Name: Zeenat **Class: TYIT** **Roll no: 578**

Assignment –1

1. Find all frequent itemsets using Apriori algorithm. Min support count = 2

cust_ID	*TID*	*items_bought* (in the form of *brand-item_category*)
01	T100	{King's-Crab, Sunset-Milk, Dairyland-Cheese, Best-Bread}
02	T200	{Best-Cheese, Dairyland-Milk, Goldenfarm-Apple, Tasty-Pie, Wonder-Bread}
01	T300	{Westcoast-Apple, Dairyland-Milk, Wonder-Bread, Tasty-Pie}
03	T400	{Wonder-Bread, Sunset-Milk, Dairyland-Cheese}

Solution:

C1

Itemset	Support count
{King's-Crab}	1
{Sunset-Milk}	2
{Dairyland-Cheese}	2
{Best-Bread}	1
{Best-Cheese}	1
{Goldenfarm-Apple}	1
{Tasty-Pie}	2
{Wonder-Bread}	2

Compare candidate support count with minimum support count

L1

Itemset	Support count
{Sunset-Milk}	2
{Dairyland-Cheese}	2
{Tasty-Pie}	2
{Wonder-Bread}	2

Name: Zeenat **Class: TYIT** **Roll no: 578**

Generate C2 candidates from L1

C2

Itemset
{Sunset-Milk, Dairyland-Cheese}
{Sunset-Milk, Tasty-Pie}
{Sunset-Milk, Wonder-Bread}
{Dairyland-Cheese, Tasty-Pie}
{Dairyland-Cheese, Wonder-Bread}
{Tasty-Pie, Wonder-Bread}

Second for count of each candidate

Itemset	Support count
{Sunset-Milk, Dairyland-Cheese}	2
{Sunset-Milk, Tasty-Pie}	0
{Sunset-Milk, Wonder-Bread}	1
{Dairyland-Cheese, Tasty-Pie}	0
{Dairyland-Cheese, Wonder-Bread}	1
{Tasty-Pie, Wonder-Bread}	1

Compare candidate support count with minimum support count

L2

Itemset	Support count
{Sunset-Milk Dairyland-Cheese}	2

Ans: Generotion of candidate itemset and Frequent itemsets, can be

The minimum support count is 2 using Apriori algorithm

Name: Zeenat **Class: TYIT** **Roll no: 578**

2. Draw a STAR schema for sales in a manufacturing company with SALESREP, CUSTOMER, PRODUCT, and TIME

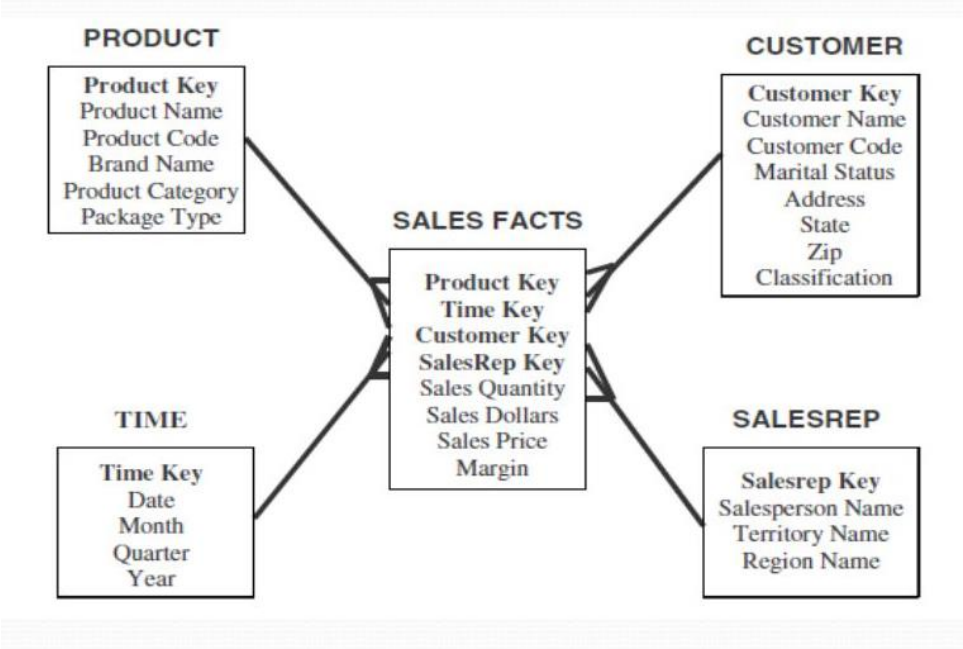
Ans:

STAR schema:

A star schema is a database organizational structure optimized for use in a data warehouse or business intelligence that uses a single large fact table to store transactional or measured data, and one or more smaller dimensional tables that store attributes about the data. Sales price, sale quantity, distant, speed, weight, and weight measurements are few examples of fact data in star schema.

The sales fact table include quantity, price, and other relevant metrics. SALESREP, CUSTOMER, PRODUCT, and TIME are the dimension tables.

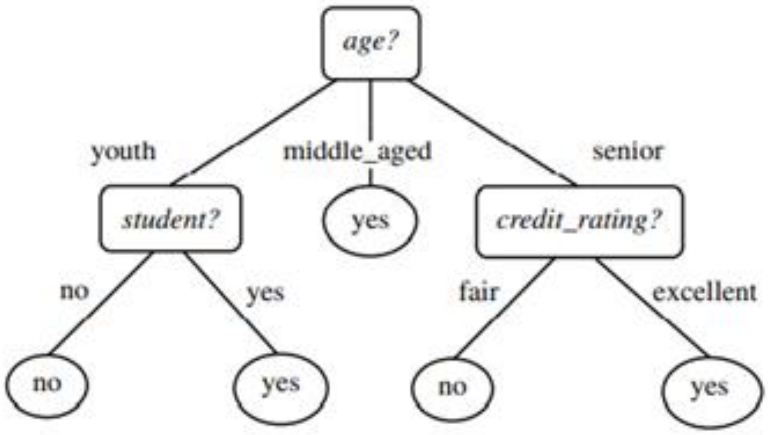
Star Schema



Name: Zeenat **Class: TYIT** **Roll no: 578**

Assignment – 2

1. Extract classification rules from a given decision tree



Ans:

Classification Rules:

Rule 1: IF age = youth AND student = no THEN buys_computer = no

Rule 2: IF age = youth AND student = yes THEN buys_computer = yes

Rule 3: IF age = middle_aged THEN buys_computer = yes

Rule 4: IF age = senior AND credit_rating = excellent THEN buys_computer = yes

Rule 5: IF age = senior AND credit_rating = fair THEN buys_computer = no

2. Explain Tree Pruning

Ans:

1. The process of adjusting Decision Tree to minimize “misclassification error” is called pruning.
2. Pruning is the method of removing the unused branches from the decision tree.
3. Some branches of the decision tree might represent outliers or noisy data.
4. Tree pruning is the method to reduce the unwanted branches of the tree.

5. This will reduce the complexity of the tree and help in effective predictive analysis.
6. It reduces the overfitting as it removes the unimportant branches from the trees.

It is of 2 types prepruning and post pruning.

1) Prepruning:

In this approach, the construction of the decision tree is stopped early. It means it is decided not to further partition the branches. The last node constructed becomes the leaf node and this leaf node may hold the most frequent class among the tuples.

The attribute selection measures are used to find out the weightage of the split. Threshold values are prescribed to decide which splits are regarded as useful. If the portioning of the node results in splitting by falling below threshold, then the process is halted.

2) Postpruning:

This method removes the outlier branches from a fully grown tree. The unwanted branches are removed and replaced by a leaf node denoting the most frequent class label. This technique requires more computation than prepruning, however, it is more reliable.

The pruned trees are more precise and compact when compared to unpruned trees but they carry a disadvantage of replication and repetition.

Repetition occurs when the same attribute is tested again and again along a branch of a tree. *Replication* occurs when the duplicate subtrees are present within the tree. These issues can be solved by multivariate splits.