JOURNAL OF SOFT COMPUTING AND DATA MINING
e-ISSN: 2716-621X

JSCDM

Vol. 5 No. 1 (2024) 104-121
https://publisher.uthm.edu.my/ojs/index.php/jscdm

Examining the Behaviors and Preferences of Online Shopping Customers Using Clustering Techniques

Randa Mokhtar Hussein[1], Khai Wah Khaw[1]*, Amira Gaber[2], XinYing Chew[3]

[1] *School of Management,*
Universiti Sains Malaysia, 11800 Minden, Penang, MALAYSIA
[2] *System and Biomedical Engineering Department,*
Cairo University, EGYPT
[3] *School of Computer Sciences,*
Universiti Sains Malaysia, 11800 Minden, Penang, MALAYSIA

*Corresponding Author: khaiwah@usm.my
DOI: https://doi.org/10.30880/jscdm.2024.05.01.009

Article Info

Received: 1 December 2023
Accepted: 25 April 2024
Available online: 21 June 2024

Keywords

Customer segmentation, clustering algorithms, PCA, k-means, agglomerative, DBSCAN, silhouette score

Abstract

In recent years, e-commerce systems have expanded and been used by a variety of businesses. The e-commerce system is an online platform for selling and promoting products to customers. Customer segmentation is a technique of putting customers into groups based on shared traits. The purpose of customer segmentation is to determine a company's target market's purchasing habits, identify trends among various customer segments, and assess customer loyalty. This information helps the business develop targeted marketing campaigns that increase the number of profitable and devoted customers it serves. It is more challenging to secure a customer base in the age of globalization and digitization since consumers have been given many options for making purchases. While obtaining clients is the key objective of any firm, the results of this study will assist the organization in developing a range of research criteria to determine the strategies it uses for marketing. Certain unsupervised machine learning techniques were employed in this study to examine customer data. Clusters develop in unsupervised machine learning methods. This business must concentrate and dedicate all its resources to providing superior customer service to its consumers. In this paper, we have employed the technique of principal component analysis (PCA) for dimensionality reduction and k-means, agglomerative, DBSCAN to determine the customer's segment. The findings of this study are that k-means algorithms gave a 0.41 silhouette score when they clustered into 5 clusters, the agglomerative algorithms gave a 0.4 silhouette score when they clustered into 4 clusters, and DBSCAN algorithms gave a 0.44 silhouette score when they clustered into 2 clusters. The best model based on the performance is k-means.

1. Introduction

The term "e-commerce," or electronic commerce, refers to a type of business in which trade is conducted over the Internet. It altered the nature of business, encouraging the creation of new economic players, new business models and new chances for both businesses and consumers.

"E-commerce" is the exchange of funds and data to carry out transactions including the selling and acquisition of products and services over the internet [18]. E-commerce platforms make it easier to find product

This is an open access article under the CC BY-NC-SA 4.0 license.

information so you can compare and make decisions [28]. Businesses need to rethink their strategies for marketing because of the digitalization of both information and non-information items that were caused by advancements in technology and the expansion of the internet. The development of an online marketplace that challenges the traditional industry has led to heightened competition [28]. Electronic commerce has grown because of businesses involving the electronic market in their strategy to expand accessibility and get access to the worldwide market. In 2022, the global e-commerce market was predicted to grow to the number of USD 5.55 trillion. Online sales made up 17.8% of total sales in 2020, but in only five years, the share of e-commerce is predicted to climb to 24.5% by 2025, 37.6% [27].

Online purchasing is currently regarded as the most common trading pattern worldwide in the realm of e-business. The purchasing preferences of customers fluctuate in such an online setting. Selling organizations consequently need this superb customer-oriented marketing technique, which uses a clustering algorithm; an unsupervised learning method to predict client online behaviors based on customer segmentation [33]. Unsupervised learning is the technique of using unconventional analysis of data to extract valuable insights. In unsupervised learning, not one attribute is more significant than any of them, in contrast to supervised learning techniques that predict a target of interest. As a result, inputs exist but no supervising output occurs with unsupervised [9].

One of the increasingly significant uses of unsupervised learning techniques in machine learning and data science is customer segmentation, often known as market segmentation. Client segmentation is the practice of dividing a client base into multiple groups based on common traits that are important for marketing, like age, gender, income, and other purchasing patterns. Businesses may better understand their target audiences, identify the correct groups to target, and create marketing strategies that work for each of those groups by using segmentation [21].

In recent years, the competition between companies has increased in the field of e-commerce. The companies began to find solutions to be leaders in this field by investing more money in marketing strategies, but this was not enough to achieve the target they aimed to reach. Thus, they started to make customer segmentation to better understand the behavior of customers so it would be easier for them to make recommendations based on the behavior of each segment.

The objectives of this study are as follows:

RO1: To develop a set of best practices for implementing customer segmentation in the e-commerce industry.

RO2: To develop machine learning models for customer segmentation.

RO3: To compare unsupervised machine learning models to determine the best performance in predicting clusters for customer segmentation.

RO4: To provide companies in the e-commerce sector with recommendations and useful insights to help them to increase their sales and achieve more profits.

2. Literature Review

2.1 Machine Learning Algorithms (ML)

These articles, ([23]; [17]) defined Machine Learning as a scientific examination of algorithms and statistical models that computer systems employ to carry out a particular task without being explicitly programmed. Several methods are used in machine learning to address data-related issues. The type of method used will depend on the type of problem you want to solve, how many variables there are, and what form of model will work best. The author highlights that because machine learning can handle complicated and large datasets, automate processes, and enhance decision-making, it has attracted a lot of attention and importance. Mahesh also discusses reinforcement learning, unsupervised learning, and supervised learning—the three primary categories of machine learning [8]. A review of several machine learning algorithms and their uses in various fields is given in this paper [23].

According to [3], unsupervised machine learning is a kind of machine learning in which a model builds up patterns and structures from the data without the need for indicated examples or clear instructions. Unsupervised learning does not need predetermined class labels or target variables, in contrast to supervised learning. Rather, their emphasis lies in identifying unknown patterns, groups, and connections within the data and finding outliers, or dividing data into useful categories. They can be applied to dimensionality reduction, association rule mining, and clustering tasks [3].

This article, [24] emphasized the importance of clustering algorithms in combining related data points and guaranteeing dissimilarity between various groups. And covered the significance of customer segmentation in marketing and business analytics ([19]; [20]). An overview of the different clustering algorithms available for machine learning-based customer segmentation is given in this paper [24].

According to [26], clustering is a technique for locating common groups within a set of data. When compared to other group entities, the entities within each group are relatively more similar to each other. Cluster-based segmentation has been widely used in data-driven studies since the 1970s, particularly in marketing research [26].

