Journal of Theoretical and Applied Information Technology
10th September 2015. Vol.79. No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A NOVEL APPROACH TO EFFICIENT EXPLORATION AND MINING OF STUDENTS' DATA

[1]ADELEKE RAHEEM AJIBOYE, [2]RUZAINI ABDULLAH ARSHAH, [3]HONGWU QIN
Faculty of Computer Systems & Software Engineering
Universiti Malaysia Pahang, Malaysia.
Lebuhraya Tun Razak, 26300 Kuantan, Pahang, Malaysia.

E-mail: [1]ajibraheem@live.com, [2]ruzaini@ump.edu.my, [3]qhwump@gmail.com

ABSTRACT

Several undergraduate students most especially the new students do have an unacceptable Grade Point Average (GPA) at the end of their first year. Research has shown that during this period, stress by this set of students sometime responsible for their poor performance as they are faced with many challenges. Having knowledge about their previous academic knowledge can bring about a workable solution to prevent this trend. In this paper, a novel approach that efficiently explores students' data for the prediction of their performance based on their historical data is proposed. A prediction system is designed in line with the concept of the Unified Modelling Language (UML) and it is implemented using PHP; while MySQL serves as the back-end. The developed model is tested and a more satisfactory result is achieved when compared with similar models emanated from using machine learning techniques. The resulting outputs of this study unveil the academic achievement of each student prior to their first registration in the university. The model developed is also found to be useful for efficient planning and quick decision making that can ensure a sustainable educational growth.
Keywords: *Educational Data Mining, UML, Predictive Model, Student's Achievement, Data Exploration*

1. INTRODUCTION

The number of students that are withdrawn from the universities annually as a result of their inability to get the acceptable GPA is worrisome; a workable solution to prevent this trend is therefore inevitable in order to bring the failure rate to the barest minimum. One of the reasons that necessitate the prediction of students' performance has been observed as to improve the targeting of interventions and support services for students that are found to be at risk of academic problems [1]. The sets of students that could not make the minimum GPA at the end of the academic session forms the bulk of dropout students each year and according to Tan and Shao [2], the key to reducing the rate of dropout is to utilize various factors associated with dropping out. This would help to identify the potential dropout among the students in order to take a targeted retention measures to avert mass failure.

The challenges in academic performance of students can be easily solved if information about their past academic achievement could be determined. Findings from such exploration would help in making decisions that can be of immense benefit to them at the early stage of their studies. A number of techniques can be used to explore the students' data and one of such approaches is the use of data mining techniques. This technique involves construction of models to unveil useful information or relevant patterns embedded in the dataset. According to Engel et al. [3], a model is an abstraction of the physical system with a certain purpose; a useful model has just the right amount of detail and represents only what is important for the task at hand. Data mining has its origin in various disciplines, the two most important are statistics and machine learning [4].

Machine learning techniques, automatically learn the pattern in the data and subsequently make intelligent decisions based on the data [5]. Sometime, it is not realistic for a machine learning technique to achieve a complete accuracy due to many factors. For instance, while using a neural network, the choice of hidden neurons and other parameters determines the possibility of achieving a good generalization. The more the number of hidden neurons, the greater the ability of the network to recognise existing patterns [6].

However, care must be taken in the choice of neurons to avoid over fitting.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Prediction is described in [7] as producing an appropriate label from a number of attributes or cartegorization for new objects, using the information gleaned from the relationship between the attribute values. It is these relationships that machine learning techniques try to understand in order to predict unknown attributes. Most of the earlier work that reported the prediction of student' performance paid much attention to the performance of students on a particular course [8] or predicting the performance of final year students [9]. Only little attention is so far focused on the students that are in the first year of their programme (fresh students). The present study, therefore, focuses on this set of students, by exploring their past data using some predictive attributes identified in their records.

The main objective of this study was to design a predictive model for prediction of academic performance of first year undergraduate students using their historical data; this is with a view to improving the management of their weaknesses and strengths. The model developed in this study has several benefits over the model developed for the same purpose using some machine learning techniques. The proposed model is designed to process thousands of students' records without sacrificing its accuracy; the developed system is user friendly and provides a statistical summary in respect of the prediction made for all the students. The rest of the paper is organized as follows: In the next Section, some related works are discussed. In Section 3, the approach used in this study is discussed, the design and implementation of the system is also illustrated. The results are displayed and discussed in Section 4, while the study concludes in Section 5.

