**A picture containing clipart

Description automatically generated**

**CST8390 - Business Intelligence and Data Analytics**

**Lab 9 - *Association Rule***

**Name: Min Li - Id: 040930563**

**Due Date:** Week 12 in corresponding lab sessions.

**Introduction**

The goal of this lab is to perform **Association Rule Mining** on Super Market dataset.

**Steps**

1. Open **Weka** and load the file **supermarket.arff** from “data” directory of **Weka**.

1. From the preprocess tab, click on the Edit button to view the instances. The “**t**” letters show which items were purchased.
2. Close Edit window and look at the **attributes**.

* 1. Number of attributes: 217.
  2. Number of instances: 4627.

1. **Find** tea, coffee, medicines and flowers and see **how many times** each of the item was purchased?

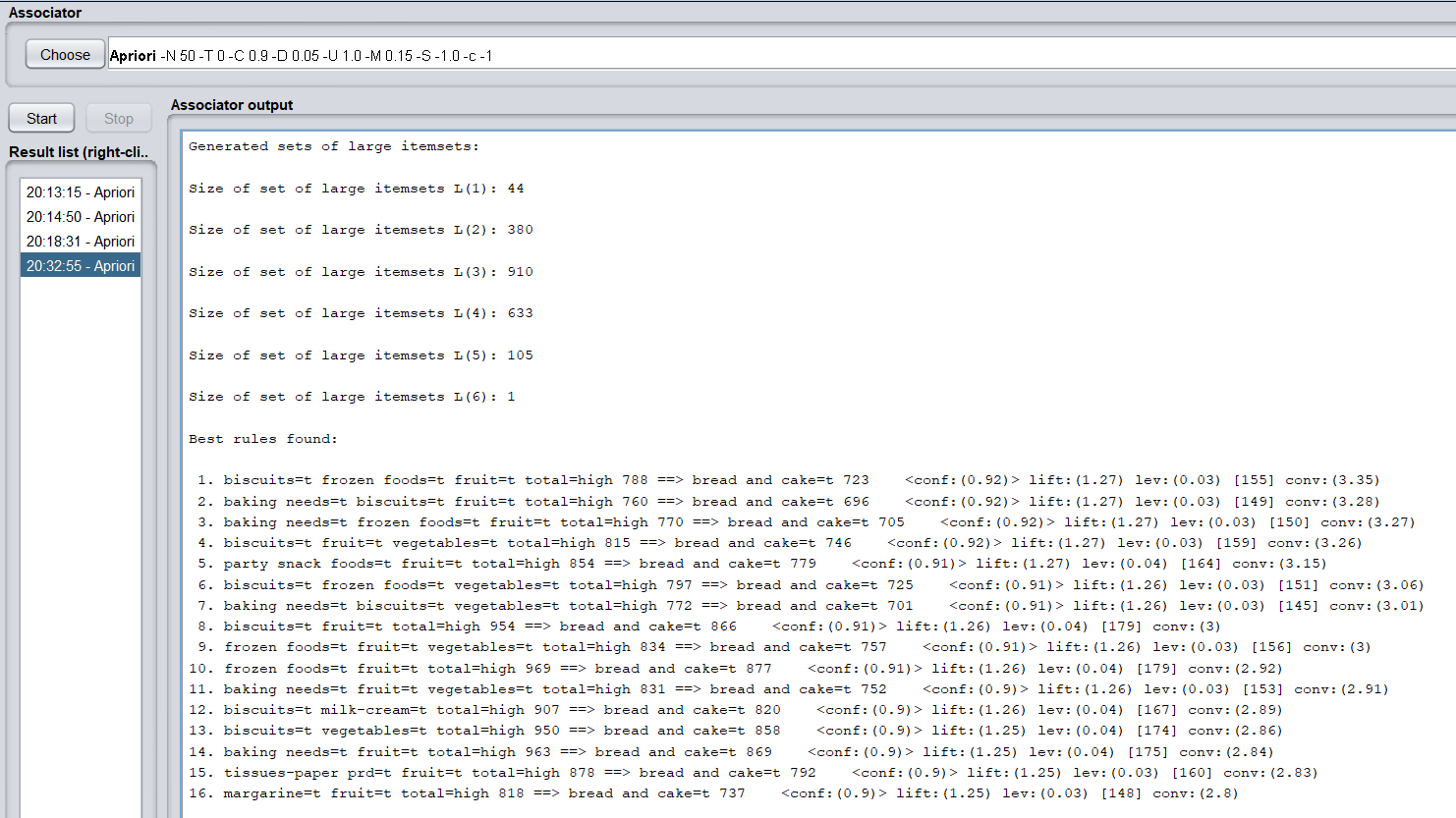
Tea: 896.

Coffee: 1094.

Medicines: 204.

Flowers: 0.

