IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 20, Issue 6, Ver. II (Nov - Dec 2018), PP 21-26
www.iosrjournals.org

Classification Algorithms for Predicting Computer Science Students Study Duration

Debby Erce Sondakh
Faculty of Computer Science/ Universitas Klabat, Indonesia)
Corresponding Author: Debby Erce Sondakh

Abstract: *The Department of Computer Science Universitas Klabat offers a bachelor program in Computer Science which should be completed within eight semesters or four years. Some students can accomplish the course in less than four years; however, others had to spend more than the specified period. This study examined students' performance (grades) in the first two semesters. The aim is to develop prediction models for duration study of computer science students. Classification techniques from the Data mining field were applied to develop the models. Three algorithms were selected: Naïve Bayes, decision tree and Support Vector Machine. The Courses, Gender, and Grades (major-grades, basic-grades, and general-grades) are the independent parameters that would predict the dependent variable, i.e. study duration. The resulting models showed that the three algorithms could develop a pretty good study duration prediction models. Decision tree and SVM performance in predicting the instances in Greater categories is better than NB. On the other hand, NB is superior in predicting instances in the Equal category. Nevertheless, the three algorithms suffered in predicting instances in Less category importance.*

Keywords: *Predictive Model, Study Duration, Classification*

Date of Submission: 10-11-2018 Date of acceptance: 25-11-2018

I. Introduction

University database stores academic data that continue to increase over time. It is a challenge for the university to utilize the data to improve the quality of managerial decisions, as well as education performance. Data mining provides the solutions for analyzing large amounts of data and transforms it into useful information and knowledge. Data mining has been used widely used as a robust tool for analyzing academic data and generate a predictive model, a model that can predict future achievement based on past performance.

Universitas Klabat (Unklab) is a private university in North Sulawesi, which has six faculties; including faculty of Computer Science, and one academy. It has an academic information system, called 'Sistem Informasi Unklab' (SIU), with a database that stores academic data of all students. However, these data have not been fully utilized, while they have the potential in providing essential knowledge about students' academic performance.

The Faculty of Computer Science offers a bachelor program, to be completed within eight semesters or four years. The fact is, some students accomplish the course in less than four hours, while some had to spend more than the specified period. This study was conducted to develop the students' academic performance prediction models based on their grades. Three classification algorithms were employed; decision tree, Naïve Bayes (NB), and Support Vector Machine (SVM). The objective is to predict students' study duration based on their academic performance, the grades. This may provide insight for the faculty management staff to properly counsel the students to improve their overall academic performance, to help them completing the course on the specified duration. This paper presents the performance of the decision tree, NB, and SVM.

II. Educational Data Mining

The growth of academic data provide challenge for a higher education institution, not only concerning data storage management but also how to utilize the data appropriately, to enable the improvement of the quality of managerial decisions and the educational performance of students and faculty members. The large number of data is indeed challenging to analyze manually; it takes a long time and complicated process. Data mining provides a solution to this problem. It is also commonly known as knowledge mining, knowledge extraction, information discovery, data analysis [1, 2]. Data mining transforms raw data into useful information and knowledge by adopting techniques and algorithms of multiple science disciplines including databases, statistics, machine learning, and artificial intelligence.

In the educational environment, data mining techniques have been commonly used to extract and retrieve valuable information related to the students, faculties, and management, also to improve the quality of

educational process and institution management. Implementation of data mining in education is called educational data mining (EDM). EDM applies data mining techniques to extract, discover, and learn the knowledge of students' behavior patterns, which have not been identified yet, that are stored in the academic database. EDM intends to identify the relationships among variables related to students learning [3], measure 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].

III. Methodology

This study adopts the hybrid model knowledge discovery process [2]. This hybrid model combines Academic research knowledge discovery models with the Cross-industry standard process for data mining (CRISP-DM), a model from the industrial field. The research was conducted in 5 steps, as depicted in Fig. 1.

