# Dataset Analysis Report

**AI & ML Internship – Task 1: Understanding Dataset & Data Types**

**Datasets Used:**

1. Titanic Dataset
2. Students Performance Dataset (student-por.csv)

**Objective:**

The objective of this task is to understand the structure, data types, and machine learning readiness of the given datasets before applying any machine learning models.

**Dataset Overview:**

- The Titanic dataset contains 891 records with 12 features related to passengers such as age, gender, passenger class, fare, and survival status.
- The Students Performance dataset contains 649 records with 33 features describing students' demographic, family, academic, and social attributes along with their grades.
- Both datasets are structured in tabular format where rows represent observations and columns represent features.

**Data Types and Feature Classification:**

- The datasets consist of both numerical and categorical features.
- Numerical features include age, fare, studytime, failures, absences, and grades (G1, G2, G3).
- Categorical features include gender, school, address, parental jobs, and family status.
- Binary features include survival status, internet access, romantic relationship status, family support, and school support.
- Ordinal features include passenger class (Pclass), studytime, family relationship quality, and health ratings.

**Missing Values and Data Quality:**

- The Titanic dataset contains missing values in columns such as Age and Cabin, which require preprocessing.
- The Students Performance dataset does not contain missing values, indicating good data quality.
- Some categorical variables show imbalance, which may affect machine learning models if not addressed properly.

**Target Variable Identification:**

- For the Titanic dataset, the target variable is Survived.
- For the Students Performance dataset, the target variable is G3, which represents the final grade.
- All remaining columns act as input features for machine learning models.

**Dataset Size and ML Suitability:**

- The Titanic dataset contains 891 samples, and the Students Performance dataset contains 649 samples.
- Both datasets are suitable for basic machine learning algorithms after appropriate preprocessing such as encoding categorical variables and scaling numerical features.

**Conclusion:**

- This task provided a clear understanding of dataset structure, feature types, missing values, and target variables.
- Both datasets are machine-learning ready after preprocessing and suitable for future modeling tasks.