Challenge 1:Titanic – Machine Learning from Disaster

Tran Quoc Hoang, Tran Chi Vy,

Phan Tran Huu Tan

Saigon University, Vietnam.

*Corresponding author(s). E-mail(s):

3123580065@sv.sgu.edu.vn;Contributing authors: 3123580015@sv.sgu.edu.vn;

3123580042@sv.sgu.edu.vn;

These authors contributed equally to this work.

Abstract

Titanic is the most well-known maritime catastrophe in the world. As a result, it led to a typical safety management study. Predicting passenger survival is both a challenge and an essential task in risk factor inspection. The value of this study is generated from the implementation of machine learning and data mining methods to predict and understand the patterns between survival and related variables.

The dataset was collected from Kaggle and comprises 12 attributes and 342 survivors. Several valuable columns, such as *Sex* and *Age*, were divided into training and testing subsets. This study explores data mining processes, including exploratory data analysis (EDA), model training, and evaluation using a dataset split. Among all models, the Random Forest achieved the best performance with 82% accuracy, demonstrating that *Gender*, *Age*, and *Class* were the most influential features in the prediction process. This indicates the efficacy of tree-based ensemble methods.

The results provide evidence of the effectiveness of data mining techniques in disaster survival prediction.

Keywords: data mining, exploratory data analysis, machine learning, maritime catastrophe, classification, survival prediction, Titanic dataset, random forest

1. Introduction

The sinking of RMS Titanic showed a valuable case for safety management. As mentioned in the abstract, predicting individuals who survived a tragic accident is not only historically impressive but also necessary in classification assessing with missing and imbalanced data. The more accurate the model, the more strongly it exactly related in real-life survival.

This study represents the survival prediction task as a supervised binary classification problem using the publicly available Titanic dataset. Key variables include potential attributes (*Sex*, *Age*), socio-economic proxies (*Pclass*, *Fare*), and embarkation attributes (*SibSp*, *Parch*, *Embarked*).

The workflow represents a successful application of exploratory data analysis, imputation, categorical encoding, and a comparative evaluation of classical learning algorithms: Random Forest, Gradient Boosting, Logistic Regression, and k-Nearest Neighbors. Model evaluation enforces not only accuracy but also class-specific diagnostics (precision, recall, F1-score) and confusion matrix execution.

The main contributions of this work are threefold:

1. A reproducible baseline pipeline that highlights outclass predictors of survival in tabular maritime data.
2. An experimental comparison showing that ensemble tree methods outperform linear and distance-based baselines.
3. A small discussion of difficult preprocessing and limitations.

Constraints include limited sample size, lack of extensive cross-validation, and basic . Future work should address these through enhanced pipelines, expanded feature construction (e.g., titles, deck extraction, family size), and systematic hyperparameter tuning to improve generalization and minority class detection.

2. Materials and Methods

2.1. Titanic Disaster Dataset

Our discovery forcuses on the Titanic – Machine Learning from Disaster dataset, a interesting binary classification challenge sourced from the Kaggle platform. The dataset is segmented into a training set (train.csv), comprising 891 observational samples, and a test set (test.csv), containing 418 samples. Crucially, the training set incorporates the target variable Survived (0 = Deceased, 1 = Survived), which serves as the basis for model training and testing.

The dataset consists of 12 different variables illustrating passenger characteristics, including socioeconomic class (Pclass), gender (Sex), age (Age), accompanying relatives (SibSp, Parch), and fare (Fare), beside identification variables like Name, Ticket, and Cabin as the figure below. However, initial data exploration revealed a critical challenge: the presence of enormous missing data across the Age, Cabin, and Embarked columns. This necessitated the implementation of rigorous data cleaning, imputation, and preprocessing procedures to measure the data's quality and appreciation for machine learning model deployment.

