PLACE DE L'ANALYSE DISCRIMINANTE DANS LE TRAITEMENT D'ENQUETES PSYCHO-SOCIALES

F. FACY*, Y. LECHEVALLIER**

* INSERM Unité 302
44 Chemin de Ronde
78170 Le Vesinet Cedex

** INRIA-Rocquencourt
Domaine de Voluceau
78153 Le Chesnay Cedex

Résumé

Les enquêtes psycho-sociales reposent sur des questionnaires fermés ayant trois volets : les données socio-démographiques et légales, la prise de drogue et les données médicales. L'objectif de l'analyse de ces questionnaires est d'expliquer, dans un cadre plurifactoriel, le volet des données socio-démographiques à partir des données médicales.

Comme ces données socio-démographiques et légales sont caractérisées par plusieurs prédicteurs nous proposons de décomposer cette analyse en deux étapes :

La première étape est une étape de classification qui nous permettra de dégager des groupes homogènes par rapport aux données sociales.

La deuxième étape est une étape de discrimination de cette typologie en fonction des données médicales. Plusieurs méthodes de discrimination seront utilisées et comparées (analyse discriminante linéaire, analyse discriminante par le modèle d'indépendance, classification par arbre).

Mots-clefs

Classification automatique, discrimination, enquêtes épidémiologiques.

1. INTRODUCTION

1.1 Cadre de l'analyse

Des enquêtes psycho-sociales portent sur l'usage et l'abus de drogues. Devant ce phénomène plurifactoriel, où l'individu la drogue et le contexte social développent des interactions, les études épidémiologiques sont organisées de façon à illustrer certaines étapes de la genèse et du développement de la toxicomanie [Min 89].

Les enquêtes épidémiologiques reposent sur des questionnaires fermés ayant trois volets : les données socio-démographiques et légales, la prise de drogue et les données médicales. Ces trois volets sont caractérisés essentiellement par des variables qualitatives. Un échantillon de 3 040 dossiers a été constitué.

L'objectif de l'analyse de ces questionnaires est de prédire le volet des données socio-démographiques et légales, à partir du volet contenant les données médicales.

1.2 Méthodologie utilisée

L'objectif des méthodes de discrimination est de prédire ou reconnaître le groupe a priori d'un individu, en fonction d'un ensemble de prédicteurs. Les prédicteurs seront les variables qualitatives associées au volet médical. Le but est d'expliquer les données socio-démographiques et légales composées de 9 prédicteurs. Comme ces données socio-démographiques et légales sont caractérisées par plusieurs variables qualitatives, nous ne pouvons pas appliquer directement les méthodes de discrimination. Nous avons donc réalisé une typologie en 5 classes de notre population et cette typologie en 5 classes formera la variable à prédire.

L'analyse comportera les deux étapes suivantes :

• La première étape est de rechercher l'existence de sous-groupes homogènes par rapport aux données sociales et pénales. Cette étape doit aussi servir à une analyse descriptive du tableau de données.

• La deuxième étape est de construire une fonction de décision à partir de cette typologie. Cette étape devra permettre une interprétation des relations entre le volet des données socio-démographiques et légales aux volets associés à la prise de drogue et aux données médicales. Cette étape est une étape de discrimination; pour cette étude trois méthodes de discrimination sont utilisées, (analyse discriminante linéaire, analyse discriminante par le modèle d'indépendance, classification par arbre).

2. ETUDE DESCRIPTIVE DE SUJETS VUS EN MILIEU CARCERAL

2.1. L'enquête

L'enquête réalisée en 1988 et 89 auprès des 16 antennes "toxicomanie" en prison [Fac 91], permet d'illustrer les traitements statistiques successivement effectués. 3099 détenus ont été identifiés comme toxicomanes par les équipes des antennes et décrits individuellement et anonymement suivant un questionnaire qui aborde les facteurs socio-démographiques et les caractéristiques des produits. (Convention entre l'INSERM, le Ministère de la Justice et la D.G.S.).

Un échantillon de 3 040 dossiers est constitué. Une description univariée et partielle des principales variables du premier volet est donnée ci-dessous afin que le lecteur ait une première vision de ce tableau.

2.2. Analyses univariées

Les sujets interrogés sont 9 fois sur 10 des hommes. Les étrangers représentent un quart de la population masculine alors que 13 % des femmes sont d'origine étrangère. Presque la moitié des sujets ont entre 20 et 24 ans et un quart de 25 à 29 ans. La population féminine est plus jeune que la population masculine. Parmi les étrangers, 12 % ont moins de 20 ans et 16 % ont plus de 30 ans.

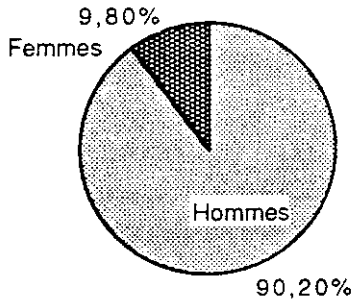

Fig 1 : Sexe

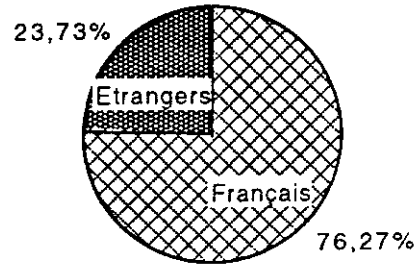

Fig 2 : Nationalité

Notre échantillon comprend 55 % de prévenus dont 46 % ont un délit en rapport avec l'ILS (infraction à la législation sur les stupéfiants).

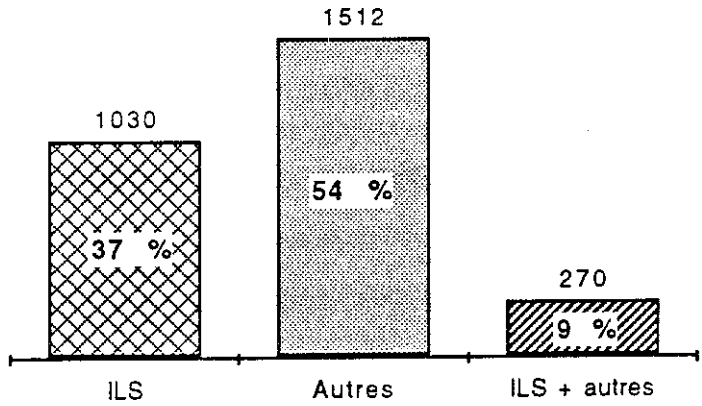

Fig 3 : Nature du délit

Plus de la moitié ont été incarcérés pour la première fois avant 20 ans. 41 % des incarcérations ont eu lieu avant l'usage de drogues et 45% des sujets sont passés devant un tribunal pour enfants.

A partir des données socio-démographiques comprenant un nombre important de variables qualitatives, l'analyse factorielle et la classification dégagent une structure de la population sous forme de typologie.

2.3. La classification automatique

Après la description des sujets pour chaque facteur relevé dans l'enquête et quelques comparaisons, le problème envisagé se pose ainsi : existe-t-il des sous-groupes homogènes dans la population toxicomane incarcérée, par rapport aux données sociales ?

La classification automatique retenue est un partitionnement en cinq classes de la population. Réalisée sur le volet des données sociales et pénales, elle permet d'obtenir une typologie des toxicomanes incarcérés.

CLASSE	1	2	3	4	5
EFFECTIF	1028 32 %	889 27 %	974 30 %	208 6 %	141 4 %
SITUATION PENALE AGE	Récidivistes 25 ans et +	Récidivistes 20-24 ans	Primaires 20-24 ans	Prévenus - 19 ans	Valeurs manquantes
ACTIVITES	42 % Diplômés ou qualifiés	45 % Aucun diplôme	47 % Diplômés ou qualifiés	60 % Aucun diplôme	
	48 % Travail épisodique	50 % Travail épisodique	34 % Travail continu		
REGION	(42 %) région parisienne	(23 %) bassin méditerranéen (12 %) Nord		Grandes agglomér.	(89 %) région parisienne

Tab 1 : Typologie des sujets incarcérés

Dans le tableau 1 sont indiqués les facteurs les plus discriminants. Cette partition, croisée avec les variables des autres volets, montre de fortes corrélations avec des usages spécifiques des produits. Une interprétation détaillée de cette typologie est dans le rapport [Fac 91].

