International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

# Integrated Supervised and Unsupervised Learning for Mall Customer Segmentation

**A. Divya[1], Deekonda Siddartha[2], Pusuluri Sai Sukeerthi[2], Shivannagari Tharun Reddy[2], Chabuksawar Arun[2]**

[1]Assistant Professor, [2]UG Scholar, [1,2]Department of IT
[1,2]Malla Reddy College of Engineering and Management Sciences, Medchal, Hyderabad

## ABSTRACT

In today's modern era, innovation has become the driving force behind everything and everyone. In this highly competitive landscape, businesses and entrepreneurs strive to outdo one another. This surge in competition has led to a sense of insecurity and tension among businesses, as they seek to attract new customers while retaining their existing ones. To achieve this, providing exceptional customer service is essential, regardless of the size of the business. Understanding the needs of customers is crucial in offering top-notch customer support and promoting the most suitable products. With the abundance of online products available, customers often find themselves puzzled about what to purchase, while businesses struggle to identify the right target audience for their specific products. This is where "Customer Segmentation" comes into play. Customer segmentation is the process of grouping customers with similar interests and shopping behavior into specific segments, while separating those with different interests and shopping patterns into other segments. By doing so, businesses can gain valuable insights into their customer base and tailor their strategies accordingly. Customer segmentation and pattern extraction are pivotal aspects of a business decision support system. Each segment represents a group of customers who likely share common interests and shopping habits. By understanding these segments, businesses can cater to their customers more effectively and offer them products and services that truly meet their needs. Therefore, this project implements optimized unsupervised learning for customer segmentation.

**Keywords**: Business analytics, Customer segmentation, Supervised learning, Unsupervised learning.

## 1 INTRODUCTION

Mall customer segmentation is a marketing and business strategy that involves categorizing a mall's customer base into distinct groups based on certain shared characteristics, behaviors, or demographics. This process allows mall operators, retailers, and marketers to tailor their services [1], products, and marketing campaigns to the specific needs and preferences of each segment, ultimately improving the overall shopping experience and increasing sales. Here's an overview of mall customer segmentation:

**Demographic Segmentation:**

- Age: Grouping customers by age ranges, such as teenagers, young adults, middle-aged, or seniors.
- Gender: Categorizing customers as male, female, or non-binary.
- Income: Classifying shoppers based on income levels, such as low, middle, or high-income groups [2].
- Family Size: Segmenting customers based on family size or household composition.

**Psychographic Segmentation:**

- Lifestyle: Grouping customers by their interests, values, and activities, such as fitness enthusiasts, fashion-conscious individuals, or eco-conscious consumers [3].
- Personality Traits: Identifying personality traits that influence shopping behavior, such as extroverted shoppers who enjoy social shopping versus introverts who prefer solo experiences.
- Values and Beliefs: Segmenting customers based on shared values and beliefs, such as eco-friendly shoppers who prioritize sustainability.

## 2. LITERATURE SURVEY

In this global pandemic known as COVID-19, new terms were created such as Work from Home (WFH) and Study from Home (SFH) that requires people to stay at home and minimize outdoor activities including shopping. Large supermarkets were also opening E-commerce system to sustain their profit during this pandemic [1]. People started using online shopping website to purchase necessary items which is very convenient in this current situation.

E-commerce system has become more popular and implemented in almost all business areas. E-commerce system is a platform for marketing and promoting the products to customer through online [2]. Customer segmentation is known as dividing the customers into groups which shares similar characteristics. The purpose of customer segmentation is to determine how to deal with customers in each category to increase the profit of each customer to the business.

When customers receive too much information or unwanted details which is not related to their regular purchase or their interest on the products, it can cause confusion on deciding their needs. This might lead their customers to give up on purchasing the items they required and effect the business to lose their potential customers. The clustering analysis will help to categorize the E-commerce customer according to their spending habit, purchase habit or specific product or brand the customers interested in. In order to process the collected data and segment the customers, an unsupervised learning algorithm is used which is known as Kmeans clustering is used [3].

