

Algorithmes d'apprentissage et modèles prédictifs

Le cas de la justice prédictive

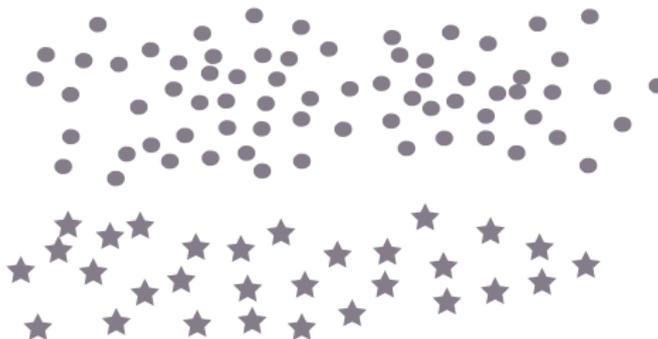
Fabien Tarissan

CNRS – ENS Paris Saclay

ENUM

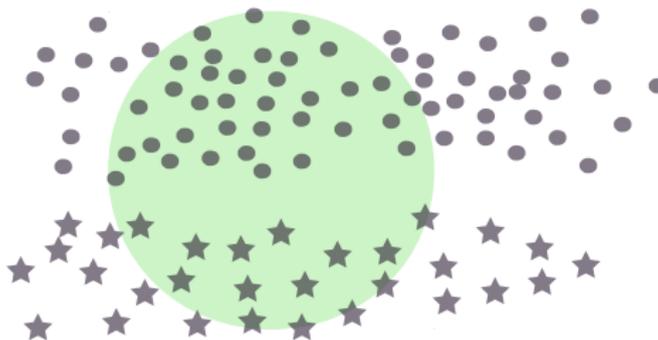
Avant de commencer

● ★ **What is looked for**

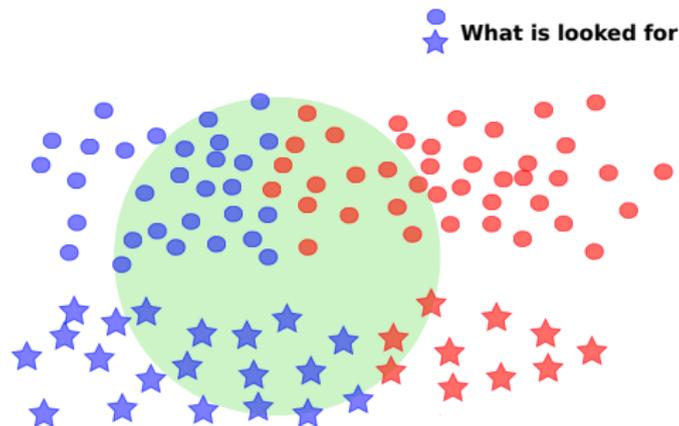


Avant de commencer

● ★ **What is looked for**

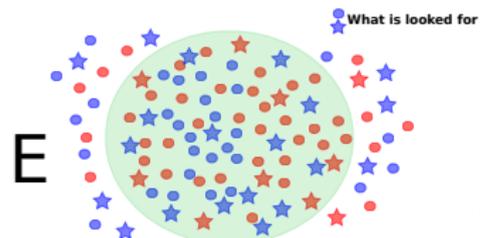
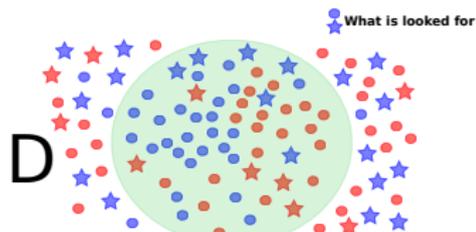
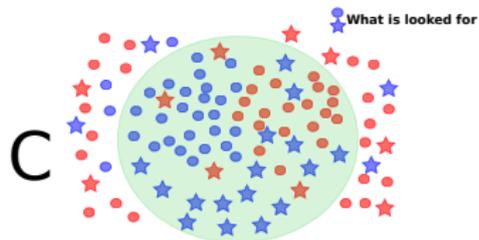
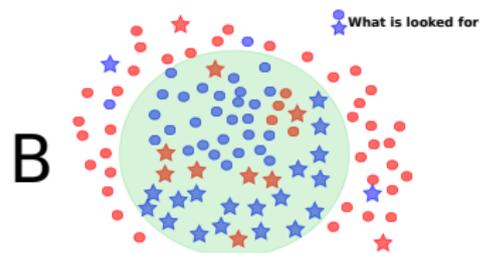
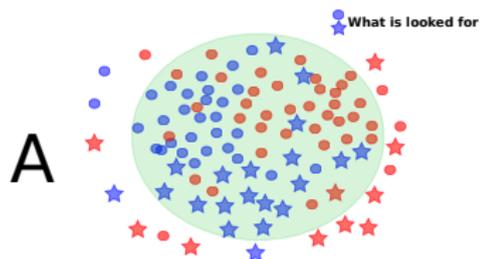


Avant de commencer



<https://www.sli.do/>

67498



Algorithmes d'apprentissage

Plan

- 1 Algorithmes d'apprentissage
 - Qu'est-ce qu'un algorithme
 - Les principes de l'apprentissage supervisé
- 2 L'apprentissage dans le contexte juridique
 - Différents acteurs pour différents modèles
 - Articles scientifiques
- 3 Évaluer un modèle prédictif
 - Définir l'équité d'un modèle
- 4 Discussions

Qu'est-ce qu'un algorithme ?

Qu'est-ce qu'un algorithme ?

Qu'est-ce qu'un algorithme ?

Qu'est-ce qu'un algorithme ?

Un algorithme est séquence d'opérations :

- simples
- non ambiguës
- qui résout un problème donné

Exemples

- une recette de cuisine (comment faire un gâteau au chocolat ?)
- un itinéraire (comment se rendre à l'ENS depuis son domicile ?)
- algorithme de Dijkstra (comment trouver un plus court chemin entre 2 points ?)

Qu'est-ce qu'un algorithme ?

Qu'est-ce qu'un algorithme ?

Un algorithme est séquence d'opérations :

- simples
- non ambiguës
- qui résout un problème donné

Exemples

- une recette de cuisine (comment faire un gâteau au chocolat ?)
- un itinéraire (comment se rendre à l'ENS depuis son domicile ?)
- algorithme de Dijkstra (comment trouver un plus court chemin entre 2 points ?)

En informatique, on s'intéresse en général à des algorithmes travaillant sur des symboles (caractères, chiffres, ...)

Qu'est-ce qu'un algorithme ?

Qu'est-ce qu'un algorithme ?

Un algorithme est séquence d'opérations :

- simples
- non ambiguës
- qui résout un problème donné

Exemples

- une recette de cuisine (comment faire un gâteau au chocolat ?)
- un itinéraire (comment se rendre à l'ENS depuis son domicile ?)
- **algorithme de Dijkstra** (comment trouver un plus court chemin entre 2 points ?)

En informatique, on s'intéresse en général à des algorithmes travaillant sur des **symboles** (caractères, chiffres, ...), qui résolvent des problèmes **génériques**

Qu'est-ce qu'un algorithme ?

Qu'est-ce qu'un algorithme ?

Un algorithme est séquence d'opérations :

- simples
- non ambiguës
- qui résout un problème donné

Exemples

- une recette de cuisine (comment faire un gâteau au chocolat ?)
- un itinéraire (comment se rendre à l'ENS depuis son domicile ?)
- **algorithme de Dijkstra** (comment trouver un plus court chemin entre 2 points ?)

En informatique, on s'intéresse en général à des algorithmes travaillant sur des **symboles** (caractères, chiffres, ...), qui résolvent des problèmes **génériques** et qui sont exécutables par des **machines**.

Comment étudier un algorithme ?

Résoudre un problème (pour un.e informaticien.ne)
c'est proposer une méthode **générique**
qui fournit la bonne réponse à **chaque instance** du problème

Analyser un algorithme, c'est donc :

- comprendre quel **problème** il cherche à résoudre
- savoir comment il **procède**
- identifier les avantages / les dangers / les biais dans sa **mise en application**

À chaque type de problème ...

... correspond une famille d'algorithmes (un paradigme)

Comment étudier un algorithme ?

Résoudre un problème (pour un.e informaticien.ne)
c'est proposer une méthode **générique**
qui fournit la bonne réponse à **chaque instance** du problème

Analyser un algorithme, c'est donc :

- comprendre quel **problème** il cherche à résoudre
- savoir comment il **procède**
- identifier les avantages / les dangers / les biais dans sa **mise en application**

À chaque type de problème ...

- Faire la somme de deux nombres ?

... correspond une famille d'algorithmes (un paradigme)

- Procédure explicite ? \implies Description séquentielle

Comment étudier un algorithme ?

Résoudre un problème (pour un.e informaticien.ne)
c'est proposer une méthode **générique**
qui fournit la bonne réponse à **chaque instance** du problème

Analyser un algorithme, c'est donc :

- comprendre quel **problème** il cherche à résoudre
- savoir comment il **procède**
- identifier les avantages / les dangers / les biais dans sa **mise en application**

À chaque type de problème ...

- Faire la somme de deux nombres ?
- ...

... correspond une famille d'algorithmes (un paradigme)

- Procédure explicite ? \implies Description séquentielle
- ... \implies Algorithmes "Diviser pour Régner", "Glouton", "Probabiliste", "Programmation Dynamique", ...

Comment étudier un algorithme ?

Résoudre un problème (pour un.e informaticien.ne)
c'est proposer une méthode **générique**
qui fournit la bonne réponse à **chaque instance** du problème

Analyser un algorithme, c'est donc :

- comprendre quel **problème** il cherche à résoudre
- savoir comment il **procède**
- identifier les avantages / les dangers / les biais dans sa **mise en application**

À chaque type de problème ...

- Faire la somme de deux nombres ?
- ...
- Étiqueter une image ? Proposer un montant compensatoire? Estimer un risque?

... correspond une famille d'algorithmes (un paradigme)

- Procédure explicite ? \implies Description séquentielle
- ... \implies Algorithmes "Diviser pour Régner", "Glouton", "Probabiliste", "Programmation Dynamique", ...

Comment étudier un algorithme ?

Résoudre un problème (pour un.e informaticien.ne)
c'est proposer une méthode **générique**
qui fournit la bonne réponse à **chaque instance** du problème

Analyser un algorithme, c'est donc :

- comprendre quel **problème** il cherche à résoudre
- savoir comment il **procède**
- identifier les avantages / les dangers / les biais dans sa **mise en application**

À chaque type de problème ...

- Faire la somme de deux nombres ?
- ...
- Étiqueter une image ? Proposer un montant compensatoire? Estimer un risque?

... correspond une famille d'algorithmes (un paradigme)

- Procédure explicite ? \implies Description séquentielle
- ... \implies Algorithmes "Diviser pour Régner", "Glouton", "Probabiliste", "Programmation Dynamique", ...
- Exemples de solutions ? \implies **Algorithmes d'apprentissage** (supervisé)

Éthique et algorithmes

Observation : les algorithmes ont pénétré énormément de systèmes qui ont un impact sur nos sociétés.

