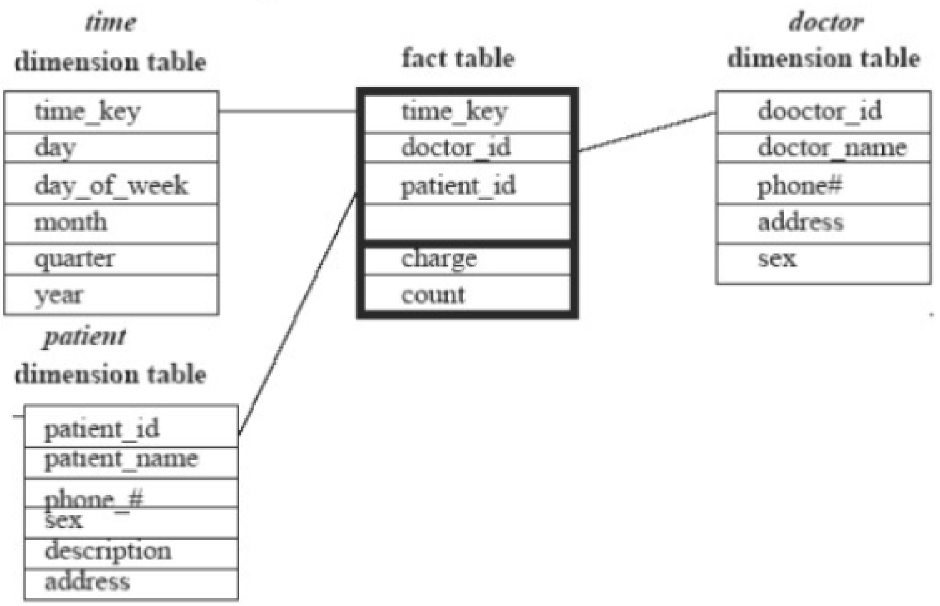
Unit 1

1) What are the differences between OLAP and OLTP? CO1 L2 10M

2) Suppose that a data warehouse consists of the three dimensions time, doctor, and patient, and the two measures count and charge, where charge is the fee that a doctor charges a patient for a visit.
Dimensions time contains: time_key, day, day_of_week, month, quarter, year
Dimension doctor contains: dooctor_id , doctor_name, phone number, address, sex
Dimension patient contains: patient_id, patient_name, phone number, sex, description, address
Draw a star schema diagram for the above data warehouse. CO1 L3 10 M

NOTE : PROVIDE EXPLANATION ABOUT STAR SCHEMA, FACT TABLE, DIMENSION TABLE ETC.



3) What are the typical OLAP operations that can be performed on a data cube? CO1 L1 10M

4) What are the characteristics of a Data Warehouse? What is the role of a metadata repository in management of a data warehouse? Give a real time application to demonstrate relation between data mart and a data warehouse. Co1 L2 5M

A data warehouse is a large and centralized repository of data that is used to support business decision-making processes. Here are some of the key characteristics of a data warehouse:

Subject-oriented: A data warehouse is organized around specific subjects or areas of interest, such as sales, customers, or products.

Integrated: Data from different sources is consolidated and integrated into a single, consistent view.

Time-variant: Data in a data warehouse is stored in a way that allows for analysis of historical trends and changes over time.

Non-volatile: Data in a data warehouse is read-only and is not modified or updated in real-time, ensuring data integrity and consistency.

A metadata repository is a database that stores information about the data in a data warehouse, including data definitions, data relationships, and data transformations. The role of a metadata repository is to provide a centralized source of information about the data in a data warehouse, making it easier to manage and maintain the warehouse.

A data mart is a subset of a data warehouse that is focused on a specific business function or department, such as sales or finance. It contains a smaller, more specialized set of data than the larger data warehouse.