2. RELATED WORK

In this section, some related studies earlier reported in the literature in the area of model construction from students' data are reviewed. The reliability of models constructed for students' performance prediction, depends on a number of factors, the data being explored and the approach for the exploration is crucial to achieving a good result. Rajaraman and Ulman[10] simply described data mining as the discovery of models for data, the choice of techniques for the data exploration and how the model is evaluated and its ability to generalize with a set of new inputs are very crucial aspect in Data Mining.

The use of machine learning techniques for model development to unveil knowledge from educational data is well reported in the literature. The study in [11], identified the techniques of data mining and machine learning as research areas where creation of models for students' academic performance predictions are gaining popularity and can help to unveil useful knowledge. The study proposed by Lye et al. [9], using the back propagation algorithm for the training of some students' data focused on developing a model that is capable of predicting students' final year examination result. Also, a comparative study for the use of machine learning techniques such as neural network techniques, and other frequently used methods such as decision tree and linear regression as reported in [12] shows that, neural network gives better prediction results.

The study carried out using the decision tree technique as reported in [13] shows that, the technique is one of the methods that can be reliably used for academic performance prediction as it generate rules, these rules can then be implemented using any suitable programming language. However, there is a tendency to experience a reduction in accuracy as the tree grows with many branches.

A study by Oladokun et al [14] used the neural network technique for construction of models to predict the students' performance in an engineering course. The network model developed in the study achieved some level of acceptable accuracy. The study is much related to the present work. Most of the researchers that explored the use of data mining techniques focused on the accuracy of the model; however, the resulting output of the model in terms of usefulness is the most important.

The technique of Bayesian network classifier was applied by Sundar [15]. The study compared a number of classifier algorithms and explored some predictive attributes such as the previous semester marks, internal marks, performance in the seminars, assignment, attendance etc., to construct a predictive model.

The present study focuses on exploration of students' data; the study was aimed at developing a predictive system capable of predicting the student academic performance for the newly admitted undergraduate students. The proposed study consistently gives much better results when compared to the use of a machine learning technique to handle a similar task. The developed model using a web scripting language- PHP is robust and gives much better result.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

3. MATERIAL AND METHODS

The predictive model is designed and implemented in this section to explore the historical data of some undergraduate students.

Table 1: The Predictive Variables And Their Normalized Values

Input variables	Normalized values	Obtainable values
Score in National Exam. (SNE)	UTME: 200-219 1 220-229 2 230-239 3 240 and above 4 DIRECT: ND/NCE/A'LEVEL 2 HND 3 BSC 4	1 – 4 2 - 4
High School Result (HSR)	GRADES: A1 6; A2, B2 5; A3, B3 4; C4 3 C5 2; C6 1	1 - 6
Qualify Exam by Institution (QEI)	SCORES (%) 50-59 1 60-65 2 66-69 3 70 and above 4	1- 4
Number of Attempts for Qualify Exam (NAQE)	1 Attempt 2 2 Attempts 1	1 - 2
Student's Age (SA)	14-20 3 21-29 2 Above 30 1	1-3
Year of last Exam. (YLE)	1 >= YLE <= 4 3 5 >= YLE <=8 2 YLE > 8 1	1 - 3

The predictive attributes here referred to as *input* are identified based on the information received through the interaction with the domain experts. Other academic /demographic variables reported in the literature [16] that are considered to have some predictive influence on students' performance prediction are also taken into consideration. The records are pre-processed into a format suitable for exploration; the achievement score is computed for each student. All the 1300 records explored in this study were collected from a public university in the north central, Nigeria. The students' high school result which is one of the predictive attributes in this study conforms to the West African Examination Council (WAEC) grading standards.

The excerpts of the data collected are normalized in accordance with the normalized values represented in table 1. The normalized data are stored in a file that consists of student identification number and the predictive attributes (see Table 2). This is how the format of the file to be explored is expected to be before it can be uploaded to the proposed system.

3.1 Design and Implementation

The proposed system is designed using the concept of the UML as shown in Figure 1. A web scripting language, PHP, is used to implement the design of the proposed model. Figure 2, illustrates the uploading of the predictive attributes and the student identification number that conforms to the format represented in Table 2. The maximum obtainable score a student can obtain after summing up the scores achieved in all the attributes is supplied in the input field before the user click the upload data button.

