MATICS : Jurnal Ilmu Komputer dan Teknologi Informasi ISSN : 1978-161X(p); 2477-2550(e)
Volume 12, No. 1(2020), pp 1-9 DOI : 10.18860/mat.v12i1.8216
Received : December 25th 2019; **Accepted** : March 4 th 2020 ; **Avalaible Online** : April 5th 2020

Comparative Analysis of 5 Algorithm Based Particle Swarm Optimization (PSO) for Prediction of Graduate Time Graduation

Achmad Noercholis, Mohammad Zainuddin

Abstract - Graduation information is very important for Higher Education involved in education. The data received by students each year is an important part as a source of information for making decisions on the Higher Education side in admitting new students. The results show the PSO-based K-NN Algorithm at k-optimum = 19 has the best performance of the 5 existing algorithms, with an Accuracy value = 74.08% and an Under Curve Area (AUC) value = 0.788. Attributes of Gender, Semester Achievment Index 1, 2, 4, 6 and 7 as well as Employment Status make a real contribution to the right graduation of students. The addition of the Particle Swarm Optimization (PSO) feature always increases the accuracy value, while the highest increase in accuracy value in the Decision Tree (C4.5) Algorithm is 5.21%, the lowest in the Vector Support Engine Algorithm of 1.79%. The K-Nearest Neighbor (K-NN) algorithm corresponds to the third order, it remains the algorithm that has the best value, the highest accuracy value, this is due to the questionable value before discussing the PSO features. For the evaluation phase, the results of the accuracy are much better if using the overall data training (Angaktan 2007-2011), both with the 2011 Force test data or the 2010-2011 Force. Timely graduation of students begins in the class of 2011, with a value of 98.18% (100% resolution), meaning students start graduating on time

Keywords: Algoritma Naive Bayes, Decision Tree (C4.5), k-Nearest Neighbor (k-NN), Neural Network, Support Vector Machine (SVM), Particle Swarm Optimization (PSO), Accuracy, Area Under the Curve (AUC).

Manuscript received March 22, 2020. This work was supported in part by Informatics Engineering Department of STIMIK ASIA Malang.

Mohammad Zainuddin is with the Informatic Engineering Departement of STIMIK ASIA Malang, Indonesia (email: mzein@asia.ac.id)

Achmad Noercholis, was with the Informatic Engineering Departement of STIMIK ASIA Malang, Indonesia (e-mail: anoercholis@asia.ac.id).

I. INTRODUCTION

Education is a social activity that enables the community to remain and thrive [1]. Higher education is one of the basic requirements in finding a job, which will prepare candidates for quality graduates and have skills in their fields. Achieving this degree requires a normal time of 3.5 to 4.5 years, but in practice many students cannot complete their studies during the specified normal time. Many factors cause this inaccuracy of graduates, these factors can be sourced from internal factors and external factors.

The prediction of timely graduation which is done at this time is only based on forecasts from the GPA data (cumulative achievement index) and IMK (cumulative mutualism index) the previous semester. Prediction is almost the same as classification and estimation, it's just that predictions are used to predict certain values that will occur in the future [1].

Meanwhile STMIK ASIA MALANG has SIS (Student Information Services) data which has not been utilized to the fullest is a pity if such a large dataset is not utilized to extract what information is contained therein. In addition, so far there has been an assumption from the Assistant Chair 1 (Academic Affairs) of STMIK ASIA MALANG that to predict graduation rates on time is enough by looking at the previous GPA and IMK data. Departing from the problems above, this research is conducted to conduct data mining of the SIS (Student Information Services) dataset so that information on timely graduation from STMIK ASIA MALANG students is obtained.

STMIK ASIA MALANG is an educational organization engaged in multimedia and technology. Information on graduation rates from students is very important to improve services that can make students comfortable and can graduate on time. The use of data mining can be used as a consideration in making further decisions about factors affecting graduation, especially factors in the student master data [2].

The prediction of a student's timely graduation is very important for the university, one of which is used as consideration for the admission of new school year students, so there are no advantages or disadvantages. So that in predicting students' graduation on time is not wrong, we need the best algorithm to support the prediction decision. The algorithm is said to be the best algorithm, one of which is marked by the greatest

accuracy value among other algorithms. On this occasion the author wants to examine which of these 5 algorithms has the greatest accuracy value which is then used as a support tool for the timely graduation prediction of students at STMIK ASIA Malang.

A. Related Research

In the table above, the average is only about predicting student graduation using one of the algorithms with or without using the selection feature. In number 3, the 3 algorithms applied only tell that the 3 methods can produce data classification of active students and graduates. In number 5, the 3 algorithms applied only tell that the 3 methods have a non-uniform prediction level between the algorithms. So it has never been known which algorithm is better than some existing algorithms, especially regarding the prediction of students' timely graduation. From a number of journals that were read by the author, many previous researchers used the Naïve Bayes Algorithm, Decesion Tree (C4.5), Neural Network, Super Vector Machine (SVM) and the K-NN Algorithm in conducting student timely prediction graduation research. The foregoing is the reason for conducting research to find which of these 5 types of algorithms produce the highest accuracy values. Furthermore, the Particle Swarm Optimization (PSO) selection feature is added to be applied to the 5 algorithms to determine the increase in accuracy and want to know which algorithm has the highest accuracy value and want to know the weighting value of each determinant attribute.

