Machine Learning Fundamentals

(for galaxy morphology classification)

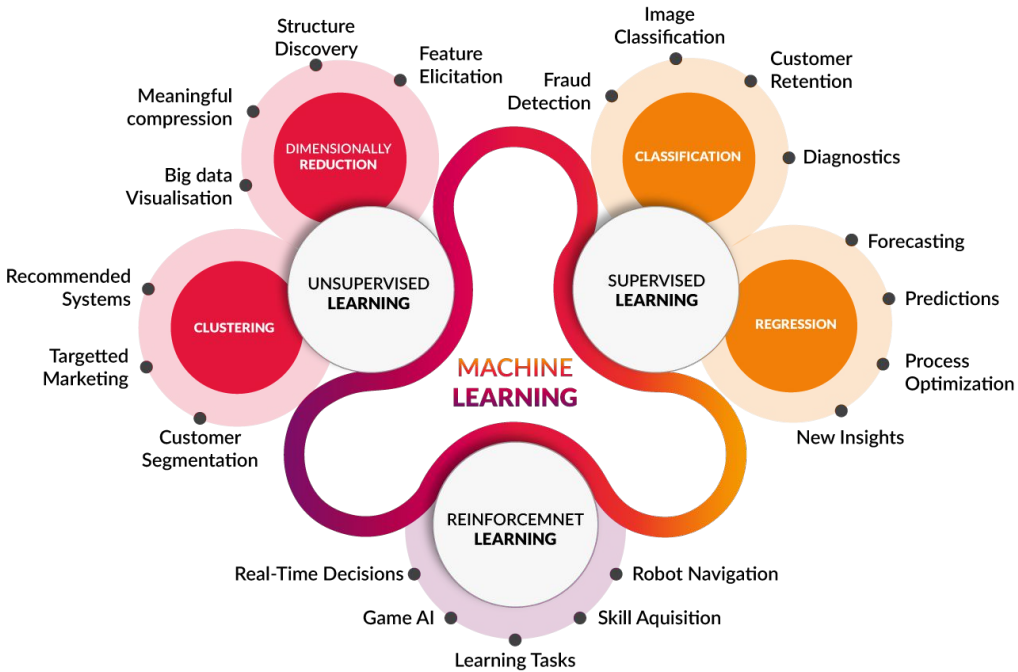
Fernando CARO

INAF - Osservatorio Astronomico di Roma

CERN Italian Teachers Program 2023

March 23, 2023

What is Machine Learning?

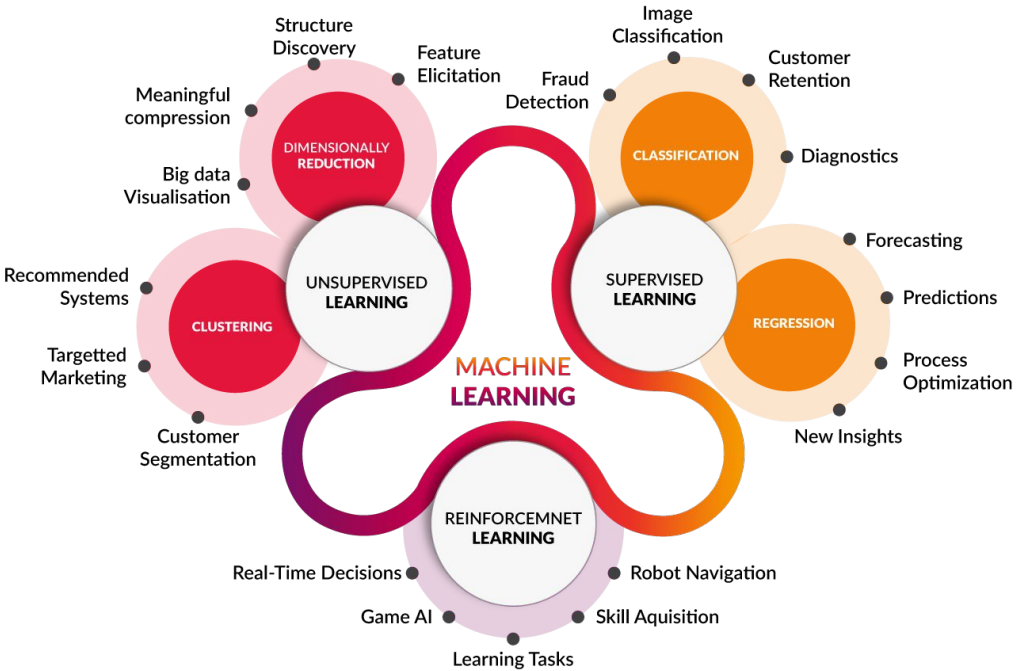


"Machine learning (ML) is a branch of artificial intelligence (AI) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy"

ML models are divided into three different main categories:

1) Supervised Learning
2) Unsupervised Learning
3) Reinforcement Learning

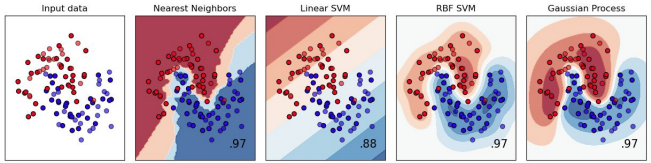
What is Machine Learning?



