KeAi
CHINESE ROOTS
GLOBAL IMPACT

International Journal of Innovation Studies

ISSN:2096-2487 | E-ISSN:2589-2975

DENSITY BASED SPATIAL CLUSTERING MODEL FOR IDENTIFYING OUTLIER STUDENTS AND SLOW LEARNERS IN HIGHER EDUCATION

Vanitha.S
Department of Computer Applications, Faculty of Science and Humanities,
SRM Institute of Science and Technology, Kattankulathur, Chennai, India
E-mail: vanithas3@srmist.edu.in
ORCID iD: https://orcid.org/0000-0001-9563-7597

Jayashree.R*
Department of Computer Applications, Faculty of Science and Humanities,
SRM Institute of Science and Technology, Kattankulathur, Chennai, India
E-mail: jayashreeram77@gmail.com
ORCID iD: https://orcid.org/0000-0002-0150-7095
*Corresponding Author

Reena Rose. R
Department of Computer Applications, Faculty of Science and Humanities,
SRM Institute of Science and Technology, Kattankulathur, Chennai, India
E-mail: reenaror@srmist.edu.in

Abstract. Predicting student performance in higher education is an emerging topic because it directly correlates with educational quality. The categorization of students' knowledge levels enables educators to support them in improving their performance. Outlier students and slow learners are two significant student groups that require varying levels of assistance from teachers. The outliers are the students whose scores lie on both edges of the performance scale. They represent those who exhibit both extremely poor or outstanding learning abilities. Slow learners are students, who learn at a slower pace than their peers, but they do not necessarily fall into the failure category; they are below-average students. Cognitive assessment is a primary approach that allows educators to organize students based on their knowledge. The periodic class test or semester examination scores indicate the student's outcome in each subject. Subsequently, teachers can provide personalized instruction, employing different methods to teach the subject and supporting students to attain the best possible results. The contemporary advances in learning analytics and machine learning help to devise efficient models that predict the student's difficulties in advance. Existing studies utilized several parameters, including demographic data, study behavior, and academic details, to categorize the student group. However, this study aims to use academic scores to categorize students using K-Means and the DBSCAN algorithm using unlabelled data. The DBSCAN outperforms the K-means model with the highest silhouette and Davies-Bouldin index.

Index Terms: Slow Learner Prediction, Clustering Outlier Students, K-means, DBSCAN, Performance Classification.

1. Introduction

The Higher Education System (HES) is the backbone of every country, dedicated to providing quality education to the learners. The preliminary focus is to sustain quality education and opportunities for students. Higher education includes various programs such as undergraduate, graduate, professional, vocational, and technical education. The undergraduate programs are the first level of higher education, and they are designed to take three or four years to award a bachelor's degree.

The HES implements several measures to ensure the well-being of students in the academic curriculum as well as facilities given by the institution. The quality of education is also evaluated by various parameters such as curriculum relevance, teaching excellence, evaluation system, technology integration, infrastructure, transport, and amenities provided to the students and faculties. It is achieved through proper admission planning and required resource allocation by the management. After the admission process, the further challenge to any educational institution is the student-success ratio. There has been a consistent decline in students' academic achievement on an annual basis, especially in undergraduate programs.

Following the outbreak of COVID-19, students' study habits underwent a sudden change as a result of the shift to online classes, where students utilize mobile phones for educational purposes. The students become addicted to spending significant amount of time on the internet and gaming programs. The advent of online learning platforms and digital technologies leads to innovation and revolution of traditional educational approaches. Even though the development, the student’s pass percentages fall yearly due to the lack of interest. The advancement of Artificial Intelligence and the availability of extensive student data enable us to analyze the progress of our next generation by frequent evaluations throughout their academic journey. This Periodic evaluation is required to improve the students’ performance by warning the weak students, teachers, and parents. This research work fulfills the academic problem statement early intervention of at-risk students, and identifying outlier students in the classroom using unsupervised machine learning.

Nowadays, Artificial Intelligence (AI) is widely involved in all back-end education processes such as predicting student performance, course recommendation systems, and admission prediction. Apart from the teaching responsibilities, instructors engage in other administrative tasks within the education system, including the preparation of study materials, creating question papers, tracking attendance, maintaining log books, assessing student performance, and analyzing class results. In the present era, instructors are using Google as a means to effectively create study materials, question papers, and particularly for self-preparation. Likewise, there are student evaluation and result analysis tools that allow for immediate results by inputting student data. Artificial Intelligence (AI) is the driving force behind these apps, ensuring the delivery of the most accurate results. To achieve that, Artificial Intelligence necessitates historical data to train the model.

The advancement of artificial intelligence and massive educational data can aid in analysing progress, improving performance, and alerting weak students, teachers, and parents to potential

problems. The K-means and DBSCAN methods are selected among various techniques to categorize students in a classroom based on their subject scores and CGPA values.
The following outlines the reason behind the selection of the algorithm and the optimal working scenario:

Case 1: Typically, the students are categorized into three groups based on their performance, which are poor, average, and good in their studies. Therefore, creating three clusters is fruitful to identify the weak students who are learning slowly in a subject. K-means clustering, which is a centroid-based method, effectively achieves this criterion.

