The Application of K-Means Clustering to Analyze and Enhance Academic Performance

Anita, Naveen Kumar, Mridula Yadav, Karishma, Yudhishthir Yadav, Reena Sharma

Assistant Professor IPEM College, Ghaziabad, U.P India
Email: anita.singh@ipemgzb.ac.in

The rapidly expanding educational database in higher education institutions is one of the important realities. These databases are growing quickly with no advantage to database management. These days, clustering techniques are widely used and important, and their significance only seems to rise with the volume of data.The K-means clustering technique is used in this paper to assess the academic performance of the students. Cluster analysis is used in this study to divide students into groups based on shared attributes. This includes the evaluation criteria for the students, such as internal class grades, GPA, assignments, lab work, and mid- and final exams. It is advised that the class instructor be informed of all of these associated details before the final exam.Teachers will benefit from this re-search by seeing a notable decrease in the dropout rate and an improvement in student performance. This work presents an ideal approach based on the K-Means Clustering algorithm utilizing the MATLABr2023b simulation, which allows educators to improve the quality of education provided to students. Based on this approach, educators can then take the appropriate action to raise students' academic performance. further comprises a thorough examination of the student performance data record results following the MATLABr2023b simulation.

Keywords: k-Meansclustering,cluster mean, centroid, Database, GPA.

1. Introduction

One important data mining technology is cluster analysis. It separates the datasets into a number of significant clusters that each represent the inherent structure of the dataset. A cluster is an assembly of data elements that are similar in common and are measured using the same type of information. By using a process called clustering, data items are grouped into different sets. Data points within a set are referred to as dissimilar sets, while those inside different sets are known as comparable sets. A popular strategy for predicting a student's academic achievement in the future is clustering. A metric of academic performance that is most frequently employed is the graded point average (GPA).

Academicians frequently use GPA as a criterion for assessing a programme in an academic setting. A student's ability to function academically and maintain a high GPA, which is a reflection of their overall academic success, may be hampered by a variety of issues. By tracking the advancement of their students' performance, faculty members can use these elements to target when creating initiatives to increase the calibre of their students' learning behaviours and academic achievement. The K-Means clustering technique using MATLABr2023b makes it feasible to identify the essential traits for predicting student success in the future.

This study presents the K-Means clustering algorithm using the straightforward and effective MATLAB r2023b simulation tool, which tracks the advancement of students' academic achievement in education.

2. LITERATURE REVIEW

One of the most important methods in data mining for analysing data sets is clustering. Over the past few decades, a number of clustering algorithms with improved performance have been used for a variety of applications.In this studyVarious tactics can be categorised as clustering approaches, such as density, model building, situation awareness in online learning, partitioning and hierarchical clustering, and others[1].This research paper presents , the K-means method has been successfully used. K-means is a partitioning approach that divides data into sets according to how close they are to one another.[2]. The original K-means method did not work well for many applications since it used Euclidean distance to replicate an assumed pattern of resemblance between data points.[3]. This Study discussed the benefits and drawbacks of the original K-means algorithm and suggested methods to increase clustering accuracy and shorten calculation times. [4]. This presents the cluster latent data representations, presented a nonlinear function that incorporated a dimensionality reduction (DR) and K-means partitioning. The combined data demonstrated improved performance in addition to the advantages of merging the two activities [5].In this paper author using data from news headlines. determined the optimal number of clusters in the K-means algorithm using both the elbow technique and conventional K-means.[6] Next, the researchers evaluated news headline clustering internally using the purity approach. The results were then used to determine the ideal amount of clusters created using the Elbow approach..[7]In This research paper, K-means technique statistically clusters students' grades and GPA from online learning courses based on past commonalities. The findings demonstrated that every group of students completed distinct assignments to review, completed tests, and dedicated more study time to the assigned chapters.when comparable student activities and backgrounds resulted in comparable performance[8].

