



FOUILLE DE DONNÉES  
ANALYSE DU JEU DE DONNÉES

Données utilisées :  
«**Contraceptive Method Choice Data Set** »

---

Etudiants : Peterson BERTELOT — Fabienne JANVIER  
PROMOTION 24 SIM  
Année Académique: 2019-2021  
Professeur: Mme. Nguyen Thi Minh Huyen

<b>Introduction</b>	<b>2</b>
<b>1- Choisir un modèle de jeu de données</b>	<b>2</b>
<b>2- Analyse descriptive du modèle</b>	<b>2</b>
<b>3. Analyse exploratoire des attributs</b>	<b>3</b>
a. Wife's age (Type : Continue)	3
b. Wife's education (Catégorie : Discrete)	3
c. Husband's education (Category : Discrete)	4
d. Number of children ever born (Category : Continue)	5
e. Wife's religion (Category : Discrete)	6
f. Wife's now working (Category: Discrete)	6
g. Husband's occupation	7
h. Standard-of-living index (Category : Discrete)	8
i. Media exposure (Category : Discrete)	9
j. Contraceptive method used (Category : Discrete)	10
<b>4. Analyse de lien entre chaque paire d'attributs</b>	<b>11</b>
a. Wife's religion avec Contraceptive method used	11
b. Wife's education avec Contraceptive method used	12
c. Number of children ever born avec Contraceptive method used	12
d. Standard-of-living index avec Contraceptive method used	13
e. Wife's now working et Contraceptive method used	14
f. Husband's occupation avec Contraceptive method used	14
<b>Analyse Factorielle</b>	<b>15</b>
Analyse Factorielle des Données Mixtes	15
B. Tableau des coordonnées	16
C. Tableau des corrélations	16
D. Tableau des moyennes conditionnelles	17
E. Tableau des vecteurs propres	17
F. Représentation Graphique	18
G. Cercle des corrélations	19
H. Moyennes conditionnelles	20
<b>6. Classification automatique</b>	<b>21</b>
6.1. L'algorithme de CAH	21
6.2. L'algorithme de K-Means	24
<b>Machine à vecteurs de support</b>	<b>28</b>
Introduction	28

2.2- Principe de la technique SVM	28
2.3- Support vecteur	29
2.4- Hyperplan	29
2.5- Marge	30
2.6- Linéarité et non-linéarité des données	30
2.6.1- Linéarité des données	30
2.6.2- Non-Linéarité des données	31
2.7- Fonctions noyaux	31
2.7.1 Type des noyaux	32
<b>3. Préparation de données</b>	<b>32</b>
3.1- Pré-traitement des données	32
3.1.1- Visualisation de donnée	33
<b>4. Construction et Évaluation du modèle</b>	<b>34</b>
4.1 Sélection des paramètres	34
4.2 - Validation du modèle	35
4.3 Matrice de confusion	35
4.3.1 Calcul du raciaux	36
4.3.2 Visualisation du resultat	36
4.4 Regression Logistique	36
4.4.1 Evaluation du modèle	37
4.4.2 Visualisation du modèle	37
4.5 Visualisation Croiser	38
<b>5. Interprétation de résultat</b>	<b>39</b>
<b>6. Conclusion</b>	<b>40</b>

## Introduction

Ce travail est réalisé dans le cadre du cours Fouille de donnée à l'IFI il a pour but de faire la présentation, l'analyse Exploratoire des variables et paires de variables du jeu de données choisi.

L'objectif est de pronostiquer la décision de moyens contraceptifs auprès de femmes indonésiennes et à quelle tranche d'âge une femme indonésienne adopte une méthode de contraception.

Pour y arriver nous nous proposons de découper ce travail en quatre petites parties à savoir :

1. Choisir un modèle de jeu de données
2. Faire une analyse descriptive de ce modèle
3. Faire analyse exploratoire des attributs le constituant
4. Faire une analyse de lien entre chaque paire d'attributs

## 1- Choisir un modèle de jeu de données

Nous avons décidé un jeu traitant un ensemble de données renfermant des renseignements sur les résultats des décisions de procédés de contraceptives chez 1472 filles Indonésiennes. Cet ensemble d'informations est un sous-ensemble de cette enquête nationale de 1987 sur la prévalence de la contraception en Indonésie. Les prélèvements sont des femmes mariées qui n'étaient pas enceintes ou qui ne savaient pas si elles l'étaient lors de l'entretien. La difficulté a pour objectif d'annoncer la décision actuelle de méthodes de contraception (pas d'utilisation, procédés sur des échéances plus longues ou moyens à court terme) d'une femme en fonction de ses spécifications techniques démographiques et socio-économiques.

**Source :** <https://archive.ics.uci.edu/ml/datasets/Contraceptive+Method+Choice>

## 2- Analyse descriptive du modèle

Nous avons mis à contribution le logiciel Tanagra pour l'analyse de notre jeu des données. TANAGRA est un logiciel gratuit d'exploration de données (Data Mining) destiné à l'enseignement et à la recherche créé en 2003. Il implémente une série de méthodes de fouille de données issues du domaine de la statistique exploratoire, de l'analyse de données, de l'apprentissage automatique et des bases de données.

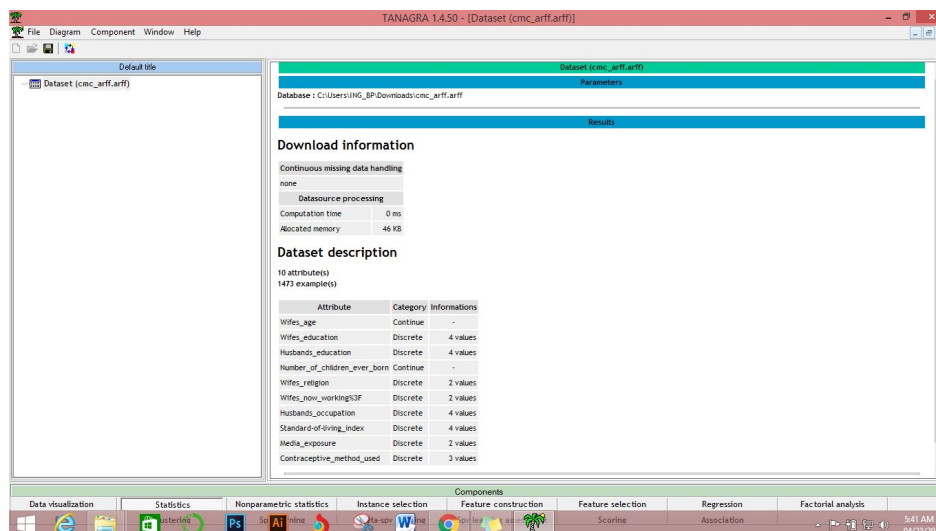


Figure 1. Information générale du modèle de jeu

La figure 1 nous donne une description générale sur le modèle choisie. Celui-ci est composé de 1473 exemples dont toutes les informations sont présentes. Dans la section suivante nous allons voir en détail les 10 attributs de ce dernier dont 2 sont continus et 8 discrets.

### 3. Analyse exploratoire des attributs

Dans cette section nous allons nous intéresser sur les attributs constituant notre modèle de jeu. Pour cela nous intéresserons à sa catégorie, nombre de valeurs possible et son utilité.

#### a. Wife's age (Type : Continue)

Cet attribut définit l'âge du sujet concerné. L'âge des filles interviewer oscille entre 16 à 49 ans, cela trouve sa cause par l'idée que, au dessus 49 ans la femme n'a plus réellement de contraceptif. Cet attribut est réellement essentiel pour la prédiction parce qu'elle offre déjà une idée sur la tranche d'âge des femmes exploitant les contraceptifs. En le comparant avec différents caractéristiques que nous allons expliquer après on a la possibilité d'avoir des bien résultats.

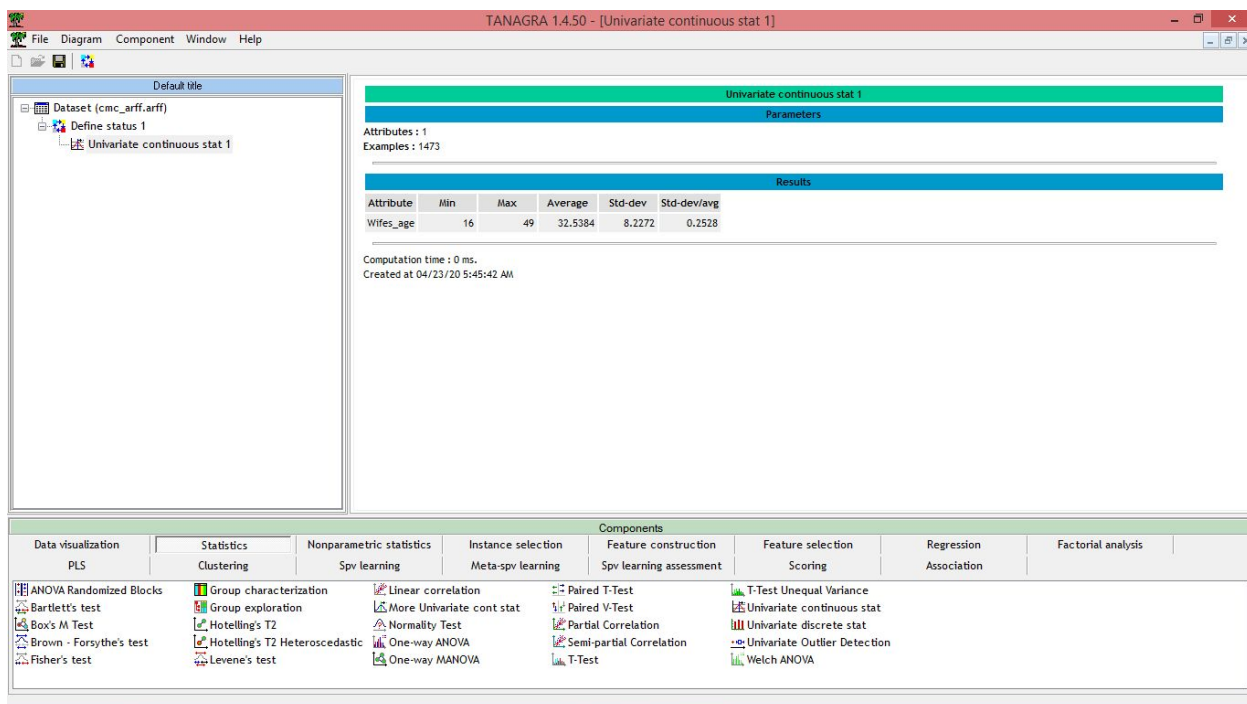


Figure 2. Attribut âge de la femme (Wife's age)

#### b. Wife's education (Catégorie : Discrete)

Cet attribut détermine le niveau d'éducation du sujet, soit son niveau d'étude. Il a 4 valeurs possible soit :

- Low (faible) : pour les femmes qui ont un niveau d'éducation inférieur
- Middle (moyen) : pour celles qui ont un niveau moyen d'éducation

- Upper middle (supérieur à la moyenne) : soit celles qui ont encore évoluée un peu plus
- High (supérieur) : soit celles qui ont un niveau supérieur ou celles qui sont suffisamment éduquées.

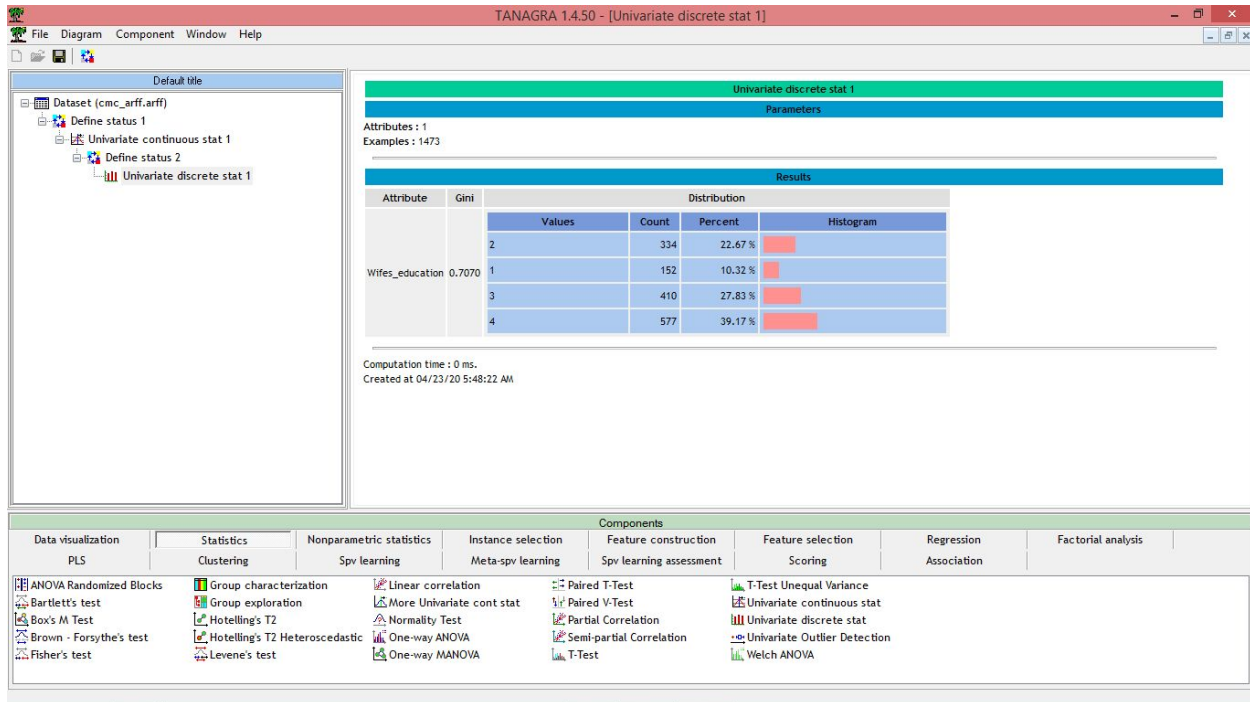


Figure 3. Attribut éducation de la femme (wife's education)

La figure 3, nous montre que parmi les qui ont été interviewé, la majorité a une bonne éducation, soit plus des femmes ont suffisamment étudié.

### c. Husband's education (Category : Discrete)

Comme l'attribut précédent, celui-ci nous informe sur l'éducation du mari. Il a aussi 4 valeurs possibles avec les mêmes significations.

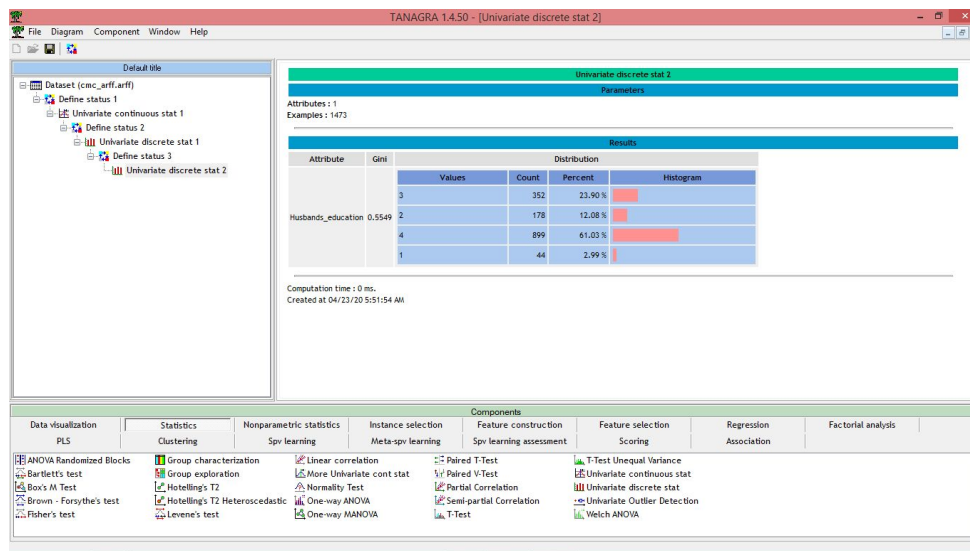


Figure 4. L'attribut éducation du mari (Husband's éducation)

Comparativement à la figure 3, on remarque clairement que dans la plupart de couples ce sont les maris qui ont une éducation très élevés sur le plan scolaire. Chose qui normale dans la majorité des sociétés.

