Mining Educational Data to Analyze Students' Performance

(A Case with University College of Science and Technology Students)

Ahmed S. J. Abu Hammad

***Abstract*—**Data mining is a new information process technology; it has gained some impressive results in higher education, such as detection of abnormal values in the result sheets of the students, a prediction about students' performance and so on. Predicting students' performance is critical for educational institutions because strategic programs can be planned for improving or maintaining students' performance during their period of studies. In this paper, data mining methods are applied at University College of Science and Technology (UCST) – Khan Younis on students enrolled. After data preparation and pre-processing, different techniques of data mining are applied to detect association, classification, clustering, and outlier detection rules. In every one of these four tasks, we offer the extracted knowledge and describe its significance in the educational domain. This paper can provide strong information support, decision support and work direction for administrators of UCST, thus, promote the comprehensive development of the educational system's and improve students' performance in UCST.

***Keywords*—**Educational Data Mining, Association Rules, Classification, Clustering, Outlier Detection, Predicting Students' Performance.

I. Introduction

Educational Data Mining (EDM), concerns with developing methods that discover knowledge from data originating from an educational context. The data can be gathered from historical and operational data reside in the databases of educational institutes. The student data can be personal or academic [8, 10].

The principal aim of institutions of higher education is to supply quality education for their students and to get better the quality of administrative decisions. One way to succeed in doing the highest level of quality in the higher education system is by discovering knowledge from educational data to study the main attributes that may impact the students' performance. The discovered knowledge can be used to offer helpful and constructive recommendations to the academic planners in institutions of higher education to promote their decision-making process, to improve students 'academic performance and reduce failure rate, to better understand students' behavior, to help instructors, to enhance teaching and many other advantages [1,9].

Educational data mining utilizes numerous techniques such as decision tree, rule induction, k-nearest neighbor, naive Bayesian and numerous others. By utilizing these techniques, numerous sorts of knowledge can be discovered such as association rules, classifications, and clustering [1,13,14].

This paper examines the educational domain of data mining utilizing a case study from the enrolled students' data collected from the University College of Science and Technology - Khan Younis. It showed what sort of data could be gathered, how could we pre-process the data, how to apply data mining techniques on the data, and finally how can we have profited from the discovered knowledge. There are numerous sorts of knowledge can be revealed from the data. In this work, we checked the most common ones which are association rules, classification, clustering and outlier detection. The Rapid Miner software is used for applying the methods to the enrolled student's data set.

The model will predict for students' performance, and relate them with other factors such as gender, address details, general secondary average, general secondary section, specialization of

Ahmed Abu Hammad, University College of Science and Technology, Khan Younis, Palestine (e-mail: asj_hammad@hotmail.com).

the student, and hours completed number. The discovered knowledge, supply a university college management with a helpful and constructive recommendation to conquer the issue of low grades of students and to progress students' academic performance.

The subsequence sections are organized as follows: section II contains related works in educational data mining. Section III illustrates the data set and the preparation and processing methods performed. Then the following section presents our experiments about applying data mining techniques to the educational data. Finally, conclusion and future work are presented.

II. Related Works

Data mining in higher education is a recent research field and this area of research is gaining popularity because of its potentials to educational institutes.

El-Halees A. [7], 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 a student' performance. He used students' data from the database course and collected all available data including personal records and academic records of students, course records and data came from the e-learning system. Then, he applied data mining techniques to discover many kinds of knowledge such as association rules and classification rules using a 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 the student.

Nghe N. et al. [11] have compared the accuracy of the decision tree and Bayesian network to predict a student's Grade Point Average (GPA) at the end of the third year of undergraduate and at the end of the first year of postgraduate from two different institutes. All data set has 20,492 and 936 complete student records respectively. The results show that the decision tree algorithm was significantly more accurate than the Bayesian network algorithm for predicting student performance. The accuracy was further improved by using re-sampling technique especially for decision tree in all cases of classes. In the same time, can reduce misclassification especially on minority class of imbalanced datasets because decision tree algorithm tends to focus on a local optimum.