Online shopping is not anymore, a new, whereas most of the business are becoming online based. There are a number of online shopping platforms keep on increasing day by day. Since most traditional business started to implement E-commerce system in their business and E-commerce system has become trending, there are more competition in the field [4]. In order for a business to sustain for a longer term and be competitive, the business should know the ways to retain their customers. For an example, if an E-commerce system continuously display a customer the products that is expensive or above their shopping budget, then the customers may decide that this E-commerce system is not suitable for them. Therefore they may look for another online shopping platforms which usually leads to high churn in E commerce platform .Customers differ in personality and have various preferences. There is evidence that inequalities in marketing exists. As a result, having the same approach and marketing for every consumer is not effective [5] and those consumers may be the most vital to the business [6]. Therefore it can be stated that inequalities in marketing and inefficiencies may cause customer churn. It is critical for a company to segment its consumers and determine the distinctions between the customer segments. For an example, the prices in the retail business in Malaysia will be increase by 50–60 percent from the 1 June 2022 onwards hence one way to keep existing customers is customer segmentation. Market segmenting according to the customer purchase behavior is important to decide the likelihood of the customer buying a specific product [7]. In this study it explains how a business can run for longer term by understanding their customer need and interest and satisfy them. The aim of this

research is to conduct customer segmentation using the customer data and grouping their customers into groups that share similar criteria. Customer segmentation is carried out to find the potential and most profitable customer groups among the total customers [8]. Therefore, this helps to reduce the risk of losing the customer by selling the wrong product to the wrong customer group. Customer segmentation shows the way for E-commerce on how to make their business customer-focused and conquer a stable position in the business world.

Rachmawati et al. [9] proclaim that for the large and sophisticated data information of today's E-commerce businesses, accurate and efficient customer segmentation management should be carried out. In this competitive and developing E-commerce business, it is important to analyse the customer need and apply market segmentation mine and analyse various target customers in the system to provide different customers with distinctive marketing methods and improve their customer loyalty and satisfaction. According to Shirole et al. [10], customer segmentation is based on discovering important differentiators that split customers into target groups. A customer segmentation model allows organizations to target specific groups of customers, allowing for more effective marketing resource allocation and the maximization of cross- and up-selling capability. Customer segmentation can also help to enhance customer service and increase customer loyalty and retention. There are several aspects of online shopping behavior can be found that can influence the strategic approach in E-commerce for longer term which are the security of seller and buyer E-commerce, optimize the re-order and up-to inventory levels of e-groceries, payment method, delivery method of the item, delivery speed of the items, the user interface of E-commerce system, wide range of products choice and reasonable price for the service and products [11].

Once the variables were selected, a clustering process was carried out to detect the product popularity based on target customer group. In this study SAPK + Kmeans clustering algorithm is used because improvised algorithm, the square error result will be smaller compared to Affinity Propagation Algorithm (AP). AP Algorithm will not provide the number of cluster or the cluster center, but then, it will consider all the samples as the exemplar which known as potential cluster center. This improved algorithm is always searching for the optimal cluster center value and maximizing the objective function value during the implementation process [13].

## 3. PROPOSED SYSTEM

Customer segmentation is a crucial task in marketing and business analytics that involves dividing a customer base into distinct groups based on certain characteristics or behaviors. This segmentation can help businesses understand their customers better and tailor their marketing strategies to specific groups. The proposed steps of mall customer segmentation using dataset preprocessing, exploratory data analysis (EDA), finding the optimal number of clusters using the Elbow Method, and applying K-Means clustering.

### Data Preprocessing

Data pre-processing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model. When creating a machine learning project, it is not always a case that we come across the clean and formatted data. And while doing any operation with data, it is mandatory to clean it and put in a formatted way. So, for this, we use data pre-processing task.

A real-world data generally contains noises, missing values, and maybe in an unusable format which cannot be directly used for machine learning models. Data pre-processing is required tasks for cleaning

the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model.

**One-Hot Encoding**: Categorical variables are one-hot encoded to convert them into a numerical format suitable for machine learning models. The code uses the pd.get_dummies() function to create binary columns for each category within categorical variables ('sex', 'cp', 'fbs', 'restecg', 'exang', 'slope', 'ca', 'thal'). This transformation allows machine learning algorithms to work with categorical data effectively.

