

Predicting Students' Academic Performances – A Learning Analytics Approach using Multiple Linear Regression

Oyerinde O. D.
University of Jos,
Computer Science Department,
PMB 2084, Bauchi Road,
Jos, Nigeria.

Chia P. A.
University of Jos,
Computer Science Department,
PMB 2084, Bauchi Road,
Jos, Nigeria

ABSTRACT

Learning Analytics is an area of Information Systems research that integrates data analytics and data mining techniques with the aim of enhancing knowledge management and learning delivery in education management. The current research proposes a framework to administer prediction of Students Academic Performance using Learning Analytics techniques. The research illustrates how this model is used effectively on secondary data collected from the Department of Computer Science, University of Jos, Nigeria. Multiple Linear Regression was used with the aid of the Statistical Package for Social Sciences (SPSS) analysis tool. Statistical Hypothesis testing was then used to validate the model with a 5% level of significance.

Keywords

Learning Analytics, Educational Data Mining, Students Academic Performance, Multiple Linear Regression.

1. INTRODUCTION

The field of learning analytics, also known as data analytics in education or educational data mining is fast gaining grounds in education management, government and industries as a result of advancement in ICT technology. The ever demanding need for knowledge and knowledge management drives the thirst for technological advancements in the aid of learning delivery [1]. With the overwhelming successes gained in Big Data analysis in the business sector, Much efforts have been taken by scholars to partake in big data era [2]. As new findings and outcomes of research crop up daily, it is evident that amongst these successes potentially identifiable, prediction of students' academic performance can have strong positive influences in knowledge management and delivery in sub-Saharan Africa thereby adding more quality to the learning experience.

Recently, researchers and developers in the education community began to explore the potentials in adopting analogous techniques for gaining insight into learning management and delivery. Two major areas currently under development as a result of this and which are oriented towards the inclusion and exploration of big data capabilities in the educational environment are Educational Data Mining (EDM) and Learning Analytics (LA) [3]. Research has shown that Data Mining techniques are widely being used in the educational field to find new and/or hidden patterns from student's data. These hidden patterns can then be used to understand some of the problems that arise in education management and delivery [4].

Predicting Students' Academic Performance (SAP) is one of the important research areas in Higher Learning Institutions. The main objectives of Prediction methods in EDM are to study the features of model that are essential for predicting SAP and to provide information about the underlying construct [5]. There have been numerous researches in constructing predictive models for various purposes in the education field [6]. The Data Mining prediction technique is used to classify the most effective factors to determine a Student's Academic Performance (SAP). These factors are applied to improve the Student's Academic Performance in a particular subjects and courses. An effective predictive model requires good input data (parameter), suitable Data Mining methods and tools for the data analysis [7].

The main aim of this research is to provide a framework for standardizing student data collections for learning providers and managers to predict students' academic performances.

2. OVERVIEW OF LITERATURE

Various data mining methods and techniques have been implemented in research cases for prediction. A comparison of machine learning methods has been carried out to predict success or failure in a course in Intelligent Tutoring Systems [8]. Other comparisons of different data mining algorithms are made to predict students final marks based on Moodle usage data [9], to predict student final grade based on features extracted from logged data [10] and to predict University students' academic performance [11].

Bayesian networks in particular have been used to predict student applicant performance [12], to model user knowledge and predict student performance within a tutoring system [13] and also to predict a future graduate's Cumulative Grade Point Average based on the students background at the time of admission [14]. Bayesian networks have also been used to model two different approaches to determine the probability a multi skill question has of being correct [15] and to predict future group performance in face-to-face collaborative learning [16]. It has also been used to predict end-of-year exam performance through student activity with online tutors [17] and to predict item response outcome [18].

