

M2 : EES – Etudes Economiques et Statistiques **Data Science Exam**

Année universitaire 2024-2025, Durée 1h 30

Partie 1 : Questions à Choix Multiple – Test d'évaluation des connaissances

Barème partie 1 : (24 questions - 6 points : Bonne réponse = + 0.25 pt, Pas de réponse = 0 pt, Mauvaise réponse = - 0.25 pt)

Répondre par **Vrai** ou **Faux** (entourer la bonne réponse) - Chaque affirmation peut être Vraie ou Fausse. Elle est considérée comme vraie si elle est TOUJOURS vraie. Par contre si l'affirmation est absurde ou si elle n'a pas de sens, elle est considérée comme fausse. Il peut y avoir plusieurs affirmations vraies ainsi que plusieurs affirmations fausses dans chaque groupe d'affirmations.

1) Analyse en Composantes Principales (ACP) normée d'un tableau de données X à n individus et p variables :

- | | |
|--|-------|
| a) l'ACP normée consiste à diagonaliser la matrice des corrélations. | V F |
| b) la somme des valeurs propres correspond à l'information totale à résumer. | V F |
| c) les coordonnées des points variables sur le plan factoriel sont les corrélations des variables avec les facteurs. | V F |
| d) les variables à synthétiser ont des unités différentes ou des variances très différentes. | V F |

2) Analyse Factorielle des Correspondances (AFC) - Test d'indépendance du chi-deux sur une table de contingence croisant deux variables qualitatives respectivement à p et à q modalités :

- | | |
|--|-------|
| a) l'AFC consiste à localiser les éventuelles liaisons entre les modalités des deux variables. | V F |
| b) le nombre de composantes principales de l'AFC est égal à $\min(p ; q)$. | V F |
| c) l'AFC est utilisée lorsque l'hypothèse nulle H_0 n'est pas rejetée. | V F |
| d) plus la p -value est petite plus les variables sont indépendantes. | V F |

3) Analyse des correspondances Multiples (ACM) de p variables qualitatives totalisant q modalités :

- | | |
|--|-------|
| a) les valeurs propres sont toutes inférieures à 1. | V F |
| b) la représentation graphique des variables est un cercle de corrélation. | V F |
| c) le tableau de Burt est la juxtaposition de $p(p-1) / 2$ tables de contingences différentes. | V F |
| d) les facteurs ou composantes principales de l'ACM sont des variables nominales. | V F |

4) Analyse Factrielle Discriminante (AFD) à p variables explicatives et une variable à expliquer à q modalités ($p > q$) :

- | | |
|---|-------|
| a) l'AFD consiste à minimiser l'inertie ou l'information Inter-classes . | V F |
| b) le nombre maximum de facteurs discriminants est égal à $(q - 1)(p - 1)$. | V F |
| c) l'AFD est une technique de classification supervisée. | V F |
| d) cette méthode ne s'applique pas lorsque les variables explicatives sont nominales. | V F |

5) Classifications : Nuées dynamiques et Classification Hiérarchique Ascendante (CHA)

- | | |
|--|-------|
| a) la méthode des nuées dynamiques est une classification hiérarchique. | V F |
| b) un individus peut appartenir à deux classes de la classification. | V F |
| c) ces techniques sont des classifications dites non supervisées. | V F |
| d) elles consistent à faire des regroupements d'individus qui se ressemblent selon leurs variables de description. | V F |

6) Modélisation

- | | |
|--|-------|
| a) l'ANOVA à 1 facteur contrôlé consiste à comparer les variances empiriques de la variable à expliquer observées sur les différents niveaux du facteur. | V F |
| b) les ANOVA sont des techniques statistiques de tests de Fisher pour comparer les variances. | V F |
| c) l'ANCOVA est une technique statistique de tests et d'estimation qui permet d'analyser l'effet de variables explicatives mixtes sur une variable continue à expliquer. | V F |
| d) En ANOVA, plus la p -value est grande, plus le modèle est significatif. | V F |

Compléter, signer et retourner ce QCM avec votre copie d'examen.

