(IJISE) 2017, Vol No. 5, Jan-Jun **e-ISSN: 2454-6402, p-ISSN: 2454-812X**

# ENHANCING THE EFFICACY OF LEARNING MANAGEMENT SYSTEM BY ANALYSING AND DEVELOPING EDUCATIONAL DATA MINING TOOLS AND TECHNIQUES , 2017

**Drishti Arora**

## ABSTRACT

*In recent years, data mining approach has been used for optimizing the mass of information on educational organizations and extracting knowledge from these data. This approach develops knowledge discovery techniques from the data in the learning environment, especially from the university students. Applying data mining techniques on the educational data will result in the extraction of useful information and knowledge from them. This information then could be used to improve the learning process and increase the efficiency and performance of students. Therefore, the purpose of this research is to provide a method for increasing the efficiency of learning management system using educational data mining. To this end, a dataset was collected from the students in an E-learning center and the techniques such as classification, clustering and association rules mining were implemented on the specified parameters. Finally, the obtained results are used to improve the management of virtual education systems.*

## 1. INTRODUCTION

Reaching a favorable situation in any educational system, whether micro or macro, requires that educational researches and evaluations be performed in the respective fields and planning and learning management be done. On the other hand, high volumes of information may lead to incorrect decisions in education fields, especially at E-learning centers. For example, a large number of student's educational problems are rooted in factors which cannot be observed by the teaching staff directly. This alone would result in wrong decisions. In recent years, data mining approach has been used in order to optimize the mass of information on educational organizations and extract knowledge from these data. This new research develops knowledge discovery techniques from the data in the learning environment, especially from the university students. The main focus of the present study is the use of data mining techniques and providing a method for analyzing the effects of educational programming on student's performance in E-learning centers. In this study, the three following main questions will be answered:

Is the course schedule including class date, exam date, • and class duration effective on student's grades level?

To what extent is student's achievement related to the • level of student's interaction with the teacher and virtual education system including presence or absence in the class, the number of

message sex changed between teacher and student, and the rate of using available resources in the system?

Are the rate of answering the questions posed by the • teacher in the virtual education system as well as quiz and mid-term exam scores related to student's final grades level?

In previous researches conducted by other researchers, the parameters used to predict and improve the performance of E-learning students included mid-term grades, final grades, class attendance records, interactions with the system, gender, place of residence, etc. The effective management factors such as course and exam schedule have not been investigated so far in any studies.

In the present study, in addition to considering the impact of management decisions, various new parameters are used to predict student's success and improvement of e-learning system performance. These parameters include the number of message sex changed between teacher and student, the number of times available resources in the system are viewed, quiz scores and answering teacher's questions.

## 2. LITERATURE REVIEW

The term data mining is synonymous with one of the terms including knowledge extraction, data collection, data verification and even data dredging which in fact describes Knowledge Discovery in Data bases (KDD). So the idea which data mining is based upon is an important process of identifying potentially useful, fresh, and ultimately understands able patterns of data. In practice, the two primary goals of data mining include prediction and description. Prediction involves the use of some variables or fields in the data set to predict unknown or future values of other variables. Description focuses on finding human-interpretable-patterns describing the data. The basis of data mining has been divided into two categories of statistics and artificial intelligence. Data mining process consists of five main stages; each stage plays an important role in achieving the desired knowledge. These stages include:

Data collection.•

Data pre-processing.•

Estimation and evaluation of model.•

Pattern discovery (data mining).•

Result validation.•

Educational data mining is an emerging scientific field which develops methods form in discovering unique learning environments data and uses these methods for a better understanding of students and the learning environment in which the processes of teaching and learning take place. Various researchers divide educational data mining methods in to different areas. The following classification can be considered as the collection of the categorizations:

Prediction.•

Clustering.•

Relationships analysis and extraction.•

Educational content mining and online interactions.•

Statistical analyses, visualization and decision support • systems.

In the early years, researches in this area have been more focused on the analysis and extraction of relationships. Statistical methods can be used for summarization and expression of descriptive characteristics of a data set. These methods can also be used for modelling the current trends in data and inferring hidden processes and patterns. In1, the difficulty level of presented problems and exercises in the virtual education system was analyzed using statistical analyses. The results of these analyses were used to improve e-learning environment.

Clustering methods classify the data based on their similarities. Clustering methods are usually used for grouping students in E-learning in terms of different features. In2, students were grouped in terms of features associated with the learning style and the results were used for improving the interactions between the same groups and preventing student's problems. In3, behavior patterns of students in an interactive learning environment were extracted using clustering methods.

In4 classification methods have been used to determine the response of different groups of students to different educational strategies. In5 student's performance and their final grades were predicted using a combination of classification methods. In6, students were classified into two groups, error-prone and not error-prone, in terms of the usage pattern and the results of this classification were used to investigate the factors which caused the errors. In7, classification methods have been used to identify students with low motivation and discovering solutions for curing them and preventing their withdrawal.

One of the common methods of classification is using decision tree production methods. Decision tree has numerous applications in the field of education and learning due to its simplicity and high interpretability. The produced decision tree can also be used to derive the If-Then-Else rules which may lead to the extraction of interesting data about the student's us age and the results.

Association rules are used for describing the relationship between different feature sin the database. The results of these algorithms are often presented as a set of rules $X \rightarrow Y$, in which X and Yare sets of attributes. In8, association rules have been used for constructing are commender-system for students. These systems recommend

different educational activities with respect to student's usage pattern and suggest shortcuts for deleting unnecessary educational resources to the students. In9, association rules were used tore solve student's problems in learning environment sand to provide advice to them. In5 a

method was offered for determining the distinctive features of students in terms of performance.

Sequential patterns are in fact specific form of association rules in which time and arrangement of object sin the data base are considered as parameters. These methods are widely used in E-learning and particularly for investigating the behavior patterns of students in the learning environment. In10, sequential patterns mining was used for the evaluation of student's activities and personalization of educational content.

Outlier detection methods look for records in the data base which does not differ much from the expected values. These methods are often used for data clean sing and removal of noise and false data. But sometimes they can be used to identify growing trends and rare cases among the data set. For example, these methods can be used to detect a normal student in e-learning. In11, these methods have been used for the detection of abnormal behavior of students in the earning environment.

