International Journal of Scientific Research and Engineering Development-– Volume 8 Issue 5, Sep-Oct 2025
Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Comparative Analysis of Clustering Algorithms for NOUL Entrance Examination Data

Pheth SONENVILAY*, Bounmy PHANTHAVONG*, Mounphine PHONEPANYA *, Souphaivy THIPPHAVONG *,Bouaketh VANNACHIT*

*(Department of Computer Science, National University of Laos, Vientiane Capital, Lao P. D. R

Email : p.sonevilay@nuol.edu.la, b.phanthavong@nuol.edu.la, mphonepanya@nuol.edu.la, s.thipphavong@nuol.edu.la, vannachit_keth@hotmail.com)

----------**********----------

Abstract:

The main objective of this study is to compare machine learning techniques for clustering, specifically using the K-Means and DBSCAN algorithms. The paper involved building, testing, and evaluating clustering models on real-world data. The dataset, collected from the National University Entrance examination (2014-2015), contains 15,602 samples. The Elbow method was applied to determine the optimal k value for K-Means, while the optimal epsilon value was used for DBSCAN. Python served as the development environment. The results showed that the K-Means model produced three clusters with a silhouette coefficient of 0.457, while DBSCAN achieved a higher silhouette score of 0.484. Therefore, DBSCAN outperformed K-Means in clustering performance.

Keywords — **K-Means, DBSCAN, Elbow, K-distance graph, Silhouette Score, Clustering, Machine Learning Technique**

----------**********----------

I. INTRODUCTION

The Clustering is one of the most widely used techniques in machine learning and data mining, aimed at grouping similar data objects without prior class labels [1]. It plays an essential role in discovering hidden patterns and structures within large datasets, which is particularly useful for educational data mining (EDM) where student-related data can be analysed to reveal trends, performance levels, or behavioural patterns [2]. Among various clustering algorithms, K-Means and DBSCAN have emerged as two of the most prominent methods due to their simplicity, scalability, and effectiveness in handling diverse datasets [3]. K-Means is a partition-based clustering algorithm that minimizes the intra-cluster variance by iteratively assigning data points to the nearest cluster centroid [4]. It is computationally efficient and works well with large datasets; however, it requires the number of clusters (k) to be defined in advance and performs poorly with noisy data or clusters of irregular shapes. To address the limitation of selecting the optimal number of clusters, the Elbow Method has been widely adopted as a heuristic to estimate the best value of k [5]. In contrast, DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a density-based algorithm that groups together closely packed data points while identifying outliers as noise [6]. Unlike K-Means, DBSCAN does not require the number of clusters to be specified beforehand and is effective in identifying clusters of arbitrary shapes, though its performance heavily depends on selecting appropriate parameter values for epsilon (ε) and the minimum number of points (MinPts) [7]. In the context of higher education, analysing entrance examination data has become increasingly important for academic institutions to understand student distribution, performance variations, and decision-making for admissions [8]. The National

University of Laos (NUOL) conducts entrance examinations every year for thousands of applicants across the country. The dataset from the academic year 2014–2015 contains 15,602 samples, which provide a rich source for data-driven analysis [9]. Applying clustering algorithms to such data allows researchers to uncover meaningful groupings, such as student performance levels or subject-specific patterns, which can ultimately support the improvement of admission processes and educational planning [10]. Recent studies have compared K-Means and DBSCAN in different application domains. Jain [3] highlighted that K-Means remains a standard reference point for clustering studies due to its simplicity and efficiency, whereas Ester et al. [6] demonstrated the robustness of DBSCAN in handling noise and discovering clusters with non-linear boundaries. In educational data mining, clustering has been applied to identify student learning behaviours, group exam scores, and evaluate curriculum effectiveness [2],[8]. The Silhouette Coefficient and Silhouette Score are commonly employed to measure the quality of clustering results, with higher values indicating better separation between clusters [11].

This paper aims to contribute to the growing body of research on clustering in educational datasets by comparing the performance of K-Means and DBSCAN on the NUOL entrance examination dataset. Python was used as the primary development environment due to its powerful libraries for data science and machine learning [12]. Specifically, the Elbow Method was used to determine the best k-value for K-Means, while DBSCAN was tuned using optimal ε and MinPts parameters. The results show that K-Means generated three clusters with a Silhouette Coefficient of 0.457, while DBSCAN achieved a Silhouette Score of 0.484, indicating superior performance. This comparative study provides valuable insights into the application of clustering algorithms for educational data mining and highlights the potential for adopting advanced data-driven approaches in higher education.

