DECISION TREES

Sebastien GONCALVES CLARO
Master Bioinformatics

Research & Development initiation

One of the most popular classification technique

- **Machine learning :** parametric supervised learning method.
- **Goal :** create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.
- **Advantage :** generate simple knowledge to solve difficult ones.
- The best to discover knowledge (outperformed by SVM and ensemble classifier).

Family of algorithm to train a classifier

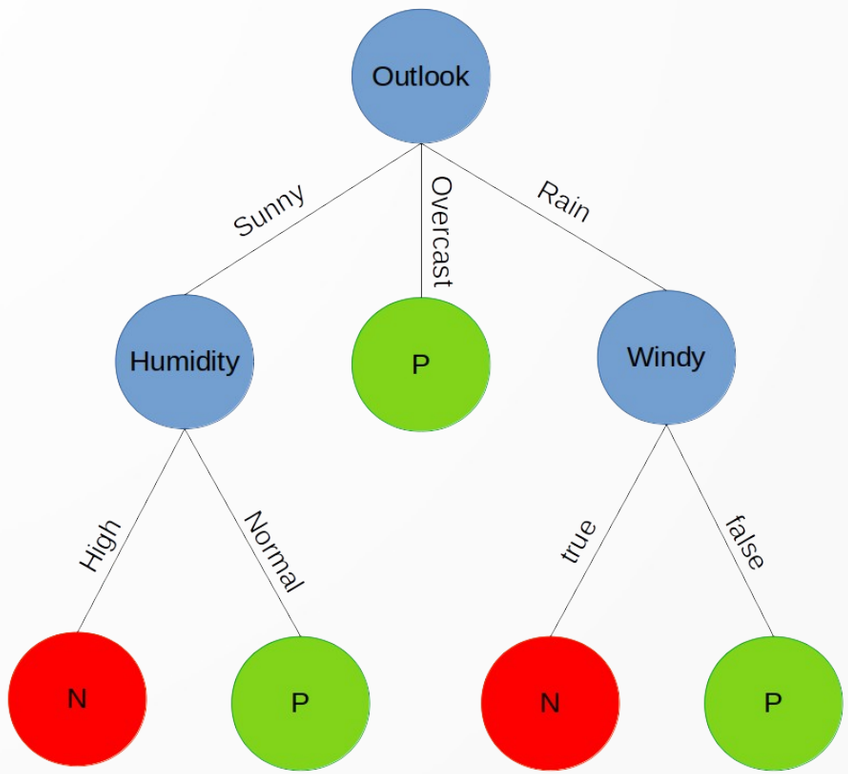
- **ID3 :** form a decision tree iteratively till all objects are classified.
- **C4.5, C5.0 :** enhancement of ID3 algorithm.
- **CART :** Classification and Regression Trees.

For the presentation : CART

From the data to the classification

No.	Outlook	Windy	Humidiy	Class
1	sunny	false	high	N
2	sunny	true	high	N
3	overcast	false	high	P
4	rain	false	normal	P
5	overcast	true	normal	P
...	...	...	...	...

Classifier



Quinlan, J. R. (1986). Induction of Decision Trees. Machine Learning, 1(1), 81–106. https://doi.org/10.1023/A:1022643204877

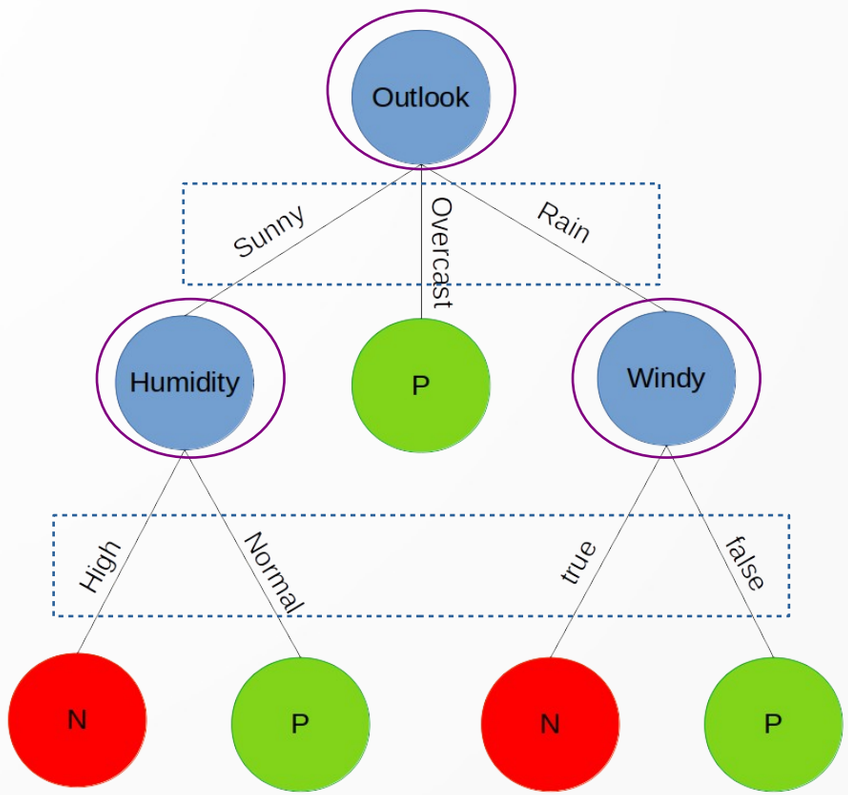
Features = Nodes

Features

No.	Outlook	Windy	Humidiy	Class
1	sunny	false	high	N
2	sunny	true	high	N
3	overcast	false	high	P
4	rain	false	normal	P
5	overcast	true	normal	P
...	...	...	...	...

Values

Classifier

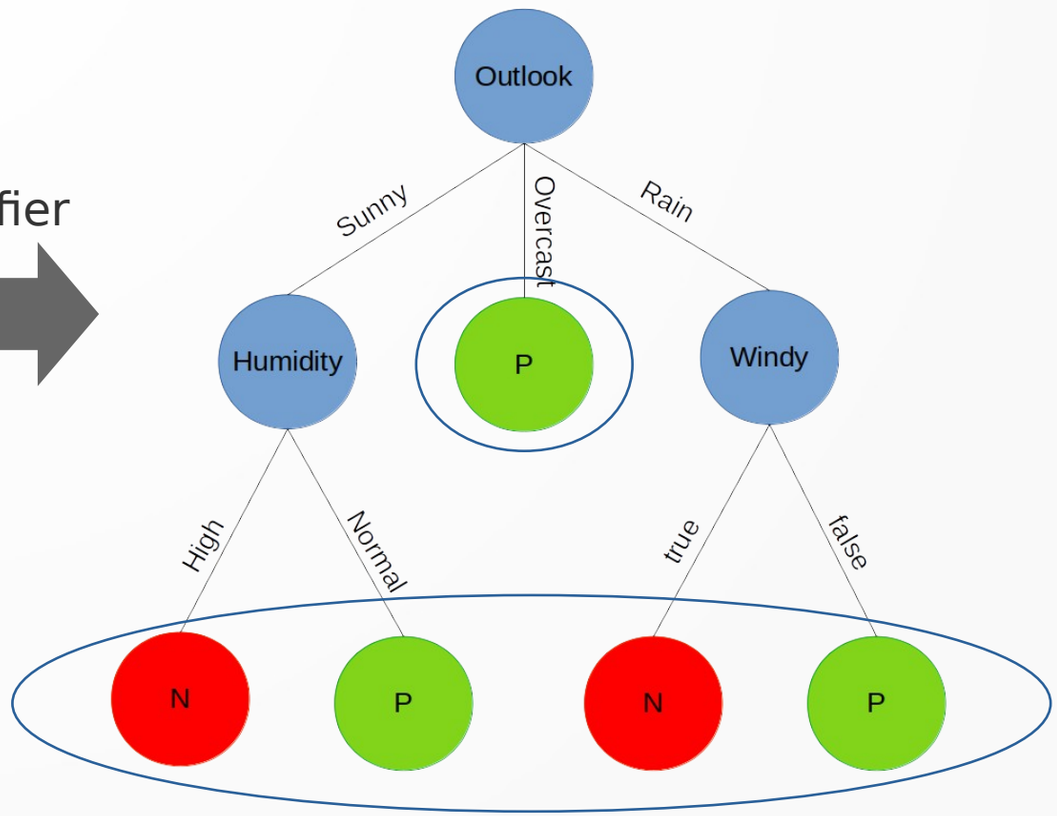


Labels = Leaves

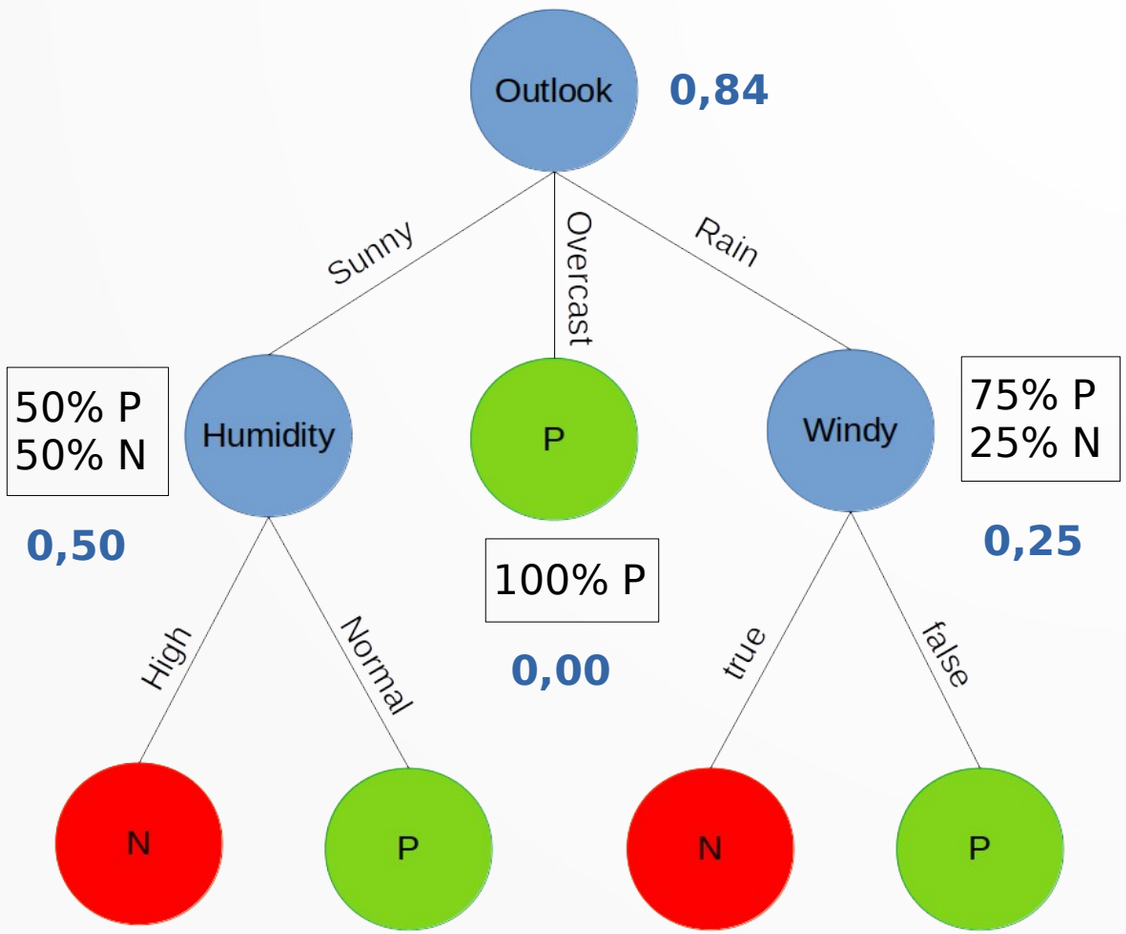
No.	Outlook	Windy	Humidiy	Class
1	sunny	false	high	N
2	sunny	true	high	N
3	overcast	false	high	P
4	rain	false	normal	P
5	overcast	true	normal	P
...	...	...	...	...

Labels

Classifier

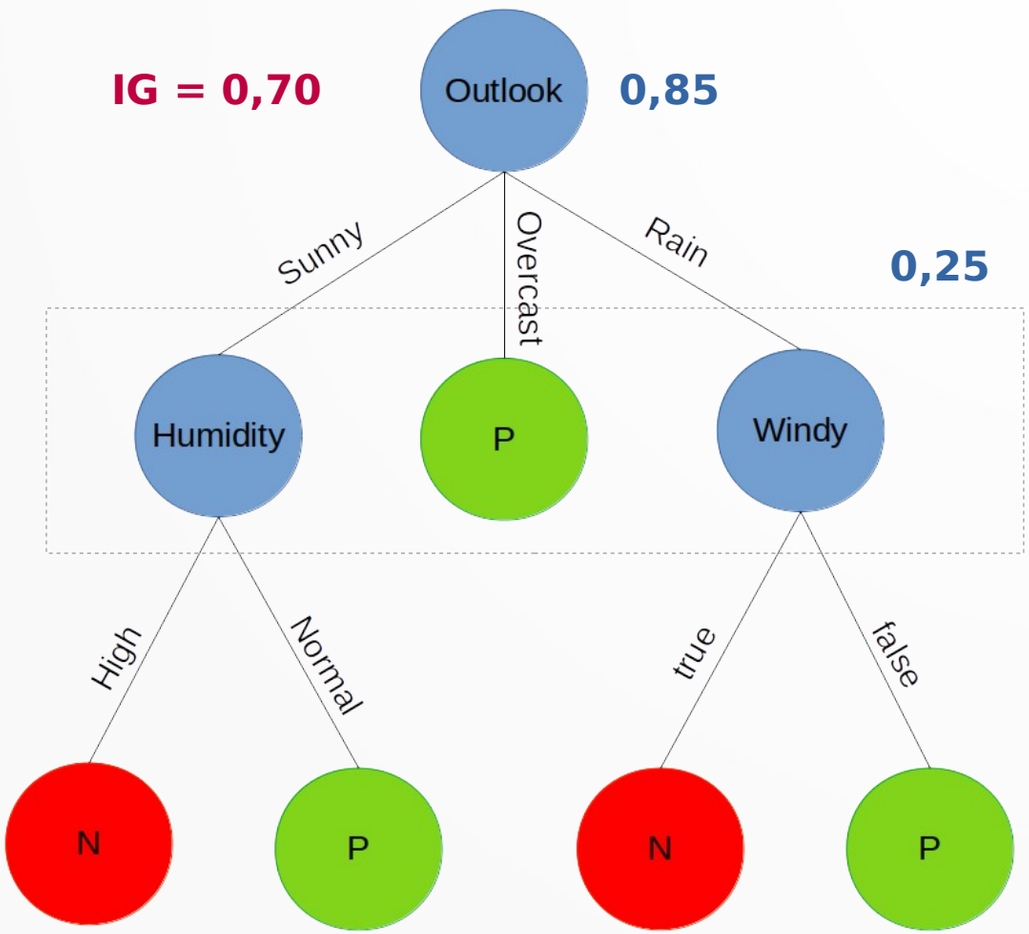


Gini impurity : potential question



- Impurity represents the chance to be incorrect if we randomly assign a label to an example.
- **Gini impurity = 1 - P(label)**
- Gini impurity : metrics between 0 and 1.
  - close to 0 stand for **purity**
  - close to 1 stand for **impurity**

Information Gain : reduce uncertainty



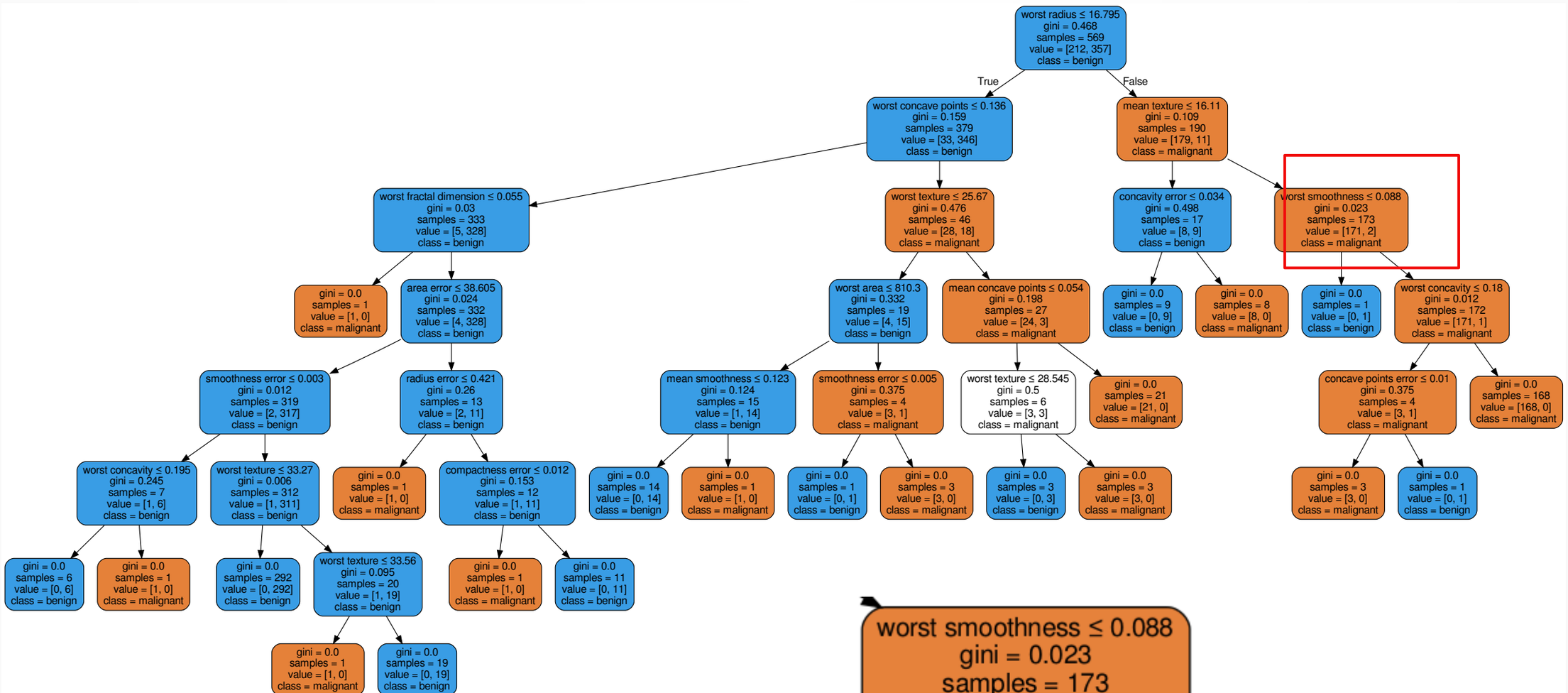
- Find the question that unmixed the most the labels.
- **IG = Gini[node] - AVG(Gini[childs])**
- **Goal :** to keep track of the best question with the best information gain (0 to 1).
- Root : « Is Humidity Normal ? » OR « Is Windy false ? » OR **« Is the Outlook Overcast ? »**

UCI dataset and Scikit-Learn visualization

Dua, D. and Karra Taniskidou, E. (2017). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

API design for machine learning software: experiences from the scikit-learn project, Buitinck et al., 2013.

Complex decision tree : Breast Cancer



worst smoothness ≤ 0.088
gini = 0.023
samples = 173
value = [171, 2]
class = malignant

References

- Bibliography
  - Quinlan, J. R. (1986). Induction of Decision Trees. Machine Learning, 1(1), 81–106. https://doi.org/10.1023/A:1022643204877
  - Stiglic, G., Kocbek, S., Pernek, I., & Kokol, P. (2012). Comprehensive decision tree models in bioinformatics. PLoS ONE, 7(3). https://doi.org/10.1371/journal.pone.0033812
  - Venkatesan, E., & Velmurugan, T. (2015). Performance Analysis of Decision Tree Algorithms for Breast Cancer Classification. Indian Journal of Science and Technology, 8(29), 1–8. https://doi.org/10.17485/ijst/2015/v8i
- Webography
  - Google developpers. Machine learning with Josh Gordon. 2016-2018 Youtube videos. https://www.youtube.com/watch?v=TF1yh5PKaqI&list=PLOU2XLYxmsIIuiBfYad6rFYQU_jL2ryal&index=10
  - API design for machine learning software: experiences from the scikit-learn project, Buitinck et al., 2013
  - Dr. William H. Wolberg, General Surgery Dept., University of Wisconsin, Clinical Sciences Center, Madison, WI 53792 wolberg@eagle.surgery.wisc.edu
  - Dua, D. and Karra Taniskidou, E. (2017). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.