Pal B. [4], utilized the classification as a data mining technique to assess a student' performance, they applied the decision tree method for classification. The point of their study is to extract knowledge that characterizes students' performance in the end semester examination. They utilized students' data including Attendance, Class test, Seminar, and Assignment marks. This study helps earlier in deciding the dropouts and students who need particular attention and allow the teacher to give appropriate advising.

Bekele R. et al. [5] have examined the Bayesian network to predict students' performance of high school students containing 8 attributes (social and personal attributes). Dataset has 514 complete student records respectively. The paper demonstrates an application of the Bayesian approach in the field of education and shows that the Bayesian network classifier has the potential to be used as a tool for prediction of student performance.

Bidgoli M. et al. [6] use the data mining classification technique to predict students' final grades based on their web-use feature. By discovering the successful patterns of students in various categories, the university can predict the final grade of every student. Therefore, it helps to identify students at risk early and allow the instructor to provide appropriate advice in a timely manner. From this case study, it can be concluded that data mining is effective in predicting a student's performances in the educational domain. The result has an impact in improving the transition rate, and the process indicator of a higher learning to institute by improving the student assessment process.

Ayesha S. et al. [2], applied a k-means clustering algorithm as a data mining technique to anticipate students' learning activities in a students' database including class quizzes, mid and final exam and assignments. This correlated information will be reported to the class teacher before the conduction of the final exam. This study helps the teachers to minimize the failing ratio by making appropriate strides at the perfect time and improve the performance of students.

Paris I. et al. [3] have compared the accuracy of data mining methods to predict students' grade. Dataset has 2427 complete records for Bachelor of Computer Science students at University Putra Malaysia (UPM) admitted from 2000 to 2004. The results show that combining different classifiers improved the prediction accuracy compare to single classifiers. The results also show that the resampling technique has not improved the accuracy of prediction in all cases. The results also show that the hidden naive Bayes method consistently outperformed.

III. Student Enrollments Dataset and Pre-processing

The admissions and registration department is currently gathering demographic, geographic, exam scores, financial information, so on., from applicants as part of the admissions and registration operation. There is too historical data available indicating the actual enrollment status of applicants along with all the other attributes that were gathered as a component of the admissions and registration operation [1].

In this study, data gathered from the UCST. The dataset contains the student enrolled data. The dataset made obtainable has 20 different attributes for each applicant including the decision result attribute. There are in all about 1173 records available. Table I shows the attributes, their types, and description that exist in the data set as taken from the source database.

TABLE I
The enrolled student's dataset description

Attribute	Description	Data Type	Selected
Seq.	A sequence for the record	Number	
Institution	A name for the institution	String	
ID	An identifier for the record	Number	
Name	Student's named	String	
DOB	Date of birth	String	
Gender	Student's gender	String	√
Nationality	Student's nationality	String	
City	Student's address details	String	√
GS Source	Source general secondary	String	
GS Year	Year general secondary	Number	
GS Avg	Average general secondary	Number	√
GS Sec	Section general secondary	Number	√
YI Join	Year institution join	Number	
YI Term	Year institution term	Number	
Std Level	Student's level	Number	
College	Student's college	String	
Specialization	Student's specialization	String	√
HC Num	Hours Completed number	Number	√
GPA	Student's a cumulative grade point average	Number	
Grade	Student's performance	String	√

An important step in the data mining process is data preprocessing. One of the challenges that face the knowledge discovery process in predicting students' performance is poor data quality. For this reason, we try to prepare our data carefully to obtain accurate and correct results. As part of the preparation and preprocessing of the data set, irrelevant and weakly relevant 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, for example, ID are not chosen to be part of the mining process; this is because they do not provide any knowledge for the data set processing, likewise, they have vast differences which make them irrelevant for data mining [1,10].

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

- Selected attributes have little missing (no more than 27 value). Then we try to fill the missing with appropriate values. So, we used to replace the missing value method. This method enables the substitution of the missing values by the minimum, maximum or average statistics calculated on the basis of existing values for all or selected attributes. Moreover, we can also replace the missing values by some pre-defined values (e.g., zero or values that we consider that provide a better fit to data). Here we the substitute of the missing values by the average calculated on the basis of existing values for all selected attributes.
- GS Avg and HC Num attributes contain many values that cannot easily identify interesting patterns in the data from which to create a model. So, we use Discretize by User Specification method. This method allows numerical attributes to be placed in bins where the boundaries of the bins are defined by the user. This converts numerical attributes into nominal ones as required by some algorithms. Here we the substitute of GS Avg attribute by classes (Excellent – Very Good – Good – Acceptable – Fail) and HC Num attribute by classes (First – Second).