- 1 Donne lieu à un débat (passionnel) entre :
 - les promoteurs : efficacité, suppression des biais humains, etc ...
 - les opposants : impacte l'économie, l'emploi, introduit de **nouveaux biais** (discriminations), impact imprévisible sur le long terme, ...
- 2 Lié à différentes problématiques en terme d'éthique
 - transparence/confiance : système APB/ParcourSup, vote électronique...
 - responsabilité : véhicule autonomes, armes létales autonomes, ...
 - vie privée: traces de navigations, plateforme en lignes, ... (cf cours 3)
 - diversité: algorithmes de classement, recommandations, ... (cf cours 2)
 - **équité/discrimination** : prise de décision (prêt bancaire, **décision juridique**)
- 3 Chacune dépend de plusieurs aspects: but, contexte, utilisateurs, **données**, ...
 - offres d'emploi attractifs (via google ad) : proposés aux hommes
 - images renvoyées pour "C.E.O." (google) : seulement 11% de femmes
 - complétion de "transgenders are" : ...

Algorithmes d'apprentissage

Une grande diversité de méthodes

Supervisé

Classification:

- K plus proches voisins
- Naïve bayésienne
- Réseaux de neurones
- Machine à vecteurs de support (SVM)
- Forêts aléatoires
- ...

Régression:

- Régression linéaire
- Régression polynomiale
- Arbres de décision
- Réseaux de neurones
- ...

Non supervisé

Clustering:

- K-means
- Analyse en Composantes Principales (ACP)
- Apprentissage profond
- ...

Apprentissage supervisé

Un algorithme d'apprentissage supervisé (simplifié) dispose :

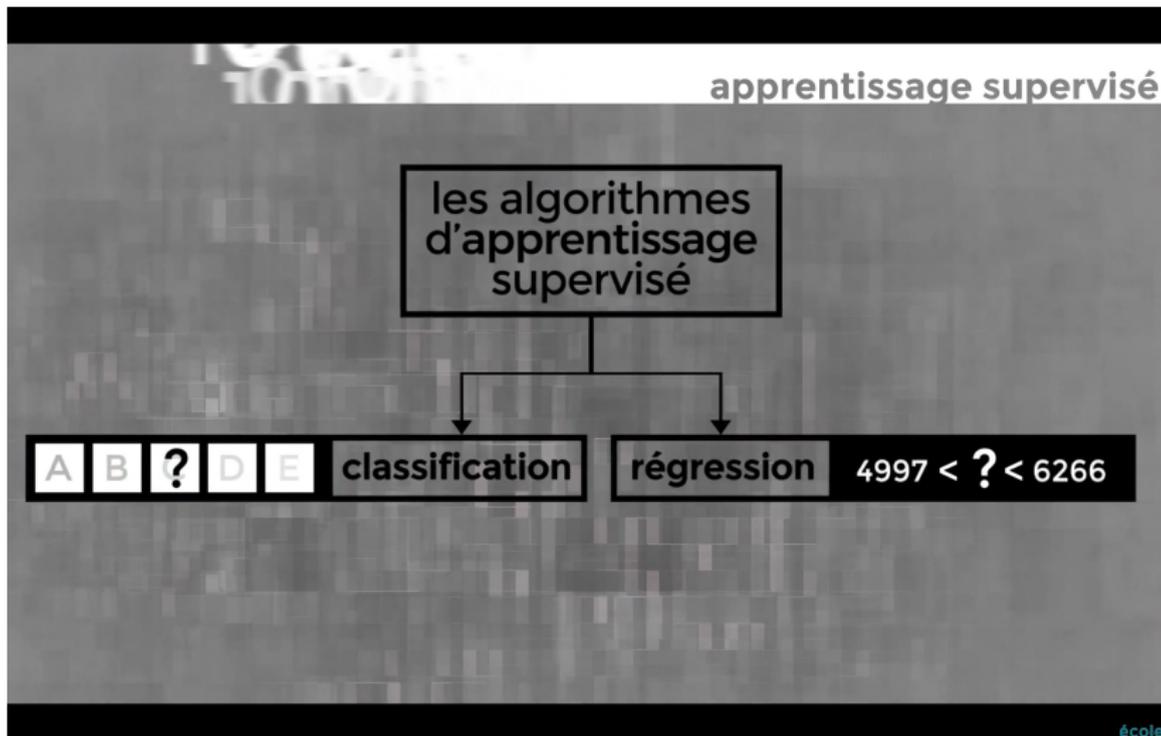
- d'une question (un critère)
- de données (observées) auxquelles poser cette question
- des réponses correctes pour ces données

Le but : avoir une procédure de décision qui associe **le plus souvent possible** la bonne réponse vis à vis de ces données.

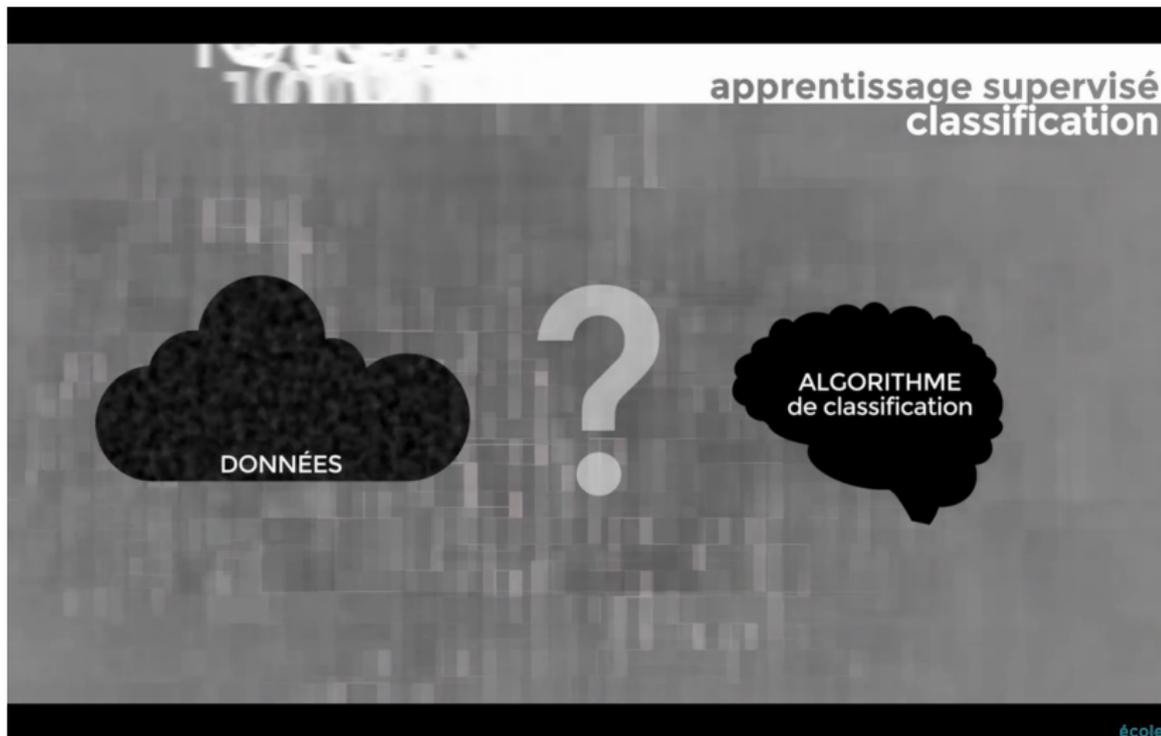
Exemple de problème à résoudre :

- étiqueter le contenu d'une image
- estimer le risque de récurrence d'un prévenu
- estimer un montant compensatoire

