Data Mining

CSE-5334

Assignment – 3

PA 3: Association Analysis - Apriori Algorithm

Harshith Pashikanti – 1001974588

Venkata Sainath Reddy Palavala – 1001949223

### Task 1: Dataset Pre-process

We have been given a dataset named “dataset\_group.csv” file with the following columns dates, transaction\_id, items as shown in figure – 1.



Figure – 1

Our task is to format the data in the pre-processing, as shown in figure - 2.

A screenshot of a computer

Description automatically generated with medium confidence

Figure - 2

First, we must be able to read the CSV file. After the file has been read, the pre-processing can begin. We were able to achieve the dataset in the desired format utilizing a loop and confirming each line with its transaction id and item value after successfully reading the file. We then removed the duplicates and sorted the information in ascending order. Finally, after performing the pre-processing, we saved the dataset as dataset.csv in a new CSV file. This new CSV file now contains the data that we require. We didn't disturb the original file this way, and we're now ready to move on.

Printing the first seven heads of the data frame. This is shown in figure – 3.

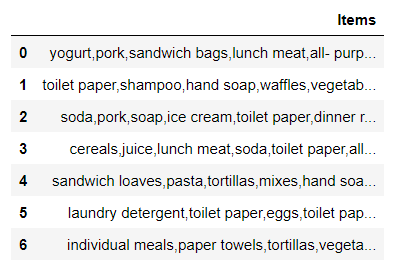


Figure – 3

### Task 2: Run apriory.py and Evaluate Results

We should be able to use the apriori method in this task. We should be able to use the file apriory.py to execute and output the results using this method.

In this work, we must utilize distinct minimal support and minimum confidence numbers. Although the default settings for minimal support and minimum confidence are 0.15 and 0.65, respectively.

Apriori Algorithm:

An algorithm known as Apriori is a common one in data mining. It's used to identify the most frequently occurring elements and meaningful associations in a dataset. The Apriori method is used for identifying the most frequently occurring itemsets in a dataset using the boolean association rule. Since it makes use of previous knowledge about common itemset features, the method is referred to as Apriori. This is achieved using an iterative technique or level-wise approach, in which k-frequent itemsets are utilized to locate k+1 itemsets. An essential feature known as the Apriori property is utilized to boost the effectiveness of level-wise production of frequent itemsets. This property helps by minimizing the search area, which in turn serves to maximize the productivity of the level-wise creation of frequent patterns.

Minimum Support:

The selection of the frequent item set that will be used in the rule-building process is the first stage in association rule mining. A threshold is employed in this stage to reject things that are not included in the frequent itemset, also known as minimal support.

Minimum Confidence:

The degree of certainty or trustworthiness associated with each detected pattern is referred to as confidence. Assume that min sup is the minimal support level. If the occurrence frequency of an itemset is more than or equal to min sup, it meets minimal support.

Case – 1:

In case 1, we consider the minimum support = 0.3 and minimum confidence = 0.8. With the minimum support value of 0.3, the number of items that will satisfy the support values increases. This can be observed in figure – 4. Because of this, the number of rules in the dataset which are generated is also increased. This can be followed in figure – 5.

