# Homework 1: Classification trees, random forests

Matej Kalc

### Decision Tree misclassification rates

| | Misclassification | Standard error |
|---|---|---|
| Train | 0% | 0% |
| Test | 20.69% | 5.40% |

**Table 1.** Misclassification rates on train and test sets using Decision Tree.

In Table 1, we observe the misclassification rate on train and test data set using Decision Tree. The standard error was calculated by using calculating the standard deviation of the misclassification of 100 decision trees with bootstrapped samples.

### Random Forest misclassification rates

| | Misclassification | Standard error |
|---|---|---|
| Train | 0% | 0% |
| Test | 1.724% | 1.738% |

**Table 2.** Misclassification rates on train and test sets using Random Forest (n=100).

In Table 2, we observe the misclassification rate on train and test data set using Random Forest (n = 100). When compared to the results of the decision tree (see Table 1), random forest is the clear winner with a much lower misclassification rate and a narrower standard error.

### Misclassification rates versus the number of trees n

Figure 1 shows misclassification rates versus the number of trees n. We observe that the plotted function has a negative trend. As the number of n increases, the misclassification decreases.

### Feature importance

In Figure 2 we observe that some features are more important than others. We also see that the features with a misclassification increase higher than 1% are also the most frequently used features in the roots of the generated decision trees. 66 out of 100 generated trees with bootstrap used the feature with the highest misclassification increase in the root node.

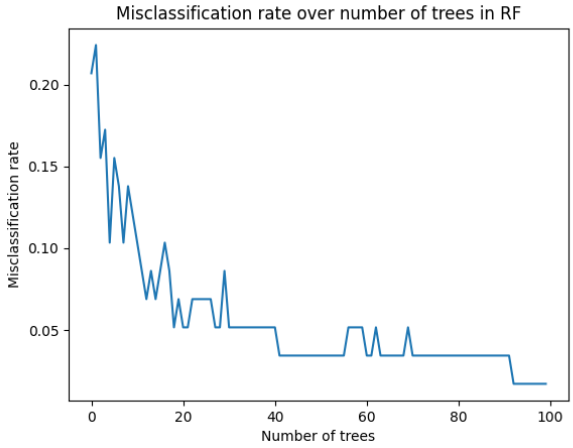


**Figure 1. Misclassification rate versus number of trees** The figure shows the misclassification rate at n trees.

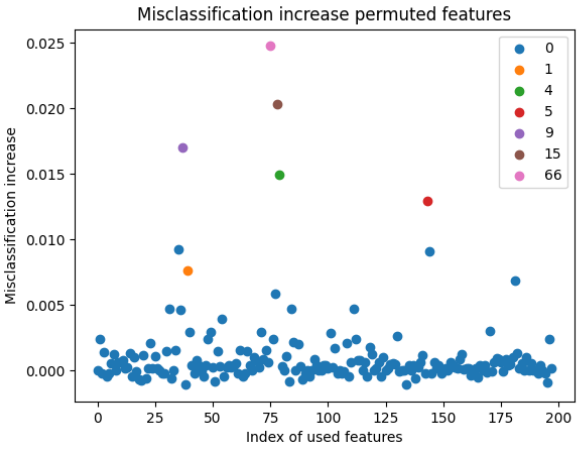


**Figure 2. Feature importance with RF** The figure shows the increase in misclassification when the feature at index i is shuffled. The color of each point indicates the number of times the specific column was the root node in a decision tree. 100 decision trees were constructed using bootstrap.