International Journal of Engineering and Applied Technology (IJEAT)
Vol. 2 | November 2019
ISSN : 2620-9632

COMPARISON C4.5 AND NAÏVE BAYES METHODS BASED ON PARTICLE SWARM OPTIMIZATION IN LEVELS OF DROP OUT STUDENTS

Dudih Gustian[1]*, Faridatun Ni'mah[1], Agus Darmawan[2]
[1]Departemen Information System, Nusa Putra University, Indonesia
[2]Departemen Mechanical Engineering, Nusa Putra University, Indonesia
*Email: dudih@nusputra.ac.id

Abstract

The high percentage of drop-out students causes a campus management problem, this is because the percentage of students graduating on time is one of the elements of accreditation assessment set by the national accreditation board of higher education. One reason why the drop out rate is still high is because the Management System has not run well, such as lecturer professionalism, campus facilities, academics and administration, student affairs, outside influence and student personality. This study aims to analyze several indicators that can cause student drop outs by comparing the C4.5 method based on particle swarm optimization and Naïve Bayes based on PSO. This study contributes to campus management in anticipating the occurrence of drop outs through indicators that occur and can predict student drop out rates through the classification process. The highest level of accuracy produced from C4.5 + PSO is around 99.32% with AUC from Naïve Bayes is 0.974 categorized as excellent classification.

Keywords: Drop out, C4.5 method, Naïve Bayes, particle swarm optimization

1 Introduction

Data previous research obtained in the academic year 2010/2012 to 2015/2016 shows that the drop-out rate of students for Civil, Electrical and Mechanical majors in the period 2010-2012 was around 21% which was obtained when registering as students until graduation, while for the period from 2014 until 2015 around 35% when they enrolled in college until they were in the third semester obtained from the department of electrical, mechanical, informatics, civil and design visual communication. Finally, for 2015-2016 the level of student dropout was obtained from the Study Program staff of around 13% which was obtained when the student registered to enter the campus until after the first semester of mid semester test. From data researchers concluded that the level of drop out is high, especially for the period between 2014 until 2015, which is around 35% (Gustian & Hundayani, 2017). This is explained through Figure 1.

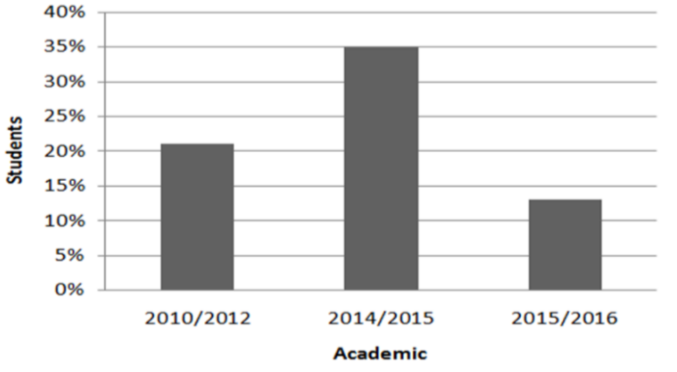

Figure 1.Comparison of the percentage of drop out rates at Nusa Putra High School of Technology in the 2010academic year up to 2015

Nusa Putra is one of the Universities in Sukabumi, that has a choice in technology with twelve study programs, namely civil engineering, design visual communication, informatics engineering, mechanical engineering, information systems, electrical engineering, management, accounting, law, and primary teacher education.

Therefore, the researchers felt the need to conduct research to find the drop-out point of students, which according to data obtained by researchers at the Nusa Putra High School of Technology was obtained by students dropping out for the Civil, Electricity and Machinery department in the period of 2010, around 21% obtained in 2012 from registration to graduation, and for the period from 2014 until 2015 around 35% when they enrolled in college until the third semester were obtained from the department of electrical, informatics, civil, mechanical, and visual communication. The latter for 2015-2016 obtained student dropout rates from six programs which reached 13% when student enrollment entered campus until after middle test in the first semester. From this data, the researchers concluded that the high level of drop out was mainly for the period between 2014 - 2015 which was around 35%.

