

Algorithmes de classification

Bassam Kurdy Ph.D





[<bassam.kurdy@apinum.fr>](mailto:bassam.kurdy@apinum.fr)

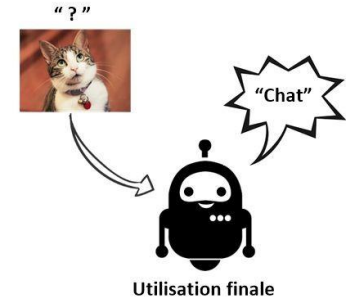
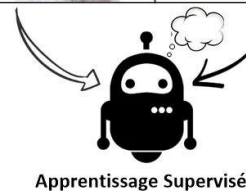
Introduction

- ❑ La classification supervisée est l'approche de machine learning **la plus utilisée et la mieux maîtrisée** à l'heure actuelle.
- ❑ La classification permet de résoudre des problèmes pratiques de la vie réelle
 - ❑ la détection de défaut d'usinage, de fraude, de maladie
 - ❑ le tri automatique de courrier, de document ou de vidéo
 - ❑ la reconnaissance d'images
- ❑ La classification permet de résoudre **les tâches où un choix est requis**.

Classification supervisée | Quésaco ?

- ❑ La classification supervisée consiste à **attribuer automatiquement une catégorie (ou une classe) à des données** dont on ne connaît pas la catégorie.
- ❑ Pour cela, **un classifieur** (algorithme de machine learning) est **entraîné** sur des données similaires ou très proches des données que l'on souhaite classer.

x	y
	"Chien"
	"Chien"
	"Chat"
	"Chien"



Crédit : machine Learnia

Algorithmes de classification supervisée

☐ Le k-plus proche voisin

- ☐ La méthode de k-proche voisins consiste à chercher dans une base de données l'exemple le plus proche de celui que l'on est entrain de traiter.

☐ L'arbre de décision

- ☐ Un outil d'aide à la décision représentant un ensemble de choix sous la forme graphique d'un arbre.
- ☐ les différentes décisions possibles sont situées aux extrémités des branches.

Algorithmes de classification supervisée

☐ Le random Forest

- ☐ On construit plusieurs arbres de décision de moins bonne qualité individuelle qui possède une vision réduite du problème.
- ☐ On réunit l'ensemble de ces estimateurs (classifieurs) pour avoir une vision globale.

☐ Le perceptron multicouches

- ☐ Un ensemble de neurones connectés.
- ☐ Il est composé d'une couche d'entrée, de n couches cachées, et d'une couche de sortie.

Algorithmes de classification supervisée

☐ Régression logistique

- ☐ un modèle statistique qui permet d'étudier les relations entre un ensemble de variables d'entrée et une variable de sortie.
- ☐ il s'agit d'un modèle linéaire généralisé utilisant une fonction logistique comme fonction de lien.