After applying the pre-processing and preparation methods, we try to analyze the data visually and figure out the grade distribution of the students which are in the pivot of the predicting students' performance, Figure 1 depicts the grade distribution of the students.

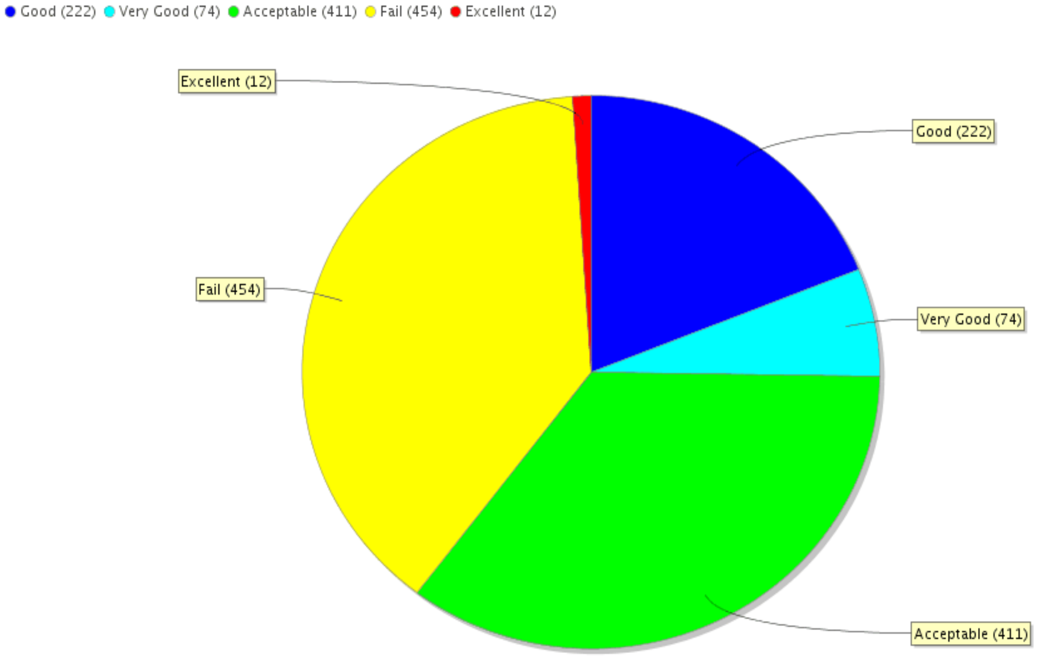


Fig. 1 The distribution of enrolled students according to their grades.

IV. Application of Data Mining Techniques to Student Enrolments 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 presents the work methodology used in this paper, which is based on the framework proposed in [8]. The methodology starts from the problem definition, then preprocessing which are debated in the introduction and the data set and preprocessing sections, then we come to the data mining methods which are an association, classification, clustering, and outlier detection, followed by the evaluation of results and patterns, finally the knowledge representation process [10, 12].

