Prediction of Students Academic Performance Using K-Means and K-Medoids Unsupervised Machine Learning Clustering Technique

[1]Prerna Joshi, [2]Pritesh Jain

[1]M.Tech Scholar, [2]Assistant Professor
Patel College of Science and Technology, Indore

***Abstract*— the capacity to screen the advancement for students' academic execution is a basic issue of the academic Group from claiming higher Taking in. An arrangement to dissecting students' comes about dependent upon group dissection and utilization standard measurable calculations with organize their scores information as stated by the level about their execution may be portrayed. In this paper, we also actualized k-mean and K-Medoids grouping algorithm for examining students' consequence information. Those model might have been consolidated for those deterministic model should dissect those students' effects of a private foundation clinched alongside % Iberia which is a great standard with screen the progression of academic execution about people for higher institutional to the reason for making an successful choice by those academic organizers.**

***Keywords*- K – Mean, K-Medoids, Clustering, Academic Performance, Algorithm.**

I. PRESENTATION

Graded perspective Normal (GPA) may be a regularly utilized pointer about academic execution. A significant number Europe, hypothetical orders had more distinction than difficult work, and speculative chemistry was situated at least GPA that if a chance to be looked after so as will keep in the degree program. Over a few Universities, those least GPA prerequisite set for the understudies may be 1. 5. Nonetheless, for whatever graduate program, A GPA from claiming 3. 0 Also over may be recognized a pointer about handy academic execution. Therefore, GPA even now stays those the vast majority regular element utilized toward those academic organizers with assess progression to a academic earth [1]. A number variables Might go about as obstructions with people achieving Also administering An helter smelter GPA that reflects their Generally speaking academic execution Throughout their residency clinched alongside college. These variables Might a chance to be focused on toward the teaching staff to Creating methodologies on move forward learner taking in also move forward their academic execution Eventually Tom's perusing method for checking those progression from claiming their execution.

Therefore, Execution assessment is a standout amongst the bases on screen the progression of scholar execution over higher institutional for Taking in. Build once this incredulous issue, grouping for understudies under diverse classifications as stated by their execution need turn into a confounded errand. With universal grouping about understudies In light of their Normal scores, it is was troublesome to get a thorough perspective of the state of the students' execution Also all the while find imperative subtle elements starting with their the long haul to the long run execution.

With that assistance about information mining methods, for example, grouping algorithm, it will be workable to find those way qualities starting with the students' execution also conceivably utilization the individual's aspects for future prediction. There bring been A percentage guaranteeing comes about from applying k-means grouping calculation with the Euclidean separation measure, the place the separation will be registered Eventually Tom's perusing finding the square of the separation the middle of each scores, summing the squares and finding the square root of the entirety of cash [6].

This paper displays k-means grouping algorithm Concerning illustration a straightforward What's more proficient device around will screen those progression of students' execution clinched alongside higher foundation.

Bunch examination Might a chance to be isolated under hierarchic grouping What's more non-hierarchical grouping systems. Samples from claiming hierarchic strategies would single linkage, complete linkage, Normal linkage, median, What's more Ward. Non-hierarchical systems incorporate k-means, versatile k-means, k-medics, also fluffy grouping. Will determine which calculation is handy is An work of the sort of information accessible and the specific reason for examination. In that's only the tip of the iceberg destination way, those solidness from claiming groups camwood a chance to be investigated in reenactment investigations [4]. The issue from claiming selecting the "best" algorithm/parameter setting will be a troublesome particular case. A great grouping calculation ideally ought to transform Assemblies for different non-overlapping boundaries, in spite of the fact that a flawless detachment camwood not normally make attained done act. Figure for value measures (indices) for example, those profile width [4] alternately those homogeneity list [5] could a chance to be used to assess the caliber about detachment got utilizing An grouping calculation. The idea for solidness of a grouping calculation might have been viewed as over [3]. Those thought behind this acceptance methodology will be that an calculation ought further bolstering make rewarded to consistency. In this paper, we executed conventional intends grouping algorithm [6] and Euclidean separation measure of comparability might have been decided to be utilized within those Investigation of the students' scores.

II. METHODOLOGY

A. Development of k-mean clustering algorithm Given a dataset of n data points x1, x2, …, an such that each data point is in Rd , the problem of finding the minimum variance clustering of the dataset into k clusters is that of finding k points $\{m_j\}$ (j=1 , 2........ K) In R^d such that

$$\frac{1}{N}\sum_{i=1}^{n}[\min_j d^2\left(x_i, m_j\right)] \qquad (1)$$

is minimized, where d $\left(x_i, m_j\right)$ denotes the Euclidean distance between x_i and m_j The points $\{m_j\}$ (j=1 , 2........ K) Are known as cluster cancroids'. The problem in Eq.(1) is to find k cluster cancroids', such that the average squared Euclidean distance (mean squared error, MSE) between a data point and its nearest cluster cancroids is minimized.

