Jurnal Teknologi

Full Paper

A FRAMEWORK FOR STUDENTS' ACADEMIC PERFORMANCE ANALYSIS USING NAÏVE BAYES CLASSIFIER

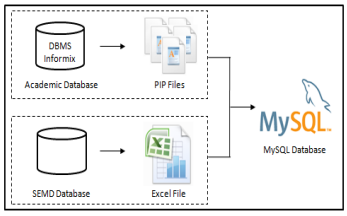
Azwa Abdul Aziz*, Nur Hafieza Ismail, Fadhilah Ahmad, Hasni Hassan

Faculty of Informatics and Computing Sultan Zainal Abidin, University Terengganu, Malaysia

*Corresponding author azwaaziz@unisza.edu.my

Article history
Received
3 December 2013
Received in revised form
2 July 2014
Accepted
25 November 2014

Graphical abstract



Abstract

Educational database of Higher Learning Institutions holds an enormous amount of data that increases every semester. Data mining technique is usually applied to this database to discover underlying information about the students. This paper proposed a framework to predict the performance of first year bachelor students in Computer Science course. Naïve Bayes Classifier was used to extract patterns using WEKA as a Data mining tool in order to build a prediction model. The data were collected from 6 year period intakes from July 2006/2007 until July 2011/2012. From the students' data, six parameters were selected that are race, gender, family income, university entry mode, and Grade Point Average. By using Naïve Bayes Classifier, it would predict the class label "Grade Point Average" as a categorical value; Poor, Average, and Good. Result from the study shows that the students' family income, gender, and hometown parameter contribute towards students' academic performance. The prediction model is useful to the lecturers and management of the faculty in identifying students with weak performance so that they will be able to take necessary actions to improve the students' academic performance.

Keywords: Higher learning institution, data mining, educational data mining, classification, Naïve Bayes Classifier, prediction, students' academic performance

1.0 INTRODUCTION

Data Mining (DM) techniques involve a large amount of data with various parameters to be explored. The advantage of this technique is its capability to discover underlying relationships and patterns that exist within the data. It combines machine learning, statistical and visualization techniques to discover and extract knowledge in such a way that humans can easily interpret [1]. Educational Data Mining (EDM) is the process of transforming raw data taken from students' information system at Higher Learning Institutions (HLI) into useful information or pattern [2]. The pattern is extracted using DM techniques which can subsequently be used as a model in predicting Students' Academic Performance (SAP).

The aim of HLI is to provide a better quality education to their students. The quality of education in HLI can be increased by discovering useful information or patterns from students' dataset to predict students' performance especially in academic aspects. A prediction model is effective to identify students with the possibility of failures in their studies. The model could serve the purpose as an alarm in detecting the problems so that the faculty management and lecturers can take an appropriate action to improve SAP in the future [3].

In HLI, SAP is measured using the Grade Point Average (GPA). GPA shows the overall SAP where it considers the average of all subject examinations' grade taken in that semester [4]. So, it shows that the GPA is a best dependent parameter in evaluating the SAP. In this study, the GPA is used as a dependent parameter by grouping the GPA values into three categorical segments which are Poor, Average, and Good.

The main objective of this paper is to use DM techniques 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 Faculty of Informatics and Computing (FIC), Sultan Zainal Abidin University (UniSZA), Terengganu, Malaysia. This research applies Naïve Bayes Classifier (NBC) on clean students' dataset to extract a hidden pattern on selected parameters that are race, gender, family income, university entry mode, and GPA.

2.0 BACKGROUND AND RELATED WORK

Nowadays, one of the major challenges in HLI is to improve and manage the educational processes to be more efficient, effective and accurate. To achieve this target, EDM is considered as the one of most suitable technique in giving additional insights to the HLI staffs such as the lecturers, students and faculty management to help them make better decisions on their educational activities especially the prediction on SAP [5].

This study was conducted to analyze the students' data to get the better understanding about SAP by employing NBC as a DM technique. A more detailed explanation about SAP and NBC as a two main components in this study will discuss in the next subsection.

