

Statistical Inference and Multi-Model Comparative Study on the Titanic Survival Rate

Yixuan Zeng^{1,a,*}, Zihuang Gong^{2,b}

¹Guanggu Future School, Wuhan, 430074, Hubei, China

²Shanghai Experimental Foreign Language School, Shanghai, 201800, China

^ayixuan.zeng916@outlook.com, ^b2561272131@qq.com

**Corresponding author*

Keywords: Statistical Inference; Arranging and Analysing Data; Titanic; Logistic Regression; Random Forest; Hard Voting

Abstract: In the early hours of April 15, 1912, the RMS Titanic—the world's largest and most luxurious ocean liner at the deeply grieved time—struck an iceberg during its maiden voyage in the North Atlantic. Referred to as “unsinkable”, the enormous ship sank completely within just 2 hours and 40 minutes. Due to a severe shortage of lifeboats and delayed rescue efforts, over 1,500 of the 2,224 passengers crew deceased, with only around 700 surviving, cementing the disaster as one of the most tragic maritime perils of the 20th century. This dissertation focuses on analyzing the key factors which influence passenger survival rates on the Titanic. By constructing a hard-voting ensemble model integrating logistic regression, random forest, and decision tree algorithms, we predict individual survival probabilities based on their unique characteristics. Our methodology systematically addresses missing data, utilizes innovative feature engineering to derive new variables, and applies a practicable model to the dataset, thereby uncovering the core determinants of survival and their underlying mechanisms. The integrated data processing and model construction demonstrate strong performance, achieving an accuracy of 89.21%. Expanding this model to maritime safety applications enables systematic collection and analysis of multi-source data, enhancing the precision of risk prediction significantly. In this context, the result not only improves disaster prevention but also contributes to the advancement of data-driven safety capabilities in modern navigation.

1. Introduction

The Titanic was the most luxurious ocean liner of the early 20th century, touted as "unsinkable." However, on its maiden voyage in 1912, it struck an iceberg and sank, resulting in over 1,500 deaths. This disaster shocked the world and became one of the most infamous maritime tragedies in history. Various approaches have been adopted to analyze the existing data of Titanic.

Therefore, in this article, we present several approaches, such as random forest, hard voting, logistic regression and other machine--learning methods, which help us to explore the correlation between the Titanic death rate and certain influencing factors, so that we can predict whether a passenger will survive in a dangerous situation. In addition, in the future, the results obtained through

these high-accuracy methods can be used by individuals who wish to take steps to ensure their survive in tragedies and avoid suffering or even death when facing major disasters.

In the following content of this paper, our group divides the dissertation into 4 parts. Firstly, Section 2 presents the data after we clean and complete it, and then we organize and define different variables as diverse types in order to investigate the major influences on death rate. After some data visualization in the section 2, section 3, we prepare to introduce some methodology we take into measure the survival rate, including logistic regression, random forest and decision tree. Eventually, Section 4 concludes and analyzes the models discussed in the main body of the paper.

2. Data Preparation and Preprocessing

2.1 Data Source and Description

The data used in this article is sourced from two sets: test and train. Merging the two data source packages into one is our first step, totally 12 groups(called: Passenger Id, Survived, P-class, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, Embarked), however, we found that some of these data are missing, there are 263 pieces of data are missing in Age, and 1 in Fare, 1014 in Cabin, and 2 in Embarked. To address the different quantities of missing data, we use the median to fill missing values in Age, the mean for Fare, the mode for Embarked. However, for data that is mostly missing and not crucial, we ignored these data in this situation such as Cabin.

2.2 Variable Classification and Coding

2.2.1 Title Extraction

First of all, we observed some of the passengers' names, and noted that the assorted titles of these names could have an effect on different survival rate. So, we made a new variable to store all sorts of titles, marked "Title". In this very same variable, we can promptly figure out a lot of rare titles. In this variable, some rare titles can be learned about quickly, such as "Rev", "Sir", "Jonkheer", "Col" and so on. Many people also have the same title, the most frequently occurring title on the Titanic is "Mr.", followed by "Miss.", "Mrs.", "Master" and others [1].

