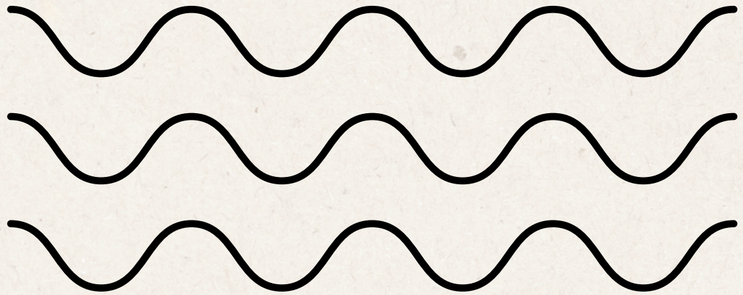
TITANIC CLASSIFIER

Agenda

01	Introduction & Objective
02	Dataset Overview & Columns Description
03	Exploratory Data Analysis
04	Data Cleaning & Null Handling
05	Feature Engineering
06	Random Forest Model & Hyperparameters
07	Model Evaluation
08	Comparison with Decision Tree
09	Deployment (Streamlit App)
10	Conclusion & Key Insights

1. Introduction:

This project aims to predict whether a Titanic passenger survived or not, based on their information.
We used a Random Forest classifier trained on the Titanic dataset.
The project includes data exploration, cleaning, feature engineering, modeling, evaluation, and a Streamlit web app deployment.



2. Dataset Overview:

The dataset is about the Titanic ship disaster (1912).
It contains information about passengers on board.
Source: Kaggle – Titanic: Machine Learning from Disaster.
Size: 891 rows × 12 columns.
Goal: Predict whether a passenger survived or not based on their information.
Target column: Survived (0 = Did Not Survive, 1 = Survived).

2. Columns Description:

01 Survived → target (0 = No, 1 = Yes).

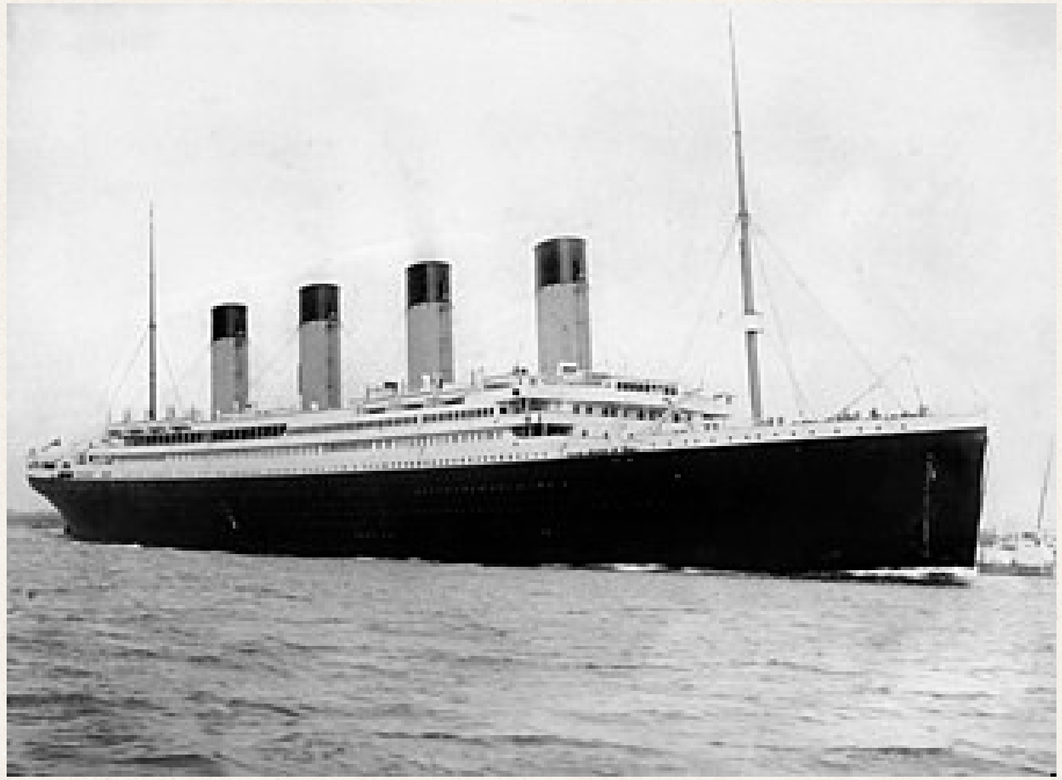
02 SibSp → siblings/spouses aboard.