2.1.1 PCA Analysis

Principal Component Analysis, or PCA for brief, is an unsupervised machine learning technique that enhances data interpretability while retaining the most information. It can be used to analyze large datasets with the greatest number of variables, features, or dimensions. For this reason, we must lower the dimensions of the data before passing features or attributes through a classifier. PCA minimizes the issue of overfitting during the model-training phase by reducing the number of features or variables. PCA is a statistical technique that transforms a set of correlated variables into uncorrelated variables by using an orthogonal transformation; as a result, the principal components that we derive should be independent of one another [5].

2.1.2 K-means Algorithm

The K-means is the most widely utilized and common algorithm for clustering data into the appropriate number of groups [30]. Within a customer base, homogeneous groups can be found using the K-means method [25]. It allows for the customization of marketing campaigns to suit each customer's preferences and needs by dividing a dataset into distinct clusters according to similarity. Iteratively assigning data points to the closest cluster centroid, the method first selects initial centroids and then updates centroids based on the mean of the allocated points [25]. The goal is to identify K cluster centroids so that the total squared distance between each cluster centroid and the data points is as small as possible. One can calculate the within-cluster sum of squares (WCSS) by dividing the total number of clusters by the sum of squared distances [16].

Centroid Determination

$$Ci = 1 /M \sum j=1 \text{ m } Xj \quad (1)$$

Euclidean Distance

$$d (p, q) = \sqrt{(p1-q1)2 + (p2-q2)2} \quad (2)$$

WCSS determined by

$$WCSS=\sum Ni=1 \sum Kk=1 rik \cdot ||xi-\mu k||2 \quad (3)$$

2.1.3 Hierarchical Clustering Algorithms

Hierarchical clustering or HCA establishes the cluster assignments. A dendrogram, or tree-like diagram with a hierarchy of points, is created by these algorithms. Using a depth-cutting dendrogram, clusters can be created. As a result, k groupings of smaller dendrograms will be produced by [15]. There are two ways to achieve hierarchical clustering: top-down clustering and bottom-up clustering. In hierarchical algorithms, the clustering process may start with a single cluster made up of individual data points. The hierarchical clustering algorithm produces a binary tree-based data structure known as a dendrogram in order to solve the clustering problem.

After the dendrogram is produced, the tree can be divided at different levels to obtain different clustering approaches for the same dataset rather than repeating the clustering process. It is both divisive (the top-down technique starts with one cluster and divides it into smaller clusters at each iteration) and agglomerative (the bottom-up approach starts with numerous clusters and combines them one by one at each iteration).

Euclidean distance

$$\sum (ai - bi) 2 \quad (4)$$

Manhattan distance

$$\sum |ai - bi| \quad (5)$$

Hierarchical clustering doesn't require a pre-defined value 'k' for clustering. It uses a dendrogram to identify what number of clusters will give the best result. The hierarchical clustering method may be an agglomerative

method or divisive method, differing on whether the breakdown is formed in a bottom-up (merging) or top-down (splitting) manner.

2.1.4 Agglomerative Hierarchical Clustering

The bottom-up strategy, also known as the agglomerative method, starts with each data point forming its own group. Up until all the groups become one group (the topmost level of the hierarchy) or until specific termination circumstances are met, it successively combines the data instances that are close to one another. The process of hierarchical clustering is often represented by a tree structure known as a dendrogram. It illustrates the grouping of data instances into a single entity. Similar data points are connected by lines whose vertical length indicates the distance between the data points. Data points are shown at the bottom ([31]; [13]).

2.1.5 DBSCAN Clustering

The Density-Based Clustering approach is the main algorithm called DBSCAN. The DBSCAN algorithm, as its name indicates, can extract unique groups or clusters of different sizes and forms from a huge amount of data that contains noise and outliers. The foundation of this technique is the idea that a cluster in the data space is an area of high point density, defined as such by the contiguous regions of low point density that divide it from other clusters. The data points that are "densely grouped" are combined into a single cluster by DBSCAN. By looking at the local density of these data points, this technique may locate clusters in big spatial datasets [10].

To determine the density estimation of the surrounding region of a particular data object, two input parameters are used: ε and minPts. To begin with, we must know the appropriate values for Eps and MinPts for each cluster, as well as at least one point from each corresponding cluster. ε is the radius of the object's local region (neighborhood), and minPts is the minimum number of data objects needed for the radius for it to be considered a cluster. With these suitable criteria, we can then get all density-reachable sites from the given point [12].

There are a few common types of unsupervised ML that are used in this study such as Principal component analysis (PCA), hierarchical clustering, k-means clustering, agglomerative hierarchical clustering, and DBSCAN. Summary of machine learning models shown in table 1:

Table 1 *Summary of machine learning models*

Machine Learning Models	Explanation
K-Means	Classify observations into mutually exclusive groups (or clusters), such that observations within the same cluster are as similar as possible, whereas observations from different clusters are as dissimilar as possible [1]; [22].
PCA	Analyze the data to identify patterns in order to reduce the dimensions of the dataset with minimal loss of information [1].
Hierarchical agglomerative clustering (HAC)	This study used the hierarchical agglomerative clustering algorithm on a Credit card dataset to perform customer segmentation. Based on the obtained results from this study, the analysts were able to promote the appropriate marketing strategies that are more profitable [12].
DBSCAN	DBSCAN algorithm identified objects based on bivalent logic [12], The existing techniques have not focused on the hybridization of DBSCAN with fuzzy [32].

2.2 Related Works Regarding Customer Segmentation by Using ML

In the marketing field, using customer segmentation is essential. Through the development of the most effective marketing plan, customer segmentation seeks to establish a relationship with the most profitable customers. Several statistical methods have been used to divide up the market, but their efficacy is significantly diminished by very large data sets. Optimizing the experimental similarity within a cluster and maximizing the dissimilarity between clusters is the goal of clustering. K-means clustering served as the basis for the segmentation that will be performed in this study; additional models will be used to confirm the findings. They also succeeded in

classifying the client into five clusters (k=5) based on the relationship between annual income and their spending score [26].

In the retailing field, using machine learning techniques with customer segmentation analysis. The retailer can see their data entirely differently. That's why retailers are searching for low-cost, straightforward ways to implement clustering and explain how it could be used to segment their customer base. This paper emphasizes customer segmentation by using machine learning algorithms such as (K-Means and Agglomerative Hierarchical Clustering). The findings are that there are five or six clusters of customers with each cluster having unique purchasing traits that define them [30].

In the telecom field, Telecommunications companies were looking to increase customer loyalty as a result of growing customer churn rates and increased competition among operators. In order to produce higher-quality clustering results, this paper investigates dimensionality reduction on an actual telecom dataset and assesses customers' clustering in reduced and latent space, relative to the original space. There are 220 features in the original dataset, which correspond to 100,000 customers. Reducing the original data's dimensionality in this manner. After that, both the original and reduced data sets are subjected to K-means clustering [2].

