Comparing Clasification Algorithm Of Data Mining to Predict the Graduation Students on Time

Imam Tahyudin*, Ema Utami, Armadyah Amborowati****
* Department of Information System, STMIK AMIKOM Purwokerto, Indonesia
** Department of Magister of Computer Engineering, STMIK AMIKOM Yogyakarta, Indonesia

Keywords:
Algorithms
Classification
Data mining
Prediction
Student
Graduation
On time

ABSTRACT

The percentage of students who graduate on time is one of the elements of accreditation of a study program. Based on data from the administrative and academic (BAAK) in 2013 showed that the graduation rate of students ontime in STMIK AMIKOM Purwokerto reached 78.80%. As the efforts to increase the percentage of students graduate on time is mining the information from the student database. Trough the student database can classify the levels of the graduation on time. The purpose of this research is to compare the several data mining classification algorithms, especially the Decision Tree (DT), Naive Bayes (NB), Artificial Neural Network (ANN), Support Vector Machine (SVM) and Logistic Regression (LR) algorithms with cross validation evaluation and T -Test to predict the graduation student on time. The method used is the comparison method. Based on the comparison of performance score and t-test, SVM algorithm is the appropriate algorithm that used to predict the student graduation on time. Level of accuracy to predict SVM algorithm is high (almost 100% with excellent clasification category). On the other hand, result of t-test of SVM algorithm is very dominant than other algorithms.

Corresponding Author:

Imam Tahyudin,
Department of Information System, STMIK AMIKOM Purwokerto, Indonesia
Email: imam.tahyudin@amikompurwokerto.ac.id

1. INTRODUCTION

Based on the assessment instrument matrix accreditation of National Accreditation Board for Higher Education (BAN-PT) [1] that the percentage of students who graduate on time is one of the elements to evaluate quality the study program. BAN-PT is the legal board in Indonesia that asses the quality of higher education. So that, every higher education must be detect student behaviors that do not graduate on time, and the influence factors.

STMIK AMIKOM Purwokerto is one of the largest higher education in Central Java, which has 3,490 students. Based on data from the administrative and academic (BAAK) in 2013 showed that the graduation rate of students STMIK AMIKOM Purwokerto on time reached 78.80%. The effort should be made to increase the percentage of students graduate on time is mining the information from a database stored in BAAK STMIK AMIKOM Purwokerto. Through the database can be extracted valuable information for consideration to improve on-time graduation. One method of data mining that can be used is the classification method [3], [4], [5].

There are many classification algorithms that can be used such as Logistic Regression Algorithm, Decision Tree, Naive Bayes, Neural Network, and Support Vector Machine (SVM) [13], [14]. Previous researches have been conducted by some researchers such as Madhu S.Shukla and Kirit R. Rathod [12] studied comparison the naive bayesian, Hoefding tree, and CVDFT algorithm. The research results show that the best algorithm is CVDFT algorithm. Vahid Alizadeh Sahzabi and Azuraliza Abu Bakr [2] on the comparison of multiple classification algorithms which data mining decision tree (J48 and LMT), Bayes Algorithm (Naive Bayes and Bayesian networks), clasification neural network (MLP = multi-layer perceptron, RBF = radial base function), and Rough Set methods. The result that ANN (MLP) have the accuracy score higher then the others . Khafiizh Hastuti [6] on a comparison of classification algorithms, Logistic Regression algorithm, decision tree, naive Bayes, neural network, with an evaluation tool that is

