# CS 573 Homework 3

Tingjun Li

**Brief Introduction**

In this assignment, I implemented both Decision Tree and Random Forest. For the Decision Tree, I implemented C4.5 algorithm which uses normalized information gain to split node.For the Random Forest, there are N different decision tree with bootstrapped sample of the original training set and split node within a random subset of features.

**Model Evaluation**

Decision Tree

| Dataset | TP | FN | FP | TN | Accuracy | Error Rate | Sensitivity | Specificity | Precision | F-1 Score |
|---|---|---|---|---|---|---|---|---|---|---|
| chess | 398 | 3 | 3 | 365 | 0.99219 | 0.00780 | 0.99251 | 0.99184 | 0.99251 | 0.99251 |
| mushroom | 1076 | 0 | 0 | 962 | 1.00000 | 0.00000 | 1.00000 | 1.00000 | 1.00000 | 1.00000 |
| nurse | 1076 | 0 | 0 | 1089 | 1.00000 | 0.00000 | 1.00000 | 1.00000 | 1.00000 | 1.00000 |

Random Forest

| Dataset | TP | FN | FP | TN | Accuracy | Error Rate | Sensitivity | Specificity | Precision | F-1 Score |
|---|---|---|---|---|---|---|---|---|---|---|
| chess | 398 | 2 | 3 | 365 | 0.99350 | 0.00650 | 0.99500 | 0.99187 | 0.99252 | 0.99376 |
| mushroom | 1076 | 0 | 0 | 962 | 1.00000 | 0.00000 | 1.00000 | 1.00000 | 1.00000 | 1.00000 |
| nurse | 1076 | 0 | 0 | 1089 | 1.00000 | 0.00000 | 1.00000 | 1.00000 | 1.00000 | 1.00000 |

**Parameters**

1. Number of trees: 100.

I chose 100 because overall, the more decision tree, the better the result. 100 is a good balance between runtime speed and test accuracy.

2. Percentage of Features to be considered: 0.6
3. Training set size: 0.8

For the second and third parameters, I chose them via a guess and correct method to get a slightly better test result. However, because our decision tree model is already performing extremely well on the test data, it is difficult to observe huge improvement from random forest to tune these two parameters.

**Conclusion**

Random forest can slightly improve the chess dataset by 0.0014 in accuracy. But this is not significant since the nature of randomness of this model. The other two test cases are already perfectly classified by the decision tree model and there are no room to improve.

Overall, I believe random forest are more robust than decision and can perform better in terms of accuracy. I think it cannot reflect in this project because our test data are too similar to our training data so that decision tree model achieved an uncommon high accuracy.