Comparative Analysis of Three Classification Algorithms in Predicting Computer Science Students Study Duration

Debby E. Sondakh
Faculty of Computer Science
Universitas Klabat
Manado, Indonesia
Email: debby.sondakh [AT] unklab.ac.id

Stenly R. Pungus
Faculty of Computer Science
Universitas Klabat
Manado, Indonesia

Abstract—**This paper aims to present a predictive model for computer science students' study duration at Faculty of Computer Science Universitas Klabat. The predictive model was developed based on students' performance (grades) in the first two semesters. Classification techniques from Data mining were applied to develop the models: Naïve Bayes, decision tree and Support Vector Machine. Comparative analysis is conducted on the three selected algorithms to find the best classification model. Moreover, this research also aims to find out the most influential subjects' grades on study duration. Courses, gender, and grades (general, basic, and major grades) serve as the independent parameters that would predict the dependent parameter i.e. study duration, which comprises of three categories: Less, Equal, and Greater. The resulting models of the three algorithms show no significant difference between Naïve Bayes and decision tree performances, while SVM has the lowest performance. Basic subjects grades found to be the most influence parameter to the students' study duration, followed by general subjects' grades, gender, and major subjects' grades parameters.**

Keywords-Predictive model, Study duration, Classification

I. INTRODUCTION

Facing the growth of academic data is a challenge for a higher education institution, not only in terms of data storage management but also how to utilize the data appropriately to improve the quality of managerial decisions as well as the educational performance of students and faculty members. The huge number of data makes it difficult to analyze them manually; it takes a long time and complicated process. Data mining; also known as knowledge mining, knowledge extraction, information discovery, data analysis [1, 2], provides solutions for this problem. To transform raw data into useful information and knowledge, data mining adopts techniques and algorithms of multiple science discipline including databases, statistics, machine learning and artificial intelligence.

In educational environment, data mining techniques have been widely used to extract and retrieve valuable information related to the students, faculties, and management, in order to improve the quality of educational process and institution management. Implementation of data mining in education is known as educational data mining (EDM). EDM is defined as the application of data mining techniques to extract, discover, and learn the knowledge of students' behavior patterns which have not been identified yet, that are stored in academic database. It aims to identify the relationships among variables related to students learning [3], measuring learning process [4], analyze and improve students performance [5, 6], making predictions [4, 5, 7, 8, 9, 10], improve student retention [11], and analyze dropout rate [12].

Universitas Klabat (Unklab) is a private university in Indonesia and faculty of Computer Science is one of the six faculties it has. Unklab has an academic information system, called Sistem Informasi Unklab (SIU), with a database that stores academic data of all students. Nevertheless, these data has not been fully utilized, while they are potentially provide valuable knowledge about students' academic performance. Faculty of Computer Science offers a bachelor program that is intended to be completed within eight semesters or four years. However, some students accomplish the course in less than four yours, while some had to spend more than the specified period. This study was conducted to develop faculty of Computer Science students' academic performance prediction models based on their grades, using three data mining classification algorithms; decision tree, Naïve Bayes, and Support Vector Machine (SVM). The models will predict students' study duration based on their academic performance, the grades. This may help faculty management staff to properly counsel the students to improve their overall academic performance, in order to complete the course on the specified duration. This paper presents the performance of decision tree, Naïve Bayes, and SVM. This paper is an extension of work originally reported in *Proceedings of the 4th International Scholars Conference.*

II. METHOD

The present study adopted the hybrid model knowledge discovery process [2]. This model combines Academic research knowledge discovery models with Cross-industry standard process for data mining (CRISP-DM), a model from

industrial field. The research has been conducted in 5 steps, as depicted in Figure. 1.

