CMR
INSTITUTE OF TECHNOLOGY

Sub:	Data Mining and Business Intelligence							Sub Code:	22MCA252
Date:	30/08/'23	Duration:	90 min's	Max Marks:	50	Sem:	II	Branch :	MCA

Note : Answer FIVE FULL Questions, choosing ONE full question from each modul

SCHEME

1 What are the various types of attributes? Explain Mean, Median, Mode, Variance, and Standard Deviation with example. [10]

- Types of attributes – 6m
- Each 1 m

2 Explain Data Generalization based on Characterization. [10]

- Concept description(4)
- Data cube(3m)
- Attribute oriented induction(3m)

3 What is Data Cleaning? Discuss various ways of handling missing values and noisy data during data cleaning.[10]

- Fill in missing values(3m)
- Identify outliers and smooth out noisy data(3m)
- Correct inconsistent data(3m)
- Resolve redundancy caused by data integration(1m)

4 Explain Data Reduction in detail. [10]

- Data Cube Aggregation(2m)
- Dimensionality reduction(2m)
- Numerosity Reduction(2m)
- Data compression(2m)
- Discretization and concept hierarchy generation(2m)

5 Explain Normalization by Min-Max score, z-score, and Decimal scaling with examples.[10]

- Min-Max score (3m)
- z-score (4 m)
- Decimal scaling (3 m)

6 Explain how correlation analysis can be performed in Numeric and Categorical data.[10]

- Chi Square (5m)
- Pearson Correlation(5m)

7 What are the various Data Compression techniques? [10]

- Lossy(5m)
- Loss Less(5m)

8 Explain Association Rule Mining (ARM) ? How Support and Confidence are used for evaluation? [10]

- ARM(4m)
- Support(2m)
- Confidence(2m)

- Example(2m)

9 Explain Apirori Algorithm with a suitable example. [10]

- Algorithm(3m)
- Explanation(2 m)
- Eample(5m)

10 Explain and construct FP Tree for given Transaction data:[10]

TId	Items
1	A B
2	B C D
3	A C D E
4	A D E
5	A B C
6	A B C D
7	B C
8	A B C
9	A B D
10	B C E

- Method(2m)
- Steps(4m)
- Solution(4m)

SOLUTION

1 What are the various types of attributes? Explain Mean, Median, Mode, Variance, and Standard Deviation with example. [10]

- Attribute Types
- An attribute is a **property of the object.**
- It also represents different **features of the object**.
  - **E.g.** Person → Name, Age, Qualification etc.
- Attribute types can be divided into four categories.
  - Nominal
  - Ordinal
  - Interval
  - Ratio
- 1) Nominal Attribute
- Nominal attributes are **named** attributes which can be **separated into discrete (individual) categories** which do not overlap.
- Nominal attributes values also called as **distinct values**.
- Example
  - 2) Ordinal Attribute
- Ordinal attribute is the **order of the values,** that's important and significant, but the differences between each one is not really known.
- Example
  - **Rankings** → 1^{st}, 2^{nd}, 3^{rd}
  - **Ratings** →
- We know that a 5 star is better than a 2 star or 3 star, but we don't know and cannot quantify–how much better it is?
  - 3) Interval Attribute

- Interval attribute comes in the form of a **numerical value** where the **difference** between points is **meaningful**.
- Example
  - **Temperature →** 10°-20°, 30°-50°, 35°-45°
  - Zero degree does not mean there is no temperature but it is very cold.
- We can not find true zero (absolute) value with interval attributes.
- 4) Ratio Attribute
- Ratio attribute is looks **like interval attribute**, but it **must have** a **true zero (absolute)** value.
- It tells us about the order and the exact value between units or data.
- Example
  - **Age Group →** 10-20, 30-50, 35-45 (In years)(can be zero cannot be below zero)
  - **Mass →** 20-30 kg, 10-15 kg
- It does have a true zero (absolute) so, it is possible to compute ratios.
- **Properties of Attribute Values**

The type of an attribute depends on which of the following properties it possesses:
–**Distinctness**: = , <>
–**Order**: < >
–**Addition**: + -
–**Multiplication**: * /
–Nominal attribute: distinctness
–Ordinal attribute: distinctness & order
–Interval attribute: distinctness, order & addition
–Ratio attribute: all 4 properties

- **Discrete and Continuous Attributes**

•Discrete Attribute
–Has only a finite or countably infinite set of values
–Examples: zip codes, counts, or the set of words in a collection of documents
–Often represented as integer variables.
–Note: binary attributes are a special case of discrete attributes
•Continuous Attribute
–Has real numbers as attribute values
–Examples: temperature, height, or weight.
–Practically, real values can only be measured and represented using a finite number of digits.
–Continuous attributes are typically represented as floating-point variables.

