ICIC Express Letters
Part B: Applications
Volume **16**, Number **12**, December **2025**

ICIC International ©2025 ISSN 2185-2766
pp. 1303–1310

PREDICTING STUDENTS' COMPLETION TIME BASED ON SEMESTERLY PERFORMANCE

Bill Kiki[1], Friska Natalia[1,*], Chang Seong Ko[2] and Sud Sudirman[3]

[1]Faculty of Engineering and Informatics
Universitas Multimedia Nusantara
Jl. Scientia Boulevard, Gading Serpong, Tangerang, Banten 15227, Indonesia
bill.kiki@student.umn.ac.id; *Corresponding author: friska.natalia@umn.ac.id

[2]Department of Industrial and Management Engineering
Kyungsung University
309, Suyeong-ro, Nam-gu, Busan 48434, Korea
csko@ks.ac.kr

[3]School of Computer Science and Mathematics
Liverpool John Moores University
Liverpool, L3 3AF, United Kingdom
s.sudirman@ljmu.ac.uk

Received January 2025; accepted April 2025

Abstract. *Recent years have seen a notable rise in higher education institutions, spurred by competitive pressures among universities. A critical measure of educational quality lies in students' completion rates, encompassing the duration of their studies as a significant component. Protracted or failed completions not only squander time but also drain financial and energy resources for individuals pursuing higher education and at the same time they can also tarnish an institution's reputation and accreditation status. In this study, we construct a predictive model for students' completion time to improve students' awareness and furnish academic advisors and staff with valuable insights to tailor academic guidance. Employing three classification algorithms – Naïve Bayes, Support Vector Machine, and Neural Network – the study optimizes these algorithms using the Particle Swarm Optimization technique to yield superior results. The Neural Network algorithm emerges as the optimal performer post-optimization, boasting accuracy, precision, specificity, sensitivity, and F1-score metrics of 0.86, 0.86, 0.86, 0.98, and 0.86, respectively, alongside superior confusion matrix outcomes. The model's efficacy will be harnessed to forecast data for active-status students and disseminated via a dedicated website.*
Keywords: Educational data mining, Naïve Bayes, Support Vector Machine, Neural Network

1. **Introduction.** Higher education providers, such as universities and colleges, play several critical roles in higher education. They are responsible for preparing future professionals to ensure continuing advancement of society as well as conducting academic research across various disciplines to advance knowledge and address societal challenges [1]. In many countries, higher education providers can be categorized into non-profit (typically state-owned) and for-profit (usually private or commercial) institutions. In Indonesia for example, there are three categories, namely public universities, private universities, and state-owned universities. The higher education business in Indonesia is thriving which is evidenced by the increasing number of universities over the years. From 2008 to 2017, the number of higher education institutions registered under the Ministry of Education, Culture, Research, and Technology rose from 2,680 to 3,276 [2]. As a result, there is intense competition between these universities to draw as many bright school-leavers as possible.

DOI: 10.24507/icicelb.16.12.1303

This drives each university to strive for excellence, aiming to enhance the quality of education provided to their students [3]. Accreditation, managed by the Indonesian National Accreditation Board for Higher Education, is commonly used as a parameter to measure university quality. One of the key components influencing accreditation is the average study period or completion time [1]. University students in Indonesia can complete their studies between 7 semesters (3.5 years) and 14 semesters (7 years) – although this upper limit has now been reduced to 10 semesters by the National Higher Education Standards body. Therefore, the average length of time for students to complete their studies is one indication of the success of the education provider in delivering their services.

Graduation entails a series of steps for students, including internships, completing a minimum number of courses, research proposal seminars, and other requirements set by the institution. However, data from the Higher Education Statistics Report in 2020 revealed a concerning dropout rate of 79.5% in private universities, totaling 478,826 students [4]. Various factors can contribute to this, including exceeding the maximum study duration, compounded by recent government regulations reducing the maximum study period for undergraduate education from 7 to 5 years [5]. Delayed completion or dropouts not only disadvantage the university but also impose financial, temporal, and emotional burdens on students [6]. Such delays can have significant global repercussions, affecting the students psychologically and impeding career advancement [7]. In light of these challenges, guidance and counseling, as forms of student support, are vital for academic success [8]. Most universities assign an academic advisor to each student to assist them in planning their studies. The role involves analysis of the student's progress and providing feedback to the student on their expected completion time. This is traditionally performed manually by the academic advisor and comes with a number of challenges. In this study, we aim to improve this process by developing a predictive model for student completion time using data mining and machine learning approaches. The predictive model serves as a reference for monitoring student progress, aiming to prevent delayed completions and dropouts.

