Assessment: Assignment 2

Student Name: Raiyan Yasawy - 041003092
Vy Tran - 041074757
Lab Professor Name: Anu Thomas

Mar 01, 2023

TITANIC SURVIVAL PREDICTION

I. DATA UNDERSTANDING

The dataset is about the passengers on the Titanic and their survival status during the ship's maiden voyage that sunk on April 15, 1912. The purpose of analyzing the dataset is to identify factors that may have influenced the survival of the passengers and to build predictive models that can be used to estimate the survival of passengers in future similar scenarios. The dataset was obtained from Kaggle, a platform for predictive modeling and analytics, and analyzed using Weka, a collection of machine learning and data mining tools.

The dataset includes information about the passengers' demographics, their travel class, cabin information, family relationships, fare paid, and other related features that may be useful in predicting survival. The methodology used for the analysis includes decision tree-based and cluster analysis techniques after data review and normalization.

Table 1. Attributes and Data types

Attributes	Description	Data type	Comment
passengerID	ID of passenger	Numeric	Not needed
Pclass	Passenger class	Nominal	
Name	Passenger name	String	Not needed
Sex	Sex	Nominal	
Age	Passenger age	Numeric	
SibSp	Number of siblings/spouses on board	Nominal	
Parch	Number of parents/children on board	Nominal	
Ticket	Ticket number	String	Not needed
Fare	Passenger fare	Numeric	
Cabin	Cabin number	String	
Embarked	Port of Embarkation	Nominal	

II. DATA PREPARATION

1.1. Data consolidation

The train and test Titanic dataset contain data for 889 and of the real Titanic passengers. Each row represents one person. The columns describe different attributes about the person including whether they survived, their age, their passenger-class, their sex and the fare they paid.

1.2. Data cleaning

From the dataset, we can see all the attributes would correlate to the survival rate of the passenger

except for PassengerID, Name and Ticket number. So, we decided to remove those attributes.

Before applying any type of data analytics on the dataset, the data must be first cleaned. There are some missing values in attributes such as Age, and Cabin that need to be handled. Since 77 % values of the Cabin attribute are missing, the attribute would be dropped from the dataset.

Table 2. Dataset after removing unnecessary attributes

Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
0	3	male	22	1	0	7.25	S
1	1	female	38	1	0	71.2833	C
1	3	female	26	0	0	7.925	S
1	1	female	35	1	0	53.1	S
0	3	male	35	0	0	8.05	S

Create new attributes:

The age attribute of the train dataset has 177 missing values, which is 20% of the dataset. To solve this issue, we put passengers into groups: NK (age not given), Child (age < 12), Youth (age < 25), Adult (age ≤ 60), and Senior (age >60) based on their Age.

Then, we create the **Relatives** attribute and calculate the total number of siblings, spouse, parents, and children. The value would be "Many" if number of relatives is greater than or equal to 3, "Few" if greater than 0, "None" otherwise.

Note that to make the classification more accurate, we need to remove the Age, SibSp and Parch attribute.

Table 3. Dataset after creating Age_group, Relatives and removing Age, SibSp, Parch attributes

Survived	Pclass	Sex	Fare	Embarked	Age_group	Relatives
0	3	male	7.25	S	Youth	Few
1	1	female	71.2833	C	Adult	Few
1	3	female	7.925	S	Adult	None
1	1	female	53.1	S	Adult	Few
0	3	male	8.05	S	Adult	None
0	3	male	8.4583	Q	NK	None
0	1	male	51.8625	S	Adult	None
0	3	male	21.075	S	Child	Many
1	3	female	11.1333	S	Adult	Few
1	2	female	30.0708	C	Youth	Few

1.3. Data processing

Group the Fare data to bins

There are several methods to calculate number of bins. We decided to use Sturge's Rule formula in

this research as it is considered a “Rule of Thumb”:

$$K = 1 + 3.322 log N = 1 + 3.322 \log(899) = 11$$

where

K = number of bins

N = number of instances

Equal Width: The Fair data has range from 0 to nearly 513, divided into 11 bins. Therefore, each bin has a fixed width of approximately 47.

Bins = [1 {0-46}, 2 {47-93}, 3 {94-140}, 4 {141-187}, 5 {188-234}, 6 {235-281}, 7 {282-328}, 8 {329-375}, 9 {376-422}, 10 {423-469}, 11 {470-516}]

