[1] **Ahmed Abdel-motlab Ahmed,**

[2] **Hythem H.Gamr Eldeen,**

[3]**Osman A.Nasr,**

[4] **Saeed Al-Khalidi,**

[5] **Elaf Ali Alsisi**

The Application of Clustering Algorithms for the Evaluation of Departmental Performance Based on Student Academic Outcomes

Abstract: - The evaluation of student knowledge is influenced by multiple factors, including their performance in academic courses, capstone projects, seminars, and group discussions. Assessments of student performance were gathered through various methods and types throughout their tenure at the university. The results derived from these evaluations serve to ascertain both departmental and collegiate performance. The present study seeks to appraise departmental performance by analyzing student outcomes from a range of assessments, specifically utilizing k-means and k-medians algorithms in the context of clustering methodologies. This study's methodology involves normalizing data and employing mass spectra mining to generate performance clusters. It was determined that the aforementioned algorithms effectively produced departmental performance clusters and served as significant clustering tools due to their established robustness against outliers. The dataset utilized was compiled from three departments over a two-year period (2021-2022) from computer science and information technology college in neelain university. The employed methodology proves to be highly effective in classifying data across any form of textual information, and it holds potential for application in similar investigative endeavors.

Keywords: Data mining; student performance; k-means, K-medians; clustering.

I. INTRODUCTION *(HEADING 1)*

According to the global university rankings (www.timeshighereducation.co.uk/world-universityrankings), multiple performance indicators are employed to assess academic performance, which are categorized into five dimensions: teaching (30%), research (30%), research citations (30%), industry innovation (2.5%), and international outlook (7.5%). The present study utilizes these indicators to evaluate the departmental performance of a school at Al-Neelain University. Data mining is characterized as a process for extracting and uncovering valuable information from extensive databases, where clustering serves as a significant technique within this data mining framework.[1] The significance of clustering lies in its utility for scenarios wherein objects or events are defined relative to specific attributes, making it advantageous to identify groups of objects based on these attributes.[2] This paper emphasizes the clustering of normalized data sets, recognizing that data is frequently normalized prior to clustering to eliminate inconsequential scale differences . In the context of clustering a text corpus, each document is typically represented as a point, with each dimension corresponding to word frequencies in the data format used for clustering. Moreover, the term "cluster centers" is commonly employed to denote prototypical points, wherein it is often preferable to normalize the cluster centers derived from normalized data. This concept has been thoroughly documented, as evidenced by Florin (2011) and Suchita and Rajeswari (2013). Florin (2011) notes that the spherical K-means algorithm is particularly apt for this purpose. As articulated by Witten and Frank (2005), under specific conditions, the application of a 1-norm distance (also referred to as Manhattan distance, denoted here as k∗k1) for measuring the distance between points is advisable. This is predicated on the fact that the cluster center minimizing the 1-norm distance to all points within that cluster corresponds to the median of that cluster; leveraging the median instead of the mean is generally regarded as being more resilient to outliers.[3] Hence, when considering these aspects, the K-medians algorithm is recognized as a robust alternative. However, employing K-medians on normalized data presents unique challenges concerning the identification of normalized locally optimal cluster centers. We propose

[1] Business Informatics , King Khalid University, Abha, Aseer, Saudi Arabia, aabdelmotlab@kku.edu.sa

[2] Information System, Baisha University, Baisha, Aseer, Saudi Arabia, hythemo@ub.edu.sa

[3] *Corresponding author Business Informatics , King Khalid University, Abha, Aseer, Saudi Arabia, oanassr@kku.edu.sa

[4] Business Informatics , King Khalid University, Abha, Aseer, Saudi Arabia, salkalidi@kku.edu.sa

[5] Business Informatics , King Khalid University, Abha, Aseer, Saudi Arabia, ealsisi@kku.edu.sa

the Manhattan Normalization (MN) algorithm, which, when combined with K-medians, effectively addresses these challenges. [4] A thorough discussion of these algorithms will be provided in the subsequent section. The current study is centered on a comparative analysis of various data mining clustering methodologies, specifically K-means clustering and K-medians. Each algorithm adopts distinct methodologies, which influences the final results based on the selected approach.[5] The primary aim of this paper is to evaluate the K-means and K-medians algorithms as clustering techniques to classify students alongside their final outcomes, thereby measuring the academic department's performance. The analysis will be performed using R software, with comprehensive details regarding this software outlined in Section 3. The structure of the paper is organized as follows: Section 2 offers an overview of clustering techniques in data mining, providing a detailed explanation of both the K-means and K-medians algorithms. Section 3 elaborates on our application study, delineating the dataset employed and the software utilized in the analysis. Section 4 encompasses the evaluation of the algorithms and the corresponding results, accompanied by a discussion of the findings. Lastly, Section 5 concludes the paper.

