Lab Number: 7

Title

Generate association rules using Apriori algorithm for a suitable primary dataset

Objective

To generate meaningful association rules from a suitable primary dataset using the Apriori algorithm, with a minimum support of 0.2 and minimum confidence of 0.7, and to analyze the frequent itemsets and discovered relationships.

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:

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

Apriori Algorithm: Apriori is a classic algorithm used to generate frequent itemsets and from them, association rules. The key features of Apriori are:

- Uses a bottom-up search (level-wise)
- Eliminates infrequent itemsets using the Apriori property
- Generates candidate itemsets and retains only those meeting the minimum support threshold
- Produces association rules that meet both support and confidence requirements

Apriori is particularly efficient for datasets with boolean or categorical attributes.

Implementation

The following steps were performed to implement the apriori 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.

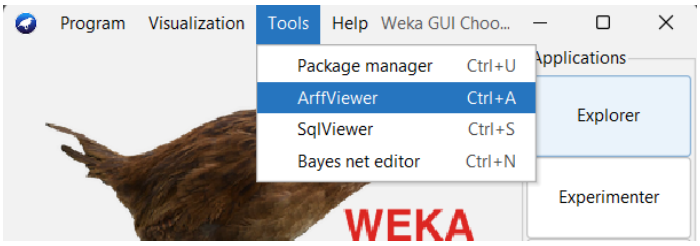


Figure 1: Opening the Arff Viewer

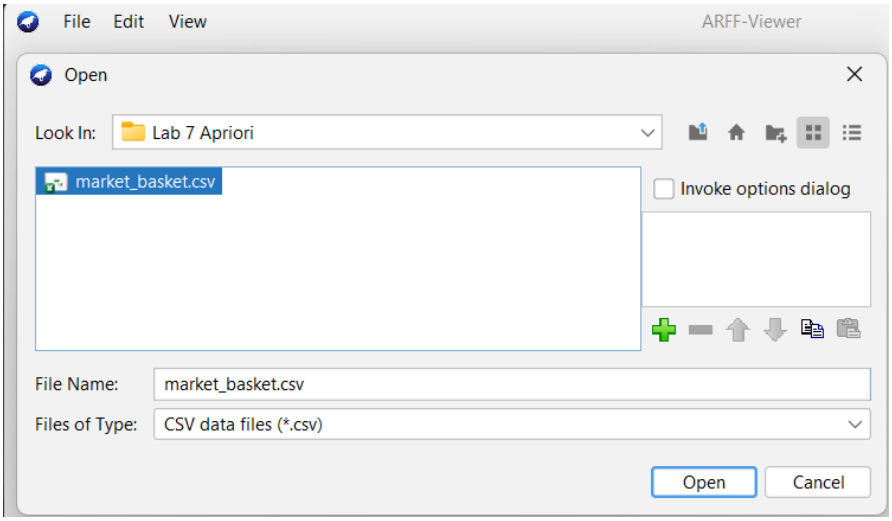
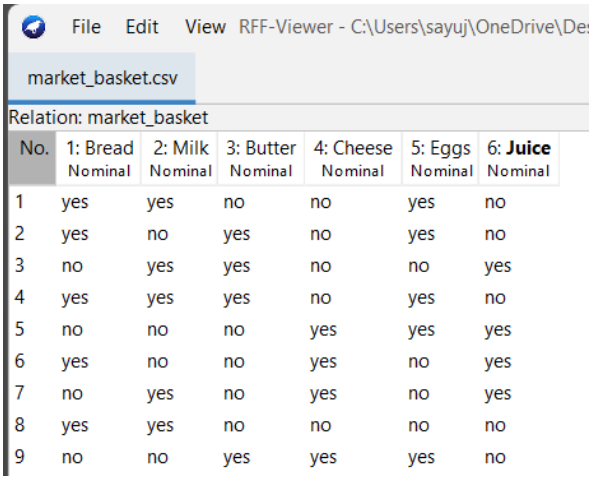