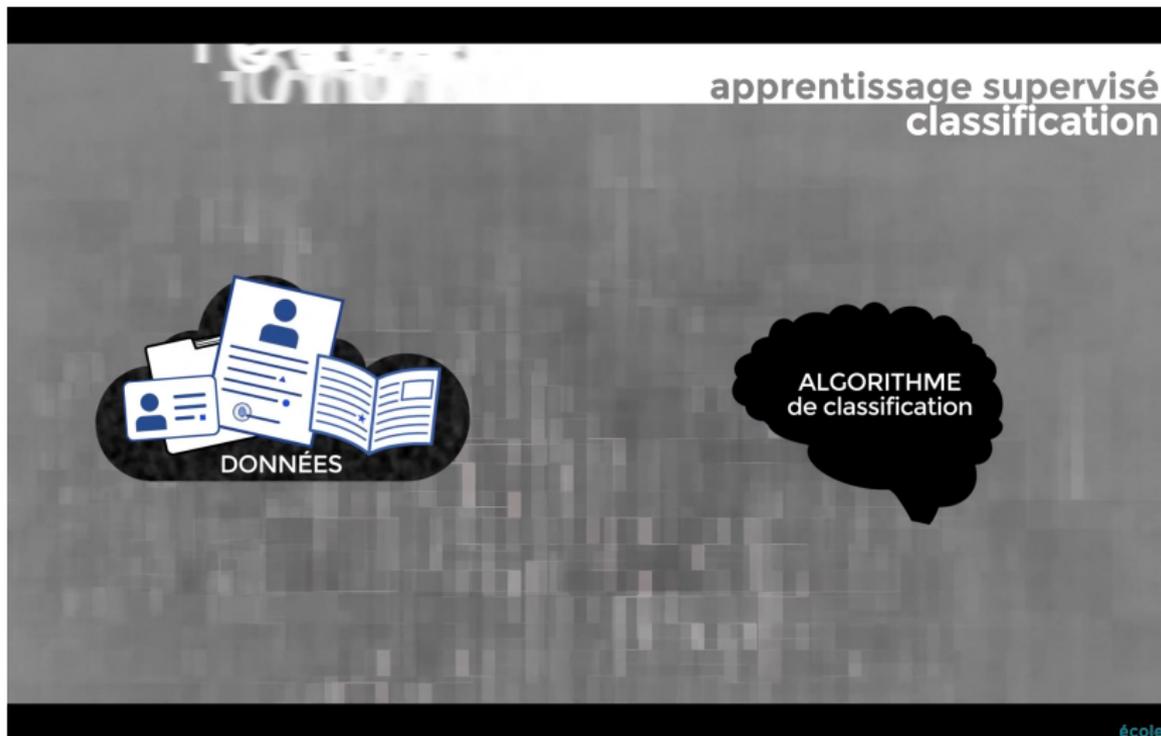
Apprentissage supervisé



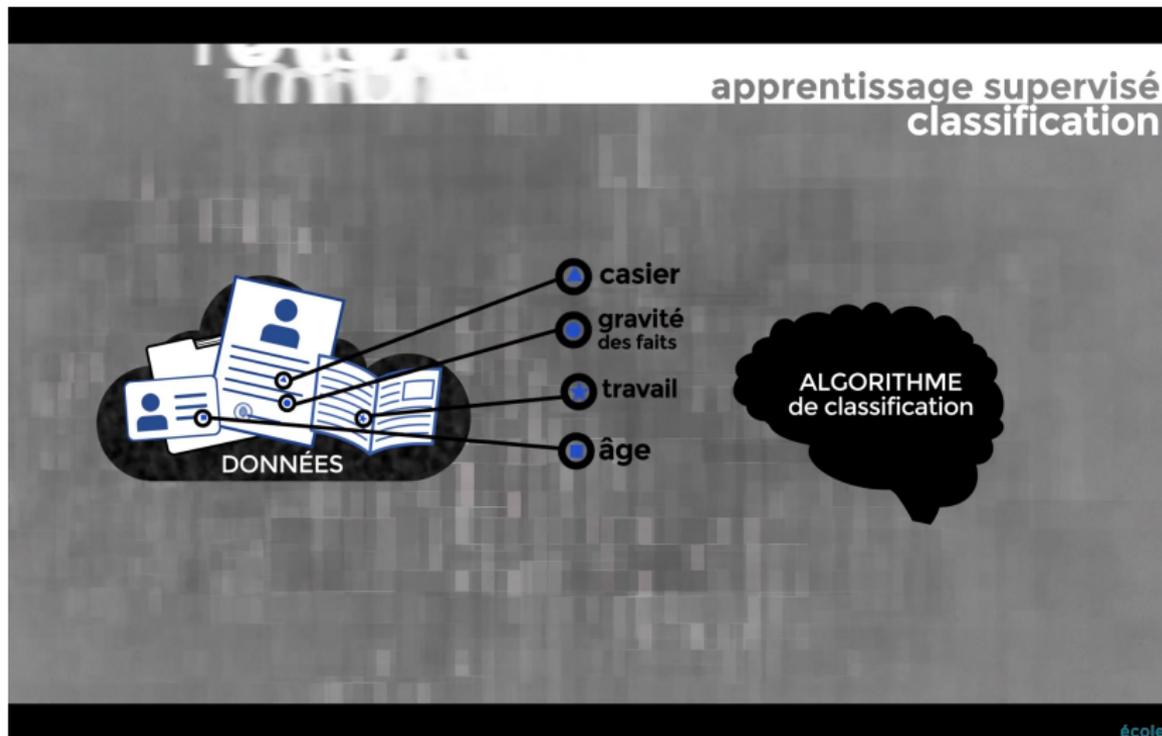
Classification



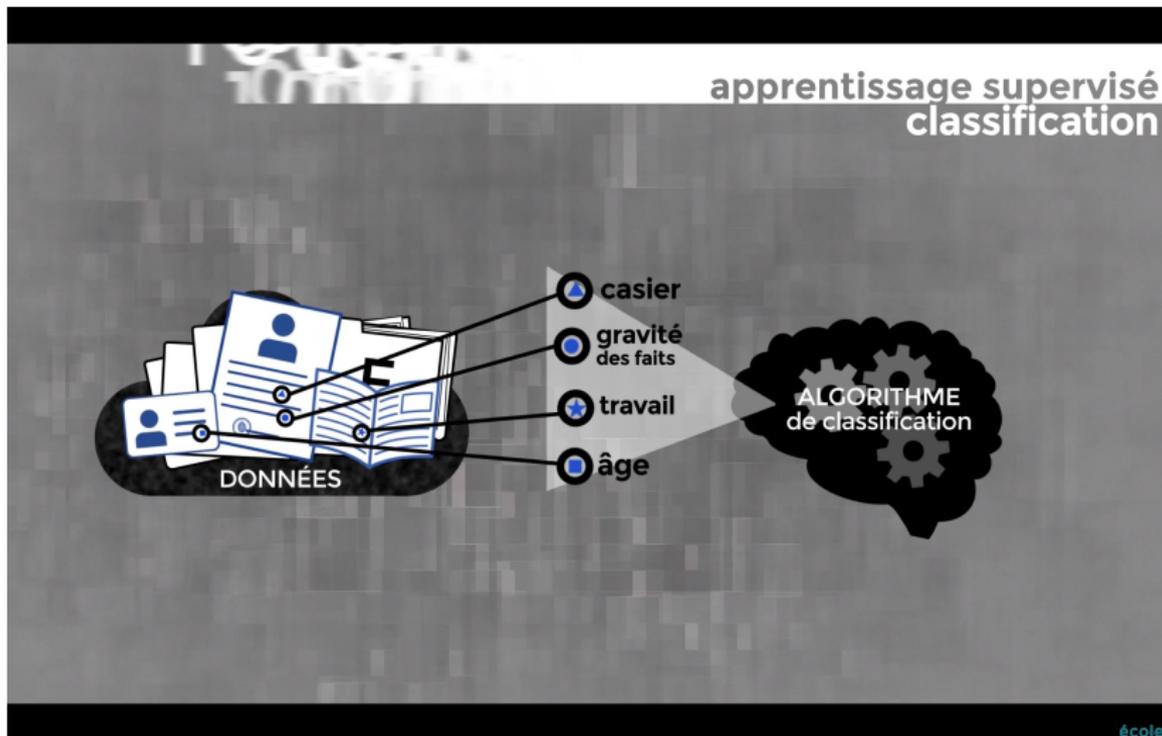
Classification



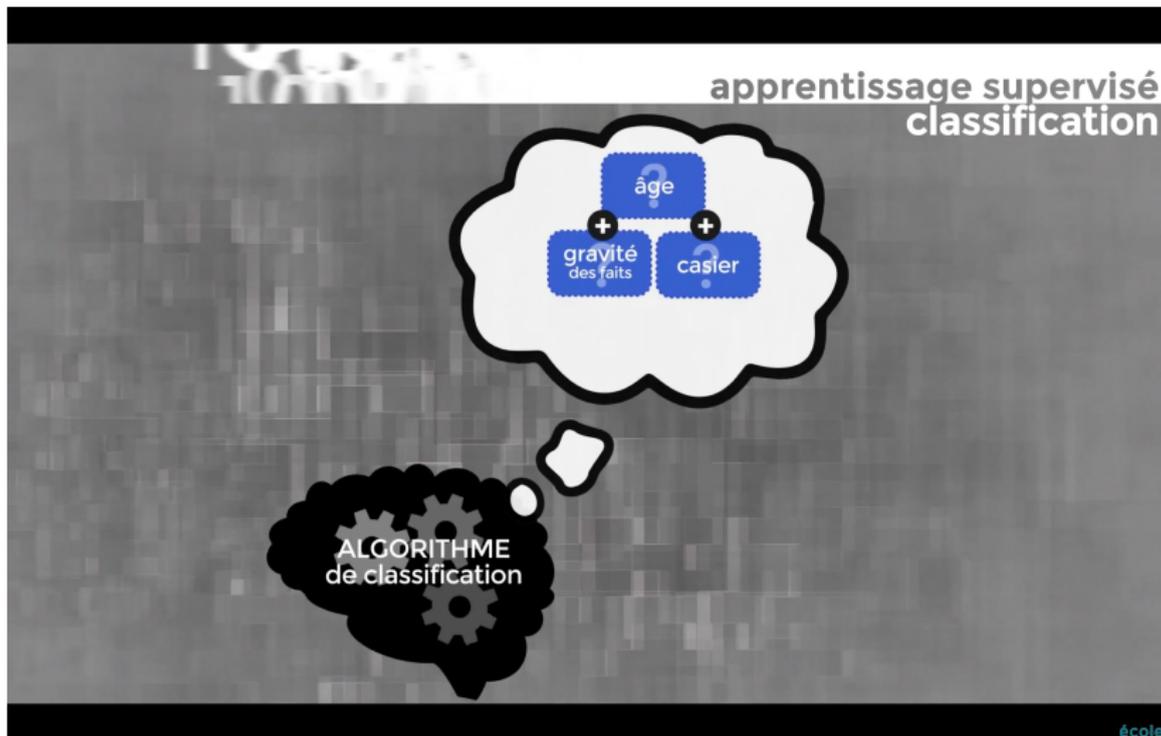
Classification



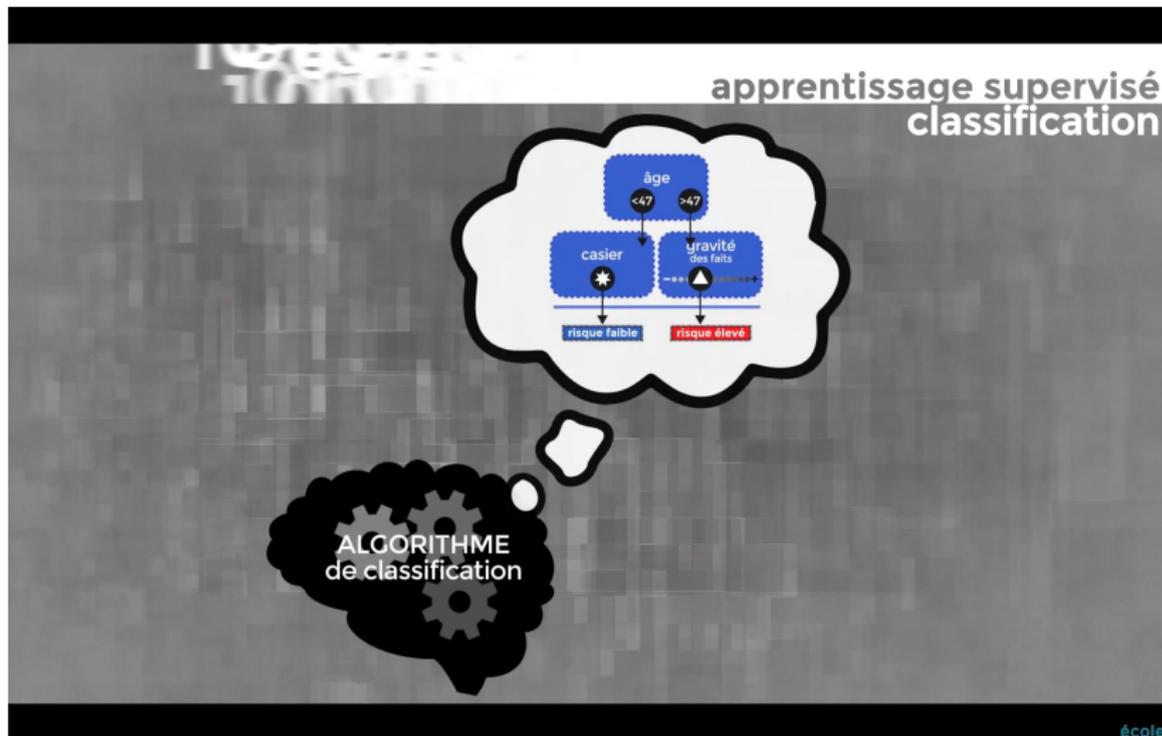
Classification



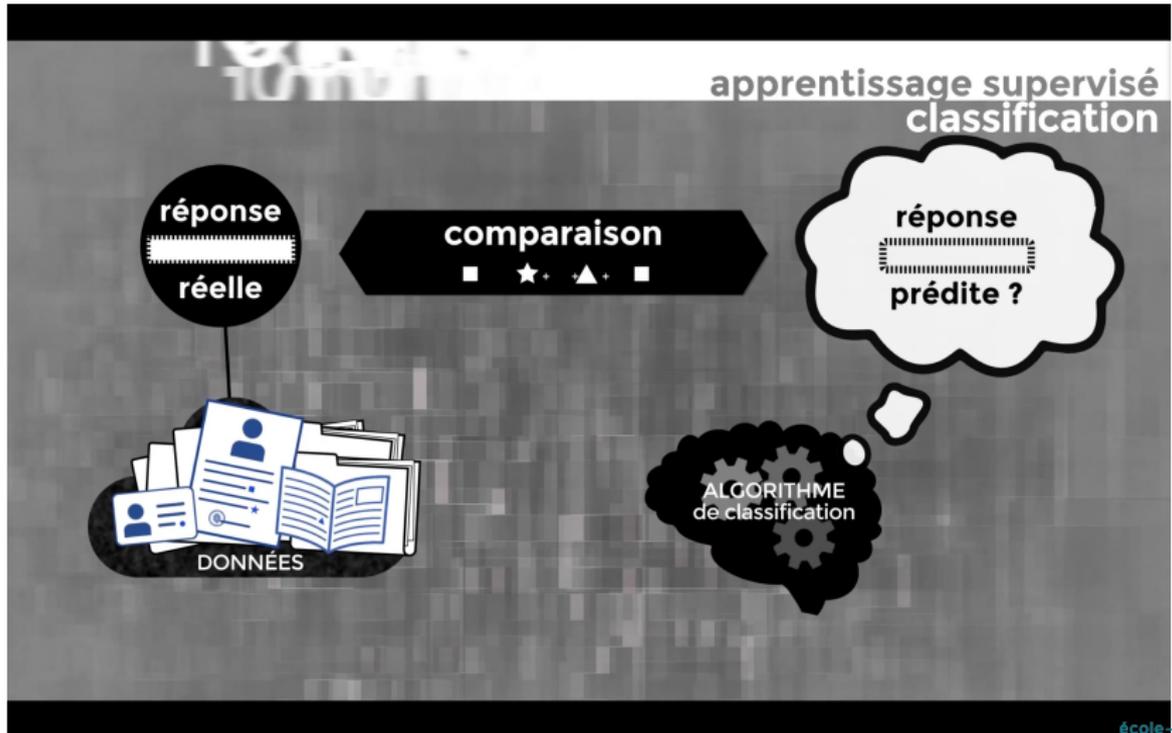
Classification



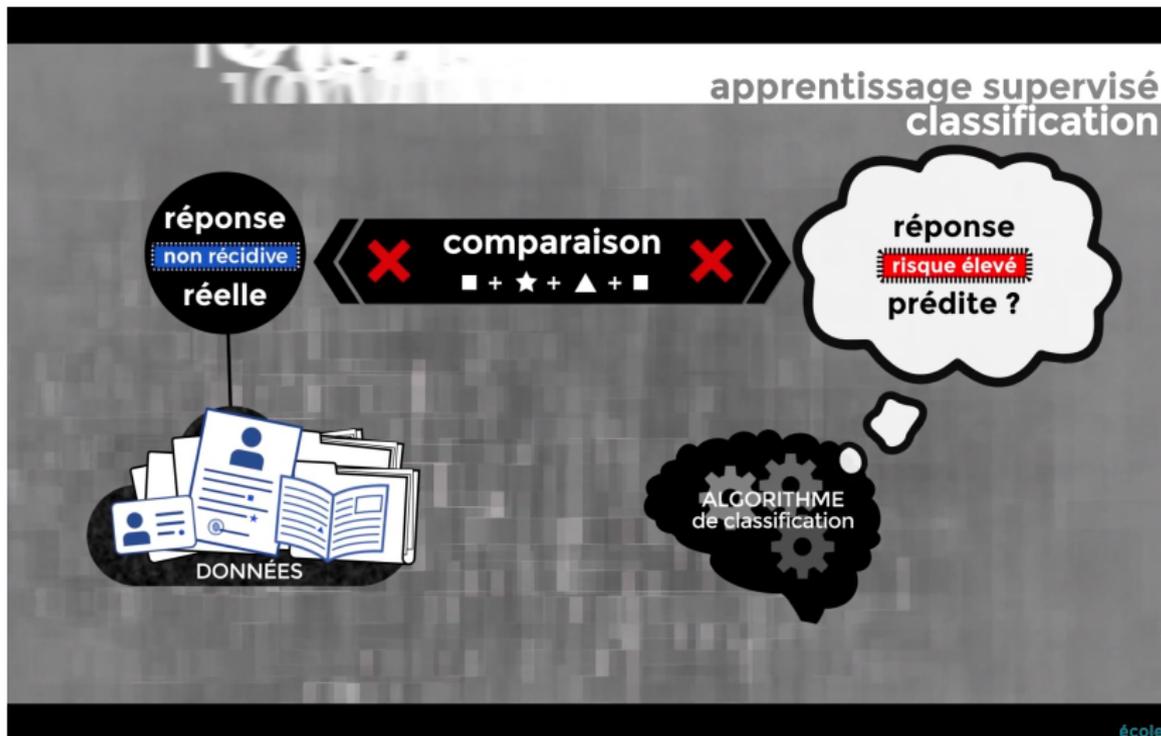
Classification



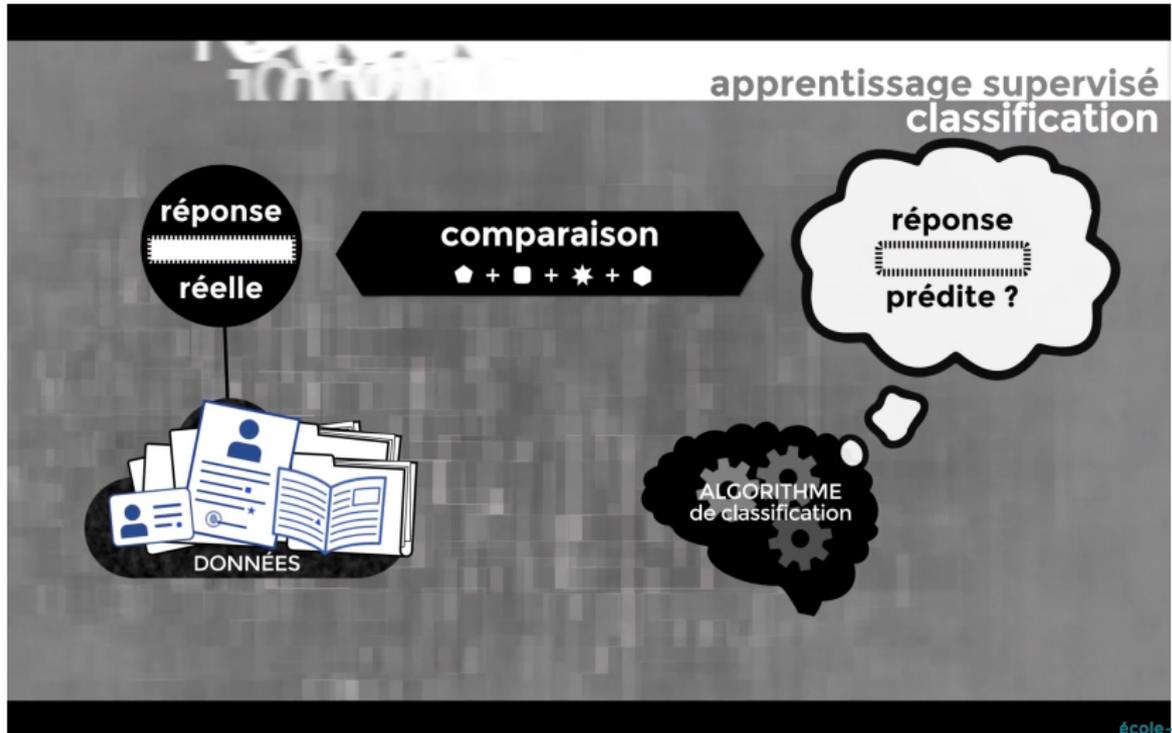
Classification



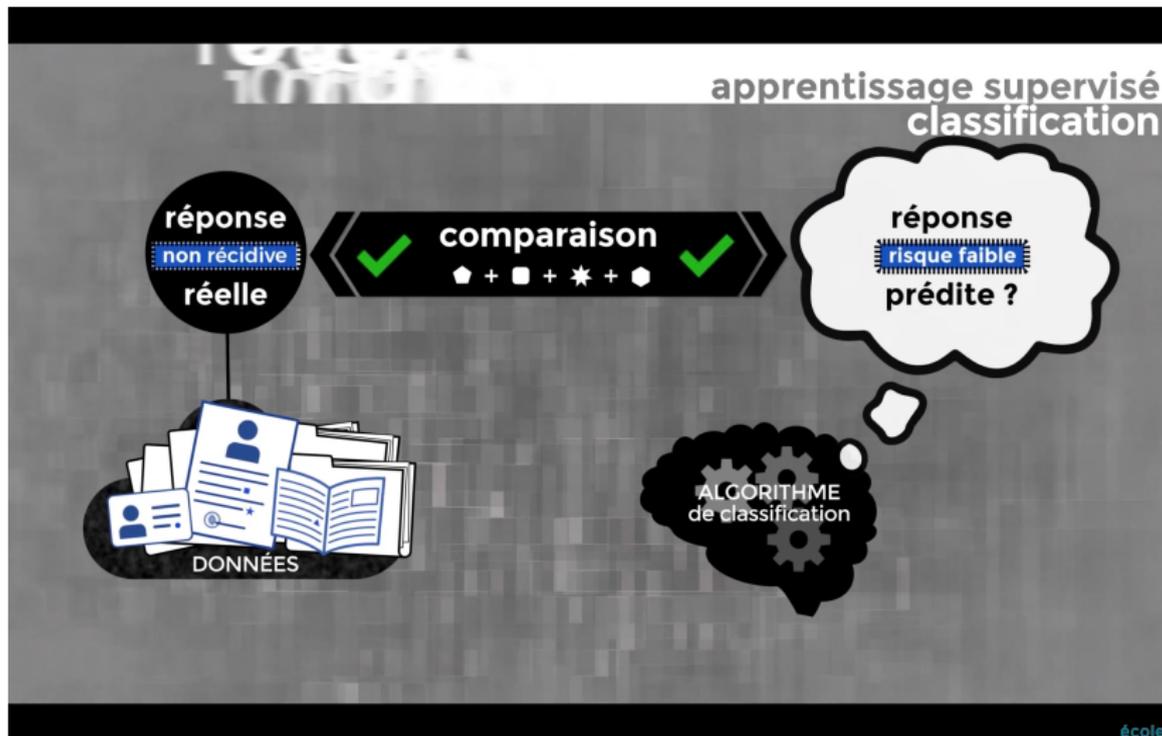
Classification



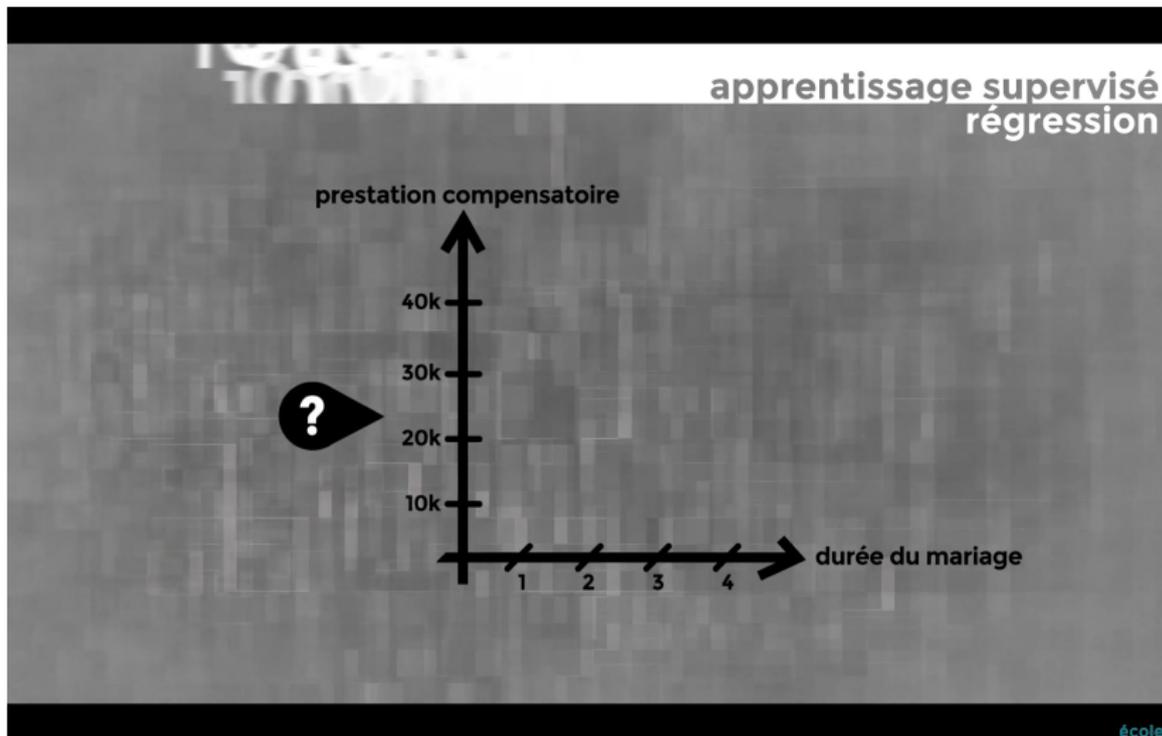
Classification



Classification

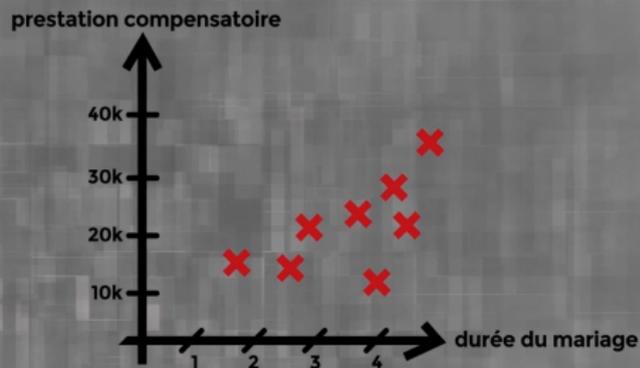


Régressions



Régressions

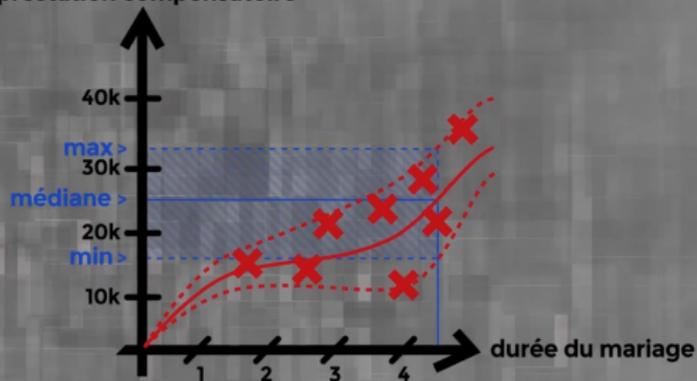
apprentissage supervisé
régression



Régressions

apprentissage supervisé
régression

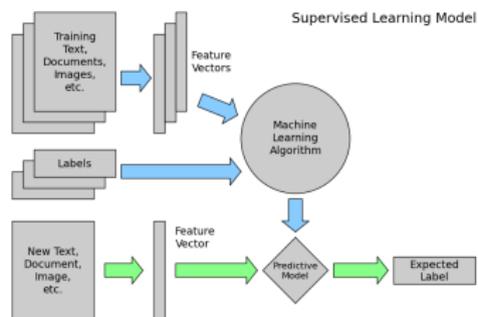
prestation compensatoire



durée du mariage

Apprentissage supervisé

- on extrait une sélection d'information (**critères**) à partir des données
- on donne des valeurs à ces critères (**poids et seuils**)
- on cherche les valeurs qui **optimisent** le succès de la décision



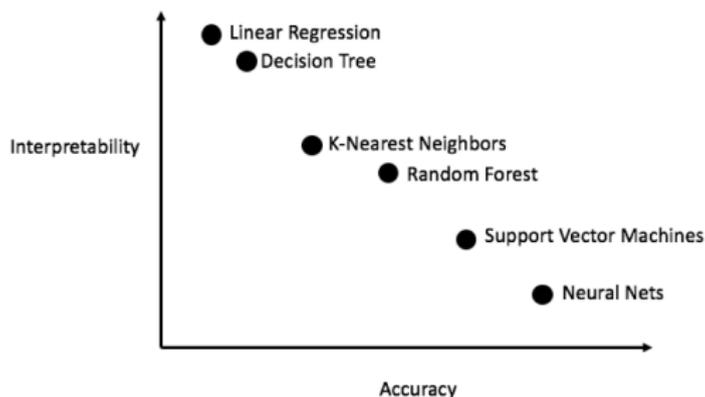
⇒ La sortie de l'algorithme c'est **la valeur de ces poids/seuils**.

Documents : affaires, jugements

Vecteurs : profil d'un prévenu (âge, domicile, travail, casier, ...), d'un contrat (durée, âge, revenu, ...)

Étiquettes : discrète (récidive) ou continue (montant compensatoire)

Précision *vs.* interprétabilité



Classification:

- K plus proches voisins
- Réseaux de neurones
- SVM
- Forêts aléatoires
- ...

Régression:

- Régression linéaire
- Arbres de décision
- Réseaux de neurones
- ...

*L'apprentissage dans le
contexte juridique*

Différents modèles (économiques)

États-Unis

Évaluation du risque lié à une liberté provisoire, liberté conditionnelle, ...

→ logiciels utilisés dans les **juridictions**

- COMPAS (Northpointe / Equivant) : *"Risk assessment tool for criminal justice practitioner"*

France

Chance de succès d'une affaire, montants (compensatoires, dommages et intérêts, ...)

→ logiciels proposés aux **cabinets d'avocats**

- PREDICTICE (2016) : *"Optimisez votre stratégie juridique."*
- CASE LAW ANALYTICS (2017) : *"Quantifier le risque juridique par l'Intelligence Artificielle."*

Royaume-Uni

Chance de succès d'une affaire, prédiction de **décisions**

→ **chatbot** / mise en relation clients ↔ avocats (plus maintenant)

- CASECRUNCH (2017) : *"Find truth in law." et "Solve law"*

Point de vue académique

Human Decisions and Machine Predictions

Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig. *The Quarterly Journal of Economics*, 2018.

- Problème : détention ou liberté provisoire ?
- Résultats (simulations) : réduction des crimes (24.7 %) ou de la surpopulation carcérale (41.9 %)

exemple : 1% des plus "dangereux" sont relâchés dans 48.5% des cas. Parmi ceux-ci, 56.3% sont absents au procès et 62.7% commettent de nouvelles infractions.

A general approach for predicting the behavior of the US Supreme Court

