Assignment :- 2 Course code: CAP446 Name: Pranav Mishra
Registration No:- 12114762 Section: D0C09
Roll No:- RD0C09A55 Date of submission: 02/12/2021

Ques1:- For the following given transaction dataset, generate rules using Apriori algorithm. Consider values as SUPPORT = 30% and CONFIDENCE = 60%.

TID	PRODUCTS			
1	Milk	Egg	Bread	Butter
2	Milk	Butter	Egg	Ketchup
3	Bread	Butter	Ketchup	
4	Milk	Bread	Butter	
5	Bread	Butter	Cookies	
6	Milk	Bread	Butter	Cookies
7	Milk	Cookies		
8	Milk	Bread	Butter	
9	Bread	Butter	Egg	Cookies
10	Milk	Butter	Bread	
11	Milk	Bread	Butter	
12	Milk	Bread	Cookies	Ketchup

Ans:- The APriori algorithm is used to generate rules to Predict the regularities in data.
In this Process, we will find the frequency of individual items in dataset and then we will extend it to larger sets as long as those itemself satisfies the given support and confidence.
The above dataset can be further splitted into the individual items to count the occurence of each item.
After counting the occurence, we will find their support and discard the items with lower support.

Given, Minimum Support = 30% and confidence = 60%.

Item	Frequency	Support
Milk	9	9/12 = 75%
Egg	3	3/12 = 25% ✗ discard
Bread	10	10/12 = 83.33%
Butter	10	10/12 = 83.33%
Ketchup	3	3/12 = 25% ✗ discard
Cookies	5	5/12 = 41%

In the above table, support is calculated using,

$$\text{Support} = \frac{\text{Frequency of data item}}{\text{Total no. of transaction}}$$

By finding the support for each data item, we find that the Egg and ketchup don't satisfy minimum support, hence, we will discard them.

Now we will find the following of item set of subset of 2 items.

Item sets	Frequency	Support
{Milk, Bread}	7	7/12 = 58.5%
{Milk, Butter}	7	7/12 = 58.5%
{Milk, Cookies}	3	3/12 = 25% (discard)
{Bread, Butter}	9	9/12 = 75%
{Bread, Cookies}	4	4/12 = 33%
{Butter, Cookies}	3	3/12 = 25% (discard)

Again, we have discarded {Milk, Cookies} and {Butter, Cookies} as it does not satisfy minimum support.

Now we will increase the itemset to subset of 3 items and find the frequency

Itemset	Frequency	support
{Milk, Bread, Butter}	6	6/12 = 50%.
{Milk, Bread, Cookies}	2	2/12 = 16.6%. → Discard
{Milk, Butter, Cookies}	1	1/12 = 8%. → Discard
{Bread, Butter, Cookies}	3	3/12 = 25%. → Discard

From the above table, we see that only one itemset satisfies the minimum support criteria, and discard all the others that do not satisfy.

The frequent itemset is - {Milk, Bread, Butter}

The subset that can be created are:

{{Milk}, {Bread}, {Butter}, {Milk, Bread}, {Milk, Butter}, {Bread, Butter}}.

Now we can generate rules on the basis of the frequent itemset that has been discovered.

formula used for confidence -

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

Association Rules:-

1. {Milk} → {Bread, Butter}

$$\text{Confidence} = \frac{\text{Supp.}(\text{Milk} \cup \text{Bread} \cup \text{Butter})}{\text{Supp.}(\text{Milk})}$$

$$= \frac{6/12}{9/12} = \frac{6}{12} \times \frac{12}{9} = \frac{6}{9} = 66.67\%$$

2. {Bread} → {Milk, Butter}

$$\text{Confidence} = \frac{\text{Supp.}(\text{Bread} \cup \text{Milk} \cup \text{Butter})}{\text{Support}(\text{Bread})}$$

$$= \frac{6/12}{10/12} = \frac{6}{12} \times \frac{12}{10} = \frac{6}{10} = 60\%$$

3. {Butter} → {Milk, Bread}

$$\text{Confidence} = \frac{\text{Supp.}(\text{Milk} \cup \text{Bread} \cup \text{Butter})}{\text{Support}(\text{Butter})}$$

$$= \frac{6/12}{10/12} = \frac{6}{12} \times \frac{12}{10} = \frac{6}{10} = 60\%$$