Those k-means calculation gives a simple strategy should execute estimated answer for eq. (1). The purposes behind the Notoriety from claiming k-means are simplicity Also Straightforwardness for implementation, scalability, pace from claiming merging What's more versatility to meager information.

Those k-means algorithm camwood be considered perfect Likewise An gradient plummet procedure, which starts during beginning bunch cancroids', and iteratively updates these cancroids' should diminishing the destination capacity Previously, eq. (1). The k-means dependably meet with An neighborhood base. Those specific nearby least found relies on the beginning group cancroids'. The issue for finding the worldwide least will be NP-complete. Those k-means calculation updates group cancroids' till neighborhood least is found. Fig. 1 indicates the summed up pseudo codes of k-means algorithm; and customary k-means calculation is introduced clinched alongside fig. 2 separately.

Preceding the k-means calculation converges, separation What's more cancroids calculations need aid carried out same time loops need aid executed An amount about times, say l, the place the certain basic l may be known as the amount from claiming k-

Means iterations. The exact esteem for l differs relying upon the starting beginning bunch cancroids' much on the same dataset. With the goal those computational period intricacy of the calculation is O(nkl), the place n will be those aggregate number about Questions in the dataset, k will be those required amount of groups we recognized and l will be the amount from claiming iterations, k≤n, l≤n [6].

```
Step 1: Accept the number of clusters to group data into and the dataset to cluster as input values
Step 2: Initialize the first K clusters
- Take first k instances or
- Take Random sampling of k elements
Step 3: Calculate the arithmetic means of each cluster formed in the dataset.
Step 4: K-means assigns each record in the dataset to only one of the initial clusters
- Each record is assigned to the nearest cluster using a measure of distance (e.g Euclidean distance).
Step 5: K-means re-assigns each record in the dataset to the most similar cluster and re-calculates the arithmetic
mean of all the clusters in the dataset.
```

Fig 1: Generalized Pseudo Code of Traditional K-Means

```
MSE = largenumber;
Select initial cluster centroids {mj}j K = 1;
Do
    OldMSE = MSE;
    MSE1 = 0;
    For j = 1 to k
        mj = 0; nj = 0;
    end for
    For i = 1 to n
        For j = 1 to k
            Compute squared Euclidean distance d2(xi, mj);
        end for
        Find the closest centroid mj to xi;
        mj = mj + xi; nj = nj+1;
        MSE1=MSE1+ d2(xi, mj);
    end for
    For j = 1 to k
```

```
nj = max(nj, 1); mj = mj/nj;
end for
MSE=MSE1;while (MSE<OldMSE)
```

Fig.2: Traditional K-Means Algorithm

2.1. K-Medoids

The K-Medoids calculation is utilized to discover Medoids in a bunch which is focus found purpose of a group. K Medoids is more vigorous when contrasted with K-Means as in K-Medoids we discover k as delegate question limit The total of dissimilarities of information objects while, K-Means utilized aggregate of squared Euclidean separations for information Objects. What's more, this separation metric diminishes clamor and anomalies.

Disadvantages of K-Means [1] calculation:

1) To discover K-Value is troublesome assignment.

2) It isn't viable when utilized with worldwide bunch.

3) If distinctive beginning allotments has been chosen than it might differ the outcome for groups.

4) Different size and diverse thickness group isn't dealt with by the calculation.

We utilized K-Medoids calculation that depends on protest delegate systems [4] to lessen the downsides of KMeans

Calculation: Medoids is the information protest of group which is most midway found. Medoids are chosen

Haphazardly from the Ky information items to shape Ky bunch and other residual information objects are put close to Medoids in a Cluster. Than process all information objects of group to discover new Medoids in rehashed design to speak to new bunch in Better way. Subsequent to finding the new Medoids tie every one of the information articles to the bunch. Area of Medoids change in like manner with every cycle. So key groups are shaped speaking to n information objects [3].

```
Input: Ky: the quantity of bunches, Dy: a Data index containing n objects.

Output: An arrangement of ky groups.

Algorithm:

x Randomly select ky as the Medoids for n information focuses.

x Find the nearest Medoids by figuring the separation between information focuses n and Medoids k and guide information questions that.

x For each Medoids m and every datum point o related to m do the accompanying:

Swap m and o to figure the aggregate cost of the arrangement than

Select the Medoids o with the most reduced cost of the arrangement.

x If there is no adjustment in the assignments rehash stages 2 and 3 on the other hand
```

Fig 3: Generalized Pseudo Code of Traditional K-Medoids

III. OUTCOMES

3.1 Data Understanding

The aim of this study applied clustering method to determine homogeneous groups in students' based on student performance. To discover the characteristics of student performance for various groups of students. A number of factors that are considered to have influence on the performance of a student were identified. These influencing factors were categorized as input variables. The variables used in this study are Gender, Nationality, State, School owner, Age. Table 1 shows the description of each variable used for clustering analysis.

