliable classification capability, as reflected by the ROC-AUC score of 0.9041 (Table 4). The confusion matrix presented in Table 5 further illustrates the distribution of correctly and incorrectly classified instances across both classes.

Table 4. Evaluation Metrics Results of Support Vector Classifier

Evaluation Matrix	Score
Average Accuracy	0.8112
Average F1 Score	0.8109
Average ROC-AUC	0.9041

Source: (Research Results, 2025)

Table 5. Results of Support Vector Classifier Confusion Matrix

	Negative (Pred)	Positive (Pred)
Negative (Actual)	309 (True Negative)	59 (False Positive)
Positive (Actual)	68 (False Negative)	237 (True Positive)

Source: (Research Results, 2025)

Nevertheless, this study has several limitations that should be acknowledged. The dataset is restricted to a single educational institution and includes only two study programs. Moreover, important non-academic factors such as financial ability and family support, which may significantly influence students' on-time graduation, were not considered in the analysis. These limitations provide opportunities for future research to explore broader datasets and additional influencing factors.

CONCLUSION

In this study, two algorithm models were compared to compare their results. The models applied and analyzed were Random Forest and Support Vector Classifier (SVC), to predict students' on-time graduation based on several correlated factors. Based on the evaluation results, the Random Forest model performed best, with an average accuracy of 91.86%, an F1 score of 91.86%, and a ROC-AUC of 97.39%. These results indicate excellent classification performance, evidenced by a balance between precision and recall. The Support Vector Classifier (SVC) model, on the other hand, performed lower, with an accuracy of 81.12%, an F1 score of 81.09%, and a ROC-AUC of 90.41%. While these results are quite good, they are lower when compared to the Random Forest model, which has relatively high accuracy.

Based on these analysis results, the Random Forest model is recommended as the best model for this study. This is supported by higher accuracy, stability, and consistency. The Random Forest model is also reliable for processing data analysis with numeric and categorical features and is more resistant to outliers and overfitting. Further research is expected to be applicable to educational institutions with larger datasets, examining non-academic factors such as financial ability and family support and factors that may impact student graduation.

REFERENCE

Darmawan, A., Yudhisari, I., Anwari, A., & Makruf, M. (2023). Pola Prediksi Kelulusan Siswa Madrasah Aliyah Swasta dengan Support Vector Machine dan Random Forest. *Jurnal Minfo Polgan*, *12*(1), 387–400. https://doi.org/10.33395/jmp.v12i1.12388

Dewi, A. C., Hermawan, A., & Avianto, D. (2024). Classification of Customers' Repeat Order Probability Using Decision Tree, Naïve Bayes and Random Forest. *Jurnal Pilar Nusa Mandiri*, *20*(1), 52–59. https://doi.org/10.33480/pilar.v20i1.5243

Diantika, S., Nalatissifa, H., Maulidah, N., Supriyadi, R., & Fauzi, A. (2024). Penerapan Teknik Random Oversampling Untuk Memprediksi Ketepatan Waktu Lulus Menggunakan Algoritma Random Forest. *Computer Science (CO-SCIENCE)*, *4*(1), 11–18. https://doi.org/10.31294/coscience.v4i1.1996

Ermamtita, E., & Hafyz, H. (2025). Prediksi Kelulusan Mahasiswa Menggunakan Metode K-Means dan Random Forest. *Jurnal Pendidikan Dan Teknologi Indonesia, 5*(1), 209–214. https://doi.org/10.52436/1.jpti.577

Hairani, H. (2021). Peningkatan Kinerja Metode SVM Menggunakan Metode KNN Imputasi dan K-Means-Smote untuk Klasifikasi Kelulusan Mahasiswa Universitas Bumigora. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, *8*(4), 713–718. https://doi.org/10.25126/jtiik.2021843428

Hermanto, D., Ricoida, D. I., Pibriana, D., & Pribadi, M. R. (2024). Analysis of Student Graduation Prediction Using Machine Learning Techniques on an Imbalanced Dataset: An Approach to Address Class Imbalance. *Scientific Journal of Informatics, 11*(3), 559–568. https://doi.org/10.15294/sji.v11i3.5528