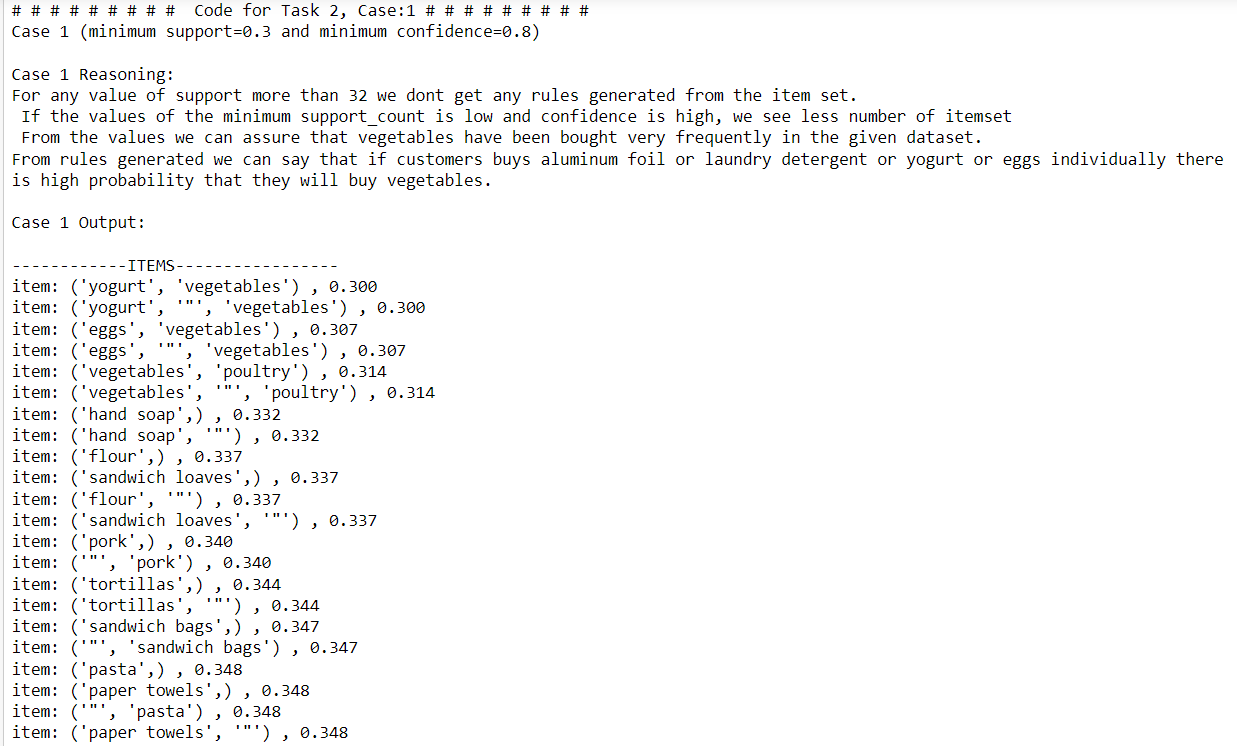


Figure – 4

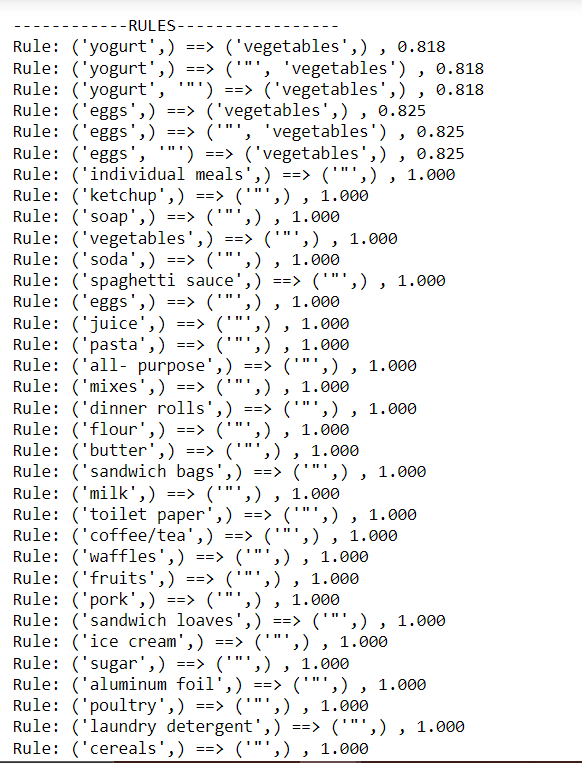


Figure – 5

Case – 2:

In case 2, we consider the minimum support = 0.2 and minimum confidence = 0.8. We see that the no. of itemset with support values more than 0.2 are less. Now we are looking into this case. The itemset that satisfies will steadily shrink as the min support value rises, resulting in fewer rules being generated. The number of occurrences in the transaction list that fit the criterion can be reduced by lowering the confidence value. Figure 6 illustrates this point.

