IOP Conf. Series: Materials Science and Engineering **649** (2019) 012018 doi:10.1088/1757-899X/649/1/012018

# eMineProve: Educational Data Mining for Predicting Performance Improvement Using Classification Method

**Jovel T. Rosado[1], Angelica P. Payne[2] and Corazon B. Rebong[3]**

[1]University of Perpetual Help System Laguna, Philippines
[2]Olivarez College, Tagaytay, Philippines
[3]Colegio de San Juan de Letran, Calamba, Philippines

rosado.jovel@uphsl.edu.ph, angiepayne5@gmail.com, corazonbrebong@gmail.com

**Abstract.** Today, Data mining has become a universal tool for converting data into useful information and knowledge. Student's academic performance is a crucial factor in building their future [3]. In this study, the classification method is selected to be applied on the students' data. Classification is a form of data analysis that extracts models describing data classes [4]. The data is collected from the Basic Education Department of the University of Perpetual Help System Laguna. The raw data set is a collection of 4, 250 data accumulated over two academic years regarding the basic information of the Grade 7 Junior High School students. The data mining technique applied was classification using Naïve Bayes. The data mining tool used to process the data into useful knowledge is RapidMiner. The main objective of this study is to predict the performance improvement of Grade 7 Junior High School students for A.Y. 2016-2017 and 2017-2018 using classification. It specifically aims to identify the general average of the Grade 7 Junior High School students when grouped according to gender, identify who perform better between male and female, identify the subject on which the students excel most, identify the subject on which students have difficulty, identify who performs best when group according to last school attended, identify the academic performance of the students based on their parent's occupation, provide a predictive analysis of data to help the decision makers create a marketing strategy for those schools where only few students enrolled. Naïve Bayes produces accuracy 92.37% that shows it is possible to obtain a good prediction model based on the academic performance of the students. Based on the data for marketing strategy, the model has an accuracy of 30.97%.

## 1. Introduction

As our society is growing in terms of population, technology advancements, literary rate and global competitiveness, education also is taking a leap as business systems do. Today, Data mining has become a universal tool for converting data into useful information and knowledge. Data Mining (DM) also known as Knowledge Discovery from Databases (KDD), is the area of discovering unique and potentially useful information from the large amount of data [1].

Educational Data Mining (EDM) is the application of Data Mining techniques on educational data. To analyze data and to resolve educational research issues is the main objective of EDM. It deals with developing new approaches to explore the educational data, and using it to better understand student learning environment [2].

Student's academic performance is a crucial factor in building their future [3]. In this study, the classification method is selected to be applied on the students' data. Classification is a form of data analysis that extracts models describing data classes [4].

The main objective of this study is to predict the performance improvement of Grade 7 Junior High School students for A.Y. 2016-2017 and 2017-2018 using classification. It specifically aims to identify the general average of the Grade 7 Junior High School students when grouped according to gender, identify who perform better between male and female, identify the subject in which the students excel most, identify the subject in which students have difficulty, identify who performs best when group according to last school attended, identify the academic performance of the students based on their parent's occupation, provide a predictive analysis of data to help the decision makers create a marketing strategy for those schools where only few students enrolled.

## 2. Related Works

A number of researchers have published work that studied various EDM techniques used to better understand and enhance student performance. Applied association rules data mining techniques on learner data to evaluate knowledge about learners' academic performance [5].

K. Shanmuga Priya and A. V. Senthil Kumar [6] applied a Classification Technique in Data Mining to improve the student's performance and help to achieve the goal by extracting the discovery of knowledge from the end semester mark.

Monika Goyal and Rajan Vohra [7] applied data mining techniques to improve the efficiency of higher education institution. If data mining techniques such as clustering, decision tree and association are applied to higher education processes, it would help to improve students' performance, their life cycle management, selection of courses, to measure their retention rate and the grant fund management of an institution

Cortez and Silva [8] attempted to predict failure in the two core classes (Mathematics and Portuguese) of two secondary school students from the Alentejo region of Portugal by utilizing 29 predictive variables.

