**Comparative Study of Data Mining Algorithms in Higher Education**

1. ApoorvaVerma 2. Dr. UshaBadhera

Lecturer Lecturer

S.S. Jain Subodh P.G. College S.S. Jain Subodh P.G. College

**Abstract**: -In the current digital era,data mining is extensively applied in various fields like marketing, agriculture, medicine, engineering and education etc. Data mining is one of the most promising and emerging technique being used in educational data mining. Presently universities, colleges and institutes are focusing more on quality education by tracing the academic performance of students. The various algorithms of data mining are being used in education to generate the results which help colleges and universities to draw better decisions that can be incorporated by academic planners.In this paper we have compared clustering algorithms K-means and K-medoids by applying them on sample datasets of undergraduatestudents, for this comparison we have implemented K-medoids and K-means clustering algorithms in Python. The results obtained prefer K-medoids over K-means cluster algorithm. This study also illustrates the practical usage of these clustering algorithms in education data mining

**Keywords: -**Educational Data Mining, Clustering, K-means, K-medoids, Academic Performance, Python.

1. **Introduction: -**Educational institutes are operating in a very complex and highly competitive environment. The main challenge for higher education institutes nowadays is to maintain and monitor the quality and the pace of education.This requires means of analyzing overall performance of their students so that they can improve further. The ability to predict/classify a student’s performance is an important aspect in educational environments.A very promising arena to attain this objective is the use of Data Mining. Data mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems. Analysts are extensively applying data mining in educational field for analyzingstudent’s performance. In the field of education, data mining techniques is used to generate and identify patterns which can be useful both for the learners and for the educators. Educational data mining (EDM) describes a research field concerned with the application of data mining, machine learning and statistics to information generated from educational settings (e.g., universities and intelligent tutoring systems). EDM refers to tools and research defined for extracting information from large repositories of data generated from learning activities in educational settings. EDM is the process of transforming raw data obtained from educational systems into useful data that can be used to make data-driven decisions. The figure1 describes the steps of data mining which can be applied in educational data mining. The foremost step is are

[Figure 1- Steps of Data Mining in EDM]

extracting or collection of dataset followed by cleaning dataset i.e. removing noise and preparing data, third step involves deciding which clustering technique should be used and last is evaluation of clustering technique applied.

The main contribution of this paper is to demonstrate the application of two important algorithms of data mining i.e. K-means and K-medoid for EDM and to draw a comparative analysis between these two algorithms.

This paper is organized as follows: Section II explains Overview of Data mining. Section III provides a review of related work. Our experimental methodology is explained in Section IV. Section V describes the algorithm applied and work carried out. Comparative analysis is in Section VI.

1. **Overview: Data Mining: -**Han and Kamber (2006) define data mining as the process of discovering ‘hidden images’, patterns and knowledge within large amount of data and making predictions for outcomes or behaviors. Data mining uses a knowledge base, analytical skills and domain knowledge to discover new trends and patterns. These trends and patterns are used to design predictive models which help the researchers to find new observations from existing dataset. Education is the most important factor for the progress of the country. It helps the people of the country to become well mannered, cultured and wise citizen of the country. Mining of educational data is known as educational data mining. Clustering is most widely used technique of data mining used for analysis and model prediction.Clustering is deals with finding or grouping unstructured data into particular groups based on grouping variable. Cluster can be defined as collection of objects having similar properties and objects of one cluster do not share any common property with objects in other clusters.

There are various techniques of clustering as shown in the below classification diagram

[Figure 2- Overview of clustering taxonomy]

**III. Related Research: -**Several researchers are working in the field of educational data mining to predict the learner’s performance academically. Carlos, Cristobal and Sebastian analyzed the data of 670 middle-school students by applying white box classification methods and predicted the drop out and failure rate with accuracy of over 90%[5]. In the very first papers of application of Data Mining techniques to Higher Education Luan, Jingdiscussvarious potential applications including marketing, alumni fund raising, to survival analysis, persistence and many others[6].

Pandey and Pal try to find out the adept teacher dealing with Students by taking a psychometric test that converts qualitative variables into quantitative and further applying association rules[7]. Umesh Kumar Pandey, Brijesh Kumar Bhardwaj, and Saurabh pal in their paper, Data Mining as a Torch Bearer in Education Sector have portrayed the roadmap of research done in EDM owing to various education sector segments[8]. The research was mostly done in order to get to inferences which may be used to improve the quality of education.

