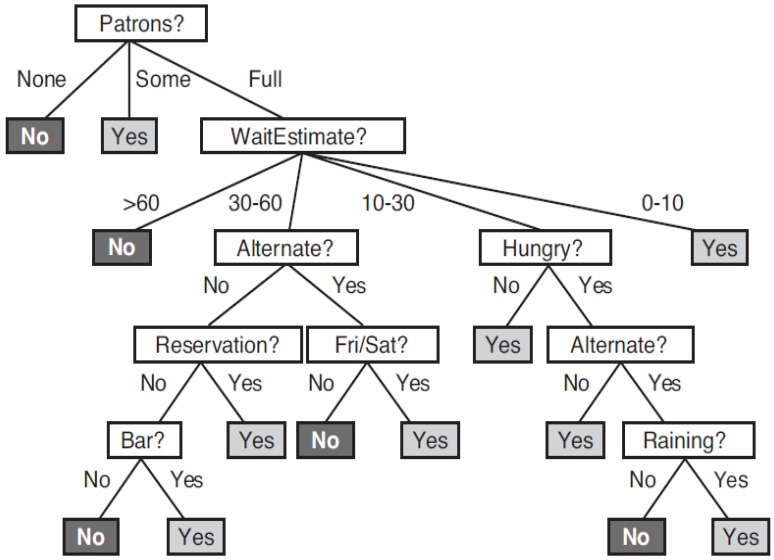
DECISION TREE

DECISION TREE

Key idea:

- Build a tree to classify data based on attribute's value
- Node
  - Where the tree splits
  - Split of attribute's value
- Leaf
  - Class

DECISION TREE



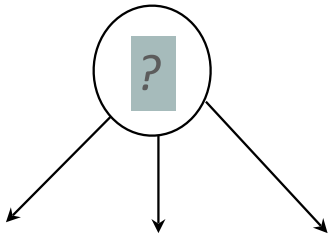
DECISION TREE: SPLITTING THE TREE

How to determine which attribute to use to split the tree?

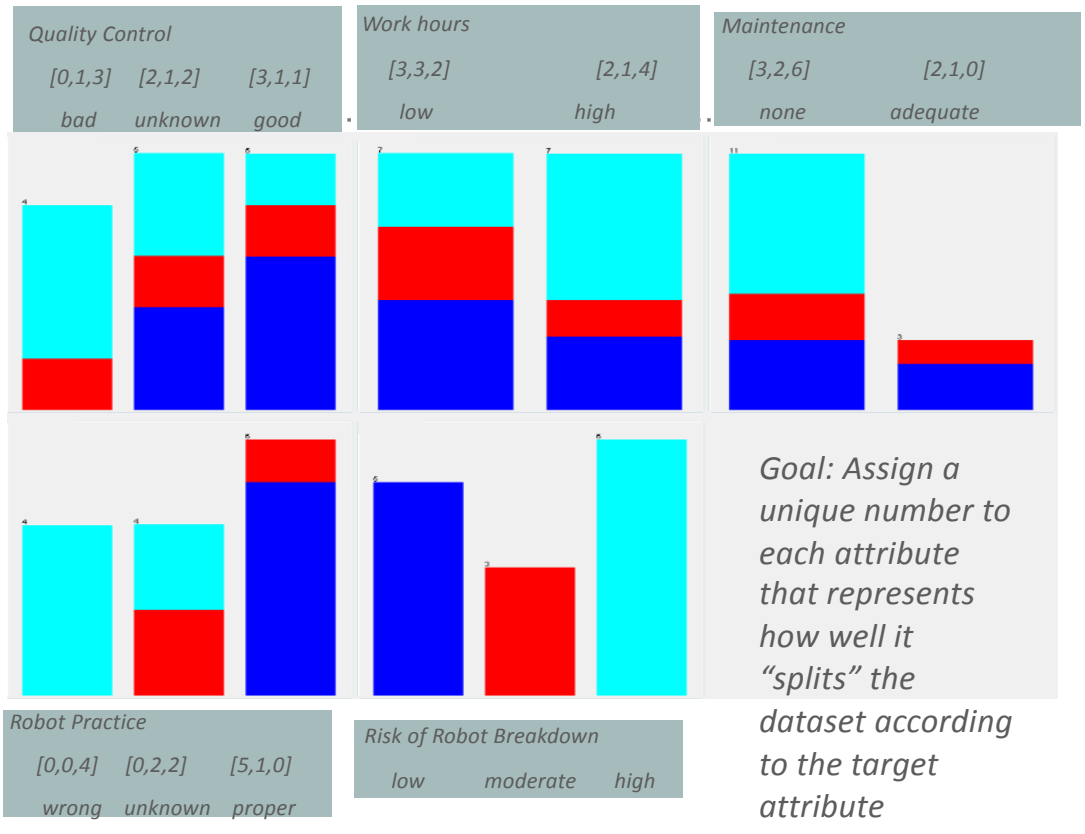
- We want to split the data into parts.
- But how to do it?

