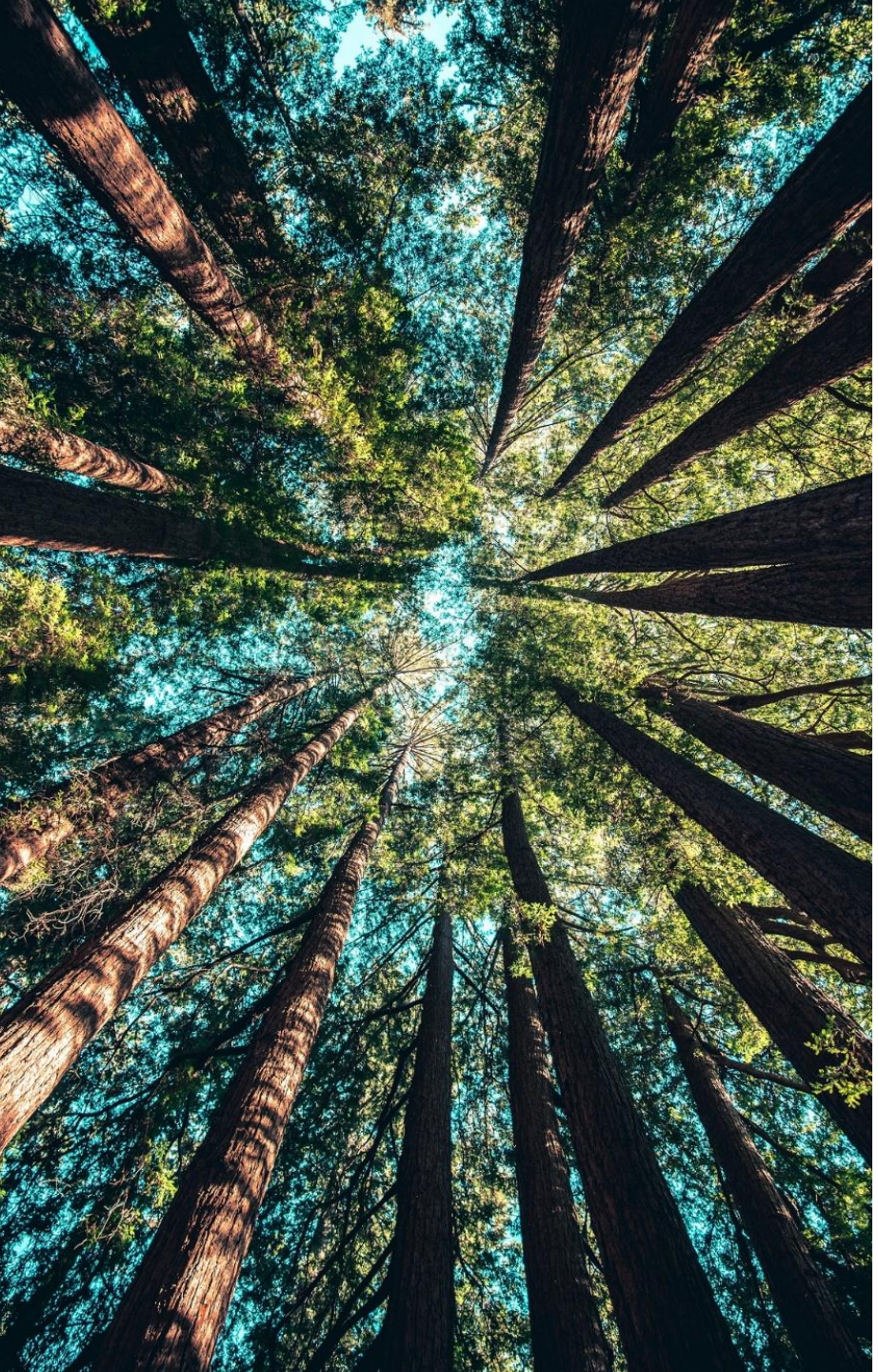


EVOLUTIONARY FOREST

An example of classification for celiac desease

(R)evolutionary T(h)ree

AGENDA

- Decision Tree
- Evolutionary Forest
- Comparison of Tree-Based Methods
  - Simulated Data
  - Celiac Dataset
- Conclusion

DECISION TREE

GOAL

Explore non linear relationship and predict Y

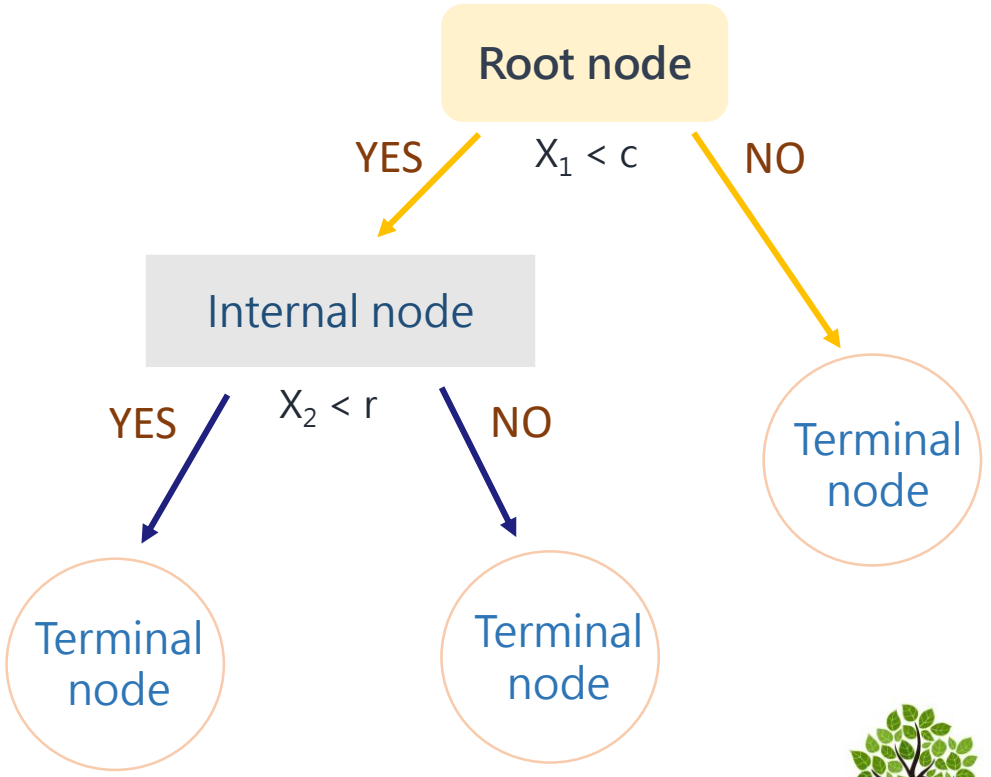
HOW

1) Partitioning the space $X=(X_1,\ldots,X_p)$ into M regions $R_m(m=1,\ldots,M)$
2) Fitting a model within every region $Y|\ X \in R_m$

