ANALYZING PERFORMANCE OF PLACEMENT STUDENTS RECORD USING DIFFERENT CLUSTERING ALGORITHM

Vasuki M

Research Scholar, Department of Computer Science and Technology,
Sathyabama Institute of Science and Technology,
Chennai, Tamil Nadu, India
dheshna@gmail.com

Dr S Revathy

Associate Professor, Department of Information Technology,
Sathyabama Institute of Science and Technology,
Chennai, Tamil Nadu, India
Ramesh.revathy@gmail.com

Abstract

In clustering problem analysis, Ensemble Cluster is proven to be a viable solution. Creating a cluster for such a comparable dataset and combining it into a separate grouping the clustering quality may be improved by using the combining clustering technique. Consensus clustering is another term for Ensemble clustering. Cluster Ensemble is a potential technique for clustering heterogeneous or multisource data. The findings of spectral ensemble clustering were utilized to reduce the algorithm's complexity. We now provide alternative clustering algorithms that have been applied to the same dataset and yielded diverse clustering outcomes. Because the many strategies were all described, it was easier to choose the most appropriate one to handle the situation at hand. To forecast the degree of student achievement in placement, clustering is created on the preprocessed information using clustering's specifically normalized k-means comparing with K-Medoids and Clarans algorithms.

***Keywords*: Consensus Clustering, K-Means, K-Medoids, Clarans.**

1. Introduction

Analyzing of Cluster is a fundamental approach for assessing variable knowledge in any field of investigation. When we use different clustering methods on the same dataset, including such k means, K-Medoids and Clarans, we obtain diverse results. The huge measure of information put away in the PC frameworks of an assortment of organizations, both public and private, has pushed the advancement of new advances for information examination and the board [1]. Information handling approaches have their starting points in setting, determined to reveal stowed away and non-insignificant connections among different sorts of information [2]. This assortment of strategies, which are utilized in an assortment of ventures as well as the scholarly climate, is gotten from standard data examination strategies and has the ability to deal with huge volumes of information [3].

For cluster result assessments, two sorts of approaches have been employed in clustering analysis. Cluster Validity Indexes (CVI) and Clustering Ensemble Algorithms are two types of clustering algorithms [4].

Cluster validity indexes can be used to determine how good clustering findings are. Another strategy for enhancing the effectiveness of both the clustering algorithm results is to combine the multiple clustering results and produce a single result. There are two fundamental stages in the cluster gathering approach [5]. They are (i) Generation and (ii) Consensus Function. For the same dataset, several clustering techniques are used to divide the data items into distinct categories. Every group is made up of the same things. The cluster ensemble method possesses the attribute of resilience, implying that the mixture procedure must outperform single clustering techniques as shown in fig 1.

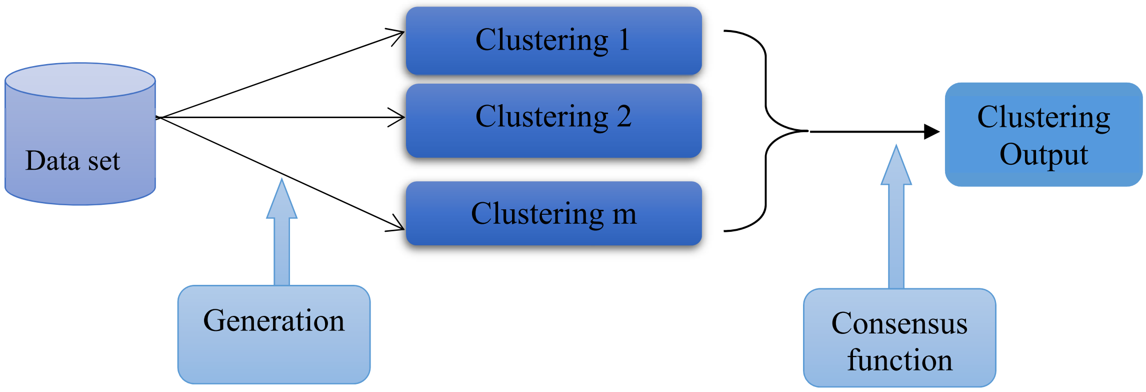

Fig. 1. Clustering Process

Student grading is a time-honored method of determining the value of a student's academic performance. Unfortunately, by not attempting to comprehend the variables that cause certain students to score greater than others in examinations, an entire underlying problem is left unsolved. In this research [6], we use final grades and a variety of socio-cultural and effort-oriented traits to try to uncover some key determinants in students' academic achievement. Different clustering methods might be utilized in the Generation Process as shown in fig 2.