3. METHODOLOGY

3.1 Data Clustering

One statistical and unsupervised method for analysing data is data clustering. In order to find hidden patterns and how they relate to one another, it is mostly employed on massive data sets to facilitate swift and effective decision-making. In essence, it is employed to group

similar data values together into homogeneous categories. Stated otherwise, the cluster analysis technique divides the large data set into clusters, which are subsets. Every cluster is made up of a collection of related data values that are grouped together inside the same cluster but differ from the data values found in other items.

3.2 Definition of K- Means Clustering Algorithm

The K-means algorithm selects K objects (data points) at random, each of which initially represents a cluster mean or centre. For each of the remaining data points or objects, an object is assigned to the cluster to which it is most comparable based on the distance between the object and the cluster mean. The new mean for each cluster is then determined. Repeat this method until the criterion function converges. This is a simple technique that has been applied to many different problem fields.

3.3 K-Means Clustering Algorithm

The following steps make up the algorithm:

1. Insert K points into the space that the clustering objects (data points) represent. The first set of centroids is represented by these positions.

2. Assign every item (data point) to the group whose centroid is nearest.

3. Recalculate the locations of the K centroids after assigning all objects (data points).

4. Continue doing Steps 2 and 3 until the centroids stop moving. As a result, the data points for the objects are divided into groups from which the metric that needs to be minimized can be determined.

5. The squared error function is the objective function that this approach seeks to minimize. The goal function is –

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$$

where $\left\| x_i^{(j)} - c_j \right\|^2$ is a chosen distance measurebetween a datapoint $x_i^{(j)}$ cluster centre c_j, is an indicator of the distance of the n data points from their respective cluster centres.

3.4Characteristics of K-Means Algorithm

- The K-means clustering algorithm looks for K partitions where the squared error functions are as little as possible.
- The algorithm's computational complexity is O(nkt), which makes it comparatively scalable and effective in processing huge data sets. Where

N= Total number of objects (data items)

K= Number of Clusters

T= Total number of iterations

- This well-known centroid-based method divides a set of n objects (data items) into k clusters with the parameter k such that the resulting intracluster similarity is high but the intercluster similarity is low.
- This algorithm mainly terminates at the local time.
- The centroid, or centre of gravity, of a cluster is the mean value of the data items in the cluster, and this value is used to quantify similarity in this clustering technique.

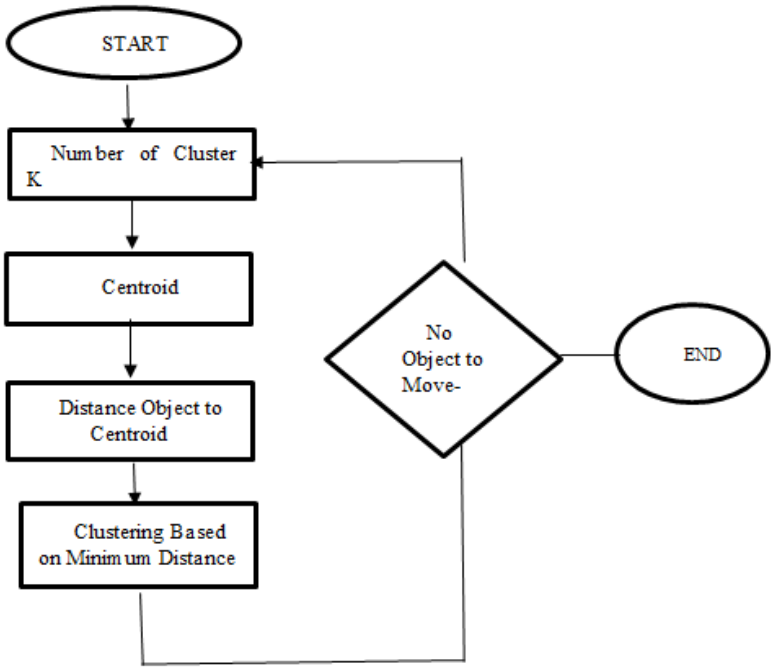

Fig. 1. Flow Chart of K-Means Clustering Algorithm.

