Multivariate statistics in R

Hannes PETER
Martin BOUTROUX
Zhe LIU

Updated schedule

- 10.09. **session 1**
- 17.09. **session 2**
- 24.09. **group work**
- 01.10. ***«modern R» with Martin*** (tidyverse)
- 08.10. **session 3**
- 15.10. **session 4**
- 22.10. autumn holidays
- 29.10. **group work**
- 05.11. **session 5**
- 12.11. **mid-term exam**, **group work**
- 19.11. **session 6**
- 26.11. **session 7**
- 03.12./10.12./17.12. **group work**
- 07.01. group presentations 1
- 14.01. group presentations 2

topics

- **data exploration** ✓
  - summary statistics
  - visualization
- **transformations**
- **resemblance metrics** ✓
  - dis/similarity, distance
- **unsupervised classification** ✓
  - cluster analysis
- **supervised classification**
  - classification and regression trees, random forest classifier
- **unconstrained ordination**
  - PCA, CA, NMDS
- **constrained ordination**
  - RDA, CCA
- **auxiliary multivariate analysis**
  - LDA, Mantel correlation, procrustes, etc…

Recap

- **data exploration**
  - summary statistics
  - visualization
- **transformations**
- **resemblance**
  - dis/similarity, distance
- **unsupervised classification**
  - cluster analysis

Recap

Classification

Unsupervised

search for main gradients and homogeneous groups in the data.

- ➢ No a priori knowledge/assumptions
- ➢ Results depend mainly structure of the dataset.

- distance/similarity metric, choice of clustering method
- assignment of samples into groups may change even with slight changes of the dataset (e.g. by adding more samples)
- examples of unsupervised methods are **cluster analysis**, TWINSPAN

Supervised

use external criteria to classify the dataset

- ➢ you supply information/rules about how to classify

- assignment of samples to groups remain the same despite changes in the structure of the dataset

- examples are classification and regression trees (**CART**), **random forest classifier**, artificial neural networks (ANN), etc.

(k-means clustering, can either be supervised or unsupervised)

Supervised classification

Classification tree analysis (**CT**) for qualitative response variables.

Regression tree analysis (**RT**) for quantitative response variables.

Classification and regression tree analysis (**CART**) combines these two procedures.

Random forest classifier combines bootstrapping and aggregating to produce more robust decision trees (useful for prediction).

Classification and Regression Trees (CART)

use two datasets

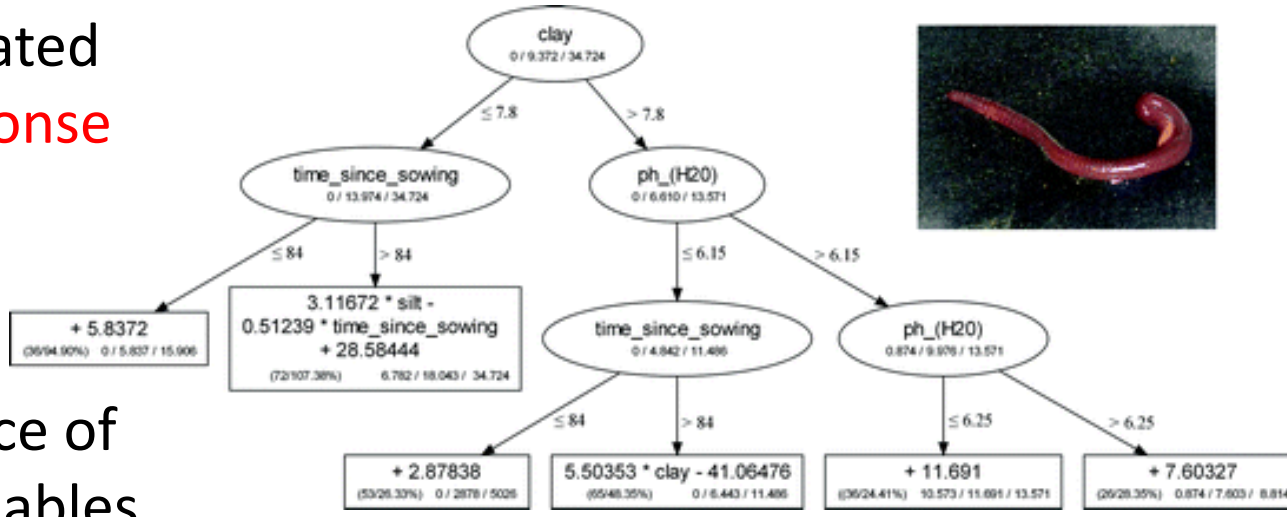
- **Univariate response**
- **Multivariate explanatory variables**
- divides the dataset into groups *(nodes)*
- applies a logical condition for each division *(binary splits)*

- Construction of a ***decision tree***
  - reads from top to bottom (divisive)
  - allows discriminating among explanatory variables

Decision Trees

Which factors determine the biomass of earthworms?

> tree-like diagram resulting from repeated splitting of the response data (dichotomous prevision model)



> shows the influence of the explanatory variables on the response at each split

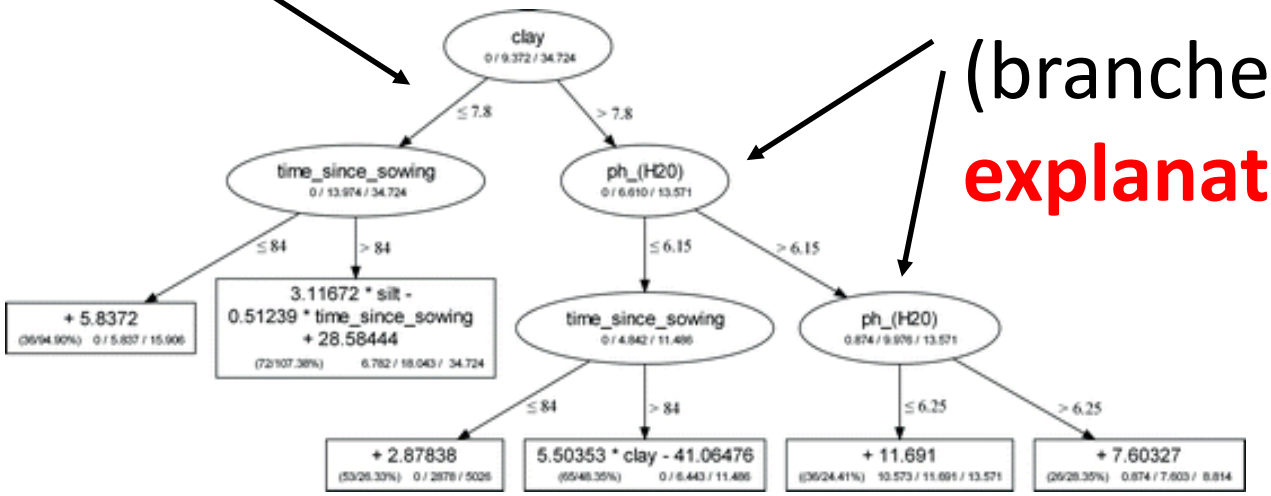
> allows prediction/decisions

Decision Tree

root node

True/False
Yes/No

internal nodes
(branches)
explanatory variables



leaf nodes/leaves
response

EPFL

an example:

What predicts whether or not a person loves Cool as Ice?

explanatory variables

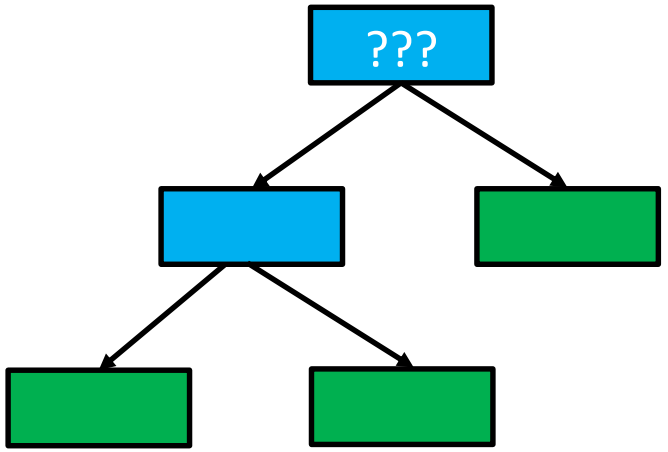
response

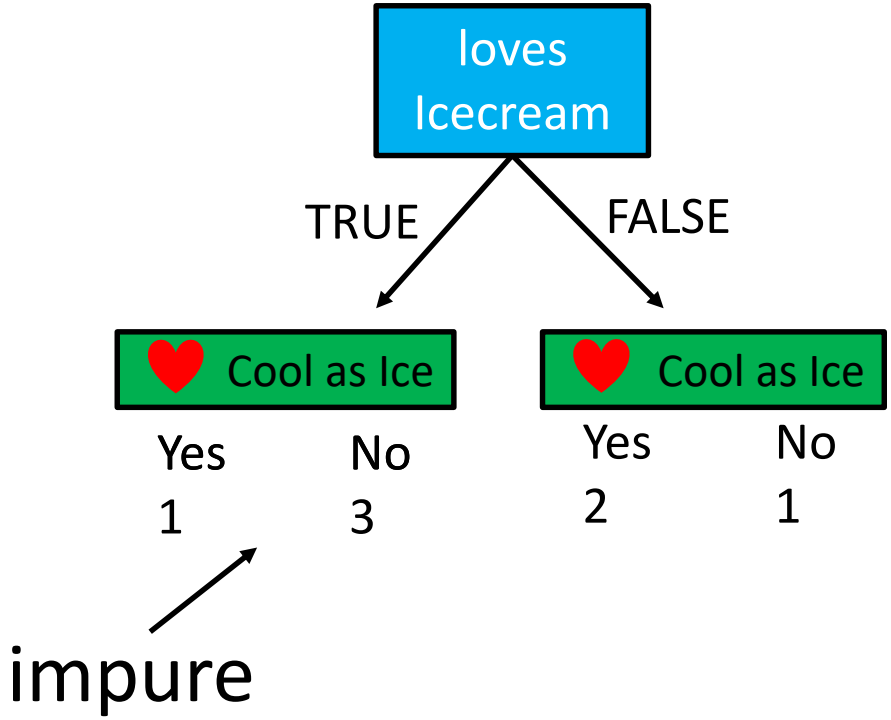
	Loves Icecream	Drinks coke	Age	Loves Cool as Ice?
Person 1	Yes	Yes	7	No
Person 2	Yes	No	12	No
Person 3	No	Yes	18	Yes
Person 4	No	Yes	35	Yes
Person 5	Yes	Yes	38	Yes
Person 6	Yes	No	50	No
Person 7	No	No	88	No

find the root…

icecream, coke, age?

	Loves Icecream	Drinks coke	Age	Loves Cool as Ice?
person1	Yes	Yes	7	No
person2	Yes	No	12	No
person3	No	Yes	18	Yes
person4	No	Yes	35	Yes
person5	Yes	Yes	38	Yes
person6	Yes	No	50	No
person7	No	No	88	No

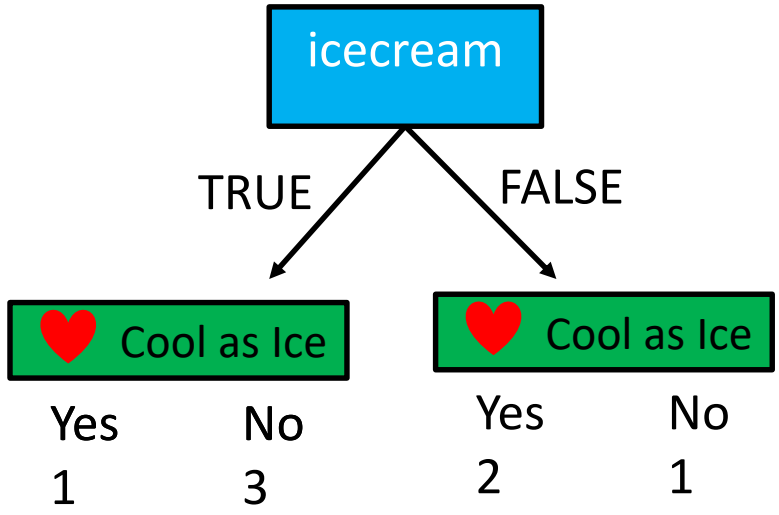


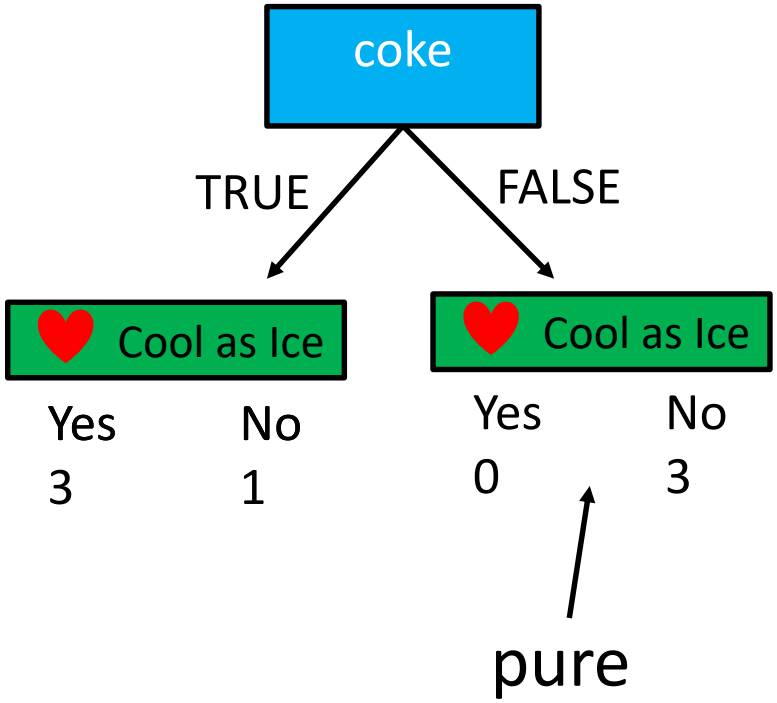


find the root…

icecream, coke, age?