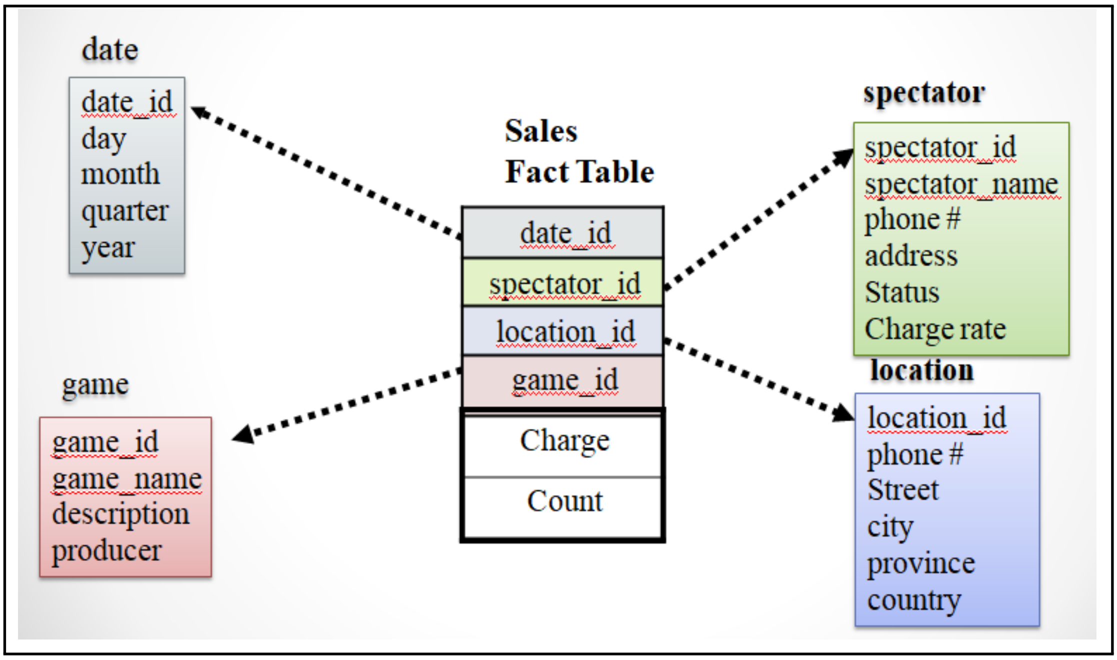
A real-time application that demonstrates the relationship between a data mart and a data warehouse could be a retail store chain that uses a data warehouse to store all of its sales data, but then creates separate data marts for each store location to analyze local trends and performance. The data warehouse provides a centralized source of data for all locations, while the data marts allow for more targeted analysis and decision-making at the local level.

5) Explain about the Three-tier data warehouse architecture with a neat diagram. CO1, L1, 10M

6) Suppose that a data warehouse consists of the four dimensions, date, spectator, location, and game, and the two measures, count and charge, where charge is the fare that a spectator pays when watching a game on a given date.

Dimension date contains date_id, day, month, quarter, and year
Dimension game contains game_id, game_name, description, producer
Dimension spectator contains spectator_id, spectator_name, status, phone , address
Dimension location contains location_id, location_name, phone number, street, city, state, country
Draw a star schema diagram for the above data warehouse. CO1 L3 10 M

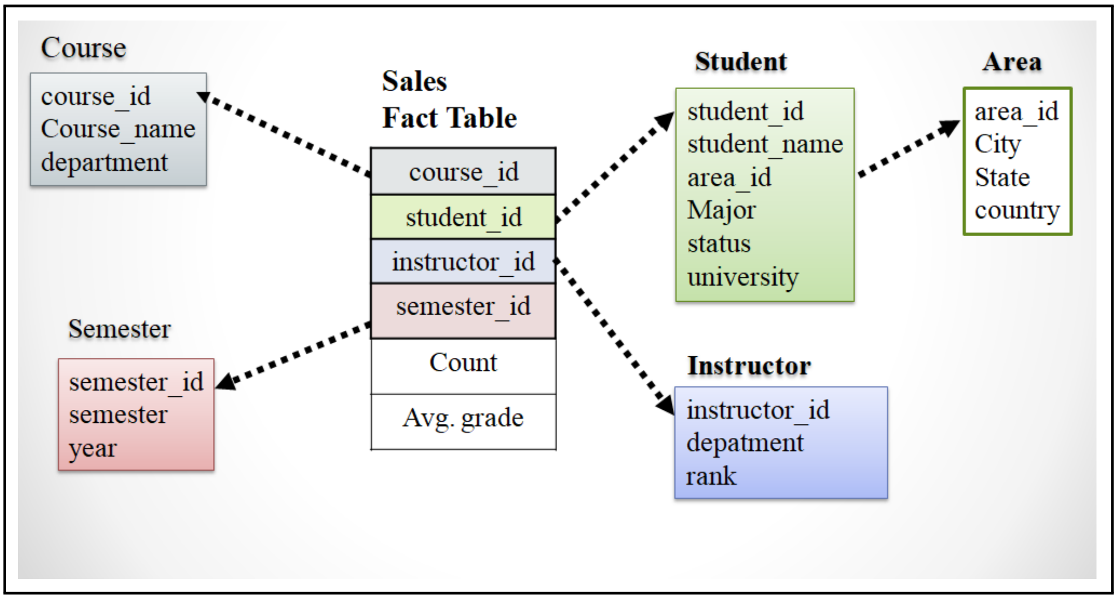
NOTE : PROVIDE EXPLANATION ABOUT STAR SCHEMA, FACT TABLE, DIMENSION TABLE ETC.
The fact table contains foreign keys to all four dimensions, as well as the measures count and charge. This schema allows for efficient querying and analysis of data across different dimensions, as well as easy integration with reporting and visualization tools.



