

# Machine Learning Professional Project: Titanic Dataset

Citizen Data Science Program – Level 1

---

## Guidelines:

1. Answer each question with a clear and concise explanation of the Operator you used. Explain why you chose that particular operator and how it helps.
2. Outline the steps you followed to arrive at the solution.
3. Provide a screenshot of your process that demonstrated the relevant portion of your answer.
4. Name the document as the {project name - your name} before submitting it.
5. Ensure that you include both your answers word document and RapidMiner process as attachments in a reply to the person who sent you the assessment.
6. Write your name and email at the bottom of this page

## Project Description

This project consists of 17 questions related to your process in RapidMiner of the Titanic dataset. The questions are designed to assess your understanding of the key concepts and your ability to apply them in practical scenarios.

You will work on the Titanic dataset using RapidMiner. The dataset contains information of passengers in Titanic. The dataset contains the following attributes:

- Age: Age of the passenger.
- Passenger Class: The class of the passenger (1st, 2nd, or 3rd class).
- Sex: Gender of the passenger.
- Number of Siblings on Board: The number of siblings or spouses aboard.
- Number of Parents or Children on Board: The number of parents or children aboard.
- Passenger Fare: The fare paid by the passenger.
- Survived (Label): Whether the passenger survived (Yes or No).

You will work with the Titanic dataset, to predict whether passengers survived or not based on various attributes.

Name: Abdullah Mohammad Al Talaq

Email: 30786223 [AlTalaqA@sabic.com](mailto:AlTalaqA@sabic.com)

---

## Section 1: Data Exploration

1. Load the Titanic dataset into RapidMiner.

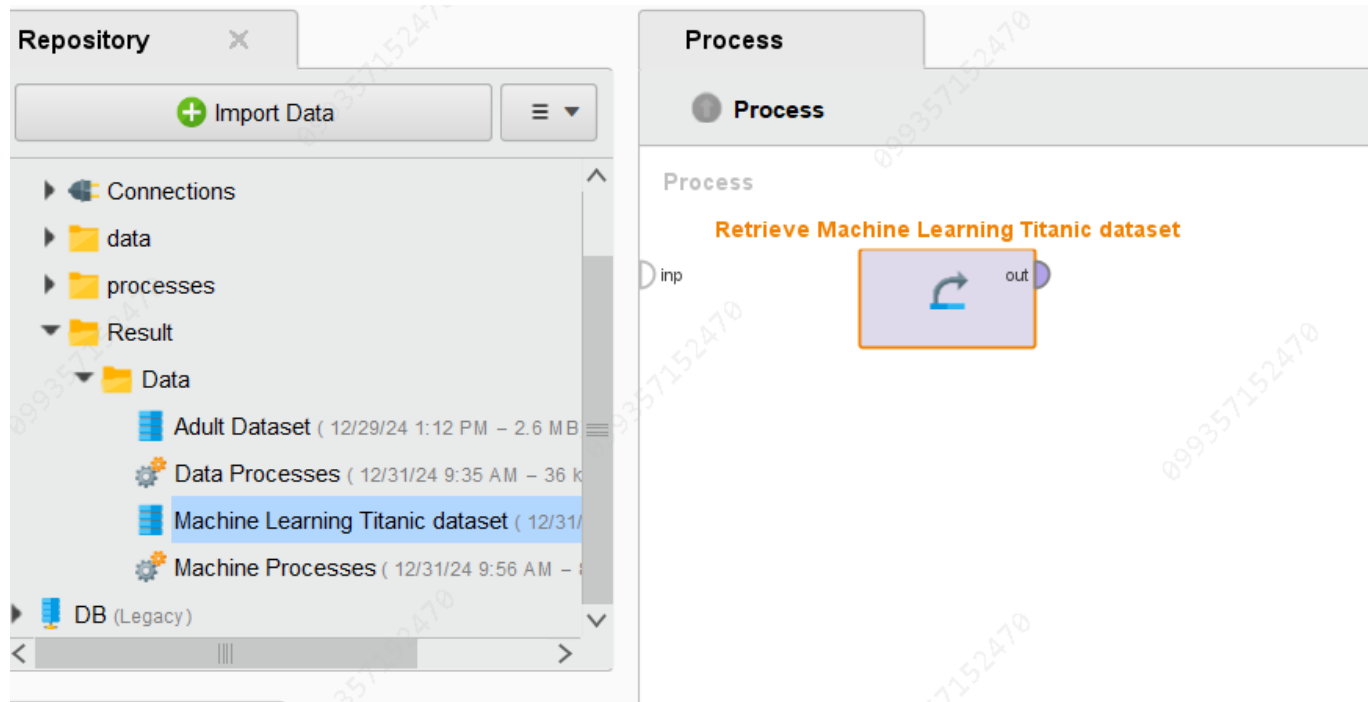



Figure 1: Importing the data.

- 1- Press Import Data button and browse data file.
  - 2- It will be presented as a box called Retrieve ML titanic dataset.
- 
2. Explore the dataset to understand its structure and the distribution of attributes. Write a brief description of what you observe.

