Research Article

Agus Santoso*, Heri Retnawati, Kartianom, Ezi Apino, Ibnu Rafi, Munaya Nikma Rosyada

Predicting Time to Graduation of Open University Students: An Educational Data Mining Study

https://doi.org/10.1515/edu-2022-0220
received September 17, 2023; accepted January 25, 2024

Abstract: The world’s move to a global economy has an impact on the high rate of student academic failure. Higher education, as the affected party, is considered crucial in reducing student academic failure. This study aims to construct a prediction (predictive model) that can forecast students’ time to graduation in developing countries such as Indonesia, as well as the essential factors (attributes) that can explain it. This research used a data mining method. The data set used in this study is from an Indonesian university and contains demographic and academic records of 132,734 students. Demographic data (age, gender, marital status, employment, region, and minimum wage) and academic (i.e., grade point average (GPA)) were utilized as predictors of students’ time to graduation. The findings of this study show that (1) the prediction model using the random forest and neural networks algorithms has the highest classification accuracy (CA), and area under the curve (AUC) value in predicting students’ time to graduation (CA: 76% and AUC: 79%) compared to other models such as logistic regression, Naïve Bayes, and k-nearest neighbor; and (2) the most critical variable in predicting students’ time to graduation along with six other important variables is the student’s GPA.

Keywords: educational data mining, machine learning, Open University, time to graduation

* Corresponding author: Agus Santoso, Department of Statistics, Universitas Terbuka, South Tangerang 15437, Banten, Indonesia, e-mail: aguss@ecampus.ut.ac.id

Heri Retnawati: Master’s and Doctoral Programs in Educational Research and Evaluation, Yogyakarta State University, Sleman 55281, Special Region of Yogyakarta, Indonesia; Department of Mathematics Education, Yogyakarta State University, Sleman 55281, Special Region of Yogyakarta, Indonesia

Kartianom, Ezi Apino: Doctoral Program in Educational Research and Evaluation, Yogyakarta State University, Sleman 55281, Special Region of Yogyakarta, Indonesia

Ibnu Rafi, Munaya Nikma Rosyada: Department of Mathematics Education, Yogyakarta State University, Sleman 55281, Special Region of Yogyakarta, Indonesia

The world’s transition to a global economy has positively impacted increasing individual interest in pursuing higher education. Individuals and governments have recognized the value of higher education in terms of competitiveness and prosperity (Alturki, Cohausz, & Stuckenschmidt, 2022). Unfortunately, this increased interest also impacts the high rate of academic failure (so-called drop-out) among students (Alturki et al., 2022). One of the key concerns of higher education institutions is that the high academic failure rate is costly for students, and subsequently has an impact on the institution and country (Kim, Choi, Jun, & Lee, 2023). At the higher education level, academic failure harms the education system (Batool et al., 2023). In developing countries like Indonesia, academic failure can divert productivity and competitiveness, exacerbate income inequality, and have long-term macroeconomic effects (Colak Oz, Güven, & Nápoles, 2023; Cruz-Jesus et al., 2020). As a result, student academic failure is a challenging problem that needs attention, particularly for the higher education institution itself.

Higher education institutions play a critical role in reducing student academic failure or in other words, promoting the academic success of their students. When it comes to academic success or failure, we first need to define what we mean by academic success or failure. We use the model proposed by York, Gibson, and Rankin (2015) in defining academic success, where academic success can be a success in terms of “academic achievement, attainment of learning objectives, acquisition of desired skills and competencies, satisfaction, persistence, and post-college performance.” (p. 5). Based on the coverage of academic success, we are interested in further exploring the academic success or failure of students in higher education from the aspect of persistence, where this aspect concerns how long students can complete their degree (York et al., 2015). According to Sarra, Fontanella, and Di Zio (2019), one essential academic goal is that many students are able to complete their studies or degrees and are able to complete them in a timely manner or at a reasonable time. As a result, higher education institutions must take preventive
actions early on to ensure that their students do not waste their time and lose interest in their studies (Kim et al., 2023). One alternative concrete action higher education institutions can take is to build a responsive system through a predictive model to predict student behavior (Alturki et al., 2022). Data containing student records in higher education institutions can help build predictive models (Cruz-Jesus et al., 2020). As did Burgos et al. (2018), they successfully built a prediction model using student records in higher education institutions and managed to reduce the academic failure rate by 14% from the previous year. However, there are few robust models for predicting student academic success; in other words, the prediction model utilized student records at higher education institutions to improve the quality of teaching and learning practices and to derive data-driven policies (Moscoso-Zea, Saa, & Luján-Mora, 2019). The most popular DM algorithms used are Random Forest (Alturki et al., 2022; Rebai et al., 2020; Xu, Wang, Peng, & Wu, 2019; Yaşçi, 2022), Naive Bayes (Sassirekha & Vijayalakshmi, 2022; Yaşçi, 2022), logistic regression, k-Nearest Neighbors (kNN) (Cruz-Jesus et al., 2020; Sassirekha & Vijayalakshmi, 2022; Yaşçi, 2022), and Neural Network (Waheed et al., 2020; Yaşçi, 2022). The DM algorithm has a prediction accuracy of 50–81% (Yağcı, 2022). However, the accuracy of the prediction model of the DM algorithm is strongly dependent on the input variables (or attributes) used, which has received less attention in prior studies (Colak Öz et al., 2023). As a result, it is crucial to identify the attributes that can significantly improve the performance of the predictive model, which implies that through the model and these attributes, student graduation time can be predicted with high accuracy.

