Clinical Data Mining

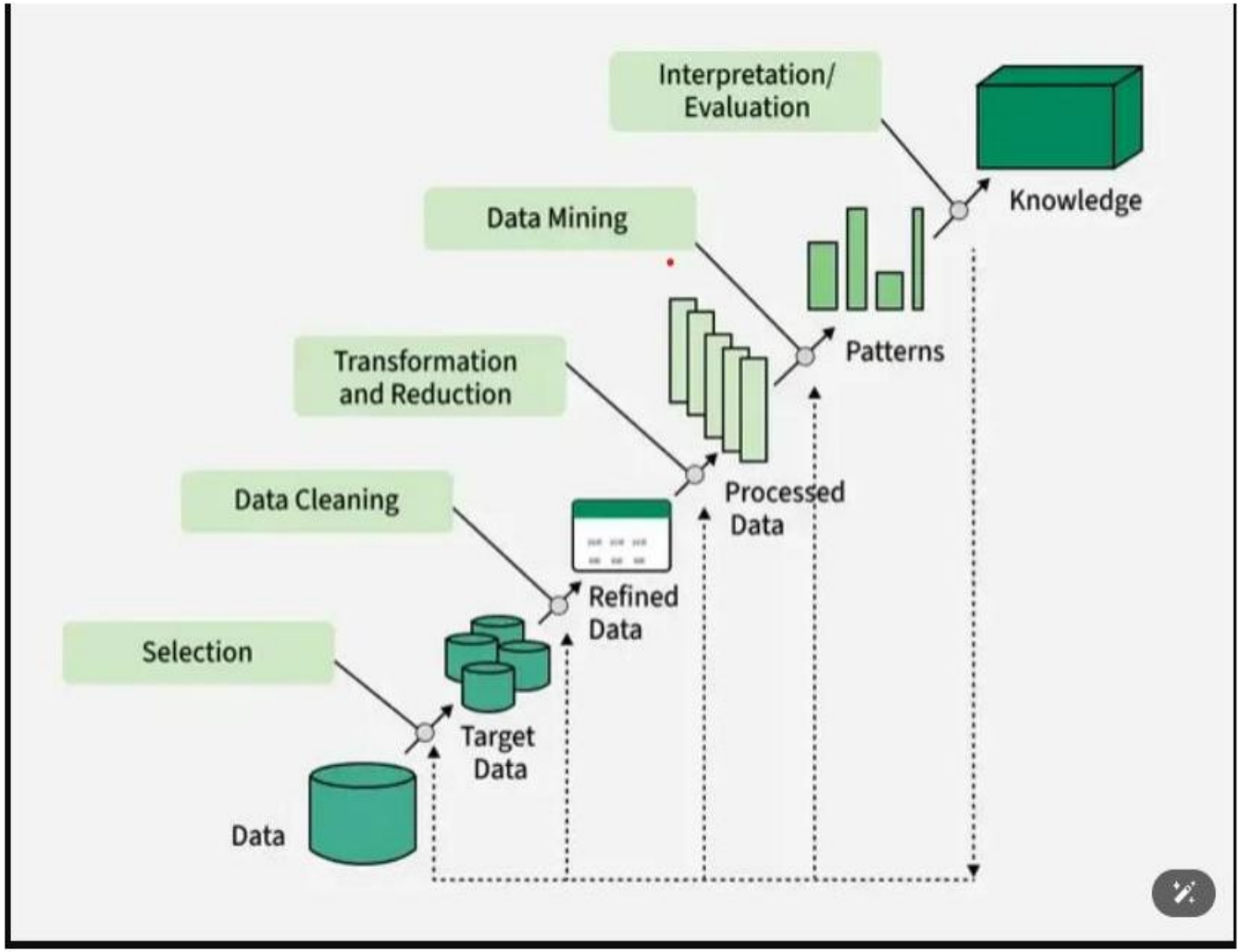
Lecture Five

By

Assist. Lect. Zainab M. Alameen

2025 - 2026

Diagram 1: key step in the Knowledge Discovery in Databases (KDD) process:



Transformation & Feature Selection:

Lecture Keys:

- Introduction.
- Data Transformation.
- Feature Selection.
- Feature Extraction.
- Why Dimensionality Reduction Matters in Healthcare.
- Discussion.

Introduction:

- Clinical datasets are often **high-dimensional**, meaning they contain many attributes (features), such as lab results, demographics, genetic data, or imaging biomarkers. Too many features can cause:
  - Increased computation time
  - Model over fitting
  - Difficulty in interpretation
- Thus, we use **data transformation** and **feature selection/extraction** to simplify data while keeping important information.

Data Transformation:

- **Transformation:** The process of converting data from one format or scale to another to improve model performance or compatibility.

➢ **Common Transformation Techniques:**

Transformation	Description	Example
Normalization	Scale data between 0-1	Glucose levels scaled between 0-1
Standardization	Mean = 0, SD = 1	Z-score transformation
Log Transformation	Reduces skewness in large ranges	Log (Blood_Pressure)
Binning / Discretization	Converts numeric to categorical	Glucose → Low/Medium/High
Encoding	Converts text to numeric	Gender → (Male=0, Female=1)

Feature Selection:

- **Feature Selection:** Selecting the **most relevant** features from the dataset that have the highest impact on prediction or classification tasks.
- **Example:**

  When predicting heart disease, selecting only key features like *Age*, *Cholesterol*, *Blood Pressure*, and *Smoking History* may improve accuracy and interpretability.

Feature Selection:

- **Main Types:**

1. Filter Methods:

- Use statistical measures like *correlation*, *ANOVA*, or *chi-square test*.
- Example: Remove highly correlated lab tests.

2. Wrapper Methods:

- Evaluate subsets of features using model performance (e.g., Recursive Feature Elimination – RFE).
- Example: Select top biomarkers that best predict diabetes.

3. Embedded Methods:

- Feature selection happens during model training (e.g., LASSO regression, Decision Tree feature importance).

Feature Extraction (Dimensionality Reduction):

- **Feature Extraction**: is transforms the original features into a new set of features that summarize the most important information.
- **Techniques:**

A. Principal Component Analysis (PCA): Converts correlated features into a smaller set of uncorrelated components and keeps most variance in fewer dimensions.

- **Example:**

  From 20 lab test results ⟶ reduce to 3 main components representing:

- Liver Function
- Kidney Function
- Blood Cell Activity

B. Linear Discriminant Analysis (LDA): Used for **classification** tasks (supervised) and maximizes separation between disease groups (e.g., healthy vs. diabetic).

C. Auto encoders (Deep Learning-based): Neural networks that compress and reconstruct data, useful in imaging or genomic datasets.

Why Dimensionality Reduction Matters in Healthcare

Benefit	Explanation
Reduces noise	Eliminates irrelevant or redundant medical features
Improves speed	Faster training and prediction
Prevents overfitting	Simpler models generalize better
Improves visualization	Allows plotting complex data in 2D/3D
Enhances interpretability	Focus on clinically meaningful indicators

Discussion

- **Dataset:** Clinical records of diabetic patients with 20 lab tests.

  Goal: Predict risk of kidney failure.

  Steps:

1. Apply **feature selection** to keep only relevant lab tests (e.g., Creatinine, Glucose, GFR).
2. Apply **PCA** to reduce them to 2 main components.
3. Visualize patient clusters — healthy vs risk group.

The End

Thanks for your listening