Classe 1 :

Cette classe a été nommée **Classe des récidivistes de plus de 25 ans.** La population de cette classe consomme principalement de l'héroïne avec une fréquence d'utilisation importante (60 % en consomment tous les jours) alors qu'ils n'étaient que 30 % à avoir la même fréquence dans la population totale.
Le cannabis est consommé par environ 50 % des personnes et l'utilisation d'autres drogues est beaucoup moins nette.

Classe 2 :

Cette classe a été nommée **Classe des récidivistes de 20 - 24 ans.** Comme pour la classe 1, 60 % des jeunes récidivistes consomment de l'héroïne mais la fréquence d'utilisation est moindre. Le cannabis est faiblement représenté et la cocaïne est prise de façon intermittente par 15 % des personnes de la classe. Enfin, 40 % des personnes interrogées, consommant du speedball ou des benzodiazépines sont dans cette classe.

Classe 3 :

Cette classe a été nommée **Classe des primaires de 20 - 24 ans.** Leur consommation est proche de la consommation moyenne de l'échantillon. Plus de 25 % des individus de cette classe consomment de l'héroïne.

Classe 4 :

Cette classe a été nommée **Classe des prévenus de moins de 19 ans**. Ce groupe se distingue de ceux précédemment étudiés par le fait que l'héroïne n'est pas la drogue la plus fréquente. Le cannabis est utilisé par 40 % des individus de la classe tous les jours. D'autres drogues caractérisent cette classe :

- les benzodiazépines et l'alcool sont consommés actuellement, avec des fréquences variables, par environ 25 % des individus,
- les solvants correspondent à la drogue utilisée au début : 12 % des individus en prenaient tous les jours.

Classe 5 :

Cette classe a été nommée **Classe des valeurs manquantes**. Les questionnaires incomplets sont particulièrement nombreux pour les sujets de cette classe.

2.4. L'analyse factorielle

Une analyse factorielle des correspondances multiples est appliquée sur le même tableau de données que celui de la classification automatique. Les 4 premiers axes expliquent 17 % de l'inertie totale. On visualise ainsi les classes de la typologie précédente sur les premiers axes factoriels

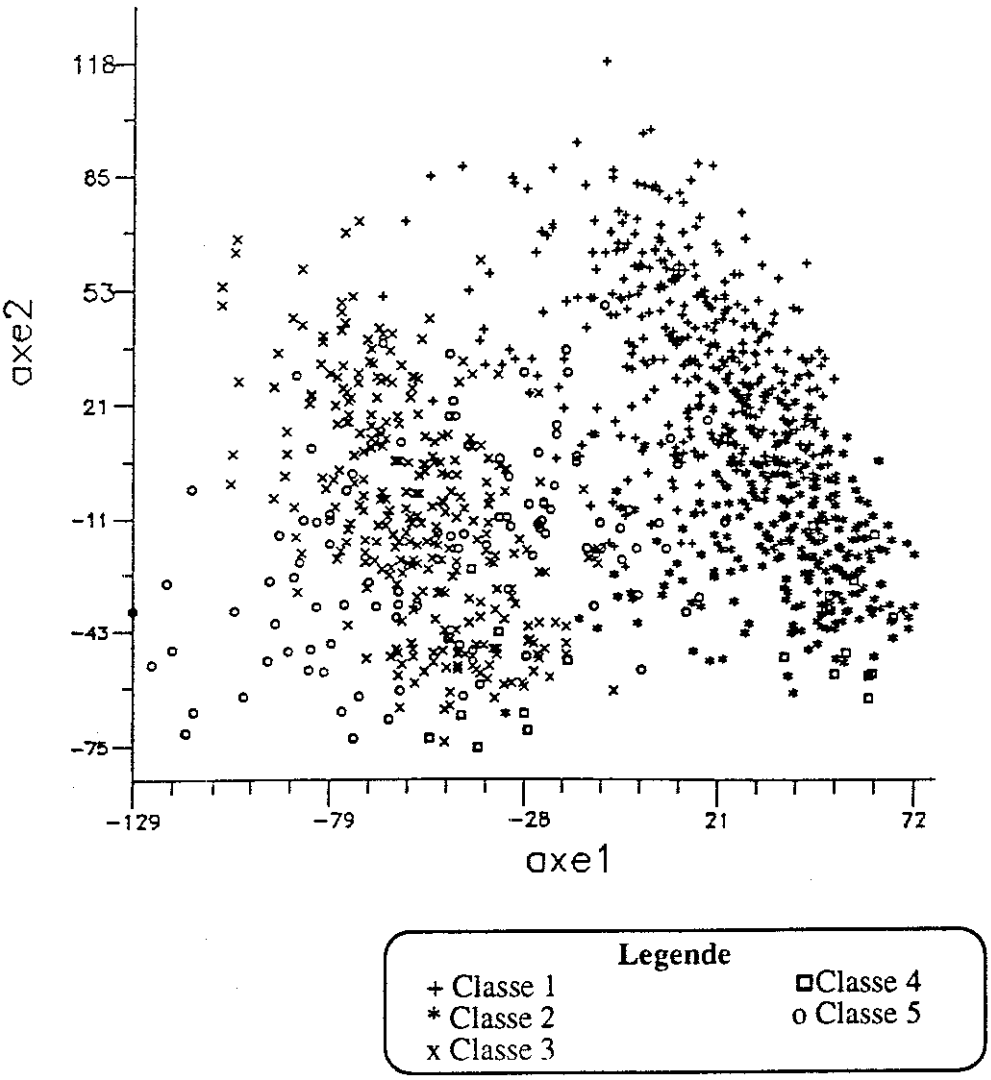

Fig 4 : Plan Factoriel des axes 1 et 2

L'axe 1 oppose deux groupes de sujets de 20-24 ans: les primaires(Classe 3) et les récidivistes (Classe 2) essentiellement marqués par des problèmes judiciaires dans l'enfance et un age de première incarcération précoce. L'axe 2 oppose des sujets vivants en couples avec des enfants et ayant plus de 25 ans (Classe 1) aux prévenus de moins de 19 ans,vivant plutôt avec les parents(Classe 4). Sur cet axe on observe une graduation de l'age, les plus jeunes sont en bas de cet axe et les plus âgés sont en haut de cet axe.