Different types of neural network models have been used in prediction as well. Back-Propagation and Feed Forward Neural Networks have been used to predict final student grades [19] and to predict the number of errors a student will make [20]. Back Propagation and Counter Propagation has been used to predict performance from test scores [21]. Radial basis functions have been used to predict students' marks (pass or fail) from Moodle logs [22] while multilayer

perception topology has been used for predicting the likely performance of an applicant being considered for admission into the university [23].

Different types of rule-based systems have also been used for prediction. Fuzzy association rules have been applied to predict student performance i.e. mark prediction in an e-learning environment [24] while key formative assessment rules have been used to predict learner performance based on the learning portfolios compiled [25]. Rule induction has been used for prediction, monitoring and evaluation of student academic performance [26] while [27], used genetic algorithm to find association rules in order to predict final grades based on features extracted from logged data in an education web-based system. [28], used grammar guided genetic programming to predict student grades in Learning Management Systems while Decision Trees have been used to predict student performance and provide timely lessons in web-based eLearning systems [29]. [30], used orthogonal search-based rule extraction algorithm to predict online students' marks.

In the same vein, several regression techniques have also been used for predication. [31], used model trees, linear regression, neural networks, support vector machines and locally weighed linear regression to predict students' marks in an open university. Linear regression prediction models have been used for predicting end-of-year accountability assessment scores [32] while a multivariable regression model was used by [33] to predict student performance from log and test scores in web-based instruction. Stepwise linear regression was used for predicting student academic performance [34] while multiple linear regression was used for predicting time to be spent on a learning page [35]. [36] identified variables that could predict success in college courses using multiple regression while [37], used regression and decision trees analysis for predicting university students' satisfaction. Linear regression was used for predicting exam results in distance education courses [38], for predicting end-of-year accountability assessment scores [39] and also for predicting the probability that the student's next response is correct [40]. Logistic regression was used for predicting when a student will get a question correct and association rules to guide a search process to find transfer models to predict a student's success [41]. Robust Ridge regression algorithm was used to predict the probability of a student giving the correct answer to a problem [42] while stepwise regression was used to predict a student's test score [43].

Correlation analyses have been applied together to predict web-student performance in on-line classes [44], to predict a student's final exam score in online tutoring [45] and for predicting high school students' probabilities of success in universities [46].

Table 1. Review of Works where Data Mining Techniques have been Implemented in Educational Systems - Source [47]

S/N	Author	Mining Task	Educational System
1	Sanjeev and Zytkow (1995)	Sequence Pattern	Traditional Education
2	Zaïane et al. (1998)	Statistic and Sequence Pattern	LCM Systems
3	Beck and Woolf (2000)	Prediction	AIWBE System

4	Becker et al. (2000)	Association and Classification	Traditional Education
5	Chen et al. (2000)	Classification	Web-based Course
6	Ha et al. (2000)	Association	Web-based Course
7	Ma et al. (2000)	Association	Traditional Education
8	Tang et al. (2000)	Text Mining	AIWBE System
9	Yu et al. (2001)	Association	Web-Based Course
10	Zaïane and Luo (2001)	Sequence Pattern	LCM System
11	Luan (2002)	Clustering and Prediction	Traditional Education
12	Pahl and Donnellan (2003)	Sequence Pattern and Statistics	LCM System
13	Shen et al. (2002)	Visualization	LCM System
14	Wang (2002)	Association and Sequence Pattern	Web-Based Course
15	Merceron and Yacef (2003)	Statistic	AIWBE System
16	Minaei-Bidgoli and Punch (2003)	Classification	Web-Based Course
17	Shen et al. (2003)	Sequence Pattern and Clustering	Web-Based Course
18	Zarzo (2003)	Statistic	Web-Based Course
19	Arroyo et al. (2004)	Prediction	AIWBE System
20	Baker et al. (2004)	Classification	AIWBE System
21	Chen et al. (2004)	Text Mining	AIWBE System
22	Freyberger et al. (2004)	Association	AIWBE System
23	Hamalainen et al. (2004)	Classification	AIWBE System
24	Heiner et al. (2004)	Statistic	AIWBE System

3. RESEARCH FRAMEWORK

A suitable framework for EDM implementation was proposed with the aim of identifying the elements needed to make the prediction. The features of the framework are:

- a. Framework for guidance.
- b. Focused on students' performances.
- c. The framework would allow other projects to be replicated
- d. Comfort factor for new adopters.
- e. Flexible and can use different data.
- f. It is reliable for people with little data mining skills.