#### d. Number of children ever born (Category : Continue)

Vu que cette enquête se faite auprès des femmes mariés ; l'attribut nombre d'enfants informe sur les enfants que possède la femme. Cette valeur varie de 0 à 16. Pour dire qu'il y a de femmes qui n'ont pas d'enfants et au maximum celle qui en a ; a 16 enfants.

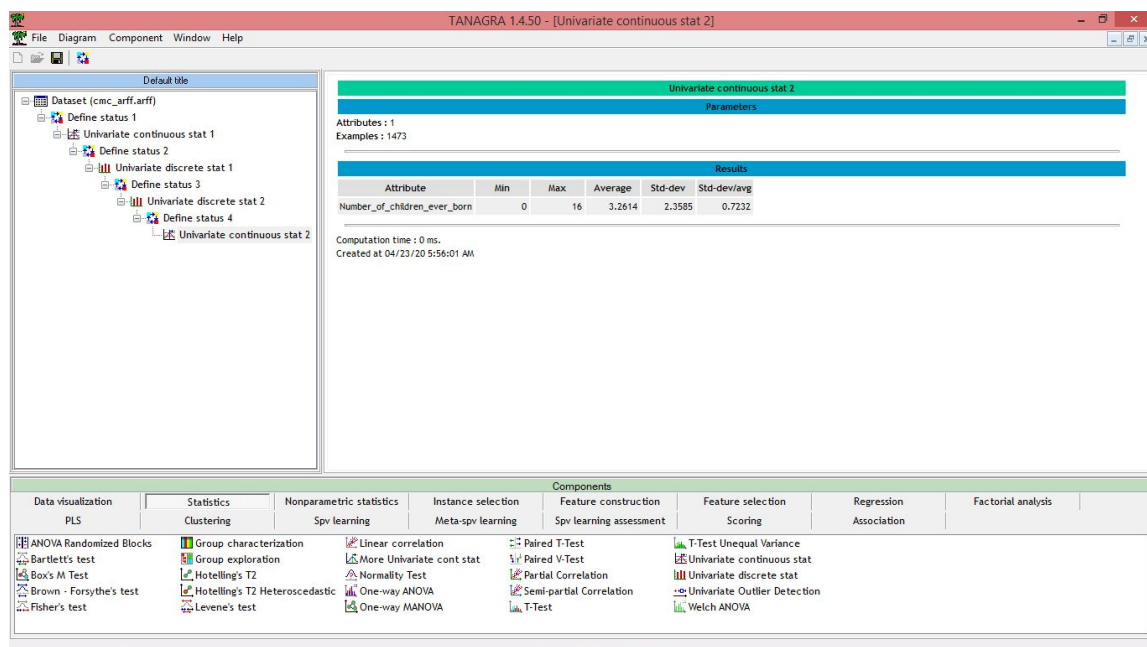


Figure 5. Attribut nombre d'enfant

#### e. Wife's religion (Category : Discrete)

Vu que la majorité la population indonésienne est musulmane; l'attribut religion de la femme prend par conséquent 2 valeurs seulement; soit Islam ou No-islam. Ainsi pour savoir la croyance du sujet concerné.

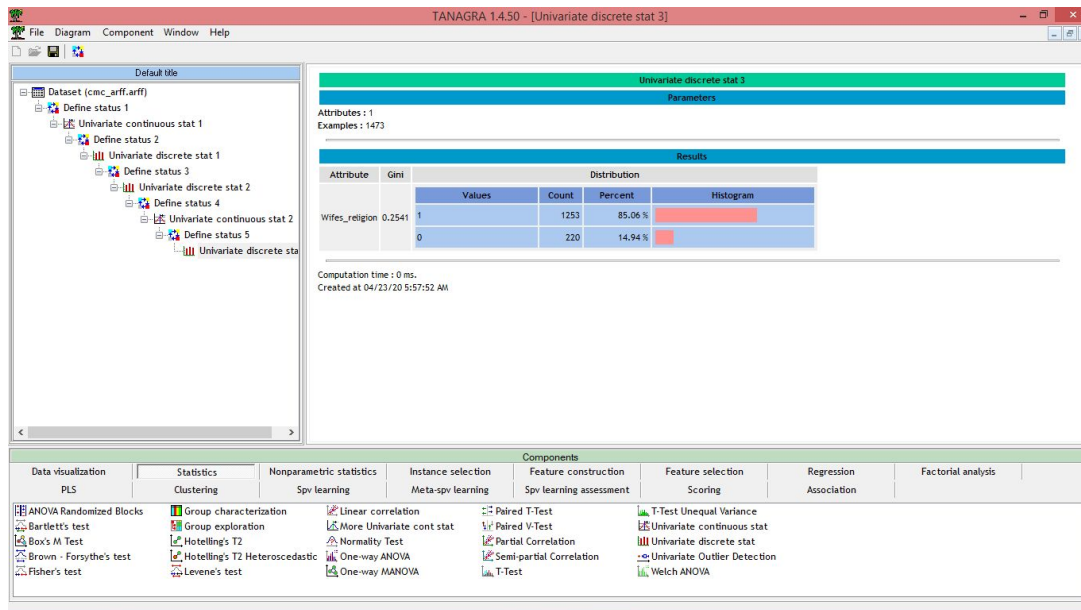


Figure 6. Attribut religion de la femme

#### f. Wife's now working (Category: Discrete)

Cet attribut fait savoir si au moment de l'entretien la femme avait une occupation. Cette a 2 valeurs possibles ; soit Yes au cas où la femme a un travail sinon c'est No.



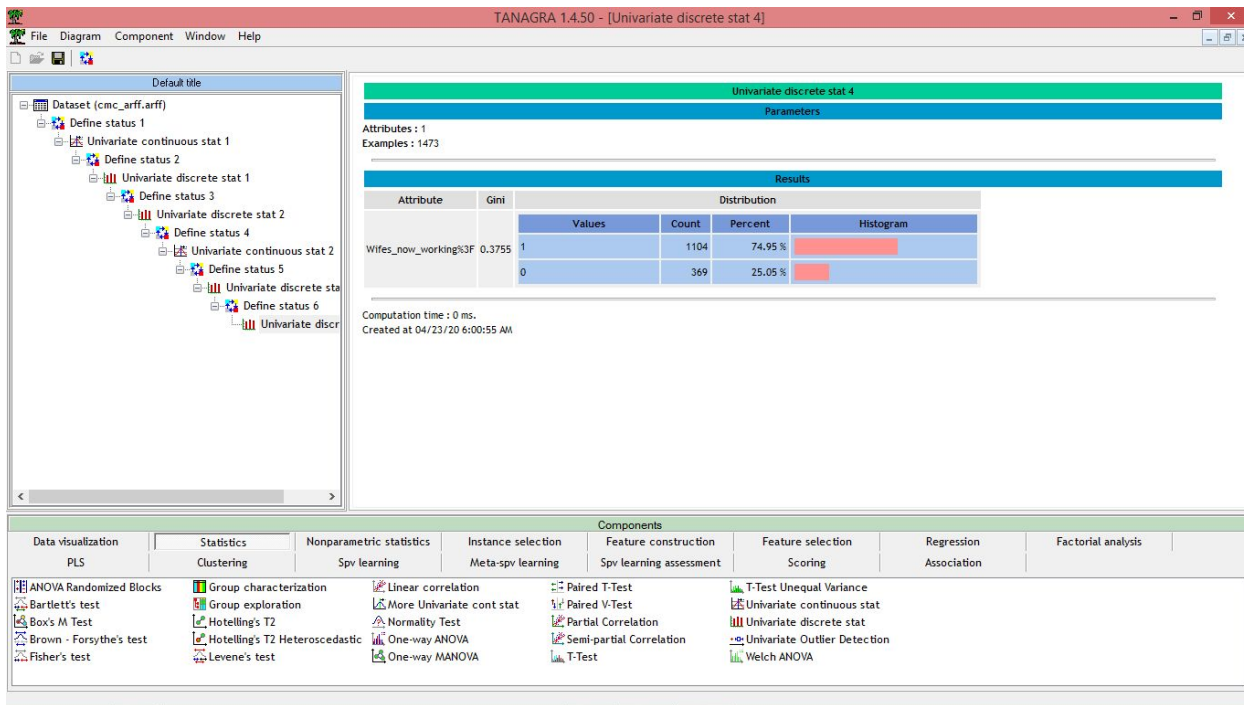


Figure 7. Attribut occupation de la femme

## g. Husband's occupation

Cet attribut contient l'occupation de du marie de la femme interviewée. Et il a 4 valeurs possibles.

1. Soit la valeur low (pour une occupation qui n'as pas assez des revenus),
2. Middle (au cas où le mari a une occupation qui fait gagner assez bien),
3. Uper middle (pour une occupation qui rapporte un peu plus encore par rapport à une occupation middle),
4. High (pour une occupation qui a un revenu très élevé).

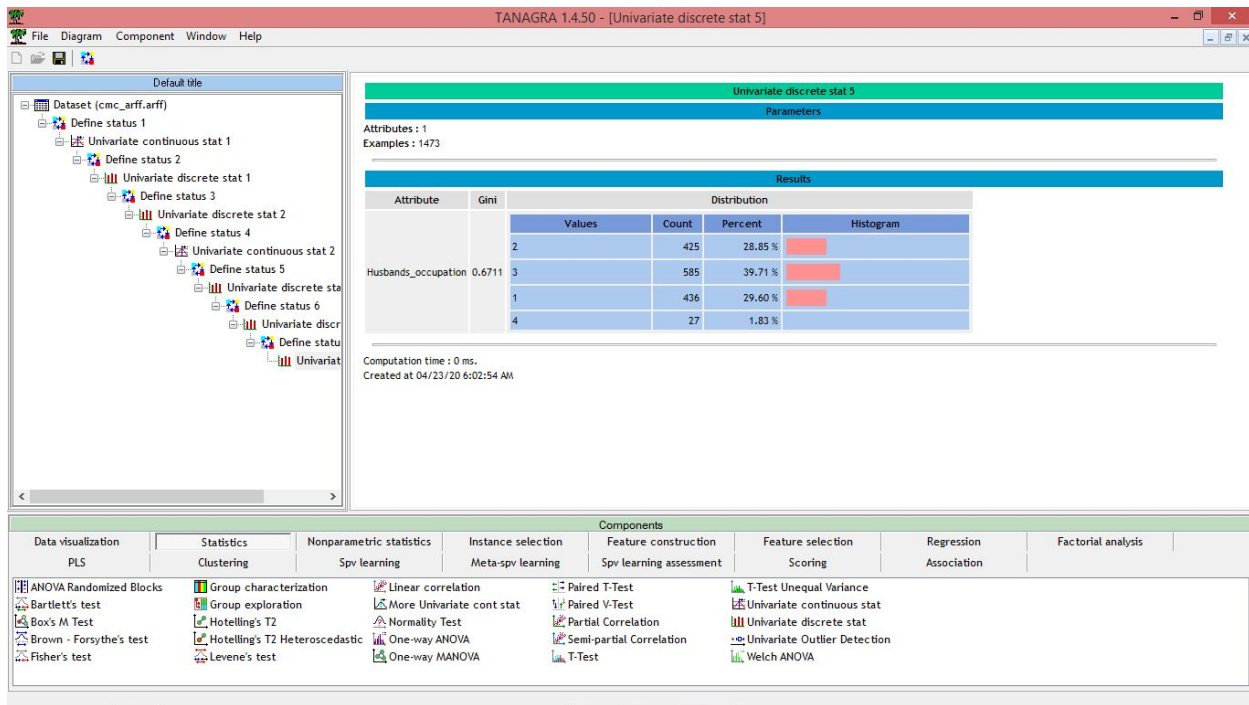


Figure 8. Attribut occupation du mari

#### h. Standard-of-living index (Category : Discrete)

Cet attribut contient l'information sur le niveau de vie de la femme ou soit du couple. Et a 4 valeurs possibles comme l'attribut précédent et aussi avec les mêmes significations.

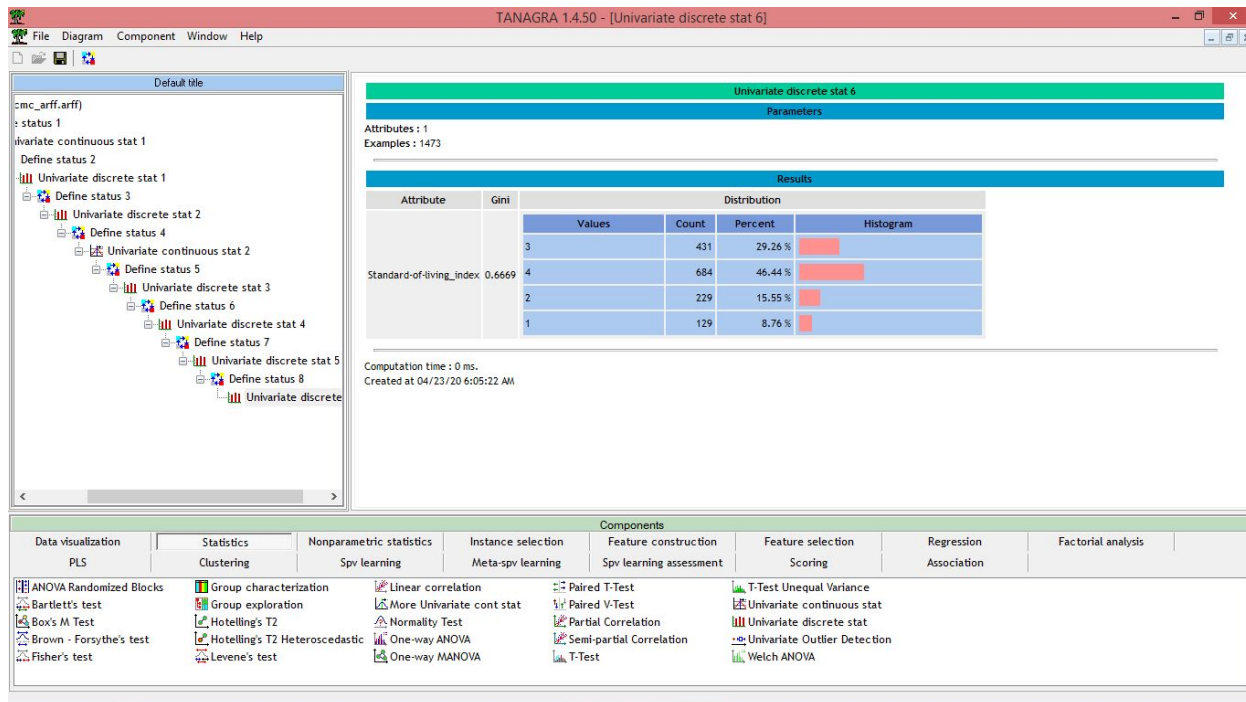


Figure 9. Attribut indice de niveau de vie

#### i. Media exposure (Category : Discrete)

Cet attribut nous informe sur l'avis du sujet concerné pour la vulgarisation des contraceptives au pr ts des m dias. Il a comme 2 valeurs possibles, soit Good pour les femmes qui sont d'accord et No-good pour celles qui ne sont pas d'accord.