## 3. INTRODUCING THE PROPOSED METHOD

In recent years, the use of data ware house and business intelligence tools have been presented as a developing area in e-learning systems and a number of researchers started to develop different learning environments to support decision making and equipping them to data ware housing and analytical processing tools12. These analyses and acquiring comprehensive knowledge of the factors affects the business (such as key performance indicators, patterns of behaviour, organization an inclination, evaluation criteria, etc.) for analysts, managers and other executives in the organization.

Businesses in elegance strategies use various tools and technologies to achieve these goals. Data warehouse are used as strategic data repository. Also, online analytical processing tools, case reports tools, and data mining algorithms are used, respectively, for the study of organizational data, data analysis and visualization, and meaningful patterns extraction from the high volume of organizational data. This paper presents are al strategy or implemented in the last mentioned area for monitoring and analysing the behaviour of students in virtual education environment. Figure 1 shows the architecture of business intelligence strategy presented in this paper. As seen in the figure, data are integrated from various sources such as registration selection system, educational planning system and learning management system and stored in a star schema in the data warehouse. The data are then available through the business intelligence and analytical processing tools. Generally, the process of designing and implementing a business intelligence strategy is a complex process that requires an accurate determination of the implementation stages and outputs of each stage.

In this study, an iterative approach similar to the strategy presented in13 has been followed which is based on five steps:

In the first stage, business requirements and a list of questions and favourable analytical reports for managers of e-learning centre were collected and they were then presented as a

requirements document. For this purpose, a series of interviews were done with managers, and academic and technical experts in e-learning centres of University of Science and Technology. The interviewees were asked to express their needs in the form of questions related to each of the entities involved in the teaching and learning process. Issues raised by stake holders included categories such as process of using the virtual education environment, learning process, course and classroom. Also, a comprehensive analysis of the implementation tools of data warehouse and analytic processing was performed at this stage. Then different strategies of business warehouse systems along with open source business intelligence tools were examined. Finally, the analytical context of Microsoft Business Intelligence Studio was selected for the purpose of implementation.

In the second stage, an integrated conceptual model of available data was created in accordance with the requirements

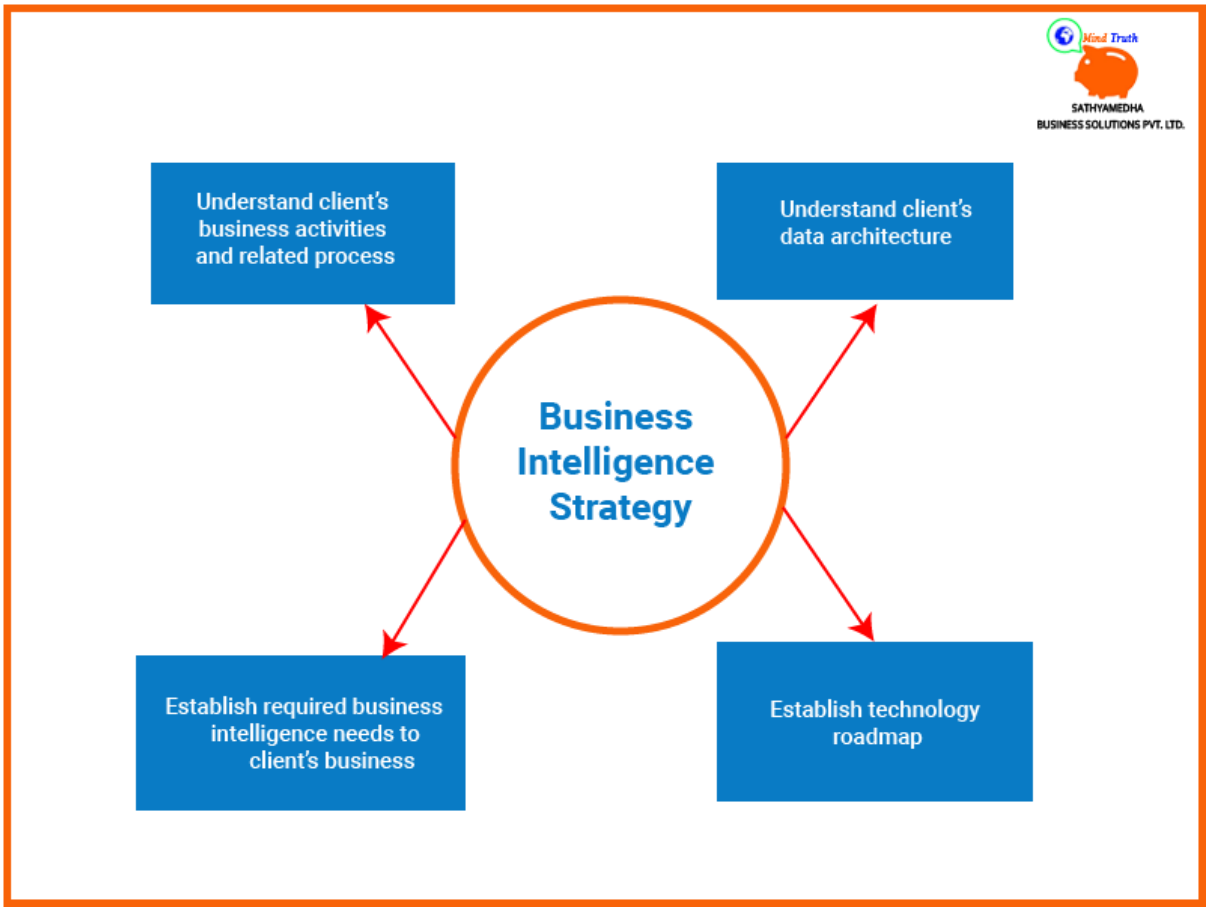


**Figure 1.** The schematic outline of business intelligence strategy presented in the research.

specified in the previous step. It was then used in the star schema form as the main form of data storage in the data warehouse. In this model, the key entities are kept in fact tables and different aspects stored in dimension able are associated with them. Six dimension tables were considered for the data warehouse in order to maintain scripture data required for analysing the

facts and examining key performance indicators from different aspects. All fact tables are associated with these tables through the foreign key:

Dim-User: This table stores student's data such as • username, name, age, gender, city, field of study, Grad Point Average (GPA) of previous academic degree, current GPA and the first and last access to the system. Meanwhile, GPA data are stored both numerically and through the rating scale (excellent, good, fair and poor) is stored.