	Loves Icecream	Drinks coke	Age	Loves Cool as Ice?
person1	Yes	Yes	7	No
person2	Yes	No	12	No
person3	No	Yes	18	Yes
person4	No	Yes	35	Yes
person5	Yes	Yes	38	Yes
person6	Yes	No	50	No
person7	No	No	88	No

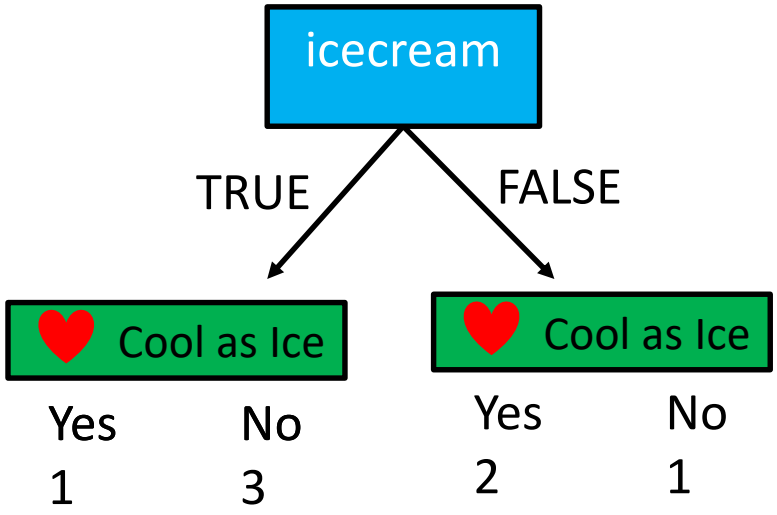




Gini Impurity

Gini = 1-(prob of *Yes*)^2-(prob of *No*)^2

range: 0 - 0.5

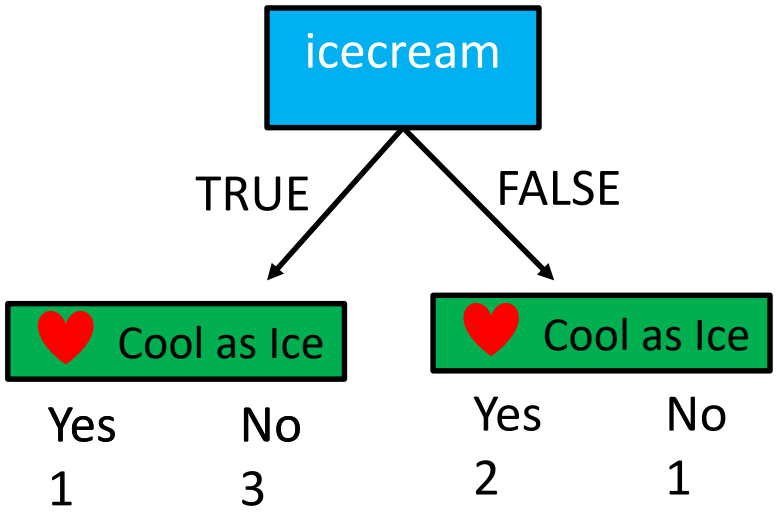


$$1 - (\frac{1}{1+3})^2 - (\frac{3}{1+3})^2 = 0.375 \qquad 0.444$$

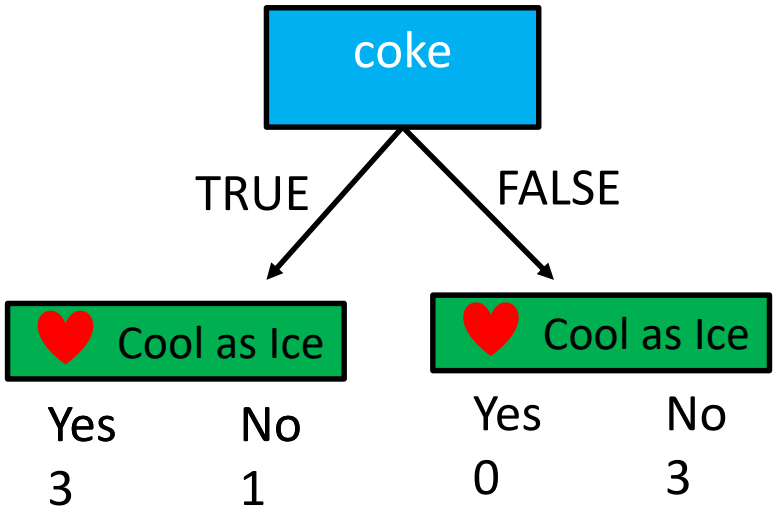
Gini Impurity:
weighted average of leaf impurities

