Intelligent Data Analysis 23 (2019) 1055–1071
DOI 10.3233/IDA-184254
IOS Press

Evaluation of data analytics based clustering algorithms for knowledge mining in a student engagement data

O.O. Oladipupo and O.O. Olugbara*
ICT and Society Research Group, Durban University of Technology, Durban 4000, South Africa

Abstract. The application of algorithms based on data analytics for the task of knowledge mining in a student dataset is an important strategy for improving learning outcomes, student success and supporting strategic decision making in higher educational institutions of learning. However, the widely used data analytics based clustering algorithms are highly data dependent, making it pertinent to find the most effective algorithm for knowledge mining in a dataset associated with student engagement. In this study, performances of five famous clustering algorithms are evaluated for this purpose. The k-means algorithm was benchmarked with 22 distance functions based on the Silhouette index, Dunn's index and partition entropy internal validity metrics. The hierarchical clustering algorithm was benchmarked with the Cophenetic correlation coefficient computed for different combinations of distance and linkage functions. The Fuzzy c-means algorithm was benchmarked with the partition entropy, partition coefficient, Silhouette index and modified partition coefficient. The k-nearest neighbor algorithm was applied to determine the optimum epsilon value for the density-based spatial clustering of applications with noise. The default parameter settings were accepted for the expectation-maximization algorithm. The overall ranking of the clustering algorithms was based on cluster potentiality using the median deviation statistics. The results of the evaluation show the well-known k-means algorithm to have the highest cluster potentiality, demonstrating its effectiveness for the task of knowledge mining in a student engagement dataset.

Keywords: Algorithm evaluation, data analytics, data clustering, knowledge mining, student engagement

1. Introduction

The advent of big data era coupled with an explosively growing mass of data and development of technology for data analytics has brought about an increasing necessity for educational enterprises to discover useful knowledge in student datasets. Amongst the intrinsic merits of knowledge mining in the education domain is to unveil hidden facts that would help to improve learning outcomes, student success and support strategic decisions pertaining to effective management of students. The educational enterprises, particularly the higher educational institutions of learning worldwide, generate, share and store a monumental volume of data at incredible rates, which present a real data management challenge. It is customarily the case that higher educational institutions are awash in data about students, staff, research activities and other daily business transactions. Moreover, many educational institutions are increasingly

*Corresponding author: O.O. Olugbara, ICT and Society Research Group, Durban University of Technology, P.O. Box1334, Durban 4000, South Africa. E-mail: oludayoo@dut.ac.za.

delivering online learning that has resulted in trillions of data being harvested and stored in various institutional datasets [1]. In particular, a student engagement dataset is crucial because student engagement is affirmed in literature as multifaceted constructs for understanding educational concerns like enrollment planning, student dropout and is positively related to academic performance [2]. In addition, it has been described as a ductile aspect of intrinsic motivation and behavior of students that are useful for learning and adjusting the institutional context [3]. Analytically, a student engagement dataset can assist students and teachers in recognizing the precarious activities as danger signs before threatening learning, academic success and full engagement [4,5]. Moreover, a student engagement dataset interconnects many other educational datasets such as admission dataset, finance dataset, academic records, research dataset and accommodation dataset.

A student engagement dataset is a set of records that reflects the quality of effort, time and energy that students, staff, faculty and institutions have committed to educational events that directly contribute to enhancing student success [6–9]. It can be monumental, depending on the size of an academic institution and learning activities that directly impact on students. In view of the importance of student engagement, the National Survey of Student Engagement (NSSE) carried out surveys in different colleges and universities all over the world in 2000. In 2017, 725 colleges and universities participated and 517,850 students completed the surveys. Since 2000, over 1,600 institutions have participated and approximately 6 million students have completed the surveys (http://nsse.indiana.edu/html/about.cfm). The main purpose of the NSSE surveys was to provide participating educational institutions with data to detect aspects of student engagement that should be enhanced through a change in policies and practices that concretely aligned with the standard practices in undergraduate education. In addition, information from the survey reports is widely used by researchers and stakeholders to learn more about how students spend their time and what they gain from their education experiences (http://nsse.indiana.edu/html/about.cfm).

Researchers and higher educational institutions have applied different methods to analyze a student engagement dataset in order to realize an outstanding purpose of detecting aspects of student engagement that should be enhanced to align with the standard practices of undergraduate education. However, many of these methods come with the intrinsic curbs that limit their wide applications for knowledge mining. For instance, they can analyze numerical variables based on groups, but cannot identify individual characteristics [10]. In addition, they can generate ambiguous results [11] and cannot reliably obtain inference that is useful for early detection of student defects. The use of methods based on data analytics is potentially valuable for improving student success and discovering useful knowledge that would help to enhance student management. Data analytics can extract meaningful knowledge in raw data and unveil hidden facts that can assist in understanding academic challenges facing students. Moreover, it is a useful device for learning and gaining intuition about student engagement [12]. This will in turn help unveil some aspects of undergraduate student education practices and activities that should be enhanced through a change in policies and practices that are congruent with the goals of undergraduate education. The unique contributions of the study at hand, lie in the above vantage position as succinctly articulated below:

(a) Five famous unsupervised clustering algorithms for data analytics are experimentally evaluated to discover the best cluster structure for knowledge mining in a student engagement dataset. The choice of unsupervised learning algorithms such as clustering over supervised learning algorithms such as neural networks is paramount because many educational datasets are often unlabeled and cannot readily be used to fit models.
(b) Since the effectiveness of the k-means algorithm is heavily dependent on the distance function utilized, 22 distance functions were tested to discover the most suitable of them for the task of knowledge mining in a student engagement dataset.

(c) The optimum number of clusters was determined based on the widely used internal validity metrics of Silhouette, NbClust and Elbow. The determination of an appropriate number of clusters is a prime problem in the application of data clustering algorithms.
(d) A methodological framework has been developed in this study to rank different data clustering algorithms based on the cluster potentiality to discover the most suitable algorithm for the task of knowledge mining in a student engagement dataset.

2. Related literature

The discussion of related literature is succinctly organized in two dimensions in order to show currency, originality, relevance and relatedness of this study with respect to the extant research and to justify the suitability of the study methods. These dimensions are methods that have been previously used to study student engagement and data clustering.

2.1. Student engagement methods

The literature reveals different methods, models, frameworks and tools that scholars have engaged to assess and understand student engagement [13,14]. These methods include self-report [15], teacher rating [16], interview [17], observation [18] and experimental sampling [19]. Moreover, various methods have been deployed to analyze a dataset associated with student engagement. They include constant comparative method [15], process-oriented analysis [20], analysis of variance [21], content analysis [22], structural equation modeling [23,24], exploratory factor analysis, confirmatory factor analysis [25], cognitive engagement model and quadrant analysis [26], correlation analysis [27], descriptive statistics and econometric analysis [28]. This current review shows that these efforts have demonstrated potentials and produced reliable reports. Nevertheless, descriptive statistics and some of the already engaged methods can only analyze numerical variables based on groups, but cannot identify individual characteristics [10]. For example, factor analysis can generate ambiguous results [11] and cannot reliably obtain inference that is useful for early detection of student defects. However, an approach of data analytics with its great potential to learn and gain intuition about data has not really been well-explored for the task of knowledge mining in a student engagement dataset to the best of our understanding, which is a gap that has inspired this study.

2.2. Data clustering methods

Data clustering is an exploratory unsupervised learning task that classifies a set of data objects into groups, such that objects are homogeneous within each group and heterogeneous between groups. Each group is referred to as a cluster, such that objects in a cluster have high similarity in properties, but are very dissimilar to objects in other groups [29]. Clustering is important for identifying hidden patterns, revealing previously unknown knowledge and giving a better understanding of each distinct cluster from a huge volume of data [12]. Different clustering algorithms have been proposed in the literature for data analytics. However, their applicability depends heavily on the type of dataset and essential requirements of the problem space [30,31]. The hiccups often associated with the methods of data analytics have been extensively reviewed in the literature [29,32,33].