Dim-Course: These tables to res course data such as • course name, course ID, course type, field of study, and teacher.

Dim-Activity: This table stores data related to each • educational activity and available learning resources in the system. Among these data are activity type, activity name, ID, address, course, start date and completion date estimation.

Dim-Date: This table stores data related to days of the • year. These include year, term, month name, month number, week, day, weekday name, weekday number, day of year and holiday.

Dim-Time: This table stores data related to time of • student connection to the system during the hours, minutes and seconds.

Dim-Session: This table stores data related to the • student's location of connection and speed of internet connection when connecting to virtual education system. These data are extracted using the IP address when connecting to the virtual classroom system which includes the connection location and whether it is inside or outside the country.

It must be mentioned that dimension tables are the main foundation of conceptual modeling. Thus, the more complete are the data stored in them, the more effective is their analysis from different aspects.

The preliminary analysis of the available data sources, their availability, quality and limitations were investigated at this stage so that they will be used for feasibility analysis in the next stage. In the third stage, the data from sources identified in the first phase are extracted and after cleaning and integration processes, they are placed in the star schema form in the data ware house. After collecting the historical data in the data warehouse, in the fourth stage, analytical cubes, dimension tables and Log_Fact table were designed in the Microsoft SQL Server environment in terms of key indicators. Then, student's us age data were transferred in to these tables once they were extracted and cleansed. Finally, analytical reports were designed for monitoring and verification. In Table 1 Logical design of Log_Fact table is shown.

In the fifth stage, reports designed in the previous stage were implemented on business intelligence tool and the user in terface required for accessing them was created in different management levels. Moreover, system testing and troubleshooting were performed at the end of the implementation and then, a program was designed for system support and maintenance. At the end of the system implementation and establishment stages, new indicators of student performance were identified according to managers' feedback. Then, changes were made in

the data warehouse design based on these new indicators. Applying some of these changes required that the design and implementation stages be performed completely. Figure 2 displays a diagram of a star schema for the implemented tables along with their relationships with each other.

**Table 1.** Logical design of Log_Fact table

| Attribute | Description |
|---|---|
| Id | Main key |
| Date_Key | The foreign key which refers to a specific date in the date dimension |
| Time_Key | The foreign key which refers to a specific time in the time dimension |
| User_Key | The foreign key which refers to a specific user in the user dimension |
| Course_Key | The foreign key which refers to a specific course in the course dimension |
| Session_Key | The foreign key which refers to a specific course in the session dimension |
| Activity_Key | The foreign key which refers to a specific activity in the activity dimension |

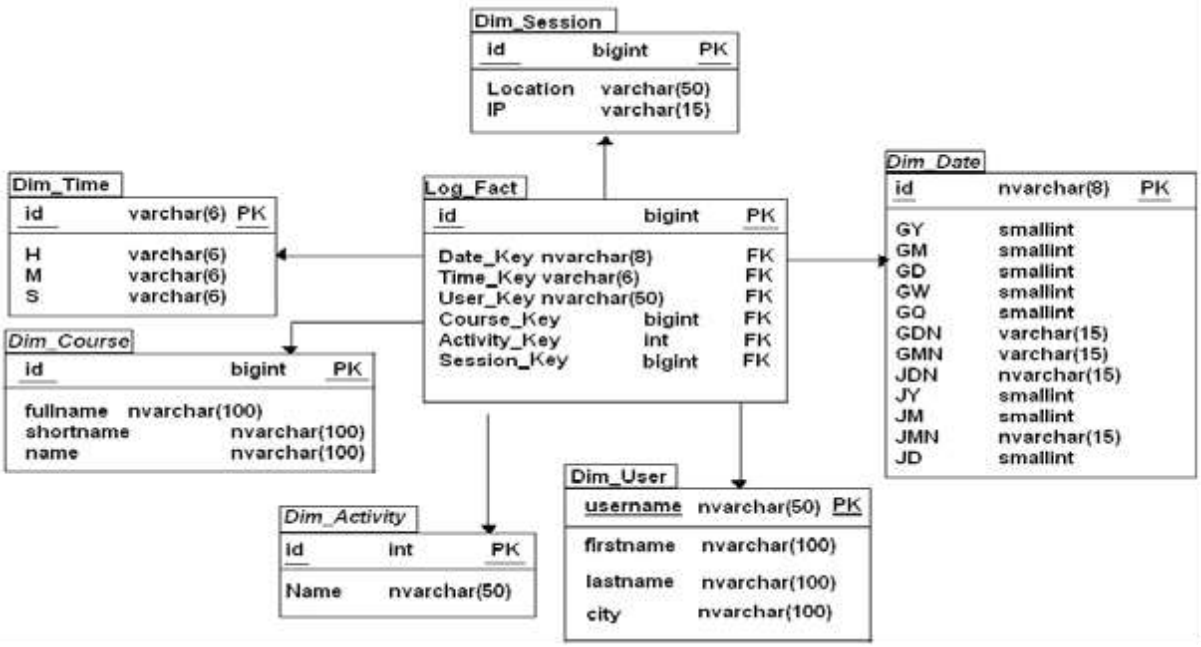


**Figure 2.** Implementation of dimension and fact tables in a star schema.

Implementation of data cube in the SSAS environment follows a straight forward process. First, it is required that various tables which are involved in creating data cubes be selected as a Data Source View. Then, the relevant dimensions are selected from the New Dimension option and different columns are selected from the dimension tables. Then, using the New Cube option and selecting the fact table, data related to the intended measures can be selected from the fact table columns and different cumulative functions can be defined for the calculation.

Meanwhile, through converting the cube file to a CSV file, different data mining techniques can be performed on the data using various data mining tools. Among the available data mining tools, Clementine tool was selected for the data analysis due to its use of up-to-date algorithms, having a user–friendly environment and a variety of reporting tools. Methods of classification, clustering, association rules discovery, and outlier analysis were among examined methods in this study.