4. {Milk, Bread} → {Butter}

$$\text{Confidence} = \frac{\text{Supp.}(\text{Milk} \cup \text{Bread} \cup \text{Butter})}{\text{Support}(\text{Milk} \cup \text{Bread})}$$

$$= \frac{6/12}{7/12} = \frac{6}{12} \times \frac{12}{7} = \frac{6}{7} = 85.7\%$$

5. {Milk, Butter} → {Bread}

$$\text{Confidence} = \frac{\text{Supp.}(\text{Milk} \cup \text{Bread} \cup \text{Butter})}{\text{Support}(\text{Milk} \cup \text{Butter})}$$

$$= \frac{6/12}{7/12} = \frac{6}{12} \times \frac{12}{7} = \frac{6}{7} = 85.7\%$$

6. {Bread, Butter} → {Milk}

$$\text{Confidence} = \frac{\text{Supp.}(\text{Milk} \cup \text{Bread} \cup \text{Butter})}{\text{Support}(\text{Bread} \cup \text{Butter})}$$

$$= \frac{6/12}{9/12} = \frac{6}{12} \times \frac{12}{9} = \frac{6}{9} = 66.67\%.$$

Since, all the above rules have confidence greater than 60%, all the rules are strong.

Thus, these were some rules generated from the given dataset using Apriori algorithm.

Ques:2:- Explain the various steps involved in Data Preprocessing by taking helps of a data set gone through the process in Rapid Miner. Attach relevant screenshot.

Ans:- Data Preprocessing:- Data preprocessing is a data mining technique which is used to transform the raw data in a useful and efficient format.

There are a lot of inconsistent data like as incomplete data, missing attributes, errors, outliers, duplicacy etc to remove all this from the data set is called data pre processing.

Steps involved in Data Preprocessing:-

Step 1:- I used the read excel operator to read all the data in excel file.

Step 2:- Then, we use trim operator to remove the blank spaces from all data set.

Step 3:- After trim operator we used the replace missing operator to assign average value to missing attributes.

Step 4:- After replace mising operator, we used the select attributes to select some relevant attributes.

Step 5:- Then, I used the set role operator to set attribute 'churn' as label that change the role of attribute in data set.

Step 6:- "Remove duplicate" operator is used to remove duplicate values from the data set.

