iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH
IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 Issue: I Month of publication: January 2023

DOI: https://doi.org/10.22214/ijraset.2023.48368

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue I Jan 2023- Available at www.ijraset.com

Mall Customer Segmentation Using Clustering Algorithm

Mr. M. Sathyanarayana[1], S. Dhanish[2], P. Shiva Kumar[3], A. NiranjanReddy[4]
[1]Assistant Professor, Dept. of Computer Science and Engineering, SNIST, Hyderabad, 501301, India
[2, 3, 4]B.Tech Scholar, Dept. of Computer Science and Engineering, SNIST, Hyderabad, 501301, India

Abstract: We live in a world where large and vast amount of data is collected daily. Analysing such data is an important need. In the modern era of innovation, where there is a large competition to be better then everyone, the business strategy needs to be according to the modern conditions. The business done today runs on the basis of innovative ideas as there are large number of potential customers who are confounded to what to buy and what not to buy. The companies doing the business are also not able to diagnose the target potential customers. This is where the machine learning comes into picture, the various algorithms are applied to identify the hidden patterns in the data for better decision making. The concept of which customer segment to target is done using the customer segmentation process using the clustering technique. In this paper, the clustering algorithm used is K-means algorithm which is the partitioning algorithm, to segment the customers according to the similar characteristics. To determine the optimal clusters, elbow methodis used.

I. INTRODUCTION

A. Introduction

Over the years, the competition amongst businesses is increased and the large historical data that is available has resulted in the widespread use of data mining techniques in extracting the meaningful and strategic information from the database of the organisation. Data mining is the process where methods are applied to extract data patterns in order to present it in the human readable format which can be used for the purpose of decision support. According to,[4] Clustering techniques consider data tuples as objects. They partition the data objects into groups or clusters, 2 so that objects within a cluster are similar to one another and dissimilar to objects in other clusters. Customer Segmentation is the process of division of customer base into several groups called as customer segments such that each customer segment consists of customers who have similar characteristics. The segmentation is based on the similarity in different ways that are relevant to marketing such as gender, age, interests, and miscellaneous spending habits. The customer segmentation has the importance as it includes, the ability to modify the programs of market so that it is suitable to each of the customer segment, support in business decision; identification of products associated with each customer segment and to mange the demand and supply of that product; identifying and targeting the potential customer base, and predicting customer defection, providing directions in finding the solutions. The thrust of this paper is to identify customer segments using the data mining approach, using the partitioning algorithm called as K-means clustering algorithm. The elbow method determines the optimal clusters.

B. Problem Statement

Customer Segmentation is the best application of unsupervised learning. Using clustering, identify segments of customers in the dataset to target the potential user base. They divide customers into various groups according to common characteristics like gender, age, interest, and spending habits so they can market to each group effectively. Use K-Means Clustering and also visualize the gender and age distributions. Then analyze their annual income and spending scores. As it describes about how we can divide the customers based on their similar characteristics according to their needs by using k-means clustering which is a classification of unsupervised machine learning.

II. EXISTING SYSTEM

The existing method is storing customer data through paperwork and computer software (digital data) is increasing day by day. At end of the day they will analyse their data as how many things are sold or actual customer count etc. By analysing the collected data they got to know who is beneficial to their business and increase their sales. It requires more time and more paperwork. Also, it is not much effective solution to find the desired customers data.

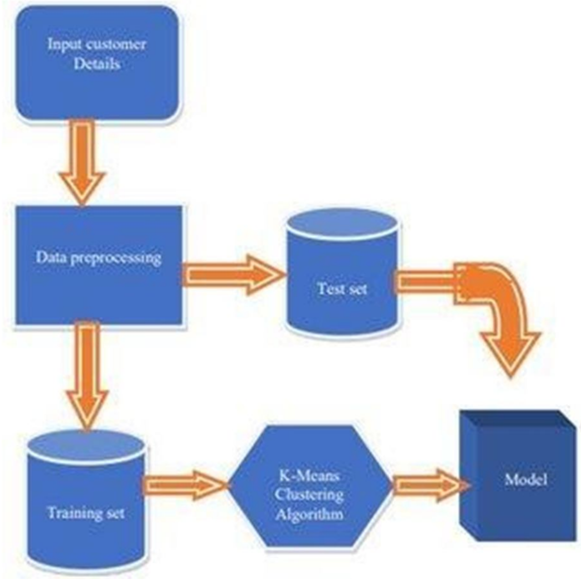
III. PROPOSED SYSTEM

A. *Proposed Method*

To overcome the traditional method i.e paper work and computerized digital data this new method will play vital role. As we collect a vast data day by day which requires more paperwork and time to do. As new technologies were emerging in today's world. Machine Learning which is powerful innovation which is used to predict the final outcome which has many algorithms. So for our problem statement we will use K-Means Clustering which groups the data into different clusters based on their similar characteristics. And then we will visualize the data.

