Volume 2 No. 2, February 2012 ISSN 2223-4985

International Journal of Information and Communication Technology Research



http://www.esjournals.org

# Mining Educational Data to Improve Students' Performance: A Case Study

**Mohammed M. Abu Tair, Alaa M. El-Halees**
Faculty of Information Technology
Islamic University of Gaza
Gaza, Palestine

## ABSTRACT

Educational data mining concerns with developing methods for discovering knowledge from data that come from educational domain. In this paper we used educational data mining to improve graduate students' performance, and overcome the problem of low grades of graduate students. In our case study we try to extract useful knowledge from graduate students data collected from the college of Science and Technology – Khanyounis. The data include fifteen years period [1993-2007]. After preprocessing the data, we applied data mining techniques to discover association, classification, clustering and outlier detection rules. In each of these four tasks, we present the extracted knowledge and describe its importance in educational domain.

**Keywords:** *Educational Data Mining, Classification, Association rules; Clustering, Outlier Detection.*

## I. INTRODUCTION

There are increasing research interests in using data mining in education. This new emerging field, called educational data mining, concerned with developing methods that extract knowledge from data come from the educational context [10]. The data can be colleted form historical and operational data reside in the databases of educational institutes. The student data can be personal or academic. Also it can be collected from e-learning systems which have a large amount of information used by most institutes [9][10].

The main objective of higher education institutes is to provide quality education to its students and to improve the quality of managerial decisions. One way to achieve highest level of quality in higher education system is by discovering knowledge from educational data to study the main attributes that may affect the students' performance. The discovered knowledge can be used to offer a helpful and constructive recommendations to the academic planners in higher education institutes to enhance their decision making process, to improve students' academic performance and trim down failure rate, to better understand students' behavior, to assist instructors, to improve teaching and many other benefits [7].

Educational data mining uses many techniques such as decision tree, rule induction, neural networks, k-nearest neighbor, naïve Bayesian and many others. By using these techniques, many kinds of knowledge can be discovered such as association rules, classifications and clustering [9][10].

This paper investigates the educational domain of data mining using a case study from the graduate students data collected from the college of Science and Technology Khanyounis. The data include fifteen years period [1993-2007]. It showed what kind of data could be collected, how could we preprocess the data, how to apply data mining methods on the data, and finally how can we benefited from the discovered knowledge. There are many kinds of knowledge can be discovered from the data. In this work we investigated the most common ones which are association rules, classification, clustering and outlier detection. The Rapid Miner software is used for applying the methods on the graduate student's data set.

From The discovered knowledge, we need to provide a college management with a helpful and constructive recommendation to overcome the problem of low grades of graduate students, and to improve students' academic performance.

This paper takes into consideration the ethical and privacy issues. Official approval from the college of science and technology - Khanyounis was obtained to have an access to the related databases for the sole use of analysis and knowledge discovery purposes. To achieve privacy, all individual and personal data are extracted from the database before applying the data mining methods.

The rest of this paper is organized as follows: Section 2 presents related works in educational data mining. Section 3 describes the data set and the preparation and processing methods performed. Section 4 reports our experiments about applying data mining techniques on the educational data. Finally we conclude this paper with a summary and an outlook for future work in Section 5.

## II. RELATED WORKS

Although, using data mining in higher education is a recent research field, there are many works in this area. That is because of its potentials to educational institutes.

Romero and Ventura [9], have a survey on educational data mining between 1995 and 2005. They concluded that educational data mining is a promising area of research and it has a specific requirements not presented in other domains. Thus, work should be oriented towards educational domain of data mining.

El-Halees [5], gave a case study that used educational data mining to analyze students' learning behavior. The goal of

his study is to show how useful data mining can be used in higher education to improve student' performance. He used students' data from database course and collected all available data including personal records and academic records of students, course records and data came from e-learning system. Then, he applied data mining techniques to discover many kinds of knowledge such as association rules and classification rules using decision tree. Also he clustered the student into groups using EMclustering, and detected all outliers in the data using outlier analysis. Finally, he presented how can we benefited from the discovered knowledge to improve the performance of student.

Al-Radaideh et al. [1], applied the data mining techniques, particularly classification to help in improving the quality of the higher educational system by evaluating student data to study the main attributes that may affect the student performance in courses. The extracted classification rules are based on the decision tree as a classification method, the extracted classification rules are studied and evaluated. It allows students to predict the final grade in a course under study.