B. Theoretical Basis

Students are an elite society where students have more complex intellectual characteristics than groups of their age who are not students, or age groups below and above them. "The intellectual characteristic is the ability of students to face, understand and find ways of solving problems more systematically [8].

Accuracy of the period of study of students is very important to note, this is because a decrease in the number of graduations will eliminate the amount of institutional income and affect the government's assessment and affect the accreditation status of the institution. According to Suhartinah & Ernastuti there are several factors that can influence student graduation including the final high school grades, Semester Achievement Index (SAI), parents 'salaries and parents' work[9]. A college usually uses an achievement index as an academic assessment, many universities set minimum standards that are difficult for students to obtain. Many variables can be used in predicting student graduation such as age, marital status, number of siblings and others.

Minister of Education and Culture Regulation No. 49/2014 which regulates the maximum undergraduate study period of five years is considered too heavy by students. For this reason, the government will return the bachelor study period to seven years. The Ministerial Regulation regulates the minimum study load of S-1 / D-4 students is 144 SKS (semester credit units). To complete the entire SKS load, S-1 / D-4 students are given a time limit of 4 to 5 years (8 to 10 semesters).

C. Naïve Bayes Algorithm

Bayesian classification is a simple probabilistic based prediction technique based on the Bayes theorem (Bayes rule) with the assumption of independence (not dependency). The strong (naïve) in other words Naïve Bayes is a model that uses "independent feature models"

In Naïve Bayes, what is meant by strong independence of features is that a feature of a data does not relate to the presence or absence of other features in the same data, for example in the case of classification of animals with attributes, earlobe, childbirth, weight and lactation. In fact, animals that are ear-leafed and breastfeeding usually breed with breeds such as monkeys, pigs, goats, horses etc., conversely animals that are ear-leafed and not breastfeeding usually breed by laying eggs such as snakes, birds, lizards etc. Here there is a dependence on breastfeeding attributes, leafy ears usually give birth to the opposite is also the same. In Bayes, this is not seen so that each feature has no relationship.

D. Naïve Bayes Classifier

Naïve Bayes is a classification with probability and statistical methods proposed by the British scientist Thomas Bayes, which predicts future opportunities based on past experience so that it is known as the Bayes Theorem. Bayesian classification is based on Bayes theorem which has the ability to classify similarly to decesion trees and neural networks. Bayesian classification is proven to have high accuracy and speed when applied to databases with large data [10].

Bayes' theorem has the following general form:

$$P(H \mid X) = \frac{P(X \mid H)P(H)}{P(X)} \quad \text{..... (1)}$$

Information:

X = Data with unknown classes
H = Data X hypothesis is a specific class
P (H | X) = H hypothesis probability based on condition X (posteriori prob.)
P (H) = Hypothesis probability H (prior prob.)
P (X | H) = Probability of X based on these conditions
P (X) = Probability of X

E. Decision Tree (C4.5) Algoritma

In general the C4.5 algorithm for constructing a decision tree is as follows [11]:

a. Select the attribute as the root
b. Create a branch for each value
c. Divide cases in branches
d. Repeat the process for each branch until all cases in the branch have the same class.

To choose an attribute as the root, based on the highest gain value of the existing attributes. To calculate the gain a formula like the following is used:

$$\text{Gain}(S,A) = \text{Entropy}(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * Entropy(S_i)$$

..... (2)

Information:

S : The set of cases
A : Attributes
N : Number of attribute part A
| Si | : Number of cases on the partition to i
| S | : Number of cases in S

Before getting the Gain value is to look for the value of Entropy. Entropy is used to determine how informative an input attribute is to produce an attribute. The basic formula of Entropy is as follows:

$$\text{Entropy}(S) = \sum_{i=1}^{n} -pi * \log_2 pi$$ ..... (3)

Information:

S : Case Set
N : Number of partitions S
Pi : Proportion from Si to S

F. K-Nearest Neighbor Algorithm

The k-Nearest Neighbor (k-NN) algorithm is a method for classifying objects based on learning data that is the closest distance to the object. K-NN is a supervised learning algorithm where the results of the new query instance are classified based on the majority of the categories in kNN. The class that appears the most will be the class that results from the classification. The purpose of this algorithm is to classify new objects based on attributes and training samples.

The k-Nearest Neighbor algorithm uses the neighbor classification as the predicted value of the new query instance. This algorithm is simple, works based on the shortest distance from the query instance to the training sample to determine its neighbors [12]

Steps to calculate the k-Nearest Neighbor method with the closest distance (euclidian) include:

a. Determine the parameter k
b. Calculate the distance between data to be evaluated with all training
c. Sort the distance formed
d. Determine the closest distance to the order k
e. Pair the appropriate class
f. Looks for the number of classes from the nearest neighbor and sets the class as the data class to be evaluated

$$d_i = \sqrt{\sum_{i=1}^{p} (x_{2i} - x_{1i})^2}$$ ……..……..(4)

Information:

x_1 = Sample data
x_2 = Test data or testing data
i = Data variable
d = Distance
p = Dimension of data

G. Neural Network Algorithm

Neural Network (NN) is a parallel distributed processor, made of simple units, and has the ability to store knowledge obtained experimentally and is ready for use for various purposes [13]. Basically, a learning system is a process of adding knowledge to NN which is continuous so that when used that knowledge will be maximally exploited in recognizing an object. Neurons are the basic part of processing a Neural Network. Below this is the basic form of a neuron.

