PERFORMANCE ANALYSIS ON STUDENTS' GPAs AND COURSE AVERAGES USING DATAMINING TECHNIQUES

Osman GÜRSOY [1]

Mehmet Akif YAMAN [2]

Emine YAMAN [3]

Abstract: Universities play very important role for an individual's success in life by giving necessary education to the people. Education offers pupils teaching skills that get ready them physically, mentally and socially for the world of work in later life. Having well educated people provide the development of country. In this paper we worked performance analysis on student s' GPAs and course averages .We think that we can contribute quality of education at the university by determining the important factors which affects students' GPAs then improve this factors. In order to find out this factors, we have used datamining tools to derive hidden attributes playing important role in education life at universities.

We prepared a collection of data including student GPA, CGPA, number of courses that a student registered per semester, course averages, students CGPA average , number of students in a course , the number of semester of students at the university , etc..

Key Words: Education and University, Data Mining, Decision Tree, Association Rules, Student Performance Analysis.

Özet: Üniversiteler, insanlara gerekli olan eğitimi verdiğinden, insanların hayatlarında önemli rol oynamaktadırlar. Eğitim, ögrencilerin daha sonraki yaşamlarında fiziksel, mental ve sosyal olarak hazir olmaları için gerekli öğrenme yetilerini tesis eder. Iyi eğitim almış insanlar ülkenin kalkınmasına yardımcı olurlar. Eğitimin bu denli önemli olmasından dolayı,eğitim kalitesini artırabilmek için, bu yayında öğrencilerin not ortalamasına ve sınıf not ortalamasına etki eden faktörler üzerinde çalıştık. Bizler, üniversitedeki eğitim kalitesinin artmasının,öğrenci notlarına etki eden faktörlerin tespit edilip ,onlar üzerinde gerekli çalışmaların yapılmasıyla sağlanabileceği düşüncesindeyiz. Bu faktörleri tespit edebilmek için, bu araştırmada veri madenciliği yöntemlerini kullandık.

Bu araştırma için hazırlamış olduğumuz veri seti, öğrencilerin not ortalaması,öğrencilerin dönem bazlı not ortalamaları, dersi alan öğrenci sayısı, öğrencilerin okulda geçirmiş olduğu dönem sayısı, öğrencilerin o dönem almış oldukları ders sayısı gibi birçok faktörü içermektedir.

Anahtar Kelimeler: Eğitim ve Üniversite, Veri Madenciliği, Karar Ağacı, İlişkilendirme Kuralları, Öğrenci Performans Analizi

[1] MSc. Student, International University of Sarajevo,IT Asisstant, ogursoy@ius.edu.ba

[2] Ph.D. Student, International University of Sarajevo, Faculty of Engineering and Natural Sciences, myaman@ius.edu.ba

[3] Ph.D. Student, International University of Sarajevo, Faculty of Engineering and Natural Sciences, eyaman@ius.edu.ba

Introduction

The quality assesstment of education and statistical factors playing role on the students performance at a higher education is very important. In this project we have studied the statistical factors playing role on student semester GPA and Course Averages at International University of Sarajevo (IUS). In order to find out this factors, we have used datamining tools to derive hidden attributes playing important role in education life at IUS.

We prepared a collection of data including student GPA, CGPA, number of courses that a student registered per a semester (Number of Courses Enrolled by a Student), course averages (CourseAvg), students CGPA Average (CGPAAvg), number of students in a course (Number of Students Enrolled in a Course), the number of semester of students at the university (Semester Number), etc. We have used the data for 10 semesters at IUS stored in the Student Information System, years between 2005 and 2009.

The data provided by the IUS, is a kind of 'no-name' data. All the student, course and lecturer information is labeled as student1..studentN, course1..courseN, lecturer1..lecturerN. Semesters and grades in the data set are real. This kind of approached is followed due to security reasons and to protect third party's privacy.

2.Data Mining and Techniques

Data mining (DM), also known as ''knowledge discovery in databases" (KDD), is the process of discovering meaningful patterns in huge databases (Han & Kamber, 2001). In addition, it is also an application that can provide significant competitive advantages for making the right decision. (Huang, Chen, & Lee, 2007).

