Journal of the ACS, Vol. 9, May 2018

Academic Educational Data Mining predictive model for early detection of students at academic risk

Asst. Prof. Dr. Mohamed EL-Zeweidy
Higher Institute for Computers and Information Systems
Al- Shorouk Academy
Cairo – Egypt
dr.mohamed.elzeweidy@sha.edu.eg

Asst. Prof. DrAhmed EL-Abbasy
Higher Institute for Computers and Information Systems
Al- Shorouk Academy
Cairo – Egypt
e.mail: dr.ahmed.elabbasy@sha.edu.eg

Abstract

In this research, we aimed at increasing college student retention by performing early detection of academic risk using data mining methods.

Predicting students' academic performance is critical for educational institutions because strategic programs can be planned in improving or maintaining students' performance during their period of studies in the institutions. Data mining technologies can be used to monitor students and simultaneously analyzing their academic behavior, thus providing a basis for implementing necessary intervention procedures, if required.

The paper describes and lays out a methodological framework to develop model that can be used to perform inferential queries on student performance using student academic records.

Preliminary results on Academic Educational Data Mining (AEDM) model development using Decision Tree data mining algorithms for classification are presented to classify students as early as possible, into three groups: 'low - risk' students who have a high probability of success; 'medium - risk' students who may succeed thanks to the measures taken by the university; and the 'high - risk' students who have a high probability of failing where several classification rules were generated.

Key words

Learning Analytics, Data Mining, SAS Enterprise Miner 12.3, decision tree, Educational data mining, knowledge discovery in databases (KDD).

1. Introduction

Academic Educational Data Mining (AEDM) has received significant attention within higher education, holds promise for improving learning processes in formal education, and beyond as well, including being highlighted in the recently released 2011 Horizon Report [4].

This interest can, in part, be traced to the work at Purdue University which has moved the field of academic analytics from the domain of research to practical application through the implementation of Course Signals. Results from initial Course Signal pilots between fall 2007 and fall 2009 have demonstrated significant potential for improving academic achievement [1].

Despite this early success, academic analytics remains an immature field that has yet to be implemented broadly across a range of institutional types, student populations and learning technologies [2].

Two distinct research communities, Educational Data Mining (EDM) and Learning Analytics and Knowledge (LAK), have developed in response.

The first workshop on Educational Data Mining was held in 2005, in Pittsburgh, Pennsylvania. This was followed by annual workshops and, in 2008, the 1st International Conference on Educational Data Mining, held in Montreal, Quebec. Annual conferences on EDM were joined by the Journal of Educational Data Mining, which published its first issue in 2009, with Kalina Yacef as Editor. The first Handbook of Educational Data Mining was published in 2010 [7].

In the summer of 2011, the International Educational Data Mining Society (IEDMS) (http://www.educationaldatamining.org/) was formed to "promote scientific research in the interdisciplinary field of educational data mining", organizing the conferences and journal, and the free open-access publication of conference and journal articles. The EDM community brings together an inter-disciplinary community of computer scientists, learning scientists, psychometricians, and researchers from other traditions. A first review of research in EDM was presented by Romero & Ventura [3], followed by a theoretical model proposed by Baker & Yacef [4]. A very comprehensive review of EDM research can be found in [6].

2. Related work

With increasing competition from the private sector and reduced funding in the public sector, many higher education (HE) institutions are giving much more attention to retention and progression of students throughout their studies. Add to this the explosion in electronic data which it is now possible to collect, and the potential for Learning Analytics is clear which leads Academic Educational Data Mining (AEDM) to be used significantly within higher education.

A number of studies have been made in education data mining for discovering different pattern to improve the student's performance [8]

Kalpesh Adhatrao, Aditya Gaykar, Amiraj Dhawan, Rohit Jha and Vipul Honrao(2013) [9] developed a system which can predict the performance of students from their previous performances using concepts of data mining techniques under Classification , analyzed the data set containing information about students, such as gender, marks scored in the board examinations of classes X and XII, marks and rank in entrance examinations and results in first year of the previous batch of students. By applying the ID3 (Iterative Dichotomiser 3) and C4.5 classification algorithms on this data, predicted the general and individual performance of freshly admitted students in future examinations , for a total of 182 students, the average percentage of accuracy achieved in Bulk and Singular Evaluations is approximately 75.275 these results developed by using rapidminer.

Brijesh Kumar Baradwaj, Saurabh Pal(2011)[13] conducted a performance analysis on 50 students whose records were taken from VBS Purvanchal University, Jaunpur (Uttar Pradesh) with the objective to study student’s performance using 8 attributes. Decision tree method was used to classify the data. Study helped teachers to improve the result of the student.

Khan [12] conducted a performance study on 400 students comprising 200 boys and 200 girls selected from the senior secondary school of Aligarh Muslim University, Aligarh, India with a main objective to establish the prognostic value of different measures of cognition, personality and demographic variables for success at higher secondary level in science stream. The selection was based on cluster sampling technique in which the entire population of interest was divided into groups, or clusters, and a random sample of these clusters was selected for further analyses. It was found that girls with high socio-economic status had relatively higher academic

achievement in science stream and boys with low socio-economic status had relatively higher academic achievement in general.