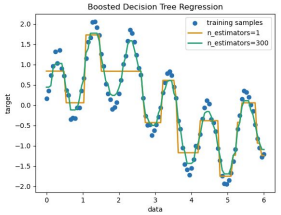
<u>Supervised learning</u> is defined by its use of labeled datasets to train algorithms to classify data or predict outcomes accurately. As input data is fed into the model, the model adjusts its weights until it has been fitted appropriately. This occurs as part of the cross validation process to ensure that the model avoids overfitting or underfitting.

Supervised learning is typically used for problems such as:

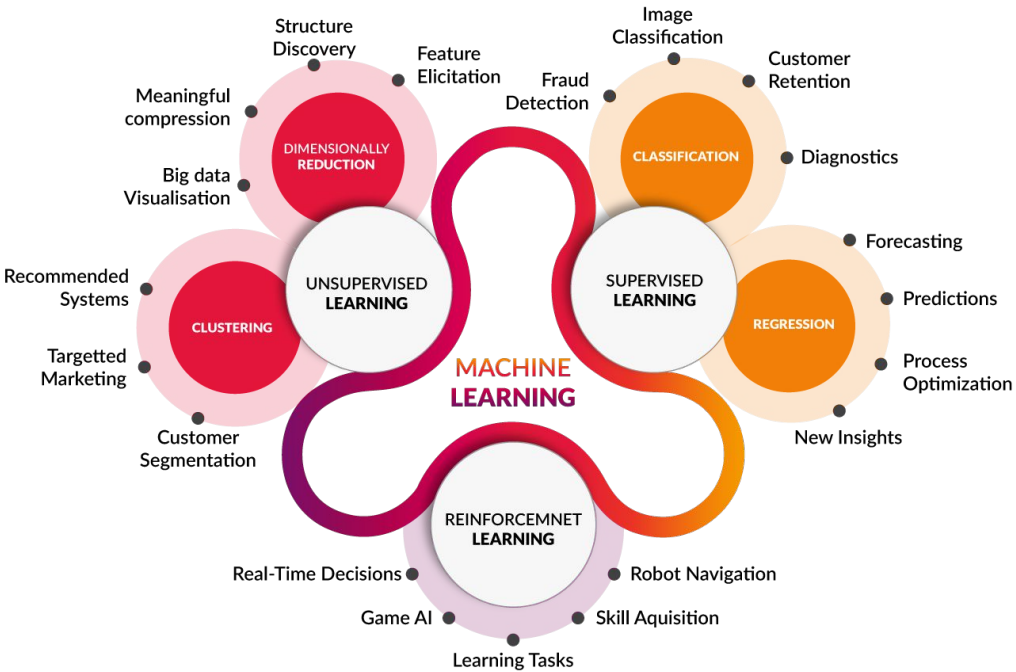
1. Classification



2. Regression



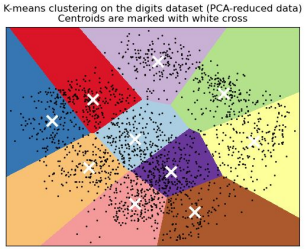
What is Machine Learning?



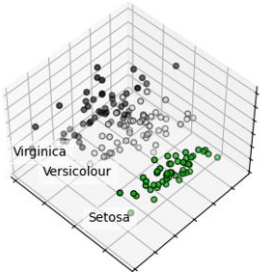
<u>Unsupervised learning</u> uses machine learning algorithms to analyze and cluster unlabeled datasets. These algorithms discover hidden patterns or data groupings without the need for human intervention.

Unsupervised learning is typically used for problems such as:

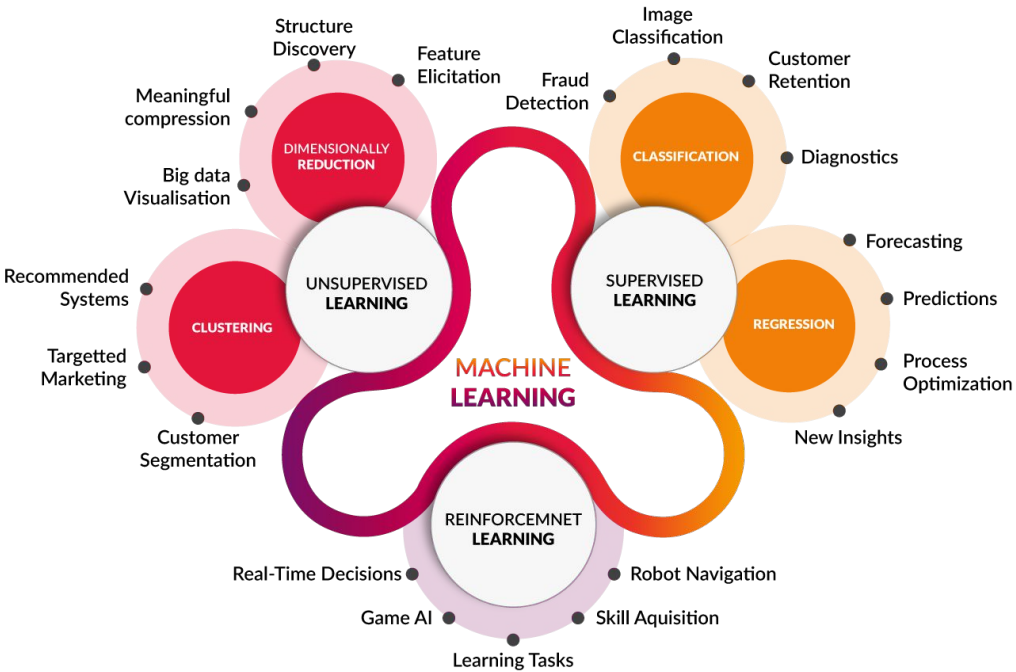
1. Clustering



2. Dimensionality Reduction



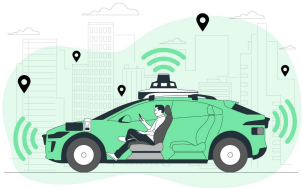
What is Machine Learning?



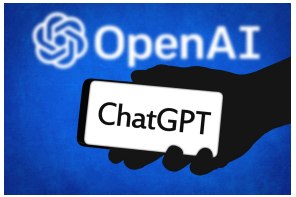
<u>Reinforcement learning</u> is a machine learning model that is similar to supervised learning, but the algorithm is not trained using sample data. This model learns as it goes by using trial and error. A sequence of successful outcomes will be reinforced to develop the best recommendation or policy for a given problem.

Reinforcement learning is now being used for problems such as:

1. Self-Driving Cars



2. Natural Language Processing (NLP)



Machine Learning: Main Concepts

For the case of **supervised machine learning** used to tackle a **classification problem**, we need to always keep in mind the following fundamental ML concepts:

- **Feature**: information used to describe an entity (for the case of a galaxy any physical properties could be used as features)

- **Label**: category to which an entity belongs to (for the case of galaxy morphology the typical labels are “spiral” and “elliptical”)

- **Training Step**: process by which the ML model (also known as **classifier**) extracts information patterns that describe the dataset provided as input (i.e. the features are mathematically connected to the labels). The dataset used in this step is typically referred to as **training set** (**training labels + training features**)

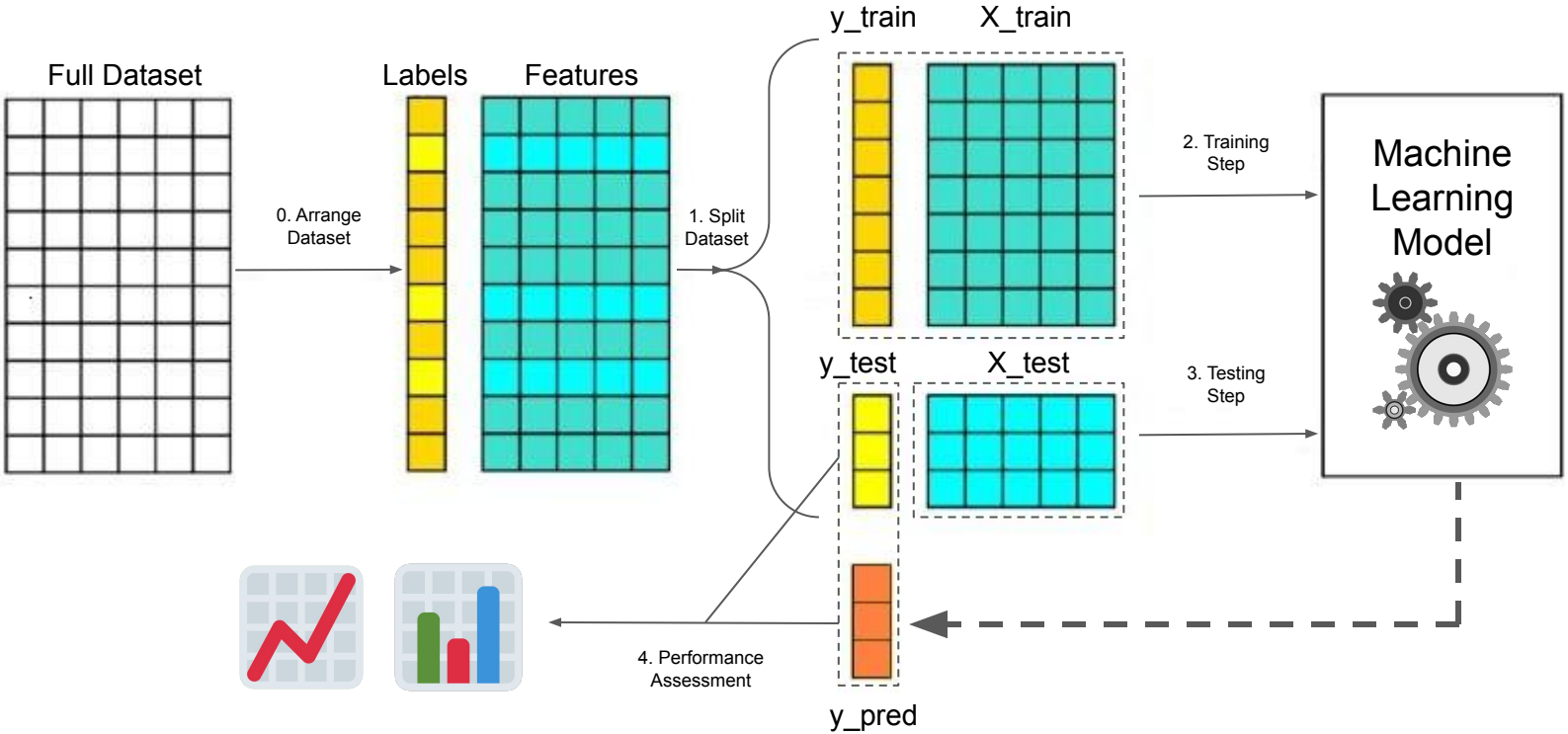
- **Testing/Prediction Step**: a list of features describing some entities, for which the corresponding labels are already known, is provided as input to the (already trained) ML model to then produce a list of **predicted labels** as output. The dataset used in this step is typically referred to as **testing set** (**testing labels + testing features**)

