Application of Clustering for Student Result Analysis

Deepshikha Aggarwal, Deepti Sharma

Abstract*: The analysis of academic performance of students is an important concern for the universities and colleges of higher learning as it is very important for planning and management of the teaching pedagogy. There is a need for the system to examine and assess the results of students in order to understand how effective the existing education system is. In this study, we have analysed the students' performance using the clustering and some other statistical tools and methods. This paper studies the actual result of university examination for the students of MCA (Masters in Computer Application), a 3 year post graduate course in information technology. In order to analyse the data k-means clustering algorithm is applied. The elbow method is used to choose appropriate number of clusters. Analysis has also been conducted gender wise to understand whether there is a pattern based on the gender of students. Academic planners can make operational decisions and future planning on the base of results attained in this research.

Index Terms*: k – *mean clustering, elbow method, academic performance, statistical algorithm

I. INTRODUCTION

Today, in most of the universities and colleges, the performance of the students academically has recently come under analysis for a number of reasons. There is number of factors involved that affects the academic performance of students. These factors may vary from one set of students to next, form one academic background to another and from one state or nation to another. At times, students work hard but their hard work may not be correlated with their result. This may lead to high failure percentage and thus increases the rate of training these graduates. Low pass percentage of students may enforce enormous cost on community in terms of low number of students graduating and reduced intake of probable users due to absence of spaces. So, in order to confirm that a big portion of labour force is highly trained, all universities and college must ensure and consider the aspects that affect the performance of the students.

To attain high profile position and more than expected salary, the higher education level is must for any organization. Many researches have been conducted to analyse the factors associated with academic performance of students in various universities but less are done for private colleges. The cost of education in private colleges is much high as compared to those of public institutions. There is no guarantee that qualifying for admission to any college will lead to success in the degree course. But the academic performance can be affected by various factors that may lead to success in any students' life. To ensure quality improvement in the universities, private universities and colleges must keep a check and assess excellence in the parts of "curriculum, teaching and academic programmes, research and scholarship, staffing, students, physical facilities, equipment and learning materials and academic environment" [14]. Accessing the factors for academic performance is also important so that students can be assisted to develop their individual academic performance in the university. This research study is intended to focus on the interaction of different factors and the role that they play to the performance of students in academics in their college or university.

This study is conducted to identify and explore the most probable factors that have a role in influencing the academic performance of students in higher education in colleges in Delhi, India. The research was conducted to identify the different factors affecting the students' performance in the university examination, and then we have established the relationship between these factors and the influence of these on the student result. We have also analysed the performance of students on the basis of gender and have been able to acquire some good results of the analysis. This paper is based on a study conducted to analyse the academic performance of students in a higher learning scenario. The sample group taken is the students of MCA course which is a post graduate level programme in information technology.

We have implemented various statistical tools for the study. The analysis of academic performance of students is an important concern for the universities and colleges. It is important for planning and management of the teaching pedagogy.

Analysis of student performance is essential to understand how effective the existing education system is. The segregation of data is done using the clustering method. K-means clustering algorithm is implemented in R [15][16]. We have used Microsoft Excel to analyse the data and plot various graphs. The different methods like clustering algorithms can be helpful in performing the analysis like the student's results for assessing the impact of different factors on performance [1]. The final goal of the study is to formulate a teaching pedagogy on the basis of the data analysis. The factors affecting the academic performance of the students in each cluster help in identifying the strong and weak areas of the current teaching pedagogy. The final result of the study is an effective teaching pedagogy that can addresses all these factors.

Revised Manuscript Received on December 22, 2018.

Deepshikha Aggarwal, Department of Information Technology, Jagan Institute of Management Studies, Delhi, India

Deepti Sharma, Department of Information Technology, Jagan Institute of Management Studies, Delhi, India.

