# Classification trees, random forests

Biljana Vitanova
*MLDS1 2024/25 , FRI, UL*
*bv7063@student.uni-lj.si*

## I. Introduction

The purpose of this project is to implement a classification tree and random forest, and further test them on an FTIR spectral dataset, report misclassification rates, and estimate uncertainty. Next, we apply permutation-based feature importance and further extend it to three features. We also implement tests to check the proper performance of the models

## II. Model Performance

### A. Uncertainty in Misclassification Rate

To estimate the uncertainty in the misclassification rate (MSR), we used two approaches:

1) Binomial SE estimate:
assuming each prediction error is an i.i.d. Bernoulli random variable. The standard error (SE) is computed as:

$$SE_{binomial} = \sqrt{\frac{MSR(1 - MSR)}{n}}$$

where $MSR$ is the misclassification rate, calculated as the proportion of misclassified samples used for the prediction.

2) Bootstrapped SE estimate:
The standard error is calculated by bootstrapping the prediction errors. This means resampling the predictions, calculating the misclassification rate $MSR_b$ for each bootrstrap sample, and computing the variance of these values. The standard error is:

$$SE_{\text{bootstrap}} = \sqrt{\frac{1}{B}\sum_{b=1}^{B}(MSR_b - M\bar{S}R)^2}$$

where $M\bar{S}R$ is the mean misclassification rate across bootstrap samples: $M\bar{S}R = \frac{1}{B}\sum_{b=1}^{B} MSR_b$

### B. Results

| Dataset | Approach | MSR |
|---|---|---|
| Train | Binomial SE<br>Bootstrap SE | $0 \pm 0$<br>$0 \pm 0$ |
| Test | Binomial SE<br>Bootstrap SE | $0.266 \pm 0.057$<br>$0.266 \pm 0.06$ |

TABLE I: Misclassification rates and standard errors for a decision tree.

| Dataset | Approach | MSR |
|---|---|---|
| Train | Binomial SE<br>Bootstrap SE | $0 \pm 0$<br>$0 \pm 0$ |
| Test | Binomial SE<br>Bootstrap SE | $0.216 \pm 0.053$<br>$0.224 \pm 0.056$ |

TABLE II: Misclassification rates and standard errors for a Random Forest with 100 trees.

### C. Misclassification Rate vs. Number of Trees

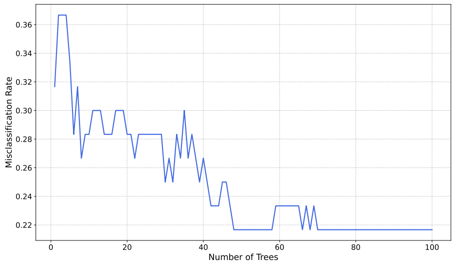

Fig. 1: The figure shows the relationship between the change in misclassification rate and the number of trees.

From the graph, we can see that the misclassification rate converges to the optimal value after 70 trees.

### D. Permutation-based variable importance

To estimate permutation-based variable importance, we used out-of-bag samples. For each variable, we measured its importance by computing the drop in accuracy after permuting its values and averaging the results across all trees.

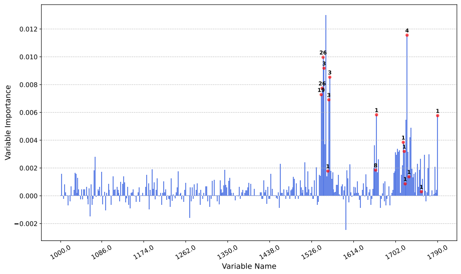

Fig. 2: The figure shows the variable importance, with red dots marking the most frequent root variables when no variable sampling is performed (non-random trees).

### E. Three-Variable Importance

| Dataset | MSR-single | MSR-triplet |
|---|---|---|
| Train | $0 \pm 0$ | $0 \pm 0$ |
| Test | $0.43 \pm 0.063$ | $0.255 \pm 0.053$ |

TABLE III: Misclassification rates and standard errors for tree built using the three most important single variables and the most important triplet (combination of three variables).

### F. Importance from Tree Structure and Unknown Data

To find the three most important variables with unknown data, we used two factors:

1) How often a variable combination appears in the trees (normalized occurrences).
2) The average depth of variables, with importance calculated as ($1 -$normalized depth), to assign higher value to shallow variable combinations.

The final score is the weighted sum of these two values. For this task, we gave more importance to depth because, as shown in previous results, RF had no significant drop in MSR. This means each tree performs well on its own.

Performance of single tree, using this approach resulted with MSR: $0.256 \pm 0.049$.

However, in general, we do not know the performance of the random forest or each individual tree, so the best approach is to use equal weights for both values.

### G. Remarks

We have written two manual tests to verify the correct construction of the tree and the handling of edge cases.

## III. Conclusion

We implemented a classification tree and a random forest, and evaluated their performance on an FTIR spectral dataset. The models were assessed using misclassification rates and uncertainty estimates, with random forests showing better generalization. Feature importance analysis showed key variables, and manual tests confirmed correct model behavior.