Ramaswami et.al.[9], developed a predictive data mining model for students' performance so as to identify the slow learners and study the influence of the dominant factors on their academic performance.

## 3. Methodology

### *3.1. Data Preprocessing Phase*

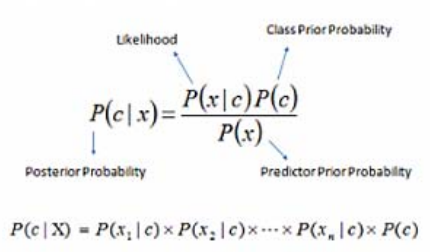
Data of 4,250 from the Grade 7 Junior High School students of the Basic Education of the University of Perpetual Help System Laguna from A.Y. 2016-2017, 2017-2018. There are total of 250 students of Grade 7 Junior High School for the two succeeding years. The data were consisting of the gender of the students, originating schools, address, parents' occupation, final ratings in all subject areas and their final average.

The dataset has to go through a preprocessing phase to clean the data from different errors and remove non-important and redundant attributes. The required data for evaluating students' performance are scattered over many tables; thus, this phase starts with attribute selection and merging multiple tables into a single table that contains the most important attributes for the required evaluation.

The last step in this phase is the transformation of the dataset into a suitable format for the mining algorithms.

### *3.2. Data Mining Phase*

In this phase, classification technique is used using Naïve Bayes are applied using RapidMiner software. RapidMiner tool is used for exploration, statistical analysis and mining the students' data. Among the algorithms, Naïve Bayes as classification technique will be used. Naïve Bayes algorithm is a simple probabilistic classifier that calculates a set of probabilities by counting the frequency and combinations of values in a given data set. The data classification process involves learning and classification. In learning the training data are analyzing by classification algorithm. In classification test data are used to estimate the accuracy of the classification rules. [10]




(1)

Equation (1) shows Bayes Theorem which provides a way of calculating posterior probability P(clx) from P(c), P(x), and P(xlc). Here;

- P(clx): posterior probability of class (c, target) given predictor (x, attributes). This represents the probability of c being true, provided x is true.
- P(c): is the prior probability of class. This is the observed probability of class out of all the
- observations.
- P(xlc): is the likelihood which is the probability of predictor-given class. This represents the probability of x being true, provided x is true.
- P(x): is the prior probability of operator. This is the observed probability of predictor out of all the observations.

Table 1 presented the variables used in data mining that include the gender, last school attended, subjects, and parents' occupation, grade per subject, general average, academic performance and marketing.

**Table 1.** Students' Record Its Attributes, Description and Possible Variables.

| Variables | Descriptions | Possible Values |
|---|---|---|
| Gender | Students gender | {Male, Female} |
| Last School Attended | The originating school of the students | {Home-grown, Transferred-in from Private, Transferred-in from Public} |
| Subjects | The subjects taken by the students | {Filipino, Araling Panlipunan, MAPEH, FCL, English, Science, Math, TLE, Financial Literacy, Personal Development, Computer} |
| POcc | Parents' Occupation | {Business, Employed} |
| Grade | Rating of the students on different subjects | {70%-100%} |
| GenAv | General Average of the students | {70%-100%} |
| APer | Final Academic Performance of the students | {90% above- Outstanding, 85%-89%-Very Satisfactory, 80%-84%-Satisfactory, 75%-79%-Fair Satisfactory, 74%-below-Beginning} |
| Marketing | Options of how students knew about perpetual | School campaign, Parent/s is/are University employee, Parent/s alumni, Through a friend, Through a relative |

## 4. Results and Discussions

The data extracted from the database of the school system was cleaned, modified and mined based on the objectives presented. The model used to analyze the data is RapidMiner employing the data mining technique which is classification using Naïve Bayes.

Figure 1 shows the general average of the students when grouped according to gender. The chart displays that there were lot of female students who had an average between 86%-87% in Grade 7. Most females have an average grade of 80%-93% while male mostly got an average of 76%-89%. However, a male student outnumbered those females who got 95% as general average.