- **Performance Assessment Step**: By comparing the predicted labels with the testing labels as function of the different parameters considered among the testing features we can characterize the performance of a classifier.

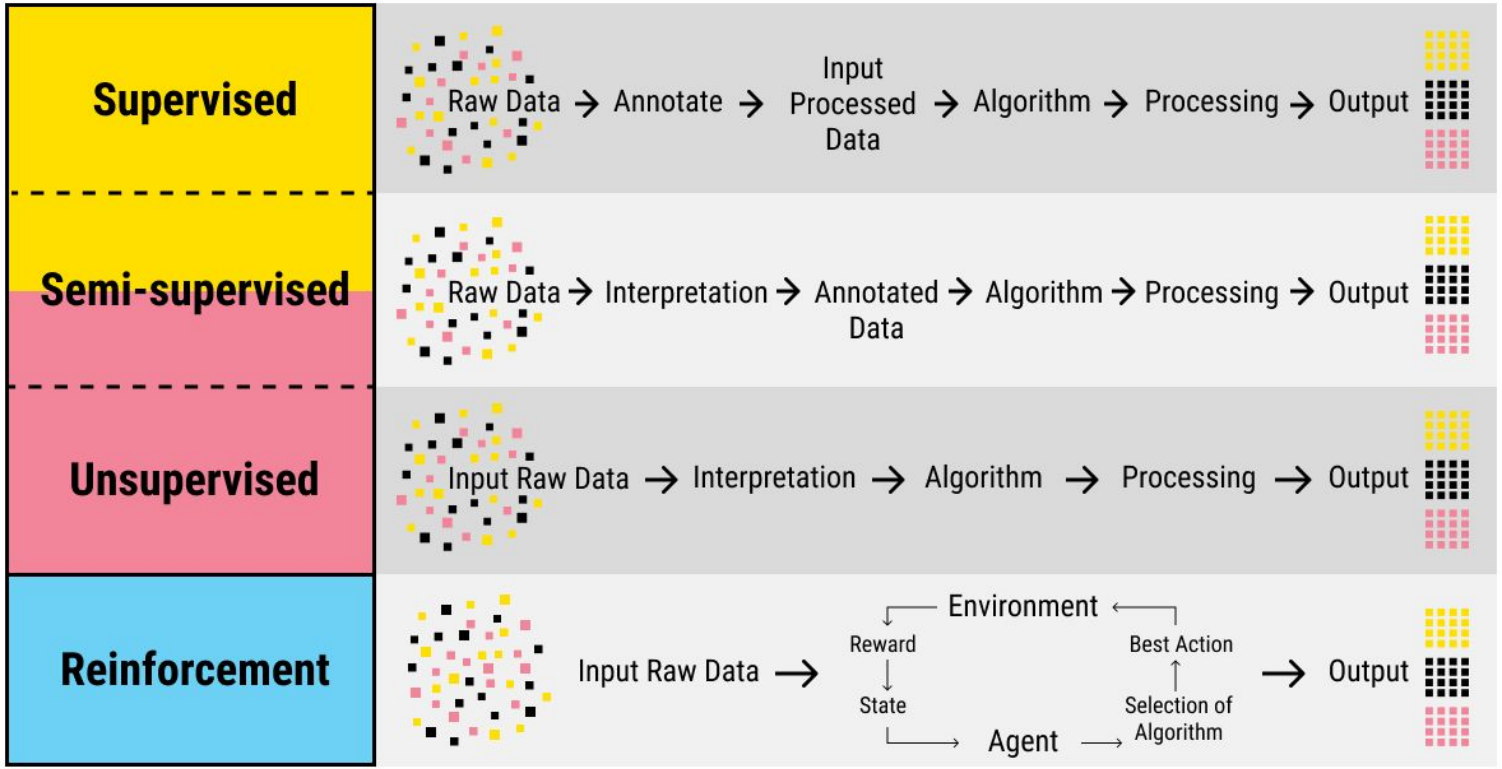
Feature 1: mag1_g = 15.56
Feature 2: mag1_r = 15.14
Feature 3: mag2_g = 18.04
Feature 4: mag2_r = 17.59
Label: “spiral”

Feature 1: mag1_g = 16.93
Feature 2: mag1_r = 15.82
Feature 3: mag2_g = 18.75
Feature 4: mag2_r = 17.60
Label: “elliptical”

