The 4th International Conference on Digital Arts, Media and Technology
and 2nd ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering

Using Classification Data Mining Techniques for Students Performance Prediction

Nongnuch Ketui*
Computer Science, Faculty of Science and Agricultural Technology, Rajamangala University of Technology Lanna Nan, Thailand
*nongnuchketui@rmutl.ac.th

Warawut Wisomka
Computer Science, Faculty of Science and Agricultural Technology, Rajamangala University of Technology Lanna Nan, Thailand
i-y@hotmail.com

Kanitha Homjun
Computer Science, Faculty of Science and Agricultural Technology, Rajamangala University of Technology Lanna Nan, Thailand
kanithahomjun@rmutl.ac.th

***Abstract*—The big problem of retired or drop out students is about academic achievement. The educational institution needs to follow up the advising system since advisers should guide the planning curriculum to their advisees. The data mining techniques are applied to anaylze the performance of the students and to impart the quality of education in the educational institutions. This paper focuses on classification models for applying in Education Data Mining. The classification models are applied to identify the suitable subject to the science students. The experiment is set to improve the student performance which comparing the performance of five classification models and then predicting the appropriated academic achievement in each major. To examine the experiment, we used 17,875 academic achievements within 483 students. Four measures; precision, recall, f-measure, and accuracy are evaluated the models. For the result, the best accuracy is Gradient Boosted Trees: GBT at 92.41% and F-measure value equals to 84.59%.**

***Index Terms*—Decision Tree, ID3, Random Tree, Gradient Boosted Trees, Education Mining**

I. Introduction

Education is an essential factor for developing a country. Academic achievement is the part evaluation in education management. An advisor system could help the student to get the performance better since the number of retired or dropped students out is less. If the curriculum planning in each semester will be proposed quickly, the number of retried students possibly decrease. Data mining technology can apply in the education for improving a learning such as the associating a knowledge gaps, the predicting the student achievement, and the discovering the suitable major. Many tools are used in data mining for analysing a large dataset and finding meaningful data called knowledge. This help us to reach the meaningful conclusion. Initially, the application of data mining were restricted to business domain but know it is extended to education and is known as EDM.

Educational data mining (EDM) deals with the application of data mining tools and techniques to inspect the data at educational institutions for deriving knowledge [1]. In order to analyze large amount of education information, the area of Knowledge Discovery in Databases (KDD) provides methods at the point of machine learning, statistics and database systems. KDD is interesting to extract the important pattern from the large databases. While an efficient algorithms is a core of application to detect the desired patterns contained within the given data.

In machine learning and data mining, the classification is the problem of recognizing to a new case belongs to a set of categories based on a training dataset which containing the category membership as known. Examples are competing a baseball game to the ”yes” or ”no” class, and assigning a major to a given student based on observed attributes of the student (sex, subject, grade, or the relative attribute, etc.).

A decision tree is used in a decision tree learning. The branches in the tree are represented theirs item value which belong to the characteristic, the attribute, or the feature. While the leave nodes show the item’s target value (class or label). The regression method is the one type of decision tree that considers the real numbers as the target variable.

This paper studies the classification techniques in data mining. Five classification models are used to predict the appropriated subject in each major and then compares the performance with four measures; (1) precision, (2) recall, (3) f-measure, and (4) accuracy scores. The purpose of this work, we introduce data mining process in Section III, represent the classification model in Section III-A. Method and procedure are shown in Section IV. Experimental results and discussions are described in Section V. Finally, conclusion and future work are represented in Section VI.

