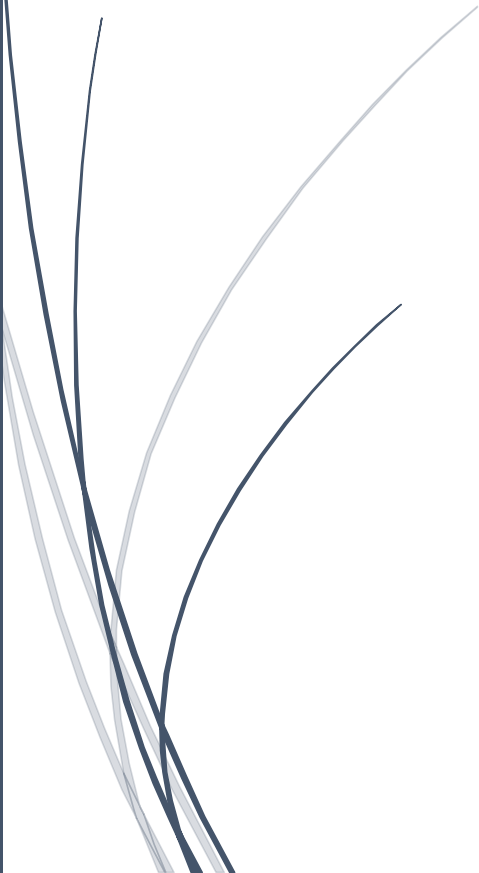


A dark blue vertical bar on the left side of the page. A blue arrow points to the right from the bar, containing the date.

3/25/2020

CST8390 Assignment 3

Decision Trees- Titanic Dataset

Several thin, curved lines in dark blue and light grey originate from the bottom left corner and curve upwards and to the right.

Rezansoff Karl
040955782

Li Min
040930563

Step 1- The Titanic Dataset

For this assignment we have been given two datasets (Titanic_Train.csv with 889 passengers, Titanic_Test.csv with 418 passengers) with data about the passengers on the Titanic such as sex, age, number of family members, fare paid, cabin, and where the passengers embarked from (all attributes listed below). Both datasets have the same attribute names, but the train dataset includes whether a passenger survived or not, while the test dataset does not. In this assignment we will use the patterns found in the train dataset to perform classification on the test dataset using Decision Trees and figure out which passengers in the test dataset survived. Overall, the purpose of this assignment is to identify which passengers from the test dataset survived, identify what sorts of passengers more likely to survive, and provide insight on what factors contributed to making a passenger more likely to survive. Afterwards, we can look at the actual survival rates from the Titanic incident and analyze the accuracy of our model.

Dataset Attributes:

PassengerId	Id given to each traveler on the ship
Pclass	The passenger class (1 st , 2 nd , 3 rd) with 1 st class being the most expensive fare paid.
Name	The Name of the passenger
Sex	Gender of the passenger (Male, Female)
Age	Numeric value for the age of the passenger. Many missing values (52 missing in test dataset) and given value "NK".
SibSp	Number of siblings and spouses traveling with the passenger
Parch	Number of parents and children traveling with the passenger
Ticket	The ticket id
Fare	The ticket fare amount
Cabin	The cabin number
Embarked	Describes 3 places passengers embarked from (S, C, Q) which are Southampton England, Cherbourg France, and Queenstown Ireland.
Survived	Describes if a passenger survived (1 for survived, 0 for deceased)

Step 2- Identifying the Relevant Attributes

Attributes Selected

Pclass	The passenger class (1 st , 2 nd , 3 rd) with 1 st class being the most expensive fare paid, this is a relevant way to categorize the passengers as a nominal attribute and we predict the higher-class passengers had a higher chance of surviving.
Sex	Gender of the passenger (Male, Female) this is relevant because there were both males and females on the ship, and we predict females and children may have had a better chance of surviving.

Embarked	Describes 3 places passengers embarked from (S, C, Q) which are Southampton England, Cherbourg France, and Queenstown Ireland. We think this is relevant as we predict the passengers who embarked from the same place may all be in the same area of the ship.
Survived	Describes if a passenger survived (1 for survived, 0 for deceased). This is necessary as this is our class variable for the decision tree or the variable we are trying to predict.