II. THE PROPOSED AND METHODOLOGY

The objectives of this paper are as follows: To analysed the entrance examination dataset and develop a clustering model for student admission at the National University of Laos (NUOL) by applying the K-Means algorithm, a compare different methods for determining the optimal number of clusters (*k values*) suitable for the K-Means model and evaluate and test the performance of the developed clustering model. For the research methodology is defined within the conceptual framework of the study, as illustrated is show in the Fig 1

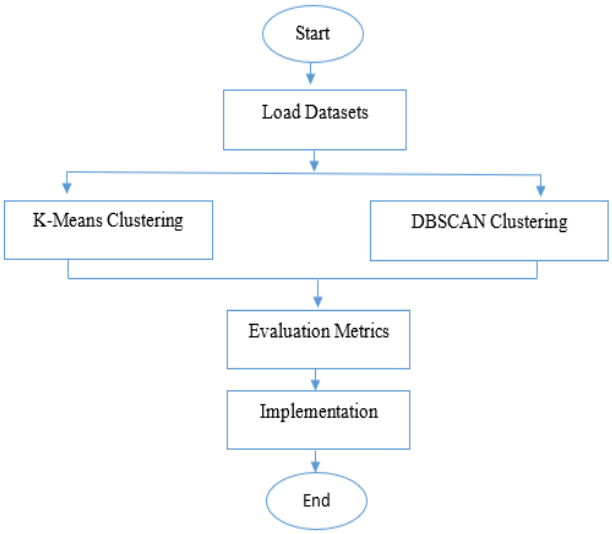

Fig. 1 Conceptual Framework

A. Input Datasets

The input dataset used in this study consists of the entrance examination records for student admission to the National University of Laos (NUOL) for the academic year 2014–2015. The dataset contains a total of 15,602 records, which have been pre-processed and organized for analysis [13]. The dataset includes information such as: (SID) Student Identification Number, (PH) Physics score, (LA) Lao Language and Literature score, (MA) Mathematics score, (GO) Geography-History score, (TOTAL) Total score.

B. Clustering Data Using K-Means

The steps of the K-Means algorithm are as follows, input: k (the predetermined number of clusters), D (the dataset). For the method 1. Calculate the initial centroids based on k selected from the dataset D. 2. Repeat the following steps until convergence: Compute the distance between each data point and the centroids using the Euclidean distance formula $dist(p,q) = \sqrt{\sum_{k=1}^{n}(p_k - q_k)^2}$ or $\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$ Assign each data point to the cluster with the nearest centroid and Recalculate the centroids based on the mean of all data points in each cluster. 3. Stop the iteration when the centroids no longer change significantly.

From these steps, it is clear that the choice of k significantly affects the performance of the K-Means algorithm. Therefore, selecting an appropriate k is crucial. The Elbow Method is widely used to determine the optimal k for K-Means. This method involves plotting the number of clusters against the sum of squared errors (SSE) and identifying the "elbow point," where the reduction in SSE starts to slow down. This point is considered the most suitable value of k for the clustering model. It shows in Fig 2