7) Write in brief about schemas in multidimensional data model. CO1 L3 10M

8) Suppose that a data warehouse for Vishnu Institute of Technology consists of the following four dimensions: student, course, semester, and instructor, and two measures count and avg grade.
Dimension course contains: course_id , course_name, department
Dimension semester contains: semester_id, semester, year
Dimension student contains: student_id, student_name, area_id, major, status, university
Dimension instructor contains: instructor_id, dept, rank

Note: Student dimension is further normalized on area_id: area_id, city, province, country
Design a snowflake schema for the above data warehouse. CO1 L3 10 M



9) A) Justify the statement “ It is better to mine knowledge from data warehouses rather than directly for data sources” . CO1, L2, 5M

B) In which situation Star Schema, Snowflake schema and fact constellation schema should be used? Justify your answer with a practical example. CO1, L2, 5M

10) Write short note on: CO1, L1, 10M

a. Roll Up
b. Drill down
c. Slice
d. Dice
c. Pivot.

Unit 2:

1)

a. What is data transformation? Consider the blood sugar values of 3 patients in a hospital: (140, 200,90). Transform these values using

B) Min-Max max normalization (new min=0, new max=1)

C) Z-score normalization. CO2 L3 5M

Data transformation is the process of converting data from one format or range to another, in order to make it more suitable for analysis or to conform to a particular data model or standard.

In the case of blood sugar values of 3 patients in a hospital (140, 200, 90), we can use two common methods of data normalization: min-max normalization and z-score normalization.

Min-Max Normalization:
Min-max normalization is a technique that scales the data to a fixed range between 0 and 1. The formula for min-max normalization is:
new_value = (old_value - min_value) / (max_value - min_value)

where min_value and max_value are the minimum and maximum values in the data set, respectively.

In our case, the minimum value is 90 and the maximum value is 200. Therefore, the min-max normalized values are:

Patient 1: (140 - 90) / (200 - 90) = 0.375

Patient 2: (200 - 90) / (200 - 90) = 1
Patient 3: (90 - 90) / (200 - 90) = 0

So, the min-max normalized values for the blood sugar levels are: 0.375, 1, and 0.

Z-Score Normalization:
Z-score normalization is a technique that standardizes the data to have a mean of 0 and a standard deviation of 1. The formula for z-score normalization is:
new_value = (old_value - mean) / standard_deviation

where mean is the average value of the data set and standard_deviation is the standard deviation of the data set.

In our case, the mean is (140 + 200 + 90) / 3 = 143.33 and the standard deviation is approximately 48.89. Therefore, the z-score normalized values are:

Patient 1: (140 - 143.33) / 48.89 = -0.067
Patient 2: (200 - 143.33) / 48.89 = 1.157
Patient 3: (90 - 143.33) / 48.89 = -1.093

So, the z-score normalized values for the blood sugar levels are: -0.067, 1.157, and -1.093.

b. What is the difference between predictive and descriptive tasks? Give 3 applications of both. CO2 L2 5M

Predictive and descriptive analytics are two types of data analysis used in business and other fields. The main difference between these two types of analysis is that predictive analytics involves predicting future outcomes, while descriptive analytics involves analyzing past and present data to understand what has happened or is happening.

Predictive analytics involves using statistical algorithms and machine learning techniques to analyze historical data and make predictions about future events. It is used to answer questions like "What will happen next?" or "What is likely to happen in the future?" Some examples of predictive analytics applications are:

Customer churn prediction: Using customer data and behavior patterns to predict which customers are most likely to leave a company.

Credit risk assessment: Analyzing past financial data to predict the likelihood of default or non-payment on a loan or credit card.

Sales forecasting: Using historical sales data and market trends to predict future sales numbers.