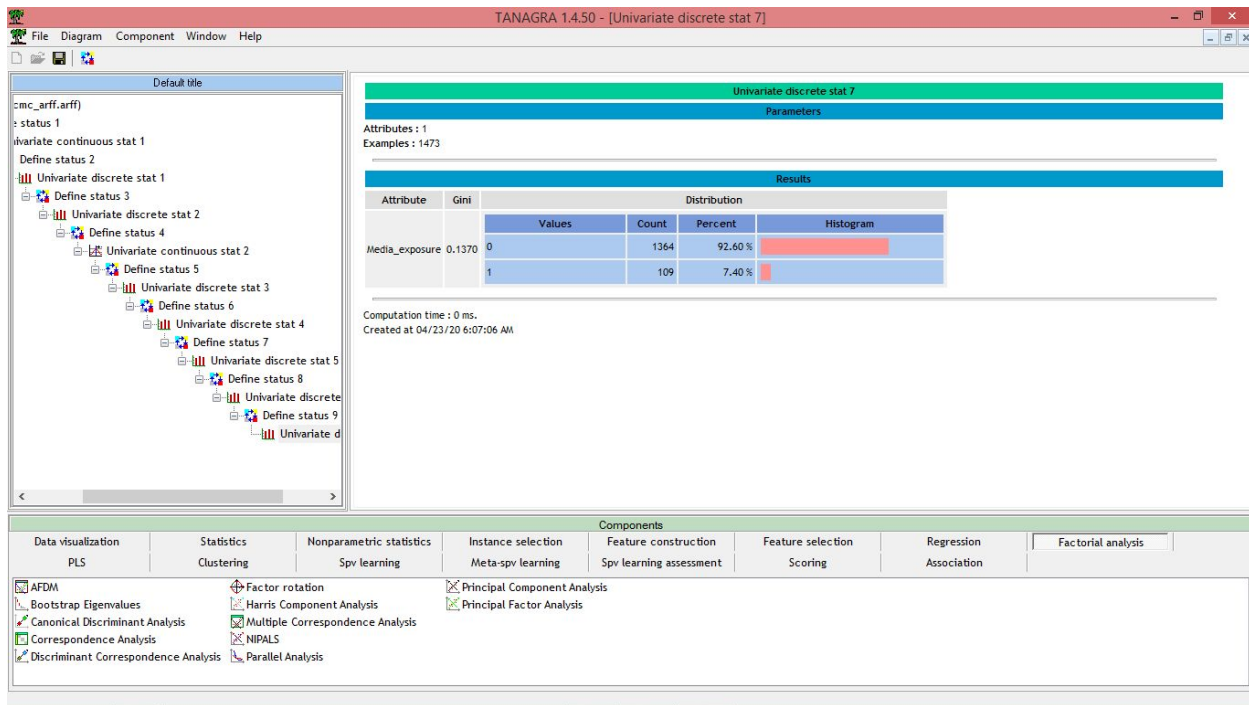


Figure 10. Exposition au media

#### j. Contraceptive method used (Category : Discrete)

Ce dernier attribut nous fait savoir sur la méthode de contraception utilisé par le sujet concerné. Il a 3 valeurs possibles soit, No-use (pour les femmes qui n'utilisent pas les contraceptives), Long-term (pour celles qui utilisent les contraceptives à long terme) et Short-term (pour celles qui les utilisent pas souvent soit à court terme).

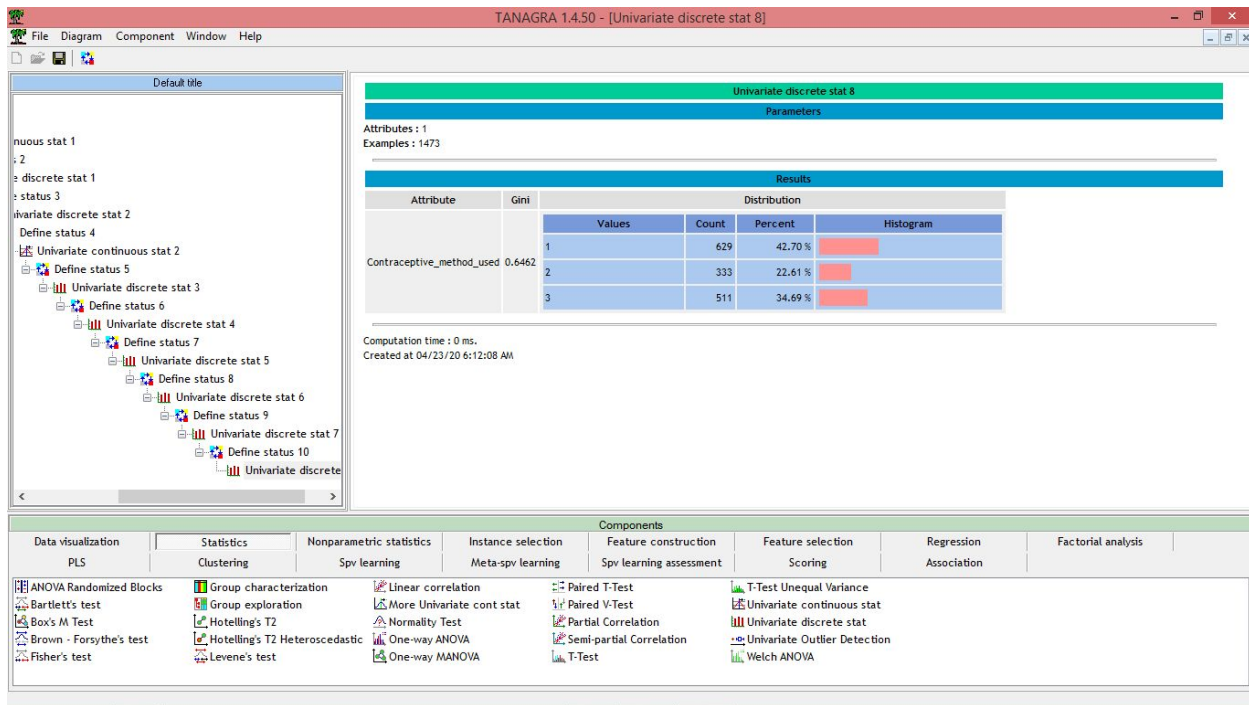


Figure 11. Attribut Méthode de contraception utilisée

#### 4. Analyse de lien entre chaque paire d'attributs

Dans ce dernier volet de notre travail d'examination d'un jeu des données ; nous allons tenter de faire une étude de lien entre chaque paire d'attributs que nous venons de voir. On fera toujours un lien entre un attribut ordinaire et l'attribut Contraceptive méthode utilisée.

##### a. Wife's religion avec Contraceptive method used

Pour ce premier lien d'analyse, nous nous intéresserons sur la méthode de contraception utilisé par rapport à la religion de la femme. Et avec le résultat que nous obtenons à la figure suivante, nous pouvons dire que les femmes musulmanes n'utilisent pas trop les contraceptifs avec une proportion de 88.1% de No-use contre 11.9% pour les femmes Non musulmanes.

Group characterization 4

Parameters

Normalization : 0

Results

Description of " Wife's religion"

Wife's religion-Islam

Examples

[ 85.1 %] 1253

Att - Desc

Test value

Group

Overall

Continuous attributes : Mean (StdDev)

Discrete attributes : [Recall] Accuracy

Contraceptive method used=No-use

2.80

[ 88.1 %]

44.2 %

42.7 %

Contraceptive method used=Short-term

1.12

[ 86.5 %]

35.3 %

34.7 %

Contraceptive method used=Long-term

-4.59

[ 77.2 %]

20.5 %

22.6 %

Wife's religion-Non-Islam

Examples

[ 14.9 %] 220

Att - Desc

Test value

Group

Overall

Continuous attributes : Mean (StdDev)

Discrete attributes : [Recall] Accuracy

Contraceptive method used=Long-term

4.59

[ 22.8 %]

34.5 %

22.6 %

Contraceptive method used=Short-term

-1.12

[ 13.5 %]

31.4 %

34.7 %

Contraceptive method used=No-use

-2.80

[ 11.9 %]

34.1 %

42.7 %

Computation time : 0 ms.

Created at 3/19/2019 5:57:40 PM

Figure 12. Lien entre la religion de la femme et la méthode contraceptive utilisée

Group characterization 5

Parameters

Normalization : 0

Results

Description of " Contraceptive method used"

Contraceptive method used=No-use

Examples		[ 42.7 %] 629	
Att - Desc	Test value	Group	Overall
Continuous attributes : Mean (StdDev)			
Discrete attributes : [Recall] Accuracy			
Wife's education=low	6.59	[ 67.8 %] 16.4 %	10.3 %
Wife's education=middle	4.20	[ 52.7 %] 28.0 %	22.7 %
Wife's education=uper middle	-0.01	[ 42.7 %] 27.8 %	27.8 %
Wife's education=high	-7.70	[ 30.3 %] 27.8 %	39.2 %

Contraceptive method used=Long-term

Examples		[ 22.6 %] 333	
Att - Desc	Test value	Group	Overall
Continuous attributes : Mean (StdDev)			
Discrete attributes : [Recall] Accuracy			
Wife's education=high	9.77	[ 35.9 %] 62.2 %	39.2 %
Wife's education=uper middle	-1.76	[ 19.5 %] 24.0 %	27.8 %
Wife's education=low	-5.19	[ 5.9 %] 2.7 %	10.3 %
Wife's education=middle	-5.73	[ 11.1 %] 11.1 %	22.7 %

Contraceptive method used=Short-term

Examples		[ 34.7 %] 511	
Att - Desc	Test value	Group	Overall
Continuous attributes : Mean (StdDev)			
Discrete attributes : [Recall] Accuracy			
Wife's education=uper middle	1.56	[ 37.8 %] 30.3 %	27.8 %
Wife's education=middle	0.67	[ 36.2 %] 23.7 %	22.7 %
Wife's education=high	-0.58	[ 33.8 %] 38.2 %	39.2 %
Wife's education=low	-2.29	[ 26.3 %] 7.8 %	10.3 %

Computation time : 0 ms.

Created at 3/19/2019 6:04:34 PM

#### b. Wife's education avec Contraceptive method used

Figure 13. Lien entre le niveau de l'éducation de la femme avec la méthode contraceptive utilisée  
La figure 13, nous montre un lien qui est entre le niveau d'éducation de la femme avec la méthode contraceptive utilisée. Nous pouvons dire que la plupart des femmes qui utilisent les contraceptives sont celles qui une bonne éducation soit celles qui ont suffisamment étudié.

#### c. Number of children ever born avec Contraceptive method used

La figure suivante nous fait voir le lien qui est entre l'attribut nombre d'enfants qu'a une femme avec la méthode contraceptive utilisée. Et nous remarquons que les femmes qui ont plus d'enfants ce sont elles qui utilisent les contraceptive à long terme puis celles qui en ont moins. Celles qui peut d'enfant soit une proportion de 2.93% n'utilise pas du tout les contraceptive.



Group characterization 6

Parameters

Normalization : 0

Results

Description of "Contraceptive method used"

Contraceptive method used-No-use

Examples		[ 42.7 %] 629	
Att - Desc	Test value	Group	Overall
Continuous attributes : Mean (StdDev)			
Number of children ever born	-4.59	2.93 (2.66)	3.26 (2.36)
Discrete attributes : [Recall] Accuracy			

Contraceptive method used-Long-term

Examples		[ 22.6 %] 333	
Att - Desc	Test value	Group	Overall
Continuous attributes : Mean (StdDev)			
Number of children ever born	4.20	3.74 (2.10)	3.26 (2.36)
Discrete attributes : [Recall] Accuracy			

Contraceptive method used-Short-term

Examples		[ 34.7 %] 511	
Att - Desc	Test value	Group	Overall
Continuous attributes : Mean (StdDev)			
Number of children ever born	1.08	3.35 (2.05)	3.26 (2.36)
Discrete attributes : [Recall] Accuracy			

Computation time : 0 ms.

Created at 3/19/2019 8:07:33 PM

Figure 14. Lien entre le nombre d'enfants de la femme avec la méthode contraceptive utilisée

#### d. Standard-of-living index avec Contraceptive method used

Comme nous l'avons dit dans section précédente ; l'indice de vie nous informe sur le niveau économique de la femme ou soit du couple, Ainsi cette figure nous montre que quel que soit le niveau de vie de la femme, la plus part n'utilise pas le contraceptive en Indonésie.

Group characterization 7											
Parameters											
Normalization : 0											
Results											
Description of "Contraceptive method used"											
Contraceptive method used-No-use				Contraceptive method used-Long-term				Contraceptive method used-Short-term			
Examples		[ 42.7 %] 629		Examples		[ 22.6 %] 333		Examples		[ 34.7 %] 511	
Att - Desc	Test value	Group	Overall	Att - Desc	Test value	Group	Overall	Att - Desc	Test value	Group	Overall
Continuous attributes : Mean (StdDev)				Continuous attributes : Mean (StdDev)				Continuous attributes : Mean (StdDev)			
Discrete attributes : [Recall] Accuracy				Discrete attributes : [Recall] Accuracy				Discrete attributes : [Recall] Accuracy			
Standard-of-living index=low	4.64	[ 62.0 %] 12.7 %	8.8 %	Standard-of-living index=high	6.16	[ 29.8 %] 61.3 %	46.4 %	Standard-of-living index=uper middle	0.90	[ 36.4 %] 30.7 %	29.3 %
Standard-of-living index=middle	2.79	[ 51.1 %] 18.6 %	15.5 %	Standard-of-living index=uper middle	-1.02	[ 20.9 %] 27.0 %	29.3 %	Standard-of-living index=middle	0.39	[ 35.8 %] 16.0 %	15.5 %
Standard-of-living index=uper middle	-0.01	[ 42.7 %] 29.3 %	29.3 %	Standard-of-living index=middle	-3.74	[ 13.1 %] 9.0 %	15.5 %	Standard-of-living index=high	-0.58	[ 33.9 %] 45.4 %	46.4 %
Standard-of-living index=high	-4.65	[ 36.3 %] 39.4 %	46.4 %	Standard-of-living index=low	-4.44	[ 7.0 %] 2.7 %	8.8 %	Standard-of-living index=low	-0.92	[ 31.0 %] 7.8 %	8.8 %

Computation time : 0 ms.  
Created at 3/19/2019 8:18:15 PM

Figure 15. Lien entre le l'indice de vie avec la méthode contraceptive utilisé

Comme nous l'avons dit dans section précédente ; l'indice de vie nous informe sur le niveau économique de la femme ou soit du couple, Ainsi la figure 15 nous montre que quel que soit le niveau de vie de la femme, la plus part n'utilise pas le contraceptive en Indonésie.

### e. Wife's now working et Contraceptive method used

Pour les femmes qui ont été soumises à l'entretien, nous trouvons que un plus nombre d'entre elles qui utilisent les contraceptives ce sont celles qui ont une occupation soit elles ont un travail. Avec une proportion de 24.1% pour le long terme et 29.8% pour le court terme au total nous avons 53.9% de femmes qui utilisent les contraceptive ont une occupation.

Group characterization 3

Parameters

Normalization : 0

Results

Description of " Contraceptive method used"

Contraceptive method used-No-use

Examples		[ 42.7 %] 629	
Att - Desc	Test value	Group	Overall
Continuous attributes : Mean (StdDev)			
Discrete attributes : [Recall] Accuracy			
Wife's now working?=Yes	1.51	[ 46.1 %]	27.0 %
Wife's now working?=No	-1.51	[ 41.6 %]	73.0 %

Contraceptive method used-Long-term

Examples		[ 22.6 %] 333	
Att - Desc	Test value	Group	Overall
Continuous attributes : Mean (StdDev)			
Discrete attributes : [Recall] Accuracy			
Wife's now working?=Yes	0.80	[ 24.1 %]	26.7 %
Wife's now working?=No	-0.80	[ 22.1 %]	73.3 %

Contraceptive method used-Short-term

Examples		[ 34.7 %] 511	
Att - Desc	Test value	Group	Overall
Continuous attributes : Mean (StdDev)			
Discrete attributes : [Recall] Accuracy			
Wife's now working?=No	2.27	[ 36.3 %]	78.5 %
Wife's now working?=Yes	-2.27	[ 29.8 %]	21.5 %

Computation time : 0 ms.

Created at 3/19/2019 5:51:47 PM

Figure 16. Lien entre l'attribut Wife's now working et Contraceptive method used

### f. Husband's occupation avec Contraceptive method used

Ici nous faisons un lien entre l'attribut Husbans's occupation qui désigne le travail du mari celui-ci peut être low, middle, upper middle ou high avec la méthode de contraception utilisé par la femme. Voyons comment Les maris qui ont une occupation de type low (soit avec moins de revenus), leurs femmes n'utilisent pas trop les contraceptive.