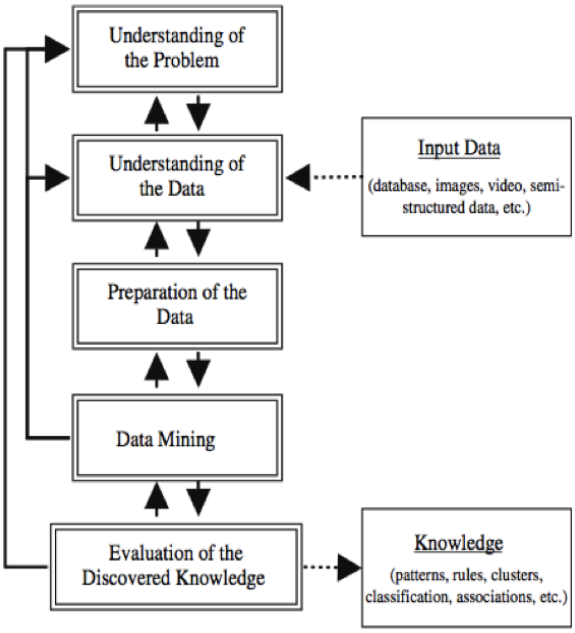

Figure 1. Methodology

A. *Understanding of the Problem Domain*.

This first step aims to understand the scope of the problem to be solved using data mining techniques, as well as determining objectives or expected output of data mining process. Universitas Klabat has SIU that manages the academic process. SIU records all students demographic and academic data, include Computer Science department students.

B. *Understanding of the Data*.

This second step did the data collection and selection. Data format and size are specified. A total of 373 data of Computer Science students, who have completed their degree, are obtained from SIU database. The data contain students' academic information from July 2003/2004 intakes to July 2012/2013 intakes. Two separate Excel files were extracted as follows:

a. Grade. This file contains information about students' registration ID, schedule ID, course code, students' data (registration number, student ID, surname, name, gender, faculty, program, date of birth), grade (number, letter), semester ID, grade input information (name, date, update), class code, lecturer ID, lecturer's name, schedule (date, room number), credits, and semester description.

b. Curriculum. This file contains information about curriculums: ID, course code, course name, credits, and course type.

C. *Preparation of the Data*.

This step includes extraction and transformation, to create student grade dataset.

a. **Data Extraction**. Grade and curriculum files were combined into a single file and five parameters were selected for this research i.e. program, gender, grade of each subject type (major, basic, and general). Then, the average grades of each subject type, from the first and second semesters, are calculated. Table I shows the parameter chosen. One parameter is added, duration, to determine the classification category.

TABLE I. PARAMETER SELECTED FOR STUDENT GRADE DATASET

Parameter	Description	Value
Program	Course offers by department of computer science	SI (Sistem Informasi), TI (Teknik Informatika)
Gender	Students gender	Male, Female
M_Grade	Average major subjects grade	0 – 4
B_Grade	Average basic subjects grade	0 – 4
G_Grade	Average general subjects grade	0 – 4
Duration	Study duration	7 – 14

b. **Data Transformation**. Data transformation stage will convert the numerical values into categorical, as shown in Table II. The six parameters are grouped into independent and dependent parameter. Independent parameters, the input for the model, are Program, Gender, M_Grade, B_Grade, and G_Grade. Dependent parameter, role as the output, is Duration.

TABLE II. TRANSFORMATION SELECTED PARAMETERS

Parameter Type	Parameter	Value
Independent	Program	SI, TI (nominal)
	Gender	M, F (nominal)
	M_Grade	Low : 0-1.99 Average : 2-2.99 High : 3-4 (nominal)
	B_Grade	Low : 0-1.99 Average : 2-2.99 High : 3-4 (nominal)
	G_Grade	Low : 0-1.99 Average : 2-2.99 High : 3-4 (nominal)
Dependent	Class (Duration)	Less : < 8 semester Equal : = 8 semester Greater : > 8 semester (nominal)

The screen shot of Weka preprocessing stage is shown in Figure 2.

