A Novel Approach for E-Commerce Customer Segmentation and Classification

1st Tanjina Saif Karim
Dept. of Computer Science and Engineering
University of Asia Pacific
Dhaka, Bangladesh
18201080@uap-bd.edu

2nd Homaira Adiba
Dept. of Computer Science and Engineering
University of Asia Pacific
Dhaka, Bangladesh
18201052@uap-bd.edu

3rd Mahbuba Haque Laka
Dept. of Computer Science and Engineering
University of Asia Pacific
Dhaka, Bangladesh
18201085@uap-bd.edu

4th Dr. Bilkis Jamal Ferdosi
Dept. of Computer Science and Engineering
University of Asia Pacific
Dhaka, Bangladesh
bjferdosi@uap-bd.edu

***Abstract*—Customer segmentation is a process that divides customers into groups based on common characteristics. The customer segmentation problem belongs to the domain of unsupervised learning, more specifically clustering. The effectiveness of customer segmentation distinctly depends on the chosen clustering algorithm. Moreover, the efficacy of a clustering algorithm is highly dependent on the dataset, type of data, utilised subspaces, and complexity, etc. However, different e-commerce or internet-based businesses collect and utilise their customer data differently and even the slightest difference in data might require a different clustering algorithm for effective customer segmentation.**

In this paper, we propose a system which consists of two modules, an unsupervised module and a supervised module. The unsupervised module will utilise unlabelled customer data and apply different categories of unsupervised clustering algorithms to find the most suitable algorithm for a given dataset. We use the acquired results to convert the unlabelled customer data into labelled data. After training a classification model using the labelled data, the supervised module can identify the groups of new customers using the trained model without further clustering. This system will work as a customer segmentation and identification system which will help businesses take data-driven decisions more efficiently.

***Index Terms*—Customer segmentation, Clustering, Classification, and Data visualization.**

I. Introduction

The e-commerce industry is currently one of the fastest-growing industries. The increasing number of e-commerce or internet-based businesses corresponds to an increase in competition and challenges. Understanding and addressing these challenges efficiently is crucial to sustaining a business in the industry. Customers are the key focus of every business. Hence companies need to understand the preferences and requirements of their respective customers. Customer segmentation is the process that divides customers into groups based on common characteristics. It plays a beneficial role in market segmentation, analysis, target marketing, advertisements, recommendation systems, etc. [1] Customer segmentation allows businesses to become more client-centric and enhances user experience. The customer segmentation problem belongs to the domain of unsupervised learning and cluster analysis that identifies similarities and differences between objects based on their characteristics. Data clustering enables us to discover patterns within data and make decisions based on them. The effectiveness of customer segmentation highly depends on the chosen clustering algorithm. But it cannot be easily determined which is the best algorithm for customer segmentation in general. Furthermore, The efficacy of a clustering algorithm highly depends on the dataset, type of data, used parameters, data dimensionality, and complexity, etc. There have been several studies on customer segmentation. R. Punhani et al. [2], proposed the application of the K-Means clustering algorithm on an e-commerce dataset to identify patterns like products having the highest sales and determining the most used payment method. In 2021, Nilashi et al. [3] developed a new method to analyse a large set of open data in social networking sites for travellers segmentation and predict tourists' choices and preferences using dimensionality reduction and deep learning techniques. Hossain et al. [4] performed a comparison between centroid-based algorithms and density-based algorithms for customer segmentation. The results imply that both algorithms can be used for the purpose of customer segmentation but the DBSCAN algorithm performed comparatively better than the K-Means algorithm for the considered dataset. In 2016, Rezaeinia and Rahmani [5] proposed a new method to increase the precision and quality of recommendation systems associated with filtering systems. Recommendations of corresponding clusters were extracted using the weighted RFM variables, expectation-maximization clustering algorithms, and their combination with the k-nearest neighbours (KNN) algorithm. The results of the proposed method were far better than the outcome of existing conventional collaborative filtering methods. The outcome of most of the previous work is confined or applicable to a particular dataset that is similar to the ones they have used. But different e-commerce or internet-based businesses collect and utilise their customer data differently. Hence, There is a need for a more tangible system. The slightest difference in data might require another clustering algorithm for effective customer segmentation. Hence, an algorithm that performs well for one customer dataset might perform worse for other datasets for the slightest difference. Moreover, the number of customers of e-commerce platforms is constantly increasing. According to the current customer segmentation approaches it is required to perform clustering or segmentation again and again to keep the system updated. Considering hundreds and thousands of customers of an e-commerce platform, such an approach might not be practical and sustainable. In this paper, we propose a system for customer segmentation that is divided into two modules which are an unsupervised module and a supervised module. Firstly, the unsupervised module will utilise unlabelled customer data and apply different categories of unsupervised clustering algorithms such as K-Means, DBSCAN, Hierarchical clustering, Affinity Propagation, etc. to form clusters. Cluster evaluation metrics such as Calinski-Harabasz Index, Davies Bouldin Index, and Silhouette's Coefficient are used to evaluate the performances of each clustering algorithm. Then the clusters will be re-evaluated using data visualization which requires human intervention. After cross-evaluation a suitable clustering algorithm is selected for a given unlabelled dataset. We reuse the acquired number of clusters to convert the unlabelled customer data into labelled customer data. In the supervised module, the labelled data is used. So, we can identify the groups of new customers using supervised classification models such as Support vector machine without further clustering. Hence, we can classify new customers into groups without performing clustering repeatedly. The proposed system will work as a customer segmentation and identification system which will help businesses take data-driven decisions more efficiently. Moreover, The proposed system can be tangible and scalable according to a company's business model and the number of customers.

