Journal of Positive School Psychology http://journalppw.com
2022, Vol. 6, No. 11, 461-467

Analysis Of Customers' Shopping Behaviour By Segmenting Customers Using K-Means Clustering Approach

Dr. Deepti Sharma[1*], Dr. Deepshikha Aggarwal[2], Dr. Archana B.Saxena[3]

[1*,2,3] *Department of Information Technology Jagan Institute of Management Studies, Sector-5,Rohini, Delhi, INDIA*

***Corresponding Author: -** Dr. Deepti Sharma

Abstract

In all trading areas today, e-commerce has become popular and most implementable method. E-commerce means promoting and marketing products to customers' electronically. Segmenting customers means dividing them into groups which share same features. The main reason of segmenting the customers is to analyse their shopping behaviour and suggesting ways to increase the profit to the business. Customers' segmentation is done to assist business to find out the customers who are profitable to the business and how to increase the sales to the business. In this study, k-means clustering approach is used to analyse customers' shopping behaviour by segmenting them in clusters. The aim of using clustering method is to find out shopping behaviour pattern within the clusters. In this research, the recommended approach investigated the similar groups and their preferences which will help business to increase profits. It also facilitates to provide different offers to the potential customers in order to achieve long term sales and profits.

Keywords: Customer segmentation, purchasing behaviour, K-Means clustering, clusters

1. Introduction

In covid-19 pandemic, new working opportunities with work from home (WFH) and study from home (STH) has started. Many large businesses also opened e-commerce websites, portals and apps to maintain their profits during this pandemic time [1] [2]. Because of this current situation, people were unable to move outside even for purchasing day to day necessary items. Thus, e-commerce played an important role where people can buy goods and services while sitting at home [3]. Customer segmentation means dividing the customers into the groups that share same features. When all customers will be divided into groups or clusters, it would be easy to analyse their shopping behaviour and thus helpful for organization to increase their sales and profit.

In this digital marketing era, the information and preferences of customers are recorded. Their shopping behaviour is stored and later used for promoting products and providing promotional offers. If customer purchases one thing, he will be shown the similar products that he can buy as per his selection [4]. But sometimes, too much mails or messages are dangerous where customer can be confused on deciding their needs. In such cases, clustering techniques will be helpful to divide the customers on the basis of their annual income, age, spending habits or purchasing specific products or brands. For collecting the data and segmenting the customers, an unsupervised machine learning algorithm is used which is known as k-means clustering [5]. The main aim of this study is to segment the customers as per their preference and similar features and then provide them various facilities and offers to increase profit of business.

2. Related Work

Due to increase in online shopping, various e commerce websites have emerged [6]. There are competitions between traditional and online businesses these days [7]. Customers plays an important role in any business. Without customers, there is no business growth and thus no profit to retain the business. Retaining customers in any business is also a great challenge [8]. Providing them opportunities, offers, customer support system as per their preference is also very important. If the wrong things will be provided to customers in terms of quality, cost, variety etc ; they will never return to that website and will think this website is not as per their taste and thus business will lose the potential customers. Business has to understand that each customer is different and has different preferences and ability to spend [9]. Thus there must be different marketing strategy for different customers [10]. Customer segmentation is done to know about the purchasing behaviour of customers to decide whether a customer will buy a specific product or not [11]. [12] Claims that to sustain business and retain in market, customer segmentation must be

carried out. As per [13] segmentation of customers depends on determining important factors that divides customers into similar groups.

3. Problem Formulation

The purpose of this study is to divide the e-commerce customers into different segments. An unsupervised learning algorithm called K-Means clustering algorithm is used for the purpose customer segmentation. The goals and offerings of this research are:

- To analyse the connection between customer spending behaviour and age of customers.
- To analyse the connection between customer spending behaviour and annual income of customers.
- To analyse the connection between customers' age and annual income.
- To understand the behaviour of customers, dividing them in segments and focusing on high profitable customers' segment.
- The business can determine appropriate product pricing to keep for different types of customers based on their spending behaviour.
- Marketing campaigns can be customized as per customers' demand and choice.
- An optimal distribution strategy can be designed.
- Specific products can be chosen for deployment
- If there are new products to be launched by company, these products can be prioritized as per customer's shopping behaviour.

4. Methods

This section elaborates the database used, objectives, algorithm and proposed methodology for the research work.