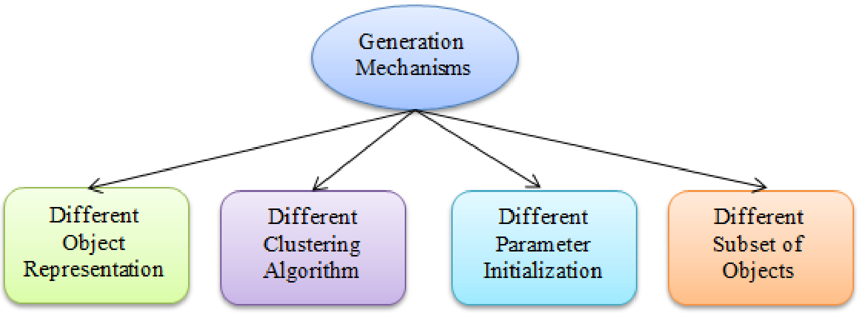
Generation Mechanisms
Different Object Representation
Different Clustering Algorithm
Different Parameter Initialization
Different Subset of Objects

Fig. 2. Types of Ensemble Generation Mechanisms

Academic data processing might be a contemporary study domain in the realm of education that examines and analyses data stored in student databases [7]. With the goal of resolving academic analysis challenges and improving the entire academic method, knowledge is studied utilizing applied math, machine learning, and data processing algorithms [8]. There has been a surge in the usage of academic coding instruments and databases collecting student data in recent years, resulting in massive data repositories that represents how students learn. Each field connects with pattern recognition in the process of understanding behavioral patterns in mechanical and also human elements of the project under observation, whether it's telecommunications, education, or politics.

Initialization sensitivity is a problem in k-means analysis [9]. Consensus clustering seeks to integrate many fundamental partitions into a single one. Consensus clustering provides dependable and high-quality results. The complexity of the k-means initiation is resolved through optimistic improvement of k-means consensus clustering. To accomplish quality clustering, Greedy and KCC were coupled. GKCC and spectral, as well as the objective function but also its standard deviations, are now available [10].

2. Related Studies

Cluster consensus function refers to integrating the results of several cluster algorithms' partitions into a single clustering. The number of clusters in the basic partitions might vary. Consensus clustering is simply a fusion problem that may be divided into two categories: (i) Utility Function and (ii) Co-Association Matrix.

The main classification makes a utility capacity that evaluates the relationship among essential parceling and the last segment, then, at that point, utilizes the utility capacity to settle a combinatorial improvement issue. The subsequent class utilizes an organizer network to decide how often a couple of events co-happen in similar groups, and afterward utilizes a diagram parcel way to deal with show up at a last agreement [11].

Educational Data Mining (EDM) is a relatively recent field of study. In the educational industry, mining techniques are utilized to obtain important information about staff or student progress habits. With the recent growth in the accessible of learning data, educational data mining has gained traction as a way of understanding and enhance teaching and learning and indeed the contexts in which it occurs [12]. For any firm, data is the most precious commodity. Extracting usable data from such a big and enormous volume of data is quite tough. Students' academic success is forecasted and evaluated using data mining techniques based on student record and forum participation. Despite the fact that various researches have been conducted throughout the world to evaluate student academic performance, there is a dearth of acceptable studies to examine aspects that might help students improve their academic performance. The goal of this study was to evaluate the factors that influence student academic achievement in Pakistan. Both basic and parallel clustering approaches are constructed and examined in this research to highlight their greatest qualities. The issues of simple algorithms are solved by parallel K-Mean algorithms, and the results of parallel algorithms are still the same, improving cluster quality, simulation time, and elapsed time. Both techniques are evaluated and compared on a dataset containing 10,000 as well as 5000 integer data elements, respectively. The datasets are assessed ten times for the shortest elapsed time, with K values ranging from 1 to 10. The proposed research is more beneficial for sifting scientific research data. Statistics from scientific studies are more accurate.

Understanding students' needs, presenting appropriate learning opportunities/resources, and improving teaching quality are all important research problems. Traditional machine learning algorithms, on the other hand, fail to deliver consistent and reliable prediction outcomes [13]. We present a bar chart ensemble machine learning approach in this research that intends to enhance the size of singular machine learning approaches by combining the results of many methods. To be more explicit, we combine supervised and unsupervised prediction approaches to create an iterative method that reproduces in a graph model and corresponds to more stable and reliable prediction outcomes. Extensive trials show that our suggested strategy is successful in forecasting more realistic student performance. In terms of prediction accuracy, our model surpasses the best standard machine learning algorithms with up to 14.8 percent.

