Joint work with:

Rong Xia
University of Michigan

Advances in Random Forests

"A Random Forest Guided Tour" (2015)

Gérard Biau, Erwan Scornet

Instead…

- Leo Breiman
- Introduction to trees and random forests
- Open questions
  - randomForest or cforest?
  - classification with unbalanced classes

Prediction

x ⟶ black box ⟶ y

Goal: accurately predict the response (y) for new predictors (x) using data

And get reliable information about the mechanism in the black box

y categorical → “classification”

y continuous → “regression”

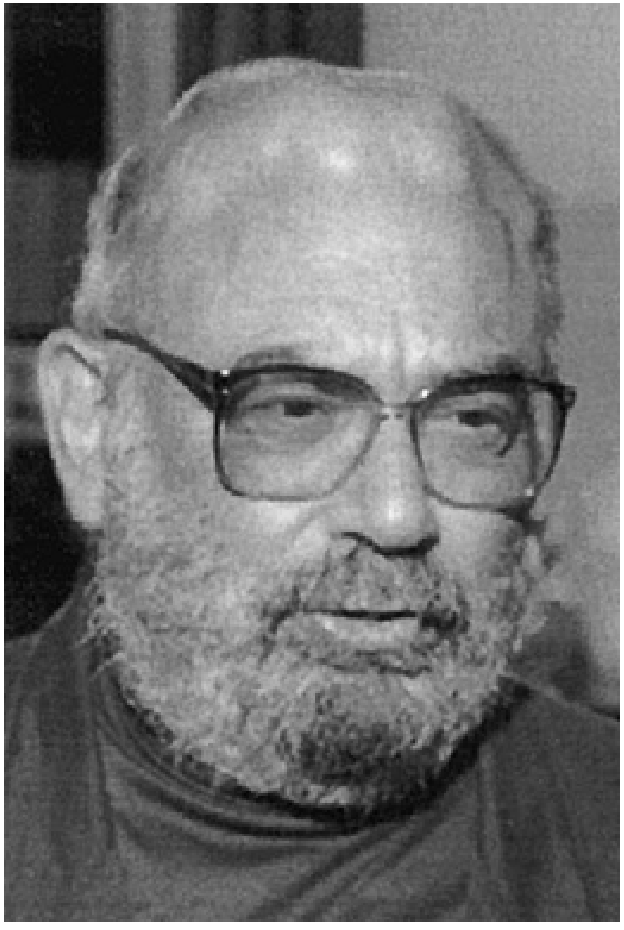
An Important Principle:

"The better the model fits the data,
the more sound the inferences about the black box are"

Breiman (2003)

“If all you have is a hammer, every problem looks like a nail” (Breiman)

Data Wizards

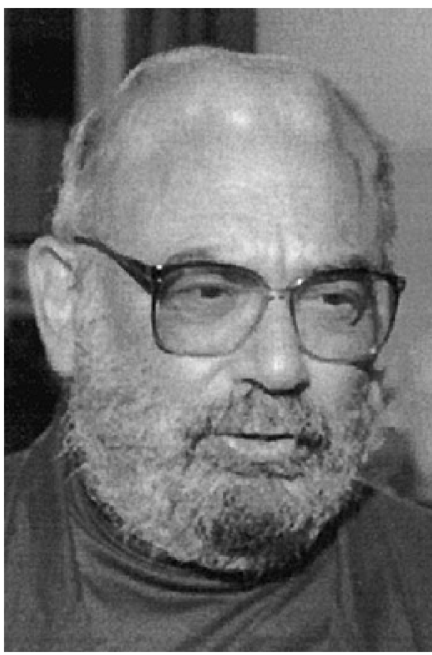
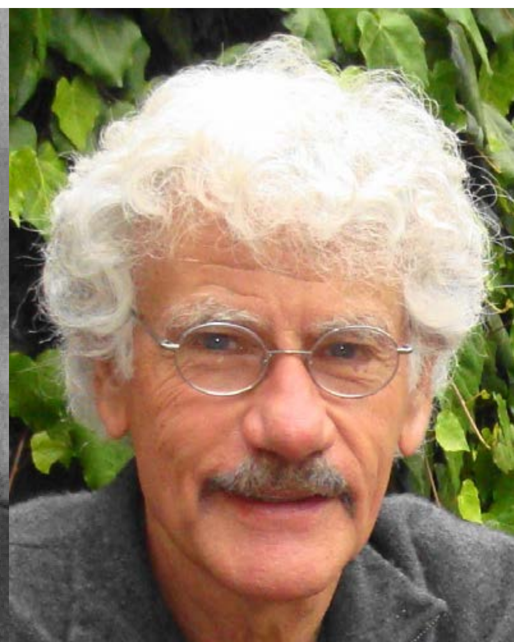


