Bit-Tech

Vol.3, No.1, August 2021
Available online at: http://jurnal.kdi.or.id/index.php/bt

Junior Class Preparedness Classification Faces A National Exam Using A C.45 Algorithm With A Particle Swarm Optimization Approach

Asep Suherman[1)*], Didi Kurnaedi[2)], Sofian Lusa[3)], Rizqi Darmawan[4)]

[1)4)]*Universitas Budi Luhur*
Jl. Cileduk Raya, Petukangan Utara, Jakarta Selatan, Indonesia
[1)]asepchepster@gmail.com
[4)]rizqidarmawan95@gmail.com

[2)]*STMIK PGRI*
Jl. Perintis Kemerdekaan II, Tangerang, Indonesia
[2)]ddk@pgri.id

[3)]*Universitas Indonesia*
Kampus Baru UI Depok, Jawa Barat, Indonesia
[3)]sofian.lusa12@ui.ac.id

Article history:

Received June 16, 2020;
Revised July 3, 2020;
Accepted July 8, 2020;
Available online August 31, 2020

Keywords: {use 4-6 keywords}

Classification
Students
National exam
C4.5
Swarm particle optimization
Try out

Abstract

These studies are counter to a trend of falling students' graduation rates on the national exam. This is because of the way students prepare their readiness to face national tests is inaccurate. On this study the hybrid method c4 algorithm.5 and the swarm particle optimization to produce a class readiness of students with high and accurate accuracy. This research suggests that by using hybridmethodC4.5 andParticle Swarm Optimizationgenerates accuracy as 97.13 %, Precisionas 96,58 %, andRecallas 100 %. Then implemented through a web-based prototype application using programming javascriptlanguage.

I. INTRODUCTION

The national examination became one of the high school students' graduation requirements for completing their high school years. Thus, many of the junior high school students took on additional study activities at the student-tutoring society to prepare for the national test they would face. At the first high school student tutoring institute getting input and trying out national exams to measure students' abilities.

After performing try out national exams, study guides will evaluate the results of try out students. Then after being evaluated the study guides did a classification of the students' results. This classification is helpful to determine the steps the study will take on the students' abilities based on the results of the national try out exam. But in fact with the student classification process which is walking now graduation rates are less students.

Therefore it takes a new class of students to produce a proper classification of students and high accuracy. Based on the problem with this research it will be proposed a new class classification by which the student classification USES a c-4 method.5 combined with a particle particle swarm optimization algorithm.

Based on a similar study carried out using a classification algorithm c4.5 for the classification of a carefully intelligent competitor, it obtained as much accuracy 81.81%, Hence the use of a c-4 method.5 and the particle swarm optimization for a class of students preparedness in the face of a national test with a student object studying ganesha operation[2].

* Corresponding author

ISSN 2622-2728 (online) 2622-271X (print)
doi:

II. RELATED WORKS/LITERATURE REVIEW (OPTIONAL)

1. Data *Mining*

Data mining refers to the knowledge of a large amount of data [7]. Almost all this data is fed by the computer applications used for handling everyday transactions mostly OLTP (Online Transaction Processing). As for the step in the data mining [16] is as follows:

a. Selection results data which will be used for the data-mining process, stored in a file, separated from the operational database.
b. Needs cleaning on the data that becomes focus knowledge discovery in discovery (KDD). The cleaning process includes removing duplicates of data examining inconsistencies, and fixing errors in the data, such as typography.
c. Data is altered or merged into a suitable format for processing in the data mining
d. Data mining is the process of finding patterns or interesting information in the selected data using a particular technique or method.
e. The information patterns resulting from the data mining process need to be displayed in a form easily understood by the concerned parties.

2. Classification

Classification is part of a data-mining technique based on machine learning that classifies one item into one set of data to another set of data [4]. Classification is one of the techniques in the data mining. The classification (taxonomy) is a process of placing certain objects or concepts into a set of categories based on objects used. A classification is a technique by looking at the behaviors and attributes of a defined group. This technique can give a new classification of data by manipulating existing data and using the results to come up with a number of rules. These rules are used on new data to be classified. This technique uses a comprehensive induction, which employs a collection of tests from a classified record to determine classes [8].

