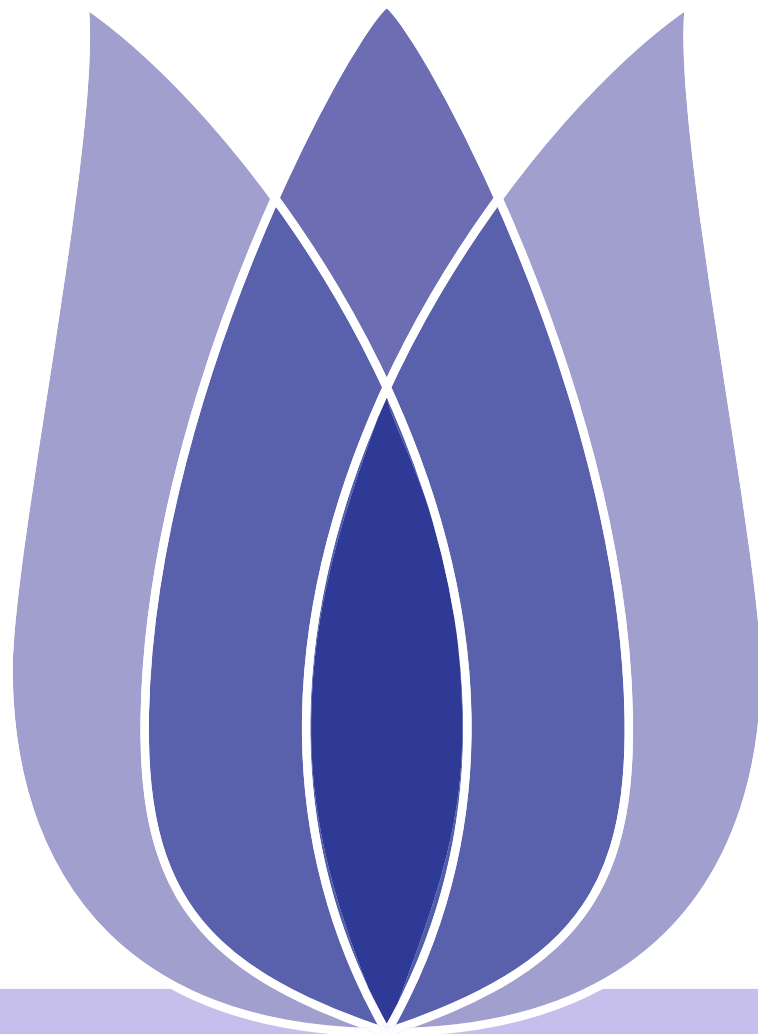


April Tabular Playground Series - Your Baseline Model

Jinzhu Liu

Nanjing University of Science and Technology
Deakin University, Australia

2023-01-28





Overview

- [Problem Definition](#)
- [Data analysis and processing](#)
- [ML Models Implementation](#)
- [Conclusion](#)

Problem Definition

Predict Whether or not A Passenger Survived the Inking of the Synthanic

Data analysis and processing

- Inquire NaN Values
- Studying Distribution of Age and Fare
- Imputing Values

ML Models Implementation

- Overall steps
- Implementing it on Test Data

Conclusion

Conclusion



- Problem Definition**
Predict Whether or not A Passenger
Survived the Inking of the Synthanic
- Data analysis and processing
- ML Models Implementation
- Conclusion

Problem Definition



Predict Whether or not A Passenger Survived the Sinking of the Titanic

Problem Definition
Predict Whether or not A Passenger
Survived the Sinking of the Titanic

Data analysis and processing

ML Models Implementation

Conclusion

Introduce

it is extremely important to start somewhere and identify it as your first standard of comparison against the progress you have. This helps you make a baseline model, get a baseline score.

- Our task is to predict whether or not a passenger survived the sinking of the Titanic.