Daniel M. Katz, Michael J. Bommarito, Josh Blackman. *Plos One*, 2017.

- Problème : prédiction des décisions (USSC) et des votes (juges)
- Résultats : prédiction correcte dans 70.2 % des décisions, 71.9 % des votes.

Predicting judicial decisions of the ECHR: a Natural Language Processing perspective

Nikolas Aletras, Dimitrios Tsarapatsanis, Daniel Preotiuc-Pietro, et al. *PeerJ in Computer Science*, 2016.

- Problème : y a-t-il violation d'un article de la Convention ?
- Résultats : prédiction correcte dans 79 % des affaires

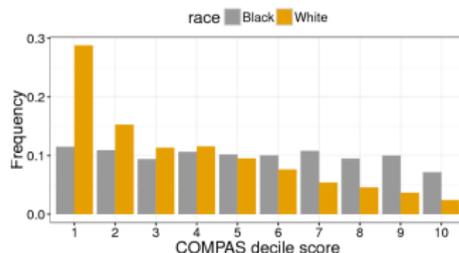
Évaluer un modèle prédictif

Un focus sur l'outil Compas

Gabriel Franco

COMPAS : Tour d'horizon d'un algorithme utilisé dans la justice américaine.

Le débat



Évaluation de COMPAS (effectuée par PROPUBLICA)

Machine bias. J. Angwin, J. Larson, S. Mattu, L. Kirchner. *ProPublica*. 2016, May.

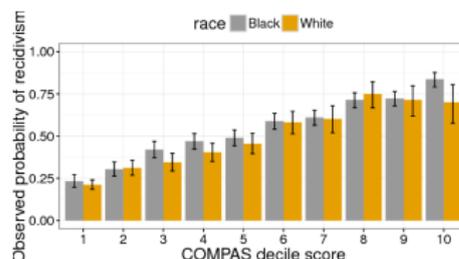
<https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>

*There's software used across the country to predict future criminals. And it's **biased against blacks.***

FPR parity Afro-américains qui n'ont **pas récidivé** sont plus souvent catégorisés à **haut risque** que les blancs (45 % vs. 23 %)

FNR parity Blancs qui ont **récidivé** sont plus souvent catégorisés à **faible risque** que les afro-américains (48 % vs. 28 %)

Le débat



Réponse de NORTHPOINTE

COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity. W. Dieterich, C. Mendoza, T. Brennan. Northpointe Inc. Research Department. 2016, July.