Mean

- Mean is the **average** of a dataset.
- To find the mean, calculate the sum of all the data and then divide by the total number of data.
- Example
  - ✔ Find out mean for **12, 15, 11, 11, 7, 13**

Median

- Median is the middle number in a dataset when the data is arranged in numerical order (Sorted Order).
- **Mode**
- **The mode is the number that occurs most often within a set of numbers.**
- **Example**

2 Explain Data Generalization based on Characterization. [10]

- Concept description
  (summarization and generalization)

- Data mining can be classified into two categories:
  - ✔ **descriptive** data mining
  - ✔ **predictive** data mining.
- Descriptive data mining **describes the data set** in a **concise and summative manner** and **presents interesting general properties** of the data. Eg. Graph/associative rules/clustering
- Predictive data mining **analyzes the data** in order to **construct one or a set of models**, and attempts to **predict the behavior of new data sets**. Eg. Classification, regression
- Database is usually storing the large amounts of data in great detail. However users often **like to view** sets of **summarized data in concise, descriptive terms**.

Data Generalization

There are two main approaches to data generalization − the **data cube approach** and **attribute orientation induction**.

Data Cube Approach:

- The data cube approach is a method of data generalization that involves creating a multi-dimensional data structure, known as a data cube, to represent the data.
- The data cube is formed by aggregating the data along different dimensions or attributes, such as time, location, or product type.
- This allows users to easily slice and dice the data to view and analyze it from different perspectives.
- One of the main benefits of the data cube approach is that it allows users to quickly and easily perform ad-hoc queries and drill down into the data to identify patterns and trends.
- It is particularly well-suited for use in data warehousing and business intelligence applications.
- Data Characterisation

Attribute Orientation Induction

- Attribute orientation induction is a method of data generalization that involves identifying and representing patterns in the data by creating a set of rules or conditions known as attribute orientations.
- These orientations are used to classify data points into different groups or categories based on their attributes or characteristics.
- One of the main benefits of attribute orientation induction is that it allows users to identify and represent complex patterns in the data in a more simplified form.
- It is particularly well-suited for use in machine learning and data mining applications.

3 What is Data Cleaning? Discuss various ways of handling missing values and noisy data during data cleaning.[10]

- I) Data Cleaning

1. **Fill in missing values**
   1. Ignore the tuple
   2. Fill missing value manually
   3. Fill in the missing value automatically
   4. Use a global constant to fill in the missing value
2. **Identify outliers and smooth out noisy data**
   1. Binning Method
   2. Clustering
3. **Correct inconsistent data**
4. **Resolve redundancy caused by data integration**

- 1) Fill missing values
- **Ignore the tuple (record/row)**:
  - Usually done when **class label is missing**.
- **Fill missing value manually**:

- Use the **attribute mean (average)** to **fill in the missing value** and **also use the attribute mean (average)** for **all samples belonging to the same class**.

- **Fill in the missing value automatically**:
  - **Predict** the **missing value** by using a **learning algorithm**:
    - Consider the attribute with the missing value as a dependent variable and run a learning algorithm (usually regression,Naive Bayes or Decision tree) to predict the missing value.
- **Use a global constant to fill in the missing value**
  - Replace **all missing attribute values** by the same constant such as a label like **"*Unknown*"** or NAN
- 2) Identify outliers and smooth out noisy data

a. **Binning method**
b. **Clustering**
c. **Regression**

- a) Binning method
- Data binning or **bucketing** is a data pre-processing technique used to **reduce the effects of minor observation errors**.
- The original data values which fall in a given small interval called **as a bin** are **replaced by a value which represents that interval**, often called the central value.
- **Steps of Binning method**
  1. **Sort the attribute values** and **partition** them into **bins**.
  2. Then smooth by **bin means**, **bin median** or **bin boundaries**.
- Binning method - Example
- Given data: **4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34**
- Step: 1
- Partition into **equal-depth [n=4]**:

Bin 1: 4, 8, 9, 15
Bin 2: 21, 21, 24, 25
Bin 3: 26, 28, 29, 34

- Step: 2
  - Smoothing by **bin means**:

  Bin 1: 9, 9, 9, 9
  Bin 2: 23, 23, 23, 23
  Bin 3: 29, 29, 29, 29

- Binning method - Example (Cont..)
- Given data: **4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34**
- Step: 1
- Partition into **equal-depth [n=4]**:

Bin 1: 4, 8, 9, 15
Bin 2: 21, 21, 24, 25
Bin 3: 26, 28, 29, 34

- Step: 2
  - Smoothing by **bin boundaries**:

  Bin 1: **4**, **4**, **4**, **15**
  Bin 2: **21**, **21**, **25**, **25**
  Bin 3: **26**, **26**, **26**, **34**

- Binning method (Cont..)
- Binning method is a **top-down splitting technique** based on a **specified number of bins**.
- It is also used as **discretization method** for data reduction and concept hierarchy generation.

- For example, attribute values can be discretized (separated) by applying equal-width or equal-frequency binning, and then replacing each value by the bin mean or median.
- It can be applied **recursively to the resulting partitions** to **generate concept hierarchies**.
- It **does not use class information**, therefore it is an **unsupervised discretization technique**.
- b) Clustering
- Clustering is a process of **partitioning a set of data** (or objects) into a **set of meaningful sub-classes**, called clusters.
- It enables the abstraction of **large amounts data** by forming **meaningful groups or categories of objects**.
- In clustering, objects in the same cluster are similar to each other and those in different clusters are dissimilar.
- Outliers may be detected by clustering.
- **Example**
  - Library (Group of Books based on different categories)
  - Cloths (By size S, M, L, XL, XXL etc.)
- Cluster Analysis
- C)Regression
- **Fitting** the data into function
- 2 attributes best fit in **linear regression**
- More than 2 attributes fit into a multi-dimensional surface.
- Regression
- 3) Correct inconsistent data
- If you have inconsistencies in your data, it can cause major problems later on.
- But with larger datasets, it can be difficult to find all of the inconsistencies.
- **It contains similarity in codes or names.**
- We can manually solve common mistakes like spelling, grammar, articles or use other tools for it.
- 4) Resolve redundancy caused by data integration
- Data redundancy occurs in database systems **which have a field that is repeated in two or more tables**.
- When customer data is duplicated and attached with each product bought, then redundancy of data is known as **inconsistency**.
- So, the entity "customer" **might appear with different values**.
- Database **normalization** prevents redundancy and makes the best possible usage of storage.
- The proper use of **foreign keys** can minimize data redundancy and reduce the chance of destructive anomalies appearing.

4 Explain Data Reduction in detail. [10]

- Data Reduction
- **Data reduction Strategies:**
  - ✔ **Data Cube Aggregation**
  - ✔ **Dimensionality reduction:** It is the process of reducing the number of random variables or attributes.
    - Attribute Subset Selection
    - Decision tree Induction
    - Wavelet Transformation
    - Principal Component Analysis (PCA)
  - ✔ **Numerosity Reduction:** This technique replaces the volume of data by alternative smaller forms of data representation.

- Parametric methods- regression
- Non Parametric methods- Histograms, sampling, data cube aggregation

✔ **Data compression:**
- Lossless data compression
- Lossy data Compression

✔ **Discretization and concept hierarchy generation**

- Data Cube Aggregation
- Multiple levels of aggregation in data cubes

✔ Further reduce the size of data to deal with

- Reference appropriate levels

✔ Use the smallest representation capable to solve the task

- Queries regarding aggregated information should be answered using data cube.
- Dimensionality Reduction

Attribute subset selection

- **Feature selection** :

✔ Select a minimum set of features such that the probability distribution of different classes given the values for those features is as close as possible to the original distribution given the values of all features.

✔ Heuristic method for feature selection
- Step-wise **forward selection**-select only required attributes
- Step-wise **backward elimination**-select all but eliminate unwanted attributes
- Combining forward selection and backward elimination

- **Decision-tree induction**

✔ A Decision tree is a flowchart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label.