2.3 Related Works Regarding Customer Segmentation in E-Commerce by Using ML

Online shopping is growing as the most common trading pattern in China's e-business landscape. According to statistics, RMB 10,632.4 billion was sold online nationally in 2019. The purchasing habits of customers fluctuate in such an online setting. Selling organizations therefore desperately need a great customer-oriented marketing strategy for data mining-based online behavior prediction. This paper uses RFM and K-means clustering algorithms to analyze customer purchase behaviors in a systematic way based on a company's online transaction data. Four categories of customers are created based on the way they make purchases ([4]; [6]; [33]).

Throughout the years, data logs were used to record customer behavior related to e-commerce access and product viewing. The length of time a customer spends viewing a product can be used as a variable in customer segmentation to determine how interested they are in it. Using information, techniques, and procedures from a customer segmentation study, this paper will examine customer segmentation. There were two types of data used for customer segmentation: internal and external. Purchase history and customer profiles were handled as internal data. These data are processed through many techniques: Purchase affinity clustering, supervised clustering, unsupervised clustering, and customer likeness clustering [29].

3. Methodology

This section outlines the steps to conduct the study. In this study, the Python programming language was used throughout the project. The end-to-end flow consists of main steps performed to identify natural groupings or clusters within the data. Offering a deeper understanding of customer segments based on their personality traits and behavioral patterns.

3.1 Data Retrieval

The dataset that was used in this project is about customer personality analysis. This dataset which is publicly available on the Kaggle website is collected to study the behavior of customers in the e-commerce sector. It's provided by someone called Dr. Omar Romero-Hernandez and it consists of 2240 observations and 29 features. The data is diverse between numerical features (26 columns) and categorical features (3 columns).

The description of each attribute is shown below:

Table 2 *Dataset description*

Names	Description
ID	Customer's unique identifier
Education	Customer's education level
Year_Birth	Customer's birth year
Income	Customer's yearly household income
Marital_Status	Customer's marital status
Teenhome	Number of teenagers in customer's household
Kidhome	Number of children in customer's household

Recency	Number of days since customer's last purchase
Dt_Customer	Date of customer's enrollment with the company
MntWines	Amount spent on wine in last 2 years
Complain	1 if the customer complained in the last 2 years, 0 otherwise
MntMeatProducts	Amount spent on meat in last 2 years
MntFruits	Amount spent on fruits in last 2 years
MntSweetProducts	Amount spent on sweets in last 2 years
MntFishProducts	Amount spent on fish in last 2 years
NumDealsPurchases	Number of purchases made with a discount
MntGoldProds	Amount spent on gold in last 2 years
AcceptedCmp2	1 if customer accepted the offer in the 2nd campaign, 0 otherwise
AcceptedCmp1	1 if customer accepted the offer in the 1st campaign, 0 otherwise
AcceptedCmp4	1 if customer accepted the offer in the 4th campaign, 0 otherwise
AcceptedCmp3	1 if customer accepted the offer in the 3rd campaign, 0 otherwise
Response	1 if customer accepted the offer in the last campaign, 0 otherwise
AcceptedCmp5	1 if customer accepted the offer in the 5th campaign, 0 otherwise
NumCatalogPurchases	Number of purchases made using a catalog
NumWebPurchases	Number of purchases made through the company’s website
NumWebVisitsMonth	Number of visits to company’s website in the last month
NumStorePurchases	Number of purchases made directly in stores

3.2 Exploratory Data Analysis (EDA)

3.2.1 Descriptive Statistics

Descriptive statistics are used to characterize or summarize features of a dataset, such as the mean, standard deviation, frequency, min and max of variables. The function that is used to describe the statistics for numerical data is df.describe (). T and the function that is used to describe the categorical data is df.describe (include='object').

J. of Soft Computing and Data Mining Vol. 5 No. 1 (2024) p. 104-121

	count	mean	std	min	25%	50%	75%	max
ID	2240.0	5592.159821	3246.662198	0.0	2828.25	5458.5	8427.75	11191.0
Year_Birth	2240.0	1968.805804	11.984069	1893.0	1959.00	1970.0	1977.00	1996.0
Income	2216.0	52247.251354	25173.076661	1730.0	35303.00	51381.5	68522.00	666666.0
Kidhome	2240.0	0.444196	0.538398	0.0	0.00	0.0	1.00	2.0
Teenhome	2240.0	0.506250	0.544538	0.0	0.00	0.0	1.00	2.0
Recency	2240.0	49.109375	28.962453	0.0	24.00	49.0	74.00	99.0
MntWines	2240.0	303.935714	336.597393	0.0	23.75	173.5	504.25	1493.0
MntFruits	2240.0	26.302232	39.773434	0.0	1.00	8.0	33.00	199.0
MntMeatProducts	2240.0	166.950000	225.715373	0.0	16.00	67.0	232.00	1725.0
MntFishProducts	2240.0	37.525446	54.628979	0.0	3.00	12.0	50.00	259.0
MntSweetProducts	2240.0	27.062946	41.280498	0.0	1.00	8.0	33.00	263.0
MntGoldProds	2240.0	44.021875	52.167439	0.0	9.00	24.0	56.00	362.0
NumDealsPurchases	2240.0	2.325000	1.932238	0.0	1.00	2.0	3.00	15.0
NumWebPurchases	2240.0	4.084821	2.778714	0.0	2.00	4.0	6.00	27.0
NumCatalogPurchases	2240.0	2.662054	2.923101	0.0	0.00	2.0	4.00	28.0
NumStorePurchases	2240.0	5.790179	3.250958	0.0	3.00	5.0	8.00	13.0
NumWebVisitsMonth	2240.0	5.316518	2.426645	0.0	3.00	6.0	7.00	20.0
AcceptedCmp3	2240.0	0.072768	0.259813	0.0	0.00	0.0	0.00	1.0
AcceptedCmp4	2240.0	0.074554	0.262728	0.0	0.00	0.0	0.00	1.0
AcceptedCmp5	2240.0	0.072768	0.259813	0.0	0.00	0.0	0.00	1.0
AcceptedCmp1	2240.0	0.064286	0.245316	0.0	0.00	0.0	0.00	1.0
AcceptedCmp2	2240.0	0.013393	0.114976	0.0	0.00	0.0	0.00	1.0
Complain	2240.0	0.009375	0.096391	0.0	0.00	0.0	0.00	1.0
Z_CostContact	2240.0	3.000000	0.000000	3.0	3.00	3.0	3.00	3.0
Z_Revenue	2240.0	11.000000	0.000000	11.0	11.00	11.0	11.00	11.0
Response	2240.0	0.149107	0.356274	0.0	0.00	0.0	0.00	1.0

Fig. 1 *Descriptive statistics for numerical columns*

This figure displays descriptive statistics for all numerical variables by using .describe ().T and it found that the Z_CostContact and Z_Revenue columns have std = zero which means that these columns have no meaning and won't give any useful insights so, they will drop from data by using .drop () function in python.

	Education	Marital_Status	Dt_Customer
count	2240	2240	2240
unique	5	8	663
top	Graduation	Married	31-08-2012
freq	1127	864	12

Fig. 2 *Descriptive statistics for categorical columns*

This figure displays descriptive statistics for categorical variables by using .describe (include='object') and it shows that the Education variable has 5 categories, and the Graduation category is the most frequent with 1127. Marital Status has 8 categories, and the Married category has the most frequency with 864.

