Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM)
Vol. 8, No. 1, March 2025, pp. 161 – 173
p-ISSN: 2614-3372 | e-ISSN: 2614-6150

Predicting Student On-Time Graduation Using Particle Swarm Optimization and Random Forest Algorithms

[1]Arif Rahman, [2]Deni Mahdiana, [3]Achmad Fauzi
[1,2,3]Master's Program in Computer Science, Faculty of Information Technology,
Universitas Budi Luhur, Indonesia.
Email: [1]211601246@student.budiluhur.ac.id,
[2]deni.mahdiana@budiluhur.ac.id, [3]2211600826@student.budiluhur.ac.id

Article Info
Article history:
Received Oct 21th, 2024
Revised Dec 24th, 2024
Accepted Jan 14th, 2024

Keyword:
Classification
Feature Selection
Particle Swarm Optimization
Random Forest
Student Graduation

ABSTRACT
Higher education plays a crucial role in human resource development and national progress. A key indicator of educational quality is students' ability to graduate on time. Delays in graduation can lower the quality of higher education. Various academic and non-academic factors influence timely graduation rates. At Universitas Islam Syekh Yusuf, the trend of students graduating beyond the expected timeframe has risen over the past three years. However, the university lacks insight into the factors contributing to these delays. This research aims to identify factors causing delayed graduation using PSO and Random Forest to predict student graduation outcomes. The application of PSO reveals key factors influencing timely graduation, including study program, student active status, student leave of absence status, inactive status for semester 1, GPA1, and credit hours in semesters 1 and 2. Evaluation results show that using PSO and Random Forest to predict timely graduation achieves high accuracy (99.63%), precision (99.77%), recall (99.65%), and F1 score (99.71%).

Corresponding Author:
Arif Rahman,
Master's Program in Computer Science, Faculty of Information Technology,
Universitas Budi Luhur, Jakarta, Indonesia.
Jl. Ciledug Raya, RT.10/RW.2, Petukangan Utara, Kec. Pesanggrahan, Kota Jakarta Selatan,
Daerah Khusus Ibukota Jakarta 12260.
Email: 2211601246@student.budiluhur.ac.id
DOI: http://dx.doi.org/10.24014/ijaidm.v8i1.33577

1. INTRODUCTION

Higher education institutions play a crucial role as providers of academic education for students and hold significant importance in the development of human resources and the advancement of a nation [1]. One of the key indicators of the success or quality of higher education is significantly influenced by students' ability to complete their studies within the designated timeframe [2]. As the number of students enrolling in higher education continues to increase, it is essential to ensure that an equivalent proportion of students successfully complete their studies on time [3]. Delays or extended time for students to graduate result in a decline in the quality of education provided by the institution [4]. In recent years, there has been a decline in the number of students graduating on time at Syekh Yusuf Islamic University, Tangerang. For the 2017 cohort, the percentage of on-time graduation was 65.13%, with 902 out of 1,385 students completing their studies within the designated period. This figure dropped to 63.94% for the 2018 cohort, with 1,018 students graduating on time out of a total of 1,592. The decline continued with the 2019 cohort, where the percentage fell to 61.64%, with 908 students graduating on time out of a total of 1,473 students.

To extract valuable insights from a set of available data, data analysis methods can be applied, particularly data mining techniques, to uncover classification patterns. This enables the creation of predictions based on the information derived from the analyzed data.

Based on previous research related to on-time graduation prediction, various algorithms have been used, such as the C4.5 Decision Tree, a popular and effective method in classification and prediction, used a limited dataset from the 2013 cohort with only 544 data points, resulting in a relatively low accuracy of 60%

[5], The Support Vector Machine (SVM) algorithm achieved 94.4% accuracy in predicting student graduation. While effective on high-dimensional datasets, SVM can be slow and computationally intensive, especially with large datasets due to complex optimization [3], Another study used the Naive Bayes algorithm and k-Fold Cross Validation technique to objectively assess the model's accuracy and reduce overfitting, achieving an accuracy of 80.19% in predicting student graduation. However, this study may not have conducted a thorough analysis of the factors influencing graduation [6], This study shows that the Random Forest and K-Nearest Neighbor (K-NN) algorithms optimized with Particle Swarm Optimization (PSO) achieved high accuracies of 97.89% and 96.74%, respectively, in predicting on-time graduation. However, there are limitations related to data quality, model complexity, and the ability to generalize the results.

The Decision Tree excels in processing numerical data but falls short in generalizing new cases [8]. SVM offers excellent performance in data classification but faces challenges in parameter selection [4]. Naïve Bayes is efficient in classification; however, its assumption of attribute independence is often unrealistic [6]. In previous studies, Random Forest demonstrated superior accuracy. As a Homogeneous Ensemble Learning algorithm, Random Forest offers low error rates and efficiency in handling large datasets [9], [10]. Higher education institutions face challenges in identifying the factors contributing to delays in student graduation and strive to understand and address these issues. PSO has emerged as an effective feature selection method, helping institutions identify the factors that influence on-time graduation [11], [12].

Although various algorithms have been used to predict on-time graduation, most studies have not employed feature selection methods to identify key factors that significantly influence on-time graduation, focusing only on optimizing classification algorithms. Existing algorithms often use general data without exploring specific factors for optimization. By applying PSO, this study identifies the most relevant features impacting on-time graduation, providing deeper insights into educational institutions

Thus, this study is expected to demonstrate that applying Particle Swarm Optimization can assist higher education institutions in identifying the factors causing students to graduate late. Additionally, implementing the Random Forest algorithm can enhance the prediction or classification of student graduation, resulting in more accurate outcomes. This enables institutions to take more effective measures to improve the on-time graduation rates of their students.