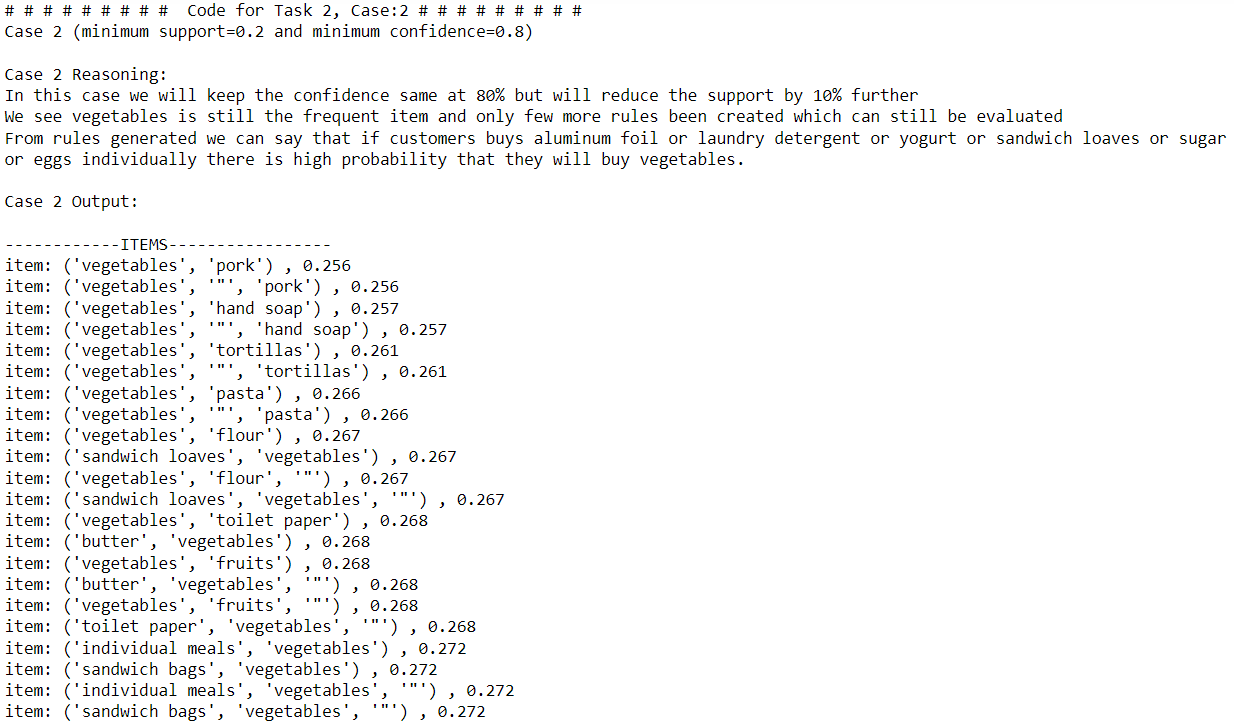


Figure – 6

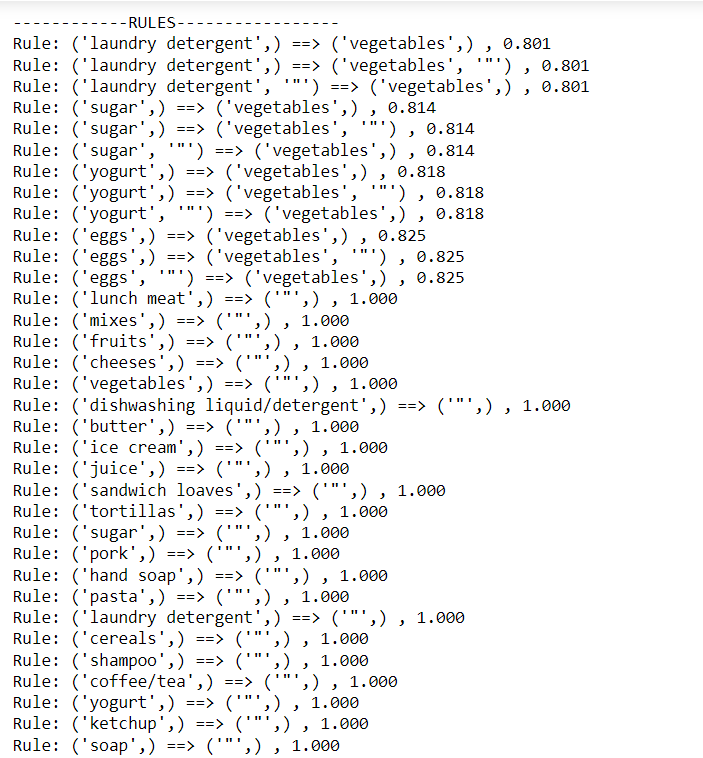


Figure – 7

Case – 3:

In case 3, we use 0.1 as the minimal support and 0.8 as the minimum confidence. The number of itemsets with support values greater than 0.15 is decreasing. This matter is still being investigated. As the minimum support value decreases, the number of itemsets that satisfy increases, resulting in a somewhat greater number of rules being generated. Figure 11 illustrates this point.