A study by Hussain et al. [9] uses Levenberg Marquardt Algorithm deep learning algorithm to predict the performance of the students based on student-related data to increase the overall performance. The data used consists of the class test, attendance, assignment, and midterm scores. Using a model that has four input variables, three hidden and one output layer, the proposed model is reported to produce a prediction accuracy of 88.6%. Machine learning is also used in [10] to predict student retention and graduation time with reported accuracy between 0.909 for retention and 0.817 for graduation time predictions. A review of the current approaches for student outcome predictions is given in [11]. The paper provides a primer on how data from a Learning Management System (LMS) can be feature-mapped and analyzed to predict students' outcomes and design necessary interventions. The paper provides an overview of popular machine learning algorithms and a review of analytic considerations such as feature generation, assessment of model performance, and sampling techniques by providing an empirical example demonstrating the ability of LMS data to predict student success.

Our study employs a predictive model by selecting the best-performing algorithm among Artificial Neural Networks (ANN), Naïve Bayes (NB), and Support Vector Machine (SVM), which is optimized using the Particle Swarm Optimization (PSO) method. ANN is chosen based on its previously reported accuracy exceeding 90% [12,13]. NB is included due to its documented accuracy performance surpassing 90% [1] whereas SVM is selected for comparison as it has demonstrated superior performance in previous research [14,15]. The PSO technique has been proven to be effective in optimizing and enhancing the performance of machine learning algorithms as indicated in [16]. The research yields insights into student performance and a predictive model for completion time based on variables such as semester, credit units, semester GPA, cumulative GPA, and total study period. The results of this study will provide valuable information for universities seeking to reduce their students' drop-out rates and for students seeking to improve their study plans.

The paper is organized as follows. In Section 2 we describe the data used and the methodology that was adopted. In Section 3, we present the experiment result and provide our analysis and discussion before concluding our findings in Section 4.

2. Material and Research Method.

2.1. Material. The data utilized in this research consists of historical data of 5,660 students at various stages (from 7th to 14th semester) of their study at Universitas Multimedia Nusantara spanning from 2014 to 2022 academic years. The data has been anonymized and none of the information could be used to link it to any of the students. The distribution of the data is shown in Table 1. Each number corresponds to the number of records (students) having data up to the specified semester. For example, there are 340 student records showing data from semesters 1 to 7, etc. Each record contains seven information for each student which is the cumulative Grade Point Average (GPA) per semester, Semester Grade Point Average (SGPA) per semester, current semester, number of credit hours taken per semester, number of credit hours passed per semester, number of credit hours failed per semester, and total semester credit units taken.

Table 1. The distribution of data across the considered semester range

Semester (up to)	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
Data size	340	3,160	1,020	440	280	40	80	300

Data preparation involves several processes including column and data anomaly removal (data cleansing), data reformatting, data augmentation, data separation, and data normalization, along with the utilization of a one-hot encoding technique on labels. Data cleansing is conducted to tidy up the data structure for usability. Reformatting is necessary as the acquired data represents each row depicting a student's semester-wise study outcomes. Null value handling is executed using the data imputation technique and indicator columns. Normalization is performed to standardize data values to a mean of 0 and a variance of 1, while the one-hot encoding technique ensures labels do not represent values 7 to 14 for graduating semesters but rather values of between 0 and 1. In the data augmentation stage, three additional pieces of information are derived from the original seven and added to the data. They are the total semester credit units passed, interim semester participation, and total remaining credits. So, in total, we have a 10-dimensional vector per student per semester.

