Students' Academic Performance Prediction using Data Mining

1st Fergie Joanda Kaunang
Computer Science Department
Universitas Klabat
Minahasa Utara, Indonesia
fergie@unklab.ac.id

2nd Reymon Rotikan
Computer Science Department
Universitas Klabat
Minahasa Utara, Indonesia
reymonr@unklab.ac.id

***Abstract*—Educational Data Mining has been an emerging topic nowadays due to the growth of educational data. This field makes it possible to develop methods in order to find out hidden patterns from educational data. The methods extracted from Educational Data Mining discipline are then used to understand students including their learning behavior as well as to predict their academic performance. This study proposes a model for predicting the academic performance of Computer Science students using Data Mining technique. The data were collected using questionnaires that contain the students' demographics, previous GPA, and family background information. Two Data Mining models (Decision Tree and Random Forest) are applied to the students' data to create the best student's academic performance prediction model. The result of this study shows that Decision Tree is the best model compared to Random Forest by receiving the highest accuracy value of 66.9%. This study shows that there are some relevant feature that influenced student's academic performance.**

Keywords—Educational Data Mining, Machine Learning, Decision Tree, Random Forest, Students' Academic Performance.

I. Introduction

Data mining is a concept for analyzing important information of certain data. It helps to extract hidden pattern and to discover relationships between parameters in a huge amount of data. Nowadays, many researchers adopt Data Mining to solve real-world problems in various area such as marketing, telecommunication, health care, medical, industrial and customer relationship. Data mining and machine learning approach can also be used and has been widely used in bioinformatics field[1-3]. Recently Data Mining has been widely used in educational field[4]. Student's academic performance has become an important part in higher learning institutions. This is because one of the key factor of a high quality learning institutions is based on the record of the students' performance[5]. Students' academic performance prediction is an important concern in the educational field especially education managements. The prediction result could provide an early warning to the students who are at risk by predicting their academic performance[6]. Moreover the prediction result can also be useful in investigating instructor's performance[7]. The Educational Data Mining can be used to develop a prediction model by exploring educational data and extract hidden pattern for predicting students' academic performance using machine learning techniques.

There are various methods to analyze and process the data in Data Mining which include Clustering or Classification, Association Rule, and Sequence Analysis[8-10]. In order to classify every item in a dataset, classification process is used. This is done to predict accurately the target class for every case in the dataset[10]. In this study, we used classification method to be applied to the students' dataset. One of popular technique for prediction is using the Decision Tree algorithm. Most studies are conducted using this method[5, 8, 11, 12] because it is easy to understand for their reasoning process and it can also be directly converted into set of IF-THEN rules[4]. However, based on Shahiri et al. [5] reviewed on the prediction methods used for student performance, Random Forest is not included. Romero et al. [4] predicted students' performance using multiple linear regression model and support vector machine (SVM).

Various previous studies regarding prediction of students' academic performance as well as students' learning behavior are conducted. Study of Meit et al. [13] showed that between male and female students, most of female students are prune to have positive learning styles and behaviors than male students. Cheewaprakobkit [14] conducted a study of 1600 students of Thailand University data to predict most important factors affecting student's academic achievement using the decision tree and neural network method. Moreover, using enrollment data consist of socio-demographic variables (gender, age, work status, class, education and disability) and study environment (course program and course block) Kovacic Z. [15] showed that ethnicity, course program, and course block are the most important attributes for prediction. Therefore, based on the literatures we see an opportunity to conduct a study in predicting students' academic performance. However, most of the previous studies didn't use Random Forest as the classification method. Therefore, this study aims to compare two different classification algorithms; Decision Tree and Random Forest in the prediction of students' academic performance. Furthermore, factors influencing students' academic performance can be seen through this study.

II. Educational Data

Educational data come with different granularity and format. It can be collected from various sources. The educational data used in this study were real-world data from

Table 1. Student related variables

Attributes	Description
Gender	Student's gender (1: Male or 2: Female)
Year	Student's academic year (numeric: from 1 to 4)
GPA	Student's previous semester GPA
Fedu	Father's education (numeric: from 0 to 4[a])
Medu	Mother's education (numeric: from 0 to 4[a])
Freetime	Free time after school (numeric: from 1 to 5[b])
Goout	Time spent going out with friends (numeric: from 1 to 5[b])
Famrel	Quality time spent with family (numeric: from 1 to 5[b])
Studytime	Weekly study time (numeric: from 1 to 4[c])

a 0 = none; 1 = elementary school; 2 = junior high school; 3 = senior high school; 4 = higher education

b 1 = very low; 2 = low; 3 = average; 4 = high; 5 = very high

c 1 = less than 2 hours; 2 = 2 – 5 hours; 3 = 5 – 10 hours; 4 = more than 10 hours

