Predict Survival on the Titanic Using MATLAB

Introduction

The sinking of the RMS Titanic is one of the most infamous shipwrecks in history. On April 15, 1912, during the maiden voyage, the Titanic sank after colliding with an iceberg, killing 1502 out of 2224 passengers and crew. This sensational tragedy shocked the international community and led to better safety regulations for ships [1].
One of the reasons that the shipwreck led to such loss of life was that there were not enough lifeboats for the passengers and crew [2]. Although there was an element of luck involved in surviving the sinking, some groups of people were more likely to survive than others.
The goal of this project is to complete the analysis of what sorts of people were likely to survive using three different data mining methods and also their modifications.
The idea as well as data for the project were given by kaggle.com [1].

First data analysis

The next features are included in the given data (features marked as bold were included in the analysis):

Feature	Include in the analysis	Modification
Passenger ID	No. It is just a set of serial numbers in the list of passengers.	NA
Passenger class	Yes. First class - 1, Second class - 2, Third class - 3	No
Sex	Yes	Yes. ‘female’ -> 0, ‘male’ -> 1
Age	Yes	No
Number of siblings/ spouses aboard	Yes	No
Number of parents/ children aboard	Yes	No
Passanger fare	Yes	No
Port of embarkation	Yes	Yes

Ticket number	No. Tickets are all different, and also include string symbols and special symbols as well as numbers.	NA
Cabin	No. Cabin number applied just for 22.9% of the data, including multiple variations for some passengers	NA

Table 1. Changes in data set and used features

All data was divided into training and validation part as 80% and 20% correspondingly. The next percentage of survived passengers is included into training, validation and all data set:

Data set	Percent of survived passengers
All data	38.4%
Validation data	38.6%
Training data	38.3%

Table 2. Percentage of survived passengers in data sets

Data visualisation

Firstly, it is interesting to know relations between some of the features.

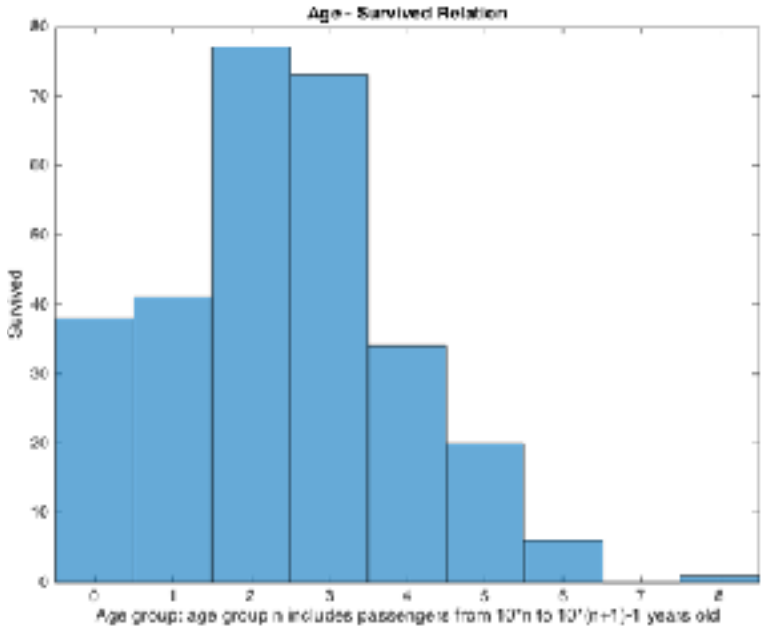
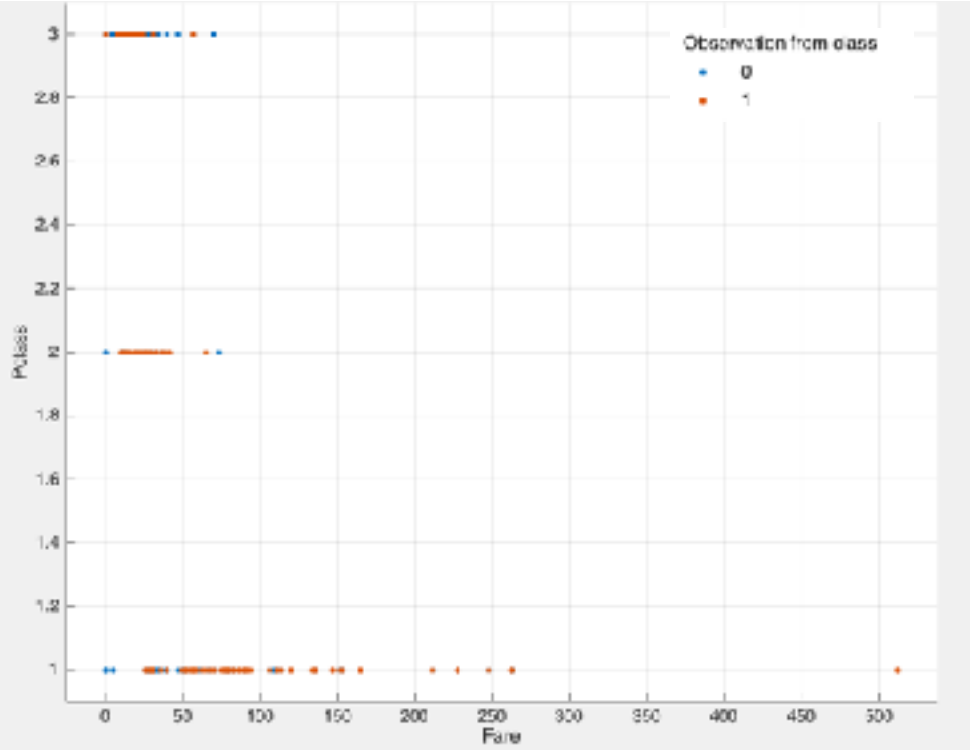