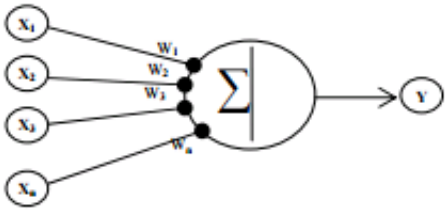

Figure 1. Communication Process Between Neurons

Figure 1 shows that NN consists of processing units in the form of neurons. Y as output receives input from neurons X1, X2, X3, ..., Xn with weights W1, W2, W3, ..., Wn. The results of the sum of all neuronal impulses are compared with certain threshold values through the activation function f of each neuron. The activation function is used as a determinant of the output of a neuron.

H. Support Vector Machine (SVM) Algorithm

According to Prasetyo [14], Support Vector Machine (SVM) is a method rooted in statistical learning theory whose results are very promising to provide better results than other methods. SVM can also work well on high-dimensional data sets, even SVM that uses the Kernel technique must map the original data from its original dimension to another relatively high dimension.

According to Y.Yin, Han & Cai [15], Support Vector Machine (SVM) is defined as a set of related learning methods that analyze data and recognize patterns, which are then used for classification and regression analysis. SVM takes a set of input data and predicts for each given input, coming from two classes which are then classified by finding the best hyperplane value.

According to Li, You & Liu [16], Support Vector Machine (SVM) is learning that leads to quadratic programming with linear constraints. Based on the risk minimization of structured principles, SVM seeks to minimize the limits on empirical errors, so that the new prediction model effectively avoids over-posing problems. In addition, the SVM model works in the high dimensional feature space formed by non-linear mapping of N-dimensional vector input x into the K-dimensional feature space (K>N) through the use of nonlinear (x) u functions

I. *Particle Swarm Optimization (PSO) Feature Selection*

Particle Swarm Optimization (PSO) was formulated by Dr. Eberhart and Dr. Kennedy in 1995. PSO is a population search method, which originated from research for the movement of groups of birds and fish in search of food [17]. Like the Genetic Algorithm, PSO performs searches using populations (called swarm) of individuals (called particles) that are updated from each iteration that is performed. To find the optimal solution, each particle moves in the direction of the previous best position (pbest) and the best global position (gbest).

In this study, PSO is used to select features (features selection) or attributes, then PSO uses the weighted attributes that have been calculated and attributes that have been selected will be predicted using the Naïve Bayes Algorithm and the k-Nearest Neighbor (k-NN) Algorithm

J. *K-Fold Cross Validation*

According to Fu [18], k-Fold Cross Validation repeats k-times to divide a set of samples randomly into k subsets that are mutually independent, each repetition leaving one subset for testing and the other subset for training. According to Vercellis [19], with k = 5 or 10 can be used to estimate the level of error that occurs, because the training data at each fold is quite different from the original training data. Overall, 5 or 10 Fold Cross Validation are both recommended and agreed upon. Calculating the accuracy value can be done using the equation:

$$\text{Akurasi} = \frac{jumlah.klasifikasi.benar}{jumlah.data.uji} x100\% \quad \ldots\ldots (5)$$

A.

K. *Confusion Matrix*

To evaluate the classification model based on the calculation of the testing object which is predicted to be true and incorrect. This calculation is tabulated into a table called confusion matrix [20]. Confusion matrix is a data set that only has two classes, one class as positive and the other class as negative. Consisting of four cells, namely True Positives (TP), False Positives (FP), True Negatives (TN) and False Negatives (FN).

Table 2. Confusion Matrix For 2 Class Models

CLASSIFICATION	PREDICTED CLASS		
		Class = YES	Class = NO
OBSERVED CLASS	Class = YES	***a*** **(True Positives – TP)**	***b*** **(False Negatives – FN)**
	Class = NO	***c*** **(False Possitives – FP)**	***d*** **(True Negatives – TN)**

To calculate the accuracy value using the formula [18]:

$$Accuracy = \frac{a + d}{a + b + c + d} = \frac{TP + TN}{TP + FN + FP + TN} \quad \ldots\ldots\ldots\ldots\ldots\ldots (6)$$

II. RESEARCH METHOD

Research methodology or stages are needed as a framework and guide the research process, so that the series of research processes can be carried out in a directed, orderly and systematic way. The following method proposed in this study, shown in Figure 1 below:

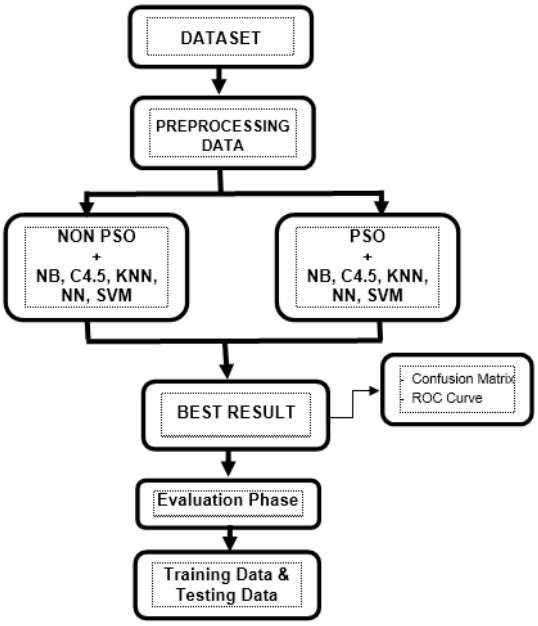

Figure 2. Proposed Method

A. *Data Collection*

This study uses Alumni data from the Department of Informatics Engineering and Computer Systems STMIK ASIA MALANG from the 2007 to 2011 batch of 1,064 records taken from BAAK. From these data only data from high school graduates as many as 845 were used as research with the reason for lectures starting from semester 1 (0 credits). Where as the data transfer from D1 / D2 / D3 was not included on the grounds that there were already several SKS that had been taken at the University before.

B. *Preprocessing Data*

There are 12 attributes carried out in this study, points 1 to 11 as determinant variables for graduation predictions and point 12 as targets or goal attributes, namely timely graduation, as listed in Table 3 below:

Table 3. Timely Attributes and Predictions of Graduation

NO.	ATTRIBUTE	INFORMATION
1	Gender	Male / Female
2	Age	17 – 37

3-9	Semester Achievment Index (1-7)	0,00 – 4,00
10	Student Employment Status	Not Yet Working
		Already Working
11	Marital status of students	Single
		Married
12	Graduation Prediction	Pass On Time
		Graduated Not On Time

Henceforth preprocessing techniques are carried out so that the quality of the data obtained is better by:

1. Data Validation, to identify and delete odd data (outlier / noise), inconsistent data, and incomplete data (missing value).
2. Data Dicretization, in the training data used in this study, selection of graduation prediction attributes is carried out by following the following rules:
   a. If the Study Duration is 3.5 - 4.5 years, it is said to graduate on time
   b. If the Study Time is > 4.5 years, it is said that Graduation is Not On Time

C. Modeling Phase

At this stage the training data processing stages are classified by the model to produce the highest accuracy value. In this phase the stages will be carried out:

1. Calculate the accuracy and AUC values of each algorithm (Naïve Bayes, Decision Tree (C4.5), k-NN and Neural Network, Support Vector Machines), and find which algorithm produces the highest accuracy and AUC values.
2. Calculate the accuracy and AUC values of each algorithm (Naïve Bayes, Decision Tree (C4.5), k-NN and Neural Network, Support Vector Machines) after adding the Particle Swarm Optimization (PSO) selection feature, and find which algorithm which produces the highest accuracy and AUC values.

D. Evaluation Phase

In this phase the best PSO-based algorithm is tested by dividing the amount of training and testing data as follows:

1. Data Training (Student Force 2007-2010) and Data Testing (Student Force 2011).
2. Training Data (Student Force 2007-2011) and Data Testing (Student Force 2011).
3. Data Training (Student Force 2007-2009) and Data Testing (Student Force 2010-2011).
4. Training Data (Students Force 2007-2011) and Data Testing (Students Force 2010-2011).

From the 4 training data testing and testing above, the one with the highest AUC accuracy and then determined as a model is selected.

III. RESULT

A. Initial Data Processing

At the stage of collecting the dataset used in this study came from the Alumni Data of the Department of Information Engineering and Computer Systems Force 2007-2011 STMIK ASIA Malang as many as 845 from pure high school graduates with 19 attributes. The original data is carried out preprocessing by filling in the blank data (Semester 1 to 7 Performance Index) due to inactive lectures (college leave) or not studying during the semester and only paying tuition fees, so the data is blank. The blank data is filled with the minimum value of each of the same attribute.

The above dataset is the initial data for which attributes have not been selected, then the selection of the required attributes is carried out, specifically for filling the objective attribute column (graduation prediction), the following references are used:

a. If the Study Period is 3.5 - 4.5 years, it is declared to be On Time
b. If the Study Time is > 4.5 years, it is declared Not Graduated On Time, finally a new dataset is obtained as in Tables 4 and 5.