Nom :	Prénom :	Signature :
-------	----------	-------------

Partie 2 : Analyses

Barème partie 2 : (Chaque question est notée sur 2.8 points = 14 points)

Recommandations : Soigner la rédaction : donner une interprétation résumée à l'essentiel, c'est-à-dire en évitant de reprendre les valeurs des tableaux des résultats obtenus. Aucun document n'est autorisé.

Base de données : *credit.sba*

Echantillon de 468 clients d'une agence bancaire. Deux principaux thèmes sont distingués totalisant 14 variables. Le premier se réfère à la signalétique du client de l'agence et le second porte sur les caractéristiques de son compte bancaire.

L'objectif est d'étudier selon certains critères (caractéristiques des comptes bancaires) comment cette agence "discrimine" ses clients. On veut également définir des profils types de clients afin de pouvoir élaborer des politiques différenciées pour chacun d'eux.

Dictionnaire des variables

1 . Type de client (appréciation de l'agence) (2 MODALITES)

BON - bon client MAUV - mauvais client

2 . Domiciliation du salaire du client à l'agence (2 MODALITES)

OUI - salaire domicilié NON – salaire non domicile

3 . Autorisation de découvert (2 MODALITES)

OUI - découvert autorisé NON - découvert interdit

4 . Interdiction de chéquier (2 MODALITES)

OUI - chéquier interdit NON - chéquier autorisé

Thème 1 - Signalétique :

5 . Genre (2 MODALITES)

FEMM - Femme HOMM - Homme

6 . Age du client (4 MODALITES)

AGE1 - moins de 23 ans AGE2 - de 23 à 40 ans

AGE3 - de 40 à 50 ans AGE4 - plus de 50 ans

7 . Situation familiale (4 MODALITES)

CELB - célibataire MARI – marié(e)

DIVO – divorcé(e) VEUF – veuf(ve)

8 . Niveau d'éducation - Diplôme (4 MODALITES)

PDIPL – Pas de diplôme DSEC – Diplôme d'études secondaires

NUNI – Niveau universitaire intermédiaire BPLU - Bac+3 et plus

9 . Profession (3 MODALITES)

CADR - cadre EMPL – employé(e) INTER – profession intermédiaire

Thème 2 - Caractéristiques du compte bancaire client :

10 . Montant de l'épargne (CONTINUE)

11 . Nombre d'années d'ancienneté client (CONTINUE)

12 . Encours moyens du compte bancaire (CONTINUE)

13 . Mouvements bancaires moyens (CONTINUE)

14 . Cumul moyen des débits bancaires (CONTINUE)

M2 : EES – Etudes Economiques et Statistiques
Data Science Exam
Année universitaire 2024-2025, Durée 1h 30

Question 1 : Analyse de la dépendance

- a) Compléter puis interpréter les 3 valeurs manquantes (A, B et C) du tableau de contingence croisant le type et le genre du client de cette agence bancaire et la valeur D de l'AFC. (cf. annexe 1).
- b) Peut-on conclure, avec un risque d'erreur $\alpha = 5\%$, qu'il y a un lien entre ces 2 variables ? Justifiez votre réponse.

Question 2 : Typologie des clients selon leur signalétique

Une analyse factorielle puis une méthode de classification ont été effectuées pour partitionner les clients de cette agence bancaire en trois classes homogènes selon leur signalétique « thème 1 ». Toutes les autres variables ont été positionnées en éléments illustratifs (cf. annexe 2).

Donner une interprétation résumée (interne et externe) de chacune des trois classes de cette classification.

Question 3 : Description des caractéristiques des comptes bancaires des clients

Une analyse factorielle a été effectuée pour synthétiser les caractéristiques bancaires des clients de cette agence bancaire « thème 2 ». Toutes les autres variables ont été positionnées en éléments illustratifs.