Attributes Added

Age group	<p>Nominal type to represent age:</p> <ul style="list-style-type: none"> • If age null, then NK • If age <12, then Child • If age < 18, then Teen • If age < 45, then Adult • If age ≤ 68, then Senior Adult • If age > 68, then Elderly <p>This is relevant as there may have been priority to board the lifeboats based on age, and we predict more children survived.</p>
Relatives	<p>Nominal type to represent number of relatives. It is the sum of Sibsp and Parch:</p> <ul style="list-style-type: none"> • If numRelatives 0, then Alone • If numRelatives <2, then Low • If numRelatives <5, then Average • else High <p>This is relevant as there may have been priority to board the lifeboats for the families as opposed to the</p>

File headers for Titanic_train_processed.csv:

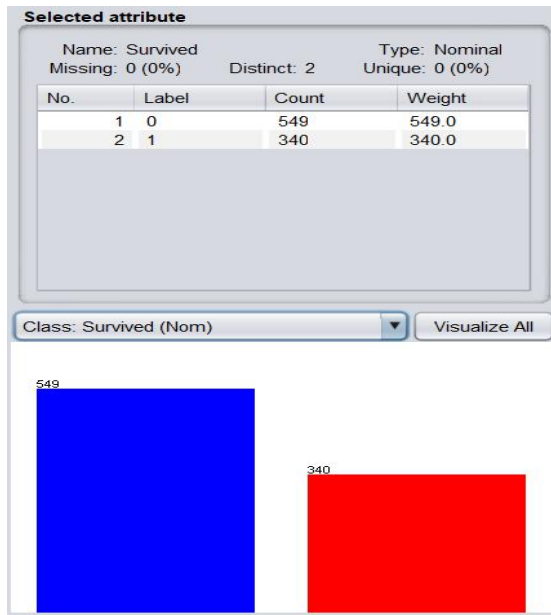
	A	B	C	D	E	F
1	Survived	Pclass	Sex	Embarked	Age Group	Relatives
2	0	3	male	S	Adult	Low
3	1	1	female	C	Adult	Low
4	1	3	female	S	Adult	Alone
5	1	1	female	S	Adult	Low
6	0	3	male	S	Adult	Alone
7	0	3	male	Q	NK	Alone
8	0	1	male	S	Senior Adult	Alone

Step 3- Loading data into Weka

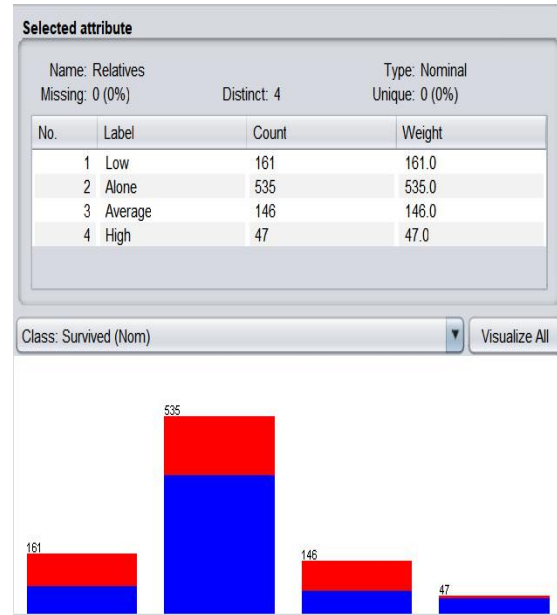
Once we opened our datasets in Weka, we changed all the attributes to nominal.

Step 4- Weka screenshots for training dataset

Distribution of the class attribute Survived:



Distribution of the attribute Relatives:



The distribution shows 549 passengers did not survive and 340 survived.

The distribution shows there are 535 passengers travelling alone, 161 passengers had less than 2 relatives, 146 passengers had less than 5 relatives, and 47 had more than 5 relatives.

