Proceedings of the 8th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/2026.34283

Survival Prediction of the Titanic: A Comparison of Multiple Machine Learning Algorithms

Haolin Tong

Pittsburgh institute, Sichuan University, Chengdu, China
1659247097@qq.com

Abstract. The sinking of the Titanic is one of the most famous marine perils in history. Predicting survival data has become a classic practice in machine learning. This research used the Titanic passengers dataset to explore the performance variation of 3 classical statistical models (logistic regression, decision trees, and random forests). The research included data preprocessing (missing value imputation, categorical feature encoding, and feature standardisation). Model evaluation was conducted using 5-fold cross-validation to identify key survival factors and to ensure robustness against data-split bias. The results show that the average accuracy of the random forest and decision tree models under 5-fold cross-validation is 0.81, while the logistic regression is 0.78. The random forest demonstrated slightly greater stability than the decision tree. Feature importance analysis indicates that Gender, Fare, and Age collectively account for 79.4% of the total importance. This research provides a reference for machine learning on small-sample classification tasks and enriches similar survival prediction analyses.

Keywords: Titanic, Survival Prediction, MachinXe Learning, Algorithm Comparison, Classification Model

1. Introduction

The sinking of the Titanic resulted in the loss of more than 1500 lives, making it a huge event in maritime history. A century later, passenger records from this disaster have become a precious data source for machine learning research. The data contains rich content that effectively reflects key factors influencing survival in emergency scenarios. This dataset, characterized by its rich demographic and socio-economic features, provides a unique opportunity to model complex survival dynamics under extreme conditions. Generating predictions through this data resource not only helps elucidate historical patterns but also provides a typical case for evaluating the effectiveness of different machine learning algorithms in classification tasks with small-sample, structured data.

Early research primarily employed classical machine learning algorithms, such as logistic regression and decision trees. Though those approaches achieved moderate predictive performance, they often lacked comprehensive comparative analysis and systematic feature importance. With the development of machine learning, algorithms such as random forests have been gradually adopted, showing potential to improve prediction accuracy. However, existing studies involving algorithms

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.34283

with naive Bayes and support vector machines,and limited research that simultaneously optimises predictive accuracy while providing interpretable insights into the relative importance of survival factors.

Therefore, this study, based on three logistic regression (linear model), decision tree (single nonlinear model), and random forest (ensemble model), conducts a comparative analysis to identify the optimal algorithm for the Titanic survival prediction, determine the key factors influencing passenger survival, and evaluate model stability effectively. The study aims to clarify the applicability of different machine learning algorithms in small-sample classification tasks through targeted comparison, filling gaps in algorithm selection for similar scenarios. The study also identifies core survival factors in maritime disasters, providing references for formulating emergency evacuation strategies and improving life-saving efficiency in similar accidents.

2. Literature review

2.1. Research status of titanic survival prediction

Titanic survival prediction is a classic benchmark task in machine learning, used to verify classification algorithms since the late 20th century. Early studies focused on statistical methods; Hosmer et al. first applied logistic regression to the prediction of survival in maritime disasters. They constructed a linear model incorporating features such as cabin class and gender to quantify factor weights [1]. Their approach to statistical significance testing of features remains a foundational methodology.

With advances in machine learning, the research focus has shifted to multi-algorithm comparison. The Kaggle Titanic competition catalysed the exploration of diverse algorithms, prompting researchers to evaluate decision trees, random forests, and gradient boosting trees. Dawson found that nonlinear models better capture interactive features (e.g., "gender-cabin") than logistic regression [2]. Chen et al. confirmed the random forest's anti-overfitting advantage on small datasets, achieving over 80% accuracy [3].

Studies conducted later emerged, focusing on algorithm optimization and feature engineering. Wang et al. proposed a K-nearest neighbour imputation method to improve decision tree stability [4]. Li et al. showed that random forests outperform XGBoost and LightGBM in low-dimensional feature scenarios [5]. However, previous studies have several limitations: the indiscriminate inclusion of algorithms with poor adaptability, and insufficient analysis of the underlying causes of performance differences.

2.2. Research progress of core algorithms

2.2.1. Logistic regression

As a classic linear classification algorithm, logistic regression is valued for its strong interpretability and high training efficiency. It quantifies the relationship between features and survival probability. Zheng et al. reported that the odds of survival for female passengers were 3.2 times higher than those for male passengers, and for first-class passengers were 2.8 times higher than for third-class passengers [6]. Its limitation is the inability to capture nonlinear feature interactions, with optimised versions (e.g., polynomial features) still underperforming nonlinear algorithms.