3.3 Data Pre-processing

This section will cover all pre-processing and transformation Steps including data cleaning, normalization, categorical variable encoding, feature scaling, and feature engineering. These procedures are crucial for guaranteeing that the data is in an appropriate format for analysis and modeling as well as for preparing the data for machine learning models.

3.3.1 Handling Missing Values

To ensure the reliability and accuracy of the analysis dataset and machine learning models, we handled missing values do not affect results.

First, we used isnull (). sum () function from pandas to check the numbers of nulls in the dataset. There were only 24 missing values in the income variable. The variable that contained missing values was a numerical

variable, so we replaced them with Median. Hence, the Median will not be affected by extreme values or outliers, we used fillna() function to fill them.

3.3.2 Handling Duplicate Rows

The duplicate rows are the rows that match perfectly across all columns. We used the .duplicated ().sum()function from pandas to check for duplicate rows in the dataset, the dataset has no duplicates, which means that all rows in the dataset are actually unique.

3.3.3 Removing Unnecessary Columns

By reducing the number of columns in a dataset, operations such as data loading, processing, and analysis can be performed more efficiently. This can lead to faster query times and improved overall performance. Some columns do not contribute any insights into the analysis such as 'ID' and 'Z_CostContact','Z_Revenue' have a standard deviation equal to 0 so, they didn't have meaning.

3.3.4 Removing Outliers

Outliers are extreme values that fall far from the rest of the data. They represent values that are significantly different from the norm. Outliers can be problematic because they can affect the results of an analysis. Some outliers represent true values from natural variation in the population. Other outliers may result from incorrect data entry. The boxplot () function is used to show numeric columns to determine the outliers in the dataset. The "Income" and the "Age" columns have outliers. We used the interquartile range (IQR). Outliers are commonly defined as any value 1.5 IQRs less than the first quartile or 1.5 IQRs greater than the third quartile. After we removed outliers the data shape became (2229, 27)

3.3.5 Feature Engineering

The feature engineering pipeline is the preprocessing steps that create new features from existing features or transform raw data into features that can be used in machine learning algorithms. In this dataset, new features are created from the existing features.

- 'Customer_for' feature was created from the 'Dt_Customer' column to count how many days that customer enrolled with the company system.
- 'age_category' feature was created from the 'Age' column to discover which category of age has high income to spend on shopping. From the figure, we can see that the older people have a high income.
- In 'Education' feature, we used .unique() function from pandas to display the unique values in education column and .replace() function from pandas to replace 'Basic', '2n Cycle', values with 'UnderGrade', 'Graduation' with 'Graduate' and 'PhD', 'Master' with 'PostGrade' as shown in figure, the graduates are highest category in dataset with 1124 followed by post-graduate category with 850.
- The .unique () function is used to display the unique values in 'Marital_Status' feature and .replace() function to replace 'Alone' with 'Single' and 'Absurd', 'YOLO' with 'NaN' to be more useful in analysis. The figure shows that the married customers are the highest category than others with 861.
- The 'Spending' feature was created by merging all 'MntWines', 'MntFruits', 'MntMeatProducts', 'MntFishProducts', 'MntSweetProducts' and 'MntGoldProds' features to calculate the total amount that customers paid in purchasing all products.
- The 'Total Purchase' feature was created by merging all 'NumDealsPurchases', 'NumWebPurchases', 'NumCatalogPurchases', 'NumStorePurchases' and 'NumWebVisitsMonth' features to calculate how many times the customer used different methods of purchases.
- The .rename() function was used to change the name of the 'Response' feature to 'AcceptedCmp6' name.
- The'TotalAccCom' feature was created from merging all 'AcceptedCmp1', 'AcceptedCmp2', 'AcceptedCmp3', 'AcceptedCmp4', 'AcceptedCmp5' and 'AcceptedCmp6' features to count how many customers that accepted each campaign. As shown in the figure, we can see that 1621 customers did not accept any campaign and 369 accepted the first campaign.

3.3.6 Encoding Categorical Variables

The encoding step is a process of converting categorical or textual data into a numerical format to be used in machine learning algorithms. In the dataset, there were 3 categorical variables; 'Education', 'age_category' and 'Marital_Status' that were converted into numerical variables by using the label encoding technique, First, the! pip install category_encoders were installed then, we used label encoding due to ordinal data in the first two

variables and the marital status was used label encoding instead of one-hot encoding because the one-hot technique increases features and the dataset has many features.

Table 3 *Encoding categorical variables*

Categorical values			Numerical values
Marital Status	**Education**	**Age Category**	
Single	Graduate	Young adult	0
Together	PostGrade	Adult	1
Married	UnderGrade	Senior	2
Divorced		Older	3
Widow			4
NaN			5

3.3.7 Feature Scaling

Feature scaling is one technique for normalizing the range of independent variables or features in a dataset. It is also referred to as data normalization in data processing and is often carried out in the data preprocessing phase. The *StandardScaler()* function was used to normalize the shape of data in range (1, -1).

3.3.8 Dimensionality Reduction

In general, the majority of clustering algorithms are not capable of processing high-dimensional data; instead, they perform better and are more accurate when the number of features is limited, or roughly less than ten attributes [13].

4. Results

In this section, the evaluation of PCA analysis, clustering model analysis and its results will be discussed; in addition, insight into the customer segments will be explained. Afterwards, a comparison of the machine learning models is presented.

4.1 PCA Analysis Results

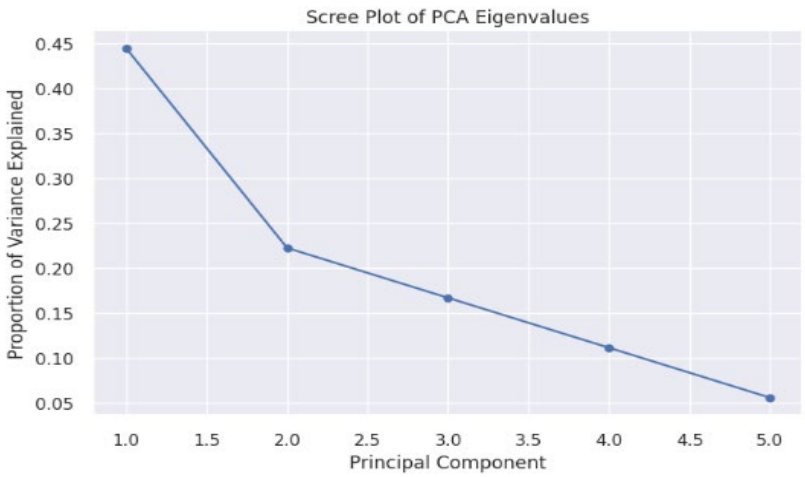


Fig. 3 *Scree plot for elbow of PCA*

Based on Figure 3, the number of PCA components = 2.