Table 4. Dataset for Algorithm Analysis Naïve Bayes, C4.5 and k-NN

NO.	GENDER	AGE	ACHIEVEMENT INDEX							Employment Status	Marital Status	GRADUATION
			S.1	S.2	S.3	S.4	S.5	S.6	S.7			
1	Male	22	2,15	2,56	2,88	2,43	2,63	2,35	2,69	Not Yet	Not Yet	Not On Time
2	Male	20	2,28	1,78	2,30	1,50	1,50	0,69	0,31	Not Yet	Not Yet	Not On Time
3	Male	23	2,38	2,69	2,57	2,62	2,81	2,14	2,00	Not Yet	Not Yet	Not On Time
4	Male	20	2,40	1,94	2,17	1,33	2,61	1,76	1,25	Not Yet	Not Yet	Not On Time
5	Male	19	2,48	2,22	2,67	2,71	2,52	2,21	3,00	Not Yet	Not Yet	Not On Time
6	Male	20	2,53	2,62	2,55	2,63	2,47	2,50	3,22	Not Yet	Not Yet	Not On Time
7	Female	20	2,55	2,74	2,88	2,76	2,65	2,95	2,93	Not Yet	Not Yet	On Time
8	Male	20	2,55	2,70	2,90	2,48	0,53	1,97	1,89	Not Yet	Not Yet	Not On Time
9	Male	20	2,60	2,75	2,66	2,90	3,17	3,29	3,50	Not Yet	Not Yet	On Time
10	Female	20	2,60	2,81	2,71	2,88	2,69	2,86	2,88	Not Yet	Not Yet	Not On Time
.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.
843	Male	18	3,65	3,50	3,71	3,50	3,43	3,25	3,06	Not Yet	Not Yet	On Time
844	Male	18	3,65	3,23	3,65	3,39	3,61	3,59	2,91	Not Yet	Not Yet	On Time
845	Male	34	3,65	3,70	3,39	3,68	3,30	3,25	3,50	Already	Already	On Time

Table 5. Datasets for Analysis of Neural Network and SVM Algorithms

NO.	GENDER	AGE	ACHIEVEMENT INDEX							Employment Status	Marital Status	GRADUATION
			S.1	S.2	S.3	S.4	S.5	S.6	S.7			
1	1	22	2,15	2,56	2,88	2,43	2,63	2,35	2,69	1	1	Not On Time
2	1	20	2,28	1,78	2,30	1,50	1,50	0,69	0,31	1	1	Not On Time
3	1	23	2,38	2,69	2,57	2,62	2,81	2,14	2,00	1	1	Not On Time
4	1	20	2,40	1,94	2,17	1,33	2,61	1,76	1,25	1	1	Not On Time
5	1	19	2,48	2,22	2,67	2,71	2,52	2,21	3,00	1	1	Not On Time
6	1	20	2,53	2,62	2,55	2,63	2,47	2,50	3,22	1	1	Not On Time
7	2	20	2,55	2,74	2,88	2,76	2,65	2,95	2,93	1	1	On Time
8	1	20	2,55	2,70	2,90	2,48	0,53	1,97	1,89	1	1	Not On Time
9	1	20	2,60	2,75	2,66	2,90	3,17	3,29	3,50	1	1	On Time
10	2	20	2,60	2,81	2,71	2,88	2,69	2,86	2,88	1	1	Not On Time
.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.
843	1	18	3,65	3,50	3,71	3,50	3,43	3,25	3,06	1	1	On Time
844	1	18	3,65	3,23	3,65	3,39	3,61	3,59	2,91	1	1	On Time
845	1	34	3,65	3,70	3,39	3,68	3,30	3,25	3,50	2	2	On Time

B. Model Experiments and Testing

Calculation of 5 Non PSO Based Algorithms

Calculation of the accuracy value of 5 Non PSO-based Algorithms obtained results as in Tables 6 to 10.

Table 6. Calculation of the Naïve Bayes Algorithm

Accuracy : 63,79% +/-5,13% (mikro:63,79%)			
	True On Time	True Not On Time	Class Precision
Prediction On Time	320	232	57,97%
Prediction Not On Time	74	219	74,74%
Class Recall	81,22%	48,56%	

Table 7. Calculation of Algorithm C4.5

Accuracy : 64,02% +/-3,21% (mikro:64,02%)			
	True On Time	True Not On Time	Class Precision
Prediction On Time	245	155	61,25%
Prediction Not On Time	149	296	66,52%
Class Recall	62,18%	65,63%	

Table 8. Calculation of the k-NN Algorithm (k = 15)

Accuracy : 69,81% +/-5,83% (mikro:69,82%)			
	True On Time	True Not On Time	Class Precision
Prediction On Time	264	125	67,87%
Prediction Not On Time	130	326	71,49%
Class Recall	67,01%	72,28%	

Table 9. Calculation of Non PSO Neural Network Algorithm (Training Cycles 400)

Accuracy : 67,44% +/-5,06% (mikro:67,46%)			
	True On Time	True Not On Time	Class Precision
Prediction On Time	249	130	65,70%
Prediction Not On Time	145	321	68,88%
Class Recall	63,20%	71,18%	

Table 10. Calculation of Algorithm Support Vector Machines Non-PSO

Accuracy : 69,10% +/-4,75% (mikro:69,11%)			
	True On Time	True Not On Time	Class Precision
Prediction On Time	257	124	67,45%
Prediction Not On Time	137	327	70,47%
Class Recall	65,23%	72,51%	

Calculation of 5 Algorithms PSO Based

Calculation of the accuracy value of 5 Algorithms PSO Based obtained results as in Tables 12 to 16

