

MICROPOL

Data driving transformation

Avril

3

ANALYSES GÉOMÉTRIQUES ET SEGMENTATIONS NON SUPERVISÉES AVEC SEGMENTATION.AI



Tanguy Le Nouvel

MICROPOL

Directeur IA & Data Science
@ tlenouvel@micropole.com



Maitre de conférences associé



AGENDA

- 01 PRÉSENTATION DE MICROPOLE
- 02 EVOLUTION DU MÉTIER DE DATA SCIENTIST
- 03 PANORAMA DES MÉTHODES ET CAS D'USAGES ASSOCIÉS
- 04 DÉMO DE SEGMENTATION.AI

MICROPOLE : CABINET DE CONSEIL SPÉCIALISÉ DANS LA DATA

CHIFFRES CLÉS

14 agences

6 pays

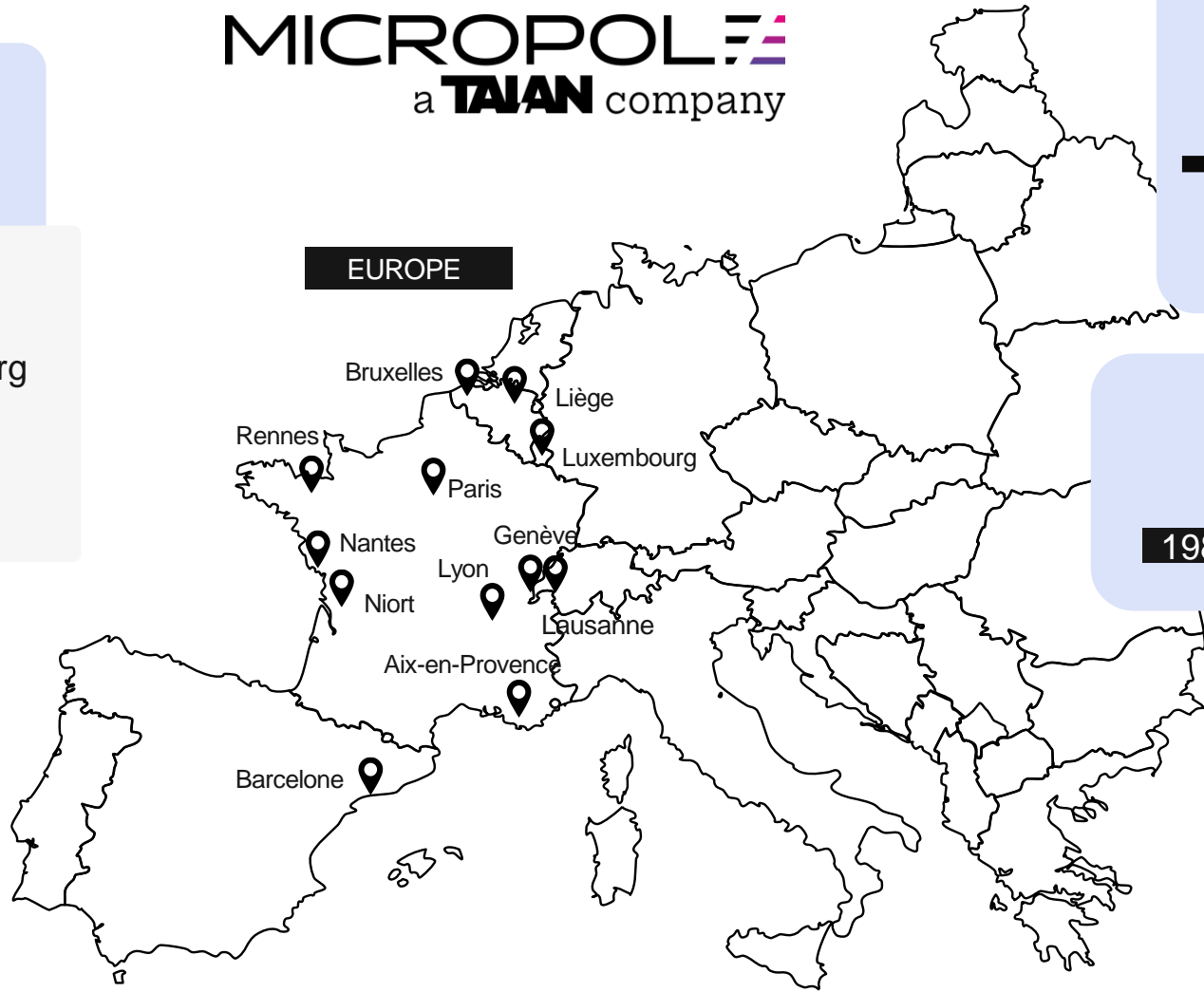
- France
- Belgique
- Luxembourg
- Suisse
- Espagne
- Chine

145 millions d'euros
DE CHIFFRE D'AFFAIRES
30% à l'international

38 ans d'expertise
1987 - 2025

1400 collaborateurs dans le monde
Dont **850** en France

MICROPOLE
a **TAIAN** company



MICROPOLE

DES EXPERTISES SUR TOUTE LA CHAÎNE DE VALEUR DATA



DATA GOVERNANCE & ARCHITECTURE COMPLIANCE

Définition des **organisations et des procédures** à mettre en place afin **d'encadrer la collecte et l'utilisation des données**.

S'organiser pour **faire face à la nouvelle ère de pilotage par les data**.

Protection données personnelles, Vigilance connaissance client KYC, Maîtrise des risques financiers.



IA & DATA SCIENCES

Exploration, croisement et analyse des données pour **comprendre, prédire et anticiper**.

Détecter des signaux faibles en avance de phase afin d'identifier des leviers de croissance **ou limiter l'impact de facteurs négatifs**.

Automatiser vos process



DATA INTELLIGENCE & BIG DATA

Extraction, stockage, transformation, transfert, analyse, exploitation des données.

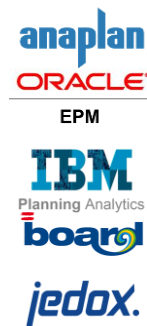
Manipuler les données, sécuriser l'information, apporter de l'intelligence, exploiter la volumétrie.



FINANCE TRANSFORMATION & PERFORMANCE

Structuration et croisement des données à des fins de pilotage et d'analyse.

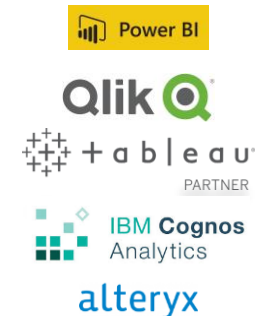
Refléter le mode de pilotage de l'organisation, structurer les données pour les faire parler, **simuler le champ du possible.**



DATA VISUALISATION & BUSINESS INTELLIGENCE

Visualisation et restitution intelligente des données.

Rendre **accessible et intelligible les données, faciliter la prise de décision**, donner envie et générer de l'adhésion.



FOCUS INTELLIGENCE ARTIFICIELLE & DATA SCIENCE



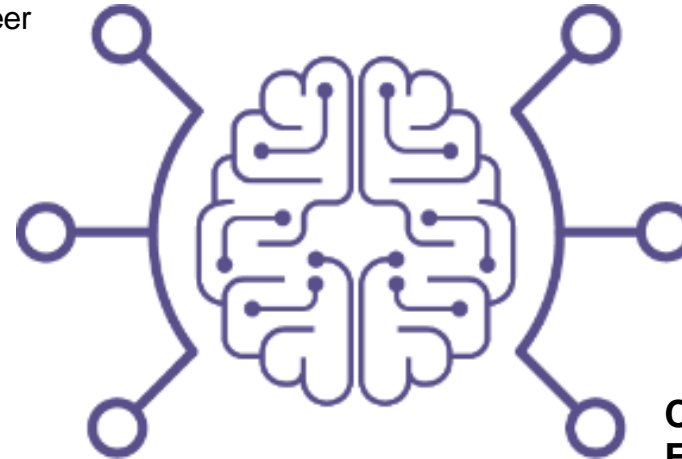
120 Experts
Data Scientists / Statisticiens /
Data engineer / ML Engineer

Plus de 250 projets / + 100 clients
• réalisés depuis 2009



Innovation et co-working

- OpenGround dédié à nos clients et partenaires
- Guilde Data Science (REX, Veille)



Chaine complète de la donnée

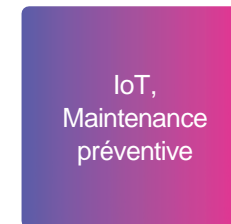
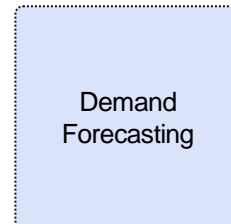
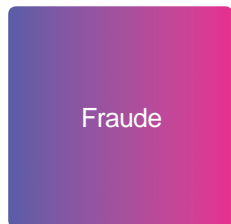
- Data intégration / Data Quality / Data Prep
- Machine Learning / NLP / Deep Learning / Gen AI
- Industrialisation / APIisation
- Monitoring

R&D et développement d'algorithmes sur-mesure

SCORING.AI SEGMENTATION.AI
FOZZY.AI FRAUD.AI

**Conseil, Expertise, Délégation,
Formation & Développement de solutions sur-mesure**

Plusieurs pôles de compétences et une expertise reconnue



EVOLUTION DU MÉTIER DE DATA SCIENTIST ET OPPORTUNITÉS



FAITS MARQUANTS

- **Démocratisation de l'accès aux algorithmes de Machine Learning et de Deep Learning via l'avènement de l'Open Source** (avant il fallait déboursier des dizaines voire des centaines de milliers d'euros pour y avoir accès)
- **Open Source = Programmation** : on est revenu à l'ère du tout programmation (CF avant Windows...). Les logiciels à base d'interface ont quasiment disparu à de rares exceptions.
- **Explosion de l'usage du langage Python** malgré les réticences de certaines DSI
- **Explosion du cloud** : tout nouvel acteur peut monter une plate-forme data en très peu de temps et disposer des meilleurs standards du marché
- **Tout est API / Service temps réel** : explosion du nombre de modèles mis en production
- **Nouveaux process CI/CD, MLOPS adapté aux projets IA / ML**
- **Nouvelles réglementations : GDPR, AI ACT** avec des amendes salées
- Le métier de data scientist n'est sans doute plus le job le plus sexy du 21^{ème} siècle !

