

# Apprentissage automatique (ML)

MGL7320 - Ingénierie logicielle des systèmes d'IA



# Outline

- Données
  - Exploration des caractéristiques des données
  - Faire face à (certains) problèmes liés aux données
- Modèles
  - Explorez les performances de différents modèles
  - Optimization des modèles
- Évaluation du modèle
  - Choisissez les mesures de qualité appropriées
  - Établissement d'un modèle de référence
  - Comprendre/expliquer le modèle

# Qu'est-ce que l'apprentissage automatique?

Est un champ d'étude de l'intelligence artificielle qui se fonde sur des approches mathématiques et statistiques pour...

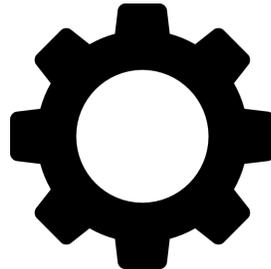
...donner aux machines la capacité **d'apprendre** à partir de **données** et prend une **décision**.

**Idée clé** : apprendre automatiquement sans être programmé encore et encore

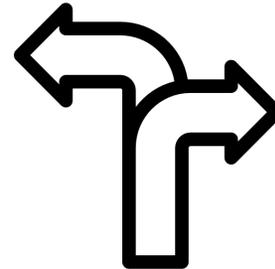
# Aperçu d'un système de ML « typique »



Données

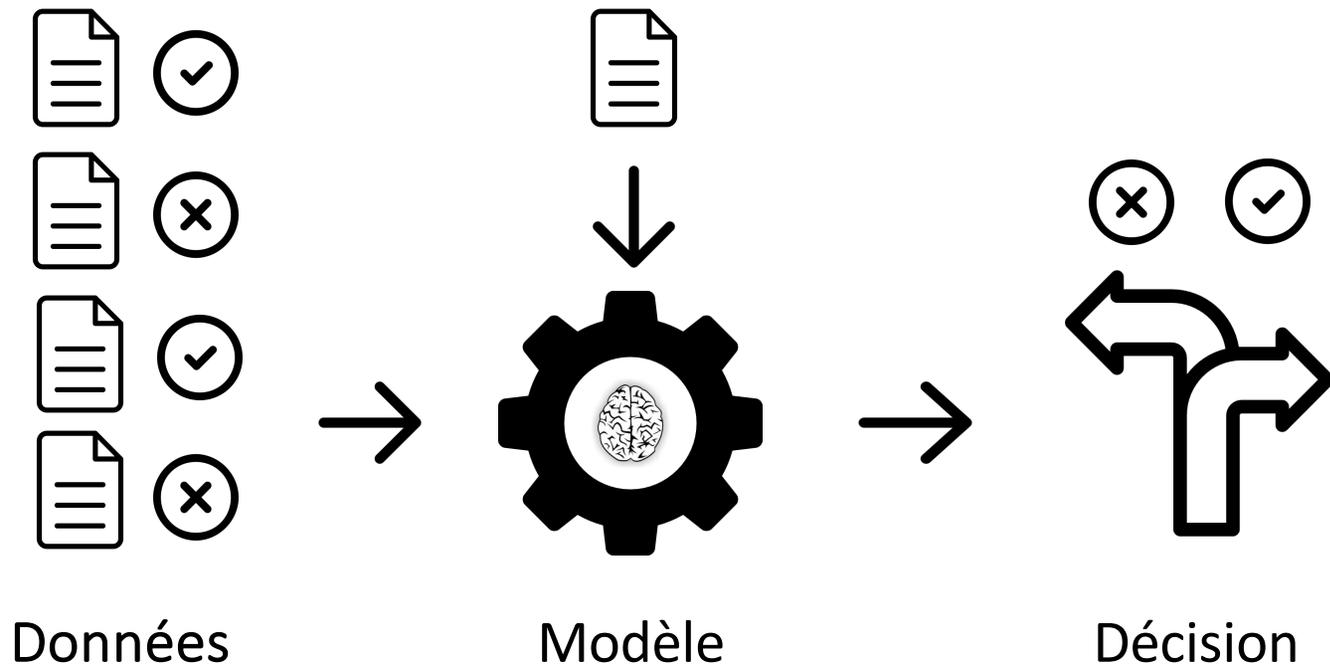


Modèle



Décision

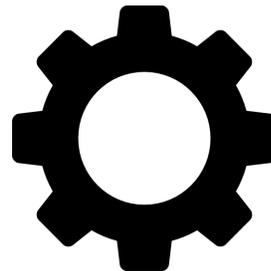
# Aperçu d'un système de ML « typique »



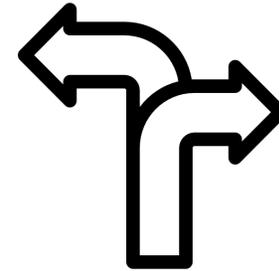
# Aperçu d'un système de ML « typique »