II. Related Works

There have been several studies on customer segmentation. Punhani et al. [2] proposed the application of clustering algorithms for effective customer segmentation. The paper used the K-means clustering algorithm on an e-commerce dataset to identify patterns like products having the highest sales and determining the most used payment method.

Nilashi et al. [3], proposed a method for segmenting travelers' data from social networking sites to predict tourists' choices and preferences using dimensionality reduction and deep learning techniques. A deep belief network, self-organizing map, and latent Dirichlet allocation were performed in a text-based dataset crawled from the travel-based website TripAdvisor.

In 2016, Rezaeinia and Rahmani [5], proposed a new method to increase the precision and quality of recommendation systems associated with filtering systems. Recommendations of corresponding clusters are extracted using weighted RFM variables, expectation-maximization clustering algorithms, and their combination with the k-nearest neighbors (KNN) algorithm. The proposed method is compared with the outcome of the traditional collaborative filtering method, and it is seen that the effect is far better than the outcome of existing conventional collaborative filtering methods.

Hossain et al. [4], performed customer segmentation by applying clustering techniques such as centroid-based K-means and density-based clustering algorithm DBSCAN. The author interpreted that DBSCAN showed more valuable insights than K-means. As it provided extra information about customers through the identified outliers.

Jun Wu et al. [6], proposed a new method to conduct customer segmentation and value analysis by using online sales data. The proposed method uses RFM model which uses recency, frequency, and monetary values and K-means clustering algorithm. Every dimension of customer information is analyzed using the RFM model and K-means algorithm to classify target customers. Customers are classified into four groups based on their purchase behaviors. On this basis, different CRM (customer relationship management) strategies are brought forward to gain a high level of customer satisfaction. This proposed method shows the results of some key performance indices such as the growth of active customers, total purchase volume, and total consumption amount.

III. Methodology

This section presents the proposed process of customer segmentation and identification system. The proposed methodology is shown in fig 1. The system consists of two main modules: an unsupervised module for customer segmentation and a supervised module for customer classification or identification. The unsupervised module consists of steps such as subspace selection, clustering using different algorithms, performance evaluation of the clustering algorithms, cross-evaluation and selection of suitable algorithms. Then, we convert the unlabelled data into labelled data using the acquired clusters from the selected clustering algorithm. In the supervised module we train the classification model using the labelled data and predict groups for new customers using the trained model. We can classify new customer data and determine its corresponding cluster without performing clustering repeatedly. Hence, the system can be reused multiple times while performing clustering only once as we create a trained model to identify customers' corresponding clusters or groups.

The proposed system is implemented step by step as follows:

A. Subspace Selection

Subspace selection refers to the selection of one or more dimensions. [7] It can be done manually or automatically. Manual selection can be done by the user where they can select the number of subspaces through an interface. The combination of all subspaces can also be selected. Lastly, the user can specify the dimensions to form the subspaces according