There is a difficulty in providing a direct categorization of clustering methods because of overlapping in the categorization [29]. However, data clustering methods could be diametrically categorized into

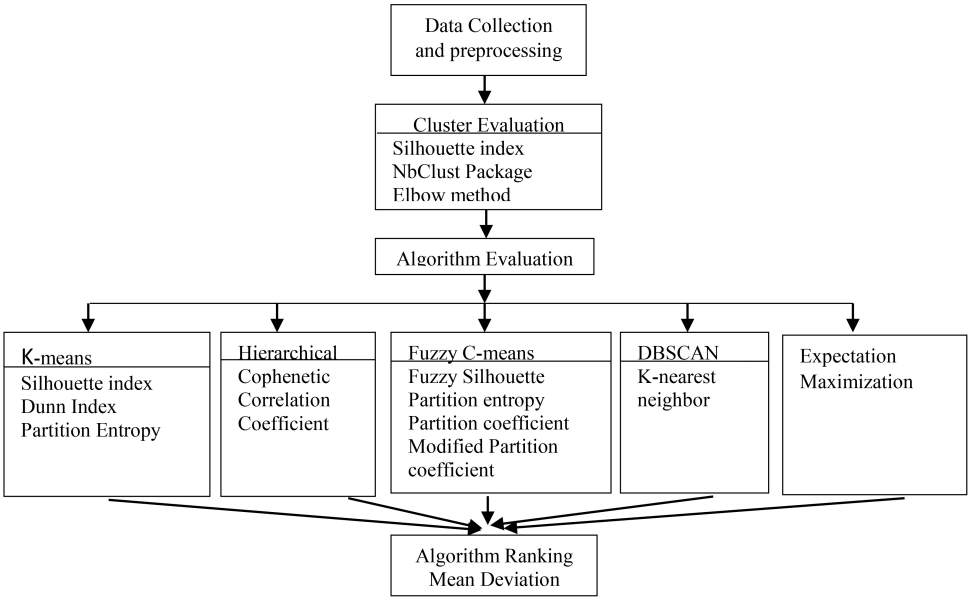


Fig. 1. The methodological framework with the standard indices for ranking data clustering algorithms.

hard clustering and soft clustering. In hard clustering, one data point is mapped exactly into one cluster, while in soft clustering, a data point can be represented in different clusters with the specified membership degrees. In [34], hard clustering methods were further categorized based on the approach of cluster structure modeling. The categories of structural modeling include, connectivity, centroid, distribution and density methods. The connectivity methods such as the hierarchical clustering algorithm, build models based on distance connectivity. The centroid methods such as the k-means algorithm, represent each cluster by a single mean vector. The distribution methods such as the Expectation-Maximization (EM) algorithm, model clusters based on statistical distributions. The density methods such as the Density Based Spatial Clustering of Applications with Noise (DBSCAN) and Ordering Points To Identify Clustering Structure (OPTICS) define clusters as connected dense regions in a data space.

Like other methods of data analytics, clustering algorithms have been broadly investigated and tested on diverse datasets in different application domains with promising results. These application domains include, gene expression data [33], biological data [35], sensor network data [36], medical dataset [37,38], food data [39], student academic data [40], image segmentation [41], market sales data [42,43], network traffic data [44] and residential electricity consumption [45]. However, the implementation performances of clustering algorithms on different datasets differs most likely because of the inherent data dependency issue [29,31,46]. Data dependency is a critical issue in data analytics research that always necessitates experimental comparisons of different methods on a particular dataset to discover the most appropriate one [31,47]. Consequently, the overarching objective of this study, was to evaluate five famous clustering algorithms for the task of knowledge mining in a student engagement dataset and rank the algorithms according to their performances. The clustering algorithms are the k-means, Hierarchical, DBSCAN, EM and Fuzzy c-means as shown in Fig. 1.

3. Material and methods

The methodological framework of this study as shown in Fig. 1 is divided into four essential com-

ponents for evaluating and ranking different data clustering algorithms. These components, which are further expounded are data collection, cluster evaluation, algorithm evaluation and algorithm ranking.

3.1. Data collection

The dataset used in this study was based on the data collected from the 2016 South African online survey on student engagement carried out by the Durban University of Technology (DUT) in South Africa. The DUT is one of the five South African universities that is participating in the Siyaphumelela Saide project focusing on student success at South African universities. The project is funded by the Kresge foundation to the tune of 2.9 million US dollars to support data analytics capability at South African universities. The principal goal of the Kresge foundation is to significantly improve student learning outcomes and success using data (https://www.siyaphumelela.org.za/about.php). The online survey was based on the NSSE framework for student engagement, financial stress and socio-demographic characteristics of students. Data on student engagement and socio-demographic characteristics were used in this study to relate student characteristics with engagement level.

The engagement framework presents four dimensions, which are the academic challenge, learning with peers, experience with staff and the campus environment. These dimensions are further sub-divided to form 10 engagement practices, which are the higher-order learning, reflective and integrative learning, learning strategies, quantitative reasoning, collaborative learning, discussions with diverse others, student-staff interaction, effective teaching practices, quality of interaction and supportive environment. The 10 engagement practices are together measured with 47 activities for proper evaluation of each engagement practice. Moreover, 41 variables describe information on socio-demographic characteristics of students. In total, 88 variables representing the dataset fields were used for a knowledge mining task. The dataset contains 1399 examples that represent perceptions of male and female students of the DUT and they constitute the dataset records. All fields of the dataset are of numeric data type, which makes data preprocessing or processing simple. A lot of records were missing on certain fields, which has resulted in the exclusion of respondents with incomplete information. Data analytics have been performed on 904 records with 88 fields after removing the missing records from the engagement dataset of this study.

3.2. Cluster evaluation

The general problem often associated with the partitioning based data clustering algorithms is to ascertain an appropriate number of clusters [48]. The cluster validity indices of the Silhouette index, NbClust package and Elbow method were examined using the k-means, Hierarchical and Fuzzy c-means algorithms. Moreover, cluster evaluation was applied to obtain the optimum number k of clusters for the k-means algorithm. The Elbow method is based on the sum of squared error (SSE) to measure clustering compactness. The goal of the method is to choose a small value of k at the knee point called Elbow that has a low SSE. It represents a point of diminishing returns when k is increased [49]. However, the Elbow method is sometimes ambiguous, which is the important reason for investigating other methods such as the Silhouette index [31] and NbClust package [50] to complement study results. The Nbclust package evaluated 30 indices for cluster validity over a number of clusters. The aggregate result is represented on a bar chart plot with the recommended number of clusters.