Current trends in predictive modeling for student academic performance or success are more likely to focus on developed countries with comprehensive databases. For example, they were using personality and presence attributes (Alturki et al., 2022; Jeno, Danielsen, & Raahelm, 2018; Mohd Khairy, Adam, & Yaakub, 2018; Roslan & Chen, 2023), culture (Moore, Dev, & Goncharova, 2018), learning strategies, perceptions of social support, motivation, socio-demographics, health conditions, and academic performance (Musso et al., 2020). These attributes are challenging to find in developing countries (Alturki et al., 2022), such as Indonesia. Attributes that have proven significant and may be available in developing countries are gender (Cruz-Jesus et al., 2020), age, employment status, previous GPA score (Alturki et al., 2022), parental income (Sarra et al., 2019), and marital status (Colak Öz et al., 2023). In addition, in this study, student’s place of residence and the provincial minimum wage were added as a substitute for cultural and family income attributes, respectively, which are currently lacking in research. Therefore, research related to predictive models in developing countries based on demographic variables and academic data is still urgent.

In addition to focusing on the context of developing countries, the investigation of predictive models for student graduation time should also focus on the context of open higher education institutions that provide online and distance education. Most existing studies (e.g., Aiken, De Bin, Hjorth-Jensen, & Caballero, 2020; Moraga-Pumarino...
et al., 2023; Suhaimi et al., 2019; Witteveen & Attewell, 2021) investigating predictive models for student academic success have centered on the context of regular higher education institutions that provide almost exclusively face-to-face on-campus learning activities. Few studies have investigated models that predict students’ academic success in open higher education institutions, especially when it is associated with the time it takes them to complete their studies. Studies on predictive models of student academic success in open higher education institutions that we found focused more on academic outcomes such as GPA in terms of various variables, of which the time of access to learning materials available in the learning management system, study time, and professional status of students are the three most important variables in predicting student academic achievement (Purwoningsih, Santoso, Puspitasari, & Hasibuan, 2021).

Given that open higher education has its own unique characteristics that distinguish it from regular higher education and that different higher education institutions have different student populations, student services, styles of instruction, and degree programs (Aiken et al., 2020), models that predict student academic success at open higher education institutions in one country in terms of time to graduation are likely to differ from those at regular higher education institutions and from open higher education institutions in other countries.

The current study, in general, aims to test the DM algorithms on the prediction model of student's study status in terms of their time of graduation and whether their graduation time is still reasonable. Specifically, this study aims to identify the most critical factors (attributes) that explain students’ time to graduation in developing countries such as Indonesia and in open higher education institutions such as Universitas Terbuka or Indonesia Open University. Two research questions address these general and specific objectives: (1) what kind of prediction model with the DM algorithm is best suited for predicting students’ time to graduation in developing countries such as Indonesia and in an open higher education institution in terms of DM algorithm performance indicators? and (2) what are the most critical factors (attributes) that can explain students’ time to graduation in developing countries like Indonesia and in an open higher education institution based on the most suitable prediction model? It is hoped that answering these two research questions can contribute to the advance of current research which focuses on using demographic information and academic data stored in the university database to predict students’ academic success in terms of their time to graduation in developing countries and open higher education institutions.

1 Methods

1.1 Type of the Study

The current study used DM method. There are two reasons to explain the use of the DM method. First, the DM method is able to create a predictive model by analyzing data in the database (predictive model; Yağcı, 2022). Second, the DM method can describe behavior (descriptive model; Yağcı, 2022). These two reasons are strongly related to the research’s objectives, which intend to create a prediction model and identify most critical factors (attributes) that can determine students’ academic success based on their time to graduation or to complete their degree.