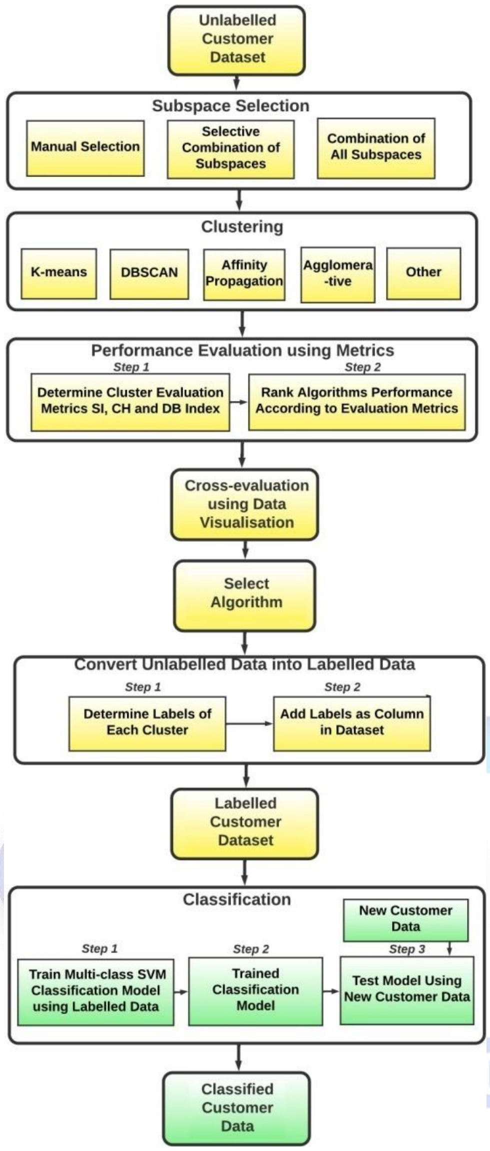

Unlabelled Customer Dataset
Subspace Selection
Manual Selection
Selective Combination of Subspaces
Combination of All Subspaces
Clustering
K-means
DBSCAN
Affinity Propagation
Agglomera -tive
Other
Performance Evaluation using Metrics
Step 1
Determine Cluster Evaluation Metrics SI, CH and DB Index
Step 2
Rank Algorithms Performance According to Evaluation Metrics
Cross-evaluation using Data Visualisation
Select Algorithm
Convert Unlabelled Data into Labelled Data
Step 1
Determine Labels of Each Cluster
Step 2
Add Labels as Column in Dataset
Labelled Customer Dataset
Classification
New Customer Data
Step 1
Train Multi-class SVM Classification Model using Labelled Data
Step 2
Trained Classification Model
Step 3
Test Model Using New Customer Data
Classified Customer Data

Fig. 1. Proposed Methodology

to the following equation. The equation of combination is as follows:

$$_nC_r = \frac{n!}{r!(n-r)!} \tag{1}$$

Where,
$_nC_r$ = number of combinations
n = total number of objects in the set
r = number of choosing objects from the set

B. Clustering

In the clustering phase of the unsupervised module different categories of clustering algorithms are applied to form clusters. The algorithms are centroid-based K-means, density-based algorithm DBSCAN, Agglomerative Hierarchical Clustering and Exemplar-based Affinity Propagation algorithm. For simplicity, we have considered the best algorithm of each category. This part can be easily scaled by adding more clustering algorithms. The clustering algorithms are described below,

1) K-means: K-Means is one of the simplest centroid-based unsupervised learning algorithm. This algorithm tries to minimise the variance of data points within a cluster. The success of this algorithm depends on the initial conditions set by the user or programmer. It clusters n objects into k groups. Let X = x_i where i = 1, 2, 3,..., n be the set of d-dimensional points and C = C_k, k = 1, 2, 3, ..., k where, k represents the centroids of the clusters.

2) DBSCAN: DBSCAN refers to density-based spatial clustering of applications with noise. It groups the points that are close to each other based on a euclidean distance measurement and a minimum number of points based on a set of data points. Moreover, It can create a separate cluster for outliers which is usually labelled as -1. The DBSCAN algorithm uses two parameters:

- minPts: It is the bare minimum of points or threshold to cluster together in order to classify a region as dense.
- eps(ε): It is a distance unit used to identify the points that are close to a given point.

3) Agglomerative: In Agglomerative algorithm, every data point is initially considered as an individual cluster. At every step, it merges the nearest pairs of the cluster in a bottom-up approach. This process continues till one cluster is formed. Single linkage, commonly referred to as the nearest neighbour method, is one of the simplest agglomerative hierarchical clustering techniques.
In the single linkage method, $D(r, s)$ is computed as: $D(r, s) = Min\ d(i, j)$, where, object i is in cluster r and object j is in cluster s. Every potential item pair (i, j) has a distance between them that is calculated. The distance between clusters r and s is said to be the distance with the lowest value among these distances.

4) Affinity Propagation: The Affinity Propagation algorithm identifies exemplars among data points and forms clusters of data points around the exemplars. It works by concurrently considering all data points as potential exemplars and exchanging messages between data points until a good set of exemplars and clusters are generated [8]. Moreover, It is not required to specify the number of clusters beforehand.
Let, x_1 through x_n be a sequence of data points, with no assumptions about their internal structure. Let, s be a function that quantifies the similarity between any two points, and $s(i, j)$ is greater than s. In $s(i, k)$, x_j is more similar to x_k. Here, the negative squared distance of the two data points is considered. For points x_i and x_k,

$$s(i, k) = -||x_i - x_k||^2 \tag{2}$$

The diagonal of s, i.e. $s(i, i)$ is used to represent the instance preference, meaning how likely a particular instance is to become an exemplar. When it is set to the same value for all inputs, it controls how many classes the algorithm produces. A value close to the minimum possible similarity produces fewer classes, while a large enough number produces many.

C. Performance Evaluation using Metrics

The performance of each clustering algorithm will be measured using cluster evaluation metrics such as Silhouette's Coefficient, Calinski-Harabasz Index and Davies Bouldin Index. The metrics are as follows.

1) Silhouette's Coefficient: Silhouette analysis is used to study the separation distance between the clusters formed by an algorithm. The distance between the clusters can be calculated by different types of distance metrics such as Euclidean, Manhattan, etc. [9]. The Silhouette Coefficient is calculated by using the mean of the intra-cluster and nearest cluster distance for all the data points. It ranges from [-1,1]. A value closer to +1 or a higher value of Silhouette Coefficients means that there is more separation between the clusters. If the value is 0 it indicates that the data point is on the decision boundary or very close between two neighbouring clusters. However, a negative value indicates that data points might have been assigned to the wrong cluster [10]. It is denoted as follows:

$$\text{Silhouette Score} = \frac{(b - a)}{max(a, b)} \tag{3}$$

Where,
a = average intra-cluster distance and b = average inter-cluster distance.