< > ⚠️ Age

Summary

 Number

Missing: 0.00%

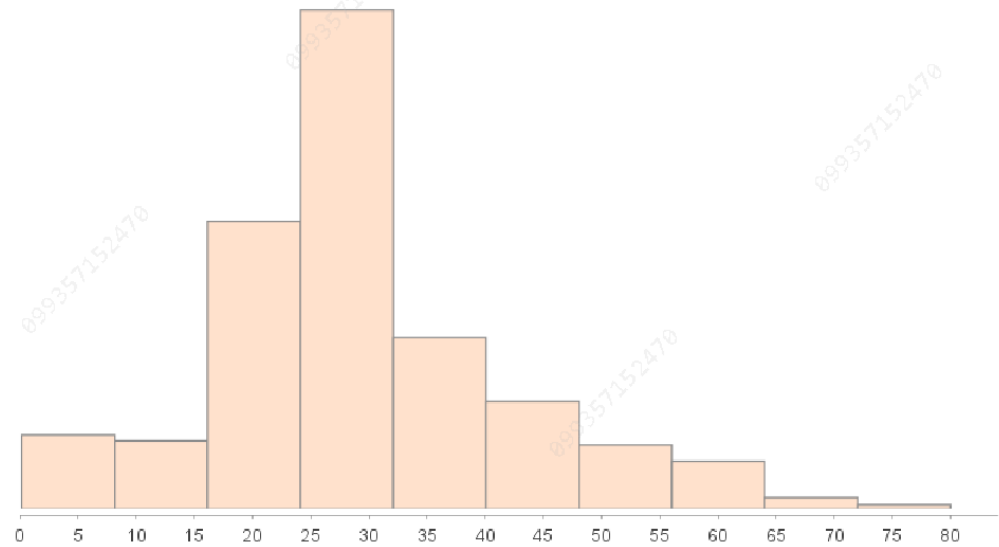
Infinite: 0.00%

ID-ness: 7.86%

Stability: 19.43%

Valid: 72.71%

Distribution



Statistics

Name	Value
Minimum	0.167
Maximum	80
Average	29.932
Standard Deviation	13.265

< > No of Parents or Children on Board

Summary

#

Number

Missing: 0.00%

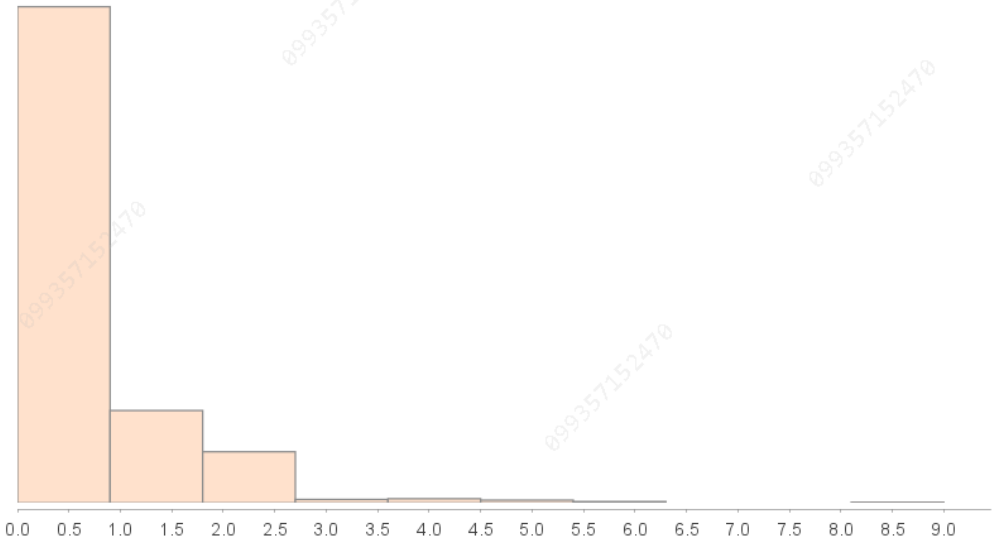
Infinite: 0.00%

ID-ness: 0.87%

Stability: 75.98%

Valid: 23.14%

Distribution



Statistics

Name	Value
Minimum	0
Maximum	9
Average	0.386
Standard Deviation	0.859

< > No of Siblings or Spouses on Board

Summary

#

Number

Missing: 0.00%

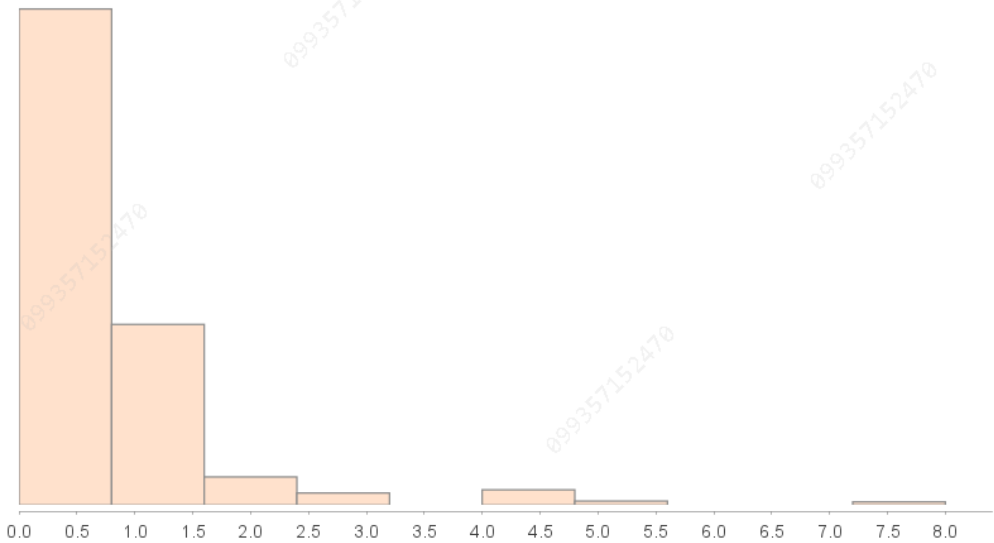
Infinite: 0.00%

ID-ness: 0.76%

Stability: 67.14%

Valid: 32.10%

Distribution



Statistics

Name	Value
Minimum	0
Maximum	8
Average	0.514
Standard Deviation	1.014

## < > Passenger Class

### Summary

Category



Missing: 0.00%

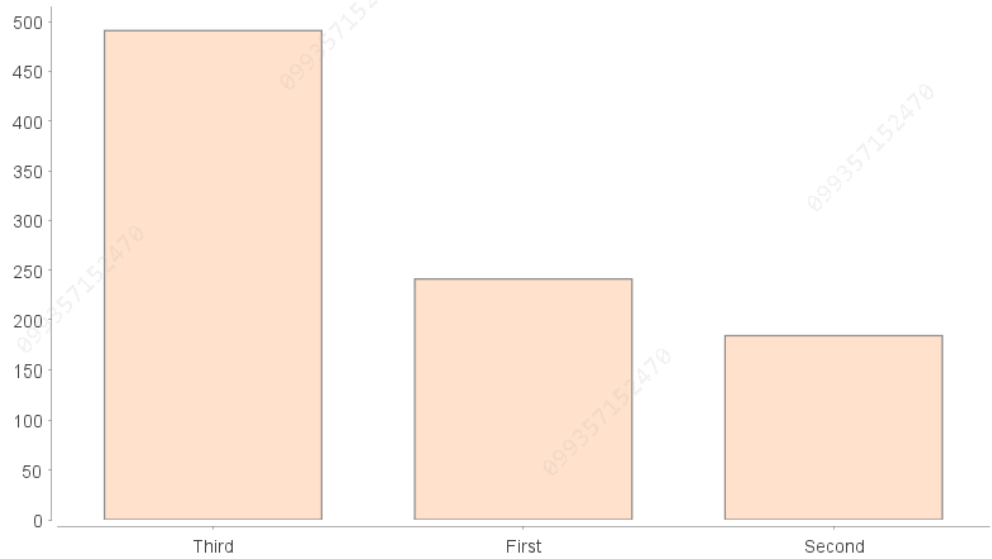
Infinite: 0.00%

ID-ness: 0.33%

Stability: 53.60%

Valid: 46.07%

### Top Values



### 3 Distinct Values:

Value	Count	Percentage
Third	491	53.60%
First	241	26.31%
Second	184	20.09%

<

>

Passenger Fare

Summary

#

Number

Missing: 0.00%

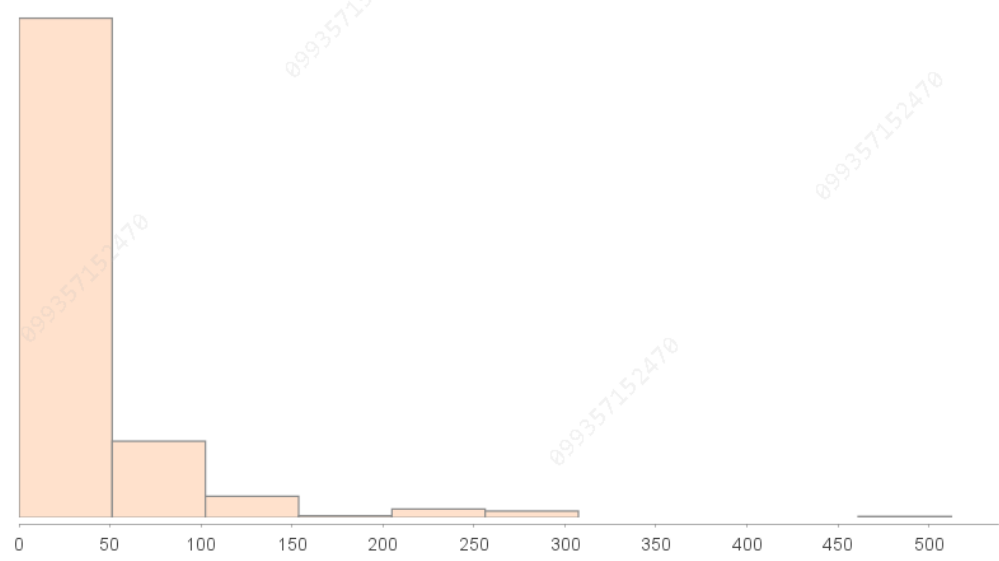
Infinite: 0.00%

ID-ness: 3.38%

Stability: 4.91%

Valid: 91.70%

Distribution



Statistics

Name	Value
Minimum	0
Maximum	512.329
Average	33.564
Standard Deviation	50.239

< > ⚠ Sex

Summary

Category

Missing: 0.00%

Infinite: 0.00%

ID-ness: 0.22%

Stability: 64.85%

Valid: 34.93%

Top Values



2 Distinct Values:

Value	Count	Percentage
Male	594	64.85%
Female	322	35.15%



< > Survived

Summary