As noted, the input for the data mining section is different data cubes which were produced in the preprocessing section. On the other hand, the output for this section can as charts or text files that using them in an evaluating recommender system can provide a tool for the educational evaluation of virtual students.

## 4. ASSOCIATION RULE MINING

The purpose of association rule mining is to discover interesting relationships among items in a set of items. A priori algorithm is one of the most important algorithms of association rule mining. Based on this property, ak –items et is frequent only if trouble shooting sets are frequent. The result obtained from this property is that the super-patterns of anon-periodic pattern are also non-periodic. This property will allow the A priori-based algorithm to set Asiden on-periodic k-item set sin (k + 1)-item sets.

In association rules discovery, two measurement criteria, namely, support and confidence, are applied for reducing and controlling the results and mining the desired results. The support value for a candidate is calculated as follows:

$$\text{Support} = P(A \cap B) = \frac{\text{number of records which inclue A, B}}{\text{number of total records}}$$

$$\text{Confidence} = P(B \mid A) = \frac{P(A \cap B)}{P(A)} = \frac{\text{number of records which inclue A, B}}{\text{number of records which inclue A}}$$

### 4.1 Clustering

Clustering means finding groups of objects in such a way that the object sin a group is similar as much as possible and also are different from the objects in other groups. Students can be grouped based on different characteristics, for example, class participation using this technique. Each group should be treated appropriately. In this study, the K-Means algorithm was used for students clustering. In a simple variant of this method, first some points are randomly selected as for the number of required clusters. Then, they are attributed to one of these clusters in the data with respect to the resemblance (similarity) level, and thus new clusters are produced. By repeating this procedure, new centres are calculated through data averaging in each repetition

and the data are attributed to new clusters again. This process is continued until no change occurs in the data. The following function is proposed as the objective function.

$$j = \sum_{j-1}^{k} / \sum_{i-1}^{n} || x_i^{(j)} - c_j^{2} ||^{/2}$$

Where || || xxis the criterion of the distance between points and cj is the centre of j-the cluster.

4.2 Classification

Classification problem aims to identify the characteristics that indicate the group to which each instance belongs. This pattern is used for the understanding the existing data and also predicting the behavior of new data. In data mining, classification models are created through examining previously classified data and a predictive (pattern) is discovered inductively. One of the most important data classification methods are decision trees. Decision tree is a method for displaying a set of rules that will lead to a category or value. The main components of a decision tree are: decision nodes, branches and leaves. The decision tree uses a technique for the correct displaying of the data or knowledge i.e. applying the rules in which if-the n rules are used as follows:

if Condition then Conclusion

In this paper, CART algorithm has been used for generating decision tree rules.

4.3 Outlier Analysis

The common problem in data mining is finding outliers and anomalies in the data base. Outliers are points which are rare to happen in a data model. Since outliers are rare, they could be indicative of bad data, malicious contents or faulty collection. There are numerous ways to identify outliers. One way is identifying based on the model in which it is assumed that the data has a parametric distribution. Such methods do not work well in areas for high dimension and finding the true model is difficult. To overcome these limitations, researchers have turned to non-parametric methods which use the distance between appoint and it's nearest neighbour as a criterion of abnormality.

## 5. RESULTS

The first question that was raised after the evaluation of student's participation in different learning activities is: which activities are more effective on students learning? In other words, is it possible to identify different activities in terms of their importance in educational effectiveness? The second question that arises in this context is: is it possible to extract rules for correlating the grades of quizzes, prerequisite courses and previous GPA to student's final grades? For this purpose, in addition to association rules extraction, decision tree mining technique has been implemented and the results are presented. Finally, the role of other effective factors was investigated. They included hardware, used facilities, student's data in the learning process, and the use of virtual education system.

If GPA of each semester is considered as the assessment scale for students learning, the effect of each key indicators of student's performance on determining their final grades is evaluated. In this regard from the available data in the data warehouse related to student's use of virtual education system, a table was extracted from the learning activities of 227 students in the fields of computer engineering and industrial engineering over the four semesters along with their GPA. Then, the knowledge discovery methods were to be applied on this table. Table 2. shows the design diagram of this table.

| Field name | Description |
|---|---|
| Course1 | Course 1 grade |
| Course2 | Course 1 grade |
| Course3 | Course 1 grade |
| Course4 | Course 1 grade |
| Resource View | The average number of viewing the educational content |
| Virtual Classroom | The average number of participation in virtual classrooms |
| Archive View | The average number of viewing archives of recorded sessions |
| Forum Read | The mean number of Forum view |
| Forum Post | The average number of discussions in the forum |
| Discussion Post | The average number of responses to the questions posed in the forum |
| Assignment View | The mean number of assignment view |
| Assignment Upload | Mean of responses to assignment |
| Average4 | Average of four courses under study |
| Average Term | Term GPA |

In order to obtain an intuitive understanding and a more accurate analysis of student performance based on the final grade C5.0 algorithm was used to build decision trees with different objectives; Apriori algorithm was also used for discovering association rules. However, Apriori algorithm was not efficient enough in examining the effect of student's activities on the final grades, and as a result, use full association rules were not discovered. Figure 2 shows the first three levels of decision tree extracted from student's data along with the Avg target node which is the parametric display of Average Term value. At the highest level of the tree, the grades of four courses will have a significant impact on GPA and this is a very obvious conclusion. In the next three levels of the tree, important influencing factors are, respectively, the average of assignment view and the average of forum view. Based on analyzing this level, it can be stated that if the ratio of assignment views is 4, most likely the student's GPA level will be between good and fair. Following this level, the average of forum view is considered as another influencing factor. Based on this node if the forum view value is from 2 to3, most likely student's GPA in the end of the semester will be fair and/or from 12 to15. The next effective factor on the decision tree is the average of presence in virtual classes and viewing there sources in the virtual education system. According to this tree, if the

student's presence in the virtual class room is good (mapped to value 2) or low (mapped to value 4), his final grade point average is predicted to be, respectively, fair and weak, he predicted. Also, if there source view value is very low; the student's GPA is predicted to be weak.