Donner une interprétation résumée, interne (variables actives) et externe (variables illustratives) des résultats obtenus sur le premier plan factoriel des graphiques 1 et 2 (cf. annexe 3).

Question 4 : Modèle décisionnel

Une analyse a été effectuée pour discriminer le type de client selon les caractéristiques de son compte bancaire.

Résumer les principaux résultats de cette analyse : qualité du modèle, tableau de classement puis déterminer et interpréter les caractéristiques bancaires qui discriminent significativement et séparent au mieux les deux classes de clients (Bon / Mauvais) de cette agence. (cf. annexe 4).

Question 5 : Modèle prédictif

Une analyse a été effectuée pour analyser l'effet des variables classes d'âges et genre sur le montant de l'épargne du client. Résumer les principaux résultats de ce modèle complet (Les conditions d'application de ce modèle ont été vérifiées) (cf. annexe 5).

M2 : EES – Etudes Economiques et Statistiques
Data Science Exam
Année universitaire 2024-2025, Durée 1h 30

Annexes – Resultats

Annexe 1 : analyse de la dépendance - Type de client & Genre

Effectif			
% en ligne			
% en colonne			
Bon	A	104	237
	B	43,9	100,0
	C	45,2	50,6
Mauvais	105	126	231
	45,5	54,5	100,0
	44,1	54,8	49,4
Ensemble	238	230	468
	50,9	49,1	100,0
	100,0	100,0	100,0

Nombre de cases d'effectif théorique < 5	0
Nombre de degrés de liberté	1
Khi-2 (k)	5,322
P-value	0,021

AFC - Information à résumer - Somme des valeurs propres = D

Annexe 2 : Typologie des clients selon leur signalétique

Classe 1/3 (Effectif : 165 - Pourcentage : 35,26%)

Libellé de la variable	Modalité caractéristique	% de la modalité dans la classe	% de la modalité dans l'échantillon	Valeur-Test	Probabilité
Classes âges des clients	<= 40ans	93,94	50,85	99,990	0,000
Situation familiale	célibataire	82,42	35,47	99,990	0,000
Niveau Education - Diplôme	Universitaire-Intermédiaire	11,52	6,20	3,237	0,001
Niveau Education - Diplôme	Pas-de-diplôme	9,70	5,13	3,010	0,001
Niveau Education - Diplôme	Etudes-Secondaires	77,58	69,66	2,676	0,004
Profession	autre prof.	40,61	32,69	2,575	0,005
Autorisation de découvert	Découvert autorisé	50,91	43,16	2,396	0,008
Type de client	Mauvais client	56,97	49,36	2,335	0,010
Profession	employé	58,18	50,64	2,314	0,010
Interdiction de chéquier	Chéquier interdit	15,76	11,32	2,052	0,020
Interdiction de chéquier	Chéquier autorisé	84,24	88,68	-2,052	0,020
Type de client	Bon client	43,03	50,64	-2,335	0,010
Autorisation de découvert	Découvert interdit	49,09	56,84	-2,396	0,008
Classes âges des clients	> 50ans	6,06	23,08	-6,828	0,000
Profession	cadre	1,21	16,67	-7,436	0,000
Niveau Education - Diplôme	BAC+3-Plus	1,21	19,02	-8,161	0,000
Classes âges des clients	40-50ans	0,00	26,07	-11,030	0,000
Situation familiale	marié	4,85	47,86	-99,990	0,000

**N.B. : les variables illustratives sont représentées en caractères gras et italique.*

Variable caractéristique	Moyenne dans la classe	Moyenne générale	Écart-type (N) dans la classe	Écart-type (N) général	Valeur-Test	Probabilité
Encours moyens	699,152	759,177	296,349	381,244	-2,508	0,006
Montant de Epargne	439,848	1040,788	1418,949	2884,767	-3,318	0,000
Années Ancienneté	4,717	6,174	3,860	5,354	-4,335	0,000
Mouvements bancaires moyens	4063,303	5277,009	3369,823	3950,827	-4,894	0,000

Classe 2/3 (Effectif : 200 - Pourcentage : 42,74%)