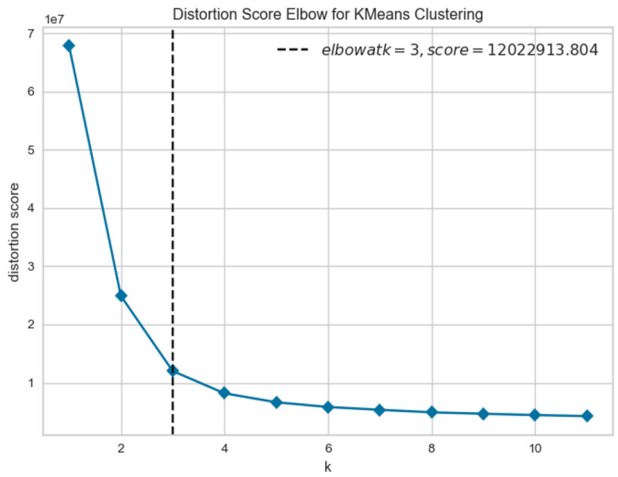

Fig. 2 Determining the Optimal k Using the Elbow Method

C. Clustering Data Using DBSCAN

The steps of the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm are as follows: 1. Find Neighbour points to Identify neighbouring points within the specified radius ε (Epsilon) and determine the core points, which are points having at least *MinPts* neighbours. 2. Create New Clusters: For each core point, if it has not been assigned to any cluster yet, create a new cluster. 3. Assign Neighbouring Points: Recursively add all points that are density-reachable from the core point to the same cluster. 4. Repeat: Continue the process for all points that have not yet been visited. Points that cannot be assigned to any cluster are considered noise points.

From these steps, it can be observed that the value of **ε** (Epsilon) plays a crucial role in the clustering process and greatly affects the algorithm's performance. Therefore, determining the most appropriate ε is essential. The K-distance graph method is commonly used to select the optimal ε value for DBSCAN. It shows in Fig 3.

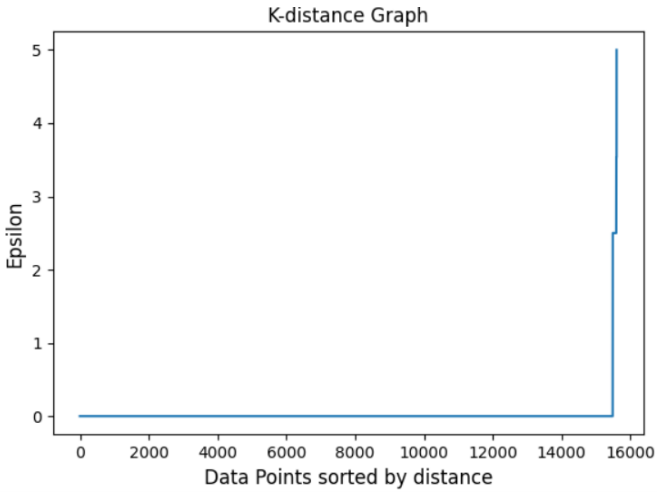

Fig. 3 illustrates the optimal value of ε (Epsilon) determined using the K-distance graph method.

D. Cluster Evaluation

The cluster evaluation, the Silhouette Coefficient method was applied, calculated using the formula: $Sw_i = \frac{bi-ai}{\max\{ai,bi\}}$ where:

- a_i= average distance between the data point I and all other points in the same cluster.
- b_i= lowest average distance between the data point iii and all points in any other cluster (nearest neighbouring cluster).

Based on the calculation using this formula, the Silhouette Coefficient results were: K-Means: 0.457 and DBSCAN: 0.484

III. RESULTS AND DISCUSSION

The results of the clustering experiments on the student selection dataset show that both **K-Means** and **DBSCAN** can organize the data into **three optimal clusters**, as evaluated using the **Silhouette Coefficient** method. The Silhouette Coefficient values were **0.457** for K-Means and **0.484** for DBSCAN. It shows in Fig 4(a) and 4(b)