$$(\frac{4}{4+3})*0.375 + (\frac{3}{4+3})*0.44 = 0.405$$

find the root…



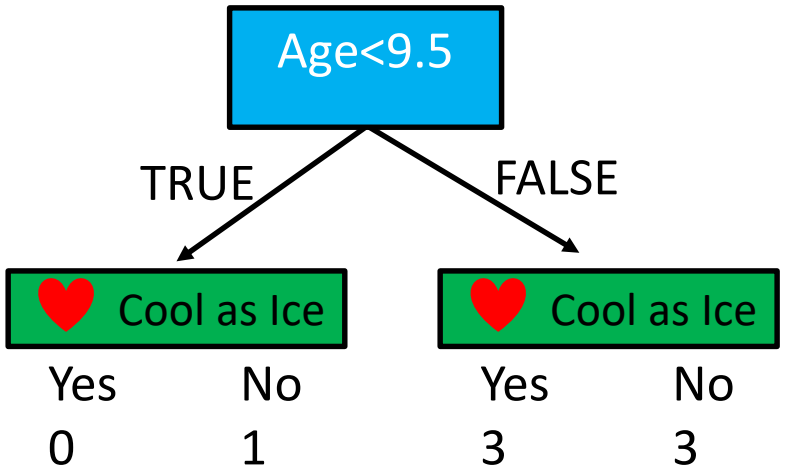
Gini Impurity: 0.405



Gini Impurity: 0.214

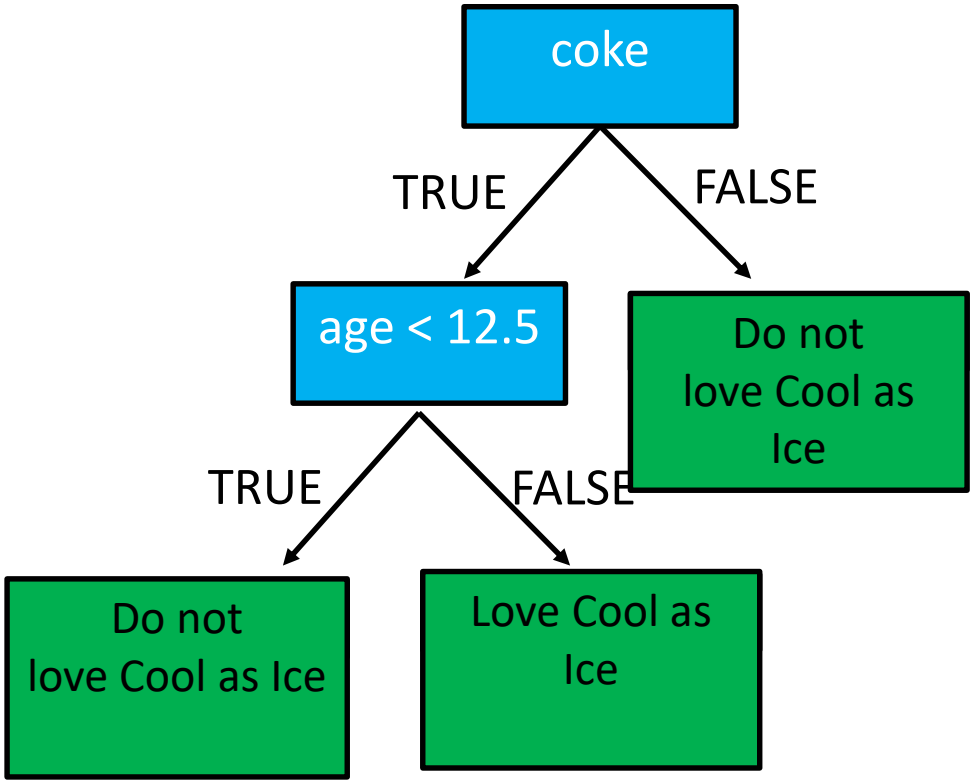
Gini Impurity of numerical data?

	Age	Loves Cool as Ice?	
9.5	7	No	0.429
15	12	No	0.343
26.5	18	Yes	0.476
36.5	35	Yes	0.476
44	38	Yes	0.343
66.5	50	No	0.429
	88	No	



Gini Impurity: 0.429

keep splitting impure nodes…



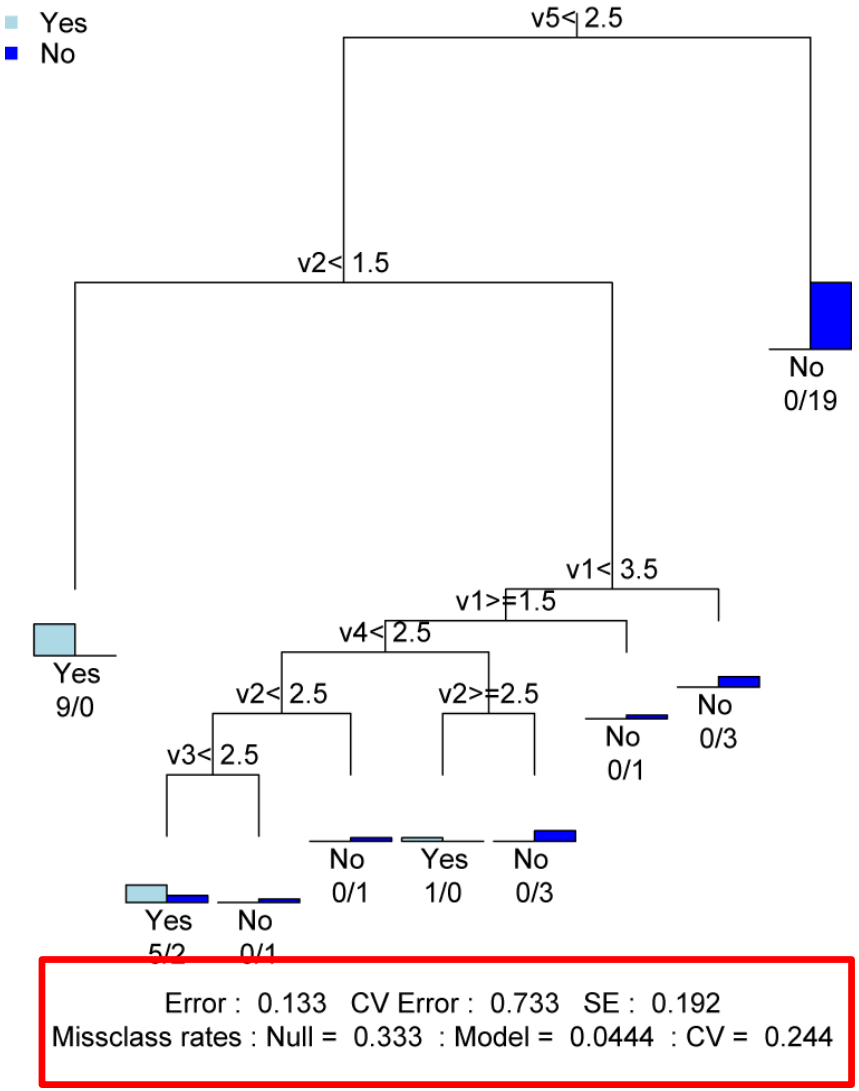
«majority vote»

Model error and Model selection

Model error (Error)

- misclassifications (errors) of the terminal leaves
- trends to 0 with increasing tree size

- *Prediction error, Relative error or Cross Validation Error (CV)*
  - sum of the errors of the terminal branches achieved by cross-validation
  - measures uncertainty of forecasting
  - *reaches a minimum* for an optimal tree size (=> pruning)



