Clustering and Visualization of Students Academic Performance using k-means ++ and PCA

Rahul Saini[1*], RitikSahu[1], Sourabh Dave[1] and Sachin Solanki[1]

[1]*Department of Computer Science and Engineering, Medi-Caps University, Indore-453331, M.P., India.*

[*]*Corresponding authors email: rahulss2899@gmail.com*

Abstract :. Clustering is an Unsupervised Machine Learning technique that involves the grouping of data points. We can use a clustering method to classify each data point given a set of data points into homogeneous-subgroup. This project focuses on the grouping of different students based on their performance in different fields. The system groups students into different clusters in which each cluster contains students that shares the most similar performance level. In this paper, we used K- means ++ Algorithm to train and group the students into clusters. The output when visualize with the help of PCA (Principal Component Analysis) shows efficient clusterization of the data points. The model uses the ensemble learning technique hence, both hierarchical as well as k- means clustering has been used to predict the best possible clusters.

Keywords: Clustering, K means ++, PCA.

Introduction

Clustering is a widely used technique used to analyse un-labelled data. It can be performed by dividing the data sets into different clusters based on their features. Each cluster contains a centroid point which defines the mean or average performance of that respective cluster.

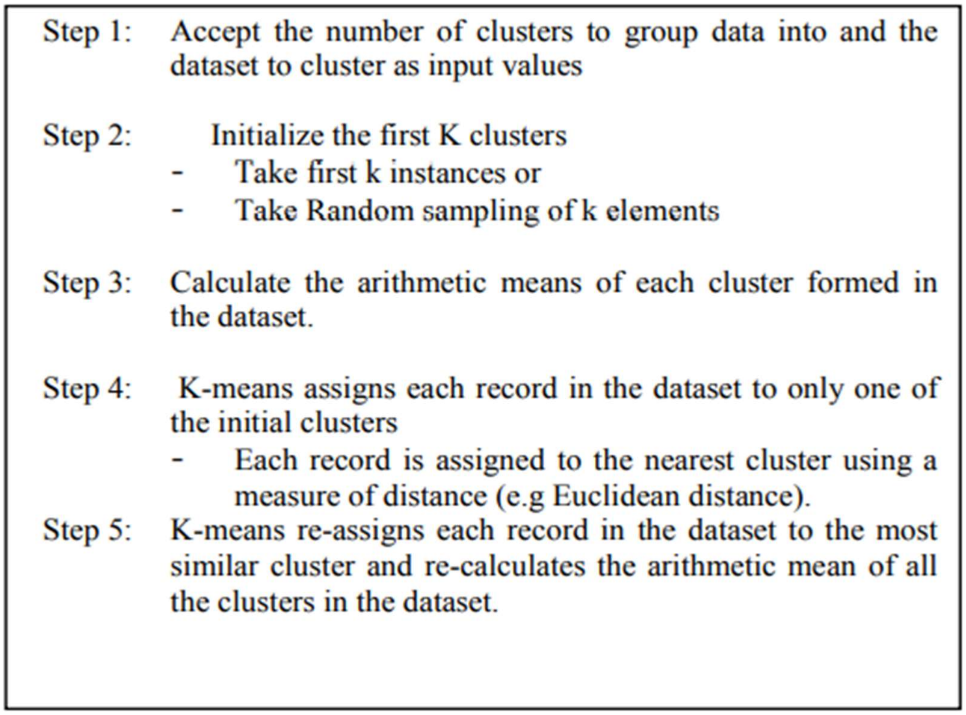
The motive of creating this model is to solve the general problem faced by universities during admitting thousands of students. If a university gave admission to a student that depends upon more than 5 criteria then it is practically very hard to differentiate between two students. To understand this we can take an example as if we consider two students who are interested in a particular university and university consider performance in a total of 15 criteria like marks in high school, scholarship, and marks in other recognized exams. Now if both the students perform above 90% in all the 14 criteria but one student underperforms in only one exam, then it is unfair to judge his discernment based on just one underperformance. There can be many combinations in which students can argue over admittance, hence our model maps the higher dimensional data into lower dimensional data and groups students on overall performance instead based on subject wise.

Clustering can be classified into hierarchical and non-hierarchical clustering. The hierarchical clustering is further divided into single linkage, complete linkage, median and ward. While non-hierarchal clustering includes k-means, k means ++ and k medoids. According to the characteristics of datasets, we choose which algorithm to use that give a favourable result. Moreover, simulation studies can be performed to measure the invariability of clusters [4].

