International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Impact Factor 2025: 7.089

Predicting Semester Performance of College Students Using Data Mining Techniques and WEKA Analysis

M. Naresh[1], K. Murali[2], B. Mamatha[3], S. Upendra[4], K. Sreenivasulu[5], C. Dwarakanath Reddy[6], B. Sarojamma[7]

[1]Department of Statistics, S V University, Tirupati, India

[2]Academic Consultant, Department of Statistics, S V University, Tirupati, India

[3]Lecturer, Department of Mathematics, Sri Padmavathi Women's Degree & PG College, Tirupati, India

[4]Research Scholar, Department of Statistics, S V University, Tirupati, India

[5]Department of Statistics, D.K. Govt. College for Women (A), Nellore, India

[6]Assistant Professor, Department of Statistics, Siddartha Academy Group of Institutions, Tirupati, India

[7]Professor, Department of Statistics, S V University, Tirupati, India
Corresponding Author Email: *saroja14397[at]gmail.com*

Abstract: ***Over the years, several statistical tools have been used to analyse and predict students' performance from different point of view. One of the biggest challenges for higher education. Today is to predict the paths of students through the educational process. Successful students' result prediction in early course stage depends on many factors. Data mining techniques could be used for this kind of job. Data mining techniques are widely used in educational field to find new hidden patterns from student's data. The hidden patterns that are discovered can be used to understand the problem arise in the educational field. Data Mining (DM), or Knowledge Discovery in Databases (KDD), is an approach to discover useful information from large amount of data. Data mining techniques apply various methods in order to discover and extract patterns from stored data Based on collected students' information; different data mining techniques need to be used. For the purpose of this project WEKA data mining software is used for the prediction of semester wise student's marks based on parameters in the given dataset. The dataset contains information about different students from one college of 5 courses in the overall semesters.***

Keywords: Data mining techniques, Clustering methods: Canopy, EM, Hierarchical, Simple-k means.

1. Introduction

Now a days, data mining is playing a vital role in educational institutions and one of the most important areas of research with the objective of finding meaningful information from the data stored in huge dataset. Educational data mining (EDM) is a very important research area which helpful to predict useful information from educational database to improve educational performance, better understanding and to have better assessment of the students learning process. Data Mining or knowledge discovery has become the area of growing significance because it helps in analyzing data from different perspectives and summarizing it into useful information.

2. Methodology

There are numerous Machine learning algorithms in the literature and some of the important and popular algorithms based on Clustering.

2.1 Canopy Clustering:

Canopy Clustering is a very simple, fast and surprisingly accurate method for grouping objects into clusters. All objects are represented as a point in a multidimensional feature space. The algorithm uses a fast approximate distance metric and two distance thresholds T1 > T2 for processing.

Step-1: Go to weka software
Step-2: Select data by using open file.
Step-3: Go to cluster and select Canopy from choose option.
Step-4: Click start button to run the program.

2.2 EM Clustering:

For clustering, EM makes use of the finite Gaussian mixtures model and estimates a set of parameters iteratively until a desired convergence value is achieved.
Step-1: Go to weka software
Step-2: Select data by using open file.
Step-3: Go to cluster and select EM from choose option.
Step-4: Click start button to run the program.