**Standardization**: Standard Scaler is applied to scale numeric features, ensuring that they have a mean of 0 and a standard deviation of 1. The 'Standard Scaler' from scikit-learn is used to standardize specific numeric features ('age', 'trestbps', 'chol', 'thalach', 'oldpeak'). Standardization is a common preprocessing step to bring features to a similar scale, which can improve the performance of some machine learning algorithms. This transformation is important for several reasons:

1. **Equal Scaling**: StandardScaler scales each feature to have the same scale. This is crucial for algorithms that are sensitive to the scale of features, such as gradient-based optimization algorithms (e.g., in neural networks) and distance-based algorithms (e.g., k-means clustering).
2. **Mean Centering**: By subtracting the mean from each data point, StandardScaler centers the data around zero. This can help algorithms converge faster during training and improve their performance.
3. **Normalization**: Scaling by the standard deviation normalizes the data, ensuring that features have comparable variances. This can prevent certain features from dominating others in the modeling process.
4. **Interpretability**: Standardized data is more interpretable because it puts all features on a common scale, making it easier to compare the relative importance of feature.

**Dataset Splitting**

In machine learning data pre-processing, we divide our dataset into a training set and test set. This is one of the crucial steps of data pre-processing as by doing this, we can enhance the performance of our machine learning model. Suppose if we have given training to our machine learning model by a dataset and we test it by a completely different dataset. Then, it will create difficulties for our model to understand the correlations between the models. If we train our model very well and its training accuracy is also very high, but we provide a new dataset to it, then it will decrease the performance. So we always try to make a machine learning model which performs well with the training set and also with the test dataset. Here, we can define these datasets as:

**Training Set**: A subset of dataset to train the machine learning model, and we already know the output.

**Test set**: A subset of dataset to test the machine learning model, and by using the test set, model predicts the output.

After running KMeans for each value of K, we'll have a set of SSD values. Now, we need to create a plot to visualize these values. On the x-axis, we'll have the values of K (the number of clusters), and on the y-axis, we'll have the corresponding SSD values.

**Analyze the Elbow Point**

In the resulting plot, we'll typically see a curve that resembles an "elbow." The point where the SSD starts to decrease at a slower rate is the "elbow point." This point represents the optimal number of clusters.

- If the elbow point is clear, it's relatively easy for us to determine the optimal K value. In the example above, it would be around K=3 or K=4.
- If the elbow is not very pronounced, we might need to use additional methods or domain knowledge to make a decision.

**Advantages of proposed system**

K-Means clustering has several advantages, which make it a popular choice for various data clustering tasks:

- Simplicity and Speed: K-Means is relatively easy to understand and implement. It's computationally efficient and can handle large datasets with ease, making it suitable for real-time or batch processing.
- Scalability: K-Means scales well with the number of data points and can handle high-dimensional data efficiently. This scalability is particularly valuable when dealing with big data.
- Versatility: K-Means can be applied to a wide range of data types and is not limited to any specific domain. It is commonly used in areas such as customer segmentation, image compression, document clustering, and more.
- Interpretability: The results of K-Means are easy to interpret. Each cluster represents a group of similar data points, allowing for meaningful insights and straightforward visualizations.
- Deterministic Results: Given the same initial conditions, K-Means will produce the same results. This determinism is useful for reproducibility and consistency in data analysis.
- Efficient for Large Datasets: K-Means doesn't require storing the entire dataset in memory during processing, making it memory-efficient and suitable for datasets that don't fit into memory.
- Robustness: K-Means can handle noisy data and outliers to some extent. However, preprocessing steps like outlier removal may be necessary for better results.
- Parallelization: K-Means is parallelizable, and various libraries and tools offer parallel implementations, which can significantly speed up the clustering process on multi-core processors or distributed computing environments.
- Initialization Methods: Advanced initialization methods like K-Means++ help improve convergence and reduce the chances of getting stuck in suboptimal solutions.
- Consistent Clusters: In most cases, K-Means produces relatively stable clusters over multiple runs, especially when using K-Means++ initialization.
- Quantization and Compression: K-Means can be used for image compression and data quantization, reducing data storage requirements while preserving essential information.

## 4. RESULTS AND DISCUSSION

This Python code is a script for performing data analysis and clustering on a dataset of mall customers. It uses the Pandas library for data manipulation, Matplotlib and Seaborn for data visualization, and Scikit-learn for K Means clustering. The dataset used for customer segmentation or analysis to understand spending patterns and preferences. The combination of age, gender, income, and spending score provides valuable insights into different customer segments and can be used for targeted marketing strategies. Figure 1 Displays a subset of the dataset, showing a few rows of data to give an overview of the information available. Figure 2 present a high-level summary or overview of the entire dataset, possibly including null value count & datatype of the columns that provide insights into the characteristics of the customers. Figure 3 provides statistics (such as mean, standard deviation,