4.1 Cluster Algorithm

A cluster is a group of similar things occur together and work together. Clustering can further be used for segmentation of customers to perform analysis on similar type of group together. Literature also reveals [14] that clustering can be used for customer segmentation. K- means clustering is an unsupervised learning algorithm which groups the dataset into different clusters. Here k stands for the number of clusters that need to be created in the process. For example, if the value of k=2, means 2 clusters will be created. If value of k=3, three clusters will be created and so on. K-means clustering algorithm is faster as compared to other algorithms in computation. The rate of misclassification of data is also reduced while using k means algorithm. One of the key uses of k-means clustering is segmentation of customers. The current research work also uses k-means algorithm because of the above reasons.

4.2 Proposed Methodology

The methodology proposed can be broadly divided into 4 steps. They are explained in detail below:

Step 1: Exploratory Data Analysis and Pre-processing of Data

In statistics, an approach of analyzing data sets to summarize their main characteristics is known as Exploratory data analysis (EDA). It is often used to create graphs and perform visualisation of data.
In this research work, EDA helps in recognizing distinctive customers, customers' annual income and their spending habits while shopping, customers' age and their spending habits and their age and annual income. Pre-processing of data is also done in mismatch of data and to check null values.

Step 2: K-means Clustering Algorithm

In k means clustering, "k" is related to the number of clusters or groups created. With K means clustering, user has to define the number of clusters required and algorithm will create it. In k-means clustering, "means" bit refers to the fundamental process by which the data get associated with the clusters. If in k-means model, value of k is equal to 5, then the algorithm will create five random points called "centroids". Thus, specify number of clusters, initialize centroids by shuffling the dataset. Keep changing the centroids, until there is no change to the centroids.

Step 3: Calculating Silhouette Score

For each observation that belongs to the clusters, Silhouette score is calculated. To calculate Silhouette score, mean distance between observation and all other data points of same clusters are calculated. This distance is also called a mean intra-cluster distance.

Step 4: To solve the problem, following steps need to be performed:

1. **Importing Libraries** : There are various libraries which will be needed for performing customer segmentation using k means clustering. Numpy, pandas, matplotlib, seaborn are few of them.
2. **Data Exploration:** It will help in knowing your data, describing it, knowing rows and columns, converting required columns in numeric values and filling the missing values.

3. **Data Visualization:** Generating the plots based on different columns and finding if there is any correlation between columns.
4. **Clustering using K-Means:** Clustering will be used for segmenting the customers. It can be done on different columns say age and spending score, annual income and spending score in this research work.
5. **Selection of Clusters:** Once segmentation is done, N clusters are selected based on inertia. It is calculated where Squared distance between Centroids and data points is less.
6. **Making an optimal number of clusters:** Finally, calculate Silhouette score to find the optimal number of clusters.

The best clusters formed after performing above points will be taken for doing further predictions about the data and ways will be suggested to improve shopping behaviour among customers.

4.3 Database

The dataset used for this research has 5 fields. It consists of 200 rows storing the customers' data. Sample data set is shown in Fig 1 and it's information in fig 2 respectively. Data set describes Gender, age, annual income and customers' spending score. Spending score means average spending amount by customers. Data is pre-processed and all missing fields are filled up. Gender field is converted into numeric value (male=1 and female=0).

