Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.778023

PREDICTIVE CLUSTERING OF STUDENT DATA USING THE K-MEANS CLUSTERING ALGORITHM: A COMPREHENSIVE ANALYSIS

Gautam Appasaheb Kudale and Dr. Sandeep Singh Rajpoot
Department of Computer Application,
Dr. A. P. J. Abdul Kalam University, Indore (M.P.), India 452010
Corresponding Author Email : gaukudale@gmail.com

Abstract:
Academic performance prediction is a challenging task in today's education system, and its analysis plays a crucial role in enhancing the quality of education and supporting decision-making. Evaluating students' performance is of utmost importance for educational institutions as it enables academic leaders to make informed decisions using the vast amount of available data and various algorithms. Clustering, a process of grouping objects based on their similarities is employed in this research to analyze students' performance. Specifically, the K-Means clustering algorithm is utilized to cluster students based on their academic characteristics. This research paper explores the application of data clustering, focusing on the K-Means algorithm, to evaluate students' performance. The findings of this study contribute to the understanding of student performance analysis and provide valuable insights for academic leaders to make data-driven decisions.

Keywords: Clustering, K-Means Clustering, Students' Academic Performance

1. INTRODUCTION

Clustering analysis holds significant importance in the field of education and student data analysis.Clustering techniques in student performance analysis offer numerous benefits to educators and academic institutions. By identifying clusters of students with similar characteristics, such as learning styles, preferences, or aptitudes, educators can design personalized learning plans and provide targeted resources and activities to enhance student engagement and achievement. Clustering algorithms also enable early identification of at-risk students, helping institutions intervene with timely support and interventions to improve outcomes and increase retention rates. Moreover, clustering analysis facilitates the formation of diverse and balanced student groups for collaborative learning, fostering peer-to-peer knowledge sharing and collaboration. It also aids in curriculum planning and resource allocation by identifying clusters of students with similar learning needs or subject preferences, allowing institutions to optimize their course offerings and develop targeted interventions. Finally, clustering analysis provides valuable insights for academic leaders and policymakers, supporting data-driven decision-making and enhancing the overall quality of education and institutional effectiveness. It empowers educational institutions to leverage the available data and improve the educational experience and outcomes for all students.

The K-Means clustering algorithm can be highly relevant in predicting student clusters by grouping students based on their similarities and patterns. It is a popular unsupervised machine learning algorithm that aims to partition a dataset into distinct clusters. The algorithm assigns data points to clusters in a way that minimizes the variance within each cluster while maximizing the variance between clusters. In the context of predicting student clusters, K-

Means can be applied to various educational datasets, such as student performance records, demographic information, or engagement metrics.

It is important to note that while K-Means clustering can provide valuable insights and facilitate decision-making in education, it should not be seen as a definitive solution. Clustering algorithms are sensitive to input data and assumptions made during the analysis. Therefore, it is crucial to consider other factors, such as expert knowledge, domain expertise, and qualitative information, to validate and interpret the results effectively. The objective of this research is to explore the application of the K-means clustering algorithm for evaluating students' academic performance, to assess the effectiveness of data clustering in analyzing students' performance and identifying meaningful clusters.

2. LITERATURE REVIEW

Clustering techniques have been widely employed in student data analysis to gain insights into student behavior, performance patterns, and learning preferences. Several studies have explored the application of clustering algorithms in this context. Here is a brief review of existing literature on clustering techniques and their applications in student data analysis:

1. Huang, J., & Li, C. (2018). Cluster analysis of e-learning students' data based on K-means algorithm. Proceedings of the International Conference on Information Science and Education, 1-6. This research paper presents a study on applying the K-Means algorithm to cluster e-learning students based on their learning behaviors. It analyzes the clusters to understand student engagement and performance, aiming to improve personalized learning support.
2. Garg, S., & Sharma, M. (2019). Clustering Algorithms for Educational Data Mining: A Comprehensive Review. International Journal of Intelligent Systems and Applications in Engineering, 7(6), 8-19. This review paper provides an overview of various clustering algorithms used in educational data mining, including K-Means, DBSCAN, and hierarchical clustering. It discusses their applications in student performance analysis, course recommendation systems, and identifying student profiles.
3. Wang, S., Chen, H., Zhang, Y., & Wang, X. (2019). An adaptive learning recommendation approach based on K-means clustering. IEEE Access, 7, 28267-28276. The study utilized the K-Means algorithm to cluster students based on their learning preferences and behaviors. The clustering results were used to generate personalized learning recommendations for individual students.
4. Elsalamouny, E. H., AbdelSalam, M. A., & El-Gayar, E. (2021). Clustering-based recommender system for enhancing student performance in e-learning environments. Journal of King Saud University-Computer and Information Sciences, 33(2), 183-192. The study presents a clustering-based recommender system that uses the K-Means algorithm to cluster students based on their learning behaviors and preferences. The system provides personalized recommendations to enhance student performance in e-learning environments.
5. Zheng, Y., & Zhao, W. X. (2021). Understanding student performance: A clustering-based approach. IEEE Access, 9, 40110-40122. This paper proposes a clustering-based approach using the K-Means algorithm to understand student performance based on