In this study we benefit from the following data mining techniques :

1. **Association Rules: Association** Rules Mining Techniques are used to find interesting relations between attributes in a database. Associton rules can have one or more output attributes but the traditional production rules can not. Moreover, an output attribute for one rule can be an input for another rule (Roiger,Geatz, 2003).Association rule mining discovers relationships among attributes in databases, producing if-then statements concerning attribute-values (Agarwal, Imielinski, & Swami, 1993).

   An X -> Y association rule explains a relative correlation between items (attribute-value) in a database with values of support and confidence. The confidence of the rule is the percentage of transactions that contains the consequence in transactions that contain the antecedent. The support of the rule is the percentage of transactions that contains both antecedent and consequence in all transactions in the database.(Christobal, Sebastian & Enrique, 2007)

   Association rule mining is a famous technique for market basket analysis, which provides at finding buying patterns for supermarket, mail-order and other customers. With using mining association rules, marketing analysts try to match sets of products that are frequently bought together, so that certain other items can be inferred from a shopping cart containing particular items. Association rules can often be used to design marketing promotions, for example, by appropriately arranging products on a supermarket shelf and by directly suggesting to customers items that may be of interest (Chen, 2007).

2. **Decision Tree :** Decision trees are powerful classification algorithms that are becoming increasingly more popular with the growth of data mining in the field of information systems applications in mediçine and healthcare. As the name implies, this technique recursively separates observations in branches to construct a tree for the purpose of improving the prediction accuracy (Delen, Fuller,McCann,Ray, 2009).

There are many type of algorithms for decision trees.These algorithms usually occupy a greedy strategy that grows a decision tree by making a series of locally optimum decisions about which attribute to us efor partitioning the data(Tan, Steinbach & Kumar, 2006). The most famous decision tree algorithms are Hunt's algorithm, ID3, C4.5, CART and Microsoft Decision Tree Algorithm which we used in that work.

The Microsoft Decision Trees algorithm is a classification and regression algorithm which is provided by Microsoft SQL Server Analysis Services to work in predictive modeling of both discrete and continuous attributes.

For discrete attributes, the algorithm tries to make predictions based on the relationships between input columns in a dataset. The values, known as states, of those columns are used to predict the states of a column that you designate as predictable. Specifically, the algorithm identifies the input columns that are correlated with the predictable column.

As a second it can work on continuous attributes, the algorithm uses linear regression to determine where a decision tree splits.

If you have more than one column set to predictable in dataset, or if the input data contains a nested table that is set to predictable, the algorithm builds a separate decision tree for each predictable column.

