FACTEURS DE RISQUE ET PREDICTION DU DIABÈTE DE TYPE 2 EN ALGÉRIE : UNE NOUVELLE APPROCHE UTILISANT LE DATA MINING

Nora **LOUNICI MOSBAH**[*]
Khadidja **SADI**[*]

Received: 02/05/2019 / **Accepted**: 24/06/2020 / **Punlished**: 25/11/2020
Corresponding authors: noralounici@yahoo.fr

RÉSUMÉ

Dans cet article, nous utilisons des outils de data mining pour extraire les facteurs de risque du diabète Type 2 (DT2) et prédire la survenue de la maladie par l'élaboration de règles prédictives. L'étude compare l'efficacité de deux modèles d'apprentissage supervisé : *les arbres de décision et le bayésien naïf*. En analyse uni-variée, sept variables étaient pertinentes : le sexe, l'âge, l'IMC, le taux de cholestérol, l'HTA, l'activité physique et les ATCF. L'analyse multivariée a montré que la masse corporelle et l'activité physique sont les principaux facteurs de risque du DT2 et à un degré moindre l'âge et le taux de cholestérol. La classification par les deux modèles a donné une précision de 94. 5% pour les arbres de décision et de 96,47% pour le bayésien. Le plus prédictif des deux modèles étant les arbre de décision, avec une aire sous la courbe ROC de 0,964, un taux d'erreur estimé à 10 ,44% et une capacité à détecter les vrais diabétiques de 90,5%.

MOTS CLÉS

Data Mining, Diabète, facteurs prédictifs, arbres de décision, bayésien naïf.

JEL CLASSIFICATION: I19, C11, C15, C55, D81, D83.

[*] Laboratoire LASAP, Ecole nationale Supérieure de la Statistique et de l'Economie Appliquée (ENSSEA), Algérie. noralounici@yahoo.fr et sadikh00@gmail.com

RISK FACTORS AND PREDICTION OF TYPE 2 DIABETES IN ALGERIA: A NEW APPROACH USING DATA MINING

ABSTRACT

In this article, we use data mining tools to extract risk factors for Type 2 Diabetes (DT2) and predict the occurrence of the disease by developing predictive rules. The study compares the effectiveness of two supervised learning models: decision trees and Naïf Bayesian. In univariate analysis, seven variables were pertinent: gender, age, BMI, cholesterol, hypertension, physical activity and ATCF. The multivariate analysis showed that body mass and physical activity, are the main risk factors for DT2 and to a lesser degree age and cholesterol rate. Classification by both models yielded an accuracy of 94.5% for the decision trees and 96.47% for the Bayesian. The most predictive of the two models was the decision tree, with an area under the ROC curve of 0.964, an estimated error rate of 10.44% and an ability to detect true diabetics of 90.5%

KEY WORDS :

Data Mining, Diabetes, factors Predictive, Decision Trees, Naive Bayesian.

JEL CLASSIFICATION: I19, C11, C15, C55, D81, D83.

عوامل الخطر والتنبؤ بمرض السكري من النوع 2 في الجزائر: نهج جديد باستخدام تقنية DATA MINING

ملخص

في هذه لمقالة، تم استخدام أدوات استخراج البيانات (Data Mining) لاكتشاف عوامل ظهور مرض السكري من النوع 2 (DT2) و التنبؤ بظهور المرض من خلال وضع قواعد تنبؤي . تقارن الدراسة فعالية نموذجي *Decision tree* و *Naïve Bayes* . *classifier* في تحليل متغير واحد ، تم التوصل إلى سبعة متغيرات ذات صلة: الجنس، العمر ، مؤشر كتلة الجسم ، مستوى الكوليسترول ، ارتفاع ضغط الدم ، النشاط البدني وتاريخ العائلة (ATCF) .أظهر تحليل متعدد المتغيرات أن كتلة الجسم والنشاط البدني هما عاملان رئيسيين من عوامل الخطر لـ DT2 ، يتبعان بمستويات العمر والكوليسترول .أعطى تصنيف النموذجين دقة 94.5٪ بتقنية *Decision tree* و 96,47% بتقنية *Naïve Bayes* . *classifier* و قد تبين أن تقنية *Decision tree* أكثر تنبؤاً ، بمساحة تحت منحنى (0.964) ROC ، معدل الخطأ يقدر بـ 10.44٪ والقدرة على اكتشاف مرضى السكري بنسبة 90.5٪.

كلمات مفتاحية :

استخراج البيانات، مرض السكري، العوامل التنبؤية، *Decision tree* ، *Naïve Bayes classifier*

تصنيف جال: I19, C11, C15, C55, D81, D83.

INTRODUCTION

Le diabète très répandu de nos jours est une maladie caractérisée par une hyperglycémie chronique qui survient lorsque le pancréas ne produit pas assez d'insuline ou lorsque l'organisme n'utilise pas correctement l'insuline qu'il produit. Cette étude concerne le DT2 qui se manifeste en général après 40 ans. La principale raison en est notre mode de vie qui associe mauvaise hygiène alimentaire et manque d'activité physique. Plusieurs facteurs ont été évoqués, ils s'expliquent par les variations de l'espérance de vie, l'environnement, le mode de vie, les prédispositions génétiques et l'augmentation de l'espérance de vie[1] (A. Guerin -Dubourg 2004).

Cette pathologie est considérée comme un sérieux problème de santé publique, en raison de la hausse rapide du nombre de patients (Santos F. 2015). La Fédération Internationale du Diabète (FID)[2] et l'Organisation Mondiale de la Santé (OMS) soulignent la prévalence astronomique du diabète qui touche désormais plus de 425 millions de personnes, dont un tiers âgées de plus de 65 ans. Selon les statistiques de la Fédération algérienne des associations de diabétiques[3], les campagnes de sensibilisation et de dépistage précoce destinées aux personnes atteintes de diabète ont permis de recenser un nombre estimé en Algérie à 5 millions, soit 14 % de la population, avec un taux d'incidence alarmant de 20.000 cas enregistrés par an. Mais ces statistiques restent incertaines, car il y a beaucoup de personnes atteintes et non diagnostiquées.

Ainsi, la statistique, comme l'analyse de données, joue un rôle important dans les domaines médicaux, en particulier en épidémiologie et en médecine. La littérature ayant trait à ce sujet est abondante.

Cependant la plupart des méthodes sont empiriques et classiques. L'analyse de données doit nécessairement étendre les outils de l'Analyse classique à ceux du Data Mining (Saporta G. 2006).

[1] https://www.inserm.fr/information-en-sante/dossiers-information/diabete-type-2

[2] file:///C:/Users/user/Downloads/French-6th%20(1).pdf

[3] http://www.elmoudjahid.com/fr/actualites/130734, publié le : 19-11-2018

La fouille de données (data mining) consiste en l'utilisation des techniques et algorithmes afin d'extraire à partir d'une masse de données, des connaissances pouvant servir de support au processus de décision (Piatesky Shapiro G. & al. 1996) et (Tufféry S. 2012). Les techniques de data mining ont été appliquées dans les soins de santé pour acquérir une compréhension approfondie des données médicales, pour sélectionner des stratégies thérapeutiques après un diagnostic précis et pour choisir des soins médicaux appropriés en fonction du pronostic de la maladie (Fontbonne. A 2010).

Des études ont déjà été réalisées avec différentes méthodes d'analyse et d'extraction de données dans le cadre d'essais cliniques sur le diabète, on peut mentionner (Nourizadeh A. & al 2013) et (Sankaranarayanan S. 2014). Des approches déjà adoptées dans ce cadre se sont tournées vers des techniques telles que les K-plus proches voisins (Hung-Chun Lin & al 2011), un système d'aide à la décision clinique basé sur l'OLAP pour le diagnostic de patients proposé par (Bagdi R. & al 2012) a permis aux auteurs de catégoriser les maladies avec une probabilité élevée, faible ou moyenne ou encore les réseaux de neurones, utilisés pour analyser des échantillons de sang et d'urine, suivre les niveaux de glucose chez les diabétiques et détecter les conditions pathologiques (Stanfford G. C. & al 1984).

En 2017, à Kénitra au Maroc, (Zeghari L. & al 2017) ont mené une étude sur 2227 diabétiques. L'objectif était l'étude de l'obésité et le contrôle glycémique sur des diabétiques de différent type (type 1, 2 et gestationnel). Les résultats qui en découlent ont montré que l'ensemble des diabétiques présentent des valeurs de l'IMC et du contrôle glycémique, supérieures aux normes.

Une autre étude multiethnique (Gregory L. & al 2008) d'une cohorte impliquant 6814 a montré l'impact négatif de l'excès pondéral et de l'obésité de participants et que l'hypertension et le diabète étaient plus fréquents chez les participants obèses malgré l'utilisation de médicaments antihypertenseurs et/ou antidiabétiques.

Par ailleurs, un des objectifs d'un projet nommé DREAM (Sheen A.J. & al 2003) qui s'est déroulé en France en 2003 était de recueillir des don-

nées épidémiologiques concernant les caractéristiques des patients diabétiques de type 2 suivis en médecine générale. Au total, 163 patients diabétiques ont été inclus : 84 hommes et 79 femmes, 59 ± 10 années d'âge ; 5,2 ± 6,1. L'indice de masse corporelle (IMC) était de 30,7 ± 5,8 kg/m2, confirmant la forte prévalence de l'excès de poids et de l'obésité dans cette population. Dans ce groupe 60 % avouaient une sédentarité marquée, 23 % un tabagisme actif. L'HTA était de 141 ± 16 / 82 ± 10 mm Hg, avec 63 % des patients prenant au moins un médicament antihypertenseur et le taux de cholestérol total était de 229 ± 59 mg/dl.

L'objectif principal à travers cette étude est donc, d'élaborer des modèles prédictifs basés sur un processus de fouille de données assez novateur pour extraire les facteurs d'influence du diabète non insulinodépendant chez les adultes et par la même occasion prévoir la classe d'un nouveau patient selon qu'il soit diabétique ou non

La suite de cet article est organisée de la façon suivante : la **section1** est consacrée à une description statistique de notre base de données, la **section2** explicite l'aspect théorique des algorithmes d'apprentissage. Nos résultats sont ensuite synthétisés en **section 3**. Après une discussion effectuée en **section 4**, nous concluons et indiquons les perspectives de ce travail dans la **section 5.**

1- NOTRE APPROCHE

Notre approche comprend deux étapes. La première porte sur la présentation et l'épuration de la base de données. Dans la seconde étape nous appliquons respectivement la technique des arbres de décision et la méthode de la classification bayésienne naïve.

1.1- Présentation des données

L'enquête a été réalisée en 2016 dans la ville de Relizane (située au nord-ouest de l'Algérie)[4]. Les informations ont été recueillies par le biais d'un questionnaire et collectées lors d'entrevues réalisées en face à face, à partir d'examens physiques et de prélèvements biologiques, disponibles au niveau de l'association des diabétiques et depuis la

[4] L'enquête a été réalisée par une étudiante de dernière année de Master option statistique et économie appliquée

clinique mobile « Changing diabetes », organisée par le ministère de la santé et de la réforme hospitalière en collaboration avec le laboratoire Novo Nordisc[5].

Les candidats intéressés à participer ont été contactés et interrogés sur les facteurs de risque évoqués dans la littérature (données sociodémographiques du patient, antécédents familiaux, existence d'autres pathologies comme l'hypertension, tabac, régime alimentaire, etc.). Au total, un échantillon de 134 individus d'âge ≥ 33 ans a été constitué.

Nous commençons par présenter dans ce qui suit, la méthodologie, le plan de l'enquête et la base de données.

1.2- Echantillonnage

L'échantillon a été tiré selon un sondage aléatoire stratifié de manière à respecter la proportion de diabétiques et de non diabétiques ainsi que la proportion de la population selon l'âge et le sexe. L'enquête concerne l'état de santé des diabétiques avant la détection de la maladie.

1.3- Description des attributs de la base de données

Les données proviennent d'une enquête réalisée auprès de personnes venant consulter. D'un long questionnaire, nous avons extrait 16 variables explicatives et une variable cible : **diagnostic** (diabétique, non diabétique) (Voir **Tableau 1** en Annexe).

• *Préparation et nettoyage des données*

Afin de sélectionner les variables les plus significatives pour les modèles, nous commençons par épurer les données : élimination de certaines valeurs, traitement des valeurs manquantes, détection des valeurs aberrantes et bien d'autres types d'incohérences qui peuvent gêner l'analyse.

Ensuite, nous nous sommes intéressés aux variables. Nous avons commencé par éliminer les variables poids et taille qui sont corrélées à

[5] L'objectif principal de cette dernière est de faire un dépistage précoce du diabète en Algérie

IMC, ainsi que les variables cystite, mycoses qui ne présentent aucun intérêt pour cette analyse (les patients n'en sont pas atteints). Les variables hyperglycémie pendant la grossesse, avoir un bébé plus de 4 kg, atteinte rénale et atteinte cardiaque, ont également été supprimées (la modalité positive concerne peu d'individus). Les variables HTA SBP et HTA DBP ont été regroupés en HTA (oui/non). Notre base de données est désormais constituée de 134 instances et 8 variables. Pour améliorer les performances de classification, nous avons discrétisé[6] la variable âge en 5 classes d'amplitudes égales. La variable continue taux de cholestérol a aussi été discrétisée en deux intervalles.

1.4- Etude exploratoire

Les résultats des attributs quantitatifs sont exprimés en moyennes ± écart-type. La saisie des données a été effectuée sous Excel et l'analyse exploratoire sous xlstat. De l'ensemble des 134 patients questionnés, près des trois quarts ont plus de 50 ans (72%) et seulement 6% ont 70 ans et plus, dont 65% sont des femmes. Le tableau ci-dessous indique la répartition par sexe selon l'âge.

Tableau n°1: structure de l'échantillon par âge et par sexe.

sexe Age	Hommes		Femmes		Total
	Effectif	%	Effectif	%	
[30,40[	0	0	5	5.75	5
[40,50[	12	25.53	21	24.13	33
[50,60[	15	31.92	32	36.78	47
[60,70[	16	34.04	27	31.03	43
[70,80[	4	8.51	2	2.29	6
Total	47	100	87	100	134

Source : Elaboré par nous même à l'aide du logiciel XLSTAT

Sur les 134 cas, on compte 87 femmes et 47 hommes d'âge moyen 55.40 ± 9.18 (min = 33; max = 79) et dont 3/4 des patients avaient plus de 50 ans. Leur niveau d'étude ne dépasse pas l'enseignement secondaire. L'IMC moyen vaut 32.21 ± 3.96 (Tableau 2). Dans notre échantillon, près de 96 % de personnes interrogées sont en surpoids dont 73% sont obèses

[6] Réaliser un découpage en classes

avec un indice de corpulence dépassant 30 kg/m2. D'après les différentes études, le critère international d'embonpoint est un IMC de 25. Quand l'IMC dépasse 30 il y a obésité et donc risque accru de diabète.

On note cependant la présence de l'antécédent familial du diabète et de l'hypertension artérielle (HTA) chez respectivement 43 (32%) et 46 (34%) des sujets enquêtés. Seuls 5 personnes interrogées ont déclaré avoir une activité physique régulière.

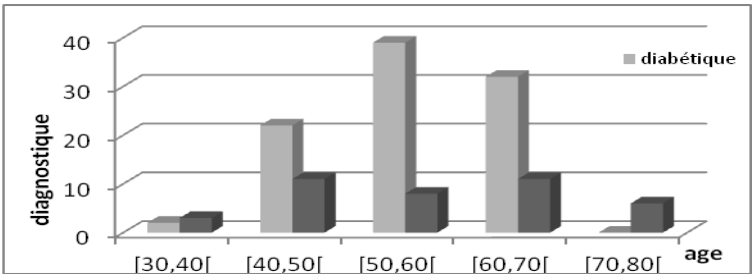
Tableau 2 : Caractéristiques des attributs continues

Attributs	Moyenne	Mode	Médiane	Min	Max
IMC en kg/m^2	32.21 ± 3.96	31.64	32.43	22.86	43
Age (ans)	55.40 ± 9.18	62	55	33	79
τ cholestérol(g/l)	1.98 ± 0.43	1.82	1.95	1.10	3.25
HTA-SBP(mmHg)	131.92 ± 19.95	130	130	80	180
HTA-DBP (mmHg)	78.88 ± 10.13	80	80	50	103
Poids(kg)	83.70 ± 11.91	81	84	56	109
Taille(m)	1.61 ± 0.08	1.60	1.60	1.46	1.80

Source : Elaboré par nous même à l'aide du logiciel XLSTAT

Sur la **Figure 1**, on observe que la tranche d'âge la plus touchée par la maladie, les deux sexes confondus, concerne les [50 - 70]ans.

Figure 1 : Répartition des diabétiques/non diabétique selon l'âge



Source : Elaboré par les auteurs à l'aide du logiciel XLSTAT

D'après les tableaux ci-dessous, nous observons que les personnes ayant un indice de masse corporelle (IMC) supérieure à 30, concerne tout particulièrement les diabétiques. En revanche, les principaux facteurs de risque pour que des complications affectent le cœur du diabétique sont, entre autres, l'hypertension et un taux élevé de cholestérol.

Dans notre jeu de données, le pourcentage des diabétiques ayant un taux de mauvais cholestérol (47%) ne varie pas de façon significative selon les groupes d'âge les plus touchés par la maladie.

Tableau 3 : Répartition des âges selon l'IMC et le τ de cholestérol

Diagnostic / IMC	Diabétique	Non Diabétique	Σ
[18.5-24.9]	///	5	5
[25-29.9]	///	31	31
>30	95	3	98
Σ	95	39	134

Diabétiques			
	Cholestérol		Σ
Age	BON	MAUVAIS	
[30-40]	2		2
[41-50]	14	14	28
[51-60]	22	13	35
[61-70]	12	18	30
[71-80]			
Σ	50	45	95

Source : Elaboré par les auteurs à l'aide du logiciel XLSTAT

On parle d'hypertension artérielle lorsque les chiffres tensionnels sont supérieurs à 140/90 mmHg[7]. S'agissant des sujets diabétiques de cette étude, 35 sur les 95 détectés se sont avérés hypertendus (37%) soit environ un tiers, une proportion plus élevée que chez les non-diabétiques (28%). Parmi les 95 patients déclarés diabétiques, 22 (23%) ont au moins un des parents diabétique et 32 (33,7%) de sœur ou frère diabétique. D'après la littérature, s'agissant des ATCD familiaux, si l'un des deux parents est diabétique de type 2, le risque de transmission à la descendance est de l'ordre de 30 %[8]. Ce qui est confirmé par nos résultats.

2- ASPECT THEORIQUE DES ALGORITHMES D'APPRENTISSAGE

2.1- Les arbres de décision

Les arbres de décision **(**Lounis H. 2006) sont des méthodes d'apprentissage supervisé qui offrent de nombreux avantages (facilement lisibles, interprétables en un ensemble de règles simples et présentant en général de bonnes performance, etc.). Elles sont couramment

[7] Diabetes, hypertension, and cardiovascular disease: an update. Sowers JR, Epstein M, Frohlich ED Hypertension. 2001 Apr; 37(4):1053-9.

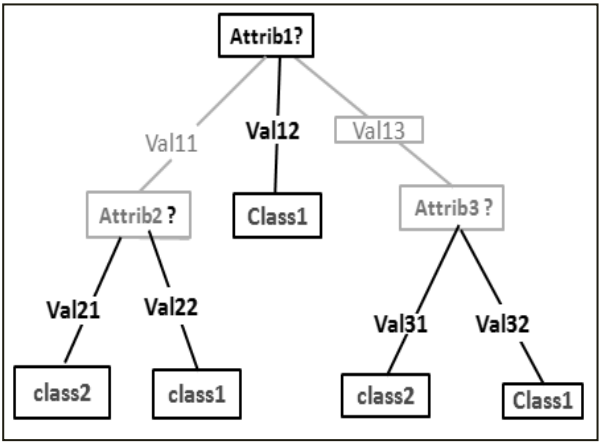
[8] Pr. A. Grimaldi; http://www.chups.jussieu.fr/polys/diabeto/diabeto.pdf

utilisées dans de nombreux domaines notamment en biologie (Geurts P. & al 2009) et comptent de nombreuses approches : ID3 et C4.5 (Quinlan J. R 1993), CART introduite par Breiman (Breiman L. & al 1984), QUEST (Shih. WY Loh & YS Shih 1997), etc.[9] La méthode C4.5 est la plus connue. C'est celle que nous présentons dans ce qui suit.

C4.5 est une méthode de discrimination qui a pour but de construire, à partir d'un échantillon, des groupes qui soient les plus homogènes possible au sens de la variable à expliquer à partir des variables prédictives (explicatives).

Les nœuds de l'arbre correspondent aux variables testées et les branches aux différentes modalités de la variable retenue, quant aux extrémités de l'arbre (les feuilles), elles indiquent les différents groupes engendrés (Figure 2). Une nouvelle observation est classifiée en parcourant l'arbre et en empruntant le chemin correspondant aux valeurs des attributs de cette observation.

Figure 2 : Schéma d'un arbre de décision n-aire



Concrètement l'algorithme examine successivement toutes les variables explicatives, en considérant les différentes divisions possibles,

[9] Hamilton, Howard J. (2006). Decision Trees. Lecture notes (Knowledge Discovery in Databases), University of Regina.

pour en sélectionner la meilleure. La sélection se fait à l'aide d'un critère de type « variance résiduelle ». Parmi toutes les possibilités de segmentation, l'algorithme conserve la variable qui produit les nœuds-fils les plus homogènes possibles pour la variable prédictive. Cette procédure est réitérée sur les nœuds-fils ainsi générés.

