

CLASSIFICATION ET ESTIMATION DE LA VALEUR

« La science des données ne remplace pas la modélisation statistique et l'analyse des données, elle les augmente. »

(P. Boily)

« Les données ne sont pas des renseignements, les renseignements ne sont pas des connaissances, la connaissance n'est pas la compréhension, la compréhension n'est pas la sagesse. »

(Attribué à Cliff Stoll dans Nothing to Hide: Privacy in the 21st Century de Keeler, 2006)

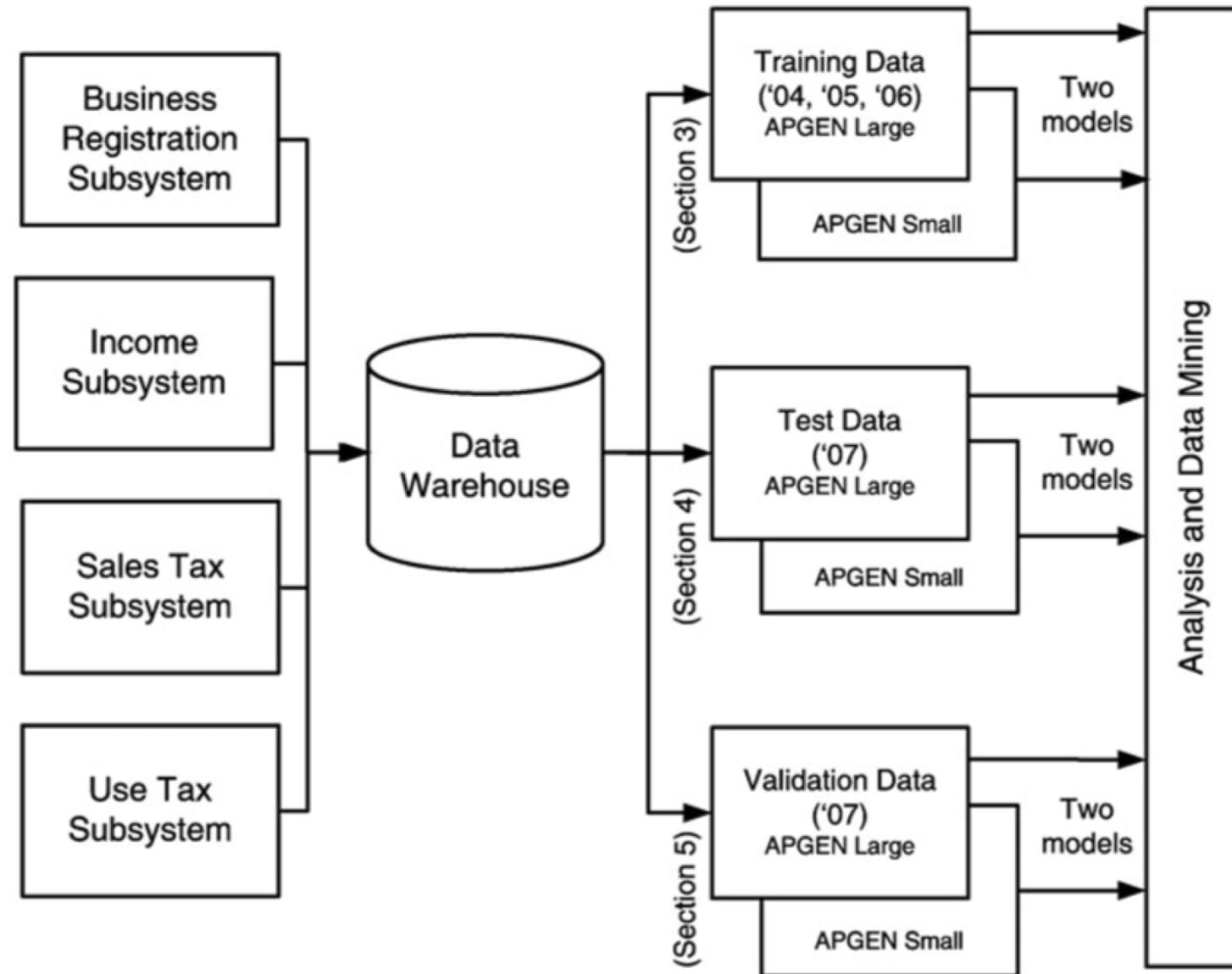
ÉTUDE DE CAS: VÉRIFICATION FISCALE DU MINNESOTA

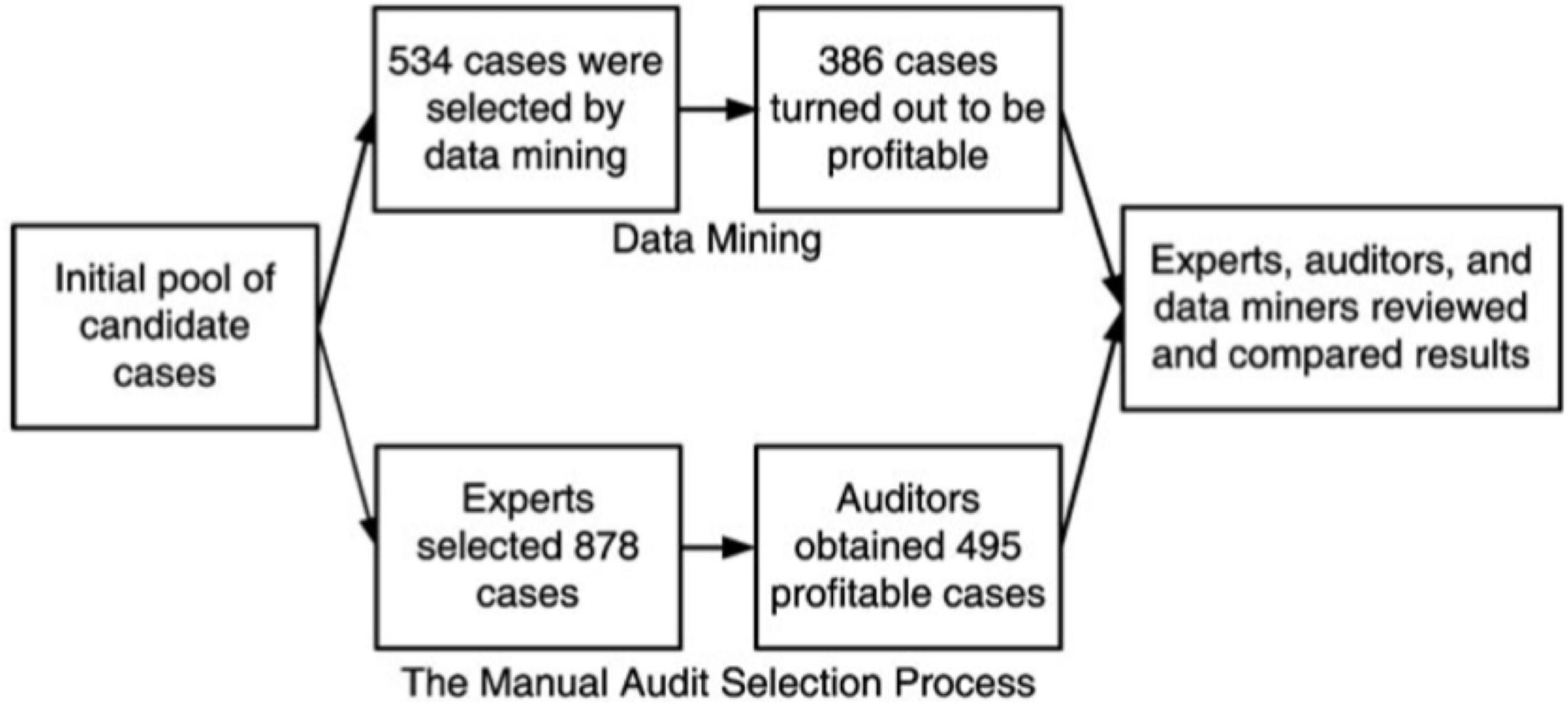
Les écarts importants entre les recettes dues (en théorie) et les recettes perçues (en pratique) sont problématiques pour les gouvernements.