*We strongly reject the conclusion that the COMPAS risk scales are racially biased against blacks. ProPublica [...] did not take into account the **differential base rates** of recidivism for blacks and whites. ProPublica [...] wrongly defined classification terms and **measures of discrimination**.*

faux positifs Afro-américains catégorisés à **haut risque** ont la même probabilité de ne **pas récidiver** que les blancs (37 % vs. 41 %)

faux négatifs Afro-américains catégorisés à **faible risque** ont la même probabilité de **récidiver** que les blancs (35 % vs. 29 %)

Évaluer un modèle prédictif

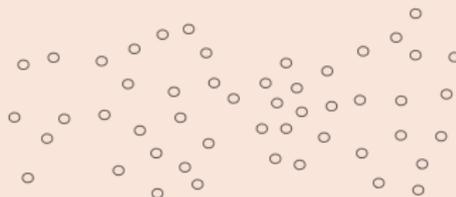
Le problème

- Considérons une tâche de classification binaire
→ récidive / non récidive
- Considérons un modèle prédictif (issu par exemple d'un algorithme d'apprentissage supervisé).
→ risque récidive élevé / faible

Comment évaluer la qualité du modèle ?

Matrice de confusion

● Ce qui est recherché



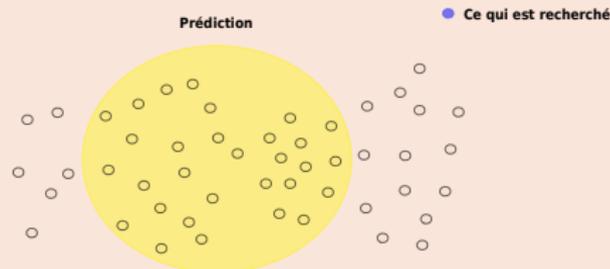
Évaluer un modèle prédictif

Le problème

- Considérons une tâche de classification binaire
→ **récidive** / **non récidive**
- Considérons un modèle prédictif (issu par exemple d'un algorithme d'apprentissage supervisé).
→ risque récidive **élevé** / **faible**

Comment évaluer la qualité du modèle ?

Matrice de confusion



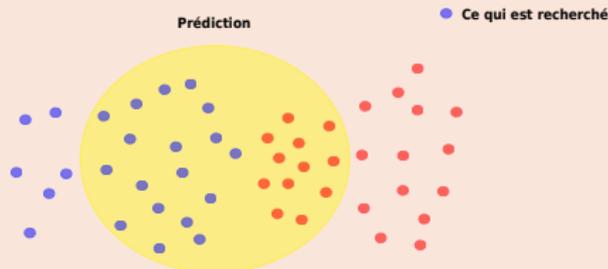
Évaluer un modèle prédictif

Le problème

- Considérons une tâche de classification binaire
→ **récidive** / **non récidive**
- Considérons un modèle prédictif (issu par exemple d'un algorithme d'apprentissage supervisé).
→ risque récidive **élevé** / **faible**

Comment évaluer la qualité du modèle ?

Matrice de confusion



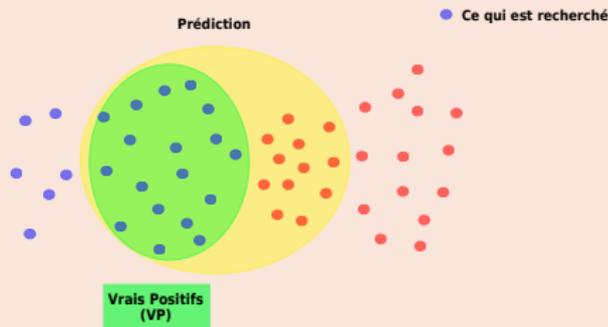
Évaluer un modèle prédictif

Le problème

- Considérons une tâche de classification binaire
→ récidive / non récidive
- Considérons un modèle prédictif (issu par exemple d'un algorithme d'apprentissage supervisé).
→ risque récidive élevé / faible

Comment évaluer la qualité du modèle ?

Matrice de confusion



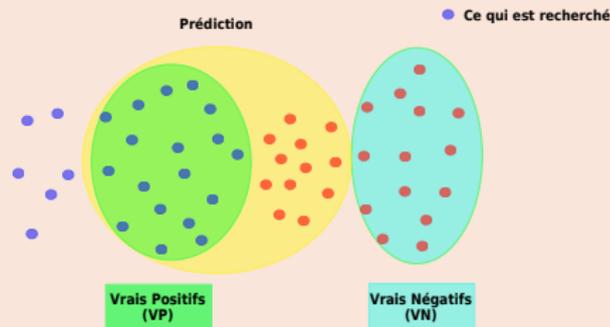
Évaluer un modèle prédictif

Le problème

- Considérons une tâche de classification binaire
→ **récidive** / **non récidive**
- Considérons un modèle prédictif (issu par exemple d'un algorithme d'apprentissage supervisé).
→ risque récidive **élevé** / **faible**

Comment évaluer la qualité du modèle ?

Matrice de confusion



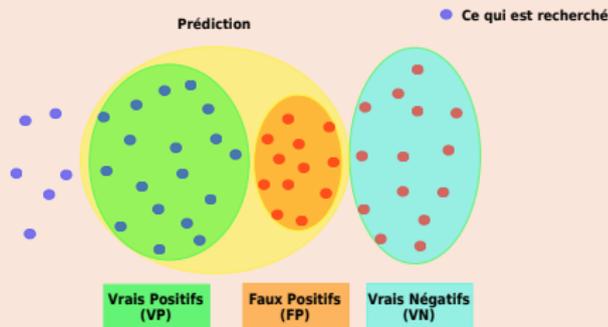
Évaluer un modèle prédictif

Le problème

- Considérons une tâche de classification binaire
→ **récidive** / **non récidive**
- Considérons un modèle prédictif (issu par exemple d'un algorithme d'apprentissage supervisé).
→ risque récidive **élevé** / **faible**

Comment évaluer la qualité du modèle ?

Matrice de confusion



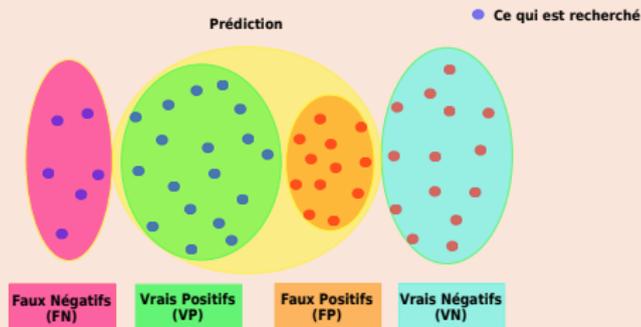
Évaluer un modèle prédictif

Le problème

- Considérons une tâche de classification binaire
→ récidive / non récidive
- Considérons un modèle prédictif (issu par exemple d'un algorithme d'apprentissage supervisé).
→ risque récidive élevé / faible

Comment évaluer la qualité du modèle ?

Matrice de confusion



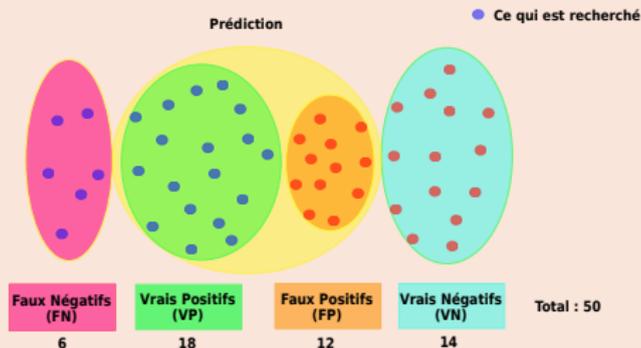
Évaluer un modèle prédictif

Le problème

- Considérons une tâche de classification binaire
→ **récidive** / **non récidive**
- Considérons un modèle prédictif (issu par exemple d'un algorithme d'apprentissage supervisé).
→ risque récidive **élevé** / **faible**

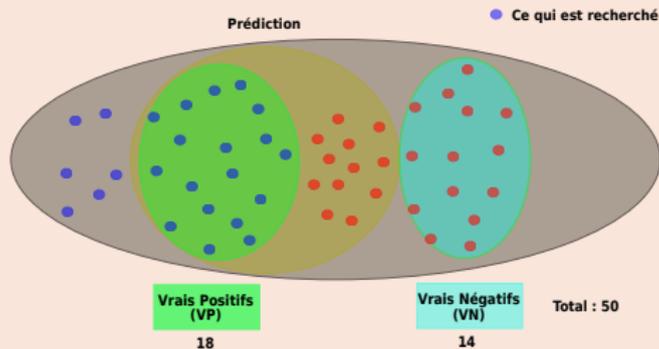
Comment évaluer la qualité du modèle ?

Matrice de confusion



Évaluer un modèle prédictif

Matrice de confusion

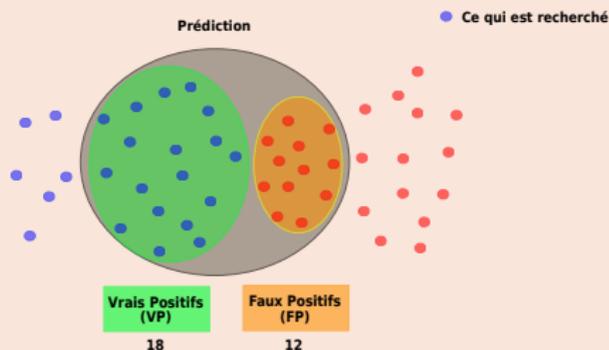


Quantifier la qualité (succès) du modèle

$$\text{Accuracy ACC} = \frac{VP+VN}{P+N} = \frac{\text{bonne predictions}}{\text{total}} = \frac{32}{50} = 0.64 \quad \text{exactitude / justesse}$$

Évaluer un modèle prédictif

Matrice de confusion



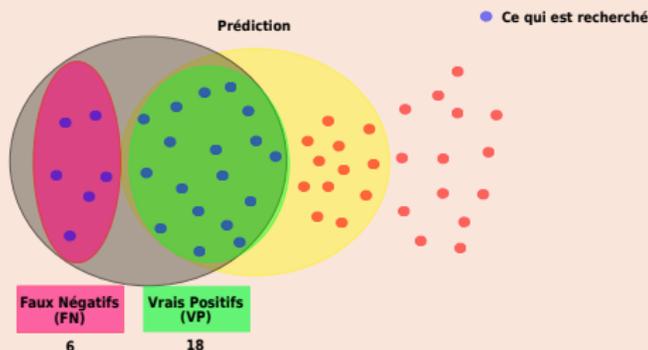
Quantifier la qualité (succès) du modèle

Accuracy $ACC = \frac{VP+VN}{P+N} = \frac{\text{bonne predictions}}{\text{total}} = \frac{32}{50} = 0.64$ exactitude / justesse

Precision $PR = \frac{VP}{VP+FP} = \frac{\text{positifs predicts}}{\text{nb predictions}} = \frac{18}{30} = 0.6$ précision