Case 2: In addition to the general classification, there is another category of students that includes the extremely poor and the top performers in each subject. They are known as outliers. DBSCAN is used to identify outlier students since it is efficient in anomaly detection. In other words, this model can simply distinguish students with different study behaviors. Predicting the least and most outstanding students is crucial to motivate them to reach the next level. Supporting the weak students will reduce the failure percentage, and encouraging the best-performing learners will enable them to become rank holders. The following circumstances also explain the appearance of outlier students:

- Despite the simplicity of the question paper and subject, some students may fail the course. They are the least-performers because the peer's score distribution will be good or excellent.
- Despite the challenging nature of the question paper and the high level of subject difficulty, only a small number of students will achieve the highest score in the subject. These students are the top performers in the subject, while the other's score distribution falls into the average or below-average categories. The research questions and contribution of this paper are as follows:

1. Which is the suitable clustering technique when the difficulty level of the subject and question paper is easy, when the students' knowledge level is good in the subject?
2. Which is the right model to identify the outlier students, when the difficulty level of the subject and question paper is high, and students' knowledge level is not distributed evenly in the classroom?
3. Does the number of features impact the precision of the clustering model?

The remaining paper contains the following sections: section 2 illustrates the related work, and section 3 comprises the dataset, methodology, and metrics employed for this experiment. The study results are discussed in section 4, and the conclusion and future perspectives are added in the last section.

2. Literature Survey

2.1 A Clustering models for student performance classification

The related work implemented for this topic is reviewed as follows: Omolewa et al., [1] applied k-means clustering technique to classify the students by collecting the dataset from the UCI machine learning repository. After applying k-means clustering, two clearly defined clusters were identified. In this study, the slow learners are not separated individually. Santosa et al.,[2] utilized the K-means clustering method to forecast the grade point average of learners based on different student profile to assist in the student admission process. The clustering process involves the use of various academic scores including four features (X1-X4) for students admitted based on merit, and nine attributes (X1 - X9) for students admitted through normal

admission. Prakash et al.,[3] investigate the efficacy of a student performance prediction model (ESVM) using an effective clustering technique. The evaluation results demonstrate a significant boost in classification tasks compared to existing systems which uses manual labeling.

Wang [4] suggested K-means clustering algorithm as a technique to assist instructional administrators in discovering the learning characteristics of students. This approach tackles challenges in assessing student performance caused by differences in course complexity using noisy dataset. Fida et al., [5] formulated a model by integrating the classification and clustering approaches to achieve more accurate prediction outcomes. The primary goal of this study is to create a smarter dataset by reducing irrelevant features. This analysis determines the implications of particular features and the influence of eliminating wrongly classified data to achieve 93% accuracy in predicting student performance. Kolawole et al.,[6] introduced a modified K-means clustering method which is designed to efficiently and accurately assess student performance using only the assessment score and attendance. Nafuri et al., [7] suggests a clustering-based prediction model to categorize students according to their academic score in higher education using three unsupervised models (K-means, BIRCH, and DBSCAN). The result shows that the optimized k-means demonstrated superior performance compared to all other models.

Omar et al., [8] proposed a method applying the k-means clustering and elbow method to assess student performance based on the academic grade and GPA to yield more precise outcomes. The goal of this research is to identify the college students who are not performing well academically to prevent them from being placed on academic probation or facing possible termination. Henderi et al., [9] constructed a model using the k-means method for evaluating the academic performance of students. The result shows that the student learning outcomes yielded three different clusters: excellent, satisfying, and poor with the following percentages respectively 40%, 44%, and 16%. Vasuki et al.,[10] created a model utilizing the K-means method to assess the learner's performance in soft skills such as Aptitude, English, Programming Logic, and Coding. Moreover, other clustering techniques, such as various implementations of kcc++, genetic k-means, and greedy optimized k-means, have also experimented with this dataset. Alhazmi et al.,[11] combine the clustering and classification methods to determine the influence of students' early-stage performance on their GPA. The result shows that the educational systems have the potential to reduce the probability of students experiencing failures during the initial phases.

Setiabudi et al., [12] investigate the correlation between student activities and grades in two different courses. The K-Means method was employed to cluster the students, observing students' activity patterns in teaching. The results show that students' performance increases with the number of activities and points earned. The number of clusters is decided based on the Smallest Sum of Squared Error (SSE) compared to the average. Nur et al.,[13] identifies the significance of unsupervised clustering method in labeling for better student classification using SVM. The objective of this study is to discover essential attributes inside the student groups and construct a prediction model for each group to forecast the success of students. This study specifically examines the importance of students' academic and non-academic features.

Critical Analysis: Researchers commonly use unsupervised clustering techniques such as K-means, fuzzy C-means, DBSCAN, and BIRCH to group students. The K-means clustering is

primarily used in many studies [1, 2–10, 12, 13], followed by fuzzy C-means overlapping clustering. There are mainly two purposes utilized in the existing studies: i) labeling the students' category to predict the student's class using a supervised classification model [1, 2, 3, 5, 12, 13] and ii) grouping the students based on their similarity features to analyze the patterns [4]. The current study follows the second one. Prakash et al. (2021) and Nur et al. (2023) demonstrated that the classification accuracy is improved when labels were assigned using the clustering method instead of manual labeling [3, 13]. Also, the existing research [7] utilizes non-academic details, which makes it impractical from an educator's point of view to collect all the student details. A previous study [14] provided evidence for this. The related work identified a research gap in distinguishing outlier students and slow learners from the general classification of poor, average, and good-performing students. This objective will fill the research gap identified to assist educators in motivating slow learners.

