International Journal of Innovation and Applied Studies
ISSN 2028-9324 Vol. 33 No. 1 Jun. 2021, pp. 105-121

http://www.ijias.issr-journals.org/

# Contribution à l'analyse factorielle discriminante des données symboliques: Application aux maladies cardiovasculaires

***Joseph Kasiama Ngi-Onkor[1], Rostin Mabela Matendo[2], Ruffin-Benoît M. Ngoie[3], and Jean Jacques Katshitshi[4]***

[1]Mathématiques et Info., Fac. Sc., U.P.N., KIN I, RD Congo

[2]Dpt. Mathématiques et Info., Fac. Sc., Université de Kinshasa, RD Congo

[3]Dpt. Mathématiques et Info., Sec. Sc. Ex., I.S.P. MBANZA-NGUNGU, RD Congo

[4]Dpt. A.I.A., Fac. Lettres & Sc. Hum., Université de Kinshasa, RD Congo



**ABSTRACT:** This article suggests a generalized method of discrimination which extends the classical Discriminating Factorial Analysis (FDA) to symbolic objects. This method is based on the adaptation of the classical Bayesian rule of discrimination to symbolic objects. This adaptation is done taking into account various elements, namely: a certain « density » measure on the observation space of symbolic objects; discriminant functions giving an idea of the « similarity » which exists between an observation and the individuals of the formation whole. This rule depends on the formation data and is typically built of the in view the minimization of the overall error rate.

The purpose of this study is to solve a capital medical problem. Indeed, several cases of sudden deaths are noted these last years in the whole world and more particularly in our country the Democratic Republic of Congo (RDC), due to Cerebral Vascular Accidents (CVA) or acute coronary events (Heart attacks). The evolution and the prevailence of these cardiovascular diseases present a certain number of real and urgent problems to policy makers and other medical officials. Many undertaken epidemiologic studies these 20 last years led to the identification of the principal Factors of cardiovascular risk (FCVR), opening the way to preventive treatment. It is on the basis of these factors of risk that we designed a tool of decision-making aid medical.

This generalized method of discrimination therefore makes it possible to produce decisions concerning whether or not a data point belongs to a predefined class, by using formation sets, from an assigning algorithm of the symbolic objects to classes that we suggest here.

**KEYWORDS:** discrimination method, discriminant factoriel analysis, symbolic objects, density function, bayesian rule, decision rule, discriminant function, formation set, factors of cardiovascular risk.

**RESUME:** Cet article propose une méthode généralisée de discrimination qui étend l'Analyse Factorielle Discriminante (AFD) classique aux objets symboliques. Cette méthode est basée sur l'adaptation de la règle bayésienne classique de discrimination aux objets symboliques. Cette adaptation s'effectue en tenant compte de plusieurs éléments, notamment: une certaine mesure de « densité » sur l'espace d'observation des objets symboliques; des fonctions discriminantes donnant une idée de la « similarité » qui existe entre une observation et les individus de l'ensemble de formation. Cette règle dépend des données de formation et est typiquement construite en vue de la minimisation d'un taux d'erreur global.

Le but de cette étude est de résoudre un problème médical capital. En effet, plusieurs cas de morts subites sont constatés ces dernières années dans le monde entier et plus particulièrement dans notre pays la République Démocratique du Congo (RDC), dues aux Accidents Vasculaires Cérébraux (AVC) ou aux Evénements coronariens aigus (Crises cardiaques). L'évolution et la prévalence de ces maladies cardiovasculaires présentent un certain nombre de problèmes réels et urgents aux décideurs politiques et autres responsables médicaux. De nombreuses études épidémiologiques menées ces 20 dernières années ont conduit à l'identification des

principaux Facteurs de risque cardiovasculaire (FRCV), ouvrant la voie du traitement préventif. C'est sur base de ces facteurs de risque que nous avons conçu un outil d'aide à la décision médicale.
Cette méthode généralisée de discrimination permet donc de produire des décisions concernant l'appartenance ou non d'un point de données à une classe prédéfinie, en utilisant des ensembles de formation, à partir d'un Algorithme d'assignation des objets symboliques à des classes que nous proposons ici.

**MOTS-CLEFS:** méthode de discrimination, analyse factorielle discriminante, objets symbolique, fonction de densité, règle bayésienne, règle de décision, fonction discriminante, ensemble de formation, facteurs de risque cardiovasculaire.

## 1 INTRODUCTION

Les bases de données qui se développent partout dans le monde prenant parfois des tailles gigantesques contiennent des concepts sous-jacents. Ces derniers sont associés aux catégories issues de produits cartésiens de variables qualitatives ou de classifications automatiques. Ces concepts constituent alors des unités d'étude d'un niveau de généralité supérieur aux données initiales (classiques) qui ne tiennent pas compte de la variation des instances de ces concepts.

Les principales méthodes de traitement utilisées par Data Mining sont la classification automatique, la hiérarchisation, le partitionnement, les arbres de décision, l'analyse factorielle, etc. Elles sont appliquées à des données numériques « classiques » - quantitatives ou qualitatives (prix, âge, rang, adresse, etc.). L'unité statistique de premier niveau de l'analyse de données numérique est l'individu avec ses différentes propriétés.

L'analyse des données symboliques s'applique à des données plus complexes – quantitatives, qualitatives, mais aussi des variables à valeurs multiples (couleur – rouge, noir, vert, etc.), de type intervalle (tranches d'âge), munies d'une conjonction de règles et taxonomies [Did87], [Did95]. L'analyse des données symboliques passe de l'individu à une unité de niveau supérieur – le concept, tout en conservant la variation interne des individus. Chaque individu est associé à un concept (par exemple une personne est associée à la ville dans laquelle elle habite). Un concept est défini par son intention (ses propriétés caractéristiques) et son extension (l'ensemble des individus qui satisfont ces propriétés). Les concepts sont modélisés à l'aide d'objets symboliques et organisés dans des tableaux de données symboliques [4], [12].

Afin d'agréger les données, l'analyse symbolique génère des descriptions des classes d'individus obtenues par généralisation. Les objets symboliques permettent de lier les descriptions aux classes qu'elles résument.

Les objets symboliques permettent aussi de décrire les concepts par leurs propriétés communes et de calculer leurs extensions (si la ville est un concept, l'ensemble des villes est son extension).

Un tableau de données symboliques autorise plusieurs valeurs par case, ces valeurs étant parfois pondérées et liées entre elles, ainsi que des intervalles et des histogrammes.

Le but de l'analyse factorielle discriminante sur des objets symboliques [1] est de décrire et visualiser sur les plans factoriels les relations entre un ensemble de prédicteurs (variables) et une variable classificatoire qui identifie une partition d'objets dans des groupes (classes), de définir et valider une règle discriminante de décision (règle de classification) basée sur des combinaisons linéaires des variables de prédicteurs afin de classifier un nouvel objet symbolique dans la classe d'adhésion/appartenance. La méthode de AFD symbolique est basée sur une procédure numérique-symbolique-numérique qui consiste en une transformation numérique des descripteurs d'objets symboliques et une interprétation symbolique des résultats.

L'idée générale de cette étude est de construire, à partir d'une base de données relationnelle, un tableau de données symboliques muni éventuellement de règles et de taxonomies. Le but étant de décrire des concepts résumant un vaste ensemble de données et d'analyser ensuite ce tableau pour en extraire des connaissances par des méthodes d'analyse des données symboliques.

Dans la cadre de cet article, nous allons nous appuyer sur le logiciel SODAS ([5], [13]) pour extraire les données concentrées dans une base de données relationnelle de type ACCESS, afin d'appliquer les méthodes d'analyse des données symboliques, ici l'AFD symbolique.

La démarche symbolique d'une analyse de données avec SODAS suit les étapes suivantes:

a. Disposer d'une base de données relationnelle (ORACLE, ACCESS, FOXPRO, …).

b. Définir ensuite un contexte par:

- Des unités statistiques de premier niveau (habitants, familles, entreprises, accidents, …);
- Des variables qui les décrivent;
- Des concepts (villes, groupes socio-économiques, scénario d'accident, …).

Chaque unité statistique du premier niveau est associée à un concept (par exemple, chaque habitant est associé à sa ville). Ce contexte est défini par une requête de la base.

c. Construire un tableau de données symboliques dont les nouvelles unités statistiques sont les concepts décrits par généralisation des propriétés des unités statistiques de premier niveau qui leur sont associés.