II. MINING CLUSTERING

As stated earlier, a major issue concerning data mining analysis is clustering. This term refers to the process of grouping the data into classes or clusters, so that objects within a cluster have high similarity in comparison to one another but are very dissimilar to objects in other clusters [6] In other words, cluster is a collection of data objects, such that the objects within a group are similar (or related) to one another and different from (or unrelated to) the objects in other groups.[7] The unsupervised assignment of elements, which is a problem of any given set of group or cluster points, is the objective of cluster analysis. There are many approaches to deal with this problem, including statistical issues[8], machine learning, and mathematical programming [7]. In fact, there are a variety of clustering tools that can be explored. The subsections that follow provide a review of the algorithms that are used in this study.

A. K-means algorithm

This algorithm is a well-known clustering technique. It is widely and frequently used when we need to find cluster centers that minimize the total of the squared 2-norm distance (also known as Euclidean distance), denoted here by k∗k2 from each point to its closest cluster center. Since finding globally optimal cluster centers is an NP-hard problem, K-means may be used to find a local solution. In order to carry out this algorithm, the number of clusters should be chosen to find an initial set of cluster centers. This is because the K-means algorithm splits objects into a data set with a fixed number of K disjoint subsets. Moreover, for each cluster, the splitting algorithm maximizes homogeneity [12]. Hand et al., 2001 noted that there are many different routes for selecting initial cluster centers. However, in this article, the work holds regardless of which technique is used. In what follows, we utilize the K-means framework, following the mathematical justifications given by Barros and Verdejo (2000). Now, let x1,x2,...,xn be a set of points. We can split them into K disjoint clusters that minimize the objective function:

$$\Phi(\{\pi_j\}_{j=1}^{k}) = \sum_{j=1}^{k} \sum_{x \in \pi_j} \| x - c_j \|_2^2, \tag{1}$$

where for each cluster j, cj is the center of each cluster, which is defined as the point for which $\sum_{x \in \pi_j} \| x - c_j \|_2^2$ is minimized. This has been proven to be accomplished by choosing cj to be the centroid of cluster j, expressed by:

$$c_j = \frac{1}{n_j} \sum_{x \in \pi_2} x, \tag{2}$$

where nj denotes the number of points in cluster j. Therefore, we can find the clusters that fix the following minimization problem:

$$\{\pi_i^*\}_{j=1}^{K} = argminQ(\pi_{j\,j=1}^{K}) \tag{3}$$

In order to find these clusters, we start with an intermittent splitting of the data $\{\pi_j^{(0)}\}_{j=1}^{K}$. We additionally define the cluster centers associated with this partitioning as $\{c_j^{(0)}\}_{j=1}^{K}$, then the index of iteration is specified as t = 1. Thereafter, the following steps are needed: (1) for each point, find the cluster center with the closest Euclidean distance. This gives a new partitioning as follows:

$$\pi_j^t = \{x : \|x - c_j^t\|_2^2 \leq \| \, x - c_j^t \, \|_2 \, 2, 1 \leq i \leq K\}, j = 1, ...K, \tag{4}$$

where ties between clusters are resolved by random assignment to one of the optimal centers. Note that this is guaranteed not to increase the objective Q, since each point is assigned to its closest center. (2) Compute the new set of cluster centers {ctj}j=1K by computing the mean (centroid) of each cluster. Since the centroid is the point that minimizes the total distances from all points, this step is also guaranteed not only to increase the objective Q. (3) If a stopping criterion is met, report {φtj}j=1K as the final partitioning and {ctj}j=1K as the final cluster centers. Otherwise, increment t by 1, and go to step 1 above. A variety of stopping conditions are available. One common condition is to stop when the difference between successive values of the objective Q is less than a small tolerance. Finally, we submit that the K-means algorithm is ultimately a local optimization algorithm for minimizing clustering error, where clustering error is defined as the total squared Euclidean distance from each point to its closest center. The objective Q never increases from one iteration to the next. Since K-means is applied to a finite number of points, the algorithm must therefore terminate.

