Project Design Phase-II
Data Flow Diagram

Date	23 October 2023
Team ID	Team-592709
Project Name	Wholesale Customer Segmentation Analysis Using ML
Maximum Marks	4 Marks

Data Flow:

Data Collection: This phase entails gathering a diverse dataset of products that are bought by customers from sources such as customer surveys and interviews, social media monitoring, or other means. The collected data are then stored in a raw data repository.

Data Cleaning: Clean the data by removing duplicates, missing values, and outliers. This step ensures that the data is accurate and reliable.

Data Transformation: Transform the data into a suitable format for clustering analysis. This step involves converting the data into a numerical format that can be used for clustering.

Feature Selection: Select the most relevant features that can help in clustering. This step involves identifying the features that are most important for clustering and discarding the rest.

Clustering Algorithm Selection: Choose an appropriate clustering algorithm based on the type of data and business requirements. There are several clustering algorithms available, such as K-means, Hierarchical Clustering, and DBSCAN.

Model Training: Train the clustering model on the selected features. This step involves running the selected clustering algorithm on the data to create clusters.

Model Evaluation: Evaluate the model's performance using metrics such as silhouette score, elbow method, and gap statistic. This step helps in determining the optimal number of clusters and assessing the quality of the clusters.

Cluster Interpretation: Interpret the clusters by analyzing the characteristics of each cluster. This step involves identifying the common characteristics of each cluster and understanding what makes them unique.

Segmentation Strategy Development: Develop a segmentation strategy based on the insights gained from cluster interpretation. This step involves using the insights gained from cluster interpretation to develop a segmentation strategy that can be used to target specific customer groups.

Flow in the model:

Data is sourced from the Data Repository, and it undergoes a transformation in the Data Preprocessing phase to become pre-processed data.

The pre-processed data flows to the Model Training process, where the model is trained using this data. The trained model can be used for classification.

In the Model Evaluation process, a distinct dataset is employed to gauge the model's accuracy and performance.

Once the model is ready, it can be deployed from the Trained Model Repository, making it accessible to users.

The data flow begins with the collection of data from a variety of sources. The data is then cleaned and transformed into a format that is suitable for analysis. The data is then reduced to make it more manageable and to improve the performance of the data analysis algorithms. The preprocessed data is then analyzed to identify patterns and trends and to develop insights into customer needs and preferences. The insights gained from the data analysis are then used to inform the product development process.

Here is an example of how the data flow might work in practice for a wholesale business that sells clothing:

The wholesale business collects data from a variety of sources, such as customer surveys, sales data, and social media.The data is cleaned to identify and correct errors, such as missing values, duplicate entries, and inconsistencies in formatting.

The data is transformed into a format that is suitable for analysis, such as converting the data to numerical values and creating new variables such as customer satisfaction.
The data is reduced to make it more manageable and to improve the performance of the data analysis algorithms. For example, the wholesale business may remove customers who have only purchased a small number of items, or they may sample the data to create a smaller dataset.

The preprocessed data is analyzed to identify patterns and trends, and to develop insights into customer needs and preferences. For example, the wholesale business may identify the most popular clothing items, the colors and styles that customers prefer, and the features that customers value most.

The insights gained from the data analysis are then used to inform the product development process. For example, the wholesale business may decide to develop a new line of clothing in a popular color or style, or they may add a new feature to an existing product.

By having a well-defined data flow, wholesale businesses can ensure that their data is processed correctly and that the insights gained from the data analysis are accurate. This information can then be used to develop products that are more likely to be successful in the marketplace.

Data Flow Diagram: