SHILLONG COLLEGE

Boyce Road, Laitumkhrah

Shillong, Meghalaya-793003

A PROJECT REPORT

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS OF

THE DEGRESS OF

BACHELOR OF COMPUTER APPLICATION (BCA)

Application of a Simple Clustering technique for prediction of Students' Academic Performance (Data Mining)

BY

STUDENT NAME: GREFFERSON MALNGIANG

ROLL NO: P1400013

REGN.NO: 9495 of 2012-2013

DEPARTMENT OF COMPUTER SCIENCE AND APPLICATION

SHILLONG COLLEGE

NORTH EASTERN HILLS UNIVERSITY

CERTIFIED THAT THIS IS BONIFIED RECORD OF THE PROJECT

ENTITILED

Application of a Simple Clustering technique for prediction of Students' Academic Performance (Data Mining)

SUBMITTED FOR THE FULLFILMENT FOR THE AWARD OF DEGREE OF

BCA (Bachelor Of Computer Application)

BY

STUDENT NAME: GREFFERSON MALNGIANG

ROLL NO: P1400013

REGN NO: 9495 OF 2012-2013

GUIDE	HEAD OF DEPARTMENT	EXAMINER
Sir J. Swer	Mrs A. Mitri	

Viva voice held on: 02/04/2016

Acknowledgement

I take this opportunity to express a deep sense of gratitude and deep regards to my guide Sir Jutang Swer for his exemplary guidance, monitoring and encouragement throughout the thesis.

I am highly in debt to SHILLONG COLLEGE for giving this opportunity in fulfilment of my Bca degree course. Further some I am obliged to staff members of COMPUTER SCIENCE DEPARTMENT for their crucial role and the valuable information provided by them in their respected fields. I am grateful for giving me the permission to use all the facilities and necessary materials with outmost cooperation during the period of my endeavour.

Last but not the least, I feel thankful to all the personages for their guidance, encouragement and feedbacks from time without which this project would not be possible.

Table of Contents

SYNOPSIS

Application of a Simple Clustering technique for prediction of Students' Academic Performance (Data Mining)

Introduction:-

The ability to monitor the progress of students' academic performance is a critical issue to the academic community of higher learning. A system for analyzing students' results based on cluster analysis and uses standard statistical algorithms to arrange their scores data according to the level of their performance is described. In this project, we implemented a Simple Clustering technique for analyzing students' result data. The student's performance plays important role in success of any institution. With the significant increase in number of students and institutions, institutions are becoming increasingly performance oriented and are accordingly setting goals and developing strategies for their achievements.

Approach: With the help of data mining methods, such as clustering Algorithm, it is possible to discover the key characteristics from the students' performance and possibly use those characteristics for future prediction. By taking on a class 11 Science 2015 result of the Shillong College Science students as per the subjects Maths, Physics and Chemistry, the proposed algorithm will be used to analyze the data.

Existing System: K-means Clusteringwith the Euclidean distance measure.

Methodology

By taking the dataset of 129 students and calculate the average of three subjects i.e, maths, physics and chemistry.

Simple Clustering Algorithm:

Step 1: Sort the data in ascending order. Assume that there are n numerical data. Suppose we have ascending

numerical data as follows:

$d_{1,0}=d_{1,1}=\ldots. < d_{2,0}=d_{2,1}=\ldots. < \ldots. < d_{n-1,0}=d_{n-1,1}=\ldots. < \ldots d_{n,0}=d_{n,1}=\ldots.$

where $d_{i,0},d_{i,1},\ldots$ denote the numerical data with the same value, $1<=i<=n$

Then we calculate the Eps (i.e average difference) as follows:

$$\text{Eps} = \frac{\sum_{i=1}^{n-1}(d_{i+1,0} - d_{i,0})}{n-1},$$

where "Eps" denotes the average of the differences between any two adjacent data.

Step 2: Using Eps, we determine whether two adjacent data can be put into a cluster. Suppose we take x_i and x_j as two adjacent data.

If $\mathbf{x_j - x_i <= Eps}$, then we put x_j into the cluster where x_i belongs to otherwise we create a new cluster for x_j. We keep on comparing every two adjacent data until all the data are clustered.

The overall performance is evaluated by applying deterministic model where the group assessment in each of the cluster size is evaluated by summing the average of the individual scores in each cluster.

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

N = the total number of students in a cluster and

n = the dimension of the data

Implement Language: C++

Dataset

Physics	Chemistry	Maths
45	51	31
56	56	40
51	47	16
46	47	17
46	48	10
41	47	11
62	59	35
57	48	38
66	58	33
54	52	30
82	77	84
47	46	24
54	42	30
48	46	18
48	48	12
48	46	15
70	60	63
52	56	30
49	47	19
46	43	25
51	61	39
54	47	12
54	43	21
69	67	45
57	49	18

44	46	16
47	46	21
60	54	53
59	55	40
54	48	45
52	51	24
47	46	25
45	45	20
46	50	32
48	61	25
47	57	22
51	51	23
48	57	15
46	55	16
48	54	15
40	46	13
51	44	20
50	48	14
43	46	7
44	47	30
59	52	33
45	52	14
54	51	31
49	50	30
49	49	4
45	50	5
62	61	33
46	55	8
42	37	1
39	40	4
47	46	10

44	49	17
52	55	17
56	52	38
61	63	20
48	50	15
46	49	30
59	65	37
51	46	23
41	49	30
51	48	18
42	50	14
52	46	25
49	49	30
49	49	15
53	56	43
48	53	30
48	49	18
65	60	61
42	49	17
45	49	12
48	48	17
62	76	55
46	48	11
48	48	22
48	47	10
59	56	22
46	46	12
50	48	30
49	47	24
48	46	22
48	48	16

53	47	9
43	48	1
57	51	30
44	47	8
43	47	12
39	48	12
46	48	19
50	46	14
48	48	14
47	47	21
49	62	33
61	66	30
48	48	19
53	46	7
49	46	30
49	48	18
47	46	13
39	46	6
48	48	13
53	53	45
49	51	24
50	54	30
51	52	5
69	80	84
38	41	10
47	48	21
43	45	7
39	46	13
65	46	36
42	46	8
40	44	9

46	46	11
47	46	16
48	49	14
39	46	23
52	47	13
48	45	21
46	48	5
50	52	30
46	42	6
42	47	5
44	49	16

