Applied Mathematical Sciences, Vol. 9, 2015, no. 129, 6415 - 6426
HIKARI Ltd, www.m-hikari.com
http://dx.doi.org/10.12988/ams.2015.53289

The Prediction of Students' Academic Performance Using Classification Data Mining Techniques

Fadhilah Ahmad*, Nur Hafieza Ismail and Azwa Abdul Aziz

Faculty of Informatics and Computing
Universiti Sultan Zainal Abidin (UniSZA), Kuala Terengganu, Malaysia
*Corresponding author

Abstract

Data Mining provides powerful techniques for various fields including education. The research in the educational field is rapidly increasing due to the massive amount of students' data which can be used to discover valuable pattern pertaining students' learning behaviour. This paper proposes a framework for predicting students' academic performance of first year bachelor students in Computer Science course. The data were collected from 8 year period intakes from July 2006/2007 until July 2013/2014 that contains the students' demographics, previous academic records, and family background information. Decision Tree, Naïve Bayes, and Rule Based classification techniques are applied to the students' data in order to produce the best students' academic performance prediction model. The experiment result shows the Rule Based is a best model among the other techniques by receiving the highest accuracy value of 71.3%. The extracted knowledge from prediction model will be used to identify and profile the student to determine the students' level of success in the first semester.

Keywords: Educational data mining; Decision Tree; Naïve Bayes; Rule Based; students' academic performance

1. Introduction

Data Mining (DM) concept is to extract hidden pattern and to discover relationships between parameters in a vast amount of data. There are many achievements of DM techniques in many areas such as engineering, education, marketing, medical, financial, and sport. It shows the DM technique's ability in

providing the alternative solution for decision makers in solving problem arise in particular areas. The exploration data in educational field using DM techniques are called as Educational Data Mining (EDM). EDM is concerned with extracting a pattern to discover hidden information from educational data.

Nowadays, the Institutions of Higher Learning (IHL) database contains so much information about their students. The information is kept increasing by times, but there is no action taken to gain knowledge from it. DM is the suitable techniques in managing the IHL data to discover new information and knowledge about students. DM consists of machine learning, statistical and visualization techniques to discover and extract knowledge in such a way that humans can easily interpret [1].

DM provides various methods for analysis process which include classification, clustering, and association rule. Classification, which is one of the prediction types classifies data (constructs a pattern) based on the training set and uses the pattern to classify a new data (testing set). Clustering is the process of grouping records in classes that are similar, and dissimilar to records in other classes. In relationship mining, the goal is to discover the relationship exist between parameters [2, 3, 4].

In this study, the classification method is selected to be applied on the students' data. This research aims to do a comparative analysis among the three selected classification algorithms; Decision Tree (DT), Naïve Bayes (NB), and Rule Based (RB). The comparative analysis is done to discover the best techniques to develop a predictive model for SAP. The patterns obtained will use to predict the first semester of the first year in two Bachelor of Computer Science (BCS) courses; Bachelor of Computer Science with specialization in Software Development (BCSSD) and Science with specialization in Network Security (BCSNS) at the Faculty of Informatics and Computing (FIC), Universiti Sultan Zainal Abidin (UniSZA), Terengganu, Malaysia. This pattern will be used to improve the SAP and to overcome the issues of low grades obtained by students.

There are several studies conducted using students' data comes from IHL Malaysia and these study become the main guideline of this research [5, 6, 7, 8, 9]. All of these studies conducted to find the relationship between independent parameters and dependent parameter selected in their studies. Mostly, the Cumulative Grade Point Average (CGPA), Grade Point Average (GPA), students' grade, and students' mark are normally used as a predictive parameter (dependent parameter) to measure the Students' Academic Performance (SAP) in particular courses or subjects.

In this study, the students' GPA is selected as a dependent parameter. The GPA values of the first semester of the first year BCS students are categories into three different classes; *poor, average*, and *good*. The other parameters used are race,

gender, family income, university entry mode, and Malaysia Certificate of Education (SPM) grades in three subjects; Malay Language, English, and Mathematics. The WEKA tool used to conduct the experiment process. WEKA is an open source machine learning software written in Java that's widely used by many researchers in various fields of studies [10].