Clustering is a common method for analyzing data and forming distinct comparable groupings. Fuzzy, rough & rough fuzzy clustering are the most extensively deployed robust soft clustering algorithms. The key characteristic of soft clustering causes the rough & fuzzy sets to be combined. The Rough Fuzzy C-Means (RFCM) technique incorporates rough set lower and boundary estimates, as well as fuzzy memberships of fuzzy sets into the c-means algorithm; nevertheless, the widely used RFCM requires more processing. To prevent this, the Fuzzy to Rough Fuzzy Link Element (FRFLE) is proposed in this study, which is employed as a key aspect in conceptualizing the rough fuzzy clustering as from fuzzy clustering outcome. Experiments with synthetic, standard, and other benchmark datasets demonstrate the FRFLE value automation process, followed by a comparison of the outcomes of ordinary RFCM versus RFCM employing FRFLE [14]. Furthermore, the results of the performance analysis reveal that the suggested RFCM method employing FRFLE takes less time to compute than existing RFCM techniques.

In all cluster applications, cluster validation is a must-have approach. Several validation approaches are used to assess cluster structure correctness. The most common approaches are geometric, with distance and membership as the only criteria for validation. This validity metric has been used in traditional rough set methods. The suggested index may be used with various clustering techniques, extending its use in corporate data mining [15].

Because clustering is unstructured learning because there are many clusters’ methods in practice, which resulting for which clustering scheme should be picked for our goal, we will look at the evolution and relevance of clustering approaches. We'll look at four different clustering methods: crisp, Juzzy, rough, and rough fuzzy. These clustering approaches have been developed, and their relative relevance has been discussed. For a better perspective, the best clustering approach for any of these three has indeed been selected. The outcomes of the experiment with the sampling dataset demonstrate the significance of clustering techniques.

3. Proposed Methodology

3.1. *Consensus of K-Means Clustering*

On both complete as well as inadequate straightforward parceling, K-means is receptive to introduction; Utility incorporates that capacity for KCC. Probes different genuine informational collections show that KCC is extremely effective and comparable to best-in-class techniques for grouping productivity; additionally, KCC shows solid flexibility with significant missing qualities for fragmented basic apportioning. Consensus Clustering (CC) is essentially an issue of combinatorial strengthening.

The current writing might be generally independent into two arrangements: CC having induced points (CCIO) versus CC with explicit objectives (CCEO). Techniques in CCIO don't characterize worldwide objective capacities. All things being equal, chart-based calculations, co-affiliation grid-based procedures, relabeling and casting a ballot strategies, transformative calculations, and different heuristics are especially worked to distinguish satisfactory responses in the authentic methodologies.

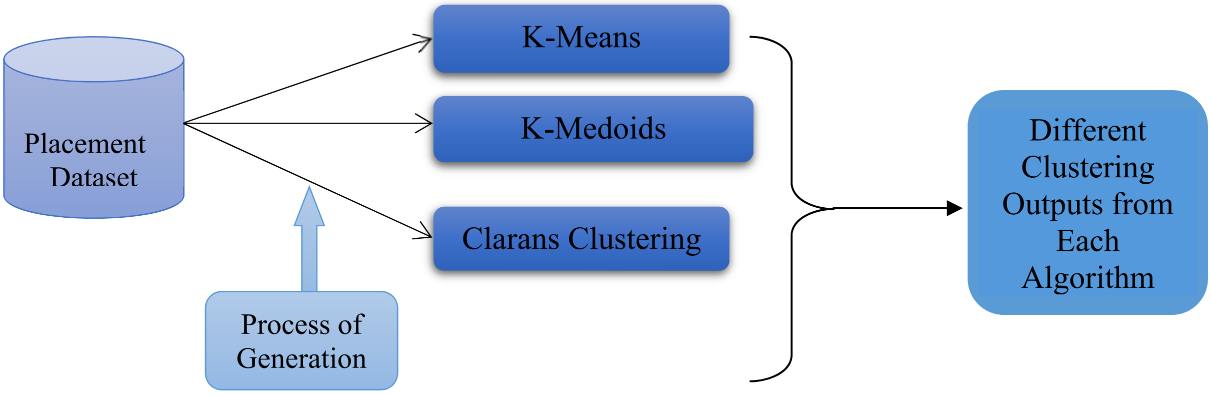

Fig. 3. Ensemble Generation employing several Process of algorithms

Other solutions for diverse goal functions are K-means utility functions, which constitute the basic Clustering framework. K-means is quite stable, even if there are only a few high-quality basic partitions or if the basic partitions are severely incomplete.