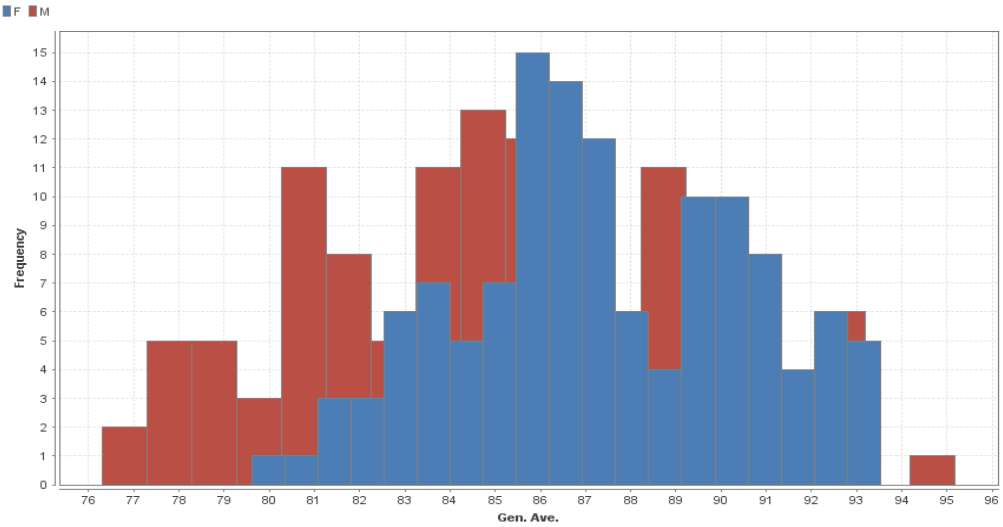



**Figure 1.** General Average of the Students When Grouped According to Gender.

Figure 2 shows the academic performance of the Grade 7 students based on their gender. The chart explained that mostly of the female students were very satisfactory in terms of academic performance compared to those male students, these were the students who got a general average of 85%-89%. Outstanding students were mostly female also and those who were satisfactory and fairly satisfactory were boys.

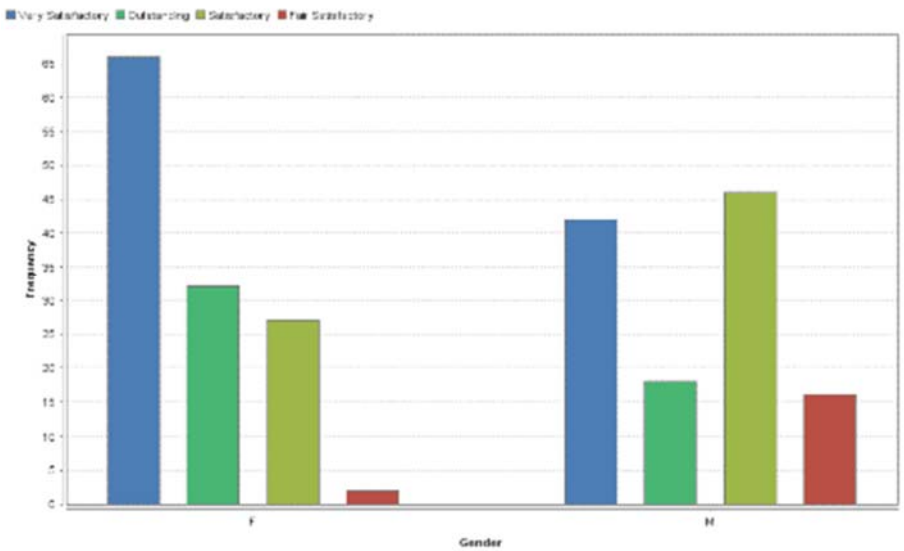



**Figure 2.** Academic Performance of Male and Female.

Figure 3 shows the distribution of the grades per subject. The chart indicates the subject that most of the students excel most is Personal Development. Based on the general average of the students, Personal Development subject has the highest percentage of passing compared to other subjects while students had difficulty in Araling Panlipunan, most of them got low grades in the said subject.