their engagement, performance, and demographic data. The clustering results help in identifying student groups with different performance patterns and characteristics.

6. Pereira, C., & Costa, P. J. (2021). Clustering higher education students' academic trajectories: An approach based on K-means algorithm. Expert Systems with Applications, 182, 115055. The study utilizes the K-Means algorithm to cluster higher education students based on their academic trajectories. The clustering approach helps in identifying distinct groups of students with similar patterns of course enrollment and academic performance.
7. Masrom, M. A., &Husin, M. R. (2021). Students clustering using k-means algorithm for e-learning recommendation. International Journal of Recent Technology and Engineering, 10(2), 3250-3255. This paper presents a student clustering approach based on the K-Means algorithm for generating personalized e-learning recommendations. The clustering results help in identifying student groups with similar learning preferences, enabling tailored recommendations.

Clustering techniques, particularly the K-Means algorithm, have found diverse applications in student data analysis within the field of education. These studies emphasize the use of clustering algorithms such as K-Means, DBSCAN, and hierarchical clustering to identify student profiles, behavior patterns, and predict performance. They showcase the algorithm's effectiveness in grouping students based on their characteristics and behaviors, enabling personalized interventions, recommendation systems, and adaptive learning environments. The research highlights the ongoing relevance and application of K-Means clustering in education, emphasizing its utility in personalization, recommendation systems, and understanding student performance.

3. THE K-MEANS CLUSTERING ALGORITHM

FLOW CHART OF K MEANS CLUSTERING ALGORITHM

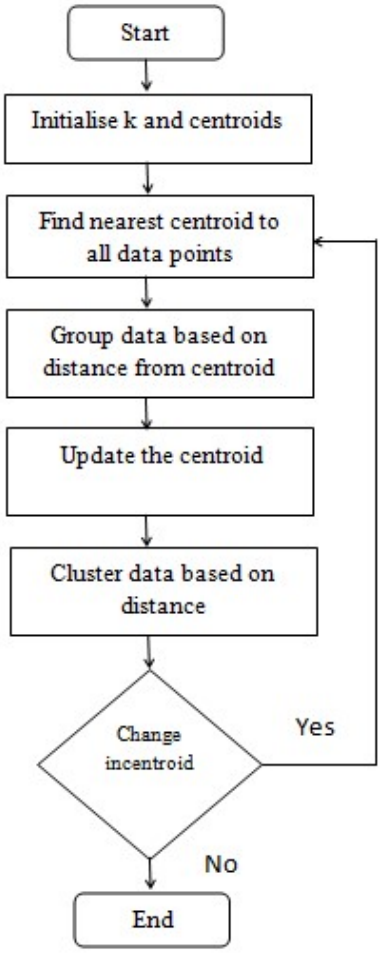


Figure 1: Generalised Pseudocode of Traditional k-means [5,8,9,18,25,27]

GENERALISED PSEUDOCODE OF TRADITIONAL K-MEANS

Following figure shows traditional generalized Pseudocode for k-means

Step 1: Accept the number of clusters to group data into and the dataset to cluster as input values

Step 2: Initialize the first K clusters

- Take first k instances or

- Take Random sampling of k elements

Step 3: Calculate the arithmetic means of each cluster formed in the dataset.

Step 4: K-means assigns each record in the dataset to only one of the initial clusters - Each record

is assigned to the nearest cluster using a measure of distance (e.g Euclidean distance).

Step 5: K-means re-assigns each record in the dataset to the most similar cluster and re-calculates

The arithmetic mean of all the clusters in the dataset.