3. Methods and Materials

3.1 Dataset Description

This study collects student academic information from 1102 students at a multidisciplinary university. It includes the students' first four semester examination grades for six subjects, as well as their CGPA. This research concentrates only on the undergraduate students due to their adolescent nature compared to the postgraduate students. The undergraduate programs consist of various faculties, such as Arts and Science, Management Studies, and Law, comprising different courses. Table 1 describes the dataset information.

Table 1. Student Academic Information

Number of Students	Faculties	Courses	Academic Features
1102 Undergraduate Students	Faculty of Science and Humanities, Faculty of Management, Faculty of Law.	BCA BSC BCOM BA BBA BL	Paper1, paper2, paper3, paper4, paper5, paper6, CGPA

3.2 Pre-processing

Data pre-processing is mandatory in machine learning models because it ensures the quality and suitability of the collected data to provide valuable results. It comprises several steps as follows:

Step 1: Data Encoding: Data encoding refers to the conversion of categorical variables into numerical ones. The student's grade values are converted into grade points. The random generation method then creates the corresponding numerical scores based on the grade rank.

Step 2: Normalization: Normalization is the process of rescaling data from the original range to a new range of values with a relative distance. The range specifies the lowest and highest values of the particular column. After encoding the data, this study applies MinMaxScalar function, and the reason for choosing is as follows: It is beneficial where all features

contribute equally to the distance metrics. Here, the attributes selected for this study are student's exam scores out of 100.

The MinMaxScalar is sensitive to outliers, but the collected dataset does not have outlier values. The range can be of any value. Every feature automatically fixes its minimum and maximum values to the fixed range. This study ranges the values 0 to 1. It preserves the relation-ship between the data without altering the distribution's shape. The minimum and maximum value of the column is utilized to calculate each element of the dataset as follows:

$$y = ((x — min)) / ((max — min)) \quad (1)$$

Where y - refers output value after scaling a data, x denotes input data, min represent the minimum value of a column, and max refers the maximum value of the column.

Step 3: Data Transformation: This study utilizes the Principle Component Analysis (PCA) method to transform the students' semester examination scores of all subjects into two-dimensional data. PCA is a dimensionality reduction technique that converts high-dimensional information into a lower-dimensional form. It also maintains as much data variability as possible. This transformation helps to visualize the patterns between the clusters in a reduced dimension. The following are the steps taken in the PCA transformation:

- Normalize the data
- Compute the Covariance Matrix
- Calculate the Eigenvalues and eigenvectors
- Selecting the Principle components
- Transform data

3.3 K-Means

K-Means is a centroid-based clustering algorithm that splits data into K unique and non-overlapping subsets, known as clusters [3, 6]. The objective is to define the best categories of student groups in the dataset to identify the weak, average, and best-performing students.

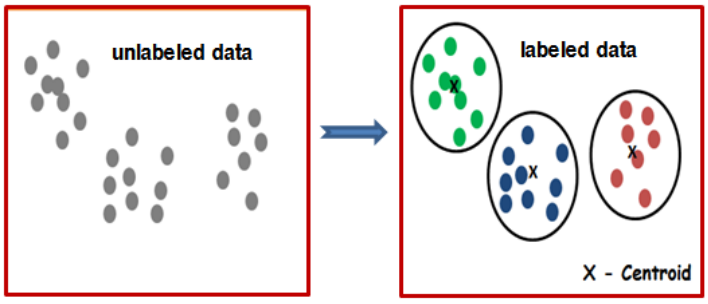

Fig. 1.K – means clustering with centroids

Figure. 1 shows the K-means clustering model working methodology and the steps followed in that algorithm is given below:

Initialization: It encloses two processes; selecting the k number of clusters and initializing the k center points. The elbow method is a simple technique for deciding the appropriate number of clusters for the dataset. Then, the K-means++ optimization algorithm randomly initializes the first centroid point and picks subsequent centroids to construct the clusters. It computes the distance for every data point before concluding the centroids.

Assigning Points to Clusters: Each data point in the dataset is assigned to the nearest centroid based on the distance estimated by the Euclidean distance measure algorithm.

Updating centroid: Once all the data points are assigned to the clusters, the position of the centroid in each cluster is recalculated. The new centroid represents the mean of all the data points in that cluster.

Repeating Steps 2 and 3: Continue the assignment and update process until there are no significant changes to the centroids. It helps refine centroids' positions and allocate data points to clusters.

Finalizing Clusters: The clusters are finalized, once the model converges. That means there is little or no shift in the centroid position.

3.4 Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

DBSCAN is a density-based clustering method [7] known for its effectiveness in detecting clusters of various shapes and managing noise in the dataset. So, this clustering technique is utilized in this experiment to identify the outlier students who stand on both edges of the knowledge level. DBSCAN is parameter-sensitive because the cluster construction depends on max_dist and min_samples. Figure 2 shows the working principle of the DBSCAN algorithm and the steps followed is given below:

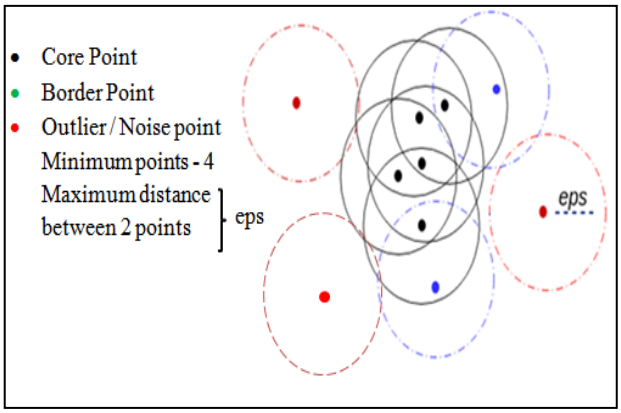

Fig. 2 DBSCAN key points to create cluster

Parameter selection (max_dist(ε), min_samples): The epsilon (ε) refers to the allowed distance between two points to be considered neighbours. The min_samples denote the minimum number of data points that need to be in the given radius to form the cluster.