03 Parch → parents/children aboard.

04 Embarked → port of embarkation (C, Q, S).

05 Pclass → passenger class (1st, 2nd, 3rd).

06 Cabin → cabin number (77% missing → removed).



07 Fare → ticket price.

08 Age → age in years.

09 Sex → male / female.

10 Ticket → ticket number (removed).

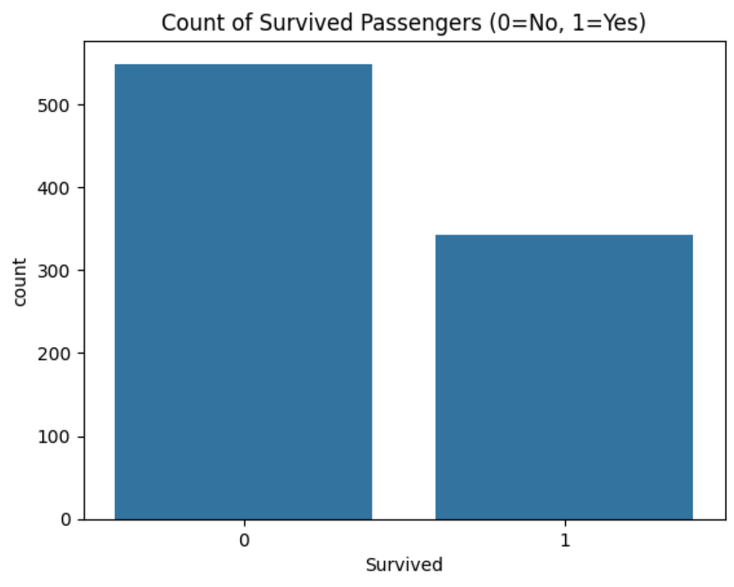
11 PassengerId → unique id (removed, not useful).

12 Name → passenger name (removed).

3.Exploratory Data Analysis (EDA):

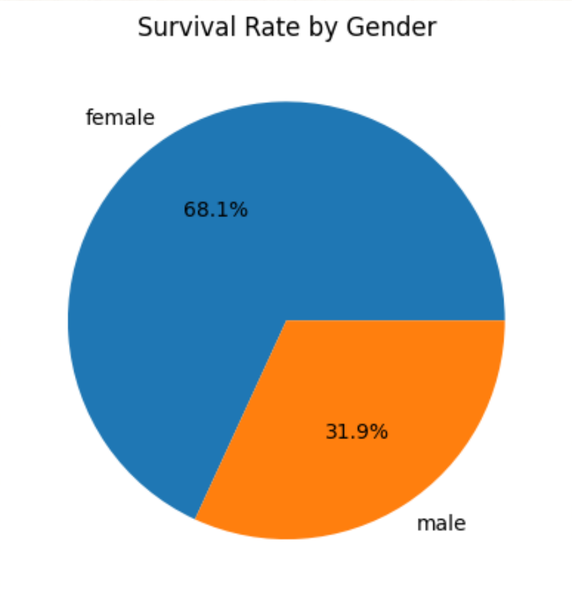
- **Survival counts:**

  -Survived: 342
  -Did not survive: 549
  -Visualized using count plot.