3.5 Formula to compute the Euclidean Distance

The distance between two points in the plane with coordinates (x, y) and (a, b) is given by-

dist ((x, y), (a,b)) = $\sqrt{(x-a)^2+(y-b)^2}$(1)

Table 1. Student Database

Rollno	GPA	Class Internals	Attendance	Assignment	Lab Performance
1	365	19	20	YES	GOOD
2	353	18	18	YES	GOOD
3	197	17	18	YES	GOOD
4	134	15	13	YES	GOOD
5	15	12	8	NO	POOR
6	24	10	4	YES	POOR
7	24	10	4	NO	POOR
8	278	20	18	YES	GOOD

This study uses the MATLAB r2023b simulation to demonstrate the application of k-means clustering. In table 1. The "student data" contained in student-data.m, a comma-separated format, served as the basis for the sample data set utilised in this example. Higher education institutions provided the student data figures. This work makes the assumption that the necessary data pre-processing has been carried out. Student-means.m is the data file that was produced. The student data set will determine how the K-means algorithm clusters the data. Three categories are used to classify the cluster: above average, average, and weak. We can determine the pace of progress of the student data set by closely examining the progression. Additionally, we are able to determine the causes of above average, average, weak, and level performance. It provides the exact result. And based on the cluster outcome, we can improve the student's academic performance.

Table 2. GPA Grade of the Students

Performance	GPA	No of Students	Percentage
Above Average	>=3.50	30	50
Average	2.00>=<3.50	25	41.66
Weak	<=2.00	5	8.33

In Table 2. By utilising the K-means clustering algorithm on the training data, we are able to classify the students into three groups based on their new grades: "Above Average," "Average," and "Weak." The previous semester's grade—which includes both internal and external assessments—is used to compute the new grade. The accompanying graph and table are shown below.

4. RESULT

When we use the simulation on MATLAB r2023b for the clustering method to the student data-based. Then we group the students into three categories. One is above average, the second is medium and the Last is weak. A graphical representation of these three categories is given below

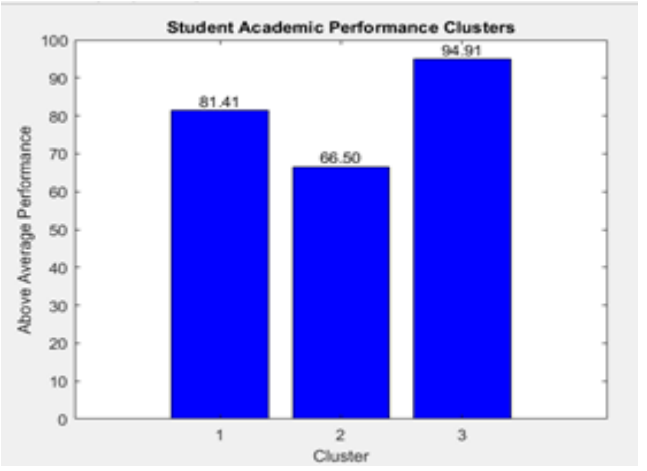

Fig. 2. Above average Performance of the Students

Fig -2 The simulation graph categorises the students in the above-average category. In this three cluster groups are made of 81.41 %, 66.50% and 94.91%. These cluster groups of students are good in academic performance. that shown need not to be taken special care to students.