This paper, the authors offer a solution to that problem, by combining the C4.5 method with the Naïve Bayes method and each based on PSO. Where through the screening process with the four methods above, it is expected to help campus management in overcoming the classic problems that always occur especially in Indonesia, so as to improve the quality and level of trust of the community. By comparing the four methods, it is expected to increase the accuracy of the existing classification methods and be able to precisely rank the indicators causing the drop out.

This research contributes to the campus management in taking preventive actions through appropriate decisions based on problems that have been sorted, so that they can provide good and intensive services for students who are predicted to drop out.

2 Literature Review

Clarify a few of the markers examined in foreseeing the level of drop out incorporate lesson cooperation, participation, semester exam and so on with the number of information sets around 450. Utilizing comparison of data mining algorithms to decide the finest exactness, that's C.45 with 55.52%, Random Tree 54.11 %, Random Forest 61.97%, ID3 79.23%, Chaid 49.50 and Decision Stump 50.95% (CEME, 2017).

The provisional direct assistance to the community (BLSM) could be a program of giving cash assistance to target family (RTS), which could be a destitute family (RTM), which is stipulated by the government in expansion to the fuel price hike. This study is based on the case that the dissemination of BLSM isn't the correct target and subjective interface. This BLSM is for the destitute who can not bear financially, but still numerous wealthy individuals who too get it extraordinarily in sub locale of Cicurug. Choice bolster framework (DSS) to decide the community coordinate help recipients in Sub Locale Cicurug with C4.5 method is one of the over subject arrangements. Data mining method is chosen because it can create models and criteria that easily deciphered by the classification of information preparing and information testing with genetic alogrithm (GA) so that it can be a better form of the method. Method C4.5 with accreditation value

92.92% from training data and 84.21% from testing data, C4.5 based on PSO have value 97.35% from training data and 98,68% from testing data, and GA-based C4.5 has accreditation data of 94.69% from training data and 90.67% from testing data. Can be concluded C4.5 based on PSO is exceptionally great (Gustian & Hundayani, 2017).

Giving a picture of two variables that can cause drop out from students, the inside variables of the students and the destitute curriculum on campus. From the factual information of HEI (Higher Instruction Institution) appears from 2000 - 2005 with average Drop out rate come to 22% from 12% in state colleges and 26% from private colleges. In expansion, agreeing to research Oscar Hipolito of Instituto lobo based on the number of sensu division of education in 2009, the level of budgetary misfortunes due to Drop out come to USD 9 billion with an expanding drift rate. This consider was conducted from 2000 - 2003 in colleges (HEI) with a few markers measured counting adolescence of understudies, need of introduction of majors, budgetary issues, imminent work, solid accentuation at colleges, impromptu relational unions and pregnancies and so forward. With choice tree Calculation utilized in foreseeing the level of drop out (Villwock, Appio, & Andreta, 2015).

Entitled "Graduation Prediction of Gunadarma University Students Using Algorithm and Naïve Bayes C4.5 Algorithm" with parameters assessed namely NEM SMA, 1st semester IP and 2nd semester IP, semester DNU GPA and 2, parent's salary and parent's work. The prediction of accuracy with the C4.5 methode 85.7% and error 14.3% while the Naïve Bayes algorithm 80.85% and error 19.05% (Suhartinah, 2010).

Describes the level of drop out students in eletro university engineering eidhoven, dutch about 40% of the total students in the majors. One of the main factors why it can be happening it is because in the curriculum that is considered difficult for students, therefore in the Netherlands there is a legal obligation that each campus should provide support for its students needed to evaluate their study options. The following stages are used in the evaluation process for the progress of the electronics department, including: using some of the indicators of the current review process, identifying specific success factors for the eletro department, identifying any data that can result in improved quality of the indicators in detail , Taking into account the views of each side of each student group and lastly modifying the results of the processed indidator in order to produce a better prediction early for the new student candidate later. Data sets collected during the years 200 2009 containing information on all students involved in electrical engineering program with the number of 648 students in the first year and VWO (pre-university education) or from the polytechnics that cooperate with the campus. In this research used some data mining classification algorithm between CART with 68% accuracy rate, OneR accuracy rate 68%, J48-M2 accuracy 70%, J48-M10 accuracy 69, Naïve Bayes accuracy rate 71%, Logic regression accuracy 69 %, Jrip 70% accuracy and 65% Random Forest for preuniversity education. The total accuracy of the dataset is CART accuracy of 79%, OneR accuracy of 75%, J48-M2 80% accuracy rate, J48-M10 80% accuracy rate, Naïve Bayes accuracy rate 75%, Logic regression accuracy rate 79%, Jrip 77% accuracy rate and Random Forest 79% (Dekker, Pechenizkiy, & Vleeshouwers, 2009).