The framework is based on the Markov principle which states that, to predict the future, the past is irrelevant given the present.

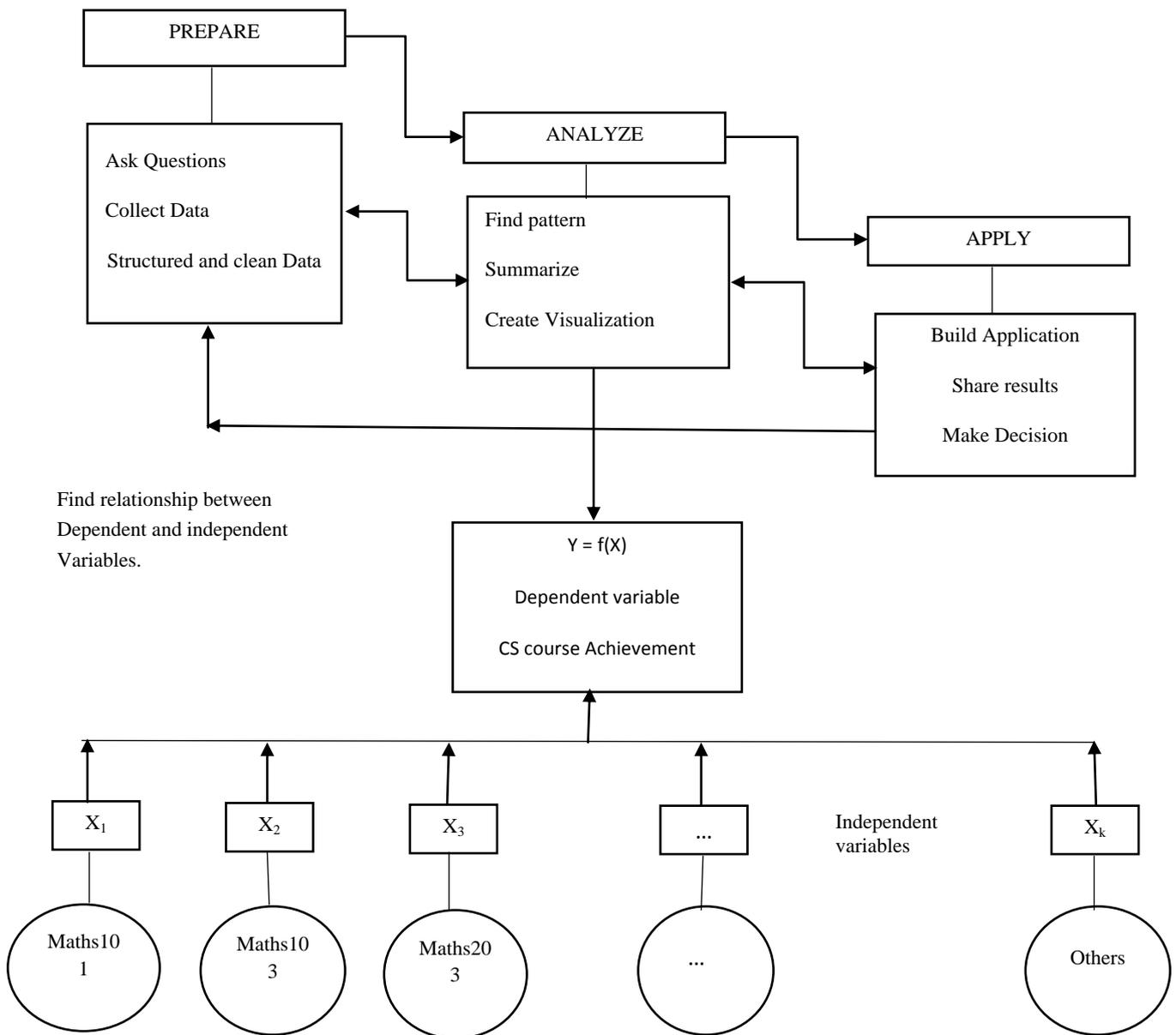


Figure 1: The research framework with extracted elements needed for prediction

4. METHODOLOGY

The objective of prediction is to estimate the unknown value of a variable that describes the student. In education the values normally predicted are performance, knowledge, score or mark. This value can be numerical/continuous value (regression task) or categorical/discrete value (classification task) [48]. Regression analysis finds the relationship between a dependent variable and one or more independent variables [49]. The type of data used in this research is secondary data. It was extracted from the computer science department, University of Jos.

This research is meant to predict students’ academic performances (SAP) using Regression analysis. In statistics, regression analysis is a statistical process for estimating the relationship between a dependent variable and one or more independent variables. Specifically, regression technique helps one to understand how typical value of the dependent variable (or ‘criterion variable’) changes when any one of the independent variables is varied, while the other independent variables are held fixed.

Multiple Linear Regression can be defined as the relationship between a dependent variable and two or more independent variables. The theoretical assumption is that, for every one-unit change in the independent variable, there is a consistent and uniform change in the dependent variable. The statistical model for multiple linear regression is given by:

$$Y = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_k X_{ik} + E$$

Where Y is Dependent variable, $X_{i1}, X_{i2}, \dots, X_{ik}$ are independent variables

β_0 = constant value and $\beta_1, \beta_2 \dots \beta_k$ = regression coefficients