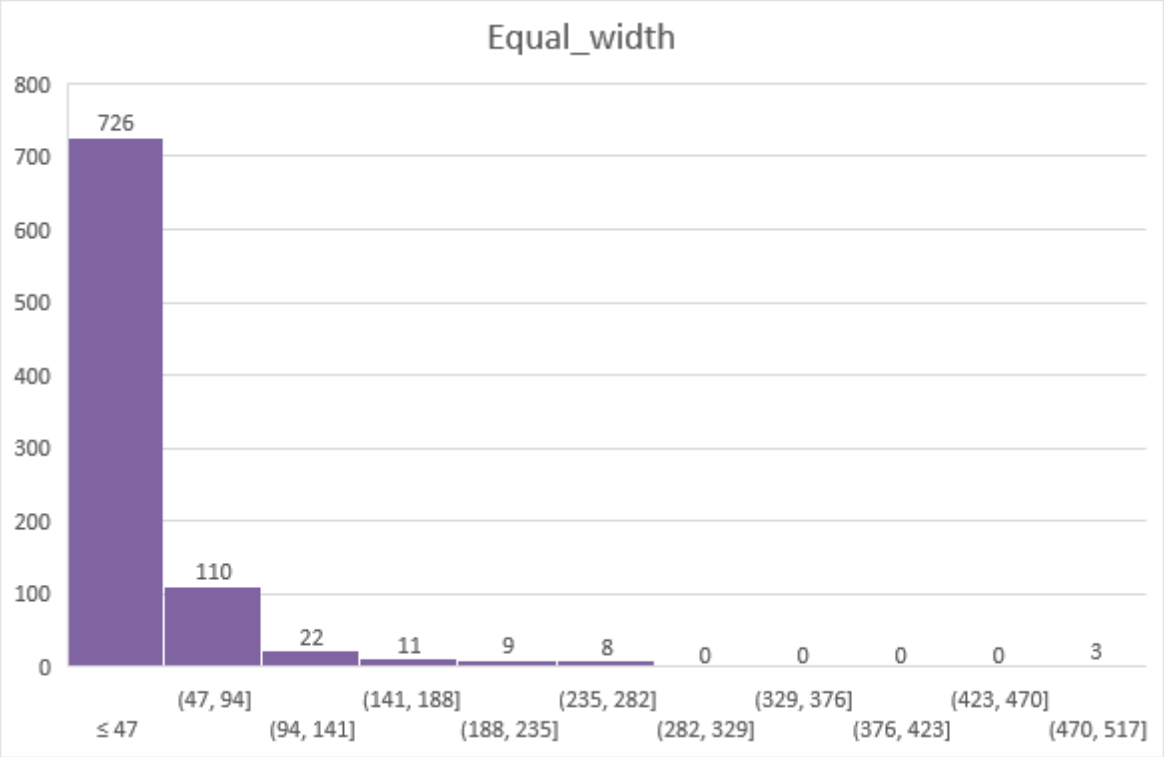

Figure 1. Equal_width Histogram

By using Equal Width approach, we can see there are 4 empty bins.

Although Equal width is better for graphical representations (histograms) and is more intuitive, it might have problems if the data is not evenly distributed, or has outliers, as we will have many empty, useless bins.

Equal Frequency: Equal frequency guarantees that every bin contains the roughly the same amount of data, which is usually preferable if you have to then use the data in any kind of model/algorithm as bins will be more significative in representing the underlying distribution.

We would use the Equal-frequency method for this analysis because this dataset is not evenly distributed. It leads to the result of 4 empty bins, which might affect the survival prediction.

The Fare data is put into bins using formula:

$$= ROUNDDOWN(PERCENTRANK(\$D\$2:\$D\$890, \$D2) * 11, 0) + 1$$

where

D2:D890: The range that contain the dataset to evaluate (Fare values)

D2: The value to calculate the percentage rank

First, get the percentage rank of the current cell ($D2) out of all the cells being binned ($D2:$D$2001). This will be a value between 0 and 1, so to convert it into bins, just multiply by the total number of bins (11). Then, use *ROUNDDOWN* to chop off the decimals. Finally, add 1 to the end of the formula as we want the bins to start at 1 rather than 0.

Note that the bins are not exactly equal in size due to the repeated continuous fare values.