The dataset used is a dataset of students of Nusaputra University consisting of 151 training data and 77 students. The variables measured were professionalism of lecturers, campus facilities, student personality, outside influences, academic finance and student affairs using the C4.5 method and Naïve Bayes method then each of these methods was optimized with PSO and compared the results to obtain optimal values. The implementation of this research is used to help users in classification using website-based applications.

In Figure 2, it appears that there is a lot of drop out problem so it is necessary to do research with existing variables. Datasets taken from 2010-2016 are divided into training data and testing data. This research resulted in a decision support system with data mining classification with confusion metrix and ROC curve measurement. Systems that have been created are tested using SQA tests.

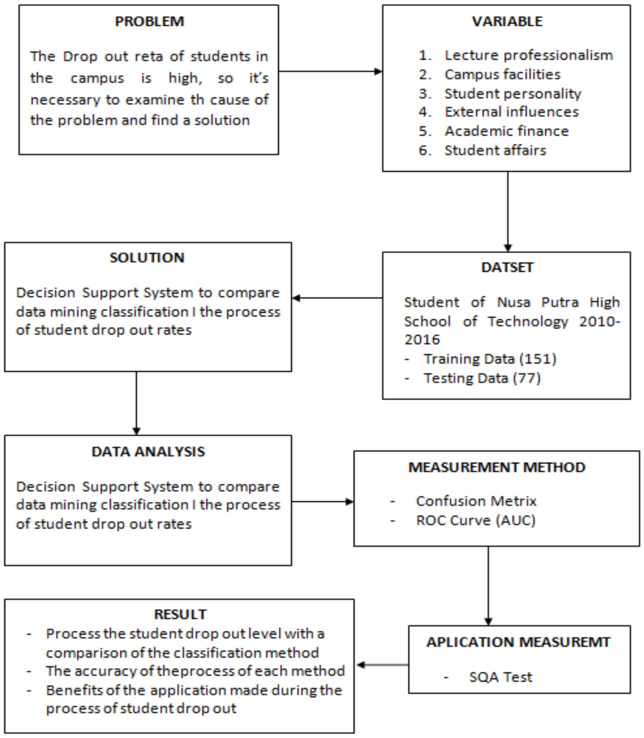
PROBLEM
The Drop out reta of students in the campus is high, so it's necessary to examine th cause of the problem and find a solution
VARIABLE
1. Lecture professionalism
2. Campus facilities
3. Student personality
4. External influences
5. Academic finance
6. Student affairs
SOLUTION
Decision Support System to compare data mining classification I the process of student drop out rates
DATSET
Student of Nusa Putra High School of Technology 2010-2016
- Training Data (151)
- Testing Data (77)
DATA ANALYSIS
Decision Support System to compare data mining classification I the process of student drop out rates
MEASUREMENT METHOD
- Confusion Metrix
- ROC Curve (AUC)
RESULT
- Process the student drop out level with a comparison of the classification method
- The accuracy of theprocess of each method
- Benefits of the application made during the process of student drop out
APLICATION MEASUREMT
- SQA Test

Figure 2.Thinking framework

3 Methodology

3.1 C4.5 Method

In this study using the method C4.5 and Naïve Bayes in the classification process and comparing to the results of the C4.5 based on PSO and Naïve Bayes based on the PSO are as follows:

a. Select the trait as root
b. Create a department for each value
c. Cases within the branch
d. Repeat the method for each department until all cases within the department have the same class

To select the property as root, based on the most noteworthy pick up esteem of the property. To calculate pick up, utilize the taking after formula.