Données



Modèle



Décision

# Le rôle des données

... la pierre angulaire de tout système AI / ML

Données CRM

Dossiers des élèves

Registres des ventes

**Habituellement numérique**

ID	Name	Phone
1	Alice	555-000-0000
2	Bob	666-000-0000

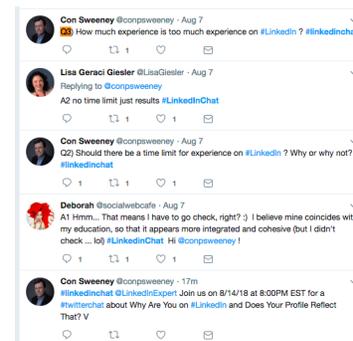
**Données structurées**

Médias sociaux

L'audio

Les articles

**Texte sous forme libre**



**Données non structurées**

# Données structurées vs non structurées

## Données structurées

### Avantages

- Typiquement quantitatif
- Traité à la machine
- Facile à analyser

### Désavantages

- Fournit des informations limitées

## Données non structurées

### Avantages

- Typiquement qualitatif
- Généré par l'homme
- Fournit des informations significatives

### Désavantages

- Très, très difficile à analyser
- Non structuré -> structuré

# Une mise en garde sur les données

- ... vos données peuvent **biais**er considérablement votre système d'IA

Alex Micu  
@axelk

Follow

PDF becomes 4th most popular religion

Google

how do I convert to

how do i convert to **judaism**

how do i convert to **islam**

how do i convert to **catholicism**

how do i convert to **pdf**

Press Enter to search.

9:15 AM - 15 Jul 2015

60,170 Retweets 47,565 Likes

678 60K 48K

# Facteurs importants à considérer à propos des données (partie 1)

## Collecte de données :

- **D'où** obtiendrez-vous les données?
- Les données collectées sont-elles **fiables**?
- **Représentez-vous** correctement le groupe observé?

# Facteurs importants à considérer à propos des données (partie 2)

## Nettoyage/traitement des données :

- Y a-t-il des valeurs **aberrantes** dans les données?
- Comment gérer les **valeurs manquantes**?
- Devez-vous mieux **structurer** certaines données?
- Avez-vous besoin de **convertir** ou de **regrouper** des données ?

# Facteurs importants à considérer à propos des données (partie 3)

## Étiquetage des données :

- Comment les données sont-elles étiquetées ?
- Les étiquettes sont-elles correctes?

# Facteurs importants à considérer à propos des données (partie 4)

- **Règle 80/20:** 80% d'efforts sont consacrés à la collecte et à la préparation de données, 20% à l'apprentissage automatique
- **Données vs Analyse:**
  - La plupart des données sous leur forme brute ne sont pas utiles.
  - Les données deviennent intéressantes lorsque vous les utilisez pour créer des analyses.

# Pratique: Rapport du Credit

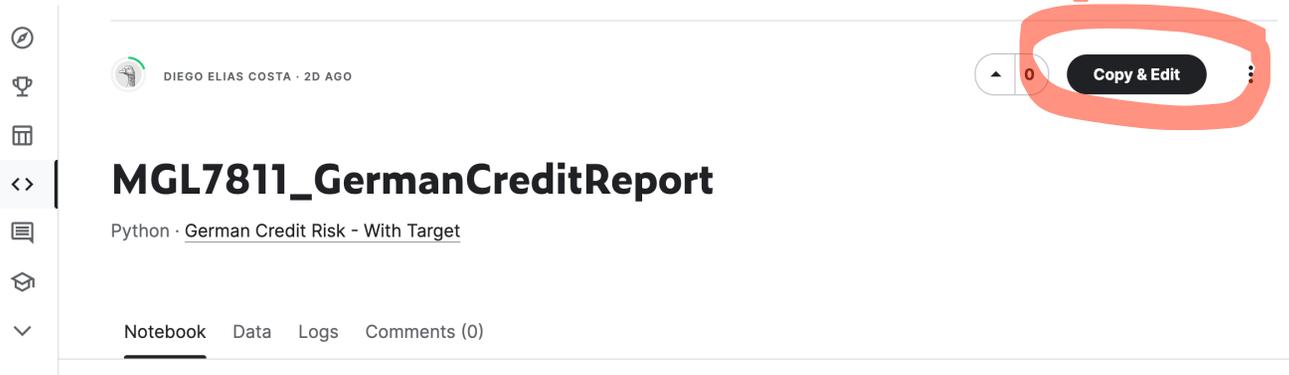
## Scénario

- Les utilisateurs de la banque demandent un crédit pour un achat
- La banque a beaucoup d'informations sur chaque client
- Les analystes utilisent les informations du client pour classer la demande dans:
  - Good (faible risque de défaut de paiement)
  - Bad (risque élevé de défaut de paiement)
- Est-ce que cela peut être automatisé (IA) ?



# Ouverture du notebook

1. Accéder au notebook dans Kaggle (disponible sur Moodle)
  - <https://www.kaggle.com/diegoeliascosta/mgl7811-germancreditreport>
2. Cliquez sur Copy & Edit



# Quelle est la qualité de notre jeu de données?



Explorez les caractéristiques du jeu de données pour répondre aux questions suivantes:

- De combien de données disposez-vous?
- Avez-vous des données manquantes (valeurs Nan)?
- Quelle est la distribution de la variable cible (label)?
- Quels sont les types d'entités dans le jeu de données ?

