Survival Prediction on the Titanic Dataset: A Machine Learning Approach

Elaborated By :
Fatma BEN REJEB

Contents

Google Colab Link

1 Introduction

This project explores the application of machine learning models to predict survival outcomes from the Titanic dataset. The process includes exploratory data analysis (EDA), data preprocessing, model selection, and evaluation. The goal is to build models that can accurately predict whether a passenger survived the Titanic disaster based on key features.

2 Problem Statement

The RMS Titanic sank in 1912, and many factors influenced whether a passenger survived, such as age, sex, and passenger class. The task is to use historical data to build predictive models that estimate the likelihood of survival. This classification problem serves as a benchmark for data science and machine learning techniques.

3 Data Overview

The dataset contains detailed information about passengers aboard the Titanic, including demographic, ticketing, and survival-related variables. Below is a list of all the variables in the dataset along with their descriptions:

- **Pclass** – Ticket class (a proxy for socio-economic status):
  1 = First class, 2 = Second class, 3 = Third class
- **Survived** – Target variable indicating survival status:
  1 = Survived, 0 = Did not survive
- **Name** : Full name of the passenger, including title (e.g., Mr., Mrs., Miss)
- **Sex** : Gender of the passenger (male or female)
- **Age** : Age in years (can include decimals for children, e.g., 0.83)
- **SibSp** : Number of siblings or spouses aboard the Titanic
- **Parch** : Number of parents or children aboard the Titanic
- **Ticket** : Ticket number
- **Fare** : Fare paid for the ticket (in British pounds)
- **Cabin** : Cabin number

- **Embarked** – Port of embarkation:
  C = Cherbourg, Q = Queenstown, S = Southampton
- **Boat** : Lifeboat number the passenger escaped in
- **Body** : ID number assigned to recovered bodies (blank if body not found)
- **Home.dest** : Final destination of the passenger

During the Exploratory Data Analysis (EDA) phase, it was found that several columns contained a large proportion of missing values or were not relevant to the survival prediction task. Therefore, only a subset of meaningful and complete features — `pclass`, `sex`, `age`, and `survived` — were retained for further analysis and model building.

4 Methodology

4.1 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) was conducted to understand the structure of the dataset and identify potential issues such as missing values or irrelevant features. Several visualizations were created to explore the distributions and relationships between variables, including age histograms, survival counts, class and gender bar charts, and a heatmap for missing values.

During this phase, columns with a significant amount of missing data —such as `cabin`, `boat`, and `body`—were identified and removed. Variables deemed irrelevant to the prediction task, like `name`, `ticket`, and `home.dest`, were also excluded from further analysis. We kept only the relevant columns: *pclass*, *sex*, *age*, and *survived*. The `age` column contained a small number of missing values, and these rows were dropped to maintain the quality and consistency of the dataset.