Step 1:

$$Gain(\mathrm{S,A}) = \mathrm{Entrophy(s)} - \sum_{i=1}^{n}\left(\frac{|S_i|}{|S|} \times Entrophy\right) \quad (1)$$

Where,

S : Set of cases

A : Attribute

N : Number of partition attributes A

$|S_i|$: Number of cases on the i-partition

|S| : Number of cases in S

Some time recently getting the Pick up esteem is to discover the Entropy esteem. Entropy is utilized to decide how instructive an quality is to produce properties. The essential Entropy equation is as follows.

Step 2:

$$Gain(\mathrm{S,A}) = \mathrm{Entrophy(s)} - \sum_{i=1}^{n}\left(p_i \times \log\left(p_i\right)\right) \quad (2)$$

Where,

S : Set of Cases

N : Number of partitions S

Pi : Proportion of Si to S

3.2 Naïve Bayes

Naïve Bayesian Classifier is one of the problem solving algorithms included in the Classification Method in Data Mining. Naïve Bayesian Classifier adopts the science of statistics by using probability theory to solve a case of supervision learning, meaning that in the dataset there are labels, classes or targets as a reference [6].

$$P(\mathrm{H}\,|\,\mathrm{X}) = \frac{P(\mathrm{X}\,|\,\mathrm{H}) \cdot \mathrm{P(H)}}{P(\mathrm{X})} \quad (3)$$

Where,

X : Data with unknown classes

H : The X data hypothesis is a specific class

P (H|X) : Probability of hypothesis H based on condition X (posterior manipulation)

P (H) : Probability of hypothesis H (prior probability)

P (X|H) : Probability X is based on conditions in hypothesis H

P (X) : Probability X

3.3 Particle Swarm Optimization

$$V_i(t) = V_i\left(t-1\right) + c_1 r_1 \left[X_{pbest_i} - X_i(t)\right] + C_2 r_2 \left[X_{Gbest} - X_i(t)\right] \quad (4)$$

$$X_i(t) = X_i\left(t-1\right) + V_i(t) \quad (5)$$

Where:

X : Particle position

V : Particle velocity

W : Weight of inertia

c_1, c_2 : Acceleration coefficient

P : Number of particles in the herd

r_1, r_2 : Random value in the range (0.1)

3.4 SQA Evaluation Results

In Table 1 below explained metric measurement of measurement software with software quality assurance, where there are eight standard measurements with its function and percentage value as a benchmark measurement. more details are displayed in Table 1.

Table 1. Metric SQA

Metric	Description	Precentage
Auditability	Meet standards or not	1.25/100
Accuracy	Accuracy of computing	1.25/100
Completeness	Completeness	1.25/100
Error Tolernce	Tolerance is wrong	1.25/100
Expandability	Soft development	1.25/100

Operability	Easy to operate	1.25/100
Simplicity	Ease to be understood	1.25/100
Training	Ease of learning Help facilities	1.25/100

Respondent score = <audibility score> * 0.10 + <accuracy score> * 0.10 + completeness score> * 0.15 + <error tolerance score> * 0.10 + <execution efficiency score> * 0.10 + < operability score> * 0.15 + <Simplicity score> * 0.15 + <learning value> * 0.15 Evaluation based on respondents' average criteria / score.

4 Results and Discussion

4.1 Classification with C4.5 Method

In the Figure 3, it appears that there are 15 rules of the model tree, with the number of indication classes surviving as many as 6 rules and the indication class dropping out as many as 9 rules. But the note is the thickness of the rule, where when the value of student variables <3,500, academic and financial, campus facilities <= 3,500, student personality <2,500 then the chance of indication of drop out is likely. Likewise, the value of the indication of drop out has a different color gradation by having different assessments on each variable, this indicates that the more blue gradations than the red color means the possibility of an indication of student drop out is large. For student variables occupy the top of the tree, this is because the variable has the first order that has an influence on students followed by academic and financial.