# Quelle est la qualité de notre jeu de données?



Explorez les caractéristiques du jeu de données pour répondre aux questions suivantes:

- De combien de données disposons-nous?
  - 1000 enregistrements + 9 attributs + 1 variable cible (Risk)
- Avons-nous des données manquantes (valeurs Nan)?
  - Oui, Savings Account + Checking Account
- Quelle est la distribution de la variable cible (etiquete)?
  - Déséquilibré ~70% good credit / 30% bad credit
- Quels sont les types d'entités dans le jeu de données
  - 4 variables numériques + 5 variables catégorielles

# Analyse de la distribution et de la relation des attributs



Explorez la distribution des attributs:

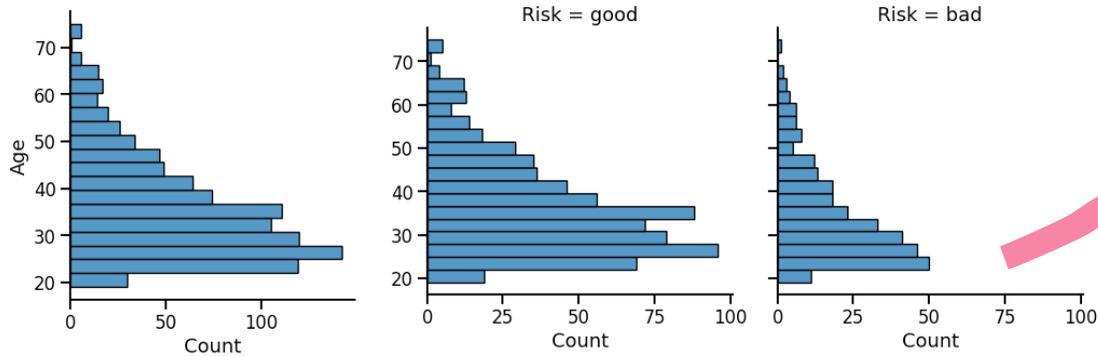
- Avez-vous un ensemble de données biaisé?
- Comment certains attributs se rapportent-ils au bon/mauvais crédit ?

Exemples d'analyses:

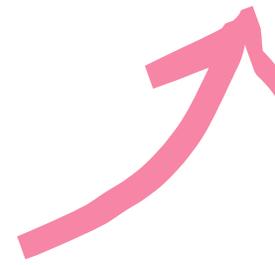
- Age + Sex **vs** Risk
- Age + Checking Account **vs** Risk
- Age + Saving Account **vs** Risk
- Age + Jobs **vs** Risk

# Analyse de la distribution et de la relation des attributs

- Exemple de analyses (Âge)

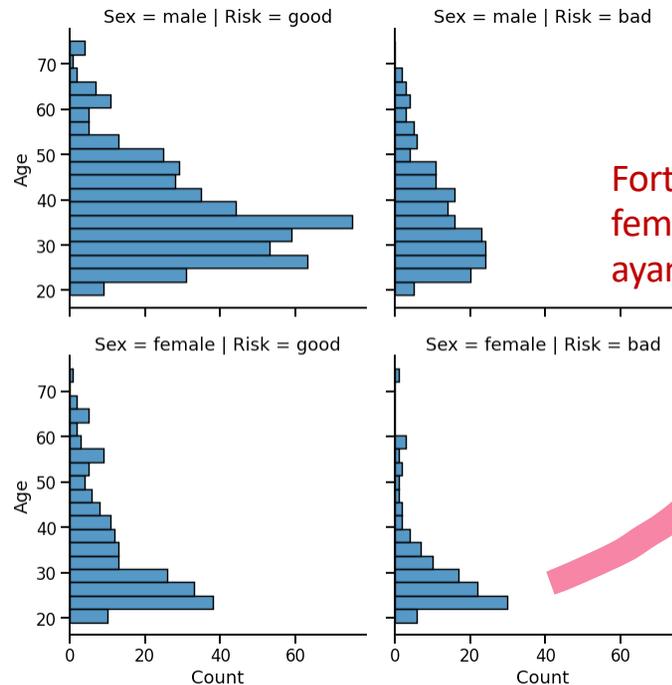
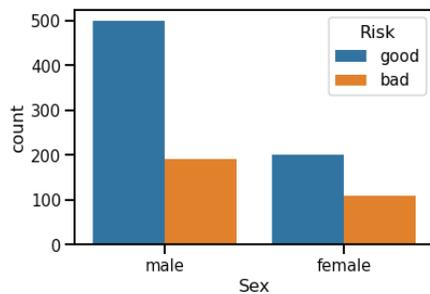
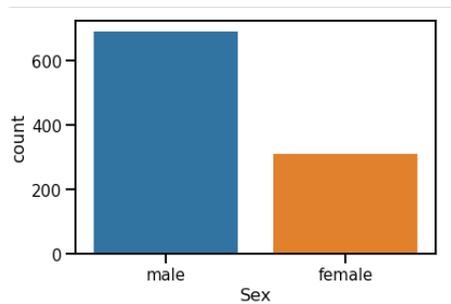


Les dossiers classés comme mauvais sont concentrés sur les jeunes (<30)



# Analyse de la distribution et de la relation des attributs

- Exemple de analyses (Sex)