“Wizardry in pursuit of the goal of gathering and analyzing data to answer interesting questions” (Breiman, 2003)

Classification and Regression Trees

Pioneers:

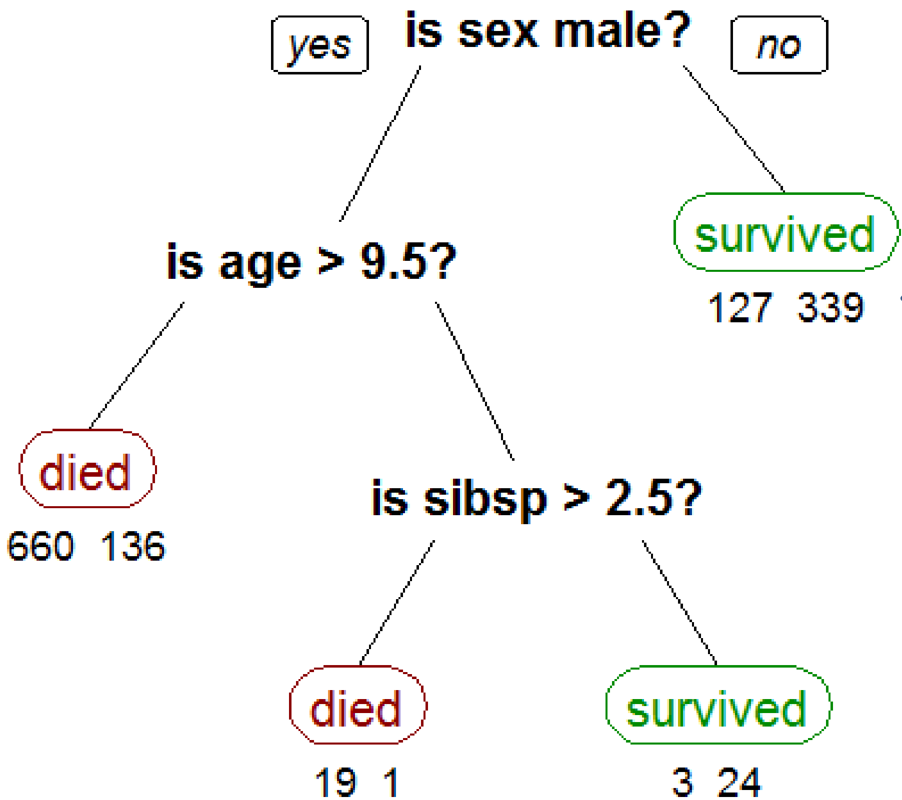
- Morgan and Sonquist (1963)
- Breiman, Friedman, Olshen, Stone (1984) ***CART***
- Quinlan (1993) *C4.5*