B. *System Architecture*

Initially we will see the dataset and then we will perform exploratory data analysis which deals with the missing data, duplicates values and null values. And then we will deploy our algorithm k-means clustering which is unsupervised learning in machine learning.



As in order to find the no of clusters we use elbow method where distance will be calculate through randomly chosen centres and repeat it until there is no change in cluster centres. Thereafter we will analyse the data through data visualization. Finally we will get the outcome.

C. *Algorithm*

1) *K-Means Clustering*

a) KMeans algorithm in an iterative algorithm that tries to partition the dataset into K predefined distinct non overlapping sub groups which are called as cluster.
b) Here K is the total no ofclusters.
c) Every point belongs to only onecluster.
d) Clusters cannot overlap.

2) *Steps of Algorithm*

a) Arbitrarily choose k objects from D as the initial cluster centers.
b) Repeat.
c) Assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster.
d) Update the cluster means, i.e. calculate the mean value of the objects for each cluster.
e) Until no change.

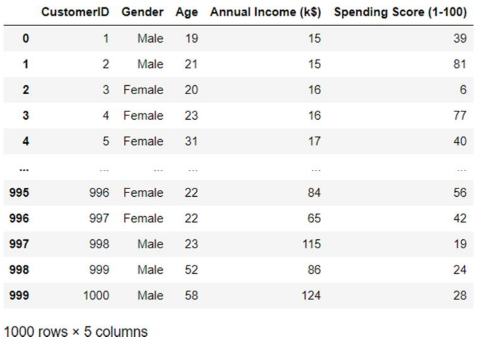
IV. METHODOLOGY

1) First of all we will import all the necessary libraries or modules(pandas, numpy, seaborn).
2) Then we will read dataset and anyalse whether it contains any null values, missing values and duplicate values. So we will fix them by dropping or fixing the value with their means, medians etc which is technically named as Data Preprocessing.
3) We will deploy our model algorithm K-Means Clustering, which divides the data into group of clusters based on similar characteristics. To find no.of clusters we will use elbow method.
4) Finally, we will visualize our data using matplot, which concludes the customers divided into groups who are similar to each other on their group.

V. IMPLEMENTATION AND ANALYSIS

A. Overview Of A Dataset

This is a mall customer segmentation data which contains 5 columns and 1500 rows.



	CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)
0	1	Male	19	15	39
1	2	Male	21	15	81
2	3	Female	20	16	6
3	4	Female	23	16	77
4	5	Female	31	17	40
...	...	...	...	...	...
995	996	Female	22	84	56
996	997	Female	22	65	42
997	998	Male	23	115	19
998	999	Male	52	86	24
999	1000	Male	58	124	28

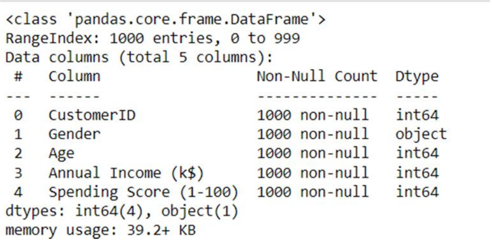
1000 rows × 5 columns

B. Exploratory Data Analysis

It deals with the data preprocessing, whether it contains any missing values or null values. There after we will see the information and description of the dataset.