Baradwaj and Pal [3], applied the classification as data mining technique to evaluate student' performance, they used decision tree method for classification. The goal of their study is to extract knowledge that describes students' performance in end semester examination. They used students' data from the student' previous database including Attendance, Class test, Seminar and Assignment marks. This study helps earlier in identifying the dropouts and students who need special attention and allow the teacher to provide appropriate advising.

Shannaq et al. [11], applied the classification as data mining technique to predict the numbers of enrolled students by evaluating academic data from enrolled students to study the main attributes that may affect the students' loyalty (number of enrolled students). The extracted classification rules are based on the decision tree as a classification method, the extracted classification rules are studied and evaluated using different evaluation methods. It allows the University management to prepare necessary resources for the new enrolled students and indicates at an early stage which type of students will potentially be enrolled and what areas to concentrate upon in higher education systems for support.

Chandra and Nandhini [4], applied the association rule mining analysis based on students' failed courses to identifies students' failure patterns. The goal of their study is to identify hidden relationship between the failed courses and suggests relevant causes of the failure to improve the low capacity students' performances. The extracted association rules reveal some hidden patterns of students' failed courses which could serve as a foundation stone for academic planners in making academic decisions and an aid in the curriculum re-structuring and modification with a view to improving students' performance and reducing failure rate.

Ayesha et al. [2], used k-means clustering algorithm as a data mining technique to predict students' learning activities in a students' database including class quizzes, mid and final exam and assignments. These correlated information will be conveyed to the class teacher before the conduction of final exam. This study helps the teachers to reduce the failing ratio by taking appropriate steps at right time and improve the performance of students.

## III. THE GRADUATE STUDENTS DATA SET AND PREPROCESSING

The data set used in this paper contains graduate students information collected from the college of Science and Technology – Khanyounis for a period of fifteen years in period from 1993 to 2007. The graduate students data set consists of 3360 record and 18 attribute. Table 1 presents the attributes and their description that exist in the data set as taken from the source database.

### Table 1: The Graduate Students Data Set Description

| Attribute | Description | Selected |
|---|---|---|
| Student_ID | The student ID | |
| Student_Name | The name of the student. | |
| Gender | The gender of the student; male or female. | √ |
| Date_of_Birth | The date of birth for the student. | |
| Place_of_Birth | The Place of birth for the student. | |
| Speciality | The specialty of the student. | √ |
| Enrollment_Year | The year of enrollment in the college. | |
| Graduation_Year | The year of graduation from the college. | |
| City | The City for the student. | √ |
| Location | The location in the city | |
| Address | The address of the student. | |
| Telephone_Number | The Telephone number for the student. | |
| Matriculation_GPA | GPA for the student in the Matriculation. | √ |
| Secondary_School_Type | The type of the Matriculation; Adabi or Ilmi | √ |
| Matriculation_Obtained_Place | The place of obtained Matriculation; in Palestine or other country. | |
| Matriculation_Year | The obtained year of the Matriculation. | |
| College_GPA | GPA for the student in the College. | |
| Grade | The grade for the student; Excellent, Very good, Good, or Average | √ |

The college of Science and Technology in Khanyounis grants their graduates a bachelor degree and diploma in twenty four technical scientific specialty, including three specialties for bachelor degree in Information Technology, Engineering of Buildings, and Medical Laboratory Sciences, and twenty one specialty for diploma which are English Language, Medical Laboratories, Pharmacy, Health monitoring, Secretarial and medical record, Software and databases, Programming and Systems Analysis, Computer networks and the Internet, Architecture, Civil Engineering, Computer Engineering, Technology of Office Equipments, Technology of radio and television, Business Administration, Accounting, Financial management and banking, Marketing and Management

Products, and Management and office automation, Office Management and Secretarial, Interior Design and Graphical Design.

As part of the data preparation and preprocessing of the data set and to get better input data for data mining techniques, we did some preprocessing for the collected data before loading the data set to the data mining software, irrelevant attributes should be removed. The attributes marked as selected as seen in Table 1 are processed via the rapid miner software to apply the data mining methods on them. The attributes such as the Student_Name or Student_ID, etc. are not selected to be part of the mining process; this is because they do not provide any knowledge for the data set processing and they present personal information of the students, also they have very large variances or duplicates information which make them irrelevant for data mining.