Expectation-Maximization is the approach followed by the k-means algorithm to rectify the problem faced in dealing with unlabelled datasets. The model is written in python language and uses libraries and functions provided by sklearn, NumPy and Pandas. As it is an unsupervised learning algorithm it does not require prior testing of data and works very efficiently. For visualization PCA (Principal Component Analysis) is used as it reduce the dimensionality of the data to be able to observe in a 2-d plane. As clustering is an unsupervised learning algorithm one can only measure the accuracy by how good the overlapping boundaries are set. To measure how well our model has clusterized the data silhouette width[4] or the homogeneity index is used that asses the quality of separation using the clustering algorithm[3].

Proposed Methodology

K-means algorithm dissolute the dataset into **Kpre-defined** non-overlapping subgroups in which subsequent data point belongs to a particular cluster or group. It groups the data points which are most similar while also trying to make as many as clusters possible.It groups data points into clusters with the goal of minimising the mean squared distance between the data points and the cluster's centroid. To make our data points more homogeneous within the clusters one has to abate the variance between clusters. The succeeding steps will summarize the function of the k-means algorithm.



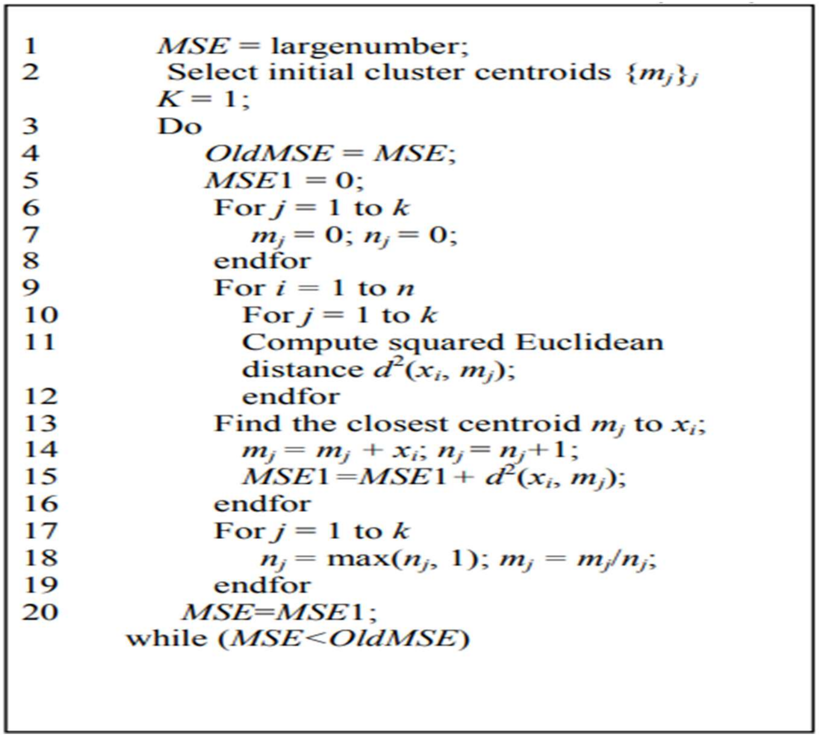
```
Step 1:  Accept the number of clusters to group data into and the
         dataset to cluster as input values

Step 2:     Initialize the first K clusters
         -  Take first k instances or
         -  Take Random sampling of k elements

Step 3:  Calculate the arithmetic means of each cluster formed in
         the dataset.

Step 4:   K-means assigns each record in the dataset to only one of
         the initial clusters
         -  Each record is assigned to the nearest cluster using a
            measure of distance (e.g Euclidean distance).
Step 5:  K-means re-assigns each record in the dataset to the most
         similar cluster and re-calculates the arithmetic mean of all
         the clusters in the dataset.
```

Fig 1-Generalised Pseudocode of K means algorithm [11]



```
MSE = largenumber;
 Select initial cluster centroids $\{m_j\}_j$
K = 1;
Do
    $OldMSE = MSE$;
    $MSE1 = 0$;
     For $j = 1$ to $k$
        $m_j = 0$; $n_j = 0$;
     endfor
     For $i = 1$ to $n$
       For $j = 1$ to $k$
       Compute squared Euclidean
       distance $d^2(x_i, m_j)$;
       endfor
     Find the closest centroid $m_j$ to $x_i$;
       $m_j = m_j + x_i$; $n_j = n_j + 1$;
       $MSE1 = MSE1 + d^2(x_i, m_j)$;
     endfor
     For $j = 1$ to $k$
        $n_j = \max(n_j, 1)$; $m_j = m_j/n_j$;
     endfor
  $MSE = MSE1$;
while ($MSE < OldMSE$)
```

Fig 2- K means Algorithm [6]

As the K means algorithm is vulnerable to initialization trap a more versatile algorithm k means ++ is used in this model which allows the model to select the initial points randomly and consider the best of it. The model will hence allow us to categorize the students into different clusters, and the number of clusters can be decided by using the elbow method shown in fig.3. It can be observed from fig 3 that the graph shows a significant drop in the value of WCSS (Within- Cluster Sum of Square) till cluster 2 and we can observe an elbow like shape emerging between 2 and 3, hence we can assume that 3 is a good number cluster for the model.