SPLITTING RULE

Top down greedy approach

→ **recursive binary splitting**



DIFFERENT TREES

REGRESSION TREE	CLASSIFICATION TREE
Quantitative variable response	Qualitative variable response
Mean	Mode
RSS	Error rate: • Misclassification error rate • Gini index • Cross entropy

ALGORITHMS

- Local optimization: CART
- Global optimization: Random Forest
- Stochastic optimization: Evolutionary Forest

EVOLUTIONARY FOREST

GOAL

Maintain the simple tree structure and offer better performance (in terms of predictive accuracy and/or complexity) than commonly-used recursive partitioning algorithms.

Evolutionary algorithms are inspired by natural Darwinian evolution employing concepts such as inheritance, mutation, natural selection and crossing over.

ALGORITHM

1) Initialise the population: let $\theta_n = (v_n, s_n) \in \Theta$ a single tree to be initialised
2) At each iteration and for each tree selected:
   - change the tree through 5 **variation operators**
   - evaluate the new tree through the **evaluation function**

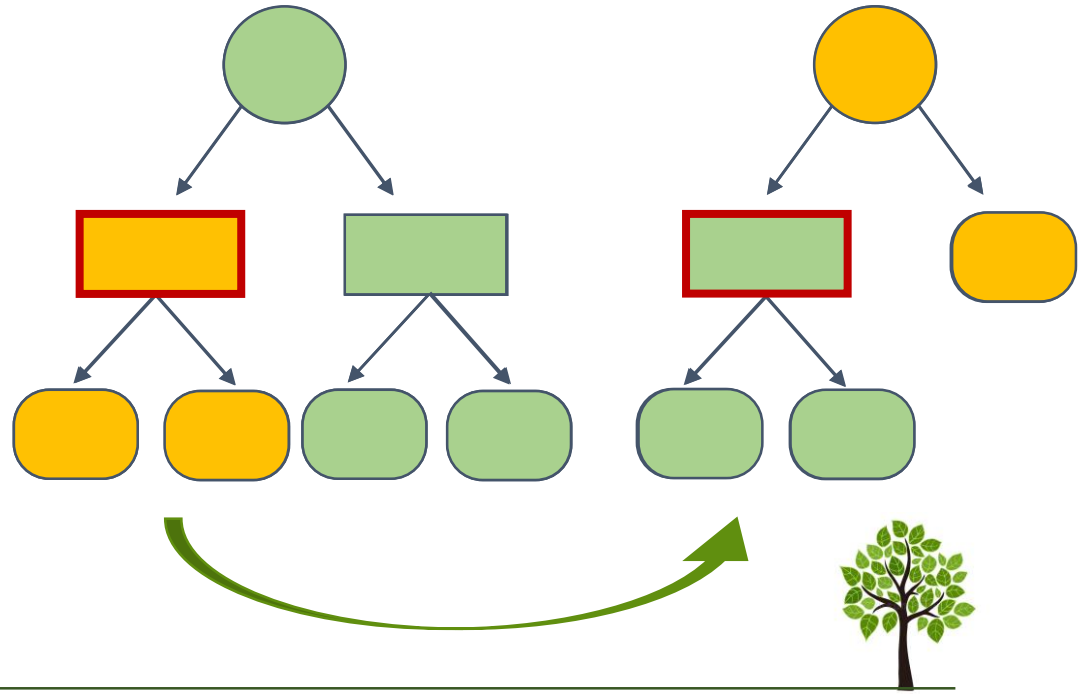
VARIATION OPERATORS

Mutation operators:

- **Split**: select randomly terminal node and assigns randomly generated splitting rule
- **Prune**: select randomly an internal node and prune it into a terminal one
- **Major split rule mutation**: chooses randomly an internal node, changes splitting rule (with splitting variables) and split point
- **Minor split rule mutation**: changes the split point only through the splitting rule

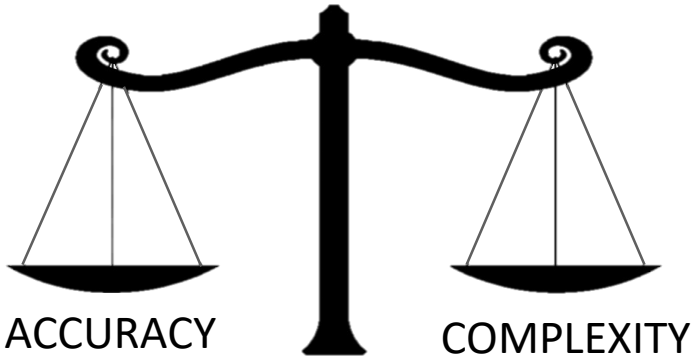
Crossover operator:

Exchanges subtrees between two trees randomly



EVALUATION FUNCTION

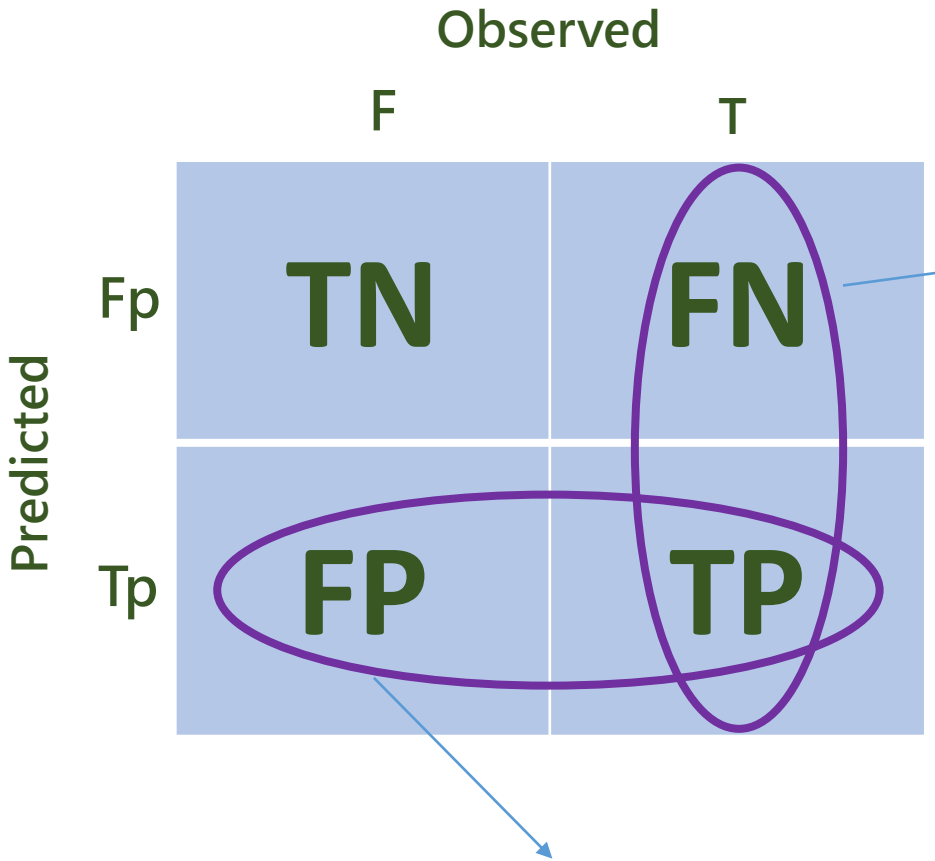
EvalFun= Loss Function + Comp(θ)



$$EvalFun_{class} = 2N \cdot MC\big(f(X,\theta)\big) + \alpha \cdot M \cdot logN$$
$$EvalFun_{reg} = 2N \cdot MSE\big(f(X,\theta)\big) + 4\alpha \cdot (M+1) \cdot logN$$

max accuracy
min complexity

METRICS TO COMPARE ALGORITHMS

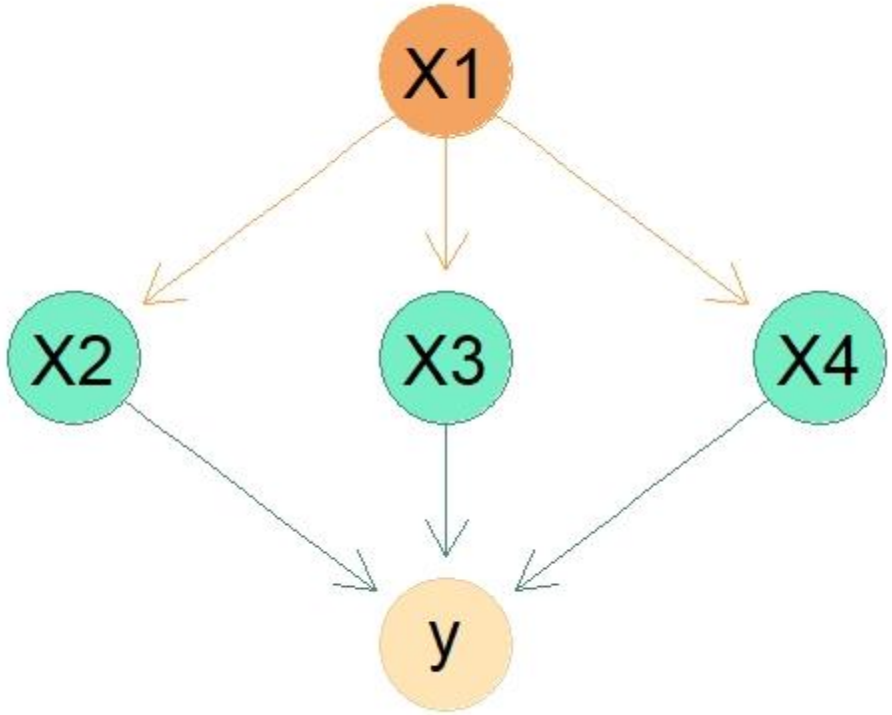


Sensitivity (Recall) = TP/(TP+FN)

$$\text{F-measure} = \frac{2*(\text{PRECISION}*\text{SENSITIVITY})}{\text{PRECISION}+\text{SENSITIVITY}}$$

Positive predicted value (Precision) = TP/(TP+FP)

