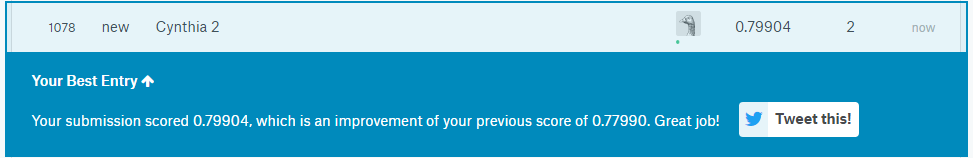
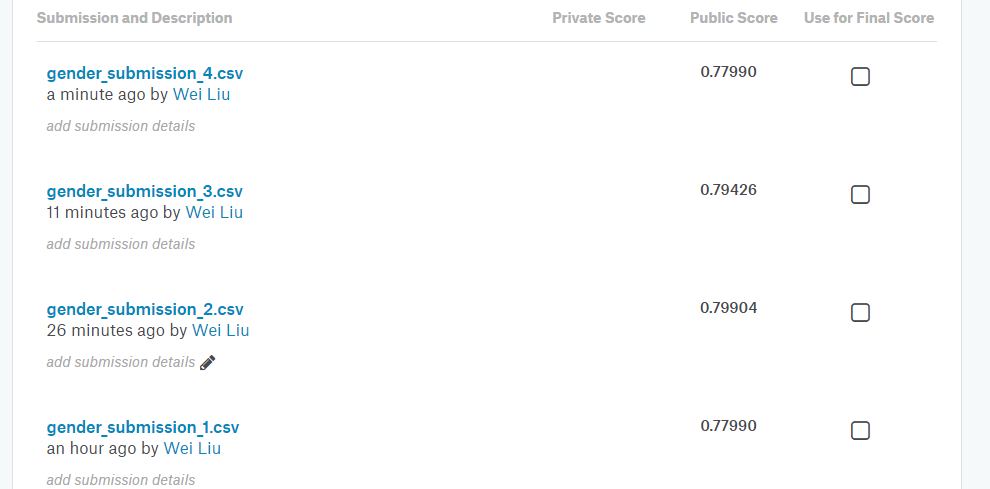
P4

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **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 |

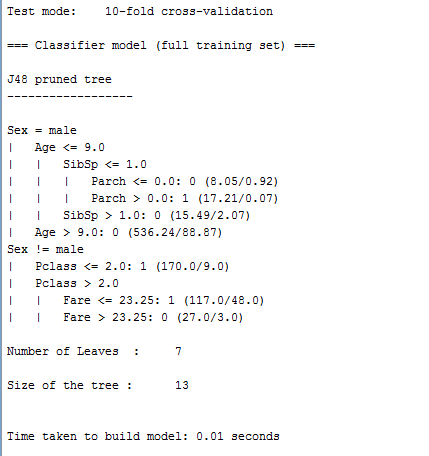
**Screenshot of the best score ranking:**

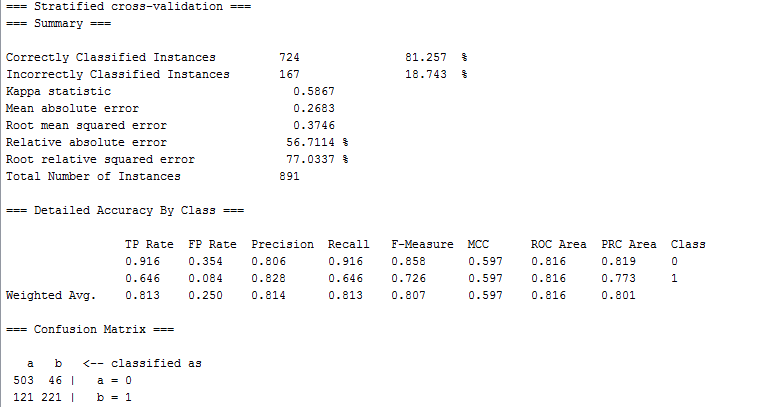
(I cannot preseve score ranking for each model, because Kaggle only save the best ranking I’ve got)



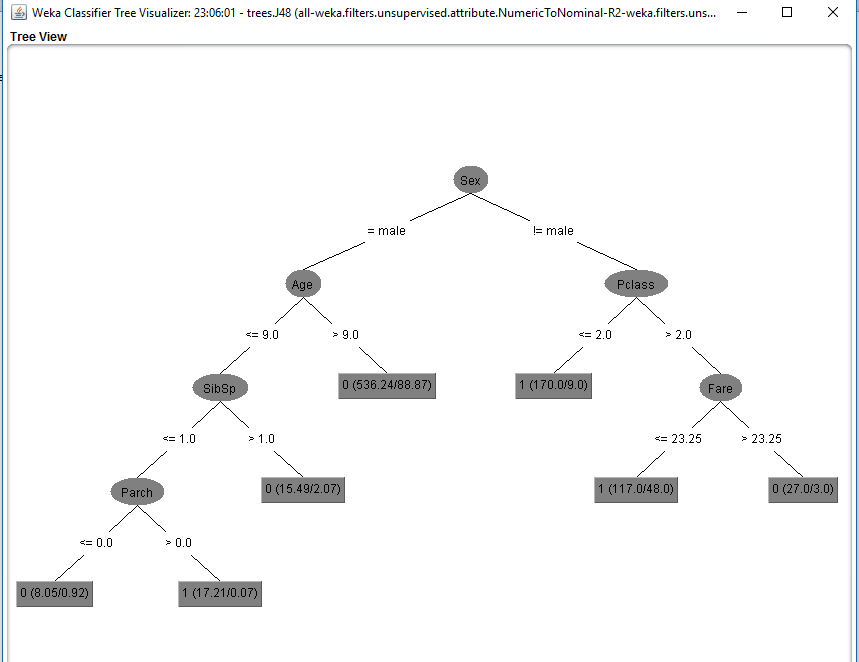
**Screenshot of the submissions:**

**The optimal J48 model:**





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.