On the Design of a DSS for Academic Achievement Prediction

Dr. Mohamed EL-Zeweidy, Dr. Ahmed ElAbbasy

melzeweidy1@gmail.com, ahmed_elabbassy@yahoo.com
Higher Institute for Computer & Information Technology,
Al Shorouk Academy
Cairo – Egypt

Abstract:

This paper tries to examine the relationship between students' overall academic performance *(GPA), students' grade of each subject of the first semester and their the high school grade, then comparing the obtained results to highlight which is more likely to be predicted from the high school grade, would it be the GPA or the grade of each subject by itself. This is done using the Decision Trees algorithm for predicting the academic performance of the first semester for the undergraduate engineering students at the Modern Academy for Engineering (MAE) by using the high school grade as the only input. The data-mining tools were able to achieve levels of accuracy for predicting student performance:*

Decision Trees score for the {pass, fail} set scored 72% for the "Mechanics" which was the least one while the highest score was for "Chemistry" with a score 89%, and as for the GPA grade the score was 80%. For {excellent, very good, good, pass, fail, very bad, absent} set, the score was much less for all of them and had a wide range of variance, it reached a minimum of 34% for "Physics" and a maximum of 62% for "English" while the GPA grade scored 42%.

In this analysis, the Decision Tree was more accurate predicting at the {pass, fail} than at the {excellent, very good, good, pass, fail, very bad, absent} data sets. The results of these case studie give insight into techniques for accurately predicting student performance.

Keywords: predicting the academic performance, Decision Tree, admission system

1. Introduction and related work

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.

The main objective of the admission system is to determine candidates who would likely do well in the university. The quality of candidates

admitted into any higher institution affects the level of research and training within the institution, and by extension, has an overall effect on the development of the country itself, as these candidates eventually become key players in the affairs of the country in all sectors of the economy.

Accurately predicting student performance is useful in many different contexts in universities. For example, identifying exceptional students for scholarships is an essential part of the admissions process in undergraduate and postgraduate institutions, and identifying weak students who are likely to fail is also important for allocating limited tutoring resources as well as strategic programs can be planned in improving or maintaining and assisting students' performance during their period of studies in the institutions.

Since institutes all over the world wants to be sure they are selecting the cream of the crop, many have tried to work on ways for predicting academic performance for their applicants or students. One of those was [8], where they tried to find if the performance is affected by age, gender, Caribbean Examination Council (CXC) qualification, aptitude test score and experience.

Another study that was made by [7], that examines the relationship between students' overall academic performance (GPA) and matriculation requirements performance in first year courses in the Bachelor of Science and Information Technology (BSCIT) program at UTECH. While the study of [4] compares the accuracy of Decision Tree and Bayesian Network algorithms for predicting the academic performance of undergraduate and postgraduate students at two very different academic institutes. They used admissions information, such as academic institute and GPA to predict GPA at the end of the first year. The data-mining tools were able to achieve similar levels of accuracy for predicting student performance: 73/71% for {fail, fair, good, very good} and 94/93% for {fail, pass} at the two institutes respectively.

The project of [10] showed different ways in which student performance statistics can be used to obtain information which may be used in assessing the individual student, course, program and the department in

terms of their performances. A number of data warehousing and data mining concepts were applied to obtaining the required results.

Another very good study was made by [3] where artificial neural networks ANN were used to predict the cumulative Grade Point Averages (CGPA) by using ten inputs which include: UME score, O/level results in Mathematics, English Language, Physics, and Chemistry, Age of student at admission, Time that has elapsed between graduating from secondary school and gaining university admission, Parents educational status, Zonal location of student's secondary school, Type of secondary school attended (privately owned, State or federal government owned), Location of university and place of residence, and Student's Gender. Similar to the previous study but using less inputs is [5] where CGPA was predicted by the students' demographic profile and the CGPA of the first semester, the study compared the accuracy of three predictive models which were Artificial Neural Networks, Decision trees and Linear Regression, and showed that the artificial neural network outperformed the other two with accuracy more than 80%.

On the other hand many others did not use any artificial intelligence for the prediction but used simple statistics depending on other variables, like [1] who used the admission test with the gender, [2] used the SAT with all its divisions like writing, verbal, math etc. [9] used Miller Analogies Test to predict the GPA, and [6] used Graduate Management Admission Test (GMAT) and Undergraduate Grade Point Average (UGPA) for predicting Graduate Student Academic Performance.