II. Related Work

The way how to discover the useful information from the large dataset is called Knowledge Discovery in Databases (KDD). Educational Data Mining (EDM) is the part of KDD which helps the student improving the performance. Many data mining approaches are used to study in EDM such as the classification algorithm, the association rule discovery, or the clustering. For the classification algorithm, many researchers reviewed that the decision tree algorithm is quite to use for predicting on student’s academic performance [3]. They implemented various decision tree algorithm ID3, J48/C4.5, random tree, multilayer perception, rule based and random forest with Weka tool. To evaluate the performance percentage

978-1-5386-8072-8/19/$31.00 ©2019 IEEE

split method or cross validation method is used. The results could improve the student's academic achievement. In the research [4], they studied the the capabilities of information mining within the context of the next Education Institute and tries to get new specific information by applying data processing techniques to academic knowledge of Technological academic Institute of Athens. They used the decision tree algorithm to analyze the student's questionnaires within the categories inside the analysis method of every department of the Institute. Other hands, Ahmed and his team [5] focused on the classification models to predict the final grade of students. They selected the decision tree (ID3) method for their experiment. Including to the research [6] of Kaur and his team, they studied the classification and prediction based data mining algorithms to predict slow learners in education sector. They applied applied on various classification algorithms such as Multilayer Perception, Naïve Bayes, SMO, J48 and REPTree using WEKA potential tool. The latest research is created by the new framework, called cloud-service decision tree classification for education platform [7]. This research studied the multi-objective weight self-adaptation form and proposed the cloud-service classification with decision tree algorithm. Then, their algorithm is verified by the measures.

III. Data Mining

Data Mining (DM) or Knowledge Discovery in Database (KDD) is a process to find patterns and relation from large data sets with the various methods such as machine learning, statistics, and database systems [16]. There are three processes of data mining as the described following:

1) Preprocessing: prepare the suitable dataset and will arrange in a useful format. A noise data should be eliminated from raw dataset. The remaining dataset which having a relation, will be used for creating a model.
   - Data cleaning: eliminate a unimportant dataset or an error data.
   - Data integration: collect the dataset from multiple sources.
   - Data selection: select the relative attribute which can be predicted the good performance.
   - Data transformation: adjust the format of attributes that is easy to use or predict.
2) Modeling: use the selected data for creating a model. A problem or business understanding is guided to choose the appropriate technique. Then, the model should be evaluated with the new cases.
3) Postprocessing: deploy the model to the real situation as the application.

A. Classification Methods

The classification [8] is called supervision which uses the training data (observations, measurements, etc.) are set by labels showing the class of the observations. The training set classifies a new data as the following class. The categorical class is labelled with the model construction based on the training set within a classifying attribute. Then, a new data is predicted by training model. There are many methods for classification. We can study four examples in our research as the following;

1) Decision Tree: a collection of nodes which the nodes describe the characteristic, branches display the testing result, and the leaf nodes represent the specific attribute [9].
2) Decision Tree Weight-based: a concept of the decision tree with weight-based works as the decision tree method but it concentrates the attribute test condition instead of the information gain [10].
3) Iterative Dichotomiser 3 (ID3): an algorithm [11] is the original of C4.5 algorithm which generating a decision tree from dataset. The machine learning and natural language processing domains quite use for analysing.
4) Random Tree: The Random Tree operator [12] focuses on a random subset of attributes before it is applied. The subset ratio parameter as the stochastic process represents the size of the subset [13].
5) Gradient Boosted Trees: GBT algorithm is a machine learning technique for producing a regressive or classified model in the form of an ensemble. A gradient boosted model is the forward-learning ensemble method that obtain predictive results through improved performances [14].

IV. Method and Procedure

In this research, we study the classification methods to identify the suitable subject to each major in science students and then evaluate the performance with the precision, recall, f-measure, and accuracy. The experiment in this research, we focus on the data mining process such as preprocessing, modeling, and postprocessing which are described below;

A. Preprocessing

- Data cleaning: 17,875 records within 483 persons examine the performance of academic achievement, we only focus on the students of Faculty of Science and Agricultural Technology from Rajamangala University of Technology Lanna Nan from 2009 to 2013 since the curriculum is the same period.
- Data selection: the registration data is used to analyze such as year, semester, subject id, subject name, department id, department name, grade, and credit are used. However, the experimental dataset only selects the records which having the active status ('Y'). Grade 'I' is eliminated while the missing values are set to 'N/A'.
- Data transformation: A to F grade level are changed into three group i.e., 'High' equals A, B+, or B level, 'Middle' displays C+ or C, and 'Low' represents D+, D, or F. Then, we prepare the dataset in a excel file for RapidMiner tool [18].