1. Click on the “Associate” tab. The **Apriori algorithm** should already be selected but click on the **text field** to **edit** the parameters. Find the **lowerBoundMinSupport**. This is the minimum support percentage that is required to create the rule sets. Set it to 0.25 (i.e., 25%). Set the “numRules” to 15, to print out the **top 15 rules** that are found. Click “Ok” to close the window and then click “Start” to run the algorithm.
2. The algorithm should run for a number of seconds and then return with **no rules**. That means that no rules were found that have a minimum support of 20%. Lower the support to 15% and run it again. Set numRules to 50. How many **rules** were generated this time? 16.



1. The rules are **sorted** from highest lift to lowest. The lift tells you **how often** the rules are related, or the strength of the rule. Which rules have the **highest lift**? Rule1,2,3,4,5.

1. biscuits=t frozen foods=t fruit=t total=high 788 ==> bread and cake=t 723 <conf:(0.92)> lift:(1.27) lev:(0.03) [155] conv:(3.35)

2. baking needs=t biscuits=t fruit=t total=high 760 ==> bread and cake=t 696 <conf:(0.92)> lift:(1.27) lev:(0.03) [149] conv:(3.28)

3. baking needs=t frozen foods=t fruit=t total=high 770 ==> bread and cake=t 705 <conf:(0.92)> lift:(1.27) lev:(0.03) [150] conv:(3.27)

4. biscuits=t fruit=t vegetables=t total=high 815 ==> bread and cake=t 746 <conf:(0.92)> lift:(1.27) lev:(0.03) [159] conv:(3.26)

5. party snack foods=t fruit=t total=high 854 ==> bread and cake=t 779 <conf:(0.91)> lift:(1.27) lev:(0.04) [164] conv:(3.15)

1. Lower the support now to 10% and re-run the algorithm. Since more rules are included in the search, this time it should take a long time to run. What is the **highest lift** now that was found and what are the rules? The highest lift is 1.3. Rule1,2,3.

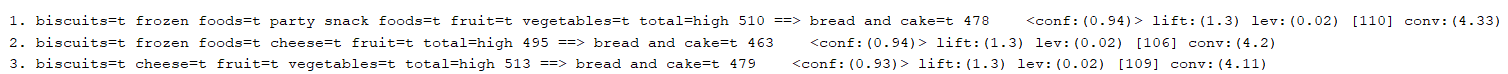
**Example**:

If you get:

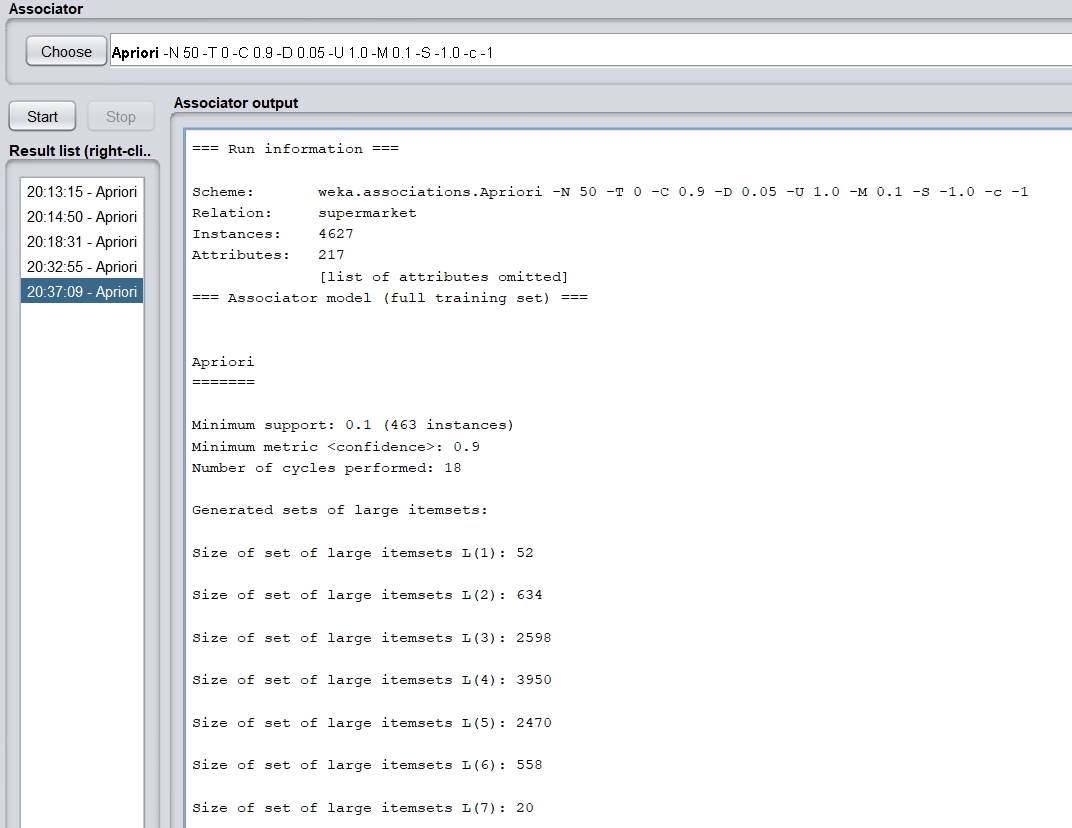
*frozen foods=t fruit=t total=high 969 ==> bread and cake=t 877 <conf:(0.91)> lift:(1.26) lev:(0.04) [179] conv:(2.92),*