cross validation, confusion matrix, ROC Curve and T-Test. The results obtained are the Logistic Regression algorithm is an algorithm which is the most dominant right but its accuracy is the lowest value. P. Nancy and R. Geetha Ramani [8] on the comparison of several methods of data mining classification Rnd Tree especially, ID3, K-NN, C-RT, CS-CRT, C.45 and CS-MC4. The research results show that the algorithm Rnd Tree has the smallest error rate. Milan Kumari, Sunila Godara [7] on a comparison of classification algorithms, namely RIPPER classifier data mining, decision tree, artificial neural network (ANN), support vector machine (SVM). The result is the SVM algorithm is an algorithm that produces the smallest error and The greatest accuracy. Aman kumar Sharma, and Suruchi Sahni [11] comparison classification algorithms data mining, especially ID3, J48, simple CART and Decision tree on spam email data set. The result is the J48 algorithm has the highest accuracy value. Neslihan Dogan and Saturn Tanrikulu [15] on the comparison of multiple data mining algorithms, especially CHAID, MLP, Logistics, airs and Naive Bayesian were applied to the data set. Resulting conclusion that the best model used for prediction is regression model. Further research conducted by Silvia Marselina Suhartinah and Ernastuti [10] on the application of data mining to predict graduation using Naive Bayes and C4.5 algorithms. The result is an algorithm C4.5 has a greater degree of accuracy than Naive Bayes. Based on the above reserach, we are interested to compare multiple algorithms generated in an earlier study produced the greatest degree of accuracy the Decision Tree (DT), Naive Bayes (NB), Artificial Neural Network (ANN), Support Vector Machine (SVM) and Logistic Regression (LR) algorithms with the evaluation parameters are cross validation and T-Test.

2. RESEARCH METHOD

This research using the comparative research method with experimental approach [9]. The analyse method is clasification of data mining method. The software used is rapid miner 5.0. The research process that will be carried out as shown in Figure 1.

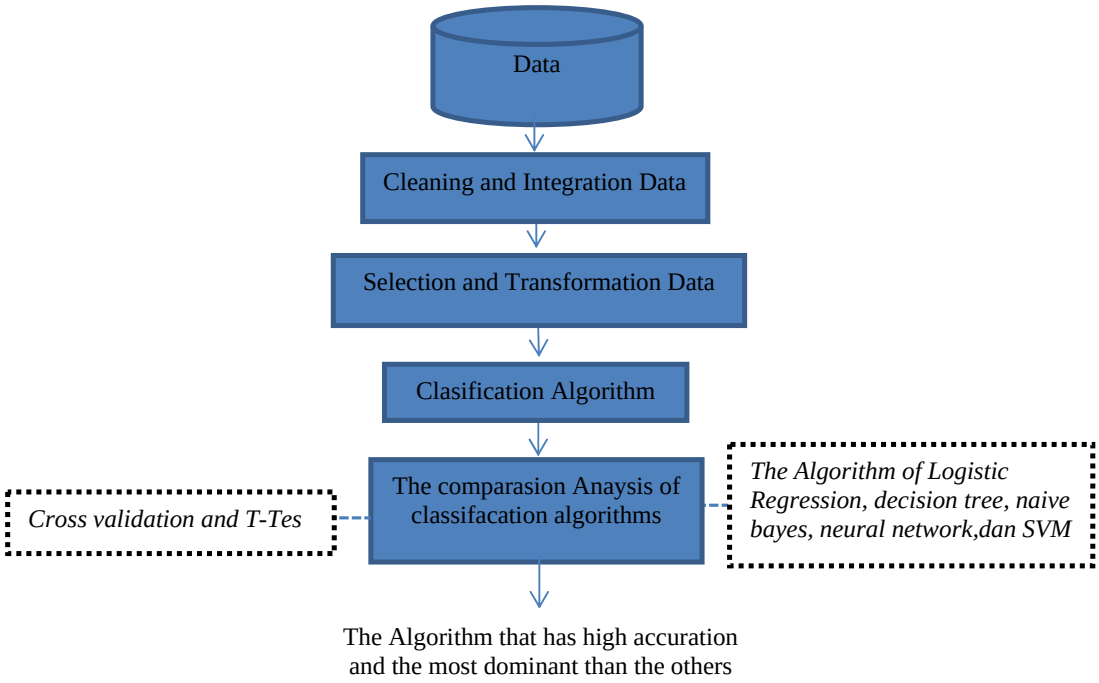

Figure 1. The Flowchart of the Reseach Steps

According to the figure 1. the data have cleaned from the blank or noise data will be integrated. After that will be selected and transformed to appropriate the format needed. And then, use some of the classification algorithm and compare them with the cross validation and t-test.