1) *Information of the dataset*

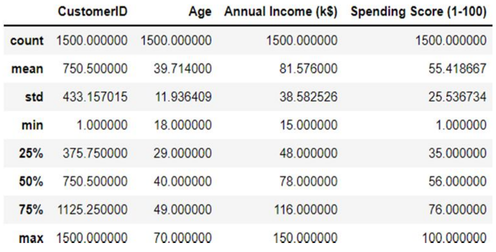
#df.info()



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 5 columns):
 #   Column                  Non-Null Count  Dtype 
---  ------                  --------------  ----- 
 0   CustomerID              1000 non-null   int64 
 1   Gender                  1000 non-null   object
 2   Age                     1000 non-null   int64 
 3   Annual Income (k$)      1000 non-null   int64 
 4   Spending Score (1-100)  1000 non-null   int64 
dtypes: int64(4), object(1)
memory usage: 39.2+ KB
```

2) *Description of the Data*

#df.describe()

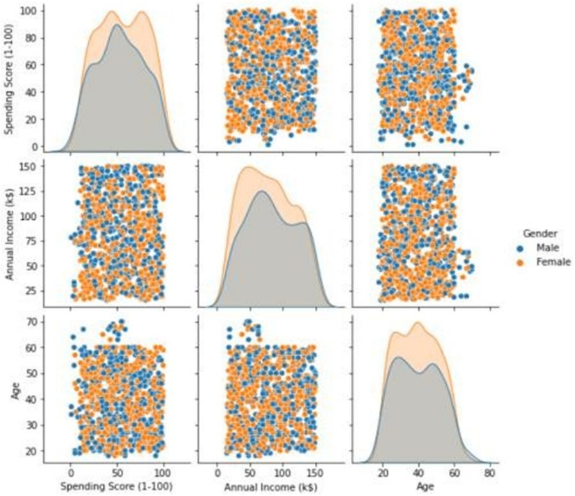


	CustomerID	Age	Annual Income (k$)	Spending Score (1-100)
count	1500.000000	1500.000000	1500.000000	1500.000000
mean	750.500000	39.714000	81.576000	55.418667
std	433.157015	11.936409	38.582526	25.536734
min	1.000000	18.000000	15.000000	1.000000
25%	375.750000	29.000000	48.000000	35.000000
50%	750.500000	40.000000	78.000000	56.000000
75%	1125.250000	49.000000	116.000000	76.000000
max	1500.000000	70.000000	150.000000	100.000000

It describes about the count which counts the no of rows in it, mean of the columns, standard deviations, maximum and minimum and percentiles etc.

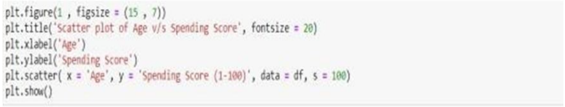
C. Seaborn Pairplot

sns.pairplot(df, vars = ['Spending Score (1-100)', 'Annual Income (k$)', 'Age'], hue = "Gender")

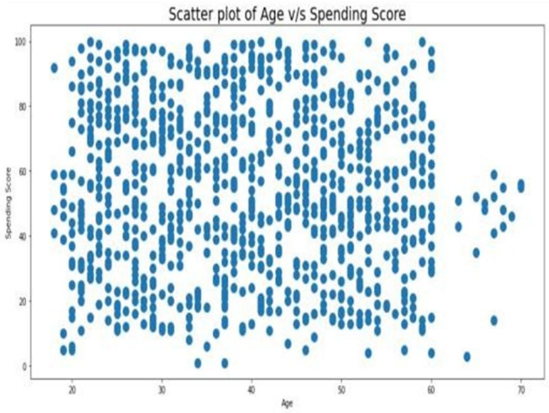


D. Age vs Spending Score

As we will use scatterplot and labelled x-axis as Age and y-axis as Spending Score(1-100)



