Dataset Analysis Report

Task: Understanding Dataset & Data Types (AI & ML Internship)

1. Introduction

The objective of this task is to understand the structure, data types, and machine learning readiness of datasets before applying any modeling techniques. Proper data understanding helps identify target variables, feature types, missing values, and potential issues such as imbalance or data quality problems.

2. Student Performance Dataset Analysis

The Student Performance dataset contains information related to students' academic results and demographic attributes. It includes both numerical and categorical features such as gender, address, parental education, lunch type, test preparation course, and exam-related attributes. The dataset contains 395 records with 33 features. The target variable selected for analysis is *address*, which represents the residential area of students and can be used for classification-based machine learning tasks. The dataset has minimal missing values and is suitable for predictive modeling after basic preprocessing.

3. Titanic Dataset Analysis

The Titanic dataset consists of passenger information such as age, sex, class, fare, and survival status. For this analysis, the target variable considered is *Age*, which is a numerical variable and can be used for regression-based machine learning tasks. Some features contain missing values, and there is a slight class imbalance in other categorical attributes. Despite this, the dataset is widely used for supervised learning and exploratory data analysis.

4. Feature Types and Data Quality

Both datasets include numerical and categorical variables. Numerical features provide measurable values, while categorical features represent labels that may require encoding. Missing values in certain columns highlight the importance of data preprocessing steps such as imputation before applying machine learning algorithms.

5. Conclusion

The analysis confirms that both datasets are suitable for machine learning applications. Understanding dataset structure, feature types, and target variables ensures better model performance and reliable predictions.