**ML\_Project\_Data\_Preprocessing**

This system implements various techniques like imputation, scaling, encoding, and filtering to ensure the data is clean, standardized, and ready for efficient model training.

**Dataset**

Employee.csv

**Key Components**

This repository contains a data preprocessing system designed to enhance the quality, reliability, and usability of data for machine learning models. It includes various techniques for data exploration, cleaning, analysis, encoding, and scaling to ensure that the data is ready for efficient machine learning processes.

**1. Data Exploration**

In this step, the following tasks were completed:

* Explored the dataset by listing the unique values in each feature and calculating its length.
* Performed statistical analysis and renamed the columns for better clarity and understanding.

**2. Data Cleaning**

This phase focused on addressing the quality of the data by:

* Identifying missing and inappropriate values, followed by applying appropriate treatments.
* Removing all duplicate rows to ensure data integrity.
* Detecting and handling outliers.
* Replacing the value 0 in the "age" column with NaN (Not a Number).
* Addressing null values in all columns using various methods (such as removal or replacing with mean/median/mode).

**3. Data Analysis**

The analysis involved:

* Filtering the data to include only those records with age > 40 and salary < 5000.
* Visualizing the relationship between age and salary through a chart.
* Counting the number of people from each place and representing this data visually.

**4. Data Encoding**

For making categorical data suitable for machine learning algorithms:

* Converted categorical variables into numerical representations using techniques such as one-hot encoding and label encoding.

**5. Feature Scaling**

After encoding the categorical features, the dataset was scaled to normalize the data for better model performance:

* Applied **StandardScaler** to scale the features to have a mean of 0 and a standard deviation of 1.
* Used **MinMaxScaler** to scale the features to a range between 0 and 1.

**Technologies Used**

* Python
* Pandas
* Numpy
* Matplotlib
* Scikit-learn