ASSOCIATION RULE MINING

1) OBJECTIVES:

ASSOCIATION RULE MINING:

Association rule-mining is a data mining approach used to explore and interpret large transactional datasets to identify unique patterns and rules. During transactions, these patterns define interesting relationships and interactions between different items.

DOMAIN OF ASSOCIATION RULE MINING:

The association rules are useful for analysing and predicting customer behaviour. They plan an important part in customer analysis, market based analysis, product clustering, catalog design and store layout.

BENEFITS FROM ASSOCIATION RULE MINING:

- Applying the algorithms to supermarkets, the scientists were able to discover links between different items purchased, called association rules, and ultimately use that information to predict the likelihood of different products being purchased together.
- For retailers, association rule mining offered a way to better understand customer purchase behaviours. Because of its retail origins, association rule mining is often referred to as market basket analysis.

USING PATTERNS IN ASSOCIATION RULE MINING:

Frequent Pattern Mining (aka Association Rule Mining) is an analytical process that finds frequent patterns, associations, or causal structures from data sets. Given a set of transactions, this process aims to find the rules that enable us to predict the occurrence of a specific item based on the occurrence of other items in the transaction.

2) DATASET:

The Online Retail dataset is downloaded from UCI machine learning repository. (https://archive.ics.uci.edu/ml/machine-learning-databases/00352/Online%20Retail.xlsx)

The dataset contains the following columns

- **InvoiceNo** : The 'InvoiceNo' contains the transaction ID or transaction number, which is used to uniquely identify the transactions.
- **StockCode** : The 'StockCode' contains item code or item ID which uniquely identifies the product/item.
- **Description** : The 'Description' contains the product names for the product ID specified in the StockCode.
- **Quantity** : The 'Quantity' contains the quantity for which the product was bought for that transaction.
- **InvoiceDate** : The 'InvoiceDate' contains the date when the product was bought or the transaction had happened.
- **UnitPrice** : The 'UnitPrice' contains the price of the product for unit quantity.
- **CustomerID** : The 'CustomerID' contains the ID which uniquely identifies the customer and it displays the customer who made that transaction.
- **Country** : The 'Country' contains the country where the transaction had happened.

PREPROCESSING:

- The actual dataset contains 541910 records (out of which, there are 45222 unique Invoices/Transactions and 4224 unique Items).
- This rule used the first 10000 records from the main dataset (out of which, there are 512 unique Invoices/Transactions and 1985 unique Items).
- The main fields in this dataset which are used for association mining are InvoiceNo and StockCode. So these fields have to be spliced out from the used dataset.
- It is found that the some transactions contained same products mentioned twice or thrice. So they have to be made unique (i.e each transaction has only unique items).

REASONING:

- This dataset will be perfect for association rule mining as retail datasets generally are preferred.
- The preprocessing is done to use only transaction ID and the items present in that transactions as the other fields are unnecessary for this mining process.
- It is assumed that the transaction should contain only unique products and their quantity is not considered for this association rule.

3) RULE MINING PROCESS:

The parameters considered for this rule mining process are

1. Support
2. Confidence

PARAMETER SETTING:

For the parameter setting process, the top most five frequently occurring items are found via the code snippet below.

```
frequentItems=simplifiedDataFrame[‘Description']
.value_counts()[:5].sort_values(ascending=False)
.to_dict()
```

The result is displayed below. The result contains the frequent items and their occurrence count.

```
{'HAND WARMER UNION JACK': 59,
'HAND WARMER SCOTTY DOG DESIGN': 58,
'WHITE HANGING HEART T-LIGHT HOLDER': 57,
'HAND WARMER OWL DESIGN': 57,
"PAPER CHAIN KIT 50'S CHRISTMAS": 52}
```

SUPPORT:

Support refers to the default popularity of an item and can be calculated by finding number of transactions containing a particular item divided by total number of transactions.

The minimum support used for this program is based on the support given by the above five frequent items which are shown below. The support is obtained by transactions containing particular item divided by total transactions.

```
The support of HAND WARMER UNION JACK = 0.105
The support of HAND WARMER SCOTTY DOG DESIGN
= 0.094
The support of HAND WARMER OWL DESIGN = 0.104
The support of WHITE HANGING HEART T-LIGHT
HOLDER = 0.107
The support of PAPER CHAIN KIT 50'S CHRISTMAS
= 0.092
```

CONFIDENCE:

Confidence refers to the likelihood that an item B is also bought if item A is bought. It can be calculated by finding the number of transactions where A and B are bought together, divided by total number of transactions where A is bought.

The minimum confidence is kept based on the confidence found out between top two frequent items (i.e HAND WARMER UNION JACK and HAND WARMER SCOTTY DOG DESIGN) which is shown below.

```
confidenceOfA_B =
round(noOfTransactionsContainingAandB/
noOfTransactionsContainingItemA,3)
print("The confidence of HAND WARMER UNION
JACK and HAND WARMER SCOTTY DOG DESIGN = " +
str(confidenceOfA_B))

The confidence of HAND WARMER UNION JACK and
HAND WARMER SCOTTY DOG DESIGN = 0.407
```

There is also a third parameter called LIST. For this program, the list parameter is not considered. Hence by default it is made one for all the rule settings.