IMPACTS

- « **Informatisation** » du métier de **data scientist** : on passe plus de temps à coder (CF Gestion des variables catégorielles, sorties de résultats / graphiques), parfois moins de temps à interpréter / analyser / restituer / expliquer
- **Difficultés de recrutement** : nouveaux profils plutôt à l'aise en programmation mais avec des lacunes en statistique et un manque d'intérêt pour la data et la compréhension métier
- **Avec l'open source, les entreprises deviennent leur propre éditeur**
- **Besoin accru de gouvernance** pour cadrer les usages, les environnements technologiques, les processus / méthodologies projet, les bonnes pratiques de développement, la conformité, le monitoring des modèles / services mis en production
- **Certains oublient l'essentiel** : compréhension des besoins métiers et des données

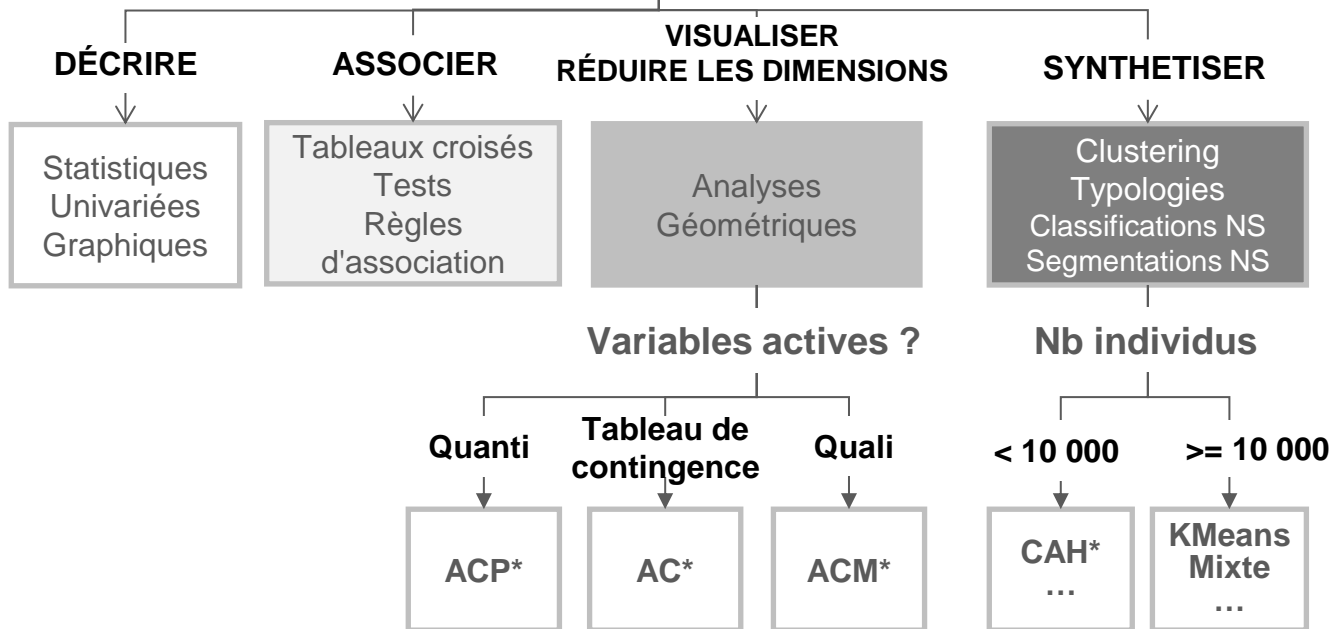
PANORAMA DES MÉTHODES ET CAS D'USAGES ASSOCIÉS

PANORAMA DES MÉTHODES

CADRE NON SUPERVISÉ APPROCHE EXPLORATOIRE

DECRIRE, ASSOCIER, VISUALISER, EXPLORER, SYNTHETISER

TYPE D'EXPLORATION

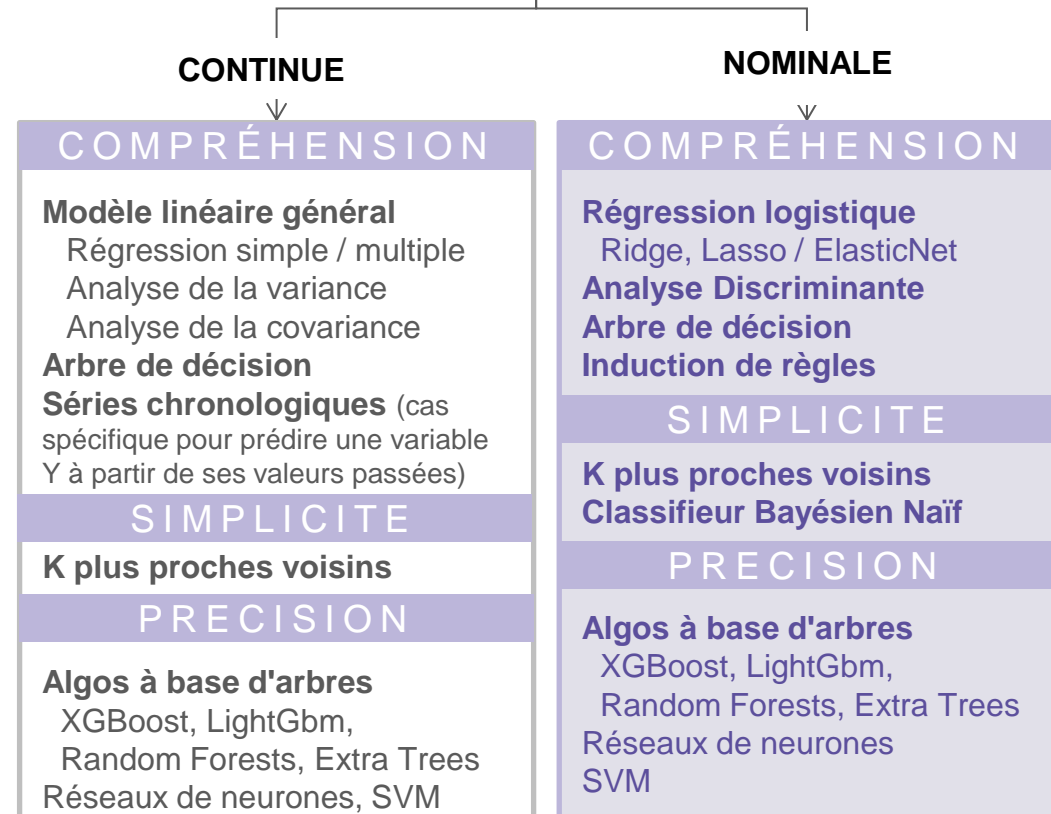


ACP : Analyse en Composantes Principales
 AC : Analyse des Correspondances
 ACM : Analyse des Correspondances Multiples
 CAH : Classification Ascendante Hiérarchique

CADRE SUPERVISÉ APPROCHE PREDICTIVE

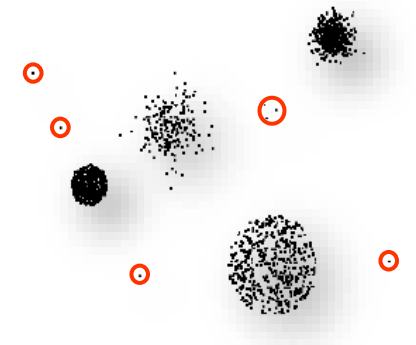
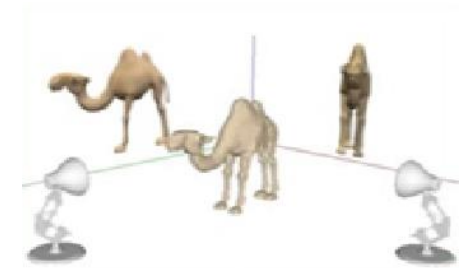
MODELISER, CLASSER, SCORER, PREDIRE

TYPE DE LA VARIABLE À EXPLIQUER

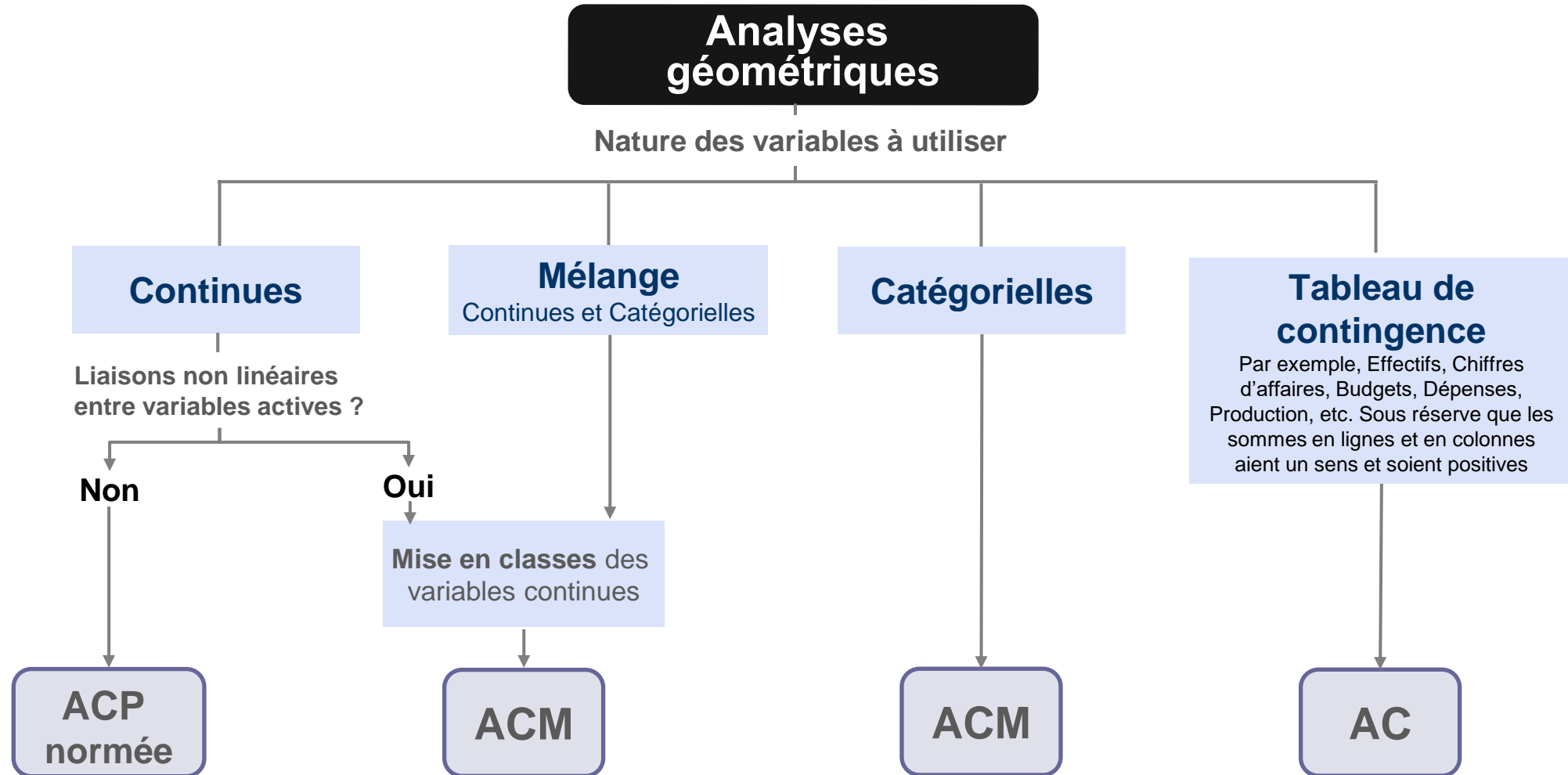


MÉTHODES D'ANALYSE GÉOMÉTRIQUE

- **Pour explorer, visualiser l'essentiel des informations contenues dans un tableau de données en quelques « photos »**
- **Pour visualiser et comprendre les ressemblances / différences entre individus (clients, patients, départements, points de vente, produits) selon un ensemble de critères**
- **Pour visualiser les liaisons et oppositions entre variables**
 - Corrélations entre variables avec l'ACP (Cercle des corrélations)
 - Correspondances entre modalités avec l'ACM
 - Correspondances entre profils lignes et colonnes avec l'AFC
- **Pour réduire l'information** en ne conservant que quelques axes ou en éliminant certaines variables corrélées entre-elles (ACP avant régression multiple par exemple)
- **Ces analyses constituent souvent une étape intermédiaire avant la classification non supervisée**



COMMENT CHOISIR LA BONNE MÉTHODE ?