4. EXPERIMENTAL ANALYSIS AND DISCUSSION

From the experimental point of view, the dataset is created by importing basic data set of students from Kaggle. Data set name students-performance which contains the attributes "gender", "race/ethnicity", "parental level of education", "lunch", "test preparation course", "math score", "reading score", "writing score" with one thousand records in it. All the above information will be consolidated as a whole form into complete dataset for the proposed methodology.

In this research, all the pre-processing on the data is done by using different libraries from Python such as Pyspark etc.Student academic performance is predicted based on multiple input attributes. Algorithms such as, K-means are used on the input attributes to generate a classification model in-order to predict academic performance of students.

After applying the K means algorithm, for Elbow method, we got the value of K and its corresponding cost. We have plotted Elbow graph which is used to predict the students' performance. Again, after applying the same K means algorithm, we got the value of K and its corresponding Silhouette score, with which we could plot Silhouette graph.

K-means algorithm implementation

The k-means algorithm is implemented as i) for all attributes in dataset and ii)for last 4 columns in dataset; following steps are implemented:

1. Vector assembler is used to assemble the data in to single columnvector which yielded thedf_features.
2. Then StandardScaler is used it creates another column i.e., df_standardized are generated. StandardScaler removes the mean and scales each feature/variable to unit variance.
3. PCA features are extracted using in new column i.e., pcaFeatures are generated.

Vector assembler: -The Vector Assembler is a feature transformation tool commonly used in machine learning and data processing pipelines. It is typically applied in the context of feature engineering, where it combines multiple input columns or features into a single vector column. This process is useful when you want to consolidate multiple features and treat them as a single feature vector, which is often required by machine learning algorithms.

Standard Scaler: -The StandardScaler is a popular feature transformation technique used in machine learning to standardize or normalize numerical features. It rescales the features so that they have zero mean and unit variance. This transformation is often performed as a preprocessing step before applying machine learning algorithms that require standardized input.

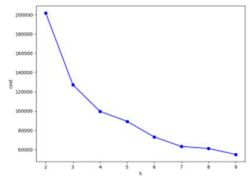
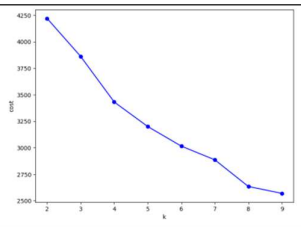
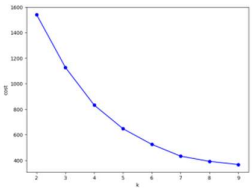
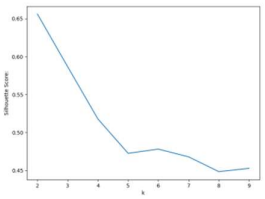
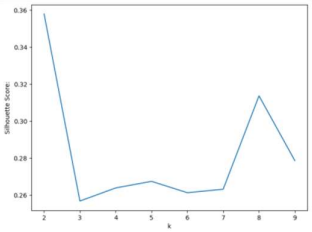
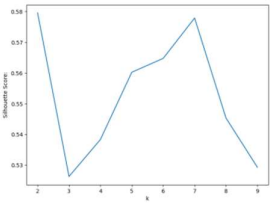
700 Records	A) Vector Assembler	B) StandardScaler	C) PCAFeatures
Elbow Plot			
Silhouette Plot			

FIG 2: ELBOW PLOT AND SILHOUETTE SCORE PLOT FOR ALL COLUMNS

Principal Component Analysis: - PCA is a popular unsupervised learning technique for reducing the dimensionality of data. It increases interpretability yet, at the same time, it minimizes information loss. It helps to find the most significant features in a dataset and makes the data easy for plotting in 2D and 3D.

Above steps are used to create input and the elbow plot and silhouette score plot id plotted to identify proper clustering.