The following steps are performed as part of the preparation and preprocessing of the data set:

- The data set contains 51 missing values in various attributes from 3360 record, the records with missing values are ignored from the data set since it doesn't consider a large amount of data. The numbers of records are reduced to 3314 record.
- The Matriculation GPA attribute in the data set contains a large number of continuous values (GPAs). So for efficient later processing, simplified data description and understanding for data mining results, we discretized this attribute to categorical one. For example, we grouped all GPAs into five categorical segments; Excellent, Very good, Good, Average and Poor.

After applying the preprocessing and preparation methods, we try to analyze the data visually and figure out the distribution of values, specifically the grade of students. Figure 1 depicts the distribution of graduate students in period from 1993 to 2007 according to their grades, it is apparent from the figure that the average students present about 54% (1796 record) of the data set.

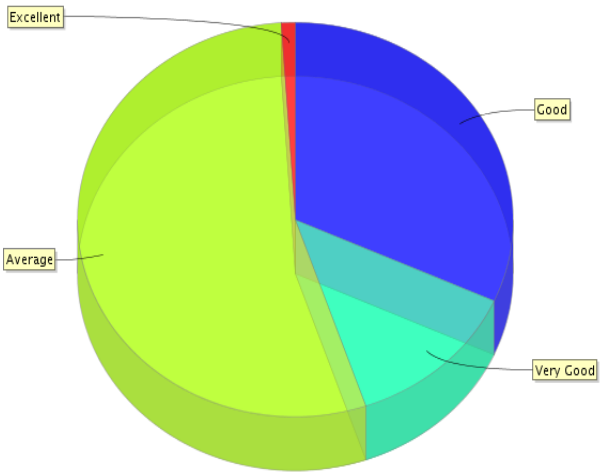


**Figure 1: The distribution of graduate students according to their grades.**

# IV. APPLICATION OF DATA MINING TECHNIQUES TO GRADUATE STUDENTS DATASET: RESULTS AND DISCUSSION

Before applying the data mining techniques on the data set, there should be a methodology that governs our work. Figure 2 depicts the work methodology used in this paper, which is based on the framework proposed in [6]. The methodology starts from the problem definition, then preprocessing which are discussed in the introduction and the data set and preprocessing sections, then we come to the data mining methods which are association, classification ,clustering, and outlier detection, followed by the evaluation of results and patterns, finally the knowledge representation process.

In this section; we describe the results of applying the data mining techniques to the data of our case study, for each of the four data mining tasks; Association, classification, clustering and outlier detection, and how we can benefited from the discovered knowledge.

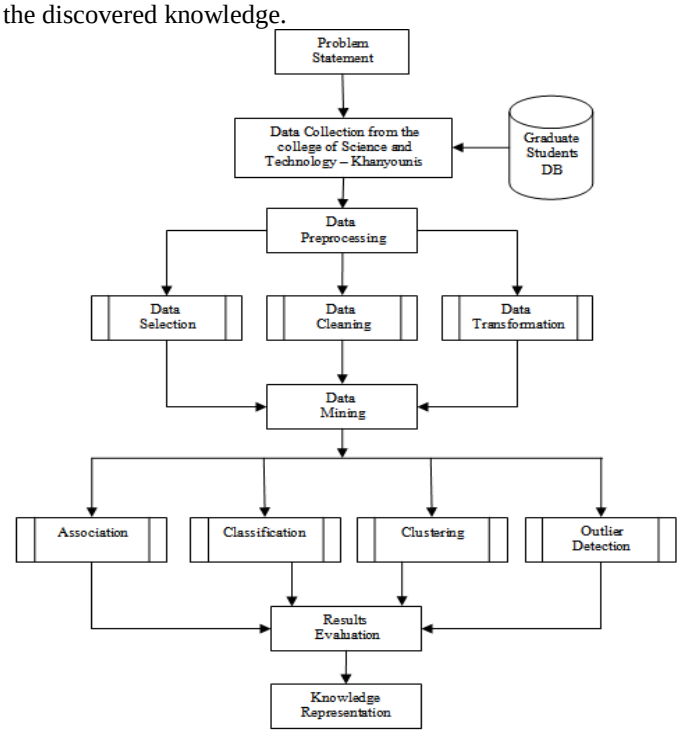


**Figure 2: Data Mining Work Methodology.**

## A. Association Rules

Mining association rules searches for interesting relationships among items in a given data set. It allows finding rules of the form If antecedent then (likely) consequent where antecedent and consequent are itemsets [6]. Itemset is set of one or more items. In our data set an example of item is: Grade = Average. Because, we are looking for items that characterize the grade of students, consequent has one item which is Grade = z where z is one value of the student grade such as Excellent,

Very good, Good, Average. As part of association method, FP-Growth algorithm is applied to the data set.