The purpose of classification is to:

a. Found a model of the training data that distingucts record into a corresponding category or class, the model is then used to classify the record, which the class has not previously known at testing set.
b. Making decisions by predicting a case, based on the results of classified data obtained.

3. *DecisionTree*

Conceptually decision tree is one of the decision analysis techniques. The trie itself was first introduced in the 1960's by fredkin. Trie or digital tree is derived from retrival as it is intended. Etymology of this is pronounced as' tree '. Although it is similar to the use of the word 'try' but it is aimed to distinguish it from the general tree [10]. In computer science, trie, or prefix tree is a data structure with a representation of associative tree used to store the associative array that's string. Decision trees are often used in classifications and predictions. This tree of decision is simple but is the way of representation of good knowledge [13].

The decision tree is also useful for exploring the data, finding the hidden connections between a number of potential input variables and a target variable. The decision tree combines between data exploration and modeling, so it's a very good first step in the modeling process even when made the ultimate model of some other techniques [9]. Stage of decision tree:

a. The construction
   Of this stage of the tree begins with the formation of roots (located at the top). Then the data is broken using attributes suitable for use as a sheet.
b. Tree trimming
   Identifies and discards unnecessary branches on established trees. This is because decisions made by trees can be large so that they can be simplified with pruning by the values of trustworthiness. Tree planting was done in addition to reducing tree size to reduce the rate of bad predictions in new cases of split solution and solution. Inequality there are two approaches:
   1) Pre-pruning is to stop building a subtree early (by deciding not to further partition the data training). When it ceases, the nodes turn into the last leaf. This latest node becomes the most common class among the subsets of samples.
   2) Postpruning, that is, simplified the tree by disposing of a few subtree branches after the tree was built. The uninterrupted node will be the most frequent of the classes.
   3) Forming decision rules makes decision-making rules out of established trees. They can be in the form of if-then derived from the decision tree by tracing from the root to the leaf. For each node and branch, if determined, then the value sheet is put. Once all rules are set, rules can be simplified or combined.

4. C4.5 Algorithm

C4.5 algorithmA group of decision tree algorithms. These algorithms have input in the form of training operations and grading tests. Training projects of sample data that will be used to build a tree that has been authenticated. Whereas Christianity is field-field data that we will later use as a parameter in administering data classification [11]. Steps to form a decision tree with a c4 algorithm.5 which is [1]:

a. prepares training data drawn from historical data or past data made into specific classes.
b. counts the roots of a tree. Choosing attributes as roots is based on the highest gain value of available attributes. Before calculating the gain value of attributes, count first the entropy value. Good choice of attributes is a enabling attribute to get the smallest decision tree to scale. Or attributes that could separate objects according to their classes. Heuristic attributes chosen are those that produce the most "cleanness" (cleanest) knots. The measure of purity is manifest with the degree of purity, and to calculate it, it can be done with the concept of entropy, entropy states the impurity of a mass of objects. The formula for calculating entropy value is as follows.

Entropy (S) = $\sum_{i=1}^{n} -$ $p_i \log_2 (p_i)$
Information:
S = Case assembly
N = Amount of partition S
p_i = The proportion of S_i as S

c. Information gain is one of the astonishing selection measures that's used to select the ratio of each node on the tree. Attributes with the highest gain information selected as a test of the attributes of a node. In the process, Mr. Gain calculations can happen or missing value. Calculating the value gain with the equation:

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

S : Case assembly
A : atribute
n : Amount of partitionA
|Si| : The number of cases on the partition to -i
|S| : Amount of partitionS

d. Of attributes as root being derived from step 3 is created a branch for each value.
e. At each branch that has not pointed to a particular class was repeated from step 1 to step 3 until all branches have pointed to a class and the process is complete.