2.1 Students' Academic Performance (SAP)

There are increasing research interests in extracting a pattern on predicting SAP in HLI for various educational purposes. Prediction on SAP will allow HLI to study what features of a model are important for prediction and to get the hidden information about underlying construct [6].

To understand and identify the most effective factors that influence the SAP, Pandey and Pal [7] had conducted studies on 600 students' data to find out hidden information using NBC to predict the SAP. Selected parameters used in their study were gender, language medium, program, and division obtained (dependent parameter). The prediction model generated after the DM process will help the institution to reduce the dropout and improve the HLI level of performance.

Mohammed and Alaa [8] applied several of EDM techniques and methods to improve SAP and overcome the problem of students' with poor grades. The EDM tasks such as association rules, classification, clustering, and outlier detection were applied to their students' data set. Six parameters from the 18 parameters that are gender, specialty, hometown, matriculation GPA, secondary school type, and grade obtained (dependent parameter) were the main parameters analyzed and examined using a DM process.

Suhem, Zaim and Fatima [9] used Apriori algorithm and K-means clustering in the educational databases to profile and group the SAP based on various parameters; exam scores, grades of team work, attendance, and practical exams. This study helped the educational institution to predict academic trends and patterns by categorizing the students into good, satisfactory, or poor group. It allows the lecturers to identify hidden patterns about students' learning styles and behaviors.

In another study involving first year students of school engineering at the National Autonomous University of Mexico (UNAM), model predictions of academic performance were extracted using NBC [10]. The students' socio-demographic and academic information were among the several parameters taken from the records of first semester students. 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). After the evaluating the results, they obtained a model of nearly 60% accuracy.

2.2 Naïve Bayes Classifier (NBC)

There are various DM techniques such as association rules, classifications, and clustering can be efficiently used for educational data to extract hidden patterns to better understand students' study behaviors. NBC is a famous technique in a DM classification method that is frequently used in a variety of experiments to predict the SAP at the HLI [2, 7, 10, 11, and 12].

Many researchers used NBC to develop a model for prediction because it shows the excellent ability during the DM process. For example, Brijesh & Pal conducted a study to develop a predictive NBC model for SAP so as to identify the difference between high and slow learners student [11]. 300 students' data were used in this study to get the pattern and a model prediction was built based on it. The advantage of NBC is ease of use because only required one time scan on the training data. It also only requires a small number of training data to do estimation for each parameter.

Sharma & Mavani [12] used NBC for predicting students' academic results. The prediction model was developed based on a record of 120 students. Out of 120 total records, 60% (72 records) was used as a training data. The remaining 48 record was used as a testing data and the success of predictive model was 70% where 70% of the predicted results matched the actual results of the students. Compared to the other classification techniques, NBC has the least error rates during prediction process. Also, the predicted result using NBC can be easily interpreted into understandable language.

From the literatures, NBC shows a high accuracy of prediction in critical fields such as medical, fraud detection etc. [13, 14] become the motivation to use NBC in this study. The proposed framework for predicting SAP at UniSZA is presented in the next section.

3.0 PROPOSED FRAMEWORK FOR PREDICTING SAP

This section presents the proposed framework in predicting SAP using EDM technique. It shows the EDM process in developing a predictive model to predict SAP in Bachelor of Computer Science with specialization in Software Development at the FIC, UniSZA [15]. Figure 1 illustrates the data collection, data transformation and patterns extraction stages of this study.