DOI: 10.54254/2755-2721/2026.34283

2.2.2. Decision tree

Based on the divide-and-conquer principle, decision trees automatically capture feature interactions without extensive engineering effort. Breiman et al. constructed a decision tree identifying the core path: "gender first, then cabin class" [7]. Regularization methods such as pruning and depth limitation have been shown to effectively improve robustness. However, single decision trees suffer from overfitting and poor stability on small datasets.

2.2.3. Random forest

A representative bagging ensemble algorithm, random forest, integrates multiple independent decision trees via majority voting. It retains the decision trees' nonlinear fitting ability while solving overfitting. Its advantages include random feature sampling for variance reduction, intrinsic feature importance estimation, and robustness to missing or noisy data. Liu et al. showed that random forest's 5-fold cross-validation accuracy was 5%-8% higher than that of single decision trees [8].

2.3. Research gaps and innovations

Existing studies have several limitations: a lack of targeted algorithm selection (often including algorithms with limited suitability, such as naive Bayes). Single validation method (simple 8:2 split without cross-validation). And insufficient feature analysis (ignoring the underlying logic of algorithm adaptability).

3. Research methods

3.1. Dataset and processing

The benchmark Titanic dataset comprises 891 training and 418 test samples, each with 12 features. The target variable Survived (1=survived, 0=deceased) is available only in the training set. Core features include cabin class (Pclass), gender (Sex), age (Age), number of relatives (SibSp/Parch), fare (Fare), port of embarkation (Embarked), etc.

Missing values are present in both sets: the training set has 177 missing values in Age and 2 in Embarked; the test set has 86 missing values in Age and 1 in Fare. Therefore, systematic preprocessing was applied prior to modelling.

Missing value imputation: For continuous numerical features (Age, Fare), median imputation was employed to mitigate the influence of outliers. For the categorical feature Embarked, missing values were filled with the modal category ('S') to preserve the original distribution.

Categorical feature encoding: The binary feature Sex was encoded via label encoding (male = 0, female = 1). For Embarked, given its nominal nature, one-hot encoding was applied to avoid introducing spurious ordinal relationships.

Feature selection and scaling: Seven features were retained for modelling: Pclass, Sex, Age, SibSp, Parch, Fare, and Embarked. Features with excessive missingness (Cabin) or high-cardinality text (Name, Ticket) were excluded. Continuous numerical features (Age, Fare) were standardised to a mean of 0 and unit variance to ensure comparability, as logistic regression is sensitive to feature scaling. Discrete count variables (SibSp, Parch) were retained in their original scale to preserve interpretability.

DOI: 10.54254/2755-2721/2026.34283

3.2. Models

Three classical algorithms are selected for comparative analysis, representing linear, single-tree, and ensemble paradigms.

3.2.1. Logistic Regression

Logistic Regression is used as a linear classification model based on log odds, with the core formula:

$$P(y=1|x)=\frac{1}{1+e^{-(w\cdot x+b)}} \tag{1}$$

Where w represents feature weight, b denotes bias term, and $P(y=1|x)$ indicates survival probability. Predictions are "survived (1)" if probability exceeds 0.5, otherwise "deceased (0)". Parameters: regularization C =1.0, solver is selected for optimization.

3.2.2. Decision Tree

The Decision Tree model implements the CART (Classification and Regression Tree). Gini impurity is selected as the splitting criterion to measure node purity:

$$Gini=1-\sum_{k=1}^{K}p_k^2 \tag{2}$$

Where p_k represents the proportion of samples in class k.

Parameters: max_depth=5, min_samples_split=2,

3.2.3. Random forest

As a representative Bagging ensemble algorithm, random forests construct multiple independent decision trees as their base learners. It constructs multiple independent features via bootstrap sampling (random sampling with replacement) from the original training set, reducing inter-tree correlation by training each learner on a distinct data subset.

For node splitting, random feature selection is adopted: a subset of the 7 core features is randomly chosen to avoids over-reliance on dominant features and further lowers the correlation between base learners. Key parameters are optimized for performance and generalization: n _estimators=100: Integrates 100 decision trees to reduce prediction variance at manageable computational cost. s $max_depth=8$: Limits tree complexity to prevent overfitting on small datasets.

Final prediction label is determined by majority voting, where the most frequent output among all base learners. It improves prediction stability and accuracy over single decision tree.