	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

Fig 1: Sample data set

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200 entries, 0 to 199
Data columns (total 5 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   CustomerID              200 non-null    int64
 1   Gender                  200 non-null    object
 2   Age                     200 non-null    int64
 3   Annual Income (k$)      200 non-null    int64
 4   Spending Score (1-100)  200 non-null    int64
dtypes: int64(4), object(1)
memory usage: 7.9+ KB
```

Fig 2: Information about data set

5. Data Visualization

5.1 Count Plot of Gender

A bar graph of the count of Gender is shown in Fig 3. It shows that there are 112 males and 88 females in the data set used.

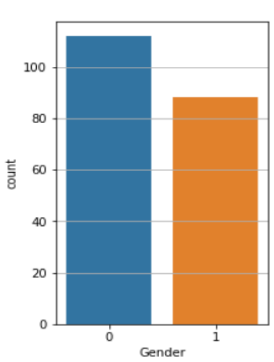

Fig 3: Count plot of Gender

5.2 Correlation Between Columns

Correlation between different columns is shown in Fig 4. It can be observed that there is hardly any correlation between columns.

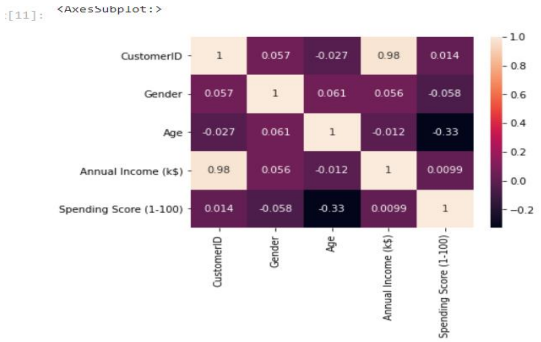

Fig 4: Correlation between Columns

5.3 Plot of Age

Fig 5 below shows the age graph of customers. It can be seen that participants are in the age group of 20-70 and maximum participants are in the age group of 32-38.

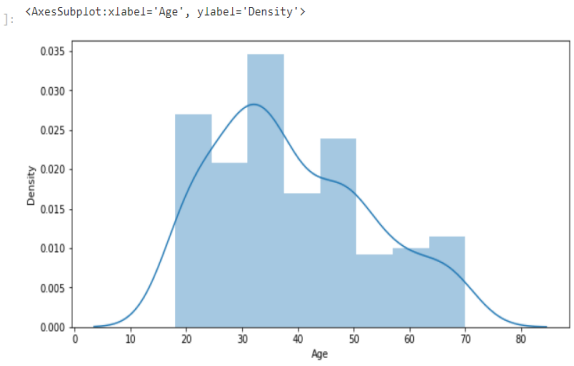

Fig 5: Plot of Age

5.4 Plot of Annual Income

The annual income of customers' participants is 20 to 140K. The maximum customers lie in the range of 50 to 60 and 70 to 80K annual income.

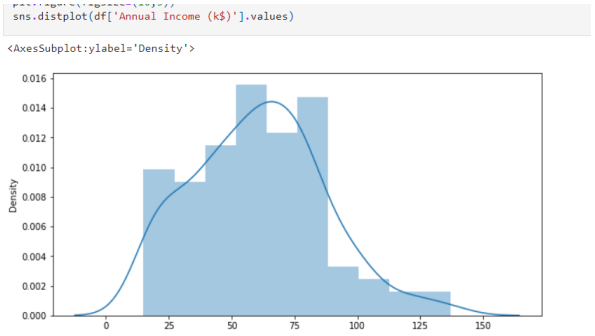

Fig 6: Plot of Annual Income

5.5 Plot of Spending score

The maximum spending score is in between 40 to 60 is being shown in Fig 7 below.

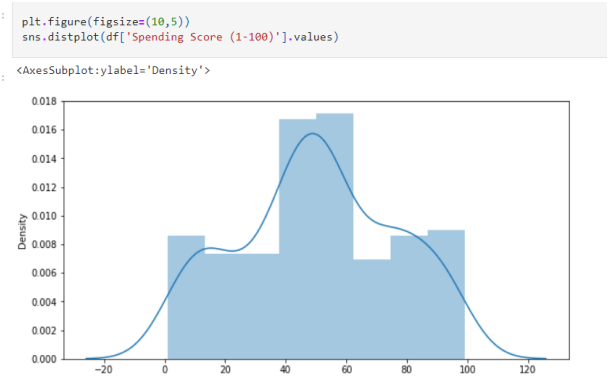

Fig 7: Plot of Spending Score

6. Creating Clusters Between Different Columns

Clustering is an unsupervised machine learning method which is used to identify and group similar data points in data set. Clustering helps to classify data into structures that can be used to understand and manipulate the data easily.

6.1 Annual Income and Spending Score

Fig 8 shows a cluster of customers' annual income and their spending score. Customers can be divided as per low spending, medium spending, and high spending score. It can be observed that people with annual income between 40 and 70k spend maximum on shopping.

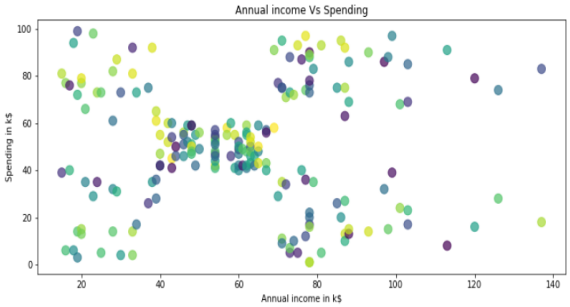

Fig 8: Cluster of Annual Income and Spending Score

6.2 Age and Spending

Fig 9 shows a cluster of customers' age and their spending score. It can be observed that people aged 30 and 50 spend maximum on shopping. There are few people in the age group above 60.

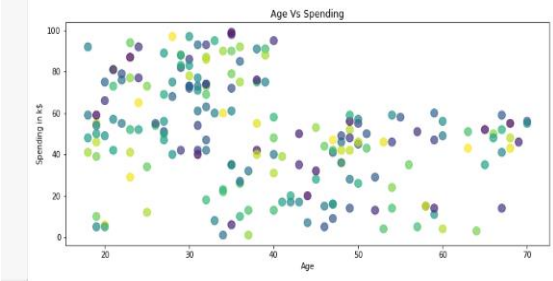

Fig 9: Cluster of Age and Spending Score

6.3 Age and Annual Income

Fig 10 shows a cluster of customers' annual income and age. It can be witnessed that people in the age of 30, 40 and 50 have maximum annual income between 40 and 70k.

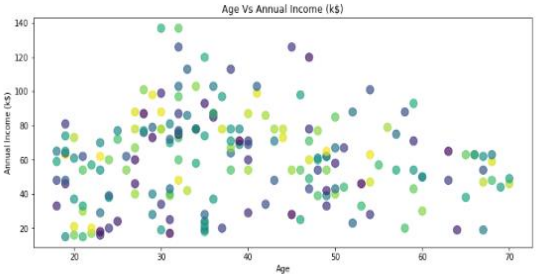

Fig 10: Cluster of Age and Annual Income

6.4 Age and Annual Income wrt Gender

Fig 11 shown below is the cluster of age and annual income with respect to gender. This cluster has a mix combination of males and females. More males in the age group of 30 and 50 have greater annual income. Females in the age group of 40 have maximum annual income.

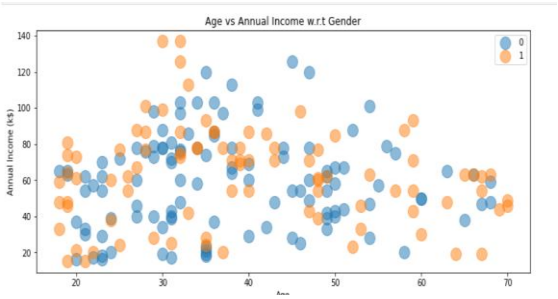

Fig 11: Cluster of Age and Annual Income wrt Gender

6.5 Annual Income and Spending wrt Gender

Fig 12 shown below is the cluster of annual income and spending score with respect to gender. This cluster has a mix combination of males and females. More males with annual income between 40 and 60K in the age group of 30 and 50 have greater annual income. Females in the age group of 40 have maximum annual income.

Fig 12: Cluster of Annual Income and Spending wrt Gender

7. Clustering Using K -means

7.1 Segmentation Using Age and Spending Score

Fig 13 shows the segmentation of age and spending score. Inertia is squared distance between centroids and data points. From this, 'N' clusters will be selected based on inertia. The elbow method is used to find out the optimal cluster which is '4' in this case.