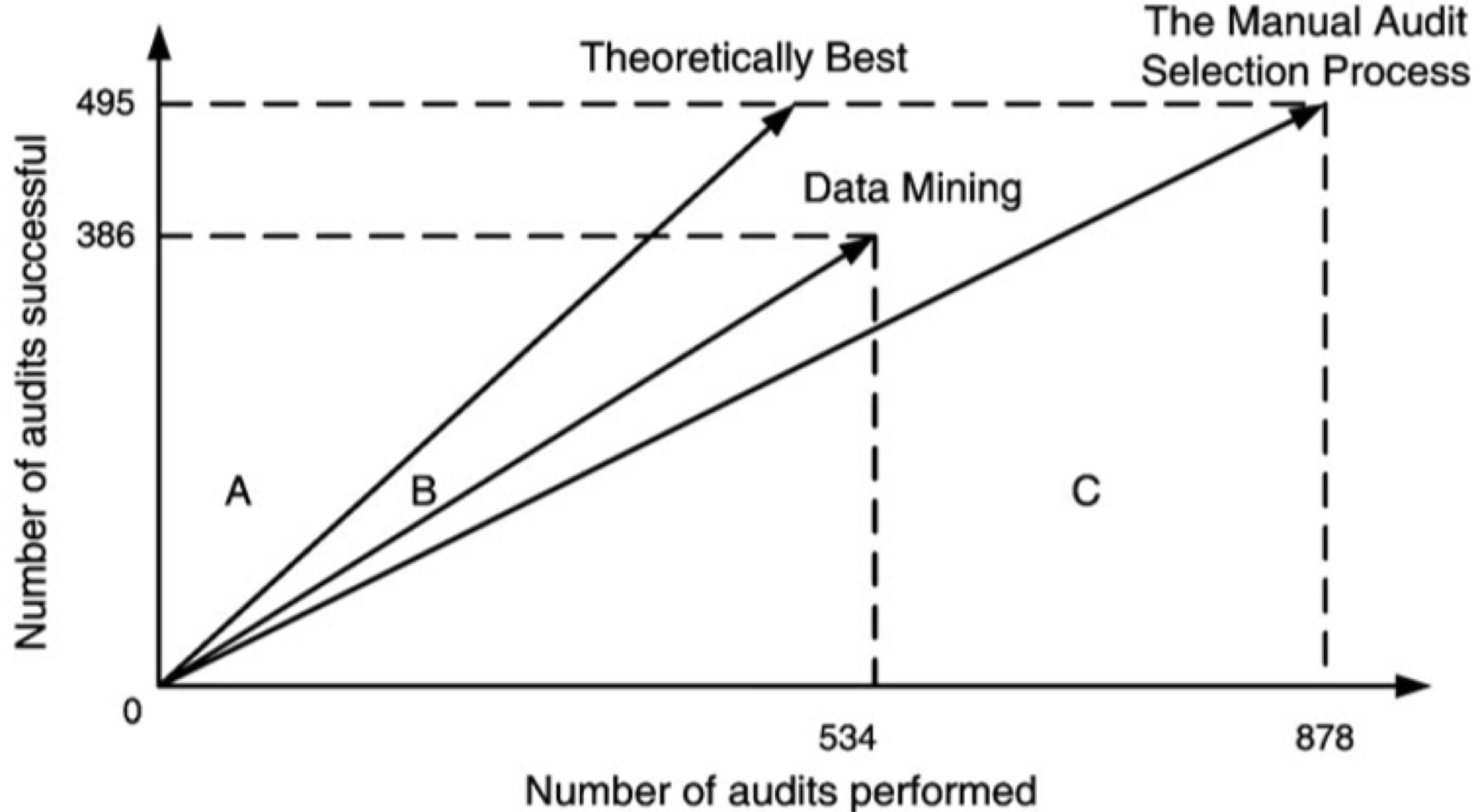
Les agences du revenu mettent en œuvre diverses stratégies de détection de la fraude (comme les examens de vérification) pour combler cette lacune.

Les audits d'entreprise sont coûteux – est-ce qu'il existe des algorithmes qui permettent de prédire si un audit sera probablement un succès ou un gaspillage de ressources?

Une agence de recouvrement de l'impôt devrait-elle chercher à maximiser ses revenus et ses profits ou assurer la conformité à la loi?







APERÇU DE LA CLASSIFICATION

En **classification**, un ensemble d'échantillons de données (l'ensemble de **formation**) est utilisé pour déterminer les règles et les tendances qui divisent les données en groupes ou classes prédéterminés (apprentissage supervisé; analyse prédictive).

Les données de formation sont habituellement constituées d'un sous-ensemble de données **étiquetés** (cible) choisies **au hasard**.

L'**estimation de la valeur** (régression) s'apparente à la classification lorsque la variable cible est numérique.

APERÇU DE LA CLASSIFICATION

Dans la phase d'**essai**, le modèle est utilisé pour assigner une catégorie aux observations pour lesquelles l'étiquette est cachée, mais au bout du compte connue (l'ensemble d'**essais**).

Le rendement d'un modèle de classification est évalué sur l'ensemble d'essais, **jamais** sur l'ensemble de formation.

Questions techniques :

- la sélection des caractéristiques à inclure dans le modèle
- la sélection de l'algorithme
- etc.

APPLICATIONS

Médecine et sciences de la santé

- prédire quel patient risque de subir une deuxième crise cardiaque mortelle dans les 30 jours en fonction de facteurs de santé (tension artérielle, âge, problèmes de sinus, etc.)

Politiques sociales

- prédire la probabilité d'avoir besoin d'une aide au logement pour les personnes âgées à partir de données démographiques et de réponses à des sondages

Marketing et affaires

- prédire quels clients sont susceptibles de changer de fournisseur de téléphonie mobile en fonction de la démographie et de l'utilisation

EXEMPLE

Scénario :

Une compagnie d'assurance automobile a un service d'enquête sur les fraudes qui étudie jusqu'à 30 % de toutes les demandes d'indemnisation, mais elle perd toujours de l'argent en raison de demandes frauduleuses.

Questions : peut-on prédire

- si une réclamation est susceptible d'être frauduleuse?
- si un client est susceptible de commettre une fraude dans un avenir proche?
- si une demande d'assurance est susceptible de donner lieu à une réclamation frauduleuse?
- le montant dont une réclamation sera réduite si elle est frauduleuse?

Training Set (with labels)

	Y_1	Y_2	...	Y_p	■
01	$x_{01,1}$	$x_{01,2}$...	$x_{01,p}$	■
04	$x_{04,1}$	$x_{04,2}$...	$x_{04,p}$	■
10	$x_{10,1}$	$x_{10,2}$...	$x_{10,p}$	■
21	$x_{21,1}$	$x_{21,2}$...	$x_{21,p}$	■
22	$x_{22,1}$	$x_{22,2}$...	$x_{22,p}$	■
23	$x_{23,1}$	$x_{23,2}$...	$x_{23,p}$	■
25	$x_{25,1}$	$x_{25,2}$...	$x_{25,p}$	■
29	$x_{29,1}$	$x_{29,2}$...	$x_{29,p}$	■
...
**	$x_{**,1}$	$x_{**,2}$...	$x_{**,p}$	■

Testing Set (with labels)

	Y_1	Y_2	...	Y_p	■
02	$x_{02,1}$	$x_{02,2}$...	$x_{02,p}$	■
03	$x_{03,1}$	$x_{03,2}$...	$x_{03,p}$	■
05	$x_{05,1}$	$x_{05,2}$...	$x_{05,p}$	■
06	$x_{06,1}$	$x_{06,2}$...	$x_{06,p}$	■
07	$x_{07,1}$	$x_{07,2}$...	$x_{07,p}$	■
08	$x_{08,1}$	$x_{08,2}$...	$x_{08,p}$	■
09	$x_{09,1}$	$x_{09,2}$...	$x_{09,p}$	■
11	$x_{11,1}$	$x_{11,2}$...	$x_{11,p}$	■
...
@@	$x_{@@,1}$	$x_{@@,2}$...	$x_{@@,p}$	■

Predictions

	■	a	p
02	■	■	■
03	■	■	■
05	■	■	■
06	■	■	■
07	■	■	■
08	■	■	■
09	■	■	■
11	■	■	■
...
@@	■	■	■