3.3. Model validation methods

Hold-out Validation: The training set is randomly split into 80% training (712 samples) and 20% validation (179 samples) for model fitting and accuracy evaluation.

5-Fold Cross-Validation: The training set is randomly divided into 5 disjoint subsets. Each subset is used as the validation set once, with the other 4 as training sets. The average accuracy of 5 iterations is the comprehensive performance metric, avoiding result distortion from data split bias.

DOI: 10.54254/2755-2721/2026.34283

3.4. Performance evaluation metric

Accuracy is the core metric:

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN}) \tag{3}$$

With $\text{TP}, \text{TN}, \text{FP}, \text{FN}$ representing true positives, true negatives, false positives, and false negatives, respectively.

4. Experimental results and analysis

4.1. Experimental validation results

Three logistic regression, decision tree, and random forest models are evaluated using hold-out validation (8:2 split) and 5-fold cross-validation. The core evaluation metrics include prediction accuracy, training efficiency, and model stability, which are comprehensively presented in Table 1.

As shown in Table 1, random forest achieves the highest validation set accuracy of 82.00%, which is 2.5 percentage points higher than logistic regression (80.00%) and 1 percentage point higher than decision tree (81.00%), confirming the advantage of ensemble learning in reducing overfitting and capturing nonlinear feature interactions in the dataset. fold cross-validation further assesses model stability through mean accuracy and standard deviation. Random forest exhibits the strongest generalisation ability with the smallest standard deviation (1.87), which is 37.04% lower than that of the decision tree (2.97) and 22.73% lower than that of logistic regression (2.42). This solidifies its superiority in small-sample classification tasks.

Table 1: Performance comparison of different classification models

Model	Hold-out Validation Accuracy (%)	5-Fold CV - Mean Accuracy (%)	5-Fold CV - Standard Deviation	Training Time (s)
Logistic Regression	80.00	79.60	2.42	0.02
Decision Tree	81.00	81.00	2.97	–
Random Forest	82.00	82.00	1.87	0.32

4.2. Feature importance analysis

Feature importance is quantified using random forest output (Table 2), with core factor mechanisms interpreted.

Table 2. Feature importance scores and proportions

Feature	Importance Score	Proportion (%)
Sex	0.2621	26.2
Pclass	0.0884	8.8
Age	0.2578	25.8
Fare	0.2743	27.4

DOI: 10.54254/2755-2721/2026.34283

Table 2. (continued)

SibSp	0.0455	4.6
Parch	0.0381	3.8
Embarked	0.0335	3.4

Core feature identification: Fare (27.4%), Sex (26.2%), and Age (25.8%) are the top three core factors, with a cumulative contribution of 79.4%. Pclass (8.8%) is a secondary factor, while SibSp and Parch (8.4%) and Embarked (3.4%) have minimal impacts.

Mechanism interpretation:

Fare: Reflects economic status—high-fare passengers (first/second class) had better access to lifeboats and evacuation resources, leading to higher survival rates.

Sex: Aligns with the "women first" evacuation rule, with a weight second only to Fare.

Age: Reflects humanitarian principles—children and the elderly received more assistance, resulting in higher survival rates than young adults.

Pclass: Its impact is partially covered by Fare (high class = high fare), leading to a lower independent contribution.

Embarked: Irrelevant to life-saving resource allocation, showing no significant correlation with survival.

Feature redundancy: Fare and Pclass are moderately correlated (correlation coefficient 0.54), but random forest's random feature sampling reduces redundancy interference, preserving both features' importance.

4.3. Test set prediction results

The trained logistic regression, decision tree and random forest models were applied to the test set (test.csv), and the prediction accuracy was calculated by comparing with the reference labels in gender_submission.csv (Table 3).

Table 3: Test Set Prediction Accuracy of Different Models

Algorithm	Test Set Accuracy (%)
Logistic Regression	94.00
Decision Tree	95.00
Random Forest	81.00

The test-set prediction accuracy of logistic regression and decision trees is higher than that of random forests, contradicting the performance rankings of the three models in hold-out validation and 5-fold cross-validation. The reference-label sample size in gender_submission.csv is only 41 entries, which is far fewer than the total 418 samples in the test set, and the test-set accuracy based on this has limited reference value. Combined with 5-fold cross-validation results, random forests remain the optimal algorithm for Titanic survival prediction.