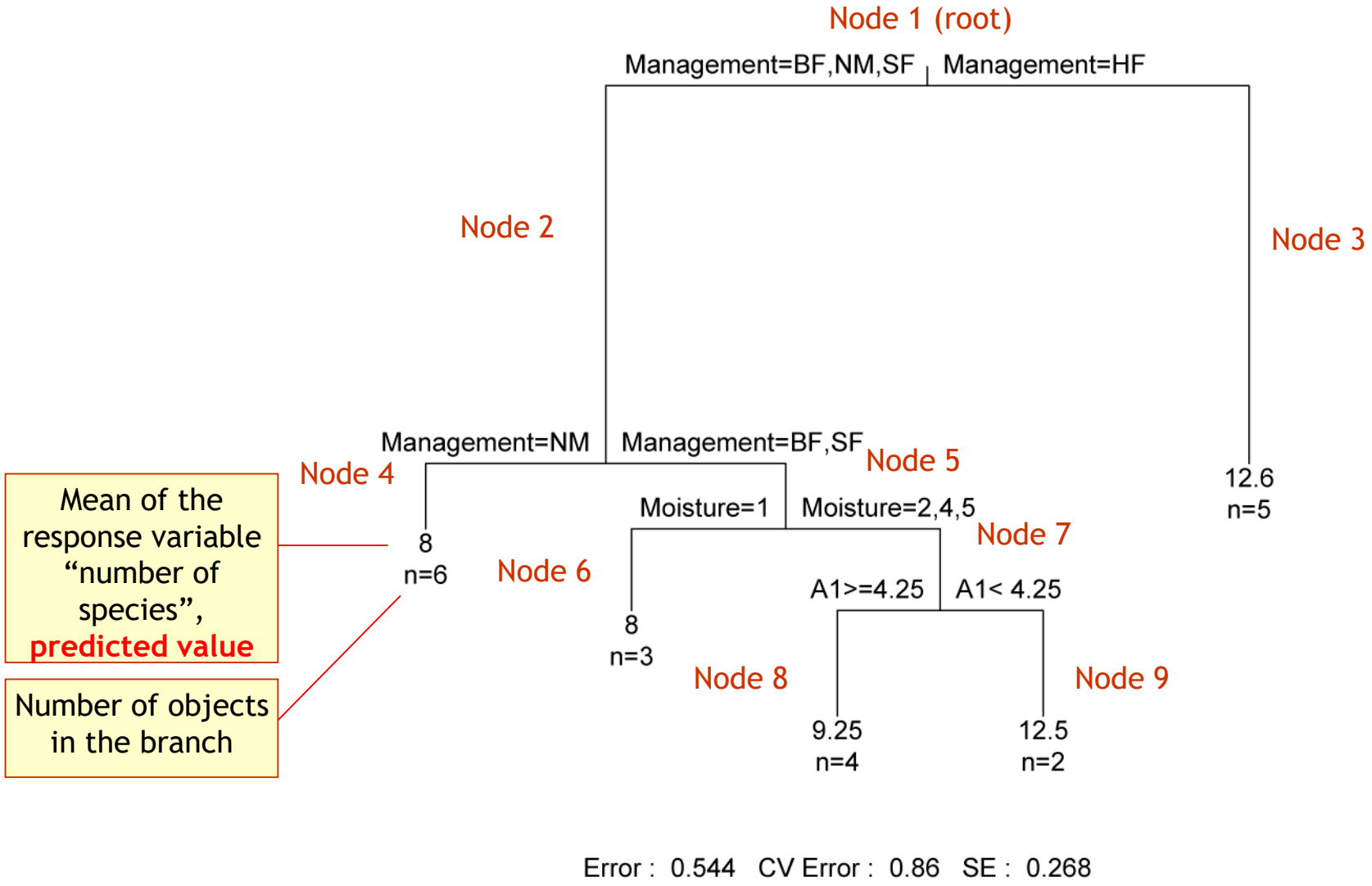
dune vegetation dataset (R::vegan)

- cover (Braun-Blanquet classes, 0 - 9) of 30 plant species at 20 dune meadow sites (2 x 2 m) in the Netherlands Batterink & Wijfels 1983
  - species names are abbreviated (4 genus + 4 species letters, ex: Achimill = Achillea millefolium)

- dune.env
  - 20 observations of environmental data
    - A1: thickness of soil A1 horizon (numeric)
    - Moisture: soil moisture with levels 1-5
    - Management: biological farming (BF), Hobby Farming (HF), NM (Nature and Conservation Management), SF (Standard farming) (factors)
    - Use: land-use with levels Hayfield, Haypastu, Pasture
    - Manure: factor with levels 0-4

Univariate regression tree

How does the number of plant species depend on management strategy?

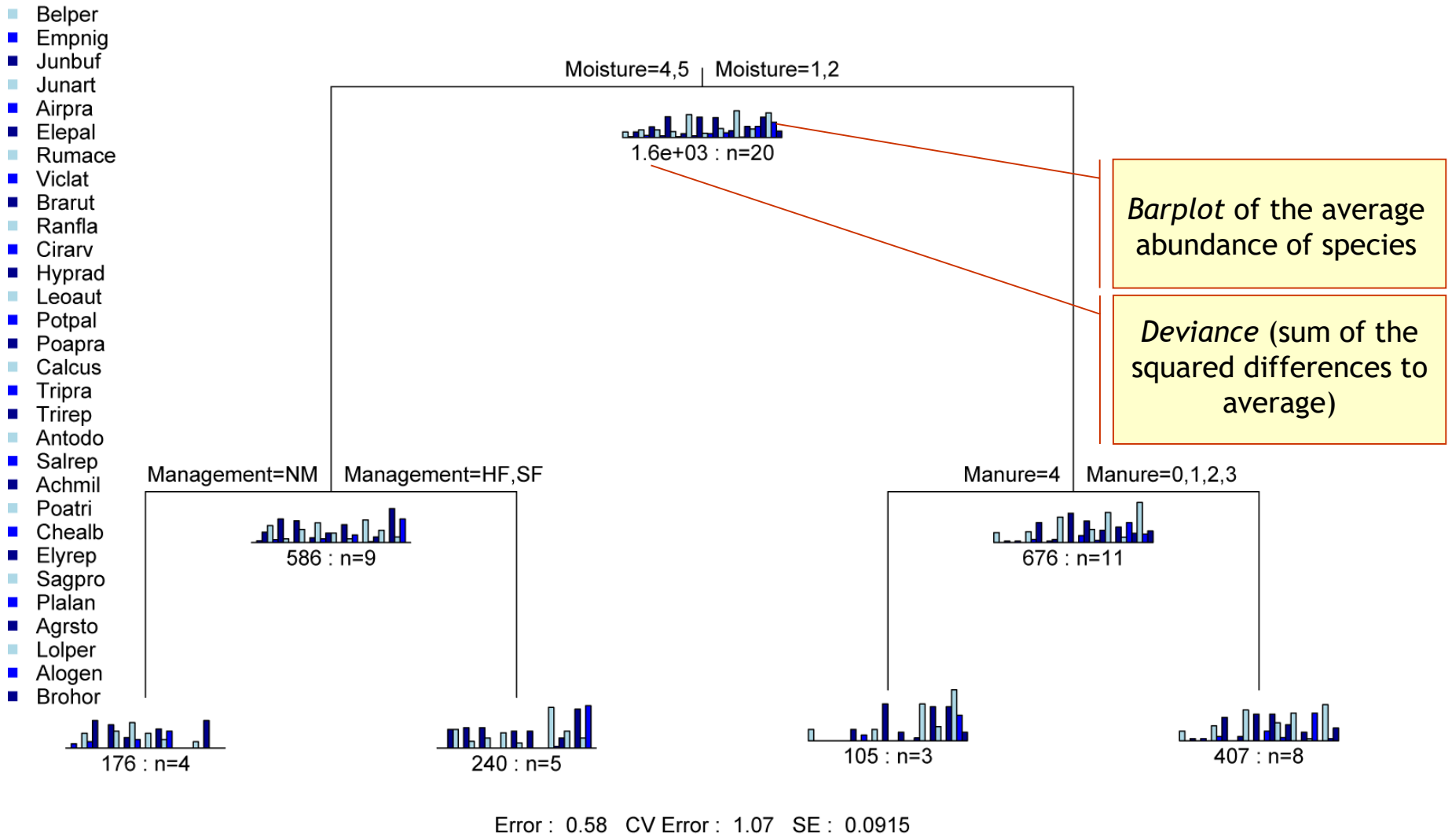


Error : 0.544 CV Error : 0.86 SE : 0.268

EPFL

multivariate regression tree

How do species abundances depend on management strategy?

