Titanic Survival Prediction

Santosh Kumar
Data Analyst / Data Science Intern
Machine Learning Project

Problem Statement

The Challenge

Predict whether a passenger survived the Titanic disaster using passenger demographic and travel data.

Primary Goal: Build an accurate and explainable machine learning model that can identify survival patterns and provide actionable insights from historical passenger data.

Dataset Overview

Data Source

Kaggle Titanic Dataset - one of the most popular datasets for classification problems

Sample Size

891 passenger records with complete survival outcomes

Key Features

Passenger class, Age, Gender, Fare, Family relationships, Port of embarkation

Target Variable

Survived: Binary outcome (0 = No, 1 = Yes)

Data Cleaning

01

Missing Value Imputation

Handled missing values in Age using median imputation and filled missing Embarked values with the most frequent port

02

Feature Elimination

Dropped Cabin column due to 77% missing data, which would introduce excessive bias

03

Column Optimization

Removed non-predictive columns including Ticket numbers, raw Name strings, and PassengerId to reduce noise

Feature Engineering

New Features Created

FamilySize

Combined SibSp and Parch to capture total family members aboard, revealing survival patterns based on family dynamics

Title Extraction

Extracted social titles (Mr., Mrs., Miss, Master) from passenger names to capture social status and age group indicators

Categorical Encoding

Applied One-Hot Encoding to transform categorical variables into numerical format for model compatibility

Exploratory Data Analysis

Key Survival Patterns Discovered

Gender Impact

Females showed significantly higher survival rates (74%) compared to males (19%), reflecting "women and children first" protocol

Class Advantage

First-class passengers had 63% survival rate versus 24% in third class, indicating socioeconomic influence on outcomes

Age Factor

Children under 12 demonstrated higher survival probability, benefiting from prioritized evacuation procedures

Family Dynamics

Passengers with 1-3 family members survived more than solo travelers or large families, suggesting optimal group support

Model Building

Model Selection Strategy

Logistic Regression

Baseline linear classifier for interpretable probability estimates and feature importance analysis

Random Forest Classifier

Ensemble method capturing non-linear patterns through multiple decision trees with bagging

Support Vector Machine

Kernel-based classifier optimizing decision boundaries in high-dimensional feature space

Training Strategy: 80/20 train-test split with cross-validation for robust performance assessment

Best Model Selection

Random Forest Emerged as Winner

Highest Accuracy

Achieved 83% accuracy on test set, outperforming other algorithms by 4-7 percentage points

Non-linear Relationships

Successfully captured complex interactions between features like age, gender, and class without manual feature crosses

Overfitting Resistance

Ensemble averaging and max-depth constraints prevented overfitting, maintaining consistent validation performance

Model Evaluation

Metrics Used

- Accuracy: Overall correctness
- Precision & Recall: Class balance
- F1-Score: Harmonic mean
- Confusion Matrix: Error analysis
- ROC-AUC: Threshold performance

Deployment & Customization

Interactive Filtering

Built a user-friendly prediction interface allowing stakeholders to explore survival probabilities across different passenger segments.

Passenger Class Filter

First, Second, or Third class selection

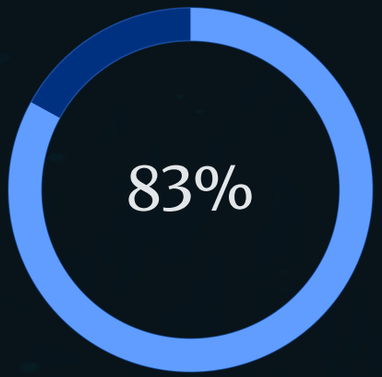
Age Range Slider

Customizable age brackets for analysis

Gender Selection

Male or Female demographic filtering

Key Outcomes



Model Accuracy

Strong predictive performance on unseen data

Pipeline Complete

End-to-end ML workflow from data to deployment

Deployment Platform: Streamlit provides interactive real-time predictions with an intuitive interface for non-technical users

Thank You

Questions & Discussion

Thank you for your time and attention. I'm happy to discuss the methodology, results, or potential improvements to this machine learning solution.

Santosh Kumar | Data Analyst / Data Science Intern