Evaluating Student's Performance Using k-Means Clustering

[1]Rakesh Kumar Arora, [2]Dr. Dharmendra Badal
[1]Dept. of Computer Science, Krishna Engineering College, Ghaziabad, UP, India
[2]Dept. of Mathematical Science & Computer Applications, Bundelkhand University, Jhansi, UP, India

Abstract
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. A system to analyze the performance of students using k-means clustering algorithm coupled with deterministic model is being described in this paper. The result of analysis will assist the academic planners in evaluating the performance of students during specific semester and steps that need to be taken to improve students' performance from next batch onwards.

Keywords
Data Mining, Education, k-Means Clustering, Deterministic Model

I. Introduction

In India ranking of students is mainly done on basis of marks obtained in exams. Most of the Indian universities follow the common pattern of grouping the students in the category of pass with distinction for students scoring 75% and above, first division for students scoring from 60% to 74.9%, second division for students scoring from 50% to 59.9%, third division having the range from 40% to 49.9%.Many Universities set a minimum set of marks that should be maintained in order to continue in the degree program. In some University, the minimum requirement set for the students is 40% while in other cases it may be 33%. Therefore, marks obtained in examination is still remains the most common factor used by the academic planners to evaluate progression in an academic environment. With traditional approach of grouping students based on their average scores, it is difficult to obtain a comprehensive view of the state of the students' performance and simultaneously discover important details from their time to time performance [1].

A very promising tool to attain valuable information about student's performance is the use of data mining. Data mining techniques are used to discover hidden information, patterns and relationships of large amount of data, which is very much helpful in decision making. With the help of data mining methods, such as clustering, decision tree or association analysis it is possible to discover the key characteristics from the students' performance and possibly use those characteristics for future prediction. This paper presents k-means clustering algorithm as a simple and efficient tool to monitor the progression of students' performance in higher institution [1].

II. Methodology

K-Means clustering generates a specific number of disjoint, flat (non-hierarchical) clusters. The K-Means method is numerical, unsupervised, non-deterministic and iterative [2]. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main idea is to define k centroids, one for each cluster. Associate each point belonging to a given data set and to the nearest centroid. After all the points in the data set are over, the first step is completed and an early grouping is done. Re-calculate k new centroids as barycenters of the clusters resulting from the previous step. After k new centroids has been calculated, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop the k centroids change their location step by step until no more changes are done. This algorithm aims at minimizing an objective function, in this case a squared error function. The objective function

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$$

where $\left\| x_i^{(j)} - c_j \right\|^2$ is a chosen distance measure between a data point $x_i^{(j)}$ and the cluster centre c_j, is an indicator of the distance of the n data points from their respective cluster centre's.[3]

Algorithmic steps for k-means clustering (4)
Let X = $\{x_1,x_2,x_3,\ldots\ldots,x_n\}$ be the set of data points and V = $\{v_1,v_2,\ldots\ldots,v_c\}$ be the set of centers.
1. Randomly select 'c' cluster centers.
2. Calculate the distance between each data point and cluster centers.
3. Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers..
4. Recalculate the new cluster center using:

$$v_i = \left(\frac{1}{c_i}\right) \sum_{j=1}^{c_i} x_i$$

where, 'c_i' represents the number of data points in i^{th} cluster.
5. Recalculate the distance between each data point and new obtained cluster centers.
6. If no data point was reassigned then stop, otherwise repeat from step 3.

The analysis using k-means clustering is being done with the help of Tanagra tool. TANAGRA is free DATA MINING software for academic and research purposes. It proposes several data mining methods from exploratory data analysis, statistical learning, machine learning and databases area. TANAGRA is an "open source project" as every researcher can access to the source code, and add his own algorithms, as far as he agrees and conforms to the software distribution license [5].

III. Results

The model was applied on the on students of Department of Computer Science and Engineering of reputed Engineering College of Ghaziabad. The analysis is being performed on the basis of marks obtained by students of batch 2009-2013 from semester III to semester VII. The semester I and II are not used for analysis since during these semesters the focus is more on generalized subjects rather than in specialization. The number of students involved in analysis is 118 and dimensions (total number of subjects) are 10. The results generated using Tanagra software for semester III to semester VII for k=4 (clusters) is shown in fig. 1.

The overall performance is evaluated by applying deterministic model in eq. 1 (1) where the group assessment in each of the cluster size is evaluated by summing the average of the individual scores in each cluster.

