PREDICTION OF STUDENTS' PERFORMANCE FOR PLACEMENT USING A CLUSTERING TECHNIQUE

Dr Kalpana Salunkhe[1], Dr Madhuri Prashant Pant[2]

Abstract
To make administrative decisions and deliver high-quality education, it is essential to analyze student academic performance in educational institutions. The amount of information relating to educational institutions is growing quickly. The management will be able to make academic decisions with the use of machine learning from these vast volumes of data. Predicting a student's academic success early on in their course will assist academia in identifying the merit students and in concentrating more attention on creating remedial programmes for the poorer students to boost their performance. This also helps in their placements. Placements are a very crucial point for all academic institutions. In this paper, the K Means clustering technique is used forcategorization of students' data.

Keywords – Machine learning, analysis, K-Means algorithm, Cluster, students' data, elbow method, students'data, prediction, centroid.

[1*]Sadhu Vaswani Institute of Management Studies for Girls, Pune, India, Salunkhekalpana96@gmail.com
[2]Vishwakarma University, Pune, Department of Computer Science

*__Corresponding Author:__ Dr Madhuri Pant
*Vishwakarma University, Department of Computer Science, Pune, India, pantmadhuri123@gmail.com

DOI: 10.48047/ecb/2023.12.si10.00225

1. Introduction

Most academics utilize classification or clustering algorithms to identify implicit patterns in educational data.

Supervisory and unsupervisory strategies are used in data mining. The use of supervisory approach is advantageous because the data already have established class labels. Contrarily, unsupervised methods do not have labeled data. Researchers took interest in clustering approaches. Market research, pattern recognition, data analysis, image processing, academic performance, and intrusion detection are just a few of the numerous applications where clustering is widely employed. [2]

Clustering or cluster analysis is a machine learning technique that groups an unlabeled data set. It can be defined as: "A method of grouping data points into different clusters consisting of similar data points." Objects with possible similarities are divided into a group that has less or no similarity to another group. It does this by looking for some similar pattern. in unlabeled data sets, such as shape, size, color, behavior, etc., and divides them according to the presence and absence of similar patterns. It is an unsupervised learning method, so the algorithm is unsupervised. It processes an unlabeled dataset. After applying this clustering method, each cluster or group is assigned a cluster ID. Machine Learning system can use this identifier to simplify the processing of large and complex data sets. [3]

If you have a bunch of anonymous data, it's very likely that you're using some sort of unsupervised learning algorithm. Clusters are especially useful when you're exploring data you know nothing about. It may take some time to figure out which type of clustering algorithm works best, but once you do, you'll gain valuable insight into your data. You may find relationships you never thought possible.

2. K-means Clustering Algorithm

It is the most used clustering algorithm. It is a centroid based algorithm and the simplest unsupervised learning algorithm. This algorithm tries to minimize the variance of the data points in a cluster. This is also how unsupervised machine learning is introduced to most people. K-means is best used for smaller data sets because it iterates over all data points. This means that it takes more time to classify the data points if there are many of them in the dataset.

2.1 K-Means Clustering Algorithm Implementation Method

K-Means is an unsupervised approach that is used to cluster data objects. The K-Means clustering algorithm divides the "n" data objects into "k" clusters (groups), with each data object assigned to the cluster with the closest mean. Each group's data items are extremely cohesive, while the objects in the other group are disjunct. Using the sum of squares, the K Means algorithm generates "k" different groups of elements. The algorithm's input parameter is the number of centroids. The distance between each element and its centroid is then calculated. The estimated distances between a data element and each centroid are compared, and the data element is assigned to the nearest centroid. As a result, each data element is assigned to one of the centroids. Initially, the unsupervised algorithm usedfor clustering data objects is K-Means. The K Means clustering algorithm divides "n" data objects into "k" clusters (groups), with each data object assigned to the cluster with the closest mean. Each group's data items are extremely cohesive, whereas the objects in the other group are disjunct.

The K-means algorithm uses the sum of squares to generate "k" different groups of elements. The number of centroids serves as the algorithm's input parameter. After that, the distance between each element and its centroid is computed. The estimated distances between a data element and each centroid are compared, and the data element is assigned to the closest centroid. As a result, each data element is allocated to a centroid. Initially. [7]

First, K clusters are constructed by allocating data elements to their respective centroids. The centroid of allocated data elements in each cluster is then recalculated. Calculate the distance between each data element with the new centroids once more and reassign the data element to the nearest centroid. This process is repeated until no data element is assigned to a new centroid, which means that the centroids of "n-1" iterations are equal to the centroids of "n" iterations.