Group characterization 8

Parameters

Normalization : 0

Results

Description of " Contraceptive method used"

Contraceptive method used-No-use

Examples

[ 42.7 %] 629

Att - Desc

Test value

Group

Overall

Continuous attributes : Mean (StdDev)

Discrete attributes : [Recall] Accuracy

Husband's occupation=middle

2.15

[ 47.1 %]

31.8

28.9 %

Husband's occupation=upper middle

0.88

[ 44.1 %]

41.0

39.7 %

Husband's occupation=high

0.58

[ 48.1 %]

2.1

1.8 %

Husband's occupation=low

-3.25

[ 36.2 %]

25.1

29.6 %

Contraceptive method used=Long-term

Examples

[ 22.6 %] 333

Att - Desc

Test value

Group

Overall

Continuous attributes : Mean (StdDev)

Discrete attributes : [Recall] Accuracy

Husband's occupation=low

7.83

[ 35.8 %]

46.8

29.6 %

Husband's occupation=high

-0.51

[ 18.5 %]

1.5

1.8 %

Husband's occupation=middle

-2.35

[ 18.6 %]

23.7

28.9 %

Husband's occupation=upper middle

-5.00

[ 15.9 %]

27.9

29.7 %

Contraceptive method used=Short-term

Examples

[ 34.7 %] 511

Att - Desc

Test value

Group

Overall

Continuous attributes : Mean (StdDev)

Discrete attributes : [Recall] Accuracy

Husband's occupation=upper middle

3.47

[ 40.0 %]

45.8

39.7 %

Husband's occupation=high

-0.15

[ 33.3 %]

1.8

1.8 %

Husband's occupation=middle

-0.17

[ 34.4 %]

28.6

28.9 %

Husband's occupation=low

-3.51

[ 28.0 %]

23.9

29.6 %

Computation time : 0 ms.

Created at 3/19/2019 8:24:53 PM

Figure 17. Lien entre l'occupation du mari avec méthode de contraceptive utilisée



## 5. Analyse Factorielle

Nous allons faire une analyse factorielle de notre modèle de données. Etant donné que notre modèle est composé de données mixtes (attributs qualitatifs et quantitatifs) comment nous l'allons décrit dans le début de ce travail, nous nous offrons donc de faire une analyse factorielle des données mixte (AFDM).

### A. Analyse Factorielle des Données Mixtes

Cette analyse prend en entrée toutes les variables de notre modèle (2 variables continues et 8 discrètes)

1. Interprétation de résultat (Tableau des valeurs propres) Il indique la variance expliquée par les axes. Nous avons  $p = 19$  facteurs car nous avons 2 variables quantitatives et 8 variables qualitatives avec respectivement 4, 4, 2, 2, 4, 4, 2, 3 modalités. Dans le tableau de données utilisé pour les calculs internes, nous avons  $2 + 4 + 4 + 2 + 2 + 4 + 4 + 2 + 3 = 27$  colonnes. Mais, nous avons introduit artificiellement de la colinéarité. En effet la somme des indicatrices d'une variable qualitative vaut systématiquement 1. De fait, le nombre de valeurs propre non nulles est en réalité égale à :  $2 + [(4 - 1) + (4 - 1) + (2 - 1) + (2 - 1) + (4 - 1) + (4 - 1) + (2 - 1) + (3 - 1)] = 19$ . Confirmé par les résultats ici, la somme des valeurs propres est bien égale à 19, et les 19 premiers axes expliquent 100.

#### Eigen values

Matrix trace = 19.00

Axis	Eigen value	% explained	Histogram	% cumulated
1	2.682348	14.12%		14.12%
2	1.932456	10.17%		24.29%
3	1.360019	7.16%		31.45%
4	1.173104	6.17%		37.62%
5	1.123397	5.91%		43.53%
6	1.027789	5.41%		48.94%
7	1.014232	5.34%		54.28%
8	0.996952	5.25%		59.53%
9	0.967791	5.09%		64.62%
10	0.919392	4.84%		69.46%
11	0.893641	4.70%		74.16%
12	0.871694	4.59%		78.75%
13	0.832729	4.38%		83.13%
14	0.743225	3.91%		87.05%
15	0.689756	3.63%		90.68%
16	0.584766	3.08%		93.75%
17	0.521716	2.75%		96.50%
18	0.342382	1.80%		98.30%
19	0.322613	1.70%		100.00%
Tot.	19.000000	-	-	-

Tanagra a mis en surbrillance les trois premiers. Pour une prise approximative de 10 sélectionne les axes portés pour une valeur propre supérieure au seuil. Les trois premiers axes (3.33, 1.80 et 1.19) passent haut la main ce seuil. Nous les conservons pour l'interprétation.

## B. Tableau des coordonnées

Le tableau des coordonnées décrit l'impact des variables, qu'elles soient quantitatives ou qualitative, dans la définition des axes. Pour les premières, il s'agit du carré du coefficient de corrélation linéaire ; pour les secondes les valeurs correspondent au carré du rapport de corrélation.

### Squared Correlation (Communalities)

Attribute	Axis_1			Axis_2			Axis_3			Axis_4			Axis_5		
	Coord.	CTR (%)	QLT % (Tot. %)	Coord.	CTR (%)	QLT % (Tot. %)	Coord.	CTR (%)	QLT % (Tot. %)	Coord.	CTR (%)	QLT % (Tot. %)	Coord.	CTR (%)	QLT % (Tot. %)
Wifes_age (*)	0.005998	0.2 %	1 % (1 %)	0.626660	32.4 %	63 % (63 %)	0.105260	7.7 %	11 % (74 %)	0.010699	0.9 %	1 % (75 %)	0.003825	0.3 %	0 % (75 %)
Wifes_education (**)	0.691372	25.8 %	23 % (23 %)	0.325932	16.9 %	11 % (34 %)	0.176084	12.9 %	6 % (40 %)	0.096852	8.3 %	3 % (43 %)	0.388806	34.6 %	13 % (56 %)
Husbands_education (**)	0.606524	22.6 %	20 % (20 %)	0.179352	9.3 %	6 % (26 %)	0.231674	17.0 %	8 % (34 %)	0.160939	13.7 %	5 % (39 %)	0.342725	30.5 %	11 % (51 %)
Number_of_children_ever_born (*)	0.032133	1.2 %	3 % (3 %)	0.360389	18.6 %	36 % (39 %)	0.366243	26.9 %	37 % (76 %)	0.013673	1.2 %	1 % (77 %)	0.020125	1.8 %	2 % (79 %)
Wifes_religion (**)	0.134703	5.0 %	13 % (13 %)	0.016372	0.8 %	2 % (15 %)	0.008001	0.6 %	1 % (16 %)	0.179591	15.3 %	18 % (34 %)	0.067883	6.0 %	7 % (41 %)
Wifes_now_working%3F (**)	0.009206	0.3 %	1 % (1 %)	0.002893	0.1 %	0 % (1 %)	0.087332	6.4 %	9 % (10 %)	0.221520	18.9 %	22 % (32 %)	0.003845	0.3 %	0 % (32 %)
Husbands_occupation (**)	0.377845	14.1 %	13 % (13 %)	0.097027	5.0 %	3 % (16 %)	0.130118	9.6 %	4 % (20 %)	0.339731	29.0 %	11 % (31 %)	0.083866	7.5 %	3 % (34 %)
Standard-of-living_index (**)	0.415830	15.5 %	14 % (14 %)	0.050386	2.6 %	2 % (16 %)	0.091358	6.7 %	3 % (19 %)	0.076855	6.6 %	3 % (21 %)	0.153058	13.6 %	5 % (26 %)
Media_exposure (**)	0.223271	8.3 %	22 % (22 %)	0.153540	7.9 %	15 % (38 %)	0.062821	4.6 %	6 % (44 %)	0.000743	0.1 %	0 % (44 %)	0.008583	0.8 %	1 % (45 %)
Contraceptive_method_used (**)	0.185465	6.9 %	9 % (9 %)	0.119906	6.2 %	6 % (15 %)	0.101129	7.4 %	5 % (20 %)	0.072501	6.2 %	4 % (24 %)	0.050681	4.5 %	3 % (26 %)
Var. Expl.	2.682348	-	14 % (14 %)	1.932456	-	10 % (24 %)	1.360019	-	7 % (31 %)	1.173104	-	6 % (38 %)	1.123397	-	6 % (44 %)

(\*) Square of correlation coefficient  
(\*\*) Correlation ratio

La somme des valeurs en lignes vaut « 1 » pour les variables quantitatives, « nombre de modalités -1 » pour les qualitatives. Ainsi, le pourcentage en ligne (d'information de la variable qui est retranscrite par l'axe. Pour la variables Husband's education par exemple, 77 pourcent de l'information qu'elle véhicule est située par le premier axe. Il y a très peu de chances qu'elle pèse significativement sur les autres. La somme des valeurs en colonnes est égale à la valeur propre associé à l'axe. Nous pouvons y voir la contribution de la variable dans la définition du facteur.

## C. Tableau des corrélations

Ce tableau précise le sens des relations entre les variables quantitatives et les facteurs.

### Continuous Attributes - Correlation (Factor Loadings)

Attribute	Axis_1	Axis_2	Axis_3	Axis_4	Axis_5
Wifes_age	0.077447	0.791618	0.324438	-0.103438	0.061843
Number_of_children_ever_born	-0.179257	0.600324	0.605180	0.116931	0.141862

## D. Tableau des moyennes conditionnelles

Ce tableau positionne les modalités sur les axes factoriels. Nous avons également un indicateur sur leurs contributions. Elles dépendent à la fois de l'écartement avec l'origine Wife's education, de l'effectif et de la valeur propre associées à l'axe. La somme des conditions des modalités doit être égale à la contribution de la variable.

Discrete Attributes - Conditional means and contributions

Attribute		Axis_1			Axis_2			Axis_3			Axis_4			Axis_5		
-		Mean	CTR (%)	v.test	Mean	CTR (%)	v.test	Mean	CTR (%)	v.test	Mean	CTR (%)	v.test	Mean	CTR (%)	v.test
Wifes_education	2	-1.2285	4.76	-15.584	-0.4756	1.37	-7.108	0.0847	0.09	1.510	0.2312	0.88	4.435	1.0716	20.63	21.005
	1	-2.4031	8.28	-19.096	2.0041	11.10	18.762	-0.7173	2.87	-8.005	0.1648	0.20	1.980	-0.7010	4.02	-8.607
	3	-0.2416	0.23	-3.515	-0.7080	3.74	-12.135	0.6971	7.31	14.243	-0.5420	5.94	-11.924	-0.6703	9.91	-15.069
	4	1.5159	12.51	28.496	0.2505	0.66	5.547	-0.3554	2.68	-9.384	0.2079	1.23	5.910	0.0406	0.05	1.181
	Tot.	-	25.77	-	-	16.87	-	-	12.95	-	-	8.26	-	-	34.61	-
Husbands_education	3	-1.0844	3.91	-14.235	-0.6294	2.54	-9.735	0.9198	10.93	16.957	-0.6323	6.94	-12.552	-0.0424	0.03	-0.860
	2	-2.1705	7.91	-18.851	0.6066	1.19	6.207	-0.5662	2.09	-6.906	0.7076	4.40	9.293	1.2702	15.45	17.047
	4	0.9785	8.12	28.687	-0.0026	0.00	-0.091	-0.1823	1.10	-7.507	0.1471	0.96	6.519	-0.1142	0.63	-5.174
	1	-2.5365	2.67	-10.427	2.6353	5.55	12.762	-1.3429	2.91	-7.752	-0.8087	1.42	-5.027	-2.4660	14.39	-15.664
	Tot.	-	22.61	-	-	9.28	-	-	17.03	-	-	13.72	-	-	30.51	-
Wifes_religion	1	-0.2519	0.75	-14.081	-0.0745	0.13	-4.909	0.0437	0.09	3.432	0.1923	2.29	16.259	-0.1157	0.90	-9.996
	0	1.4345	4.27	14.081	0.4245	0.72	4.909	-0.2489	0.50	-3.432	-1.0954	13.02	-16.259	0.6590	5.14	9.996
	Tot.	-	5.02	-	-	0.85	-	-	0.59	-	-	15.31	-	-	6.04	-
Wifes_now_working%3F	1	-0.0908	0.09	-3.681	-0.0432	0.04	-2.064	0.1992	1.61	11.338	0.2947	4.73	18.058	-0.0380	0.09	-2.379
	0	0.2718	0.26	3.681	0.1293	0.11	2.064	-0.5961	4.81	-11.338	-0.8818	14.15	-18.058	0.1137	0.26	2.379
	Tot.	-	0.34	-	-	0.15	-	-	6.42	-	-	18.88	-	-	0.34	-
Husbands_occupation	2	-0.3257	0.43	-4.859	0.0561	0.02	0.986	0.0643	0.06	1.346	-0.9833	20.27	-22.182	0.4653	4.95	10.726
	3	-0.8208	3.72	-15.607	-0.4778	2.43	-10.704	0.1672	0.60	4.464	0.3157	2.88	9.078	-0.1798	1.02	-5.282
	1	1.5104	9.39	22.943	0.5494	2.39	9.832	-0.1031	0.17	-2.200	0.5103	5.60	11.721	-0.2353	1.30	-5.523
	4	-1.4786	0.56	-4.733	0.5987	0.18	2.258	-2.9685	8.73	-13.345	0.3968	0.21	1.921	0.3706	0.20	1.833
	Tot.	-	14.09	-	-	5.02	-	-	9.57	-	-	28.96	-	-	7.47	-
Standard-of-living_index	3	-0.1572	0.10	-2.368	-0.3512	0.97	-6.233	0.4173	2.75	8.829	0.2203	1.03	5.018	-0.6188	8.88	-14.405
	4	0.9548	5.88	20.826	0.3300	1.35	8.480	-0.1039	0.27	-3.182	-0.0799	0.22	-2.637	0.3406	4.27	11.479
	2	-1.3499	3.94	-13.568	-0.2306	0.22	-2.731	0.0220	0.00	0.311	-0.5150	3.00	-7.827	-0.0008	0.00	-0.012
	1	-2.1412	5.58	-15.540	-0.1672	0.07	-1.429	-0.8826	3.69	-8.995	0.6022	2.31	6.609	0.2629	0.48	2.949
	Tot.	-	15.50	-	-	2.61	-	-	6.72	-	-	6.55	-	-	13.62	-
Media_exposure	0	0.2188	0.62	18.129	-0.1540	0.59	-15.034	0.0826	0.34	9.616	-0.0083	0.00	-1.046	0.0278	0.06	3.554
	1	-2.7376	7.71	-18.129	1.9269	7.36	15.034	-1.0340	4.28	-9.616	0.1044	0.06	1.046	-0.3474	0.71	-3.554
	Tot.	-	8.32	-	-	7.95	-	-	4.62	-	-	0.06	-	-	0.76	-
Contraceptive_method_used	1	-0.5967	2.11	-12.067	0.1590	0.29	3.788	-0.4290	4.25	-12.183	-0.3166	3.11	-9.682	-0.2034	1.40	-6.355
	2	1.2332	4.78	15.613	0.6366	2.45	9.495	0.2868	1.01	5.100	0.0695	0.08	1.331	-0.1129	0.23	-2.208
	3	-0.0691	0.02	-1.181	-0.6105	3.46	-12.281	0.3411	2.18	8.179	0.3444	2.99	8.892	0.3239	2.88	8.545
	Tot.	-	6.91	-	-	6.20	-	-	7.44	-	-	6.18	-	-	4.51	-

Prenons le cas de la variable Husband's education sur le troisième axe. Le carré du rapport de corrélation est  $\text{COORD2}(\text{Husband's education}) = 0.231674$ . Sa contribution est donc de  $\text{CONTRIB2}(\text{Husband's education}) = 0.231674 / 1.360019 \times 100 = 17.03$ . Voyons maintenant la modalité (3 = upper middle) pour le cas où le mari a une éducation supérieure à la moyenne. Elle correspond à 352 observations. Sa contribution est obtenue par  $\text{CONTRIB2}(\text{Husband's education} = 3) = [352 \times (-0.9198 - 0)^2] / [1473 \times 1.3600192] \times 100 = 10.93$ .

