**cDAF+**

**Option : Système Intelligent et Multimédia**

Niveau : **Master I**

**Cours : Fouille de Données**

Professeur : NGUYEN Thà Minh Huyen

Titre du devoir

**Rapport du TP2 & TP3 du cours de Fouille de Données**

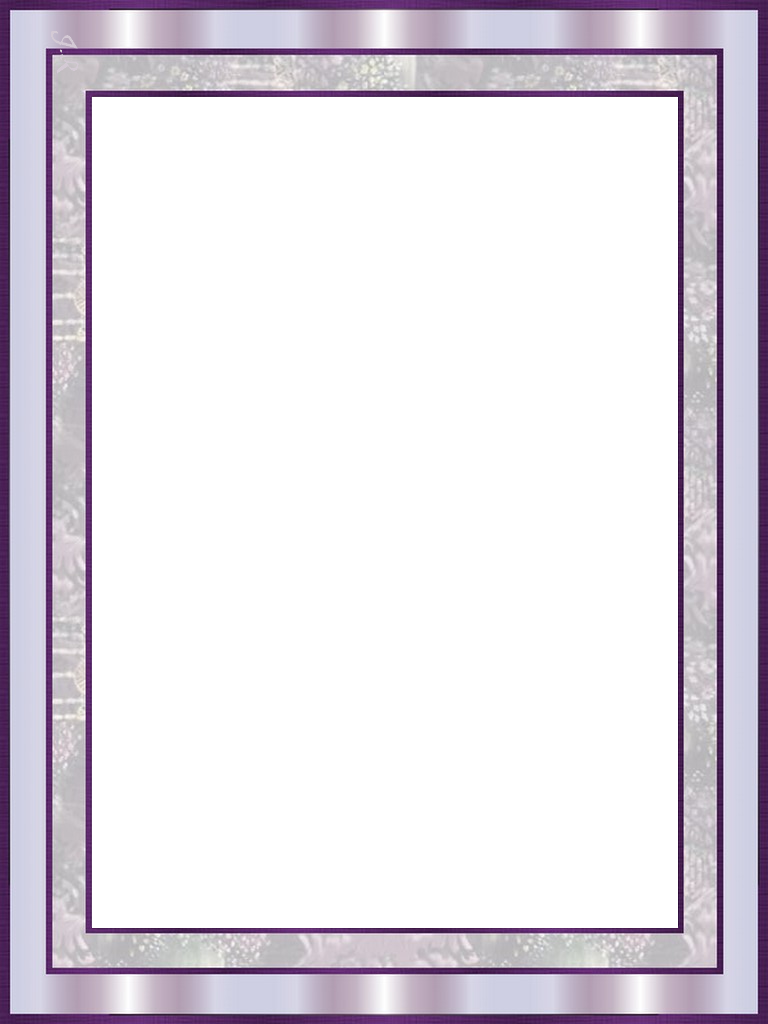
* **Analyse Factorielle**
* **Clustering**

**Préparé par le Binôme 13 :**

**Mongetro GOINT & Myderson SEMEURAND**

TECHNOPLUS-SYSTEM

Date: 19 Mars 2018



**1**

Universite nationale du vietnam (unv) / Institut Francophone pour l’Innovation (IFI)

**Table des matières**

[Tableaux et figures 2](#_Toc509307646)

[Analyses Factorielles 4](#_Toc509307647)

[1.- Application des méthodes factorielles à notre jeu de données “Credit Scoring” 4](#_Toc509307648)

[1.1.- ACP – Analyses en Composante Principale : 4](#_Toc509307649)

[2.- Analyse Factorielle des Correspondances 8](#_Toc509307650)

[2.1.- Les résultats numérique de l’Analyse Factorielle des Correspondances (AFC) 8](#_Toc509307651)

[2.2.- Représentation graphique de l’Analyse Factorielle des Correspondances (AFC) 9](#_Toc509307652)

[Clustering 11](#_Toc509307653)