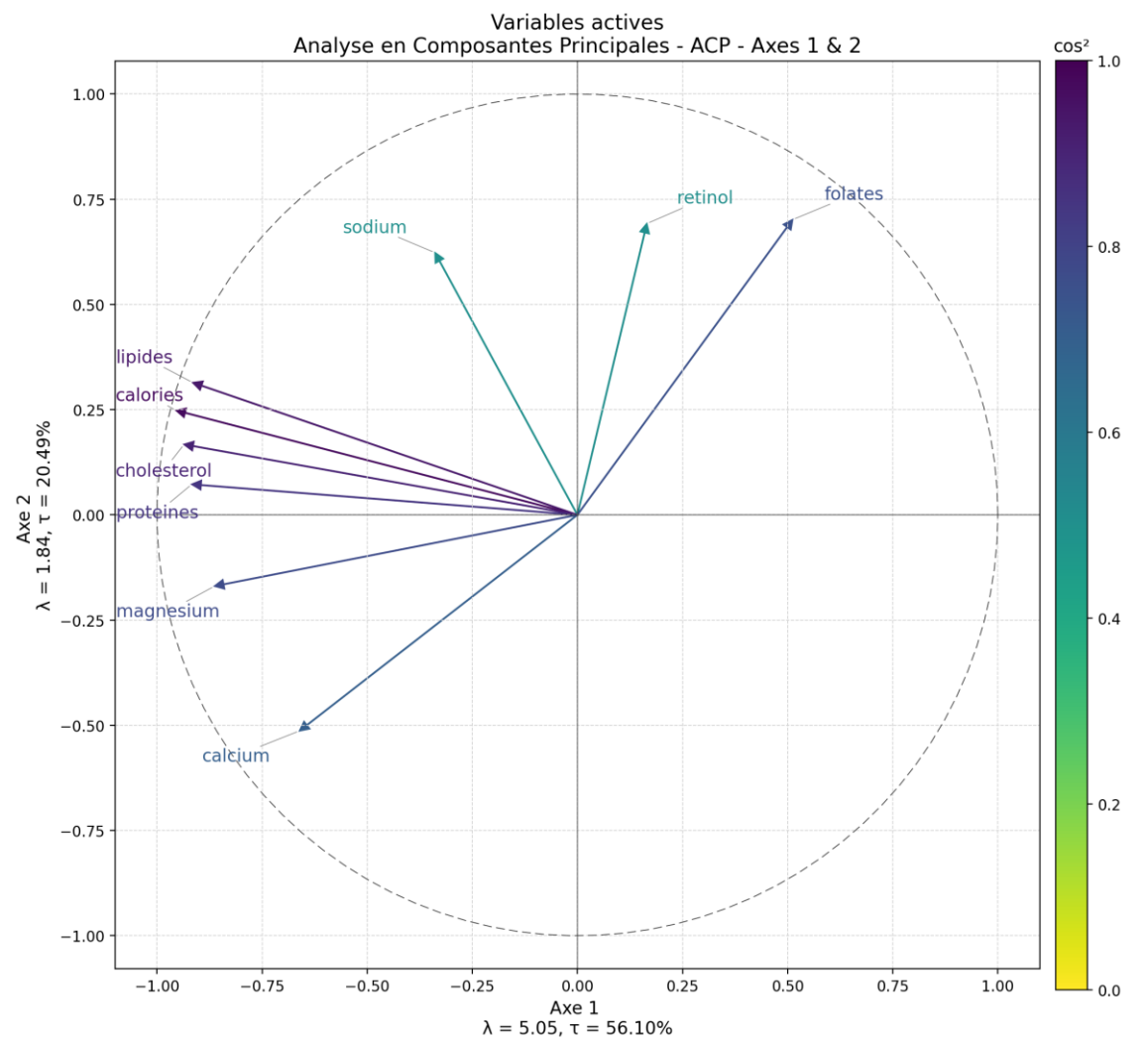
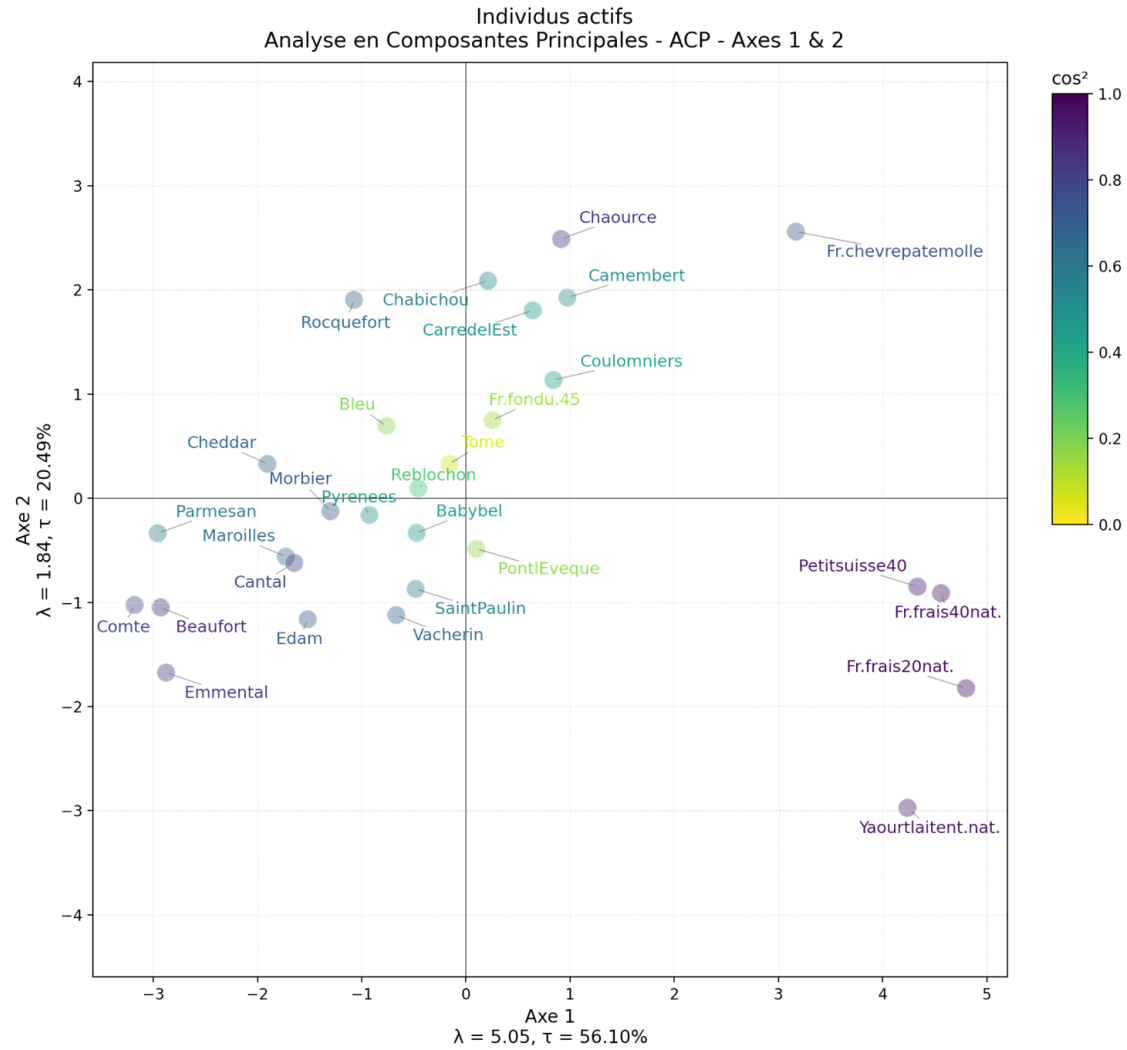
EXEMPLE SUR DES FROMAGES

ACP
normée

Fromages	calories	sodium	calcium	lipides	retinol	folates	proteines	cholesterol	magnesium
Carre del Est	314	353.5	72.6	26.3	51.6	30.3	21	70	20
Babybel	314	238	209.8	25.1	63.7	6.4	22.6	70	27
Beaufort	401	112	259.4	33.3	54.9	1.2	26.6	120	41
Bleu	342	336	211.1	28.9	37.1	27.5	20.2	90	27
Camembert	264	314	215.9	19.5	103	36.4	23.4	60	20
Cantal	367	256	264	28.8	48.8	5.7	23	90	30
Chabichou	344	192	87.2	27.9	90.1	36.3	19.5	80	36
Chaource	292	276	132.9	25.4	116.4	32.5	17.8	70	25
Cheddar	406	172	182.3	32.5	76.4	4.9	26	110	28
Comte	399	92	220.5	32.4	55.9	1.3	29.2	120	51
Coulomniers	308	222	79.2	25.6	63.6	21.1	20.5	80	13
Edam	327	148	272.2	24.7	65.7	5.5	24.7	80	44
Emmental	378	60	308.2	29.4	56.3	2.4	29.4	110	45
Fr. chevrepate molle	206	160	72.8	18.5	150.5	31	11.1	50	16
Fr. fondu.45	292	390	168.5	24.0	77.4	5.5	16.8	70	20
Fr. frais 20 nat.	80	41	146.3	3.5	50	20	8.3	10	11
Fr. frais 40 nat.	115	25	94.8	7.8	64.3	22.6	7	30	10
Maroilles	338	311	236.7	29.1	46.7	3.6	20.4	90	40
Morbier	347	285	219	29.5	57.6	5.8	23.6	80	30
Parmesan	381	240	334.6	27.5	90	5.2	35.7	80	46
Petitsuisse 40	142	22	78.2	10.4	63.4	20.4	9.4	20	10
Pont l'Évêque	300	223	156.7	23.4	53	4	21.1	70	22
Pyrenées	355	232	178.9	28.0	51.5	6.8	22.4	90	25
Reblochon	309	272	202.3	24.6	73.1	8.1	19.7	80	30
Roquefort	370	432	162	31.2	83.5	13.3	18.7	100	25
Saint Paulin	298	205	261	23.3	60.4	6.7	23.3	70	26
Tome	321	252	125.5	27.3	62.3	6.2	21.8	80	20
Vacherin	321	140	218	29.3	49.2	3.7	17.6	80	30
Yaourt laitent. nat.	70	91	215.7	3.4	42.9	2.9	4.1	13	14

