Automated Student Advisory using Machine Learning

Walid Mohamed Aly
College of Computing and Information Technology
Arab Academy for Science, Technology & Maritime Transport
Alexandria, Egypt

Osama Fathy Hegazy
Computer Science Department
Cairo Higher Institute for Engineering, Computer Science and Management
Cairo, Egypt

Heba Mohmmed Nagy Rashad
Computer Science Department
Cairo Higher Institute for Engineering, Computer Science and Management
Cairo, Egypt

ABSTRACT
Educational data mining is a specific data mining field applied to data originating from educational environments, it relies on different approaches to discover hidden knowledge from the available data. Among these approaches are machine learning techniques which are used to build a system that acquires hidden knowledge from previous data. Machine learning can be applied to solve different regression, classification, clustering and optimization problems. In our research, we propose a "Student Advisory Framework" that utilizes classification and clustering. This system can be used to guide the first year university students to the more suitable educational track. The classification phase will predict the department which is most likely to be chosen by a student and the clustering phase will recommend a department to student by showing his expected rate of success for each department, this recommendation aims to decrease the high rate of academic failure for first year students. Our approach is tested using a real case study from "Cairo Higher Institute for Engineering, Computer Science, and Management" using data collected for a period within 12 years from 2000 – 2012.

Keywords
Classification, Clustering, Educational Data Mining (EDM), Machine Learning,Higher Education system

1. INTRODUCTION
Within recent few years, the number of educational institutes that adopted an information system has been growing very quickly; consecutively the amount of data available in each educational institute database has also increased. Educational data mining is intuitively applied to discover hidden information from this data that would improve the quality of the whole educational system. Educational data mining can be applied to discover patterns in data sets to automate the decision making process for learners, students and administrators.

Educational data mining methods [1] belong to a diversity of literatures. These literatures include-among others- machine learning, information visualization and computational modeling.

Machine learning approaches include neural networks, naive Bayesian, K-nearest neighborhood, decision tree, support vector machine (SVM), linear regression, and rule induction. All these techniques can be used to discover association rules, classification, clusters, and outliers within educational data sets.

This paper uses machine learning techniques (classification – clustering) to develop an intelligent student advisory framework. This framework improves the student's performance and the quality of the education by reducing the failure rate of first year students. One of the main reasons for this high failure rate is the incorrect selection of the student's department.

The proposed framework captures information from the data sets which stores the academic achievements of current students before enrolling to higher education, the dataset also include students' first year grade after enrolling in a certain department. After acquiring all the relevant information, a new student can utilize the intelligent system to receive commendation of a certain department in which he/she would likely succeed. The framework also predicts the department which is most likely to be chosen by student.

The remaining parts of this paper are organized as follows; Section 2 presents samples of related works in educational data mining, section 3 presents the basic information of machine learning with a special concern on the algorithms used in paper. The proposed intelligent framework for a student advisory system is introduced in section 4. The case study that is used to demonstrate the proposed framework is presented in section 5. Finally, the last section gives a concluding remark.

2. RELATED WORK
Many researchers have contributed to the field of data mining in higher education. In this section, the researchers will give an overview on a few representative works.

Abu Tair and El-Halees gave a comprehensive case study from the higher education stage [1]. The main purpose of their study is to show how useful data mining can be in the educational domain, their research discovered many kinds of knowledge from the graduate student dataset using different educational data mining techniques, this acquired knowledge included classification, clusters, association rule, and outlier detection.

M. Sukanya, S. Biruntha, S. Karthik and T. Kalaikumaran applied the Bayesian classification technique on existing higher education students [2]. The main goal of their study is to predict the number of upcoming students in the next year based on the number of enrolled students in the previous years. This study helps decision makers to manage the resources and staff they need to administer the outcomes of a student. This study helps also the teachers to identify at early

stage the students that need more attention to facilitate taking the correct action at the suitable time to reduce the failure in the academic approach and improve the student's academic performance.

Md. Hedayetul Islam Shovon and MahfuzaHaque[3] implemented k-means cluster algorithm. The main goal of their study is to improve the quality of the education by dividing the students into groups according to their characteristics using the application which have implemented.

