Data Preprocessing and Feature Engineering

1. Introduction:

Data preprocessing is the first and most essential step in the machine learning workflow. It involves cleaning and transforming raw data into a usable format. Feature engineering enhances model performance by creating or modifying input features.

2. Data Cleaning:

- Remove duplicate records
- Handle missing values (drop or impute)
- Fix inconsistencies and outliers

Example:

df.dropna(), df.fillna(mean), df.duplicated()

3. Handling Missing Values:

Techniques include:

- Mean/Median/Mode Imputation
- Interpolation
- Using models (KNN imputation)
- Deletion (if data is sparse)

4. Encoding Categorical Variables:

- Label Encoding: Assigns numeric values to categories
- One-Hot Encoding: Creates binary columns for each category

Useful libraries: scikit-learn, pandas

5. Feature Scaling:

- Standardization: (X - mean)/std
- Normalization: Scales data to [0, 1]

Helps models like SVM, KNN, and neural networks perform better.

6. Feature Engineering:

- Combining features
- Creating new ratios or interactions
- Extracting date/time components
- Polynomial features for non-linear patterns

7. Feature Selection:
- Filter Methods: Correlation, Chi-square
- Wrapper Methods: RFE (Recursive Feature Elimination)
- Embedded Methods: Lasso Regression, Tree-based importance

8. Summary:
High-quality data and well-engineered features greatly enhance model accuracy. Proper preprocessing ensures the model learns from meaningful patterns, not noise.