Dataset Analysis Report – Titanic Dataset

The Titanic dataset contains information about passengers who travelled on the Titanic ship. Each row represents one passenger, and each column provides a specific detail about that passenger.

1. The dataset consists of **891 rows and 12 columns**, making it a moderately sized dataset suitable for basic machine learning analysis.
2. Initial exploration was done by viewing the first and last few records of the dataset. This helped in understanding the overall structure and type of data present in each column.
3. The dataset includes different types of data, such as **numerical, categorical, ordinal, and binary** features, which are commonly used in machine learning problems.
4. **Numerical features** such as *Age* and *Fare* represent measurable values that can be used directly for statistical analysis.
5. **Categorical features** include *Sex* and *Embarked*, which represent text-based categories and require encoding before applying machine learning models.
6. The ***Pclass*** column is an **ordinal feature** because it represents passenger class in a specific order (1st, 2nd, and 3rd class), while the *Survived* column is a **binary feature** containing only two values: 0 and 1.
7. Analysis of missing values showed that the *Age* column has several missing entries, and the *Cabin* column contains a large number of missing values that may affect model performance.
8. The **target variable** in this dataset is *Survived*, which indicates whether a passenger survived or not. The distribution of this variable shows a slight class imbalance.
9. Overall, the Titanic dataset is suitable for machine learning tasks after basic preprocessing steps such as handling missing values and encoding categorical data. This analysis highlights the importance of understanding data before building any machine learning model.