I.J. Modern Education and Computer Science, **2016, 11, 36-42**
Published Online November 2016 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijmecs.2016.11.05

Modeling and Predicting Students' Academic Performance Using Data Mining Techniques

Ahmed Mueen
King Abdulaziz University, Saudi Arabia, Jeddah
Email: mueen@kau.edu.sa

Bassam Zafar
King Abdulaziz University, Saudi Arabia, Jeddah
Email: bzafar@kau.edu.sa

Umar Manzoor
King Abdulaziz University, Saudi Arabia, Jeddah
Email: uelahi@kau.edu.sa

***Abstract*—**The main objective of this study is to apply data mining techniques to predict and analyze students' academic performance based on their academic record and forum participation. Educational Data Mining (EDM) is an emerging tool for academic intervention. The educational institutions can use EDU for extensive analysis of students' characteristics. In this study, we have collected students' data from two undergraduate courses. Three different data mining classification algorithms (Naïve Bayes, Neural Network, and Decision Tree) were used on the dataset. The prediction performance of three classifiers are measured and compared. It was observed that Naïve Bayes classifier outperforms other two classifiers by achieving overall prediction accuracy of 86%. This study will help teachers to improve student academic performance.

***Index Terms*—**Educational Data Mining, Classification, Academic performance prediction, Knowledge Discovery

I. Introduction

Recent advancement in various fields has led to the collection of large amount of data and usually the data is stored in different formats like records, files, images, sound, and videos. The collected data is used in decision making processes, however, as the amount of data is huge that makes managing / analyzing data complex and challenging. Using data for better decision making needs proper method of extracting knowledge from large repositories. Data mining techniques can be used to discover valuable and meaningful knowledge from large amount of data. Data mining is a powerful analytical tool that gives critical information and knowledge, which can help to improve decision making processes [11]. Data Mining enables researcher to make very useful discoveries from data, these discoveries are very important especially in businesses as the key decision are taken based on these discoveries. Different methods and algorithms are used in data mining to extract patterns from stored data. Xindong Wu in [25] highlighted the most influential algorithms in the field of data mining; the authors analyzed these algorithms effectiveness in different domains and also outlined future research directions.

Data mining is considered to be a new paradigm, but due to its significance in decision making, it has been successfully applied to a variety of domains including education. Recently, there are growing research interests in using data mining in education [4, 26], this new field is called Educational Data Mining (EDM). The objective of EDM is to develop new methods to explore educational data to determining the usefulness of learning systems [18], analysis learner academic performance [9], and developing an early warning system [17]. Prediction and analysis of student academic performance is vital for student academic progress [4, 21] and is a difficult / challenging task due to the influence of different factors which effect student performance such as family factor, psychological profile, previous schooling, prior academic performance, and student interaction with their classmates and teachers [2].

According to Baker [5], educational data mining algorithms differs from traditional data mining algorithms because educational data hierarchy is different from traditional data hierarchy. In recent years, researchers have proposed novel approaches for educational data mining and it is emerged as independent research area. Educational data mining current methods can be broadly classified into five categories, one of which is predication which usually deal with predicting the output value based on input data. Predication can be classified into three broad categories namely 1) classification, 2) regression and 3) density estimation. Popular classification algorithm includes support vector machine, neural network, naïve bayes, Decision Tree and the predication is either a binary or categorical variable [23]. In this research, we have used three different data mining classification algorithms (Naïve Bayes, Neural Network,

and Decision Tree) for predicting the performance of undergraduate students. For this purpose, we collected students' data from two undergraduate courses; the prediction performance of three classifiers are measured and compared. Experimental results show that Naïve Bayes classifier outperforms other two classifiers by achieving overall prediction accuracy of 86%.

The paper is organized as follows. The first section discusses the background related to data mining classification and existing research related to Predicting students' performance. This section is followed by the discussion of the proposed prediction performance model. In Section 4, a performance analysis of proposed solution on different test cases is presented. At the end conclusion is drawn and we outline some questions for future research.

II. Related Works

Various algorithms and techniques are used for knowledge extraction from educational databases, these techniques and methods in data mining required some brief to have better understanding.