2.2.2 Fare Distribution and Grouping

From this distribution chart of Fare, we can clearly observe that among different ticket prices, the most frequent fare on the cruise ship is 8. We speculate that Fare has some relationship with Survived, Age and P-class. In the Fare data, we divide these prices into three categories based on their size: cheap, medium, and expensive. Based on this statistical chart, the overall trend shows that the higher the fare, the fewer buyers there will be, and the P-class may also be higher. Their high social status may also reduce their mortality rate in this disaster, and they may enjoy more happiness compared with others which may not be the same rich as them.

2.2.3 Ticket Uniqueness and Data Validation

The ticket codes of passengers are basically unique and non-repeating, and the various ticket codes have been categorized into a variable called "Ticket", which can verify the reality of the data [2].

2.2.4 Family Size Construction and Survival Implications

Through analysing these existing data, it is not very difficult for us to discover that some people would like to choose to take on this journey with their important people who get a close relationship

with them. This behavior is easy understanding and therefore we wonder this factor whether can create effects of survival rate or not. We put two variables (“Parch”, “SibSp”,) into one piece, named “Family Size”, it includes whether a passenger has any family or spouse on board the ship.

In figure 1, we divided these families into different types based on the distinct number of family members, they are Single, Small, and Large. Small group indicates survival rate about 60% higher than death rate, but unfortunately other families who are in the single and large size have more risk to decease. This interprets that family bonds may make significant efforts for ensuring your safety when you meet a huge disaster. However, because of poor calmness and organizational skills, they might sacrifice more people to safe others in large families.