[3.- Classifications 11](#_Toc509307654)

[3.1.- Répartition des individus 11](#_Toc509307655)

[3.2.- Part d’explication de la partition dans l’inertie totale 11](#_Toc509307656)

[3.4.- Groupe de caractérisation 12](#_Toc509307657)

[3.5.- La projection des individus suivant leur cluster dans le premier plan factoriel 12](#_Toc509307658)

[3.6.- HAC 13](#_Toc509307659)

[Conclusion **Error! Bookmark not defined.**](#_Toc509307660)

# Tableaux et figures

[Figure 1: Importance des composantes principales 5](#_Toc509307497)

[Figure 2 : Corrélation entres les variables et les axes principaux 5](#_Toc509307498)

[Figure 3 : Plan factoriel sur le premier axe 6](#_Toc509307499)

[Figure 4 : Plan factoriel du premier axe modifié 7](#_Toc509307500)

[Figure 5 : Cercle des corrélations entre les variables dans le premier plan factoriel 7](#_Toc509307501)

[Figure 6 : Test CHI 2 d’indépendance 8](#_Toc509307502)

[Figure 7 : Valeurs propres calculées 8](#_Toc509307503)

[Figure 8 : AFC dans le premier plan factoriel 9](#_Toc509307504)

[Figure 9 : Coordonnées factorielles des profils lignes 10](#_Toc509307505)

[Figure 10 : Coordonnées factorielles des profils colonnes. 10](#_Toc509307506)

[Figure 11 : Répartition des individus dans les clusters 11](#_Toc509307507)

[Figure 12 : Part d’explication de la partition dans l’inertie totale. 11](#_Toc509307508)

[Figure 13 : Coordonnées des centres de chacun des cinq clusters. 12](#_Toc509307509)

[Figure 14 : Caractérisation des clusters 12](#_Toc509307510)

[Figure 15 : Projection des individus par cluster 13](#_Toc509307511)

[Figure 16 : Sélection du nombre optimal de clusters 13](#_Toc509307512)

[Figure 17 : Résultat du Clustering 14](#_Toc509307513)

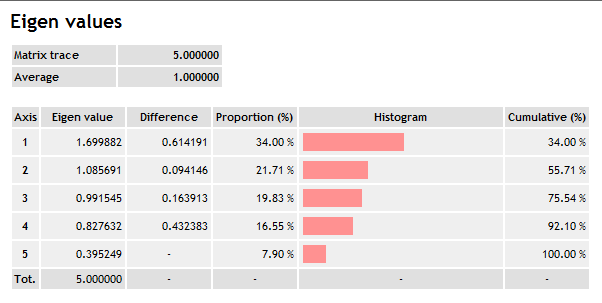
[Figure 18 : Coordonnées des centres des clusters 14](#_Toc509307514)

# Analyses Factorielles

## 1.- Application des méthodes factorielles à notre jeu de données “Credit Scoring”

### 1.1.- ACP – Analyses en Composante Principale :

Notre objectif en réalisant cette analyse en composante principale est de considérer l’ensemble des données contenant 24 observations et 5 variables continues afin de résumer l’information disponible à l’aide de variables synthétiques appelées composantes principales. Au moyen de la fonction « Principal Analysis » située sous l’onglet « Factoriel Analysis », nous réalisons l’analyse en composante principale. Ci-dessous, les résultats obtenus après cette opération :

[](#_Valeurs_propres)Tableau 1 : Valeurs propres

#### 1.1.1.- Valeurs propres

D’après le tableau ci-dessous [(voir le tableau 1)](#_Valeurs_propres), l’inertie décrite par le premier axe principal est ƛ1=1.70 et la part de celle décrite par ce même axe principal est ƛ1/p=0.34. Autrement dire, la première composante principale décrite à elle seule 34% (soit un peu moins de la moitié) de la variance totale.