4.2 ROC Curve

The Recipient Working Characteristic (ROC) bend is an curiously bend of war pull between affectability and specificity at different crossing point focuses. From this ROC strategy we'll get the esteem of Region Beneath Bend (AUC).

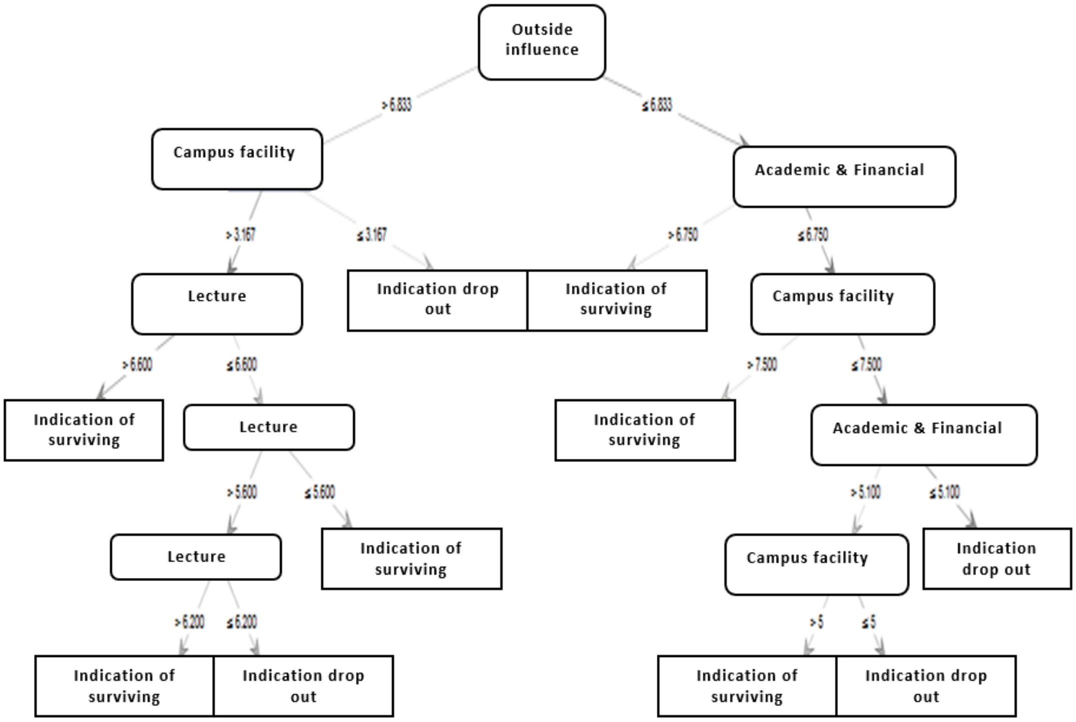

Figure 3.Decision tree from training data

4.3 Comparison of Methods

Within the Table 2 appeared underneath, that C4.5 methode with preparing information has an accuracy value of 99.32% whereas testing has an accuracy value of 84.00% showing that the exactness of training data is more noteworthy than testing data. Naïve Bayes calculation with accuracy value is training data 89.73% and testing data 89.33% appears that the accuracy esteem of training data is more noteworthy than testing data. C4.5 methode optimized with PSO has an accuracy value of 99.32% and 93.33% testing data and Naïve Bayes methode PSO optimization has an accuracy of 91.10% from training data and 90.67% from testing data.

It was concluded that all training data methode had higher accuracy values than the value of testing data. With this in conclusion C4.5 + PSO appears the most elevated esteem.

