Dataset Analysis Report – Titanic Dataset

Task: Understanding Dataset and Data Types
Dataset Used: Titanic Dataset
Objective:
The objective of this task is to explore the Titanic dataset, understand its structure, identify different types of data, detect data quality issues, and evaluate its suitability for machine learning.

1. Dataset Overview

The Titanic dataset contains information about passengers who traveled on the Titanic. Each row represents a single passenger, and each column represents a specific attribute such as age, gender, ticket class, and survival status. The dataset is commonly used for classification problems in machine learning.

2. Data Types Identified

After analyzing the dataset using df.info() and manual inspection, the following types of data were identified:

- **Numerical Features:**
  Examples include *Age*, *Fare*, and *SibSp*. These features contain numeric values and can be directly used for mathematical computations.
- **Categorical Features:**
  Examples include *Sex*, *Embarked*, and *Name*. These features represent categories and need to be encoded before applying machine learning algorithms.
- **Binary Feature:**
  The *Survived* column is a binary feature with values 0 and 1, indicating whether a passenger survived or not.
- **Ordinal Feature:**
  The *Pclass* column is an ordinal feature since it has an order (1st class, 2nd class, 3rd class).

3. Target Variable

The target variable in this dataset is Survived.
This column indicates whether a passenger survived (1) or did not survive (0). Since the target variable is binary, this is a classification problem.

4. Missing Values

Upon analyzing missing values using df.isnull().sum(), it was observed that some columns contain missing data. For example, the *Age* and *Cabin* columns contain null values. Missing values can negatively affect model performance, so they must be handled using techniques such as mean/median imputation, mode filling, or removing rows/columns with excessive missing values.

5. Data Distribution and Imbalance

The distribution of the target variable (*Survived*) was analyzed using value_counts(). It was observed that the number of passengers who did not survive is higher than those who survived. This shows that the dataset is imbalanced. Data imbalance can lead to biased predictions and may require resampling techniques.

6. Dataset Size and Suitability

The Titanic dataset contains a sufficient number of records and features for learning and practicing machine learning concepts. Although it is not very large, it is ideal for beginners to understand data preprocessing, feature engineering, and classification models.

7. Conclusion

In conclusion, the Titanic dataset contains a mix of numerical, categorical, binary, and ordinal features. Some columns have missing values that need preprocessing. The target variable is clearly defined as *Survived*, making it a classification problem. Despite slight class imbalance, the dataset is suitable for basic machine learning tasks and exploratory data analysis.