Ainsi, chaque concept est décrit par des variables dont les valeurs peuvent être des histogrammes, des intervalles, des valeurs uniques (éventuellement munies de règles et de taxonomies) etc., selon le type de variables et le choix de l'utilisateur.

Le module DB2SO (Data Base TO Symbolic Objects) permet d'extraire ce tableau de données symboliques à partir de la base de données relationnelle.

d. Créer un fichier d'objets symboliques sur lequel des méthodes d'analyse des données symboliques peuvent s'appliquer dans le logiciel SODAS (histogrammes des variables symboliques, classification automatique, analyse factorielle, analyse discriminante, visualisations graphiques,…).

Cet article est organisé de la façon suivante: la section 2 propose un algorithme d'assignation des objets symboliques à des classes, basé sur les fonctions de densité et l'adaptation de la règle bayésienne classique de discrimination aux objets symboliques. La section 3 est consacrée à la construction d'une base de données relationnelle sous le logiciel MS-ACCESS 2013. Nous présentons d'abord la structure de la base, suivie de la description des variables d'étude. Nous terminons cette section par la création des requêtes SQL. La section 4 concerne la construction d'un tableau de données symboliques à partir de la base de données relationnelle, le module DB2SO permet d'extraire ce tableau. La section 5 présente les résultats de l'AFD symbolique. Enfin, la section 6 donne l'interprétation des résultats de l'analyse tout en faisant ressortir les connaissances, avant de conclure cet article.

## 2 MÉTHODE GÉNÉRALISÉE DE DISCRIMINATION

### 2.1 PROBLÈME

Chaque individu $u \epsilon E$ est décrit par p variables symboliques $Y_1, \ldots, Y_p$ avec des domaines (espaces d'observation) Yl, …, YP. Les variables $Y_j$ peuvent être quantitatives, qualitatives, multivaluées, probabilistes, types intervalles, etc., tel que la gamme Bj de $Y_j$ est généralement plus complexe que dans le cas classique (où Bj est identique à l'espace d'observation Yj, par exemple, Bj = Yj = $\mathbb{R}$ ou {0, 1}).

Alors, les valeurs du vecteur des données symboliques $X \coloneqq (Y_1, \ldots, Y_p)$ appartiennent au produit cartésien B = $\mathop{\mathbf{X}}_{j=1}^{p}$ Bj (qui, dans le cas classique, peut être identique, par exemple, à $\mathfrak{X} = \mathbb{R}^p$ ou $\{0, 1\}^p$).

L'analyse discriminante essaye de donner une solution au problème suivant [1]:

Etant donné g échantillons, $G_1, \ldots, G_g$ de Y appartenant à g classes $\Pi_1, \ldots, \Pi_g$, et soit B l'espace de formation, avec $G_k: x_{k1}, \ldots, x_{kn_k} \epsilon$ B de $\Pi_k$ (1)

Le $k^{ème}$ échantillon de formation de taille $n_k$, $k = 1, \ldots, g$.

Nous voulons utiliser ces "ensembles de formation" pour assigner à une de ces classes un seul nouveau point de données $x \epsilon$B. Comment pouvons-nous accomplir cette tâche ?

### 2.2 DÉMARCHE

La règle bayésienne classique de discrimination donne une solution au problème d'analyse discriminante pour des données classiques. Comme pour la plupart des méthodes de classification, cette règle de discrimination a besoin d'être adaptée au problème d'objets symboliques [Did 87, 88, 89], dans lequel nous pouvons avoir des informations de différents types: quantitatif, qualitatif,

probabiliste, type intervalle… au sein d'une conjonction des événements. Pour effectuer une telle adaptation, on a besoin d'une certaine mesure de " densité " sur l'espace $\mathfrak{X}$ d'observation des objets symboliques et des fonctions discriminantes.

L'estimation de densité des grains est un outil qui permet au statisticien de construire une densité sur n'importe quel échantillon de données. Il existe plusieurs méthodes pour l'estimation de densité, notamment les méthodes paramétriques et non paramétriques ([6], [11], [2]).

Pour notre étude, nous allons utiliser la méthode paramétrique pour l'estimation de densité en considérant les matrices de covariance égales.

#### 2.2.1 RÈGLE DE DÉCISION ET FONCTIONS DISCRIMINANTES [1]

L'identification de la classe d'où une observation $x \in$ B tire son origine exige la définition d'une fonction d, appelée une règle de décision:

$$d: \mathrm{B} \longrightarrow \{1, \dots, g\}: x \longmapsto d(x) \quad (2)$$

Cette règle dépend des données de formation $x_{11}, \dots, x_{1n_1}, \dots, x_{g1}, \dots, x_{gn_g}$ et est typiquement construite en vue de la minimisation d'un taux d'erreur global.

Chaque règle d détermine une partition $(P_1, \dots, P_g)$ de l'espace B des valeurs de données possibles:

$$P_k = \{x \mid d(x) = k\}, 1 \leq k \leq g \quad (3)$$

La règle d sera souvent définie avec l'aide de g fonctions discriminantes $h_k(x)$, k = 1…, g. La fonction $h_k(x)$ est associée avec la $k^{ème}$ population $\Pi_k$ et donne une idée de la "similarité" qui existe entre x et les individus du $k^{ème}$ échantillon de formation: $G_k = \{x_{k1}, \dots, x_{kn_k}\}$.

Alors, d est spécifiée par la partition de B.

$$P_k = \{x \mid h_k(x) \geq h_i(x), \forall i = 1, \dots, g\}, 1 \leq k \leq g \quad (4)$$

S'il y a plus d'une classe maximisant $h_k(x)$:

$$\exists\, i \neq k: h_i(x) = h_k(x) \geq h_j(x)\ pour\ j = 1, \dots, g,$$

Quelques choix arbitraires sont possibles: L'individu x peut être assigné au hasard à une des classes $\Pi_k$, peut être assigné à la classe $\Pi_k$ avec index minimum k ou peut rester non classifié, etc..,

Pour des données classiques avec, par exemple, B = $\mathfrak{X} = \mathbb{R}^p$, les fonctions discriminantes peuvent être déterminées par des approches probabilistes. Dans le cas symbolique, seules des méthodes empiriques sont connues.

#### 2.2.2 LE CADRE PROBABILISTE CLASSIQUE [1]

Nous considérons le cas classique où les classes $\Pi_1, \dots, \Pi_g$ sont décrites par g densités de probabilité $f_1(x), \dots, f_g(x)$ pour X qui prend ses valeurs x dans $\mathfrak{X} = \mathbb{R}^p$. En affectant les données x à des classes différentes, quelques erreurs peuvent se produire.

Supposons que $p_{ki}(d), 1 \leq k \leq g$, Que $1 \leq i \leq g$, soit la probabilité qu'un individu (avec le vecteur de données X) qui appartient à la classe $\Pi_k$ est assigné à la classe $\Pi_i$. Étant donné que certaines erreurs peuvent être moins importantes que d'autres, un coût $c_{ki}, 1 \leq k \leq g, 1 \leq i \leq g$, peut être associé à l'assignation à la classe i d'une observation de la classe k. Nous supposons ici $c_{kk} = 0$, c.à.d., pas de coûts qui sont attribués à une classification correcte.

Pour chaque population $\Pi_k, 1 \leq k \leq g$, nous définissons le risque $R_k(d)$ comme la valeur attendue des coûts attribués à une fausse assignation pour des individus du groupe k:

$$\mathrm{R}_k(d) = \sum_{i=1}^{g} c_{ki}\, p_{ki}(d), 1 \leq k \leq g \quad (5)$$

Ainsi, nous obtenons un vecteur de risque de dimension g:

$$\vec{R}(d) = (R_1(d), \dots, R_g(d))' \quad (6)$$

La spécification d'une "meilleure" règle est un problème critique. Évidemment, nous pouvons exclure des règles de décision $d'$ pour lesquelles il existe une règle d uniformément meilleure dans le sens suivant:

Une règle $d'$ est dominée par d $\Leftrightarrow \begin{cases} \forall\, k\epsilon\,\{1,\dots,g\}\colon R_k\,(d)\ \leq R_k\,(d') \\ \exists\ \tilde{k}\epsilon\{1,\dots,g\}\colon R_{\tilde{k}}\,(d)\ \leq R_{\tilde{k}}\,(d') \end{cases}$ (7)

Par conséquent, nous restreignons l'ensemble D de toutes les règles de décision à $\widetilde{D}$, l'ensemble des règles d sans une règle de domination $d'$:

$$\widetilde{D} = \{d\ \epsilon\ D\colon \nexists\ \tilde{d}\ \epsilon\ D\ tel\ que\ \tilde{d}\ domine\ d\} \quad (8)$$

### 2.3 La Règle Bayésienne (De Thomas Bayes)

Supposons que dans les hypothèses énoncées à la section 2.2.2., nous connaissons les probabilités a priori $\boldsymbol{p_k}$, $1 \leq k \leq g$, pour X provenant de la population $\Pi_k$. Le risque bayésien est défini comme le coût global attendu [1]:

$R(d) = \sum_{k=1}^{g} p_k\, R_k(d)$ (9)

c.à.d., une moyenne pondérée des risques $R_1(d), \dots, R_g\,(d)$.