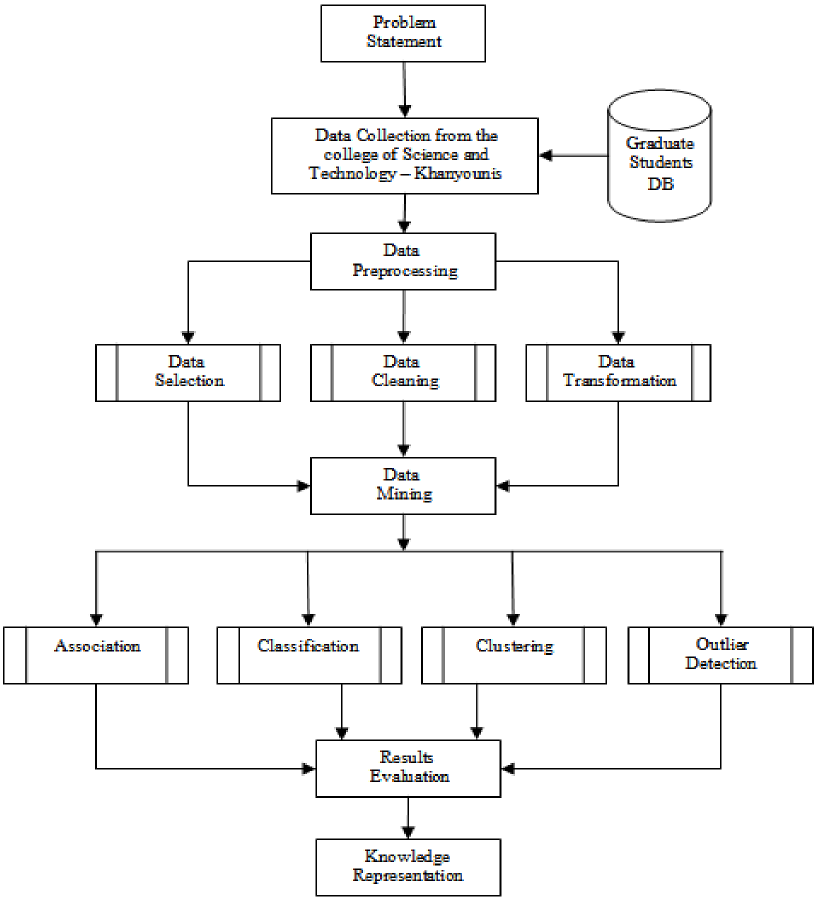


Fig. 2 Data mining work methodology [8].

In this section; we illustrate the results of applying the data mining techniques to the data of our case study, for every one of the four data mining tasks; association, classification, clustering and outlier detection, and how we can profit from the discovered knowledge.

4.1. Association method

Association method is the discovery of association rules showing attribute-value conditions that occur frequently together in a given set of data. Association analysis is widely used for data analysis. More formally, association rules are of the form X=>Y, i.e.,"A1^----^Am → B1^----^Bn", where Ai (for i to m) and Bj (j to n) are attribute-value pairs. The association rule X=>Y is interpreted as database tuples that satisfy the conditions in X are also likely to satisfy the conditions in Y " [10, 12].

As part of the association method, before applying the FP-Growth algorithm requires the change of nominal attributes into binomial attributes, the FP-Growth algorithm is applied to student enrollment data set with min confidence = 0.89, Table II illustrates some useful rules extracted from student enrolment data ordered by confidence.

TABLE II
ASSOCIATIONS RULES FOR STUDENT ENROLMENT DATA.

#	RULE	CONFIDENCE
1	[GS Avg = Weak, Specialization = Business] --> [GS Sec = Literary]	0.965
2	[Specialization = Business] --> [GS Sec = Literary]	0.961
3	[City = Khan Younis, Specialization = Business] --> [GS Sec = Literary]	0.957
4	[City = Khan Younis, GS Sec = Literary, Gender = Male] --> [GS Avg = Weak]	0.937
5	[GS Sec = Literary, Gender = Male, HC Num = Second] --> [GS Avg = Weak]	0.929
6	[GS Sec = Literary, Gender = Male] --> [GS Avg = Weak]	0.927
7	[City = Khan Younis, GS Sec = Literary, Specialization = Business] --> [GS Avg = Weak]	0.915
8	[City = Khan Younis, Specialization = Business] --> [GS Avg = Weak]	0.909
9	[GS Sec = Literary, Specialization = Business] --> [GS Avg = Weak]	0.906
10	[Specialization = Business] --> [GS Avg = Weak]	0.902

Rules #1, #2, #3 and #4, can be used to predict the general secondary section of the student. For example, from rule #1, we understand that there is a general secondary section = Literary of the student if he is general secondary average = Weak and specialization = Business. Rules #5, #6, #7, #8, #9, #10, #11 and #12 provide with better understanding for general secondary average. For example, from rule #5, we understand that there is a general secondary average = Weak of the student if he is city = Khan Younis, general secondary section = Literary, and gender = male.

