iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** VI **Month of publication:** June 2021

DOI: https://doi.org/10.22214/ijraset.2021.35140

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VI Jun 2021- Available at www.ijraset.com

Data Mining Approach for Customer Segmentation

Anshumala Jaiswal[1], Deepika Achla[2], Urmila Rana[3]
[1, 2, 3]*B.Tech, Computer Science and Engineering (8th Semester), Government Engineering College, Raipur*

Abstract: ***In Marketing world, rapidly increasing competition makes it difficult to sustain in this field, marketers have to take decisions that satisfy their customers. Growth of an organization is highly depended on right decisions by the organization. For that, they have to collect deep knowledge about their customer's needs. Substantial amount of data of customers is collected daily. To manage such a huge data is not a piece of cake. An idea is to segment customers in different groups and go through each group and find the potential group among pool of customers. If it is done manually, it will require lot of human efforts and also consume lot of time. For reducing the human efforts, machine learning plays an important role. One can find various patterns which is used to analyze customers database using machine learning algorithms. Using clustering technique, customers can be segmented on the basis of some similarities. One of the best procedures for clustering technique is by using K-means algorithm.***

The k-means clustering algorithm is one of the widely used data clustering methods where the datasets having "n" data points are partitioned into "k" groups or cluster [1].in this paper. K is number of clusters or groups or segments and elbow method is used for determining value of K.

Keywords: Customer Segmentation, clustering, K-Means Algorithm, Elbow method

I. INTRODUCTION

Competition among businesses is increased so rapidly in recent years. To stand in a good position in competitive marketing field, making right choices matters so much, because if it is able to satisfy customer's needs then it will help to increase revenue of the company, which helps hugely in the company's growth. While taking decision, requirements of potential customers have to be kept in mind. To find potential customers among pool of customers is very difficult. Everyday huge amount of data is stored and from these data, extraction of meaningful information is very crucial. To find potential customers among pool of customers is very difficult. Everyday huge amount of data is stored and from these data, extraction of meaningful information is very crucial. Data mining techniques are used for this. Data mining is a method of extracting meaningful data from raw datasets. This process involves analyzing data patterns using some methods and converting extracted data into format which is easily readable by human. Clustering technique is widely used in data mining which is used to divide data points into clusters such that data points are related to one another in some way within a cluster and different from data points of other clusters. segmentation of customer is a process of dividing customers into groups which is also called as segments on the basis of some similar characteristics in many ways, may it be behavioral or may it be geographical characteristics. customer segmentation is very beneficial for finding profitable customers among different segments. It is helpful for prize optimization by understanding financial status of the customers. It helps to enhance competitiveness in market by increasing customer retention, which increases revenue generation. It also supports business decisions. Aim of this paper is to divide customers into groups using data mining approach, which includes K-means algorithm. K -means algorithm divides customer into K segments and for determining value of K elbow method is used. In Demographic segmentation customers can be divided into groups on the basis of their age, gender, income etc. In psychographic, division of customer is on thew basis of characteristics like interests and priorities of customers.

II. LITERATURE REVIEW

A. Customer Segmentation

It is process of splitting the customers into different groups, each group have customers which share some similar characteristics. The purpose of segmentation is to customize the products, services, and marketing messages for each segment [2]. Segmentation is divided into four types i.e., Demographic, psychographic, Behavioral and geographic. In behavioral segmentation customers are divide by observing there spending habits, purchasing habits etc. In geographic segmentation, division is on the basis of geographic location of customers, it can be their zip code, city or country. Customer segmentation works as decision support for an organization, it increases retention of customer in an organization, revenue of organization, competitiveness. It is beneficial for company's growth.

B. Clustering

Clustering is process of forming clusters of data points, which share some relationship with one another within a cluster and different from other cluster's data points. Clustering is the unsupervised classification of patterns (observations, data items, or feature vectors) into groups (clusters) [3]. Clustering is a technique used in data mining. There are different algorithms available for clustering, k-means algorithm is one of them.

C. K-Means Algorithm

It is iterative process of finding K best centroids for K clusters and forming clusters by using distance formula. Initially value of k is defined but the optimum value of k can be found further through elbow method.

Let us assume there is a dataset D, which contains m data points. and there are K clusters i.e., N1, N2 up to NK and their respective centroids as C1, C2, up to CK. K-means algorithm splits m data points into K clusters. Let m data points are x1, x2, up to xm. The decision whether a data point xi belongs to a cluster or not is based on the distance factor. If data point xi is closest to centroid Cj then it belongs to cluster j where value of i and j is from 1 to K.

Algorithm: INPUT: Dataset D containing m data point, k: No of clusters to be formed. OUTPUT: K cluster set

Method:

choose k random centroids for k clusters from m data points

Repeat for remaining data points until centroid is not changing.

Compute distance between daata points and all centroids.

Allocate each data point to a cluster , if it is closest to centroid of that cluster.

for all clusters, recalculate centroid by taking average of data points(including centroid) belongs to that cluster.

