Machine Learning

- Gradient Descent Algorithm
- Linear Regression
- Non-Linear Regression
- Logistic Regression
- Decision Trees
  - Regression Trees
  - Classification Trees
- Clustering Algorithms
  - K-Means
  - Hierarchical clustering
  - DB-Scan
  - Mean Shift
  - GMM
- Support Vector Machine

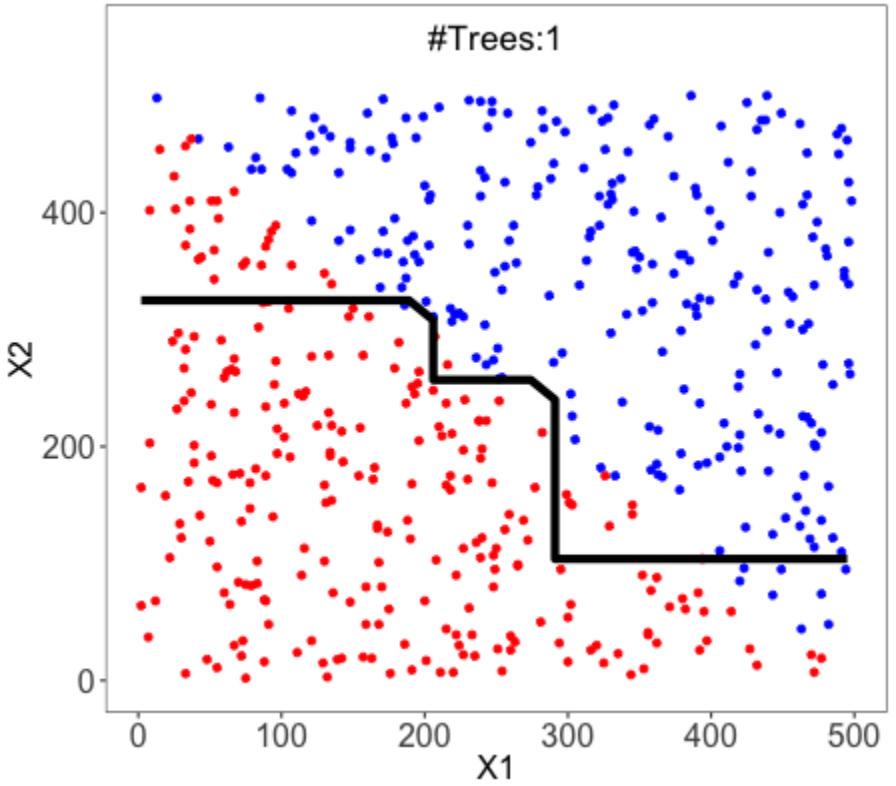
Deep Learning

- MLP
- CNN

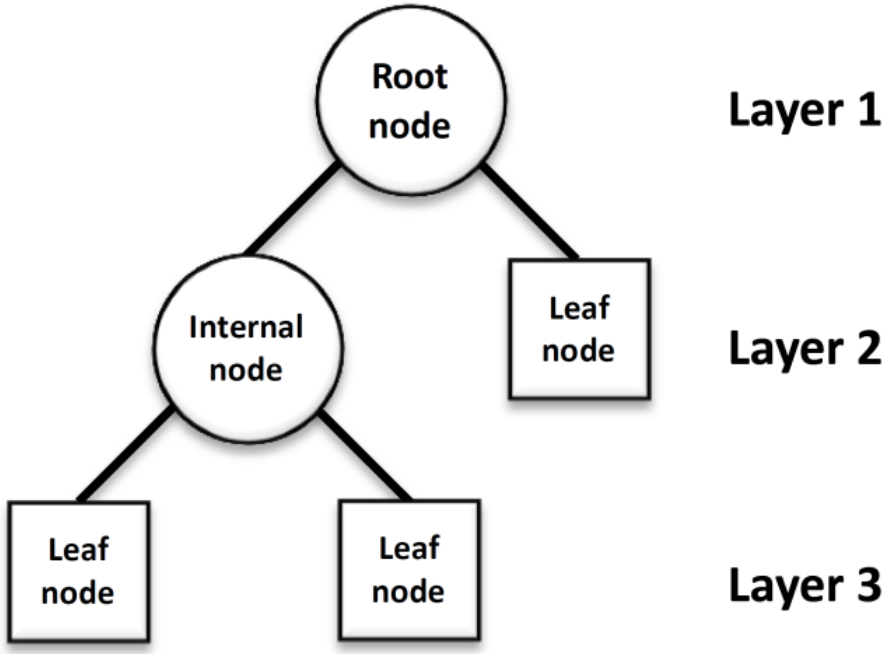
Datasets

- Breast Cancer Wisconsin
- MIMIC-III
- Framingham Heart Study
- Alzheimer’s Disease Neuroimaging Initiative
- Drug discovery
- Microbiome

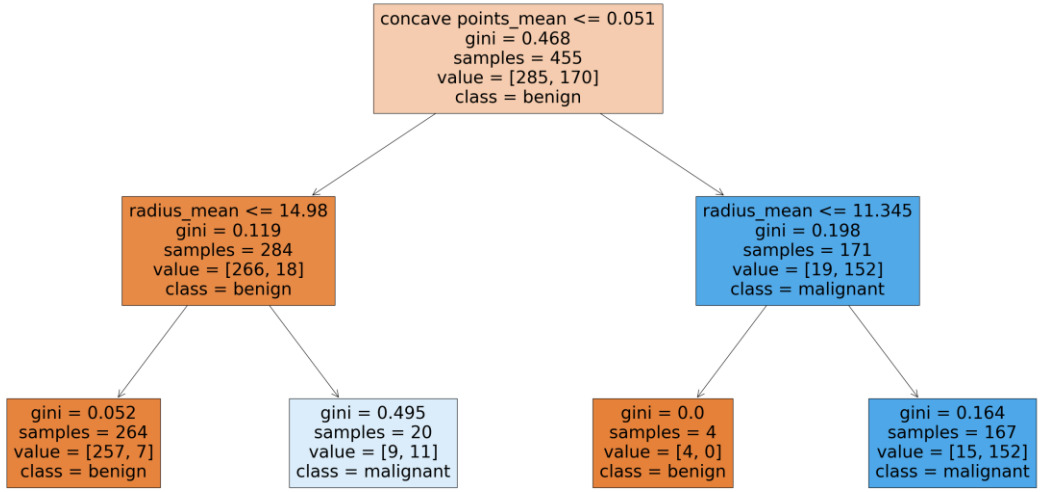
Tree Based Models

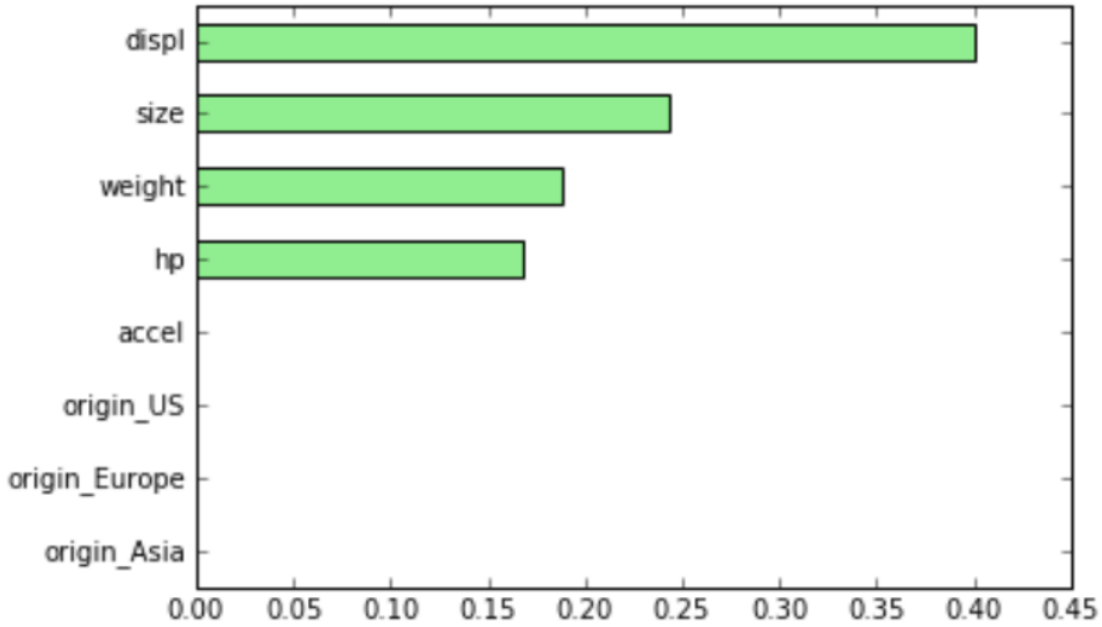


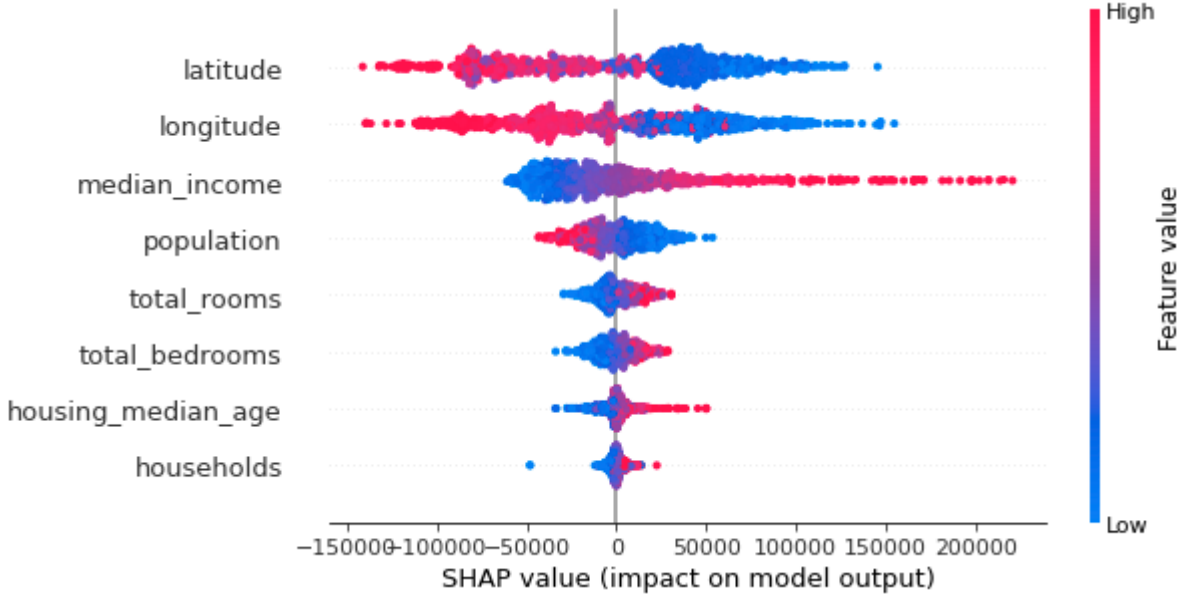
C4 - DT growing algorithm



Feature importance



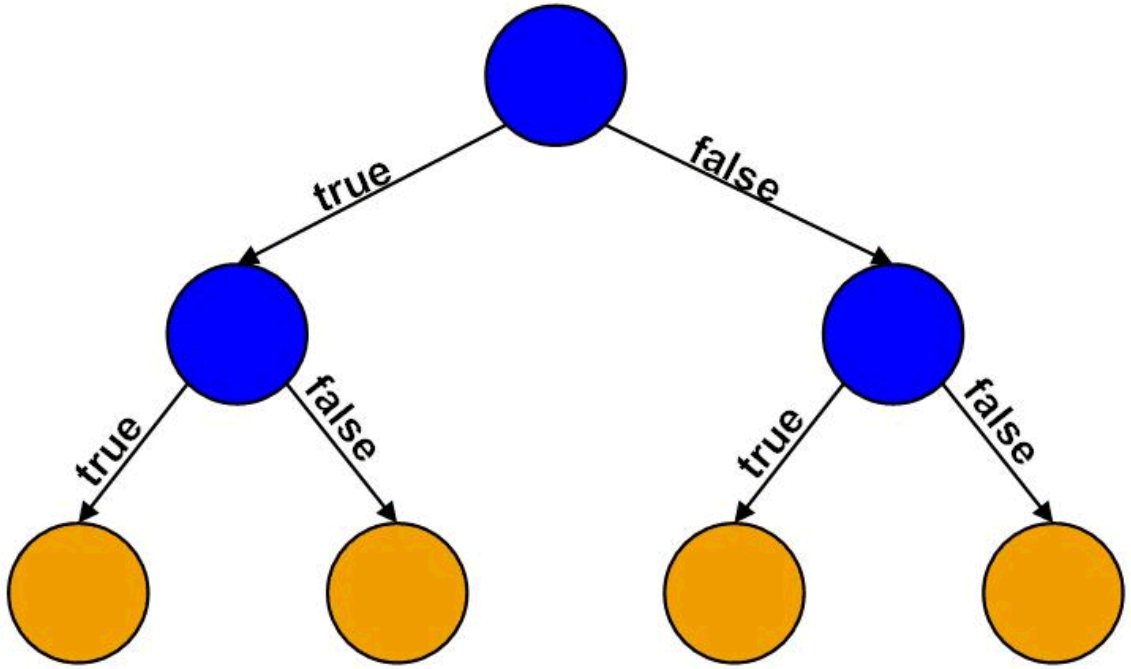




Decision Tree

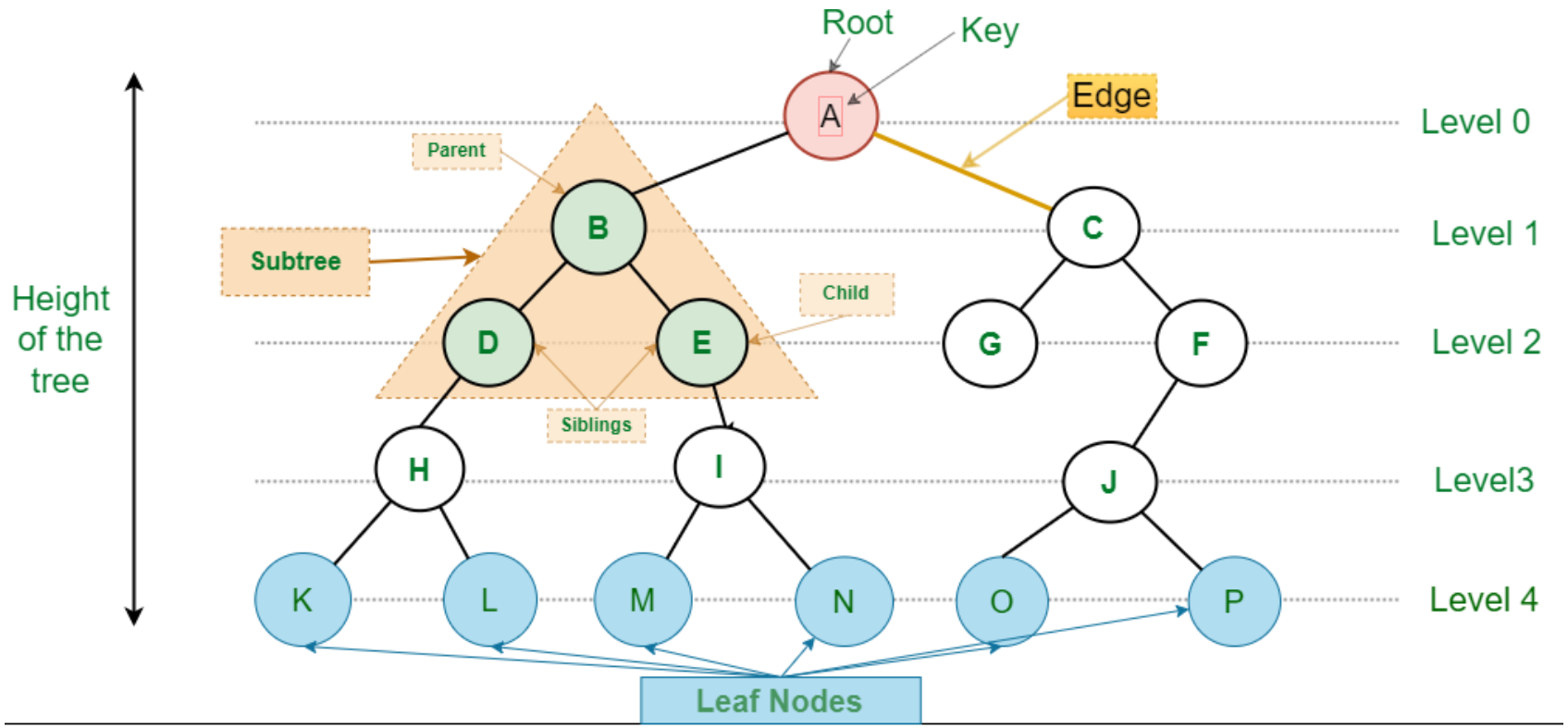
GOAL: Learn if else-questions with each question involving one feature and one split-point
HOW:

- Divides the feature-space (N-dimensional) into regions where all instances in one region are assigned to only one class-label (discrete or continue).



Decision Tree building blocks

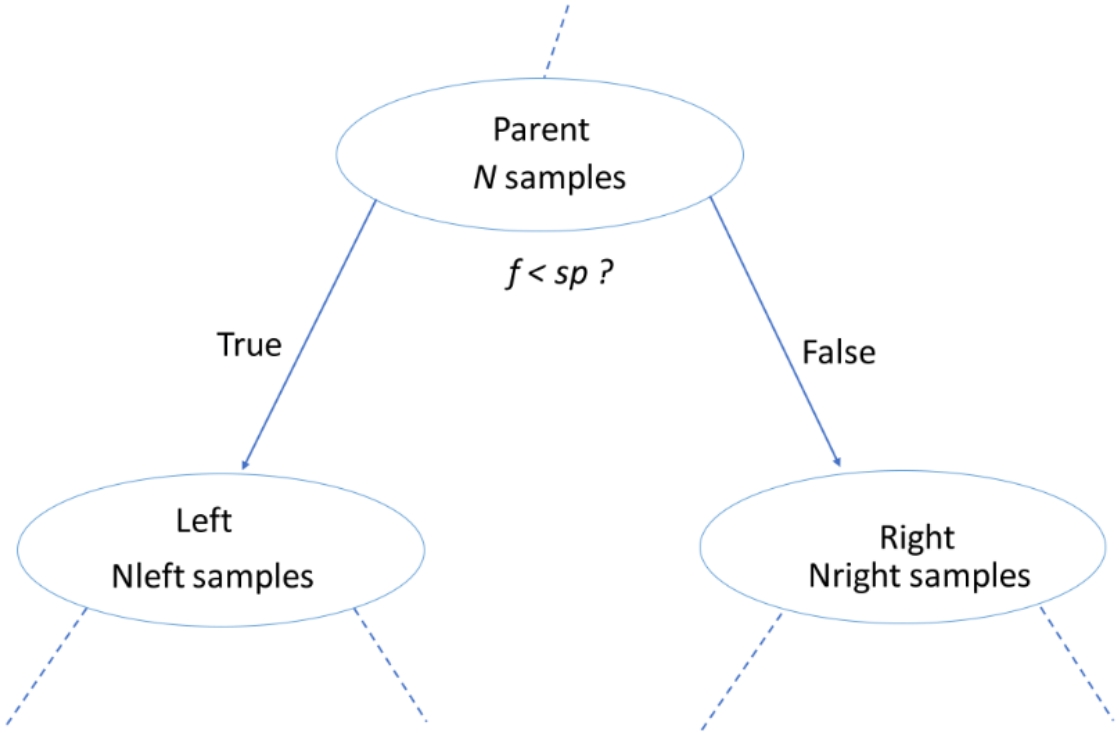
- **DT**: data structure consisting of a hierarchy of nodes (individual units).
- **NODE**:
  - **ROOT**: no-parent node
  - **INTERNAL NODE**: question giving rise to two children nodes
  - **LEAF**: prediction(discrete or continue), no children nodes.



Decision Tree construction/learning

- Nodes are grown recursively.
- The obtention of an internal node or a leaf depends on the state of its predecessors.
- At each node, split the data based on:
  - Feature f and split-point sp to maximize criteria (info-gain).

The algorithm therefore evaluates all variables on some statistical criteria and then chooses the variable that performs best on the criteria.



Decision Tree data science perspective

1. Flow of information through the Decision Tree

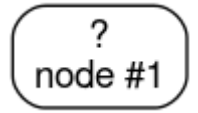
2. How does Decision Trees select which variable to split on at decision nodes?

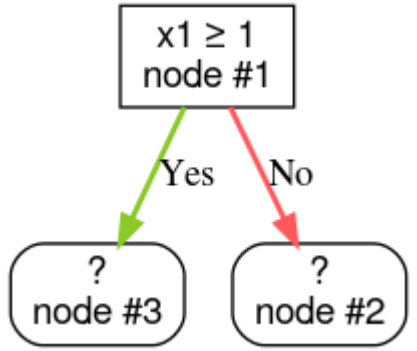
3. How does it decide that the tree has enough branches and that it should stop splitting?

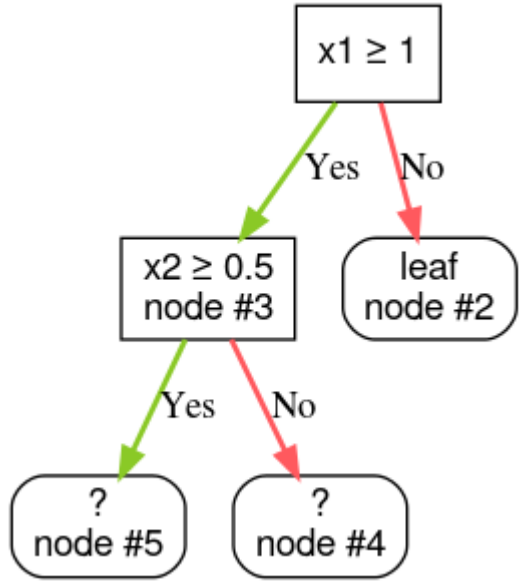
Now let us look at a simplified toy example to understand the above process more concretely.

DT growing algorithm

```
def train_decision_tree(training_examples):
  root = create_root() # Create a decision tree with a single empty root.
  grow_tree(root, training_examples) # Grow the root node.
  return root

def grow_tree(node, examples):
  condition = find_best_condition(examples) # Find the best condition

  if condition is None:
    # No satisfying conditions were found, therefore the grow of the branch stops.
    set_leaf_prediction(node, examples)
    return

  # Create two childrens for the node.
  positive_child, negative_child = split_node(node, condition)

  # List the training examples used by each children.
  negative_examples = [example for example in examples if not condition(example)]
  positive_examples = [example for example in examples if condition(example)]

  # Continue the growth of the children.
  grow_tree(negative_child, negative_examples)
  grow_tree(positive_child, positive_examples)
```







Information

Information theory is a subfield of mathematics concerned with transmitting data across a noisy channel.

Information theory is a field of study concerned with quantifying information for communication.

Measurements of information are widely used in artificial intelligence and machine learning, such as in the construction of decision trees and the optimization of classifier models.

The intuition behind quantifying information is the idea of measuring how much surprise there is in an event. Those events that are rare **(low probability)** are more surprising and therefore have more information than those events that are common **(high probability).**

Information

.

- **Low Probability Event**: High Information (*surprising*).
- **High Probability Event**: Low Information (*unsurprising*).

> *The basic intuition behind information theory is that learning that an unlikely event has occurred is more informative than learning that a likely event has occurred.*

— Page 73, Deep Learning, 2016.

Rare events are more uncertain or more surprising and require more information to represent them than common events.

Information

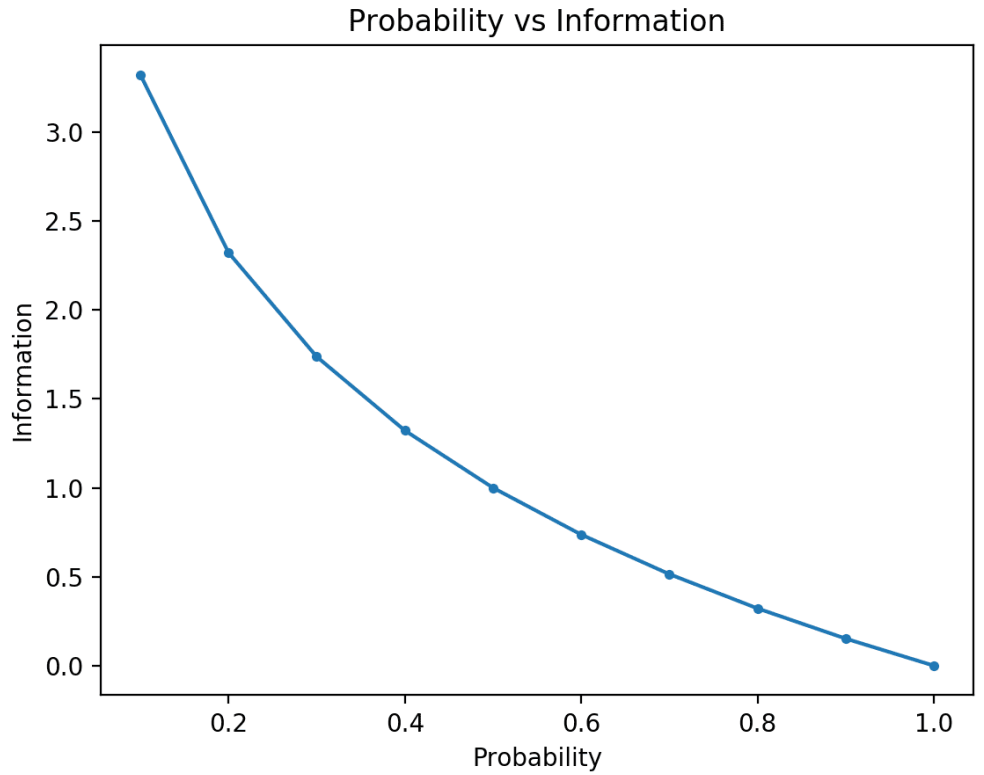
.

The calculation of information is often written as *h()*; for example:

- h(x) = -log(p(x))

The negative sign ensures that the result is always positive or zero.

Information will be zero when the probability of an event is 1.0 or a certainty, e.g. there is no surprise.



Entropy and Information Gain

Entropy is a scientific concept as well as a measurable physical property that is most commonly associated with a state of disorder, randomness, or uncertainty.

In information theory, the entropy of a random variable is the average level of “information”, “surprise”, or “uncertainty” inherent to the variable’s possible outcomes.

In the context of Decision Trees, entropy is a measure of disorder or impurity in a node.

El nodo se divide donde se maximiza InfoGain

$$E = -\sum_{i=1}^{n} p_i \, log_2(p_i)$$

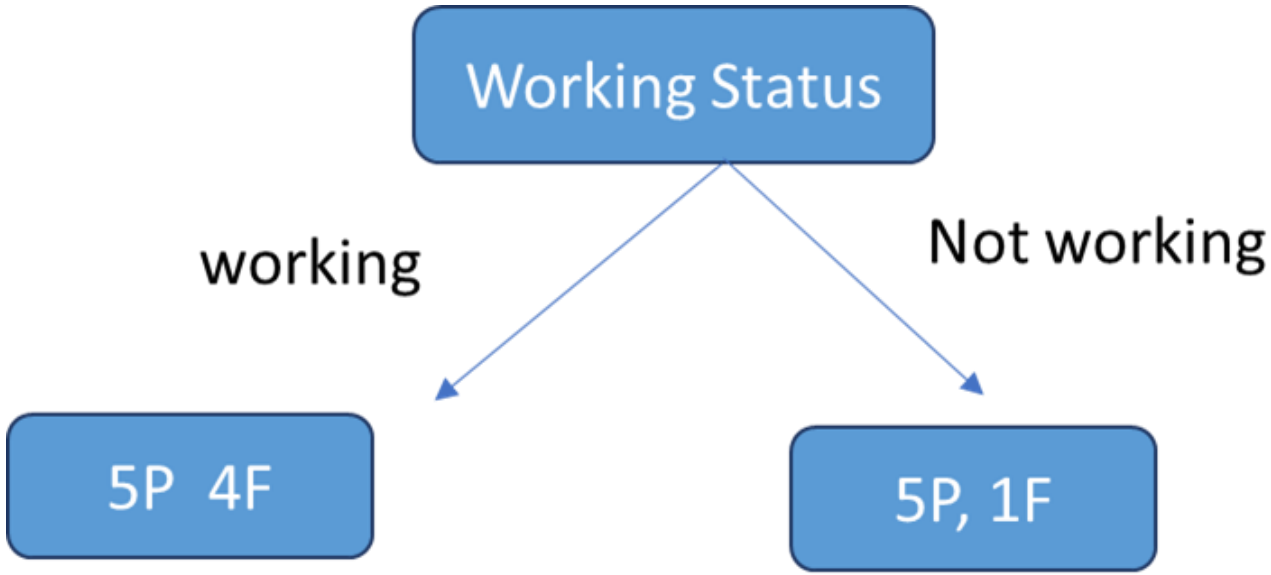
$$Information\ Gain = Entropy_{\ parent} - Entropy_{\ children}$$

$$Information\ Gain = E_{parent} - AvgE_{child}$$

Entropy and Information Gain

$$Information\ Gain = E_{parent} - AvgE_{child}$$

Resp srl no	Target variable	Predictor variable	Predictor variable	Predictor variable
	Exam Result	Other online courses	Student background	Working Status
1	Pass	Y	Maths	NW
2	Fail	N	Maths	W
3	Fail	y	Maths	W
4	Pass	Y	CS	NW
5	Fail	N	Other	W
6	Fail	Y	Other	W
7	Pass	Y	Maths	NW
8	Pass	Y	CS	NW
9	Pass	n	Maths	W
10	Pass	n	CS	W
11	Pass	y	CS	W
12	Pass	n	Maths	NW
13	Fail	y	Other	W
14	Fail	n	Other	NW
15	Fail	n	Maths	W



Entropy and Information Gain

$$Information\ Gain = E_{parent} - AvgE_{child}$$

Resp srl no	Target variable	Predictor variable	Predictor variable	Predictor variable
	Exam Result	Other online courses	Student background	Working Status
1	Pass	Y	Maths	NW
2	Fail	N	Maths	W
3	Fail	y	Maths	W
4	Pass	Y	CS	NW
5	Fail	N	Other	W
6	Fail	Y	Other	W
7	Pass	Y	Maths	NW
8	Pass	Y	CS	NW
9	Pass	n	Maths	W
10	Pass	n	CS	W
11	Pass	y	CS	W
12	Pass	n	Maths	NW
13	Fail	y	Other	W
14	Fail	n	Other	NW
15	Fail	n	Maths	W

P_{pass} = Probability of passing/Total no. of instances = 9/15

P_{fail}= Probability of failing/Total no. of instances = 6/15

$$E = -\sum_{i=1}^{n} p_i \log_2(p_i)$$

$$E = -(P_{pass} \log_2(P_{pass}) + P_{fail} \log_2(P_{fail}))$$

$$\text{Average Entropy} = \frac{(n_{subnode\ 1})}{n\_parent} E\_subnode1 + \frac{(n_{submode\ 2})}{n_{parent}} E\_subnode2$$

Entropy and Information Gain

$$Information\ Gain = E_{parent} - AvgE_{child}$$

Resp srl no	Target variable	Predictor variable	Predictor variable	Predictor variable
	Exam Result	Other online courses	Student background	Working Status
1	Pass	Y	Maths	NW
2	Fail	N	Maths	W
3	Fail	y	Maths	W
4	Pass	Y	CS	NW
5	Fail	N	Other	W
6	Fail	Y	Other	W
7	Pass	Y	Maths	NW
8	Pass	Y	CS	NW
9	Pass	n	Maths	W
10	Pass	n	CS	W
11	Pass	y	CS	W
12	Pass	n	Maths	NW
13	Fail	y	Other	W
14	Fail	n	Other	NW
15	Fail	n	Maths	W

P_{pass} = Probability of passing/Total no. of instances =

P_{fail}= Probability of failing/Total no. of instances =

$$E = -(P_{pass}\ log_2(P_{pass}) + P_{fail}\ log_2(P_{fail})$$

$$\text{Average Entropy} = \frac{(n_{subnode\ 1})}{n\_parent} E\_subnode1 + \frac{(n_{submode\ 2})}{n_{parent}} E\_subnode2$$

Entropy and Information Gain

$$Information\ Gain = E_{parent} - AvgE_{child}$$

Resp srl no	Target variable	Predictor variable	Predictor variable	Predictor variable
	Exam Result	Other online courses	Student background	Working Status
1	Pass	Y	Maths	NW
2	Fail	N	Maths	W
3	Fail	y	Maths	W
4	Pass	Y	CS	NW
5	Fail	N	Other	W
6	Fail	Y	Other	W
7	Pass	Y	Maths	NW
8	Pass	Y	CS	NW
9	Pass	n	Maths	W
10	Pass	n	CS	W
11	Pass	y	CS	W
12	Pass	n	Maths	NW
13	Fail	y	Other	W
14	Fail	n	Other	NW
15	Fail	n	Maths	W

$$E = -\left[\frac{8}{15}log_2\left(\frac{8}{15}\right) + \frac{7}{15}log_2\left(\frac{7}{15}\right)\right]$$

$$E=-(.5333*\ (-0.9069)\ + .4667*\ (-1.0995)\ ) = 0.9968$$

$$Entropy_{\_working} = -\left[\frac{3}{9} * log_2\left(\frac{3}{9}\right) + \frac{6}{9}log_2\left(\frac{6}{9}\right)\right] = 0.9183$$

$$\text{Entropy\_Notworking} = \left[\frac{5}{6} * log_2(6) + \frac{1}{6}log_2\left(\frac{1}{6}\right)\right] = .6500$$

$$\text{Entropy\_working\_status} = \left[\frac{9}{15} * 0.9183 + \frac{6}{15}0.6500\right] = 0.8110$$

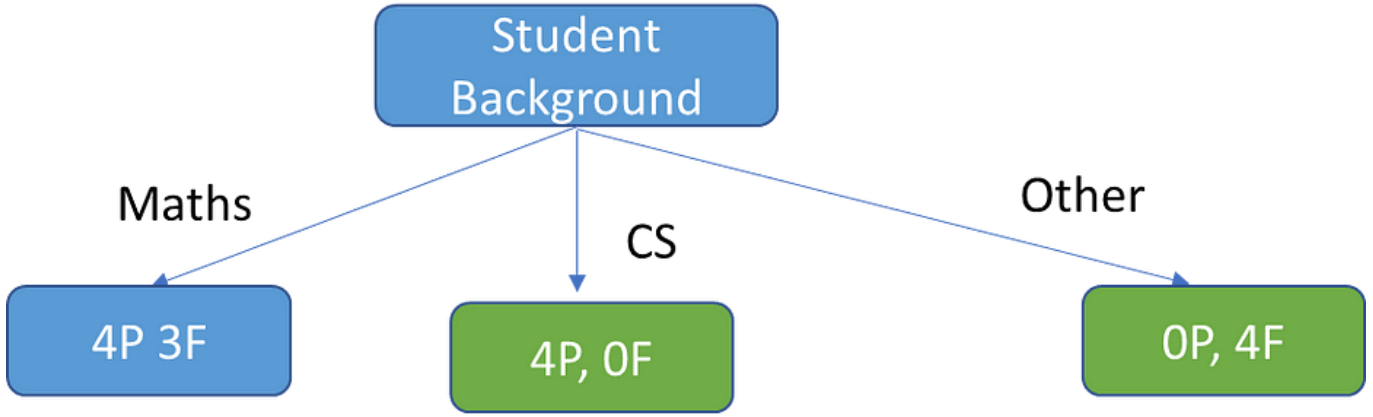
$$Information\ Gain = \ 0.9183–0.8119 = .1858$$

Entropy and Information Gain

$$Information\ Gain = E_{parent} - AvgE_{child}$$

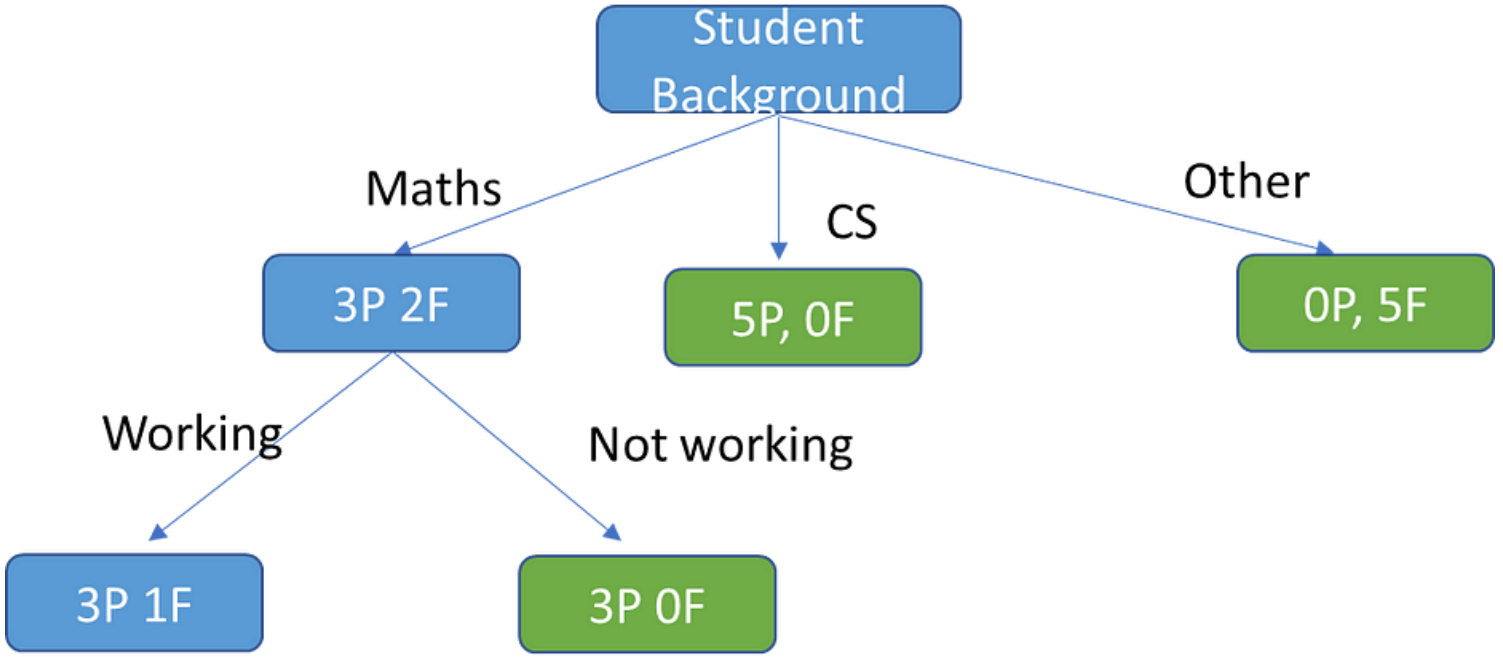
Resp srl no	Target variable	Predictor variable	Predictor variable	Predictor variable
	Exam Result	Other online courses	Student background	Working Status
1	Pass	Y	Maths	NW
2	Fail	N	Maths	W
3	Fail	y	Maths	W
4	Pass	Y	CS	NW
5	Fail	N	Other	W
6	Fail	Y	Other	W
7	Pass	Y	Maths	NW
8	Pass	Y	CS	NW
9	Pass	n	Maths	W
10	Pass	n	CS	W
11	Pass	y	CS	W
12	Pass	n	Maths	NW
13	Fail	y	Other	W
14	Fail	n	Other	NW
15	Fail	n	Maths	W

	Entropy Node	Average Entropy	Information Gain
Parent	0.9968		
working	0.9183	0.8110	
Not_work	0.6500		0.1858
Bkgrd_Ma	0.9852	0.4598	
Bkgrd_CS	0.0000		
Bkgrd_oth	0.0000		0.5370
online_co	0.9544	0.9688	
online_no	0.9852		0.0280



Resp srl no	Target variable	Predictor variable	Predictor variable	Predictor variable
	Exam Result	Other online courses	Student background	Working Status
1	Pass	Y	Maths	NW
2	Fail	N	Maths	W
3	Fail	y	Maths	W
4	Pass	Y	CS	NW
5	Fail	N	Other	W
6	Fail	Y	Other	W
7	Pass	Y	Maths	NW
8	Pass	Y	CS	NW
9	Pass	n	Maths	W
10	Pass	n	CS	W
11	Pass	y	CS	W
12	Pass	n	Maths	NW
13	Fail	y	Other	W
14	Fail	n	Other	NW
15	Fail	n	Maths	W

	Entropy Node	Average Entropy	Information Gain
Bkgrd_Ma	0.9852		
working	0.8113	0.4636	0.5216
Not_worki	0.0000		
online_co	0.9183	0.9533	0.0319
online_no	1.0000		



Gini Index

The Gini Index or Impurity measures the probability for a random instance being misclassified when chosen randomly

$$Gini = 1 - \sum_{i=1}^{j} P(i)^2$$

Calculates the amount of probability of a specific feature that is classified incorrectly when selected randomly.

The Gini index has a maximum impurity is 0.5 and maximum purity is 0, whereas Entropy has a maximum impurity of 1 and maximum purity is 0.

Se divide el nodo de menor pureza

Gini Index

Resp srl no	Target variable	Predictor variable	Predictor variable	Predictor variable
	Exam Result	Other online courses	Student background	Working Status
1	Pass	Y	Maths	NW
2	Fail	N	Maths	W
3	Fail	y	Maths	W
4	Pass	Y	CS	NW
5	Fail	N	Other	W
6	Fail	Y	Other	W
7	Pass	Y	Maths	NW
8	Pass	Y	CS	NW
9	Pass	n	Maths	W
10	Pass	n	CS	W
11	Pass	y	CS	W
12	Pass	n	Maths	NW
13	Fail	y	Other	W
14	Fail	n	Other	NW
15	Fail	n	Maths	W

$$Gini = 1 - \sum_{i=1}^{j} P(i)^2$$

$$Gini_{maths} = 1 - \left(\frac{4}{7}\right)^2 - \left(\frac{3}{7}\right)^2 = .4897$$

$$Gini_{CS} = 1 - \left(\frac{4}{4}\right)^2 - \left(\frac{0}{4}\right)^2 = 0$$

$$Gini_{others} = 1 - \left(\frac{0}{4}\right)^2 - \left(\frac{4}{4}\right)^2 = 0$$

Bibliography

https://machinelearningmastery.com/what-is-information-entropy/
https://towardsdatascience.com/decision-trees-explained-entropy-information-gain-gini-index-ccp-pruning-4d78070db36c
https://medium.com/analytics-steps/understanding-the-gini-index-and-information-gain-in-decision-trees-ab4720518ba8
https://github.com/kamperh/data414
https://polakowo.io/datadocs/docs/machine-learning/tree-based-models
https://towardsdatascience.com/why-random-forests-outperform-decision-trees-1b0f175a0b5
https://www.statistics.cool/post/why-do-random-forests-work/
https://luisvalesilva.com/datasimple/random_forests.html
https://developers.google.com/machine-learning/decision-forests/growing?hl=es-419