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

minimum, maximum, quartiles) for the numerical attributes in the dataset, offering a quantitative summary of the data.

| | 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 |
| **...** | ... | ... | ... | ... | ... |
| **195** | 196 | Female | 35 | 120 | 79 |
| **196** | 197 | Female | 45 | 126 | 28 |
| **197** | 198 | Male | 32 | 126 | 74 |
| **198** | 199 | Male | 32 | 137 | 18 |
| **199** | 200 | Male | 30 | 137 | 83 |

200 rows × 5 columns

Figure 1: Sample dataset used for mall customer segmentation.

```
<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
```

Figure 2: Summary of mall customer segmentation dataset

| | CustomerID | Age | Annual Income (k$) | Spending Score (1-100) |
|---|---|---|---|---|
| **count** | 200.000000 | 200.000000 | 200.000000 | 200.000000 |
| **mean** | 100.500000 | 38.850000 | 60.560000 | 50.200000 |
| **std** | 57.879185 | 13.969007 | 26.264721 | 25.823522 |
| **min** | 1.000000 | 18.000000 | 15.000000 | 1.000000 |
| **25%** | 50.750000 | 28.750000 | 41.500000 | 34.750000 |
| **50%** | 100.500000 | 36.000000 | 61.500000 | 50.000000 |
| **75%** | 150.250000 | 49.000000 | 78.000000 | 73.000000 |
| **max** | 200.000000 | 70.000000 | 137.000000 | 99.000000 |

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

Figure 3: summary statistics for the numeric columns

Figure 4 Displays a count plot, visually representing the distribution of genders in the dataset. It shows how many individuals are categorized as male and how many as female. Figure 5 presents a visualization, likely a histogram, showing the distribution of ages in the dataset. It provides insights into the age demographics of the customers.

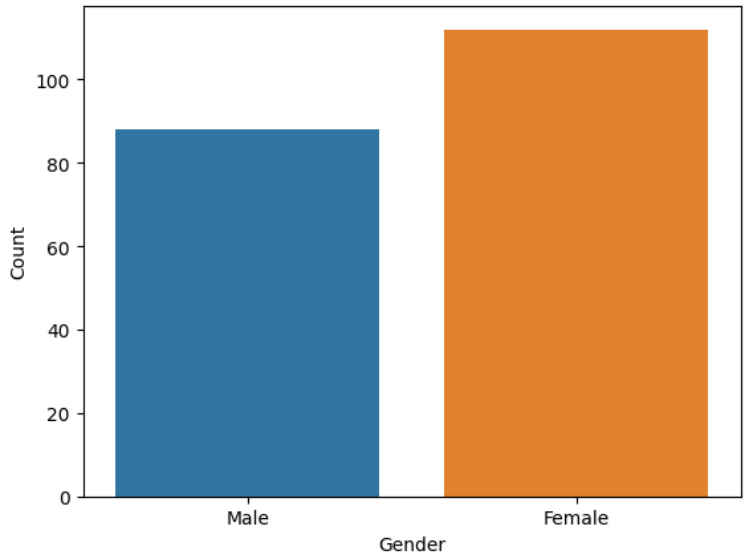


Figure 4: count plot for gender column of a dataset

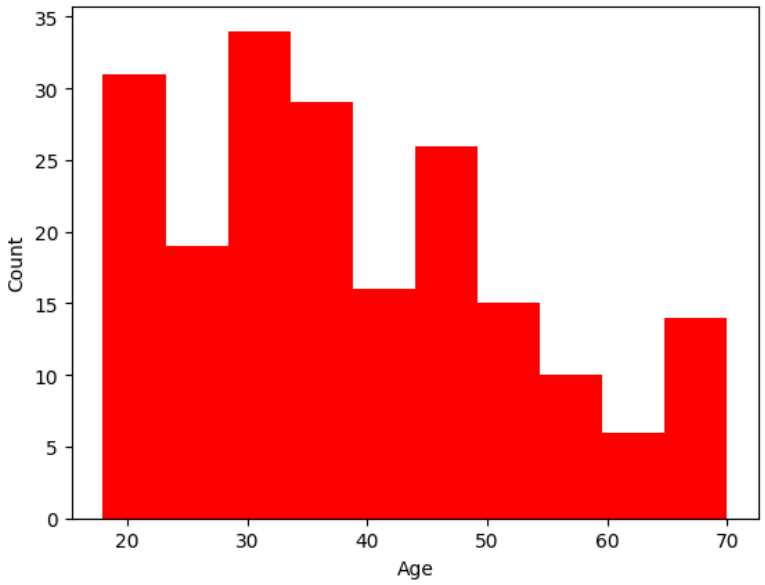


Figure 5: Distribution of ages in the dataset.