2. MATERIAL AND METHOD

In order to facilitate the research, the use of the Cross Industry Standard Process for Data Mining (CRISP-DM) methodology has been implemented. This research process will be divided into six stages, as shown in Figure 1, which include Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. Below is an explanation of the activities to be carried out at each of these stages.

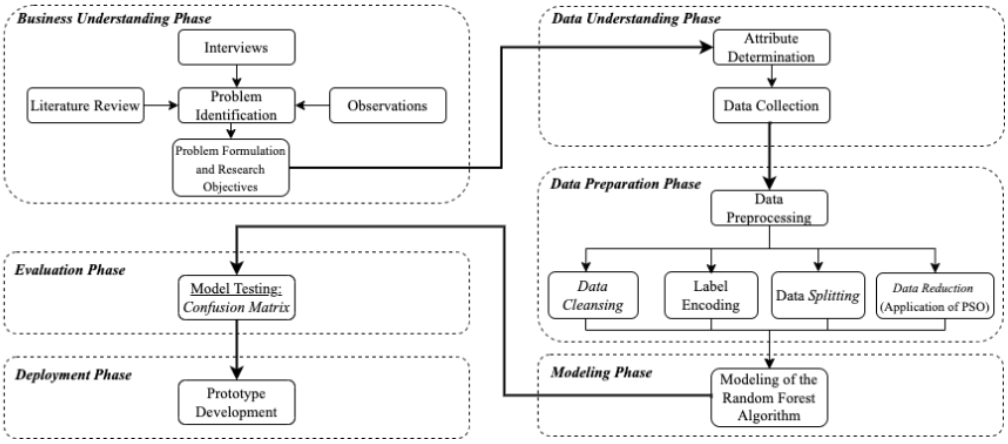

Figure 1. Research Methodology Steps

2.1. Data Mining

Data mining is an area with practical relevance in applying concepts to real-world situations, not just in theoretical contexts [13]. According to [14], data mining is the process of discovering patterns or interesting information in selected data using specific techniques or methods. The selection of these techniques must align with the overall goals and processes of Knowledge Discovery in Database (KDD). Data mining plays a crucial role in extracting valuable information from large datasets and is now widely used in various fields such as medicine, engineering, business, and education to uncover important information hidden in historical data for future decision-making [15].

The Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology is a model that outlines the stages in a data mining project cycle. The following are the stages, as shown in Figure 2.

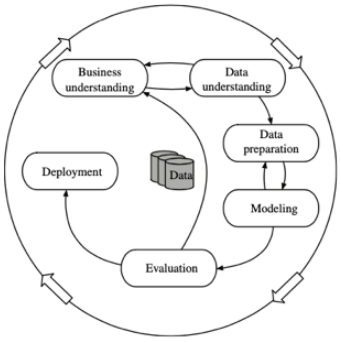

Figure 2. CRISP-DM [13]

1. Business understanding: Explore business objectives, assess whether data mining can assist, and identify the data needed to build a relevant model.
2. Data understanding: Create and analyze the initial dataset to ensure processing feasibility. If data quality is insufficient, new data collection or a review of the business context may be necessary.
3. Data Preparation and Modeling: Prepare raw data for algorithm processing, often in conjunction with model creation. This process is iterative, adjusting techniques based on modeling results.
4. Evaluation: Evaluate the model to ensure optimal performance. If inadequate, revising business objectives or data strategies may be required. If accurate, the model is ready for deployment.
5. Deployment: Integrate the model into the software system, often requiring technical adjustments such as re-implementation using the appropriate programming language.

2.2. Random Forest

According to Schonlau and Zou [16], Random Forest is one of the best-performing machine learning algorithms, which combines multiple decision trees to improve accuracy and prevent overfitting. This algorithm creates a forest of decision trees, where each tree is trained based on a random subset of data and features, and the final prediction is made by combining the predictions from all the trees [17]. Random Forest, as an ensemble method based on Decision Trees, addresses overfitting by combining the results of various models through a voting mechanism on random bootstrap samples. Figure 3 shows how Random Forest operates and generates the final class through the aggregation of results from all decision trees [18].

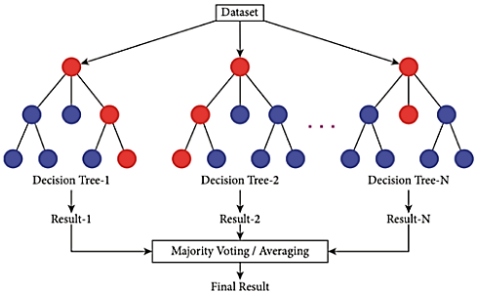

Figure 3. Classification Illustration of Random Forest

Random Forest maps the attributes of the class, allowing it to perform classification on previously unseen data. Below are the stages in testing the performance of the Random Forest algorithm [19]:

1. Observe the labels in the data, if all the labels are the same, a leaf node will be created with the label value from all the data.
2. Calculate the information value using all the available data, applying the following formula 1.

$$\text{info (D)} = -\sum_{i=1}^{m} \text{pi} * \log_2 (\text{pi}) \quad (1)$$

The above formula represents the probability that a tuple in dataset D will belong to a certain class, assuming that the entropy of D is the average amount of information needed to identify a tuple in D.

If the value of A is discrete, then dataset D will be partitioned based on the number of values that A has, so that each branch will have a homogeneous class. After the first branching, the number of possible branches is measured using the formula in Formula 2.

$$\text{info A(D)} = \sum_{v}^{e} \frac{|Dj|}{|D|} * \text{info A(Dj)} \quad (2)$$

3. The information value is calculated using the following formula
4. Attributes are evaluated based on their ability to form a decision tree partition weight $\frac{|Dj|}{|D|}$ which measures the information required to classify a tuple in partition A. The best partition has the smallest infoA(D). For continuous attributes, the candidate split point is determined by sorting the data, calculating the average between two consecutive data points, and then selecting the split point with the smallest infoA(D). The gain is calculated using formula (2), and the attribute with the highest gain becomes the branch of the decision tree.

$$\text{gain (A)} = \text{info (D)} - \text{info A(D)} \quad (3)$$

5. After a branch is formed in the decision tree, the calculations will continue as in steps 1 to 4. However, if the branch has reached the maximum allowed number of branches, a leaf node will be created with the majority value of the data in that branch.