En considérant le critère de Kaiser-Guttman proposé et utilisé par notre outil de travail [Tanagra](http://eric.univ-lyon2.fr/~ricco/tanagra/fr/tanagra.html) pour la définition des composantes principales les plus significatives, nous observons que deux des cinq valeurs propres sont importantes et pour la suite, ce sont les seules qui seront retenues. [(voir la figure 1)](#_Valeurs_propres)

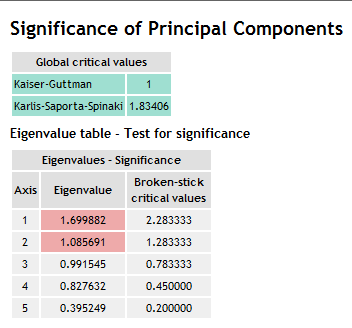
[](#_Valeurs_propres)

Figure 1: Importance des composantes principales

#### 1.1.2.- Corrélation entre les variables et les axes principaux

Cette partie des résultats [(voir la figure 2)](#_Corrélation_entre_les) indique la corrélation entre les variables et les axes principaux.

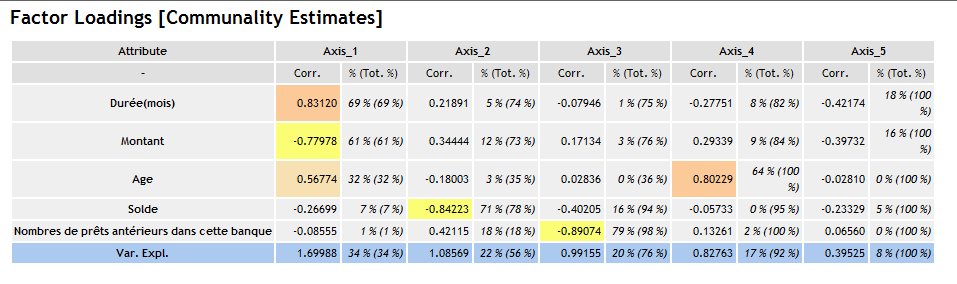


Figure 2 : Corrélation entres les variables et les axes principaux

A remarquer qu’ici, le premier axe est fortement corrélé positivement avec les variables : Durée(mois) et Age, et, négativement avec la variable : Montant. Ce qui indique qu’il prend en compte une bonne partie des données quantitatives de notre « Credit Scoring ». Mais cependant, l’évolution de la variable associée à cet axe est opposée à celle de la variable : Montant.

Le premier axe ainsi observé est déterminant dans l’identification d’un dossier « bon » ou « mauvais » concernant la fidélité des clients chez l’établissement de crédit.

Les trois autres axes suivants par contre, sont peu corrélé et le cinquième est quant à lui faiblement corrélé.

#### 1.1.3.- Plans factoriels

La popularité de l’ACP est en majeure partie liée à sa capacité de fournir des représentations graphiques, permettant d’apprécier par visualisation, les proximités entre les observations.

Dans notre cas, nous projetons les observations dans le premier plan factoriel où nous visons d’associer les identifiants aux points. Nous utilisons pour cela le composant SCATTERPLOT

WITH LABEL (onglet DATA VISUALIZATION) que nous plaçons en dessous de l’ACP. Nous le paramétrons de manière à avoir en abscisse le premier facteur, en ordonnée le second facteur.

Notons qu’il est très aisé de passer d’un plan factoriel à un autre.

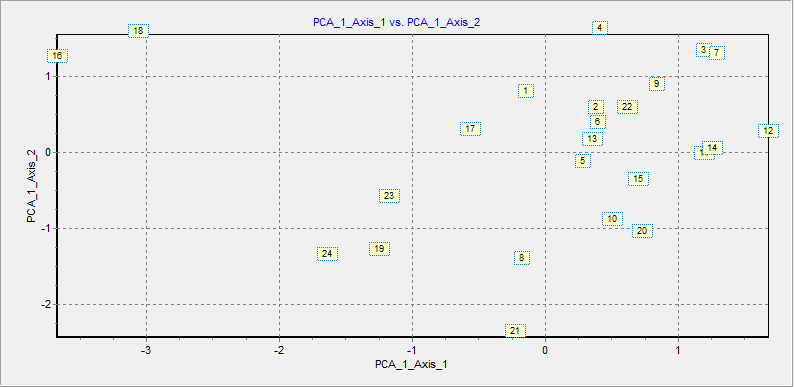
[](#_1.1.3.-_Plans_factoriels)

