Lab Number: 8

Title
Generate association rules using FPGrowth for a suitable primary dataset.

Objective
To apply the FP-Growth algorithm on a suitable primary dataset in WEKA to generate frequent itemsets and association rules, using predefined support and confidence thresholds, and to analyze the resulting rule set.

IDE/Tools Used
Weka 3.8.6

Theory
Primary Dataset: A primary dataset represents raw, original data collected directly from transactional or operational systems. For association rule mining, the dataset typically consists of multiple transactions where each transaction contains a set of items. These datasets are especially suitable for market basket analysis and co-occurrence pattern discovery.

Association Rule Mining: Association rule mining identifies patterns, correlations, and co-occurrences among items within large datasets. It discovers rules of the form:

$$A \Rightarrow B$$

Where:
A = antecedent (items on the left side)
B = consequent (items predicted to occur with A)

Each rule is evaluated using:

- **Support**

Support measures how frequently an itemset appears in the dataset. It tells you how common or popular the itemset is across all transactions and calculated by formula:

$$\text{Support } (A \Rightarrow B) = \frac{Number\ of\ transactions\ containing\ A \cup B}{Total\ Transaction}$$

- **Confidence**

Confidence measures how often the rule is true, meaning how likely the consequent is to occur when the antecedent occurs. It reflects the reliability of an association rule and us calculate using the formula:

$$\text{Confidence } (A \Rightarrow B) = \frac{Support\ (A \cup B)}{Support(A)}$$

- **Lift**

Lift measures how much more likely the consequent is to occur with the antecedent compared to randomly. Lift > 1 indicates a positive association, meaning the items occur together more often than expected by chance. It is calculated using the formula:

$$\text{Lift } (A \Rightarrow B) = \frac{Confidence\ (A \Rightarrow B)}{Support(B)}$$

FP-Growth Algorithm: FP-Growth (Frequent Pattern Growth) is an efficient algorithm for mining frequent itemsets without generating candidate itemsets explicitly, unlike Apriori. Key characteristics of FP-Growth algorithm includes:

- Uses a compressed data structure called an FP-Tree
- Avoids repeated database scans (only two passes needed)
- Very efficient for large datasets
- Extracts frequent itemsets using a recursive divide-and-conquer pattern
- Generates association rules from frequent itemsets meeting minimum support and confidence

FP-Growth is generally faster and more scalable than Apriori because it eliminates the need to generate and test a large number of candidate itemsets.

Implementation

The following steps were performed to implement the FP-Growth algorithm in WEKA.

1. Conversion of dataset from csv to arff format

Firstly, open the arff viewer from tools and then open the csv file in the viewer. Visualize the dataset and save it as arff format.

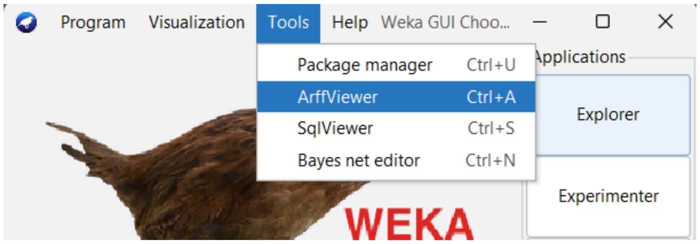
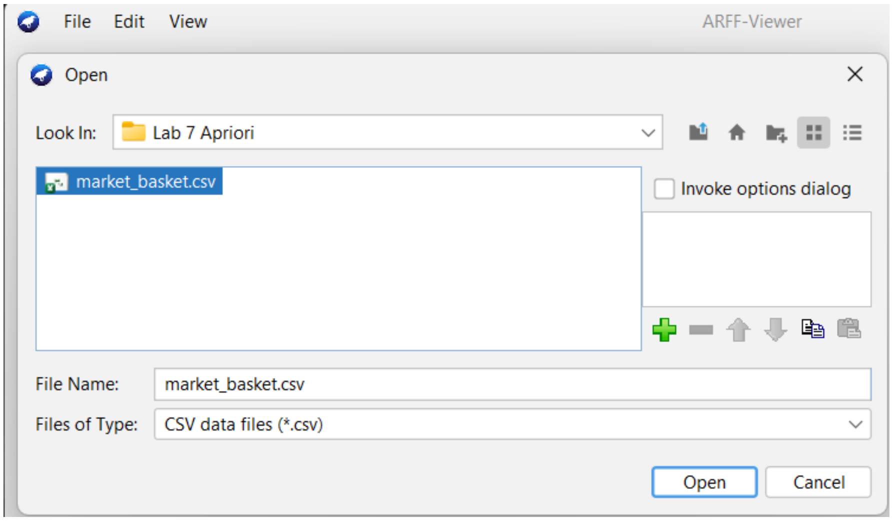


