# Les arbres de décision

Dr Emmanuel Chazard
MCU-PH Université Lille 2
http://emmanuel.chazard.org

# Data Mining ou fouille de données

- Définition
  - =KDD (Knowledge discovery in databases)
  - Mise en évidence de connaissances jusqu'alors inconnues dans des bases de données de grand dimension, à l'aide de méthodes dérivées de statistiques, du data management et de l'intelligence artificielle
- Deux types :
  - Data mining non supervisé :
    - Recherche de groupes de patients, d'association de variables
    - « Il n'y a que des X »
    - Exemples : Règles d'association, classification hiérarchique ascendante, nuées dynamiques /K means, analyse en composantes principales, etc.
  - Data mining supervisé :
    - Expliquer une variable par les autres (quand elle est connue)
    - Prédire la valeur de cette variable (quand elle est inconnue, nouvel échantillon)
    - « Il y a un seul Y (variable à expliquer), et plusieurs X (variables explicatives) »
    - Exemples : arbres de décision, régressions, analyse discriminante, réseaux de neurones, etc.

# Que sont les “big data” (base de grande dimension)

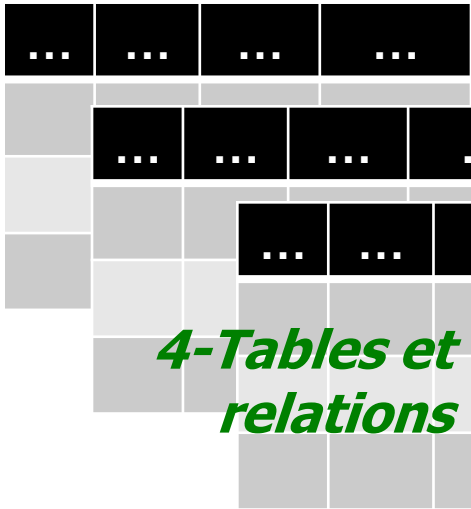
- 5 dimensions de la « grandeur » d’une base :

***2-Variables***

***1-Enregistrements***

| id | age | gender | diagnosis | … |
|---|---|---|---|---|
| 123 | 23 | M | I10 | … |
| 125 | 78 | M | K37 | … |
| 245 | 13 | F | M61.2 | … |
| 278 | 24 | M | I41 | … |
| 324 | 65 | F | I48 | … |
| 350 | 34 | F | F20.2 | … |
| … | … | … | … | … |

***3-Variables qualitatives multivaluées***




***4-Tables et relations***

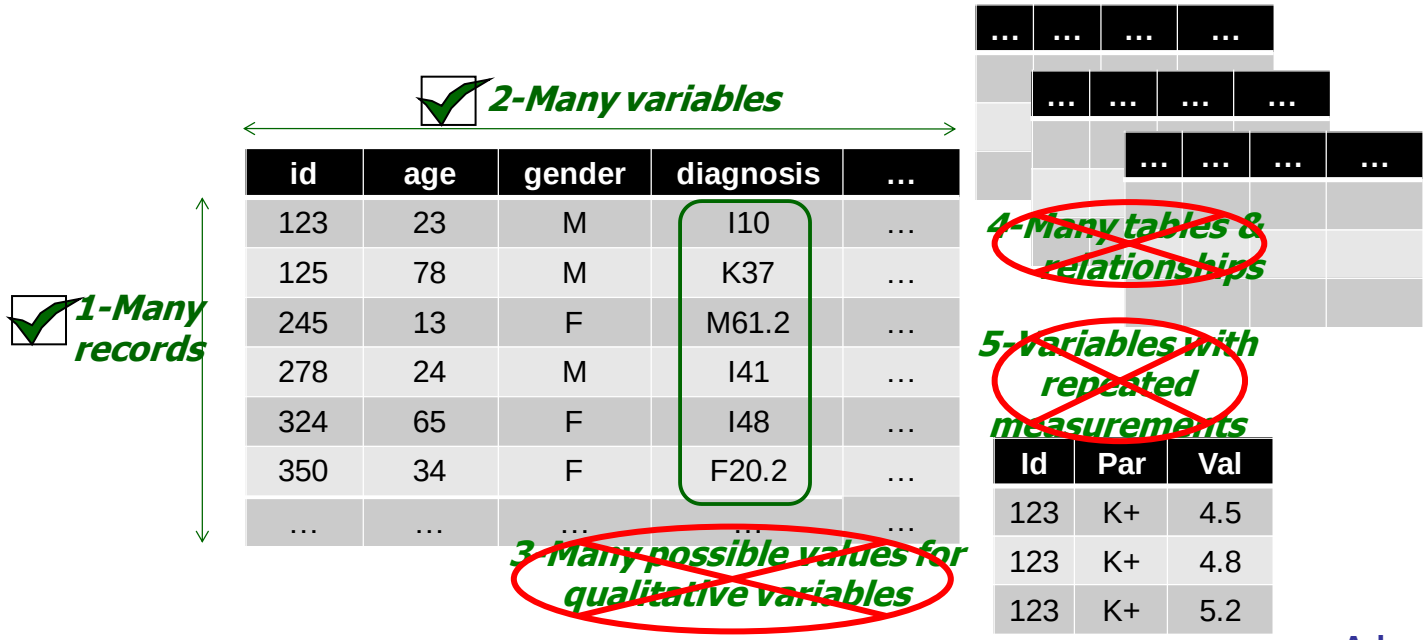
***5-Mesures répétées***

| Id | Par | Val |
|---|---|---|
| 123 | K+ | 4.5 |
| 123 | K+ | 4.8 |
| 123 | K+ | 5.2 |

Dr Emmanuel Chazard
MCU-PH Université Lille 2
http://emmanuel.chazard.org

# Les bases de grandes dimensions doivent être simplifiées

- Obligatoire avant l'analyse statistique
- Il reste donc des tables avec :
  - Beaucoup de lignes (nombreux enregistrements)
  - Beaucoup de colonnes (nombreuses variables)



# Data mining supervisé : quelle variable Y à expliquer ? Exemples des méthodes

- Variable quantitative discrète ou continue
  - Ex : poids {65, 67, 72…}
  - Régression linéaire
  - Arbre de régression
- Variable binaire
  - Ex : malade {0, 1}
  - Régression logistique
  - Arbre de classification
- Variable qualitative
  - Ex : couleur des cheveux {brun, blond…}
  - Régression logistique multinomiale
  - Arbre de classification

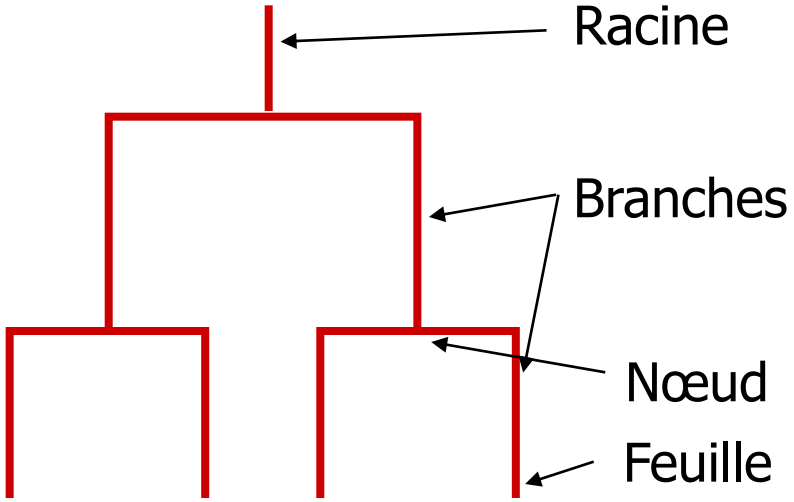
# Data mining supervisé : quelle variable Y à expliquer ? Exemples des méthodes

- Variable correspondant à un décompte
  - Ex : nombre de patients vus en consultation
  - Régression de Poisson
  - Arbre de Poisson
- Variable correspondant à la survenue d'un événement non répété, en prenant en compte le délai :
  - Ex : survenue ou non du décès, au bout d'un temps connu
  - La variable à prédire est de fait le hazard ratio, qui traduit le sur-risque instantané que l'événement se produise
  - Modèle de Cox (analyse de survie)
  - Arbre de survie

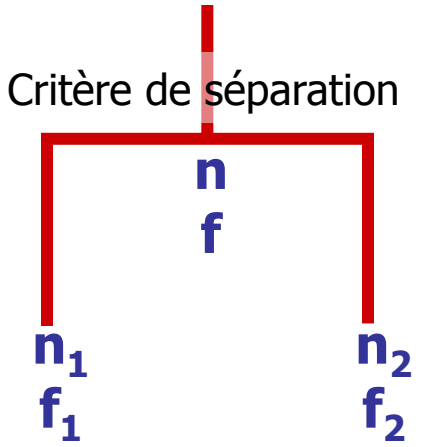
# Exemple introductif : arbre de classification