4.1.1 Plots

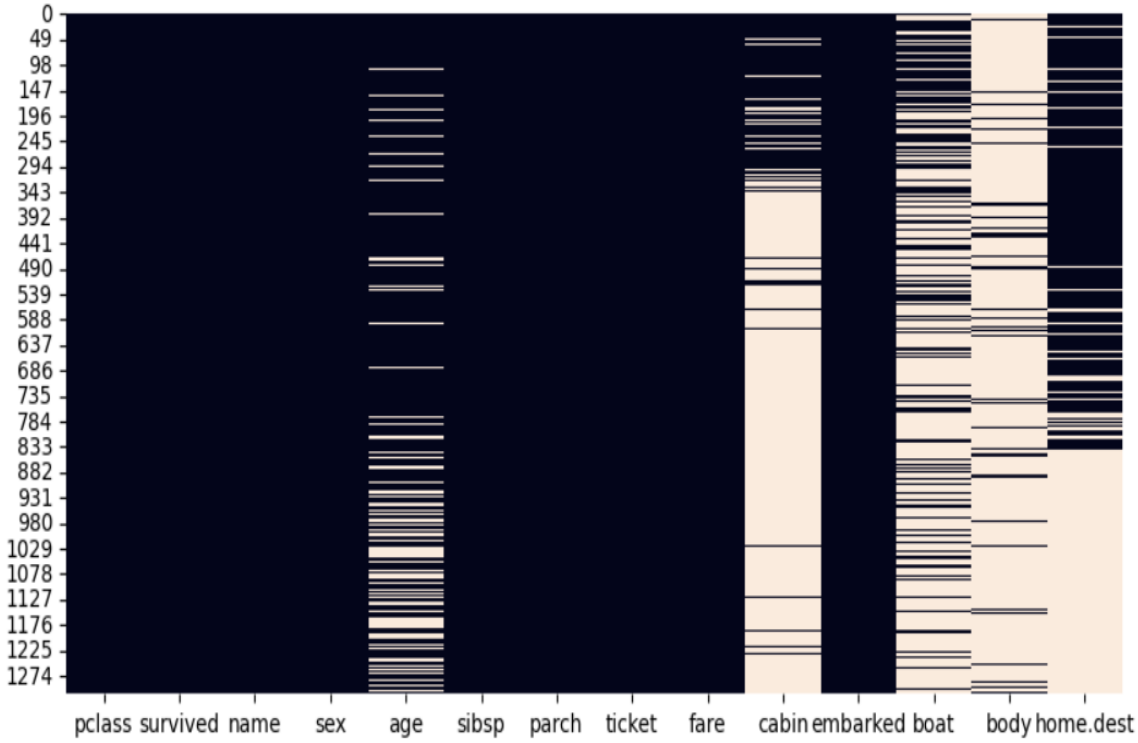

Figure 1: Heatmap Showing Missing Values Across Features

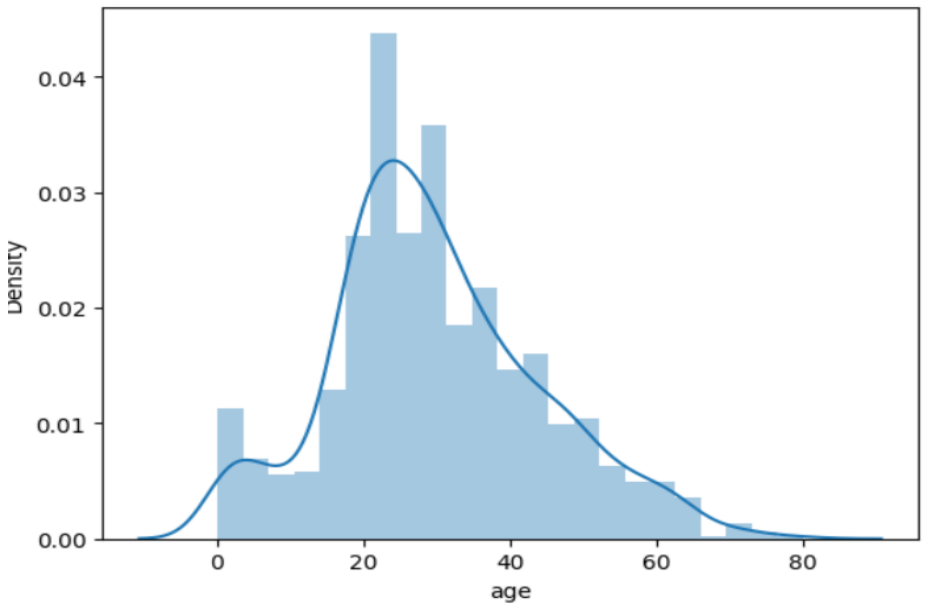

Figure 2: Distribution of Passenger Ages

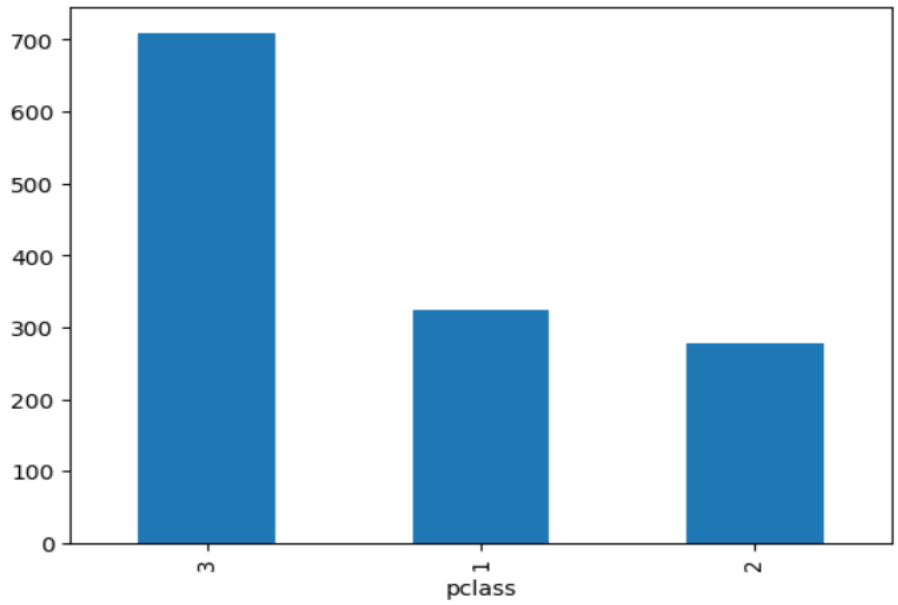

Figure 3: Passenger Count per Ticket Class

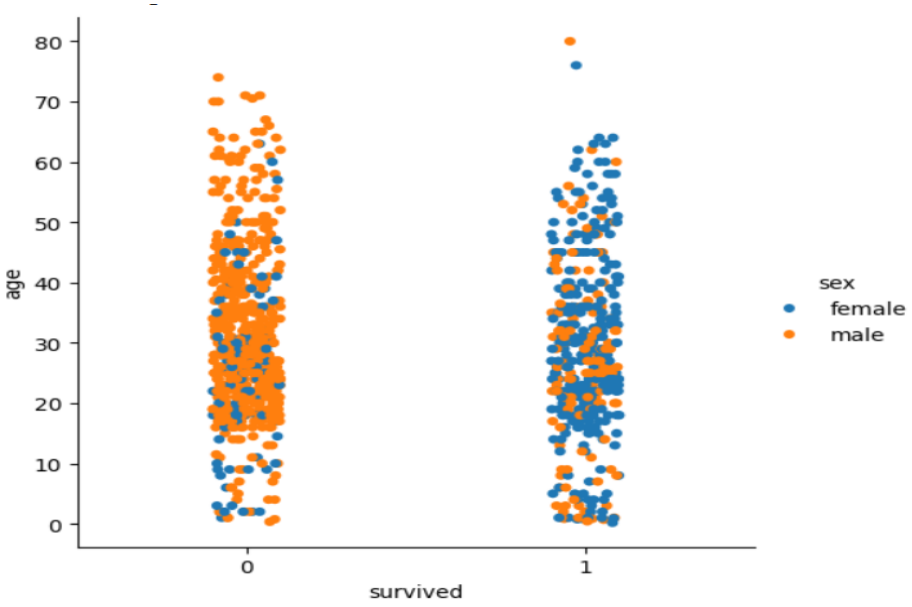

Figure 4: Age Distribution of Survivors and Non-Survivors by Gender

4.1.2 Insights

- The variables `Cabin`, `Boat`, `Body`, and `Home.Dest` contain a high percentage of missing values and are not crucial for survival prediction; these will be considered for removal during preprocessing.
- The `Age` variable has missing values; rows with missing age entries will be dropped in the preprocessing step.
- Most passengers fall within the age range of 18 to 32 years, indicating that the majority were young adults.
- The dataset includes 323 passengers in first class, 277 in second class, and 709 in third class.