This paper is organized into several sections. The background and related works section briefly describe the previous works on SAP and classification techniques. Followed by, the current problem section and a proposed framework for predicting SAP section. Next, the results of the three prediction algorithms are compared in result and discussion section. Finally, the conclusion, limitation, and future work for this study were discussed in the conclusion section.

2. Background and Related Work

IHL faces a major challenge in order to improve and manage the organization to be more efficient in managing students' activities. To achieve this target, DM is considered as the one of most suitable technique in giving additional insights to the IHL community to help them make better decisions in educational activities [11]. There are various previous studies conducted to predict the SAP by using DM techniques. The next subsections will present the other author is works and selected classification techniques applied in this study. A more detailed explanation about SAP and classification method will discuss in the next subsection.

2.1 Students' Academic Performance (SAP)

The SAP prediction on will allow IHL to study what features of a model are important for prediction and to get the hidden information in students' data [2]. There are a lot of researches conducted to develop an SAP prediction model for particular courses or subjects. These studies used various types of students' data with a variety of parameters to identify and classify their students [12, 13, 14].

The SAP prediction of Introductory Engineering Course is done to understand and identify the students' level of performance. For example, if the result of the prediction shows there are some students that will perform poorly in the course, so the lecturers can take appropriate action to help those students. The additional exercise, assignment, or lesson given by lecturers may help the students to improve their understanding in subject taken [12].

The study is also conducted in Malaysia using students' data taken from University Malaysia Pahang (UMP) database management system. The 1000 of student records with three courses in the Faculty of Computer System and Software Engineering, UMP contained students' personal, academic, and course information. The students' grade is selected as a predictor parameter and was divided

into five categories which are *excellent, very good, good, average,* and *poor*. The result indicated that the proposed model is suitable to be used as an SAP prediction [13].

The students' information such as exam scores, grades of team work, attendance, and practical exams are used for profiling and grouping the SAP using selected DM algorithms. The output from analysis process will help the institution to predict academic trends and patterns by categorizing the students into *good, satisfactory,* or *poor* group. It allows the lecturers to get a better understanding about students' learning styles and behaviors [15].

The study involving first year students of school engineering at the National Autonomous University of Mexico (UNAM) is conducted using students' socio-demographic and previous academic information. The data were divided into three categories; students who passed none or up to two courses (low group), students who passed three or four courses (middle group), and students who passed all five courses (high group). The extract patterns from the experiment will allow the IHL to predict academic performance of the new students so that the lecturers will know the level of the new students' preparedness at admission [16].

2.2 Classification Techniques

DM is the process of extracting useful information and knowledge from large data stores or sets. It involves the use of data analysis tools to discover previously unknown patterns and relationships in large data sets. DM not only has the abilities in collecting and managing data, but also has the capability to conduct the analysis and predicting tasks.

Many studies have applied DM methods to predict SAP using popular methods such as classification, clustering, and association rule [11]. The primary goal of using DM techniques in educational field is to develop a prediction model for the students' overall performance in selected courses. The students' performance in prior courses is used as predictor parameter. The extracted model will assist the lecturers to identify the students' problems in order to enhance the students' level of performance in academic [17].

DT is one of the most popular techniques in EDM because it provides an intuitive and human friendly explanation for decision makers to make further action [18]. This technique was applied to the students' data in previous researches to classify students into successful and unsuccessful students' category. So, the lecturers can provide extra learning lesson to the students who are less potential to be successful [7, 8, 11, 13, 19].

NB uses the Bayes' probability theory which assumes the effect of parameter value of a given class is independent of the values of the other parameters. It rep-

resents a predictive approach to make predictions on values of data using know results found from different data [20]. Also, the output from the prediction model using NB can be easily interpreted into the understandable human language [16, 19, 20, 21]. The generated predictive model will help the faculty staff in managing the students' dropout and to predict the SAP of new intake students [22].

RB is a technique for classifying records using a collection of "IF…THEN…" rules. IF-THEN rules will represent the extract knowledge from a dataset in a form that is easy to understand. This gives the chance to the researchers or the domain experts to analyze and validate that knowledge, and combine it with the existing information [23]. The researchers have discovered that a set of IF-THEN classification rules produces has a high level knowledge representation and can be used directly for decision making [10, 24].