1.2 Data Set

The data set used in this study was obtained from an open higher education institution, which regularly keeps all student data in electronic form. These data can be of various types and volumes, from student demographics to academic achievement. This study collected data from an Academic Information System (AIS) database of the Universitas Terbuka (UT) (or Open University), a public university that is the only university in Indonesia that organizes open and distance learning. The data contain demographic and academic records of 132,734 students who started their studies in the academic year of 2014 (semester 1) to 2019 (semester 1). The data set we obtained is presented in a spreadsheet format which contains information that students input when registering and student academic achievements which include: (1) student ID; (2) name; (3) place and date of birth; (4) age; (5) address or place of residence of the student; (6) gender; (7) marital status; (8) semester and year of the student’s initial and final registration; (9) employment or occupational status; (10) study program; (11) GPA or cumulative GPA; (12) number of credits taken; (13) graduation status; and (14) code of the Unit Program Belajar Jarak jauh Universitas Terbuka (UPBJJ-UT) (Open University Distance Learning Program Unit) that is primarily responsible for student learning.

1.3 Measures

1.3.1 Predictor Variables

There are two criteria used to guide the selection of independent variables in this study, namely (1) the variable

...
must be available in the AIS database; (2) the variable is a potential, predominant, or significant predictor in predicting the academic success of students based on previous studies (Alyahyan & Düşteğör, 2020; Purwoningsih et al., 2021; Suhaimi et al., 2019). Based on these two criteria, we have successfully identified six predictor variables; they are age, gender, marital status, address or place of residence of the student, cumulative GPA, and student employment or occupational status. In order to operationalize the student address or place of residence variable, we categorized the variable into three regional groups based on time zones in Indonesia, namely, West, Central, and East. Given two criteria, seven variables were selected as independent variables. Furthermore, given previous studies showing that family income affects students’ academic success (Alyahyan & Düşteğör, 2020; Suhaimi et al., 2019; Yildiz & Börekci, 2020), we considered using family income as a predictor variable. However, our data set does not provide information on this. As an alternative to this issue, we decided to use the Regional/Provincial Minimum Wage (UMR/UMP) information that we obtained from Statistics Indonesia (Badan Pusat Statistik (BPS), the Central Bureau of Statistics) to represent the family income variable. We assigned the amount of family income of the student based on the UMR/UMMP according to the student’s address or place of residence in the year when the student made the initial registration. Thus, a total of seven predictor variables were used in this study to predict students’ academic success based on the time it took them to complete their studies or their time to graduation. These seven variables were grouped into demographic factors and academic data. Table 1 shows a description of each independent or predictor variable.

### Table 1: Description of independent variables

<table>
<thead>
<tr>
<th>Category</th>
<th>Variable</th>
<th>Description</th>
<th>Type</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demographics</td>
<td>Age</td>
<td>Student’s age</td>
<td>Numerical</td>
<td>≥20 years old</td>
</tr>
<tr>
<td></td>
<td>Gender</td>
<td>Student’s gender</td>
<td>Categorical (Binary)</td>
<td>0 = Female; 1 = Male</td>
</tr>
<tr>
<td></td>
<td>Marital</td>
<td>Student’s marital status</td>
<td>Categorical (Binary)</td>
<td>0 = Single; 1 = Married</td>
</tr>
<tr>
<td></td>
<td>Employ</td>
<td>Student’s employment status</td>
<td>Categorical (Multi)</td>
<td>0 = Others; 1 = Civil servant teacher; 2 = Private school teacher; 3 = Honorary teacher; 4 = Not identified; 5 = Not a teacher</td>
</tr>
<tr>
<td>Region</td>
<td>Student place of residence categorized by time zone in Indonesia</td>
<td>Categorical (Multi)</td>
<td>1 = IDN-East; 2 = IDN-West; 3 = IDN-Central</td>
<td></td>
</tr>
<tr>
<td>Minimum wage</td>
<td>Regional/provincial minimum wage that applies in the place of residence of the student</td>
<td>Numerical</td>
<td>≥910,000 (IDR)</td>
<td></td>
</tr>
<tr>
<td>Academic data</td>
<td>GPA</td>
<td>Cumulative grade average point (GPA)</td>
<td>Numerical</td>
<td>1.98–4.00</td>
</tr>
</tbody>
</table>

#### 1.3.2 Dependent Variable

The time that students take to complete their studies or degree or time to graduation was used as the dependent variable in this study. The time taken by students to complete their degree ranges from 8 to 17 semesters. Time to graduation is categorized into binary data. Given that the time it takes students to complete a typical undergraduate or bachelor’s degree is 8–10 semesters, we further assigned a code of 1 (successful) for students who were able to complete their studies in a maximum of 10 semesters, while for those who took more than 10 semesters to complete their studies, we assigned a code of 0 (unsuccessful). We determined the number of semesters that students need to complete their study which represents the time to graduation based on the semester and year in which students have their initial and final registration. Since the main focus of this study is on the time to graduation of students, the data we used in this study are data on students whose graduation status is “alumni.”