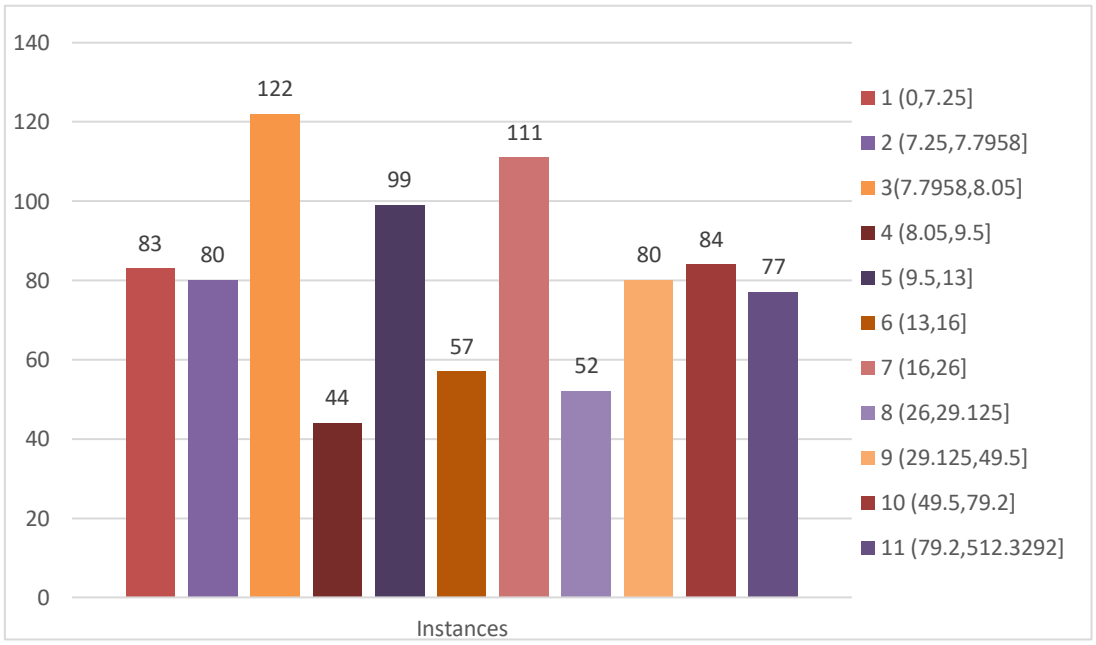

Figure 3. After binning

F8 None

	A	B	C	D	E	F	G	H	I	J
1	Survived	Pclass	Sex	Embarked	Age_group	Relatives	Fare_bin (11)			
2	0	3	male	S	Adult	None	1			
3	0	1	male	S	Adult	None	1			
4	1	3	male	S	Adult	None	1			
5	0	2	male	S	NK	None	1			
6	0	3	male	S	Youth	None	1			
7	0	2	male	S	NK	None	1			
8	0	2	male	S	NK	None	1			
9	0	2	male	S	NK	None	1			
10	0	3	male	S	Adult	None	1			
11	0	1	male	S	NK	None	1			
12	0	2	male	S	NK	None	1			

Figure 2. Titanic_train_processed file header

1.4. Data analysis

The Survived, Pclass, and Fare_bin data types are numeric by default, so we applied NumericToNominal filter to convert them to the right types.

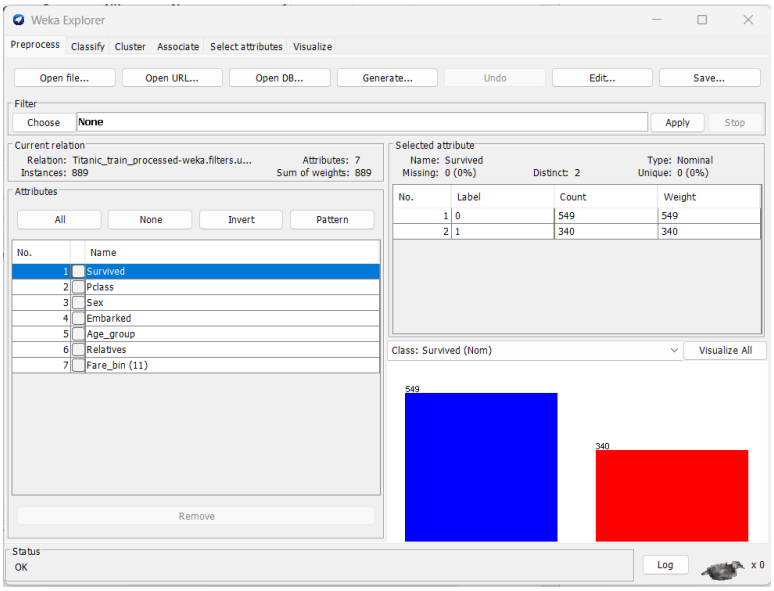

Figure 4. Distribution of the Class attribute

Age vs Survival

Figure 5 and 6 illustrates the impact of age on the rate of survival. The figures indicate that a lower age value is associated with a higher survival rate (Baby, Child), while a higher age value is associated with a lower survival rate (Senior)