3.2. *Clustering using genetic k-means*

A clever clustering technique is by all accounts the hereditary qualities GKA that additionally consolidates the genetic algorithm with the k-means calculation. The durability and high efficiency of this fusion technique are achieved. As a consequence, GKA will continue to outperform all other genetic algorithms.

3.3. *Consensus Clustering function (CC)*

Consensus Clustering (CC) in a drawn out parcel work space. In an exceptionally proficient variation of K-means propelled by K-medoid, the K centres are introduced utilizing the first K-1 focuses, and the leftover one is insatiably scanned involving for introduction of K-means.

These mediator segments can be utilized for agreement combination. The whole methodology looks like voracious K-means. The centroids are utilized in each progression to voraciously look for an additional one place in the past stage, and afterward K-means are utilized to alter the current centroids. Into a result, the original data as well as basic partitions are blended as new data in order to produce following basic partitions. As a result, we provide a new fundamental partition generating technique that achieves an end-to-end ensemble clustering process by tightly coupling the following fusion. CC progressively adds new centres, overcoming the K mean's initialization sensitivity.

The benefits of CC are divided into three categories.
1. It consolidates greedy K-means with CC for an exact and high clustering.
2. The first information as well as essential segments is utilized to assemble future fundamental parts, making the consensus cluster only one methodology.
3. CC addresses the responsiveness issue with K-means in-statement and produces a steady, top-notch outcome.

Consensus Clustering Algorithm

Input: Placement Dataset
Output: Clustering Classification

```
begin
for n=1 to i
construct the matrix n Tree of size N * M
for each classifier
for each individuals x, y ϵ {1, 2, …. n}
Q(a,b) =Q(a,b)+$\frac{1}{classifier}$ ($s_a^u$=$s_b^u$)
$s_a^u$=$s_b^u$ displays a yield indication for each
s represents individuals
Dist (Distance) = $\sqrt{1 - nTree}$
$Q_j$ = CC (N Dist R(Dist), i)
end
```

The feature just at bottom from the set of rating will be removed first in the suggested approach. At each level of the feature elimination procedure, the classification accuracy will be verified. Validation is done to see if the concealed function is really redundant or just noisy. Validation may be calculated by evaluating accuracy rate before and after removing the function. If the classification accuracy and during the eradication of a feature is higher, the value will be kept; otherwise, it will be deleted. The iteration is complete when the correctness of the current subset's classification outweighs that of the preceding subset's classification.

4. Experimental Results

In the given results, visualized data from the given placement dataset based on their performance metrics and compared it with K-means, K-Medoid and Clarans algorithm as shown in below figures and tables.