Table 11. Calculation of Naïve Bayes Algorithm + PSO

Accuracy : 65,92% +/-1,85% (Mikro:65,92%)			
	True On Time	True Not On Time	Class Precision
Prediction On Time	328	222	59,64%
Prediction Not On Time	66	229	77,63%
Class Recall	83,25%	50,78%	

Table 12. Calculation of C4.5 Algorithm + PSO

Accuracy : 69,23% +/-0,04% (mikro:69,23%)			
	True On Time	True Not On Time	Class Precision
Prediction On Time	228	94	70,81%
Prediction Not On Time	166	357	68,26%
Class Recall	57,87%	79,16%	

Table 13. Calculation of the k-NN Algorithm (k = 19) + PSO

Accuracy : 74,08% +/-0,62% (Mikro:74,08%)			
	True On Time	True Not On Time	Class Precision
Prediction On Time	277	102	73,09%
Prediction Not On Time	117	349	74,89%
Class Recall	70,30%	77,38%	

Table 14. Calculation of Neural Network Algorithm (Training Cycles 100) + PSO

Accuracy : 72,55% +/-1,62% (Mikro:72,54%)			
	True On Time	True Not On Time	Class Precision
Prediction On Time	252	90	73,68%
Prediction Not On Time	142	361	71,77%
Class Recall	63,96%	80,04%	

Table 15. Calculation of Support Vector Machines Algorithms + PSO

Accuracy : 70,89% +/-4,75% (mikro:70,90%)			
	True On Time	True Not On Time	Class Precision
Prediction On Time	268	120	69,07%
Prediction Not On Time	126	331	72,43%
Class Recall	68,02%	73,39%	

IV. DISCUSSION

Based on Tables 6 to 10, from 5 Non PSO Based Algorithms it is known that the k-Nearest Neighbor (k = 15) algorithm produces the best test with an accuracy value of 69.81% and Area Under the Curve (AUC) of 0.764, as shown in Table 16 below.

Table 16. Recapitulation of Test Results 5 Non PSO Algorithms

Type of Testing	Naive Bayes	Decision Tree (C4.5)	k-NN (k=15)	NN	SVM
Accuracy	63,79 %	64,02 %	**69,81 %**	67,44 %	69,10 %
Precision	58,18 %	61,64 %	68,09 %	66,55 %	67,45 %
Recall	81,22 %	62,16 %	66,97 %	63,22 %	65,23 %
AUC	0,736	0,633	**0,764**	0,745	0,754

Based on Tables 11 to 15, from 5 PSO Based Algorithms it is known that the k-Nearest Neighbor (k = 19) algorithm produces the best test with an accuracy value of 74.08% and Area Under the Curve (AUC) of 0.788, as shown in Table 17 below.

Table 17. Recapitulation of Test Results for 5 Algorithms PSO Based

Type of Testing	Naive Bayes	Decision Tree (C4.5)	k-NN (k=19)	NN	SVM
Accuracy	65,92 %	69,23 %	**74,08 %**	72,55 %	70,89 %

Precision	59,77 %	71,56 %	73,12 %	74,62 %	69,07 %
Recall	83,25 %	57,87 %	70,30 %	63,96 %	68,02 %
AUC	0,721	0,687	**0,788**	0,780	0,778

A. Increased Accuracy Value

Based on Table 16 and Table 17 finally obtained a count of increasing the value of accuracy as in Table 18 below:

Table 18. Increase in Accuracy Value

ALGORITHM	NON PSO	+ PSO	Increased Accuracy Values
Naive Bayes	**63,79 %**	**65,92 %**	**2,13 %**
k-NN	**69,81 %**	**74,08 %**	**4,27 %**
Neural Network	**67,44 %**	**72,55 %**	**5,11 %**
Decision Tree (C4.5)	**64,02 %**	**69,23 %**	**5,21 %**
Support Vector Machines	**69,10 %**	**70,89 %**	**1,79 %**

From Table 18 the results obtained increase the highest accuracy value lies in the Decision Tree (C4.5) Algorithm of 5.21%, the lowest on the Vector Machines Support Algorithm of 1.79%. In the non PSO Algorithm, the k-NN Algorithm becomes the best algorithm with the highest accuracy value of 69.81%, likewise after the addition of the PSO feature, the highest accuracy value is also 74.08%, although the increase in accuracy is lower than the C4 Algorithm. 5 (5.21%) or Neural Network Algorithm (5.11%). The highest increase in accuracy value is not necessarily a PSO based algorithm which has the highest accuracy value, this is influenced by the accuracy value before the addition of PSO features.

B. Determination of the Best Algorithm

From Table 17 above, it can be concluded that the PSO-based k-NN algorithm produces the best test with the highest accuracy value of 74.08% (at k-optimum = 19) and the Area Under the Curve (AUC) value = 0.788. To make it easier to read the above data, here is a graph display of accuracy and AUC values:

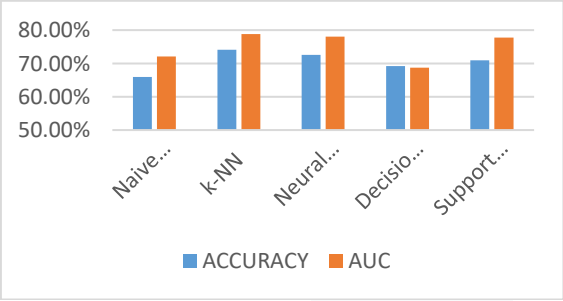

Figure 3. Accuracy Value Graph and AUC

From the graph in Figure 3 above, it is known that the order of accuracy values from lowest to highest is the Naive Bayes Algorithm (NB), Decision Tree (C4.5), Support Vector Machines (SVM), Neural Network (NN) and k-NN.

