

Decision trees and Random Forests

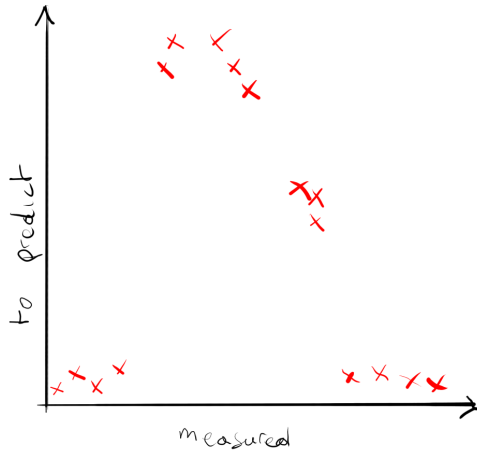
AI for ecologists

Paul Tresson

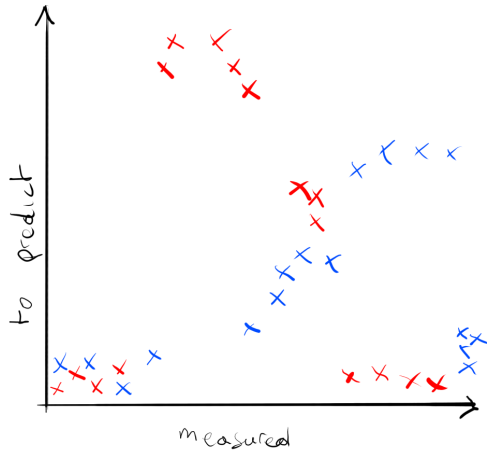
20/05/25

Introduction

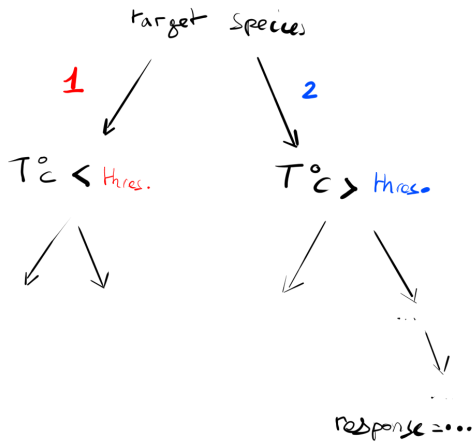
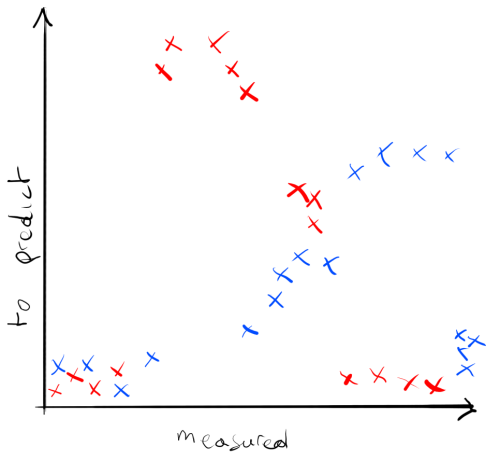
Motivation



Motivation



Motivation



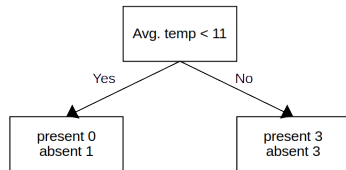
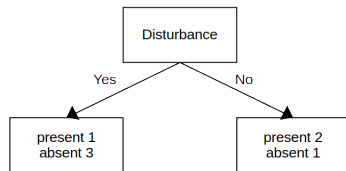
Decision Trees

Simple example

Disturbance	Habitat	Avg. temp.	Presence
Yes	Shrubs	10	0
Yes	Forests	12	0
No	Shrubs	18	1
No	Shrubs	25	1
Yes	Shrubs	28	1
Yes	Forests	30	0
No	Forests	33	0

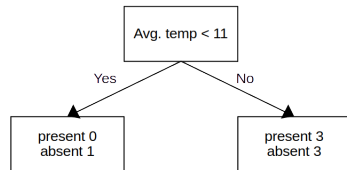
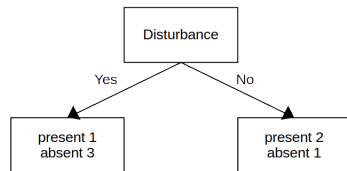
Simple example

Disturbance	Habitat	Avg. temp.	Presence
Yes	Shrubs	10	0
Yes	Forests	12	0
No	Shrubs	18	1
No	Shrubs	25	1
Yes	Shrubs	28	1
Yes	Forests	30	0
No	Forests	33	0



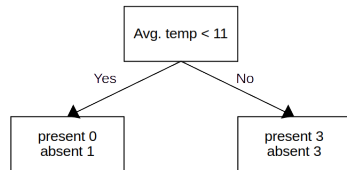
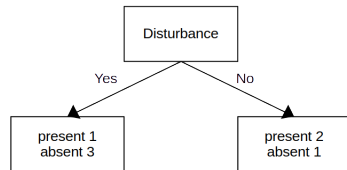
Gini impurity

$$\sum_{i=1}^J \left(p_i \sum_{k \neq i} p_k \right) = 1 - \sum_{i=1}^J p_i^2$$



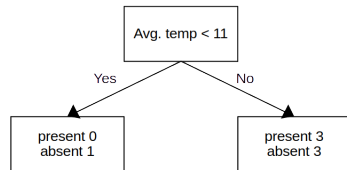
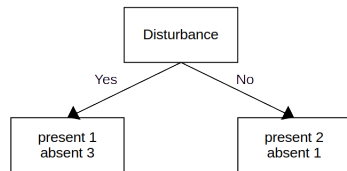
Gini impurity

$$1 - \left(\frac{1}{1+3}\right)^2 - \left(\frac{3}{1+3}\right)^2 = 0.375$$



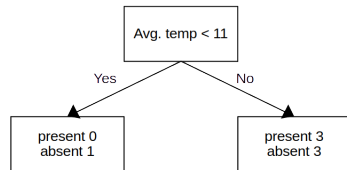
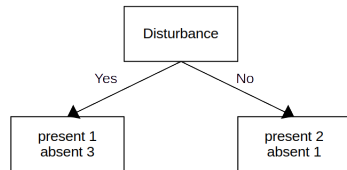
Gini impurity

$$\text{Leaf Gini} = \left(\frac{4}{4+3}\right)0.375$$



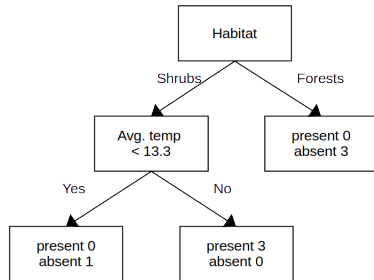
Gini impurity

$$1 - \left(\frac{0}{0+1}\right)^2 - \left(\frac{1}{0+1}\right)^2 = 0$$



Building the tree

Disturbance	Habitat	Avg. temp.	Presence
Yes	Shrubs	10	0
Yes	Forests	12	0
No	Shrubs	18	1
No	Shrubs	25	1
Yes	Shrubs	28	1
Yes	Forests	30	0
No	Forests	33	0



Random Forests

RF advantages and drawbacks

Advantages

- different inputs

Drawbacks

RF advantages and drawbacks

Advantages

- different inputs
- different outputs

Drawbacks

RF advantages and drawbacks

Advantages

- different inputs
- different outputs
- \approx explainable

Drawbacks

RF advantages and drawbacks

Advantages

- different inputs
- different outputs
- \approx explainable
- pretty fast

Drawbacks

RF advantages and drawbacks

Advantages

- different inputs
- different outputs
- \approx explainable
- pretty fast
- seasoned

Drawbacks

RF advantages and drawbacks

Advantages

- different inputs
- different outputs
- \approx explainable
- pretty fast
- seasoned

Drawbacks

- need to test hyper-parameters

RF advantages and drawbacks

Advantages

- different inputs
- different outputs
- \approx explainable
- pretty fast
- seasoned

Drawbacks

- need to test hyper-parameters
- **need for rich descriptors**

Decendants

- Gradient Boosting
- XGBoost

Usefull ressources

- `scikit-learn` docs !
- StatQuest

Thanks for you attention !

Let's practice !

References i