Figure 3 depicts a sample of association rules discovered from data for students with Average grade, with their support, confidence, and lift.

```
[Matriculation_GPA = Poor, City = Khanyounis, Secondary_School_Type = Adabi] --> [Grade = Average]
(support: 0.195, confidence: 0.757, lift: 1.396)

[City = Khanyounis, Secondary_School_Type = Adabi, Gender = Male] --> [Grade = Average]
(support: 0.101, confidence: 0.754, lift: 1.391)

[Matriculation_GPA = Poor, Secondary_School_Type = Adabi, Gender = Male] --> [Grade = Average]
(support: 0.155, confidence: 0.745, lift: 1.375)

[Matriculation_GPA=Poor, City=Khanyounis, Secondary_School_Type=Adabi, Gender=Female]-->[Grade=Average]
(support: 0.105, confidence: 0.731, lift: 1.349)

[Matriculation_GPA = Poor, Secondary_School_Type = Adabi] --> [Grade = Average]
(support: 0.301, confidence: 0.731, lift: 1.348)

[Matriculation_GPA = Poor, Secondary_School_Type = Adabi, Gender = Female] --> [Grade = Average]
(support: 0.146, confidence: 0.716, lift: 1.321)

[Matriculation_GPA = Poor, Specialty = Business Administration] --> [Grade = Average]
(support: 0.103, confidence: 0.712, lift: 1.314)

[Matriculation_GPA = Poor, City = Khanyounis, Gender = Male] --> [Grade = Average]
(support: 0.133, confidence: 0.706, lift: 1.303)

[Secondary_School_Type = Adabi, Gender = Male] --> [Grade = Average]
(support: 0.176, confidence: 0.699, lift: 1.298)

[City = Khanyounis, Secondary_School_Type = Adabi] --> [Grade = Average]
(support: 0.232, confidence: 0.677, lift: 1.250)
```

**Figure 3: The Resulting Association rules**

These rules are sorted by lift metric. The lift value is the ratio of the confidence of the rule and the expected confidence of the rule and it is used to measuring the interestingness of the rule [12]. The lift value of greater than 1 indicates a positive correlation between antecedent and consequent and the occurrence of one implies the occurrence of the other. For example the first rule with lift is 1.396 means there is a high positive correlation between the antecedent poor Matriculation, Khanyounis City, Adabi secondary school type, and the consequent graduate student Grade Average. With the lift value, we can measure the importance of a rule. The first rule, with the highest lift which means highest correlation is the most important, and so on.

To interpret the rules in the association rules model, the first rule means that of the graduate students under study, 19% (support) are have gotten poor in the Matriculation, from khanyounis city, Adabi in the secondary school type, and have gotten the Grade Average. There is an 75% probability (confidence, or certainty) that a student with Poor in the Matriculation, and from khanyounis city, and with adabi in the sechondary school type will get the Grade Average, and so on.

## B. Classification

Classification is a data mining task that predicts group membership for data instances [6]. In this paper, the classification approaches are used to predict the Grade of the graduate student and there are four grades (Excellent, Very good, Good, and Average) and how other attributes affect them.

Two classification methods are used which are Rule Induction and Naïve Bayesian classifier. A Rule-based classifier extracts a set of rules that show relationships between attributes of the data set and the class label. It uses a set of IF-THEN rules for classification. Rules are easier for humans to understand. Naïve Bayesian classifier is a technique for estimating probabilities of individual variable values, given a class, from training data and to then allow the use of these probabilities to classify new entities [6].

Figure 4 depicts the rules that resulted from applying the Rule Induction classification algorithm on the Grade of the graduate student as a target class. As it is seen from the figure, the attributes that influence the category of the target class are the Secondary_School_Type, the Matriculation_GPA, the City, the Gender and the Speciality, the model presented in figure 4 has an accuracy of 71.25% which is acceptable accuracy and we suggest using Rule Induction algorithm for predicting the grade of the student. To interpret the rules in the Rule Model, the first rule says that, if the Secondary_School_Type equal Adabi and the Matriculation_GPA equal Poor, the grade of the student can be predicted as Average, and so on.