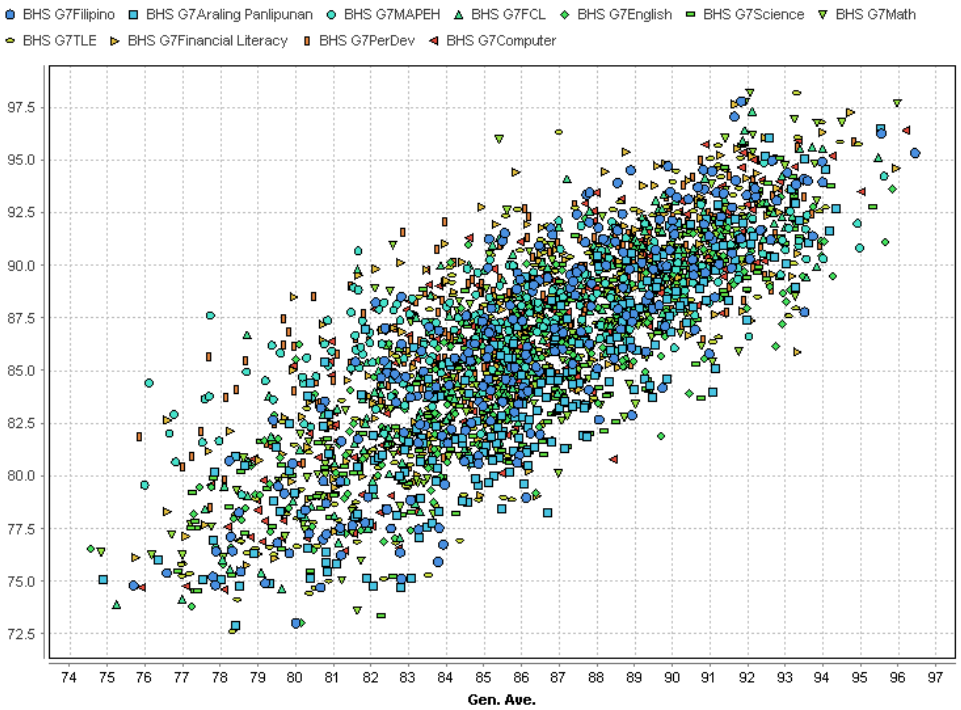



**Figure 3.** Grades of Students Per Subject.

Figure 4 shows the general average of students when grouped according to last school attended. The chart displays that the students who transferred-in from private schools had highest average obtained in Grade 7 compared to those who were homegrown and transferred-in from public.

