CST8502 - Lab 3
Classification by Decision Trees

Student Name: Alvin Litani Liauw
Student Number: 41118874

Q1-2

Attribute Name	description	Correct Data Type	Data quality issues
PassengerID	ID of passengers	Nominal	Missing values of “62” and “830”
Age	Age of passengers in years	Numeric	Missing data in some rows
Cabin	Cabin where passenger stayed	String	Missing values
Ticket	Passengers’ ticket number	String	Inconsistent formatting
Fare	Fare paid by passenger	Numeric	Some entries has 4 decimal places even though currencies only have 2 decimal places

Q3. Screenshot

72	72	0	3	'Goodwin, Miss. Lillian Amy'	female	16	5	2	CA 2144	46.9		S
73	73	0	2	'Hood, Mr. Ambrose Jr'	male	21	0	0	S.O.C. 14879	73.5		S
74	74	0	3	'Chronopoulos, Mr. Apostolos'	male	26	1	0	2680	14.4542		C
75	75	1	3	Liauw, Mr. Alvin Litani'	male	37	7	4	41118	11	18	S
76	75	1	3	'Bing, Mr. Lee'	male	32	0	0	1601	56.4958		S
77	76	0	3	'Moen, Mr. Sigurd Hansen'	male	25	0	0	348123	7.65	F G73	S
78	77	0	3	'Staneff, Mr. Ivan'	male		0	0	349208	7.8958		S

Q8. Screenshot

Two columns for bins – equal width (range of fare) and equal frequency (instance/bins)

	Pclass	Sex	AgeGroup	Relatives	FareGroupWidthBin	FareGroupFrequencyBin	Embarked	Survived
2	3	male	Youth	Few	1	1	S	0
3	1	female	Adult	Few	1	6	C	1
4	3	female	Adult	None	1	2	S	1
5	1	female	Adult	Few	1	6	S	1
6	3	male	Adult	None	1	2	S	0
7	3	male	NK	None	1	2	Q	0
8	1	male	Adult	None	1	5	S	0
9	3	male	Child	Many	1	4	S	0
10	3	female	Adult	Few	1	3	S	1
11	2	female	Teen	Few	1	5	C	1
12	3	female	Child	Few	1	4	S	1
13	1	female	Adult	None	1	5	S	1
14	3	male	Youth	None	1	2	S	0
15	3	male	Adult	Many	1	5	S	0

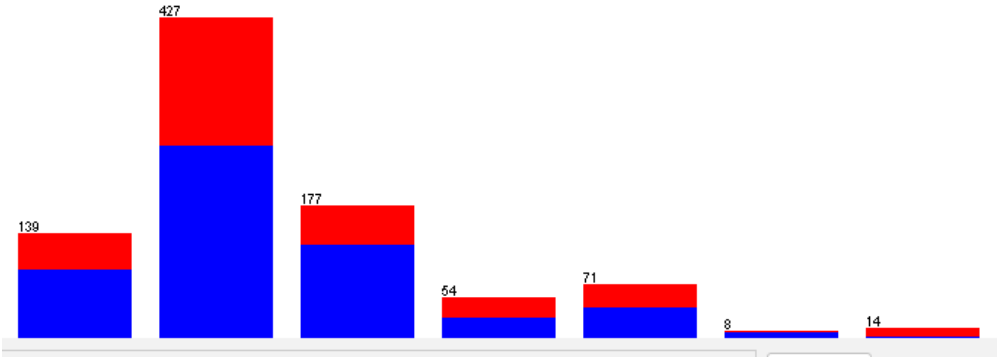
Q10. Screenshots

Selected attribute

Name: AgeGroup
Missing: 0 (0%)
Distinct: 7
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	Youth	139	139
2	Adult	427	427
3	NK	177	177
4	Child	54	54
5	Teen	71	71
6	Senior	8	8
7	Baby	14	14

Class: Survived (Nom) | Visualize All

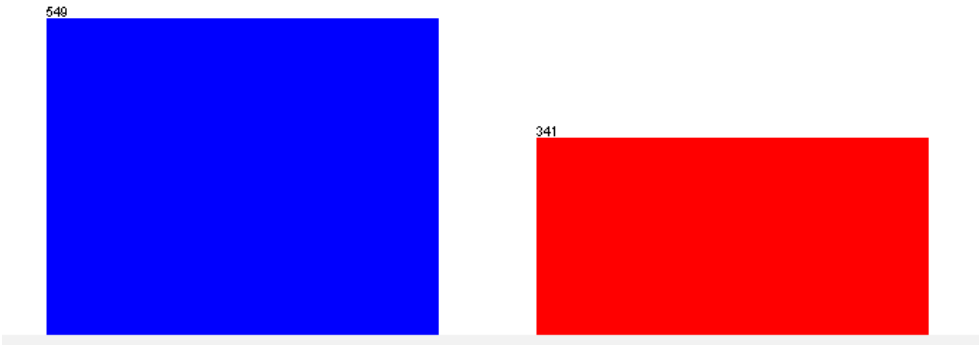


Selected attribute

Name: Survived
Missing: 0 (0%)
Distinct: 2
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	0	549	549
2	1	341	341