Pandey and Pal [11] conducted study on the student performance based by selecting 600 students from different colleges of Dr. R. M. L. Awadh University, Faizabad, India. By means of Bayes Classification on category, language and background qualification, it was found that whether new comer students will performer or not.

Edin Osmanbegović , Mirza Suljić (2012) [10], In this paper different techniques of data mining suitable for classification have been compared: Bayesian classifier, neural networks and decision trees. Neural networks have in many areas shown success in solving problems of prediction, approximation, function, classification and pattern recognition. Their accuracy was compared with decision trees and with the Bayesian classifier. This work is based on the survey conducted on students of the Faculty of Economics, in Tuzla, academic year 2010-2011, in which, aside from the demographic data, the data about their past success and success in college have been collected. This analysis was conducted after the training and testing of the algorithms, making it possible to draw conclusions on possible predictors of students' success ,the accuracy of NB is 76.65% , MLP is 71.20% and J48 is 73.93% , the results indicate that the naive bayes classifier outperforms in prediction decision tree and neural network methods, to develop this results using WEKA software , the sample of students used in this study is 257.

3. Methodological framework

Academic Educational Data Mining (AEDM) considered in our work are based on supervised learning (classification) techniques given that labeled training data is available (data sets used for training purposes carry both input features describing student characteristics, as well as student academic performance).

The goal is to build an early predictive model to discriminate between students in good standing and students that are not doing well to classify students as early as possible, into three classes : ‘low-risk’ students who have a high probability of success; ‘medium-risk’ students who may succeed thanks to the measures taken by the university; and the ‘high-risk’ students who have a high probability of failing in all subjects in the first term of the Academic year 2013-2014 among first and second year students.

The results of the model took place at the 8th week of the Academic Term (before the final exam by 8 weeks) thus each student has been informed by his risk level before the final exam early enough.

The Study was done on the data set of 2295 students and each student has 12 attributes. Data were analyzed using decision trees to predict student risk class depending on their academic performance (Lecture attendance- Section attendance- Midterm Score) and based on the student pre-university data (Student's Interest -High school type -High school Percentage).

Data partitioned into Two categories: the training, validation .

The training set during the learning phase allows the system to observe the type of relationships between input data and output, In this experiment the training set is 70 % of the collected data set. The validation data set is used to check the degree of learning of the model in order to determine if the model is converging correctly for adequate generalization ability, in another word a validation data set with the aim of assessing their predictive accuracy and consistency with the results obtained for the training data set. This phase involves an iterative process of fitting different versions of models of training and testing data set, each time evaluating their predictive performance and then chooses the best one based on their performance, also used to evaluate the performance (accuracy) of the model. In this experiment, the validation data set is 30% of the data set. Accuracy was evaluated using the misclassification rate (MISC) and average squared error (ASE) methods for the validation data set in SAS Enterprise Miner 12.3.

Our methodological framework consists of five phases, namely Collect data, Rescale/Transform Data, Partition Data, Train Model, and Evaluate Model using Test Data. The first four phases deal with preparing the input data used to build (train) and subsequently evaluate (test) model. Trained and tested model can then be used to score incoming data.

Collect Data: data used is extracted from the student records system of the Higher Institute of Computer and information technology at El-Shorouk Academy (SHA). Identifying student information is removed during the data extraction process. Logs data of individual course events tracked by each of the tools used by an instructor in a given course shell (e.g. 12 attributes Assignments, Assessments) as well as scores (grade contributions) on gradable events recorded by the Gradebook tool.

Rescale/Transform Data (Data Cleansing): Data is recoded / processed according to specific needs of the classification model building process. The end product is a data set of 2295 students that collects data of each course taken by each student in a given semester, augmented with (Lecture attendance- Section attendance- Midterm Score) and based on the student pre-university data (Student's Interest -High school type -High school Percentage).

The target (class) variable named Academic Risk establishes a threshold of questionable academic performance (e.g. 1 defines good academic performance "Low Risk"; 2 2 defines "Medium Risk"; and 3 defines poor academic performance "High Risk").

This stage is also concerned with the removal of outliers, handling of missing data, and addressing the issue of variability among courses in terms of assessment and student activity. The aforementioned variability is dealt with by replacing counts with ratios computed with respect to the average metric for the full course. An aggregated score is derived from partial (Gradebook) scores on gradable events. Once again the purpose is to shave variability across courses and compute a metric that can be used to make early predictions on student academic performance a few weeks into the semester.

Partition Data: input data is randomly divided in two datasets: a training data set, and a test data set. The training data set is used to build the model. Model is then tested using test data to compute a realistic estimate of the performance of the model on unobserved data. We use a ratio of 70% of the data used for training, and 30% testing, following standard data mining practice.