Table 2.Comparison of methods

Method	Training Data	Testing Data	ROC Training	ROC Testing
C4.5	99.32%	84.00%	0.490	1.000
Naïve Bayes	89.73%	89.71%	0.974	0.978
C4.5+PSO	99.32%	93.33%	0.954	1.000
Naïve Bayes+PSO	91.10%	91.10%	0.958	0.958

Showing up from the decision tree figure, there are 15 rules of the model tree, with the number of sign classes holding as numerous as 6 rules and the sign course dropping out as numerous as 9 rules. But the note is the thickness of the run the show, where when the value of student factors <3,500, scholarly and monetary, campus offices <= 3,500, student identity <2,500 at that point the chance of sign of drop out is likely. Moreover, the esteem of the sign of drop out encompasses a distinctive color degree by having distinctive evaluations on each variable, this demonstrates that the more blue degrees than the ruddy color implies the plausibility of an sign of understudy drop out is expansive. For student variables occupy the best of the tree, this can be since the variable has the primary arrange that has an influence on understudies taken after by scholarly and monetary. At that point in section classification with C4.5 and Naïve Bayes methods, the method of shaping a Tree model of training data, the accuracy of data.

4.4 System Implementation

For implementation using a website application, Users can only enter parameters and the results of the analysis can only be seen by the academic section.

a. Use Case Diagram

The Figure 4, consist of 2 actors, admin and student, where the admin acts as the manager of the entire system from the data of students who fill out questionnaires and reports. Students as the second actor can register and fill out the questionnaire.

In the Figure 4, consists of 2 actors, namely admins and students, where the admin acts as the manager of the entire system of student data that fills out questionnaires and reports. Students as the second actor can registrsi and fill out questionnaires.

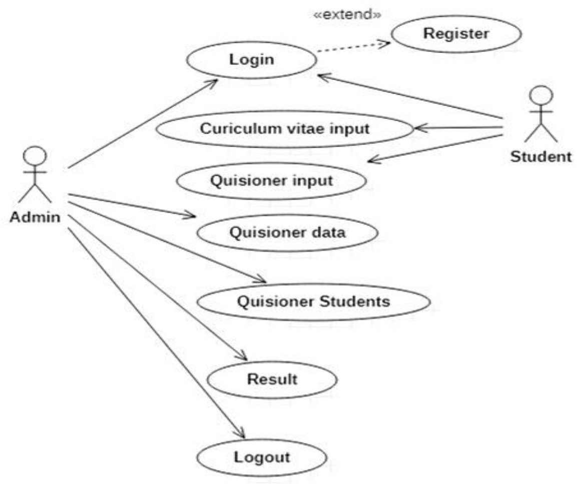

Figure 4.Use case diagram of eligibility of student drop out recipients

b. Class Diagram

The images listed above have 5 objects consisting of registers, students, quisioners, logins and users. Where registers and quisioners can be accessed by students, while the admin is the manager of the student register and quisioner data.

The Figure 5 shown 5 objects consisting of registers, students, quisioner, login and user. Where registers and quisioner can be accessed by students, while admin is the manager of the register data and student quisioner.

4.5 Software testing results prototype

To ensure that the software made has minimum quality standards, one method for quantifying software quality quantitatively is the SQA (*Software Quality Assurance*) method. To ensure that the software is designed for the SQA (Software Quality Assurance) method.