**IV. Research Methodology:-**The dataset of students pursuing collected. It was analyzed using K \_means and K-medoids algorithm to draw the results and comparative analyses of these algorithms based on performances of students. Structure of the data set used for research.

|  |  |  |
| --- | --- | --- |
| **S.no.** | **Column Name** | **Attributes** |
| 1 | Name | Contains name of the student |
| 2 | Father’s Name | Contains Father’s Name of the student |
| 3 | Gender | Contains two values(M=Male & F=Female) |
| 4 | Board | Contains two values(RBSE and CBSE) |
| 5 | Reg. No. | Contains registration number of the student |
| 6 | Course | Course being pursued by student |
| 7 | 12thResult | Percentage obtained by students in 12th |
| 8 | 1stYear | Percentage obtained by student in first year |

[Table1- Structure of Dataset used for Research]

The figure below gives a diagrammatic approach of the methodology used in the paper

[Figure 3- Experimental Procedure]

On this data set we have applied K-Means algorithm first and then K-medoid algorithm and then by the results obtained we have made a comparative analysis between the two algorithms.

**V. Methodology**

**K-Means Algorithm:-**

K means algorithm is the simplest unsupervised learning algorithm used to solve clustering problems. It does so by dividing the objects and making clusters of the object having similar properties. 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 Euclidean distance between an object and all the nearby centroid is calculated as per the formula given below

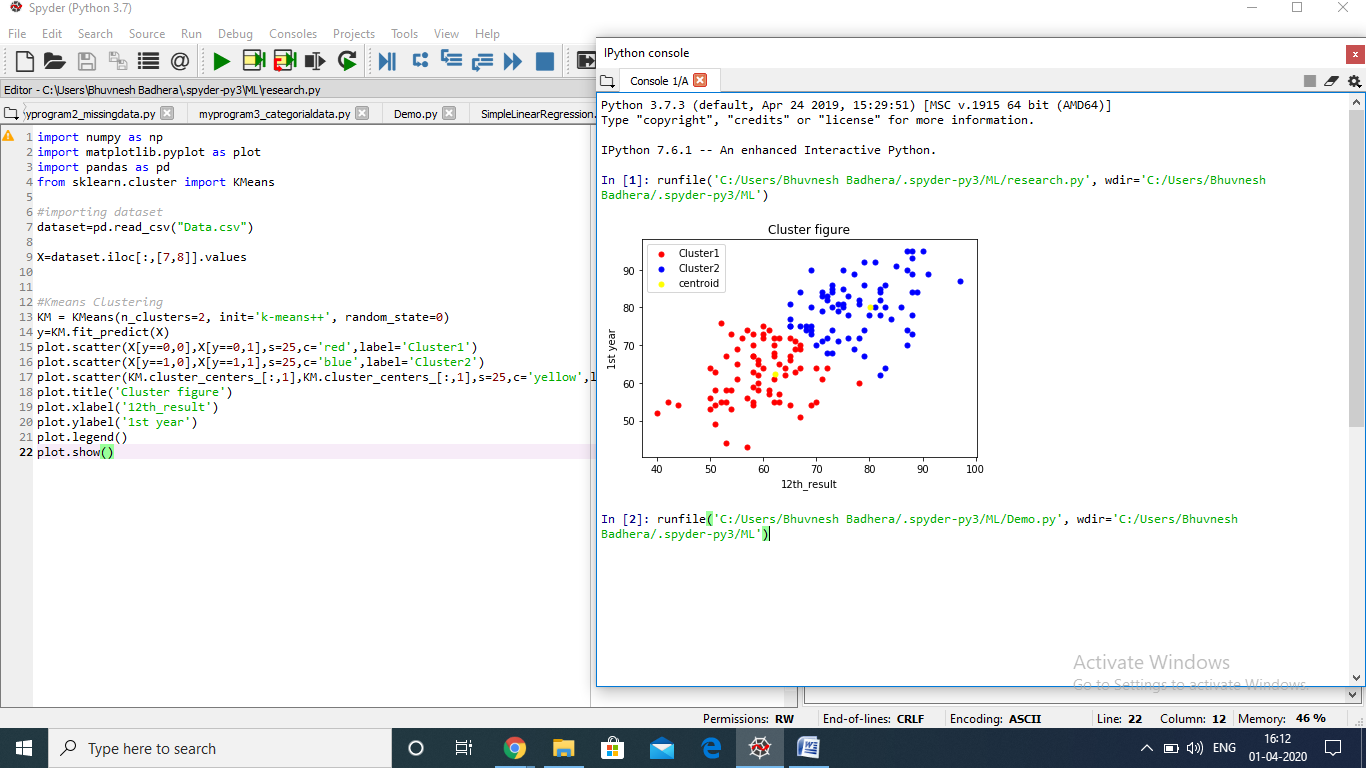
Where ||Xi (j) – Cj ||2 is the nearest distance measure between a data point xij and the Centroid Cj, and it indicates the distance between data points from their Centroid. The time complexity of the K-means algorithm is subjected to the formula; O (ndk+1).