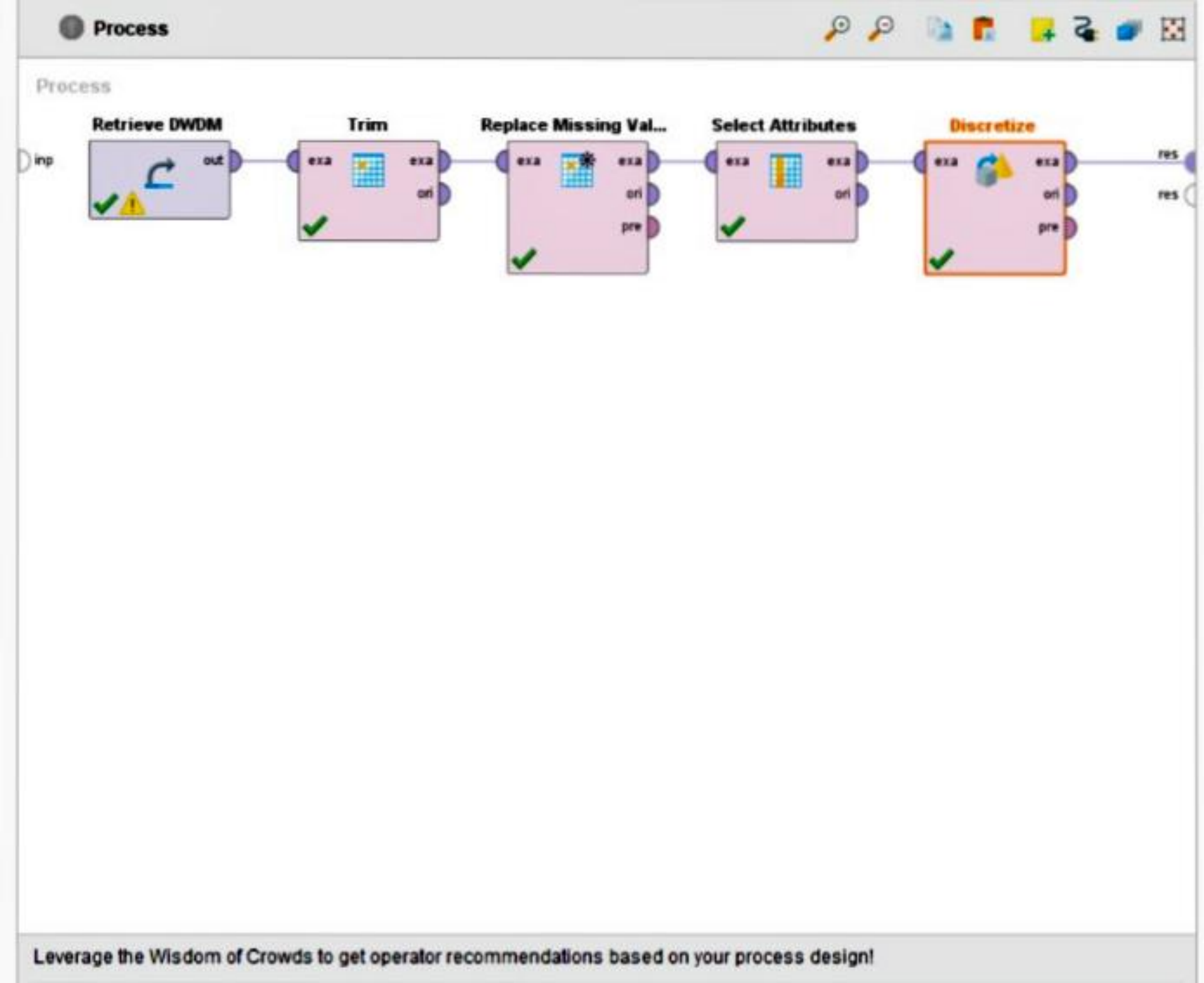
Trim operator :- I use trim operator to remove the blank spaces from ~~the~~ all the dataset.

Replace mising operator :- After the trim operator I use the replace mising value by some average, minimum, maximum value from the dataset.

Select attributes :- We use this operator to select only important and relevant attributes from dataset for a good understanding of data set.

Set role :- This operator is used to select one attributes as label and one of the basis of this dataset will give result.

Filter example :- This operator is used for remove the attributes that contains missing value.


Process
Process
Retrieve DWDM
Trim
Replace Missing Val...
Select Attributes
Discretize
inp
out
exa
ori
pre
res
Leverage the Wisdom of Crowds to get operator recommendations based on your process design!

Row No.	PRODUCT...	D	TID
1	Milk	Bread	1
2	Milk	Egg	2
3	bread	ketchup	3
4	Milk	Butter	4
5	Bread	Cookies	5
6	Milk	Butter	6
7	Milk	Butter	7
8	Milk	Butter	8
9	bread	Egg	9
10	Milk	Bread	10
11	Milk	Butter	11
12	Milk	Cookies	12

Ques:-3 Prepare FP Growth tree and all following transaction.
Given Minimum support=7

Transaction	Product
1	beer, wine, cheese
2	beer, Potato, chips
3	eggs, flour, butter, cheese
4.	eggs, flour, butter, beer, Potato chips
5.	wine, cheese
6.	Potato chips
7.	eggs, flour, butter, wine, cheese
8.	eggs, flour, butter, beer, Potatochips
9.	wine beer
10.	beer, Potato chips
11.	butter, eggs
12.	beer, Potatochips
13.	flour, eggs
14.	beer, Potato chips
15.	eggs, flour, butter, wine, cheese,
16.	beer, wine, Potato chips, cheese
17.	wine cheese.
18.	beer, Potato chips
19.	wine, cheese
20.	beer, Potato chips

Ans:- for preparing FP growth tree, first of all, we ~~know~~ have to find the occurrence or frequency of the itemsets.