B. K-medians algorithm

Now, we turn our attention to reviewing the K-medians algorithm, which is fitted when we need to minimize the total 1-norm distance from each point to its nearest cluster center. More details of this algorithm can be found in Tan et al., (2002) as well as Witten and Frank (2005). K-medians is quite similar to K-means, meaning it plays a very close role to that based on the K-means algorithm. However, there have been reported some differences between them. These differences are provided below. Since we now work with 1-norm distance instead of squared Euclidean distance, our objective is stated as:

$$\Phi()\{\pi_j\}^k j = 1 = \overset{k}{\underset{j=1}{XX}} \overset{}{\underset{x \in \pi_j}{}} \mathrm{k}\, x - c_j \,\mathrm{k}_1 . \tag{5}$$

We start with a partitioning of the data as that from the K-means algorithm, taking into our account the following steps. (1) We initially set t = 1, then for each point, find the closest cluster center as measured via 1-norm distance. (2) Compute the new set of cluster centers {ctj}j=1K by computing the median of the cluster. In other words, for each dimension, compute the median value for that dimension's overall points in the cluster. The median has been used because it is the point that minimizes the total 1-norm distance from all points [12]. (3) Terminate if the stopping condition is met. Increment t and go to step 1 otherwise. In a similar fashion to the K-means algorithm, steps 1 and 2 of the K-medians algorithm are guaranteed not only to increase the objective Q.

III. APPLICATION STUDY

A. Data description

The student data set that has been used for clustering in this paper consists of 108 records divided into three classes that represent three departments in the School of statistic, Faculty of Mathematical Science and Statistics. The data includes all the academic information regarding the student in the school. Particularly, the analyzed data set consists of the following five numeric variables and a nominal column.

Deg1: represents students' degree in the first year,

Deg2: represents the students degree in the second year,

Deg3: represents the students degree in the third year,

Deg4: represents the students degree in the forth year,

Deg5: represents the students' degree in the fifth year and department factor with three levels as the class label.

A summary of numeric columns is displayed in Table 1. Figure 1 (a) shows the distribution of students in departments and (b) shows the distribution of students marks in the departments.

Table 1: Dataset summary

Measurement	Deg1	Deg2	Deg3	Deg4	Deg5
Min.:	52.18	52.41	52.077	52.10	54.34
1stQu.:	59.81	59.96	61.34	59.28	61.80
Median:	67.06	64.77	66.69	64.60	67.14
Mean:	65.57	65.35	66.73	66.59	68.24
3rd Qu.:	70.38	70.19	72.38	73.12	72.92
Max.:	78.77	87.62	80.92	89.07	92.29

