

# Decision trees and Random Forests

AI for ecologists

---

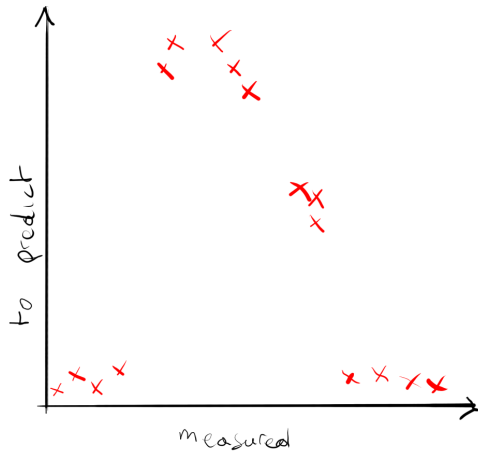
Paul Tresson

20/05/25

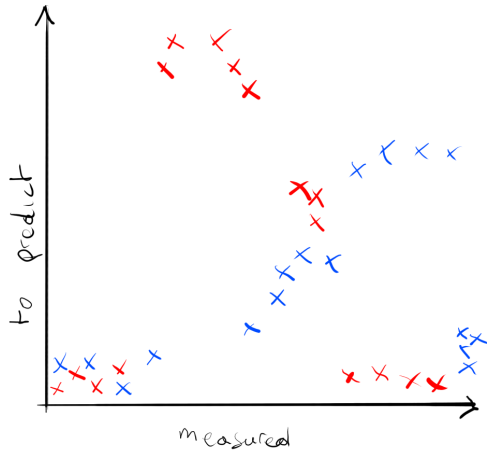
# Introduction

---

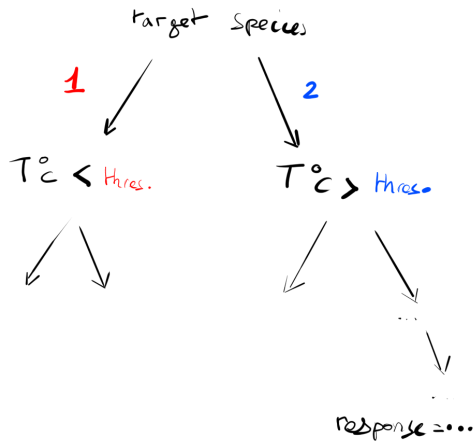
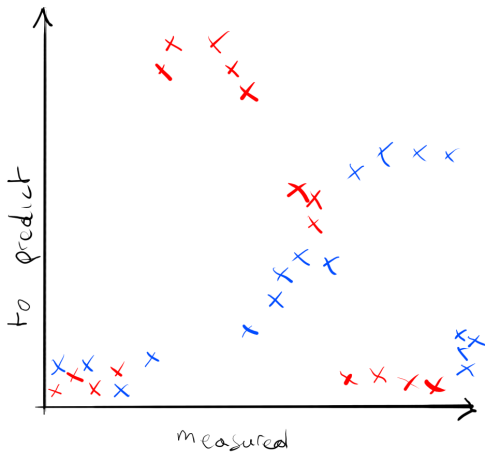
# Motivation



# Motivation



# Motivation



# Decision Trees

---

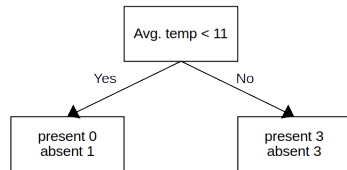
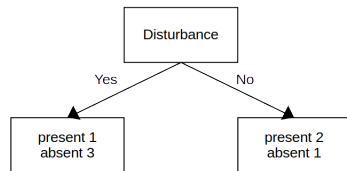
## Simple example

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

Adapted from StatQuest

## Simple example

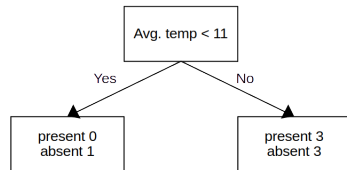
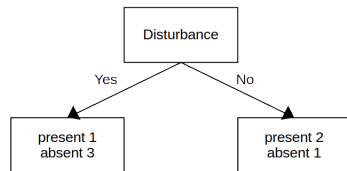
Disturbance	Habitat	Avg. temp.	Presence
Yes	Shrubs	10	0
Yes	Forests	12	0
No	Shrubs	18	1
No	Shrubs	25	1
Yes	Shrubs	28	1
Yes	Forests	30	0
No	Forests	33	0





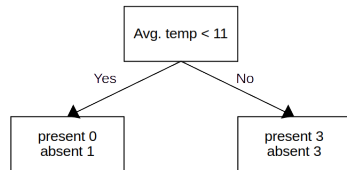
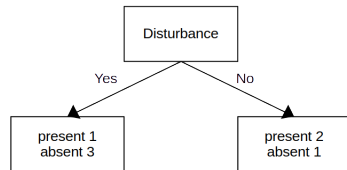
# Gini impurity

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



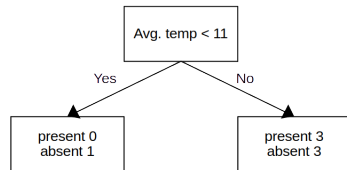
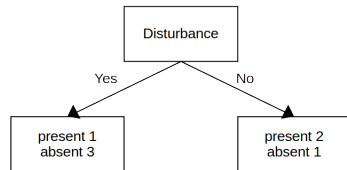
# Gini impurity