2.3. Feature Selection

Feature selection is a method of knowledge discovery that helps understand problems by analyzing the most important features. The goal of feature selection is to improve the development of classifiers by identifying significant features, while also helping reduce the computational load [20]. Feature selection is applied in various fields as a tool to remove irrelevant and redundant features. It simplifies datasets by reducing their dimensions and determining important features without compromising prediction accuracy [21] ultimately producing an optimal feature set for classification task [22].

2.4. Particle Swarm Optimization

Particle Swarm Optimization (PSO) is an algorithm inspired by the behavior of insect swarms such as ants, termites, bees, or birds. PSO attempts to mimic the social interactions of these organisms. Social interactions involve individual actions and the influence of other members within the group. Essentially, particles cannot move suddenly but instead, move toward the best position based on their personal experience and the experiences of all other particles [23]. The model of particle movement is illustrated in Figure 4.

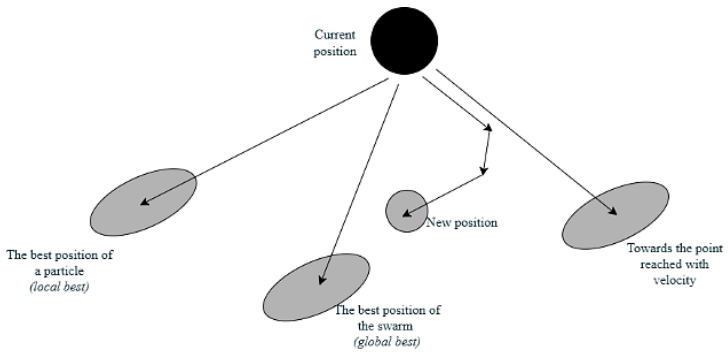

Figure 4. PSO Migration Model

The term particle in PSO refers to entities like birds in a flock, where each particle behaves independently but is influenced by the collective behavior of its group. If one particle discovers an optimal path, the others will follow it. PSO, proposed by Kennedy and Eberhart, is an optimization algorithm based on swarm intelligence aimed at finding optimal solutions in a multidimensional space, with each particle having its own position and velocity [24]. n feature selection, PSO is used to choose the optimal attributes for classification, where each particle represents a candidate solution in the search space through a binary string; a value of 1 indicates a selected component and a value of 0 means it is ignored [25]. The initialization process of particles is done randomly, and these particles move within the search space to find the optimal feature subset by updating their positions x and velocities i, as expressed by equations 4 and 5 [26]. In 1997, Kennedy

and Eberhart modified PSO into Binary Particle Swarm Optimization (BPSO), where the position of particles is represented by a binary string [25].

$$x_i = \{x_{i1}, x_{i2}, \ldots, x_{iD}\}, \tag{4}$$

Where D represents the dimension of the main search space.

$$v_i = \{v_{i1}, v_{i2}, \ldots, v_{iD}\}. \tag{5}$$

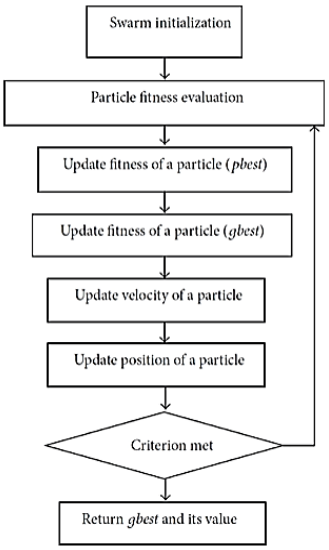

Figure 5. PSO Workflow for Feature Selection

In Figure 5 according to [26] the stages of Particle Swarm Optimization are explained as follows.

1. Swarm initialization, initializing the position and velocity of each particle randomly.

2. Particle Fitness Evaluation, Assessment of how well a particle performs in the search space.
   if fitness of x_i > $pbest_i$
   $pbest_i=x_i$
   if fitness of $pbest_i$ >$gbest_i$
   $gbest_i=pbest_i$

3. Update The Velocity of Particle i that is, update the velocity of particle i

$$v_{id}^{t+1} = w * v_{id}^{t} + c_1 * r_{1i} * (p_{id} - x_{id}^{t}) + c_2 * r_{2i} * (p_{gd} - x_{id}^{t}) \tag{6}$$

   Update the position of particle i that is, update the velocity of particle

$$x_{id}^{t+1} = x_{id}^{t} + v_{id}^{t+1} \tag{7}$$

4. If the stopping criteria are not met, continue with Steps 2 and 3.
5. Return best and its fitness values.

With the steps outlined above, the PSO algorithm continuously optimizes the position of particles in the search space to achieve the best solution based on the given objective function.

2.5. Confusion Matrix

A Confusion Matrix is a fundamental concept in machine learning used to evaluate the accuracy of a model by comparing the predictions made by the model with the actual values [27] In this context, the predicted data falls into two classes: positive class and negative class [23].

Table 1 shows a representation of the confusion matrix. True positives and false positives are abbreviated as TP and FP, while false negatives and true negatives are abbreviated as FN and TN [28]. The classification model predicts the class for each data point, providing a predicted label (positive or negative) for

each sample. Thus, at the end of the classification process, each sample is categorized into one of four situations: actual positives correctly predicted as positive, referred to as true positives (TP); actual positives incorrectly predicted as negative, referred to as false negatives (FN); actual negatives correctly predicted as negative, referred to as TN and actual negatives incorrectly predicted as positive, referred to as FP.

