Data Mining & Analytics

Aditya Raj
RA2311003030555
CSE(Core)-I

What is Data Mining?

Think of it as **KDD** (Knowledge Discovery in Databases). Basically, it’s the "clever" part of finding hidden patterns and trends in giant piles of data so businesses can actually make smart, data-driven decisions instead of just guessing.

Analysis vs. Analytics (The "When")

They sound the same, but they look at different things:

- **Data Analysis**: Looks at the **past**. It takes apart what already happened to find out *why* it happened (like reviewing last year’s sales).
- **Data Analytics**: Looks at the **future**. It uses that past data to build models and predict what *will* happen next (like figuring out when to launch a new product to make the most profit).

How Knowledge is Actually Discovered (The KDD Steps)

It’s not magic; it’s a process:

1. **Data Cleaning**: Deleting "noise" and mistakes.
2. **Data Integration**: Merging data from different places.
3. **Data Selection/Transformation**: Picking the right data and formatting it (like scaling numbers).
4. **Data Mining**: Running algorithms to find patterns.
5. **Pattern Evaluation**: Deciding which patterns actually matter.
6. **Knowledge Presentation**: Visualizing everything so people can understand it.

What Kind of Patterns Can We Find?

Characterization: Summarizing general traits of a group (e.g., "Our big spenders are usually 40-50 years old").

Association Analysis: Seeing what things happen together (e.g., people who rent action movies often buy popcorn).

Classification: Sorting data into labeled categories (e.g., flagging a credit card user as "safe" or "risky").

Clustering: Grouping things that are similar when we don’t have labels yet.

Outlier Analysis: Finding the "weird" stuff, like potential fraud.

The "Dirty" Data Problem (Preprocessing)

Real-world data is usually messy—missing values, human errors, or just plain "noise".

- **Cleaning**: We either ignore missing rows or fill them in with averages.
- **Reduction**: Making the data smaller (so it’s faster to process) without losing the main info.
- **Normalization**: Putting everything on the same scale (like 0 to 1) so one big number doesn't mess up the whole math model.

Why it’s Hard to Do (Challenges)

Privacy: It’s creepy if companies know too much about your habits without you knowing.

Performance: Processing terabytes of data is slow and expensive.

Complexity: Real data isn't just numbers; it's also videos, photos, and maps.

Real-World Uses

- **Health**: Detecting fraud in insurance or predicting patient needs.
- **Shopping**: "Market Basket Analysis" helps stores know what to put on sale together.
- **Education**: Predicting how students will perform so they can get help sooner.

THANK YOU