Améliorer la performance des modèles avec des **Méthodes d'ensemble**

Atelier #13

Combiner des modèles en *"ensemble"*

Quand combiner des modèles?

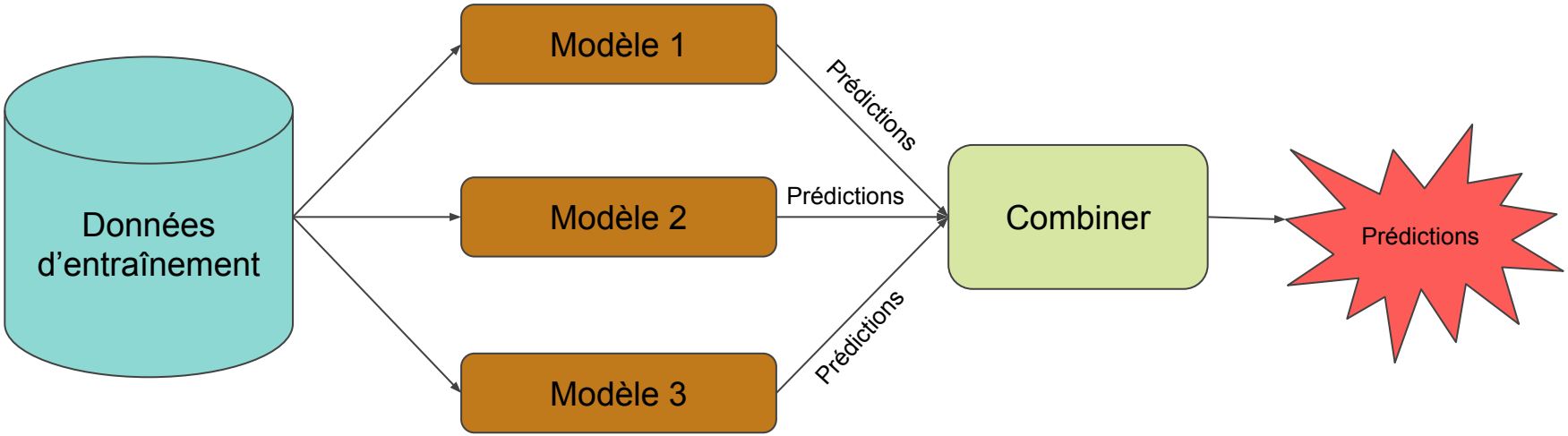
- La combinaison de plusieurs modèles (*"model ensembling"*) est pertinent lorsque plusieurs d'entres eux ont des performances (e.g. "accuracy") similaires et qu'il est difficile d'en sélectionner qu'un seul au final.
- L'utilisation de méthodes d'ensemble permet d'éviter le sur-ajustement d'un modèle unique.

Différentes méthodes utilisées :

On va voir 3 principales méthodes de combinaison de modèles :

1. Bagging
2. Boosting
3. Votes / *Voting*

Qu'est ce que l'apprentissage par ensemble?



Pourquoi combiner des modèles?

Avantages

- Meilleure précision (*accuracy*) et réduit l'erreur associé aux modèles.
- Diminue les chances de sur-ajustement des modèles.
- Peut soit réduire le biais ou la variance des modèles.

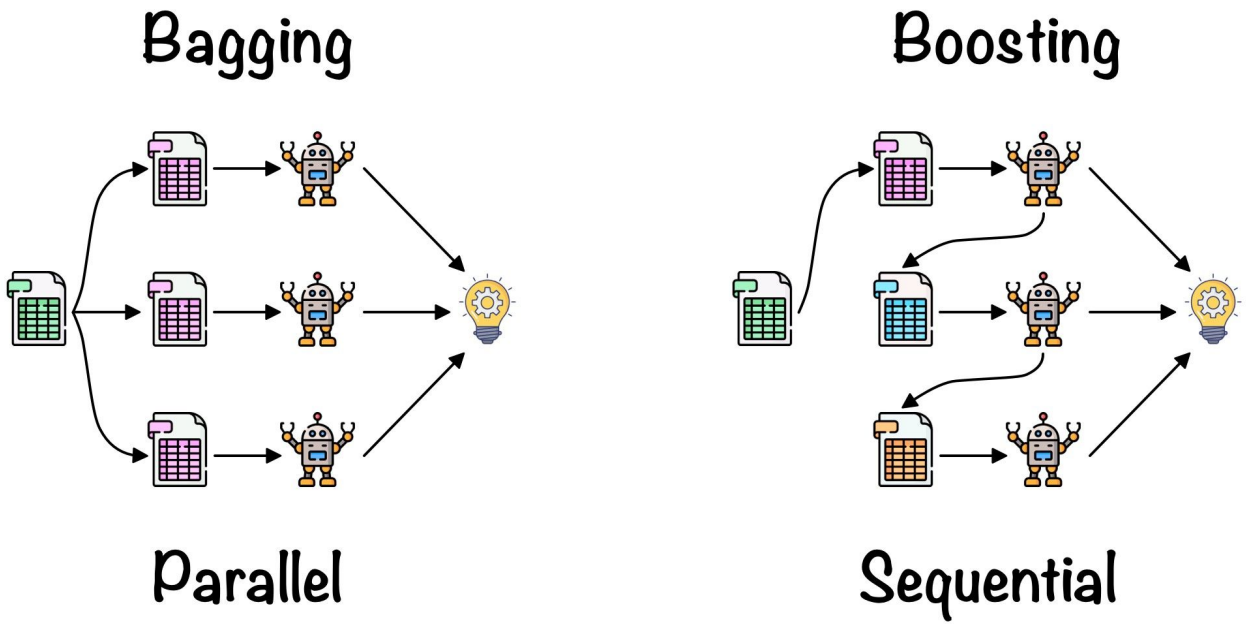
Désavantages

- Les modèles perdent en interprétation directe.
- Demandent plus de pouvoir computationnel pour être effectué.

3 méthodes d’ensembles

1. **Bagging** : Construire plusieurs modèles de même type à partir de différent sous-ensemble du jeu de données.
2. **Boosting** : Construire plusieurs modèles de même type qui vont apprendre à corriger les erreurs des modèles précédemment roulés.
3. **Voting** : Construire plusieurs modèles de types différents qui vont être combinés à l'aide d'une statistique simple pour maximiser la performance.

Modèle simple vs Bagging vs Boosting



https://towardsdatascience.com/ensemble-learning-bagging-boosting-3098079e5422

Bagging

Fonctionnement

- Bagging = **B**ootstrap **Agg**regat**ing** (donc dur à traduire directement au français)
- Construire plusieurs modèles de même type à partir de différent sous-ensemble du jeu de données.
- Faire une “moyenne” de la prédiction de chaque modèle construit.
- Combine des modèles de même type.
- Un modèle en exemple = Forêt aléatoire / “*Random forest*”

Avantage : Réduire la variance et permet de réduire les erreurs de prédictions

Désavantage : Prends plus de temps pour être exécuter.

Boosting

Fonctionnement

- Le boosting se concentre à sur les observations les plus difficiles à prédire.
- Utilisé pour la classification et la régression.
- Combine des modèles de même type.

Avantage : Réduit le biais.

Désavantage : Augmente la variance.

Lien : https://scikit-learn.org/stable/modules/classes.html#module-sklearn.ensemble

Système de votes / *“Voting”*

Fonctionnement

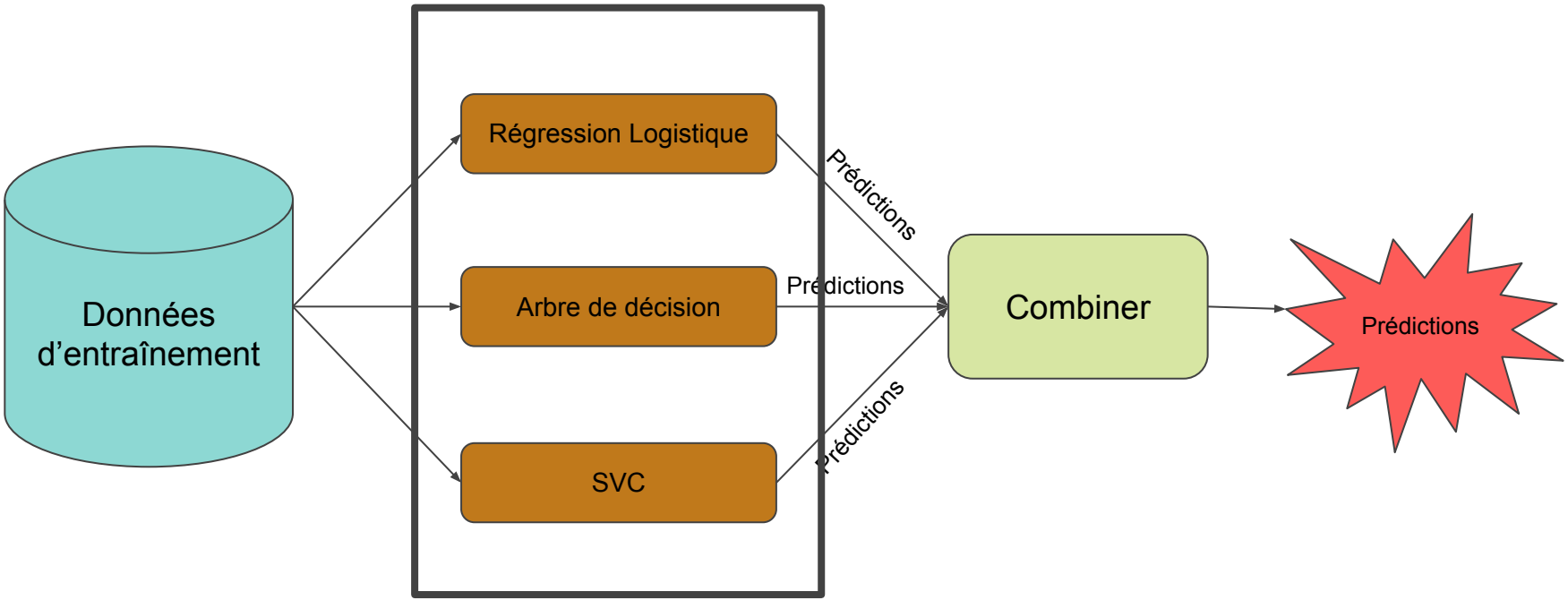
- Cette méthode permet de regrouper des modèles qui sont conceptuellement différents dans un système de vote pour faire une “moyenne” qui compense les faiblesses de chacuns.

Avantage : Permet d’obtenir la meilleure précision (“accuracy”) ou tout autre métrique.

Désavantage : Le modèle final perd en interprétabilité.

Lien : https://scikit-learn.org/stable/modules/ensemble.html#voting-classifier

Diagramme pour la méthode du vote



Choisir une méthode d’ensemble à priori n’est pas évident, mais considérer des modèles utilisant certaines de ces méthodes est pratique courante.