### 1.4 Data Analysis

#### 1.4.1 Building the Model and Implementing the DM Algorithm

The model to be built in this study is a prediction model of students’ study status based on time to graduation (i.e., successful or unsuccessful) using the R packages: “psych” (Revelle, 2023), “caret” (Kuhn, 2008), and “DALEX” (Biecek,
This model was developed by analyzing the training data and was validated using 10-fold cross-validation. Five DM models (i.e., random forest (RF), neural networks (NN), logistic regression (LR), Naïve Bayes (NB), and k-nearest neighbor (kNN) are applied in this study to predict student time to graduation. These five models have better efficiency and accuracy when used for prediction purposes (Delen, 2010; Huang & Fang, 2013; Nandeshwar, Menzies, & Nelson, 2011; Vandamme, Meskens, & Superby, 2007).

1.4.1.1 RF
RF is a supervised DM algorithm. This algorithm builds several decision trees and combines them. In addition, this algorithm uses the voting mechanism of several decision trees to improve the limitations of a single decision tree and obtain more accurate predictions (Biecek, 2018; Hannaford, Cheng, & Kunes-Connell, 2021). Each tree in the RF provides a class prediction, and the class with the highest votes becomes the model’s prediction. One of the main advantages of RF compared to other techniques is its ability to tackle or at least reduce overfitting (Cruz-Jesus et al., 2020).

1.4.1.2 NN
NN are a series of DM algorithms that endeavor to recognize the underlying relationships in a data set by imitating the information processing of the human brain (Alturki et al., 2022; Cruz-Jesus et al., 2020). The process occurs in two phases. First, the network is trained with paired data to determine the input–output mapping. Then, the weights of the connections between neurons are fixed, and the network is used to determine the classification of a new data set. The main problem when using NN is that the final model is a black box consisting only of weights on the connections between neurons (Cruz-Jesus et al., 2020). Therefore, the readability of the model is very limited.

1.4.1.3 LR
LR is a supervised DM algorithm. This algorithm attempts to differentiate between classes (categories) by examining the relationship between existing independent features (Alturki et al., 2022; Yaşcı, 2022). In this study, binary logistic regression is used when the dependent feature has only two possible outcomes, and multinomial logistic regression, where the dependent feature has three possible outcomes (Boehmke & Greenwell, 2019).

1.4.1.4 NB
NB is a supervised DM algorithm. This algorithm assumes that the features are mutually independent (Alturki et al., 2022). This algorithm is based on Bayes’ theorem, which states that if event B has occurred, then we can find the probability of event A, and is represented as follows:

\[ P(A|B) = \frac{P(B|A) \times P(A)}{P(B)} \]

1.4.1.5 kNN
kNN is a supervised DM algorithm. This algorithm estimates the probability that a data point will belong to a group based on measuring the distance between the classification example and the closest training example in the feature space (Cruz-Jesus et al., 2020). This algorithm is very simple to implement, and its performance is determined solely by choosing k parameters (Cruz-Jesus et al., 2020). Using the “caret” package (Kuhn, 2008), in this study we set \( k = 5 \), which is the default value of the number of neighbors under the package.

1.4.2 Evaluation Criteria
Train and test data were used to evaluate the model’s effectiveness in predicting students’ study status in terms of time to graduation. The process of evaluating the prediction model is carried out in two steps: the learning step on the data train and the prediction or classification step on the test (unseen data). In the prediction step, the model classifies the invisible data into one of the following classes: true-positive (TP), false-positive (FP), true-negative (TN), and false-negative (FN), as shown in the confusion matrix (Table 2). Based on the data from the confusion matrix, four metrics (i.e., classification accuracy, F-Score, precision, and recall) are used to evaluate the efficiency of the prediction model.

1.4.2.1 Classification Accuracy (CA)
CA is the ratio of correctly predicted observations (TP + TN) to the total number of observations (TP + TN + FP + FN).

<table>
<thead>
<tr>
<th>Actual</th>
<th>Negative (0)</th>
<th>Predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Negative (0)</td>
<td>Positive (1)</td>
</tr>
<tr>
<td>TN</td>
<td>FN</td>
<td></td>
</tr>
<tr>
<td>Positive (1)</td>
<td>FP</td>
<td>TP</td>
</tr>
</tbody>
</table>
Accuracy is interesting if TP and TN are more important than FP and FN. However, accuracy may not be very informative when one of the binary categories is predominant (so-called unbalanced label; Biecek & Burzykowski, 2021; Yağcı, 2022). For example, if the test data contain 90% success, a model that always predicts success will achieve an accuracy of 0.9, although it is arguable that this is not a very useful model. There are situations where FP or FN might be more of a concern. Therefore, other measures focused on spurious results (FP and FN) may be interesting. CA is formulated as follows:

\[ CA = \frac{TP + TN}{TP + TN + FP + FN}. \]