Table 1. Standard Confusion Matrix

	Positive Prediction	Negative Prediction
Actual Positive	True Positive TP	False Negative FN
Actual Negative	False Positive FP	True Negative TN

The confusion matrix aids in assessing the performance of the classification model by providing insights into accuracy, precision, recall, and F-Measure [28]. Below are the formulas for calculating each performance indicator.

1. Recall
   Recall is the ratio of the number of TP to the total number of positive observations. Recall measures how effectively the model identifies all positive observations. The precision measurement can be seen in Formula 8.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{8}$$

2. Precision
   Precision is the ratio of the number of TP to the total number of positive predictions. Precision measures how effectively the model correctly predicts positive outcomes. The precision measurement can be seen in Formula 9.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{9}$$

3. Accuracy
   Accuracy It is the ratio of the number of correct predictions to the total number of predictions. This metric describes how accurately the model classifies the data. The accuracy measurement can be seen in Formula 10.

$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{FP} + \text{FN} + \text{TN})} \tag{10}$$

4. F-Measure
   F-measure is the harmonic mean of precision and recall, calculated using Formula 11.

$$\text{F} - \text{ Measure} = 2\text{x}\frac{(\text{ Precision x Recall })}{(\text{ Precision} + \text{ Recall })} \tag{11}$$

3. RESULTS AND ANALYSIS

3.1. Business Understdaning

In the business understanding stage, the main focus is on identifying the problem and establishing the research objectives. Three methods are employed: literature review, interviews, and observations. The literature review involves examining various sources such as articles, journals, and books related to student graduation prediction using data mining methods, particularly PSO and Random Forest. Interviews are conducted with relevant parties such as the Academic Administration Office, Academic Advisors, and the IT Service Unit to identify the factors causing students to not graduate on time and to gather supporting data. Observations involve analyzing data from the literature review and interviews, focusing on student profiles, learning activities, and learning outcomes to identify patterns and factors that influence graduation

3.2. Data Understanding

Following the business understanding stage is the data understanding stage, where these steps are applied by utilizing the results from the literature review, interviews, and observations. This data understanding phase consists of two main methods: data collection and attribute selection.

1. Attribute Determination

In the initial dataset, as shown in Table 1, after consulting the literature review and conducting interviews with the Academic Administration Office, academic advisors, and faculty, there are 15 attributes identified, including 14 predictor attributes and 1 target attribute, before undergoing the data preparation stage where selection will be made based on relevant attributes. The attribute selection is divided into two factors. The first factor is student identity characteristics, including gender, age, and the name of the chosen study program. The second factor pertains to the academic aspects of the students, including Student Active Semester Status (SAS), Student Leave of Absence Status (SLAS), Inactive Semester Status (ISS), GPA from semesters 1 to 4, and the number of credits hour from semesters (CH) 1 to 4.

Table 2. Attributes in the Initial Data

No	Atribut	Value
1	Study program	Categorical
2	Gender	Categorical
3	Age	Numerical
4	Student Active Status	Numerical
5	Student Leave of Absence Status	Numerical
6	Inactive Status for Semester 1	Numerical
7	GPA 1	Numerical
8	GPA 2	Numerical
9	GPA 3	Numerical
10	GPA 4	Numerical
11	CH 1	Numerical
12	CH 2	Numerical
13	CH 3	Numerical
14	CH 4	Numerical
15	Remarks	Categorical

2. Data Collection

After formulating the problem and establishing the research objectives, the next step is the collection of preliminary data. The data was obtained from literature reviews, interviews with the Academic Administration Office, academic advisors, the PSTI team, and observations related to the factors influencing timely graduation. The collected data includes the graduation status of students from Syekh Yusuf Islamic University for the years 2021, 2022, and 2023, with admission years 2017, 2018, and 2019, sourced from the Academic Information System (SINA). The focus of the academic data is the student's status at the end of the normal study period (8th semester), which is categorized into two labels: on-time graduation and late graduation, with a total of 4,450 records.

Table 3. Sample Dataset of Graduation

Study Program	Gender	AGE	SAS	SLAS	ISS	GPA1	GPA2	GPA3	GPA4	CH1	CH2	CH3	CH4	Remarks
Administrasi Publik	M	22	1	1	1	0.00	0.00	0.00	0.00	21	0	0	0	Late graduation
Akuntansi	F	20	7	0	0	3.79	3.70	3.25	3.92	19	20	20	19	On time graduation
Ilmu Hukum	M	20	7	0	0	3.90	3.90	3.75	3.76	20	20	22	21	On time graduation
Ilmu Komunikasi	F	21	2	4	1	2.79	2.84	0.00	0.00	19	19	0	0	Late graduation
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
Teknik Sipil	M	22	1	0	1	3.18	0.00	0.00	0.00	19	0	0	0	Late graduation

3.3. Data Preparation

At this stage, data preprocessing is performed on the initial data to make it more relevant for further analysis. The data obtained comes from the database of the Academic Information System of Syekh Yusuf Islamic University (SINA) in .csv format, which is then converted to .xlsx format for easier processing. At this stage, three methods are employed: data cleansing, data reduction, and data splitting. Below is a description of these three methods.

1. Data Cleansing

At this stage, data cleansing is performed on the timely graduation data, where there are NULL values in attributes such as GPA from semesters 1 to 4 and credits hour from semesters (ch) 1 to 4, which can affect data analysis during model evaluation. To address the missing values, the NULL values in these attributes are replaced with the number 0, as NULL is considered equivalent to 0. This substitution ensures that the data is more complete and supports a more accurate analysis.

Table 4. Sample Dataset of Graduation with NULL Values

Study Program	Gender	AGE	SAS	SLAS	ISS	GPA1	GPA2	GPA3	GPA4	CH1	CH2	CH3	CH4	Remarks
Administrasi Publik	F	19	7	0	0	2.70	2.92	3.13	3.50	20	19	20	19	On time graduation
Administrasi Publik	F	20	2	0	0	3.00	3.68	0.00	null	21	19	17	null	Late graduation
Manajemen	M	20	1	0	0	0.00	0.00	null	null	20	0	null	null	Late graduation
Manajemen	M	18	1	0	0	3.65	0.00	null	null	20	20	null	null	Late graduation
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
Administrasi Publik	F	20	5	3	0	3.30	3.74	3.52	3.40	20	19	22	21	Late graduation

Table 5. Sample Dataset of Graduation with NULL Values That Has Been Cleansed

Study Program	Gender	AGE	SAS	SLAS	ISS	GPA1	GPA2	GPA3	GPA4	CH1	CH2	CH3	CH4	Remarks
Administrasi Publik	F	19	7	0	0	2.70	2.92	3.13	3.50	20	19	20	19	On time graduation
Administrasi Publik	F	20	2	0	0	3.00	3.68	0.00	null	21	19	17	null	Late graduation
Manajemen	M	20	1	0	0	0.00	0.00	null	null	20	0	null	null	Late graduation
Manajemen	M	18	1	0	0	3.65	0.00	null	null	20	20	null	null	Late graduation
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
Administrasi Publik	F	20	5	3	0	3.30	3.74	3.52	3.40	20	19	22	21	Late graduation

2. Data Transformation

Next, there is the data transformation stage, where the values of variables that are in text or categorical form will be transformed using label encoding techniques. The data will be altered or consolidated into a format suitable for the data mining process. The variables being transformed include the values of the study program, gender, and remarks. This is done to enable the algorithms or models used to perform more effectively, as these algorithms or models require attributes with numerical values. As shown in Tables 6, 7, and 8, this includes the initial variable data and the resulting variable data after transformation.

Table 6. Label Encoding of the Study Program Attribute

Attribute	Label Encoding
Administrasi Publik	0
Akuntansi	1
Ilmu Hukum	2
Ilmu Komunikasi	3
Manajemen	4
Pendidikan Agama Islam	5
Pendidikan Bahasa Inggris	6
Teknik Industri	7
Pendidikan Ekonomi	8
Teknik Informatika	9
Teknik Kimia	10
Teknik Sipil	11

In Table 6, the results of the label encoding transformation stage on the program study attribute can be seen. At this stage, the names of the study programs, which were previously in categorical text form, have been converted into a numerical format. Administrasi Publik is represented as 0, Ilmu Komunikasi as 1, Ilmu Hukum as 2, Pendidikan Agama Islam as 3, Teknik Kimia as 4, Teknik Industri as 5, Teknik Informatika as 6, Teknik Sipil as 7, Pendidikan Ekonomi as 8, Pendidikan Bahasa Inggris as 9, Manajemen as 10, and Akuntansi as 11.

Table 7. Label Encoding of the Gender Attribute

Attribute	Label Encoding
M (Male)	0
F (Female)	1

Table 7 shows the results of the data transformation with label encoding, where the values of the gender attribute have been encoded such that male is represented as 0 and female as 1. Thus, each study program and gender has been assigned a unique numerical value to facilitate the subsequent data analysis process.

Table 8. Label Encoding of Remarks

Attribute	Label Encoding
On-time graduation	0
Late graduation	1

Table 8 shows the results of the data transformation with label encoding, where the values of the attribute have been encoded such that on-time graduation waktu is represented as 0 and no on-time graduation as 1. Thus, each study program and gender has been assigned a unique numerical value to facilitate the subsequent data analysis process. Below, Table 9 presents the sample data resulting from the data transformation stage.

Table 9. Sample Student Graduation Data Resulting from Data Transformation

Study Program	Gender	AGE	SAS	SLAS	ISS	GPA1	GPA2	GPA3	GPA4	CH1	CH2	CH3	CH4	Remarks
0	M	22	1	1	1	0.00	0.00	0.00	0.00	21	0	0	0	1
1	F	20	7	0	0	3.79	3.70	3.25	3.92	19	20	20	19	0
2	M	20	7	0	0	3.90	3.90	3.75	3.76	20	20	22	21	0
3	F	21	2	4	1	2.79	2.84	0.00	0.00	19	19	0	0	1
4	F	19	7	0	0	3.63	3.36	3.65	3.44	20	22	23	24	0
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
11	M	22	1	0	1	3.18	0.00	0.00	0.00	19	0	0	0	1

3. Data Splitting

At this stage, the dataset of students who graduate on time, consisting of 4,450 records, is divided using the data-splitting method. The data is separated with a ratio of 70:30, where 70% of the data is used for training and 30% for testing, as well as with ratios of 80:20 (80% training, 20% testing) and 90:10 (90% training, 10% testing). This division is carried out to optimize the model's performance by training it on the majority of the data and then testing its performance.

4. Data Reduction

The next stage is data reduction, during which the Particle Swarm Optimization (PSO) algorithm is applied. In the PSO method, particles act as representations of candidate solutions in the search space, forming a population known as a swarm. This swarm is initialized by randomly distributing values of 1 and 0. Each particle will be selected if its main component value is 1, while components with a value of 0 are ignored. In this example, PSO is used for feature selection with the following parameters: a population size of 5, a maximum of 30 iterations, an inertia weight of 1.0, acceleration constants c1 and c2 set to 1.0, and random values for speeds r1i and r2i derived from the training data.