The Modern Academy for Engineering MAE has had thousands of applicants per year over the last few years. Of this number approximately one thousand is accepted. The academy offers a Bachelor of Engineering in many majors like computer, mechanics, civil, architect …etc.

And since the selection of students solely depend on the high school grade, this study tries to find out how much does the high school grade is suitable or related alone by itself to the academic performance of each subject in the first semester at the academy by comparing the results of three algorithms which are: Decision Trees, Logistic Regression and Artificial Neural Networks algorithms.

In the following section the overall methodology of the research will be described, from selection of a data- mining platform to modeling of the academic performance prediction problem. Next, the results of the prediction algorithms will be compared and finally, the conclusions.

2. The proposed system

The proposed system will use the high school grade to try predicting both, the overall academic performance (GPA), and the grade of each subject by itself of the first semester, and then these results will be compared to determine which is more likely to be predicted. This step will be done using the Decision Trees algorithm for the undergraduate engineering students at the Modern Academy for Engineering (MAE) by using the high school grade as the only input.

3. Methodology

This section describes the process followed to collect and analyze the academic performance data. First the selection of a data-mining tool, followed by the difficult task of preparing the data for analysis. Then presenting the model of the academic performance prediction, and how the parameters of the prediction algorithms were used in a way to improve the results.

3.1 Data-mining tool selection

Microsoft SQL server with Microsoft Business Intelligence Development Studio were selected due to the wide range of users who uses Microsoft products taking into consideration the good user friendly interface and the good support and online help.

3.2 Data Preparation

The basic selection of data included the student number, the high school grade, grades for each subject and the total grade for the first semester.

The next step was to gather, analyze and prepare the historical data from the academic records of the academy. A total of 1314 records were collected for students admitted in 2006/2007. After cleaning and validating the data, they were reduced to 878 complete records. Figures

1, 2 shows a distribution of some of the data that should be predicted: the student's actual grades at the end of the 1st semester of undergraduate at MAE. The figure represents the classes {excellent, very good, good, pass, fail, very bad, absent} for the "Physics" subject.

4. Data preparation phase

Data was collected from Modern Academy for Engineering as excel sheets, the data was for the preparatory year and contained 1314 students with their grades in all 16 subjects and their high school grades and types which were 12 types, as shown in Figure 3, and Since the source is not clean enough to be loaded into the warehouse, therefore two methods were used to clean our data sample, namely:

1. Delete (ignore) record:

Some students are absent all the year or deceased and others do not have any data for the high school so the best solution was to delete the record.

2. Fill in the missing value manually:

Some students' records might have some missing data, but this can be solved by either calculating the data from other fields or by filling the missing values by the mean of the other similar records.

The records after cleaning and validation were reduced to 878 complete records, Figure 3 shows a sample of the original data collected from MAE, and Figure 4 shows the grades distribution of the actual data listed in Table 1

Figure 5 shows a sample of the original data collected from MAE, and Figure 4 shows the grades distribution of the actual data listed in Table 2.

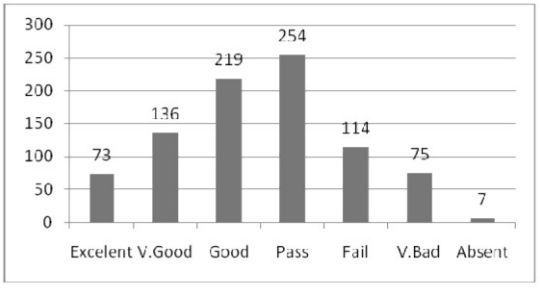


Figure 1
DISTRIBUTION OF ACTUAL GRADES FOR
"PHYSICS" SUBJECT

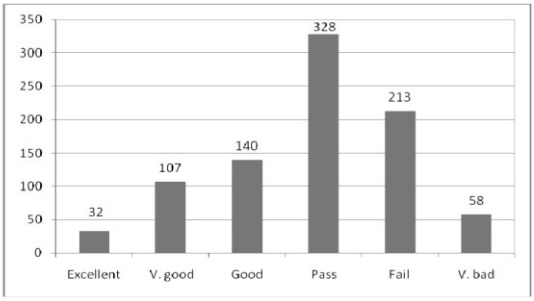
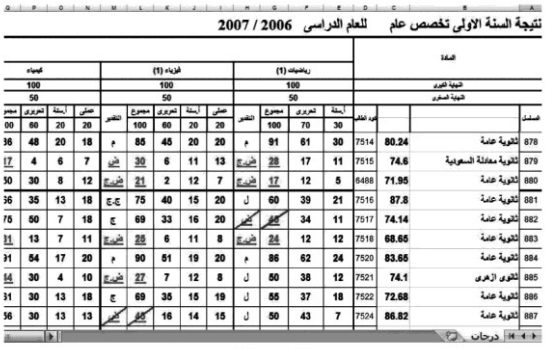