Performance Evaluation

Deployment

Classifier

Model

Classes

SYSTÈMES DE CLASSIFICATION

Régression logistique

Réseaux neuronaux

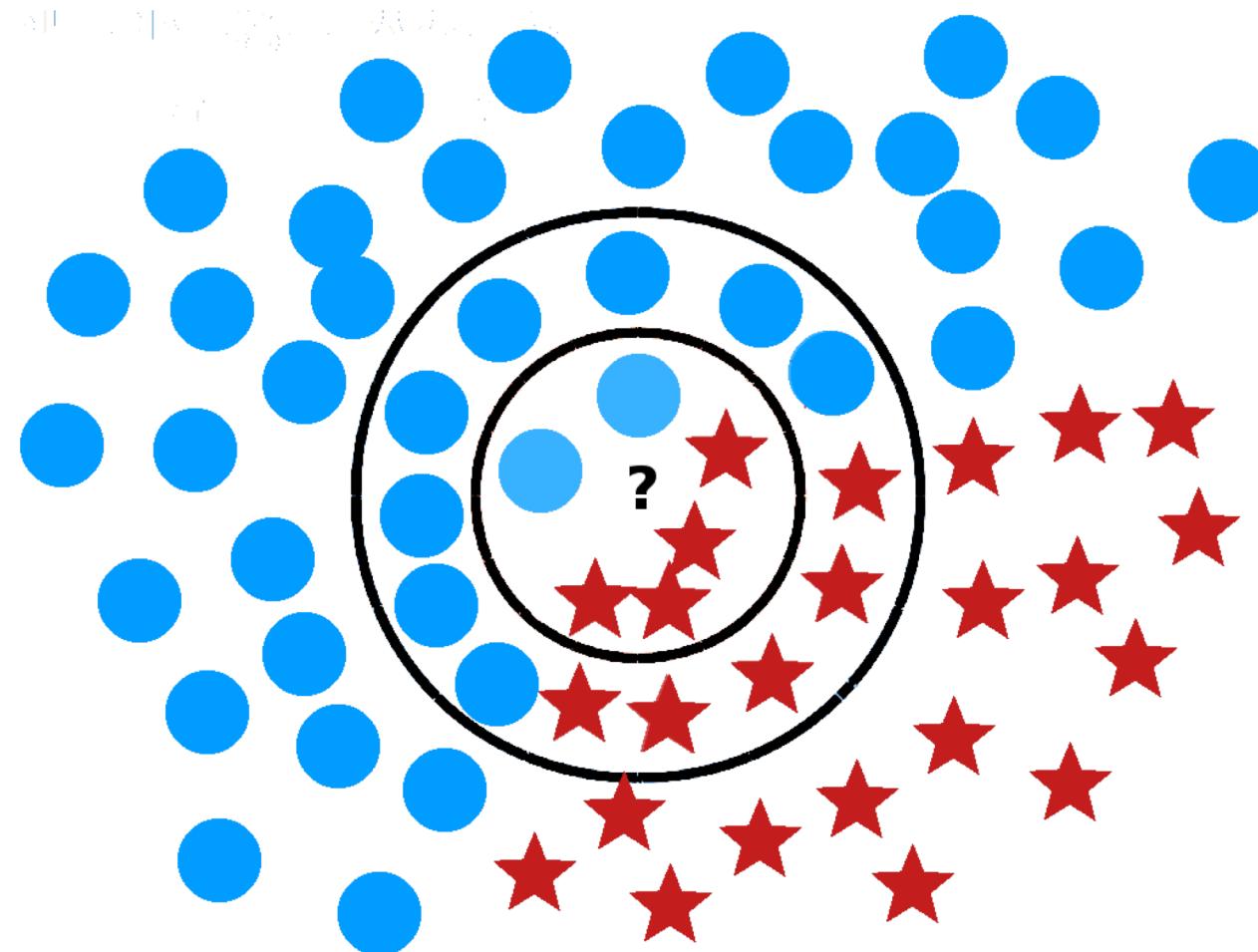
Arbres de décision

Classificateur bayésien naïf

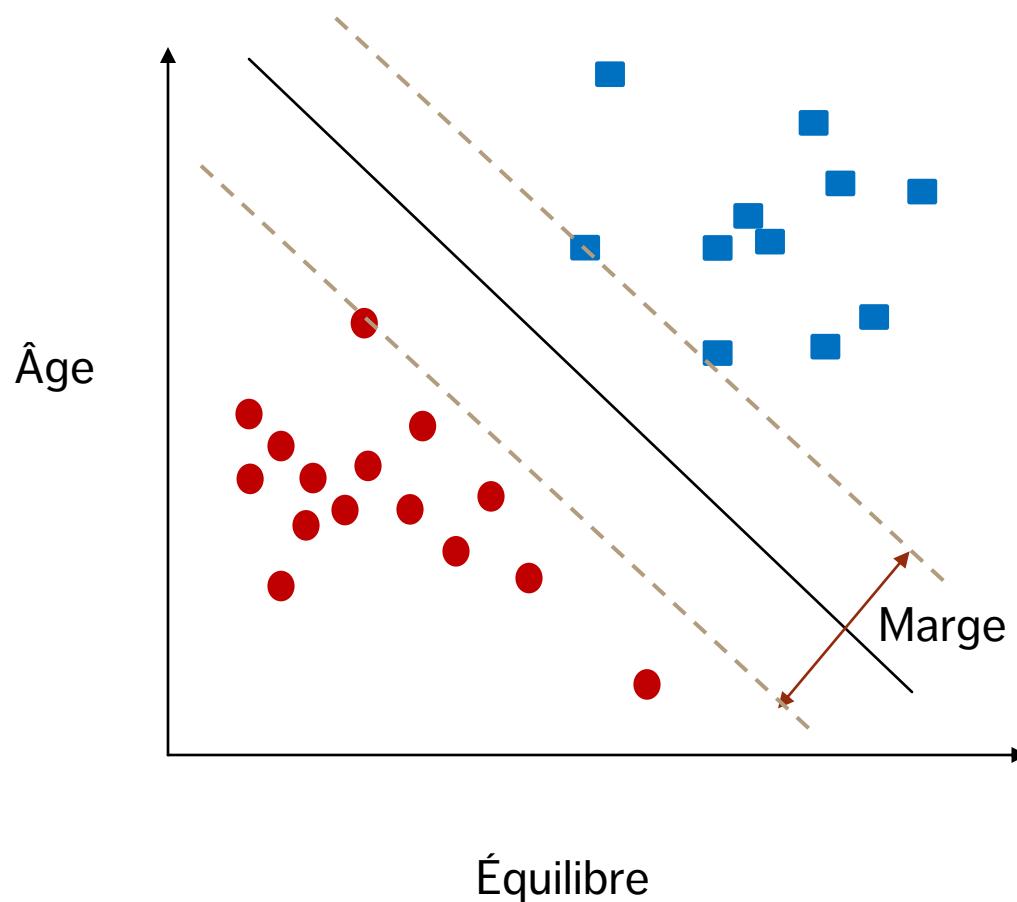
Machines à vecteurs de support