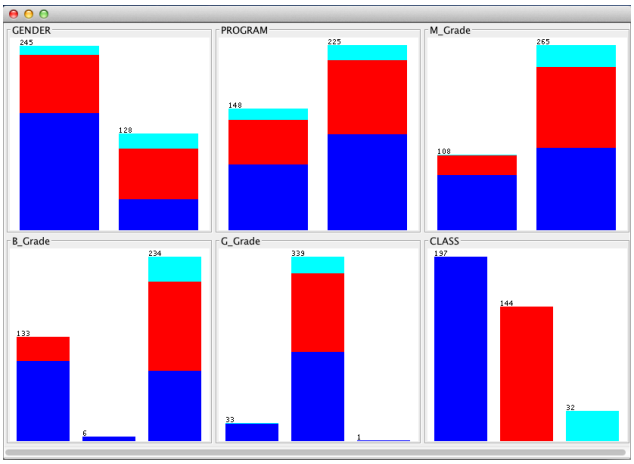

Figure 2. Data Distribution – Preprocessing Step

c. **Data mining**. At this stage, dataset is analyzed using Weka tool to obtain the predictive models. Three algorithms were compared. Decision tree is a famous classification algorithm. It decomposes the data into a hierarchical structure called tree. Decision tree classifier comprises of *internal nodes* that stores the attributes, *branches* come out of an internal node as the conditions represent one attribute value, and *leaf nodes* represent the category or class [13]. Naïve Bayes is a probabilistic classifier that utilize mixture model, a model that combine terms probability with category, to predict object category probability [14]. It is based on Bayes probability theory that assumes the effect of an attribute value of a given class is independent from the values of other attributes [12]. SVM aims to find a boundary, called decision surface or decision hyperplane, which separates two groups of vectors/classes. The system was trained using positive and negative samples from each category, and then calculated boundary between those categories. Data are classified by first calculating their vectors and partition the vector space to determine where the data vector is located. The best decision hyperplane is selected from a set of decision hyperplane $\sigma_1, \sigma_2, \ldots, \sigma_i$ in vector space $|T|$ dimension that separate the positive and negative training data. The best decision hyperplane is the one with the widest margin [15].

d. **Evaluation of the Discovered Knowledge**. The resulting model from data mining algorithms is further evaluated to interpret the hidden valuable knowledge in it.

III. Result and Discussion

Experimental results are discussed in this section. This study's goal is to develop a study duration predictive model of computer science students, based on their performance in the first two semesters, using input parameters as per Table II. They are analyzed using data mining classification techniques: decision tree, Naïve Bayes, and SVM. WEKA data mining tool is used for the performance evaluation.

TABLE III. Decision Tree Classifier Performance

Class	TP Rate	FP Rate	Precision	Recall	F-1	ROC
GREATER	0.7	0.3	0.72	0.7	0.68	0.745
EQUAL	0.65	0.39	0.51	0.65	0.57	0.645
LESS	0	0	0	0	0	0.729
Weighted Avg.	0.62	0.31	0.58	0.62	0.6	0.705

TABLE IV. Naïve Bayes Classifier Performance

Class	TP Rate	FP Rate	Precision	Recall	F-1	ROC
GREATER	0.61	0.21	0.77	0.61	0.68	0.757
EQUAL	0.76	0.46	0.51	0.76	0.61	0.678
LESS	0	0.003	0	0	0	0.757
Weighted Avg.	0.62	0.287	0.603	0.62	0.6	0.727

TABLE V. SVM Classifier Performance

Class	TP Rate	FP Rate	Precision	Recall	F-1	ROC
GREATER	0.69	0.386	0.668	0.69	0.68	0.652
EQUAL	0.58	0.37	0.49	0.58	0.53	0.629
LESS	0	0	0	0	0	0.457
Weighted Avg.	0.59	0.35	0.54	0.59	0.57	0.626

The performance of decision tree, Naïve Bayes, and SVM are given in Table III, IV, and V. To classify the study duration correctly from training dataset, accuracy and error rates are calculated. Table VI presents the performance comparison of the three algorithms via values of weighted average. The values show no significant difference between decision tree and Naïve Bayes accuracies. Both algorithms are better than SVM for the chosen dataset.