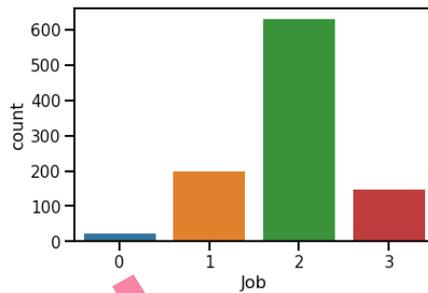


Forte proportion de jeunes femmes classées comme ayant un mauvais crédit

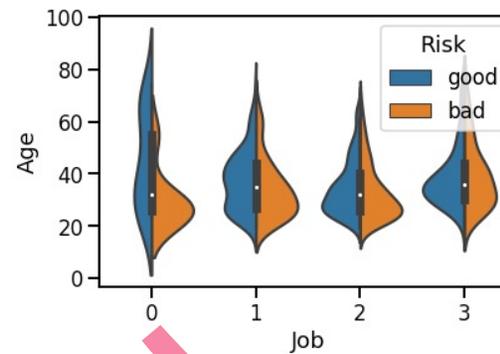


# Analyse de la distribution et de la relation des attributs

- Exemple d'analyses (Job)



Très peu de données sur les clients sans emploi

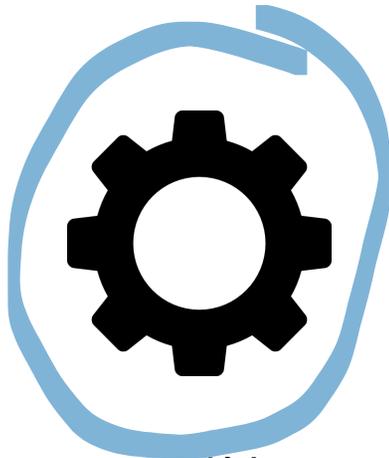


Souvent classé comme mauvais crédit

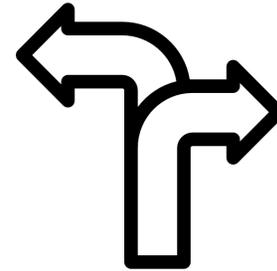
# Aperçu d'un système de ML « typique »



Données



Modèle



Décision

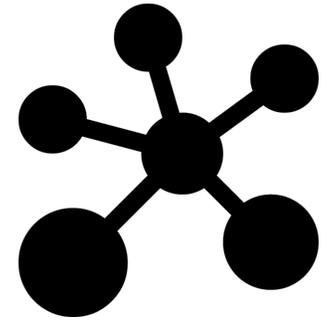
# Main Categories of ML Models

**Modèles d'apprentissage supervisé:** Le modèle s'entraîne sur un jeu de données d'entraînement **étiquetées**. Les prédictions se produisent sur des données inédites.

**Modèles d'apprentissage non supervisé:** Les données **ne sont pas étiquetées**. Le modèle regroupe des points de données similaires.

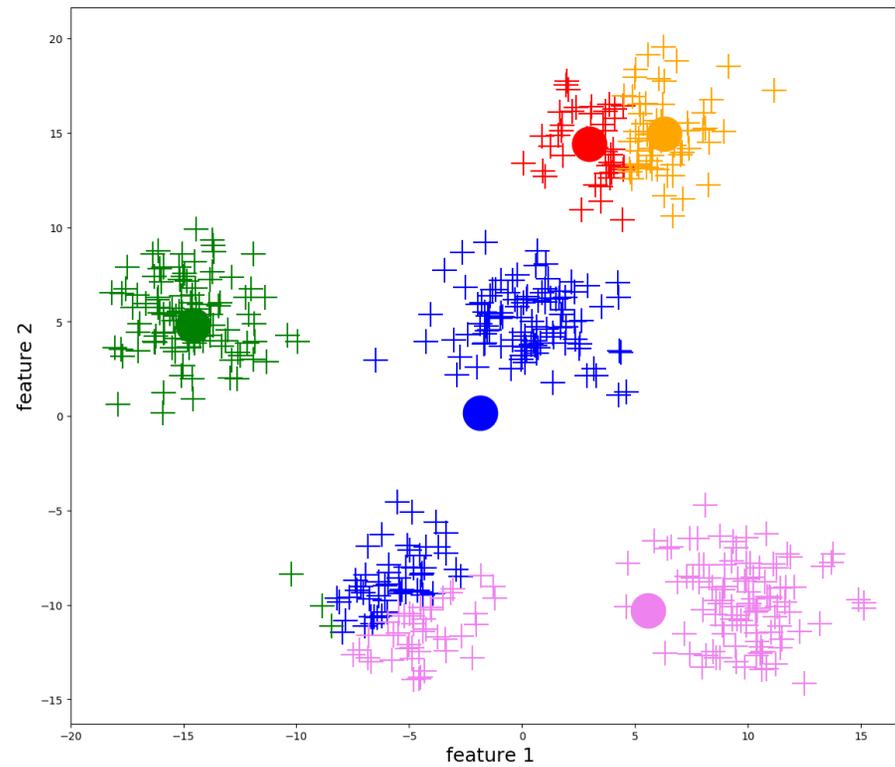
# **Exemple de modèles d'apprentissage automatique**

