Applied Geodata Science I

Session 12

Prof. Dr. Benjamin Stocker

12.05.2025

GECO

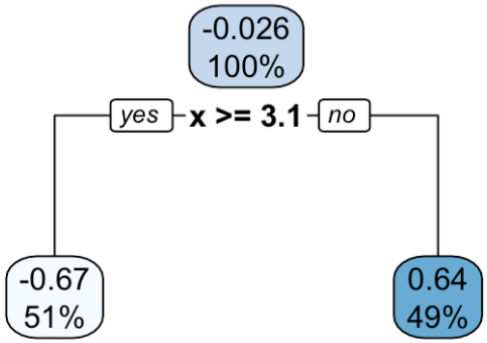
u^b

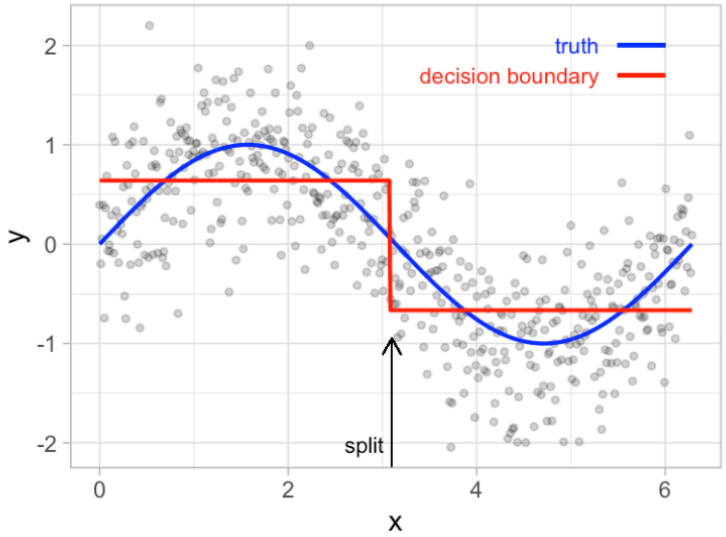
UNIVERSITÄT BERN

What's the weight of Field Marshall?