Computer Science Department of Universitas Klabat collected from questionnaires distributed to all the students of the related department. The questionnaire itself had been reviewed by lecturers who are professionals at this fields. The final version of the questionnaire consisted of 7 questions in a single A4 paper and distributed to 300 students. 51 answers were removed due to lack of details which is important for the result of the study. Eventually, the data was integrated into one dataset consist of 249 records with 3 different classes. The dataset of this study is a composite of background and demographic data, personal data, and past grade data provided in Table 1. The reason behind the selection of students' demographic data as the attribute is because male and female students have different learning styles[5]. The background and demographic data include gender, family relationship, father educational status, mother educational status. Personal data of this study consist of time they spent with their friends, weekly study time, and their free time after school. For past grade data, we asked them to write down their GPA of last semester of their study (previous semester before this study is conducted). The students' GPA is further selected as dependent parameter which are categorized into three different classes: 1 = poor, 2 = average, and 3 = good. This is done by doing a data normalization to scale the GPA of the student. We scaled the students' GPA data according to Table 2.

Table 2. Students' GPA category (class)

Students' GPA	Category (class)
0 – 2.32	Poor
2.33 – 3.66	Average
3.67 – 4.0	Good

III. Prediction Model

The analytical flowchart of this work is described in this section using selected classification techniques. The flowchart shows the steps involved in developing the prediction model. Figure 1 shows the process flow of obtaining the knowledge as the result of this study.

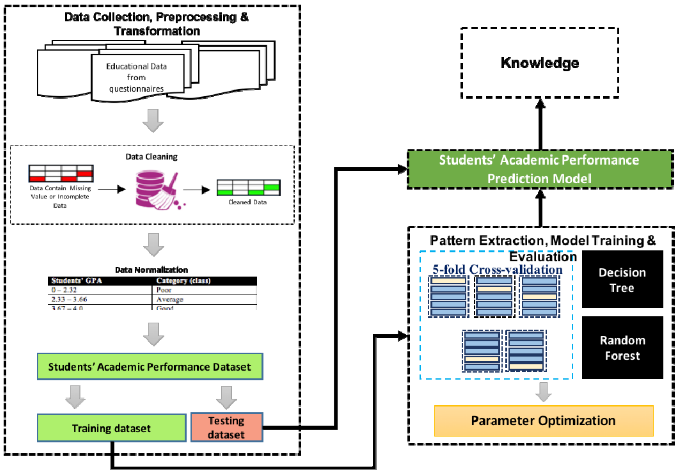
Data Collection, Preprocessing & Transformation
Educational Data from questionnaires
Data Cleaning
Data Contain Missing Value or Incomplete Data
Cleaned Data
Data Normalization
Students' GPA
Category (class)
0 – 2.32
Poor
2.33 – 3.66
Average
3.67 – 4.0
Good
Students' Academic Performance Dataset
Training dataset
Testing dataset
Knowledge
Students' Academic Performance Prediction Model
Pattern Extraction, Model Training & Evaluation
5-fold Cross-validation
Decision Tree
Random Forest
Parameter Optimization

Figure 1 Students' Academic Performance Prediction Model Flow Chart

A. Data Collection, Preprocessing and Transformation

The dataset used in this study were collected from students of Computer Science Department of Universitas Klabat. The detail of how the educational data were collected has already explained in previous section. As mentioned in Educational Data Section, the data attributes used in this study are students' GPA and students' background and demographic data. These attributes are chosen based on previous studies which they consider students' GPA affects their future educational and career mobility. While the background and demographic data is because students, female and male, have different learning style[5].

After the data were collected, the data cleaning process was done to remove missing or incomplete data resulting in cleaned data. This cleaned data is then being normalized where the numerical values of students' GPA were transformed into categorical class based on Table 2. After that, the final students' academic performance dataset was split into two different parts which are training dataset and testing dataset using 70:30 ratio. 70% of training dataset and 30% of testing dataset.

B. Pattern Extraction, Model Training and Evaluation

The next stage of the prediction model flow is pattern extraction, model training and evaluation process. The training dataset from the previous stage were used to build the model using classification techniques. In this stage the data mining process was performed. In order to optimized the parameter of the classifier, we conduct 5 and 10 fold cross validation. However, this k-fold cross validation is conducted to reduce the bias in our experiments as well as to help us to further choose the best fold to build our prediction model. Furthermore, the testing dataset used to measure and validate the students' academic performance prediction model. The result acquired from this stage will be then evaluated and depicted as knowledge.

