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)

OBJECTIFS D'APPRENTISSAGE

Se familiariser avec les concepts fondamentaux associés à la classification et à l'estimation des valeurs, et avec certains algorithmes communémment utilisés.

Se familiariser avec un algorithme de croissance spécifique pour les arbres de décision.

Se familiariser avec les critères d'évaluation de la performance en ce qui à traît à la classification et à l'estimation des valeurs.

CONTENU

1. Étude de cas: Vérification fiscale du Minnesota
2. Principes de base de la classification
3. Algorithmes de classification
4. Notes et validation
5. Exemple: Cyphose

ÉTUDE DE CAS : VÉRIFICATION FISCALE DU MINNESOTA

CLASSIFICATION ET ESTIMATION DE LA VALEUR

Sélection de la vérification fiscale basée sur l'exploration des données : une étude de cas d'un projet pilote du « Minnesota Department of Revenue »

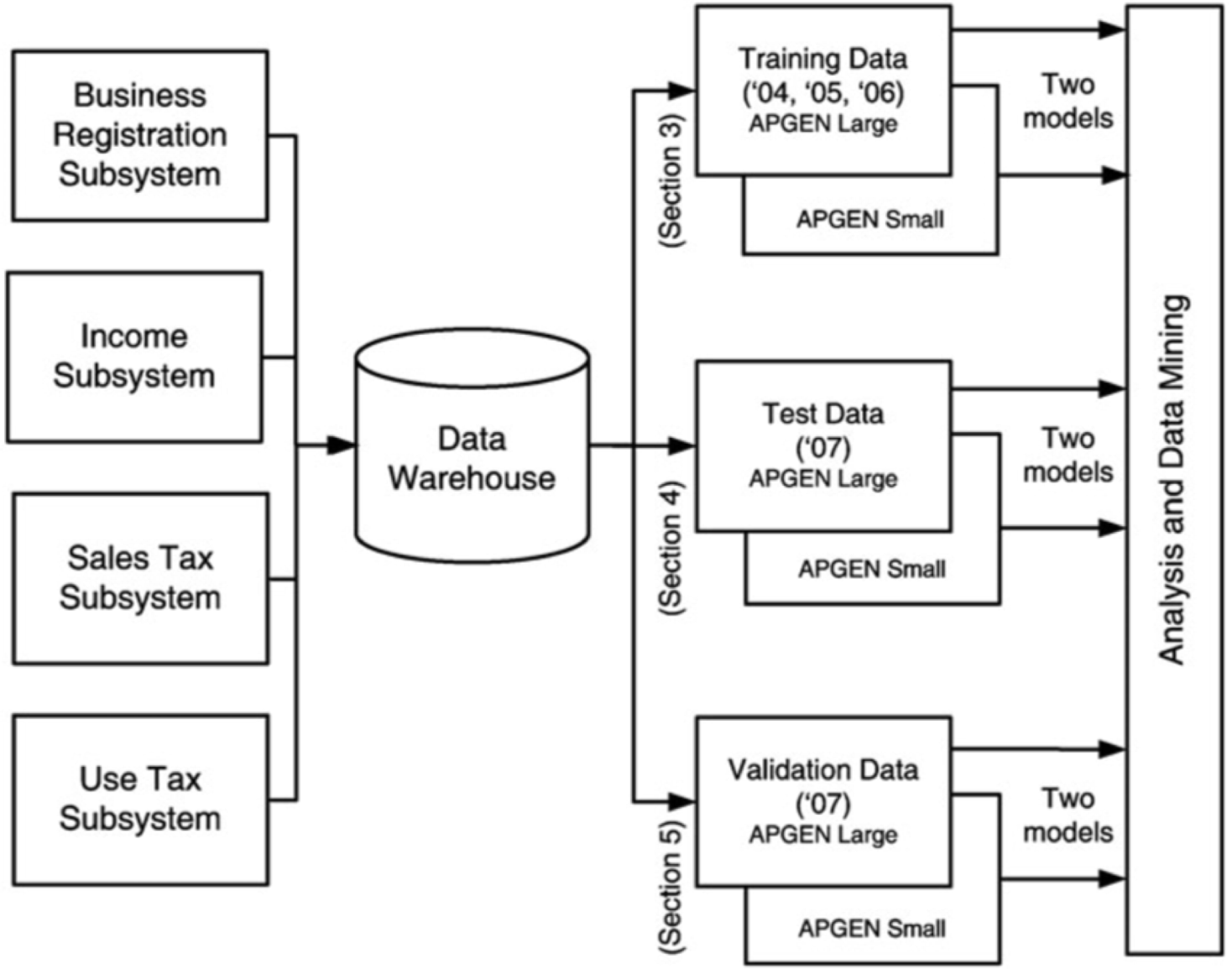
(Hsu, W., Pathak, N., Srivatsava, J., Tschida, Bjorklund, E. [2013], *Real Word Data Mining Applications*, Annals of Information Systems, v.17, Springer).

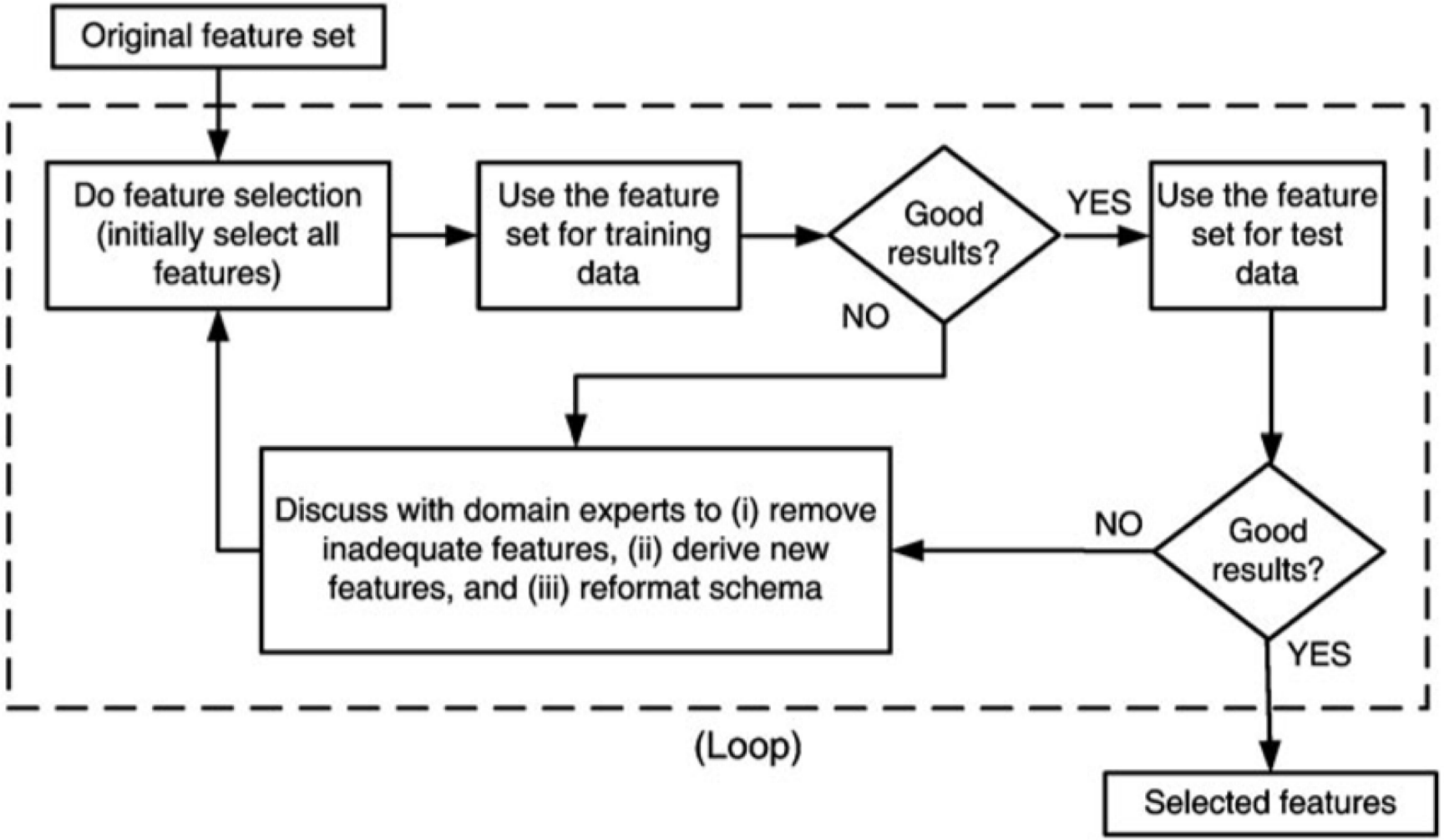
CONTEXTE

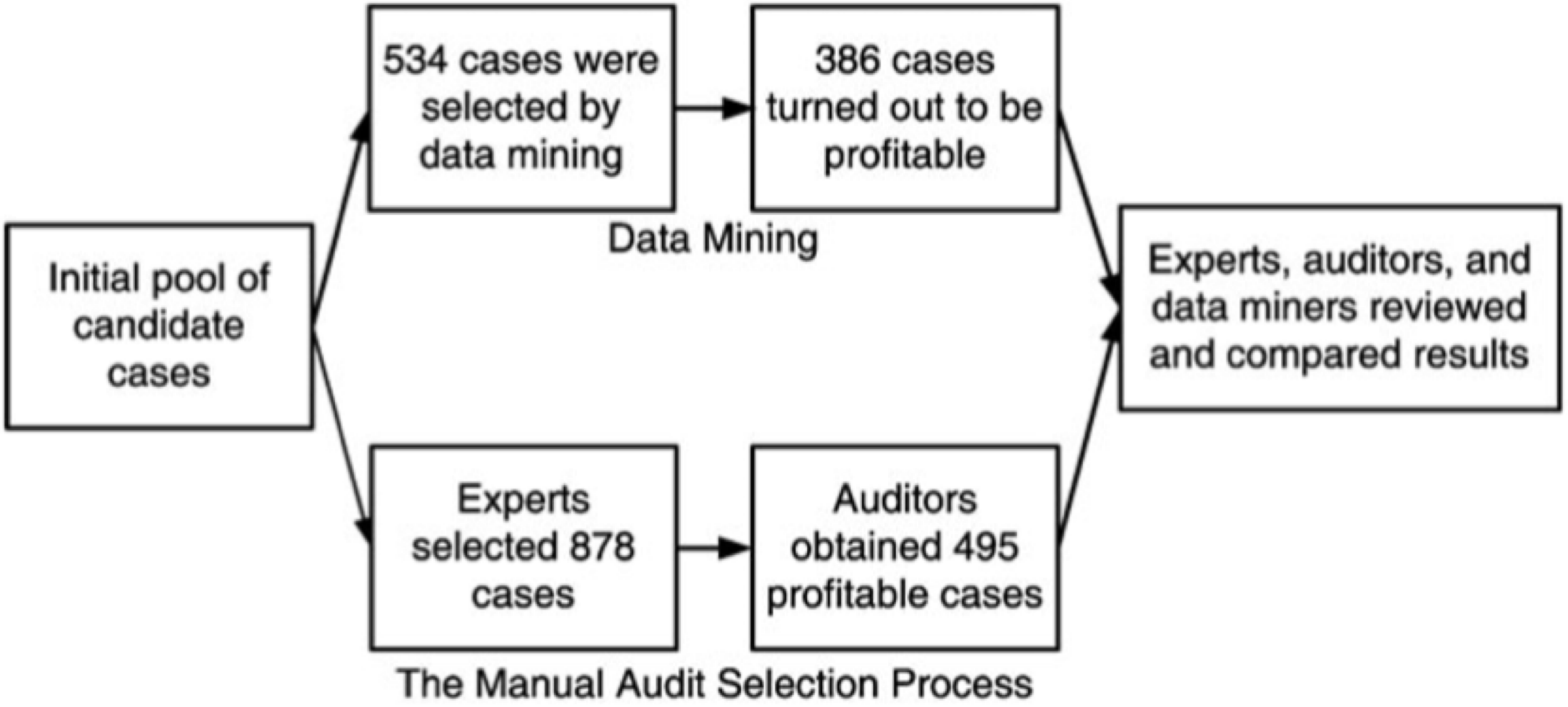
Les écarts importants entre les recettes dues (en théorie) et les recettes perçues (en pratique) sont problématiques pour les gouvernements.

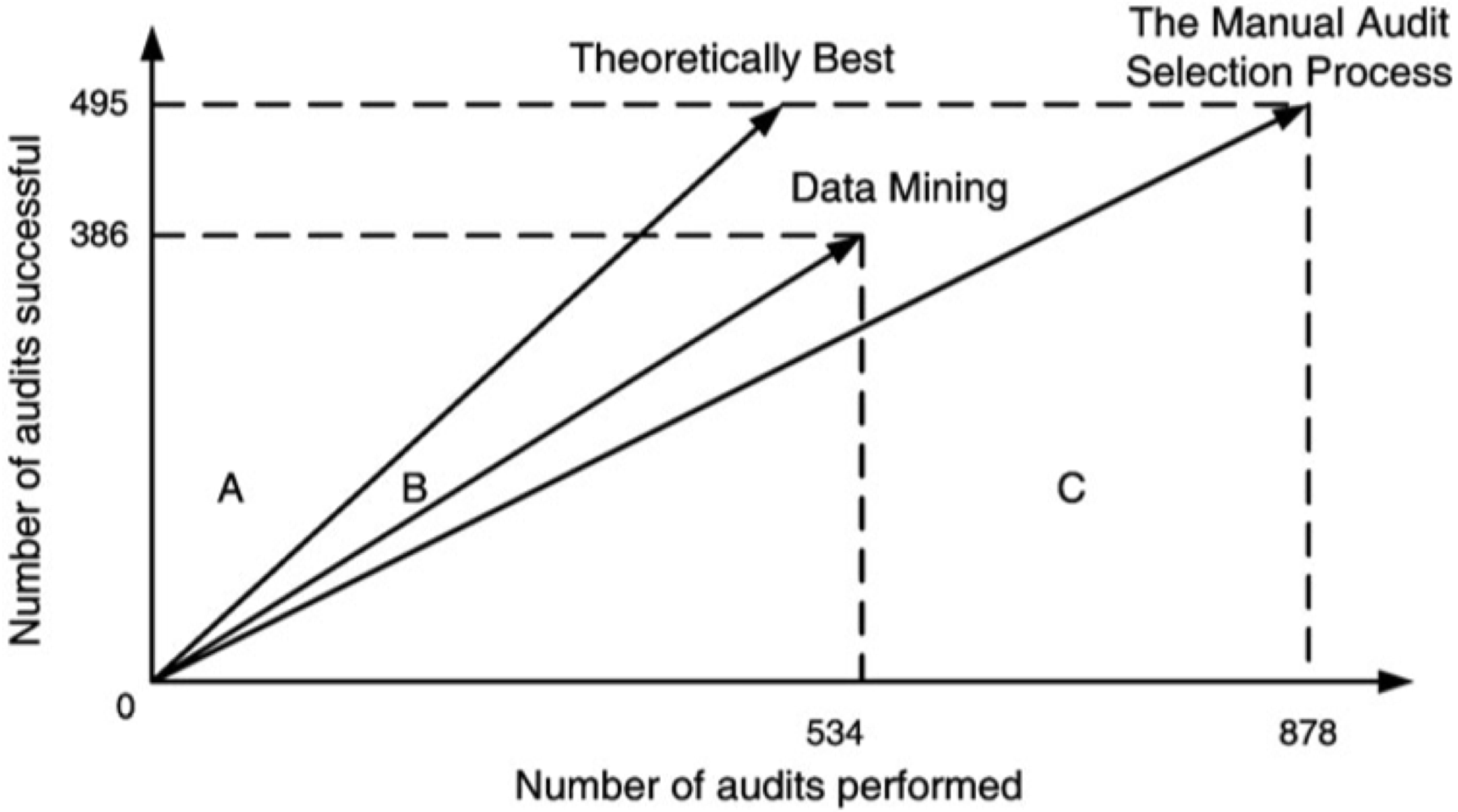
Les agences du revenu mettent en œuvre diverses stratégies de détection de la fraude (comme les examens de vérification) pour combler cette lacune.