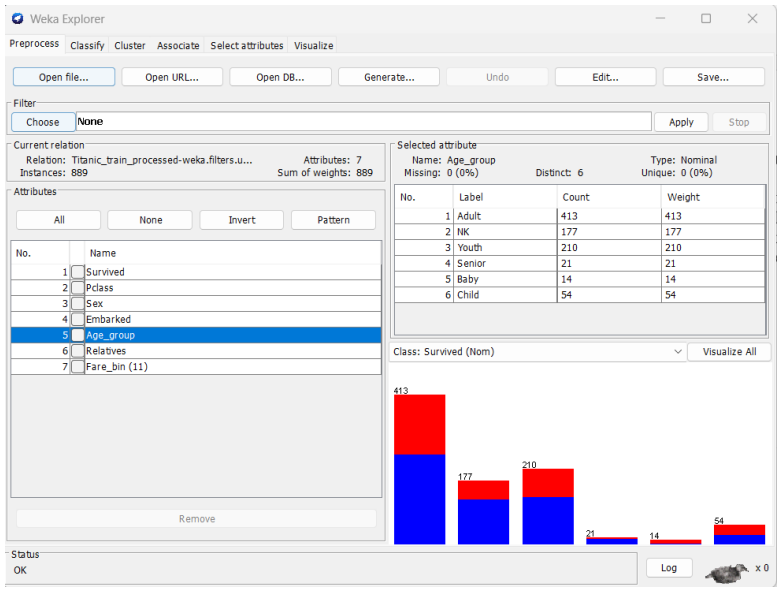

Figure 5. Distribution of the Age_group attribute

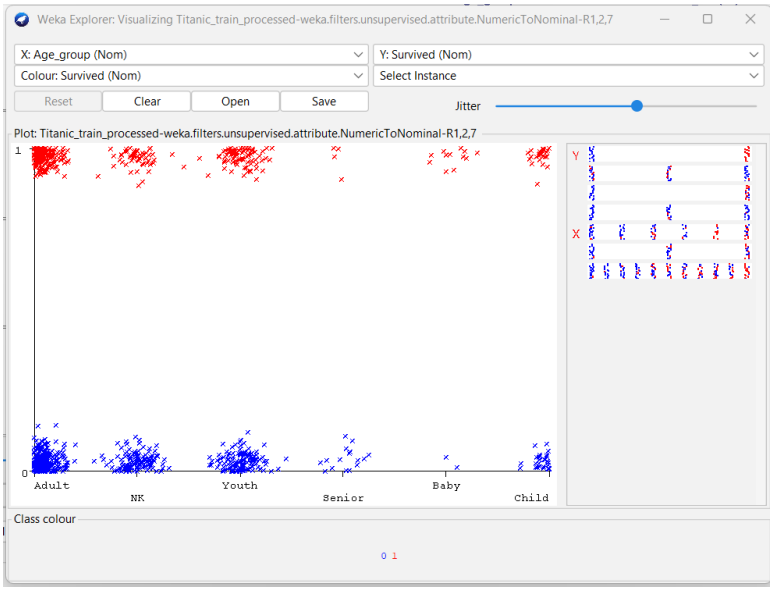

Figure 6. Visualization

Table 4. Age_group vs survival rate

Age_group	Survival rate
Adult	40.19%
NK	29.38%
Baby	85.71%
Youth	37.62%
Senior	19.05%
Child	50.00%

Sex vs Survival

It is visually evidence that there is a strong correlation between the sex of the passengers and their survival chances, which is also reflected in the J48 tree which will be discuss later in the research. Figures 7 and 10 reveal that females have a higher probability of surviving compared to males, with survival rates of 74.4% and 18.9% respectively.