1.4.2.2 Precision

Precision represents the ratio of correctly predicted observations (TP) to the total predicted positive observations (TP + FP). The precision value is in the \([0,1]\) (Yağcı, 2022). The precision is high if the sum of FP is low. Therefore, precision is useful when the penalties for committing Type I errors (FP) are high (Biecek & Burzykowski, 2021). Precision is formulated as follows:

\[ \text{Precision} = \frac{TP}{TP + FP}. \]

1.4.2.3 Recall

Recall represents the ratio of correctly predicted positive observations (TP) to the total observations in the actual class (TP + FN). Recall values are in the \([0,1]\) (Yağcı, 2022). Recall is high if the number of FN is low. Therefore, recall becomes useful when penalties for committing Type II errors (FN) are high (Biecek & Burzykowski, 2021). Recall is formulated as follows:

\[ \text{Recall} = \frac{TP}{TP + FN}. \]

1.4.2.4 F-Score (F1)

F1 tends to give low scores if both precision and recall are low, and high scores if both precision and recall are high. For example, if the precision is 0, then the F1 will also be 0 regardless of the recall value. Therefore, it is a useful measure if we have to find a balance between precision and recall. F1 is formulated as follows:

\[ F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2TP}{2TP + FP + FN}. \]

Besides using the information in the confusion matrix table, the model’s effectiveness for predicting student study status in terms of time to graduation is evaluated based on the area under the curve (AUC) information from the receiver operating characteristics (ROC) curve. The AUC–ROC curve is used to evaluate the performance of a classification problem. AUC–ROC is a widely used metric to evaluate the performance of machine learning (ML) algorithms, especially in cases where there are unbalanced data sets (Yağcı, 2022), and describes how well the model executes predictions (Wang, King, & Leung, 2023). The wider the area covered, the better the ML algorithm discriminates between the given classes. AUC has an ideal value of 1.

2 Results

2.1 Descriptive Statistics

Table 3 presents descriptive statistics which contain information on mean, standard deviations, and student proportions based on demographic and academic data. From 2018 to 2022, it was recorded that 74% of students succeeded in completing their studies, and the remaining 26% were declared unsuccessful, with a mean GPA of 2.95. The mean age of these students is 34 years, and most are female. Most of them come from the western part of Indonesia (79%) with single status (55%) and work as teachers (69%). The mean minimum wage for jobs in their work areas is less than IDR 2,000,000.

<table>
<thead>
<tr>
<th>Category</th>
<th>Feature</th>
<th>Mean (SD) or proportions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demographics</td>
<td>Age</td>
<td>34 (8.65)</td>
</tr>
<tr>
<td></td>
<td>Gender</td>
<td>Female: 74%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Male: 26%</td>
</tr>
<tr>
<td></td>
<td>Marital</td>
<td>Married: 45%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Single: 55%</td>
</tr>
<tr>
<td></td>
<td>Employ</td>
<td>Civil servant teacher: 7%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Private school teacher: 38%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Honorary teacher: 24%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Not a teacher: 26%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Others: 4%</td>
</tr>
<tr>
<td></td>
<td>Region</td>
<td>IDN-West: 79%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>IDN-East: 2%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>IDN-Central: 19%</td>
</tr>
<tr>
<td>Academic Data</td>
<td>Minimum wage</td>
<td>IDR 1,692,812 (IDR 549,774)</td>
</tr>
<tr>
<td></td>
<td>Study duration</td>
<td>≤10 Semester: 74%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>&gt;10 Semester: 26%</td>
</tr>
</tbody>
</table>
2.2 Model Evaluation

2.2.1 Confusion Matrix

The confusion matrix reflects the current situation in the data set and the number of correct/false predictions from the model. Model performance is calculated based on the number of observations classified correctly and those classified incorrectly. The rows show the actual number of samples in the test set, while the columns represent estimates from the model. In Table 2, TP and TN show the number of observations predicted as 1 (positive) when they should be in class 0 (negative). FN indicates the number of observations predicted as 0 (negative) when they should be in class 1 (positive).

Table 4 shows the confusion matrix for the five DM algorithms. In a confusion matrix for each algorithm with dimensions of $2 \times 2$, the main diagonal denotes the percentage of instances predicted correctly, and elements other than the main diagonal denote the percentage of prediction errors. Table 4 shows that based on the RF algorithm, 65.6% of those whose actual study period was less than or equal to 10 semesters (successful) and 10.7% of those with more than 10 semesters (not successful) were predicted correctly. Based on the NN algorithm, 67.3% of those whose actual study period was less than or equal to 10 semesters (successful) and 8.8% of those with more than ten semesters (unsuccessful) were predicted correctly. Furthermore, based on the NB algorithm, 41.6% of those whose actual study period was less than or equal to 10 semesters (successful) and 22.2% of those with more than ten semesters (unsuccessful) were predicted correctly. Lastly, based on the kNN algorithm, 64.9% of those whose actual study period was less than or equal to 10 semesters (successful) and 10.8% of those with more than 10 semesters (unsuccessful) were predicted correctly.