Descriptive analytics, on the other hand, involves analyzing past and present data to understand what has happened or is happening in a business or other field. It is used to answer questions like "What happened?" or "What is happening now?" Some examples of descriptive analytics applications are:

Business intelligence reporting: Creating reports and dashboards to provide insights into past and present business performance.

Web analytics: Analyzing website traffic and user behavior to understand how visitors interact with a website.

Supply chain management: Analyzing supply chain data to identify bottlenecks and areas for improvement in the production and distribution process.

In summary, the main difference between predictive and descriptive analytics is that predictive analytics is focused on making predictions about future events, while descriptive analytics is focused on analyzing past and present data to understand what has happened or is happening.

2) What are the different kinds of data on which data mining is generally applied? CO2 L1 10M

Data mining is the process of discovering patterns, trends, and insights from large and complex data sets. Data mining can be applied to different types of data, including:

Relational data: This is data that is organized in a tabular format, with columns representing attributes or variables and rows representing individual data points or observations. Relational data is commonly found in databases and spreadsheets.

Transactional data: This is data that is generated by transactions or events, such as point-of-sale transactions or online clicks. Transactional data is often recorded in log files or other event-based systems.

Time-series data: This is data that is recorded over time, such as stock prices or weather data. Time-series data can be used to identify trends and patterns over time.

Text data: This is data that is in the form of natural language text, such as emails, social media posts, or customer reviews. Text data can be analyzed using natural language processing techniques to extract insights and sentiments.

Graph data: This is data that represents relationships between entities, such as social networks or road networks. Graph data can be analyzed to identify patterns and communities within the network.

Data mining techniques can be applied to any type of data, as long as it is organized in a way that can be analyzed by the algorithms. The choice of data mining technique depends on the type of data and the research questions being asked.

3)

a. What is data mining? What are its applications? CO2, L1, 5M

Data mining is the process of extracting useful and meaningful insights and patterns from large, complex datasets using advanced computational and statistical techniques. It involves exploring and analyzing data from multiple angles and using different perspectives to uncover previously unknown correlations and relationships.

The applications of data mining are diverse and span across different industries and domains. Some of the most common applications of data mining include:

Business and Marketing: Data mining can be used to identify customer preferences, market trends, and purchase patterns, which can be used to make informed business decisions and marketing strategies.

Healthcare: Data mining can be used to identify patterns and relationships in medical data to improve patient outcomes, identify disease risk factors, and develop personalized treatment plans.

Finance: Data mining can be used to identify fraudulent transactions, detect anomalies in financial data, and predict financial market trends.

Manufacturing: Data mining can be used to optimize production processes, identify quality control issues, and improve supply chain management.

Telecommunications: Data mining can be used to analyze call records, identify network issues, and improve service quality.

Sports: Data mining can be used to analyze player performance, predict game outcomes, and develop game strategies.

Social Media: Data mining can be used to analyze social media posts, identify trends and sentiments, and target ads to specific audiences.

In summary, data mining is a powerful tool that can be used to extract insights and knowledge from large, complex datasets, and has applications across many industries and domains.

b. Motivating challenges of data mining CO2, L1, 5M

4) What is the need of sampling for data reduction? Write a short note on Sampling with replacement, Sample without replacement, stratified sampling, and progressive sampling. CO2, L1 10 M

5)

a. What is the need to reduce the number of features? What are the approaches for feature subset selection? CO2, L1, 5M

b. What is the difference between Nominal, Binary, Ordinal and Numeric attributes? Consider a student dataset having columns – Gender, B.Tech marks, Pincode, Coding skills (in grade),Placement status. Specify the type of attribute. CO2, L3, 5M

6)

a. Consider a dataset

Roll No	Car colour	Tyres Used
1	Red	MRF
2	Yellow	Apollo
3	Red	Apollo
4	Yellow	MRF

Find the proximity matrix for the above data. Consider Car colour and Tyres used as nominal attributes. CO2, L3, 7M

Proximity measures for Nominal Attributes

Nominal attributes can have two or more different states e.g. an attribute 'color' can have values like 'Red', 'Green', 'Yellow', etc. Dissimilarity for nominal attributes is calculated as the ratio of total number of mismatches between two data tuples to the total number of attributes.

Let M be the total number of states of a nominal attribute. Then the states can be numbered from 1 to M. However, the numbering does not denote any kind of ordering and can not be used for any mathematical operations.