4.2. Classification method

Classification is the processing of finding a set of models (or functions) which describe and distinguish data classes or concepts, for the purposes of being able to use the model to predict the class of objects whose class label is unknown [10]. In this paper, the classification approaches are used to predict students' performance and present how other attributes affect them. The derived model may be represented in various forms, such as decision trees, rule-based classifier, k-nearest neighbors, or naive Bayesian classifiers. Here we apply five classification techniques above on our data set.

4.2.1. Decision trees

Decision trees have become one of the most powerful and popular approaches in knowledge discovery and data mining [12], Figure 3 depicts the decision tree that resulted from applying the decision tree classification algorithm on the grade as a target class.

As it is seen from the Figure 3, the attributes that influence the category of the target class are GS Avg, City, GS Section, and HC Num, the model presented in Figure 3 has the accuracy of 41.88%. To interpret the rules in the decision tree, the most left branch of the decision tree says that, if the average general secondary = Acceptable and city = Gaza, then grade= Good.

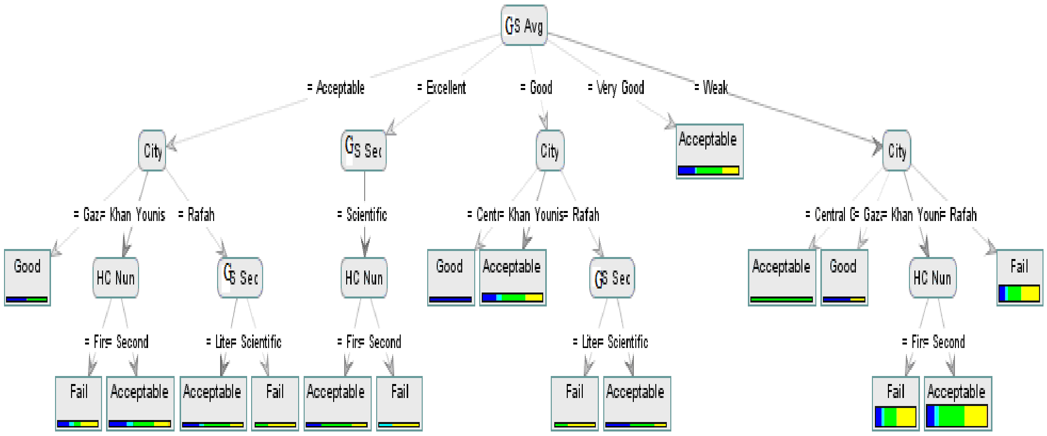


Fig. 3 Decision tree classification for Grade as a target class.

4.2.2. Rule-based classifier

A rule-based classifier is a technique for extracting a set of rules that demonstrate the relationships between the attributes of a dataset and the class label using a collection of" if ... then ..." rules [10]. Table III depicts some of the rule-based classifiers that resulted from applying the rule-based classifier algorithm on the grade as a target class.

As it is observed from Table III, the attributes that influence the category of the target class (grade) are Specialization, GS Avg, City, GS Section, and HC Num, the model presented in table III has the accuracy of 61.88%. To interpret the rules in the rule-based classifier, for example from rule #1 we understand that If Specialization = Graphic Design and HC Num = First then Grade = Acceptable.

TABLE III
THE RULE-BASED CLASSIFIER FOR GRADE AS A TARGET CLASS.

#	RULE
1	If Specialization = Graphic Design and HC Num = First then Grade = Acceptable
2	If Specialization = Graphic Design and GS Sec = Literary then Grade = Acceptable
3	If Specialization = Computer networks and the Internet Graphic and GS Sec = Literary then Grade = Acceptable
4	If City = Gaza and GS Avg = Weak then Grade = Acceptable
5	If Specialization = Medical laboratories and HC Num = Second then Grade = Acceptable

