CUSTOMER DATA SEGMENTATION

INTRODUCTION:

Online shops are facing fierce competition as a result of the e-commerce sector's explosive expansion. Understanding and addressing customer satisfaction levels is critical for firms looking to acquire a competitive advantage. E-commerce businesses may make wise decisions to promote client loyalty and improve services by determining the elements that impact consumer happiness.

This project's objective is to analyse customer data and ascertain the degree of happiness among e-commerce consumers by applying unsupervised learning techniques, including DBSCAN, hierarchical clustering, and k-means clustering. It's finds hidden patterns in the data and group clients based on similarities by using these clustering techniques. Because of its user-friendly interface, which makes it simpler to visualise and explain the findings, k-means clustering is particularly pleasant to use.

RELATED WORK:

Customer data segmentation for satisfaction in the e-commerce space has been studied in the past using a variety of techniques and approaches. Previous research has investigated the use of machine learning methods to predict customer happiness, including decision trees and neural networks. These supervised learning techniques, however, depend on labelled data, which isn't always accessible or easily used in practical situations. Consequently, unsupervised learning methods like clustering algorithms have become more and more well-liked for objectively assessing consumer happiness.

In the realm of customer satisfaction analysis, DBSCAN, hierarchical clustering, and k-means clustering have drawn a lot of interest lately. The goal of the density-based clustering method DBSCAN is to locate dense areas in the data and categorise outliers differently. In contrast, either agglomerative or divisive methods can be used in hierarchical clustering to produce a hierarchy of clusters. Finally, by minimising the sum of squares inside each cluster, k-means clustering divides the data into k unique clusters. Through an examination and comparison of different clustering approaches, then which method is most suited for our particular e-commerce dataset is decided.

METHODOLOGY:

This section will detail the methodical process used to apply unsupervised learning techniques to the topic of e-commerce consumer happiness. Preprocessing the client data in the first place entailed resolving missing values and eliminating any unnecessary or redundant properties. After that, label encoding was used to make sure that every variable/ features data was on the same/ machine understandable format, allowing for equitable understanding across various characteristics.

The pre-processed data was then subjected to the three unsupervised learning techniques: k-means clustering, hierarchical clustering, and DBSCAN. To maximise performance, each algorithm's hyperparameters were carefully chosen. The quality and efficiency of the clustering were assessed by utilising suitable evaluation measures, such as output accuracy performance and silhouette scores, to analyse the outcomes produced by each approach.

For every clustering algorithm, classification reports were made to help with interpretation and comprehension of the findings. A more thorough examination of consumer behaviour and satisfaction levels was made possible by these report assistance in spotting clusters and outliers.

RESULTS AND DISCUSSION:

1) Data Collection:

In order to train the E-commerce Customer Behaviour model on satisfaction level, pertinent data must be gathered during the data collecting phase. Customer demographics, past purchases, satisfaction scores, and other pertinent data may be included in this. The data gathered from the Kaggle website.

2)Dataset:

The model is to be trained on 350 distinct data entries with 11 columns: Customer ID, Gender, Age, City, Membership Type, Total Spend, Items Purchased, Average Rating, Discount Applied, Days Since Last Purchase, and Satisfaction Level. In order to manage missing values, encode categorical variables, and get the dataset ready for training, it is preprocessed and cleaned.

3) Data Preparation:

The gathered dataset is processed and made ready for model training during the data preparation stage. Managing missing data, utilising LabelEncoder to encode categorical variables, StandardScaler to scale numerical features, and train_test_split to divide the data into training and testing sets are some of the steps.

4) Model Selection:

Three clustering techniques are used for this project:

Applications with Noise: A Density-Based Spatial Clustering Approach (DBSCAN)
Agglomerative clustering, or hierarchical clustering
K-Means clustering

After applying each algorithm to the training set, measures including the silhouette score, accuracy, precision, recall, and F1 score are used to assess how well the algorithm performed. The algorithm that performs the best is chosen for more examination.

The findings of the use of the DBSCAN, hierarchical clustering, and k-means clustering approaches are presented in depth in this section. Different clusters were successfully found using DBSCAN, which showed clear patterns in the customer data. Outliers were successfully identified by the algorithm, enabling the identification of unhappy clients who might need more assistance and attention. A hierarchical structure of clusters was produced using hierarchical clustering, emphasising the connections and commonalities among various client groups. In contrast, k-means clustering divides the data into a preset number of groups, making it possible to clearly separate and classify clients according to their satisfaction levels.

When the three clustering methods were compared, it became clear that k-means clustering outperformed the other two in terms of accuracy and result clarity. The intuitive k-means clustering interface makes it simpler to visualise the consumer segments and to comprehend the resulting clusters. The choice of algorithm should be carefully studied based on the unique requirements of the e-commerce organisation, since each clustering approach has advantages and limits of its own, by comparing the performance and testing in UI then finalised with k-means clustering technique.

CONCLUSION AND FUTURE WORK:

In summary, this study effectively used k-means clustering, hierarchical clustering, and DBSCAN unsupervised learning techniques to analyse and assess e-commerce consumers' satisfaction levels. Customer data was clustered to reveal hidden patterns and structures that offered important insights into the behaviour.

Because of its simplicity and effectiveness, k-means clustering, and hierarchical clustering is a good method for analysing consumer happiness in e-commerce, according to the findings of the comparison of the three clustering algorithms. Still, more investigation is needed to determine the possibilities of other unsupervised learning strategies and how well they work with various e-commerce datasets. Furthermore, investigating ensemble approaches and implementing more sophisticated data pretreatment procedures may improve the clustering analysis's resilience and accuracy.

All things considered; this study advances the field of e-commerce customer satisfaction analysis by offering a thorough grasp of the use of unsupervised learning methodologies. The information gathered may help companies make data-driven decisions that will increase consumer happiness, improve services, and provide them a competitive edge in the e-commerce sector.

References:

[1]https://www.kaggle.com/datasets/uom190346a/e-commerce-customer-behavior-dataset?resource=download

[2] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html

[3] https://www.datacamp.com/tutorial/introduction-hierarchical-clustering-python

[4] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html