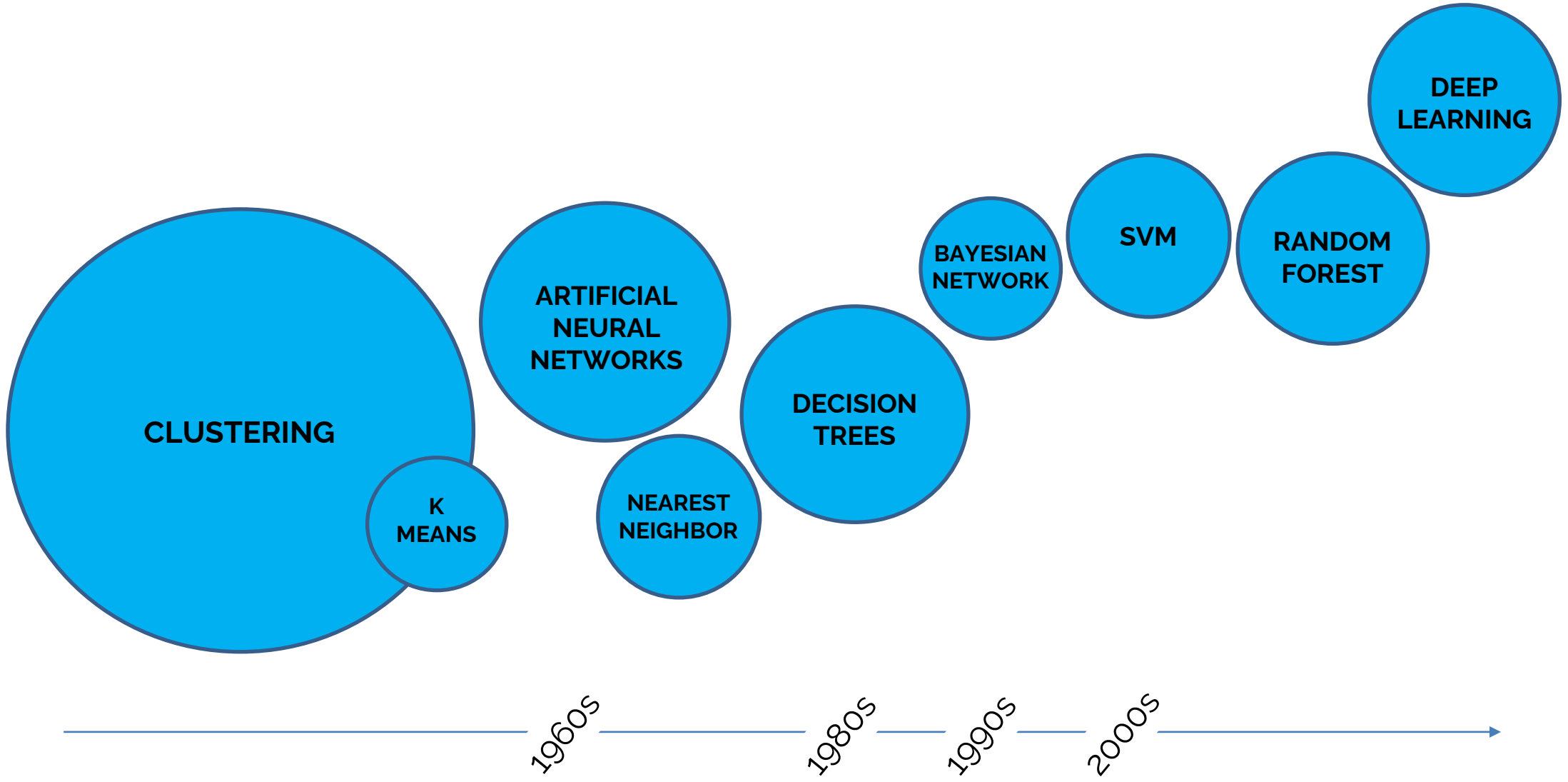
Machine learning: basi e sue applicazioni

Christian Salvatore
Scuola Universitaria Superiore IUSS Pavia

christian.salvatore@iusspavia.it


CLUSTERING
K MEANS
ARTIFICIAL NEURAL NETWORKS
NEAREST NEIGHBOR
DECISION TREES
BAYESIAN NETWORK
SVM
RANDOM FOREST
DEEP LEARNING
1960s
1980s
1990s
2000s