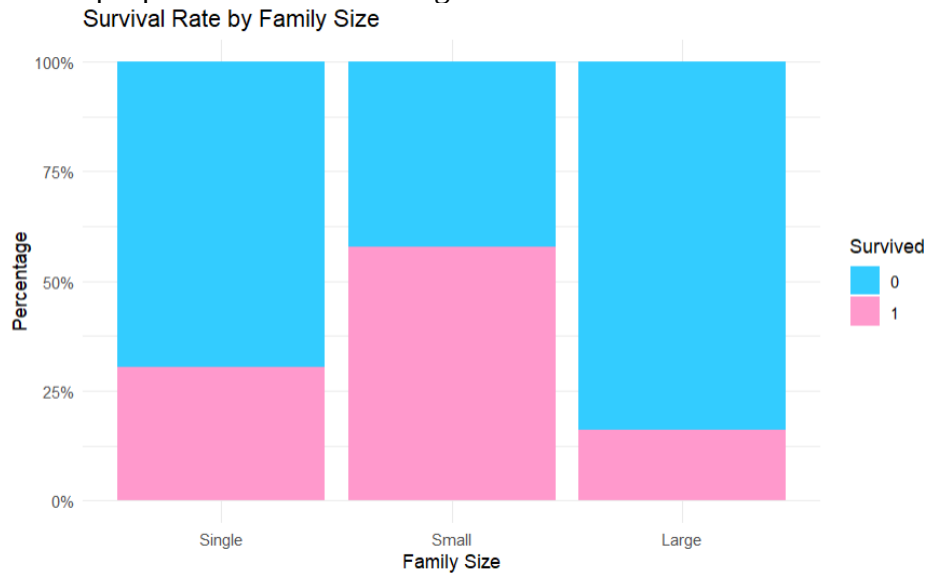


Figure 1: Survival Rate by Family Size

2.3 Exploratory Data Visualization

2.3.1 Family Size, Sex Disparities in Survival Outcomes

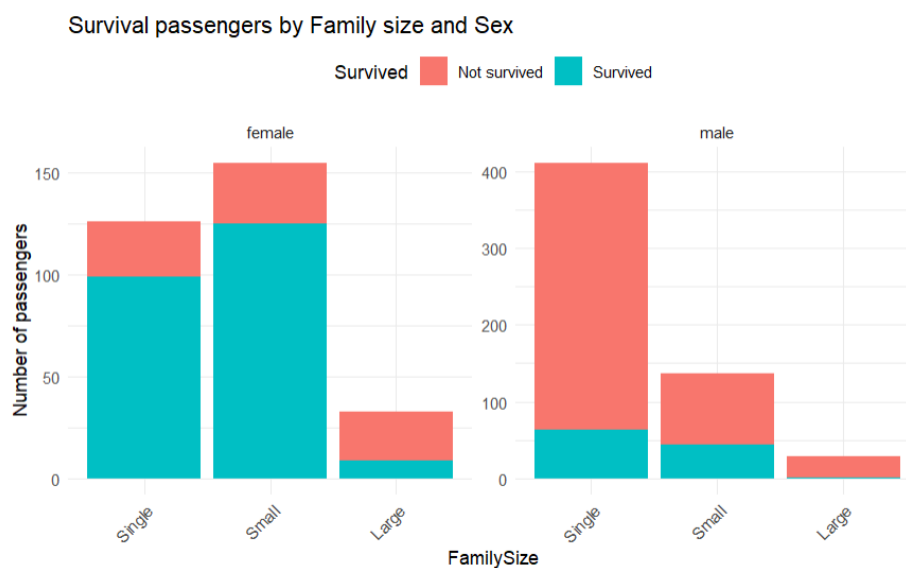


Figure 2: Survival Passengers by Family Size and Sex

There is a significant difference in survival rates between males and females, as shown in figure 2, where the survival rate of males is much lower than that of females. This indicates male who are be single have the highest risk to decease in this disaster because of the loss of families' assistance. And male would like to rescue other female got rid of this tragedy and acted as a hero, this is another reason for male to have high death rate.

2.3.2 Title, Age Disparities in Survival Outcomes

In the group of children, different titles have little effect on survival rate. In tremendous contrast, as shown in figure 3, titles make huge influences on survival rate in adult group caused by enormous difference between survival and death number of people in the title of "Mr.". And this condition also appears in the title "Mr." of teens and seniors. The title of people which is called "Miss." or "Ms." usually have lower death rate compared with the "Mr.". Because maybe young men chose to safe their wife or parents. The titles in passenger names (such as "Mr.", "Mrs.", "Master") imply age, gender, and social status, which directly affect the priority of rescue. In terms of age, children and teens may gain more rescue supply at the first time.

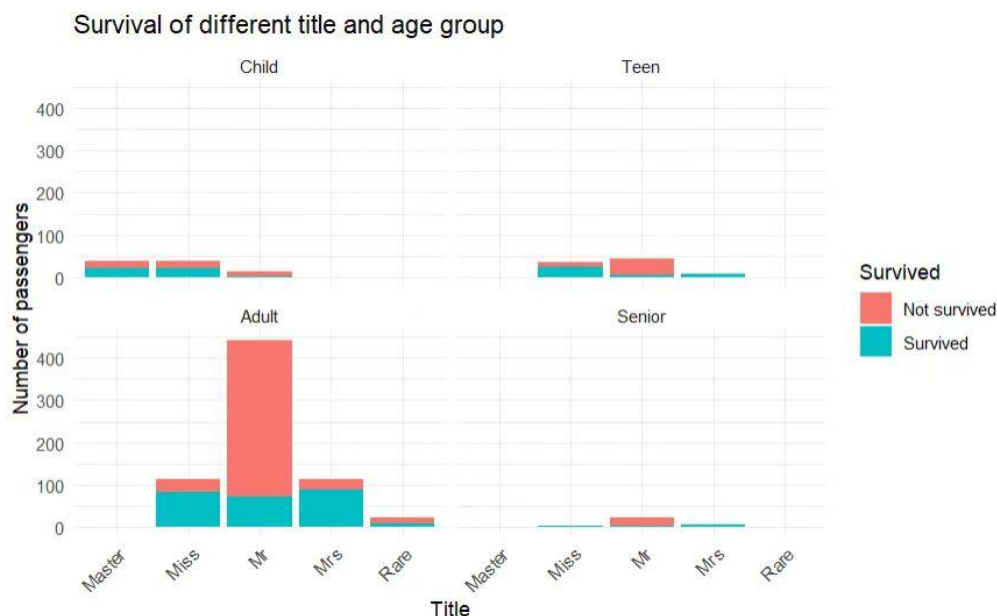


Figure 3: Survival of Different Title and Age Group

2.3.3 Survival Disparities across P-Class

At the second and third P-class, there is not a very huge difference between the number of death and survival people. But at the first P-class, it is very direct to observe that the survival rate in it is much higher than others. As shown in figure 4, this situation can explain by different social statuses, as well as the unequal distribution of social resource, people who sit at the Second and Third P-class are hardly possible to obtain the same rescue supply.