Since each record has varying widths (between 7 and 14), some of the data is padded with repeating information to make their width equal. In addition, since we also run a short semester between odd-numbered semesters and even-numbered semesters after the first year (e.g., between the 3rd and 4th semester, and the 5th and 6th semester) the total width of each record becomes 20. Following this, data is split into training and testing sets utilizing the k-fold cross-validation method [17]. Each model training employs the *StratifiedKFold()* function from the *sklearn* library for data splitting, ensuring consistent proportions of data in each fold. A total of 10 folds are chosen for cross-validation as it yields the best performance compared to 3, 5, or 15 folds, considering model generalization and data distribution balance across iterations.

2.2. Research method. The methodology employed in this study is based on the cross-industry standard process for data mining method as illustrated in Figure 1. The process commences with the business understanding phase including background and problem identification followed by a proposal of possible solutions as described in the previous

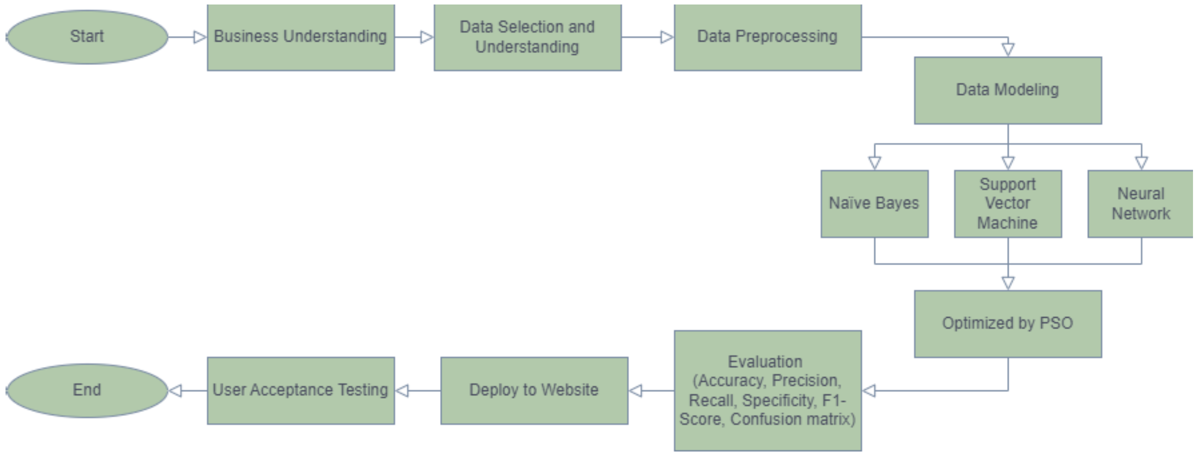


FIGURE 1. The overview of the adopted methodology

introduction section, previously. The data understanding, selection, and preprocessing stages have been elaborated in the previous section.

The process then proceeds to data model creation, where three classification models are developed using the Naïve Bayes, Support Vector Machine, and Neural Network algorithms. The comparison of the performance results of these algorithms reveals an imbalance, as certain algorithms by default yield superior outcomes for specific datasets, while others require hyperparameter tuning. Consequently, optimization of these three algorithms is carried out using PSO, resulting in performance enhancements for all three algorithms.

The performance of the models is compared by using several predefined metrics, including accuracy, precision, recall (or sensitivity), specificity, F1-score, and confusion matrix. The formulas for calculating these metrics are given below:

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

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

$$Recall\ (or\ Sensitivity) = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$F1 = \frac{Precision \times Recall}{Precision + Recall}$$

where TP, TN, FP, and FN are the number of True Positive, True Negative, False Positive, and False Negative, respectively.

The model exhibiting the best performance is selected and directly employed by implementing it on the active student data. Evaluation is also conducted on the prediction system implemented on the website, employing the User Acceptance Testing concept through questionnaire dissemination utilizing the Likert scale. The questionnaire will be distributed to two parties: students and academic advisors or staff members.

Following the determination of the best algorithm based on evaluation metrics, the final stage involves deployment, where the chosen model is utilized to predict the completion semester of active students and obtain prediction results, which are then displayed on the website. The prediction process is performed on the backend rather than directly on the website to enhance effectiveness and efficiency.