- On s'intéresse aux séjours qui présentent une hyperkaliémie
  - Echantillon = tous les séjours avec une hyperkaliémie
  - N=100
- Parmi eux :
  - certains sont recontrôlés le lendemain => c'est bien ! (Y=0 ; 70% des séjours)
  - d'autres non => c'est très mal ! (Y=1 ; 30% des séjours)
  - Nous cherchons à comprendre pourquoi cet incident survient : Y est la variable à expliquer
- On dispose pour ce faire de 987 variables explicatives $X_i$ chez ces patients
  - $X_1$ = patient transféré {0;1}
  - $X_2$ = l'hyperkaliémie est mesurée un dimanche {0;1}
  - $X_3$ = patient insuffisant rénal {0;1}
  - $X_4$ = patient diabétique {0;1}
  - …
  - $X_{987}$ (…) {0;1}

# Principes des arbres binaires




Unité de base : le nœud sépare un groupe en deux groupes, un avec une f augmentée, l'autre avec f diminuée.



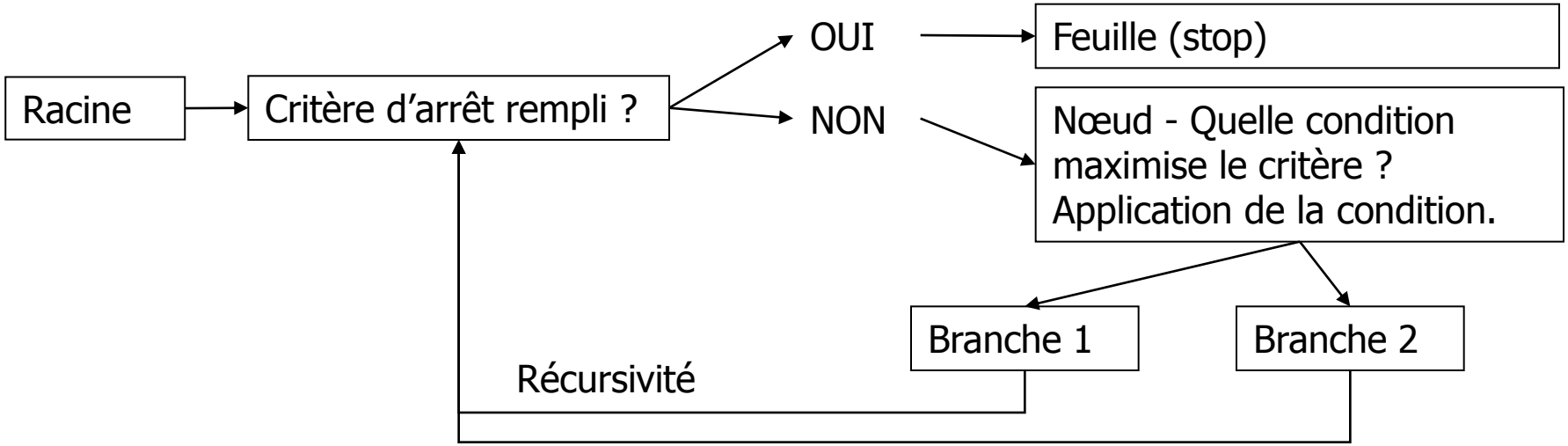

$n=n_1+n_2$

$f=(n_1/n).f_1 + (n_2/n).f_2$

But : $f_1 << f << f_2$

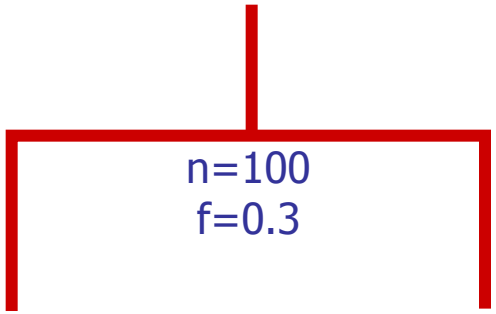
# Principes des arbres binaires

- Il s'agit donc d'un algorithme récursif : le produit d'une fonction est à son tour soumis à cette fonction
- Préalable :
  - Définir un critère de choix de la meilleure condition pour les nœuds
  - Définir un critère d'arrêt : si rempli, la branche devient une feuille. Sinon, l'algorithme continue.
- Déroulement récursif :




# Approche par l'exemple

Première itération : nous cherchons la condition qui séparera le mieux les Y=1 (à droite) et les Y=0 (à gauche)

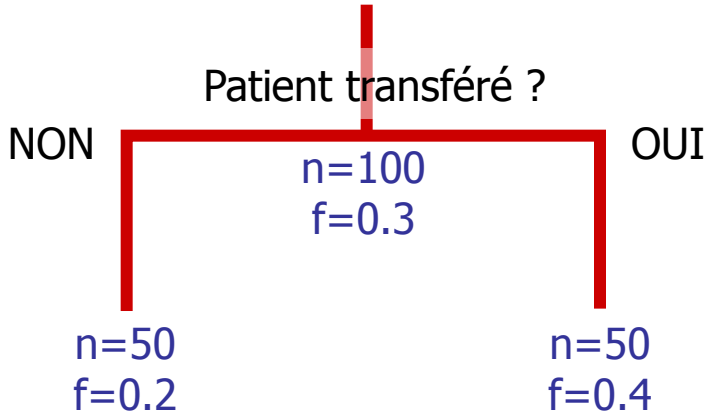



- Echantillon : n=100, f=0.3
- Variable à expliquer Y : séjour déviant ou non {0;1}
- Variables explicatives $X_i$ :
  - $X_1$ = patient transféré {0;1}
  - $X_2$ = dimanche {0;1}
  - $X_3$ = insuffisant rénal {0;1}
  - $X_4$ = diabétique {0;1}
  - …

| Variable | P du Chi² vs Y |
|---|---|
| X1 | $1^{E}$-04 |
| X3 | $1^{E}$-01 |
| X2 | 0.15 |
| X4 | 0.5 |
| … | … |

# Approche par l'exemple

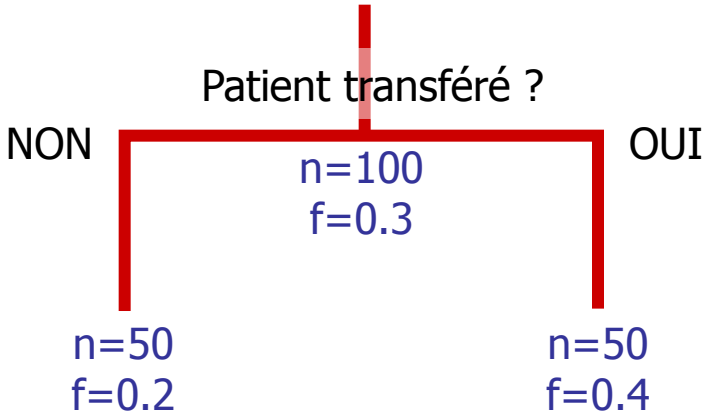
On choisit X1, qui est la plus fortement associée à Y.




- Echantillon : n=100, f=0.3
- Variable à expliquer Y : séjour déviant ou non {0;1}
- Variables explicatives $X_i$ :
  - $X_1$ = patient transféré {0;1}
  - $X_2$ = dimanche {0;1}
  - $X_3$ = insuffisant rénal {0;1}
  - $X_4$ = diabétique {0;1}
  - …

# Approche par l'exemple

Itération maintenant sur la branche de droite.

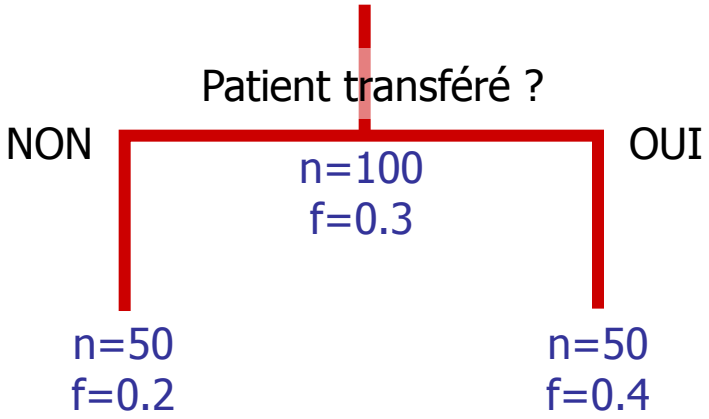



- Echantillon : n=100, f=0.3
- Variable à expliquer Y : séjour déviant ou non {0;1}
- Variables explicatives $X_i$ :
  - $X_1$ = patient transféré {0;1}
  - $X_2$ = dimanche {0;1}
  - $X_3$ = insuffisant rénal {0;1}
  - $X_4$ = diabétique {0;1}
  - …

| Variable | P du Chi² vs Y |
|---|---|
| X3 | 0.21 |
| X2 | 0.35 |
| X4 | 0.60 |
| … | … |

# Approche par l'exemple

Itération maintenant sur la branche de gauche

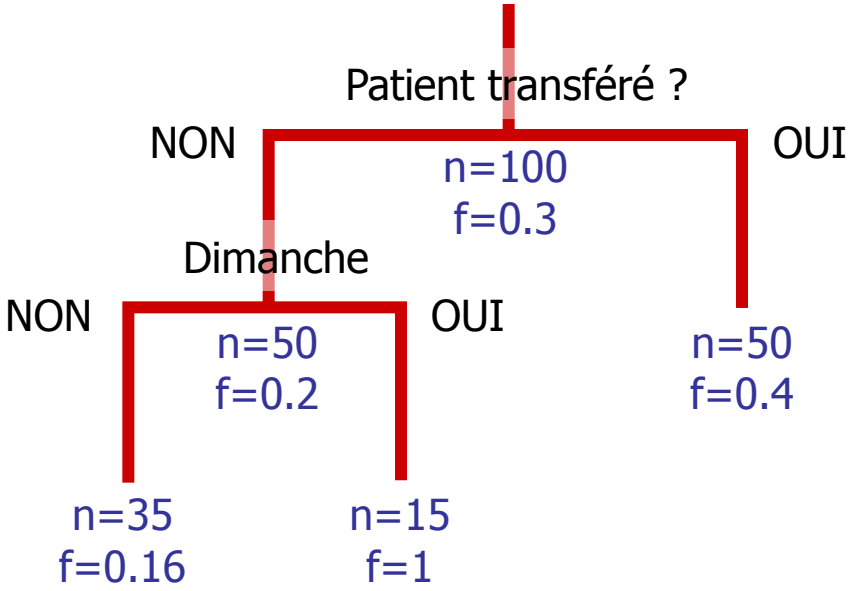



- Echantillon : n=100, f=0.3
- Variable à expliquer Y : séjour déviant ou non {0;1}
- Variables explicatives $X_i$ :
  - $X_1$ = patient transféré {0;1}
  - $X_2$ = dimanche {0;1}
  - $X_3$ = insuffisant rénal {0;1}
  - $X_4$ = diabétique {0;1}
  - …

| Variable | P du Chi² vs Y |
|---|---|
| X2 | $1^{E}$-02 |
| X3 | 0.36 |
| X4 | 0.70 |
| … | … |

# Approche par l'exemple

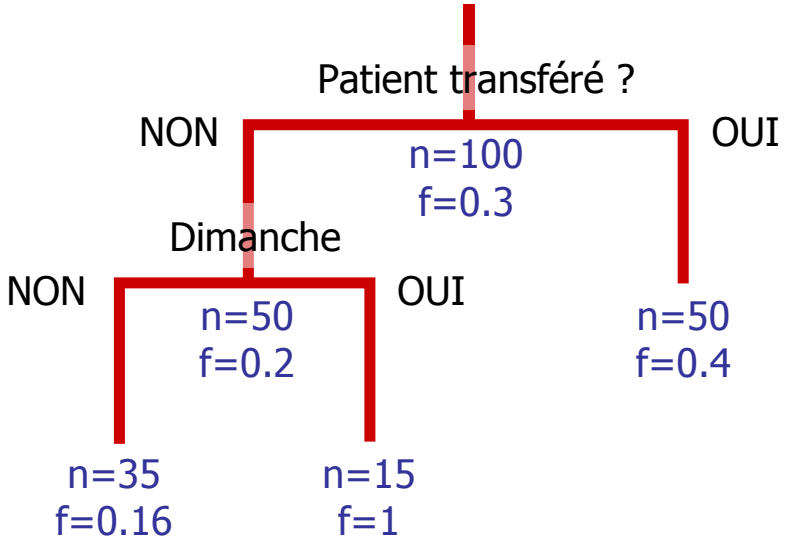
On choisit X2.
Peut-on encore continuer ?