A. Classification

One of the common technique used in data mining is classification. In a classification technique, model is built from pre-classified examples to assign label or class to a record. There are two parts of classification technique training part and testing part. In training, model is constructed using part of the data known as training set, which know all the attribute even the classes. After building a model, it is used to define a label or class to new record where class attribute is unknown. There are many techniques to design a model or classifier like Decision Tree (DT), Neural Network (NN), Naïve Bayes and Support Vector Machines [1]. In this study, we have used Decision Tree, Naïve Bayes, ad Neural Network. Decision Tree is very powerful data mining technique [9]. It has tree-shaped structures that comprises by nodes and branches where internal node offerings a decision based on attribute value and the branch of an internal node represents the choice made in the node, and leaf node is the end, which represents the class to be assigned. The well-known algorithms for building decision trees are ID3, C4.5 [20], and CART [7]. The main difference among these algorithms is split criterion that corresponds to an entire attribute. The most well-known split criterion are Information Gain, Gain Ratio, and Gini Index.

Equations (1) is used for the calculations of Information Gain, which is the split criterion in ID3. When an attribute *A* splits the set *S* into subsets *Si.*

$$Gain(\text{S, A}) = \text{E(S)-I(S, A)} = \text{E(S)-}\sum\frac{|S_i|}{|S|}.\text{E}\ (S_i) \qquad (1)$$

The extension of the information gain that reduces its bias towards multi-valued attributes which is used in C4.5 algorithm. This algorithms use Gain Ratio and its calculations given in (2).

$$\text{GainRatio(S,A)} = \frac{\text{Gain(S,A)}}{\text{SplitIfo(S,A)}} \qquad (2)$$

Gini Index is used in CART (Classification and Regression Trees) algorithm and the impurity is calculated by (3).

$$Gini(s) = 1 - \sum_i p_i^2 \qquad (3)$$

Neural Network algorithms inclined by the performance of the human brain. Neurons are used for information processing, which are interconnected to sense the propagation of signals. These networks of neurons become very useful for solving problems of classification and prediction [8]. Architecture of multi-layer perceptron which consists of an input layer, output layer and hidden layer is shown in "Fig. 1".

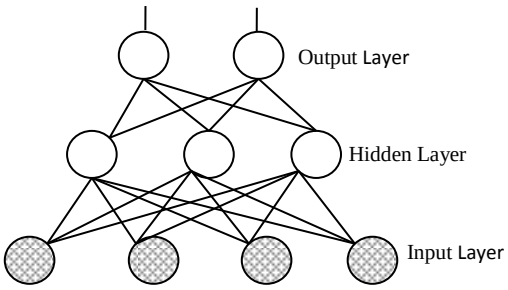

Fig.1. Artificial neural network diagram

All neurons of the hidden layer get fed from the neurons of input layers. There can be more than one hidden layers and connections between these layers are established by connecting the nodes of given layers to all neurons of next layer. The interconnection between the nodes and different layers of neurons are connected by a string of connection scalar weights which are updated during the learning process. Outputs are obtained from the output layer. Neural Network is slow but it can stand noisy data even there is no relation between variables and classes. That is why Neural Network can be used in any complex classification problems.

Whereas, Naïve Bayes technique uses probabilistic relationship between the classes and their attributes. Classifying a record depends on its attributes values that can be used as the probability of record of being from the particular class and then record is assigned to the class with largest probability [17]. The Naive Bayes classifier is based on Bayes' theorem which calculate the probability that *x* belongs to class *c* given in (4).

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)} \qquad (4)$$

P (c|x) = probability of instance x being in class c
P (x|c) = probability of generating instance x given class c
P(c) = probability of occurrence of class c
P(x) = probability of instance x occurring

B. Predicting Student's Performance

Predicting students' academic achievement is a critical part in higher learning institutions. Understanding the factors that affect student performance is a difficult research task due to many different aspects such as cultural, social, previous academic performance, interaction with teachers, etc. [21]. Several researchers have been working on these factors and they had produced promising results. For instance, some researchers investigated the impact of socio-economic status [14]. Some others studied the connection between student academic performance and their parent behaviors [3] while others looked into the efficiency of teacher to improve student academic performance [6, 13]. It is also noticed that due to Learning Management System (LMS) such as Blackboard, Moodle, WebCT etc. most of the recent research conducted on EDM has been applied to web-based education [21]. These system provide information about student assessments, activities in forums, and how many times students access teaching resources, which is very important information in predicting student performance and help teacher to detect course weaknesses [10]

III. Methodology

The method suggested in this paper to improve prediction of students' academic performance is belongs to the process of Data Mining. There are four main stages in this method, Data collection, preprocessing, classification, and interpretation (see Fig 2). Data collection is gathering all information available on students considering factors affect student performance.