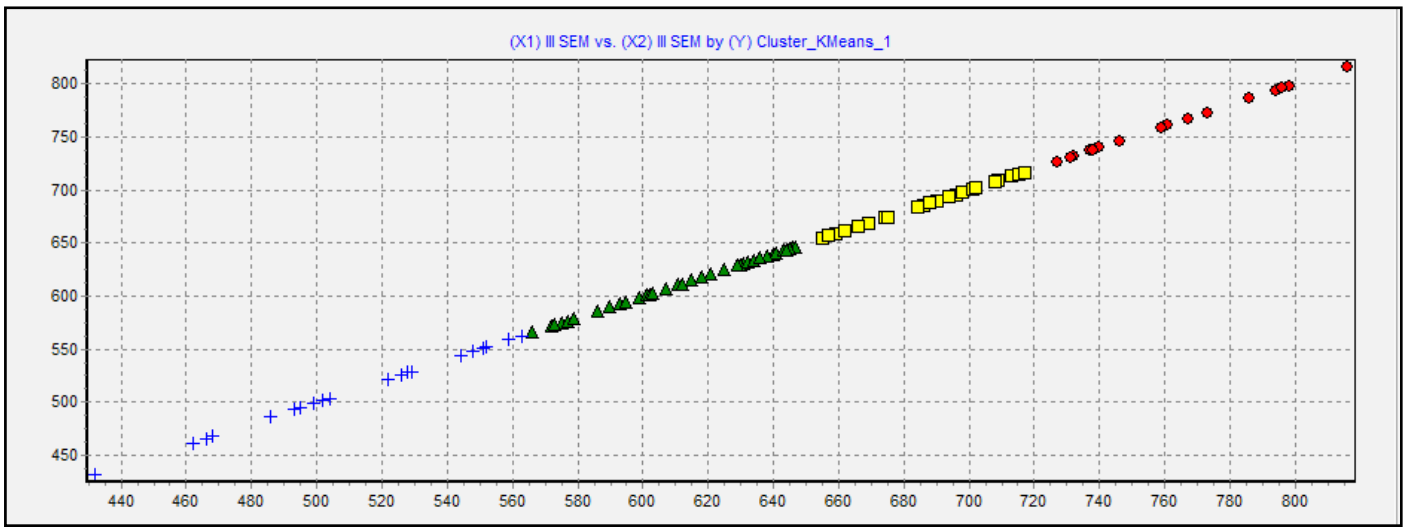
$$\frac{1}{N}\left(\sum_{j=1}^{N}\left(\frac{1}{n}\sum_{i=1}^{N}x_i\right)\right) \quad (1)$$

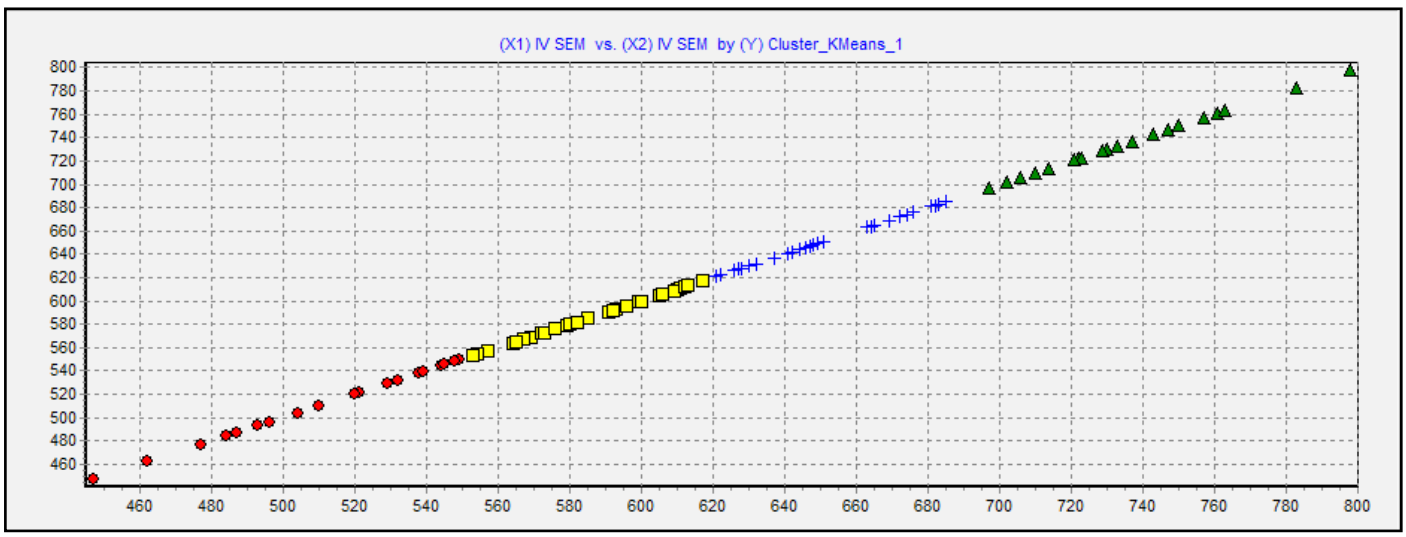
N = the total number of students in a cluster and
n = the dimension of the data