From the previous studies, the three classification techniques were chosen for this study are DT, NB, and RB.

3. Current Problem

The IHL goal is to provide the finest quality of education to their students. To achieve that goal, the new discovery about students' learning behaviours and the factors contribute to the students' success should be exposed for the community benefits. To discover hidden information and knowledge from the students' data, a few elements such as parameters, methods, and tools need to be identified and considered in order to produce the best model prediction of SAP. The prediction on SAP can be used as a guideline for the faculty management and lecturers to prevent students from dropout [22]. The objective of this study is to get the patterns of SAP focusing on the first semester of the first year BCS at the FIC, UniSZA, Malaysia.

At the beginning of the semester for new students, a lecturer faces difficulty to know and analyse the student's performance because there is lacking of information about students' previous background. All the information about students is stored in a database at Academic Department, UniSZA and Student Entry Management Department (SEMD), Ministry of Higher Education based in different location (Kuala Lumpur, Malaysia). The parameters used are GPA, race, gender, family income, university entry mode, and SPM grades in three subjects; Malay Language, English, and Mathematics. The study is made to determine whether or not the selected parameters contribute to the SAP.

Besides, this study is also conducted to find out the relationship between the independent parameters and the dependent parameter. The discovered pattern can be used by lecturers to make a prediction on SAP among first semester of the first year bachelor students at FIC. The SAP prediction is very important to provide more information about the students to the lecturers. Therefore, the lecturers would

know how the first year students' will perform in their academic from the earliest moment. The extracted model can work as a helping tool for lecturers to plan the teaching materials in order to improve students' performance, and to decrease the failure rate in computer science course.

4. Proposed Framework for Predicting SAP

This section will present the proposed framework in producing a model prediction by using selected classification techniques. The framework shows the steps involved in developing models to predict SAP of the first semester BCS students at the FIC, UniSZA. Fig. 1 illustrates the three main stages involved in this study; Data Collection and Integration, Data Transformation and Patterns Extraction. The detail explanation about all stages will be described in the next subsections.