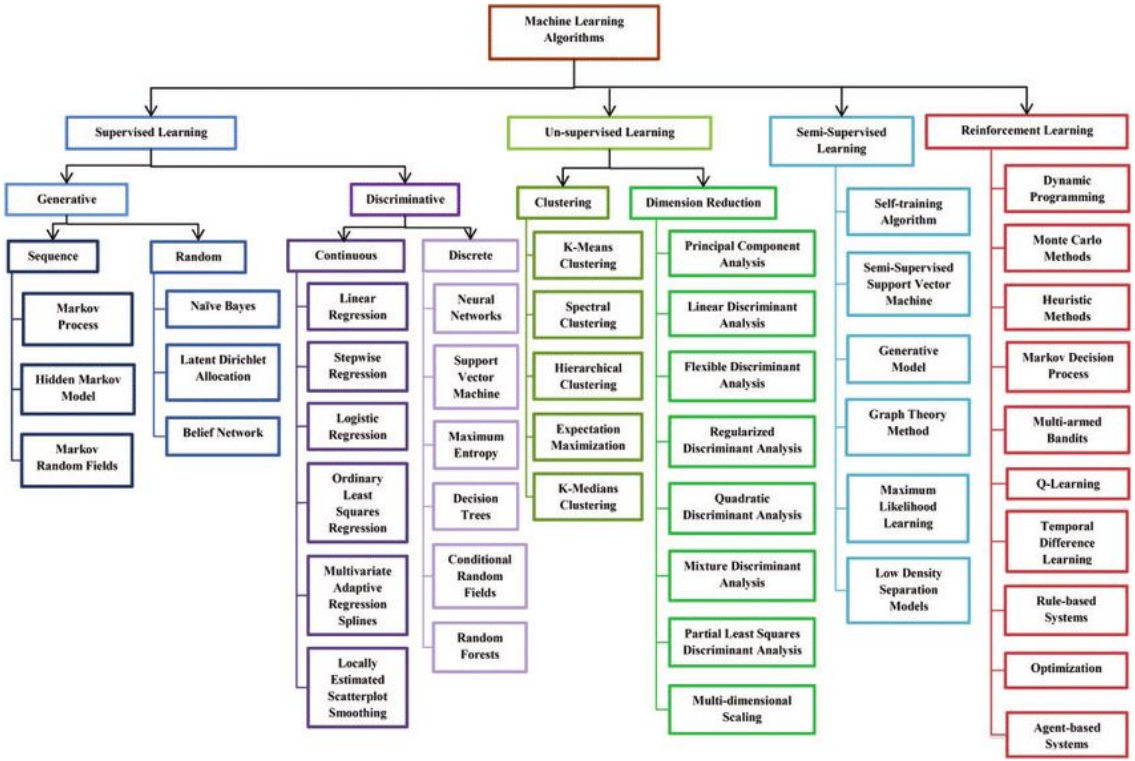
Machine Learning: Main Concepts



Machine Learning: Main Concepts



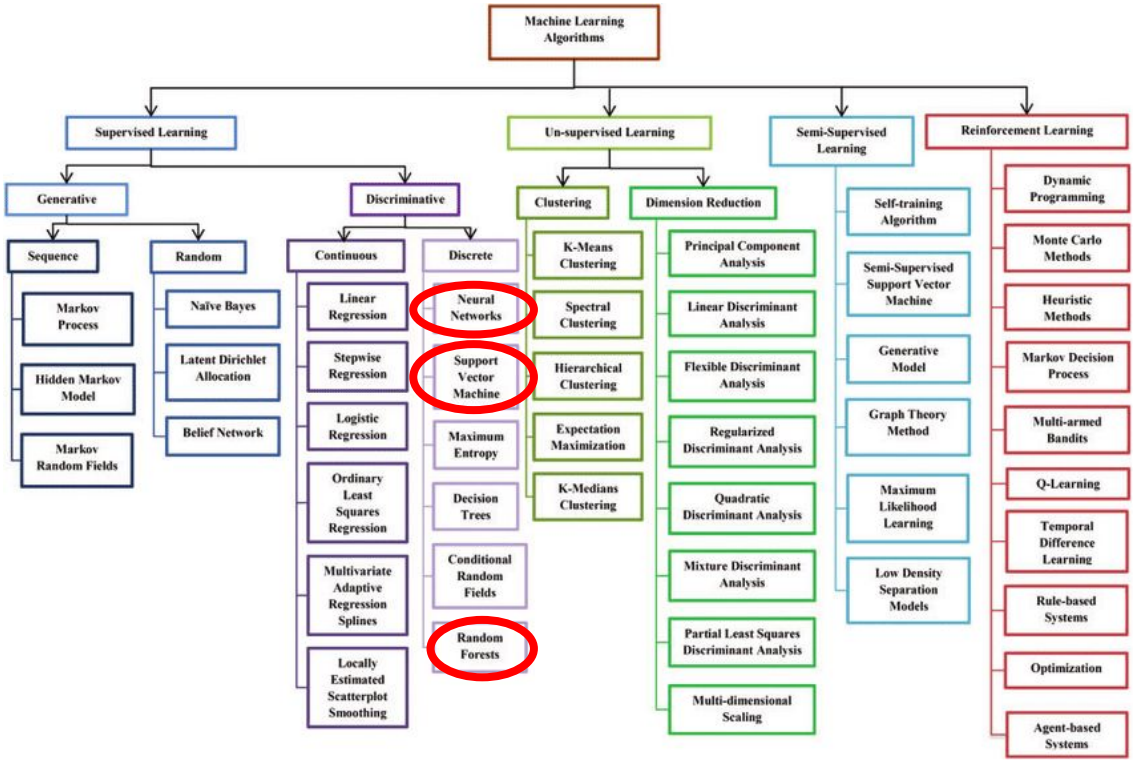
ML Algorithms: A Taxonomical Classification



