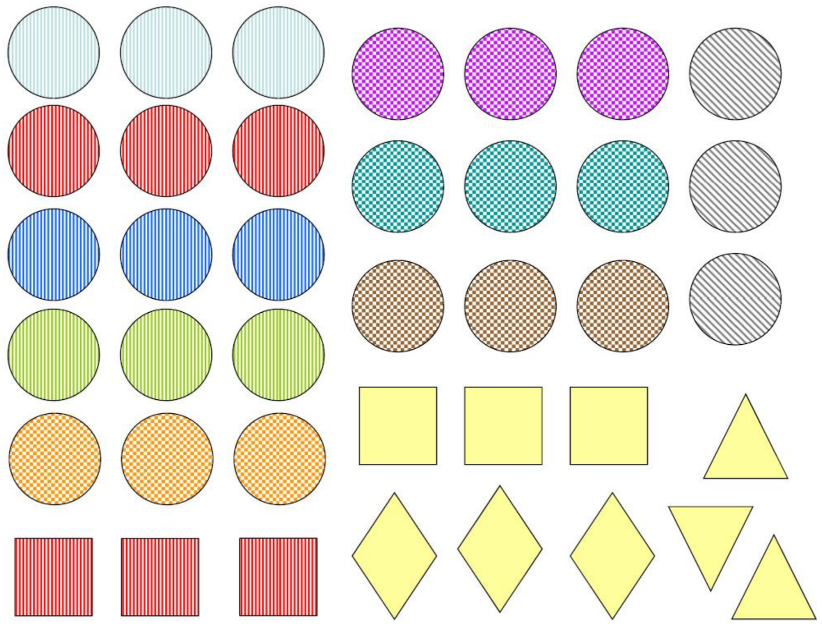
SMOOTH INTRODUCTION TO AUTOMATIC CLASSIFICATION

MasterClass Brasilia Nov2023

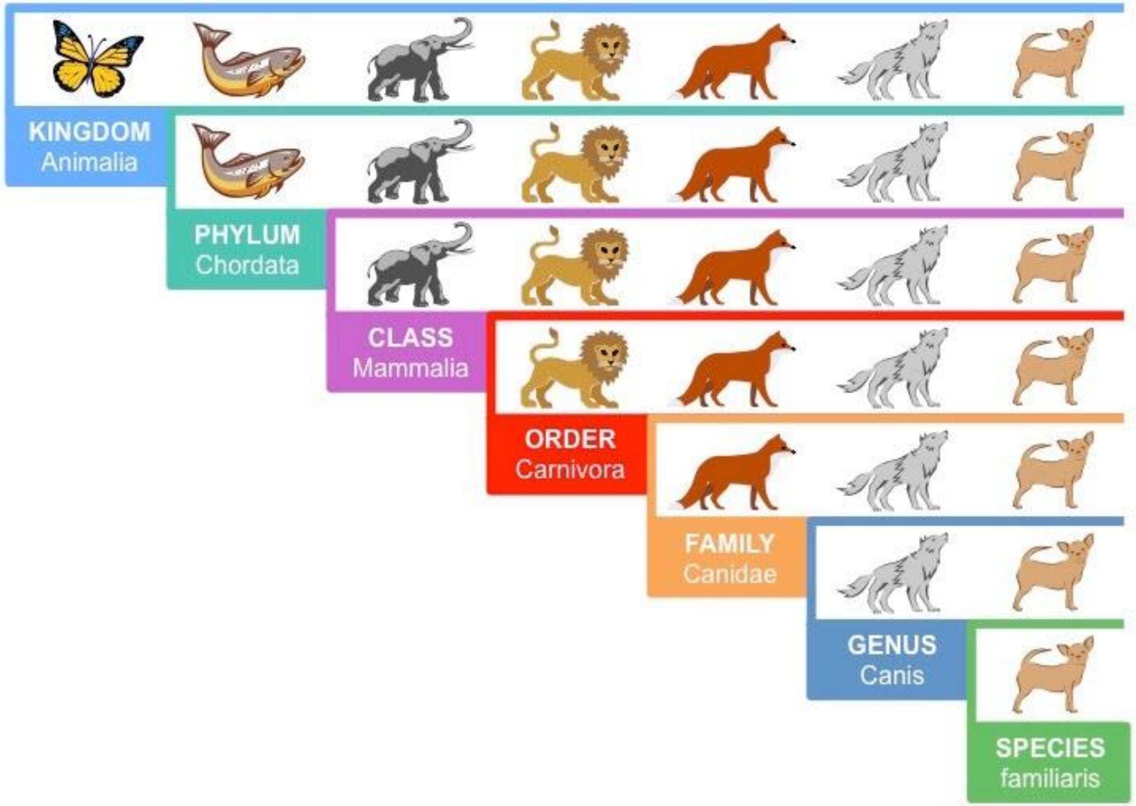
Marie-Dominique Devignes
LORIA, Nancy

CLASSIFICATION AS A HUMAN SKILL

Games for children in kindergarten



The tree of life: example de hierarchical classification



CLASSIFICATION : MATHEMATICAL DEFINITION

Definition : action of grouping objects into groups or classes on the basis of shared properties (shape, color, etc.)