II. LITERATURE REVIEW

Learning always leads to performance. It is the main pointer to show learning is resulted. [5] explains about various categories of the outcomes of learning. They are defined as verbal information, intellectual skills, attitudes and motor skills. Learning process has an impact and importance for these outcomes. Various learning errands are required to show various learning results. Grade Point Average (GPA) is a main measure for finding students' performance academically [6]. It includes all the learning results anticipated of a student in his/her semesters' subjects. According to [7], an important interpreter of any students' graduate performance is GPA. Thus, it can be considered as the main measure of student's current and future studies. As per [8], putting higher demands on student learner is more important as compared to external forces. According to [9], the institutional guidelines and rules must be implemented properly to enhance the studying environment. It is needed that institution's plans and procedures must be directed towards students' accountability and they must be active enough to participate in their own college and university for promotion. [10] did not consider other variables like students' internal nature which would be obtained from students' assessments or results. [11] studied both the students' characteristics and self-reported theories to identify the aspects that add to the academic presentation of students at the university. Self-efficacy, assigned goals, self-goals and abilities are four major causes the students' academic performances. According to [12], students' age is important for qualification which is essential for academic performance. [13] through path analysis observed and explained gender to have a direct effect on the academic performance of third year students.

III. METHODOLOGY

The following research objectives were formulated to analyse the student result:

- To identify categories of students as per the performance in exams.
- To identify factors affecting the performance of students in each category.
- To perform Gender wise performance analysis of the student result.
- Proposing an improved teaching pedagogy on the basis of the research outcomes.

A. Data Demographics

The data used for the study is the actual result of university examination for the students of MCA (Masters in Computer Application), a 3 year post graduate course in information technology.

TABLE 1: DATA DEMOGRAPHICS

Total number of students:	117
Number of subjects in first semester:	09
Number of subjects in second semester:	10
Number of male students:	63
Number of female students:	54

