Decision Trees

Elizabet Doliar
Lecture VI
June 26th, 2019

What is a decision tree?

Overall:
A decision tree is a map of the possible outcomes of a series of related choices.

What is a decision tree?

Overall:
A decision tree is a map of the possible outcomes of a series of related choices.

Technically speaking:
Decision tree is a type of supervised learning algorithm (having a pre-defined target variable) that is mostly used in classification problems.

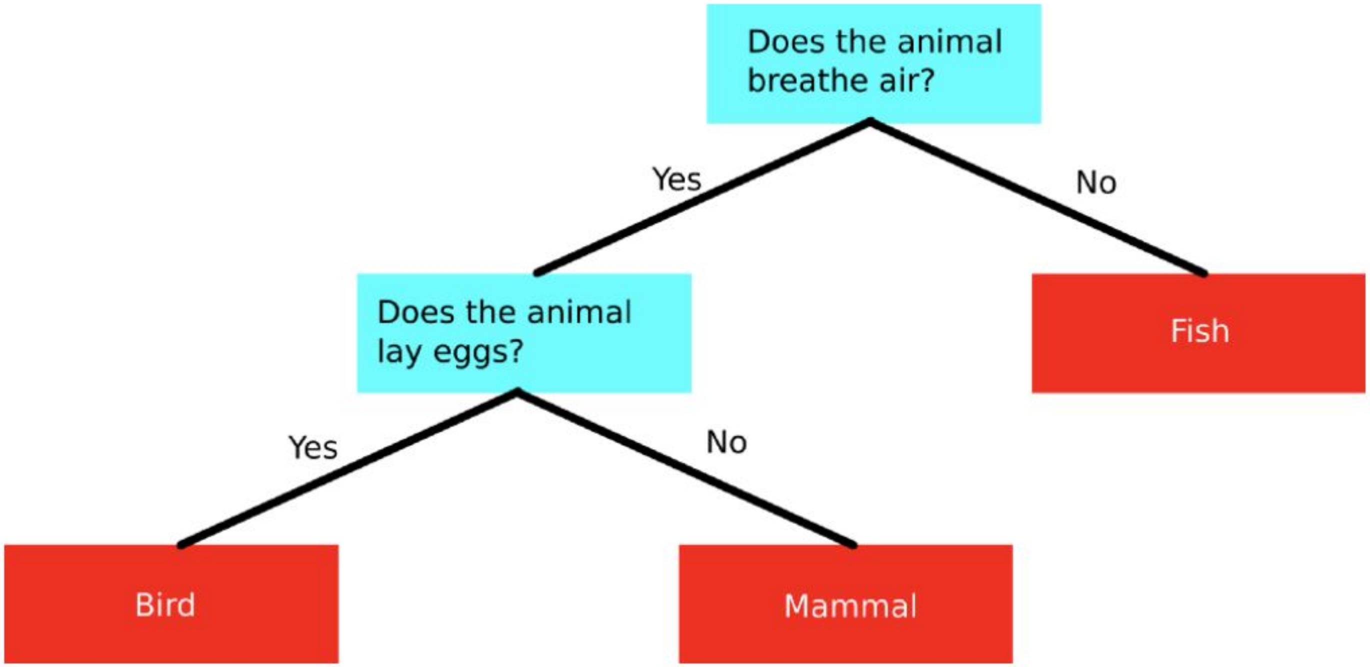
What is a decision tree?

Overall:
A decision tree is a map of the possible outcomes of a series of related choices.

Technically speaking:
Decision tree is a type of supervised learning algorithm (having a pre-defined target variable) that is mostly used in classification problems.

In terms of Methodology:
It analyses a data set in order to construct a set of rules, or questions, which are used to predict a class.