The performance evaluated by using eq. (1) for each cluster for semester III to semester VII is shown below in Table 1.

Table 1: Performance oxf students

Semester	Cluster	Cluster Size	Performance
III	1	25	51.63
	2	44	61.16
	3	31	68.73
	4	18	75.88
IV	1	24	51.37
	2	35	58.65
	3	33	65.17
	4	26	73.47
V	1	25	58.28
	2	29	66.88
	3	34	72.19
	4	30	78.97
VI	1	21	59.18
	2	23	65.41
	3	38	70.21
	4	36	77.40
VII	1	17	56.10
	2	48	64.19
	3	37	71.11
	4	16	79.05





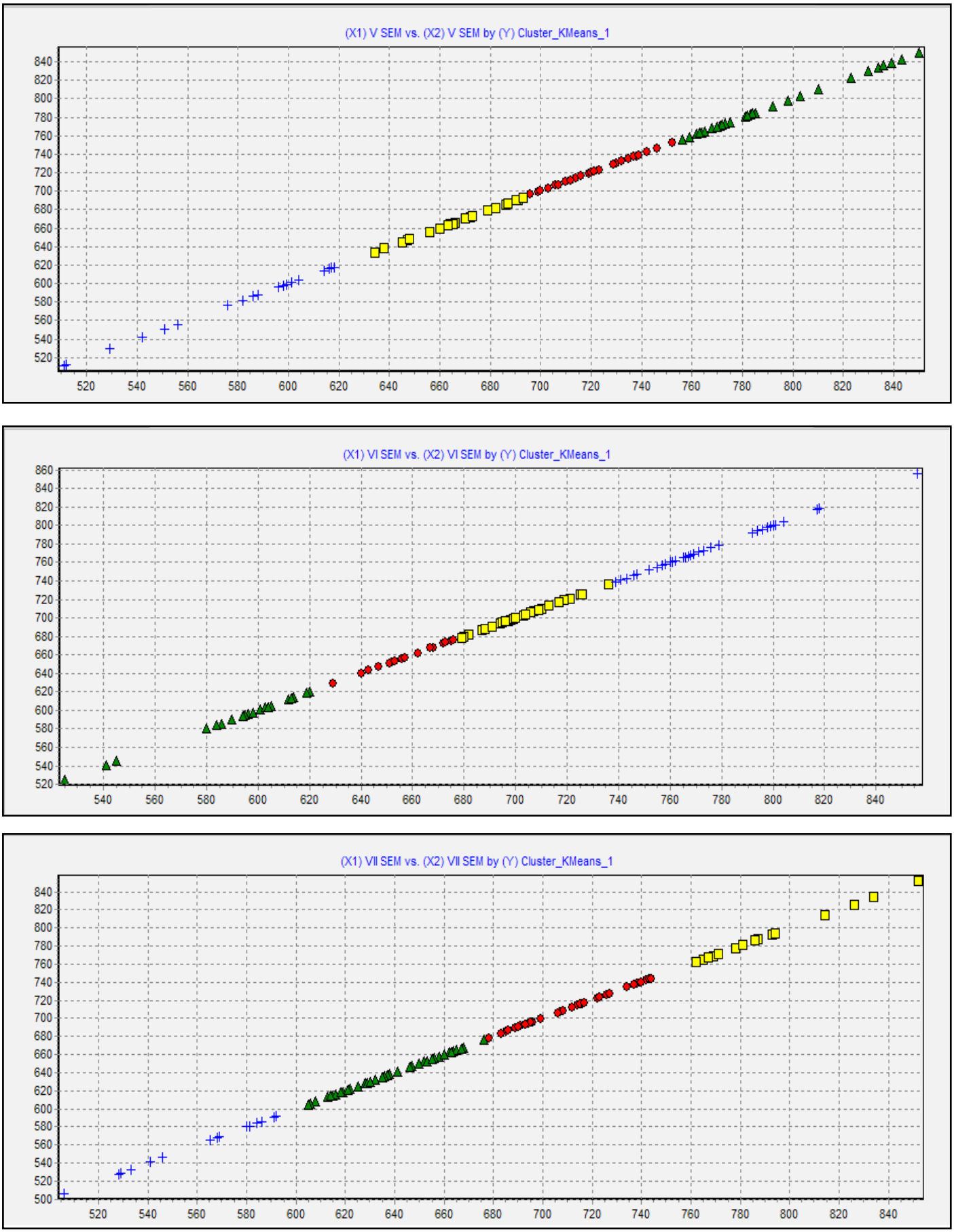


Fig. 1: Clusters Generated Using Tanagra Software from Semester III to Semester VII

The performance of students from semester III to semester VII is graphically shown as

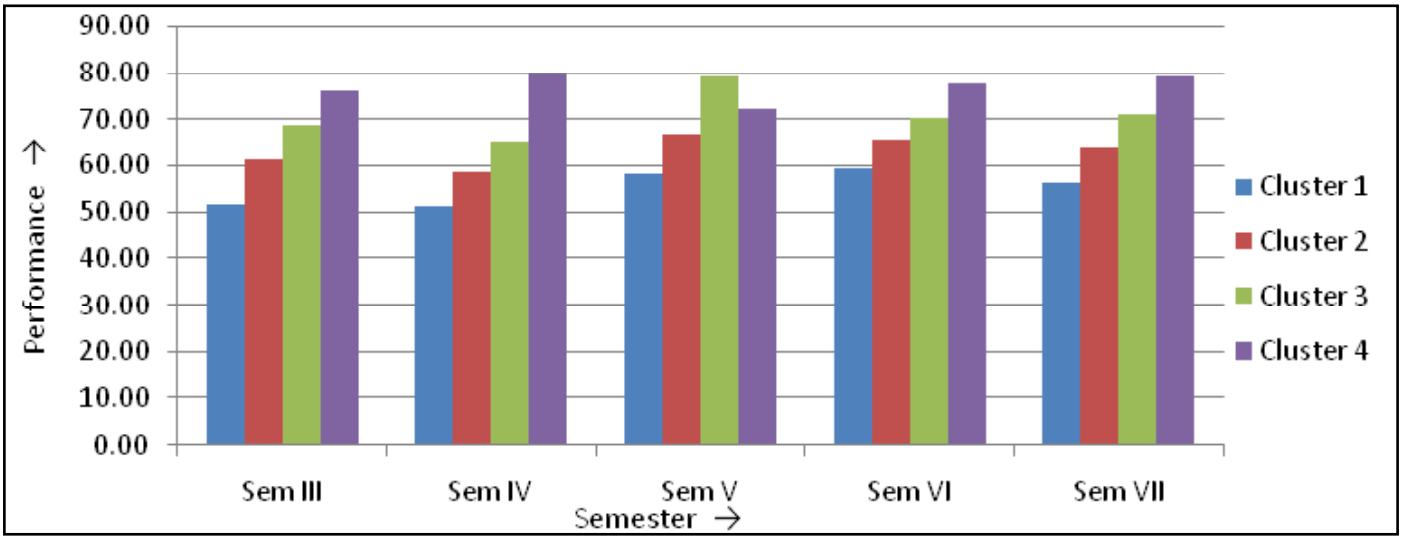


Fig. 2: Performance of Students from Semester III to Semester VII

To compare the performance of students, the grades are being associated with various percentage intervals.

Table 2: Performance Index

Percentage	Grades
75 and above	A
70-74	B
65-69	C
60-64	D
55-59	E
50-54	F
Below 50	G

In Table 1 for semester III, the overall performance for cluster 1 is 51.63%, for cluster 2 is 61.16%, for cluster 3 is 68.73% and for cluster 4 is 75.88%. If this performance is compared with performance index table as depicted in Table 2, the analysis shows that 25 students out of 118 belongs to F grade, 44 students belong to D grade , 31 students belong to C grade and 18 students belong to A grade.