Figure 1: Opening the Arff Viewer



Look In: Lab 7 Apriori

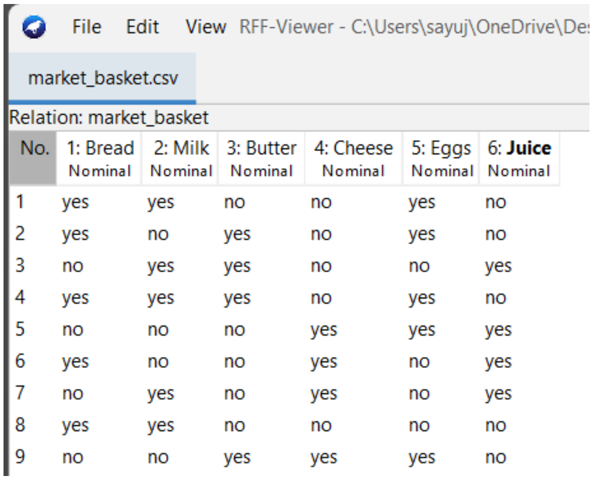
market_basket.csv

□ Invoke options dialog

File Name: market_basket.csv

Files of Type: CSV data files (*.csv)

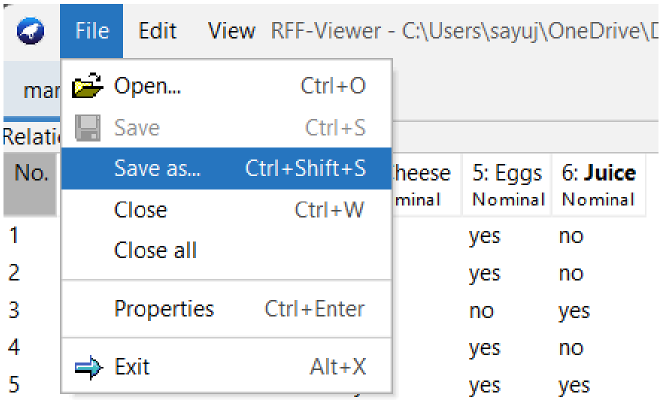
Figure 2: Selecting the csv file



Relation: market_basket

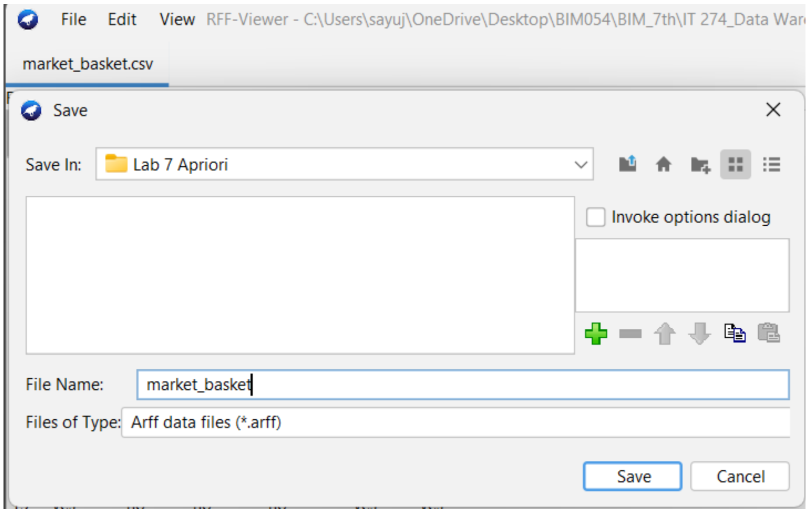
No.	1: Bread Nominal	2: Milk Nominal	3: Butter Nominal	4: Cheese Nominal	5: Eggs Nominal	6: Juice Nominal
1	yes	yes	no	no	yes	no
2	yes	no	yes	no	yes	no
3	no	yes	yes	no	no	yes
4	yes	yes	yes	no	yes	no
5	no	no	no	yes	yes	yes
6	yes	no	no	yes	no	yes
7	no	yes	no	yes	no	yes
8	yes	yes	no	no	no	no
9	no	no	yes	yes	yes	no