- *Notations formelles*

Soit E l'échantillon concerné par le problème d'apprentissage décrit sous forme d'un tableau individus × variables. Chaque individu i est caractérisé par p variables prédictives : x_1, ..., x_p et une variable à prédire Ci (la classe).

Les algorithmes inductifs d'arbres de décision génèrent un modèle représenté sous la forme d'un arbre de décision, pour prédire la classe d'un nouvel individu. Le chemin qui mène de la racine de l'arbre à une feuille représente une règle de décision. L'ensemble de ces règles forme le modèle de prédiction qui permet de prédire la valeur de la variable à expliquer pour un nouvel individu dont on ne connaît que les variables explicatives. Une première difficulté concerne le choix des variables pertinentes. Ce choix est fait sur la base d'un critère de séparation. Parmi les critères les plus fréquemment utilisés figurent : L'entropie de Shannon pour C4.5, GINI pour CART (Santos F. 2015).

Pour un nœud donné N_d et une variable à expliquer discrète à k modalités ($m_1, m_2, ..., m_i, ..., m_k$), la fonction entropie qui est en réalité une entropie moyenne est définie par :

$H(N_d) = H(m_1, m_2, ..., m_k) = -\sum_{mi=1}^{k} p_{mi,N_d} \times log_2(p_{mi,N_d})$

Pratiquement, p_{mi,N_d} est estimé par la proportion $\frac{n_{m_{i,N_d}}}{n_{N_d}}$; où n_{mi,N_d} est l'effectif dans le nœud N_d de la modalité m_i et n_{N_d} l'effectif total du nœud N_d.

➔Entropie de Shannon du nœud N_d: $\boxed{H(N_d)=-\sum_{mi=1}^{k} \frac{n_{m_{i,N_d}}}{n_{N_d}} \times log_2\left(\frac{n_{mi,N_d}}{n_{N_d}}\right)}$

La stratégie de recherche de l'algorithme C4.5 consiste pour chaque variable potentiellement candidate, à effectuer un partitionnement des

observations. A la suite de quoi, l'indicateur de qualité $H(N_d)$ est calculé et la variable qui sera retenue est celle qui optimise ce critère et donc discrimine au mieux la population d'apprentissage vis-à-vis des classes de la variable cible. Un sous arbre est alors construit pour chaque sous-population non encore discriminée.

Les performances des arbres de décision reposent essentiellement sur leur taille. On cherche à construire l'arbre le plus parcimonieux possible. Déterminer un tel arbre revient à minimiser autant que possible l'erreur de classification, ce qui revient à contrôler sa taille et faire les meilleures prédictions sur de nouvelles données.

En effet, construire un arbre trop spécialisé est souvent sujet à des problèmes de sur-apprentissage et donc à des feuilles contenant très peu d'individus et très dépendantes des données d'apprentissage. Il existe deux stratégies pour contrôler la taille de l'arbre. On arrête la procédure de segmentation lorsqu'il n'y a plus de division admissible des nœuds, c'est-à-dire lorsque l'effectif dans le segment terminal est inférieur à un chiffre initialement fixé (généralement fixé à 5 sujets), ou que la valeur Y est la même pour tous les individus (c'est-à-dire lorsque la variance est nulle).

Il est intéressant de noter que les attributs apparaissant dans l'arbre sont les attributs pertinents pour le problème considéré.

2.2- Le bayésien naïf

Ce modèle est largement utilisé pour les problèmes de classification et a donné de très bons résultats dans différents domaines d'application **(**Rish I. 2005) et **(**Domingos P. & Pazzani M. 1997). Le bayésien naïf est une méthode d'apprentissage supervisée probabiliste basée sur l'application du théorème de Bayes et s'appuie sur une hypothèse dite « naïve ». On le qualifie de naïf car il présuppose une hypothèse très forte : les variables explicatives sont deux à deux indépendantes[10] conditionnellement à la variable à prédire. Cette hypothèse introduit certes

[10] Pour pouvoir appliquer cette méthode, on présuppose que les valeurs d'attributs sont indépendantes, sachant une classe.

un biais dans les prédictions, mais fournit des résultats étrangement bons pour des problèmes de classification (Domingos P. & al 1996).

Le bayésien est un modèle linéaire **(**Parent E. & Bernier J. 2007), il présente d'excellentes performances (au vu de sa simplicité) en prédiction, notamment dans le cas où toutes les variables sont discrètes. Il peut être appliqué sur de très grandes bases de données et son efficacité est comparable à celle des autres techniques d'apprentissage comme l'analyse discriminante et la régression logistique même lorsque l'hypothèse d'indépendance est violée **(**Ripley B. D. 1996).

- *Notations formelles*

Formellement, étant donné un jeu de données, où $X = (x_1, \ldots\ldots, x_j)$ l'ensemble des variables prédictives et $Y = (y_1, y_2, \ldots, y_k)$ la variable à prédire comportant k modalités. Le classifieur naïf de Bayes nécessite simplement en entrée l'estimation des probabilités conditionnelles par variable P(xj |Y) et les probabilités à priori P(Y). La probabilité conditionnelle jointe P(X|Y) étant difficilement estimable on utilise la version naïve (hypothèse de l'indépendance des variables).

L'estimation des paramètres pour les modèles bayésien s naïfs repose sur le maximum de vraisemblance. La classe d'un nouvel individu ω est déterminée par la règle de décision qui consiste à classer ω dans la classe Y pour laquelle P(Y = y/X) est maximale. C.-à-d.

$$\hat{y}(\omega) = \arg\max_k P(Y = y_k/X) \quad \Leftrightarrow \quad \hat{y}(\omega) = \arg\max_k \left(\frac{P(Y=y_k) * P(X/Y=y_k)}{P(X=x)}\right)$$

Comme on cherche à extraire le maximum de $\hat{y}(\omega)$ selon y, et que le dénominateur ne dépend pas de la classe y, nous pouvons simplifier la règle d'affectation :

$$\hat{y}(\omega) = \arg\max_k \left(P(Y = y_k) * P(X/Y = y_k)\right)$$

Sous l'hypothèse imposée de l'indépendance des variables nous pouvons écrire la probabilité à posteriori de chaque classe comme suit : $P(X/Y = y_k) = \prod_{j=1}^{J} P(x_j/Y = y_k)$;

$$P(Y = y_k/X) = P(Y = y_k) * \prod_{j=1}^{J} P(x_j/Y = y_k) \quad (1)$$

On cherche à maximiser cette expression :

$$y^*_k = \arg\max_k \left(P(Y = y_k) * \prod_{j=1}^{J} P(x_j / Y = y_k)\right) \quad (2)$$

Généralement à cette étape nous passons par les logarithmes[11] car le produit de plusieurs valeurs inférieures à 1(les probabilités conditionnelles estimées) peut conduire à des dépassements de capacités. La règle de décision de (1) devient alors :

$$y^*_k = \arg\max_k \left(\ln P(Y = y_k) + \sum_{j=1}^{J} \ln P(x_j / Y = y_k)\right)$$

Notre premier objectif est d'extraire les facteurs de risque responsables les plus probables de la maladie. Or cette hypothèse ne répond pas à cette question. Par conséquent, nous dérivons un modèle explicite à partir de bayésien naïf selon l'approche proposée dans **(**Domingos P. & Pazzani M. 1997) qui consiste au préalable à créer un tableau disjonctif complet à partir de notre tableau de données.