3.3. Algorithm evaluation

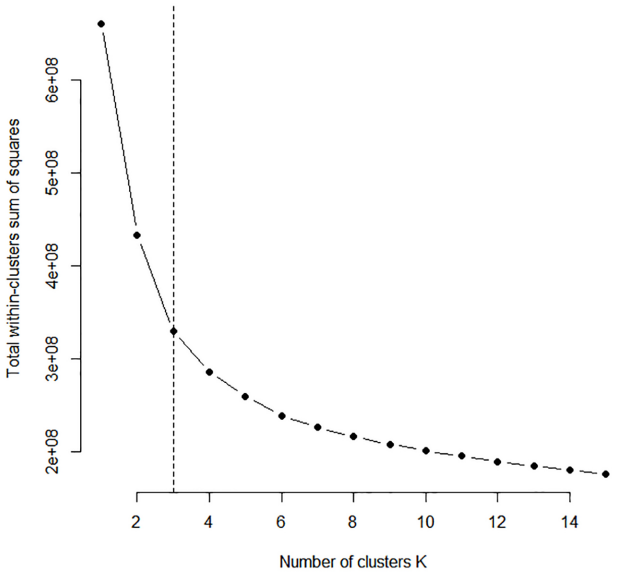
In this study, five different clustering algorithms have been investigated, evaluated and ranked based

on cluster potentiality. The primary reason for selecting these clustering algorithms is because they are famous, they belong to the category of cluster structure modeling and their implementations are readily available in many data analytics software tools. To determine the best cluster structure for the k-means algorithm, 22 distance functions from different families of similarity measures have been investigated [51] in the application domain of student engagement. The potentiality of each distance function with respect to the k-means algorithm was evaluated based on the Dunn, Silhouette and partition entropy internal cluster validity indices. The partition entropy measures the disorderliness within a partition, which implies that the lowest entropy is an indication of the best clustering performance [52]. The Silhouette index [53] measures the quality of clustering by computing the average Silhouette of observations for different values of clusters. That is, it determines how well each object lies within its cluster with a high average Silhouette width indicating a good clustering. The optimal number of clusters is the one that maximizes the average Silhouette over a range of possible values for k. The Dunn's index [54,55] identifies the clusters that are well separated and compact. The goal was to maximize the inter-cluster distance while minimizing the intra-cluster distance concomitantly. A large Dunn's index implies that a compact and well-separated cluster exists. Although there are other external validity methods such as the Rand index, Jaccard and Purity, they were not considered appropriate because the dataset of this study is unlabeled. There is no class label for external evaluation of the clustering results, which is an important property that discriminates supervised learning from unsupervised learning. An unsupervised learning method is required because it is an exploratory phase that is necessary to obtain an initial knowledge of the study area [45].

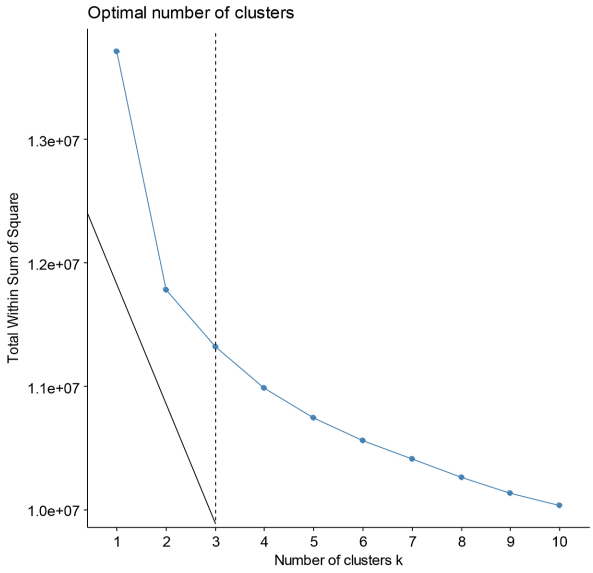
The hierarchical clustering algorithm engages linkage and distance functions for data clustering. Different combinations of 8 linkage functions with 5 distance functions were experimentally evaluated to determine the best performed hierarchical clustering method. The efficiency of the function combinations was evaluated using the Cophenetic correlation coefficient (CPCC), which is a widely used metric in literature [31,56]. The CPCC is the Pearson correlation between the actual distance and predicted distance based on a particular hierarchical configuration. A value of 0.75 or above needs to be achieved for a clustering to be considered useful [57]. The optimal value of epsilon was determined for the DBSCAN algorithm using the k-nearest neighbor distances in a matrix of points [58]. This calculates the average distance of every point to its k-nearest neighbors, where the value of k corresponds to the value of MinPts specified by a user. The k-distance graph is plotted in ascending order to determine the "knee" that corresponds to the optimal epsilon, which is a threshold where the plot shows a strong bend or K-distance curve. In accordance with what entails in literature, the MinPts value of 6 has been used in this study because of the size of the dataset. The default parameter settings were accepted for the EM algorithm that maximizes the likelihood function to estimate the underlying model parameters [59]. Five widely used distance functions were evaluated over 100 maximum iterations for the Fuzzy c-means algorithm. The simulation was repeated 100 times for each distance function, after which the Fuzzy Silhouette index, partition coefficient, partition entropy and modified partition coefficient were considered for comparison. For the best Fuzzy c-means performance, the Fuzzy Silhouette index and partition coefficient should be maximized, while the partition entropy and the modified partition coefficient should be minimized [52,53,60].

3.4. Algorithm ranking

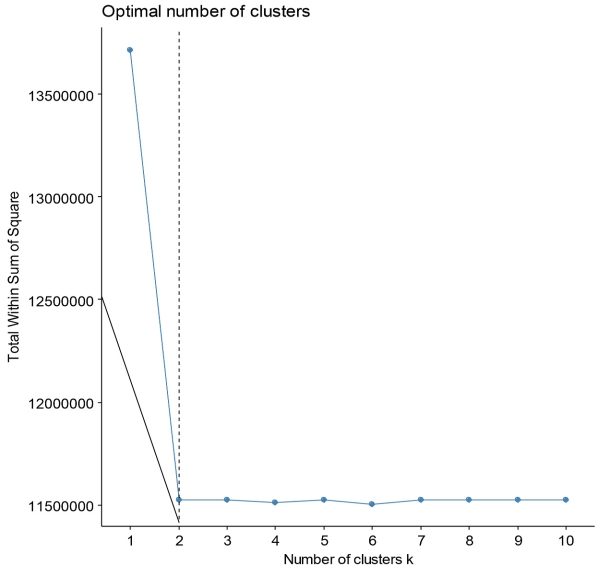
The overall ranking of five clustering algorithms has been done using the average number of data points per cluster [31]. This was achieved using the cluster potentiality based on the mean deviation



(a) K-means Elbow method plot



(b) Hierarchical Elbow method plot

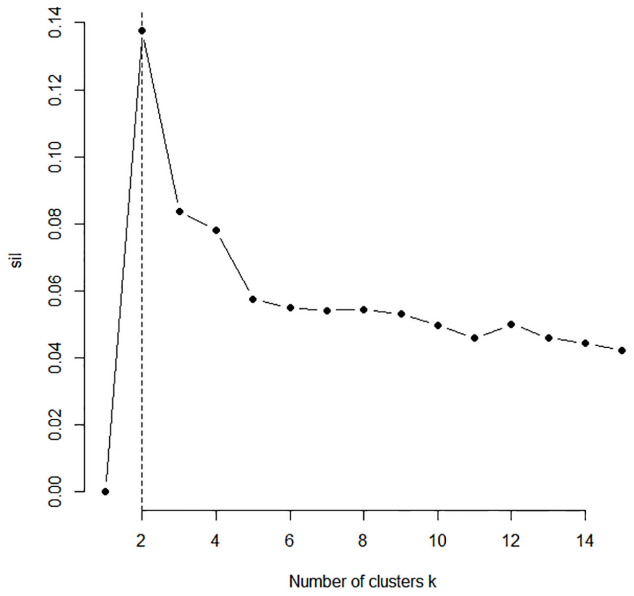


(c) Fuzzy c-means Elbow method plot

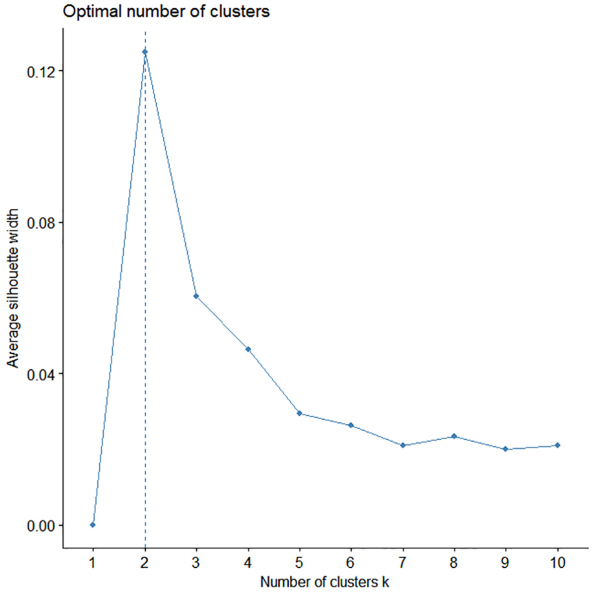
Fig. 2. The within the sum of square error plot for k-means, Hierarchical and Fuzzy c-means clustering algorithms.