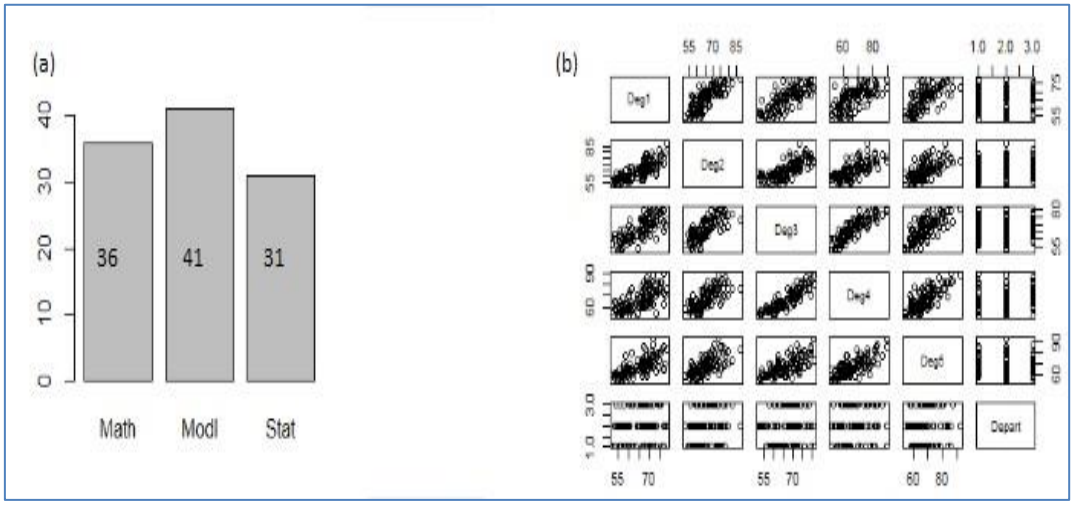

Figure 1: Students distribution

B. The R software

is a free software environment for statistical computing and graphics. This software presents a wide variety of statistical and graphical techniques. Also, it can be extended easily via packages. there are around 4000 packages available in the CRAN package repository[11]. R is widely used in both academia and industry. Note that this software would be a primary issue in the current analysis for implementing both, the K-means and K-medians algorithms[12]. In order to assist users in getting suitable R packages to implement, the CRAN Task Views can be considered as good guidance. Particularly, these views give collections of packages for different tasks. Below, we outline several task views linked with data mining are:

- Machine Learning & Statistical Learning;
- Cluster Analysis & Finite Mixture Models;
- Time Series Analysis;
- Multivariate Statistics; and

IV. Results and discussion

As mentioned earlier, the purpose of this article is to present a comparative study of two algorithms dealing with the clustering data mining analysis. The findings are summarized in the following seven tables. Here, we have to make it clear that we have firstly removed the class Depart to implement the K-means algorithm. The number of clusters has been set to 3. The results are shown below. Table 2 presents the cluster means, under different clusters sizes, 15, 42, 51. The clustering result is then compared with the class label (Depart) to check whether or not similar objects are grouped together.

Table 2: Clusters

Cluster no	Deg1	Deg2	Deg3	Deg4	Deg5
1	72.86733	75.37867	77.23200	81.66933	80.48267
2	69.50048	68.68619	70.56214	69.91071	70.75952
3	60.17725	59.65784	60.48569	59.42431	62.56627

Table 3: Dataset summary

Depart	1	2	3
Math	3	12	21
Modl	5	15	21
Stat	7	15	9

Based on the findings in Tables 1 and 2, we find the following: (1) the performance of students in Cluster 1 is increasing and their grades are excellent. In this Cluster, there are 7 students in the Stat Department represented 23% of the whole Department students, 3 students in the Math Department representing 8% of the Department students and 5 students in the Modl Department representing 12% of the Department students. (2) The performance of students in Cluster 2 is swinging but their grades are generally good. In this Cluster, there are 15 students in the Stat Department representing 48% of the Department students, 12 students in the Math Department representing 34% of the Department students and 15 students in the Modl Department representing 29% of the Department students. (3) As well, the performance of students in Cluster 3 is again swinging and their grades maybe are good. In this Cluster, there are 9 students in the Stat Department representing 29% of the Department students, 21 students in the Math Department representing 58% of the Department students and 21 students in the Modl Department representing 51% of the Department students. Overall, more than 50% of students in both the departments, Math and Modl, are in cluster 3, which is to say that 50% of students in the Stat department are in cluster 2. Higher percentages of students in the best cluster (cluster 1) belong to the Stat department as well as the Modl department. Finally, compared to other departments, the performance of students in the Stat department was the best.

The results of the K-medians algorithm are given in Tables 4 and 5. Again, the number of clusters was specified to be 3, and the clustering results are compared with the class label (Depart).

Table 4: Cluster medians

Cluster no	Deg1	Deg2	Deg3	Deg4	Deg5
1	67.36	71.05	71.95	75.85	75.00
2	66.67	62.07	65.77	61.00	68.14
3	56.59	56.21	56.97	57.41	60.86

Table 5: Cluster result and class label

Cluster	Math	Modl	Stat
1	9	16	19
2	16	11	8
3	11	14	4

By looking at Table 4, we see that, as we expected, the K-medians results were similar to those obtained by the K-means results, specifically in terms of the performance process. Table 5 reveals that more than 61% of students of the Stat department students were in the best cluster (cluster 1). Thereafter, 39% of students from the Modl department students were in the same cluster, and finally the Math department. The K-medians algorithm findings were better when compared to those based on the K-means results (this situation can be justified by some previous studies, see for example, Tan et al., 2002), however, there are overlapping in the clusters. This overlapping is clearly shown in figure 4. Therapy, it would be better to group the students into two clusters, and to minimize the overlap between the clusters. Tables 6 and below shows the results when the number of clusters was set to 2.