EXEMPLE SUR DES FROMAGES

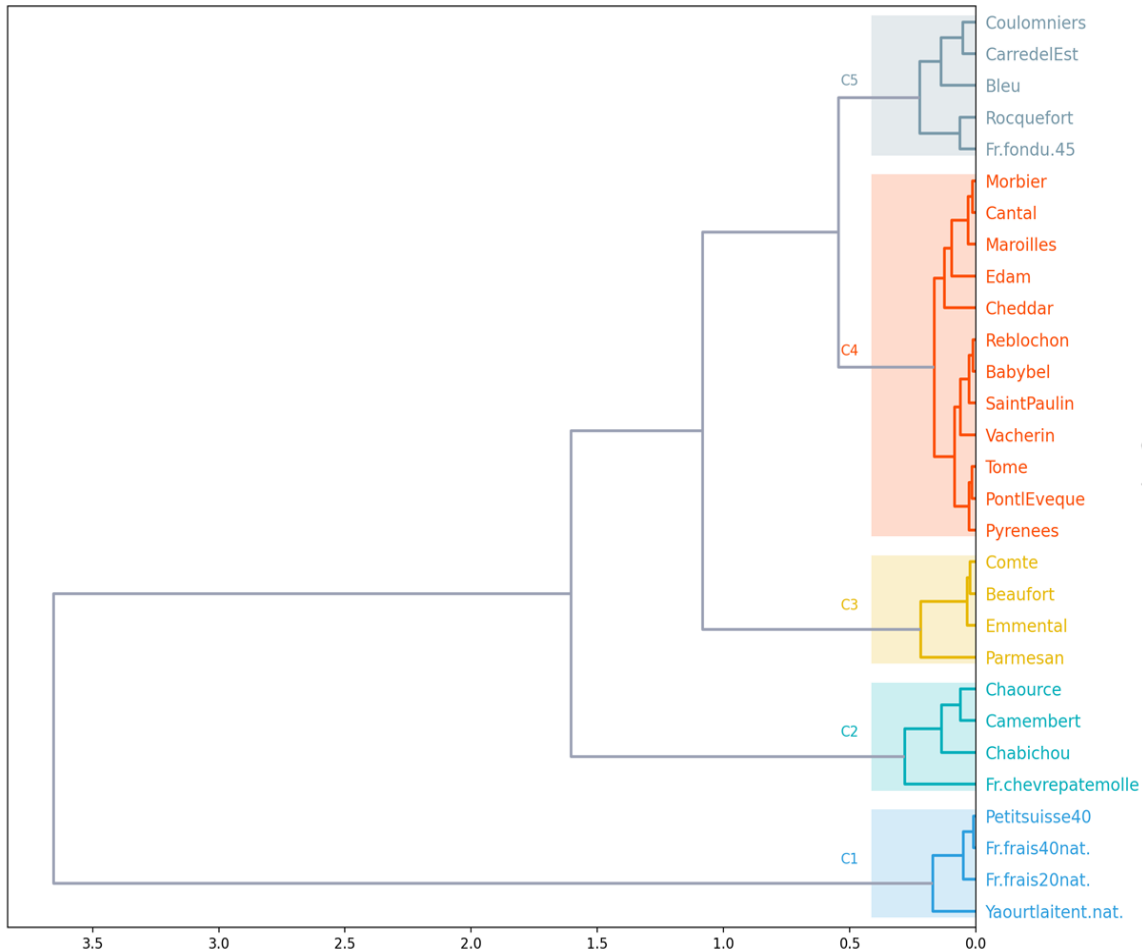
ACP
normée



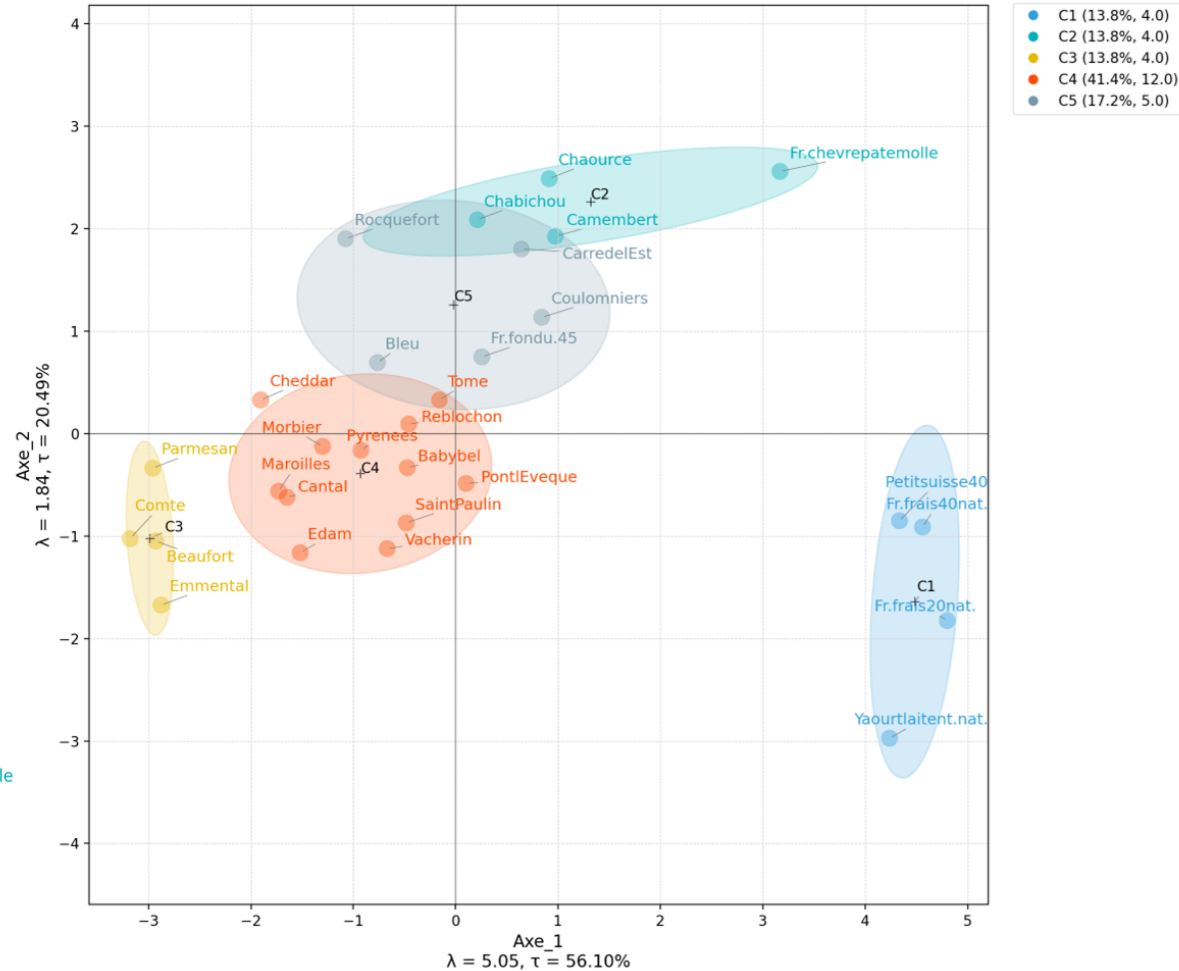
SEGMENTATION DE FROMAGES

CAH

CAH en 5 classes

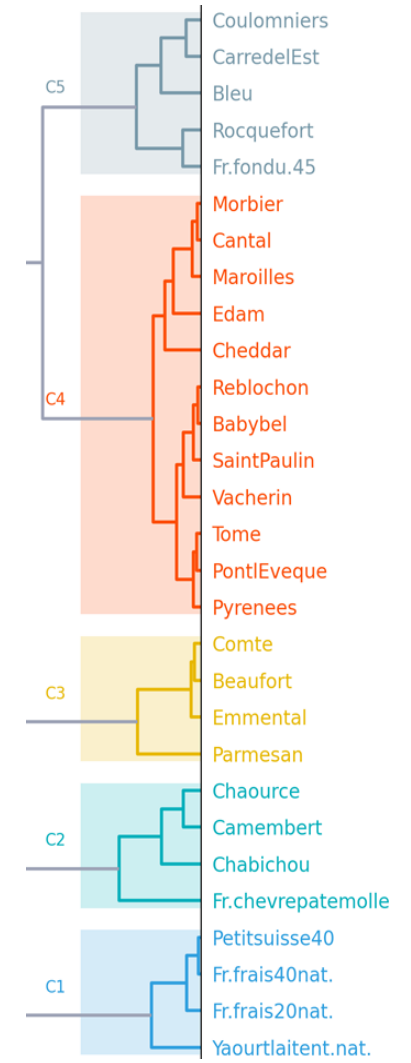
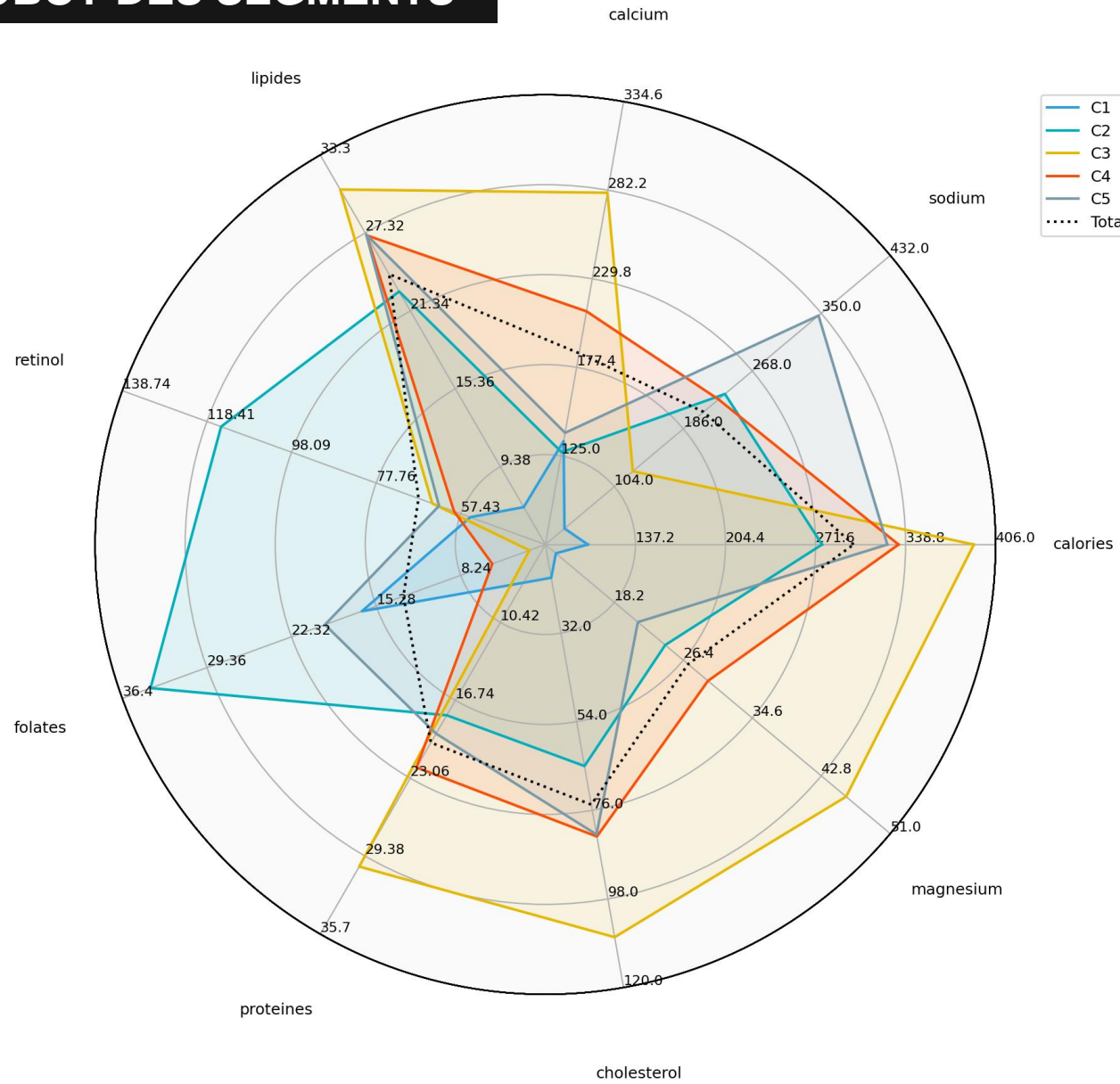