Libellé de la variable	Modalité caractéristique	% de la modalité dans la classe	% de la modalité dans l'échantillon	Valeur-Test	Probabilité
Situation familiale	marié	84,00	47,86	99,990	0,000
Classes âges des clients	40-50ans	53,00	26,07	11,665	0,000
Niveau Education - Diplôme	Etudes-Secondaires	86,00	69,66	6,736	0,000
Profession	employé	60,50	50,64	3,601	0,000
Niveau Education - Diplôme	Universitaire-Intermédiaire	3,50	6,20	-1,933	0,027
Niveau Education - Diplôme	Pas-de-diplôme	2,00	5,13	-2,537	0,006
Situation familiale	veuf	0,50	4,06	-3,426	0,000
Situation familiale	divorcé	6,00	12,61	-3,701	0,000
Classes âges des clients	> 50ans	13,50	23,08	-4,223	0,000
Niveau Education - Diplôme	BAC+3-Plus	8,50	19,02	-5,067	0,000
Classes âges des clients	<= 40ans	33,50	50,85	-6,442	0,000
Profession	cadre	3,50	16,67	-6,981	0,000
Situation familiale	célibataire	9,50	35,47	-10,563	0,000

**N.B. : les variables illustratives sont représentées en caractères gras et italique.*

Variable caractéristique	Moyenne dans la classe	Moyenne générale	Écart-type (N) dans la classe	Écart-type (N) général	Valeur-Test	Probabilité

Classe 3/3 (Effectif : 103 - Pourcentage : 22,01%)

Libellé de la variable	Modalité caractéristique	% de la modalité dans la classe	% de la modalité dans l'échantillon	Valeur-Test	Probabilité
Niveau Education - Diplôme	BAC+3-Plus	67,96	19,02	99,990	0,000
Profession	cadre	66,99	16,67	99,990	0,000
Classes âges des clients	> 50ans	68,93	23,08	11,635	0,000
Situation familiale	veuf	14,56	4,06	5,195	0,000
Situation familiale	divorcé	28,16	12,61	4,844	0,000
Type de client	Bon client	60,19	50,64	2,089	0,018
Interdiction de chéquier	Chéquier autorisé	94,17	88,68	1,897	0,029
Interdiction de chéquier	Chéquier interdit	5,83	11,32	-1,897	0,029
Type de client	Mauvais client	39,81	49,36	-2,089	0,018
Classes âges des clients	40-50ans	15,53	26,07	-2,719	0,003
Profession	autre prof.	13,59	32,69	-4,810	0,000
Situation familiale	célibataire	10,68	35,47	-6,252	0,000
Profession	employé	19,42	50,64	-7,252	0,000
Classes âges des clients	<= 40ans	15,53	50,85	-8,292	0,000
Niveau Education - Diplôme	Etudes-Secondaires	25,24	69,66	-10,655	0,000

**N.B. : les variables illustratives sont représentées en caractères gras et italique.*

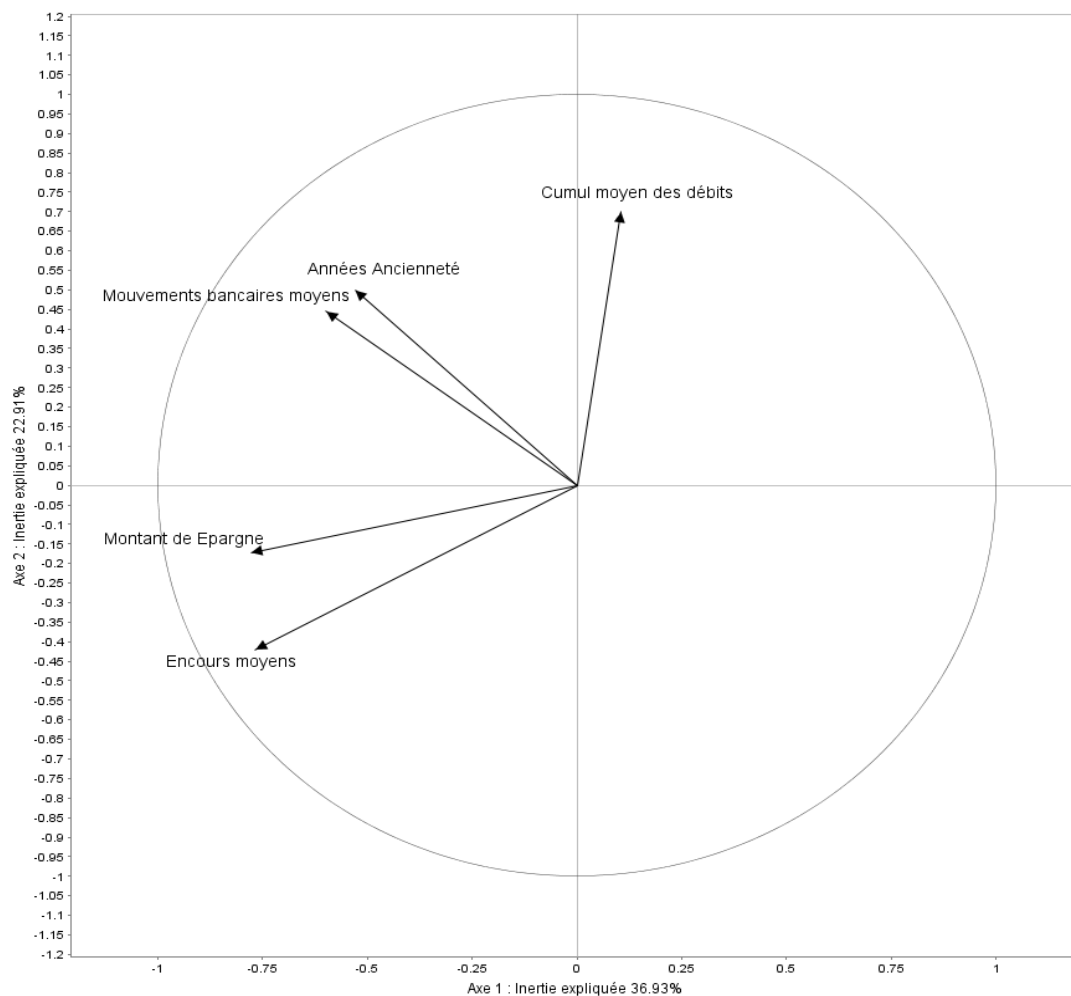
Variable caractéristique	Moyenne dans la classe	Moyenne générale	Écart-type (N) dans la classe	Écart-type (N) général	Valeur-Test	Probabilité
Années Ancienneté	8,437	6,174	6,033	5,354	4,847	0,000
Mouvements bancaires moyens	6685,631	5277,009	4320,558	3950,827	4,089	0,000
Encours moyens	854,126	759,177	470,645	381,244	2,856	0,002
Montant de Epargne	1749,893	1040,788	3759,092	2884,767	2,819	0,002

Annexe 3 : Description des caractéristiques bancaires

Axe	Variance de l'axe (valeur propre)	% de variance expliquée	% de variance expliquée cumulé
1	1,846	36,93	36,93
2	1,145	22,91	59,84
3	0,940	18,79	78,63
4	0,673	13,47	92,10
5	0,395	7,90	100,00
Total	5,000	100,00	100,00