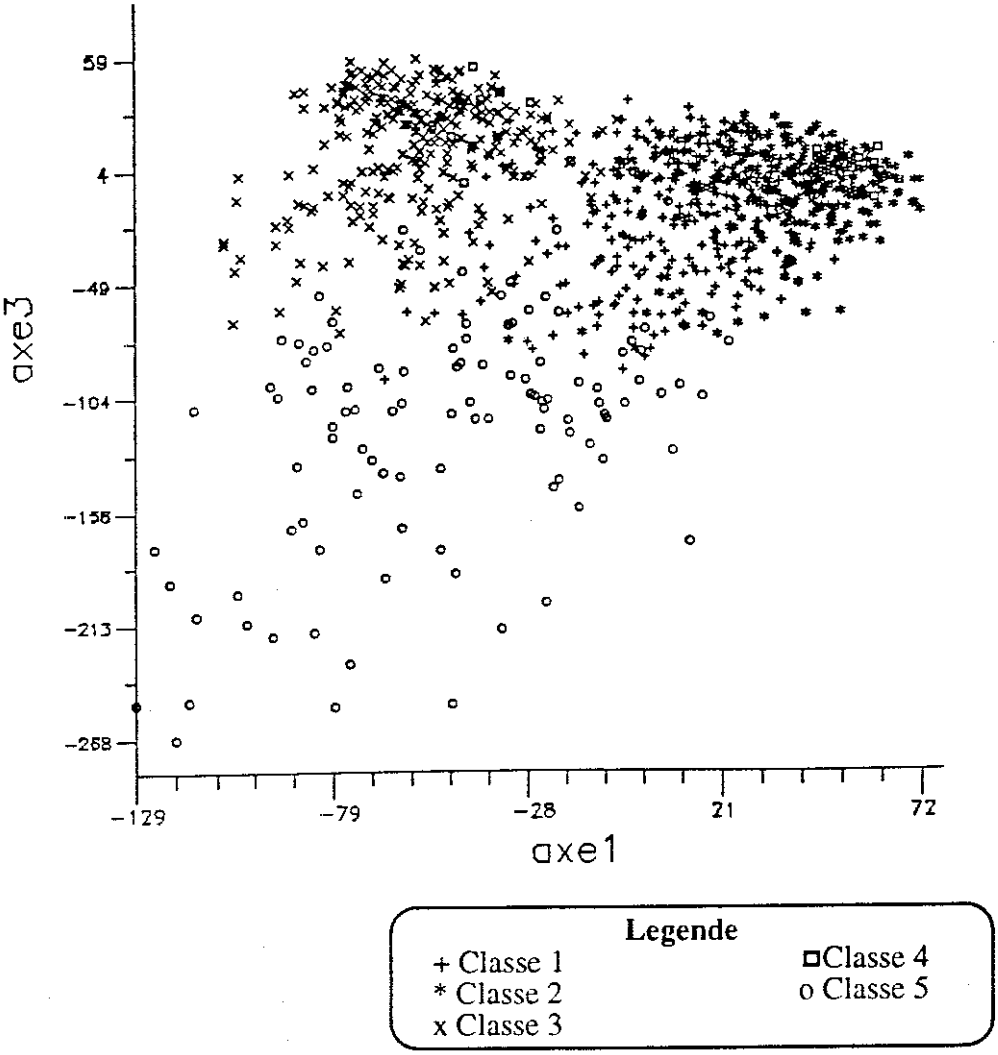

Fig 5 : Plan Factoriel des axes 1 et 3

L'axe 3 est caractérisé par les données manquantes et on peut voir sur ce plan se détacher la classe 5 où se trouvent les questionnaires les plus incomplets.

2.5. Discussion

L'intérêt de l'approche typologique est de suggérer l'existence de sous-groupes bien typés, par rapport à des données sociales et pénales.

Ainsi à travers ces résultats, on peut rejeter l'idée d'un "profil" de toxicomane délinquant. Il en existe au moins quatre types, bien définis par rapport à leur situation pénale, entre prévenus, délinquants primaires et récidivistes. L'importance des mesures judiciaires est ici manifesté par rapport aux comportements des toxicomanes et l'influence des parquets est suggéré par les corrélations avec les zones géographiques.

Par rapport aux individus, la différenciation des usages de produits est nette suivant les classes, les jeunes prévenus sont concernés par le cannabis et les récidivistes plus âgés par l'héroïne.
La dernière classe (marquée par les non réponses) montre les limites de l'enquête, qui du fait de la méthode épidémiologique, ne peut atteindre la finesse d'une enquête clinique. Il est à noter que c'est plus en Région Parisienne que se rencontre le problème de manque d'informations [OCR 91].

3. ANALYSE DISCRIMINANTE

3.1. Les principes [CDG 89]

Le but de l'analyse discriminante est de construire une fonction de décision qui identifie chaque individu à un groupe a priori en fonction d'informations contenues dans les descripteurs. Ainsi à chaque individu est associé un ensemble de descripteurs, dans cette étude cet ensemble de descripteurs est surtout caractérisé par des variables qualitatives. Dans l'espace défini par ces variables, qui sera appelé **espace de représentation**, chaque individu de la population à étudier est associé à un point de cet espace. Une variable qualitative, représentant les groupes a priori, sera la variable à expliquer.

Le rôle d'une méthode discriminante est construire une fonction de décision qui identifie, à partir des variables descriptives, chaque individu à un groupe a priori qui sera sa **classe d'affectation**. Cette identification est parfaite si chaque individu est affecté son groupe a priori qui sera sa classe d'affectation. Ainsi une méthode de discrimination construit, pour chaque groupe a priori, une **région** dans l'espace de représentation défini par les variables descriptives. La seule contrainte est que les régions soient toutes disjointes ; c'est-à-dire qu'un point de l'espace des variables descriptives ne peut appartenir qu'à une seule région donc qu'à un seul groupe a priori. Les méthodes utilisées dans cette analyse suivront les principales étapes suivantes.

3.2. Les principales étapes

La première étape est la construction du tableau de données :

Définition et sélection de l'ensemble de base : c'est l'ensemble des individus qui vont être utilisés par la méthode de discrimination. Dans cette analyse l'ensemble de base est constitué de la totalité des toxicomanes incarcérés

Sélection de la variable à expliquer : cette variable doit être une variable qualitative et elle caractérise l'ensemble de groupes à reconnaître appelés groupe a priori. Dans cette étude il s'agit de prédire le volet des données socio-démographiques et légales à partir du volet des données médicales. Ce volet est constitué de plusieurs variables qualitatives aussi un compromis acceptable est d'utiliser notre typologie en cinq classes comme variable à expliquer.

Sélection des variables descriptives : ces variables doivent servir à l'identification des individus. Ces variables seront issues des variables du volet contenant les données médicales.

La deuxième étape est le choix de la méthode discriminante :

Notre ensemble de variables descriptives est constitué en majorité de variables qualitatives. Généralement les méthodes de discrimination ne sont applicables que sur un ensemble homogène de variables aussi nous allons réaliser, pour chaque méthode sélectionnée, les transformations de variables ou codages nécessaire à homogénéisation de ce tableau. Ce codage est une étape très importante car il s'agit de rendre cohérent le tableau de données sur lequel la méthode de discrimination sera appliquée. C'est aussi une étape délicate car il faut transformer le type de variable sans perdre les informations utiles contenues dans cette variable.

L'objectif d'une méthode de discrimination est de construire une fonction de décision dont le **taux d'erreur de classement** est le plus faible possible. Cependant l'utilisateur veut aussi comprendre la décision prise par la méthode ce qui implique une fonction de décision la plus simple que possible. Ces deux critères sont contradictoires car le premier entraîne généralement l'utilisation de méthodes ,

dites "Boites noires", pour atteindre le meilleur taux de reconnaissance, le deuxième privilégie le caractère explicatif de la fonction de décision à la performance de la méthode.

Le deuxième objectif est d'avoir une fonction de décision la plus généralisable que possible c'est à dire indépendante de l'échantillon servant à sa construction, c'est la validation de la fonction de décision.

La troisième étape est le choix de la statégie de validation de la fonction de décision:

La fonction de décision crée un partitionnement de l'espace de représentation en régions de telle manière que le tableau de contingence, calculé entre la variable qualitative associée aux groupes a priori et celle associée à ces régions, soit le plus proche possible d'un tableau ayant les éléments non diagonaux nuls. Ce tableau de contingence sera appelé **tableau de classement** et il permet d'évaluer l'efficacité de la fonction de décision. La performance de la fonction de décision ne peut pas être mesurée à partir de l'échantillon utilisé pour sa construction car il est toujours possible, dans l'espace de représentation, de construire des régions où il n'y a aucun individu mal classé. Il faut donc contrôler sa performance en utilisant des individus n'ayant pas servi à sa construction, ces individus forment l'**échantillon test**. Une stratégie simple de validation consiste a construire sur l'échantillon test le tableau de contingence entre les groupe a priori et les classes d'affectation puis à partir de ce tableau de classement on peut calculer le taux d'erreur de classement et le risque de Bayes associés à cette fonction de décision.