Category

Missing: 0.00%

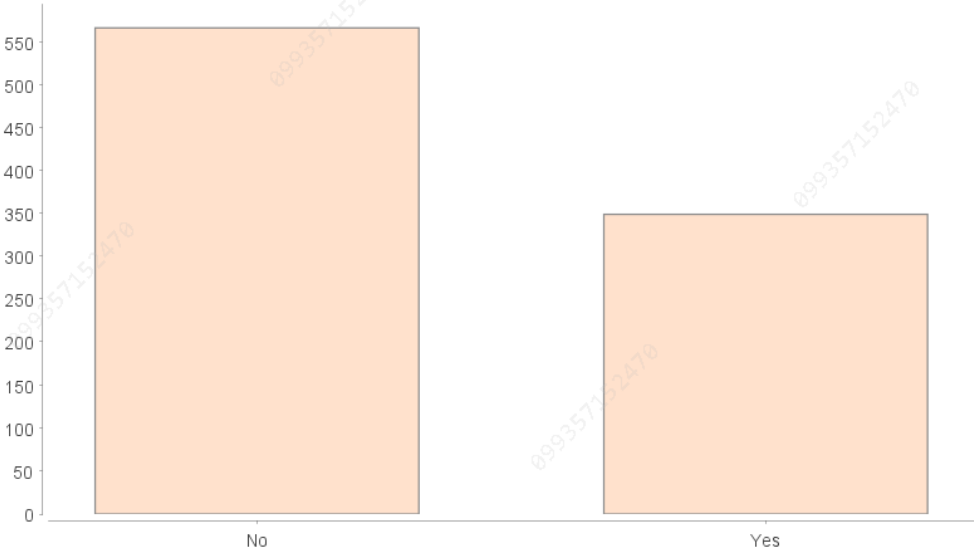
Infinite: 0.00%

ID-ness: 0.22%

Stability: 61.90%

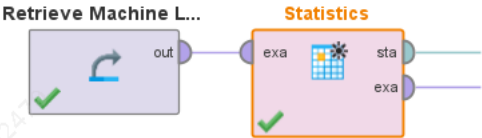
Valid: 37.88%

Top Values



2 Distinct Values:

Value	Count	Percentage
No	567	61.90%
Yes	349	38.10%



Using this operation lets me show all graphs above.

Dataset contains 916 records, and 7 attributes.

Attributes:

- Age: Age of the passenger.
- Passenger Class: The class of the passenger (1st, 2nd, or 3rd class).
- Sex: Gender of the passenger.
- Number of Siblings on Board: The number of siblings or spouses aboard.

- Number of Parents or Children on Board: The number of parents or children aboard.
- Passenger Fare: The fare paid by the passenger.
- Survived (Label): Whether the passenger survived (Yes or No).

3. Calculate basic statistics (mean, median, standard deviation) for the 'Age' attribute.

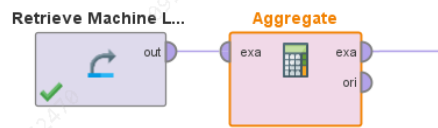


Figure 2: Aggregate operator.

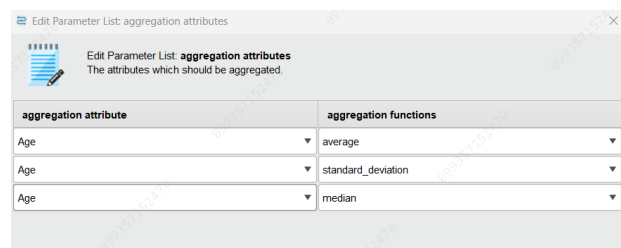


Figure 3 : Selecting aggregation attributes.

Table 1: Aggregation result.

Feature	Average	Stander Deviation	Median
Age	29.932	13.265	29.881

1- Select the aggregate operation will let me use different functions in statistic on the selected attribute.

## Section 2: Data Preprocessing

1. Encode categorical attributes ('Sex') into numerical values.



Figure 4 : Encode operation.

Sex
0
0
1
0
1
0
1
0
0

Figure 5 : Operation result.

- 1- By using nominal to numerical operation
2. Split the data into training and testing sets (80% train, 20% test).

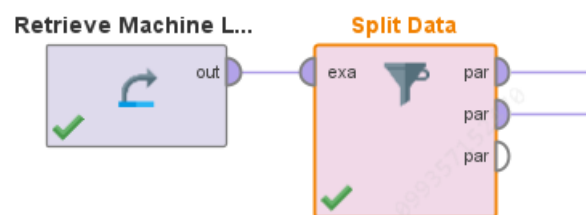


Figure 6 : Split data operation.