Based on the research process above the data used are secondary data. Data obtained from the Administration and Academic (BAAK) STMIK AMIKOM Purwokerto is a parent student biographical data, when first registered in 2005 to 2009 as many as 1,286 students. The number of obsevations are 651 students. They are students who have graduated from the program of study information systems and computer

engineering in 2009 - 2013 period I - VIII. Supporting data in this research was obtained from the literature such as journals, reference books, proceedings, and so forth. Attribute data is used as shown in Table 1.

Table 1. Data attributes used

Atribut	**Explanation**
NIM	Id
Shift	Special Shift, Shift I, Shift II, Shift III
Sex	Male, female
Home Town	Purwokerto, Other purwokerto
Senior High School statue	Public, Private
The major of Senior High School	IPA, IPS, Technique, other
The Major in High education	Computer engineering, information system
Student statue	New, Transfer
Marriage statue	Maried, unmaried
Age	< 20; 20 – 30; > 30
IPK	< 2,75; 2,75 – 3,5; > 3,5
Graduation predicat	Prise, very satisfaction, satisfaction
Graduation period	I, II, III, IV, V, VI, VII, VIII
The study acuration	On time, late

3. RESULT AND ANALYSIS

3.1. Output Performance based on the value of Confusion Matrix

Table 2. Output confusion matrix DT algorithm

Table View / Plot View

accuracy: 80.01% +/- 2.94% (mikro: 80.00%)

	true TEPAT	true TELAT	class precision
pred. TEPAT	433	72	85.74%
pred. TELAT	51	59	53.64%
class recall	89.46%	45.04%	

Table 3. Output confusion matrix NB algorithm

Table View / Plot View

accuracy: 75.16% +/- 5.88% (mikro: 75.12%)

	true TEPAT	true TELAT	class precision
pred. TEPAT	396	65	85.90%
pred. TELAT	88	66	42.86%
class recall	81.82%	50.38%	

Table 4. Confusion matrix output ANN algorithm

Table View / Plot View

accuracy: 100.00% +/- 0.00% (mikro: 100.00%)

	true TEPAT	true TELAT	class precision
pred. TEPAT	484	0	100.00%
pred. TELAT	0	131	100.00%
class recall	100.00%	100.00%	

Table 5. Confusion matrix output SVM algorithm

Table View / Plot View

accuracy: 100.00% +/- 0.00% (mikro: 100.00%)

	true TEPAT	true TELAT	class precision
pred. TEPAT	484	0	100.00%
pred. TELAT	0	131	100.00%
class recall	100.00%	100.00%	

Table 6. Output LR confusion matrix algorithm

Table View / Plot View

accuracy: 100.00% +/- 0.00% (mikro: 100.00%)

	true TEPAT	true TELAT	class precision
pred. TEPAT	484	0	100.00%
pred. TELAT	0	131	100.00%
class recall	100.00%	100.00%	

Based on the result in Table 2 – 6 the accuracy score of DT algorithm is 80,01%, so it can predict the graduation on time 80,01%. The accuracy score of NB algorithm is 75,16%. It is the lowest accuracy to predict then the others. ANN, SVM and LR accuracy rates are almost 100%. So that the ANN, SVM, and LR algorithms can predict the graduation student on time is almost 100%.