Dans notre application l'ensemble test est constitué de 20 % de la population et **l'ensemble d'apprentissage**, utilisé pour construire la fonction de décision, est constitué des individus n'ayant pas été pris dans l'ensemble test. Les méthodes de rééchantillionnage ne sont pas indispensables car notre ensemble de base est important

3.3. Analyse discriminante linéaire

C'est la méthode classique en discrimination. Dans le cadre bayésien la surface de discrimination obtenue avec égalité des probabilités à priori, égalité des coûts, normalité des densités et égalité de matrice de variances-covariances des groupes a priori est un hyperplan. La fonction de décision associée consiste à affecter un individu au groupe a priori dont le centre de gravité, avec la distance euclidienne, est le plus proche.

Le caractère qualitatif de la majorité des variables rend inopérante l'analyse discriminante linéaire. Aussi une solution est la factorisation de ces variables, en utilisant l'analyse factorielles des correspondances, puis la sélection des facteurs les plus discriminants, et enfin réaliser une analyse discriminante linéaire sur ces facteurs. Le principal inconvénient de cette stratégie est que la fonction de décision basée sur les facteurs est une "boite noire" pour l'utilisateur. Mais une transcription, à partir des variables du questionnaire, de cette fonction de décision consiste à affecter à chaque modalité significative une valeur ou un score. La valeur de cette nouvelle fonction de décision est égale, pour chaque individu, à la somme des scores associés aux modalités prises par cet individus. Bien que réalisée, en utilisant les programmes MULTC et DISC de la bibliothèque MODULAD [MOD 87], nous ne donnerons pas les résultats dans cet article.

3.4. Analyse discriminante par le modèle d'indépendance [Mha 91]

Les variables descriptives utilisées par cette méthodes doivent être qualitatives ou binaires. La discrimination par le modèle d'indépendance suppose que les variables sont indépendantes conditionnellement aux groupes a priori. On montre que la règle déduite est linéaire mais différente de la règle obtenue par la discrimination linéaire. Dans notre cas nous avons pris comme hypothèse l'équiprobabilité des groupes a priori afin d'être, dans le cadre bayesien, équivalent à l'hypothèse utilisée dans la méthode de segmentation, qui est de prendre les coûts inversement proportionnels aux probabilités a priori. Une version de ce programme sera bientôt disponible dans la bibliothèque MODULAD.

3.5. Classification par arbre, Segmentation

3.5.1. Introduction

Morgan et Sonquist [MoS 63] ont mis au point le premier programme de segmentation AID (Automatic Interactions Detector). On trouve dans Bouroche et Tenenhaus [BoT 70] une présentation des premières méthodes de segmentation. Brieman, Friedman, Olsen et Stone [BrF 84] proposent le programme CART qui introduit la notion d'élagage; leur livre est une référence très importante. Dans le domaine de l'apprentissage Quilan [Qui 86] a introduit la segmentation avec le programme ID3.

L'objectif des méthodes de segmentation est de construire une fonction de décision qui peut être représentée par un arbre binaire. Cette construction est réalisée de manière récursive; à chaque noeud de l'arbre la méthode découpe en deux classes la population en optimisant un critère d'évaluation de cette coupure. Cette division s'arrête quand l'un des critères d'arrêt est vérifié. Alors ce noeud est appelé **segment terminal** et il est étiqueté en lui affectant un des groupes a priori.

L'affectation de nouveaux individus se fait de manière très simple et rapide : à chaque noeud de l'arbre est associé, par construction, un test, si la réponse à ce test est positive l'individu est dirigé vers la partie droite de l'arbre sinon il est dirigé vers la partie gauche et ceci jusqu'à ce qu'il arrive dans une feuille qui est un segment terminal. Alors sa classe d'affectation est le groupe a priori associé à ce segment terminal. La dimension et la profondeur de l'arbre binaire de décision dépend du choix des critères d'arrêt. Les outils nécessaires à la construction d'un arbre de décision sont :

- la définition d'un ensemble de questions binaires,
- un critère d'évaluation de la segmentation,
- une règle d'arrêt de la construction de l'arbre,
- une règle d'affectation de chaque segment terminal à un groupe a priori.

Les méthodes de segmentation ont deux représentations principales :

- l'une sous forme d'un **arbre binaire** par exemple :

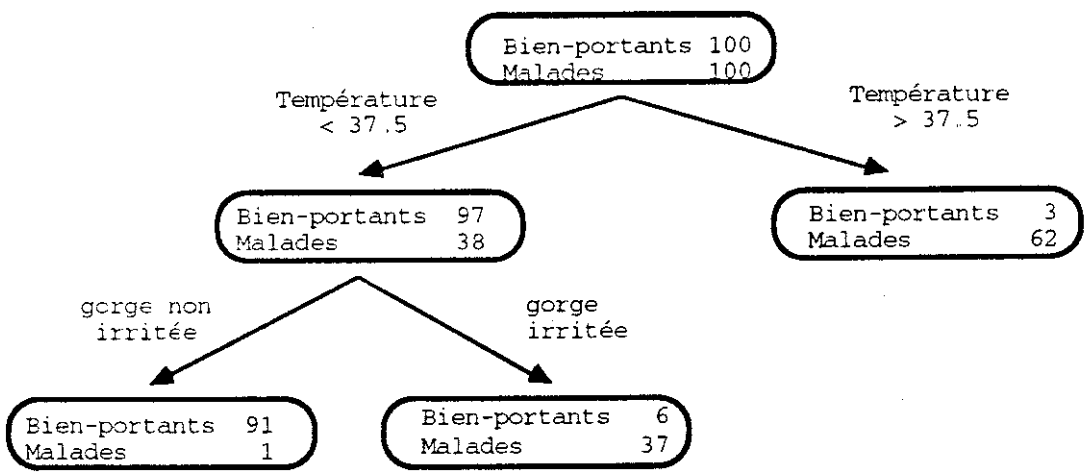

Fig 6 : Arbre de décision

Cette représentation est réalisée à partir d'un échantillon fictif de 200 patients, 100 bien-portants et 100 atteints d'une maladie donnée, décrits par un certain nombre de caractéristiques médicales. Cet arbre binaire est constitué, de deux questions binaires (température < 37.5) et (gorge irritée) qui sont associées aux deux noeuds de cet arbre, et de trois segments terminaux qui sont les feuilles de cet arbre. Un de ces segments terminaux est associé au groupe a priori "Sain" et les deux autres au groupe a priori "Malade".

L'affectation d'un individu se fait à l'un des segment terminaux en partant de la racine de l'arbre et en le parcourant suivant les réponses aux questions associées aux noeuds de cet arbre binaire.

• l'autre sous la forme de **règles de production** :

Le groupe a priori "Malade" est caractérisé par deux règles :
Règle 1 : [température > 37.5]
Règle 2 : [température < 37.5] **ET** [gorge irritée]

Le groupe a priori "Sain" est caractérisée par la règle suivante :
Règle 3 : [température < 37.5] **ET** [gorge non irritée]

Pour chaque individu une et une seule de ces trois règles est vérifiée ; il est alors affecté au groupe a priori associé à la règle qui est vraie. Cette écriture de la fonction de décision montre le pouvoir explicatif de cette méthode car les règles sont facilement interprétables par l'utilisateur.