Imran, B., Hambali, H., Subki, A., Zaeniah, Z., Yani, A., & Alfian, M. R. (2022). Data Mining Using Random Forest, Naïve Bayes, and Adaboost Models for Prediction and Classification of Benign and Malignant Breast Cancer. *Jurnal Pilar Nusa Mandiri*, *18*(1), 37–46. https://doi.org/10.33480/pilar.v18i1.2912

Kartini, Fetty Tri Anggraeny, Aang Kisnu Darmawan, Anik Anekawati, & Ivana Yudhisari. (2023). Early Prediction for Graduation of Private High School Students with Machine Learning Approach. *Technium: Romanian Journal of Applied Sciences and Technology*, *16*, 129–136. https://doi.org/10.47577/technium.v16i.9971

Law, T. J., Ting, C. Y., Ng, H., Goh, H. N., & Albert, Q. (2024). A Review of Graduate on Time Prediction. *International Journal on Advanced Science, Engineering and Information Technology*, *14*(6), 1957–1966. https://doi.org/10.18517/ijaseit.14.6.17475

Lestari, W., Abdullah, A. S., Amin, A. M. A., Nurfaridah, Sukotjo, C., Ismail, A., Ibrahim, M. S. M., Insani, N., & Utomo, C. P. (2024). Artificial intelligence to predict pre-clinical dental student academic performance based on pre-university results: A preliminary study. *Journal of Dental Education*, *88*(12), 1681–1695. https://doi.org/10.1002/jdd.13673

Mu'tashim, M. L., & Zaidiah, A. (2023). Klasifikasi Ketepatan Lama Studi Mahasiswa Dengan Algoritma Random Forest Dan Gradient Boosting (Studi Kasus Fakultas Ilmu Komputer Universitas Pembangunan Nasional Veteran Jakarta). *Seminar Nasional Mahasiswa Bidang Ilmu Komputer Dan Aplikasinya (SENAMIKA)*, *4*(1), 155–166.

Ngaeni, N. S., Kusrini, K., & Kusnawi, K. (2024). Analisis Kombinasi Algoritma K-Means Clustering dan TOPSIS Untuk Menentukan Pendekatan Strategi Marketing Berdasarkan Background Target Audiens. *Journal of Computer System and Informatics (JoSYC)*, *5*(2), 393–403. https://doi.org/10.47065/josyc.v5i2.4948

Oon Wira Yuda, Darmawan Tuti, Lim Sheih Yee, & Susanti. (2022). Penerapan Penerapan Data Mining Untuk Klasifikasi Kelulusan Mahasiswa Tepat Waktu Menggunakan Metode Random Forest. *SATIN - Sains Dan Teknologi Informasi*, *8*(2), 122–131. https://doi.org/10.33372/stn.v8i2.885

Purba, S. E. M. (2025). A Comparative Study of Drug Prediction Models using KNN, SVM, and Random Forest. *Journal of Information Systems and Informatics, 7*(1), 378–392. https://doi.org/10.51519/journalisi.v7i1.1013

Puspa, G., Rachmadhan Amri, M. F., & Nugraha, M. P. (2025). Pemanfaatan Random Forest untuk Prediksi Ketepatan Waktu Kelulusan Mahasiswa Studi Kasus: Institut Desain dan Bisnis Bali. *JIKO (Jurnal Informatika Dan*

Komputer), *9*(2), 413. https://doi.org/10.26798/jiko.v9i2.1886

Sabita, H., & Trisnawati, S. (2023). Perbandingan Algoritma Support Vector Machine dan AdaBoost Dalam Memprediksi Waktu Kelulusan Mahasiswa. *TEKNIKA*, *17*(2), 359–372. https://doi.org/10.5281/zenodo.8220872

Satrio Junaidi, Valicia Anggela, R., & Kariman, D. (2024). Klasifikasi Metode Data Mining untuk Prediksi Kelulusan Tepat Waktu Mahasiswa dengan Algoritma Naïve Bayes, Random Forest, Support Vector Machine (SVM) dan Artificial Neural Nerwork (ANN). *Journal of Applied Computer Science and Technology*, *5*(1), 109–119. https://doi.org/10.52158/jacost.v5i1.489

Wicaksono, S. A., Wijaksana, C. P., Ngaeni, N. S., & Utami, E. (2023). Systematic Literature Review of the Development of Financial Information Systems in Churches. *International Journal of Computer and Information System (IJCIS)*, *4*(3), 178–182. https://doi.org/10.29040/ijcis.v4i4.143