(MD) statistics. The lower the MD value, the better the cluster structure, which suggests less deviation of the objects in clusters, as compared to the average number of objects in the corresponding clusters taken collectively [31]. The k-means algorithm with the Pearson correlation coefficient gave the best result, according to the ranking information and is nominated as a suitable method for clustering a student engagement dataset.

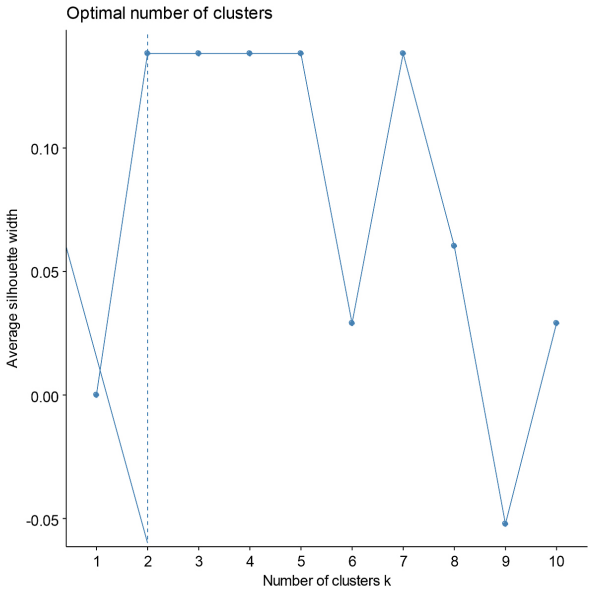
In terms of its operational mode, the k-means algorithm uses as input a dataset of $m \times n$ dimensions, where "m" is the number of records and "n" is the number of fields of the dataset. It must know a priori the number k of clusters to use in minimizing the SSE within each cluster using an appropriate distance



(a) K-means average Silhouette width plot



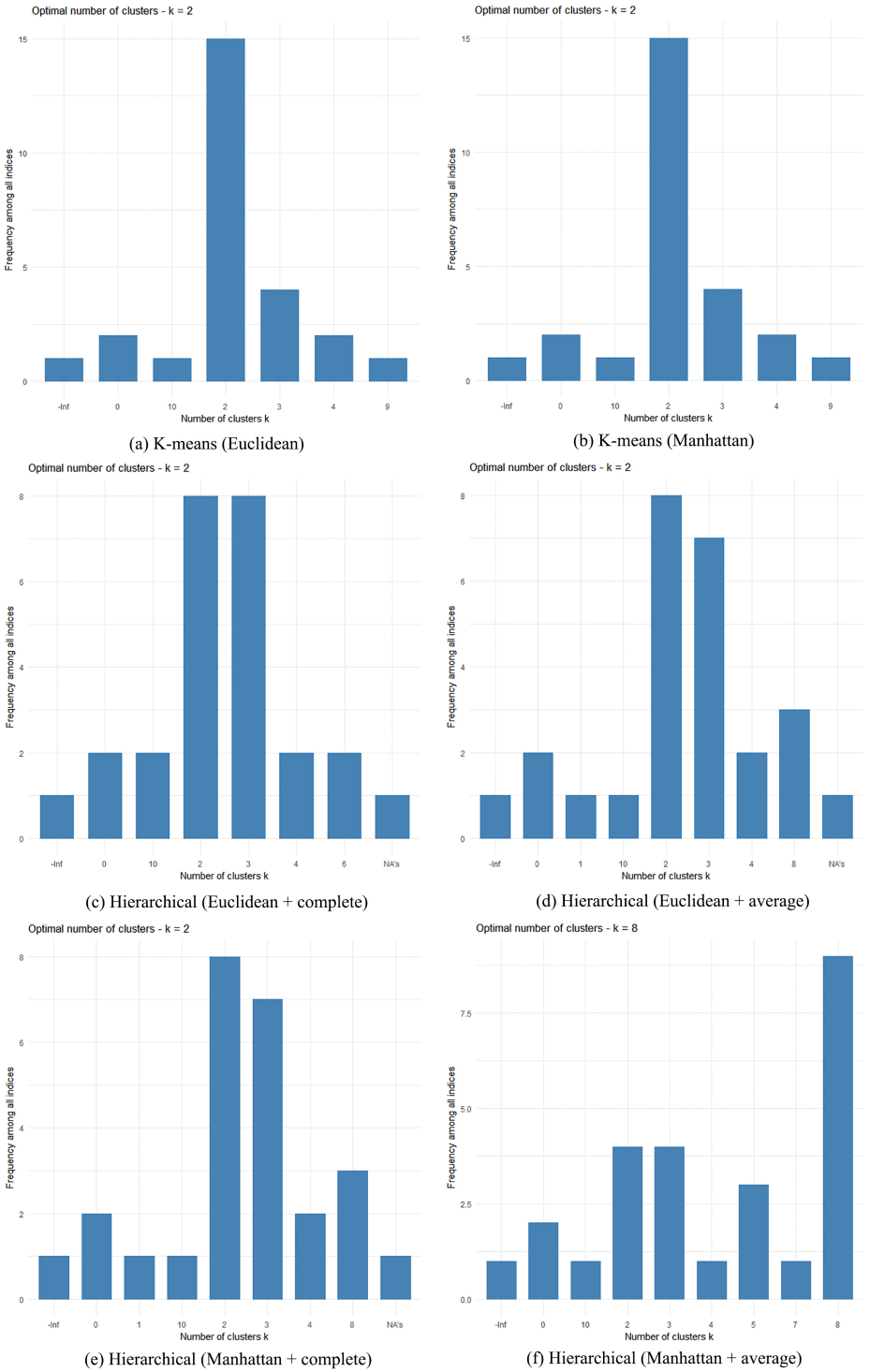
(b) Hierarchical average Silhouette width plot



(c) Fuzzy c-means average Silhouette width plot

Fig. 3. The average silhouette width for k-means, Hierarchical and Fuzzy c-means clustering algorithms.

function. The k-means clustering algorithm is usually implemented in 3 steps. In step 1, k vectors are randomly chosen from the dataset as the initial centroid of each cluster, forming k clusters because each cluster has initially a unique data point. In step 2, the algorithm iterates until it finds a stable state, each of the "m" vectors is assigned to a cluster with the smallest centroid distance. In step 3, the centroid of each cluster is recomputed. If the centroids of the current iteration match those of the previous iteration, the algorithm reaches a steady state and terminates faithfully. However, if the centroids do not match



(a) K-means (Euclidean)

(b) K-means (Manhattan)

(c) Hierarchical (Euclidean + complete)

(d) Hierarchical (Euclidean + average)

(e) Hierarchical (Manhattan + complete)

(f) Hierarchical (Manhattan + average)

Fig. 4. The NbClust graphs representing frequency of 30 indices cluster evaluation output.