Interpretation

Variable Notes

- pclass: A proxy for socio-economic status (SES) 1st = Upper 2nd = Middle 3rd = Lower
age: Age is fractional if less than 1. If the age is estimated, is it in the form of xx.5
sibsp: Describes the number of Siblings and spouses accompanying passengers on the Titanic
parch: Describes the number of parents and children travelling with passengers on the Titanic



TULIP

Team for Universal Learning and Intelligent Processing



Predict Whether or not A Passenger Survived the Inking of the Synthanic

Problem Definition
Predict Whether or not A Passenger
Survived the Inking of the Synthanic

Data analysis and processing

ML Models Implementation

Conclusion

	Passengerid	Survived	Pclass	Name	Sex	Age	SibSp
0	0	1	1	Oconnor, Frankie	male	NaN	2
1	1	0	3	Bryan, Drew	male	NaN	0
2	2	0	3	Owens, Kenneth	male	0.33	1
3	3	0	3	Kramer, James	male	19.00	0
4	4	1	3	Bond, Michael	male	25.00	0

	Parch	Ticket	Fare	Cabin	Embarked
0	0	209245	27.14	C12239	S
1	0	27323	13.35	NaN	S
2	2	CA457703	71.29	NaN	S
3	0	A.10866	13.04	NaN	S
4	0	427635	7.76	NaN	S



[Problem Definition](#)

[Data analysis and processing](#)

[Inquire NaN Values](#)
[Studying Distribution of Age and Fare](#)
[Imputing Values](#)

[ML Models Implementation](#)

[Conclusion](#)

Data analysis and processing



Inquire NaN Values

- Problem Definition
- Data analysis and processing
- Inquire NaN Values**
- Studying Distribution of Age and Fare
- Imputing Values
- ML Models Implementation
- Conclusion

PassengerId	0
Survived	0
Pclass	0
Name	0
Sex	3292
Age	0
SibSp	0
Parch	0
Ticket	4623
Fare	134
Cabin	67866
Embarked	250

- list out the columns holding Numerical Values - 1.Age 2.Fare
- The remaining columns do not hold numerical values. Let’s explore the distribution of the numerical values a bit before we replace their NaN values



Studying Distribution of Age and Fare

Fare

- Problem Definition
- Data analysis and processing
- Inquire NaN Values
- Studying Distribution of Age and Fare
- Imputing Values
- ML Models Implementation
- Conclusion