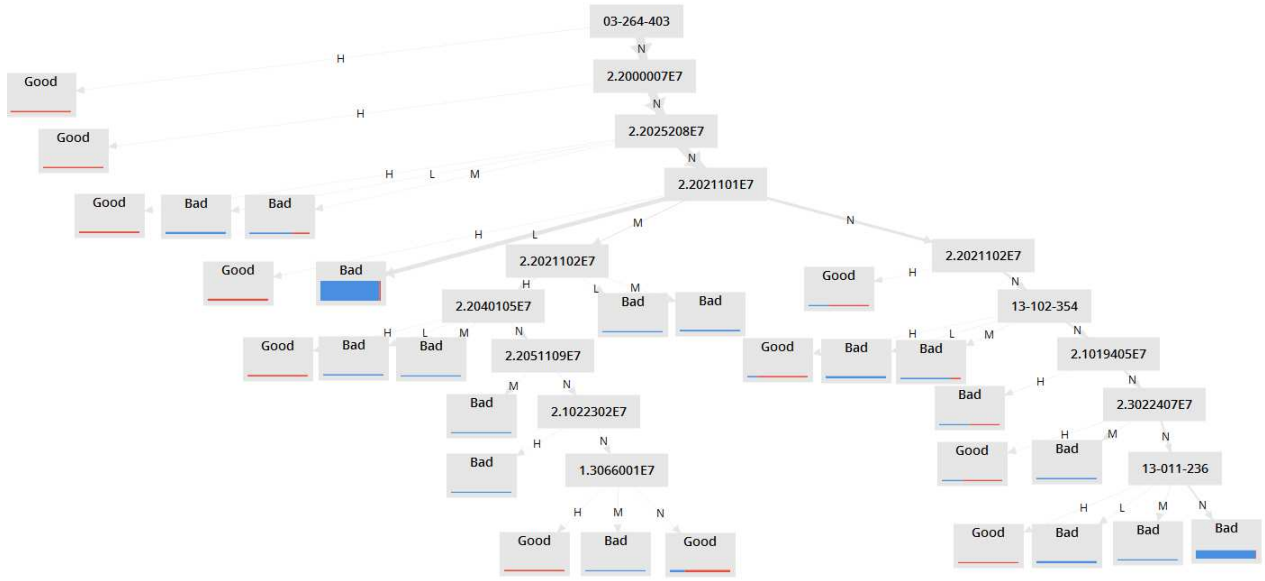

Fig. 1. The Tree Model from Rapid Miner Environment

TABLE I
THE PERFORMANCE OF FIVE CLASSIFICATION MODELS

Model	Precision	Recall	F-measure	Accuracy
DT	86.75	77.03	81.60	91.03
ID3	80.84	**79.74**	80.29	89.66
Random Tree	42.07	50.00	45.69	84.14
GBT	**92.62**	77.85	**84.59**	**92.41**
DT-Weight	42.07	50.00	45.69	84.14

B. Modeling

In this experiment, five classification methods (Decision Tree, Decision Tree Weight-based, ID3, Random Tree, and GBT) used to create the models and then compare the performance with the measures. Here, the model will be shown in Tree format and generated into the rules later. Each of leaf node is classified into two classes (Good and Bad). Here, the *good* class means the GPA of student which is over than 40% at the top ranking. The opposite of the *good* class, the *bad* class equals to the accumulated GPA which is under the 40% of the below ranking. The 40% value can be changed since it depends on the frequency of data. In this experiment, the dataset is divided by a cross-validation test called 10-fold cross-validation. The training set is used to create the model and the testing set predicts the class, respectively. The tree model shows in Figure 1. It is changed the tree to rules as shown in Figure 2. We can use the rules to implement or predict the new case (the student achievement). After getting the models, the performance measures used in our experiment are precision, recall, f-measure, and accuracy [19]. Eq. (1) shows the calculation of F-measure value.

$$F - measure = \frac{2 \times \ Precision \times Recall}{Precision + Recall} \tag{1}$$