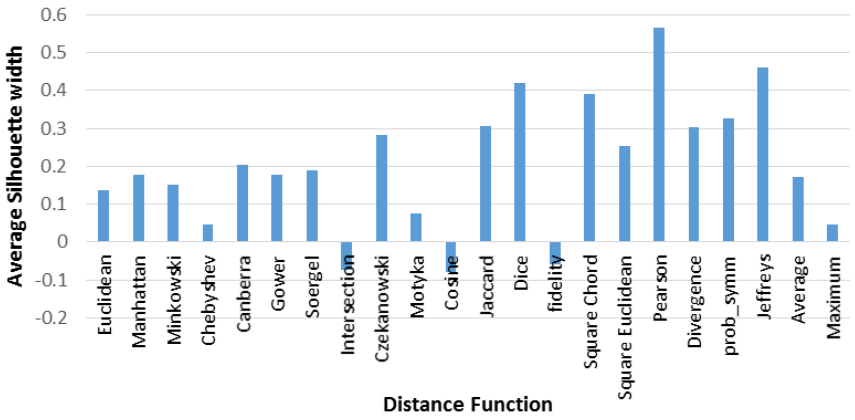


Fig. 5. The Bar chart for average Silhouette width versus distance function for k-means algorithm.

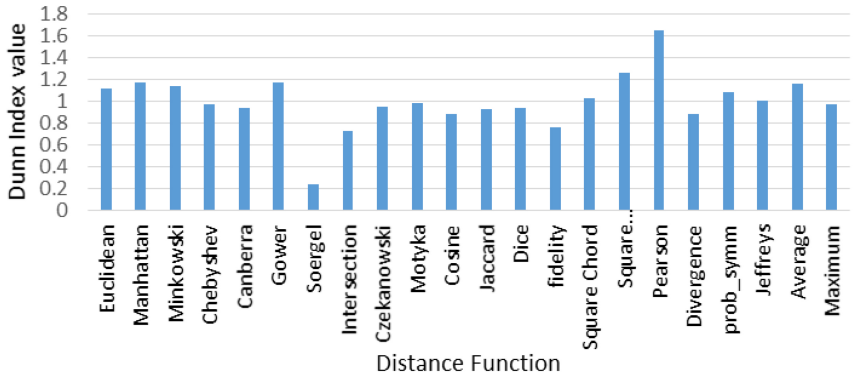


Fig. 6. The Bar chart for average Dunn index value versus distance function for k-means algorithm.

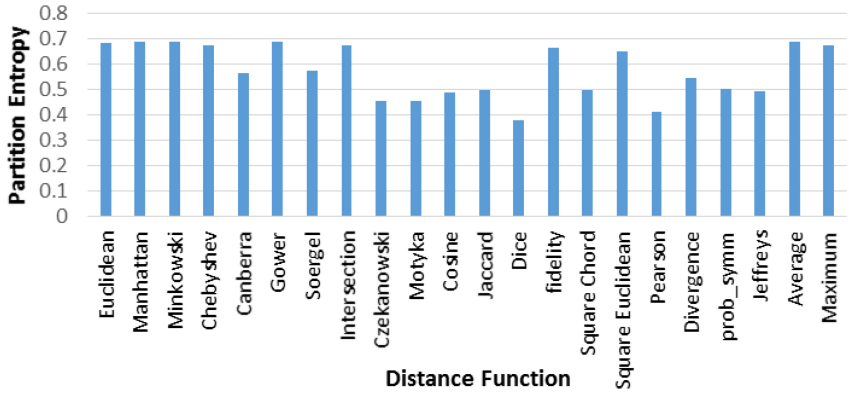
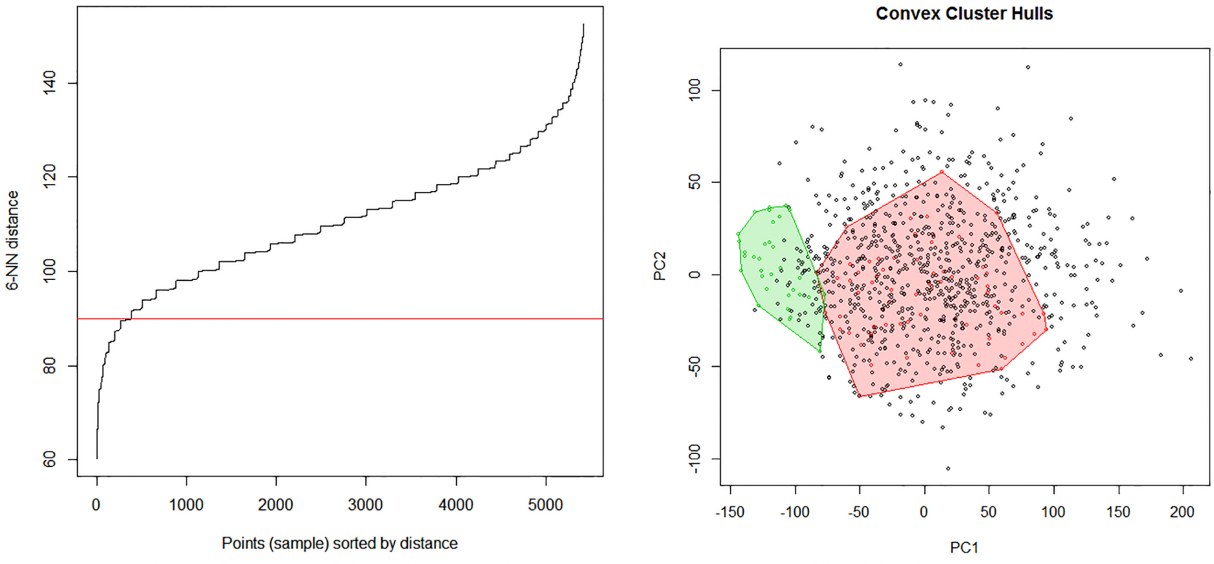


Fig. 7. Bar chart for partition entropy versus distance function for k-means algorithm.

the iteration, the cycle continues by performing steps 2 and 3. The members of each vector are finally obtained as an output. The clustering was implemented using the R studio data analytics environment with more than 750 experiments performed and $m = 904$, $n = 88$, $k = 2$ in this study.

Table 1
The cluster validity indices for k-means algorithm with 22 distance functions where the higher the Silhouette value, Dunn's index and the smaller the entropy value, the better the clustering result

Distance function family	Distance function	Cluster validity indices			Cluster structure
		Silhouette Width	Dunn's Index	Entropy	
Minkowski	Euclidean	0.1381785	1.1172550	0.6837099	390,514
	Manhattan	0.1763842	1.1695550	0.6888239	494,410
	Minkowski	0.1510747	1.1355600	0.6888239	494,410
	Chebyshev	0.0472641	0.9687154	0.6761914	535,369
L1	Canberra	0.2045962	0.9358604	0.5659517	675,229
	Gower	0.1763842	1.1695550	0.6888239	494,410
	Soergel	0.1898759	0.2324031	0.5764540	666,238
Intersection	Intersection	−0.0742717	0.7295518	0.6753595	367,537
	Czekanowski	0.2839971	0.9450667	0.4581940	155,749
	Motyka	0.0758611	0.9848424	0.4581940	749,155
Inner product	Cosine	−0.0795503	0.8800908	0.4882058	173,731
	Jaccard	0.3077045	0.9225273	0.4976297	725,179
	Dice	0.4201069	0.9361561	0.3789171	790,114
Fidelity/square chord	Fidelity	−0.0590359	0.7641854	0.6669539	349,555
	Square chord	0.3901254	1.0264250	0.5007090	723,181
Square L2	Square Euclidean	0.2535445	1.2634900	0.6524962	324,580
	Pearson	**0.5657380**	**1.6531840**	**0.4157387**	**772,132**
	Divergence	0.3033252	0.8782081	0.5459778	213,691
	prob_symm	0.3277734	1.0834260	0.5037577	183,721
Shannon entropy	Jeffreys	0.4599001	1.0015700	0.4945196	727,177
Combination	Average	0.1708036	1.1623960	0.6890276	411,493
	Maximum	0.0472641	0.9687154	0.6761914	535,369



(a) The k-nearest neighbor distances plot (b) DBSCAN cluster structure plot

Fig. 8. The k-nearest neighbor distances in a matrix of points plot for DBSCAN and cluster structure plot.