Classificateurs des voisins les plus proche

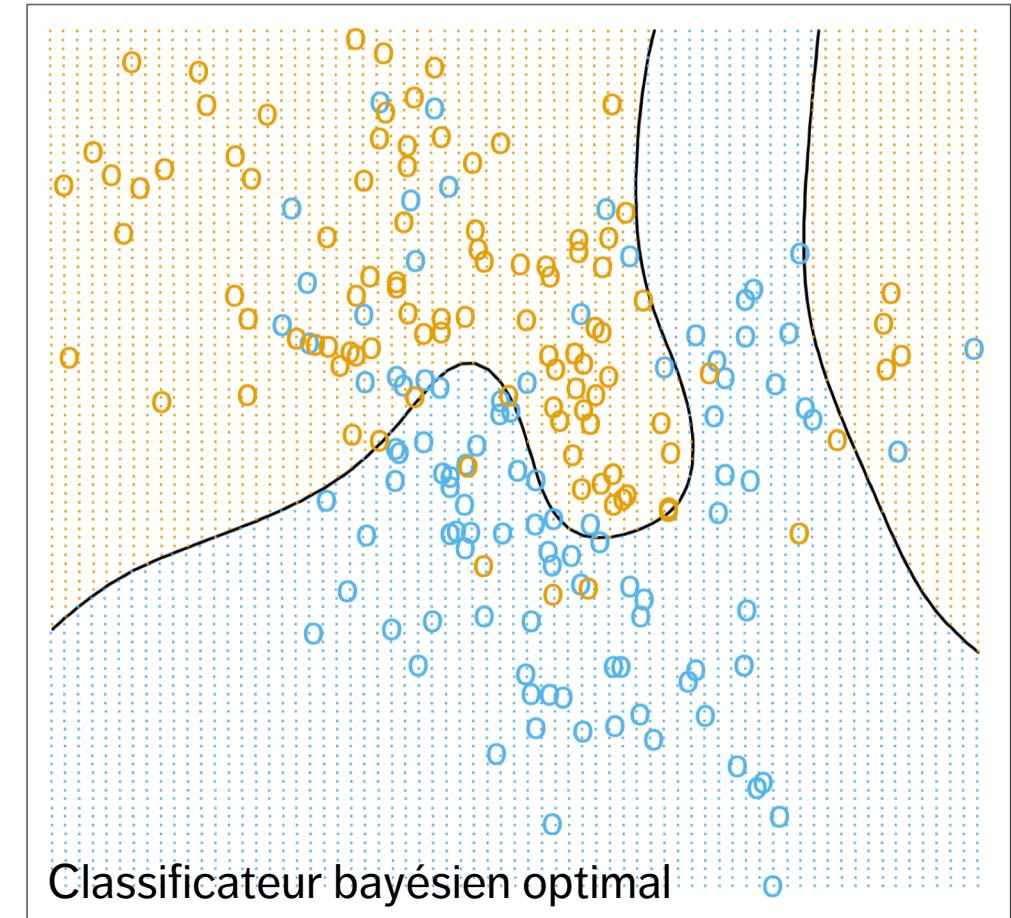
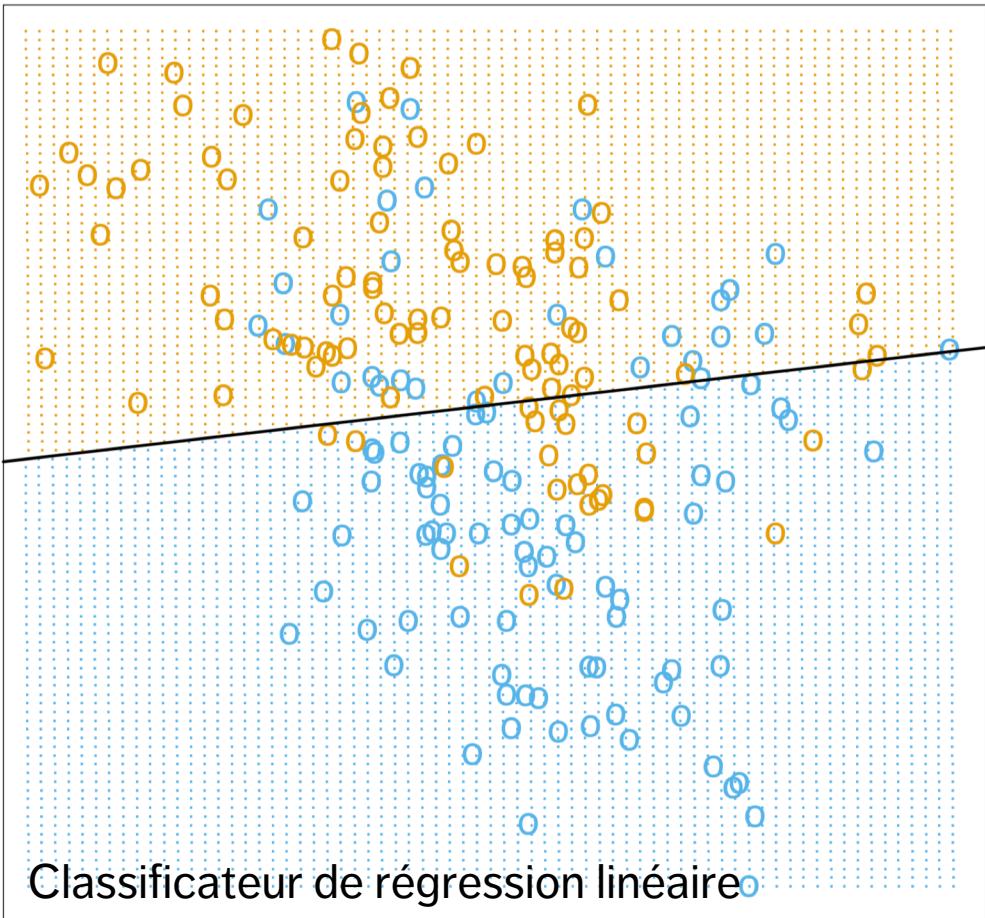
k VOISINS LES PLUS PROCHES



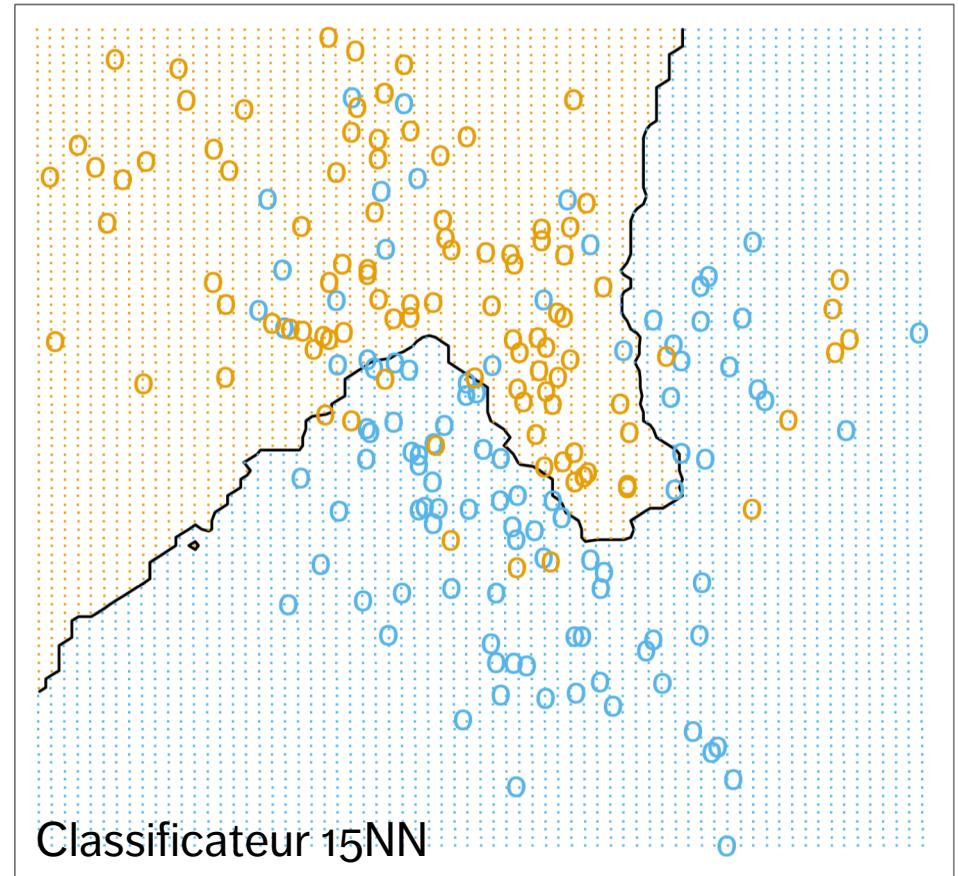
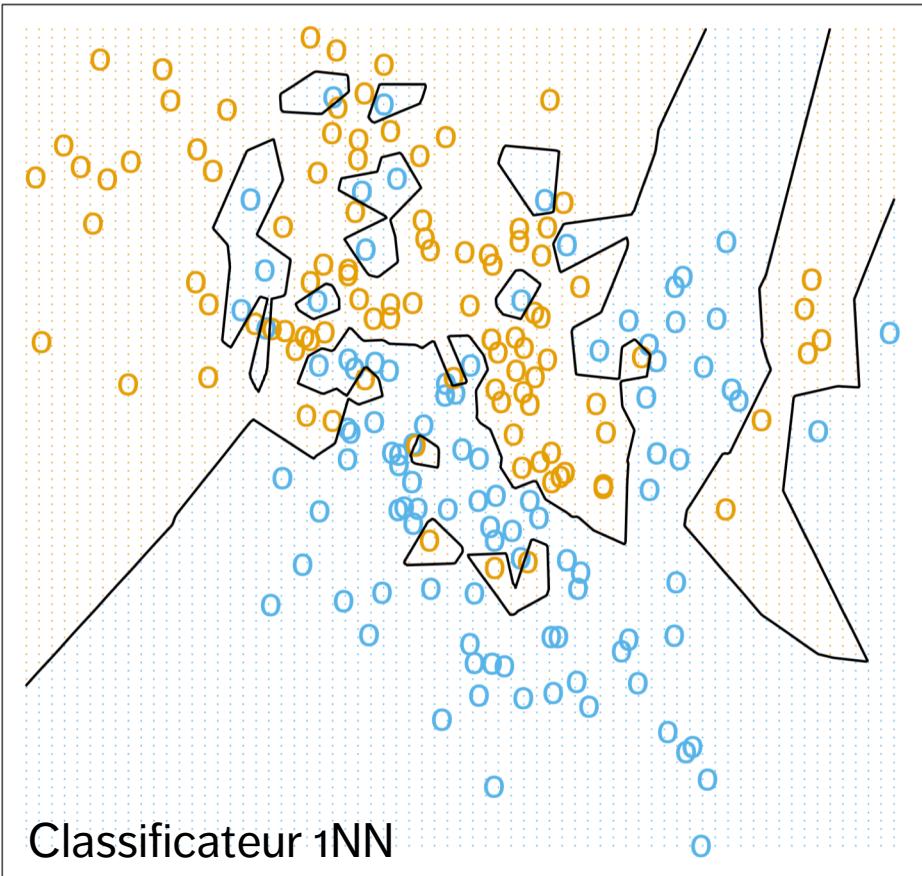
MACHINE À VECTEURS DE SUPPORT



ILLUSTRATION

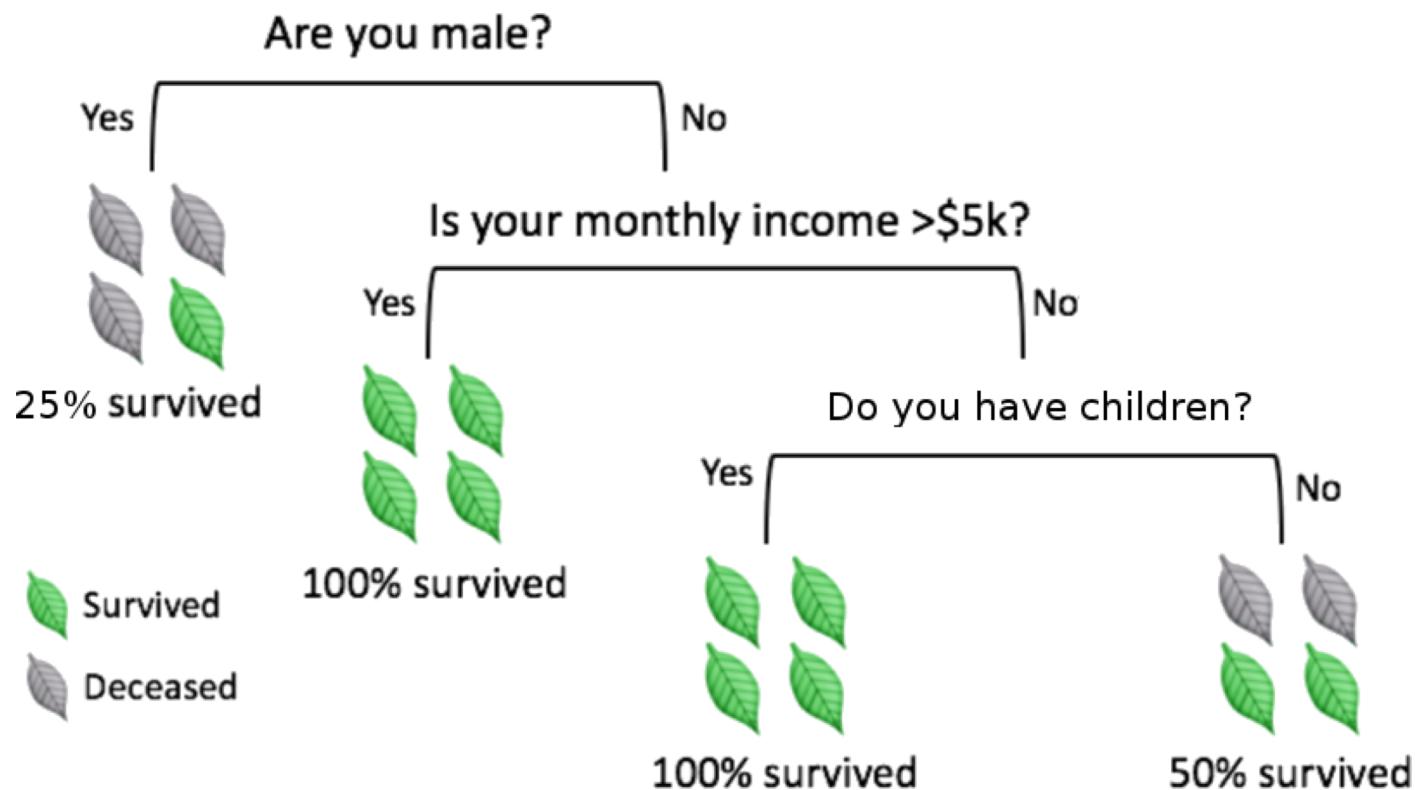


ILLUSTRATION



ARBRES DE DÉCISION

Les arbres de décision constituent probablement la plus **intuitive** de ces méthodes : la classification s'effectue en suivant un tracé le long d'un arbre, de ses **racines**, à travers ses **branches**, pour se terminer dans ses **feuilles**.



ARBRES DE DÉCISION

Afin d'effectuer une **prédition** pour une nouvelle instance, suivez le tracé le long de l'arbre, lisant la prédition directement une fois qu'une feuille est atteinte.

Créer l'arbre et suivre le tracé peuvent **prendre du temps** s'il y a trop de variables.

L'exactitude des prévisions peut être préoccupante pour les arbres dont la croissance n'est **pas contrôlée**. Dans la pratique, le critère de **pureté** au niveau des feuilles est lié à de mauvais taux de prédition pour de nouvelles instances.

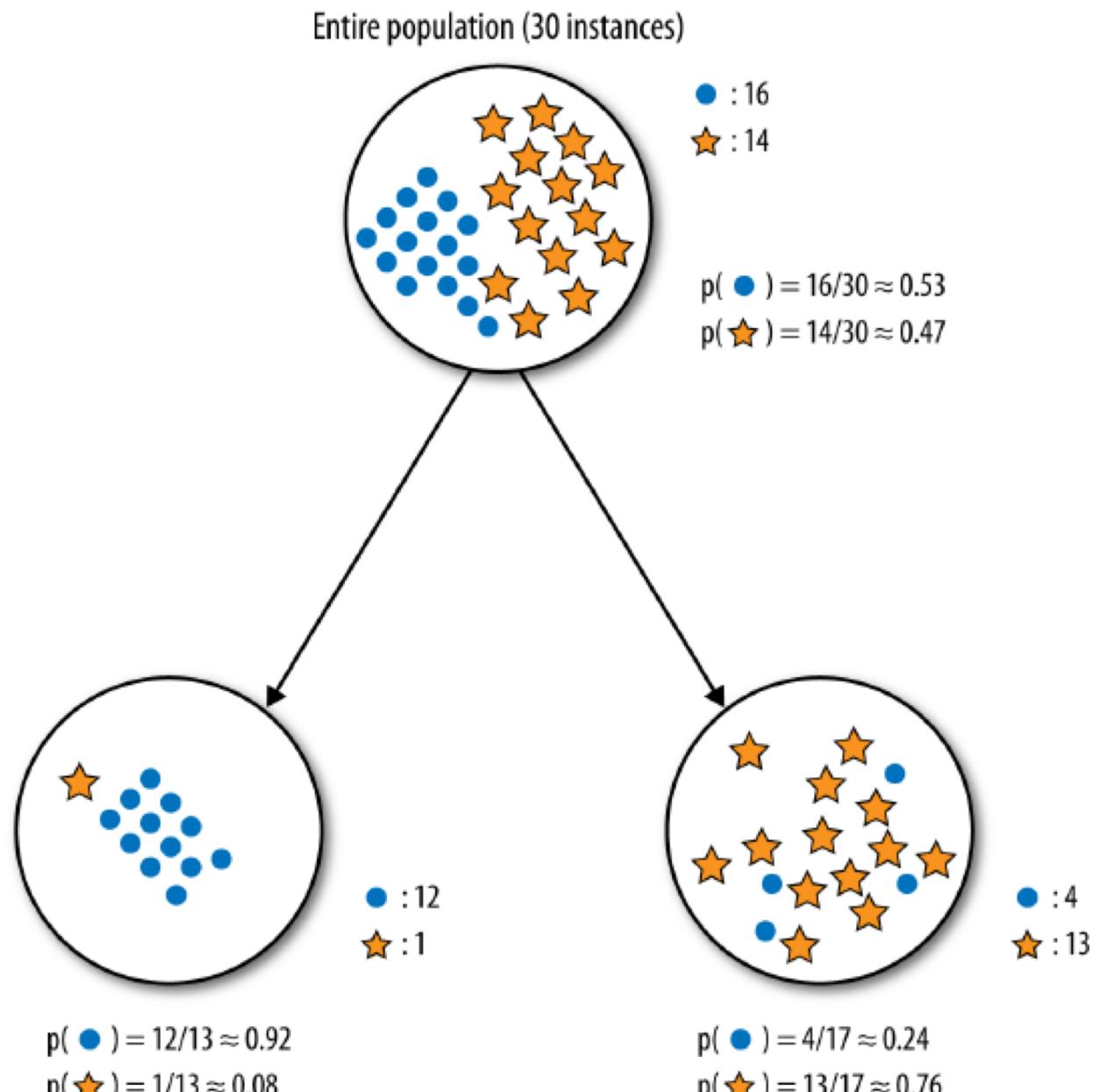
- d'autres critères sont souvent utilisés pour « élaguer » les arbres, ce qui peut conduire à des feuilles **impures** (c.-à-d. avec une entropie non triviale).

ALGORITHME DE L'ARBRE DE DÉCISION (ID3)

Tâche : développer un arbre de décision à l'aide d'un ensemble de formation (un sous-ensemble de données pour lequel la classification correcte de la cible est connue).

Aperçu :

1. Répartir les données de formation (« **parents** ») en sous-ensembles (« **enfants** »), en utilisant les différents niveaux d'un attribut particulier.
2. Calculer le **gain d'information** pour chaque sous-ensemble
3. Sélectionner la répartition la **plus avantageuse**
4. Répéter l'opération pour chaque nœud jusqu'à ce que certains des critères de **feuille** soient respectés (chaque élément de la feuille a la même classification?)



$$= -\frac{1}{30} \log \frac{1}{30} - \frac{1}{30} \log \frac{1}{30} \approx 0.99$$

$$E(L) = -p_o \log p_o - p_* \log p_*$$

$$= -\frac{12}{13} \log \frac{12}{13} - \frac{1}{13} \log \frac{1}{13} \approx 0.39$$

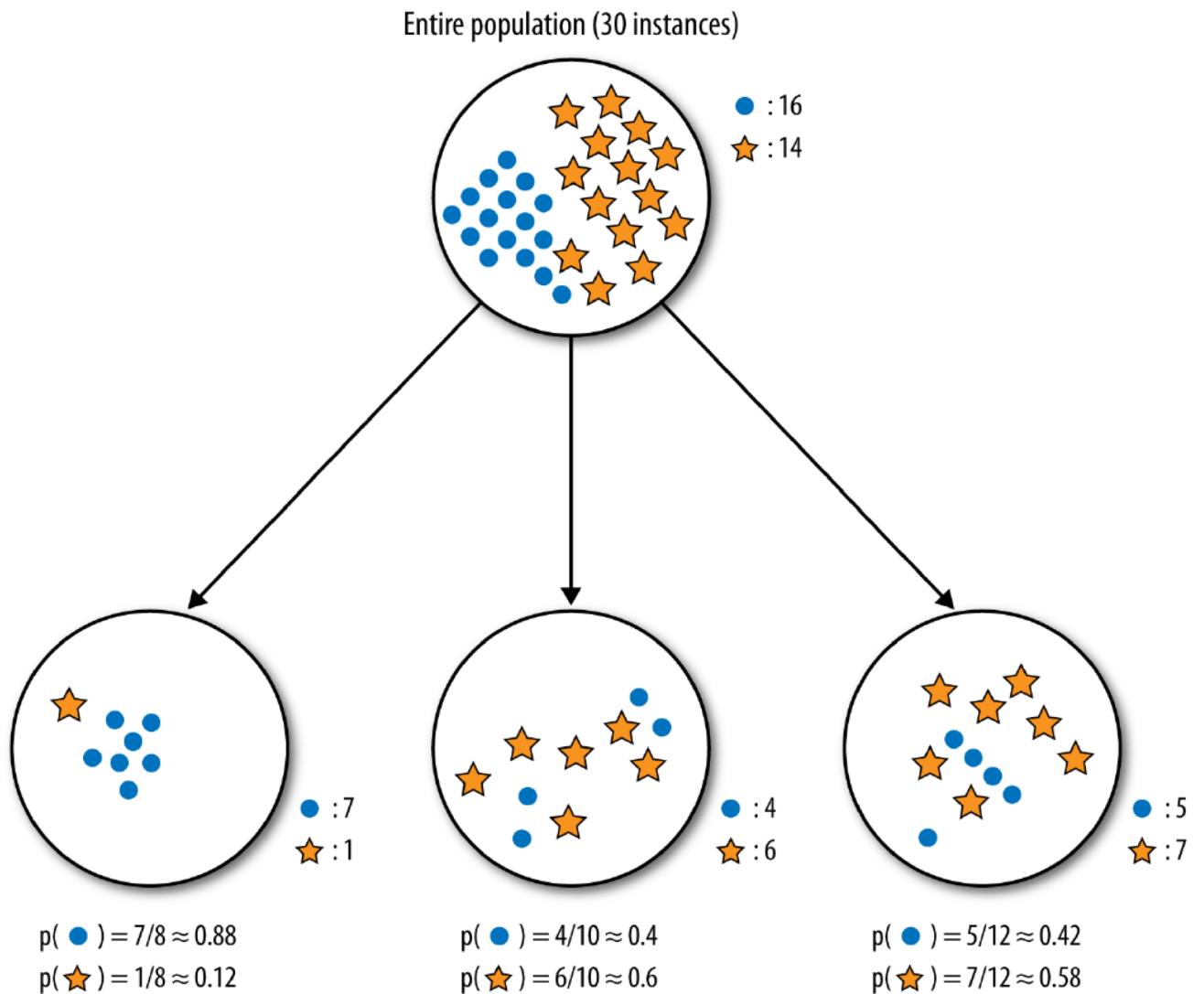
$$E(R) = -p_o \log p_o - p_* \log p_*$$

$$= -\frac{4}{17} \log \frac{4}{17} - \frac{13}{17} \log \frac{13}{17} \approx 0.79$$

$$\text{IG} = E(S) - \frac{1}{30}[q_L E(L) + q_R E(R)]$$

$$\approx 0.99 - \frac{1}{30}[13(0.39) + 17(0.79)]$$

$$\approx \mathbf{0.37}$$



$$\begin{aligned}
 E(L) &= -p_o \log p_o - p_* \log p_* \\
 &= -\frac{7}{8} \log \frac{7}{8} - \frac{1}{8} \log \frac{1}{8} \approx 0.54 \\
 E(C) &= -p_o \log p_o - p_* \log p_* \\
 &= -\frac{4}{10} \log \frac{4}{10} - \frac{6}{10} \log \frac{6}{10} \approx 0.97 \\
 E(R) &= -p_o \log p_o - p_* \log p_* \\
 &= -\frac{5}{12} \log \frac{5}{12} - \frac{7}{12} \log \frac{7}{12} \approx 0.98 \\
 \text{IG} &= E(S) - \frac{1}{30}[q_L E(L) + q_C E(C) + q_R E(R)] \\
 &\approx 0.99 - \frac{1}{30}[8(0.54) + 10(0.97) + 12(0.98)] \\
 &\approx \mathbf{0.13}
 \end{aligned}$$

POINTS FORTS ET POINTS FAIBLES DES ARBRES DE DÉCISION

Modèle « **boîte blanche** »

Peut être utilisé avec des ensembles de données **incomplets**

Sélection des variables **intégrée**

Ne fait **aucune hypothèse** concernant

Pas aussi précis que les autres algorithmes (habituellement)

Pas robuste

Particulièrement vulnérable au **sure-ajustement**

L'apprentissage optimal de l'arbre de décision est **NP-complet**

Biaisé envers les variables catégoriques qui ont un grand nombre de niveaux

REMARQUES CONCERNANT LES ARBRES DE DÉCISIONS

Méthode de fractionnement

- gain d'information, impureté de Gini, réduction de la variance, etc.

Algorithmes communs

- Dichotomiseur itératif 3, C4.0, C4.5, CHAID, MARS, arbres d'inférences conditionnelles, CART

Les arbres de décision peuvent également être combinés entre eux à l'aide d'algorithmes de boosting (**AdaBoost**) ou des **forêts aléatoire**, offrant ainsi un type de procédure de vote (apprentissage par ensembles).

AUTRES FACTEURS À PRENDRE EN CONSIDÉRATION

La classification est liée à l'estimation des probabilités

- des approches fondées sur des modèles de régression pourraient s'avérer fructueuses

Des évènements rares (souvent plus intéressants ou importants) continuent de nuire aux tentatives de classification

- les données historiques du réacteur nucléaire de Fukushima avant la fusion ne pouvaient pas être utilisées pour en savoir plus sur les fusions

Pas de théorème de la passe droite : aucun classificateur ne fonctionne mieux pour toutes les données.

Avec des données massives, les algorithmes doivent aussi tenir compte de l'efficacité.

ÉVALUATION DU RENDEMENT

Les classificateurs sont évalués sur l'ensemble du test.

Idéalement, un bon classificateur aurait des taux élevés de **Vrais positifs** (TP) et **Vrais négatifs** (TN), et des taux faibles de **Faux positifs** (FP, erreur de type I) et **Faux négatifs** (FN, erreur de type II).

Les paramètres d'évaluation ne signifient pas grand-chose en soi : le contexte exige une comparaison avec d'autres classificateurs et d'autres paramètres d'évaluation.

		Predicted		Total	79.0%
Actuals	A	54	10		
	B	6	11	17	21.0%
Total	60	21	81		
	74.1%	25.9%			

		Predicted		Total	66.7%
Actuals	A	54	0		
	B	16	11	27	33.3%
Total	70	11	81		
	86.4%	13.6%			

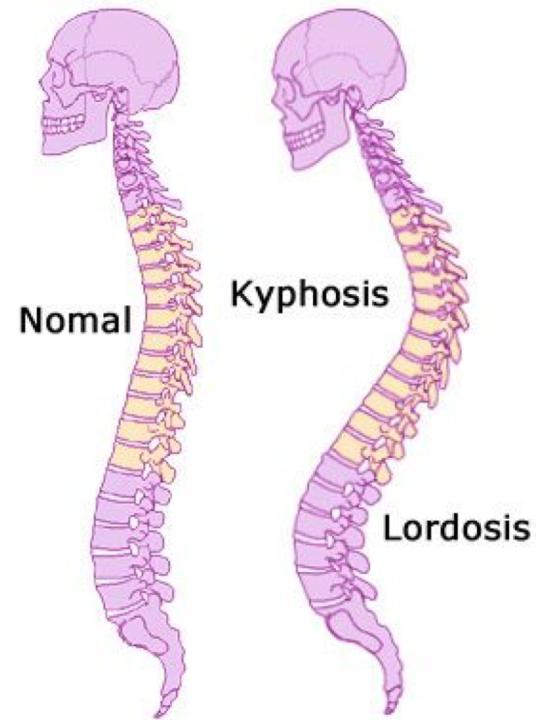
Classification Rates		Performance Metrics	
Sensitivity:	0.84	Accuracy:	0.80
Specificity:	0.65	F1-Score:	0.87
Precision:	0.90	Informedness (ROC):	0.49
Negative Predictive Value:	0.52	Markedness:	0.42
False Positive Rate:	0.35	M.C.C.:	0.46
False Discovery Rate:	0.10	Pearson's chi2:	0.01
False Negative Rate:	0.16	Hist. Stat:	0.10
Classification Rates		Performance Metrics	
Sensitivity:	1.00	Accuracy:	0.80
Specificity:	0.41	F1-Score:	0.87
Precision:	0.77	Informedness (ROC):	0.41
Negative Predictive Value:	1.00	Markedness:	0.77
False Positive Rate:	0.59	M.C.C.:	0.56
False Discovery Rate:	0.23	Pearson's chi2:	0.33
False Negative Rate:	0.00	Hist. Stat:	0.40

EXEMPLE – ENSEMBLE DE DONNÉES SUR LA CYPHOSE

La cyphose est une condition médicale liée à la courbure convexe excessive de la colonne vertébrale. La chirurgie corrective de la colonne vertébrale est parfois pratiquée sur des enfants.

L'ensemble de données comprend 81 observations et 4 attributs :

- **cypsose** (absente ou présente après l'opération)
- **âge** (au moment de l'opération, en mois)
- **nombre** (de vertèbres concernées)
- **point de départ** (vertèbre supérieure opérée)

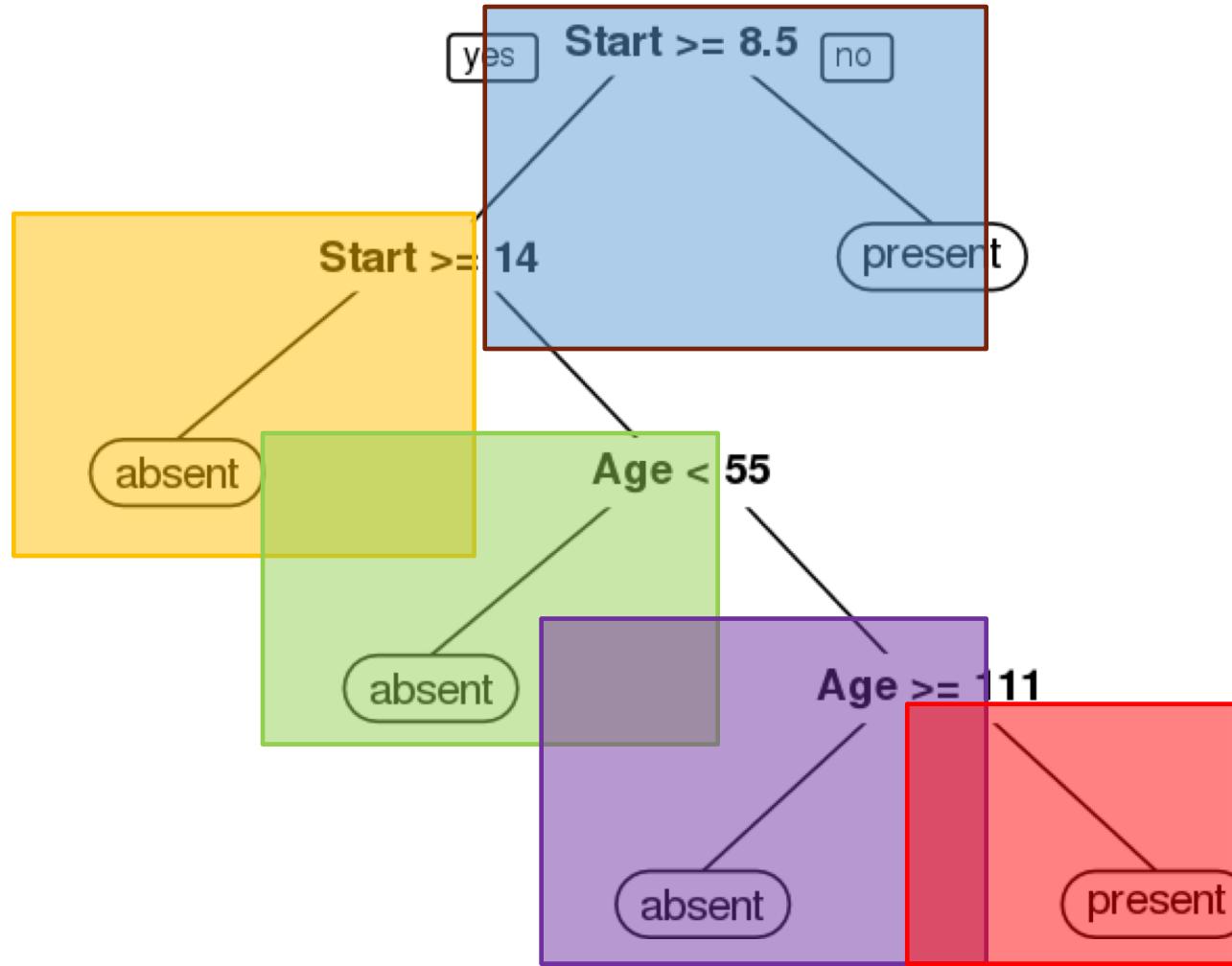
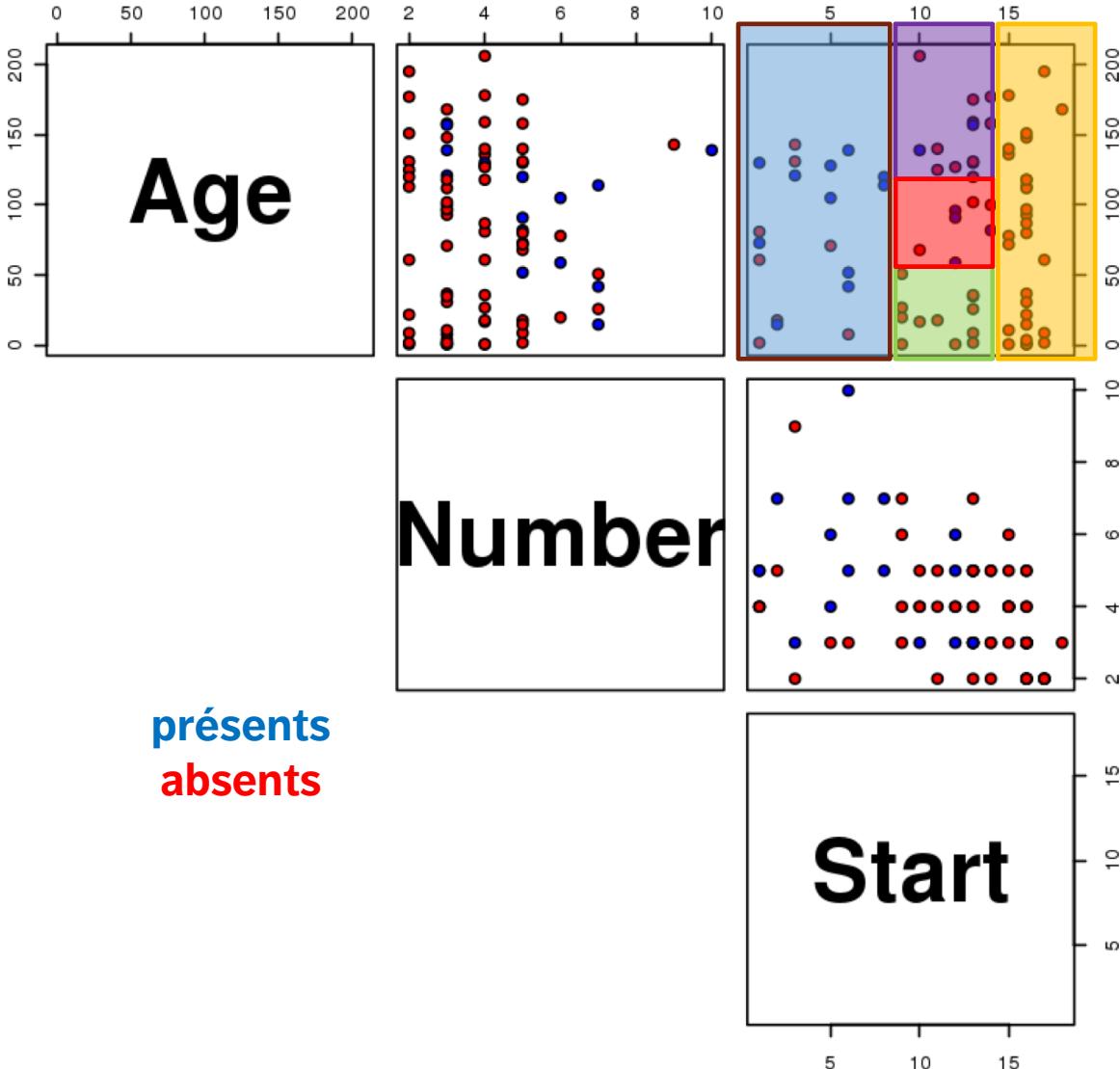


EXEMPLE – ENSEMBLE DE DONNÉES SUR LA CYPHOSE

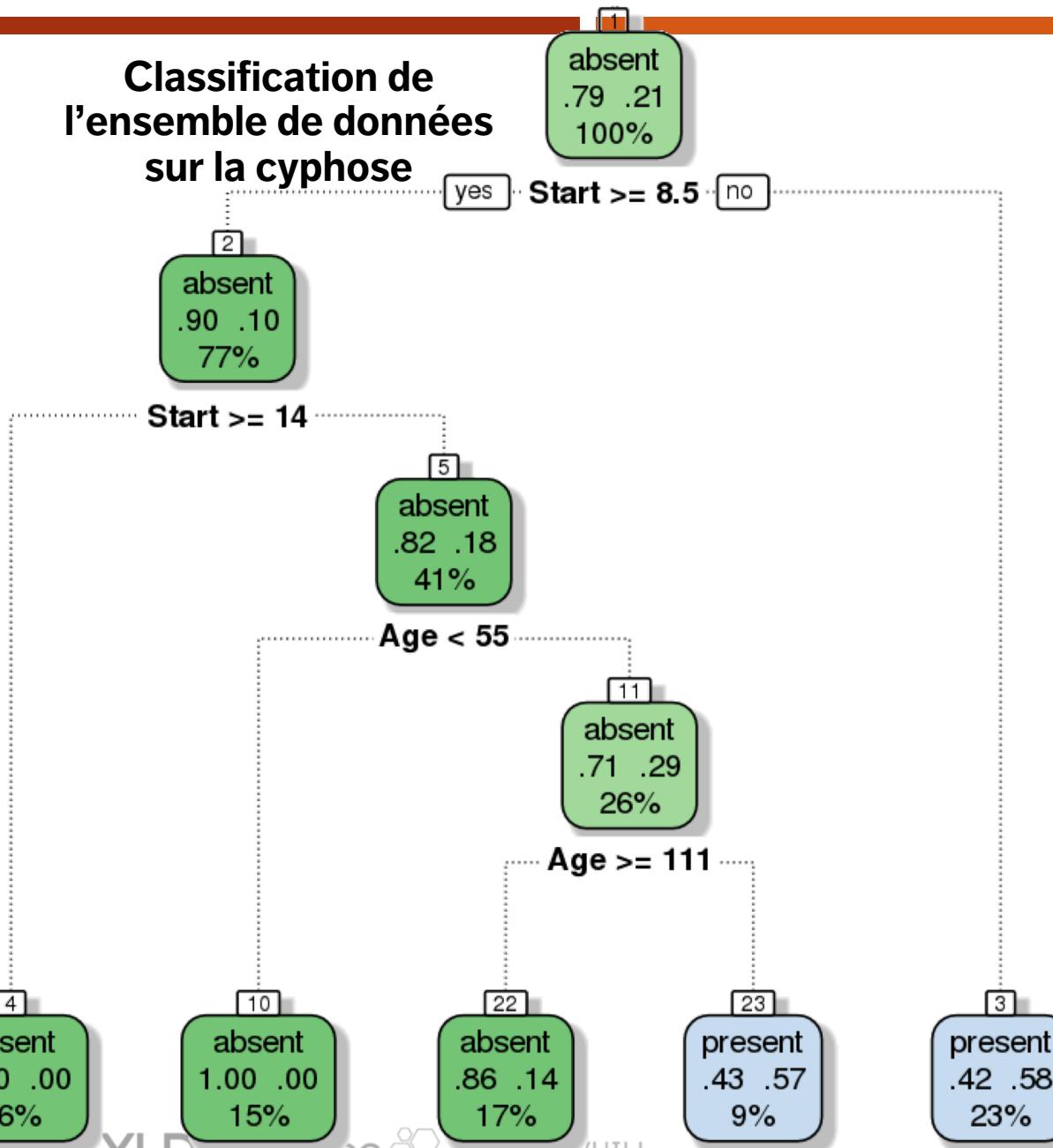
La question d'intérêt pour cet ensemble de données naturelles est de savoir comment les trois attributs explicatifs pourraient influer sur le succès de l'opération.

Nous utilisons l'implémentation rpart de CART pour générer des arbres de décision candidats.

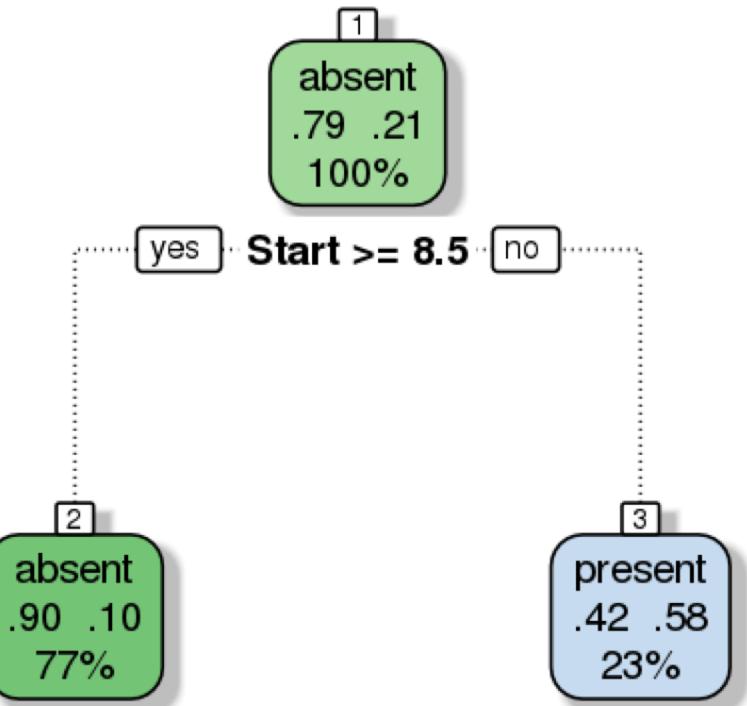
Strictement parlant, il ne s'agit pas d'une tâche supervisée prédictive puisque nous traitons l'ensemble des données comme un ensemble de formation (il n'y a pas d'observations d'essais à distance pour l'instant).



Classification de l'ensemble de données sur la cyphose



Classification élaguée de l'ensemble de données sur la cyphose



EXEMPLE - ENSEMBLE DE DONNÉES SUR LA CYPHOSE

Nous utilisons un modèle sur 50 observations (choisies au hasard) et évaluons le rendement sur les 31 autres observations.

		Predicted		Total	83.9%
		A	B		
Actuals	A	23	3	26	83.9%
	B	3	2	5	16.1%
Total		26	5	31	
		83.9%	16.1%		

Classification Rates		Performance Metrics	
	Sensitivity: 0.88		Accuracy: 0.81
	Specificity: 0.40		F1-Score: 0.88
	Precision: 0.88		Informedness (ROC): 0.28
	Negative Predictive Value: 0.40		Markedness: 0.28
	False Positive Rate: 0.60		M.C.C.: 0.28
	False Discovery Rate: 0.12		Pearson's chi2: 0.00
	False Negative Rate: 0.12		Hist. Stat: 0.00

ÉVALUATION DU RENDEMENT

Dans le problème de l'évaluation catégorique et numérique, une mesure de rendement isolée ne justifie pas suffisamment la validation du modèle, à moins qu'il n'ait d'abord été normalisé.

Il y a (beaucoup) plus à dire sur le thème du choix du modèle.