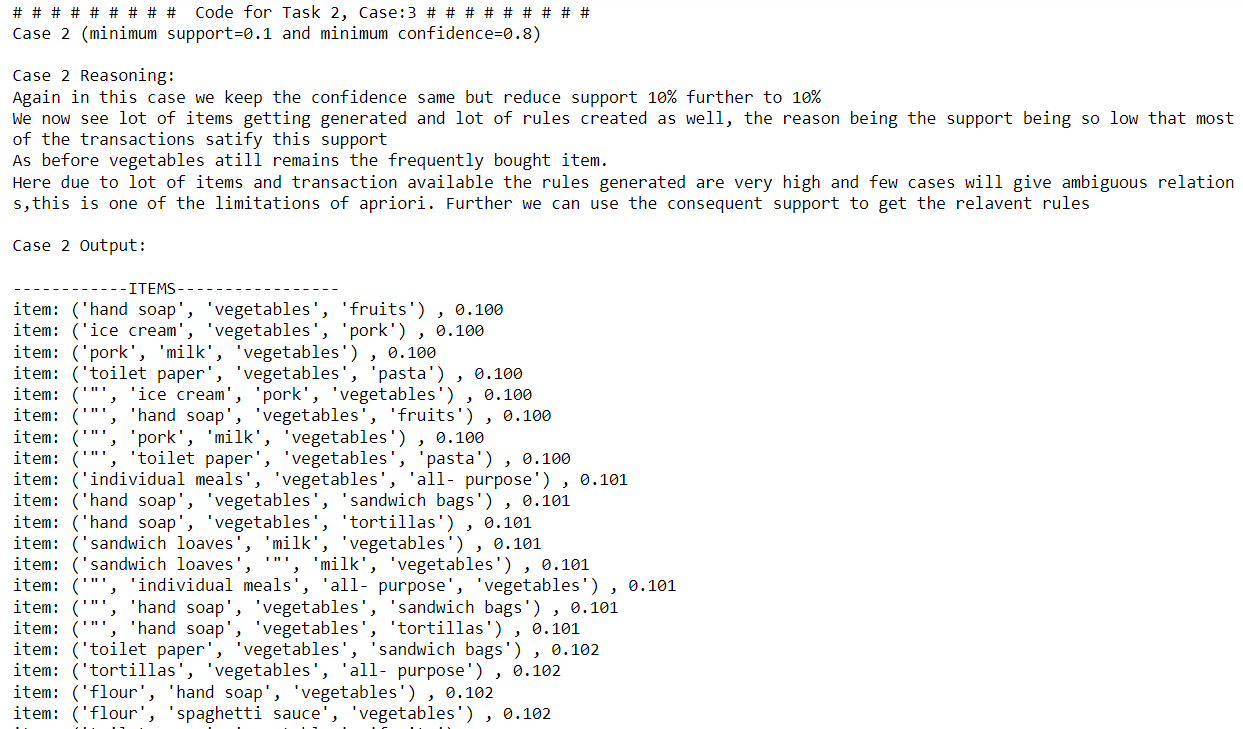


Figure – 8



Figure – 9

* Task for the Sample Dataset:

Case – 1:

In case 1, we consider the minimum support = 0.4 and minimum confidence = 0.9. With minimum support value = 0.4 and confidence value = 0.9, the number of item sets that satisfy the minimum support value increases. Since the confidence value is = 0.9, the number of rules that can be generated is lower using these itemsets. This can be observed in figure – 10

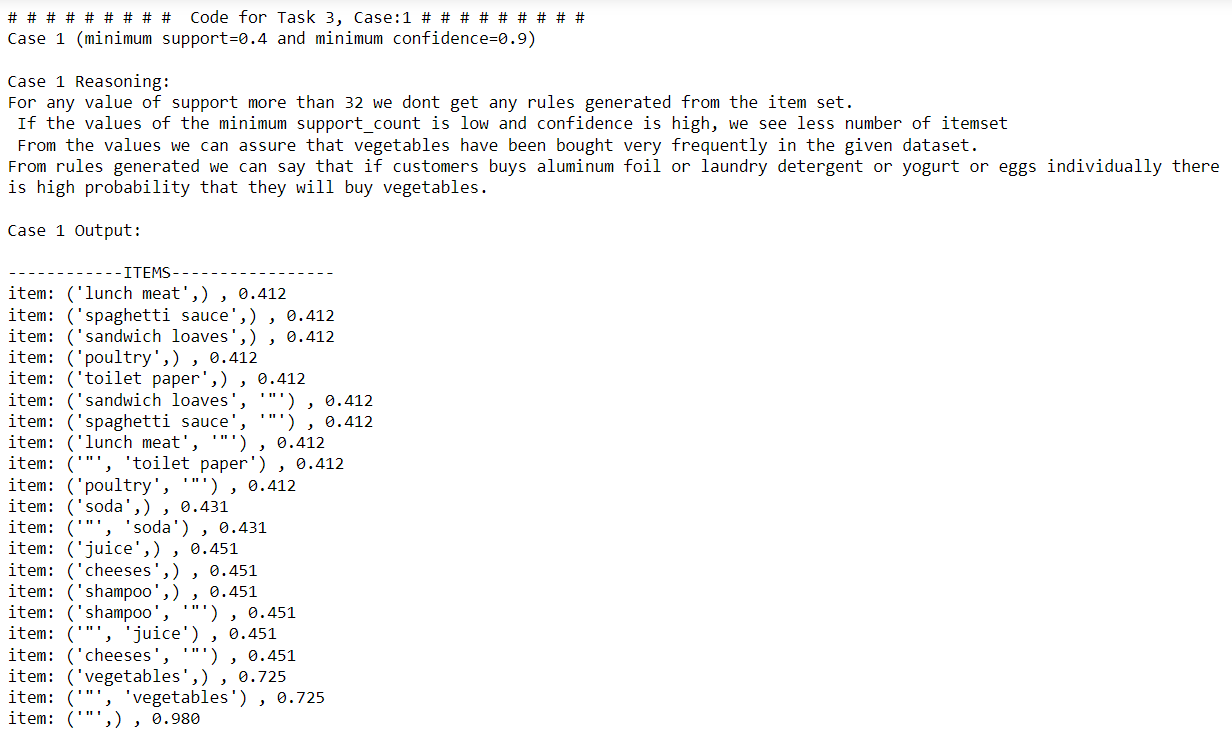


Figure – 14

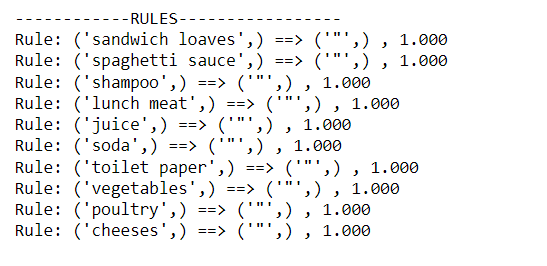


Figure – 11

Case – 2:

In case 2, we consider minimum support = 0.4 and minimum confidence = 0.6. With minimum support value = 0.4 and minimum confidence value = 0.6, the number of itemsets that satisfy the minimum support value increases. Thereby increasing the number of rules that can be generated using this itemset. We can also observe that there isn't much difference compared to case 1. This can be followed in figure – 12.

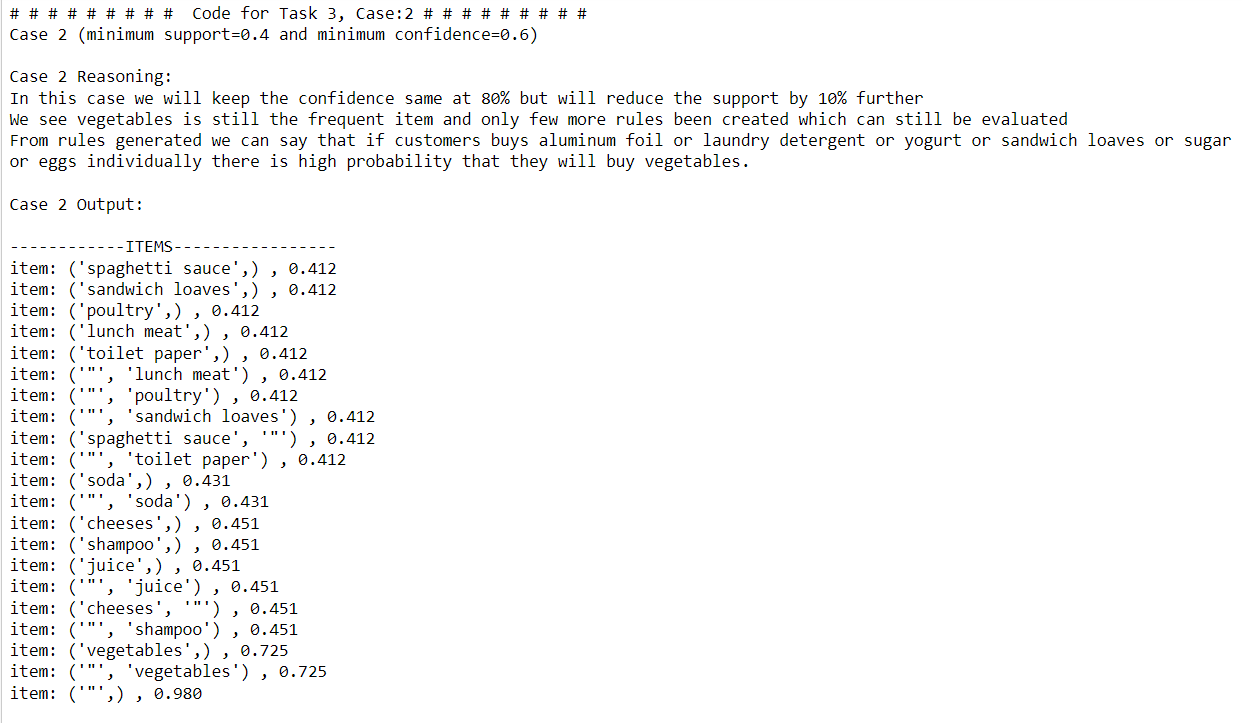


Figure – 12

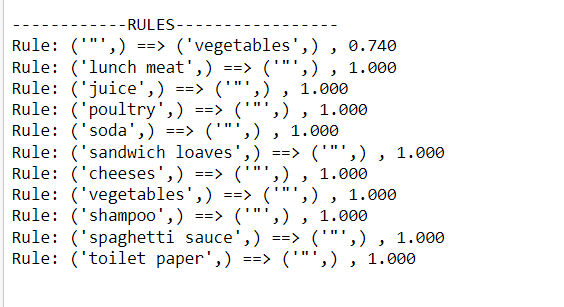


Figure – 13

Case – 3:

In case 3 we are considering the minimum support =0.15 and minimum confidence = 0.8. With minimum support value = 0.15 and minimum confidence value = 0.8. The number of itemsets that will satisfy the minimum support value increases. Thereby increasing the number of rules that can be generated using these item sets. We can also observe that there isn't much difference compared to case 1. The values range between 0.3 to 0.71 mainly. This can be followed in figure – 14.

