Outliers in Machine Learning

➢ What are Outliers?

Outliers are **data points that deviate significantly** from the rest of the dataset. They lie far outside the expected range or pattern of the data.

- **Example:**
  In a dataset of student scores where most lie between 40–90, a score of **5 or 1000** would be an outlier.

➢ Why Outliers Matter in ML?

- **Skew results**: They can distort **mean**, **variance**, and statistical assumptions.
- **Mislead models**: Especially sensitive models like **linear regression, k-NN**, or **SVM**.
- **Cause overfitting or poor generalization**.

➢ Common Causes of Outliers:

- Data entry errors
- Measurement errors
- Natural variation or rare events
- Fraud or anomalies (which can sometimes be useful!)

➢ How to Detect Outliers?

1. **Statistical methods**:
   1. Z-Score
   2. IQR (Interquartile Range)
2. **Visualization**:
   1. Box plots
   2. Scatter plots
   3. Histograms
3. **Model-based**:
   1. Isolation Forest
   2. DBSCAN
   3. One-Class SVM

➢ How to Handle Outliers?

- **Remove** (if due to error or clearly irrelevant)
- **Cap or clip** extreme values

- **Transform** data (e.g., log transform)
- **Impute** with more reasonable values
- **Model with outliers in mind** (e.g., use robust algorithms)

➢ Key Takeaway:

Outliers can either be **noise** or **valuable signals** depending on the context. Detecting and handling them wisely ensures **better model performance and reliability**.