In order to select the relevant variables that could predict performances effectively, a list of potential attributes were identified from discussions with lecturers, students, opinion from some educators (the Dean, the Head Of Department), and review of literature. Table 2 gives a summary of the discussions. Due to accessibility of data, MTH204, MTH205, MTH103, and CS201 were the only available data as at the

time of conducting this study. The relevant attribute chosen was total score of students in each course.

The data analysis process adopted is a step-by-step approach to making sense of data available. The process is made up of the following paradigms:

Prepare	Analyze	Apply
Ask Question	Find pattern	Model development
Collect data	Summarize data	Share results
Organize data	Create visualization	Make decision
Cleanse data		

From above figure, the forward-backward arrow between prepare and analyze indicate that one can move forward and backward during data preparation and analysis through evaluation and validation of model. Secondly, between analyze and apply again a two-way arrow. This could also allow the analyst to make corrections when the results of analysis cannot be communicated positively. One can also traverse from apply to prepare, when one discovers the data collected has not been structured and cleaned properly. However, one cannot move from prepare to apply, since "Analyze" is the "processor" of the system.

5. ANALYSIS

Patterns and relationships are found by comparing or correlating the independent variables and dependent variable. Statistical Packages for Social Scientists (SPSS) was used following the regression equation below:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

Where Y is the dependent variable = CS201

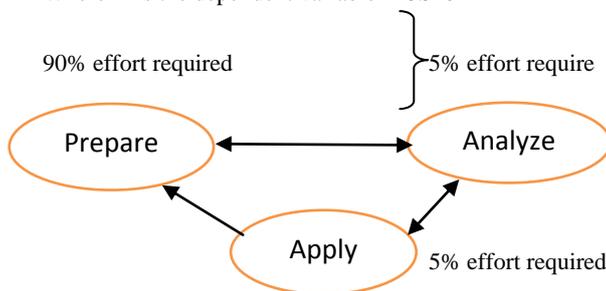


Figure 2: Workflow diagram of a customized framework.

