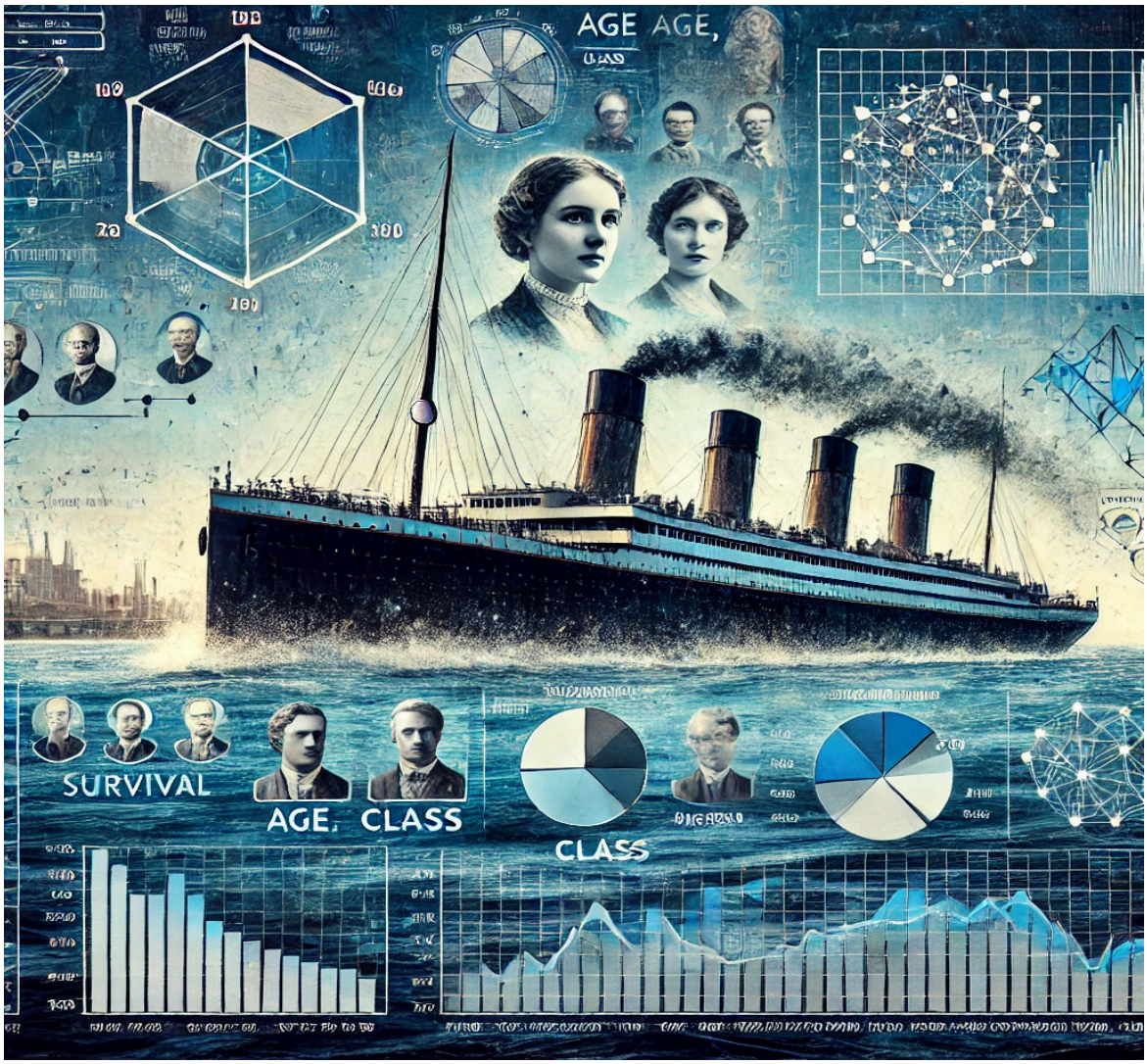
Predicting Titanic Survivors

-Dinakaran.k

Overview:



This project aims to predict **Titanic survivors** using **Machine Learning (Random Forest Classifier)**. The dataset includes **passenger details such as age, gender, class, fare, and family size**, which help determine survival likelihood.

Key Features:

- **Dataset Used:** Titanic Dataset (`Titanic-Dataset.csv`)
- **Data Preprocessing:** Handled missing values and engineered new features.
- **Feature Engineering:** Created **FamilySize** and **IsAlone** features for better predictions.
- **Machine Learning Model:** Implemented **Random Forest Classifier** for survival prediction.
- **Performance Metrics:** Evaluated accuracy, precision, recall, and F1-score.
- **Custom Passenger Prediction:** Allows input of new passenger details to predict survival.

Workflow & Implementation:

1. Data Preprocessing

- **Handled Missing Values:**
    - Replaced **missing age values** with the **median age**.
    - Filled missing **embarkation points** with the most frequent value.
    - Replaced missing **fare values** with the median.

2. Feature Engineering

- **Family Size:** `SibSp + Parch` (Total number of relatives on board).
- **IsAlone:** 1 if no family members, else 0.
- **One-Hot Encoding:** Converted **Sex** and **Embarked** into numeric values.

3. Model Training & Evaluation

- **Algorithm Used: Random Forest Classifier**.
- **Train-Test Split:** 80% training, 20% testing.
- **Feature Scaling:** Standardized numerical values for better model performance.

4. Model Performance

- **Accuracy: 82.7%**
- **Precision: 80.3%**
- **Recall: 77.0%**
- **F1-Score: 78.6%**

Real-Time Prediction for a New Passenger

Input Passenger Details:

```
Pclass: 1
Age: 32
SibSp: 1
Parch: 0
Fare: 50
FamilySize: 1
IsAlone: 0
Sex_male: 1
Embarked_Q: 0
Embarked_S: 1
```

Prediction Output:

```
The passenger is predicted to have survived.
```

Challenges and Solutions:

- **Challenge:** Missing values for age and embarkation.
    - **Solution:** Used **median imputation** for missing age and **mode imputation** for embarkation.
- **Challenge:** Feature selection for better accuracy.
    - **Solution:** Created **FamilySize** and **IsAlone** features to enhance the model.
- **Challenge:** Overfitting risk in Random Forest.
    - **Solution:** Used **regularization techniques** and **limited tree depth**.

Progress and Next Steps:

Accomplishments:

- Successfully trained a **Titanic survival prediction model**.
- Achieved **82.7% accuracy** using **Random Forest**.
- Implemented **custom passenger survival predictions**.

Next Steps:

- Improve prediction accuracy by **hyperparameter tuning**.
- Implement **other classifiers (Logistic Regression, XGBoost)** for comparison.
- Deploy as a **web application for interactive predictions**.

Conclusion:

The **Titanic Survivor Prediction Model** efficiently predicts passenger survival using **Random Forest Classifier**. This project demonstrates how **data preprocessing, feature engineering, and machine learning** can be applied to real-world datasets for meaningful predictions.