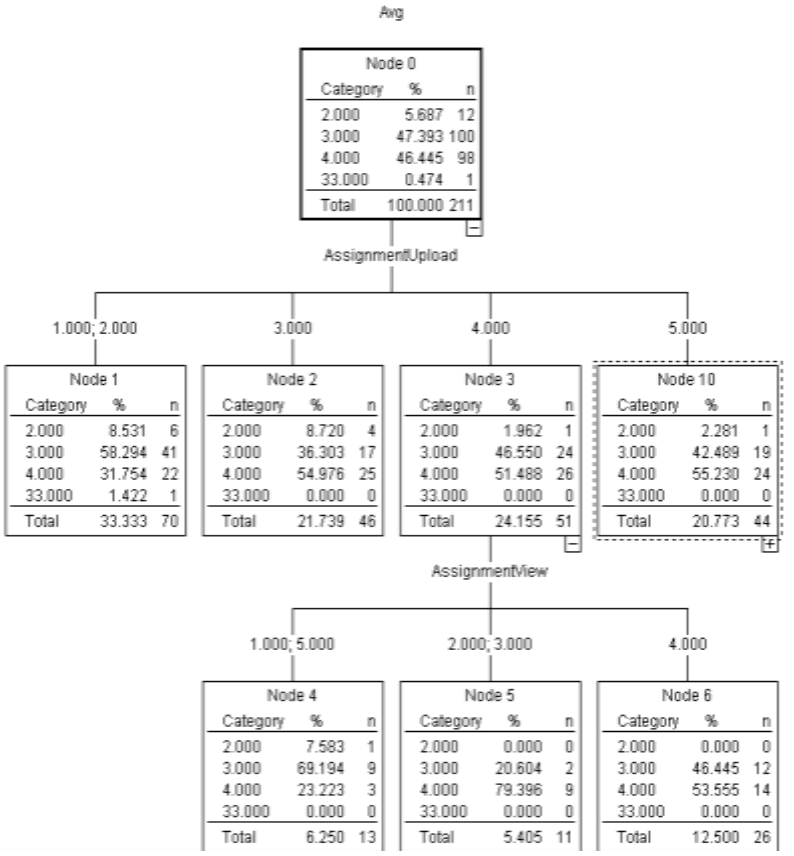


**Figure 3.** The first three levels soft decision tree (part 1).

In the following, factors which may affect the level of student activities have been added to Table 2. They include the gender of the student, student's residence code (triple regions), employment status (employed or unemployed), the speed of used internet connection, and using a laptop. Figure 3 shows the decision tree which was generated based on the selected field sand the Resource View target node, i.e. the average of viewing the educational resources. Based on the generated decision tree, the most important influencing factors on viewing the educational content are the amount of viewing the contents of the forum, the speed of internet connection and student's gender. Accordingly, those students who have good internet connection speed (Map 2) are more likely to view educational resources. Also, female students viewed the educational resources in a greater proportion.

Figure 4 shows the decision tree which was generated based on the selected fields and the Archive View target node, i.e. the average of viewing archives of recorded sessions.

The third level of the tree shows the direct relationship between employment status and the use of computers along with the amount of viewing archives of recorded sessions. According to

this tree, employed students (map0) will view the archives of recorded sessions many times. Also, students who use laptops will view the archives of recorded sessions numerous times, i.e. more than 50%. Figure 5 shows the decision tree which was generated based on the selected fields and the Virtual Classroom target node, i.e. or the average of participation in virtual classrooms.

The speed of internet connection is on the highest level of the tree as the most important factor for participating in virtual classroom sessions. As it can be observed medium to high speed connections are reasons for more participations in classroom. Also, employed students participate less in classrooms.

