Unit – 3

Introduction to Data Mining

Syllabus

- What is data mining?
- Decision Trees – introduction, Types, Advantages and Disadvantages,
- Types of Sources of Data in Data Mining ,
- Data mining Techniques.
- Examples of Data mining applications.

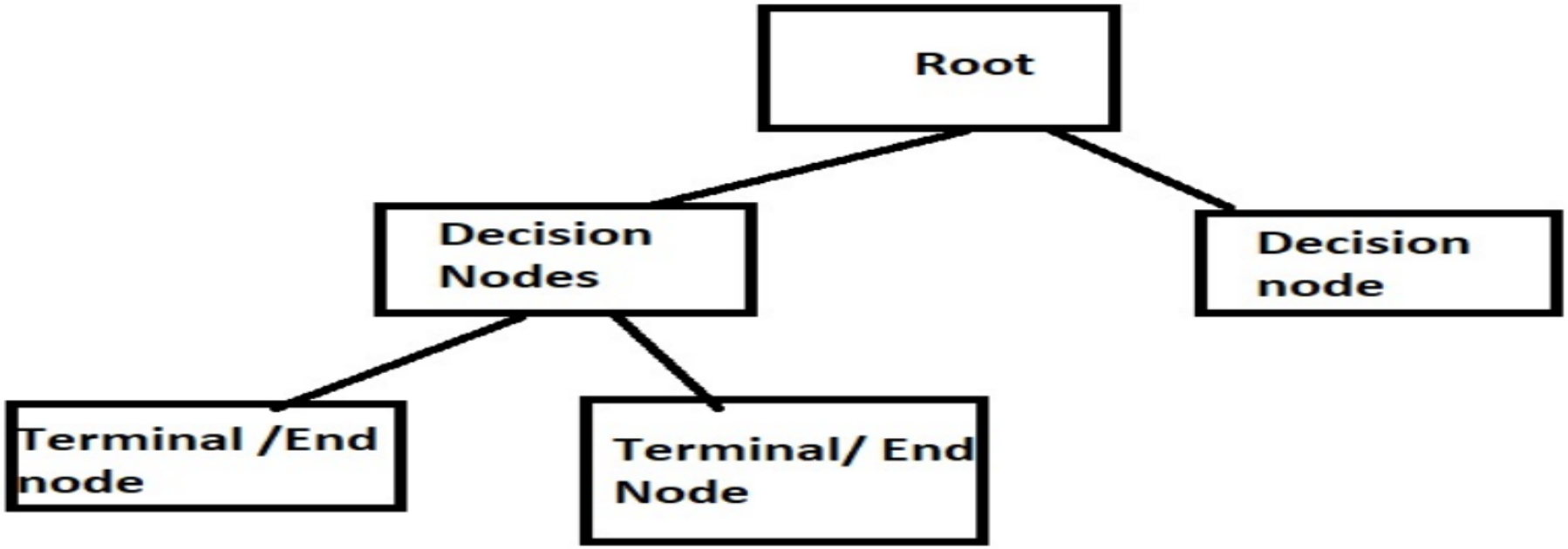
Data Mining

- Data mining refers to extracting or “mining" knowledge from large amounts of data.
- Data mining is the process of extracting valuable and useful information from large data sets, often using statistical and computational techniques.
- The main goal of data mining is to identify patterns, relationships, and trends within the data that can be used to make informed decisions or predictions.
- It can be applied to a wide range of fields, including business, healthcare, finance, marketing, and science.
- Mining of gold from rocks or sand is referred to as gold mining rather than rock or sand mining. Thus, data mining should have been more appropriately named **knowledge mining from data**, which is unfortunately somewhat long.

Decision Trees

- A decision tree is a graphical representation of a decision-making process that helps in identifying the possible outcomes of a particular course of action.
- A decision tree typically starts with a single node, which branches into possible outcomes. Each of those outcomes leads to additional nodes, which branch off into other possibilities. This gives it a tree-like shape.
- In a decision tree, each internal node represents a decision based on a particular attribute or feature, and each leaf node represents a possible outcome
- It is a popular ML Algorithm that is used for prediction in various fields such as business, healthcare, finance etc.