Source Code

```
#include<iostream>
#include<fstream>
#include<math.h>
#include<stdlib.h>
using namespace std;
main()
{
    float
    eps,distance,avg,sum=0,sum1=0,sum2=0,temp,value[200][3],sub[200][3],over[200],avg1[2
    00],ag,count1;
      int i,j,n,id=1,c=3,count,freq[175],apr;
     /*char stdn_no[50];*/
     cout<<endl;
     //using ifstream to input file stream from a text file
     ifstream ifl("pro1.data");
```

```
    ifl>>n;
    cout<<"No. of Datas="<<n<<endl;

    cout<<"The average of students in three subjects i.e, Maths, Physics and Chemistry:\n";
  for(i=0;i<n;i++)
  ifl>>sub[i][0]>>sub[i][1]>>sub[i][2];

    for(i=0;i<n;i++)
    {
    for(j=0;j<c;j++)
            {
            sum=sum+sub[i][j];
       avg=(sum/3);
       value[i][0]=avg;
       }

   // printf("Sum of the %d row is = %f\n", i, sum);
   // printf("Average of the row %d = %f\n", i, avg);
   sum=0;avg=0;
      }
       for(i=0;i<n;i++)
    {
     cout<<value[i][0]<<endl;
    }

for(i=0;i<n-1;i++)
  for(j=0;j<n-1-i;j++)
 {
  if ((value[j][0]) > (value[j+1][0]))
  {
```

```
    temp= value[j][0];
    value[j][0]=value[j+1][0];
    value[j+1][0]=temp;
   }
  }

      cout<<endl<<"Sorted list is given below:"<<endl;

      for(i=0;i<n;i++)
  cout<<value[i][0]<<endl;
  value[0][1]=id;

  for(i=1;i<n;i++)
  {
  distance=(value[i][0]-value[i-1][0]);
  //cout<<distance<<endl;
  }

  for(i=1;i<n;i++)
  {
   sum1=sum1+(value[i][0]-value[i-1][0]);
   eps=(sum1)/(n-1);
  }

cout<<endl<<"Eps(Radius)="<<eps<<endl;
 for(i=1;i<n;i++)
 {
  if ((value[i][0]-value[i-1][0])<=eps)
  value[i][1]=id;
  else
```

```
  value[i][1]=++id;
 }

cout<<endl<<"The different clusters are clustered by their id no. as follow:"<<endl;

 for(i=0;i<n;i++)
 cout<<value[i][0]<<"\tid="<<value[i][1]<<endl;
 cout<<endl;

for(i=0; i<n; i++)
  {
    count = 1;
    for(j=i+1; j<n; j++)
    {
      if(value[i][1] == value[j][1])
      {
        count++;
        freq[j] = 0;
      }
    }

    if(freq[i]!=0)
    {
      freq[i] = count;
    }
  }
cout<<endl<<"The different cluster sizes is given below :"<<endl;
  for(i=0; i<n; i++)
  {
    if(freq[i]!=0)
```

```
    {
      cout<<"Cluster No."<<value[i][1]<<"\t=\t"<<freq[i]<<endl;
    }
  }
cout<<endl;

for(i=0;i<n;i++)
 {
  sum2=0;
  count1=0;
  for(j=0;j<n;j++)
   {
    if(value[i][1]==value[j][1])
     {
       sum2=sum2+value[j][0];
       count1++;
       over[i]=sum2;
     }
     ag=(sum2/count1);
     avg1[i]=ag;
  }
 }

 for(i=0;i<n;i++)
 {
  cout<<value[i][1]<<"\t"<<over[i]<<"\t\tAverage="<<avg1[i]<<endl;
 }
}
```

EXPERIMENTAL RESULTS:

```
No. of Datas=129
The average of students in three subjects i.e, Maths,Physics and Chemistry:
42.3333
50.6667
38
36.6667
34.6667
33
52
47.6667
52.3333
45.3333
81
39
42
37.3333
36
36.3333
64.3333
46
38.3333
38
50.3333
37.6667
39.3333
60.3333
41.3333
35.3333
38
55.6667
51.3333
49
42.3333
39.3333
36.6667
42.6667
44.6667
```

42
41.6667
40
39
39
33
38.3333
37.3333
32
40.3333
48
37
45.3333
43
34
33.3333
52
36.3333
26.6667
27.6667
34.3333
36.6667
41.3333
48.6667
48
37.6667
41.6667
53.6667
40
40
39
35.3333
41
42.6667
37.6667
50.6667
43.6667
38.3333

62
36
35.3333
37.6667
64.3333
35
39.3333
35
45.6667
34.6667
42.6667
40
38.6667
37.3333
36.3333
30.6667
46
33
34
33
37.6667
36.6667
36.6667
38.3333
48
52.3333
38.3333
35.3333
41.6667
38.3333
35.3333
30.3333
36.3333
50.3333
41.3333
44.6667
36
77.6667

```
29.6667
38.6667
31.6667
32.6667
49
32
31
34.3333
36.3333
37
36
37.3333
38
33
44
31.3333
31.3333
36.3333

Sorted list is given below:
26.6667
27.6667
29.6667
30.3333
30.6667
31
31.3333
31.3333
31.6667
32
32
32.6667
33
33
33
33
33
33.3333
```

34
34
34.3333
34.3333
34.6667
34.6667
35
35
35.3333
35.3333
35.3333
35.3333
35.3333
36
36
36
36
36.3333
36.3333
36.3333
36.3333
36.3333
36.3333
36.6667
36.6667
36.6667
36.6667
36.6667
37
37
37.3333
37.3333
37.3333
37.3333
37.6667
37.6667
37.6667
37.6667