There are some requirements in the form of hardware and software that must be met for smooth running of the proposed model. The model is tested and run satisfactorily on a computer with the following configurations: 4GB of RAM, Quad-Core Intel Core i7 @ 3.4 GHz processor and 500GB hard disk. There is the possibility of experiencing slow performance if configurations fall below the listed values. The installed software for the running of this model include: Windows 8, WampServer, and Macromedia Dreamweaver 8.

In order to predict the performance of the students, the present study adopts knowledge-driven approach whereby, the domain expert provides the information in the form of rules that can assist in the interpretation of the overall score achieved by each student. In practice, to effectively use the technique of data mining to make an acceptable decision, a cooperative effort of humans and computers is very crucial, as best decisions are

Journal of Theoretical and Applied Information Technology
10th September 2015. Vol.79. No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

only achievable by balancing the knowledge of human experts in describing problems and goals with the search capabilities of the computer [4].

From the normalized data earlier shown in Table 1, the maximum total obtainable score is 46; the score achieved by each student is further standardized to a percentage, based on the formula in (1). For instance, if a student achieved the following values in the input attributes: SNE= 2; HSR=24, i.e. the value achieved in 5 subjects; NEA=2; Age=2; STI=2; YLE=2 this gives a total of 34, using equation (1), the student has gotten 34/46 x 100 = 73.9% .

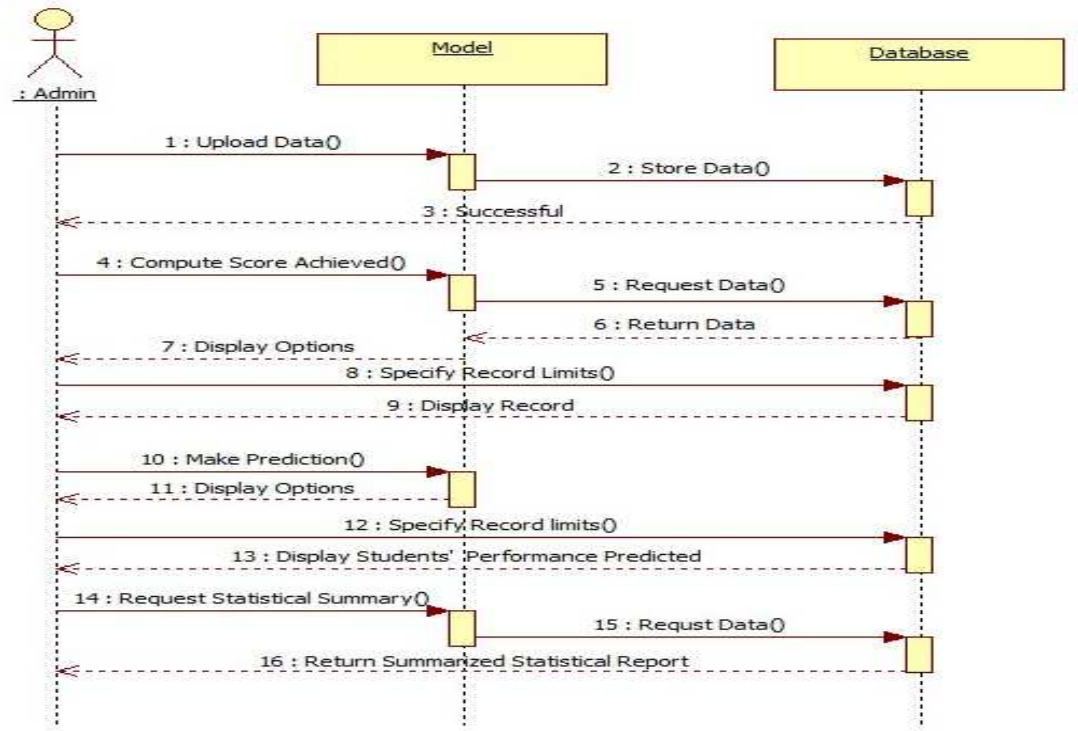

Figure 1: Sequence diagram showing user interaction with the proposed system

$$\frac{v}{Y} \; x \; 100 \qquad (1)$$

where v is the value obtained, and Y is the total obtainable value. The likely performance of each student is therefore, predicted based on a number of rules.