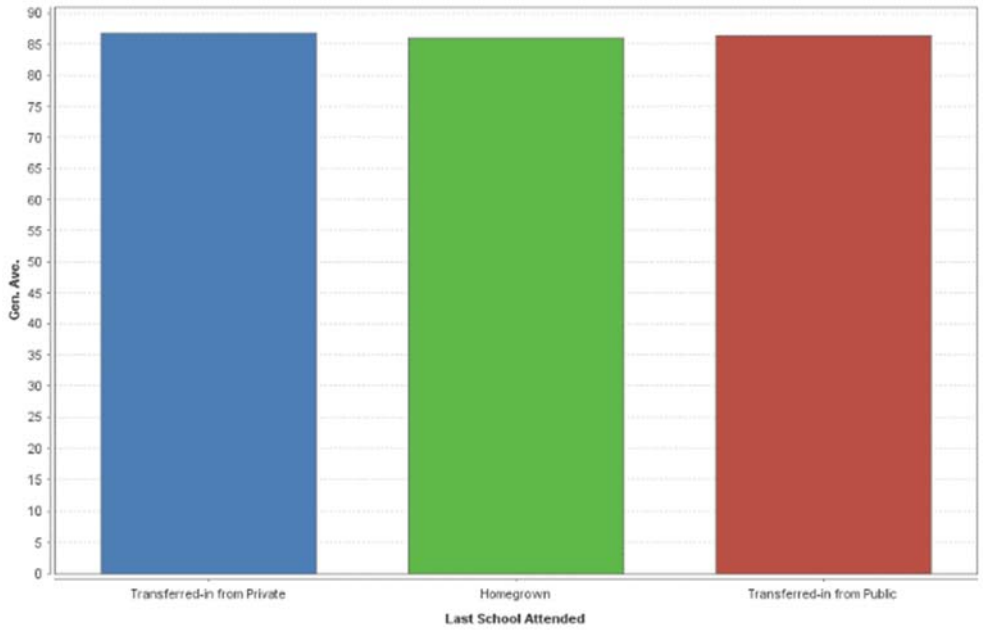



**Figure 4.** General Average When Grouped According to Last School Attended.

As indicated in figure 5, students whose parents were businessmen had obtained higher general average in Grade 7 compared to those students whose parents were employees. Businessmen parents had time freedom compared to those parents who are employees. They had all the resources that they can provide to their children; time and money.

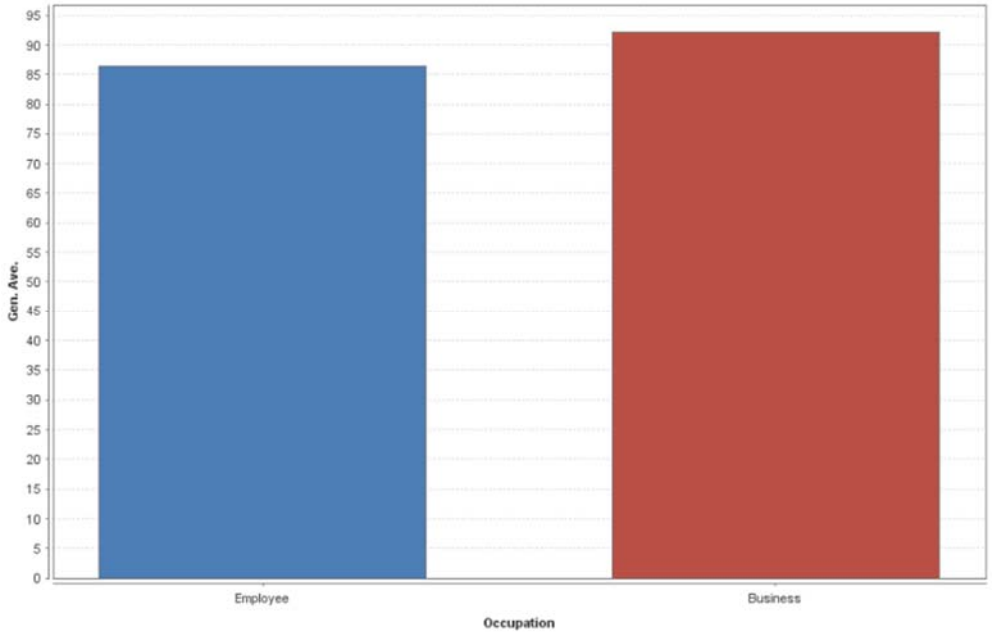



**Figure 5.** General Average When Grouped According to Parents' Occupation.

The figure shows the predictive analysis of the students' data in terms of their last school attended. The data shows that the accuracy of the model is 35.73% based on the last school attended. Class precision for Transferred-in from private school is 40.74%, 40.21% from the Homegrown school and a

class precision of 13.64% is the Transferred-in from the Public school. The results show that there is a high possibility of analysis that students who will enroll will come from the private schools.