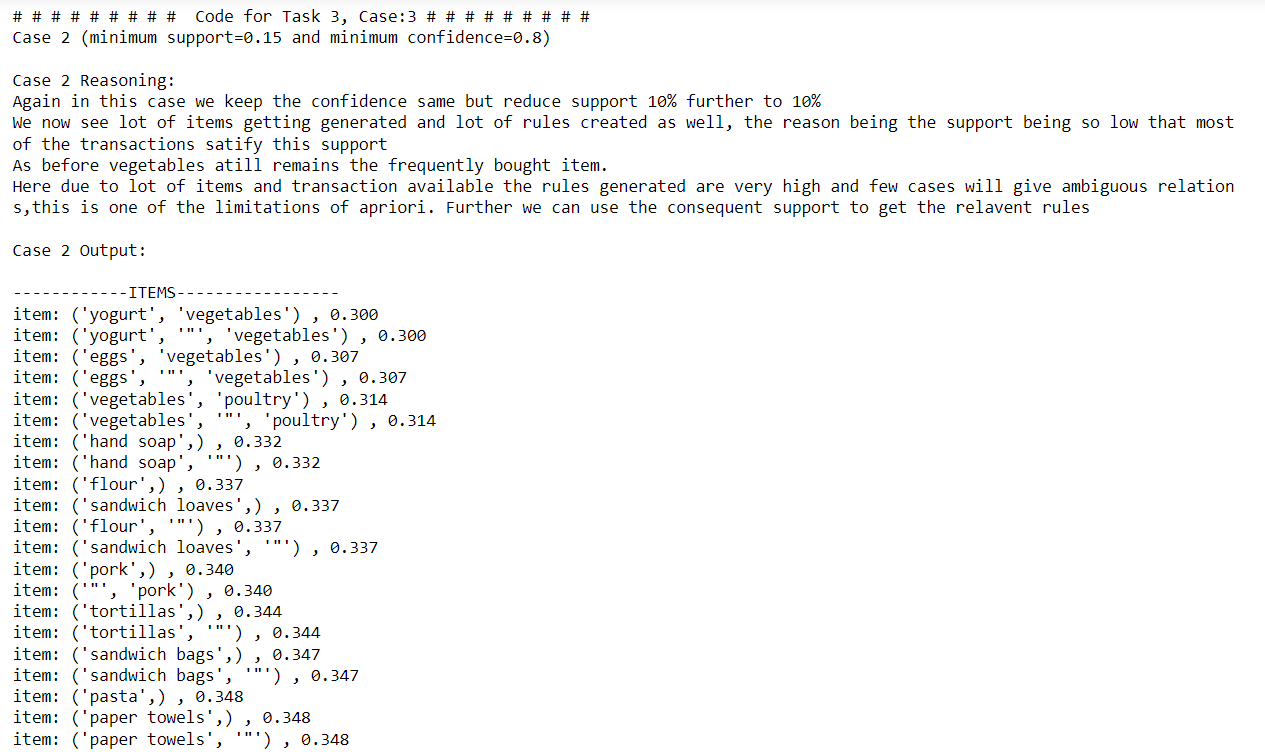


Figure – 14

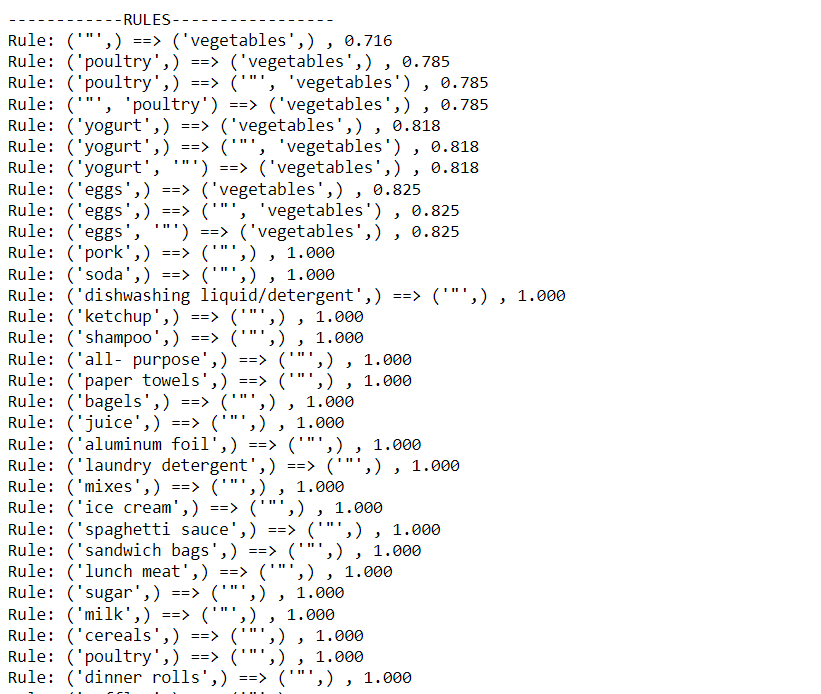


Figure – 15

**Solving theoretically using the formula for calculating the small data set:**

**Case1:**

Consider Support = 50% and confidence = 75%

|  |  |
| --- | --- |
| Transaction ID | Items Purchased |
| 1 | Bread, Cheese, Egg, Juice |
| 2 | Bread, Cheese, Juice |
| 3 | Bread, Milk, Yogurt |
| 4 | Bread, Juice, Milk |
| 5 | Cheese, Juice, Milk |

Step 1: Find frequent itemset and their support

|  |  |  |
| --- | --- | --- |
| Item | Frequency | Support (In %) |
| Bread | 4 | 4/5 = 80% |
| Cheese | 3 | 3/5 = 60% |
| Egg | 1 | 1/5 = 20% |
| Juice | 4 | 4/5 = 80% |
| Milk | 3 | 3/5 = 60% |
| Yogurt | 1 | 1/5 = 20% |

Support(Item) = Frequency of Item / No. of Transactions

Step 2: Remove all the items whose support is below the given minimum support

|  |  |  |
| --- | --- | --- |
| Item | Frequency | Support (In %) |
| Bread | 4 | 4/5 = 80% |
| Cheese | 3 | 3/5 = 60% |
| Juice | 4 | 4/5 = 80% |
| Milk | 3 | 3/5 = 60% |

Step 3: Now form the two items candidate set and write their frequencies.

|  |  |  |
| --- | --- | --- |
| Item Pair | Frequency | Support (In %) |
| Bread, cheese | 2 | 40% |
| Bread, Juice | 3 | 60% |
| Bread, Milk | 2 | 40% |
| Cheese, Juice | 3 | 60% |
| Cheese, Milk | 1 | 20% |
| Juice, Milk | 2 | 40% |

Step 4: Remove all the items whose support is below the given minimum support.

|  |  |  |
| --- | --- | --- |
| Item Pair | Frequency | Support (In %) |
| Bread, Juice | 3 | 3/5 = 60% |
| Cheese, Juice | 3 | 3/5 = 60% |

Step 5: Generate Rules

For rules we consider item pairs:

1. (Bread, Juice)

Bread -> Juice and Juice -> Bread

1. (Cheese -> Juice)

Cheese -> Juice and Juice -> Cheese