37.6667
38
38
38
38
38.3333
38.3333
38.3333
38.3333
38.3333
38.3333
38.6667
38.6667
39
39
39
39
39.3333
39.3333
39.3333
40
40
40
40
40.3333
41
41.3333
41.3333
41.3333
41.6667
41.6667
41.6667
42
42
42.3333
42.3333
42.6667
42.6667

42.6667
43
43.6667
44
44.6667
44.6667
45.3333
45.3333
45.6667
46
46
47.6667
48
48
48
48.6667
49
49
50.3333
50.3333
50.6667
50.6667
51.3333
52
52
52.3333
52.3333
53.6667
55.6667
60.3333
62
64.3333
64.3333
77.6667
81

Eps(Radius)=0.424479

The different clusters are clustered by their id no. as follow:

26.6667 id=1
27.6667 id=2
29.6667 id=3
30.3333 id=4
30.6667 id=4
31 id=4
31.3333 id=4
31.3333 id=4
31.6667 id=4
32 id=4
32 id=4
32.6667 id=5
33 id=5
33 id=5
33 id=5
33 id=5
33 id=5
33.3333 id=5
34 id=6
34 id=6
34.3333 id=6
34.3333 id=6
34.6667 id=6
34.6667 id=6
35 id=6
35 id=6
35.3333 id=6
35.3333 id=6
35.3333 id=6
35.3333 id=6
35.3333 id=6
36 id=7
36 id=7
36 id=7
36 id=7
36.3333 id=7

36.3333 id=7
36.3333 id=7
36.3333 id=7
36.3333 id=7
36.3333 id=7
36.6667 id=7
36.6667 id=7
36.6667 id=7
36.6667 id=7
36.6667 id=7
37 id=7
37 id=7
37.3333 id=7
37.3333 id=7
37.3333 id=7
37.3333 id=7
37.6667 id=7
37.6667 id=7
37.6667 id=7
37.6667 id=7
37.6667 id=7
38 id=7
38 id=7
38 id=7
38 id=7
38.3333 id=7
38.3333 id=7
38.3333 id=7
38.3333 id=7
38.3333 id=7
38.3333 id=7
38.6667 id=7
38.6667 id=7
39 id=7
39 id=7
39 id=7
39 id=7
39.3333 id=7

```
39.3333 id=7
39.3333 id=7
40      id=8
40      id=8
40      id=8
40      id=8
40.3333 id=8
41      id=9
41.3333 id=9
41.3333 id=9
41.3333 id=9
41.6667 id=9
41.6667 id=9
41.6667 id=9
42      id=9
42      id=9
42.3333 id=9
42.3333 id=9
42.6667 id=9
42.6667 id=9
42.6667 id=9
43      id=9
43.6667 id=10
44      id=10
44.6667 id=11
44.6667 id=11
45.3333 id=12
45.3333 id=12
45.6667 id=12
46      id=12
46      id=12
47.6667 id=13
48      id=13
48      id=13
48      id=13
48.6667 id=14
49      id=14
49      id=14
```

50.3333 id=15
50.3333 id=15
50.6667 id=15
50.6667 id=15
51.3333 id=16
52 id=17
52 id=17
52.3333 id=17
52.3333 id=17
53.6667 id=18
55.6667 id=19
60.3333 id=20
62 id=21
64.3333 id=22
64.3333 id=22
77.6667 id=23
81 id=24

The different cluster sizes is given below :

Cluster No.1	=	1
Cluster No.2	=	1
Cluster No.3	=	1
Cluster No.4	=	8
Cluster No.5	=	7
Cluster No.6	=	13
Cluster No.7	=	45
Cluster No.8	=	5
Cluster No.9	=	15
Cluster No.10	=	2
Cluster No.11	=	2
Cluster No.12	=	5
Cluster No.13	=	4
Cluster No.14	=	3
Cluster No.15	=	4
Cluster No.16	=	1
Cluster No.17	=	4
Cluster No.18	=	1
Cluster No.19	=	1
Cluster No.20	=	1
Cluster No.21	=	1
Cluster No.22	=	2
Cluster No.23	=	1
Cluster No.24	=	1