```
#Selecting N Clusters based in Inertia (Squared Distance b
plt.figure(figsize = (12 ,5))
plt.grid(True)

plt.plot(np.arange(1 , 11) , inertia , 'o', alpha = 0.9)
plt.plot(np.arange(1 , 11) , inertia , '-' , alpha = 0.8)

plt.xlabel('Number of Clusters') , plt.ylabel('Inertia')
plt.show()
```

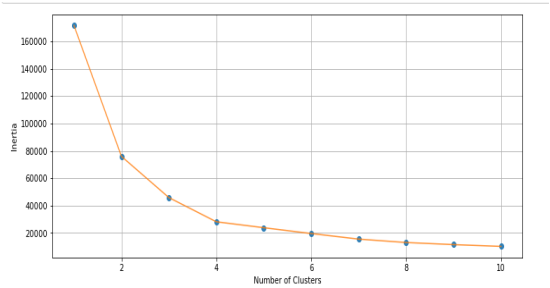

Fig 13: Segmentation using Age and Spending Score

7.1.1 Calculating Silhautte Score

Fig 13 shown the segmentation of customers using Age and spending score. The next step is to calculate silhautte score to find the optimal number of clusters. Fig 14 shows the silhautte score for the clusters with k = 2 to 8. The optimal cluster can be 4 where average silhautte score is 0.499.

