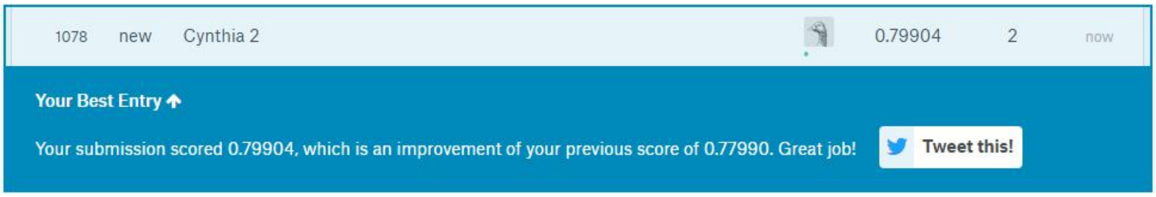
P4

Model	Data pre-processing steps	J48 parameters	Evaluation on training data	Evaluation on test data
1	Delete the attribute of Passenger name	J48 -C 0.25 -M 2	80.0224 %	0.77990
2	Delete Passenger name, ID, Cabin	J48 -C 0.1 -B -M 2	81.257 %	0.79904
3	Based on pretreatment of 2nd model, perform normalization on all numerical attributes	J48 -C 0.1 -B -M 2 -A	81.257 %	0.79426
4	Same pre-processing as 3rd model	J48 -U -M 2	79.6857 %	0.77990

Screenshot of the best score ranking:

(I cannot preseve score ranking for each model, because Kaggle only save the best ranking I've got)