4.2.3. k-Nearest Neighbors

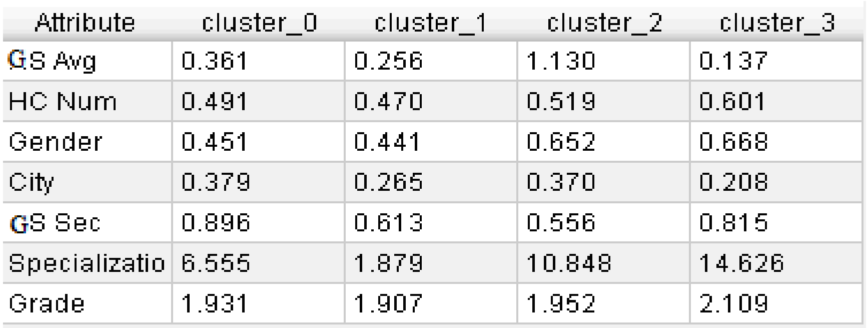
K-nearest neighbor is a supervised learning algorithm where the result of new instance query is classified based on the majority of K-nearest neighbor category. The purpose of this algorithm is to classify a new object based on attributes and training samples. The classifiers do not use any model to fit and only based on memory. Given a query point, we find K number of objects or (training points) closest to the query point. The classification is using majority vote among the classification of the K objects [12]. The classification model that resulted from applying the K-NN algorithm; the model has five classes with the accuracy of 54.37%.

4.2.4. Naive Bayesian Classifiers

Naive Bayesian Classifiers or Naive Bayes 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 [10], Naive Bayes on our dataset presents the accuracy of 58.50%.

4.3. Clustering method

Unlike classification and prediction, which analyze class-labeled data objects, clustering analyzes data objects without consulting a known class label. In general, the class labels are not present in the training data simply because they are not known to begin with. Clustering can be used to generate such labels. The objects are clustered or grouped based on the principle of maximizing the intra-class similarity and minimizing the interclass similarity. 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 [10,12]. The k-means algorithm requires the change of nominal attributes into numerical. The clustering method produced a model with four clusters. Figure 4 demonstrates the resulting "Centroid Table" where from the figure we can see the average value of each attribute in each cluster; for example, the cluster labeled "Cluster_0" has an average of grade: 1.931 and this cluster has 348 items which represent to about %29.67 of the records. The centroid of a cluster represents the most typical case.



Attribute	cluster_0	cluster_1	cluster_2	cluster_3
GS Avg	0.361	0.256	1.130	0.137
HC Num	0.491	0.470	0.519	0.601
Gender	0.451	0.441	0.652	0.668
City	0.379	0.265	0.370	0.208
GS Sec	0.896	0.613	0.556	0.815
Specializatio	6.555	1.879	10.848	14.626
Grade	1.931	1.907	1.952	2.109

Fig. 4 Resulting clusters after applying the k-means algorithm.

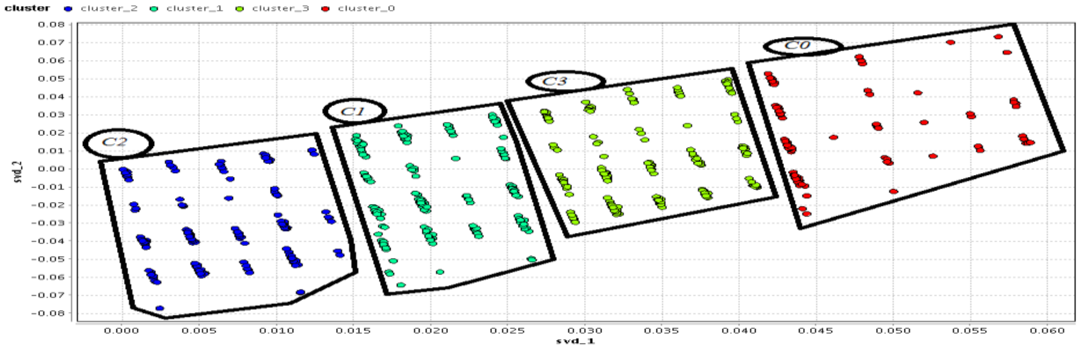


Fig. 5 Clusters distribution plot with SVD applied.

Figure 5 depicts a graphical representation of the clusters distribution after applying the Single Value Decomposition (SVD) method, which reduces the number of attributes to two in order to easily plot the clusters.