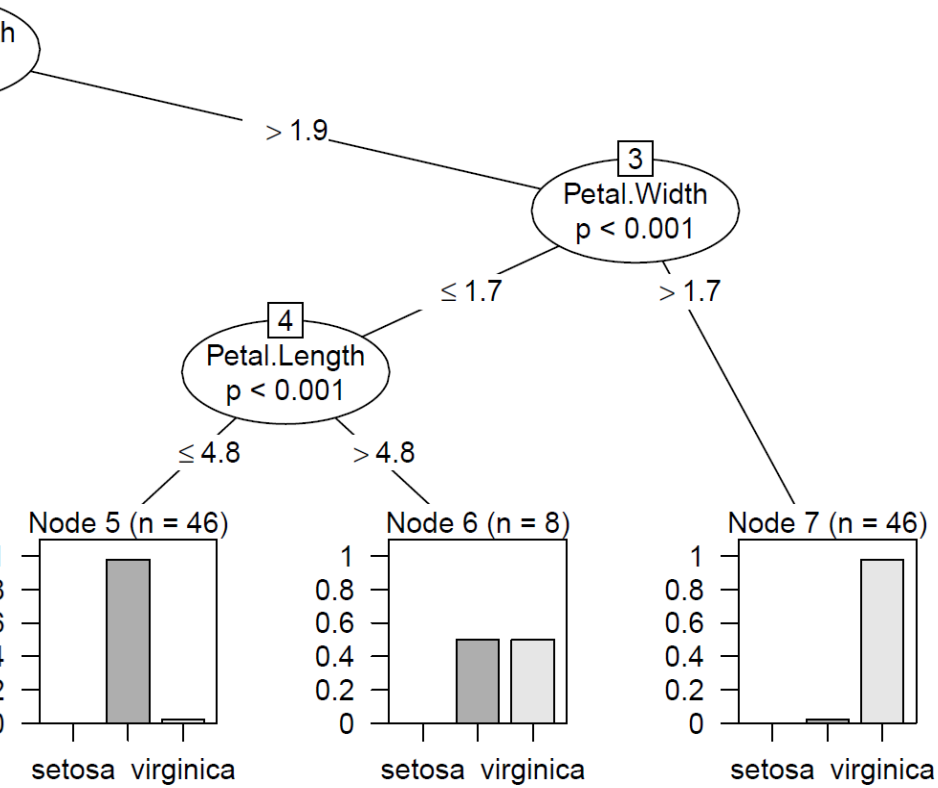


Error : 0.58 CV Error : 1.07 SE : 0.0915

EPFL

Implementation in R

partykit::ctree (Conditional Inference Trees)



- Combines recursive binary partitioning with permutation tests (Bonferroni-adjusted p-value for each node)
- Stops when no significant associations (regression relationship) between any predictor and the response can be found (no tree pruning necessary)
- Can handle weighted predictors

But see also: rpart

Advantages of CART methods

- powerful tool for *data mining*
- easy to build and interpret decision trees
- robust and flexible technique
  - All sorts of variables (binary, multi-class, ordered)
  - Accepts missing data
  - No assumptions on the variable distribution and relationships

BUT: CART suffer from inaccuracy when predicting new data!

=> Random forest classifiers

Random Forest Classifiers

- Use a large number of decision trees (=forest) each built with a different, random sub-sample of the dataset (bootstrapping) and only a subset of explanatory variables

- Uses all decision trees for making a prediction. Criteria: the most frequent pattern (majority vote).

Random Forest Classifiers

- Random Forests are generated using bootstrapping
  Bootstrap: resampling the dataset with replacement for constructing the decision tree
- For each node, only a subset of explanatory variables are taken randomly
  - iterate using different numbers of explanatory variables
- Generate many trees -> provides flexibility for classifying new data
- run data through all trees and measure the outcomes, aggregate all outcomes to make a decision (prediction).
- **B**ootstrap and **agg**regate => “bagging”
- use out-of-bag samples to evaluate random forest classifier performance

example: Use V1, V2, BMI, Sport to predict health state

Health state	V1	V2	BMI	Sport?
Ill	0	1	25	No
Well	1	1	45	No
Well	1	0	65	Yes
Ill	0	1	32	Yes

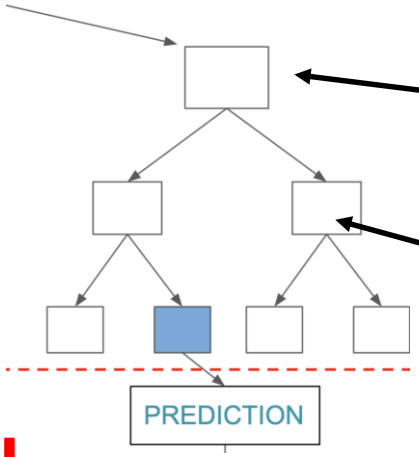
bootstrap samples

Health state	V1	V2	BMI	Sport?
Ill	0	1	25	No
Ill	0	1	32	Yes
Well	1	0	65	Yes
Well	1	0	65	Yes