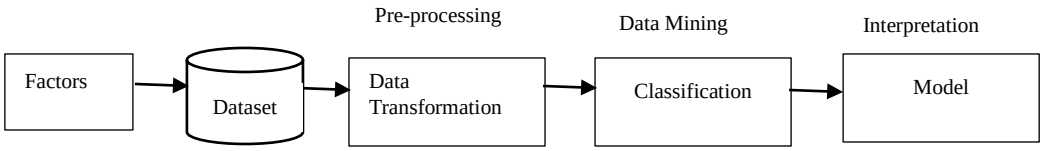

Fig.2. Method proposed for improving the prediction of students' academic performance

These information can be collected from different sources of data available and combined in to the dataset. In pre-processing stage data cleaning, attributes selection, dimensionality reduction, and data partitioning are applied to get better prediction. Whereas, in classification stage Data Mining algorithms are used for the classification of data. Normally, at this stage different Data Mining algorithms are executed with different variables and compared to select algorithm which produce best results. Finally, in interpretation stage models obtained from previous stage are analyzed to predict student performance.

A. Collecting Data

In this paper, we have collected learning information for the undergraduate students who had taken the Programming Fundamental and Advanced Operating System courses from August 2014 to May 2015. The two courses were delivered over a period of two semesters by the same lecturer. All student learning activities were collected from the LMS used in this study. Information retrieve from LMS including teaching material access duration, student academic performance information (assignments, quizzes, and tests), and discussion on forums. These influencing aspects are considered as input variables. After collection of data transformation activity was performed. This step transform data format from source data system to destination data system. In our situation data is converted from the text file into standard format required by the WEKA tool [24]. The tool we have used for Data Mining.

B. Pre-processing Data

Pre-processing data is a necessary step for preparing the dataset before applying classification techniques. It is important to note that this task directly affect the result due to the quality and reliability of available information. In this task, careful analysis of variable and their corresponding values is carried out to eliminate any abnormalities. In this study, we applied three main pre-processing tasks.

Feature selection. We thoroughly analyze our dataset to identify attributes which have greater impact on our output variable. Although, we do not have large number of attributes but still some features are not related to student performance.

Weka provides several feature selection algorithms. We have used ranking algorithm to select appropriate attributes.

Imbalanced data. Data is imbalanced when number of instances in one class is much smaller than the number of instances in other class. Therefore, during the training stage classifier take more sample from the classes which have bigger number of instances. Due to this, at test stage classifiers are less sensitive to the classes which have smaller number of instances. There is wide range of data balancing algorithms available in Weka. We have used SMOTE to solve data imbalanced problem.

Data transformation. This pre-processing tasks is to integrate the data obtained from different sources into one single dataset. Then convert the format of the source data file into the format of destination data file. We have converted our data file into .ARFF format of Weka.

C. Data Mining and Prediction model

To predict student academic performance we have used three well know classification techniques: decision trees, artificial neural networks, and Naïve Bayes. We have selected these classification technique based on their reputation in newly published data mining literature and their superiority in prediction type problems. In DT a tree is constructed by recursively separating observations into branches to achieve highest prediction accuracy [20]. To construct tree different mathematical algorithms are used. The benefit of DT is that the created rules are effortlessly detected and interpreted because of its tree-like format, which decrease the chance of errors arising in a problem. In this study, we have used C4.5 algorithm, which is highly rank algorithm in data mining research [25]. Whereas, Neural Networks (NN) have the outstanding capability to develop meaning from complex data. Multi-Layer Perception (MLP) is the most famous NN architecture learning network model used for academic classification objectives. We used MLP for this study with back-propagation type supervised-learning algorithm. MLP is shown to be a robust function estimator for prediction problems [16]. Naïve Bayes is one of the simplest density approximation approach from which a classification method can be constructed. A Naïve Bayes classifier classifies based on prior knowledge. Naïve Bayes classifier incorporate independence expectations which do not actually work in the real world, but still many complex problem have successfully been solved using this classifier [15].