TABLE 3.1.: Dataset Description

Attributes	Attribute-Details	Class-Values
Gender	Student gender	Male /Female
Nationality	Student nationality	Omani / Non Omani
State	Student Origin	Urban/Rural
School Owner	School Types	Public/Private
Age	Student Age	Group 1 (17 -18 years old) Group 2 (19 – 21 years old) Group 3 (22 years old and above)
Result	Final Student Result	Percent Subject wise

3.2 Data Preparation

During this phase, we applied some pre-processing for the collected data to prepare it for the mining techniques. At first, we eliminated some irrelevant attributes, e.g. student name, nationality, and campus. Then, we re-arranged the student data so that each student has the following attributes: Student ID, Gender, Age, Nationality, State, school ownership, final result. In the final step, we discredited the numerical attributes to categorical ones. For example, we grouped the final result into six groups: excellent, very good, good, fair, poor and very poor. In the same way, we discredited the students' state into two groups: urban and rural. Also, we discredited the students' age into three groups: (17- 18), (19- 21) and (22 above). The Oman education data are extracted and loaded for initial data understanding and summarization. The data reveals certain characteristics of Oman education data. Clustering represents a collection of records that are similar to another, and dissimilar to records in other clusters. Association in a cluster is determined based on some measure of distance between cases.

We connected the model on the information situated (academic effect about one semester) of a school to Nigeria. Those result created will be demonstrated to tables 2, 3, 4, and 5, individually. In table 2, for k = 3; to group 1, those group measure will be 25 and the general execution is 62. 22. Also, the group sizes and the in general exhibitions to group numbers 2 Also 3 need aid 15, 29 and 45. 73 and 53. 03, consciously. Comparative analyses also hold for tables 3 what's more 4. The graphs would create on figures 3, 4 what are more 5, respectively, the place the generally speaking execution is plotted against the bunch size.

Table 5 indicates the measurement of the information set (Student's scores) in the structure n by m matrices, the place n is the rows (# from claiming students) Also m will be those section (# about courses) advertised by each learner.

The in general execution is assessed toward applying deterministic model on eq. 2 [7] the place the aggregation evaluation to every of the bunch size is assessed toward summing the Normal of the distinct scores clinched alongside every bunch.

$$\frac{1}{N}\left(\sum_{i=1}^{N}\left(\frac{1}{N}\sum_{i=1}^{N} x_i\right)\right) \qquad (2)$$

Where

N = the total number of students in a cluster and

n = the dimension of the data

Table 3.2: Performance Index

70 and above	Excellent
60-69	Very Good
50-59	Good
45-49	Very Fair
40-45	Fair
Below 45	Poor

3.3 Result

Over figure 3, the in general execution to bunch extent 32 may be 62. 22% same time the general execution for bunch extent 6 will be 45. 73% also bunch measure 12 need those in general execution for 53. 03%. This investigation demonstrated that, 32out about 79 people needed a "Very Good" execution (62. 22%), same time 6 out about 79 scholars needed execution in the locale about extremely “Fair” execution (45. 73%) and the remaining 12 people required An “Good” execution (53. 03%) Concerning illustration delineated in the execution list clinched alongside table 1.

Figure 4 indicates the patterns for execution examination as follows; general execution to group size 24 is 50. 08% same time the generally execution for group span 6 may be 65. 00%. Group extent 12 needs those in general execution of 58. 89%, same time group extent 8 is 43. 65%. Those patterns in this dissection shown that, 24 people fall in those area from claiming “Good” execution list On table 1 over (50. 08%), same time 6 scholars need execution in the locale for “Very Good” execution (65. 00%). 12 scholars need An “Good” execution (58. 89%) Also 8 people required execution for “Fair” effect (43. 65%).

For figure 5, those in general execution for bunch span 22 will be 49. 85%, same time those in general execution to bunch size 7 may be 60. 97%. Bunch span 12 needs that general execution of 43. 65%, same time those bunch extent 7 need generally execution of 64. 93% what’s more group measure 2 need generally execution about 55. 79%. This execution examination shown that, 22 people crossed through with “Good” execution locale (49. 85%), same time 7 understudies needed “Very Good” execution comes about (60. 97%). 12 learners fall in the locale from claiming “Fair” execution list (43. 65%), 7 understudies were in the district for “Very Good” execution (64. 93%) and the remaining 2 understudies required “Good” execution (55. 79%).

Table 3.3: K-Means = 3

Sub1	Sub2	Sub3	Sub4	Sub5	K_means_cluster_labels	points belonging to cluster
61.34375	59.46875	63.125	55.1875	47.15625	0	32
91.83333	86.83333	91.33333	92	92.5	1	6
18.5	26.75	22.25	24.66667	36.08333	2	12

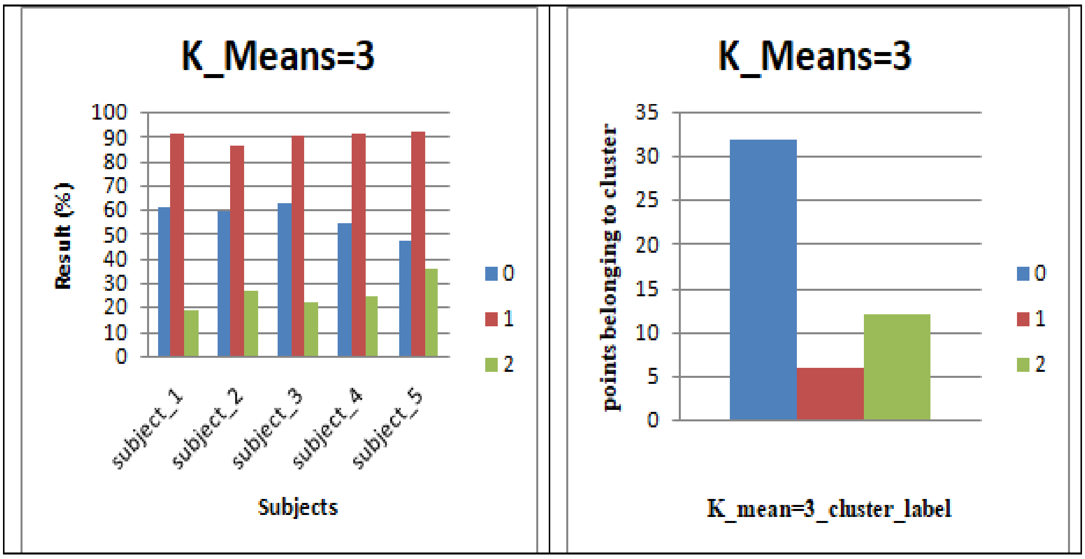


Fig. 3: Overall Performance Versus Cluster Size (# of Students) K-Means = 3

Table 3.4: K-Means = 4

sub1	sub2	sub3	sub4	sub5	K_means_cluster_labels	points belonging to cluster
72.5	62.5	75.5	36	24	0	24
91.833333	86.833333	91.333333	92	92.5	1	6
57.625	58.458333	59	61.583333	54.875	2	12
18.5	26.75	22.25	24.666667	36.083333	3	8

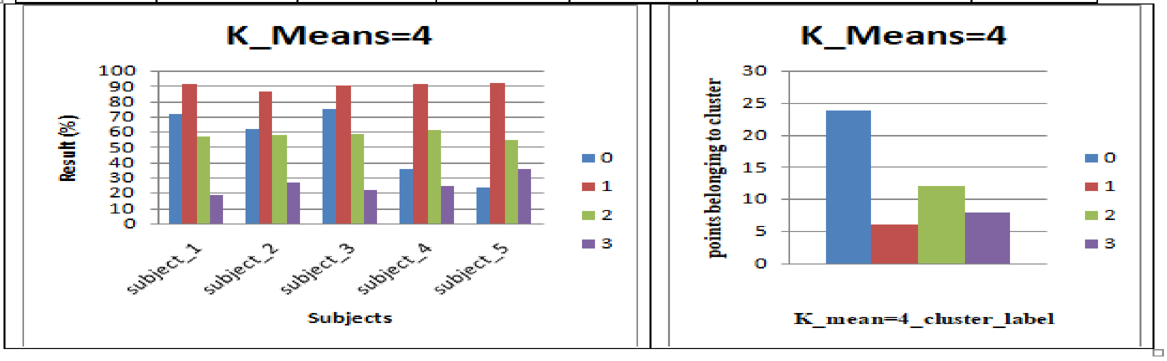


Fig. 4: Overall Performance Versus Cluster Size (# of Students) K-Means = 4

Table 3.5: K-Means = 5

sub1	sub2	sub3	sub4	sub5	K_means_cluster_labels	points belonging to cluster
88	94	96	25	31	0	22
90.285714	84.285714	90.857143	86.571429	89.571429	1	7
18.5	26.75	22.25	24.666667	36.083333	2	12
56	57.818182	57.227273	61.636364	55.181818	3	7
67.714286	53.428571	68.714286	43.714286	23	4	2

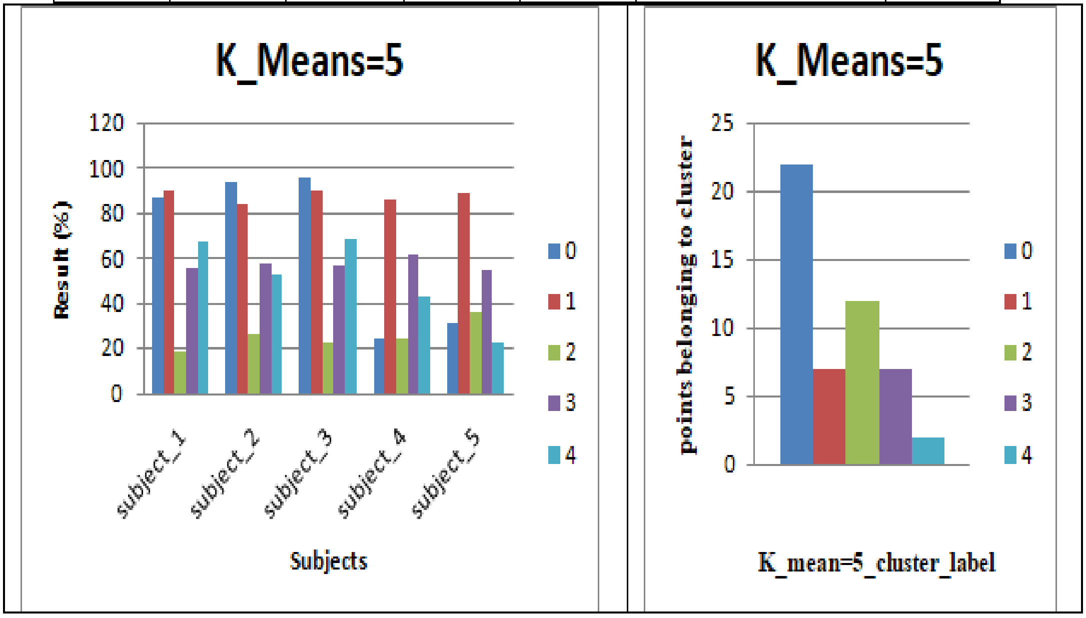
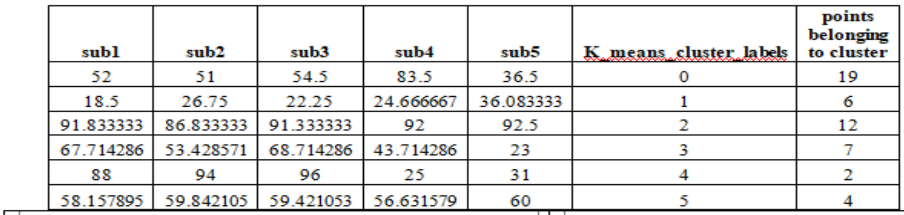


Fig. 5: Overall Performance Versus Cluster Size (# of Students) K-Means = 5

Table 3.6: K-Means = 6



sub1	sub2	sub3	sub4	sub5	K_means_cluster_labels	points belonging to cluster
52	51	54.5	83.5	36.5	0	19
18.5	26.75	22.25	24.666667	36.083333	1	6
91.833333	86.833333	91.333333	92	92.5	2	12
67.714286	53.428571	68.714286	43.714286	23	3	7
88	94	96	25	31	4	2
58.157895	59.842105	59.421053	56.631579	60	5	4