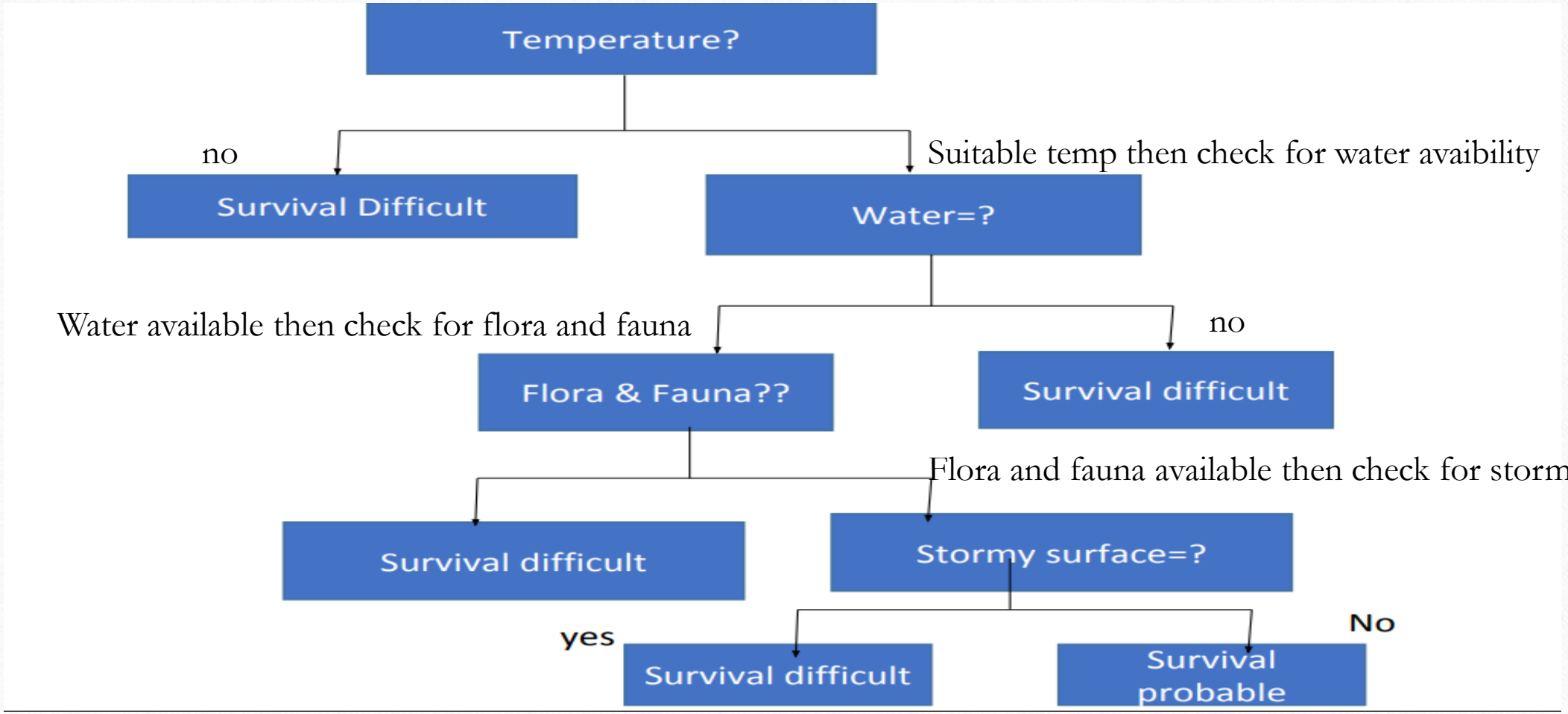
Structure of a Decision Tree



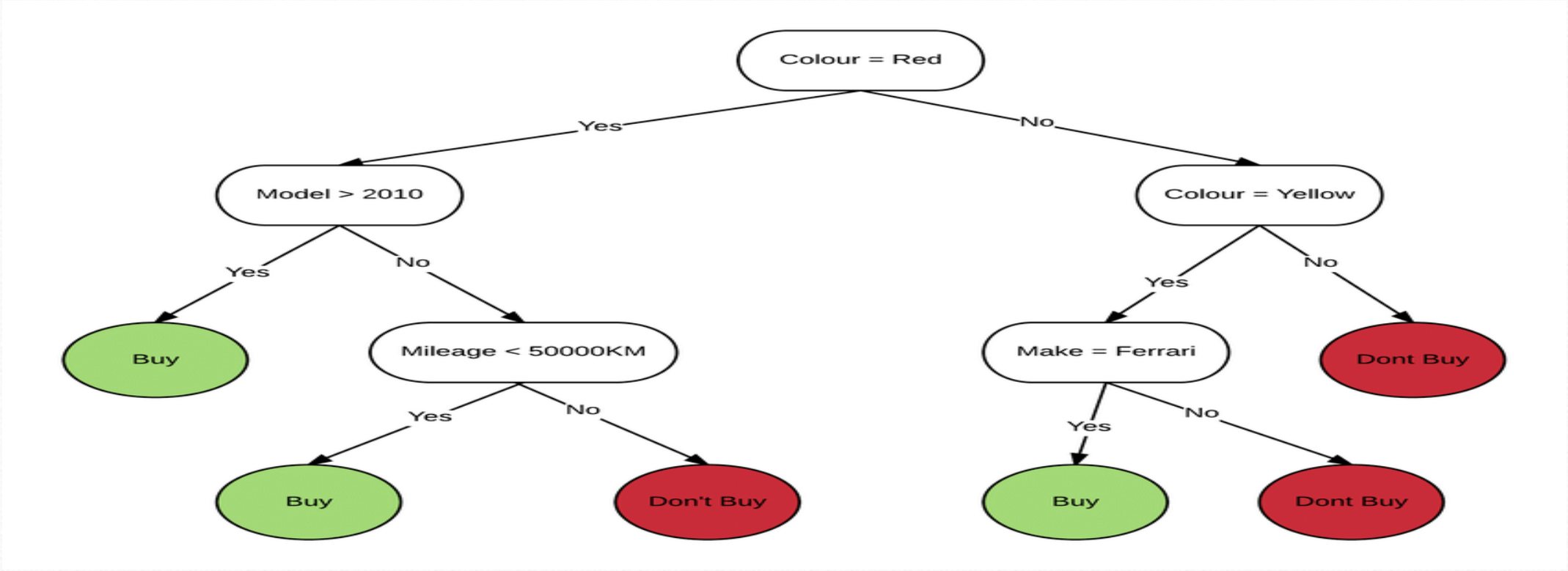
Types of Nodes in a Decision Tree

- There are three different types of nodes: chance nodes, decision nodes, and end nodes.
- A **Chance Node**, represented by a circle, shows the probabilities of certain results.
- A **Decision Node**, represented by a square, shows a decision to be made.
- An **End Node** shows the final outcome of a decision path.

Decision Tree: To check whether survival is possible in a new habitat.