The different cluster sizes is given below :

1	26.6667	Average=26.6667
2	27.6667	Average=27.6667
3	29.6667	Average=29.6667
4	250.333	Average=31.2917
4	250.333	Average=31.2917
4	250.333	Average=31.2917
4	250.333	Average=31.2917
4	250.333	Average=31.2917
4	250.333	Average=31.2917
4	250.333	Average=31.2917
4	250.333	Average=31.2917
5	231	Average=33
5	231	Average=33
5	231	Average=33
5	231	Average=33
5	231	Average=33
5	231	Average=33

5	231	Average=33
6	452.667	Average=34.8205
6	452.667	Average=34.8205
6	452.667	Average=34.8205
6	452.667	Average=34.8205
6	452.667	Average=34.8205
6	452.667	Average=34.8205
6	452.667	Average=34.8205
6	452.667	Average=34.8205
6	452.667	Average=34.8205
6	452.667	Average=34.8205
6	452.667	Average=34.8205
6	452.667	Average=34.8205
6	452.667	Average=34.8205
7	1690.33	Average=37.563
7	1690.33	Average=37.563
7	1690.33	Average=37.563
7	1690.33	Average=37.563
7	1690.33	Average=37.563
7	1690.33	Average=37.563
7	1690.33	Average=37.563
7	1690.33	Average=37.563
7	1690.33	Average=37.563
7	1690.33	Average=37.563
7	1690.33	Average=37.563
7	1690.33	Average=37.563
7	1690.33	Average=37.563
7	1690.33	Average=37.563
7	1690.33	Average=37.563
7	1690.33	Average=37.563
7	1690.33	Average=37.563
7	1690.33	Average=37.563
7	1690.33	Average=37.563
7	1690.33	Average=37.563
7	1690.33	Average=37.563
7	1690.33	Average=37.563

7	1690.33	Average=37.563
7	1690.33	Average=37.563
7	1690.33	Average=37.563
7	1690.33	Average=37.563
7	1690.33	Average=37.563
7	1690.33	Average=37.563
7	1690.33	Average=37.563
7	1690.33	Average=37.563
7	1690.33	Average=37.563
7	1690.33	Average=37.563
7	1690.33	Average=37.563
7	1690.33	Average=37.563
7	1690.33	Average=37.563
7	1690.33	Average=37.563
7	1690.33	Average=37.563
7	1690.33	Average=37.563
7	1690.33	Average=37.563
7	1690.33	Average=37.563
7	1690.33	Average=37.563
8	200.333	Average=40.0667
8	200.333	Average=40.0667
8	200.333	Average=40.0667
8	200.333	Average=40.0667
8	200.333	Average=40.0667
9	629.667	Average=41.9778
9	629.667	Average=41.9778
9	629.667	Average=41.9778
9	629.667	Average=41.9778
9	629.667	Average=41.9778
9	629.667	Average=41.9778
9	629.667	Average=41.9778
9	629.667	Average=41.9778
9	629.667	Average=41.9778
9	629.667	Average=41.9778
9	629.667	Average=41.9778
9	629.667	Average=41.9778

9	629.667	Average=41.9778
9	629.667	Average=41.9778
10	87.6667	Average=43.8333
10	87.6667	Average=43.8333
11	89.3333	Average=44.6667
11	89.3333	Average=44.6667
12	228.333	Average=45.6667
12	228.333	Average=45.6667
12	228.333	Average=45.6667
12	228.333	Average=45.6667
12	228.333	Average=45.6667
13	191.667	Average=47.9167
13	191.667	Average=47.9167
13	191.667	Average=47.9167
13	191.667	Average=47.9167
14	146.667	Average=48.8889
14	146.667	Average=48.8889
14	146.667	Average=48.8889
15	202	Average=50.5
15	202	Average=50.5
15	202	Average=50.5
15	202	Average=50.5
16	51.3333	Average=51.3333
17	208.667	Average=52.1667
17	208.667	Average=52.1667
17	208.667	Average=52.1667
17	208.667	Average=52.1667
18	53.6667	Average=53.6667
19	55.6667	Average=55.6667
20	60.3333	Average=60.3333
21	62	Average=62
22	128.667	Average=64.3333
22	128.667	Average=64.3333
23	77.6667	Average=77.6667
24	81	Average=81