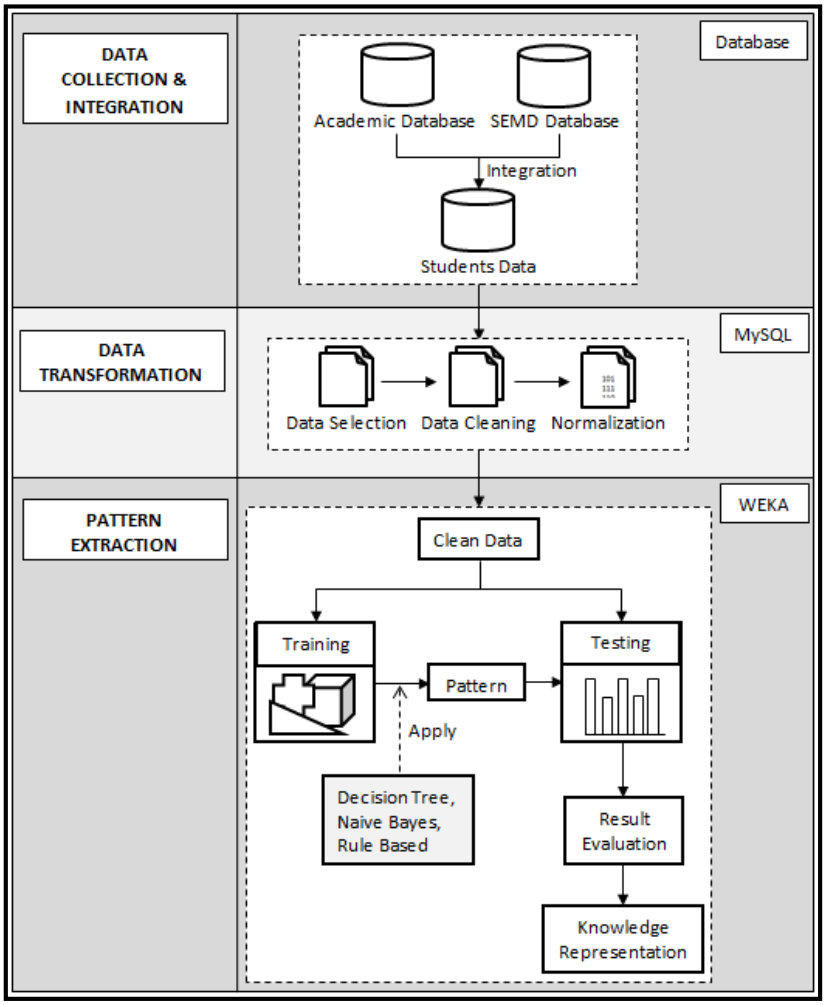

Figure 1: Framework of SAP prediction.

4.1 Data Collection and Integration

The 497 data of BCSSD students FIC, UniSZA from July 2006/2007 intakes until July 2013/2014 intakes were collected. The data were collected from two different data sources that contain all information about UniSZA students. Firstly, the data was collected from the database of Academic Department, UniSZA that stored in Informix Database Management System (DBMS).

4.2 Data Transformation

The data transformation stage was performed to improve the quality of input data to produce the better and quality results. In this stage, all transformation process was also handled using MySQL/PHP statement and programming codes. This stage consists of three phases which are data selection, data cleaning, and data normalization.
In data selection phase, only nine parameters were selected for the mining process. Those parameters were selected based on the literature review done on previous work. The data selection is conducted using MySQL/PHP statements and coding. In this paper, those processes are interpreted in Relational Database Schema to better understand the flow of the process. The steps involved are:

- Step 1:

Select the BCSSD program students' data from the FIC-STUDENT_PROFILE (fic-sp) table that contains all the information of FIC students. In FIC-LOOKUP_PROGRAM table, the BCSSD program is coded as C10. So, all the data with C10 code is selected in the FIC-STUDENTS_PROFILE table and extracted data were saved in the new created table of FIC-STUDENTS_PROFILE_C10.

- Step 2:

In this step, the students' matric number (ID) from

FIC-STUDENTS_PROFILE_C10 (c10) table were matched with the students' ID from FIC-GPA&CGPA (gpa) table to obtain the students' GPA of first semester and the new selected data were saved in the new table named FIC-C10_SEM1.

- Step 3:

The students' addresses in table FIC-C10_SEM1 (c10_1) were matched with TOWN_LOCATION (tl) table to categories the students' hometown location whether is a town or rural. After that, the new added data were saved in the new table named C10_SEM1_TOWN.

- Step 4:

By using students' ID as a matching key, all the parameters in SEMD table were combined with all parameters in C10_SEM1_TOWN (c10-town) table and saved in the new table named C10_SEM1_TOWN2.

- Step 5:

Finally, nine parameters were selected to be mined. The parameters are gender, race, hometown (ht), GPA, family income (fi), university mode entry (ume), and SPM grades in three subjects; Malay Language (bm), English (bi), and Mathematics (math).

Many parameters have been used to identify the factors that influence the students' achievement in academic. So that, in this study, the parameter that contributed to the students' success will be identified. Next, the data cleaning process will remove the missing or incomplete data.

After the cleaning process, only 399 from 497 data can be used for mining since 98 data were removed due to missing values of several parameters. The next phase is a data normalization process where the numerical values such as GPA parameter were transformed into nominal or categorical class. The GPA was grouped into three classes; *good, average*, and *poor*. The SPM grades were categories into nine groups (0-8) based on the previous (2000-2008) and the current (starting from 2009)

4.3 Pattern Extraction

In this stage, the WEKA open source tool is used to conduct the experiment. This stage consists of five phases; training, pattern, testing, result evaluation and knowledge representation. In this stage, the cleaned data were divided into two parts; training set and testing set. The training set used to build the model or pattern from the classification techniques and testing set used to validate the models. After that, the result obtained will be evaluated and represented as knowledge. The three selected classification algorithms for this process are; DT (J48), NB, and RB (PART).

5. Results and Discussions

In this section, the experiment result from the DM process is represented. The accuracy value obtained shows how good the extraction model can predict a new data. Two types of data splitting used were percentages and fold cross validation. For the percentages 10:90, the training data set is 10%, while the testing data set is

90% of the total data. In fold cross validation, the data were divided into 3, 5, or 10 subset. Based on the experimentation, RB shows the highest accuracy value of 71.3% in 80:20 percentages test option compared the other techniques. NB shows the best accuracy value of 67.0% in 80:20 percentages test option. While the DT displays the best accuracy value of 68.8% in 80:20 of percentage test option. From the three selected algorithms for the experiment, the model prediction extracted by RB displays the highest accuracy value. The confusion matrix table is then constructed which contains the information about actual and predicted classifications. The matrix shows the prediction is successful for the *good*, *average* and *poor* categories.

From the investigation, we found out the classification algorithms can achieve the highest prediction accuracy under the following circumstances:
1. Involved a lot of data in the mining process, sometimes thousands.
2. The dataset prepared for analyzing contains just a few noisy or incomplete data.

6. Conclusion and Future Work

The amount of data stored in an educational database at IHL is increasing rapidly by the times. In order to get the knowledge about student from such large data and to discover the parameter that contributed to the students' success, the classification techniques are applied to the students' data. This study also conducts a comparative analysis of three classification techniques; DT, NB, and RB using WEKA tool. The experimental result shows that the RB has the best classification accuracy compared to NB and DT. The model will allow the lecturers to take early actions to help and assist the *poor* and *average* category students to improve their results. The limitation of this study is the small size of data due to incomplete and missing value in the collected data. In the future, this study will be expanded by adding more data from different years or different institutions in order to increase the accuracy of the prediction.

Acknowledgments. The authors would like to thank Universiti Sultan Zainal Abidin (UniSZA) especially the Information Technology Center, and the Academic Department for providing the data and other resources for this research.

References

[1] Y. Zhang, S. Oussena, T. Clark, & H. Kim, Use Data Mining to Improve Student Retention in Higher Education – A Case Study. ICEIS - *12th International Conference on Enterprise Information Systems 2010,* (2010). http://dx.doi.org/10.5220/0002894101900197

[2] R. B. Sachin, & M. S. Vijay, A Survey and Future Vision of Data Mining in Educational Field, *2012 Second International Conference on Advanced Comput-*

ing & Communication Technologies, (2012), 96–100. http://dx.doi.org/10.1109/acct.2012.14

[3] A. A. Aziz, N. H. Ismail, & F. Ahmad, Mining Students' Academic Performance, *Journal of Theoretical and Applied Information Technology*, **53** (2013), no. 3, 485–495.

[4] R. S. J. Baker, Data Mining for Education, Advantages Relative to Traditional Educational Research Paradigms, (2010).

[5] C.-T. Lye, L.-N. Ng, M. D. Hassan, W.-W. Goh, C.-Y. Law, & N. Ismail, Predicting Pre-university Student's Mathematics Achievement, *Procedia - Social and Behavioral Sciences*, **8** (2010), 299–306. http://dx.doi.org/10.1016/j.sbspro.2010.12.041

[6] M. F. M. Mohsin, M. H. A. Wahab, M. F. Zaiyadi, & C. F. Hibadullah, An Investigation into Influence Factor of Student Programming Grade Using Association Rule Mining, *International Journal on Advances in Information Sciences and Service Sciences*, **2** (2010), no. 2, 19–27. http://dx.doi.org/10.4156/aiss.vol2.issue2.3

[7] M. Wook, Y. H. Yahaya, N. Wahab, M. R. M. Isa, N. F. Awang, & H. Y. Seong, Predicting NDUM Student's Academic Performance Using Data Mining Techniques, *2009 Second International Conference on Computer and Electrical Engineering*, (2009), 357–361. http://dx.doi.org/10.1109/iccee.2009.168

[8] N. M. Norwawi, S. F. Abdusalam, C. F. Hibadullah, & B. M. Shuaibu, Classification of Student's Performance in Computer Programming Course According to Learning Style, *2009 2nd Conference on Data Mining and Optimization*, (2009), 37–41. http://dx.doi.org/10.1109/dmo.2009.5341912

[9] H. Othman, Z. M. Nopiah, I. Asshaari, N. Razali, M. H. Osman, & N. Ramli, (2009), A Comparative Study of Engineering Students on Their Pre-University Results with Their First Year Performance at Fkab, UKM. Seminar Pendidikan Kejuruteraan dan Alam Bina (PeKA'09).