Évaluer un modèle prédictif

Matrice de confusion



Quantifier la qualité (succès) du modèle

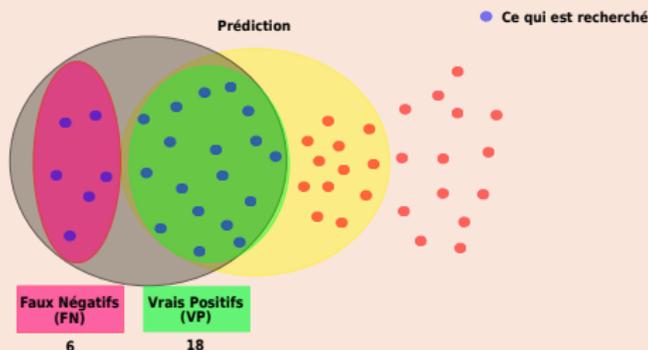
Accuracy $ACC = \frac{VP+VN}{P+N} = \frac{\text{bonne predictions}}{\text{total}} = \frac{32}{50} = 0.64$ exactitude / justesse

Precision $PR = \frac{VP}{VP+FP} = \frac{\text{positifs predicts}}{\text{nb predictions}} = \frac{18}{30} = 0.6$ précision

Recall $RC = \frac{VP}{VP+FN} = \frac{\text{positifs predicts}}{\text{nb positifs}} = \frac{18}{24} = 0.75$ rappel / sensibilité

Évaluer un modèle prédictif

Matrice de confusion



Quantifier la qualité (succès) du modèle

Accuracy $ACC = \frac{VP+VN}{P+N} = \frac{\text{bonne predictions}}{\text{total}} = \frac{32}{50} = 0.64$ exactitude / justesse

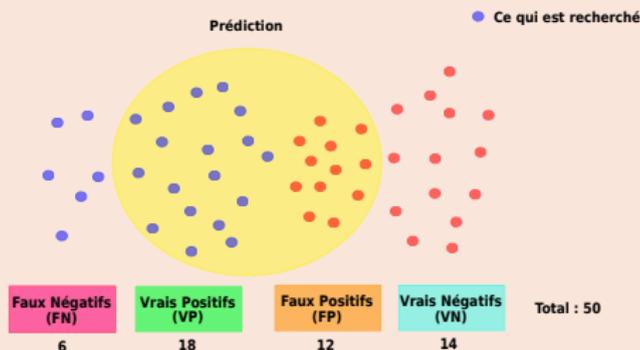
Precision $PR = \frac{VP}{VP+FP} = \frac{\text{positifs predicts}}{\text{nb predictions}} = \frac{18}{30} = 0.6$ précision

Recall $RC = \frac{VP}{VP+FN} = \frac{\text{positifs predicts}}{\text{nb positifs}} = \frac{18}{24} = 0.75$ rappel / sensibilité

F-score $F = \frac{2 \times PR \times RC}{PR+RC} = 0.67$

Évaluer un modèle prédictif

Matrice de confusion



Quantifier la qualité (succès) du modèle

Accuracy $ACC = \frac{VP+VN}{P+N} = \frac{\text{bonne predictions}}{\text{total}} = \frac{32}{50} = 0.64$ exactitude / justesse

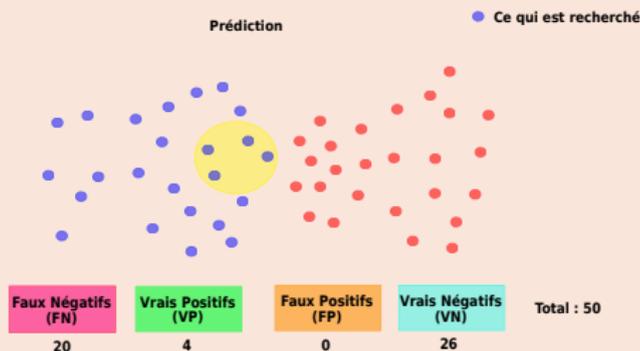
Precision $PR = \frac{VP}{VP+FP} = \frac{\text{positifs predicts}}{\text{nb predictions}} = \frac{18}{30} = 0.6$ précision

Recall $RC = \frac{VP}{VP+FN} = \frac{\text{positifs predicts}}{\text{nb positifs}} = \frac{18}{24} = 0.75$ rappel / sensibilité

F-score $F = \frac{2 \times PR \times RC}{PR+RC} = 0.67$

Évaluer un modèle prédictif

Matrice de confusion



Quantifier la qualité (succès) du modèle

Accuracy $ACC = \frac{VP+VN}{P+N} = \frac{\text{bonne predictions}}{\text{total}} = \frac{30}{50} = 0.6$ exactitude / justesse

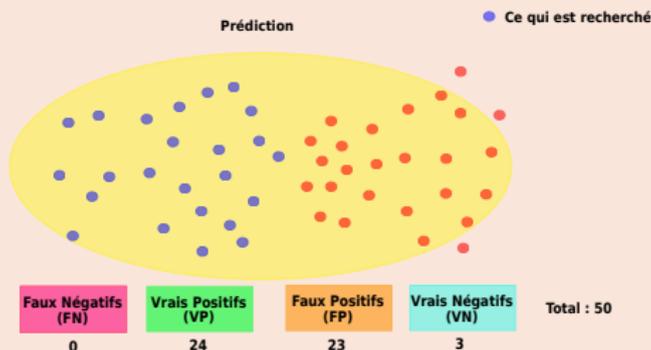
Precision $PR = \frac{VP}{VP+FP} = \frac{\text{positifs predicts}}{\text{nb predictions}} = \frac{4}{4} = 1.0$ précision

Recall $RC = \frac{VP}{VP+FN} = \frac{\text{positifs predicts}}{\text{nb positifs}} = \frac{4}{24} = 0.17$ rappel / sensibilité

F-score $F = \frac{2 \times PR \times RC}{PR+RC} = 0.27$

Évaluer un modèle prédictif

Matrice de confusion



Quantifier la qualité (succès) du modèle

Accuracy $ACC = \frac{VP+VN}{P+N} = \frac{\text{bonne predictions}}{\text{total}} = \frac{27}{50} = 0.54$ exactitude / justesse

Precision $PR = \frac{VP}{VP+FP} = \frac{\text{positifs predicts}}{\text{nb predictions}} = \frac{24}{47} = 0.51$ précision

Recall $RC = \frac{VP}{VP+FN} = \frac{\text{positifs predicts}}{\text{nb positifs}} = \frac{24}{24} = 1.0$ rappel / sensibilité

F-score $F = \frac{2 \times PR \times RC}{PR+RC} = 0.68$

Différentes notions d'équité

Le problème

- Considérons une tâche de classification binaire (observation O)
→ récidive / non récidive
- Considérons un modèle prédictif M (issu par exemple d'un algorithme d'apprentissage supervisé).
→ risque récidive élevé / faible

Différentes notions d'équité

Le problème

- Considérons une tâche de classification binaire (observation O)
→ récidive / non récidive
- Considérons un modèle prédictif M (issu par exemple d'un algorithme d'apprentissage supervisé).
→ risque récidive élevé / faible
- Considérons deux groupes G distincts : \circ et \star
→ blancs / afro-américains
→ hommes / femmes

Comment évaluer si les groupes sont traités de manière équitable par le modèle ?

Différentes notions d'équité

Le problème

- Considérons une tâche de classification binaire (observation O)
→ récidive / non récidive
- Considérons un modèle prédictif M (issu par exemple d'un algorithme d'apprentissage supervisé).
→ risque récidive élevé / faible
- Considérons deux groupes G distincts : \circ et \star
→ blancs / afro-américains
→ hommes / femmes

Comment évaluer si les groupes sont traités de manière équitable par le modèle ?

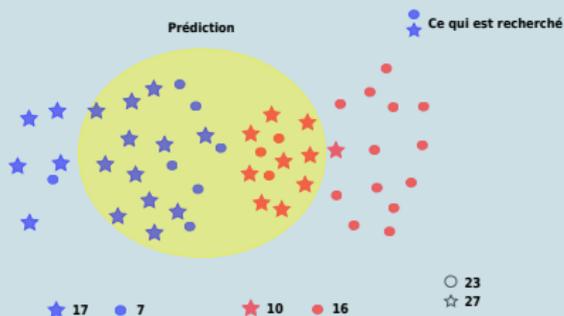
Plusieurs propositions (*group fairness*) :

- Parité démographique (*Demographic parity*)
- Égalité en opportunité (*Equal opportunity*)
- Parité en faux-positif (*False positive rate parity*)
- Égalité des chances (*Equalized odds*)
- Parité en taux de prédiction (*Predictive rate parity*)
- ...

Quelle relation entre ces notions ?

Parité démographique

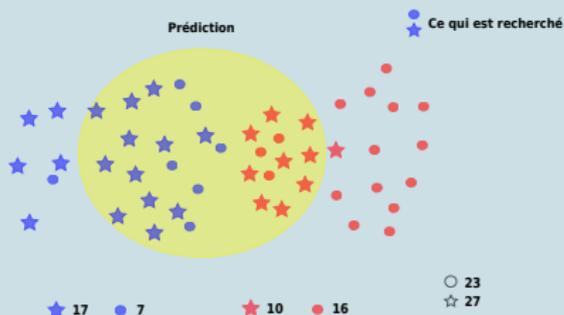
Matrice de confusion



Demographic parity (ou statistical parity)

Parité démographique

Matrice de confusion



Demographic parity (ou statistical parity)

Un modèle est réputé préserver la parité démographique si :

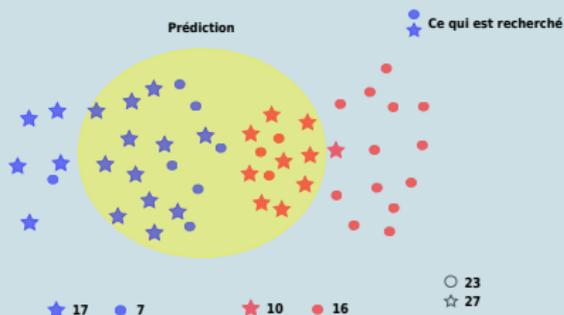
$$P(M = \text{●} \mid G = \text{○}) = P(M = \text{★} \mid G = \text{☆})$$

$$P(M = \text{●} \mid G = \text{○}) = P(M = \text{★} \mid G = \text{☆})$$

La probabilité qu'un individu reçoive une prédiction (**positive** ou **négative**) doit être la même quelque soit le groupe auquel il appartient.

Parité démographique

Matrice de confusion



Demographic parity (ou statistical parity)

Un modèle est réputé préserver la parité démographique si :

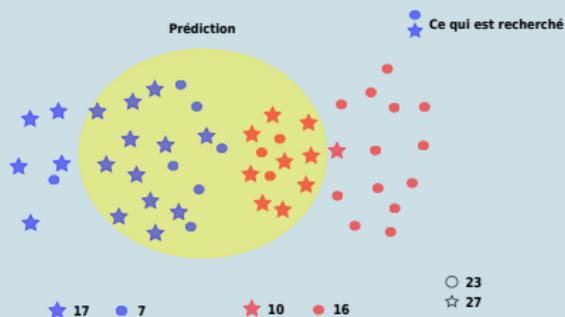
$$0.39 \leftarrow P(M = \bullet \mid G = \circ) = P(M = \star \mid G = \star) \rightarrow 0.78$$

$$0.61 \leftarrow P(M = \bullet \mid G = \star) = P(M = \star \mid G = \circ) \rightarrow 0.22$$

La probabilité qu'un individu reçoive une prédiction (**positive** ou **négative**) doit être la même quelque soit le groupe auquel il appartient.