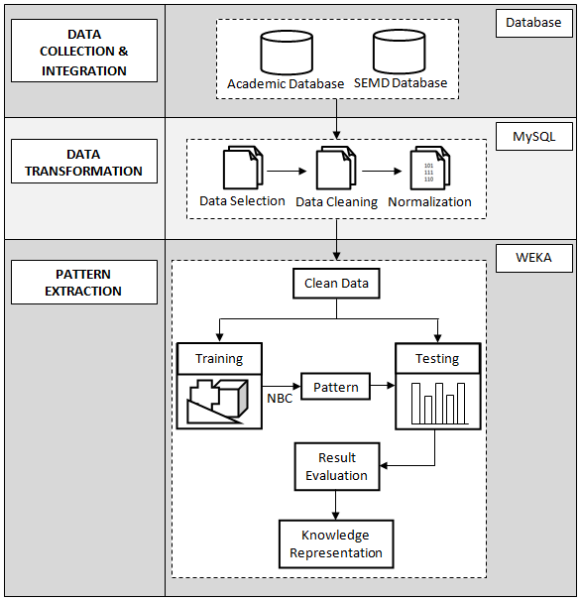


Figure 1 Framework of SAP prediction

3.1 Data Collection and Integration

Data of 245 bachelor students of FIC, UniSZA from July 2006/2007 intakes until July 2011/2012 intakes were collected for this study. Figure 2 shows GPA of the first semester as a dependent parameter in predicting SAP. The three colours (, ,) in the figure below represent the GPA category (Average, Good, Poor).

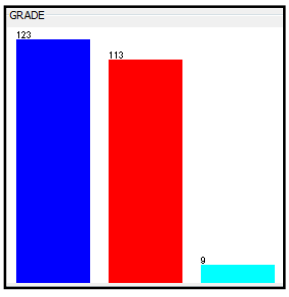


Figure 2 Distribution of GPA of 1st semester

Figure 3 shows the flow of the data collection process with a detail view. At this stage, the data were collected from two different data sources where the data contains information about the Bachelor of Computer Science students, FIC, UniSZA. Firstly, the data was collected from the Academic Department of the UniSZA database that is stored in Informix Database Management System (DBMS). Five separate WordPad files were extracted as follows:

FIC-LOOKUP_PROGRAM.pip
This file contains information about all programs offered by FIC (program's name, code, shortcut name, and total credit hours for each program).
FIC-LOOKUP_SUBJECT_CODE.pip
All subjects' code for each FIC program is listed here.

FIC-GPA&CGPA.pip
This file holds the information about students' GPA, Cumulative Grade Point Average (CGPA), session, and total credit hours for each semester.

FIC-MARK.pip
This file contains the information on students' grade and marks obtained for each subject taken during their studies.

FIC-STUDENT_PROFILE.pip
This file contains the students' demographic information such as name, matrix number, identification number, address, birth date, session entry dates, etc.

Secondly, the additional parameters are taken from the Student Entry Management Department (SEMD), Ministry of Higher Education database or known as a University Center Unit (UCU) based in different location (Kuala Lumpur, Malaysia). The data retrieved from this database were saved in an excel file. The data contains the information about students' family income, university entry mode, Malaysian Certificate of Education (SPM) grades in several subjects; Malay Language, English, Mathematics, Chemistry, Physic, Science, and Biology.

Next, all the students' information taken from two different data sources was combined in a single database called "students' data" using MySQL/PHP programming.

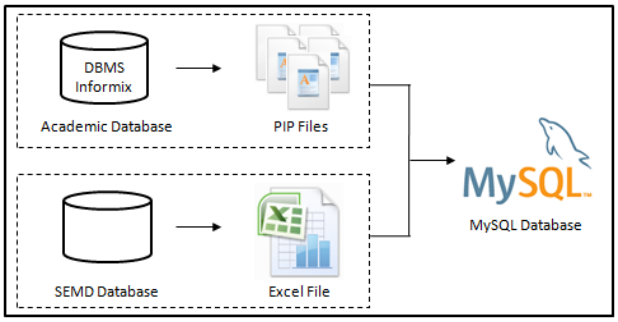


Figure 3 Flow of data collection extraction process

The Relational Database Schema interpreted from MySQL/PHP programming used during the data integration process for this research is shown below. The steps in the process were:

Step 1:

Select the Bachelor of Computer Science program students' data from the raw data that contains all the FIC students' information taken from the Academic Department database. The FIC raw data named is "student_FIC" and Bachelor of Computer Science program code is "c10". The new extracted data were saved in the new table named "c10".

$$\sigma_{code=c10}(\text{student\_FIC}) \quad (1)$$

Step 2:

The students' ID from c10 table were matched with the students' ID from Mark_FIC table to obtain the students' GPA during 1st semester and the new selected data were saved in the new table named "c10_1".

