Project: Task 1 - Understanding Dataset & Data Types

Dataset Selected: Titanic Dataset Tools Used: Python, Pandas, NumPy

1. Dataset Overview The Titanic dataset contains information about the passengers aboard the Titanic. The goal is typically to predict survival based on passenger characteristics.

Total Rows: 891

Total Columns: 12

2. Data Structure & Types Upon inspecting the dataframe using df.head() and df.info(), the features were categorized as follows:

Numerical: Age (Continuous), Fare (Continuous), SibSp (Discrete), Parch (Discrete).

Categorical: Sex (Nominal), Embarked (Nominal).

Ordinal: Pclass (1st, 2nd, 3rd class).

Binary/Target: Survived (0 = No, 1 = Yes).

Mixed/Text: Name, Ticket, Cabin.

3. Statistical Summary Using df.describe(), we observed:

The average age of passengers is approximately 29.7 years.

The fare prices vary significantly, ranging from 0 to 512, indicating potential outliers.

The survival rate is roughly 38% (mean of the Survived column is 0.3838).

4. Data Quality & Missing Values Significant data quality issues were identified:

Cabin: Contains a very high number of missing values (>77%). This column may need to be dropped or heavily engineered.

Age: Contains roughly 177 missing values. Imputation (filling with median/mean) will be required.

Embarked: Has only 2 missing values, which can be easily filled.

5. ML Readiness

Target Variable: Survived is the clear target for a classification problem.

Imbalance: There is a slight imbalance (approx. 60% died vs. 40% survived), but it is not severe enough to require complex resampling techniques immediately.

Conclusion: The dataset is suitable for Machine Learning after handling missing values in Age and Cabin and converting categorical variables (like Sex and Embarked) into numerical format.