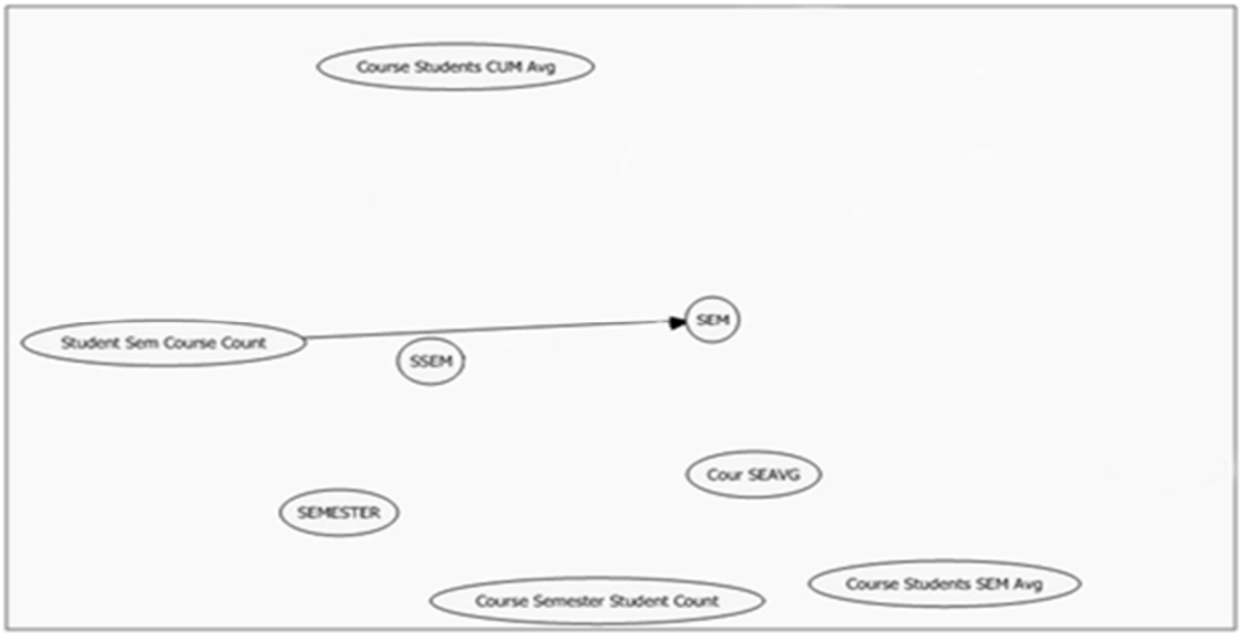
As a overview, we can say that to able to use Microsoft decision tree algorithms we need at least one input column which might be discrete or continuous, a key column which should be a unique across the all column, and at least one prediction column.

3.Performance Analysis

3.1. Performance Factors on Students' GPA

Dependency network of our data showed that some attributes are effecting the students' GPA as well as CGPA. In Figure 1 , it is shown that the most important affecting factor for the student GPA is number of courses that a student enrolled in a particular semester.

Figure 1: Number of Courses that Student Registered



The GPAs of students at IUS show that students with high GPAs usually are those who also enrolled in high number of courses. We think that the policy for course enrollment at IUS by 2009, regarding the number of courses that can be enrolled by a student in a particular semester has a great effect on this result.

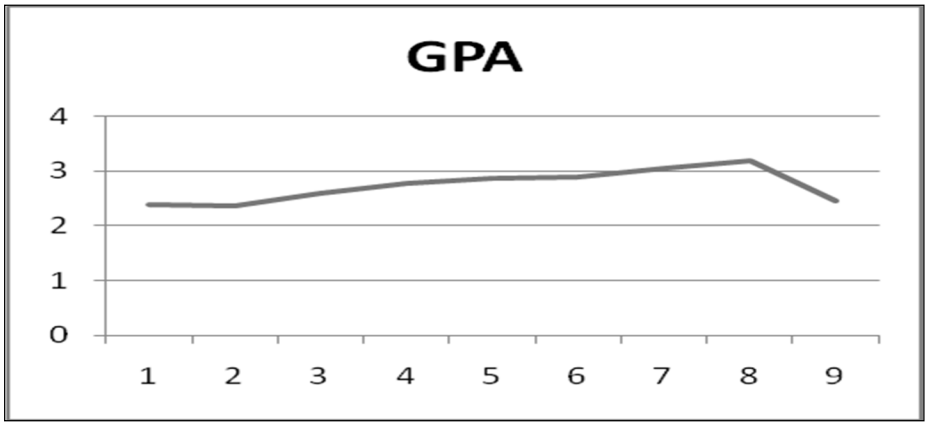
> ***"****... A student who has a grade point average of 2.25 or higher may be allowed to take a maximum of 36 ECTS credits (6 courses) and students with a grade point average of 3.00 or higher may be allowed to take 42 ECTS credits (7 courses) in one term. The ECTS credit load limit can be changed when necessary upon the approval of the student advisor and the Dean.* ***". (IUS Regulation, Article 15)***

Second most important factor on a student GPA is his/her semester number at IUS. It shows that students are improving their GPA while they are getting experience at the university. As a result of Micsosoft's Decision Tree Algorithm within the SQL Analysis Server gives the following formula for GPA, based on the student semester.