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



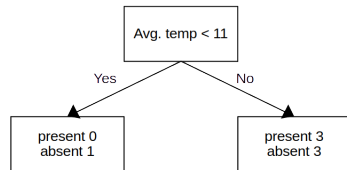
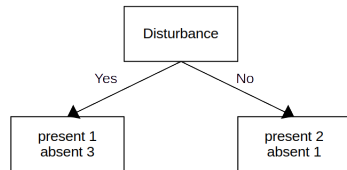
# Gini impurity

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



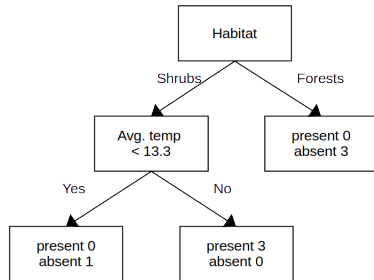
# Gini impurity

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

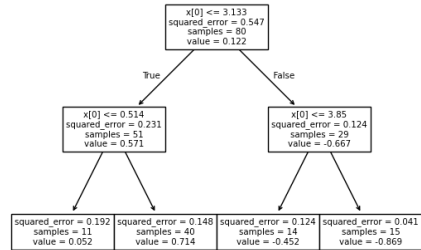
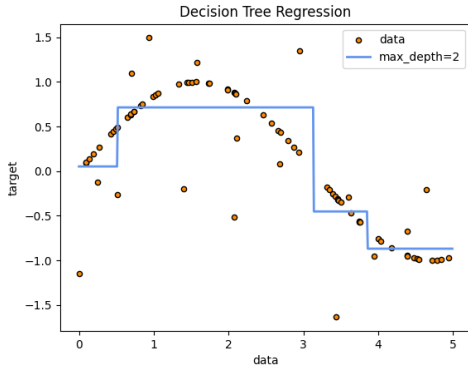


## Building the tree

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

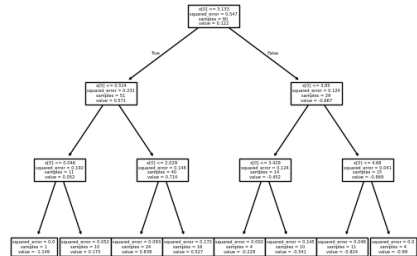
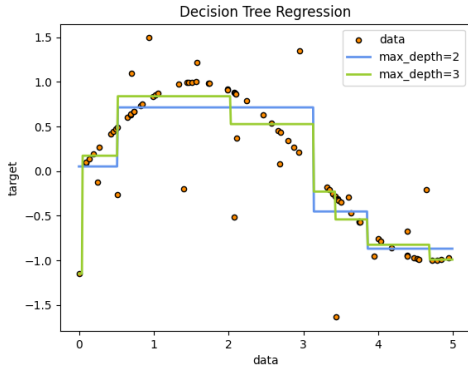


# Regression trees



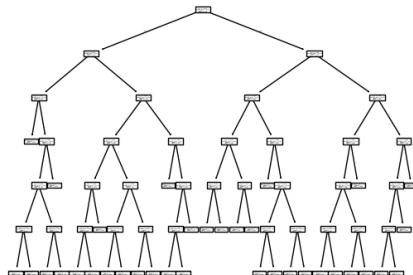
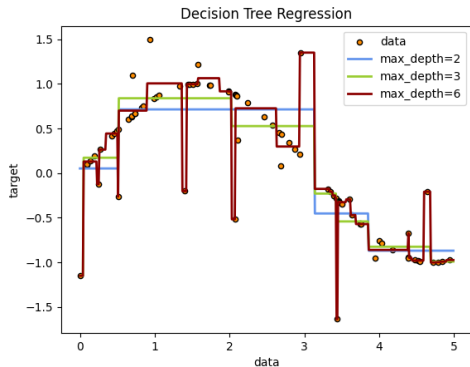
Adapted from sklearn documentation

# Regression trees



Adapted from sklearn documentation

# Regression trees



Adapted from sklearn documentation



# Regression trees

Non-linear data, multiple outputs !

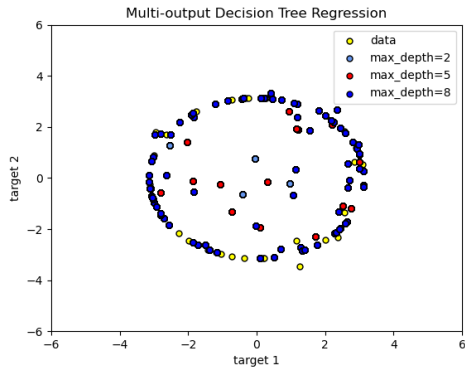


Figure from sklearn documentation

# Random Forests

---

# Main idea

Why not several trees ?

# Boostraping

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

# Boostraping

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

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

## Subset variables

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

## Subset variables

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

Disturbance	<b>Habitat</b>	<b>Avg. temp.</b>	Presence
Yes	<b>Shrubs</b>	<b>10</b>	0
Yes	<b>Forests</b>	<b>12</b>	0
Yes	<b>Forests</b>	<b>12</b>	0
No	<b>Shrubs</b>	<b>18</b>	1
No	<b>Shrubs</b>	<b>25</b>	1
Yes	<b>Shrubs</b>	<b>28</b>	1
Yes	<b>Shrubs</b>	<b>28</b>	1

## Subset variables

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



## RF advantages

- different inputs

## RF advantages

- different inputs
- different outputs

## RF advantages

- different inputs
- different outputs
- $\approx$  explainable

# RF advantages

- different inputs
- different outputs
- $\approx$  explainable
- pretty easy to fit

## RF advantages

- different inputs
  - different outputs
  - $\approx$  explainable
  - pretty easy to fit
- **seasoned and reliable**

## RF drawbacks

- need to test hyper-parameters

## RF 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