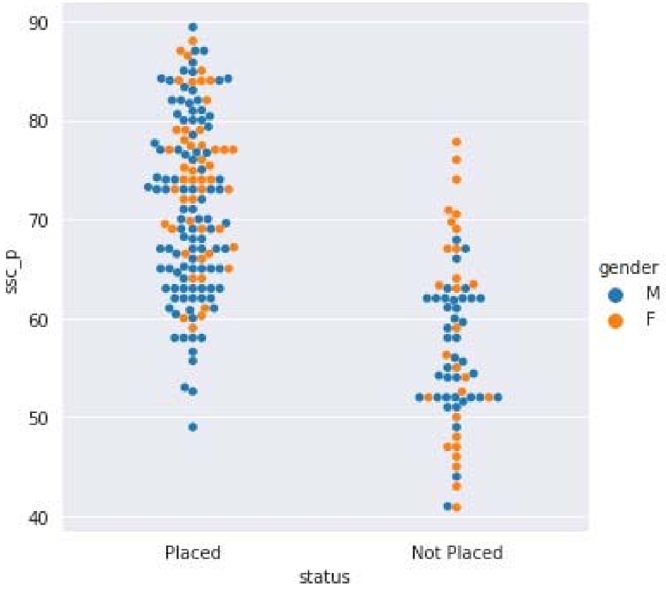

Fig. 4. Students Placed based on SSC

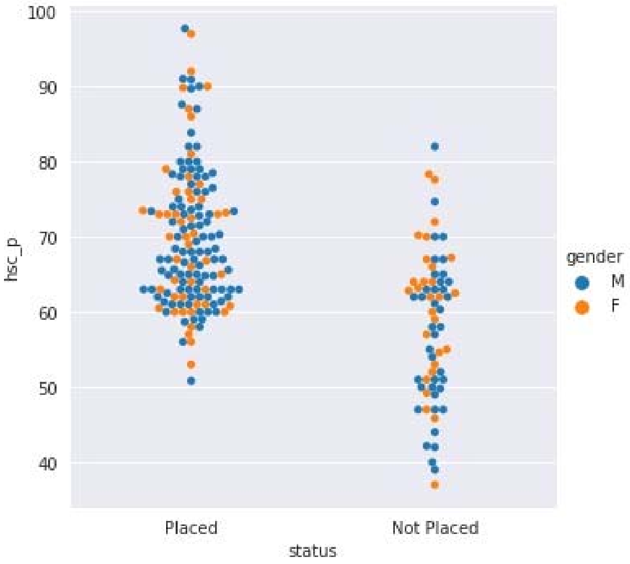

Fig. 5. Students Placed based on HSC

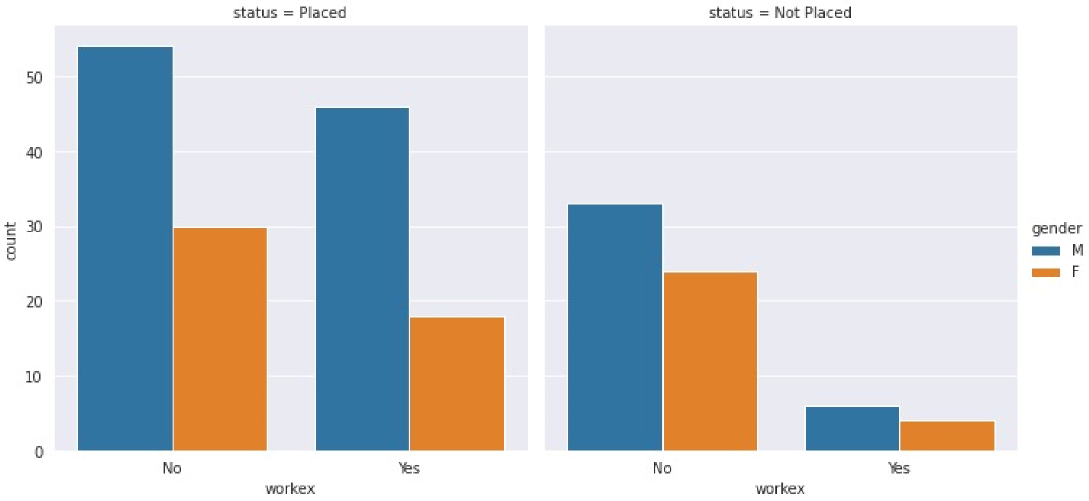

Fig. 6. Students placed counts

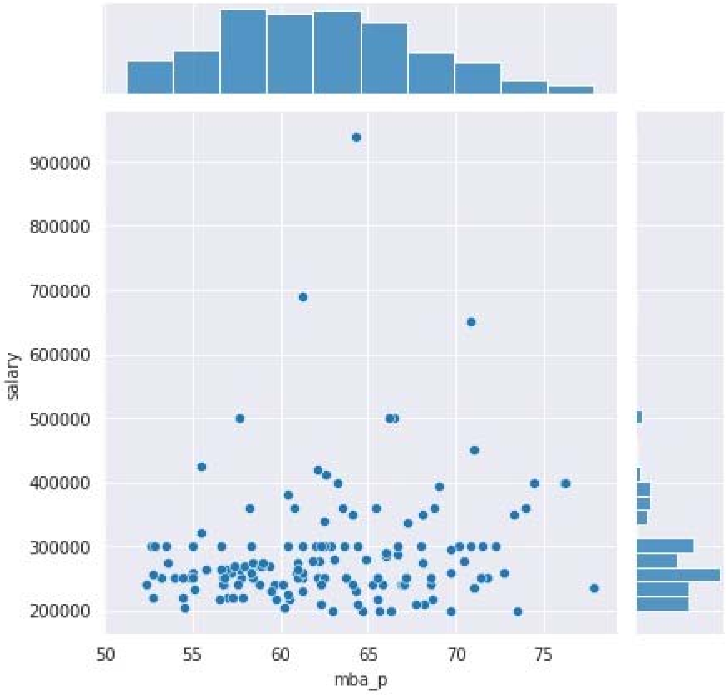

Fig. 7. MBA students placed vs Salary

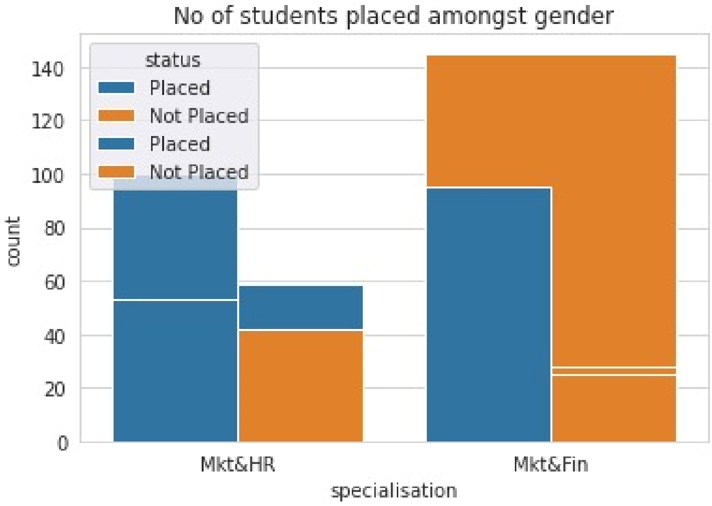