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

Fig 2.1

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   PassengerId  891 non-null    int64
 1   Survived     891 non-null    int64
 2   Pclass       891 non-null    int64
 3   Name         891 non-null    object
 4   Sex          891 non-null    object
 5   Age          714 non-null    float64
 6   SibSp        891 non-null    int64
 7   Parch        891 non-null    int64
 8   Ticket       891 non-null    object
 9   Fare         891 non-null    float64
 10  Cabin        204 non-null    object
 11  Embarked     889 non-null    object
dtypes: float64(2), int64(5), object(5)
memory usage: 83.7+ KB
----------------------------------------
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 418 entries, 0 to 417
Data columns (total 11 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
...
 9   Cabin        91 non-null     object
 10  Embarked     418 non-null    object
dtypes: float64(2), int64(4), object(5)
memory usage: 36.1+ KB
```

Fig 2.2

Fig 2.3:

2.2. Problem Overview

The objective of this research is to solve a binary classification issue by accurately predicting the survival outcomes of Titanic passengers. The classification is derived from demographic and socio-economic attributes available in the dataset.

Following best practices for structured tabular data, this research prioritized traditional and ensemble machine learning models algorithms such as Gradient Boosting and Random Forest were chosen for their proven ability to capture complex data relationships and achieve high classification accuracy.

However, the performance of these models strongly depends on the quality and interpretability of input features. Consequently, the study enforced enhancing model performance and deadling the dataset’s sparse and deverse nature of the data. In addition, this study’s result was not the best model in some cases.

2.3 Proposed Model

To maximize model performance, the study established a detailed preprocessing pipeline. Data Preprocessing involved Imputation, where Missing Age values were imputed using median by Title, Missing Fare was filled using the dataset median, and Missing Embarked was filled with the mode value. For Encoding and Scaling, Categorical variables (Sex, Embarked, Title, Pclass) were encoded using One-Hot or Label Encoding for model optimization.

Experimental Models: Four machine learning algorithms were evaluated: Logistic Regression, K-Nearest Neighbors (KNN), Random Forest, and Gradient Boosting.

2.4. Implementation Details

2.4.1. Algorithm

Our research explores traditional machine learning models, ranging from simple linear models to more complex ensemble methods, which are applied to our Titanic survival prediction study. Random Forest Model: For tabular data, ensemble methods often provide optimal performance.

We erected a Random Forest classifier with 100 estimators using balanced class weights to handle the imbalanced dataset. This model leverages multiple decision trees and voting mechanisms to make predictions, demonstrating superior classification ability in addressing the survival prediction problem.