$$c10 \bowtie_{c10.id=mark.id} (\sigma_{sem=1} \text{mark\_FIC}) \quad (2)$$

Step 3:

The students' address in table c10_1 was matched with location table to categorize the students' hometown location into "Town" or "Rural". After that, the selected data were saved in the new table named "c10_sem1".

$$c10\_1 \bowtie_{c10\_1.hometown=(\%.location.class,\%)} \text{town} \quad (3)$$

Step 4:

By using students' ID as a matching key, all the parameters in SEMD table were combined with all parameters in c10_sem1 table and saved in the new table named "c10_sem1_town".

$$c10\_sem1 \bowtie_{c10\_sem1.id=SEMD.id} \text{SEMD} \quad (4)$$

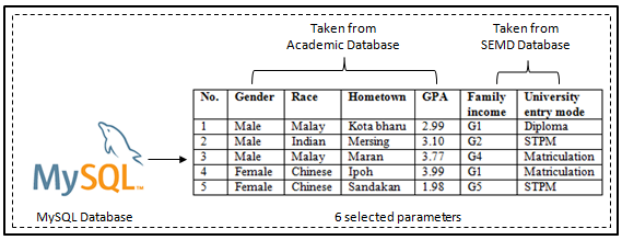
Step 5:

Finally, six parameters were selected to be mined. The parameters are gender, race, hometown, GPA, family income and university mode entry.

$$\prod_{gender,race,hometown,GPA,income,admission} (c10\_sem1\_town) \quad (5)$$

3.2 Data Transformation

The data transformation stage was performed to improve the input data quality for mining. In this stage, all the transformation processes were handled using MySQL/PHP programming. This stage consists of three phases which are data selection, data cleaning, and data normalization. Figure 4 displays the data selection for this study.



No.	Gender	Race	Hometown	GPA	Family income	University entry mode
1	Male	Malay	Kota bharu	2.99	G1	Diploma
2	Male	Indian	Mersing	3.10	G2	STPM
3	Male	Malay	Maran	3.77	G4	Matriculation
4	Female	Chinese	Ipoh	3.99	G1	Matriculation
5	Female	Chinese	Sandakan	1.98	G5	STPM

Figure 4 Data selection

In data selection phase, only six parameters were selected for the mining process. Four parameters taken from the Academic Department, UniSZA database are students' gender, race, hometown, and GPA. The other two parameters were extracted from SEMD database are students' family income and, university entry mode. All this six selected parameters were extracted and saved in one table.

Next, the data cleaning process will remove the missing or incomplete data. Figure 5 shows the example of incomplete data that were removed during the data cleaning process. Based on the figure, data no. 3 didn't have a value of family income, data no. 6 has incomplete race value, and data no. 8 has a missing value in GPA field. So, all of this 3 data are removed to improve the quality of data for better mining results. After the cleaning process, only 176 from 245 data can be used for mining since 69 data were removed due to missing values of several parameters.

No.	Gender	Race	Hometown	Family income	University entry mode	GPA	
1	Male	Malay	Merang	G1	Diploma	2.99	
2	Male	Indian	Kuching	G2	STPM	3.10	
3	Male	Malay	Besut		Matriculation	3.77	← Missing family income value
4	Female	Chinese	Pekan	G1	Matriculation	3.99	
5	Female	Chinese	Serdang	G5	STPM	1.98	
6	Male	xxxx	Labuan	G1	STPM	2.66	← Incomplete race value
7	Male	Indian	Bachok	G2	STPM	3.05	
8	Male	Malay	Bentong	G4	Matriculation		← Missing GPA value
9	Male	Chinese	Chukai	G1	Matriculation	2.78	
10	Female	Chinese	Ajil	G5	STPM	3.98	

Sample of the data

Figure 5 Data cleaning

The next phase is a data normalization process where the numerical values such as GPA parameter were transformed into nominal or categorical class as shown in Table 1. The GPA was grouped into three classes; Good, Average, and Poor. While, the hometown parameter value was transformed into Rural or Town category based on the students' address. The six selected parameters were divided into two types which are independent and dependent parameter. Independent parameter became input of the model used in methods' equation or rules to predict the dependent parameter as an output.