ExampleSet (183 examples,0 special attributes,7 regular attributes)

Figure 7 : Split result (Test)

ExampleSet (733 examples,0 special attributes,7 regular attributes)

Figure 8 : Split result (Train)

- 1- By using Split Data.
- 2- Split to two data (80% Train, 20% Test)
3. Verify if the age column's values are valid or if there might have been data entry errors and handle them appropriately.

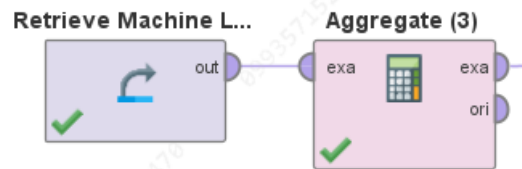


Figure 9 : Aggregate operation.

Row No.	Age	count(Age)
1	0.167	1
2	0.667	1
3	0.750	2
4	0.833	1

Figure 10 : Result of aggregate operation.

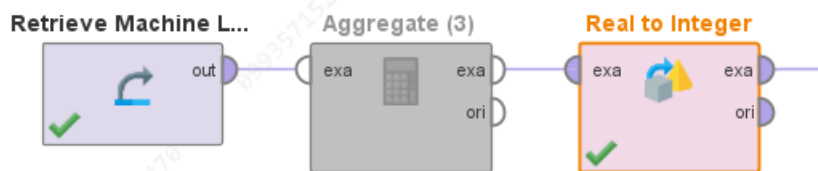


Figure 11 : Real to Integer operation.

Parameters

Real to Integer

attribute filter type
single

attribute
Age

☐ invert selection

☐ include special attributes

☒ round values

Figure 12 : Real to Integer parameters.

- 1- Using the aggregate operator to combine all occurrence for each value in age so we can notice if there is a problem or not.
- 2- Notice the type of the variables in results and found its Real which well contains all possible numbers and in age we want integer values with no comma.
- 3- Use real to integer operator so we can make it int and round 0 values to nearest number.
- 4- Determine an appropriate technique for handling the missing values if there are any.

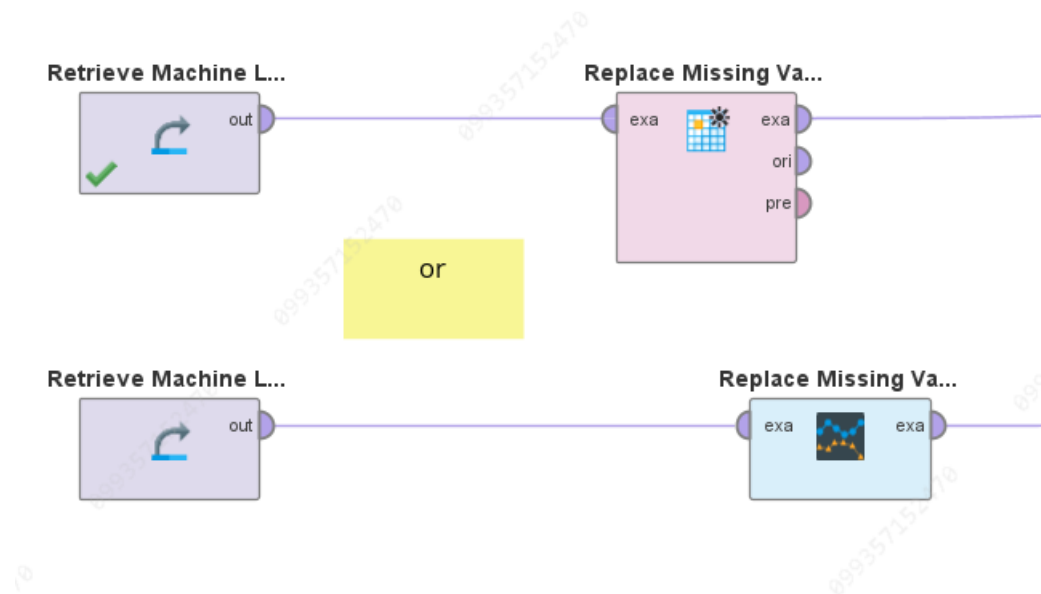


Figure 13 : Fill null values operations.

- 1- Replace missing values by using average values.
- 2- Replace missing values by using time series (previous value of record)