Examine core points, border points, and noise. The core point refers to the point with the minimum number of neighbours and within the ε distance. The border point is within the core point's range and lacks sufficient neighbours. Finally, an outlier, or noise point, is a point that is not within the ε distance and lacks sufficient neighbours.

Cluster Construction: Choose a data point. Choose a randomly selected point that has not yet undergone verification.

Neighbourhood Check: Retrieve all the data points within the ε radius of the selected point.

Core Point Check: If it is a core point, create a cluster; otherwise, label it a border point and move to the next point.

Expand Cluster: Include all of the neighbours. If the neighbour is the core point, repeat the neighbourhood check and add all the nearest points within the ε distance.

Repeat: Continue following these steps until each point is verified.

Noise Management: The points that are neither core nor border points are noise points. Thus, the core clusters are created by eliminating outliers.

3.5 Performance Evaluation

Silhouette Score. The Silhouette Score [7] is a measure used to assess the effectiveness of a clustering technique based on the compactness of individual clusters (intra-cluster distance) and the separation between clusters (inter-cluster distance). Eq. (1) explains the score calculation.

$$s_i = \frac{b_i - a_i}{max(b_i, a_i)} \qquad (1)$$

Where refers inter cluster distance and refers intra cluster distance. The overall Silhouette score for the whole dataset is determined as the average of the silhouette scores for all data points in the dataset. The score will always range from -1 to 1, with 1 indicating superior clustering.

Davies Bouldin Index (DBI). DBI is another metric for clustering. Similarity is determined by the ratio of distances within clusters to distances between clusters. Clusters that are more distant and less spread out will lead to a higher score. The lowest possible score is 0, where smaller values suggest more effective grouping. Eq. (2) explains the score calculation.

$$DBI = \frac{1}{k}\sum_{i=1}^{k} \underset{j \neq i}{max} \frac{s_i + s_j}{M_{ij}} \qquad (2)$$

Where indicates intra-cluster compactness refers to the average distance of all data points within the clusters to the centroid. And denotes inter-cluster separation, which refers to the distance between the centroids of clusters i and j. The k represents the number of clusters.

3. Results and Discussion

The student's performance in each subject will be different, so the minimum score will change arbitrarily. In some subjects, 50 to 70 marks are considered good, but in others, it may be 70 to 90. The min-max scalar function provides a solution for fixing each subject's minimum and maximum scores in a range. It helps to predict the poor performers based on the subject's lowest score rather than a constant value. For example, 40 represent an undergraduate's pass mark. This experiment developed three different sets of models with various numbers of features, and the results are as follows:

4.1 Identifying slow learners and outstanding students based on CGPA

This experiment clusters the 1102 students based on their CGPA value. The elbow method helps to determine the optimal number of clusters (K) for grouping the students. The graphical approach involves identifying a point on the plot where the within-cluster sum of squares (WCSS) begins to fall at a slower rate, forming an "elbow" shape. Figure. 3 represents the elbow method with inertia, also known as WCSS. The curve bends where the number of clusters is 3, which is the optimal number of clusters. Figure. 4(a) displays the K-means result, classifying the students into three groups.

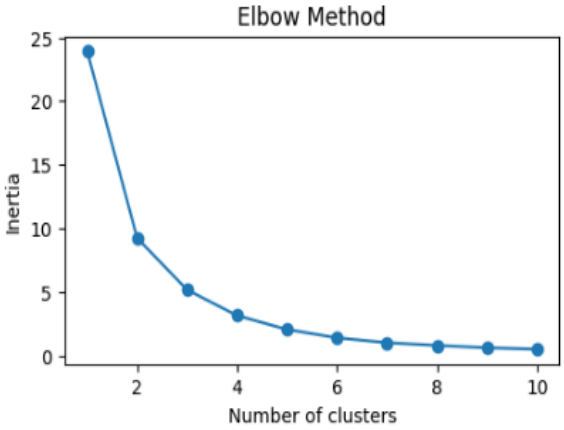

Fig.3. Elbow method showing optimal number of clusters

In Figure 4(a). the 0.0 on the x-axis represents the lowest CGPA score, while the 1.0 represents the highest score. Violet, green, and yellow colors classify the low, average, and best performers, respectively. The k-means output shows the labels assigned to each student category. Accordingly, 0 is for 103 slow learners, 1 is for 624 best scores and 2 for the 375 average students.