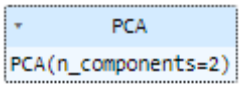


Fig. 4 *PCA*

4.2 Validation of Clustering Models

For data modeling, the evaluation of the clustering outputs is crucial. Either external criteria (evaluating the findings in relation to a pre-specified structure) or internal criteria (evaluating the results in relation to information pertaining to the data alone) might serve as the basis for the validation.). There are different measures for the internal validation of clustering such as; Davies-Bouldin Index, Silhouette coefficient and Dunn index [7].

4.3 Clustering Models Evaluation

In this section, we have used two evaluation methods; silhouette score and the Davies Bouldin Score to measure accuracy and performance for each model. This is a metric used to assess the efficacy of a clustering technique.

4.3.1 K-means Clustering Algorithms

In k-means clustering algorithms, we calculate silhouette score based on the number of clusters (k) from SSE calculations by using silhouette score () function, and we found that when k= 5, we got the highest silhouette score with a value equal to 0.41.

Elbow method for K-means:

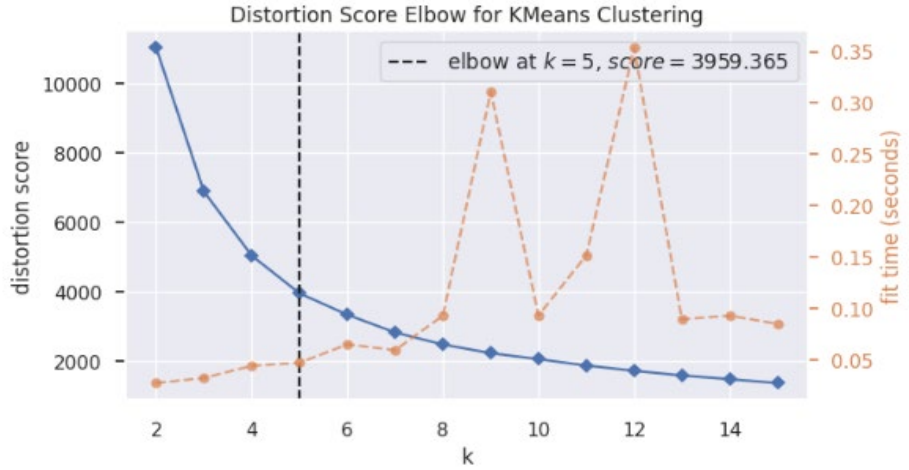


Fig. 5 *Elbow for K means*

The above figure displays the elbow method for k-means, and we decided that the dataset will be clustered into 5 clusters.

3D representation of clusters:

J. of Soft Computing and Data Mining Vol. 5 No. 1 (2024) p. 104-121

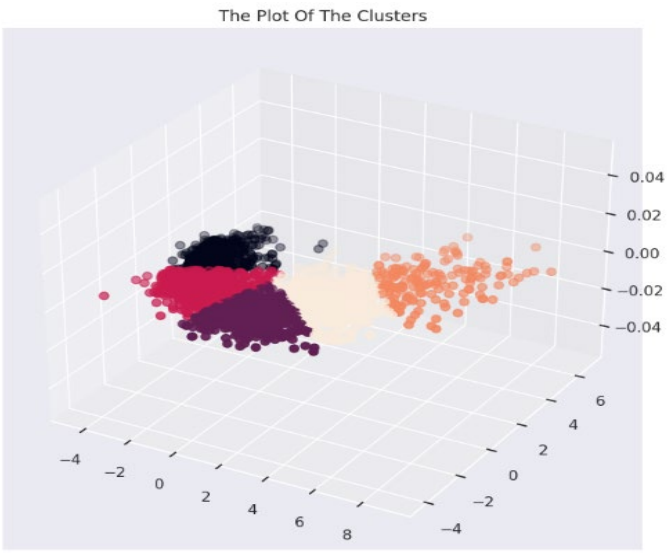


Fig. 6 *3D plot for k means clusters*

This figure represents the 5 clusters for k-means algorithms in a 3D plot.
Count of each cluster:

Fig. 7 *Distribution of clusters for k means*

This figure displays the number of customers that are included in each cluster, and we can see that cluster 1 contains 633, cluster 2 contains 426, cluster 3 contains 572, cluster 4 contains 170 and cluster 5 contains 428.
Income vs Spending scatter plot:

Fig. 8 *Scatter plot for k means*

4.3.2 Hierarchical Clustering Algorithms

In hierarchical clustering algorithms, we used a dendrogram for visualization which is a tree-like diagram. Figure 33 represents the dendrogram used to divide a cluster of data into many different clusters.
Dendrogram:

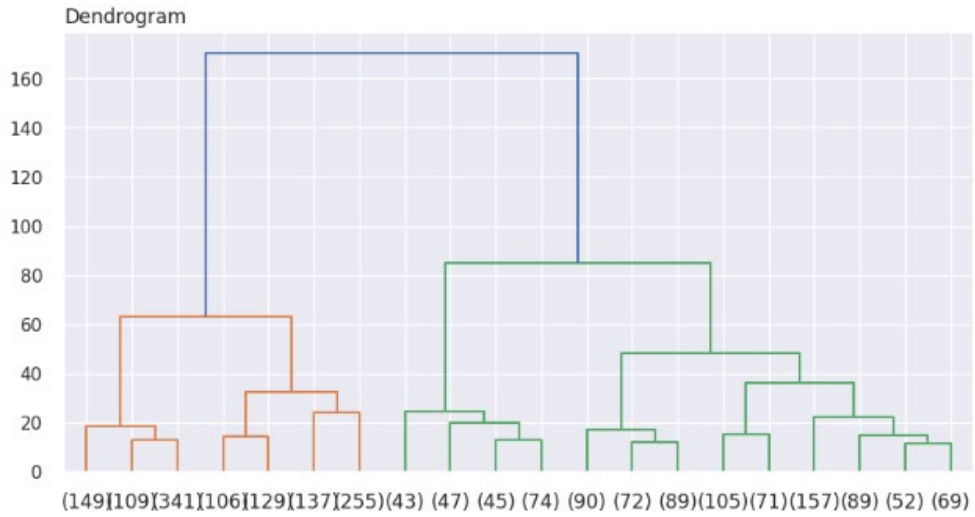


Fig. 9 *Dendrogram*

From this diagram, it is better when the dendrogram is cut when k =4.

4.3.3 Agglomerative Hierarchical Clustering

In the agglomerative hierarchical method, we calculate silhouette score by using *.silhouette_score ()* function from sklearn.metrics and we found that when k= 4, we got the highest silhouette score with a value equal to 0.40.

J. of Soft Computing and Data Mining Vol. 5 No. 1 (2024) p. 104-121

Count of each cluster:

Fig. 10 *Distribution of clusters for agglomerative*

This figure is a counting plot for agglomerative clusters, there are 794 in cluster 1, there are 627 in cluster 2, there are 209 in cluster 3 and there are 599 in cluster 4.

