Model Prediction of Academic Performance for First Year Students

Ernesto Pathros Ibarra García[1] Pablo Medina Mora[2]
School of Engineering Department of Teaching Support
National Autonomous University of Mexico (UNAM)
Mexico city, Mexico
[1]*patrotsky@yahoo.com,* [2]*pabme@unam.mx*

***Abstract*—The aim of this paper was to obtain a model to predict new students' academic performance taking into account socio-demographic and academic variables. The sample contained records of first semester students at a School of Engineering from a range of students' generations. The data was divided into three groups: students who passed none or up to two courses (low), students who passed three or four courses (middle), and students who passed all five courses (high). By using data mining techniques, the Naïve Bayes classifier and the Rapidminer software, we obtained a model of almost 60% accuracy. This model was applied to predict the academic performance of the following generation. After checking the results of the predictions, 50% were classified as correct. However, we observed that, for students of certain engineering majors of high and low groups, the model's accuracy was higher than 70%.**

Keywords: Academic Performance, Prediction Model, Data Mining.

I. INTRODUCTION

For educational institutions it is important to know the level of the new students' preparedness at admission. This helps to take decisions about further measures such as to detect which students require support, to predict student performance [13], to detect at-risk students [2] who may be successful but who need extra attention or specific individual care in order to succeed [3].

The following questions arise: Can the academic behavior of a student be predicted with prior information before entering school? If so, what tools are available? What is the prediction accuracy level that can be achieved with these tools? What prediction accuracy degree are we satisfied with? How useful are these predictions for teachers and tutors?

A. Previous Work

Several studies have applied data mining techniques to predict students' academic performance. In [1], linear regression models were used in order to find out which variables are more likely associated with academic performance; they found that prior academic performance was the most important variable. In [4], logistic regression was used to predict the academic success/failure with 70% effectiveness, agreeing on the fact that, previous academic performance is the most important variable besides attendance and class participation. In [7], statistical regression models were used to predict the academic performance of first year students with an R^2=0.476 effectiveness while the most important variable was the diagnostic test in mathematics. In [3], decision trees were also used with effectiveness between 75% and 80% whereupon the most important variables were linear algebra and calculus. Also in [9], decision trees were used with effectiveness of 96.47%; the author found a simple optimal rule structure based solely upon academic and professional background. In [10], the accuracies obtained by using Bayesian networks and decision trees were compared and showed that the latter were better, reaching 86% and 74% effectiveness; they found that the attributes with the highest information gain are the Cumulative Grade Point Average for the 2nd year, English Skill, and Institute Rank. In [2], neural networks, regression, and classification trees were compared. In this case, the former model was better, with results of 66.67% and 71.11% effectiveness. In [11], artificial neural networks were also utilized and obtained 70% effectiveness. In [12], they compared neural networks and traditional statistical techniques where the former model performed better, with results of 72.14% effectiveness; they found that the undergraduate academic results and test score are the most important variables in measuring the academic performance. In [14], the effectiveness among decision trees, neural networks and linear discriminant analysis was compared. In this case, the latter model performed best with 57.35% effectiveness; they found that 20% of the variables showed significant correlations with academic success.

In all of these papers, except in [10], the data sets are considerably small and most of them predict only two classes of the label attribute. In this paper, we work with a bigger dataset and a three class label attribute.

The employment of data mining techniques has proved to be a better tool, since they have been utilized successfully in different studies and research projects. Now they also are applied to educational data as exemplified in [13].

At admission, the students at a School of Engineering are assigned to a tutor as part of the New Era Tutoring Program. Tutors are provided with a report about each of their students which includes their diagnostic test results and some answers from the socio-demographic survey.

In the area of educational support, several linear prediction models on students' academic performance have been tested. These models have been limited as they only predict a small number of cases correctly. With the use of data mining techniques, we are trying to look for a model that is able to predict correctly a larger number of cases.

978-0-7695-4605-6/11 $26.00 © 2011 IEEE
DOI 10.1109/MICAI.2011.28

The main objective was the prediction of academic performance of students in the first semester of the 2011 generation. Achieving this goal will respond to the questions presented in the introduction.

