

UNIT 1 – INTRODUCTION TO DATA MINING (Easy English Notes)

What is Data Mining?

- Data Mining means finding useful knowledge from large amounts of data.
- It is also called **Knowledge Discovery in Databases (KDD)**.
- Simply: **Data → Knowledge → Action → Goal**

Example:

- Netflix: collects user ratings (data) → learns what movies you like (knowledge) → suggests new movies (action) → you keep using Netflix (goal).
-

Why Data Mining?

- Because **we have too much data** and not enough knowledge.
- “We are drowning in data but starving for knowledge.”
- Data mining helps us find **important patterns** and **useful information** from big data.

Example:

Google searches, YouTube videos, Tweets – too much data every second!

Data mining helps find only the **useful information**.

KDD Process (Knowledge Discovery in Databases)

Steps in KDD:

1. **Selection** – Choose the needed data from databases.
 2. **Preprocessing** – Clean the data (remove wrong or missing data).
 3. **Transformation** – Convert data into proper format.
 4. **Data Mining** – Apply intelligent methods to find patterns.
 5. **Pattern Evaluation** – Find which patterns are interesting or useful.
 6. **Knowledge Presentation** – Show results using graphs or charts.
-

On What Kind of Data?

1. **Relational Databases** – Tables of rows and columns (e.g., SQL Server, Oracle).
 2. **Data Warehouses** – Collected data from many sources (e.g., D-Mart, Stock Market).
 3. **Transactional Databases** – Records of transactions (e.g., Flipkart orders).
 4. **Other Data Types** – Maps (spatial), multimedia (video, text, image), web data, etc.
-

What Kinds of Patterns Can Be Mined?

Two types:

1. **Descriptive** – Describes data (e.g., trends, clusters).
2. **Predictive** – Predicts future or unknown values (e.g., sales prediction).

Examples:

- **Characterization:** Summarize features of a group.
 - **Discrimination:** Compare two groups (e.g., loyal vs new customers).
 - **Frequent Patterns:** Items that often appear together (milk & bread).
 - **Association:** Relationship between items (if buy computer → buy software).
 - **Correlation:** Find relation strength (TV ads ↑ = sales ↑).
 - **Classification:** Predict categories (spam or not spam).
 - **Regression:** Predict numerical values (house price by size).
 - **Clustering:** Group similar data (similar customers).
 - **Outlier Detection:** Find unusual data (abnormal attendance).
-

Are All Patterns Interesting?

No!

Many patterns may not be useful.

We check **interestingness** using:

- **Objective Measures:** Based on math (support, confidence, lift).
 - **Subjective Measures:** Based on user interest or domain knowledge.
-

Technologies Used in Data Mining

1. **Statistics** – For data analysis and prediction.
 2. **Machine Learning** – Lets computers learn from data.
 - *Supervised learning:* uses labeled data (classification).
 - *Unsupervised learning:* uses unlabeled data (clustering).
 - *Semi-supervised & Active learning:* mix or user feedback.
 3. **Database Systems & Warehouses** – Manage large datasets efficiently.
 4. **Information Retrieval (IR):** Searching and retrieving text or multimedia data (like Google).
-

Applications of Data Mining

1. Business Intelligence (BI):

- Find sales trends, customer preferences, competitor analysis.
- Used in marketing, performance tracking, prediction.

2. Web Search Engines:

- Use data mining to search, rank, and recommend results.
 - Handle huge, fast-changing web data.
-

Data Mining Issues

1. Mining Methodology:

- Handle noisy or incomplete data.
- Combine techniques from different fields (like NLP + Data Mining).

2. User Interaction:

- Easy and interactive interfaces for users.
- Include background knowledge and visualization.

3. Efficiency & Scalability:

- Must work fast on large data.
- Use parallel/distributed algorithms.

4. Diversity of Databases:

- Handle different data types (images, streams, networks).