Er. Rimmy Chuchra, M. tech [4] gave a case study from the higher education university. Their study was based on applying neural networks on the existing student dataset from the university database to build decision trees that can be used to evaluate the performance of students.

Brijesh Kumar Bhardwaj and Saurabh Pal [5] applied Baysian classification on the student database from the higher education stage. This study aimed at identifying those students who needed more attention to reduce the drop out ratio and take action at a right time.

Md. Hedayetul Islam Shovon and Mahfuza Hague [6] applied a hybrid procedure that was based on decision trees and data clustering. The main goal of their study is to predict the GPA, this kind of knowledge reduces the drop out ratio and improve the performance of the students.

3. ADOPTED MACHINE LEARNING TECHNIQUES

Machine learning aims at building an intelligent system which will be intelligent enough to determine a decision or calculate output based on new inputs after passing the learning phase and being fed with a set of training data.

According to the definition of Tom Mitchell [7]: "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E".

Learning can be a supervised learning where the correct output in the training set is made available. Supervised learning is used to solve regression or classification problems. Examples of classification problems include identifying an email as a spam, face recognition and hand writing recognition. While regression problems include building a model for a system that can be used to predict the output value of the system for a given input.

Another type of learning is unsupervised learning where the exact output is unknown. This type of learning is used typically to solve clustering problems. Semi-supervised learning stands between both supervised and unsupervised learning as it uses both labeled and unlabeled data during learning.

Two of machine learning techniques are described in the next section.

3.1 C4.5

C4.5[8] is a supervised learning algorithm for producing decision trees , it was proposed by Ross Quinlan as an extension of the basic "Iterative Dichotomiser 3" (ID3) algorithm. C4.5 is considered a statistical classifier as it can deal with both continuous and discrete attributed and data set with missing attributes values.

C4.5 determines the successive nodes in the decision tree based on the Kullback–Leibler divergence(KLD) as the splitting criterion, Kullback–Leibler divergence of A from B is a figure of merit measuring the information gained when B is used to find an approximate value of A, that is why KLD is also known as information gain.

The information gain is generally calculated using the following equation:

$$IG(T,a) = H(T) - H(T|a) \qquad (1)$$

Where
IG(T,a): information gain of parameter a when calculating T

H: information entropy.

Base cases can be identified when none of the attributes offers any information gain, or only one class exists and all the data points are already labeled with this class

The standard C4.5 algorithm is as follows:

1. Read set S of examples described by continuous or discrete attributes.
2. Identify base cases.
3. Find the attribute which has the highest informational gain (Abest).
4. Divide S into S1, S2, S3... according to the values of Abest.
5. Repeat the steps for S1, S2, and S3 etc. . . .

CART (Classification and Regression Trees) is much similar to C4.5, however, it can be used to solve regression problems by using numerical variables as target, CART does not also compute rule sets.

3.2 k-Means Clustering

K-means clustering [9] is an unsupervised learning algorithm. It is one of the partitioning clustering procedures. It is dependent on distance-based that split "n" data set into the specific predetermined number of clusters in which each cluster is associated with a centroid.

The following is the pseducode for applying the k-means algorithm:

1. Select K points as the initial Centroids.
2. Repeat
   a. Form K clusters by assigning data points to nearest centroid, the distance between the data point and the centroid is calculated as Euclidean distance, which calculates the distance between two artesian points q and p in n dimensional space by the following equation:-

$$d(q,p) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \qquad (2)$$

   Where $\mathbf{p} = (p_1, p_2, ..., p_n)$ and $\mathbf{q} = (q_1, q_2, ..., q_n)$
   b. Calculate the new mean of each cluster
   c. Update the Centroid of each cluster to be equal to the mean of the cluster
3. Until the Centroids do not change or a maximum iteration number is reached.

Although there is no guarantee that the optimum clustering will be reached, the K-means will converge to a local

optimum point in a reasonable amount of time. The clustering problem can be solved by k means in time $O(n^{dk+1} \log n)$, where *n* is the number of entities to be clustered, d is the dimension of problem space and k is number of clusters.