select n explanatory variables

Health state	V1	BMI
Ill	0	25
Ill	0	32
Well	1	
Well	1	

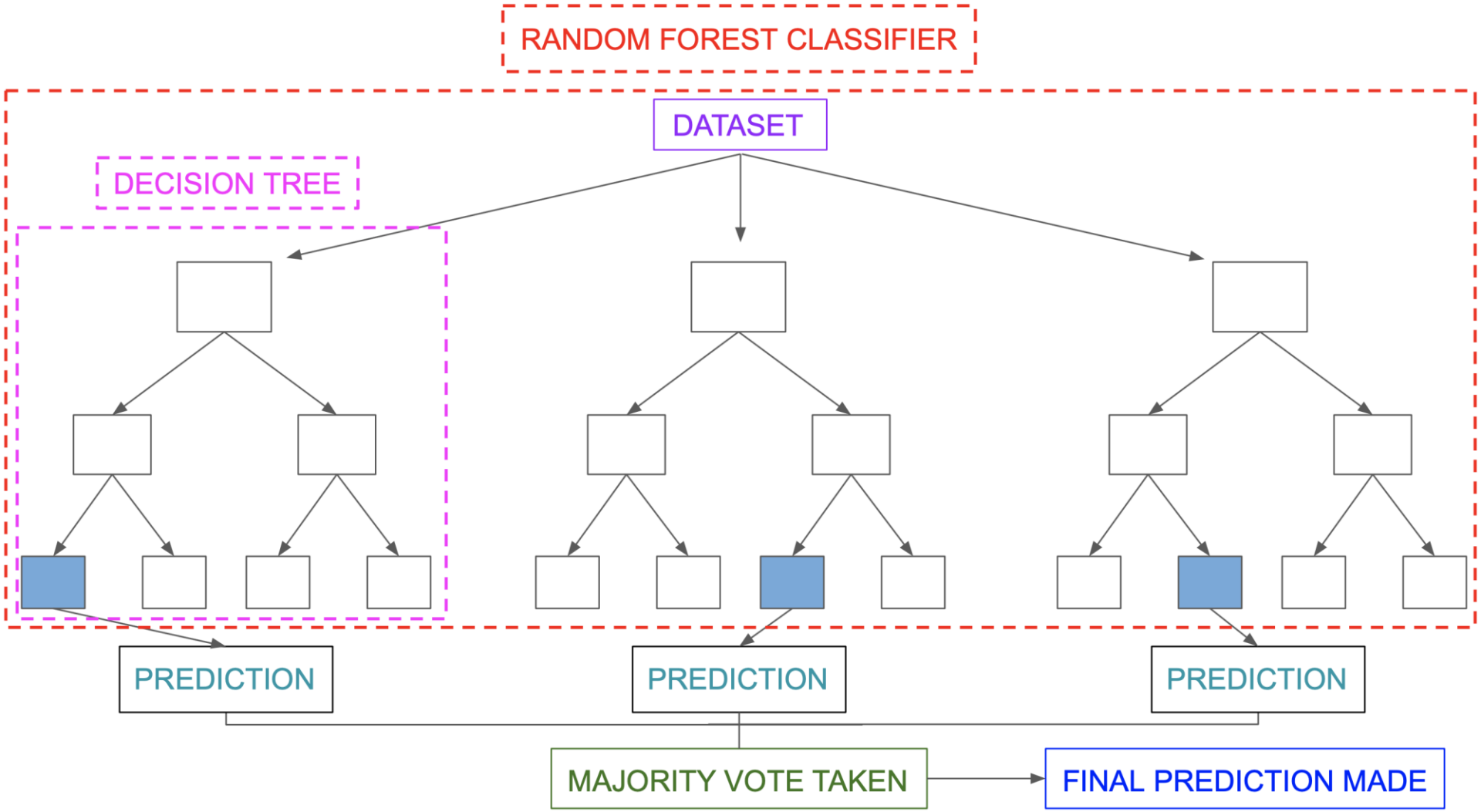
Health state	V2	Sport?
Ill	1	No
Ill	1	Yes
Well	0	Yes



predict first node

predict second node

EPFL


RANDOM FOREST CLASSIFIER
DATASET
DECISION TREE
PREDICTION
PREDICTION
PREDICTION
MAJORITY VOTE TAKEN
FINAL PREDICTION MADE

EPFL

Out-of-bag samples (OOB)

Health state	V1	V2	BMI	Sport?
Ill	0	1	25	No
Well	1	1	45	No
Well	1	0	65	Yes
Ill	0	1	32	Yes

Health state	V1	V2	BMI	Sport?
Ill	0	1	25	No
Ill	0	1	32	Yes
Well	1	0	65	Yes
Well	1	0	65	Yes

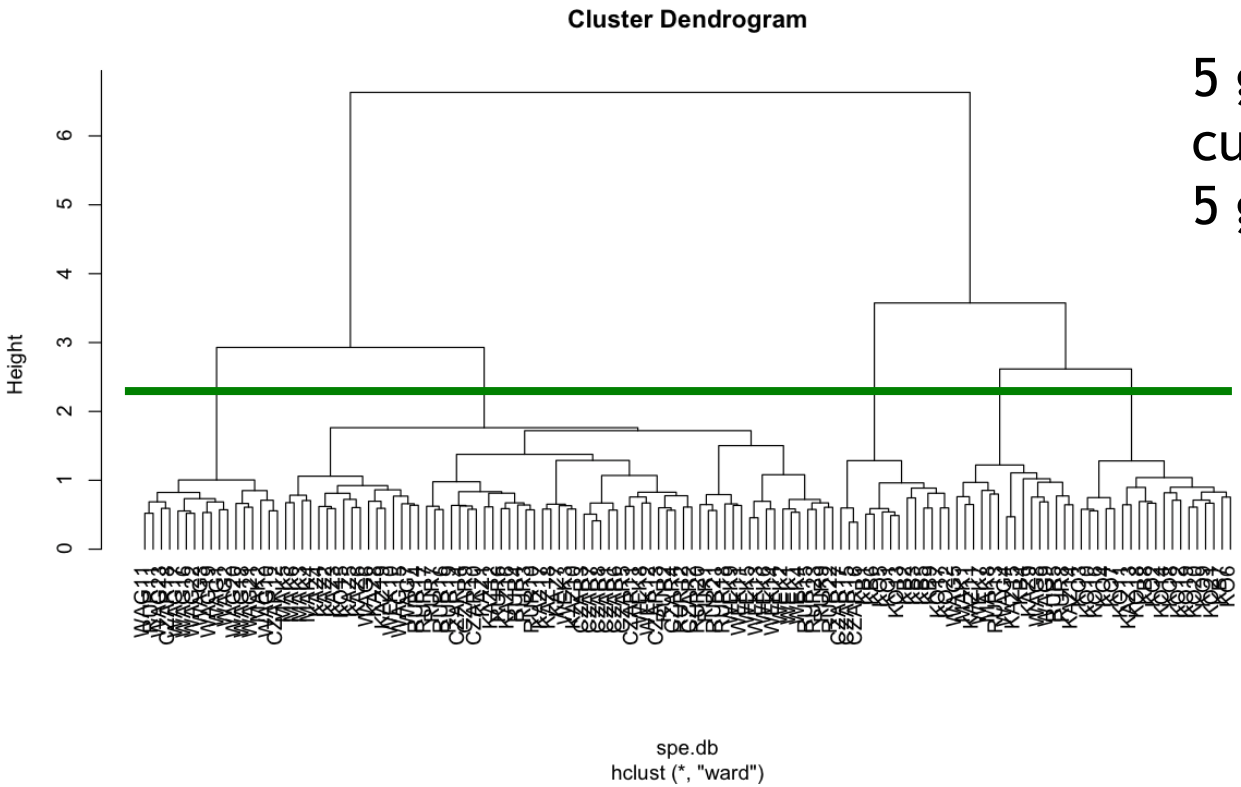
Typically ~1/3 of samples are not included by bootstrapping into random forest classifier tree generation

$\Rightarrow$ Use these samples to evaluate random forest classifier performance: (OOB error = fraction correctly classified OOB samples)

Example: microbes

Dataset: microbes and environmental variables (nutrient availability) in a wetland

132 observations, 70 species (spe) and 15 environmental variables (env)