3.5.2. Règle d'affectation d'un segment terminal

Nous avons choisi comme règle d'affection de chaque segment terminal à un groupe a priori la règle bayésienne dont le principe est le suivant :

On note Πj les probabilités a priori associées aux groupes a priori et $C(i / j)$ le coût de mauvais classement d'un individu du groupe a priori j dans la classe d'affectation i. Pour un individu, représenté par ses p valeurs $\underline{x} = (x_1, \ldots x_p)$ et par son groupe a priori j (noté $Y(\underline{x}) = j$) nous décidons son affectation au groupe a priori i en utilisant la fonction de décision d (noté $d(\underline{x}) = i$). La classe d'affectation i est l'ensemble des individus ayant été affectés au groupe a priori i par la fonction de décision d.
Le coût de mauvais classement de cette fonction est égal à :

$$R(d) = \sum_{j=1}^{k} \Pi j \left[\sum_{i \neq j} C(i / j)\, P[d(\underline{x}) = i / Y(\underline{x}) = j] \right]$$

La fonction de décision d est appelée règle de Bayes si son coût est minimal. Le programme DNP de la bibliothèque MODULAD suppose, par hypothèse, que les coûts sont inversement proportionnels aux probabilités a priori. Dans ce cas l'affectation d'un segment terminal t au groupe a priori i se fait de la manière suivante :

L'ensemble des régions associées aux segments terminaux réalise un partionnement de l'espace de représentation. Ainsi l'arbre de décision définit une fonction δ qui associe à tout vecteur $\underline{x}$ de l'espace de représentation un segment terminal t. La règle de Bayes, appliquée au segment t, consiste à assigner à ce segment le groupe a priori i qui minimise l'expression suivante :

$$\sum_{j \neq i} C(i / j).P[Y(\underline{x})=j / \delta(\underline{x})=t].$$

Comme $P[Y(\underline{x})=j / \delta(\underline{x})=t] = \Pi_j .P[\delta(\underline{x})=t / Y(\underline{x})=j]/P[\delta(\underline{x})=t]$

d'où

$$\sum_{j \neq i} C(i / j).\Pi_j .P[\delta(\underline{x})=t / Y(\underline{x})=j]/P[\delta(\underline{x})=t]$$

et comme $C(i / j).\Pi_j$ est constant par hypothèse, il faut choisir le groupe a priori i tel que la probabilité $P[\delta(\underline{x})=t / Y(\underline{x})=i]$ soit maximale. L'estimation de cette probabilité est obtenue par N_{it} / N_i ; où N_{it} est le nombre d'individus du groupe a priori i dans le segment t et N_i est l'effectif du groupe a priori i. Ainsi l'affectation se fait en choisissant la fréquence, conditionnellement aux groupes a priori, la plus élevée.

Dans ces conditions il est facile de voir que le risque de Bayes est estimé par :

$$\sum_{j \neq i} N_{ij} / N_i$$

où N_{ij} est le nombre d'individus du groupe a priori i dans la classe d'affectation j, le tableau de classement de la fonction de décision d contient ces valeurs

3.5.3. Choix du critère d'évaluation

Ce critère mesure la pertinence de la question binaire vis-à-vis du problème de discrimination. Le choix de ce critère influe peu sur les performances de la méthode, ceci est montré très clairement dans l'étude comparative de Mingers [Min 89]. De plus il existe des relations entre ces critères, une liste de ces critères et une étude des liaisons fonctionnelles entre ces critères est réalisée dans [CeL 90].

Cependant on peut remarquer que ces critères sont souvent fondés sur la notion d'impureté développée par Breiman [BrF 84]. Ces critère utilisent la structure d'arbre de la règle de décision et calculent la réduction d'impureté lorsque l'on remplace dans l'arbre de décision le segment terminal t par deux nouveaux segments terminaux t_g et t_d. Ces segments terminaux sont les fils de gauche et de droite du segment t obtenus en répondant à la question binaire q. Traditionnellement le noeud gauche est associé à la réponse positive à cette question binaire, celui de droite est associé à la réponse négative. Cette réduction d'impureté s'écrit : $I(q, t) = i(t) - P(t_g)\, i(t_g) - P(t_d)\, i(t_d)$ où les valeurs $P(t_g)$ et $P(t_d)$ sont les proportions des individus mis dans les segments t_g et t_d et i la fonction d'impureté définie sur les sous-ensembles de l'ensemble des individus

Nous avons choisit un autre type de critère qui est directement lié au risque de Bayes. Ce critère a été proposé par Friedman [Fri 77] et Celeux et Lechevallier [CeL 82]. Dans ce cas les variables doivent être continues ou binaires. Si les variables sont binaires il n'y qu'une seule coupure possible; dans le cas où la variable est continue alors les questions binaires $(X \leq x)$? imposent le découpage de cette variable X en deux demi-droite représentant chacune un groupe a priori.

Dans ce cadre on choisi la coupure c de manière à rendre minimum le coût R de mauvais classement. Si le nombre de groupes a priori est égal à deux le coût R de la fonction de décision d_c associé à la coupure c s'écrit ainsi :

$$R(c) = \Pi_1 C(2 / 1) P[d_c(\underline{x}) = 2 / Y(\underline{x}) = 1] + \Pi_2 C(1 / 2) P[d_c(\underline{x}) = 1 / Y(\underline{x}) = 2]$$

Si on suppose que la région $]-\infty ; c[$ est associée au groupe a priori 1 nous avons :

$$P[d_c(\underline{x}) = 1 / Y(\underline{x}) = 2] = P[x \in\ ]-\infty ; c[/ Y(\underline{x}) = 2] = F_2(c)$$

où F_2 est la fonction de répartition de la variable X pour le groupe a priori 2. D'où :

$$R(c) = \Pi_2 C(1 / 2) F_2(c) + \Pi_1 C(2 / 1)(1 - F_1(c))$$

et comme $\Pi_1 C(2/1) = \Pi_2 C(1/2) = \alpha$ on a : $R(c) = \alpha(1 + F_2(c) - F_1(c))$

Ainsi rechercher le minimum de R est équivalent à rechercher l'écart maximal entre les fonctions de répartitions F_1 et F_2. Cet écart est la distance de Kolmogorov-Smirnov entre les deux groupes a priori. On choisit la région discriminante R_1 du groupe a priori 1 égale à $]-\infty ; c[$ si $F_1(c) > F_2(c)$) sinon on choisit comme région discriminante l'intervalle $[c ; -\infty [$.

Lorsque le nombre de groupe a priori est supérieur à deux, il faut rechercher pour chaque variable et pour chaque coupure le meilleur regroupement $\mathcal{A}$ des groupes a priori d'où :

$$D(c_j) = \sup_{A \in \mathcal{A}} \sup_{x} |\hat{F}_{\bar{A}}(x) - \hat{F}_A(x)|$$

où $\mathcal{A}$ désigne l'ensemble des partitions en 2 classes de $(P_1,\ldots,P_k)$ et $\bar{A}$ désigne le complémentaire de A dans $\mathcal{A}$.

La fonction de répartition de A est égale à :

$$F_A(x) = \frac{1}{\Pi_A} \sum_{P_i \in A} \pi_i \hat{F}_i(x)$$

3.5.4. Elagage de l'arbre

Afin de réaliser un arbre généralisable et robuste (c'est-à-dire indépendant de l'échantillon qui a permis sa construction) Brieman [BrF 84] a développé une stratégie d'élagage de l'arbre. En simplifiant, cette stratégie consiste à construire un arbre le plus grand possible à partir d'un ensemble d'apprentissage et puis en utilisant un ensemble test, indépendant de l'ensemble d'apprentissage, d'éliminer les noeuds non significatifs. Un noeud est jugé non significatif s'il entraîne, sur l'ensemble test, une augmentation du risque de Bayes. Sur l'ensemble d'apprentissage le risque de Bayes est toujours décroissant. L'article de Gueguen et Nakache [GuN 88] présente une application médicale utilisant cette stratégie.

En conclusion l'avantage des arbres de décision réside essentiellement dans la lisibilité des règles d'affectation d'un individu à un groupe a priori. L'inconvénient majeur est le caractère unidimensionnel de la segmentation mais ceci devient un avantage si le nombre de variables explicatives est très grand car, à la fin, l'arbre de décision est basé sur un sous ensemble réduit de ces variables.