Figure 2: Selecting the csv file



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

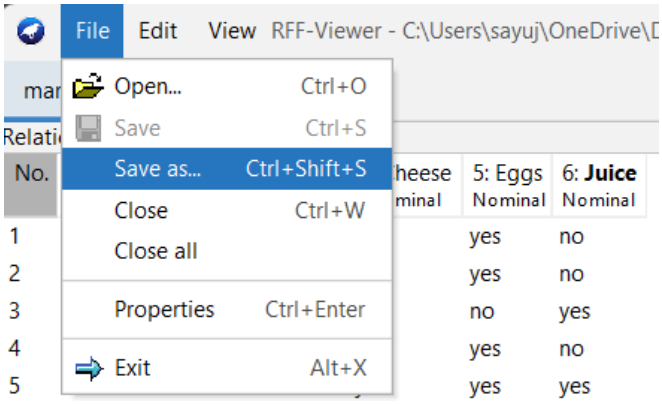


Figure 4: Option to save as arff

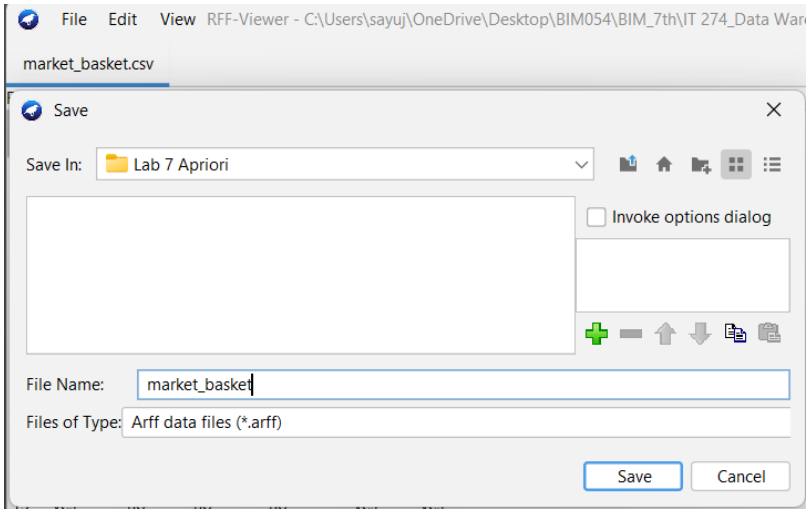


Figure 5: Saving as arff format

1. 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.

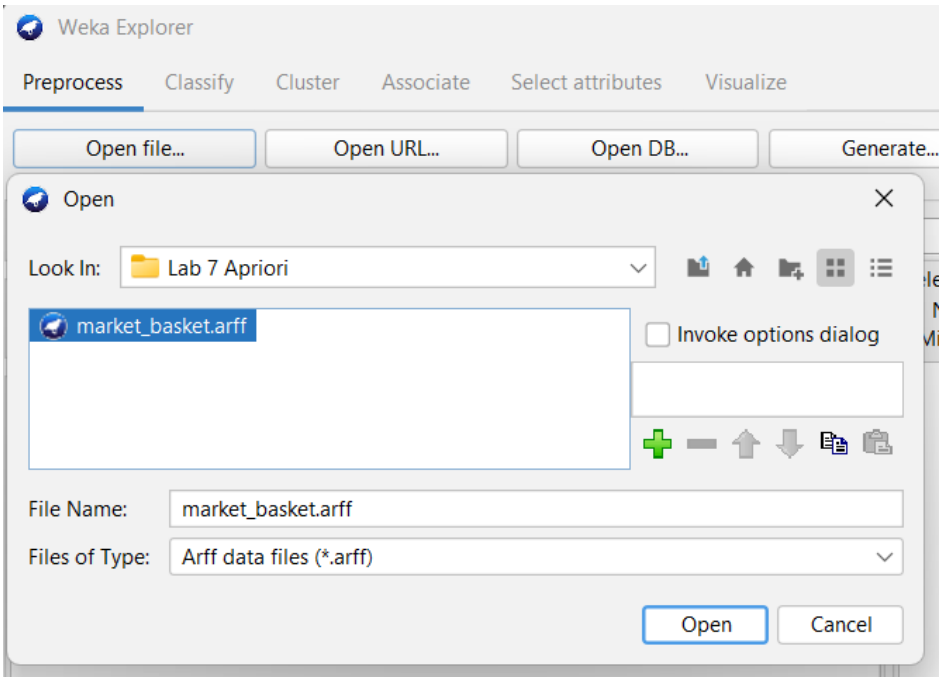


Figure 6: Opening the dataset

2. Running the Apriori Algorithm

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

- lowerBoundMinSupport: 0.2
- mertricType: Confidence
- minMetric: 0.7, 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.