- Gender distribution is skewed, with 843 male passengers and 466 female passengers.
- Survival was significantly higher among female passengers, highlighting gender as an important factor in survival.

4.2 Data Preprocessing

Following the initial EDA and feature selection, the dataset was prepared for modeling. This phase involved cleaning, encoding, and exploring the filtered data to ensure it was suitable for machine learning algorithms.

- Categorical variables were identified, and the only remaining one—`sex`—was encoded numerically (male = 0, female = 1).
- Rows with missing values in the `age` column were dropped to maintain data consistency.

After removing irrelevant or highly incomplete columns, additional exploration was conducted to better understand the relationships between variables. Grouped statistics were computed, such as:

- Average survival rate and age by gender.
- Combined averages by gender and passenger class.
- Distribution of passenger class among individuals under 18 years of age.

These groupings helped confirm the influence of features like gender, class, and age before proceeding to the modeling phase.

4.2.1 Survival Analysis by Gender, Class, and Age

sex	pclass	survived	age
0	2.300912	0.205167	30.585233
1	2.048969	0.752577	28.687071

		survived	age
sex	pclass		
0	1	0.350993	41.029250
	2	0.145570	30.815401
	3	0.169054	25.962273
1	1	0.962406	37.037594
	2	0.893204	27.499191
	3	0.473684	22.185307

4.2.2 Insights

The grouped statistics provided deeper insights into the roles of gender, class, and age in survival outcomes. Key findings include:

- On average, **75% of female passengers** survived, compared to only **20% of male passengers**.
- Among **females in first class**, **96% survived**, while only **47% survived** in third class.
- Among **males**, **35% of those in first class** survived, compared to just **16% in third class**.
- **First-class passengers** had a significantly higher chance of survival, regardless of gender.
- **68% of the children** (passengers under 18) were traveling in **third class**, suggesting age and class are intertwined in survival dynamics.

4.3 Modeling

To predict passenger survival, several classification algorithms were implemented, including:

- Decision Tree Classifier
- Random Forest Classifier
- AdaBoost Classifier

- Support Vector Machine (SVM)
- k-Nearest Neighbors (KNN)

Each model was trained on the preprocessed dataset using a supervised learning approach. A custom evaluation function was used to assess performance, which involved fitting the model, generating predictions on a test set, and displaying both the confusion matrix and the classification report.

To visualize model learning behavior, training and validation scores were plotted using learning curves with 4-fold cross-validation and an F1-score metric.

Additionally, a personalized prediction function was created to simulate survival outcomes for new individual profiles based on their passenger class, gender, and age. When supported by the model, survival probabilities were also displayed to give a clearer sense of predictive confidence.

After initial training and evaluation, hyperparameters for each model were optimized using Grid Search Cross-Validation. Once tuned, the models were re-evaluated and used again with the survival prediction function to observe any improvements in both accuracy and interpretability.

4.3.1 Model Performance Visualization

```
DecisionTree
[[124  14]
 [ 21  51]]
              precision    recall  f1-score   support

           0       0.86      0.90      0.88       138
           1       0.78      0.71      0.74        72

    accuracy                           0.83       210
   macro avg       0.82      0.80      0.81       210
weighted avg       0.83      0.83      0.83       210

RandomForest
[[122  16]
 [ 18  54]]
              precision    recall  f1-score   support

           0       0.87      0.88      0.88       138
           1       0.77      0.75      0.76        72

    accuracy                           0.84       210
   macro avg       0.82      0.82      0.82       210
weighted avg       0.84      0.84      0.84       210

AdaBoost
[[118  20]
 [ 17  55]]
              precision    recall  f1-score   support

           0       0.87      0.86      0.86       138
           1       0.73      0.76      0.75        72

    accuracy                           0.82       210
   macro avg       0.80      0.81      0.81       210
weighted avg       0.83      0.82      0.82       210
```