Build Predictive Model: We train our model with the training dataset, using different statistical and machine learning processes. We chose the C4.5 decision trees classifier as it is one of the state of the art robust classification methods that can deal with both categorical and continuous features. The *C4.5* [5] *decision tree* algorithm is a non-parametric classifier that learns rules from data.

Model Evaluation: Trained model is evaluated using the test data to measure the predictive performance derived from the confusion matrix that yields counts of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). Given the unbalanced nature of classes, the overall accuracy (TP+TN)/(TP+TN+FP+FN) is not a good metric for evaluating the classifier, as it is dominated by the student-in-good-standing class (TN+FP). We therefore appeal to two other accuracy metrics: sensitivity (TP/(TP+FN)), which measures

the ability of the classifier to detect the class of interest (academic risk); and specificity (TN/(TN+FP)) that measures the number of false alarms raised by the classifier.

4. Experimental setup

A 2295 data sample corresponding to the first term of the Academic year 2013-2014 undergraduate students of the Higher Institute of Computer and information technology at El-Shorouk Academy (SHA) for the first and second year students.

Initially the size of the data set is 2295 , 1138 of it for the first year students and 1157 for the second year students. All the subjects for the first year students are (Introduction to computer -Discrete Math- Math1 - Technical English 1 - physics) and the subjects for the second year students are (Math 3-data structures-Object Oriented Programming- Logic Design - Introduction To Accounting) total number of subjects are 10 subjects 5 of them for the first year students and another 5 for the second year students. Initially data is collected in an excel sheet from IT department this data is (student code-subject name-student year-high school score-high school percentage-student status-midterm score-lecture attendance-section attendance).In order to collect the interest of students in computer science and type of high school ,a questionnaire was prepared and distributed manually to the undergraduate students of both first and second year students.

Table 1 shows a sample of the data set used in this study.

Table 1. Sample Data Set

Code	Subject_Name	Student_Year	High_School	Percentage	Student_Status	Lecture_Attendance	Section_Attendance	Midterm_Score	Interest	High_School_Type	Risk_Level	Risk_Level_Code
3.13E+08	Introduction To Computer	First	344	83.902439	Residual	50	0	6	in between	English	High	3
3.13E+08	Introduction To Computer	First	300	73.1707317	Residual	12.5	12.5	3			High	3
3.13E+08	Introduction To Computer	First	373	90.9756098	Residual	12.5	25	7	in between	Arabic	Medium	2
3.13E+08	Introduction To Computer	First	353	86.097561	Newcomer	62.5	75	7	interest	English	Medium	2
3.14E+08	Introduction To Computer	First	353.5	86.2195122	newcomer	50	87.5	7			Medium	2
3.14E+08	Introduction To Computer	First	358.5	87.4390244	newcomer	62.5	87.5	7	interest	Arabic	Medium	2
3.14E+08	Introduction To Computer	First	368.5	89.8780488	newcomer	75	75	4	in between	English	Medium	2
3.14E+08	Introduction To Computer	First	363.5	88.6585366	newcomer	50	62.5	8			Medium	2
3.14E+08	Introduction To Computer	First	374.5	91.3414634	newcomer	62.5	87.5	8	interest	English	Medium	2
3.14E+08	Introduction To Computer	First	351.5	85.7317073	newcomer	50	75	7	interest	Arabic	Medium	2
3.14E+08	Introduction To Computer	First	358	87.3170732	newcomer	25	62.5	3	interest	English	Medium	2
3.14E+08	Introduction To Computer	First	376	91.7073171	newcomer	50	37.5	8	interest	Arabic	Medium	2
3.14E+08	Introduction To Computer	First	359	87.5609756	newcomer	62.5	75	8	interest	Arabic	Medium	2
3.14E+08	Introduction To Computer	First	353	86.097561	newcomer	50	87.5	5			Medium	2
3.14E+08	Introduction To Computer	First	344	83.902439	newcomer	50	87.5	7			Medium	2
3.14E+08	Introduction To Computer	First	356.5	86.9512195	newcomer	37.5	75	6			Medium	2
3.14E+08	Introduction To Computer	First	358	87.3170732	newcomer	62.5	50	6			Medium	2
3.14E+08	Introduction To Computer	First	345.5	84.2682927	newcomer	50	75	6	in between	Arabic	Medium	2
3.14E+08	Introduction To Computer	First	348.5	85	newcomer	75	75	7			Medium	2
3.14E+08	Introduction To Computer	First	350.5	85.4878049	newcomer	62.5	87.5	8	interest	Arabic	Medium	2
3.14E+08	Introduction To Computer	First	368.5	89.8780488	newcomer	37.5	62.5	7	interest	Arabic	Medium	2
3.14E+08	Introduction To Computer	First	343.5	83.7804878	newcomer	50	75	5	interest	Arabic	Medium	2
3.14E+08	Introduction To Computer	First	373	90.9756098	newcomer	62.5	75	10	interest	Arabic	Low	1
3.14E+08	Introduction To Computer	First	357	87.0731707	newcomer	50	75	5	in between	Arabic	Medium	2
3.14E+08	Introduction To Computer	First	352	85.8536585	newcomer	62.5	62.5	8	in between	English	Medium	2
3.14E+08	Introduction To Computer	First	358	87.3170732	newcomer	37.5	50	7	interest	English	Medium	2
3.14E+08	Introduction To Computer	First	350	85.3658537	newcomer	62.5	75	9	interest	English	Low	1
3.14E+08	Introduction To Computer	First	367.5	89.6341463	newcomer	62.5	75	4	interest	Arabic	Medium	2
3.14E+08	Introduction To Computer	First	361.5	88.1707317	newcomer	62.5	75	6	not interest	Arabic	Medium	2
3.14E+08	Introduction To Computer	First	347.5	84.7560976	newcomer	12.5	37.5	5	interest	English	Medium	2
3.14E+08	Introduction To Computer	First	349	85.1219512	newcomer	50	87.5	8	interest	English	Low	1

