**ASSIGNMENT 1**

**Saurabh Vijay Isane  
122B1F035**

**Exploratory Data Analysis (EDA) and Preprocessing for Titanic Dataset**

**1. Introduction**

Exploratory Data Analysis (EDA) is the process of analyzing datasets to summarize their key characteristics, often with visual methods. Preprocessing prepares data for modeling by handling missing values, encoding categorical features, and normalizing data. This document outlines the steps for EDA and preprocessing using the Titanic dataset.

**2. Steps in EDA**

**2.1 Loading the Dataset**

* The dataset is loaded using the Pandas library.
* The first few rows are displayed to understand the structure of the data.
* info() and describe() methods provide an overview of data types, missing values, and statistical summaries.

**2.2 Checking for Missing Data**

* Missing values are identified using isnull().sum().
* A heatmap is generated to visualize missing values in different features.

**2.3 Handling Missing Values**

* The 'Age' column is imputed with the median value.
* The 'Embarked' column is imputed with the mode value.
* The 'Cabin' column is dropped due to a high percentage of missing values.

**2.4 Dropping Unnecessary Columns**

* The 'Name' and 'Ticket' columns are removed as they do not contribute to predictive modeling.

**2.5 Correlation Analysis**

* A correlation matrix is created to identify relationships between numerical features.
* A heatmap is plotted to visualize correlations.

**2.6 Data Visualization**

* Histograms are used to explore distributions of numerical features.
* Pair plots help identify relationships between key variables.

**3. Data Preprocessing**

**3.1 Encoding Categorical Variables**

* 'Sex' is label-encoded since it has binary categories (Male/Female).
* 'Embarked' is one-hot encoded to convert categorical values into numerical form.

**3.2 Feature Scaling (If Required)**

* Scaling methods like Min-Max Scaling or Standardization can be applied to normalize numerical features.

**3.3 Final Data Check**

* The processed dataset is reviewed to ensure all necessary transformations are complete.

**4. Conclusion**

EDA and preprocessing are crucial steps in data analysis and machine learning. The Titanic dataset was analyzed for missing values, feature importance, and correlations. Preprocessing steps such as handling missing values, encoding categorical variables, and dropping irrelevant columns were applied to prepare the data for modeling.

This structured approach ensures that the dataset is clean, well-prepared, and ready for further analysis or predictive modeling.