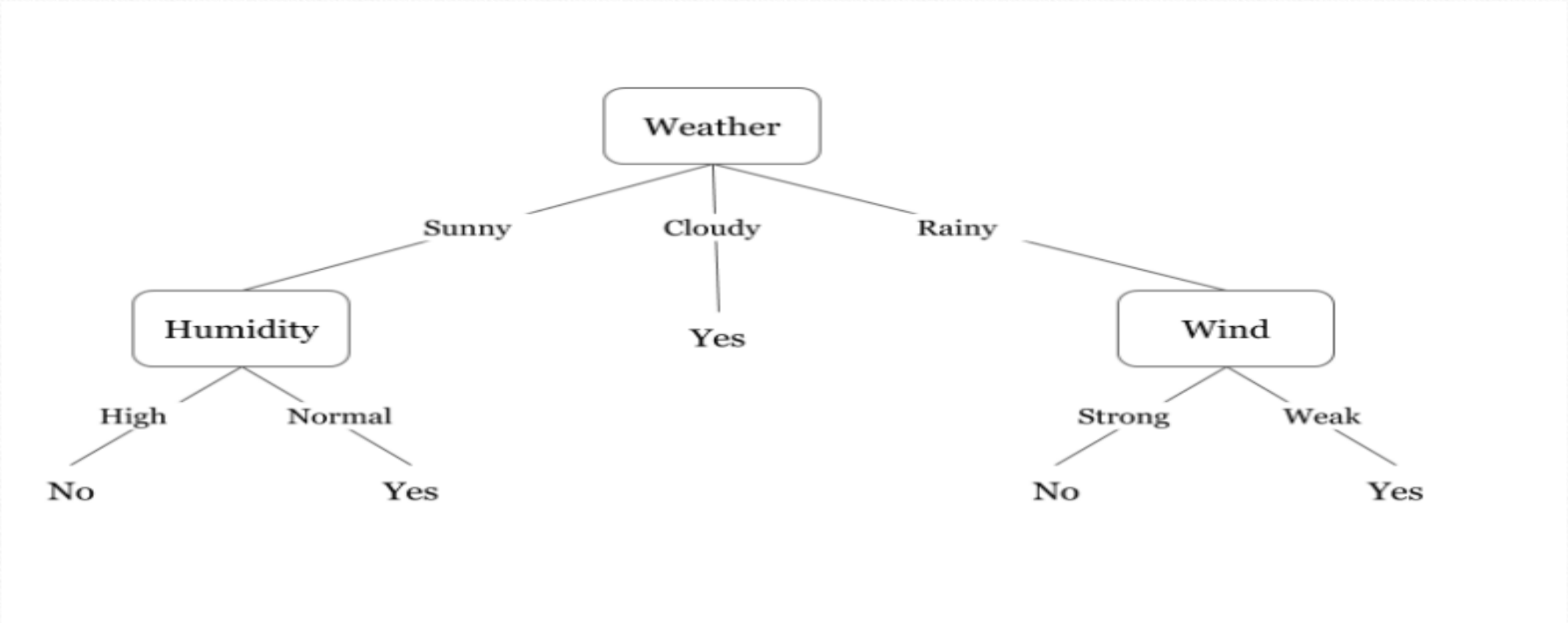
Decision Tree: Lets consider a person's preference for buying a car. If the color is red, then further constrains like built year and mileage is considered. If not, then the brand of the vehicle is kept in mind. Wherever these conditions are not met, the car is not bought. On the other hand, it would be bought if it is red and newer than 2010, red car with good mileage, or a yellow Ferrari.



Decision Tree: Let's assume we want to play badminton on a particular day — say Saturday — how will you decide whether to play or not.

Day	Weather	Temperature	Humidity	Wind	Play?
1	Sunny	Hot	High	Weak	No
2	Cloudy	Hot	High	Weak	Yes
3	Sunny	Mild	Normal	Strong	Yes
4	Cloudy	Mild	High	Strong	Yes
5	Rainy	Mild	High	Strong	No

Decision Tree: Let's assume we want to play badminton on a particular day — say Saturday — how will you decide whether to play or not.



Types of a Decision Tree

- Categorial Variable Decision Tree

A categorical variable decision tree includes categorical target variables / output variables that are divided into categories. For example, the categories can be yes or no. The categories mean that every stage of the decision process falls into one of the categories, and there are no in-betweens.

- Continuous Variable Decision Tree

A continuous variable decision tree is a decision tree with a continuous target variable. For example, the income of an individual whose income is unknown can be predicted based on available information such as their occupation, age, and other continuous variables.

Advantages of a Decision Tree

- Decision trees generate understandable rules
- Decision trees perform classification without requiring much computation
- Decision trees are capable of handling both continuous and categorical variables
- Decision trees provide a clear indication of which fields are most important for prediction or classification

Disadvantages of a Decision Tree

- Overfitting: Decision trees can easily overfit the training data.
- Instability: Decision trees can be sensitive to small changes in the data or model parameters.
- Bias: Decision trees can have high bias if the model is too simple.
- Complexity: Decision trees can become complex and difficult to interpret.
- Overemphasis on certain features: Decision trees can overemphasize certain features in the data.