In this paper, in subsection *II A*, we talk of the utilized variables which are listed in more detail at the end of this paper. Then, in subsection *II B*, we talk about the software used to perform data mining techniques, the experiments carried out with different variables, and the validations that helped to determine the degree of accuracy or effectiveness of the obtained models. In subsection *II C*, we discuss the model validation. Subsequently, in subsection *II D*, we explain the sources from which the data are obtained by briefly describing the process of data collection and data cleansing. Also, we mention the number of records that we used in order to carry out the training. Finally, in section *III* we discuss the results of the predictions of the best model obtained after applying it to predict the academic performance of students in the 2011 generation. The tutors' general opinion on these results is showed as well and we talk about our future work in section *IV*.

II. METHODS

A. The Model

The dependent variable, also known as the target or label variable, is *aprob_c* (the number of passed courses), which can take on three values: {L, M, H} (Low, Middle, High) as also raised in [14]. This was the result of discretizing *Low* as 2 or less passed courses, *Middle* as 3 or 4 passed courses and *High* as 5 courses passed. The student ID was defined as the identifying variable.

We worked with several variables that have been used in other works, such as: *age of student at admission* in [2], [9] and [11]; *gender* in [2], [7], [9], [10] and [11]; *parents educational status* in [2] and [11]; *whether the student work or not* in [12] and [9]; *reason for choosing an engineering major* in [1] and [4]; *academic preparedness* in [1] and [9]; *engineering major chosen* in [9] and [10]; *type of secondary school attended* and *university location* in [11]; *results in mechanics* in [7], *mathematics* in [2],[7], and [11]; *physics* and *chemistry* in [11]; *income* in [10]. All of these variables, along with the other variables we worked with, plus the identifying and dependent variables, equals 57 variables in total (see Table 1 in appendix). We also show the variables, which have worked as the best predictors according to the model performance results, in a confusion matrix (see Table 2 in appendix).

B. Experiments

We designed and implemented a database to integrate all the available information about the School of Engineering's new students. The data was collected from the socio-demographic survey and diagnostic test administered to new students annually as well as from the students' background information.

Based on the data type we have - both nominal and numerical -, we tested different classification algorithms such as *k-NN*, *IBk*, *decision trees*, and *naïve Bayes*. We noticed in previous experiments that the latter could obtain the best results.

The naïve Bayes model is tremendously appealing because of its simplicity, elegance, and robustness [15]. It is part of the top 10 algorithms according to the International Conference on Data Mining. Moreover, it can handle with both numerical and categorical data.

Rapidminer has an operator that optimizes the feature selection and chooses those variables that help best to describe the model by obtaining the best possible performance of a given model. The operator is called *Optimize Selection (evolutionary)*. This operator uses a genetic algorithm for feature selection which simulates the mutation (it connects and disconnects features) and crossover (exchanges properties). The following shows the optimizer's parameters and values that we selected:

- Minimum number of attributes: 4
- Population size: 40
- Maximum number of generations: 35
- Maximal fitness: infinity
- Selection scheme: tournament
- Tournament size: 0.4
- Dynamic selection pressure: enabled
- Keep best individual: enabled
- P initialize: 0.5
- P mutation: -1.0
- P crossover: 0.5
- Crossover type: shuffle

We trained and tested several models (see *figure 1* in appendix) by modifying the optimizer's parameters (showed above) and by selecting different variables; we discretized the data where appropriate in order to achieve the highest accuracy, which is the largest number of cases classified as correct. For example, in Junior High GPA we first tried with 4 different values, the top grade was included in a group. When we separated the top grade from the rest, the model's accuracy improved slightly.

Of the 55 variables (57, if we count the identifier and the label variables), the optimizer selected 30 (shown in Table 1 underlined). These are the variables that help best to explain the model.

At first, we worked with fewer variables since we did not yet have those related to age at admission, shift, and all the other diagnostic test results (mechanics, chemistry, algebra, geometry, thermodynamics, trigonometry, calculus, and electromagnetism). Without these variables, we got a model of 58.18% accuracy (L=59.19%, M=47.85%, H=66.32%). When we added the diagnostic test results, the accuracy improved to 58.48% (L=60.20%, M=48.78%, H=64.35%). Still, we obtained another improvement when we included all these variables and only the diagnostic results averages were excluded. We reached 58.64% accuracy (L=60.58%, M=50.17%, H=63.93%). According to the last two results, we observed that the group *M* improved without affecting so much the performance of the groups *L* and *H*. The latter accuracy is the maximum value of accuracy achieved. The squared error is 0.320 +/- 0.324.