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?**


Business Registration Subsystem
Income Subsystem
Sales Tax Subsystem
Use Tax Subsystem
Data Warehouse
(Section 3)
(Section 4)
(Section 5)
Training Data ('04, '05, '06) APGEN Large
APGEN Small
Test Data ('07) APGEN Large
APGEN Small
Validation Data ('07) APGEN Large
APGEN Small
Two models
Two models
Two models
Analysis and Data Mining


Original feature set
Do feature selection (initially select all features)
Use the feature set for training data
Good results?
YES
NO
Use the feature set for test data
Discuss with domain experts to (i) remove inadequate features, (ii) derive new features, and (iii) reformat schema
NO
Good results?
YES
(Loop)
Selected features


534 cases were selected by data mining
386 cases turned out to be profitable
Data Mining
Initial pool of candidate cases
Experts, auditors, and data miners reviewed and compared results
Experts selected 878 cases
Auditors obtained 495 profitable cases
The Manual Audit Selection Process


Theoretically Best
The Manual Audit Selection Process
Data Mining
495
386
Number of audits successful
A
B
C
0
534
878
Number of audits performed

	Predicted as good	Predicted as bad
Actually good	386 (Use tax collected) R = $5,577,431 (83.6 %) C = $177,560 (44 %)	109 (Use tax lost) R = $925,293 (13.9 %) C = $50,140 (12.4 %)
Actually bad	148 (costs wasted) R = $72,744 (1.1 %) C = $68,080 (16.9 %)	235 (costs saved) R = $98,105 (1.4 %) C = $108,100 (26.7 %)

DISCUSSION

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

PRINCIPES DE BASE DE LA CLASSIFICATION

CLASSIFICATION ET ESTIMATION DE LA VALEUR

« De nos jours, nous sommes beaucoup trop enclins à diviser les gens en catégories permanentes, oubliant qu'une catégorie n'existe que pour son but particulier et qu'elle doit être oubliée dès que ce but est atteint. »

(Dorothy L. Sayers, Are Women Human? Astute and Witty Essays on the Role of Women in Society)

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

AUTRES UTILISATIONS

Prédire qu'un objet appartient à une classe particulière.

Organiser et regrouper les instances en catégories.

Améliorer la détection d'objets pertinents

- évitement : « cet objet représente un véhicule se déplaçant dans notre direction »
- suivi : « il est peu probable que cet emprunteur manque à ses engagements hypothécaires »
- degré : « ce chien a 90 % de chances de vivre jusqu'à l'âge de 7 ans »

En l'absence de données d'essai, la classification peut être **descriptive**, mais pas prédictive.

EXEMPLES

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?

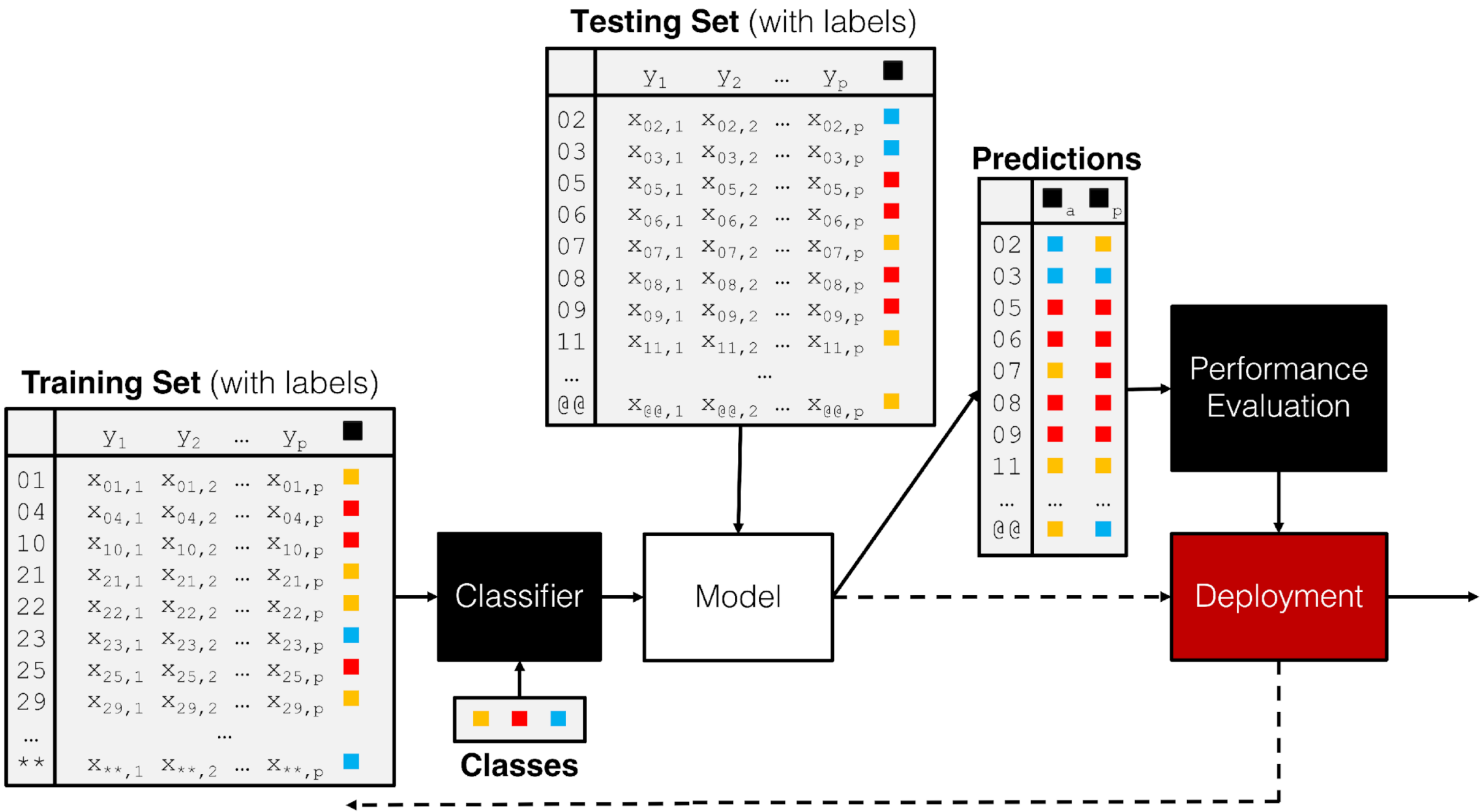
EXEMPLES

Scénario :

Il a été déterminé que les clients qui font un grand nombre d'appels au numéro de service à la clientèle d'une compagnie de téléphonie mobile sont susceptibles d'annuler leur abonnement. L'entreprise désire réduire le taux d'annulation de l'abonnement.

Questions : peut-on prédire

- la valeur totale à vie d'un client?
- quels sont les clients les plus susceptibles d'annuler leur abonnement dans un avenir proche?
- quelle est l'offre de fidélisation à laquelle un client particulier répondra le mieux?


Testing Set (with labels)
Predictions
Training Set (with labels)
Performance Evaluation
Classifier
Model
Deployment
Classes

EXERCICE

Comment utiliseriez-vous les techniques de modélisation statistique standard pour répondre à ces questions?

ALGORITHMES DE CLASSIFICATION

CLASSIFICATION ET ESTIMATION DE LA VALEUR

« La diversité des problèmes pouvant être abordés par les algorithmes de classification est importante et couvre de nombreux domaines. [...] Il est difficile de discuter en détail de toutes les méthodes dans un seul livre. »

(C.C. Aggarwal, Data Classification: Algorithms and Applications)

SYSTÈMES DE CLASSIFICATION

Régression logistique

- modèle classique
- influencée par l'inflation de la variance et le processus de sélection des variables

Réseaux neuronaux

- difficile à comprendre
- demande que toutes les variables soient du même type
- plus facile à utiliser depuis la rétropropagation (règle de la chaîne)

Arbres de décision

- peut sur-ajuster les données si elles ne sont pas « élaguées » correctement (manuellement?)

SYSTÈMES DE CLASSIFICATION

Classificateur bayésien naïf

- assez performant pour les applications d'exploration de texte (filtre antipourriel)
- les hypothèses ne sont pas souvent rencontrées dans la pratique

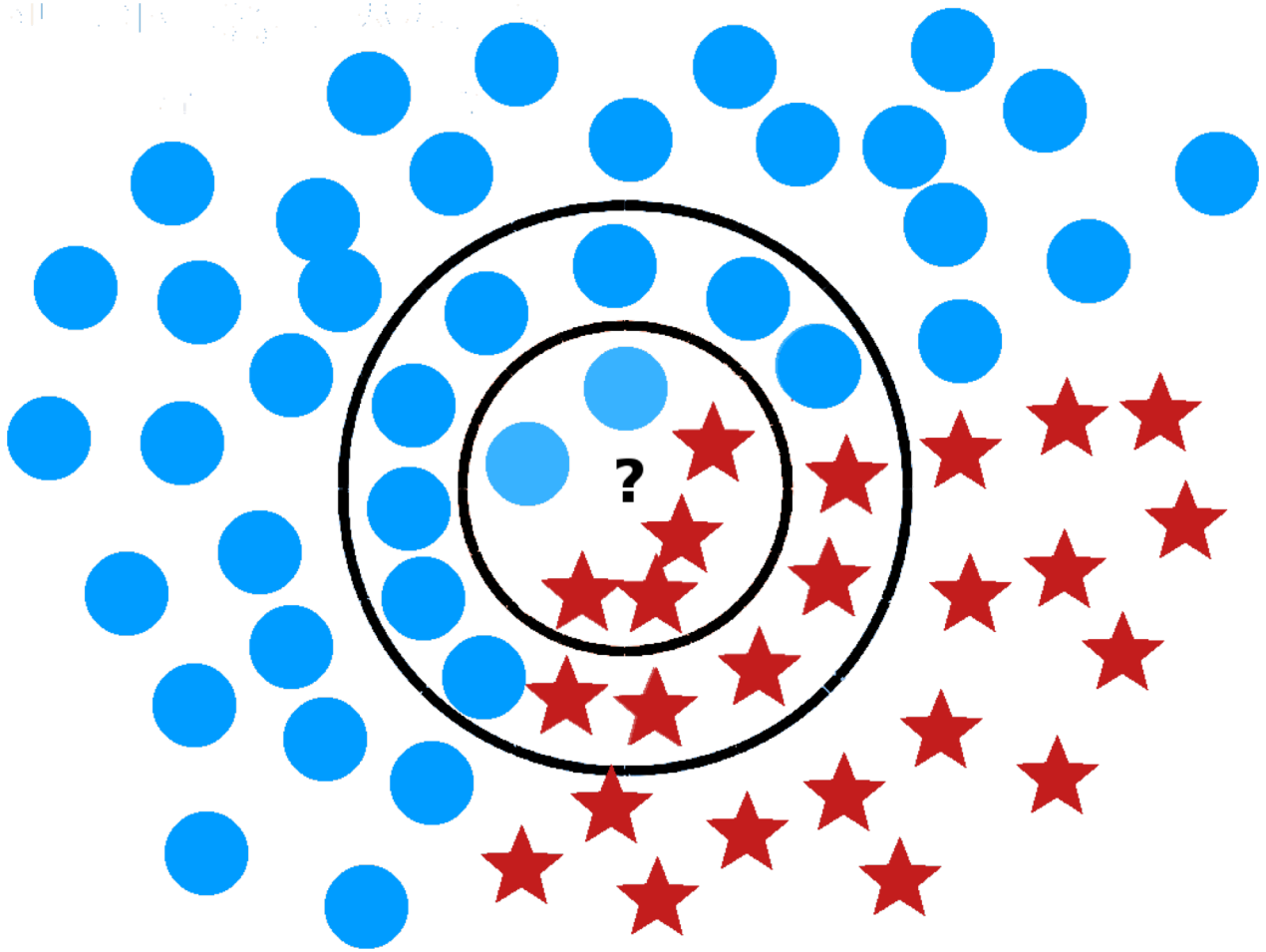
Machines à vecteurs de support

- peut être difficile à comprendre (limites non linéaires)
- peut aider à atténuer les difficultés liées aux mégadonnées

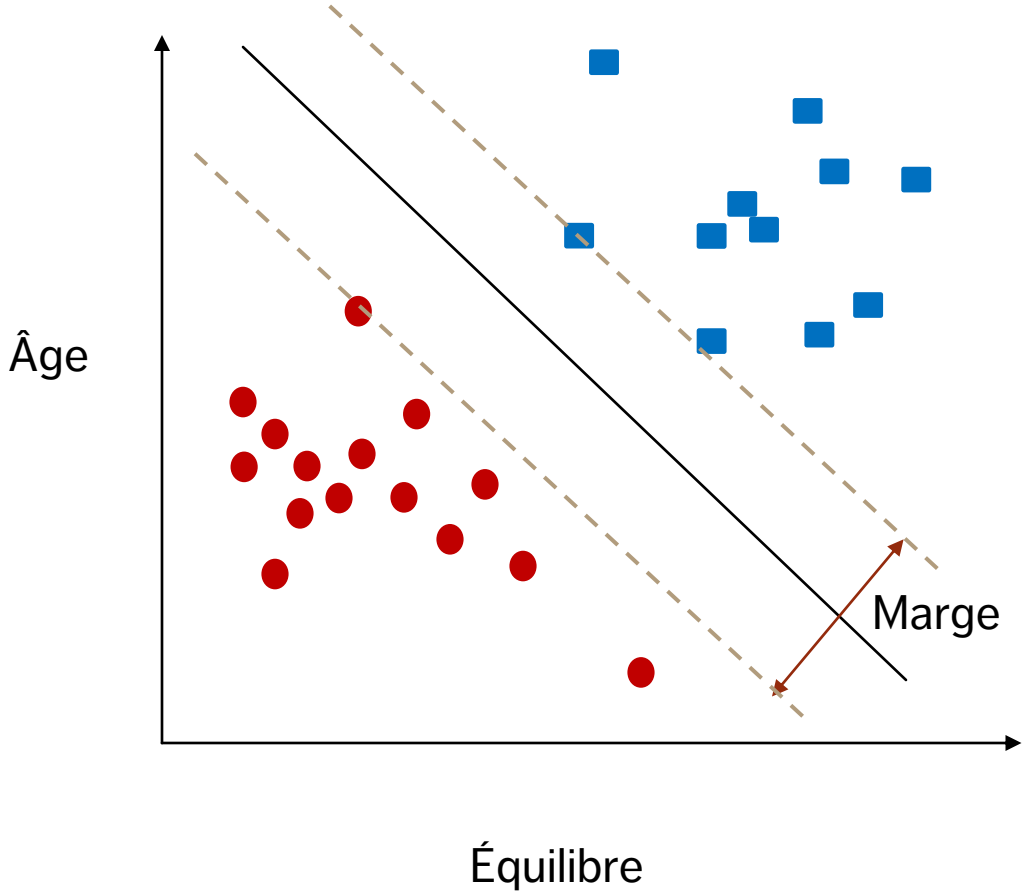
Classificateurs des voisins les plus proche

- nécessitent très peu d'hypothèses au sujet des données
- pas très stables (l'ajout de points peut modifier substantiellement la limite)

K VOISINS LES PLUS PROCHE

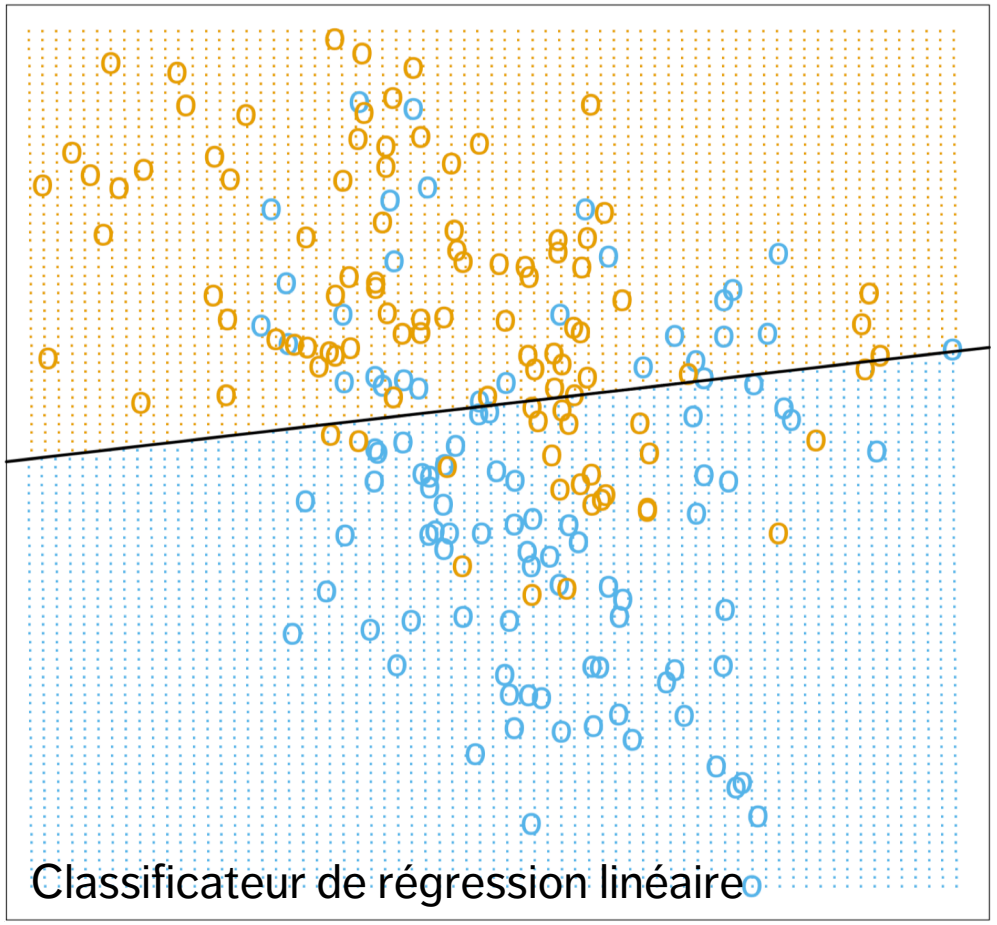


MACHINE À VECTEURS DE SUPPORT

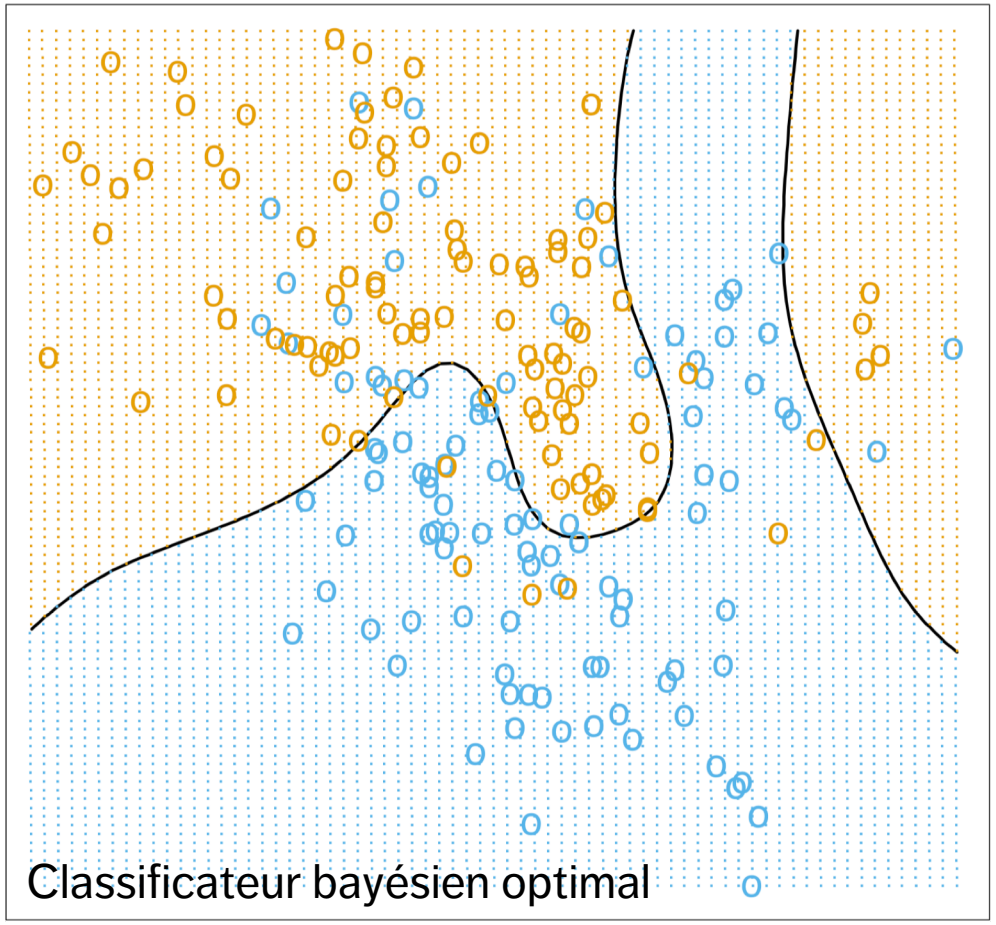


[Illustrations tirées de *The Elements of Statistical Learning*]

ILLUSTRATION



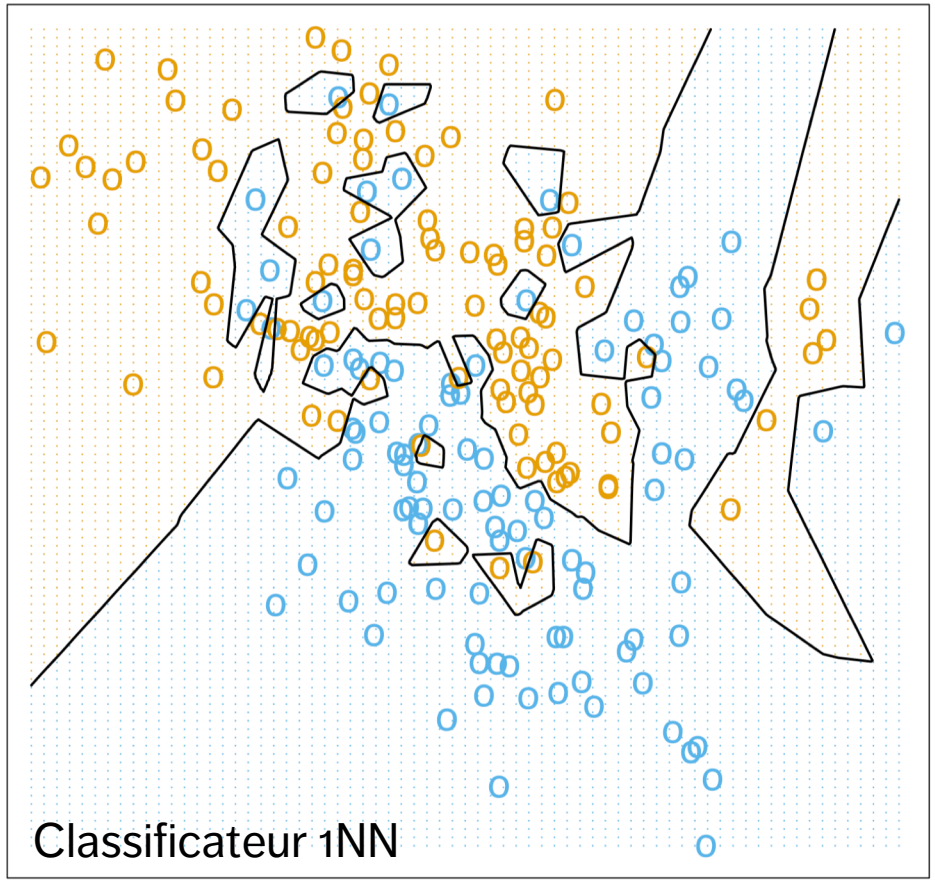
Classificateur de régression linéaire



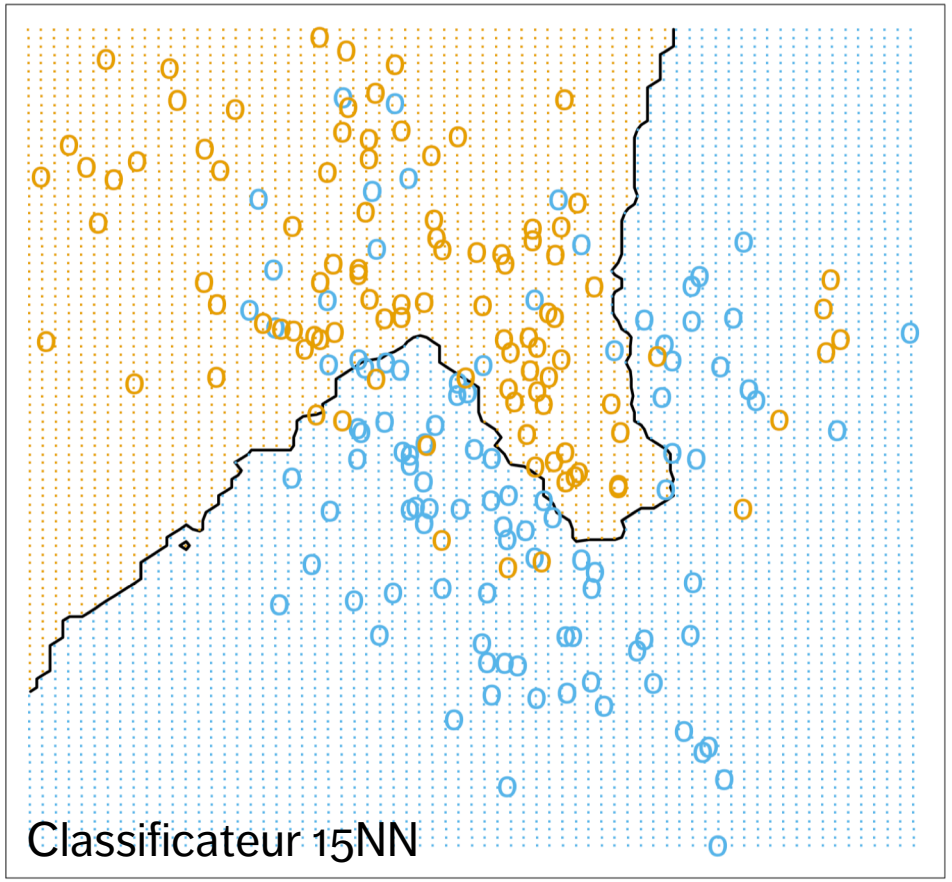
Classificateur bayésien optimal

[Illustrations tirées de *The Elements of Statistical Learning*]

ILLUSTRATION



Classificateur 1NN

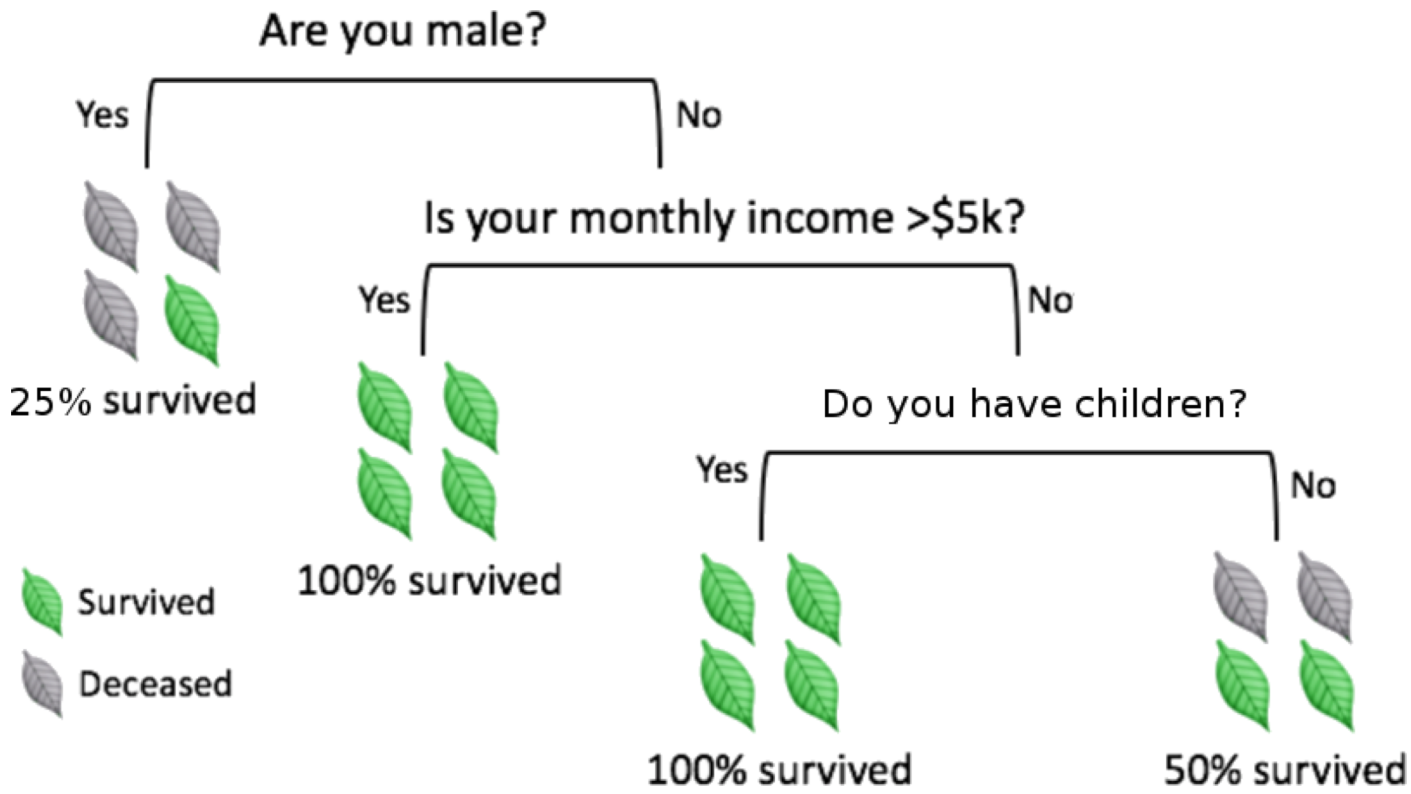


Classificateur 15NN

[adapté de A. Ng, K. Soo, Numsense!, Surviving a Disaster]

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 **prédiction** pour une nouvelle instance, suivez le tracé le long de l'arbre, lisant la prédiction 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 prédiction 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?)

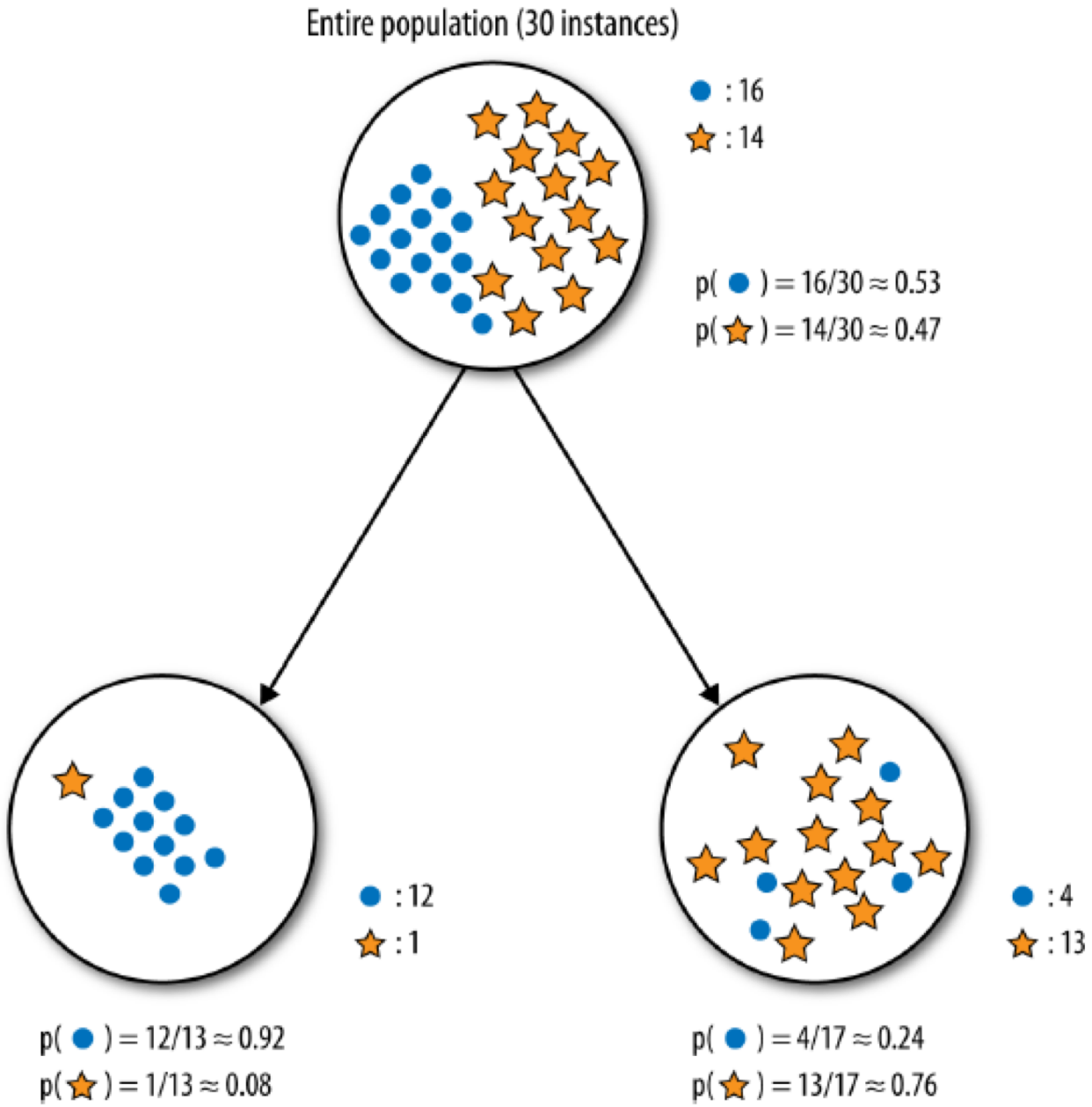
INFORMATION GAIN

L'**entropie** est une mesure du désordre de l'ensemble S. Soit p_i la proportion des observations de S appartenant à la catégorie i, pour $i = 1, \dots, n$. L'entropie de S est

$$E(S) = -p_1 \log p_1 - p_2 \log p_2 - \cdots - p_n \log n.$$

Si l'**ensemble parent** S contient m observations, et que ces dernières sont réparties en k **sous-ensemble enfants** $C_1, \dots, C_k$ contenant $q_1, \dots, q_k$ observations chacuns, le **gain d'information** produit par la répartition des observations est

$$\mathrm{IG}(S; C) = E(S) - \frac{1}{m}[q_1 E(C_1) + \cdots + q_k E(C_k)].$$



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

$$E(L) = -p_{\circ}\log p_{\circ} - p_{*}\log p_{*}$$

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

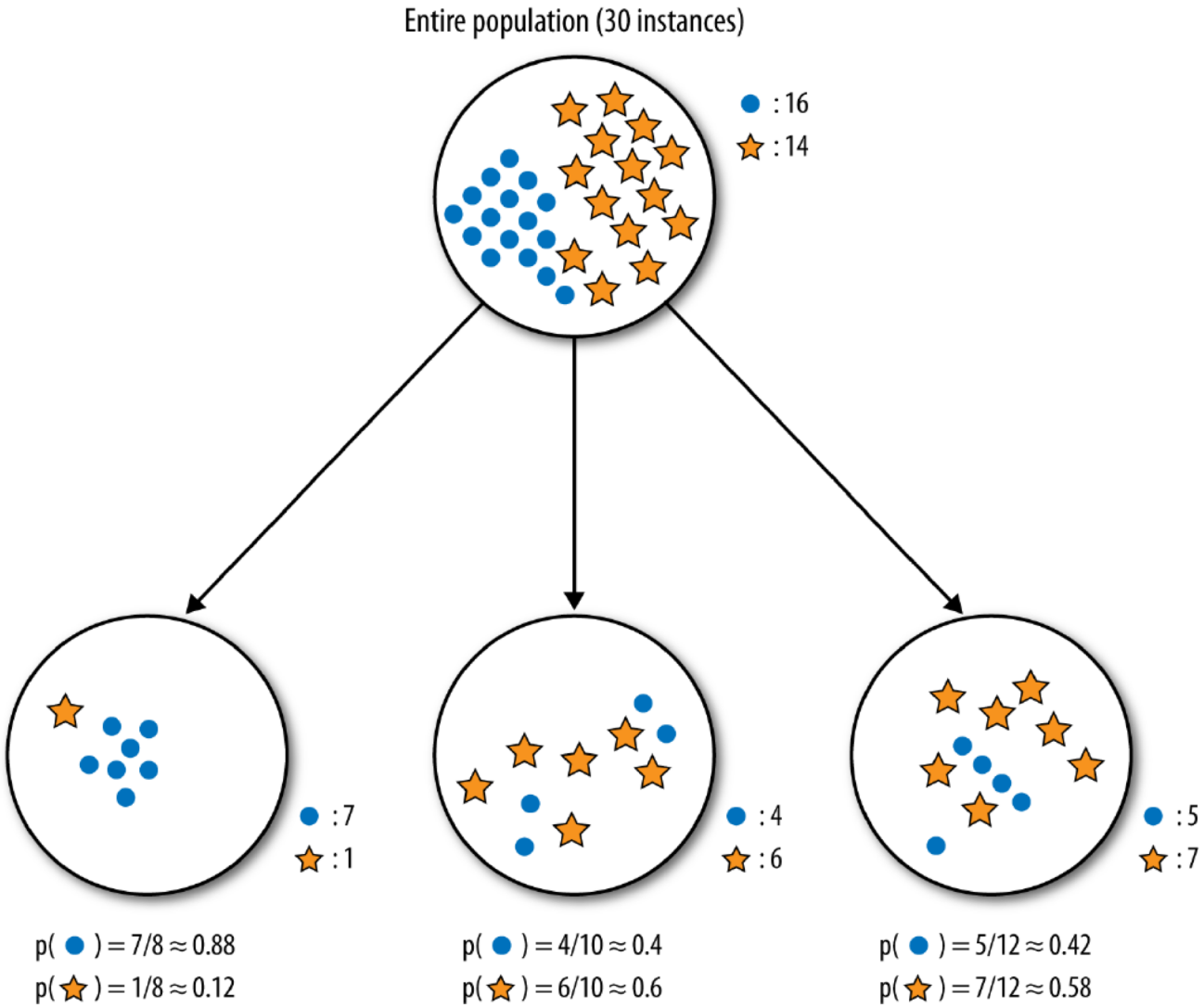
$$E(R) = -p_{\circ}\log p_{\circ} - p_{*}\log p_{*}$$

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

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

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

$$\approx \mathbf{0.37}$$



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

$$= -\frac{7}{8}\log\frac{7}{8} - \frac{1}{8}\log\frac{1}{8} \approx 0.54$$

$$E(C) = -p_\circ \log p_\circ - p_* \log p_*$$

$$= -\frac{4}{10}\log\frac{4}{10} - \frac{6}{10}\log\frac{6}{10} \approx 0.97$$

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

$$= -\frac{5}{12}\log\frac{5}{12} - \frac{7}{12}\log\frac{7}{12} \approx 0.98$$

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

$$\approx 0.99 - \frac{1}{30}[8(0.54) + 10(0.97) + 12(0.98)]$$

$$\approx \mathbf{0.13}$$

DISCUSSION

Quelle est la répartition la plus avantageuse?

En quoi le choix des algorithmes dépend-il des données et des types de données disponibles, et de l'objectif de la prévision?

EXERCICE

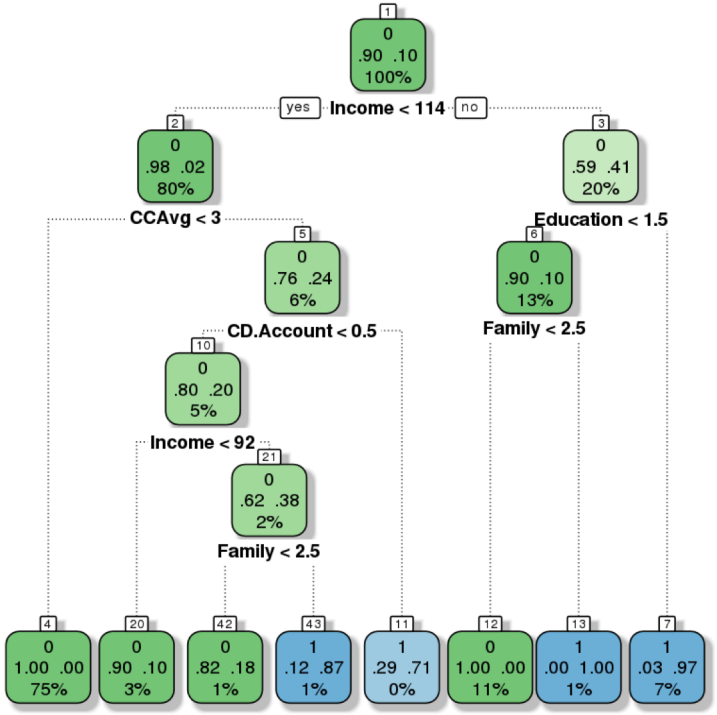
UniversalBank is looking at converting its **liability** customers (i.e., customers who only have deposits at the bank) into **asset** customers (i.e., customers who have a loan with the bank). In a previous campaign, *UniversalBank* was able to convert 9.6% of 5000 of its liability customers into asset customers. The marketing department would like to understand what combination of factors make a customer more likely to accept a personal loan, in order to better design the next conversion campaign.

UniversalBank's dataset contains data on 5000 customers, including the following measurements: age, years of professional experience, yearly income (in thousands of USD), family size, value of mortgage with the bank, whether the client has a certificate of deposit with the bank, a credit card, etc.

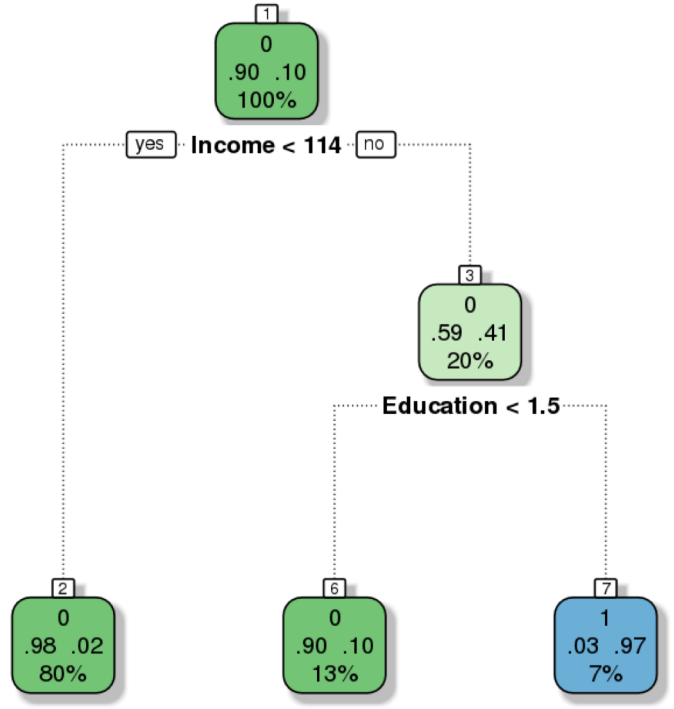
EXERCICE

We build 2 decision trees on a training subset of 3000 records to predict whether a customer is likely to accept a personal loan (1) or not (0).

Tree *A*



Tree *B*



EXERCICE

Explore the `UniversalBank.csv` dataset. Can you come up with a reasonable guess as to what each of the variables represent?

1. How many variables are used in the construction of tree A? Of tree B?
2. Is the following decision rule valid or not for tree A:
   IF (Income ≥ 114) AND (Education ≥ 1.5)
   THEN (Personal Loan $= 1$)?
3. Is the following decision rule valid or not for tree B:
   IF (Income < 92) AND (CCAvg ≥ 3) AND (CD.Account < 0.5)
   THEN (Personal Loan $= 0$)?

EXERCICE

4. What prediction would tree A make for a customer with:
   - yearly income of 94,000$USD (Income = 94),
   - 2 kids (Family = 4),
   - no certificate of deposit with the bank (CD.Account = 0),
   - a credit card interest rate of 3.2% (CCAvg = 3.2), and
   - a graduate degree in Engineering (Education = 3).
5. What about tree B?

NOTES ET VALIDATION

CLASSIFICATION ET ESTIMATION DE LA VALEUR

« Des arbres, des forêts et des jungles au hasard, oh mon ... !!! »

(inconnu)

POINTS FORTS DES ARBRES DE DÉCISION

Modèle « **boîte blanche** »

- les prédictions peuvent toujours être expliquées en suivant les tracés appropriés

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

Sélection des variables **intégrée**

- les fonctions moins pertinentes n'ont pas tendance à être utilisées comme fonctions de répartition

Ne fait **aucune hypothèse** concernant

- l'indépendance, la variance constante, les distributions sous-jacentes, la colinéarité

CONTRAINTES DES ARBRES DE DÉCISION

Pas aussi précis que les autres algorithmes (habituellement)

Pas robuste : de petits changements dans l'ensemble de données de formation peuvent conduire à un arbre complètement différent, avec des prédictions complètement différentes

Particulièrement vulnérable au **sur-ajustement** en l'absence d'un « élagage »

- les procédures « d'élagage » sont typiquement alambiquées

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.

ÉVALUATION DU RENDEMENT

sensitivity = $TP/(TP + FN)$

specificity = $TN/(FP + TN)$

precision = $TP/(TP + FP)$

recall = $TP/(TP + FN)$

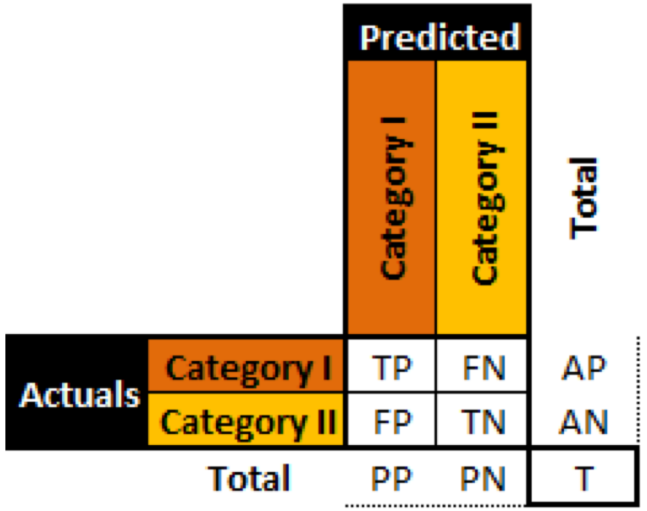
negative predictive value = $TN/(TN + FN)$

false positive rate = $FP/(FP + TN)$

false discovery rate = $FP/(FP + TP)$

false negative rate = $FP/(FN + TP)$

accuracy = $(TP + TN)/T$



Other metrics:

F_1-score, ROC AUC, informedness, markedness, Matthews' Correlation Coefficient (MCC), etc.

		Predicted			Total
		A	B		
Actuals	A	54	10	64	79.0%
	B	6	11	17	21.0%
Total		60	21	81	
		74.1%	25.9%		

Classification Rates	
Sensitivity:	0.84
Specificity:	0.65
Precision:	0.90
Negative Predictive Value:	0.52
False Positive Rate:	0.35
False Discovery Rate:	0.10
False Negative Rate:	0.16

Performance Metrics	
Accuracy:	0.80
F1-Score:	0.87
Informedness (ROC):	0.49
Markedness:	0.42
M.C.C.:	0.46
Pearson's chi2:	0.01
Hist. Stat:	0.10

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

Classification Rates	
Sensitivity:	1.00
Specificity:	0.41
Precision:	0.77
Negative Predictive Value:	1.00
False Positive Rate:	0.59
False Discovery Rate:	0.23
False Negative Rate:	0.00

Performance Metrics	
Accuracy:	0.80
F1-Score:	0.87
Informedness (ROC):	0.41
Markedness:	0.77
M.C.C.:	0.56
Pearson's chi2:	0.33
Hist. Stat:	0.40

DISCUSSION

De façon générale, combien de modèles prédictifs devraient être construits pour faire face aux tâches supervisées?

EXERCICE

(*UniversalBank*, continued)

The confusion matrices for the predictions of trees A and B on the remaining 2000 testing observations are shown here.

Tree A

		Predicted			
		A	B	*Total*	
Actuals	A	1792	19	1811	90.55%
	B	18	171	189	9.45%
Total		1810	190	2000	
		90.50%	9.50%		

Tree B

		Predicted			
		A	B	*Total*	
Actuals	A	1801	10	1811	90.55%
	B	64	125	189	9.45%
Total		1865	135	2000	
		93.25%	6.75%		

EXERCICE

6. Using the appropriate matrices, compute the 9 performance evaluation metrics for each of the trees (on the testing set).
7. If customers who would not accept a personal loan get irritated when offered a personal loan, what tree should *UniversalBank*'s marketing group use to help maintain good customer relations?

EXEMPLE : CYPHOSE

CLASSIFICATION ET ESTIMATION DE LA VALEUR

« Mon malheur est que je ressemble encore trop à un homme.
J'aimerais être une bête comme cette chèvre. »

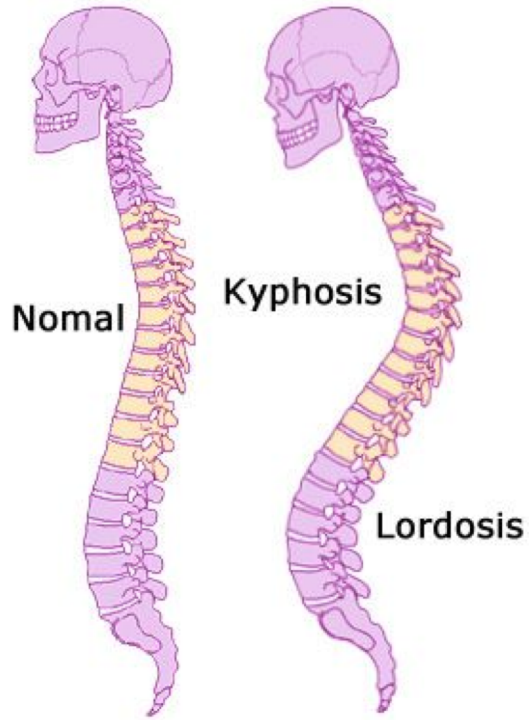
(V. Hugo, *Le bossu de Notre-Dame*)

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 :

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

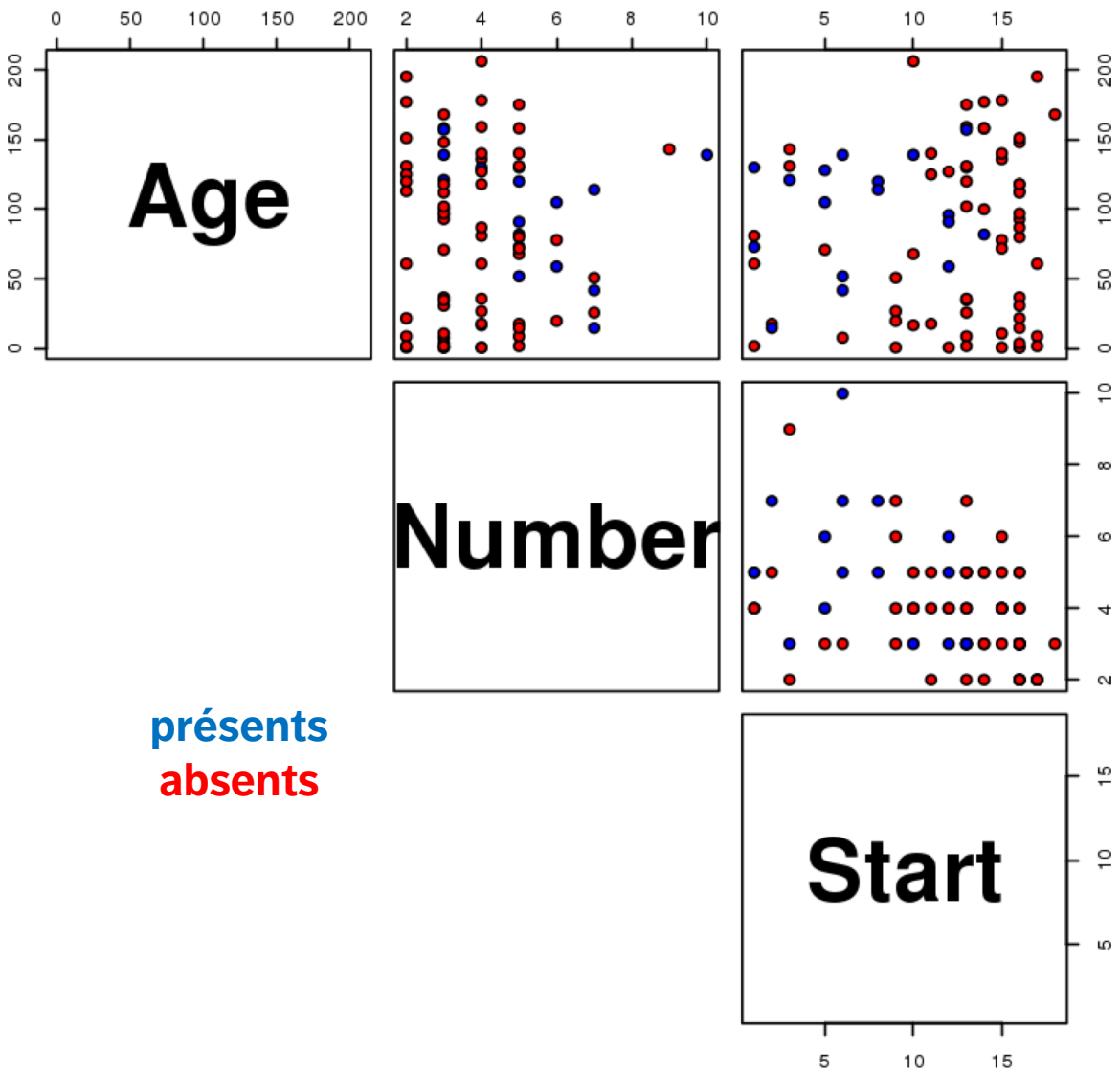


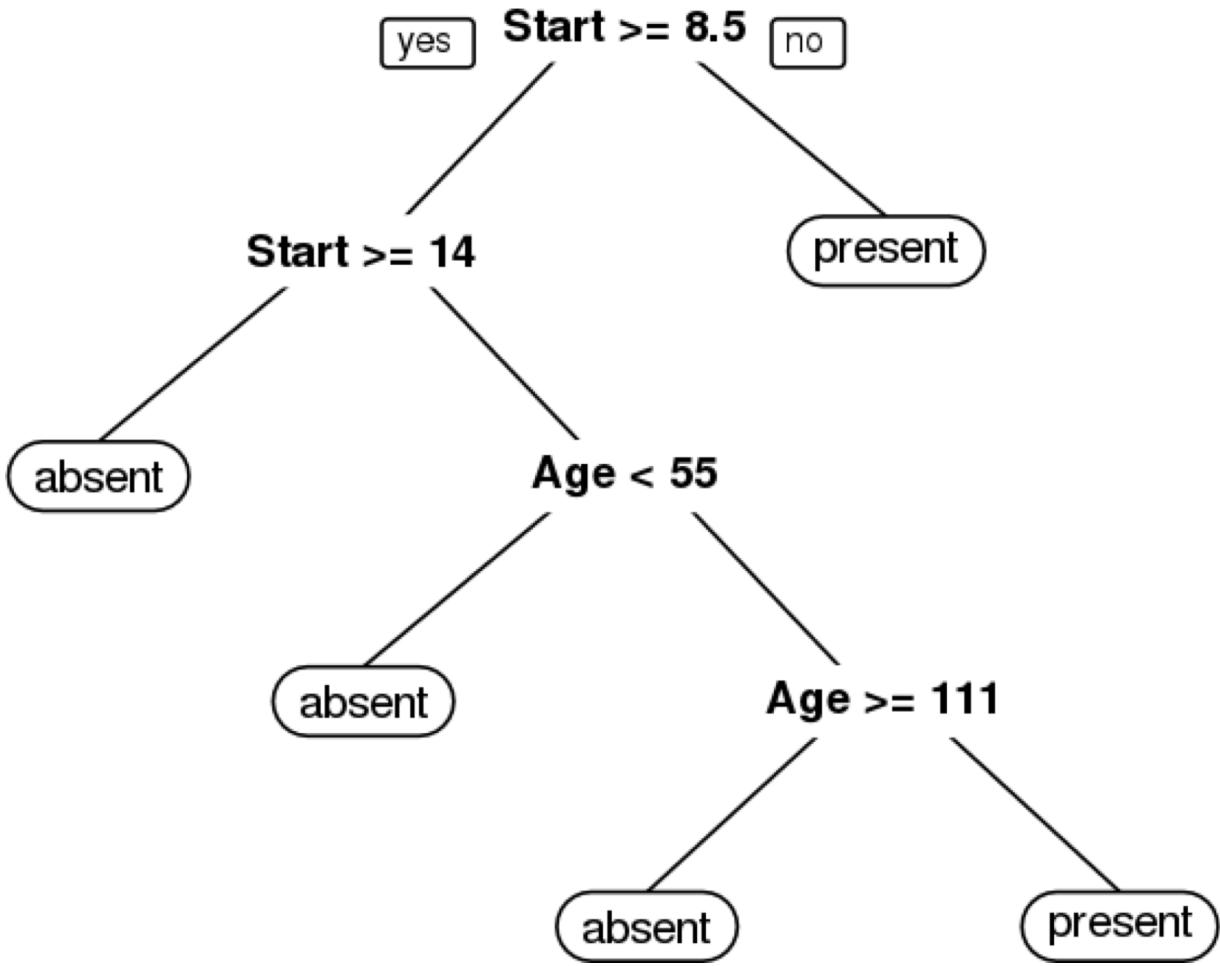
EXEMPLE – ENSEMBLE DE DONNÉES SUR LA CYPHOSE

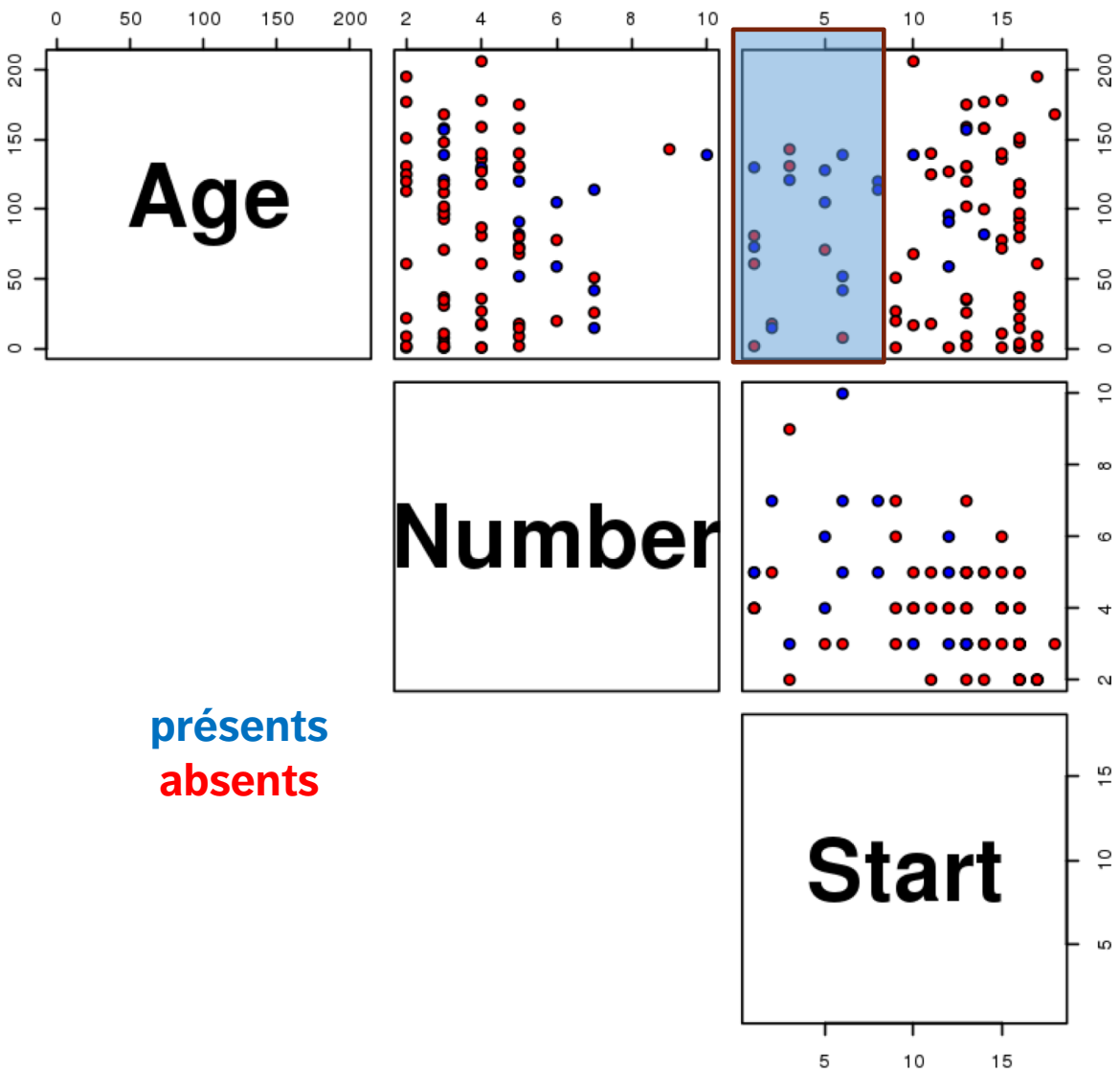
La question d'intérêt pour cet ensemble de données naturelles est de savoir comment les trois attributs explicatifs pourraient influer sur le succès de l'opération.