2.2.2 ROC

Performance comparison between the proposed models based on the AUC, CA, $F_1$, precision, and recall metrics is presented in Table 5. All proposed models have more than 75% accuracy or are in the high category, except for the NB model. It means a high correlation exists between the predicted and actual data in all proposed models (except NB). Specifically, the RF model achieves the best results compared to other models regarding precision metrics, with percentages of 76.3 and 81.1 on invisible data (test). It can be interpreted that the RF model can predict students who complete their studies effectively with low type I (false positive) prediction errors. On the other hand, the NN model achieves the best results compared to other models in terms of recall metrics, with respective percentages of 76.3 and 81.1 on invisible data (test). It can be interpreted that the NN model is comparatively effective in predicting the possibility of students completing their studies with low type II errors (false negatives). Overall, the RF and NN models have almost the same values (not significantly different) when viewed from the CA, $F_1$, precision, and recall metrics.
recall metrics, so it can be said that these two models have the best performance in predicting student study status based on time to graduation compared to other models.

The AUC–ROC metric was used for further evaluation of the performance of the ML algorithm, especially in the case of imbalanced data. Table 5 shows that the RF and NN models have almost the same AUC values (not significantly different) and are the highest compared to the other models, with respective percentages of 79.5 and 79.3. The same results are also shown in Figure 1. Thus, the RF and NN models have the best performance based on the AUC metric in predicting student study status based on time to graduation compared to other models.

### 2.3 Model Explanation

#### 2.3.1 Variable Importance

Variable importance is a measure of the extent to which certain variables contribute to the predictions of the dependent variable. Figure 2 shows the variable importance of the best models, namely RF and NN. In the RF model, demographic variables such as gender, marital status, and region seem to have a less significant contribution in explaining student time to graduation. Demographic variables such as age, employment status, and minimum wage as well as academic data such as GPA seem to be the most important variables in explaining student study status based on time to graduation. These four variables have a greater contribution than the other three variables. In the NN model, demographic variables such as gender and region are also variables with a small contribution in explaining the success of student studies based on their time to graduation, coupled with the minimum wage variable which in the RF model is actually the most important variable. Demographic variables

![Feature Importance](image1)

**Figure 2:** Variable importance.
such as age and employment or occupational status and academic data variables such as GPA seem to consistently be the most important variables in the NN model in explaining student study status based on time to graduation. In addition to these variables, the marital status variable, which previously had a not-so-large contribution to the RF model, became the most important variable in explaining student time to graduation. Overall, all predictor variables (i.e., demographics and academic data) used in the RF and NN models have a contribution (we found no variables with negative RMSE values) with different magnitudes in explaining student study status in terms of time to graduation.

2.3.2 Partial Dependence

Partial dependence is used to calculate the dependence of the final prediction on a certain variable. Two sets of variables were examined, i.e., demographic information and academic data. This demographic information is related to the unique attributes of students such as gender, age, marital status, employment, region, and minimum wage. Figure 3 depicts that the RF and NN models predict a high probability of success for students who are female, single, and do not have a permanent job to complete their studies in no more than 10 semesters.

Figure 3: Partial dependence profile for demographics information.
Students from regions with a minimum wage of more than IDR 3,000,000 also demonstrated a high probability of success in completing their studies. However, the RF and NN models have different predictions in terms of age. The RF predicts a high probability of success for students aged 30–50 to complete their studies, while those over 50 have a lower probability of success in completing their studies. It is different from the NN predicting students over 50 years of age have a higher probability of completing their studies in no more than 10 semesters.

The next focus is on academic data relating to prior students' academic grades as represented by GPA. The dependence plot of the GPA variable is presented in Figure 4. The figure demonstrates that the RF and NN models predict a high probability of success for students who have a GPA of more than 3.0 to complete their studies in no more than 10 semesters.

3 Discussion

The current study proposed a new model to predict student study status in terms of time to graduation in developing countries such as Indonesia and in an open higher education institution based on the DM algorithm, taking data on demographics, GPA, and student study duration as data sources. The performance of the RF, NN, LR, NB, and kNN algorithms was examined and compared to predict student time to graduation. In comparing the performance of the algorithms, two parameters are used. The first parameter is the prediction of student study status in terms of time to graduation based on demographic information and prior cumulative GPA. Yağcı (2022) also used the same parameter in predicting student final examination grades. The second parameter is a comparison of ML algorithm performance indicators based on AUC, CA, F1, recall, and precision metrics. This parameter is carried out especially on imbalanced data sets (Biecek & Burzykowski, 2021). Apart from comparing the performance of the ML algorithms, this research also demonstrated important variables that contribute to predicting student time to graduation.