Figure 3: Visualization of the dataset



No.	heese minal	5: Eggs Nominal	6: **Juice** Nominal
1		yes	no
2		yes	no
3		no	yes
4		yes	no
5		yes	yes

Figure 4: Option to save as arff


File
Edit
View
RFF-Viewer - C:\Users\sayuj\OneDrive\Desktop\BIM054\BIM_7th\IT 274_Data War
market_basket.csv
Save
□ Invoke options dialog
Save
Cancel

Save In: Lab 7 Apriori

File Name: market_basket

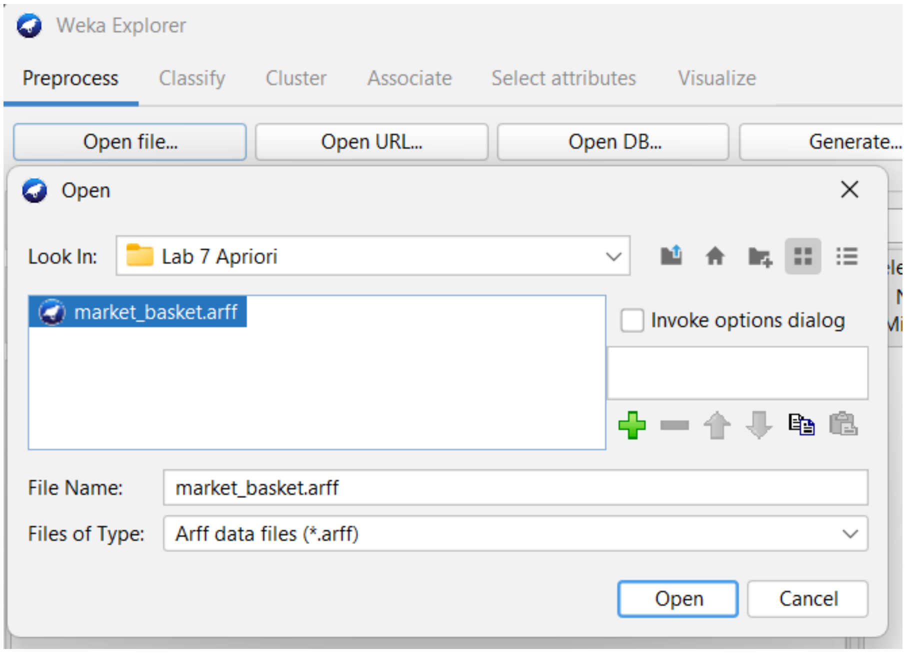
Files of Type: Arff data files (*.arff)

Figure 5: Saving as arff format

2. Loading the Dataset

A market basket style primary dataset was used, containing information about items purchased together in individual transactions. The dataset includes six binary (yes/no) attributes: Bread, Milk, Butter, Cheese, Eggs, and Juice. Each row represents one shopping transaction, indicating whether a particular item was purchased in that transaction.

The dataset consists of 50 transactions, making it suitable for association rule mining algorithms such as Apriori and FP-Growth, both of which require nominal or binary attributes. After preparing the dataset in arff format, it was loaded into WEKA and inspected through the Preprocess tab. The attributes were confirmed to be nominal and transaction-based, meaning no additional preprocessing or removal of columns (such as an ID field) was required.



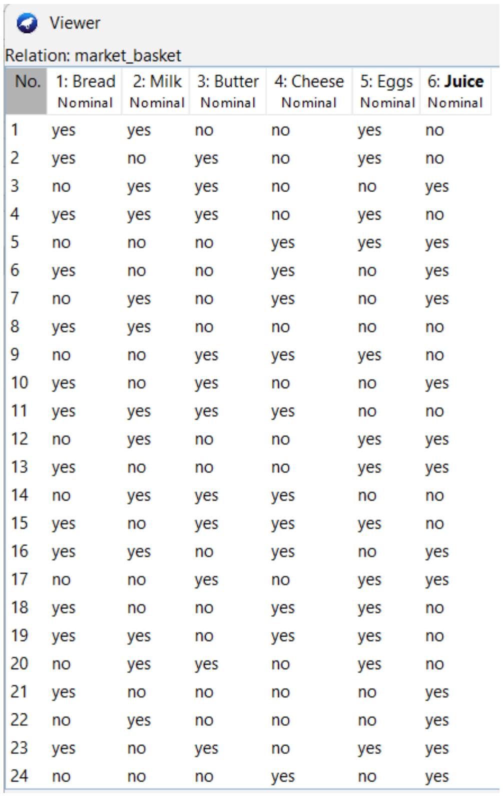
Look In: Lab 7 Apriori

market_basket.arff

File Name: market_basket.arff

Files of Type: Arff data files (*.arff)

