Dataset Analysis Report

Dataset: Titanic – Machine Learning from Disaster (Kaggle)

Domain: Supervised Learning — Binary Classification

Objective: Predict passenger survival status (0 = Died, 1 = Survived)

1.Dataset Overview

The Titanic dataset contains demographic, travel, and fare-related information of passengers aboard the RMS Titanic. It is widely used as an introductory classification dataset for machine learning tasks.

Total Records: 418

Total Features: 12

Target Variable: Survived

Machine Learning Task: Classification (Binary)

2.Feature Information

Numerical: Age, SibSp, Parch, Fare, PassengerId

Categorical: Sex, Embarked, Ticket, Name, Cabin

Ordinal: Pclass (1 → 2 → 3)

Target: Survived

3.Missing Values

Age: 86 missing

Fare: 1 missing

Cabin: 327 missing

Others: 0 missing

Observations:

- Cabin has high missing values and may be dropped or simplified.
- Age requires imputation (median recommended).
- Fare has minimal missing values and can be filled easily.

4.ML Suitability Assessment

Advantages:

- Target variable available.
- Mix of categorical & numerical variables.
- Real-world application.

Challenges:

- Missing data in key fields.
- Categorical encoding required.
- Slight target imbalance.

Overall Suitability:

The dataset is suitable for machine learning after preprocessing (imputation + encoding).

5. Recommended Preprocessing Steps

- Impute missing Age and Fare.
- Drop or transform Cabin.
- Encode Sex, Embarked, and Pclass.
- Normalize numerical features if required.

Conclusion:

The Titanic dataset provides a practical classification problem involving mixed data types and realistic preprocessing challenges.

After handling missing values and categorical encoding, it can be effectively used to build predictive ML models for survival prediction.