Confidence(A -> B) = Support(A U B) / Support(A)

Therefore,

1. Confidence (Bread -> Juice) = Support(Bread U Juice) / Support(Bread)

= 3/5 \* 5/4 = 75%

1. Confidence (Juice -> Bread) = Support(Juice U Bread) / Support(Juice)

= 3/5 \* 5/4 = 75%

1. Confidence (Cheese -> Juice) = Support(Cheese U Juice) / Support(Cheese)

= 3/5 \* 5/3 = 100%

1. Confidence (Juice -> Cheese) = Support(Juice U Cheese) / Support(Juice)

= 3/5 \* 5/4 = 75%

All the above rules are good because the confidence of each rule is greater than or equal to minimum confidence given in the problem.

**Case2:**

Consider Support = 52.5% and confidence = 70%

|  |  |
| --- | --- |
| Transaction ID | Items Purchased |
| 1 | Bread, Cheese, Egg, Juice |
| 2 | Bread, Cheese, Juice |
| 3 | Bread, Milk, Yogurt |
| 4 | Bread, Juice, Milk |
| 5 | Cheese, Juice, Milk |

Step 1: Find frequent itemset and their support

|  |  |  |
| --- | --- | --- |
| Item | Frequency | Support (In %) |
| Bread | 4 | 4/5.25 = 76% |
| Cheese | 3 | 3/5.25 = 57% |
| Egg | 1 | 1/5.25 = 19% |
| Juice | 4 | 4/5.25 = 76% |
| Milk | 3 | 3/5.25 = 57% |
| Yogurt | 1 | 1/5.25 = 19% |

Support(Item) = Frequency of Item / No. of Transactions

Step 2: Remove all the items whose support is below the given minimum support

|  |  |  |
| --- | --- | --- |
| Item | Frequency | Support (In %) |
| Bread | 4 | 4/5.25 = 76% |
| Cheese | 3 | 3/5.25 = 57% |
| Juice | 4 | 4/5.25 = 76% |
| Milk | 3 | 3/5.25 = 57% |

Step 3: Now form the two items candidate set and write their frequencies.

|  |  |  |
| --- | --- | --- |
| Item Pair | Frequency | Support (In %) |
| Bread, cheese | 2 | 38% |
| Bread, Juice | 3 | 57% |
| Bread, Milk | 2 | 38% |
| Cheese, Juice | 3 | 57% |
| Cheese, Milk | 1 | 19% |
| Juice, Milk | 2 | 38% |