TABLE VI. Algorithms Performance Comparison - Accuracy

Parameter	Decision Tree	Naive Bayes	Support Vector Machine
Correctly Classified	62%	62%	59%
TP Rate	0.62	0.62	0.59
FP Rate	0.31	0.29	0.35
Precision	0.58	0.6	0.54
F-1	0.6	0.6	0.57
ROC	0.705	0.727	0.626

Table VII depicts the error report of the three algorithms. Three measurements were analyzed i.e. the Kappa statistic, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). Kappa statistic is a chance-corrected measure of agreement between the classification and the true classes. It calculate the difference between how much agreement is actually present (called 'observed agreement) compared to how much agreement would be expected to be present by chance alone (called 'expected agreement) [16]. Kappa values of the three models belong to 'fair agreement', see Kappa scale [16].

These indicate that the resulting models are not good enough in predicting study duration in this case study.

TABLE VII. ALGORITHMS' ERROR REPORTS

Statistic	Decision Tree	Naive Bayes	Support Vector Machine
Kappa	0.3	0.31	0.23
MAE	0.32	0.31	0.33
RMSE	0.41	0.4	0.43

MAE is a statistical measure of how far the prediction from actual value. It is the average of absolute magnitude of the individual errors, and slightly smaller than RMSE. RMSE calculates the differences between values predicted by a model and the values actually observed from the thing being modeled. It is used to measure the accuracy and is ideal if it is small. In Table VII NB get the lowest RMSE 0.4; which means NB accuracy is the highest.

Table VIII reports the significant test result, using *t-paired* test with 5% level of significance. Naïve Bayes acts as the test base. The parameters tested refer to the accuracy and error rate measurements in Table VI and Table VII. Symbol ^{v} (*victory*) indicates a classifier is superior to the base, * indicates a lower classifier performance, and " " (unmark) states that the significance test cannot determine whether the classifier performance is better or poorer than the other. Overall, significant test results show no difference with the previous test. For SVM we get lower accuracy percentage, precision, AUC, and Kappa statistic. Decision tree wins against NB in terms of TP-Rate and FP-Rate, but lost in precision.

TABLE VIII. T-TEST RESULT

Parameter	Naive Bayes	Decision Tree	Support Vector Machine
Correctly Classified	62.55	62.44	57.78*
TP Rate	0.62	0.71^{v}	0.69
FP Rate	0.2	0.29^{v}	0.41^{v}
Precision	0.79	0.73*	0.67*
F-1	0.69	0.72	0.66
AUC	0.77	0.76	0.64*
Kappa	0.32	0.31	0.21*
MAE	0.31	0.32	0.34^{v}
RMSE	0.40	0.41	0.43^{v}

To determine the parameter that most influence students' study duration feature selection is conducted by applying Information Gain (IG) calculation using WEKA. Table X presents the IG for each parameter. B_Grade parameter has highest IG value of 0.144 bits, it shows that B_Grade is the most influencing parameter for study duration in this case study. B_Grade is followed by G_Grade with IG 0.079 bits, Gender 0.072 bits, M_Grade 0.063 bits, and Program 0.001 bits as the less influence parameter of students' study duration.

TABLE IX. INFORMATION GAIN

Attributes	IG
B_Grade	0.144
G_Grade	0.079
Gender	0.072
M_Grade	0.063
Program	0.001

IV. CONCLUSION

Data mining techniques have been widely used in educational environment. This research's goal is to apply data mining technique to analyze the department of Computer Science of Unklab students' performance in terms of study duration based on their grades in the first two semesters. Three classification algorithms were applied, namely decision tree, Naïve Bayes, and Support Vector Machine. The resulting models of the three algorithms show no significant difference between Naïve Bayes and decision tree performances, while SVM has the lowest performance. Basic subjects grades found to be the most influence parameter to the students' study duration, followed by general subjects' grades, gender, and major subjects' grades parameters.