2.2 In K-means clustering, the distance measure is Euclidean distance.

Assume the elements are:

X= (x1, x2, x3,...) and Y= (y1, y2, y3,...).

$D(X, Y) = (x1-y1)^2 + (x2-y2)^2 + \ldots$

$$\sqrt{\sum_{i=1}^{n}(xi - yi)} \quad \ldots\ldots\ldots\ldots \text{ eq. (1) [6][1]}$$

The distance between each actual or observed data point and the centroid is computed using eq. (1). The data element is then assigned to the

centroid.

with the shortest distance. The mean of all data points in that group is the centroid. Each centroid with a set of data components is referred to as a cluster.

2.3 K-means Algorithm

1. Accept as input values for the number of clusters to which data is arranged.
2. Create the first K clusters.

a. Select the first k instances; otherwise.
b. Select a random sample of k elements.

3. Determine the arithmetic means of each cluster in the dataset.
4. K-means assigns each record in the dataset to one of the initial clusters.

a. Using a distance measure (e.g., Euclidean distance), each record is assigned to the nearest cluster.
b. K-means reassigns each record in the dataset to the most similar cluster and recalculates the arithmetic mean of all clusters in the dataset.

2.4 K-Means Clustering Algorithm Implementation Using Python

Various committees, such as the NBA (National Board of Accreditation), NAAC (National Assessment and Accreditation Council) place a high value on students' academic performance. Various organizations that inspect colleges, like management institutions, place a high value on students' academic performance. The dataset for the experiment was collected from the college's Administration Department. The data set is initially normalized by manual verification. Below steps are taken by researcher to cluster the students' data.

Step 1

The researcher had considered the data of 180 students of MBA and MCA. Following is the code snippet to visualize the data using python.

```
import numpy as np
import pandas as pd
df=pd.read_csv('student1_clustering_modified.csv
')
df.head()
shape of data is (180, 2) Output
    cgpa  iq
0   5.13   88
1   5.90  113
2   8.36   93
3   8.27   97
4   5.45  110
     ..   ...  ...
175      8.46  98
176      8.94  115
177      5.87  108178   4.99   88
179      8.91  115
[180 rows x 2 columns]
```

Step 2

To determine the number of clusters to be formed, the Elbow method is used. In the elbow method, we calculate WCSS (Cluster Sum of Square). WCSS is the list of the square roots of sum of squares between each data value and the centroid value.

Here the researcher has considered the data of 180 students. Which is nothing but a student clustering.csv. For clustering purposes, we will use pyplot library. Library SKlearn cluster contains the KMeans algorithm.

To calculate the value for square root of sum of all squares of the difference between the actual value and centroid value for both the cgpa and iq, we will run the loop for all data. Here we need to create the km object and we pass the dataset as df.

Now the entire dataset is trained and now we can use one attribute km i e 'inertia_' to insert values into list wcss. wcss= km.insertia_

Python code for above is as below:

```
from sklearn.cluster import KMeans
wcss= []
#empty list
for i in range (1,11):
km=KMeans(n_clusters=i)
km.fit_predict(df)
wcss.append(km.inertia_)
wcss
#list containing square root of squares of variance.
Output
 [27422.967222222218,
 3837.5269999999996,
 2129.7680419177113,
 620.5295666996047,
 476.4367059085609,
 365.298649636924,
 290.4253176772653,
 220.86491930359497,
 183.61208650311045,
 165.80745936149512]
```

The above list shows that data is largest at top and rapidly decreases towards the bottom. import matplotlib. pyplot as plt

plt. Plot (range (1,11), wcss)

If we plot WCSS with k (no. of clusters), the curve looks like an elbow.