- Distinguish from « ranking » (sometimes known as classification also) which consists of finding an order between objects (from largest to smallest for example)

Basic Concepts

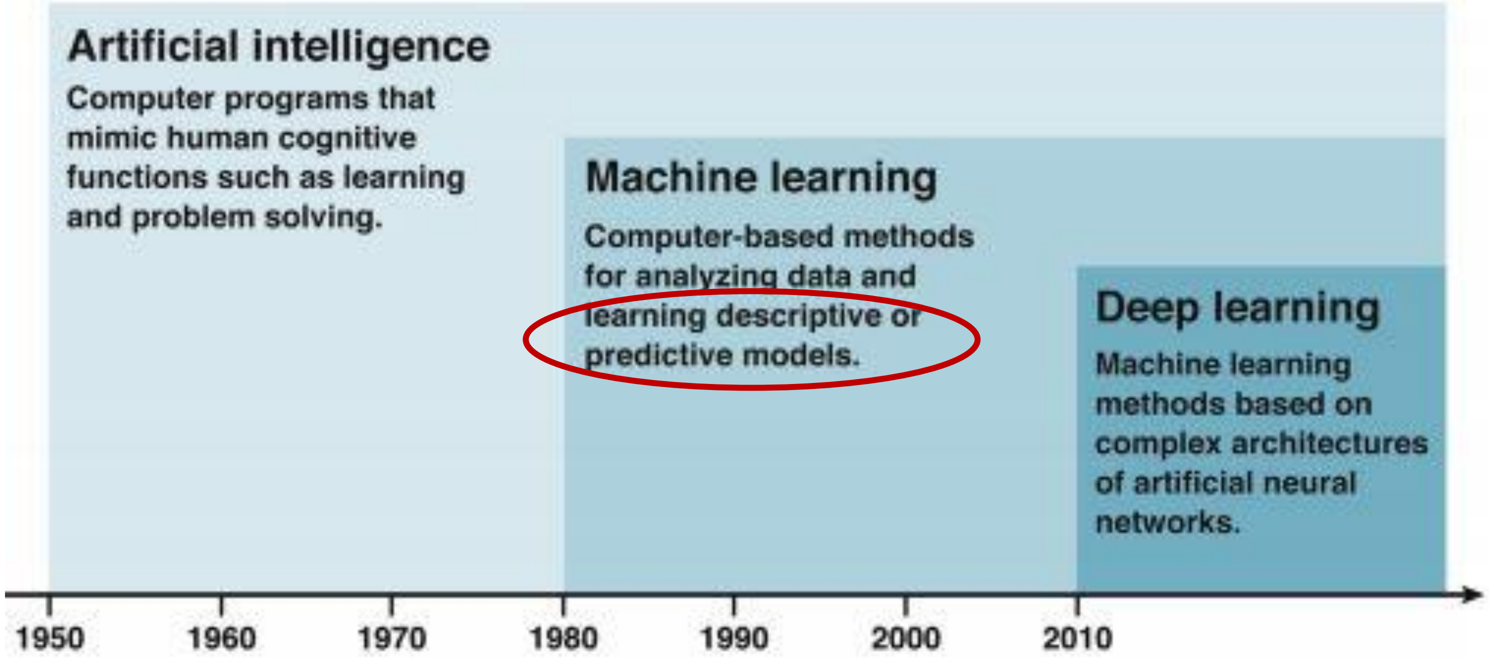
- **Objects** to be classified: organisms, protéines, molecules, patients …
  - Mathematical notation X_i, dataset = $\{X_1, X_2, X_3 \ldots\ldots X_n\}$
- **Properties** of these objects: « **descriptors** », « features », « **attributes** », « **variables** »
  - Classical mathematical representation as a vector $X = (d_1, d_2, d_3 \ldots. d_m)$
  - Descriptors can be numeric (quantitative : 53, 22, 3.5, etc.) ou nominal (qualitative : blue, yellow, etc.)
- **Classes** : groups of objects sharing common properties
  - Mathematical notation Y_i Ex: Y_0 = « healthy », Y_1 = « sick »
  - **Binary classifcation** : only 2 classes
  - **Multiclass classification** : objects can be grouped in more than two classes
- **Dataset :**
  - Matrix (Objects X Properties) : n rows, m columns (without Class) or (m+1) column (with Class)

n objects

m descriptors per object

2 classes or more

CLASSIFICATION AND ARTIFICIAL INTELLIGENCE



Classification tasks belong to artificial intelligence (AI)

- Finding classes in a datset = learning a « descriptive model »
- Finding the class of a new given object = requires a « predictive model »