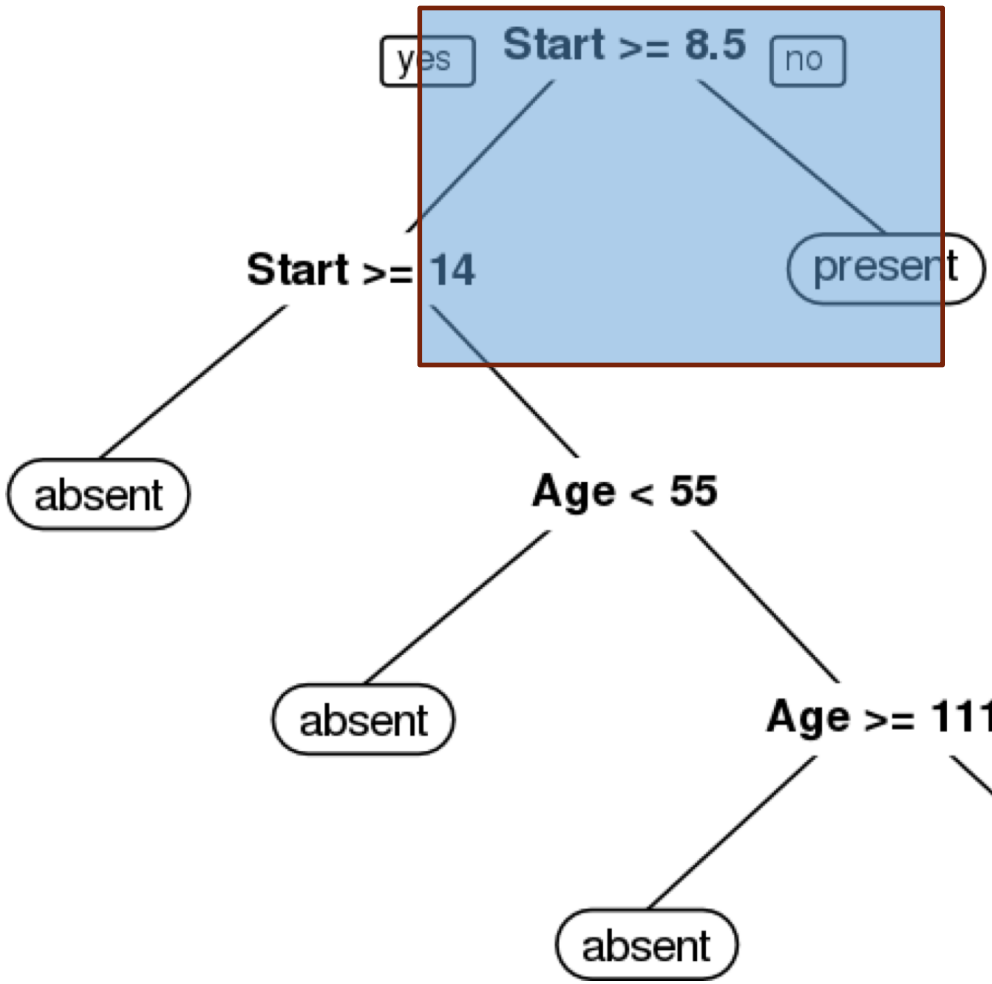
Nous utilisons l'implémentation rpart de CART pour générer des arbres de décision candidats.

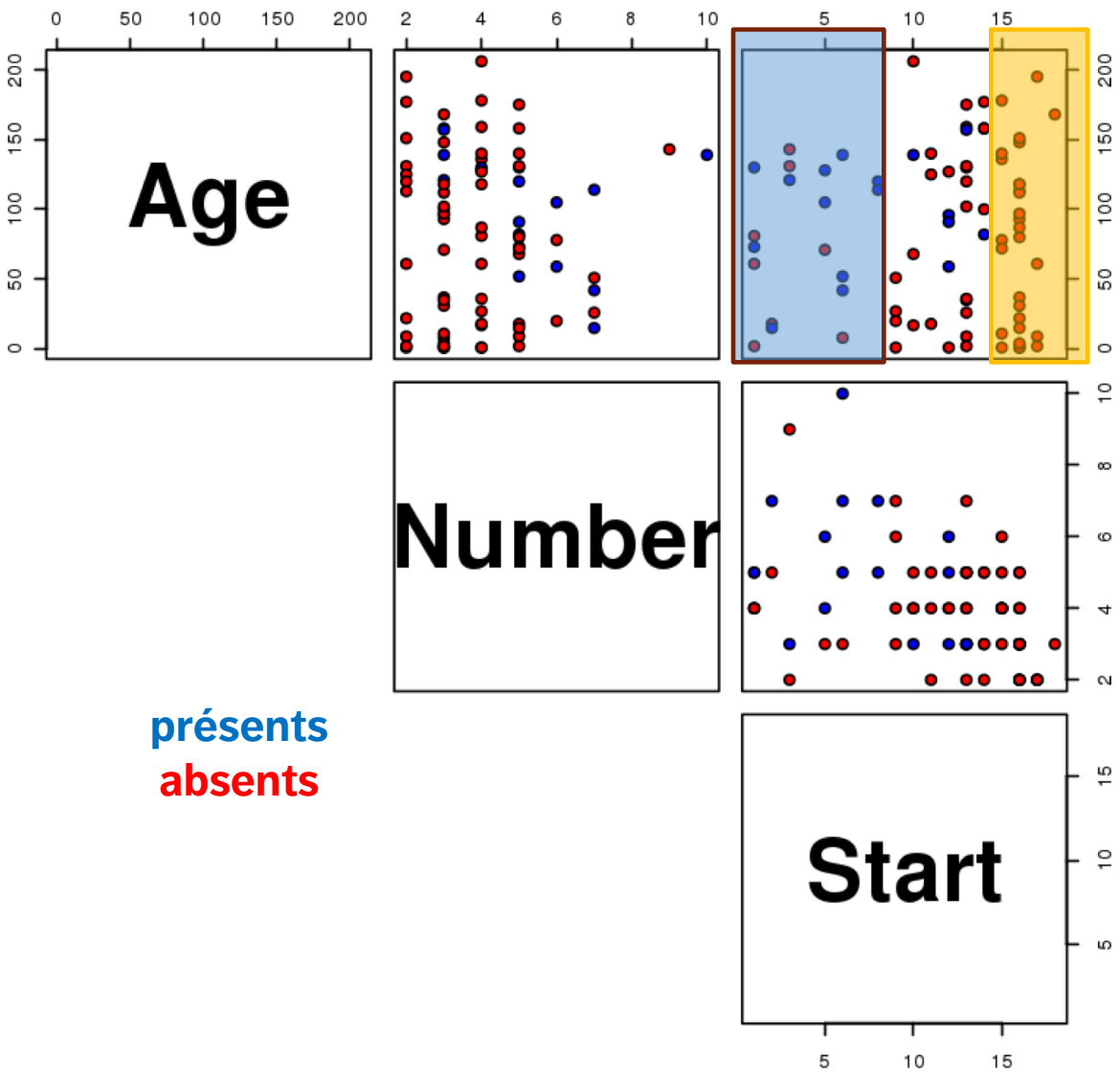
Strictement parlant, il ne s'agit pas d'une tâche supervisée prédictive puisque nous traitons l'ensemble des données comme un ensemble de formation (il n'y a pas d'observations d'essais à distance pour l'instant).

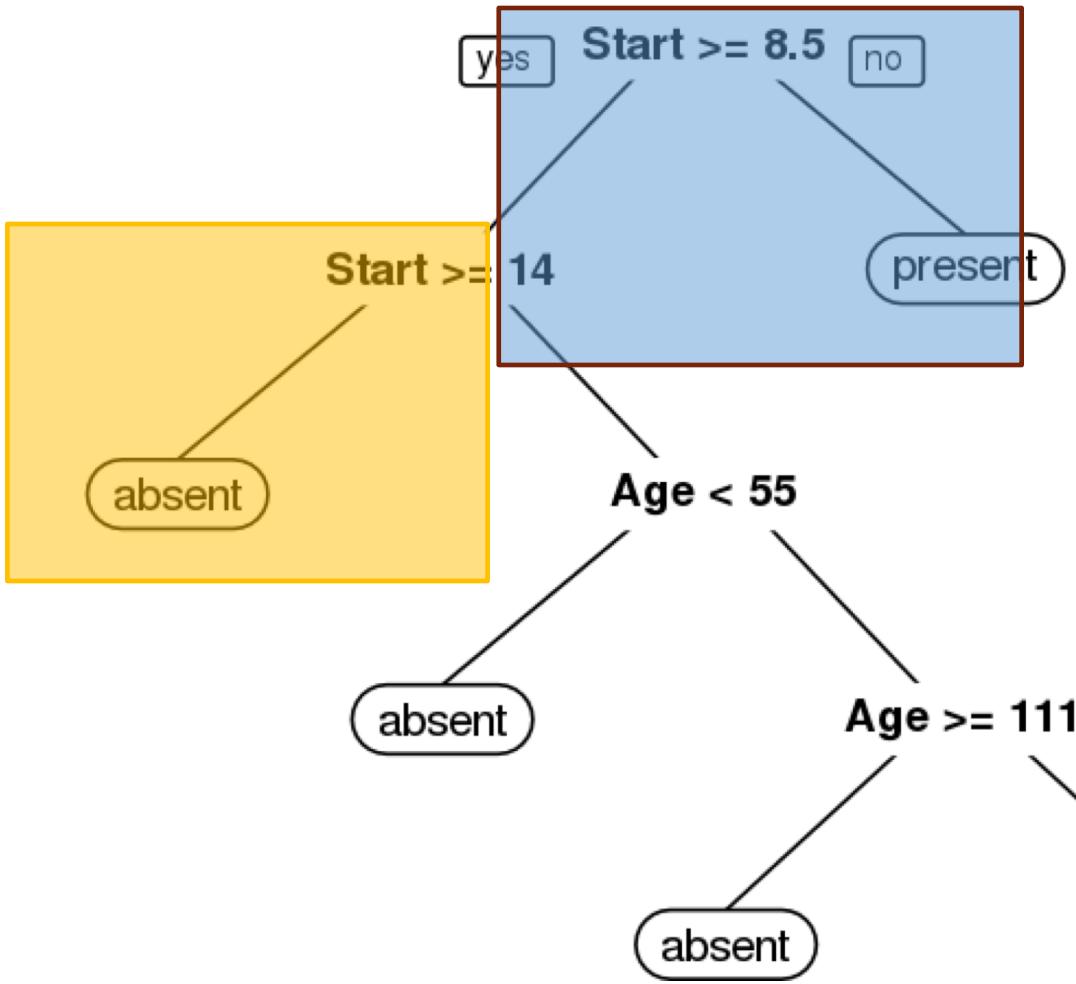

Age
Number
Start
présents
absents

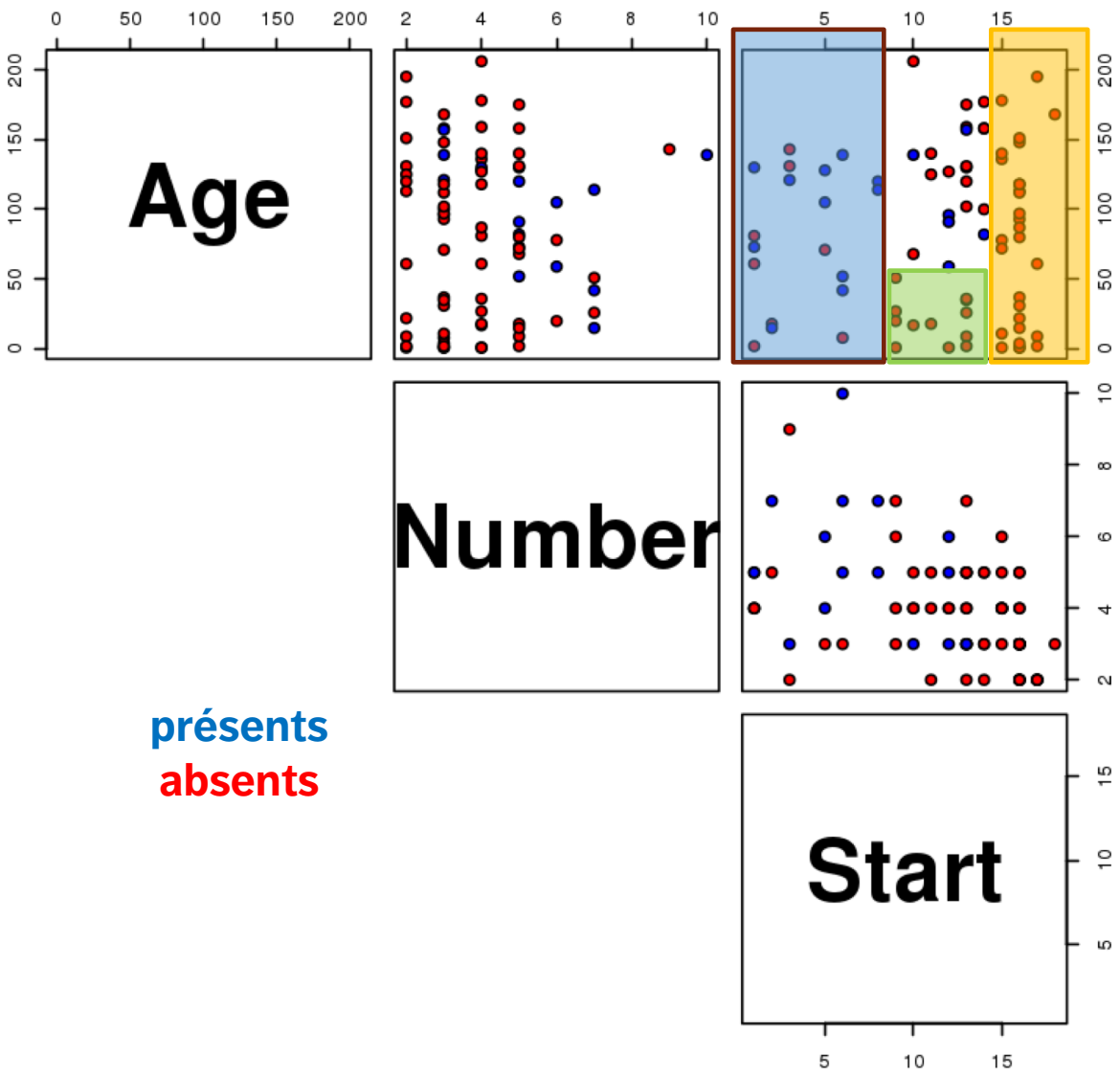

Start >= 8.5
yes
no
Start >= 14
present
absent
Age < 55
absent
Age >= 111
absent
present