Example

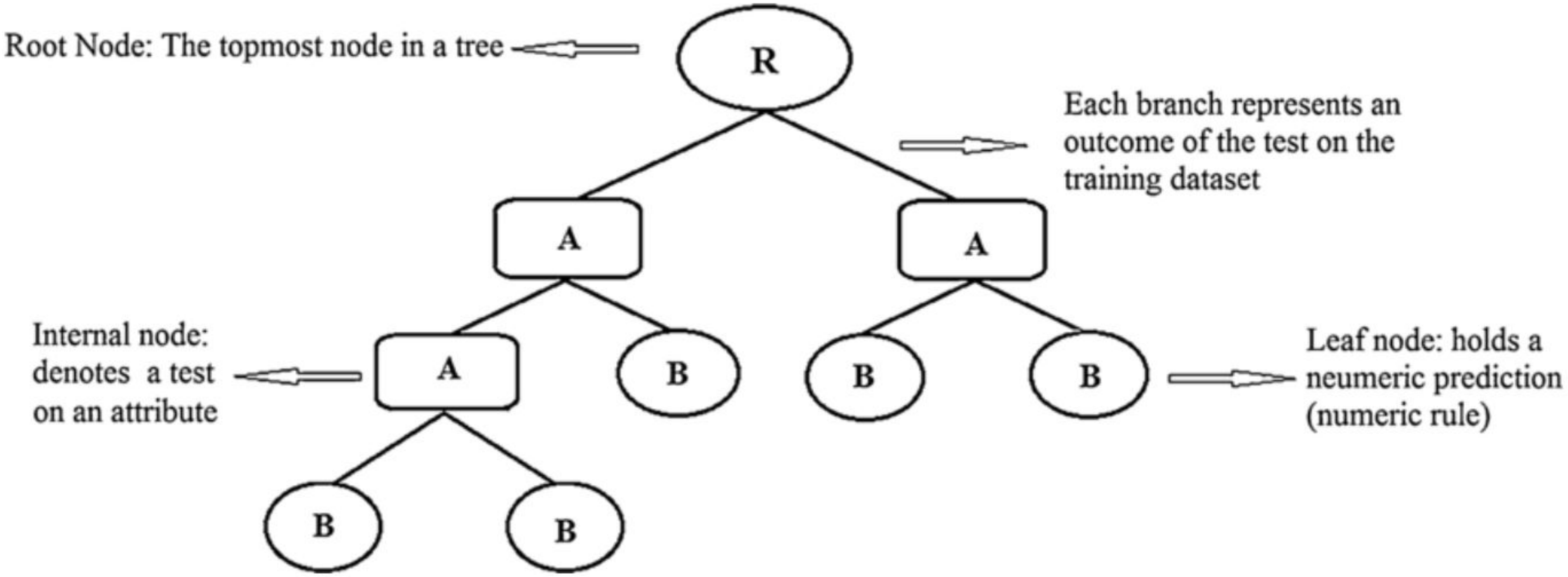


Does the animal breathe air?
Yes
No
Does the animal lay eggs?
Fish
Yes
No
Bird
Mammal

Exercise: select the odd one out

- Loan approval
- Determination of likely buyers of a product using demographic data to enable targeting of limited advertisement budget
- Help with prioritization of emergency room patient treatment using a predictive model based on factors such as age, blood pressure, gender, location and severity of pain, and other measurements
- Evaluation of trends; making estimates, and forecasts
- Predicting election results based on average age, income, previous election results