2.3 Hierarchical Clustering:

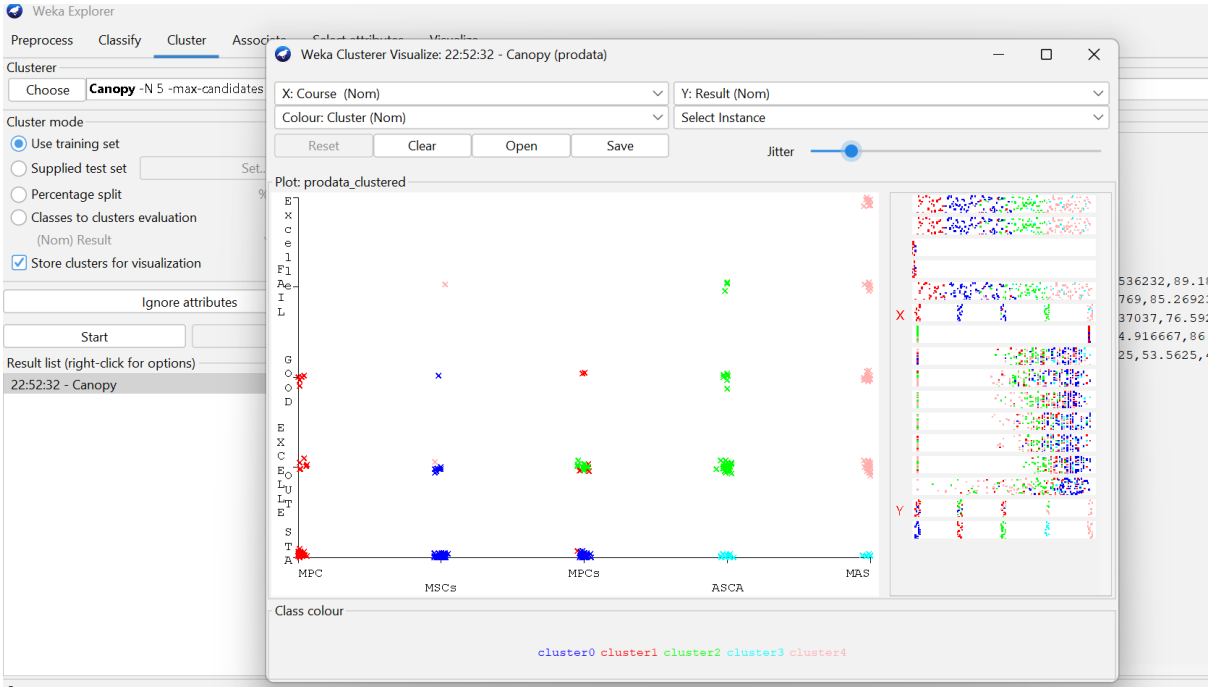
Hierarchical clustering is mainly focussing on building of hierarchy of clusters, i.e., cluster tree and it is represented in a dendrogram. It is either merging smaller clusters into

larger clusters or splitting larger clusters into smaller ones. A clustering of the data items is obtained through cutting a dendrogram at a desired level.
Step-1: Go to weka software
Step-2: Select data by using open file.
Step-3: Go to cluster and select Hierarchical from choose option.
Step-4: Click start button to run the program.

2.4 Simple-k means clustering:

K-means clustering is a simple unsupervised learning algorithm. In this, the data objects ('n') are grouped into a total of 'k' clusters, with each observation belonging to the cluster with the closest mean. It defines 'k' sets, one for each cluster k n (the point can be thought of as the centre of a one or two-dimensional figure). The clusters are separated by a large distance.
Step-1: Go to weka software
Step-2: Select data by using open file.
Step-3: Go to cluster and select simple k means from choose option.
Step-4: Click start button to run the program.

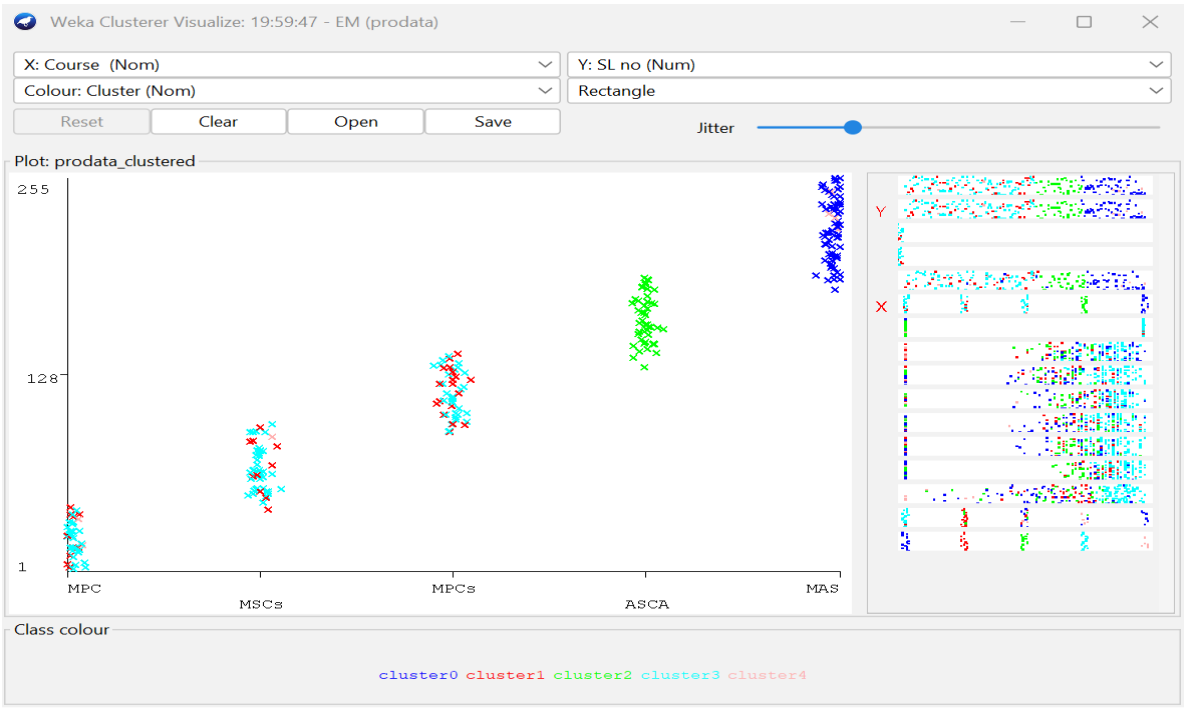
Emphirical Investigations:
Canopy Algorithm: This canopy algorithm is tested with project dataset in WEKA tool; it produces five different clustered instance clusters 0: 76(30%), cluster 1: 50(20%), cluster 2: 50(20%), cluster 3: 19(7%), cluster 4: 60(24%), and Time taken to build model (full training data): 0.01 seconds.