C. Postprocessing

In this process, the experiment considers the subjects which are suitable to the science students in order to the condition or class that occurring from a root node to the leaf node. If the number of the good class is more than the bad, it is possible that the new student who has the same record may gets the good achievement as well. In Figure 1, the number of good class equals to 15 which is less than the bad class (11 classes). The number is quite similar so that if the student will attend the class carefully, the performance will be increase or good. We can start the output from the root to the leaf node. For example, in the Figure 1 shows the root node is the subject code.03264403. If the students get the grade in high level (B, B+, and A) of the subject code.03264403 (Farm Field and Plant Food Management) then they will get the good achievement or this subject is suitable to the science students. In opposite of the high grade, the students have not studied in the 03264403 (N value) but they get the high achievement in the 22000007. The output class is shown the good. The other case, the students have not studied the subject code.03264403 and 22000007 and gotten the middle grade (C and C+) in the subject code.22025208 so that the class will quite return the bad class. For the Figure 2, the rules of those conditions are generated respectively as below;

1) Class Good: If 03264403 in {High}

Tree

```
03-264-403 = H: Good {Bad=0, Good=5}
03-264-403 = N
|   2.2000007E7 = H: Good {Bad=0, Good=4}
|   2.2000007E7 = N
|   |   2.2025208E7 = H: Good {Bad=0, Good=9}
|   |   2.2025208E7 = L: Bad {Bad=14, Good=0}
|   |   2.2025208E7 = M: Bad {Bad=5, Good=2}
|   |   2.2025208E7 = N
|   |   |   2.2021101E7 = H: Good {Bad=0, Good=13}
|   |   |   2.2021101E7 = L: Bad {Bad=224, Good=5}
|   |   |   2.2021101E7 = M
|   |   |   |   2.2021102E7 = H
|   |   |   |   |   2.2040105E7 = H: Good {Bad=0, Good=7}
|   |   |   |   |   2.2040105E7 = L: Bad {Bad=5, Good=0}
|   |   |   |   |   2.2040105E7 = M: Bad {Bad=2, Good=0}
|   |   |   |   |   2.2040105E7 = N
|   |   |   |   |   |   2.2051109E7 = M: Bad {Bad=2, Good=0}
|   |   |   |   |   |   2.2051109E7 = N
|   |   |   |   |   |   |   2.1022302E7 = H: Bad {Bad=2, Good=0}
|   |   |   |   |   |   |   2.1022302E7 = N
|   |   |   |   |   |   |   |   1.3066001E7 = H: Good {Bad=0, Good=6}
|   |   |   |   |   |   |   |   1.3066001E7 = M: Bad {Bad=2, Good=0}
|   |   |   |   |   |   |   |   1.3066001E7 = N: Good {Bad=4, Good=12}
|   |   |   |   |   2.2021102E7 = L: Bad {Bad=3, Good=0}
|   |   |   |   |   2.2021102E7 = M: Bad {Bad=8, Good=0}
|   |   |   |   2.2021101E7 = N
|   |   |   |   |   2.2021102E7 = H: Good {Bad=1, Good=2}
|   |   |   |   |   2.2021102E7 = N
|   |   |   |   |   |   13-102-354 = H: Good {Bad=1, Good=5}
|   |   |   |   |   |   13-102-354 = L: Bad {Bad=15, Good=0}
|   |   |   |   |   |   13-102-354 = M: Bad {Bad=5, Good=1}
|   |   |   |   |   |   13-102-354 = N
|   |   |   |   |   |   |   2.1019405E7 = H: Bad {Bad=1, Good=1}
|   |   |   |   |   |   |   2.1019405E7 = N
|   |   |   |   |   |   |   |   2.3022407E7 = H: Good {Bad=1, Good=2}
|   |   |   |   |   |   |   |   2.3022407E7 = M: Bad {Bad=5, Good=0}
|   |   |   |   |   |   |   |   2.3022407E7 = N
|   |   |   |   |   |   |   |   |   13-011-236 = H: Good {Bad=0, Good=2}
|   |   |   |   |   |   |   |   |   13-011-236 = L: Bad {Bad=14, Good=0}
|   |   |   |   |   |   |   |   |   13-011-236 = M: Bad {Bad=3, Good=0}
|   |   |   |   |   |   |   |   |   13-011-236 = N: Bad {Bad=89, Good=1}
```