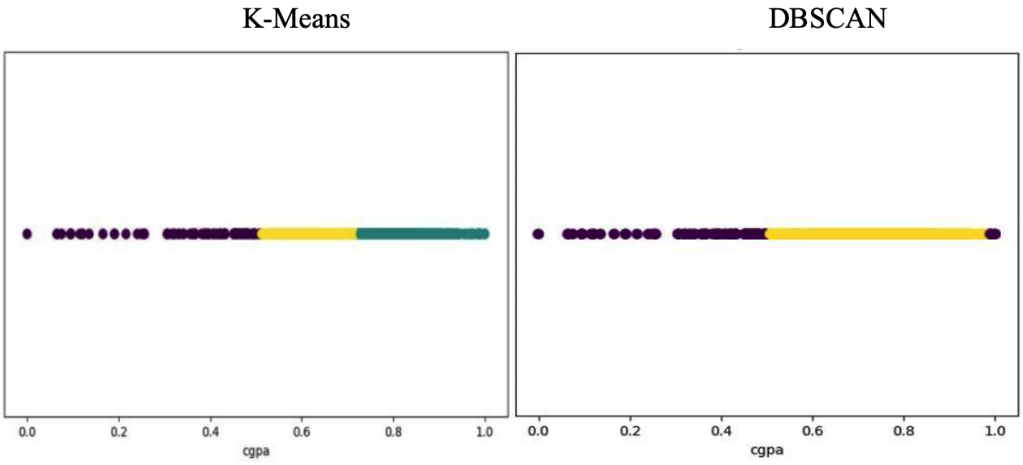

Fig.4. (a) K-means clustering using CGPA (b) DBSCAN clustering using CGPA

Figure 4(b). presents the DBSCAN results. The two violet colour edges represent outlier students, including poor and best-performing students, whereas the yellow shows that core or normal students have more peers. Out of 1102 students, 105 falls into the outlier category, leaving 997 in the remaining group. The DBSCAN output shows core and outlier students with the labels assigned 0 and -1, respectively.

Output: **K-Means** **DBSCAN**

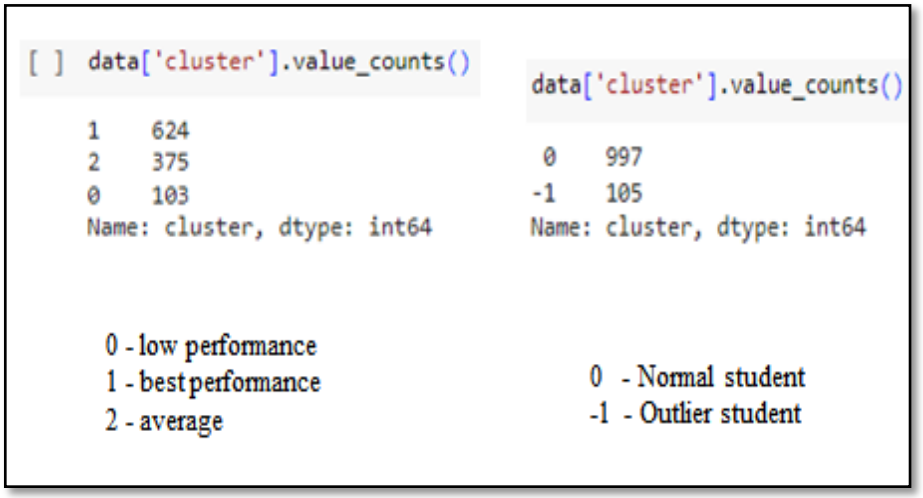

Table 2. Clustering the students using one feature.

Number of Students – 1102, Feature - CGPA						
Metrics			**Hyper-parameters**			
Model	**Silhouette Score**	**DBI**	**Total Clusters**	**Inertia**	**Epsilon(ε)**	**Min Samples**
K-means	0.54	0.58	3	5.15	-	-
DBSCAN	0.64	0.60	-	-	0.1	300

Table 2 shows the metrics and hyper-parameter details of this model. In the hyper-parameter setting, the number of clusters and inertia belong to K-means clustering, whereas the epsilon (ε) and minimum samples are for DBSCAN. The elbow method finds the optimal number of clusters for K-means. However, the maximum distance between two data points and the minimum samples required to create a cluster for DBSCAN is selected through trial and error. The results show that the silhouette score for K-means and DBSCAN is 0.54 and 0.64, respectively. The DBI scores achieved by K-means and DBSCAN are 0.58 and 0.60, respectively.

4.2 Identifying slow learners using subject score and CGPA

In this experiment, student performance in a particular subject is identified by comparing the CGPA of 118 students using the semester score details. This model helps to identify the students' overall performance of the semester and the particular subject. Figure. 5(a) shows the K-means result. The output shows the labels assigned for each class: 0, 1, and 2, which refer to the best, lowest, and average performers, respectively.