Choosing the value of K affects the efficiency of clustering, one of the heuristics used to calculate an efficient value of K is to use the elbow method [8], the elbow method choses the minimum value of K that offers a reasonable percentage of variance explained, this percentage should have a minor increase with higher value of K.

Another simpler heuristic is to choose k based on the number of n data points to be in the range of $\sqrt{\frac{n}{2}}$.

The following figure represents a flowchart of the standard K-Means Algorithm

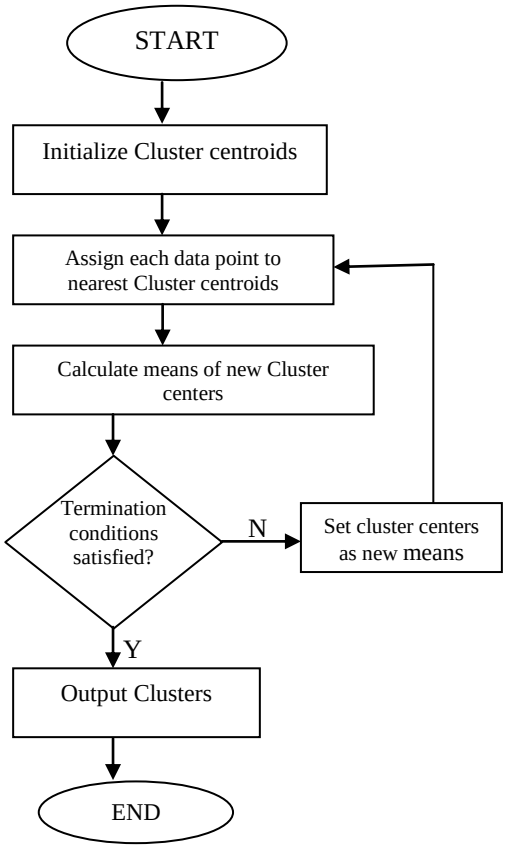

Fig 1: K-means Clustering Flowchart

4. INTELLIGENT FRAMEWORK FOR A STUDENT ADVISORY SYSTEM

4.1 The Framework Description

The proposed framework apply classification to predict the department that the new student will be willing to enroll, the framework also apply clustering to recommend the students enrollment to a certain department. These two machine learning approaches (Classification and clustering) acquire hidden knowledge by learning from an educational data set that includes attributes representing:

- Student academic grades before entering college.
- Department chosen by student.
- Student grade in first year.

Data preprocessing is applied for normalization and feature selection followed by the learning phase where classification and clustering rules are learned, after that the framework can be running and capable of receiving new students records.

The main processes of the frame work are explained within the following points.

4.2 Learning to Predict Using Classification

In this process, a classification algorithm is applied on the educational data set to build an efficient classifier. The role of the classifier is to predict the department which the student is likely to choose for enrollment. The steps in this phase are as follows:

1. Use the training data set, and apply different decision tree classification algorithms (ID3, C4.5, and CART) with the Department attribute as the target class.
2. Record the set of rules for the classification algorithm with highest performance, different measure to evaluate performance are mentioned in section 5.2.

The holdout validation method was adopted to avoid over fitting of the classifier, 70 % of the dataset were used for training the classifier, and the rest of the dataset (30%) were used to test the classifier performance.

4.3 Learning to Recommend Using Clustering

In this process, a clustering algorithm is applied on the educational data set to divide student records into a number of clusters based on marks' similarity. The steps in this phase are as follows:

1. Apply k-means clustering algorithm on student records. The department and the first year grade will not be affecting how clusters are created.
2. Measure the rate of success in each cluster for all departments.

4.4 Request an Output from the System

A user can ask the system to acquire a prediction and a recommendation for a certain educational department. The steps of this phase can be summarized as follows:

1) The new student will enter his/her data.
2) The purposed system will read the data and validate its soundness.
3) Predict the department according to rules declared by classification phase.
4) Identify the cluster to which the student belongs
5) Recommend a department for student as the department with the highest the rate of success in student cluster.

Figure 2 shows a block diagram for the proposed framework.

