Phase-2 Submission – Data Analytics Project

Institution: Government College of Engineering, Dharmapuri

Department: Electronics and Communication Engineering (ECE)

GitHub Repository Link: https://github.com/kaviya-0809/Customer-purchasing-.git

GitHub Repository Link: [Insert your GitHub repo link here]

# Student Names and Register Numbers:

- Vaishanavi M

- Kaviya V (613523106012)

- Bavana A (613523106003)

- Risha (613523106032)

# 1. Problem Statement

Retailers often struggle to identify customer purchasing patterns from large volumes of transactional data. This project mines association rules from transaction logs to uncover hidden relationships between products, helping in better store arrangement, product bundling, and targeted promotions. It focuses on market basket analysis to drive business strategy improvements.

# 2. Project Objectives

- Identify frequent itemsets in transactions.

- Generate strong association rules using the Apriori algorithm.

- Derive insights for bundling and product placement.

- Create visualizations for pattern recognition.

- Original objectives remain consistent post initial data understanding.

# 3. Project Workflow (Flowchart)

Data Collection → Data Cleaning → Data Transformation → EDA → Frequent Itemset Generation → Association Rule Mining → Visualization → Insights & Documentation

# 4. Data Description

- Dataset Source: Online Retail Dataset (Kaggle)

- Type: Public, Static

- Rows & Columns: ~500,000+ transactions, 8 columns

- Attributes: Invoice No, Stock Code, Description, Quantity, Unit Price, Invoice Date, Customer ID, Country

- Format: Structured

- Preprocessing Required: Yes (missing values, duplicates, formatting)

# 5. Data Preprocessing

- Removed missing CustomerID entries and blank product names.

- Dropped duplicate entries.

- Filtered only valid quantity and price values (> 0).

- Transformed into a transactional matrix (one-hot encoded).

- No outliers as primary focus is categorical association.

# 6. Exploratory Data Analysis (EDA)

- Identified top-selling products using bar plots.

- Visualized co-occurring products with heatmaps.

- Frequency distribution of most bought items.

Key Insights:

- Certain product pairs are frequently bought together.

- Lift and confidence values support bundling strategy.

- Network graphs show product clusters (cross-selling opportunities).

# 7. Tools & Technologies

- Language: Python

- IDE: Google Colab / Jupyter Notebook

- Libraries: pandas, numpy, seaborn, matplotlib, mlxtend, networkx

- Optional Tools: plotly, Streamlit (for UI if required)

# 8. Team Members and Contributions

|  |  |
| --- | --- |
| Team Member | Role and Responsibility |
| Vaishanavi M | Team Leader & Data Collection Lead: Dataset sourcing, team coordination, milestone tracking |
| Kaviya V | Data Cleaning & Preprocessing: Removing noise, restructuring data for mining |
| Bavana A | EDA & Visualization: Graphs, frequency analysis, heatmaps, and plots |
| Risha | Modeling & Rule Mining: Implemented Apriori algorithm, interpreted results, derived rules |