Table 1 Selected parameters from students data

Parameter type	Parameter	Category
Independent parameter	Gender	Male, Female
	Race	Malay, Chinese, Indian, Kedayan, European, Khek (Hakka), Indonesian Malay, Hainan, Kensiu
	Hometown	Town, Rural
	University entry mode	STPM, Diploma, Matriculation
	Family income	G1 - RM1– RM500 G2 - RM501– RM1000 G3 - RM1001 – RM2000 G4 - RM2001 – RM3000 G5 - RM3001 – RM4000 G6 - RM4001 – RM5000 G7 - RM5001 – RM7500 G8 - RM7501 – RM10000 G9 - RM10001 and above
Dependent parameter	GPA in 1st semester	Poor - 0.00 – 1.99 Average - 2.00 – 2.99 Good - 3.00 – 4.00

3.3 Pattern Extraction

In pattern extraction stage, WEKA DM open source tool was used to predict the SAP based on the GPA obtained at the 1st semester. The "arff" file format was created to be used in the NBC and to generate predictive patterns for SAP. This stage consists of five phases that are training, pattern, testing, result evaluation and knowledge representation. In this stage, the cleaned data were divided into two parts; training data and testing data. The NBC was applied to the training data to estimate the prior probability $P(X_i)$ for each class using the GPA category (Poor, Average, and Good) by calculating the occurrence of each class in the training data. All parameters (race, gender, hometown, family income, and university entry mode) C_1, C_2, C_3,…., C_i can be calculated to determine $P(C_i)$. After that, the posterior probability $P(C_i \mid X_i)$ can be estimated by calculating how often each parameter occurs in the each GPA category in the training data.

$$P(C_i \mid X_i) = P(X_i \mid C_i)\, P(C_i) \,/\, P(X_i) \quad (1)$$

To classify the target class, the conditional and prior probabilities generated from the training set were used to make the prediction on training data by estimating $P(Y_i \mid C_i)$ by

$$P(Y_i \mid C_i) = \prod_{k=1}^{n} (x_{ij} \mid C_i)$$
$$= P(\mathbf{x}_1 \mid C_i) * P(\mathbf{x}_2 \mid C_i) * \ldots\ldots * P(\mathbf{x}_n \mid C_i) \quad (2)$$

To calculate $P(Y_i)$, Y_i the likelihood in each class can be estimated since the probability that Y_i is in a class is the product of the conditional probabilities for each parameter value. The class with the highest probability value was chosen for the target class [11] and the output received from the testing data process will detect a pattern. The detected pattern was evaluated by estimating the pattern accuracy and becomes useful knowledge to improve SAP.

4.0 RESULTS AND DISCUSSIONS

In this section, the experiment result from the DM process is represented. During the experiment, the accuracy values were recorded in a table. The accuracy value obtained shows how good the extraction model can predict a new data. Table 2 displays the mining results of SAP using WEKA tool. Two types of data splitting used in the mining process were percentages and fold cross validation. For 10:90, the training data is 10% of all data sets and 90% of all data sets used for testing data. For fold cross validation, the data sets were divided into 3, 5, or 10 subset the holdout method was repeated based on subset's number. For 3 fold cross validation, one of the 3 subset is used as the testing data and others 2 subsets were used to develop the training data. The testing data accuracy is average in 3 trails.

Table 2 The mining result

		NBC Accuracy
Percentages Training : Testing	***10:90***	52.5%
	20:80	53.2%
	30:70	51.2%
	40:60	51.9%
	50:50	55.7%
	60:40	54.3%
	70:30	54.7%
	80:20	54.3%
	90:10	55.6%
Fold Cross Validation	***3***	**57.4%**
	5	51.7%
	10	53.4%

From Table 2, the 3 fold cross validation was chosen as the best model by 57.4% accuracy obtained. The Figure 6 shows the detailed results of NBC by applying 3 fold cross validation splitting data technique during the mining process.