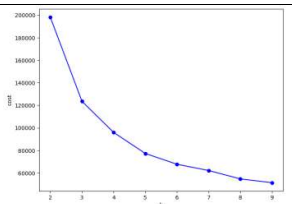
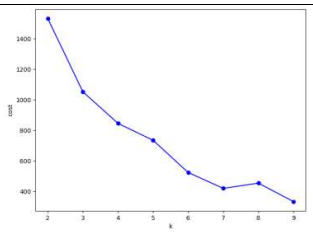
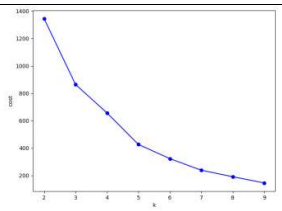
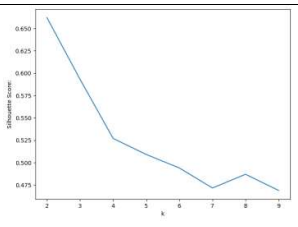
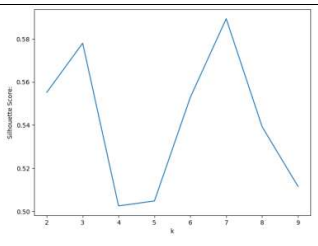
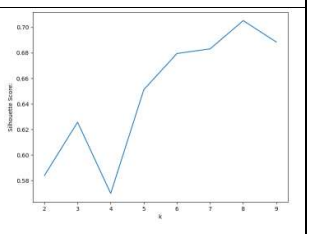
700 Records	A) Vector Assembler	B) Standard Scaler	C) PCAFeatures
Elbow Plot			
Silhouette Plot			

FIG 3. ELBOW PLOT AND SILHOUETTE SCORE PLOT FOR LAST FOUR COLUMNS

The Elbow Method: This is one of the most popular methods to determine the optimal number of clusters. It is a little bit simple approach. In this method, we calculate the cost which consists of sum of squared distances of points to their nearest centres. The drawback of Elbow method is that sometimes we cannot get the optimal value of k. We can get the ambiguous value of k. In such ambiguous situation, we have to use the Silhoutte method.

The Silhoutte method: This method estimates a value which shows how a point is closer to its own cluster as compared to other clusters. The value of silhouette coefficient is between -1 to 1 [4].

One cannot bypass the Elbow method and consider only The Silhoutte one. The Elbow method is used to get a rough estimate of k whereas Silhoute value method is used to get the exact value of k. Both the methods conjunctively form a tool for us to take confident decision for the determination of value of k.

In the above plots i.e. for all columns in Fig. 2A and 2B i.e. Graph of Vector Assemblerand Standard Scaler, we see elbow plot in which we have to see the first or most significant turning point of the curve which is visible as an elbow which suggest the right number of clusters [4]. In this case, it can be 4 or 5. The Silhouette score reaches its global maximum at the optimal k. This should ideally appear as a peak in the silhouette values versus-k plot. But there is no clarity with the Silhoutte plot. There is no a clear maximum or minima visible. In the plots Fig. 2C i.e. Graph of PCAFeatures in elbow plot it can be 4 or 5 and with the Silhoutte plot it is 5 and as it is present in list of elbow point, so we can select 5 as a number of clusters for these.

In the above plots i.e. for all last four columns in Fig. 3A and 3B i.e. Graph of Vector Assembler and Standard Scaler, we see elbow plot in which we have to see the first or most significant turning point of the curve which is visible as an elbow which suggest the right number of clusters [4]. In this case, it can be 4 or 5. The Silhouette score reaches its global maximum at the optimal k. This should ideally appear as a peak in the silhouette values versus-k plot. But there is no clarity with the Silhoutte plot. There is no a clear maximum or minima visible. In the plots Fig. 3C i.e. Graph of PCAFeatures in elbow plot it can be 4, 5 or 6 and with the Silhoutte plot it is 5 and as it is present in list of elbow point, so we can select 5 as a number of clusters for these.

5. CONCLUSION

In this paper we have taken student dataset of 700 records from Kaggle, we have applied K means algorithm on the dataset. Then vector assembler is used to assemble the data in to single (column) vector i.e., df_features are generated. Then StandardScaler is used it creates another column i.e., df_standardized are generated. PCA features are extracted using in new column i.e., PCAFeatures are generated. The df_features, df_standardized and pcaFeatures are used to create input and the elbow plot and silhouette score plot id are plotted to identify proper clustering. It is observed that for all attributes as well as for last four columns also the PCA features results looks good. In both case we get five clusters i.e. K is equal to five. This K is useful in future work for predicting academic performance of students.

REFERENCES

[1] Data Mining Introductory and Advanced Topics, Margaret H. Dunhan, Pearson

[2] Data Mining Practical Machine Learning Tools and Techniques, 3rd Edition, Ian H.witten, Eibe Frank, Mark A. Hall

[3] Butkar Uamakant, "A Formation of Cloud Data Sharing With Integrity and User Revocation", International Journal Of Engineering And Computer Science, Vol 6, Issue 5, 2017

[4] Data Mining, Concepts and Techniques, 3rd Edition, Jiawei Han, Micheline Kamber, Jian Pei

[5] Prof. Prashant Sahai Saxena, Prof. M. C. Govil, "Prediction of Student's Academic Performance using Clustering," Special Conference Issue: National Conference on Cloud Computing & Big Data

[6] Bindiya M Varghese, Jose Tomy J, Unnikrishnan A, Poulose Jacob K, "Clustering student data to characterize performance patterns," (IJACSA) International Journal of Advanced Computer Science and Applications, Special Issue on Artificial Intelligence,

[7] Umakant Dinkar Butkar, Dr. Nisarg Gandhewar. (2022). ALGORITHM DESIGN FOR ACCIDENT DETECTION USING THE INTERNET OF THINGS AND GPS MODULE. Journal of East China University of Science and Technology, 65(3), 821–831. Retrieved from http://hdlgdxxb.info/index.php/JE_CUST/article/view/313

[8] Oyelade, O. J, Oladipupo, O. O., Obagbuwa, I. C., "Application of k-Means Clustering algorithm for prediction of Students' Academic Performance," (IJCSIS) International Journal of Computer Science and Information Security, Vol. 7, o. 1, 2010

[9] Rakesh Kumar Arora, Dr. Dharmendra Badal, "Evaluating Student's Performance Using k-Means Clustering," International Journal of Computer Science And Technology, IJCST Vol. 4, Issue 2, April - June 2013, ISSN : 0976-8491 (Online) | ISSN : 2229-4333 (Print)

[10] Sharmila, R.C Mishra, "Performance Evaluation of Clustering Algorithms," International Journal of Engineering Trends and Technology (IJETT) - Volume4 Issue7- July 2013, ISSN: 2231-5381

[11] Ramjeet Singh Yadav, P. Ahmed, A. K. Soni and Saurabh Pal, "Academic performance evaluation using soft computing techniques," CURRENT SCIENCE, VOL. 106, NO. 11, 10 JUNE 2014

[12] Harwatia, Ardita Permata Alfiania, Febriana Ayu Wulandaria, "Mapping student's performance based on data mining approach (a case study)," The 2014 International Conference on Agro-industry (ICoA): Competitive and sustainable Agro industry for Human Welfare, Agriculture and Agricultural Science Procedia 3 (2015) 173 – 177

[13] Patel, J. and Yadav, R.S. (2015) "Applications of Clustering Algorithms in Academic Performance Evaluation." Open Access Library Journal, 2: August 2015 | Volume 2 | e1623

[14] Jyotirmay Patel, Ramjeet Singh Yadav, "Applications of clustering algorithms in academic performance evaluation"

[15] Atul Prakash Prajapati, Sanjeev Kr. Sharma, Manish Kr. Sharma, "Student's performance analysis using machine learning tools," International Journal of Scientific & Engineering Research Volume 8, Issue 10, October-2017 ISSN 2229-5518

[16] E.Venkatesan, S.Selvaragini, "Prediction of students academic performance using classification and clustering algorithms," International Journal of Pure and Applied Mathematics Volume 116 No. 16 2017, 327-333 ISSN: 1311-8080 (printed version); ISSN: 1314-3395 (on-line version)

[17] Snehal Bhogan , Kedar Sawant , Purva Naik , Rubana Shaikh , Odelia Diukar , Saylee Dessai, "Predicting student performance based on clustering and classification," IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661,p-ISSN:2278-8727, Volume 19, Issue 3, Ver. V (May-June 2017), PP 49-52

[18] Mr. Shashikant Pradip Borgavakar, Mr. Amit Shrivastava, "Evaluating student's performance using k-means clustering," International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, Vol. 6 Issue 05, May – 2017

[19] Mrs .Mary vidya john, Akshata police patil, Anjali mishra, Bindhu reddy G, Jamuna N, "Clustering technique for student performance," International Research Journal of Computer Science (IRJCS), Issue 06, Volume 6 (June 2019), ISSN: 2393-9842