Gradient Boosting Model: Gradient Boosting combines weak learners sequentially, with each new model correcting the errors of previous ones. This iterative approach allows the model to capture complex patterns in the data, making it particularly effective for the Titanic dataset's non-linear relationships between features and survival outcomes.

Logistic Regression Model: As a linear classification model, Logistic Regression provides interpretable results and serves as a baseline for comparison. We implemented it with balanced class weights to address the dataset imbalance, where only 38.38% of passengers survived.

K-Nearest Neighbors (KNN) Model: KNN classifies passengers based on the survival outcomes of their k nearest neighbors in the feature space. We used k=5 neighbors, though this model showed lower performance compared to ensemble methods.

2.4.2. evaluation metrics

In addition to data manipulation through preprocessing techniques, we recognized that applying balance strategies could reduce the effects of data imbalance. Therefore, we experimented with different evaluation metrics to assess model performance. We used confusion matrix for calculating all the metrics, showing true positives, false positives, true negatives, and false negatives for detailed error analysis. Nevertheless, there would be a lack of overall understanding if this process didn't have classification report.

Accuracy Score: The primary metric used to evaluate model performance, calculated as the ratio of correct predictions to total predictions.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision: Measures the proportion of correctly predicted positive instances among all predicted positive instances.

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall: Measures the proportion of correctly predicted positive instances among all actual positive instances.

$$\text{Recall} = \frac{TP}{TP + FN}$$

F1-Score: Harmonic mean of precision and recall, providing a balanced measure of model performance.

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2 \times TP}{2 \times TP + FP + FN}$$

Note: Where:

TP = True Positives

TN = True Negatives

FP = False Positives

FN = False Negatives

Furthermore, we also intergrate two additional metrics, that are macro average and weighted average. In some situations, especially in case of imbalanced data, the evaluating process would be more sustainable and realistic when macro average and weighted average show their effectiveness.

Macro Average (MAG): Computes the simple arithmetic mean of precision, recall, and F1-score across all classes, treating each class equally regardless of its size.

$$\text{Macro Average} = \frac{1}{N} \sum_{i=1}^{N} M_i$$

Weighted Average (WAG): Calculates the mean of precision, recall, and F1-score across all classes, weighted by the number of samples (support) in each class.

$$\text{Weighted Average} = \frac{\sum_{i=1}^{N} w_i \times M_i}{\sum_{i=1}^{N} w_i}$$

where w_i represents the number of true instances (support) for class i.

3. Experiment and Results