Step 5

In this step we saved the training dataset as titanic_train_processed.arff

titanic_train_processed.arff

```
Titanic_train_processed.arff
1 @relation Titanic_train_processed-weka.filters.unsupervised.attribute.NumericToNominal-R1-weka.filters.unsupervised.attribute.NumericToNominal-R2
2
3 @attribute Survived {0,1}
4 @attribute Pclass {1,2,3}
5 @attribute Sex {male,female}
6 @attribute Embarked {S,C,Q}
7 @attribute 'Age Group' {Adult,NK,'Senior Adult',Child,Teen,Elderly}
8 @attribute Relatives {Low,Alone,Average,High}
9
10 @data
11 0,3,male,S,Adult,Low
12 1,1,female,C,Adult,Low
13 1,3,female,S,Adult,Alone
14 1,1,female,S,Adult,Low
15 0,3,male,S,Adult,Alone
16 0,3,male,Q,NK,Alone
17 0,1,male,S,'Senior Adult',Alone
18 0,3,male,S,Child,Average
19 1,3,female,S,Adult,Average
```

Step 6- Preparing test dataset

The test dataset was prepared with the same attribute headers as the training file, and since we had to add the survived attribute all the values in this column are "?".

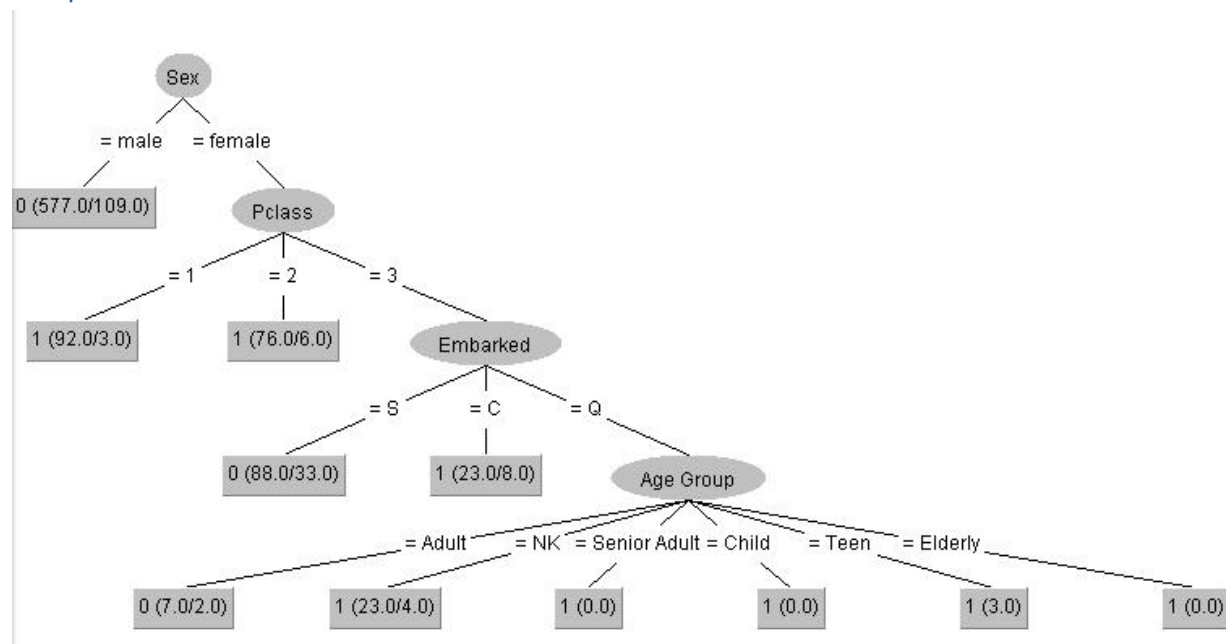
Step 7- Performing classification

Confusion Matrix:

```
=== Confusion Matrix ===  
  
  a   b  <-- classified as  
525  24 |   a = 0  
143 197 |   b = 1
```

In this step we performed classification using a J48 Decision Tree with 10-fold cross validation and the Confusion Matrix is shown above. The Confusion Matrix labelled not survived (0) as "a" and survived (1) as "b". The Confusion Matrix shows how well our model classified our samples, so for "a" it classified correctly 525/549 instances of "a" and for "b" classified 197/340 correctly. This shows our model is very inaccurate for predicting passengers who did not survive.

