



COURS DE CLUSTERING

**MASTER 2 SIAD
2024/2025**

OBJECTIFS



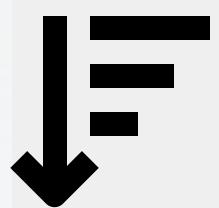
A partir de l'étude d'un cas concret, maîtriser la conception d'une typologie, et les préconisations qui en découlent.



- Prise de recul sur les données
- Interprétation des résultats
- Préconisations
- Présentation des résultats

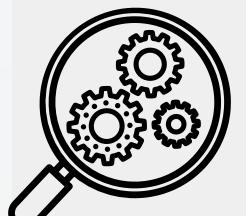


RAPPELS MÉTHODOLOGIQUES



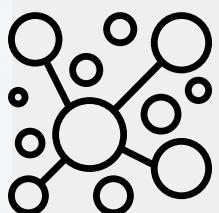
Prise en main des données et de la problématique

Tri des variables, statistiques descriptives



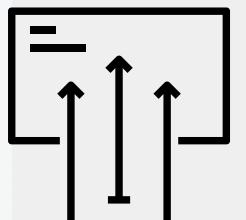
Analyse factorielle des données brutes

ACP, ACF, ACM



Classification basée sur l'analyse factorielle

CAH, Nuée dynamique ...