Statistical Measures

Cluster#	Cluster Size	Overall Performance
1	1	26.66
2	1	27.66
3	1	29.66
4	8	31.29
5	7	33
6	13	34.82
7	45	37.56
8	5	40.06
9	15	41.97
10	2	43.83
11	2	44.66
12	5	45.66
13	4	47.91
14	3	48.88
15	4	50.5
16	1	51.33
17	4	52.16
18	1	53.16
19	1	55.66
20	1	60.33
21	1	62
22	2	64.33
23	1	77.66
24	1	81

Table 1 :Performance Index

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

Graphs:

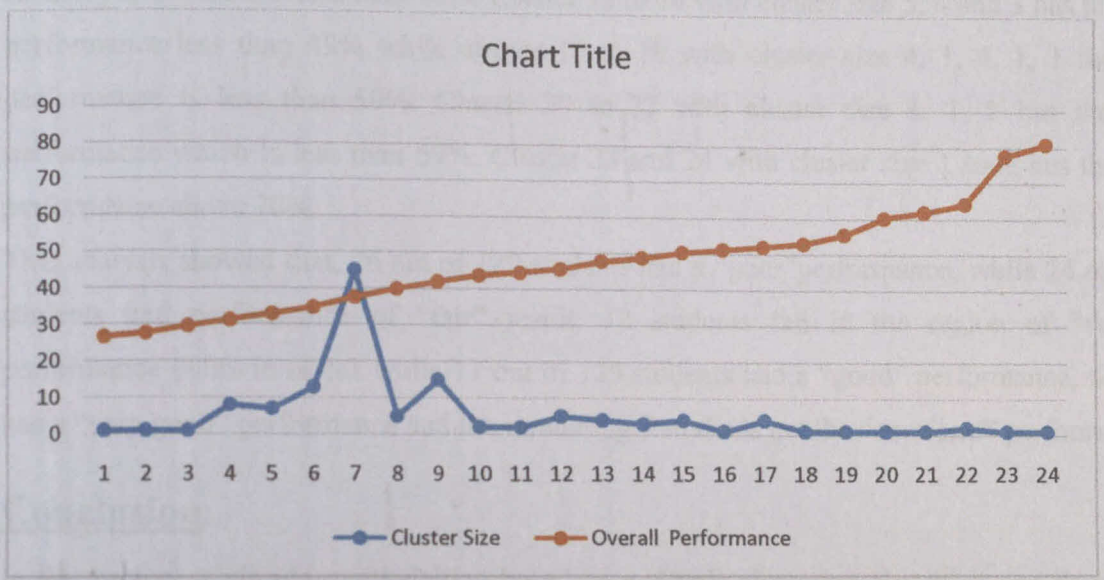
Chart Title
90
80
70
60
50
40
30
20
10
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
Cluster Size
Overall Performance