The results of this study indicate that the proposed model achieves a classification accuracy of 60–76%. RF, NN, LR, and kNN are algorithms with a very high level of classification accuracy that can be used to predict student study status based on time to graduation, while NB has a low classification accuracy. The results of this study are in line with the results of previous studies (Musso et al., 2020; Yağcı, 2022) which showed that the RF, NN, LR, NB, and kNN algorithms have very high classification accuracy in predicting student academic performance and study status. Similar results were also found by Alturki et al. (2022) that LR, RF, kNN, and NB have very high accuracy in predicting the academic performance of doctoral students. In addition, Waheed et al. (2020) also found that the performance of the LR, NN, and RF algorithms have the highest accuracy in identifying students at high risk of academic failure based on their demographic characteristics.

One of the important findings from this study is that the RF and NN algorithms have the highest classification accuracy and AUC compared to the other three proposed algorithms (such as LR, kNN, and NB). Based on these findings, it can be said that the RF and NN algorithms have more accurate results in predicting student study status using ML algorithms. The results of this research are in line with the results of previous research which also showed the performance of the RF and NN algorithms as the best compared to the LR, kNN, and NB algorithms in predicting student final exam grades using only three academic data attributes (i.e., midterm exam grades, department data, and faculty data), even the results of the two are very similar (Yağcı, 2022). Sassirekha and Vijayalakshmi (2022) also demonstrated that the performance of the RF algorithm was the best with an accuracy of up to 90% in predicting student academic progress compared to the LR, kNN, and NB algorithms. Xu et al. (2019) in their study found that the highest accuracy obtained through the NN algorithm compared to RF in predicting the academic performance of undergraduate students, with accuracy that is not much different.
In addition to finding a model that can accurately predict the time to graduation of students, the results of this study also identified the variables that provide the greatest increase in model performance. These variables were identified using the RF and NN algorithms which demonstrated the best performance. Based on these results, it can be said that demographic variables (i.e., gender, age, marital status, employment, region, and minimum wage) and academic variables (i.e., GPA) are important predictors of student study status. The results of this research are in line with the results of previous studies which show that demographic information such as wage, age, occupation, place of residence (Costa-Mendes, Oliveira, Castelli, & Cruz-Jesus, 2021; Cruz-Jesus et al., 2020), and prior academic achievements (Hoffait & Schyns, 2017) are important variables that can be used to predict academic achievement scores and student study status. In addition, the results of this study also succeeded in identifying the most important variable in predicting student study status, namely prior academic achievement (i.e., GPA). This result is in line with the findings of Yağcı (2022) and Hannaford et al. (2021), showing that prior student academic achievement has the most important contribution in predicting student study status based on their time to graduation. Furthermore, Suhaimi et al. (2019) have also noted that many studies have demonstrated GPA as an important predictor of students’ time to graduation. This is possible because when students get a low GPA, especially in the early days of their studies, they have to repeat the course they got a low grade in the following semester especially when the course is a prerequisite, and this will certainly prolong their graduation time.

Our study has reported that the most important predictor of time to graduation for students at the Open University is GPA. In addition, taking into account various factors that may affect the time to graduation, one of which is age, our study revealed that age is also considered an important predictor. Age is one of the factors that should be considered as a predictor of students’ graduation time given that the context of our current study is an Open University that embraces the educational paradigm of “life-long learning.” By adhering to this educational paradigm, on the one hand, anyone can become an Open University student regardless of their age as long as they have a high school diploma or its equivalent, and this certainly provides a wide opportunity for everyone to be able to pursue higher education. On the other hand, the wide age range of students may make it challenging to provide relevant learning facilities and support to each student to encourage them to complete their studies in a reasonable time. The wide age range of students at the Open University also has consequences for the high diversity of students’ employment status or study orientation. At the Open University, relatively young students generally expect to complete their studies as quickly as possible so that they can apply for jobs with their diplomas. Meanwhile, older students tend to have relatively stable jobs, so their studies are more focused on gaining knowledge and motivating their children to emulate them in terms of their enthusiasm for learning, and thus, their studies are less focused on an orientation to graduate as quickly as possible. The findings of a study by Sánchez-Gelabert, Valente, and Duart (2020) demonstrate similarly that when it comes to the online university context, older students tend to focus their studies on opportunities to improve theoretical knowledge or the acquisition of new knowledge and concepts. Meanwhile, younger students tend to focus their studies more on improving their chances of getting a job, developing their careers, and increasing practical knowledge that is highly beneficial to their work (Sánchez-Gelabert et al., 2020).