In the Cluster visualization we choose Course on X-axis and Result no on Y-axis.

Blue colour shows Cluster 0, Red colour shows Cluster 1, Green colour shows Cluster 2, Sky-blue colour shows Cluster 3, Light pink colour shows Cluster 4.

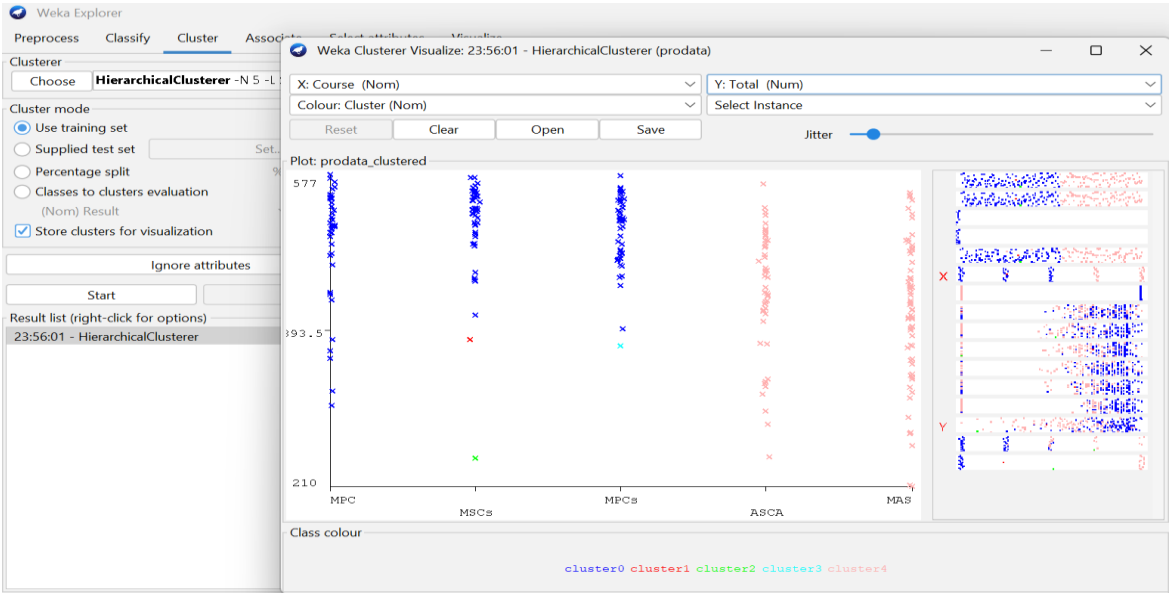
EM Clustering: This EM algorithm is tested with project dataset in WEKA tool. It produces five different clustered instance clusters 0: 63(25%), cluster 1: 38(15%), cluster 2: 49(19%), cluster 3: 98(38%), cluster 4: 7(3%), and Time taken to build model (full training data): 0.01 seconds.



In the Cluster visualization we choose Course on X-axis and SL no on Y-axis.

Blue colour shows Cluster 0, Red colour shows Cluster 1, Green colour shows Cluster 2, Sky-blue colour shows Cluster 3, Light pink colour shows Cluster 4.

Hierarchical Cluster: This Hierarchical Cluster algorithm is tested with project dataset in WEKA tool; it produces five different clustered instance clusters 0: 136(53%), cluster 1: 1(0%), cluster 2: 1(0%), cluster 3: 1(0%), cluster 4: 116(45%).