A decision tree

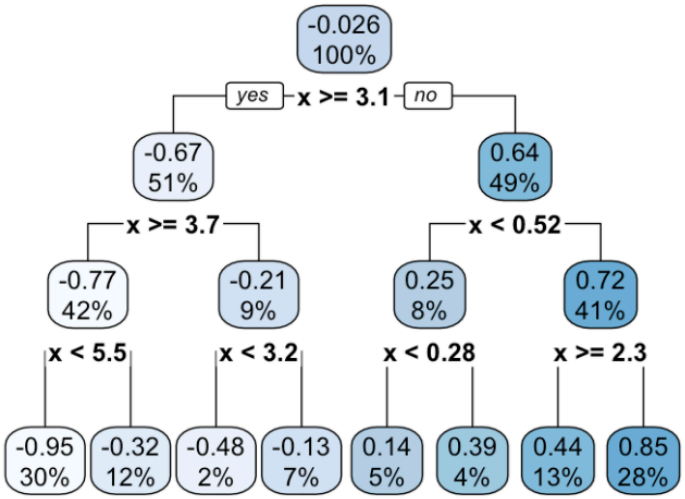
A decision tree

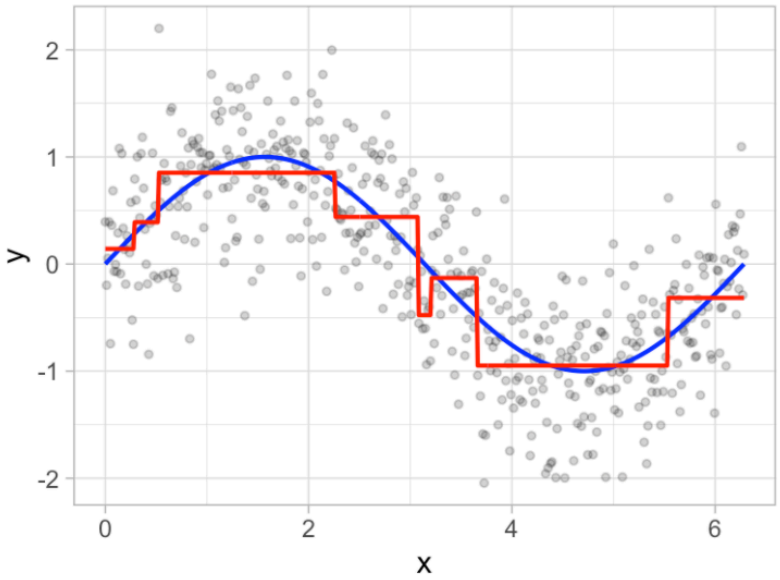




Bradley & Boehmke *Hands On Machine Learning in R*

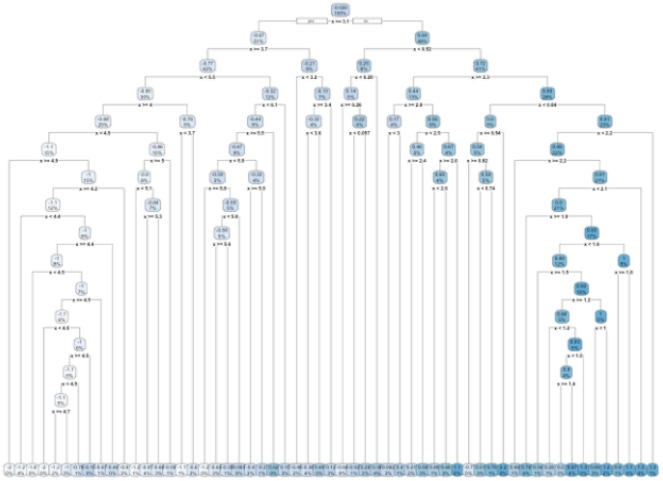
Tree growth

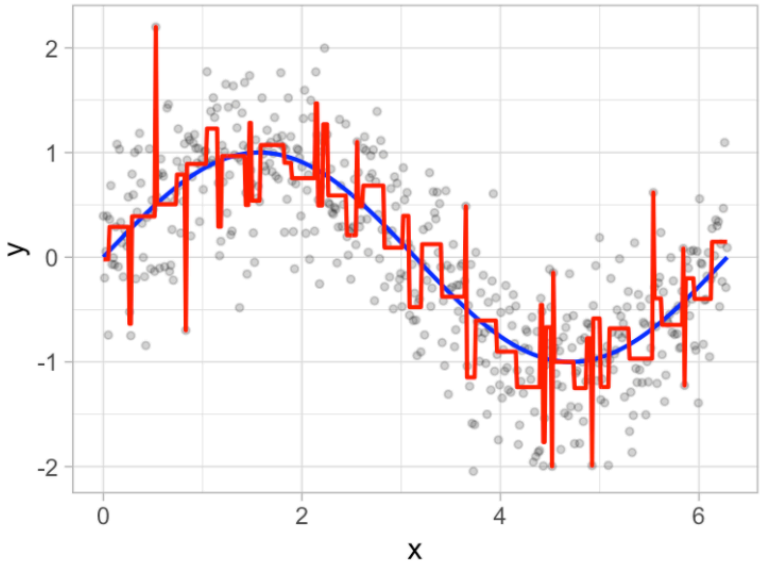




Bradley & Boehmke *Hands On Machine Learning in R*

Too deep a tree (overfitted)





Bradley & Boehmke *Hands On Machine Learning in R*

From a tree to a forest

By generating randomness and averaging across samples for a robust estimate:

- Subsampling predictors considered for each decision (branch) within a tree
- Subsampling number of training data points for each tree

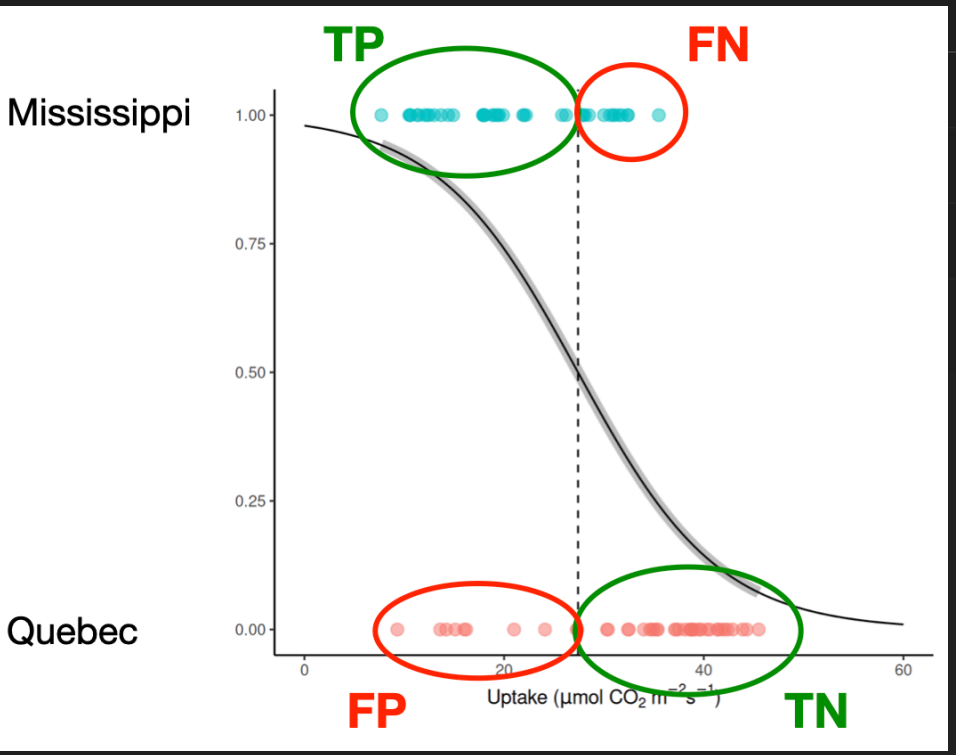
Hyperparameters in Random Forst

- number of trees
- *mtry*: number of predictors considered at each branch
- *min.node.size*: the number of data points at the “bottom” of each decision tree (leaf nodes)
- *splitrule*: the function applied for determining the goodness of a decision (default: SSE for regression, Gini impurity for classification).

Gini impurity

For a node with classes $k \in \{1, 2, \ldots, K\}$, and proportion p_k of class k, the **Gini impurity** is:

$$G = 1 - \sum_{k=1}^{K} p_k^2$$



Left group:

- Class A: 2/2 = 1.0
- Class B: 0/2 = 0.0

$$G_{\text{left}} = 1 - (1.0^2 + 0.0^2) = 0$$

Right group:

- Class A: 0/2 = 0.0
- Class B: 2/2 = 1.0

$$G_{\text{right}} = 1 - (0.0^2 + 1.0^2) = 0$$

$$G_{\text{split}} = \frac{2}{4} \cdot G_{\text{left}} + \frac{2}{4} \cdot G_{\text{right}} = 0.5 \cdot 0 + 0.5 \cdot 0 = 0$$