TITANIC DATASET REPORT

"Exploratory Data Analysis and Survival Prediction on Titanic Dataset"

OBJECTIVE

To explore the Titanic dataset, perform data cleaning and visualization, identify patterns affecting survival, and build a predictive model to estimate passenger survival on the test dataset.

DATASET INFORMATION

Source : https://www.kaggle.com/c/titanic/data

Files Used :

- train_csv : for training the model
- test_csv : for prediction
- gender_submission.csv : baseline prediction file

TOOLS & LIBRARIES USED

- Python
- Pandas
- Seaborn
- Matplotlib
- Scikit-learn

DATA CLEANING & PREPROCESSING

- Missing Values Handles:
  - Age
  - Embarked
  - Cabin
  - Fare
- Feature Transformation:
  - Categorical Encoding
  - Convert age from float to int

EXPLORATORY DATA ANALYSIS (EDA)

- **Descriptive Statistics:**
  - Used .info(), .describe(), and .value_counts() to understand data structure
- **Key Visualizations & Observations:**
  - **Pairplot :** Higher class passengers (Pclass 1) and females had higher survival rates.

- **Heatmap :** Pclass and Fare moderately correlated with survival and Age had weak correlation.
- **Histogram of Age :** Most passengers aged between 20–40.
- **Boxplot: Age vs Pclass :** Higher-class passengers were generally older.
- **Scatterplot: Fare vs Age (colored by survival) :** Many survivors paid higher fares and were younger.

MODEL TRAINING

- **Algorithm Used:** Logistic Regression
- **Target Variable:** Survived
- **Features Used:** Pclass, Sex, Age, Fare, Embarked

PREDICTIONS ON TEST DATA

- Applied the same preprocessing to test.csv
- Used trained model to predict survival
- Output saved as titanic_predictions.csv

SUMMARY & FINDINGS

- **Gender:** Females had a significantly higher survival rate.
- **Class:** First-class passengers had better survival odds.
- **Age:** Children and young adults had slightly higher survival.
- **Fare:** Higher fare correlated with survival (often linked to class).
- **Embarked:** Port of embarkation had some influence.

CONCLUSION

The Titanic dataset reveals how socio-economic factors like class and gender influenced survival. With basic preprocessing and machine learning, we can predict survival outcomes with good accuracy. The model outperformed the gender baseline submission from Kaggle.