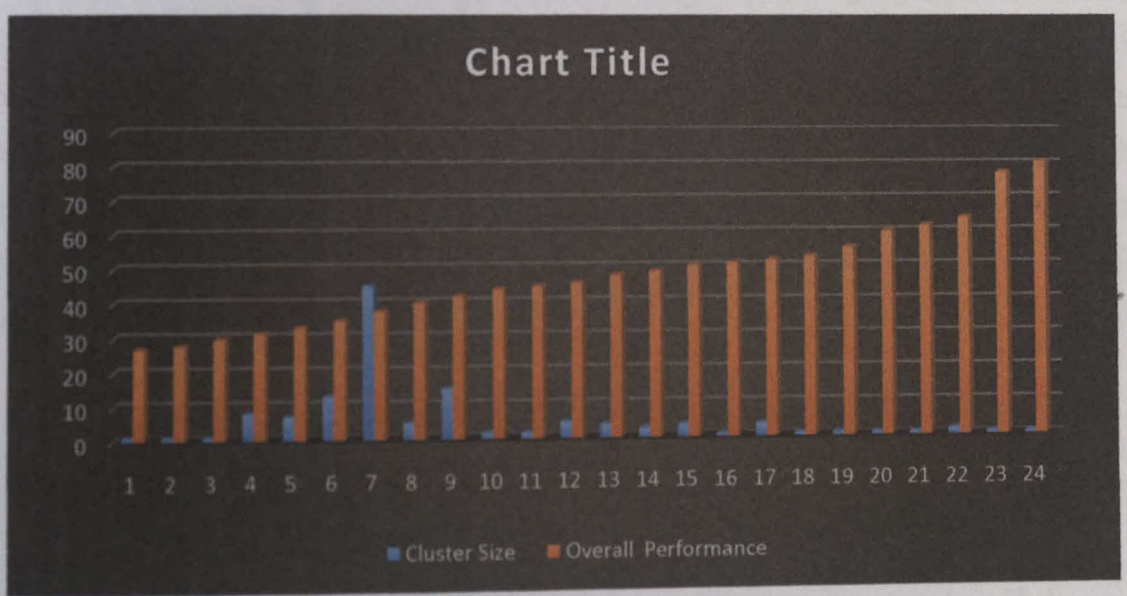
Chart Title
90
80
70
60
50
40
30
20
10
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
Cluster Size
Overall Performance

Explanation:

From the above data we can see that there are 24 number of clusters each of which contains different cluster sizeswith corresponding overall performance. Based on the trends above the graphs it depicts that cluster number 1 to 7 with cluster size of 1, 1, 1, 8, 7, 13 and 45 has the overall performance which is less than 40% while cluster 8 to 11 with cluster size 5, 15, 2, 2 the overall performance is less than 45%. Cluster 12 to 14 with cluster size 5, 4 and 3 has the overall performance less than 49% while cluster 15 to 19 with cluster size 4, 1, 4, 1, 1 the overall performance is less than 59%. Cluster 20 to 22 with cluster size 1, 1, 2 has the overall performance which is less than 69%. Cluster 23 and 24 with cluster size 1 each has the overall performance above 70%.

This analysis showed that, 76 out of 129 students had a "poor"performance, while 24 out of 129 students had performance of "fair" result. 12 students fall in the region of "very fair" performance index in table1 while 11 out of 129 students had a "good" performance, 4 students has a "very good" performance and the remaining 2 students get the "excellent" performance.

Conclusion:

In this project, a simple methodology based on a simple clustering algorithm and deterministic model is being used to evaluate the performance of students in institutions. This methodology will assist the academic planners to monitor student's performance during each term. Hence this model will play important role for academic planners to determine the reasons for decline in performance of students during particular semester and steps that need to be taken to improve performance from next academic session. However, this project can be further enhance according to needs.

Bibliography:

1. Evaluate Student's Performance Using K-means Clustering, Deterministic model by Rakesh Kumar Arora, Dr. Dhanmendra Badal
2. Application of K-means Clustering algorithm for student's academic performance by Oyelade O. J, Oladipupo O.O and Obagbuwa I. C.
3. S. Sujit Sansgiry, M. Bhosle, and K. Sail, "Factors that affect academic performance among pharmacy students," American Journal of Pharmaceutical Education, 2006.
4 J. O. Omolehin, J. O. Oyelade, O. O. Ojeniyi and K. Rauf, "Application of Fuzzy logic in decision making on students' academic performance," Bulletin of Pure and Applied Sciences, vol. 24E(2), pp. 281-187, 2005.