Figure 6: Opening the dataset



Relation: market_basket

No.	1: Bread	2: Milk	3: Butter	4: Cheese	5: Eggs	6: Juice
	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal
1	yes	yes	no	no	yes	no
2	yes	no	yes	no	yes	no
3	no	yes	yes	no	no	yes
4	yes	yes	yes	no	yes	no
5	no	no	no	yes	yes	yes
6	yes	no	no	yes	no	yes
7	no	yes	no	yes	no	yes
8	yes	yes	no	no	no	no
9	no	no	yes	yes	yes	no
10	yes	no	yes	no	no	yes
11	yes	yes	yes	yes	no	no
12	no	yes	no	no	yes	yes
13	yes	no	no	no	yes	yes
14	no	yes	yes	yes	no	no
15	yes	no	yes	yes	yes	no
16	yes	yes	no	yes	no	yes
17	no	no	yes	no	yes	yes
18	yes	no	no	yes	yes	no
19	yes	yes	no	yes	yes	no
20	no	yes	yes	no	yes	no
21	yes	no	no	no	no	yes
22	no	yes	no	no	no	yes
23	yes	no	yes	no	yes	yes
24	no	no	no	yes	no	yes

Figure 7: Visualizing the dataset

3. Running the FP-Growth Algorithm

Go to the Associate tab and choose FP-Growth from the list of association rule learners and set the following parameters:

- lowerBoundMinSupport: 0.2
- mertricType: Confidence
- minMetric: 0.5, and other parameters can remain at default

Now click on start and let the algorithm run. The output will be displayed under Associator Output section.