3.4. Modelling

In the modeling stage, the Random Forest algorithm is used to predict student graduation. Subsequently, Particle Swarm Optimization (PSO) is applied. PSO serves as a feature selection method that helps to select the parameters needed by the Random Forest algorithm, thereby enhancing its accuracy.

1. Random Forest.

The stages of Random Forest, using equations 1 and 2, as shown in Table 10, yield the gain values for each attribute.

Table 10. Calculation of Gain Values from Random Forest

Attribute	Gain Values
Study Program	0.0215
Gender	0.0374
Age	0.0404
SAS	0.7457
SLAs	0.4543
ISS	0.589
GPA1	0.3248
GPA2	0.4827
GPA3	0.5549
GPA4	0.6074
CH1	0.0022
CH2	0.2697
CH3	0.4224
CH4	0.4981

Based on Table 10, the calculations using the Random Forest formulas indicate that the attribute Student active status has the highest gain value of 0.7457. The attribute with the highest gain value will be used as a branch in the decision tree. This gain value is obtained to form several decision trees from bootstrap random samples of the training data using the bagging method or random features to create a forest of decision trees. The testing process involves making predictions with each decision tree, and the final label is determined through a voting mechanism.

The results of the predictions for the dataset of students graduating on time, using the Random Forest algorithm with an 80:20 data split and predicted with 100 decision trees, are shown in Table 11.

Table 11. Final Voting Results

Remarks	Prediction	Confidence (On Time Graduation)	Confidence (Late Graduation)
0	0.99	0.01	0.00
1	0.13	0.87	1.00
0	0.96	0.04	0.00
1	0.00	1.00	1.00
…	…	…	…
0	1.00	0.00	0.00
TP		571	
FP		2	
TN		313	
FN		4	

In the prediction results shown in Table 11, there are 571 true positives (TP) for students predicted to graduate, 2 false positives (FP) for incorrect predictions of on-time graduation, 313 true negatives for students correctly predicted not to graduate on time, and 4 false negatives for incorrect predictions of students who did not graduate on time,

2. Particle Swarm Optimization

The stages of Particle Swarm Optimization, using equations 1 and 2, and shown in Table 12, illustrate the results of feature selection through Particle Swarm Optimization.

Table 12. Hasil Seleksi Fitur dengan PSO

Attribute	Value
Study Program	0.819
Gender	0.000
Age	0.000
SAS	1.000
SLAs	1.000
ISS	1.000
GPA1	1.000
GPA2	0.000
GPA3	0.488
GPA4	0.000
CH1	1.000
CH2	1.000
CH3	0.255
CH4	0.000

It can be observed in Table 12 that the feature selection results using PSO show the following values: the attribute study program has a value of 0.819, gender has a value of 0, age has a value of 0, Student Active Status (SAS) has a value of 1, Student Leave of Absence Status (SLAS) has a value of 1, Inactive Status for Semester 1 (ISS) has a value of 1, GPA1 has a value of 1, GPA2 has a value of 0, gpa3 has a value of 0.4882, GPA4 has a value of 0, credit hour in semester 1 (CH1) has a value of 1, credit hour in semester 2 (CH2) has a value of 1, credit hour in semester 3 (CH3) has a value of 0.2554, and credit hour in semester 4 (CH4) has a value of 0. It can be concluded that the best attributes are study program, Student Active Status (SAS), student Leave of Absence Status (SLAs), Inactive Status for Semester 1 (ISS), GPA1, credit hour in semester 1 (CH1), and credit hour in semester 2 (CH2).

3.5. Evaluation

In this evaluation stage, the performance results are presented using a confusion matrix from the Random Forest model, which has been trained to identify patterns and make predictions on new or unknown data (test data). The training data was split using a data splitting method with a ratio of 70:30, where 70% of

the data was used for training and 30% for testing. Additionally, other ratios used were 80:20 (80% training, 20% testing) and 90:10 (90% training, 10% testing). The test evaluation results were compared with two other algorithms: K-NN and Naïve Bayes. The following are the evaluation results of the three models, with the notation: Prediksi Positive (PP) indicating Predicted On-Time Graduation, Prediksi Negative (PN) indicating Predicted Late Graduation, Aktual Positive (AP) indicating True On-Time Graduation, and Aktual Negative (AN) indicating True Late Graduation.

Table 13. Results Confusion Matrix

Algorithm		Split data 70:30			Split data 80:20			Split data 90:10	
Random Forest		PP	PN		PP	PN		PP	PN
	AP	859	4	AP	571	4	AP	287	1
	AN	3	469	AN	2	313	AN	2	155
K-NN		PP	PN		PP	PN		PP	PN
	AP	860	3	AP	572	3	AP	287	1
	AN	92	380	AN	58	257	AN	27	130
Naïve Bayes		PP	PN		PP	PN		PP	PN
	AP	856	7	AP	569	6	AP	285	3
	AN	2	470	AN	1	314	AN	1	256

Table 14. Comparison of Model Evaluation Results Using Confusion Matrix

Algorithm	Split data	70:30		
	Accuracy	Precision	Recall	F-Measure
Random Forest	99.48%	99.65%	99.54%	99.59%
K-NN	92.88%	90.34%	99.65%	94.77%
Naïve Bayes	99.33%	99.77%	99.19%	99.48%
	Split data	80:20		
	Accuracy	Precision	Recall	F-Measure
Random Forest	99.33%	99.65%	99.30%	99.48%
K-NN	93.15%	98.91%	98.70%	98.80%
Naïve Bayes	99.21%	99.82%	98.96%	99.29%
	Split data	90:10		
	Accuracy	Precision	Recall	F-Measure
Random Forest	99.33%	99.31%	99.66%	99.48%
K-NN	93.71%	91.40%	99.65%	95.35%
Naïve Bayes	99.10%	99.56%	98.96%	99.30%

From the evaluation results presented in Table 14, Random Forest demonstrates the best performance across all data split scenarios, achieving the highest metrics for Accuracy, Precision, Recall, and F-Measure. In the 70:30, 80:20, and 90:10 data splits, as well as in the 10-Fold Cross-Validation, Random Forest consistently recorded the highest Accuracy values, reaching 99.48% for the 70:30 split and 99.33% for both the 80:20 and 90:10 splits. Meanwhile, Naïve Bayes also performed very well, though slightly below Random Forest, with a maximum Accuracy of 99.33% in the 70:30 split. K-NN exhibited the lowest performance among the three models, with its highest Accuracy recorded at 93.71% in the 90:10 split. Therefore, Random Forest is the superior model for this dataset, followed by Naïve Bayes, and lastly K-NN.