## E. Tableau des vecteurs propres

Le tableau des vecteurs propres est utilisé pour le déploiement ; c'est-à-dire de la projection d'un nouvel individu dans le repère factoriel. Il faut center ('center') et réduire ('scale') les valeurs prises sur chaque description, puis appliquer les coefficients associés aux axes. L'objectif est de positionner un nouveau



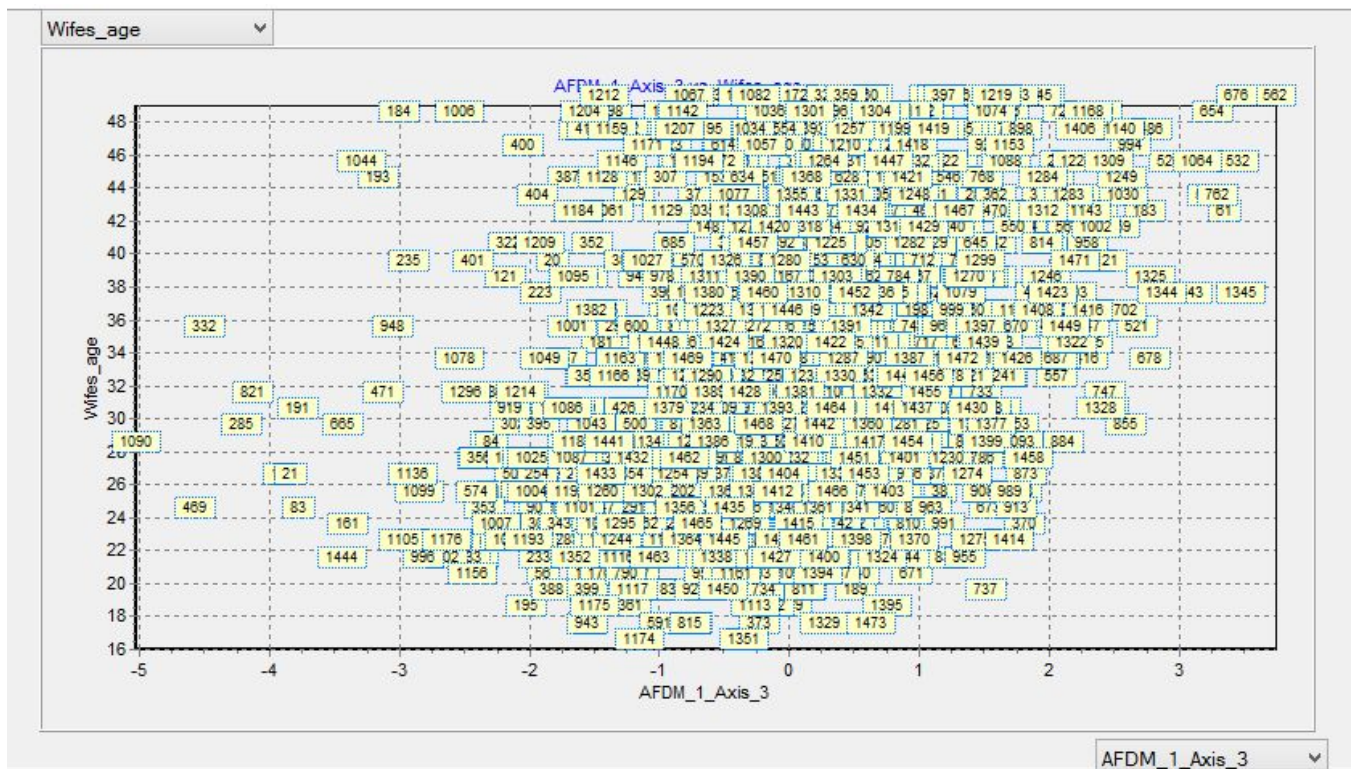
venu par rapport aux femmes déjà existantes. Il sera ainsi aisé de déduire ses caractéristiques à partir des observations situées dans son voisinage.

## Eigen vectors - Factor Scores

Attribute	Center	Scale	Axis_1	Axis_2	Axis_3	Axis_4	Axis_5
Wifes_age	32.538357	8.224452	0.047288	0.569457	0.278201	-0.095501	0.058348
Wifes_education = 2	0.226748	0.476181	-0.218091	-0.117199	0.029669	0.093846	0.454223
Wifes_education = 1	0.103191	0.321233	-0.287795	0.333144	-0.169432	0.045117	-0.200443
Wifes_education = 3	0.278344	0.527583	-0.047517	-0.193292	0.270426	-0.243754	-0.314786
Wifes_education = 4	0.391718	0.625873	0.353697	0.081116	-0.163567	0.110916	0.022644
Husbands_education = 3	0.238968	0.488844	-0.197631	-0.159224	0.330621	-0.263497	-0.018445
Husbands_education = 2	0.120842	0.347623	-0.281296	0.109123	-0.144715	0.209696	0.393056
Husbands_education = 4	0.610319	0.781229	0.284991	-0.001066	-0.104733	0.097932	-0.079423
Husbands_education = 1	0.029871	0.172832	-0.163437	0.235689	-0.170657	-0.119151	-0.379392
Number_of_children_ever_born	3.261371	2.357748	-0.109451	0.431848	0.518934	0.107959	0.133844
Wifes_religion = 1	0.850645	0.922304	-0.086605	-0.035572	0.029642	0.151211	-0.095000
Wifes_religion = 0	0.149355	0.386465	0.206683	0.084893	-0.070741	-0.360868	0.226720
Wifes_now_working%3F = 1	0.749491	0.865731	-0.029322	-0.019366	0.126831	0.217495	-0.029283
Wifes_now_working%3F = 0	0.250509	0.500509	0.050718	0.033497	-0.219380	-0.376202	0.050651
Husbands_occupation = 2	0.288527	0.537147	-0.065227	0.015590	0.025380	-0.450247	0.222490
Husbands_occupation = 3	0.397149	0.630197	-0.192845	-0.155830	0.077467	0.169610	-0.100847
Husbands_occupation = 1	0.295995	0.544054	0.306350	0.154673	-0.041253	0.236670	-0.113964
Husbands_occupation = 4	0.018330	0.135388	-0.074628	0.041944	-0.295510	0.045793	0.044660
Standard-of-living_index = 3	0.292600	0.540925	-0.031699	-0.098293	0.165969	0.101565	-0.297943
Standard-of-living_index = 4	0.464358	0.681439	0.242568	0.116370	-0.052045	-0.046436	0.206588
Standard-of-living_index = 2	0.155465	0.394291	-0.198429	-0.047059	0.006389	-0.173102	-0.000276
Standard-of-living_index = 1	0.087576	0.295933	-0.236234	-0.025598	-0.192038	0.151915	0.069261
Media_exposure = 0	0.926001	0.962290	0.078482	-0.076678	0.058464	-0.006845	0.023777
Media_exposure = 1	0.073999	0.272027	-0.277629	0.271245	-0.206816	0.024215	-0.084110
Contraceptive_method_used = 1	0.427020	0.653467	-0.145363	0.053766	-0.206109	-0.176369	-0.118303
Contraceptive_method_used = 2	0.226069	0.475467	0.218591	0.156621	0.100274	0.028185	-0.047769
Contraceptive_method_used = 3	0.346911	0.588992	-0.015183	-0.186085	0.147725	0.172923	0.169815

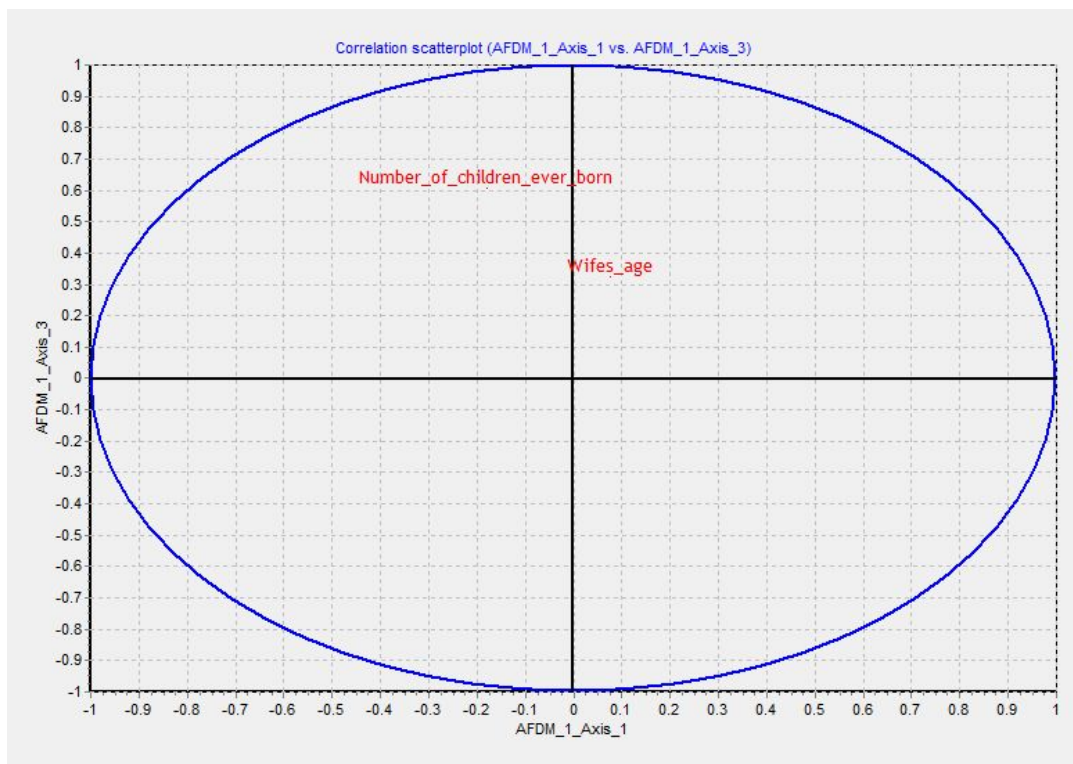
## F. Représentation Graphique

Parlons-en justement des représentations graphiques. Lorsque les observations sont étiquetées, le positionnement des individus dans le repère factoriel est particulièrement intéressant



## G. Cercle des corrélations

Par rapport au tableau des corrélations , l'outil « Cercle des corrélations » permet l'introduction des variables illustratives.

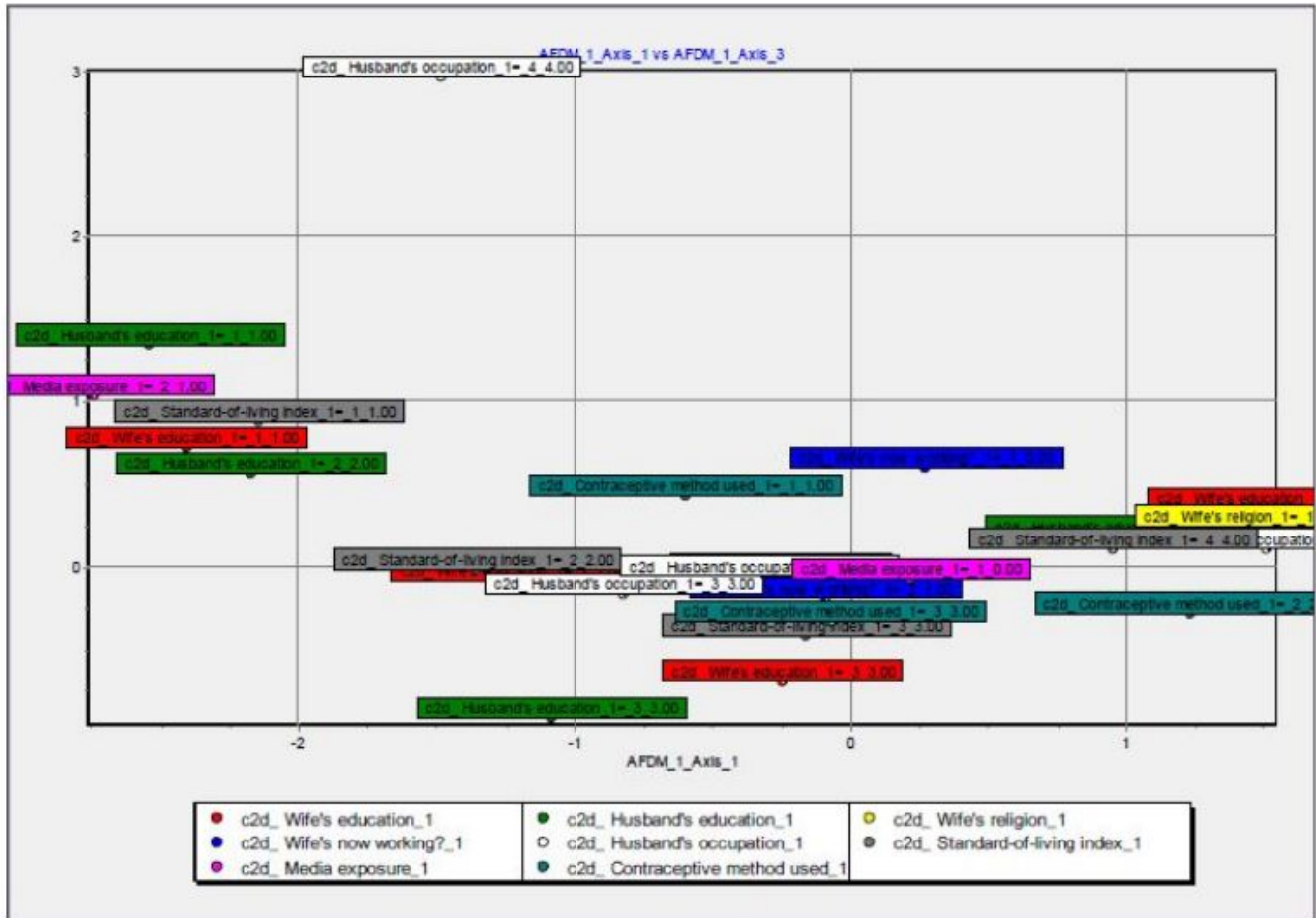




Toujours à la suite de notre analyse factorielle de données mixtes, nous plaçons en Target les deux premiers axes et en input nos deux variables quantitatives actives.

## H. Moyennes conditionnelles

De même manière, par rapport au tableau des moyennes conditionnelles, l'outil graphique de Tanagra autorise l'introduction des variables illustratives. L'interprétation des axes peut en être renforcée. Comme pour le cercle de corrélation, nous introduisons en Target les deux premier axes et en input nos deux variables quantitatives. Nous observons les positionnements des différentes modalités des variables de l'étude.



Le positionnement des modalités semble être cohérent par rapport à tous les résultats trouvés jusqu'ici. D'après la proportion de la modalité (4 = high) pour la variable Husband's occupation, nous avons vu que cette dernière a vraiment peut, c'est-à-dire peu des maris ont une occupation qui rapporte vraiment beaucoup d'argent, et par conséquent remarqué sa position. Mais aussi pour la variable Method contraceptif used ; voyons que pour la modalité (1 = no-use) elle est positionnée au-dessus des autres modalités et cela confirme sa proportion trouvée dans la partie précédente.

## 6. Classification automatique

Nous allons essayer de regrouper nos objets (les femmes) selon certains critères de ressemblance. Les diverses techniques de classification visent toutes à répartir  $n$  individus, caractérisés par  $n$  variables  $x_1, x_2, \dots, x_p$  en un certain nombre  $m$  de sous-groupes aussi homogènes que possible, chaque groupe étant bien différencié des autres. Pour y arriver nous allons utiliser deux grands algorithmes à savoir ;

1. L'algorithme de CAH
2. L'algorithme de K-Means

### 6.1. L'algorithme de CAH

La Classification Ascendante Hiérarchique (CAH) est l'une d'entre elles. On cherche à ce que les individus regroupés au sein d'une même classe (homogénéité intra-classe) soient le plus semblables possibles tandis que les classes soient le plus dissemblables (hétérogénéité interclasse). Nous choisissons la CAH pour la construction de la typologie.

#### 1. diagramme dendrogramme

Le dendrogramme est un diagramme fréquemment utilisé pour illustrer l'arrangement de groupes générés par un regroupement hiérarchique ou hiérarchisant. Ainsi sur la figure suivante le dendrogramme nous montre comment nos individus sont regroupés en trois classes.

Clustering results		
Clusters	From the	After one-pass
	dendrogram	relocation
cluster n1	851	795
cluster n2	310	350
cluster n3	312	328

#### 2. Caractérisation des groupes

Après avoir regroupé nos individus en différents groupes l'étape suivante consiste à étudier les caractéristiques de chaque groupe. L'outil GROUP CHARACTERIZATION est le plus approprié pour cela, il permet de comparer les indicateurs (moyenne ou proportion) marginaux et conditionnellement aux groupes. Pour ce faire nous plaçons en Target la variable CAH obtenue précédemment et en entrée nous plaçons nos deux variables continues avec notre variable discrète qui détermine la Méthode contraceptive utilisée par la femme.

### 1. La classe ClusterHAC2 = chac1

Nous remarquons bien que cette classe est constituée des femmes trop jeunes et quelques une ont l'âge moyen c'est leur âge varie(16 à 32).semblent avoir peu d'enfants au même pas presque. Cette classe contient 40.9 pourcent des femmes utilisant les contraceptif à long terme, 41.8 pourcent n'utilisent pas les contraceptifs et 17.4 à court terme.