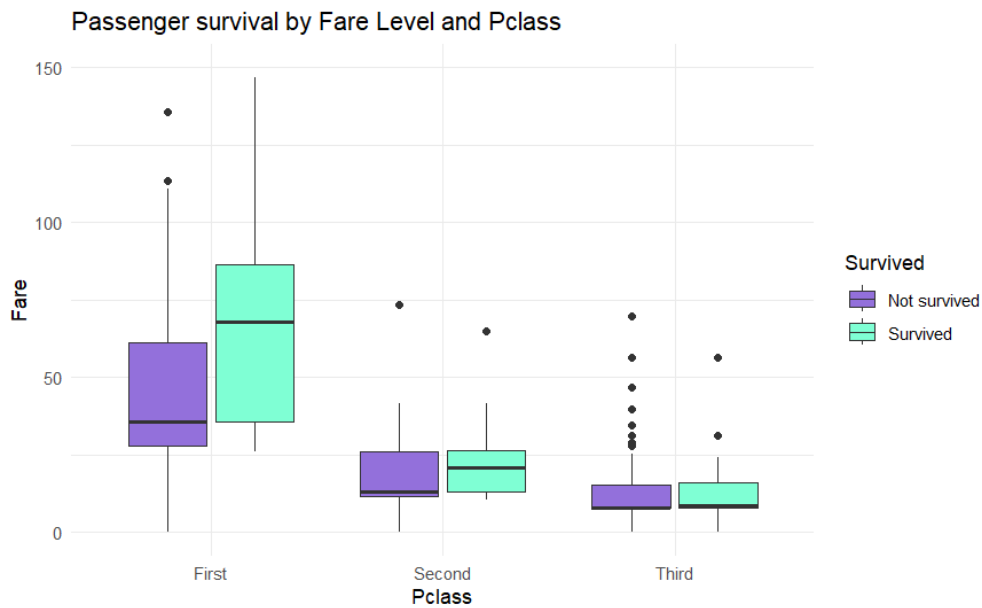


Figure 4: Passenger Survival by Fare Level and P-class

2.3.4 Combined Effects of Gender, Age, and P-class on Survival

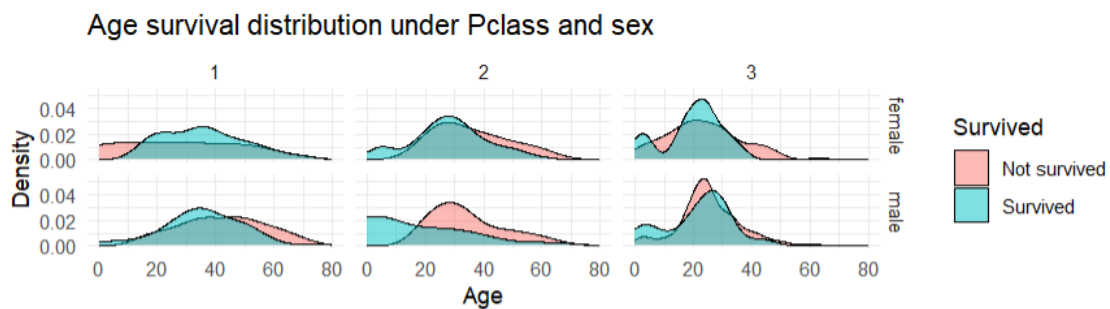


Figure 5: Age Survival Distribution Under P-class And Sex

The figure 5 indicates clearly that the higher the cabin class, the higher the survival rate, especially for women, as well as revealing the principle of prioritizing the rescue of children and women is evident. Meanwhile, the age distribution shows that children who are under 10 years old have a significant survival advantage. Higher survival rate of passengers which are on the first P-class illustrates difference between dissimilar social status will affect the rate too.