5. Particle Swarm Optimization

Particle Swarm Optimization was introduced by Dr. Eberhart and Dr. Kennedy in 1995, is an optimization algorithm that mimics the processes that occur in the life of a bird population (flock of bird) and fish (school of fish) in survival [3]. Since it was first introduced, the PSO algorithm has developed quite rapidly, both in terms of application and in terms of the development of methods used in the algorithm [5]. Because of this, they categorize algorithms as part of artificial life / artificial life [6]. This algorithm is also connected with evolutionary computing, genetic algorithms and evolutionary programming [12]

The PSO algorithm contains some of the following processes (muhammad et al., 2017):

a. Initialize
   1) Initialize a. Initialize initial velocity. On the 0th iteration, it's certain that the initial velocity of all particles is 0.
   2) Initialize the initial position of particles. On the 0th iteration, the initial position of particles was revived with equation:

$$x = x_{min} + rand[0,1] \; x \; (x_{max} - x_{min})$$

   3) Initialize pbest and gbest. In the 0th iteration, the pbest will be equated with the initial particle position value. While the gbest is chosen from one pbest with the highest fitness.

b. Update speed

To do speed updates, used the following formula:

$$v_{i,j}^{t+1} = w\ v_{i,j}^{t} + c_1.r_1\left(Pbest_{i,j}^{t} - x_{i,j}^{t}\right) + \ c_2.r_2\left(Gbest_{g,j}^{t} - x_{i,j}^{t}\right)$$

Information:
Vij = Components of individual speed to the I on d dimension
ω = parameter *inertiaweight*
*c*1 *c*2 = Learning rate, values between 0 and 1
r1, r2 = Random parameters between 0 to 1
Pbestij= Pbest (*localbest*) Individuals I in j dimensions
Gbestij= Gbest (*globalbest*) In j dimensions

c. Update the position and calculate fitness

To update the position, use the following formula:

$$x_{i,j}^{t+1} = x_{i,j}^{t} + v_{i,j}^{t+1}$$

Where:
Xij = position of individual i on d dimensions

d. Update pBestandgBest

A comparison between pBest in the previous iteration and the results of the position update. Higher fitness will be the new pBest. The latest pBest which has the highest fitness value will be the new gBest.

III. METHODS

The research steps can be seen in the research structure below.