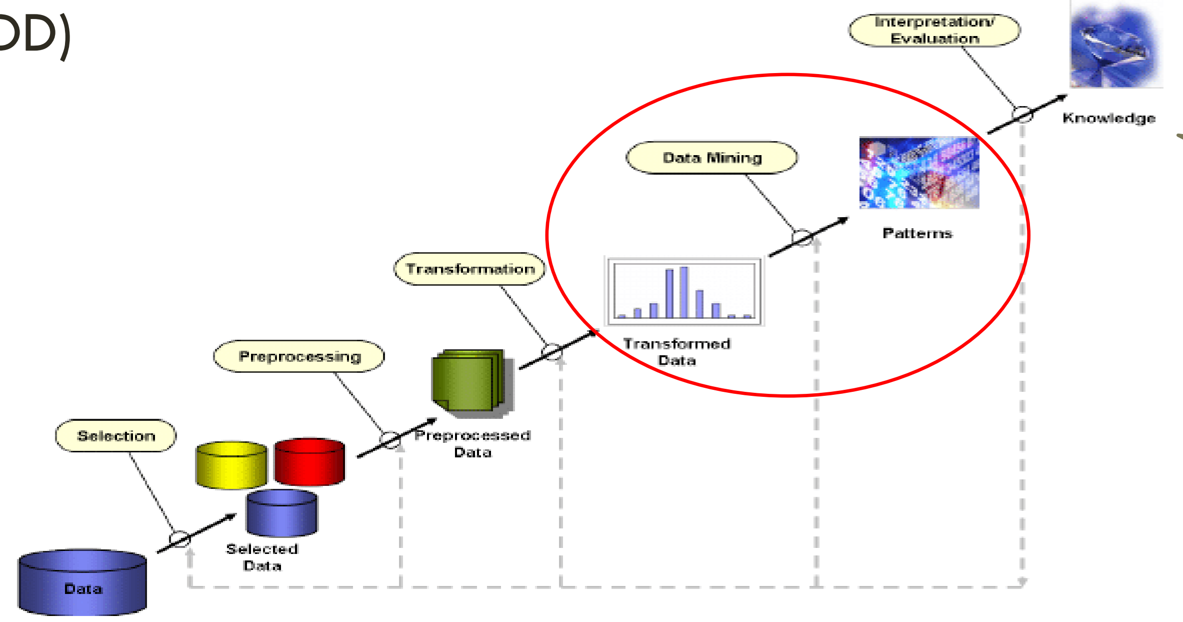
CLASSIFICATION AND KNOWLEDGE DISCOVERY

Classification is part of Data Mining (synonym of Machine Learning or Pattern Recognition)

- Attention: data mining is not restricted to « information retrieval » from huge database

Data Mining = one step in a global process known as Knowledge Discovery from Databases (KDD)

Importance of **Data Science**



Importance of **Knowledge Representation**

https://www.researchgate.net/figure/Steps-in-the-KDD-process_fig1_236373188

DEUX PARTS

I. Non supervised classification: « clustering »

II. Supervised classification

I. NON SUPERVISED CLASSIFICATION OR « CLUSTERING » (1/5)

Non supervised = the classes de each object is unknown

Goal : to find groups or classes in a set of objects. The set of groups with their shared features is a '**descriptive model**' for a dataset

- Members of a class must be as similar as possible = **cohesion of the class**
- Members of a class must be as different as possible from the members of the other classes = **discrimination** between classes.

1. Question of similarity or distance between objets
   - By défault : euclidean distance between feature vectors

For two points P and Q with coordinates (p1, p2) and (q1, q2),

$$d(P,Q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2}$$

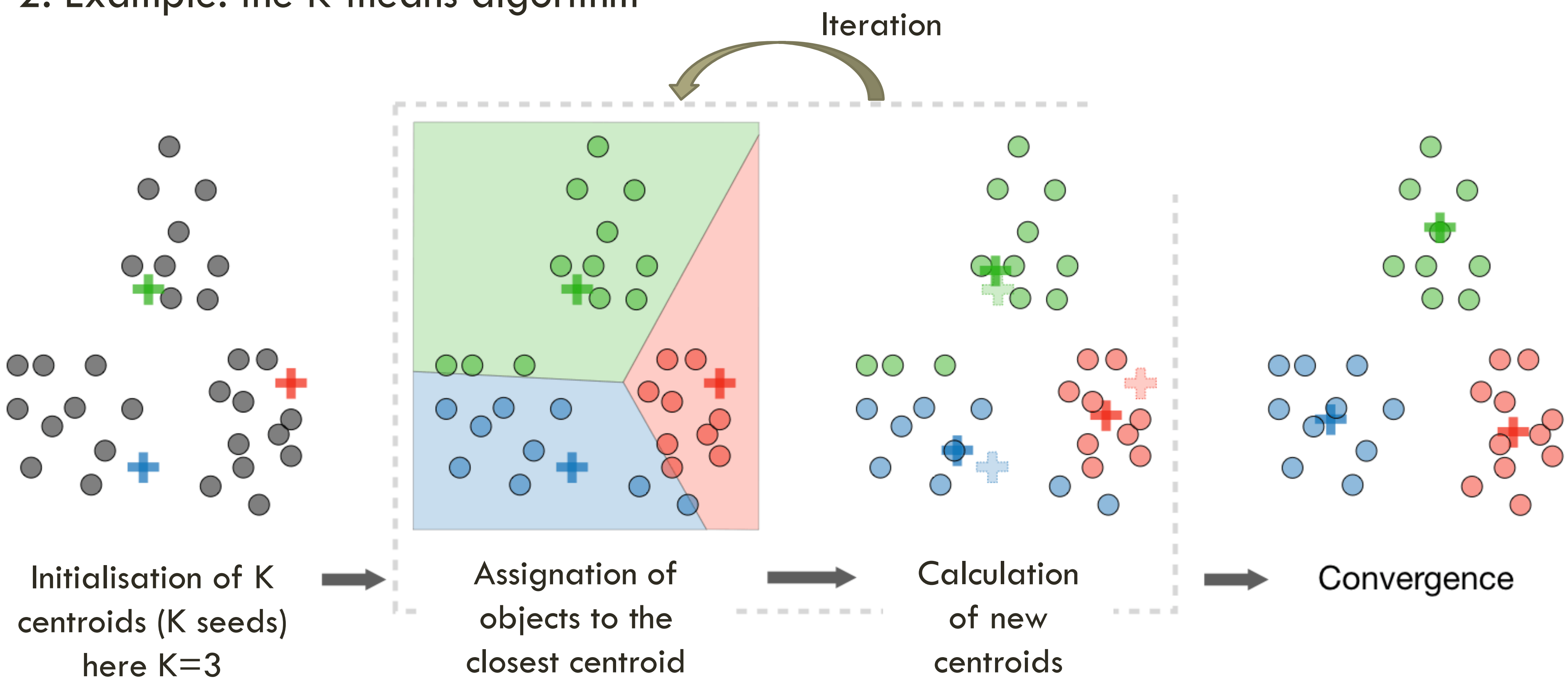
Generalisation for two points X_i et X_j de coordonnées $(d_1^i, d_2^i, \dots d_m^i)$ et $(d_1^j, d_2^j, \dots d_m^j)$.

$$d(X_i, Xj) = \sqrt{\sum_{z=1}^{m} (d_z^i - d_z^j)^2}$$

- Alternative : some classification algorithms can take as input a similarity matrix between pairs of objects

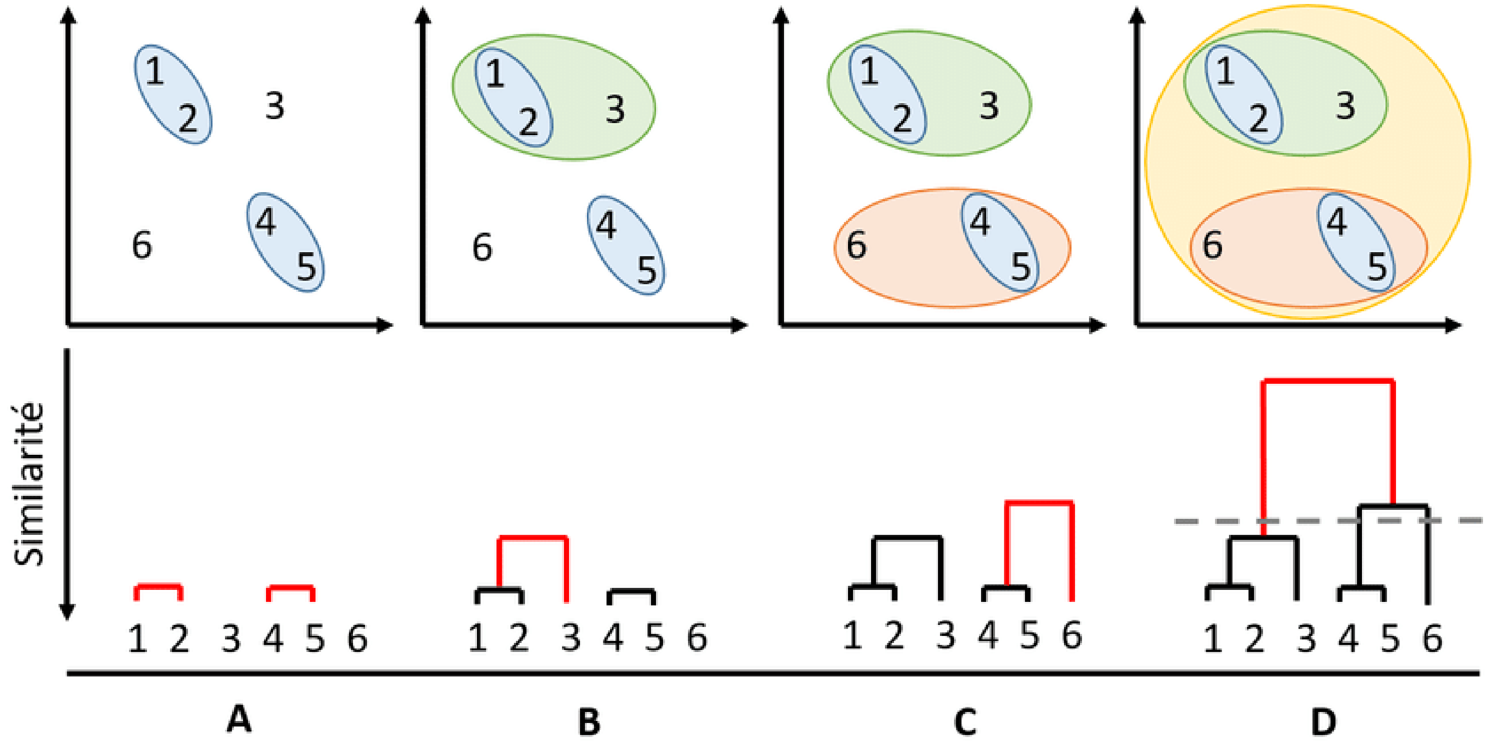
NON SUPERVISED CLASSIFICATION OR « CLUSTERING » (2/5)

2. Example: the K-means algorithm



NON SUPERVISED CLASSIFICATION OR « CLUSTERING » (3/5)

3. Example: algorithme for hierarchical ascending clustering (HAC)



From C. Rieux's thesis https://www.researchgate.net/figure/Exemple-de-Classification-Ascendante-Hierarchique-CAH-Les-6-objets-sont-finalement_fig41_325956424

NON SUPERVISED CLASSIFICATION OR « CLUSTERING » (4/5)

4. Searching the optimal number of clusters (K) by the 'elbow' method

'Elbow' : inflection point of the cohesion curve

Cohesion coefficient = within-cluster sum of squares (wss)

- For each cluster, one computes the sum of the squares of the distances between eah pair of objects
- These 'within-cluster' sums are summed for all clusters.

The cohesion coefficient is calculated and plotted for various K values

The inflection point is identified visually.

https://ichi.pro/fr/methode-silhouette-mieux-que-la-methode-elbow-pour-trouver-des-clusters-optimaux-61080390822033

NON SUPERVISED CLASSIFICATION OR « CLUSTERING » (5/5)

5. Searching the optimal number of clusters (K) by the 'silhouette' method

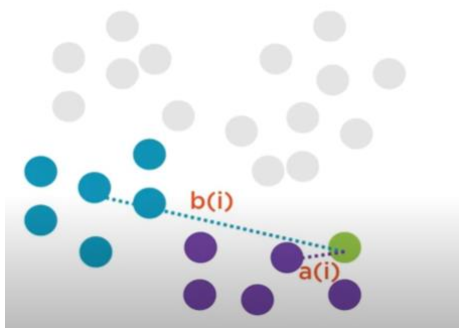
Optimal K value is determined at the maximum of average silhouette coefficient

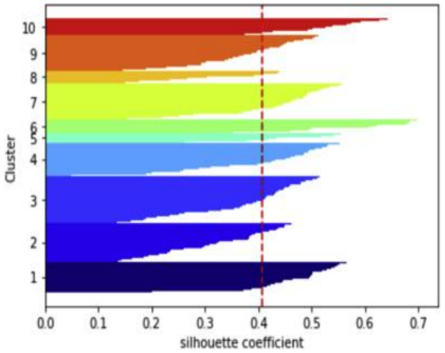
Silhouette : compromise between within-cluster cohesion and inter-cluster discrimination.

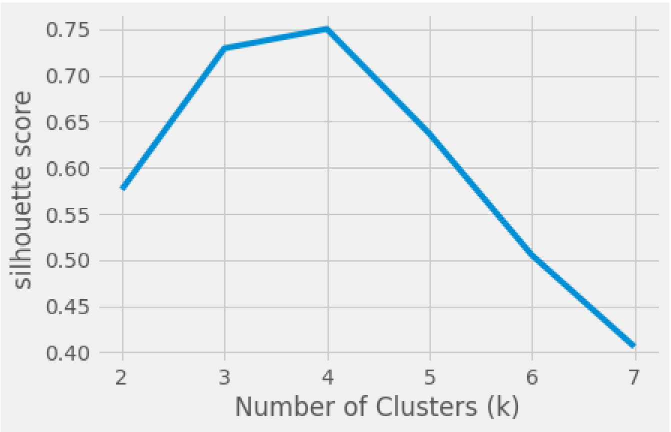
- For each object X assigned to cluster C, calculate the mean $\boldsymbol{a_X}$ of the distances of X to the other objects of C: (within-cluster mean),
- And the mean $\boldsymbol{b_X}$ of the distances of X to all objects of the closest cluster to C
- Silhouette coefficient of X, $Silh(X) = \frac{b_X - a_X}{Max(a_X, b_X)}$
- $Silh(X)$ = -1 if $\boldsymbol{b_X}$ = 0 -> poor discrimination between clusters
- $Silh(X)$ = +1 if $\boldsymbol{a_X}$ = 0 -> excellent cohesion
- Average all silhouette coefficient on all objects of each cluster, and then on all clusters

Plot the average slihouette coefficient for various K values

- Optimal K value is for maximal average silhouette coefficient







https://ichi.pro/fr/methode-silhouette-mieux-que-la-methode-elbow-pour-trouver-des-clusters-optimaux-61080390822033

https://www.sciencedirect.com/topics/computer-science/silhouette-coefficient

II. SUPERVISED CLASSIFICATION

Supervised = The class is known for each object in a training dataset

Goal : learn how to classify new objects in known classes

Method : construct a '**predictive model**' by training a '**classifier**' (neural network, Support Vector Machines, decision tree, random forests), then use this classifier to assign a class to new objects.

1. Preparing the training dataset

Objects X_i represented by a feature vector and by their class Y_i

Training dataset = $\{(X_1,Y_1), (X_2,Y_1), (X_3,Y_0), \ldots.. (X_n, Y_0)\}$

- If binary problem then Y has only two values, e.g. True and False, or + and -, or 1 and 0.
- If multiclass problem, then Y takes as many values as classes.

Attributes Objects	d_1	d_2	d_3	…	Class (+ or -)
X1	3,5	0	Jaune	…	+
X2	57,9	1	Jaune	…	-
X3	2,8	0	Jaune	…	+
…	…	…	…	…	
Xn	67,3	0	Vert	…	-

Example of training dataset with two classes + and -

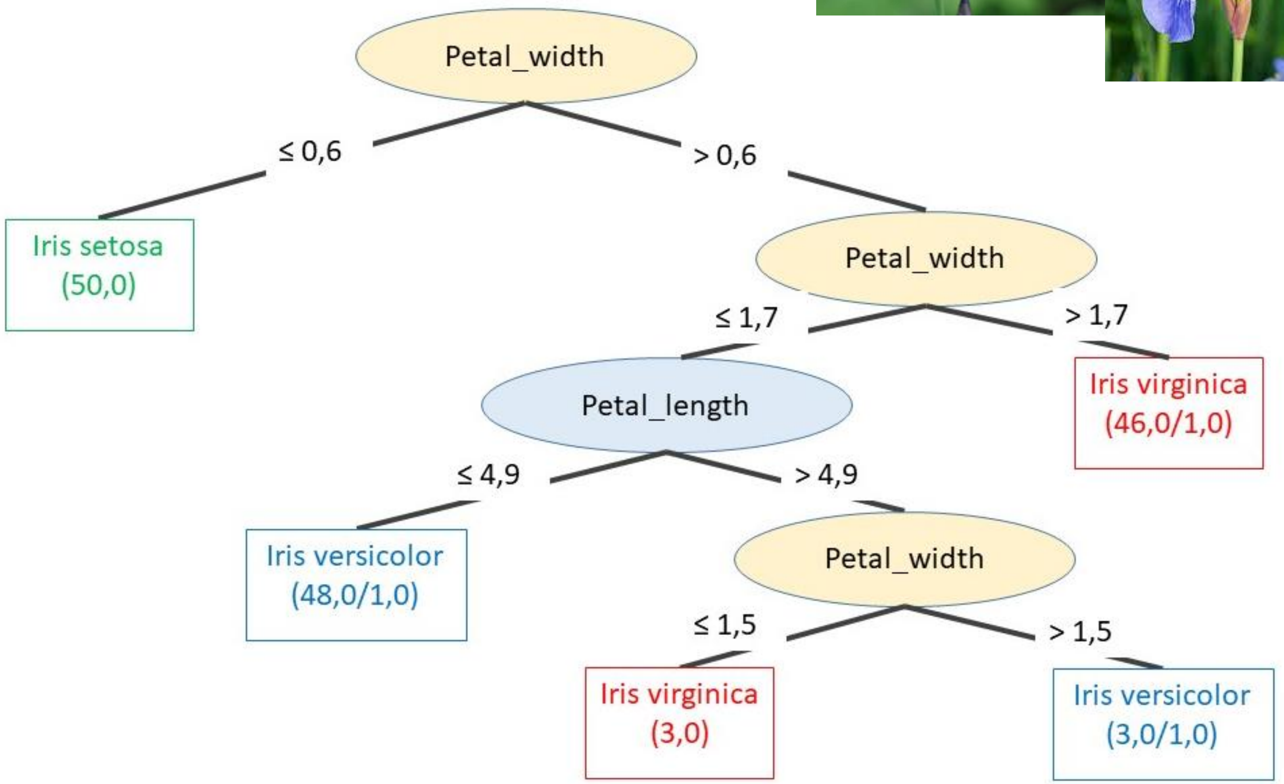
EXAMPLE OF ALGORITHM FOR SUPERVISED CLASSIFICATION : DECISION TREE

Dataset Iris.2D : 3 iris species setosa, virginica, versicolor = 3 classes

Training dataset :

- 50 samples or 'instances' of each class.
- 2 descriptors for each sample in addition to class : petal_width and petal_length

The decision tree is calculated here with the J48 algorithm.



SUPERVISED CLASSIFICATION: MODEL EVALUATION

Cross-validation

- The training dataset is divided in 10 equivalent subsets (same number of instances and same proposition of classes as in the total dataset = 10 'folds'
  - Example 150 iris samples with 50 samples from each class -> 10 folds of 15 iris with 5 samples from each class in each fold.
- One fold is kept apart (test set) and a decision tree is built with the 9 other folds. Then the decision tree is tested on the test fold (metrics described below)
- The process (training + test) is repeated 9-times with the 9 other folds taken consecutively as test set -> in total 10 decision trees are built and tested.
  - All examples have been tested at the end.
- The results of the test are compiled in a confusion matrix

```
======Confusion Matrix =========
  a    b    c   ← classified as
 49    1    0    a = Iris setosa
  0   47    3    b = Iris versicolor
  0    2   48    c = Iris virginica
```

Example of quality metrics

Well-classified ratio

(when classes are well balanced)

Sum of counts in the diagonal divided by the total number of samples

$$\frac{49+47+48}{150} = 0{,}96$$

OTHER EVALUATION METRICS

Kappa statistics : % of well-classified samples corrected by the random distribution across classes (propensity) as a function of effectifs in each class.

- P0 = sum of the diagonal (well classified) $P0 = p_{11} + p_{22}$
- Pe = probability to obtain this distribution just by chance $Pe = p_{1.}\,p_{.1} + p_{2.}\,p_{.2}$

$$\text{Kappa} = \frac{P0 - Pe}{1 - Pe}$$

== Probability matrix ==

$$\begin{array}{cc|cc} 1 & 2 & \leftarrow \text{classified as} & \\ \hline p_{11} & p_{12} & p_{1.} & 1 \\ p_{21} & p_{22} & p_{2.} & 2 \\ \hline p_{.1} & p_{.2} & & \end{array}$$

Precision, recall and F-measure (for a given class noted +)

- True positive count TP
- False positive count FP
- Precision $\frac{TP}{TP+FP}$ Recall or sensitivity $\frac{TP}{TP+FN}$
- F-mesure $\frac{2TP}{2TP+FP+FN}$

True Class	Predicted class +	Predicted class -
+	TP True Positive	FN False Negative Type II error
-	FP False Positive Type I error	TN True Negative

MCC Matthews correlation coefficient

- More complicated but takes into account also the TN
- Less optimistic than F1 measure

Area under the ROC curve (see next slide)

Sam ples	True Class	Predicted Class	Score*
X6	+	+	0,99
X30	+	+	0,99
X7	-	-	0,75
X43	-	+	0,75
X37	+	-	0,64
X12	+	+	0,64
X103	-	+	0,33
…			…

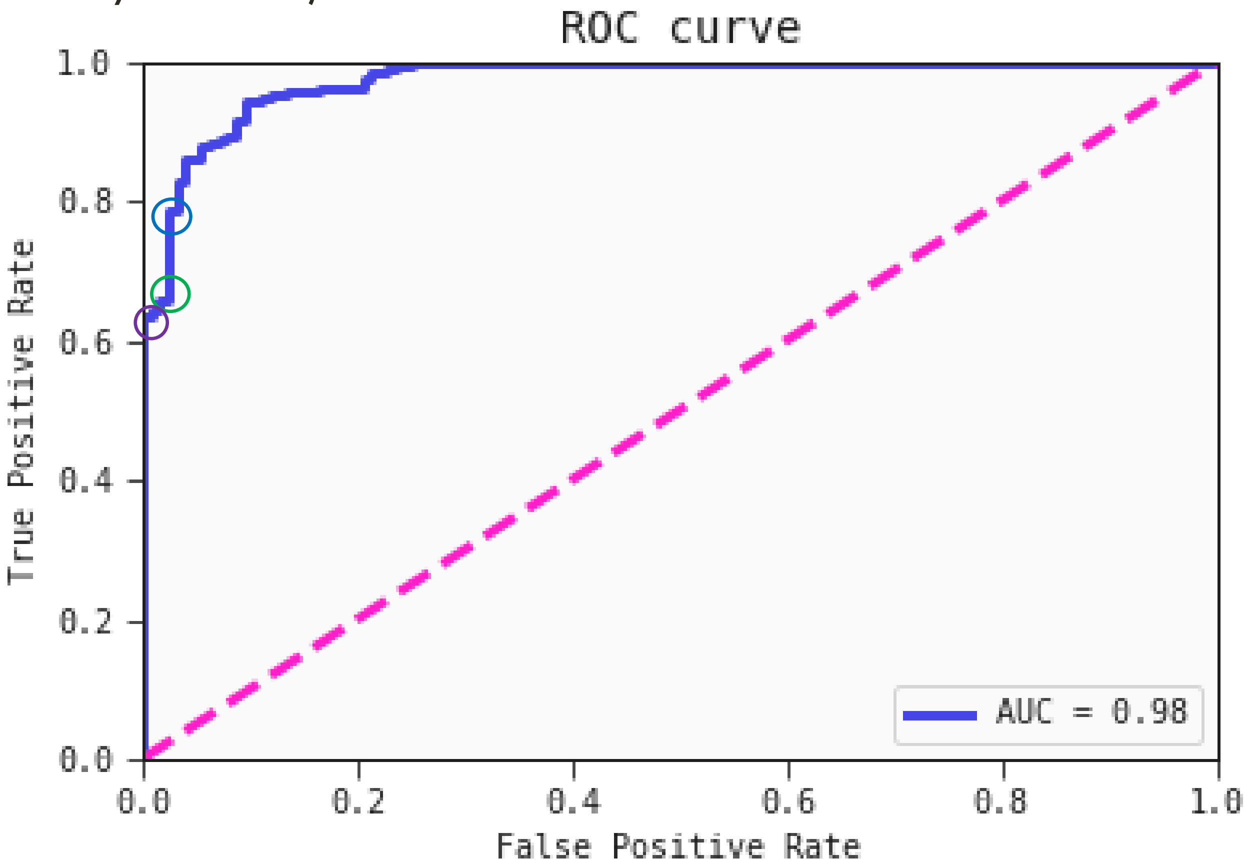
* The score reflects the degree to which the instance belongs to its class according to the prediction. For a decision tree, it can be the % of samples of the majoritary class in the leaf.

Threshold 0,9 -> confusion matrix at 0.9 -> (TPR, FPR)

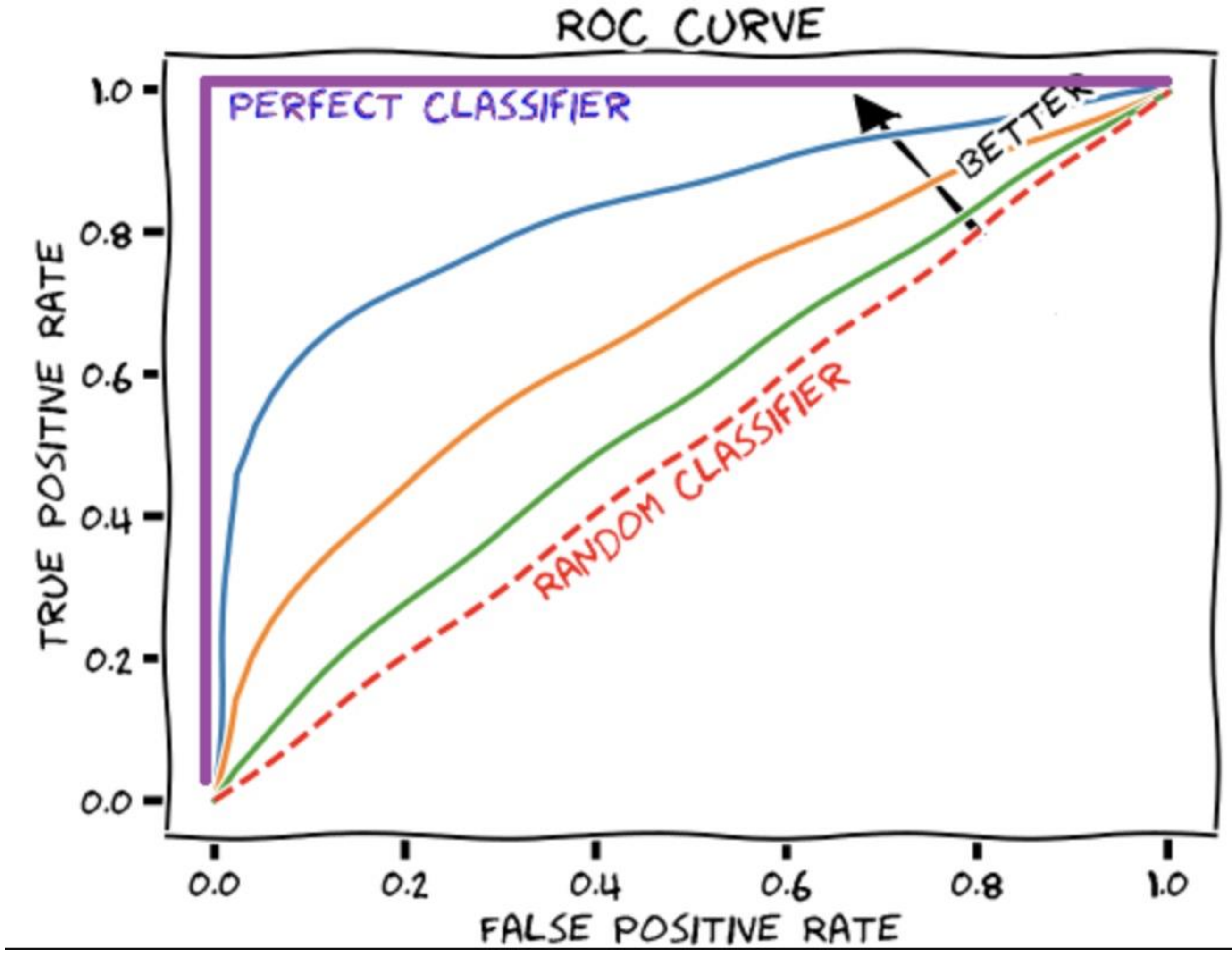
Threshold 0,7 -> confusion matrix at 0.7 -> (TPR, FPR)

Threshold 0,6 -> confusion matrix at 0.7 -> (TPR, FPR)

TPR = TP/(TP+FN)
(sensitivity or recall)



FPR = FP/(FP+TN)
(1 - specificity)



The **larger** the AUC (Area Under the ROC Curve), the **better** the classifier. Max AUC = 1

https://stats.stackexchange.com/questions/523760/regarding-roc-curve-of-good-classifier-why-tpr-and-fpr-both-increase?noredirect=1&lq=1

TO GO FURTHER

Download and use WEKA !

Software :

https://waikato.github.io/weka-wiki/downloading_weka/

User guide :

https://www.cs.waikato.ac.nz/~ml/weka/book.html

https://www.cs.waikato.ac.nz/ml/weka/Witten_et_al_2016_appendix.pdf