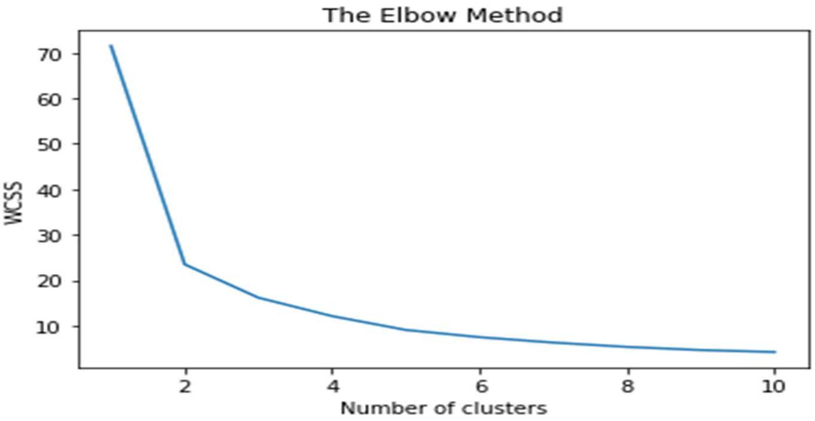


Fig 3- The Elbow Method

For visualization of the data i.e, the students after clustered are done by using PCA algorithm as the criteria for choosing students can be higher in number so it is not possible to visualise without reducing the dimensionality. We can use the PCA algorithm only to visualize the test result as if we use it as a standard machine learning algorithm it will somehow compromise the accuracy of our test result. It uses the correlation between some dimensions and tries to provide a minimum number of variables that keeps the maximum amount of variation or information about how the original data is distributed. Fig 5 is the output of the algorithm on a particular dataset.

The model is also flexible if a particular university wants to admit students while focusing on performance in particular exams we can add weights to the normalized data to concentrate the overall performance on those subjects. With the help of centroid as shown in fig 4, we can differentiate different clusters and classify students accordingly.