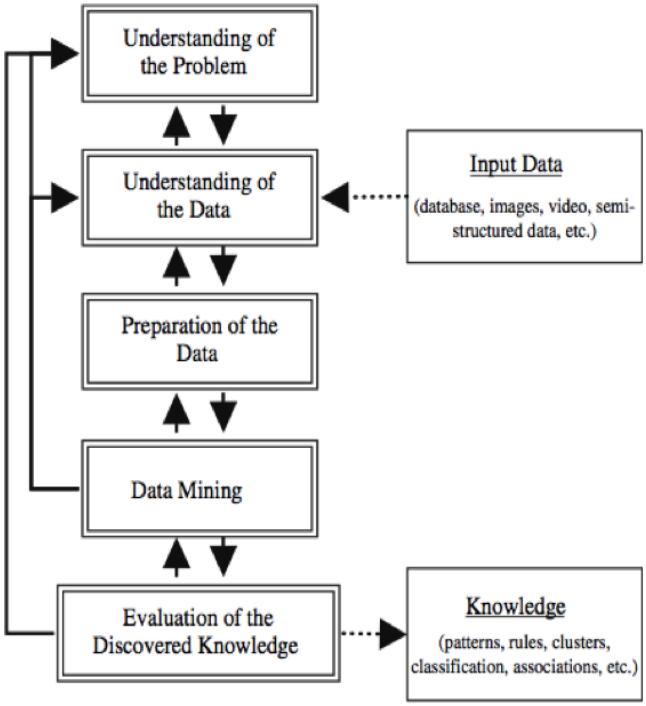

Figure 1: Methodology

1. Understanding the Problem Domain.

The objective of this first step is to understand the scope of the problem to be solved using data mining techniques, and determining the objectives or expected output of the data mining process. Unklab has SIU that deal with the academic process. SIU records students' demographic and academic data, including the Computer Science department students. This study aims to develop predictive models of the Computer Science students' study duration, based on students' performance in the first two semesters, using data mining classification techniques, which are Decision Tree, NB, and SVM. WEKA open source tool will be used. It is a tool written in Java, contains a collection of data mining algorithms used for data analysis designation.

2. Understanding the Data.

This second step manages the data collection and selection. The data format and size are specified. A total of 373 data of Computer Science students, who have completed their degree, are acquired from the 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 stores information about curriculums: ID, course code, course name, credits, and course type.

3. Data Preparation.

This step encounters the extraction and transformation process, to create student grade dataset.

a. Data Extraction. In this step, the Grade and curriculum files were combined into a single file. Five parameters were selected for this study, i.e. program, gender, grade of each subject type (major, basic, and general). Finally, the average grades of each subject type, from the first and second semesters, are calculated. Table 1 presents the parameter chosen. One parameter is added, Duration, to determine the classification category

b. Data Transformation. In Data transformation the numerical values are converted into categorical, as shown in Table 2. The six parameters are grouped into the independent and dependent parameter. Independent parameters, the input for the model, are Program, Gender, M_Grade, B_Grade, and G_Grade. The dependent parameter, role as the output, is Duration.

Table 1: 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

Table 2: Transformation Selected Parameter

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)

4. Data mining.

At this stage, the dataset is analyzed using Weka tool to obtain the predictive models. Table 3 shows the classifiers used to construct the models. The decision tree is a popular classification algorithm. It divides the data into a hierarchical structure called tree. Decision tree classifier comprises *internal nodes* that store 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 utilizes a mixture model, a model that combine terms probability with the category, to predict object category probability [14]. It is based on the Bayes probability theory that assumes the effect of an attribute value of a given class is independent of the values of other attributes [12]. SVM aims to find a boundary, called decision surface or decision hyperplane. It 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].

Table 3: WEKA Classifiers for Building the Models

Algorithm	Classifier
Decision Tree	java weka.classifiers.trees.Id3
Naïve Bayes	java weka.classifiers.bayes.NaiveBayesMultinomial
SVM	java weka.classifiers.functions.SMO

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

IV. Results

At this stage, models were constructed to predict study duration (less than 4 years, equal to 4 years, or more than four years) based on students' major, basic, and general subjects grades in their first and second semesters. The models were constructed using three classification algorithms, decision tree, naïve Bayes, and SVM. The confusion matrix describes the performance of the classification model. The line in confusion matrix represents the actual class, while the columns represent the prediction class.

1. Decision Tree

Table 4 shows, the number correctly classified instances in class Greater is 137, 70% of the total instances in the Greater category. Instances in the Equal category that are correctly classified are 93 (65% of the total instance in Equal category). Decision tree classifier failed to predict instances in the Less category.

Table 4: Confusion Matrix – Decision Tree

n=373	Prediction			Total Correctly Classified	Percent Correctly Classified
	Greater	**Equal**	**Less**		
Greater	137	59	0	137	70%
Equal	51	93	0	93	65%
Less	2	29	0	0	0%

Based on the resulting confusion matrix we can calculate the *true positive rate* (TPR), *false positive rate* (FPR), *true negative rate* (TNR), and *false negative rate* (FNR). These parameters will then determine model accuracy, precision, recall (equal to TPR), f-measure (F-1), and receiver operator characteristics (ROC). From Table 5 we can see the rates of each class. Decision tree produces a model that classifies instances in Greater category with TPR equal to 0.7, precision 0.72, F-1 0.68, and ROC 0.745. Instances in Equal category are classified with TPR, precision, F-1 and ROC values respectively 0.65, 0.51, 0.57, and 0.645. Correspond to Table 4, the correctly classified percentage of instances in Less category equal to 0%, so the value of TPR, precision, and F-1 are zero as well.

Table 5: Decision Tree Classifier Performance

Class	TP Rate	FP Rate	TN Rate	FN Rate	Precision	F-1	ROC
GREATER	0.7	0.3	0.7	0.3	0.72	0.68	0.745
EQUAL	0.65	0.39	0.61	0.35	0.51	0.57	0.645
LESS	0	0	1	1	0	0	0.729

2. Naïve Bayes

As Table 6 shown, the number correctly classified instances in class Greater is 121, 61% of the total instances in Greater category. Instances in the Equal category that are correctly classified are 109 (76% of the total instance in the Equal category). As well as the decision tree, NB classifier also unable to predict instances in the Less category. Table 7 shows NB resulting a model predicts instances in Greater category with recall value equal to 0.61, precision 0.77, F-1 0.68, and ROC 0.757. Instances in the Equal category are classified with TPR, precision, F-1 and ROC values respectively 0.76, 0.51, 0.61, and 0.678.