Fig. 1 Age-Survived passengers relation

While analyzing the data, the hypothesis of strong connection between 'Passenger fare' and 'Passenger class' features and, as a result, unavailability for one of them came to me. Therefore, the Figure 2 is shown to reject the hypothesis.



Fug. 2 Passenger fare - passenger class relation, where true class (Survived) is marked as '1'

Here it can be clearly seen that the passenger was able to buy a ticket for any of the classes in the same range of fare.

Methods and performance

Thinking about methods which is better suit to this project, I chose Naive Bayes, k-NN and Decision Tree methods. In the beginning of this paragraph I am going to present the performance each model showed for the set of features selected in Table 1.

1. Naive Bayes

Accuracy on the training data is **79.9%.** To evaluate the model, prediction for validation data was involved. Validation data was used only as an accuracy measure as usual.
The accuracy on the validation data is **76.5%** according to the confusion matrix shown below.

		Real		
		0	1	Totall
Prediction	0	86	24	110
	1	18	51	69
	Totall	104	75	179

Table 3. Confusion matrix on the validation data for Naive Bayes method

2. k-NN

The maximum accuracy was found while the number of neighbors is equal to 15. Accuracy on the training data set is **78.5%.** To evaluate the model, the prediction for ***validation data*** was involved.
The accuracy on the validation data is **77.7%** according to the confusion matrix shown below.

		Real		
		0	1	Totall
Prediction	0	100	10	110
	1	30	39	69
	Totall	130	49	179

Table 4. Confusion matrix on the validation data for k-NN method

3. Decision Tree

The corresponding accuracy on the training data is **82.7%.** To evaluate the model, the prediction for validation data was involved.
Confusion matrix on the ***validation data***:

		Real		
		0	1	Totall
Prediction	0	94	16	110
	1	22	47	69
	Totall	116	63	179

Table 5. Confusion matrix on the validation data for decision tree method

The accuracy on the validation data is **78.8%**
Figure connected with Fig.2 but for Decision Tree performance is shown in APPENDIX (Fig. A2).

4. Modifications of the models

Comparing these three methods on the performance on validation data, I decided to create one more (connected with others) feature and try Decision Tree method on the new set of features.
The additional feature is called 'Child' and contains logical values when '1' means that the passenger is less then 16 years old and '0' is the opposite.

The performance on the validation data for Decision Tree method with the addition feature is **76.5%**. The accuracy is given by confusion matrix shown below.

		Real		
		0	1	Totall
Prediction	0	88	22	110
	1	20	49	69
	Totall	128	71	179

Table 6. Confusion matrix on the validation for the decision tree method with addition 'Child' feature

Modification of the 'Sex' feature in a way to add additional '2' for the passenger who is less then 16 years old (in this case 'Child' feature was excluded) doesn't improve the accuracy too. The corresponding accuracy on the validation data is **74.3%**

That is why, for the first set of features (Table 1) I found the rating of top features from the most important to the less important (the tree visualization is shown in APPENDIX (Fig. A1)):

1) Sex
2) Passenger class
3) Passenger fare
4) Age
5) Number of parents/children aboard
6) Port of embarkation
7) Number of siblings/spouses abroad

As the last step to improve the model I tried k-NN method (showed the second performance success on the validation data) with the reduced number of features. The first five features in rating shown above were used. The corresponding accuracy on the validation data was **77.9%**. This means that k-NN model was slightly improved by correctly decreasing the number of features.