Etant donnée une variable prédictive X à J modalités, on peut dériver J indicatrices comme suit :

La formule (1) $\Leftrightarrow$ $\mathbf{d}(y_k, X) = ln\, P(Y = y_k) + \sum_{j=1}^{J} ln\, P(X = x_j / Y = y_k) \times I_j$

$\Leftrightarrow$ $\mathbf{d}(y_k, X) = \mathbf{a_{0,k}} + \sum_{l=1}^{L} \mathbf{a_{l,k}} \times \mathbf{I_l}$

Ainsi, on obtient une combinaison linéaire d'indicatrices exactement comme avec la régression logistique. Il s'agit précisément de la fonction de classement $d(y_k, X)$ pour la modalité k. On doit ensuite généraliser le modèle additif aux J variables explicatives. Le *jème* descripteur Xj prend Lj modalités, nous lui associons donc Lj indicatrices. La fonction de classement qui répond au second objectif, la prédiction de la classe d'un nouvel individu s'écrit :

$$d(y_k, X) = ln P(Y = y_k) + \sum_{j=1}^{J} lnP(x_j = Lj / Y = y_k) + \sum_{j=1}^{J} \sum_{l=1}^{Lj-1} ln \frac{p(x_j = l / Y = y_k)}{p(x_j = Lj / Y = y_k)} \times I_l^{\,j}$$

où $I_l^{\,j}$ est l'indicatrice l de la variable x_j.

[11] http://eric.univ-lyon2.fr/~ricco/cours/slides/naive_bayes_classifier.pdf

3- EXPERIMENTATIONS

Nous poursuivons deux objectifs dans cette expérimentation. D'une part, nous souhaitons extraire les facteurs de risque tout en vérifiant si la qualité de la prédiction de l'arbre construit est comparable à celui du classifieur bayésien naïf.

D'autre part, nous appliquons les règles de classification établies lors de la phase d'apprentissage pour en déduire la classe, a priori inconnue, d'un nouveau patient. Cette prédiction est réalisée pour chaque algorithme. Afin de mettre les deux algorithmes sur un pied d'égalité, tous les attributs continus ont été discrétisés. Le taux d'erreur est estimé par la procédure de validation croisée.

3.1- Facteurs de risque et prédiction du diabète par les arbres de décision

L'apprentissage par arbres de décision est un thème de recherche très apprécié en data mining ces dernières années **(**Piatesky Shapiro G. & al. 1996), essentiellement pour leur représentation graphique et leur intelligibilité. A tout arbre de décision, on associe tout naturellement un ensemble de règles relatives aux différents chemins qui façonnent l'arbre. Pour exécuter l'algorithme, nous appliquons la règle de l'apprentissage supervisé par validation croisée (cross validation).

La validation croisée consiste à répéter l'estimation de l'erreur sur plusieurs échantillons de validation pour en calculer une moyenne **(**Ripley B. D. 1996). Pour ce faire, on subdivise les données en K blocs (généralement 10) et on répète K fois le processus en considérant l'ensemble des K-1 blocs comme échantillon d'apprentissage et le test sur le K-ième bloc restant. Ceci permet de diminuer la variance du taux de prédiction estimé. L'arbre, nous l'avons construit avec l'algorithme J48 (équivalent de C4.5. sous WEKA[12]).

Nous avons divisé aléatoirement notre ensemble d'apprentissage en 10 portions distinctes, soit approximativement 13 instances pour chaque portion. L'erreur obtenue par validation croisée en 10 blocs est

[12] Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann et Ian H Witten (2009). « The WEKA data mining software: an update ». In : ACM SIGKDD explorations newsletter 11.1, p. 10–18 (cf. p. 66, 100, 168)

de 18%. Nous jugeons cette erreur un peu élevée. Etant donnée la taille réduite de notre échantillon, nous décidons de diminuer le nombre de portions à 5.

D'après le Tableau 4, on obtient une estimation de l'erreur égale à 10,44%, soit près de 89,56% de bien classées, ce qui est acceptable. Concernant les autres indicateurs de qualité, un tableau comparatif sera présenté dans la partie discussion dans le but de confronter les résultats des deux modèles.

Tableau 4 : Résultats d'évaluation (arbre de décision)

J48 unpruned tree

Number of Leaves : 9

Size of the tree : 14

=== Summary ===

Correctly Classified Instances	120	89.5522 %
Incorrectly Classified Instances	14	10.4478 %
Total Number of Instances	134	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	PRC Area	Class
	0,905	0,128	0,945	0,905	0,925	0,964	0,981	diabetique
	0,872	0,095	0,791	0,872	0,829	0,964	0,894	Non_diabetique
Weighted Avg.	0,896	0,118	0,900	0,896	0,897	0,964	0,956	

=== Confusion Matrix ===

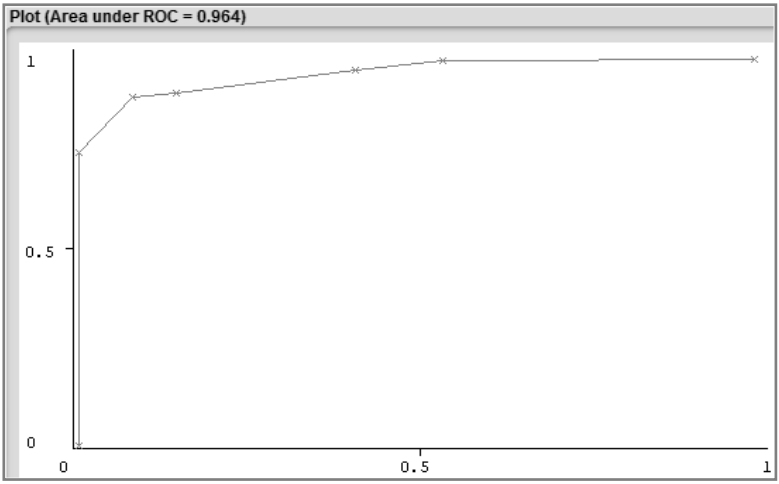
a	b	<-- classified as
86	9	a = diabetique
5	34	b = Non_diabetique

Source : Elaboré par les auteurs à l'aide du logiciel WEKA

Nous nous intéressons à présent à la courbe ROC pour évaluer la qualité du classifieur obtenu (pourcentage d'instances non rejetées bien classées en fonction du pourcentage des instances rejetées).

La courbe ROC est un estimateur de l'efficacité globale du modèle à bien discriminer les groupes. Le modèle est meilleur lorsque son AUC[13] est proche de 1.

Figure 3: Courbe ROC pour les arbres de décision



Source : Elaboré par les auteurs à l'aide du logiciel WEKA

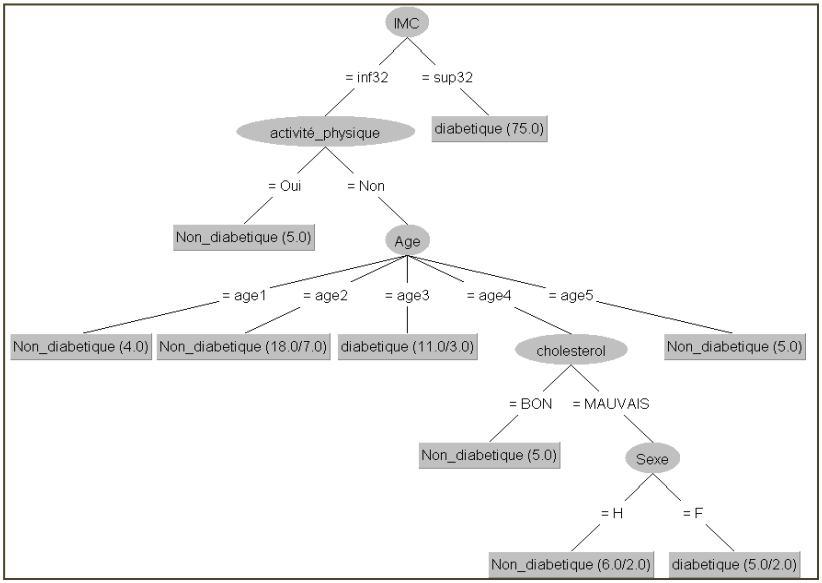
La Figure 3 nous donne la courbe ROC du modèle (courbe en trait plein) et révèle que l'aire de cette courbe vaut : AUC = 0.964, ce qui dénote d'une très bonne robustesse du modèle.