IV. Experimental Setup

During experiment all the classification algorithms were executed using 10-fold cross validation to train the model. Each dataset was divided into ten corresponding subsets, nine were used for training the model, and one subset was used for model testing. The data set used was composed of records of students enrolled from August 2014 to May 2015.

Furthermore, Blackboard platform was used to provide on-line resources and discussion forum. Students were required to submit their assignment and do quizzes through Blackboard. Students were given instruction to use discussion forum for their queries about subject material, theory, and exercise. They were allow to discuss with the instructor or with other students. When course ended students were needed to sit for written exam. There were 60 students: 41 passed (68.33%) and 19 failed (31.66%). Each student record contains several input variable as shown in Table I.

Table 1. Variables used in this study

Source	Variable
General	Age, section, number of students in section, type of program, hours spent studying daily, methods of study used, city of birth, transport method, distance to the college, subjects interest, motivation level, difficulty doing homework, facilities in college, having home tuition, level of father education, level of mother education, attendance.
Forum	forum login time, logout time, forum join rate, forum reply, write messages, read messages, total number of words write, time spent
Academic	Grade Point Average (GPA), quiz1, quiz2, quiz average, Assignment submit, Assignment delay, labtest1, labtest2, labtest average, final exam grade, total time spent.

We carried out different experiments to get our objectives. Our first objective was to predict student academic performance. The second objective was to reduce number of attributes. And the last objective is to compare classification accuracy of different classifiers. The data mining tool we have used is Weka 3.6 open-source data mining software (www.cs.waikato.ac.nz/ml/weka).

The performance of classification model are measured by evaluating the correctness of the classification decision of the classifier. The table shows these counts are commonly known as confusion matrix. The terms used in confusion matrix are:

(TP): Number of True Positives (Classifier correctly labeled record as positive).
(TN): Number of True Negatives (Classifier correctly labeled record as negative).
(FP): Number of False Positives (Classifier incorrectly labeled record as positive).
(FN): Number of False Negatives (Classifier incorrectly labeled record as negative).

To evaluate and compare classifier performance we used accuracy, precision, recall, and specificity. Using confusion matrix they are measured with (5), (6), (7), and (8).

Accuracy (proportion of total number of correct prediction):

$$\frac{\mathrm{TP+TN}}{\mathrm{P+N}} \tag{5}$$

Precision (proportion of correct positive observations):

$$\frac{TP}{TP+FP} \tag{6}$$

Recall (proportion of positives correctly predicted as positive):

$$\frac{TP}{P} \tag{7}$$

Specificity (proportion of negatives correctly predicted as negative):

$$\frac{TN}{N} \quad (8)$$

V. Result

We compared, tested, and analyzed dataset with three classifiers. Those classifiers are Naïve Bayes, Multilayer Perception and C4.5 (J48). All three classifier were tested on all 38 available attributes. We used tenfold cross validation that means dataset was randomly divided into 10 subsets of same size. Table II, shows the result of our first experiment using all the attributes with an average of 10 executions.

Table 2. Classifier comparison result using all attributes

Classifier	Accuracy	Precision	Recall	Specificity
Naïve Bayes	86.0%	88.4	85.8	86.3
Multilayer Perception	82.7%	82.5	86.3	79.1
C4.5	79.2%	81.4	78.0	80.2

Accuracy rate which represent the effectiveness of the classifier shows Naïve Bayes performs better than other two. Naïve Bayes is also the winner in precision which shows the predicative power. According to recall which represents the sensitivity, Multilayer Perception performs better. In specificity again Naïve Bayes outperforms others. In second experiment, we applied feature selection algorithms then perform ranking process in which each algorithm select a list of attributes. Finally, an attribute selected by more algorithms considered best attribute. In our case, we selected attributes which has frequency more than three. Table III, shows our best seven attributes.

Table 3. Best seven attributes and descriptions

Attributes	Descriptions
GPA	Student average of Grade Points obtained in all the subjects previously
TestAvg	Two test average
Assignmentsub	Did student submit assignment or not
ParticipationRate	Reply discussion messages /total discussion messages)
Attendance	Good or bad attendance
LabTestAvg	Two lab test average
FinalGrade	Final exam marks

Again three classifiers are executed on this reduced dataset using 10 fold cross-validation. The result of this reduced dataset can be seen in Table IV. Naïve Bayes is still the best classifier even in the reduced dataset according to accuracy, precision, and specificity. But recall showed Multilayer perception perform superior than other two. Furthermore, if we compare table 2 and table 4, we can see slight improvement and decrease in values, which shows the results in general are very similar.