Modélisation de la classification

Modèle d'affectation des individus aux classes



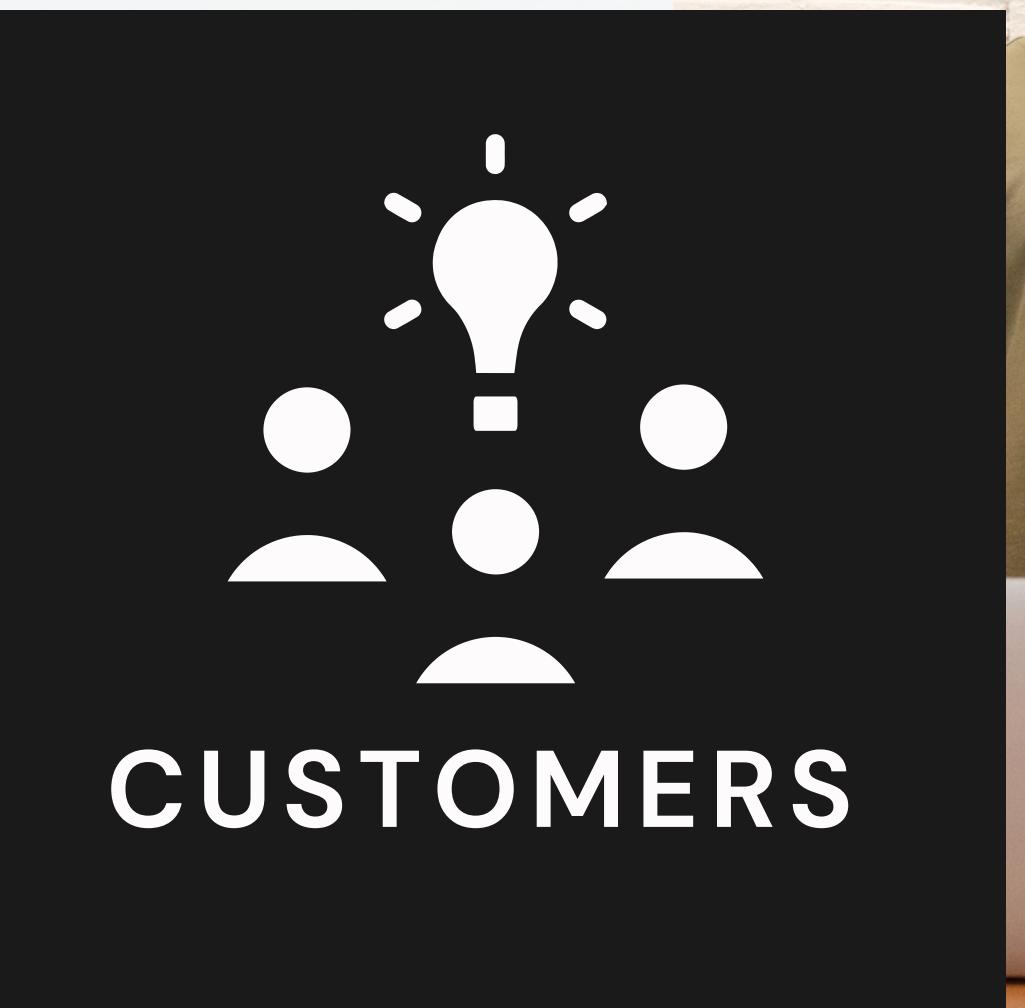
ETUDE DE CAS

Problématique

Le service marketing de la banque en ligne 'BanqTout' souhaite dresser un profil des clients qui se connectent sur leur site web:

- pour améliorer sa connaissance client
- pour personnaliser son site web selon les différents profils de ses internautes

Pour cela vous disposez d'un fichier contenant la liste des internautes avec leurs caractéristiques



DONNÉES DISPONIBLES

Variable	Signification
Num_cli	Identifiant de l'internaute
connexion_mois	Nombre de connexion web sur un mois
source_web	L'origine de la connexion web
typ_client	Type de crédit détenus par l'internaute
montant_credit	Montant du crédit principal (donnée vide si l'internaute n'est plus client)
activite	Activité du crédit principal (actif si la dette est positive, inactif si crédit remboursé)
anciennete	Ancienneté du crédit principal détenu par le client
utilisation_credit	Agrégat donnant le taux de remboursement du crédit principal(+++= reste beaucoup à rembourser, -- = crédit remboursé)
assurance	Crédit principal assuré ou non
logmt	Type de logement
revfyr	Montant des revenus
tr_age	Age
sitfam	Situation familiale
dept	Département de l'internaute

MÉTHODOLOGIE

Construction de la typologie en 3 étapes

Etape technique permettant de faire une première connaissance des liens entre les variables. Permet d'éliminer des variables gênantes, apporte une base de données idéale pour la classification.

**ANALYSE
FACTORIELLE**

C'est l'étape qui crée la typologie « théorique ». Permet d'identifier les groupes de clients, de choisir le nombre de groupes.

CLASSIFICATION

C'est l'étape qui rend « opérationnelle » la typologie. Elle permet via une segmentation ou une analyse discriminante de donner les règles d'affectation de chaque client à son groupe.

AFFECTATION



2 OCT

Compréhension de la
problématique
Analyse descriptive du
fichier

9 OCT

Réalisation
de l'ACM
et interprétation

16 OCT

Comparaison des 3
types de classification
Interprétation des
résultats

23 OCT

Modélisation de la
classification
Interprétation des
résultats
Préconisations

LIVRABLES

Présentation



- 1 support visuel à destination de l'équipe marketing
 - objectif
 - présentation des clusters
 - préconisations
- Le support doit être vulgarisé, c'est à dire qu'il va droit au but, qu'il est clair et efficace. Vous pouvez utiliser les zones de commentaires pour préciser vos idées.

Un rapport avec

- une partie portant sur votre étude, et l'interprétation des résultats
 - une partie technique, avec le détails de vos stats, les explications des décisions que vous avez prises.
- L'objectif est de prévoir qu'une autre personne peut reprendre votre travail des mois plus tard et comprendre votre raisonnement

Rapport



Nom des documents sous la forme : **Gx_By_NOM1_NOM2_presOurrapport** avec les noms par ordre alphabétique, x numéro du groupe, y le numéro du binôme à envoyer à maureen.dhondt@hotmail.fr

