UNIVERSITÉ NATIONALE DU VIETNAM À HANOÏ
INSTITUT DE LA FRANCOPHONIE POUR L'INNOVATION

Option : Systèmes Intelligents et Multimédia (SIM)
Promotion: 22 Année académique 2017-2018

Module : Fouille de données et de recherche d'informations.

RAPPORT DE TRAVAUX PRATIQUES(TP1 TP2 TP3 TP4) DU GROUPE 10.

Présenté par

ZONGO Sylvain

OUBDA Raphaël Nicolas W

Professeur: NGUYEN Thi Minh Huyen

Table of Contents

1 Introduction 3

1 Analyse des variables simples et liens des paires de variables **4**

1 Base de données. 4

1.1 Choix de la base de données 4

1.2 Objectif de la base de données. 4

2 Analyse détaillée du jeu de données Fertility 4

2.1 Description des attributs 5

2.2 Normalisation des variables de la base de données 5

3 Les types de variables 6

3.1 Analyse des paires de variables. 7

3.2 Interprétation des résultats. 10

2 Analyse factorielle **11**

1 Analyse de l'attribut age et consommationAlcool. 11

2 La corrélation 12

3 Analyse des composants principaux. 12

3 Clustering **14**

1 Clustering avec la méthode de K-means 14

2 Clustering avec la méthode de VARCLUST1 de l'onglet clustering de tanagra 15

4 Etudes de deux algorithmes et leur application à notre jeux de données **18**

1 Etudes des algorithmes 19

1.1 LDA 19

1.2 SVM 20

2 Comparaison des resultats de LDA et de SVM 22

2.1 Utilisation de la fonction Scoring 22

2.2 ROC CURVE 23

3 Conclusion 24

Références bibliographiques 24

List of Figures

1.1 Visualisation des données. 6
1.2 Visualisation des attributs. 7
1.3 lien entre Saison maladie 8
1.4 lien entre fièvre et saison 8
1.5 lien entre fumer et interventionChirurgicale 9
1.6 lien entre fumer et consommationAlcool 9

2.1 Visualisation du résultat des variables age et consommationAlcool. 11
2.2 Corrélation. 12
2.3 Analyse des composants principaux. 13

3.1 K-means avec l'option VARCLUST1 de l'onglet clustering de tanagra 14
3.2 VARCLUST1 de l'onglet clustering de tanagra 15
3.3 regroupement des attributs 16
3.4 Hiérachie des groupes 17

4.1 repartition des données d'apprentissage et de test 18
4.2 repartition des données d'apprentissage et de test 19
4.3 répartition des données d'apprentissage et de test 20
4.4 Apprentissage avec SVM 21
4.5 Test avec SVM 22
4.6 Caption 22
4.7 répartition des données d'apprentissage et de test 23
4.8 24

1 Introduction

De nos jours la Data Science (data mining) est perçue comme ensemble des techniques d'exploration de données permettant d'extraire des connaissances sous forme de modèles de description. Ainsi divers outils sont utilisés pour faciliter cette exploration parmi lesquels nous avons Tanagra, Weka, R, SPSS etc. Ce TP1 nous a été donné afin de faire une prise de main avec l'un des outils. Nous allons donc utiliser pour le traitement d'une base de données que nous choisirons sur le site `http://archive.ics.uci.edu/ml/datasets.html` afin de répondre à cette perspective. Le jeu de données que nous avons choisi est : Fertility Data Set. Nous ferons l'exploration de ce jeu de données à l'aide de l'outil Tanagra ce jeu de données relate les données sur la fertilité des hommes en fonction de la concentration des spermatozoïdes. Cette prédiction est liée aux données sociodémographiques, aux facteurs environnementaux, à l'état de santé et aux habitudes de vie. Ainsi pour notre travail, nous avons fixé comme objectif de prédire la fertilité d'un homme à partir de diagnostiques.
Notre travail comporte quatre grandes parties. La première partie est l'analyse des paires de variable, la deuxième partie l'analyse factorielle, la troisième partie le K-means et nous terminerons par l'étude de deux algorithmes de prédiction ainsi qu'à leur compraison.

Chapitre 1

Analyse des variables simples et liens des paires de variables

1 Base de données.

1.1 Choix de la base de données

Nous avons travaillé sur le jeu de données **«Fertility»**. Les données sont issues du site `http://archive.ics.uci.edu/ml/datasets/Fertility`. Le jeu de données comporte dix(10) attributs et contient 100 données.

1.2 Objectif de la base de données.

Les taux de fécondité ont considérablement diminué au cours des deux dernières décennies, en particulier chez les hommes. Il a été décrit que les facteurs environnementaux, ainsi que les habitudes de vie, peuvent affecter la qualité du sperme. C'est de ce là dont traite la base de données Fertility, qui a pour but de prédire la fertilité chez les hommes de dix-huit(18) à trente six(36) ans. De ce fait nous avons travaillé sur le jeu de données Fertility afin de prédire la fertilité chez les hommes. Cette base est issues d'un échantillonnage de cent(100) volontaires sains qui ont fourni leur sperme analysé selon les critères de l'OMS 2010. La concentration des spermatozoïdes est liée aux données socio-démographiques, aux facteurs environnementaux, à l'état de santé et aux habitudes de vie des individus . Cette base a un jeu de données très interessante dans la mesure où son objectif est de prédire la fertilité d'un homme à partir de diagnostiques suivants des facteurs environnemental et comportemental[2].

2 Analyse détaillée du jeu de données Fertility

Le problème posé dans notre jeu de données est de prédire la fertilité d'un homme à partir de diagnostiques suivant des facteurs environnemental et comportemental. Notre étude s'est porté sur 100

individus volontaires sains de sexe masculin de 18 à 36 ans. Nous avons 10 attributs dans notre base de données.

2.1 Description des attributs

- **saison:** il détermine la saison dans laquelle l'analyse a été réalisée, il peut prendre les valeurs suivantes: 1) l'hiver, 2) le printemps, 3) l'été, 4) l'automne. Ces valeurs sont normalisées suivant les valeurs : -1, -0,33, 0,33, 1. Ainsi l'attribut **saison** a donc 4 valeurs.

- **age:** cet attribut donne l'âge de l'individu au moment de l'analyse. Cette valeur 18-36 est la plage d'âge de la population étudiée qui est normalisée à l'intervalle (0, 1) ce qui signifie que dans la base de données l'âge est dans l'intervalle[0,1];

- **maladieInfantile:**Il précise si l'individu a eu à souffrir d'une maladie infantile (c'est à dire Varicelle, rougeole, oreillons, polio) 1) oui, 2) non. La réponse est normalisée avec les valeurs binaires (0, 1);

- **accidentTraumatisme:** il précise si l'individu a eu un accident traumatisme avant l'analyse.. 1) Oui, 2) Non. La réponse est normalisée avec les valeurs binaires (0, 1);

- **interventionChirurgicale:**précise si l'individu a eu a subir une opération chirurgicale. 1) oui, 2) non. La réponse est normalisée avec les valeurs binaires (0, 1);

- **fièvre:** cet attribut précise si l'individu a eu une fièvre élevée au cours de la dernière année 1) il y a moins de trois mois, 2) il y a plus de trois mois, 3) non. La normalisation permet d'assigner respectivement les valeurs suivantes (-1, 0, 1);

- **consommationAlcool:** précise la fréquence de consommation d'alcool de l'individu 1) plusieurs fois par jour, 2) tous les jours, 3) plusieurs fois par semaine, 4) une fois par semaine, 5) rarement ou jamais (0, 1);

- **fumeur:** montre si l'individu a l'habitude de fumer. Habitude de fumer 1) jamais, 2) occasionnel 3) quotidien. (-1, 0, 1);

- **nbrHAssisJour:** précise le nombre d'heure que l'individu passe en étant assis par jour. Nombre d'heures passées assis par jour ene-16 (0, 1);

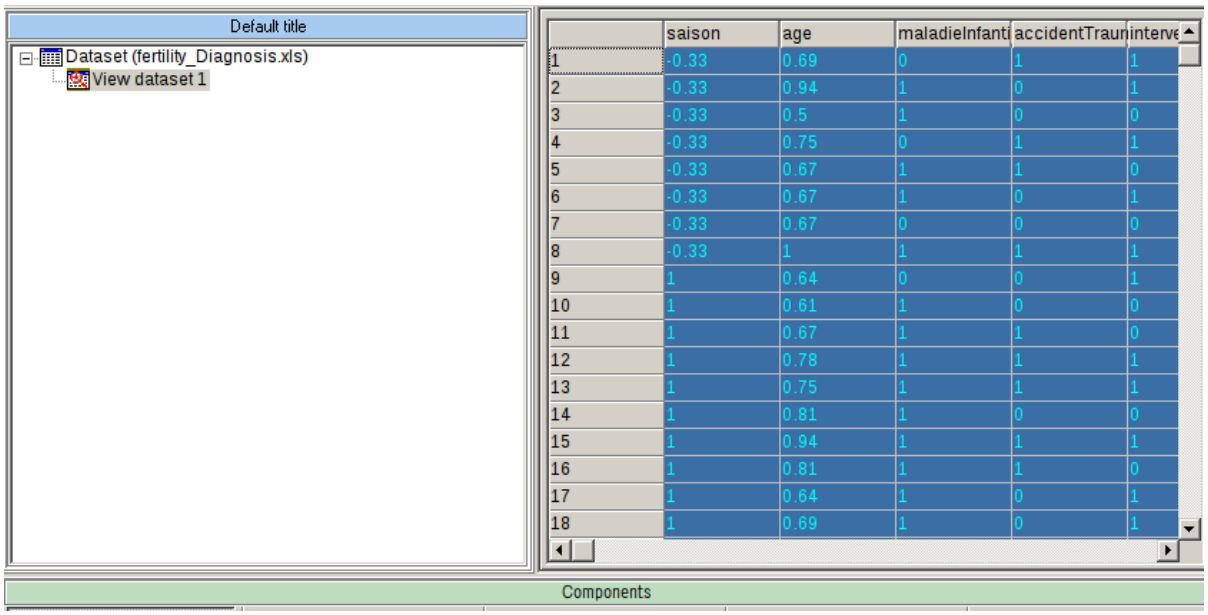
- **diagnostic:** est la variable de sortie c'est à dire à prédire elle montre si l'individu est fertile ou pas . Diagnostic normal (N), modifié (O).