C. Model Validation

In order to measure the model's performance, the data was split in two groups: the training set (70%) which was used to train the model, and the test set (30%), utilized to test the model in order to measure its accuracy by means of a confusion matrix.

We obtained a model of almost 60% accuracy. Although it is not high enough we considered it acceptable since we are working on a social phenomenon. *See table 2*.

D. Dataset

Given the institutional nature of the University, there are several sources of information about future students. These sources include the General School Administration (DGAE) – which receives the general data of the students -, the socio-demographic survey, and the diagnostic test. The survey and the test are administered to the School of Engineering's students at admission. The responses and the data are collected and stored in a database designed for the purpose of information analysis.

Such data was mainly stored in spreadsheet documents, which were not related. Then, we cleansed the data because there were repeated records and no standardized values. The database facilitated the data processing and organization.

The data sample includes students of the twelve engineering majors taught at the UNAM School of Engineering (see majors list in the appendix table 4) of 2008, 2009, and 2010 generations.

From this sample, 2112 records belong to the 2008 generation; 2177 belong to the 2009 generation, and 2295 belong to the 2010 generation, giving a total of 6584 records.

From the same records, we know that the students:

- Passed 2 or less courses: 2224 records
- Passed 3 or 4 courses: 2285 records
- Passed 5 courses: 2075 records

in the first semester (see Figure 2 in appendix).

III. Results

A. The Model's Accuracy

Right after the end of the first semester of the 2011 generation, we could check the results of the model's predictions. The accuracy of it obtained was 50.39% (see Table 3), which is lower than the 58.63% that we obtained during the model's validation.

Nevertheless, we found interesting results when examining the confusion matrices for each group. For example, when middle level students were classified as high, what they had mostly in common was that their self-perception in major guidance was in doubt and their grades in algebra were quite good or good (*figure 6*). On the other hand, when middle level students were classified as low, their self-perception in major guidance is in doubt as well. However, their grades in algebra or calculus are low (*figure 7*). We found this by using k-means clustering. This case will need deeper analysis to try to determine which variables define better a middle level student.

Low, Middle, and High groups are described in the next subsections (see also figures 3, 4, and 5 in appendix).

B. Better Classification for the Low and High Groups in Some Cases

For predictions of low-performing students who chose electrical and electronic engineering and geophysics engineering, the accuracy was 71.80% and 70.75% respectively. In the case of the students of computer engineering and petroleum engineering the accuracy was 61.73% and 60.47% respectively. See *Figure 3*.

For predictions of high-performing students of mechatronics engineering, the accuracy was 77.72%. In the case of students of telecommunications engineering, geological engineering, mechanical engineering, electrical and electronic engineering, civil engineering, and industrial engineering the accuracy was higher than 60%. We also observed that in the case of predicting mining and metallurgical engineering, the accuracy was null. See *Figure 5*.

C. Middle Group: Unpredictable

As for the predictions of students belonging to the middle group, the accuracy percentages were low. The highest accuracy achieved was 60.29% for mechatronics engineering (see *Figure 4*). So far, predicting this group seems to be difficult. Hence, we have to search for other ways to improve the model's performance in this case.

D. Data Mining and the Software Used

Data mining is the process of discovering interesting knowledge from large amounts of data stored in databases, data warehouses, or other information repositories [5].

Rapidminer [8] is open source software that has a big variety of algorithms to perform data mining as well as tools to import data in several formats, tools for data processing and data cleansing, as well as tools to create graphs, statistics and reports.

IV. Discussion

A. The Percentage Accuracies

At the first glance, we have noticed the model's low percentage accuracy. However, if we observe this result from different perspectives, we see relatively high percentage accuracies for some groups along with some engineering majors.

The information obtained may well be valuable:

First, we proved that mechatronics engineering students are high performance. We can practically say that, if a student chooses this major engineering, the probability to succeed in the first semester is high.

Second, if the model predicts poor performance in the cases of geophysics engineering, and electrical and electronic engineering, we should definitely pay attention to this issue.

As for the middle group, we see that the model has problems to classify the students' academic performance since it is the line between succeeding and failing. This makes the model relatively unstable.

B. User's Opinions: Tutors

We administered a survey to the tutors about the usefulness of the predictions. Of about 190 tutors, 56 answered the survey (which was set online).