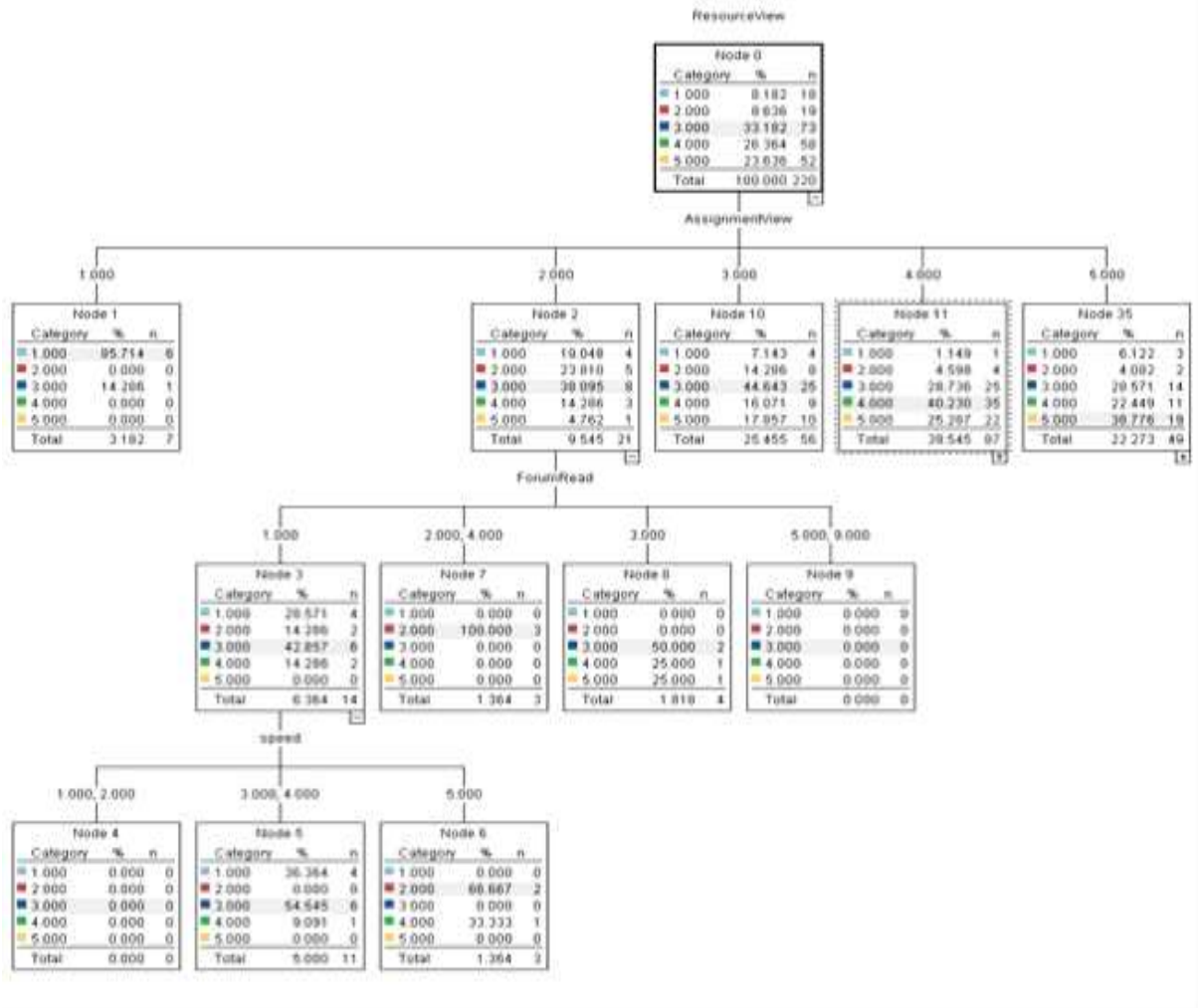


**Figure 4.** The generated decision tree with the Resource View target node.

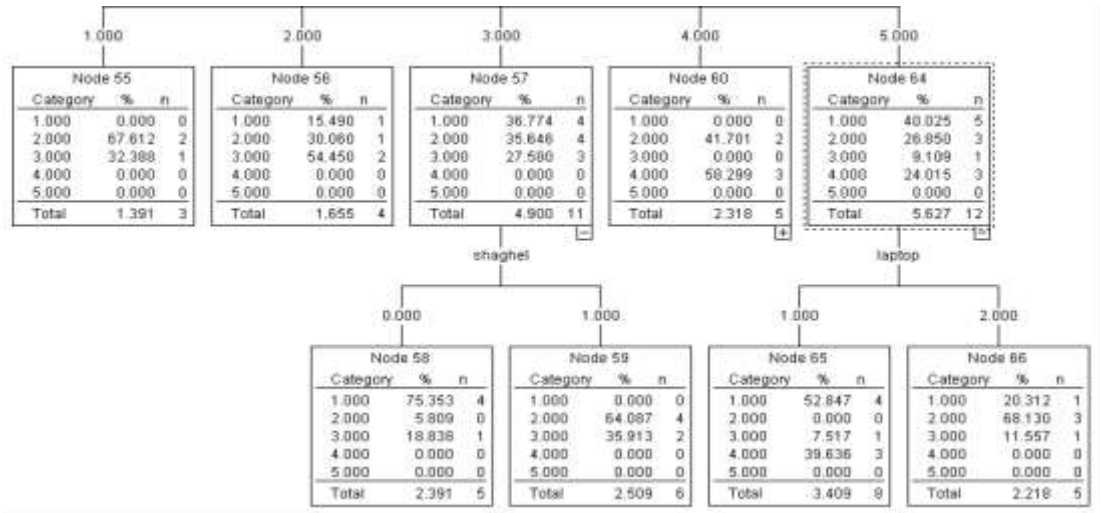


**Figure 5.** The generated decision tree with the Archive View target node.

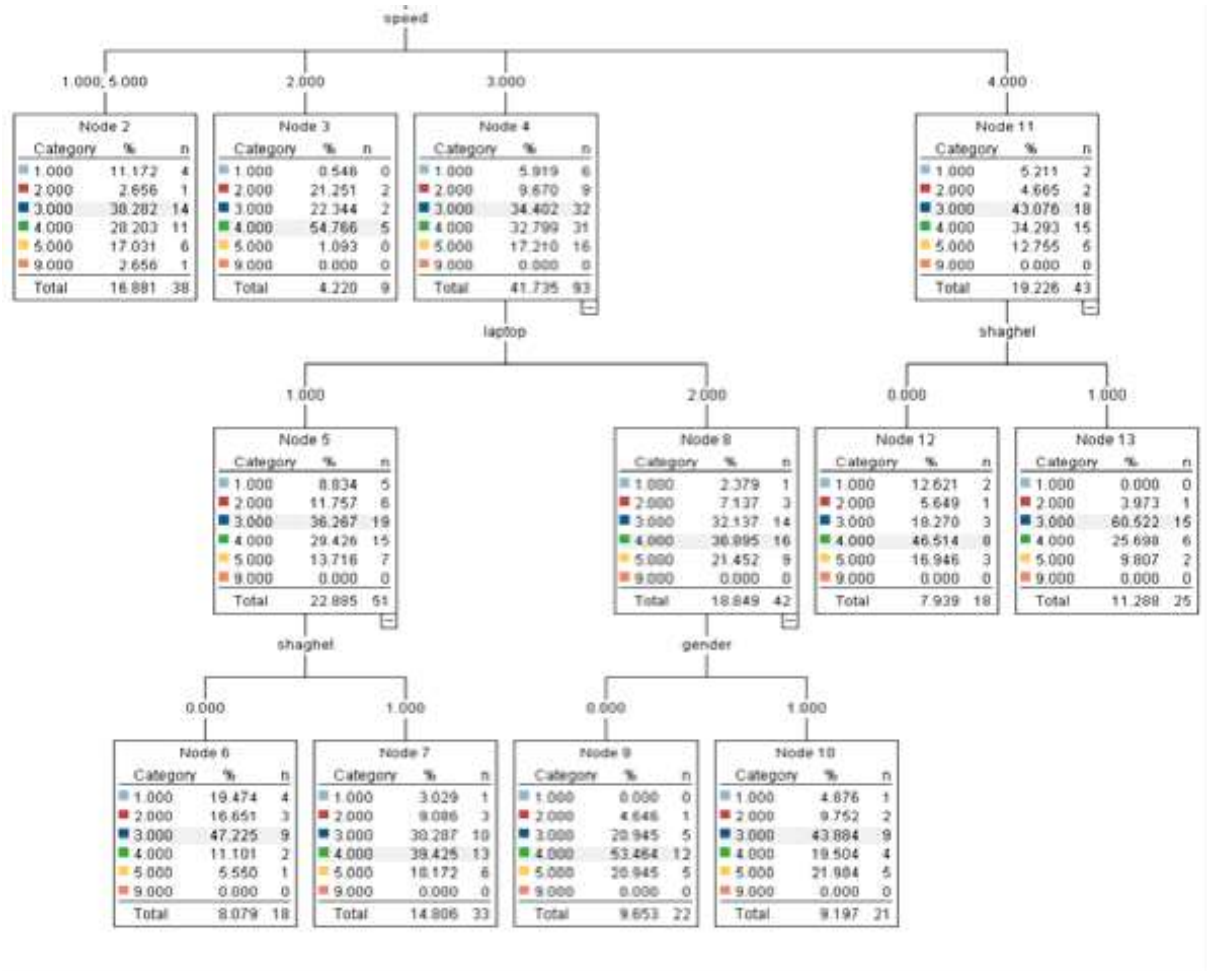


**Figure 6.** The generated decision tree with the Virtual Classroom target node.

Figure 6 shows the decision tree which was generated based on the selected fields and the Assignment View target node, i.e. the average number of assignment view. According to this

tree, the number of assignment views is positively related to the city of residence, speed of internet connection, and student's gender. Students who have access to high speed internet connection will view the assignments in a greater proportion. Also, male students are most likely to view the assignments.