```
plt.figure(1 , figsize = (15 , 7))
plt.title('Scatter plot of Age v/s Spending Score', fontsize = 20)
plt.xlabel('Age')
plt.ylabel('Spending Score')
plt.scatter( x = 'Age', y = 'Spending Score (1-100)', data = df, s = 100)
plt.show()
```



From the plot we observed that it varies from low annual income with low expenditure or spending money to high annual income with high expenditure.

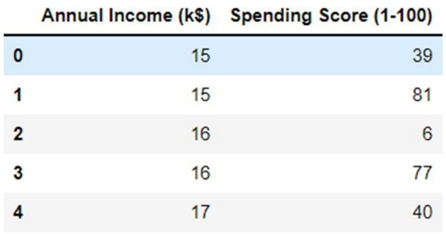
E. Elbow Method

The elbow method is based on the observation that increasing the number of clusters can help to reduce the sum of within-cluster variance of each cluster. This is because having more clusters allows one to capture finer groups of data objects that are more similar to each other. To define the optimal clusters, Firstly, we use the clustering algorithm for various values of k. This is done by ranging k from 1 to 10 clusters. Then we calculate the total intra-cluster sum of square. Then, we proceed to plot intra-cluster sum of square based on the number ofclusters.

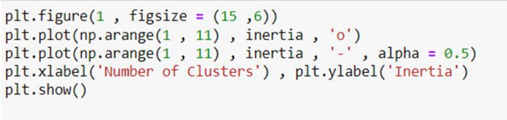
The plot denotes the approximate number of clusters required in our model. The optimum clusters can be found from the graph where there is a bend in the graph.

First we will consider the data Xwhich as only two columns they are annual income and spending score.

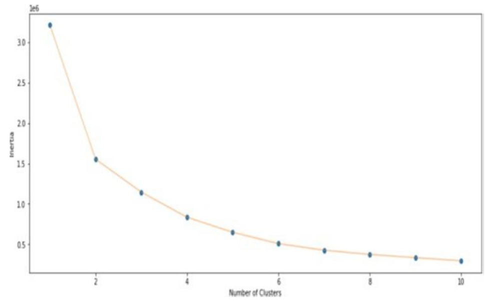
X=df[['Annual Income (k$)','Spending Score (1-100)']]
X.head()



	Annual Income (k$)	Spending Score (1-100)
0	15	39
1	15	81
2	16	6
3	16	77
4	17	40

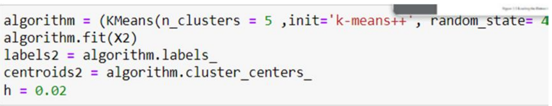


```
plt.figure(1 , figsize = (15 ,6))
plt.plot(np.arange(1 , 11) , inertia , 'o')
plt.plot(np.arange(1 , 11) , inertia , '-' , alpha = 0.5)
plt.xlabel('Number of Clusters') , plt.ylabel('Inertia')
plt.show()
```

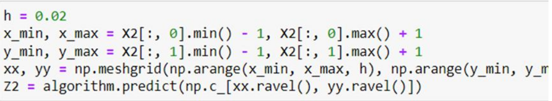


So from the graph we observed thatthe at 5 there is bend and it can be considered as k which is no of clusters. Therefore, k=5 i.e no of clusters are equal to 5.

F. Fitting the Algorithm

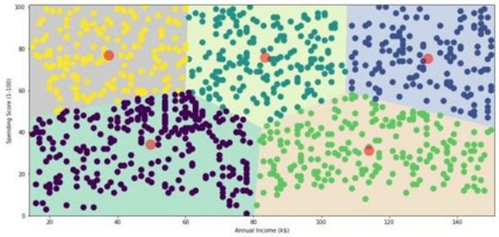


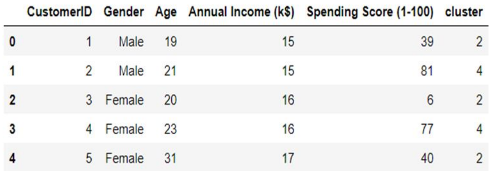
```
algorithm = (KMeans(n_clusters = 5 ,init='k-means++', random_state= 4
algorithm.fit(X2)
labels2 = algorithm.labels_
centroids2 = algorithm.cluster_centers_
h = 0.02
```