Il apparaît que:

$$\mathrm{R}_k(d) = \sum_{i=1}^{g} c_{ki}\, p_{ki}(d) = \sum_{i=1}^{g} c_{ki} \int_{P_i} f_k\,(x)dx \quad pour\ k = 1, \dots, g$$

et par conséquent:

$$R(d) = \sum_{k=1}^{g} p_k \sum_{i=1}^{g} c_{ki} \int_{P_i} f_k\,(x)dx = \sum_{i=1}^{g} \int_{P_i} [\sum_{k=1}^{g} p_k\, c_{ki} f_k(x)]dx$$

Ainsi, R (d) est minimal pour la partition $P = (P_1, \dots, P_g)$ donnée par:

$P_i = \{\, x\colon \sum_{k=1}^{g} p_k c_{ki}\, f_k(x) \leq \sum_{k=1}^{g} p_k c_{kj}\, f_k(x)\ pour\ j = 1, \dots, g\}$ (10)

Si les coûts de disclassification sont censés être égaux:

$$c_{ki} = c > 0\ \forall\ k, i = 1, \dots, g, k \neq i\ et\ c_{kk} = 0\ pour\ k = 1, \dots, g,$$

Nous avons:

$\sum_{k=1}^{g} p_k c_{ki}\, f_k(x) = c\left[\sum_{k=1}^{g} p_k f_k\,(x) - p_i f_i\,(x)\right]$ (11)

Par conséquent, nous trouvons (laissant tomber les termes constants):

$P_i = \{x \mid p_i f_i(x) \geq p_j f_j(x), \forall j = 1, \dots, g\}$, i =1, …, m (12)

Ceci prouve que dans ce cas, les fonctions discriminantes définissant la règle bayésienne sont données par:

$h_k(x) = p_k f_k\,(x)$, pour k = 1, …, g (13)

Ainsi, la probabilité a posteriori est donnée par:

$q_k(x) = \frac{p_k f_k(x)}{\sum_{i=1}^{g} p_i f_i(x)}$, pour $k = 1, \dots, g$ (14)

Une fois ce calcul effectué pour chacune des populations, le vecteur de données individuel xx pourra alors être affecté à la population $\Pi_k$ qui maximise la probabilité a posteriori $q_k(x)$.

### 2.4 Estimation De Densité [6], [11], [2]

Le problème de classification est résolu par une des règles de décision classique (ici la règle bayésienne) si les densités $f_k(x), 1 \leq k \leq g$, sont connues. En réalité, ces densités sont inconnues, mais comme énoncé au début, nous supposons que nous avons donné un échantillon $G_k$ de chaque population $\Pi_k$. Ces échantillons (avec $n = n_1 + \dots + n_g$ vecteurs de données en tout) seront utilisés pour estimer les g fonctions de densité.

Il existe plusieurs méthodes pour l'estimation de densité, notamment les méthodes paramétriques, les méthodes de grain uniforme et d'autres grains pour résoudre des problèmes impliquant des données classiques mélangées comprenant les composants

binaires et continus. Ces méthodes des grains sont dites non paramétriques. Pour notre étude, nous allons utiliser la méthode paramétrique pour l'estimation de densité en considérant les matrices de covariance égales.

En effet, en Analyse discriminante des données quantitatives classiques, des hypothèses paramétriques de normalité multivariable peuvent être faites. Les données de l'échantillon sont supposées être extraites de g populations multivariables normales de moyennes $\mu_k$ et avec des matrices de covariance égales $\sum_k = \Sigma\ (k = 1, \dots, g)$, de densité:

$$f_k(x) = \varphi(x;\ \mu_k, \Sigma) = \frac{1}{\sqrt{2\pi}^p\ |\Sigma|^{\frac{p}{2}}} e^{-\frac{1}{2}(x-\mu_k)\ '\Sigma^{-1}\ (x-\mu_k)} \quad (15)$$

Les ensembles de formation sont alors utilisés pour estimer les paramètres:

$$\hat{u}_k = \bar{x}_k = \frac{1}{n_k}\sum_{i=1}^{n_k} x_{ki}\ \ pour\ k = 1, \dots, g \quad (16)$$

$$S = \hat{\Sigma} = \frac{1}{n-g}\ \sum_{k=1}^{g}\sum_{i=1}^{n_k}\ (x_{ki} - \bar{x}_k)\ (x_{ki} - \bar{x}_k)\ ' \quad (17)$$

Ces estimateurs sont substitués en des densités $\varphi(x;\ \mu_k;\ \Sigma)$.

### 2.5 DÉTERMINATION DES PROBABILITÉS A PRIORI [1]

Au moyen de la construction des grains généralisés, le problème de discrimination peut être résolu pour des données symboliques par la règle bayésienne si les probabilités a priori appartenant aux classes $\Pi_k$ sont données. Si ces probabilités a priori ne sont pas données, quelques choix sont possibles. Des probabilités a priori égales peuvent être utilisées:

$$\hat{p}_k = \frac{1}{g} \text{ pour } k = 1, \dots, g \quad (18)$$

Nous pouvons aussi considérer les proportions observées dans les ensembles de formation donnés ($n = n_1 + \ \dots + n_g$ points de données symboliques à partir de $g$ classes). Alors, nous obtenons les probabilités a priori suivantes:

$$\hat{p}_k = \frac{n_k}{n} \text{ pour } k = 1, \dots, g \quad (19)$$

### 2.6 UN ALGORITHME D'ASSIGNATION DES OBJETS SYMBOLIQUES À DES CLASSES

Définissons d'abord quelques notations et terminologie utiles de cet algorithme:

$h_k\ (x)$: fonction discriminante associée à la $k^{ème}$ population $\Pi_k$, définissant la Règle de décision, k = 1…, g

$f_k\ (x)$: densité de probabilité associée à la $k^{ème}$ population $\Pi_k$, k = 1…, g.

$c_{ki}$: coût associé à l'assignation à la classe i d'une observation de la classe k, k, i = 1…, g

$c_{kk}$: coûts qui sont attribués à une classification correcte, k = 1…, g

$P_i$: partition de l'espace de formation B, $i = 1, \dots, g$

$p_k$: probabilité a priori d'appartenance d'un individu à la classe $\Pi_k$, $k = 1, \dots, g$

$\mu_k$: moyenne (ou centroïde) de la classe $\Pi_k$, $k = 1, \dots, g$

$\Sigma_k$: matrice de covariance de la classe $\Pi_k$, $k = 1, \dots, g$

$q_k\ (\mathrm{x})$: probabilité a posteriori d'appartenance d'un individu x à la classe $\Pi_k$, $k = 1, \dots, g$

En utilisant les ensembles de formations donnés ($n = n_1 + \ \dots + n_g$ points de données symboliques à partir de $g$ classes), notre algorithme calcule, pour classifier un vecteur de données symboliques x (input), les estimations $\hat{I}_k\ (x)$, k = 1…, g. Utilisant les probabilités a priori estimées également, nous obtenons les valeurs requises à maximiser:

$\hat{p}_k \hat{I}_k\ (x)$, pour $k = 1, \dots, g$

Par la normalisation de ces valeurs, nous avons finalement un ensemble de coefficients similaires aux probabilités a posteriori des classes. Les données output de notre logiciel comprennent l'ensemble de ces coefficients pour chaque point de données x à classifier.