Step 8- Decision Tree Visualized



Looking at the visual of our Decision Tree, we can observe a visual representation of our rules and the accuracy of each rule. One rule does have a high misclassification rate (female -> 3rd class -> embarked S) which only correctly classified 72.7% of the passengers for that rule or leaving 33 passengers misclassified. Overall, we are satisfied with the results for classifying female passengers. When it comes to classifying males, our model only has one rule and that is to just classify males as "not survived" and is correct for 84% of the males. While this rule is more accurate than the worst performing rule on the

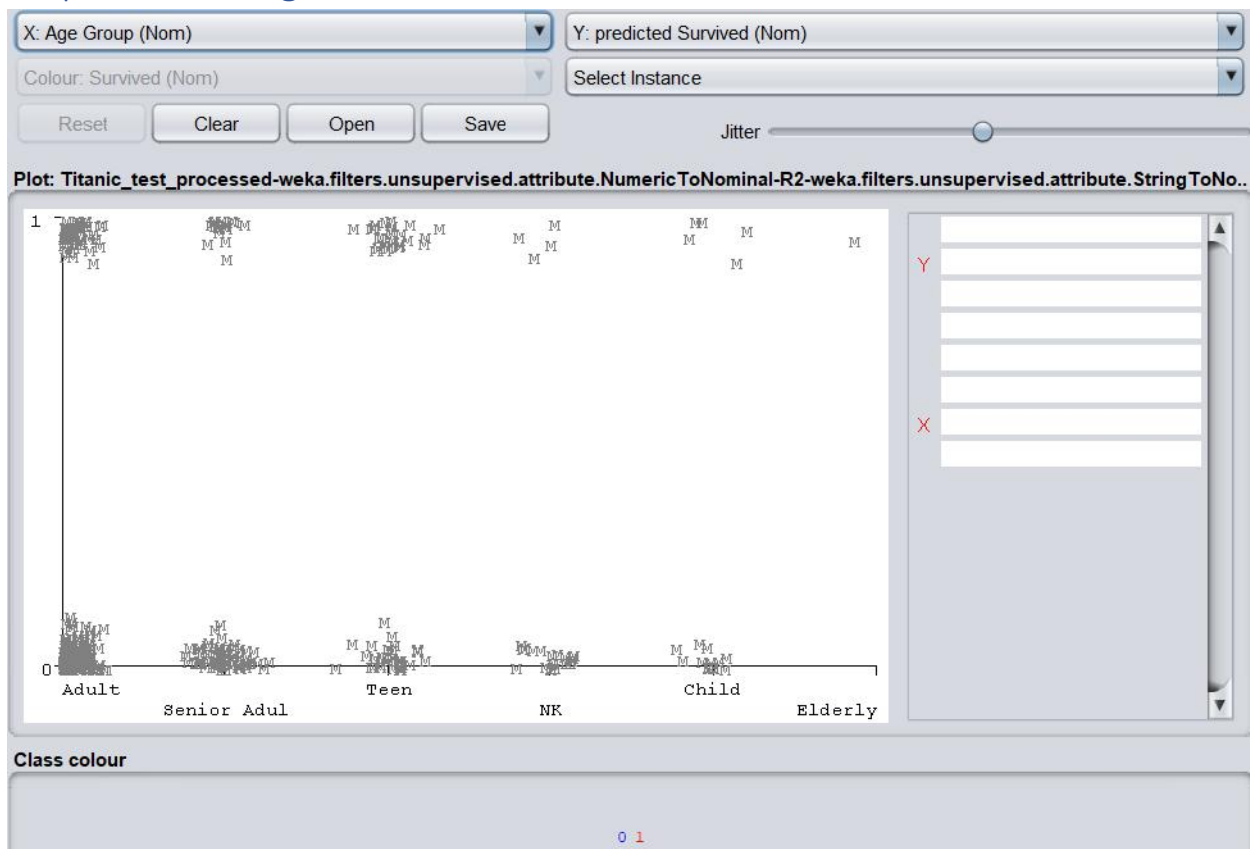
female side, this rule leaves 109 passengers misclassified, which is significant for our dataset of 889 instances.

