AI & ML INTERNSHIP – TASK 1
Understanding Dataset & Data Types

Name: Jeshwanth
Dataset Used: Titanic Dataset
Tools Used: Python, Pandas, Google Colab

1. Introduction

→This task focuses on understanding the structure and data types of a dataset before applying machine learning techniques. The Titanic dataset was analyzed to identify feature types, missing values, and overall suitability for machine learning.

2. Dataset Overview

→The Titanic dataset contains information about passengers such as age, gender, ticket class, fare, and survival status. It consists of multiple numerical and categorical attributes and is commonly used for binary classification problems.

3. Feature Type Identification

→The dataset includes different types of features. Numerical features include Age, Fare, SibSp, and Parch. Categorical features include Sex, Embarked, Cabin, Name, and Ticket. The feature Survived is a binary variable, while Pclass is an ordinal feature representing passenger class.

4. Data Understanding & Statistics

→The dataset was explored using df.info() and df.describe() to understand data types, missing values, and statistical summaries. Columns such as Age and Cabin contain missing values. Statistical measures like mean, minimum, and maximum were observed for numerical features.

5. Target Variable & ML Suitability

→The target variable in the Titanic dataset is Survived, which indicates whether a passenger survived or not. Other features act as input variables. The dataset is suitable for machine learning as it has a clear target variable and meaningful features.

6. Data Quality Issues

→The dataset contains missing values, especially in the Age and Cabin columns. Cabin has a large number of missing entries, which may affect analysis. The target variable Survived also shows class imbalance. These issues need to be handled during preprocessing.

7. Conclusion

→This task helped in understanding the importance of data exploration before applying machine learning models. By analyzing feature types, data quality, and target variables, the dataset's readiness for machine learning was evaluated.