

# Task 1: Data Cleaning & Preprocessing Report

Dataset Used: Titanic Dataset

Objective:

Learn how to clean and prepare raw data for Machine Learning using Python.

Tools Used:

- Python
- Pandas
- NumPy
- Matplotlib
- Seaborn
- Scikit-learn

Step 1: Data Exploration

- Loaded dataset using pandas.
- Used head(), info(), and describe() to understand structure.
- Checked missing values using isnull().sum().

Step 2: Handling Missing Values

- Filled 'Age' column using median.
- Filled 'Embarked' column using mode.
- Dropped 'Cabin' column due to excessive missing values.

Step 3: Encoding Categorical Variables

- Applied Label Encoding to 'Sex' column.
- Used One-Hot Encoding for 'Embarked' column.

Step 4: Feature Scaling

- Standardized numerical columns (Age, Fare, SibSp, Parch) using StandardScaler.

Step 5: Outlier Detection & Removal

- Visualized outliers using boxplots.
- Applied IQR method to remove extreme values from 'Fare' column.

Conclusion:

The dataset was successfully cleaned and preprocessed.  
It is now ready for Machine Learning model building.