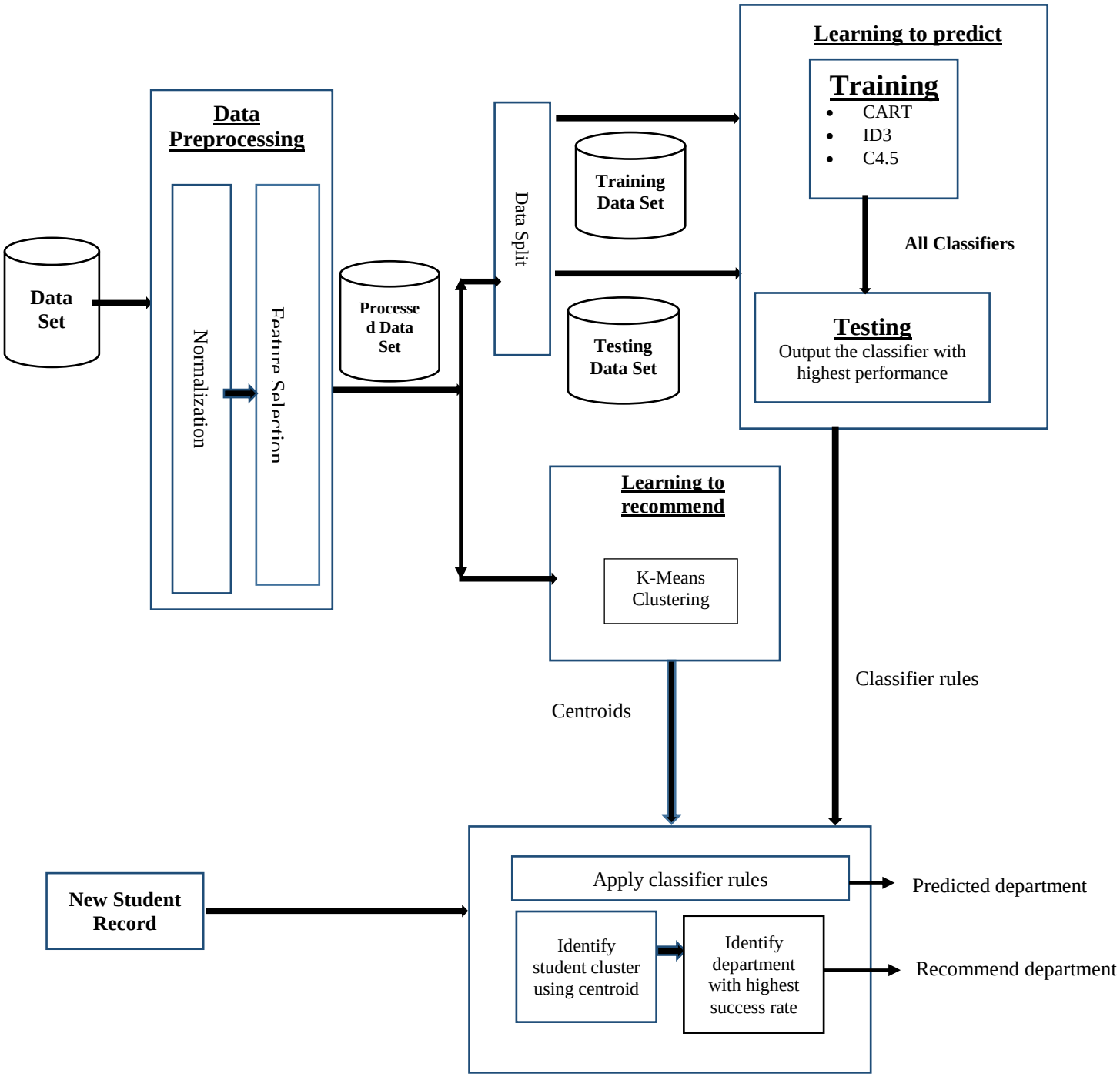

Fig 2: Framework Block Diagram

5. CASE STUDY

The Student Data used in this case study is obtained from "Cairo Higher Institute for Engineering, Computer Science, and Management" (CHI) which is located in Cairo, Egypt. CHI has four departments:

- Management Information System (MIS)
- Computer Science (CS)
- Architecture Engineering (AE)
- Computer Engineering (CE)

5.1 Data Set

The student data is collected from CHI during the period from 2000 to 2012 to form a data set known as (CHISDS). CHISDS includes 1866 records, each record has 21 attributes.