- We need some math here.
- It is called “Entropy”

DECISION TREE: ENTROPY



We need to pick the attribute that BEST split the data into groups.



DECISION TREE: ENTROPY

We need a function to calculate a number from the data.

f([0,1,3],[2,1,2],[3,1,1]) = something

There are several functions, but we will use "Entropy"

DECISION TREE: ENTROPY

Entropy([p,q,…,z])

$= - (p/m)\log_2(p/m) – (q/m)\log_2(q/m) - … - (z/m)\log_2(z/m)$

where m = p+q+…+z

DECISION TREE: ENTROPY

Entropy([p,q,…,z])

$= - (p/m)\log_2(p/m) – (q/m)\log_2(q/m) - … - (z/m)\log_2(z/m)$

where m = p+q+…+z

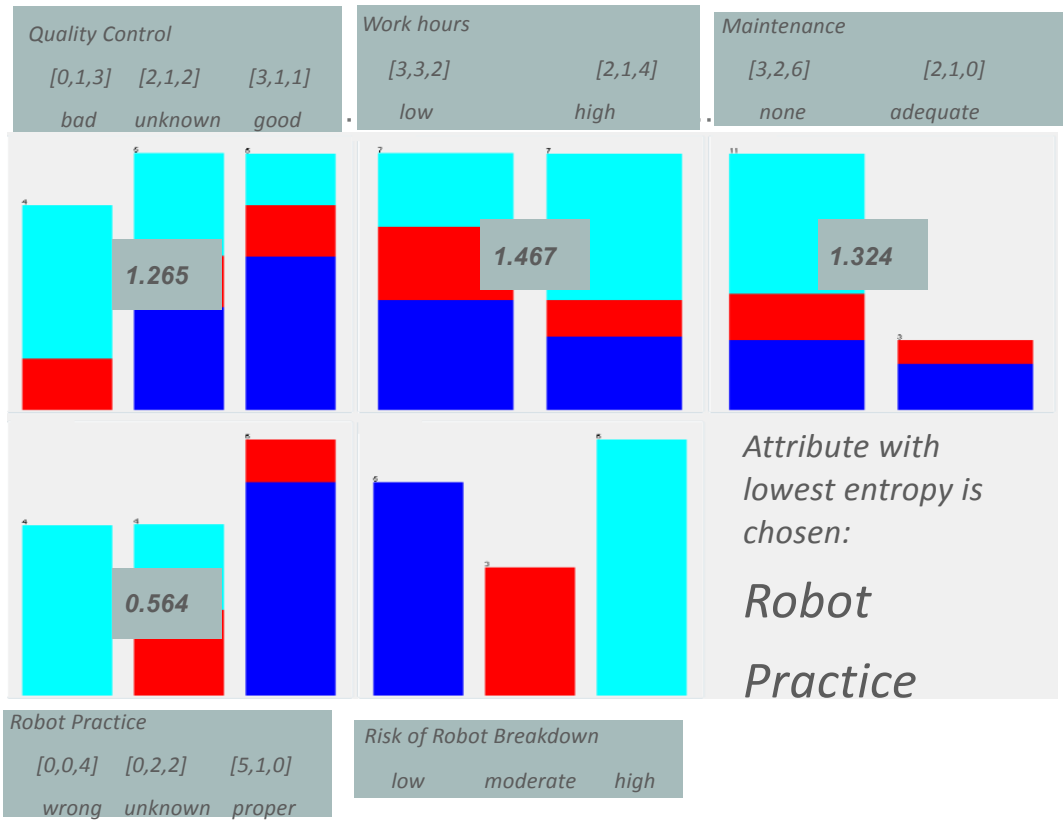
f([0,1,3],[2,1,2],[3,1,1])

= Entropy([0,1,3],[2,1,2],[3,1,1])

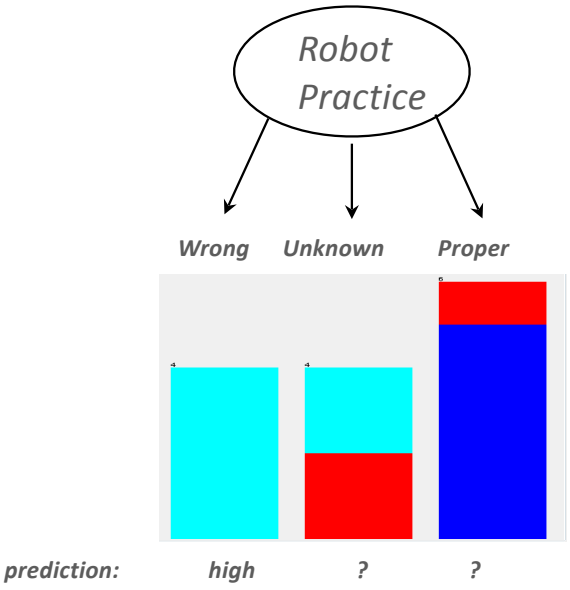
*= (4/14)*Entropy([0,1,3]) + (5/14)*Entropy([2,1,2]) + (5/14)*Entropy([3,1,1])*

= (4/14)[-0 -1/4 $\log_2(1/4)$ -3/4 $\log_2(3/4)$]*
+ (5/14)[-2/5 $\log_2(2/5)$ -1/5 $\log_2(1/5)$ -2/5 $\log_2(2/5)$]*
+ (5/14)[-3/5 $\log_2(3/5)$ -1/5 $\log_2(1/5)$ -1/5 $\log_2(1/5)$]*

= 1.265



DECISION TREE: ENTROPY



DECISION TREE: OVERFITTING

Do we split until the end?

- Yes and No
- It depends on the dataset, and our goal.

Sometimes, tree splits too much, and describe noises as useful information

- Overfitting
- We build a model from training set, which may or may not represent the whole population.
- We need to be as specific and as general as possible.
- It is both Arts and Sciences.

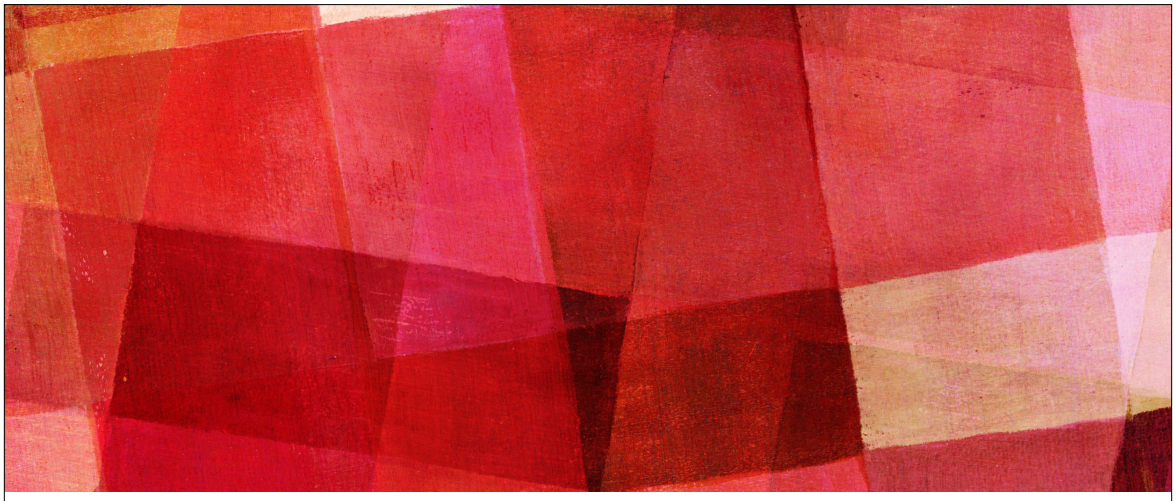
DECISION TREE: PRUNING

How to prevent Overfitting?

- Pruning

Key idea:

- Trim the tree and group up result
- Diabetes dataset in Weka



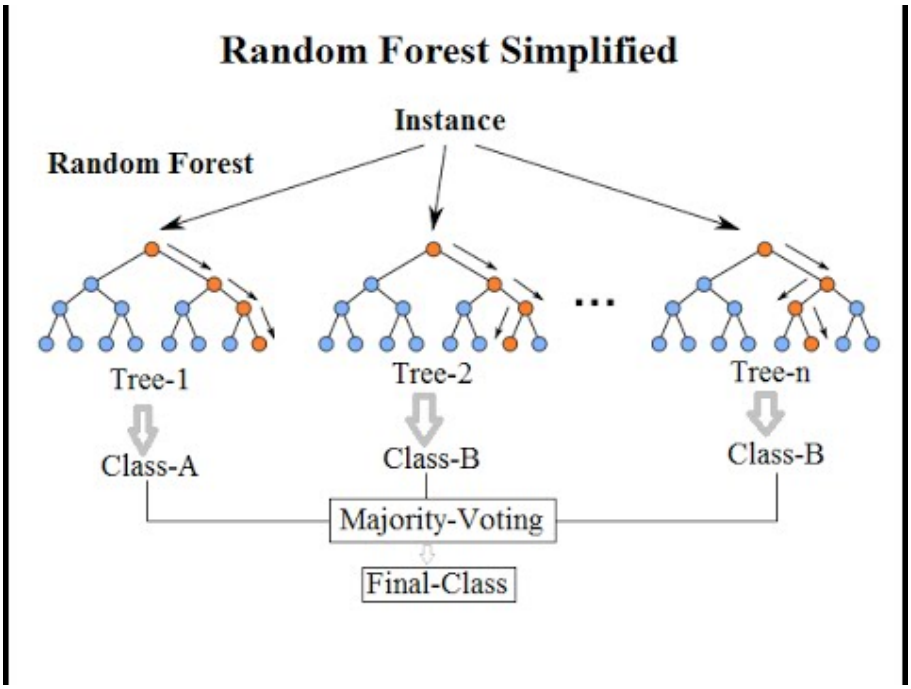
RANDOM FOREST

DECISION TREE

- Decision tree can output one model.
- However, one tree may not cover every details.
- So we grow more tree it becomes a forest.

RANDOM FOREST

- Ensemble Learning
  - Using multiply machine learning algorithms to obtain better result
- We randomly sample attributes with replacement, then construct a decision tree many times.
- After we get many trees, each tree classifies an object, then vote the result.



Retrieved from: https://i.ytimg.com/vi/ajTc5y3OqSQ/hqdefault.jpg