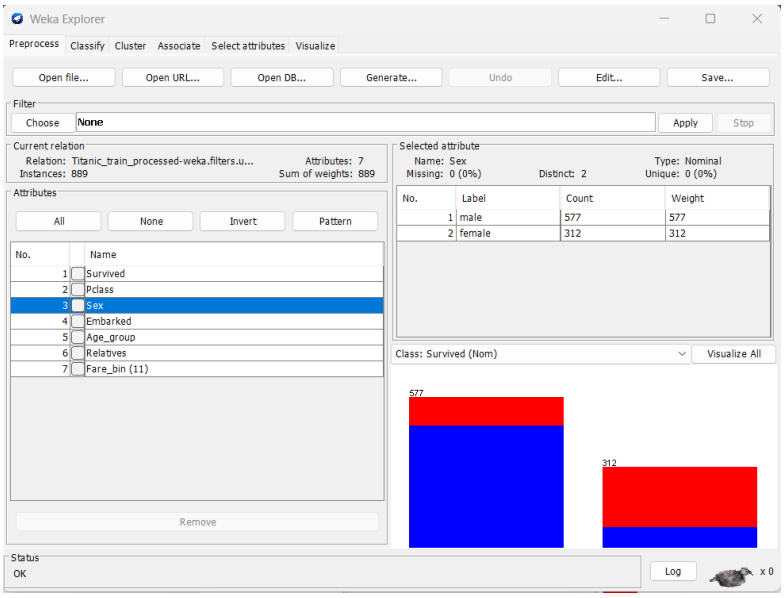

Figure 7. Distribution of the Sex attribute

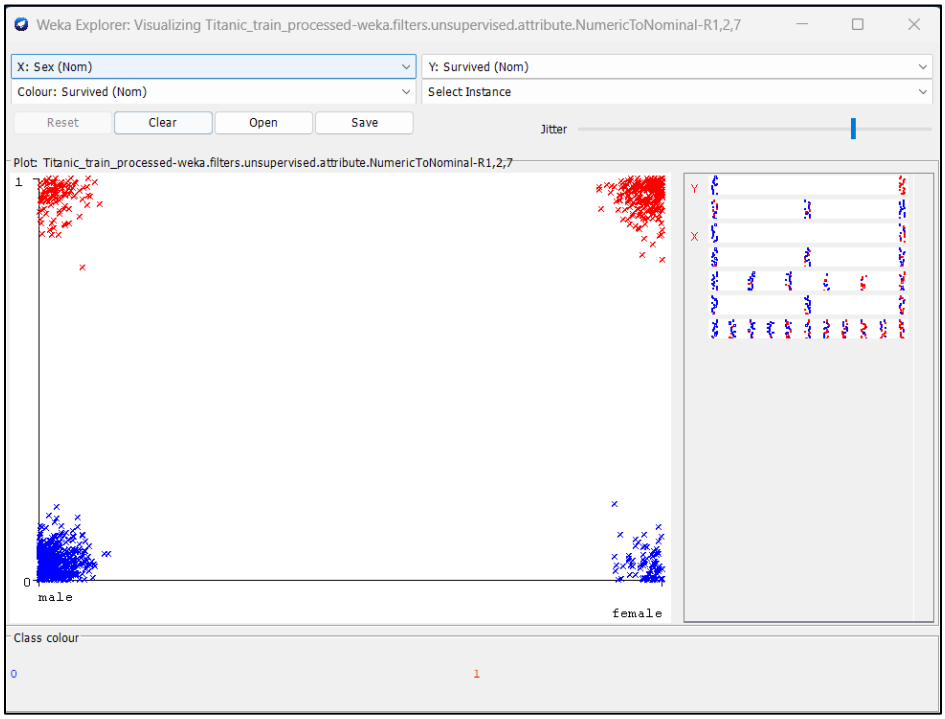

Figure 8. Survived vs Sex Classification

Pclass vs survival

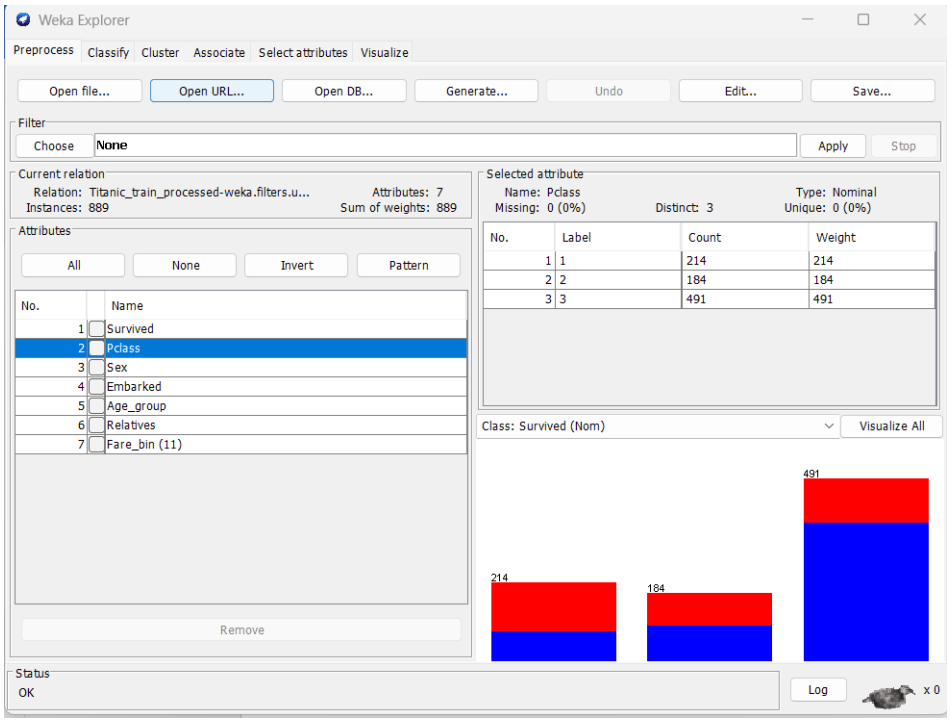

Figure 9. Pclass distribution

We can see that Passengers who were travelling in the third class is less likely to survive.

Embarked vs Survival rate

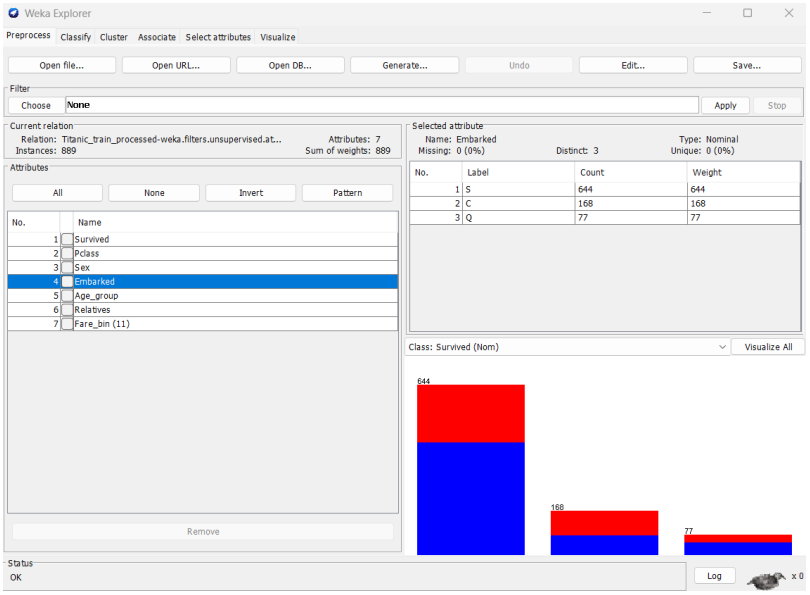

Figure 10. Embarked distribution

These appear to be a correlation between survival rate and where passengers embarked. Those who embarked at port C are more likely to survive than those embarked at port Q or S.