During the data selection step, only those attributes that are required for data mining were selected. The input attributes "Predictors" selected are (MIDTERM SCORE-LECTURE ATTENDANCE-SECTION ATTENDANCE-HIGH SCHOOL PERCENTAGE-STUDENT INTEREST-HIGH SCHOOL TYPE-STUDENT STATUS-SUBJECT NAME). The only output attribute "Target" is (STUDENT RISK LEVEL). Table 1 shows the list of attributes with their description and possible values.

Table 2 Summarize the Predictors and Target Attribute.

Name	Description	Possible values
Midterm Score	Is the score of the midterm exam of the student in a particular subject	0,1,2,3,4,5,6,7,8,9,10
Lecture Attendance	This is the Percentage of the student attendance in lecture. Represent the total number of days the student attended in lecture for every subject	From 0....to 100
Section Attendance	This is the percentage of the attendance in section. Represent the total number of days the student attended in Section for every subject	From 0.....to 100
Student Status	indicates if the student is a fresh one for the year or he is a residual one Ex -newcomer -Residual	- New Comer - Residual
Subject Name	Is the name of the subject which student taken in first year and second year	Introduction to computer , math3 , physics , discrete math, technical english1,math1,data structures ,Object Oriented Programming, Logic Design , Introduction to accounting
High School percentage	Is the percentage of the high school total grade.	From 0.....to 100
High School grade	Is the score which student get in the high school	From 0....to 450
Risk Level	Is the category(class)of the performance of student in a particular subject.	- High - Medium - Low

Table 2: list of all attributes with their description and possible values

Table 3. Features (predictors and target) in input dataset

Feature Type	Feature Name
Predictors	MIDTERM SCORE, LECTURE ATTENDANCE, SECTION ATTENDANCE, HIGH SCHOOL PERCENTAGE, STUDENT INTEREST, HIGH SCHOOL TYPE, STUDENT STATUS, SUBJECT NAME
Target	STUDENT RISK LEVEL ACADEMIC_RISK (1 = at risk; 0 student in good standing)

Experiments were conducted using SAS Enterprise Miner 12.3. For each of these tools a flow of execution was developed to perform the experiments. The experiments followed these guidelines:

(i) Out of the input dataset, generate five different random partitions (70% for training, 30% for testing) by varying the random seed

(ii) Balance each training dataset by oversampling records with class ACADEMIC_RISK=1

For each balanced training dataset for the classification algorithms C4.5 Decision Tree, train a predictive model.

For the purpose of this experimental work

(iii) Using each corresponding test dataset , evaluate each classifiers' performance by measuring their predictive performance. (sensitivity, specificity)

(iv) Produce summary measures (mean and standard error)

5. Results and discussions

Data was analyzed visually Using SAS Enterprise Guide and figure out the distribution of values ,Figure 2 shows the distribution of nominal attributes.

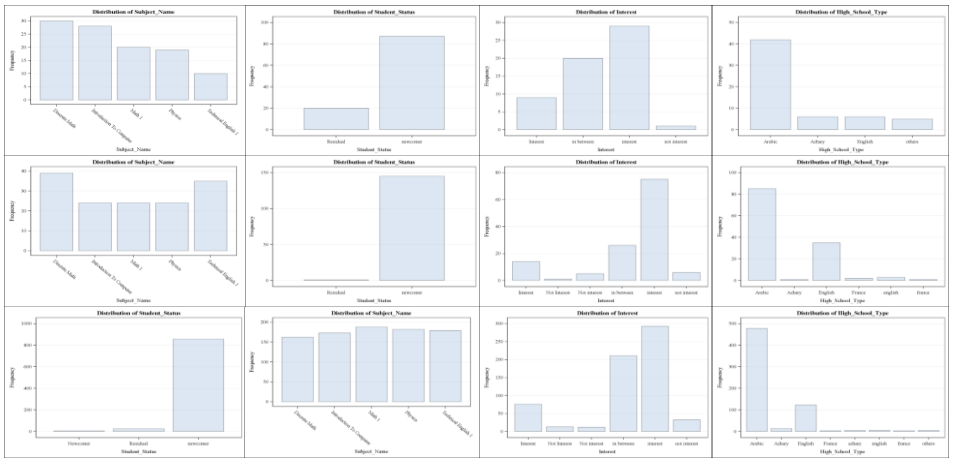


Figure 2: The distribution of nominal attributes

After analyzing the data , we feed the pruned excel sheet data set which contain all input attributes to the decision tree to rank predictors according to the strength of their relationship with depended or outcome variable (risk level).The outcome of the variable selection would be a rank list of predictors according to their importance for further analysis of the depended variable with the other methods for classification. Results of variable selection are presented in table 1.