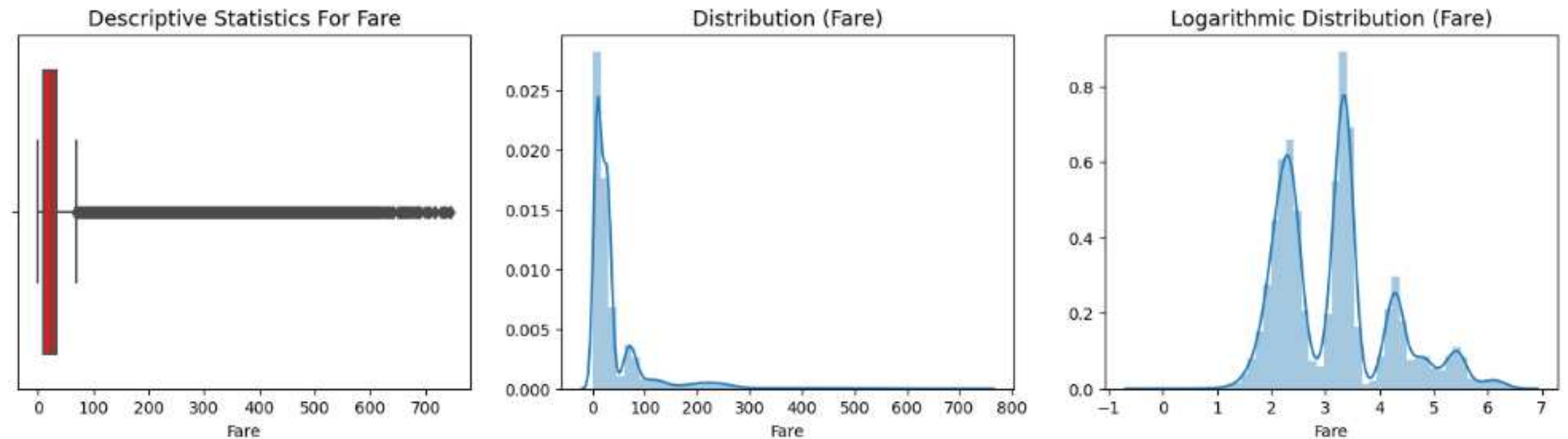


Figure 1: Descriptive Statistics For Fare

- From the boxplot , we can see that anything above 100 looks like an outlier and that there are a lot of outliers. The suggested median seems to be somewhere between 0-100.
- The Distribution Plot suggests that the data is left skewed.
- From the Logarithmic Distribution,we can assume that these three peaks are a result of price distribution on those three classes.



Fare distribution for each class

- Problem Definition
- Data analysis and processing
- Inquire NaN Values
- Studying Distribution of Age and Fare
- Imputing Values
- ML Models Implementation
- Conclusion