2.2 Normalisation des variables de la base de données

— Les variables numériques, tels que l'âge, sont normalisés sur l'intervalle (0-1).

— Les variables avec seulement deux attributs indépendants sont arrangés au préalable avec des valeurs binaires (0, 1).
— Les variables avec trois attributs indépendants, tels que **«Les vaccins reçus»** , **«Une fièvre élevée l'année dernière»** et **«l'habitude de fumeur»**, sont préalablement arrangés en utilisant les valeurs ternaires (-1, 0, 1).
— Les variables avec quatre attributs indépendants, tels que **«saison où l'analyse a été effectuée »** ou **«état civil »**, sont préétabli en utilisant les quatre valeurs de distance différentes et égales (-1, -0,33, 0,33, 1).

Notre jeu de données contient 100 obseravtions et 10 attributs. Il n'existe pas de valeurs manquantes dans notre base de données FERTILITY, cependant ce jeu de données possède des valeurs (N/A) c'est dire des valeurs non applicables ou non compatibles. Ces lignes représentent les résultats recueillis lors de l'analyse du sperme. Sur Tanagra à l'aide de l'option View data set de l'onglet Data visualization nous avons visualisé les différentes données de notre jeu de données [1].



	saison	age	maladieInfanti	accidentTraun	interve
1	-0.33	0.69	0	1	1
2	-0.33	0.94	1	0	1
3	-0.33	0.5	1	0	0
4	-0.33	0.75	0	1	1
5	-0.33	0.67	1	1	0
6	-0.33	0.67	1	0	1
7	-0.33	0.67	0	0	0
8	-0.33	1	1	1	1
9	1	0.64	0	0	1
10	1	0.61	1	0	0
11	1	0.67	1	1	0
12	1	0.78	1	1	1
13	1	0.75	1	1	1
14	1	0.81	1	0	0
15	1	0.94	1	1	1
16	1	0.81	1	1	0
17	1	0.64	1	0	1
18	1	0.69	1	0	1

Figure 1.1 – Visualisation des données.

3 Les types de variables

Notre jeu de données comporte deux types de variables:

- **Variables qualitatives:** diagnostic, fumeur, maladieInfantile, accidentTraumatisme, interventionChirurgicale, Fievre, saison
- **Variables quantitatifs:** age, consommationAlcool, nbrHAssisJour.

Variables input: maladieInfantile, accidentTraumatisme, interventionChirurgicale, Fievre, fumeur, saison, age, consommationAlcool, nbrHAssisJour.

Variables output : diagnostic.

Après l'importation de notre jeu de données sur Tanagra nous pouvouns visualiser les différents attributs.

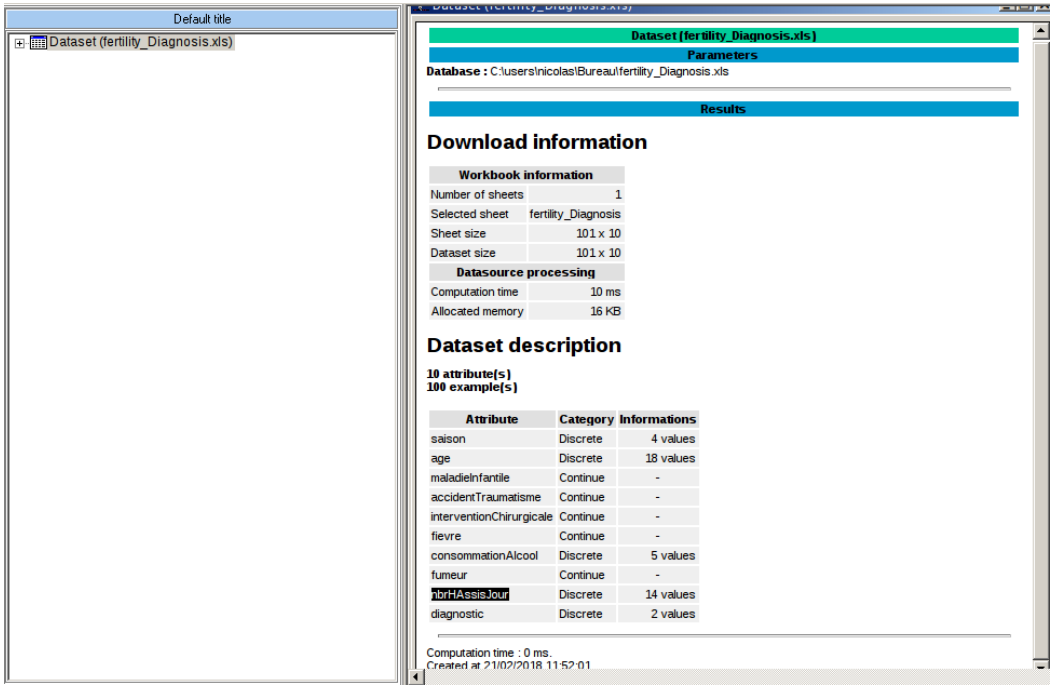


Figure 1.2 – Visualisation des attributs.

3.1 Analyse des paires de variables.

Hypothèses

Supposons qu'il existe une relation entre les variables saisons et maladieInfantile.

Supposons qu'il existe une relation entre la variable fumeur et la variable interventionChirurgicale.

Supposons qu'il existe une relation entre les variables accidentTraumatisme et interventionChirurgicale

Supposons qu'il existe une relation entre les variables fievre et saison.

Test de Khi-deux

Dans cette partie nous avons transformmés les variables discrètes(age,consommationAlcool, nbrHAssis-Jour et les variables binaires (maladieInfantile,accidenttraumatisme,interventionChirugicale) en variables continues.

1. Test de Khi-deux entre les variables saison et maladieIfantile

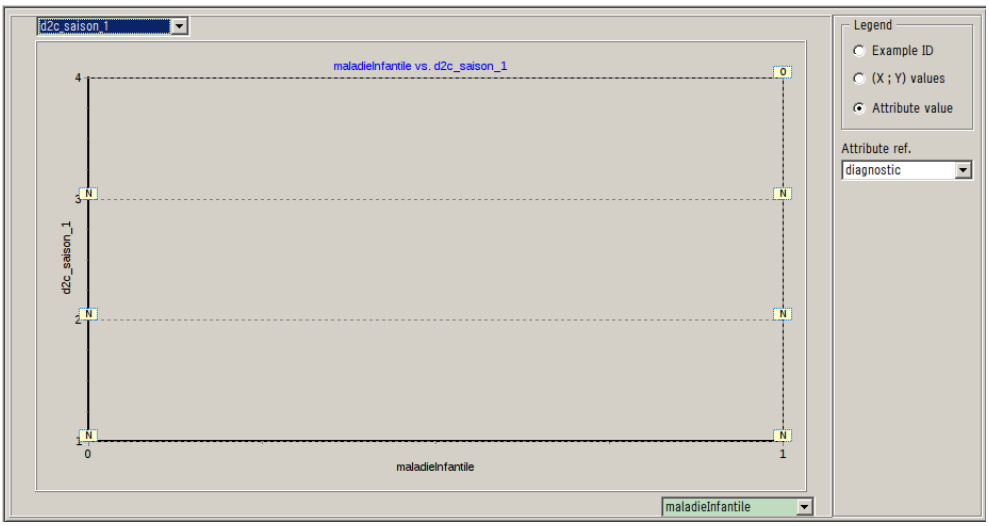


Figure 1.3 – lien entre Saison maladie

Interpretation: Les individus qui ont eu des maladies infantiles en automne ont eu une modification de leurs spermatozoides. Ce qui signifie qu'il ya une relation entre ces deux variables.