Égalité en *opportunité*

Matrice de confusion



Equal opportunity

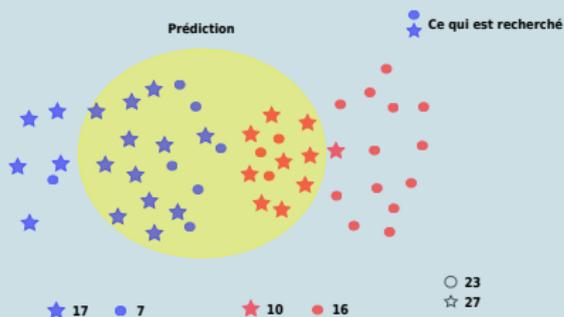
Un modèle est réputé préserver l'égalité en opportunité si :

$$P(M = \bullet | G = \circ, O = \bullet) = P(M = \star | G = \star, O = \star)$$

La probabilité qu'un individu **positif** reçoive une prédiction **positive** doit être la même quelque soit le groupe auquel il appartient.

Égalité en *opportunité*

Matrice de confusion



Equal opportunity

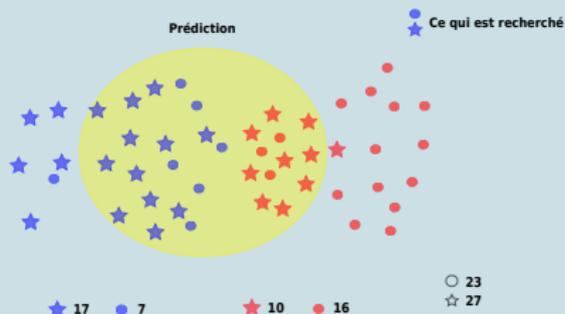
Un modèle est réputé préserver l'égalité en opportunité si :

$$0.86 \leftarrow P(M = \bullet | G = \circ, O = \bullet) = P(M = \star | G = \star, O = \star) \rightarrow 0.71$$

La probabilité qu'un individu **positif** reçoive une prédiction **positive** doit être la même quelque soit le groupe auquel il appartient.

Parité vis-à-vis des faux-positifs

Matrice de confusion



False positive rate parity

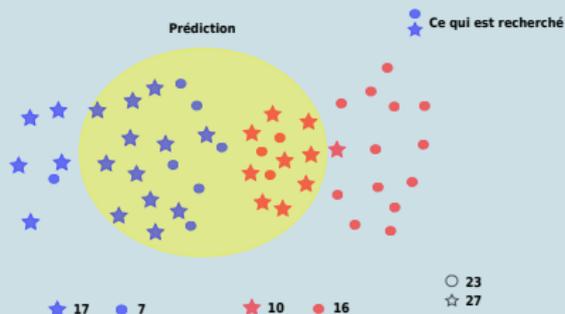
Un modèle est réputé préserver la parité en faux-positif si :

$$P(M = \bullet | G = \circ, O = \bullet) = P(M = \star | G = \star, O = \star)$$

La probabilité qu'un individu **néгатif** reçoive une prédiction **positive** doit être la même quelque soit le groupe auquel il appartient.

Parité vis-à-vis des faux-positifs

Matrice de confusion



False positive rate parity

Un modèle est réputé préserver la parité en faux-positif si :

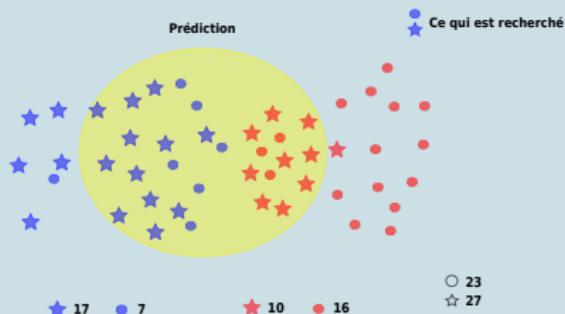
$$0.19 \leftarrow P(M = \bullet | G = \circ, O = \bullet) = P(M = \star | G = \star, O = \star) \rightarrow 0.90$$

La probabilité qu'un individu **néglatif** reçoive une prédiction **positive** doit être la même quelque soit le groupe auquel il appartient.

C'est exactement ce qui est mis en défaut par PROPUBLICA !!

Parité vis-à-vis des prédictions

Matrice de confusion



Prédictive rate parity

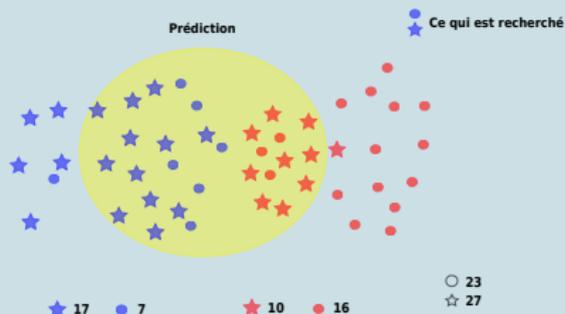
Un modèle est réputé préserver la parité en taux de prédiction si :

$$P(O = \bullet | G = \circ, M = \bullet) = P(O = \star | G = \star, M = \star)$$

La probabilité qu'un individu prédit comme **positif** soit effectivement **positif** (resp. négatif) doit être la même quelque soit le groupe auquel il appartient.

Parité vis-à-vis des prédictions

Matrice de confusion



Prédictive rate parity

Un modèle est réputé préserver la parité en taux de prédiction si :

$$0.67 \leftarrow P(O = \bullet \mid G = \circ, M = \bullet) = P(O = \star \mid G = \star, M = \star) \rightarrow 0.57$$

La probabilité qu'un individu prédit comme **positif** soit effectivement **positif** (resp. négatif) doit être la même quelque soit le groupe auquel il appartient.

C'est exactement ce qui est mis en avant par NORTHPOINT/EQUIVANT !!

Que faire de ces notions ?

Toutes ces notions d'équité sont souhaitables mais ...

Que faire de ces notions ?

Toutes ces notions d'équité sont souhaitables mais ...

Théorème d'impossibilité

À moins d'avoir un modèle prédictif parfait (oracle) ou que les groupes aient exactement la même proportion d'individus positifs (*equal base rate*), un modèle ne peut pas satisfaire à la fois :

- la parité démographique
- l'égalité des chances (liée à la propriété mise en avant par PROPUBLICA)
- la parité en taux de prédiction (mis en avant par EQUIVANT)

Que faire de ces notions ?

Toutes ces notions d'équité sont souhaitables mais ...

Théorème d'impossibilité

À moins d'avoir un modèle prédictif parfait (oracle) ou que les groupes aient exactement la même proportion d'individus positifs (*equal base rate*), un modèle ne peut pas satisfaire à la fois :

- la parité démographique
- l'égalité des chances (liée à la propriété mise en avant par PROPUBLICA)
- la parité en taux de prédiction (mis en avant par EQUIVANT)

Pire : ces trois propriétés sont incompatibles **deux à deux** !

Discussion

Impact sur la justice

- Comment les juges utilisent ces outils ? (affaire *Paul Zilly*)
- Quelles recours ? (affaire *State vs. Loomis*)

Réflexions en vrac

- **Théorème d'impossibilité** : pas seulement pour l'apprentissage ...
- sources de biais (23) : *selection bias, information bias, confounding factors*, ...
- **Intégrer** de l'équité dans les modèles : approches *preprocessing, postprocessing* et *fairness-aware algorithms*.
- Biais cognitifs : choix des critères, concepts d'*automation bias, moral buffer*, ...
- Outils présentés comme utiles au niveau **individuel** mais basés sur calcul **global**
- Corrélation n'est pas causalité.

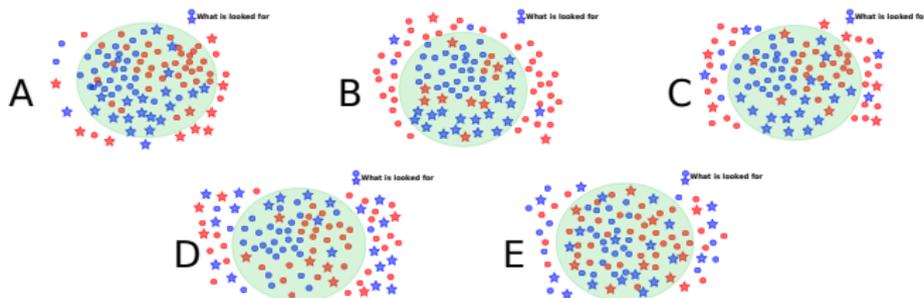
Constat pas très rassurant mais et si nous pouvions :

- **contrôler la distance** à une équité optimale pour les différentes notions ? (en relâchant les contraintes)
Cela va de pair avec une **perte en efficacité** ...
- mettre les praticiens (juges, avocats, ...) en position d'**experts** ?
Cela va de pair avec **un coût** (énergie et investissement dans la formation) ...

Quel type de juge êtes-vous ?

<https://www.sli.do/>

67498

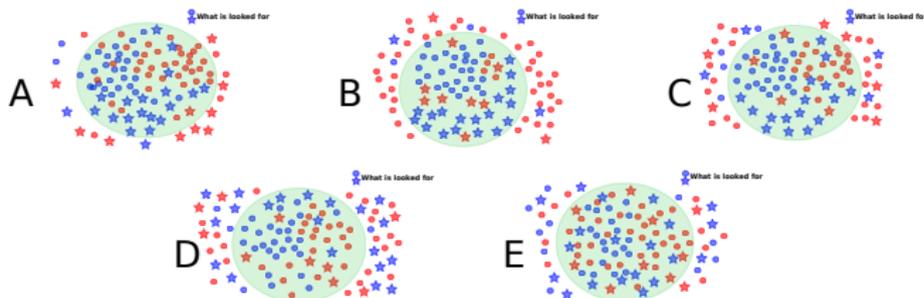


Modèle	Acc	Pr	Recall	F-score	Dem Parity	FP Parity	PR Parity
A							
B							
C							
D							
E							

Quel type de juge êtes-vous ?

<https://www.sli.do/>

67498



Modèle	Acc	Pr	Recall	F-score	Dem Parity	FP Parity	PR Parity
A	0.60	0.56	0.92	0.69	0.90 0.63	0.87 0.10	0.43 0.95
B	0.83	0.78	0.92	0.84	0.47 0.87	0.12 0.80	0.85 0.69
C	0.69	0.64	0.88	0.74	0.65 0.79	0.49 0.56	0.57 0.78
D	0.57	0.56	0.64	0.60	0.66 0.37	0.50 0.50	0.57 0.55
E	0.45	0.46	0.66	0.55	0.71 0.70	0.75 0.80	0.40 0.62

Questions?

<http://tarissan.complexnetworks.fr/>