IV. Results and Discussions

A. Environments and Evaluation Measures

We conducted our study using WEKA [16] in order to evaluate the proposed prediction model. To evaluate the quality of the classification, four different metrics were used: Accuracy, Precision (Specificity), Recall (Sensitivity), and F-Measure [17, 18]. Accuracy is the ratio of the total number of correctly predicted instances to the total number of instances. Precision is the ratio of correctly predicted positive instances to the total predicted positive instances. Recall is the ratio of correctly predicted positive instances to the all instance in actual class. Finally, we compute the F-measure to combine precision and recall in order to show the harmonic mean. We calculate these four metrics based on the following equation using the confusion matrix.

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$S_p = \frac{TN}{TN + FP} \tag{2}$$

$$S_n = \frac{TP}{TP + FN} \tag{3}$$

$$F - measure = 2\frac{Sp * Sn}{Sp + Sn} \tag{4}$$

B. Results

After the classification techniques were applied to the dataset, there are different results obtained. Table 3 shows the classification method results using 5 and 10-fold cross validation.

Table 3 Classification Method Results Using K-fold Cross Validation

Evaluation Metrics	Decision Tree		Random Forest	
	5-fold CV	10-fold CV	5-fold CV	10-fold CV
Accuracy	62.85%	**66.85%**	61.14%	61.14%
Precision	0.543	**0.622**	0.572	0.573
Recall	0.629	**0.669**	0.611	0.611
F-measure	0.565	**0.632**	0.588	0.589

Decision Tree classifier generates better results compared to Random Forest classifier. It gives 66.85% of accuracy which indicates that there are 117 students are correctly classified to the right class (Poor, Average, and Good) out of 175 students in the training dataset. In addition, the predictive performance of the built training model yielding a stable performance of 0.622, 0.669, and 0.632 of precision, recall, and F-measure value respectively. This makes Decision Tree is the best model to our prediction model.

Furthermore, in order to validate our prediction model, using the testing dataset we conduct an independent test using Decision Tree algorithm only since it shows better results than Random Forest. As seen in Table 4, our model yields 66.2% of accuracy, 0.594 precision value, 0.635 of recall, and 0.614 of F-measure. This shows that our model provide a fair results. The precision and recall performance shows balance performance which indicates that there is no overfitting of the training dataset to the testing dataset.

Table 4 Classification Results Using Testing Dataset

Evaluation Metrics	Decision Tree
Accuracy	**63.51%**
Precision	**0.594**
Recall	**0.635**
F-measure	**0.614**

C. Discussions

The results of our experiment reveal that the Decision Tree model generates a more accurate prediction than Random Forest model. Moreover, we found that factors influencing students' academic performance are as follows:

1. Number of time spent with family.
2. Parents' Education (father and mother's education background).
3. Number of time spent going out with friends.
4. Number of study time per week.

The results from the test are: the Decision Tree algorithm consider year attribute as the most relevant features as it appears at the root of the tree. There are some interesting knowledge that can be extracted from the tree on Figure 2. For example the third and fourth year students showed average academic performance over all for there is no other inputs considered are used. Another interesting knowledge is that father and mother's education plays important roles to the first year students to achieve good academic performance. It also showed that the female first year students tend to attain better academic performance than the male students. Moreover, another attribute like study time and family relation play important role for students with average academic performance.

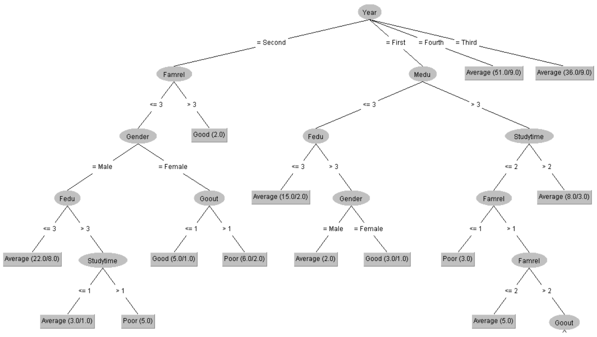

Figure 2 Decision Trees

However, different interesting knowledge obtained for second year students. It showed that father's education plays more important role than mother's education. Moreover, the tree showed that those male students who spent less than two hours of study time weekly tend to achieve average academic performance. Nevertheless, for female students going out time become the most influential attribute. Those who spent less time going out with friends tend to achieve good academic performance.

Our experiment shows that gender also plays an important role in predicting students' academic performance. This reveal that our study is in harmony with the study of Meit et al. [13] which shows that male and female students have different learning behavior. Moreover, our other attributes such as family relationship, go out time, study time, as well as parents' educational background are also affect the academic performance of a student.