```
h = 0.02
x_min, x_max = X2[:, 0].min() - 1, X2[:, 0].max() + 1
y_min, y_max = X2[:, 1].min() - 1, X2[:, 1].max() + 1
xx, yy = np.meshgrid(np.arange(x_min, x_max, h), np.arange(y_min, y_m
Z2 = algorithm.predict(np.c_[xx.ravel(), yy.ravel()])
```

As here we initialized the kmeans as km with 5 clusters and we will fit it. There after we will predict the data and store it in y. And then we will add new column named as Cluster and data as y.





	CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)	cluster
0	1	Male	19	15	39	2
1	2	Male	21	15	81	4
2	3	Female	20	16	6	2
3	4	Female	23	16	77	4
4	5	Female	31	17	40	2

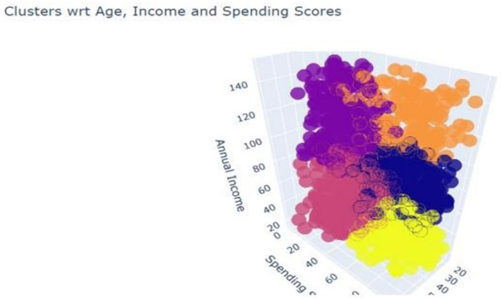
```
trace1 = go.Scatter3d( x= df['Age'],
y= df['Spending Score (1-100)'], z= df['Annual Income (k$)'], mode='markers',
marker=dict(
color = df['cluster'], size= 10,
line=dict(
color= df['cluster'], width= 12
),
opacity=0.8
)
)
```

So from the figure we observed that each customer is labelled with cluster which is based on their characteristics.

G. Visualization the Clusters

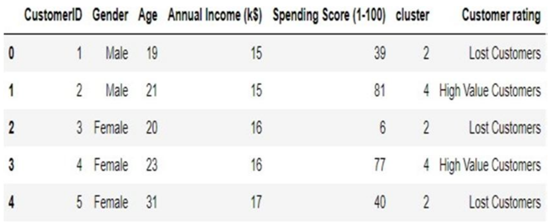
Visualizing the clusters based on Annual Income and Spending Score of the customers. As here we plot a graph named as Clusters of Customers to visualize the data in terms of groups or cluster.

```
        import plotly as py
        import plotly.graph_objs as go
      data = [trace1] layout = go.Layout(
         title= 'Clusters wrt Age, Income and Spending Scores',
         scene = dict(
             xaxis = dict(title = 'Age'),
              yaxis  =  dict(title        = 'Spending Score'),
              zaxis  =  dict(title          = 'Annual Income')
           )
        )
        fig           =           go.Figure(data=data, layout=layout)
        py.offline.iplot(fig)
```



H. Naming the Clusters