[20] Noel Varela , Edgardo Sánchez Montero , Carmen Vásquez , Jesús García Guiliany , Carlos Vargas Mercado , Nataly Orellano Llinas , Karina Batista Zea , and Pablo Palencia, "Student performance assessment using clustering techniques," © Springer Nature Singapore Pte Ltd. 2019 Y. Tan and Y. Shi (Eds.): DMBD 2019, CCIS 1071, pp. 179–188, 2019. https://doi.org/10.1007/978-981-32-9563-6_19

[21] N.Valarmathy, S.Krishnaveni, "Performance evaluation and comparison of clustering algorithms used in educational data mining," International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-7, Issue-6S5, April 2019

[22] Lubna Mahmoud Abu Zohair, "Prediction of Student's performance by modelling small dataset size," Abu Zohair International Journal of Educational Technology in Higher Education (2019) 16:27 https://doi.org/10.1186/s41239-019-0160-3

[23] Mrs. Bhawna Janghel, Dr. Asha Ambhaikar, "Performance of student academics by k-mean clustering algorithm," International J. Technology. January – June, 2020; Vol. 10: Issue 1, ISSN 2231-3907 (Print), ISSN 2231-3915 (Online)

[24] Marzieh Babaie, Mahdi Shevidi Noushabadi, "A review of the methods of predicting students' performance using machine learning algorithms," Archives of Pharmacy Practice ¦ Volume 11 ¦ Issue S1 ¦ January-March 2020

[25] Dr. G. Rajitha Devi, "Prediction of student academic performance using clustering," International Journal of Current Research in Multidisciplinary (IJCRM) ISSN: 2456-0979 Vol. 5, No. 6, (June'20), pp. 01-05

[26] Dewi Ayu Nur Wulandari; Riski Annisa; Lestari Yusuf, Titin Prihatin, "An educational data mining for student academic prediction using k-means clustering and naïve bayes classifier," journal Pilar Nusa Mandiri Vol 16, No 2 September 2020

[27] Yann Ling Goh, Yeh Huann Goh, Chun-Chieh Yip, Chen Hunt Ting, Raymond Ling Leh Bin, Kah Pin Chen, "Prediction of students' academic performance by k-means clustering," Peer-review under responsibility of 4th Asia International Multidisciplinary Conference 2020 Scientific Committee

[28] Revathi Vankayalapati, Kalyani Balaso Ghutugade, Rekha Vannapuram, Bejjanki Pooja Sree Prasanna, "K-means algorithm for clustering of learners performance levels using machine learning techniques," Revue d'Intelligence Artificielle Vol. 35, No. 1, February, 2021, pp. 99-104

[29] Rina Harimurti, Ekohariadi, Munoto, I. G. P Asto Buditjahjanto, "Integrating k-means clustering into automatic programming assessment tool for student performance analysis," Indonesian Journal of Electrical Engineering and Computer Science Vol. 22, No. 3, June 2021, pp. 1389~1395 ISSN: 2502-4752, DOI: 10.11591/ijeecs.v22.i3.pp1389-1395

[30] Rui Shang , Balqees Ara, Islam Zada, Shah Nazir , Zaid Ullah, and Shafi Ullah Khan, "Analysis of simple k-mean and parallel k-mean clustering for software products and organizational performance using education sector dataset," Hindawi Scientific Programming Volume 2021, Article ID 9988318, 20 pages https://doi.org/10.1155/2021/9988318

[31] Bao Chong, "K-means clustering algorithm: a brief review," Academic Journal of Computing & Information Science ISSN 2616-5775 Vol. 4, Issue 5: 37-40, DOI: 10.25236/AJCIS.2021.040506

[32] Said Abubakar Sheikh Ahmed, "Evaluating students' performance of social work department using k-means and two-step cluster "a case study of mogadishu university"," Mogadishu University Journal, Issue 7, 2021, ISSN 2519-9781

[33] Zhihui Wang, "Higher education management and student achievement assessment method based on clustering algorithm," Hindawi Computational Intelligence and Neuroscience Volume 2022, Article ID 4703975, 10 pages https://doi.org/10.1155/2022/4703975

[34] Ahmad Fikri Mohamed Nafuri , Nor Samsiah Sani, Nur Fatin Aqilah Zainudin , Abdul Hadi Abd Rahman and Mohd Aliff, "Clustering analysis for classifying student academic performance in higher education," Appl. Sci. 2022, 12, 9467. https://doi.org/10.3390/app12199467