Let m be total number of matches between two tuple attributes and p be total number of attributes, then the dissimilarity can be calculated as,

$d(i,j)=p-mp$

We can calculate similarity as,

$s(i,j)=1-d(i,j)$

b. What is the difference between Data Matrix and Dissimilarity Matrix. CO2, L1, 3M

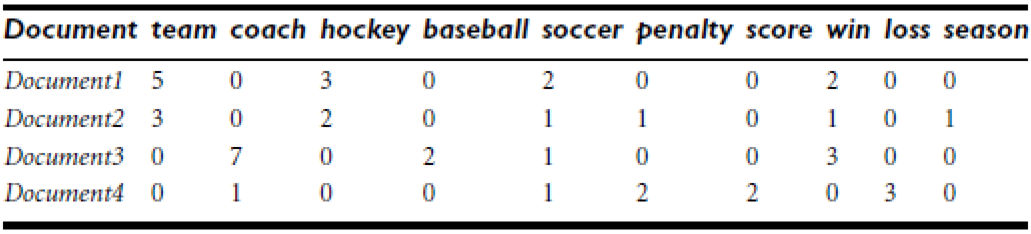
A Data Matrix is a rectangular table of data where the rows represent individual samples or observations, and the columns represent variables or features. Each cell in the table contains a value or measurement for the corresponding sample and feature. Data matrices are commonly used in various fields such as statistics, biology, and computer science to represent and analyze complex data sets.

On the other hand, a Dissimilarity Matrix (also called a distance matrix) is a square matrix that quantifies the dissimilarity or distance between pairs of samples or observations. Each element of the matrix represents the distance or dissimilarity between two samples or observations. Dissimilarity matrices are commonly used in various fields such as biology, ecology, and computer science to compare and cluster similar samples or observations based on their similarity or dissimilarity.

In summary, a Data Matrix represents the raw data, whereas a Dissimilarity Matrix represents the distance or dissimilarity between the samples or observations based on some measure or metric.

7)

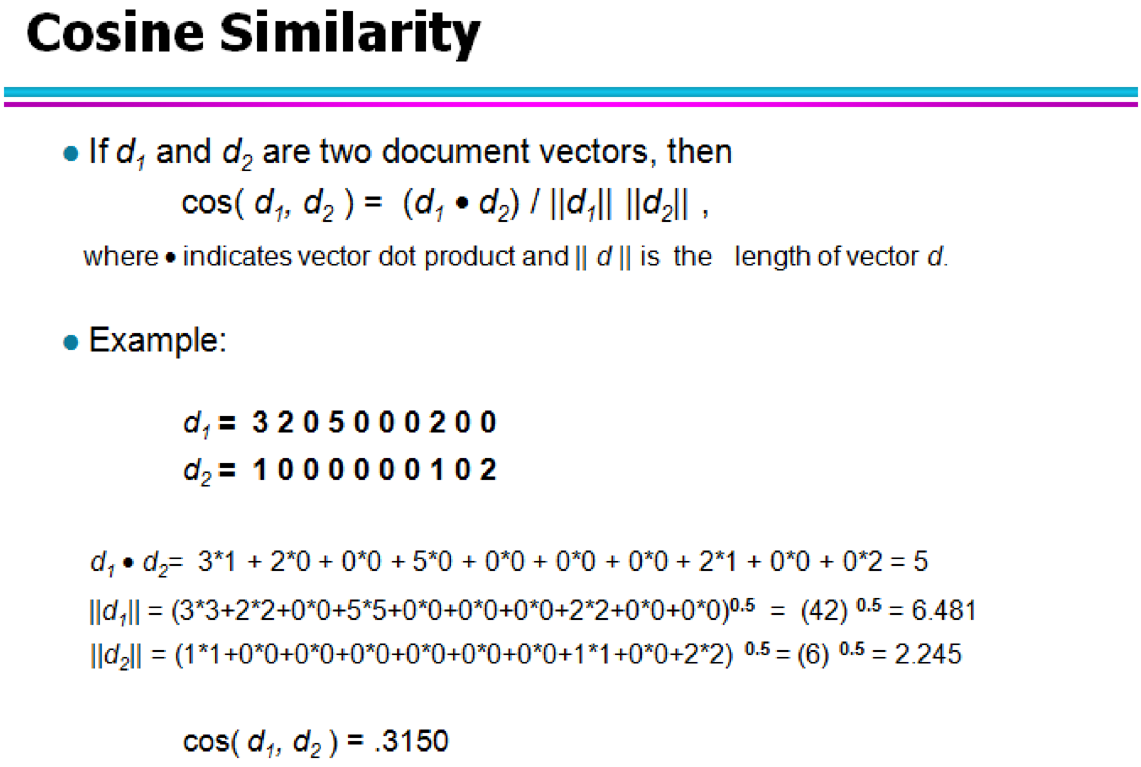
a. Consider the dataset below



Document	team	coach	hockey	baseball	soccer	penalty	score	win	loss	season
Document1	5	0	3	0	2	0	0	2	0	0
Document2	3	0	2	0	1	1	0	1	0	1
Document3	0	7	0	2	1	0	0	3	0	0
Document4	0	1	0	0	1	2	2	0	3	0