Veille individuelle  **20min**

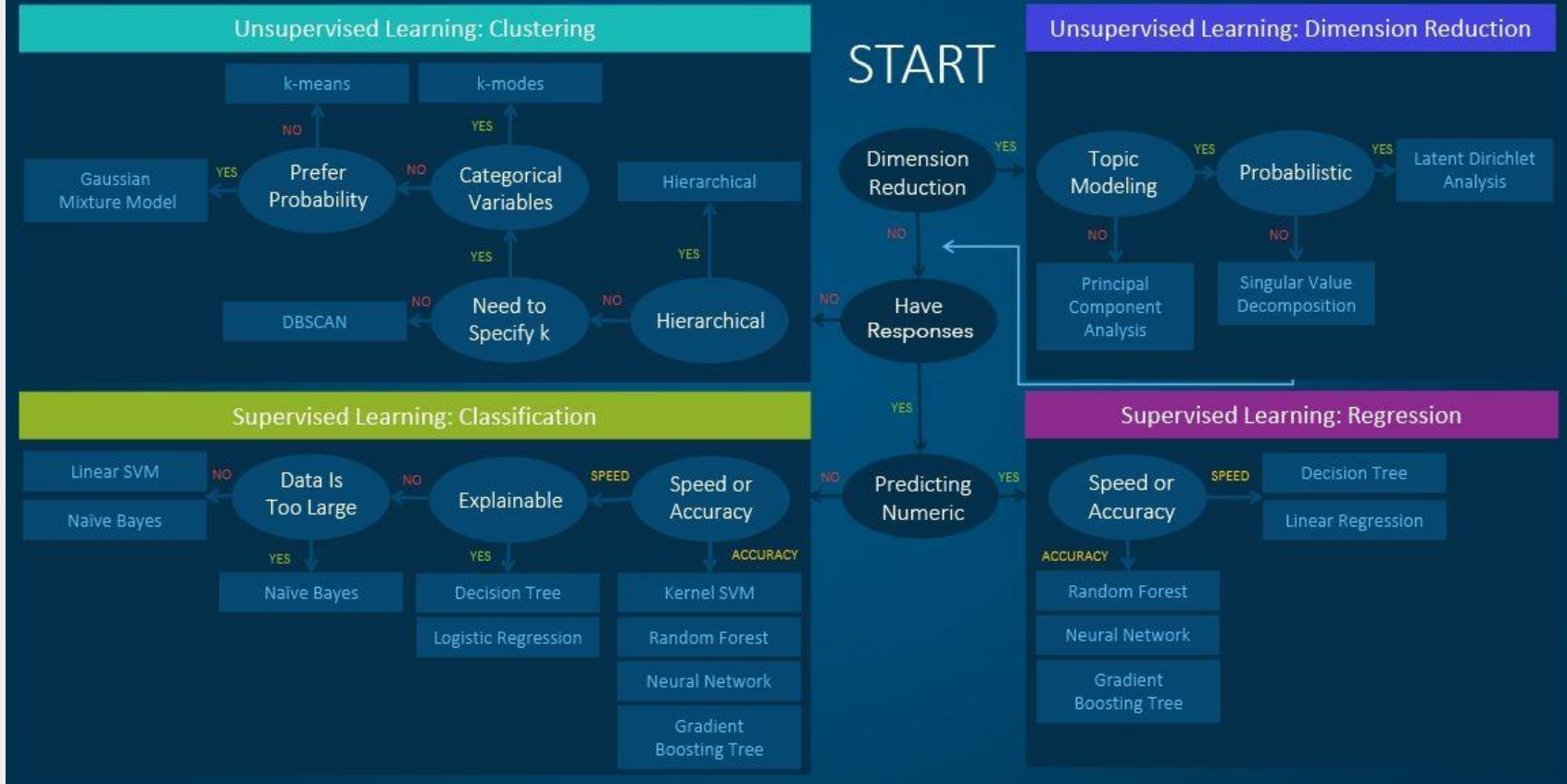
Algorithmes de classification supervisée

- ❑ Lister les algorithmes de classification les plus utilisés
- ❑ Comment choisir “le meilleur” algorithme ?



- ❑ <https://www.kdnuggets.com/2020/05/guide-choose-right-machine-learning-algorithm.html>
- ❑ [Sélectionner un algorithme d'apprentissage automatique - Azure Machine Learning | Microsoft Learn](#)

Machine Learning Algorithms Cheat Sheet



Source : SAS Algorithm Flowchart

Régression logistique | Quésaco

- ❑ La régression logistique est utilisée pour estimer une valeur **discrète** (**classe** ou **catégorie**).
- ❑ La régression logistique est un **modèle statistique** qui permet d'étudier les relations entre un ensemble d'entrée **X** et une variable de sortie **y**.
- ❑ Un modèle de régression logistique permet aussi de **prédire la probabilité** qu'un événement arrive (valeur de 1) ou non (valeur de 0) à partir de l'**optimisation des coefficients de régression**.
- ❑ Lorsque la valeur prédite est supérieure à un seuil, l'événement est susceptible de se produire, alors que lorsque cette valeur est inférieure au même seuil, il ne l'est pas.

Régression logistique | Quésaco

Pourquoi régression logistique ?