Figure 3 : Plan factoriel sur le premier axe

En activant l’option LEGEND / ATTRIBUTE VALUE, et en choisissant la référence « Dossier Prêt », nous obtenons la carte des points étiquetés par leurs identifiants. À noter que cette option est pratique que si les observations sont peu nombreuses. Mais, à mesure que les observations augmentent, le risque d’avoir un nombre de point illisible devient évidente.

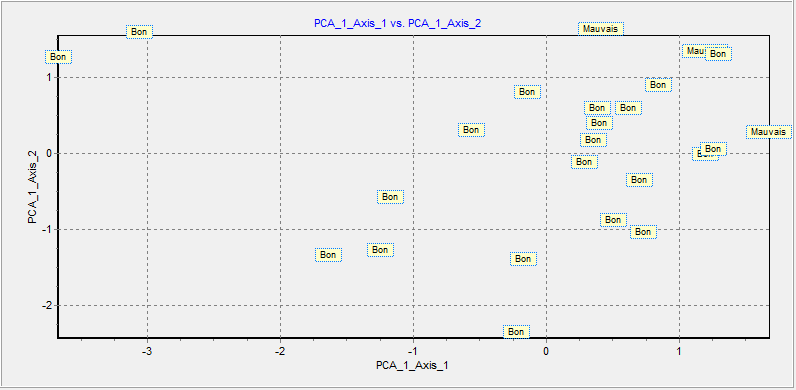
[](#_1.1.3.-_Plans_factoriels)

Figure 4 : Plan factoriel du premier axe modifié

#### 1.1.4.- Cercle des corrélations

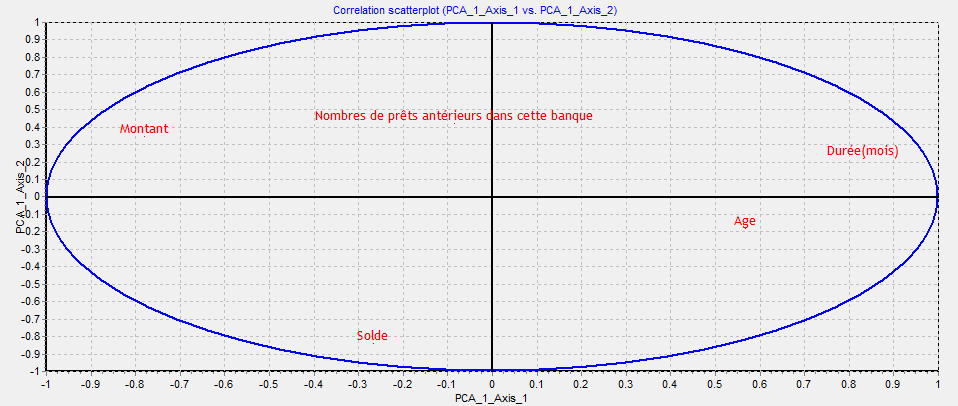


Figure 5 : Cercle des corrélations entre les variables dans le premier plan factoriel

Sur le cercle des corrélations ci-dessus, nous pouvons constater des variables comme (Durée et Age) sont très proches du cercle et du seconde axe (position des valeurs négatives), ce qu’en quelque sorte confirme leur forte corrélation positive avec la seconde composante principale. En opposition évidemment à la variable (Montant).

Quant à elle la variable (Nombre de prêts antérieurs dans cette banque), elle est plus proche du centre cercle du cercle et ne donne aucune information sur sa corrélation avec les deux premières composantes principales. Le cas de notre variable (Solde) situé proche du centre de notre cercle mais ne présentant aucune information sur sa corrélation avec les deux premières composantes principales.

