The Combination of Naive Bayes and Particle Swarm Optimization Methods of Student's Graduation Prediction

Evi Purnamasari[1], Dian Palupi Rini[2], Sukemi[3]

[1,2,3] Faculty of Computer Science, Universitas Sriwijaya, Jl. Srijaya Negara Bukit Besar, Palembang 30139, Indonesia

ARTICLE INFO

ABSTRACT

This research conducted classification testing on the study case of student graduation prediction in a university. It aims to assist the university in maintaining academic development and in finding solutions for improving timely graduation. This study combined two methods, i.e., Naive Bayes and Particle Swarm Optimization, to produce a better level of accuracy. The Naive Bayes method is a statistical classification method used to predict a student's graduation in this study. That will be further enhanced using the Particle Swarm Optimization method to produce a better level of accuracy. There are 10 (ten) samples in this study randomly selected from the alumni data of UIGM students in 2011-2014. From the test results, this research resulted in an accuracy value of 90% from the Naive Bayes algorithm testing, after testing the Naive Bayes with Particle Swarm Optimization, which produced an accuracy value of 100%. The conclusion obtained from the results is the Naive Bayes method has a higher accuracy value if combined with Particle Swarm Optimization. Thus the university can more easily predict whether or not the students graduate on time for the upcoming graduation period. The results of this test prove that to predict student graduation using the Naive Bayes method with Particle Swarm Optimization is appropriate.

Corresponding Author:

Evi Purnamasari, Dian Palupi Rini, Sukemi
Universitas Sriwijaya, Jl. Srijaya Negara Bukit Besar, Palembang 30139, Indonesia
Email: evipurnama8@gmail.com, dprini@unsri.ac.id, sukemi@ilkom.unsri.ac.id

1. INTRODUCTION

Problems with graduation rates are experienced at many universities, one of which is at IGM University. Graduation rates cannot be predicted yet, so there is no management to avoid a drop in graduation because all of this will have an impact on university accreditation, so graduation predictions need to be done. The quality of a university is influenced by many factors; one of them is the success of students who graduate within the stipulated time. The higher numbers of students who graduate timely fashion give effect to the accreditation of the department in a university [1].

Several factors that influence the number of a college graduate in a university are required to be analyzed in order to make the anticipation policy so that all students can timely graduate. There are many factors that influence graduation rate, namely lower academic skills, lecture program, achievement index, or other factors [2].

Some former researches related to some applications on the prediction of timely graduation have been conducted in several fields, which can be explained as follows. The research applied C 4.5 algorithm from the performance calculation of algorithms, resulting in the prediction of this student graduation algorithm C 4.5 reaching 82% of accuracy level [3]. The accuracy level of dataset testing, which used the Naive Bayes method towards the college graduation rate in this research, was 80.72% of 1.162 of datasets used for training data and 587 datasets for testing data in a university using 6 attributes of student data [4].

Research on student performance using the Naive Bayes Method showed that the accuracy rate was 85% of students will timely graduate. The use of Naive Bayes would be better if it applied more training data [5]. A

study related to the predictions of the study duration of a college student using the Naive Bayes Algorithm revealed that the accuracy of dataset testing has resulted in an average accuracy rate of 85.17% [6].

Research on student performance prediction with Naive Bayes calculation result obtained for this classification using Naive Bayes method showed the accuracy value of this research reached 70% by using several attributes, including Indonesian national exam scores and English national exam scores, the average score of national exam, any achievement related to languages, and the number of books read each month [7].

Research on Student Graduation Prediction using Naive Bayes Classifier (NBC) the main objective of this research is to produce a system on the NBC algorithm that can predict a student's success to be able to complete the study period in a timely manner based on student performance, which is the student's GPA from semester 1 to semester 4. Researchers conduct trials with 2 stages. Namely, the first stage is a literature review, and the second stage determines the prediction of Computer Science student achievement using the NBC method [8].