GPA = 2.522+0.121*(StudentSEM-3.145)

We also checked this result through SQL query in order to see the real result in our data and it proved this idea that students are improving their GPA while he/she is having experiences at the university except those who exceed the normal education time. Figure 2 shows the graph of GPA averages based on student semesters at IUS.

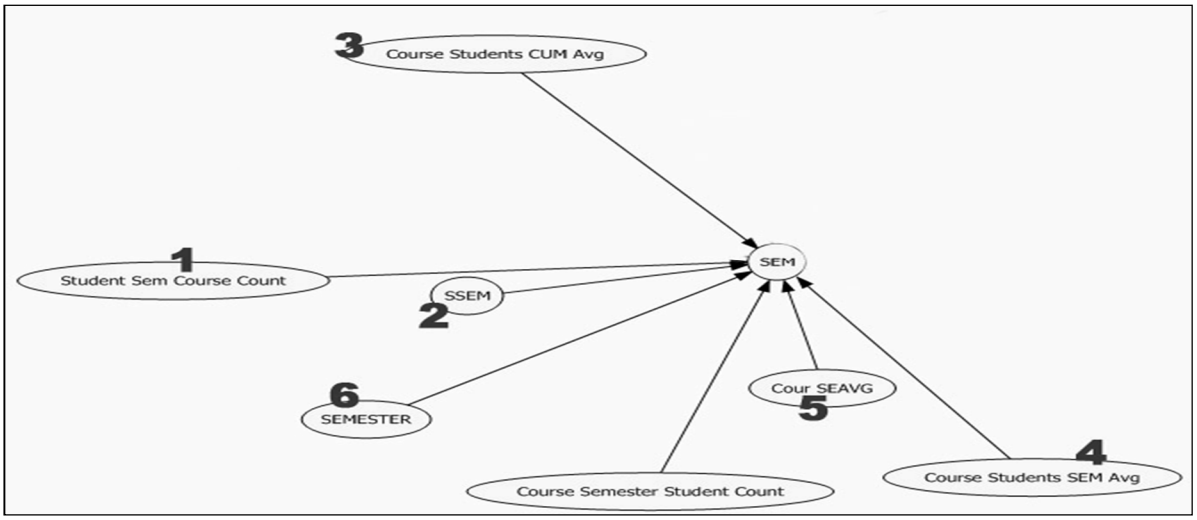
Figure 2: GPA Average Based on Students Semesters at IUS



Remaining dependency network in our data is shown below in ascending order. Figure 3 also shows this remaining dependency network in our data.

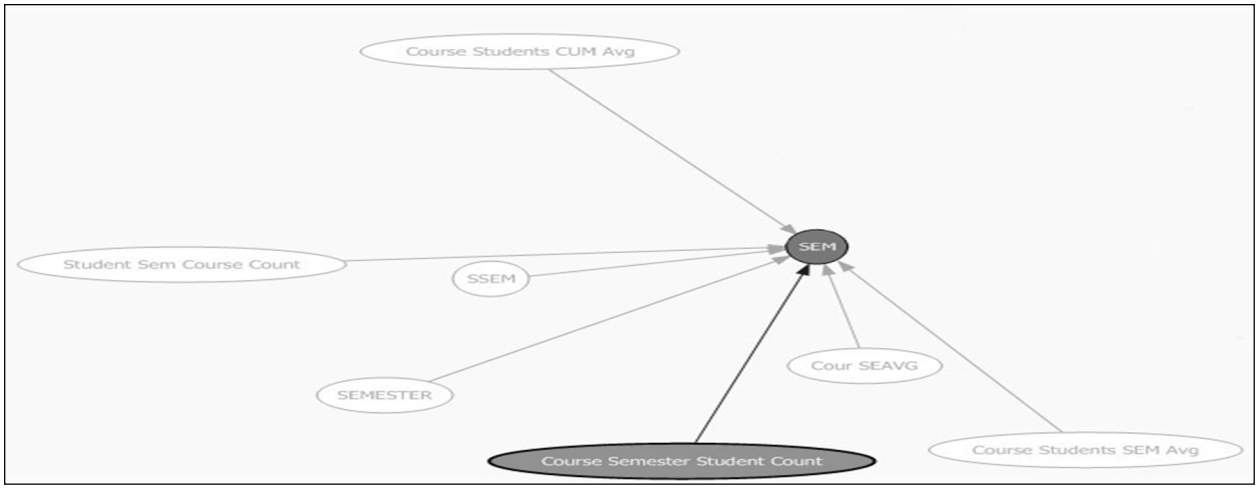
3- CGPAA of the students in courses
4- GPAA of the students in courses
5- Course Averages.
6- Semester (Spring or Fall)