Table 4. Classifier comparison using best attributes

Classifier	Accuracy	Precision	Recall	Specificity
Naïve Bayes	85.7.0%	89.3	84.6	89.1
Multilayer Perception	81.4%	84.0	86.5	83.2
C4.5	80.5%	80.9	79.3	78.4

Additionally, we have analyzed our dataset to identify factors which cause student to loss his academic status due to academic performance. We have found that poor performance of student was due to lack of participation in on-line discussion forum. Students who were not discussing in the forum with other students or with instructor perform poor and loss academic status. In contrast, those students who pass the course were very active in discussion forum with the instructor and classmates.

VI. Conclusion

This work is an effort to use Data Mining techniques to predict and analyze students" academic performance. Three techniques Decision tree (C4.5), Multilayer Perception, and Naïve Bayes were used. All these techniques were applied on student's data collected from under graduate courses conducted in duration of two semesters. In this study, three classification models were built to predict student academic performance. Results shows that Naïve Bayes classifier outperforms other two classifier by obtaining the overall prediction accuracy of 86%. This research assist teachers to early detect student who is expected to fail the course. Instructor can provide special attention to those student and help them to enhance their academic performance. There are number of studies conducted in this regards identifying different factors such as student personal factor, family factor, or instructor factor. Many factors or combination of different factors effect student performance and also it vary from one country to other country, one institution to other institution, one culture to other and one group of students to other group of students. We feel instructor role is important is this regards. He has to be more interactive with student, provide proper guidance and motivate the student. In our case, we have not allocate any marks to use forum that is why students were not interested to use forum. We suggest instructor should motivate student to use forum or allocate some marks for the usage of forum. More appropriately percentage of marks should be given according to the messages student posted on the forum.

Finally, for further research we like to carry out more experiments with bigger dataset including different courses and different educational levels. We would also like to develop a automate system to analyze all the factor automatically.

REFERENCES

[1] Alkhasawneh and R. Hobson, "Modeling student retention in science and engineering disciplines using neural networks," in *Proc. IEEE Global Eng. Edu. Conf. (EDUCON)*, pp. 660–663, Apr. 2011.

[2] Araque, F., Roldan, C., & Salguero, A. "Factors influencing university dropout rates," Journal of Computer & Education, 53, pp.563–574, 2009.

[3] Attaway, N. M., & Bry, B. H. "Parenting style and black adolescents' academic achievement," Journal of Black Psychology, 30, pp.229–247, 2004.

[4] Baker, R.S., Corbett, A.T., Koedinger, K.R. "Detecting Student Misuse of Intelligent Tutoring Systems," Proceedings of the 7th International Conference on Intelligent Tutoring Systems, pp.531-540, 2004.

[5] Baker, R.S.J.d. "Data Mining for Education". In McGaw, B., Peterson, P., Baker, E. (Eds.) *International Encyclopedia of Education (3rd edition)*, vol. 7, pp. 112-118, 2010.

[6] Bassam Zafar, Ahmed Mueen, Mohammad Awedh, Mohammad Balubaid, "Game-based learning with native language hint and their effects on student academic performance in a Saudi Arabia community college" Computer in Education vol. 1, no. 4, pp. 371-384, 2014.

[7] Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. "Classification and Regression Trees," Belmont, CA: Wadsworth International Group 1984.

[8] Calvo-Flores, M. D., Galindo, E. G., Jimenez, M. P., & Pineiro, O. P. "Predicting students' marks from Moodle logs using neural network models," Current Developments in Technology-Assisted Education, 1, pp. 586–590, 2006.

[9] Charu, C. A. "An Introduction to Data Classification" Data Classification, Chapman and Hall/CRC, pp. 1-36, 2014.

[10] Cristóbal Romero, Manuel-Ignacio López, Jose-María Luna, Sebastián Ventura. "Predicting students' final performance from participation in on-line discussion forums", Journal of computer and Education vol. 68, pp.458-472, 2013.

[11] Essa, A., & Ayad, H. "Student success system: Risk an alytics and data visualization using ensembles of predictive models," Paper presented at the 2nd international conference on learning analytics and knowledge, Vancouver 2012.