A Classification Tree

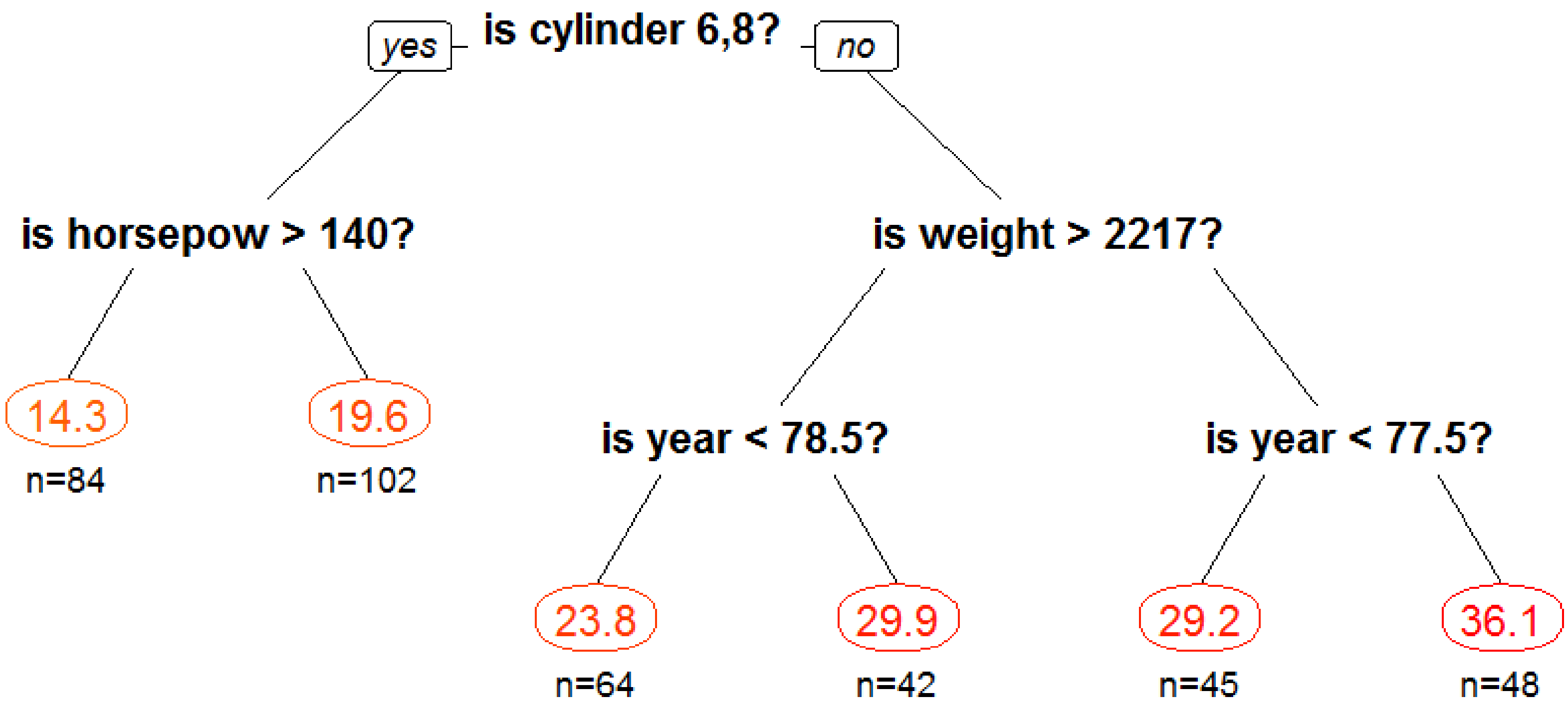
“yes”
go left

660 died
136 survived



127 died
339 survived

A Regression Tree



Advantages of Trees

- Work for both classification and regression
- Handle categorical predictors naturally
- No formal distributional assumptions
- Can handle highly non-linear interactions and classification boundaries
- Handle missing values in the variables

Disadvantages: inaccuracy, instability

Random Forests

Take a bootstrap sample from the data
Fit a classification or regression tree } Repeat

At each node:

1. Select ***mtry*** variables **at random** out of all ***M*** possible variables (independently at each node)
2. Find the best split on the selected ***mtry*** variables
3. Grow the trees big

Combine by

- voting (classification)
- averaging (regression)

Random Forests

Take a bootstrap sample from the data
Fit a classification or regression tree

} Repeat

At each node:

1. Select mtry variables at random out of all M possible variables (independently at each node)
2. Find the best split on the selected mtry variables
3. Grow the trees big

Combine by

- voting (classification)
- averaging (regression)

Random Forests

Idea: most of the trees are good for most of the data and make mistakes in different places

More formally (Breiman, 2001) the trees have

- high strength
- low correlation

Variable Importance

Two measures:

- Gini criterion
  – rough-and-ready
- Permutation importance
  – recommended

Advantages of Random Forests

- Usually (a lot) more accurate than trees
- Built-in estimates of accuracy
- Automatic variable selection
- Variable importance
- Work well “off the shelf”
- Handle “wide” data

Disadvantages of Random Forests

- Forests are inscrutable

Disadvantages of Random Forests

- Forests are inscrutable

- Bias in variable importance if categorical predictors have different numbers of levels and/or predictors are mixed categorical and continuous (Strobl et al. 2007, Boulesteix 2012) $\longrightarrow$ cforest

Biased Variable Importance

Strobl et al. (2007) report that when predictors have unequal scales it

> *"severely affects the reliability and interpretability of the variable importance measure"*

Gini Criterion for Classification Splits

- CART and Random Forests use Gini
- Gini is known to favor many-level categoricals and continuous variables over categoricals with only a few levels

Simulations

- 1000 trees in each forest
- 100 observations in the training set
- 1000 observations in an independent test set
- 100 repetitions
- replace = FALSE for cforest
- default parameters unless otherwise noted

Examples 1 and 2

$x1 \sim M(2)$

$x2 \sim M(2)$

$x3 \sim M(4)$

$x4 \sim M(10)$

$x5 \sim U(0, 1)$

$x6 \sim N(0, 1)$

Example 1: (main effect)

$y = \text{Bernoulli}(p)$

$p = .3$ if $x1 = 0$

$p = .7$ if $x1 = 1$

Example 2: (interaction)