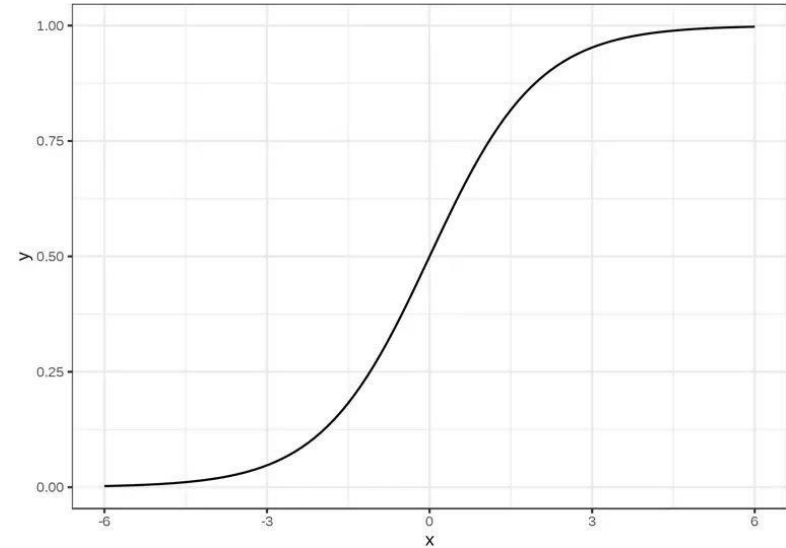
- ❑ “**Régression**” : on cherche à montrer une relation de dépendance entre une variable à expliquer et des variables explicatives. Cette dépendance s’exprime en terme de probabilité d’appartenir à chacune des classes.
- ❑ “**Logistique**” : la loi de probabilité est modélisée à partir d’une loi logistique.

Régression logistique | Quésaco

Mathématiquement

- ❑ Considérons une entrée
 $X = x_1, x_2, \dots, x_n$
- ❑ La régression a pour objectif de trouver une fonction h telle que :
$$y = \begin{cases} 1 & \text{si } h(X) \geq \text{seuil} \\ 0 & \text{si } h(X) < \text{seuil} \end{cases}$$
- ❑ On utilise la fonction **sigmoïde**

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$



Représentation de la fonction sigmoïde

Régression logistique | Quésaco

Mathématiquement

- ❑ La fonction h qui définit la régression logistique s'écrit alors :

$$\forall (X \in \mathbb{R}^n) \quad h(X) = \frac{1}{1 + e^{-\sum_{i=1}^n \theta_i x_i}}$$

- ❑ Le problème de classification par régression logistique est un problème d'**optimisation**, où, à partir de données, nous essayons d'**obtenir le meilleur jeu de paramètre** qui permettent à la courbe sigmoïde de **coller au mieux aux données**.
- ❑ C'est dans cette étape qu'intervient l'**apprentissage automatique**.

Régression logistique | Types

- ☐ Il existe plusieurs types de régression logistiques :
- ☐ **Régression logistique binaire**
 - ☐ oui/non
- ☐ **Régression logistique multinomiale** : la variable dépendante a trois catégories nominales ou plus.
 - ☐ prédiction de la catégorie d'iris
- ☐ **Régression logistique ordinale** : la variable dépendante a trois catégories ordinales ou plus.
 - ☐ la notation d'un produit de 1 à 5

Régression logistique | ChatGPT



qu'est ce que la régression logistique



La régression logistique est un algorithme de classification utilisé pour résoudre les problèmes de classification binaire. Il utilise une fonction de probabilité logistique pour modéliser la probabilité qu'un élément appartienne à une certaine classe. Il peut également être utilisé pour résoudre les problèmes de classification multi-classes en utilisant des techniques dites de "one-vs-all" ou "softmax".