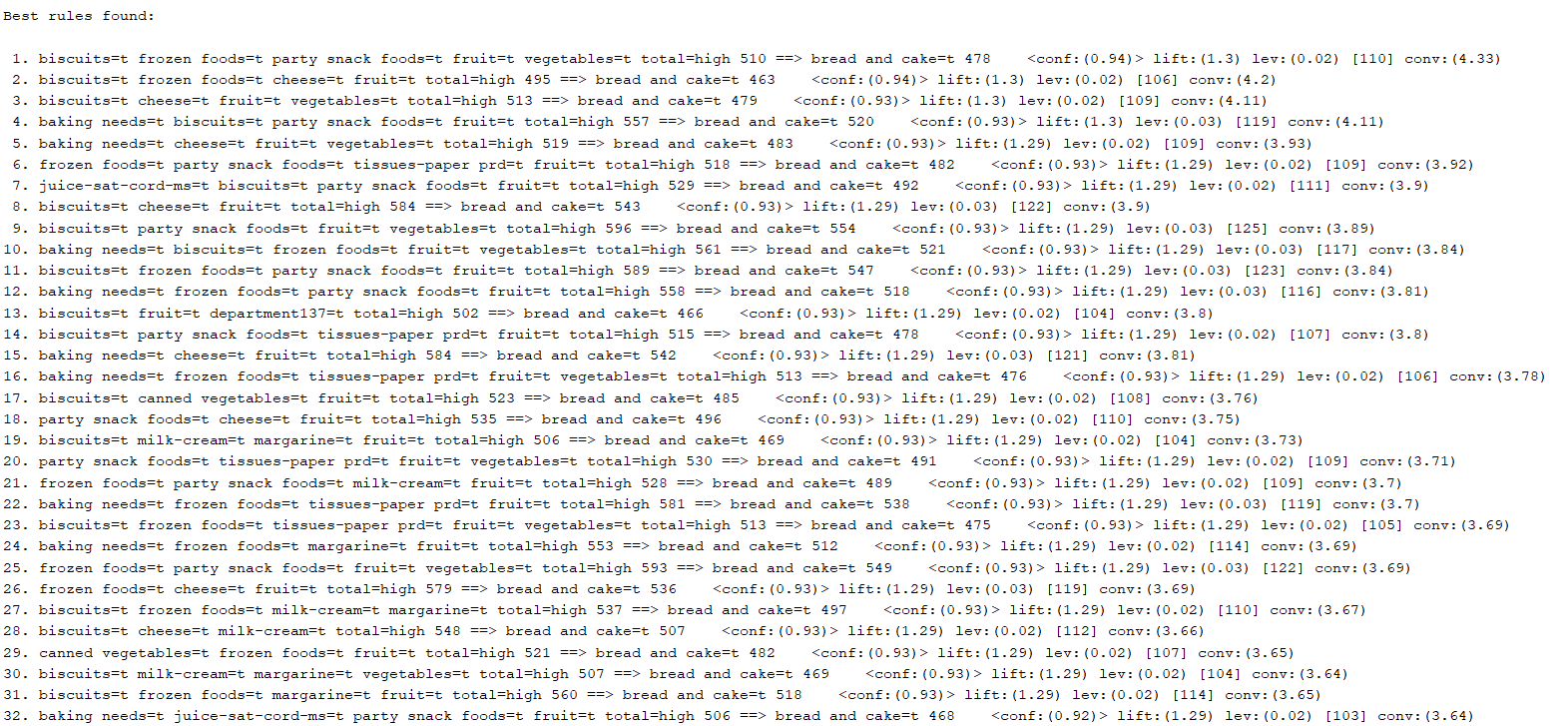
you need to writer the rule as:

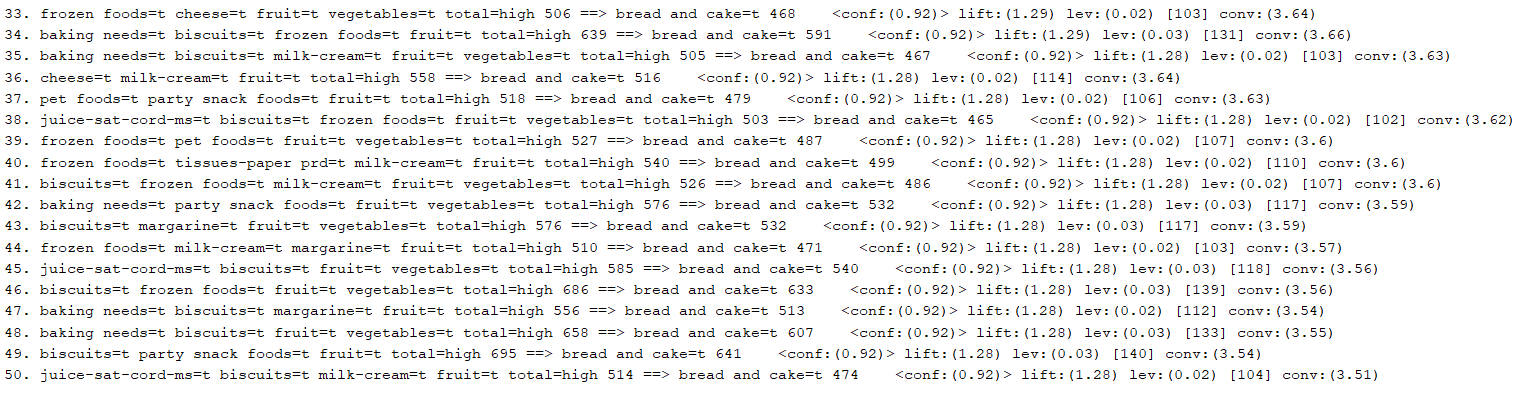
***frozen foods, fruit ==> bread and cake (conf: 0.91, lift: 1.26).***



1. biscuits, frozen foods, party snack foods, fruit, vegetables ==> bread and cake (conf: 0.94, lift:1.3).
2. biscuits, frozen foods, cheese, fruit ==> bread and cake (conf: 0.94, lift:1.3).
3. biscuits, cheese, fruit, vegetables ==> bread and cake (conf: 0.93, lift:1.3).







**REMEMBER:**

In order to get grades, you need to upload filled-in answer document and screenshots from steps 6-8.

***FOR YOUR ANALYSIS:***

***\* Option 1****: Use your own words to explain* ***Association Rule Mining*** *and situations where apply it.*

***\* Option 2****: What is the strategy to identify association rules in a specific scenario?*

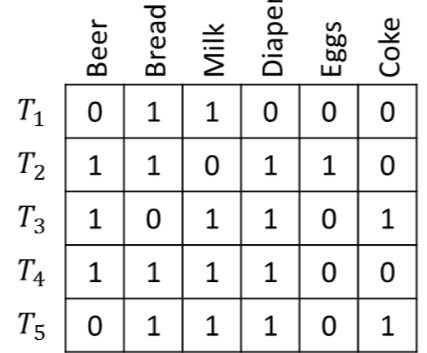
***Option 1****:* **Association Rule Mining** is one of the ways to find patterns in data. It finds: 1.features (dimensions) which occur together 2. features (dimensions) which are “correlated”. What does the value of one feature tell us about the value of another feature? For example, people who buy diapers are likely to buy baby powder. Or we can rephrase the statement by saying: If (people buy diaper), then (they buy baby powder). Note the if, then rule. This does not necessarily mean that if people buy baby powder, they buy diaper. In General, we can say that if condition A tends to B it does not necessarily mean that B tends to A. Watch the directionality!

We can use Association Rules in any dataset where features take only two values i.e., 0/1. Some examples are here: 1. **Market Basket Analysis** is a popular application of Association Rules. 2. People who visit webpage X are likely to visit webpage Y. 3. People who have age-group [30,40] & income [>$100k] are likely to own home.

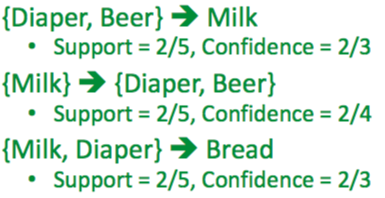
The measures of effectiveness of the rule are as Follows: 1. Support 2. Confidence 3. Lift 4. Others: Affinity, Leverage. We are going to discuss Support and Confidence in more detail using an example dataset.



The above dataset can also be represented like this:



**Support** means how much historical data supports your rule and **Confidence** means how confident are we that the rule holds. Support can be calculated as the fraction of rows containing both A and B or joint probability of A and B. Among rows containing A, Confidence is the fraction of rows containing B or conditional probability of B given A.



**Lift** is the ratio Confidence is to Support. If the lift is < 1 then A and B are negatively correlated else positively correlated and if it is equal to 1 it is not correlated.

Reference: <https://towardsdatascience.com/association-rule-mining-be4122fc1793>