Fig. 8. Number of students placed in Marketing and Finance

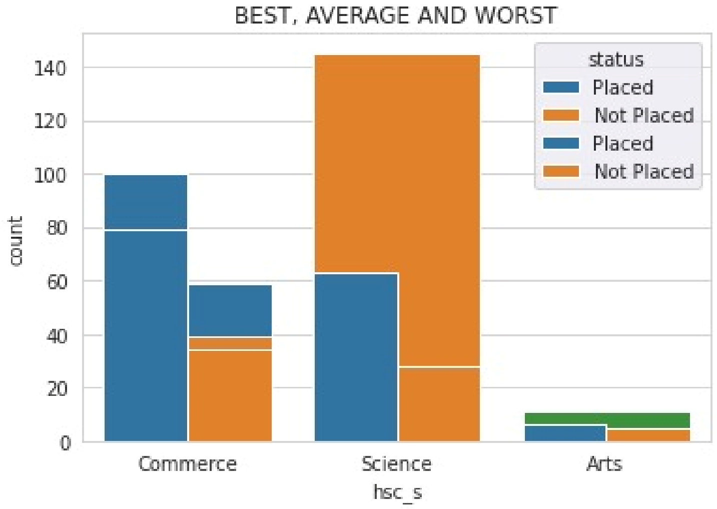

Fig. 9. Best, Average and Worst case of placed

Attributes	K-Means	K-medoids	Clarans
Mean distance of cluster (intra)	0.896	0.045	0.025
Mean distance of cluster (Inter)	0.79	1.55	3.25
No of iterations (to attain optimally)	17	24	35
Square error	2.83%	1.64%	0.86%