### Section 3: Data Visualization

1. Create a bar chart to show the distribution of passenger classes. Provide a screenshot of the chart.

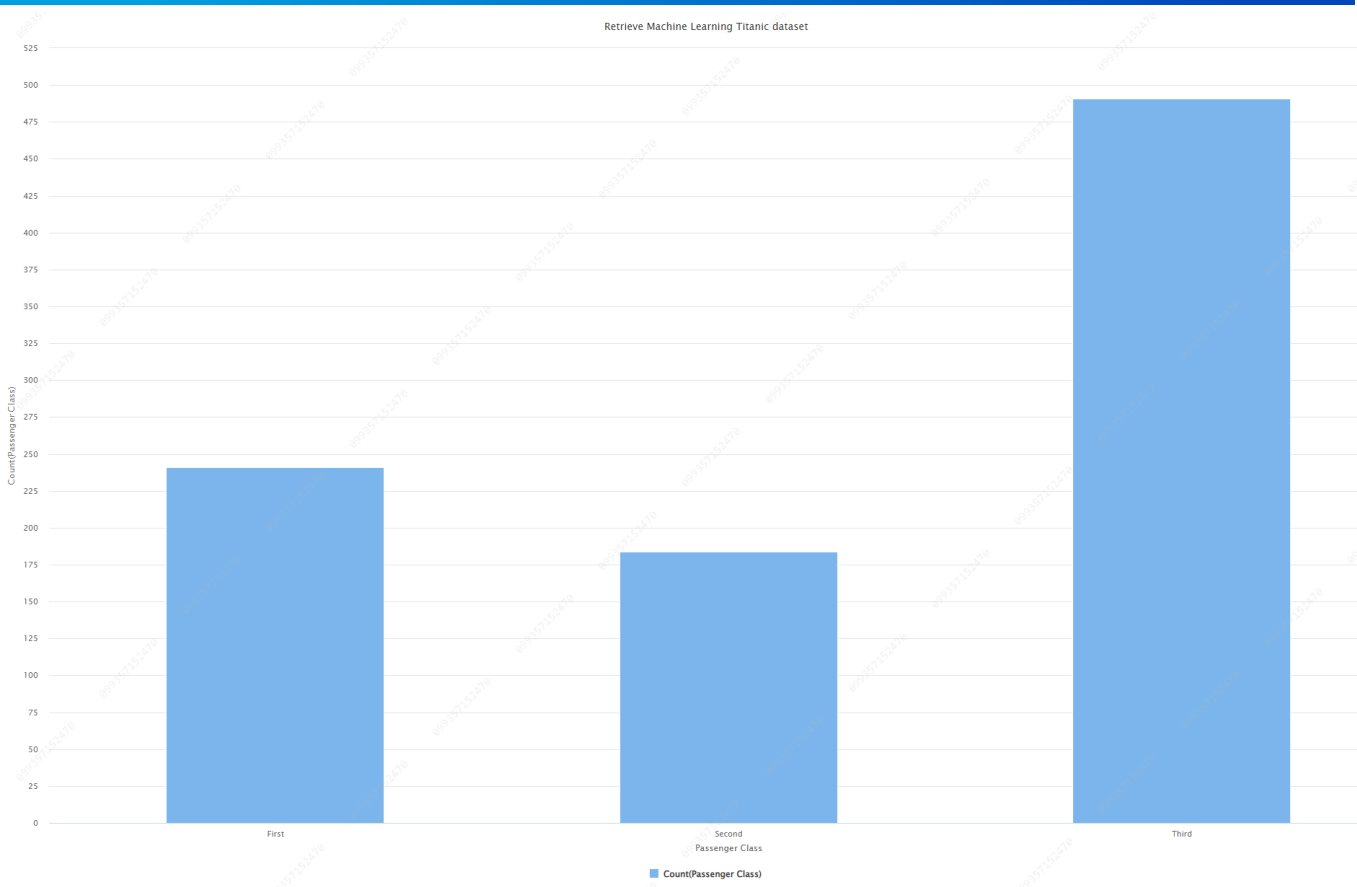


Figure 14 : Bar chart (Passenger Class)

2. Create a histogram to visualize the age distribution of passengers. Provide a screenshot of the histogram.

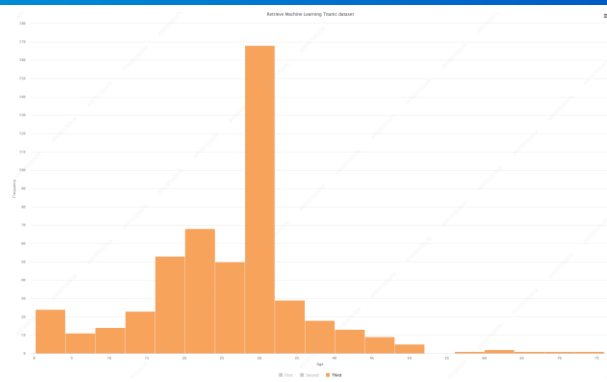


Figure 15 : Age (Third type)

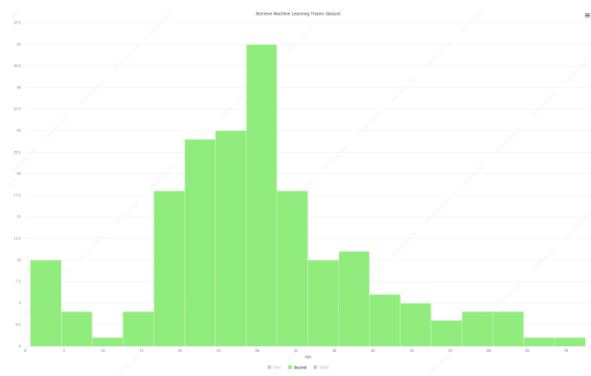


Figure 16 : Age (Second type)

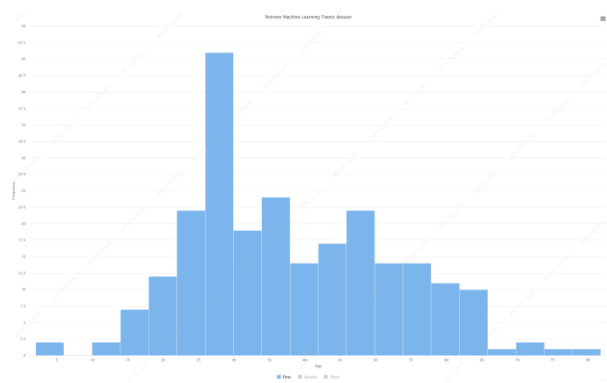


Figure 17 : Age (First type)

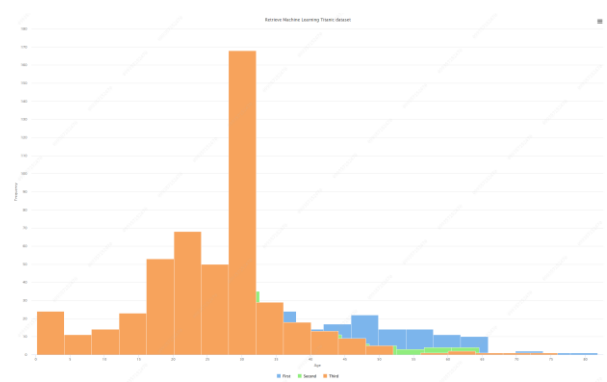


Figure 18 : Age (All)

1. Build and train a classification model using the decision tree algorithm Describe the prediction process.

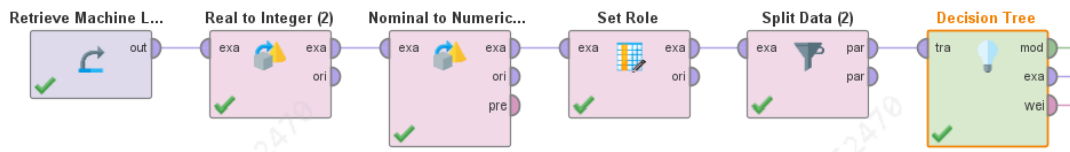


Figure 19 : Decision Tree operation.

attribute	weight ↑
Sex	0.053
Passenger Class	0.106
No of Siblings or Spouses on Board	0.108
No of Parents or Children on Board	0.212
Passenger Fare	0.241
Age	0.281

Figure 20 : Attributes weights.

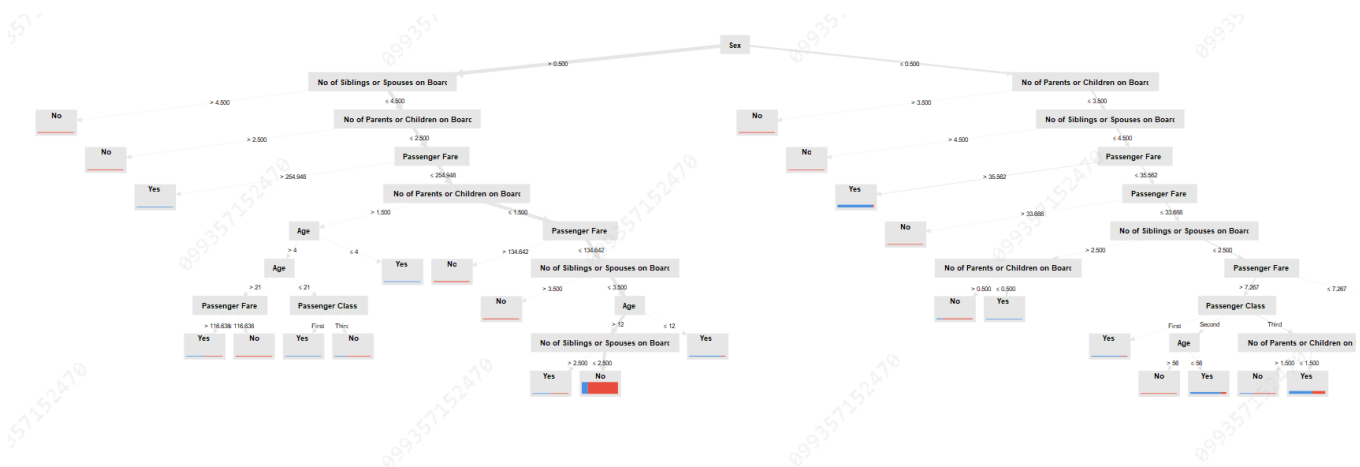


Figure 21 : Model graph.