Some of them expressed doubts about the model's effectiveness in terms of forecasting. They also commented that the predictions should be used carefully, in a complementary way and not in a deterministic way.

On the other hand, they commented on the advantages of this kind of information, since they have more elements to point out the guidance and attention toward their tutees. They considered that this information provides an overview of the students' preparedness.

V. Conclusions and Future Work

Predictions for the low and high groups have significant percentage accuracy in some cases, exceeding 70% if the naïve Bayes classifier is used. This shows that it is possible to obtain a good prediction model. For example, it can be used to detect low performing students and take appropriate decisions even before the courses start and, hence, to revert their academic standing. It can also be used to detect high performing students in order to channel them to personalized educational program services. Given the complexity of human behavior, the model is limited for academic performance prediction. However, it is a useful tool, besides others, such as interviews and daily monitoring, that contributes to prognosis and, as such, it directs the tutor's work for the benefit of the students.

In some cases, the number of correct predictions was significant, which allows us to create prediction models with high percentage accuracy for some cases.

We will search for new variables in order to include them in the training set, such as the information about the students' study habits, which might help to improve the current model's performance. We will try other data mining algorithms and optimizers as well.

Acknowledgment

We thank the UNAM School of Engineering for its support. We also thank Christian Jordan and Iván Vladimir Meza for their thoughtful comments on our paper. We would like to thank the independent international reviewers for their comments and very useful critics on our paper.

References

[1] Byrne M, Flood B. Examining the relationships among background variables and academic performance of first year accounting students at an Irish University, Journal of Accounting Education, Volume 26, Issue 4, December 2008, Pages 202-212.

[2] Chun-Teck Lye, Lik-Neo Ng, Mohd Daud Hassan, Wei-Wei Goh, Check-Yee Law, Noradzilah Ismail (2010). Predicting Pre-university Student's Mathematics Achievement. Procedia - Social and Behavioral Sciences, Volume 8, International Conference on Mathematics Education Research 2010 (ICMER 2010), Pages 299-306

[3] Dekker, G.W., Pechenizkiy, M., Vleeshouwers, J.M. (2009). Predicting Students Drop Out: A Case Study. In International Conference on Educational Data Mining, Cordoba, Spain, 41-50.

[4] García, M.V., Alvarado, J.M. y Jiménez, A. (2000). La predicción del rendimiento académico: Regresión lineal versus regresión logística. Psicothema, 12, 248-252.

[5] Han, J., Kamber, M. Data Mining Concepts and Techniques. Morgan Kaufmann, San Francisco (2006)

[6] Kirsten McKenzie; Robert Schweitzer. Who Succeeds at University? Factors predicting academic performance in first year Australian university students. Higher Education Research & Development, 1469-8366, Volume 20, Issue 1, 2001, Pages 21 – 33

[7] Lee, S., Harrison, M., Pell, G., & Robinson, C. (2008). Predicting performance of first year engineering students and the importance of assessment tools therein. Engineering Education: Journal of the Higher Education Academy Engineering Subject Centre, 3(1).

[8] Mierswa, I.: RapidMiner. http://rapid-i.com. accessed 25.09.2011.

[9] Moore J. S. An expert system approach to graduate school admission decisions and academic performance prediction, Omega, Volume 26, Issue 5, October 1998, Pages 659-670.

[10] N. Thai Nge, P. Janecek, and P. Haddawy, "A comparative analysis of techniques for predicting academic performance" in 37th ASEE/IEEE Frontiers in Education Conference, 2007.

[11] Oladokun, V.O., A.T. Adebanjo, and O.E. Charles-Owaba. 2008. "Predicting Students' Academic Performance using Artificial Neural Network: A Case Study of an Engineering Course". Pacific Journal of Science and Technology. 9(1):72-79.

[12] Paliwal M.,Kumar U. A. A study of academic performance of business school graduates using neural network and statistical techniques, Expert Systems with Applications, Volume 36, Issue 4, May 2009, Pages 7865-7872.

[13] Romero, C., Ventura, S.: Educational data mining: a review of the state-of-the-art. IEEE Trans. Syst. Man Cybern. C Appl. Rev. 40(6), 601–618 (2010)

[14] Vandamme, J. -P. , Meskens, N. and Superby, J. -F.(2007) 'Predicting Academic Performance by Data Mining Methods', Education Economics, 15: 4, 405 — 419, First published on: 26 June 2007 (iFirst)

[15] X. Wu, V. Kumar, J.R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G.J. McLachlan, A.F.M. Ng, B. Liu, P.S. Yu, Z.-H. Zhou, M. Steinbach, D.J. Hand, and D. Steinberg, "Top 10 Algorithms in Data Mining," Knowledge and Information Systems, vol. 14, no. 1, pp. 1-37, 2008.