Class: Survived (Nom) | Visualize All



Q11. Screenshot

titanic_train_processed.arff

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

@attribute Pclass {1,2,3}
@attribute Sex {male,female}
@attribute AgeGroup {Youth,Adult,NK,Child,Teen,Senior,Baby}
@attribute Relatives {Few,None,Many}
@attribute FareGroupWidthBin {1,2,3,4,6}
@attribute FareGroupFrequencyBin {1,2,3,4,5,6}
@attribute Embarked {S,C,Q}
@attribute Survived {0,1}

@data
3,male,Youth,Few,1,1,S,0
1,female,Adult,Few,1,6,C,1
3,female,Adult,None,1,2,S,1
1,female,Adult,Few,1,6,S,1
3,male,Adult,None,1,2,S,0
3,male,NK,None,1,2,Q,0
1,male,Adult,None,1,5,S,0
3,male,Child,Many,1,4,S,0
3,female,Adult,Few,1,3,S,1
```

Q12. Screenshot

Titanic_test_processed.arff

```
@relation Titanic_test-weka.filters.unsupervised.attribute.NumericToNomina

@attribute Pclass {1,2,3}
@attribute Sex {male,female}
@attribute AgeGroup {NK,Adult,Child,Youth,Teen,Baby,Senior}
@attribute Relatives {None,Few,Many}
@attribute FareGroupWidthBin {1,2,3,4,6}
@attribute FareGroupFrequencyBin {1,2,3,4,5,6}
@attribute Embarked {S,C,Q}
@attribute Survived {0,1}

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

Q13.

Using the fare binned into equal width,

a) Confusion matrix

a	b	<-- classified as

518	31	| a = 0
140	201	| b = 1

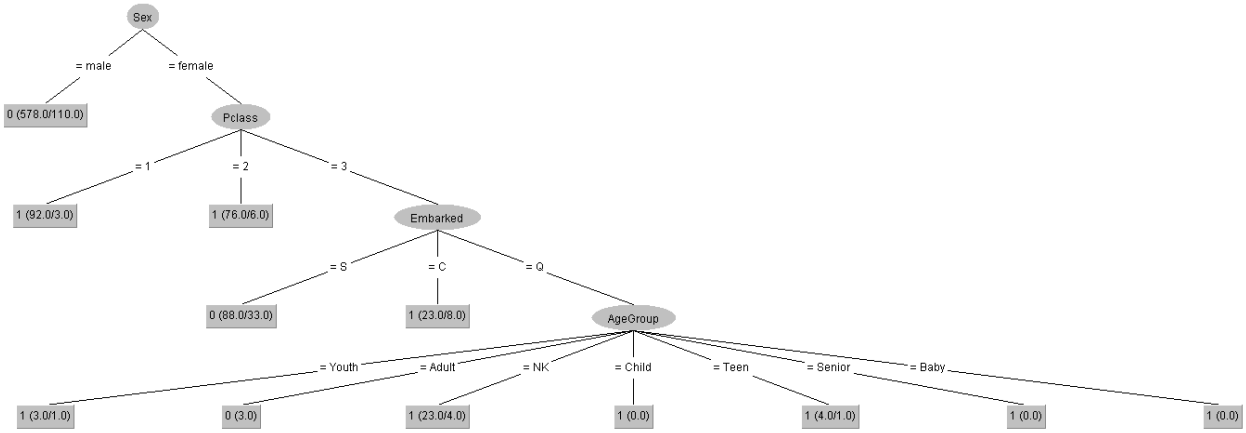
b) Accuracy

Accuracy = 80.7865 %

c) F1 measure

F-measure = 0.798

d) Tree



e) Rules

A male would be more likely to not survive. Out of 578 males, 110 did survive.

A female who was in passenger class 1 would be more likely to survive. Out of 92 females in class 1, 3 did not survive.

A female who was in passenger class 2 would be more likely to survive. Out of 76 females in class 1, 6 did not survive.

For a female who was in passenger class 3 and embarked in Southampton, she would be more likely to not survive. Out of 88 females from Southampton, 33 survive.

For a female who was in passenger class 3 and embarked in Cherbourg, she would be more likely to survive. Out of 23 females from Cherbourg, 8 did not survive.