As for further research, a more comprehensive analysis of each subject included in basic type can be done to find out the specific subject that most influence students' study duration.

REFERENCES

[1] J. Han & M. Kamber, *Data Mining Concepts and Techniques*, 2nd Ed., Morgan Kauffman Publisher, USA, 2006.

[2] K. J. Cios, et.al., *Data Mining A Knowledge Discovery Approach*, Springer, New York, USA, 2007.

[3] B. K. Baradwaj dan S. Pal, "Mining Educational Data to Analyze Students' Performance", International Journal of Advanced Computer Science and Applications, Vol. 2, No. 6, 2011.

[4] M. Durairaj dan C. Vijitha, "Educational Data Mining for Prediction of Student Performance Using Clustering Algorithms", International Journal of Computer Science and Information Technologies, Vol. 5, No.4, 2014.

[5] A. A. Aziz, N. H. Ismail, dan F. Ahmad, "First Semester Computer Science Students' Academic Performances Analysis by Using Data Mining Classification Algorithms", in *Proceeding of the International Conference on Artificial Intelligence and Computer Science* (AICS 2014), Bandung, Indonesia, 2014.

[6] K. S. Priya dan A. V. S. Kumar, "Improving the Student's Performance Using Educational Data Mining", International Journal of Advanced Networking and Applications, Vol. 04, No. 04, pp. 1680-1685, 2013

[7] A. O. Ogunde dan D. A. Ajibade, "A Data Mining for Predicting University Students' Graduation Grades Using ID3 Decision Tree Algorithm", Journal of Computer Science and Information Technology, Vol. 2, No.1, pp. 21-46, Maret 2014.

[8] G. S. Abu-Oda dan A. M. El-Halees, "Data Mining in Higher Education: University Student Dropout Case Study", International Journal of Data Mining & Knowledge Management Process (IJDKP), Vol. 5, No. 1, Januari 2015

[9] D. Kabakcieva, "Predicting Student Performance by Using Data Mining Methods for Classification", Cybernetic and Information Technologies, Vol. 13, No. 1, pp. 61-72, 2013, doi:10.2478/cait-2013-0006.

[10] A. B. Ahmed & I. S. Elaraby, "Data Mining: A Prediction for Student's Performance Using Classification Method", World Journal of Computer Application and Technology, Vol. 2, No. 2, pp. 43-47, 2014, doi: 10.13189/wjcat.2014.020203

[11] Y. Zhang, S. Oussena, T. Clark & H. Kim, "Use Data Mining to Improve Student Retention in Higher Education", in *Proceeding of the 125h International Conference on Enterprise Information System*, Madeira, Portugal, June 2010.

[12] S. Pal, "Mining Educational Data Using Classification to Decrease Drop Out Rate of Student", International Journal of Multidisciplinary Sciences and Engineering, Vol. 3 No. 5, pp.35-39, May 2012.

[13] C. C. Aggarwal & C. X. Zhai, "A Survey of Text Classification Algorithms", in Mining Text Data, Springer Science Business Media, 2012.

[14] S. Ramasundaram and S.P. Victor, "Algorithms for Text Categorization: A Comparative Study", World Applied Sciences Journal, vol. 22, pp. 1232-1240, 2013.

[15] F. Sebastiani, "Machine Learning in Automated Text Categorization", ACM Computing Surveys, vol. 34, pp. 1-47, March 2002.

[16] A. J. Viera, J. M. Garrett, "Understanding Interobserver Agreement: The Kappa", Family Madicine, vo.37, pp. 360-363, May 2005.