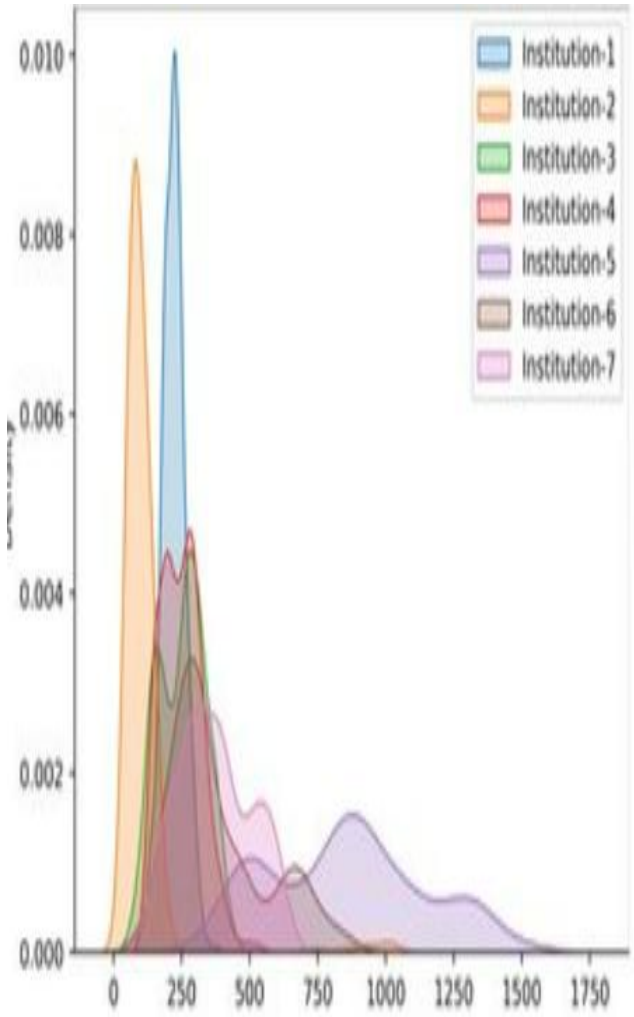
Clinical Data Mining

Lecture Four

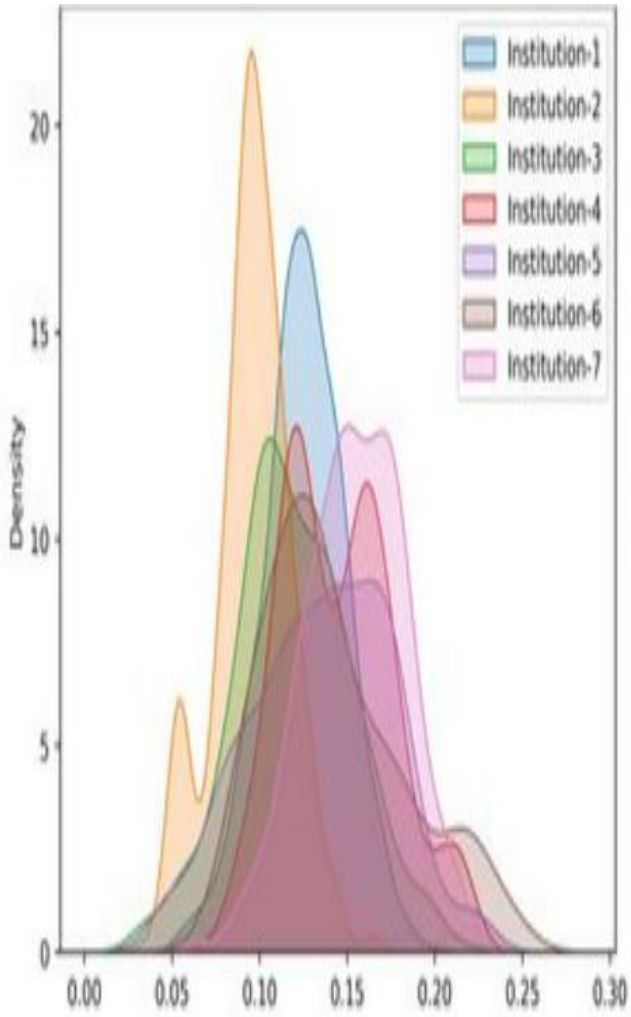
By

Assist. Lect. Zainab M. Alameen

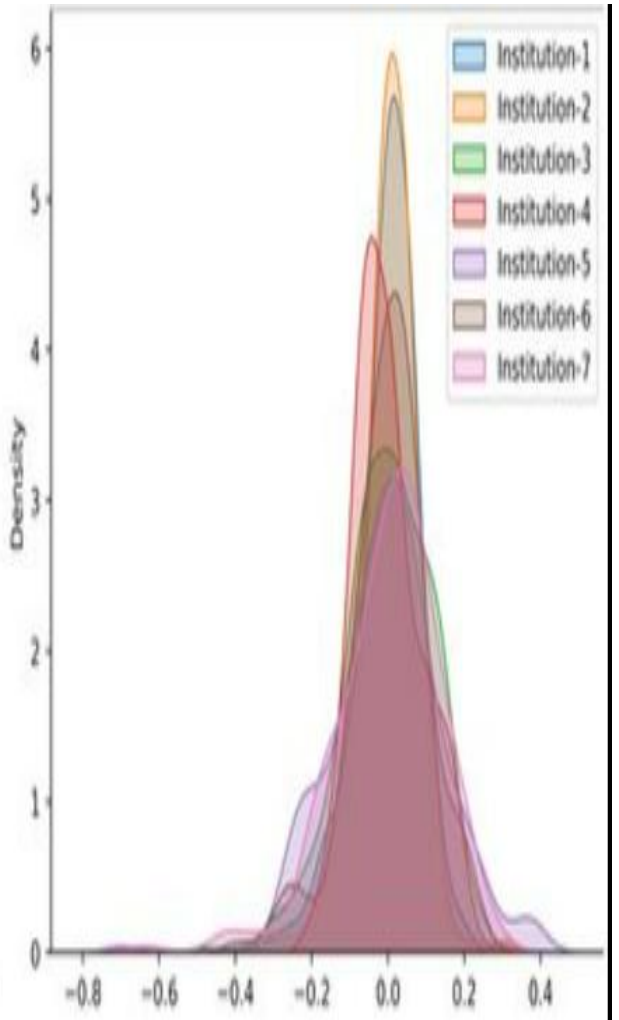
2025 - 2026



Original data distribution



Min-Max normalization



Z-Score normalization

Data Preprocessing and Cleaning II

Lecture Keys:

- Introduction.
- What is Normalization?
- Normalization common types.
- Discretization.
- Types of Discretization.
- Homework (Assignment2).

Introduction:

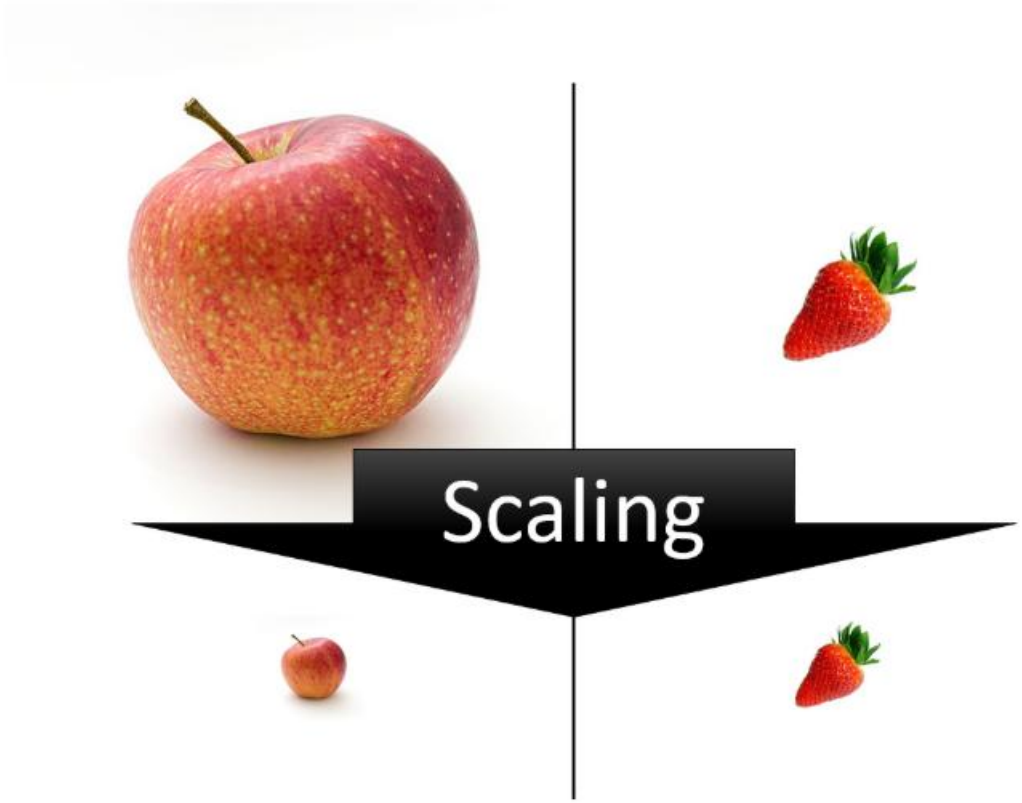
- In clinical data mining, preprocessing is a critical step to ensure data quality and improve the performance of data mining algorithms.