2) Calinski-Harabasz Index: The Calinski-Harabasz index is the ratio of the sum of between-cluster dispersion and the sum of inter-cluster dispersion for all clusters. It is based on the principle of variance ratio. A higher score implies better performance. It is also known as the Variance Ratio Criterion [11]. It is calculated using the equation given below,

$$CH(k) = \frac{B(k)}{W(k)} \cdot \frac{(n - k)}{(k - 1)} \tag{4}$$

Where,
n = number of data points, k = number of clusters, $W(k)$ = within cluster variation, and $B(k)$ = between cluster variation.

3) Davies Bouldin Index: Davies Bouldin index is based on the principle of intra-cluster and inter-cluster distances. It is generally used for deciding the number of clusters in which the data points should be labelled. A lower value implies better performance. It is calculated as the average similarity of each cluster [12]. It is mainly calculated using the following equation which is given below,

$$DB = \frac{1}{N}\sum_{i=1}^{N} D_i \tag{5}$$

Where,
D_i is the ith cluster's worst similarity score which is the largest across all other clusters, and finally, DB index is the averaged D_i across N clusters.

D. Cross-evaluation using Data Visualisation

In this step, we graphically represent the data which are basically clusters aquired from different clustering algorithms. Human intervention is required to re-evaluate the outcomes of the unsupervised module as humans are very good at recognizing visual patterns. In terms of customer segmentation visually understanding the patterns in customer data is as important as performing efficient clustering or segmentation. As the clustering or segmentation has to be meaningful and interpretable. An expert or user cross-validates the acquired clustering outcomes because only a human can make data driven decision and plan according to a business's requirement. When the clustering outcomes are satisfactory and interpretable, the expert or user approves the acquired results. The expert can also disapprove the results acquired using the evaluation metrics and select another suitable algorithm for a given dataset provided that the observation is accurate.

E. Selecting Algorithm

The selection of the clustering algorithm is done based on the cluster evaluation metric and also data visualisation. Data visualisation is used to cross-evaluate the outcomes of the clusters by a human expert or user. The clusters of each defined subspace will be evaluated using Silhouette Coefficient, Calinski-Harabasz Index and Davies Bouldin Index. All the applied clustering algorithms are compared based on these three evaluation metrics. The algorithm which satisfied all three metrics and surpasses the results of the rest of the algorithm is chosen as the suitable algorithm for the given dataset. Then, cross-evaluation is done based on data visualisation and the chosen algorithm is validated. The labels of each cluster of the selected algorithm is considered for the next step.

F. Conversion of Data

In this step the unlabelled customer data is transformed into labelled data. During clustering, each data point is assigned to a particular cluster and each cluster has a different label. We utilize this label data and add an extra column in the given dataset. The values of label denote which customer belongs to which cluster. Generally, the customers belonging to the same cluster exhibit the same characteristics. This adds an extra dimension in the dataset which adds an informative label to provide more context to the dataset. Hence, we can classify the customers now as a machine learning model can learn from the labelled data.

G. Classification

This step belongs to the supervised module of our proposed system. In this step, we utilize the labelled dataset acquired from the previous step. As the dataset now consists of extra meaningful information which is called label. Now supervised classification algorithms can be applied to this dataset. Now we train a machine learning model which will identify the clusters of new customers very easily. For classification, Multi-class Support Vector Machine model is used. The model is explained as follows:

1) Multi-class Support Vector Machine (SVM): Support Vector Machine (SVM) is one of the most popular supervised learning algorithms. It is applicable for classification as well as regression problems. But when there are more than two classes it becomes a multi-class classification problem. Hence, it is used in this proposed system. SVM generally supports binary classification and separates data points into two classes [13]. The same principle is utilized after dividing the multi classification problem into multiple binary classification problems to perform classification among multiple class [14]. SVM algorithm creates a decision boundary that can separate n-dimensional space into classes which can easily classify new data point in the correct category. In order to classify data points from m classes data set there are two approaches named One-to-Rest approach where the classifier can use m SVMs where each SVM would predict membership in one of the m classes. On the other hand, in One-to-One approach the classifier uses $\frac{m(m-1)}{2}$ SVMs.

H. Dataset

Two datasets have been used for this research.

- Mall Segmentation Dataset[1].
- Marketing Campaign Dataset[2].

TABLE I
DATASET DETAILS

Dataset Name	Attributes	Row	Used attributes of the Dataset
Mall Segmentation	5	200	Age, Annual Income, Spending Score
Marketing Campaign	29	2240	Age, Income, Total-Spend

Both datasets contain data about customers, their characteristics, and purchase behaviour which are the core of this research work. Among all the attributes few are selected to form the subspaces for further analysis.

IV. PERFORMANCE ANALYSIS

This section evaluates the performance of the proposed system. Cluster evaluation metrics are used to evaluate the acquired clusters in the clustering step. Then, we cross-evaluate the outcome using data visualization as the clusters need to be interpretable as well as efficient. After cross evaluation we finalise a suitable algorithm for a given dataset. Finally, we label the dataset using the acquired labels of the selected algorithm and train our classification model.