Table 2: Variable importance

Orbs	NAME	RULES#	IMPORTANCE	VIMPORTANCE	RATIO
1	Midterm Score	2	1	1	1
2	Student Status	2	0.68882	0.34642	0.50292
3	Section Attendance	1	0.34808	0.16853	0.48416

Table 2,summarize the variable name and label, the number of rules (or splits) in the tree that involve the variable (NRULES), the importance

of the variable computed with the training data (IMPORTANCE), the importance of the variable computed with the validation data (VIMPORTANCE), and the ratio of VIMPORTANCE to IMPORTANCE.

The tree indicated that The most effective attributes(predictors) were: Midterm Score , student Status and Section attendance . The midterm score was the strongest attribute , then the student status ,then section attendance. The attributes such as : student interest , high school percentage , high school type , lecture attendance didn't show any clear effect for predicting student risk level , i.e. haven't any importance (the importance of them are 0) .

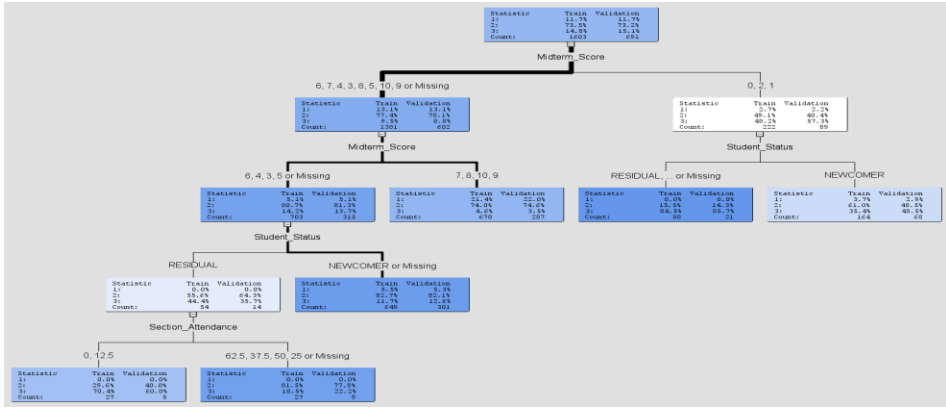


Figure 3: A decision tree diagram for predicting student risk level

Figure 3 , shows the decision tree for the model ,The number of levels of that tree are 4 levels and the number of leaves are 6 leaves.

The knowledge represented by decision tree can be extracted and represented in the form of IF-THEN rules. Classification tree rules can be easy to explain and used with the newly enrolled student. Rules for the decision tree for the model are given for all six terminals.

Table 3: Classification Rules Generated By Decision Tree for the predictive model.

Node=5 **IF** Midterm Score IS One OF : 7 ,8,10, 9 **THEN** Tree Node Identifier= 5 , Number Of Observations= 678 , Predicted: Risk Level= “MEDIUM" with probability=0.74.
Node=6 **IF** student status Is One OF : RESIDUAL , FROM OUTSIDE FIRST TIME Or Missing **AND** Midterm Score Is One Of: 0,2,1 **THEN** Tree Node Identifier=6 , Number Of Observations=58, Predicted: Risk Level=" HIGH" with probability=0.84.
Node=7 **IF** Student Status IS One OF: NEWCOMER **AND** Midterm Score Is One OF: 0,2,1 **THEN** Tree Node Identifier=7 , Number Of Observations=164 Predicted: Risk Level= "MEDIUM" with Probability=0.61
Node=9 **IF** Student Status IS One Of: NEWCOMER Or Missing **AND** Midterm Score Is One Of : 6,4,3,5 or Missing **THEN** Tree Node Identifier=9 , Number Of Observations=649, Predicted: Risk level=" MEDIUM" with probability=0.83.
Node=14 **IF** Student Status IS One OF: RESIDUAL **AND** Section Attendance Is One Of: 62.5 , 37.5 , 50 , 25 , or missing **THEN** Tree Node Identifier=14, Number Of Observations=27 , Predicted: Risk Level="MEDIUM" with probability=0.81.
Node=15 **IF** Student Status Is One OF: RESIDUAL **AND** Section Attendance Is One Of: 0 , 12.5 **AND** Midterm Score Is One Of : 6,4,3,5, or Missing **THEN** Tree Node Identifier=15 , Number Of Observations=27, Predicted: Risk Level="HIGH" with Probability=0.70.