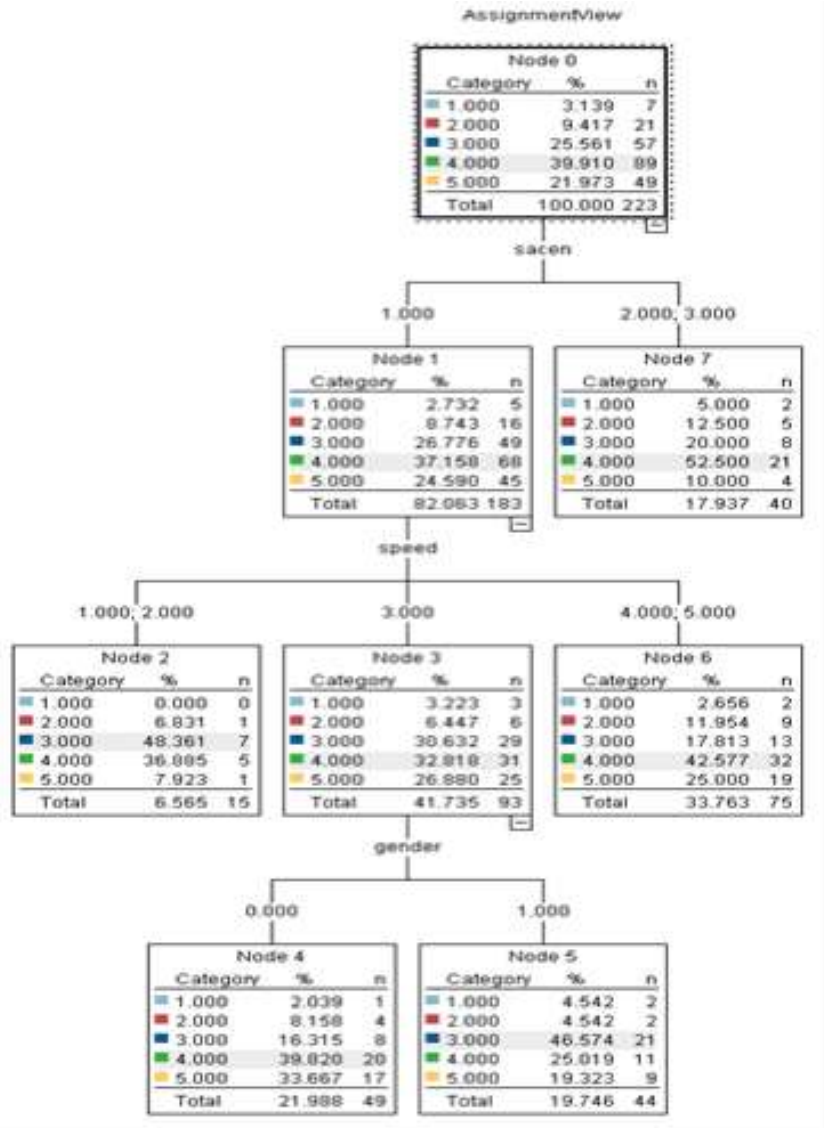


**Figure 7.** The generated decision tree with the Assignment View target node.

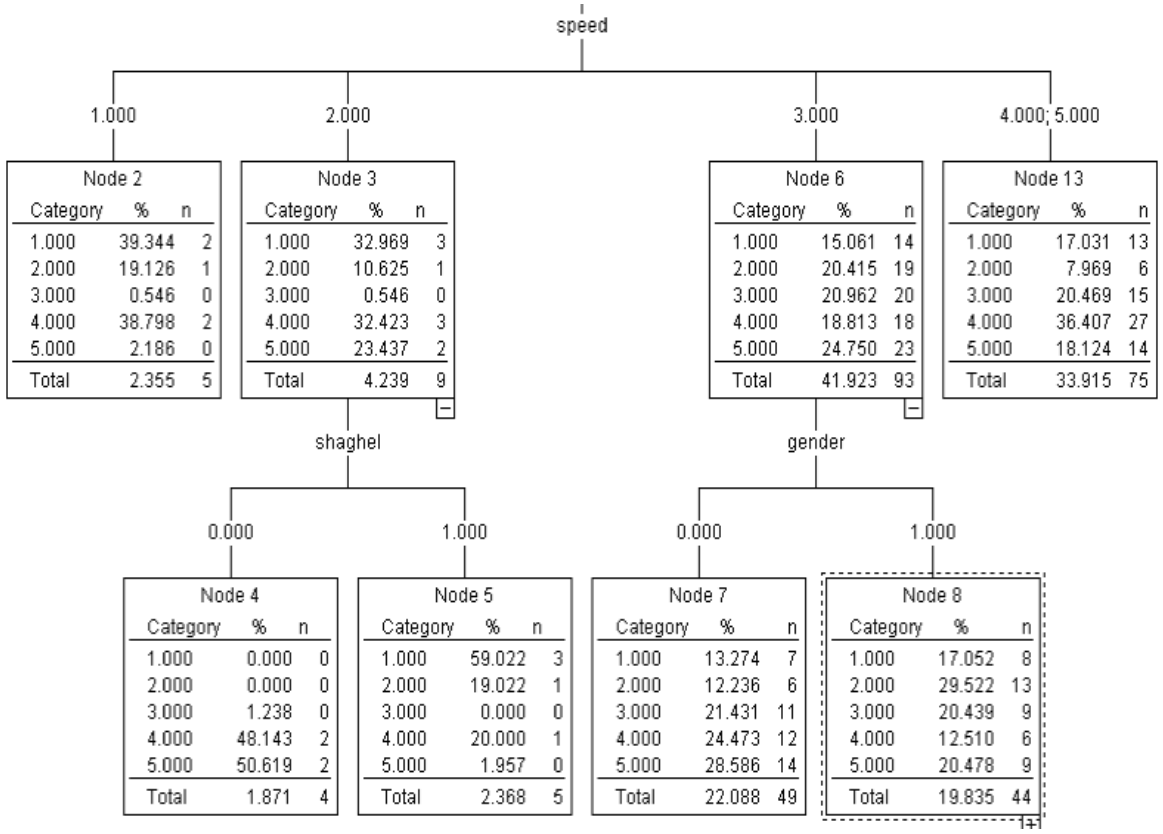


**Figure 8.** The generated decision tree with the Assignment Upload target node.

Figure 7 shows the decision tree which was generated based on the selected fields and the Assignment upload target node, i.e. the average of responses to the uploaded assignments. According to this tree, the number of assignment views is positively related to the employment status and student's gender. Employed students will respond to fewer assignments. Also, female students are more active in responding the assignments.

In the following, the results of implementing the decision tree technique were analyzed in order to predict virtual student's final grades and indeed, their GPA. Here, some of the features of the data warehouse were chosen as influencing parameters and input of data mining algorithms, and the results were obtained and investigated based on them. These parameters include GPA of the current term, prerequisite math grade, prerequisite physics grade, and gender, field of study, Diploma GPA, previous semester GPA, quiz scores, midterm exam score, and having failed courses in the previous semester.

Continuous parameters were mapped and then data mining techniques were applied on the data. Student's grades and GPA were classified in four categories of more than 17, from 15 to 17, from 12 to 15, and below 12, and each category was mapped, respectively, 4, 3, 2 and 1, which represent excellent, good, fair and poor performance. The highest level of the generated tree is an indicator of the direct relationship and enormous effect of fail or not fails in the previous semester parameter. Based on this level, it is predicted that students who have failed courses of the previous semester will gain a fair or good GPA in the current semester grade point

average or good will and students who did not have fail courses will mostly have good or very good GPA. In the next level of the tree, quiz and midterm scores mean is predicted as the next effective parameter. Accordingly, students whose quiz scores mean is excellent, will have an excellent GPA. Also, students whose quiz scores mean is good or fair, are predicted to have a fair GPA. In the next level of the tree, prerequisite math and physics grades are identified as effective parameters. According to their most considerable results, it can be found that the students who did not pass the prerequisite physics course are most likely to have an excellent GPA.

**Table 3.** Some of the most important discovered association rules

| | Association rule | Attributes | Analysis |
|---|---|---|---|
| 1 | Avg_avg → Mantaghe1,Sanaye | Sup = 0.5<br>Conf = 0.8 | 50% of students who have a fair GPA average live in District 1andstudy industrials. |
| 2 | Avg_good → Laptop, Computer | Sup = 0.5<br>Conf = 0.7 | 50% of students who have a good GPA, study computer and use a laptop. |