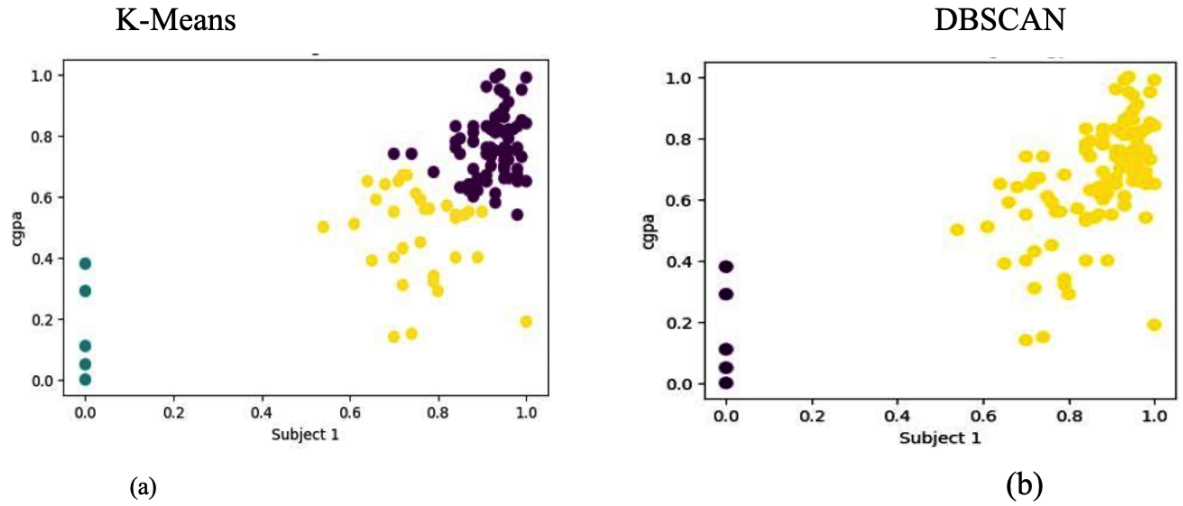

Fig. 5(a). K-means using 2 features (b) DBSCAN using 2 features

Figure. 5(b) describes the DBSCAN result, which classifies five poor-performing students as outliers on the subject. In the output, the labels 0 and -1 refer to core and weak students, respectively. The results show that both methods accurately predict the subject's least performers. These students lack interest in the subject compared to others plotted on the right side.

Output: K-Means DBSCAN

```
data['cluster'].value_counts()
0    80
2    33
1     5
Name: cluster, dtype: int64
```

0 - best
1 - low
2 - average

```
data['cluster'].value_counts()
 0    113
-1      5
Name: cluster, dtype: int64
```

0 - Normal student
-1 - Outlier student

Table 2 displays the result and hyper-parameter details of this model. The outcomes reveal that the silhouette score for K-means and DBSCAN is 0.51 and 0.76, respectively. The DBI scores attained by K-means and DBSCAN are 0.64 and 0.29, respectively.

Table 2. Clustering the students using two features.

Number of Students – 118, Features – Subject 1 score, CGPA						
Metrics			**Hyper-parameters**			
Model	**Silhouette Score**	**DBI**	**Total Clusters**	**Inertia**	**Epsilon (ε)**	**Min Samples**
K-means	0.51	0.64	3	2.2	-	-
DBSCAN	0.76	0. 29	-	-	0.3	40

4.3 Identifying slow learners based on all subject of a semester

This experiment is to identify the student category based on the performance of all subjects in a semester. The principal component analysis (PCA) is used to reduce the student features into two (2D) dimensions and visualize the patterns.

Figure. 6(a) shows the K-means result; yellow represents 29 slow learners, green displays 54 average students, and violet for 35 best-performing students. The labels in the output 0, 1, and 2 refer to the best, average, and low performers. Likewise, in DBSCAN output, 0 and -1 refer to core and outlier students, respectively. Fig. 6(b) shows the result; the violet depicts the 26 outliers, of which two outstanding students are located in the top centre of the figure, and the remaining are dull students in the semester, located in the right and bottom centres.

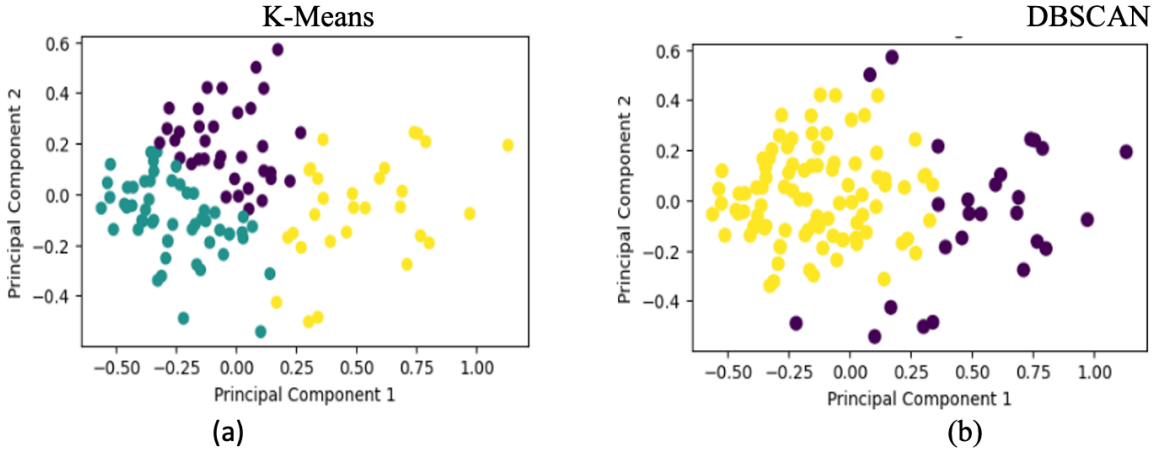

Fig. 6(a) K-means using 6 features (b) DBSCAN using 6 features

Output:

K-Means

```
data['cluster'].value_counts()

1    54
0    35
2    29
Name: cluster, dtype: int64
```

0 - Best
1 - Average
2 - Low

DBSCAN

```
data['cluster'].value_counts()

 0    92
-1    26
Name: cluster, dtype: int64
```

0 - Normal student
-1 - Outlier student

Table 3. Clustering the students using six features.