La Figure 4 montre que la variable qui sépare le mieux les instances de chaque classe est IMC. Parmi les 134 patients, 75 dont l'IMC est supérieur à 32 sont diabétiques. La seconde bifurcation indique que les personnes qui pratiquent une activité physique encourent moins de risque de contracter le diabète.

En poursuivant sur les branches de droite, les patients âgés de 50 à 60 ans et femmes âgées de 61 à 70 ans et souffrant de mauvais cholestérol ont un risque plus élevé d'avoir cette maladie.

[13] AUC : correspond à l'aire sous la courbe ROC

Figure 4 : Modèle d'arbre de décision



Source : Elaboré par les auteurs à l'aide du logiciel WEKA

Le modèle d'arbre de décision J48 (Figure 4) a finalement retenu 5 variables considérées dans l'ordre comme les plus pertinentes :

- L'IMC
- L'activité physique
- L'âge du patient
- Le taux de cholestérol
- Le sexe.

Enfin cet arbre de décision peut être décrit par la règle :

si *le patient ne pratique aucune activité physique, son IMC > 32, son âge compris entre 50 et 70 ans et si en plus c'est une femme souffrant de mauvais cholestérol* **alors** *il est à haut risque de développer un diabète.*

3.2- Extraction des facteurs de risque par le classifieur bayésien naïf

A présent, nous appliquons la technique bayésien naïf de classification. L'idée générale est de calculer la probabilité de chaque

classe C_k, sachant que les attributs (x1, ..., xn) d'un individu donné ont été observés. Pour calculer ces probabilités, nous avons repris les variables explicatives retenues dans le modèle des arbres de décision. Les modalités de la variable cible **Class** sont (diabétique/ non diabétique). Les traitements ont été réalisés avec le logiciel Tanagra[14].

Les différentes robabilités conditionnelles sont engendrées par les tableaux croisés de chaque variable explicative avec la variable Class. Cependant, comme nous l'avons déjà signalé, ces tableaux ne nous sont d'aucune utilité dans ce cadre-là. Pour déterminer les facteurs de risque potentiels du DT2, nous dérivons du bayésien naïf un modèle explicite. Plus clairement, on s'intéresse aux fonctions de classement. Les résultats sont donnés dans le tableau 5.

Tableau 5 : Fonctions de classement (bayésien Naïf) « Tanagra »

Supervised Learning 1 (Naive bayes)		
Results		
	Classification functions	
Descriptors	**Diabétique**	**Non-diabétique**
Sexe = H	-0,662376	-0,446287
age = [51-60]	0,216223	-0,405465
age = [61-70]	0,066691	0,080043
age = [71-80]	-3,36296	-0,693147
age = [30-40]	-2,268684	-0,875469
IMC = > 32	1,286211	-3,688879
Cholestérol = MAUAIS	-0,103184	-0,146603
Activité physique à Non	4,564348	1,763589
ATCD familiaux = Nom	0,803495	0,550046
Constant	-9,923307	-6,592311

Source : Elaboré par les auteurs à l'aide du logiciel Tanagra

Nous obtenons deux régressions linéaires selon les classes (diabétique, non diabétique) avec lesquelles nous pouvons effectuer des prévisions :

$$\mathbf{d(diabétique, X) = -9.923 + 0{,}216 \times Age\ = [51-60] + 0{,}0667 \times Age = [61\text{-}70] + \cdots + 4{,}564 \times (Activité\ physique = Non) + 0{,}803 \times (ATCD\_familiaux = Non)}$$

[14] R. Rakotomalala , TANAGRA, (2005) : Une Plate-Forme d'Expérimentation pour la Fouille de Données", Revue MODULAD, 32, 70-85

d(Non_diabétique, X) = -6.592 - 0,405 × Age $= [51 - 60] + 0,0800 \times$ Age = [61-70] $+ \cdots +$ 1,763 × (Activité physique = Non) + 0,550 × (ATCD_familiaux = Non)

Les X étant des indicatrices (0/1). Pour l'individu à classer nous prenons comme valeurs les indicatrices suivantes (1 ; 1 ; 1… 0; 0; 0).

d(diabétique, X) = $-9.923 + 0,216 \times 1 + 0,0667 \times 1 - 3,36 \times 1 - 2,268 \times 1 + 1,28 \times 1 + 0$ = -13,988
d(Non_diabétique, X)= $-6.592 - 0,405 \times 1 + 0,0800 \times 1$ - 0,693 × $1 - 0,875 \times 1 - 3,689 \times 1 + 0$ = -12,174

Comme la variable cible est binaire, nous calculons une fonction score d(X) unique comme pour la régression logistique. Elle est définie par : Score = d(diabétique, X) - d(Non_diabétique, X)

d(X) = d(diabétique, X) - d(Non_diabétique, X) = -13,988 +12,174 = -1,81

La règle d'affectation serait alors :

si d(X(ω)) > 0 $\hat{y}$ (ω) = + sinon $\hat{y}$ (ω) = - ⇒ X est non diabétique.

Afin d'extraire les variables les plus intéressantes, nous appliquons la méthode CFS[15] proposée dans Tanagra. Cette méthode est basée sur le calcul des corrélations de chaque variable explicative avec la variable cible et les corrélations croisées entre les variables sélectionnées, qui sont ensuite classées par ordre décroissant.

La méthode CFS FORWARD[16] repose sur le calcul d'une mesure globalede « mérite » d'un sous-ensemble de p variables qui tient compte à la fois de leur pertinence et de leur redondance. Le *merit* recherche les prédicteurs les plus en relation avec la cible et les moins liés entre eux. Cette méthode s'appuie sur l'algorithme pas à pas «forward». A chaque étape, l'algorithme choisit la variable qui maximise le merit. Une variable est jugée intéressante si sa liaison avec la

[15] 8 M. Hall, S. Lloyd, « Feature subset selection: a correlation based filter approach », in 1997 Int. Conf. On Neural Information Processing and Intelligent Information Systems, pp/ 855-858, Springer, 1997.

[16] Il s'agit de la méthode STEPDISC (Stepwise Discriminant Analysis) adaptée aux variables qualitatives. Elle repose le critère du LAMBDA de WILKS.

cible Y surpasse sa liaison moyenne avec les prédicteurs déjà sélectionnés. L'expression du merit est donnée par :

$$\text{merit} = \frac{p \times \overline{s_{Y,X}}}{\sqrt{p + p \times (p-1) \times \overline{s_{X,X}}}}$$

Où $\overline{s_{Y,X}}$ est la moyenne des corrélations entre les variables prédictives et la variable cible. $\overline{s_{X,X}}$ représente la moyenne des corrélations croisées entre les variables prédictives. Nous exécutons le composant permettant de calculer le merit dans tanagra. Trois (3) variables significatives sur sept(7), au sens de cette statistique sont sélectionnées (**Tableau 6**) :

L'IMC, l'activité physique et l'âge avec respectivement pour merit (0,498 ; 0,125 ; 0,139).