This is how a typical taxonomical classification of the most commonly used ML algorithms looks like

This type of classification is based on the learning approach considered (supervised, unsupervised, etc) and the type of problem to be addressed (classification, clustering, etc)

ML Algorithms: A Taxonomical Classification



This is how a typical taxonomical classification of the most commonly used ML algorithms looks like

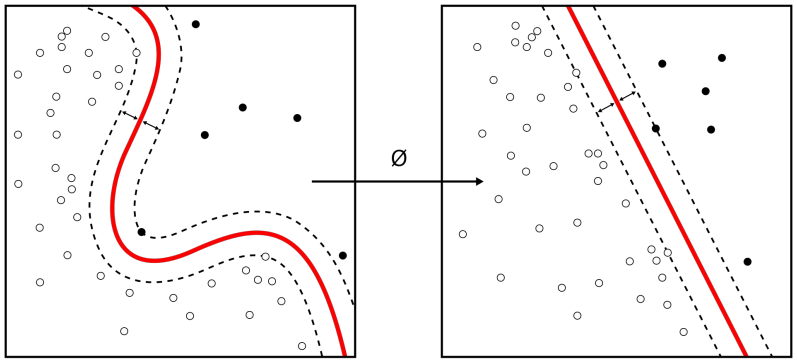
This type of classification is based on the learning approach considered (supervised, unsupervised, etc) and the type of problem to be addressed (classification, clustering, etc)

Now, we will quickly review three of the most popular classification algorithms used in the literature and in practical applications:

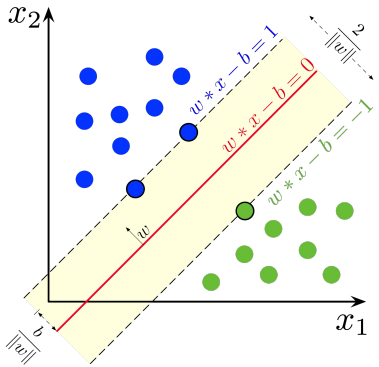
1. Support Vector Machines (SVM)
2. Random Forest (RF)
3. Neural Networks (NN)

ML Algorithms: Support Vector Machine (SVM)

*The **Support Vector Machine** algorithm constructs a **hyperplane or set of hyperplanes** in a high or infinite-dimensional space which is used to provide a **linear solution** for a classification problem. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training-data point of any class (so-called functional margin), since in general the larger the margin, the lower the generalization error of the classifier*



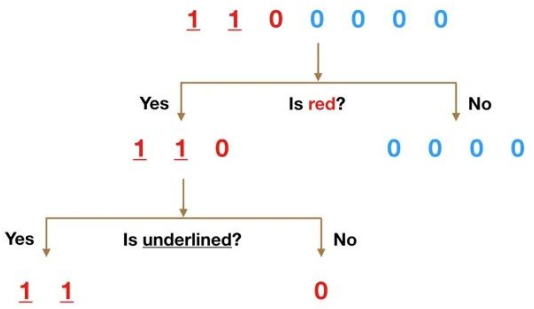
Re-projection of data in other parameter space where linear is feasible



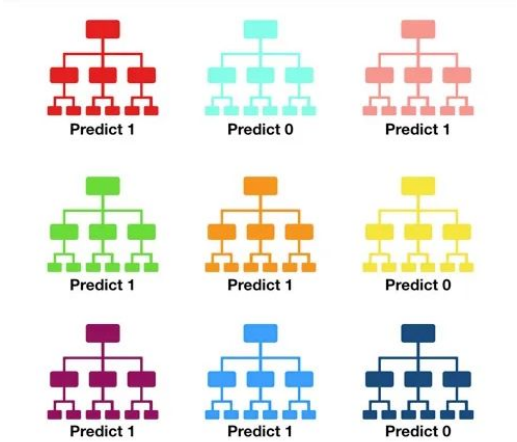
Margin computation for optimal separation between two classes

ML Algorithms: Random Forest (RF)

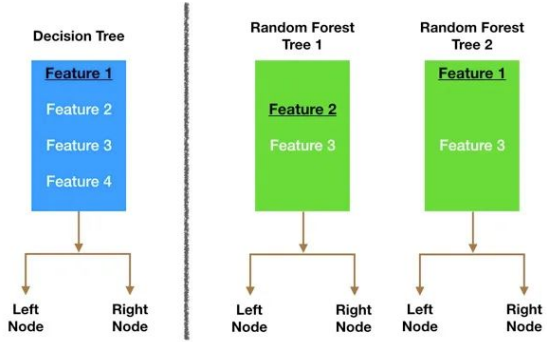
*The **Random Forest** is a classification algorithm consisting of many decisions trees. It uses **bagging** and **feature randomness** when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree.*