3. **Experiment Result, Analysis, and Discussion.** The performance results of the three algorithms optimized using the PSO algorithm indicate that the Neural Network algorithm outperforms Naïve Bayes and Support Vector Machine algorithms in classifying student data used in this study. The application of educational data mining in building a predictive model for student completion time using classification algorithms involves creating and comparing three classification algorithm models: Naïve Bayes, Support Vector Machine, and Neural Network, all optimized using the Particle Swarm Optimization algorithm within the CRISP-DM framework stages: business understanding, data understanding, data preparation, modeling, evaluation, and deployment.

Based on the evaluation metrics obtained from the three algorithms as shown in Table 2, the Neural Network demonstrates the highest evaluation metrics compared to Naïve Bayes and Support Vector Machine. The average values of accuracy, precision, sensitivity, specificity, and F1-score for the Neural Network are 0.86, 0.86, 0.86, 0.98, and 0.86, respectively. These values are higher than those of Naïve Bayes, which are 0.63, 0.61, 0.63, 0.90, and 0.54, and Support Vector Machine, which are 0.80, 0.74, 0.54, 0.95, and 0.59, respectively. Further evaluation is conducted using a confusion matrix, where the Neural Network algorithm demonstrates the highest total true positives and true negatives, as well as the fewest false positives and false negatives, in 7 out of 8 outcome classes, specifically in the classes of semester 7, 9, 10, 11, 12, 13, and 14, as shown in Figure 2. The Neural Network model developed from the first fold is identified as the best model and is saved for predicting new data from actively enrolled students.

Table 2. Comparison of classification performance results of the three algorithms

Metrics	Naïve Bayes	SVM	Neural Network
Accuracy	0.63	0.80	**0.86**
Precision	0.61	0.74	**0.86**
Recall/Sensitivity	0.63	0.54	**0.86**
Specificity	0.90	0.95	**0.98**
F1-score	0.54	0.59	**0.86**

```
[   0  340    0    0    0    0    0    0]
[   0 3160    0    0    0    0    0    0]
[   0  853  160    7    0    0    0    0]
[   0  272   14  154    0    0    0    0]
[   0  172    9   15   80    4    0    0]
[   0   24    2    2    4    8    0    0]
[   0   49    3    4    8    4    6    6]
[   0  182   10   18   30   15   26   19]
```

(a)

```
[ 193  145    0    2    0    0    0    0]
[  31 3093   33    1    0    0    0    2]
[   3  374  627    5    8    0    0    3]
[   1  123   51  236   15    0    1   13]
[   0   52   51   27  118    0    1   31]
[   0    3    4    0   13    8    0   12]
[   0    6   12    8    8    1   10   35]
[   0   15   12    1   13    0    6  253]
```

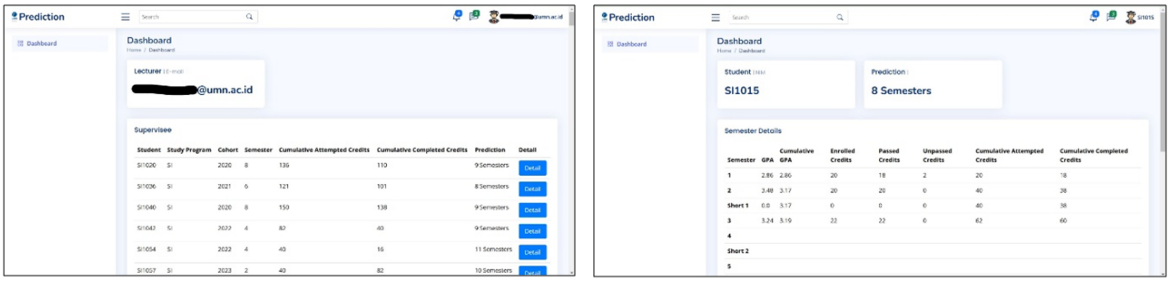
(b)

```
[ 216  123    1    0    0    0    0    0]
[  57 3017   64   13    5    0    1    3]
[   2  214  775   16   10    1    1    1]
[   1   54   36  335   10    2    0    2]
[   0   21   35   14  191    3   10    6]
[   0    0    3    1    6   30    0    0]
[   0    3    8    0   10    2   42   15]
[   0    3    2    0    3    1    5  286]
```