Tableau 6 : Calcul du merit

Selected attribute	MERIT(S)
IMC	0,498238
activité physique	0,125529
Age	0,139126

Par conséquent, les variables les plus significatives admises comme facteurs de risque sont :

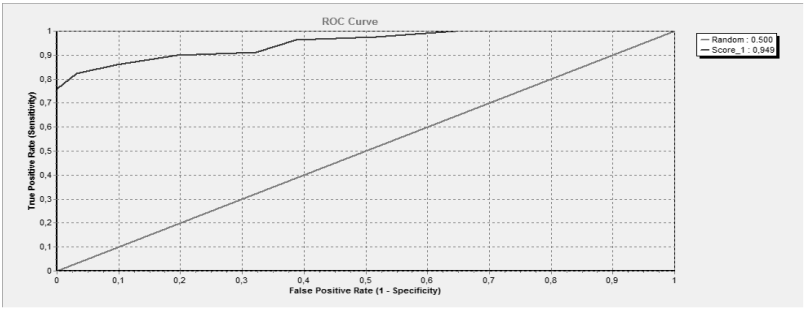
l'IMC; activité physique et l'âge

- **Evaluation du modèle**

Tableau 7 : Performances du bayésien

Error rate			**0,1194**			
Value prediction			**Confusion matrix**			
value	Recall	1-predision		Diabétique	Non-diabétique	Sum
Diabétique	0,8632	0,0353	Diabétique	82	13	95
Non diabétique	0,9231	0,2653	Non diabétique Non diabétique	3	36	39
			Sum	85	49	134

Figure 5 : Courbe ROC pour le bayésien Naïf



Source : les auteurs sous Tanagra

Le Tableau 7 indique une estimation de l'erreur égale à 11,94%, soit près 88% des instances originales classées correctement ce qui est acceptable. L'AUC qui permet de déterminer la qualité globale de la prédiction est proche de 1, elle vaut 0,949.

Le **Tableau 8** présente un récapitulatif des résultats obtenus par les deux modèles. Le taux d'erreur moyen, l'AUC et quelques critères de qualités. Nous remarquons que les résultats découlant des deux méthodes sont assez proches. Les arbres de décision sont toutefois légèrement plus efficaces que le bayésien en termes de performance prédictive.

Les taux d'erreurs obtenus pour les deux méthodologies sont proches, l'arbre de décision détecte mieux les diabétiques avec une sensibilité de 90,5%. A l'inverse le bayésien naïf a tendance à mieux déceler les non diabétiques avec une spécificité égale à 92,3%. Par conséquent, les meilleurs résultats en précision sont obtenus avec le bayésien naïf et en rappel (sensibilité) par les arbres de décisions. Enfin, on peut aussi remarquer que contrairement au bayésien naïf, les arbres de décision font apparaitre en plus les attributs taux de cholestérol et sexe.

Tableau 8 : Performances de prédiction des deux modèles

	Arbre de décision	**bayésien naïf**
Echantillon apprentissage	Cross validation	
Technique	C4.5	bayésien + CS-FORWARD
τ de bon classement	89,55%	88%
τ d'erreur	10,45	11,94%

Précision(diabétiques)	94,5%	**96,47%**
Rappel(diabétiques)	**90,5%**	86,32%
Spécificité(Non diabétiques)	87,2%	92,3%
AUC(diabétique)	96,4%	94,9%
Variables explicatives Retenues	IMC	IMC
	Activité φsique	Activité φsique
	Age	Age
	τ de cholestérol	
	sexe	

Source : Auteurs

4- DISCUSSION

Notre étude présente quelques limites qu'il faut signaler notamment la faible participation de la gente masculine. La ville étant conventionnelle et conservatrice par nature, beaucoup d'hommes ont refusé de répondre au questionnaire.

L'étude a confirmé l'importance de l'IMC comme facteur de risque dans la population des DT2 **(**Procopiou M. 2005)**.** On note, cependant que, le tabagisme est connu pour être un facteur de risque aggravant, malgré cela, les modèles que nous avons appliqués n'ont pas pu mettre en évidence ce paramètre. Contraint par des données collectées auprès de femmes essentiellement non fumeuses, ce facteur de risque a été négligé dans cette étude. Il est par conséquent, nécessaire de conduire d'autres enquêtes pour inclure ce facteur dans le but de mettre en place des politiques de prévention efficaces et pérennes.

Le DT2 a été observé chez 70,9% des personnes qui se sont présentées pour ce faire dépister. Ce travail a montré que cinq (5) variables sont les facteurs déterminants du DT2 pour ce jeu de données (l'IMC, l'activité physique, l'âge, le taux de cholestérol et le sexe). Ces résultats proches de ceux approuvés par (Lounici A. & al 2007), qui par l'application de la régression logistique sur l'étude de 1108 patients dont l'IMC moyen est de 30 ± 5. Le diabète a été dépisté chez 110 patients (9,9 %). Le modèle de régression a fait ressortir les facteurs : âge, TT, DT2 chez l'un des parents, notion d'hyperglycémie transitoire et HTA chez les proches parents et l'obésité abdominale. Des résultats semblables sont également rapportés par (Lounici A. & al 2007) et **(** James R. W. 2002) **.**

(Kabamba AT, & al 2014) a pour sa part mis en évidence unediminution significative du cholestérol-HDL directement liée à la baisse du bon cholestérol avec comme conséquences directes la survenue des complications chez les diabétiques connus de type 2.

Finalement, sur le plan méthodologique, il faut souligner que le choix du modèle pour aider le médecin dans sa prise de décision est primordial. L'intérêt des arbres de décision réside dans son interprétabilité et la lisibilité de ses résultats, ils constituent un outil de régression et de classification incontournable en data mining. De plus, ils permettent d'établir une hiérarchie des variables explicatives par ordre d'importance. Le classifieur bayésien naïf pour sa part, est un outil de classification efficace en pratique pour de nombreux problèmes réels, comme souligné en section 3.2.

CONCLUSION ET PERSPECTIVES

La fouille de données est une nouvelle discipline qui consiste à extraire des connaissances cachées dans les données en utilisant différentes techniques empruntées aux statistiques, à l'informatique et aux mathématiques. Les méthodes de fouille de données ont été utilisées dans de nombreux domaines de la recherche clinique. Bien que les défis de confidentialité représentent les véritables freins à l'utilisation de ces techniques, l'exploration de données reste un outil puissant.

Une approche de data mining est présentée dans cet article, évaluant la capacité de détection facteurs de risque du DT2 au moyen de deux méthodes de data mining, le Arbres de décision et le bayésien Naïf. Le travail présenté avait un double objectif. Le premier est d'identifier les variables ayant le plus d'impact sur les patients diabétiques. Le second objectif est la construction d'un modèle de prédiction. L'ensemble des règles découlant de ces deux modèles peuvent être utilisées pour prédire la maladie sur de nouveaux sujets.