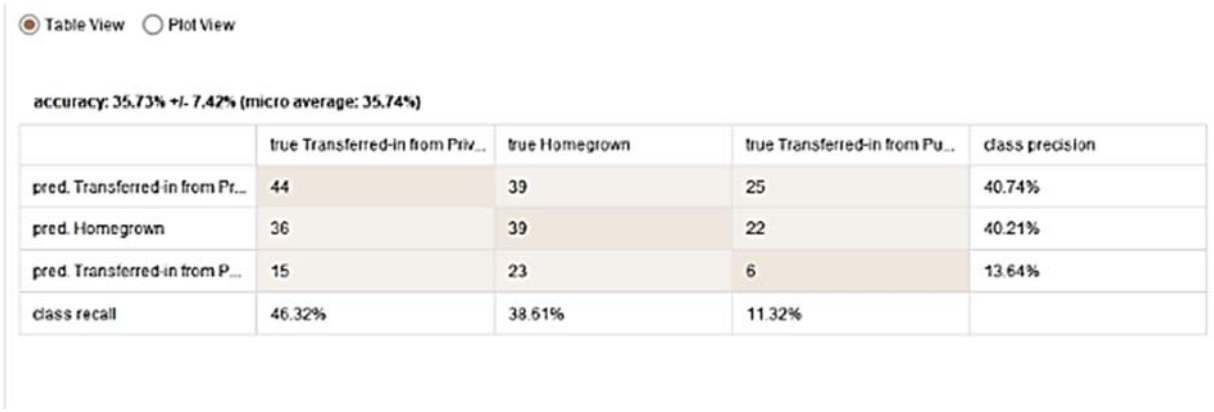


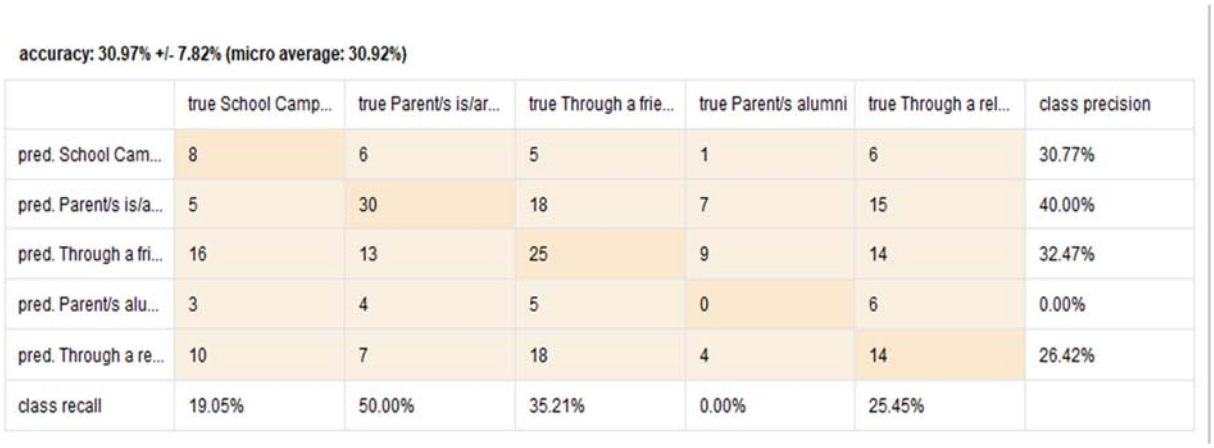
Table View Plot View

accuracy: 35.73% +/- 7.42% (micro average: 35.74%)

| | true Transferred-in from Priv... | true Homegrown | true Transferred-in from Pu... | class precision |
|---|---|---|---|---|
| pred. Transferred-in from Pr... | 44 | 39 | 25 | 40.74% |
| pred. Homegrown | 36 | 39 | 22 | 40.21% |
| pred. Transferred-in from P... | 15 | 23 | 6 | 13.64% |
| class recall | 46.32% | 38.61% | 11.32% | |

**Figure 6.** Class Precision in Terms of Last School Attended.

The figure shows the class precision and the class recall of data based on the marketing strategy used by the school. The data sets used here were school campaign, parent/s is/are University employee, through a friend, parent/s alumni and through a relative. These were the options given as to what how the students know about Perpetual. Class precision refers to the ratio of correctly predicted positive observations to the total predicted observations which is shown that Parent/s is/are University employee has the highest percentage of precision of 40%, the same thing is with the class recall which has a percentage of 50%. Recall is the ratio of correctly predicted positive observations to the all observations in actual class-yes. Thus the prediction of data shows that mostly of the students enrolled in Perpetual is because their parents are employees.



