RESEARCH ARTICLE

STUDENT ACADEMIC PERFORMANCE USING DATA MINING TECHNIQUES

L.Pandeeswari, M.Phil. (Computer Science)   **K.Rajeswari,** M.Sc., M.Phil.

Research Scholar, Assistant Professor in Computer Science

Vivekanandha College for Women, Unjanai, Tiruchengode, India

pandieswariap22@gmail.com

Abstract: Data and Information or Knowledge has a significant role on human activities. Data mining is the knowledge discovery process by analyzing the large volumes of data from various perspectives and summarizing it into useful information. Assessment as a dynamic process produces data that reasonable conclusions are derived by stakeholders for decision making that expectedly impact on students' learning outcomes. The data mining methodology while extracting useful, valid patterns from higher education database environment contribute to proactively ensuring students maximize their academic output. This paper develops a methodology by the derivation of performance prediction indicators to deploying a simple student performance assessment and monitoring system within a teaching and learning environment by mainly focusing on performance monitoring of students' continuous assessment (tests) and examination scores in order to predict their final achievement status upon graduation. Based on various data mining techniques (DMT) and the application of machine learning processes.

Keyword: data mining, methodology, data mining techniques, machine learning process

1. Introduction

The highly inter-disciplinary field of Educational Data Mining (EDM) has resulted from a fusion of many different areas, some which include Machine Learning, Cognitive Science and Psychometrics? The main task in EDM is to construct computational models and tools to mine data that originated in an educational setting. With rapidly increasing data repositories from different educational contexts (paper tests, e-learning, Intelligent tutoring system etc) good practices in EDM can potentially answer important research questions about student learning. Recently educational institutions targets activities within its organizations with ERP tools to handle and store huge data available in educational process for hidden patterns. The objective of this study are (i) prediction of first year engineering student's performance (ii) find out association between the different factors influencing grades and (iii) compare different prediction algorithms for classifying students. This study is more useful for identifying weak students and the identified students can be individually assisted by the educators so that their performance is better in future.

2. Existing Issue

The aim of HSC is to provide quality education to their students. The quality of education in HSC can be increased by discovering new knowledge or patterns for prediction of students' performances especially in the academic aspect. The prediction on SAP can be used as a guideline for the faculty management or educators to prevent students from dropout The objective of this study is to get the patterns of SAP focusing on the first semester of the first year Bachelor of Computer Science with specialization in Software Development at the FIT, At the beginning of the semester for new students, educators face difficulty to know and analyses the student's performance because there are lack of information about their students' previous background. All the information about students is stored in a database at Academic Department, Student Entry Management Department (SEMD), Ministry of Higher Education based in different location (Kuala Lumpur, Malaysia). The selected parameters from Academic Department, database are gender, hometown, race, and GPA. The parameters on students' university entry mode and family income are taken from different database which is located in SEMD database. 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 (GPA). The discovered pattern can be used by educators the first year bachelor students

at FIT. The development of effective systems to predict SAP is very important to provide more information about the students to the educators. Therefore, the educators would know how the first year students' will perform in their academic from the earliest moment. A prototype system will be developed using the discovered patterns that are extracted from the DM process. In additional, the system can work as a helping tool for educators to plan the teaching materials in order to improve students' performance, and to decrease the failure rate in computer science course. Artificial Neural Network (ANN) is based on the human brain architecture that consists of multiple processing layers connected with nodes. ANN method is used in the educational field for predicting and classifying SAP.

3. Proposed Framework for Predicting SAP

This section will present the new proposed framework in predicting SAP by using EDM technique. The proposed framework is based on DM model development phases. This model is selected for two reasons: 1) Using the same steps in extracting patterns from databases. 2) The parameters are collected from two different locations. The newly proposed framework takes into consideration for different parameters, DM method, and tool. The framework consists of four stages, which are data collection, data transformation, pattern extraction, and prototype development.

The detail explanation about all stages will be described in the next subsections.
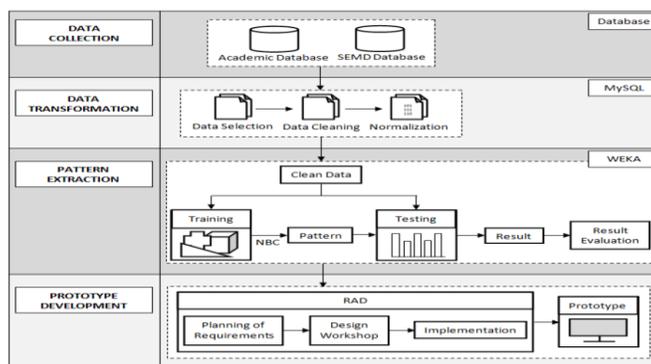

Fig. SAP PROCESSING

4. METHODOLOGY

4.1. Data Collection:

For this study real world data's are collected from first year engineering students. A sample of 464 students was taken from a group of colleges. Students were grouped in a classroom they were briefed clearly about the questionnaire and it took on average half an hour to fill this questionnaire. Selection of students was at random. The primary data was collected using a questionnaire. Which include questions (i.e. with predefined options) related to several personal, socio-economic, psychological and school and college related variables that were expected to affect student performance. The questionnaire was reviewed by professionals and tested on a small set of 50 students in order to get a feedback. The final version contained 23 questions in a single A4 sheet and it was answered by more than 700 students. Latter we selected a sample of 464 from the whole. All questionnaires were filled with the response rate of 100% out of which 316 were females and 184 were males. The secondary data such as semester mark details, attendance percentage, and class test performance were collected from the college and from the directed website.