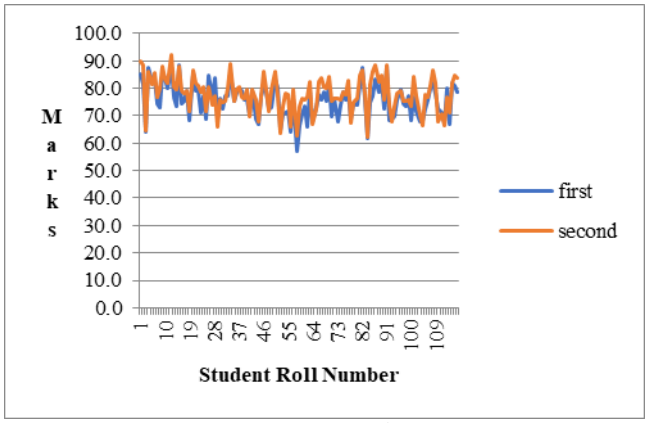

Fig. 2 Overall Result of Students

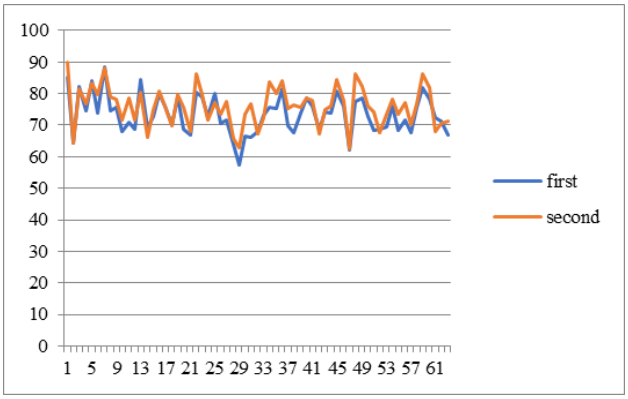

Fig3 Result of Male Students

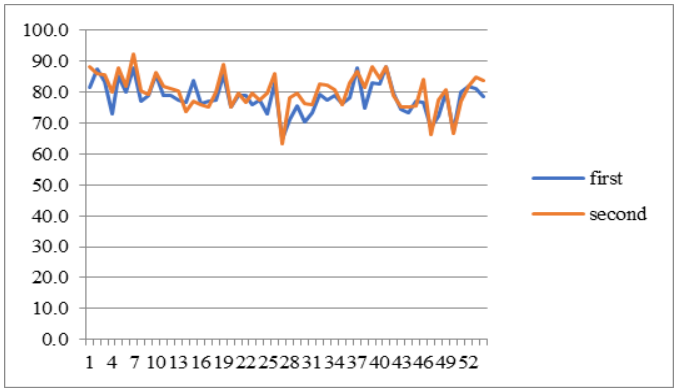

Fig. 4 Result of Female Students

B. Research Methodology

In this study, we are analysing the result of students in order to formulate the future teaching pedagogy on the basis of students' performance in the first year of study in the MCA (Masters in Computer Application), a 3 year post graduate course in information technology. For the purpose of conducting this analysis we have used clustering as the statistical tool. Clustering is defined as a set of techniques that are used to find subgroups of observations within a given data set which in our study is the result of students who appeared in the university exam for the first and second semester of the MCA course.

When we cluster the data, we get the results as data in the same cluster to be similar whereas data in different clusters is not similar. Clustering is a method of unsupervised machine learning. We have chosen this method for our dataset because this method is used to find relationships between the collected observations. In our case the observations are the factors affecting the performance of students in the university exams. This method can establish relationships between observed values without being trained by a response variable as it is an unsupervised method. Clustering allows us to identify which parts of the dataset are similar, and hence enable the researchers to categorize the data in order to perform the required analysis.

In this paper we have used a research methodology based on collection and analysis of data in order to propose an effective teaching pedagogy for the institutions of higher education. This analysis can be used as an effective tool by the colleges in order to understand the relationship between various factors affecting the students in their academic performance. This in turn will help the academic policy makers to provide an effective teaching methodology accommodation all the important factors.

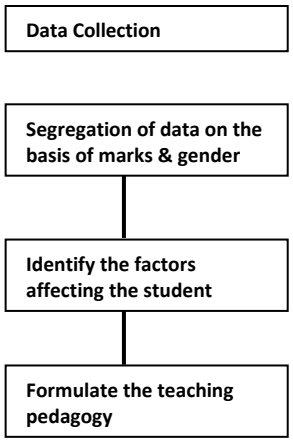

Fig 1 Research Methodology

C. K-Means Clustering

K-means clustering is a method of unsupervised machine learning and is effectively used for partitioning a given dataset into k groups or *k* clusters, where *k* represents the number of groups or clusters. The number of clusters is calculated by the analyst to derive an effective grouping of data. The k-means clustering method segregates the data into clusters in such a way that the data values within the same cluster are similar to each other. In k-means clustering, each cluster is represented by its centre point or mean which is also termed as the centroid and is calculated as mean value of the data within that cluster.

D. Estimating Number of Clusters

The first step was to estimate the number of clusters required for an effective analysis of the data. We have used the Elbow method for this purpose. The k-means clustering works by defining the clusters in such a way that the total variation within a given cluster is minimum. The total within cluster sum of squares is denoted by WSS and it indicates how compact of the cluster is and our aim while using the clustering method is to keep the WSS as small as possible. This leads to formation of effective clusters for the data analysis. The Elbow method is a method that considers the total WSS as a function of the number of clusters suitable for the particular dataset and enables us to choose the appropriate number of clusters in such a way that adding more clusters does not have any impact on the data analysis results.

minimize($k\sum k=1W(Ck)$)

where Ck is the kth cluster and W(Ck) is the within-cluster variation. We have used the following algorithm to define the optimal clusters:

- For different values of k, the k-means clustering is computed.
- For each cluster *k*, we have to calculate the total within the cluster sum of square (WSS)
- The next step is to plot the curve for the values of WSS thus obtained for all the clusters *k*.
- The location of a bend (knee) in the plot is the indicator of the appropriate number of clusters.