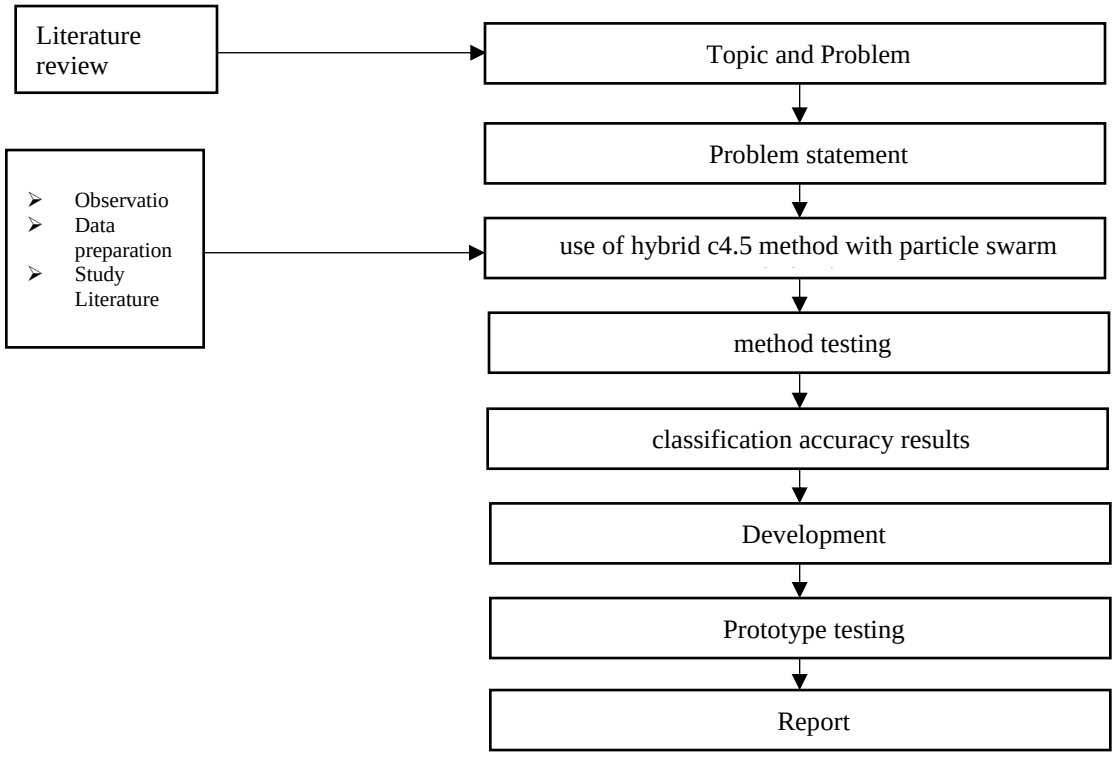

Figure III-1 Research Steps

1. Determination of Topics and Identification of Problems
   The first step in this research is to conduct a literature study to determine the topic to be used, after that the problem is identified so that in this research it is clear what will be done.

2. Formulation of the Problem
   At this stage the formulation of the problems that will be examined in this research is carried out, so that it is clear what problems will be solved in this study.
3. Use of hybrid C4.5 method with Particle Swarm Optimization
   The use of hybrid C4.5 method with Particle Swarm Optimization was chosen based on the results of literature studies that have been carried out, with the hope that the method will produce accuracy and classification according to needs.
4. Testing Method
   The method used is then tested in order to determine the level of success of the method used.
5. Classification Accuracy Results
   At this stage, the accuracy of the method that has been through the testing process is produced.
6. Making Prototype Application
   Making an application based on the method that has been used. So the result of the prototype application is the implementation of the algorithm used.
7. Application Prototype Testing
   After the prototype application is complete, then the application is tested to ensure the application runs smoothly and there are no bugs.

IV. RESULTS

A. Data Training Preparation

In this study the data that will be used is the data of the try out results of Ganesha Operation Middle School student tutoring for the last 3 years. During the 3 year period, 1,737 records of try out data were obtained from 1737 students. From these data, attributes will be selected to be used in the process of calculating the C4.5 algorithm and Particle Swarm Optimization. These attributes are as follows:

Table IV-1 Selection of attributes

KK* MATA PELAJARAN				**KK* RATA NILAI**	**LULUS / TDK LULUS**
IND	**ING**	**MAT**	**IPA**		
A	A	A	A	Sangat Baik	LULUS / TDK LULUS
B	B	B	B	Baik	
C	C	C	C	Cukup	
D	D	D	D	Kurang	

Information:
IND : Bahasa Indonesia
ING : Bahasa Inggris
MAT : Matematika

From this table, it is explained that the attributes to determine graduation are in the form of subject completeness criteria which are divided into IND, ING, MAT, and Natural Sciences. And the criteria for completeness criteria average value. Then the amount of data and the grouping of SMP try out results are calculated as follows:

Table IV-2 Number of data from junior high school try out results

		JML SISWA	**TDK LULUS (S1)**	**LULUS (S2)**
TOTAL		1737	327	1410
IND				
	A	85	3	82
	B	582	73	509
	C	795	155	640

		JML SISWA	TDK LULUS (S1)	LULUS (S2)
	D	275	96	179
ING				
	A	357	37	320
	B	918	125	793
	C	394	109	285
	D	68	56	12
MAT				
	A	30	2	28
	B	294	12	282
	C	803	65	738
	D	610	248	362
IPA				
	A	10	0	10
	B	246	10	236
	C	865	92	773
	D	616	225	391
RATAAN NILAI				
	Kurang	190	190	0
	Cukup	1114	137	977
	Baik	421	0	421
	Sangat Baik	12	0	12

From this table, information was obtained that from 1737 there were 327 students who did not graduate and 1410 students who passed. Then the number is broken down again based on predetermined attributes. After grouping the data, it then enters the hybrid processing method of the C4.5 algorithm with Particle Swarm Optimization.