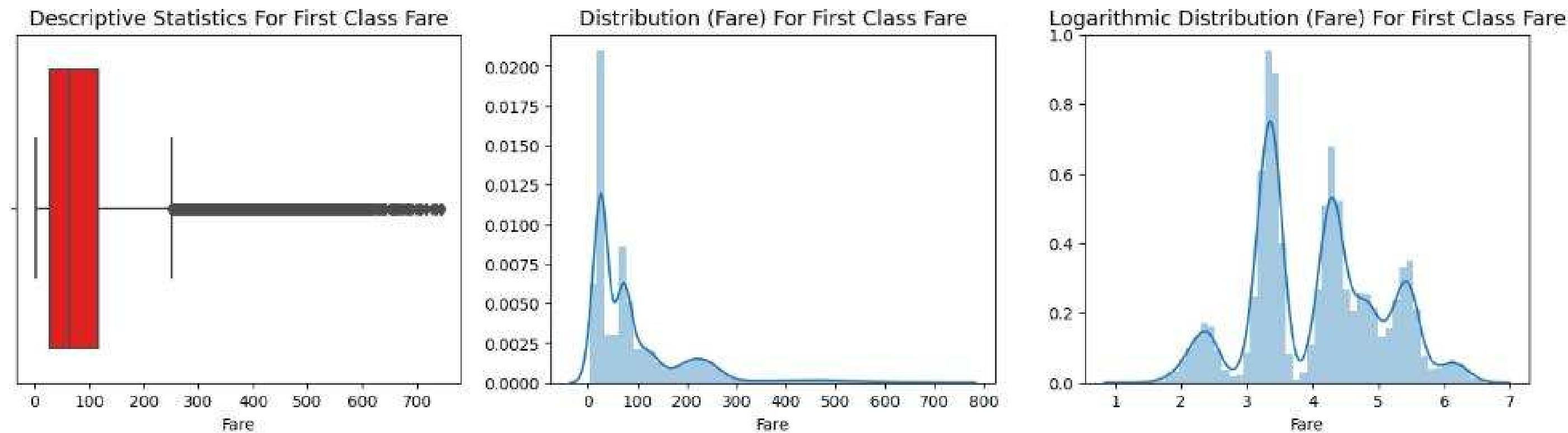


Figure 2: First Class Fare

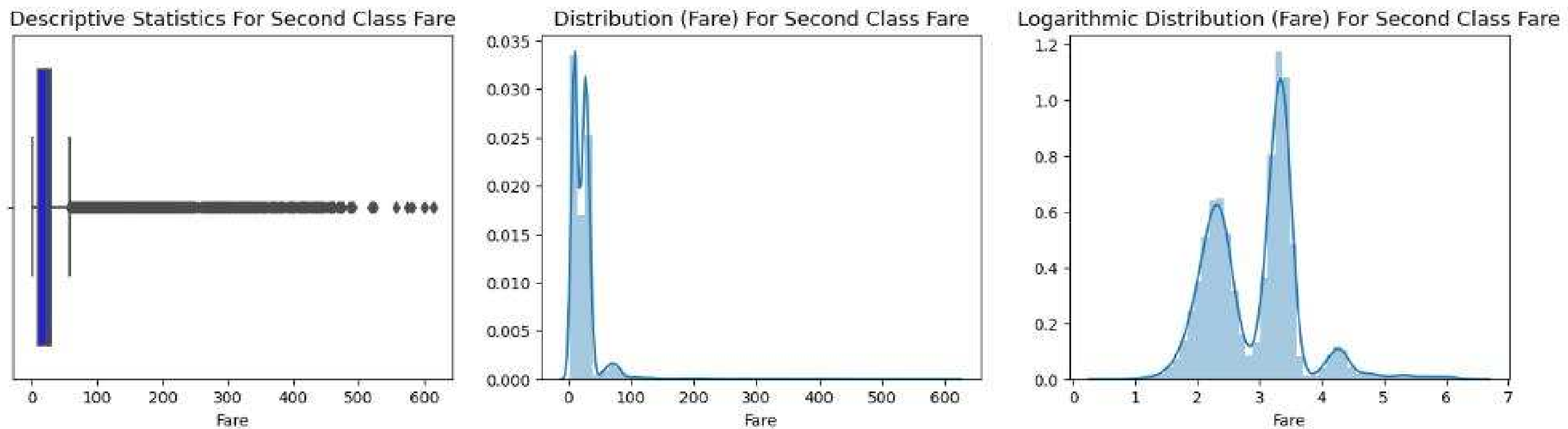
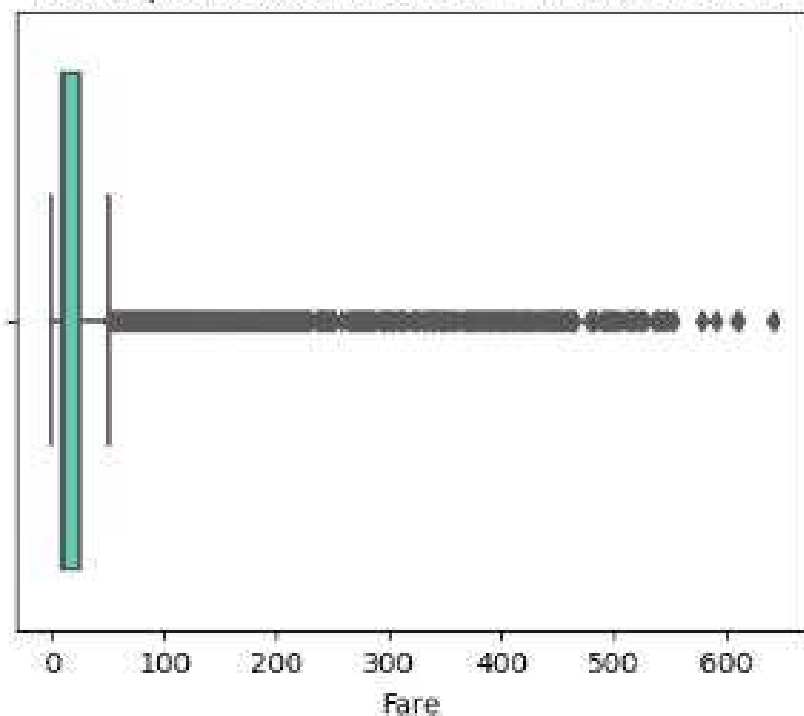


