📊 Binning and Binarization | Discretization Techniques

These techniques help us turn continuous numerical values into simpler, structured categories — making data easier to understand or use in models.

➢ Why Do We Use Binning and Binarization?

- Machine learning models sometimes **perform better** with **categorized (discretized)** data.
- Binning **groups** continuous numbers into **bins/intervals** (like age groups: 0–18, 19–30...).
- Binarization converts values into **0 or 1**, making data **simpler and more uniform**.

➢ What is Discretization (Binning)?

Discretization = Converting continuous numerical values into **discrete categories or bins**.
E.g.,

- Ages: 12, 21, 35 → "Teen", "Adult", "Senior"
- Incomes: Grouped into “Low”, “Medium”, “High”

➢ Types of Binning

1 Equal Width Binning (Uniform Binning)

- Divides the range of values into **equal-sized intervals**
- Simple but can result in **imbalanced bins**

Example: Income 0–100k split into 5 bins of width 20k each

---

2 Equal Frequency Binning (Quantile Binning)

- Each bin has **roughly the same number of data points**
- Based on percentiles (25%, 50%, etc.)

Better for keeping distribution even

---

3 KMeans Binning

- Uses **KMeans clustering** to group similar values into bins
- More **data-driven**, adapts to natural clusters
- Often more accurate but **computationally heavier**

---

4 Custom/Domain-Based Binning

- Manual bins based on **expert knowledge or business rules**
- **E.g.,** Medical risk categories based on blood pressure ranges

---

➢ After Binning: Encoding

After creating bins, we often need to **encode** them into numbers so ML models can use them:

- Label encoding: Bin A → 0, Bin B → 1…
- One-hot encoding: Each bin gets its own binary column

What is Binarization?

- Converts numerical values into **binary format (0 or 1)**
- Based on a **threshold**
    1. If value > threshold → 1
    2. Else → 0

Example: Turn all temperatures > 30°C to 1 (hot), rest to 0
This is useful for:

- Creating flags
- Simplifying features
- Preparing data for binary classifiers

➢ Summary Table

Technique	Purpose	Example
Equal Width Binning	Uniform intervals	0–20, 21–40, 41–60
Quantile Binning	Same number of items in each bin	Quartiles, Deciles
KMeans Binning	Data-based bin grouping	Clustering-based categories
Custom Binning	Manual or domain-specific groups	“Low”, “Medium”, “High”
Binarization	Convert to 0 or 1	Above 50 = 1, else 0