Two prior requirements of the algorithm

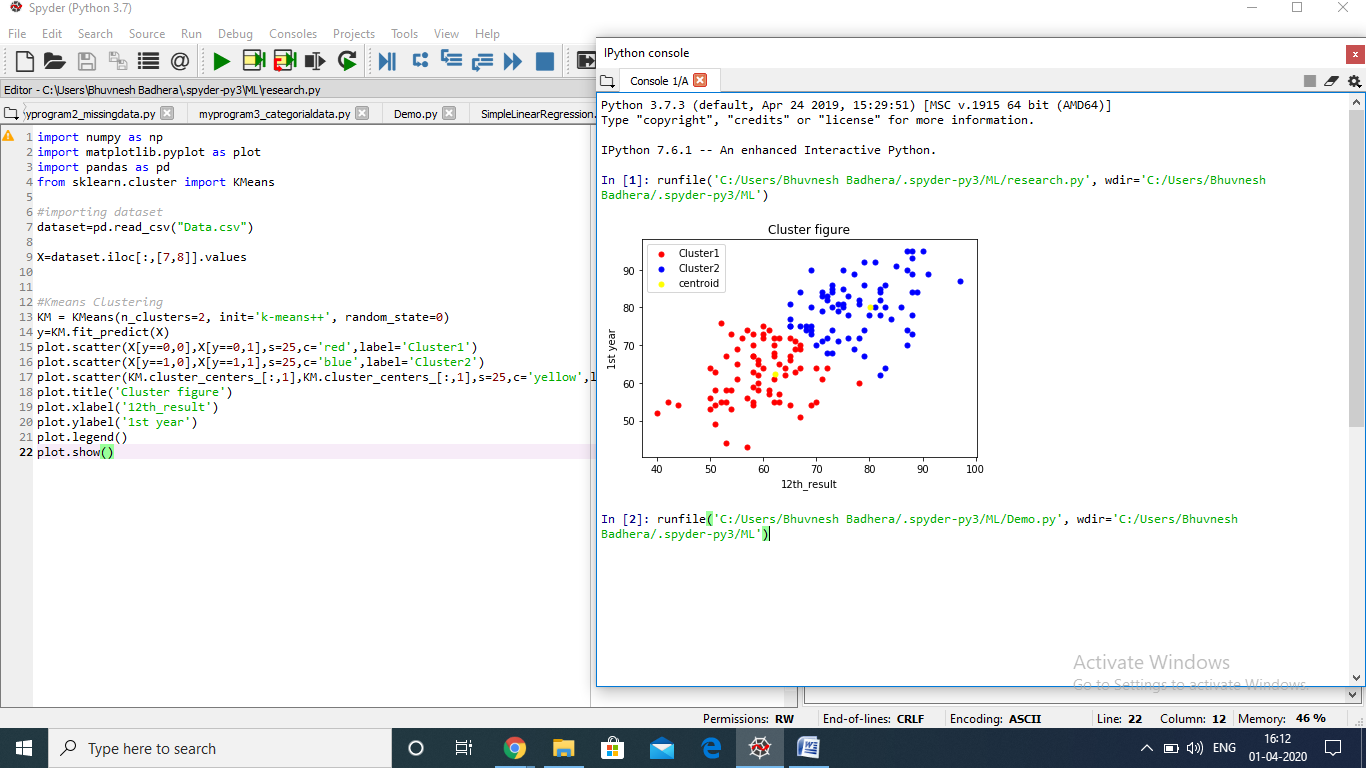
1. The data points
2. Value of k- the number of clusters.

In this paper we have implemented the algorithm on the student data set .We have used python language for the same.

1. We imported our data file(DATA.csv)
2. Then we created data frames.
3. Then we applied elbow method to find the optimum number of clusters, i.e. Value of k.
4. Then using the K means algorithm we generated the clusters which is shown below in the figure.



[Figure 4- Clustering Figure(1)]

****

**[** Figure 5- Clustering Figure obtained applying K-Means algorithm using Python]

The above two figures describes the clusters obtained using K means algorithm using Python language.