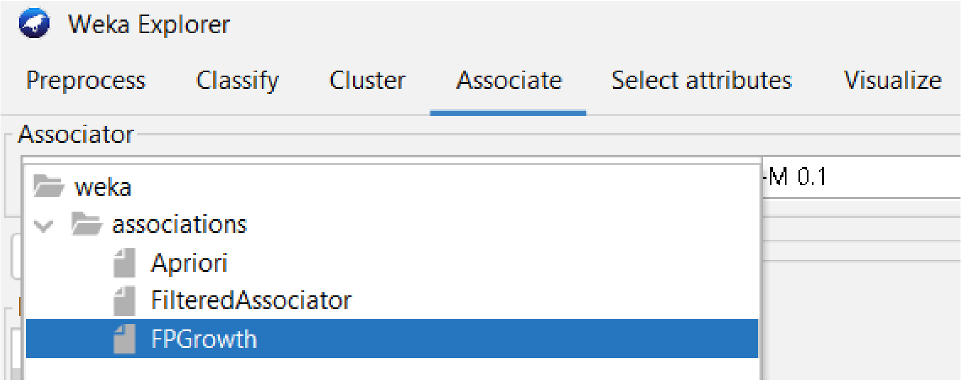
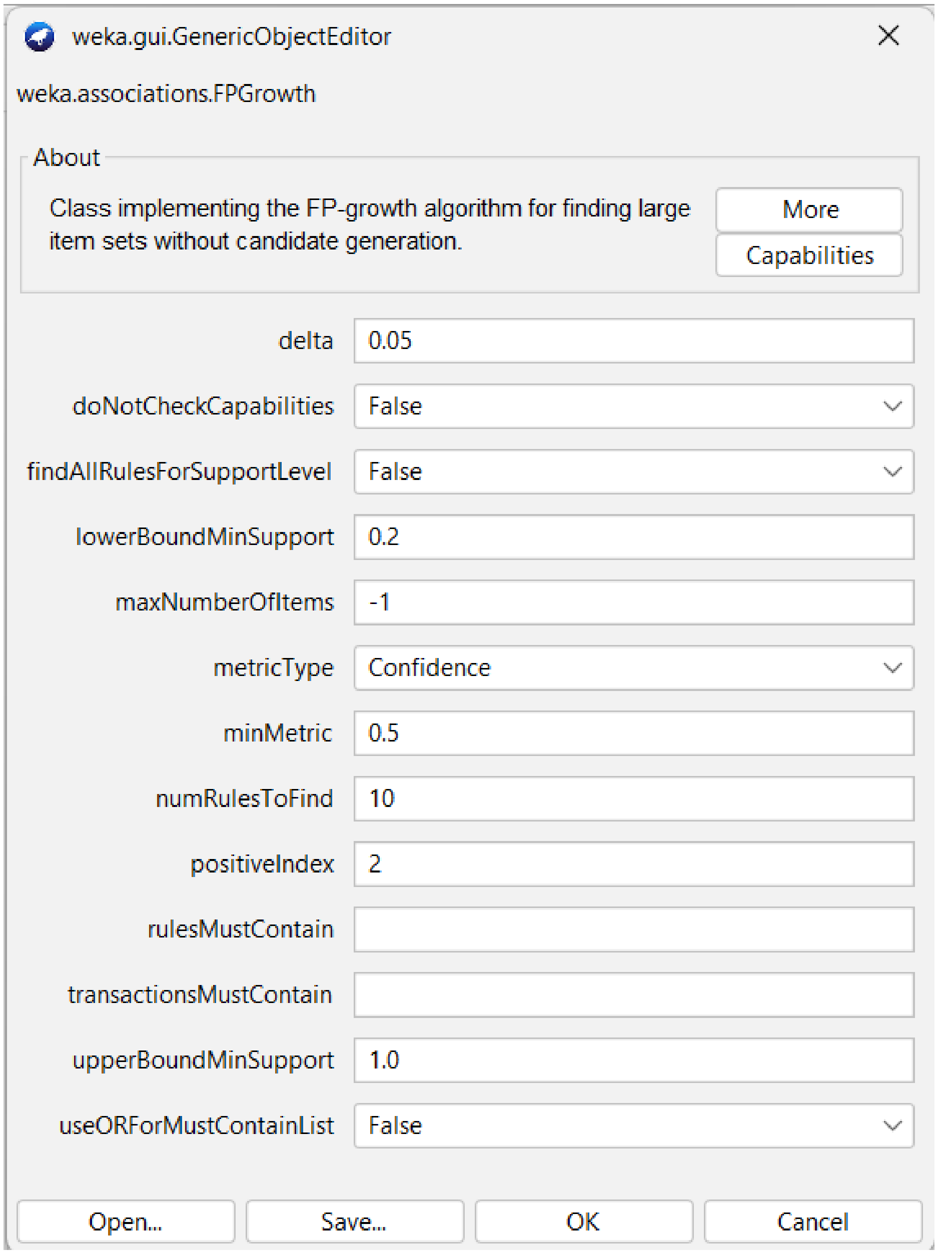


Figure 8: Choosing the FP-Growth from associate tab

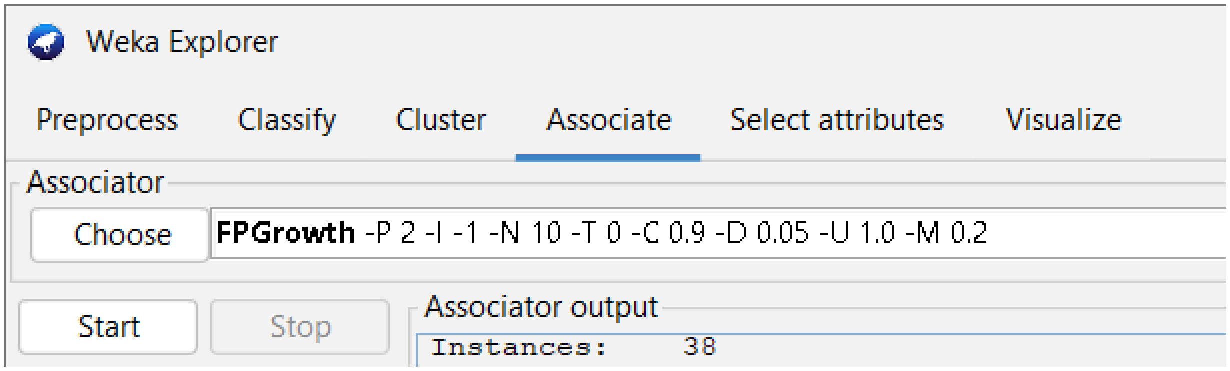


weka.associations.FPGrowth

Class implementing the FP-growth algorithm for finding large item sets without candidate generation.