X_1 = Student total score in Math103

X_2 = Student total score in Math204 the predictors

X_3 = Student total score in Math205

and $\beta_0, \beta_1, \beta_2, \beta_3$ are attributes associated with the variables.

Hypothesis Testing was used with the Null hypothesis, H_0 , being that there is no relationship between the predictors (Mathematics Courses) and the predicted (CS 201) variables. The Alternate hypothesis, H_1 , therefore is that to a reasonable level of significance, in this case 5%, there is a relationship between the predictors and the predicted.

H_0 : students' performance in CS201 is NOT determined by their performance in Maths103, Maths203 and Maths205

H_1 : students' performance in CS201 is determined by their performance in Maths103, Maths203 and Maths205.

With $\alpha = 0.05$ Level of significance.

H_0 : $\beta_i = 0$ for all $i = 0, 1, 2, 3 \dots$ whenever the parameters are not significant

H_1 : $\beta_i \neq 0$ parameters are significant; leads to rejection of null hypothesis.

(Where H_0 and H_1 are null and alternative hypothesis respectively and β_i the parameters used in the model.)

If the p-value is less than 5% level of significance we accept the alternate hypothesis and reject the null hypothesis; otherwise the hypothesis remain in its default state (null hypothesis). The results below were gotten from the SPSS viewer.

Model Summary^{c,d}

Model	R	R Square ^a	Adjusted R Square	Std. Error of the Estimate
1	.943 ^a	.890	.883	18.208

- a. Predictors: mth103, mth204, mth205
- b. For regression through the origin (the no-intercept model), R Square measures the proportion of the variability in the dependent variable about the origin explained by regression. This CANNOT be compared to R Square for models which include an intercept.
- c. Dependent Variable: cs201
- d. Linear Regression through the Origin

ANOVA^{c,d}

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	126018.412	3	42006.137	126.706	.000 ^a
	Residual	15581.588	47	331.523		
	Total	141600.000 ^a	50			

- a. Predictors: mth103, mth204, mth205
- b. This total sum of squares is not corrected for the constant because the constant is zero for regression through the origin.
- c. Dependent Variable: cs201
- d. Linear Regression through the Origin

Coefficients^{a,b}

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	mth204	.500	.148	.298	3.367	.002
	mth205	.537	.126	.396	4.260	.000
	mth103	.453	.149	.314	3.043	.004

- a. Dependent Variable: cs201
- b. Linear Regression through the Origin

Figure 3: Image of SPSS Multiple Regression tables

In the model summary table above, R-square is .890, meaning 89.0% of the variation from the dependent (CS201) can be explained by variation in the independent variables (Maths103, Maths204, and Maths205). The remaining 11% can be explained by other factors that are not in the model. With these results, it can be concluded that the mathematics courses have strong predictive powers, as these variables accounts for about 89.0% of the variation in CS201 performance. The Anova table above reveals that the regression equation is significant. It implies that, at least one parameter of the model is significant.

From the above coefficients table, we see that the math 204 parameter, that is, β_1 , is significant, since the p-value is 0.02, the math205 parameter, which is, β_2 , is significant since the p-value is 0.00 and the math103 parameter as well is significant since the p-value is 0.04. Given the coefficients ($\beta_1 = 0.298$, $\beta_2 = 0.396$, $\beta_3 = 0.314 \neq 0$), with the betas not equal to zero (A coefficient of 0 means that the values of the dependent variable do not consistently differ as the values of the independent variable increase.), we shall therefore reject the null hypothesis and accept the alternate hypothesis, which says;

Ha: students' performance in CS201 is determined by their performance in Maths103, Maths203 and Maths205

The ANOVA table and the table of coefficients have shown that the model is significant since all the p-values less than 0.05 indicated that the parameters for determining the performance of computer science students that sat for CS201 is significant. Therefore, the valid model is given by

$$Y = 0.298X_1 + 0.396X_2 + 0.314X_3 + E \quad (\text{where } E, \text{ the random error is usually neglected})$$

Where Y = the predicted score of student per percentage increase in math's courses

X_1 = the total score of student in math204,

X_2 = the total score of student in math205,

X_3 = is the total score of student in math103

The table below shows the variables of ten students that we shall use in the model to predict their performance in CS201.