Research on detection cases, with the Classification method using the Naive Bayes algorithm and Improved Particle Swarm Optimization (PSO), aims at classification to help in the prediction of release in online learning. This study concludes that a simple modification to test standard PSO with Naive Bayes through front attribute selection using accuracy estimates in online learning systems. The calculation result of the trial shows that the use of variables does not appear to be related to a class attribute, which tends to reduce the accuracy and reliability of the classification model. However, when using the hybrid PSO algorithm, it clearly tends to find a small subset of the attributes of PSO with Naive Bayes. Naive Bayes is combined with the PSO method because hybrids have been proven to be the best in the ability to choose features without reducing an accuracy value in the classification method [9].

Research that discusses a prediction with the Naive Bayes algorithm and Particle Swarm Optimization, the authors submit that the proposed model is very effective and efficient to improve the accuracy of the Naive Bayes algorithm by using particle optimization in the feature selection section that achieves the same performance or maybe better classification. Researchers have successfully developed new algorithms that maximize classification performance and can minimize the number of features. The conclusion from the results of the researchers, analyzing this algorithm is automatically able to evolve feature choices with fewer features and improve classification performance compared to using all features in the dataset [10].

Research on Particle Swarm Optimization (PSO) based algorithm with Naive Bayes algorithm suggested a new model diagnose a disease based on a combination of PSO algorithms and Naive Bayes classification, due to the fact that the presence of algorithms to choose the best data training and avoid selecting those who are decreasing classification accuracy. This test obtained an accuracy rate of 97.95%, and according to the results obtained that in improving classification accuracy is not always necessary to present a new classification method by selecting the best training data and eliminating inappropriate training data, but the classification accuracy rate can increase [11].

Studies using the Naive Bayes classification method is often used as a classification method to analyze the college graduation rate the dataset, but those have different results, although those researches resulted in a high accuracy rate and some low accuracy value. There are several researchers who have classified Naive Bayes with Particle Swarm Optimization to generate an accuracy rate. Researchers who have made improved Particle Swarm Optimization (PSO) generated Improved Particle Swarm Optimization (IPSO). This improvement aims to strengthen PSO capability in developing results efficiently [12]. The researcher proposed Hybrid Naive Bayes testing and Improved Particle Swarm Optimization to obtain optimal accuracy rate in classifying case study of college graduate data.

2. RESEARCH METHOD

In this stage, this study is conducted by reviewing several papers related to this research. When it is found that some issues have not been resolved from previous research, a literature review is conducted to figure out those issues that will be analyzed. This study uses the Naive Bayes method and is improved using PSO in analyzing or classification case studies of UIGM student graduation data so as to produce better accuracy.

2.1. Collecting Data

In this stage, the collection data techniques are performed by the researcher to collect datasets. In collecting data, there are secondary and primary data sources. This research used datasets from Indo Global Mandiri University data, which was the secondary data on SIMAK through the Bureau of Academic Administration, which had 11 departments of undergraduate student data consisting of graduation years from 2011 to 2014 with the total of 893 Data. The training dataset is shown in Table 1.

Table 1. Training Dataset

Gender	City	School Type	Employment Status	Study Course	IPS 1	IPS 2	IPS 3	IPS 4	OT/NOT
M	Palembang	NON SMA	Work	Afternoon	3.23	2.55	3.25	3.13	On Time
M	Palembang	SMA	Work	Afternoon	3.45	3.36	3.38	3.25	On Time
M	Palembang	SMA	Nor Work	Morning	3.05	2.64	2.35	0.53	Not On Time
M	Outside Palembang	SMA	Not Work	Morning	3.23	2.91	2.9	2.29	Not On Time

2.2. Naive Bayes Algorithm Model

Naive Bayes is a classifier or statistical grouping that can be used to predict the possibility of membership in a class. Based on the Bayes Theorem is able to classify something similar to a decision tree [13].

Naive Bayes algorithm is one of the algorithms in the classification method. Naive Bayes is a classification with a probability, and the statistical method that has been raised by the British scientist Thomas Bayes is a prediction of future opportunities based on experience from the past, so it is known as the Bayes Theorem. The theorem is then combined with Naive, where it is assumed that the conditions between attributes are independent. The Naive Bayes classification has assumed that having or not having certain characteristics of a class will not have anything to do with the characteristics of another class [6].