5. Data Mining & Society:

- Privacy protection is important.
 - Avoid misuse of personal data.
 - “Invisible mining” happens when online stores collect user data automatically.
-

Attributes

- **Attribute:** Property or feature of an object.
- Example:
 - Person → Name, Age
 - Book → Author, Price
 - Computer → Model, RAM

Types of Attributes:

1. Quantitative (Numeric)

- *Discrete*: Countable (e.g., number of students).
- *Continuous*: Measurable (e.g., height, temperature).

2. Qualitative (Categorical)

- *Nominal*: Names or categories (e.g., color).
- *Ordinal*: Ordered values (e.g., 1st, 2nd, 3rd).
- *Binary*: Two values (yes/no).
 - *Symmetric*: both equal (male/female).
 - *Asymmetric*: not equal (positive/negative test).

Extra:

- **Interval**: Difference matters, no true zero (e.g., temperature).
 - **Ratio**: True zero exists (e.g., weight, age).
-



Statistical Measures

- **Mean**: Average value.
Formula: $(\text{Sum of all values}) \div (\text{Number of values})$
 - **Median**: Middle value when data is sorted.
 - If even count \rightarrow average of two middle numbers.
 - **Mode**: Most frequent value.
 - One mode \rightarrow unimodal
 - Two \rightarrow bimodal
 - Three or more \rightarrow multimodal
 - **Range**: Difference between largest and smallest value.
 - **Standard Deviation (σ)**: Shows how much data values vary from the mean.
-



Data Distribution

- **Symmetric**: Mean = Median = Mode
 - **Positively Skewed**: Mean > Median
 - **Negatively Skewed**: Mean < Median
-

Quantiles

- Divide data into equal parts.
 - **Q1 (25%)** → first quartile
 - **Q2 (50%)** → median
 - **Q3 (75%)** → third quartile
 - **IQR (Interquartile Range)** = Q3 – Q1
 - **Five-number summary:** Minimum, Q1, Median, Q3, Maximum
 - **Boxplot:** Graph showing these five numbers with whiskers.
-

Data Matrix vs Dissimilarity Matrix

- **Data Matrix:**
 - Rows → data objects
 - Columns → attributes
 - Example: Student table with Name, Age, Marks
 - Called **two-mode matrix** (objects × attributes)
 - **Dissimilarity Matrix:**
 - Shows distance/difference between pairs of data points
 - Called **one-mode matrix**
-

Dissimilarity of Numeric Data

Used to measure how different two data points are.

1. Euclidean Distance:

$$\sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \dots}$$

2. Manhattan Distance:

$$|x_1 - y_1| + |x_2 - y_2| + \dots$$

3. Minkowski Distance:

$$(|x_1 - y_1|^q + |x_2 - y_2|^q + \dots)^{1/q}$$

4. Supremum Distance:

Maximum of all absolute differences.

Example:

$x = [22, 1, 42, 10]$, $y = [20, 0, 36, 8]$

→ Euclidean ≈ 6.708, Manhattan = 11, Minkowski ($q=3$) ≈ 6.118, Supremum = 6

description of this topic from ppt

Box Plot (in Easy English)

What is a Box Plot?

- A **Box Plot** (also called a **Box-and-Whisker Plot**) is a **graph** that shows how data is **spread or distributed**.
 - It helps us see:
 - the **minimum and maximum values**
 - the **median (middle value)**
 - the **quartiles (Q1 and Q3)**
 - and **outliers (unusual data points)**
-

Parts of a Box Plot

1. **Minimum** → Smallest data value (except outliers)
2. **Q1 (First Quartile)** → Value below which 25% of data lies
3. **Median (Q2)** → Middle value (50% point)
4. **Q3 (Third Quartile)** → Value below which 75% of data lies
5. **Maximum** → Largest data value (except outliers)

Interquartile Range (IQR)

$$\text{IQR} = Q3 - Q1$$

It shows the range of the middle 50% of the data.

Five-Number Summary

A box plot is drawn using these 5 values:

Minimum, Q1, Median (Q2), Q3, Maximum

How to Draw a Box Plot

1. Arrange data in ascending order.
2. Find **Q1, Q2 (Median), Q3**.
3. Calculate **IQR = Q3 – Q1**.
4. Find **lower limit = Q1 – 1.5 × IQR**, and **upper limit = Q3 + 1.5 × IQR**.
 - Values outside this range are **outliers**.
5. Draw a box from Q1 to Q3.

