Mini-Project: Data Preprocessing and Analysis with Scikit-Learn

For AI Students

December 24, 2024

1 Objective

This mini-project aims to familiarize students with preprocessing numerical and categorical data, detecting and handling outliers, imputing missing values, and creating new features. Students will use a real-world dataset from Kaggle to apply these techniques.

2 Dataset

You will use the **"House Prices - Advanced Regression Techniques"** dataset available on Kaggle. This dataset contains information about house features and their sale prices. You can download the dataset from this link: https://www.kaggle.com/c/house-prices-advanced-regression-techniques/dataHouse Prices Dataset.

3 Project Statement

3.1 Loading the Data

- Load the data from the `train.csv` file.

3.2 Exploratory Data Analysis (EDA)

- Perform exploratory data analysis to understand the structure of the data, the types of features, and missing values.

3.3 Data Preprocessing

- **Rescaling and Standardization**: Rescale and standardize numerical features.
- **Handling Missing Values**: Impute missing values using appropriate methods (e.g., KNN or mean/median).
- **Handling Outliers**: Detect and handle outliers in numerical features.
- **Encoding Categorical Features**: Encode categorical features using techniques like one-hot encoding or label encoding.

3.4 Creating New Features

- Create new features from existing features (e.g., interaction between two features or polynomial features).

3.5 Modeling

- Train a regression model (e.g., Linear Regression or Random Forest) to predict house prices.
- Evaluate the model's performance using appropriate metrics (e.g., RMSE).