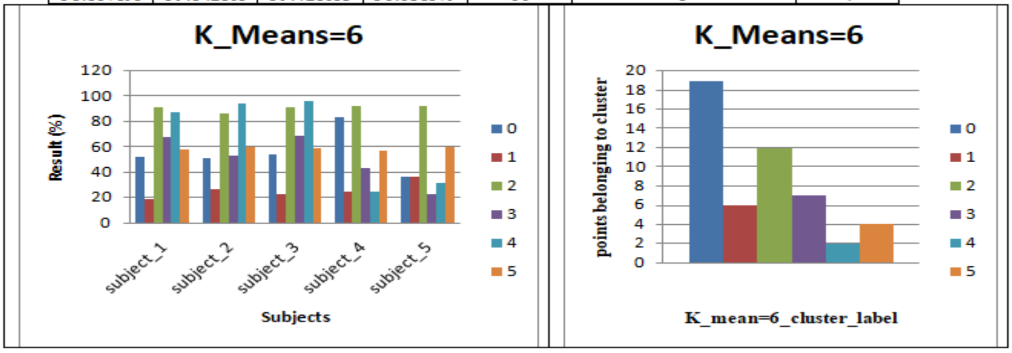


Fig. 6: Overall Performance Versus Cluster Size (# of Students) K-Means = 6

Table 3.7: K-Medoids = 3

subject_ 1	subject_ 2	subject_ 3	subject_ 4	subject_ 5	K_medoids=3_cluster_label	points belonging to cluster
55	59	51	58	51	0	32
16	27	40	23	41	1	6
96	88	98	94	99	2	12

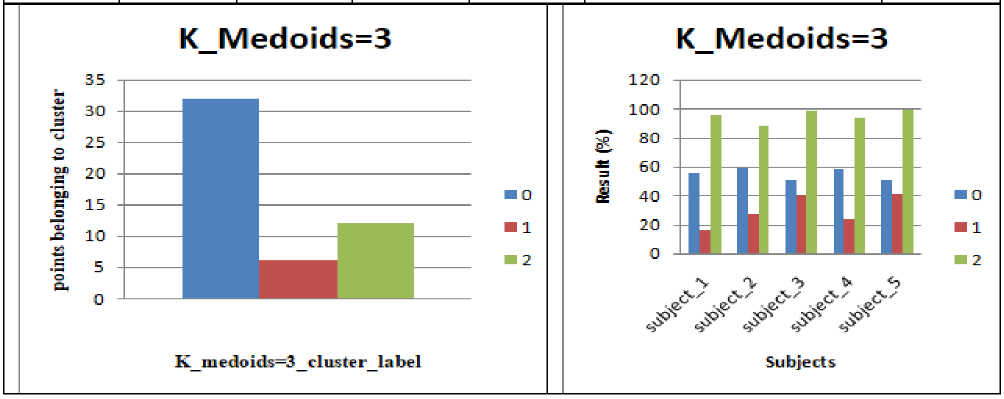


Fig. 7: Overall Performance Versus Cluster Size (# of Students) K-Medoids = 3

Table 3.8: K-Medoids = 5

subject_1	subject_2	subject_3	subject_4	subject_5	K_medoids_cluster_labels	points belonging to cluster
16	27	40	23	41	0	22
81	54	72	35	32	1	6
96	88	98	94	99	2	10
51	58	51	55	58	3	12

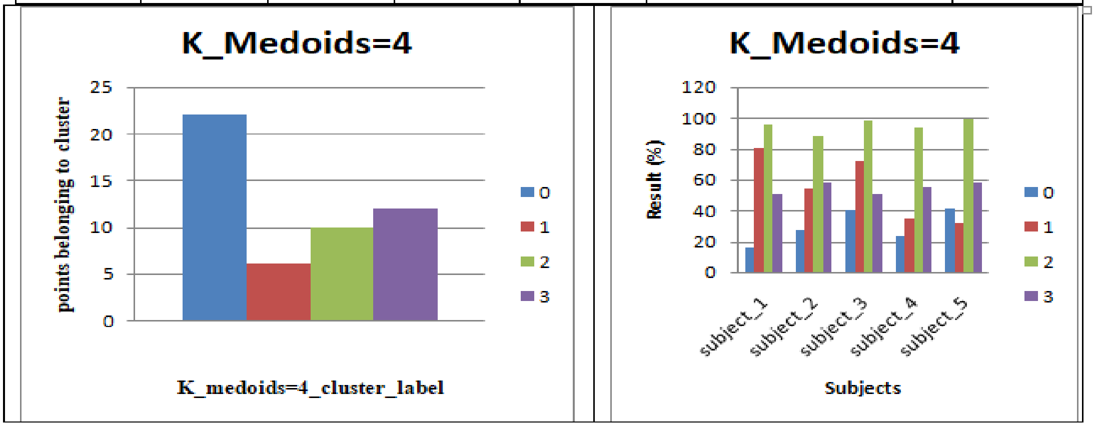


Fig. 8: Overall Performance Versus Cluster Size (# of Students) K-Medoids = 4

Table 3.9: K-Medoids = 4

subject_1	subject_2	subject_3	subject_4	subject_5	K_medoids_cluster_labels	points belonging to cluster
81	54	72	35	32	0	12
23	48	22	15	34	1	13
51	58	51	55	58	2	5
64	69	58	72	62	3	12
96	88	98	94	99	4	8