Number of Students – 118, Features – paper 1, paper 2, paper 3, paper 4, paper 5, and paper 6.						
Metrics			**Hyper-parameters**			
Model	**Silhouette Score**	**DBI**	**Total Clusters**	**Inertia**	**Epsilon(ε)**	**Min Samples**
K-means	0.18	1.77	3	7.3	-	-
DBSCAN	0. 32	1. 43	-	-	0.3	40

Table 3 shows the result and hyper-parameter details of this model. The result exhibit that the silhouette scores for K-means and DBSCAN is 0.18 and 0.32, respectively. The DBI scores reached by K-means and DBSCAN are 0.77 and 0.43, respectively. The Figure.7 compares the model's performance based on the number of features using the silhouette and Davies-Bouldin index. The DBSCAN model using two features gives a better result than the other combinations with Silhouette - 0.76, and DBI - 0.29.

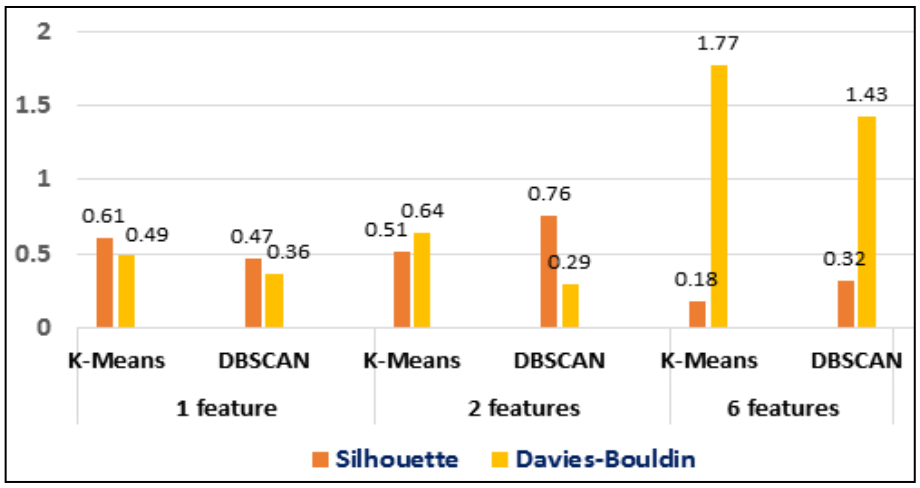

Fig.7. Clustering model's feature-wise comparison

Figure. 8 shows the current model using six features surpasses the existing system with the highest silhouette scores of 0.18 and 0.32 for K-means and DBSCAN, respectively. The DBI scores for K-means and DBSCAN are 1.77 and 1.43, respectively. The existing work [7] uses non-academic details with several features to group the students. However, this study demonstrates that academic data alone can accurately predict a student's current status, even with limited features.

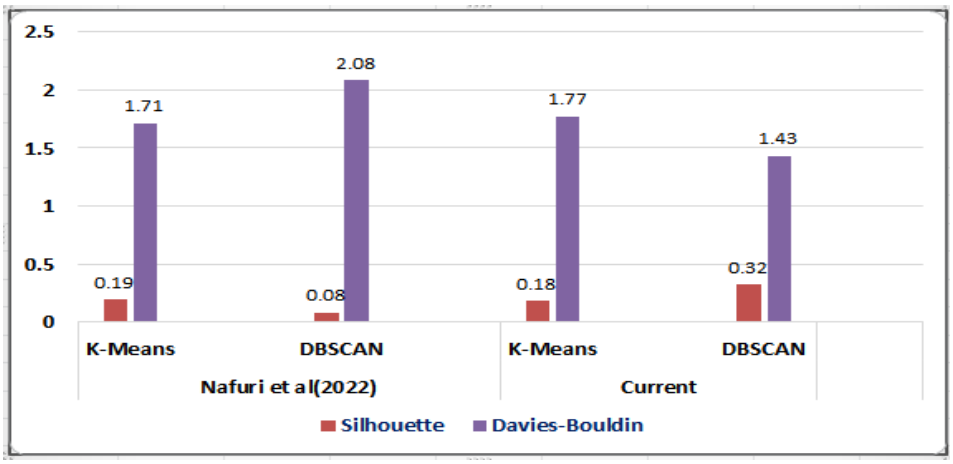

Fig. 8. Comparison of existing and proposed method

In response to the first research question (RQ1), K-means works best when the data distribution centers at a specific point, having closer samples. As per the results in Figures. 4(a), 5(a) and 6(a), the K-means performs better than DBSCAN for general classifications of students such as poor, average, and good in studies, apart from outliers. To answer the second question (RQ2), DBSCAN is unsuitable for general classification because it concentrates only on noise and core

data points, which have an uneven distribution. The result in Figures. 4(b), 5(b) and 6(b) shows two categories of students based on the density and distance between the points. Here, the density refers the study behaviour of the learners. For the third question (RQ3), a smaller number of features led the model to cluster efficiently. Figure. 7 proves that the model with two features reached the highest accuracy.