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

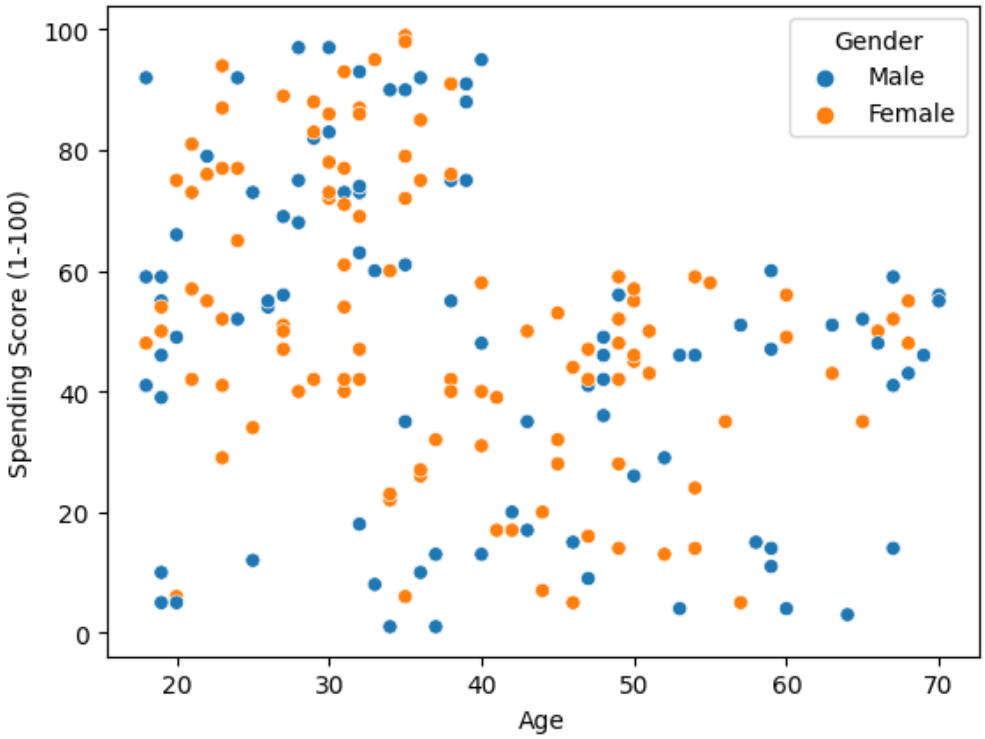


Figure 6: Scatter plot to check Relationship between age and spending score

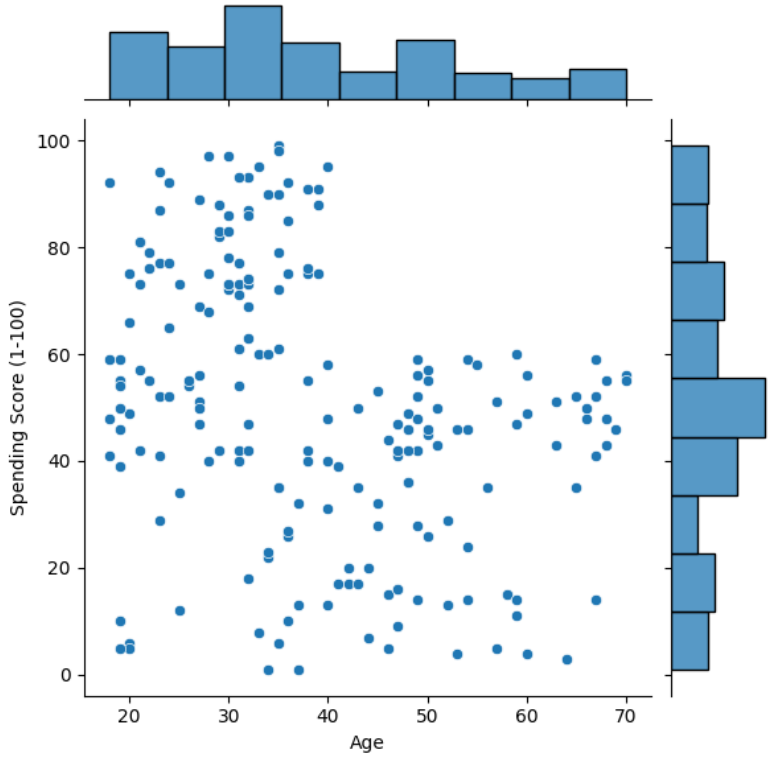


Figure 7: joint plot to check the relationship between age and spending score

## 5 CONCLUSION AND FUTURESCOPE

In conclusion, customer segmentation is a valuable technique for businesses to better understand and cater to the diverse needs and preferences of their customer base. Perform the dataset preprocessing to

clean, select relevant features, scale data, and encode categorical variables to prepare the dataset for analysis. Use summary statistics and visualizations to gain insights into the data, understand feature distributions, correlations, and detect outliers. Determine the most suitable number of clusters (K) by running KMeans clustering with a range of K values and plotting the sum of squared distances (SSD) for each K. The "elbow point" on the graph, where the SSD starts to decrease at a slower rate, indicates the optimal K. After determining the optimal K, apply KMeans clustering to segment customers into distinct groups based on their similarity. This involves iterative assignment of data points to clusters and updating cluster centers until convergence.

**REFERENCES**

[1] Sayyida, S.; Hartini, S.; Gunawan, S.; Husin, S.N. The Impact of the Covid-19 Pandemic on Retail Consumer Behavior. Aptisi Trans. Manag. (ATM) 2021, 5, 79–88.
[2] Bhaskara, G.I.; Filimonau, V. The COVID-19 pandemic and organisational learning for disaster planning and management: A perspective of tourism businesses from a destination prone to consecutive disasters. J. Hosp. Tour. Manag. 2021, 46, 364–375.
[3] Nie, F.; Li, Z.; Wang, R.; Li, X. An Effective and Efficient Algorithm for K-means Clustering with New Formulation. IEEE Trans. Knowl. Data Eng. 2022. Available online: https://ieeexplore.ieee.org/abstract/document/9723527/ (accessed on 2 May 2022).