Cluster_HAC_2=c_hac_1			
Examples		[ 54.0 %] 795	
Att - Desc	Test value	Group	Overall
Continuous attributes : Mean (StdDev)			
Number of children ever born	-21.86	2.02 (1.30)	3.26 (2.36)
Wife's age	-31.98	26.21 (4.09)	32.54 (8.23)
Discrete attributes : [Recall] Accuracy			
c2d_ Contraceptive method used_2=_3_3.00	5.4	[ 63.6 %] 40.9 %	34.70%
c2d_ Contraceptive method used_2=_1_1.00	-0.79	[ 52.8 %] 41.8 %	42.70%
c2d_ Contraceptive method used_2=_2_2.00	-5.21	[ 41.4 %] 17.4 %	22.60%

Voyons que 63.6% de femmes utilisant les contraceptifs à long terme se trouvent de cette classe. Nous pouvons dire que, dans cette classe les femmes utilisant les contraceptifs à long terme ce sont elles qui ont déjà l'âge moyen avec 3 à 6 enfants. Car dans cette classe personne n'a plus de six enfants. Pour celle n'utilisant pas les contraceptifs sont celles qui sont encore trop jeunes et ont besoin d'avoir les enfants.

### 2. La classe ClusterHAC2 = chac2

Dans cette deuxième classe nous trouvons qu'il y a plus des femmes qui ont l'âge avancé soit 33 ans et plus sans enfant ou peu (entre 0 et 5 ans) et qui ont besoin d'avoir des enfants. Presque la moitié des femmes de cette classe n'utilisent pas les contraceptifs. Et environs 30% l'utiliser juste pour le court terme (peut-être pour espacer). Seulement 16.6% des femmes utilisant les contraceptives à long terme font parties de cette classe.

Cluster_HAC_2=c_hac_2			
Examples		[ 23.8 %] 350	
Att - Desc	Test value	Group	Overall
Continuous attributes : Mean (StdDev)			
Wife's age	20.3	40.33 (4.62)	32.54 (8.23)
Number of children ever born	-3.96	2.83 (1.21)	3.26 (2.36)
Discrete attributes : [Recall] Accuracy			
c2d_ Contraceptive method used_2=_2_2.00	3.05	[ 30.0 %] 28.6 %	22.60%
c2d_ Contraceptive method used_2=_1_1.00	1.92	[ 26.2 %] 47.1 %	42.70%
c2d_ Contraceptive method used_2=_3_3.00	-4.68	[ 16.6 %] 24.3 %	34.70%



Il est évident que, une femme qui déjà plus de 33 ans et n'a pas encore d'enfant elle n'a pas vraiment besoins de contraceptif car elle cherche à tout pris à avoir des enfants. C'est ce qui nous donne la deuxième classe.

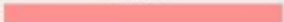

### 3. La classe ClusterHAC2 = chac3

Dans cette dernière classe se trouvent les femmes les plus âgées et celles qui ont l'âge moyen et qui ont déjà beaucoup d'enfants. C'est pour cela que les proportions de l'utilisation des contraceptif semble être presque égale avec 29% à court terme, 40% n'utilisent pas et 30% à long terme. Cela s'explique par le fait que, une femme plus âgée qui n'est plus dans la période de procréation même elle a déjà beaucoup d'enfants elle peut ne plus utiliser le contraceptif, le 40% des femmes n'utilisant pas les contraceptifs car elles ne sont plus dans la période de procréation. Le 30% à long-terme c'est évidemment les femmes ayant déjà beaucoup d'enfants et elles sont toujours procréatives.

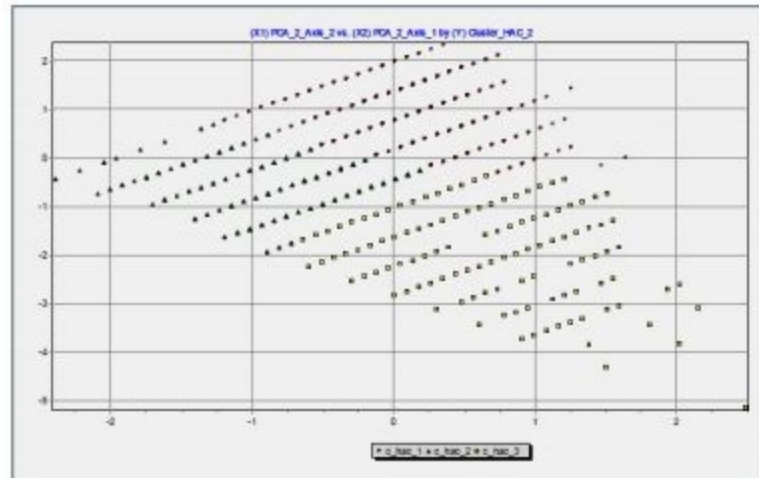
Cluster_HAC_2=c_hac_3			
Exemples		[ 22.3 %] 328	
Att - Desc	Test value	Group	Overall
Continuous attributes : Mean (StdDev)			
Number of children ever born	30.24	6.73 (1.83)	3.26 (2.36)
Wife's age	17.54	39.57 (5.41)	32.54 (8.23)
Discrete attributes : [Recall] Accuracy			
c2d_ Contraceptive method used_2=_2_2.00	3.12	[ 28.5 %] 29.0 %	22.60%
c2d_ Contraceptive method used_2=_1_1.00	-1.02	[ 21.0 %] 40.2 %	42.70%
c2d_ Contraceptive method used_2=_3_3.00	-1.68	[ 19.8 %] 30.8 %	34.70%

### 4. Analyse de Composante Principales

Pour visualiser les groupes et mieux les situer les uns par rapport aux autres, nous allons les projeter dans le premier plan factoriel. Nous ajoutons donc un composant PCA (Principal Component Analysis) dans notre diagramme.

Eigen values					
Matrix trace		2.000000			
Average		1.000000			
Axis	Eigen value	Difference	Proportion (%)	Histogram	Cumulative (%)
1	1.540126	1.080252	77.01 %		77.01 %
2	0.459874	-	22.99 %		100.00 %
Tot.	2.000000	-	-	-	-

Le premier axe résume 77% de l'information. Cela laisse à penser que nous obtiendrons une représentation satisfaisante des proximités entre les observations.



Pour bien clarifier ce que nous avons dit sur la caractérisation dans le point précédent, essayons de projeter nos observations suivant l'âge de la femme et le nombre d'enfants qu'elle a déjà

## 6.2. L'algorithme de K-Means

K-means est un algorithme de classification automatique (apprentissage non supervisé). Son objectif étant la création des sous-groupes homogènes des exemples. Les individus du même sous-groupe sont similaires. Les individus dans différents sous-groupes sont aussi différents que possible. Considérant toujours notre jeu, les variables actives qui vont participer à la création des groupes sont ; Wife's age (celle qui détermine l'âge de la femme concernant l'utilisation de contraceptif) et aussi Number of children ever born (le nombre d'enfants déjà mis au monde par cette femme). La variable illustrative utilisée sera Contraceptif méthode used qui a trois valeurs possibles (no-use, long-term ou short-term).

### 1. Normalisation des variables actives

Nous voulons normaliser les variables avant d'appliquer l'approche K-Means. L'objectif est d'éliminer les disparités d'échelle entre les variables. Nous ajoutons le composant STANDARDIZE (FEATURE CONSTRUCTION) dans le diagramme.

Attribute standardization	
Src att	New att
Wife's age	std_Wife's age_1
Number of children ever born	std_Number of children ever born_1
Computation time : 0 ms.	
Created at 5/1/2019 12:59:33 PM	

## 2. L'approche K-Means

Nous nous servons des variables normalisées que nous plaçons en entrée. Et nous laissons Tanagra détecter lui-même les classes nécessaires. Vu que les variables utilisées sont déjà normalisées ce n'est plus nécessaire d'utiliser la distance normalisée. 2 Cfr

## Global evaluation

Within Sum of Squares	948.4368
Total Sum of Squares	2944.0001
R-Square	0.6778

## Cluster size and WSS

Clusters	3		
Cluster	Description	Size	WSS
cluster n°1	c_kmeans_1	649	272.7969
cluster n°2	c_kmeans_2	276	184.4412
cluster n°3	c_kmeans_3	548	391.1988

## R-Square for each attempt

Number of trials	5
Trial	R-square
1	0.677841
2	0.674681
3	0.677541
4	0.677841
5	0.677541

## Cluster centroids

Attribute	Cluster n°1	Cluster n°2	Cluster n°3
std_Wife's age_1	-0.914769	1.016150	0.371579
std_Number of children ever born_1	-0.645869	1.611255	-0.046601

Use [GRIPEP-LYNAC-TEDIGAT-3.0P](#) for detailed copyright notice.

Le TSS(Total Sum Square) est de 2944.0001 ; le WSS(Within Sum of Square) est 948.4368. Le BSS (Between Sum of Square) expliqué par le partitionnement est  $BSS = TSS - WSS$  d'où  $2944.0001 - 948.4368 = 1995.5633$ . Le ratio résultant est  $1995.5633 / 2944.0001 = 67.78\%$  Nous avons 649 exemples dans la première classe, 276 dans la deuxième classe et 548 dans la troisième. Dans la partie inférieure de la figure, la section CLUSTERS CENTROIDS donne la moyenne de chaque variable en fonction des clusters (classes).

3. Interprétation des groupes Nous sommes maintenant dans l'étape majeure du processus de classification ; nous voulons interpréter les groupes. Quelles sont les caractéristiques de chaque cluster ? Qu'est-ce qui se différencie les uns des autres ? Nous jugeons moins nécessaire

d'utiliser le composant View Dataset parce que notre jeu contient beaucoup de données et nous n'avons pas non plus de label pour identifier chacun des exemples.

#### 4. La première classe ClusterKMeans2 = ckmeans1

la première classe de K-Means est très proche de la première classe de HAC. Cela étant dit, dans cette classe nous retrouvons plus des femmes qui sont moins âgées et avec celle qui ont l'âge moyen c'est-à-dire de 16 à 34 ans. 50 pourcent des femmes utilisant les contraceptifs à long-terme se trouvent dans cette classe, ceci nous pouvons dire que, les femmes de classe qui ont déjà l'âge moyen soit 28 à 34 ans et qui 4 à 6 enfants n'ont plus vraiment besoin de faire d'autres enfants. Dans cette classe il y a 44% de femmes n'utilisant pas les contraceptifs et cela s'explique par le fait que les plus jeune femmes de cette classe qui n'ont pas encore d'enfant ou qui en ont moins n'ont pas vraiment d'utiliser le contraceptifs vu qu'elles ont besoin de faire les enfants.

Cluster_KMeans_2=c_kmeans_1			
Examples		[ 44.1 %] 649	
Att - Desc	Test value	Group	Overall
Continuous attributes : Mean (StdDev)			
Number of children ever born	-21.99	1.74 (1.16)	3.26 (2.36)
Wife's age	-31.15	25.01 (3.47)	32.54 (8.23)
Discrete attributes : [Recall] Accuracy			
c2d_ Contraceptive method used_2=_3_3.00	3.51	[ 50.3 %] 39.6 %	34.70%
c2d_ Contraceptive method used_2=_1_1.00	1.05	[ 45.6 %] 44.2 %	42.70%
c2d_ Contraceptive method used_2=_2_2.00	-5.23	[ 31.5 %] 16.2 %	22.60%

ClusterKMeans2 = ckmeans1

Cluster_KMeans_2=c_kmeans_2			
Examples		[ 18.7 %] 276	
Att - Desc	Test value	Group	Overall
Continuous attributes : Mean (StdDev)			
Number of children ever born	29.68	7.06 (1.82)	3.26 (2.36)
Wife's age	18.72	40.90 (5.46)	32.54 (8.23)
Discrete attributes : [Recall] Accuracy			
c2d_ Contraceptive method used_2=_2_2.00	1.69	[ 21.9 %] 26.4 %	22.60%
c2d_ Contraceptive method used_2=_1_1.00	0.83	[ 19.7 %] 44.9 %	42.70%
c2d_ Contraceptive method used_2=_3_3.00	-2.35	[ 15.5 %] 28.6 %	34.70%

Remarquons que, c'est dans cette classe où on retrouve peu de proportion des observations avec seulement 21% des femmes utilisant les contraceptifs à court-terme sont dans cette classe et aussi 19% pour celles qui n'utilisent pas sont dans cette classe et 15% pour le long-terme

## 5. ClusterKMeans2 = ckmeans3

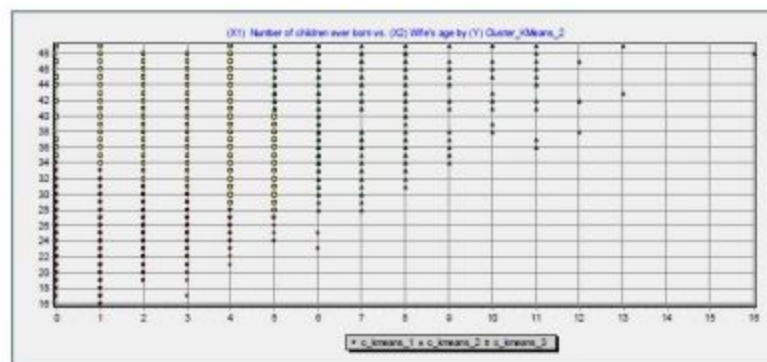
Contrairement à la deuxième classe, dans cette classe il y a de femmes avec âge moyen et âge avancé mais avec peu d'enfants c'est-à-dire moins de 5 enfants. Voyons que plus de 30 pourcent de femmes utilisent les contraceptifs à long-terme, cela s'explique par leur âge avancé et ne veulent plus faire des enfants, cela se marie avec le 39% de celles qui n'utilisent plus le contraceptif parce qu'elles n'ont plus l'âge de procréation.

Cluster_KMeans_2=c_kmeans_3			
Examples		[ 37.2 %] 548	
Att - Desc	Test value	Group	Overall
Continuous attributes : Mean (StdDev)			
Wife's age	16.88	37.24 (5.32)	32.54 (8.23)
Number of children ever born	-1.38	3.15 (1.29)	3.26 (2.36)
Discrete attributes : [Recall] Accuracy			
c2d_ Contraceptive method used_2= 2_2.00	4.01	[ 46.5 %] 28.3 %	22.60%
c2d_ Contraceptive method used_2= 3_3.00	-1.71	[ 34.2 %] 31.9 %	34.70%
c2d_ Contraceptive method used_2= 1_1.00	-1.74	[ 34.7 %] 39.8 %	42.70%

La plupart de femmes utilisent les contraceptifs à court-terme se trouvent dans cette classe avec 46.5%

## 6. Représentation graphique

La représentation graphique est un autre moyen de mettre en évidence les résultats de la classification. Le nuage de points est un outil très utile dans ce contexte. Nous pouvons positionner les groupes en fonction de deux variables simultanément. Ainsi, nous pouvons vérifier s'il existe des interactions entre les variables.



## CONCLUSION

Dans notre travail, nous avons présenté d'un modèle de jeu des données et analyse des attributs de ce dernier (première partie), une analyse factorielle de données mixtes (deuxième partie) et une classification automatique avec l'algorithme de K-Means et CAH dans la dernière partie avec l'outil d'analyse des données Tanagra. Par rapport à l'objectif que nous nous sommes fixé, nous disons être satisfait des résultats.

# Machine à vecteurs de support

## Introduction

Les machines à vecteurs supports (ou séparateurs à vaste marges) sont des techniques d'apprentissage supervisé dont le but est de résoudre les problèmes de discrimination, c'est-à-dire déterminer la classe à laquelle appartient un individu (individu est employé au sens de constituant d'un ensemble), ou de régression, c'est-à-dire prédire la valeur numérique d'une variable. Le succès de cette méthode est justifié par les solides bases théoriques qui la soutiennent.

Les SVM apportent une approche très intéressante de l'approximation statistique. Souvent, le nombre des exemples pour l'apprentissage est insuffisant pour que les estimateurs fournissent un modèle avec une bonne précision. D'un autre côté, l'acquisition d'un grand nombre d'exemples s'avère être souvent très coûteuse et peut même mener à des problèmes de sur-apprentissage dans le cas où la capacité du modèle est très complexe. Pour ces deux raisons, il faut arriver à un compromis entre la taille des échantillons et la précision recherchée.