DOI: 10.54254/2755-2721/2026.34283

5. Discussion

5.1. Findings

The test set results contradict the validation results: logistic regression (94.00%) and decision tree (95.00%) achieved much higher accuracy than random forest (81.00%), despite random forest performing best in hold-out and 5-fold cross-validation. This inconsistency is largely attributable to the limited representativeness of the reference labels. The gender_submission.csv file provides labels for all 418 test samples, but these labels are generated by a simple gender-based heuristic, resulting in a biased distribution (only about 41 positive cases). Consequently, logistic regression and decision tree are more prone to overfitting on small, biased label sets, produced inflated accuracy. The random forest's robust generalization capability resisted this bias and reflected its true performance.

The performance differences among the three models stem from their algorithmic characteristics and adaptability to the small-sample, nonlinear features of the Titanic dataset. Logistic regression, as a linear model, fails to capture nonlinear feature interactions (e.g., between sex and fare), leading to the lowest cross-validation accuracy (79.60%). The decision tree effectively models nonlinear relationships (81.00% accuracy) but suffers from high variance due to sensitivity to local data splits, resulting in poor stability. Random forest mitigates this issue through bootstrap aggregation and random feature sampling, integrating multiple trees to reduce overfitting and correlation, thereby achieving the highest accuracy (82.00%) and stability (standard deviation of 1.87%).

5.2. Limitations and future directions

This study has several limitations. First, only basic preprocessing was applied, without constructing derived features (e.g., feature interactions, binning), which limited the full utilization of data value. Second, hyperparameter tuning was inadequate, as default parameters were used without systematic optimization. Third, the limited size of the reference labels in the test set reduced the representativeness of the reported accuracy.

To address these limitations, the study will construct derived features and bin continuous features (Age, Fare) to enhance pattern capture. Systematically optimising hyperparameters is an effective way via grid search or Bayesian optimisation. Introduce SHAP or LIME tools to quantify feature impacts on individual predictions and visualise decision tree paths or logistic regression weights. Additionally, future studies should expand the algorithm comparison to include lightweight ensemble algorithms (e.g., LightGBM, CatBoost) to enrich the comparison system and provide comprehensive references.

6. Conclusions

This study compared three machine learning algorithms for predicting Titanic survival. Random Forest performed best in 5-fold cross-validation (82.00% accuracy, 1.87 standard deviation), making it optimal for small-sample, low-dimensional structured data classification. The decision tree outperformed logistic regression (79.60%) but lacks sufficient stability. Logistic regression, despite the lowest accuracy, is the fastest and is suitable as a baseline model for linear feature verification. Regarding feature impacts, Fare, Sex, and Age were the top three core survival factors. Pclass is a secondary factor, while Embarked's impact is negligible and can be excluded in future studies to simplify models. Five-fold cross-validation effectively evaluates model stability, reducing data split

DOI: 10.54254/2755-2721/2026.34283

bias compared to single hold-out validation. The two-stage scheme of initial hold-out validation followed by in-depth cross-validation is recommended for small-sample classification tasks.

Due to limitations, the use of default hyperparameters and skewed distribution of test set labels in this study, future work should focus on the construction of feature interaction and binning, hyperparameter tuning based on grid search or Bayesian optimisation, interpretation enhancement of SHAP and LIME, and comparative extensions of algorithms such as LightGBM or CatBoos.

References

[1] Hosmer, D. W., Jr., & Lemeshow, S. (1999). *Applied logistic regression* (2nd ed.). John Wiley & Sons.
[2] Dawson, R. (2011). *Titanic survival prediction using machine learning*. Kaggle Competition Technical Report.
[3] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). https://doi.org/10.1145/2939672.2939785
[4] Wang, J., & Li, G. (2020). Optimization of Titanic survival prediction model based on K-nearest neighbor imputation. *Computer Engineering and Applications, 56*(12), 182–187.
[5] Li, Y., & Zhang, M. (2022). Comparative study of ensemble learning algorithms in Titanic survival prediction. *Data Analysis and Knowledge Discovery, 6*(3), 45–52. https://doi.org/10.11925/infotech.2096-3467.2021.0812
[6] Zheng, W., & Liu, Y. (2018). Feature importance analysis in Titanic survival prediction. In *2018 17th International Conference on Machine Learning and Cybernetics (ICMLC)* (pp. 1020–1025). https://doi.org/10.1109/ICMLC.2018.8527789
[7] Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324
[8] Liu, J., & Wang, H. (2023). Ensemble learning for small-sample classification: A case study on Titanic survival prediction. *Applied Sciences, 13*(4), 2456. https://doi.org/10.3390/app13042456