# K-means/K-moyennes (Non-supervisée)



- **Idée** : regrouper les données non étiquetées en K clusters
- **Comment?**
  - User provides as input K, the number of clusters
  - Centroids are picked and distance is measured between each data point
  - Iterate until distance is minimized and K clearly defined clusters emerge

# K-means/K-moyennes



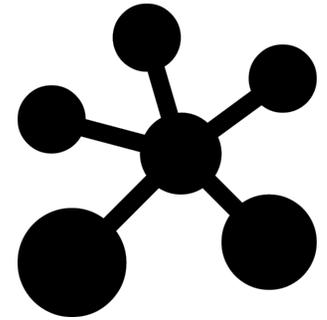
# K-means/K-moyennes

- **Avantages**

- Pas besoin de données étiquetées
- Algorithme simple

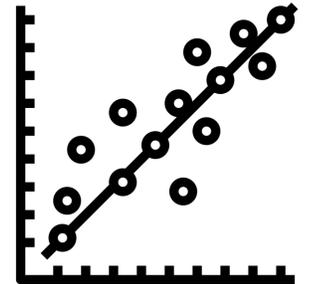
- **Disavantages**

- K doit être déterminé a priori
- Les clusters devront toujours être étiquetés par la suite

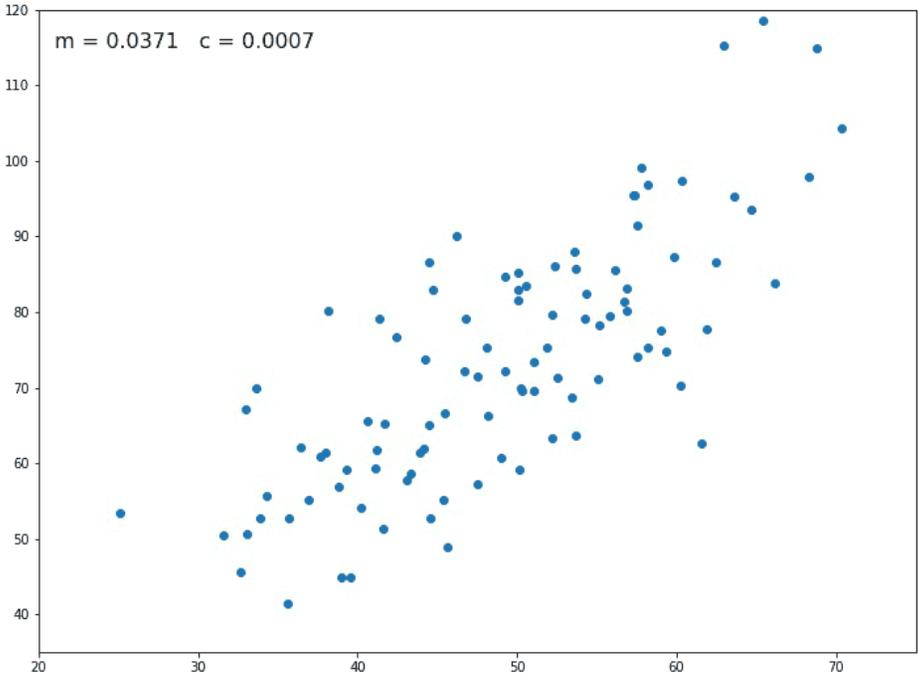


# Régression Lineaire (Supervisée)

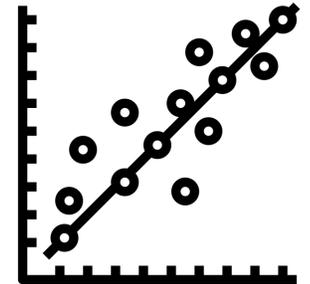
- **Idée** : Utiliser un modèle statistique pour représenter la relation entre 2 variables (ou plus)
- **Comment?**
  - Utilisez une partie des données et ajustez une ligne
  - Choisissez la ligne pour minimiser l'erreur
  - Le résultat est une valeur, p. ex., la taille, le prix, etc.