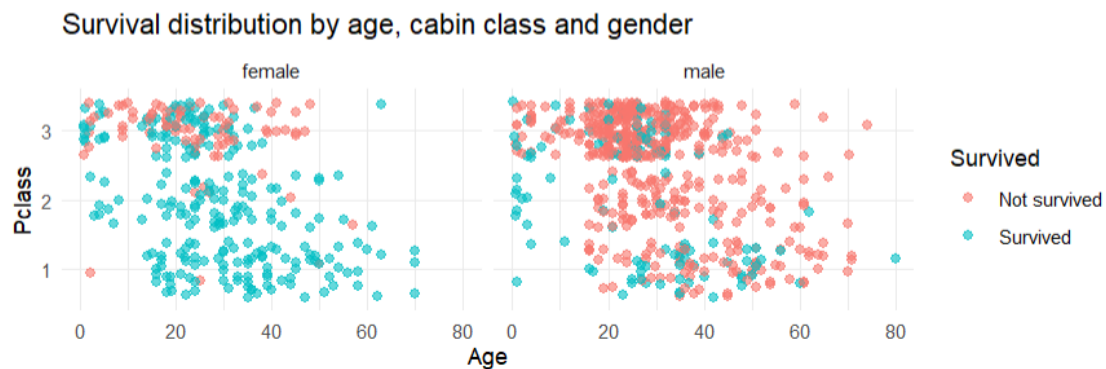


Figure 6: Survival Distribution by Age, Cabin class And Gender

As shown in figure 6, the comprehensive analysis can globally demonstrate the impact of cabin class, age, and gender on survival rates during the Titanic disaster, revealing the rescue principle of "women and children first" and social class differences at that time.

3. Statistical Modelling Approach

The major thought we have to build up an entire model is combining 4 individual models and put their outputs together; therefore, we would have a number of accurate and logical results which can be practised and carried out in the real life. For random forest, it can make sure the results will be more effective than only use few decision trees, and the logistic regression consists of two types, adopts stepwise and forward stepwise.

3.1 Multicollinearity Diagnostics among Predictors

3.1.1 Correlation Analysis

The explanatory variables are assumed to be independent. Correlation analysis revealed high R-values between certain variables in the correlation matrix, indicating the presence of multicollinearity. This may occur because these variables are either mathematically derived from one another or measure similar underlying constructs. For example, Fare and P-class have a relatively high r value because the two variables have some certain correlation and include the same group of people.

3.1.2 Chi-Square Matrix

The explanatory variables have no multicollinearity between them [3]. Putting two variables into Chi-square test and check out their P-value, it is easy for us to find out that the P-value is more than 0.05 so that the null hypothesis will be accepted. Different types of data have some association between them and for the survival rate, something may be explained by strong relationship among different variables.

3.2 Logistic Regression Model Specification

Logistic regression is a supervised learning algorithm for binary classification that models the probability of an outcome using a sigmoid function. It estimates the relationship between independent variables and the log-odds of the dependent variable through maximum likelihood estimation. The model outputs probabilities between 0 and 1, which are then impeded to make class predictions. Key advantages include interpretability through coefficient analysis and efficiency with linear decision boundaries. However, it assumes linearity between features and log-odds, requires absence of multicollinearity, and may underperform with complex nonlinear relationships. In this paper, we need to use it to produce some accurate predictions and find out results.

3.3 Decision Tree Classification via Recursive Partitioning

A decision tree is a non-parametric supervised learning algorithm that recursively partitions data into subsets based on feature values to maximize homogeneity. It uses impurity measures (Gini index/entropy) to select optimal splits, forming a tree-like structure of decision nodes and leaf nodes. Simple to interpret yet prone to overfitting, it handles both classification and regression tasks. To mitigate overfitting, pruning techniques which remove branches that contribute little to predictive accuracy may be applied. Furthermore, the inherent instability of individual decision trees motivates ensemble approaches such as Random Forests, which aggregate multiple trees to achieve greater

stability and generalisation performance. In this study, while the single decision tree is examined as an interpretable baseline, its ensemble extension is also considered to capture potential non-linear interactions that may be overlooked by logistic regression.

3.4 Random Forest Ensemble via Bootstrap Aggregation

Random Forest is an ensemble learning method that constructs multiple decision trees through bootstrap sampling and feature randomization. It aggregates predictions via majority voting in the classification type or averaging in the regression type [4], enhancing accuracy and reducing overfitting. Robust to noise and missing data, it provides feature importance metrics while maintaining interpretability.

4. Empirical Findings and Model Performance

4.1 Feature Selection and Variable Filtering

Before constructing the predictive model, it is essential to perform preliminary feature screening to eliminate redundant variables and identify key influencing factors. In this study, correlation analysis was employed to initially explore the linear association strength among numerical features.