In semester IV, the overall performance for cluster 1 is 51.37%, for cluster 2 is 58.65%, for cluster 3 is 65.17% and for cluster 4 is 73.47%. If this performance is compared with performance index table as depicted in Table 2, than the analysis shows that 24 students out of 118 belongs to F grade, 35 students belong to E grade , 33 students belong to C grade and 26 students belong to B grade. There is significant decrease in performance of students in semester IV as compared to semester III. The table 2 clearly shows that the performance of students in cluster 2 has been downgraded from grade D to grade E. Similarly, in cluster 4 the performance of students has been degraded from grade A to grade B.

In semester V, the overall performance for cluster 1 is 58.28%, for cluster 2 is 66.88%, for cluster 3 is 72.19% and for cluster 4 is 78.97%. If this performance is compared with performance index table as depicted in Table 2, than the analysis shows that 25students out of 118 belongs to E grade, 29 students belong to C grade, 34 students belong to B grade and 30 students belong to A grade. The performance is greatly enhanced in semester V as compared to semester IV. The Table 2 clearly shows that there are no students left in grade F. The average marks have increased in semester V as majority of students are now in grade A and grade B.

In semester VI, the overall performance for cluster 1 is 59.18%, for cluster 2 is 65.41%, for cluster 3 is 70.21% and for cluster 4 is 77.40%. If this performance is compared with performance index table as depicted in Table 2, than the analysis shows that 21students out of 118 belongs to E grade, 23 students belong to C grade, 38 students belong to B grade and 36 students belong to A grade. The performance is further improved in semester VI as more number of students are placed in grade A and grade B.

In semester VII, the overall performance for cluster 1 is 56.10%, for cluster 2 is 64.19%, for cluster 3 is 71.11% and for cluster 4 is 79.05%. If this performance is compared with performance index table as depicted in Table 2, than the analysis shows that 17 students out of 118 belongs to E grade, 48 students belong to D grade , 37 students belong to B grade and 16 students belong to A grade. The performance has declined in semester VII as numbers of students are being placed in grade D. Though the performance of students in cluster 3 and cluster 4 has increased but the number of students is greatly reduced in these two clusters.

IV. Conclusion

In this paper, a simple methodology based on k-means clustering algorithm and deterministic model is being used to evaluate the performance of students in higher institutions. This methodology will assist the academic planners to monitor student's performance during each semester. 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.

References

[1] Oyelade O. J, Oladipupo O. O, Obagbuwa, I. C , "Application of k-Means Clustering algorithm for prediction of Students' Academic Performance", International Journal of Computer Science and Information Security, Vol. 7, 2010

[2] [Online] Available: http://www.improvedoutcomes.com/docs /WebSiteDocs/Clustering/K_Means_ Cluste ring_Overview. htm

[3] [Online] Available: http://www.home.deib.polimi.it/ matteucc/Clus tering/tutorial_html/kmeans.html

[4] [Online] Available: https://sites.google.com/site/datacluster ing algorithms /k-means-clustering-algori thm

[5] [Online] Available: http://www.eric.univ-lyon2.fr/~ricco/tanagra/ en/tanagra.html
[6] Arora K Rakesh, Badal Dharmendra,"Location wise student admission analysis", International Journal of Computer Science, Information Technology and Security, Dec 2012.
[7] Arora K. Rakesh, Gupta K. Manoj,"Data Mining: Scope Out Valuable Resources From Mountains Of Information", IITM Buisness Review Journal, July 10

Rakesh Kumar Arora is currently working in Department of Computer Science at Krishna Engineering College, Mohan Nagar, Ghaziabad, U.P, India. He has more than 10 years of teaching expierence in reputed institutes. He has no. of papers in International Journals and Conferences to his credit.

Dr. Dharmendra Badal is currently working in Department of Mathematical Sciences and Computer Applications at Bundelkhand University, Jhansi, U.P, India. He has more than 20 years of expierence at Bundelkhand University. He is also handling the additional responsibilities of Computer Head and Controller of Examination at Bundelkhand University. He was also associated as director at SRI institutions, Datia. He had presided no. of conferences and has no. of papers in International Journals and Conferences to his credit.