- 1- Set age to Integer so overcome miss values like 0.785 to round it to 1.
- 2- Set Gender to numerical (change male=1 female=0).
- 3- Set Role (Survived was selected) and split the data into 80% 20%.
- 4- Run the model.

2. Use the model to make predictions on the test dataset.



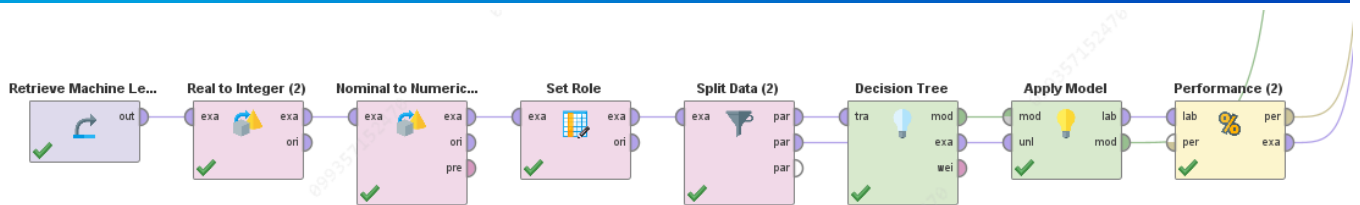


Figure 22 : Model operation.

Survived	prediction(Survived)	confidence(Yes)	confidence(No)
Yes	Yes	0.963	0.037
No	No	0.170	0.830
Yes	Yes	0.963	0.037
Yes	No	0.170	0.830
Yes	No	0.170	0.830
Yes	Yes	0.963	0.037
Yes	Yes	0.963	0.037
No	No	0.170	0.830
No	No	0.170	0.830
Yes	Yes	0.963	0.037
No	No	0.170	0.830
No	No	0.170	0.830
Yes	No	0.170	0.830
Yes	Yes	0.963	0.037
Yes	Yes	0.963	0.037
Yes	Yes	0.963	0.037
Yes	No	0.170	0.830
Yes	Yes	1	0
Yes	Yes	0.963	0.037
Yes	Yes	0.963	0.037
No	Yes	1	0
Yes	No	0	1
Yes	Yes	0.963	0.037
No	No	0.170	0.830
Yes	Yes	0.963	0.037
Yes	No	0.170	0.830
No	No	0.170	0.830

Figure 23 : Sample of the model's prediction.

3. Evaluate the model's accuracy, precision, recall, and F1-score. Provide a screenshot of the evaluation results.

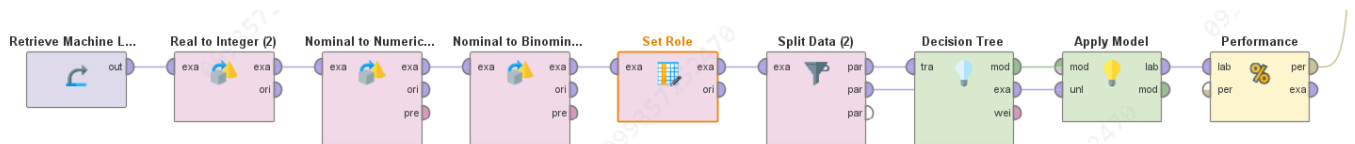


Figure 24 : Model operation.

accuracy: 78.14%

	true Yes	true No	class precision
pred. Yes	49	19	72.06%
pred. No	21	94	81.74%
class recall	70.00%	83.19%	

precision: 81.74% (positive class: No)

	true Yes	true No	class precision
pred. Yes	49	19	72.06%
pred. No	21	94	81.74%
class recall	70.00%	83.19%	

recall: 83.19% (positive class: No)

	true Yes	true No	class precision
pred. Yes	49	19	72.06%
pred. No	21	94	81.74%
class recall	70.00%	83.19%	

f\_measure: 82.46% (positive class: No)

	true Yes	true No	class precision
pred. Yes	49	19	72.06%
pred. No	21	94	81.74%
class recall	70.00%	83.19%	

Figure 25 : Evaluation results.

The steps:

- 1- Used real to int operator for age.
- 2- Used nominal to numerical to change the value of sex from male and female to 1 and 0.
- 3- Set Survived as binominal.
- 4- Set survived as a label role.
- 5- Split the data into 80% 20%.
- 6- Use the model and run the 80% train on it and apply the model to compare the model performance 20% hide set of the data. The it predicts same results the more the accuracy will be higher.

## Section 5: Model Comparison

1. Build and train another algorithm of your choice to compare it with the decision tree model.

Selected model: Logistic Regression

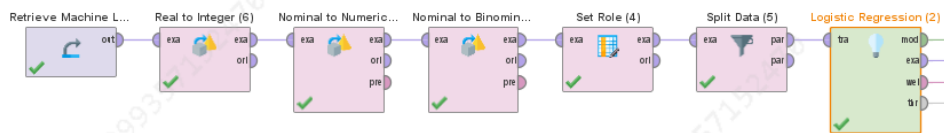


Figure 26 : Model Operation.

## Threshold

Threshold: 0.36457345458409945

first class: Yes

second class: No

if confidence(No) > 0.36457345458409945 then No

else Yes

Figure 27 : Threshold.

attribute	wei... ↓
Passenger Class = Third	1.875
Sex	1.276
Passenger Class = Second	1.235
Age	0.371
No of Siblings or Spouses on Board	0.218
No of Parents or Children on Board	0.134
Passenger Class = First	0
Passenger Fare	-0.041

Figure 28 : Attribute weights.

- Use the trained model to make predictions on the test dataset. Describe the prediction process.