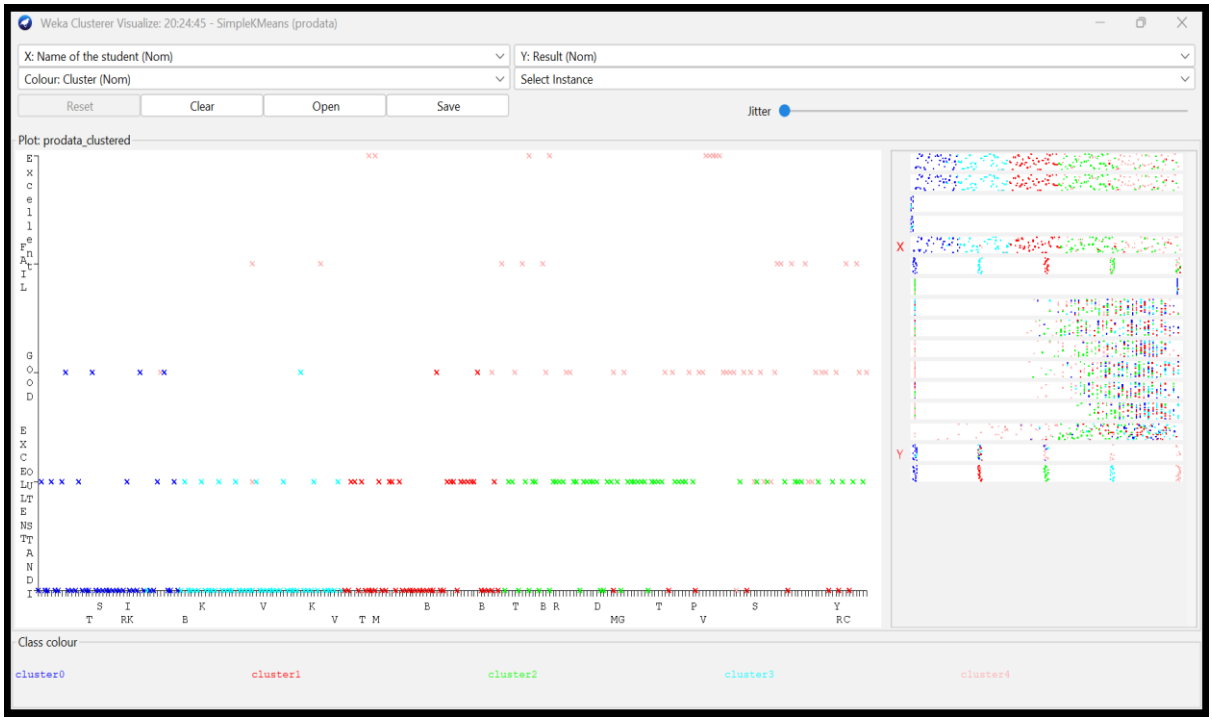


In the Cluster visualization we choose Course on X-axis and Total no on Y-axis.

Blue colour shows Cluster 0, Red colour shows Cluster 1, Green colour shows Cluster 2, Sky-blue colour shows Cluster 3, Light pink colour shows Cluster 4.

Simple-k means clustering: K means clustering is a simple cluster analysis method. The number of clusters can be set using the setting tab. The centroid of each cluster is calculated as the mean of all points within the clusters. With the increase in the number of clusters, the sum of square errors is reduced. The objects within the cluster exhibit similar characteristics and properties. The clusters represent the class labels.

This Simple -k meansClusterer algorithm is tested with project dataset in WEKA tool; it produces five different clustered instance clusters 0: 40(16%), cluster 1: 56(22%), cluster 2: 56(22%), cluster 3: 49(19%), cluster 4: 54(21%).



In the Cluster visualization we choose Name of the student on X-axis and Result no on Y-axis. Blue colour shows Cluster 0, Red colour shows Cluster 1, Green colour shows Cluster 2, Sky-blue colour shows Cluster 3, Light pink colour shows Cluster 4.

In all four-algorithm result is generated on the basis of similar objects and time to create that clusters. The Best algorithm found as Simple k-Means clustering. It is taking less time and good accuracy than other clustering algorithm to find similar clusters through weak tool for student data set.

The following table represents the analysis process of all four algorithms and the accuracy results are shown in table.