- **Wavelet transformation (Video class)**
- **PCA-Principal Component Analysis (Video class)**

✔ Numerosity Reduction

- **Parametric methods**

✔ Assume the data fits some model, estimate model parameters, store only the parameters, and discard the data (except possible outliers). If we are fitting the data in an equation say y=mx+c then we need not store y data. Eg. Regression

- **Non-parametric methods**

✔ Do not assume models

✔ Using histograms, clustering, sampling, Binning

5 Explain Normalization by Min-Max score, z-score, and Decimal scaling with examples.[10]

- Data Transformation (Cont..)

4. **Normalization**

Normalization is **scaling technique** or a **mapping technique**.
With normalization, we can find **new range from an existing range**.
There are three techniques for normalization.
Min-Max Normalization

This is a simple normalization technique in which we fit given data in a pre-defined boundary, or a pre-defined interval [0,1].

Normalization by Decimal scaling

In this technique we move the decimal point of values of the attribute.

Z-score Normalization

In this technique we move the decimal point of values of the attribute.

1) Min-max normalization

Min max is a technique that helps to **normalizing the data**.

It will **scale the data between 0 and 1**.

Example

$$\text{Formula}: V' = \frac{v - Min_A}{Max_A - MinA}(NewMax_A - NewMin_A) + NewMin_A$$

2) Decimal scaling

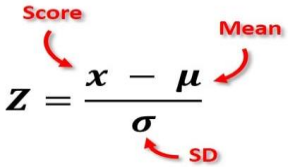
In this technique we move the decimal point of values of the attribute.

This movement of decimal points totally depends on the **maximum value among all values** in the attribute.

Value V of attribute A can be normalized by the following formula

Normalized value of attribute

3) z-Score



$$Z = \frac{x - \mu}{\sigma}$$

Eg.

The formula for calculating a z-score is z = (x-μ)/σ, where x is the raw score, μ is the population mean, and σ is the population standard deviation.

z-score is equal to 0, it is on the mean.

A positive z-score indicates the raw score is higher than the mean average. For example, if a z-score is equal to +1, it is 1 standard deviation above the mean.

A negative z-score reveals the raw score is below the mean average. For example, if a z-score is equal to -2, it is two standard deviations below the mean.

Example,

Given the Data, Calculate the Mean and SD for the data.

Mean=40, SD=10.

Note: Z-score more than +3 or -3 is known as an outlier.

6 Explain how correlation analysis can be performed in Numeric and Categorical data.[10]

- **Redundant data may be able to be detected by correlational analysis.(for numerical data)**

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Where,

r = Pearson Correlation Coefficient

x_i = x variable samples

y_i = y variable sample

$\bar{x}$ = mean of values in x variable

$\bar{y}$ =mean of values in y variable

$r_{x,y}$ =0 means no correlation

$r_{x,y}$ >0 means Positive correlation

$r_{x,y}$ <0 means negative correlation

- **Correlational analysis for nominal/categorical data**

Chi-Square: $\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$

	male	female	Total
fiction	250 (90)	200 (360)	450
non_fiction	50 (210)	1000 (840)	1050
Total	300	1200	1500

Note: Are *gender* and *preferred_reading* correlated?

Using Eq. (3.1) for χ^2 computation, we get

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840}$$
$$= 284.44 + 121.90 + 71.11 + 30.48 = 507.93.$$

For this 2×2 table, the degrees of freedom are $(2-1)(2-1) = 1$. For 1 degree of freedom, the χ^2 value needed to reject the hypothesis at the 0.001 significance level is 10.828

7 What are the various Data Compression techniques? [10]

Lossless data compression

✓ Lossless data compression is used to compress the files **without losing an original file's quality and data**. Simply, we can say that in lossless data compression, file size is reduced, but the quality of data remains the same.

Lossy Data Compression

✓ Lossy data compression is used to compress larger files into smaller files. In this compression technique, some specific amount of **data and quality are removed (loss) from the original file**. It takes less memory space from the original file due to the loss of original data and quality.

	Lossless data compression	Lossy data compression
1.	In Lossless data compression, there is no loss of any data and quality.	In Lossy data compression, there is a loss of quality and data.
2.	In lossless, the file is restored in its original form.	In Lossy, the file does not restore in its original form.
3.	Lossless data compression algorithms are Run Length Encoding, Huffman encoding, Arithmetic encoding, etc.	Lossy data compression algorithms are: Transform coding, Discrete Wavelet Transform, etc.
4.	Lossless compression is mainly used to compress text, sound and images.	Lossy compression is mainly used to compress audio, video, and images.
5.	As compare to lossy data compression, lossless data compression holds more data.	As compare to lossless data compression, lossy data compression holds less data.
6.	File quality is high in the lossless data compression.	File quality is low in the lossy data compression.