Table 6: Confusion Matrix – Naïve Bayes

n=373	Prediction			Total Correctly Classified	Percent Correctly Classified
	Greater	**Equal**	**Less**		
Greater	121	75	1	121	61%
Equal	35	109	0	109	76%
Less	1	31	0	0	0%

Table 7: Naïve Bayes Classifier Performance

Class	TP Rate	FP Rate	TN Rate	FN Rate	Precision	F-1	ROC
GREATER	0.61	0.21	0.8	0.39	0.77	0.68	0.757
EQUAL	0.76	0.46	0.54	0.24	0.51	0.61	0.678
LESS	0	0.003	1	1	0	0	0.757

3. SVM

Table 8 and Table 9 show the result of SVM classifier. The number correctly classified instances in class Greater is 137, 70% of the total instances in Greater category. Instances in Equal category that are correctly classified are 83 (58% of the total instance in Equal category). This model is also failed to predict the Less category's instances.

Table 8: Confusion Matrix – SVM

n=373	Prediction			Total Correctly Classified	Percent Correctly Classified
	Greater	**Equal**	**Less**		
Greater	137	60	0	137	70%
Equal	61	83	0	83	58%
Less	7	25	0	0	0%

SVM classifier performance is shown in Table 9. Model resulting using SVM algorithm can predicts instances in Greater category with recall value equal to 0.69, precision 0.67, F-1 0.68, and ROC 0.652. Instances in Equal category are classified with TPR 0.58, precision 0.5, F-1 0.58 and ROC 0.629.

Table 9: SVM Classifier Performance

Class	TP Rate	FP Rate	TN Rate	FN Rate	Precision	F-1	ROC
GREATER	0.69	0.39	0.61	0.3	0.67	0.68	0.652
EQUAL	0.58	0.37	0.63	0.42	0.5	0.58	0.629
LESS	0	0	1	1	0	0	0.457

V. Conclusion

Data mining techniques have been widely used in the educational environment. This ongoing research's goal is to apply data mining, specifically the classification techniques, to examine the Department of Computer Science of Unklab students' performance regarding study duration by looking at their grades in the first two semesters. Three classification algorithms were chosen, namely decision tree, Naïve Bayes (NB), and Support Vector Machine (SVM). The resulting models of the three algorithms set forth fine study duration prediction models. Decision tree and SVM predict the instances in Greater categories superior to NB. On the other hand, NB is superior in predicting instances in the Equal category. Even so, the three algorithms failed to predict instances in the Less category.

As for further research, analyzing the overall performance of each algorithm is needed to compare and find the best algorithm. The study duration prediction model will be constructed using the resulting model of the best one.

References

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

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

[3]. B. K. Baradwaj dan S. Pal, Mining Educational Data to Analyze Students' Performance, *International Journal of Advanced Computer Science and Applications, 2(6),* 2011, 63-39.

[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*, 5(4), 2014, 5987-5991.

[5]. Aziz, N. H. Ismail, dan F. Ahmad, First Semester Computer Science Students' Academic Performances Analysis by Using Data Mining Classification Algorithms, *Proc. of the International Conference on Artificial Intelligence and Computer Science* (AICS 2014), Bandung, Indonesia, 2014. 100-109.

[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, 4(4),* 2013, 1680-1685.

[7]. 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, 2(1),* 2014, 21-46.

[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, 5(1),* 2015, 15-27.

[9]. D. Kabakcieva, Predicting Student Performance by Using Data Mining Methods for Classification, *Cybernetic and Information Technologies 13(1),* 2013, 61-72.

[10]. B. Ahmed & I. S. Elaraby, Data Mining: A Prediction for Student's Performance Using Classification Method, *World Journal of Computer Application and Technology, 2(2),* 2014, 43-47.

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

[12]. S. Pal, Mining Educational Data Using Classification to Decrease Dropout Rate of Students, *International Journal of Multidisciplinary Sciences and Engineering, 3(5),* 2012. 35-39.
[13]. C. C. Aggarwal & C. X. Zhai, A Survey of Text Classification Algorithms, *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, 22,* 2013, 1232-1240.
[15]. F. Sebastiani, Machine Learning in Automated Text Categorization, *ACM Computing Surveys, 34,* 2002, 1-47.

IOSR Journal of Computer Engineering (IOSR-JCE) is UGC approved Journal with Sl. No. 5019, Journal no. 49102.

Debby Erce Sondakh. " Classification Algorithms for Predicting Computer Science Students Study Duration." IOSR Journal of Computer Engineering (IOSR-JCE) 20.6 (2018): 21-26.