Generally, rules consist of two parts: the IF part, which is usually referred to as the antecedent-premise or condition and the THEN part, usually referred to as the consequent- conclusion or action. According to [6], a rule can have multiple antecedents joined by the keywords AND, OR and sometimes a combination of both. The rules that guided the prediction of students' performance in this study, which is subsequently implemented using the web scripting language, PHP are:

IF score achieved <50 and risk status ='unacceptable', THEN performance predicted is '***Poor***'.

IF score achieved >=60 and the score achieved <70 and risk status ='low risk', THEN performance predicted is '***Good***'.

IF score achieved >=70 and the score achieved <=100 and risk status ='risk-free', THEN performance predicted is '***Excellent***'.

A prompt message is displayed to confirm the success or otherwise of the uploaded file as shown in Figure 2. The successful upload implies that, the data to be explored has been properly stored in the MySQL database that serves as the backend.

The present study considers six predictive attributes; the proposed model is designed to accommodate 20 attributes, this is for the purpose of generalizing. The next link on the model's sidebar shown in Figure 2 is the computation of the achievement by all the students using the predictive attributes in the uploaded file.

Journal of Theoretical and Applied Information Technology
10th September 2015. Vol.79. No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

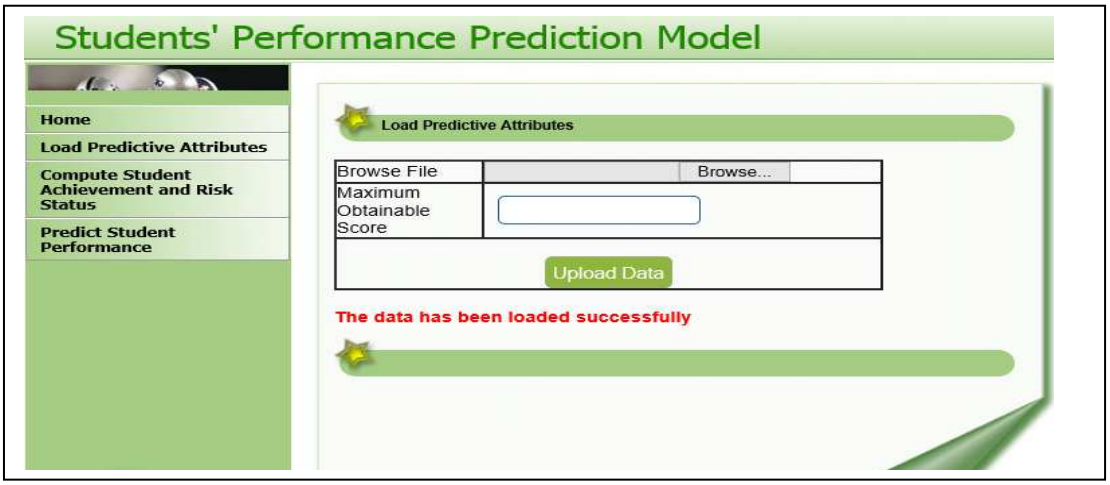

Figure 2: Interface For Uploading Of The Normalized Predictive Attributes

Relevant PHP codes process the data of the uploaded file to aid classification of the processed records. For instance, in order to determine those that appears to have excellent performance; the model prompts a message to supply the desired limit of records as shown in Figure 3; here, 20 is inputted. The statistical summary can also be determined in order to have a broad information about the whole data being explored.

Table 4 shows the best 20 students and gives a summary that unveils the risk status associated with each students' achievement. The students' performance prediction link on the side bar consolidates what has been processed in the previous link to make some inference. The performance predicted in respect of each student is directly related to the total score a student obtained and the status of risk attached to such an achievement as illustrated in Table 3.

The last link on the side bar predicts the students' performance. The link triggers the performance predicted and append same to the outputs results of computing the students' achievement score and the associated risk status.

