Using K-Means to Determine Learner Typologies for Project-based Learning: A Case Study of the University of Education, Winneba

Delali Kwasi Dake
Department of ICT Education
University of Education, Winneba Ghana

Esther Gyimah
Department of ICT Education
University of Education, Winneba Ghana

ABSTRACT
Currently, academic instructors in Ghana have some difficulty in grouping students for projects-based courses because of increasing student numbers. One of the recent challenges educational institutions and instructors are facing is the explosive growth of educational data and how to use this data to improve the quality of teaching. K-means clustering is an unsupervised Data Mining technique for grouping large datasets with insightful similarity patterns to expose hidden trends and behavior in each cluster. The purpose of this research is to apply K-means clustering algorithm to analyze students' clusters for centered project-based learning. This research uses K clusters of 20. The clustering gave a low within cluster Sum of Square Error (SSE) of 3.60889. Clusters 1 and 6 have the highest member set of 32 each whiles clusters 8 and 9 have the lowest member set of 2. The results show that the K-means clustering algorithm is effective in grouping learners based on similar characteristics that indicate their performance. Assessments can also be tailored to suit all categories of learners for efficient results in project-based courses.

Keywords
K-means, Clustering, Educational Data Mining, Data Mining, Project-Based Learning.

1. INTRODUCTION
Data in higher education is increasing rapidly with less or no benefits to academic counsellors, students and management. Educational Data Mining (EDM) is a fast growing research area with advanced machine learning techniques to mine and better understand students learning behaviours and course re-design in academia [1][2]. The ideal motive of using data mining techniques in EDM is to expose hidden data patterns which serve as a predictive tool in Education.

One fast course re-design trend that academicians are pursuing is the concept of project-based learning by grouping students per study groups or for a project purpose. Project-based Learning (PBL) is either groups or individual learning where students' study efforts over a certain period of time and are assessed through active participation [3]. The idea behind this concept is to improve learner participation in group projects and help learners develop relevant skills [4]. However, [5] mention that within the same learning setting, students may exhibit different achievement levels, project-based learning when deployed, empowers learners to participate in interdisciplinary and collaborative activities.

Using K-Means Clustering algorithm is a better data mining technique for grouping students if improving learners' academic performance is of high relevance to academic authorities. This unsupervised algorithm can form the basis for grouping large data sets into number of clusters or groups. The relevance is to mine the educational data and develop grouping patterns with each training data set and a test data. These learner typologies will help academic counsellors as a reflective practice on which clusters to give more attention to because of peculiar properties.

This paper investigates EDM in higher education using a case study from data collected in the Department of ICT Education, University of Education, Winneba. The discovered knowledge will help provide academicians and University management with needed recommendations for improving project-based learning or learner grouping.

2. STATEMENT OF PROBLEM
Clustering is an unsupervised data mining technique used to identify groups with similar data characteristics [6]. Yet, in educational research, cluster analysis has been underutilized [7]. Learner typologies have been done in different formats by academic authorities but with little machine learning techniques to influence project-based course design and impact the academic performance of learners. This has led to a subjective way of grouping learners with no supervised learning mechanisms.

Effective learning designs are important for higher educational institutions [8] and with the emergence of big data and data mining, prediction and clustering will help re-design course projects based on useful information from training sets.

3. REVIEW OF LITERATURE
K-means clustering algorithm has been applied by various researchers to group learners based on learning styles and behaviors. [9] studied the coupling relationships of user attributes and propose a Coupled User-Clustering Algorithm (CUCA) for web-based learning using Taylor-like expansion to represent integrated coupling correlations of both intra-coupled and inter-coupled relationships. The Taylor-like expansion helps to cluster users using a spectral clustering algorithm which when applied in web-based learning systems can efficiently capture learners' behaviors and group them for personalized learning.

In another study, [10] presented a review paper on clustering in educational data mining. From their extensive review, they established that by using clustering techniques such as K-means Clustering Algorithm, it is easy to group students based on similar learning styles. They proposed working towards generating a unified clustering approach that can be easily applied to any educational data with less overhead delays.