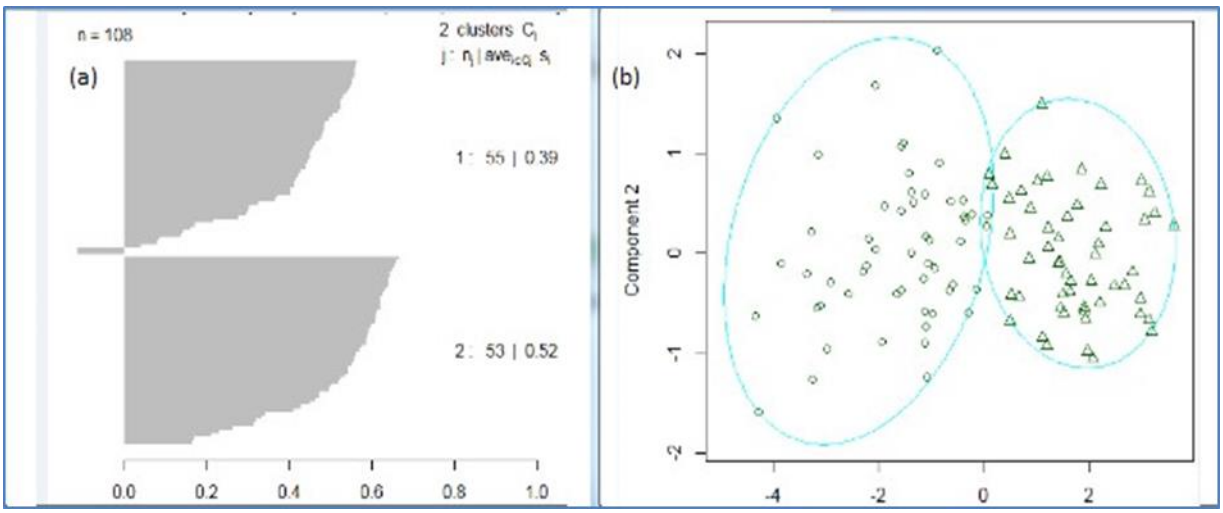

Figure 2: Clusters in K-medians algorithm

Table 6: Cluster medians

Cluster no	Deg1	Deg2	Deg3	Deg4	Deg5
1	68.97	68.82	71.37	69.38	71
2	62.08	59.59	61.74	60.55	65

Table 7: Two cluster result and class label Cluster Math Modl Stat

Cluster no	Math	Modl	Stat
1	13	20	22
2	23	21	9

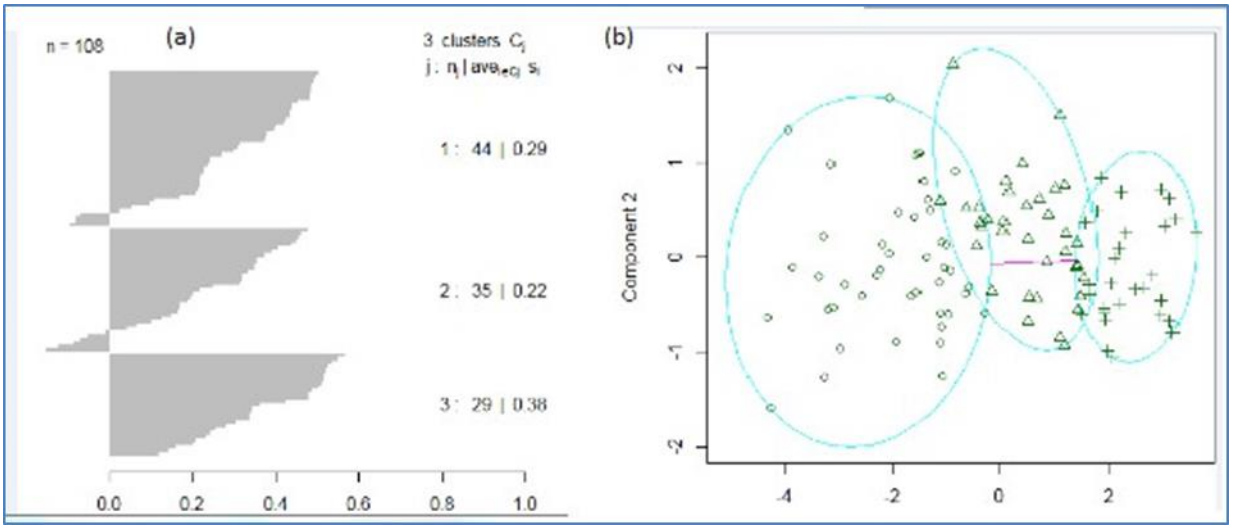

Figure 3: Two clusters in K-medians algorithm

From figure 4, it is clear that there is a little overlapping in the clusters and there are 72% of students of Stat department students located in the best cluster (cluster 1), followed by 49% students of the Modl department students in the same cluster and finally 36% in the Math department. These percentages are presented in figure 4.