Not all the attributes will be used in the data mining process, some of the attributes in the data set such as the Student ID, Student Name, Address, or Home Phone Number present personal information that do not expand any knowledge for the data set under processing. The selected attributes are shown in Table I.

Table 1. Dataset Metadata

Attribute	Data Type	Range
Secondary Stage Type	Discrete	9 values (SSA1,SSA2,...SSA9)
Total Marks	Continues	0-420
High school English Marks	Continues	0-50
High school Math Marks	Continues	0-100
High school Physics Marks	Discrete	0-50
First Year Grade	Discrete	8values (A,B+,B,,C+,C,D+,D,F)
Department	Discrete	4 values (AE,CE,CS,MIS)

5.2 Results for Predicition Using Classification

A number of decision trees classification algorithms (C4.5, CART and ID3) were used individually to build an efficient classifier. The C4.5 proved to be the most efficient and robust.

A typical way to demonstrate the efficiency of a classifier is by using the confusion matrix which is shown in table 2.

Table 2. Confusion Matrix

Predicted	Actual Positive	Actual Negative
predicted positive	TP	FN
predicted negative	FP	TN

The confusion matrix has four categories [10]:

1. True positives (TP) are examples correctly labeled as positive.
2. False positives (FP) refer to negative examples incorrectly labeled as positive.
3. True negatives (TN) correspond to negative examples correctly labeled as negative.
4. Finally, false negatives (FN) refer to positive examples incorrectly labeled as negative.

The confusion matrix is used to evaluate the performance of the proposed technique by calculating precision, recall, error rate, f-measure and accuracy, the equations for calculating these performance indices are as follows[11]:-

$$\textbf{TP} = \frac{\text{TP}}{\text{TP+FN}} \quad (3)$$

$$\textbf{FP} = \frac{\text{FP}}{\text{FP+TN}} \quad (4)$$

$$\textbf{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (5)$$

$$\textbf{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (6)$$

$$\textbf{F-measure} = \frac{2*(\text{precision}*\text{recall})}{\text{precision}+\text{recall}} \quad (7)$$

$$\textbf{Accuracy} = \frac{(\text{TP+TN})}{(\text{TP+FP+TN+FN})} \quad (8)$$

$$\textbf{Error Rate} = 1 - \text{Accuracy} \quad (9)$$

Table 3 shows the average of F-measure percentage, Accuracy, and Error Rate for different classification algorithm used.

Table 3. Performance Evaluation of Classifiers

Classificatin Technique	F-Measure	Accuracy(AC)	Error Rate
C4.5	0.9607	0.98028	0.01972
CART	0.94869	0.97371	0.02629
ID3	0.86227	0.93127	0.06873

Applying the C4.5 algorithm as classifier resulted on the most efficient classification of recommended department. The F-measure for C4.5 classification for each department is as shown in Table 4.

Table 4. F-Measure for C4.5 Classifier

Classified Department	F-Measure
MIS	0.95015
CS	0.98660
AE	0.94902
CE	0.95742

5.3 Results for Recommendation Using Clustering

Applying the K-means algorithm on the available data set-with a maximum number of 10 iterations-resulted on having four different clusters. The evaluated Centroids are shown in table 5.

Evaluation of clustering efficiency is measured by the "within cluster sum of squares" (WCSS) and was equal to 2593.2097. WCSS is calculated for K clusters in a p dimension space using the following equation:-

$$WCSS = \sum_{k=1}^{K} \sum_{i \in C_k} \sum_{j=1}^{p} (x_{ij} - \bar{x}_{kj})^2 \quad (10)$$

Where C_K is the set of observations in the *k*th cluster and $\bar{x}_k$ is the centroid of the k cluster.