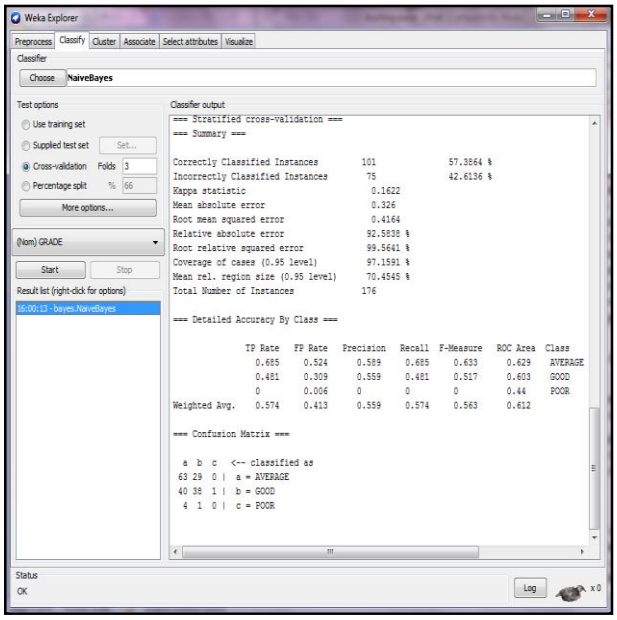


Figure 6 NBC Result of Student's Dataset produced by WEKA

The confusion matrix contains information about actual and predicted classifications done by WEKA. The confusion matrix resulted from 3 fold cross validation model of NBC is shown in Table 3 below.

Table 3 The output of NBC 3 fold cross validation model

	Classified as				
	Poor	***Average***	***Good***	***Total data***	***Prediction Success***
Poor	0	4	1	5	0%
Average	0	63	29	92	68.5%
Good	1	40	38	79	48.1%

From Table 3, about 57.4% of data were classified correctly. The average category shows the highest accuracy which is 68.5%, only 48.1% of good category is correctly classified and 0% shows of poor category. This prediction model gives a better classification for average student category and failed to predict the poor student category.

Moreover, from the confusion matrix, the chance prediction of a poor student category into a good category is 20%. It is a quite risky situation if we classify the students into the wrong category. We anticipate that the accuracy could be improved by adding more data in our next analysis.

An experiment on each parameter in this study was conducted using WEKA DM tool to identify the parameters that contributed to the students' academic success. The Table 4 shows the probability values for six parameters that were used for SAP prediction model construction.

Table 4 The parameter probability value

Parameter	Probability
Race	51.7%
Gender	52.3%
Hometown	52.3%
Family Income	56.8%
University Entry Mode	49.4%

From Table 4, we found out that family income parameter gives the highest influence on the SAP for Bachelor of Computer Science students, FIC, UniSZA. It was also revealed that the female students are better in their studies compared to male students. The result also indicated that female students with the low family income range RM1-RM3000 perform better in their studies. The results of prediction SAP using NBC obtained in this study were compared with the prediction result that also used NBC in past research. From the investigation, we found out that NBC achieved the highest prediction accuracy than other DM methods 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.

The extraction model will allow the lecturers to take early actions to help and assist the poor and average category students from getting low GPA at the end of the each semester.

5.0 FUTURE WORK

In future, this study can be expanded by adding more data to achieve more accurate results and to extract more hidden knowledge from students' data. In fact, the data from 2012/2013 will be added to the existing data when available. For further analysis, we plan to do a comparative analysis with other classification DM techniques. Furthermore, we aim to develop a simple prototype system to predict SAP so that it could be used as a decision making tool in providing appropriate assistance to the students to improve their academic performance.

6.0 CONCLUSION

The amount of data stored in an educational database at HLI 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, NBC as a DM method are applied to the students' data. Six parameters were selected in this study which is race,

gender, hometown, family income, university entry mode, and GPA. The mining result reveals that NBC gives the 57.4% accuracy in prediction. Prediction on average category students, gives the highest accuracy 68.5% of actual data. This study also discovers that the students' family income, gender, and hometown parameter contribute towards students' academic performance. The limitation of this study is the small size of data due to incomplete and missing value. For the next experiment, more data will be added so the accuracy of prediction model can be improved for faculty benefits in monitoring the SAP.