  After handling **missing values**, **inconsistencies**, and **duplicates**, the next essential tasks are:

- **Normalization**
- **Discretization**

What is Normalization:

- **Normalization:** is the process of **scaling numerical data** into a specific range, often **[0, 1]** or **[-1, 1]**, to ensure that variables contribute equally to the analysis.
- **Why it's important in clinical data:**
  - Different lab measurements may have different units (e.g., blood glucose vs. blood pressure).
  - Prevents features with large ranges from dominating the analysis.
  - Improves the performance of algorithms (like k-means clustering or logistic regression).


Scaling

Normalization:

- **Common methods:**

1. Min–Max Normalization

$$X_{norm} = X - X_{min} / X_{max} - X_{min}$$

2. Z-score Normalization

$$X_{norm} = X - \mu / \sigma$$

1. Min–Max Normalization

- **Advantages:**
  - Simple and easy to apply.
  - Keeps all values within the same range.
- **Disadvantages:**
  - Sensitive to outliers (extreme values can distort scaling).
  - **Example (conceptual):**

    Blood glucose levels between 85 and 240 mg/dL will be scaled between 0 and 1.

Min–Max Normalization

Patient_ID	Blood_Glucose	Cholesterol
1	85	180
2	120	220
3	160	200
4	200	250
5	240	300

Patient_ID	Glucose_MinMax	Cholesterol_MinMax
1	0.00	0.00
2	0.18	0.29
3	0.38	0.14
4	0.58	0.50
5	1.00	1.00

2. Z-Score Normalization (Standardization)

- Converts data to have **mean = 0** and **standard deviation = 1**.
- Formula:
- $Xz=(X-\mu)/\sigma$
- μ = mean, σ (sigma) = standard deviation.
- **Advantages:**
- Works well when data contains outliers.
- Useful when algorithms assume normally distributed data.

Z-Score Normalization (Standardization)

- **Example (conceptual):**

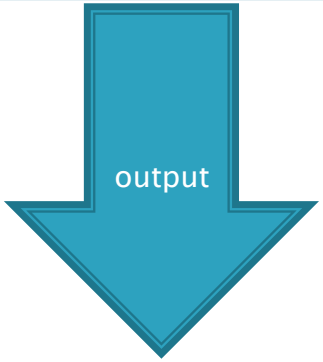
  If a patient's cholesterol level is 250 mg/dL and the mean cholesterol is 200 mg/dL ($\sigma = 40$),

  then:

- **$Z=(250-200)/40=1.25$** (meaning the value is 1.25 standard deviations above the mean).

Z-Score Normalization (Standardization)

Patient_ID	Blood_Glucose	Cholesterol
1	85	180
2	120	220
3	160	200
4	200	250
5	240	300



Patient_ID	Glucose_Zscore	Cholesterol_Zscore
1	-1.26	-1.10
2	-0.63	-0.27
3	0.00	-0.68
4	0.63	0.41
5	1.26	1.63

Comparison Table:

Method	Formula	Range	Sensitive to Outliers	Suitable for
Min-Max	$(X-Xmin)/(Xmax-Xmin)$	0-1	Yes	Neural networks, decision trees
Z-Score	$(X-\mu)/\sigma$	Unbounded	No	Regression, clustering

Discretization

- Discretization is the process of **converting continuous attributes** into **categorical attributes**.
- **Example:**
- Age (continuous) ⟶ Age Groups (0–18, 19–35, 36–60, 60+)
- Blood pressure values ⟶ “Low”, “Normal”, “High”
- **Benefits:**
- Simplifies models and improves interpretability.
- Useful in algorithms that work better with categorical data.
- Helps in identifying trends and patterns in medical datasets.
- **Techniques:**
- **Equal-width binning**
- **Equal-frequency binning**
- **Clustering based discretization**

Data Binning

Grouped into

Large Continuous Data

Small Discrete Bins

WallStreetMojo

Types of Discretization:

- **Equal-Width Binning:**

  The range of values is divided into intervals of equal size.

  Example: glucose 70–200 divided into 3 equal bins ⟶ [70–113.3], [113.3–156.6], [156.6–200].

- **Equal-Frequency Binning:**

  Each bin has (approximately) the same number of samples.

  Example: sorting glucose values and dividing them so each bin has equal number of patients.

- **Custom (Domain-Based) Binning:**

  Bins are based on **medical standards or expert rules**.

  Example:

- < 100 mg/dL ⟶ *Normal*
- 100–125 mg/dL ⟶ *Pre-diabetic*
- ≥ 126 mg/dL ⟶ *Diabetic*

Homework

Assignment Title: Understanding Discretization in Clinical Data

Questions:

- Write a short report discussing the following points:
- **Define** discretization and explain its importance in clinical data mining.
- **Explain** the main types of discretization methods:
- Provide a **real-world clinical scenario** (e.g., blood glucose or blood pressure data) and describe how discretization can make the data more useful for analysis or classification.
- Discuss **advantages and possible limitations** of using discretization in medical data preprocessing.

The End

Thanks for your listening