InternsElite Mini Project-

Product Clustering Using K-Means

1. Introduction:

Understanding purchasing patterns is crucial for effective product management. In this project, we applied K-Means clustering to the **Online Retail II UCI Dataset** to segment products based on their sales behaviour.

K-Means is an unsupervised learning algorithm that groups similar items together, which helps businesses tailor their marketing strategies and manage inventory more efficiently.

Why K-Means Clustering?

K-Means clustering identifies groups of products that share similar purchasing patterns. By clustering products, businesses can make informed decisions about product placements, promotions, and stock management.

2. Problem Statement:

The objective of this project was to use K-Means clustering to categorize products from the **Online Retail II UCI Dataset**.

The goal was to group products with similar sales behaviours to enhance marketing and inventory strategies.

3. Methodology:

Data Analysis

1. Loading Data: We started by loading the dataset, which includes transaction details such as product codes and quantities.
2. Cleaning Data: We removed duplicates and handled missing values, focusing on valid transactions.
3. Feature Engineering: We calculated total quantity purchased and total revenue for each product to use as features for clustering.

K-Means Clustering

1. **Selecting Features**: We used total quantity and total revenue to represent each product.
2. **Scaling Features**: Standardized features to ensure fair contribution to the clustering process.
3. **Building the Model**: Applied K-Means clustering to create groups of similar products.
4. **Evaluating Clusters**: Assessed the clusters using the silhouette score to ensure meaningful grouping.

4. Results and Discussions:

Optimal Number of Clusters

By using the Elbow Method, we identified 5 as the optimal number of clusters. This was determined by observing where the inertia reduction rate slowed down.

Cluster Insights

- **Cluster 0: High revenue and moderate quantity.**
- **Cluster 1: Low revenue but high quantity.**
- **Cluster 2: Moderate values in both revenue and quantity.**
- **Cluster 3: Very high revenue with low quantity.**
- **Cluster 4: Low revenue and quantity.**

Evaluation

A silhouette score of 0.78 indicates that the clusters are well-separated and cohesive. The K-Means algorithm effectively grouped products into distinct clusters.

5. Conclusions:

K-Means clustering successfully identified product segments based on purchasing behaviour. These clusters can guide targeted marketing and inventory strategies, enhancing business decision-making.

6. References:

1. **Online Retail II UCI Dataset**: Kaggle
2. **Scikit-learn Documentation**: scikit-learn.org
3. **Introduction to Data Mining**: Tan, P.-N., Steinbach, M., & Kumar, V. (2005). Addison-Wesley.