Table 3, Shows the classification rules generated by decision tree for the model, The rules for the model are given for all six terminals.

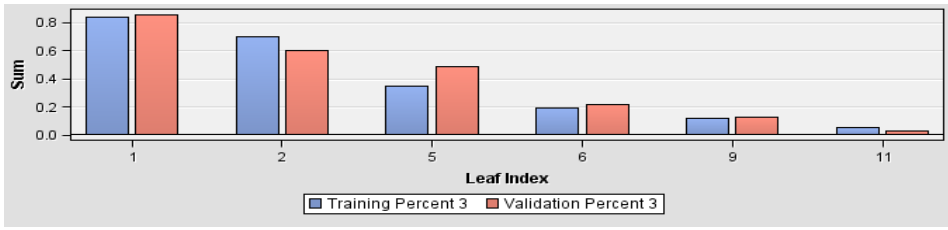


Figure 4: Leaf statistics

Figure 4 ,shows the leaf Statistics of the model, A Leaf Statistics bar chart in which the height of each bar equals the percentage of donors in

the leaf for both the training and validation data. The order of the bars is based on the percentage of donors (1's) in the training data.

Node#	Depth	Training Observations	Training Average	Validation Observations	Validation Average
5	2	678	0.05	287	0.03
9	3	649	0.12	301	0.13
7	2	164	0.35	68	0.49
6	2	58	0.84	21	0.86
15	4	27	0.7	5	0.6
14	4	27	0.19	9	0.22

Table 4: Tree Leaf Report

Table 4, illustrating figure 4 ,i.e. Shows the tree leaf report of predictive model, These report shows that the tree has six leaves, The leaves in the table are in order from the largest number of training observations to the fewest training observations. Each node have the number of it , depth of tree on that node , number of observations of training and validation on that node , and average number of training and validation on it.

Cumulative lift is the ability of the model to learn data. Whenever the percentile of the data are increasing whenever the cumulative lift is decreasing because of the fast learning of the model to the data. Figure 5 shows the Cumulative lift for the model.

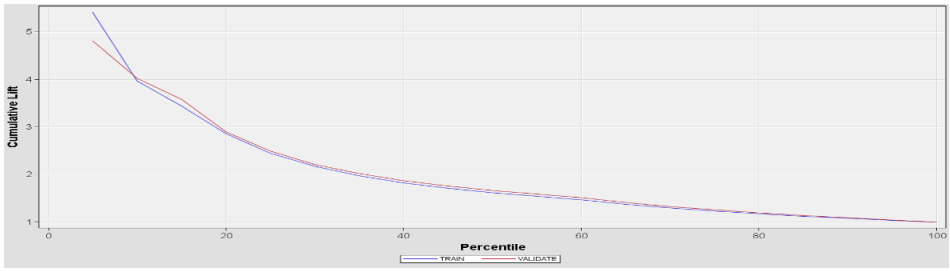


Figure 5: The Cumulative lift for the model

Table 5: The Assessment Score Ranking for the training set

Percentile	Cumulative lift	Number of observations
5	5.42026	81
10	3.96914	80
15	3.44229	80
20	2.85924	80
25	2.44617	80
30	2.1705	80
35	1.97135	81
40	1.82398	80
45	1.70927	80
50	1.61745	80
55	1.54228	80
60	1.46112	80
65	1.37259	80
70	1.2958	81
75	1.23011	80
80	1.17261	80
85	1.12186	80
90	1.07673	80
95	1.03635	80
100	1	80

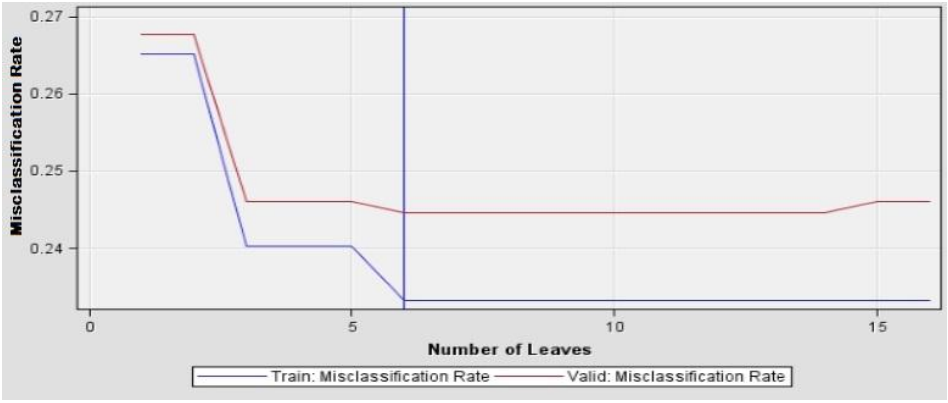
Table 5, Shows the value of cumulative lift and the number of observations on each percentile of data for the training data set.