Figure 2
DISTRIBUTION OF ACTUAL GRADES FOR
FIRST SEMESTER "GPA"



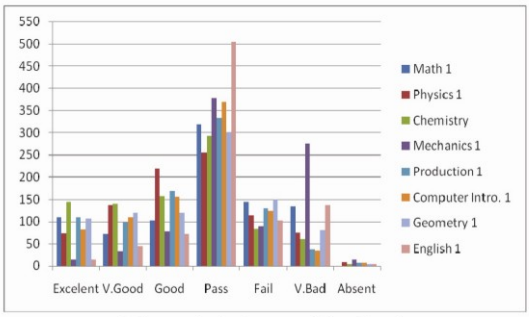
نتيجة السنة الاولى تخصص عام للعام الدراسى 2007 / 2006

المسلسل			المادة															
			النهاية الكبرى															
			النهاية الصغرى															
			رياضيات (1)					فيزياء (1)					كيمياء					
			50					50					100					
			أعلى	أسئلة	تحريرى	مجموع	التقدير	أعلى	أسئلة	تحريرى	مجموع	التقدير	أعلى	أسئلة	تحريرى	مجموع	التقدير	
			30	20	70	100		20	20	60	100		20	20	60	100		
878	ثانوية عامة	80.24	7514	30	61	91	م	20	20	45	85	م	18	20	48	96		
879	ثانوية معادلة السعودية	74.6	7515	11	17	28	ض.ج	13	11	6	30	ض	7	6	4	17		
880	ثانوية عامة	71.95	6488	5	12	17	ض.ج	7	12	2	21	ض.ج	12	8	30	50		
881	ثانوية عامة	87.8	7516	21	39	60	ل	20	15	40	75	ج-ج	18	13	35	66		
882	ثانوية عامة	74.14	7517	11	34	45	كثير	20	16	33	69	ج	18	7	50	75		
883	ثانوية عامة	68.65	7518	12	12	24	ض.ج	8	11	6	25	ض.ج	11	7	13	31		
884	ثانوية عامة	83.65	7520	24	62	86	م	20	19	51	90	م	20	17	54	91		
885	ثانوى أزهرى	74.1	7521	12	38	50	ل	8	12	7	27	ض.ج	10	4	30	64		
886	ثانوية عامة	72.68	7522	18	37	55	ل	19	15	35	69	ج	18	13	30	61		
887	ثانوية عامة	86.82	7524	7	43	50	ل	15	14	16	45	كثير	13	13	30	56		

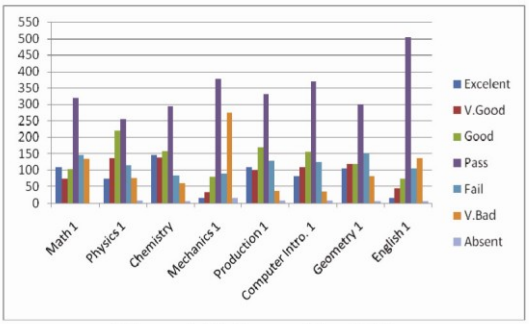
Figure 3
Sample of the original data collected from MAE

Subject \ Grade	Ex.	V.G.	G.	P.	F.	V.B.	A.
Math	109	72	102	318	144	133	0
Physics	73	136	219	254	114	75	7
Chemistry	144	139	157	292	83	59	4
Mechanics	14	32	78	377	89	274	14
Production	109	98	168	332	128	37	6
Comp. Intro.	81	109	155	369	124	34	6
Geometry	105	119	119	300	150	80	5
English	14	45	72	504	103	136	4

Table 1
ACTUAL DESRIBUTION OF FIRST SEMESTER GRADES



(Figure 4-a) Grouped by Grade



(Figure 4-b) Grouped by Subject

Figure 4

GRADES DESTRIBUTION for the {excellent, very good, good, pass, fail, very bad, absent} set

5. Results And Analysis

The proposed system objective is to determine which is more likely to be predicted from the high school grade, would it be the overall academic performance (GPA) or the grade of each subject by itself. Decision Trees algorithm was selected for the task of predicting the academic performance of the first semester for the undergraduate engineering students at the Modern Academy for Engineering (MAE) by using the high school grade as the only input.