Output Fig 1: Elbow Curve

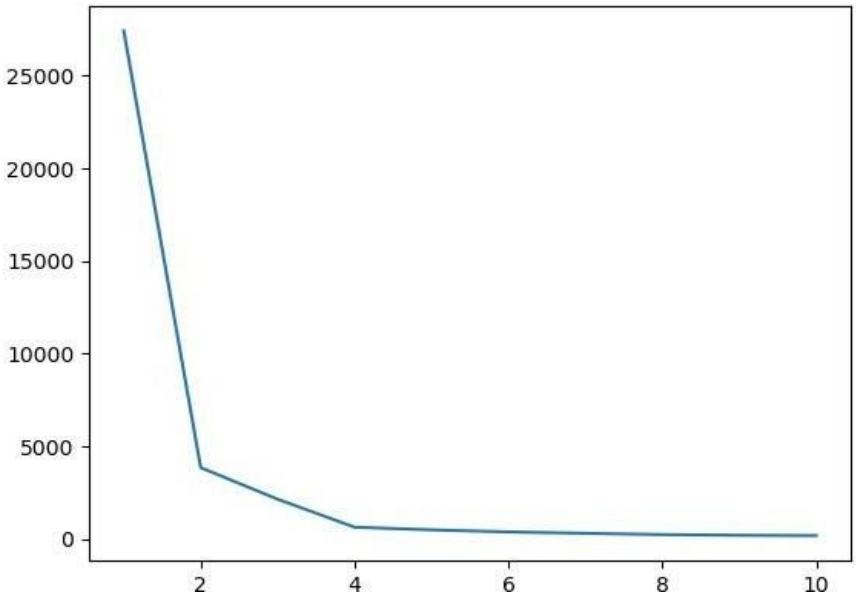

Fig 1: Elbow Curve

As the number of clusters increases, the value of 'wcss' starts to decrease. The 'wcss' value is the largest when K = 2.5 Analyzing the Graph

We see that the graph changes rapidly at some point, thus forming an elbow shape. From this point the graph moves almost parallel to the X-axis. The value of K corresponding to this point is the optimal value of K, that is, the optimal number of clusters. From the above graph it is seen that optimal value of k =4.

Now we will create one numpy array to store the cluster wise data points of students' cgpa and iq. Python code snippet is-
X=df.iloc[:].values km=KMeans(n_clusters=4)
y_means=km.fit_predict(X)
When the value of y_means is printed,it shows the output below. It is nothing but the data points in a cluster y_means. Output-
array ([2, 3, 0, 0, 3, 3, 0, 1, 3, 0, 2, 3, 0, 2, 3, 0, 3,
0, 3, 3, 0, 2,
0, 2, 2, 0, 2, 1, 0, 3, 1, 3, 1, 3, 0, 0, 1, 3, 2, 3, 2, 0,
0, 2,
1, 1, 0, 3, 1, 3, 2, 2, 1, 0, 1, 3, 3, 1, 3, 1, 3, 0, 0, 1,
2, 1,
0, 2, 3, 0, 3, 1, 0, 2, 3, 1, 3, 1, 2, 0, 0, 1, 3, 2, 1, 2,
1, 3,
1, 3, 1, 1, 0, 2, 0, 0, 1, 0, 2, 1, 3, 2, 2, 1, 2, 2, 0, 2,
1, 1,
0, 1, 3, 3, 0, 1, 0, 3, 1, 2, 2, 3, 0, 1, 0, 2, 0, 3, 2, 0,
0, 3,
2, 2, 3, 1, 3, 2, 0, 0, 0, 2, 3, 2, 2, 1, 2, 1, 3, 2, 1, 2,
1, 1,
2, 0, 3, 1, 3, 0, 2, 1, 3, 0, 1, 2, 3, 2, 2, 1, 1, 3, 1, 2,
2, 0,
1, 3, 2, 1])

Cluster no 1 data points are printed with the code snippet:
X[y_means==1,1]
Output
array ([115., 119., 117., 118., 118., 116., 116., 119.,
116., 115., 115.,
117., 118., 113., 116., 118., 117., 121., 116., 117.,
117., 117.,
114., 118., 118., 119., 118., 118., 117., 118., 117.,
119., 118.,
118., 117., 117., 117., 116., 118., 119., 117., 119.,
120., 117., 115., 115.])

Now we will take all four cluster's data points with below code snippet:
Python code: plt.title("Cumulative Grade Point Average VS Intelligence Quotient")
plt.xlabel("Intelligence Quotient")
plt.ylabel("Cumulative Grade Point Average")
plt.scatter(X[y_means==0,0],X[y_means==0,1],color="blue")
plt.scatter(X[y_means==1,0],X[y_means==1,1],color="green")
plt.scatter(X[y_means==2,0],X[y_means==2,1],color="red")
plt.scatter(X[y_means==3,0],X[y_means==3,1],color="orange") Output.