B. Hybrid Method Method C4.5 Algorithm & Particle Swarm Optimization

In processing using the C4.5 algorithm the following classification results are obtained:

```
KK* RATA NILAI = Baik: LULUS {TDK LULUS=0, LULUS=421}
KK* RATA NILAI = Cukup
|  KK ING = A
|  |  KK IPA = B
|  |  |  KK MAT = C: LULUS {TDK LULUS=0, LULUS=1}
|  |  |  KK MAT = D
|  |  |  |  KK IND = B: LULUS {TDK LULUS=1, LULUS=3}
|  |  |  |  KK IND = C: TDK LULUS {TDK LULUS=2, LULUS=2}
|  |  |  |  KK IND = D: TDK LULUS {TDK LULUS=1, LULUS=0}
|  |  KK IPA = C: LULUS {TDK LULUS=5, LULUS=73}
|  |  KK IPA = D
|  |  |  KK IND = B
|  |  |  |  KK MAT = B: TDK LULUS {TDK LULUS=2, LULUS=0}
|  |  |  |  KK MAT = C: LULUS {TDK LULUS=3, LULUS=11}
```

| | | | KK MAT = D: TDK LULUS {TDK LULUS=7, LULUS=7}
| | | KK IND = C: LULUS {TDK LULUS=7, LULUS=34}
| | | KK IND = D: LULUS {TDK LULUS=2, LULUS=9}
| KK ING = B
| | KK MAT = B: LULUS {TDK LULUS=3, LULUS=36}
| | KK MAT = C: LULUS {TDK LULUS=25, LULUS=309}
| | KK MAT = D
| | | KK IND = A: LULUS {TDK LULUS=0, LULUS=10}
| | | KK IND = B
| | | | KK IPA = B: TDK LULUS {TDK LULUS=3, LULUS=2}
| | | | KK IPA = C: LULUS {TDK LULUS=9, LULUS=31}
| | | | KK IPA = D: LULUS {TDK LULUS=7, LULUS=29}
| | | KK IND = C: LULUS {TDK LULUS=23, LULUS=108}
| | | KK IND = D: LULUS {TDK LULUS=3, LULUS=39}
| KK ING = C: LULUS {TDK LULUS=19, LULUS=263}
| KK ING = D
| | KK IPA = C
| | | KK IND = A: TDK LULUS {TDK LULUS=3, LULUS=0}
| | | KK IND = B: TDK LULUS {TDK LULUS=12, LULUS=1}
| | | KK IND = C: LULUS {TDK LULUS=0, LULUS=3}
| | | KK IND = D: LULUS {TDK LULUS=0, LULUS=2}
| | KK IPA = D: LULUS {TDK LULUS=0, LULUS=4}
KK* RATA NILAI = Kurang: TDK LULUS {TDK LULUS=190, LULUS=0}
KK* RATA NILAI = Sangat Baik: LULUS {TDK LULUS=0, LULUS=12}

Then the results are described through the decision tree as follows.

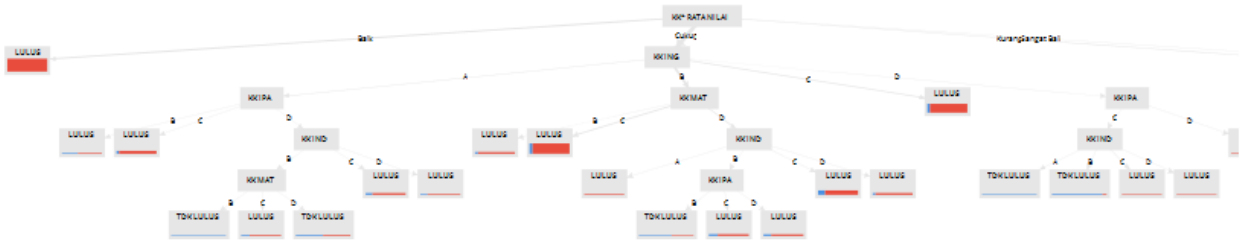

Figure2. Decision Tree

After these steps are taken, the next step is to process optimization with the Particle Swarm Optimization algorithm using the rapid miner application. The design that was made to process the Particle Swarm Optimizaton algorithm can be seen in the image below.