Table 1. Comparison for Different Algorithms

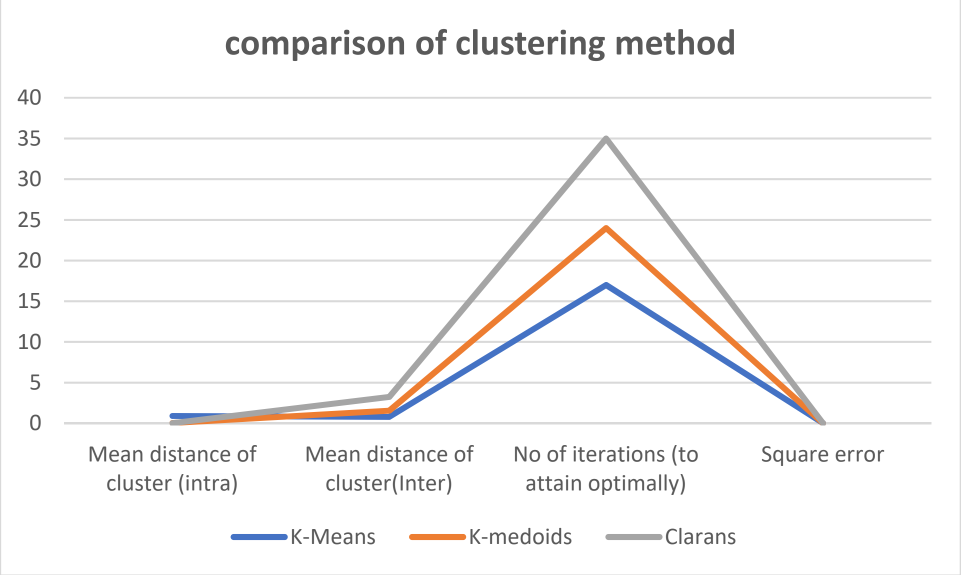

Fig. 10. Comparison Chart for Clustering Algorithms

No. of clusters	k-means time complexity	k-medoid time complexity	Clarans
1	1000	1000	2000
2	5000	3000	6000
3	15000	9000	20000
4	30000	15000	35000

Table 2. Different number of clusters for Time complexity

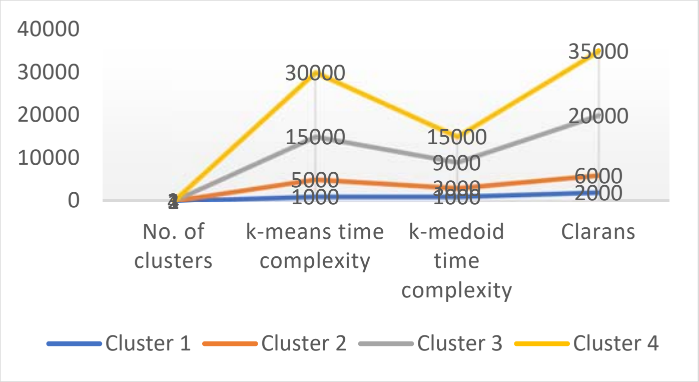

Fig. 11. Comparison Chart for Time Complexity in Clustering Algorithms

No. of iterations	k-means time complexity	k-medoid time complexity	Clarans
2	2000	1000	500
4	4000	3000	2000
6	6000	4000	3000
8	8000	5000	2000

Table 3. Number of iterations for space complexity

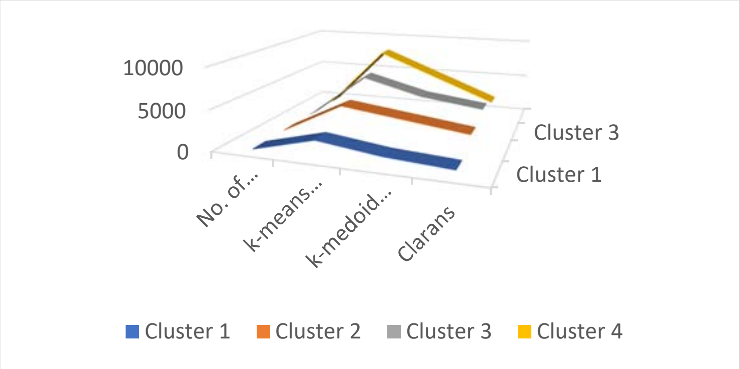

Fig. 12. Comparison Chart for Space Complexity in Clustering Algorithms

No. of cluster	k-means time complexity	k-medoid time complexity	Clarans
5	100	600	700
10	300	700	800
15	500	800	900
20	900	900	1000

Table 4. Different number of clusters for space complexity

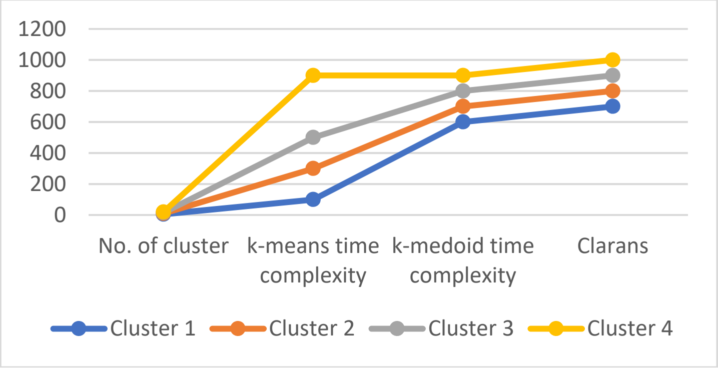

Fig. 13: Comparison Chart for Clustering Algorithms using Space complexity

5. Conclusion and Future Enhancement

In this paper, the K-means algorithm is utilized to segment a position dataset into clusters, characterizing the degree of execution in different delicate abilities including such fitness, English, programming rationale, and coding abilities. The cluster execution is resolved utilizing R-Tool and standardized utilizing the mean/standard deviation equation. For the situation dataset, Rn and NMI were processed as outside estimations. For the equivalent dataset, a few K-Means, K-Medoids, and Clarans algorithms were utilized. Afterward, the registered bunch execution might be thought about, and future investigations can anticipate which approach creates astounding clusters.