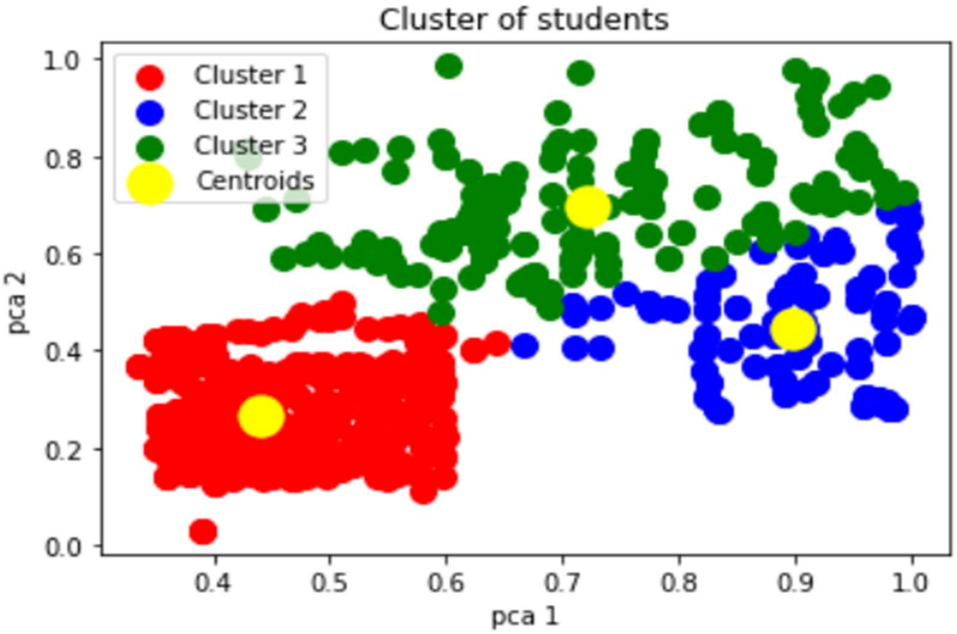


Fig 4- Cluster of Students

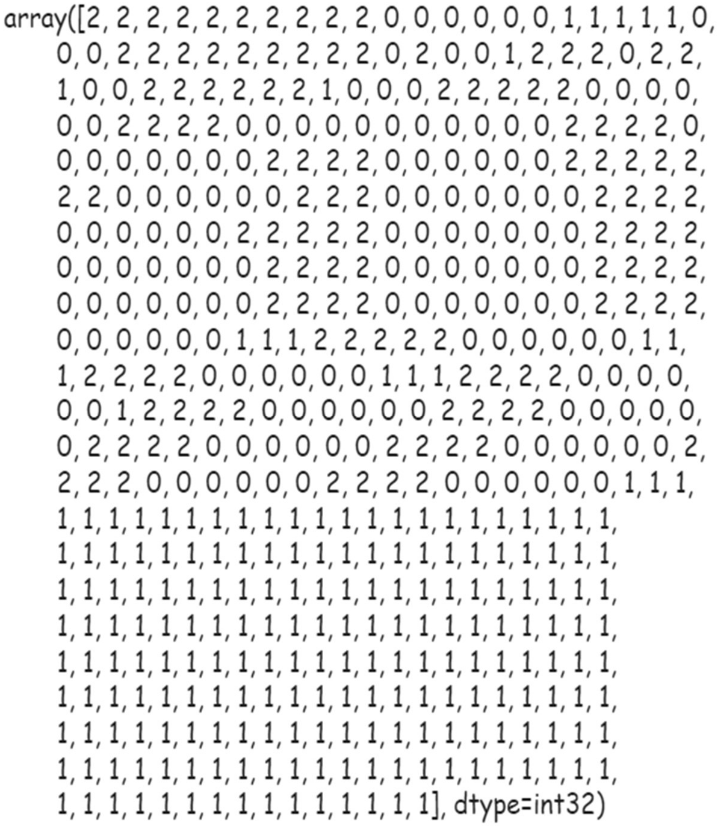
Results & discussion

The k means ++ algorithm performance between training and testing using a variety of parameters including running time, which is the amount of time it takes the software to complete the task. For checking how organized the clusters are formed we have used the Elbow method for selecting the number of clusters, and Silhouette analyses[10] which are used to determine the degree of separation between clusters.

Coefficient .: (b[i]-a[i]) / max(a[i],b[i])

The mean distance between all data points in the same cluster is a[i], whereas the mean distance between all data points in the nearest cluster is b[i].. Therefore, the more the coefficient is close to 1 the more efficient will be our model.

The training data set we have used to compute the clusters contains information of 900 students and their respective marks/performance in 15 exams/ criteria. The data set is further normalized to increase the algorithm efficiency. The output we get from this algorithm is a matrix that contains values from 0 to n-1 (where n is the number of clusters) as shown in fig 5.



```
array([2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 0,
       0, 0, 2, 2, 2, 2, 2, 2, 2, 2, 2, 0, 2, 0, 0, 1, 2, 2, 2, 0, 2, 2,
       1, 0, 0, 2, 2, 2, 2, 2, 2, 1, 0, 0, 0, 2, 2, 2, 2, 2, 0, 0, 0, 0,
       0, 0, 2, 2, 2, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 2, 2, 2, 0,
       0, 0, 0, 0, 0, 0, 0, 2, 2, 2, 2, 0, 0, 0, 0, 0, 0, 2, 2, 2, 2, 2,
       2, 2, 0, 0, 0, 0, 0, 0, 2, 2, 2, 0, 0, 0, 0, 0, 0, 0, 2, 2, 2, 2,
       0, 0, 0, 0, 0, 0, 2, 2, 2, 2, 2, 0, 0, 0, 0, 0, 0, 0, 2, 2, 2, 2,
       0, 0, 0, 0, 0, 0, 0, 2, 2, 2, 2, 0, 0, 0, 0, 0, 0, 0, 2, 2, 2, 2,
       0, 0, 0, 0, 0, 0, 0, 2, 2, 2, 2, 0, 0, 0, 0, 0, 0, 0, 2, 2, 2, 2,
       0, 0, 0, 0, 0, 0, 1, 1, 1, 2, 2, 2, 2, 2, 0, 0, 0, 0, 0, 0, 1, 1,
       1, 2, 2, 2, 2, 0, 0, 0, 0, 0, 0, 1, 1, 1, 2, 2, 2, 2, 0, 0, 0, 0,
       0, 0, 1, 2, 2, 2, 2, 0, 0, 0, 0, 0, 0, 2, 2, 2, 2, 0, 0, 0, 0, 0,
       0, 2, 2, 2, 2, 0, 0, 0, 0, 0, 0, 2, 2, 2, 2, 0, 0, 0, 0, 0, 0, 2,
       2, 2, 2, 0, 0, 0, 0, 0, 0, 2, 2, 2, 2, 0, 0, 0, 0, 0, 0, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1], dtype=int32)
```

Fig 5-Shows the output array of students

The centroid of each cluster is thus the mean performance of that particular cluster. Hence we can compute the average performance of each cluster while considering the centroid only. Fig 6 shows the number of students in 3 clusters.