Terminology



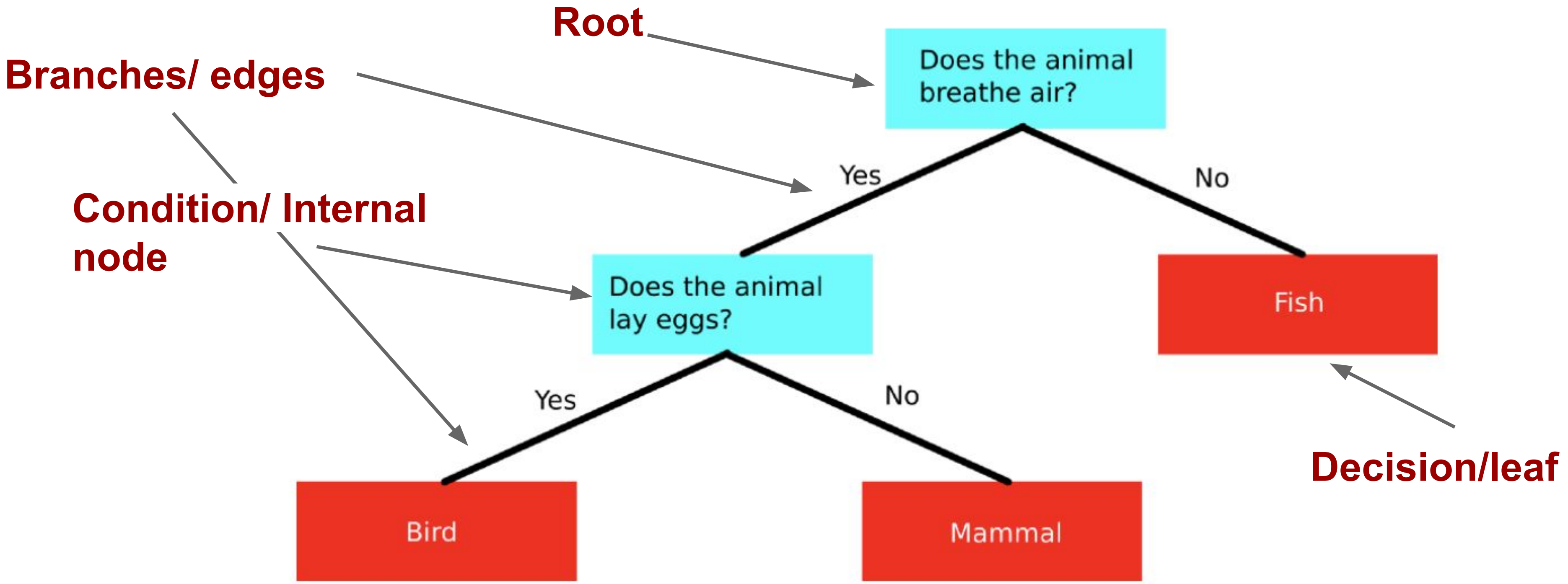
Root Node: The topmost node in a tree

Each branch represents an outcome of the test on the training dataset

Internal node: denotes a test on an attribute

Leaf node: holds a neumeric prediction (numeric rule)

Terminology



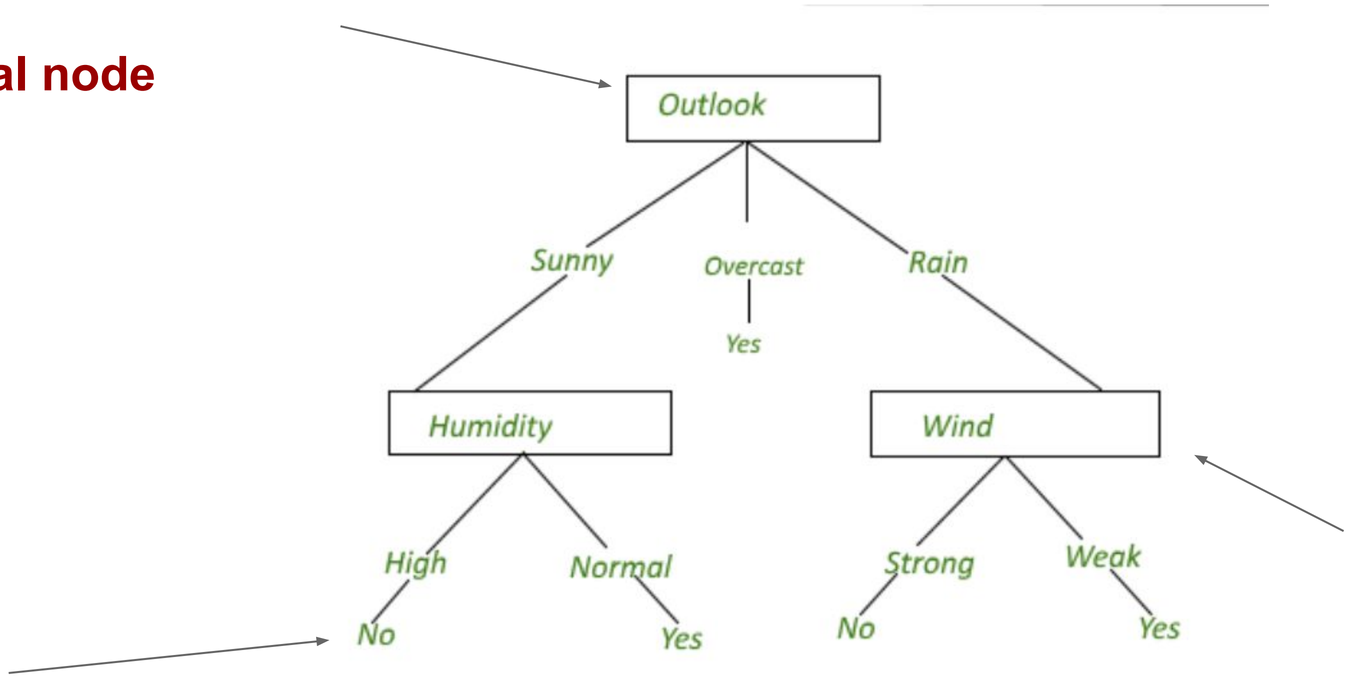
Exercise: Should you play tennis today?