4. APPLICATION A L'ETUDE DES SUJETS VUS EN MILIEU CARCERAL

4.1. Sélection des groupes a priori

La typologie précédente en 5 classes sera la variable à prédire. Cependant deux classes de la typologie n'offrent que peu d'intérêt ce sont les classes 4 et 5. La classe 4 comporte une population jeune et son effectif est faible. La classe 5 comprend les questionnaires ayant beaucoup de non réponses et donc non informatifs. A partir des classes 1, 2 et 3 sélectionnés nous allons réaliser une discrimination. Après ces analyses nous verront qu'il est difficile de reconnaître ces 3 classes. Les classes 2 et 3 forment une population homogène en fonction de l'age et il est intéressant d'analyser s'il existe une différence de comportement ; pour cela, on va réaliser une discrimination entre ces deux classes.

4.2. Les principaux résultats par la discrimination par modèle d'indépendance

4.2.1. Résultats de la discrimination en 3 classes

La variable à expliquer représente les 3 classes (1, 2, 3) de la typologie. Les variables explicatives concernent le volet médical. Après application de la méthode de classement, les concordances et discordances sont analysées par classe pour les échantillons d'apprentissage et test.

		Groupes a priori		
		1	**2**	**3**
Classes d'affectation	**1**	338	232	159
	2	138	167	97
	3	284	286	339

Tab 2 : Tableau de classement sur l'ensemble d'apprentissage

Les taux de bon classement apparents sont de 44,5 % pour la classe 1, de 24.4 % pour la classe 2 et de 57.0 % pour la classe 3.

		Groupes a priori		
		1	**2**	**3**
Classes d'affectation	**1**	77	70	41
	2	29	44	29
	3	56	63	101

Tab 3 : Tableau de classement sur l'ensemble test

Le taux de bon classement de la classe 1 est de 47.5 % , celui de la classe 2 est de 24.9 % et celui de la classe 3 est de 59.1 % . En utilisant la stratégie de la validation croisée nous avons :

		Groupes a priori		
		1	**2**	**3**
Classes d'affectation	**1**	407	293	200
	2	154	153	116
	3	379	367	481

Tab 4 : Tableau de classement par la validation croisée

Le taux de bon classement de la classe 1 est maintenant de 43.3 %, celui de la classe 2 est de 18.8 % et celui de la classe 3 est de 60.3 %. Ces taux sont très semblables à ceux obtenus en utilisant la stratégie de l'ensemble test. Il est clair qu'il y a une très mauvaise reconnaissance de la classe 2; par contre les classes 1 et 3 sont assez bien reconnues.

4.2.2. Résultats de la discrimination en 2 classes

La variable à expliquer comprend maintenant uniquement les classes 2 et 3 de la typologie. L'ensemble des variables explicatives est identique à l'ensemble précédent.

		Groupes a priori	
		2	**3**
Classes d'affectation	**2**	362	206
	3	325	395

Tab 5 : Tableau de classement sur l'ensemble d'apprentissage

Le taux de bon classement apparent de la classe 2 est de 52.7% , celui de la classe 3 est de 65.7%

		Groupes a priori	
		2	**3**
Classes d'affectation	**2**	87	55
	3	88	92

Tab 6 : Tableau de classement sur l'ensemble test

Le taux de bon classement de la classe 2 est de 49.7%, celui de la classe 3 est de 62.6%. En utilisant la stratégie de la validation croisée le taux de bon classement de la classe 2 est égal à 50.4%, celui de la classe 3 est égal à 64.0%.

		Groupes a priori	
		2	**3**
Classes d'affectation	**2**	410	287
	3	403	510

Tab 7 : Tableau de classement par la validation croisée

Comme pour la discrimination précédente des taux obtenus sur l'ensemble test et ceux obtenus par la validation croisée sont très proches. Dans ce cadre nous avons une bonne séparation entre les deux classes, cependant nous ne pouvons expliquer le faible taux de reconnaissance de la classe 2 dans la première discrimination

4.3. Les principaux résultats par la segmentation

En sélectionnant le critère associé à la distance de Kolmogorov-Smirnov nous avons transformé toutes les variables qualitatives sélectionnées en variables binaires en utilisant le codage disjonctif complet.

Ce codage crée une variable binaire pour chaque modalité de la variable qualitative à coder. Par exemple la variable Overdose ayant trois modalités (Oui, Non et Non Réponse) a été transformée en trois variables binaires : Overdose = Oui, Overdose = Non et Overdose = Non Réponse.

4.3.1. Résultats de la discrimination en 3 classes

Après l'élagage la fonction de décision construit 8 segments terminaux et l'arbre de décision est le suivant :

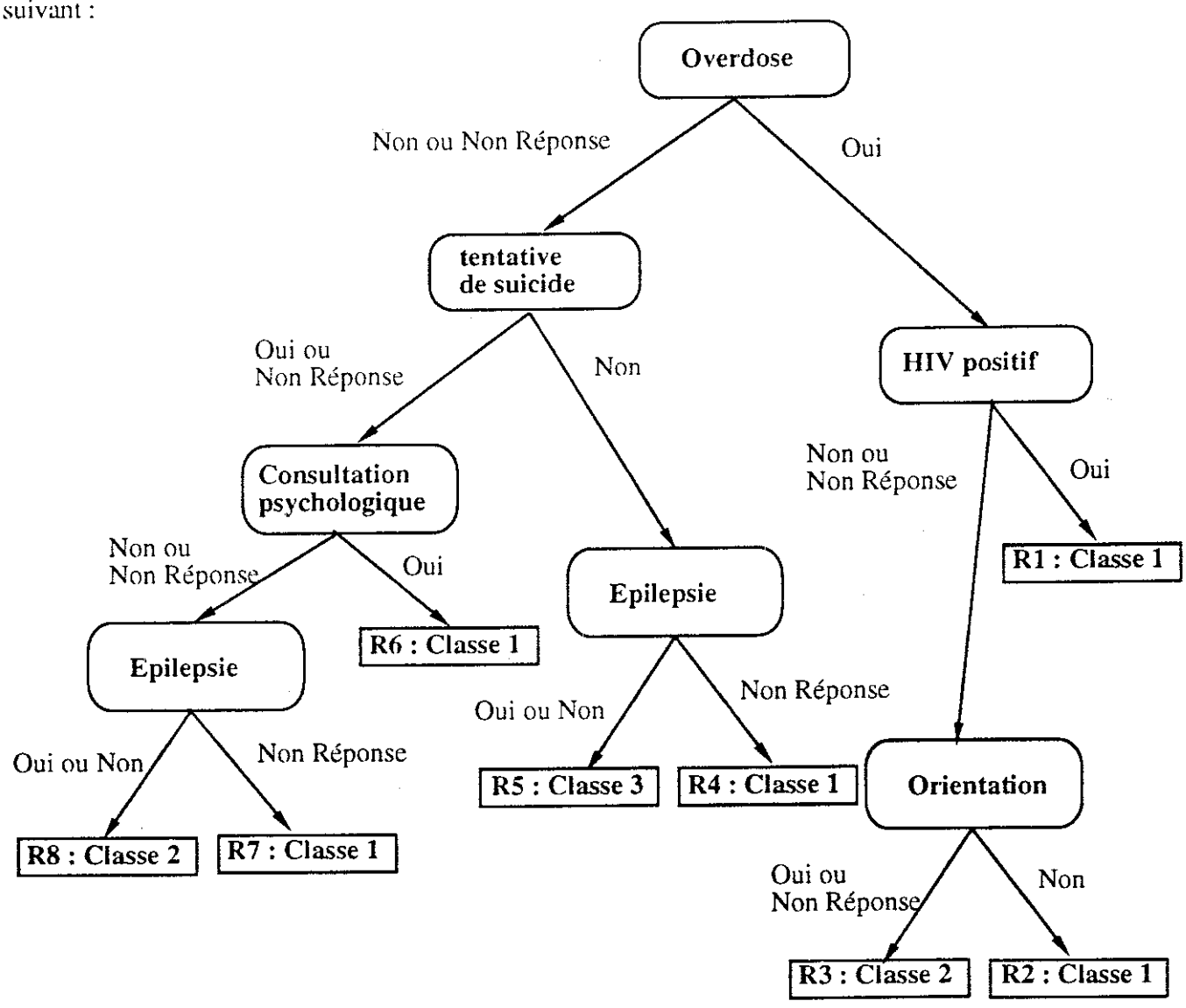

Fig 7 : Arbre binaire de la discrimination en 3 classes

Les segments R1, R2, R4, R6 et R7 caractérisent la classe 1; les segments R3 et R8 caractérisent la classe 2; le segment R5 caractérise la classe 3.