```
df['Customer rating']=np.where(df['cluster']==2, "Lost            Customers"        ,(np.where( df['cluster'] ==3, "top value
Customer", (np.where( df['cluster'] == 0, "Medium Value Customer",np.where(df['cluster']== 1,'Low Value Customers', 'High
Value Customers'))))))
        df.head(5)
```



	CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)	cluster	Customer rating
0	1	Male	19	15	39	2	Lost Customers
1	2	Male	21	15	81	4	High Value Customers
2	3	Female	20	16	6	2	Lost Customers
3	4	Female	23	16	77	4	High Value Customers
4	5	Female	31	17	40	2	Lost Customers

So from the above one we observed that the there are 5 clusters which are named as 0, 1, 2, 3, 4.

I. Clustering Segmentation Model

- *Cluster 0: Medium value customers*

This group is having the high balance and medium purchase frequency.

- *Cluster 1: Low value customers*

This group is having the high balance and low purchase frequency.

- *Cluster 2: Lost customers*

This group is having the medium and low balance and low purchase frequency.

- *Cluster 3: Top value customers*

This group is having the highest balance and high purchase frequency.

- *Cluster 4: High value customers*

This group is having the low balance and high purchase frequency

VIII. CONCLUSION

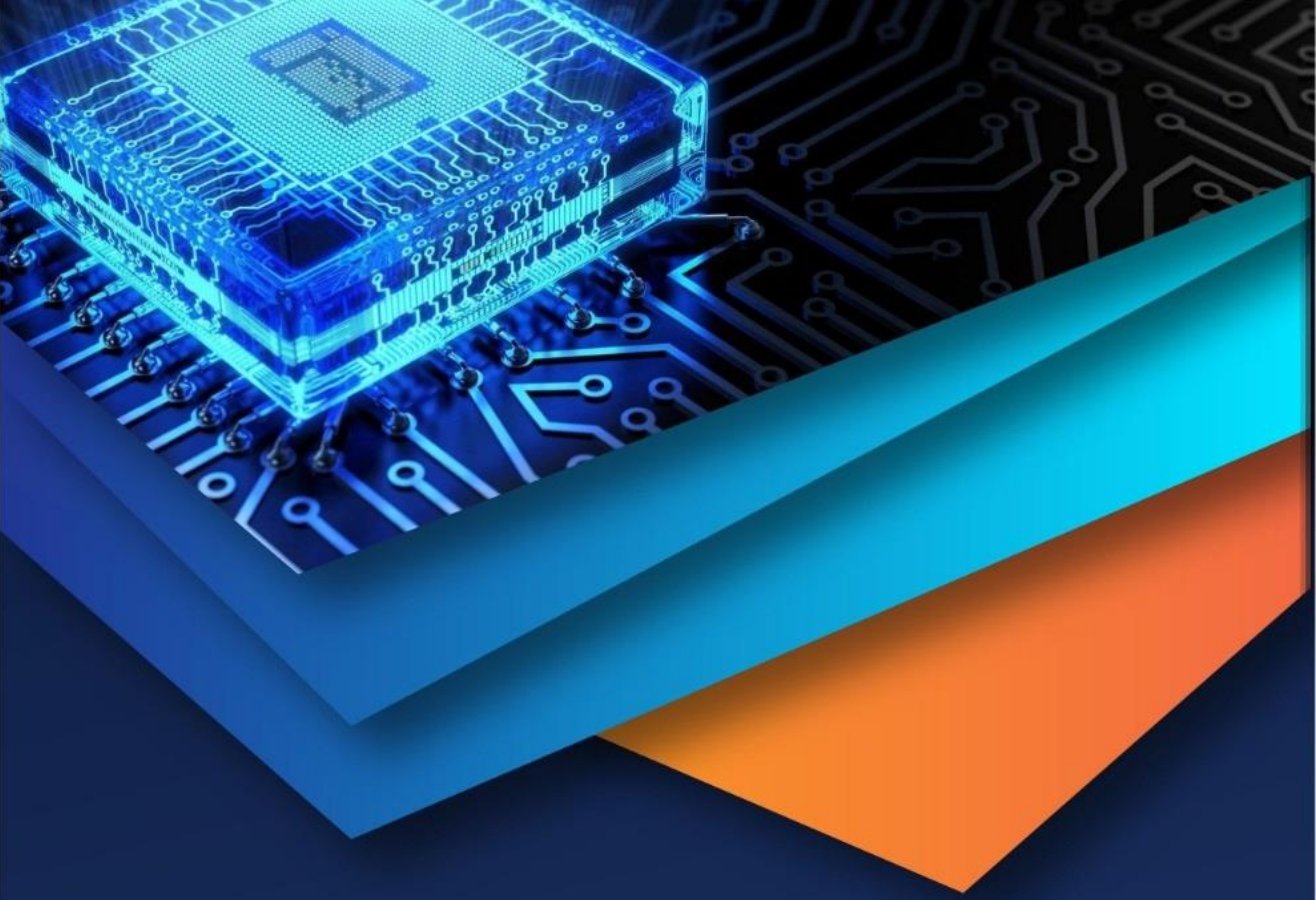
So we concluded that the ,

1) The Highest income , high spending can be target these type of customers as they earn more money and spend as much as they want.
2) Highest income, low spending can be target these type of customers by asking feedback and advertising the product in a better way.
3) Average income, Average spending may or may not be beneficial to the mall owners of this type of customers.
4) Low income, High spending can be target these type of customers by providing them with low-cost EMI's etc.
5) Low income, Low spending don't target these type of customers because they earn a bit and spend some amount of money.

So high income, high spending are the most beneficial ones to the mall owners which increases the owner's business. Using market segmentation, companies are able to identify their target audiences and personalize marketing campaigns more effectively. This is why market segmentation is key to staying competitive. It allows you to understand your customers, anticipate their needs, and seize growth opportunities.

REFERENCES

[1] Cooil, B., Aksoy, L. & Keiningham, T. L. (2008), 'Approaches to customer segmentation', Journal of Relationship Marketing 6(3-4), 9–39.
[2] D. Aloise, A. Deshpande, P. Hansen, and P. Popat, "The Basis Of Market Segmentation" Euclidean sum-of-squares clustering," Machine Learning, vol. 75, pp. 245- 249, 2009.
[3] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R.Silverman, and A. Y. Wu, "An efficient K-means clustering algorithm," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 24, pp. 881-892, 2002.
[4] Bhatnagar, Amit,Ghose, S. (2004), 'A latent class segmentation analysis of e-shoppers', Journal of Business Research 57, 758–767.
[5] Marcus, C. (1998), 'A practical yet meaningful approach to customer segmentation approach to customer segmentation', Journal of Consumer Marketing15, 494–504.



doi crossref

10.22214/IJRASET

INDEX COPERNICUS

45.98

ISRA JIF

IMPACT FACTOR: 7.129

TOGETHER WE REACH THE GOAL

IMPACT FACTOR: 7.429

ISI

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY