Groupe Biopuces


# Elise Maigné

**elise.maigne@inrae.fr**


February 28, 2024

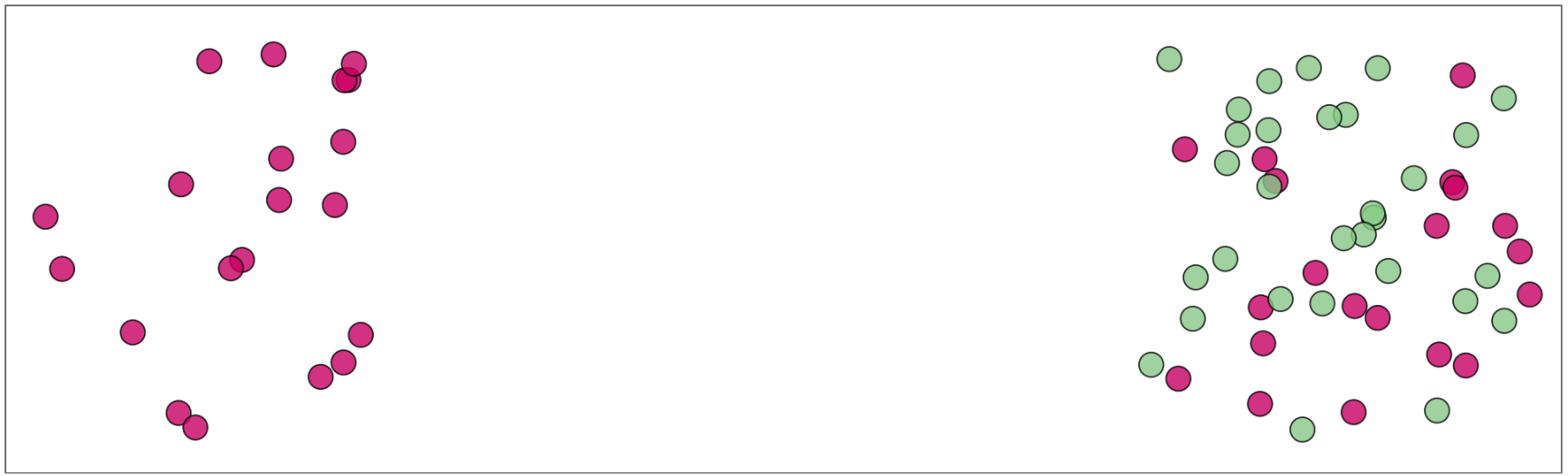
RÉPUBLIQUE FRANÇAISE
*Liberté*
*Égalité*
*Fraternité*

INRAE

# Modèle descriptif VS modèle prédictif

## Analyse descriptive/explicative

Faire un modèle pour expliquer (**décrire**) ce qu'il y a dans les données (la force d'une relation, l'intensité de phénomènes observés sur les données, …).



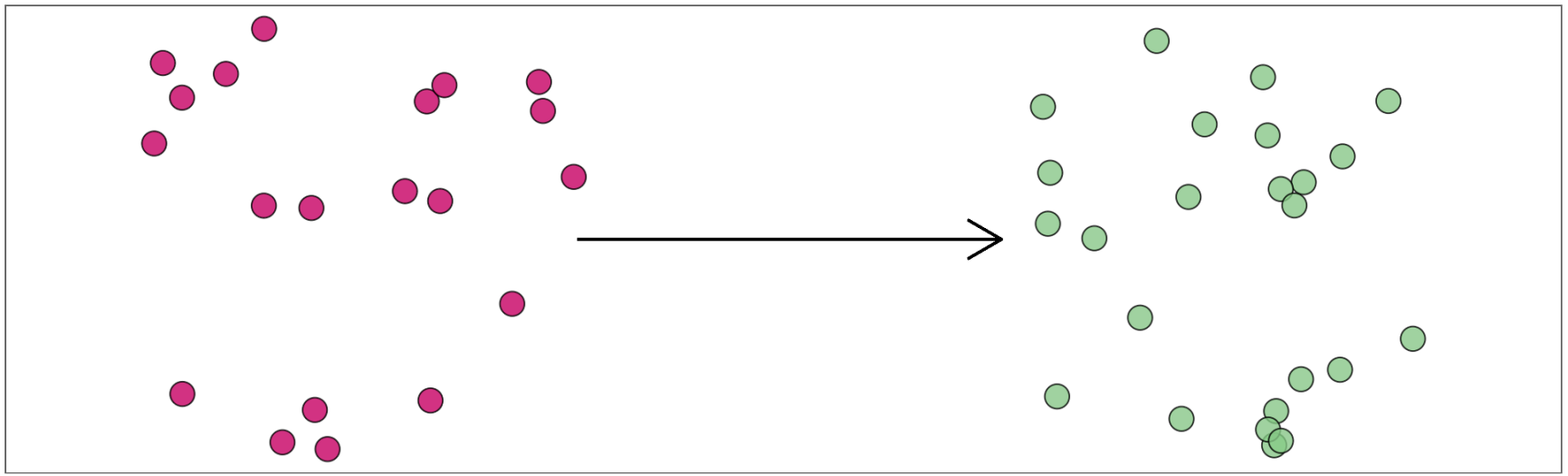
RÉPUBLIQUE FRANÇAISE
*Liberté Égalité Fraternité*

INRAE

# Modèle descriptif VS modèle prédictif

## Analyse prédictive

Construire un modèle pour catégoriser (**prédire**) sur un individu une caractéristique inconnue.



RÉPUBLIQUE FRANÇAISE
*Liberté*
*Égalité*
*Fraternité*

INRAE

# Qu’est-ce que le surapprentissage ?

**Objectif** : séparer ce qui est de la tendance et ce qui est du bruit.

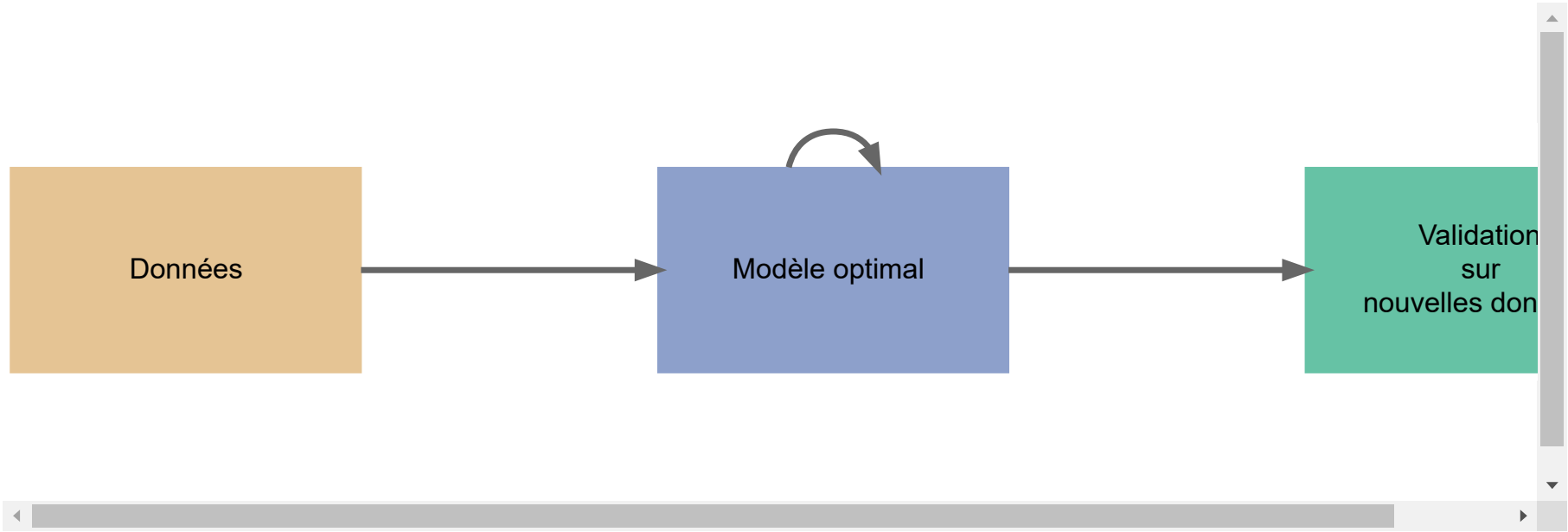
# Qu’est-ce que le surapprentissage ?

plus de biais, moins de variance

plus de variance, moins de biais

On cherche le modèle **le plus simple** qui donne les meilleurs résultats.

# Le processus du machine learning




## Phase d'entrainement

On ajuste un modèle jusqu'à avoir de bons résultats (ajustement des paramètres, choix des variables, …).
Processus itératif.

## Phase de test

# On teste/valide ce modèle..

# Processus général de machine learning

# Processus général de machine learning

# Processus général de machine learning

Classiquement 60%/40%, 70%/30% ou 80%/20%.

- **train** : l'échantillon sur lequel on va ajuster le modèle.
- **test** : l'échantillon sur lequel on va vérifier que le modèle s'applique bien à un autre jeu de données.

# Processus général de machine learning

L’échantillon test ne doit JAMAIS être utilisé pendant la phase d’entraînement
N’empêche pas le sur-apprentissage !

# Processus général de machine learning

L'échantillon de **validation** va nous servir à valider le modèle, pendant la phase d'entraînement.
N'empêche toujours pas le sur-apprentissage !

## Processus général de machine learning

Processus générique machine learning (Source **Boehmke and Greenwell 2019**)

## Validation croisée

- **Holdout** (Train/Test/Validation seulement pour grands jeux de données, et sous certaines conditions) (voir **Molinaro, Simon, and Pfeiffer 2005**; **Hawkins, Basak, and Mills 2003**)

Exemples de techniques de validation croisée :

- **Repeated random sub-sampling**
- **k-fold**
- **LOOCV** (leave-one-out cross-validation)
- **Bootstrapping**
- …

Voir par exemple :

```
1  ?caret::trainControl # Voir argument method
```

RÉPUBLIQUE FRANÇAISE
*Liberté Égalité Fraternité*

INRAE

# Validation croisée exemple k-fold

Exemple k-fold. (Source **Boehmke and Greenwell 2019**)

# La théorie VS la vie

## La théorie VS la vie

On trouve au choix :

- petits échantillons
- classes déséquilibrées
- plus de variables que d'individus

On peut contrôler :

- le tirage aléatoire
- le processus de validation (validation croisée)
- le choix des méthodes

# Contrôle du tirage aléatoire

# Comment diviser les données ?

Pour la division train/test :

## Simple tirage aléatoire

On va tirer aléatoirement 30% (ou 20% ou 40%) des données totales pour former l'échantillon test. Le reste formera l'échantillon d'apprentissage.

## Tirage aléatoire stratifié

On veut que la distribution de la variable à expliquer (cible) soit la même dans les échantillons. Le tirage se fait soit par quantile (var. continue) soit par classe (var. catégorielle). Parfois utile par exemple dans des distributions déséquilibrées.

Dans tous les cas toujours fixer une graine pour pouvoir reproduire le tirage (par ex. en R : `set.seed()`).

RÉPUBLIQUE FRANÇAISE
*Liberté Égalité Fraternité*

INRAE

# Déséquilibre des classes

## Déséquilibre des classes

Exemple dans un problème de classification on essaie de prédire Y qui est distribué 1: 5% et 0: 95%.

---

Cas de déséquilibre des classes : une classe est largement minoritaire.

# Déséquilibre des classes - Rééchantillonnage de la phase d'entrainement

# Déséquilibre des classes - Rééchantillonnage de la phase d'entrainement

# Déséquilibre des classes - Rééchantillonnage de la phase d'entrainement

Selon les modèles on n'est pas obligé d'avoir 50/50. Ex. Arbres : 5/10% peuvent suffire.

## Déséquilibre des classes - Rééchantillonnage de la phase d'entrainement

- UpSampling (ou OverSampling) : clonage aléatoire (plutôt pour méthodes linéaires)

  Et aussi :

- **SMOTE** (Synthetic Minority Oversampling TEchnique - (**Chawla et al. 2002**)), **SMOTE-NC** si variables catégorielles,

  Des individus ressemblant à ceux de la classe minoritaire sont générés. 2 paramètres, $k$ et $\alpha$. moyenne pondérée d'un voisin parmi les k plus proches.

- **ROSE** (Random Over Sampling Examples - (**Menardi and Torelli 2014**)).

  Technique basée sur le bootstrap et des méthodes à noyaux. Génère des individus artificiels autour des individus. 2 paramètres à ajuster.

```
1  ?caret::trainControl # Voir argument sampling
```

## Déséquilibre des classes - Autres méthodes

Peut être pris en compte directement par l'algorithme (à ne pas combiner avec un rééquilibrage).

- rééchantillonnage interne à la volée, sur-pondération de la classe minoritaire (poids), …

Pris en compte sur l'erreur de classification :

- on n'utilise pas la MSE mais plutôt recall, precision, ou F1 score
- on ajuste le calcul à postériori (ex. Balanced error rate, …)

Dans tous les cas, une méthode mal utilisée peut accroître les biais du modèle.

RÉPUBLIQUE FRANÇAISE
*Liberté Égalité Fraternité*

INRAE

# Peu d’observations

# Peu d’observations

Risque aigu.

## Peu d'observations

- Augmenter la taille des données
- Rechercher des modèles simples : regrouper classes, diminuer le nombre de variables.
- Pour modèles d'arbres : diminuer la profondeur max, …
- Attention aux aberrants
- Faire de la sélection de variables
- LOOCV est bien adaptée quand il y a peu de données

RÉPUBLIQUE FRANÇAISE
*Liberté Égalité Fraternité*

INRAE

## Peu d’observations - LOOCV

LOOCV [Source : https://biol607.github.io/lectures/crossvalidation.html#31]

## Peu d'observations

- Utiliser des approches ensemblistes (combiner plusieurs modèles - bagging, boosting, stacking, …) pour diminuer la variance.

---

RÉPUBLIQUE FRANÇAISE
*Liberté*
*Égalité*
*Fraternité*

INRAE

## Peu d’observations

- Utiliser des approches ensemblistes (combiner plusieurs modèles) pour diminuer la variance.

---

# Hyperparamètres

# Hyperparamètres

**Paramètres** = estimés par le modèle $\neq$ **Hyperparamètres** fixés avant de faire tourner le modèle.

Exemple :

$y = \alpha + \beta x + \epsilon$

Puis on introduit une pénalité g(x) plus ou moins forte (> 0) :

$y = \alpha + \beta x + \epsilon + \lambda g(x)$

A combien doit-on fixer $\lambda$ ?

Ou $mtry$ dans les randomforests, ...

RÉPUBLIQUE FRANÇAISE
*Liberté Égalité Fraternité*

INRAE

## Hyperparamètres - Optimisation

Les différents paramètres sont à ajuster pendant le processus d'entrainement, le taux de rééchantillonnage peut l'être aussi.
**Techniques :**

- Essai de plusieurs valeurs. Lourde et attention à la reproductibilité mais relativement rapide.
- Recherche par grille (espace fini)
- L'optimisation bayesienne (espace des paramètres)

Il existe des optimiseurs pour les hyperparamètres (voir par exemple `{tune}`).

RÉPUBLIQUE FRANÇAISE
*Liberté Égalité Fraternité*

INRAE

# Hyperparamètres - Optimisation

Source : https://leandeep.com/tradeoff-biais-variance/

Trouver un compromis entre biais et variance.

# Processus général de machine learning

Processus générique machine learning (Source **Boehmke and Greenwell 2019**)

## Références

Boehmke, Brad, and Brandon M Greenwell. 2019. *Hands-on Machine Learning with r*. CRC press.

Chawla, Nitesh V, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. “SMOTE: Synthetic Minority over-Sampling Technique.” *Journal of Artificial Intelligence Research* 16: 321–57.

Hawkins, Douglas, Subhash Basak, and Denise Mills. 2003. “Assessing Model Fit by Cross-Validation.” *Journal of Chemical Information and Computer Sciences* 43 (March): 579–86. **https://doi.org/10.1021/ci025626i**.

Menardi, Giovanna, and Nicola Torelli. 2014. “Training and Assessing Classification Rules with Imbalanced Data.” *Data Mining and Knowledge Discovery* 28: 92–122.

Molinaro, Annette M., Richard Simon, and Ruth M. Pfeiffer. 2005. “Prediction error estimation: a comparison of resampling methods.” *Bioinformatics* 21 (15): 3301–7. **https://doi.org/10.1093/bioinformatics/bti499**.

RÉPUBLIQUE FRANÇAISE
*Liberté*
*Égalité*
*Fraternité*

INRAE