The first step in building the system is to create the data mining model. Figures 5, 6 show the mining structure for Subjects, and for the GPA.

After creating the model, the next step is to train the model. Training the model means running the model against training data set using a particular algorithm. Training is usually the most time-consuming step. The algorithm may iterate over the training dataset a few times to find the hidden patterns. The training process was done automatically by the used engine which splits the data for training and testing.

The last step was to use the model in the predictions process on new datasets. The accuracy score of the prediction is shown in figures 7 to 10.

The accuracy of the results for the MAE predictions using the selected algorithm is shown in the following figures. The results show that the predictions for the {pass, fail} set are noticeably more accurate than for the {excellent, very good, good, pass, fail, very bad, absent} set, which is expected given the much larger number of grades to be predicted. The results also show that the selected algorithm scored an average of 82% for all subjects for the {pass, fail} set, while for the other set the average score was 44%, as shown in Tables 2, 3.

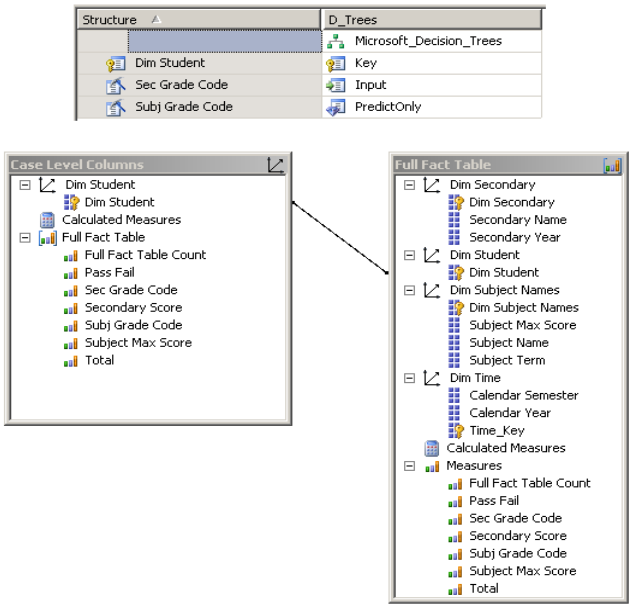


Figure 5 Mining Structure for Subjects Using Decision Trees Algorithm

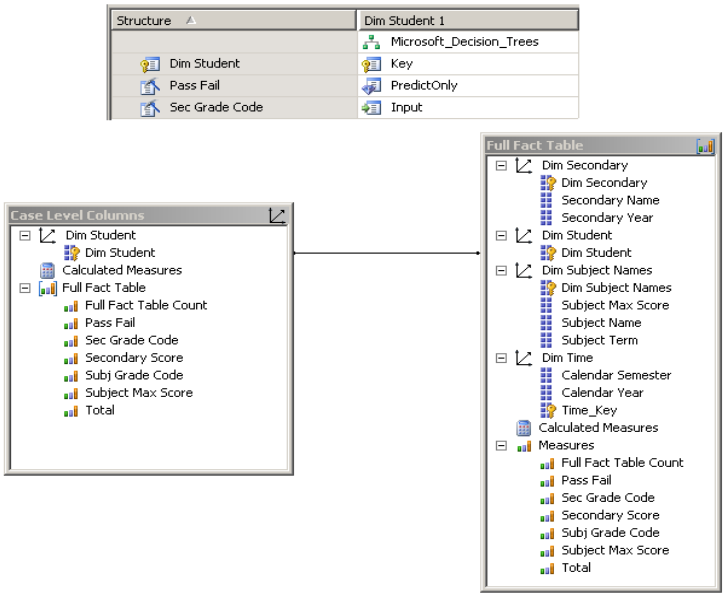


Figure 6 Mining Structure for the Total Grade Using Decision Trees Algorithm

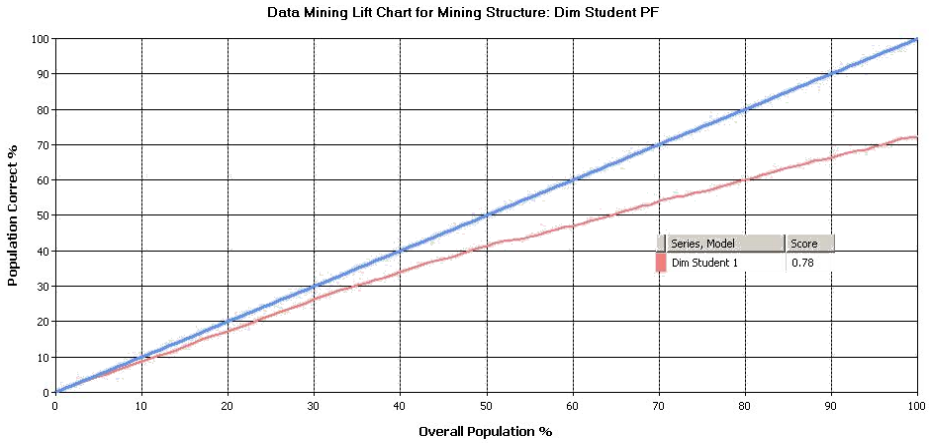


Figure 7 Accuracy Score Of "MATH" For The {Pass, Fail} Set Using Decision Trees Algorithm

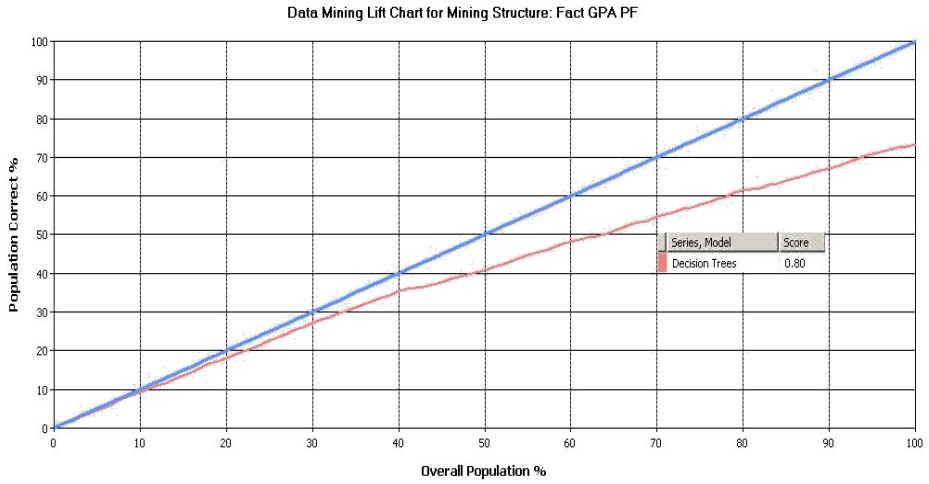


Figure 8 Accuracy Score Of "Total Grade" For The {Pass, Fail} Set Using Decision Trees Algorithm

Math	Physics	Chemistry	Mechanics	Production	Computer Intro.	Geometry	English	Average	Total Grade
78%	85%	89%	72%	85%	87%	77%	81%	82%	80%

Table 2 Decision Trees Accuracy Score For The {Pass, Fail } Set

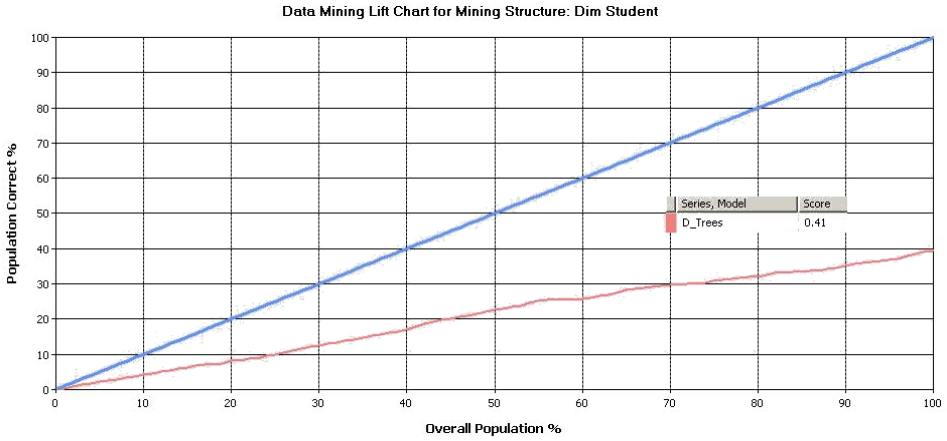


Figure 9 Accuracy Score Of "Math" For The {Excellent, Very Good, Good, Pass, Fail, Very Bad, Absent} Set Using Decision Trees Algorithm

