Conclusion

Pour conclure, il n'existe aucun algorithme universellement meilleur permettant de résoudre tous les problèmes de classification ou de prédiction. On appelle ce phénomène le *no free lunch theorem*. Chaque jeu de données, chaque contexte d'analyse, chaque objectif de modélisation est unique. Il n'y a donc pas de solution pré-établie : c'est donc au statisticien ou au data scientist de faire preuve de discernement, d'esprit critique, et de créativité pour adapter les méthodes aux données et à la question posée.

Cela dit, on peut dégager certains principes qui peuvent nous guider dans une démarche rigoureuse :

1. Commencer par une exploration descriptive des données, en étudiant les distributions, les relations entre variables, les valeurs manquantes, etc.
2. Tirer parti de l'expertise métier, en consultant les personnes qui connaissent le contexte des données (experts du domaine, utilisateurs finaux, etc.).
3. S'inspirer des méthodes ayant déjà donné de bons résultats dans des situations similaires, en gardant toutefois un regard critique.

Une des principales difficultés en pratique est souvent liée à la dimensionnalité du problème : le nombre de modèles ou d'approches possibles augmente rapidement avec le nombre de variables disponibles. Il devient alors crucial de simplifier l'espace de recherche, e.g. réduire la dimension avec une ACP, qui permettent de résumer l'information tout en limitant le bruit. Une approche fréquente consiste à appliquer des méthodes d'analyse à un sous-ensemble de variables soigneusement choisies, ou à utiliser les composantes principales comme nouvelles variables explicatives. Mais là encore, aucune règle générale ne permet d'identifier automatiquement le "bon" sous-ensemble…

Enfin, on peut noter que plusieurs questions centrales en analyse de données ont été laissées de côté dans ce cours. Pourtant, celles-ci sont omniprésentes en pratique. Parmi ces questions :

— Comment définir les variables à utiliser comme prédicteurs (*feature engineering*) ?
— Comment repérer et gérer les données aberrantes (*outliers) ?
— Comment scinder les données entre apprentissage, validation et test ?
— Que faire en présence de données manquantes ?

— Les données analysées sont-elles représentatives de la population cible ?

Ces questions soulignent que l'analyse de données est autant un art qu'une science. Il ne suffit pas d'appliquer mécaniquement des algorithmes : il faut formuler des hypothèses claires, évaluer leurs limites, et rester attentif aux enjeux éthiques, sociaux et pratiques.