Table 2: Format Of The Uploaded Predictive Attributes

Reg no	Attr_1	Attr_2	Attr_3	Attr_4	Attr_5	Attr_6
12CC01						
12CC02						
12CC03						
12CC04						
12CC05						
12CC06						
12CC07						
12CC08						
12CC09						
.....						
.....						
.....						
12CC1300						

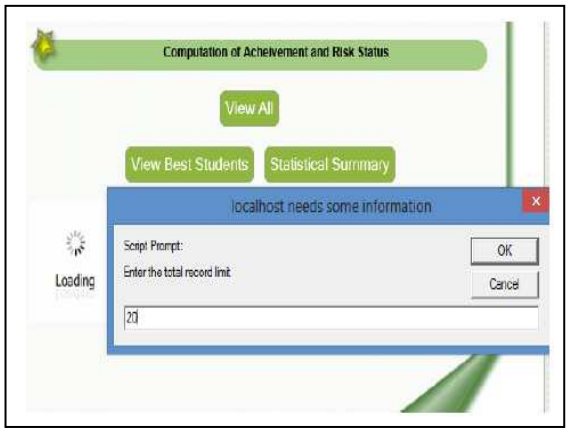

Figure 3: Prompts Message To Receive The Number Of Best Students To Display

4. RESULTS AND DISCUSSION

In this section, some outputs of the proposed model are represented. Table 3 shows the results of students' achievement computed based on their historical data, and the risk associated with each achievement. It is possible to have thousands of students' data processed at a click of a button with the proposed system; the performance of all the students whose data were uploaded are predicted, but due to space constraint of this paper, only the first 20 students are represented in Table 4.

Journal of Theoretical and Applied Information Technology
10th September 2015. Vol.79. No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 3: Excerpts Of The Computed Students' Achievement Scores And Their Associated Risk

S/N	REGISTRATION NO	SCORE ACHIEVED	RISK STATUS
1	12CC658	93.02	Risk Free
2	12CC949	93.02	Risk Free
3	12CC696	90.7	Risk Free
4	12CC920	90.7	Risk Free
5	12CC682	90.7	Risk Free
6	12CC027	90.7	Risk Free
7	12CC067	90.7	Risk Free
8	12CC752	88.37	Risk Free
9	12CC076	88.37	Risk Free
10	12CC760	88.37	Risk Free
11	12CC779	88.37	Risk Free
12	12CC084	88.37	Risk Free
13	12CC1083	88.37	Risk Free
14	12CC505	88.37	Risk Free
15	12CC1219	88.37	Risk Free
16	12CC1146	86.05	Risk Free
17	12CC148	86.05	Risk Free
18	12CC744	86.05	Risk Free
19	12CC591	86.05	Risk Free
20	12CC767	86.05	Risk Free