Conclusion

Three main models and their modifications were used. The overall performance of the models is shown below.

Method	Accuracy on the validation data
Naive Bayes	76.5%
k-NN	77.7%
Decision Tree	78.8%
Decision Tree with additional feature	76.5%
Decision Tree with 'Sex' feature modification	74.3%
k-NN with less number of features (Decision Tree + k-NN)	77.9%

Table 7. Models performance

I chose Decision Tree model as a final one based on the Table 7.
This project is an active competition on kaggle.com [1]. Therefore, there is a test file with test data on the website. The submission on [1] was done with the performance **0.75120.** So, accuracy for the test data in precent is **75.12%.**

References

[1] Kaggle link to the competition: https://www.kaggle.com/c/titanic
[2] Wikipedia link: https://en.wikipedia.org/wiki/RMS_Titanic

APPENDIX

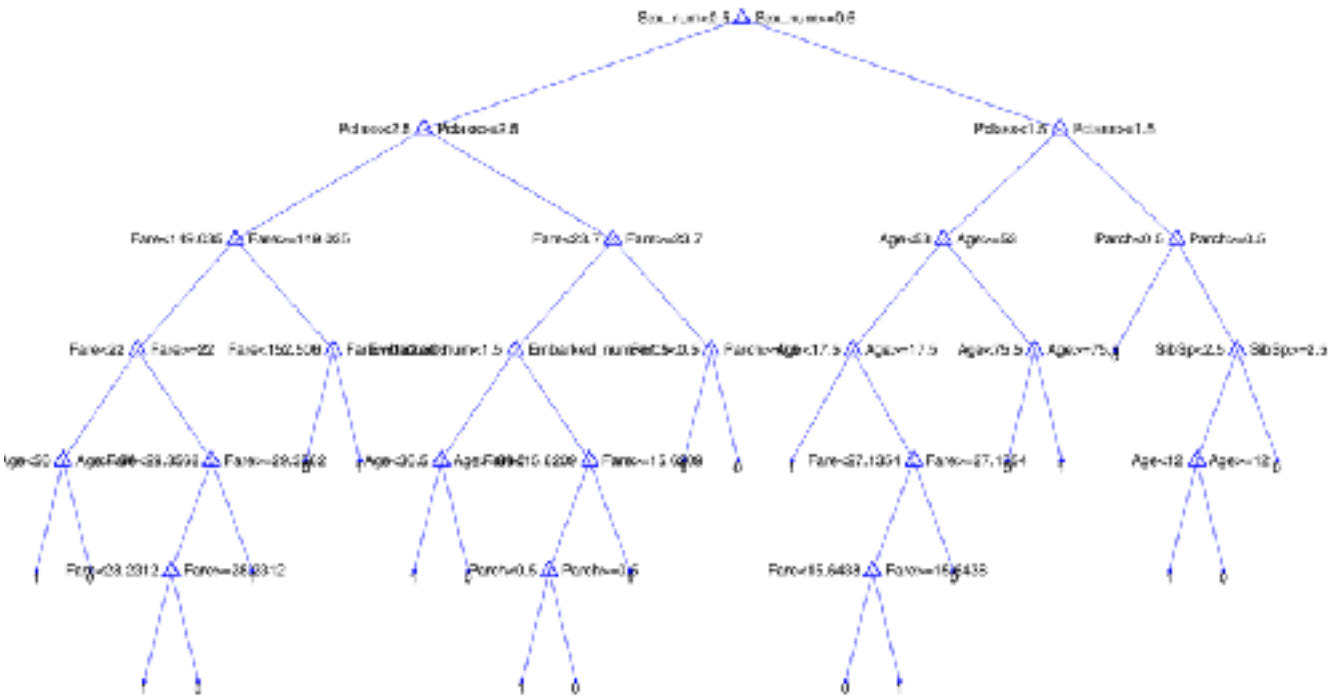


Fig. A1 Decision tree for the set of features from Table 2.

Fig. A2 Passenger fare - passenger class relation, where true class (Survived) is marked as '1'. Results for Decision Tree model.