Pour la classe 1 nous avons les règles de production suivantes :

Règle 1 associée au segment R1:
[Overdose = Oui]
ET [HIV positif = Oui]
Cette règle est vérifiée par 14% des individus de la classe 1.

Règle 2 associée au segment R2 :
[Overdose = Oui]
ET [[HIV positif = Non] OU [HIV positif = Non Réponse]]
ET [Orientation = Non]
Cette règle est vérifiée par 12% des individus de la classe 1.

Règle 4 associée au segment R4 :
[[Overdose = Non] OU [Overdose = Non Réponse]]
ET [Tentative de suicide = Non]
ET [Epilepsie= Non Réponse]

Règle 6 associée au segment R6 :
[[Overdose = Non] OU [Overdose = Non Réponse]]
ET [[Tentative de suicide = Oui] OU [Tentative de suicide = Non Réponse]]
ET [Consultation = Oui]

Règle 7 associée au segment R7 :
[[Overdose = Non] OU [Overdose = Non Réponse]]
ET [[Tentative de suicide = Oui] OU [Tentative de suicide = Non Réponse]]
ET [[Consultation = Non] OU [Consultation = Non Réponse]]
ET [Epilepsie= Non Réponse]
Les règles 4, 6 et 7 sont chacune vérifiées par environ 10% des individus de la classe 1.

Pour la classe 2 nous avons les règles suivantes :

Règle 3 associée au segment R3 :
[Overdose = Oui]
ET [[HIV positif = Non] OU [HIV positif = Non Réponse]]
ET [[Orientation = Oui] OU [Orientation = Non Réponse]]

Règle 8 associée au segment R8 :
[[Overdose = Non] OU [Overdose = Non Réponse]]
ET [[Tentative de suicide = Oui] OU [Tentative de suicide = Non Réponse]]
ET [[Consultation = Non] OU [Consultation = Non Réponse]
ET [[Epilepsie = Oui] OU [Epilepsie= Non]]
Les règles 3 et 8 sont chacune vérifiées par environ 10% des individus de la classe 2.

Pour la classe 3 nous avons la règle suivante :

Règle 5 associée au segment R5 :
[[Overdose = Non] OU [Overdose = Non Réponse]]
ET [Tentative de suicide = Non]
ET [[Epilepsie = Oui] OU [Epilepsie= Non]]
Cette règle représente la classe d'affectation la plus importante.

La performance de la méthode est mesurée par les tableaux suivants :

		Groupes a priori		
		1	2	3
Classes d'affectation	1	411	281	206
	2	133	151	122
	3	202	235	299

Tab 8 : Tableau de classement sur l'ensemble d'apprentissage

Le taux de bon classement apparent de la classe 1 est de 55 % , celui de la classe 2 est 22.6 % et celui de la classe 3 est 47.7 % .

		Groupes a priori		
		1	2	3
Classes d'affectation	1	101	62	62
	2	40	23	24
	3	53	61	84

Tab 9 : Tableau de classement sur l'ensemble test

Le taux de bon classement de la classe 1 est de 52.1 %, celui de la classe 2 est 15.7% et celui de la classe 3 est de 49.4 %. La performance de cette méthode est identique à la méthode précédente. De même la stratégie décrite en 3.3, permettant d'utiliser l'analyse discriminante linéaire, donne des résultats semblables. Cependant la segmentation donne une explication à la non reconnaissance de la classe 2.

La population du segment R6, aussi bien obtenue par l'ensemble d'apprentissage que par l'ensemble test, est composée d'individus de la classe 1 et 2 dans la même proportion. La proportion de la classe 1 étant légèrement plus grande alors ce segment est affecté à la classe 1. La segmentation de R6 n'est pas possible à partir des variables du volet des données médicales; ce partage serait facile en introduisant l'age.

4.2.2. Résultats de la discrimination en 2 classes :

Les segments S1, S2 et S3 caractérisent la classe 2 ; le segment S4 caractérise la classe 3. %2 est la fréquence qu'un individu appartenant à la classe 2 soit affecté à cet segment terminal, il en est de même pour %3 vis à vis de la classe 3.

Pour la classe 2 nous avons les règles de production suivantes :

Règle 1 associée au segment S1:
[Overdose = Oui]

Règle 2 associée au segment S2 :
[[Overdose = Non] OU [Overdose = Non Réponse]]
ET [HIV =Positif]

Règle 3 associée au segment S3:
[[Overdose = Non] OU [Overdose = Non Réponse]]
ET [[HIV = Non Positif] OU [HIV = Non Réponse]]
ET [Consultation =Oui]

Pour la classe 3 nous avons les règles suivantes :

Règle 4 associée au segment S4:
[[Overdose = Non] OU [Overdose = Non Réponse]]
ET [[HIV = Non Positif] OU [HIV = Non Réponse]]
ET [[Consultation =Non] OU [Consultation =Non Réponse]]