DATE DE RENDU DU DOSSIER

A choisir entre vous

Date à rendre le 9/10

Date max : 20/12

PARTIE 1

Compréhension de la problématique

Analyse descriptive de la base de données

RECUEIL DU BESOIN



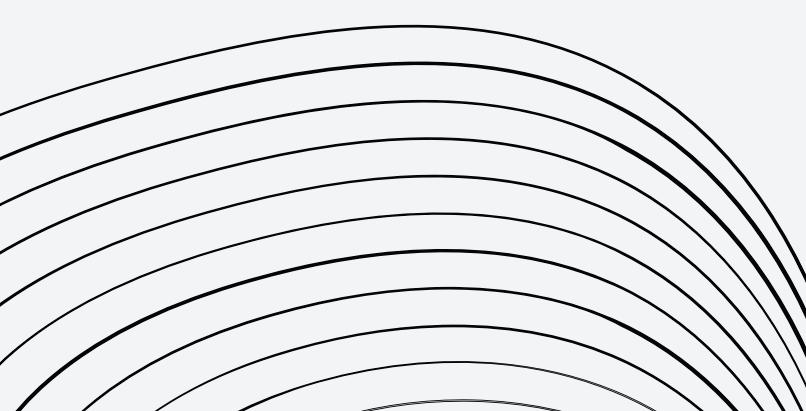
Partie la plus importante de l'étude

Si on ne cerne pas le besoin correctement, on n'y répondra pas

TOUJOURS avoir en tête la problématique

On demande quels seront les usages du clustering pour orienter l'étude et cadrer le besoin

Cela permet de faire le tri a priori des données intéressantes dans la construction du modèle



ANALYSE DESCRIPTIVE

Analyse à une dimension

Permet de prendre connaissance de la base, des modalités des variables
Si on ne comprend pas une variable/ une modalité, on demande au métier de l'expliquer

Analyse à deux dimensions

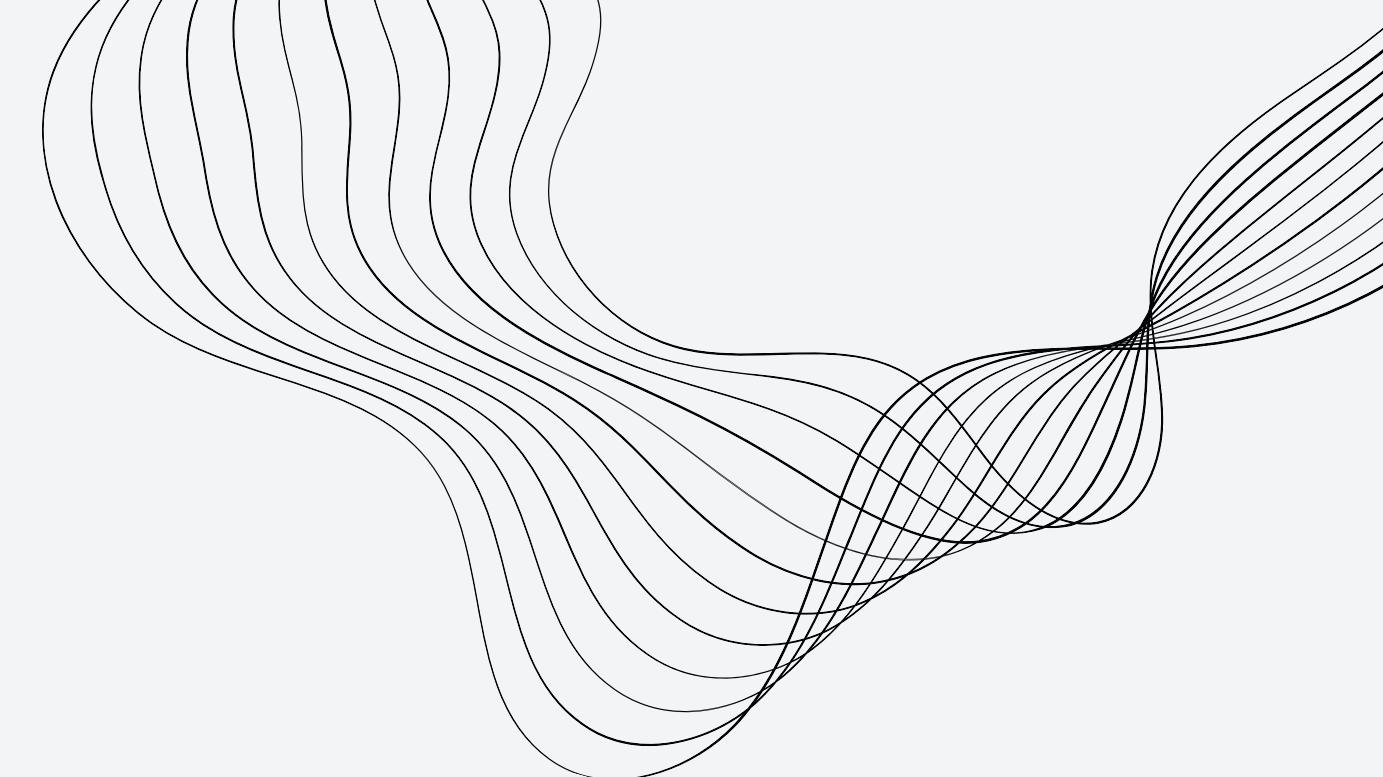
Croisements pertinents des variables entre elles pour répondre à l'objectif de l'étude. Cette étape permet de préparer le clustering et de regrouper des modalités de variables si nécessaire.



