# Machine Learning - CPS803

# Assignment 4 Report

Rohithraj Baskaran
501217980

## 1. Background

Data points like order frequency, time of purchase, previous order intervals, and product categories are captured in the dataset used for this analysis, which offers insightful information about E-Commerce consumer behaviour. This information, which was taken from an open-source dataset, shows the patterns of internet buyers over time. Since it directly influences user experience improvements, inventory management, and customized marketing strategies, consumer behaviour analysis in e-commerce is of great importance.

Order frequency (order_number), order day of the week (order_dow), purchase time (order_hour_of_day), and the most popular departments and goods purchased are among the quantitative and categorical elements included in the dataset. Understanding user clustering patterns will enable the division of clients into relevant categories, which is the aim of this study. Businesses can use this segmentation to target particular user groups with customized marketing.

Knowing the clustering results in this context might help firms respond to inquiries such as: Which clientele groups are most likely to place repeat orders? Which buying habits such as frequency or time are most prevalent? What differences exist in product preferences between clusters? K-Means and DBSCAN clustering algorithms are used in this analysis to find these patterns in customer behaviour.

## 2. Methods

Data preprocessing, feature engineering, addressing missing values, clustering techniques, and visualization/evaluation are the five main steps in the methodology. Here is a thorough analysis:

1. **Data preprocessing**: To ensure consistency, all numerical features were transformed to numeric types, with errors forced into NaN. Using LabelEncoder, categorical data like department, reordered, and department_id were transformed into numerical representations.
2. **Feature Engineering**: Data was grouped by user_id to produce aggregated user-level features. For every user, metrics like the most frequent product category, mean order time, and maximum order count were calculated. To guarantee consistent scale, StandardScaler was used to standardize features for both methods.
3. **Handling Missing Values**: Using the SimpleImputer technique, missing values in numeric columns were imputed using each column's median value. When it comes to skewed data distributions, median imputation is resilient.
4. **Clustering Algorithms**:
    a. **K-Means Clustering:** The dataset was divided into four clusters, each with a randomly initialized centroids. Given the nature of consumer segmentation, the selection of four groupings was heuristic.

   b. **DBSCAN (Density-Based Spatial Clustering)**: Clusters were found using data density. The cluster analysis does not include noise points (designated as -1).

5. **Visualization and Evaluation**: The high-dimensional feature space was reduced to two dimensions for display using Principal Component Analysis (PCA). Plotting of the results revealed cluster distributions for DBSCAN and K-Means.Metrics such as the silhouette score for K-Means were used to evaluate the quality of the clustering.

## 3. Results

The results for the DBSCAN clustering was a bit disappointing as it was unclear and not well separated. Visualization using PCA shows the separation of clusters, confirming the presence of distinct customer segments.  DBSCAN identified some customers who don't fit into any of the other segments, possibly representing outliers or users with inconsistent purchase patterns. DBSCAN emphasized noise and crowded areas, whereas K-Means clusters were well-separated.

K-Means clustering revealed 4 distinct customer segments:

- Cluster 0: This cluster has a moderate reorder rate, frequent orders, and a comparatively low average order number. This implies that users in this market may be brand-new to the site or make regular, modest transactions.
- Cluster 1: Less frequent orders, a larger average order number, and a greater reorder rate are characteristics of this cluster. These clients could be devoted and rarely buy bigger things.
- Cluster 2: This cluster has a modest reorder rate, extremely few orders, and the highest average order number. Consumers in this market group might occasionally place large orders or buy in bulk.
- Cluster 3: This cluster has the lowest reorder rate, little orders, and a low average order number. These clients may be infrequent purchasers or product experimenters.

DBSCAN clustering revealed 3 main clusters and some noise points (outliers):

- Cluster 0: This group represents new or recurring small-order buyers and shares characteristics with K-Means Cluster 0.
- Cluster 1: This group is similar to K-Means Cluster 1 and probably consists of devoted clients who place larger, infrequent orders.
- Cluster 2: This group, which includes those who buy infrequently or frequently, is comparable to K-Means Cluster 2.

## 4. Conclusions

To sum up, the clustering analysis provided insightful information about e-commerce customer behaviour. Potential target groups for specialized marketing tactics were highlighted by K-Means' successful identification of four unique consumer categories with varying order frequency and sizes. Cluster 1 concentrated on devoted, larger orders, while Cluster 0 reflected frequent, smaller transactions. Cluster 3 represented infrequent buyers, while Cluster 2 was linked to infrequent mass purchases. Despite identifying three primary clusters, DBSCAN's findings were less definitive because some noise points pointed to outliers. All things considered, these results provide useful information that companies can use to improve marketing tactics based on consumer buying trends and target customers more effectively.

References:

[1] Hunter “Supermarket dataset for predictive marketing 2023” Kaggle, 2023 https://www.kaggle.com/datasets/hunter0007/ecommerce-dataset-for-predictive-marketing-2023