Titanic Dataset Analysis Report

1. Dataset Overview

The Titanic dataset contains **891 records and 12 features**, representing passenger information such as demographic details, ticket class, fare, and survival status. The dataset is commonly used for **binary classification** problems in machine learning.

---

2. Feature Types

- **Numerical Features:** PassengerId, Age, SibSp, Parch, Fare
- **Categorical Features:** Name, Sex, Ticket, Cabin, Embarked
- **Ordinal Feature:** Pclass (1st, 2nd, 3rd class)
- **Binary Feature (Target):** Survived (0 = Did not survive, 1 = Survived)

---

3. Target Variable & Input Features

- **Target Variable:** Survived
- **Input Features:** Pclass, Sex, Age, SibSp, Parch, Fare, Embarked

Columns such as **PassengerId, Name, and Ticket** do not directly contribute to prediction and may be excluded or feature-engineered.

---

4. Statistical Summary & Distribution

- Average passenger age is **~29.7 years**, with values ranging from **0.42 to 80 years**.
- The mean fare is **32.20**, indicating the presence of high-fare outliers.
- **61.6% of passengers did not survive**, while **38.4% survived**, showing moderate class imbalance.
- **Male passengers dominate the dataset (64.8%)**, and most passengers embarked from **Southampton (72.4%)**.

---

5. Data Quality Issues

- **Missing Values:**
  - Age: 177 missing values
  - Cabin: 687 missing values (major data loss)
  - Embarked: 2 missing values
- **Categorical features** require encoding before model training.
- Presence of **class imbalance** in the target variable.

---

6. Dataset Size & ML Suitability

With **891 rows**, the dataset is suitable for **traditional machine learning algorithms** such as Logistic Regression, Decision Trees, and Random Forests. However, it is **not suitable for deep learning models** due to limited data size.

---

7. Conclusion (ML Readiness)

The Titanic dataset is well-structured and appropriate for supervised machine learning tasks. Although preprocessing steps such as handling missing values, encoding categorical variables, and addressing imbalance are required, the dataset provides a strong foundation for building and evaluating classification models.