3.2. Output Performance based on the value of the ROC (Receiver Operating Characteristic) or AUC (Area Undercurve)

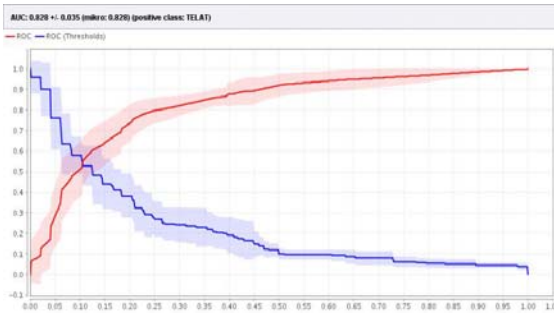

Figure 2. Output AUC DT algorithm

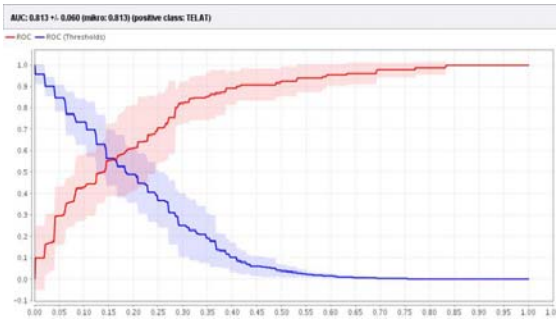

Figure 3. Output AUC of NB algorithm

Figure 4. Output AUC of ANN algorithm

Figure 5. Output AUC of SVM algorithm

Figure 6. Output AUC of LR algorithm

Based on the figure 2 – 6 the AUC value of DT and NB algorithms are 0.828 and 0,813, they mean can predict the graduation student on time with the good clasification category. The AUC value of ANN, SVM and LR algorithms are 1.00. This result indicates that they can predict the graduation student on time with the excellent clasification category.

3.3. Output T-test

To determine the best algorithm is used to predict the on time student graduation rates using a statistical t-test. Based on Table 3.6. the hypotesis that the SVM algorithm is the most dominant among the others is accepted because it's p-value (0.00) are all less than alpha (0.05). The t-test results obtained by t-test results matrix in Table 7.

Table 7. Matrix T-test

	DT	NB	ANN	SVM	LR
DT		0,97	0,81	**0,00**	0,60
NB	0,97		0,76	**0,00**	0,07
ANN	0,81	0,76		**0,00**	0,05
SVM	**0,00**	**0,00**	**0,00**		**0,00**
LR	0,60	0,07	0,05	**0,00**	

3.4. Performance Comparison of Data Mining Classification Algorithms

Based on the results obtained classification accuracy rate each performance outcome algorithm to predict the on time graduation of students as shown in Table 8.

Table 8. Comparison of Data Mining Classification Algorithms performance

	Algoritma DT	**Algoritma NB**	**Algoritma ANN**	**Algoritma SVM**	**Algoritma LR**
Accuracy	80,01%	75,16%	100%	100%	100%
AUC	0,828 *(Good clasification)*	0,813 (*Good clasification)*	1,00 *(excellent clasification)*	1,00 *(excellent clasification)*	1,00 *(excellent clasification)*
T-Test	Dominant	No dominant	Dominant	The most dominant	dominant

Based on the comparison Table 8. There are three algorithms that have the same accuracy and value of the same classification category based on evaluation by using confusion matrix and ROC Curve. They are ANN, SVM and LR algorithms. They mean that the algorithms can predict the graduation on time with the accuration almost 100% and as the excellent clasification category. But the t-test results show that the SVM algorithm is an algorithm which is most dominant among the others.

4. CONCLUSION

Based on the results of the comparison showed that the ANN, SVM and LR algorithms can predict the graduation rates of students on time with an accuracy rate almost 100% and as the excellent classification categories. But the algorithm which has higher t-test value from the others is SVM algorithm. Thus SVM algorithm is the best algorithm that be can used to predict the graduation student on time. Suggestions for further studies, researchers need to try other classification algorithms such as K-Nearest Neighbourhood, ID3, CHAID and etc. It also needs to develop the type of evaluation used as delong Pearson.

ACKNOWLEDGEMENTS

This reseach has been completed, researchers thanks to the help and support from various parties that could not be mention one by one. In particular thanks to my little family, children and my wife. Researchers would like to thank to the leader of STMIK AMIKOM Purwokerto which has provided the opportunity to complete this research. Furthermore, researchers would like to thank the parties in STMIK AMIKOM Yogyakarta who have provided support, guidance and cooperation in completing this research. Hopefully this research can be useful.

REFERENCES

[1] Anonim, The 4 th Book of Assessment Instrument Matrix Accreditation Program National Accreditation Board for Higher Education , 2008.

[2] Barnaghi, P. Mamani, SV. Aliza, AB. Azuraliza, "Comparative Study for Various Methods of Classification," *International conference and computer network (ICICN 2012). IPCSIT vol. 27 (2012) ©(2012) IACSIT*, 2012

[3] M.H. Dunham, "Data mining," *Introductory and Advanced Topics*. Prentice Hall. 2002.