In a related work, [11] proposed a clustering approach using K-means clustering algorithm to group students in courses that require that students' project assignments are done in

groups. The algorithm grouped the students based on topics that suited their similar interests and preferences and made use of the weighted formula that ensured that the maximum number of students allowed for a group was not exceeded. The results showed that 73% of the learners were satisfied with their groups because they received one of their top three priority choices for their preferred topic.

[12] in their research proposed a grouping method based on two clustering techniques: similarity student clustering and complementation student clustering to automatically group students in Foreign Language Learning (FLL) for effective group based learning.

They argued that the manual grouping method was not accurate because students were not grouped based on their knowledge levels and the method also failed to capture changes in students' knowledge during the learning process and adjust the groups automatically. Hence with the proposed clustering method, a student profile is built to model students' knowledge levels, which is updated automatically based on the examination results of each student. The proposed clustering methods and the manual grouping methods were then evaluated for learning effectiveness. The results showed that the proposed clustering methods enhanced the effectiveness of group-based learning.

In another related research, [13] implemented an Automated Group Decomposition Program (AGDP) tool that grouped students using the K-means clustering algorithm for effective collaborative learning. The tool can be used to group students into either heterogeneous groups, based on the students' knowledge of the subject; or homogeneous groups based on characteristics of communication skills, fluency in using computers and group work attitude or a mixed grouping of both heterogeneous and homogeneous characteristics. Then final heterogeneous groups are formed so that students can share the skills learnt in their groups with other students in their new groups thereby enhancing effective collaborative learning.

4. METHODOLOGY

Academic Instructors at the University of Education, Winneba (UEW) are focusing on project-based learning as a way to improve learner-centered education. Constructivists' approaches emphasize learners actively constructing their own knowledge rather than passively receiving information transmitted to them from teachers and text books. From a constructivist perspective, knowledge cannot simply be given to students: students must construct their own meanings [14].
The Department of ICT Education under the University started implementing project-based group learning from 2015 to Date. Key departmental courses: Database Management Systems, C++, Java, Computer Networking and Visual Literacy were used as prototypes. The selection criteria for grouping the students were based on incremental index numbers and subjective student's performance with no analytical statistics.

One of the Researchers of this study was the Instructor for the Database Management Systems second year class for the academic year 2016/2017 and the results after the project-based grouping were not satisfactory. With observation, the researcher realized that brilliant learners normally do the work for the group. The projects actually did little impact on improving the strengths of the group due to the generic nature of the tasks and the groupings.

4.1 Data Source

The dataset for this study is taken from the Database Management Systems continuous assessment of second year students for the 2017/2018 academic year with a population of 259 students in the Department of ICT Education of the University. The course is project-based and require students to be grouped to work on projects for the semester. The K-means Clustering Algorithm is used to group students into learner typologies based on their performance in all three assessment categories.

Table 1. Attribute Description of Dataset for Database Course Groupings

No	Attribute	Description	Values
1	Quiz	Entity – Relationship Diagram	<40
2	Test	Core Database Concept	<20
3	Practical Database Design Test	Hands on Database Practical Test	<30

These are attribute with instance records used as basis to group students. Each attribute has a specific purpose on the learners' ability to understand the Database subject and class groupings are based on the marks scored by the students in each instance.

4.1.1 Quiz

The quiz assessment was based on Entity-Relationship (ER) Diagram design. E-R diagram is part of the conceptual framework in database design. It relates to Entities and Attributes in a database design and the relationship between entity instances. The diagrammatic representation of cardinalities using crow's foot notation was added to the E-R design. Learners were introduced to SmartDraw and tested using it.

4.1.2 Test

The test assessment exposed the learner to core database concepts especially in data abstraction, data model, entity types, attribute types and normalization. These concepts form the basic rule in any Database Management System. Normalization for instance introduced the learner to database anomalies and functional dependencies. Normal Forms up to Third Normal Form (3NF) was explained to learners and later tested in design concepts.