5 Conclusions

This experimental study clusters the least and best-performing students using the DBSCAN model, which is sensitive to the noise data. Additionally, the k-means algorithm identified the slow-learning students in each subject to improve their CGPA scores. This study categorizes the learners based on their performance in each subject and overall CGPA score. First, the collected dataset is pre-processed by scaling and data transformation using the Min-Max scalar and the PCA technique, respectively. Next, the clustering models were employed by involving three distinct sets of features to evaluate the efficiency of model. The number of optimal clusters is identified using the elbow method. Finally, the results revealed that the increasing number of features led to the complexity of classifying the student group. Moreover, the DBSCAN performs better than K-means with a score of 0.76 (silhouette) and 0.29 (DBI) using two features and surpasses the existing system with the model using six features. The future scope is to implement the classification model using this labelled dataset and compare the outcome with manual labelling.

Conflicts of Interest

The authors declare that there is no conflict of interest concerning the publication of this paper.

Author Contributions

The author's contributions to this paper are as follows: Conceptualization, methodology, validation, formal analysis, investigation, resources, data curation, writing review, original draft preparation, and visualization have been done by Vanitha.S. The supervision and editing have been done by Jayashree. R.

References

[1] O. T. Omolewa, A. T. Oladele, A. A. Adegun, and R. O. Ogundokun, “Prediction of Student’s Academic Performance using k-Means Clustering and Multiple Linear Regressions,” Journal of Engineering & Applied Sciences/Journal of Engineering and Applied Sciences, vol. 14, no. 22, pp. 8254–8260, Oct. 2019, doi: 10.36478/jeasci.2019.8254.8260.

[2] R. G. Santosa, Y. Lukito, and A. R. Chrismanto, “Classification and prediction of students’ GPA using K-means clustering algorithm to assist student admission process,” Journal of Information Systems Engineering and Business Intelligence, vol. 7, no. 1, p. 1, Apr. 2021. doi:10.20473/jisebi.7.1.1-10.

[3] K. P. Prakash and K. Selvakumari, “An Intelligent Clustering Technique for Analysing the Performance of Students during Lockdown Period of Covid-19,” vol. 12, no. 9, pp. 2499–2512, Apr. 2021, doi: https://doi.org/10.17762/turcomat.v12i9.3733.

[4] Z. Wang, “Higher Education Management and Student Achievement Assessment Method Based on Clustering Algorithm,” Computational Intelligence and Neuroscience, vol. 2022, pp. 1–10, Jul. 2022, doi: https://doi.org/10.1155/2022/4703975.

[5] S. Fida, N. Masood, N. Tariq, and F. Qayyum, "A Novel Hybrid Ensemble Clustering Technique for Student Performance Prediction," Journal of universal computer science, vol. 28, no. 8, pp. 777–798, Aug. 2022, doi: https://doi.org/10.3897/jucs.73427.
[6] K. K. Kolawole, Dr. A. A. O, S. J. A, and B. W. Adebayo, "Performance evaluation student result using k-means clustering," International Journal of Communication and Information Technology, vol. 3, no. 1, pp. 01–05, Jan. 2022, doi: https://doi.org/10.33545/2707661x.2022.v3.i1a.35.
[7] A. F. M. Nafuri, N. S. Sani, N. F. A. Zainudin, A. H. A. Rahman, and M. Aliff, "Clustering Analysis for classifying student academic performance in higher education," Applied Sciences, vol. 12, no. 19, p. 9467, Sep. 2022, doi: 10.3390/app12199467.
[8] T. Omar, A. K. Alzahrani, and M. Zohdy, "Clustering approach for analyzing the student's efficiency and performance based on data," Journal of Data Analysis and Information Processing, vol. 08, no. 03, pp. 171–182, Jan. 2020, doi: 10.4236/jdaip.2020.83010.
[9] Henderi, A. Sunarya, Zakaria, S. L. Nurmika, and N. Asmainah, "Evaluation model of students learning outcome using k-means algorithm," Journal of Physics: Conference Series, vol. 1477, p. 022027, Mar. 2020, doi: https://doi.org/10.1088/1742-6596/1477/2/022027.
[10] M. Vasuki et al., "Evaluating Students Placement Performance Using Normalized K-Means Clustering Algorithm," Turkish Journal of Computer and Mathematics Education (TURCOMAT), vol. 12, no. 11, pp. 3785–3792, May 2021.
[11] E. Alhazmi and A. Sheneamer, "Early Predicting of Students Performance in Higher Education," IEEE Access, pp. 1–1, 2023, doi: https://doi.org/10.1109/ACCESS.2023.3250702.
[12] D. H. Setiabudi and M. Santoso, "Effect of students' activities on academic performance using clustering evolution analysis," CommIT (Communication and Information Technology) Journal/Commit Journal, vol. 17, no. 2, pp. 209–219, Sep. 2023, doi: 10.21512/commit.v17i2.9053.
[13] Nur, A. Majid, and ShahnorbanunSahran, "Identification of Student Behavioral Patterns in Higher Education Using K-Means Clustering and Support Vector Machine," Applied sciences, vol. 13, no. 5, pp. 3267–3267, Mar. 2023, doi: https://doi.org/10.3390/app13053267.
[14] Vanitha and Jayashree, "ED-NET: Multivariate Time Series approach for uncovering student learning outcome in higher education using blended deep learning technique," International Journal of Intelligent Engineering and Systems, vol. 17, no. 2, pp. 526–543, Apr. 2024, doi: 10.22266/ijies2024.0430.43.