A. Performance Evaluation using Metrics

Evaluating the performance of clustering algorithms is not similar to the evaluation of supervised learning algorithms. Supervised algorithms are evaluated based on counting the number of errors or the precision and recall. Whereas, clusters are evaluated based on similarity or dissimilarity measures where ground truths are unknown. One of the most popular cluster evaluation metrics Silhouette Coefficient [10], Calinski-Harabasz Index [11] and Davies Bouldin Index [12] are considered in this research. The analysis and comparison based on both datasets are as follows:

- Mall Dataset: The following parameters are chosen to form the subspaces. Later on, clusters are formed and analysed based on formed subspaces.
  - x = Age, y = Annual Income, z = Spending Score

TABLE II
OUTCOME OF ALGORITHMS WITH CLUSTER EVALUATION METRICS

Algorithm	Parameters	Cluster	SI	CH Index	DB Index
K-Means	x,z	4	0.49974	332.563	0.68693
	y,z	5	0.55393	247.359	0.57256
DBSCAN	x,z	5	0.09769	41.627	3.33028
	y,z	5	0.29065	44.527	2.14070
Agglomerative	x,z	5	0.16156	126.837	3.50679
	y,z	5	0.54904	241.582	0.57658
Affinity	x,z	6	0.24480	181.424	3.28338
	y,z	6	0.31251	196.557	2.43585

- Market Campaign Dataset: The following parameters are used to form the subspaces and further analysis.
  - x = Age, y = Income, z = Total-Spend

TABLE III
OUTCOME OF ALGORITHMS WITH CLUSTER EVALUATION METRICS

Algorithm	Parameters	Cluster	SI	CH Index	DB Index
K-Means	x,z	5	0.02414	1201.593	23.89924
	y,z	5	0.53784	6426.918	0.45857
DBSCAN	x,z	6	-0.61949	2.27800	3.30819
	y,z	2	-0.61399	2.16400	2.29342
Agglomerative	x,z	5	0.17529	1310.304	16.57432
	y,z	5	0.48736	5087.492	0.46258
Affinity	x,z	0	-	-	-
	y,z	0	-	-	-

According to the values of the cluster evaluation metrics mentioned in III the performance orders are as follows:

- Mall Dataset:
  K-Means - Agglomerative - Affinity Propagation - DBSCAN
  Hence, K-Means performed best for this dataset. So, this dataset will be labelled according to the clusters acquired from K-Means. The labelling will be done separately for each subspace which is formed.
- Marketing Campaign Dataset:
  Agglomerative - K-Means - DBSCAN - Affinity Propagation
  Agglomerative has outperformed the other algorithms.

[1] https://www.kaggle.com/datasets/vjchoudhary7/customer-segmentation-tutorial-in-python

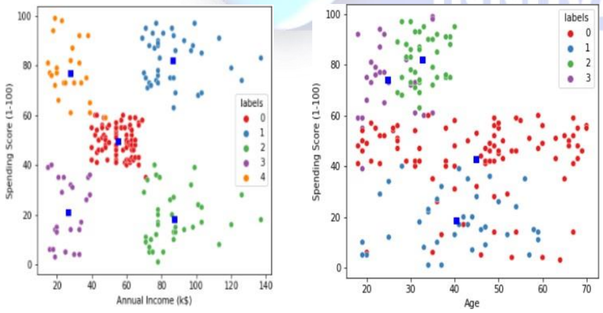
[2] https://www.kaggle.com/datasets/rodsaldanha/arketing-campaign

This dataset will be labelled according to the clusters acquired from the agglomerative algorithm. It will be done separately for each considered subspace.

B. Cross-evaluation of Mall Dataset

Data visualization is the practice of representing data in a meaningful and visual way so that viewers can interpret and understand easily. In this section, the outcomes of the clustering algorithms are visualised. The visualisation is shown using scatter plots of the found clusters. The interpretation of the outcomes are also discussed.

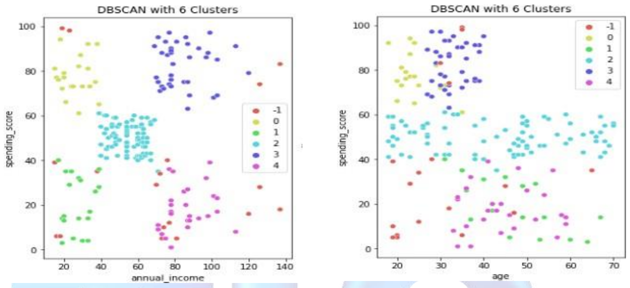
1) Visualisation of K-Means: In fig 2(a), five clusters were found in Annual Income vs Spending Score subspace and in fig 2(b) four clusters are found in Age vs Spending Score subspace by K-Means. From fig 2(a), it can be observed that the red cluster denotes the customers having moderate income have moderate spending scores. The orange cluster indicates the customers having lower annual income but a higher spending score. On contrary, the blue cluster indicates customers having higher annual income and higher the spending score. Moreover, the purple clusters indicates the customers having lower annual income and also lower spending score. Lastly, the green cluster indicates the customers having higher annual income but lower spending score. Hence, meaningful insights about customers are gained from the relationship between annual income and spending score. From fig 2(b), it can be observed that the green cluster denotes customers aged 30-40 years old who spend the most . Moreover, the purple cluster indicates that young customers spend more who are in the age range of 20-30 years old. But no specific relation is found in the red and blue cluster.