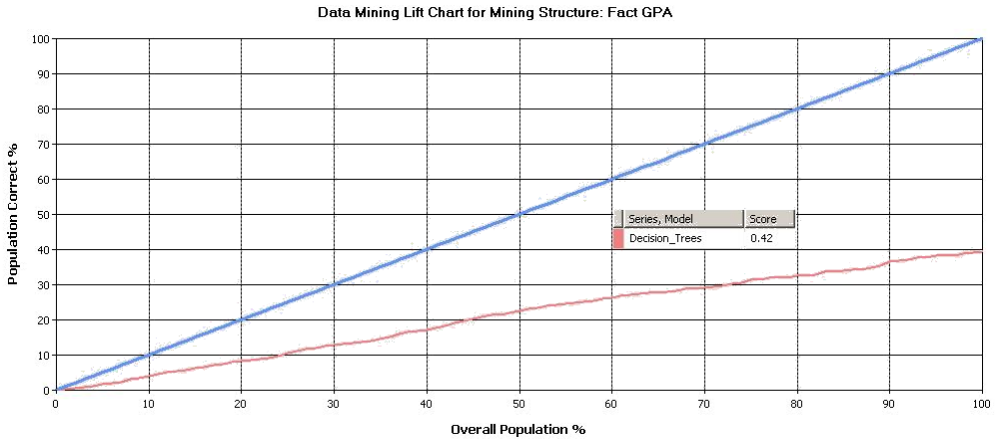


Figure 10 Accuracy Score Of "Total Grade" For The { Excellent, Very Good, Good, Pass, Fail, Very Bad, Absent} Set Using Decision Trees Algorithm

Math	Physics	Chemistry	Mechanics	Production	Computer Intro.	Geometry	English	Average	Total Grade
41%	34%	39%	50%	42%	46%	35%	62%	44%	42%

Table 3 Accuracy Score For The {Excellent, Very Good, Good, Pass, Fail, Very Bad, Absent} Set Using Decision Trees Algorithm

6. Conclusion

The research work is devoted to develop an efficient model that can provide knowledge discovery and data mining to discover the influence of high school grades over the success level of the student for each subject in a particular faculty.

The objective of the article is to build a system that help undergraduate students to choose the best field or branch of study that suits their skills and abilities depending on their high school grade.

On the other hand this system can be used in more than one faculty so that the student can know which faculty he/she is more likely to succeed in and have better achievements in it.

Moreover the system provides the administration with the information needed to help the students who need academic assistance.

This goal is achieved through the following:

Students' data has been collected from the Modern Academy for Engineering, the original data was found in two sources, the first was on paper documents and the other was in an Oracle database. Data from both sources was extracted and put into the form of Excel sheets to be ready for the staging process which includes data cleaning, integration and transformation.

The basic selected data included the student number, the high school grade, grades for each subject and the total grade for the first semester. Also other data that existed in the original database and where thought of to be important for future work and in the same time will not put any overload on the work, the data includes the type of high school and the details of the subjects grade details (practical score, course work and final exam grades).

A total of 1314 records were collected for students admitted in 2006/2007. The data is for the preparatory year and contains students with their grades in all 16 subjects (8 subjects per semester) and their high school grades and types which are 12 types, after cleaning and validating the data, they were reduced to 878 complete records.

The research tried to find which is more likely to be predicted from the high school grade, would it be the total grade or the grade of each subject by itself. This is done using the Decision Trees algorithm for predicting the academic performance of the first semester for the undergraduate engineering students at the Modern Academy for Engineering by using the high school grade as the only input. The data-mining tools were able to achieve levels of accuracy for predicting student performance that was discussed in the previous chapter.

From the results shown in tables 2 and 3, high school grade is good for predicting the {pass, fail} for almost all subjects, but to predict the excellent students it did not produce high enough score that can be depended on, so it is clear that more factors are needed to be taken into consideration. As for the total grade, it scored a little less than the average score of the subjects where it scored 80% for the {pass, fail} set, and 42% for the {excellent, very good, good, pass, fail, very bad} set, while the average score for subjects was 82%, 44% for the two sets respectively.