Table 2
CPCC values for the hierarchical clustering of various combinations of linkages and distance functions

Linkage function	Distance function	CPCC	Cluster structure (cluster_1, cluster_2)
Single	Euclidean	0.3837549	903, 1
	Manhattan	0.3016219	903, 1
	Minkowski	0.3525002	903, 1
	Canberra	**0.8786360**	903, 1
	Maximum	0.2706975	903, 1
Complete	Euclidean	0.4425751	721, 183
	Manhattan	0.4153056	810, 94
	Minkowski	0.4903237	810, 94
	Canberra	**0.5752210**	901, 3
	Maximum	0.2108464	903, 1
Average	Euclidean	0.5596330	894, 10
	Manhattan	0.5341758	882, 22
	Minkowski	0.5619265	882, 22
	Canberra	**0.8823338**	902, 2
	Maximum	0.4027319	903, 1
Centroid	Euclidean	0.4637284	903, 1
	Manhattan	0.3834603	903, 1
	Minkowski	0.4478218	903, 1
	Canberra	**0.8999532**	**903, 1**
	Maximum	0.4291382	903, 1
Median	Euclidean	0.3520795	903, 1
	Manhattan	0.2736511	903, 1
	Minkowski	0.3122837	903, 1
	Canberra	**0.8346402**	903, 1
	Maximum	0.3042970	903, 1
Ward1	Euclidean	**0.4017131**	493, 411
	Manhattan	0.3755526	589, 315
	Minkowski	0.4000701	589, 315
	Canberra	0.1151776	347, 557
	Maximum	0.1640220	362, 542
Ward2	Euclidean	0.4124021	523, 381
	Manhattan	0.4094657	551, 353
	Minkowski	0.3973250	551, 353
	Canberra	**0.4786026**	799, 105
	Maximum	0.1771125	377, 527
Mcquitty	Euclidean	0.4256226	755, 149
	Manhattan	0.4383421	594, 310
	Minkowski	0.4905333	594, 310
	Canberra	**0.7836350**	903, 1
	Maximum	0.3146729	903, 1

4. Results and discussion

The Silhouette index, NbClust package and Elbow method were faithfully engaged in a series of experiments to determine the best possible number of clusters using the k-means, hierarchical, DBSCAN, EM and Fuzzy c-means algorithms. The experiments were repeated 100 times to estimate the average Silhouette values with the k-means, hierarchical clustering and Fuzzy c-means algorithms. The results are shown in Fig. 2a–c, Fig. 3a–c and Fig. 4a–f. Specifically, in Fig. 2a and b, the Elbow method for the k-means and hierarchical clustering suggests 3 clusters, while in Fig. 2c, the Fuzzy c-means algorithm indicates 2 clusters. From Fig. 3a–c, the Silhouette with the k-means, Hierarchical clustering and Fuzzy

Table 3
Fuzzy c-means performance with five distance functions

Distance function	Average computing time	Average iteration rate	Partition entropy	Partition coefficient	Fuzzy Silhouette index	Modified partition coefficient	Cluster structure
Euclidean	**3.0800**	**30.2100**	**0.6932**	**0.5000**	**0.3293**	**0.0000**	**447,457**
Manhattan	2.7300	31.1200	0.6932	0.5000	0.3217	0.0000	449,455
Chebyshev	2.0100	24.4900	0.6932	0.5000	0.1641	0.0000	547,357
Cosine	0.9600	9.7600	0.6932	0.5000	0.1831	0.0000	471,433
Minkowski	5.4800	45.0700	0.7335	0.3150	0.3300	0.3700	451,453

Table 4
Number of data objects obtained by using 5 clustering algorithms in two clusters

Clustering algorithm	Cluster 1	Cluster 2	Media deviation	Ranking
K-means	772	132	129.34	I
Hierarchical clustering	903	1	391.34	II
Fuzzy c-means	447	457	520.66	III
Expectation-maximization	*NaN*	*NaN*	*NaN*	*NaN*
DBSCAN	*NaN*	*NaN*	*NaN*	*NaN*
Average number of objects per cluster	707.33	196.67		

c-means suggests 2 clusters. In Fig. 4a–e, the NbClust package suggests 2 clusters, while in Fig. 4f, the NbClust package for the average hierarchical clustering using the Manhattan distance function suggests 8 clusters. Hence, using the principle of majority voting, 2 clusters have been taken as optimal.

Figures 5–7 respectively give the bar charts of the average Silhouette width, Dunn's index and Partition entropy against 22 distance functions. In Figs 5 and 6, the Pearson correlation has the highest average Silhouette width and Dunn's index respectively. In Fig. 7, the Dice has the smallest partition entropy (0.38) followed by the Pearson correlation (0.41). This result suggests the Pearson correlation as the best distance function for k-means clustering. Figure 8a and b shows the k-nearest neighbor distance graph for computing optimal epsilon for the DBSCAN. The plot of the corresponding cluster structure shows outliers, core and border objects respectively. From Fig. 8a, the knee for k-nearest neighbor distance plot was found at 90, which suggests the best value for epsilon, while the MinPts value is 6. In Fig. 8b, the DBSCAN cluster structure (804, 67 and 33) plot shows the overcrowding of outliers around the two clusters. The cluster structure shows 2 clusters and 804 noise points because of the capability of the DBSCAN algorithm to discover noise in a dataset [59], making it inappropriateness for the knowledge mining task considered.

The result of 3 internal validity metrics used to evaluate the best cluster structure for the k-means algorithm against 22 distance functions is presented in Table 1. The result of the CPCC for hierarchical clustering is presented in Table 2, where it is evident that a combination of Canberra distance with a centroid linkage function has the highest value of CPCC. This is the best possible parameter for achieving an ideal hierarchical clustering for the task of knowledge mining considered. In the experimental analysis of the Fuzzy c-means algorithm, 5 different distance functions were evaluated over 100 maximum iterations with 2 clusters. Results were collected after 100 times of repeating the algorithm for each distance function. The final experimental result is shown in Table 3, where it can be seen that the Cosine distance function has recorded the smallest average execution time of 0.96 s and average converging iteration rate of 9.76. The Minkowski distance function shows the largest average execution time of 5.4 s and average converging iteration rate of 45.07. This result indicates that engaging Fuzzy c-means algorithm with the Cosine distance function utilizes less computing time and converges faster when compared to others, irrespective of its performance. Nevertheless, the Euclidean distance function gives the best performance

for all the indices evaluated for the Fuzzy c-means algorithm when compared to other algorithms. This recommends Euclidean as a good distance function to apply in Fuzzy c-means algorithm for the task of knowledge mining considered in this study.

The overall result of ranking five data clustering algorithms based on the average number of data objects per cluster is shown in Table 4. The result shows that k-means algorithm has the lowest value of mean deviation, which suggests it with the Pearson correlation as the best ranked algorithm for the task of knowledge mining in a student engagement dataset.

5. Findings

The findings of this study have shown that a student engagement dataset could be clustered into two categories, which likely indicate positive and negative perceptions of students on engagement. The findings show that when benchmarking the k-means algorithm with different distance functions, some distance functions, majorly from the Shannon's entropy family [51] such as Kullback-Leibler, K-divergence, Topsoe, Jensen difference and Jensen-Shannon broke down when applied to the dataset of this study. This observed curb can be attributed to the fact that these functions are logarithmic in nature, which are mainly defined for positive and non-zero values. It is possible to overcome this inherent curb by transforming the functions to realize the shifted equivalent functions that would cater for negativity and singularity, which is not considered in this study. The Pearson correlation is the best choice of a distance function for k-means clustering (Table 1). This finding is contrary to what exists in [31], where Cosine stood out as the best distance function for the task of knowledge mining in manufacturing firms as a result of the nature of the dataset investigated. It was also found to be the fastest distance function in terms of convergence with the k-means and K-medoids algorithms when applied to a large dimensional continues dataset in [30].