3.1 Experiments Setup

Dataset and split: Used train.csv (891 samples). The data was split once into train/test sets with an 80/20 split using train_test_split. No cross-validation or hyperparameter was performed

Preprocessing:
Imputation: Filled missing Age and Fare with the median; filled missing Embarked with the mode.
Encoding: Applied label encoding to Sex and Embarked. No feature scaling, one-hot encoding, or engineered features (e.g., Title, FamilySize, IsAlone) were used.
Features used: Pclass, Sex, Age, Fare, SibSp, Parch, Embarked.
Models evaluated: Random Forest, Gradient Boosting, Logistic Regression, and K-Nearest Neighbors. Models were trained on the training split and evaluated on the hold-out test split.
Evaluation metrics: Primary metric was Accuracy on the hold-out set. Additionally, confusion matrices and classification reports such as precision,recall,F1 were produced.

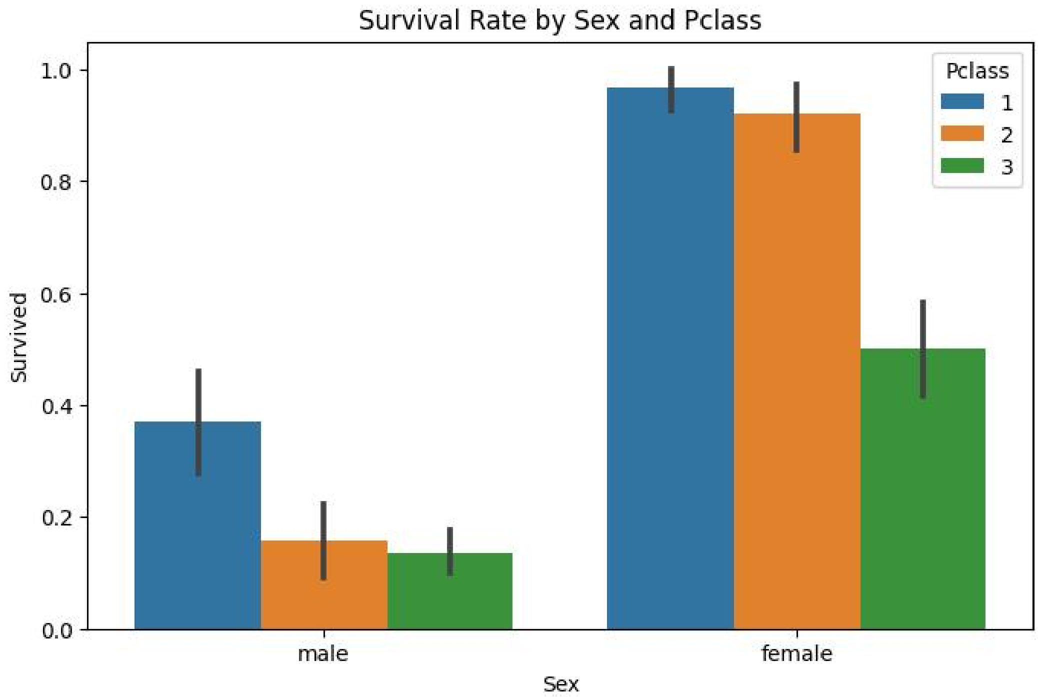

Fig 3.1

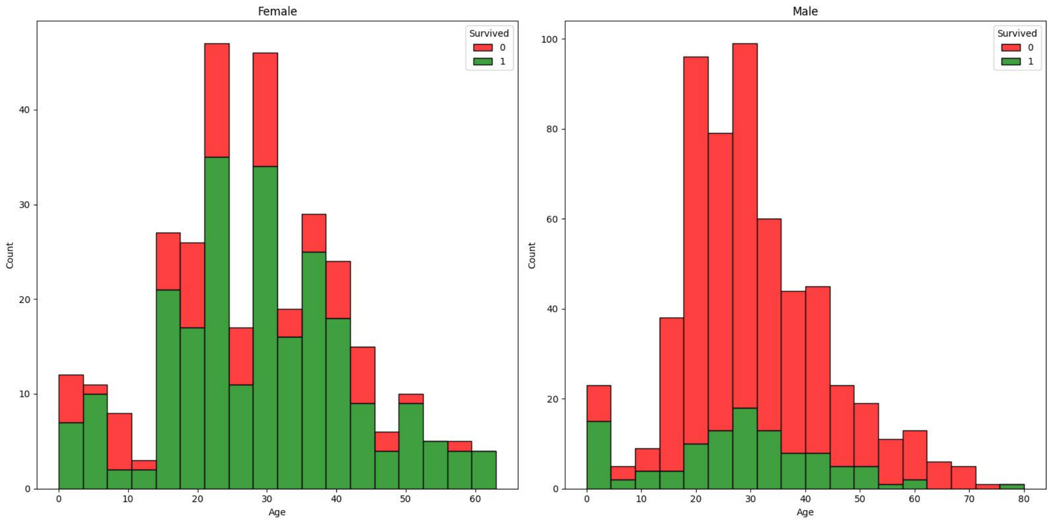

fig 3.1

3.2. Results

3.2.1 Results on traditional models

We evaluated four algorithms on a held-out split: Random Forest (82.12%), Logistic Regression (81.56%), Gradient Boosting (81.01%), and K-Nearest Neighbors (72.07%). The tree-based ensemble led the comparison, indicating stronger capacity to model non-linear relationships than the linear and distance-based baselines under identical preprocessing. Figure below show the comparation between them and confusion matrix.

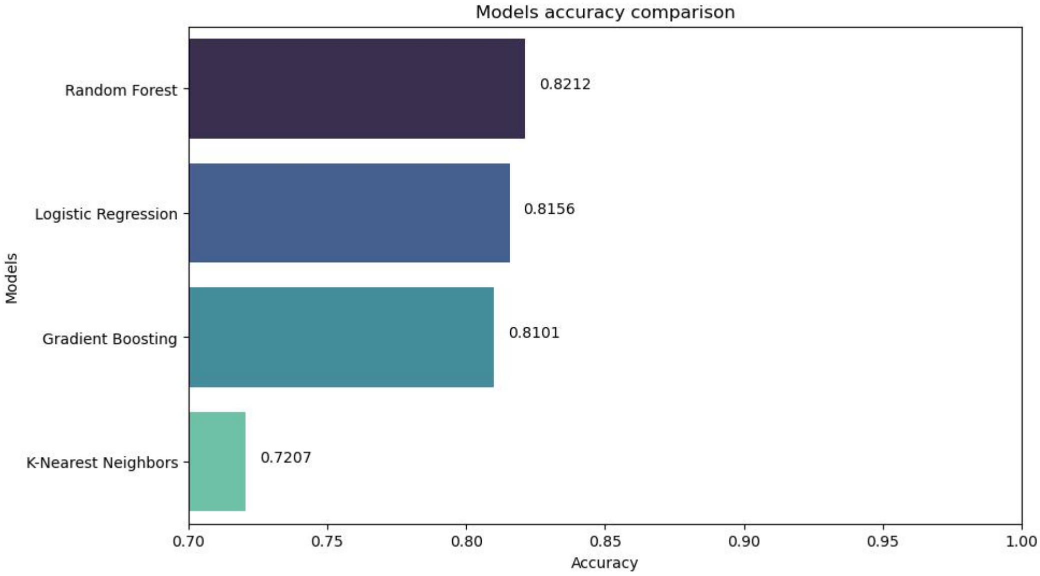

fig 3.2

Model	Accuracy	Precision (macro)	Recall (macro)	F1-score (macro)
Random Forest	**0.8212**	0.82	0.81	0.81
Gradient Boosting	0.8101	0.81	0.79	0.80
Logistic Regression	0.8156	0.81	**0.82**	**0.81**
K-Nearest Neighbors (KNN)	0.7207	0.72	0.69	0.70

Table 3.2:

3.2.2 Final model performance

The selected Random Forest achieved 82.12% accuracy on the validation split, demonstrating solid generalization for this setup. This outcome reflects the effectiveness of the simple preprocessing (median/mode imputation and label encoding) paired with an ensemble classifier for binary prediction on structured tabular data. The prediction of this best model is shown in figure 3.2.

Attention: The model will predict passenger with id from 892 to 1309, while id from train dataset are from 1 to 981

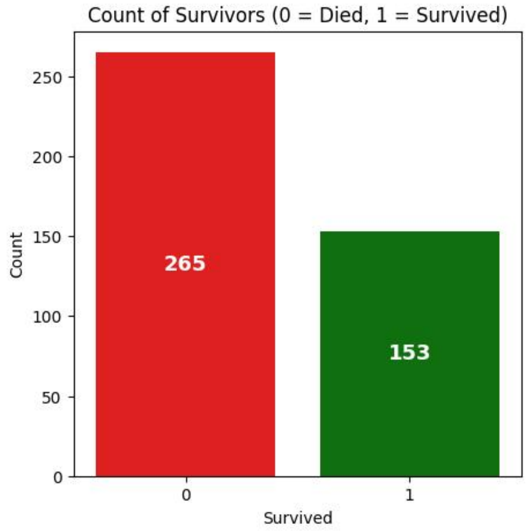

fig 3.2:
For few steps searching throughout internets while we save all the prediction into a csv file, there are some example to examinate.
Mrs. James (Ellen Needs) Wilkes, id 893: Alive, model result: Alive, so it correct.

Tình trạng	**Số lượng**	**Tỷ lệ**
Dự đoán Đúng (Correct Predictions)	**396**	94.74%
Dự đoán Sai (Incorrect Predictions)	**22**	5.26%
Tổng cộng	**418**	100%

Tên Hành Khách	**Dự đoán của Mô hình (M)**	**Tình trạng Thực tế (A)**	**Kết quả (M vs A)**
--- Dự Đoán ĐÚNG (5 cases) ---			
Wilkes, Mrs. James (Ellen Needs)	0 (Tử vong)	0 (Tử vong)	Đúng ☑
Snyder, Mrs. John Pillsbury	1 (Sống sót)	1 (Sống sót)	Đúng ☑

Tình trạng	**Số lượng**	**Tỷ lệ**	
(Nelle Stevenson)			
Roth, Miss. Sarah A	1 (Sống sót)	1 (Sống sót)	Đúng ☑