CLASSIFICATION

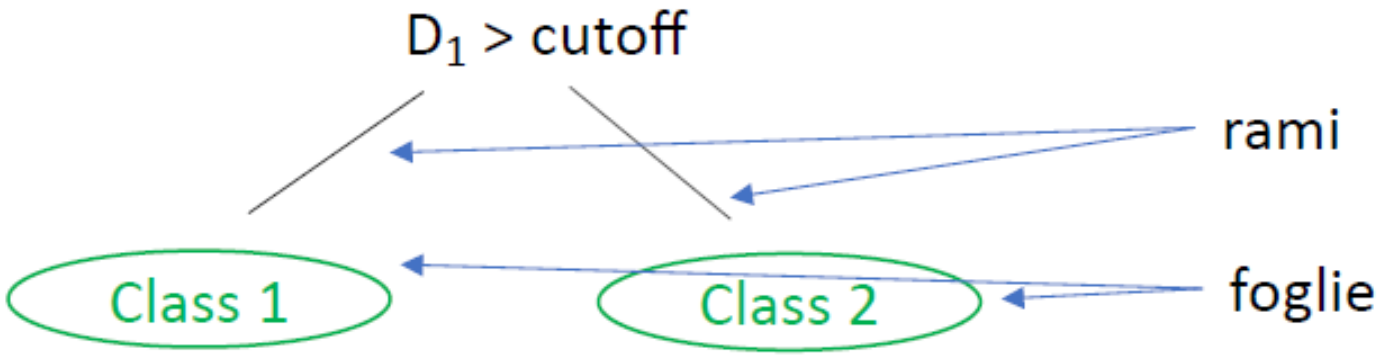
DECISION TREE

Decision Tree

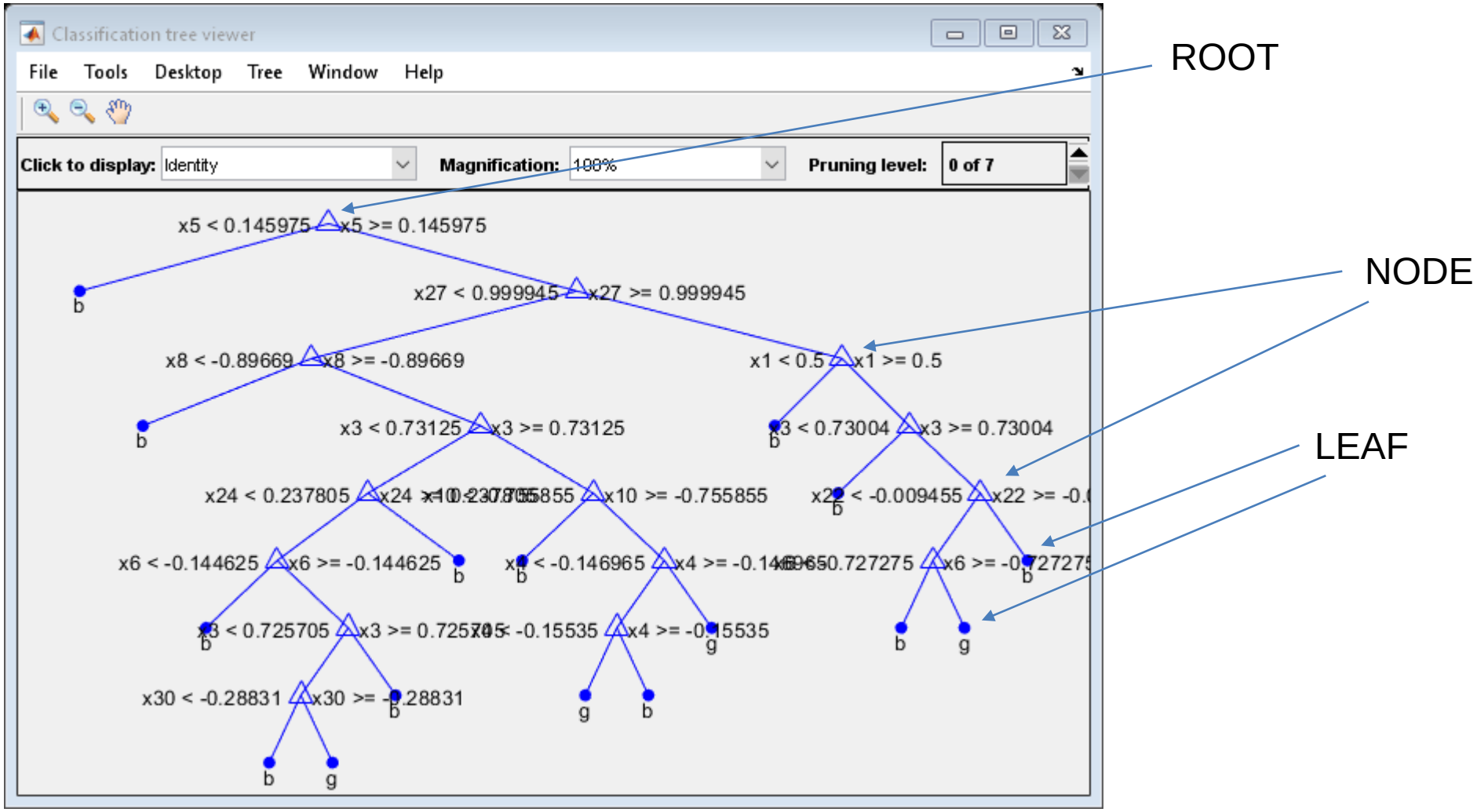
A decision tree or classification tree is a binary tree in which each node divides patterns based on a criterion on a single feature (or dimension).

In these tree structures, the leaf nodes represent the class labels and the branches represent the set of characteristics that lead to those classifications.