PARTIE 2

Analyse des Correspondances Multiples

L'ACM

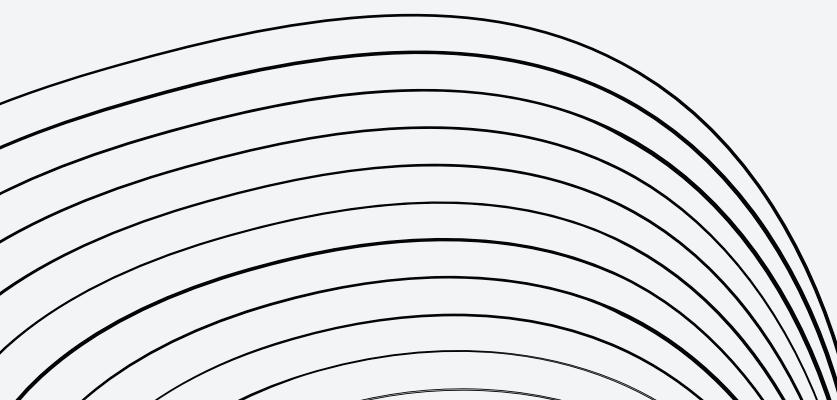


Choix des variables

En se basant sur l'étape 1 on choisit les variables que l'on souhaite faire rentrer dans la construction du modèle et on s'assure d'avoir éliminer les modalités rares

Réalisation de l'ACM : proc corresp sous SAS

- Choix du nombre d'axes à retenir + description des 4 premiers
- Analyse des dimensions retenues



ACM : QUESTIONS PRATIQUES

Histogramme des valeurs propres : Combien de dimensions je retiens ?

- Soit un nombre communément admis (ex : 10 axes)

- Soit le nombre d'axes qui restitue 50% de l'inertie totale

- Soit le nombre d'axes dont l'information restituée est > à l'information moyenne de chaque axe (1/nb VP)