Classification en 5 classes - Axe 1 & 2



PORTRAIT-ROBOT DES SEGMENTS

CAH



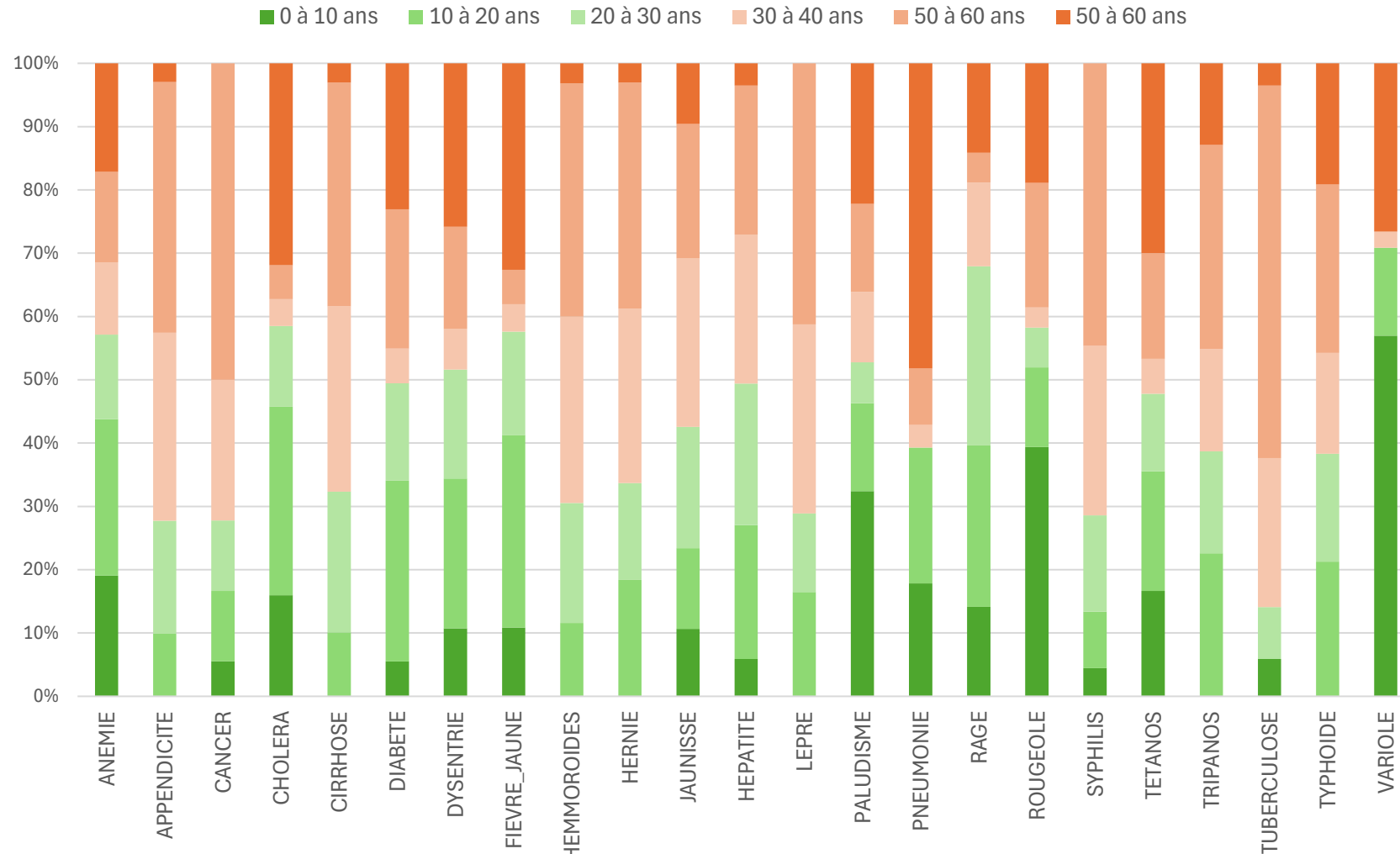
EXEMPLE D'ANALYSE D'UN TABLEAU DE CONTINGENCE

AC

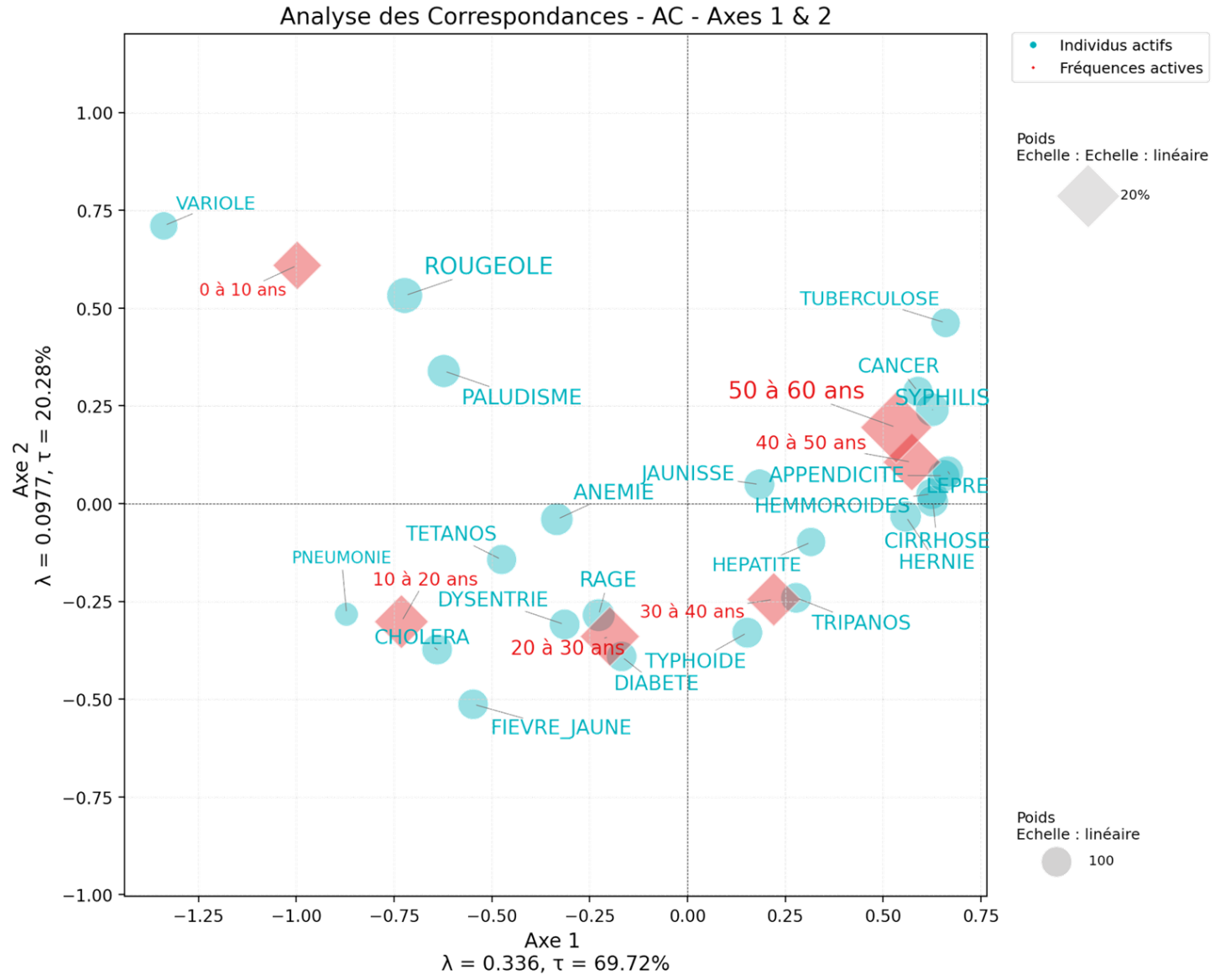
Maladie	0 à 10 ans	10 à 20 ans	20 à 30 ans	30 à 40 ans	40 à 50 ans	50 à 60 ans
ANEMIE	20	18	26	14	12	15
APPENDICITE	0	3	10	18	30	40
CANCER	5	0	10	10	20	45
CHOLERA	15	30	28	12	4	5
CIRRHOSE	0	3	10	22	29	35
DIABETE	5	21	26	14	5	20
DYSENTRIE	10	24	22	16	6	15
FIEVRE_JAUNE	10	30	28	15	4	5
HEMMOROIDES	0	3	11	18	28	35
HERNIE	0	3	18	15	27	35
JAUNISSE	10	9	12	18	25	20
HEPATITE	5	3	18	19	20	20
LEPRE	0	0	16	12	29	40
PALUDISME	35	24	15	7	12	15
PNEUMONIE	10	27	12	0	2	5
RAGE	15	15	27	30	14	5
ROUGEOLE	50	24	16	8	4	25
SYPHILIS	5	0	10	17	30	50
TETANOS	15	27	17	11	5	15
TRIPANOS	0	12	21	15	15	30
TUBERCULOSE	5	3	0	7	20	50
TYPHOIDE	0	18	20	16	15	25
VARIOLE	45	21	11	0	2	0