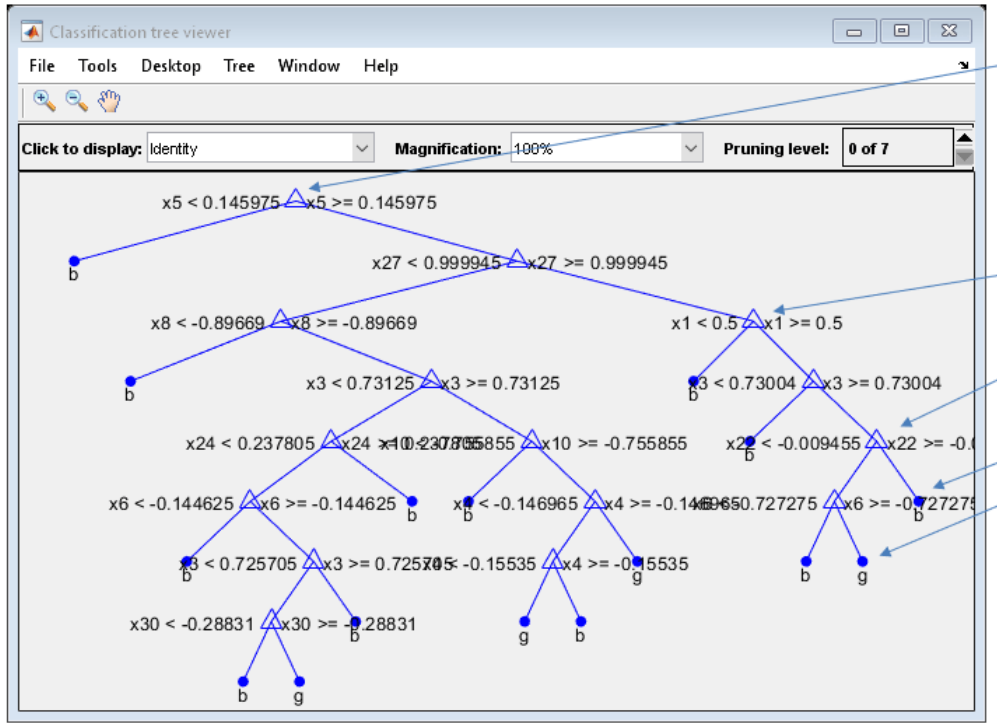
Tree algorithms in which the target variable can take a discrete set of values are called classification trees, those in which the target variable can take continuous values (usually real numbers) are called regression trees.