Naive Bayes formula used is:

$$P(H|X) = \big(P(X|H)P(H)\big)/\big(P(X)\big) = P(X|H) \times P(H) / P(X) \tag{1}$$

Where
X   = Data with unknown classes
H   = The data X hypothesis is a more specific class
P (*H* | X)   = H hitosesis probability based on condition X (posteriori probability)
P (*H*)   = Hitosesis probability H (prior probability)
P (X | *H*)   = Probability X is based on the conditions in hypothesis H
P (*H*)   = Probability X

2.3 PSO

Particle Swarm Optimization (PSO) is a Bioinspired algorithm that can be used to solve problems and find solutions given space by calculating the probability of formation of the particles that give space. The calculation is carried out until the best solution is found [14].

PSO General Algorithm:
1. For {every particle}
2. Particle initialization
3. End
4. Do
5. For {every particle}
6. Calculate fitness value
7. If the fitness value is better than PBest in history then
8. set GBest equal to the current value
9. End
10. Select the Pbest of all particles as GBest
11. For {every particle}
12. Calculate the particle velocity according to the equation (i)
13. Update the position of the particles according to the equation (ii)
14. End

Particle swarm optimization is a metaheuristic that was introduced by Kennedy and Eberhart in 1995, whose algorithmic process was motivated by social behavior in animals. PSO is a population-based stochastic approach to solving continuous and discrete optimization problems. At each iteration, the velocity of the

particle is adjusted stochastically according to the effect on the best solution and the best solution of its neighbors, then calculates new points to be evaluated [15].

2.4 PSO Based Naive Bayes Model

The steps that will be performed in implementing PSO combination algorithm in finding a solution to optimal as follows:

1. Initialization
   In the initial stage of PSO initialization usually use code binary to simplify a problem
   - Initialization of Initial Velocity
     In initial velocity of all particles is 0.
   - Initialization of initial particle position
     The zero Iteration on the early position of particle rising using the equation as follows:

$$x = x|min + rand[0,1] \times (x|max + x|min) \tag{2}$$

   - Initialization PBest and Gbest

2. Update Velocity
   Calculate new velocity using

$$\vee j(i) = \vee j(i-1) + c1r1[Pbest, j - xj(i-1)] + c2r2[Gbest, j - xj(i-1)] \tag{3}$$

3. Position Update
   Calculate function using

$$f(x) = 1 / (1 + e^{\wedge}(-X)) \tag{4}$$

4. PBest and Gbest Update

3. RESULTS AND DISCUSSION

This research applied the classification method using the Naive Bayes algorithm to produce a better accuracy rate from the previous research, which used hybrid NB – improved PSO. NB – Improved PSO is a combination of Naive Bayes and improved Particle Swarm Optimization. The main component of the proposed algorithm is the NB algorithms application for the PSO optimization method.

NB is to determine the accuracy level of UIGM student's graduation based on SIMAK UIGM data, using student graduation data with Naive Bayes. This model calculates the number of timely and not timely graduate alumni using attributes; they are IP Semester 1, IP Semester 2, IP Semester 3, IP Semester 4, gender, course program, origin, Senior High School origin SMA, employment status.

From Table 1 training dataset, there are 9 attributes. There are two class values, i.e., On time and not, to determine the data class requires the calculation of possible values that can be known from the training data.

Calculate P(C) value for i = on time and not on time

P (status = on time)	= 5/10	= 0.5
P (status = not on time)	= 5/10	= 0.5
P (status = gender = L|status = On time)	= 4/5	= 0.8
P (status = gender = L|status = Not on time)	= 4/5	= 0.8
P (status = gender = P|status = on time)	= 1/5	= 0.2
P (status = gender = P|status = not on time)	= 1/5	= 0.2
P (status = city = palembang|status = on time)	= 2/5	= 0.4
P (status = city = palembang |status = not on time)	= 3/5	= 0.6
P (status = city = outside palembang|status = on time)	= 3/5	= 0.6
P (status = city = outside palembang|status = not on time)	= 2/5	= 0.4
P (status = school type = SMA|status = on time)	= 3/5	= 0.6
P (status = school type = SMA|status = not on time)	= 4/5	= 0.8
P (status = school type = NON SMA|status = on time)	= 2/5	= 0.4