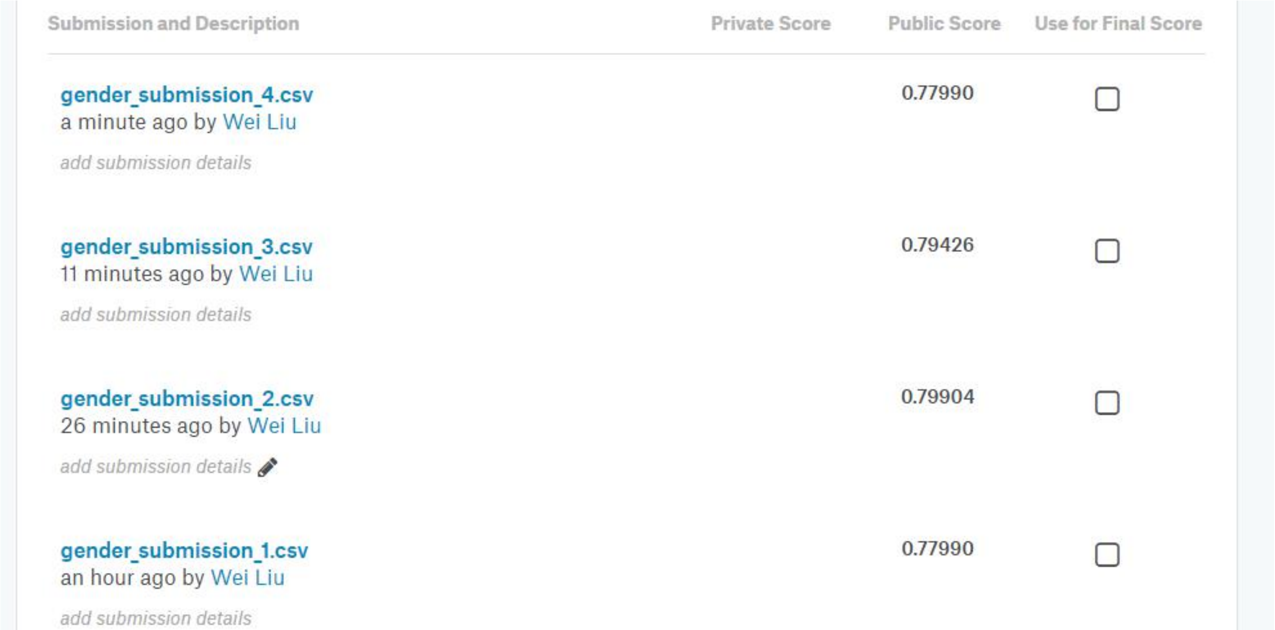


1078 new Cynthia 2 0.79904 2 now

Your Best Entry ↑

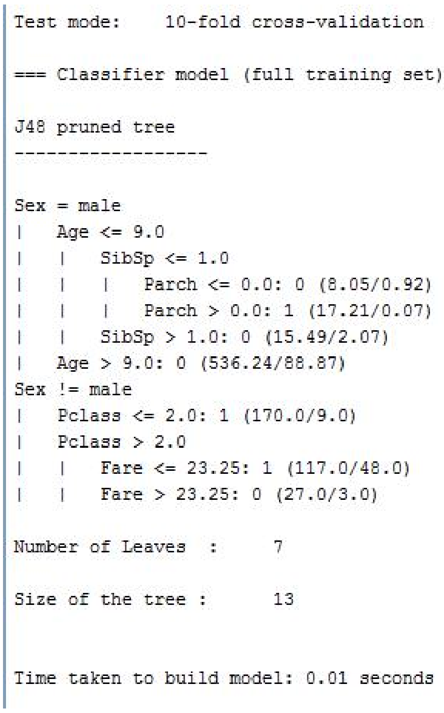
Your submission scored 0.79904, which is an improvement of your previous score of 0.77990. Great job!

Screenshot of the submissions:

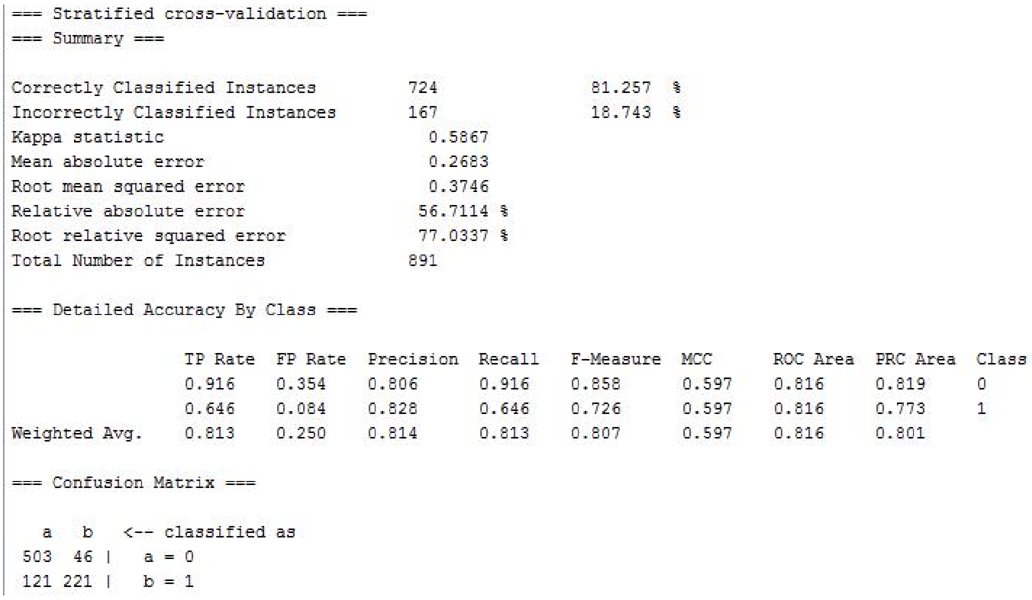


Submission and Description	Private Score	Public Score	Use for Final Score
gender_submission_4.csv a minute ago by Wei Liu add submission details		0.77990	□
gender_submission_3.csv 11 minutes ago by Wei Liu add submission details		0.79426	□
gender_submission_2.csv 26 minutes ago by Wei Liu add submission details		0.79904	□
gender_submission_1.csv an hour ago by Wei Liu add submission details		0.77990	□

The optimal J48 model:

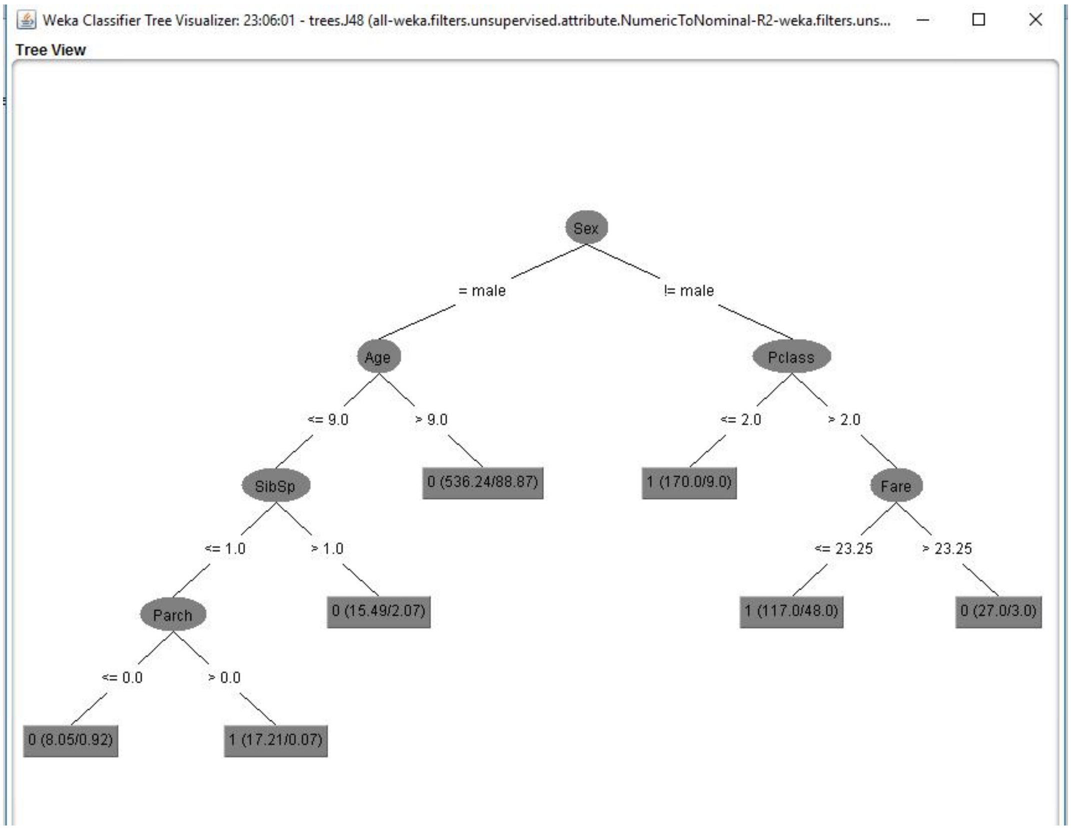


```
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
------------------

Sex = male
|   Age <= 9.0
|   |   SibSp <= 1.0
|   |   |   Parch <= 0.0: 0 (8.05/0.92)
|   |   |   Parch > 0.0: 1 (17.21/0.07)
|   |   SibSp > 1.0: 0 (15.49/2.07)
|   Age > 9.0: 0 (536.24/88.87)
Sex != male
|   Pclass <= 2.0: 1 (170.0/9.0)
|   Pclass > 2.0
|   |   Fare <= 23.25: 1 (117.0/48.0)
|   |   Fare > 23.25: 0 (27.0/3.0)

Number of Leaves  :     7

Size of the tree :      13


Time taken to build model: 0.01 seconds
```



```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         724               81.257  %
Incorrectly Classified Instances       167               18.743  %
Kappa statistic                          0.5867
Mean absolute error                      0.2683
Root mean squared error                  0.3746
Relative absolute error                 56.7114 %
Root relative squared error             77.0337 %
Total Number of Instances              891

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.916    0.354    0.806      0.916    0.858      0.597    0.816     0.819     0
                 0.646    0.084    0.828      0.646    0.726      0.597    0.816     0.773     1
Weighted Avg.    0.813    0.250    0.814      0.813    0.807      0.597    0.816     0.801

=== Confusion Matrix ===

   a   b   <-- classified as
 503  46 |   a = 0
 121 221 |   b = 1
```

The visualized J48 tree:



What I have found for J48 classifier:

1. The data preprocessing is equally important versus the parameters determination of the models. For example, the deletion of the useless attributes; Normalization of the numerical attributes; Converting of the numerical attributes to nominal attribute (in this case, convert “survived” (0/1) to nominal attribute is essential); the treatment of the missing data and etc.
2. Normalize the numerical data in this case seems not helpful to improve the performance of the model.
3. It was shown that pruning or not is critical for the J48 classifier performance.

From this model, we can see some other interesting facts:

1. For male/female passenger, the critical factor for surviving are age and Pclass respectively.
2. The children have a much higher chance to survive.
3. For the children, more siblings may decrease the survive opportunity, but existence from parents is a positive factor for surviving.
4. For the Female with different Pclass, the fare is a good predictor, since it may related with the location of the passenger when the disaster was happening.