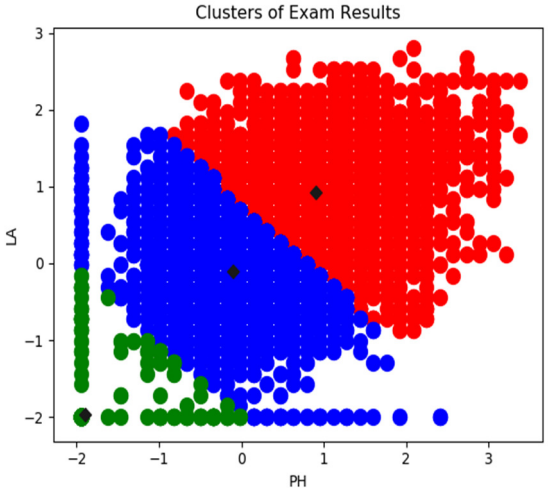

Fig. 4(a) Cluster of Exam Results

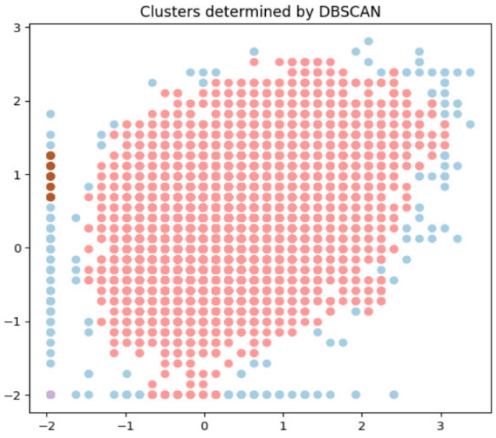

Fig. 4(b) Cluster determined by DBSCAN

IV. CONCLUSION

Through data analysis and model construction, the results can be summarized as follows: the value of *k* significantly affects the performance of the K-means algorithm. Therefore, using the Elbow method, $k = 3$ was determined to be the most suitable number of clusters. As a result, $k = 3$ was selected for the model.

For the DBSCAN model, it was observed that determining the value of Ɛ (Epsilon) is crucial and directly impacts the model's performance. The appropriate Ɛ value for this model was identified using the K-distance graph method, which suggested Ɛ = 0.3 as the optimal parameter for DBSCAN.

After comparing both models, the clustering results showed 3 groups, with the Silhouette Coefficient values being 0.457 for K-Means and 0.484 for DBSCAN. This indicates that DBSCAN performs better than K-Means. In the future, additional methods or diverse techniques may be applied for further comparison in order to achieve clearer clustering results.

V. FUTURE WORK

For future paper, several directions can be explored to further improve clustering performance and gain deeper insights from the dataset: Application of Additional Clustering Techniques Other clustering algorithms such as Hierarchical Clustering, Gaussian Mixture Models (GMM), or advanced density-based methods can be applied to compare performance with K-Means and DBSCAN. Optimization of Parameters More systematic parameter tuning methods, such as Grid Search, Bayesian Optimization, or automated approaches, can be used to identify optimal values of *k*, Ɛ, and MinPts to enhance clustering accuracy. High-Dimensional and Large-Scale Data The models can be extended and tested on higher-dimensional or large-scale datasets to evaluate scalability and robustness in more complex real-world applications.

ACKNOWLEDGMENT

The author would like to express sincere gratitude to the Faculty of Computer Science, University of Laos, for providing the opportunity, resources, and academic support throughout the course of this research.

Special thanks are extended to the professors, lecturers, and staff members whose guidance and teaching have laid the foundation for this work. The author also wishes to thank fellow classmates and project collaborators for their cooperation, encouragement, and shared knowledge during the development of this smart home system.

Appreciation is also given to family and friends for their unwavering support, patience, and motivation throughout the research and writing process.

REFERENCES

[1] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. San Francisco, CA, USA: Morgan Kaufmann, 2012.

[2] R. Baker and K. Yacef, "The state of educational data mining in 2009: A review and future visions," *Journal of Educational Data Mining*, vol. 1, no. 1, pp. 3–17, 2009.

[3] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–666, Jun. 2010.

[4] T. Kanungo et al., "An efficient k-means clustering algorithm: Analysis and implementation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 881–892, 2002.

[5] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. Springer, 2009.

[6] M. Ester, H. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining (KDD)*, Portland, OR, USA, 1996, pp. 226–231.

[7] M. Schubert et al., "DBSCAN revisited, revisited: Why and how you should (still) use DBSCAN," *ACM Transactions on Database Systems (TODS)*, vol. 42, no. 3, pp. 1–21, 2017.

[8] C. Romero and S. Ventura, "Educational data mining: A review of the state of the art," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 40, no. 6, pp. 601–618, 2010.

[9] [9] NUOL Entrance Examination Committee, "National University of Laos Entrance Examination Dataset 2014–2015," Internal Report, Vientiane, Laos, 2015.

[10] A. Peña-Ayala, "Educational data mining: A survey and a data mining-based analysis of recent works," *Expert Systems with Applications*, vol. 41, no. 4, pp. 1432–1462, 2014.

[11] P. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.

[12] W. McKinney, *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*, 2nd ed. O'Reilly Media, 2017.