**L'algorithme s'écrit alors:**

Soit g échantillons, $G_1, \ldots, G_g$ de Y appartenant à g classes $\Pi_1, \ldots, \Pi_g$, et soit B l'espace de formation, avec $G_k$: $x_{k1}, \ldots, x_{kn_k} \in$ B de $\Pi_k$ le $k^{ème}$ échantillon de formation de taille omoo $n_k$, $k = 1, \ldots, g$. Les probabilités ou les proportions théoriques de ces classes sont notées $p_1, \ldots, p_g$; dans cette optique bayésienne, il s'agit des probabilités a priori.

**A. Les étapes**

**Etape 0:** Prétraitement des données (Mise au pont)

- Ouverture du fichier des données symboliques (fichier de type.sds)
- Pour chaque échantillon $G_k = \{x_{k1}, \ldots, x_{knk}\}$ de $\Pi_k$, $k = 1, \ldots, g$

**Etape 1:** Calcul des probabilités a priori $p_k$

$$p_k = P(\Pi_k) = \frac{1}{n_k}, k = 1, \ldots, g$$

**Etape 2:** Calcul des moyennes $\mu_k$ et des matrices de covariance $\Sigma = \Sigma_k$

$$\mu_k = \bar{x}_k = \frac{1}{n_k}\sum_{i=1}^{n_k} x_{ki} \ pour \ k = 1, \ldots, g$$

$$S = \Sigma = \frac{1}{n-g}\sum_{k=1}^{g}\sum_{i=1}^{n_k} (x_{ki} - \bar{x}_k)(x_{ki} - \bar{x}_k)'$$

**Etape 3:** Calcul des densités de probabilité $f_k(x)$

$$f_k(x) = \varphi(x;\ \mu_k, \Sigma) = \frac{1}{\sqrt{2\pi}^p |\Sigma|^{\frac{p}{2}}} e^{-\frac{1}{2}(x-\mu_k)'\Sigma^{-1}(x-\mu_k)}$$

Si non fin de fichier, aller à l'étape 0.

**Etape 4:** Détermination de la partition $P_i$ de l'espace de formation B en considérant le Risque bayésien minimal:

$$P_i = \{ x: \sum_{k=1}^{g} p_k c_{ki} f_k(x) \ \leq \sum_{k=1}^{g} p_k c_{kj} f_k(x) \ pour \ j = 1, \ldots, g\}$$

Si $c_{ki} = c > 0 \ \forall \ k, i = 1, \ldots, g, k \neq i \ et \ c_{kk} = 0 \ pour \ k = 1, \ldots, g,$

alors :

$$\sum_{k=1}^{g} p_k c_{ki} f_k(x) = c\left[\sum_{k=1}^{g} p_k f_k(x) - p_i f_i(x)\right]$$

En laissant tomber les termes constants, on a:

$$P_i = \{x \mid p_i f_i(x) \geq p_j f_j(x), \forall j = 1, \ldots, g\}, i = 1, \ldots, g$$

**Etape 5:** Calcul des fonctions discriminantes $h_k(x)$ définissant la règle bayésienne

$h_k(x) = p_k f_k(x)$ pour k = 1, ..., g

**Etape 6:** Calcul des probabilités a posteriori $q_k(x)$:

$$q_k(x) = \frac{p_k f_k(x)}{\sum_{i=1}^{g} p_i f_i(x)}, \text{pour } k = 1, \ldots, g$$

La règle de Bayes peut-être interprétée de la façon qu'un vecteur de données individuel xx est assigné à la population $\Pi_k$ qui maximise la probabilité a posteriori.

**B. Calcul de la complexité**

| Etapes | Nombre d'opérations |
|---|---|
| Etape 0 | 1+ $g^2$ |
| Etape 1 | $g$ |
| Etape 2 | 3+ $4n_k g$ |
| Etape 3 | $13g$ |
| Etape 4 | 1+ $3g$ |
| Etape 5 | $2g$ |
| Etape 6 | $(2+2g)g$ |

Dès lors, en posant:

$$f(g) = 1 + g^2 + g + 3 + 4n_k g + 13g + 1 + 3g + 2g + (2+2g)g$$

$$f(g) = 3g^2 + 4n_k g + 21g + 5, \text{ où } k = 1, \dots, g$$

En posant: $g = n$;

On obtient une complexité donnée par:

$C(n) = 3n^2 + 4n\epsilon + 21n + 5 = O\ (n^2)$

## 3 CONSTRUCTION DE LA BASE DE DONNÉES

### 3.1 STRUCTURE DE LA BASE [8]

Pour notre étude, nous avons créé la base de données RCVPatients.mdb sous le logiciel Microsoft Access 2013. Il s'agit d'une base de données relationnelle construite à partir de données des tableaux 4.1, 4.2 et 4.3 issus de notre article intitulé: « Extraction de connaissances dans les données médicales par l'approche de l'analyse discriminante: Application au risque cardiaque et aux accidents vasculaires cérébraux », publié dans les Annales de la Faculté des Sciences de l'Université de KINSHASA, Volume 1 (2014), Editions Ita'yala Printer, KINSHASA, 2014.

Cette base de données contient des informations concernant 150 patients indemnes ou non de tout événement cardiovasculaire, prélevées dans le Registre de Laboratoire de Biologie clinique de l'Institut National de Recherche Biomédicale (INRB) et celles recueillies par nous même avec notre Médecine ambulatoire.

Ces données (informations) ont porté sur les paramètres (variables d'étude) appelées aussi Facteurs de Risque Cardio-Vasculaire (FRCV) [7], [9], [10].

Dans le schéma ci-dessous (Figure 1), nous avons un aperçu des différentes tables de notre base de données ainsi que la relation entre elles.

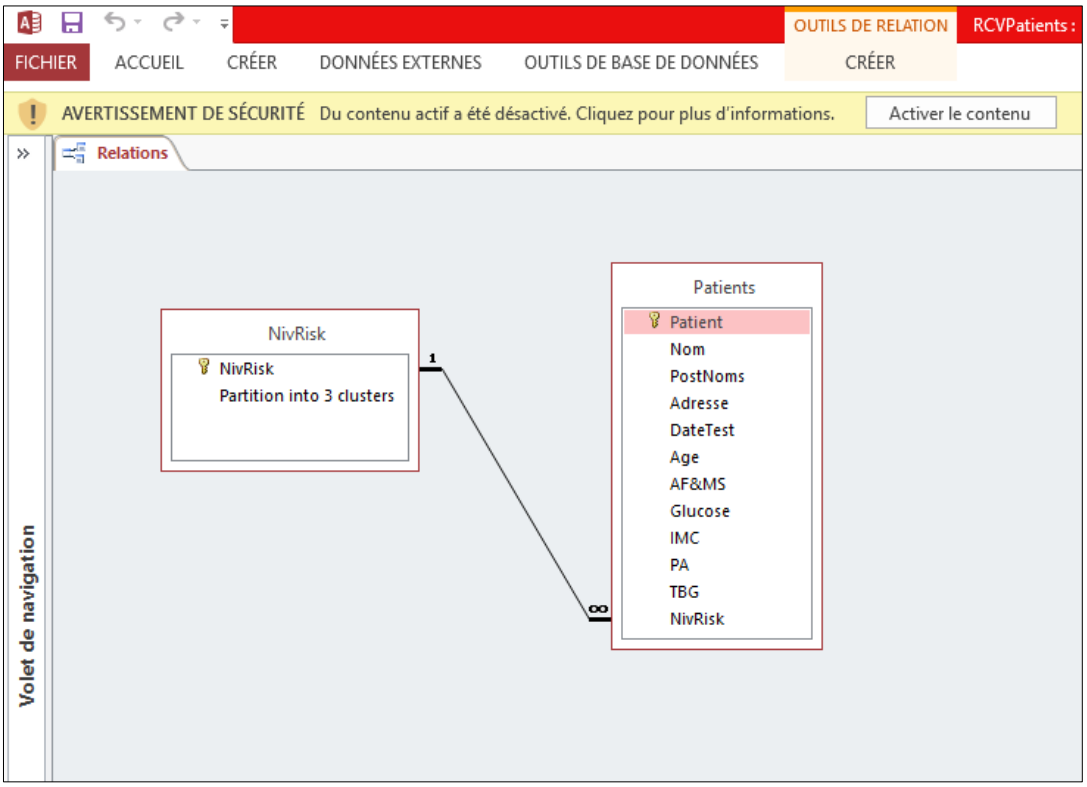



***Fig. 1. Schéma relationnel de la base de données***

Notre base est donc constituée de deux tables:

- La table Patients pour les individus; elle regroupe les informations telles que l'identifiant de l'individu, son adresse, la date de test, les résultats des examens cliniques sur les FRCV, etc. Elle forme un ensemble de 11 variables, dont 5 sont de type qualitative et 6 de type quantitative.
- La table NivRisk décrit les différents concepts; elle contient toutes les informations concernant les groupes à risque de patients que nous avons noté ici NivRisk (Niveau de Risque).