Fig. 2. The Ruled Tree Model of Each Subject

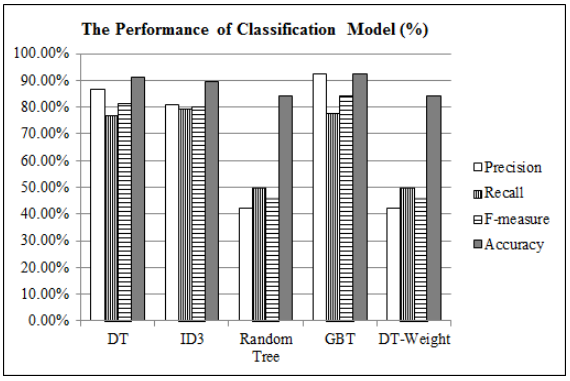

Fig. 3. A Graph of Precision, Recall, F-measure, and Accuracy Score

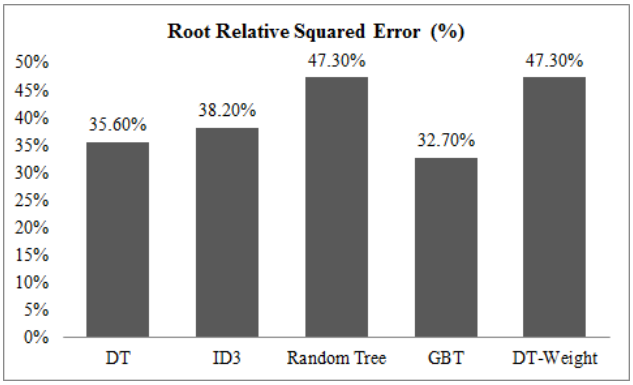

Fig. 4. Comparison of the Root Relative Squared Error

2) Class Good: If 03264403 in {None} and 22000007 in {High}

3) Class Bad : If 03264403 in {None} and 22000007 in {None} and 22025208 in {Middle}.

4) Class Bad : If 03264403 in {None} and 22000007 in {None} and 22025208 in {None} and 22021101 in {Low}.

V. Result and Discussion

To examine the root relative square error and compare the performance of five models in the classification for each major. Each of the classification model should consider the root relative square error (see in Figure 4) since the root relative square error can describe the relied performance of each model in the percentage. According to Figure 3, it compares the performance of each classification models. Each measures are evaluated and displayed in the figure. Here, the light bar shows the precision, the vertical bar displays the recall, the horizontal bar represents the f-measure, and the gray bar illustrates the accuracy value. In the Figure 4 shows that the GBT has minimum root relative square error in order to the maximum accuracy (see the gray bar in Figure 3). Both of the random tree algorithm and decision tree with weight based get the maximum root relative square error. So that, the precision, recall, f-measure, and accuracy are the same scores.

The performance of five classification models is represented in Table I. The experiment found that the GBT algorithm gets the highest precision while the random tree and decision tree with weight-based are lowest. The recall score, the maximum is ID3 (79.74%) while the recall of decision tree (77.03%) is close to the GBT (77.85%). To consider the f-measure score which is calculated from the ratio of precision and recall scores. The result displayed that the highest performance is the GBT (84.59%) while the decision tree equals to the second ranking (81.60%). According to the accuracy, the maximum score equals to 92.41% of the GBT algorithm. Finally, the gradient boosted tree: GBT is the best classification model for considering the suitable subject to each major.

VI. Conclusion and Future Work

In our research, we applied the data mining classification techniques for improving the student performance. We focus on the classification for predicting the suitable subject in the faculty of science and agricultural technology student. Then, the famous evaluations; the precision, recall, f-measure, and accuracy scores are used to measure the performance of each model. To use 17,875 records of student registration in Faculty of Science and Agricultural Technology, Rajamangala University of Technology Lanna Nan. The classification results illustrated that the best model is the gradient boosted tree: GBT algorithm at 92.41%, and the decision tree model get the second ranking (91.03%). While the lowest is both of the random tree and decision tree with weight-based. In the future works, we will evaluate with other data mining technique for improving the student performance.