It is observed from Table 2 that most of the cluster structures generated by the hierarchical clustering algorithm have unequal cluster sizes. This could be as a result of how the algorithm divides data into clusters. A cluster structure is majorly dependent on the chosen distance and linkage functions [31]. The average and centroid methods were found to give the best performance in terms of CPCC and generated unequal cluster sizes. This behavior was also observed in [56], using a different dataset. The Ward method on the contrary, gave almost equal sizes of cluster structures, but the performance was bad in terms of CPCC. The Ward performance could be as a result of outliers in the dataset of this study as highlighted under the review [56]. The Canberra distance function performed outstandingly with all the linkage functions investigated for the hierarchical clustering with a high CPCC close to 1, except with the Ward linkages (Table 2). The plot of DBSCAN cluster structure (Fig. 8b), reveals the occurrence of noise in the investigated dataset, which because of its sensitivity to noise makes it unable for the task of knowledge mining in a student engagement dataset. The Euclidean distance function was able to maximize the Silhouette index and minimize the modified partition coefficient when compared to other distance functions considered for Fuzzy c-means clustering algorithm (Table 3).

In the overall ranking, the k-means algorithm provides the best cluster, followed by the hierarchical clustering algorithm (Table 4). The Fuzzy c-means algorithm is the least performed clustering algorithm, while the EM algorithm could not give a stable result. The DBSCAN algorithm was found incongruous (Fig. 8), because of its high sensitivity to noise in the investigated dataset [61]. The unstable performance of the EM algorithm is suspected to be as a result of its capability to detect outliers. This intrinsic limitation could be improved by evaluating outliers in a student engagement dataset to identify those records that could possibly bias result before engaging a data clustering algorithm.

6. Conclusion

In general, focusing on the objective of this study, selecting the best clustering algorithm for the task of knowledge mining in a student engagement dataset is quite challenging because the performance of a clustering algorithm is majorly data dependent. The objective of this study was to experimentally evaluate the performances of five famous clustering algorithms to determine the best one for the task of knowledge mining in a student engagement dataset. Clusters have been evaluated using the Silhouette index, Nbluster package and Elbow method to find the best possible number of clusters in a student engagement dataset in accordance with what exists in literature. All the three methods have found two clusters in the dataset of this study. Three cluster validity indices of the Dunn, Silhouette and partition entropy have been fittingly applied to benchmark the k-means clustering algorithm with 22 different distance functions because of the unlabeled nature of the investigated dataset, while the Pearson correlation stood out as the best. The Cophenetic correlation coefficient was used to benchmark 8 linkage functions and different distance functions used with the hierarchical clustering algorithm in harmony with the previous works. In addition, for other algorithms, different indices have been implemented to get the best cluster from them.

The methodology of this study becomes a plausible device to identify factors impacting on student engagement and locate the aspects of student engagement practices and activities to be improved. This study has allowed to identify student engagement perceptions by social-demographic characteristics of students. In this way, we were able to identify valuable information that allows to recommend alternative student engagement practices and activities using the appropriate criteria that can influence policy change in undergraduate education. In future work, it is prudent to carry out an evaluation of outlier to identify records that could possibly bias results of knowledge mining in a student engagement dataset. In addition, practical application of the k-means clustering algorithm for the detection of useful patterns that could inform the building of a student engagement theory is highly desirable.

References

[1] B. Daniel, Big data and analytics in higher education: Opportunities and challenges, *British Journal of Educational Technology* **46**(5) (2015), 904–920.
[2] D.J. Shernoff, S. Kelly, S.M. Tonks, B. Anderson, R.F. Cavanagh, S. Sinha and B. Abdi, Student engagement as a function of environmental complexity in high school classrooms, *Learning and Instruction* **43** (2016), 52–60.
[3] K. Salmela-Aro, J. Moeller, B. Schneider, J. Spicer and J. Lavonen, Integrating the light and dark sides of student engagement using person-oriented and situation-specific approaches, *Learning and Instruction* **43** (2016), 61–70.
[4] E. Wagner and P. Ice, Data changes everything delivering on the promise of learning analytics in higher education, *EDUCAUSE Review* (2012), 33–42.
[5] J.A. Schmidt, J.M. Rosenberg and P.N. Beymer, A person-in-context approach to student engagement in science: Examining learning activities and choice, *Journal of Research in Science Teaching* **55**(1) (2018), 19–43.
[6] K. Krause and H. Coates, Students' engagement in first-year university, *Assessment and Evaluation in Higher Education* **33**(5) (2008), 493–505.
[7] G.D. Kuh, How to help students achieve, *Chronicle of Higher Education* **53**(41) (2007), B12–B13.
[8] C.S. Johnson and S. Delawsky, Project-based learning and student engagement, *Academic Research International* **4** (2013), 1–11.
[9] G.M. Elmore and E.S. Huebner, Adolescents' satisfaction with school experiences: relationships with demographics, attachment relationships, and school engagement behaviour, *Psychology in the Schools* **47**(6) (2010), 525–537.
[10] J.K. Sharma, Fundamental of business statistics, 2nd Edition, Vikas Publish House, PVT Ltd. India. 2014, 7–8.
[11] V. Trninić, I. Jelaska and J. Štalec, Appropriateness and limitations of factor analysis methods utilized in psychology and kinesiology: Part II., *Fizička Kultura* **67**(1) (2013), 1–17.
[12] P. Kang and S. Cho, K-means clustering seeds initialization based on centrality, sparsity, and isotropy, in: *International Conference on Intelligent Data Engineering and Automated Learning*, 2009, pp. 109–117.