In addition to the age factor, our study has reported that based on RF and NN algorithms, minimum wage and employment status, respectively, become important factors after GPA in predicting students’ graduation time in Open University. It has been pointed out that those with a higher minimum wage and no job, who can focus more on their studies, tend to have a higher chance of completing their studies within a reasonable time. Meanwhile, those with a lower minimum wage or who are working (so-called working students), so studies may not be their main focus, will tend to take a longer time to graduate because they will be more likely to take fewer courses or credits each semester. At the Open University, students are facilitated to take a few courses or credits each semester, even taking just one course is allowed and facilitated. The study conducted by Ecton, Heinrich, and Carruthers (2023) showed that working students tend to take fewer courses or credits which results in a lower percentage chance of graduating in a reasonable time compared to non-working students. Even though working students take a longer time to graduate, their performance in terms of GPA still shows similar results or is not much different from non-working students who can complete their studies faster or in a reasonable time (Ecton et al., 2023).

Overall, the RF and NN models, as the best ones proposed in this study, are able to predict student study status based on time to graduation with an accuracy of 76%. Accordingly, it can be said that the student’s time to graduation can be predicted with these models in the future. Predictions on a student’s future study status in terms of time to graduation provide an opportunity for students to evaluate the work methods they use in order to improve their academic performance (Yağcı, 2022). A similar finding was also demonstrated by Bernacki, Chavez, and Uesbeck (2020).
where their early warning prediction model based on malleable behavior managed to correctly identify 75% of students who were not successful in a course or could not fulfill the prerequisites for more advanced courses. In addition, Burgos et al. (2018), through a prediction model that they built via a tutoring action plan, managed to reduce the student academic failure rate by 14% from the previous year. Furthermore, the most important variables in predicting academic achievement that have been identified can be useful for policymakers and other stakeholders to identify appropriate and most cost-effective interventions (Wang et al., 2023). Since age and cumulative GPA are the most important variables related to the teaching and learning process in predicting student graduation time, improving the quality of student learning assistance through weekly tutorials, learning resources or modules, and open educational resources/massive open online courses tailored to the diverse needs and characteristics of students such as age and occupational is an important thing to do by open higher education institutions. Quality improvements in these areas are expected to increase students’ GPA and cumulative GPA by increasing their engagement in available learning services and their motivation to learn and subsequently provide equal opportunities for every student to complete their studies or degree in a reasonable time regardless of their age, occupation, or region of residence.

4 Conclusion

This study contributes to current developments that utilize demographic information and GPA recorded in university databases to predict student study status in terms of student time to graduation in developing countries such as Indonesia. In terms of methodology, this study contributes to providing additional insight into predicting students’ time to graduation by utilizing ML approaches or algorithms such as RF, NN, logistic regression (LR), NB, and kNN. In terms of obtained results, this study contributes to complementing previous literature, which focused more on comparing the accuracy of ML models and did not highlight critical factors based on demographics information and student GPA, which could later be used as an intervention basis to treat students who have the potential to fail to complete their studies within a reasonable time.

The results of this study indicate that the RF and NN models have the highest classification accuracy and AUC values (CA: 76% and AUC: 79%) compared to other models, such as LR, NB, and kNN, in predicting students’ study status based on their time to graduation. The NB model is the lowest, with a classification accuracy of 64% and an AUC of 78%. The RF and NN models were able to identify one of the most critical variables in predicting students’ time to graduation along with six other important variables, namely, the student’s GPA. In conclusion, the RF and NN models proposed in this research can be used to predict students’ time to graduation with high accuracy, and policymakers and other stakeholders can use the most critical variables identified as intervention materials to deal with students who have the potential to fail to complete their studies within a reasonable time in the future. Furthermore, given that this study only focused on five algorithms (i.e., RM, NN, LR, NB, and kNN) in constructing the model to predict students’ time to graduation, while a number of studies have demonstrated that other algorithms such as xgboost (Aiken et al., 2020) have very promising performance these days, we suggest future studies to consider such algorithm to be used in predicting students’ time to graduation and compare their performance with the five algorithms. In addition, while our study used the default value available in the “caret” package for the $k$ value in the kNN algorithm, we expect future studies to use $k$ values other than the default from the package to predict students’ time to graduation.

Funding information: This work was financially supported by the Institute for Research and Community Service (or Lembaga Penelitian dan Pengabdian kepada Masyarakat, LPPM), Universitas Terbuka, Indonesia, through Applied Research scheme under the Annual Work Plan and Budget for fiscal year 2023 (grant number: B/549/UN31.LPPM/PT.01.03/2023).

Conflict of interest: The authors state no conflict of interest.

References