Acknowledgment

We would like to give a big clap to registration staffs at Education Division of Rajamangala University of Technology Lanna Nan (RMUTL Nan) for preparing the raw dataset. We give a special thank to Computer Center of Rajamangala University of Technology Lanna Nan (RMUTL Nan) for managing the database server in order to clapping my hand to lecturers and students at Science department, RMUTL Nan for all kind helps.

References

[1] M. Al-Razgan, A. S. Al-Khalifa, and H. S. Al-Khalifa, “Educational data mining: A systematic review of the published literature 2006- 2013”, In Proceedings the 1st International Conference on Advanced Data and Information Engineering, 2013, pp.711–719.
[2] R.h Agrawal and R. Srikant, “Fast algorithms for mining association rules”. Proceedings of the 20th International Conference on Very Large Data Bases, VLDB, pp. 487–499, Santiago, Chile, September 1994.
[3] Kaur, Hardeep. “A Literature Review from 2011 TO 2014 on Student’s Academic Performance Prediction and Analysis using Decision Tree Algorithm.” Journal of Global Research in Computer Science 9.5 (2018): 10-15.
[4] Rao, NV Krishna, et al. “A Review on Data Mining Approach used in Education Data Mining using Decision Tree Algorithm.” (2018): 1735-1738.
[5] Ahmed, Abeer Badr El Din, and Ibrahim Sayed Elaraby. “Data Mining: A prediction for Student’s Performance Using Classification Method.” World Journal of Computer Application and Technology 2.2 (2014): 43-47.
[6] Kaur, Parneet, Manpreet Singh, and Gurpreet Singh Josan. “Classification and prediction based data mining algorithms to predict slow learners in education sector.” Procedia Computer Science 57 (2015): 500-508.
[7] Chao, Wang, and Wang Junzheng. “Cloud-service decision tree classification for education platform.” Cognitive Systems Research 52 (2018): 234-239.
[8] J. Han, M. Kamber, Data Mining Concepts and Techniques, Morgan Kaufmann, 2000.
[9] B. Kamiński, M. Jakubczyk, P. Szufel. “A framework for sensitivity analysis of decision trees”. Central European Journal of Operations Research, 2017.
[10] Markus Hofmann, Ralf Klinkenberg,“RapidMiner: Data Mining Use Cases and Business Analytics Applications (Chapman and Hall/CRC Data Mining and Knowledge Discovery Series)”, CRC Press, October 25, 2013.
[11] J. R. Quinlan. “Induction of Decision Trees”. Mach. Learn. 1, 1 (Mar. 1986), 81–106.
[12] Joseph L. Doob. “Stochastipoic processes”. Wiley. p. 46 and 47, 1990.
[13] Ionut Florescu. “Probability and Stochastic Processes”. John Wiley and Sons. pp. 294 and 295, 7 November 2014, ISBN 978-1-118-59320-2.
[14] J. H. Friedman. “Greedy Function Approximation: A Gradient Boosting Machine”. February 1999 (PDF).
[15] C. Shearer, “The CRISP-DM model: the new blueprint for data mining”, Journal of Data Warehousing, vol.5, 2000, pp.13–22.
[16] “Data Mining Curriculum”, ACM SIGKDD. 2006-04-30. Retrieved 2014-01-27.
[17] G. Piatetsky-Shapiro, Discovery, analysis, and presentation of strong rules, in Piatetsky-Shapiro, Gregory; and Frawley, William J.; eds., “Knowledge Discovery in Databases”, AAAI/MIT Press, Cambridge, MA, 1991.
[18] M. Hofmann and R. Klinkenberg. “Rapidminer: Data Mining Use Cases and Business Analytics Applications”. Chapman and Hall/CRC, 2013.
[19] Powers, David M W. “Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation” (PDF). Journal of Machine Learning Technologies. 2 (1): 37–63, 2011.