Dataset Analysis Report

Understanding Dataset & Data Types

Introduction

Understanding a dataset is an essential step before applying any machine learning techniques. It helps in identifying the structure, feature types, and potential data quality issues. In this task, the Titanic Dataset and Students Performance Dataset are analyzed using Python and Pandas to evaluate their suitability for machine learning applications.

Dataset Overview

The Titanic Dataset contains passenger information such as age, gender, ticket class, and survival status. It consists of approximately 891 rows and 12 columns.

The Students Performance Dataset includes student background details and exam scores, with around 1000 rows and 8 columns. Both datasets are structured and well-organized.

Feature Types

The datasets include different data types:

- Numerical: Age, Fare, Math Score, Reading Score, Writing Score
- Categorical: Sex, Embarked, Gender, Lunch
- Ordinal: Passenger Class, Parental Education Level
- Binary: Survived, Test Preparation Course

Identifying these features helps in selecting suitable preprocessing techniques.

Target Variable

In the Titanic dataset, Survived is the target variable used for classification.

In the Students Performance dataset, Math Score (or average score) is considered the target variable for regression analysis.

Data Quality Observations

The Titanic dataset contains missing values in columns like Age and Cabin, and also shows class imbalance in survival outcomes. The Students Performance dataset is relatively clean with minimal missing values. Some categorical features require encoding before modeling.

Conclusion

Both datasets are suitable for machine learning tasks after basic preprocessing. The Titanic dataset is appropriate for classification problems, while the Students Performance dataset is suitable for regression analysis. This analysis ensures better model selection and improved prediction performance in future steps.