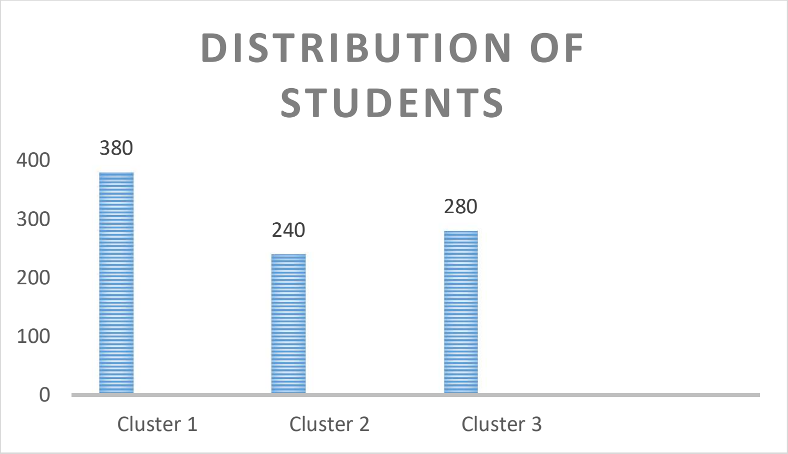


Fig 6-Distribution of Students

Conclusion

To categorize students per subject marks criteria were used earlier in which many students who performed excellent overall but underperform in one by a margin are not considered by the universities, also it was a tiresome task to compare students if the parameters of their comparison were too big. Therefore, to outrun these problems the current unsupervised machine learning model has been used which calculate the altogether performance of students. This model can visualize the data to make it possible for the people of admission departments and students to observe their position in the pool of different students.

References

[1] . Li, Youguo & Wu, Haiyan. (2012). A Clustering Method Based on K-Means Algorithm. Physics Procedia. 25. 1104-1109. 10.1016/j.phpro.2012.03.206.

[2]. S. Na, L. Xumin and G. Yong, "Research on k-means Clustering Algorithm: An Improved k-means Clustering Algorithm," 2010 Third International Symposium on Intelligent Information Technology and Security Informatics, 2010, pp. 63-67, doi: 10.1109/IITSI.2010.74.

[3].Rousseeuw P. J, "A graphical aid to the interpretation and validation of cluster analysis," Journal of Computational Appl Math, vol 20, pp. 53– 65, 1987.

[4]. Sharmir R. and Sharan R., "Algorithmic approaches to clustering gene expression data," In current Topics in Computational Molecular Biology MIT Press; pp. 53-65, 2002.

[5].K. P. Sinaga and M.-S. Yang, "Unsupervised *K*-means clustering algorithm," *IEEE Access*, vol. 8, pp. 80716–80727, 2020.View at: Publisher Site | Google Scholar

[6].M. S. Yang and K. P. Sinaga, "A feature-reduction multi-view *k*-means clustering algorithm," *IEEE Access*, vol. 9, p. 1, 2019.View at: Google Scholar

[7]. J. Song, X. Li, and Y. Liu, "An optimized *k*-means algorithm for selecting initial clustering centers," *International Journal of Security and Its Applications*, vol. 9, no. 10, pp. 177–186, 2015.View at: Publisher Site | Google Scholar

[8]. C. Xia, J. Hua, and W. Tong, "Distributed *K*-Means clustering guaranteeing local differential privacy," *Computers & Security*, vol. 90, pp. 101699.1–101699.11, 2020.View at: Publisher Site | Google Scholar

[9]. 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

[10]. Wang, Fei & Franco-Penya, Hector-Hugo & Kelleher, John & Pugh, John & Ross, Robert. (2017). An Analysis of the Application of Simplified Silhouette to the Evaluation of k-means Clustering Validity. 10.1007/978-3-319-62416-7_21.

[11]. Oyelade, Jelili & Oladipupo, Olufunke & Obagbuwa, Ibidun. (2010). Application of k Means Clustering algorithm for prediction of Students Academic Performance. International Journal of Computer Science and Information Security. 7.