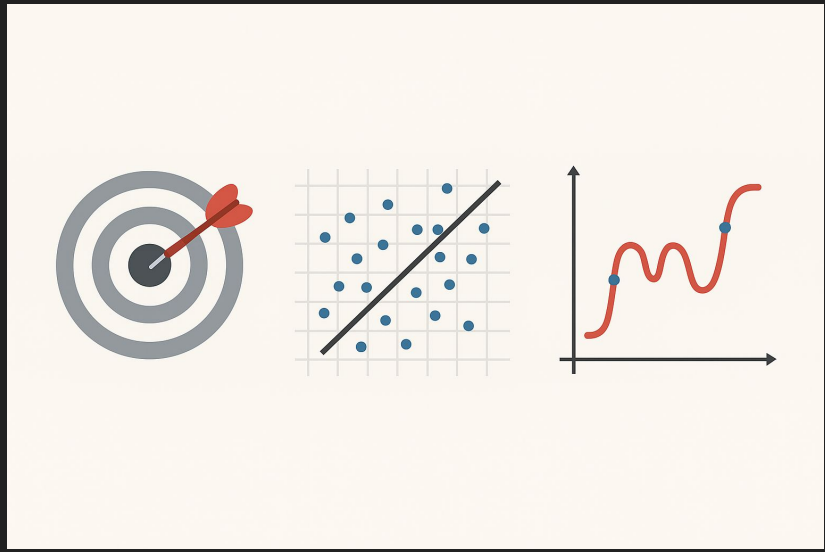
Biais, Bruit, Overfitting et Underfitting en Machine Learning



Comprendre leurs impacts sur la performance des modèles

Biais

Erreur systématique due à des suppositions simplificatrices du modèle

Bruit

Variabilité aléatoire dans les données rendant plus difficile l'apprentissage du modèle

Surapprentissage / Sous-apprentissage

Modèle trop complexe ou trop simple induisant un défaut d’apprentissage.

Biais (Bias)

- Orientation ou direction qui s'écarte de la ligne droite ou verticale.
- Manière particulière d'aborder, d'envisager quelque chose sous un certain angle.
- Erreur systématique affectant le résultat d'une mesure ou évaluation et menant à une estimation incorrecte.

Impact : Un biais élevé entraîne un sous-apprentissage.

Biais (Bias)

- Le **biais** se réfère à l'erreur systématique introduite par un modèle lorsqu'il fait des hypothèses trop simplifiées sur la réalité (par exemple, une approximation linéaire pour des données qui suivent une relation non linéaire).

- Un **modèle à fort biais** sous-ajuste les données, car il ne capture pas toute la complexité du problème. Cela peut se produire lorsqu'un modèle est trop simple ou trop restrictif (par exemple, un modèle linéaire pour des données non linéaires).

- **Exemple** : Un modèle de régression linéaire qui essaie de prédire des données qui suivent une courbe quadratique aura un biais élevé.

Algorithme : un risque de biais à chaque étape

Biais techniques

Variable omise
Les données non mesurables ou manquantes ne seront pas traitées.

Bases de donnée
L'algorithme apprend à partir de données de mauvaise qualité.

Sélection
Les données sélectionnées dans la base de données ne sont pas représentatives de la réalité.

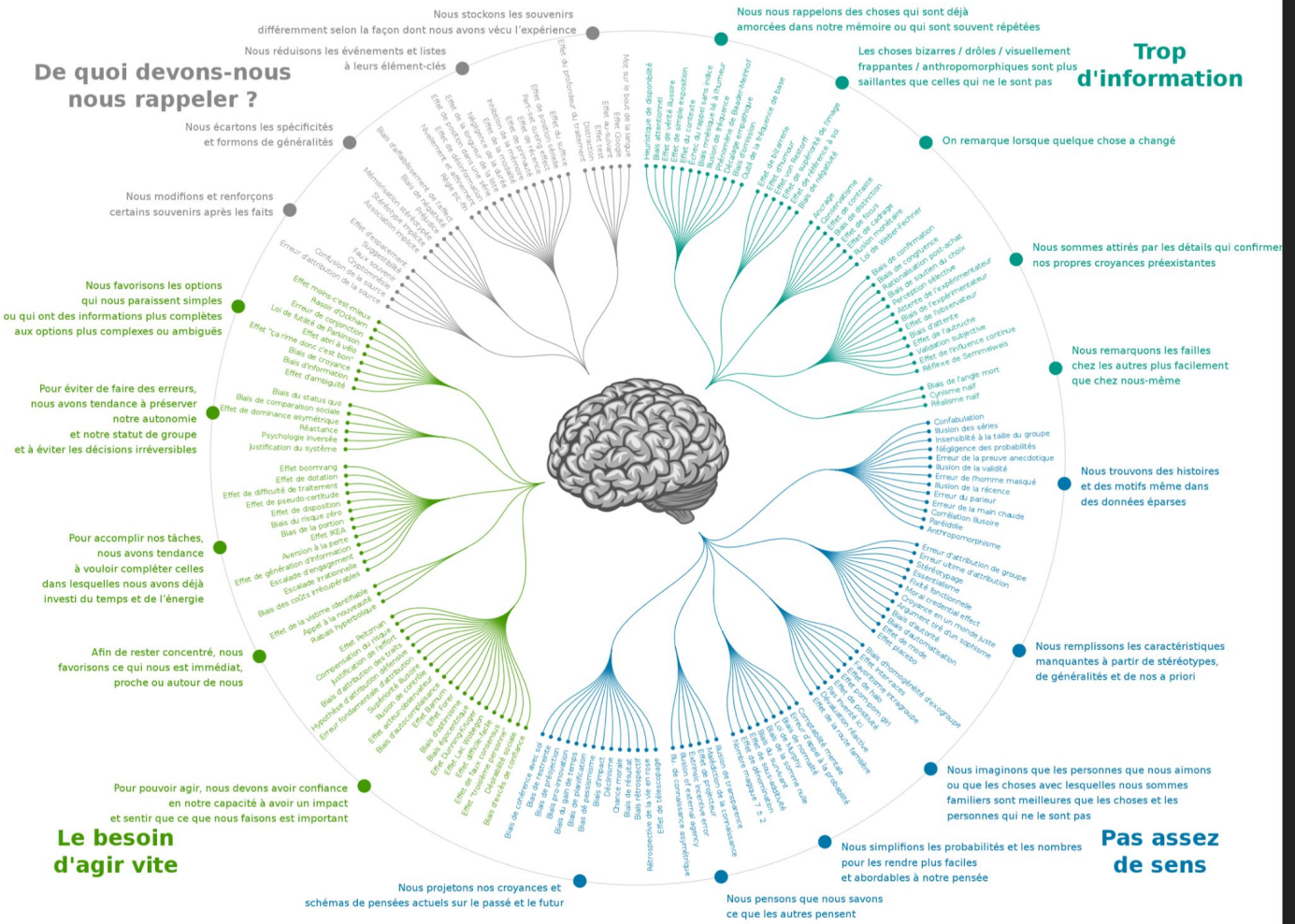
Biais de société

Économiques
Nos habitudes de consommation ou les besoins économiques de l'entreprise orientent l'algorithme.

Cognitifs
Nos visions du monde influencent nos raisonnements et la manière dont nous traitons l'information.

Émotionnels
Nos émotions distordent nos jugements et nos décisions.

LE CODEX DES BIAIS COGNITIFS



Bruit (Noise)

- Variations aléatoire dans les données qui ne reflètent pas un vrai signal.
- Erreurs de mesure, données incohérentes, fluctuations imprévisibles...
- Informations introduisant des erreurs, des distorsions ou des inexactitudes dans un ensemble de données.

Impact : Un bruit élevé entraîne un sous-apprentissage.

Bruit (Noise)

- Le **bruit** est l'erreur qui vient des fluctuations aléatoires dans les données elles-mêmes. Il est indépendant du modèle, et même un modèle parfait ne pourrait pas le réduire. C'est une composante aléatoire qui introduit une variabilité dans les observations.

- Le bruit est souvent considéré comme une erreur inhérente aux données, qui ne peut pas être expliquée ou capturée par le modèle.

Équipe #1

Équipe #2

Équipe #3

Équipe #4

Equipe 1 : Pas de biais, peu de bruit.

Equipe 2 : Biais important, peu de bruit.

Equipe 3 : Pas de biais, beaucoup de bruit.

Equipe 4 : Biais important, bruit important.

Variance

- Mesure de la sensibilité du modèle aux variations des données d'entraînement.
- Capacité d’un modèle à fluctuer en fonction des données qu’il apprend.
- Une variance élevée signifie que le modèle s’adapte trop aux particularités des données d'entraînement, y compris le bruit.

Impact : Une variance élevé entraîne un surapprentissage rendant le modèle inefficace sur de nouvelles données.

Variance

- La **variance** mesure la sensibilité d'un modèle aux fluctuations ou variations dans les données d'entraînement. Un modèle avec **haute variance** s'ajuste très bien aux données d'entraînement, mais cela peut aussi entraîner un **surapprentissage** (overfitting), où le modèle devient trop spécifique aux données d'entraînement et échoue à généraliser sur de nouvelles données.

- Un **modèle à faible variance** sera plus robuste aux variations dans les données d'entraînement et aura tendance à mieux généraliser, mais il risque d'ignorer certaines relations dans les données.

Lien entre biais, variance et bruit

- **Biais élevé, variance faible** : Le modèle fait des approximations simplistes (sous-ajustement), mais il est stable. Il fait des erreurs systématiques en raison de son incapacité à capturer toute la complexité des données.

- **Biais faible, variance élevée** : Le modèle s'ajuste de manière trop précise aux données d'entraînement (surapprentissage), ce qui signifie qu'il est sensible aux fluctuations dans ces données et risque de ne pas généraliser sur de nouvelles données.

- **Bruit** : Il n'est pas lié au modèle, mais plutôt à la nature des données elles-mêmes. Il est souvent impossible à réduire, quel que soit le modèle.

Surapprentissage

- Le modèle apprend trop bien les détails du dataset d'entraînement, y compris le bruit.

Impact : Mauvaise généralisation aux nouvelles données.

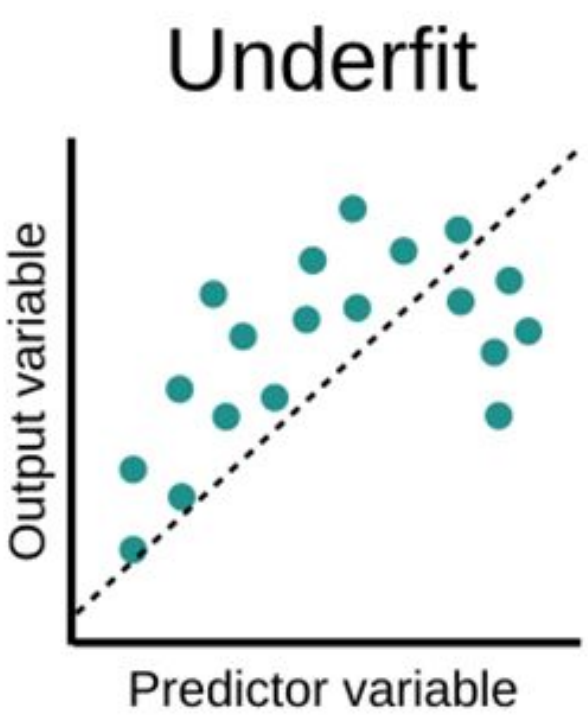
Par exemple le modèle démontre 100% de précision en entraînement, mais échoue sur de nouvelles données.

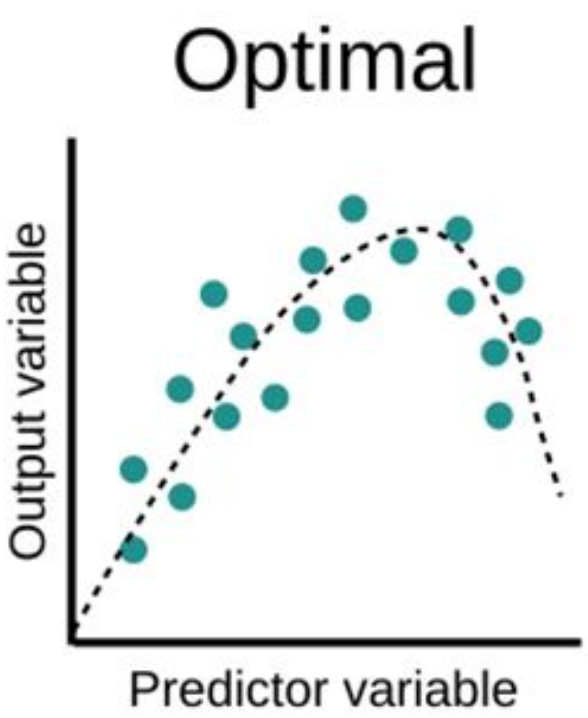
Sous-apprentissage

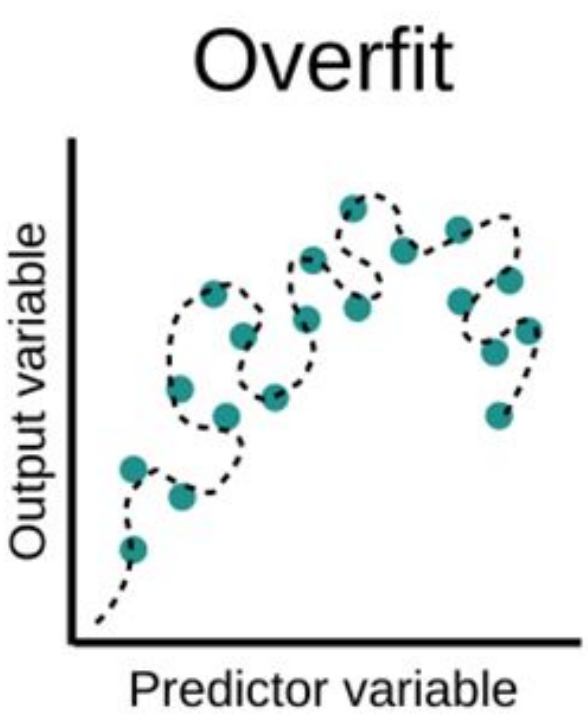
- Le modèle souffre d’un biais trop élevé, ce qui signifie qu’il simplifie excessivement les relations entre les données et ne parvient pas à apprendre les tendances pertinentes.

Impact : Mauvaise performance aussi bien sur les données d'entraînement que sur les nouvelles données.

Par exemple, le modèle obtient une précision faible dès l'entraînement, indiquant qu'il n'a pas suffisamment appris.







- Le modèle, trop simpliste, ne parvient pas à capturer les tendances dans les données.
- Il possède un biais élevé, signifiant qu'il fait de fortes approximations et ignore des relations importantes dans les données.
- Mauvaise performance sur les données d'entraînement que sur les nouvelles.

- Ce modèle capture l'essence des données sans trop s'adapter aux variations spécifiques.
- Il atteint un bon compromis entre biais et variance, lui permettant de bien généraliser sur de nouvelles données.

- Le modèle possède une variance trop élevée, ce qui signifie qu'il est trop sensible aux petites fluctuations des données d'entraînement.
- Il performe très bien sur l'entraînement, mais généralise mal sur de nouvelles données.

Concept	Définition	Impact sur l'apprentissage	Défauts d'apprentissage
Biais	Simplification excessive du modèle qui l'empêche de bien apprendre les relations dans les données.	Un biais élevé entraîne un sous-apprentissage (le modèle ne capture pas la structure des données.)	Sous-apprentissage : le modèle est trop simpliste pour détecter les tendances.
Variance	Sensibilité aux variations des données d'entraînement, capacité du modèle à s'adapter aux fluctuations des données.	Une variance élevée entraîne un surapprentissage (le modèle s'adapte trop aux données d'entraînement, y compris le bruit.)	Suraprentissage : le modèle est trop complexe et ne généralise pas bien.
Bruit	Variation aléatoire dans les données qui ne reflètent pas un vrai signal.	Un bruit élevé peut induire du suraprentissage, car le modèle risque d'apprendre des fluctuations non pertinentes.	Le bruit est souvent capté par un modèle trop complexe, augmentant la variance et le suraprentissage.
Sous-apprentissage (Underfiting)	Le modèle est trop simple pour capturer les tendances des données.	Mauvaise performance sur l'entraînement et la généralisation.	Associé à un biais élevé et un faible variance.
Surapprentissage (Overfiting)	Le modèle est trop complexe et s'adapte trop aux données d'entraînement, y compris le bruit.	Bonne performance sur l'entraînement mais mauvaise généralisation aux nouvelles données.	Associé à une variance élevée et une sensibilité excessive aux détails du dataset.
Modèle optimal	Bon équilibre entre biais et variance, capturant les tendances principales sans s'adapter aux fluctuations aléatoires.	Bonne généralisation aux nouvelles données.	Ni sous-apprentissage, ni surapprentissage : modèle bien régularisé.