P (status = school type = NON SMA|status = not on time)   = 1/5   = 0.2

P (status = employment status = work|status = on time)   = 5/5   = 1
P (status = employment status = work|status = not on time)   = 0/5   = 0

P (status = employment status = not work|status = on time)   = 0/5   = 0
P (status = employment status = not work|status = not on time)   = 5/5   = 1

P (status = study program = morning|status = on time)   = 0/5   = 0
P (status = study program = morning|status = not on time)   = 5/5   = 1

P (status = study program = afternoon|status = on time)   = 5/5   = 1
P (status = study program = afternoon|status = not on time)   = 0/5   = 0

P (status = IPS 1 = above 3|status = on time)   = 5/5   = 1
P (status = IPS 1 = above 3|status = not on time)   = 5/5   = 1

P (status = IPS 1 = under 3|status = on time)   = 0/5   = 0
P (status = IPS 1 = under 3|status = not on time)   = 0/5   = 0

P (status = IPS 2 = above 3|status = on time)   = 3/5   = 0.6
P (status = IPS 2 = above 3|status = not on time)   = 2/5   = 0.4

P (status = IPS 2 = under 3|status = on time)   = 2/5   = 0.4
P (status = IPS 2 = under 3|status = not on time)   = 3/5   = 0.6

P (status = IPS 3 = above 3|status = on time)   = 5/5   = 1
P (status = IPS 3 = above 3|status = not on time)   = 1/5   = 0.2

P (status = IPS 3 = under 3|status = on time)   = 0/10   = 0
P (status = IPS 3 = under 3|status = not on time)   = 4/10   = 0.8

P (status = IPS 4 = above 3|status = on time)   = 3/5   = 0.6
P (status = IPS 4 = above 3|status = not on time)   = 1/5   = 0.2

P (status = IPS 4 = under 3|status = on time)   = 2/5   = 0.4
P (status = IPS 4 = under 3|status = not on time)   = 4/5   = 0.8

After calculating the Probability value from the amount of data, a Particle value is randomly determined between 0 and 1, so the values in Table 2 are obtained.

Table 2. Particle Random Value between 0 and 1

Particle	Gender	City	School Type	Employment Status	Study Program	IPS 1	IPS 2	IPS 3	IPS 4
1	1	1	1	1	1	1	0	0	1
2	0	1	0	1	0	1	0	0	0
3	1	0	0	0	1	1	1	0	1
4	0	1	1	1	1	1	1	1	1
5	0	0	0	0	0	1	1	1	1
6	0	1	0	0	0	1	0	0	1
7	0	0	0	0	1	1	1	0	1
8	1	1	1	1	1	0	0	0	1
9	1	0	1	1	0	1	1	0	0
10	1	1	1	1	0	1	1	1	1

After getting random Particle values between 0 and 1, the particles are counted on time and not on time using these random values where the results are obtained in Table 3 and Table 4.

Table 3. Calculation on-Time Particle

Particle	Gender	City	School Type	Employment Status	Study Program	IPS 1	IPS 2	IPS 3	IPS 4	Total
				on time						
1	0.8	0.4	0.4	0.4	1	1	0	0	0.6	**0.511111111**
2	0	0.4	0	0.4	0	1	0	0	0	**0.2**
3	0.8	0	0	0	0	1	0.4	0	0.4	**0.288888889**
4	0	0.6	0.6	1	0	1	0.4	0	0.4	**0.444444444**
5	0	0	0	0	0	1	0.6	1	0.4	**0.333333333**
6	0	0.4	0	0	0	1	0	0	0.4	**0.2**
7	0	0	0	0	0	1	0.4	0	0.4	**0.2**
8	0.2	0.6	0.4	0.4	1	0	0	0	0.4	**0.333333333**
9	0.8	0	0.6	0.4	0	1	0.6	0	0	**0.377777778**
10	0.2	0.6	0.4	1	0	1	0.6	1	0.6	**0.6**