| 3 | Laptop → Mantaghe1 , Bikar | Sup = 0.5<br>Conf = 0.8 | 50% of students who use laptops, live in district 1 and are not employed. |
| 4 | Avg_bad → Mantaghe1, Laptop | Sup = 0.3<br>Conf = 0.4 | 30% of students who have a poor GPA, live in District 1anduse a laptop. |
| 5 | Bikar, Laptop → Mojarad | Sup = 0.7<br>Conf = 0.7 | 70% of students who are unemployed and use a laptop, are single. |
| 6 | Shaghel → Mantaghe1 | Sup = 0.3<br>Conf = 0.8 | 30% of students who are employed live in District 1. |

In the following, potential significant patterns from among other parameters which may influence on the performance of virtual education students were analysed using association rules mining and Apriori algorithm techniques. Parameters such as the current semester GPA, area of residence, marital status, employment status, and gender, field of study, laptop use, and cell phone use were studied in this phase.

According to the proposed model, the discovered association rules were extracted from virtual student's data base using Clementine tool. Assuming the minimum support and confidence of 3.0, 222 significant rules were discovered from student's database. As noted above, it is practically impossible to analyse the rules due to the selection of low value of confidence and support. Hence, various values of confidence and support were tested during the second implementation of the association rules mining algorithm so that favourable results would be obtained. Assuming the minimum support and confidence of 0/4, 85 significant rules were discovered from student's database. In Table 3, some of the most important discovered association rules are presented.

## 6. CONCLUSIONS

The strategy presented in this study is specifically intended for analysing the data related to student's use of virtual education environments and its relationship with their final grades. In

this study, actual data of students in thee-learning centre of University of Science and Technology have been used. First, data warehouse was developed using these data which provided a suitable context for studying the behaviour of the students and their participation in any of the educational activities. Then, using analytical tools and business intelligence strategies, procedures were developed for the easy access to this data by the centre's directors, and further studies were carried out on the student's behaviour and learning pattern in the learning environment. In the first step, effective parameters on student's educational activities were selected from the database and then, along with the semester GPA, they were used in generating the decision trees. In the next step, it was attempted to predict the final GPA using student's educational records and data mining techniques. Finally, student's data were analysed through discovering interesting association rules by Apriori algorithm. All the analysis and pattern discovery stages in this study were implemented using Clementine data mining tool. The obtained results can be presented as a strategy for improving virtual education systems to the officials and directors of these systems.