Figure 3: Second Class Fare

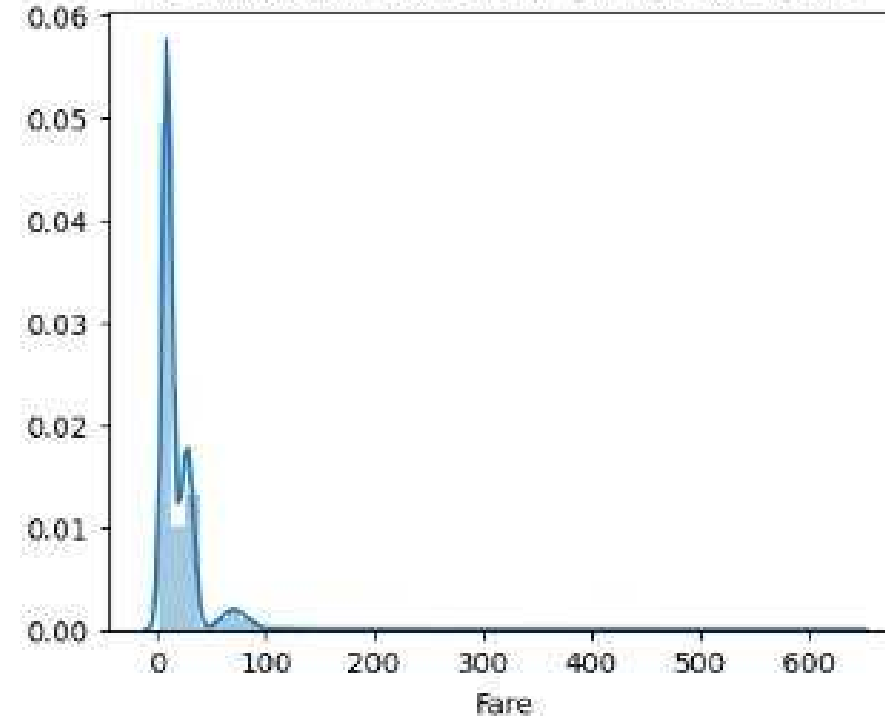


- [Problem Definition](#)
- [Data analysis and processing](#)
- [Inquire NaN Values](#)
- [Studying Distribution of Age and Fare](#)**
- [Imputing Values](#)
- [ML Models Implementation](#)
- [Conclusion](#)

Descriptive Statistics For Third Class Fare



Distribution (Fare) For Third Class Fare



Logarithmic Distribution (Fare) For Third Class Fare

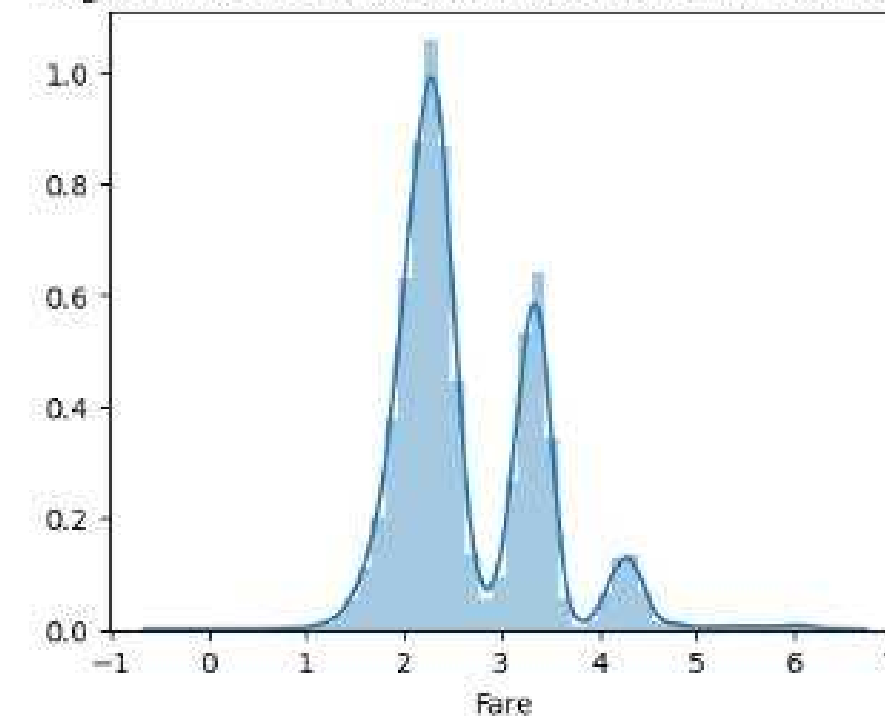


Figure 4: Third Class Fare

As you can see from the graph, Fare does not have a fixed price for any class, so it is intended to scale the data and then use the average to impute.