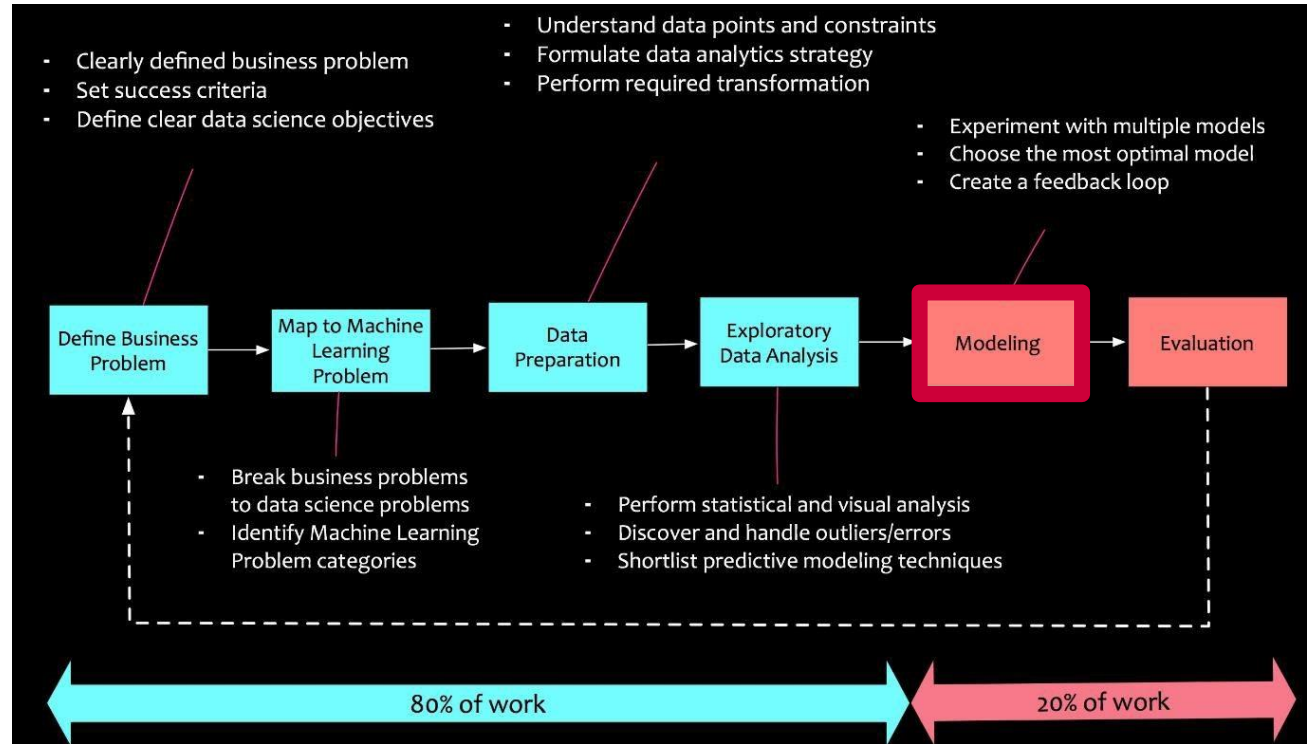


La régression logistique est basée sur la méthode des moindres carrés pour ajuster les coefficients d'un modèle linéaire aux données d'entraînement. Une fois que le modèle est entraîné, il peut être utilisé pour prédire la probabilité d'appartenance à une classe pour de nouveaux éléments de données.

La régression logistique est souvent utilisée lorsque les données sont linéairement séparables, c'est-à-dire qu'il existe une droite ou un hyperplan qui sépare les différentes classes. Elle est également utilisée pour résoudre les problèmes de classification binaire et multi-classes, pour des problèmes d'analyse de risque, de marketing, de biologie, de médecine, etc.

Mon premier modèle de ML

Cycle de vie d'un projet IA



source :