# Régression Lineaire



# Régression Lineaire



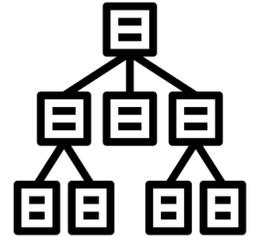
- **Avantages**

- Modèle simple et explicable
- Très populaire, même aujourd'hui

- **Disavantages**

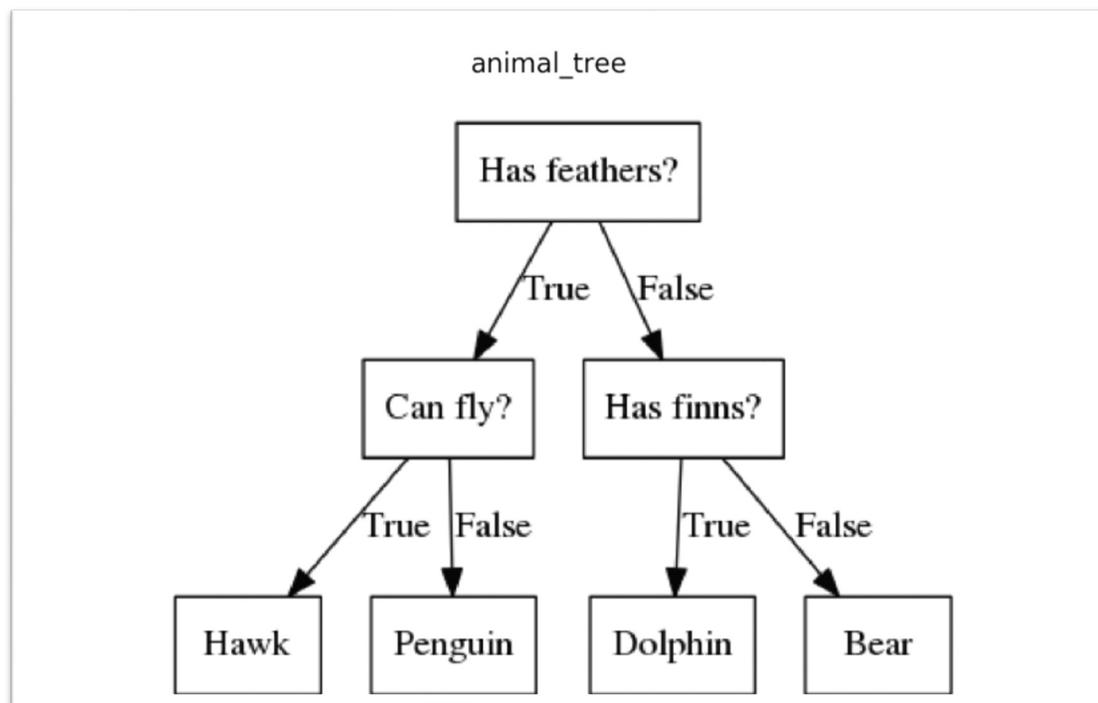
- Suppose une **relation linéaire** entre les variables explicatives et de réponse
- Nécessité d'examiner attentivement la distribution et l'indépendance des données d'entrée

# Arbres de décision (Supervisé)



- **Idée** : utiliser une structure arborescense d'organigramme pour représenter la relation entre les entités et les résultats
- **Comment?**
  - Sélectionnez le meilleur attribut pour diviser les données en sous-ensembles
  - Répète récursivement pour chaque division
  - Nodes -> attributs
  - Branches -> règles de décision
  - Leafs -> les résultats

# Arbres de décision



<https://towardsai.net/p/programming/decision-trees-explained-with-a-practical-example>

# Arbres de décision

- **Avantage**

- Les prédictions sont faciles à expliquer
- Aucune hypothèse sur la distribution des données
- Peut capturer des modèles non linéaires

- **Disavantage**

- Biaisé avec des ensembles de données déséquilibrés
- Moins précis que les autres algorithmes

# Différents modèles pour différents problèmes

- Regroupement de données non étiquetées
  - Non-supervisée (K-means)
- Prédiction de la valeur suivante (continu)
  - Problème de régression (Régression Linéaire)
- Prédiction de la meilleure classe/décision
  - Problème de classification (Arbre de Décision)

# Facteurs importants à considérer

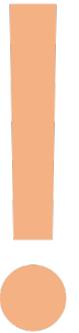
- **Étiquetage des données** : disposez-vous de données étiquetées de bonne qualité
  - Modèles supervisés vs non supervisés
- **Hypothèses du modèle**: y a-t-il des hypothèses précises sur les données ou le modèle?
- **Performance**: Le modèle fonctionne-t-il bien pour le problème en question?
  - Overfitting vs Underfitting
- **Explicabilité**: les décisions sont explicables?

# Préparation des données

Les attributs viennent avec différents formats:

1. Comment gérer les valeurs manquantes?
2. Comment encoder des attributs catégorielles?
3. Comment extraire les informations les plus pertinentes à partir de données brutes?

Nous allons parcourir ce processus ensemble.



# Quels modèles sont les plus performants?

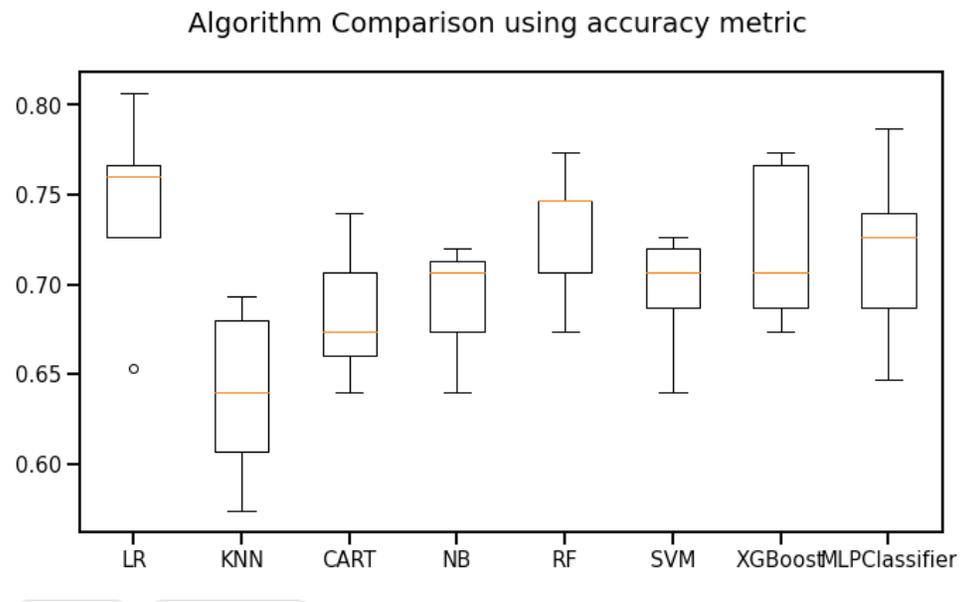


Explorez comment certains modèles performant:

1. Choisissez un modèle
2. Exécutez l'entraînement et signalez la performance
3. Lisez leur documentation respective et essayez d'affiner certains de ses paramètres

# Quels modèles sont les plus performants?

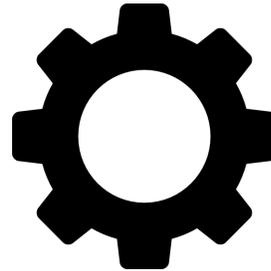
Accuracy + paramètres par défaut



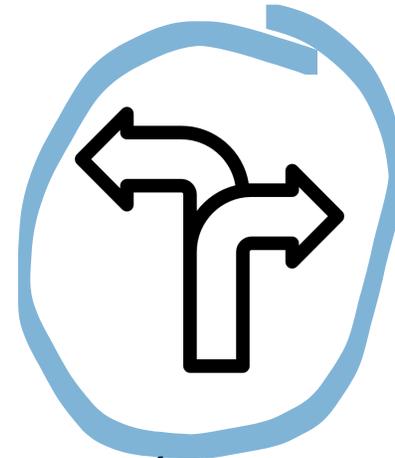
# Aperçu d'un système de ML « typique »



Données



Modèle



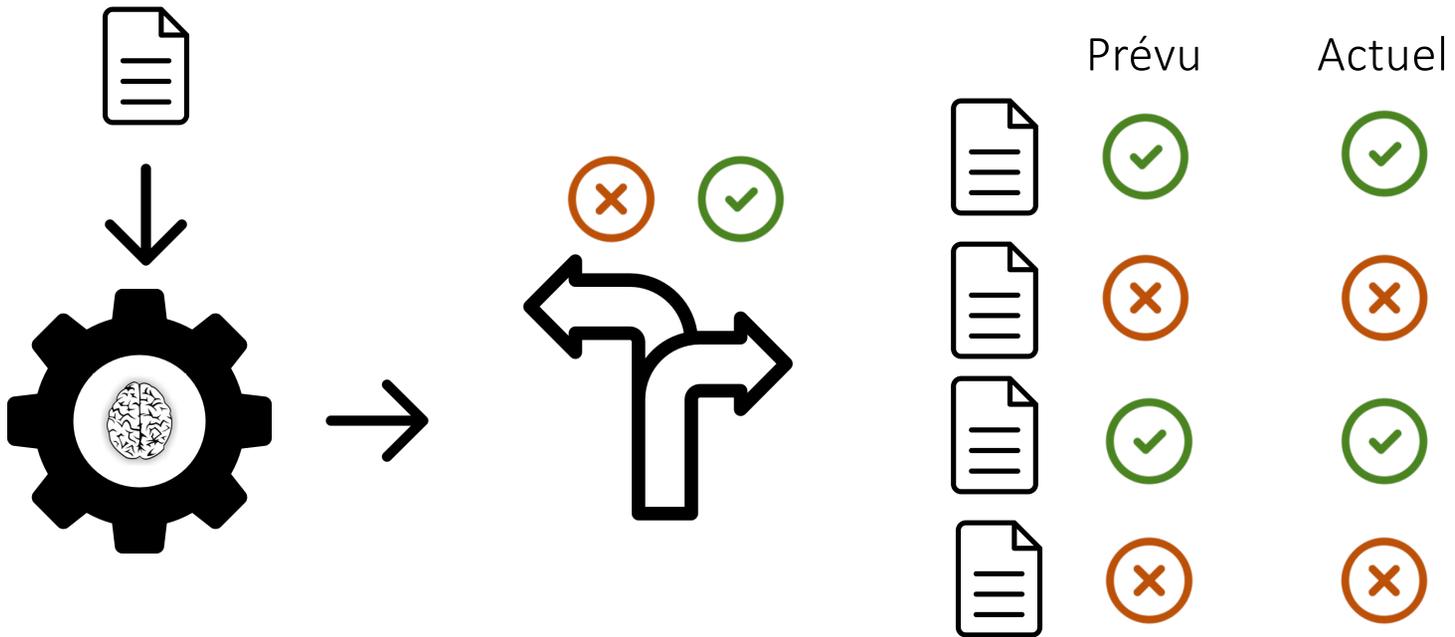
Décision

# Quelle est la performance **réelle** de notre modèle ?

Nous n'avons exploré que la performance sur les données de formation:

1. Choisissez le meilleur modèle que vous avez évalué
2. Évaluer les performances de l'ensemble de tests
3. Comparez les performances avec certaines lignes de base (??)