EXEMPLE D'ANALYSE D'UN TABLEAU DE CONTINGENCE

Répartition des cas de décès par classe d'âge pour chaque maladie



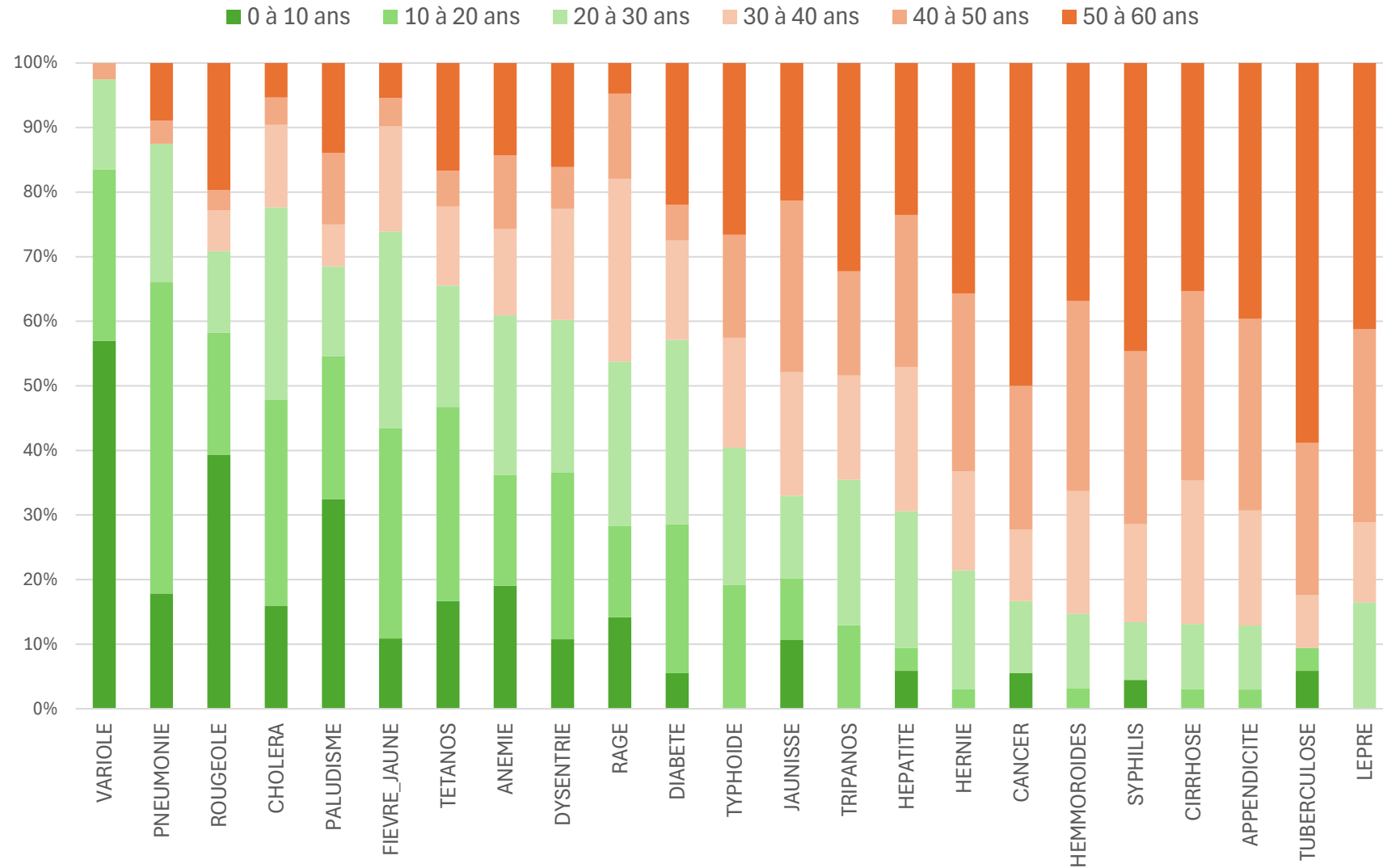
EXEMPLE D'ANALYSE D'UN TABLEAU DE CONTINGENCE



EXEMPLE D'ANALYSE D'UN TABLEAU DE CONTINGENCE

AC

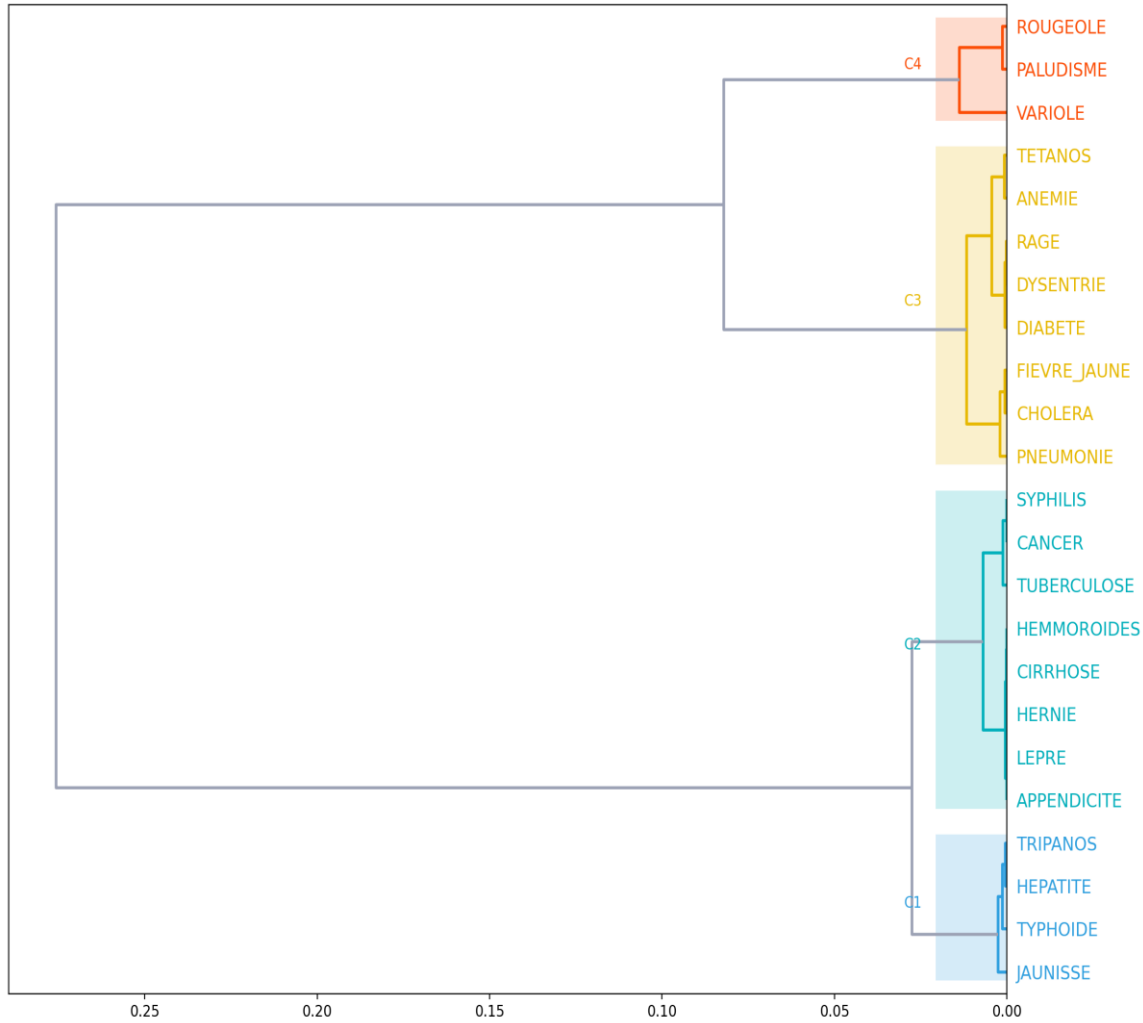
Répartition des cas de décès par classe d'âge pour chaque maladie



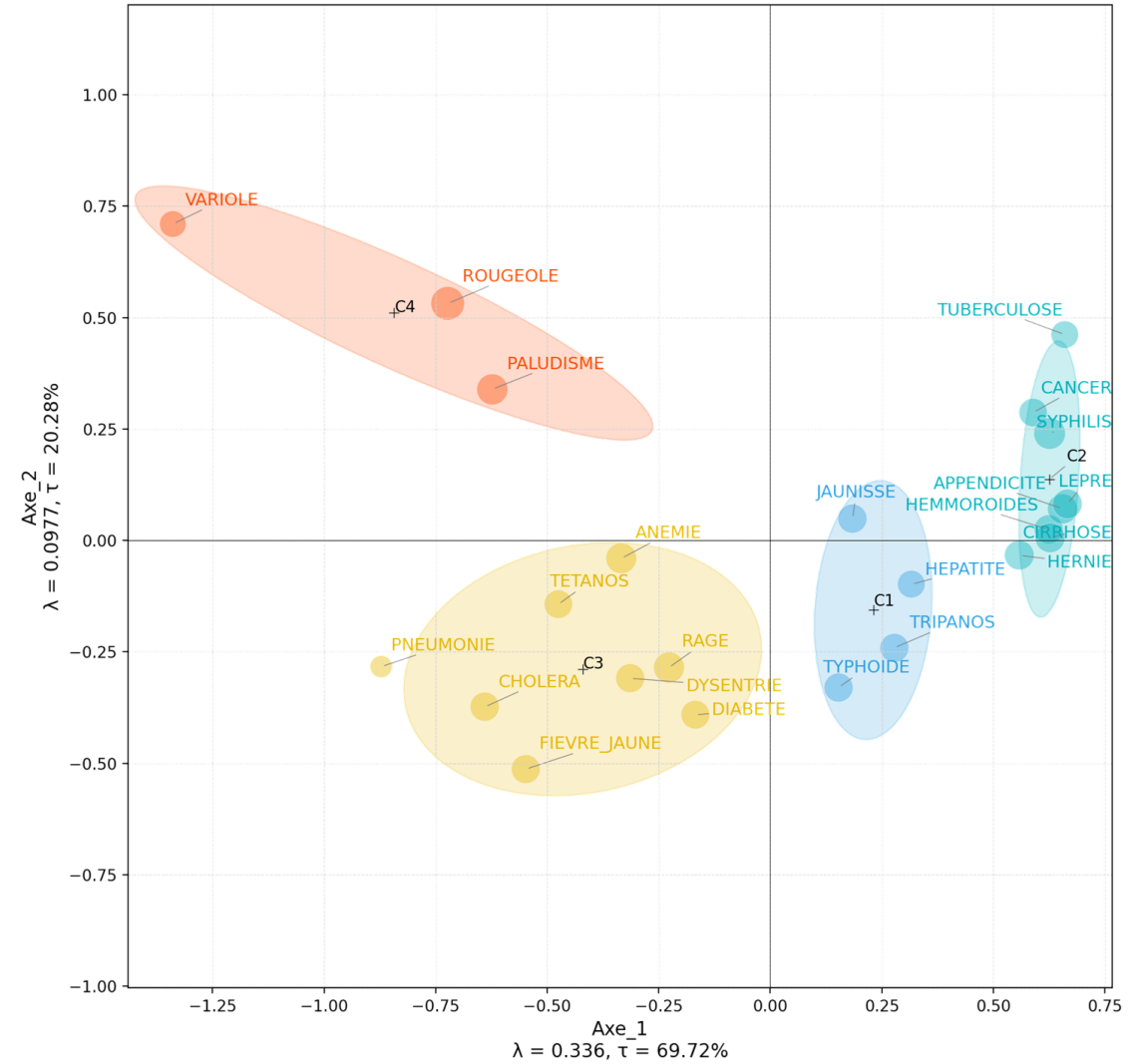
En triant les maladies selon leur coordonnée sur l'axe 1 de l'analyse des correspondances, on obtient un graphique beaucoup plus interprétable !

SEGMENTATION DES MALADIES

CAH en 4 classes



Classification en 4 classes - Axe 1 & 2



QUELQUES CAS D'USAGES

SEGMENTATION
PRODUITS

SEGMENTATION
COMPORTEMENTALE

SEGMENTATION
ATTITUDINALE

SEGMENTATION
DE PARCOURS DE
NAVIGATION

SEGMENTATION DE
POINTS DE VENTES

DÉTECTION
D'ANOMALIES

SEGMENTATION
AFFINITAIRE

SEGMENTATION
MULTICANALE



DÉTECTION DE
MULTICOLINÉARITÉ

RÉDUCTION DE
DIMENSIONS

ENQUÊTES
D'OPINIONS /
SATISFACTION

The logo for MICROPOL, featuring the word "MICROPOL" in a bold, black, sans-serif font, followed by a stylized graphic element consisting of three horizontal lines of varying lengths, colored in shades of purple and blue.

MICROPOL

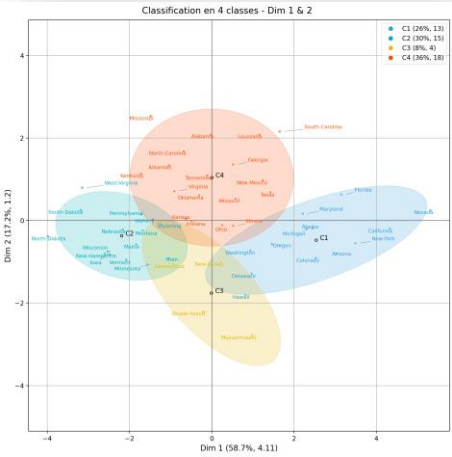
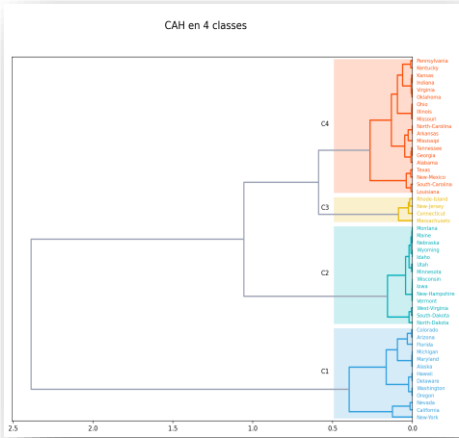
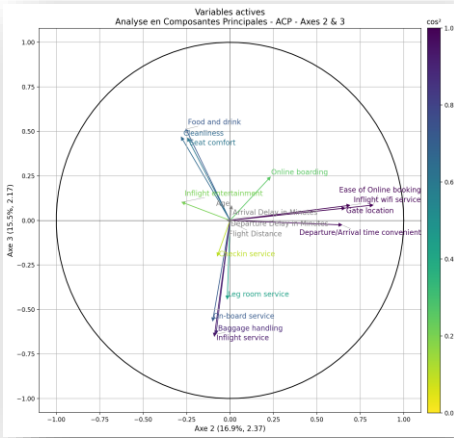
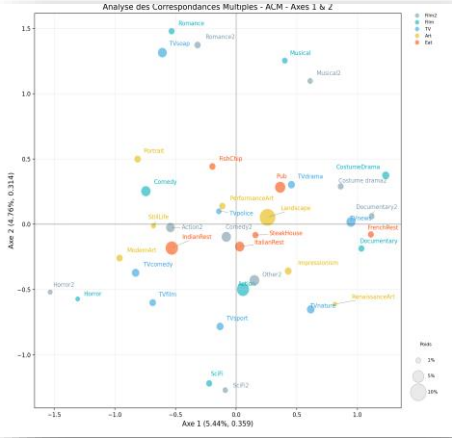
The logo for SEGMENTATION.AI, featuring the word "SEGMENTATION.AI" in a bold, black, sans-serif font. The letter "T" is replaced by a blue icon of three interconnected nodes, resembling a simple neural network or data structure.