8 Explain Association Rule Mining (ARM) ? How Support and Confidence are used for evaluation? [10]
Association rule mining finds interesting associations and relationships among large sets of data items. This rule shows how frequently a itemset occurs in a transaction. A typical example is a Market Based Analysis.

Market Based Analysis is one of the key techniques used by large relations to show associations between items.It allows retailers to identify relationships between the items that people buy together frequently.

Given a set of transactions, we can find rules that will predict the occurrence of an item based on the occurrences of other items in the transaction.

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Before we start defining the rule, let us first see the basic definitions.

Support Count( ) – Frequency of occurrence of a itemset.
Here ({Milk, Bread, Diaper})=2

Frequent Itemset – An itemset whose support is greater than or equal to minsup threshold.
Association Rule – An implication expression of the form X -> Y, where X and Y are any 2 itemsets.
Example: {Milk, Diaper}->{Beer}

Rule Evaluation Metrics –

- **Support(s) –**
  The number of transactions that include items in the {X} and {Y} parts of the rule as a percentage of the total number of transaction.It is a measure of how frequently the collection of items occur together as a percentage of all transactions.
- **Support = (X+Y) total –**
  It is interpreted as fraction of transactions that contain both X and Y.
- **Confidence(c) –**
  It is the ratio of the no of transactions that includes all items in {B} as well as the no of transactions that includes all items in {A} to the no of transactions that includes all items in {A}.
- **Conf(X=>Y) = Supp(X Y) Supp(X) –**
  It measures how often each item in Y appears in transactions that contains items in X also.
- **Lift(l) –**
  The lift of the rule X=>Y is the confidence of the rule divided by the expected confidence, assuming that the itemsets X and Y are independent of each other.The expected confidence is the confidence divided by the frequency of {Y}.
- **Lift(X=>Y) = Conf(X=>Y) Supp(Y) –**
  Lift value near 1 indicates X and Y almost often appear together as expected, greater than 1 means they appear together more than expected and less than 1 means they appear less than expected.Greater lift values indicate stronger association.
-

Example – From the above table, {Milk, Diaper}=>{Beer}
s= ({Milk, Diaper, Beer}) |T|

= 2/5

= 0.4

c= (Milk, Diaper, Beer) (Milk, Diaper)

= 2/3

= 0.67

l= Supp({Milk, Diaper, Beer}) Supp({Milk, Diaper})*Supp({Beer})

= 0.4/(0.6*0.6)

= 1.11

9 Explain Apirori Algorithm with a suitable example. [10]

- **Purpose**: The Apriori Algorithm is an influential algorithm for mining frequent itemsets for boolean association rules.
- **Key Concepts**:
  - **Frequent Itemsets**: The sets of item which has minimum support (denoted by L_i for ith-Itemset).
  - **Apriori Property**: Any subset of frequent itemset must be frequent.
  - **Join Operation**: To find L_k, a set of candidate k-itemsets is generated by joining L_{k-1} itself.
  - Find the frequent itemsets: the sets of items that have minimum support – A subset of a frequent itemset must also be a frequent itemset **(Apriori Property)**
  - i.e., if {AB} is a frequent itemset, both {A} and {B} should be a frequent itemset – Iteratively find frequent itemsets with cardinality from 1 to k (k-itemset)
  - Use the frequent itemsets to generate association rules.
- **The Apriori Algorithm : Pseudo code**
  - **Join Step**: C k is generated by joining Lk-1with itself
  - **Prune Step**: Any (k-1)-itemset that is not frequent cannot be a subset of a frequent k-itemset
  - Pseudo-code:

    C_k: Candidate itemset of size k

    L_k: frequent itemset of size k

    L_1= {frequent items};

    for (k = 1; L_k != ∅; k++) **do begin**

    C_{k+1} = candidates generated from L_k;

    for each transaction t in database do

Increment the count of all candidates in

C_{k+1} That are contained in t

L_{k+1} = candidates in C_{k+1} with min_support

end

return $\cup_k L_k$;

TID	List of Items
T100	I1, I2, I5
T100	I2, I4
T100	I2, I3
T100	I1, I2, I4
T100	I1, I3
T100	I2, I3
T100	I1, I3
T100	I1, I2 ,I3, I5
T100	I1, I2, I3

Example

- Consider a database, **D**, consisting of 9 transactions.
- Suppose min. support count required is **2** (i.e. min_sup = 2/9 = 22 %)
- Let minimum confidence required is **70%**.
- We have to first find out the frequent itemset using Apriori algorithm.
- Then, Association rules will be generated using min. support & min. confidence.