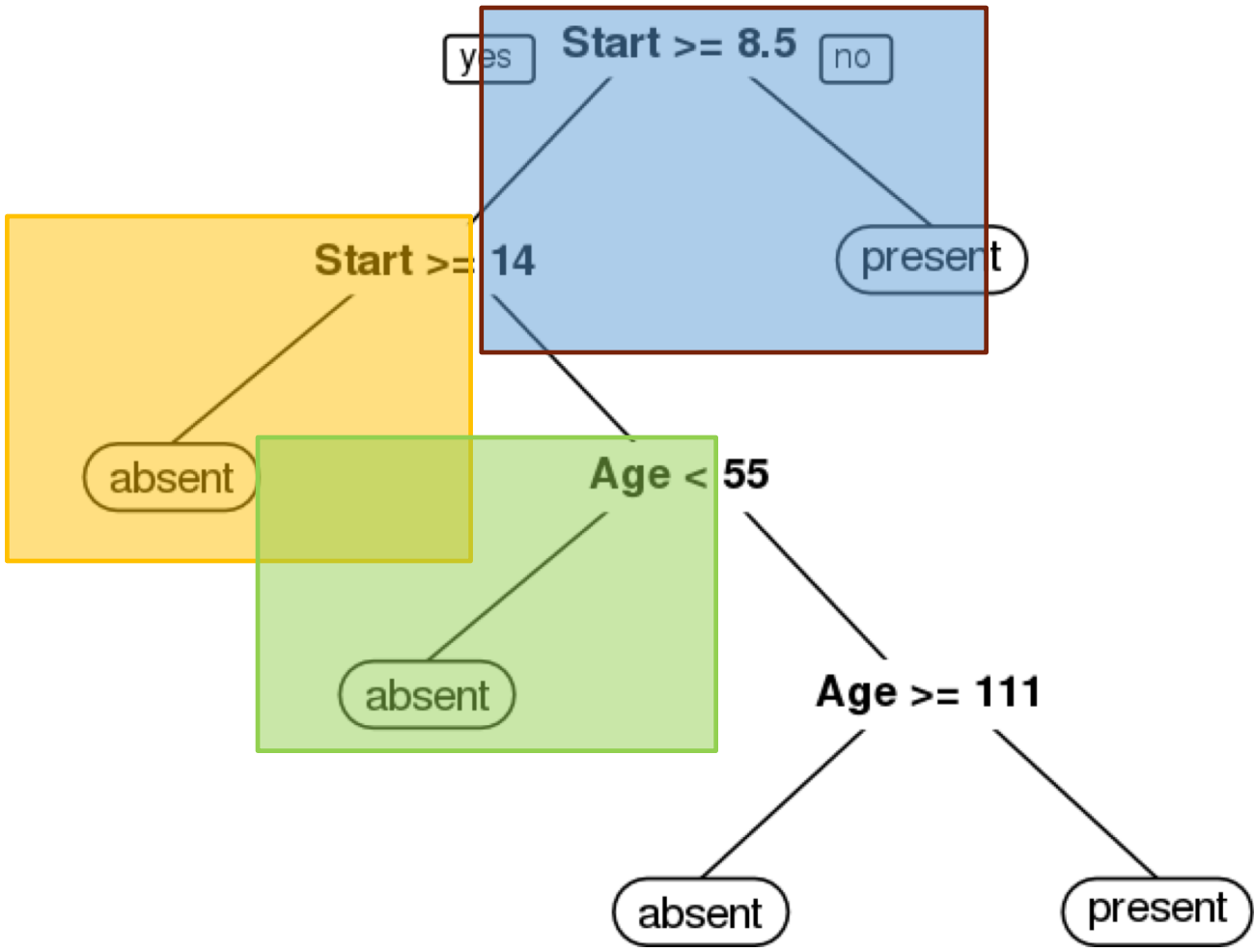

Age
Number
Start
présents
absents

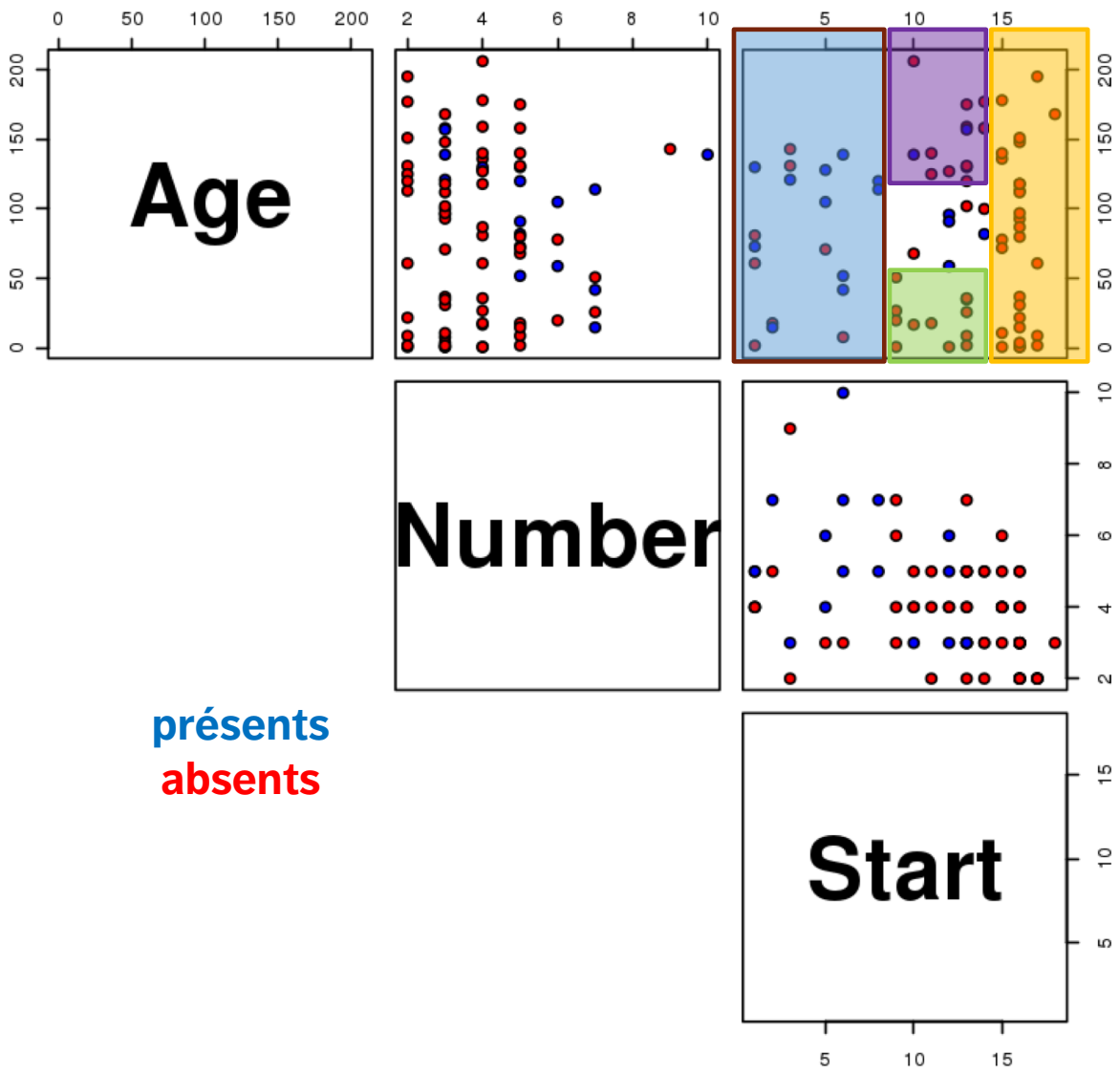

Start >= 8.5
yes
no
Start >= 14
present
absent
Age < 55
absent
Age >= 111
absent
present

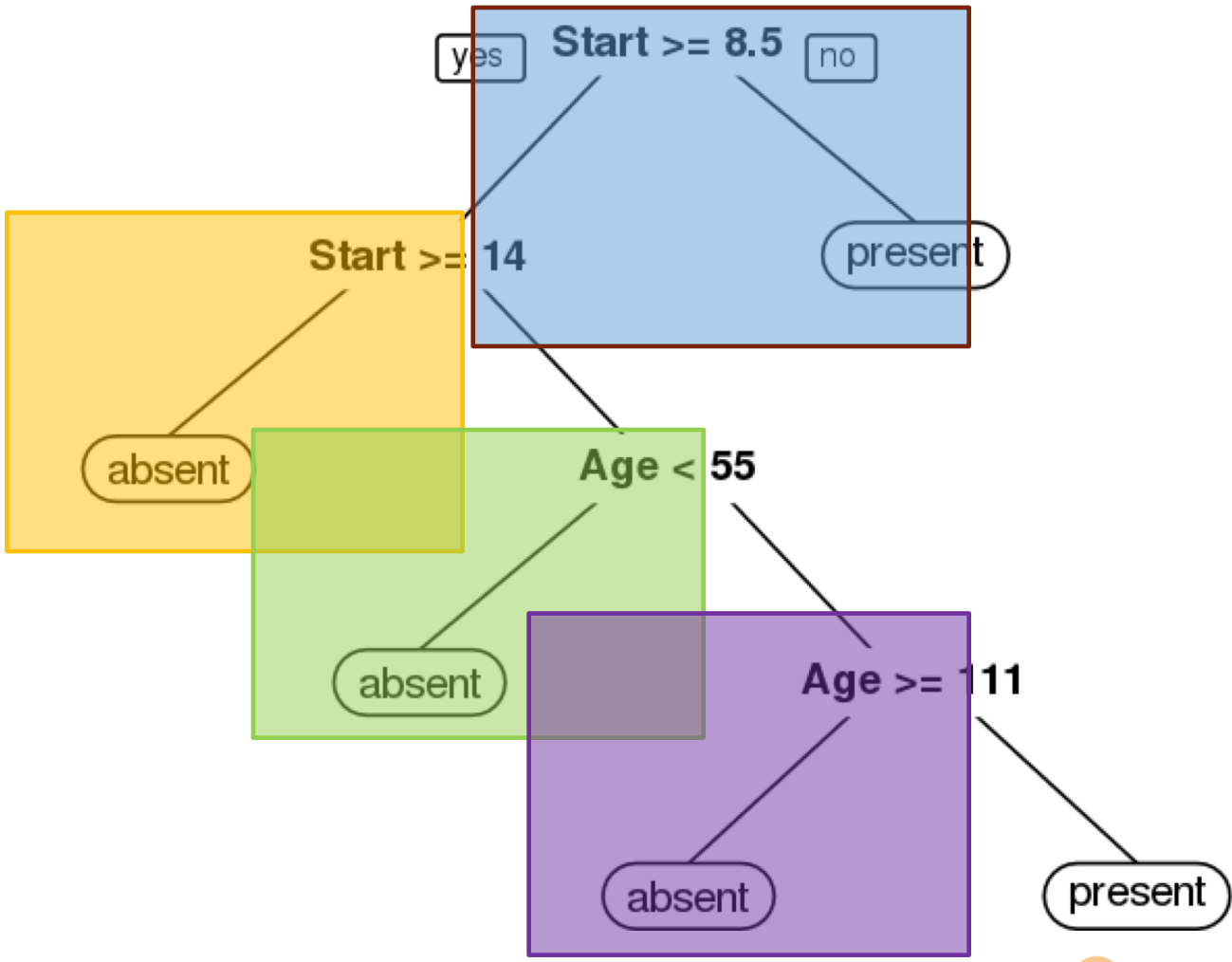

Age
Number
Start
présents
absents


Start >= 8.5
yes
no
Start >= 14
present
absent
Age < 55
absent
Age >= 111
absent
present