Income vs Spending scatter plot:

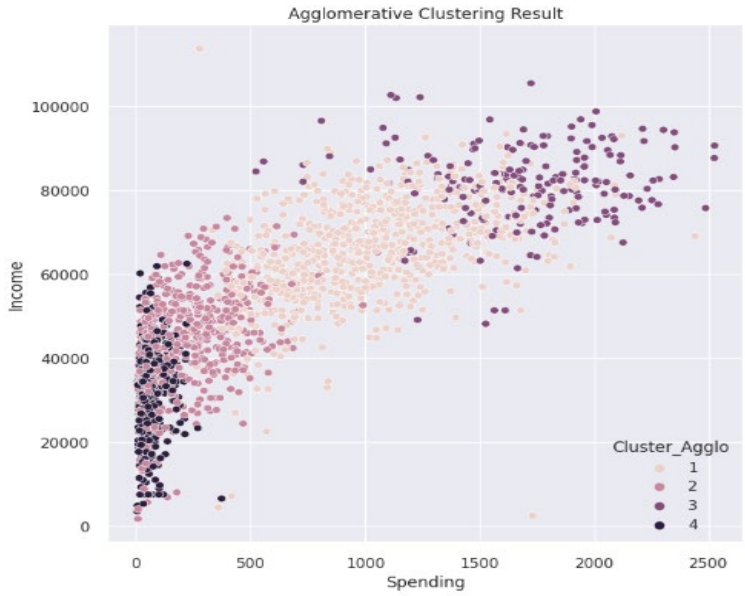


Fig. 11 *Scatter plot for agglomerative*

4.3.4 DBSCAN

The DBSCAN algorithm does not need to predetermine the number of clusters for performing the clustering. We used two parameters which are the epsilon and minPoints. We calculate silhouette score by using the *.silhouette_score ()* function from sklearn.metrics and we found that, when k= 2, we got the highest silhouette score with a value equal to 0.44 when parameters are eps= 0.8 and min_samples= 12.

Count of each cluster:

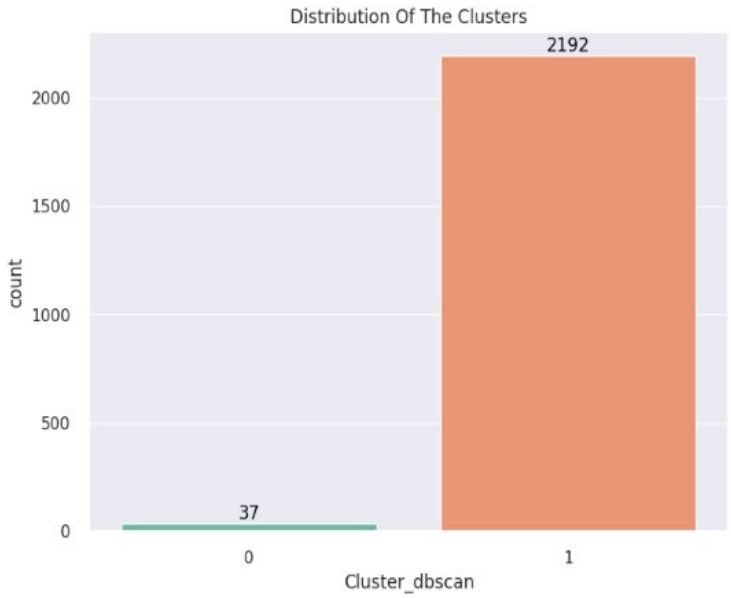


Fig. 12 *Distribution of clusters for DBSCAN*

This figure is the number of customers in each cluster, in cluster 0, there are 37 and in cluster 1, there are 2192.

Income vs spending scatter plot:

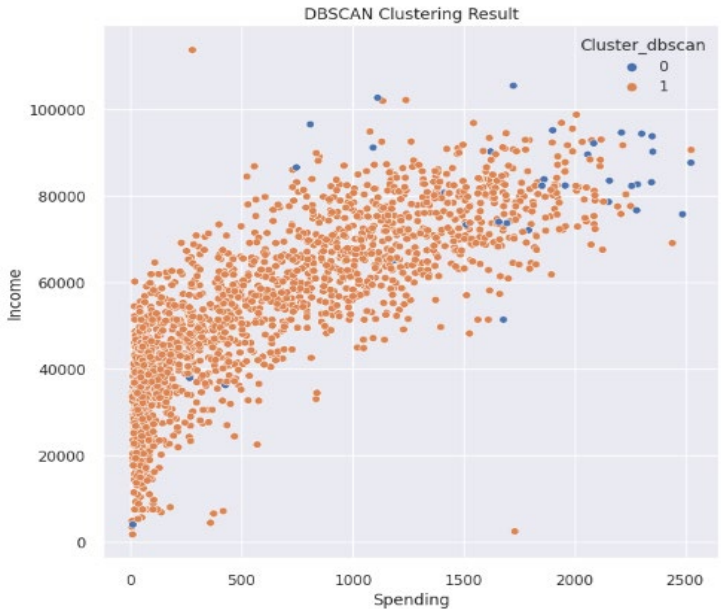


Fig. 13 *Scatter plot for DBSCAN*

4.4 Clustering Models Comparison

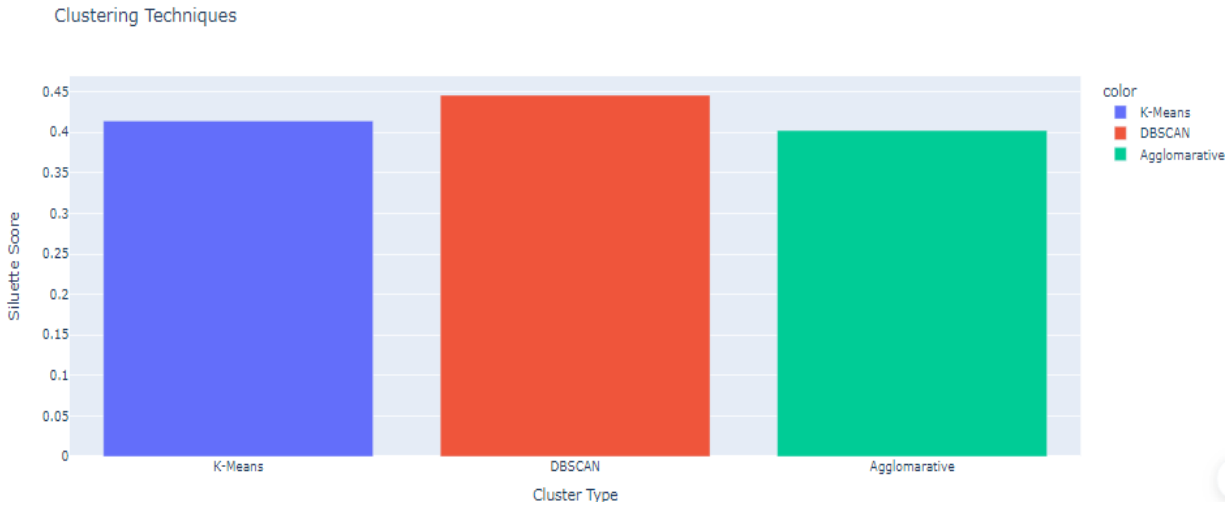
By comparing silhouette score results that we got from k-means, agglomerative and DBSCAN clustering algorithms, we found out that DBSCAN had the highest score with a value equal to 0.44 followed by k-means with a value equal to 0.41 and agglomerative with value equal 0.4 but DBSCAN clustered dataset into only 2 clusters, is don't perform well and this is not useful from business perspective, so we build our analysis in this study based on k-means and consider it as champion model.
Based on k-means clustering algorithms, the data clustered into 5 clusters.