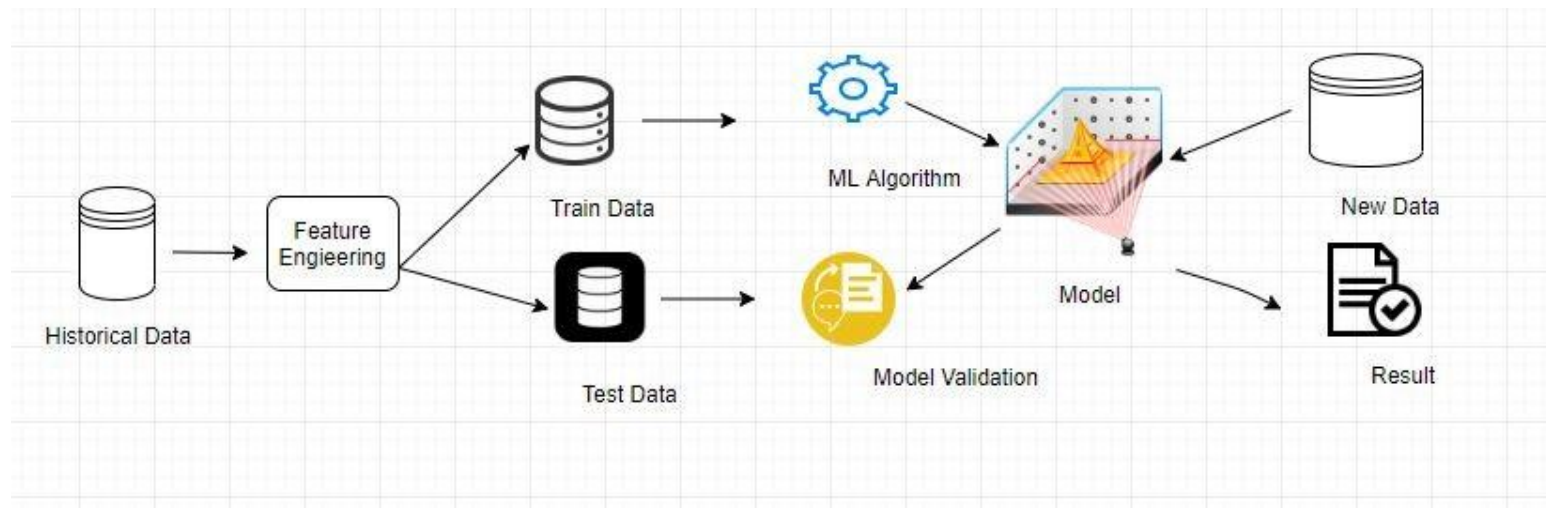
[Data Science Simplified Part I : Principles and Process - Pradeep Menon](#)

Étapes clés



7 étapes vers l'apprentissage automatique

Source : [7 Steps to Machine Learning: How to Prepare for an Automated Future](#)



Étapes de création d'un modèle ML
Source : [Machine Learning Workflow](#)

Etape de construction d'un modèle

1. Importer les données
2. Séparation des données en sous ensemble d'entraînement et un sous ensemble de test. -> ***train_test_split***
3. construction du modèle -> ***LogisticRegression***
4. Entrainement du modèle avec le sous ensemble d'entraînement -> ***fit***
5. Prédications -> ***predict***
6. Evaluation du modèle -> ***score***