4.4. Outlier method

A database may contain data objects that do not comply with the general behavior of the data and are called outliers. The analysis of these outliers may help in fraud detection and predicting abnormal values [12], so in this paper, we apply two outlier methods, one of which is based on distance-based approach, the other is the density-based approach.

4.4.1. Distance-based approach

A popular method of identifying outliers is by examining the distance to an example's nearest neighbors, and the result of applying this method is to flag the records either to be the outlier or not, with true or false [10]. Figure 6 depicts a graphical representation of the region of detected outliers, after applying (SVD) method.

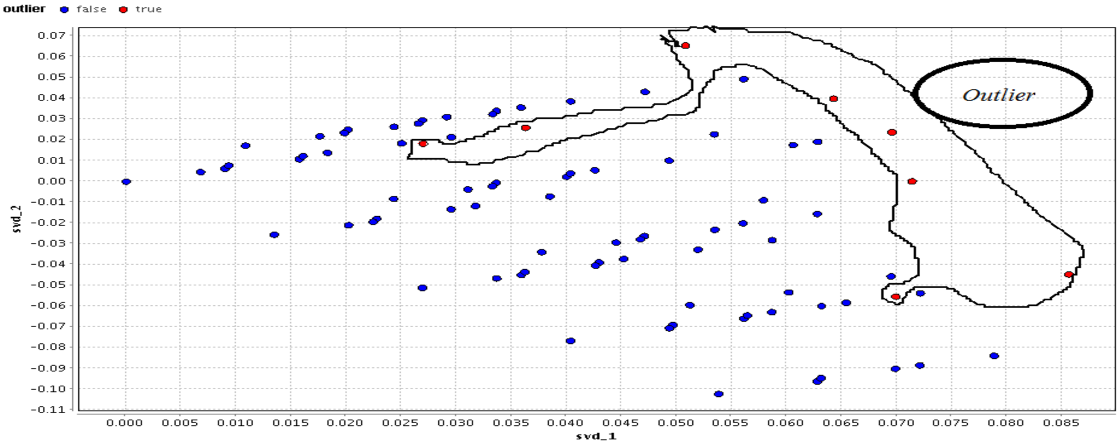


Fig. 6 Outlier (Distance-Based) plot with SVD applied.

4.4.2. Density-based approach

Compute local densities of particular regions and declare instances in low-density regions as potential outliers [12]. Figure 7 depicts a graphical representation detected outliers, by using Local Outlier Factor (LOF) approach, after applying (SVD) method, which represents too large percentage of the outlier.

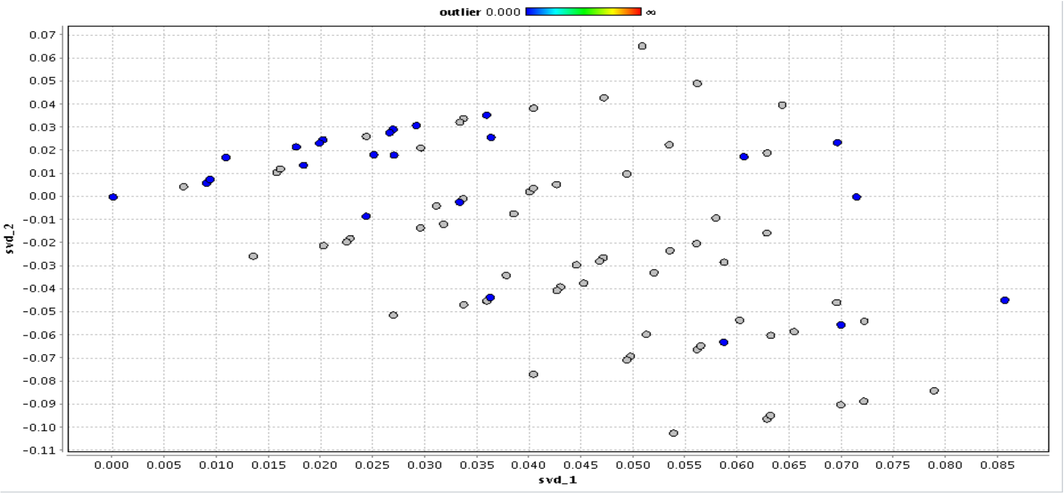


Fig. 7 Outlier (LOF) plot with SVD applied.