Figure 3: Remaining Dependency Network of the data



Finally, we have also found an interesting relation between GPA and number of students in the courses. As we may suppose that less number of students in a course may increase the student GPA but it seems that there is no relation between number of students in the courses and student GPA at IUS. Figure 4 shows this unexpected relationship in order 7.

Figure 4: Number of Students Enrolled in a Course has minimum affect on Student GPA



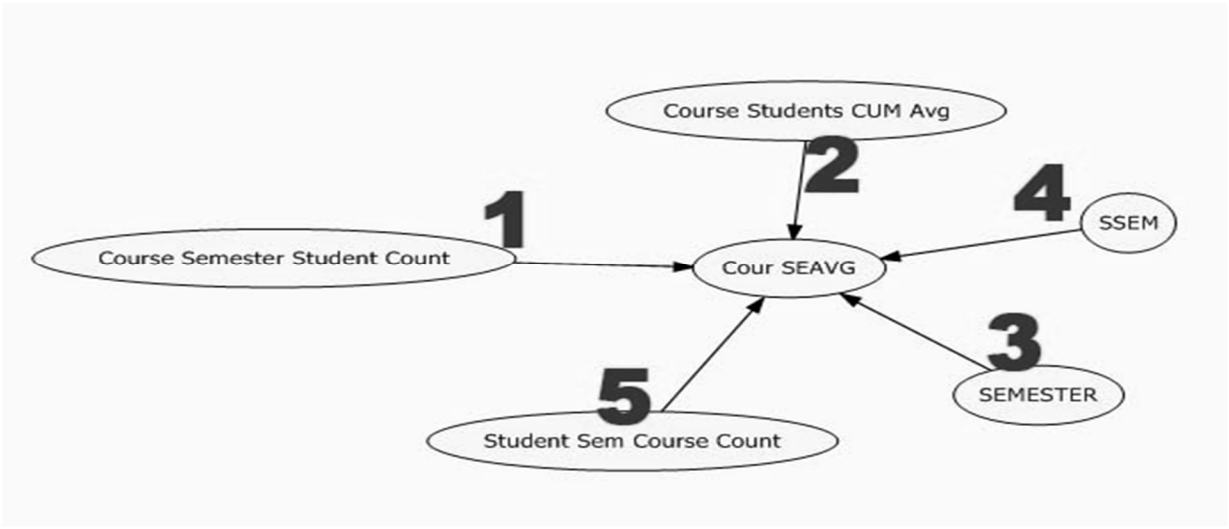
3.2.Performance Factors on Course Averages

Now we are ready to find out the fundamental factors affecting the course averages at IUS. Our dependency network for course averages shows us that the most important factor on a course average is the number of students enrolled in that course. We see that high number of students in a course has a negative effect on course average. As a result of Micsosoft's Decision Tree Algorithm within the SQL Analysis Server gives the following formula for Course Average, based on the number of students enrolled in that course in a particular semester.