VI. Appendix

TABLE I. Independent Variables. The Variables Chosen (30) by the Optimizer are Underlined.

Information on School History	Results in Diagnostic Test
• Elementary school type • Junior high school type • High school of origin (of the UNAM or outside) • High school type or Institution type • CCH high school • ENP high school • Elementary school GPA • Junior high school GPA • High school GPA • High school years • Admission type to the School of Engineering	• Algebra • Trigonometry • Euclidean geometry • Analytic geometry • Calculus • Mechanics • Thermodynamics • Electromagnetism • Chemistry
Socio-demographic Information	**Self-Perception**
• Engineering major • Gender • Age at admission • Shift in high school • Mother's educational status • Father's educational status • Household income • Whether the student works • Whether somebody else can help the student if he/she stops working • Parents' situation • Number of siblings • Number of persons who support the household budget • Number of persons living in the household • How he/she gets to the university • Time taken to get to the university	• Self-perception in mathematics • Self-perception in major guidance • Student self-perception • Main reason to study engineering
Goods and services in his/her household	
• Fridge • Washing machine • Drying machine • Dish machine • Water boiler • Telephone line • Cell phone	• Video recorder • Cable TV • Sound equipment • Microwave oven • Computer • Internet • Family car • Own car • Service staff

TABLE II. Confusion Matrix of the Best Model Obtained (Accuracy: 58.64%, Kappa: 0.397).

	true L	*true M*	*true H*	*Class precision*	*Totals*
pred. L	435	223	60	60.58%	718
pred. M	165	294	127	50.17%	586
pred. H	73	169	429	63.93%	671
class recall	64.64%	42.86%	69.64%		
totals	673	686	616		1975

TABLE III. Confusion Matrix of the Model's Forecasts (Accuracy: 50.39%, Kappa: 0.296).

	true L	*true M*	*true H*	*Class precision*	*Totals*
pred. L	476	305	103	53.85%	884
pred. M	249	373	200	45.38%	822
pred. H	64	175	400	62.60%	639
class recall	60.33%	43.73%	56.90%		
totals	789	853	703		2345

TABLE IV. List of Engineering Majors Offered at the UNAM School of Engineering

- Civil Engineering (ICi)
- Mining and Metallurgical Engineering (Imm)
- Geological Engineering (IGl)
- Petroleum Engineering (IPe)
- Geophysical Engineering (IGf)
- Computer Engineering (ICo)
- Telecommunications Engineering (Ite)
- Geomatics Engineering (IGm)
- Mechatronics Engineering (IMt)
- Mechanical Engineering (IMe)
- Industrial Engineering (IIn)
- Electrical and Electronic Engineering (IEe)

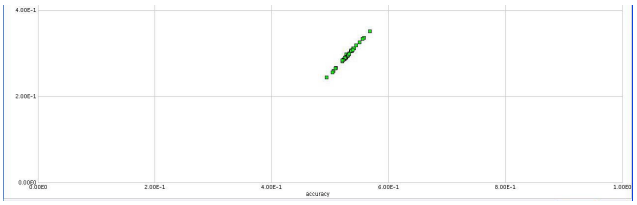

Figure 1. Optimizer's graph of all models tested (kappa - accuracy).

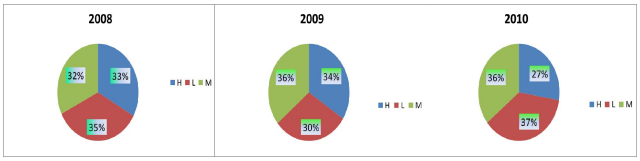

Figure 2. Percentage of students who passed 2 or less, 3 or 4 and 5 courses per generation.