Table 6: The Assessment Score Ranking For the validation Set

Percentile	Cumulative lift	Number Of Observations
5	4.81567	35
10	4.02004	35
15	3.58573	34
20	2.89406	35
25	2.49014	34
30	2.21227	35
35	2.0193	34
40	1.87014	35
45	1.75739	34

50	1.66447	35
55	1.58862	35
60	1.51109	34
65	1.41157	35
70	1.32867	34
75	1.25468	35
80	1.19178	34
85	1.13462	35
90	1.08525	34
95	1.03977	35
100	1	34

Table 6,Shows the values of cumulative lift and the number of observations on each percentile of data for validation data set ,Table 5 and table 6 illustration figure 5.



The Percentage of the correctly classified instances is often called accuracy or sample accuracy of a model.

Figure 6: The accuracy Of the model based on Misclassification rate

Figure 6, Shows the accuracy of model for classification applied on data sets using Misclassification Rate , the accuracy of the model based on misclassification rate is 76%.

Table 7:Misclassification Rate of leaves On training and validation sets

Number Of Leaves	Misclassification Rate on training	Misclassification Rate On validation
1	0.2651279	0.267727931
2	0.2651279	0.267727931
3	0.2401747	0.24602026
4	0.2401747	0.24602026
5	0.2401747	0.24602026
6	0.2333125	0.244573082
7	0.2333125	0.244573082
8	0.2333125	0.244573082
9	0.2333125	0.244573082
10	0.2333125	0.244573082
11	0.2333125	0.244573082
12	0.2333125	0.244573082
13	0.2333125	0.244573082
14	0.2333125	0.244573082
15	0.2333125	0.24602026
16	0.2333125	0.24602026

Table 7,illustrating Figure 6 ,i.e. shows the misclassification rate for training and validation sets on each leaf.

Figure 7: The Accuracy of the model based on Average squared Error

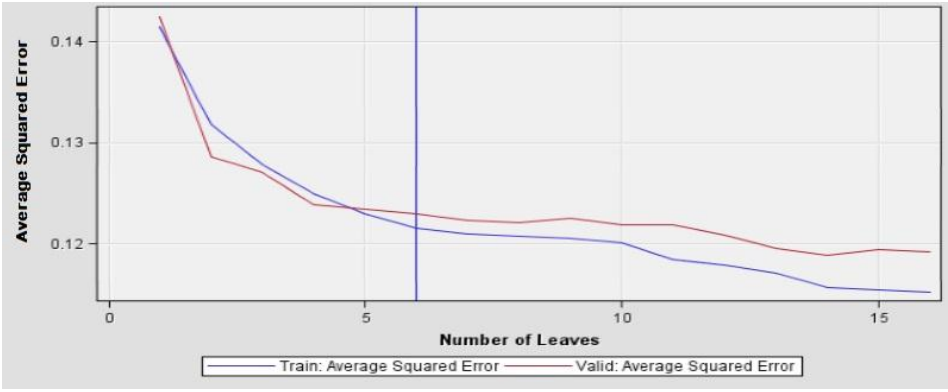


Figure 7, Shows the accuracy of a model for classification applied on data sets using Average squared Error , the accuracy of the model based on Average squared Error is 88%.

Table 8: Average Squared Error On Training and Validation Sets

Number Of Leaves	Average Squared Error On Training Set	Average Squared Error On Validation Set
1	0.141437	0.142465
2	0.131859	0.128631
3	0.127857	0.12705
4	0.124975	0.123922
5	0.123066	0.123468
6	0.121557	0.122986
7	0.121041	0.122396
8	0.120828	0.122133
9	0.12056	0.122546
10	0.120091	0.121929
11	0.118468	0.121936
12	0.117973	0.120966
13	0.117222	0.119627
14	0.115731	0.118974
15	0.115468	0.11948
16	0.115351	0.11929

Table 8,illustrating Figure 7 ,i.e. shows Average Squared Error for training and validation sets on each leaf.

Table 9: Fit Statistics of a model

Fit Statistics	Statistics Label	Train	Validation
MISC	Misclassification Rate	0.23	0.24
ASE	Average Squared Error	0.12	0.12

Table 9, Shows the fit statistics (Misclassification rate and average squared error) for the predictive model on training and validation sets.

To compute the accuracy of a model based on misclassification rate and average squared error ,the accuracy = 1- misclassification rate , and =1- average squared error.

Table 10: Accuracy Of a Model on validation and training sets based on misclassification

Data Set	Accuracy based on misclassification rate	Accuracy based on average squared error
Validation Set	76 %	88 %
Training Set	77%	88%

Table 10, Shows the accuracy percentage of a model based on misclassification rate and average squared error on a validation and training sets.

Table 11: All results of a model on a validation set.

Model Name	MISC	ASE	MISC Accuracy	ASE Accuracy	Tree leaves #	Tree levels #
Model	0.24	0.12	76 %	88 %	6	4

Table 11: Summarize all results of a model on a validation set, Number of tree levels of the model are 4 levels, Number of tree leaves of a model are 6 leaves, MISC(Misclassification rate) is 0.24 , ASE(Average Squared Error) is 0.12 ,Accuracy of a model based on MISC=1-MISC=76%, And accuracy of a model based on ASE = 1-ASE=88%.