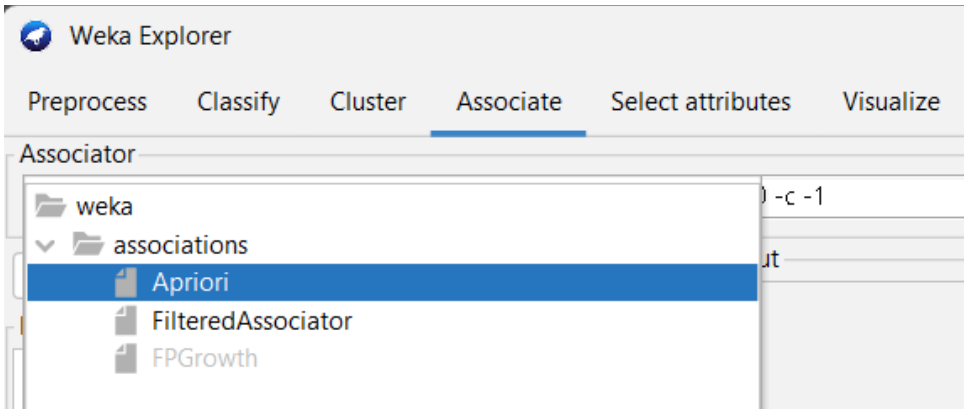


Figure 7: Choosing the Apriori from associate tab

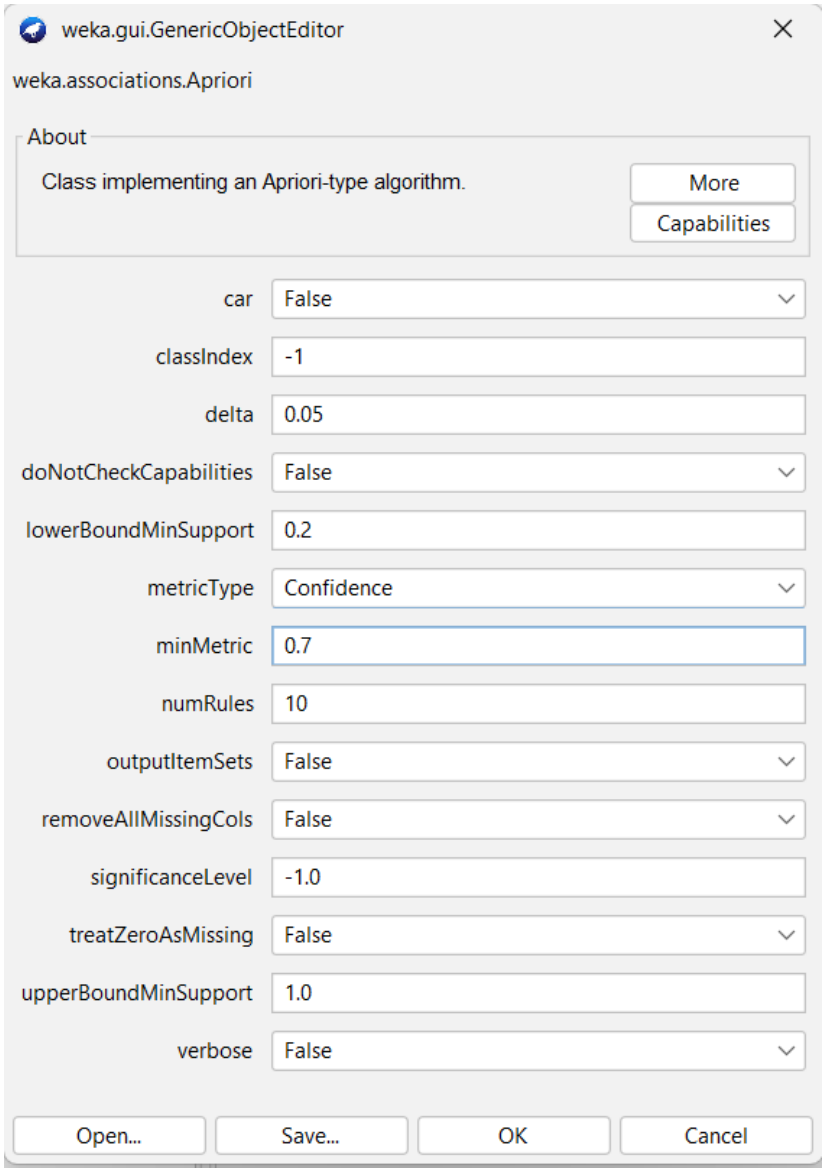


Figure 8: Setting parameters of Apriori

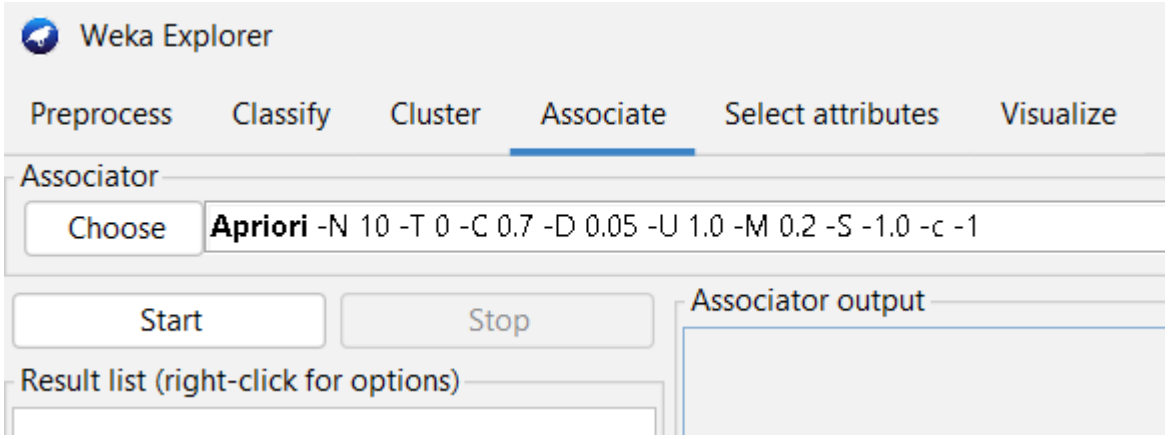


Figure 9: Starting the algorithm

Output

```
=== Run information ===

Scheme:       weka.associations.Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1
-S -1.0 -c -1
Relation:     market_basket
Instances:    38
Attributes:   6
              Bread
              Milk
              Butter
              Cheese
              Eggs
              Juice
=== Associator model (full training set) ===


Apriori
=======

Minimum support: 0.1 (4 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 18

Generated sets of large itemsets:

Size of set of large itemsets L(1): 12

Size of set of large itemsets L(2): 60

Size of set of large itemsets L(3): 132

Size of set of large itemsets L(4): 30

Best rules found:

 1. Bread=no Milk=no Butter=no 4 ==> Juice=yes 4     <conf:(1)> lift:(1.81)
lev:(0.05) [1] conv:(1.79)
 2. Bread=no Cheese=yes Juice=yes 4 ==> Butter=no 4    <conf:(1)> lift:(1.81)
lev:(0.05) [1] conv:(1.79)
 3. Milk=yes Butter=yes Cheese=yes 4 ==> Juice=no 4    <conf:(1)> lift:(2.24)
lev:(0.06) [2] conv:(2.21)
 4. Milk=yes Butter=yes Eggs=yes 4 ==> Juice=no 4     <conf:(1)> lift:(2.24)
lev:(0.06) [2] conv:(2.21)
 5. Cheese=yes Eggs=yes Juice=yes 4 ==> Butter=no 4    <conf:(1)> lift:(1.81)
lev:(0.05) [1] conv:(1.79)
 6. Butter=yes Cheese=yes Eggs=yes 4 ==> Juice=no 4    <conf:(1)> lift:(2.24)
lev:(0.06) [2] conv:(2.21)
```

Discussion

The Apriori algorithm successfully generated strong association rules from the market_basket dataset using a minimum support of 0.1 and a minimum confidence of 0.9. The algorithm explored multiple cycles (18 in total) and identified numerous large itemsets across different sizes, with the largest itemsets of size 3 being the most abundant (132).

The rules discovered indicate highly reliable co-occurrences. All top rules achieved perfect confidence (1.0), meaning that the consequent always occurred whenever the antecedent was present in the dataset. For example, the rule Bread=no Milk=no Butter=no ⇒ Juice=yes appeared 4 times with confidence 1.0 and a lift of 1.81, showing a strong positive association between the absence of these items and the purchase of juice. Similarly, combinations like Milk=yes Butter=yes Cheese=yes ⇒ Juice=no and Milk=yes Butter=yes Eggs=yes ⇒ Juice=no also had confidence 1.0 with the highest lift values (2.24), indicating that customers who bought these items together were highly unlikely to purchase juice.

The patterns revealed by the algorithm reflect meaningful relationships in a market basket context. Items such as milk, butter, cheese, and eggs tend to co-occur and show inverse associations with juice in certain combinations. The high lift and leverage values further confirm that these associations are significantly above what would be expected by chance, demonstrating the algorithm's effectiveness in uncovering actionable insights from transactional data.

Conclusion

In this experiment, the Apriori algorithm was applied to a market basket dataset to identify strong association rules using minimum support of 0.1 and minimum confidence of 0.9. The results produced multiple high-confidence rules with significant lift values, highlighting clear and interpretable co-occurrence patterns among items such as milk, butter, cheese, eggs, and juice. The analysis demonstrates that Apriori is an effective tool for discovering reliable and meaningful associations in categorical transactional datasets, which can be leveraged for decision-making in retail, marketing strategies, and product placement.