```
print(f"for clisters= {n}, avg silhautte score is {silhautte_avg}")
```

```
for clisters= 2, avg silhautte score is 0.4692341232501655
for clisters= 3, avg silhautte score is 0.45300127016521263
for clisters= 4, avg silhautte score is 0.49973941540141753
for clisters= 5, avg silhautte score is 0.46553524067755037
for clisters= 6, avg silhautte score is 0.4376185638584134
for clisters= 7, avg silhautte score is 0.42313509747504796
for clisters= 8, avg silhautte score is 0.4304921688137185
```

Fig 14: Calculating Silhouette Score

7.2 Segmentation Using Annual income and spending score

Fig 15 shows the segmentation of annual income and spending score. Inertia is squared distance between centroids and data points. From this, 'N' clusters will be selected based on inertia. The elbow method is used to find out the optimal cluster which is '5' in this case.

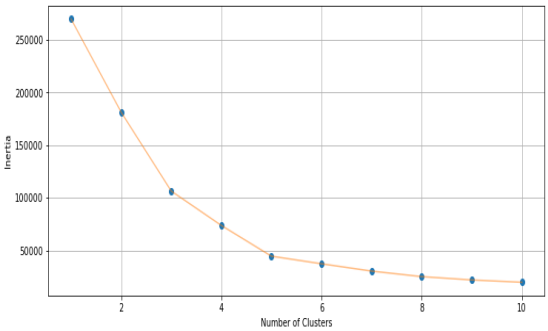

Fig 15: Segmentation using Annual Income and Spending Score

7.2.1 Calculating Silhautte Score

Fig 15 shown the segmentation of customers using Annual income and spending score. The next step

is to calculate silhautte score to find the optimal number of clusters. Fig 16 shows the silhautte score for the clusters with k = 2 to 8. The optimal cluster can be 5 where average silhautte score is 0.55.

```
    print(f"for clisters= {n}, avg silhautte score is {silhautte_avg}")

for clisters= 2, avg silhautte score is 0.2968969162503008
for clisters= 3, avg silhautte score is 0.46761358158775435
for clisters= 4, avg silhautte score is 0.4931963109249047
for clisters= 5, avg silhautte score is 0.553931997444648
for clisters= 6, avg silhautte score is 0.5376203956398481
for clisters= 7, avg silhautte score is 0.5270287298101395
for clisters= 8, avg silhautte score is 0.4572211842776841
```

Fig 16: Calculating Silhouette Score

7.3 Segmentation Using Age, Annual Income and Spending Score

Fig 17 shows the segmentation of annual income and spending score. Inertia is squared distance between centroids and data points. From this, 'N' clusters will be selected based on inertia. The elbow method is used to find out the optimal cluster which is '6' in this case.

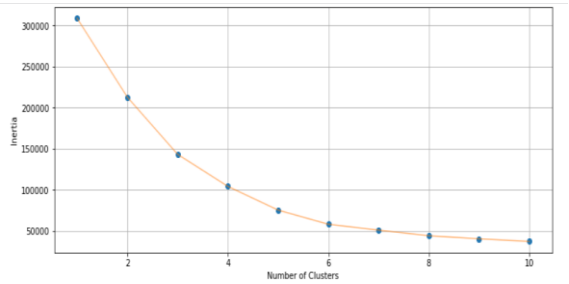

Fig 17: Segmentation using Age, Annual Income and Spending Score

7.3.1 Calculating Silhautte Score

Fig 17 shown the segmentation of customers using Annual income and spending score. The next step is to calculate silhautte score to find the optimal number of clusters. Fig 18 shows the silhautte score for the clusters with k = 2 to 8. The optimal cluster can be 6 where average silhautte score is 0.45.