$y = 1$ if $x1 = x2$

$y = 0$ otherwise

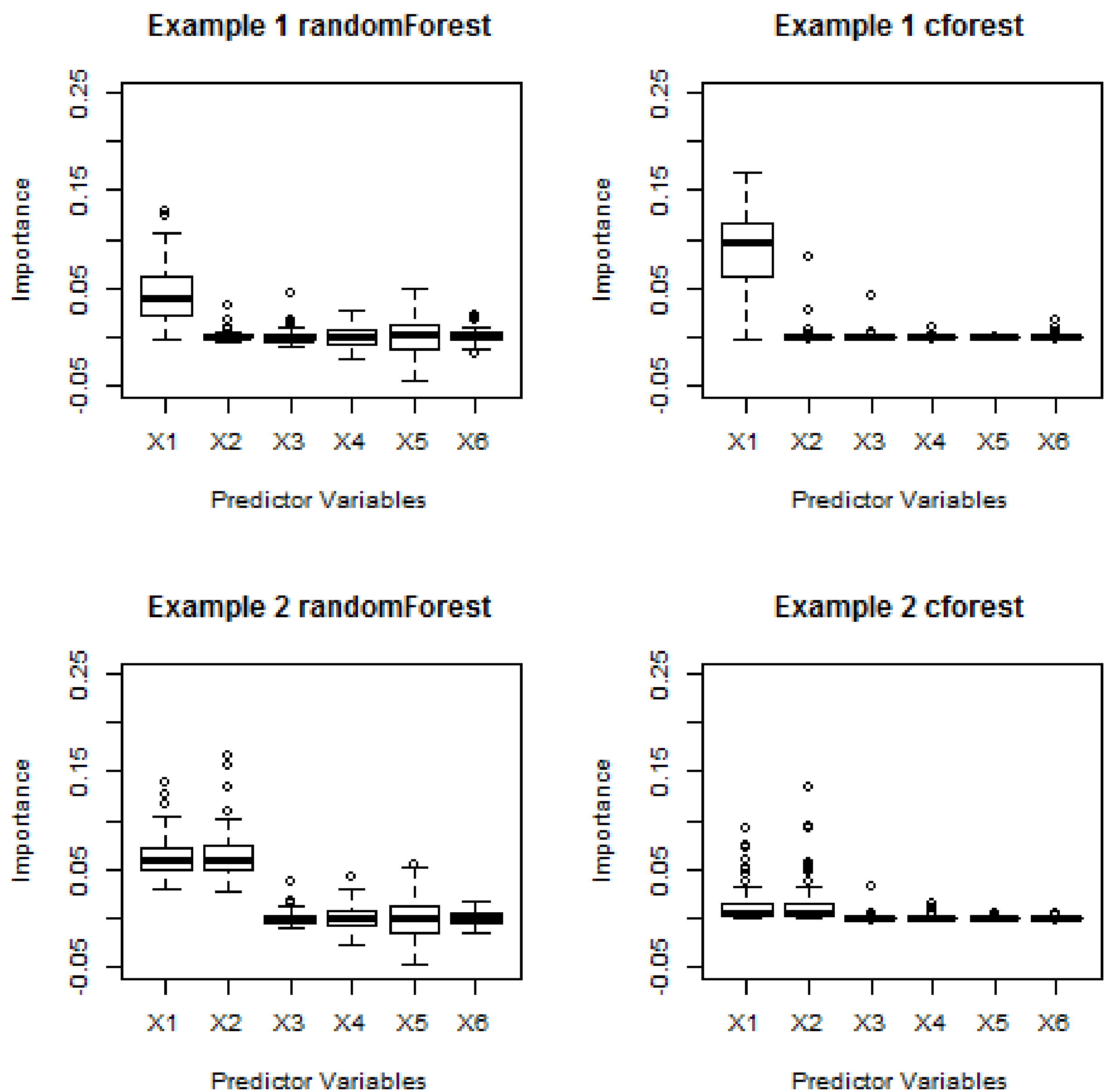
M is multinomial

% Error rates example 1

mtry	random forest	cforest	random forest	cforest
	mean		SE	
1	39.1	45.1	.4	.8
2	41.2	39.3	.4	.9
3	41.6	35.4	.4	.9
4	41.8	32.9	.4	.7
5	42.1	31.7	.4	.5

% Error rates example 2

mtry	random forest	cforest	random forest	cforest
	mean		SE	
1	17.7	49.8	0.5	0.2
2	24.8	48.7	0.5	0.5
3	36.1	47.1	0.6	0.9
4	40.3	44.8	0.5	1.2
5	42.2	41.6	0.5	1.6



Example 1 randomForest
Example 1 cforest
Example 2 randomForest
Example 2 cforest
Importance
0.25
0.15
0.05
-0.05
X1
X2
X3
X4
X5
X6
Predictor Variables

Examples 3 and 4

$x1, x2, \ldots, x6 \sim N(0, 1)$

Example 3: (main effect)

$y = 0$ if $x1 > 0$

$y = 1$ otherwise

Example 4: (interaction)

$y = 0$ if $x1 * x2 > 0$

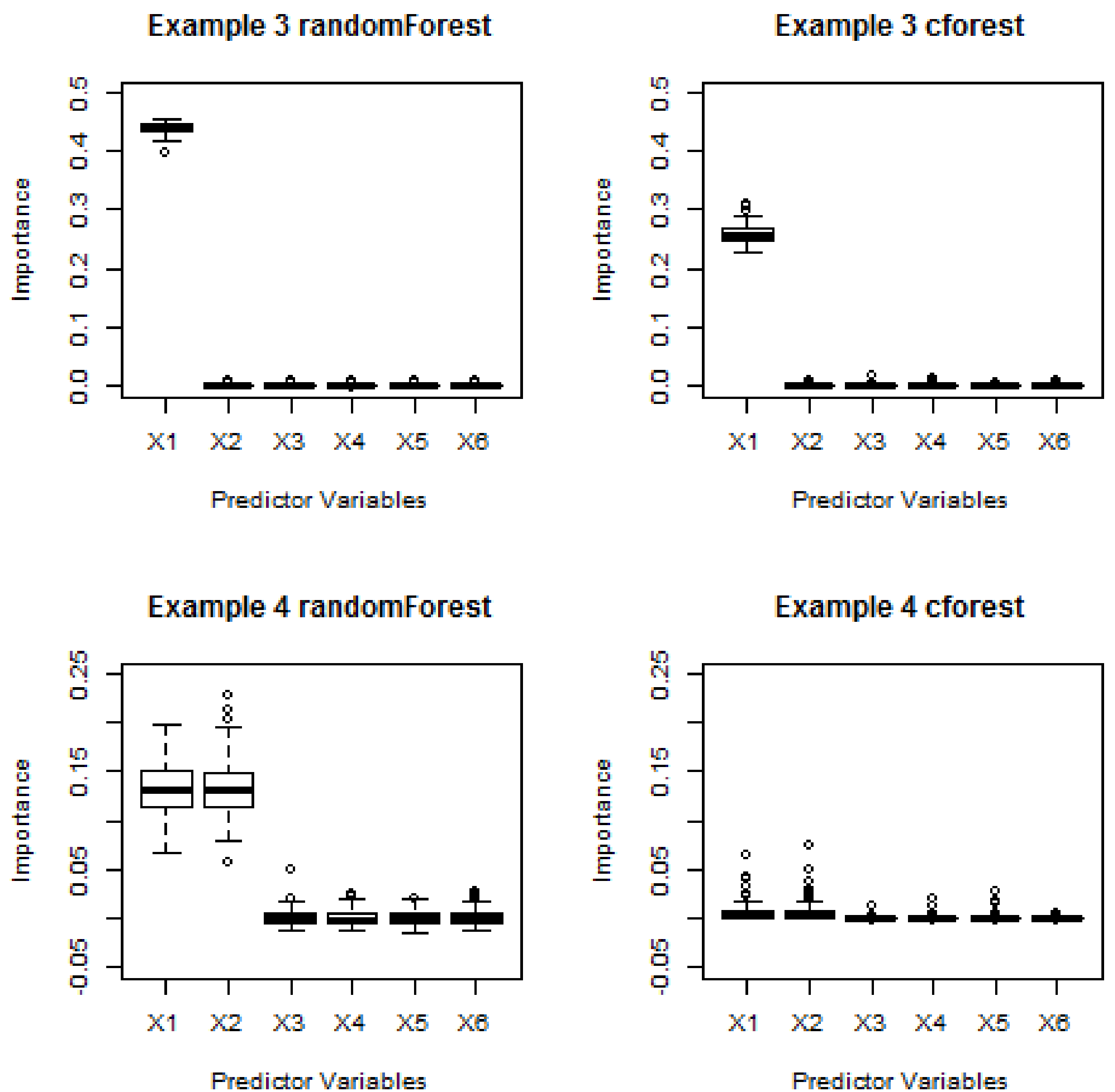
$y = 1$ otherwise

% Error rates example 3

mtry	random forest	cforest	random forest	cforest
	mean		SE	
1	.61	13.86	.05	2.10
2	.45	2.42	.04	.72
3	.43	1.21	.04	.13
4	.41	1.10	.04	.12
5	.41	1.06	.04	.12

% Error rates example 4

mtry	random forest	cforest	random forest	cforest
	mean		SE	
1	28.5	50.0	.4	.1
2	21.8	49.8	.5	.2
3	17.6	48.8	.6	.4
4	14.7	48.0	.7	.5
5	13.1	47.0	.7	.7


Example 3 randomForest
Importance
0.0
0.1
0.2
0.3
0.4
0.5
X1
X2
X3
X4
X5
X6
Predictor Variables
Example 3 cforest
Importance
0.0
0.1
0.2
0.3
0.4
0.5
X1
X2
X3
X4
X5
X6
Predictor Variables
Example 4 randomForest
Importance
-0.05
0.05
0.15
0.25
X1
X2
X3
X4
X5
X6
Predictor Variables
Example 4 cforest
Importance
-0.05
0.05
0.15
0.25
X1
X2
X3
X4
X5
X6
Predictor Variables

% correct, mtry = 3

Example	random forest	cforest
	% correct	
1	80	97
2	98	89
3	100	100
4	100	68

References

Leo Breiman, *Random Forests*, Machine Learning, 2001, 45:5-32

Anne-Laure Boulesteix et al., *Overview of Random Forest Methodology and Practical Guidance with Emphasis on Computational Biology and Bioinformatics*, WIREs Data Mining Knowl Disc 2012, 2:493-507

Carolin Strobl et al., *Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a solution,* BMC Bioinformatics 2007, 8:25