Data preprocessing

_AICVS

Data Preprocessing in Python using Pandas, NumPy, and Scikit-learn (PTS)

Data preprocessing is a critical step in the data analysis and machine learning pipeline to ensure data is clean, consistent, and ready for modeling. Below are the main steps involved, with examples using popular Python libraries.

1. Data Collection and Loading

Load the data into a Python environment using libraries like Pandas.

2. Handling Missing Data

Missing values can significantly affect the model's performance. Common techniques include:

- Removing rows/columns with missing values.
- Imputing missing values with mean, median, or mode.

3. Data Encoding

Convert categorical variables into numerical format using techniques like **One-Hot Encoding** or **Label Encoding**.

.

4. Feature Selection and Engineering

Identify the most important features and/or create new features.

5. Splitting the Dataset

Divide the dataset into training and testing sets.

6. Handling Outliers

Detect and handle outliers using methods like the **IQR Rule** or **Z-Score**.

Questions

1. **What is an outlier in a dataset?**
2. **Name two common techniques to handle missing data.**
3. **What are the common steps involved in data preprocessing?**