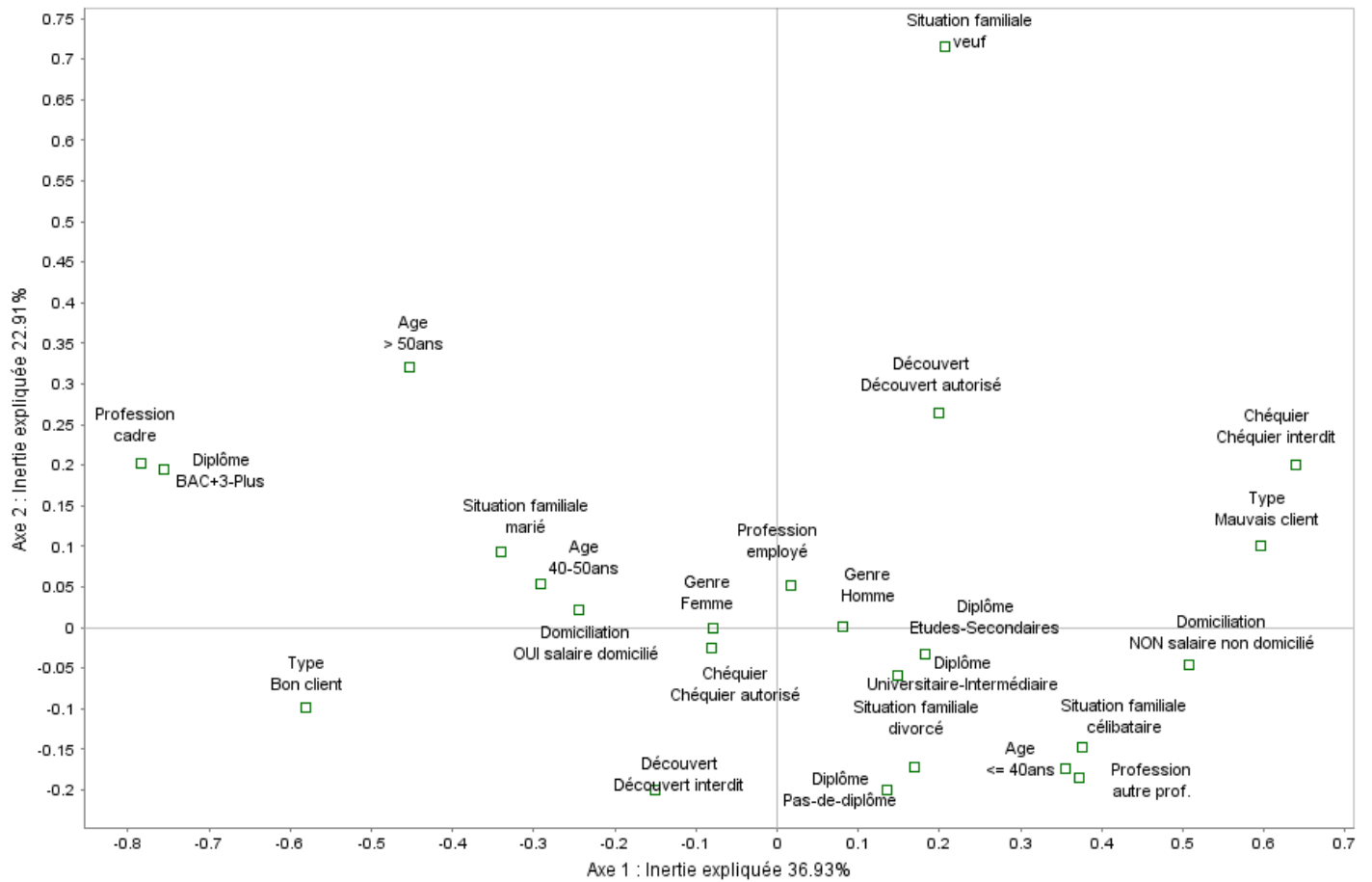
Libellé de la variable	Coordonnées des variables actives		Corrélations Variables actives - Facteurs	
	Axe 1	Axe 2	Axe 1	Axe 2
Montant de Epargne	-0,779	-0,172	-0,779	-0,172
Années Ancienneté	-0,530	0,500	-0,530	0,500
Encours moyens	-0,768	-0,421	-0,768	-0,421
Mouvements bancaires moyens	-0,599	0,445	-0,599	0,445
Cumul moyen des débits	0,106	0,700	0,106	0,700