La plupart des techniques de l'apprentissage machine possèdent un grand nombre de paramètres d'apprentissage à fixer par l'utilisateur (Structure d'un réseau de neurones, coefficient de mise à jour du gradient, . . .). De plus, avec ces méthodes, le nombre de paramètres à calculer par l'algorithme d'apprentissage est en relation linéaire, voire exponentielle, avec la dimension de l'espace d'entrée

SVM est donc une méthode de classification particulièrement bien adaptée pour traiter des données de très haute dimension telles que les textes et les images. L'idée principale des SVM consiste à projeter les données dans un espace de plus grande dimension appelé, espace de caractéristiques, tant que les données non linéairement séparables dans l'espace d'entrée deviennent linéairement séparables dans l'espace de caractéristiques.

En appliquant dans cet espace la technique de construction d'un hyperplan optimal séparant les deux classes, on obtient une fonction de classification qui dépend d'un produit scalaire des images des données de l'espace d'entrée dans l'espace des caractéristiques.

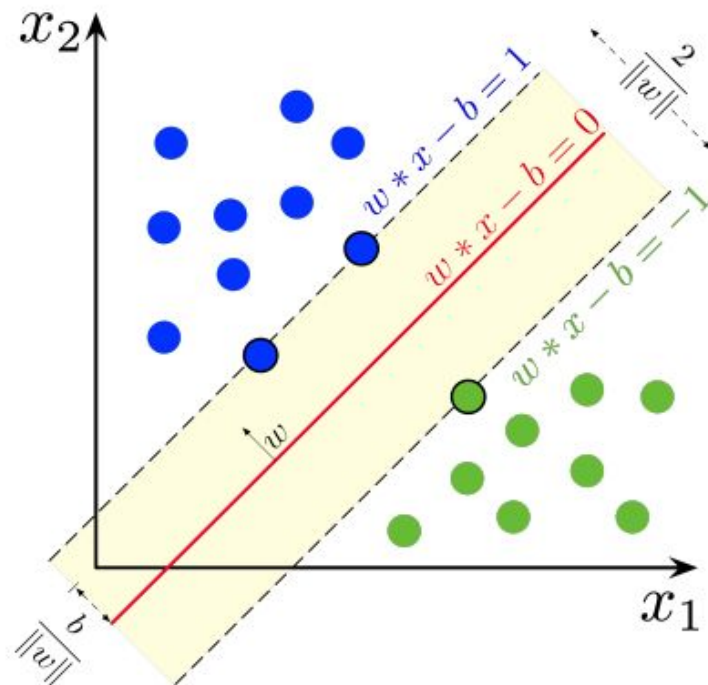
## 2.2- Principe de la technique SVM

Cette technique est une méthode de classification à deux classes qui tente de séparer les exemples positifs des exemples négatifs dans l'ensemble des exemples. La méthode cherche alors l'hyperplan qui sépare les exemples positifs des exemples négatifs, en garantissant que la marge entre le plus proche des positifs et des négatifs soit maximale. Cela garantit une généralisation du principe car de nouveaux exemples pourraient ne pas être trop similaires à ceux utilisés pour trouver l'hyperplan mais être situés d'un côté ou l'autre de la frontière

L'intérêt de cette méthode est la sélection de vecteurs supports qui représentent les vecteurs discriminant grâce auxquels est déterminé l'hyperplan. Alors les exemples utilisés lors de la recherche de l'hyperplan ne sont plus utiles et seuls ces vecteurs supports sont utilisés pour classer un nouveau cas, ce qui peut être considéré comme un avantage pour cette méthode.

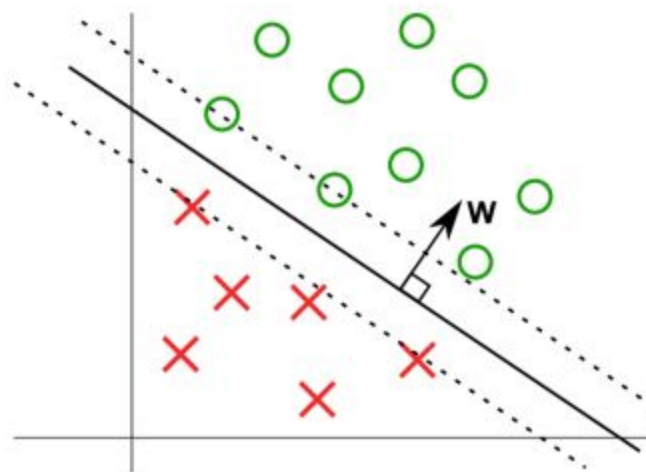
### 2.3- Support vecteur

Les vecteurs de support sont les points de données les plus proches de l'hyperplan. Ces points définissent mieux la ligne de séparation en calculant les marges. Ces points sont plus pertinents pour la construction du classificateur.



### 2.4- Hyperplan

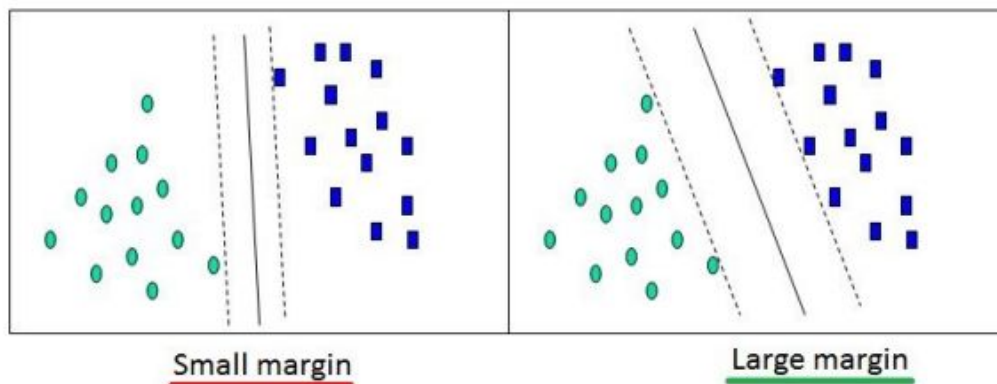
Un hyperplan est un plan de décision qui sépare un ensemble d'objets ayant différentes appartenances à une classe.





## 2.5- Marge

Une marge est un espace entre les deux lignes des points de classe les plus proches. Ceci est calculé comme la distance perpendiculaire de la ligne pour supporter les vecteurs ou les points les plus proches. Si la marge est plus grande entre les classes, elle est considérée comme une bonne marge, une marge plus petite est une mauvaise marge.



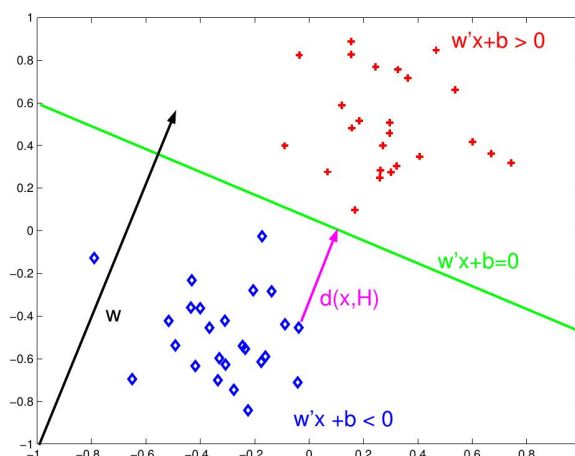
## 2.6- Linéarité et non-linéarité des données

### 2.6.1- Linéarité des données

La linéarité des données permet de trouver facilement le classificateur linéaire. Dans la plupart des problèmes réels il n'y a pas de séparation linéaire possible entre les données, le classificateur de marge maximale ne peut pas être utilisé car il fonctionne seulement si les classes de données d'apprentissage sont linéairement séparables.

Un classificateur est dit linéaire lorsqu'il est possible d'exprimer sa fonction de décision par une fonction linéaire en  $x$ . Dans la suite, nous supposons que les exemples nous sont fournis dans le format vectoriel.

Notre espace d'entrée  $X$  correspond donc à  $\mathbb{R}^n$  ou  $n$  est le nombre de composantes des vecteurs contenant les données.  $h(x) = w \cdot x + b$   $n=1 \quad w_i x_i + b$

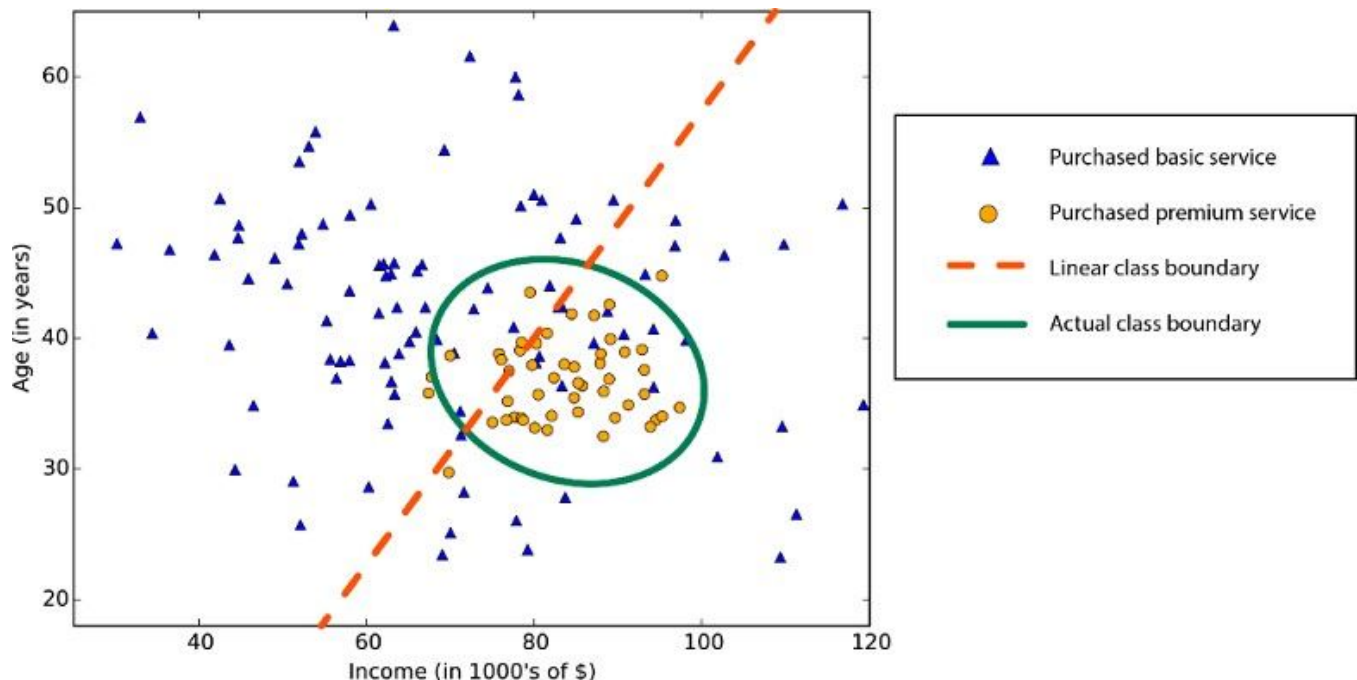




### 2.6.2- Non-Linéarité des données

Pour surmonter les inconvénients des cas non linéairement séparable, l'idée des SVM est de changer l'espace des données. La transformation non linéaire des données peut permettre une séparation linéaire des exemples dans un nouvel espace. On va donc avoir un changement de dimension. Ce nouvel espace est appelé « espace de redescription ». En effet, intuitivement, plus la dimension de l'espace de redescription est grande, plus la probabilité de pouvoir trouver un hyperplan séparateur entre les exemples est élevée.

Le classificateur de marge maximale ne peut pas être utilisé dans la plupart des problèmes réels : si les données ont été affectées par le bruit, il n'y a pas de séparation linéaire entre elles. Dans ce cas, le problème d'optimisation ne peut pas être résolu. Pour surmonter cette nouvelle contrainte, nous allons introduire une notion de « tolérance » faisant appel à une technique dite des variables ressort (slack variables).



### 2.7- Fonctions noyaux

Le classificateur à marge maximale que nous venons de présenter, permet d'obtenir de très bons résultats lorsque les données sont linéairement séparables. L'intérêt principal d'un classificateur de ce type réside dans le fait que l'on en contrôle facilement la capacité et donc le pouvoir de généralisation. L'idée retenue dans SVM va dans un autre sens ; on va tenter de trouver une transformation (mapping) de l'espace d'entrée vers un autre espace appelé « espace de redescription » (feature space) dans lequel les données sont linéairement séparables.

La dimension de l'espace des caractéristiques est généralement très élevée. Cela ne pose pas de problème pour notre classificateur à marge maximale vu que sa formulation duale ne le nombre de variables à déterminer en fonction de la taille de l'ensemble d'apprentissage.

### 2.7.1 Type des noyaux

1. Noyau linéaire : Un noyau linéaire peut être utilisé comme produit scalaire normal sous deux observations données. Le produit entre deux vecteurs est la somme de la multiplication de chaque paire de valeurs d'entrée.  $K(x, x_i) = \sum(x * x_i)$ .
2. Noyau RBF : Le noyau de la fonction radiale est une fonction du noyau couramment utilisée dans la classification des machines à vecteurs de support. RBF peut mapper un espace d'entrée dans un espace à dimensions infinies.  $K(x, x_i) = \exp(-\gamma * \sum((x - x_i)^2))$ .

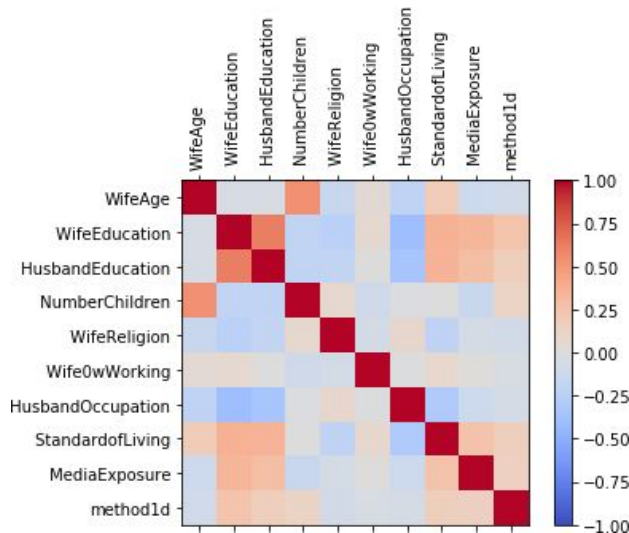
## 3. Préparation de données

Le jeu de données sur lequel nous avons appliqué les expérimentations contient 1473 observations sans aucun manque et il est décrit par 2 attributs continus et 8 discrets. Les attributs continus contiennent l'âge de la femme qui varie de 16 ans à 49 ans et nombre d'enfants que possède la femme qui aussi varie de 0 à 16 enfants.

Numero	Attribut	Type	Description
1	Wife age	continue	Tranche age d'une femme variant 16 à 49 ans
2	wife education	Discrete	Information sur education de la femme
3	Husband education	Discrete	Information sur l'education du mari
4	Number of children	Continue	Nombre d'enfant d'une femme
5	Wife religion	Discrete	la femme est musulmane ou non
6	Wife working	Discrete	Savoir si la femme travail ou pas
7	Husband occupation	Discrete	Type d'emploi du mari
8	Standard living	Discrete	Niveau de vie de la famille
9	Media exposure	Discrete	la femme aime si choix soit parler au media ou non
10	contraceptive methode	Discrete	attribut de prediction

### 3.1- Pré-traitement des données

Les bons résultats qu'un classificateur automatique peut fournir reposent, en grande partie, sur la phase de prétraitement. Les données issues d'un mauvais prétraitement vont mettre en péril la qualité du classificateur. Cette phase consiste en une succession de traitements sur les données brutes afin d'extraire de l'information et de ne garder que celle qui est utile à la classification. voici la figure ci-dessous qui montre la corrélation des nos variables :



### - Pourquoi nous avons utilisé ce schéma cette méthode de corrélation?

Dans notre graphique, lorsque la corrélation est égale à 0 ou proche de 0, la couleur est grise. Le rouge le plus foncé indique une corrélation positive parfaite, tandis que le bleu le plus sombre indique une corrélation négative parfaite.