Table 1 presents the Pearson correlation coefficient matrix for the four core numerical variables: Age, SibSp, Parch, and Fare. As shown in the table, the correlations among all variable pairs are generally weak, with no absolute correlation coefficient exceeding 0.45, indicating the absence of severe multicollinearity. Notably, SibSp (number of siblings/spouses) and Parch (number of parents/children) exhibit a correlation coefficient of 0.4087, suggesting a weak positive relationship. This is logically consistent, as passengers with larger family sizes tend to have more accompanying relatives across both categories. In addition, Age shows weak negative correlations with SibSp (-0.2837) and Parch (-0.2042), implying that younger passengers may be more likely to travel with larger family groups. Overall, the low inter-variable correlations indicate that multicollinearity is not a concern, and all four variables can be preliminarily retained for subsequent modeling.

Table 1: Variable relationship between numerical values

	Age	SibSp	Parch	Fare
Age	1.0000000	-0.2837086	-0.2041874	0.1074368
SibSp	-0.2837086	1.0000000	0.4086644	0.1832983
Parch	-0.2041874	0.4086644	1.0000000	0.2226807
Fare	0.1074368	0.1832983	0.2226807	1.0000000



Figure 7: Heatmap of the Feature Correlation Matrix

Figure 7 visualizes the above correlation matrix using a heatmap. In this figure, color intensity

corresponds to the magnitude of the correlation coefficient, with the diagonal cells representing perfect positive correlation (coefficient=1) with the variable itself. The heatmap clearly reveals that Fare has correlation coefficients below 0.23 with all other three variables, indicating that fare is relatively independent of the other demographic features and may contribute additional incremental information to the model. Based on this preliminary screening, further feature selection methods—such as feature importance ranking and Recursive Feature Elimination (RFE)—will be applied in subsequent steps to determine the final optimal input variable set.

Table 2 presents the p-value matrix for the pairwise independence tests. Each cell in this matrix represents the p-value from the Chi-square test between the corresponding two variables, with the null hypothesis assuming that the two variables are independent. At a significance level of $\alpha = 0.05$, if $p < 0.05$, the null hypothesis is rejected, indicating a statistically significant correlation between the two variables; if $p \geq 0.05$, the null hypothesis cannot be rejected, suggesting that the variables may be independent. As observed from Table 2, the majority of variable pairs yield p-values below 0.05, indicating widespread correlations among categorical variables. Notably, P-class exhibits p-values far below 0.05 with almost all other variables, suggesting that passenger class is significantly associated with multiple features.

Table 2: Pairwise P-value for Independence Tests

	Title	Sex	Age Group	Family Size	Fare	Category	Embarked
P-class	0.000000e+00	5.836577e-10	2.303567e-02	2.500778e-06	9.989261e-11	1.212998e-112	3.448800e-19
Title	5.836577e-10	0.000000e+00	1.388877e-150	1.935387e-67	4.116176e-43	1.965986e-18	8.856999e-05
Sex	2.303567e-02	1.388877e-150	0.000000e+00	3.781651e-03	3.545855e-14	1.152319e-09	2.018321e-02
Age Group	2.500778e-06	1.935387e-67	3.781651e-03	1.470636e-33	0.000000e+00	3.751332e-05	8.224403e-01
Family Size	9.989261e-11	4.116176e-43	3.545855e-14	1.470636e-33	0.000000e+00	1.826064e-41	2.911225e-05
Fare Category	1.212998e-112	1.965986e-18	1.152319e-09	3.751332e-05	1.826064e-41	0.000000e+00	4.177160e-13
Embarked	3.448800e-19	8.856999e-05	2.018321e-02	8.224403e-01	2.911225e-05	4.177160e-13	0.000000e+00

To further investigate the predictive power of each categorical variable with respect to the target variable (Survival), the Weight of Evidence (WOE) and Information Value (IV) were calculated for each variable. WOE measures the ability of a particular feature category to distinguish between survival and non-survival outcomes, while IV evaluates the overall predictive strength of the variable. Generally, an IV value below 0.02 indicates extremely weak predictive power, $0.02 \leq IV < 0.1$ indicates weak to moderate, $0.1 \leq IV < 0.3$ indicates moderate, $0.3 \leq IV < 0.5$ indicates strong, and $IV \geq 0.5$ indicates extremely strong predictive power.

Tables 3-11 present the WOE and IV calculation results for each categorical variable, including P-class, Sex, Age Group, SibSp, Parch, Fare, Embarked, Fare Category, and Title.