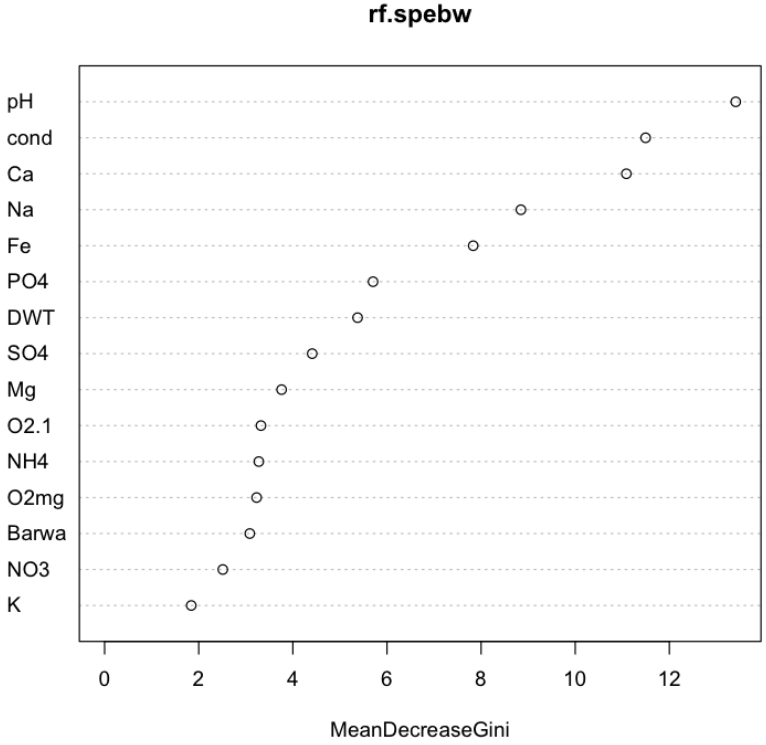


5 groups => building a cutree object with k = 5 groups (spebw.g)

Example: microbes

rf.env = randomForest(spebw.g~., env, ntree=500, mtry=10, proximity=T)

ntree: number of decision trees
mtry: number of variables selected for each node in the tree
proximity: keep estimates of closeness of pairs of samples

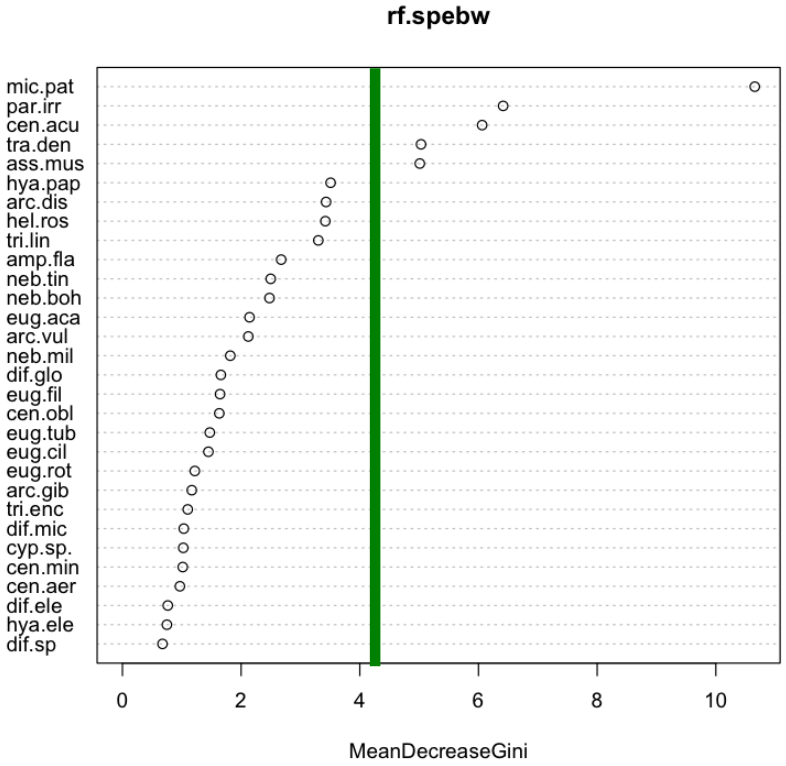


Variable Importance:
Shows the most important environmental variabels.

OOB estimate of error rate: 27.27%
(36 missclassified out of 132 samples)

Example: microbes

rf.spe = randomForest(spebw.g~., spe, ntree=500, mtry=10, proximity=T)



Variable Importance:
Shows the most important species.

OOB estimate of error rate: 9.85%
(13 misclassified out of 132 samples)

EPFL

Classification and regression trees

+ Useful for qualitative and quantitative prediction
+ Fast
+ accepts missing data
+ **Visual discrimination**
+ **Tree easy to interpret**

+/- Pruning of the tree

- No option for selecting variables
- Requires a low CV Error for prediction

Random Forest

+ Useful for qualitative and quantitative prediction
+ **Provides variable importance (ranking)**
+ **Robust (bootstraping)**
+ accepts missing data (in explanatory variables and in new predictions)

- individual decision trees are not directly interpretable (‘Black Box’)
- sometimes slow

nature climate change

LETTERS

https://doi.org/10.1038/s41558-018-0393-5

Widespread loss of lake ice around the Northern Hemisphere in a warming world

Sapna Sharma[1,11]*, Kevin Blagrave[1,11], John J. Magnuson[2,11], Catherine M. O'Reilly[3,11], Samantha Oliver[4], Ryan D. Batt[5], Madeline R. Magee[2,6], Dietmar Straile[7], Gesa A. Weyhenmeyer[8], Luke Winslow[9] and R. Iestyn Woolway[10]

Ice provides a range of ecosystem services—including fish harvest[1], cultural traditions[2], transportation[3], recreation[4] and regulation of the hydrological cycle[5]—to more than half of the world's 117 million lakes. One of the earliest observed impacts of climatic warming has been the loss of freshwater ice[6], with corresponding climatic and ecological consequences[7]. However, while trends in ice cover phenology have been widely documented[2,6,8,9], a comprehensive large-scale assessment of lake ice loss is absent. Here, using observations from 513 lakes around the Northern Hemisphere, we identify lakes vulnerable to ice-free winters. Our analyses reveal the importance of air temperature, lake depth, elevation and shoreline complexity in governing ice cover. We estimate that 14,800 lakes currently experience intermittent winter ice cover, increasing to 35,300 and 230,400 at 2 and 8 °C, respectively, and impacting up to 394 and 656 million people. Our study illustrates that an extensive loss of lake ice will occur within the next generation, stressing the importance of climate mitigation strategies to preserve ecosystem structure and function, as well as local winter cultural heritage.

in some winters[2,11]. This transitional period from annual winter ice to permanent loss of ice cover may endure for decades[2]. The factors influencing whether or not ice forms are well known; previous research has indicated that air temperature, wind speed, and lake size are essential components to ensure that vertical heat transfer is sufficient to cool surface water temperatures to 0 °C[12,13]. Precipitation[12], snow cover[14], cloud cover, solar radiation[14], distance to coastline[15] and regional differences[7,16] can govern the timing of ice formation and ice growth during the winter season. However, previous research has not identified how the interactions between features such as climate and lake shape (area and depth) will dictate when and where the threat of lake ice loss is greatest. We provide the first global estimate of how many lakes are likely to lose annual winter ice cover as the climate warms.

We used updated lake ice cover records for 346 lakes in North America, 136 lakes in Europe, and 32 lakes in Asia to evaluate the threat of lake ice loss[17] (Supplementary Fig. 1). Lakes were designated as annual or intermittent winter ice-covered lakes. Annual ice-covered lakes experienced complete ice cover every winter, whereas intermittent ice-covered lakes had one or more winters