[10] Kabakchieva, D., Predicting Student Performance by Using Data Mining Methods for Classification, *Cybernetics and Information Technologies*, **13** (2013), no. 1, 61–72. http://dx.doi.org/10.2478/cait-2013-0006

[11] S. Prakash, K. S. Ramaswami, & C. A. Post, Fuzzy K- Means Cluster Validation for Institutional Quality Assessment, *Communication and Computational Intelligence (INCOCCI), 2010 International Conference*, (2010), 628–635.

[12] S. Huang, & N. Fang, Work in Progress - Prediction of Students' Academic Performance in an Introductory Engineering Course, *In 41st ASEE/IEEE Frontiers in Education Conference*, (2011), 11–13. http://dx.doi.org/10.1109/fie.2011.6142729

[13] S. Sembiring, M. Zarlis, D. Hartama, & E. Wani, Prediction of student academic performance by an application of data mining techniques, *2011 International Conference on Management and Artificial Intelligence*, **6** (2011). 110–114.

[14] P. Golding, L. Facey-Shaw, & V. Tennant, Effects of Peer Tutoring, Attitude and Personality on Academic Performance of First Year Introductory Programming Students, *36th ASEE/IEEE Frontiers in Education Conference*, (2006), 7–12. http://dx.doi.org/10.1109/fie.2006.322662

[15] S. Parack, Z. Zahid, & F. Merchant, Application of data mining in educational databases for predicting academic trends and patterns, *2012 IEEE International Conference on Technology Enhanced Education (ICTEE)*, (2012), 1–4. http://dx.doi.org/10.1109/ictee.2012.6208617

[16] E. P. I. García, & P. M. Mora, Model Prediction of Academic Performance for First Year Students, *2011 10th Mexican International Conference on Artificial Intelligence*, (2011), 169–174. http://dx.doi.org/10.1109/micai.2011.28

[17] A. T. Chamillard, Using student performance predictions in a computer science curriculum, *ACM SIGCSE Bulletin*, **38** (2006), no. 3, 260. http://dx.doi.org/10.1145/1140123.1140194

[18] D. T. Larose, *Discovering Knowledge in Data: An Introduction to Data Mining*, Hoboken, NJ, USA: Wiley, 2005. http://dx.doi.org/10.1002/0471687545

[19] J. Shana, & T. Venkatachalam, Identifying Key Performance Indicators and Predicting the Result from Student Data, *International Journal of Computer Applications*, **25** (2011), no. 9, 45–48. http://dx.doi.org/10.5120/3057-4169

[20] U. Kumar, & P. S. Pal, Data Mining: A prediction of performer or underperformer using classification, International Journal of Computer Science and Information Technologies (IJCSIT), **2** (2011), no. 2, 686–690.

[21] M. Sharma, Development of Predictive Model in Education System: Using Naïve Bayes Classifier, *International Conference and Workshop on Emerging Trends in Technology (ICWET 2011) – TCET*, Mumbai, India, (Icwet), (2011), 185–186. http://dx.doi.org/10.1145/1980022.1980064

[22] S. Pal, Mining Educational Data Using Classification to Decrease Dropout

Rate of Students, *International Journal of Multidisciplinary Sciences and Engineering*, **3** (2012), no. 5, 35–39.

[23] E. Frank, & I. H. Witten, Generating Accurate Rule Sets without Global Optimization, In: Proc. Of The 15th Int. Conference on Machine Learning.

[24] C. Romero, S. Ventura, P. G. Espejo, & C. Hervás, (2008). Data Mining Algorithms to Classify Students, *in: The 1st International Conference on Educational Data Mining Montréal*, Québec, Canada, (1998), 8–17.

Received: April 15, 2015; Published: November 2, 2015