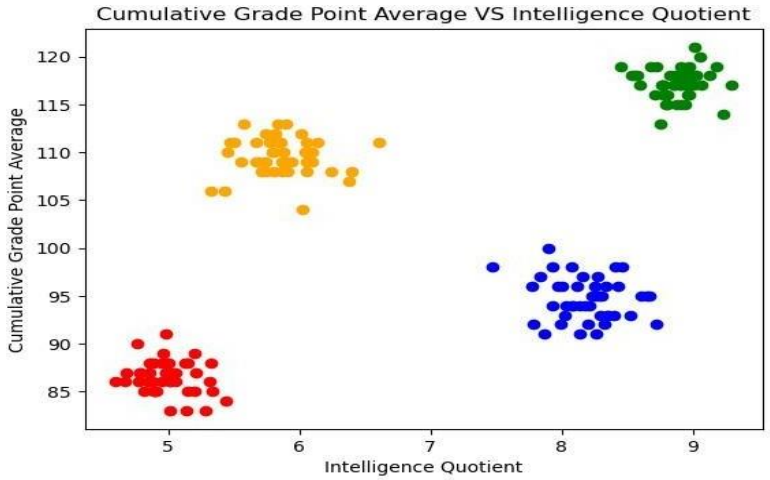

Fig clusters

2.6 Plot Predictions

Green colored cluster 0, students' data indicates that the students are intelligent and their CGPA is also good. So, placement officer need not make many efforts to place them. Much guidance is not required for such students. Such students get placed with the least effort.

If we focus on the blue colored cluster 1, it indicates that these students are hard workers. These students are not very intelligent, but they have more CGPA due to their sincere nature. Such students also do not require much guidance while making their placement.

Orange colored cluster 2 includes the students whose IQ is more than enough but their CGPA is less. Such students are not sincere. They just chill out. They are in a relaxed mood. So, placement officer has to motivate them before their examination. They are not sincere. Some assessment type of efforts must be get done from such students. They may be asked to attempt extra assignments, tutorials, presentations, group discussions, so that their interest in studies can be increased.

Red colored cluster 3 includes students' data whose 'cgpa' and 'iq' both are less. So, placement officer has to make many efforts on these students to increase CGPA also to increase their course related knowledge. Extra efforts like couching classes, notes, practice tests can be conducted for such students to bring them into the mainstream of placement.

3. Conclusions

In this study, we used the k-means clustering technique to analyze student data and forecast the placements of students whose academic performance is bad, average, or good, excellent. If teachers take adequate measures to enhance students' academic performance from bad or average to good, then placements can be increased. This simple analysis study demonstrates that a suitable machine learning technique on student data may be efficiently employed for hidden knowledge./ Information, which can be used for decision making by an educational institution's management and academic department. We hope that the information obtained because of the machine learning and data clustering techniques will be useful.

REFERENCES

1. A Seetharam Nagesh, Ch V S Satyamurty, Application of Clustering algorithm for Analysis of Student Academic Performance, International Journal for Computer and Engineering, Vol 6 Issue 1, E ISSN:2347-2693.
2. Dr Kajal Rai, Students Placement Prediction Using Machine Learning Algorithms, South Asia Journal of Multidisciplinary Studies SAJMS, Vol 8, Issue 5.
3. Er. Arpit Gupta, Er.Ankit Gupta 2Er. Amit Mishra Faculty of Engineering and Technology (IT department), Mahatma Gandhi Chitrakoot Gramodaya Vishwavidyalaya, *International Journal of Advance Technology & Engineering Research (IJATER) ISSN NO: 2250-3536 Vol. 1, Issue 1, November 2011.*
4. Tashfin Ansari, Dr. Almas Siddiqui, Awasthi G. K, Clustering Analysis using an Unsupervised Machine Learning Method, *International Journal of Scientific Research in Computer Science Engineering and Information Technology, DOI:10.32628/CSEIT12173174.*
5. Camila Maione a, Donald R. Nelson b, Rommel Melgaço, Barbosa, Research on social data by means of cluster analysis. *Applied Computing and Informatics Volume 15, Issue 2, July 2019, Pages 153-162.*

6. Oyelade, O. J, Obagbuwa, I. C, Oladipupo, O. O, Application of k-Means Clustering algorithm for prediction of Students' Academic Performance, *International Journal of Computer Science, and Information Security (IJCSIS), Vol. 7, Issue No. 1, 2010.*

7. Sudhir Singh, Nasib Singh Gill, Analysis and Study Of K-Means Clustering Algorithm, *International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, Vol. 2Issue 7, July – 2013.*