Find the cosine similarity between Document 1 and Document 2. CO2, L3, 7M

Cosine Similarity Example:



Cosine Similarity

- If d_1 and d_2 are two document vectors, then

$$\cos(d_1, d_2) = (d_1 \bullet d_2) / ||d_1|| \; ||d_2|| ,$$

where $\bullet$ indicates vector dot product and $|| d ||$ is the length of vector d.

- Example:

$$d_1 = 3\ 2\ 0\ 5\ 0\ 0\ 0\ 2\ 0\ 0$$
$$d_2 = 1\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 2$$

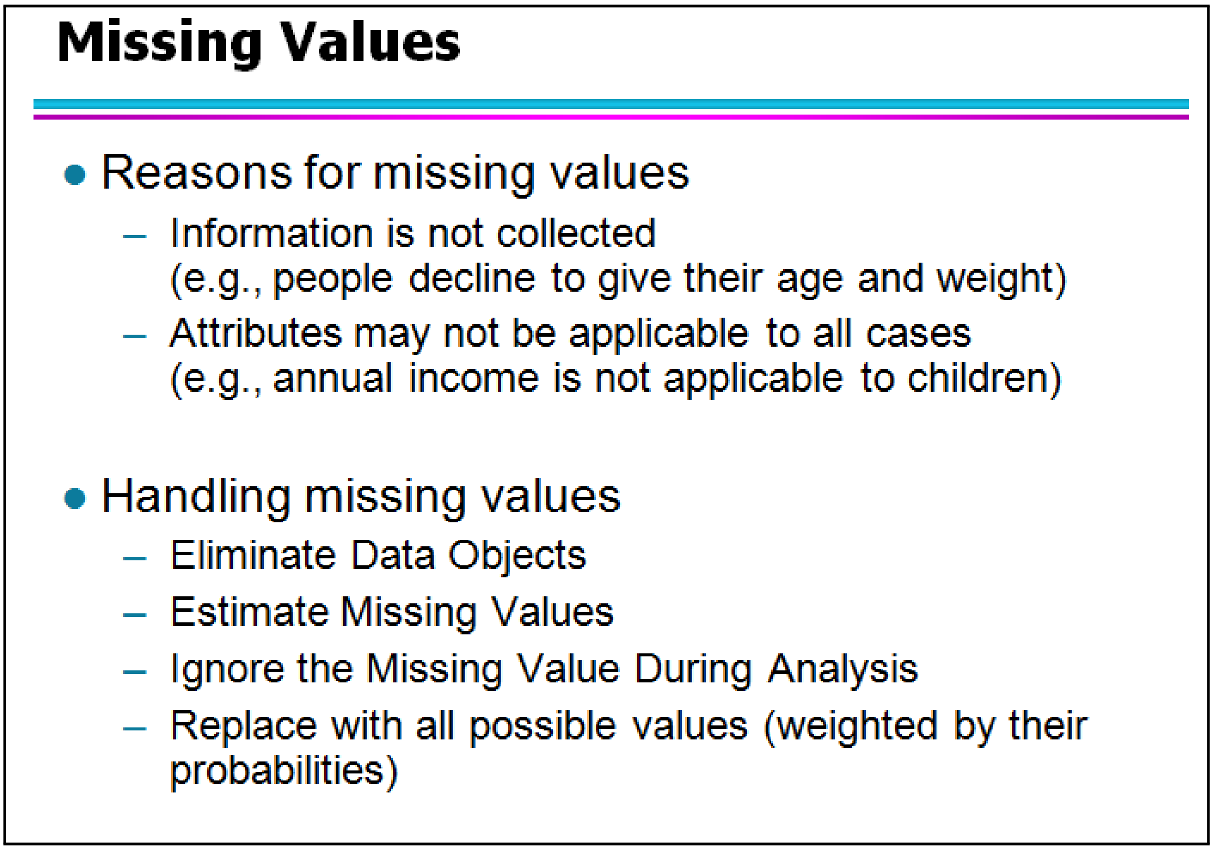
$d_1 \bullet d_2 = 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5$