SEGMENTATION.AI

**Plateforme d'analyse de
données multidimensionnelles
et de classifications non
supervisées**

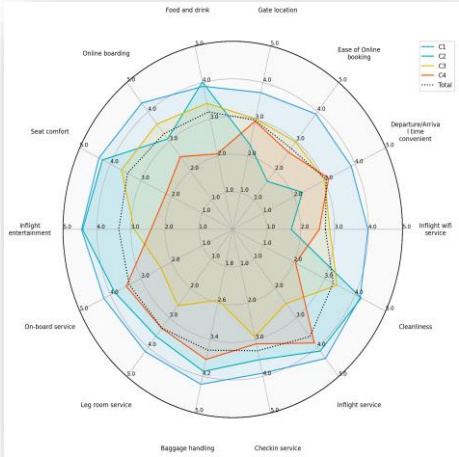


SEGMENTATION.AI



Tableaux croisés - Distribution des variables continues (mises en classes) au sein des différents clusters

Variables	C1 (26%, 13)	C2 (30%, 15)	C3 (18%, 9)	C4 (26%, 13)	Total
age	8,1%	38,8%	0,0%	2,0%	4,8%
age	17,00-24,30	15,0%	28,0%	0,0%	6,1%
age	24,30-31,60	18,6%	74,2%	0,0%	25,0%
age	31,60-38,90	25,4%	8,1%	38,6%	30,1%
age	38,90-46,20	14,9%	0,0%	28,6%	27,0%
age	46,20-53,50	0,0%	0,0%	22,9%	11,1%
age	53,50-60,80	5,8%	0,0%	12,2%	5,9%
age	60,80-68,10	5,8%	0,0%	11,1%	5,9%
age	68,10-75,40	2,1%	0,0%	3,2%	0,2%
age	75,40-82,70	0,5%	0,0%	0,2%	0,0%
age	82,70-90,00	0,3%	0,0%	0,7%	0,0%
educational	1,00-2,50	0,5%	0,4%	2,1%	0,0%
educational	2,50-4,00	0,7%	0,0%	2,7%	0,4%
educational	4,00-5,50	1,9%	2,8%	8,2%	0,1%
educational	5,50-7,00	2,0%	3,7%	5,0%	0,8%
educational	7,00-8,50	2,7%	8,7%	5,0%	0,9%
educational	8,50-10,00	25,9%	42,3%	45,5%	13,9%
educational	10,00-11,50	22,8%	35,8%	21,4%	8,6%
educational	11,50-13,00	2,4%	2,2%	1,5%	0,0%
educational	13,00-14,50	10,4%	3,7%	4,8%	18,1%
educational	14,50-16,00	7,3%	0,3%	6,8%	37,3%
hours per week	1-10	1,8%	0,0%	0,0%	0,0%
hours per week	10-20	9,9%	12,7%	4,2%	1,2%
hours per week	20-30	4,1%	11,6%	1,6%	2,1%
hours per week	30-40	4,3%	6,7%	62,5%	32,4%
hours per week	40-50	19,3%	19,3%	7,9%	0,0%
hours per week	50-60	16,8%	4,0%	6,9%	28,3%
hours per week	60-70	8,7%	0,4%	1,9%	18,9%
hours per week	70-80	2,0%	0,0%	0,5%	6,1%
hours per week	80-90	0,6%	0,0%	0,1%	3,7%
hours per week	90-100	0,0%	0,0%	0,3%	1,3%

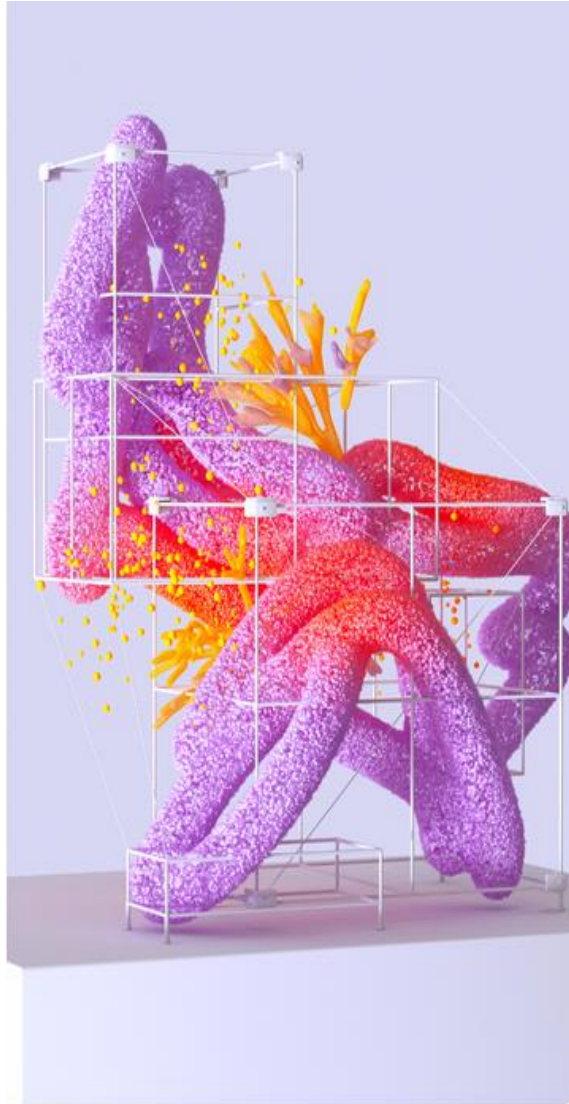


Logiciel **NoCode** permettant de mettre en œuvre et d'industrialiser des **analyses de données multidimensionnelles** et des **segmentations non supervisées** en un temps record.

Résultats et graphiques automatiquement édités et archivés dans **Excel**

Aides à l'interprétation des axes et portraits-robots des segments mettant en avant leurs caractéristiques significatives

Déploiement simplifié et automatisé sous forme d'**API Rest** ou de traitement **batch**



01 Solution logicielle NoCode d'analyses de données multidimensionnelles et de segmentations non supervisée, développée en langage Python



02

Analyses de données multidimensionnelles préalables de type : AC, ACP, ACM, ACMspé

Pondération des individus et/ou les variables

03

Classifications non supervisées par CAH, KMEANS et classification mixte

04

Edition automatique de multiples graphiques représentant l'analyse multidimensionnelle et les résultats des classifications

05

Sorties Excel complètes intégrant l'ensemble des résultats dont la description des axes, la description des clusters



06

Déploiement automatique par calcul de distances (Mahalanobis ou géométrique) et ou modèle prédictif



07

Détection d'anomalies et explicabilité des anomalies

RÉPERTOIRE DE PROJET ET FICHER À IMPORTER

VISUALISATION ET RÉDUCTION DE DIMENSIONS

CLASSIFICATION NON SUPERVIS...
 ON

DEPLOIEMENT
 OFF

RESULTATS

VISUALISATION ET RÉDUCTION DE DIMENSIONS

- Choix de la méthode
- Rôle des variables
- Rôle des individus
- Variables catégorielles
- Sorties graphiques
- Sorties Excel**
- Export csv
- Inversion des axes

RUN

ÉDITION DES SORTIES EXCEL

Nombre d'axes à éditer dans Excel

Gestion des individus

Édition des individus dans Excel

Édition des individus extrêmes dans Excel

Nombre d'écarts-types pour leur identification

Critères de sélection pour la description des axes

Variables actives : Contribution moyenne Autre seuil en % :

Individus actifs : Contribution moyenne Autre seuil en % :

Éléments supplémentaires

Taille de l'effet (Cohen) Seuil D de Cohen

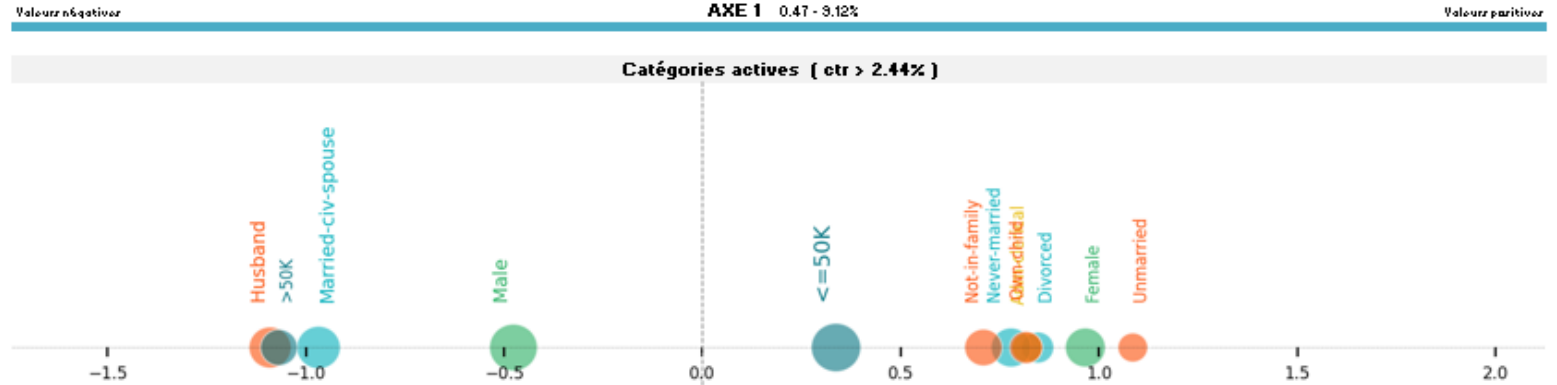
Significativité (pvalue) Seuil de pvalue