# L'Évaluation de la performance du modèle



# L'Évaluation de la performance - Accuracy

Actuel	Prevu		Actuel	Prevu	
		TP			TN
		FP			TN
		TP			TN
		TN			TN
		FN			TN
		TN			FP

# L'Évaluation de la performance - Accuracy

Actuel    Prevu

✓	✓
✗	✓
✓	✓
✗	✗
✓	✗
✗	✗

$$\text{Accuracy: } (TP+TN)/(TP+FP+TN+FN) \\ = 4/6 = 66.67\%$$

Actuel    Prevu

✗	✗
✗	✗
✗	✗
✗	✗
✗	✗
✗	✗
✓	✗

$$\text{Accuracy: } (TP+TN)/(TP+FP+TN+FN) \\ = 5/6 = 83.34\%$$

Pourquoi le modèle ci-dessous montre-t-il une meilleure performance?

# L'Évaluation de la performance - Accuracy

Actuel    Prevu

✓	✓
✗	✓
✓	✓
✗	✗
✓	✗
✗	✗

**Accuracy:**  $(TP+TN)/(TP+FP+TN+FN)$   
 $= 4/6 = 66.67\%$

**Precision:**  $TP/(TP+FP) = 2/3 = 66.67\%$

**Recall:**  $TP/(TP+FN) = 2/3 = 66.67\%$

Actuel    Prevu

✗	✗
✗	✗
✗	✗
✗	✗
✗	✗
✗	✗
✓	✗

**Accuracy:**  $5/6 = 83.34\%$

**Precision:**  $TP/(TP+FP) = 0\%$

**Recall:**  $TP/(TP+FN) = 0\%$

# Quand utiliser différentes mesures?

- Accuracy
  - Très instructif dans les ensembles de données équilibrés.
- Precision
  - La précision de la décision est la priorité
- Recall (rappel)
  - Trouver tous les cas positifs est la priorité
- F1 score
  - Moyenne harmonique entre precision et recall
  - Valeurs égales precision et recall

$$F_1 = 2 \cdot \frac{(\text{précision} \cdot \text{rappel})}{(\text{précision} + \text{rappel})} \in [0, 1].$$

Avec, pour rappel :

- $\text{précision} = VP / (VP + FP)$
- $\text{rappel} = VP / (VP + FN)$

<https://fr.wikipedia.org/wiki/F-mesure>

# Quelle est la performance **réelle** de notre modèle ?

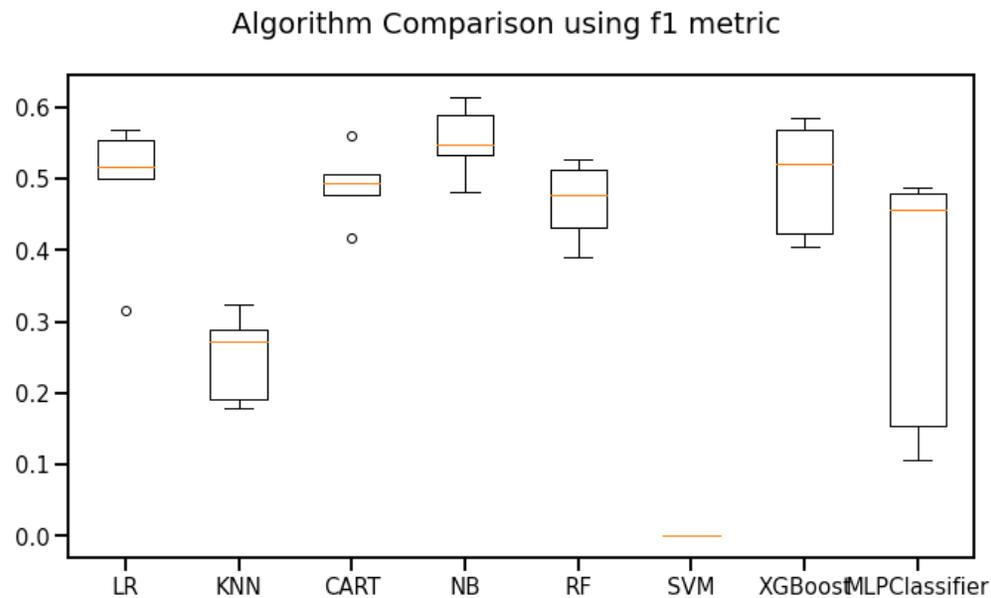
Nous n'avons exploré que la performance sur les données de formation:

1. Choisissez le meilleur modèle que vous avez évalué
2. **Choisissez une mesure de performance appropriée**
3. Évaluer les performances de l'ensemble de tests
4. Comparez les performances avec certaines lignes de base (??)

# Quels modèles sont les plus performants?



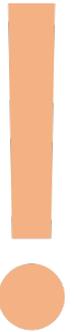
F1 + paramètres par défaut



# Comprendre le modèle

Vous devez toujours inspecter (et apprendre) avec le modèle:

- Les attributs plus pertinents
- La courbe probabiliste par chaque attribut
- Expliquer certaines prédictions



# Message à retenir

- La construction de systèmes d'IA nécessite un examen attentif
- Les données sont plus importantes que les algorithmes
- Choisissez les bons algorithmes, car la plupart ont de nombreuses hypothèses complexes
- Valider à l'externe et rechercher les biais potentiels