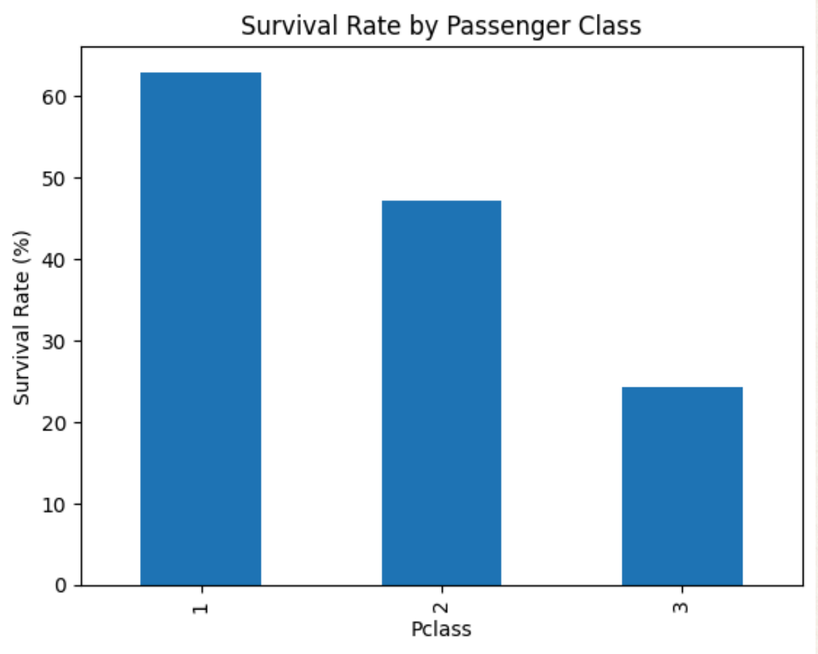
- **Gender distribution in survival:**

  -Female: 68.1% survived
  -Male: 31.9% survived
  -Visualized with a pie chart.



- **Survival across passenger classes (Pclass):**

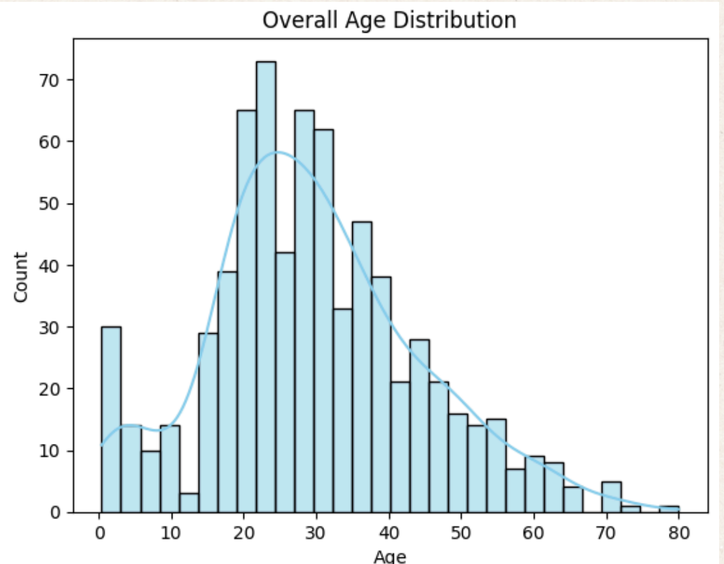
  -Class 1: >60% survived
  -Class 2: >40% survived
  -Class 3: >20% survived
  -Visualized with a bar chart.

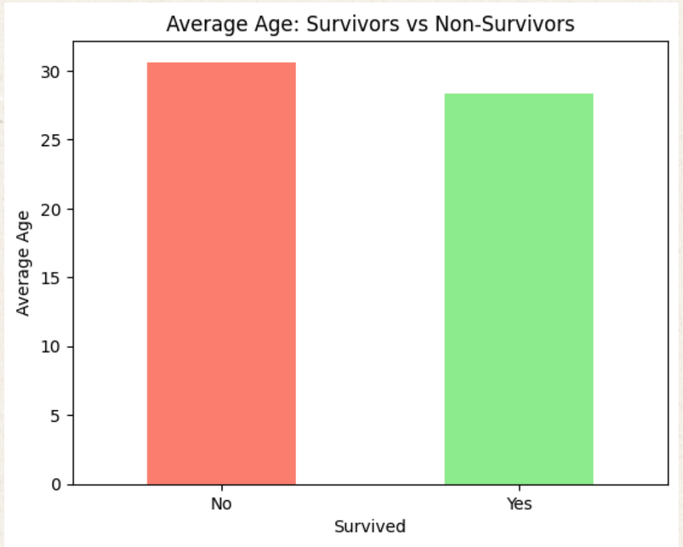


3.Exploratory Data Analysis (EDA):

- **Age distribution:**

-Most frequent ages: 20–25
-Least frequent: 80
-Average age of survivors: 28.34
-Average age of non-survivors: 30.63
-Visualized with histograms and bar plots.



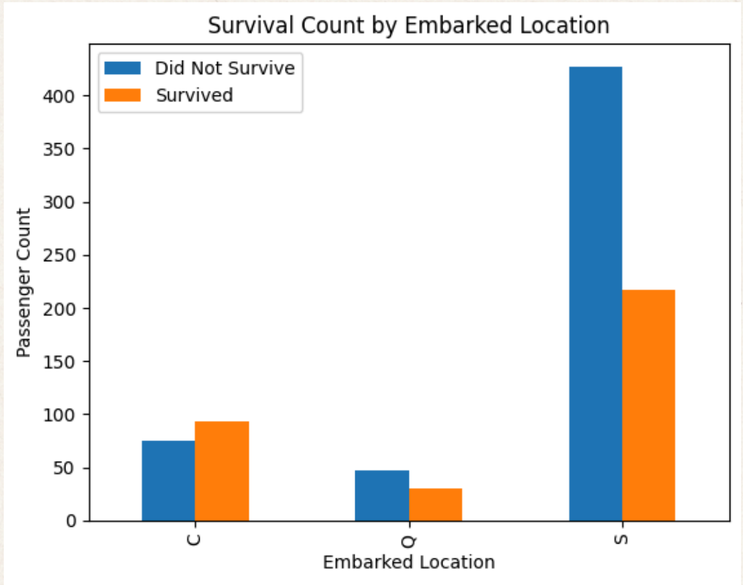


- **Family effect:**
-Passengers with family onboard had higher survival rates.

- **Fare effect:**

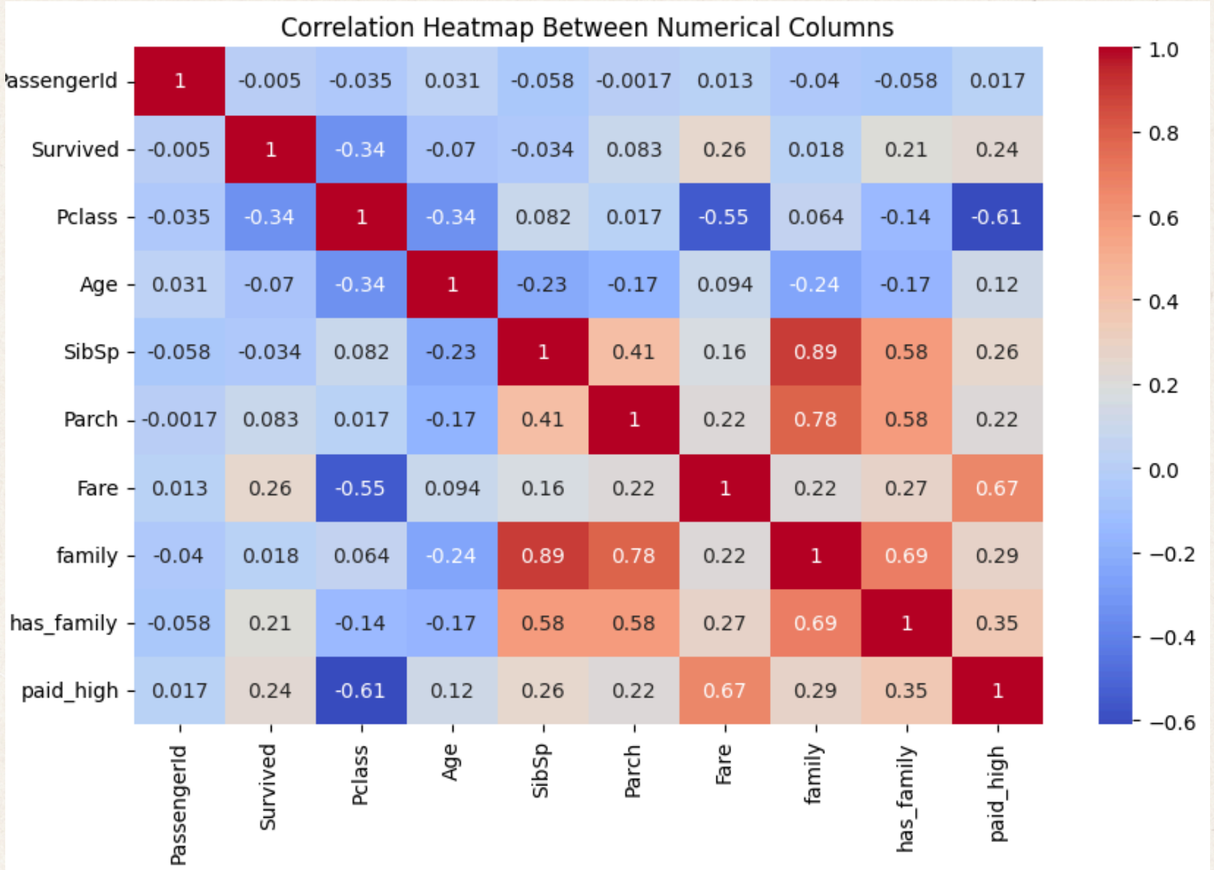
-Higher fare passengers survived more than low fare passengers.

- **Embarked location effect:**
-C: 55.36% survival
-Q: 38.96% survival
-S: lowest survival
-Visualized with bar charts.



3.Exploratory Data Analysis (EDA):

- **heatmap to display the corelation between numerical data:**



-column passenger id doesnt have any strong relation with any column so it isnot an important column ,column survived doesnt have any strong relation with any column because it needs to convert into category datatype ,column pclass has strong relation with Paid_high (-0.61) and Fare (0.55) and it has weak relation with other columns,column Age doesnt have any strong relation with any column (but its an important column) ,column SibSp has strong relation with Family (0.89) and has_family (0.58) and it has weak relation with other columns ,column Parch has strong relation with Family (0.78) and has_family (0.58) and it has weak relation with other columns ,column Fare has strong relation with Paid_high (0.67) and pclass (-0.55) and it has weak relation with other columns,column Family has strong relation with SibSp (0.89) ,has_family (0.69) and Parch (0.78) and it has weak relation with other columns ,column has_family has strong relation with SibSp (0.58) ,Parch (0.58) and family (0.69) and it has weak relation with other columns ,column paid_high has strong relation with pclass (-0.61) and Fare (0.67) and it has weak relation with other columns

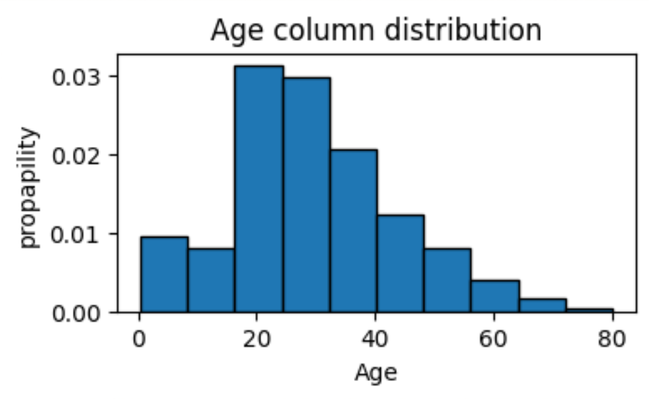
4.Data Cleaning & Null Handling:

- Null values:

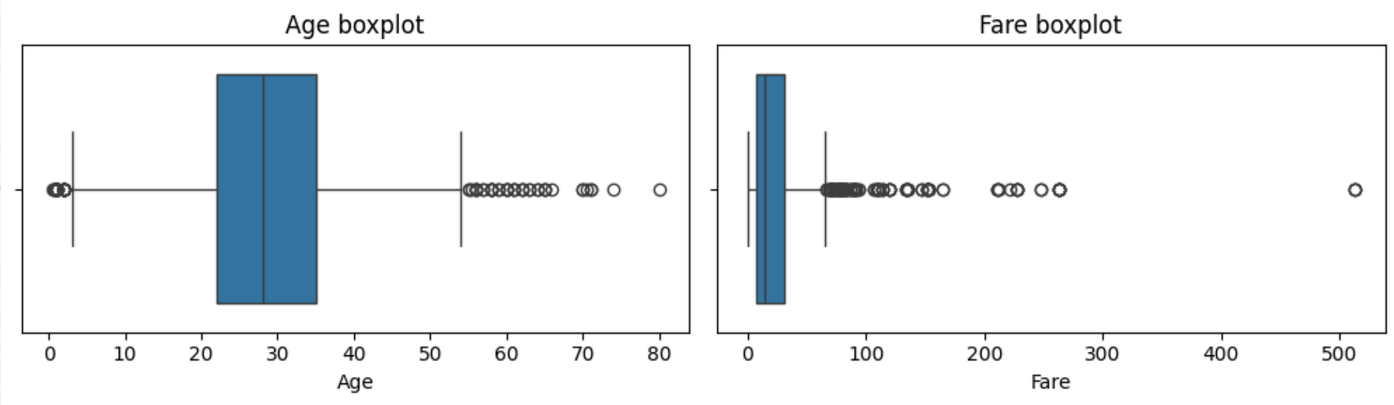
-Embarked nulls dropped (0.22%)

-Cabin dropped (77.1% missing)

-Age nulls filled with median(because it doesn’t have a normal distribution



- Duplicates: none
- Outliers: Detected in Age and Fare, removed successfully



- Unnecessary-columns removed: PassengerId, Name, Ticket.
- Data type conversion: categorical columns → Pclass, SibSp, Parch, Sex, Embarked, Survived.