```
for clisters= 2, avg silhautte score is 0.29307334005502633
for clisters= 3, avg silhautte score is 0.383798873822341
for clisters= 4, avg silhautte score is 0.4052954330641215
for clisters= 5, avg silhautte score is 0.44406692047430008
for clisters= 6, avg silhautte score is 0.45205475380756527
for clisters= 7, avg silhautte score is 0.43949619264530887
for clisters= 8, avg silhautte score is 0.4349105351263195
```

Fig 18: Calculating Silhouette Score

8. Recommendations: Future Work and Scope

Segmentation of customers based on their buying pattern and spending habits is a challenging task. Retaining customers is another major apprehension in both online and offline shopping. The research was performed based on understanding on e-commerce platform. In the present work, k means clustering was evaluated using Silhouette analysis with different number of clusters. Based on Silhouette score, optimal solution is found. An optimal solution means to find out how many clusters are optimal and who the targeted customers are. Based on those targeted customers, an effective marketing approach can be designed which will help to increase the profit of the company.

The current research findings and interpretation notify the decision makers to establish purchasing patterns and making decisions for effective management based on data analysis. This study targets to help e-commerce stakeholders to compare the structure of market and group them in small parts to take decisions effectively. People can be divided as low spending, medium spending and high spending behaviour. The business can determine appropriate product pricing (low to high) to keep for different types of customers based on their spending behaviour. When customers are divided into clusters, marketing campaigns can be customized as per customers' demand and choice. Another area, where e-commerce stakeholders can focus is designing an optimal distribution strategy. Specific products can be chosen for deployment. Also, if there are new products to be launched by company, these products can be prioritized as per customer's shopping behaviour.

Therefore, this research allows us to divide a unit from E-commerce performance data from multiple category stock and purchasing histories. In addition, it would capture the dissimilarity that can be perceived between the high profitable segments and low profitable segment.

9. References

[1] S. Sayyida, S. Hartini, S. Gunawan and S. Husin, "The Impact of the Covid-19 Pandemic on Retail Consumer Behavior," Aptisi Trans. Manag. (ATM) , p. 79–88, 2021.

[2] D. A. A. B. s. Deepti Sharma, "Augmenting Customer Satisfaction in Smartphone Based Online Shopping," International Journal of Future Generation Communication and Networking, vol. 13, no. 4, pp. 3609-3615, 2020.

[3] G. Bhaskara and V. Filimonau, "The COVID-19 pandemic and organisational learning for

disaster planning and management: A," . J. Hosp. Tour. Manag. , vol. 46, p. 364–375, 2021.
[4] S. V. a. V. R. Kayalvily Tabianan, "K-Means Clustering Approach for Intelligent Customer Segmentation Using Customer Purchase Behavior Data," Sustainability, p. 7243, 2022.
[5] F. Nie, Z. Li, R. Wang and X. Li, "An Effective and Efficient Algorithm for K-means Clustering with New Formulation," in IEEE Trans., 2021.
[6] D. Aggarwal, "Mobile technology adoption by Indian consumer," International journal of recent technology & engineering, vol. 5, 2019.
[7] P. Brandtner, F. Darbanian, T. Falatouri and C. Udokwu, " Impact of COVID-19 on the customer end of retail supply chains: A big," Sustainability, vol. 13, p. 1464, 2021.
[8] D. Aggarwal, "Determinants for Consumer Attitude towards Technology Enabled Grocery Procurement," International Journal of Engineering and Advanced Technology, vol. 9, no. 3, 2020.
[9] M. M. P. P. Anitha, "RFM model for customer purchase behavior using K-Means algorithm," Journal of King Saud University –Computer and Information Sciences, vol. 34, pp. 1785-1792, 2022.
[10] D. R. Khong, "How Marketing Causes Inequality," Asian J. Law Policy, vol. 1, p. 83–86, 2021.
[11] S. Janardhanan and R. Muthalagu, "Market segmentation for profit maximization using machine learning algorithms.," in J. Phys. Conf., 2021.
[12] I. Rachmawati, "Collaboration Technology Acceptance Model, Subjective Norms and Personal Innovations on Buying Interest," Int. J. Innov. Sci. Res. Technol., vol. 5, p. 115–122, 2020.
[13] R. Shirole, L. Salokhe and S. Jadhav, "Customer Segmentation using RFM Model and K-Means Clustering.," Int. J. Sci. Res. Sci. Technol., vol. 8, pp. 591-597, 2021.
[14] W. Q. S. Abdallah, "Customers Segmentation in the Insurance Company (TIC) Dataset," Procedia Computer Science, vol. 144, pp. 277-290, 2018.