Table 4. Calculation not on Time Particle

Particle	Gender	City	School Type	Employment Status	Study Program	IPS 1	IPS 2	PS 3	IPS 4	Total
				not on time						
1	0.8	0.6	0.2	0	0	1	0	0	0.2	**0.31111111**
2	0	0.6	0	0	0	1	0	0	0	**0.17777778**
3	0.8	0	0	0	1	1	0.6	0	0.8	**0.46666667**
4	0	0.4	0.8	1	1	1	0.6	0.8	0.8	**0.71111111**
5	0	0	0	0	0	1	0.4	0.2	0.8	**0.26666667**
6	0	0.6	0	0	0	1	0	0	0.8	**0.26666667**
7	0	0	0	0	1	1	0.6	0	0.8	**0.37777778**
8	0.2	0.4	0.2	0	0	0	0	0	0.8	**0.17777778**
9	0.8	0	0.8	0	0	1	0.4	0	0	**0.33333333**
10	0.2	0.4	0.2	1	0	1	0.4	0.2	0.2	**0.4**

Naive Bayes's fitness results are obtained on time, and particle calculations are not on time where the fitness is checked between the graduation status data with the Naive Bayes system graduation status, so the fitness value in Table 5 is obtained.

Table 5. Naive Bayes Fitness Value

Particle	Fitness	Graduation Status System NB	Graduation Status Data	Fitness Value
1	0.377777778	On-Time	On-Time	1
2	0.488888889	On-Time	On-Time	1
3	0.088888889	Not On Time	Not On Time	1
4	0.177777778	Not On Time	Not On Time	1
5	0.111111111	On-Time	On-Time	1
6	0.355555556	Not On Time	Not On Time	1
7	0.266666667	Not On Time	Not On Time	1
8	0.244444444	On-Time	On-Time	1

Particle	Fitness	Graduation Status System NB	Graduation Status Data	Fitness Value
9	0.444444444	On-Time	On-Time	1
10	0.355555556	On-Time	Not On Time	0
		Accuracy		**90 %**

The results of the Naive Bayes fitness value obtained an accuracy value of 90%. Then accuracy value result from Naive Bayes will be optimized using Particle Swarm Optimization. So it will be obtained from a combination of those two methods in Table 6.

Table 6. Naive Bayes Fitness Value using Particle Swarm Optimization

NPM	Year	Name	NB+PSO Result	Real Result	Fitness Value
2012620010	2012	Anrie Vernando	On-Time	On-Time	1
2012620014	2012	Muhammad Ajie Danu Wijaya	On-Time	On-Time	1
2012620017	2012	Rizki Septiandi	Not On Time	Not On Time	1
2012620021	2012	M. Barkah	Not On Time	Not On Time	1
2012620022	2012	Hamzah Kurniawan	On-Time	On-Time	1
2012620026	2012	M. Roffy Agung Setia	Not On Time	Not On Time	1
2012620027	2012	R. Gusti Prima Dyora	Not On Time	Not On Time	1
2012620031	2012	Nurhasanah	On-Time	On-Time	1
2012620033	2012	M. Reza Baskara	On-Time	On-Time	1
2012710002	2012	Jerisca	Not On Time	Not On Time	1
		Accuracy			**100%**

The results of combining the two Naive Bayes methods with PSO obtained an accuracy value of 100%. Here proves that Naive Bayes, combined with PSO, further increases the value of accuracy.

4. CONCLUSION

Based on the test results that have been done that the method Naive Bayes generates an accuracy value of 90% after combining with Particle Swarm Optimization produces a higher accuracy value of 100%. In connection with it, researchers advise for subsequent further studies to use the merging method for better results.

