Dataset Analysis Report (Task 1)

Dataset Name

Titanic Dataset

Objective

The objective of this analysis is to understand the structure, data types, and overall quality of the Titanic dataset and to evaluate whether it is suitable for machine learning tasks, especially classification.

Dataset Overview

The Titanic dataset contains information about passengers who traveled on the Titanic. Each row represents one passenger, and each column represents a specific attribute such as age, gender, ticket class, and survival status.

- Total Records: 891 rows
- Total Features: 12 columns

Data Types

The dataset contains different types of data:

- **Numerical Data:** Age, Fare, SibSp, Parch
  These columns contain numeric values and are useful for statistical analysis.

- **Categorical Data:** Sex, Embarked, Pclass
  These columns represent categories or groups.

- **Binary Data:** Survived
  This is the target variable with two possible values (0 = Not Survived, 1 = Survived).

Target Variable

- **Survived** is the target variable.
  It indicates whether a passenger survived the disaster or not. This makes the problem a **classification problem**.

Input Features

The input features include Age, Sex, Pclass, Fare, SibSp, Parch, and Embarked. These features are used to predict the target variable.

Data Quality and Missing Values

The dataset contains missing values in some columns:

- Age has missing values.
- Cabin has a large number of missing values.
- Embarked has a few missing values.

These missing values must be handled before applying machine learning models.

Dataset Suitability for Machine Learning

- The dataset is structured and well-organized.
- It represents a real-world problem.
- The target variable is clearly defined.
- Data cleaning and preprocessing are required.

Conclusion

The Titanic dataset is suitable for machine learning, especially for classification tasks. After handling missing values and encoding categorical variables, it can be effectively used to build predictive models.