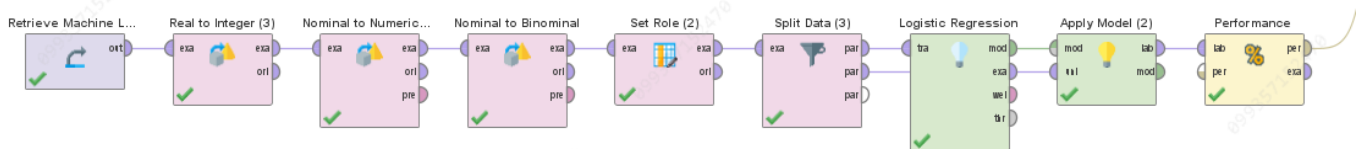


Figure 29 : Model Operation.

Survived	prediction(S...
Yes	Yes
No	No
Yes	Yes
Yes	No
Yes	No
Yes	Yes
Yes	Yes
No	No
No	No

Figure 30 : Sample of the output.

- Compare the accuracy of the trained model with the decision tree model. Write a description of the comparison.

Algorithm	Accuracy
Decision Tree	78.14
LR	80.33%

Figure 31 : Accuracy Table.

1. Evaluate the model's accuracy, precision, recall, and F1-score. Provide a screenshot of the evaluation results.

accuracy: 80.33%

	true Yes	true No	class precision
pred. Yes	45	11	80.36%
pred. No	25	102	80.31%
class recall	64.29%	90.27%	

precision: 80.31% (positive class: No)

	true Yes	true No	class precision
pred. Yes	45	11	80.36%
pred. No	25	102	80.31%
class recall	64.29%	90.27%	

recall: 90.27% (positive class: No)

	true Yes	true No	class precision
pred. Yes	45	11	80.36%
pred. No	25	102	80.31%
class recall	64.29%	90.27%	

f\_measure: 85.00% (positive class: No)

	true Yes	true No	class precision
pred. Yes	45	11	80.36%
pred. No	25	102	80.31%
class recall	64.29%	90.27%	

Figure 32 : Evaluation results.

## Section 5: Conclusion

1. Summarize the results of your analysis. Provide insights into the factors that may influence a passenger's survival on the Titanic. Use your findings and visualizations to support your conclusions.

Attribut...	Passen...	Sex	Survived	Age	No of Si...	No of P...	Passen...
Passeng...	1	0.099	0.281	-0.375	0.069	0.010	-0.577
Sex	0.099	1	0.548	0.103	-0.078	-0.201	-0.175
Survived	0.281	0.548	1	0.056	0.019	-0.066	-0.234
Age	-0.375	0.103	0.056	1	-0.219	-0.121	0.178
No of Sib...	0.069	-0.078	0.019	-0.219	1	0.376	0.164
No of Pa...	0.010	-0.201	-0.066	-0.121	0.376	1	0.238
Passeng...	-0.577	-0.175	-0.234	0.178	0.164	0.238	1

Figure 33 : Correlation matrix.

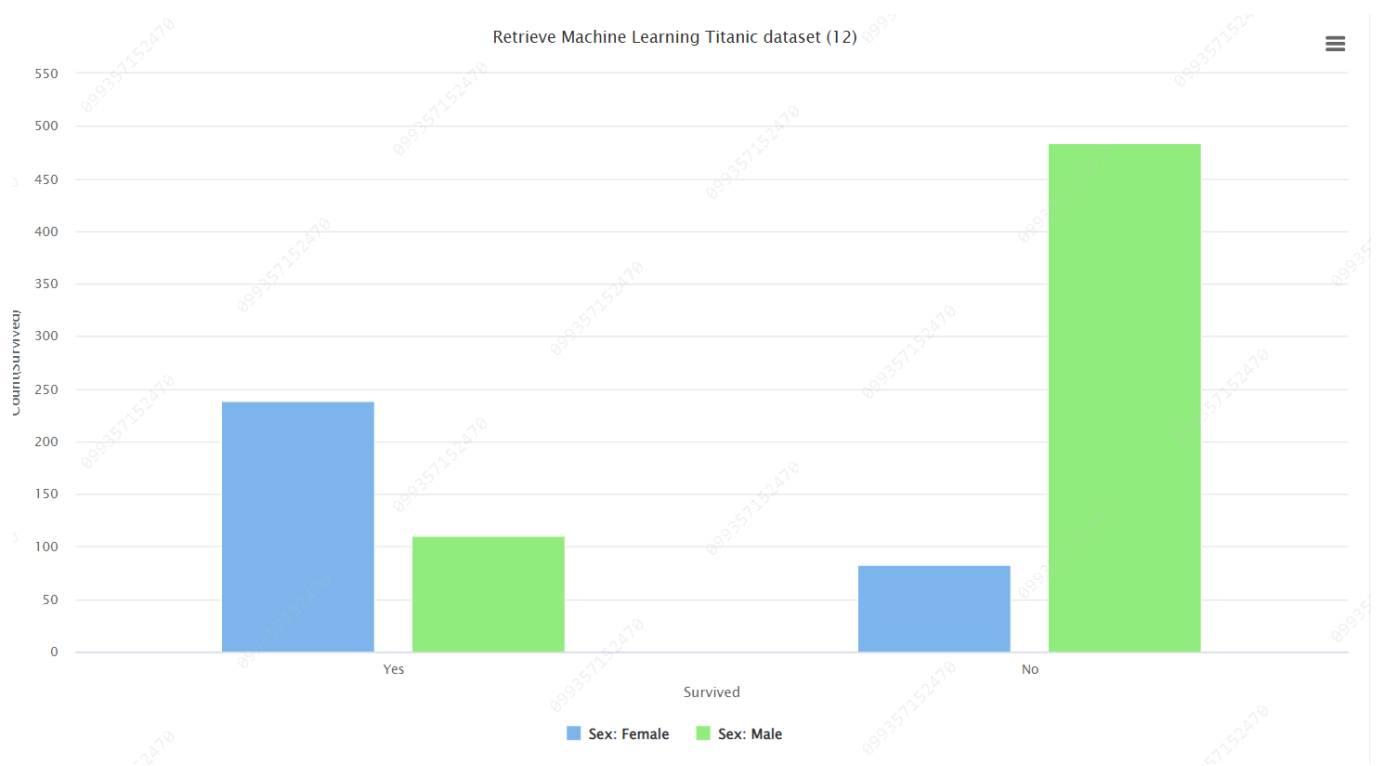


Figure 34 : Bar chart (Gender survived).