Table 3: P-class Table

P-class	N	Percent	WOE	IV
1	162	0.22720900	1.0019996	0.2360596
2	153	0.2145863	0.3798041	0.2680357
3	398	0.5582048	-0.6339524	0.4707607

Table 4: Sex Table

Sex	N	Percent	WOE	IV
Female	257	0.3604488	1.3958476	0.7052462
Male	456	0.6395512	-0.9314524	1.1758586

Table 5: Age Table

Age	N	Percent	WOE	IV
[0,5]	142	0.1991585	0.330292708	0.022379042
[6,12]	115	0.1612903	-0.355307266	0.041741563
[13,18]	167	0.2342216	-0.107112966	0.04439342
[19,24]	145	0.2033661	0.065906199	0.04528323
[25,30]	144	0.2019635	-0.009931878	0.04530313

Table 6: SibSp Table

SibSp	N	Percent	WOE	IV
[0,0]	482	0.6760168	-0.1632075	0.017634712
[1,8]	231	0.3239832	0.3239186	0.05263437

Table 7: Parch Table

Parch	N	Percent	WOE	IV
[0,0]	541	0.7587658	-0.1668170	0.020668022
[1,6]	172	0.2412342	0.4946282	0.08195066

Table 8: Fare Table

Fare	N	Percent	WOE	IV
[0,7.88]	142	0.1991585	-0.7633732	0.1021249
[7.88,10.46]	132	0.1851332	-1.0327061	0.2657287
[10.5,20.52]	152	0.2131837	0.1798505	0.2727517
[20.58,35.5]	143	0.2005610	0.2608065	0.2867367
[36.75,512.33]	144	0.2019635	0.9526745	0.4768265

Table 9: Embarked Table

Embarked	N	Percent	WOE	IV
<NA>	2	0.002805049	0.00000000	0.00000000
C	127	0.178120617	0.64503480	0.07721477
Q	60	0.084151473	-0.00405239	0.07721615
S	524	0.734922861	-0.17631350	0.09954886

Table 10: Fare Category Table

Fare Category	N	Percent	WOE	IV
Cheap	238	0.3338008	-0.827911677	0.1985636
Expensive	230	0.3225806	0.716063194	0.3709448
Mid	245	0.3436185	-0.002613748	0.3709471

Table 11: Title Table

Title	N	Percent	WOE	IV
Master	31	0.04347826	0.93090364	0.0391056
Miss	152	0.21318373	1.18432111	0.3450958
Mr	407	0.57082749	-1.17085643	0.9725454
Mrs	102	0.14305750	1.70532494	1.3748816
Rare	21	0.02945302	-0.01413651	1.3748875

Above all, these data are called IV value which can evaluate the predictive strength. By quantifying how well a feature separates target classes, IV helps analysts select meaningful variables during feature engineering. It's calculated through Weight of Evidence analysis, where values below 0.02 suggest useless predictors, 0.02-0.1 indicate weak predictive power, 0.1-0.3 show medium strength, and scores above 0.3 reveal strong predictors though potentially overfitting.

4.2 Logistic Regression Results and Coefficient Interpretation

Logistic regression is a statistical model for binary relationship between features and the log-odds of outcomes, optimized via maximum likelihood. After executing the algorithm, this model remains a variable: P-class, Sex, Age, Family Size, and Title. They are all negatively correlated with survival rate.

These table above this can show the correlation between variables easily and clearly, and the accuracy of this model is 81.21%.

4.3 Decision Tree Classification and Pruning Results

The decision tree was initially executed using default parameters for baseline evaluation. As shown in figure 8, the model automatically generated a tree structure with 6 depth levels and 7 leaf nodes,

employing the Gini impurity algorithm for branch splitting. As a metric for measuring dataset purity, the Gini coefficient guides the algorithm to partition features by maximizing purity in child nodes, achieving a final model accuracy of 89.12%.

This tree can be time-efficiency to work out a relatively accurate result. Based on these results, when the maximum depth of the tree is in the middle layer, performance is better and accuracy is higher and the performance of Gini coefficient is greater than information gain. Compared to the original tree, the branches of this model are cut after the Ticket, so the previous tree may have the following issues: as one grows older, boarding and ticket prices do not really bring a lot of information. So, this tree will calculate time and depth to reduce the complexity of the model.

