# Code Explanation

This document provides a detailed explanation of the Python code for data preprocessing, feature scaling, outlier removal, and modeling using Random Forest Regressor. The code is structured to handle training and testing data, apply transformations, and predict values for a given dataset.

## 1. Data Loading

The function `get\_file(path)` reads a CSV file and loads it into a Pandas DataFrame. The train and test datasets are read using this function.

## 2. Column Naming

The function `set\_col\_name(N, data\_name)` renames the columns based on whether the data is training or testing data. The 'Age' column is missing in the test dataset, as it is the target variable.

## 3. Data Visualization

The function `show\_details\_col(N, text)` plots a boxplot of numerical columns to identify outliers and data distribution.

## 4. Data Scaling

The function `scaler(N)` applies feature scaling using StandardScaler and RobustScaler. StandardScaler is used for general normalization, while RobustScaler is applied to handle outliers.

## 5. Removing Unnecessary Columns

The function `remove\_col(N)` drops non-numeric categorical columns such as 'Sex', 'Site\_of\_recording', and 'Population\_area', which are not needed for modeling.

## 6. Outlier Handling

The functions `remove\_outliers\_iqr(data, threshold)` and `apply\_winsorization(series, limits)` use interquartile range (IQR) and winsorization methods to remove and adjust outliers, ensuring robust model performance.

## 7. Data Normalization

The function `normalize(N, text)` performs the entire data preprocessing pipeline, including:

- Handling missing values  
- Applying feature scaling  
- Removing and replacing outliers  
- Visualizing processed features

## 8. Model Training

The function `Model(X1, X2, y1, y2)` trains a Random Forest Regressor on the training set and evaluates it using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R² score.

## 9. Making Predictions

The function `Test(N, model, output\_path)` applies the trained model on the test dataset to predict the missing 'Age' values and saves the results to a CSV file.

## 10. Feature Importance

The function `show\_feature\_importances(model, X\_train)` plots the importance of each feature as determined by the Random Forest model, helping in feature selection and analysis.

پویا حاجی صادقی