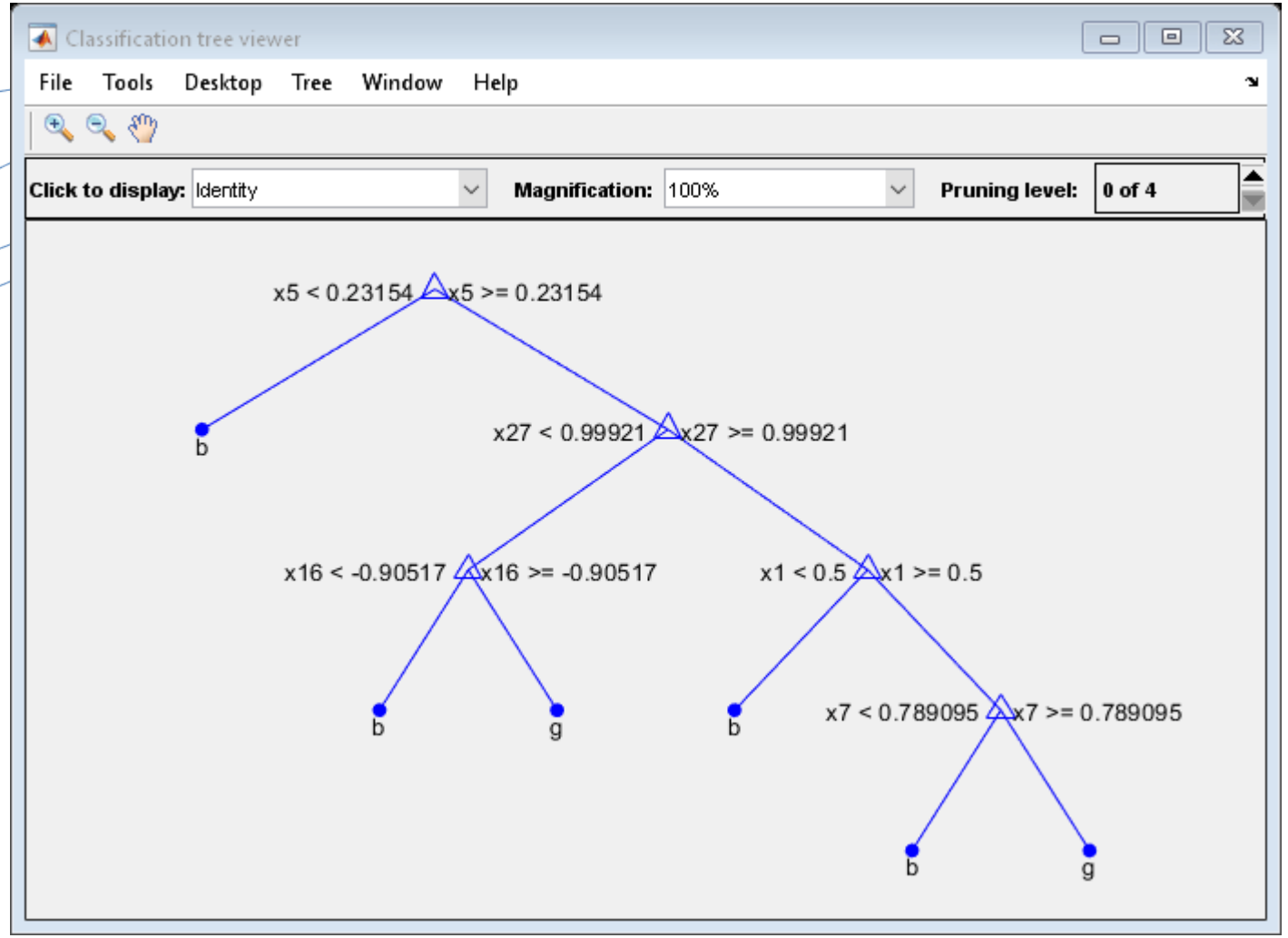
Decision Tree



Decision Tree



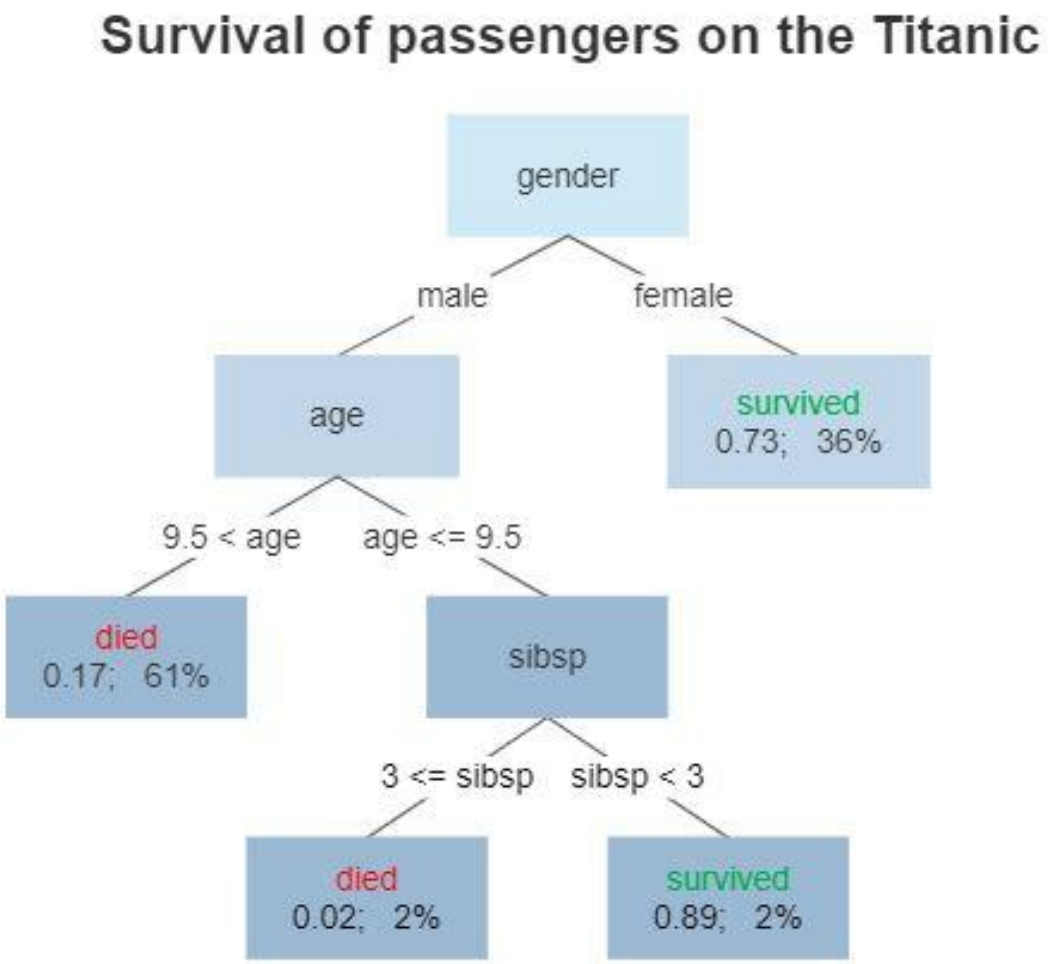
ROOT



The predicate that is associated with each internal node (on the basis of which the data is distributed) is called the split condition.

It is useful to set a minimum number of samples required to bolt.

Decision Tree



Decision Tree | Objective Function

In a good classification tree, leaf nodes should be as pure as possible (i.e., contain only data that belongs to a single class).

A parameter that defines the impurity criterion is the Gini index or Gini impurity, which can be considered as an objective function.

> It is a measure of how often a randomly chosen element from the set would be incorrectly labeled if it was randomly labeled according to the distribution of labels in the subset

$$\mathrm{I}_G(p) = \sum_{i=1}^{J} \left(p_i \sum_{k \neq i} p_k \right)$$

Decision Tree | Objective Function

The information gain is estimated as the difference between the impurity of the parent node and the sum of the impurities of the child nodes: the lower the impurity of the child nodes, the higher the information gain (decrease in Gini index).

The Gini index reaches its minimum (zero) when the node belongs to a single category. Intuitively, Gini's impurity can be considered as a criterion for minimizing the probability of an incorrect classification.

Decision Tree | Objective Function

Another parameter that defines the impurity criterion is the entropy index

$$I_E(i) = -\sum_{j=1}^{m} f(i,j) \log f(i,j)$$

The Gini index and the entropy index are the parameters that are usually used to guide the construction of the tree.

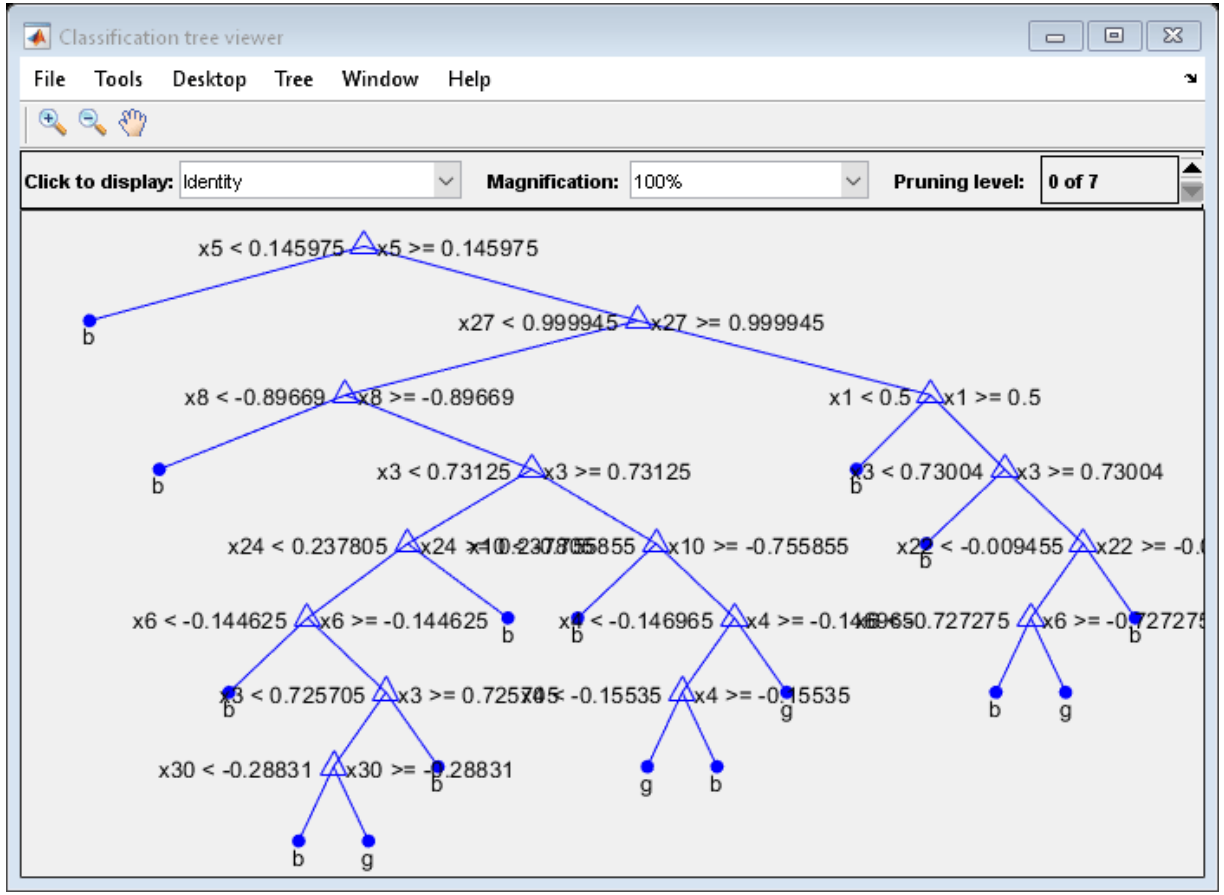
Decision Tree | Objective Function

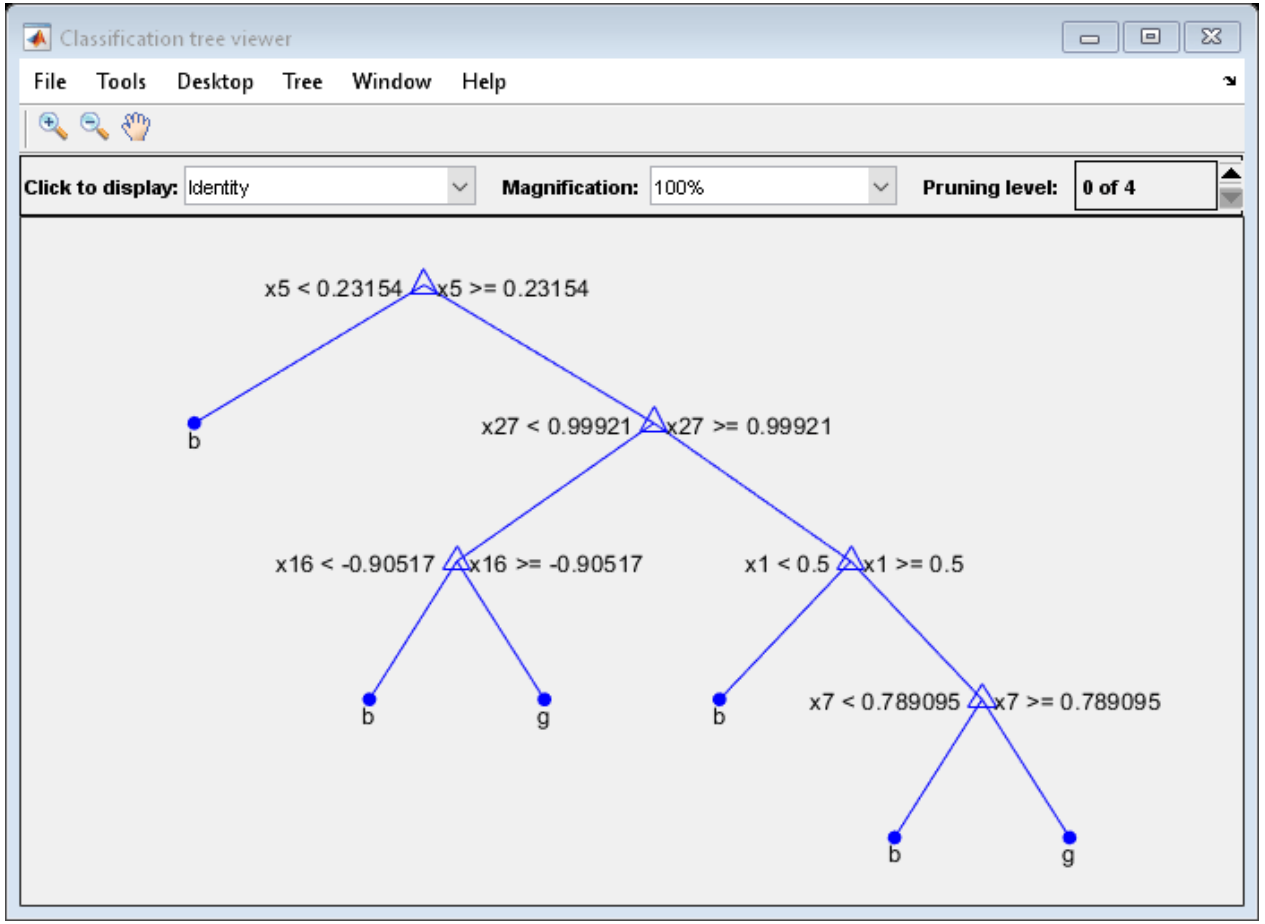
The classification error rate that is used to perform an optimization of the tree is known as the halting or **pruning** process, in order to determine the maximum depth of the tree (max depth) .

The growing the depth of a tree (or its size) does not directly affect the goodness of the model.

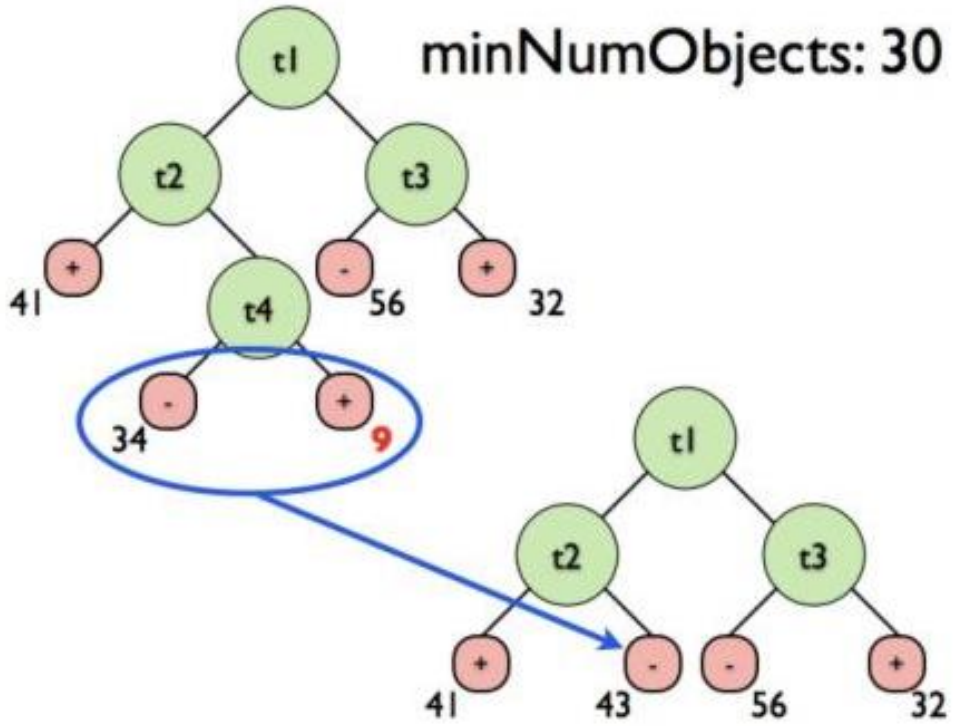
In fact, an excessive growth of the tree size could only lead to a disproportionate increase in computational complexity compared to the benefits regarding the accuracy of the predictions / classifications.

Decision Tree | Objective Function



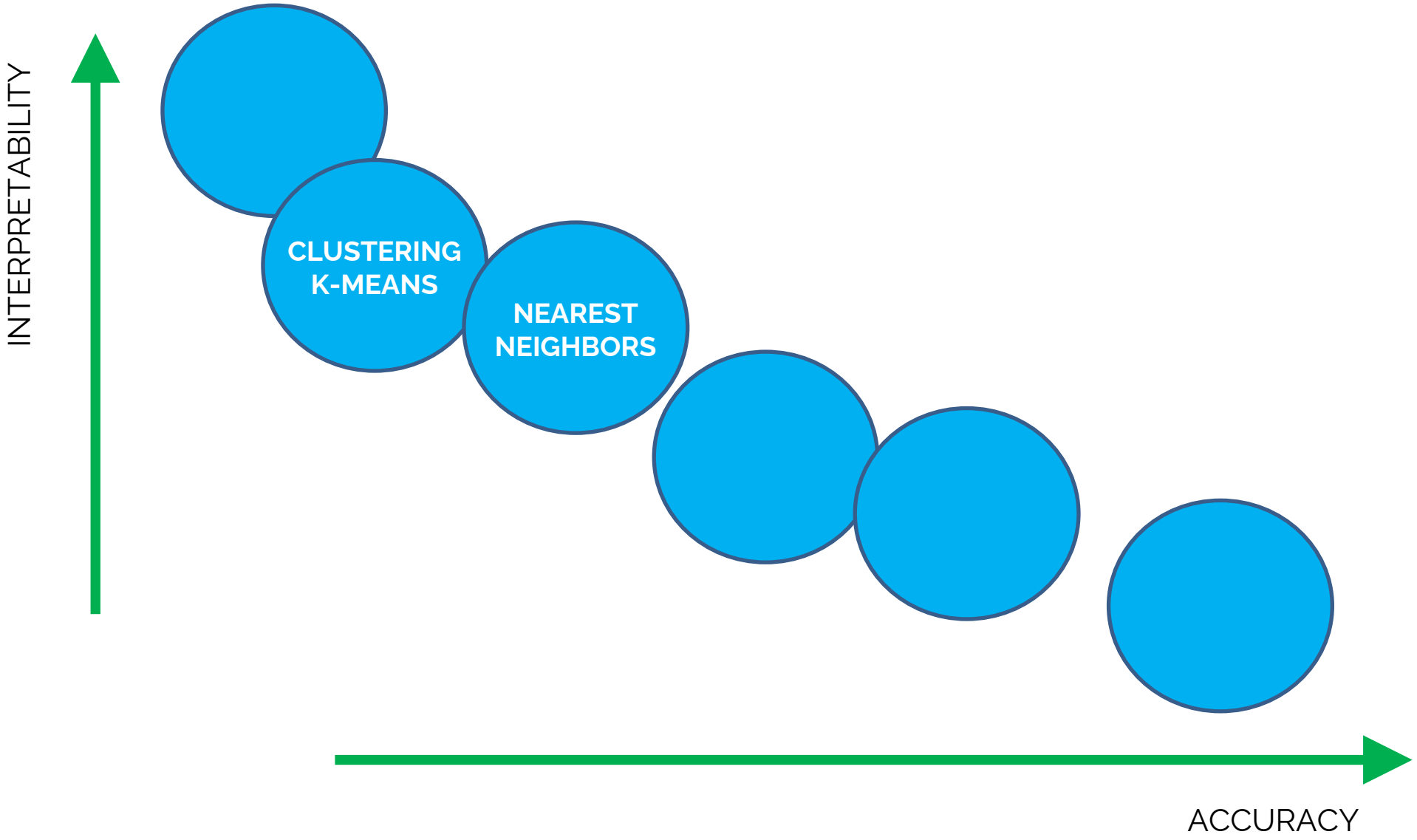


Decision Tree | Objective Function

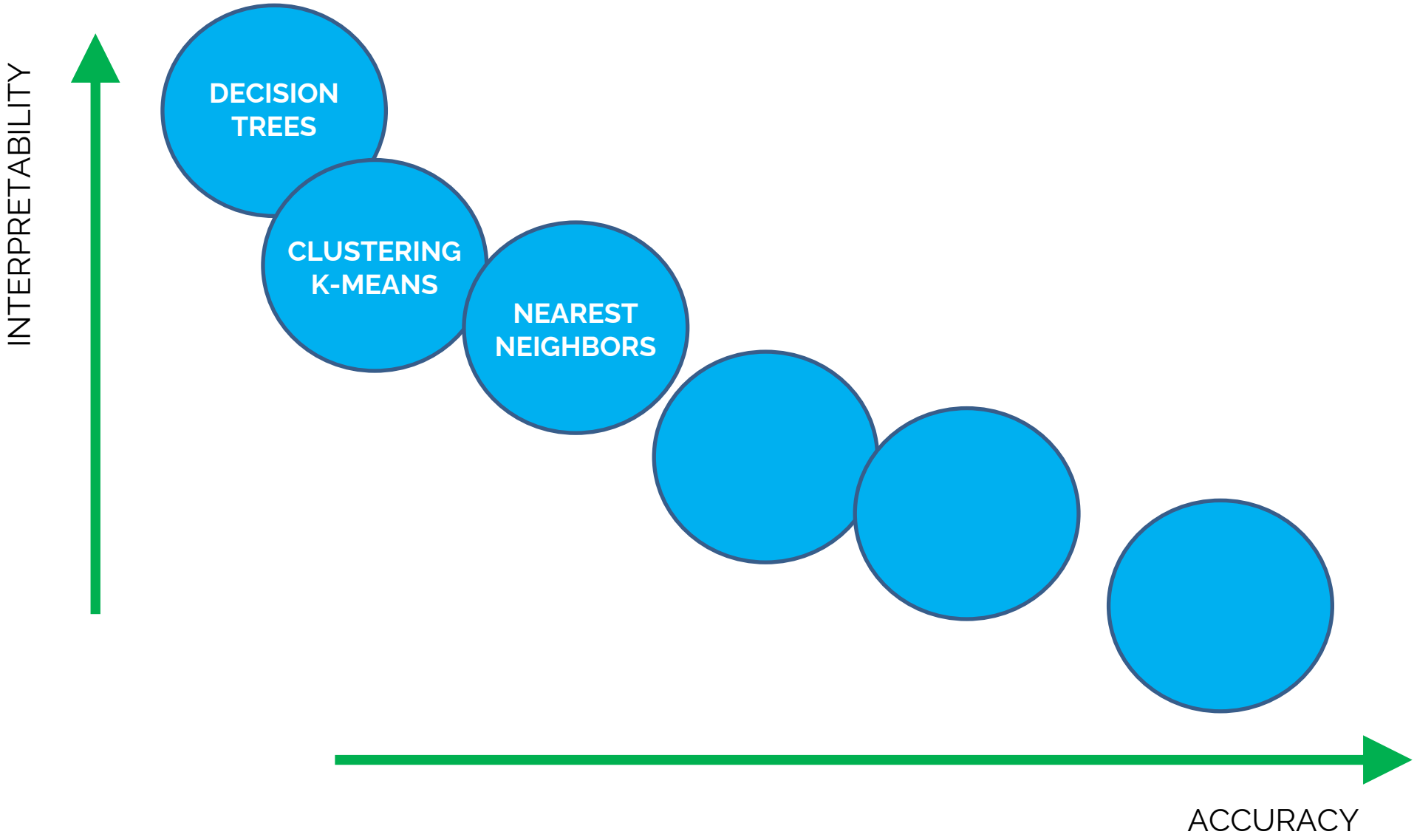


REF: Study of Various Decision Tree Pruning Methods with their Empirical Comparison in WEKA

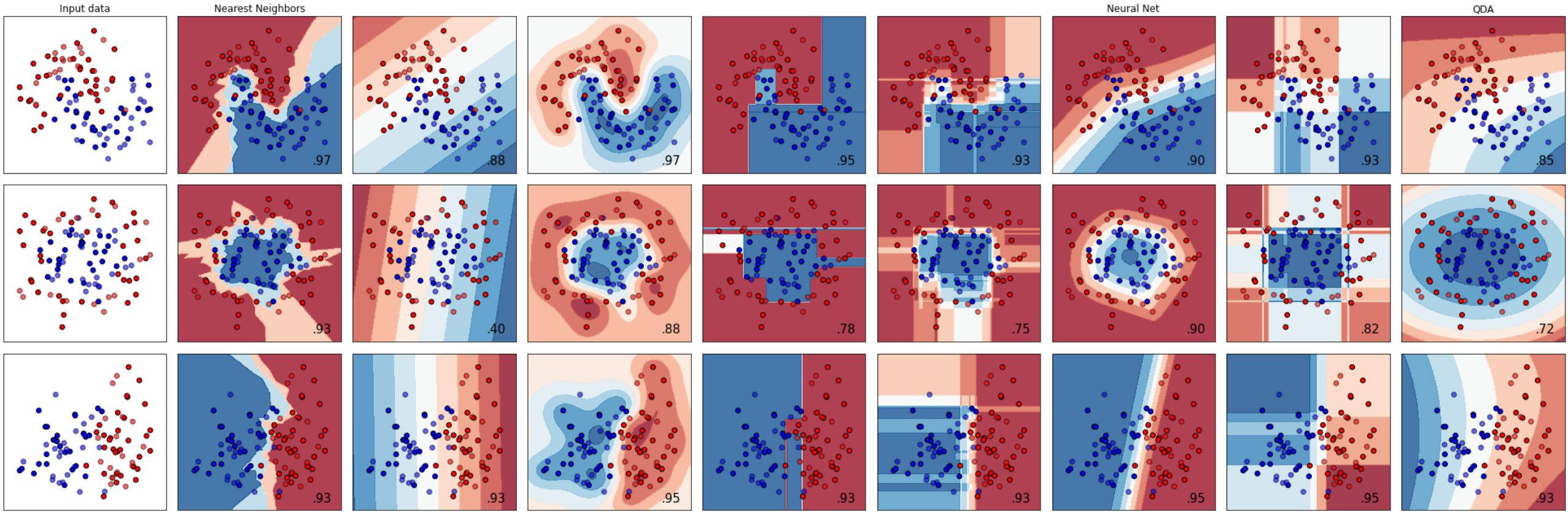
Interpretability-Accuracy TRADEOFF



Interpretability-Accuracy TRADEOFF



Which is which | Decision Function



Which is which | Decision Function