4.1.3 Practical Database Design Test

The practical test was mainly in the area of Structured Query Language (SQL) and MySQL Database Management System was adopted for the course. Learners were tested practically on using SQL Data Definition Commands and Data Manipulation Commands. Evaluation was done based on practical Database creation, relationship between tables and manipulation of records.

4.2 K-Means Clustering Algorithm

Clustering is a data mining technique for grouping data into classes with identical characteristics and seeks to identify homogenous groups of data instances based on the values of their attributes [15][16].

Cluster analysis divides data into groups that are useful and share common characteristics. With clustering, data groupings will have a high intra-cluster similarity and low inter-cluster similarity [17]. K-Means is a partitioned clustering algorithm with unsupervised classification of patterns that determines all clusters at once with its derivative as K-medoid. K-means partitions n instances into k clusters where $k < n$ and each instance belongs to the cluster with the nearest mean. This algorithm aims at minimizing the squared error objective function.

K-means is similar to Gaussians expectation-maximization algorithm [18] in that both algorithms attempt to find cluster centroids in a data. This algorithm operates on data points into k clusters using Euclidean distance between the two points.

$$d(\vec{x}, \vec{y}) = \sqrt{\sum_{n=1}^{N}(x_n - y_n)^2} \quad (1)$$

$d(\vec{x}, \vec{y})$, denotes the Euclidean distance between the data point $\{x_n\}$ and the cluster centroid in vector space $\{y_n\}$.

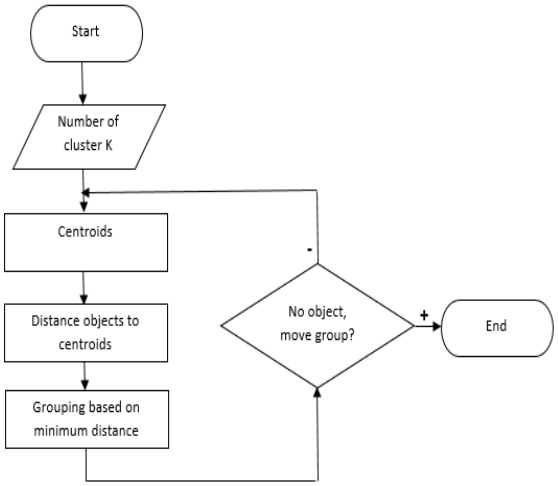


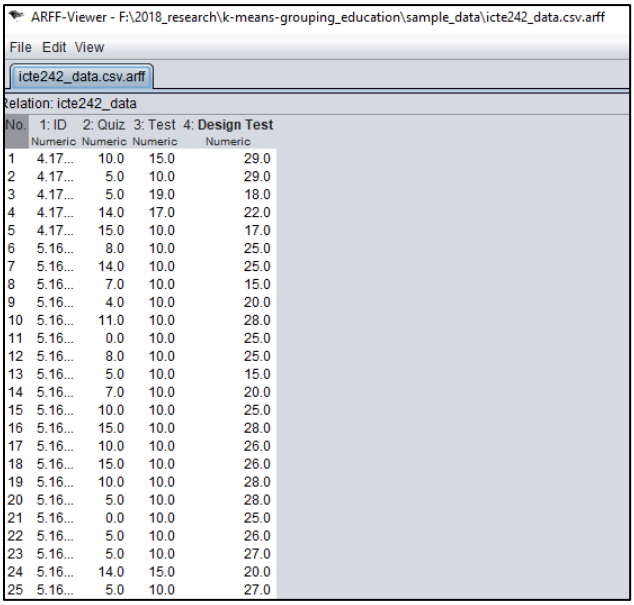
Fig 1: Logic behind the K-means Algorithm in creating homogenous groups from number of clusters

4.2.1 Algorithm Steps

- Specify *k*, the desired number of clusters
- Choose *k* points at random as cluster centers (centroid)
- Assign all instances to their closes cluster centers (Euclidean distance)
- Calculate the centroid, the mean of instances of each cluster
- These centroids are the new cluster enters
- Continue until the cluster centers don't change