Table 2: Five students' variable.

X ₁	X ₂	X ₃
40	8	5
19	40	50
27	54	5
41	23	16
48	8	34

Table 3: Regression results for inputted values

Y	=	$\beta_1 X_1$	+	$\beta_2 X_2$	+	$\beta_3 X_3$
16.658		11.92		3.168		1.57
37.202		5.662		15.84		15.7
31		8.046		21.384		1.57
26.35		12.218		9.108		5.024
28.148		14.304		3.168		10.676

From the above table of values, it can be concluded that for students with these scores' variables (40, 8, and 5) in mathematics, our best prediction for percentage performance in CS201 is 16.658. The same predictive patterns goes to all the other four students.

6. CONCLUSION AND FUTURE WORK

The study was able to achieve the objective of building a model to predict students' academic performance. The analysis showed that students who perform well in mathematics courses have better chances of acquiring excellence in other Computer Science courses.

However, this research was not without difficulties. These are highlighted below:

- The availability and accessibility of authentic data for analysis.
- There was no central data repository where data can be readily available.
- Inconsistencies in determining which students' attributes contributes to academic performances

Based on the limitations of the research work, the following recommendations are put forward for future research in this endeavor. First, the research is heavily dependent on the availability/accessibility of accurate and reliable data. Therefore, a research in building central data warehouse/repository could cushion that limitation. Secondly, there should be intervention programs that brings student and educators in close mediation. A face to face discussion can spark up the students' desires to do more.

The future direction of this research is to provide a standardized data collection method, based on identified metrics, parameters and indicators, and build an Educational Data Warehouse. Applying highlighted Data Mining tools and methods, Education Providers and Managers will be able to predict students' academic performances. This will assist in adequate placement in tertiary education courses, enable proper guidance of students in achieving their academic goals and objectives, and also maximize utility in the overall learning experience for both the learners and the learning providers and managers.

7. ACKNOWLEDGMENTS

We would like to acknowledge and appreciate the efforts and contributions of Prof F.O Bankole of the School of Computing, University of South Africa towards this research.

8. REFERENCES

- [1][47][48] Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert systems with applications*, 33(1), 135-146.
- [2] Baker, R. & Yacef, K. (2009). The State of Educational Data Mining in 2009: A Review and Future Visions. *Journal of Educational Data Mining (JEDM)*, 1(1), 3-17.
- [3] Papamitsiou, Z., & Economides, A. (2014). Learning Analytics and Educational Data Mining in Practice: A Systematic Literature Review of Empirical Evidence. *Educational Technology & Society*, 17 (4), 49-64.
- [4] Aziz, A. A., Ismail, N. H., & Ahmad, F. (2013). Mining Students' Academic Performance. *Journal of Theoretical & Applied Information Technology*, 53(3).
- [5] Sachin, R. B., & Vijay, M. S. (2012). A survey and future vision of data mining in educational field. In *Advanced Computing & Communication Technologies (ACCT), 2012 Second International Conference on* (pp. 96-100). IEEE.
- [6][7] Arsad, P. M., Buniyamin, N., Ab Manan, J. L., & Hamzah, N. (2011). Proposed academic students' performance prediction model: A Malaysian case study. In *Engineering Education (ICEED), 2011 3rd International Congress on* (pp. 90-94). IEEE.
- [8] Hämäläinen, W., Vinni, M. (2006). Comparison of machine learning methods for intelligent tutoring systems. In *international conference in intelligent tutoring systems, Taiwan*, 525-534.
- [9] Romero, C., Ventura, S., Hervás, C., Gonzales, P. (2008). Data mining algorithms to classify students. In *International Conference on Educational Data Mining, Montreal, Canada*, 8-17.
- [10] Minaei-bidgoli, B., Kashy, D.A., Kortmeyer, G., Punch, W.F. (2003). Predicting student performance: an application of data mining methods with an educational Web-based system. In *International Conference on Frontiers in Education*, 13-18.
- [11] Ibrahim, Z., Rusli, D. (2007). Predicting students' academic performance: comparing artificial neural network, decision tree and linear regression. In *Annual SAS Malaysia Forum, Kuala Lumpur*, 1-6.
- [12][14] Haddawy, P., Thi, N., Hien, T.N. (2007). A decision support system for evaluating international student applications. In *Frontiers In Education Conference, Milwaukee*, 1-4.
- [13] Pardos, Z., Heffernan, N., Anderson, B., Heffernan, C. (2007). The Effect of Model Granularity on Student Performance Prediction Using Bayesian Networks. In *International Conference on User Modeling, Corfu, Greece*, 435-439.
- [15] Pardos, Z., Beck, J.E., Ruiz, C., Heffernan, N. (2008). The Composition Effect: Conjunctive or Compensatory? An Analysis of Multi-Skill Math Questions in ITS. In *International Conference on Educational Data Mining, Montreal*, 147-156.

