Introduction To Data Preprocessing

Data Preprocessing: It is a crucial step in data analysis and machine learning pipeline. **It involves tranforming of raw data into use-able format data, making it simplified & enchaning the quality of data, which increases the performance.**

Importance Of Data Preprocessing:

- Improves model accuracy
- Reduces complexity
- Handles Missing Values
- Enchances Data quality

Key Steps in Data Preprocessing:

1. **Data Cleaning:** It involves identifying the errors in dataset, Like:
    - Handling Missing values: Techniques like imputation (filling missing data using mean, median, mode) or deletion (removing records with missing values).
    - Removes Duplicates.
    - Outlier Detection and treatment
    - Noise Reduction.
2. **Data Tranformation:** modifies the dataset into a suitable format for analysis.
    - Normalization
    - Standardization
    - Encoding Categorial Variables
    - Feature Engineering.
3. **Data Spiltting:** It involves spiltting the dataset into separate subsets to train and evaluate models effectively.
    - Training Set: Used to train model (70%-80% of dataset).
    - Test Set: Used to evaluate the final model's performance (30% - 20% of dataset is used)
    - This method helps in preventing Overfitting.
4. **Data Normalization:** Normalization is the process of scaling numbers to a common range (usually 0-1) to compare and analyze data easily.
    - Min-Max Scaling: Rescales features to a range from 0 and 1.
        - **Formula: $X' = X - \min(X) / \max(X) - \min(X)$**
    - Z-score Normalization **(Standardization)**: Center the data around zero with standard deviation of one.

        - **Formula: Z = X−μ / σ**
    - Normalization helps improve convergence speed in algorithms.
    - ensures that all features contribute equally to distance calculations.
5. **Data Batching**: refers to dividing dataset into smaller batches during training, which allows for efficient processing and memory management.
    - Memory Management
    - Faster convergence.
6. **Data Shuffling:** Data shuffling involves randomly rearranging the order of samples in a dataset before training.
    - Reduces Bias.
    - Enhances Generalization.
    - **Generally this step is done before spiltting the dataset for train & test.**
7. **Overfitting**: When model learns training set too well but also learns noise and outliers which overall effects the performance like
    - High Training Accuracy
    - Low Testing accuracy.
8. **Underfitting:** When models are too simple and doesn't learn much on training set which effects both the training and test accuracy.
    - Low Training accuracy
    - Low Testing accuracy.