Journal of Education and Vocational Research
Vol. 3, No. 5, pp. 132-137, May 2012 (ISSN 2221-2590)

Measuring Students' Performance with Data Mining

Jean Pierre Atanas
Advanced University Program, the Petroleum Institute, Umm Al Naar, Abu Dhabi, United Arab Emirates
jpierre@pi.ac.ae

Abstract: Understanding the true reasons behind students' failure, and bringing preventive measures to this issue at early stages are invaluable in the educational learning process. Preventing problems such as language deficiency or misclassification of the students in the appropriate academic levels is primordial for any educational institution. Many factors influence the learning process of the students, such as the demographic characteristics, educational background as well as language barrier. This work highlights the most preponderant factors affecting students' advancement in the learning process and provides support to academic administrators. It uses some of state of the art classification and regression algorithms in the application domain of predicting students' progress. Datasets were filtered and trained using predictive algorithms. It is shown that Science learning and English language skills are highly correlated. Datasets are not always suitable for data mining unless it is preprocessed and well adapted to the context being studied. A tool has been developed to preprocess the data provided that feeds into Weka Data Mining Software to profile students' performance.

Keywords: *Education, correlation, measuring performance, prediction, students' profiling*

1. Introduction

In recent years, researchers have started to apply machine-learning techniques to predict students' performance and their learning styles (Pattanasri, Mukunoki, & Minoh, 2012).Tracking students' performance is a preventive process and could be applied at early stages in order to identify poor performers and hence apply different learning styles. Remedial actions can be taken, on the spot, by tutors or administrators, to overcome this issue and provide additional help to minimize groups at risk. University placement and entrance exams administered to students are not always conclusive and predictive of the students' performance as many factors can intervene throughout the academic year. Diagnosis of students' performance is a dynamic process. It becomes more accurate as new curriculum information is entered during the academic year. Many machine learning algorithms have been used for the purpose of predicting performance(Kotsiantis, Pierrakeas, & Pintelas, preventing student dropout in distance learning systems using machine learning techniques, 2003);(Kotsiantis, Pierrakeas, & Pintelas, 2004);(Xenos, Pierrakeas, & Pintelas, 2002). None has been identified as the best algorithm for all cases (Mitchell, 2011). Accuracy of the outcomes depends on the quality of the data itself, the preprocessing phase, algorithms being used as well as the attribute being predicted. The following sections describe in brief the University Program and policies, the correlations between language skills and Science, an experiment results for all of the tested algorithms to profile students' performance, a comparative study among learning algorithms and a conclusion. The AUP (i.e., Advanced University Program) is a preparatory program for one year, and is designed to help students develop the knowledge, study skills and work habits that are needed to prepare them to be successful at an excellent engineering university. While studying in this program, learners have the opportunity to earn university credit for courses in mathematics, chemistry and physics. Students who meet the requirements may apply these credits to their Bachelor degree, allowing them to proceed through this program in less time. English language and computing courses have been designed to help our students acquire the language, technological, and analytical skills required to meet entrance requirements fixed by the university and assist them in their future studies.

The AUP English course is a one-year modular program that helps high school graduates prepare for the four-year engineering degree program through a set of four modules in the fall and one in the spring. The AUP English course also works to support AUP Math and Science to ensure student success in all subject areas. The

course consists of different eight-week modules. Depending on their performance in an initial placement test, students are expected to complete between one and five modules of language study within the allotted year. AUP Mathematics and Science courses run for a period of one year (i.e., two semesters) and lessons are scheduled for five hours per week mathematics and eight hours per week Physics and Chemistry. Students are assessed on a regular basis. Weekly quizzes are the norm, and writing tests in addition to a final exam are a feature of students' continuous assessment. Progress grades are collected weekly. To exit the program, students are required to achieve a minimum of 500 on the TOEFL or a minimum of 5 in IELTS at the end of the year as well as successfully complete the English course of study with a minimum of C. These requirements serve as criteria for entrance into the Freshman Arts and Sciences Program. At the end of the academic year, students are classified into four categories: "Failed", "Passed", "Terminated" or "Withdrew". The basic features or attributes along with their respective values are presented in the following table 1. The set of attributes was divided into three groups: The registry class, the tutors' class and the assessment class. The registry class represents attributes collected from the admission office of the university. The tutors' class corresponds to the weekly assessments done per semester in all subjects: English: five modules in total, Mathematics, Physics and Chemistry (two levels each, AUP and AP levels). Finally the class attribute represent the final average assessment of the year based on cumulative GPA's obtained from both semesters and classified into four categories: Terminated, Withdrew, Pass, Fail.