2. Testt de Khi-deux entre les variables fièvre et saison

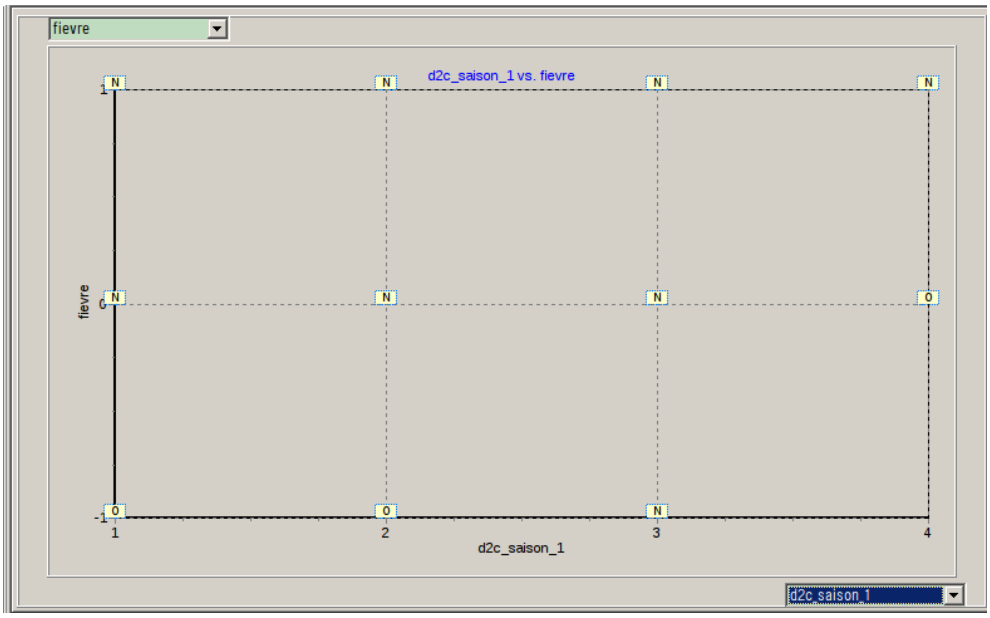


Figure 1.4 – lien entre fièvre et saison

Interprétation: Les individus ayant eu une forte fièvre il y a au plus trois mois en Hiver, au Printemps et en Automne présentent une modification de leurs spermatozoides. Notre hypothèse est donc vérifiée. Il ya une relation entre ces deux variables.

3. Test de Khi-deux entre les variables fumeur et interventionChirurgicale

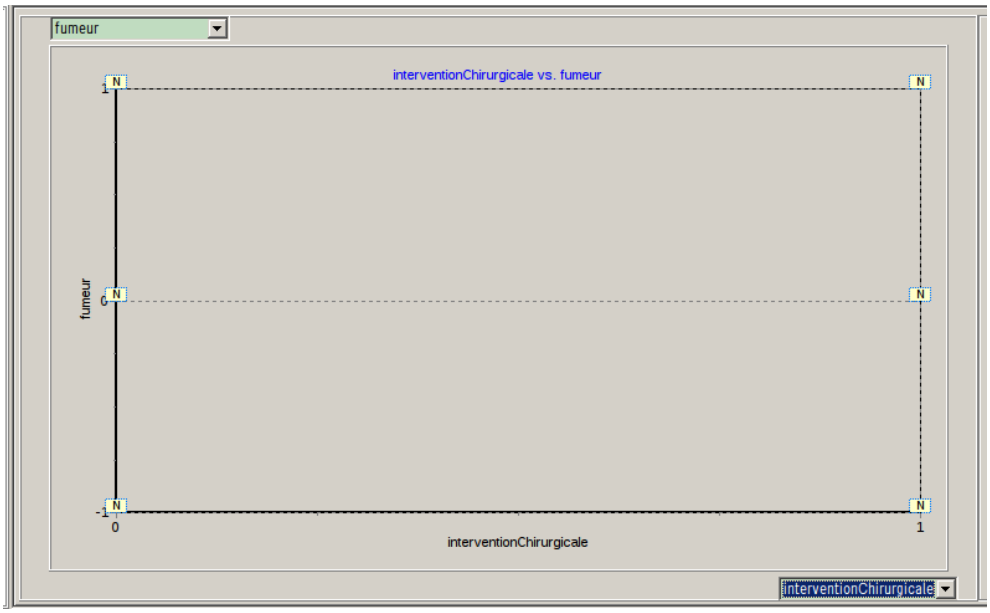


Figure 1.5 – lien entre fumer et interventionChirurgicale

Interpretation: Nous observons qu'aucun individus ne présente aucune modfication de ses spermatozoides. Ce qui montre que notre hypothèse est rejettée.

4. Test de Khi-deux entre les variables fumeur et interventionChirurgicale

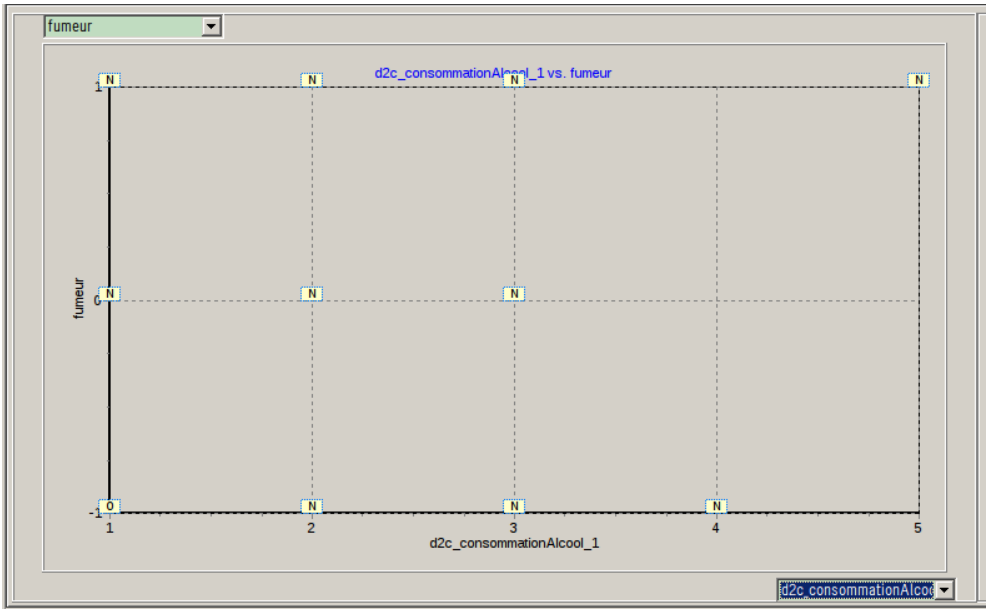


Figure 1.6 – lien entre fumer et consommationAlcool

Interpretation: Nous observons que les individus qui fument quotiennement et qui ne boivent pas ont leurs spermatozoides modifiés.

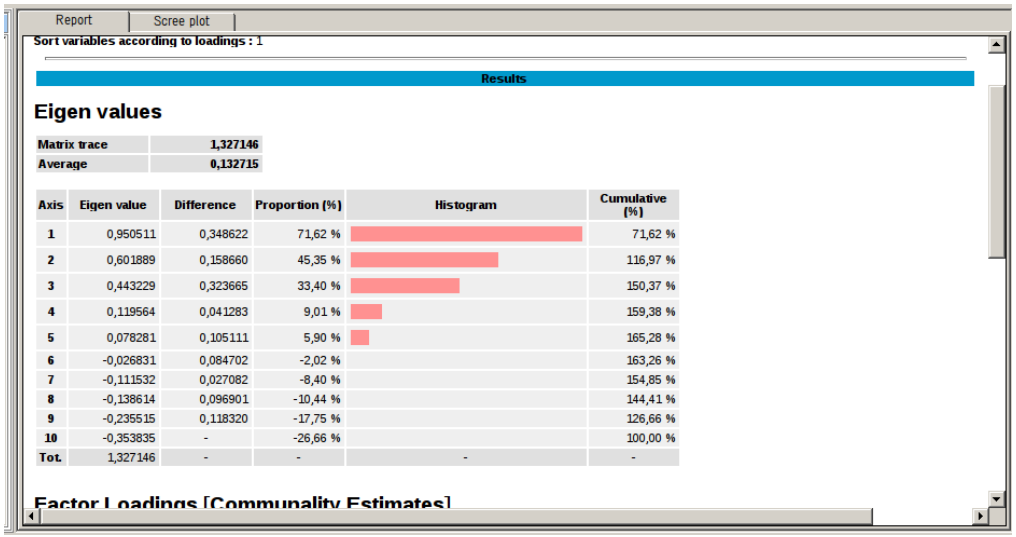
3.2 Interprétation des résultats.

A l'issue de la figure ci-dessus nous pouvons dire que les variables maladieInfantile et accidentTraumatisme représentent les 65.06% des informations. De ce fait ces variables sont les facteurs principaux agissant sur la concentration des spermatozoïdes dans le sperme.

Chapitre 2

Analyse factorielle

1 Analyse de l'attribut age et consommationAlcool.


Report | Scree plot

Sort variables according to loadings : 1

Results

Eigen values

Matrix trace	1,327146
Average	0,132715

Axis	Eigen value	Difference	Proportion (%)	Histogram	Cumulative (%)
1	0,950511	0,348622	71,62 %		71,62 %
2	0,601889	0,158660	45,35 %		116,97 %
3	0,443229	0,323665	33,40 %		150,37 %
4	0,119564	0,041283	9,01 %		159,38 %
5	0,078281	0,105111	5,90 %		165,28 %
6	-0,026831	0,084702	-2,02 %		163,26 %
7	-0,111532	0,027082	-8,40 %		154,85 %
8	-0,138614	0,096901	-10,44 %		144,41 %
9	-0,235515	0,118320	-17,75 %		126,66 %
10	-0,353835	-	-26,66 %		100,00 %
Tot.	1,327146	-	-	-	-

Factor Loadings [Communality Estimates]

Figure 2.1 – Visualisation du résultat des variables age et consommationAlcool.