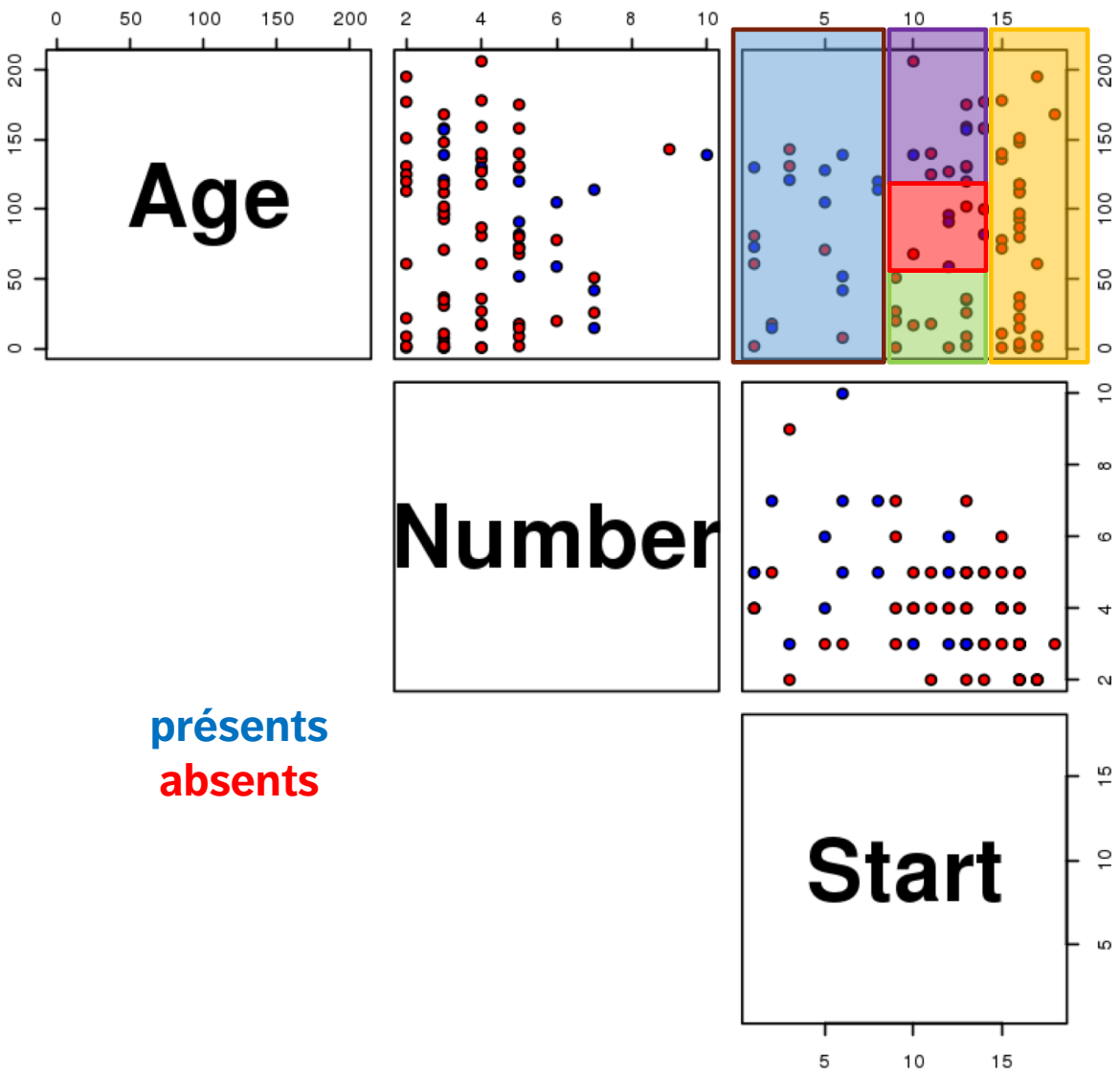

Age
Number
Start
présents
absents

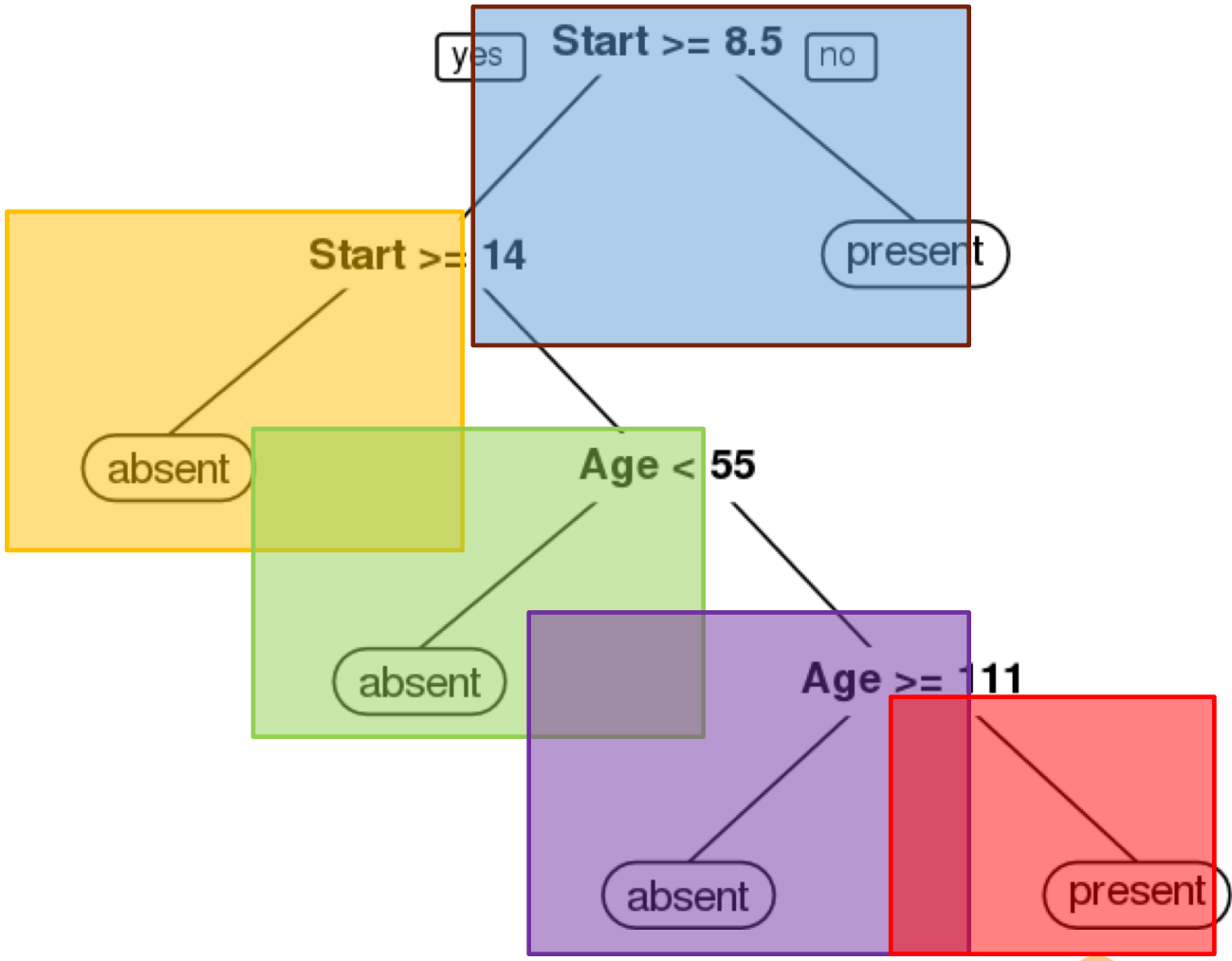

Start >= 8.5
yes
no
Start >= 14
present
absent
Age < 55
absent
Age >= 111
absent
present



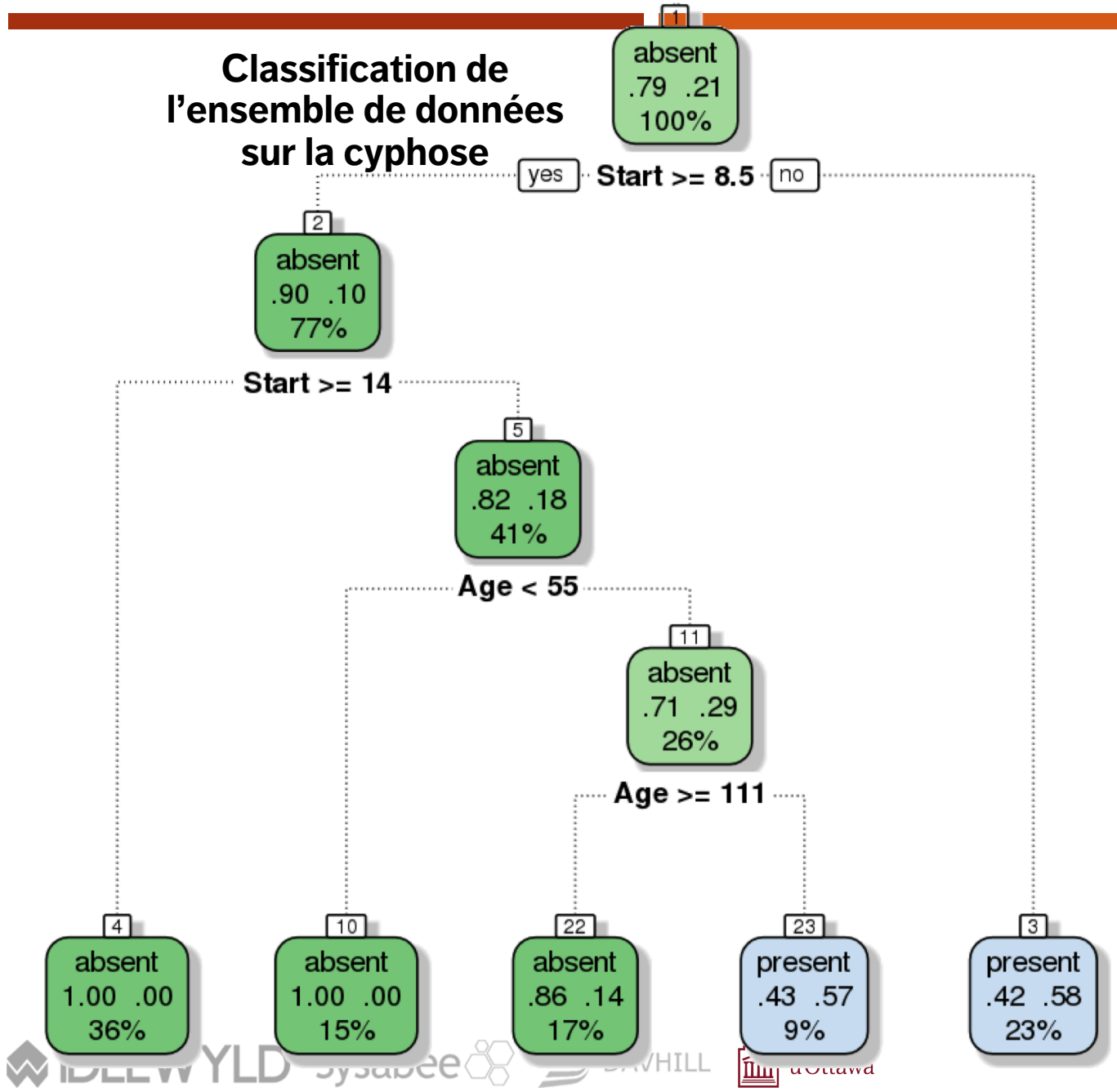


data-action-lab.com

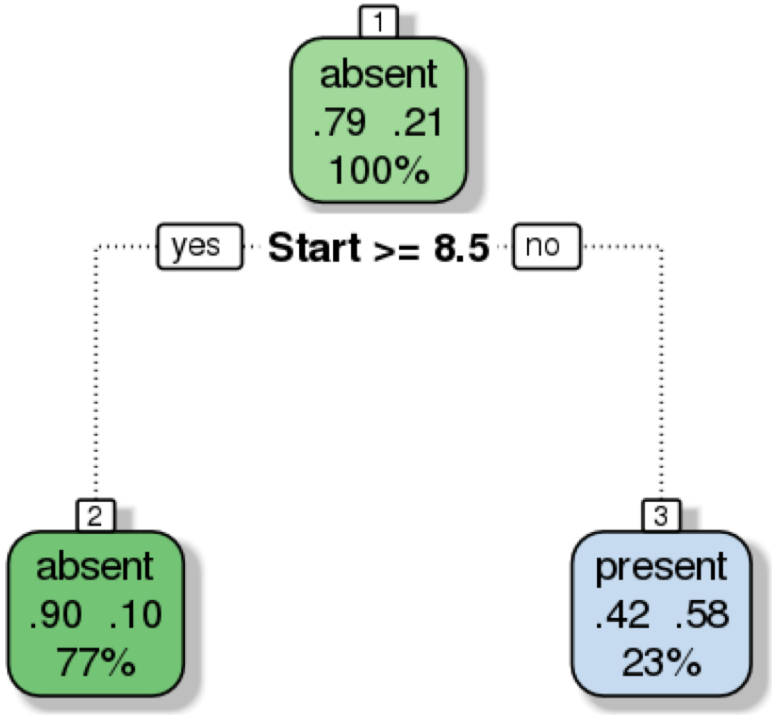

Age
Number
Start
présents
absents


Start >= 8.5
yes
no
present
Start >= 14
absent
Age < 55
absent
Age >= 111
absent
present

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

Classification Rates	
Sensitivity:	0.88
Specificity:	0.40
Precision:	0.88
Negative Predictive Value:	0.40
False Positive Rate:	0.60
False Discovery Rate:	0.12
False Negative Rate:	0.12

Performance Metrics	
Accuracy:	0.81
F1-Score:	0.88
Informedness (ROC):	0.28
Markedness:	0.28
M.C.C.:	0.28
Pearson's chi2:	0.00
Hist. Stat:	0.00

DISCUSSION

Est-ce un bon modèle?

La plupart des paramètres de rendement ne se généralisent pas aux cas multinominaux.

MCC: 69.7%
Accuracy: 78.3%
Pearson: 0.13161
Hist: 30.0%

Actuals			Predicted: Maltreatment: Unfounded	Predicted: Maltreatment: Suspected	Predicted: Maltreatment: Substantiated	Predicted: Risk: No	Predicted: Risk: Yes	Predicted: Risk: Unknown	Total	
Maltreatment		Unfounded	4,577	-	-	198	6	-	*4,781*	*29.2%*
		Suspected	-	965	-	29	2	-	*995*	*6.1%*
		Substantiated	-	-	6,187	116	35	2	*6,339*	*38.7%*
Risk		No	894	-	763	949	19	9	*2,632*	*16.1%*
		Yes	123	-	520	122	111	5	*880*	*5.4%*
		Unknown	212	-	303	184	21	24	*745*	*4.6%*
	Total		*5,805*	*965*	*7,772*	*1,597*	*194*	*40*	***16,372***	
			35.5%	*5.9%*	*47.5%*	*9.8%*	*1.2%*	*0.2%*		

ÉVALUATION DU RENDEMENT

Si y est une valeur numérique approximée par $\hat{y}$, on évalue le rendement à l'aide de:

- **erreur quadratique moyenne** et **erreur absolue moyenne**

$$\text{MSE} = \text{mean}\{(\hat{y}_i - y_i)^2\},\ \text{MAE} = \text{mean}\{|\hat{y}_i - y_i|\}$$

- **erreur quadratique moyenne normalisée** et **erreur absolue moyenne normalisée**

$$\text{NMSE} = \frac{\text{mean}\{(\hat{y}_i - y_i)^2\}}{\text{mean}\{(\bar{y} - y_i)^2\}}, \text{NMAE} = \frac{\text{mean}\{|\hat{y}_i - y_i|\}}{\text{mean}\{|\bar{y} - y_i|\}}$$

- **pourcentage moyen d'erreur** $\text{MAPE} = \text{mean}\left\{\frac{|\hat{y}_i - y_i|}{y_i}\right\}$

- **corrélation** $\rho_{\hat{y},y}$

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.

RÉFÉRENCES

CLASSIFICATION ET ESTIMATION DE LA VALEUR

DOCUMENTATION SUPPLÉMENTAIRE

Méthodes d'estimation de la valeur
https://www.data-action-lab.com/wp-content/uploads/2019/03/Value-Estimation-Methods.pdf

Régression logistique
https://www.data-action-lab.com/wp-content/uploads/2019/03/Logistic-Regression.pdf

Classification naïve bayésienne
https://www.data-action-lab.com/wp-content/uploads/2019/03/Naïve-Bayes-Classification.pdf

RÉFÉRENCES

Kitts, B., Zhang, J., Wu, G., Brandi, W., Beasley, J., Morrill, K., Ettedgui1, J., Siddhartha, S., Yuan, H., Gao, F., Azo, P., Mahato, R. (sous presse), Click Fraud Detection: Adversarial Pattern Recognition over 5 Years at Microsoft, Annals of Information Systems Special Issue on Data Mining in Real-World Applications.

Kitts, B. (2013), The Making of a Large-Scale Ad Server, in Data Mining Case Studies Workshop and Practice Prize 5, Proceedings of the IEEE Thirteenth International Conference on Data Mining Workshops (ICDMW 2013), Décembre, Dallas, TX, IEEE Press.

Fefilatyev, S., Kramer, K., Hall, L., Goldgof, D., Kasturi, R., Remsen, A., Daly, K. (2011), Detection of Anomalous Particles from Deepwater Horizon Oil Spill Using SIPPER3 Underwater Imaging Platform, in Data Mining Case Studies IV, Proceedings of the Eleventh IEEE International Conference on Data Mining, Vancouver, Canada

Kitts, B. (2005), Product Targeting From Rare Events: Five Years of One-to-One Marketing at CPI, Marketing Science Conference, Atlanta, Juin 2005.

RÉFÉRENCES

https://algobeans.com/2016/07/27/decision-trees-tutorial/

https://fr.wikipedia.org/wiki/Arbre_de_d%C3%A9cision_(apprentissage)

https://fr.wikipedia.org/wiki/Analyse_pr%C3%A9dictive

https://fr.wikipedia.org/wiki/R%C3%A9gression_multivari%C3%A9e_par_spline_adaptative

https://en.wikipedia.org/wiki/Evaluation_of_binary_classifiers

Aggarwal, C.C. (éd.) [2015], Data Classification: Algorithms and Applications, CRC Press.

Leskovec, J., Rajamaran, A., Ullman, J.D. [2014], Mining of Massive Datasets, Cambridge Press.

Provost, F., Fawcett, T. [2013], Data Science for Business, O'Reilly.

RÉFÉRENCES

Classification naïve bayésienne (Wikipédia)

Zhang, H. (2014) The optimality of Naïve Bayes

Domings, P., and Pazzani, M. (1997) Beyond independence: Conditions for the optimality of the simple Bayesian classifier

Markham, K. Scikit-learn video #3: Machine learning first steps with the Iris dataset

http://www.ee.columbia.edu/~vittorio/BayesProof.pdf

https://www.cs.cmu.edu/~epxing/Class/10701-08s/Lecture/lecture3-annotated.pdf

http://www.cogsys.wiai.uni-bamberg.de/teaching/ss05/ml/slides/cogsysII-9.pdf