[4] Brandtner, P.; Darbanian, F.; Falatouri, T.; Udokwu, C. Impact of COVID-19 on the customer end of retail supply chains: A big data analysis of consumer satisfaction. Sustainability 2021, 13, 1464.
[5] Khong, D.W.K. Rents: How Marketing Causes Inequality by Gerrit De Geest. Asian J. Law Policy 2021, 1, 83–86.
[6] Manero, K.M.; Rimiru, R.; Otieno, C. Customer Behaviour Segmentation among Mobile Service Providers in Kenya using K-Means Algorithm. Int. J. Comput. Sci. Issues 2018, 15, 67–76.
[7] Janardhanan, S.; Muthalagu, R. Market segmentation for profit maximization using machine learning algorithms. J. Phys. Conf. Ser. 2020, 1706, 012160.
[8] Dawane, V.; Waghodekar, P.; Pagare, J. RFM Analysis Using K-Means Clustering to Improve Revenue and Customer Retention. In Proceedings of the International Conference on Smart Data Intelligence (ICSMDI 2021), Online, 29–30 April 2021.
[9] Rachmawati, I.K. Collaboration Technology Acceptance Model, Subjective Norms and Personal Innovations on Buying Interest Online. Int. J. Innov. Sci. Res. Technol. 2020, 5, 115–122.
[10] Shirole, R.; Salokhe, L.; Jadhav, S. Customer Segmentation using RFM Model and K-Means Clustering. Int. J. Sci. Res. Sci. Technol. 2021, 8, 591–597.
[11] Ekren, B.Y.; Mangla, S.K.; Turhanlar, E.E.; Kazancoglu, Y.; Li, G. Lateral inventory share-based models for IoT-enabled E-commerce sustainable food supply networks. Comput. Oper. Res. 2021, 130, 105237.
[12] Sinaga, K.P.; Yang, M.S. Unsupervised K-means clustering algorithm. IEEE Access 2020, 8, 80716–80727.
[13] Kuruba Manjunath, Y.S.; Kashef, R.F. Distributed clustering using multi-tier hierarchical overlay super-peer peer-to-peer network architecture for efficient customer segmentation. Electron. Commer. Res. Appl. 2021, 47, 101040.