Interprétation:
Nous constatons que la plupart des volontaires c'est à dire 92% ont l'age compris entre 18-29 ans. Les volontaires dont l'age est compris entre 30-35 représentent 8% de la population étudiée. Cette observation est faite suivant la variable age. Selon les observations suivant la consommation alcool, 39% des volontaires consomment rarement l'alcool, 40% consomment une fois par semaine, 19% consomment plusieurs fois par semaine, 2% consomment l'alcool tous les jours.

2 La corrélation

Dans cette partie nous allons donner les dépéndances entre les différentes variables.

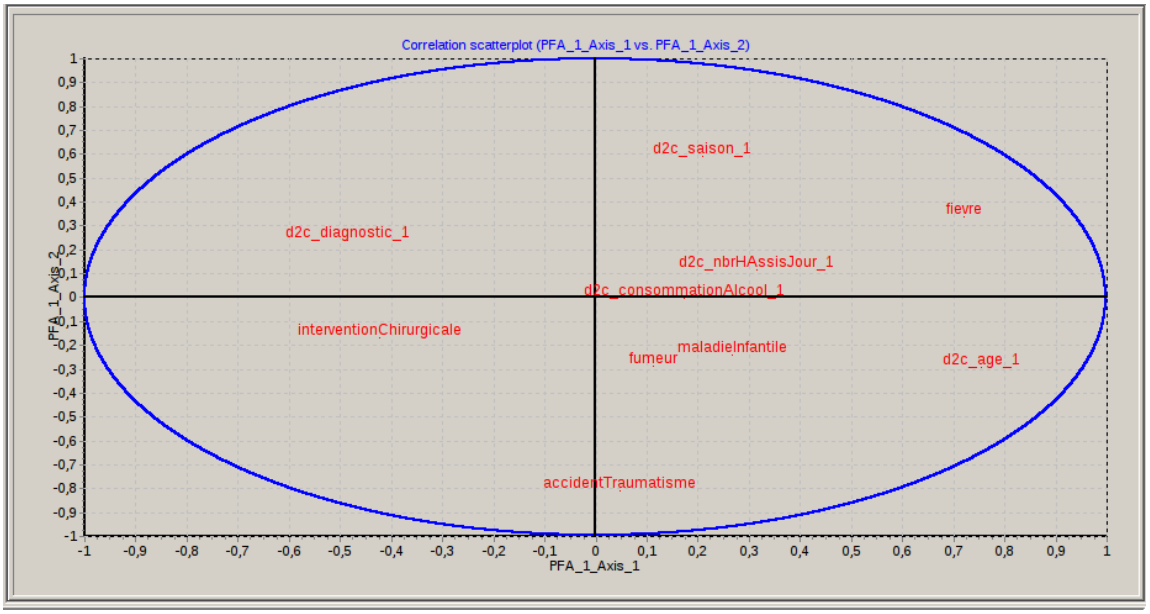
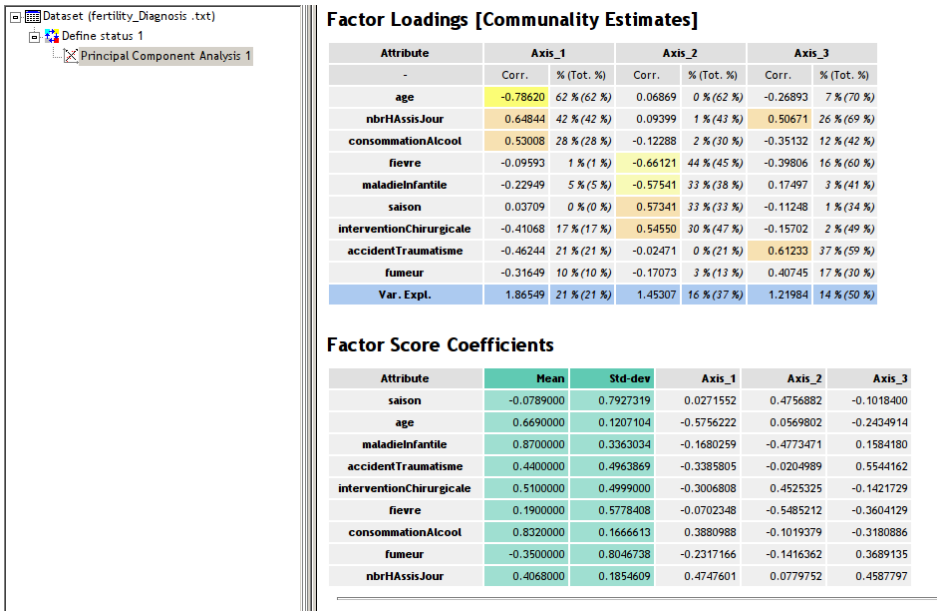


Figure 2.2 – Corrélation.

Interprétation:
le diagramme nous motre que les variables fumer et maladieInfantile sont fortement corrélées positivement, il existe donc une dependance très etroite entre les deux.
accidentTraumatisme et interventionChirurgicale sont fortement collorées positivement dont la plus part des individus qui ont eu un accident traumatisseme ont subi une intervention chirurgicale.
NbrHAssis et consommationAlcool sont fortement corrélées, ce qui s'assayent pendant long temps consomment assez l'alcool. Les variables maladieInfantile et fievre sont corellées négativement.

3 Analyse des composants principaux.

L'analyse des composants principaux(ACP) a pour objectif de trouver un ensemble d'axes de faible dimension qui résument les données. Nous allons appliquer l'ACP à notre jeu de données sur trois axes. Nous obtenons la figure ci-dessous:



Factor Loadings [Communality Estimates]

Attribute	Axis_1		Axis_2		Axis_3	
-	Corr.	% (Tot. %)	Corr.	% (Tot. %)	Corr.	% (Tot. %)
age	-0.78620	62 % (62 %)	0.06869	0 % (62 %)	-0.26893	7 % (70 %)
nbrHAssisJour	0.64844	42 % (42 %)	0.09399	1 % (43 %)	0.50671	26 % (69 %)
consommationAlcool	0.53008	28 % (28 %)	-0.12288	2 % (30 %)	-0.35132	12 % (42 %)
fievre	-0.09593	1 % (1 %)	-0.66121	44 % (45 %)	-0.39806	16 % (60 %)
maladieInfantile	-0.22949	5 % (5 %)	-0.57541	33 % (38 %)	0.17497	3 % (41 %)
saison	0.03709	0 % (0 %)	0.57341	33 % (33 %)	-0.11248	1 % (34 %)
interventionChirurgicale	-0.41068	17 % (17 %)	0.54550	30 % (47 %)	-0.15702	2 % (49 %)
accidentTraumatisme	-0.46244	21 % (21 %)	-0.02471	0 % (21 %)	0.61233	37 % (59 %)
fumeur	-0.31649	10 % (10 %)	-0.17073	3 % (13 %)	0.40745	17 % (30 %)
Var. Expl.	1.86549	21 % (21 %)	1.45307	16 % (37 %)	1.21984	14 % (50 %)

Factor Score Coefficients

Attribute	Mean	Std-dev	Axis_1	Axis_2	Axis_3
saison	-0.0789000	0.7927319	0.0271552	0.4756882	-0.1018400
age	0.6690000	0.1207104	-0.5756222	0.0569802	-0.2434914
maladieInfantile	0.8700000	0.3363034	-0.1680259	-0.4773471	0.1584180
accidentTraumatisme	0.4400000	0.4963869	-0.3385805	-0.0204989	0.5544162
interventionChirurgicale	0.5100000	0.4999000	-0.3006808	0.4525325	-0.1421729
fievre	0.1900000	0.5778408	-0.0702348	-0.5485212	-0.3604129
consommationAlcool	0.8320000	0.1666613	0.3880988	-0.1019379	-0.3180886
fumeur	-0.3500000	0.8046738	-0.2317166	-0.1416362	0.3689135
nbrHAssisJour	0.4068000	0.1854609	0.4747601	0.0779752	0.4587797

Figure 2.3 – Analyse des composants principaux.

Interprétation

Nous constatons que sur l'axe 1, les variable Nombre de jours assis(nbAssis) et la variable comsommation d'alcool (consommationAlchool) sont fortement corrélés positivement donc il ya un lien entre ces deux variables.

Sur l'axe 2 les variables saison et intervention chirurgicale sont fortement corrélées positivement donc l'intervention chirurgicale est liée à la saison. Sur ce même axe les variables fièvre et maladie infantile sont corrélées négativement. Donc si l'individus n'a pas souffert de maladie infantile il a moins de fièvre.

Sur l'axe 3 nous constatons que nombre de jours assis(nbJoursAssis) et accidentTraumatissement sont fortement corrélés positivement donc plus l'individus passe plus de temps entant assis plus il est plus exposé aux accident traumatissement. Ainsi nous avons un lien entre nombre de jours assis et la variable accidentTraumatissment.