Despite the accuracy of our Decision Tree, we could not figure out how to modify the parameters to create a more accurate model, so we decided to move forward anyways and see how it performs with the test dataset.

Step 10

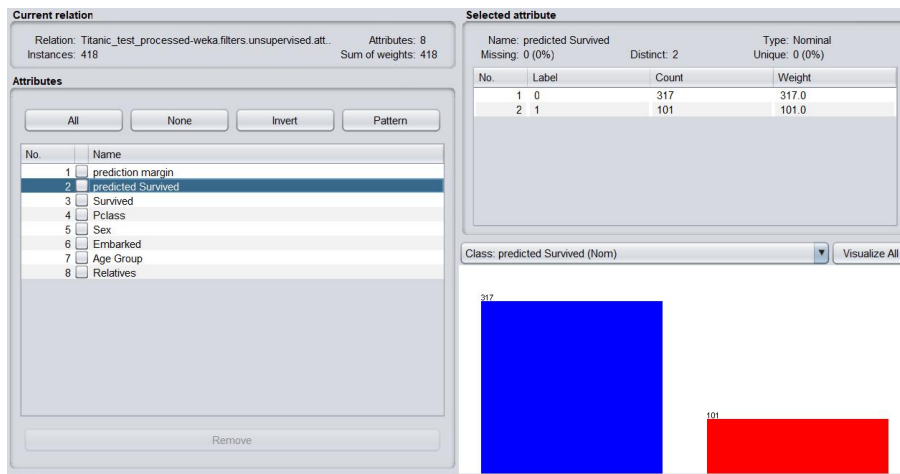
For this step instead of using Cross-Validation, we chose the option to supply a test set and supplied it the Titanic test dataset.

Step 11- Running Decision Tree for the test set



Step 12

- Total instances in the test file: 418
- Number of persons predicted to survive (1): 101
- Number of persons predicted not to survive (0): 317
- Percentage of predicted survival: 24% (101survived out of 418 passengers)



Step 13- Comparing to the actual results

Our predicted results are different from the actual results in the reference. Our percentage of predicted survival is only 24% while the actual survival percentage was 37%. We recognized early on when making our Decision Tree that our model was not accurate for classifying males who survived, so we expect this to skew our results, but we're hopeful we could get good results on classifying the women. Below are more comparisons to the actual results:

- Our predicted 1st class survivors is 47% (50 out of 107) and the actual result is 61%, difference of 14%.
- Our predicted 2nd class survivors is 32% (30 out of 93) and the actual result is 42%, difference of 10%.
- Our predicted 3rd class survivors is 10% (21 out of 218) and the actual result 24%, difference of 14%.
- Our predicted female survivors was 66% (101 out of 152) and the actual result 75%, difference of 9%.
- Our predicated male survivors was 0% (0 out of 266) and the actual result 20%.

The percentage of males in each class are:

- 1st Class: 53% Male
- 2nd Class: 68% Male
- 3rd Class: 67% Male

Despite, the 2nd Class having the highest percentage of males and a higher misclassification rate (from the decision tree), 2nd class survivors was one of our best predictions (only off by 10%). Overall, our best prediction was in predicting female survivors which we expected.

So in our opinion, the first reason is probably we need to change some of the parameters to get closest percentage according to the actual survival percentage; the second reason is that after comparing 1st, 2nd, 3rd class, female, and male predicted survivors with the actual 1st, 2nd, 3rd class, female, and male survivors, we got different predicted results from the actual results; the last but is the most important reason, we think this is because our training dataset was not the full dataset (Titanic had 2,300 people on board), so we may have been missing some information.

For improving the accuracy of our model, I think we should have modified some of the parameters of our decision tree. Another thing we could have tried is using more data to train our model so we could more accurately classify men or tried removing the outliers from our dataset that may have been skewing our results. Lastly, I think another approach would have been trying out a few different types of Decision Trees, such as Random Forest and making comparisons on which model would work best for our situation.