SIMULATED DATA



SETTING

- Diamond DAG
- n=500
- Y ~ binomial

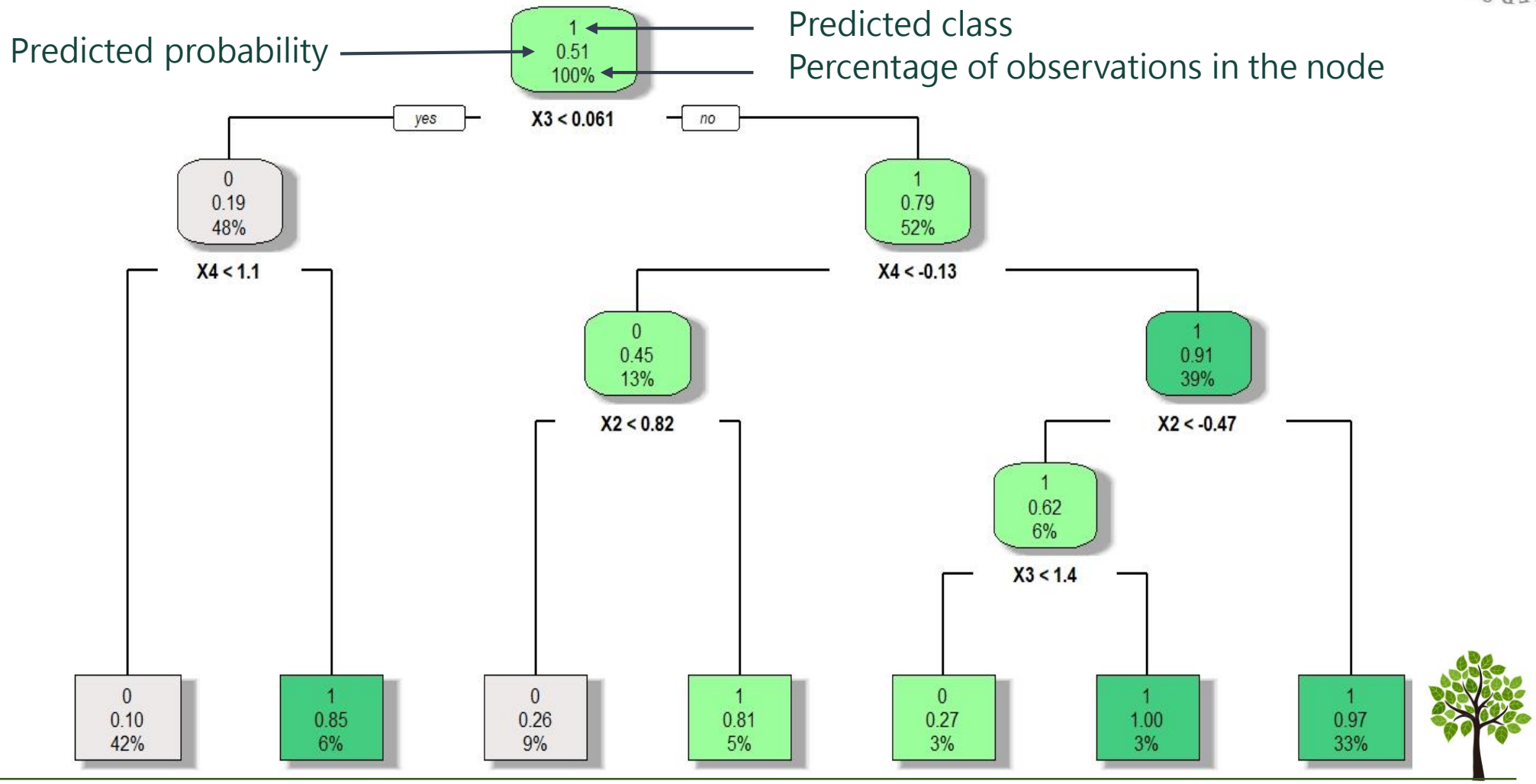
ANALYSIS

- Logit Regression
- CART Model
- Evolutionary Forest
- Random Forest

LOGIT

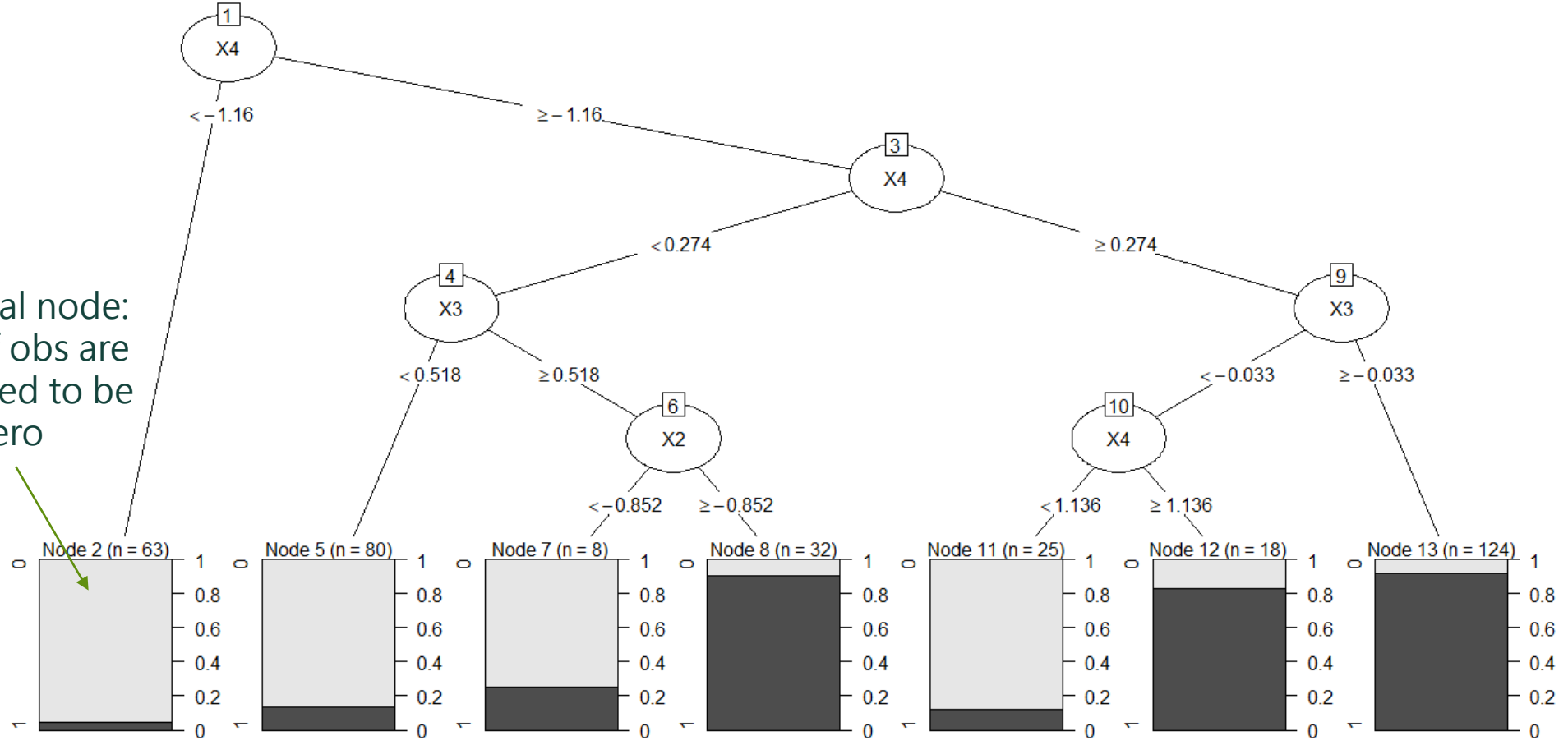
```
glm(formula = y ~ X, family = binomial(link = "logit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1740  -0.4331  -0.0348   0.4244   2.7455

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.2891     0.1443  -2.004   0.0451 *
XX1          -0.2995     0.3015  -0.993   0.3206
XX2           0.8516     0.1712   4.976 6.50e-07 ***
XX3           1.2472     0.1731   7.207 5.71e-13 ***