Table 5. Embarked vs Survival rate

Embarked	Survival rate
S	33.70%
C	55.36%
Q	38.96%

Fare vs Survival rate

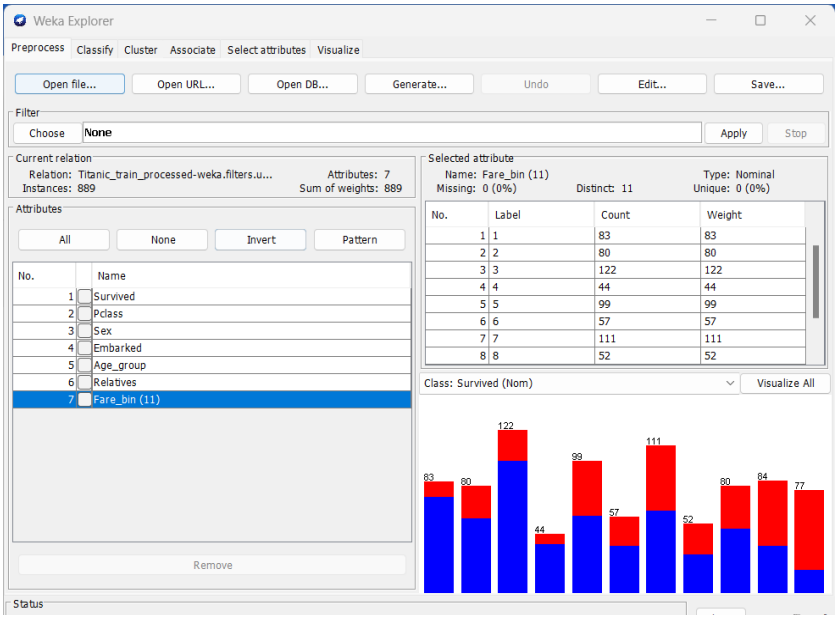

Figure 11. Fare distribution

Table 5. Fare_bin vs survival rate

Fare_bin	Survival rate
1	13.25%
2	30.00%
3	18.85%
4	63.64%
5	41.41%
6	38.60%
7	44.14%
8	44.23%
9	40.00%
10	58.33%
11	76.62%

Based on the table above, we can conclude that the survival rate increases with a higher fare.

```
@relation 'Titanic_train_processed-weka.filters.unsupervised.attribute.NumericToNominal-R1,2,7'

@attribute Survived {0,1}
@attribute Pclass {1,2,3}
@attribute Sex {male,female}
@attribute Embarked {S,C,Q}
@attribute Age_group {Adult,NK,Youth,Senior,Baby,Child}
@attribute Relatives {None,Few,Many}
@attribute 'Fare_bin (11)' {1,2,3,4,5,6,7,8,9,10,11}

@data
0,3,male,S,Adult,None,1
0,1,male,S,Adult,None,1
1,3,male,S,Adult,None,1
0,2,male,S,NK,None,1
0,3,male,S,Youth,None,1
0,2,male,S,NK,None,1
0,2,male,S,NK,None,1
0,2,male,S,NK,None,1
0,3,male,S,Adult,None,1
0,1,male,S,NK,None,1
0,2,male,S,NK,None,1
0,2,male,S,NK,None,1
0,1,male,S,Adult,None,1
```

Figure 12. Normalized Trained Dataset in ARFF format

DECISION TREE CLASSIFICATION

Decision tree is a supervised learning algorithm. It is capable of handling both categorical and continuous input and output variables. Using Weka, we generated a J48 Tree (C.45 implementation) which resulted in the classifier output. The J48 Tree diagram shown in figure 8 below illustrates the classification path that the data suggests.