$||d_1|| = (3*3+2*2+0*0+5*5+0*0+0*0+0*0+2*2+0*0+0*0)^{0.5} = (42)^{0.5} = 6.481$

$||d_2|| = (1*1+0*0+0*0+0*0+0*0+0*0+0*0+1*1+0*0+2*2)^{0.5} = (6)^{0.5} = 2.245$

$$\cos(d_1, d_2) = .3150$$

b. What are the different ways to deal with missing values? CO2, L1 3M,



Missing Values

- Reasons for missing values
  - Information is not collected
    (e.g., people decline to give their age and weight)
  - Attributes may not be applicable to all cases
    (e.g., annual income is not applicable to children)

- Handling missing values
  - Eliminate Data Objects
  - Estimate Missing Values
  - Ignore the Missing Value During Analysis
  - Replace with all possible values (weighted by their probabilities)

8)

a. What is the difference between Classification and regression? Give examples of classification and regression in terms of education domain. CO2 L1 5M

b. Demonstrate computation of the following measures for similarity/dissimilarity among data: a) Euclidean distance b) Manhattan measure CO2 L2 5M

9) Consider the Table below.

Sl.No	Test (Ordinal)
1	Excellent
2	Fair
3	Good
4	Excellent

The attribute “Test” is an ordinal attribute. Compute the dissimilarity matrix for the above table. CO2, L3, 10M

10) Demonstrate computation to calculate

a. Dissimilarity between Binary Symmetric attributes
b. Dissimilarity between Binary Asymmetric attributes. CO2, L3, 10M

Unit 3:

1) Mention different characteristics to construct a Decision tree. CO3 L1 5M

A decision tree is a machine learning algorithm that is used for both regression and classification tasks. It is a tree-like model where each internal node represents a decision on an input attribute, each branch represents the outcome of the decision, and each leaf node represents a class label or a numerical value.

The characteristics used to construct a decision tree are:

Impurity measures: Impurity measures such as entropy, Gini Index, or Classification error are used to evaluate the quality of a split in the decision tree. The split with the lowest impurity measure is chosen as the best split.

Attribute selection measures: Attribute selection measures such as Information Gain, Gain Ratio, or Chi-Square Test are used to determine the importance of each input attribute in the decision tree. The attribute with the highest score is selected as the root node of the decision tree.

Splitting criteria: Splitting criteria such as binary split, multiway split, or rule-based split are used to divide the dataset into subsets based on the selected input attribute. The decision tree recursively splits the subsets until all the samples in a subset belong to the same class or have the same numerical value.

Pruning techniques: Pruning techniques such as pre-pruning or post-pruning are used to avoid overfitting and improve the generalization performance of the decision tree. Pre-pruning involves stopping the tree construction early, while post-pruning involves removing branches that lead to low-performance leaves.

Handling missing values: Missing values can be handled by imputation techniques such as mean imputation, median imputation, or regression imputation. Alternatively, decision trees can handle missing values by assigning a probability weight to each branch based on the probability of the missing value.