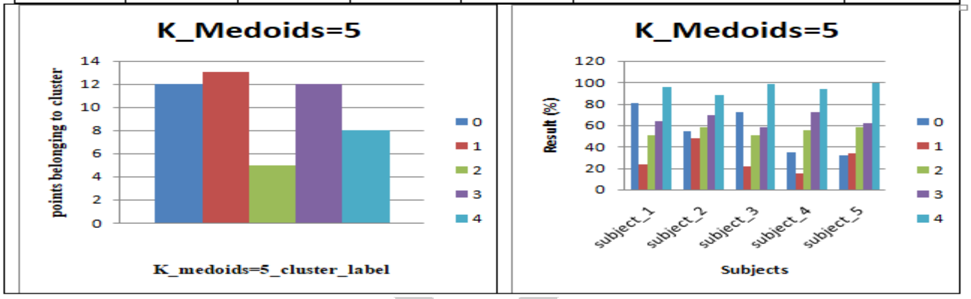


Fig. 9: Overall Performance Versus Cluster Size (# of Students) K-Medoids = 5

Table 3.10: K-Medoids = 6

subject_1	subject_2	subject_3	subject_4	subject_5	K_medoids_cluster_labels	points belonging to cluster
55	59	51	58	51	0	15
81	54	72	35	32	1	10
96	88	98	94	99	2	5
22	24	35	11	30	3	12
64	69	58	72	62	4	6
88	94	96	25	31	5	2

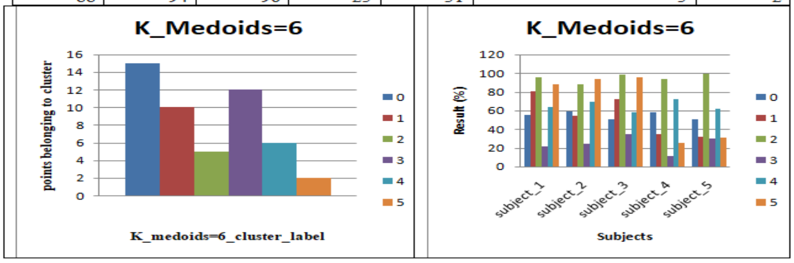


Fig. 9: Overall Performance Versus Cluster Size (# of Students) K-Medoids = 6

3.4 Compression Study Existing Work Verses Proposed Work

Table 3.11 Compression Study Existing Work Verses Proposed Work

Parameters	Existing Work	Proposed Work	
	Decision Trees	K-Mean	K-Medoids
All Subject Percent	Above 68%	Above 71 %	Above 73%

IV. CONCLUSION

In this work, we given a basic Furthermore qualitative technique on analyze those predictive force of grouping algorithm and the Euclidean separation concerning illustration and measure of comparability separation. We showed our technique utilizing methods grouping calculation [6] Also consolidated for those deterministic model in [7] on a information situated from claiming private school outcomes with nine courses advertised for that semester to each scholar for downright amount about 79 students, What's more produces those numerical understanding of the outcomes to the Execution assessment. This model progressed around a few of the impediments of the existing methods, for example, model created Eventually Tom's perusing [7] Furthermore [8]. These models connected fluffy model to foresee students' academic execution on two dataset best (English dialect what's more Mathematics) for auxiliary Schools comes about. Additionally those exploration fill in by [9] best gives information mining structure to Students' academic execution. The examination by [10] utilized harsh set hypothesis as a arrangement methodology on examine scholar information the place the Rosetta toolkit might have been used to assess the learner information will portray different dependencies between those qualities and the person status the place the uncovered examples would clarified for plain English.

Therefore, this grouping calculation serves as a great k_medoid compare to k-mean on screen that progression of students' execution on higher organization. It also enhances the choice making toward academic organizers should screen the candidates' execution semester by semester toward enhancing on the future academic brings about the subsequence academic session.

REFERENCES

[1] Trcka N, Pechenizkiy M, Van der Aalst WMP. Process Mining from Educational Data (Chapter 9); 2011.

[2] Van der Aalst W, Weijters T, Maruster L. Workflow mining: discovering process models from event logs. IEEE Trans Knowl Data Eng 2004, 16:1128–1142.