(a) (b)

Fig. 2. Clusters found using K-Means clustering (a)Annual Income vs Spending Score (b) Age vs Spending Score

2) Visualisation of DBSCAN: In fig 3(a), five clusters are found in Annual Income vs Spending Score subspace and in fig 3(b) five clusters are found in Age vs Spending Score subspace by DBSCAN. From fig 3(a), it can be observed that the olive cluster indicates the customers having lower annual income but a higher spending score. The green cluster shows the customers having lower annual income and spending score. Moreover, the blue cluster shows the customers having higher annual income and spending score. Whereas, the purple cluster specifies the customers having higher annual income but a lower spending score and the sky blue cluster denotes the customers having moderate income have moderate spending scores. From fig 3(b), it can be observed that the olive cluster indicates that young customers spend more who are in the age range of 20-30 years old. Whereas, the blue cluster indicates the customers who are 30-40 years old and spend more. But no specific relation is found in the blue and purple color cluster. In addition, DBSCAN could identify the outliers present in the red cluster. Hence, this gives us extra information because outliers can provide insightful information of data.

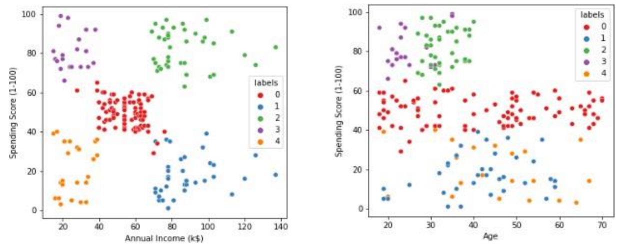


(a) (b)

Fig. 3. Clusters found using DBSCAN (a) Annual Income vs Spending Score (b) Age vs Spending Score

3) Visualisation of Agglomerative: In fig 4(a), five clusters are found in the Annual Income vs Spending Score subspace and in fig4(b) five clusters are found in the Age vs Spending Score subspace by Agglomerative Hierarchical method. From fig 4(a), it can be observed that the red cluster denotes customers having moderate income and moderate spending score. Whereas, the orange cluster indicates customers having lower annual income and lower spending score. The purple cluster indicates the customers having lower annual income but a higher spending score. Moreover, The green cluster indicates the customers having higher annual income and spending score. The blue cluster indicates the customers having higher annual income but a lower spending score. From fig 4(b), It is visible that there are no distinct groups in terms of customers' age and spending score using Agglomerative method.

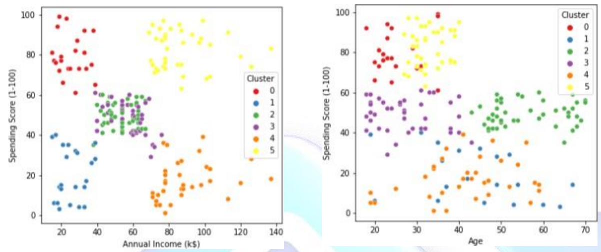
4) Visualisation of Affinity Propagation: In fig 5(a) and fig 5(b), six clusters are found in Annual Income vs Spending Score subspace and Age vs Spending Score subspace respectively by Affinity Propagation. From fig 5(a), it can be observed that most of the customers having moderate-income (40-60k) have moderate spending scores (40-60) and from fig 5(b), it is visible that young customers spend more who are in the age range of 20-30 years old. Moreover, for cluster 1,4 (Blue,Orange) there are no distinct groups in terms of customers' age and spending score. In addition this method



(a) (b)

Fig. 4. Clusters found using Agglomerative Hierarchical method (a) Annual Income vs Spending Score (b) Age vs Spending Score

found more clusters in both the spaces but the quality of the clusters in terms of cluster separation seems poor.

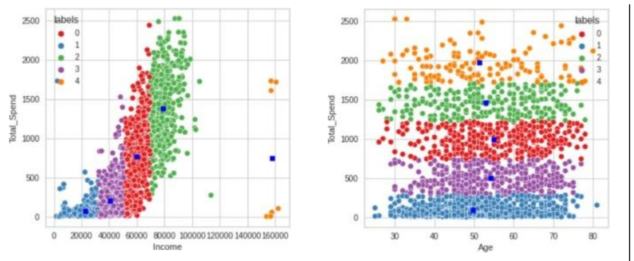


(a) (b)

Fig. 5. Clusters found using Affinity Propagation method (a) Annual Income vs Spending Score (b) Age vs Spending Score

C. Cross-evaluation of Marketing Campaign Dataset

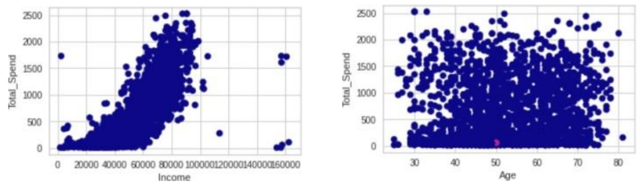
1) Visualisation of K-Means: In fig 6(a), five clusters were found in the Income vs Total-Spend subspace. Similarly, in fig 6(b) five clusters were also observed in the Age vs Total-Spend subspace by K-Means. This method found centralized clusters in both subspaces. It refers to poor separation between clusters and also overlapping of clusters.



