Handling Mixed Variables in Feature Engineering

Real-world datasets often contain a **mix of categorical and numerical variables**. To prepare them for machine learning models, we must treat them properly and consistently.

➢ What Are Mixed Variables?

- **Numerical Variables**: Quantitative values (e.g., age, income, salary)
- **Categorical Variables**: Qualitative values (e.g., gender, color, product type)

Example:

In a dataset about customers:

- Age, Income → Numerical
- Gender, City → Categorical

➢ How to Handle Them Together

1 Preprocessing Numerical Variables

- **Scaling** is key:
  1. **StandardScaler** (mean = 0, std = 1)
  2. **MinMaxScaler** (range = 0 to 1)
  3. Ensures fair treatment in models like SVM, KNN, and gradient descent-based models

---

2 Preprocessing Categorical Variables

- **Encoding** is needed:

1. **One-Hot Encoding**: Converts categories into binary columns

E.g., Gender → Male: [1,0], Female: [0,1]

2. **Label Encoding**: Assigns integer values (only when order matters)

E.g., Size → Small: 0, Medium: 1, Large: 2

---

3 Combining Both Types

- After encoding and scaling, both variable types can be **combined into a single feature matrix**.
- This allows ML models to treat all inputs uniformly during training.

➢ *Why Is This Important?*

If not handled properly:

- Models may misinterpret categorical variables as numerical
- Features on different scales can dominate or be ignored
- Leads to **bias**, **poor accuracy**, or **model instability**

➢ Best Practices

Variable Type	Preprocessing Needed
Numerical	Scaling (Standard/MinMax)
Categorical	Encoding (One-Hot or Label)
Mixed Dataset	Apply both, then combine