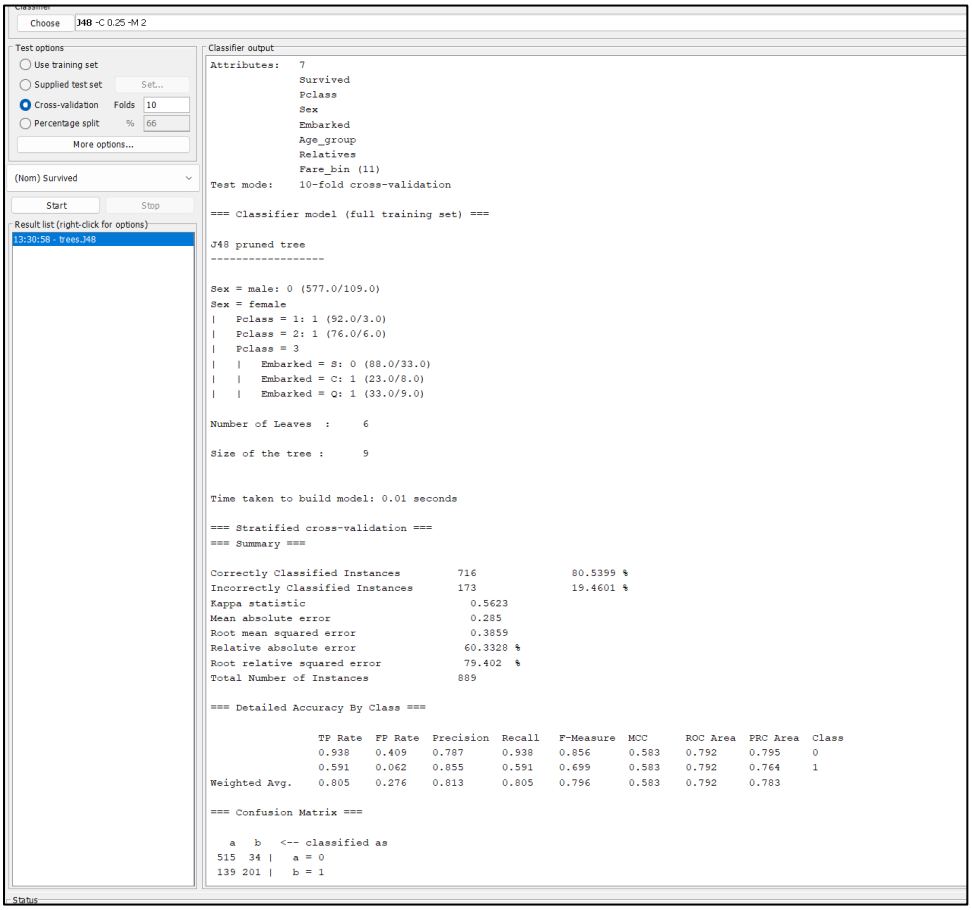

Figure 13. J48 Classify view

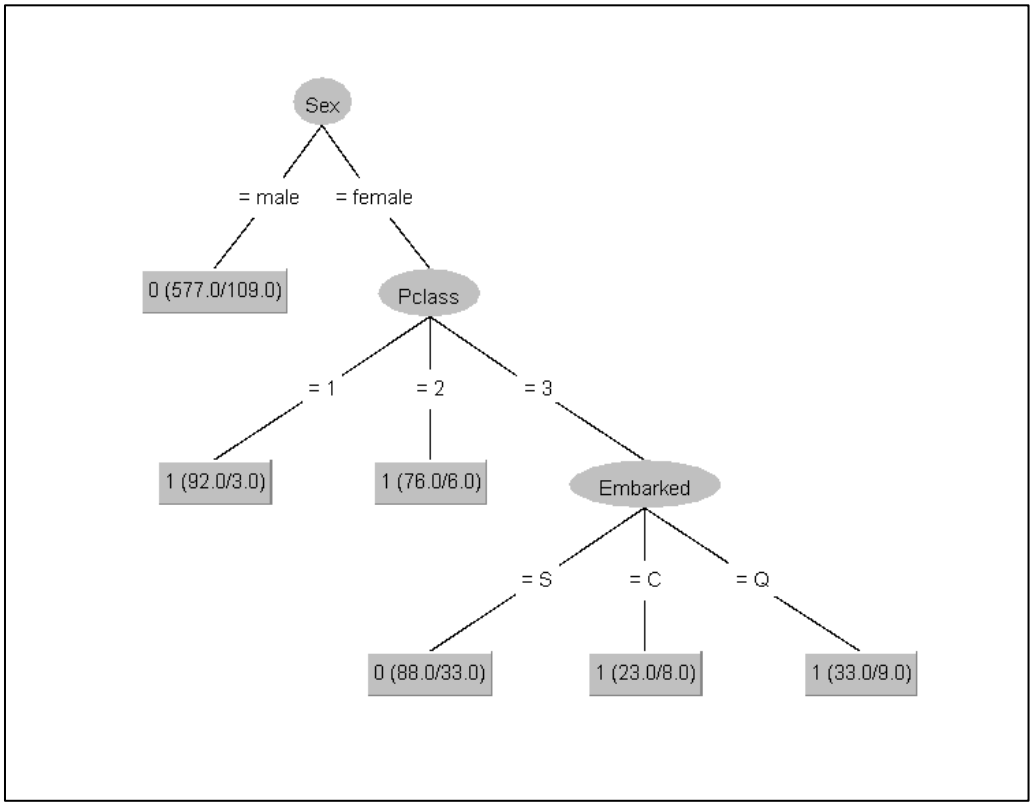

Figure 14. J48 Classifier with 10-fold

Figure 15. Confusion matrix

Survived	TP	FP	Number of misclassifications
0	0.938	0.409	34
1	0.591	0.062	139

Accuracy: 80.5%

Based upon the outcome of the J48 analysis it was clear that the most significant association with regards to survival was related to Sex; in that just being Female was the most significant classifier. Relatives and Fares do not play a significant role in J48 classifiers prediction process.

III. MODELING & EVALUATION

Prepare the test set: In order to group Fare values of the test into correct bins, we calculate the maximum fare value of each bin in the trained dataset. Then apply IF formula for the each of the fare value of the test dataset: "If fare <=7.25, then bin 1

Else if fare <=7.79, then bin 2
Else if fare <=8.05, then bin 3
Else if fare <=9.5, then bin 4
Else if fare <=13, then bin 5
Else if fare <=16, then bin 6
Else if fare <=26, then bin 7
Else if fare <=29.125, then bin 8
Else if fare <=49.5, then bin 9
Else if fare <=79.2, then bin 10
Else bin 11"

Bins	Max_fare_value
1	7.25
2	7.7958
3	8.05
4	9.5
5	13
6	16
7	26
8	29.125
9	49.5
10	79.2
11	512.3292

```
@relation Titanic_test-weka.filters.unsupervised.attribute.NumericToNominal-Rfirst-last

@attribute Survived {0,1}
@attribute Pclass {1,2,3}
@attribute Sex {male,female}
@attribute Embarked {S,C,Q}
@attribute Age_group {NK,Adult,Child,Youth,Baby,Senior}
@attribute Relatives {None,Few,Many}
@attribute 'Fare_bin (11)' {1,2,3,4,5,6,7,8,9,10,11}

@data
?,1,male,S,NK,None,1
?,1,male,S,Adult,None,1
?,3,male,S,Child,Few,1
?,3,male,C,NK,None,1
?,3,male,C,NK,Few,1
?,3,male,S,Youth,Few,1
?,3,female,Q,Adult,None,1
?,3,female,S,Adult,Few,1
?,3,male,S,NK,None,1
?,3,male,S,NK,None,1
?,3,male,S,Youth,None,1
?,3,male,C,Youth,None,1
?,3,female,C,Adult,None,1
?,3,male,C,Youth,None,1
?,3,male,C,Adult,None,1
?,3,male,C,Adult,None,1
?,3,male,C,Youth,None,1
?,3,male,C,NK,None,1
```