Lors de l'évaluation de la corrélation entre toutes les entités, la méthode « corr () » inclut la corrélation de chaque entité avec elle-même, qui est toujours 1, c'est pourquoi ce type de graphique a toujours la diagonale rouge du haut à gauche. en bas à droite. Outre la diagonale, le reste des carrés montre une corrélation entre différentes caractéristiques, ce qui permet de constater facilement que les variables wifeAge, wifeworking, NumberChildrem sont corrélés.

le "corr()" est très facile à utiliser et très puissant pour les premières étapes de l'analyse des données (préparation des données), en faisant un graphique de ses résultats en utilisant matplotlib ou tout autre utilitaire de traçage en python, vous aurez une meilleure idée des données tant que vous puissiez prendre des décisions pour les prochaines étapes de la préparation et de l'analyse des données.

#### 3.1.1- Visualisation de donnée

Notre dataset n'a pas de données manquante , d'où ça nous épargnée de gérer les données manquante, La figure ci dessous le démontre :

Entrée [362]: `print(df)`

	Unnamed: 0	WifeAge	Wifelworking	Enfants	Methode
0	0	24	1	3	0
1	1	45	1	10	0
2	2	43	1	7	0
3	3	42	1	9	0
4	4	36	1	8	0
5	5	19	1	0	0
6	6	38	1	6	0
7	7	21	0	1	0
8	8	27	1	3	0
9	9	45	1	8	0
10	10	38	0	2	0
11	11	42	1	4	0
12	12	44	0	1	0
13	13	42	0	1	0
14	14	38	1	2	0
15	15	26	1	0	0
16	16	48	1	7	0
17	17	39	1	6	0
18	18	37	1	8	0
19	19	39	1	5	0
20	20	26	0	1	0
21	21	24	0	0	0
22	22	46	1	1	0

## 4. Construction et Évaluation du modèle

Notre modèle de prédiction est basé sur l'algorithme du SVM , cependant notre modèle est construit à partir d'une fonction noyau gaussienne , ayant pour paramètre  $\gamma = 0.1$  et pour paramètre  $c = 1000000$ . En effet , Le choix de ce modèle n'est pas aléatoire, il découle de la haute performance de cette fonction au regard des autres.

### 4.1 Sélection des paramètres

On va ici effectuer le choix des paramètres  $\gamma$  et  $C$  à optimiser conjointement

1.  $\gamma$  : Le  $\gamma$  est un paramètre du noyau RBF et peut être considéré comme la "propagation" du noyau et donc de la région de décision. Lorsque le  $\gamma$  est faible, la «courbe» de la limite de décision est très basse et la région de décision est donc très large. Lorsque le  $\gamma$  est élevé, la «courbe» de la limite de décision est élevée.
2.  $C$  : “C” est un paramètre de l'apprenant SVC et constitue la sanction pour la classification erronée d'un point de données. Lorsque  $C$  est petit, le classificateur accepte les points de données mal classés (biais élevé, faible variance). Lorsque “C” est grand, le classificateur est fortement

pénalisé pour les données mal classées et par conséquent, se penche en arrière pour éviter tout point de données mal classifié (faible biais, variance élevée).

```
Entrée [367]: from sklearn.svm import SVC
              svm = SVC(kernel='rbf', random_state = 0 ,gamma=.01, C=1000000)
              svm.fit(x_train, y_train)

Out[367]: SVC(C=1000000, cache_size=200, class_weight=None, coef0=0.0,
              decision_function_shape='ovr', degree=3, gamma=0.01, kernel='rbf',
              max_iter=-1, probability=False, random_state=0, shrinking=True, tol=0.001,
              verbose=False)
```

## 4.2 - Validation du modèle

Ce point s'applique également à toute autre méthode d'apprentissage supervisé. Il s'agit de l'étude de la performance du modèle construit, c'est-à-dire on cherche ici à estimer les erreurs de classification de notre modèle, donc, la probabilité qu'il prédit correctement la classe de

Ce point s'applique également à toute autre méthode d'apprentissage supervisé. Il s'agit de l'étude de la performance du modèle construit, c'est-à-dire on cherche ici à estimer les erreurs d'une donnée. Intuitivement, on peut réaliser que cette erreur correspond la probabilité de le classifieur faille dans la prédiction de la classe d'une nouvelle donnée.

## 4.3 Matrice de confusion

La matrice de confusion expose la qualité d'un modèle. Chacune de ses colonnes représente le nombre d'occurrences d'une classe prédite par le classifieur, tandis que chacune de ses lignes représente le nombre d'occurrences d'une classe cible (classe désirée ou classe de sortie attendue). Les données considérées peuvent être issues de l'ensemble d'apprentissage ou d'un ensemble de données réservées au test.

	VD Predite=0	VD Predite=1
VD Rel= 1	72	1 FP
VD Rel=1	3 FN	72

Voici la visualisation de notre matrice de confusion en python :

```

              precision    recall  f1-score   support

      0      0.96      0.99      0.97        73
      1      0.99      0.96      0.97        75

 accuracy      0.97        148
 macro avg      0.97      0.97      0.97        148
 weighted avg      0.97      0.97      0.97        148

[[72  1]
 [ 3 72]]
```

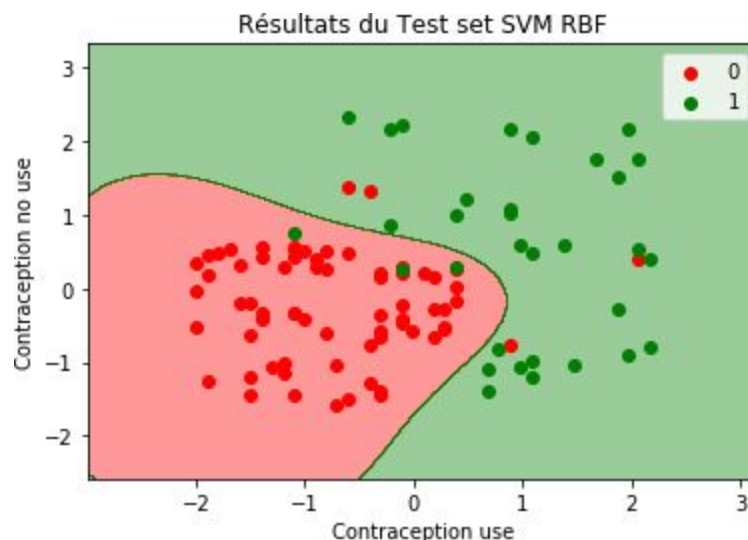
En effet notre modèle a été évalué sur nos données test qui est de 10 pourcent de notre dataset en occurrence 148 sur 1473, cependant nous avons eu - une (1) observation qui est False Positive car le modèle a prédit pour le choix d'utilisation de la méthode contraceptive et que la valeur réel l'observation est non a l'utilisation de la méthode contraceptive, - 3 observations False négative cad le modèle a prédit pour le non choix d'utilisation de la méthode contraceptive et que les valeurs de observations réelles est pour l'utilisation de la méthode contraceptive, -72 observations dont les valeurs de prédiction sont pour le non choix de la méthode de contraceptive et les valeurs réelles sont pour le non choix de la méthode contraceptive -72 observations dont les valeurs de prédiction sont pour le choix de la méthode de contraceptive et que les valeurs réelles sont pour le choix de la méthode de contraceptive.

#### 4.3.1 Calcul du raciaux

1. Accuracy rate(Taux d'observation) L'Accuracy est le nombre d'observation correcte prédit par le modele  $AR = 144/148 = 97.2$  pourcent.
2. Error rate( Taux d'observation incorrect) L'ER est le nombre d'observation pour laquelle le modele se tromper sur le nombre d'observation réel  $ER = 4/148 = 0.27$

#### 4.3.2 Visualisation du resultat

Voici la figure ci dessous qui montre comment les observations de notre modele se comporte par rapport a la methode du SVM.

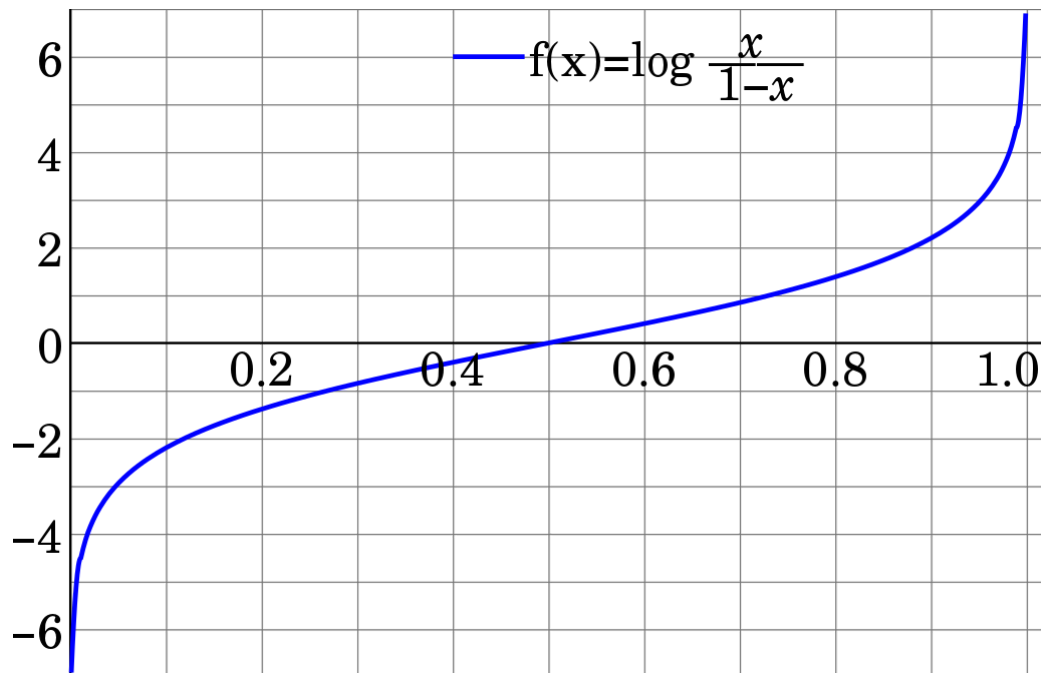


#### 4.4 Regression Logistique

La régression logistique ou modèle logit est un modèle de régression binomiale. Comme pour tous les modèles de régression binomiale, il s'agit de modéliser au mieux un modèle mathématique simple à des observations réelles nombreuses. La régression logistique constitue un cas particulier de modèle linéaire généralisé. Elle est largement utilisée en apprentissage automatique.



Voici la présentation de la courbe de la régression Logistique dans la gure ci-dessous :



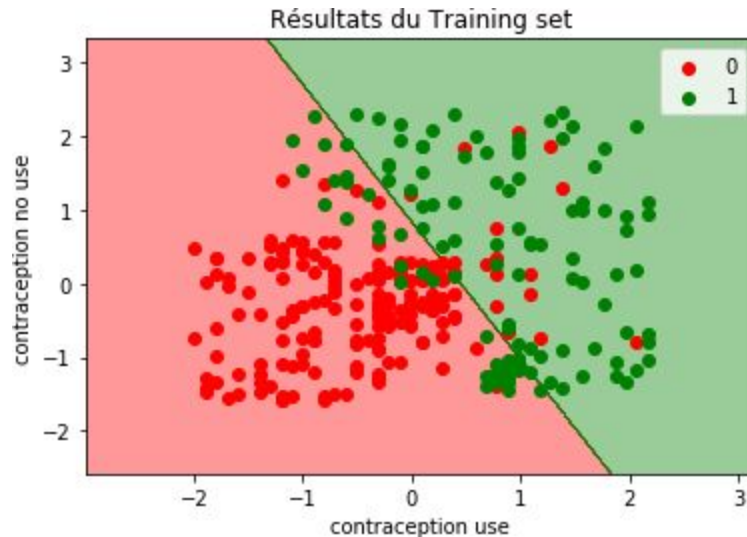
#### 4.4.1 Evaluation du modèle

En effet ,nous avons évalué le modèle sur base de la matrice de confusion , voici le résultat dans la figure ci-dessous :

	precision	recall	f1-score	support
0	0.70	0.64	0.67	73
1	0.68	0.73	0.71	75
accuracy			0.69	148
macro avg	0.69	0.69	0.69	148
weighted avg	0.69	0.69	0.69	148
[[47 26] [20 55]]				

#### 4.4.2 Visualisation du modèle

Nous avons utilisé la librairie matplotlib de python pour faire notre visualisation du modèle voici la gure ci-dessous qui le visualise :



#### 4.5 Visualisation Croiser

Nous avons essayé avec différent taille de test set , ainsi nous avons remarqué que lorsqu'on diminue la taille du test set, l'Accuracy rate cad le nombre des observations correct prédit par le modèle augmente , et Et l'erreur rate car le nombre d'observation pour laquelle le modele se tromper sur le nombre d'observation réel diminue.

En effet , voici la figure qui le démontre :

<u>Dataset</u> Taille	<u>SVM</u>	<u>Regression Logistique</u>
T = 10 %	Accuracy rate:97%, Erro rate: 0.7%	Accuracy: 70%, error:30%
T = 20 %	Accuracy rate:92%, Erro rate: 0.8%	Accuracy: 65%, error:35%
T = 30 %	Accuracy rate:90%, Erro rate: 10%	Accuracy: 62%, error:38%
T = 40 %	Accuracy rate:87%, Erro rate: 12%	Accuracy: 60%, error:40%
T = 50 %	Accuracy rate:85%, Erro rate: 15%	Accuracy: 59%, error:41%
Moyenne	Accuracy: 90%, error: 7.7%	Accuracy: 63.2%, error:36.8%



## 5. Interprétation de résultat

En effet voici l'interprétation de notre matrice de confusion :

	precision	recall	f1-score	support
0	0.96	0.99	0.97	73
1	0.99	0.96	0.97	75
accuracy			0.97	148
macro avg	0.97	0.97	0.97	148
weighted avg	0.97	0.97	0.97	148

[[72 1]
[ 3 72]]

Notre modèle a été évalué sur nos données test qui est de 10 pourcent de notre dataset en occurrence 148 sur 1473, cependant nous avons eu une (1) observation qui est False Positive car le modèle a prédit pour le choix d'utilisation de la méthode contraceptive et que la valeur réel l'observation est non à l'utilisation de la méthode contraceptive, 3 observations False négative cad le modèle a prédit pour le non choix d'utilisation de la méthode contraceptive et que les valeurs de observations réelles est pour l'utilisation de la méthode contraceptive, 72 observations dont les valeurs de prédiction sont pour le non choix de la méthode de contraceptive et les valeurs réelles sont pour le non choix de la méthode contraceptive 72 observations dont les valeurs de prédiction sont pour le choix de la méthode de contraceptive et que les valeurs réelles sont pour le choix de la méthode de contraceptive tandis que pour la méthode de regression logistique nous avons eu une 26 observations sont False Positive cad le modèle a prédit pour le choix d'utilisation de la méthode contraceptive et que la valeur réel l'observation est non a l'utilisation de la méthode contraceptive, 20 observations False négative cad le modèle a prédit pour le non choix d'utilisation de la méthode contraceptive et que les valeurs de observations réelles est pour l'utilisation de la méthode contraceptive, 47 observations dont les valeurs de prédiction sont pour le non choix de la méthode de contraceptive et les valeurs réelles sont pour le non choix de la méthode contraceptive 55 observations dont les valeurs de prédiction sont pour le choix de la méthode de contraceptive et que les valeurs réelles sont pour le choix de la méthode de contraceptive.

## 6. Conclusion

La réalisation d'un programme d'apprentissage par SVM se ramène à résoudre un problème d'optimisation impliquant un système de résolution dans un espace de dimension conséquente. L'utilisation de ces programmes revient surtout à sélectionner une bonne famille de fonctions noyau et à régler les paramètres de ces fonctions. Ces choix sont le plus souvent faits par une technique de validation croisée, dans laquelle on estime la performance du Système en la mesurant sur des exemples n'ayant pas été utilisés en cours d'apprentissage.

Cependant dans ce travail, nous avons construit un modèle basé sur la méthode de la Machine à vecteurs de support (SVM) pour la prédiction sur le choix de la méthode e la méthode contraceptive d'une femme en fonction de ses caractéristiques démographique et socio-économiques.