- Problem Definition
- Data analysis and processing
- Inquire NaN Values
- Studying Distribution of Age and Fare
- Imputing Values
- ML Models Implementation
- Conclusion

Age

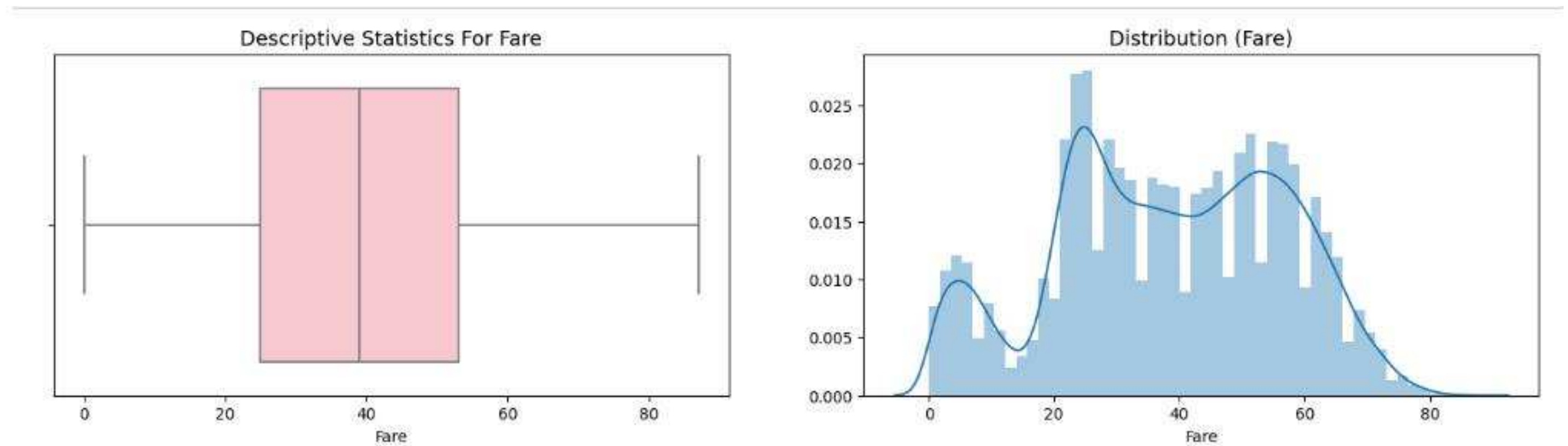


Figure 5: Descriptive Statistics For Age

- It can be seen from the figure that the missing value can be filled by the median.



Imputing Values

- [Problem Definition](#)
- [Data analysis and processing](#)
- [Inquire NaN Values](#)
- [Studying Distribution of Age and Fare](#)
- [Imputing Values](#)
- [ML Models Implementation](#)
- [Conclusion](#)

- We will use **KNN** to perform missing interpolation for Embarked.

PassengerId	0
Survived	0
Pclass	0
Name	0
Sex	0
Age	0
SibSp	0
Parch	0
Ticket	4623
Fare	0
Cabin	67866
Embarked	0

Transformation and Correlations With Target Variable

- Problem Definition
- Data analysis and processing
- Inquire NaN Values
- Studying Distribution of Age and Fare
- Imputing Values
- ML Models Implementation
- Conclusion

