Task 1 – Dataset Understanding & Data Types

1. Objective

- Understand the structure, data types, and ML readiness of the Titanic Dataset and Students Performance Dataset.
- Identify target variables, feature types, missing values, and preprocessing needs.

2. Tools Used

- Python (Pandas, NumPy)
- Jupyter Notebook

3. Titanic Dataset

- **Rows / Columns:** ~891 rows, 12 columns
- **Target Variable:** Survived (Binary: 0 = No, 1 = Yes)
- **Feature Types:**
  - Numerical: Age, Fare, SibSp, Parch
  - Categorical: Name, Ticket, Cabin, Embarked
  - Ordinal: Pclass
  - Binary: Sex
- **Data Quality Observations:**
  - Missing values in Age, Cabin, and Embarked
  - Cabin has many missing values
  - Some categorical columns have high cardinality (Name, Ticket)
- **ML Suitability:**
  - Suitable for classification
  - Requires preprocessing: handle missing values, encode categorical variables, feature scaling.

4. Students Performance Dataset

- **Rows / Columns:** ~1000 rows, 8 columns
- **Target Variable:** Math Score (Numerical / Regression), or overall performance (Classification)
- **Feature Types:**
  - Numerical: Math score, Reading score, Writing score
  - Categorical: Race/ethnicity, Parental level of education
  - Binary: Gender, Lunch, Test preparation course
- **Data Quality Observations:**
  - No missing values
  - Well-structured and clean
- **ML Suitability:**
  - Suitable for regression or classification
  - Requires encoding of categorical features
  - No major cleaning required.

Conclusion

- Titanic dataset: classification-ready after preprocessing.
- Students Performance dataset: regression/classification-ready with minimal preprocessing.
- Both datasets provide a strong foundation for understanding data structure, types, and ML readiness.

NAME: S. Sri Mullai Harinie