Types of Sources of Data in Data Mining

- Flat Files
- Relational Databases
- Data Warehouse
- Transactional Databases
- Multimedia Databases
- Spatial Databases
- Time Series Databases
- World Wide Web(WWW)

Sources Contd…

1) Flat Files

- Flat files is defined as data files in text form or binary form with a structure that can be easily extracted by data mining algorithms
- Data stored in flat files have no relationship or path among themselves, like if a relational database is stored on flat file, then there will be no relations between the tables.
- This can include email messages, social media posts, customer reviews, and other unstructured data.
- Application: Used in Data Warehousing to store data, Used in carrying data to and from server, etc.

Sources Contd…

2) Relational Databases

- A Relational database is defined as the collection of data organized in tables with rows and columns.
- Physical schema in Relational databases is a schema which defines the structure of tables.
- Logical schema in Relational databases is a schema which defines the relationship among tables.
- Standard API of relational database is SQL.
- Application: Data Mining, ROLAP model, etc.

Sources Contd…

3) Data WareHouse

- A Data Warehouse is defined as the collection of data integrated from multiple sources that will queries and decision making.
- There are three types of Data Warehouse : Enterprise Data Warehouse, Data Mart and Virtual Warehouse.
- Two approaches can be used to update data in Data Warehouse : Query-driven Approach and Update-driven Approach.
- Application: Business decision making, Data mining, etc.

Sources Contd…

4) Transactional Databases

- Transactional databases is a collection of data organized by time stamps, date, etc to represent transaction in databases.
- Transactional databases are a type of database that records the transactions or interactions between entities, such as customers and products, in a business or organization.
- This type of database has the capability to roll back or undo its operation when a transaction is not completed or committed.
- Follows ACID property of DBMS.
- Application: Banking, Distributed systems, Object databases, etc.

Sources Contd…

5) Multimedia Databases

- Multimedia databases consists audio, video, images and text media.
- They can be stored on Object-Oriented Databases.
- They are used to store complex information in a pre-specified formats.
- This can include medical images, satellite images, phone call recordings, music files, etc.
- Application: Digital libraries, video-on demand, news-on demand, musical database, etc.

Sources Contd…

7) WWW Databases

- WWW refers to World wide web is a collection of documents and resources like audio, video, text, etc. which are identified by Uniform Resource Locators (URLs) through web browsers, linked by HTML pages, and accessible via the Internet network.
- It is the most heterogeneous repository as it collects data from multiple resources.
- It is dynamic in nature as Volume of data is continuously increasing and changing.
- Application: Online shopping, Job search, Research, studying, etc.

Data Mining Techniques

1. CLASSIFICATION ANALYSIS
2. ASSOCIATION RULE LEARNING
3. ANOMALY OR OUTLIER DETECTION
4. CLUSTERING ANALYSIS
5. REGRESSION ANALYSIS

Classification Analysis

- Classification analysis is a type of data mining technique used to predict the class or category of a target variable based on one or more input variables.
- This analysis is used to retrieve important and relevant information about data, and metadata. It is used to classify different data in different classes.
- Classification is similar to clustering in a way that it also segments data records into different segments called classes.
- Classification Algorithms can be used to predict the class of the target variable for new data based on the learned patterns and relationships.
- A classic example of classification analysis would be our Outlook email. In Outlook, they use certain algorithms to characterize an email as legitimate or spam.

Association Rule Learning

- Association rule learning is a type of data mining technique used to identify patterns and relationships between variables in large data sets.
- This technique can help you unpack some hidden patterns in the data that can be used to identify variables within the data and the concurrence of different variables that appear very frequently in the dataset.
- The goal of association rule learning is to identify sets of items that frequently occur together, known as itemset, and to generate rules that describe the relationships between these itemset.
- Association rules are useful for examining and forecasting customer behavior. It is highly recommended in the retail industry analysis.
- This technique is used to determine shopping basket data analysis, product clustering, catalog design and store layout.