(a) (b)

Fig. 6. Clusters found using K-Means clustering (a) Income vs Total-Spend (b) Age vs Total Spend

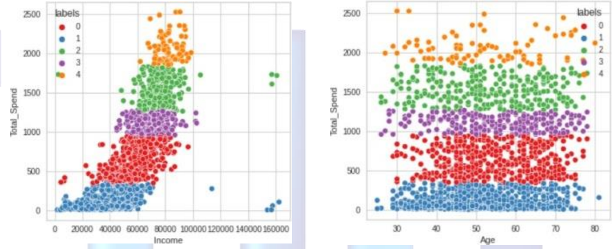
2) Visualisation of DBSCAN: In fig 7(a), two clusters were found in Income vs Total-Spend subspace, and in fig 7(b) six clusters were found in Age vs Total-Spend subspace by DBSCAN. But the six clusters are overlapping and it seems that the outliers are prevalent throughout all the clusters. In addition, DBSCAN has identified the outliers present in the data. Hence, this gives us extra information because outliers can provide insightful information on data.



(a) (b)

Fig. 7. Clusters found using DBSCAN (a) Income vs Total-Spend (b) Age vs Total-Spend

3) Visualisation of Agglomerative: In fig 8(a), five clusters were found in the Income vs Total-Spend subspace, and in fig 8(b) five clusters were found in the Age vs Spending Score subspace by Agglomerative method. As like the clusters of K-Means this is centralized but the performance is somewhat better than the K-Means according to the evaluation metrics.



(a) (b)

Fig. 8. Clusters found using Agglomerative Hierarchical method (a) Income vs Total-Spend (b) Age vs Total-Spend

4) Visualisation of Affinity Propagation: Affinity Propagation has given the worst performance in this market campaign dataset. Zero clusters are found in both subspaces Income vs Total-Spend and Age vs Total-Spend subspace.[Fig. 9]

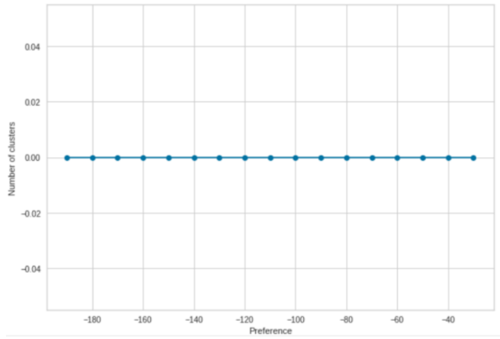


Fig. 9. Zero Clusters found using Agglomerative Hierarchical method in both subspaces of Income vs Total-Spend, and Age vs Total-Spend

D. Classification

To perform classification a Multi-class Support Vector Machine model is trained using the acquired labelled data.

As there are multiple clusters which resemble classes in this research. The selected subspaces are considered as feature variables and newly added attribute label is considered as target variable for classification. Different kernel functions such as linear function, polynomial function, radial basis function and sigmoid function are used for this purpose. A kernel function takes two data points x_n and x_m and calculates a distance measure of both data points. The distance measure is higher for closer datapoints and lower for data points which are further apart. This distance score helps to transform the data points to a higher-dimensional mapping. It reduces the computational cost and time which is useful for huge amounts of data. Moreover, It does not require more complex transformation. This method ensures a fast convergence and a good classification performance [15]. The datasets were split into 70% training data and 30% testing data. The testing data was later on used to predict the labels and also determine classification accuracy.

1) Classification Accuracy: The classification accuracy is measured based on the kernel functions for each subspace. We have considered two subspaces from each dataset.

- Mall Dataset
  The considered subspaces are (x, z) and (y, z).
  Where, x = Age, y = Annual Income and z = Spending Score.
  The dataset is labelled using the cluster labels acquired from K-Means. As K-Means outperformed the other algorithms according to the evaluation metrics and visualization. For (x, z) subspace, 4 clusters were found labelled as 0-3 and 5 clusters were found based on (y, z) subspace labelled as 0-4.
- Marketing Campaign Dataset
  The considered subspaces are (x, z) and (y, z).
  Where, x = Age, y = Income and z = Total Spend.
  The dataset is labelled using the cluster labels acquired from Agglomerative clustering algorithm. As it outperformed the other algorithms according to the evaluation metrics and visualization. For (x, z) subspace, 5 clusters were found labelled as 0-4 and similarly, 5 clusters were found based on (y, z) subspace labelled as 0-4.

