Sélection des variables

Atelier #7

Avantages de sélectionner des variables

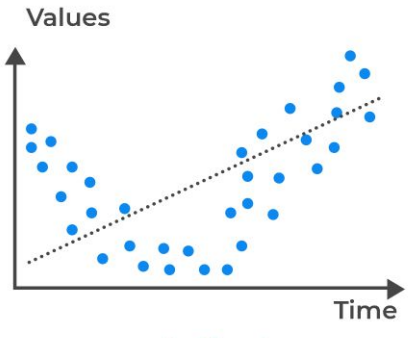
Pourquoi sélectionner les variables à utiliser dans un algorithme?

Pourquoi cette étape est si importante?

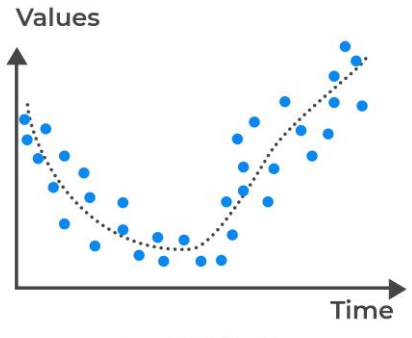
Voici les avantages :

1. Réduit le **SUR-APPRENTISSAGE** ou **SUR-AJUSTEMENT** ou "***l'overfitting***"
2. Améliore la **PRÉCISION** des modèles
3. Réduit le **TEMPS D'ENTRAÎNEMENT**

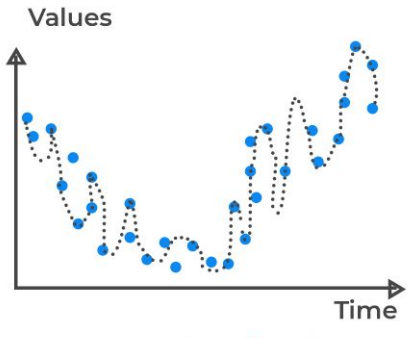
Généralisation vs Sur-ajustement *(“underfitting vs overfitting”)*



Généralisé

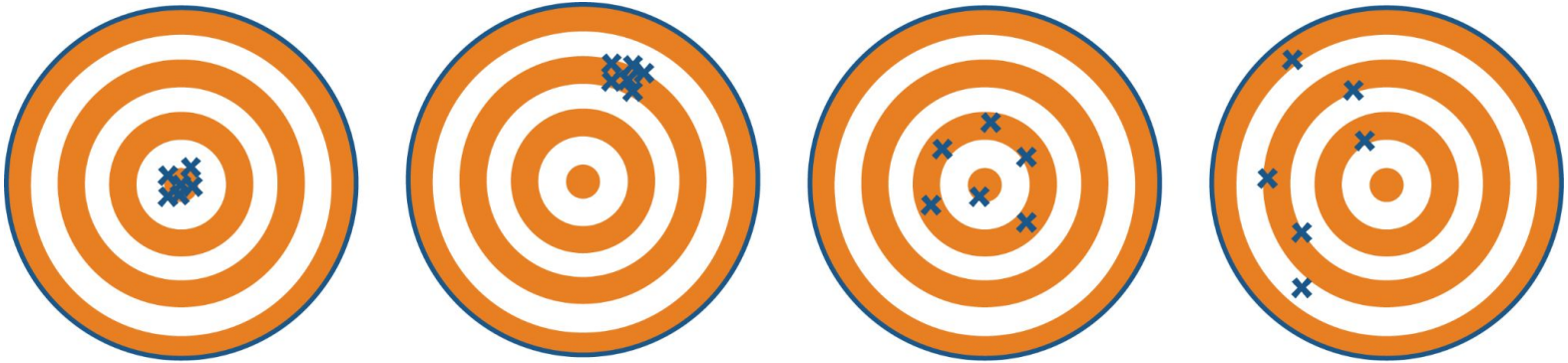


Équilibré



Sur-ajusté

Précision vs Exactitude



Forte Exactitude
Forte Précision

Faible Exactitude
Forte Précision

Forte Exactitude
Faible Précision

Faible Exactitude
Faible Précision

Précision dans un contexte de données

Précision = 40%

Originale	Prédiction
0	1
1	1
0	1
0	1
1	1

Précision = 100%

Originale	Prédiction
0	0
1	1
0	0
0	0
1	1

Temps et ressources d’entraînement

Temps : Il est important d’optimiser le temps nécessaire pour faire rouler un algorithme.

Ressources : Les algorithmes sont roulés sur des serveurs clouds, qui sont coûteux à utiliser, ou locaux, qui affectent la performance de l’ordinateur.

	Temps	Ressources
Petit jeu de données	➖	➖
Grand jeu de données	➕	➕

Méthodes de sélection

On va voir 3 méthodes de sélection :

1. Sélection univariée
2. Élimination récursive des variables
3. Importance des variables

Note : Il existe plusieurs autres méthodes pour sélectionner des variables, mais on se concentrera sur celles-ci dans le cadre de ce cours.

Méthodes de sélection de variables

3 méthodes de sélection automatique de variable avec Scikit-learn :

1. Sélection univariée des variables
2. Élimination récursive des variables (RFE - Recursive feature elimination)
3. Importance des variables