**Assignment 3 Report**

**Data Mining**

**CSE 5334 Section 001**

**Student Name:**

**Ahir Kunjal (1001877263)**

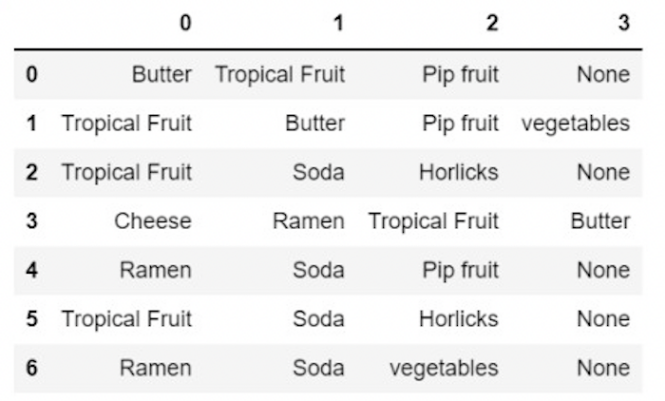
**Sheth Devanshi (1001959019)**

**Gope Pankaj (1001990547)**

**Professor: Dr Elizabeth D Diaz**

**Using Weka**

The screenshot of the small data set that we have formed is shown below:



The data set is about the groceries taken from grocery store.

The frequency of the items in the item sets is shown below:

|  |  |
| --- | --- |
| **Item** | **Frequency** |
| Tropical Fruit | 5 |
| Butter | 2 |
| Pip fruit | 3 |
| vegetables | 2 |
| Cheese | 1 |
| Soda | 4 |
| Ramen | 2 |
| Horlicks | 2 |

The support of an item is the frequency of that particular item in the transaction list.

We can find the commonly bought together items by setting a threshold on the support. If the support threshold for an item is 50% or 0.5 it means that the item has occurred in half of the total transactions of the dataset.

The maximum value of the support threshold can be 1 which means that the item exists in all the transactions. The support for every individual item is shown below:

83

33

50

33

16

66

33

33

If we keep the support as S1 = 0.1 and confidence as C1 = 0.7, then items Cheese and butter are selected as they have a support threshold of more than 0.6.

Now we will pair two of them together and find the combined support.