Step 1: Generating 1-itemset Frequent Pattern

Scan D for count of each candidate

Itemset	Sup.Count
{I1}	6
{I2}	7
{I3}	6
{I4}	2
{I5}	2

C_1

Compare candidate support count with minimum support count

Itemset	Sup.Count
{I1}	6
{I2}	7
{I3}	6
{I4}	2
{I5}	2

L_1

- **The set of frequent 1-itemsets, L_1**, consists of the candidate 1- itemsets satisfying minimum support.
- In the first iteration of the algorithm, each item is a member of the set of candidate.

Step 2: Generating 2-itemset Frequent Pattern

Generate C_2 candidates from L_1

Itemset
{I1, I2}
{I1, I3}
{I1, I4}
{I1, I5}
{I2, I3}
{I2, I4}
{I2, I5}
{I3, I4}
{I3, I5}
{I4, I5}

C_2

Scan D for count of each candidate

Itemset	Sup. Count
{I1, I2}	4
{I1, I3}	4
{I1, I4}	1
{I1, I5}	2
{I2, I3}	4
{I2, I4}	2
{I2, I5}	2
{I3, I4}	0
{I3, I5}	1
{I4, I5}	0

C_2

Compare candidate support count with minimum support count

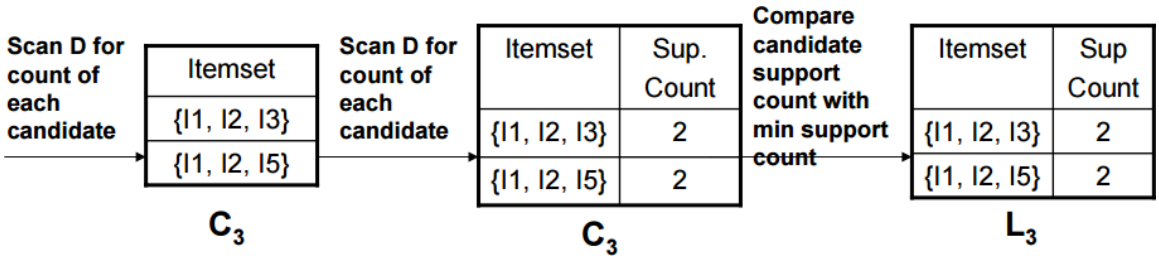
Itemset	Sup Count
{I1, I2}	4
{I1, I3}	4
{I1, I5}	2
{I2, I3}	4
{I2, I4}	2
{I2, I5}	2

L_2

- To discover the set of frequent 2-itemsets, L_2, the algorithm uses L_1 Join L_1 to generate a candidate set of 2-itemsets, C_2.
- Next, the transactions in D are scanned and the support count for each candidate itemset in C_2 is accumulated (as shown in the middle table).
- The set of frequent 2-itemsets, L_2, is then determined, consisting of those candidate 2-itemsets in C_2 having minimum support.

- **Note**: We haven't used Apriori Property yet.

Step 3: Generating 3-itemset Frequent Pattern



10 Explain and construct FP Tree for given Transaction data:[10]

TId	Items
1	A B
2	B C D
3	A C D E
4	A D E
5	A B C
6	A B C D
7	B C
8	A B C
9	A B D
10	B C E

Minimum Support >= 2

TID	Items
1	A B
2	B C D
3	A C D E
4	A D E
5	A B C
6	A B C D
7	B C
8	A B C
9	A B D
10	B C E

Header

Item	Support
B	8
A	7
C	7
D	5
E	3

FP Tree, After the 10th Transaction,

TID	Items
1	B A
2	B C D
3	A C D E
4	A D E
5	B A C
6	B A C D
7	B C
8	B A C
9	B A D
10	B C E