Branches/ edges

Condition/ Internal node

Root

Decision/leaf



Types of Decision Trees

1. **Categorical Variable Decision Tree:** Decision Tree which has categorical target (dependent) variable.

2. **Continuous Variable Decision Tree:** Decision Tree has continuous target (dependent) variable.

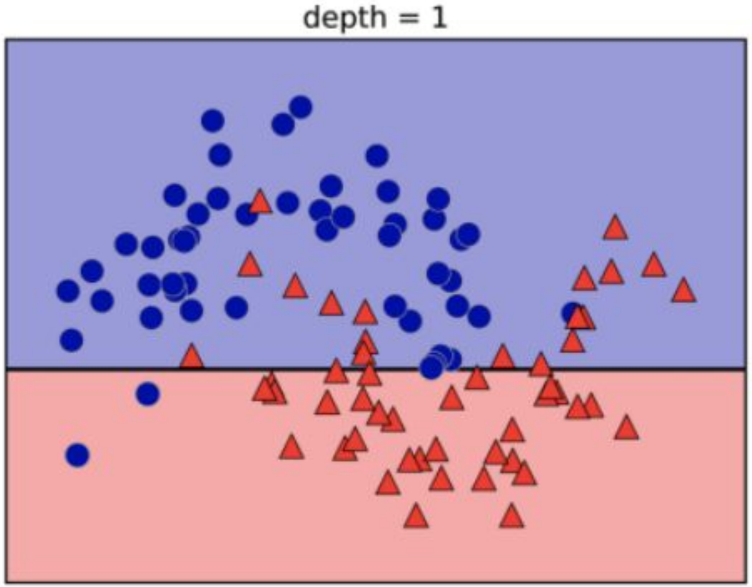
*Target variable = What are we trying to predict?

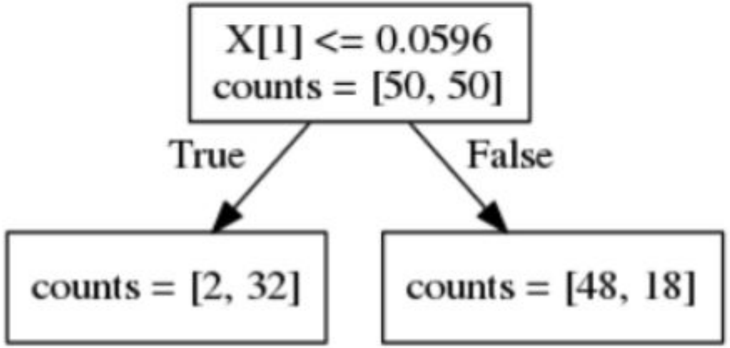
Methodology

Top-Down Approach:

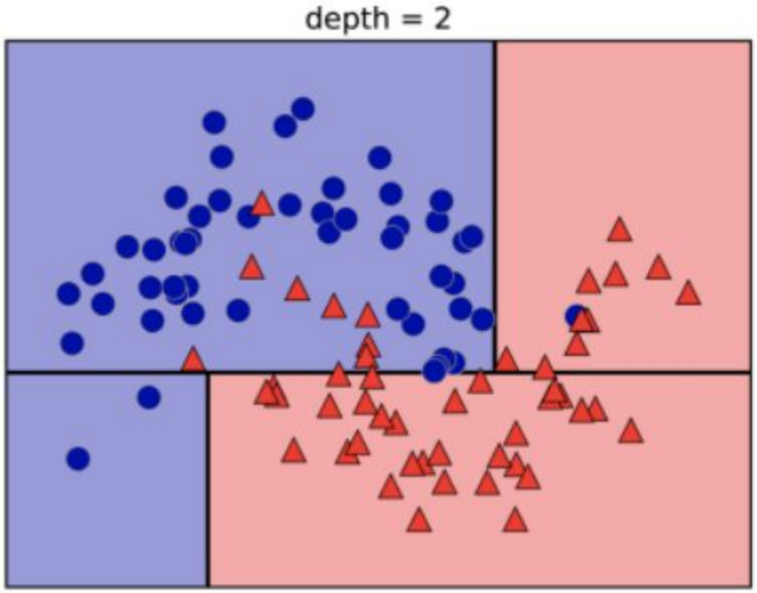
1. Start at the top of the tree (select the root node (feature) to split on)
2. Split the training set into distinct and non-overlapping regions/**subsets**
3. Repeat 1 & 2 -- this splitting process is continued until a user defined stopping criteria is reached or every data point is classified

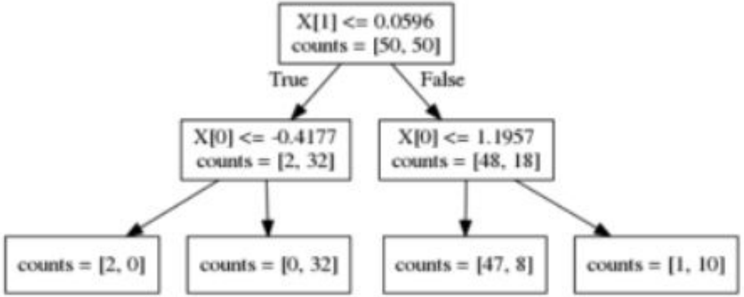
Behind the scenes I





Behind the scenes II





X[1] <= 0.0596
counts = [50, 50]

True

False

X[0] <= -0.4177
counts = [2, 32]

X[0] <= 1.1957
counts = [48, 18]

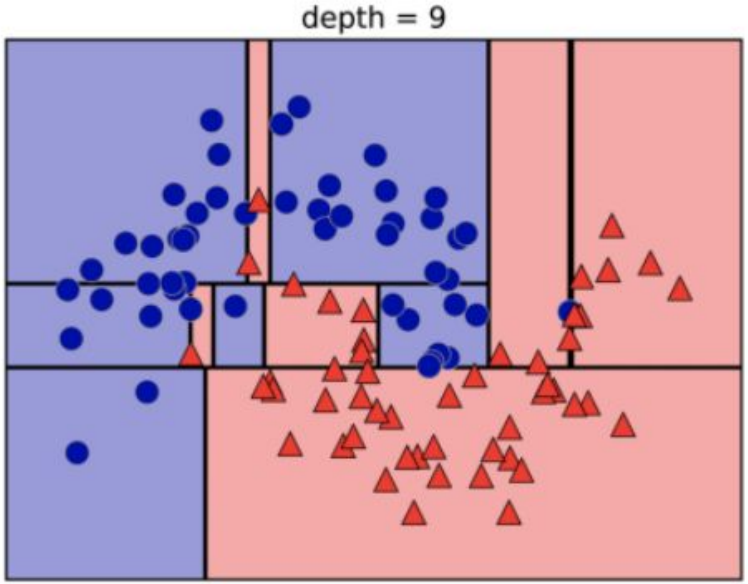
counts = [2, 0]

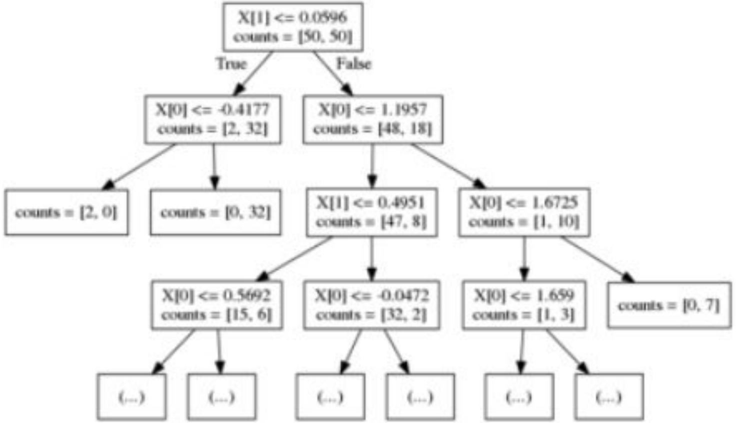
counts = [0, 32]

counts = [47, 8]

counts = [1, 10]