Variable Name	Description	Domain
SEX	student's sex	{male, female}
COMM	student's community	{OC, BC, MBC, SC, ST}
REL	student's religion	{ hindu, christian, muslim }
SA	student's living area	{urban, rural}
SD	student's department	{ mechanical, computer science, EEE, ECE, CIVIL, IT }
PQ	parent's qualification	{ illiterate, schooling, degree/diploma }
F-OCC	father occupation	{daily wages, farmer, weaver, ex serviceman, government, business, private}
M-OCC	mother occupation	{house wife, daily wages, farmer, weaver, private, government }
MOS	student's schooling medium of study	{ tamil, english }
SSLC-TOTAL	student's 10th total	{ <250, 251 – 350, 351 – 450, >451 }
SSLC-PERCEN	student's 10th percentage	{O – 90% - 100%, A – 80% - 89%, B – 70% - 79%, C – 60% - 69%, D – 50% - 59%, E – 40% - 49%, F - < 40%}
HSC-TOTAL	student's 12th total	{ <250, 251 – 350, 351 – 450, >451 }
HSC-PERCEN	student's 12th percentage	{O – 90% - 100%, A – 80% - 89%, B – 70% - 79%, C – 60% - 69%, D – 50% - 59%, E – 40% - 49%, F - < 40%}
HSC-CUT	student's 12th cut off	{ <50, 51-100, 101 -150, 151 – 200 }
SQ	student's quota	{ management , counselling }
SSP	student's staying place	{ day scholar, hostel }
SATT	student's college attendance percentage	{<50, 50, 60, 70, 80, 90}
G-OBT	grade obtained	{O – 90% - 100%, A – 80% - 89%, B – 70% - 79%, C – 60% - 69%, D – 50% - 59%, E – 40% - 49%, F - < 40%}

4.2. Algorithms:

Although many classification models exist, only some have been selected within the scope of this study. The selected algorithms are Naive Bayesian algorithm; MLP, SMO, J48, REP tree, RANDOM tree and Decision table are used. The Naïve Bayesian model defines the classification problem with respect to probabilistic idioms, and supplies statistical methods to classify the instances based on probabilities. Multilayer preceptor is a type of artificial neural network algorithm which regards the human brain as the modeling tool. It provides a generic model for learning real, discrete and vector target values. The ability to understand the hidden model is hard and training times may be long. In decision tree algorithms, the classification process is summarized by a tree. After the model is built, it is applied to the database

4.3. Implementation:

First, data cleaning was applied on the datasets. According to the missing data analysis, missing data have been removed from the datasets. Other than missing data analysis, datasets were also cleaned to remove noisy data. Unnecessary space characters or other spelling mistakes were also cleaned in the datasets. Another usual step in data pre-processing is data discretisation. Although some algorithms are said to perform better when the numerical input variables are discredited, in this study numerical variables have not been put into binned intervals in order to maintain the same conditions for all algorithms. Once the data pre-processing steps have been completed, the dataset have been used to run the classification algorithms Naïve Bayesian algorithm, MLP, SMO, J48, REP tree, RANDOM tree and Decision table. For all algorithms, splitting the data into train and test splits has been selected as the validation method. 66% of the data has been set as the training part and the rest has been set as the testing part.

4.4. Weka Data Mining Software:

WEKA is open source software issued under the GNU General Public License. WEKA has been utilized as the tool to run different classification algorithms. The algorithms can either be applied directly to a dataset or called from your own Java code. WEKA contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes.

730

5. Conclusion

The HSC main objective is to provide the best quality education to its students and to improve the quality of managerial decisions. Prediction of SAP helps educators to get a better understanding or big picture on how the students will perform in their studies. The prediction model acts like a warning system to detect potential weak students, so that the educators can take an appropriate action towards them. For an instant, they can give advices to prevent failure in the examinations or early desertion of studies. The proposed framework for predicting SAP based on the selected parameters and NBC is presented. This study will be preceded with the implementation of the framework in order to test for its applicability and prediction accuracy

6. Future Work

The future scope could be dynamic updating of the database by the user before applying the algorithm to generate the patterns of interest rather than performing it on statistical historical data. Moreover the data mining can be made as constraint-based mining wherein the rules can be generated based on the constraints provided by the user.

REFERENCES

[1] Chandra, E. and Nandhini, K. (2010) 'Knowledge Mining from Student Data', European Journal of Scientific Research, vol. 47, no. 1, pp. 156-163.

[2] El-Halees, A. (2008) 'Mining Students Data to Analyze Learning Behavior: A Case Study', The 2008 international Arab Conference of Information Technology (ACIT2008) – Conference Proceedings, University of Sfax, Tunisia, Dec 15- 18.

[3] Faouzi Mhamdi, Mourad Elloumi, "A New Survey On knowledge Discovery And Data Mining" December 2007.

[4] Usama Fayyad, "Data Mining and Knowledge Discovery in Databases: Implications for Scientific Databases" Proceedings of the Ninth International Conference on Scientific and Statistical Database Management IEEE Computer Society Washington, DC, USA, 1997.