Vu la structure de notre base de données relationnelle, nous pouvons dégager facilement les concepts, les individus et les variables de descriptions choisis.

Nos concepts sont donc les groupes à risque des patients, les individus sont les patients et les variables de description sont les Facteurs de Risque Cardio-Vasculaire (FRCV) des patients et leurs identités.

### 3.2 VARIABLES ET REQUÊTES

#### 3.2.1 EXPLICATION DES CHAMPS DE DESCRIPTION (8(, (7(, (9(, (10(.

Les individus sont les patients qui sont indemnes ou non de tout événement cardio-vasculaire, ils sont représentés par leurs identifiants et leurs numéros d'ordre dans la base de données et sont décrits par les variables suivantes:

- Nom de l'individu
- PostNoms de l'individu
- Adresse de l'individu
- Date de test (DateTest) de l'individu
- Age: - Homme âgé de 50 ans ou plus
  - Femme âgée de 60 ans ou plus ou ménopausée
- Antécédents familiaux et/ou Mort-subite (AF&MS):
  - avant l'âge de 55 ans chez le père/frère
  - avant l'âge de 65 ans chez la mère/sœur

- Diabète traité ou non: la fourchette normale de glycémie (taux de glucose dans le sang) varie de 4,1 à 6,6 mmol/l.
- Obésité abdominale: le poids est idéal quand l'Indice de Masse Corporelle (IMC) est compris entre 19 et 25 kg/m. L'IMC est égal au poids (en kilos) divisé par la taille (en mètre) au carré, soit IMC = $\frac{Poids}{Taille^2}$
- Hypertension artérielle: une tension est considérée comme normale si la pression artérielle (PA) systolique est inférieure à 140 mmHg, et si la pression artérielle diastolique est inférieure à 90 mmHg. Soit 14/9 (cm de mercure).
- Tabagisme (TBG) en cours: la relation dose / effet (complications ischémiques) est continue et se manifeste dès la première cigarette quotidienne dans les études épidémiologiques puissantes. Même le tabagisme passif accroît le risque de complication vasculaire ischémique. D'où ne pas fumer !
- Cholestérol (CHLT): la fourchette normale de cholestérol total varie de 3,6 à 6,7 mmol/l.
- Niveau de risque (NivRisk) de l'individu.

N.B.: Notre Médecine ambulatoire n'ayant pas de matériel approprié pour tester le cholestérol, nous avons tenu compte de ce dernier seulement dans la constitution des groupes à risque homogènes (NivRisk). Dans le cas contraire (résultats de laboratoire), il doit faire partie intégrante des unités statistiques.

Nous avons 9 concepts qui sont constitués des différents niveaux de risque, répartis en 3 classes contenant chacune 3 groupes à risque. Les variables de description des concepts sont les suivantes:

- Haut risque
- Risque moyen
- Risque faible
- Likolo
- Katikati
- Nase
- Nazulu
- Nakatikati
- Nansi

Comme pour l'AFD classique où plusieurs individus composent une même classe, en AFD-OS on peut aussi avoir plusieurs objets symboliques dans une classe. En réalité, ces classes représentent respectivement ici les 3 groupes à risque: Haut risque, Risque moyen et Risque faible. C'est pourquoi nous avons dans chaque classe des objets comme: Likolo, (qui signifie Haut en Lingala), Nazulu (qui signifie Haut en Kikongo), Katikati (qui signifie Moyen en Lingala), Nakatikati (qui signifie Moyen en Kikongo), Nase (qui signifie Faible en Lingala) et Nansi (qui signifie Faible en Kikongo). Tous ces objets sont exprimés en langues nationales, pour éviter simplement la confusion des termes; et ces objets appartiennent à l'ensemble Test. Dans l'ensemble de formation nous avons les trois objets: Haut risque, Risque moyen et Risque faible.

### 3.2.2 CRÉATION DES REQUÊTES

Pour pouvoir, par la suite, utiliser notre base de données avec SODAS, il nous faut écrire sous Access des requêtes pour extraire l'information de la base et l'alimenter dans le fichier.SDS, fichier source pour les analyses statistiques.

Les requêtes utilisées sont au nombre de deux: la requête individu_concept et la requête concept_description.

**a) La requête individu_concept**

Cette requête va nous permettre de renvoyer les individus que nous avons choisis, définis comme individus de premier ordre, leurs caractéristiques, ainsi que les concepts associés. Nous obtenons ainsi un tableau, avec en première colonne l'individu, en seconde le concept, et ensuite les variables de descriptions souhaitées pour l'étude:

La construction de la requête individu_concept:

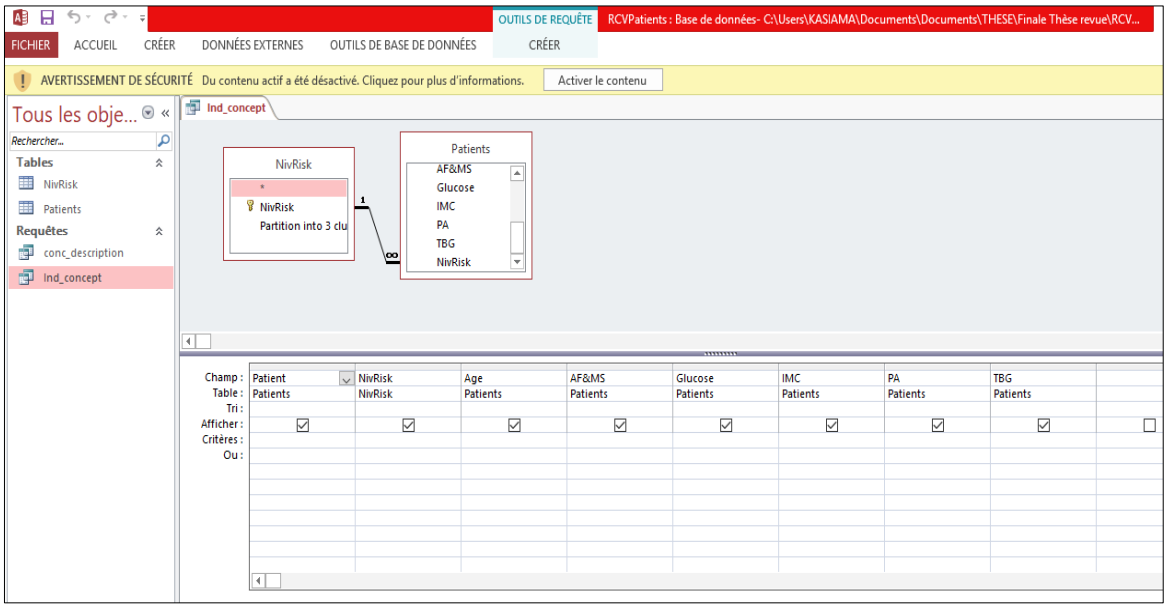



***Fig. 2. Requête individu_concept***

Voici en SQL la requête ci-dessous:

**SELECT** Patients.Patient, NivRisk.NivRisk, Patients.Age, Patients.AF&MS, Patients.Glucose, Patients.IMC, Patients.PA, Patients.TBG

**FROM** Patients

Cette requête donne le résultat suivant:

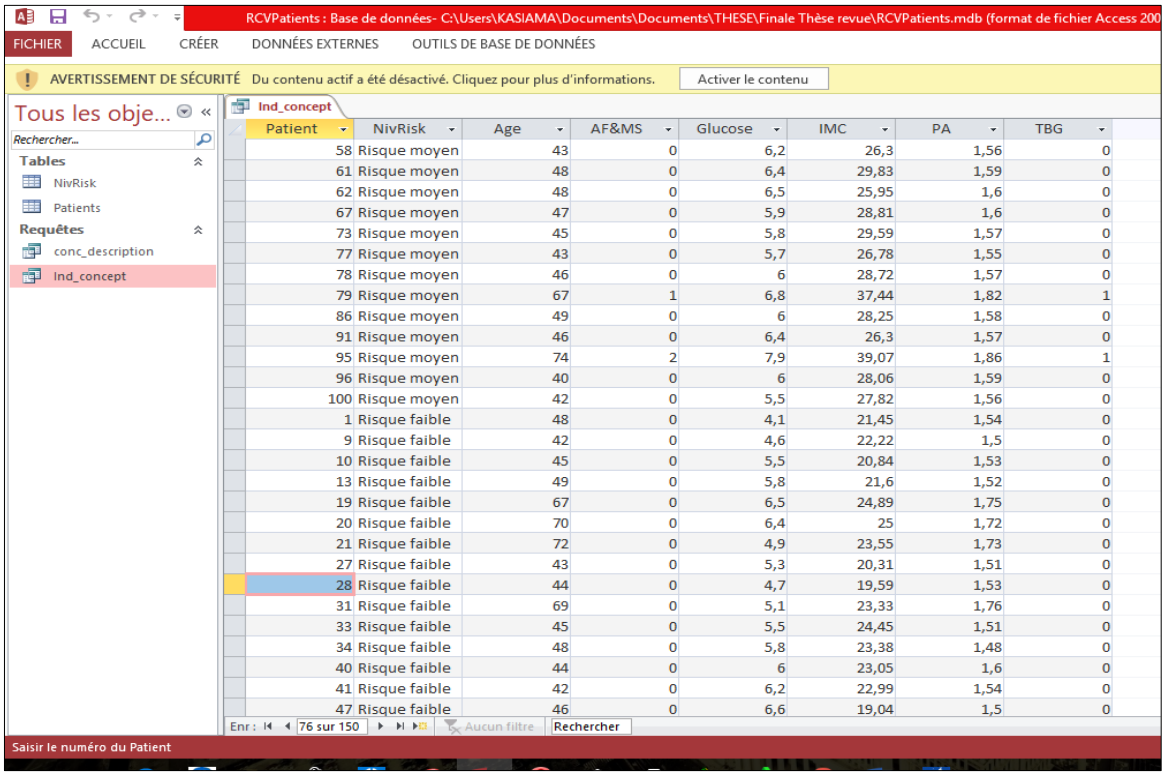



***Fig. 3. Résultat de la requête individu_concept***

N.B.: Comme vous pouvez le constater, nous avons laissé tomber les variables: Nom, PostNoms, Adresse et DateTest; cela ne va rien influencer, car l'individu est bien représenté par son numéro d'ordre en première colonne.

**b) La requête concept_description**

La requête concept_description permet d'ajouter des colonnes de description du concept dans SODAS. Elle permet ainsi de réaliser ce que l'on appelle des «add single».

La construction de la requête concept_description:

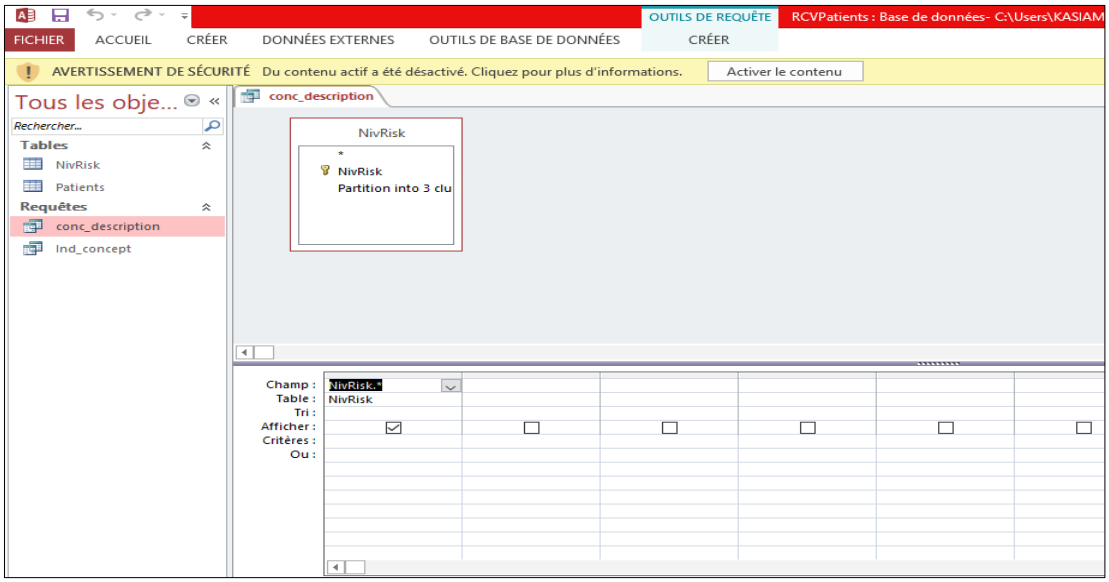



***Fig. 4. Requête concept_description***

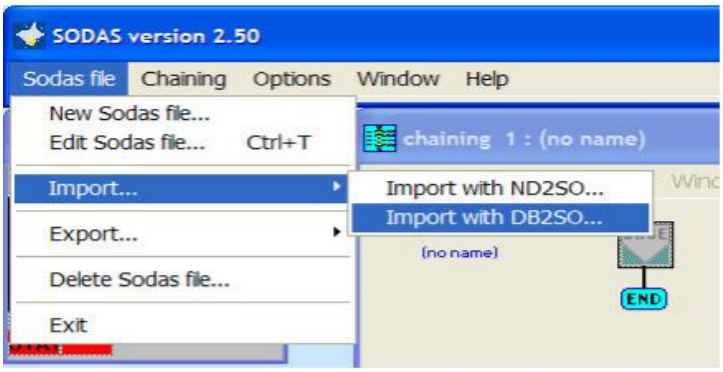
Voici en SQL la requête ci-dessous:

**SELECT** NivRisk.NivRisk

**FROM** NivRisk

## 4 GÉNÉRATION DES DONNÉES SYMBOLIQUES AVEC DB2SO

Le module DB2SO permet la génération d'un tableau des données symboliques à partir d'une base relationnelle. Il est accessible à partir du menu Sodas file → Import → DB2SO.




Ce SGBD inclut le driver ODBC permettant l'accès de DB2SO à la base de données relationnelle.

Ici, nous allons importer notre base de données ainsi que les deux requêtes créées précédemment dans DB2SO, afin de pouvoir utiliser SODAS pour analyser notre base. Ces deux requêtes nous ont donc permis de disposer les données de manière exploitable pour DB2SO, et par la même SODAS.

Donc, après avoir invoqué DB2SO dans SODAS et après avoir sélectionné les différentes fonctions (commandes) du logiciel ainsi que la Base de données RCVPatients.mdb et les deux requêtes, DB2SO crée un nouveau fichier SODAS de type *.sds (ici RCVPatients.sds). Ce fichier sera la base de toutes les applications SODAS.

Au final, le module DB2SO fournit une synthèse des éléments créés, le résultat est présenté dans la figure 5 ci-après:

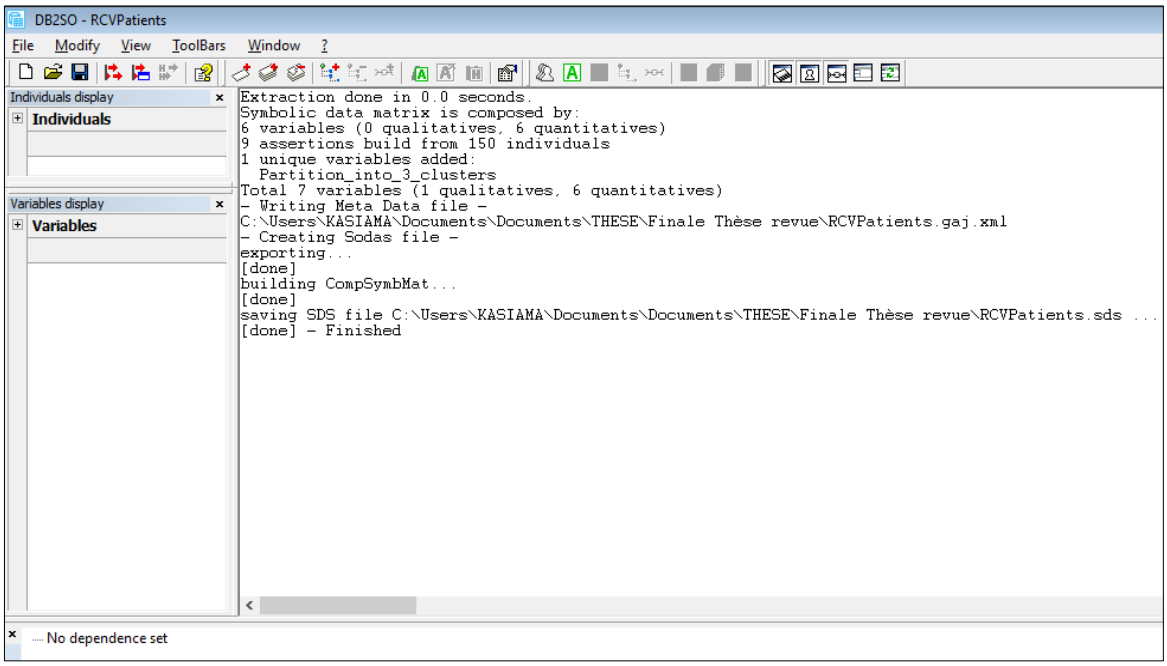



***Fig. 5. Résumé des données symboliques***

Nous étudions donc une population de 150 individus. Hormis les 4 variables que nous avons laissé tomber (Nom, PostNoms, Adresse et DateTest), les 7 variables dont nous disposons sont composées de 6 variables quantitatives et 1 variable qualitative.

Les variables quantitatives sont:

- Age
- Antécédents familiaux et/ou Mort-subite (AF&MS):
- Glucose
- Indice de Masse Corporelle (IMC)
- Pression artérielle (PA)
- Tabagisme (TBG)
- Niveau de risque (NivRisk) de l'individu.

Elles seront représentées sous SODAS par des variables intervalles.

L'unique variable qui reste est qualitative et est représentée sous SODAS en tant que variable modale. Il s'agit de:

- Partition_into_3_clusters

Le tableau suivant nous donne le contenu du tableau de données symboliques: les concepts et les variables qui les décrivent.

***Tableau 1. Extrait du tableau de données symboliques***

| | Age | AFMS | Glucose | IMC | PA | TBG | Partition_into_ |
|---|---|---|---|---|---|---|---|
| Haut risque | [ 50.00 : 75.00 ] | [ 1.00 : 2.00 ] | [ 6.70 : 8.60 ] | [ 17.96 : 46.08 ] | [ 1.58 : 1.85 ] | [ 1.00 : 1.00 ] | Clust |
| Risque moyen | [ 40.00 : 74.00 ] | [ 0.00 : 2.00 ] | [ 5.50 : 7.90 ] | [ 25.30 : 39.07 ] | [ 1.55 : 1.86 ] | [ 0.00 : 1.00 ] | Clust |
| Risque faible | [ 40.00 : 72.00 ] | [ 0.00 : 0.00 ] | [ 4.10 : 6.60 ] | [ 19.04 : 25.00 ] | [ 1.48 : 1.76 ] | [ 0.00 : 0.00 ] | Clust |
| Likolo | [ 58.00 : 79.00 ] | [ 0.00 : 2.00 ] | [ 4.20 : 8.50 ] | [ 18.29 : 46.08 ] | [ 1.42 : 1.89 ] | [ 1.00 : 1.00 ] | Clust |
| Katikati | [ 40.00 : 48.00 ] | [ 0.00 : 1.00 ] | [ 5.80 : 6.60 ] | [ 25.39 : 29.34 ] | [ 1.53 : 1.65 ] | [ 0.00 : 1.00 ] | Clust |
| Nase | [ 43.00 : 72.00 ] | [ 0.00 : 2.00 ] | [ 4.20 : 7.10 ] | [ 19.87 : 24.97 ] | [ 1.48 : 1.73 ] | [ 0.00 : 1.00 ] | Clust |
| Nazulu | [ 56.00 : 78.00 ] | [ 1.00 : 2.00 ] | [ 6.80 : 8.50 ] | [ 18.55 : 41.15 ] | [ 1.73 : 1.87 ] | [ 1.00 : 1.00 ] | Clust |
| Nakatikati | [ 40.00 : 44.00 ] | [ 0.00 : 0.00 ] | [ 6.10 : 6.60 ] | [ 26.12 : 29.34 ] | [ 1.57 : 1.58 ] | [ 0.00 : 0.00 ] | Clust |
| Nansi | [ 44.00 : 69.00 ] | [ 0.00 : 0.00 ] | [ 4.20 : 6.50 ] | [ 19.87 : 26.83 ] | [ 1.48 : 1.72 ] | [ 0.00 : 0.00 ] | Clust |

## 5 PRÉSENTATION DES RÉSULTATS DE L'AFD-OS

Dans notre analyse, nous nous sommes décidés de considérer tous les descripteurs de type intervalle (Age, AF/MS, Glucose, IMC, PA et TBG). Pour ces données, nous avons appliqué notre algorithme (logiciel) dont voici les résultats:

***Tableau 2. Les valeurs propres et % des sommes des carrées***

| N° | Inertia | Percentage of expl.inertia | Cumulated % of inertia |
|---|---|---|---|
| 1 | 0.02455 | 80.944 | 80.944 |
| 2 | 0.00502 | 16.555 | 97.498 |
| 3 | 0.00076 | 2.502 | 100.000 |

***Tableau 3. Les coordonnées d'intervalle des éléments d'OS***

| | Factor 1 | Factor 2 | Factor 3 |
|---|---|---|---|
| Haut risque | [-0.025508; 0.011760] | [-0.029862;-0.000141] | [-0.014470; 0.004285] |
| Risque moyen | [-0.011680; 0.038903] | [-0.027025; 0.024005] | [-0.006180; 0.016989] |
| Risque faibl | [-0.024481; 0.010325] | [ 0.014414; 0.030969] | [-0.000904; 0.011823] |
| Likolo | [-0.034739; 0.010371] | [-0.031599; 0.009037] | [-0.013996; 0.012885] |
| Katikati | [-0.004185; 0.038877] | [-0.011542; 0.023891] | [-0.004696; 0.012035] |
| Nase | [-0.029581; 0.031180] | [-0.016161; 0.029580] | [-0.012645; 0.011192] |
| Nazulu | [-0.023528; 0.008421] | [-0.030331;-0.004318] | [-0.012193; 0.005038] |
| Nakatikati | [ 0.008670; 0.015411] | [ 0.019102; 0.022468] | [ 0.009582; 0.011759] |
| Nansi | [-0.021919; 0.012613] | [ 0.014376; 0.029425] | [ 0.000076; 0.012560] |

