Task 1: Understanding Dataset & Data Types

Dataset Analysis Report

Abhijeet Kumar

1. Dataset Loading and Structure

The Titanic dataset was loaded using the Pandas library in Python. It contains 891 rows and 12 columns, where each row represents an individual passenger and each column describes a specific attribute related to that passenger. Viewing the first and last few records helped in understanding the overall structure of the dataset, including column names, data types, and the kind of information stored in each feature.

2. Identification of Feature Types

After inspecting the dataset, the features were categorized as follows:

Numerical features: Age, Fare, SibSp, and Parch.

Categorical features: Sex, Embarked, Ticket, Cabin, and Name.

Ordinal feature: Pclass, since it represents passenger class in an ordered manner (1st, 2nd, 3rd).

Binary feature: Survived, which indicates whether a passenger survived or not.

3. Data Types, Missing Values, and Summary Statistics

The `df.info()` function revealed that some columns contain missing values. The Age column has several missing entries, Cabin has a large number of missing values, and Embarked has a few missing records. Statistical analysis using `df.describe()` provided useful insights such as mean, median, and standard deviation for numerical columns. For instance, the average age of passengers is around 30 years, and the Fare column shows significant variation, indicating the presence of outliers.

4. Unique Values in Categorical Columns

Examining unique values in categorical columns helped in understanding data distribution. The Sex column contains male and female categories, Embarked includes C, Q, and S ports, and Pclass has three values corresponding to passenger classes. This analysis also highlighted imbalance in some categories, such as a higher number of male passengers compared to females.

5. Target Variable and Input Features

The target variable for machine learning is **Survived**. The input features considered suitable for prediction include Pclass, Sex, Age, SibSp, Parch, Fare, and Embarked. These features have logical relevance to survival outcomes and are commonly used in predictive modeling.

6. Dataset Size and Suitability for Machine Learning

With 891 observations, the dataset is relatively small but sufficient for learning and applying basic machine learning algorithms such as logistic regression, decision trees, and support vector machines. While it is not ideal for deep learning models, it is well suited for classification tasks and concept understanding.

7. Data Quality Observations

The dataset contains missing values that require preprocessing, particularly in the Age and Cabin columns. The Cabin column has too many missing entries and may need to be dropped or heavily engineered. Additionally, the target variable shows class imbalance, with more non-survivors than survivors. Categorical variables also require encoding before applying machine learning models.

Overall, the Titanic dataset is clean, well-structured, and appropriate for introductory to intermediate machine learning tasks, with manageable preprocessing requirements.