DS203: Assignment 5

Deadline: 20th March, 11:55 PM

Datasets:

- Credit Card Users
- UCI Wholesale

1 Core Tasks

Complete these steps for both datasets:

1.1 Data Preparation

1. Complete the following:
   - Select the top 5 features with highest variance (after standardization)
   - Apply StandardScaler to normalize all features
   - Create a correlation heatmap of the selected features
   - Report the percentage of missing values and how you handled them

1.2 Clustering Implementation

2. Use K-Means clustering:
   - Use the elbow method to find the optimal K (test K from 2 to 10)
   - Report the silhouette score for the optimal K
   - Create a scatter plot showing the clusters (use first two PCA components)
3. Use Hierarchical clustering:
   - Use Ward linkage method
   - Cut the dendrogram to produce the same number of clusters as determined in K-Means
   - Report the silhouette score for this clustering
4. Use DBSCAN:
   - Use eps=0.5 and min_samples=5
   - Report the number of clusters found and percentage of outliers
   - Report the silhouette score (excluding outliers)

2 Comparative Analysis

Answer the following specific questions (max 3 sentences each):

1. Which algorithm produced the highest silhouette score for your dataset?
2. Which method's cluster count seems most appropriate?
3. What percentage of data points were identified as outliers by DBSCAN? What do these outliers represent in your dataset?
4. For the best-performing algorithm, calculate and report the mean values of each feature for each cluster.

3 Algorithm Comparison

Complete this table in your report:

Metric	**K-Means**	**Hierarchical**	**DBSCAN**
Number of clusters			
Silhouette score			
Execution time (seconds)			
Handles outliers (Yes/No)			

Submission Requirements

Your submissions should include the following:

a. **A PDF document** with all the above analyses and comments. Ensure that you include the required figures and tables (i.e., metrics data) in your report, along with the explanations and analysis.

b. **Your Python source file** (`.py` file). Please **DO NOT** upload Jupyter Notebooks – they get bulky! All important information (tables, plots, etc.) should be presented in the report.

c. **File Naming:**

   - The name of the PDF should be: `E5-your-roll-number.pdf`
   - The name of the Python source file should be: `E5-your-roll-number.py`

d. Upload the PDF and the source file to the assignment submission point **E5**.