Example of decision tree

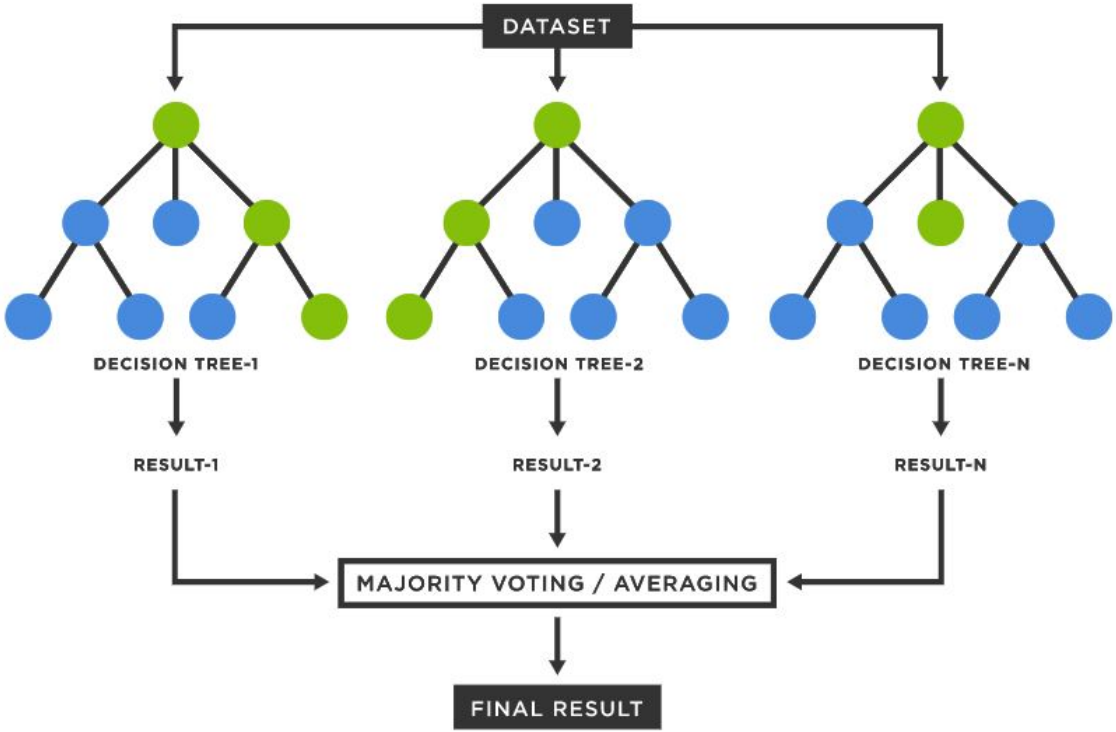


For this RF the final prediction is 1 based on the majority vote of six decision trees predicting 1 and three predicting 0



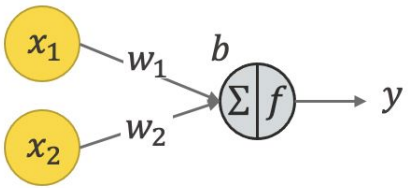
Example of feature randomness

ML Algorithms: Random Forest (RF)



ML Algorithms: Neural Networks (NN)

*A **Neural Networks** consist of a network of functions, typically called neurons, which allows the computer to learn, and to fine tune itself, by analyzing new data. **Each neuron is a function which produces an output, after receiving one or multiple inputs**. Those outputs are then passed to the **next layer of neurons**, which use them as inputs of their own function, and produce further outputs. Those outputs are then passed on to the next layer of neurons, and so it continues until **every layer of neurons** has been considered, and the terminal neurons have received their input. Those terminal neurons then output the final result for the model.*



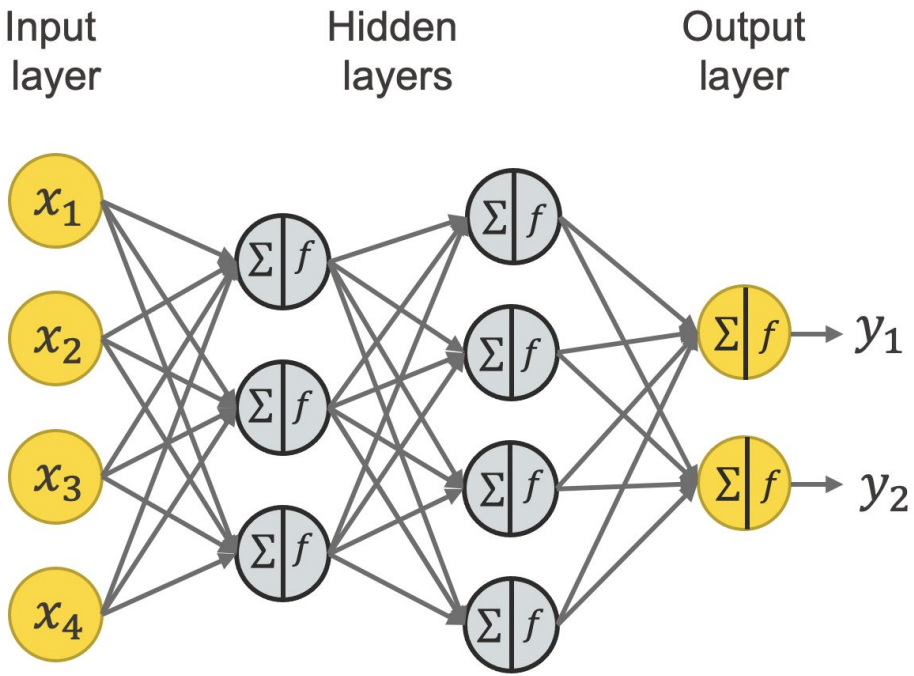
Example of a neuron that receives two parameters as input to produce its output