Travail en groupe



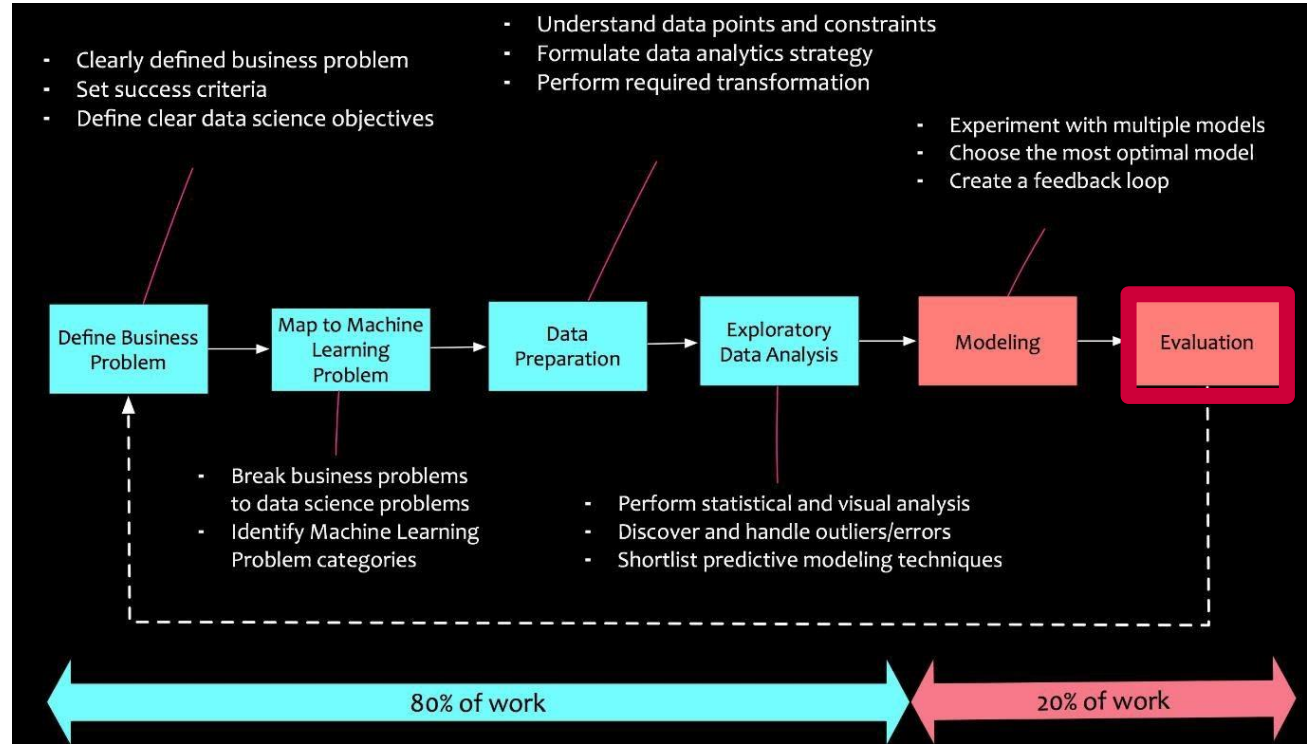
30min

Modèle de régression logistique

- ❑ Workshop :
Régression logistique
- ❑ Partie I uniquement
- ❑ Niveau : Imiter



Cycle de vie d'un projet IA



source :

[Data Science Simplified Part I : Principles and Process - Pradeep Menon](#)

Veille individuelle



45min

Evaluation du modèle

- ❑ Comment évaluer un modèle de classification ?
 - ❑ la matrice de confusion
 - ❑ l'exactitude
 - ❑ la précision
 - ❑ le rappel
 - ❑ F1 score
 - ❑ la courbe roc



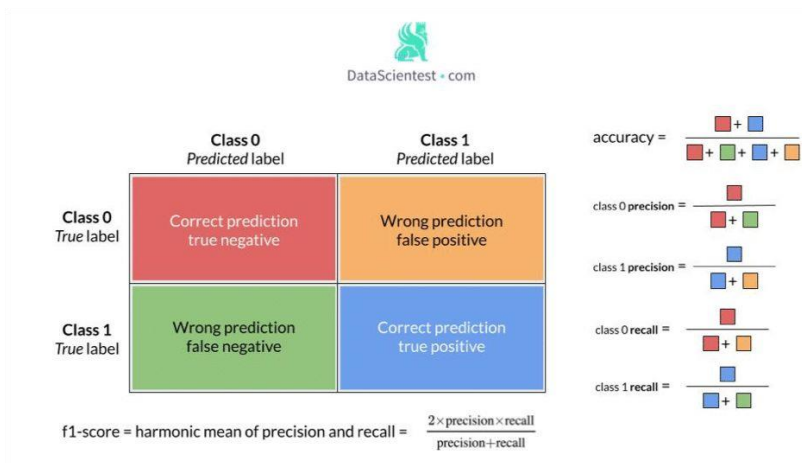
<https://datascientest.com/comment-gerer-les-problemes-de-classification-desequilibree-partie-i#:~:text=Qu'est%20ce%20qu'une,m%C3%A9trique%20utilis%C3%A9e%20pour%20l'%C3%A9valuer.>

<https://docs.microsoft.com/fr-fr/azure/machine-learning/component-reference/evaluate-model#metrics-for-classification-models>

[Quel sens métier pour les métriques de classification ? - OCTO Talks !](#)

Evaluation d'un modèle ML

- ❑ Métriques pour les modèles de **classification**
 - ❑ **L'exactitude (accuracy)** : mesure l'adéquation d'un modèle de classification sous forme de proportion de résultats réels sur le nombre total de cas.
 - ❑ La **précision (precision)** : correspond à la proportion de résultats réels sur tous les résultats positifs.