Anomaly or Outlier Detection

- Anomaly or outlier detection is a data mining technique used to identify data points that deviate significantly from the normal or expected behavior of the data and can be indicative of errors, fraud, or other unusual occurrences.
- Anomalies are also known as outliers, novelties, noise, deviations and exceptions. Often they provide critical and actionable information.
- These types of items are statistically aloof as compared to the rest of the data and hence, it indicates that something out of the ordinary has happened and requires additional attention.
- This technique can be used in a variety of domains, such as intrusion detection, system health monitoring, fraud detection, fault detection, event detection in sensor networks, and detecting eco-system disturbances.
- Analysts often remove the anomalous data from the dataset top discover results with an increased accuracy

Clustering Analysis

- Clustering analysis is a type of data mining technique used to group similar objects or data points together into clusters based on their characteristics or attributes.
- The goal of clustering is to identify natural groupings or patterns in the data that may not be immediately apparent.
- Clustering analysis is the process of discovering groups and clusters in the data in such a way that the degree of association between two objects is highest if they belong to the same group and lowest otherwise.
- One common application of clustering analysis is in customer segmentation, where it is used to group customers with similar characteristics or buying habits together. This information can be used by companies to tailor marketing messages and product offerings to different customer segments.

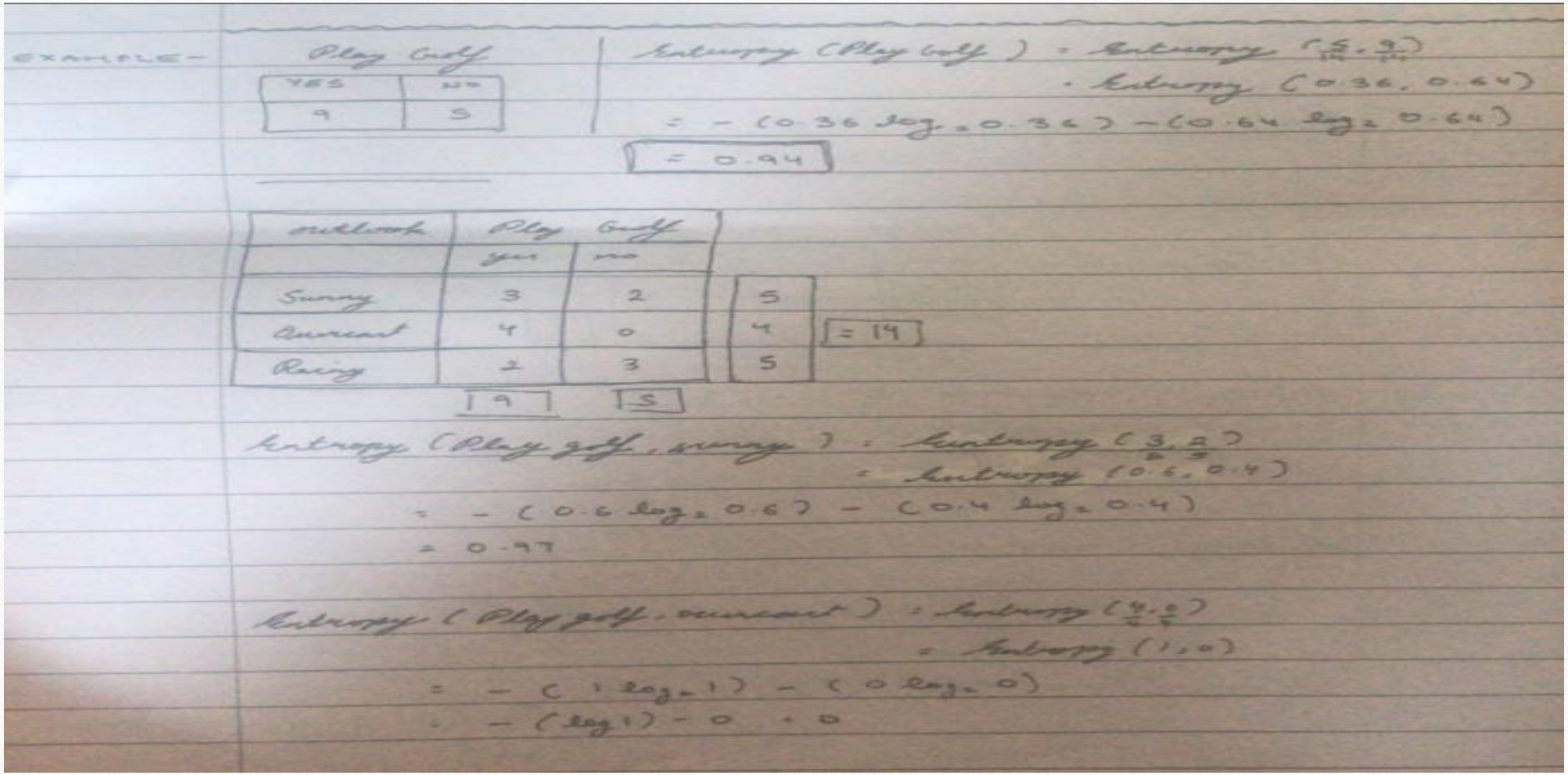
Regression Analysis

- Regression analysis is a data mining technique used to identify the relationship between one or more independent variables and a dependent variable.
- The goal of regression analysis is to understand how changes in the independent variables impact the dependent variable, and to use this information to make predictions about future outcomes.
- This means one variable is dependent on another, but it is not vice versa.
- It is generally used for prediction and forecasting.