3.6. Development

The Deployment process is carried out after testing and evaluating several algorithms, namely Random Forest, K-NN, Naïve Bayes, and Random Forest + PSO. Based on the evaluation results, the Random Forest + PSO algorithm has proven to be superior in predicting on-time student graduation. Below is a display of the web-based prototype developed using Python programming language with the Streamlit framework.

Figure 6. Dashboard

Figure 7. Prediction Page

4. CONCLUSION

The findings of this research indicate that the application of the PSO method successfully identified the factors influencing on-time graduation for students, namely the study program, student enrollment status, leave of absence status, inactive status in the first semester, GPA1, CH1, and CH2. Furthermore, the use of the PSO and Random Forest algorithms to predict on-time graduation yielded the best performance, with an accuracy of 99.63%, precision of 99.77%, recall of 99.65%, and an F1 score of 99.71%. For future research, it is recommended to incorporate additional variables that may impact student graduation to enhance model accuracy, as well as to test other algorithms different from those used in this study.

However, this study has several limitations that need to be considered. First, the data used only includes certain variables, meaning it does not account for all factors that could potentially affect students' timely graduation. Second, the models employed in this study are limited to the PSO and Random Forest algorithms, and other algorithms that may perform better with the given dataset have not been tested.

For future research, it is recommended to identify and include additional variables that may influence student graduation, such as economic factors, involvement in academic activities, and others. Additionally, exploring other ensemble learning algorithms such as AdaBoost, Gradient Boosting, and XGBoost is necessary to compare their performance with the algorithms used in this study. Researchers should also consider using cross-validation techniques to ensure a more robust and generalizable model for a broader dataset.

REFERENCES

[1] I. Irawan, R. Qisthiano, M. Syahril, and P. M. Jakak, "Optimasi Prediksi Kelulusan Tepat Waktu: Studi Perbandingan Algoritma Random Forest dan Algoritma K-NN Berbasis PSO," *Jurnal Pengembangan Sistem Informasi dan Informatika*, vol. 4, no. 4, 2023, doi: 10.47747/jpsii.v4i4.1374.

[2] M. Ridwan, "Sistem Rekomendasi Proses Kelulusan Mahasiswa Berbasis Algoritma Klasifikasi C4.5," *Jurnal Ilmiah Informatika*, vol. Volume 2, 2017, doi: 10.35316/jimi.v2i1.460.

[3] E. Haryatmi and S. P. Hervianti, "Penerapan Algoritma Support Vector Machine Untuk Model Prediksi Kelulusan Mahasiswa Tepat Waktu," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 5, no. 2, 2021, doi: 10.29207/resti.v5i2.3007.

[4] Suhardjono, G. Wijaya, and A. Hamid, "Prediksi Waktu Kelulusan Mahasiswa Menggunakan SVM Berbasis PSO," *Bianglala Informatika*, vol. 7, no. 2, 2019, doi: 10.31294/bi.v7i2.6654.

[5] C. N. Dengen, K. Kusrini, and E. T. Luthfi, "Implementasi Decision Tree Untuk Prediksi Kelulusan Mahasiswa Tepat Waktu," *SISFOTENIKA*, vol. 10, no. 1, 2020, doi: 10.30700/jst.v10i1.484.

[6] D. A. Putra and M. Kamayani, "Prediksi Kelulusan Mahasiswa Tepat Waktu Menggunakan Metode Naive Bayes di Program Studi Teknik Informatika UHAMKA," *Prosiding Seminar Nasional Teknoka*, vol. 5, 2020, doi: 10.22236/teknoka.v5i.331.

[7] S. Devella, Y. Yohannes, and F. N. Rahmawati, "Implementasi Random Forest Untuk Klasifikasi Motif Songket Palembang Berdasarkan SIFT," *JATISI (Jurnal Teknik Informatika dan Sistem Informasi)*, vol. 7, no. 2, 2020, doi: 10.35957/jatisi.v7i2.289.

[8] B. A. Arifiyani and R. S. Samosir, "Sistem Simulasi Prediksi Profil Kelulusan Mahasiswa Dengan Decison Tree," *Kalbiscentia*, vol. 5, no. 2, 2018.

[9] L. Breiman, "Random Forests," *Mach Learn*, vol. 45, pp. 5–32, 2001, Accessed: Jan. 30, 2024. [Online]. Available: https://doi.org/10.1023/A:1010933404324

[10] I. D. Mienye, Y. Sun, and Z. Wang, "An improved ensemble learning approach for the prediction of heart disease risk," *Inform Med Unlocked*, vol. 20, 2020, doi: 10.1016/j.imu.2020.100402.

[11] B. Xue, M. Zhang, and W. N. Browne, "Particle swarm optimisation for feature selection in classification: Novel initialisation and updating mechanisms," *Applied Soft Computing Journal*, vol. 18, 2014, doi: 10.1016/j.asoc.2013.09.018.