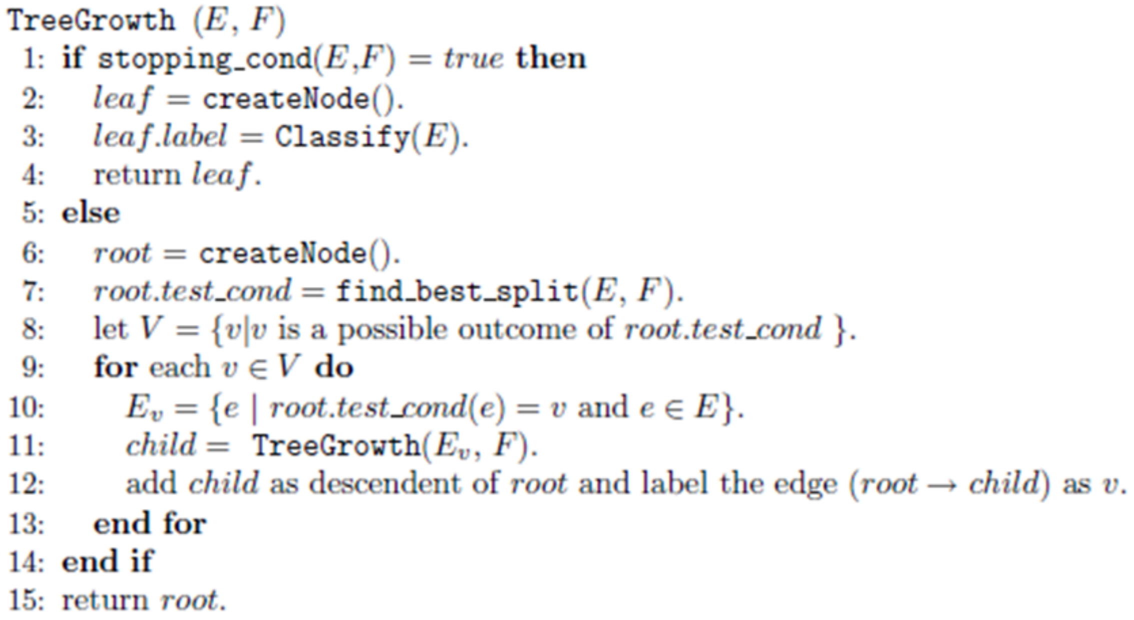
By considering these characteristics, a decision tree can be constructed to make accurate predictions and perform well on unseen data.

2) What is meant by Classification? What are applications of classification models?

CO3 L1 5M

3) Write the algorithm of Decision Tree Induction CO3 L1 5M

Algorithm for decision tree induction
Note: E is training records , F is the attribute set.



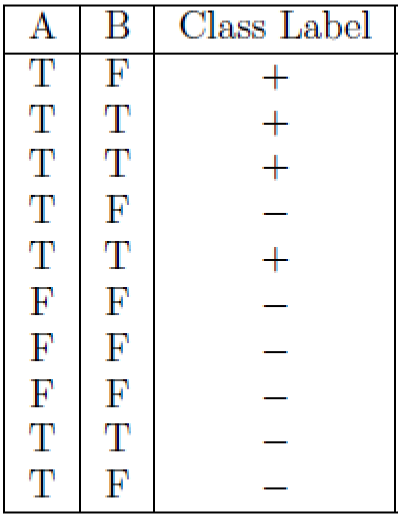
```
TreeGrowth ($E$, $F$)
if stopping_cond($E$,$F$) = $true$ then
    $leaf$ = createNode().
    $leaf.label$ = Classify($E$).
    return $leaf$.
else
    $root$ = createNode().
    $root.test\_cond$ = find_best_split($E$, $F$).
    let $V = \{v | v$ is a possible outcome of $root.test\_cond\}$.
    for each $v \in V$ do
        $E_v = \{e \mid root.test\_cond(e) = v$ and $e \in E\}$.
        $child$ = TreeGrowth($E_v$, $F$).
        add $child$ as descendent of $root$ and label the edge ($root \rightarrow child$) as $v$.
    end for
end if
return $root$.
```

i) The create node() function extends the decision tree by creating a new node. A node in the decision tree has either a test condition, denoted as node.test_cond, or a class label, denoted as node.label.
ii) The find.best_split () function determines which attribute should be selected as the test condition for splitting the training records.
iii) The classify() function determines the class label to be assigned to a leaf node.
iv) The stopping_cond() function is used to terminate the tree-growing process by testing whether all the records are classified or not.

4) What are the Methods for expressing attribute test conditions? CO3 L1 5M

5) What are Measures for selecting the best split? CO3 L1 5M

6) Consider the table below.



A	B	Class Label
T	F	+
T	T	+
T	T	+
T	F	−
T	T	+
F	F	−
F	F	−
F	F	−
T	T	−
T	F	−

The table has 2 attributes “A” and “B” and the class labels. Which Attribute has the highest gain. Use Gini index to calculate the impurity. CO3 L3 5M

7) What do you mean by classification, training data, test data, Learning algorithm and model. CO3 L3 5M

8) Write short notes on:
   i. Gini Index
   ii. Entropy CO3 L3 5M

9) Describe the general approach to solving a classification problem with a suitable example. List and explain the design issues of decision tree induction. CO3 L2 5M

10) Consider the table below

Coding skills	Communication skills	Placed/ Not Placed
A	A	Placed
A	B	Placed
B	A	Not Placed
B	B	Not Placed

Out of the 2 attributes “Coding skills” and “Communication skills”, Which Attribute has the highest gain. Use Gini index to calculate the impurity. CO3 L3 5M