REFERENCES

[1] A. Aryasanti. Sistem Komparasi Naive Bayes dan Decision Trees untuk Menentukan Klasifikasi Kegagalan Studi Mahasiswa," *Jurnal TICOM*, Vol. 6, No. 3, 2018, pp. 103-111. Online

[2] A. Romadhona, S. Suprapedi, and H. Himawan, "Prediksi Kelulusan Mahasiswa Tepat Waktu Berdasarkan Usia, Jenis Kelamin dan Indeks Prestasi Menggunakan Algoritma Decision Tree," *Jurnal Teknologi Informasi Cyberku*, Vol. 13, No. 1, pp. 69–83, 2017. Google Scholar

[3] I. P. Astuti, "Prediksi Ketepatan Waktu Kelulusan Dengan Algoritma Data Mining C4.5," *Fountain of Informatics Journal*, vol. 2, no. 2, pp. 41-45, 2017. DOI: 10.21111/fij.v2i2.1067

[4] Salmu, S., & Solichin, A. "Prediksi Tingkat Kelulusan Mahasiswa Tepat Waktu Menggunakan Naïve Bayes: Studi Kasus UIN Syarif Hidayatullah Jakarta," *Prosiding Seminar Nasional Multidisiplin Ilmu Universitas Budi Luhur*, pp. ICT-701-709, April 2017. Research Gate

[5] S. Jayaprakash, E. Balamurugan, and V. Chandar, "Predicting Students' Academic Performance Using Naïve Bayes Algorithm," *8th Annual International Applied Research Conference*, 2015. Google Scholar

[6] M. W. Amelia, A. S. M. Lumenta, and A. Jacobus, "Prediksi Masa Studi Mahasiswa dengan Menggunakan Algoritma Naïve Bayes," *Jurnal Teknik Informatika*, vol. 11, no. 1, pp. 2017. DOI: 10.35793/jti.11.1.2017.17652

[7] U. Pujianto, E. N. Azizah and A. S. Damayanti, "Naive Bayes using to predict students' academic performance at faculty of literature," *2017 5th International Conference on Electrical, Electronics and Information Engineering (ICEEIE)*, Malang, 2017, pp. 163-169. DOI: 10.1109/ICEEIE.2017.8328782

[8] E. Sugiharti, "Predictive Evaluation of Performance of Computer Science Students of Unnes Using Data Mining Based on Naive Bayes Classifier (NBC)," *Journal of Theoretical and Applied Information Technology*, vol. 95, no. 4, pp. 902–911, 2017. Google Scholar

[9] P. V. P. Sundar and D. A. V. S. Kumar, "A Hybrid Classification Method for Disengagement Detection in Online

          Learning," *International Journal of Education and Information Studies*, vol. 5, no. 1, pp. 67–74, 2015. Google Scholar

[10] S. A. Pattekari, and A. Parveen, "Prediction System for Heart Disease Using Naive Bayes," *International Journal of Advanced Computer and Mathematical Sciences*, vol. 3, no. 3, 290–294. 2012. Google Scholar

[11] N. K. Ghanad and S. Ahmadi, "Combination of PSO Algorithm and Naive Bayesian Classification for Parkinson Disease Diagnosis," *Advances in Computer Science: An International Journal*, vol. 4, no. 4, 119–125, 2015. Google Scholar

[12] W. F. Mahmudy, "Improved Particle Swarm Optimization untuk Menyelesaikan Permasalahan Part Type Selection dan Machine Loading pada Flexible Manufacturing System (FMS)," *Konferensi Nasional Sistem Informasi*, 2015. Google Scholar

[13] A. Kesumawati and D. T. Utari, "Predicting patterns of student graduation rates using Naïve bayes classifier and support vector machine," AIP Conference Proceedings, vol. 2021, 2018. DOI: 10.1063/1.5062769

[14] S. Kaur and M. Goyal, "Fast and robust Hybrid Particle Swarm Optimization Tabu Search Association Rule Mining (HPSO-ARM) algorithm for Web Data Association Rule Mining (WDARM)," *International Journal of Advance Research in Computer Science and Management Studies*, vol. 2, no. 9, 448–451, 2014. Google Scholar

[15] H. Abadlia, N. Smairi, and K. Ghedira, "A hybrid Immigrants schema for particle swarm optimization algorithm," *Procedia Computer Science*, vol. 126, pp. 105–115, .2018. DOI: 10.1016/j.procs.2018.07.214