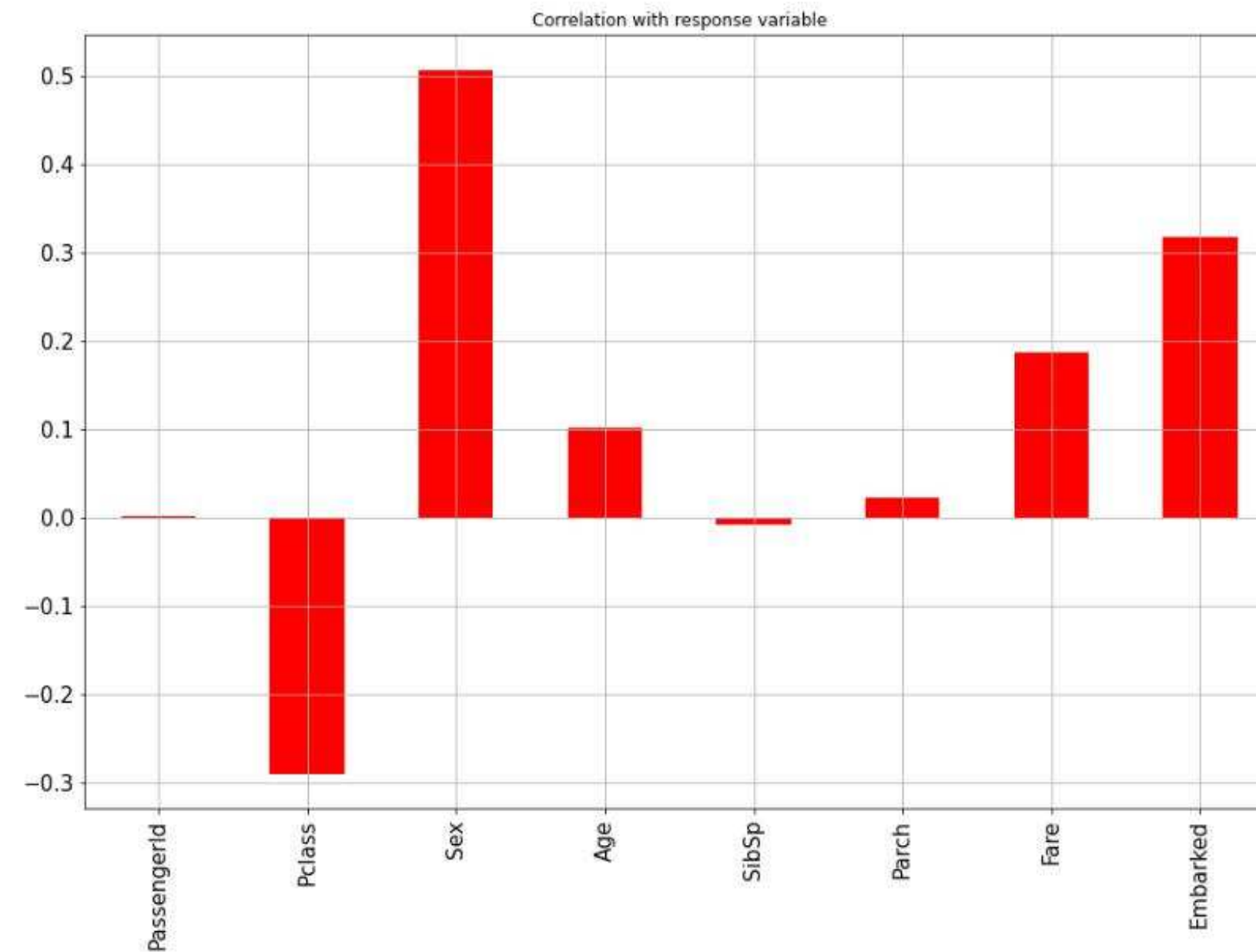


Figure 6: Transformation and Correlations With Target Variable

- from this we can see that variables **SibSp**, **Parch** , **PassengerId** have a very small correlation coefficient as compared to others. We will be dropping these variables.



Problem Definition
Data analysis and processing
Inquire NaN Values
Studying Distribution of Age and Fare
Imputing Values
ML Models Implementation
Conclusion

- Scale the data of Age and Fare so that all variables are in almost the same range.

	Pclass	Sex	Age	Fare	Embarked
0	1	0	0.034609	−0.241265	0.0
1	3	0	0.034609	−0.439429	0.0
2	3	0	−2.112537	0.393176	0.0
3	3	0	−1.075888	−0.443884	0.0
4	3	0	−0.742739	−0.519758	0.0



- [Problem Definition](#)
- [Data analysis and processing](#)
- [ML Models Implementation](#)**
- [Overall steps](#)
- [Implementing it on Test Data](#)
- [Conclusion](#)

ML Models Implementation



Overall steps

- Problem Definition
- Data analysis and processing
- ML Models Implementation
- Overall steps
- Implementing it on Test Data
- Conclusion

- First, we set up several classifier models to make a prediction. Then we use the training set to fit the model I built. After fitting, I use the fitted model to predict the remaining data in the training set and calculate its accuracy, weight, etc. Then fuse multiple groups of models, stack the fused model with the logistic regression model, and then fit the training set to get the prediction score. Use this model to predict our test set and see the prediction results of our test set.