It is important to know that classification rules are different than rules generated from association. Association rules are characteristic rules (it describes current situation), but classification rules are prediction rules (it describes future situation) [5].

```
if Secondary_School_Type = Adabi and Matriculation_GPA = Poor then Average
if Matriculation_GPA = Poor and Gender = Male then Average
if Matriculation_GPA = Poor and City = Khanyounis then Good
if Secondary_School_Type = Adabi and Matriculation_GPA = Average then Average
if Matriculation_GPA = Average and Gender = Female then Good
if Gender = Male and Speciality = English Language then Good
if Gender = Male and Speciality = Business Administration then Good
if Gender = Female and Speciality = Computer networks and the Internet then Very good
if Speciality = Programming and Systems Analysis and Matriculation_GPA = Good  then Average
if Gender = Female and Speciality = Financial management and banking then Very good
if Gender = Male and Matriculation_GPA = Average then Good
if Gender = Female and Speciality = Business Administration then Very good
if Matriculation_GPA = Very good and City = Gaza then Very good
if Secondary_School_Type = Adabi and Matriculation_GPA = Good  then Average
if Speciality = Software and databases and Matriculation_GPA = Excellent then Very good
if Gender = Female and Matriculation_GPA = Good then Very good
if City = Al Sharqiya then Good
if Speciality = Pharmacy and City = Rafah then Average
if Gender = Male and City = Khanyounis then Good
if Speciality = Architecture and Matriculation_GPA = Very good  then Very good
if City = Al Wosta Region and Speciality = Medical Laboratories then Very good
if Speciality = Technology of Office Equipments then Good
if Matriculation_GPA = Poor and Speciality = Medical Laboratories then Average if City = Rafah
and Speciality = Software and databases then Very good
if Matriculation_GPA = Good and City = North Gaza then Very good
if Matriculation_GPA = Poor and Speciality = Accounting then Average
if Matriculation_GPA = Poor and Speciality = Programming and Systems Analysis then Very good
if Gender = Male and Speciality = Technology of radio and television then Good
if Matriculation_GPA = Good and Speciality = Architecture then Very good
if Gender = Male and Matriculation_GPA = Very good then Good
else Average
```

**Figure 4: The Resulting classification rules.**

Figure 5 presents the Distribution model for label attribute Grade that resulted from applying the Naïve Bayesian classifier, the model has an accuracy of 67.50% which is acceptable accuracy and we suggest using Naïve Bayesian method for predicting the grade of the student.

```
SimpleDistribution

Distribution model for label attribute Grade

Class Good (0.318)
5 distributions

Class Very good (0.130)
5 distributions

Class Excellent (0.010)
5 distributions

Class Average (0.542)
5 distributions
```

**Figure 5: The Distribution model for label attribute Grade**

The benefit of these two methods is that it can predict low grades on time. For example the college management can predict Average students from the beginning and they may work on them to improve their performance before the graduation.

## C. Clustering

Clustering is a data mining task that finds groups of objects so that objects that belong to one cluster are more similar to each other than to objects belonging to different cluster. The objective of clustering is to find high-quality clusters such that the Inter-cluster distances are maximized and the Intra-cluster distances are minimized. The clustering method applied in this paper is the k-means; the objective of this k-means test is to choose the best cluster center to be the centroid [6]. The k-means algorithm requires the change of nominal attributes into numerical. The clustering method produced a model with four clusters. Figure 6 presents the resulting centroid Table where from the figure we can see the average value of each attribute in each cluster; e.g. the cluster labeled Cluster 2 has an average of the Grade attribute of 1.886 and this cluster has 1032 items which represents about %31 of the records. The centroid of a cluster represents the most typical case.

| Attribute | cluster_0 | cluster_1 | cluster_2 | cluster_3 |
|---|---|---|---|---|
| Matriculation_GPA | 0.414 | 0.862 | 0.500 | 0.308 |
| Gender | 0.372 | 0.445 | 0.621 | 0.517 |
| Speciality | 5.108 | 1.316 | 9.900 | 15.335 |
| City | 2.191 | 2.176 | 2.374 | 2.498 |
| Secondary_School_Type | 0.769 | 0.097 | 0.748 | 0.796 |
| Grade | 1.895 | 1.511 | 1.886 | 1.929 |
| Number of Records | 769 | 1032 | 1033 | 480 |
| Percentage | 23% | 31% | 31% | 15% |

**Figure 6: The resulting centroid Table after applying k-means algorithm.**

Figure 7 depicts a graphical representation of the clusters distribution after applying Single Value Decomposition method, which reduces the number of attributes to two in order to easily plot the clusters.