Course AVG = 2.548-0.006*(Number of Students Enrolled in the Course-41.981)

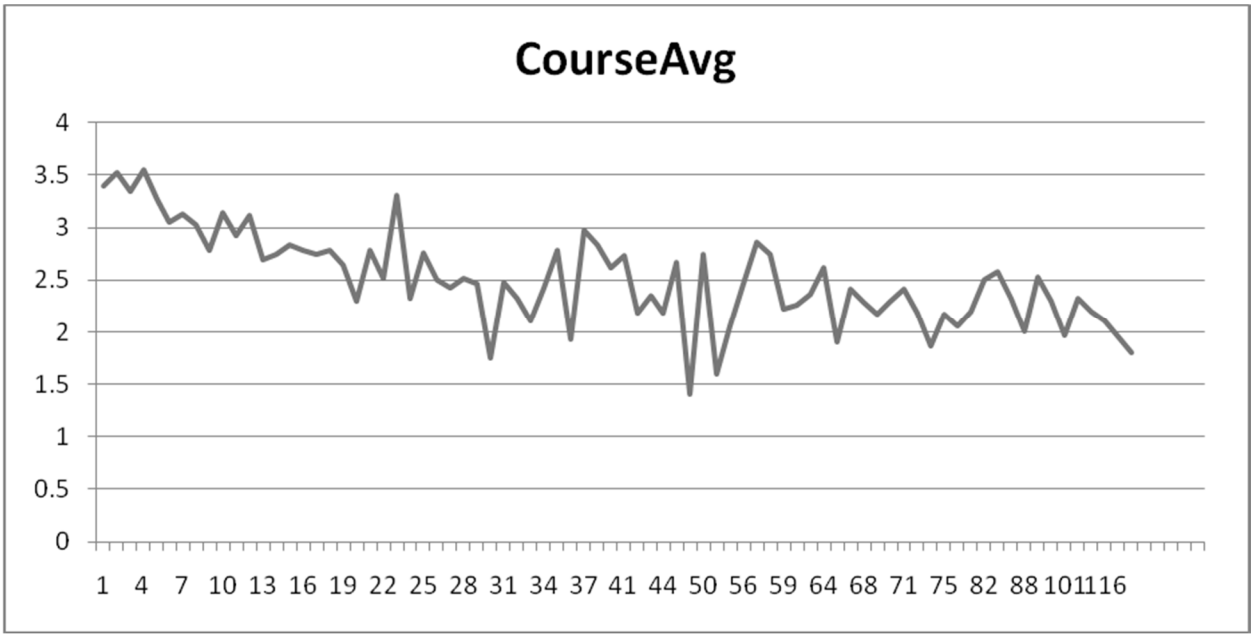
Figure 5 shows the entire dependency network regarding course average factors.

Figure 5: Dependency Network on Course Averages



We also checked this result through SQL query in order to see the real result in our data and it proved this idea that high number of students in a course has a negative effect on course average.

Figure 6: Course Averages based on number of students enrolled in courses



Remaining dependency network for course average is shown below in ascending order. Figure 5 also shows this remaining dependency network in our data.

2- CGPAA of the students in courses
3- Semester (Spring, Fall, Summer)
4- Student Semester Number.
5- Student Semester Course Count

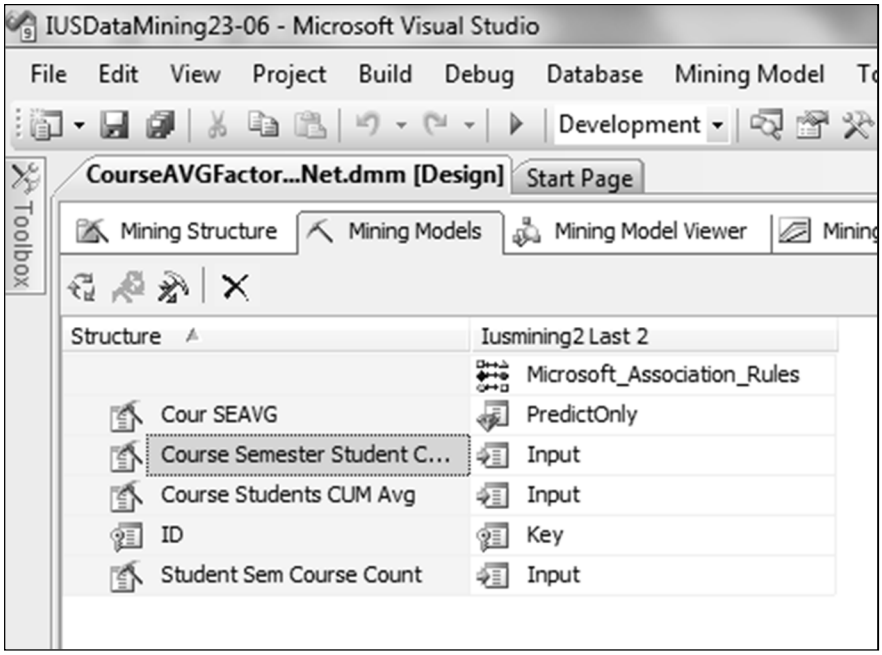
3.3. Performance Analysis Through Association Rules