Step One - LogisticRegression Model.

- Problem Definition
- Data analysis and processing
- ML Models Implementation
- Overall steps
- Implementing it on Test Data
- Conclusion

- We will first split them and then fit the models.

	precision	recall	f1-score	support
0	0.80	0.77	0.78	11336
1	0.71	0.74	0.73	8664
accuracy			0.76	20000
macro avg	0.76	0.76	0.76	20000
weighted avg	0.76	0.76	0.76	20000

Figure 7: LogisticRegression



Step Two - DecisionTreeClassifier Model

- Problem Definition
- Data analysis and processing
- ML Models Implementation
- Overall steps
- Implementing it on Test Data
- Conclusion

	precision	recall	f1-score	support
0	0.72	0.74	0.73	11336
1	0.64	0.62	0.63	8664
accuracy			0.69	20000
macro avg	0.68	0.68	0.68	20000
weighted avg	0.68	0.69	0.68	20000

Figure 8: LogisticRegression



Step There - KNeighborsClassifier Model

- Problem Definition
- Data analysis and processing
- ML Models Implementation
- Overall steps
- Implementing it on Test Data
- Conclusion

	precision	recall	f1-score	support
0	0.76	0.79	0.77	11336
1	0.71	0.68	0.69	8664
accuracy			0.74	20000
macro avg	0.74	0.73	0.73	20000
weighted avg	0.74	0.74	0.74	20000

Figure 9: LogisticRegression



Step Four - RandomForestClassifier Model

- Problem Definition
- Data analysis and processing
- ML Models Implementation
- Overall steps**
- Implementing it on Test Data
- Conclusion

	precision	recall	f1-score	support
0	0.74	0.76	0.75	11336
1	0.68	0.66	0.67	8664
accuracy			0.72	20000
macro avg	0.71	0.71	0.71	20000
weighted avg	0.72	0.72	0.72	20000

Figure 10: LogisticRegression



Step Five - Stacking Classifier

- Problem Definition
- Data analysis and processing
- ML Models Implementation
- Overall steps
- Implementing it on Test Data
- Conclusion

- we try Stacking Classifier and Logistic Regression on the test data.

	precision	recall	f1-score	support
0	0.79	0.80	0.79	11336
1	0.73	0.72	0.72	8664
accuracy			0.76	20000
macro avg	0.76	0.76	0.76	20000
weighted avg	0.76	0.76	0.76	20000

Figure 11: LogisticRegression



Implementing it on Test Data

- Problem Definition
- Data analysis and processing
- ML Models Implementation
- Overall steps
- Implementing it on Test Data
- Conclusion

- First, let’s see if there are missing values in the test, and fill in the missing values in the test in the same way as train. Finally, the well-fitting model is used to make predictions.

PassengerId	0
Survived	0
Pclass	0
Name	0
Sex	0
Age	3487
SibSp	0
Parch	0
Ticket	5181
Fare	133
Cabin	70831
Embarked	277



[Problem Definition](#)

[Data analysis and processing](#)

[ML Models Implementation](#)

[Conclusion](#)

Conclusion

Conclusion



Conclusion

Problem Definition
Data analysis and processing
ML Models Implementation
Conclusion
Conclusion

- Finally, the highest accuracy in LogisticRegression Model is 0.76.
- A relatively basic Kaggle project was selected, the purpose is to be familiar with the Kaggle project, deeply analyze and understand each line of the project process, this project has done more processing on the step of data feature processing, and learned a lot from it
- The work can also be further refined to improve the accuracy of prediction, for example, in the process of processing the age column, the age can be segmented according to the size of the age, and it is felt that the size of the age has a certain relationship with the size of the final survival rate.





Contact Information

jinzhu Liu
Nanjing University of Science and Technology
Deakin University, Australia

-  JINZ_LIU@NJUST.EDU.CN
-  TEAM FOR UNIVERSAL LEARNING AND INTELLIGENT PROCESSING