References

[1] Fang-Xiang Wu, (2008), ”Genetic weighted k-means algorithm for clustering large-scale gene expression data” BMC Bioinformatics, 9(Suppl 6):S12, DOI https://doi.org/10.1186/1471-2105-9-S6-S12

[2] Diyar Qader Zeebaree, Habibollah Haron, Adnan Mohsin Abdulazeez and Subhi R. M (2017)” Combination of K-means clustering with Genetic Algorithm: A review”. Zeebaree International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 24 pp. 14238-14245 © Research India Publications. http://www.ripublication.com

[3] Dataram Soni Madhulatha D.C. Wyld et al (2011), ” Comparison between K-Means and K-Medoids Clustering Algorithms “. (Eds.): ACITY 2011, CCIS 198, pp. 472–481. © Springer-Verlag Berlin Heidelberg 2011

[4] P. IndiraPriya, Dr. D.K. Ghosh (2013), “A Survey on Different Clustering Algorithms in Data Mining Technique”, International Journal of Modern Engineering Research (IJMER) Vol.3, Issue.1, pp-267-274 ISSN: 2249-6645.

[5] Jyotismita Goswami (2015), “A Comparative Study on Clustering and Classification Algorithms” International Journal of Scientific Engineering and Applied Science (IJSEAS) - Volume-1, Issue-3, ISSN: 2395-3470.

[6] Pooja Kumari; Praphula Kumar Jain; Rajendra Pamula (2018), ” An efficient use of ensemble methods to predict students’ academic performance” 4th International Conference on Recent Advances in Information Technology (RAIT)

[7] M. Shovon and M. Haque (2012), An approach of Improving Student Academic Performance by using K-means clustering Algorithm and Decision tree vol. 3 pp. 8.

[8] L. Juhanak et al. (2017), "Using process mining to analyze students' quiz-taking behavior patterns in a learning management system" in Computers in Human Behavior.

[9] Alejandro Pena-Ayala (2014), "Educational data mining: A survey and a data mining-based analysis of recent works" in Expert Systems with Applications pp. 1432-1462.

[10] Hashmia Hamsa Simi Indiradevi and Jubilant J. Kizhakkethottam (2016), "Student academic performance Prediction Model Using Decision Tree and Fuzzy genetic algorithm" Procedia Technology vol. 25 pp. 326-332.

[11] Shang, R., Ara, B., Zada, I., Nazir, S., Ullah, Z., & Khan, S. U. (2021). Analysis of simple K-mean and parallel K-mean clustering for software products and organizational performance using education sector dataset. Scientific Programming.

[12] Wang, Y., Ding, A., Guan, K., Wu, S., & Du, Y. (2021). Graph-based Ensemble Machine Learning for Student Performance Prediction. arXiv preprint arXiv:2112.07893.

[13] Subramanion, R., Balasubramanian, P., & Noordeen, S. (2017). Enforcement of Rough Fuzzy Clustering Based on Correlation Analysis. International Arab Journal of Information Technology (IAJIT), 14(1).

[14] Revathy, S., Parvathavarthini, B., & Caroline, S. S. (2016). Decision Theory, an Unprecedented Validation Scheme for Rough-Fuzzy Clustering. International Journal on Artificial Intelligence Tools, 25(02), 1650003.

[15] Revathy, S., & Parvathavarthini, B. (2011). Integrating rough clustering with fuzzy sets.

Authors Profile

Mrs. M Vasuki Research Scholar from the Department of computer science and Technology of Sathyabama Institute of Science and Technology, Chennai India. Her research interest includes Data Mining, Machine Learning, and Big Data. She has published many publications in refereed journals

Dr S Revathy is an Associate Professor from the Department of Information Technology working in Sathyabama Institute of Science and Technology, Chennai India. Her research interest includes Data Mining, Machine Learning, Data Analytics and Big Data. She has published more than thirty publications in refereed journals. She is the reviewer of many refereed journals.