[13] J.A. Fredricks and W. McColskey, The measurement of student engagement: A comparative analysis of various methods and student self-report instruments, in: S.L. Christenson et al. (eds.), *Handbook of Research on Student Engagement*, 2012, pp. 763–782.
[14] E. Skinner, Using community development theory to improve student engagement in online discussion: A case study, *ALT-J* **17**(2) (2009), 89–100.
[15] A.C.K. Wong, Understanding students' experiences in their own words: Moving beyond a basic analysis of student engagement, *The Canadian Journal of Higher Education* **45**(2) (2015), 60–80.
[16] A. Wigfield, J.T. Guthrie, K.C. Perencevich, A. Taboada, S.L. Klauda, A. McRae et al., Role of reading engagement in mediating the effects of reading comprehension instruction on reading outcomes, *Psychology in the Schools* **45** (2008), 432–445.
[17] P. Blumenfeld, J. Modell, W.T. Bartko, W.G. Secada, J.A. Fredricks, J. Friedel and A. Paris, School engagement of inner-city students during middle childhood, *Developmental Pathways Through Middle Childhood. Rethinking Contexts and Diversity as Resources* **27** (2005), 145–170.
[18] P. Witkowski and T. Cornell, An investigation into student engagement in higher education classrooms, *In Sight: A Journal of Scholarly Teaching* **10** (2015), 56–67.
[19] D.J. Shernoff and J.A. Schmidt, Further evidence of an engagement-achievement paradox among US high school students, *Journal of Youth and Adolescence* **37**(5) (2008), 564–580.
[20] S. Järvelä, M. Veermans and P. Leinonen, Investigating student engagement in computer-supported inquiry: A process-oriented analysis, *Social Psychology of Education* **11**(3) (2008), 299–322.
[21] R.L. Miller, R.F. Rycek and K. Fritson, The effects of high impact learning experiences on student engagement, *Procedia-Social and Behavioral Sciences* **15** (2011), 53–59.
[22] T. Petty and A.A. Farinde, Investigating student engagement in an online mathematics course through windows into teaching and learning, *Journal of Online Learning and Teaching* **9**(2) (2013), 261–270.
[23] J. Hamari, D.J. Shernoff, E. Rowe, B. Coller, J. Asbell-Clarke and T. Edwards, Challenging games help students learn: An empirical study on engagement, flow and immersion in game-based learning, *Computers in Human Behavior* **54** (2016), 170–179.
[24] K.C. Manwaring, R. Larsen, C.R. Graham, C.R. Henrie and L.R. Halverson, Investigating student engagement in blended learning settings using experience sampling and structural equation modeling, *The Internet and Higher Education* **35** (2017), 21–33.
[25] F.H. Veiga, Assessing student engagement in school: Development and validation of a four-dimensional scale, *Procedia-Social and Behavioral Sciences* **217** (2016), 813–819.
[26] P. Himmele and W. Himmele, Total participation techniques: Making every student an active learner, ASCD. 2017.
[27] South African Survey of Student Engagement (SASSE), Institutional Report. 2016.
[28] B. Schreiber and D. Yu, Exploring student engagement practices at a South African university: Student engagement as reliable predictor of academic performance, *South African Journal of Higher Education* **30**(5) (2016), 157–175.
[29] J. Han, M. Kamber and J. Pei, Data mining: concepts and techniques, Morgan Kaufmann. 2011.
[30] A.S. Shirkhorshidi, S. Aghabozorgi and T.Y. Wah, A comparison study on similarity and dissimilarity measures in clustering continuous data, *PloS One* **10**(12) (2015), e0144059.
[31] V. Bhatnagar, R. Majhi and P.R. Jena, Comparative performance evaluation of clustering algorithms for grouping manufacturing firms, *Arabian Journal for Science and Engineering* (2017), 1–13.
[32] B. Singla, K. Yadav and J. Singh, Comparison and analysis of clustering techniques, in: *Information Technology, ITSim 2008. International Symposium*, on. 3, 2008, pp. 1–3.
[33] J. Oyelade, I. Isewon, F. Oladipupo, O. Aromolaran, E. Uwoghiren, F. Ameh and E. Adebiyi, Clustering algorithms: Their application to gene expression data, *Bioinformatics and Biology Insights* **10** (2016), 237–253.
[34] V. Estivill-Castro, Why so many clustering algorithms: A position paper, *ACM SIGKDD Explorations Newsletter* **4**(1) (2002), 65–75.
[35] W.H. Au, K.C. Chan, A.K. Wong and Y. Wang, Attribute clustering for grouping, selection, and classification of gene expression data, *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **2**(2) (2005), 83–101.
[36] A.A. Abbasi and M. Younis, A survey on clustering algorithms for wireless sensor networks, *Computer and Communications* **30** (2007), 2826–2841.
[37] N. Esfandiari, M.R. Babavalian, A.M.E. Moghadam and V.K. Tabar, Knowledge discovery in medicine: Current issue and future trend, *Expert Systems with Applications* **35** (2014), 4434–4463.
[38] B. Zheng, S.W. Yoon and S.S. Lam, Breast cancer diagnosis based on feature extraction using a hybrid of k-means and support vector machine algorithms, *Expert Systems with Applications* **41** (2014), 1476–1482.
[39] M. Phanich, P. Pholkul and S. Phimoltares, Food recommendation system using clustering analysis for diabetic patients, in: *IEEE International Conference on Information Science and Applications (ICISA)*, 2010, pp. 1–8.
[40] O.J. Oyelade, O.O. Oladipupo and I.C. Obagbuwa, Application of k-means clustering algorithm for prediction of stu-

dents' academic performance, *International Journal of Computer Science and Information Security* **7**(1) (2010), 292–295.
[41] O.O. Olugbara, E. Adetiba and S.A. Oyewole, Pixel intensity clustering algorithm for multilevel image segmentation, *Mathematical Problems in Engineering* (2015), 19 pages.
[42] S. Arora, K. Singha and S. Sahney, Understanding consumer's showrooming behaviour: Extending the theory of planned behavior, *Asia Pacific Journal of Marketing and Logistics* **29**(2) (2017), 409–431.
[43] B. Civic and D. Cilimkovic, Characteristics of consumers' behavior in shopping of food products in the market of bosnia and herzegovina, *Research in World Economy* **8**(2) (2017), 49–58.
[44] B. Eriksson, P. Barford and R.D. Nowak, Network discovery from passive measurements, in: *Proceedings of the ACM SIGCOMM 2008 Conference on Applications, Technologies, Architectures and Protocols for Computer Communications*, Seattle, WA, USA, August 17–22. 2008, pp. 291–302.
[45] C. Pedro, D. Barbero, I. Martini and C. Discoli, Application of the k-means clustering method for the detection and analysis of areas of homogeneous residential electricity consumption at the Great La Plata region, Buenos Aires, Argentina, *Sustainable Cities and Society* **32** (2017), 115–129.
[46] B.R. Sharmaa and A. Paula, Clustering algorithms: Study and performance evaluation using Weka tool, *International Journal of Current Engineering and Technology* (2013), 1094–1094.
[47] D. Kabakchieva, Predicting student performance by using data mining methods for classification, *Cybernetics and Information Technologies* **13**(1) (2013), 61–72.
[48] X. Shao, H. Lee, Y. Liu and B. Shen, Automatic K selection method for the K-means algorithm, in: *Systems and Informatics (ICSAI), 4th International Conference on*, 2017, pp. 1573–1578.
[49] T. Thinsungnoena, N. Kaoungkub, P. Durongdumronchaib, K. Kerdprasopb and N. Kerdprasopb, The clustering validity with Silhouette and sum of squared errors, in: *Proceedings of the 3rd International Conference on Industrial Application Engineering*, 2015, pp. 44–51.
[50] M. Charrad, N. Ghazzali, V. Boiteau and A. Niknafs, NbClust: An R package for determining the relevant number of clusters in a data Set, *Journal of Statistical Software* **61**(6) (2014), 1–36.
[51] S.H. Cha, Comprehensive survey on distance/similarity measures between probability density functions, *International Journal of Mathematical Models and Methods in Applied Science* **4**(1) (2007), 300–307.
[52] J.C. Bezdek, Cluster validity with fuzzy sets, *J. Cybernet* **3** (1974), 58–72.
[53] P.J. Rousseeuw, Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, *Journal of Computational and Applied Mathematics* **2** (1987), 53–65.
[54] J.C. Dunn, Well separated clusters and optimal fuzzy partitions, *J. Cybernet* **4** (1974), 95–104.
[55] Z. Ansari, M.F. Azeem, W. Ahmed and A.V. Babu, Quantitative evaluation of performance and validity indices for clustering the web navigational sessions, *World of Computer Science and Information Technology Journal (WCSIT)* **1**(5) (2011), 217–226.
[56] S. Saraçli, N. Doğan and I. Doğan, Comparison of hierarchical cluster analysis methods by Cophenetic correlation, *Journal of Inequalities and Applications* **1** (2013), 203–210.
[57] NCSS, LLC. NCSS Statistical Software. NCSS.com [online] Chapter 445 Hierarchical clustering/Dendrograms. http://ncss.wpengine.netdna-cdn.com/wpcontent/themes/ncs.zs/pdf/Procedures/NCSS/Hierarchical_ClusteringDendrograms.pdf.
[58] S. Kurumalla and P.S. Rao, K-nearest neighbor based DBSCAN clustering algorithm for image segmentation, *Journal of Theoretical and Applied Information Technology* **92**(2) (2016), 395–402.
[59] W.H. Gui and H.N. Zhang, Asymptotic properties and expectation-maximization algorithm for maximum likelihood estimates of the parameters from Weibull-Logarithmic model, *Applied Mathematics-A Journal of Chinese Universities* **31**(4) (2016), 425–438.
[60] M.F. Saad and A.M. Alimi, Validity Index and number of clusters, *International Journal of Computer Science Issues (IJCSI)* **9**(1) (2012), 52–57.
[61] L. Duan, Density-based clustering and anomaly detection, in: *Business Intelligence-Solution for Business Development*, InTech, 2012.