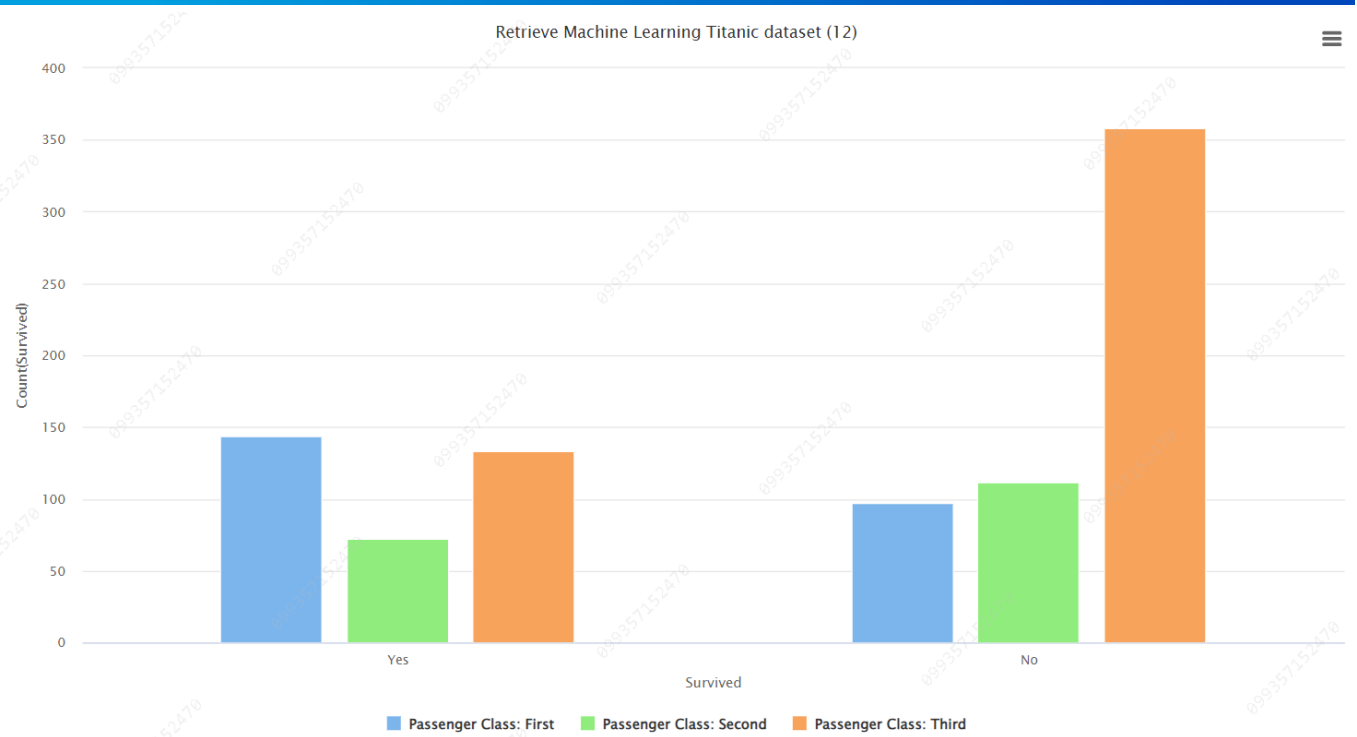


Figure 35 Bar chart (Class survived).

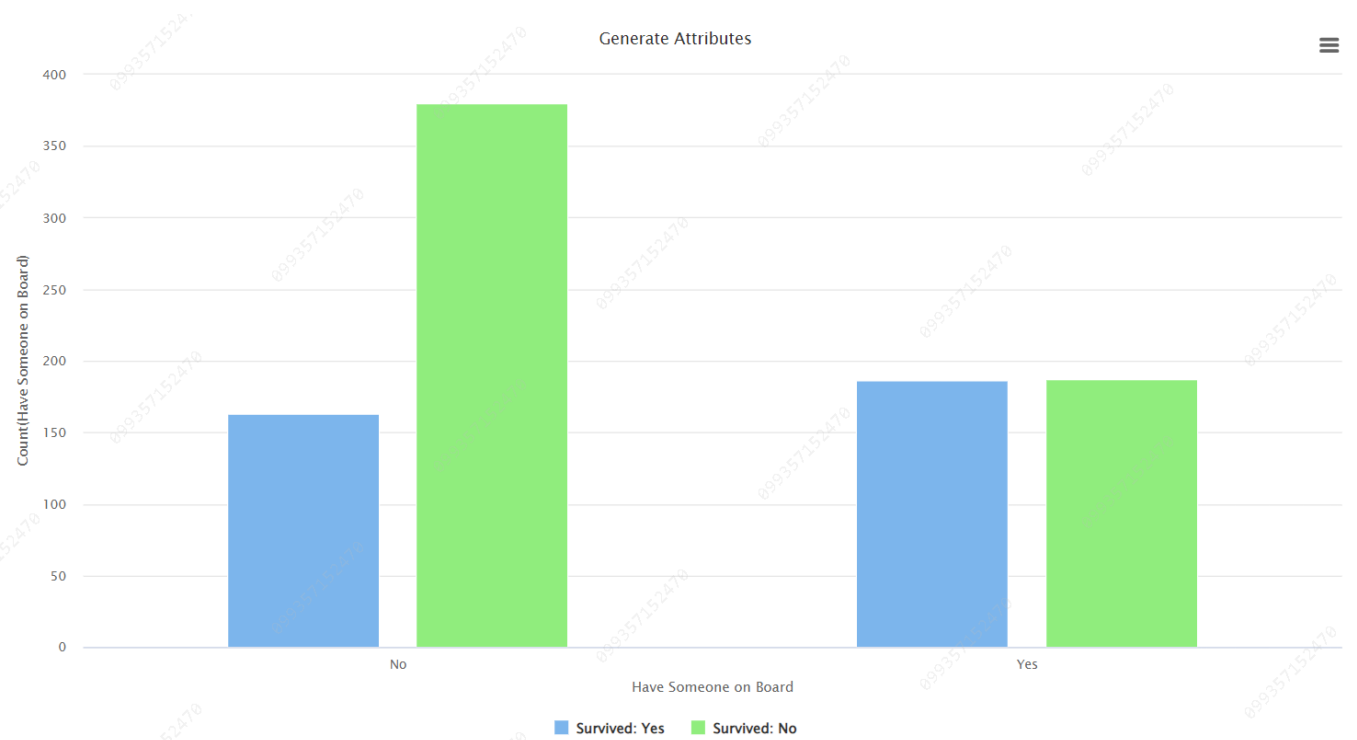


Figure 36 : Bar chart (On board survived).

The graphs show the relation between features. Applied correlation matrix, and the focusing on which has high relation with survived. Found that sex has high impact, and Passenger class. Down below applied multiple bar chart. First bar shows that which gender survived most and see that most of survived were female. Second bar shows the higher the passenger classes the possibility of surviving would be higher. Last bar I Made new attributes collect all records who has someone on board and

presented as yes while no one presented as no. those who survived and has someone on board are higher than those who has no one. While the passengers who has no one on board is highest bar on all of them. We can consider these things:

- 1- Females enter the boats firstly so that's why they survived more.
- 2- Those who are in first class survived more maybe because they are Politicians, celebrities, rich people so they enter the boats firstly.
- 3- Those who has someone on board survived more. Can be mothers and their Childrens mostly.
- 4- Last class, male and came alone. Your chance to survive would be weak. Probably you will be arranging the movements of passengers to the boats that's why they survived less then others