2. Data Analysis and Results

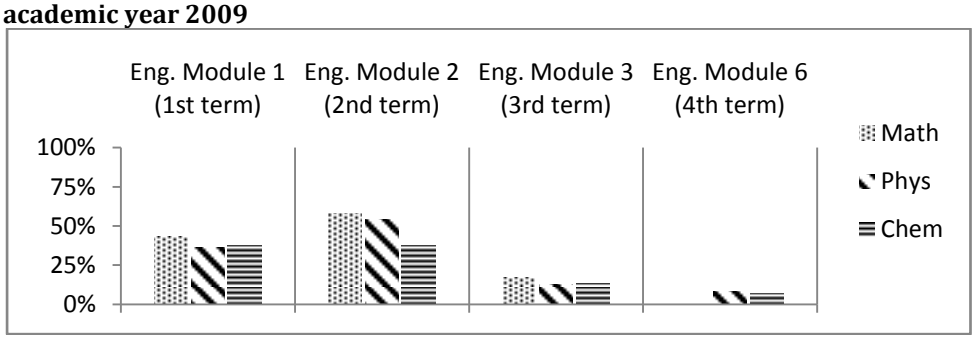
The data set comprises male students within the age interval of 18-20 years. Most of them are living on the same campus, having homogeneous learning habits and cultural background. Analysis of the registry attributes showed that the ratio of national students to expatriate who passed, in 2009academic year, is 33%. Furthermore, the percentage of national students with drawn, failed or terminated, for diverse reasons, was about 60% in the same year. Another important factor was the students' educational background. The classification predicts that 85% of students coming from private schools are more likely to pass at the end of the preparatory program, whereas 80% of the students coming from public schools failed at the end of the academic year. This ratio is very high and one can raise a question about the learning efficiency at these public schools. According to a survey conducted by the Abu Dhabi Education Council many teachers in public schools had not taken any professional development courses(TheNational, 2010), furthermore, more than half, 51 % felt such workshops were seldom conducted and rarely followed up. Students are more likely to struggle with the English in their first semester. Their knowledge of the English language is very limited. This weakness must have been carried forward from schools. Students in high schools have shown both low-achievement in English and a negative attitude towards Science subjects where most of the Science courses were taught in Arabic. At the university level, many methods, strategies and pedagogies are mandated by instructors and educators to help students accelerate the language learning so important to their future. This achievement will be highlighted in the following paragraph. With all the effort of the English department to overcome the situation by innovative means (Eid, 2009), 28% of national students failed the TOEFL or the IELTS at the end of the preparatory year2009. Due to Arts and Science requirements for entering the freshman year, students will be terminated unless they score minimum 500 or above in the TOEFL or equivalent.

Starting from the first semester, Science and Math are introduced to them according to their level in the entrance exam. A net correlation was seen between English skills and learning Sciences which could be considered as causally transparent outcomes. As outlined above in the previous paragraph, English language skills among students were so weak to follow up scientific concepts with the instructors. Progress in Math and Science was not perceptible until the second semester where specific modules in English were designed and offered to accelerate the language proficiency of the students in the first semester. Results of this outcome were effective in the second semester. Figure 1 shows strong correlation between Math, Physics and Chemistry and English, during each term, that is, in the first semester where English Modules 1 and 2 were offered. Students had not yet enough command of English to understand the basic scientific concepts and terminologies. The correlation is even higher in the second term of the first semester suggesting that English skills are more required as students advance in the Science curriculum where definitions and terminologies are required. It is not before the beginning of the second semester that the correlation drops by a factor of 3. English language is no more a barrier to achieve good performances in other fields of Science.