Inertia and Chi-Square Decomposition

Singular Value	Principal Inertia	Chi-Square	Percent	Cumulative Percent	3	6	9	12	15
0.75697	0.57300	51570	13.22	13.22	*****	*****	*****	*****	*****
0.50438	0.25440	22896	5.87	19.09	*****	*****	*****	*****	*****
0.46513	0.21635	19471	4.99	24.09	*****	*****	*****	*****	*****
0.41036	0.16839	15155	3.89	27.97	*****	*****	*****	*****	*****
0.39894	0.15915	14324	3.67	31.65	*****	*****	*****	*****	*****
0.38317	0.14682	13214	3.39	35.03	*****	*****	*****	*****	*****
0.36869	0.13593	12234	3.14	38.17	*****	*****	*****	*****	*****
0.35426	0.12550	11295	2.90	41.07	*****	*****	*****	*****	*****
0.35169	0.12368	11131	2.85	43.92	*****	*****	*****	*****	*****
0.34792	0.12105	10895	2.79	46.71	*****	*****	*****	*****	*****
0.34184	0.11685	10517	2.70	49.41	*****	*****	*****	*****	*****
0.34057	0.11599	10439	2.68	52.09	*****	*****	*****	*****	*****
0.33783	0.11413	10272	2.63	54.72	*****	*****	*****	*****	*****
0.33458	0.11195	10075	2.58	57.30	*****	*****	*****	*****	*****
0.33417	0.11167	10050	2.58	59.88	*****	*****	*****	*****	*****
0.33306	0.11093	9984	2.56	62.44	*****	*****	*****	*****	*****
0.33177	0.11007	9907	2.54	64.98	*****	*****	*****	*****	*****
0.33079	0.10942	9848	2.53	67.51	*****	*****	*****	*****	*****
0.32992	0.10885	9796	2.51	70.02	*****	*****	*****	*****	*****
0.32843	0.10787	9708	2.49	72.51	*****	*****	*****	*****	*****
0.32487	0.10554	9499	2.44	74.94	*****	*****	*****	*****	*****
0.32313	0.10441	9397	2.41	77.35	*****	*****	*****	*****	*****
0.31747	0.10079	9071	2.33	79.68	*****	*****	*****	*****	*****
0.31614	0.09994	8995	2.31	81.99	*****	*****	*****	*****	*****
0.31129	0.09690	8721	2.24	84.22	*****	*****	*****	*****	*****
0.30077	0.09047	8142	2.09	86.31	***	***	***	***	***
0.29136	0.08489	7640	1.96	88.27	***	***	***	***	***
0.28868	0.08333	7500	1.92	90.19	***	***	***	***	***
0.28613	0.08187	7368	1.89	92.08	***	***	***	***	***
0.26760	0.07161	6445	1.65	93.73	***	***	***	***	***
0.25761	0.06636	5973	1.53	95.26	***	***	***	***	***
0.24837	0.06169	5552	1.42	96.69	**	**	**	**	**
0.23662	0.05599	5039	1.29	97.98	**	**	**	**	**
0.22159	0.04910	4419	1.13	99.11	**	**	**	**	**
0.19605	0.03843	3459	0.89	100.00	*	*	*	*	*

$$1/35 \times 100 = 2,85$$

ACM : QUESTIONS PRATIQUES

Analyse des axes : Quelles modalités sont significatives ?

On retient les modalités dont la contribution est supérieure à la contribution moyenne ($\frac{1}{nb \ total \ modalités}$)

Puis on analyse les coordonnées des modalités significatives

Contribution moyenne : 1/49

	Dim2	Contr2	Dim3	Contr3
Carte de credit enseigne partenaire	-0,608853044	0,028246942	1,444376826	0,179044542
Credit voiture	0,731526804	0,013536949	-0,476542765	0,006470199
Pas client ou ancien client	0,289362482	0,002587547	-0,190910965	0,001268583
Reserve d'argent	0,061300065	0,001161416	-0,29799996	0,030913755
1.<=1000e	-0,804249468	0,042261895	0,371507305	0,010156772
2.]1000e- 3000e]	-0,488233782	0,032654506	0,462229912	0,032965308
4.]3000e- 6000e]	0,237589744	0,007281842	-0,510071131	0,037800798