Figure 16. Normalized Test Dataset in ARFF format

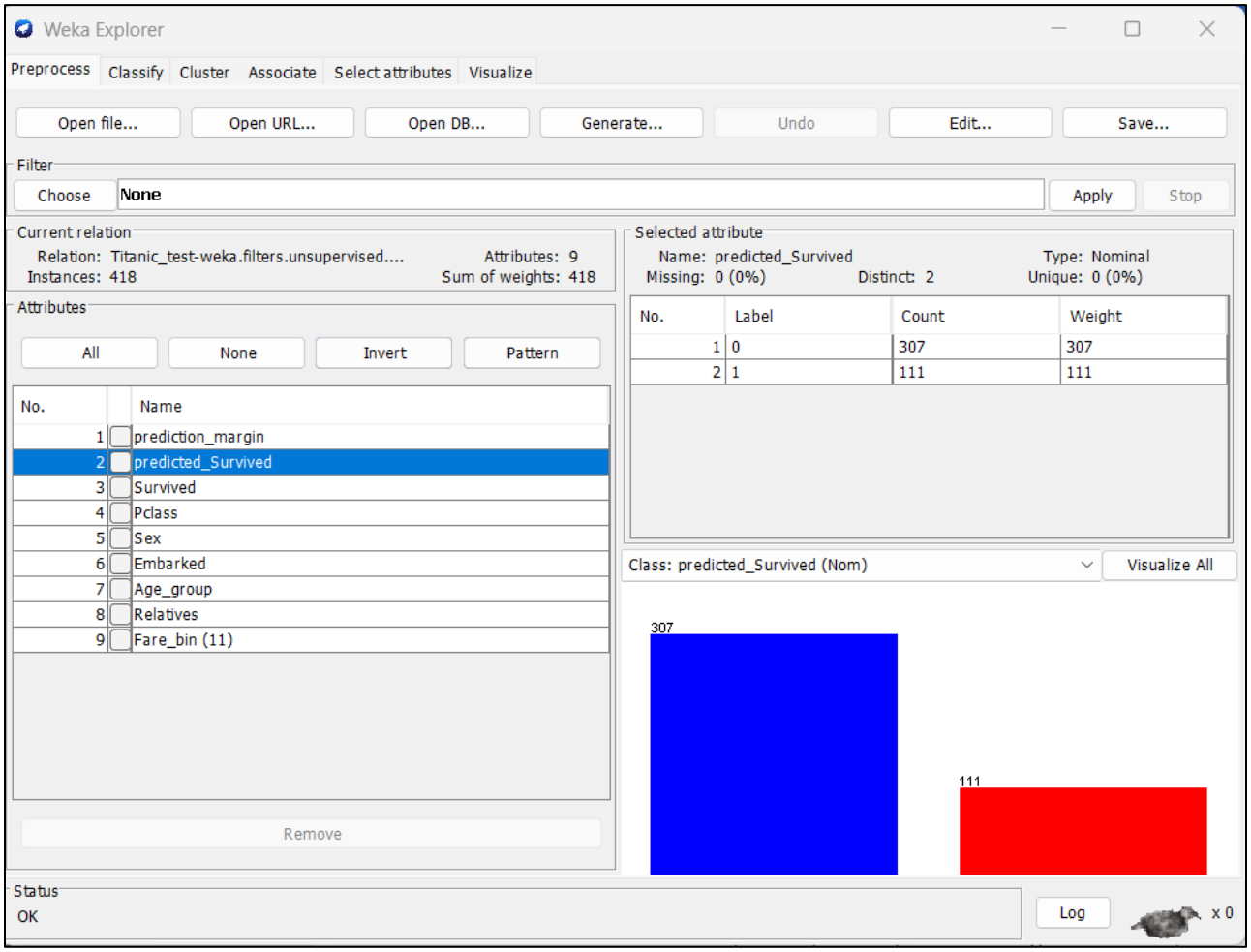

Figure 17. Test dataset predicted survived

Total instances in the test file	418
Number of persons predicted to survive	111
Number of persons predicted to not survive	307
Percentage of predicted survival	26.6%

IV. DISCUSSION OF RESULTS

Our model predicted 26.6% survivors, which is 5% less than 31.6% actual survivors of the incident.
Comparation:
Pclass vs Survival rate

Table 6. Predicted vs Actual

Pclass	Predicted_survival_rate	Actual
1	46.73%	61%
2	32.26%	42%
3	14.22%	24%

The reason for the difference:

- Overfitting: Decision trees can easily overfit the training data. This can lead to poor generalization and accuracy on new, unseen data.
- Instability: Small changes in the training data can result in a completely different decision tree, which makes decision trees less stable than other algorithms.
- Bias: Decision trees can be biased towards features with more levels or categories, which can result in a skewed decision tree.
- Difficulty in capturing complex relationships: Decision trees may struggle to capture complex relationships between features, especially when the relationships are non-linear.

V. CONCLUSION

Our initial step was to explore the train dataset, during which we gained an understanding of its characteristics and identifying important features. In the data preparation stage, we addressed missing values, converted data types, categorized Age, SibSp, Parch, and Fares values, and created a few new attributes. Then we proceeded to use Decision tree machine learning model and applied cross validation on it, analyzed its confusion matrix. Finally, we run decision tree for the test set and compared the prediction results to the actual information of the incident.

The J48 algorithm was used for data classification throughout this process. Although this technique has limitations, it can be enhanced to become a highly precise classification algorithm.