Source : [datascientest](https://datascientest.com)



	Class 0 Predicted label	Class 1 Predicted label
Class 0 True label	Correct prediction true negative	Wrong prediction false positive
Class 1 True label	Wrong prediction false negative	Correct prediction true positive

$$\text{f1-score} = \text{harmonic mean of precision and recall} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

$$\text{accuracy} = \frac{\text{red} + \text{blue}}{\text{red} + \text{green} + \text{blue} + \text{orange}}$$

$$\text{class 0 precision} = \frac{\text{red}}{\text{red} + \text{green}}$$

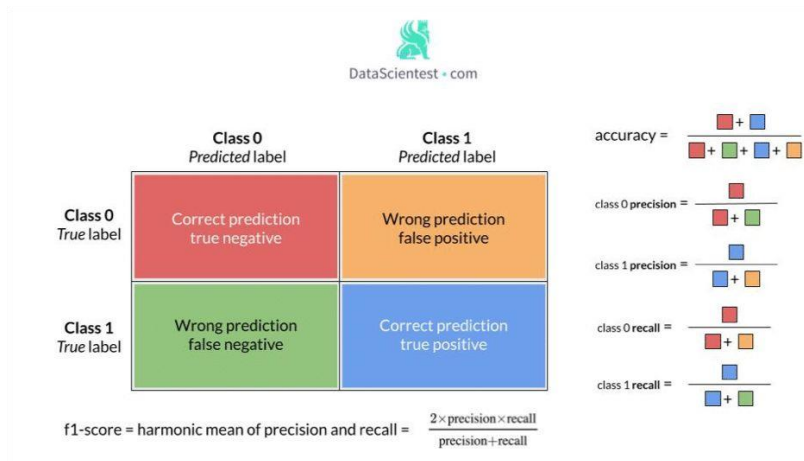
$$\text{class 1 precision} = \frac{\text{blue}}{\text{blue} + \text{orange}}$$

$$\text{class 0 recall} = \frac{\text{red}}{\text{red} + \text{orange}}$$

$$\text{class 1 recall} = \frac{\text{blue}}{\text{blue} + \text{green}}$$

Evaluation d'un modèle ML

- ❑ Métriques pour les modèles de **classification**
 - ❑ Le **rappel (recall)** : est la fraction de la quantité totale d'instances pertinentes qui ont été réellement récupérées.
 - ❑ Le **F1 Score** : la moyenne harmonique de précision et de rappel.



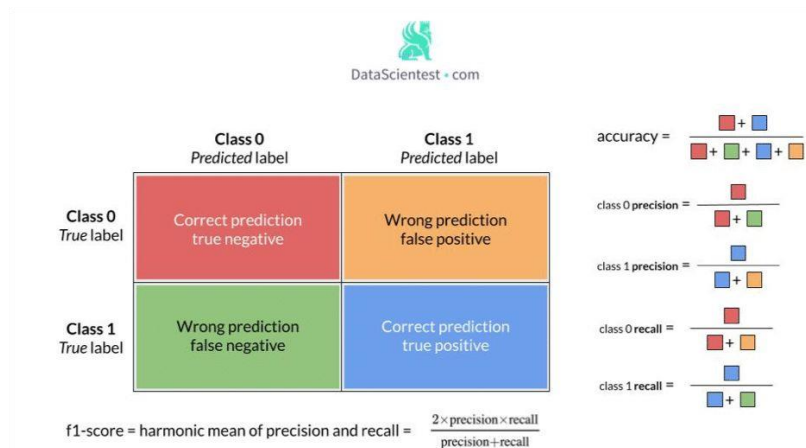
Source : [datascientest](https://datascientest.com)

Evaluation d'un modèle ML

❑ Métriques pour les modèles de classification

- ❑ **AUC** : mesure la zone sous la courbe tracée avec les vrais positifs sur l'axe y et les faux positifs sur l'axe x.

Cette métrique est utile car elle fournit un nombre unique qui vous permet de comparer les modèles de types différents.



Source : [datascientest](https://datascientest.com)

Travail en binôme



45 min

Modèle de régression logistique

- ❑ Workshop :
Régression logistique
- ❑ Partie II
- ❑ Niveau : Adapter



Avantages et Inconvénients

Avantages

- ❑ Algorithme simple, efficace et facile à mettre en oeuvre
- ❑ ne nécessite pas une grande puissance de calcul
- ❑ fournit des scores de probabilité pour les observations
- ❑ très utilisés pour le scoring

Inconvénients

- ❑ Ne peut pas résoudre le problème de non-linéarité ce qui nécessite la transformation des caractéristiques non linéaires.
- ❑ Il faut faire très attention à l'interprétabilité des paramètres