Step 4: Remove all the items whose support is below the given minimum support.

|  |  |  |
| --- | --- | --- |
| Item Pair | Frequency | Support (In %) |
| Bread, Juice | 3 | 3/5.25 = 57% |
| Cheese, Juice | 3 | 3/5.25 = 57 % |

Step 5: Generate Rules

For rules we consider item pairs:

1. (Bread, Juice)

Bread -> Juice and Juice -> Bread

1. (Cheese -> Juice)

Cheese -> Juice and Juice -> Cheese

Confidence(A -> B) = Support(A U B) / Support(A)

Therefore,

1. Confidence (Bread -> Juice) = Support(Bread U Juice) / Support(Bread)

= 3/5.25 \* 5.25/4 = 75%

1. Confidence (Juice -> Bread) = Support(Juice U Bread) / Support(Juice)

= 3/5.25 \* 5.25/4 = 75%

1. Confidence (Cheese -> Juice) = Support(Cheese U Juice) / Support(Cheese)

= 3/5.25 \* 5.25/3 = 100%

1. Confidence (Juice -> Cheese) = Support(Juice U Cheese) / Support(Juice)

= 3/5.25 \* 5.25/4 = 75%

All the above rules are good because the confidence of each rule is greater than or equal to minimum confidence given in the problem.

**Case3:**

Consider Support = 55% and confidence = 85%

|  |  |
| --- | --- |
| Transaction ID | Items Purchased |
| 1 | Bread, Cheese, Egg, Juice |
| 2 | Bread, Cheese, Juice |
| 3 | Bread, Milk, Yogurt |
| 4 | Bread, Juice, Milk |
| 5 | Cheese, Juice, Milk |

Step 1: Find frequent itemset and their support

|  |  |  |
| --- | --- | --- |
| Item | Frequency | Support (In %) |
| Bread | 4 | 4/5.5 = 72% |
| Cheese | 3 | 3/5.5 = 55% |
| Egg | 1 | 1/5.5 = 18% |
| Juice | 4 | 4/5.5 = 72% |
| Milk | 3 | 3/5.5 = 55% |
| Yogurt | 1 | 1/5.5 = 18% |

Support(Item) = Frequency of Item / No. of Transactions

Step 2: Remove all the items whose support is below the given minimum support

|  |  |  |
| --- | --- | --- |
| Item | Frequency | Support (In %) |
| Bread | 4 | 4/5.5 = 72% |
| Cheese | 3 | 3/5.5 = 55% |
| Juice | 4 | 4/5.5 = 72% |
| Milk | 3 | 3/5.5 = 55% |

Step 3: Now form the two items candidate set and write their frequencies.

|  |  |  |
| --- | --- | --- |
| Item Pair | Frequency | Support (In %) |
| Bread, cheese | 2 | 36% |
| Bread, Juice | 3 | 55% |
| Bread, Milk | 2 | 36% |
| Cheese, Juice | 3 | 55% |
| Cheese, Milk | 1 | 18% |
| Juice, Milk | 2 | 36% |

Step 4: Remove all the items whose support is below the given minimum support.

|  |  |  |
| --- | --- | --- |
| Item Pair | Frequency | Support (In %) |
| Bread, Juice | 3 | 3/5.5 = 55% |
| Cheese, Juice | 3 | 3/5.5 = 55 % |

Step 5: Generate Rules

For rules we consider item pairs:

1. (Bread, Juice)

Bread -> Juice and Juice -> Bread

1. (Cheese -> Juice)

Cheese -> Juice and Juice -> Cheese

Confidence(A -> B) = Support(A U B) / Support(A)

Therefore,

1. Confidence (Bread -> Juice) = Support(Bread U Juice) / Support(Bread)

= 3/5.5 \* 5.5/4 = 75%

1. Confidence (Juice -> Bread) = Support(Juice U Bread) / Support(Juice)

= 3/5.5 \* 5.5/4 = 75%

1. Confidence (Cheese -> Juice) = Support(Cheese U Juice) / Support(Cheese)

= 3/5.5 \* 5.5/3 = 100%

1. Confidence (Juice -> Cheese) = Support(Juice U Cheese) / Support(Juice)

= 3/5.5 \* 5.5/4 = 75%

All the above rules are not equitable to the confidence of each rule. Only Rule 3 is greater than or equal to minimum confidence given in the problem.

**Task – 3: External Sources**

<https://docs.python.org/3/library/csv.html#csv.csvwriter.writerow>

<https://youtu.be/OeVH8fAfgmM>

<https://www.javatpoint.com/apriori-algorithm>

<https://stackabuse.com/association-rule-mining-via-apriori-algorithm-in-python/>

<https://github.com/asaini/Apriori>

<https://www.kaggle.com/acostasg/random-shopping-cart>