Let's remember the importance of Association Rules that are also mentioned on Microsoft MSDN web page:

> *"The Microsoft Association Rules Viewer in Microsoft SQL Server Analysis Services displays mining models that are built with the Microsoft Association algorithm. The Microsoft Association algorithm is an association algorithm for use in creating data mining models that you can use for market basket analysis. "*
>
> *(*http://msdn.microsoft.com/en-us/library/ms174768.aspx#BKMK_Rules, Last Visited Date: 30-06-2010)

First, we have checked the association rules for Course Averages by providing the following attributes in figure 7.

Figure 7: Association Rules Mining Model for Course Average



With this mining tool we find the following rules which are also interesting for this study. The number of rules that are generated by the algorithm with minimum probability of 0.40 and importance of 0.16 was 30. We present here only first 5 rules with high probability.

Figure 8: Association Rules on Course Averages

Prob.	Impor-tance	Rule
1	0.677	Number of Students Enrolled in a Course= 77 - 102, Course Students CGPA Avg = 2.4075999936 - 2.686713284 -> Course AVG= 2.5938004816 - 3.171927008
1	0.827	Course Students CGPA Avg >= 2.9824726872, Number of Students Enrolled in a Course= 24 - 48 -> Course AVG>= 3.171927008
1	0.824	Number of Students Enrolled in a Course>= 102, Course Students CGPA Avg = 2.1229584564 - 2.4075999936 -> Course AVG= 1.9008336854 - 2.1949252164
0.682	0.664	Number of Courses registered by Student< 4, Course Students CGPA Avg >= 2.9824726872 -> Course AVG>= 3.171927008
0.647	0.456	Number of Courses registered by Student< 4, Course Students CGPA Avg = 2.4075999936 - 2.686713284 -> Course AVG= 2.5938004816 - 3.171927008

Even though the probability of first 3 rules is high, the importance of these rules shows us the usefulness of these rules. So, It seems that the 2.rule is the most useful rule that the association algorithm derived from the data.

This algorithm also provided us a number of item set which the algorithm identified as frequently found together. The Figure 8 shows the percentage of item sets that occurs in the total data:

Figure 8: Percentage of Item Sets

%	Item Set
37	Number of Students Enrolled in a Course< 24
33	Number of Courses registered by Student= 7 - 8
31.2	Course Students CGPA Avg = 2.4075999936 - 2.686713284
30.2	Number of Courses registered by Student= 6 - 7
27.4	Course AVG= 2.1949252164 - 2.5938004816
26.8	Number of Students Enrolled in a Course= 24 - 48
22.6	Course AVG= 2.5938004816 - 3.171927008
21.8	Course Students CGPA Avg = 2.686713284 - 2.9824726872
20.9	Course Students CGPA Avg = 2.1229584564 - 2.4075999936
19.1	Number of Courses registered by Student>= 8
17.9	Number of Students Enrolled in a Course= 48 - 77
16.6	Course Students CGPA Avg < 2.1229584564
16.4	Course AVG= 1.9008336854 - 2.1949252164
15.5	Number of Courses registered by Student= 4 - 6
14.9	Course AVG>= 3.171927008
13.4	Course Students CGPA Avg = 2.686713284 - 2.9824726872, Number of Students Enrolled in a Course< 24
12.3	Course AVG= 2.5938004816 - 3.171927008, Number of Students Enrolled in a Course< 24
11.2	Course AVG>= 3.171927008, Number of Students Enrolled in a Course< 24
11.2	Course Students CGPA Avg = 2.4075999936 - 2.686713284, Number of Students Enrolled in a Course< 24
11	Number of Students Enrolled in a Course= 77 - 102
10.6	Number of Courses registered by Student= 7 - 8, Number of Students Enrolled in a Course< 24
10.5	Number of Courses registered by Student= 6 - 7, Number of Students Enrolled in a Course< 24
10.4	Course AVG= 2.5938004816 - 3.171927008, Course Students CGPA Avg = 2.4075999936 - 2.686713284
10.2	Course Students CGPA Avg = 2.4075999936 - 2.686713284, Number of Courses registered by Student= 7 - 8
10	Number of Students Enrolled in a Course= 24 - 48, Course AVG= 2.1949252164 - 2.5938004816
9.98	Number of Students Enrolled in a Course= 24 - 48, Course Students CGPA Avg = 2.4075999936 - 2.686713284
9.79	Number of Courses registered by Student= 6 - 7, Course Students CGPA Avg = 2.4075999936 - 2.686713284
9.63	Course AVG= 2.1949252164 - 2.5938004816, Course Students CGPA Avg = 2.4075999936 - 2.686713284
9.52	Course Students CGPA Avg >= 2.9824726872

9.18	Number of Courses registered by Student>= 8, Number of Students Enrolled in a Course< 24
9.15	Course Students CGPA Avg >= 2.9824726872, Number of Students Enrolled in a Course< 24
9.09	Number of Students Enrolled in a Course= 24 - 48, Number of Courses registered by Student= 7 - 8
9.09	Course AVG= 2.1949252164 - 2.5938004816, Number of Courses registered by Student= 7 - 8
8.65	Number of Students Enrolled in a Course= 24 - 48, Number of Courses registered by Student= 6 - 7
8.12	Course AVG= 2.1949252164 - 2.5938004816, Number of Courses registered by Student= 6 - 7

Conclusion

In this paper we have studied the performance factor on students GPA and course averages at International University of Sarajevo (IUS). We use two different mining tools within the Microsoft SQL Analysis Server. By using decision tree we concluded that the most important statistical factor on student GPAs is the number of courses that students enrolled in a particular semester at IUS.

We also concluded that the most important statistical factor on a course average at IUS is number of students that enrolled in that particular course.

We also derived some interesting rules from the data using association rules data mining algorithm.

Output of this study can help academic advisors during the registration periods and similar studies can be used to built perfect class demography aiming high success rates in the future.

References

Agarwal, R., Imielinski, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. In Proceedings of the ACM SIGMOD international conference on management of data, Washington DC, USA (pp. 1–22).

Chen, M. C. (2007). Ranking discovered rules from data mining with multiple criteria by data envelopment analysis. Expert Systems with Applications, 33, 1110–1116.

Cristobal Romero , Sebastian Ventura, Enrique Garcıa(2007). Data mining in course management systems: Moodle case study and tutorial. Computers & Education 51 (2008) 368–384

Dursun Delen , Christie Fuller, Charles McCann, Deepa Ray, (2009), Analysis of healthcare coverage: A data mining approach, Expert Systems with Applications 36 (2009) 995–1003

Han, J., & Kamber, M. (2001). Data mining: Concepts and techniques. San Francisco,CA, USA: Morgan Kaufmann.

Huang, M. J., Chen, M. Y., & Lee, S. C. (2007). Integrating data mining with casebased reasoning for chronic diseases prognosis and diagnosis. Expert Systems with Applications, 32(3), 856–867.

Richard J.Roiger, Michael W.Geatz, (2003). Data Mining , a tutorial based primer, p-49,ISBN 0-201-74128-8.

P. N. Tan, M. Steinbach, V. Kumar (2006).Introduction to Data Mining,p-151, ISBN 0-321-42052-7.