6. Conclusion

This paper reports on the goals and objectives of the Academic Analytic, providing a detailed description of the methodology used to develop predictive model in academic analytics using SAS Enterprise Miner 12.3. to early detect potential weak "at-risk" students, so that the instructors can take an appropriate action towards them. For instance, they can give advice to prevent failure in the examination or early desertion of studies. In this work, we used decision tree technique to predict the student's risk level in all subjects for first and second year students. The model was tested on a real case study of computer science department from El-Shorouk Academy, the results of this study shows that from the validation

data set the accuracy of the model based on MISC is 76%, and the accuracy of it based on ASE is 88 %. On working on performance , Many attributes have been tested and some of them are found effective on the performance prediction. The midterm score was the strongest attribute, then the student status and then the section attendance. But another attributes didn't show any clear effect on outcome variable (risk level).

Future work can be conducted to build a user-friendly software of instructors using the concept of decision tree method in which an instructor will just have to enter the details of the students and it will display the possible result (risk level) of those students in final exams and then make alert on risky students.

Also a flexible system can be made in which we can add the data set dynamically and the system updates its results automatically, another future work the same concept can be extended for student internal selection within the faculty for the different departments, future work can be performed for improving the classification accuracies by the usage of different data mining techniques such as (ANN) with more distinctive attributes

7. Acknowledgements

This research is supported by EL Shourouk Academy..

8. References

[1] Arnold, Kimberly E. "Signals: Applying Academic Analytics", *EDUCAUSE Quarterly*, vol.33, no. 1, 2010.

[2] Baepler, P., Murdoch, C.J. (2010, July). Academic Analytics and Data Mining in Higher Education. International Journal for the Scholarship of Teaching and Learning vol. 4, no. 2, pp. 1-9 (July 2010) ISSN 1931-4744 @ Georgia Southern University.

[3] George Siemens, Ryan S J.d. Baker,"Learning Analytics and Educational Data Mining: Towards Communication and Collaboration", ACM 1-58113-000-0/00/0010, *Conference'10*, Month 1–2, 2010

[4] Johnson, L., Smith, R., Willis, H., Levine, A., and Haywood, K., (2011). The 2011 Horizon Report.Austin, Texas: The New Media Consortium.

[5]. Baker, R.S.J.d., Yacef, K. (2009) The State of Educational

Data Mining in 2009: A Review and Future Visions. *Journal of Educational Data Mining*, 1 (1), 3-17.

[6]. Romero, C., Ventura, S. (2010) Educational Data Mining: A Review of the State-of-the-Art. IEEE Transaction on Systems, Man, and Cybernetics, Part C: Applications and

Reviews. *40* (6), 601-618.

[7]. Romero, C., Ventura, S., Pechenizky, M., Baker, R. (2010) *Handbook of Educational Data Mining.* 2010. Editorial Chapman and Hall/CRC Press, Taylor & Francis Group. Data Mining and Knowledge Discovery Series.

[5] Quinlan, J.R., C4.5 : programs for machine learning. The Morgan Kaufmann series in machine learning. 1993, San Mateo, Calif.: Morgan Kaufmann Publishers.

[6] U.S. Department of Education, National Center for Education Statistics. Integrated Postsecondary Education Data System, Fall 2010. Retrieved February 15, 2011 from http://nces.ed.gov/collegenavigator.

[7] Vapnik, V.N., The nature of statistical learning theory. 2nd ed. Statistics for engineering and information science. 2000, New York: Springer.

[8] Hall, M., Frank, E., Holmes, G.,Pfahringer, B., Reutemann, P., Witten, I.H. (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, 11(1)

[9] Kalpesh Adhatrao, Aditya Gaykar, Amiraj Dhawan, Rohit Jha and Vipul Honrao . PREDICTING STUDENTS' PERFORMANCE USING ID3 AND C4.5 CLASSIFICATION ALGORITHMS. International Journal of Data Mining & Knowledge Management Process (IJDKP) vol.3, no.5, September 2013.

[10] Edin Osmanbegović , Mirza Suljić . DATA MINING APPROACH FOR PREDICTING STUDENT PERFORMANCE. Economic Review – Journal of Economics and Business, vol. X, Issue 1, May 2012.

[11] U . K. Pandey, and S. Pal, "Data Mining: A prediction of performer or underperformer using classification", (IJCSIT) International Journal of Computer Science and Information Technology, vol. 2(2), pp.686-690, ISSN:0975-9646, 2011.

[12] Z. N. Khan, "Scholastic achievement of higher secondary students in science stream", Journal of Social Sciences, vol. 1, no. 2, pp. 84-87, 2005.

[13] Brijesh Kumar Baradwaj, Saurabh Pal. Mining Educational Data to Analyze Students‟ Performance, (IJACSA) International Journal of Advanced Computer Science and Applications, vol. 2, no. 6, 2011.