Table 1: Set of attributes from students' registry and from instructors' records for year 2009

Student's registry attributes	Gender	Female, Male
	Nationality	List of Countries
	Country Status	Expatriate, Nationals
	High School Average	No. : 0-100 (bins of 10)
	High School Status	Private, Public
	AP Calculus (AB)	Scores : 5,4,3,2,1
	AP Chemistry(C)	Scores : 5,4,3,2,1
	AP Physics Mech. (C)	Scores : 5,4,3,2,1
	AP Physics Elec. (C)	Scores : 5,4,3,2,1
	AUP Math	
	AP Physics (C)	
	AUP Physics	
	AP Chemistry (C)	
	AUP Chemistry	Scores[1] : A,A-,B+,B,B-,C+,C,C-,D,F,W,WI,WF,I
	English Module 1	
	English Module 2	
	English Module 3	
	English Module 6	
	Computing(C)	
	GPA	No. : 0-4
	Status	Pass, Fail
Attributes from tutors records	AP Math	
	AP Physics	
	AP Physics	
	AUP Physics	Scores : A,A-,B+,B,B-,C+,C,C-,D,F,W,WI,WF,I
	AP Chemistry	
	AUP Chemistry	
	English Spring	
	Computing	
	GPA	No. : 0-4
	Status	Pass, Fail
Assessment Class		Terminated, Withdrew, Pass, Fail

Figure 1: Correlation between Science and modules of English offered sequentially during the first academic year 2009



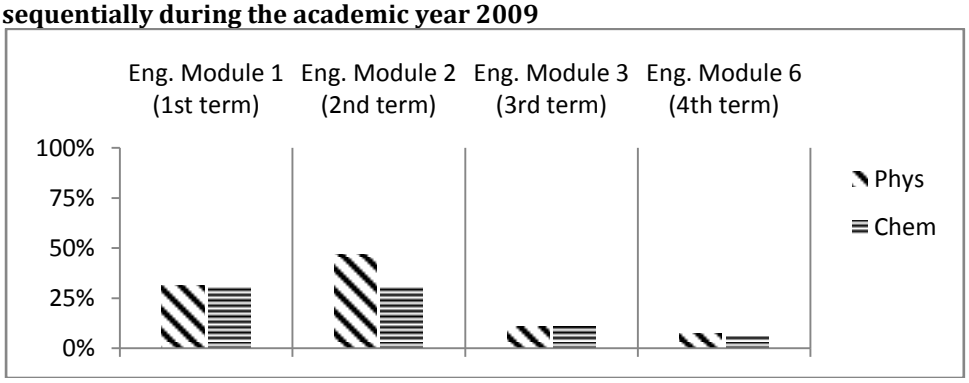
Students' performance can vary within Science courses for a given semester. Table 2 shows the correlation between Math and Science: Physics and Chemistry.

Table 2: Correlation between Math and Science (physics and chemistry) in each semester

correlation in %	Physics	Chemistry	Semester
Math	85% (±5%)	85% (±5%)	1
	85% (±5%)	80% (±5%)	2

In order to show the impact of English language on Science learning, a cross-correlation was derived from Table 2.Taking this new correlation into account, the cross-correlation graph obtained between English and Science is illustrated in the following chart. Figure 2 shows clearly the impact that English has on the Science courses. Although mathematics is considered an important tool for analyzing scientific concepts or solving problems, and has a positive impact on learning Science in general, English language remains the primordial factor indispensable to the process of learning Science. A quick look on both figures shows that cross-correlation drops in average, throughout the academic year, by a factor 4% and 3% for Physics and Chemistry respectively.