[12] T. M. Shami, A. A. El-Saleh, M. Alswaitti, Q. Al-Tashi, M. A. Summakieh, and S. Mirjalili, "Particle Swarm Optimization: A Comprehensive Survey," *IEEE Access*, vol. 10, pp. 10031–10061, 2022, doi: 10.1109/ACCESS.2022.3142859.

[13] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining, Fourth Edition: Practical Machine Learning Tools and Techniques*, Fourth Edition., vol. 4. Todd Green, 2016. doi: 10.1016/C2015-0-02071-8.

[14] Y. Mardi, "Data Mining : Klasifikasi Menggunakan Algoritma C4.5," *Edik Informatika*, vol. 2, no. 2, 2017, doi: 10.22202/ei.2016.v2i2.1465.

[15] A. Ishaq *et al.*, "Improving the Prediction of Heart Failure Patients' Survival Using SMOTE and Effective Data Mining Techniques," *IEEE Access*, vol. 9, 2021, doi: 10.1109/ACCESS.2021.3064084.

[16] M. Schonlau and R. Y. Zou, "The random forest algorithm for statistical learning," *Stata Journal*, vol. 20, no. 1, 2020, doi: 10.1177/1536867X20909688.

[17] D. A. Rachmawati, N. A. Ibadurrahman, J. Zeniarja, and N. Hendriyanto, "Implementation of The Random Forest Algorithm in Classifying The Accuracy of Graduation Time For Computer Engineering Students at Dian Nuswantoro University," *Jurnal Teknik Informatika (Jutif)*, vol. 4, no. 3, pp. 565–572, Jun. 2023, doi: 10.52436/1.jutif.2023.4.3.920.

[18] M. Y. Khan, A. Qayoom, M. S. Nizami, M. S. Siddiqui, S. Wasi, and S. M. K. U. R. Raazi, "Automated Prediction of Good Dictionary EXamples (GDEX): A Comprehensive Experiment with Distant Supervision, Machine

Learning, and Word Embedding-Based Deep Learning Techniques," *Complexity*, vol. 2021, 2021, doi: 10.1155/2021/2553199.
[19] E. Ismanto and M. Novalia, "Komparasi Kinerja Algoritma C4.5, Random Forest, dan Gradient Boosting untuk Klasifikasi Komoditas," *Techno.Com*, vol. 20, no. 3, pp. 400–410, Aug. 2021, doi: 10.33633/tc.v20i3.4576.
[20] U. M. Khaire and R. Dhanalakshmi, "Stability of feature selection algorithm: A review," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 4, 2022, doi: 10.1016/j.jksuci.2019.06.012.
[21] M. H. Aghdam and S. Heidari, "Feature selection using particle swarm optimization in text categorization," *Journal of Artificial Intelligence and Soft Computing Research*, vol. 5, no. 4, 2015, doi: 10.1515/jaiscr-2015-0031.
[22] H. S. Baruah, J. Thakur, S. Sarmah, and N. Hoque, "A Feature Selection Method using PSO-MI," in *2020 International Conference on Computational Performance Evaluation, ComPE 2020*, IEEE, Jul. 2020. doi: 10.1109/ComPE49325.2020.9200034.
[23] M. Y. Kurniawan and M. E. Rosadi, "Optimasi Decision Tree Menggunakan Particle Swarm Optimization pada Data Siswa Putus Sekolah," *Jurnal Teknologi Informasi Universitas Lambung Mangkurat (JTIULM)*, vol. 2, no. 1, 2017, doi: 10.20527/jtiulm.v2i1.13.
[24] H. Herlinda, M. Itqan Mazdadi, D. Kartini, and I. Budiman, "Implementation of Particle Swarm Optimization on Sentiment Analysis of Cyberbullying using Random Forest," *JUITA: Jurnal Informatika*, vol. 11, no. 2, pp. 301–309, 2023, doi: 10.30595/juita.v11i2.17920.
[25] E. Purnamasari, D. Palupi Rini, and Sukemi, "Seleksi Fitur menggunakan Algoritma Particle Swarm Optimization pada Klasifikasi Kelulusan Mahasiswa dengan Metode Naive Bayes," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 4, no. 3, pp. 469–475, 2017, doi: 10.29207/resti.v4i3.1833.
[26] I. Ahmad, "Feature selection using particle swarm optimization in intrusion detection," *Int J Distrib Sens Netw*, vol. 2015, 2015, doi: 10.1155/2015/806954.
[27] G. Zeng, "On the confusion matrix in credit scoring and its analytical properties," *Commun Stat Theory Methods*, vol. 49, no. 9, 2020, doi: 10.1080/03610926.2019.1568485.
[28] P. Lalwani, M. K. Mishra, J. S. Chadha, and P. Sethi, "Customer churn prediction system: a machine learning approach," *Computing*, vol. 104, no. 2, 2022, doi: 10.1007/s00607-021-00908-y.

BIBLIOGRAPHY OF AUTHORS

Arif Rahman is a student in the master's program with a concentration in Information Systems at the Faculty of Information Technology, Universitas Budi Luhur, Jakarta, Indonesia. The author's research interests include Information Systems, Data Mining, and Data Analytics. In addition to academic activities, the author is also actively involved in the professional field as a programmer, focusing on the development of information systems.

Deni Mahdiana is a lecturer at the Faculty of Information Technology, Universitas Budi Luhur, Jakarta, Indonesia. The author has completed various research, particularly in the fields of Information Systems, Decision Support Systems, and Data Mining.

Achmad Fauzi is a student in the master's program with a concentration in Information Systems at the Faculty of Information Technology, Universitas Budi Luhur, Jakarta, Indonesia. He is also an entrepreneur specializing in the development of information systems, focusing on creating desktop, web, and mobile applications. The applications he develops cater to a wide range of clients, including small, medium, and large enterprises. He has been successfully running this business since 2015.