5. RESULTS AND ANALYSIS

A total of 259 instances were taken to run the classifier in Weka environment. The Comma-Separated Values (CSV) in excel was converted to the Weka Attribute-Relation File Format (ARFF) using the ARFF-Viewer in Weka


ARFF-Viewer - F:\2018_research\k-means-grouping_education\sample_data\icte242_data.csv.arff

File Edit View

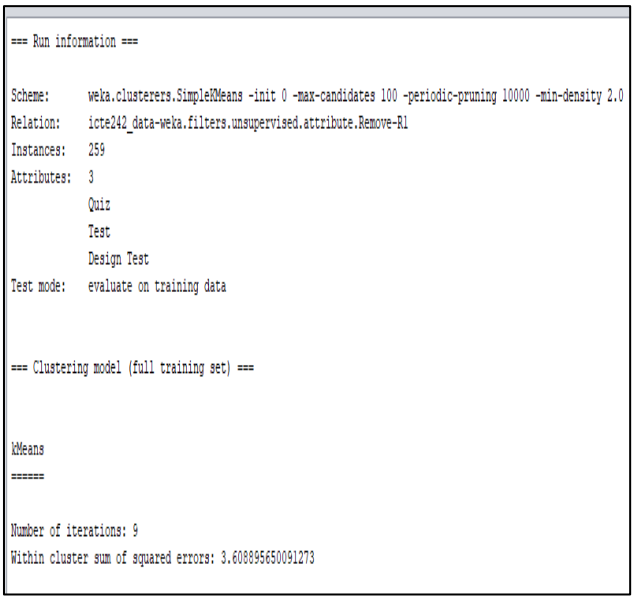
icte242_data.csv.arff

Relation: icte242_data

No.	1: ID	2: Quiz	3: Test	4: Design Test
	Numeric	Numeric	Numeric	Numeric
1	4.17...	10.0	15.0	29.0
2	4.17...	5.0	10.0	29.0
3	4.17...	5.0	19.0	18.0
4	4.17...	14.0	17.0	22.0
5	4.17...	15.0	10.0	17.0
6	5.16...	8.0	10.0	25.0
7	5.16...	14.0	10.0	25.0
8	5.16...	7.0	10.0	15.0
9	5.16...	4.0	10.0	20.0
10	5.16...	11.0	10.0	28.0
11	5.16...	0.0	10.0	25.0
12	5.16...	8.0	10.0	25.0
13	5.16...	5.0	10.0	15.0
14	5.16...	7.0	10.0	20.0
15	5.16...	10.0	10.0	25.0
16	5.16...	15.0	10.0	28.0
17	5.16...	10.0	10.0	26.0
18	5.16...	15.0	10.0	26.0
19	5.16...	10.0	10.0	28.0
20	5.16...	5.0	10.0	28.0
21	5.16...	0.0	10.0	25.0
22	5.16...	5.0	10.0	26.0
23	5.16...	5.0	10.0	27.0
24	5.16...	14.0	15.0	20.0
25	5.16...	5.0	10.0	27.0

Fig 2: Part of the dataset in Weka for Analysis

The dataset was subjected to K-Means clustering algorithm in Weka for the groupings of the Students into 20 clusters.


```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0
Relation:     icte242_data-weka.filters.unsupervised.attribute.Remove-R1
Instances:    259
Attributes:   3
              Quiz
              Test
              Design Test
Test mode:    evaluate on training data


=== Clustering model (full training set) ===


kMeans
======

Number of iterations: 9
Within cluster sum of squared errors: 3.608896650091273
```

Fig 3: K-means

In Figure 3, the cluster results were based on 3 attributes, Quiz, Test, and Design Test. The K-means algorithm iterated 9 times to give the 20 clusters for consideration using the Euclidean Distance. The K-means clustering gave a low sum of squared error (SSE) of 3.60889. The low SSE figure is an indication of a good cluster results.

Final cluster centroids:

		Cluster#									
Attribute	Full Data	0	1	2	3	4	5	6	7	8	9
	(259.0)	(4.0)	(32.0)	(6.0)	(14.0)	(17.0)	(12.0)	(32.0)	(21.0)	(2.0)	(2.0)
Quiz	10.3333	5.5	7.6563	17.1667	6.8571	13.8824	16.4167	9.0313	2.8095	16.5	10
Test	11.2054	18	10	15	10.2857	10	10	10.0625	10.1429	12.5	19.5
Design Test	23.6977	18.25	17.0313	19.3333	21.2143	27.7059	15.5833	25.9688	24.2381	17.5	16.5

10	11	12	13	14	15	16	17	18	19
(10.0)	(13.0)	(5.0)	(5.0)	(11.0)	(19.0)	(8.0)	(19.0)	(4.0)	(23.0)
18.2	11.7692	0	25	20.1818	5.7368	12.125	13.5439	29	5
14	15.4615	10	17	10.0909	10.1053	18.25	10.0634	10	10.0435
27.5	26.9231	27.6	14.8	26.7273	29	23.75	23.0367	25.5	27.3478

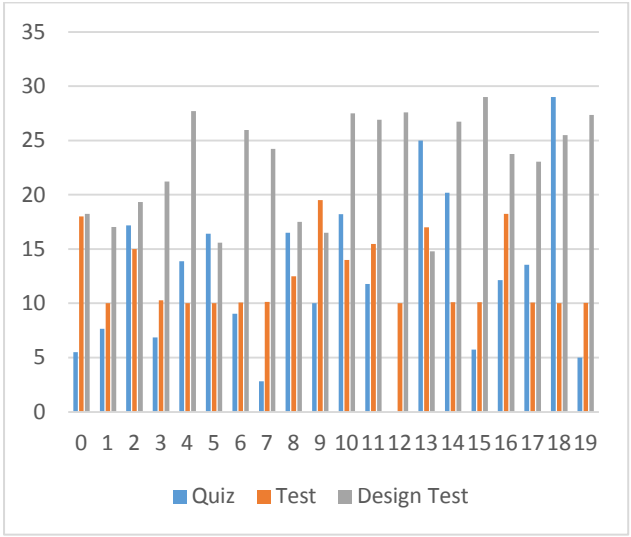


Fig 4: K-means member centroid

Figure 4 shows the centroids results calculated using the Euclidean Distance of K-means algorithm under the attributes Quiz, Test and Design Test. There were 20 clusters in general starting from 0 – 19. Each cluster has similar numeric characteristics and features.

Clustered Instances

0	4	(2%)
1	32	(12%)
2	6	(2%)
3	14	(5%)
4	17	(7%)
5	12	(5%)
6	32	(12%)
7	21	(8%)
8	2	(1%)
9	2	(1%)
10	10	(4%)
11	13	(5%)
12	5	(2%)
13	5	(2%)
14	11	(4%)
15	19	(7%)
16	8	(3%)
17	19	(7%)
18	4	(2%)
19	23	(9%)

Fig 5: Cluster member in percentages

Figure 5 above shows clustered instances percentage. Clusters 1 and 6 has the highest membership of 32 each represented as 24%. Clusters 8 and 9 has membership of 2 each with a 2% representation.

```
@relation icte242_data-weka.filters.unsupervised.attribute.Remove-R1_clustered

@attribute Instance_number numeric
@attribute 'Quiz ' numeric
@attribute Test numeric
@attribute 'Design Test' numeric
@attribute Cluster {cluster0,cluster1,cluster2,cluster3,cluster4,cluster5,cluster6,cluster7,cluster8,

@data
0,10,15,29,cluster11
1,5,10,29,cluster15
2,5,19,18,cluster0
3,14,17,22,cluster16
4,15,10,17,cluster5
5,8,10,25,cluster6
6,14,10,25,cluster17
7,7,10,15,cluster1
8,4,10,20,cluster3
9,11,10,28,cluster4
10,0,10,25,cluster7
11,8,10,25,cluster6
12,5,10,15,cluster1
13,7,10,20,cluster3
14,10,10,25,cluster6
15,15,10,28,cluster4
16,10,10,26,cluster6
17,15,10,26,cluster4
```

Fig 6: Sample Data Instances and Corresponding Clusters

Figure 6 above shows sample Data instances and corresponding cluster member numbers. The attributes instances were grouped under Quiz, Test, and Design Test.

