Lab Assignment 4 - Dataset Preprocessing

Build **Decision Tree Regression, Linear Regression and Polynomial Regression models** using **Sklearn** for predicting the Violent Crimes Per Population in USA. The objective of this assignment is to experiment with different dataset preprocessing methods.

Steps

1. **Dataset:** Download this racist dataset "Communities and Crime" from the link https://archive.ics.uci.edu/ml/datasets/Communities+and+Crime. The dataset contains 127 features and target (Violent Crimes Per Population).
2. **Dataset Preprocessing**
    - **Loading:** Read dataset file communities.data to the code.
    - **Encoding:** Use one hot encoder to convert any categorical features into numerical features.
    - **Normalization:** Since some numerical features may be in different ranges, each column can be normalized into 0 to 1 using different methods such as scaling, standardizing etc. This part you have to explore.
    - **Missing Value Imputation:** Dataset contain lot of missing values in it. Missing values are represented using "?". Those values can be predicted using different methods such as replace by global constant, mean, median, mode, value from k-nearest sample, etc. Sklearn Decision trees are capable of handling missing values by their own but still you can try all these.
    - **Splitting:** Split the dataset (using train_test_split function) into training and testing sets after the preprocessing. Use 70% for training and 30% for testing set.
3. **Training and Testing Classification Model:** Use **Sklearn** to train different models such as Decision Tree Regression, Linear Regression and Polynomial Regression with different parameters.
4. **Playing with the Model:** You can try different strategies to see whether testing accuracy improves or not. Strategies can be different 1. Normalization methods (such as min-max, z-score, division by maximum etc.) 2. Missing value imputation methods such as replace by global constant, mean, median, mode, value from k-nearest sample 3. Decision Tree parameters (such as max_depth, min_samples_split, min_samples_leaf, max_features, max_leaf_nodes). 4. Removal of some features (feature selection), 5. Different polynomial degrees. Check the model accuracy for the testing data for each setup.

Reference: Web 😊

Suggested Platform: Python: Azure Notebook/Google Colab Notebook, packages such as Numpy, Pandas

Submission: Submit your files in **Single ipython Notebook** in LMS before **Sunday 18th Aug, 11.59 pm.**

Marking: Marking is based on both **performance during the lab hours** as well as **complete submission in LMS**.