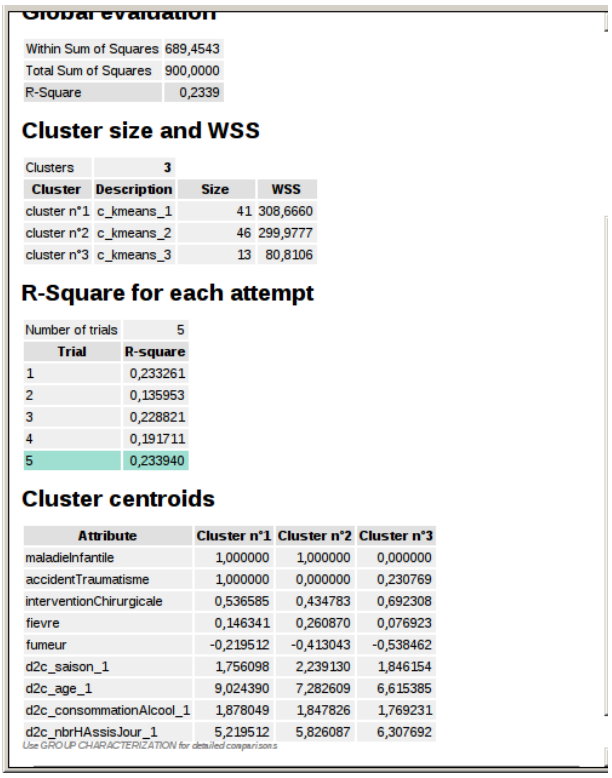
L'ACP nous a permis de regroupé nous donnée en fonction de trois axes. Ce algorithme améliorera ainsi nos résultats lors de la régresssion à travers les algorithmes LDA et SVM

Chapitre 3

Clustering

Dans cette partie nous avons transformé toutes nos variables discètes en continues. Ensuite nous avons défini un Define status dans lequel nous mettons en entrée toutes nos variables continues et en sortie notre variable de sortie diagnostic.

1 Clustering avec la méthode de K-means



Global evaluation

Within Sum of Squares	689,4543
Total Sum of Squares	900,0000
R-Square	0,2339

Cluster size and WSS

Clusters	3		
Cluster	**Description**	**Size**	**WSS**
cluster n°1	c_kmeans_1	41	308,6660
cluster n°2	c_kmeans_2	46	299,9777
cluster n°3	c_kmeans_3	13	80,8106

R-Square for each attempt

Number of trials	5
Trial	**R-square**
1	0,233261
2	0,135953
3	0,228821
4	0,191711
5	0,233940

Cluster centroids

Attribute	Cluster n°1	Cluster n°2	Cluster n°3
maladieInfantile	1,000000	1,000000	0,000000
accidentTraumatisme	1,000000	0,000000	0,230769
interventionChirurgicale	0,536585	0,434783	0,692308
fievre	0,146341	0,260870	0,076923
fumeur	-0,219512	-0,413043	-0,538462
d2c_saison_1	1,756098	2,239130	1,846154
d2c_age_1	9,024390	7,282609	6,615385
d2c_consommationAlcool_1	1,878049	1,847826	1,769231
d2c_nbrHAssisJour_1	5,219512	5,826087	6,307692

Use GROUP CHARACTERIZATION for detailed comparisons

Figure 3.1 – K-means avec l'option VARCLUST1 de l'onglet clustering de tanagra

Interprétation

Ces données ci-dessus sont les résultats obtenus à partir du traitement avec la méthode de K-means. Nous avons trois clusters cluster1 avec 50 éléments , cluster2 avec 37 éléments et cluster3 avec 13 éléments. Les distances entre les différentes attributs nous permettent de regrouper les attributs pour former des groupe. De ce fait les variables maladieInfantile, saison, fumeur, accidentTraumatissement peuvent etre regroupés dans un même groupe. Les autres variables ne peuvent pas être regroupées car la distance qui les sépare est très grande, ils forment ainsi chacune un cluster.

2 Clustering avec la méthode de VARCLUST1 de l'onglet clustering de tanagra

Report | Splitting sequence

3	1	1,0000	1,0000
4	1	1,0000	1,0000
5	1	1,0000	1,0000
6	4	1,3222	0,3305
Total		6,3222	0,7025

Cluster members and R-square values

Cluster	Members	Own Cluster	Next Closest	1-R² ratio
1	fievre	1,0000	0,0941	0,0000
2	d2c_age_1	1,0000	0,0941	0,0000
3	d2c_nbrHAssisJour_1	1,0000	0,0335	0,0000
4	interventionChirurgicale	1,0000	0,0536	0,0000
5	d2c_consommationAlcool_1	1,0000	0,0252	0,0000
6	maladieInfantile	0,1795	0,0141	0,8323
	accidentTraumatisme	0,6129	0,0179	0,3942
	fumeur	0,2191	0,0167	0,7942
	d2c_saison_1	0,3107	0,0007	0,6898

Cluster correlations -- Structure

Attribute	# membership	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
maladieInfantile	0	0,0756	0,1187	-0,0858	-0,1410	0,0377	0,4236
accidentTraumatisme	1	-0,0823	0,1337	-0,1350	0,1032	0,0389	0,7829
interventionChirurgicale	1	-0,2316	-0,1823	0,1088	1,0000	0,0157	0,0295
fievre	1	1,0000	0,3068	0,1832	-0,2316	0,1012	-0,1258
fumeur	0	-0,0075	0,1293	-0,0273	-0,0534	-0,0037	0,4681
d2c_saison_1	1	0,2344	-0,0266	-0,0826	-0,0776	-0,0020	-0,5574
d2c_age_1	1	0,3068	1,0000	0,1568	-0,1823	0,0362	0,1743
d2c_consommationAlcool_1	1	0,1012	0,0362	0,1588	0,0157	1,0000	0,0346
d2c_nbrHAssisJour_1	1	0,1832	0,1568	1,0000	0,1088	0,1588	-0,0824

Standardized Regression coefficients

Figure 3.2 – VARCLUST1 de l'onglet clustering de tanagra

Interprétation

Ces données ci-dessus sont les résultats obtenus à partir du traitement avec l'option VARCLUST1 de l'onglet clustering. Le tableau représente le CLUSTER MEMBERS AND R-SQUARE VALUES.
Nous obtenons six groupes:

1. cluster 1 composé d'un membre;
2. cluster 2 composé d'un membre;
3. cluster 3 composé d'un membre;
4. cluster 4 composé d'un membre;
5. cluster 5 composé d'un membre;

6. cluster 6 composé de quatre.

Les valeurs deR^2 permettent de comprendre la pertinence du model choisi:

— Si R^2 est proche de 1 alors le model est proche de la réalité. Ainsi 1- $R^2 = 0$, ce qui permet de dire que R^2est très proche de 1 et illustré dans les 5 groupes donc le model représente bien la réalité.

— Si R^2 est proche de 0 alors le model explique très mal la réalité. Cette différence étant non nulle alors nous pouvons dire que le model représente partiellent la réalite dans le 6eme groupe.

Pour finaliser cette interprétation les critères de clustering dépendent de la similitude et de la distance existente entre la population étudiée. En conclusion nos onze(11) attributs ont été regroupés en en six groupes:

Cluster	Members	Own Cluster	Next Closest	1-R² ratio
1	fievre	1,0000	0,0941	0,0000
2	d2c_age_1	1,0000	0,0941	0,0000
3	d2c_nbrHAssisJour_1	1,0000	0,0335	0,0000
4	interventionChirurgicale	1,0000	0,0536	0,0000
5	d2c_consommationAlcool_1	1,0000	0,0252	0,0000
6	maladieInfantile	0,1795	0,0141	0,8323
	accidentTraumatisme	0,6129	0,0179	0,3942
	fumeur	0,2191	0,0167	0,7942
	d2c_saison_1	0,3107	0,0007	0,6898

Figure 3.3 – regroupement des attributs

Nous obtenons ainsi la hiérachie des groupes ci-dessous:

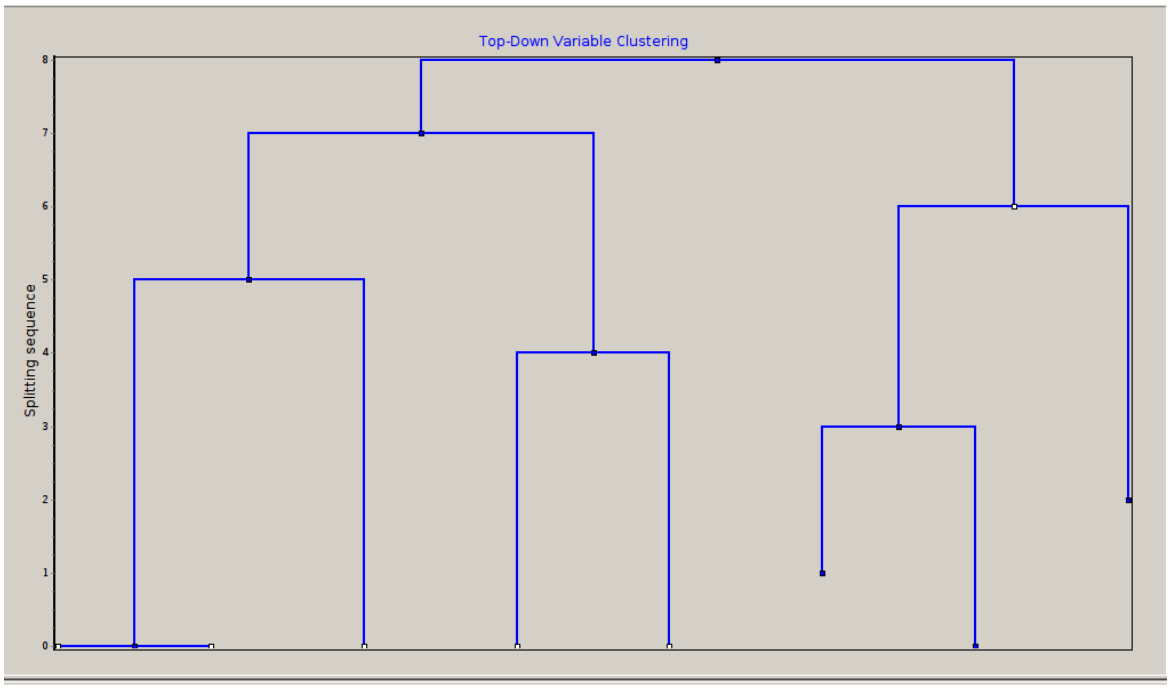


Figure 3.4 – Hiérachie des groupes

Chapitre 4

Etudes de deux algorithmes et leur application à notre jeux de données

Cette partie des TPS(TP5) fait l'objet de l'étude de deux algorithmes et leur application sur notre jeux de données afin de faire une comparaison de ces algorithmes. Nos algorithmes choisis ici sont SVM et LDA. Après une briève description de ces algoritmes nous allons les appliquer sur notre jeu de données Fertility dont l'objectif est de prédire la fertilité chez les hommes de dix-huit(18) à trente six(36) ans . Cette base a un jeu de données très interessante dans la mesure où son objectif est de prédire la fertilité d'un homme à partir de diagnostiques suivants des facteurs environnemental et comportemental. Elle comporte un jeux de données de 100 individus.

Scinder les données en données d'apprentissage et de test.

— 70 individus pour l'apprentissage

— 30 individus pour le test

Cette division de données sera appliquées aux deux algorithmes afin de les comparer. Nous mettons en input tous les variables d'entrées qui sont continues et en sortie la variable à prédire diagnostic.Pour ce faire, nous utilisons le composant SAMPLING, nous le paramétrons comme suit.

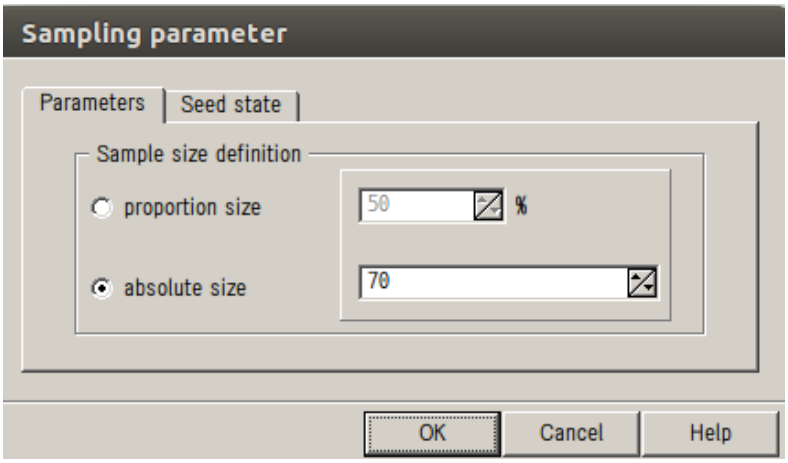


Figure 4.1 – repartition des données d'apprentissage et de test

1 Etudes des algorithmes

1.1 LDA

L'analyse discriminante linéaire (aussi LDA, en anglais : linear discriminant analysis) est une technique d'analyse discriminante prédictive. Il s'agit d'expliquer et de prédire l'appartenance d'un individu à une classe (groupe) prédéfinie à partir de ses caractéristiques mesurées à l'aide de variables prédictives. L'Analyse Linéaire Discriminante est une méthode simple de discrimination basée sur une modélisation probabiliste des données. On veut classifier des individus qui peuvent appartenir à la classe positive + ou à la classe négative (discrimination binaire). Ainsi dans notre cas nous allons classer les individus en deux groupes, les individus dont les spermatozoïdes présentent des modifications(-) et les individus dont les spermatozoïdes ne présentent aucune modification.

Apprentissage avec LDA

Pour l'apprentissage nous allons prendre 2/3 de nos individus (70 individus) comme données d'apprentissage et 1/3(30 individus) pour les tests. La figure ci-dessous présente les différents résultats obtenus après avoir utilisé l'algorithme de LDA.

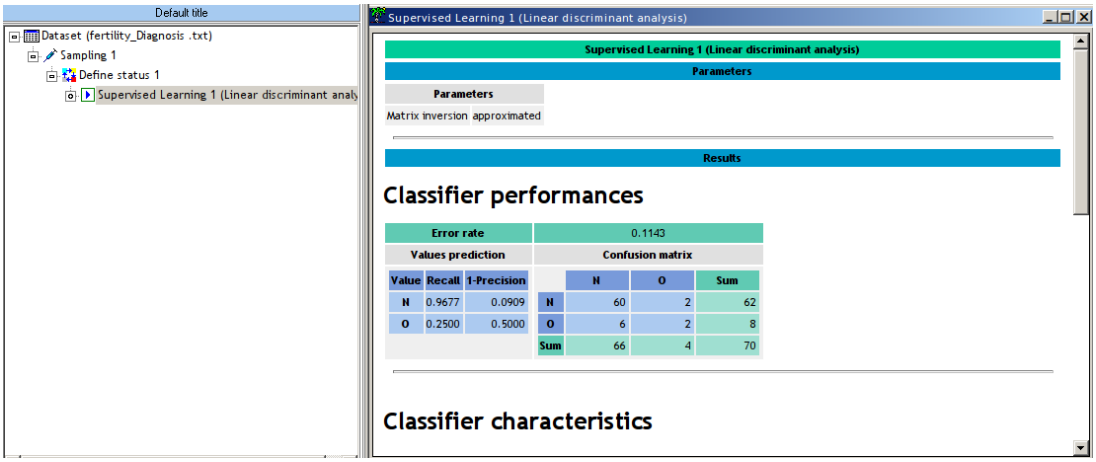


Figure 4.2 – repartition des données d'apprentissage et de test

Interprétation

Les résultats montrent que nos données ont été classifiées en deux groupes. Le premier groupe constitue les individus(O) dont les spermatozoides ont subi une modification tandisque le deuxième groupe constitue les individus dont les spermatozoides n'ont pas subi de modification. Les données ont été classées avec un taux d'erreur de 0,1143. En effet, pour les individus présentant des spermatozoides non modifiés, le taux de précision est (0,9091) et pour les spermatozoides modifiés 0,5. Donc pour les individus non modifié le taux de précision est très élevé tandisque pour les individus modifié le taux de précision est de 0,5 ce qui n'est très bon mais acceptable. En somme LDA donne un taux de précision de 0,9091 et 0,0919 de taux d'erreur. De ce fait, nous allons appliquer notre algorithme aux données de test c'est-à- dire les 30 individus restant.

Test avec LDA

La figure ci-dessous présente les résultats obtenus pour les données de test.

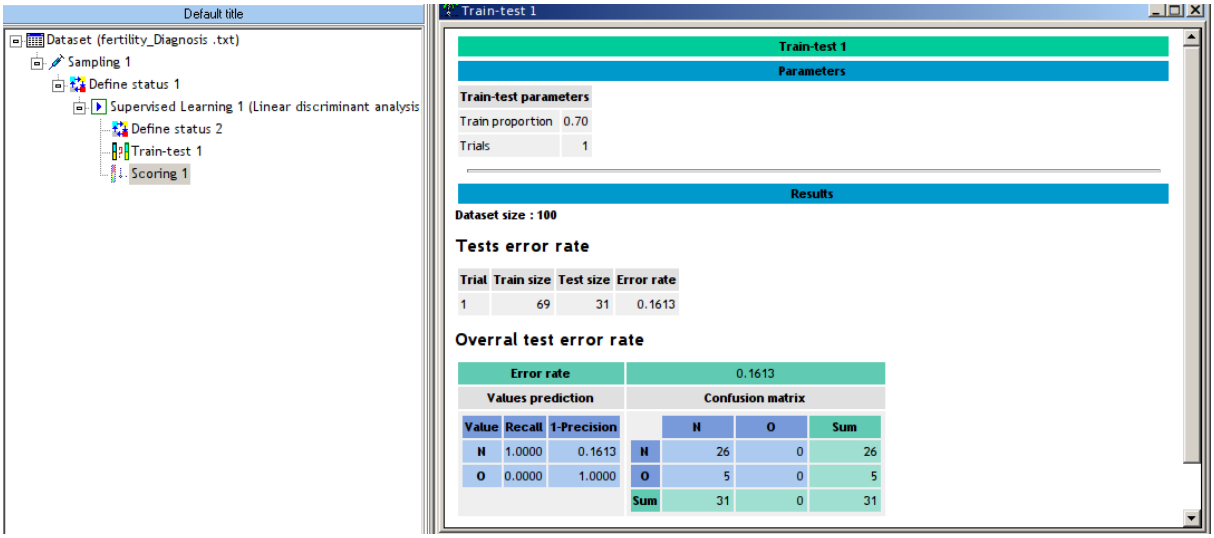


Figure 4.3 – répartition des données d'apprentissage et de test

Les résultats montrent un taux d'erreur de 0,1613 donc un taux de précision de 0,9387. En effet, nous obtenons un bon taux de précision de notre algorithme.