From the results shown in table 2 and table 3, high school grade is good for predicting the {pass, fail} for almost all subjects, but to predict the excellent students it did not produce high enough score that can be considered. As for the GPA, it scored a little less than the average score of the subjects where it scored 80%, 42% for the two sets while the average score for subjects was 82%, 44% respectively.

It is clear from the obtained results that more factors are needed to be taken into consideration beside the high school grade. Such factors may include Age of student at admission, Time that has elapsed between graduating from secondary school and gaining university admission, Parents educational status, Zonal location of student's secondary school, Type of secondary school attended (privately owned, State or federal government owned).

7. Future work

Future work can include testing the same data using more algorithms like Logistic Regression, Neural Network, Naive Bayes, Association Rules and Clustering. On the other hand, a deeper analysis can be done by testing the grades of each subject in the high school and see if it might

give a high prediction score for any subject in the collage which could then result in the selection of students not only by the high school grade but also by demanding a minimum grade in certain subjects to ensure the quality of students.

The present work uses only the "Total Grade" and the "Subjects Grades", the future work could investigate the followings:

Other metrics such as "Gender" and "Subject Grade Details", in addition to the "Total Grade" used in this work could be investigated.

To study the effect of the "high school subjects' grades" attribute according to the attribute sensitivity, this could then result in the selection of students not only by the high school grade but also by demanding a minimum grade in certain subjects to ensure the quality of students.

Gathering much more data should be of great value and should help producing more accurate prediction with higher score.

When having enough data some attributes could be investigated like the name of the high school itself or the student gender and address.

References

1. Aavo Luuk, Kersti Luuk. Predicting students' academic performance in Aviation College from their admission test results. 2008.
2. Christopher M. Cornwell, David B. Mustard and Jessica Van Parys. How Does the New SAT Predict Academic Achievement in College. 2008.
3. V.O. Oladokun, A.T. Adebanjo and O.E. Charles-Owaba. Predicting Students' Academic Performance using Artificial Neural Network - A Case Study of an Engineering Course. *The Pacific Journal of Science and Technology,* Volume 9, Number 1, 2008.
4. Nguyen Thai Nghe, Paul Janecek, and Peter Haddawy. A comparative analysis of techniques for predicting academic performance. 37th ASEE/IEEE Frontiers in Education Conference, 2007.
5. Zaidah Ibrahim and Daliela Rusli. Predicting Students' Academic Performance – Comparing Artificial Neural Network, Decision Tree and Linear Regression, 2007.

6. Nathan R. Kuncel, Marcus Crede, LISA L. Thomas. A Meta-Analysis of the Predictive Validity of the Graduate Management Admission Test (GMAT) and Undergraduate Grade Point Average (UGPA) for Graduate Student Academic Performance. Academy of Management Learning & Education, Vol. 6, No. 1, 2007.

7. Paul Golding and Opal Donaldson. Predicting Academic Performance. 36th ASEE/IEEE Frontiers in Education Conference, 2006.

8. Paul Golding and Sophia McNamarah. Predicting Academic Performance in the School of Computing & Information Technology (SCIT). 35th ASEE/IEEE Frontiers in Education Conference, 2005.

9. Nathan R. Kuncel, Sarah A. Hezlett and Deniz S. Ones. Academic Performance, Career Potential, Creativity, and Job Performance - Can One Construct Predict Them All. Journal of Personality and Social Psychology Vol. 86, No. 1, 2004.

10. Dervis Z. Deniz and Ibrahim Ersan. Academic DSS for Student, Course and Program Assessment. International Conference on Engineering Education, 2001.

11. Christopher M. Cornwell, David B. Mustard and Jessica Van Parys. How Does the New SAT Predict Academic Achievement in College. 2008.

12. Nathan R. Kuncel, Sarah A. Hezlett and Deniz S. Ones. Academic Performance, Career Potential, Creativity, and Job Performance - Can One Construct Predict Them All. Journal of Personality and Social Psychology Vol. 86, No. 1, 2004.

13. Nathan R. Kuncel, Marcus Crede, LISA L. Thomas. A Meta-Analysis of the Predictive Validity of the Graduate Management Admission Test (GMAT) and Undergraduate Grade Point Average (UGPA) for Graduate Student Academic Performance. Academy of Management Learning & Education, Vol. 6, No. 1, 2007

14. Sanjay Soni – UNISYS, Zhaohui Tang - Microsoft , Jim Yang – Microsoft. Performance Study of Microsoft Data Mining Algorithms, 2002.