## 2.- Analyse Factorielle des Correspondances

D’abord, rappelons qu’au niveau de la première partie de ce travail, nous avons réalisé une analyse factorielle des correspondances entre les attributs qualitatifs de notre « CreditScoring ». Ces relations entre les paires de variables qualitatives sont représentées dans un tableau de contingence répartis en plusieurs figures (soit Figure 15 à 27).

### 2.1.- Les résultats numérique de l’Analyse Factorielle des Correspondances (AFC)

Dans cette partie nous présentons les résultats numériques de l’ACP à l’aide du Test de CHI 2 et des variables propres, en nous basant sur les variables discrètes surtout.

#### 2.1.1.- Le Test de CHI 2

Avec les 14 variables qualitatives de notre « CreditScoring » les Test de CHI 2 sont nombreux, par contre ici nous ne présentons que l’un d’entre-deux.

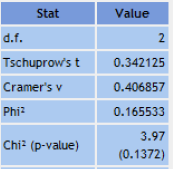


Figure 6 : Test CHI 2 d’indépendance

#### 2.1.2.- Les valeurs propres

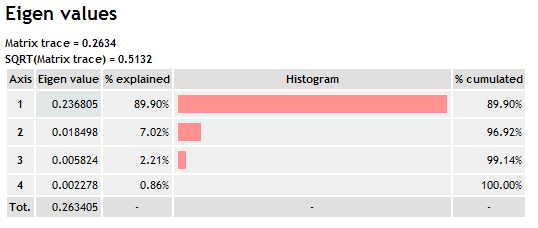


Figure 7 : Valeurs propres calculées

### 2.2.- Représentation graphique de l’Analyse Factorielle des Correspondances (AFC)

Nous ne devons pas oublier que notre outil Data Mining de travail « Tanagra » offre une fenêtre de visualisation où nous pouvons observer les graphiques suivant les différents plans factoriels à partir de l’onglet CHART ([voir la figure 8](#_2.2.2.-_Représentation_graphique)).

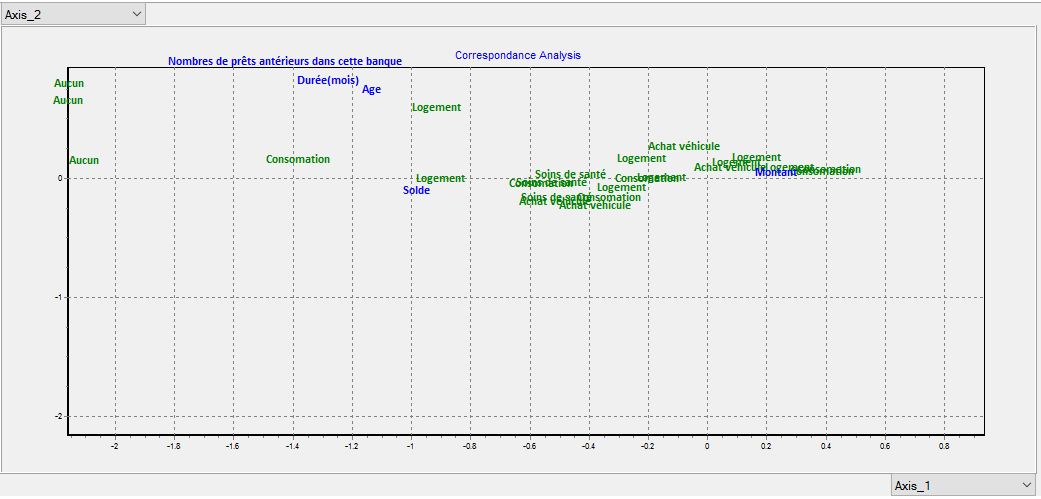


Figure 8 : AFC dans le premier plan factoriel

#### 2.2.3.- Coordonnées factorielles et interprétations des axes par les points individus :

Les points individus dans notre cas sont les différents Objectif de prêt (Aucun, Consommation, Logement, Achat de Véhicules et Soins de santé).