Représentation des variables actives



				Coordonnées des modalités		Valeurs-tests des modalités	
Libellé	Effectif	Poids	Distance à l'origine	Axe 1	Axe 2	Axe 1	Axe 2
Type de client							
Bon client	237	237	0,644	-0,581	-0,098	-9,358	-2,012
Mauvais client	231	231	0,661	0,596	0,101	9,358	2,012
Domiciliation du salaire							
Non salaire non domicilié	152	152	0,581	0,508	-0,046	5,606	-0,645
OUI salaire domicilié	316	316	0,280	-0,244	0,022	-5,606	0,645
Autorisation de découvert							
Découvert autorisé	202	202	0,391	0,199	0,265	2,760	4,659
Découvert interdit	266	266	0,297	-0,151	-0,201	-2,760	-4,659
Interdiction de chéquier							
Chéquier autorisé	415	415	0,091	-0,082	-0,026	-3,629	-1,444
Chéquier interdit	53	53	0,712	0,639	0,200	3,629	1,444
Genre							
Femme	238	238	0,095	-0,079	-0,001	-1,277	-0,027
Homme	230	230	0,098	0,082	0,001	1,277	0,027
Classes âges des clients							
<= 40ans	238	238	0,406	0,355	-0,173	5,748	-3,552
> 50ans	108	108	0,654	-0,453	0,320	-3,946	3,539
40-50ans	122	122	0,331	-0,292	0,054	-2,759	0,649
Situation familiale							
célibataire	166	166	0,431	0,376	-0,148	4,436	-2,213
divorcé	59	59	0,262	0,170	-0,171	1,025	-1,313
marié	224	224	0,366	-0,341	0,094	-5,196	1,818
veuf	19	19	0,764	0,206	0,716	0,674	2,972
Niveau Education - Diplôme							
BAC+3-Plus	89	89	0,794	-0,755	0,195	-5,817	1,912
Etudes-Secondaires	326	326	0,188	0,183	-0,033	4,405	-1,022
Pas-de-diplôme	24	24	0,541	0,136	-0,201	0,502	-0,942
Universitaire-Intermédiaire	29	29	0,235	0,149	-0,058	0,609	-0,302
Profession							
autre prof.	153	153	0,438	0,373	-0,184	4,134	-2,594
cadre	78	78	0,828	-0,783	0,202	-5,570	1,825
employé	237	237	0,079	0,017	0,052	0,273	1,073

Représentation des modalités des variables illustratives



Annexe 4 : Modèle décisionnel

TABEAU DE CLASSEMENT

GROUPES D'ORIGINE	POURCENTAGES DES CLASSEMENTS		TOTAL
	BIEN CLASSES	MAL CLASSES	
Bon client	165.00 (69.62)	72.00 (30.38)	237.00 (100.00)
Mauvais client	179.00 (77.49)	52.00 (22.51)	231.00 (100.00)
TOTAL	344.00 (73.50)	124.00 (26.50)	468.00 (100.00)

FONCTION LINEAIRE DISCRIMINANTE

VARIABLES	CORRELATIONS	COEFFICIENTS		ECARTS	RATIO
.....	VARIABLES	FONCTION	REGRESSION	TYPES	COEF/ECT
NUM LIBELLES	AVEC F.L.D.	DISC.		(RES. TYPE REG.)	
	(SEUIL= 0.09)				
10 Montant de Epargne	0.243	0.0001	0.0000	0.0000	1.80
11 Années Ancienneté	0.309	0.1137	0.0416	0.0079	5.24
12 Encours moyens	0.322	0.0011	0.0004	0.0001	3.05
13 Mouvements bancaires	0.283	0.0001	0.0000	0.0000	3.85
14 Cumul moyen des débits	-0.300	-0.0179	-0.0065	0.0009	-6.94
CONSTANTE		-0.957863	-0.354166	0.1378	-3.31
D2 = 1.47956	T2 = 173.07988	PROBA = 0.001			