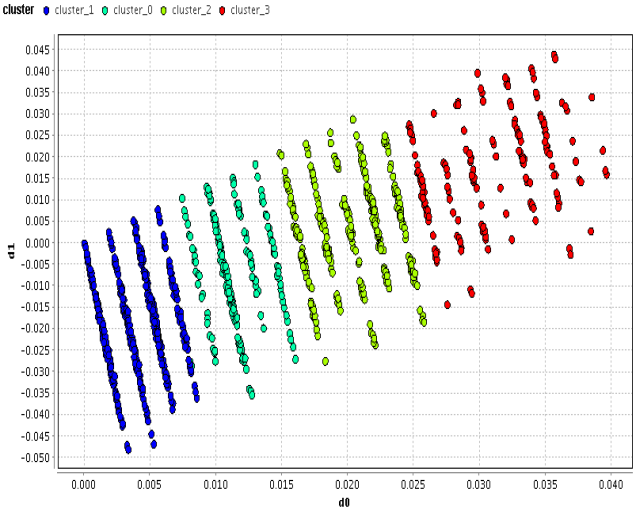


**Figure 7: Clusters distribution plot with SVD applied.**

By studying and observing each cluster, we can form a table determining the characteristics of each Cluster and comparison between all clusters as shown in Table 2.

**Table 2: Clusters characteristic with respect to selected attributes**

| Attribute / Cluster | Matriculation GPA | Perc. | Gender | Perc. | Specialty | Perc. | City | Perc. | Secondary School Type | Perc. | Grade | Perc. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Cluster 0 | Poor | 73% | Male | 63% | Architecture | 0.4% | Al Wosta | 10% | Ilmi | 23% | Good | 31% |
| | Average | 15% | Female | 37% | Civil Engineering | 6.8% | Rafah | 24% | Adabi | 77% | Very good | 8.8% |
| | Good | 9% | | | Business Administration | 81% | Gaza | 4% | | | Excellent | 0.4% |
| | Very Good | 3% | | | English Language | 5% | Khanyounis | 60% | | | Average | 60% |
| | | | | | Technology of R and TV | 7% | North Gaza | 1% | | | | |
| | | | | | | | Al Shanjiya | 1% | | | | |
| Cluster 1 | Poor | 48% | Male | 56% | Medical Laboratories | 28% | Al Wosta | 11% | Ilmi | 90% | Good | 37% |
| | Average | 24.4% | Female | 44% | Pharmacy | 30% | Rafah | 23% | Adabi | 10% | Very good | 19% |
| | Good | 22% | | | Programming and Systems Analysis | 24% | Gaza | 10% | | | Excellent | 2% |
| | Very Good | 5% | | | Architecture | 18% | Khanyounis | 50% | | | Average | 42% |
| | Excellent | 0.6% | | | | | North Gaza | 5% | | | | |
| | | | | | | | Al Shanjiya | 1% | | | | |
| Cluster 2 | Poor | 67% | Male | 38% | Accounting | 25% | Al Wosta | 5% | Ilmi | 25% | Good | 29% |
| | Average | 20% | Female | 62% | Office Management and Secretarial | 21% | Rafah | 27% | Adabi | 75% | Very good | 12.5% |
| | Good | 9.6% | | | Health monitoring | 11% | Gaza | 1% | | | Excellent | 0.5% |
| | Very Good | 3% | | | Software and databases | 24% | Khanyounis | 64% | | | Average | 58% |
| | Excellent | 0.4% | | | Secretarial and medical record | 19% | North Gaza | 0.3% | | | | |
| | | | | | | | Al Shanjiya | 2.7% | | | | |
| Cluster 3 | Poor | 78% | Male | 48% | Financial management and banking | 28% | Al Wosta | 2% | Ilmi | 20% | Good | 30% |
| | Average | 15% | Female | 52% | Marketing and Management Products | 16% | Rafah | 26% | Adabi | 80% | Very good | 8.2% |
| | Good | 5% | | | Interior Design | 17% | Gaza | 0.7% | | | Excellent | 0.6% |
| | Very Good | 1.3% | | | Technology of Office Equipments | 2.7% | Khanyounis | 67% | | | Average | 61% |
| | Excellent | 0.5% | | | Computer networks and the Internet | 11% | North Gaza | 0.3% | | | | |
| | | | | | Management and office automation | 20% | Al Shanjiya | 4% | | | | |
| | | | | | Computer Engineering | 4.3% | | | | | | |
| | | | | | Graphical Design | 1% | | | | | | |

As we note from Table 2 that the most students in cluster 0 with the male gender, the matriculation_GPA is Poor, the secondary school type is Adabi, the city is Khanyounis, the

Specialty is Business Administration and the Grade is Average, and so on.

Using the results in Table 2 , we can guide students in the four groups according to their behavior.

## D. Outlier Detection

Outlier detection discovers data points that are significantly different than the rest of the data [8]. In educational data mining outlier analysis can be used to detect students with learning problems [9]. In this paper, we used outlier analysis to detect outliers in the student data. Two outlier methods are used which are Distance-based Approach and Density-Based Approach.

### 1) Distance-based Approach

It Identifies the number of outliers in the given data set based on the distance to their k nearest neighbors, and the result of applying this method is to flag the records either to be outlier or not, with true or false [6].

Figure 8 depicts a graphical representation of the outliers distribution with the red color after applying Single Value Decomposition method, which reduces the number of attributes to two in order to easily plot the outliers.

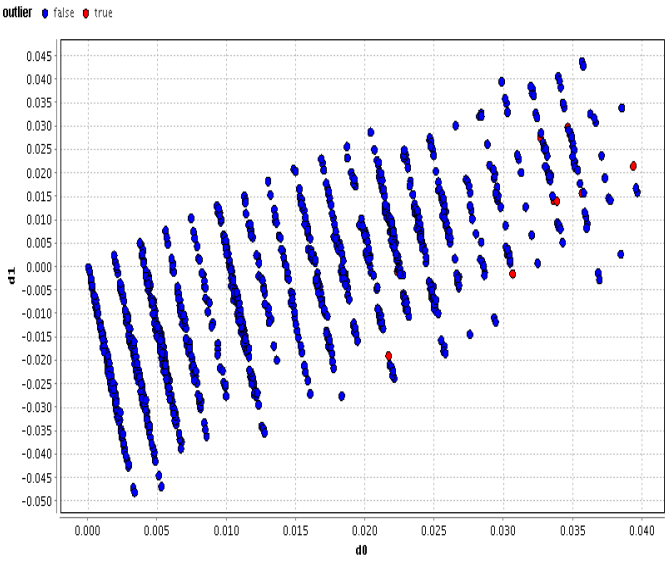


**Figure 8: Outliers (distance-based) distribution plot with SVD applied.**

The system detected 10 outliers in our data, by studying and checking some of the 10 outlier points, we found that outliers are not errors but rather it represents a rare event, for example some of the outlier points are students with excellent degree in the matriculation GPA and also with excellent degree in the college which are different than the rest of the students in The graduate students data set and they are few students.

### 2) Density-Based Approach

It Computes local densities of particular regions and declare instances in low density regions as potential outliers. The method used is Local Outlier Factor (LOF), the Basic idea of LOF is to compare the local density of a point with the densities of its neighbors, and the result of applying this method is to flag the records with a percentage of outlier. The larger score means larger possibility of being outlier [6].

Figure 9 depicts a graphical representation of the outliers after applying Single Value Decomposition method.

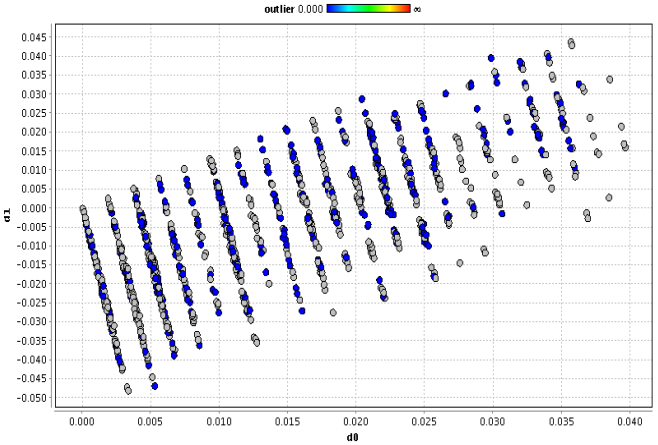


**Figure 9: Outliers (LOF) distribution plot with SVD applied.**

For each case, the college management can look at the outlier behavior of the student and try to find and understand why the irregularity happened and then analyzed for knowing the cause.