To determine the weighting of each attribute, it is applied to the k-NN + PSO algorithm (k = 19) as the algorithm that provides the best accuracy value. Table 19 shows the weighting values of each attribute, where the attributes: Age, Semester Achievment Index 3, Semester Achievment Index 5 and Marriage Status did not make a significant contribution to the timely graduation of students. Attributes: Gender, Semester Achievment Index 1, Semester Achievment Index 2, Semester Achievment Index 6, Semester Achievment Index 7 and Employment Status make a real contribution to the timely graduation of students. The Semester Achievment Index 4 attribute continued to contribute to graduation on time even though it was only 0.682.

Table 19. Attribute Weighting of the k-NN Algoritm + PSO

Attribute	Weighting
Gender	1
Age	0
Semester Achievment Index 1	1
Semester Achievment Index 2	1
Semester Achievment Index 3	0
Semester Achievment Index 4	0,682
Semester Achievment Index 5	0
Semester Achievment Index 6	1
Semester Achievment Index 7	1
Student Employment Status	1
Marital status of students	0

C. Evaluation Phase

Based on the results of the discussion above, the PSO-based k-NN algorithm is the algorithm that has the best accuracy value, then the evaluation phase is tested using the PSO-based k-NN algorithm (k = 19, number of validation 10) by dividing the amount of training data and testing data as follows:

1. Training Data (Student Force 2007-2010 = 790 data) and Data Testing (Student Force 2011 = 55 data) obtained counts:

Table 20. Calculation of Training Data for Force 2007-2010 and Data Testing for Force 2011

Accuracy : 98,18% +/-0,00% (Mikro:98,18%)			
	True On Time	**True Not On Time**	**Class Precision**
Prediction On Time	**540**	**0**	**100,00%**
Prediction Not On Time	**10**	**0**	**0,00%**
Class Recall	**98,18%**	**00,00%**	

From Table 20 above, 54 (= 540/10) students are predicted to graduate on time with the reality of graduating on time, at least there are 54 students who

are in accordance with the predictions of graduating on time. There are 1 (10/10) students who are predicted to graduate not on time with the reality of graduating on time, meaning there is 1 student who predicts graduation is not on time. By using Force Training Data for 2007-2010, as many as 790 data were able to predict graduates of 2011 Force students with an accuracy value of 98.18% (close to 100%).

2. Training Data (Student Force 2007-2011 = 845 data) and Data Testing (Student Force 2011 = 55 data) obtained counts:

Table 21. Calculation of Training Data for Force 2007-2011 and Data Testing for Force 2011

Accuracy : 98,18% +/-0,00% (Mikro:98,18%)			
	True On Time	**True Not On Time**	**Class Precision**
Prediction On Time	**540**	**0**	**100,00%**
Prediction Not On Time	**10**	**0**	**0,00%**
Class Recall	**98,18%**	**00,00%**	

With the same calculation as points 1 and 2 above, finally obtained several accuracy values with different training data and testing data, the results can be seen in Table 22 below:

Table 22. Comparison of Training Data and Data Testing Accuracy Values

NO.	TRAINING DATA	TESTING DATA	ACCURACY
1	2007-2010 (790 data)	2011 (55 data)	**98,18%**
2	2007-2011 (845 data)	2011 (55 data)	**98,18%**
3	2007-2009 (565 data)	2010-2011 (280 data)	**73,89%**
4	2007-2011 (845 data)	2010-2011 (280 data)	**82,25%**

Table 22 shows the results of the best accuracy value is in the Data Testing Force of 2011 (55 data) that is equal to 98.18%, both with Force Training Data 2007-2010 (790 data) and Data Training Force 2007-2011 (845 data), meaning prediction On-time graduation from the Class of 2011 is close to 100% with actual graduation data. The accuracy value will be better if you use the whole training data (845 data) both for the 2011 Force Testing Data or the 2010-2011 Force Testing Data, meaning that to predict the next generation of students it is better to use the whole training data, this can be seen comparing the accuracy value at the 3 is smaller than point 4 by using the same testing data (Force 2010-2011).

V. CONCLUSION

The conclusions obtained from this study are:

a. The results of the analysis using Confusion Matrix and ROC Curve can be concluded that the k-Nearest Neighbor Algorithm Based on PSO has the best performance for the classification of work fields with an Accuracy value of 74.08% and the value of Area Under The Curve (AUC) = 0.788. Attributes of Gender, Semester Achievment Index 1, 2, 4, 6 and 7 as well as Employment Status make a real contribution to the right graduation of students' time.
b. The addition of the PSO feature always increases the accuracy value, which increases the highest accuracy value in the Decision Tree Algorithm (C4.5) by 5.21%, the lowest in the Algorithm, Vector Engine Support by 1.79%.
c. Increasing the accuracy value of the k-NN Algorithm is located in the third order, remains the best value added, this value increases the value added before increasing the PSO feature to the highest of the 5 existing algorithms.
d. For the evaluation phase, the results of the accuracy are much better if using the overall data training (Angaktan 2007-2011), both with the 2011 Force test data or the 2010-2011 Force.

