

Data Preprocessing

Objective: The main objective of this project is to design and implement a robust data preprocessing system that addresses common challenges such as missing values, outliers, inconsistent formatting, and noise.

By performing effective data preprocessing, the project aims to enhance the quality, reliability, and usefulness of the data for machine learning.

Dataset:

https://drive.google.com/file/d/1F3lRf32JM8ejnXq-Cbf9y7fa57zSHGz_/view?usp=sharing

Key Components to be fulfilled:

1. Data Exploration: (Score : 2)
 - Explore the data, list down the unique values in each feature and find its length.
 - Perform the statistical analysis and renaming of the columns.
2. Data Cleaning: (Score : 2)
 - Find the missing and inappropriate values, treat them appropriately.
 - Remove all duplicate rows.
 - Find the outliers.
 - Replace the value 0 in age as NaN Treat the null values in all columns using any measures (removing/ replace the values with mean/median/mode) .
3. Data Analysis: (Score : 2)
 - Filter the data with age >40 and salary <5000 Plot the chart with age and salary
Count the number of people from each place and represent it visually
4. Data Encoding: (Score : 2)
 - Convert categorical variables into numerical representations using techniques such as one-hot encoding, label encoding, making them suitable for analysis by machine learning algorithms.
5. Feature Scaling: (Score : 2)
 - After the process of encoding, perform the scaling of the features using standard scaler and minmax scaler.

Total Score : 10