Figure 5: Classification Report of 3 models

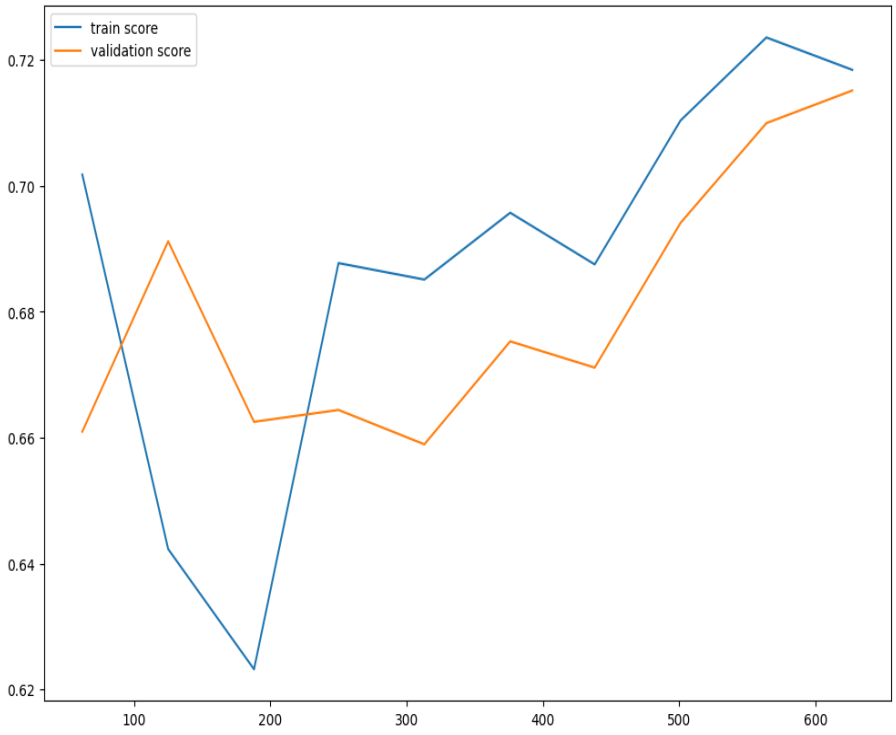

Figure 6: Learning Curve of SVM

- **DecisionTree:** You would have survived the Titanic (survival probability: 0.67)

- **RandomForest:** You would have survived the Titanic (survival probability: 0.71)

- **AdaBoost:** You would have survived the Titanic (survival probability: 0.65)

- **SVM:** You would have survived the Titanic.

- **KNN:** You would have survived the Titanic (survival probability: 0.80)

Figure 7: Prediction

4.3.2 Insights

- Among the tested models, the **Random Forest Classifier** and **SVM** achieved the highest accuracy and F1-scores, indicating strong predictive performance.
- **k-Nearest Neighbors (KNN)** performed adequately but was less effective than ensemble methods, especially for imbalanced classes.
- After hyperparameter tuning with Grid Search CV, most models saw slight improvements in performance, especially in terms of Recall.
- Gender and passenger class were consistently strong predictors of survival across all models.

- The personalized prediction function provided intuitive insights into how model decisions varied by age, class, and gender, enhancing the interpretability of results.

5 Results

The models were compared based on their predictive accuracy and performance on test data. After tuning hyperparameters, some models showed improved performance on the prediction tasks. The results demonstrated the importance of feature selection and model optimization in building effective machine learning pipelines.

6 Conclusion

This project showcases the full pipeline of a supervised learning task: from EDA and preprocessing to model training, prediction, and evaluation. Simpler models like Decision Trees were compared with ensemble methods like Random Forest and AdaBoost, and advanced algorithms like SVM. The exercise highlighted key machine learning concepts and the iterative nature of model improvement.