Decision Tree Structure

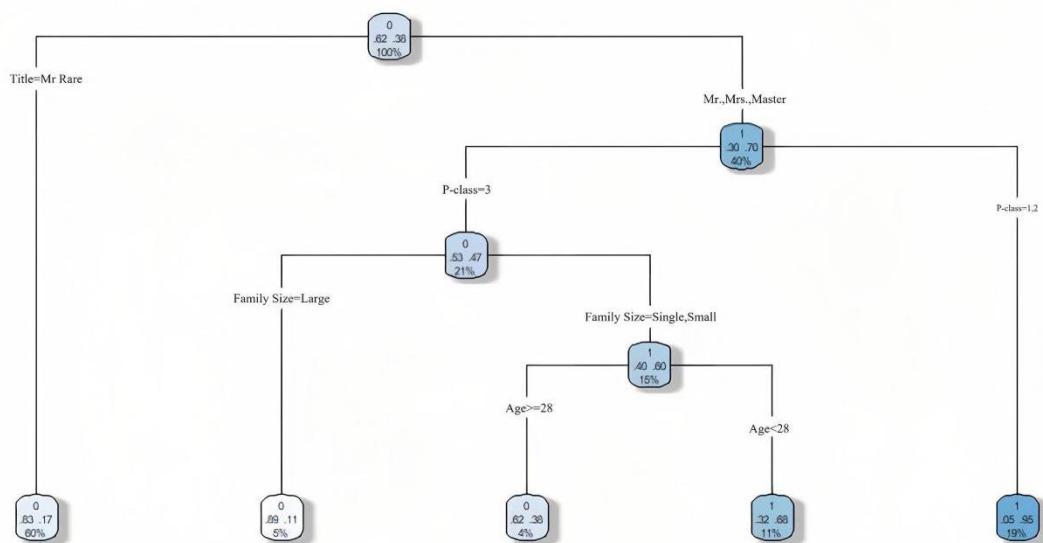


Figure 8: Decision Tree Structure

4.4 Random Forest Classification Results

Random forest is a kind of classification regression method which employs many decision trees to build up a collective forest and more importantly, to combine their outputs that weigh the possibility more logically. It usually has robustness and can help improve the accuracy of prediction. Random forest addresses the limitations of decision trees and provides a more comprehensive processing of data. Therefore, these trees are taken and the model gets an accuracy of 83.21%.

4.5 Hard Voting Results and Summary of Findings

Hard voting combines predictions from multiple machine learning models by selecting the majority class in the classification type or averaging outputs in the regression type. This ensemble method reduces individual model biases and variance, often improving overall accuracy. It works best with diverse, uncorrelated base models to maximize performance gains through collective decision-making.

In conclusion, previous data prediction can get an 82.74% accuracy which can export output to explain the cause of various results. These data have multi-collinearity and false negatives are more than false positives that is why we can use them to find out certain characteristics of variables and some accurate predictions.

5. Conclusion

Overall, in the paper we are dedicated to organise data and build up several models in order to predict a target whether are like to survive in this disaster or not. Thanks to the hard-voting strategy that combines three types of prediction models, including logistic regression, decision trees and random forests, the model achieves better predictive performance. By using these models, this hard voting system can carry out a specific prediction. Moreover, these models all play important roles, decision trees can deal with the missing data and print out a relatively accurate result that can predict the survival rate, random forest would complement the other models by capturing complex non-linear relationships. This system can also be used into the predictions of some huge disasters survival rate government want to prevent the death of people as much as possible. As well as using in the navigation field is a great choice. This theory would impact the multiply data analysing in the real life.

Limitations in this article is about fewer models that cannot work out a further result, and lack some accurate data which reveals accuracy.

References

- [1] Megan L. Risdal. (2017). *Exploring Survival on the Titanic*. Kaggle.com; Kaggle. <https://www.kaggle.com/code/mrisdal/exploring-survival-on-the-titanic>.
- [2] Hiteshp. (2018). *Head Start for Data Scientist*. Kaggle.com; Kaggle. <https://www.kaggle.com/code/hiteshp/head-start-for-data-scientist#feature-engineering>.
- [3] Gregory, D. N., & Bader, K. (2018). *Logistic and linear regression assumptions: violation recognition and control*. In *Proceedings of Midwest SAS User Group 2018 conference, Indiana, AA-091* (pp. 1-22).
- [4] Shareef, A. Q., & Kurnaz, S. (2023). *Deep Learning Based COVID-19 Detection via Hard Voting Ensemble Method*: AQ Shareef, S. Kurnaz. *Wireless Personal Communications*, 1-12.