V. Conclusions and Future Work

In this study, we have proposed a new students' academic performance prediction model. This study conducts a comparative analysis of two classification techniques between Decision Tree and Random Forest using WEKA. The results show that Decision Tree has better classification performance than Random Forest. There are interesting knowledges provided in the previous section of this paper. Overall, the results shows those factors that affect the learning behavior of students. For example, the first year students who spent too much time going out with their friends prone to show poor academic performance as well as if they spent very few time with their family. The discovered knowledge obtained from our experiments can be taken into consideration for students as their learning behavior evaluation, for lecturers as their reference of students' background as well as students' learning behavior, and for parents as well.

This study provides results that can contribute to conduct further study such as expanded dataset with more attributes for better results. More classification techniques can also be used for more comparative analysis and accurate results.

Acknowledgment

The authors would like to thank Universitas Klabat especially Computer Science Department for the permission to collect the data used in this study.

References

[1] H.-J. Kao, S.-L. Weng, K.-Y. Huang, F. J. Kaunang, J. B.-K. Hsu, C.-H. Huang, *et al.*, "MDD-carb: a combinatorial model for the identification of protein carbonylation sites with substrate motifs," *BMC systems biology,* vol. 11, p. 137, 2017.

[2] J. T. Wang, M. J. Zaki, H. T. Toivonen, and D. Shasha, "Introduction to data mining in bioinformatics," in *Data Mining in Bioinformatics*, ed: Springer, 2005, pp. 3-8.

[3] E. Frank, M. Hall, L. Trigg, G. Holmes, and I. H. Witten, "Data mining in bioinformatics using Weka," *Bioinformatics,* vol. 20, pp. 2479-2481, 2004.

[4] C. Romero and S. Ventura, "Educational Data Mining: A Review of the State of the Art," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews),* vol. 40, pp. 601-618, 2010.

[5] A. M. Shahiri, W. Husain, and N. a. A. Rashid, "A Review on Predicting Student's Performance Using Data Mining Techniques," *Procedia Computer Science,* vol. 72, pp. 414-422, 2015/01/01/ 2015.

[6] B. Guo, R. Zhang, G. Xu, C. Shi, and L. Yang, "Predicting Students Performance in Educational Data Mining," in *2015 International Symposium on Educational Technology (ISET)*, 2015, pp. 125-128.

[7] A. M. Ahmed, A. Rizaner, and A. H. Ulusoy, "Using data Mining to Predict Instructor Performance," *Procedia Computer Science,* vol. 102, pp. 137-142, 2016/01/01/ 2016.

[8] F. Ahmad, N. H. Ismail, and A. A. Aziz, *The prediction of students' academic performance using classification data mining techniques* vol. 9, 2015.

[9] D. S. D. Miss Pooja M. Dhekankar, "Analysis of Student Performance by using Data Mining Concept," *International Journal on Recent and Innovation Trends in Computing and Communication (IJRITCC),* vol. 3, pp. 2942 - 2944, May 2015.

[10] G. Kesavaraj and S. Sukumaran, "A study on classification techniques in data mining," in *2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT)*, 2013, pp. 1-7.

[11] M. M. N. Quadri and D. N. V. Kalyankar, "Drop Out Feature of Student Data for Academic Performance Using Decision Tree Techniques," *Global Journal of Computer Science and Technology,* 2010-04-05 2010.

[12] S. Natek and M. Zwilling, "Student data mining solution–knowledge management system related to

higher education institutions," *Expert Systems with Applications,* vol. 41, pp. 6400-6407, 2014/10/15/ 2014.

[13] N. J. B. L. A. E. Scott S. Meit, "Personality Profiles of Incoming Male and Female Medical Students: Results of a Multi-Site 9-Year Study," *Medical Education Online,* vol. 12, 2007.

[14] P. Cheewaprakobkit, "Study of Factors Analysis Affecting Academic Achievement of Undergraduate Students in International Program," in *Proceedings of the International MultiConference of Engineers and Computer Scientists*, 2013, pp. 13-15.

[15] Z. Kovacic, "Early prediction of student success: Mining students' enrolment data," 2010.

[16] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *ACM SIGKDD explorations newsletter,* vol. 11, pp. 10-18, 2009.

[17] D. M. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," 2011.

[18] E. A. Amrieh, T. Hamtini, and I. Aljarah, "Preprocessing and analyzing educational data set using X-API for improving student's performance," in *Applied Electrical Engineering and Computing Technologies (AEECT), 2015 IEEE Jordan Conference on*, 2015, pp. 1-5.