From the above figure we can see that red dots indicate the first cluster and the blue color dots indicate second cluster. The yellow color dots indicate the centroid value generated using the algorithm based on which clusters have been generated.

**K-Medoid Algorithm:-**

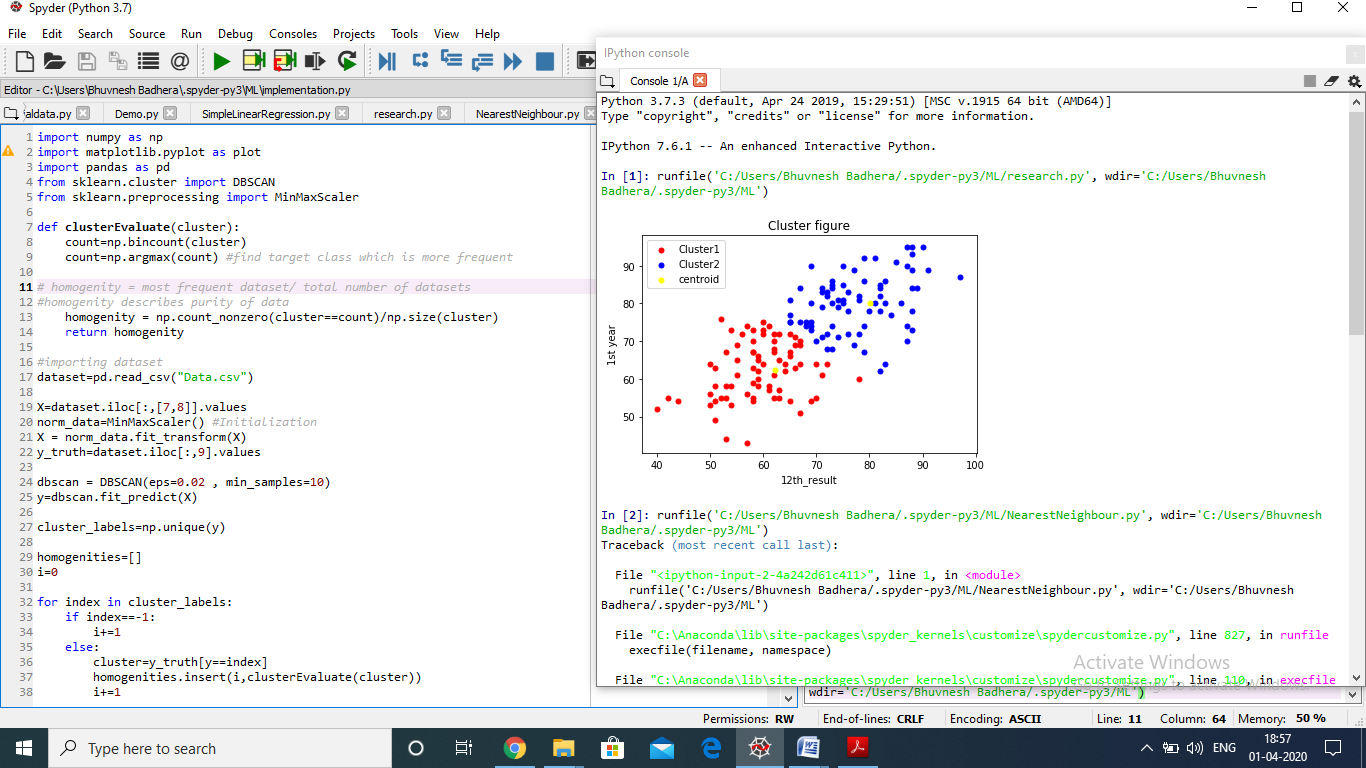
K medoid or partitioning around medoid (PAM) is the clustering algorithm. The method was proposed in 1987.A medoid can be defined as the object of a cluster whose average dissimilarity to all the objects in the cluster is minimal, that is, it is a most centrally located point in the cluster. This method minimizes the absolute distance between the data points and the centroid value rather than the square value thus k medoid is more robust to noise and outliers as compared to k-means.

The distance is calculated by the given formula

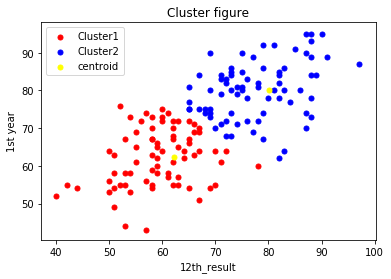
The time complexity is O(k(n-2)2).

We have used python language for implementing the k-medoid.