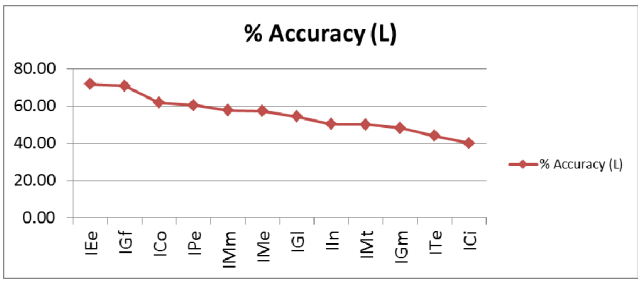

Figure 3. Accuracy percentages for the low group per engineering major.

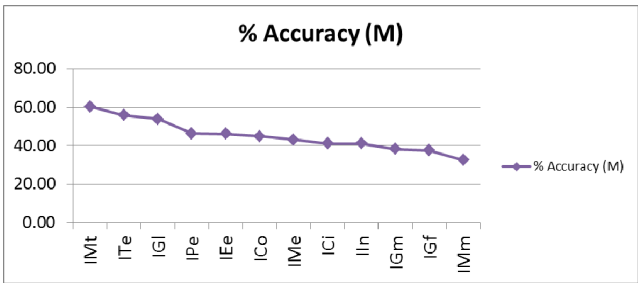

Figure 4. Accuracy percentages for the middle group per engineering major.

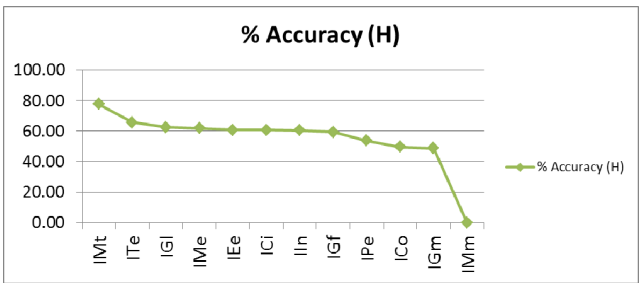

Figure 5. Accuracy percentages for the high group per engineering major.

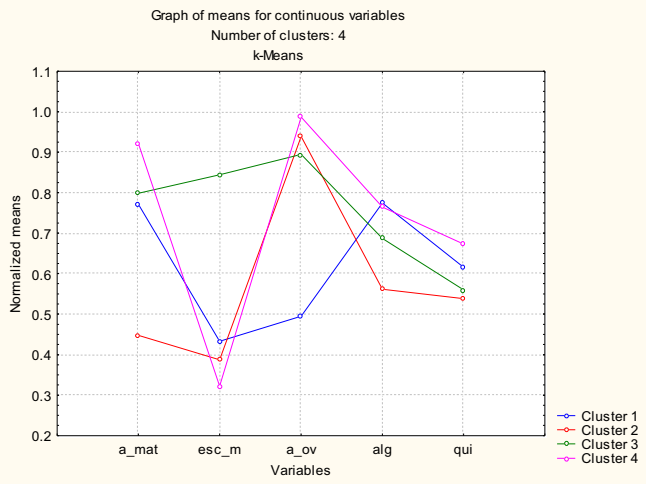

Figure 6. Clusters for middle level students classified as high. *A_ov* refers to self-perception in major guidance. *Esc_m* refers to mother's educational status and *qui* refers to chemistry.

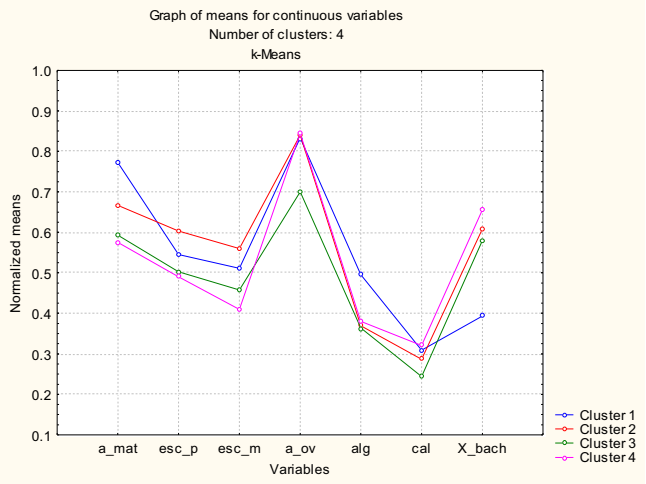

Figure 7. Clusters for middle level students classified as low. *Alg* referes to algebra, *cal* refers to calculus. *X_bach* refers to high school GPA.