Techniques de validation des modèles :

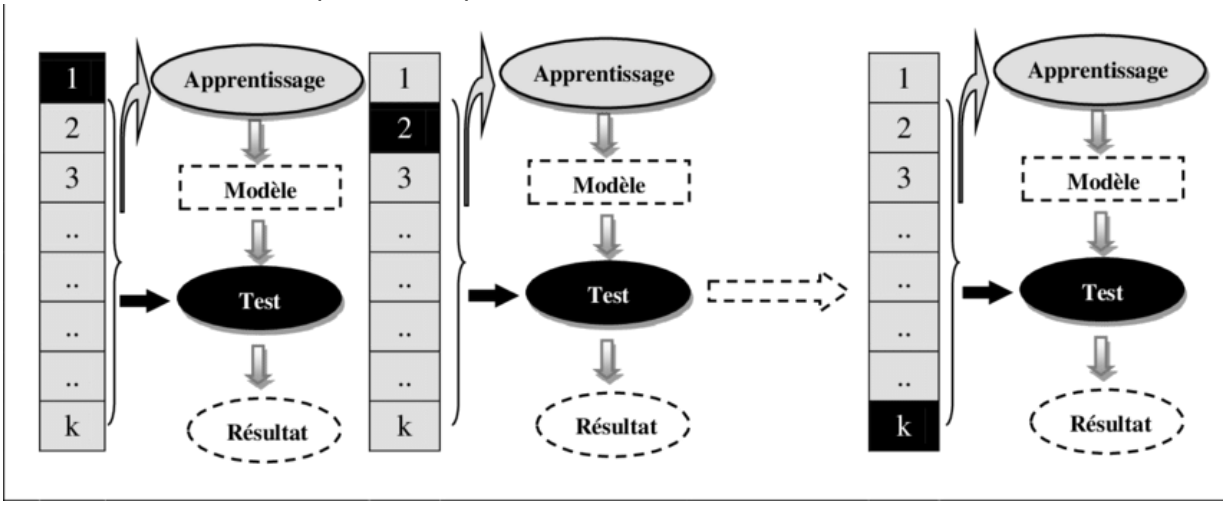
1. **Validation « TestSplit » :** on divise l'échantillon de taille n en deux sous-échantillons, le premier dit d'apprentissage (communément supérieur à 60 % de l'échantillon) et le second dit de validation ou de test. Le modèle est bâti sur l'échantillon d'apprentissage et validé sur l'échantillon de test avec un score de performance de notre choix.

2. **La validation croisée à k blocs, « *k-fold cross-validation* »** : on divise l'échantillon original en k échantillons (ou « blocs »), puis on sélectionne un des k échantillons comme ensemble de validation pendant que les k-1 autres échantillons constituent l'ensemble d'apprentissage. Après apprentissage, on peut calculer une performance de validation. Puis on répète l'opération en sélectionnant un autre échantillon de validation parmi les blocs prédéfinis. À l'issue de la procédure nous obtenons ainsi scores de performances, un par bloc. La moyenne et l'écart type des scores de performances peuvent être calculés pour estimer le biais et la variance de la performance de validation.

Tableau de répartitions des données pour une validation croisée à k=3 blocs

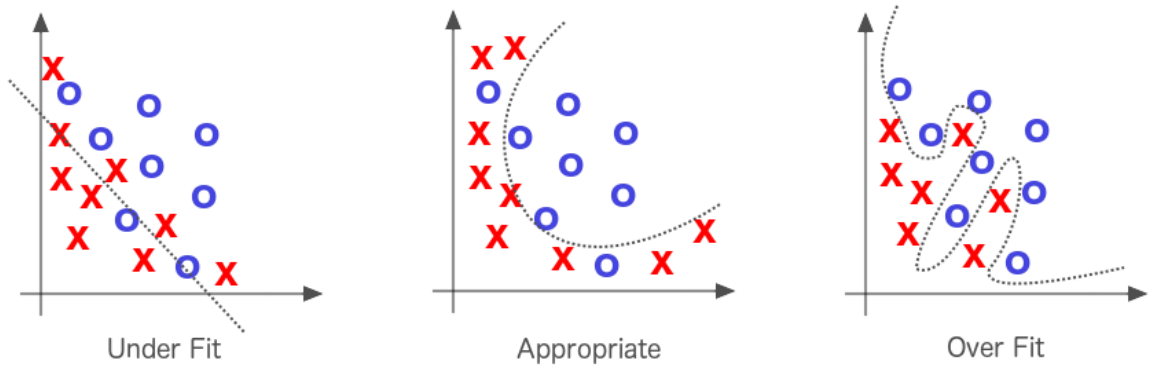
k	bloc 1	bloc 2	bloc 3
1	validation	apprentissage	apprentissage
2	apprentissage	validation	apprentissage
3	apprentissage	apprentissage	validation

- **« *leave-one-out cross-validation* » (LOOCV)** : cas particulier de la deuxième méthode où k=n, c'est-à-dire que l'on apprend sur n-1 observations puis on valide le modèle sur la énième observation et l'on répète cette opération n fois



Surapprentissage : on parle de surapprentissage (le terme anglais est overfitting) quand un modèle a trop appris les particularités de chacun des exemples fournis en exemple. Il présente alors un taux de succès très important sur les données d'entraînement (pouvant atteindre jusqu'à 100%), mais se généralise mal (performance moins bonnes sur les données de test).

Sousapprentissage : un algorithme qui n'apprend pas suffisamment de la phase d'apprentissage (mauvaise performance sur le training set)

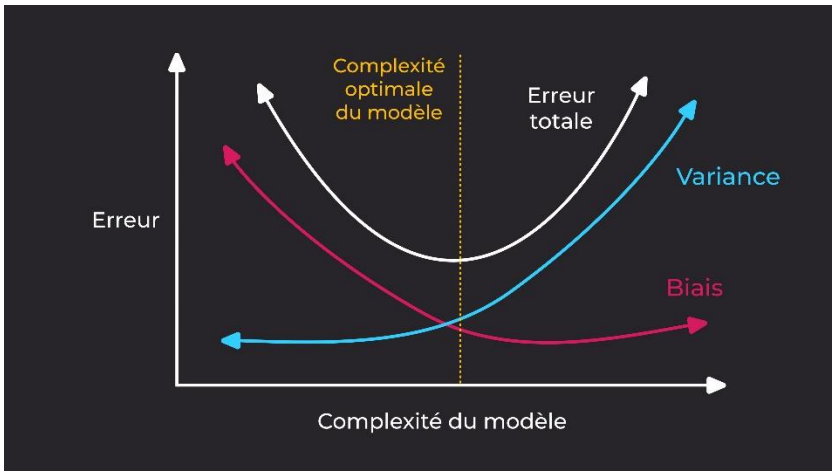


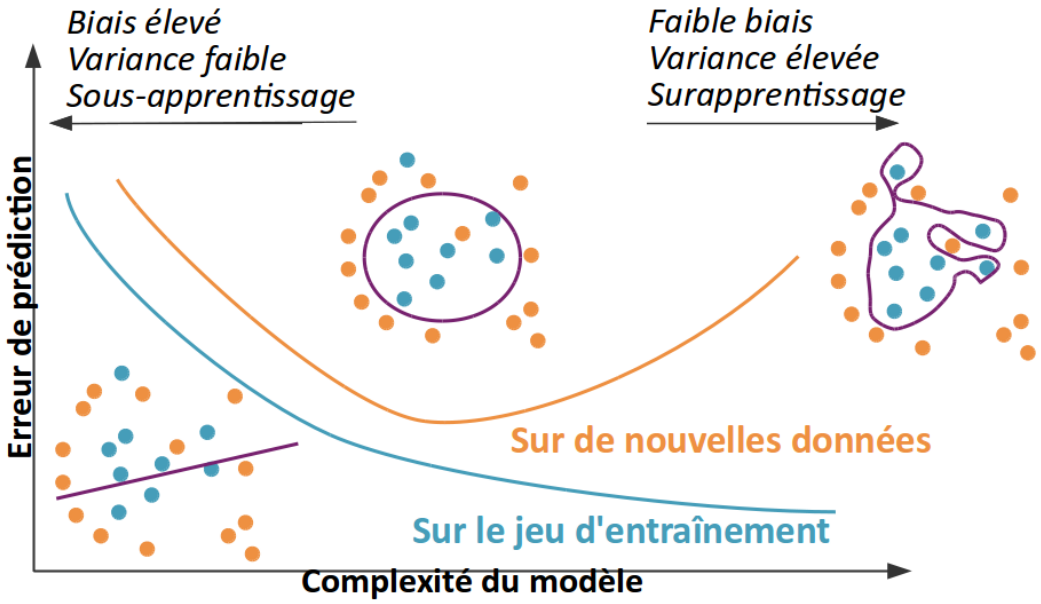
Dilemme Biais-Variance :

- Le *biais* est l'erreur provenant d'hypothèses erronées dans l'algorithme d'apprentissage. Un biais élevé peut être lié à un algorithme qui manque de relations pertinentes entre les données en entrée et les sorties prévues (sous-apprentissage).

- La *variance* est l'erreur due à la sensibilité aux petites fluctuations de l'échantillon d'apprentissage. Une variance élevée peut entraîner un surapprentissage, c'est-à-dire modéliser le bruit aléatoire des données d'apprentissage plutôt que les sorties prévues.

  - Le principe de **compromis entre biais et variance** est une des problématiques à laquelle vous serez confrontés lors de votre travail quotidien !
  - En utilisant un modèle comportant une **trop grande complexité**, dit "**à haute variance**", on peut mal capturer le phénomène sous-jacent et devenir trop dépendant aux données d'entraînement et aux petites fluctuations aléatoires, non représentatives du phénomène.
  - A contrario, il ne faut pas choisir un modèle **trop "simple"** qui biaise le résultat et ne parvient pas à capturer toute la complexité du phénomène.





Mesures de performance :

Précision (ou valeur prédictive positive) est la proportion des items pertinents pour une classe parmi l'ensemble des items proposés à cette classe,

Ra**ppel** (ou sensibilité) est la proportion des items pertinents proposés pour une classe parmi l'ensemble des items réellement pertinents pour cette classe.

- Proportion de Positifs ($TP+FN$) classés correctement :
  ➔ Sensibilité ou Rappel (Recall) $= \frac{TP}{TP+FN}$

- Proportion de Négatifs ($TN+FP$) classés correctement :
  ➔ Spécificité $= \frac{TN}{TN+FP}$

- Proportion des classés Positifs ($TP+FP$) correctement classés :
  ➔ Précision $= \frac{TP}{TP+FP}$

		Prédite	
		Positif	Négatif
Actuelle	Positif	TP	FN
	Négatif	FP	TN

Une mesure qui combine la précision et le rappel est leur moyenne harmonique, nommée F-mesure ou *F-score* :

$$F = 2 * \frac{Précision * Rappel}{\text{Précision+Rappel}}$$

Accuracy = nb instances correctement classifiés / nb total d'instances

	Yes	NO
Yes	7	2
No	4	1

$Precision_{Yes}$ = 7/7+4 =7/11 $Rappel_{Yes}$ = 7/7+2 =7/9

$Precision_{No}$ = 1/1+2 =1/3 $Rappel_{No}$ =1/1+4 =1/5

Accuracy = (7+1)/14 = 8/14