From the results presented in figure 4, each of the 20 clusters represent students groupings based on the similarity of their scores and performance in three areas: Quiz, Test and Practical Database Design Test.

5.1 General Cluster Patterns

Cluster Analysis in Education especially for project-based learning from this research is extremely important if the concept of constructivism and personalized learning is of relevance to Instructors. Cluster member analysis presents a unique opportunity for Instructors to improve the academic performance of project-groups and re-design project contents for projects groups and even further, for individuals within the clusters. From the research conducted even without individual cluster analysis, students performed well generally in the Practical Database Design Test scoring $\frac{23.697}{30}$ as compared to scores in Quiz $\frac{10.33}{40}$ and Test $\frac{10.20}{20}$ as shown in Figure 4. This can be linked to SQL tutorial given to the learners even before the semester started. This research used K-means algorithm and limited the number of clusters to 20. Increasing the cluster groupings will increase the similarity in patterns of its members. This prototype is a representation of how Academic Instructors can design projects for large class members in revealing interesting pattern groups.

5.2 Interesting Cluster Patterns

5.2.1 Cluster 1

From Figure 5, Cluster 1 has one of the highest membership of 32 students but students in this cluster had a relatively Average scores in Quiz, Test and Practical Database Design Test. Scoring $\frac{7.6563}{40}$, $\frac{10}{20}$ and $\frac{17.0313}{30}$ respectively as represented in Figure 6. The project for this cluster should cover strongly all the three aspects of the Database course and the Instructor guidance of this cluster must be intensified.

5.2.2 Cluster 18

From Figure 5, Cluster 18 has only 4 members with excellent performances in Quiz and Practical Database Design Test scoring $\frac{29}{40}$, $\frac{25.5}{30}$ respectively but an average performance of $\frac{10}{20}$ in the Test. Cluster 18's project will be more focused on the Test score comprising of the core database concepts.

5.2.3 Cluster 15

This is a 19 member cluster with almost an excellent score of $\frac{29}{30}$ in Practical Database Design Test but with a poor score in the Quiz $\frac{5.7368}{40}$ and an average Test score of $\frac{10.1053}{20}$. This cluster involves students with good understanding of SQL but relatively low understanding in E-R Diagram and Core Database concepts as shown in the Quiz and Test scores From Figure 5.

5.2.4 Cluster 12

One interesting cluster is the 5 member cluster 12 scoring $\frac{0}{40}$ in the Quiz, average of $\frac{10}{20}$ in the Test but an excellent score of $\frac{27.6}{30}$ in the Practical Database Design Test. The scores indicate that the students understood well the creation of a database using SQL with relatively average understanding in Design Concept but demonstrated very poor understanding in E-R diagram.

The cluster results confirm [10] review which states that it is easy to group students based on similar learning styles. They proposed working towards generating a unified clustering approach in Education. [11] in their paper with 73% satisfaction from learners also stressed the importance of using K-means algorithms in grouping learners based on topics that suited their similar interests and preferences and made use of the weighted formula that ensured that the maximum number of students allowed for a group was not exceeded.

This research conducted references most of the literature reviewed for this study of the importance of using clustering algorithm, especially K-means in grouping learners for either homogenous or heterogeneous projects if improving learners academic score is of much relevance to Instructors.

6. CONCLUSION AND FUTURE WORK

In this paper, the K-means clustering algorithm is used to determine similar learner typologies among students in a Database Management System class at the University of Education, Winneba. The researchers limited the number of clusters in this research to 20 and had a low sum of squared error (SSE) of 3.60889 which is an indication of a good cluster results after 9 iterations. The results of this research helped in grouping learners according to their strengths and areas of improvement in the Database Course. The results show that the K-means unsupervised learning algorithm was effective in grouping learners with similar concept score in the Database course.