delta	0.05
doNotCheckCapabilities	False
findAllRulesForSupportLevel	False
lowerBoundMinSupport	0.2
maxNumberOfItems	-1
metricType	Confidence
minMetric	0.5
numRulesToFind	10
positiveIndex	2
rulesMustContain	
transactionsMustContain	
upperBoundMinSupport	1.0
useORForMustContainList	False

Figure 9: Setting parameters of FP-Growth

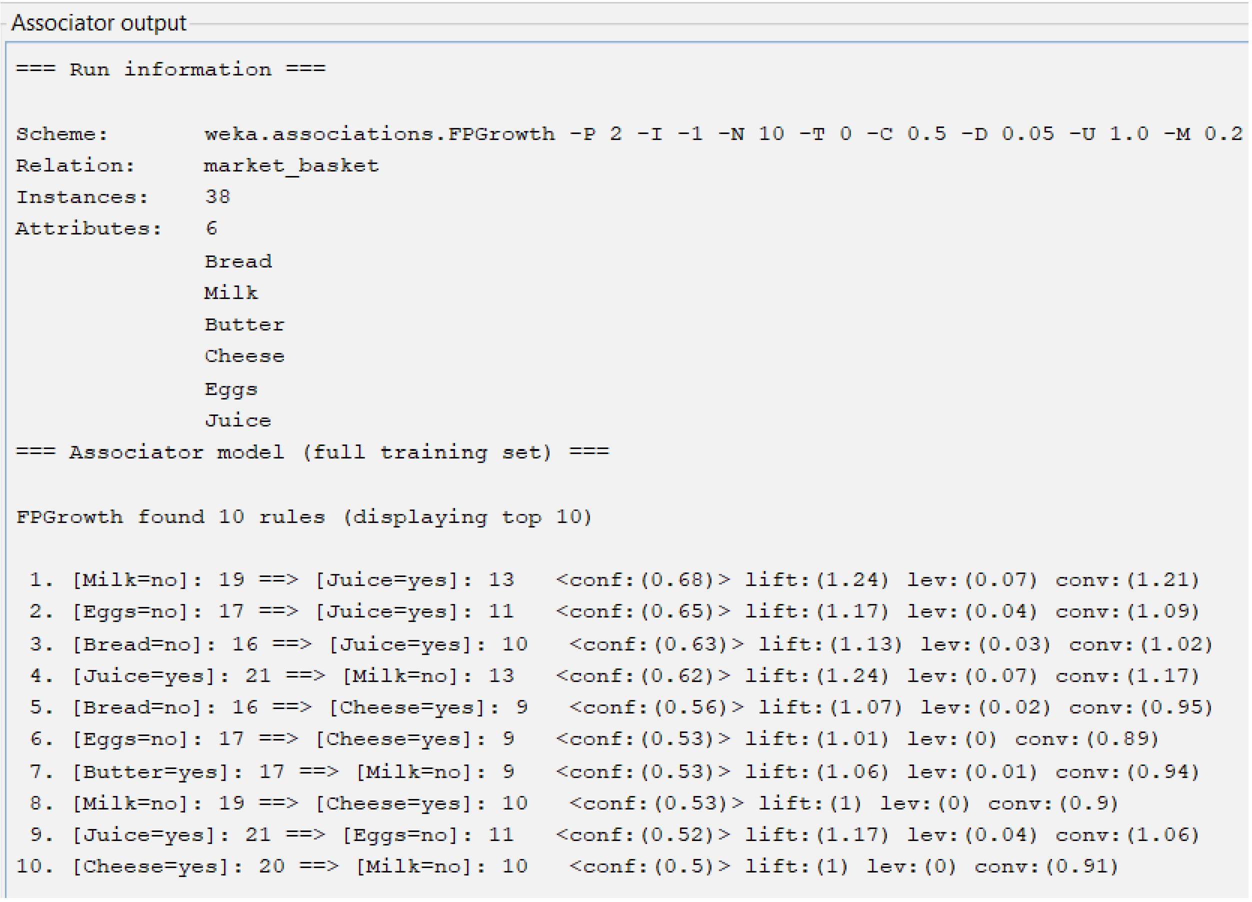


```
FPGrowth -P 2 -I -1 -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.2
```

```
Instances:     38
```