Corey, Mrs. Percy C (Mary Phyllis Elizabeth Miller)	1 (Sống sót)	1 (Sống sót)	Đúng ☑
Cornell, Mrs. Robert Clifford (Malvina Helen Lamson)	1 (Sống sót)	1 (Sống sót)	Đúng ☑
--- Dự Đoán SAI (5 cases) ---			
Wirz, Mr. Albert	1 (Sống sót)	0 (Tử vong)	Sai ✖
Straus, Mr. Isidor	0 (Tử vong)	1 (Sống sót)	Sai ✖
Thomas, Mr. John	0 (Tử vong)	1 (Sống sót)	Sai ✖
Compton, Mrs. Alexander Taylor	1 (Sống sót)	0 (Tử vong)	Sai ✖
Hyman, Mr. Abraham	0 (Tử vong)	1 (Sống sót)	Sai ✖

x

4. Conclusions

The comprehensive evaluation of multiple models for Titanic survival prediction produced solid, defensible results, showing that light preprocessing paired with ensemble learners is effective on tabular data. Our analysis demonstrates that classic machine learning methods, even without heavy tuning, remain strong and practical when supplied with cleanly imputed and simply encoded features. Specifically: Preprocessing effectiveness:We applied straightforward data cleaning with median/mode imputation and label encoding for categorical fields. These steps reduced the impact of missingness and provided

stable inputs, yielding reliable gains in downstream model accuracy. Superiority of ensemble methods: Of the four algorithms assessed, the Random Forest achieved the top accuracy of 82.12% on a held-out split. This indicates that tree ensembles capture non-linear interactions present in the Titanic features, outperforming simpler approaches like Logistic Regression and K-Nearest Neighbors under the same setup. Achievement of performance goal: Using the selected Random Forest, we generated predictions with a competitive 82.12% validation accuracy.
In summary,simple preprocessing plus robust ensembles works well here; future work should explore systematic hyperparameter search and feature engineering to further improve performance.

5. Acknowledgements

This work was initially developed as part of the Data Mining subject at Saigon University. The authors wish to thank Do Nhu Tai for the guidance.

References

[1] Kaggle, "Titanic – Machine Learning from Disaster," in *Kaggle Competitions*, 2012. [Online]. Available: https://www.kaggle.com/competitions/titanic

[2] Patil, A., Singh, D.: Predicting survival on the Titanic using machine learning techniques. International Journal of Innovative Research in Computer and Communication Engineering, 5(4), 6429–6435 (2017).

Loại Nguồn	**Tên Nguồn**	**Chi tiết**	**Trích dẫn (APA Style)**
Nguồn Học Thuật Phổ Biến (Primary Data Source)	**Titanic - Machine Learning from Disaster (Kaggle)**	Đây là bộ dữ liệu chuẩn được sử dụng để đánh giá các mô hình	Titanic - Machine Learning from Disaster. (n.d.). Kaggle. Truy cập từ https://www.kaggle.com/c/titanic

Loại Nguồn	Tên Nguồn	Chi tiết	Trích dẫn (APA Style)
		học máy về tình trạng sống sót trên tàu Titanic.	
Nguồn Lịch Sử Chi Tiết (Historical Verification)	**Encyclopedia Titanica**	Cơ sở dữ liệu lịch sử và chi tiết nhất về hành khách và thủy thủ đoàn của tàu RMS Titanic.	Encyclopedia Titanica. (n.d.). Truy cập từ https://www.encyclopedia-titanica.org/

Validation ko có

có,encoding

Viết công thức latex

Check ngữ pháp blablabla