PARTIE 3

Classification

MÉTHODES DE CLASSIFICATION

Classification	Avantage	Inconvénient
non hierarchique (Ex : nuée dynamiques, centres mobiles ...)	Rapide donc classification sur une BDD volumineuse	Nombre de classes imposé dès le départ
Hiérarchique (ex : CAH)	Choix du nombre optimal de classes	Temps de traitement long

Classification mixte pour combiner les avantages des 2 méthodes

CAH: choix du nb de classes (proc cluster, dendrogramme, CCC, R2 semi-partiel)

Nuées dynamiques (proc fastclus)

MÉTHODES DE CLASSIFICATION

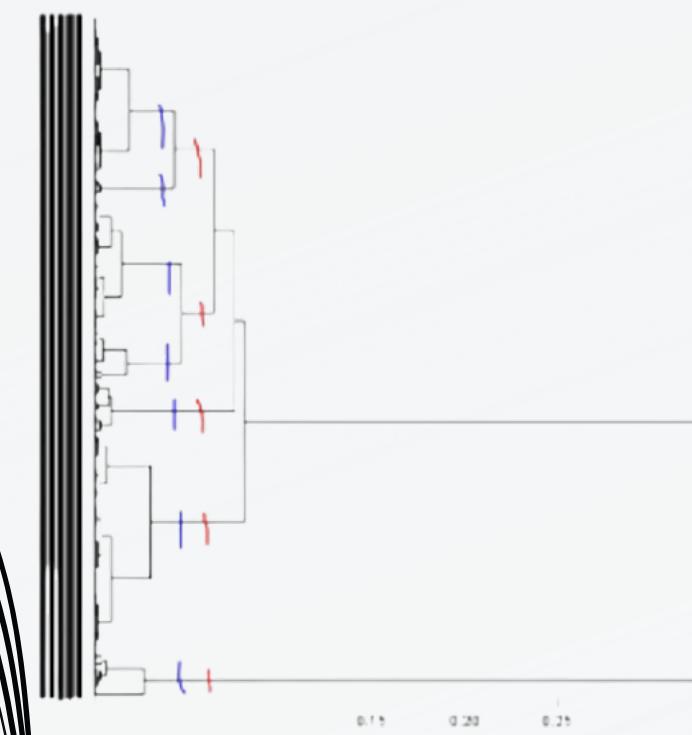


Quelle que soit la méthode
retenue, réalisation de
statistiques descriptives pour
décrire le contenu de chaque
groupe issu de la classification

