

Dataset Analysis Report

AI & ML Internship – Task 1

Introduction

This report analyzes the Titanic and Students Performance datasets to understand their structure, data types, and suitability for machine learning tasks.

Dataset Overview

1. The Titanic dataset contains passenger details used to predict survival, making it suitable for classification tasks.
2. The Students Performance dataset contains demographic data and exam scores, suitable for regression tasks.

Feature Types

Features were identified by inspecting column names and values:

- Numerical: Age, Fare, Math Score, Reading Score, Writing Score
- Categorical: Sex, Gender, Race/Ethnicity, Lunch
- Binary: Survived, Test Preparation Course
- Ordinal: Parental Level of Education

Statistical Summary

Descriptive statistics show variability in passenger age and fare in the Titanic dataset and a reasonable distribution of student exam scores.

Target Variable & ML Suitability

1. **Titanic Dataset:**

Target Variable – Survived (Binary Classification)

2. **Students Performance Dataset:**

Target Variable – Math Score (Regression)

Data Quality Observations

- Titanic dataset contains missing values and class imbalance.
- Categorical features in both datasets require encoding.
- Students dataset has minimal missing data.

Conclusion

Both datasets are suitable for introductory machine learning tasks after basic preprocessing

