TITANIC DATASET ANALYSIS

1. Introduction

This project focuses on understanding, analysing, and preparing the Titanic dataset for machine learning applications. The primary objective was to explore the dataset structure, identify data types, handle missing values, and assess its suitability for predictive modeling. In addition to exploratory analysis, the project extends to building and evaluating a basic machine learning model to predict passenger survival.

2. Dataset Overview

The Titanic dataset consists of **891 passenger records** with **12 features**, including demographic details, travel information, and survival status. Each row represents an individual passenger, and the dataset contains a mix of numerical and categorical variables.

- **Target Variable:** `Survived` (Binary: 0 = Not Survived, 1 = Survived)
- **Key Features:** Age, Sex, Pclass, Fare, SibSp, Parch, Embarked

3. Exploratory Data Analysis

Initial exploration using dataset previews, structural information, and statistical summaries revealed important insights:

- Numerical and categorical features coexist within the dataset.
- Missing values were present in `Age`, `Cabin`, and `Embarked`.
- Survival distribution indicated class imbalance, with fewer survivors than non-survivors.
- Fare values showed high variance, suggesting the presence of outliers.

This step was crucial for understanding data quality and preparing an effective pre-processing strategy.

4. Data Cleaning and Pre-processing

To make the dataset machine-learning ready, several pre-processing steps were applied:

- Missing values in `Age` were filled using the median.
- Missing values in `Embarked` were filled using the most frequent category.
- The `Cabin` column was dropped due to excessive missing data.
- Irrelevant features such as `Name`, `Ticket`, and `PassengerId` were removed.
- Categorical variables (`Sex`, `Embarked`) were encoded into numerical form.

After pre-processing, the dataset contained no missing values and all features were numerical.

5. Machine Learning Model

A **Logistic Regression** model was used due to its suitability for binary classification problems.

- The dataset was split into **80% training** and **20% testing** data.
- Model performance was evaluated using accuracy, classification report, and confusion matrix.

The confusion matrix showed strong performance in identifying both survivors and non-survivors, with a reasonable balance between correct predictions and misclassifications.

6. Results and Conclusion

The project successfully demonstrated an end-to-end data analysis and machine learning workflow. The Titanic dataset proved to be suitable for predictive modeling after proper pre-processing. The trained model achieved reliable classification performance, highlighting the importance of data understanding and cleaning before model building. Overall, this project showcases practical skills in data analysis, pre-processing, and basic machine learning implementation.