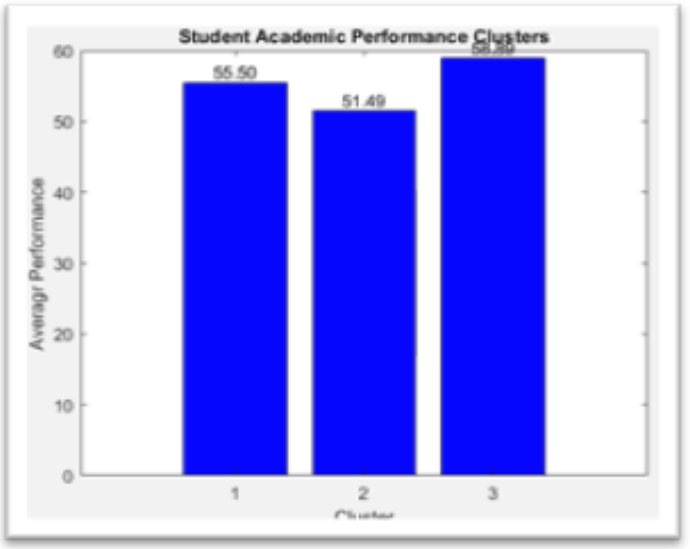

Fig. 3. Average Performance of the students

Fig -3 The simulation graph categorises the students in the average category. In this three cluster groups are made of 56.50 %, 51.49% and 59.80%. These cluster groups of students are good in academic performance. that shows the need to take care of students in the conduction of class tests, quizzes and Lab performance.

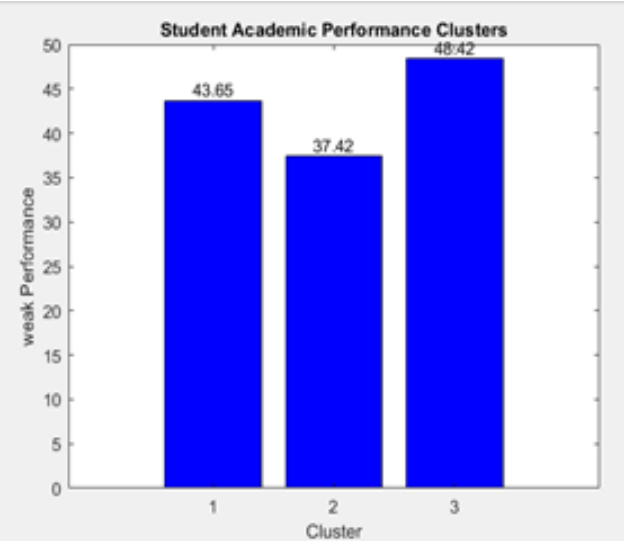

Fig .4. Weak Performance Students

Fig-4 The simulation graph categorises the students in the weak category. In these three cluster groups are made of 43.85 %, 37.42 % and 48.42%. These cluster groups of students are not good in academic performance. that shows Need a lot of practice in his/her lessons and also takes care of all the courses ct,labs,quiz , and attendance carefully.

5. CONCLUSION

In this study, we employ the k-means clustering algorithm as part of a data mining process in the student database to forecast the learning activities of the students. We anticipate that both instructors and students may find use for the material produced by the application of data mining and data clustering techniques. This effort may enhance student performance and reduce the failing ratio by implementing the appropriate measures at the appropriate times to improve the quality of education.By taking the proper actions at the correct time to improve the quality of education, this work may improve student performance and lower the failing ratio. We intend to improve our method in subsequent research to produce more accurate and valuable outputs that will help teachers enhance their students' learning objectives.

References

1. Adeyemo, S. A: The relationship between effective classroom management and students' academic achievement, European Journal of Educational Studies, 4(3), 367-381.(2012).
2. A. Burgess, C. Senior, E. MooresA 10-year case study on the changing determinants of university student satisfaction in the UK PloS One, 13, pp. 1-15(2018)
3. Bell, S. Project-based learning for the 21st century: Skills for the future. The Clearing House, 83(2), 39-43.(2010)
4. B.J. Frey, D. Dueck Clustering by passing messages Between data Points315, pp. 972-976 (2017)
5. Cothran, D. J., & Ennis, C. D. (1997). Students' and teachers' perceptions of conflict and power. Teaching and teachereducation, 13(5), 541-553 (1997).
6. Oyelade, Oladipupo &Obagbuwa, ”Applicationof K-means clustering IJCSIS,vol.7,No.1, pp-292.(2010)
7. Rousseeuw P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and appliedmathematics 20, 53-65 (1987).
8. U. Bodenhofer, A. Kothmeier, S.APCluster HochreiterAn R package for Affinity Propagation ClusteringBioinformatics, 27 , pp. 2463-2464 (2011)