III. METHODOLOGY

In this article, a mall customer dataset is used . It contains 200 rows and 5 columns . The attribute present in the dataset are CustomerId, Gender , Age , Annual income (K$), Spending Score (1-100) in that order.

A. Exploring and Preprocessing the Data

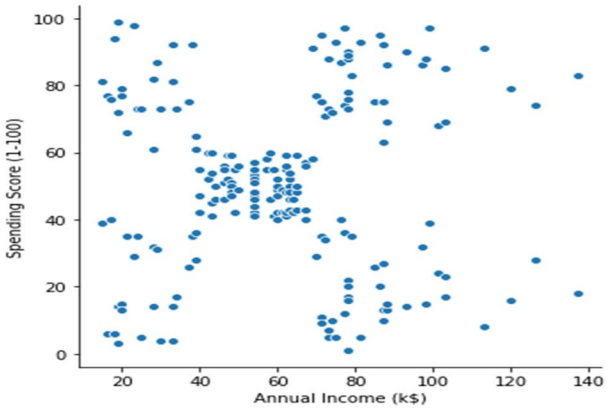
```
df =pd.read_csv("C:\\\\Users\\\\LENOVO\\\\Desktop\\Mall_Customers (1).csv")
```

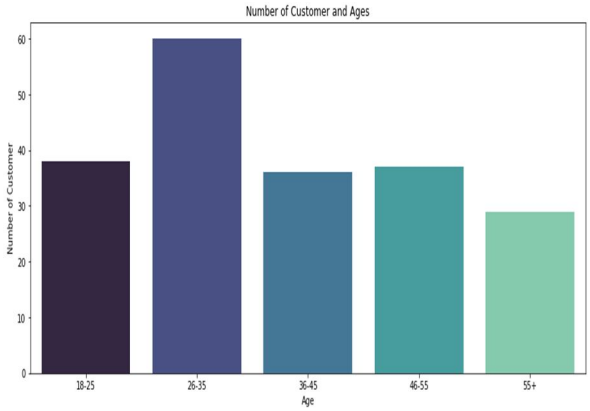
```
df.head()
```

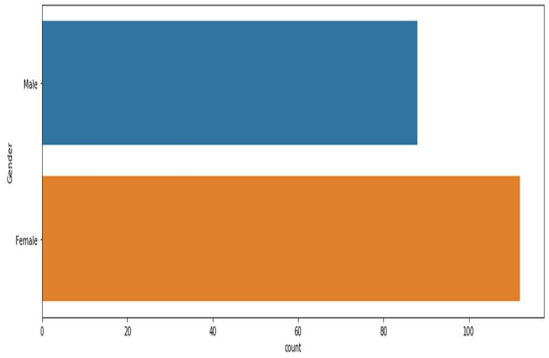
```
df.describe()
```

B. Visualizing the Dataset

```
sns.relplot(x="Annual Income (k$)", y="Spending Score (1-100)", data=df)
```



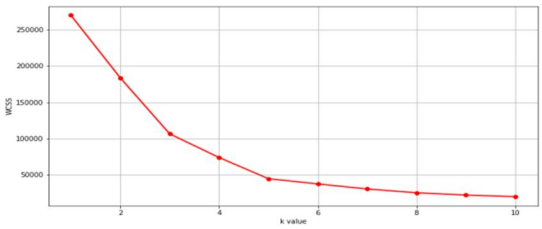




C. Elbow Method

Elbow method is used to find most suitable value for k. As the number of cluster increases, the sum of squared distance between centroid and data points within a cluster also increases. In elbow method, graph between sum of squared distance(SSE) and number of cluster is formed and the value of k is chosen at a point where there is a shape like elbow. after that point in a graph if value of k increases , sum of squared distance tends to zero.

```
X2=df.loc[:, ["Annual Income (k$)","Spending Score (1-100)"]].values

from sklearn.cluster import KMeans

wcss = []

for k in range(1,11):

    kmeans = KMeans(n_clusters=k, init="k-means++")

    kmeans.fit(X2)

    wcss.append(kmeans.inertia_)

plt.figure(figsize=(12,6))

plt.grid()

plt.plot(range(1,11),wcss, linewidth=2, color="red", marker ="8")

plt.xlabel("k value")

plt.ylabel("WCSS")

plt.show()
```



D. Clustering

```
kmeans = KMeans(n_clusters=5)

label = kmeans.fit_predict(X2)

print(label)



plt.scatter(X2[:,0],                    X2[:,1],
c=kmeans.labels_, cmap='rainbow')

plt.scatter(kmeans.cluster_centers_[:,0]
,kmeans.cluster_centers_[:,1],  color='black')

plt.title('cluster of customers')

plt.xlabel('Annual Income (k$)')

plt.ylabel('Spending Score (1-100)')

plt.show()
```

IV. CONCLUSION

A. As it is clear that, optimum value of k is 5, which is obtained by using elbow method. There are 5 clusters formed.

B. Most of the customer have Annual income between 40-60 in the unit of k$, and there spending score lies between 40-60 out of 100. Which is represented by green color in the graph.

