DATASET ANALYSIS REPORT – TITANIC DATASET

The Titanic dataset contains information about 891 passengers and includes demographic, ticket, and travel-related attributes. The dataset consists of 12 features, including both numerical and categorical variables, making it appropriate for supervised machine learning tasks.

Numerical features such as Age, Fare, SibSp, and Parch represent continuous and discrete values, while categorical features include Sex, Embarked, Cabin, and Ticket. The feature Pclass is an ordinal variable representing passenger class. The target variable is Survived, which is binary in nature and indicates whether a passenger survived the disaster.

Initial data exploration using df.info() revealed missing values in the Age, Cabin, and Embarked columns. Statistical analysis using df.describe() provided insights into data distribution, central tendency, and variability. The Cabin column contains a large number of missing values, indicating poor data completeness for that feature.

The dataset shows a mild class imbalance, with more passengers not surviving compared to those who survived. However, the dataset size is adequate for training machine learning models after applying preprocessing steps such as handling missing values and encoding categorical variables.

Overall, the Titanic dataset is suitable for classification-based machine learning problems and serves as a good dataset for understanding data types, structure, and machine learning readiness.