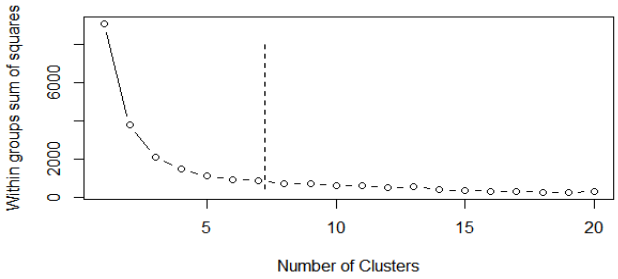

Fig 5 Estimating Number of clusters

After plotting the graph, we chose 5 as the cut off point, because while the wss does continue to decrease, but it doesn't seem to do so at a considerable rate to accept the added complexity of more clusters.

IV. DATA ANALYSIS

A. k-means Clustering of Data

k – means clustering is performed on the data for the marks in the first and second semester with 5 clusters as determined by the Elbow method. The resultant graph is as follows:

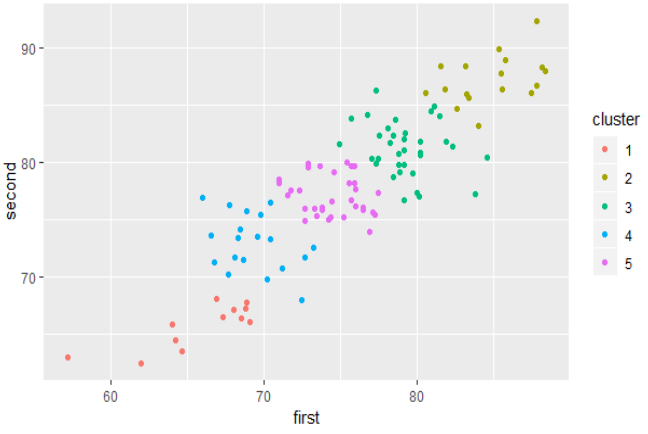

Fig 6 k-means clustering with 5 clusters

B. Cluster Analysis

The analysis of the clusters thus obtained is as follows:

TABLE 2: CLUSTER ANALYSIS

CLUSTER NUMBER	MEAN MARKS OF FIRST SEMESTER	MEAN MARKS OF SECOND SEMESTER
1	65.80	65.65
2	84.80	87.26
3	79.23	81.16
4	69.32	72.94
5	74.50	77.10

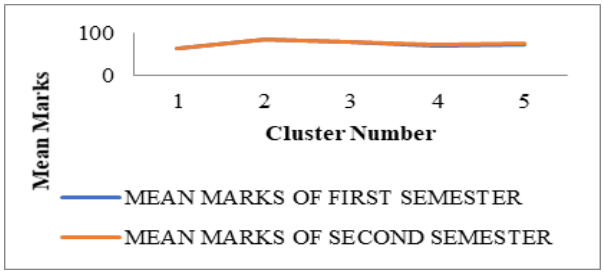

Fig 7 Cluster Analysis

TABLE 3: GENDER-WISE CLUSTER ANALYSIS

CLUSTER NUMBER	NUMBER OF MALES	NUMBER OF FEMALES
1	9	3
2	5	11
3	13	21
4	18	3
5	18	16

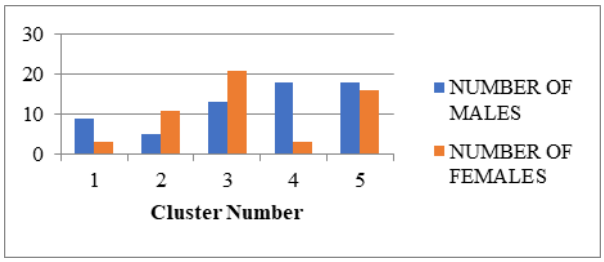

Fig 8 Gender wise Cluster Analysis

V. DISCUSSION AND CONCLUSION

In this study, we have analyzed the result of students in order to formulate the future teaching pedagogy on the basis of students' performance in the first year of study in the MCA (Masters in Computer Application), a 3 year post graduate course in information technology. For the purpose of conducting this analysis we have used k – means clustering in R as the statistical tool. Along with that we have also used Microsoft excel to plot various graphs for the analysed data. When we cluster the data, we get the results as data in the same cluster to be similar whereas data in different clusters is not similar. Using elbow method, we have found appropriate number of clusters that can be used for this study. Cluster analysis with five clusters is shown. Gender wise cluster analysis is also shown. From table 2 of cluster analysis, it can be seen that students with good performance lie in cluster 2 and cluster 3. Also, it is proved (from table 3) that more females lay in cluster 2 and 3 and thus females have better academic performance as compared to male students. In future we plan to analyse the data from different semesters and on compare the data on various parameters.