Item	Frequency	Priority
Beer	11	1
Wine	8	3
Cheese	8	4
Potato chips	10	2
egg	7	5
flour	6	Discarded } Frequency is less than support
butter	6	Discarded

Here, we have given minimum support = 7, hence, we have discarded flour and butter.

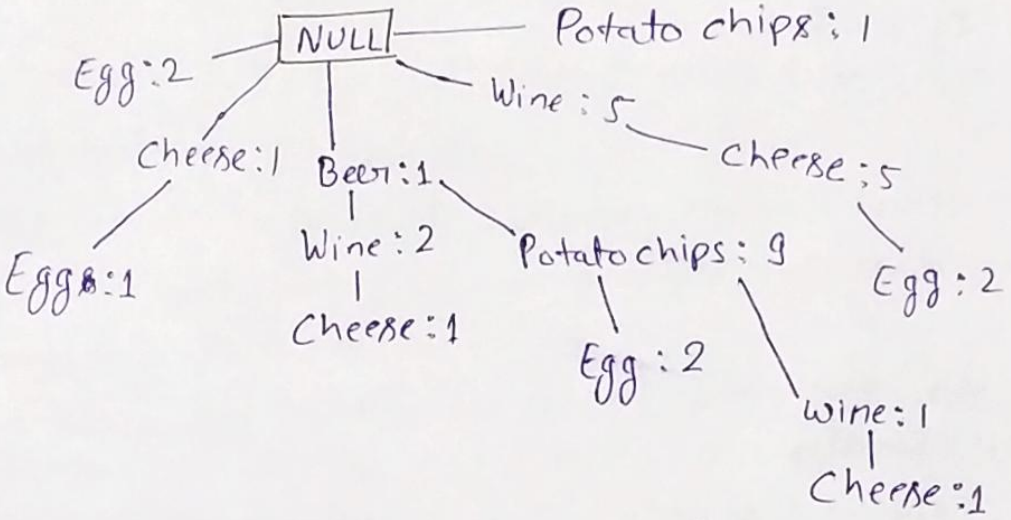
Now, we will arrange the data set with the Priority set above:-

Transaction No.	Ordered Products
1	Beer, wine, cheese
2	Beer, Potato chips
3	cheese, eggs
4.	Beer, Potato chips, egg
5.	wine, cheese
6	Potato chips
7	Wine, cheese, egg
8	Beer, Potato chips, egg
9	Beer, wine
10	Beer, Potatochips
11	eggs
12	Beer, Potato chips
13	egg
14	Beer, Potato chips
15	Wine, cheese, egg

Transaction no.	Ordered Products
16	Beer, Potato chips, wine, cheese
17	Wine, cheese
18	Beer, Potato chips
19	wine, cheese
20	Beer, Potato chips

On the basis of Prioritized ordered Products, we can create FP growth tree.

For tree, we have to create the root of tree first, which is represented by Null.



Here, In the above tree, we have Put the items with most frequency at the top, and they are in the descending order of the frequency.

We take the transactions one by one, and if any itemset of current tranaction is already Present in another branch, then this transaction branch will share the common item.

We also count the occurrence of each itemset in each iteration and increment it if it occurs in the transaction.

Now, the FP tree is created, we can mine the FP-tree, we will examine the lowest node first along with its links

From this, we can find conditional Pattern base, that is the traverse path.

Item	Conditional Pattern Base
Egg	{{cheese:1},{cheese, wine:2}{Potato chips, Beer:1}}
Cheese	{{wine, Beer:2}{wine, Potatochips, beer:1}, {wine:3}}
wine	{{Beer:2},{Potato chips, Beer:1}}
Potato chips	{{Beer:9}}
Beer	ϕ(as. it is directly connected with null)

Now, for each item, we will prepare the conditional frequent Pattern tree

It is done by taking the set of elements that is common in all the Paths in the conditional Pattern Base of that item. and calculating its support count by summing the support counts of all the Paths in the conditional Pattern Base.

Item	Conditional FP-tree
Egg	ϕ
Cheese	{wine:6} ×discarded
wine	ϕ
Potatochips	{Beer:9} ✓
Beer	ϕ

From the conditional FP tree, we can say that Beer is mostly associated with Potato chips and it also support the minimum support value.
Hence, the rule, of Potato chips with Beer will be considered valid and others will be discarded.