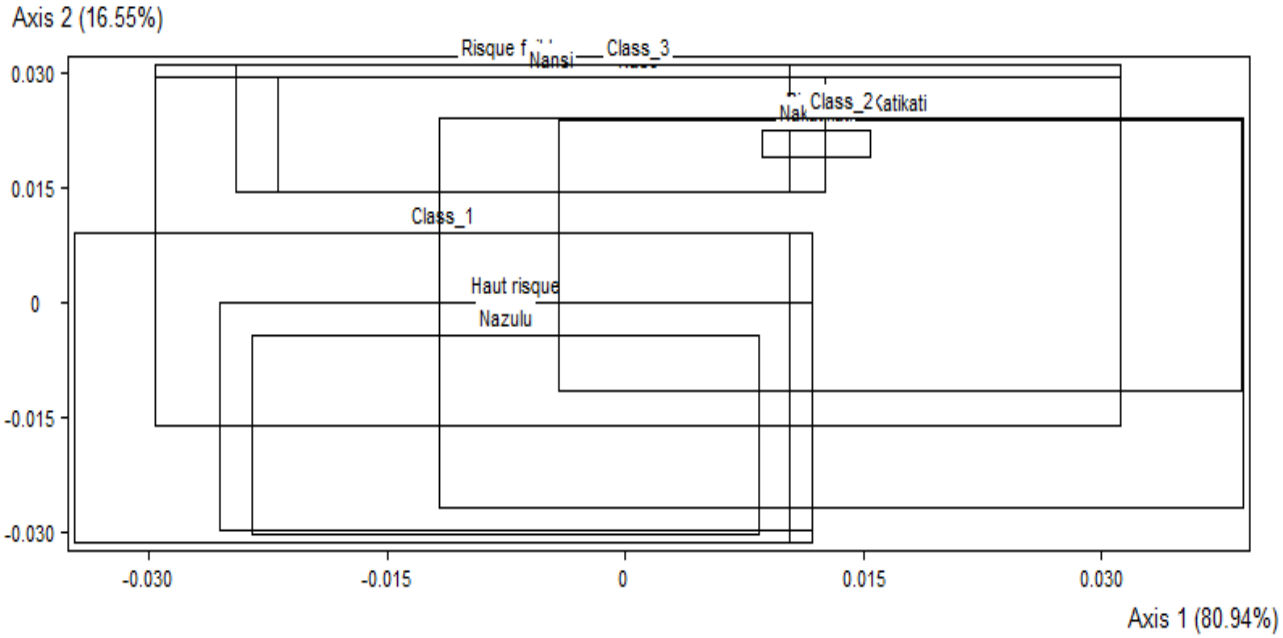



***Fig. 6. Représentation graphique des objets symboliques***

***Tableau 4. Matrice d'Affectation des Objets symboliques***

| | Class 1 | Class 2 | Class 3 |
|---|---|---|---|
| Haut risque | 1.0 | - | - |
| Risque moyen | - | 1.0 | - |
| Risque faible | - | - | 1.0 |
| Likolo | 1.0 | - | - |
| Katikati | - | 1.0 | - |
| Nase | - | 1.0(3) | - |
| Nazulu | 1.0 | - | - |
| Nakatikati | - | - | 1.0(2) |
| Nansi | - | - | 1.0 |

## 6 INTERPRÉTATION ET EXTRACTION DE CONNAISSANCES

En observant de près les résultats de l'analyse de notre Base de données (fichier RCVPatients.sds), nous constatons ce qui suit:

- Dans le Tableau 2, le logiciel a déjà sélectionné les trois premières valeurs propres qui expliquent ensemble le 100% de l'inertie totale; nous résumons donc les données par les trois premières composantes principales.
- Le logiciel a donné également les coordonnées d'intervalle des éléments d'OS (voir Tableau 3); on peut voir tout cela à travers la Figure 6.
- Le rendement graphique de l'AFD-OS est montré sur la Figure 6.
- Les résultats de la classification sont récapitulés dans la table d'assignation (voir Tableau 4) où les lignes sont les objets symboliques et les colonnes les classes; les parenthèses donnent les classes a priori des objets mal classés. Ici, les deux objets symboliques Nase et Nakatikati sont mal classés de sorte que le taux correct de classification est de 77.8 %.

## 7 CONCLUSION

Les résultats de l'AFD-OS montre à suffisance que la statistique des individus n'est pas la statistique des concepts: il y a en fait complémentarité des approches classiques et symboliques. Nous pouvons voir cela dans les faits suivants:

En codage classique, chaque concept (unité statistique) est représenté par juste un point du plan discriminant d'une AFD classique et les classes sont typiquement identifiées par leurs centroïdes correspondants, comme indiqué dans la figure ci-dessous tirée de l'article susmentionné à la section 3.1. [8]

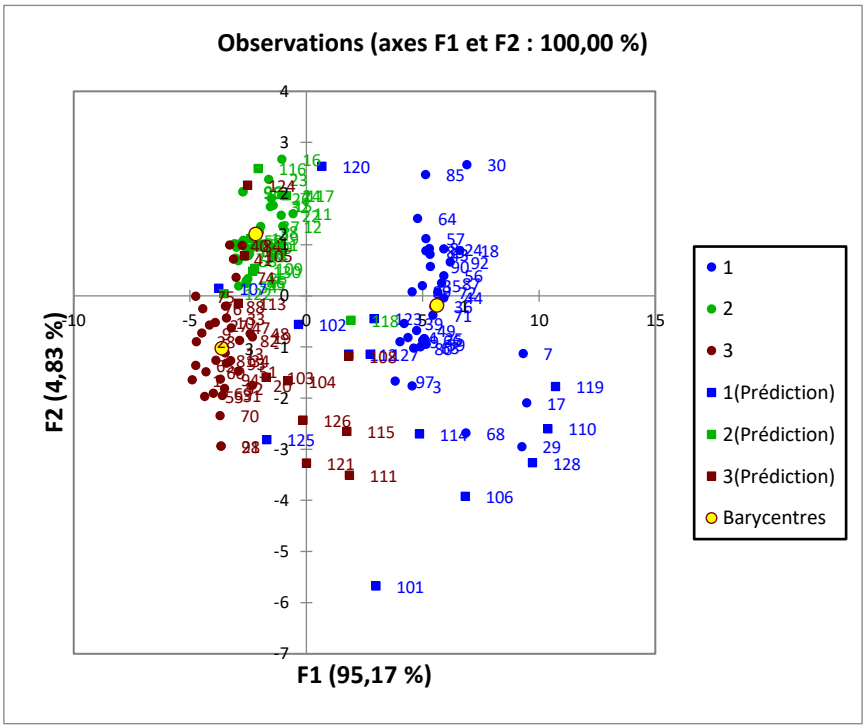



***Fig. 7. Plan discriminant 1x2 (us actives et us tests)***

Ainsi, dans cette figure, on voit apparaître des observations (us actives et us tests) et les barycentres; les 3 groupes à risque (1, 2 et 3) sont bien distingués et représentent respectivement: le haut risque, risque moyen et risque faible.

En codage symbolique d'une AFD symbolique, chaque concept est représenté par une surface, ici (voir la figure 6) un rectangle exprimant la variation du concept (de la valeur minimale à la valeur maximale prises par les individus inclus dans le concept). Les classes sont visualisées comme des rectangles comprenant les objets symboliques appartenant à la classe.

Aussi, chaque concept peut être encore décrit par une conjonction de propriétés réduite aux axes factoriels retenus; ici: Age, AF/MS, Glucose, etc.

**REFERENCES**


[1] Bock, H. H. & Diday, E. (eds). Analysis of Symbolic Data. Exploratory Methods for Extracting Statistical Information from Complex Data. Springer, Berlin, vol. 15, 2000.

[2] DEVROYE L., Non-parametric density estimation: the Ll view. Wiley, New York, 1985.

[3] DIDAY E., Introduction à l'Analyse des Données Symboliques. Rapport de Recherche INRIA N° 1074, INRIA Rocquencourt 78150, France, 1989.

[4] DIDAY E. and Kodratoff Y., Des objets de l'analyse des connaissances, Induction symbolique et numérique à partir des données, Edition Cépaduès, 1991.

[5] DIDAY E. et NOIRHOMME M., Symbolic Data Analysis and the SODAS software, Wiley, 2007.

[6] HAND D. J., Kernel Discriminant Analysis. Wiley, Chichester, 1982.

[7] KANNEL WB, DAWBER TR, KAGAN A, REVOTSKIE N, STOKES J III. Factors of risk in the development of coronary heart disease - six-year follow-up experience: the Framingham Study, Ann Intern Med, 1961; 55: 33-50.

[8] KASIAMA J. et MABELA R., Extraction de connaissances dans les données médicales par l'approche de l'analyse discriminante: Application au risque cardiaque et aux accidents vasculaires cérébraux, In Annales de la Faculté des Sciences, Volume1, Editions Ita'yalaPrinter, Université de Kinshasa, RDC, 2014.

[9] NABI H, KIVIMAKI M, DE VOGLI R, MARMOT MG ET SINGH-MANOUX A, Positive and negative affect and risk of coronary heart disease: Whitehall II prospective cohort study, BMJ, 2008, 337: a 118.

[10] Newman JD, Davidson KW, Shaffer JA et al, Observed hostility and the risk of incident ischemic heart disease: A prospective population study from the 1995 canadian Nova Scotia health Survey, J Am Coll Cardiol, 2011; 58: 1222-1228.

[11] SILVERMAN B.W., Density estimation for statistics and data analysis. Chapman and Hall, London, 1986.

[12] STEPHAN V., Construction d'objets symboliques par synthèse des résultats de requêtes SQL. Thèse de Doctorat, Université Paris IX- Dauphine, Janvier 1998.

[13] SUMMA G., MGS in SODAS. Cahiers du CEREMADE No. 9935, Universite Paris IX-Dauphine, France, 1999.