* Coordonnées factorielles

Le tableau de la figure suivantes est une démonstration du résultat des coordonnées factorielles obtenues pour chaque point individu ([voir la figure 9](#_2.2.3.-_Coordonnées_factorielles)).

**Interprétation :**

L’axe 1 présente les crédits à la consommation comme étant dominante tandis que l’axe 2 met accent sur les crédits pour achat de véhicules. Au niveau des axes 3 et 4 la remarque à faire n’est autre que quasiment le même degré d’importance de tous les points individus.

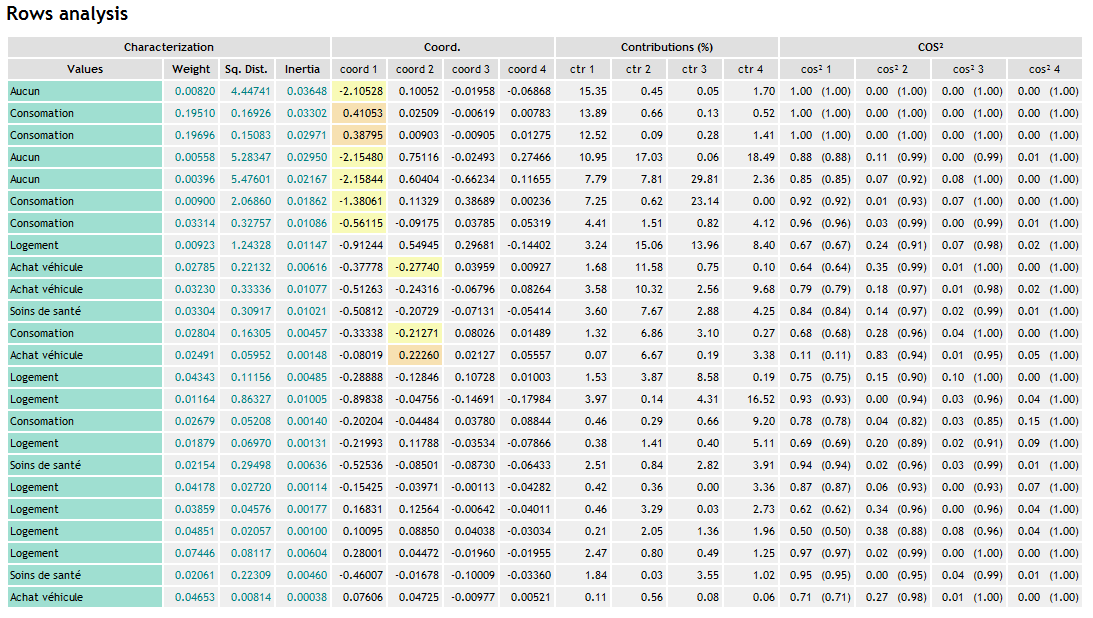


Figure 9 : Coordonnées factorielles des profils lignes

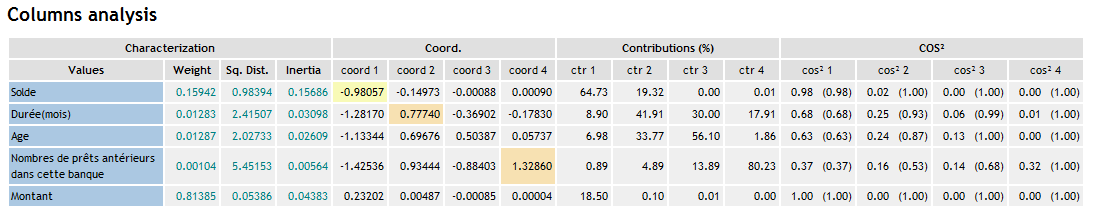


Figure 10 : Coordonnées factorielles des profils colonnes.