(c)

Figure 2. The confusion matrix result of (a) the Naïve Bayes algorithm, (b) the Support Vector Machine algorithm, and (c) the Neural Network algorithm

To enhance efficiency, the model is not run on the website since new data is not entered on the website but is instead acquired in mass each semester. The model runs independently (back-end), as the volume of new data is substantial. Predicted labels are automatically stored in the database using Python libraries. Predicted data as well as the data of actively enrolled students stored in the database are displayed on the website for two types of users: students and supervising professors. The user interface for both types of users has slight differences, with supervising professors having an additional page to view an overview of their supervised students before accessing detailed information, as shown in Figure 3.



(a) (b)

FIGURE 3. Two website screenshots showing predicted results for completion times in semesters. Screen (a) shows the first website interface for the supervisor user type and (b) shows the second website interface for the supervisor user type & shows the first website interface for the student user type.

TABLE 3. The summary of the respondents' review

Question	Score
The prediction results can provide information or insights regarding the prediction of students' completion time for students.	4.3
To what extent is the student completion time prediction system beneficial for students?	4.3
The prediction results can increase awareness regarding graduating on time for students.	4.4
Overall rating regarding the student completion time prediction system.	4.4
The prediction results can provide information or insights regarding students' completion time prediction for supervising professors or academic staff.	3
The prediction results can assist supervising professors in adjusting the guidance given to students.	3
To what extent is the student completion time prediction system beneficial for supervising professors or academic staff?	4
Overall rating regarding the student completion time prediction system.	4

The effectiveness of the web application is assessed through User Acceptance Testing by taking feedback from the 74 users via questionnaires. Five questions were posed to each respondent and they were asked to provide a score from 1 to 5 with 1 being 'strongly disagree' and 5 'strongly agree'. The questions and the feedback results are shown in Table 3. According to the table, the respondents agree that the student completion time prediction system receives an overall rating of "good" or 4 scores out of 5. The table also indicates that the prediction system is able to provide information on completion time prediction and raise awareness among students about timely completion. However, the table suggests that although the respondents are neutral (3 scores out of 5) when deciding if the system has provided insights and assistance to supervising professors in adjusting guidance for their students, it still benefits supervising professors and academic staff. Overall, the questionnaire results provide additional validation that the student completion time prediction system can offer benefits or positive impacts as an alternative to minimizing instances of late completion or dropout due to study durations exceeding the maximum duration.

4. **Conclusions.** In this paper, we presented the result of our study of implementing educational data mining to develop a predictive system for students' completion time.

We compare three classification algorithms, namely the Naïve Bayes, Support Vector Machine, and Neural Network, optimized using Particle Swarm Optimization. The CRISP-DM framework guides the data processing through six stages: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. Our experimental results show that Neural Network outperforms Naïve Bayes and Support Vector Machine in evaluation metrics, further confirmed through a confusion matrix analysis, showing Neural Network with the highest true positive and true negative rates and the lowest false positive and false negative rates across most outcome classes. The Neural Network model, identified as the best-performing algorithm, is deployed for implementation on a website, beginning with predicting completion timelines for active-status students and storing prediction labels in a database. The website interface developed using the models has been shown to have the desired outcome of the project.

REFERENCES

[1] L. Setiyani, M. Wahidin, D. Awaludin and S. Purwani, Analisis prediksi kelulusan mahasiswa tepat waktu menggunakan metode data mining Naïve Bayes: Systematic review (Prediction analysis of student graduation on time using the Naïve Bayes data mining method: Systematic review), *Fakt. Exacta*, vol.13, no.1, pp.35-43, 2020.

[2] Lokadata, *Jumlah Perguruan Tinggi (Number of Higher Education Institutions)*, 2017 (in Indonesia).

[3] A. Budiyantara, I. Irwansyah, E. Prengki, P. A. Pratama, N. Wiliani et al., Komparasi algoritma decision tree, Naïve Bayes dan K-Nearest Neighbor untuk memprediksi mahasiswa lulus tepat waktu (Comparison of decision tree, Naïve Bayes and K-Nearest Neighbor algorithms to predict students graduating on time), *J. Ilmu Pengetah. Dan Teknol. Komputer (Journal Comput. Sci. Technol.)*, vol.5, no.2, pp.265-270, 2020.