- Application of regression analysis is in medical research, where it is used to identify the relationship between various risk factors and a particular disease or health outcome.

Data Mining Applications

- Recommendation Systems: Data mining techniques are used to analyze user behavior and preferences in order to provide personalized recommendations for products and services
- Fraud Detection: Data mining algorithms can be used to detect fraudulent activity in financial transactions
- Customer Segmentation: Data mining techniques are used to group customers with similar characteristics or buying habits together, in order to tailor marketing messages and product offerings to different customer segments.
- Image Classification: Data mining algorithms can be used to classify images based on their content, such as identifying objects, people, and scenes in photographs.
- Sentiment Analysis: Data mining techniques are used to analyze social media and other text data in order to identify the sentiment and opinions of individuals and groups.

Information Gain Examples



EXAMPLE-

Play Golf	
YES	NO
9	5

Entropy (Play Golf) : Entropy $(\frac{5}{14}, \frac{9}{14})$

: Entropy (0.36, 0.64)

$= -(0.36 \log_2 0.36) - (0.64 \log_2 0.64)$

$= 0.94$

outlook	Play Golf	
	yes	no
Sunny	3	2
Overcast	4	0
Rainy	2	3
	9	5

5
4
5

= 14

Entropy (Play golf, sunny) : Entropy $(\frac{3}{5}, \frac{2}{5})$

: Entropy (0.6, 0.4)

$= -(0.6 \log_2 0.6) - (0.4 \log_2 0.4)$

$= 0.97$

Entropy (Play golf, overcast) : Entropy $(\frac{4}{4}, \frac{0}{4})$

: Entropy (1, 0)

$= -(1 \log_2 1) - (0 \log_2 0)$

$= -(\log 1) - 0 \cdot 0$

Entropy (Play golf, Rainy) : Entropy $(\frac{2}{5}, \frac{3}{5})$

: Entropy (0.4, 0.6)

$= -(0.4 \log_2 0.4) - (0.6 \log_2 0.6)$

$= 0.97$

Entropy (Sunny, overcast, rainy) =>

$: \frac{5}{14} \times 0.97 + \frac{4}{14} \times 0.0 + \frac{5}{14} \times 0.97$

$E = 0.68$

✓ **Information gain** $= 0.94 - 0.68$

$: 0.26$

Reference Questions

1. Explain Decision Tree along with diagram and relevant example

2. Write Advantages and Disadvantages of Decision Tree

3. Explain various Data mining Techniques

4. What is Data Mining? And explain Types of Sources of Data in Data Mining

5. What is Data Mining and write Examples of Data mining applications

6. Create Decision Tree for given is the below data and calculate information gain of following