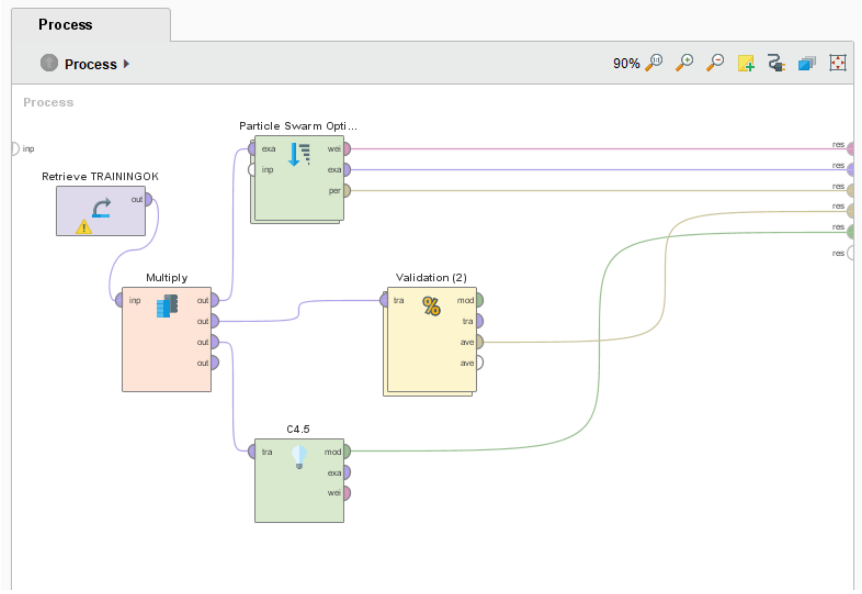

Figure IV-1 Design processes for rapid miners

Then the accuracy testing is done using a confusion matrix model to find out how accurate the resulting classifications are.

Table View Plot View

accuracy: 92.53%

	true TDK LULUS	true LULUS	class precision
pred. TDK LULUS	20	0	100.00%
pred. LULUS	13	141	91.56%
class recall	60.61%	100.00%	

Figure IV-2 Confusion Matrix Model C4.5

To calculate the accuracy of the confusion matrix above the process is as follows:

Table IV-3Confusion matrix

		Kelas Prediksi	
		Negatif	Positif
Kelas Sebenarnya	Negatif	a	b
	Positif	c	d

Information:

a. if the predicted result is negative and the actual value is negative.
b. if the prediction result is positive while the actual value is negative.
c. if the prediction result is negative while the actual value is positive.
d. if the prediction result is positive and the actual value is positive.

Precision is the proportion of cases with true positive results.

Precison = d / (d + c)
= 141 / (141 + 5)
= 96.58 %

Recall is the proportion of positive cases that are correctly identified.
Recall = d / (d + b)
= 141 / (141 + 0)
= 100 %

Accuracy is the proportion of cases identified correctly to the total number of all cases.

Accuracy = (a+d)/(a+b+c+d)
= (28 + 141) / (28 + 5 + 0 + 141)
= 97.13 %

Implementation of Application Prototype

At this stage an application prototype is created by coding javascript based on the hybrid method C4.5 and Particle Swarm Optimization which will then produce the following application.

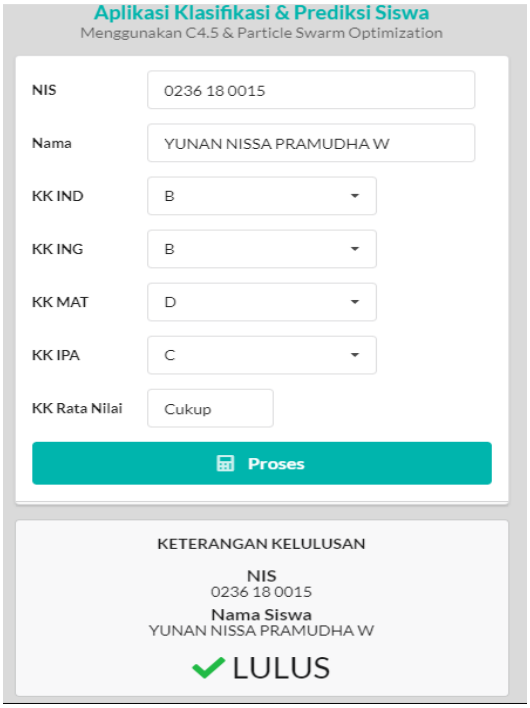

Figure IV-3 User Interface Prototype Application Using C4.5 and PSO Applications

This application prototype serves as a prediction of students' readiness to face the National Examination by using the classification of data that has been previously processed. Where in this application prototype the user simply needs to enter the NIS, as well as other attributes needed to know whether the student can pass or not on the readiness of the National Examination that will be faced.