[4] Databoks, *PTS Sumbang 79,5% Mahasiswa Putus Kuliah Pada 2020 (Private Universities Contribute to 79.5% Drop out Total in 2020)*, 2020.

[5] L. W. Wadani and A. Adil, Sistem informasi untuk mengetahui penyebab mahasiswa/i tidak lulus tepat waktu studi kasus STMIK Bumigora Mataram (Information system to find out the causes of students not graduating on time case study of STMIK Bumigora Mataram), *J. Bumigora Inf. Technol.*, vol.1, no.1, pp.60-68, 2019.

[6] M. Alkaff, E. S. Wijaya and A. Rojali, Sistem peringatan dini keterlambatan masa studi mahasiswa menggunakan metode support vector machine (Early warning system for delays in student study periods using the support vector machine method), *SMARTICS J.*, vol.6, no.2, pp.54-61, 2020.

[7] C. Aina and F. Pastore, Delayed graduation and overeducation in Italy: A test of the human capital model versus the screening hypothesis, *Soc. Indic. Res.*, vol.152, no.2, pp.533-553, 2020.

[8] N. Nurochim, Dinamika keberfungsian dosen penasehat akademik bagi mahasiswa (Dynamics of academic advisors function for students), *JPPI (Jurnal Penelit. Pendidik. Indones.)*, vol.7, no.1, pp.1-7, 2021.

[9] M. M. Hussain, S. Akbar, S. A. Hassan, M. W. Aziz and F. Urooj, Prediction of student's academic performance through data mining approach, *J. Informatics Web Eng.*, vol.3, no.1, pp.241-251, 2024.

[10] K. Okoye, J. T. Nganji, J. Escamilla and S. Hosseini, Machine learning model (RG-DMML) and ensemble algorithm for prediction of students' retention and graduation in education, *Comput. Educ. Artif. Intell.*, vol.6, 100205, 2024.

[11] C. J. Arizmendi, M. L. Bernacki, M. Raković, R. D. Plumley, C. J. Urban, A. T. Panter, J. A. Greene and K. M. Gates, Predicting student outcomes using digital logs of learning behaviors: Review, current standards, and suggestions for future work, *Behav. Res. Methods*, vol.55, no.6, pp.3026-3054, 2023.

[12] M. Anwar, Prediction of the graduation rate of engineering education students using artificial neural network algorithms, *Education*, vol.5, no.1, pp.15-23, 2021.

[13] N. M. Suhaimi, S. Abdul-Rahman, S. Mutalib, N. H. A. Hamid and A. Hamid, Review on predicting students' graduation time using machine learning algorithms, *Int. J. Mod. Educ. Comput. Sci.*, vol.11, no.7, pp.1-13, 2019.

[14] V. Riyanto, A. Hamid and R. Ridwansyah, Prediction of student graduation time using the best algorithm, *Indones. J. Artif. Intell. Data Min.*, vol.2, no.1, pp.1-9, 2019.

[15] E. Haryatmi, S. P. Hervianti et al., Penerapan algoritma support vector machine untuk model prediksi kelulusan mahasiswa tepat waktu (Application of the support vector machine algorithm for prediction models for on time student graduation), *Rekayasa Sist. Dan Teknol. Inf. (Systems Eng. Inf. Technol.)*, vol.5, no.2, pp.386-392, 2021.

[16] A. Noercholis and M. Zainuddin, Comparative analysis of 5 algorithm based particle swarm optimization (PSO) for prediction of graduate time graduation, *MATICS J. Ilmu Komput. Dan Teknol. Inf. (Journal Comput. Sci. Inf. Technol.)*, vol.12, no.1, pp.1-9, 2020.

[17] N. B. Setiawan, F. Natalia, F. V. Ferdinand, S. Sudirman and C. S. Ko, Classification of skin diseases and disorders using convolutional neural network on a mobile application, *ICIC Express Letters, Part B: Applications*, vol.12, no.8, pp.715-722, 2021.