Figure 2: Cross-correlation between Science (Physics and Chemistry) and modules of English offered sequentially during the academic year 2009



Because of this study, one can conclude that Science should not be administered to students before a complete mastery of the English language. At least two terms are needed for the students to acquire the basic skills in English and to develop their linguistic skills in order to be able to grasp the scientific concepts in physics and chemistry more easily. The program started to apply the outcome of this study in the first semester 2011.Classification of students into appropriate streams or levels proposed by the university is of major concern. The AUP department is looking for an efficient way to optimize students' classification in streams in a very short time by taking into consideration their educational, cultural and linguistic background. It is worth mentioning that the program is open to all local Emirati students that constitute approximately 75% from the whole students' intake of the academic year and the rest 25% to all other nationalities. The internal policy of the program does not really restrict the access to Emirati students to

college regardless of their lack in English and their weakness in Science. Instead, the program is challenged to prepare these students in two years maximum to have the required level to enter the freshmen year. A decision has to be taken in order to classify efficiently all the students in the appropriate streams, offered by the program, from the first beginning of the semester or during the first term. An entrance exam is administrated a week before the beginning of the semester. The data is useful but not reliable. An hour exam is sometimes not indicative of the true educational level of the student. The alternative was to turn to data mining in order to classify students with high accuracy, to be able to track them through the semester and to be able to predict poor performers. Discriminant function analysis is best suited to modeling with continuous-scaled variables as predictors. Since this data set is composed of both categorical and numeric variables, discriminant function analysis cannot handle this kind of complexity of data types in one single analysis unless tremendous data transformation, such as converting categorical variables to dummy codes, is used (Streifer & Schumann, 2005). Alternatively, classification trees and Regression algorithms were employed to generate student profiles. In the following section a brief introduction to these methodologies will be given. Classification trees are used in an approach to predicting student performance in a high-enrollment, high-impact, and core engineering course (Fang & Lu, 2009); aim to find which independent variable(s) can successively make a decisive spilt of the data by dividing the original group of data into pairs of subgroups in the dependent variable.

Table 3: Performance comparison on the selected class for all used decision trees

	TP	FP	Precision	ROC	Class
J4.8 - No pruning	0.903	0.098	90.2%	0.933	Pass
	0.667	0.071	90.4%	0.799	Fail
J4.8- pruning	0.952	0.361	72.5%	0.955	Pass
	0.75	0.091	89.2%	0.89	Fail
Simple Cart	0.903	0.082	91.7%	0.955	Pass
	0.833	0.131	86.4%	0.903	Fail
Random Tree- No pruning	0.887	0.115	88.5%	0.916	Pass
	0.625	0.101	86.1%	0.749	Fail
LAD Tree	0.935	0.016	98.3%	0.979	Pass
	0.792	0.141	84.9%	0.902	Fail
BayesNet Tree	0.903	0.016	98.3%	0.968	Pass
	0.875	0.131	87.0%	0.954	Fail
Naïve Bayes Tree	0.935	0.016	98.3%	0.976	Pass
	0.833	0.091	90.2%	0.949	Fail

Because classification trees can provide guidelines for decision making, they are also known as decision trees. It is important to note that data mining focuses on pattern recognition, hence no probabilistic inferences and Type I error are involved. In addition, unlike regression that returns a subset of variables, classification trees can rank order the factors that affect the retention rate. There are three types of splitting criteria in classification trees: Entropy, Gini, and chi-square. Entropy favors balanced or similar splits. The Gini index tends to favor the largest split or branch of the tree (Han & Kamber, 2006)whereas the chi-square measure is essentially a test of the goodness of fit (Grabmeier & Lambe, 2007).This study employs classification trees such as simple Cart, J48 (Quinlan, 1993),Random Tree (i.e., Variation of J48 that considers K randomly chosen attributes at each node and performs no pruning), LAD Tree (Holmes, Pfahringer, Kirkby, Frank, & Hall, 2002), Bayes Net tree(Pearl, 1985), and Naïve Bayestree (Friedman & Kohavi, 2002). By running all these algorithms on the selected data, most of the trees split on the attribute *"Status Fall"* which represents the GPA score of the students grouped into two main categories "Pass" for GPA >=2.0 and "Fail" otherwise. It is clear why the highest retention rate is obtained for this particular attribute. If the student succeeded at the end of the first semester, then he/she is more likely to pass at the end of the academic year. In fact, this attribute constitute an important decisional factor. Another important attribute was the Math grade in the first semester. The attributes *"AUP Math"* and "*High School Average*" figure among the highest retention rates, depending on the algorithm used. As an example, around 90% of the students having 80% or above on the "*High School Average*" are predicted to pass at the end of the academic year. However, the program administrators wish to have sometimes a prediction of the students' performances before joining the program, in that case only registry attributes will be taken into consideration. Moreover, a comparative study

among all the listed classifiers was derived to select the most appropriate one according to the parameters listed in table 3.The results show that Bayes decision trees with the LAD Tree outperform than other classifiers in performance for both binary and multiclass datasets. Moreover, ROC Area for these algorithms is close to 1. We can see also that precision of Bayes decision trees is higher than that of other classifiers.