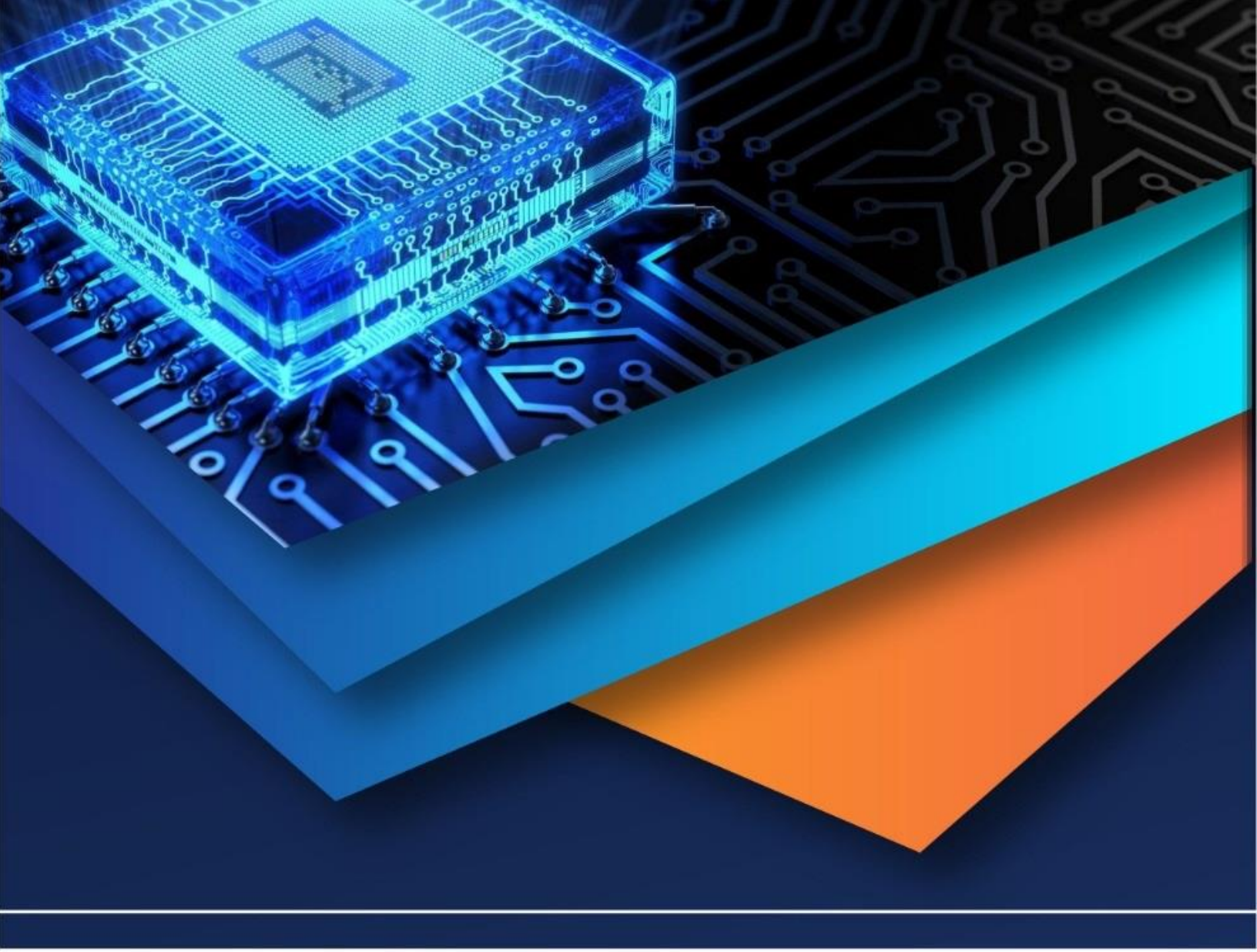
C. They are potential customers so in mall, some new scheme has to be used to attract them.

D. Count of female customers are more as compared to male customers as shown in counter plot. New products will have to be provided related to female needs, it helps to increase revenue.

E. There are a greater number of customers between age group of 25 to 35 as shown in bar graph so decision have to be taken which helps to attract their attention to buy items.

REFERENCES

[1] [Haraty RA, Dimishkieh M, Masud M. An Enhanced k-Means Clustering Algorithm for Pattern Discovery in Healthcare Data. International Journal of Distributed Sensor Networks. June 2015. doi:10.1155/2015/615740
[2] M A Syakur et al 2018 IOP Conf. Ser.: Mater. Sci. Eng. 336 012017
[3] JAIN, A. K., MURTY, M. N., & FLYNN, P(1999). Data Clustering: A Review. *ACM Computing Surveys*, *31*(3), 265–323.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY