Title page

Titanic Survival Prediction Project

This project delves into predicting passenger survival on the ill- fated RMS Titanic using machine learning. We leverage a rich dataset from the Kaggle Titanic competition, focusing on demographic and socioeconomic features to uncover patterns.

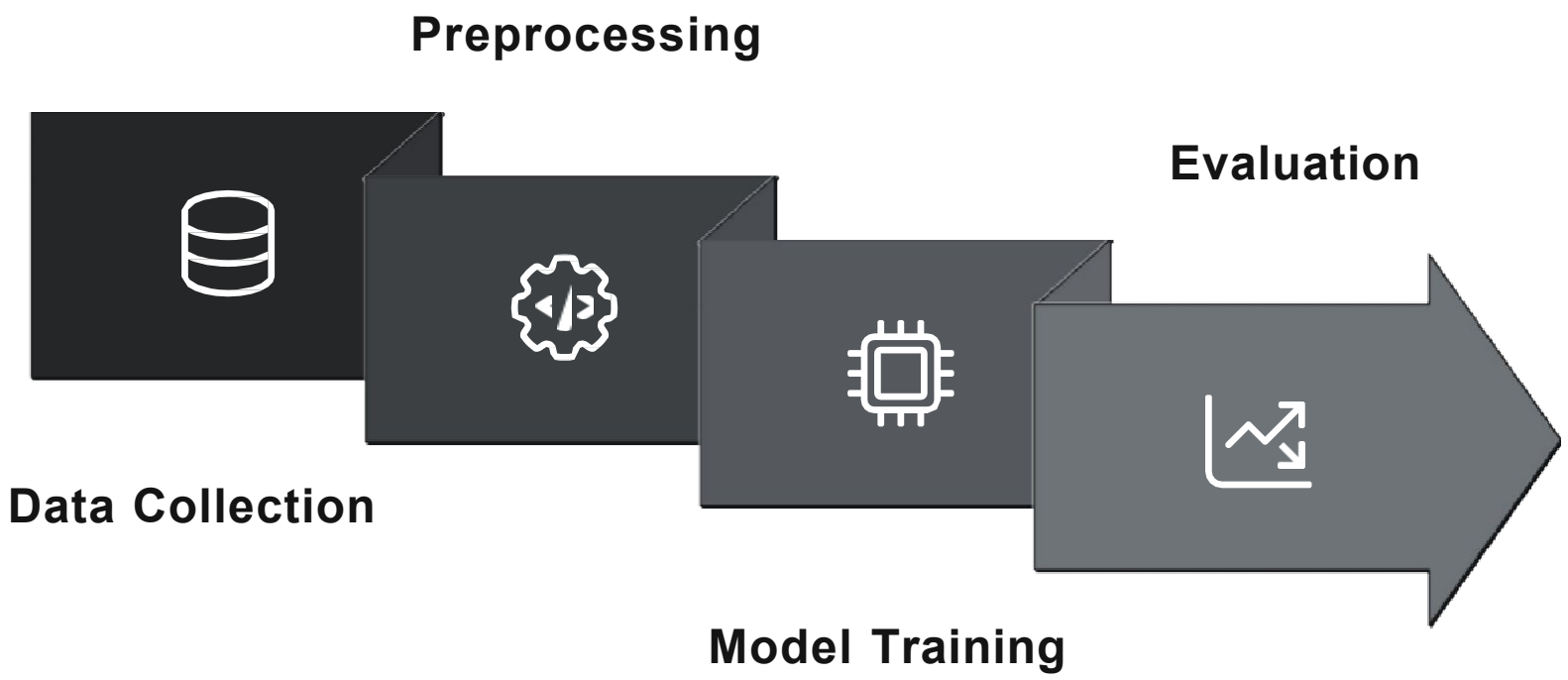
Presented by : P.Dhanush

Date :02/08/2025

Course : AI&DS

Abstract

Our primary objective was to predict Titanic passenger survival using various machine learning techniques. We utilized key features such as age, sex, passenger class, fare, and family size. After rigorous testing, both Random Forest and Logistic Regression models proved effective, with the Random Forest model achieving approximately 84% accuracy on the test data.



Introduction

The sinking of the RMS Titanic in 1912 remains one of history's most tragic maritime disasters, claiming 1,502 lives out of 2,224 passengers and crew. Understanding the factors influencing survival provides crucial historical insight and demonstrates the power of machine learning in predictive modeling. The Kaggle Titanic dataset serves as an ideal benchmark for this predictive analytics exercise.

Objectives

Model Development

Develop a robust predictive model for survival classification.

Factor Analysis

Analyze key factors influencing survival rates on board the Titanic.

Data Preparation

Perform comprehensive data cleaning and effective feature engineering.

Performance Evaluation

Evaluate model performance and interpret the results effectively.

Tools and Technologies Used

Programming S Environment

- Python programming
- language
  Jupyter Notebook or
- Google Colab
-
-

Key Libraries

Pandas: Data manipulation and analysis

NumPy: Numerical operations

Matplotlib & Seaborn: Data visualization

Scikit-learn: Machine learning algorithms

Algorithms Explored

Logistic Regression

Random Forest

XGBoost

Dataset Description

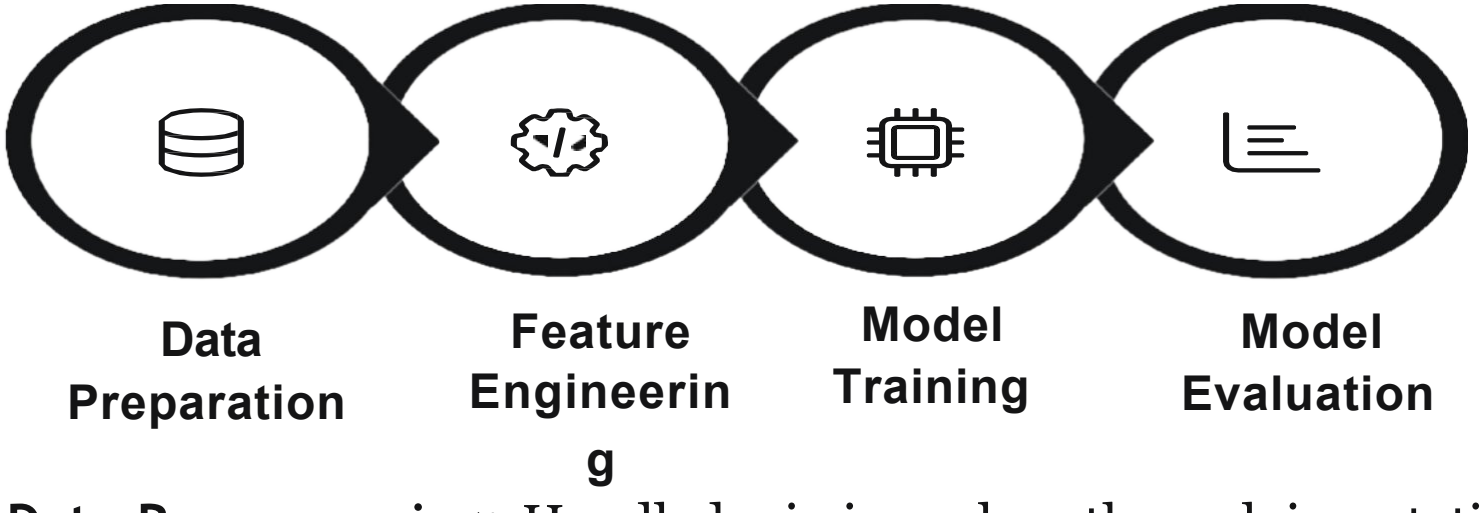
The project utilizes the Kaggle Titanic dataset, comprising both training (train.csv) and testing (test.csv) files. The training set includes 831 samples, each with 12 crucial features that describe passenger attributes and survival status.

PassengerId	Numeric	Unique identifier for each passenger.
Survived	Binary	Target variable (0 = No, 1 = Yes).
Pclass	Categorical	Passenger class (1st, 2nd, 3rd).
Sex	Categorical	Gender of the passenger.
Age	Numeric	Age in years (approx. 20% missing).
Fare	Numeric	Passenger fare.
Cabin	Categorical	Cabin number (approx. 77% missing).
Embarked	Categorical	Port of embarkation (2 entries missing).

Notable missing values were identified in 'Age', 'Cabin', and 'Embarked', which required careful handling during preprocessing.

Methodology

Our approach involved a structured machine learning pipeline to ensure accurate predictions and reliable insights.



Data Preprocessing: Handled missing values through imputation and outlier detection.

Feature Engineering: Extracted titles from names and computed family size for richer insights.

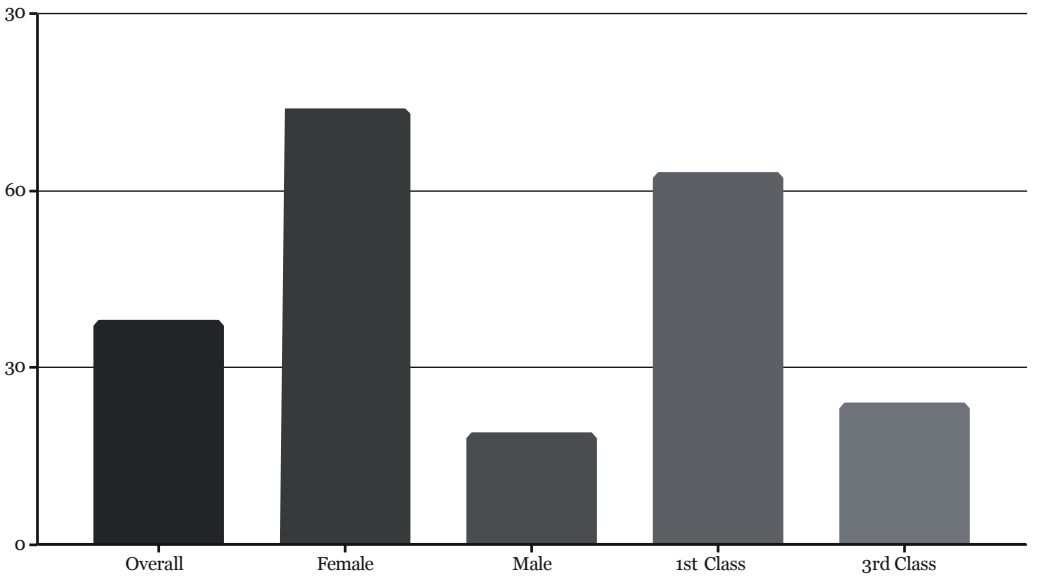
Encoding: Converted categorical variables (Sex, Embarked) into numerical representations.

- **Data Split:** Divided the dataset into training and validation sets to prevent overfitting.
- **Model Training:** Trained various models and fine-tuned hyperparameters for optimal performance.

Evaluation: Assessed models using accuracy scores and confusion matrices.

Exploratory Data Analysis (EDA)

Our EDA revealed crucial insights into survival patterns, highlighting the significant impact of demographic and socioeconomic factors.



Overall Survival Rate: Only about 38% of passengers survived.

- **Survival by Sex:** Females had a significantly higher survival rate (approximately 74%) compared to males (around 13%), reflecting the "women and children first" protocol.

- **Age Distribution:** Children generally exhibited higher survival rates, emphasizing protective measures for younger passengers.

Visualizations such as bar charts and histograms provided clear insights into these distributions.

- **Survival by Pclass:** First-class passengers had the highest survival rate (around 63%), while third-class passengers had the lowest (about 24%), indicating socioeconomic disparities.

Results and Discussion

The Random Forest model demonstrated the highest predictive accuracy among the tested algorithms, confirming the influence of key features.

Random Forest Model Accuracy: ~84%

This indicates a strong capability to predict survival based on the features utilized.

74%

Female Survival

Females consistently showed the highest survival rates across all models.

63%

1st Class Survival

Passengers in first class had significantly higher survival odds.

24%

3rd Class Survival

Third-class passengers had the lowest survival rates, despite their larger numbers.

Key Predictors: Sex, Pclass, Age, Fare, and Family Size were identified as the most influential factors.

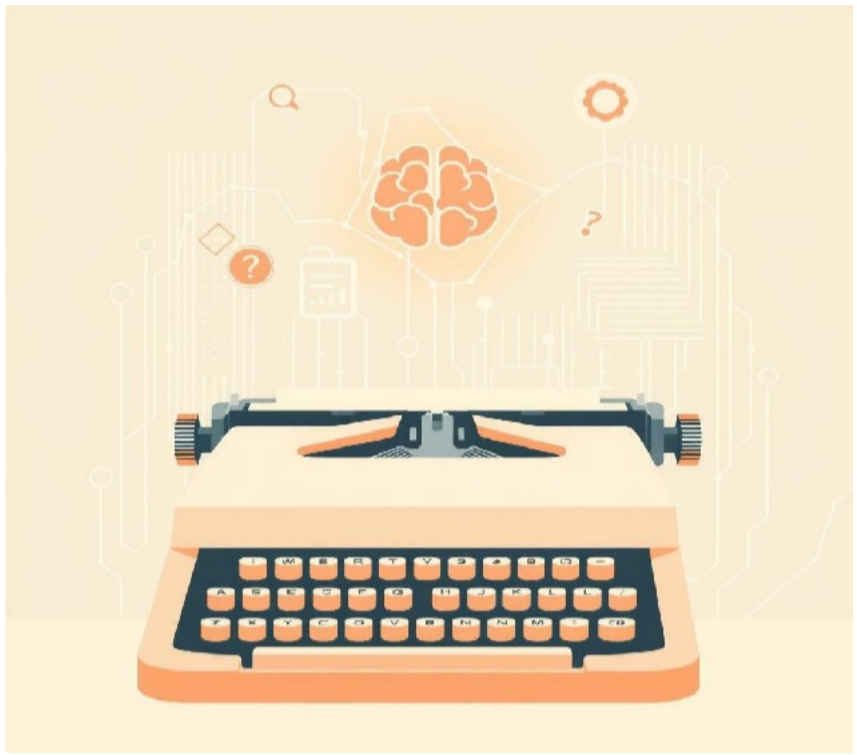
Model Limitations: Challenges included handling extensive missing data and accounting for unobserved factors (e.g., specific lifeboat assignments).

Potential Improvements: Future work could involve more advanced feature engineering, exploring ensemble methods, or leveraging deep learning architecture.

Conclusion and Future Scope

Conclusion

We successfully developed a predictive model for Titanic passenger survival with reasonable accuracy, confirming historical survival patterns tied to demographics and socioeconomic status. The project highlights the utility of machine learning in historical analysis.



Future Scope

- **Deep Learning Integration:** Explore advanced neural networks for potentially higher accuracy.
- **Additional Datasets:** Incorporate other historical maritime disaster datasets for broader insights.
- **Real-time Prediction:** Develop a real-time prediction system for modern maritime safety applications.

This project serves as a foundational step towards more complex and impactful predictive analytics in historical and contemporary contexts.