1.2 SVM

Les SVM peuvent être utilisés pour résoudre des problèmes de discrimination, c'est-à-dire décider à quelle classe appartient un échantillon, ou de régression, c'est-à-dire prédire la valeur numérique d'une variable. La résolution de ces deux problèmes passe par la construction d'une fonction qui à un vecteur d'entrée fait correspondre une sortie. Les principes de l'algoritme SVM La première consiste à définir l'hyperplan comme solution d'un problème d'optimisation sous contraintes dont la fonction objectif ne s'exprime qu'à l'aide de produits scalaires entre vecteurs et dans lequel le nombre de contraintes "actives" ou vecteurs supports contrôle la complexité du modèle. Le passage à la recherche de surfaces séparatrices non linéaires est obtenu par l'introduction d'une fonction noyau (kernel) dans le produit scalaire induisant implicitement une transformation non linéaire des données vers un espace intermédiaire (feature space) de plus grande dimension. D'où l'appellation couramment rencontrée de machine à noyau ou kernel machine . Sur le plan théorique, la fonction noyau définit un espace hilbertien, dit auto-reproduisant et isométrique par la transformation non linéaire de l'espace initial et dans lequel est résolu le problème

Forces de SVM

— Capacité à traiter de grandes dimensionnalités;
— Traitement des problèmes non linéaires avec le choix des noyaux;
— Paramétrage permet de la souplesse (ex. résistance au sur apprentissage avec C);
— Souvent performant dans les comparaisons avec les autres approches;
— Robuste par rapport aux points aberrants (contrôlé avec le paramètre
— Donne une bonne indication de la complexité du problème traité

Faiblesses de SVM

— Difficulté à identifier les bonnes valeurs des paramètres;
— Difficulté à traiter les grandes bases;
— Problème lorsque les classes sont bruitées (multiplication des points supports);
— Pas de modèle explicite pour les noyaux non linéaires (utilisation des points supports);
— Difficulté d'interprétations.

Apprentissage avec SVM

Pour l'apprentissage avec le SVM nous prenons aussi 70 individus et 30 pour les tests. La figure ci-dessous présente les résultats obtenus:

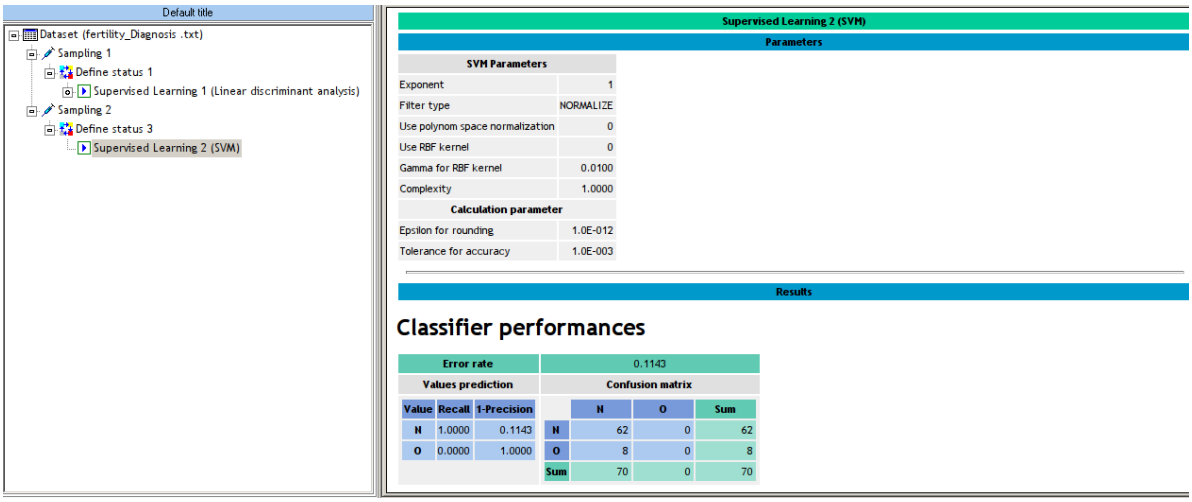


Figure 4.4 – Apprentissage avec SVM

Test avec SVM

Avec l'application de l'échantillonnage de test, nous obtenons les résultats suivants: une précision de 0.871 pour les individus dont les spermantozoides n'ont pas été modifiés. La valeur de précision 0.871 qui tend vers la valeur 1 signifie que nous otenons des résultats assez proches de la réalité pour les individus dont les spermatozoides n'ont pas été modifiés; Une précision de 0 pour les individus dont les spermantozoides ont été modifiés. Avec cet échantillonnage nous avons un taux d'erreur de 0.1290. La figure ci-dessous répresente les résultats de l'échantillon de test.

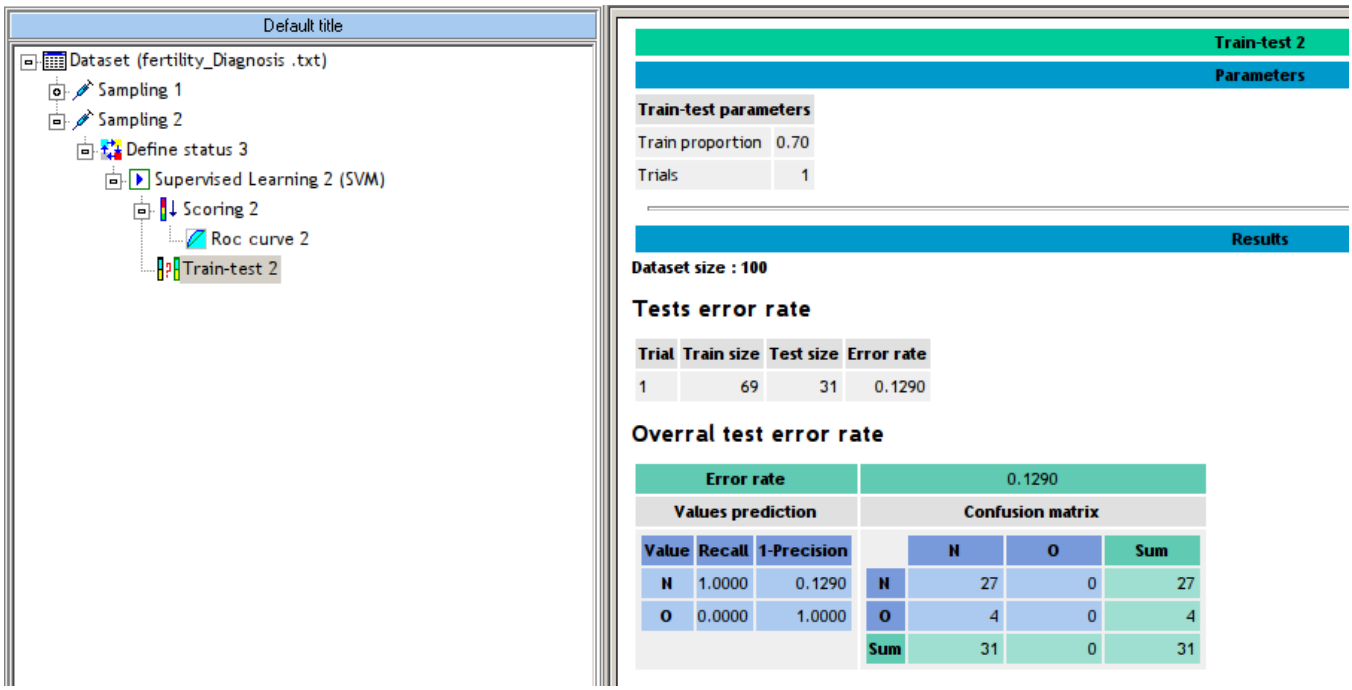


Figure 4.5 – Test avec SVM

2 Comparaison des resultats de LDA et de SVM

2.1 Utilisation de la fonction Scoring

L'objectif du scoring est d'étendre la comparaison entre les méthodes utilisées. Ainsi nous allons ici attribuer un score à chaque variable de toute la base de données qu'elle soit scindée en données d'apprentissage ou de test. Cela se fait comme suit:

— O : pour les individus dont les spermatozoides ont été modifiés (valeur positive)

— N : pour les individus dont les spermatozoides n'ont pas été modifiés (valeur négative)

