Ans to the Question no 1

Apriori Algorithm usually contains or deals with a large number of transactions. For example, customers buy a lot of goods from a grocery store. By applying this algorithm method, grocery stores can enhance their sales performance and work effectively. Can you consider online grocery store-related datasets as an example? and explain the Apriori Algorithm based on your datasets. [Theoretical]

Ans:

Consider the following database of a grocery store:

Items of the database:

itemset
(1,2,3,4)
(1,2,4)
(1,2)
(2,3,4)
(2,3)
(3,4)
(2,4)

Here we will try to appear at least three transactions with the value of 3, which will support the threshold.

By scanning the database for the first time, we obtain the following result:

item	support
(1)	3
(2)	6
(3)	4
(4)	5

All the item sets of the size 1 have the support of at least 3, so they are all frequent:

item	Support
(1,2)	3
(1,3)	1
(1,4)	2
(2,3)	3
(2,4)	4
(3,4)	3

The pairs (1,2), (2,3), (2,4), and (3,4) all meet or exceed the minimum support of 3, so they are frequent.
The pairs (1,3) and (1,4) are not. Now because (1,3) and (1,4) are not frequent, any large set which contains (1,3) or (1,4) cannot be frequent.

Item	support
(2,3,4)	2

Ans to the Question no 4

Explain the Following terms in short:
(i)Cross-Validation (ii)Feature selection

Ans:
  i.   Cross-Validation:
       Cross-Validation is a technique that involves reserving a particular sample of a dataset on which we do not train the model. Later, we test our model on this sample before finalizing it.

   Here are the steps involved in cross-validation:
   - We reserve a sample data set
   - Train the model using the remaining part of the dataset

- Use the reserve sample of the test (validation) set. This will help us in gauging the effectiveness of our model's performance. If our model delivers a positive result on validation data, go ahead with the current model. It rocks!

ii.   Feature selection:

Feature selection is a process where we automatically select those features in our data that contribute most to the prediction variable or output in which we are interested. Having irrelevant features in our data can decrease the accuracy of many models, especially linear algorithms like linear and logistic regression.

Three benefits of performing feature selection before modeling your data are:
- Reduces Overfitting: Less redundant data means less opportunity to make decisions based on noise.
- Improves Accuracy: Less misleading data means modeling accuracy improves.
- Reduces Training Time: fewer data means that algorithms train faster.