[4] F. Gorunescu, "Data Mining Concept Model and Techniques," Berlin: *Springer*. ISBN 978-3-642-19720-8. 2011.

[5] J. Han and M. Kamber, "Data Mining Concept and Tehniques," San Fransisco: *Morgan Kauffman*. ISBN 13: 978-1-55860-901-3, 2006.

[6] K. Hastuti, "Comparative Analysis of Classification Algorithm for Data Mining Prediction of Non-Active Students," *Seminar Nasional Applied Information and Communication 2012 (SEMANTIK 2012) Semarang*, June 23, 2012.

[7] M. Kumari and S. Godara, "Comparative Study of Data Mining Clasification Methods in Cardiovascular Diseases Prediction," *IJCST Vol 2. Issue 2, June 2011. ISSN : 2229 – 4333(Print) – 0976 – 8491 (online), 2011.*

[8] P. Nancy, Ramani, and R. Geetha. 2011. "A Comparation on Performance of Data Mining Algorithms in Clasification of Social Network Data," *International Journal of computer Applications (0975 – 8887) Vol. 32 No. 8 October 2011.*

[9] M. Nazir, "The Reseach Method," *Ghalia*, Indonesia

[10] S.M. Suhartinah and Ernastuti. 2010. "Graduation Prediction of Guna Darma University Students Using Algorithm Naive Bayes and C4.5 Algorithm," *Jurnal Magister Sistem Informasi*. Universitas Guna Darma. Jakarta. 2005.

[11] Sharma, K. Aman and S. Sahni, "A Comparative Study of Clasification Algorithms for Spam Email Data Analysis," *International Journal On Computer Science and Engineering (IJCSE). ISSN : 0975 – 3397 Vol. 3 No. 5 may 2011*

[12] Shukla, S. Madhu, Rathod, R. Kirit. 2013. "Stream Data Mining and Comparative Study Of Clasification Algorithms," *International journal of engineering research and applications (IJERA) ISSN : 2248 – 9622 Vol. 3, Issue 1, January – February 2013, pp. 163-168*

[13] Vercellis, Carlo. 2009. "Business Intelligence: Data Mining and Optimization for Decision Making," United Kingdom: *John Willey & Son.*

[14] Witten, Ian, H; Eibe, Frank; Hall, A.M. 2011. "Data Mining : Practical Machine Learning Tools and Technique," 3rd ed., Asma Sthepan and Burlington, Eds. United States of America, *Morgan Kaufman.*

[15] D. Neslihan and T. Zuhal, "A comparative framework for evaluating clasification algorithms," *Proceedings of The World Congress on Engineering 2010 Vol. I WCE 2010, June 30 – july 2, 2010, London, U.K.* 2010.

BIBLIOGRAPHY OF AUTHORS

	Imam Tahyudin was born in Indramayu, West Java, Indonesia, on July 12, 1983. He Received S.Si degree from Faculty of Science and Technology in 2006 and M.M. degree from faculty of Economic in 2010 from Jenderal Soedirman University, Purwokerto, Indonesia. He is currently pursuing the M.Eng. degree in Department of Computer Engineering STMIK AMIKOM Yogyakarta, Indonesia, in the field of information system. He is lecturer in the department of information system STMIK AMIKOM Purwokerto, Indonesia. His research interests are in information system management and data mining.
	Ema Utami was born in Lampung, Sumatera, Indonesia, on February 21, 1975. She received the S.Si, M.Kom and Doctoral degrees in Computer Science from Gadjah Mada University, Yogyakarta, Indonesia in 1997, 2002 and 2010 respectively. Since 1998 she has been a lecturer in STMIK AMIKOM Yogyakarta, Indonesia. Her areas of interest are Natural Language Processing, Computer Algorithms, and Database Programming.
	Armadyah Amborowati was Received S.Kom degree from STMIK AMIKOM Yogyakarta and M.Eng in MTI faculty of electricity Engineering from Gadjah Mada University, Yogyakarta, Indonesia. He is lecturer in STMIK AMIKOM Yogyakarta, Indonesia. Her research interests are database programming and data mining.