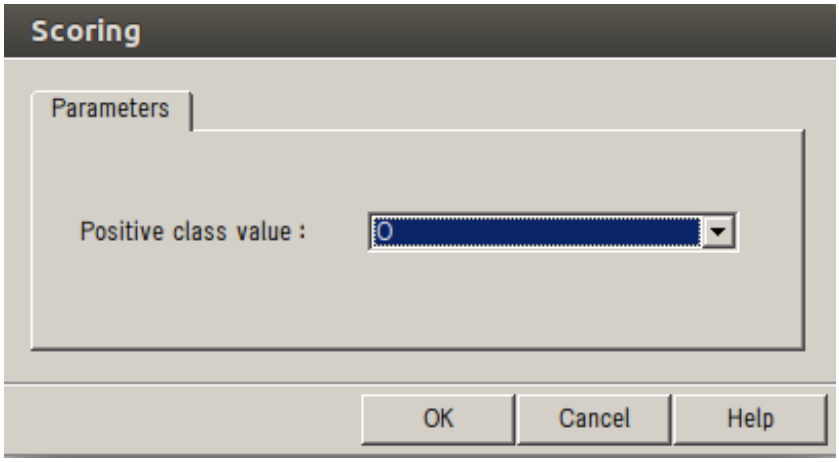


Figure 4.6 – Caption

Les résultats sont regroupés dans un tableau récapitulatif. Nous disposons :

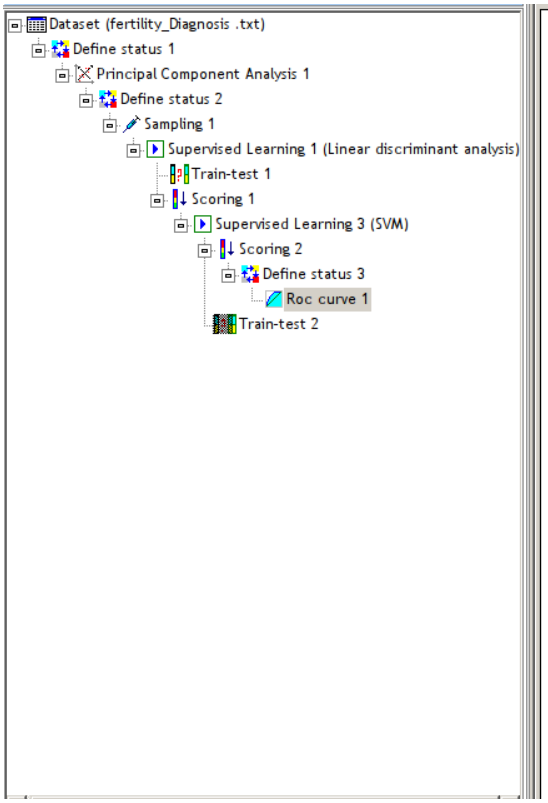
— De l'indicateur AUC calculée de manière très simplifiée à l'aide de la méthode des trapèzes.

— Pour chaque taille de cible (Vrais positifs + Faux positifs), nous disposons des taux de faux positifs et taux de vrais positifs.

— Attention, dans la majorité des cas, les scores ne sont pas comparables d'une méthode à l'autre mqis permet de mieux voir les valeurs des scores de ces algorithmes.

— Score 1 : pour LDA

— Score 2 : pour SVM



ROC Curve

Sample size : 70
Positive examples : 8
Negative examples : 62

Score Attribute	Score_1			Score_2		
AUC	0.8992			0.8427		
Target size (%)	Score	FP-Rate	TP-Rate	Score	FP-Rate	TP-Rate
0	0.7823	0.0000	0.0000	0.2699	0.0000	0.0000
5	0.6297	0.0161	0.2500	0.2697	0.0161	0.2500
10	0.2881	0.0484	0.5000	0.2691	0.0484	0.5000
15	0.2539	0.0968	0.5000	0.2691	0.0806	0.6250
20	0.1685	0.1290	0.7500	0.2689	0.1290	0.7500
25	0.1372	0.1774	0.7500	0.2689	0.1774	0.7500
30	0.1154	0.2419	0.7500	0.2688	0.2419	0.7500
35	0.1027	0.2581	1.0000	0.2687	0.2742	0.8750
40	0.0814	0.3226	1.0000	0.2686	0.3387	0.8750
45	0.0735	0.3710	1.0000	0.2685	0.3871	0.8750
50	0.0590	0.4355	1.0000	0.2684	0.4516	0.8750
55	0.0470	0.4839	1.0000	0.2683	0.5000	0.8750
60	0.0343	0.5484	1.0000	0.2682	0.5645	0.8750
65	0.0272	0.5968	1.0000	0.2681	0.6129	0.8750
70	0.0181	0.6613	1.0000	0.2678	0.6774	0.8750
75	0.0152	0.7097	1.0000	0.2677	0.7258	0.8750
80	0.0135	0.7742	1.0000	0.2674	0.7742	1.0000
85	0.0100	0.8226	1.0000	0.2671	0.8226	1.0000
90	0.0065	0.8871	1.0000	0.2670	0.8871	1.0000
95	0.0046	0.9355	1.0000	0.2668	0.9355	1.0000
100	0.0010	1.0000	1.0000	0.2661	1.0000	1.0000

Figure 4.7 – répartition des données d'apprentissage et de test

2.2 ROC CURVE

Cette courbe compare les deux algorithmes.

— Score_1: Couleur verte avec une valeur 0.899 qui correspond à la valeur de performance de l'algorithme LDA. Il donne également un taux de précision de 0,909.

— Score_2: Couleur jaune avec une valeur 0.843 qui correspond à la valeur de performance de l'algorithme SVM avec un taux de précision de 0.871.

A partir de ces résultats nous pouvons conclure que l'algorithme LDA donne des resultats mieux appréciables par rapport à l'algorithme l'algorithme SVM.

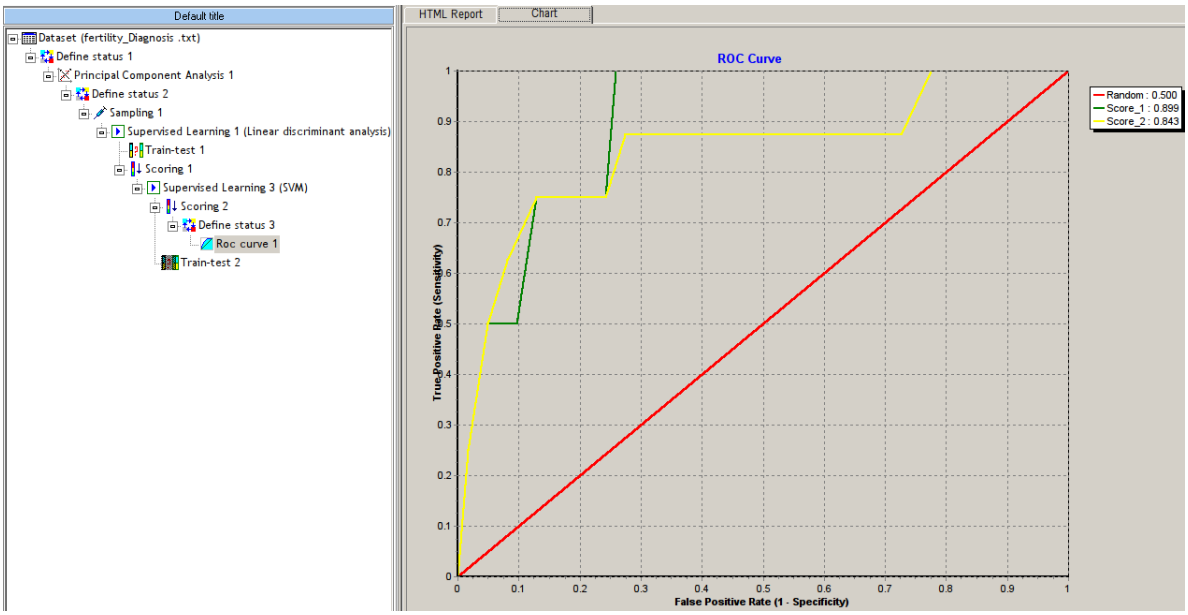


Figure 4.8

3 Conclusion

Le data mining est très important pour les entreprises, dans la mesure où elle favorise la prédiction. En effet, dans ces travaux pratiques nous avons choisi comme base de donnée fertility qui a pour but de prédire la modification des spermatozoides d'un individus en fonction des attributs d'entrés. L'analyse exploratoire de notre base de données nous a permis de mieux comprendre les attributs. De plus nous avons utilisé deux algorithmes d'apprentissage le SVM et le LDA sur notre base de données, ce qui a permis de comparer leur résultats. En effet, si les données sont peu nombreuses l'apprentissage est peu fiable et ne donne pas de résultats satisfaisants. Ainsi dans notre cas nous avons constaté que l'algorthime LDA était mieux adapter car il donne un taux d'erreurs inférieur à celui du SVM. De ce fait si les données sont peu nombreuses le LDA est mieux adapté pour l'apprentissage. A travers, ces TPs nous avons mieux compris dans la pratique les algorithmes vus pendant la théorie.Cependant nous pensons qu'avec les suggestions et les remarques ce travail pourra être amélioré.

Bibliography

[1] David Gil, Jose Luis Girela, Joaquin De Juan, M Jose Gomez-Torres, and Magnus Johnsson. Predicting seminal quality with artificial intelligence methods. *Expert Systems with Applications*, 39(16):12564–12573, 2012.

[2] Ludovic Lebart, Marie Piron, and Alain Morineau. *Statistique exploratoire multidimensionnelle: visualisation et inférences en fouilles de données*. 2006.