1. Imported the data file(Data.csv)
2. Applied elbow method to determine value of k
3. Determined the homogeneity and the value of eps
4. Applied the k medoid and generated the clusters.



[Figure6- Clustering Figure obtained by applying K- Medoid algorithm using Python]



[Figure 7- Clustering Figure]

The above two figures describes the clusters obtained using K medoid algorithm using Python language. The red dots represent first cluster and the blue dots represent the second cluster and yellow dots represent the centroid value.

**VI.Conclusion:-** This paper involves the implementation of the k means and k medoid algorithm on student data set for predicting which algorithm is much better for educational data mining between the two. For comparison of the algorithms, python code has been implemented. The reason for using python language is that it provides benefits like extensive supported libraries, user friendly data structures, learning ease and support available. The results obtained are in favour of K medoid which is better at clustering data and for educational data mining. The results k medoid are more flexible and more robustK medoid is more expensive and this is the main drawback of this algorithm.

**References:-**

1. A.Fahad, N. Alshatri, Z. Tari, A. Alamri, I. Khalil, A. Zomaya, S. Foufou, A. Bouras, "A Survey of Clustering Algorithms for Big Data: Taxonomy & Empirical Analysis" in TRANSACTIONS ON EMERGING TOPICS IN COMPUTING 2014 10.1109/TETC.2014.2330519, IEEE Transactions on Emerging Topics in Computing.
2. Aksenova, S. Sretlana., Zhang, Du and Lu, Meilin (2006) “Enrolment prediction through Data Mining” IEEE Conference Publications: 510-515.
3. Cios, K.J., Swiniarski, W.R., Pedrycz, W. and Kurgan, A. L., “Data Mining(2007): A Knowledge Discovery Approach” , Springer, New York.
4. ParneetKaura,ManpreetSinghb,Gurpreet Singh Josan “Classification and prediction based data mining algorithms to predict slow learners in education sector” in 3rd International Conference on Recent Trends in Computing 2015(ICRTC-2015), Elsevier Procedia Computer Science.
5. Cristóbal Romero, Sebastián Ventura, Pedro G. Espejo and César Hervás “ Data Mining Algorithms to Classify Students” in The 1st International Conference on Educational Data Mining Montréal, Québec, Canada, June 20-21, 2008 Proceedings.
6. Luan, Jing “Data Mining and Knowledge management in Higher Education - Potential Applications”, in Annual Forum for the Association, for Institutional Research (42nd, Toronto, Ontario, Canada, June 2-5, 2002).
7. Umesh Pandey, PSRIET, Sourabh Pal, “Mining Data to Find Adept Teachers in Dealing with Students” in I.J. Intelligent Systems and Applications, 2012, 3, 27-33.
8. Ramaswami, M., R. Bhaskaran. A CHAID Based Performance Prediction Model in Educational Data Mining. – IJCSI International Journal of Computer Science Issues, Vol. 7, January 2010, Issue 1, No 1, 10-18.
9. Umesh Kumar Pandey, Brijesh Kumar Bhardwaj, Saurabh pal “Data Mining as a Torch Bearer in Education Sector “ in Technical Journal of LBSIMDS .
10. Jaya Srivastava Ph.D. Scholar, Jaipur National University, Jaipur DrAbhay Kumar Srivastava Assistant Professor, Jaipuria Institute of Management, Lucknow “Data Mining in Education Sector: A Review” in Special Conference Issue: National Conference on Cloud Computing & Big Data
11. DorinaKabakchieva “Predicting Student Performance by Using Data Mining Methods for Classification” in CYBERNETICS AND INFORMATION TECHNOLOGIES , Volume 13, No 1,2013
12. Alok Singh Chauhan, Varun Singh Chauhan, Narendra Kumar Sharma “Application of K-means Clustering Algorithm for Mining Educational Data” in InternatIonal Journal of Computer Science and technology vol 3, Issue 4, 2012
13. Shyamala K., Rajagopalan S. P “Data Mining Model for a better Higher Educational System” in Information Technology Journal Volume 5 (3): 560-564, 2006
14. Dr. M Rajeshwari,S. Sathya,S.Deepika- Analysis of students Performance based on factors using K means.
15. MahendraTiwari, Randhir Singh “Comparative investigation of K means and K medoid algorithm on Iris data” in International Journal of Engineering Research and Development Volume 4, Issue 8 (November 2012), PP. 69-72.
16. <https://en.wikipedia.org/wiki/Data_mining>
17. <http://educationaldatamining.org/>
18. <https://www.geeksforgeeks.org/ml-k-medoids-clustering-with-example/>