J. of Soft Computing and Data Mining Vol. 5 No. 1 (2024) p. 104-121

Table 4 *Comparison between models*

	K-means	Agglomerative	DBSCAN
Silhouette score	0.41	0.40	0.44
Davies Bouldin Score	0.86	0.80	0.92

Fig. 14 *Comparison between models*



5. Discussion

In this study, we focused on conducting analysis on customers in the e-commerce sector to better understand the behavior of online shopping customers, we used different clustering algorithms for customer segmentation and based on the best performance model we divided data into 5 clusters and each cluster have special features:

About cluster 1:

From the above results, we can conclude that cluster 1 includes customers from the young adult category who have very low income and very low spending on purchases, half of them are undergraduates, some of them are graduates and a few of them are postgraduates. There are many customers who are in a relationship, a few of them are divorced and others are widows. Most customers in cluster 1 have one kid in their home and few of them have no kids, most of them don't have teenagers and few of them have one teenager.

About cluster 2:

From the above results, we can conclude that cluster 2 contains many customers from seniors and few of them from the young adult category who have average income and average spending on purchases. Half of them have a low educational level; the other half are graduates, and a few are postgraduates. There are many numbers in relationships, some of them are widowed and a few of them are divorced. There is no kid in the home for most of the customers and few of them have one kid but most of them have one teenager and few of them don't have one.

About cluster 3:

From the above results, we can conclude that cluster 3 contains customers who are seniors and a few of them are young adults. They have low average income and low spending on purchases. Most of them are undergraduates and graduates and few are postgraduates. Most of them are in relationships and widows and some of them are single and divorced. Half of them have one kid, some of them don't have and few have 2 kids but most of them have one teenager, some do not have one, and few have 2.

About cluster 4:

From the above results, we can conclude that cluster 4 contains customers who are adults; they have the highest income and highest spending on purchases. Half of them are undergraduates, some are graduates, and a few are postgraduates. There are many in relationships, some are widowed, some are divorced, and a few are single. They don't have kids or teenagers in their home.

About cluster 5:

From the above results, we can conclude that cluster 5 contains customers who are in the young adult and senior category. They are those with high average income and high average spending on purchases. Most of them are undergraduates, some of them are graduates and a few are postgraduates. Most of the customers are in a relationship, some of them are divorced, others are widows, and few are single. Most of the customers have no kids or teenagers in their home but few of them have 1 teenager.

6. Conclusion

In conclusion, the study's findings including clustering algorithms, patterns, and customer profiling, offer e-commerce companies the benefits of targeted marketing campaigns, improved products and services, customer loyalty, and opportunities for expanding channels and production. Companies should prioritize marketing efforts and allocate resources to target cluster 4 and cluster 5, characterized by high average incomes and spending on goods, by employing personalized incentives, special offers, and deals, while also addressing the needs of clusters 1 and 2 with lower incomes trough affordable options, promotions, and discounts to their purchasing behavior and expenditure. Future research should aim to increase the sample size to improve the generalizability and reliability of the findings, providing a better understanding of consumer behavior and demographic data. Exploring advanced methods like supervised learning and deep learning can extract more profound patterns and insights from consumer data beyond the unsupervised machine learning algorithms used in the current study. Additionally, conducting longitudinal analysis can help comprehend the evolving nature of consumer categories over time and adjust marketing strategies accordingly.

6.1 Limitations

Firstly, the data was limited by sample size; it consisted of 29 features and 2240 records only. These weren't sufficient to conduct this study to analyze behavior and demographic information of customers. There were other important features not included in the data such as address and gender that caused restriction in getting more insights.

The source of data was not disclosed from which company was collected, and the country which the data was collected from was not determined, which will lead to biased information about the e-commerce industry. We can't apply the findings in e-commerce in Egypt due to the changes in behaviors and demographic information for the customers.

The dataset had missing values; we couldn't replace them with the mean due to the outliers in the data. Machine learning models and statistical studies are significantly skewed by outliers, causing unreliable findings and conclusions. The accuracy of the analysis might also be impacted by missing values, particularly if they are not handled properly.

Acknowledgement

This work is funded by Ministry of Higher Education Malaysia, Fundamental Research Grant Scheme [Grant Number: FRGS/1/2022/STG06/USM/02/4], for the Project titled "Efficient Joint Process Monitoring using a New Robust Variable Sample Size and Sampling Interval Run Sum Scheme".

Conflict of Interest

Authors declare that there is no conflict of interests regarding the publication of the paper.

Author Contribution

The authors confirm contribution to the paper as follows: ***study conception and design:*** *Hussein RM, Khaw KW;* ***data collection:*** *Hussein RM, Chew XY, Khaw KW;* ***analysis and interpretation of results:*** *Hussein RM, Khaw KW, Chew XY, Gaber A;* ***draft manuscript preparation:*** *Hussein RM, Khaw KW, Gaber A. All authors reviewed the results and approved the final version of the manuscript.*

References

[1] Abdulhafedh, A. (2021). Incorporating k-means, hierarchical clustering and pca in customer segmentation. Journal of City and Development, 3(1), 12-30. https://www.researchgate.net/publication/349094412

[2] Alkhayrat, M., Aljnidi, M., & Aljoumaa, K. (2020). A comparative dimensionality reduction study in telecom customer segmentation using deep learning and PCA. Journal of Big Data, 7(1). https://doi.org/10.1186/s40537-020-0286-0

[3] Alloghani, M., Al-Jumeily, D., Mustafina, J., Hussain, A., & Aljaaf, A. J. (2019). A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science. In *Unsupervised and semi-supervised learning* (pp. 3–21). https://doi.org/10.1007/978-3-030-22475-2_1

[4] Alzami, F., Sambasri, F. D., Nabila, M., Megantara, R. A., Akrom, A., Pramunendar, R. A., Prabowo, D. P., & Sulistiyawati, P. (2023). Implementation of RFM Method and K-Means Algorithm for Customer Segmentation in E-Commerce with Streamlit. *ILKOM Jurnal Ilmiah*, *15*(1), 32–44. https://doi.org/10.33096/ilkom.v15i1.1524.32-44