- Echantillon : n=100, f=0.3
- Variable à expliquer Y : séjour déviant ou non {0;1}
- Variables explicatives $X_i$ :
  - $X_1$ = patient transféré {0;1}
  - $X_2$ = dimanche {0;1}
  - $X_3$ = insuffisant rénal {0;1}
  - $X_4$ = diabétique {0;1}
  - …

# Approche par l'exemple

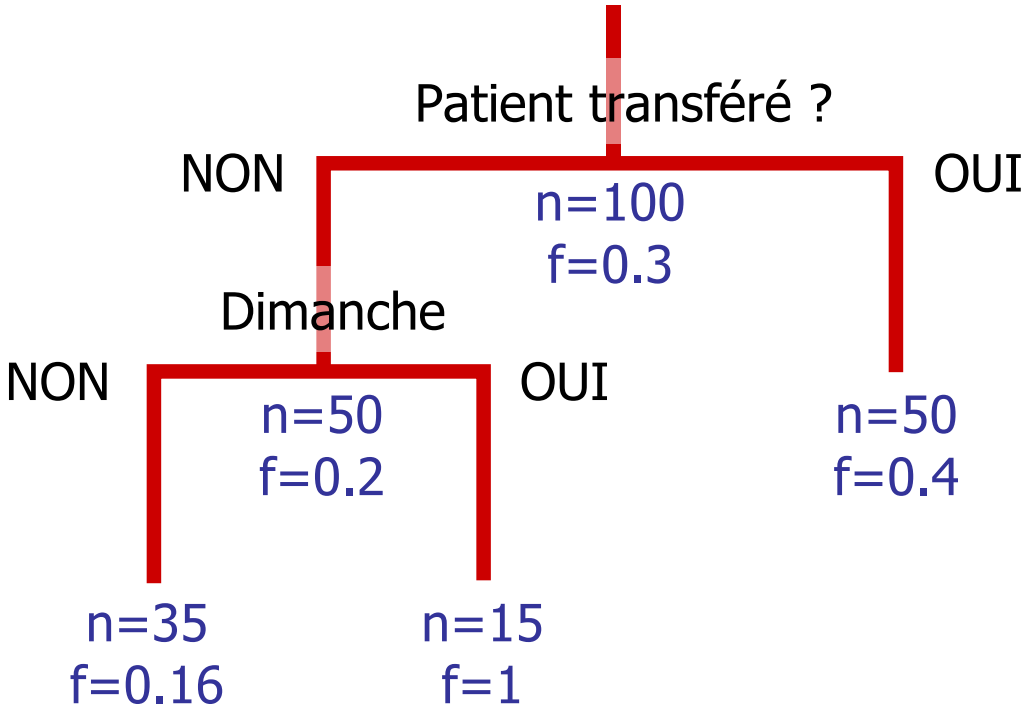
Arbre à 3 feuilles => 3 règles de classification.
De droite à gauche :




- Racine (tous les séjours) : p(Y=1)=0.3
- Patient transféré → p(Y=1)=0.4
- Patient non transféré ET mesure le dimanche → p(Y=1)=1
- Patient non transféré ET mesure autres jours → p(Y=1)=0.16

# Approche par l'exemple

- Interprétation
- Y=1 lorsque l'hyperkaliémie n'est pas recontrôlée
- Dans ce cas, quelle utilisation peut-on faire des connaissances mises en évidence ?



# Utilisation des règles issues d'un arbre ou de règles d'associations

- Sous-groupes identifiés par des règles :
  - Chaque règle est un set de conditions menant à un effet $C1 \cap C2 \cap \ldots \cap Ck \rightarrow E$
  - Soit C le set de conditions : $C = C1 \cap C2 \cap \ldots \cap Ck$
  - Effet E : ici effet binaire Y=1
  - Condition Ci construite de différents manières :
    - Soit avec Xi binaire : par exemple Xi
    - Soit avec Xi quantitatif : par exemple Xi>50
    - Soit avec Xi qualitatif : par exemple $Xi \in \{A;B\}$
- Quantités calculables :
  - Prévalence (confiance à la racine) : P(E)
  - Confiance (prévalence dans une feuille) : P(E|C)
  - Support (importance des cas de la feuille dans l'échantillon de départ) : $P(E \cap C)$
  - Lift (dans la feuille, la prévalence est multipliée par…) : P(E|C) / P(E)
  - Risque relatif (idem mais par rapport aux individus hors de la feuille ; moins pertinent si plusieurs règles sont utilisables) : $P(E|C) / P(E|\overline{C})$

# Quelles questions nous sommes-nous posées pendant la construction ? (bleu : l'exemple)

| | Méthode CHAID<br>Chi-squarred automated interaction detector | Méthode CART<br>Classification and regression trees |
|---|---|---|
| Nature de Y | Binaire, qualitative | Binaire, quantitative, survie, Poisson |
| Nature de X | Binaire, qualitative, quantitative | Binaire, qualitative, quantitative |
| Critère d'arrêt | ▪ Aucun Chi² significatif<br>▪ Y constant dans la branche<br>▪ Toutes les $X_i$ constantes dans la branche<br>*=> Critères de pré-pruning* | *Post-pruning et cross-validation* |
| Critère de choix | Condition qui maximise une quantité associée au Chi² | Indice de Gini |

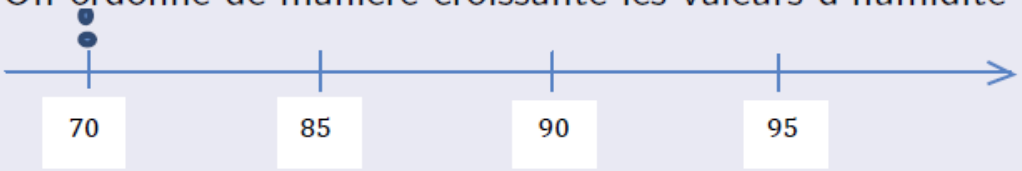
# Questions techniques : Comment utiliser différents types de X ?

- Variables binaires : telles quelles
- Variables qualitatives :
  - Binarisation par regroupement des classes en 0 ou en 1 (merge)
- Variables quantitatives :
  - Binarisation par détermination d'un seuil (cut-off)
  - Procédé :
    - Classement par ordre croissant des valeurs de X
    - Soit $n_d$ le nombre de valeurs différentes, il existe donc $n_d$-1 seuils possibles
    - On teste chaque seuil : 0 en-deçà du seuil choisi, 1 au-delà
    - On conserve le seuil qui maximise la statistique de test

# Exemple de binarisation d'une variable quantitative

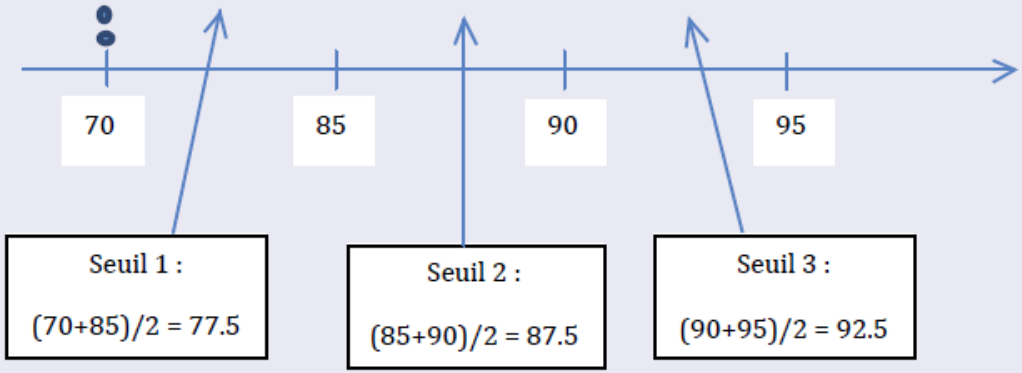
## Exemple avec la variable humidité (1)

- On ordonne de manière croissante les valeurs d'humidité :




- Il y a 5 observations dans le sommet in[soleil] et $n_d = 4$ valeurs distinctes
- Nous avons donc $n_d - 1 = 3$ seuils possibles

## Exemple avec la variable humidité (2)




Exemple emprunté à Michaël Génin

# Exemple de binarisation d'une variable quantitative

## Exemple avec la variable humidité (4)

| | Humidité <77.5 | Humidité >= 77.5 |
|---|---|---|
| Jouer=oui | 2 | 0 |
| Jouer=non | 0 | 3 |

| | Humidité <87.5 | Humidité >= 87.5 |
|---|---|---|
| Jouer=oui | 2 | 0 |
| Jouer=non | 1 | 2 |

| | Humidité <92.5 | Humidité >= 92.5 |
|---|---|---|
| Jouer=oui | 2 | 0 |
| Jouer=non | 2 | 1 |

| Seuils | Pvalue ($\chi^2$) |
|---|---|
| **77.5** | **0.0253** |
| 87.5 | 0.1360 |
| 92.5 | 0.3613 |

Exemple emprunté à Michaël Génin

# Exemple de binarisation d'une variable qualitative

## Principe

- Initialement : la segmentation d'une variable qualitative produit autant de sommets enfants que de modalités
- Possibilité de fusion des sommets enfants $\rightarrow$ limiter la fragmentation des données (faibles effectifs) et les sommets enfants "redondants"
- Comparaison des distributions de la VAE dans chaque sommet enfant et regroupement des sommets ayant des profils proches
- Test du $\chi^2$ d'équivalence distributionnelle
  - $H_0$ : les deux sommets enfants ont des profils similaires
  - $H_1$ : les deux sommets enfants ont des profils différents
- On fusionne les deux sommets enfants ayant les profils les plus proches (au sens du test) puis on réitère l'opération jusqu'à ce qu'aucune fusion ne soit possible
- Possibilité qu'aucune fusion ne se réalise
- Possibilité que tous les sommets enfants soient fusionnés $\rightarrow$ la variable de segmentation est éliminée d'office

Emprunté à Michaël Génin

# Exemple de binarisation d'une variable qualitative

## Exemple avec la variable Ensoleillement (1)

- Intégration de la possibilité de fusion
- Comparaison des sommets deux à deux :

| Sommets | $\chi^2$ | Pvalue ($\chi^2$) | Action |
|---|---|---|---|
| Soleil et couvert | 3.6 | 0.058 | - |
| Soleil et Pluie | 0.4 | 0.527 | Fusion |
| Couvert et Pluie | 2.06 | 0.151 | - |

- Risque de première espèce ($\alpha$) de 10%
- Les modalités Soleil et Pluie peuvent être fusionnées

## Exemple avec la variable Ensoleillement (2)

| Sommets | $\chi^2$ | Pvalue ($\chi^2$) | Action |
|---|---|---|---|
| (Soleil et Pluie) et Couvert | 3.1 | 0.078 | - |

Aucune fusion n'est possible $\rightarrow$ l'algorithme s'arrête !

Emprunté à Michaël Génin

# Généralités sur le pré-pruning et le post-pruning (approche simplifiée)

- Pré-pruning (exemples de méthodes) :
  - A chaque nœud, la pousse de l'arbre est susceptible de s'arrêter (le nœud devient une feuille)
  - Critères connus avant le nœud :
    - Branche déjà pure (que des Y=0 ou que des Y=1)
    - Effectif trop faible dans la branche
    - Toutes les variables $X_i$ constantes
  - Critères connus dans le nœud :
    - Aucun test significatif
    - Effectif à venir trop faible dans les branches filles
  - Critères connus en considérant l'ensemble de l'arbre déjà construit :
    - Apport du nœud dans l'arbre (arbitrage entre l'amélioration de la prédiction et la longueur de l'arbre : critère de parcimonie)

# Généralités sur le pré-pruning et le post-pruning (approche simplifiée)

- Post-pruning (exemples de méthodes) :
  - L'arbre pousse en entier, puis on le raccourcit itérativement pour améliorer le compromis entre prédiction et complexité
  - Exemples :
    - Critère de parcimonie évaluant l'effet de la suppression d'un nœud
    - Cross-validation (tirage au sort dans chaque nœud pour éviter le surajustement, i.e. la découverte d'association fortuite entre des variables)

- Ces différentes possibilité de pré-pruning ou post-pruning participent à la diversité des méthodes

# Avantages des arbres de décision

- Construction :
  - Construction facile, paramétrable
  - Natures variées de Y
  - Natures variées de X, utilisables en même temps
- Point de vue statistique :
  - Fonctionne en grande dimension (nombreuses lignes, nombreuses colonnes) sans difficulté
  - Pas besoin de filtrer les variables à inclure : nombre immense possible (même plus que d'individus), corrélations entre $X_i$ non problématiques
  - Pas de principe d'additivité des effets, au contraire effets conditionnels plus réalistes en médecine
  - Totalement non paramétrique
- Lecture
  - Arbre facile à interpréter, lecture intuitive
  - Règles faciles à réutiliser (oral, papier, tableur, base de données…)

# Inconvénients des arbres de décision

- Risque de sur-ajustement (d'où la cross-validation)
- Instabilité des arbres (mais prédiction peu altérée ; solution = random forest trees)
- Moins performants que les réseaux de neurones
- Résultats utilisables tels quels pour la prédiction, mais nécessitent un remaniement expert pour l'explication…
- Pas de prise en compte des données manquantes

# Logiciels et packages

- Sipina
  - Logiciel Libre
  - Interface graphique du type SPSS
  - Méthodes implémentées : CHAID, ID3, C4.5, Improved CHAID...
  - Possibilité de construction d'arbres en utilisant des connaissances expertes
- R - Package Rpart
  - Logiciel libre
  - Package reconnu et souvent utilisé en recherche
  - Méthode implémentée : CART
  - Rendus graphiques paramétrables