Table 2

Clustering algorithm	No. of Clusters	Clustering instances
Canopy	5	76(30%) 50(20%) 50(20%) 19(07%) 60(24%)
EM	5	63(25%) 38(15%) 49(19%) 98(38%) 7(3%)
Hierarchical	5	136(53%) 1(0%) 1(0%) 1(0%) 116(45%)
Simple-k means	5	40(16%) 56(22%) 56(22%) 49(19%) 54(21%)

The above table, the Simple-k means clustering algorithm and EM Clustering Algorithm are showing good accuracy than other cluster algorithm.

We have performed analysis with four clustering algorithms are Canopy clustering algorithm, EM Algorithm, Hierarchical Algorithm and Simple k-Means clustering algorithm. In all four-algorithm result is generated on the basis of similar objects and time to create that clusters. These Best algorithms getting the best accuracy in a short time for student dataset.

In accordance with the obtained results, it can be told that, Canopy clustering algorithm, EM Algorithm, Hierarchical Algorithm and Simple k-Means clustering algorithms are the most proper clustering method for evaluation of the students'

performances in educational data mining. For future studies, these applications can be experienced on various educational datasets with recently developed algorithms.

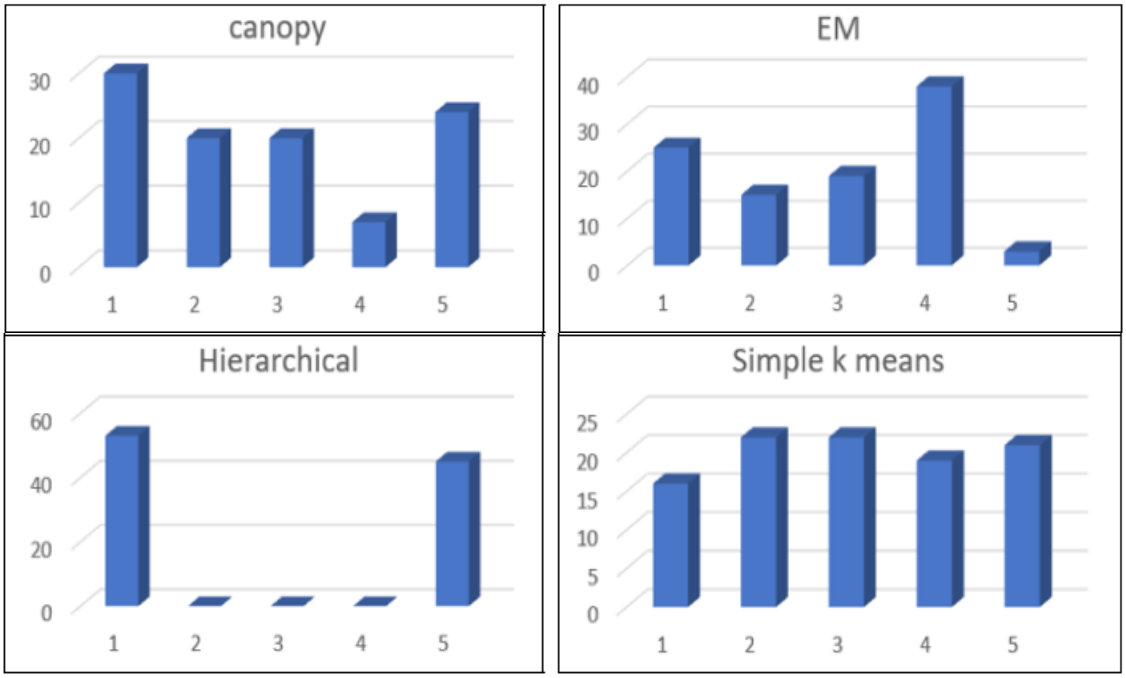


Figure 25

The above chart shows the performance accuracy of the four-clustering based on different clustering metrics. This metrics shows that Simple-k means, EM algorithms are performing good accuracy result better than other clustering.

3. Summary and Conclusions

We have performed analysis with four clustering algorithms are Canopy clustering algorithm, EM Algorithm, Hierarchical Algorithm and Simple k-Means clustering algorithm. In all four-algorithm result is generated on the basis of similar objects and time to create that clusters. These Best algorithms getting the best accuracy in a short time for student dataset.

In accordance with the obtained results, it can be told that, Canopy clustering algorithm, EM Algorithm, Hierarchical Algorithm and Simple k-Means clustering algorithms are the most proper clustering method for evaluation of the students' performances in educational data mining. For future studies, these applications can be experienced on various educational datasets with recently developed algorithms.