REFERENCES

[1] Prabowo. (2012). Aneka Teknik, Piranti dan Penerapan Data Mining: Studi Kasus Peramalan Harga Saham Industri Telekomunikasi Berbasis Jaringan Saraf Tiruan. Modul Perkuliahan Universitas Budi Luhur.

[2] Nuqson Masykur Huda. (2010). "Aplikasi Data Mining Untuk Menampilkan Informasi Tingkat Kelulusan Mahasiswa", Semarang.

[3] Asif R., Agathe M., Mahmood KP. (2015). *Predicting Student Academic Performance at Degree Level: A Case Study. I.J. Intelligent Systems and Applications*, 01, 49-61. Published Online December 2014 in MECS (http://www.mecs-press.org/). DOI: 10.5815/ijisa.2015.01.05.

[4] Ayu, Mutiara B.,H.Irwan Budiman and Andi Farmadi. (September 2015). Penerapan K-Optimal Pada Algoritma k-NN untuk Prediksi Kelulusan Tepat Waktu Mahasiswa Program Studi Ilmu Komputer FMIPA UNLAM Berdasarkan IP Sampai Dengan Semester 4. Kumpulan jurnaL Ilmu Komputer (KLIK) ISSN: 2406-7857. Volume 02, No.02.

[5] Handjaratie,Lillyan. (2015). *Prediction And Data Mapping of Students Of Engineering Faculty*, Universitas Negeri Gorontalo Using Data Mining.

[6] Nursalim, Suprapedi and H.Himawan. (April 2014). Klasifikasi Bidang Kerja Lulusan Menggunakan Algoritma K-Nearest Neighbor. Jurnal Teknologi Informasi, ISSN 1414-9999. Volume 10 Nomor 1.

[7] Anuradha,C., T.Velmurugan. (July 2015). *A Comparative Analysis on the Evaluation of Classification Algorithms in the Prediction of Students Performance. Indian Journal of Science and Technology*, Vol 8(15), DOI: 10.17485/ijst/2015/v8i15/74555, ISSN (Print) : 0974-6846. ISSN (Online) : 0974-5645,.

[8] Azwar, S. Penyusunan Skala Psikologi. Yogyakarta: Pustaka Pelajar, 2004.

[9] Hanief Muhamad M., Metri Annisa, Narendi Muhandri and Kadarsyah Suryadi. Prediksi Masa Studi Sarjana Dengan Artificial Neural Network. Internetworking Indonesia Journal. Vol.1/No.2. 2009.

[10] Kusrini, Luthfi, E.T. "Algoritma Data Mining", Andi Offset. Surabaya, 2009.

[11] Jefri. "Implementasi Algoritma C4.5 Dalam Aplikasi Untuk Memprediksi Jumlah Mahasiswa Yang Mengulang Mata Kuliah Di STMIK AMIKOM Yogyakarta", Yogyakarta, 2013.

[12] Rizal, Azwar. (2013). Perbandingan Performa antara Imputasi Metode Konvensional dan Imputasi dengan Algoritma Mutual Nearest Neighbor. Jurusan Teknik Informatika, Fakultas Teknologi Insformasi, Institut Teknologi Sepuluh Nopember. Surabaya.

[13] Rajasekaran S., GA. Vijayalakshmi Pai. “Neural Network, Fuzzy Logic and Genetic Algorithms”, Prentice-Hall of India, New Delhi, 2005.
[14] Prasetyo, E. (2012). Data Mining: Konsep dan Aplikasi Menggunakan Matlab. Indonesia: Andi Yogyakarta.
[15] Yin,Y., Han, D., & Cai,Z. (2011). Explore Data Classification Algorithm Based on SVM and PSO for Education Decision. Journal of Convergence Information Technology, 6(10), 122-128.
[16] Li, G., You, J., & Liu, X. (2015). Support Vector Machine (SVM) based prestack AVO inversion and its applications. Jounal of Applied Geophysics, 129, 60-68.
[17] S.-W. Fei, Y.-B. Miao, and C.-L. Liu. *"Chinese Grain Production Forecasting Method Based on Particle Swarm Optimization-based Support Vector Machine"*, in Recent Patents on Engineering, Shanghai, pp. 8-12. 2009
[18] Bramer, Max. (2007). *Principles of Data Mining*. London: Springer.
[19] Vercellis, Carlo. (2009). Business Intelligent: *Data Mining and Optimization for Decision Making*. Southern Gate, Chichester, West Sussex: John Willey & Sons, Ltd.
[20] Gorunescu, Florin. Data Mining: Concepts and Techniques. Verlag berlin Heidelberg: Springer, 2011.