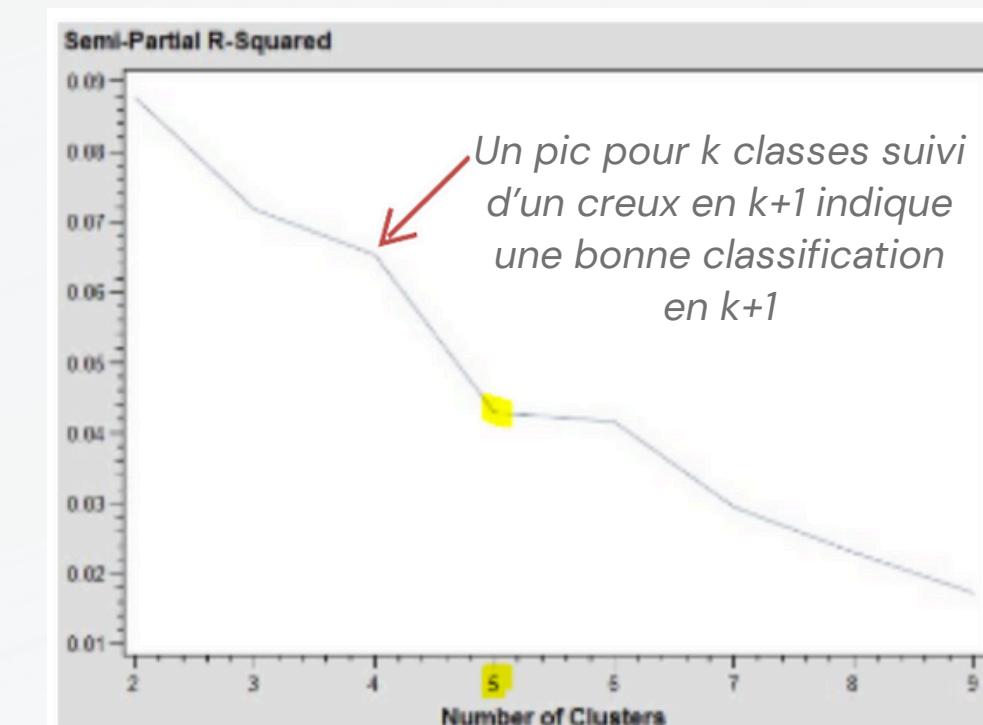
CAH : QUESTIONS PRATIQUES

Comment déterminer le nombre de classes optimal ?

Méthode intuitive
découpage du dendrogramme

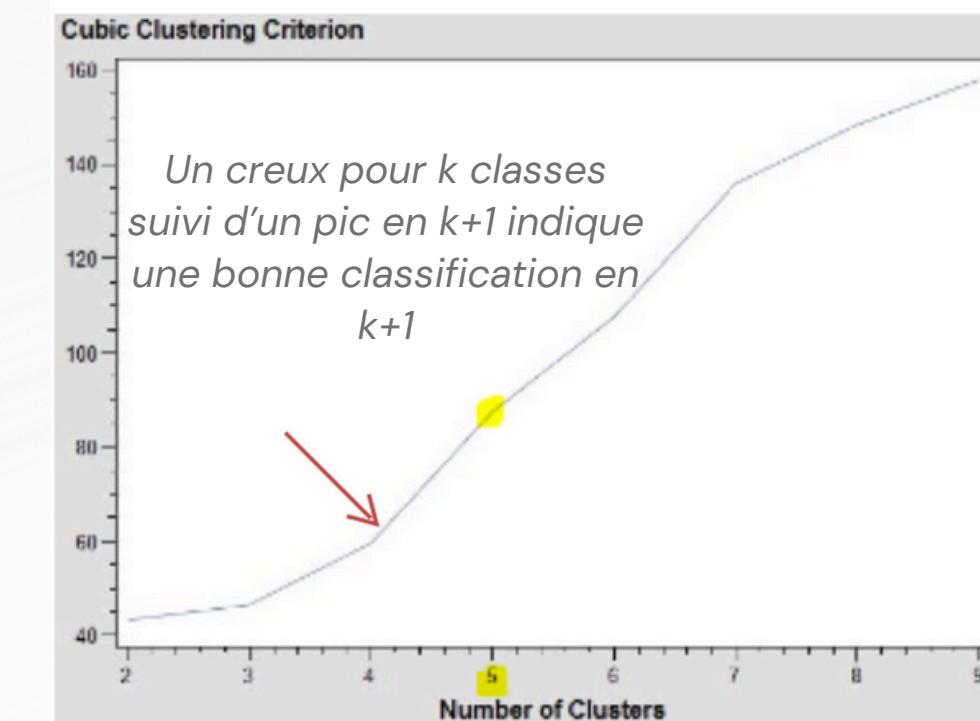


Indicateurs de mesure de qualité



Cubic clustering criterion (CCC):
si > 2 alors la classification est
bonne

R2 semi-partiel: (SPRSQ): mesure
la perte d'inertie interclasse
provoquée en regroupant 2
classes





PARTIE 4

Affection

AFFECTATION DE LA CLASSIFICATION

Un clustering se construit une fois

Une mise à jour est nécessaire pour attribuer un cluster aux nouveaux clients, mais aussi car le comportement du client a pu changer.

On peut attribuer un cluster à un client de plusieurs manières

- règles d'attribution simples
- modélisation (pour le TD méthode retenue, avec une régression logistique généralisée)