A neuron **Σ** which is defined by the **activation function f** considers a **weight w_i** for every **input parameter x_i** that receives along with a **constant offset value b** to compute its output y as:

$$y = f(w_1x_1 + w_2x_2 + \ldots + b)$$

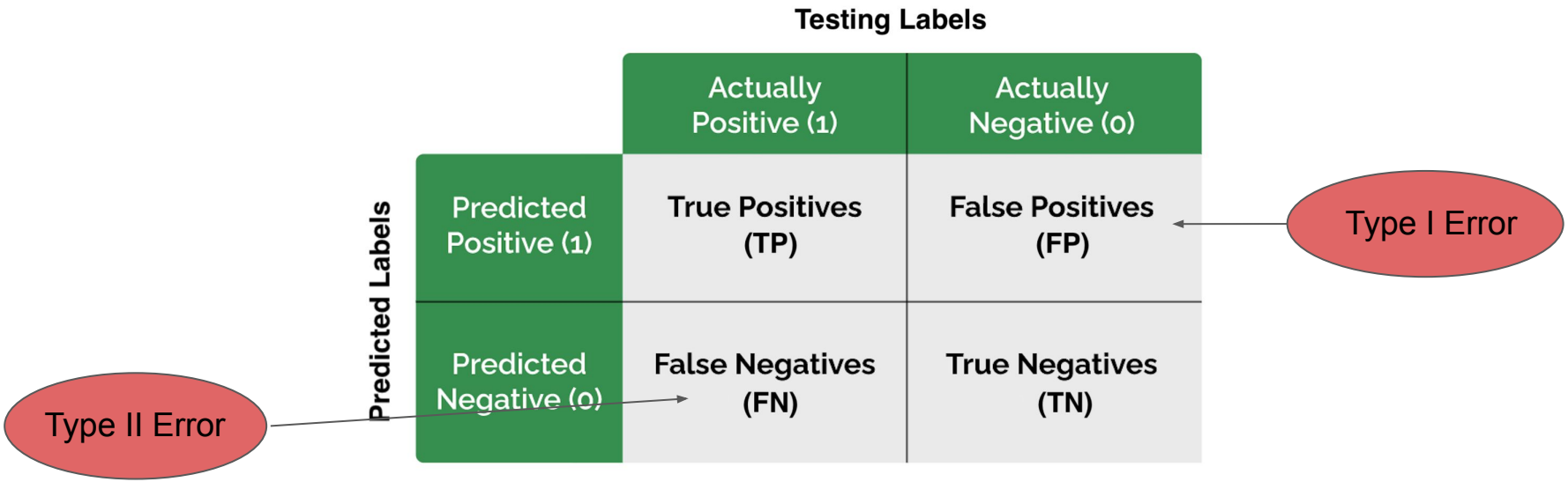
Typical alternatives for the function f are the Sigmoid function, the Hyperbolic-Tangent function and the Rectified Linear Unit (ReLU)

ML Algorithms: Neural Networks (NN)



Machine Learning: Performance Assessment

To assess the performance of a binary classifier we compare the predicted and testing labels using a **confusion matrix**.



Machine Learning: Performance Assessment

With the confusion matrix information (**TP, FP, TN, FN**) it is then possible to compute multiple **performance metrics** to characterize the classifier behavior:

sensitivity, recall, hit rate, or true positive rate (TPR)

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{P}} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

specificity, selectivity or true negative rate (TNR)

$$\mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{N}} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$$

precision or positive predictive value (PPV)

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

negative predictive value (NPV)

$$\mathrm{NPV} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FN}}$$

accuracy (ACC)

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{P} + \mathrm{N}} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

F1 score

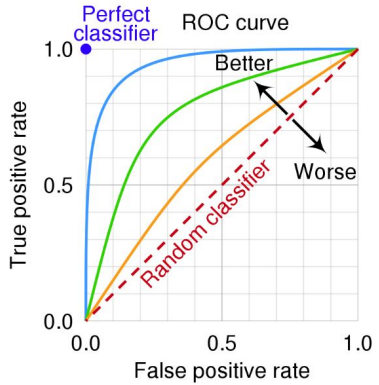
is the harmonic mean of precision and sensitivity:

$$\mathrm{F}_1 = 2 \times \frac{\mathrm{PPV} \times \mathrm{TPR}}{\mathrm{PPV} + \mathrm{TPR}} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}$$

Machine Learning: Performance Assessment

The canonical approach to summarize in a visual manner the performance of a classifier is through the Receiver Operating Characteristic (ROC) and Precision-Recall curves.

1. The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold for the classifier. The Area Under the Curve (AUC) is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve.



2. The Precision-Recall curve shows the tradeoff between precision and recall for different thresholds. A high area under the curve represents both high recall and high precision, where high precision relates to a low false positive rate, and high recall relates to a low false negative rate.