Annexe 5 : Modèle prédictif

MODALITES				
IDENT	LIBELLE	EFF.	POIDS	HISTOGRAMME DES POIDS RELATIFS
-----+-----+-----+-----+-----				
5 . Genre				
m1	- Femme	238	238.00	*****
m2	- Homme	230	230.00	*****
-----+-----+-----+-----+-----				
6 . Classes âges des clients				
m1	- <= 40ans	238	238.00	*****
m2	- > 50ans	108	108.00	*****
m3	- 40-50ans	122	122.00	*****
-----+-----+-----+-----+-----				
STATISTIQUES SUR LES 2 VARIABLES CONTINUES				
+-----+-----+-----+-----+-----				
VARIABLE		EFFECTIF	MOYENNE	MINIMUM
NUM . IDEN	LIBELLE	POIDS	ECART-TYPE	MAXIMUM
+-----+-----+-----+-----+-----				
10 .	Mont - Montant de Epargne	468	1040.79	0.00
		468.00	2884.77	21000.00
+-----+-----+-----+-----+-----				
ESTIMATION / COEFFICIENTS				
AJUSTEMENT DES MOINDRES CARRES (AVEC TERME CONSTANT)				
468 INDIVIDUS, 6 PARAMETRES (CONSTANTE EN QUEUE).				
IDEN	LIBELLE	COEFFICIENT	ECART-TYPE	STUDENT
462				
PROBA. V.TEST				
CRITERE(S)				
m1	- Femme	31.8025	140.104	0.227
* m2	- Homme	-31.8025	140.104	0.227
m1	- <= 40ans	-593.2349	176.181	3.367
m2	- > 50 ans	21.4774	211.746	0.101
* m3	- 40-50 ans	571.7575	204.692	2.793
INTERACTION(S) D'ORDRE 2				
m1	- Femme			
m1	- <= 40ans	185.1550	176.181	1.051
m1	- Femme			
m2	- > 50 ans	80.0732	211.746	0.378
* m1	- Femme			
* m3	- 40-50 ans	-265.2282	167.637	1.582
* m2	- Homme			
* m1	- <= 40ans	-185.1550	176.181	1.051
* m2	- Homme			
* m2	- > 50 ans	-80.0732	211.746	0.378
* m2	- Homme			
* m3	- 40-50 ans	265.2282	167.637	1.582
	CONSTANTE	1182.2612	140.104	8.438
TEST D'AJUSTEMENT GLOBAL				
SOMME DES CARRES DES ECARTS SCE = 3761563136.0000				
COEFFICIENT DE CORRELATION MULTIPLE ... R = 0.1849 R2 = 0.0342				
VARIANCE ESTIMEE DES RESIDUS S2 = 8141912.0000 S = 2853.4036				
TEST DE NULLITE SIMULTANEE DES COEFFICIENTS DES 5 VARIABLES :				
FISHER = 3.269 DEG.LIB = 5 462				
P.CRIT = 0.0065 V.TEST = 2.48				
ANALYSE DE LA VARIANCE (SANS EFFET DE REPETABILITE)				
NB. AJUSTEMENT DES MOINDRES CARRES SUR LE MODELE RENDU DE				
PLEIN RANG PAR LES CONTRAINTES : SOMMES NULLES DES				
COEFFICIENTS. LES SOMMES DE CARRES CORRESPONDENT A				
L'ANNULATION SUCCESSIVE DE GROUPES DE COEFFICIENTS SUR				
LE MODELE DE PLEIN RANG				
SOURCE	SOMME	FISHER	DEGRES DE	PROBA.
	DES CARRES		LIBERTE	CRITIQUE
ECARTS RESIDUELS	3761563392.000		462	
INTERACTION(S) D'ORDRE 2				
Sexe				
Classes d'âges des clients				
----->	16532480.000	1.015	2 462	0.3595
CRITERE(S)				0.36
Sexe				
----->	419584.000	0.052	1 462	0.8205
Classes d'âges des clients				
----->	113376512.000	6.963	2 462	0.0010
				3.09