REFERENCES

1. Fahim A. M., Salem A. M., Torkey F. A. and Ramadan M. A., "An efficient enhanced k-means clustering algorithm," Journal of Zhejiang University Science A., pp. 1626–1633, 2006.
2. S. Sujit Sansgiry, M. Bhosle, and K. Sail, "Factors that affect academic performance among pharmacy students," American Journal of Pharmaceutical Education, 2006.
3. Naeimeh Delavari, "Application of Enhanced Analysis Model for Data Mining Processes in Higher Educational System ", 2005 IEEE July 7 Juan Dolio, Dominican Republic ITHET 6th Annual International Conference.
4. Pooja Thakar, Anil Mehta , "Performance analysis and Prediction in Educational Data Mining", International Journal of Computer Applications (IJCA) , Volume 110, Issue 5, January 2015.
5. Gagn'e, R. (1985). The conditions of learning and theory of instruction. (4th Ed.) New York: Holt, Rinehart & Winston.
6. McKenzie, K., Gow, K. & Schweitzer, R. (2004). Exploring first-year academic achievement through structural equation modelling. Higher Education Research & Development, 23, 1, 95-112.
7. Yang, B., & Lu, D. R. (2001). Predicting academic performance in management education: An empirical investigation of MBA success. Journal of Education for Business, 77, 15-21.
8. Robbins, S. B., Allen, J., Casillas, A., Peterson, C. H. & Le, H. (2006). Unravelling the differential effects of motivation and skills, social and self-management measures from traditional predictors of college outcomes. Journal of Educational Psychology, 98, 3.
9. Tam, M. (2002). Measuring the effect of higher education on university students. Quality Assurance in Education, 10, 4, 223- 228.
10. Smith, J. & Naylor, R. (2001). Determinants of degree performance in U.K. universities: a statistical analysis of the 1993 student cohort. Oxford Bulletin of Economics and Statistics, 63, 1, 29-60.
11. Carroll, C. A., Garavalia, L. S. (2004). Factors contributing to the achievement of pharmacy students: Use of the goal-efficacy framework. American Journal of Pharmaceutical Education, 68, 4, 88.
12. Ofori, R. & Charlton, J. P. (2002). Issues and innovations in nursing education: A path model of factors influencing the academic performance of nursing students. Journal of Advanced Nursing, 38, 507-515.
13. Zeegers, P. (2004). Student learning in higher education: a path analysis of academic achievement in science. Higher Education research & Development, 23, 1, 35-56.
14. Kaberia, F. (2006). Enhancing quality and relevance in university education (Kenya). Paper presented in the fourth exhibition by Kenyan Universities, Nairobi: Kenya.
15. Niranjan Lal, Shamimul Qamar , Monika Kalra, "K- Mean Clustering Algorithm Approach for Data Mining of Heterogeneous Data" Information and Communication Technology for Sustainable Development{ICT4SD), LNNS, Springer Proceeding , Volume 10, pp.61-70 2017.
16. Niranjan Lal, Navneet Kaur, "Clustering of Social Networking Data using SparkR in Big Data" Springer Nature Singapore Pte Ltd. 2018, Communications in Computer and Information Science (CCIS), Volume 906, pp. 217–226, Nov, 2018.
17. Bouras, "A Survey of Clustering Algorithms for Big Data: Taxonomy & Empirical Analysis," IEEE transactions on emerging topics in computing, (2014).
18. S.ARORA, I.CHANA, "A survey of clustering techniques for Big Data analysis," in Confluence The Next Generation Information Technology Summit (Confluence), 5th International Conference-. IEEE, p. 59-65, (2014).