Dataset Analysis Report – Titanic Dataset (Task 1)

The Titanic dataset contains 891 rows and 12 columns, where each row represents a passenger aboard the Titanic. The dataset was loaded using the pandas.read_csv() function. Initial inspection using the head() and tail() methods helped in understanding the overall structure and format of the dataset.

The dataset includes different types of variables such as numerical, categorical, binary, and ordinal features. Numerical features include Age, Fare, SibSp, and Parch. Categorical features include Sex, Ticket, Cabin, Embarked, and Name. The Survived column is a binary feature indicating survival status, while Pclass is an ordinal feature representing passenger class.

Analysis using df.info() revealed missing values in several columns. The Age column has 177 missing values, the Cabin column has 687 missing values, and the Embarked column has 2 missing values. The dataset size is approximately 83.7 KB, making it lightweight and suitable for machine learning experiments.

Statistical analysis using df.describe() shows that the mean age of passengers is approximately 29.7 years. The Fare feature is highly right-skewed with a maximum value of 512, indicating the presence of outliers. The survival rate in the dataset is approximately 38.4 percent, indicating moderate class imbalance.

The target variable for machine learning is Survived. Input features include Pclass, Sex, Age, SibSp, Parch, Fare, and Embarked. Overall, the dataset is suitable for classification tasks but requires preprocessing steps such as handling missing values, encoding categorical variables, and scaling numerical features.