XX4           1.0166     0.1612   6.306 2.87e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

CART

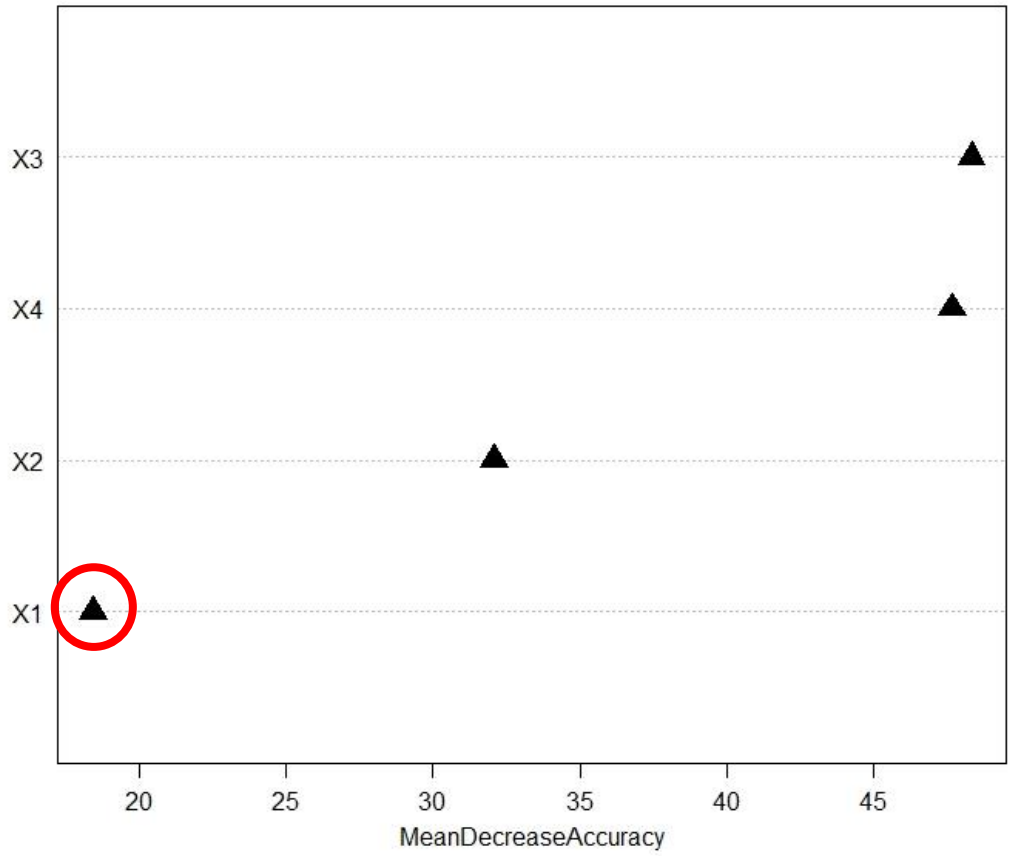

Predicted probability
1
0.51
100%
Predicted class
Percentage of observations in the node
yes
X3 < 0.061
no
0
0.19
48%
1
0.79
52%
X4 < 1.1
X4 < -0.13
0
0.45
13%
1
0.91
39%
X2 < 0.82
X2 < -0.47
1
0.62
6%
X3 < 1.4
0
0.10
42%
1
0.85
6%
0
0.26
9%
1
0.81
5%
0
0.27
3%
1
1.00
3%
1
0.97
33%

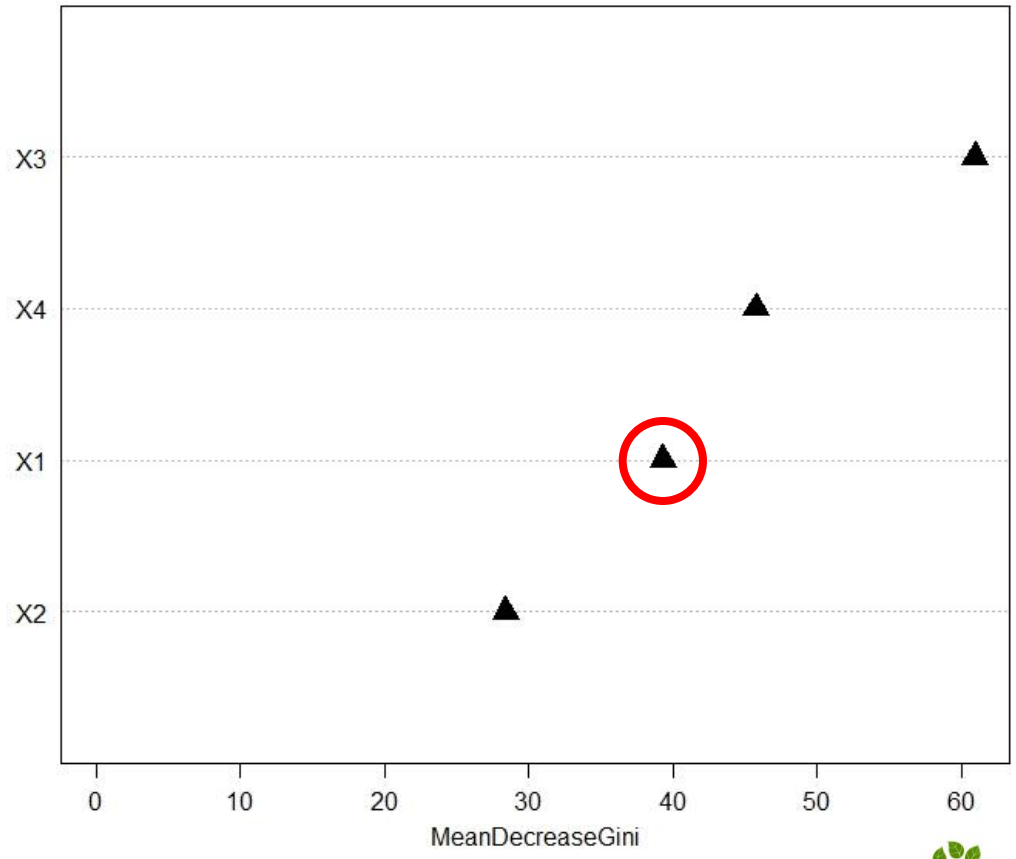
EVOLUTIONARY FOREST

Terminal node: 95% of obs are predicted to be zero



RANDOM FOREST





Variable Importance

COMPARISON

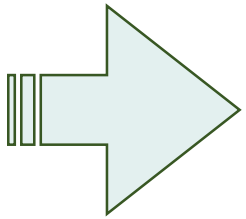
	CART	EVTREE	RANDOM FOREST
N. terminal nodes	7	7	
Misclassification	0.18	0.15	0.15
Evaluation function	223.50	190.62	

	CART	EVTREE
Positive predictive value	83%	91%
Sensitivity	75%	73%
F-measure	79%	81%

CELIAC DATASET

55 celiac desease patients
from 25 to 52 years old

40 healty subjects
from 23 to 42 years old



We join the datasets

Y
- 0 if healty
- 1 if cd patient

49 metabonomics as explanatory variables

EXPLORING THE DATA

TREE-BASED ANALYSIS

LOGISTIC REGRESSION
correlation among explanatory variables

REGULARIZATION

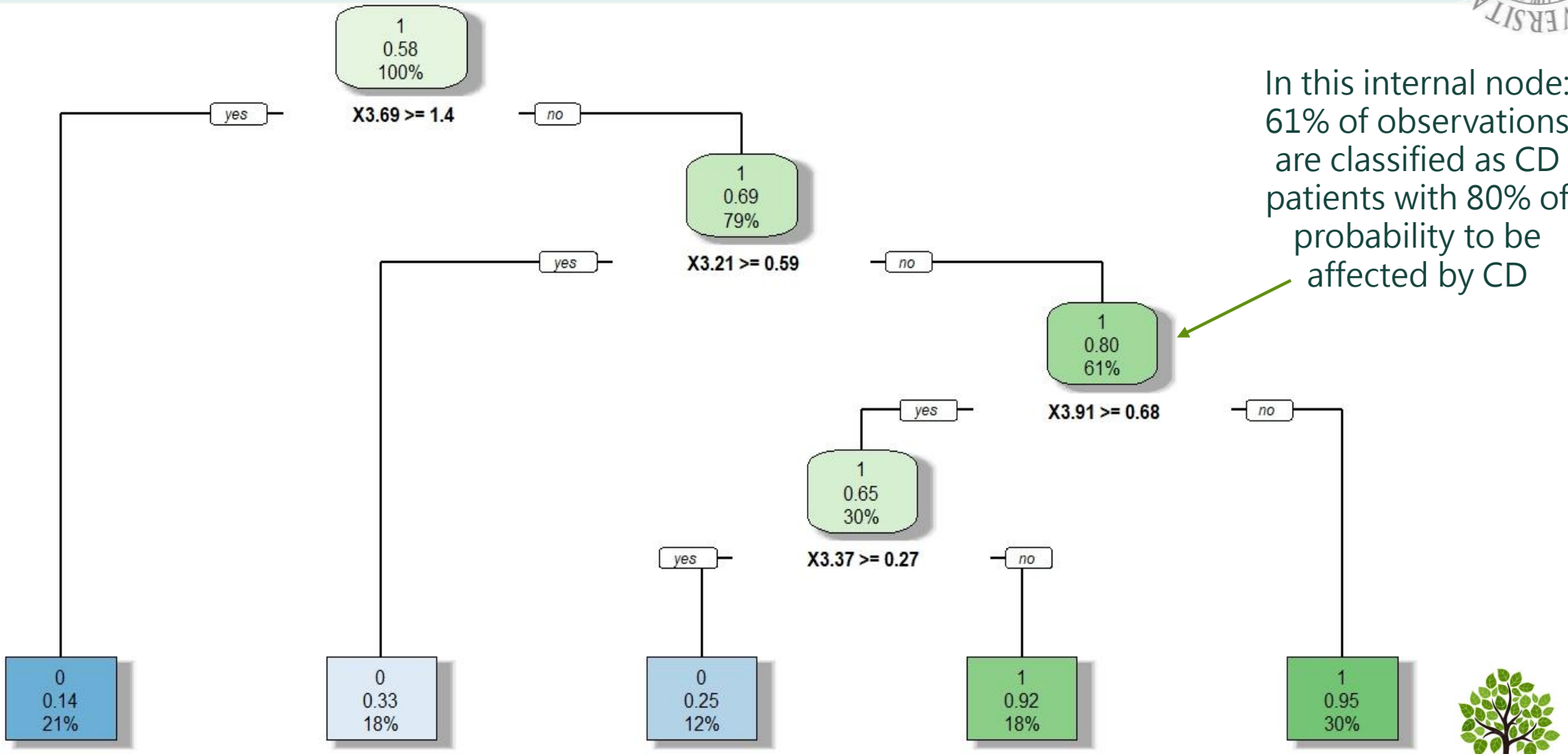
a) Lasso estimator
40 coefficients shrinked to 0
(From 49 to 9 explanatory variables)

b) Ridge estimator
All the coefficients shrinked towards zero
→ more stable estimates

SPLITTING DATASET
- 70% obs in training set – used to build trees
- 30% obs in test set – used to predict

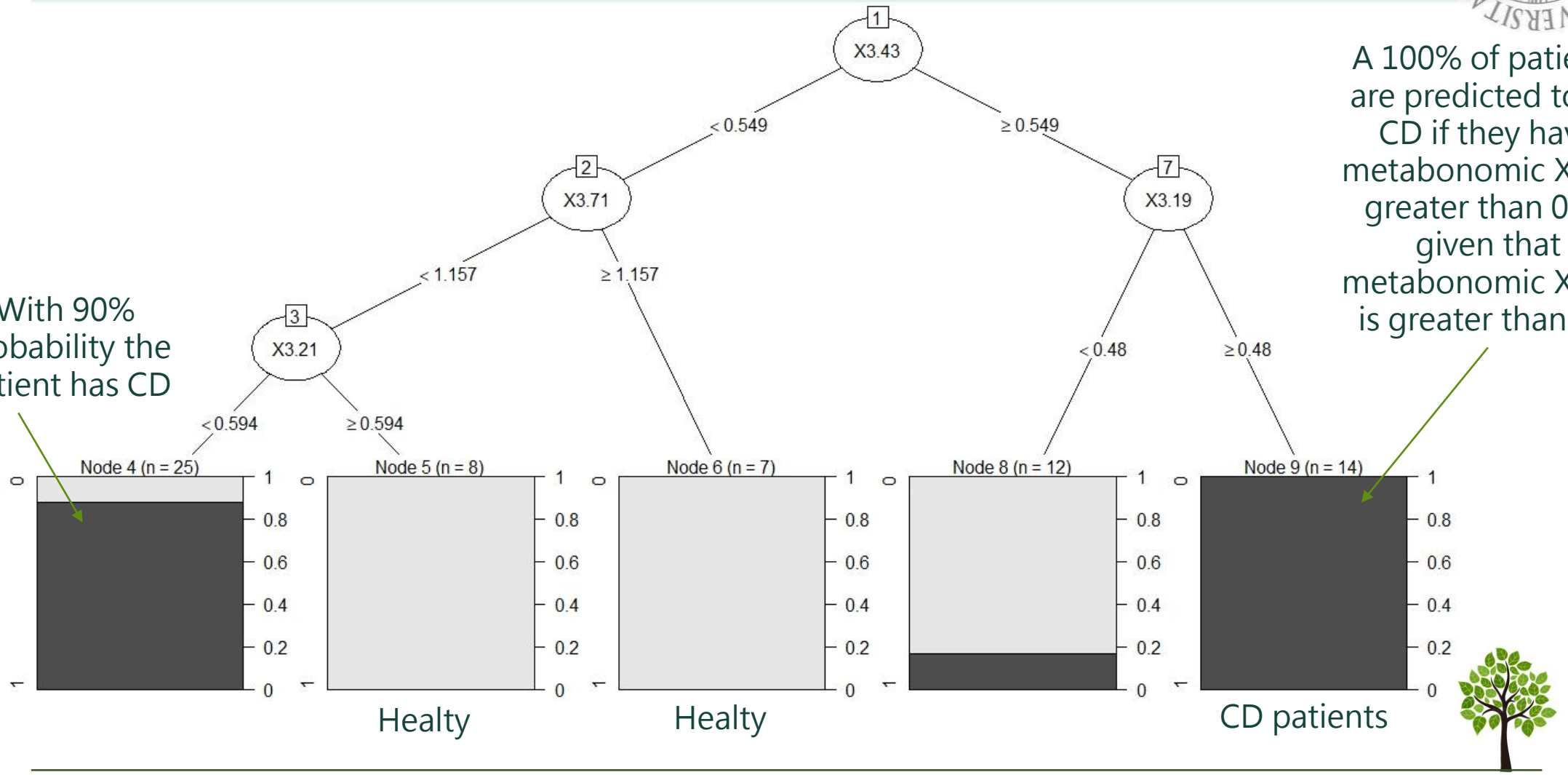
ALGORITHMS
- CART
- Evolutionary Forest
- Random Forest

CART



In this internal node: 61% of observations are classified as CD patients with 80% of probability to be affected by CD

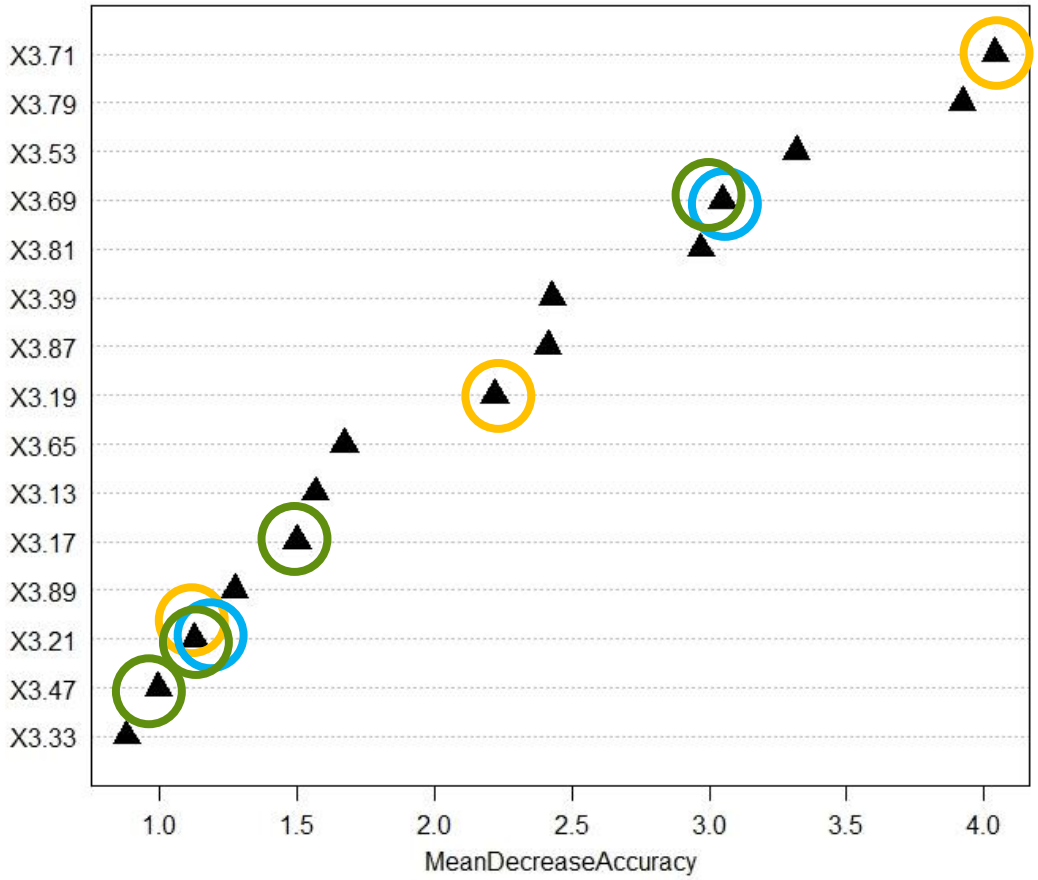
EVOLUTIONARY FOREST

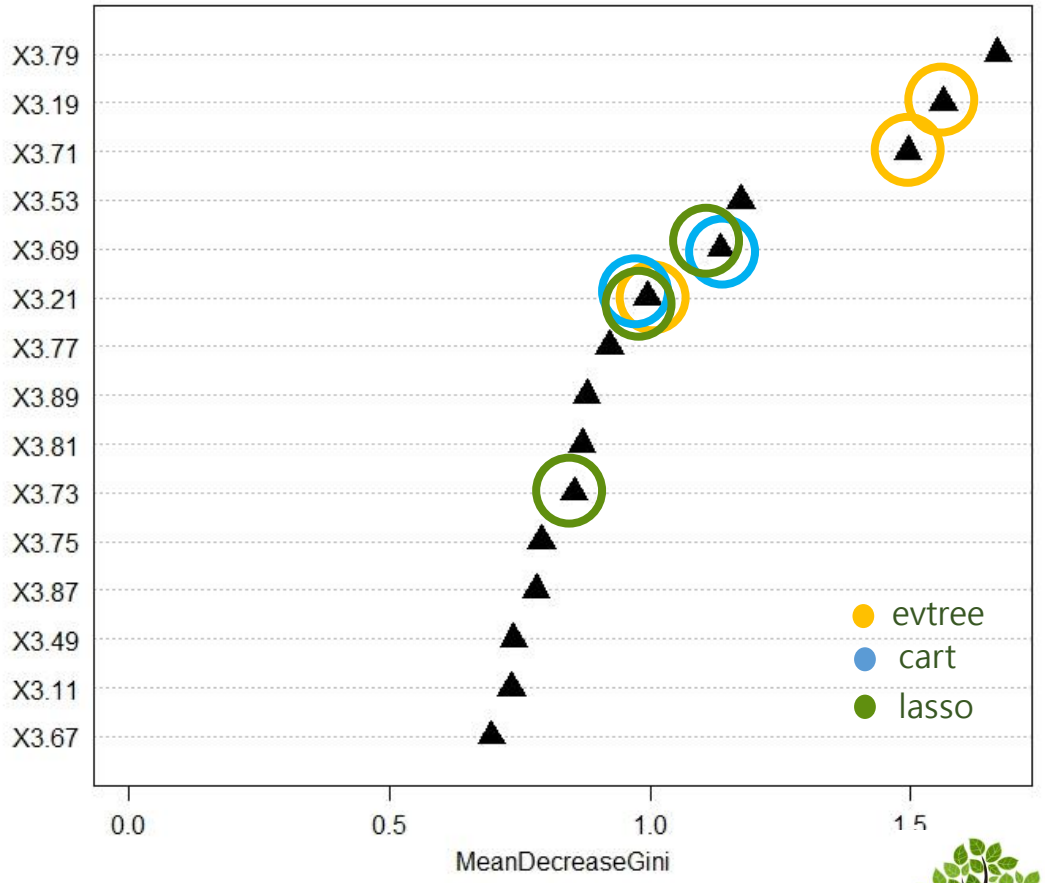


With 90% probability the patient has CD

A 100% of patients are predicted to be CD if they have metabonomic X3.19 greater than 0.48 given that metabonomic X3.43 is greater than 0.5

RANDOM FOREST



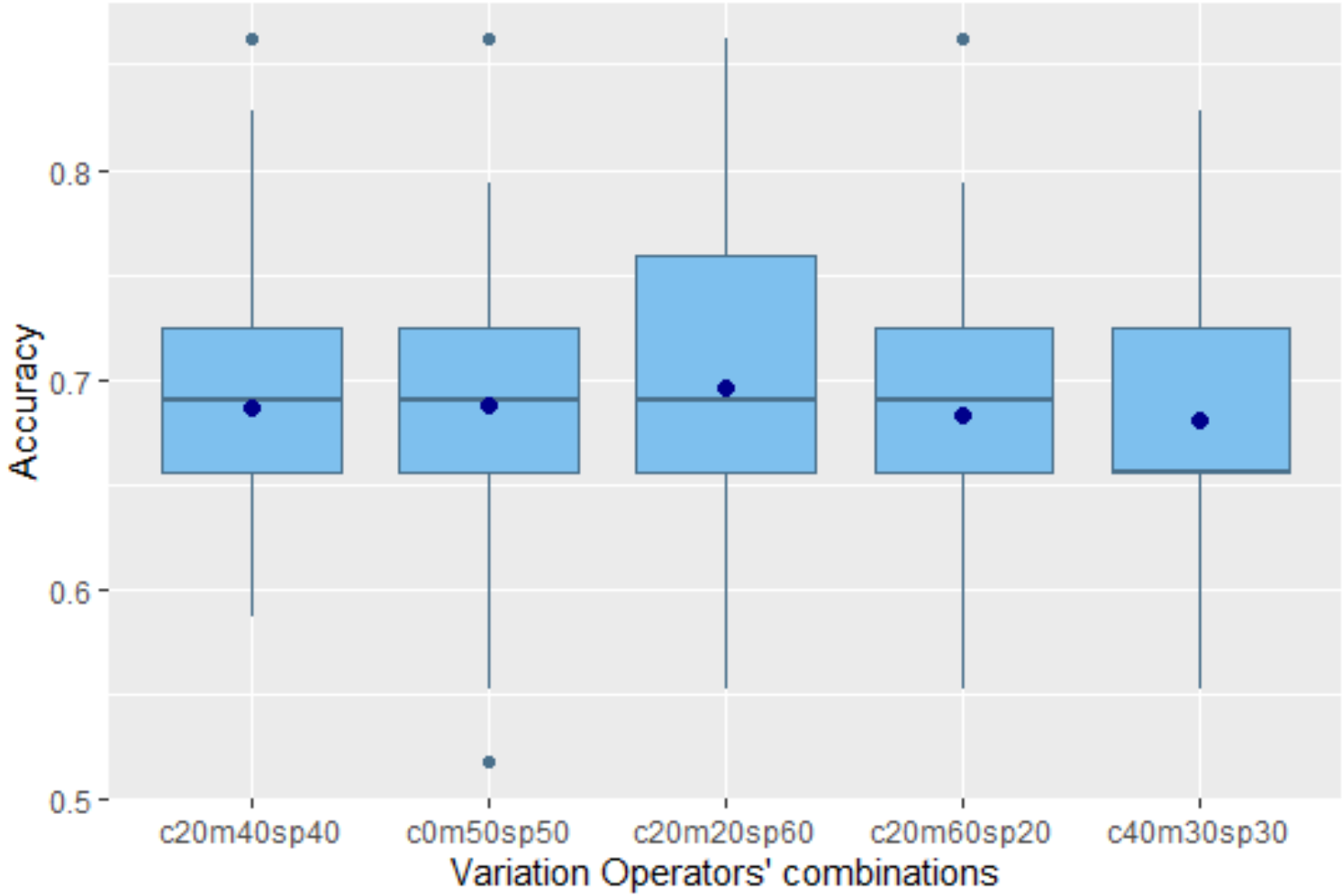


COMPARISON

	CART	EVTREE	RANDOM FOREST
N. terminal nodes	5	5	
Misclassification	0.31	0.31	0.24
Evaluation function	81.73	81.73	

	CART	EVTREE
Positive predictive value	75%	72%
Sensitivity	70%	76%
F-measure	72%	74%

COMPARISON
(VARIATION OPERATORS)



CONCLUSION

Simulated Data

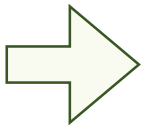
- Evtree works better than CART
- both did not include X_1 as a split variable

Celiac Data

- CART has higher correct classification percentage
- Evtree outperforms CART

Evtree

- ✓ **complement to CART**
- ✓ global partitioning method



another viewpoint

THANK YOU
FOR YOUR ATTENTION!