[5] Ananda, L. G. (2023). A STUDY OF CUSTOMER PERSONALITY ANALYSIS USING MACHINE LEARNING ALGORITHMS. *Statistics and Data Science*, 108.
[6] Anitha, P., & Patil, M. M. (2022). RFM model for customer purchase behavior using K-Means algorithm. *Journal of King Saud University-Computer and Information Sciences*, *34*(5), 1785-1792. https://doi.org/10.1016/j.jksuci.2019.12.011
[7] Atienza, R. (2020). *Advanced Deep Learning with TensorFlow 2 and Keras: Apply DL, GANs, VAEs, deep RL, unsupervised learning, object detection and segmentation, and more*. Packt Publishing Ltd.
[8] Aurélien Géron. 2019. Hands-On Machine Learning with ScikitLearn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. O'Reilly.
[9] Brett Lantz. 2019. Machine Learning with R. Packt Publishing Ltd
[10] Bushra, A. A., & Yi, G. (2021). Comparative analysis Review of pioneering DBSCAN and successive Density-Based clustering algorithms. *IEEE Access*, *9*, 87918–87935. https://doi.org/10.1109/access.2021.3089036
[11] Deng, Y., & Gao, Q. (2020). A study on e-commerce customer segmentation management based on improved K-means algorithm. *Information Systems and e-Business Management*, *18*, 497-510. Sd. https://doi.org/10.1007/s10257-018-0381-3
[12] Ester, M., Kriegel, H., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial Databases with Noise. *Knowledge Discovery and Data Mining*, 226–231. https://www.aaai.org/Papers/KDD/1996/KDD96-037.pdf
[13] Han, J., Kamber, M., & Pei, J. (2012). Data mining: concepts and techniques. *Choice Reviews Online*, *49*(06), 49–3305. https://doi.org/10.5860/choice.49-3305
[14] Hung, P. D., Lien, N. T. T., & Ngoc, N. D. (2019, March). Customer segmentation using hierarchical agglomerative clustering. In *Proceedings of the 2nd International Conference on Information Science and Systems* (pp. 33-37). https://doi.org/10.1145/3322645.3322677
[15] Jayaratne, S. D. (2023). *Customer Segmentation Using Machine Learning* (Doctoral dissertation).
[16] John, J. M., Shobayo, O., & Ogunleye, B. (2023). An Exploration of Clustering Algorithms for Customer Segmentation in the UK Retail Market. *Analytics*, *2*(4), 809-823. https://doi.org/10.3390/analytics2040042
[17] Joung, J., & Kim, H. (2023). Interpretable machine learning-based approach for customer segmentation for new product development from online product reviews. *International Journal of Information Management*, *70*, 102641. https://doi.org/10.1016/j.ijinfomgt.2023.102641
[18] Kong, X.T.; Zhong, R.Y.; Zhao, Z.; Shao, S.; Li, M.; Lin, P.; Chen, Y.; Wu, W.; Shen, L.; Yu, Y.; et al. Cyber physical ecommerce logistics system: An implementation case in Hong Kong. Comput. Ind. Eng. 2020, 139, 106170. https://doi.org/10.1016/j.cie.2019.106170
[19] Kumar, A. (2022). Customer segmentation of shopping mall users using K-Means clustering. In *Advances in electronic commerce series* (pp. 248–270). https://doi.org/10.4018/978-1-6684-5727-6.ch013
[20] Li, Y., Qi, J., Chu, X., & Mu, W. (2023). Customer segmentation using K-means clustering and the hybrid particle swarm optimization algorithm. *The Computer Journal*, *66*(4), 941-962. https://doi.org/10.1093/comjnl/bxab206
[21] Liu, H., Huang, Y., Wang, Z., Li, K., Hu, X., & Wang, W. (2019). Personality or Value: A comparative study of psychographic segmentation based on an online review enhanced Recommender System. *Applied Sciences*, *9*(10), 1992. https://doi.org/10.3390/app9101992
[22] Lone, H., & Warale, P. (2022). Cluster Analysis: Application of K-Means and Agglomerative Clustering for Customer Segmentation. *Journal of Positive School Psychology*, 7798-7804.
[23] Mahesh, B. (2020). Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).[Internet]*, *9*(1), 381-386.
[24] Monil, P., Darshan, P., Jecky, R., Vimarsh, C., & Bhatt, B. R. (2020). Customer segmentation using machine learning. *International Journal for Research in Applied Science and Engineering Technology (IJRASET)*, *8*(6), 2104-2108. http://doi.org/10.22214/ijraset.2020.6344
[25] Ogunleye, B., Maswera, T., Hirsch, L., Gaudoin, J., & Brunsdon, T. (2023). Comparison of topic modelling approaches in the banking context. *Applied Sciences*, *13*(2), 797. https://doi.org/10.3390/app13020797
[26] Pradana, M. G., & Ha, H. T. (2021). Maximizing strategy improvement in mall customer segmentation using k-means clustering. *Journal of Applied Data Sciences*, *2*(1), 19-25. https://doi.org/10.47738/jads.v2i1.18

[27] Rita, P., & Ramos, R. F. (2022). Global research trends in consumer behavior and sustainability in E-Commerce: A bibliometric analysis of the knowledge structure. *Sustainability*, *14*(15), 9455. https://doi.org/10.3390/su14159455

[28] Rosário, A., & Raimundo, R. (2021). Consumer marketing strategy and E-commerce in the last decade: a literature review. *Journal of theoretical and applied electronic commerce research*, *16*(7), 3003-3024. https://doi.org/10.3390/jtaer16070164

[29] Sari, J. N., Nugroho, L. E., Ferdiana, R., & Santosa, P. I. (2016). Review on customer segmentation technique on ecommerce. *Advanced Science Letters*, *22*(10), 3018-3022. https://doi.org/10.1166/asl.2016.7985

[30] Shirole, R., Salokhe, L., & Jadhav, S. (2021). Customer Segmentation using RFM Model and K-Means Clustering. *International Journal of Scientific Research in Science and Technology*, 591–597. https://doi.org/10.32628/ijsrst2183118

[31] Shmueli, G., Patel, N. R., & Bruce, P. (2006). *Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner*. http://ci.nii.ac.jp/ncid/BB05276247

[32] Snehalatha, N., S, M. K., & Kachroo, V. (2023). CUSTOMER SEGMENTATION AND PROFILING FOR E-COMMERCE USING DBSCAN AND FUZZY C-MEANS. *Proceedings on Engineering Sciences*, *5*(3), 539–544. https://doi.org/10.24874/pes05.03.016

[33] Wu, J., Shi, L., Lin, W. P., Tsai, S. B., Li, Y., Yang, L., & Xu, G. (2020). An empirical study on customer segmentation by purchase behaviors using a RFM model and K-means algorithm. *Mathematical Problems in Engineering*, *2020*, 1-7.D https://doi.org/10.1155/2023/9873736