**Interprétation :**

Sur l’axe 1, le solde est le point individu à remarquer tandis que l’axe 2, et 4 sont respectivement dominés par la durée du prêt, le nombre de prêts antérieurs d’un de cette banque.

# Clustering

## 3.- Classifications

Pour illustrer la méthode **K-Means**, nous choisissons de classifier les objectifs suivants la similarité de leurs individus (valeurs des variables quantitatives). Pour cela, nous utilisons le composant **K-Means** de l’onglet **Clustering** de Tanagra avec la variable Objectif en Target et les cinq variables continues en Input. Les paramètres utilisés sont :

* 5 Clusters
* 10 Itérations au maximum
* 5 Essais

### 3.1.- Répartition des individus

Nous utilisons tous les variables Continues de notre « CreditScoring ».

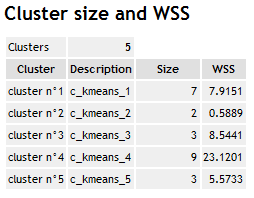


Figure 11 : Répartition des individus dans les clusters

Remarque pertinente : Le cluster No 2 est celui contenant le moins d’individu soit 2 au total, tandis que le No 4 est celui du plus grand nombre, soit 9 au total.

### 3.2.- Part d’explication de la partition dans l’inertie totale

D’après la figure suivante, la partition explique 61.88% de l’inertie totale.

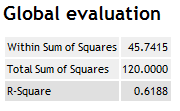


Figure 12 : Part d’explication de la partition dans l’inertie totale.

#### 3.3.- Coordonnées des centres de chacun des cinq clusters sont présentés ci-dessous :

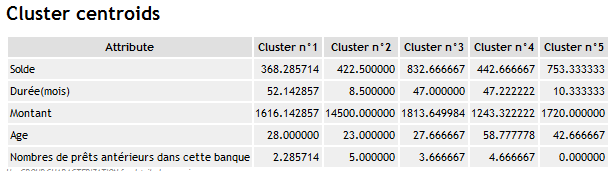


Figure 13 : Coordonnées des centres de chacun des cinq clusters.

### 3.4.- Groupe de caractérisation

Pour être explicite, nous présentons la répartition des individus au sein des différents groupes (clusters) à partir d’une caractérisation de ces derniers en utilisant l’outil **Group Characterization** de **Tanagra.**

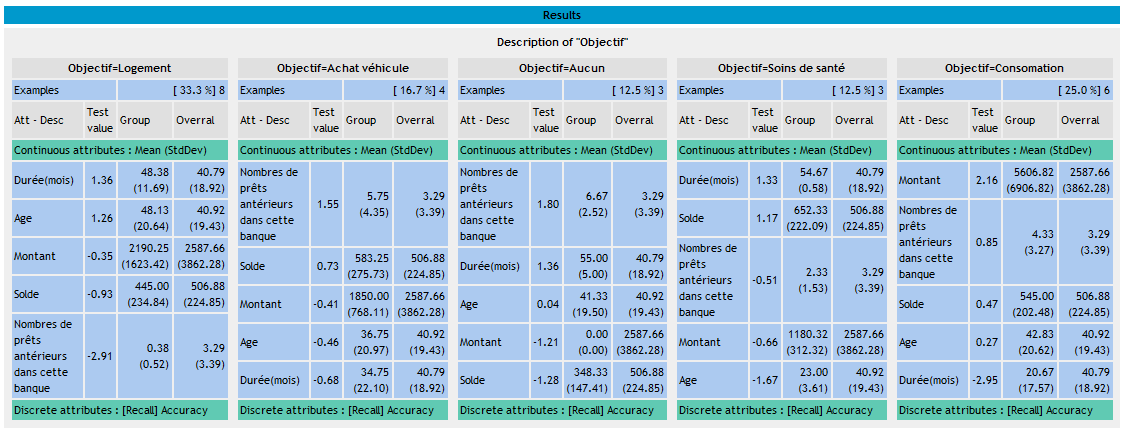


Figure 14 : Caractérisation des clusters

### 3.5.- La projection des individus suivant leur cluster dans le premier plan factoriel

Elle est décrite dans la figure suivante :