[12] Fayyad.U, Piatesky, G. Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," AAAI Press , Massachusetts Institute Of Technology The MIT Press 1996,

[13] Gerber, S. B., & Fin, J. D. "Teacher aides and students" academic achievement Educational Evaluation and Policy Analysis, 23(2), pp.123–143, 2001.

[14] Goddard, R. D., Sweetland, S. R., & Hoy, W. K. "Academic emphasis of urban elementary schools and student achievement in reading and mathematics: A multilevel analysis," Educational Administration Quarterly, 36(5), pp.683–702, 2000,

[15] Hand D.j., & Yu, K. "Idiot's Bayes- not so stupid after all?" *International Statistical Review*, 69(3), pp.385- 399, 2001.

[16] Hornik, K., Stinchcombe, M., & White, H. (1990). "Universal approximation of an unknown mapping and its derivatives using multilayer feedforward network,"Neural Networks, 3, pp.359–366, 1990.

[17] Hongbo, D., Yizhou, S., Yi, C., & Jiawei, H. "Probabilistic Models for Classification," Data Classification, Chapman and Hall/CRC, (pp. 65-86), 2014.

[18] Kotsiantis, S. B. "Use of machine learning techniques for educational proposes: A decision support system for forecasting students grades," Artificial Intelligence Review, 37(4), pp.331–344, 2012,.

[19] Mostow, J., Beck, J., Cuneo, A., Gouvea, E., & Heiner, C. "A Generic Tool to Browse Tutor-Student Interactions: Time Will Tell," Proceedings of the 12th International Conference on Artificial Intelligence in Education (AIED 2005), pp. 884-886, 2005 Amsterdam,.

[20] Quinlan, J. R. "*C4.5:* programs for machine Learning," Morgan Kaufmann Publishers Inc, 1993.

[21] Romero, C., & Ventura, S."Data mining in Education," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 3(1), pp.12–27, 2013.

[22] Tang, T., McCalla, G. "Smart recommendation for an evolving e-learning system: architecture and experiment", International Journal on E-Learning, vol. 4, issue1, pp.105–129, 2005.

[23] U Manzoor, S Nefti "An agent based system for activity monitoring on network-ABSAMN" Expert Systems with Applications 36 (8), pp. 10987-10994, 2009.

[24] Weka, (2015). Retrieved from http://www.cs.waikato.ac.nz/ml/weka/.

[25] [25] Xindong Wu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus F. M. Ng, Bing Liu, Philip S. Yu, Zhi-Hua Zhou, Michael Steinbach, David J. Hand, Dan Steinberg "Top 10 algorithms in data mining," Knowl. Inf. Syst. 14(1), pp.1-37, 2008.

[26] Zaïane, O. "Building a recommender agent for e- learning systems". Proceedings of the International Conference on Computers in Education, pp.55–59, 2002.

Authors' Profiles

Dr. Ahmed Mueen is an assistant professor in King Abdulaziz University Jeddah. His research interests include Machine learning, Data Mining, image classification, Game-based learning, information retrieval, and image processing**.**

Dr. Bassam Zafar received his BS in Electronic Engineering and Communication, his MS in Information Technology and his PhD in Computer Science from De Montfort University, Leicester, UK, Manchester, UK, in 2003, 2004 and 2008 respectively. He is currently working as an Associate Professor at the Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudia Arabia.

Dr. Umar Manzoor received his BS in Computer Science and his MS in Computer Science from the National University of Computer and Emerging Sciences, and his PhD in Multi-Agent Systems from the University of Salford, Manchester, UK, in 2003, 2005 and 2011, respectively. In 2006, he joined the National University of Computer and Emerging Sciences, Islamabad, Pakistan, as a Lecturer and promoted after as an Assistant Professor. In 2012, he was promoted as an Associate Professor; currently, he is working at King Abdulaziz University, Jeddah, Saudi Arabia. He has published extensively in the area of multi-agent systems, autonomous systems, behaviour monitoring and network management/monitoring.

How to cite this paper: Ahmed Mueen, Bassam Zafar, Umar Manzoor,"Modeling and Predicting Students' Academic Performance Using Data Mining Techniques", International Journal of Modern Education and Computer Science(IJMECS), Vol.8, No.11, pp.36-42, 2016.DOI: 10.5815/ijmecs.2016.11.05