TABLE IV
CLASSIFICATION ACCURACY

Kernel Function	Subspace	Mall Dataset	Marketing Campaign Dataset
Linear	x, z	1.000	0.998
	y, z	0.950	0.993
Polynomial	x, z	0.975	0.991
	y, z	0.975	0.993
Radial Basis	x, z	0.600	0.562
	y, z	0.300	0.429
Sigmoid	x, z	0.350	0.510
	y, z	0.375	0.154

From table IV, it is visible that the polynomial kernel function has achieved the best accuracy for Mall Dataset. It is 97.5% for both (x, z) and (y, z) subspace. On the other hand, for Marketing Campaign Dataset, linear kernel function has the best accuracy which is 99.8% for (x, z) subspace and 99.3% for (y, z) subspace. Hence, the trained model is able to classify new customer data properly.

V. CONCLUSION

The proposed system is a comprehensive system for businesses to classify and identify their customers. It is implemented in two modules: unsupervised module and supervised module. According to the unsupervised module, four clustering algorithms are applied to two different unlabelled customer datasets. We aquired two different suitable clustering algorithms for the two different customer datasets. K-means was selected for Mall Dataset as it outperformed the other clustering algorithms for this dataset. Similarly, Agglomerative was selected for Marketing Campaign Dataset. According to the proposed system, the datasets were labelled according to the acquired clusters from the selected suitable algorithm. As the datasets become labelled, Multi-class SVM is applied to classify customer data into groups. Moreover, utmost 97.5% accuracy is achieved for Mall Dataset and a maximum of 99.8% accuracy is acquired for Marketing Campaign Dataset using different kernel functions. This system can be fabricated according to the requirements of a business in order to make it more fruitful. It is a tangible and scalable system for customer segmentation, identification, and analysis. The effectiveness of the system will not drastically reduce with the increasing number of customers as it does not require clustering every time new customers are added to the system. It can help businesses make business decisions efficiently, create effective marketing strategies, address customers' needs, and most importantly create loyal customers. This system can also be useful to classify prospective customers which can be beneficial for a business. The future aspect of this work is to optimize the computation while implementing multiple clustering algorithms at once and to reduce the complexity. Moreover, It would have been more fruitful to use data belonging to an actual e-commerce platform but It could not be used as it is proprietary.

REFERENCES

[1] Camilleri, M. A. (2018). Market Segmentation, Targeting, and Positioning. In Travel Marketing, Tourism Economics and the Airline Product (Chapter 4, pp. 69-83). Springer, Cham, Switzerland.

[2] Punhani, R., Arora, V. P. S., Sabitha, S., and Kumar Shukla, V. (2021). Application of Clustering Algorithm for Effective Customer Segmentation in E-Commerce. 2021 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE). doi:10.1109/iccike51210.2021.9410713

[3] Nilashi, M., Samad, S., Minaei-Bidgoli, B., Ghabban, F. (2021). Online Reviews Analysis for Customer Segmentation through Dimensionality Reduction and Deep Learning Techniques. Arabian Journal for Science and Engineering, 46(9), 8697–8709. doi:10.1007/s13369-021-05638-z

[4] Hossain, A. S. M. S. (2017). Customer segmentation using centroid-based and density-based clustering algorithms. 2017 3rd International Conference on Electrical Information and Communication Technology (EICT). doi:10.1109/eict.2017.8275249

[5] Rezaeinia, S. M., and Rahmani, R. (2016). Recommender system based on customer segmentation (RSCS). Kybernetes, 45(6), 946–961. doi:10.1108/k-07-2014-0130

[6] 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

[7] Ferdosi, B,, Buddelmeijer, H., Trager, S., Wilkinson, M and Roerdink, J. (2010). Finding and visualizing relevant subspaces for clustering high- dimensional astronomical data using connected morphological operators. IEEE Symposium Density-basedlytics Science and Technology (VAST '10). 2010. 35-42. 10.1109/VAST.2010.5652450.

[8] Xiaoman, W., Kun, Y., Xingying, H., Jun, X., and Li, I. (2017, November). Analysis of power large user segmentation based on affinity propagation and K-means algorithm. In 2017 IEEE Conference on Energy Internet and Energy System Integration (EI2) (pp. 1-5). IEEE.

[9] Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

[10] Lin, Wilfred. (2021). Silhouette Coefficient for Clustering Tutorial.

[11] Calinski, T., and Harabasz, J. (1974). A dendrite method for cluster analysis. Communications in Statistics - Theory and Methods, 3(1), 1–27. https://doi.org/10.1080/03610927408827101

[12] Davies, D. L., and Bouldin, D. W. (1979). A Cluster Separation Mea- sure. IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI–1(2), 224–227. https://doi.org/10.1109/tpami.1979.4766909

[13] Franc, V., and Hlavac, V. (2002). Multi-class support vector machine. Object Recognition Supported by User Interaction for Service Robots. https://doi.org/10.1109/icpr.2002.1048282

[14] G. Lefait and T. Kechadi, "Customer Segmentation Architecture Based on Clustering Techniques," (2010).Fourth International Conference on Digital Society, 2010, pp. 243-248, doi: 10.1109/ICDS.2010.47.

[15] Nguyen, HN., Ohn, SY. (2006). Unified Kernel Function and Its Training Method for SVM. In: King, I., Wang, J., Chan, LW., Wang, D. (eds) Neural Information Processing. ICONIP 2006. Lecture Notes in Computer Science, vol 4232. Springer, Berlin, Heidelberg. https: //doi.org/10.1007/11893028 88