Acknowledgement

This research has been conducted to fulfil the needs for the certification of the Master of Computer Science (Data Mining) research study. A lot of thanks for the support and contribution from supervisor Associate Professor Dr. Fadhilah Ahmad, and Mr. Azwa bin Abdul Aziz. Also big thanks to Information Technology Center, UniSZA by providing a students' data for this research.

References

[1] Zhang, Y., S. Oussena, T. Clark, and H. Kim. 2010. Using Data Mining To Improve Student Retention In Higher Education-A Case Study. In *12th International Conference on Enterprise Information Systems (ICEIS)*. Portugal. June 8-12, 2010.

[2] Kumar, S. A. and M. N. Vijayalakshmi. 2012. Mining of Student Academic Evaluation Records in Higher Education. In *2012 International Conference on Recent Advances in Computing and Software Systems*. Chennai, India. April 25-27, 2012. 67-70.

[3] Lye, C. T., L. N. Ng, M. D. Hassan, W. W. Goh, C. Y. Law, and N. Ismail. 2010. Predicting Pre-university Student's Mathematics Achievement. In *Procedia of Social and Behavioral Sciences*. 8: 299-306.

[4] Yadav, S. K., B. Bharadwaj, and S. Pal. 2012. Data Mining Applications : A Comparative Study for Predicting Student's performance. *International Journal Of Innovative Technology & Creative Engineering*. 1(12): 13-19.

[5] Kumar, S. P. and K. S. Ramaswami. 2010. Fuzzy K- Means Cluster Validation for Institutional Quality Assessment. In *International Conference Communication and Computational Intelligence (INCOCCI)*.Erode. December 27-29, 2010. 628-635.

[6] Sachin, R. B. and M. S. Vijay. 2012. A Survey and Future Vision of Data Mining in Educational Field. In *2nd International Conference on Advanced Computing & Communication Technologies*. Rohtak, Haryana. January 7-8, 2012. 96-100.

[7] Kumar, U. and P. S. Pal. 2011. Data Mining : A Prediction of Performer or Underperformer Using Classification. *International Journal of Computer Science and Information Technologies (IJCSIT)*. 2(2): 686-690.

[8] Tair, M. M. A. and A. M. El-halees. 2012. Mining Educational Data to Improve Students' Performance : A Case Study. *International Journal of Information and Communication Technology Research*. 2(2): 140-146.

[9] Parack, S., Z. Zahid, and F. Merchant. 2012. Application of Data Mining in Educational Databases for Predicting Academic Trends and Patterns. In *IEEE International Conference on Technology Enhanced Education (ICTEE)*. Kerala. January 3-5, 2012. 1-4.

[10] García, E. P. I. and P. M. Mora. 2011. Model Prediction of Academic Performance for First Year Students. In *10th Mexican International Conference on Artificial Intelligence*. Puebla. November 26-December 4, 2011 169-174.

[11] Bhardwaj, B. K. and S. Pal. 2011. Data Mining : A prediction for Performance Improvement Using Classification. *International Journal of Computer Science and Information Security (IJCSIS)*. 9(4): 136-140.

[12] Sharma, M. 2011. Development of Predictive Model in Education System : Using Naïve Bayes Classifier. In *Proceedings of the International Conference and Workshop on Emerging Trends in Technology (ICWET 2011)*. Mumbai, India. Febuary 25-26, 2011. 185-186.

[13] Bhuvaneswari, R. and K. Kalaiselvi. 2012. Naive Bayesian Classification Approach in Healthcare Applications. *International Journal of Computer Science and Telecommunications*. 3(1): 106-112.

[14] Balaniuk, R., P. Bessiere, E. Mazer, and P. Cobbe. 2012. Risk based Government Audit Planning using Naïve Bayes Classifiers. In Grana M. *et al*. (Eds.). *Advances in Knowledge-Based and Intelligent Information and Engineering Systems*. 1313-1323.

[15] Aziz, A. A., N. H. Ismail, and F. Ahmad. 2013. Mining Students' Academic Performance. *Journal of Theoretical and Applied Information Technology*. 53(3): 485-495.