3. Conclusion

This paper aims to prevent misclassification of students in a pre-University program. It brings preventive measures once the student is enrolled in the program. It can detect students at risk starting from the first term, in order for the instructors to take appropriate and corrective measures. It has been shown that language skills are primordial for learning Science in general. Mathematics has been found to be highly correlated with Science. Cross-correlation was used to filter out the influence of Math on Science learning. The structure of the dataset along with the classifiers used filled the gap between empirical prediction of student performance and the existing classification and regression techniques. Generally, the education domain offers many interesting and challenging applications for data mining. An educational institution often has many diverse and varied sources of information. There are the traditional databases (i.e., students' information, teachers' information, class and schedule information, alumni information). Secondly, many diverse interest groups in the educational domain give rise to many interesting mining requirements. Finally, data mining present a powerful tool very efficient to predict poor performers and offers ways to encounter these educational issues.

References

Eid, C. (2009). Elementary transcribing. Aston: Aston University.

Fang, N. & Lu, J. (2009). Work in progress: a decision tree approach to predicting student performance in a high enrollement, high impact, and core engineering course. 39th ASEE/IEEE Frontiers in Education Conference, (84-86). San Antonio, Texas.

Friedman, N. & Kohavi, R. (2002). Bayesian Classification. In W. Klosgen, & J. Zytkow, Handbook of data mining and knowledge discovery(282-288). London: Oxford University Press.

Grabmeier, J. & Lambe, L. (2007). Decision trees for binary classification variable grow equally with gini impurity measure and Pearson's chi-square test. *International journal of business intelligence and data mining,* 2(2), 213-226.

Han, J. & Kamber, M. (2006). Data mining: Concepts and techniques (2nd Ed.). Boston, MA.

Holmes, G., Pfahringer, B., Kirkby, R., Frank, E. & Hall, M. (2002). Multiclass alternating decision tree. 13th proceedings of the European Coference on machine learnig (161-172). Helsinky: Springer-Verlag.

Kotsiantis, S., Pierrakeas, C. & Pintelas, P. (2003). preventing student dropout in distance learning systems using machine learning techniques. In V. Palade, R. Howlett, & L. Jain. *Lecture notes in Artificial intelligence,* 2774, 267-274). Berlin, Heidelberg: Springer-Verlag.

Kotsiantis, S., Pierrakeas, C. & Pintelas, P. (2004). Predicting students' performance in distance learning using machine learning techniques. *Applied Artificial Intelligence,* 18(5), 411-426.

Mitchell, B. (2011). Machine learning: International edition. London: McGraw Hill.

Pattanasri, M., Mukunoki, M. & Minoh, M. (2012). Learning to estimate slide comprehension in classrooms with Support Vector Machines. IEEE Transactions on Learning Technologies, 52-61.

Pearl, J. (1985). Bayesian Network:A model of self-activated meory for evidential reasoning. Proceedings of te th conferece of the cognitive science society (329-334). Irvine,California: niversity of California.

Quinlan, J. (1993). C4.5: Programs for machine learning. London: Morgan Kaufmann.

Streifer, P. A. & Schumann, J. A. (2005). Using data mining to identify actionable information:breaking new ground in data -driven decision making. *journal of education for students at risk,* 10(1), 281-293.

The National (2010). News. Retrieved December 15, 2011, from EdArabia: http://www.edarabia.com/22742/1460-teachers-of-uae-public-schools-do-not-hold-university-degrees/.

Xenos, M., Pierrakeas, C. & Pintelas, P. (2002). A survey on student dropout rates and dropout causes concerning the students in the course of informatics of the Open Hellenic University. Computers & Education, 361-377.