This research modeled a classifier using training dataset from 2017/2018 Database Course from the University of Education, Winneba. With this clustering, class labels are now generated for each cluster member for the Database Course, making predictive classification possible for Test Data. A future research will be in the area of cluster member predictions using classification techniques.

7. REFERENCES

[1] Baradwaj, B. K., & Pal, S. (2012). Mining educational data to analyze students' performance. *arXiv preprint arXiv:1201.3417*.

[2] Romero, C., & Ventura, S. (2010). Educational data mining: a review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *40*(6), 601-618.

[3] Bilgin, I., Karakuyu, Y., & Ay, Y. (2015). The effects of project based learning on undergraduate students' achievement and self-efficacy beliefs towards science teaching. *Eurasia Journal of Mathematics, Science & Technology Education*, *11*(3), 469-477.

[4] Perera, D., Kay, J., Koprinska, I., Yacef, K., & Zaïane, O. R. (2009). Clustering and sequential pattern mining of online collaborative learning data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(6), 759-772.

[5] Han, S., Capraro, R., & Capraro, M. M. (2015). How science, technology, engineering, and mathematics (STEM) project-based learning (PBL) affects high, middle, and low achievers differently: The impact of student factors on achievement. *International Journal of Science and Mathematics Education*, *13*(5), 1089-1113

[6] Shovon, M., Islam, H., & Haque, M. (2012). An Approach of Improving Students Academic Performance by using k means clustering algorithm and Decision tree. *arXiv preprint arXiv:1211.6340*.

[7] Bahr, P. R. (2010). The bird's eye view of community colleges: A behavioral typology of first-time students based on cluster analytic classification. *Research in Higher Education*, *51*(8), 724-749.

[8] Paquette, G., Léonard, M., Lundgren-Cayrol, K., Mihaila, S., & Gareau, D. (2006). Learning design based

on graphical knowledge-modelling. *Educational Technology & Society*, *9*(1), 97-112

[9] Niu, K., Niu, Z., Zhao, X., Wang, C., Kang, K., & Ye, M. (2016). A Coupled User Clustering Algorithm for Web-based Learning Systems. In *EDM* (pp. 175-182).

[10] Dutt, A., Aghabozrgi, S., Ismail, M. A. B., & Mahroeian, H. (2015). Clustering algorithms applied in educational data mining. *International Journal of Information and Electronics Engineering*, *5*(2), 112.

[11] Akbar, S., Gehringer, E. F., & Hu, Z. (2018). Improving formation of student teams: a clustering approach. In *Proceedings of the 40th International Conference on Software Engineering: Companion Proceeedings* (pp. 147-148). ACM.

[12] Li, L., Luo, X., & Chen, H. (2015). Clustering Students for Group-Based Learning in Foreign Language Learning. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, *9*(2), 55-72

[13] Sarkar, A., Seth, D., Basu, K., & Acharya, A. (2015). A new approach to collaborative group formation. *International journal of computer applications*, *128*(3).

[14] Stage, F. K., Muller, P. A., Kinzie, J., & Simmons, A. (1998). Creating learning centered classrooms: What does learning theory have to say f Washington. *DC: The George Washington University, Graduate School of Education and Human Development and Association for the Study of Higher Education*.

[15] Ester, M., Frommelt, A., Kriegel, H. P., & Sander, J. (1998). Algorithms for characterization and trend detection in spatial databases. *In Proc. 4th Int. Conf. on Knowledge Discovery and Data Mining* (pp. 44-50). New York City, NY

[16] Kaufman, L and Rousseeuw, J. P. (2005). Finding Groups in Data: An Introduction to Cluster Analysis (Wiley Series in Probability and Statistics). Wiley-Interscience

[17] MacQueen, J. B. (1967). Some Methods for classification and Analysis of Multivariate Observations. *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability* (pp. 281–297). University of California Press.

[18] Jian, B., & Vemuri, B.C. (2005). A robust algorithm for point set registration using mixture of Gaussians. Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1, 2, 1246-1251 Vol. 2.