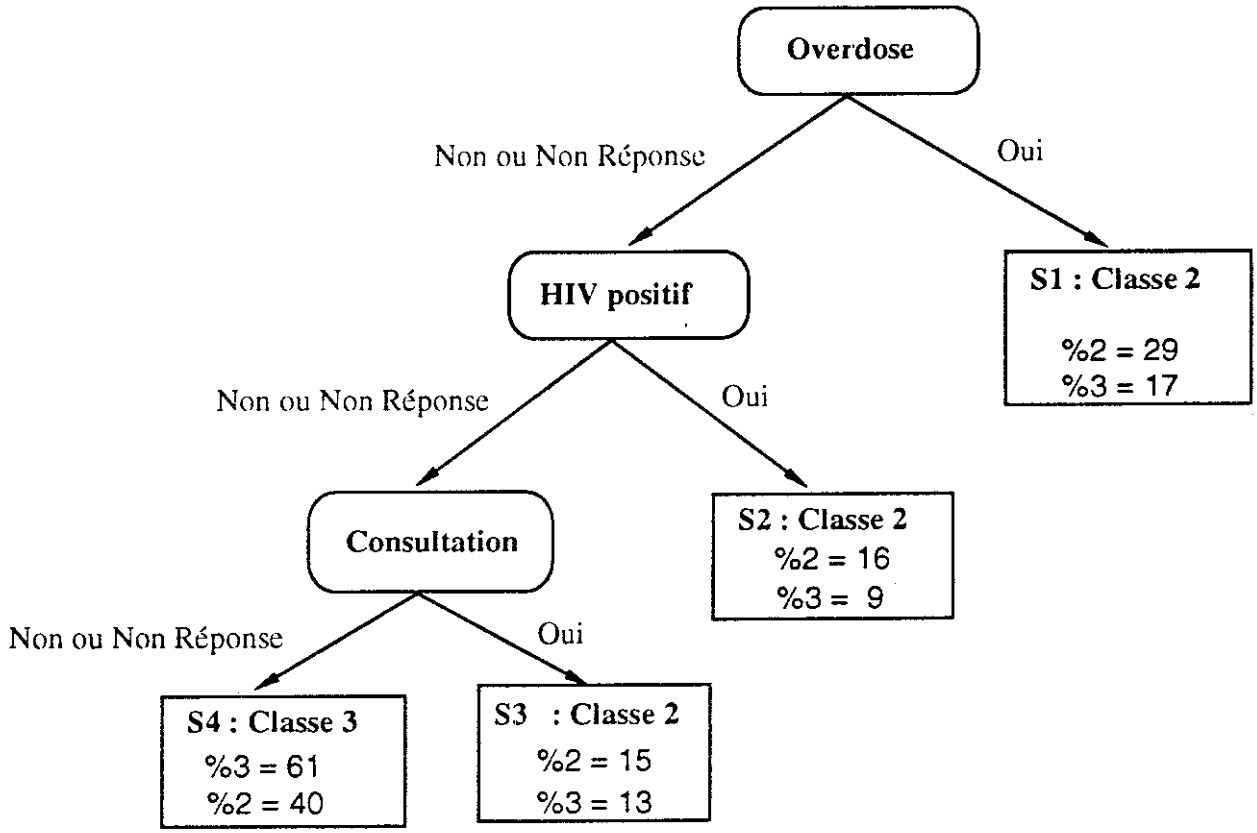

Fig 8: Arbre binaire de la discrimination en 2 classes

La performance de la méthode est mesurée par les tableaux suivants :

		Groupes a priori	
		2	**3**
Classes d'affectation	**2**	388	258
	3	252	387

Tab 10 : Tableau de classement sur l'ensemble d'apprentissage

Le taux de bon classement apparent de la classe 2 est de 60.6% , celui de la classe 3 est de 60%

		Groupes a priori	
		2	**3**
Classes d'affectation	**2**	93	59
	3	72	98

Tab 11 : Tableau de classement sur l'ensemble test

Le taux de bon classement de la classe 2 est de 56.4% et celui de la classe 3 est de 61.4% Nous pouvons observer que la performance de cette méthode est identique à la méthode de discrimination par modèle d'indépendance.

Dans notre étude le comportement de trois méthodes de discrimination utilisées est très semblable. Il faut noter que le fait d'ajouter la classe 1 entraîne une diminution importante du taux global de reconnaissance. Cependant cette diminution n'est pas équivalente pour chaque classe; la classe 3 possède toujours le même pouvoir d'identification par contre la classe 2 ne peut plus être reconnue. Ceci est dû au comportement hétérogène des individus de la classe 1 vis à vis des classes 2 et 3, mais l'arbre de décision peut en donner l'interprétation suivante.

L'arbre de décision marque l'hétérogénéité de la classe 1 en lui affectant 5 segments terminaux. La variable Overdose est très importante pour séparer la classe 2 de la classe 3; par contre elle ne caractérise pas la classe 1 car les nombreux segments associés à cette classe se répartissent équitablement entre la partie gauche et droite de l'arbre de décision. L'ensemble des individus ayant répondus positivement à la question Overdose (partie droite de l'arbre) de la classe 1 sont difficilement séparables des individus de la classe 2 ayant répondus de la même manière à cette question.

Dans la partie gauche de l'arbre, formée des individus ayant répondus négativement à la question Overdose, se trouvent deux ensembles d'individus de la classe 1. Le premier ensemble, représenté par le segment R4, s'oppose à la classe 3 (segment R5) en utilisant la question concernant l'épilepsie. Le deuxième ensemble, représenté par les segments R6 et R7, est difficilement discriminé des individus de la classe 2 (segment R8) par les deux questions associées à la consultation et à l'épilepsie.

5. CONCLUSION

5.1. Sur les résultats

La typologie montre l'existence de sous-groupes bien différenciés sur un plan socio-démographique et pénal dans une population de sujets incarcérés repérés comme usagers de drogue. La discrimination complète la typologie dans la recherche explicative de comportement d'usage de drogue et de délinquance.

Les comparaisons de classement des sujets par ces différentes méthodes montre un certain "pouvoir" de prédiction des variables médicales liées à la toxicomanie par rapport aux conduites de récidive délictueuse, pour les classes 2 et 3, quand l'âge est contrôlé. Plus les difficultés de santé liées à la toxicomanie sont présentes avec consultation psychologique ou psychiatrique, plus il y a de risque de récidive, entre 20 et 24 ans.

5.2. Sur les méthodes

Les méthodes de classement complètent la classification, par la recherche de règles d'affectation. Du fait du taux de bon classement obtenu, on peut supposer que l'introduction d'autres variables améliorerait la discrimination, à partir de données sociales ou psychologiques par exemple [DoF 91].

Par ailleurs, il serait intéressant de poursuivre cette étude dans une direction explicative, en prenant compte la chronologie des difficultés médicales et des problèmes judiciaires. L'introduction de données chronologiques poserait d'autres problèmes pour l'application des analyses discriminantes, peut être identiques à ceux rencontrés lors des analyses typologiques [FaG 84] De façon générale, la recherche typologique, éprouvée par une méthode explicative, constitue ainsi une démarche de l'épidémiologie dans la définition d'indices ou d'échelle de gravité des comportements, bases indispensables pour toute action de prévention.

6. BIBLIOGRAPHIE

[BrF 84] Breiman L., Friedman J.H., Ohlsen R.A., Stone C.J. (1984) - *Classification and Regression Trees*. Wadsworth

[BrT 70] Bouroche J.P. et Tenenhaus N. (1970) - *Quelques méthodes de segmentation*. RAIRO 4.2, pp 29-42

[DoF 91] Dormas O., Facy F. et Vilamont B. (1991) - *Toxicomanes incarcérés : approche psychométrique*, Revue Européenne de Psychologie Appliquée, vol 41, n°2 pp 85-91.

[CeL 82] Celeux G., et Lechevallier Y. (1982) - *Méthodes de segmentation non paramétriques*. R.S.A. Vol. n° 30 n° 4, pp 39-53.

[CeL 90] Celeux G. et Y. Lechevallier (1990) - *Arbre segmentation*. Dans : *Analyse discriminante sur variables continues*. Collection Didactique 7, INRIA.

[CDG 89] Celeux G., Diday E., Govaert G., Y. Lechevallier, et H. Ralambondrainy (1989) - *Classification automatique des données, Environnement statistique et Informatique*. Dunod, pp 237-242.

[Fac 91] F. FACY et coll. (1991) - *Toxicomanes incarcérés* , Rapport INSERM-DGS-91 Paris.

[Fri 77] Friedman J.H. (1977) - *A recursive partitioning decision rule for non parametric classification*. IEEE Trans. on Comp C 26-4, pp 404-408.

[Gue 88] Gueguen A. et Nakache J.P. (1988) - *Méthode de discrimination basée sur la construction d'un arbre de décision binaire*, R.S.A. 36 n°1, pp 19-38.

[MIN 89] Ministère de la santé (1989). SESI - *Enquête toxicomanie*, Nov. 89.

[MoS 63] Morgan J.N. et Sonquist J.A (1963) - *Problems in the analysis of survey data and a proposal.*, J. Amer. Statist. Assoc , 58, pp 415-435.

[MOD 87] MODULAD (1987) - *Bibliothèque Fortran pour l'analyse des données* . INRIA.

[OCR 91] OCRTIS (1991) - *Usage et trafic de drogues en France*, Statistiques de l'année 1991. Ministère de l'Intérieur

[Min 89] Mingers J. (1989) - *An empirical comparison of Selection measures for Decision-tree induction. Machine Learning 3*, pp 319-342.

[Qui 86] Quinlan J.R. (1986). - *Induction of decision trees*. Machine Learning 1, pp 81-106

[Mkh 91] Mkhadri A.(1991) - *Classification et discrimination des données qualitatives : Discrimination Multinomiale Régularisée*, Thèse de l'Université Paris 6.

[FaG 84] Facy F., Govaert G. et Laurent F. (1984) - *Analyse typologique de données chronologiques*, R.S.A. Vol 32, N° 3.