Table 5. Centroids of each Cluster

Attribute	Cluster#1	Cluster#2	Cluster#3	Cluster#4
Total Marks	346	352	335	268
English Marks	30	31	41	30
Math Marks	75	82	80	53
Physics Marks	23	35	33	28

The rate of success in each cluster is calculated and used to recommend the department for the student where he will have higher chance of success. Table 6 shows the ratio of successes in each department distribution over these three clusters. N/A is shown if no students in this cluster enrolled to a certain department

Table 6. Success Ratio

	Cluster#1	**Cluster#2**	**Cluster#3**	**Cluster#4**
MIS	83.33%	75%	70.3%	56.65%
CS	76.33%	71.43%	64.71%	N/A
AE	82.65%	81.96%	64.98%	3.22%
CE	N/A	83.07%	61%	0%

The Results show the existence of different ratio of successes in each department in each cluster. The rate of success for the predicted department from the classification phase is of the most importance as a wrong decision from student might increase his fail possibility.

6. CONCLUSION

This paper proposes an "Automated Student Advisory framework" to improve the students and institutes educational performance.

In this work, we used C4.5 algorithm to predict the likely department for the first year university students. We used also, k-mean cluster algorithm to divide the students into number of clusters and determine the rate of success in each cluster for each department. The rate of success in each cluster is calculated and used to recommend the department for the student where he will have higher chance of success.

The framework was tested on a real case study from Cairo Higher Institute for, Engineering, Computer Science and Management. Results demonstrated the usage and efficiency of the proposed framework.

Future work might include improving the framework by removing outliers and applying the filter or wrapper technique for accurate feature selection.

7. REFERENCES

[1] A.M. El-Halees, and M.M. Abu Tair, "Mining educational data to improve students' performance: A case study," International Journal of Information and Communication Technology Research, 2011, pp. 140-146.

[2] S.Karthik M.Sukanya, S.Biruntha and T.Kalaikumaran, "Mining Data mining: Performance improvement in education sector using classification and clustering algorithm," ICCCE In Proceedings of the International Conference on Computing and Control Engineering, 2012.

[3] Mahfuza Haque Md. Hedayetul Islam Shovon. "Prediction of student academic performance by an application of k- means clustering algorithm". International Journal of Advanced Research in Computer Science and Software Engineering, 2(7):353–355, July2012.

[4] M. tech Er. Rimmy Chuchra. "Use of data mining techniques for the evaluation of student performance: a case study". International Journal of Computer Science and Management Research, 1(3):425–433, October 2012.

[5] Brijesh Kumar Bhardwaj and Saurabh Pal. Data mining:" A prediction for performance improvement using classification". (IJCSIS) International Journal of Computer Science and Information Security, 9(4), April,2011.

[6] Md. Hedayetul Islam Shovon and Mahfuza Haque." An approach of improving student's academic performance by using k-means clustering algorithm and decision tree". (IJACSA) International Journal of Advanced Computer Science and Applications, 3(8):146–149, August,2012.

[7] T. M. Mitchell. Machine Learning. McGraw-Hill Companies, New York, USA, 1997, ISBN 0-07-042807-7.

[8] J.Han and M.Kamber.Data Mining.Concepts and Techniques, Simon Fraser University, Morgan Kaufmann publishers,ISBN 1-55860-489-8,2001.

[9] ShiNa, Shi Na, Liu Xumin, and Guan Yong. Research on k-means clustering algorithm: An improved k-means clustering algorithm. In Intelligent Information Technology and Security Informatics (IITSI), 2010 Third International Symposium on, pages 63–67, 2010.

[10] Powes, D.M.W,"Evaluation: From Precision, Recall and F-measure To Roc,Informedness, Markedness & Correlation" , Journal of Machine Learning Technologies, ISSN: 2229-3981 & ISSN: 2229-399X, Volume 2, Issue 1, pp-37-63, 2011.

[11] Fabrice Guillet, Howard J. Hamilton.Quality Measures in Data Mining.Studies in Computational Intelligence,ISBN-10: 3540789820 & ISBN-13: 978-3540789826, Springer; 1 edition p.p-140-141, 29, April 2008.