Market Basket Analysis

Sanjana V. Choudhari (3BR23CD086)  
Shirisha T.S. (3BR23CD088)  
Guide: Mr. Bharath

# What is Market Basket Analysis?

Market Basket Analysis (MBA) is a data mining technique used by retailers to understand the purchase behavior of customers by discovering associations between different items that customers place in their shopping baskets.

# Key Concepts

• Itemset: A collection of one or more items.  
• Support: Frequency of an itemset appearing in the dataset.  
 - support(A) = Transactions containing A / Total transactions.  
• Confidence: Likelihood that item B is bought when item A is bought.  
 - confidence(A → B) = Support(A ∪ B) / Support(A)  
• Lift: Measures how much more likely B is bought when A is bought compared to B being bought independently.  
 - lift(A → B) = Confidence(A → B) / Support(B)

# Applications

• Retail: Product placement, promotions, inventory management.  
• E-commerce: Recommendation systems (e.g., “Customers who bought this also bought…”).  
• Banking: Cross-selling financial products.

# Popular Algorithms

## 1. Apriori Algorithm

• Key Idea: Uses a bottom-up approach (starts from single items, expands to larger sets).  
• Generates candidate itemsets, then filters them using support threshold.  
• Pros: Simple and easy to implement. Useful for small to medium datasets.  
• Cons: Generates many candidates → slow on large datasets. Needs multiple database scans.

## 2. FP-Growth (Frequent Pattern Growth)

• Key Idea: Builds a compact tree structure (FP-tree) instead of generating candidates.  
• Uses recursive pattern mining from the FP-tree.  
• Pros: Faster than Apriori for large datasets. Needs only two database scans.  
• Cons: Tree structure is complex to implement manually.

## 3. ECLAT (Equivalence Class Transformation)

• Key Idea: Uses vertical data format (item → transaction ID list).  
• Finds frequent itemsets using intersection of transaction IDs.  
• Pros: Efficient for datasets with many frequent itemsets. Performs well when transactions are short but numerous.  
• Cons: Uses more memory due to TID-lists. Less popular, limited library support.

# Source Code and Implementation

• Libraries used: pandas, numpy, matplotlib, seaborn, mlxtend, networkx, warnings.  
• Dataset: A list of transactions where each list represents a customer's basket.  
• Transaction Encoding: Binary matrix of transactions (items present = True).  
• Frequent Itemset Mining: FP-Growth used to find itemsets with ≥ 60% support.  
• Association Rule Generation: Extracts rules with Support, Confidence, Lift.  
• Post-processing: Converts frozensets to strings, handles NaNs.

# Visualizations

1. Top 10 Frequent Itemsets (Bar Chart): Shows commonly bought item combinations.  
2. Support vs Confidence (Scatter Plot): Plots rule strength with Lift as color/size.  
3. Heatmap of Lift: Matrix showing strength of associations between item pairs.  
4. Network Graph of Association Rules: Graph with itemsets as nodes and rules as arrows.

# Conclusion

Visualizations and rule metrics help convert raw transaction data into actionable insights for cross-selling, shelf organization, and personalized product recommendations.