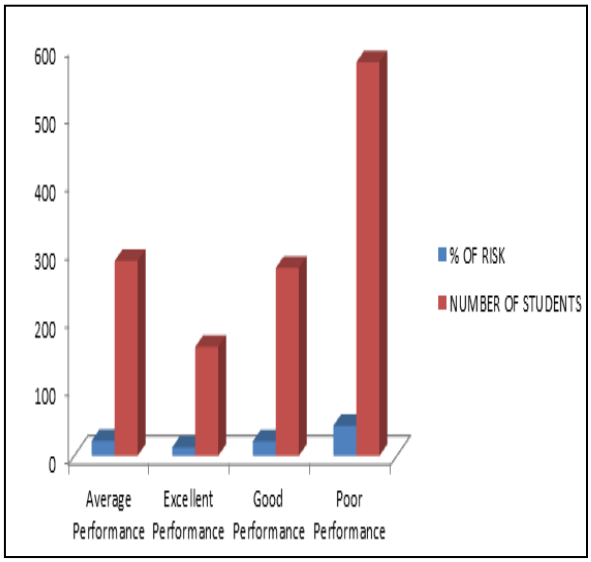

Figure 4: Summary Report Of The Students' Performance Prediction

Table 4: Students' Performance Prediction

	REGISTRATION NO	SCORE ACHIEVED	RISK STATUS	PERFORMANCE PREDICTED
1	12CC001	48.84	Unacceptable	Poor Performance
2	12CC002	53.49	High Risk	Average Performance
3	12CC003	58.14	High Risk	Average Performance
4	12CC004	79.07	Risk Free	Excellent Performance
5	12CC005	65.12	Low Risk	Good Performance
6	12CC006	83.72	Risk Free	Excellent Performance
7	12CC007	46.51	Unacceptable	Poor Performance
8	12CC008	30.23	Unacceptable	Poor Performance
9	12CC009	55.81	High Risk	Average Performance
10	12CC010	62.79	Low Risk	Good Performance
11	12CC011	48.84	Unacceptable	Poor Performance
12	12CC012	53.49	High Risk	Average Performance
13	12CC013	72.09	Risk Free	Excellent Performance
14	12CC014	62.79	Low Risk	Good Performance
15	12cc015	67.44	Low Risk	Good Performance
16	12CC016	39.53	Unacceptable	Poor Performance
17	12CC017	76.74	Risk Free	Excellent Performance
18	12CC018	39.53	Unacceptable	Poor Performance
19	12CC019	53.49	High Risk	Average Performance
20	12CC020	76.74	Risk Free	Excellent Performance

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

In order to validate the model developed in this study, the model is tested using different sizes of students' historical data and the output results are found to be consistently accurate and adhered to all the rules listed earlier.

4.1 Comparison of Results

The proposed model can be compared to the similar model created for the same task using machine learning methods, such as neural network techniques. Neural Networks are widely reported in the literature for the modelling of students' performance prediction [12], [14], [2], [17], [18] and [19].

A neural network is a biologically inspired non-linear parallel computing paradigm for information processing and exploratory analysis; it uses neurons that are interconnected by signal channels and fine-tuned by a training algorithm [20].

In the process of using neural network for model construction, the dataset to be modelled iterates for a specified number of epochs and subsequently converges; the relationship between the input attributes and the target attribute is thus established. Our comparison are based on the following criteria:

Generalization: The present model maintains the correctness of its outputs irrespective of the size of students' historical data that is being explored. Our previous findings reported in [21] show that, the use of ANNs for construction of predictive models can achieve good generalization provided the training data sufficiently span through the input space. Generalization is the ability of an Artificial Neural Network or a predictive model to produce correct results from the data on which it has not been trained [22].

Accuracy: This is the degree at which a model produces the outputs that tally with the desired target or the level of conformity to the rules being implemented. In the present study, the implementation of the rules using PHP is free of syntax error and semantically correct; it, therefore, gives accurate outputs. ANNs model can also achieve a similar degree of accuracy if properly configured and the data is properly trained.

Rule-based: Neural network model does not generate rules. The processing involved in the mapping of the input attributes to the desired target is carried out within the hidden layer in an opaque manner, the proposed model is designed and coded based on certain clearly understood rules.

Size of data: The resulting outputs of simulating network model build from data using neural network techniques, are sometimes affected by the size of the data available for training. For instance, if too few data that does not cut across the input attributes are trained, the outputs of such model when simulated with the new untrained input becomes inaccurate. Regardless of the size of the dataset, the proposed model is consistently accurate.

Transparency: The major processing task of ANN is done at the hidden layer; how the input transform to the output is not clearly revealed. The training process is mostly described as opaque [6]; the operation within the hidden layer (s) is not transparent and seems like a black box.

The comparison of the proposed model with the use of neural network techniques for the same task is summarized in Table 5.

Table 5: Comparison Of The Predictive Model Constructed Using Neural Network Technique And The Proposed Model

Techniques	Features				
	Generalization	Accuracy	Rule-based for clear understanding	Independent of data size	Transparency
Neural network	✔	✔	x	x	x
Proposed model	✔	✔	✔	✔	✔

Note : ✔ denotes **supported**

x denotes **not supported**

5. CONCLUSIONS AND FUTURE STUDY

This study explored the historical data of 1300 undergraduate students. The data have been

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

captured prior to their first registration in the university and the exploration is done with a view to determining their past academic achievement in order to predict their future performance. A prediction model is designed and the implementation is done using a web scripting language, PHP. Testing carried out to establish the efficacy of the model show that, it consistently displays accurate outputs regardless of the size of the data being explored.

This study has shown an alternative means by which the academic performance of undergraduate students can be predicted; and comparison of this study with the use of machine learning technique based on several metrics shows that, the proposed model generalizes well and performs much better.

In the near future, we hope to extend this study to encompass the integration of mechanisms for the construction of a more robust prediction model that can interact well with the students' portal for real-time exploration in order to strengthen the institution's information system.

ACKNOWLEDGMENT

The authors would like to thank the Universiti Malaysia Pahang for funding this research through the grant: GRS130374.

REFERENCES

[1]. McKenzie, K. and R. Schweitzer, *Who succeeds at university? Factors predicting academic performance in first year Australian university students.* Higher education research and development, 2001. 20(1): p. 21-33.

[2]. Tan, M. and P. Shao, *Prediction of Student Dropout in E-Learning Program Through the Use of Machine Learning Method.* International Journal of Emerging Technologies in Learning (iJET), 2015. 10(1): p. pp. 11-17.

[3]. Engels, G., et al., *Process Modeling Using UML*. 2008.

[4]. Kantardzic, M., *DATA MINING: Concepts, Models, Methods, and Algorithms*. 2nd ed. 2011: John Wiley & Sons Inc.

[5]. Han, J., M. Kamber, and J. Pei, *DATA MINING Concepts and Techniques*. 3rd ed. 2012: Morgan Kaufman, Elsevier Inc. USA.

[6]. Negnevitsky, M., *Artificial Intelligence A Guide to Intelligent Systems*. 2011: Pearson Education Limited.

[7]. Skillicorn, D., *Understanding Complex Datasets: Data Mining with Matrix Decompositions*. 2007, USA: Taylor & Francis Group.

[8]. Li, K.F., D. Rusk, and F. Song. *Predicting Student Academic Performance*. in *Complex, Intelligent, and Software Intensive Systems (CISIS), 7th International Conference* 2013. IEEE.

[9]. Lye, C.-T., et al., *Predicting Pre-university Student's Mathematics Achievement.* Procedia-Social and Behavioral Sciences, 2010. 8: p. 299-306.

[10]. Rajaraman, A. and J.D. Ullman, *Mining of Massive Datasets*. 2012, New York, USA: Cambridge University Press.

[11]. Thammasiri, D., et al., *A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition.* Expert Systems with Applications, 2014. 41(2): p. 321-330.

[12]. Ibrahim, Z. and D. Rusli. *Predicting students" Academic Performance: Comparing Artificial Neural Network, Decision Tree and Linear Regression*. in *Proceedings of the 21ª Annual SAS Malaysia Forum, Kuala Lumpur, Malaysia*. 2007.

[13]. Bunkar, K., et al. *Data mining: Prediction for performance improvement of graduate students using classification*. in *Wireless and Optical Communications Networks WOCN; 9th International Conference* 2012. IEEE.

[14]. Oladokun, V., A. Adebanjo, and O. Charles-Owaba, *Predicting students' academic performance using artificial neural network: A case study of an engineering course.* The Pacific Journal of Science and Technology, 2008. 9(1): p. 72-79.

[15]. Sundar, P.P., *A Comparative Study For Predicting Student's Academic Performance Using Bayesian Network Classifiers.* IOSR Journal of Engineering (IOSRJEN) e-ISSN, 2013: p. 2250-3021.

[16]. Thiele, T., et al., *Predicting students' academic performance based on school and socio-demographic characteristics.* Studies in Higher Education, 2015(ahead-of-print): p. 1-23.

[17]. Arsad, P.M., N. Buniyamin, and J.-l.A. Manan. *Prediction of engineering students' academic performance using Artificial Neural Network and Linear Regression: A comparison*. in *Engineering Education (ICEED), 2013 IEEE 5th Conference* 2013. IEEE.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[18]. Chen, J.-F. and Q.H. Do, *Training Neural Networks To Predict Student Academic Performance: A Comparison of Cuckoo Search and Gravitational Search Algorithms.* International Journal of Computational Intelligence and Applications, 2014. 13(01).

[19]. Mohd Arsad, P., N. Buniyamin, and J.-l. Ab Manan. *Neural Network and Linear Regression methods for prediction of students' academic achievement*. in *Global Engineering Education Conference (EDUCON), 2014 IEEE*. 2014. IEEE.

[20]. Chattamvelli, R., *Data Mining Methods*. 2009, Oxford U.K.: Alpha Science.

[21]. Ajiboye, A.R., et al., *Evaluating the Effect of Dataset Size on Predictive Model Using Supervised Learning Technique.* International Journal of Software Engineering & Computer Science (IJSECS), 2015. 1(1).

[22]. Lykourentzou, I., et al., *Early and dynamic student achievement prediction in e-learning courses using neural networks.* Journal of the American Society for Information Science and Technology, 2009. 60(2): p. 372-380.