accuracy: 30.97% +/- 7.82% (micro average: 30.92%)

| | true School Camp... | true Parent/s is/ar... | true Through a frie... | true Parent/s alumni | true Through a rel... | class precision |
|---|---|---|---|---|---|---|
| pred. School Cam... | 8 | 6 | 5 | 1 | 6 | 30.77% |
| pred. Parent/s is/a... | 5 | 30 | 18 | 7 | 15 | 40.00% |
| pred. Through a fri... | 16 | 13 | 25 | 9 | 14 | 32.47% |
| pred. Parent/s alu... | 3 | 4 | 5 | 0 | 6 | 0.00% |
| pred. Through a re... | 10 | 7 | 18 | 4 | 14 | 26.42% |
| class recall | 19.05% | 50.00% | 35.21% | 0.00% | 25.45% | |

**Figure 7.** Predictive Analysis of Marketing Strategy.

## 5. Conclusions

In this study, a model was developed based on some selected input variables. Data mining classification algorithm Naïve Bayes was applied to predict the academic performance of students based on the previous years' database. The tool Rapid miner is used for exploration, statistical analysis and mining of student data. Cross-Validation operator is used to perform a cross-validation process. From the above analysis, the researchers concluded that the Naïve Bayes produces accuracy 92.37% that shows it is possible to obtain a good prediction model. The proposed methodology can be adopted to predict the performance of students and help the teachers and the management to enhance the quality of learning and student's academic performance by doing important decision at the right time. The predictive

analysis based on the data could help the school intensify the marketing strategies to encourage lot of students from the community. In the future, the study can be enhancing by including the data with more information about the students and of higher quality which might help to improve the current model performance and also to obtain more accurate student performance.

**References**

[1] Arockiam L, Charles S, Arulkumar, et al. 2010 *Int. J. Computer Sci. Eng.* Deriving Association between Urban and Rural Students Programming Skills, **2(3)** 687-90.
[2] Baker R S, Corbett A T and Koedinger K R 2004 *7th Int. Conf. Intell. Tutoring Syst.* Detecting Student Misuse of Intelligent Tutoring Systems, **3220** 531-40.
[3] Baker R S, Corbett A T and Koedinger K R 2004 Detecting Student Misuse of Intelligent Tutoring Systems. Proceedings of the International Conference on Intell. Tutoring Syst. 3220 531-40.
[4] Romero C and Ventura S 2010 *IEEE Trans. Syst. Man Cyber. C. (Appl. Rev.)* Educational Data Mining: A Review of the State-of-the-Art, **40(6)** 601-18.
[5] AZIZ A. A., ISMAIL N.H., AHMAD F. and HASAN H 2015 A Framework for Students' Academic Performance Analysis using Naïve Bayes Classifier in Proceedings of the ICIDM.
[6] Priya K S and Senthil Kumar A V 2013 *Int. J. Adv. Netw. Appl.* Improving the Student's Performance Using Educational Data Mining, **4(4)** 1680-5.
[7] Goyal M and Vohra R 2012 *Int. J. Comput. Sci. Issues* Applications of Data Mining in Higher Education, **9(2)**.
[8] Cortez P and Silva A 2008 *EUROSIS-ETI* Using Data Mining to Predict Secondary School Student Performance.
[9] Ramaswami M, Bhaskaran R and CHAID A 2010 *Int. J. Comp. Sci. Issues* Based Performance Prediction Model in Educational Data Mining, **7(1)**.
[10] Dimitoglou G, Adams J A and Jim C M 2012 *J. Comp.* Comparison of the C4.5 and a Naïve Bayes Classifier for the Prediction of Lung Cancer Survivability, **4(8)**.