V. RESULTS DISCUSSION AND ANALYSIS

The data set of 1173 students used in this study was obtained at University College of Science and Technology (UCST) – Khan Younis enrolled student.

When analyzing the results of applying the association and classification methods, it is apparent as discussed in sections 4 that there exists a direct relationship between the grade and the specification. The classification model can be used to predict the categories of the target class of grade. The accuracy of the models suggests using Rule-Based classifier for predicting the grade since it has the accuracy of 61.88%, while decision tree has the accuracy of 41.88%, the Naive Bayesian algorithm has the accuracy of 58.50%, and K-NN algorithm has the accuracy of 54.37%. The low accuracy values of the classification models that resulted would recommend using alternate classification algorithm or enhance the quality of data set during the preprocessing phase.

When analyzing clustering and outlier results, the same knowledge can be induced by both methods. Figure 5, 6 and 7 present graphs resulted from the clusters and outlier methods, it is apparent that the cluster labeled "cluster_0", which represents 29.67% of the total records. Also, the outliers based on distance presented in Figure 6, and the outliers based on density presented in Figure 7, which represents a too large percentage of the outlier.

VI. CONCLUSION AND FUTURE WORK

In this paper data mining methods are applied at University College of Science and Technology (UCST) – Khan Younis enrolled students. The goal of this paper is to present how data mining can help solve low students' performance in UCST, by discovering patterns and building a classification model that relates grade as well as other affecting factors. Knowledge discovery processes were applied which included pre-processing, data mining, patterned evaluation, and knowledge representation. The classification model has resulted which is a basis for predicting students' performance, but the model's accuracy is not high. The paper addresses the issue of data quality as the main limitation that undermines producing an accurate classification model or generating useful patterns out of the data mining methods.

In future experiments, we want to measure the compressibility of each classification model and use data with more information about the students (i.e. profile and curriculum) and of higher quality (complete data about students that have done all the course activities). In this way, we could measure how the quantity and quality of the data can affect the performance of the algorithms.

References

[1] AbuTair M., and El-Halees A., "Mining Educational Data to Improve Students' Performance: A Case Study", *International Journal of Information and Communication Technology Research*, Volume2 No.2, February2012, ISSN2223-4985.

[2] Ayesha S., Mustafa T., Sattar A., and Inayat M., "Data Mining Model for Higher Education System", *European Journal of Scientific Research*, vol. 43, no. 1, pp. 24-29, 2010.

[3] Paris I., Affendey L., and Mustapha N., "Improving Academic Performance Prediction Using Voting Technique in Data Mining", *World Academy of Science, Engineering and Technology*, 2010.

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

[5] Bekele R., and Menzel W., "A Bayesian Approach to Predict Performance of a Student (BAPPS): A case with Ethiopian students", *in Proceedings of the International Conference on Artificial Intelligence and Applications (AIA-2005)*, Vienna, Austria, 2005.

[6] Bidgoli M., Kashy D., Kortemeyer G., and PunchBehrouz W., " Predicting Student Performance: An Application of Data Mining Methods with the Educational Web-based System lon-capa", *33rd ASEE/IEEE Frontiers in Education Conference*, IEEE, 2003.

[7] El-Halees A., "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, 2008.

[8] Han J. and Kamber M., "Data Mining: Concepts and Techniques", *Morgan Kaufmann*, 2000.

[9] Kumar V., and Chadha A., "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, 2011.

[10] Mannila H., "Data Mining: Machine Learning, Statistics, and Databases", *IEEE*, 1996.

[11] Nghe N., Janecek P., and Haddawy P., "A Comparative Analysis of Techniques for Predicting Academic Performance", *ASEE/IEEE Frontiers in Education Conference*, pp. T2G7-T2G12, 2007.

[12] Olson D., and Delen D., "Advanced Data Mining Techniques ", *ISBN 978-3-540-76917-0*, 2008.

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

[14] Romero C., Ventura S., and García E., "Data Mining in Course Management Systems: Moodle Case Study and Tutorial", *Computers & Education*, vol. 51, no. 1, pp. 368-384, 2008.