## V. CONCLUSION AND FUTURE WORK

In this paper, we gave a case study in the educational data mining. It showed how useful data mining can be used in higher education particularly to improve graduate students' performance. We used graduate students data collected from the college of Science and Technology in Khanyounis. The data include fifteen years period [1993-2007]. We applied data mining techniques to discover knowledge. Particularly we discovered association rules and we sorted the rules using lift metric. Then we used two classification methods which are Rule Induction and Naïve Bayesian classifier to predict the Grade of the graduate student. Also we clustered the students into groups using K-Means clustering algorithm. Finally, we used outlier detection to detect all outliers in the data, two outlier methods are used which are Distance-based Approach and Density-Based Approach. Each one of these tasks can be used to improve the performance of graduate student.

Our future work include applying data mining techniques on an expanded data set with more distinctive attributes to get more accurate results. Also, experiments could

be done using more data mining techniques such as neural nets, genetic algorithms, k-nearest Neighbor, and others. Finally, the used preprocess and data mining algorithms could be embedded into the college system so that any one using the system can benefited from the data mining techniques.

## REFERENCES

[1] Al-Radaideh, Q., Al-Shawakfa, E. and Al-Najjar, M. (2006) 'Mining Student Data Using Decision Trees', The 2006 International Arab Conference on Information Technology (ACIT'2006) – Conference Proceedings.

[2] Ayesha, S. , Mustafa, T. , Sattar, A. and Khan, I. (2010) 'Data Mining Model for Higher Education System', European Journal of Scientific Research, vol. 43, no. 1, pp. 24-29.

[3] Baradwaj, B. and Pal, S. (2011) 'Mining Educational Data to Analyze Student s' Performance', International Journal of Advanced Computer Science and Applications, vol. 2, no. 6, pp. 63-69.

[4] Chandra, E. and Nandhini, K. (2010) 'Knowledge Mining from Student Data', European Journal of Scientific Research, vol. 47, no. 1, pp. 156-163.

[5] El-Halees, A. (2008) 'Mining Students Data to Analyze Learning Behavior: A Case Study', The 2008 international Arab Conference of Information Technology (ACIT2008) – Conference Proceedings, University of Sfax, Tunisia, Dec 15- 18.

[6] Han, J. and Kamber, M. (2006) Data Mining: Concepts and Techniques, 2nd edition. The Morgan Kaufmann Series in Data Management Systems, Jim Gray, Series Editor.

[7] Kumar, V. and Chadha, A. (2011) 'An Empirical Study of the Applications of Data Mining Techniques in Higher Education', International Journal of Advanced Computer Science and Applications, vol. 2, no. 3, pp. 80-84.

[8] Mansur, M. O. , Sap, M. and Noor , M. (2005) 'Outlier Detection Technique in Data Mining: A Research Perspective', In Postgraduate Annual Research Seminar.

[9] Romero, C. and Ventura, S. (2007) 'Educational data Mining: A Survey from 1995 to 2005', Expert Systems with Applications (33), pp. 135-146.

[10] Romero, C. , Ventura, S. and Garcia, E. (2008) 'Data mining in course management systems: Moodle case study and tutorial', Computers & Education, vol. 51, no. 1, pp. 368-384.

[11] Shannaq, B. , Rafael, Y. and Alexandro, V. (2010) 'Student Relationship in Higher Education Using Data Mining Techniques', Global Journal of Computer Science and Technology, vol. 10, no. 11, pp. 54-59.

[12] Sheikh, L. , Tanveer, B. and Hamdani , S. (2004) 'Interesting Measures for Mining Association Rules', IEEE-INMIC – Conference Proceedings.

## AUTHORS PROFILE

**Mohammed M. Abu Tair** Received his B.Sc. in Information Technology Systems from the Islamic University of Gaza in 2005; currently he is pursuing his M.Sc. in Information Technology from the Islamic university of Gaza. He works as computer administrative in the College of Science and Technology, KhanYounis. His research activities are in the area of data mining and web services. He holds differenet technical certificates:MCSE, CCNA.

**Alaa M. El-Halees** is an associate professor in computing and supervisor of graduate studies of the faculty of Information Technology Department at Islamic University of Gaza, Palestine. He holds PhD degree in data mining in 2004, M.Sc. degree in software development in 1998 from Leeds Metropolitan University, UK. He received his B.Sc. degree in computer engineering in 1989 from University of Arizona, USA. His research activities are in the area of data mining, in particular text mining, machine learning and e-learning.