[3] Romero C, Ventura S. Educational data science in massive open online courses. WIREs Data Mining Knowl Discov 2017, 7:1–12.

[4] Weijters AJMM, van Der Aalst WM, De Medeiros AA. Process mining with the heuristics miner-algorithm. In: Technische Universiteit Eindhoven, Tech. Rep. WP. Vol 166, 2006, 1–34.

[5] Romero C, Cerezo R, Bogarín A, Sánchez-Santillán M. Educational process mining: a tutorial and case study using moodle data sets. In: Data Mining and Learning Analytics: Applications in Educational Research. Hoboken, NJ: John Wiley & Sons; 2016, 1–28

[6] Mans RS, van der Aalst W, Vanwersch RJ. Process Mining in Healthcare: Evaluating and Exploiting Operational Healthcare Processes. Berlin, Germany: Springer; 2015, 17–26.

[7] Trcka N, Pechenizkiy M. From local patterns to global models: towards domain driven educational process mining. In: Ninth International Conference on Intelligent Systems Design and Applications, IEEE, Pisa, Italy, 2009, 1114–1119.

[8] Reimann P, Markauskaite L, Bannert M. E-research and learning theory: what do sequence and process mining methods contribute? Br J Educ Technol 2014, 45:528–540.

[9] Bergenthum R, Desel J, Harrer A, Mauser S. Modeling and mining of learnflows. Trans Petri Nets Other Models Concurrency 2012, 5:22–50.

[10] Perez-Rodriguez R, Caeiro-Rodriguez M, AnidoRifon L. Enabling process-based collaboration in Moodle by using aspectual services. In: Ninth IEEE International Conference on Advanced Learning Technologies, IEEE, Riga, Latvia, 2009, 301–302.

[11] Van der Aalst WM, Guo S, Gorissen P. Comparative process mining in education: An approach based on process cubes. In: International Symposium on DataDriven Process Discovery and Analysis. Springer Berlin Heidelber, Riva del Garda, Italy; 2013, 110–134.

[12] Vidal JC, Vázquez-Barreiros B, Lama M, Mucientes M. Recompiling learning processes from event logs. Knowledge-Based Syst 2016, 100:160–174

[13] Van der Aalst WM. Process Mining: Data Science in Action. Berlin, Germany:Springer; 2016.

[14] Barreiros BV, Lama M, Mucientes M, Vidal JC. Softlearn: a process mining platform for the discovery of learning paths. In: 14th International Conference on Advanced Learning Technologies. IEEE, Athens, Greece; 2014, 373–375.

[15] Pooja Kewat , Roopesh Sharma , Upendra Singh , Ravikant Itare, "Support Vector Machines Through Financial Time Series Forecasting ", Electronics, Communication and Aerospace Technology (ICECA), 2017 International conference of IEEE , 20-22 April 2017 ,pp. 1-7.

[16] Yashika Mathur ,Pritesh Jain ,Upendra Singh, "Foremost Section Study And Kernel Support Vector Machine Through Brain Images Classifier ", Electronics, Communication and Aerospace Technology (ICECA), 2017 International conference of IEEE , 20-22 April 2017 ,pp.1-4.

[17] Vineeta Prakaulya ,Roopesh Sharma ,Upendra Singh, "Railway Passenger Forecasting Using Time Series Decomposition Model ", Electronics, Communication and Aerospace Technology (ICECA), 2017 International conference of IEEE , 20-22 April 2017 ,pp.1-5.

[18] Sonal Sable ,Ankita Porwal ,Upendra Singh , "Stock Price Prediction Using Genetic Algorithms And Evolution Strategies ", Electronics, Communication and Aerospace Technology (ICECA), 2017 International conference of IEEE , 20-22 April 2017 ,pp.1-5.

[19] Rohit Verma ,Pkumar Choure ,Upendra Singh , "Neural Networks Through Stock Market Data Prediction" , Electronics, Communication and Aerospace Technology (ICECA), 2017 International conference of IEEE , 20-22 April 2017 ,pp.1-6.

[20] Dinesh Bhuriya ,Girish Kaushal ,Ashish Sharma," Stock Market Predication Using A Linear Regression ", Electronics, Communication and Aerospace Technology (ICECA), 2017 International conference of IEEE , 20-22 April 2017 ,pp. 1-4.

[21] Ashish Sharma ,Dinesh Bhuriya ,Upendra Singh, "Survey Of Stock Market Prediction Using Machine Learning Approach" , Electronics, Communication and Aerospace Technology (ICECA), 2017 International conference of IEEE , 20-22 April 2017 ,pp.1-5.