Le premier modèle "arbre de décision " donne en sortie de meilleurs résultats que la classification bayésienne, qui manque quelque peu de flexibilité sur cet échantillon. Le taux d'erreur global des arbres de décision est estimé à 10,45% avec une aire sous la courbe ROC égale à 0,964. Cinq variables, dont l'âge, le cholestérol, l'IMC adulte, l'activité

physique et le sexe sont des facteurs les plus importants pour prédire le diabète. Par rapport aux deux modèles, seuls l'âge, l'activité physique et le cholestérol sont les facteurs de risque communs.

Par ailleurs, la démarche adoptée dans ce cadre-là, ainsi que les résultats qui en découlent ont été utilisés dans ce contexte pour la première fois et s'avèrent être intéressants pour l'exploration des données selon l'avis des professionnels de la santé, même si elles restent insuffisantes pour une meilleure appréciation. Par conséquent, d'autres améliorations sont possibles.

Il serait certainement plus judicieux d'inclure les différents types de diabètes pour ensuite faire la distinction. De plus, il serait intéressant d'envisager une base de données plus consistante et plus adéquate et intégrer d'autres méthodes de prédiction telle que les SVM et la régression logit afin de mieux valider les résultats obtenus et rendre la démarche proposée plus générique et donc applicable à d'autres maladies.

Pour finir, ce travail montre la nécessité de considérer de nouvelles approches statistiques pour l'analyse des données en épidémiologie. Le data mining est une des alternatives crédibles à comparer aux outils conventionnels.

Références bibliographiques

Guerin-Dubourg A., (2004). « *Etude des modifications structurales et fonctionnelles de l'albumine dans le diabète de type 2 : identification de biomarqueurs de glycoxydation et de facteurs de risque de complications vasculaires*», THESE de Doctorat de l'université de la Réunion.

Santos F., (2015). « *Arbres de décision* », CNRS, UMR 5199 PACEA, 1-5

Piatesky Shapiro G., & **al., (1996).** « *From Data Mining to Knowledge Discovery: An Overview Advances in Knowledge Discovery and Data Mining*», AAAI Press MIT Press.

Tufféry S., (2012). « *Data mining et statistique décisionnelle, l'intelligence des données*», éd TECH NIP, 4éme édition, Paris, p 611.

Fontbonne A., (2010). «*Epidémiologie des états diabétiques* ». Liv .Masson; éd .France.3-7.

Saporta G., (2006). «*Probabilités, analyse des données et statistique* », Technip.

Nourizadeh A., & al. (2013). «*Analyse exploratoire de données sur l'insulinothérapie chez les patients âgés diabétiques de type 2*», Studia Informatica Universalis 11 (3), 32-49.

Sankaranarayanan S., (2014). «*Diabetic Prognosis through Data Mining Methods and Techniques*», International Conference on 162-166.

Hung-Chun Lin & al. (2011). «An Application of Artificial Immune Recognition System for Prediction of Diabetes Following Gestational Diabetes», J. *Medical Systems* 35(3): 283-289

Bagdi R. & al. (2012). « Diagnosis of Diabetes Using OLAP and Data Mining Integration», *International Journal of Computer Science & Communication Networks*, Vol 2(3), 314 -322

Stanfford G. C., & al., (1984). « *Recent improvements in and analytical applications of advanced ion-trap technology* », Intl. J. Mass Spectrometry Ion Processes. 60 : 85-98

Zeghari L., & al. (2017). «The overweight, the obesity and the glycemic control among diabetics of the provincial reference center of diabetes (CRD), Kenitra, Morocco», *Pan African Medical Journal*.; 27:189

Gregory L., & al. (2008). « *The Impact of Obesity on Cardiovascular Disease Risk Factors and Subclinical Vascular Disease: The Multi-Ethnic Study of Atherosclerosis»,* Intern Med.; 168(9):928-935. doi:10.1001/archinte.168.9.928

Sheen A.J., & al. (2003). «*Optimalisation de la prise en charge du patient diabétique de type 2 : Résultats de l'étude « DREAM » en médecine générale*», Rev Med Liege 2003 ; 58 : 3 : 139-146.

Lounis H., (2006). «*Arbres de décision. Notes de cours (Séminaire sur l'apprentissage automatique)* », Programme de Doctorat en Informatique Cognitive, Université du Québec à Montréal.

Geurts P., & al., (2009). « Supervised learning with decision tree-based methods in computional and systems biology», *Molecular BioSystems* 5(12):1 593-605 ·

Quinlan J. R., (1993). *« C4.5: Programs for Machine Learning»*, San Francisco. Morgan Kaufman

Breiman L., & al., (1984). « Breiman, L., Friedman, J., Charles, J. S., & Olshen , R. (1984). chapitre 11. In *Classification and Regression Trees (Anglais)* (illustrée, réimprimée, révisée). Wadsworth, New York. Taylor & Francis.

Shih. WY Loh & YS Shih (1997). «Split selection methods for classification trees», Institute of Statistica Sinica, Academia *Sinica,Vol. 7*, No. 4, pp. 815-840.

Rish I., (2001). «*An empirical study of the naive bayes classier* », IJCAI-01 workshop on Empirical Methods in AI.

Domingos P., & Pazzani M., (1997). «On the optimality of the simple Bayesian classier under zero-one loss », *Machine Learning, 29*, 103130.

Domingos P., & al., (1996). «*Beyond independence: Conditions for the optimality of the simple bayesian classifier*», International Conference on Machine Learning.

Parent E., & Bernier J., (2007). « *Le raisonnement bayésien, modélisation et Inférences* », Springer-Verlag France

Ripley B. D., (1996). «*Pattern Recognition and Neural Networks*», Cambridge University Press, Cambridge, United Kingdom.

Procopiou M., (2005). «Dépistage et diagnostic du diabète de type 2 : quels tests ? », *Rev Med Suisse* ; volume 1. 30418

Lounici A., & al., (2007). « Facteurs prédictifs cliniques du diabète de type 2 dépisté : l'obésité abdominale définie selon les seuils de l'ATP III est plus discriminative que celle de l'IDF», *Diabetes & amp; Metabolism* Vol 33, N° Spe1 - p. 129

James R. W., (2002). « Particularités de la dyslipidémie du diabète», *Rev Med Suise* ; volume -2. 21994

Kabamba AT, & al., (2014). « Decrease in HDL cholestérol indicator of oxidative stress in type 2 diabetes», *Pan Afr Med*; 19:140.

ANNEXES

Tableau 1 - Caractéristiques des attributs de la BD

Attribut	Type	Modalités	Attribut	Type
Sexe	Qualitatif	F / H	Age (ans)	Discret
Niveau_scolarité	Qualitatif	Oui/Non	Poids (kg)	Discret
Atteinte rénale	Qualitatif	Non / Oui	Taille (m)	Discret
Atteinte_Cardiaque	Qualitatif	Non/ Oui	Cholestérol (g/l)	Continue
ATC Familiaux	Qualitatif	Non/ Oui	IMC (kg/m^2): (taille*2)/poids	Continue
Tabagisme	Qualitatif	Non/ Oui	HTA SBP mmHg	Discret
Sédentarité	Qualitatif	Non/ Oui	HTA DBP mmHg	Discret
Cystite	Qualitatif	Non / Oui	Hyperglycémie (grossesse)	Qualitatif
Mycoses	Qualitatif	Non/ Oui	Avoir bébé plus de 4 kg	Qualitatif
Diagnostic	Qualitatif		diabétique / non diabétique	

Source : variables recueillies par les auteurs à partir du questionnaire