6. Draw a line inside the box at the **Median (Q2)**.
 7. Extend “**whiskers**” to show minimum and maximum values.
-

Diagram (Description) from ppt

|-----|=====|-----|

Min Q1 Median Q3 Max

Or with labels:

Outlier Min Q1 Median Q3 Max Outlier

-----|-----|=====|=====|-----|-----

- The **box** shows Q1 to Q3 (middle 50% of data).
 - The **line in the box** is the **median (Q2)**.
 - The **whiskers** show the range of remaining data.
 - The **stars/dots** (if any) show **outliers**.
-

Example

Data: 5, 7, 8, 12, 13, 14, 18, 21, 23

- 1 **Median (Q2)** = 13
- 2 **Q1** = 8
- 3 **Q3** = 18
- 4 **IQR** = $Q3 - Q1 = 18 - 8 = 10$
- 5 **Lower limit** = $Q1 - 1.5 \times IQR = 8 - 15 = -7$ (no data below this)
- 6 **Upper limit** = $Q3 + 1.5 \times IQR = 18 + 15 = 33$ (no data above this)
- ✓ So no outliers.

Draw box from 8 → 18 with median line at 13.

In Short for Exam

- **Box Plot** shows data spread using **five-number summary**.
 - **IQR** = $Q3 - Q1$
 - **Middle line** = **Median**
 - **Whiskers** = min & max values
 - **Outliers** = values outside $(Q1 - 1.5 \times IQR)$ or $(Q3 + 1.5 \times IQR)$
-

Would you like me to **draw and show an actual box plot diagram image** for this example (so you can visualize it for exam)?

You said:

quantiles

ChatGPT said:

Perfect 🌟 — here's the **Quantiles** topic in **simple English** so you can easily write it in your theory exam.

Quantiles (Easy English Notes)

What are Quantiles?

- **Quantiles** are numbers that **divide your data into equal parts**.
- They show how **data is spread (distributed)** in a dataset.
- Each quantile tells us what **percentage of data** lies **below** a certain value.

 In simple words:

Quantiles help us **understand how values are spread** — from smallest to largest.

Types of Quantiles

Quantiles are named based on **how many parts** the data is divided into:

Quantile Type Divides Data Into Example

Quartiles	4 parts	Q1, Q2, Q3
Deciles	10 parts	D1, D2, ... D9
Percentiles	100 parts	P1, P2, ... P99
2-Quantiles	2 parts	Median (Q2)

Quartiles in Detail

1 Q1 (First Quartile)

→ 25% of data lies **below** Q1.

→ Also called **25th percentile**.

2 Q2 (Second Quartile)

→ 50% of data lies **below** Q2.

→ This is the **Median or 50th percentile**.

3 Q3 (Third Quartile)

→ 75% of data lies **below** Q3.

→ Also called **75th percentile**.

Example

Data (sorted):

5, 7, 8, 12, 13, 14, 18, 21, 23

- **Q2 (Median) = 13**
- **Q1 = value between 7 and 8 = 7.5**
- **Q3 = value between 18 and 21 = 19.5**

So:

$Q_1 = 7.5$

$Q_2 = 13$

$Q_3 = 19.5$

Interquartile Range (IQR)

- $IQR = Q_3 - Q_1$
- It shows the **spread of the middle 50% of the data**.

From example:

$$IQR = 19.5 - 7.5 = 12$$

Five-Number Summary (used in Box Plot)

1. Minimum
2. Q1
3. Median (Q2)
4. Q3
5. Maximum

These five values are used to **draw a Box Plot**.

In Short for Exams

- Quantiles divide data into equal parts.
- Common types: **Quartiles (4 parts), Deciles (10), Percentiles (100)**.
- **$Q_1 = 25\%$, $Q_2 = 50\%$, $Q_3 = 75\%$** .

- **IQR = Q3 – Q1** (spread of middle 50%).
- Used to describe **data distribution** and to make **Box Plots**.