V. DISCUSSION

The suggestions for further research are as follows

1. For tutoring management it is recommended to use this research as a classification method used to classify students' readiness in facing the National Examination in order to obtain a classification with high accuracy.
2. It is technically recommended for researchers to further add attributes that affect the National Examination in hopes of increasing the accuracy of classifying students' readiness with the hybrid method.
3. From the software development side, it is recommended to develop an application prototype into a whole application that uses a database and a mobile version.
4. For further research, especially for tutoring institutions or other academic practitioners, it is expected to be able to perform another hybrid method algorithm to produce higher classification accuracy compared to the results of this study.

VI. CONCLUSIONS

Based on the description, explanation, and testing that have been carried out, found several conclusions as follows:

1. This study resulted in the classification of student readiness using the hybrid method algorithm C4.5 and Particle Swarm Optimization with an accuracy level of 97.13% with a precision value of 96.58% and a 100% recall in its classification.
2. The prototype of the classification application and prediction of students' readiness to face the National Examination are used to accelerate the delivery of solutions to students who are predicted to pass and not pass to increase the number of graduations.

REFERENCES

[1] Anam, C. dan Santoso, H. B. (2018) "Perbandingan Kinerja Algoritma C4 . 5 dan Naive Bayes untuk Klasifikasi Penerima Beasiswa," Jurnal Ilmiah Ilmu-Ilmu Teknik, 8(1), hal.13–19.

[2] Ardiansyah, D. dan Walim, W. (2018) "Algoritma c4.5 untuk klasifikasi calon peserta lomba cerdas cermat siswa smp dengan menggunakan aplikasi rapid miner," Jurnal Inkofar, 1(2), hal.5–12.

[3] Chen& Shih (2013) "Solving University Course Timetabling Problems Using Constriction Particle Swarm Optimization with Local Search," Algorithms 2013, 6, 227-244; doi:10.3390/a6020227

[4] Engelbrecht (2006) "A study of particle swarm optimization particle trajectories," Information Sciences 176 (2006) 937–971

[5] Haupt RL, Haupt SE: Practical Genetic Algorithms. seconded edition. New Jersey; 2004.

[6] Hsieh et al. (2007) "Cross-Searching Strategy for Multi-objective Particle Swarm Optimization," IEEE Congress on Evolutionary Computation (CEC 2007)

[7] JiaweiHan ,Micheline Kamber, Data Mining: Concepts and Techniques, 2nd edition, 2006.

[8] JianweiHan (2001) " CMAR: accurate and efficient classification based on multiple class-association rules." Proceedings 2001 IEEE International Conference on Data Mining DOI: 10.1109/ICDM.2001.989541

[9] KusrinidanLuthfi, E. T. 2009. Algoritma Data Mining.Edisi 1.AndiOffset.Yogyakarta.

[10] McLachlan · Angus Ng · Bing Liu · Philip S. Yu ·Zhi-Hua Zhou · Michael Steinbach · David J. Hand · Dan Steinberg Top 10 algorithms in data mining © Springer-Verlag London Limited 2007

[11] Quinlan J.R., (1986).Induction of Decision Trees, Machine Learning, pp81-106.

[12] Rajesh, K., and Sheila Anand. "Analysis of SEER dataset for breast cancer diagnosis using C4. 5 classification algorithm." International Journal of Advanced Research in Computer and Communication Engineering 1.2 (2012): 2278-1021

[13] Surjeet K. Y., Saurabh P., (2012).Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification" WCSIT,ISSN: 2221-0741 Vol. 2, No. 2, 51-56