- [16] Stevens, R., Giordani, A., Cooper, M., Soller, A., Gerosa, L., Cox, C. (2005). Developing a Framework for Integrating Prior Problem Solving and Knowledge Sharing Histories of a Group to Predict Future Group Performance. In International Conference on Collaborative Computing: Networking, Applications and Work sharing, Boston, 1-9.
- [17] Ayers E., Junker B.W. (2006). Do skills combine additively to predict task difficulty in eighth grade mathematics? In AAAI Workshop on Educational Data Mining: Menlo Park, 14-20.
- [18] Desmarais, M.C., Gagnon, M., Meshkinfram, P. (2006). Bayesian Student Models Based on Item to Item Knowledge Structures In Conference on Technology Enhanced Learning, Crete, Greece, 1-10.
- [19] Gedeon, T.D., Turner, H.S. (1993). Explaining student grades predicted by a neural network. In International conference on Neural Networks, Nagoya, 609-612.
- [20] Wang, T., & Mitrovic, A. (2002). Using neural networks to predict student's performance. In *Computers in Education, 2002. Proceedings. International Conference on* (pp. 969-973). IEEE.
- [21] Fausett, L.V., Elwasif, W. (1994). Predicting performance from test scores using back propagation and counter propagation. In IEEE World Congress on Computational Intelligence, Paris, France, 3398-3402
- [22] Delgado, M., Gibaja, E., Pegalajar, M.C., Pérez, O. (2006). Predicting Students' Marks from Moodle Logs using Neural Network Models. In International Conference on Current Developments in Technology-Assisted Education, Sevilla, Spain, 586-590.
- [23] Oladokun, V.O., Adebajo, A.T., Charles-owaba, O.E. (2008). Predicting student's academic performance using artificial neural network: A case study of an engineering course. In Pacific Journal of Science and Technology, 9,1, 72-79.
- [24] Nebot, A., Castro, F., Vellido, A., Mugica, F. (2006). Identification of fuzzy models to predict students performance in an e-learning environment. In International Conference on Web-based Education, Puerto Vallarta, 74-79.
- [25] Chen, C., Chen, M., Li, Y. (2007). Mining key formative assessment rules based on learner profiles for web-based learning systems. In IEEE International Conference on Advanced Learning Technologies, Japan, 1-5.
- [26] Ogor, E.N. (2007). Student Academic Performance Monitoring and Evaluation Using Data Mining Techniques. In Electronics, Robotics and Automotive Mechanics Conference, Washington, DC, 354-359.
- [27] Shangping, D., & Ping, Z. (2008, April). A data mining algorithm in distance learning. In *Computer Supported Cooperative Work in Design, 2008. CSCWD 2008. 12th International Conference on* (pp. 1014-1017). IEEE.
- [28] Zafra, A., Ventura, S. (2009), Predicting student grades in learning management systems with multiple instance programming. In International Conference on Educational Data Mining, Cordoba, Spain, 307-314.
- [29] Chan, C.C. (2007). A Framework for Assessing Usage of Web-Based eLearning Systems. In International Conference on innovative Computing, Information and Control, Washington, DC, 147- 151.