Using the fare binned into equal frequency,

a) Confusion matrix

```
  a     b       <-- classified as
509    40     |   a = 0
137   204     |   b = 1
```

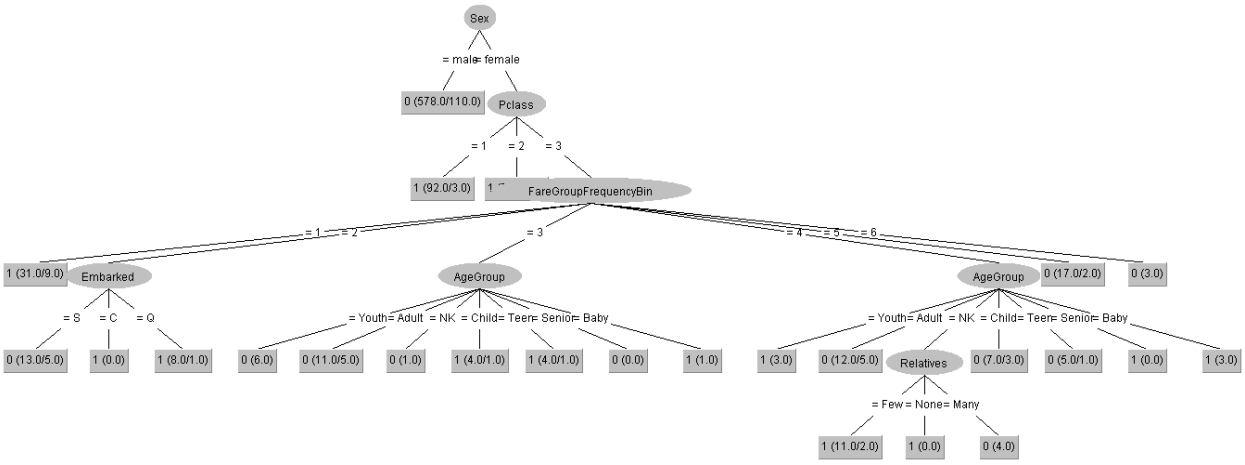
b) Accuracy

Accuracy = 80.1124 %

c) F1 measure

F-measure = 0.793

d) Tree



e) Rules

A male would be more likely to not survive. Out of 578 males, 110 did survive.

A female who was in passenger class 1 would be more likely to survive. Out of 92 females in class 1, 3 did not survive.

A female who was in passenger class 2 would be more likely to survive. Out of 76 females in class 1, 6 did not survive.

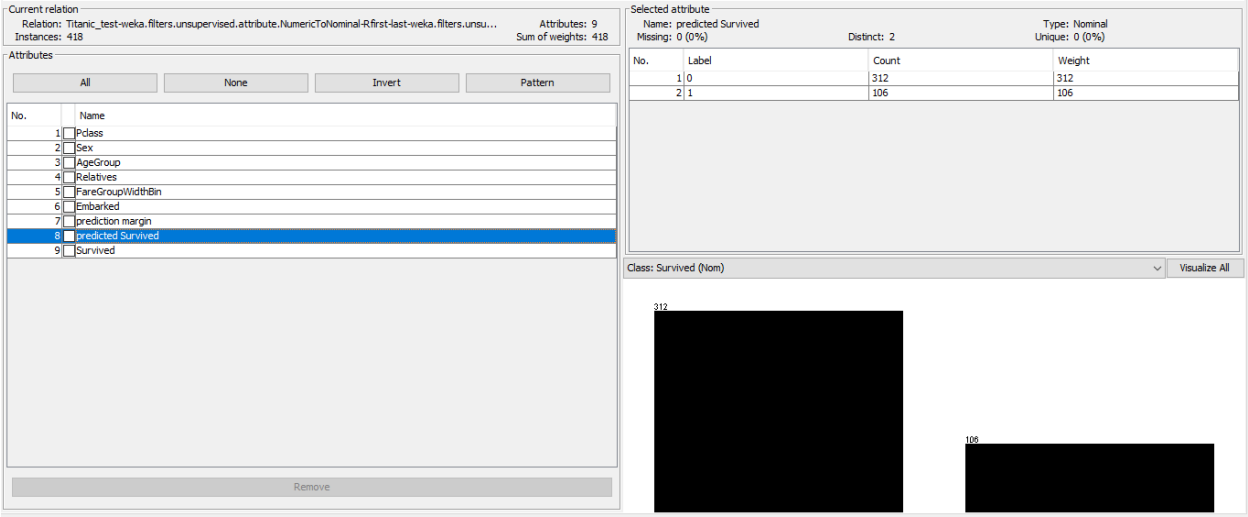
For a female who was in passenger class 3 and in the fare frequency bin 1, she would be more likely to survive. Out of 31 females, 9 did not survive.

For a female who was in passenger class 3 and in the fare frequency bin 5, she would be more likely to not survive. Out of 17 females, 2 did survive.

Q16. Screenshot

For equal width fare bin prediction,

1. Total instances in the test file: 418
2. Number of persons predicted to survive (1): 106
3. Number of persons predicted not to survive (0): 312
4. Percentage of predicted survival: 106/418 * 100% = 25.36%



For equal frequency fare bin prediction,

5. Total instances in the test file: 418
6. Number of persons predicted to survive (1): 118
7. Number of persons predicted not to survive (0): 300
8. Percentage of predicted survival: 118/418 * 100% = 28.23%