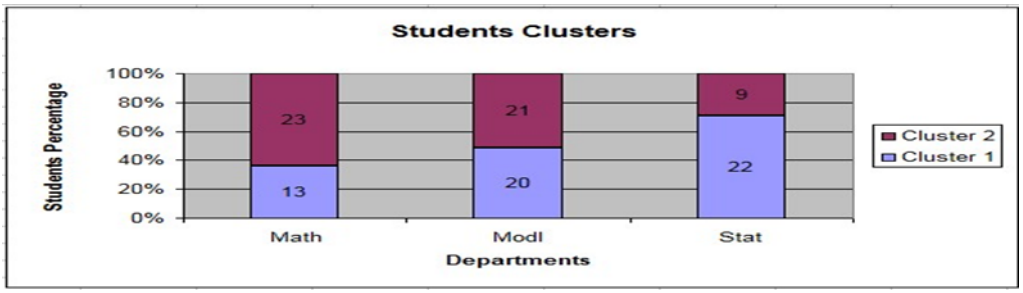

Figure 4: Students percentage in each departmen

V. CONCLUSION

In this article, we have studied the comparison of two clustering techniques applied to a user-made students' data set from the School of Statistics, Al Neelain University. The clustering techniques compared included the K-means and K-medians algorithms. The goal was to apply these algorithms in order to cluster the students and their final results to evaluate the department's performance. The current analysis has been done using the R software. We have finally achieved the following basic concluding remarks: First, there are extreme similarities regarding the students' degrees with computing the means and medians. Second, students' performance in the Stat department is better than the other departments considered in this study, followed by the student's performance in the Modl department and finally those students who were based in the Math department. Finally, the best classification of students is divided into two clusters with K-medians algorithm.

REFERENCES

[1] A. Khan and S. K. Ghosh, "Student performance analysis and prediction in classroom learning: A review of educational data mining studies," Education and information technologies, 2021. [HTML]

[2] C. F. Rodríguez-Hernández and E. Cascallar, "Socio-economic status and academic performance in higher education: A systematic review," *Educational Research*, vol. 2020, Elsevier. uantwerpen.be

[3] S. Caeiro, L. A. Sandoval Hamón, R. Martins, et al., "Sustainability assessment and benchmarking in higher education institutions—A critical reflection," *Sustainability*, vol. 12, no. 1, 2020. mdpi.com

[4] M. Yağcı, "Educational data mining: prediction of students' academic performance using machine learning algorithms," Smart Learning Environments, 2022. springer.com

[5] R. Hasan, S. Palaniappan, S. Mahmood, A. Abbas, "Predicting student performance in higher educational institutions using video learning analytics and data mining techniques," *Applied Sciences*, vol. 10, no. 1, 2020. mdpi.com

[6] Florin, G. (2011). Data mining: concepts, models and techniques. Springer-Verlag. Berlin Heidelberg.

[7] Jiawei, H., Micheline, K. and Jian P. (2012). Data mining concepts and techniques, Third edition. Elsevier Inc: USA.

[8] Barros, B. and Verdejo, M. F. (2000). Analysing student interaction processes in order to improve collaboration: the degree approach. International Journal of Artificial Intelligence in Education, 11, 221-241.

[9] Tan, P.-N., Kumar, V., and Srivastava, J. (2002). Selecting the right interestingness measure for association patterns. In KDD ’02: Proceedings of the 8th ACM SIGKDD

[10] Linh, N. T. and Chau, C. C. (2013). Application of CURE data clustering algorithm to batangas state university student database. International Journal on Advances in Computing and Communication Technology, 4, 108-115.

[11] Venables, W. N., Smith, D. M. and R Development Core Team (2010). An introduction to R. R foundation for statistical computing, Vienna, Austria. ISBN 3-900051-12-7.

[12] Raviya, K. H. and Dhinoja, K. (2013). An Empirical Comparison of K-Means and DBSCAN Clustering Algorithm. PARIPEX Indian Journal of Research, 4, 153-155.