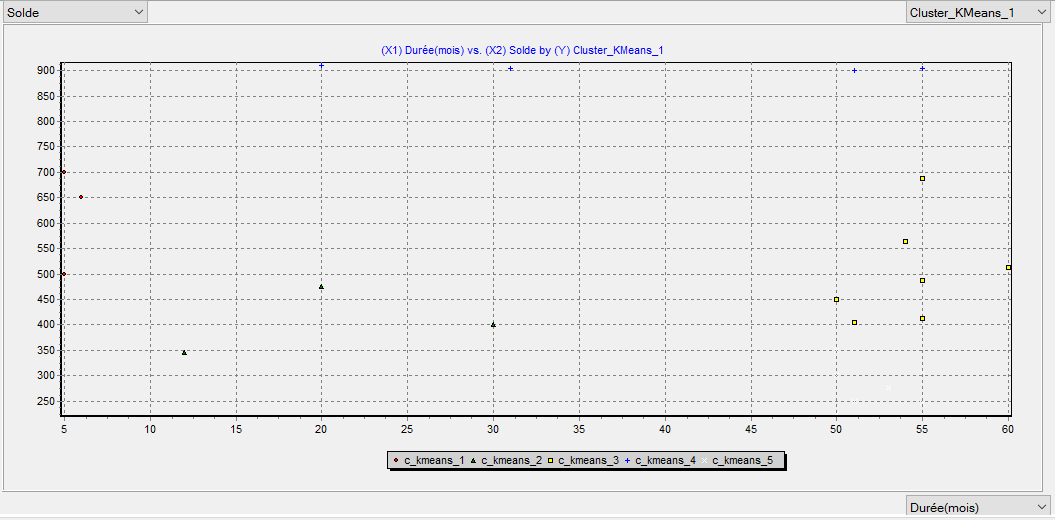


Figure 15 : Projection des individus par cluster

### 3.6.- HAC

Pour illustrer la classification hiérarchique, nous utilisons toutes nos variables Continues et une variable discrète (Objet). Nous appliquons l’outil HAC de l’onglet Clustering de Tanagra.

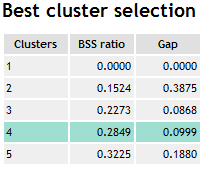


Figure 16 : Sélection du nombre optimal de clusters

* Les quatre clusters obtenus après réorganisation des données sont présentés à la figure 17 et les coordonnées de leurs centres à la figure 18.

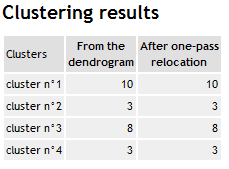


Figure 17 : Résultat du Clustering

**Interprétation :**

Dans la figure 17 où nous apparaissent les quatre clusters, nous pouvons remarquer que les valeurs restent même au départ et après relocation. En plus, le cluster no1 est celui ayant la plus grande valeur.

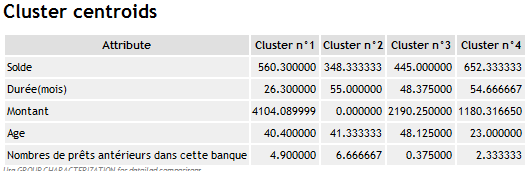


Figure 18 : Coordonnées des centres des clusters

**Interprétation :**

Après avoir représenté les clusters, il est à remarquer que dans la figure 18 où représentent les coordonnées de leurs centres, l’attribut **Montant** est celui ayant la plus grande valeur dans tous les clusters, sauf dans le no 2. Alors que l’attribut **Nombre de prêts antérieurs dans cette banque** est celui ayant la plus petite valeur dans tous les clusters, sauf dans le no 2.