Behind the scenes III





Advantages and Disadvantages

Advantages:

1. Easy to Understand
2. Useful in Data exploration
3. Less data cleaning required
4. Data type is not a constraint (can handle both numerical and categorical variables)

Advantages and Disadvantages

Disadvantages:

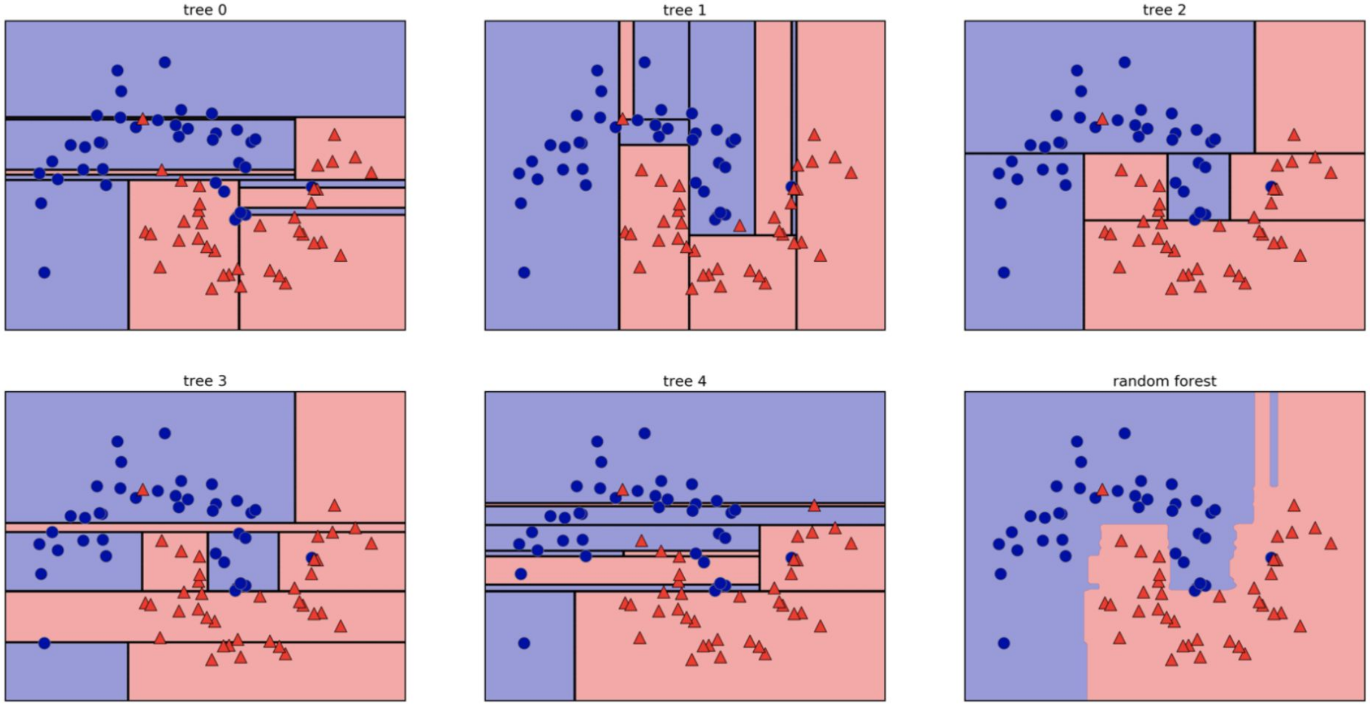
1. May suffer from overfitting
2. Decision trees can be unstable
3. Not fit for continuous variables
4. *Greedy* algorithms cannot guarantee to return the globally optimal decision tree.

Random Forest

The random forest is a model made up of many decision trees.
Two key concepts that gives it the name *random*:

1. Random sampling of training data points when building trees
2. Random subsets of features considered when splitting nodes

Example


tree 0
tree 1
tree 2
tree 3
tree 4
random forest