CHOICE OF ALGORITHM:

The algorithm chosen for this association mining for the retail dataset is **Apriori algorithm.**

Apriori is an algorithm for frequent item set mining and association rule learning over relational databases. Apriori is designed to operate on databases containing transaction data. Apriori uses the following steps.

1. Set a minimum value for support and confidence. This means that we are only interested in finding rules for the items that have certain default existence (e.g. support) and have a minimum value for co-occurrence with other items (e.g. confidence).
2. Extract all the subsets having higher value of support than minimum threshold.
3. Select all the rules from the subsets with confidence value higher than minimum threshold.
4. Order the rules by descending order of Lift.

TIME REQUIRED:

Based on the resulting rules set, the time taken to produce the result varies largely when the apriori algorithm is used.

EXPERIMENT:

There are five iterations/experiments made for this airport algorithm with different support and confidence values resulting in different rule sets.

For all of these five iterations, the lift value is kept as 1 because only support and confidence is considered.

4) RESULTING RULES:

ITERATION-1:

The minimum support = 0.001
The minimum confidence = 0.1
The minimum length = 2

The result size is 467138 and the last rule in the result set is shown below.

```
Rule: ZINC WILLIE WINKIE CANDLE STICK -> ZINC
METAL HEART DECORATION
Support: 0.001953125
Confidence: 0.125
Lift: 4.923076923076923
```

ITERATION-2:

The minimum support = 0.01
The minimum confidence = 0.15
The minimum length = 2

The result size is 108729 and the last rule in the result set is shown below.

```
Rule: 10 COLOUR SPACEBOY PEN -> HAND WARMER
BIRD DESIGN
Support: 0.01171875
Confidence: 0.46153846153846156
Lift: 5.495527728085868
```

ITERATION-3:

The minimum support = 0.02
The minimum confidence = 0.2
The minimum length = 2

The result size is 108729 and the last rule in the result set is shown below.

```
Rule: WOOD 2 DRAWER CABINET WHITE FINISH -> 
WHITE METAL LANTERN
Support: 0.021484375
Confidence: 0.4583333333333333
Lift: 21.333333333333332
```

ITERATION-4:

The minimum support = 0.05
The minimum confidence = 0.5
The minimum length = 2

The result size is 6 and all the six rules in the result set are shown below.

```
Rule: HAND WARMER BIRD DESIGN -> HAND WARMER 
OWL DESIGN
Support: 0.052734375
Confidence: 0.627906976744186
Lift: 6.065818341377797
============================================
Rule: HAND WARMER RED RETROSPOT -> HAND WARMER 
OWL DESIGN
Support: 0.0546875
Confidence: 0.5283018867924528
Lift: 6.440251572327044
============================================
```

```
Rule: HAND WARMER SCOTTY DOG DESIGN -> HAND
WARMER OWL DESIGN
Support: 0.068359375
Confidence: 0.660377358490566
Lift: 7.044025157232704
============================================
Rule: KNITTED UNION FLAG HOT WATER BOTTLE ->
RED WOOLLY HOTTIE WHITE HEART
Support: 0.05859375
Confidence: 0.7317073170731707
Lift: 8.9198606271777
============================================
Rule: WHITE HANGING HEART T-LIGHT HOLDER ->
KNITTED UNION FLAG HOT WATER BOTTLE
Support: 0.05078125
Confidence: 0.6341463414634146
Lift: 5.903325942350333
============================================
Rule: PAPER CHAIN KIT VINTAGE CHRISTMAS ->
PAPER CHAIN KIT 50'S CHRISTMAS
Support: 0.0546875
Confidence: 0.5957446808510638
Lift: 6.932301740812378
============================================
```

ITERATION-5:

The minimum support = 0.065
The minimum confidence = 0.5
The minimum length = 2

The result size is 1 and the rule is shown below.

```
Rule: HAND WARMER SCOTTY DOG DESIGN -> HAND
WARMER OWL DESIGN
Support: 0.068359375
Confidence: 0.660377358490566
Lift: 7.044025157232704
```

SELECTION OF RULES:

For clients, the rules obtained in iteration-4 and iteration-5 should be shown (i.e 6+1=7 rules) because they give the minimum number of rules (only 7 from almost 5 lakh possible rule combinations) with the minimum support and confidence.

5) RECOMMENDATIONS:

Based on the rules presented to the client, the client should use the rules to organise the items present in the rule together which indicates that these two items are bought together more frequently. By keeping these two items together, the number of transactions can be increased thereby increasing profit for the client.

———————————————————————————————————

The dataset can be found via the below link

https://github.com/KaranrajMokan/College-works/blob/main/Data%20Mining/Worksheet-2/Online%20Retail.xlsx

The working python file can be found via the below link

https://github.com/KaranrajMokan/College-works/blob/main/Data%20Mining/Worksheet-2/ws2.py

———————————————————————————————————

- By
Karanraj M
17PW18