Figure 10: Starting the algorithm

Output



```
=== Run information ===

Scheme:       weka.associations.FPGrowth -P 2 -I -1 -N 10 -T 0 -C 0.5 -D 0.05 -U 1.0 -M 0.2
Relation:     market_basket
Instances:    38
Attributes:   6
              Bread
              Milk
              Butter
              Cheese
              Eggs
              Juice
=== Associator model (full training set) ===

FPGrowth found 10 rules (displaying top 10)

 1. [Milk=no]: 19 ==> [Juice=yes]: 13   <conf:(0.68)> lift:(1.24) lev:(0.07) conv:(1.21)
 2. [Eggs=no]: 17 ==> [Juice=yes]: 11   <conf:(0.65)> lift:(1.17) lev:(0.04) conv:(1.09)
 3. [Bread=no]: 16 ==> [Juice=yes]: 10   <conf:(0.63)> lift:(1.13) lev:(0.03) conv:(1.02)
 4. [Juice=yes]: 21 ==> [Milk=no]: 13   <conf:(0.62)> lift:(1.24) lev:(0.07) conv:(1.17)
 5. [Bread=no]: 16 ==> [Cheese=yes]: 9   <conf:(0.56)> lift:(1.07) lev:(0.02) conv:(0.95)
 6. [Eggs=no]: 17 ==> [Cheese=yes]: 9   <conf:(0.53)> lift:(1.01) lev:(0) conv:(0.89)
 7. [Butter=yes]: 17 ==> [Milk=no]: 9   <conf:(0.53)> lift:(1.06) lev:(0.01) conv:(0.94)
 8. [Milk=no]: 19 ==> [Cheese=yes]: 10   <conf:(0.53)> lift:(1) lev:(0) conv:(0.9)
 9. [Juice=yes]: 21 ==> [Eggs=no]: 11   <conf:(0.52)> lift:(1.17) lev:(0.04) conv:(1.06)
10. [Cheese=yes]: 20 ==> [Milk=no]: 10   <conf:(0.5)> lift:(1) lev:(0) conv:(0.91)
```

Figure 11: Output of the FP-Growth Algorithm

Discussion

The FP-Growth algorithm was applied to the market_basket dataset with a minimum support of 0.2 and a confidence threshold of 0.5. Unlike Apriori, FP-Growth efficiently compressed the dataset into a frequent pattern tree (FP-tree) and generated the association rules without needing to repeatedly scan the dataset for candidate itemsets.

The algorithm discovered 10 strong association rules. The top rules highlight notable patterns in customer purchasing behavior. For instance, [Milk=no] ⇒ [Juice=yes] had a confidence of 0.68 and a lift of 1.24, indicating that when milk was not purchased, juice was slightly more likely to be bought. Similarly, [Eggs=no] ⇒ [Juice=yes] with a confidence of 0.65 suggests that juice purchase is moderately associated with the absence of eggs in the basket. Other rules, such as [Bread=no] ⇒ [Cheese=yes] (confidence 0.56, lift 1.07) and [Butter=yes] ⇒ [Milk=no] (confidence 0.53, lift 1.06), reflect weaker but still meaningful associations.

The rules overall suggest that the absence of staple items like milk, eggs, and bread often coincides with the presence of other items like juice or cheese. While the confidence values are moderate rather than perfect, the lift values above 1 indicate positive associations stronger than random chance. These insights reveal actionable patterns for product placement, promotions, and inventory management. FP-Growth proved particularly effective in quickly identifying these rules without the overhead of candidate generation used in Apriori.

Conclusion

In this lab, the FP-Growth algorithm was successfully applied to the market_basket dataset to extract frequent itemsets and generate association rules. The top 10 rules revealed several moderate but meaningful co-occurrence patterns, such as the tendency for juice to be purchased when milk or eggs are absent, and associations between bread and cheese. FP-Growth demonstrated efficiency in handling the dataset and producing actionable insights with fewer iterations than Apriori. These results confirm that FP-Growth is a reliable and efficient tool for uncovering hidden relationships in transactional data, which can support marketing strategies, cross-selling decisions, and store layout optimization.