33

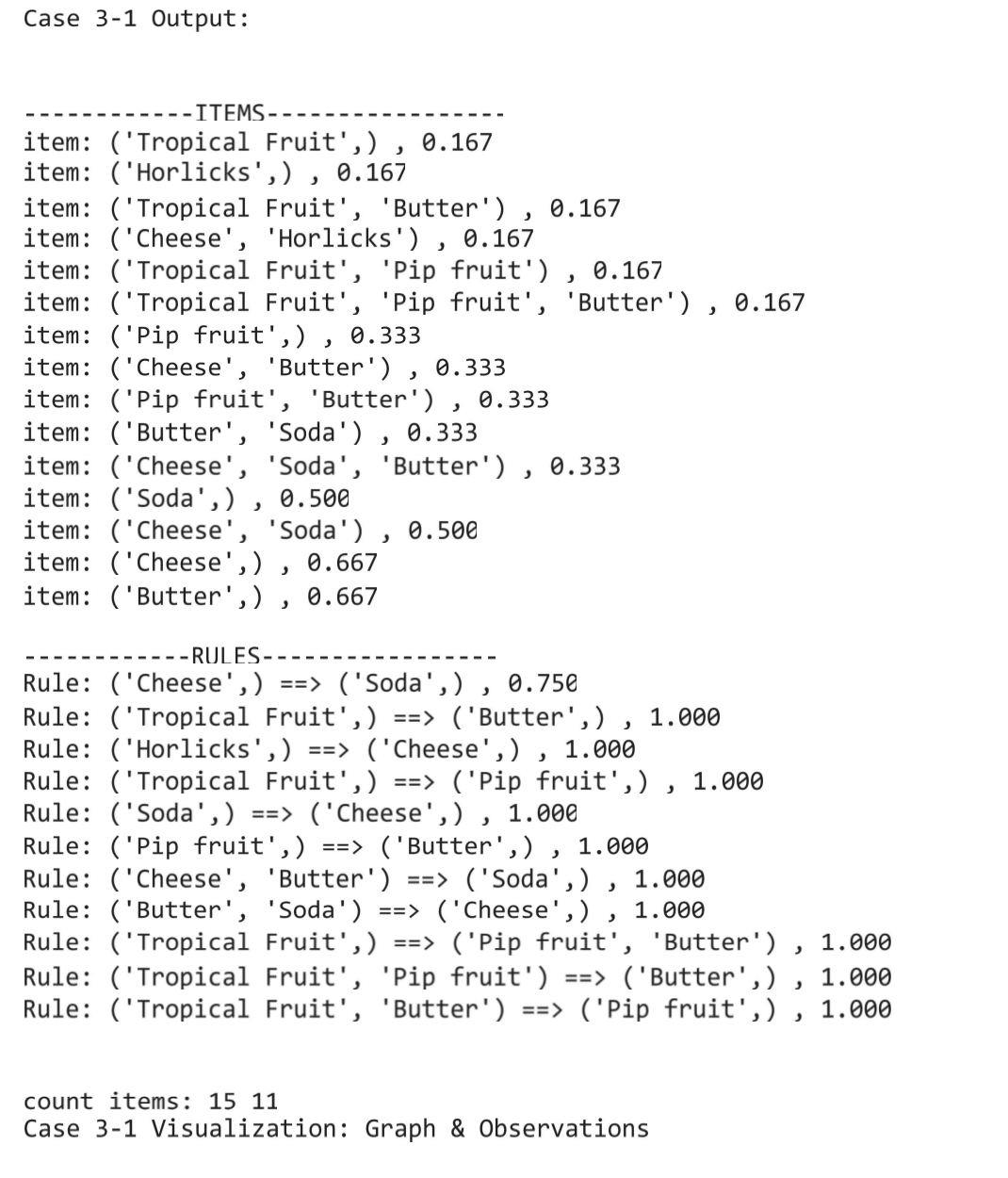
33

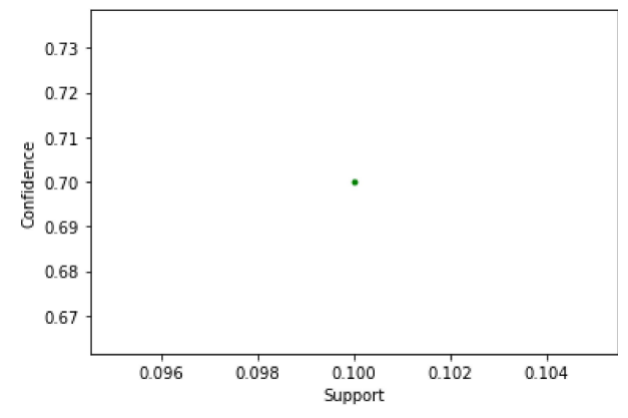
We can see that the support threshold of the cheese and butter pair and the butter and soda pair is less than the set threshold value of 0.6 hence they are not considered.

Hence, we conclude that only the items (cheese and butter) are frequent.

Now we will generate the association rule:

Since both the association rules are above the confidence level of 0.8, we can say that they are strong confidence rules.





Similarly solving for different value of support S2 = 0.4 and confidence C2 = 0.7

667

667

5

Since only the combination of (mobile cover, mobile and earphones) has support equal to 0.5.

Now we will generate the association rules.

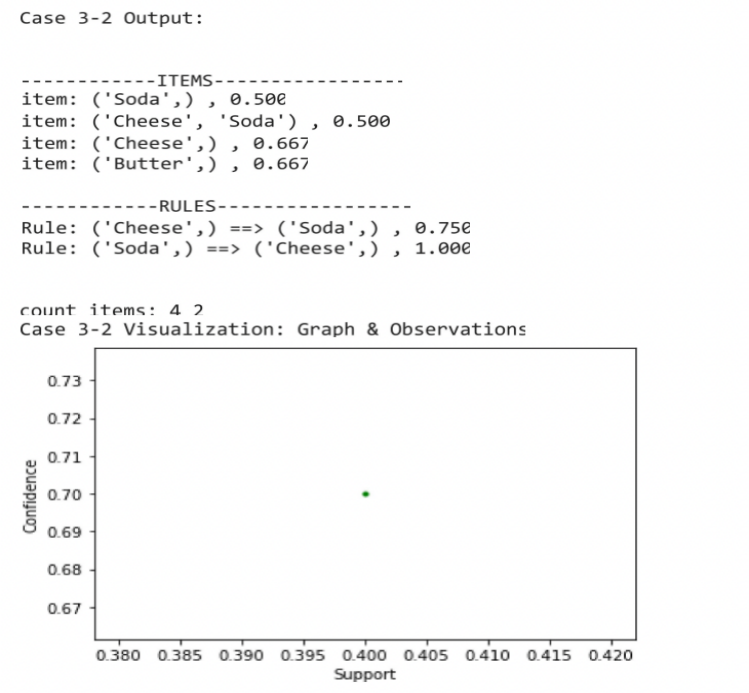
Rule 1:

0.75

Rule 2:

1.0

Hence, we can say that the Rule 2is the strong association rules.



Now we will take the value of support as S3 = 0.2 and confidence as C3 = 0.9.

33

33

33

667

667

5

Now we will generate the association rules.

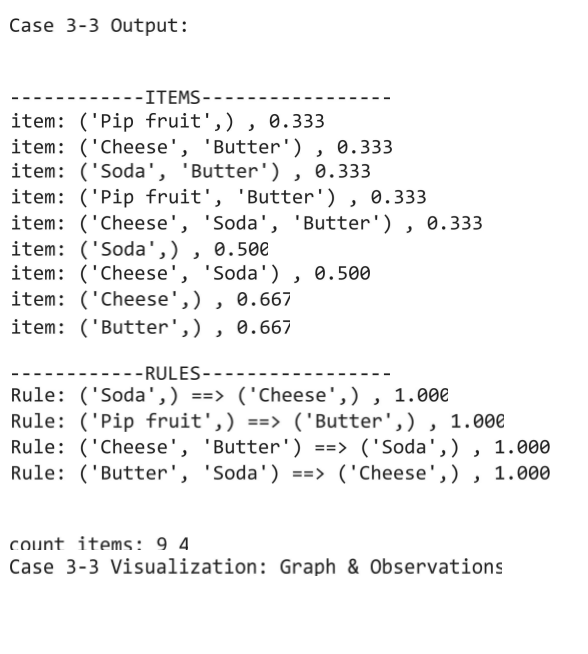
Rule 1:

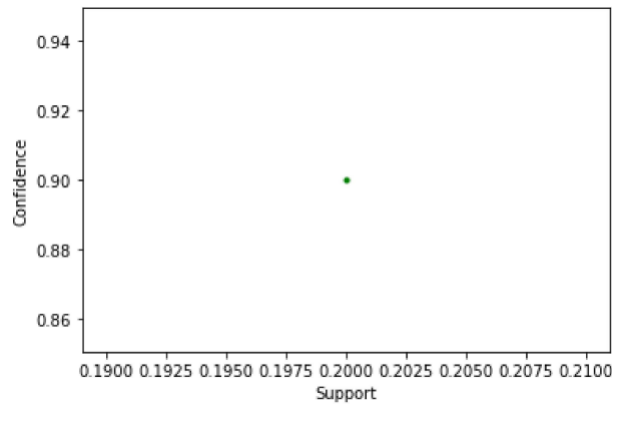
0.1

Rule 2:

1.0

Hence, we can say that the Rule 1 and 2 is the strong association rules.





Observations

The benefit of this strategy is that we can readily trim the goods that are not frequently purchased together. However, the difficulty with this is that it takes a long time to compute since it must scan over a full data collection. As a result, while dealing with bigger data sets, this approach becomes slightly more time intensive.

Here, we can observe that as the support threshold values declined, so did the number of frequent item sets and rules. This means that for a large data collection, we need a somewhat higher support threshold number to prevent needless rules from being produced as a result of low support threshold.

However, we cannot have a very high threshold number because this would result in certain critical association rules being disregarded.

References

* Introduction to Data Mining by Pang Ning, Michael Steinbach and Vipin Kumar
* <https://www.geeksforgeeks.org/apriori-algorithm/>
* <https://towardsdatascience.com/the-apriori-algorithm-5da3db9aea95>