5.Feature Engineering :

- Created new columns during EDA: family, has_family, age_group, paid_high
- Dropped them afterward to avoid redundancy
- Encoded categorical features using LabelEncoder: Sex, Embarked
- Normalized numerical features with MinMaxScaler (values scaled 0–1)

6.Random Forest Model & Hyperparameters:

- Split data: 80% training, 20% testing
- Random Forest Classifier:

n_estimators = 500

max_depth = 4

max_features = 'sqrt'

- Accuracy: 0.8371 → best performance
- Checked overfitting/underfitting:

Training Accuracy: 0.8439

Testing Accuracy: 0.8371

- Difference is small → no overfitting/underfitting

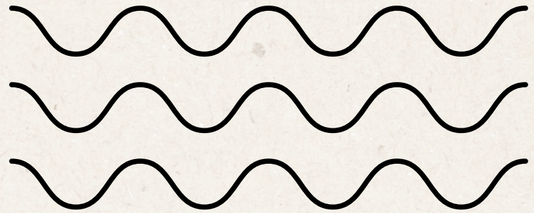
- Hyperparameter Tuning:

n_estimators effect:

10 → 0.8371

100 → 0.8258

500 → 0.8371



max_depth effect:
2 → underfitting

4 → best performance

10 → overfitting

None → overfitting and worst testing accuracy

7.Model Evaluation:

- FIRST 10 PREDICTIONS: 1
  WRONG PREDICTION (ROW 5)

- CONFUSION MATRIX:

-TRUE NEGATIVES: 99

-TRUE POSITIVES: 50

-FALSE POSITIVES: 10

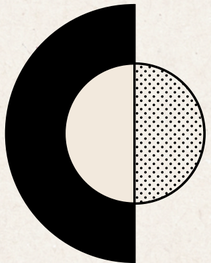
-FALSE NEGATIVES: 19

- METRICS:

-PRECISION: 0.8333

-RECALL: 0.7246

-F1 SCORE: 0.7752



- FEATURE IMPORTANCE:

-SEX: 0.4599

-PCLASS: 0.1642

-FARE: 0.1619

- CROSS-VALIDATION:

-SCORES: [0.7640, 0.8202, 0.8315, 0.7921, 0.8531]

-AVERAGE ACCURACY: 0.8122 → MODEL IS STABLE, GENERALIZES WELL

8.Comparison with Decision Tree:

Decision Tree Accuracy: 0.8202

Random Forest Accuracy: 0.8371 → better because ensemble of trees reduces overfitting

9.Deployment (Streamlit App):

Model saved as titanic_model.pkl using joblib

Streamlit app for user input:

Enter passenger details → predict survival

Displays probability (%) and Gauge meter visualization

Includes emojis for visual feedback (😊 Survived, 😢 Did Not Survive)

10.Conclusion & Key Insights:

Key Findings from EDA:

Female passengers had higher survival rate (68.1%) than males (31.9%).
Passengers in 1st class survived more than 2nd and 3rd class.
Children (<16 years) had the highest survival rate (59%).
Passengers with family onboard or who paid higher fares had higher chances of survival.
Embarkation location mattered: C > Q > S in survival rates.

Model Insights

Random Forest performed better than a single Decision Tree.
Best parameters: n_estimators = 500, max_depth = 4.
Top 3 important features: Sex, Pclass, Fare.
Accuracy: ~83.7%, F1 Score: 0.775 → model is reliable and generalizes well.

Deployment:

Model deployed using Streamlit for interactive predictions.
Users can input passenger info and get survival probability, with emoji feedback (😊 Survived, 😢 Did Not Survive).

Conclusion:

The project successfully predicts survival of Titanic passengers with good accuracy.
Provides insights about which factors influenced survival the most.
Demonstrates a full ML pipeline: EDA → Cleaning → Feature Engineering → Modeling → Evaluation → Deployment.

THANK YOU