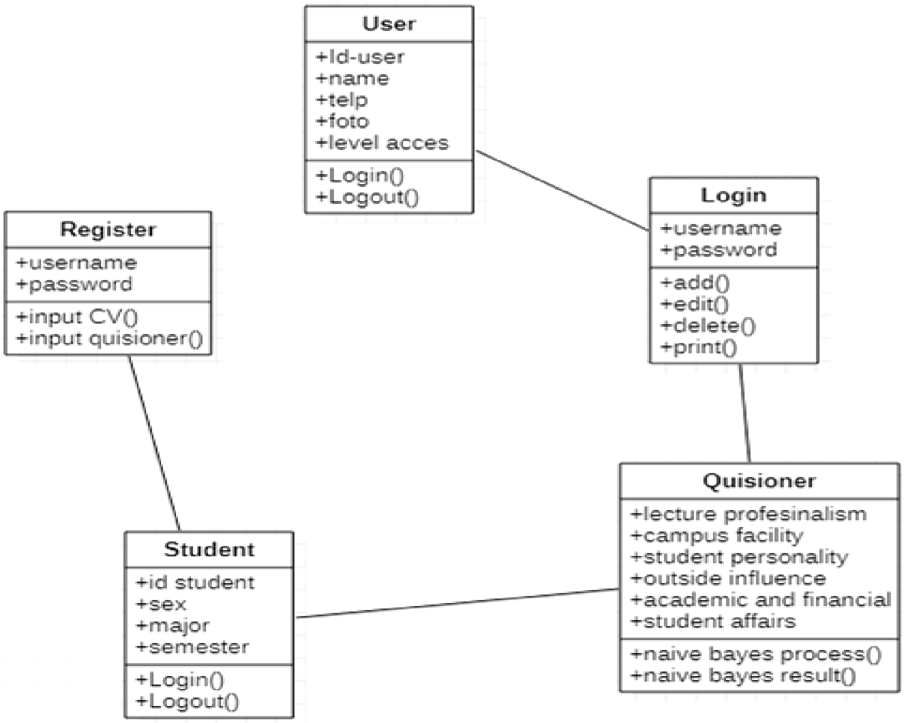

Figure 5.Class diagram of students drop out receiver feasibility test

Table 3.Results of SQA evaluation

Audience	Metric Score								Score
	1	2	3	4	5	6	7	8	
#1	84	87	85	75	90	95	85	78	84.87
#2	83	85	90	70	80	80	75	87	81.25
#3	86	95	82	75	87	78	86	90	84.87
#4	80	95	90	82	76	77	79	84	82.87
#5	80	85	90	70	80	80	75	85	

In the Table 3 above is the result of a questionnaire conducted on 5 observers who acted as users and were taken randomly. Score = <82.6> * 0.125 + <89.4> * 0.125 + <87.4> * 0.125 + <74.4> * 0.125 + <82.6> * 0.125 + <82> * 0.125 + <80 > * 0.125 + <84.8> * 0.125.The average score produced is 82.9, while the optimal value for a software that meets quality standards based on the SQA test is 82.9.

5 Conclusion

This study concludes that it can assist management in predicting student drop outs, with several existing criteria. A comparison of classification methods is used to produce the best model using measurements of Confusion matrix and ROC Curve. The SQA method is used to provide information on how much benefit the application has made. The C4.5 method gives better results than Naïve Bayes on training data, but Naïve Bayes is better at testing data. After both of these methods are optimized with PSO, C4.5 method is better in accuracy both training and testing data with a value of 99.32% and 93.33%. After being tested by several users, a value of 82.9 is obtained. This illustrates that this application can be categorized quite well in assisting management in predicting student drop rates.

References

Basuny, A. M., Arafat, S. M., & Ahmed, A. A. (2012). Vacuum frying: an alternative to obtain high quality potato chips and fried oil. *Banat's Journal of Biotechnology, 3*(5).

Belkova, B., Hradecky, J., Hurkova, K., Forstova, V., Vaclavik, L., & Hajslova, J. (2018). Impact of vacuum frying on quality of potato crisps and frying oil. *Food Chemistry, 241*, 51-59.

CEME, N. (2017). Student performance prediction and risk analysis by using data mining approach. *Journal of Intelligent Computing Volume, 8*(2), 49.

Dekker, G. W., Pechenizkiy, M., & Vleeshouwers, J. M. (2009). Predicting Students Drop Out: A Case Study. *International Working Group on Educational Data Mining*.

Gustian, D., & Hundayani, R. D. (2017). *Combination of AHP Method with C4. 5 in the level classification level out students.* Paper presented at the 2017 International Conference on Computing, Engineering, and Design (ICCED).

Su, Y., Zhang, M., & Zhang, W. (2016). Effect of low temperature on the microwave-assisted vacuum frying of potato chips. *Drying Technology, 34*(2), 227-234.

Suhartinah, E. M. S. (2010). Graduation Prediction Of Gunadarma University Students Using Algorithm And Naive Bayes C4. 5 Algorithm. *Fac. Ind. Technol. Gunadarma Univ*.

Villwock, R., Appio, A., & Andreta, A. A. (2015). Educational data mining with focus on dropout rates. *International Journal of Computer Science and Network Security (IJCSNS), 15*(3), 17.