- [30] Etchells, T.A., Nebot, A., Vellido, A., Lisboa, P.J.G., Mugica, F. (2006). Learning what is important: feature selection and rule extraction in a virtual course. In European Symposium on Artificial Neural Networks, Brussels, Belgium, 401-406.
- [31] Kotsiantis, S.B., Pintelas, P.E., (2005). Predicting Students' Marks in Hellenic Open University. In IEEE international Conference on Advanced Learning Technologies, Washington, DC, 664-668.
- [32][39] Anozie, N., Junker, B.W. (2006). Predicting end-of-year accountability assessment scores from monthly student records in an online tutoring system. In Educational Data Mining AAAI Work-shop, California, 1-6
- [33] Yu, C.H., Jannasch-pennell, A., Digangi, S., Wasson, B. (1999). Using On-line interactive statistics for evaluating Web-based instruction. In Journal of Educational Media International, 35, 157-161.
- [34] Golding, P., Donalson, O. (2006). Predicting Academic Performance. In Frontiers in Education Conference. San Diego, California, 21-26.
- [35] Arnold, A., Scheines, R., Beck, J.E., Jerome, B. (2005). Time and Attention: Students, Sessions, and Tasks. In AAAI2005 Workshop on Educational Data Mining, Pittsburgh, 62-66.
- [36] Martinez, D. (2001). Predicting Student Outcomes Using Discriminant Function Analysis.
- [37] Thomas, E. H., & Galambos, N. (2004). What satisfies students? Mining student-opinion data with regression and decision tree analysis. *Research in Higher Education*, 45(3), 251-269.
- [38] Myller, N., Suhonen, J., Sutinen, E. (2002). Using Data Mining for Improving Web-Based Course Design. In International Conference on Computers in Education, Washington, 959- 964.
- [40] Beck, J.E., Woolf, B.P. (2000). High-level student modeling with machine learning. In Fifth International Conference on Intelligent Tutoring Systems, Alagoas, Brazil, 584-593.
- [41] Freyberger, J., Heffernan, N.T., Ruiz, C. (2004). Using Association Rules to Guide a Search for Best Fitting Transfer Models of Student Learning. In Workshop Analyzing Student-Tutor Interaction Logs to Improve Educational Outcomes, Alagoas, Brazil, 1-10.
- [42] Cetintas, A., Si, L., Xin, Y.P., Hord, C. (2009). Predicting correctness of problem solving from low-level log data in intelligent tutoring systems. In International Conference on Educational Data Mining, Cordoba, Spain, 230-238.
- [43] Feng, M., Heffernan, N., Koedinger, K. (2005). Looking for sources of error in predicting student's knowledge. In: AAAI'05 workshop on Educational Data Mining, 1-8.
- [44] Wang, A.Y., Newlin, M.H. (2002). Predictors of web-based performance: the role of self-efficacy and reasons for taking an on-line class. In Computers in Human Behavior Journal, 18, 151-163.

- [45] Pritchard, D., Warnakulasooriya, R. (2005). Data from a Web-based Homework Tutor can predict Student's Final Exam Score. In World Conference on Educational Multimedia, Hypermedia and Telecommunications, Chesapeake, 2523-2529.
- [46] McDonald, B., (2004). Predicting student success. In Journal for Mathematics Teaching and Learning, 1-14.
- [49] Draper, N.R., Smith, H. (1998). Applied Regression Analysis. Wiley.