Valeur par défaut

Description des axes

Décomposition de la variance

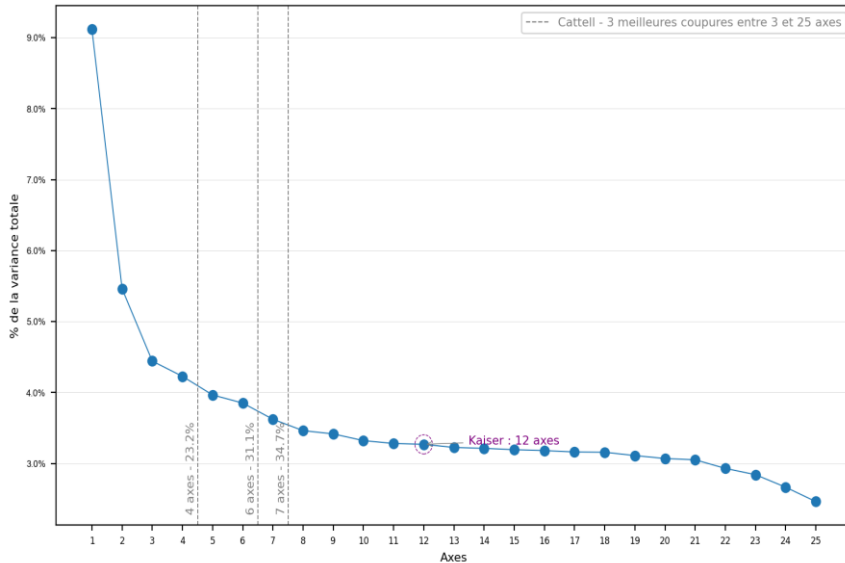
Axe	Variance	Taux de variance (en %)	Taux cumulés (en %)	Taux de variance modifiés (en %)	Taux modifiés cumulés (en %)	Critère de Kaiser
Axe 1	0.4727	9.12%	9.12%	75.81%	75.81%	0
Axe 2	0.2832	5.46%	14.58%	12.09%	87.90%	0
Axe 3	0.2304	4.44%	19.02%	4.08%	91.99%	0
Axe 4	0.2192	4.23%	23.25%	2.94%	94.92%	0
Axe 5	0.2057	3.97%	27.21%	1.79%	96.71%	0
Axe 6	0.1997	3.85%	31.06%	1.38%	98.09%	0
Axe 7	0.1879	3.62%	34.69%	0.72%	98.81%	0
Axe 8	0.1796	3.46%	38.15%	0.38%	99.19%	0
Axe 9	0.1772	3.42%	41.57%	0.31%	99.50%	0
Axe 10	0.1723	3.32%	44.89%	0.17%	99.67%	0
Axe 11	0.1702	3.28%	48.17%	0.13%	99.80%	0
Axe 12	0.1696	3.27%	51.44%	0.12%	99.92%	0
Axe 13	0.1673	3.23%	54.67%	0.08%	100.00%	0
Axe 14	0.1666	3.21%	57.88%	0.00%	100.00%	0
Axe 15	0.1657	3.19%	61.07%	0.00%	100.00%	0
Axe 16	0.1650	3.18%	64.26%	0.00%	100.00%	0
Axe 17	0.1640	3.16%	67.42%	0.00%	100.00%	0
Axe 18	0.1638	3.16%	70.58%	0.00%	100.00%	0
Axe 19	0.1614	3.11%	73.69%	0.00%	100.00%	0
Axe 20	0.1592	3.07%	76.76%	0.00%	100.00%	0
Axe 21	0.1583	3.05%	79.81%	0.00%	100.00%	0
Axe 22	0.1520	2.93%	82.74%	0.00%	100.00%	0
Axe 23	0.1474	2.84%	85.58%	0.00%	100.00%	0
Axe 24	0.1382	2.67%	88.25%	0.00%	100.00%	0
Axe 25	0.1277	2.46%	90.71%	0.00%	100.00%	0
Axe 26	0.1152	2.22%	92.93%	0.00%	100.00%	0
Axe 27	0.1079	2.08%	95.01%	0.00%	100.00%	0
Axe 28	0.0924	1.78%	96.79%	0.00%	100.00%	0
Axe 29	0.0866	1.67%	98.46%	0.00%	100.00%	0
Axe 30	0.0586	1.13%	99.60%	0.00%	100.00%	0
Axe 31	0.0186	0.36%	99.95%	0.00%	100.00%	0
Axe 32	0.0023	0.04%	100.00%	0.00%	100.00%	0
Axe 33	0.0000	0.00%	100.00%	0.00%	100.00%	0
Total	5.1859	100.00%	100.00%	100.00%	100.00%	0



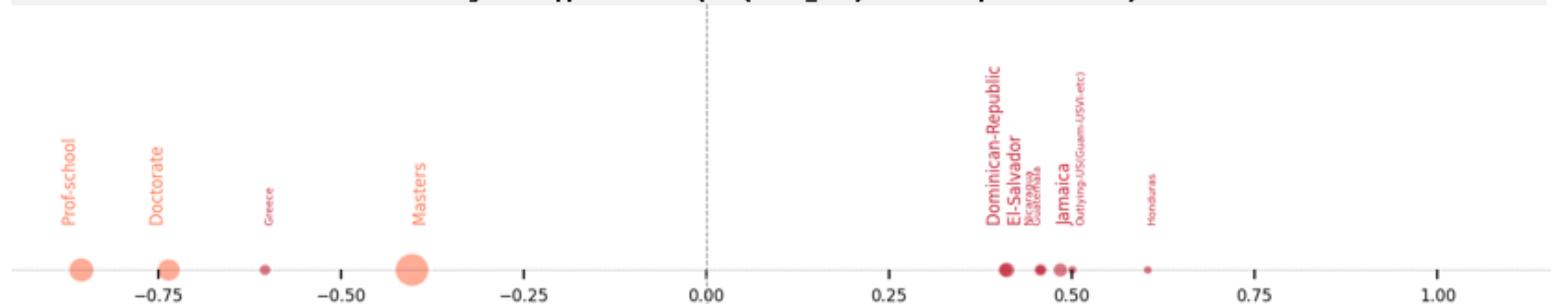
Variable	Categorie	Coord	Ctr
relationship	Husband	-1.09	16.95%
income	>50K	-1.07	9.62%
marital-status	Married-civ-spouse	-0.97	15.16%
gender	Male	-0.48	5.38%

Variable	Categorie	Coord	Ctr
relationship	Unmarried	1.08	4.35%
gender	Female	0.96	10.84%
occupation	Other-service	0.86	2.60%
marital-status	Divorced	0.84	3.41%
occupation	Adm-clerical	0.82	2.71%
relationship	Own-child	0.81	3.63%
marital-status	Never-married	0.78	7.04%
relationship	Not-in-family	0.71	4.54%
income	<=50K	0.34	3.03%

Diagramme des variances des axes - 25 premiers axes



Catégories supplémentaires (abs(effect_size) >= 0.4 AND pvalue <= 0.025)



Variable	Categorie	Coord	effect_size	pvalue
education	Prof-school	-0.86	-0.86	< 0.001
education	Doctorate	-0.74	-0.74	< 0.001
native-country	Greece	-0.60	-0.60	< 0.001
education	Masters	-0.40	-0.40	< 0.001

Variable	Categorie	Coord	effect_size	pvalue
native-country	Honduras	0.60	0.60	0.0035
native-country	Outlying-US(Guam-USVI)	0.50	0.50	0.0083
native-country	Jamaica	0.48	0.48	< 0.001
native-country	Guatemala	0.46	0.46	< 0.001
native-country	Nicaragua	0.46	0.46	< 0.001
native-country	El-Salvador	0.41	0.41	< 0.001
native-country	Dominican-Republic	0.41	0.41	< 0.001

[Retour Menu](#)

Tableaux croisés - Distribution des variables catégorielles au sein des différents clusters

% colonnes	C1 - 11.19%	C2 - 41.73%	C3 - 4.73%	C4 - 6.39%	C5 - 35.96%	Total	Statut
relationship							
	Chi² : 96491.78 - Chi2_pvalue : 0.00e+00 - V de Cramer : 0.70						
Husband	0,0%	8,7%	0,5%	80,4%	87,8%	40,4%	Active
Not-in-family	27,5%	47,1%	0,0%	8,4%	7,0%	25,8%	Active
Other-relative	2,8%	6,0%	0,0%	1,5%	0,5%	3,1%	Active
Own-child	4,8%	32,5%	0,0%	6,8%	2,7%	15,5%	Active
Unmarried	64,6%	5,7%	0,0%	2,7%	2,0%	10,5%	Active
Wife	0,3%	0,0%	99,4%	0,3%	0,0%	4,8%	Active
gender							
	Chi² : 19082.90 - Chi2_pvalue : 0.00e+00 - V de Cramer : 0.63						
Female	79,3%	43,5%	99,4%	2,3%	3,6%	33,2%	Active
Male	20,7%	56,5%	0,6%	97,7%	96,4%	66,8%	Active
marital-status							
	Chi² : 52579.80 - Chi2_pvalue : 0.00e+00 - V de Cramer : 0.52						
Divorced	32,9%	19,8%	0,0%	2,9%	4,1%	13,6%	Active
Married-AF-spouse	0,0%	0,0%	1,6%	0,0%	0,0%	0,1%	Active
Married-civ-spouse	0,5%	10,1%	98,4%	81,4%	88,1%	45,8%	Active
Married-spouse-absent	0,2%	3,0%	0,0%	0,1%	0,0%	1,3%	Active
Never-married	17,5%	66,2%	0,0%	14,3%	6,9%	33,0%	Active
Separated	23,4%	0,8%	0,0%	0,5%	0,4%	3,1%	Active
Widowed	25,5%	0,1%	0,0%	0,8%	0,5%	3,1%	Active
occupation							
	Chi² : 37214.33 - Chi2_pvalue : 0.00e+00 - V de Cramer : 0.44						
?	9,2%	6,9%	8,4%	0,0%	4,0%	5,8%	Passive
Adm-clerical	19,2%	14,0%	22,6%	0,3%	6,7%	11,5%	Active
Armed-Forces	0,0%	0,0%	0,0%	0,0%	0,1%	0,0%	Passive
Craft-repair	7,1%	9,2%	2,2%	17,3%	18,6%	12,5%	Active
Exec-managerial	10,2%	8,3%	14,9%	12,5%	17,7%	12,5%	Active
Farming-fishing	0,3%	0,2%	0,4%	45,5%	0,0%	3,1%	Active
Handlers-cleaners	2,6%	5,9%	1,3%	0,0%	4,0%	4,2%	Active
Machine-op-inspct	2,4%	13,4%	5,1%	0,0%	0,2%	6,2%	Active
Other-service	23,6%	12,6%	11,7%	2,0%	4,2%	10,1%	Active
Priv-house-serv	4,4%	0,0%	0,0%	0,0%	0,0%	0,5%	Active
Prof-specialty	8,3%	11,8%	20,4%	10,3%	14,4%	12,6%	Active
Protective-serv	0,2%	0,1%	0,3%	0,0%	5,4%	2,0%	Active
Sales	11,1%	11,5%	8,5%	12,1%	11,3%	11,3%	Active
Tech-support	0,5%	6,1%	3,0%	0,0%	0,6%	3,0%	Active
Transport-moving	1,0%	0,2%	1,1%	0,0%	12,8%	4,8%	Active
income							
	Chi² : 8167.95 - Chi2_pvalue : 0.00e+00 - V de Cramer : 0.41						
<=50K	94,6%	91,4%	52,7%	72,9%	56,2%	76,1%	Active
>50K	5,4%	8,6%	47,3%	27,1%	43,8%	23,9%	Active

TRAITS CARACTÉRISTIQUES DES SEGMENTS

	Variable	Modalité	% de la modalité dans la Classe	% de la modalité dans l'échantillon total	Lift	% de la Classe dans la modalité	H de Cohen	Taille de l'effet (H de Cohen)	pvalue
C1	relationship	Unmarried	64,07%	10,49%	6,11	79,80%	1,20		0,0000
C1	gender	Female	72,49%	33,15%	2,19	28,58%	0,81		0,0000
C1	marital-status	Separated	23,29%	3,13%	7,44	97,19%	0,65		0,0000
C1	marital-status	Widowed	23,23%	3,11%	7,47	97,69%	0,65		0,0000
C1	income	<=50K	93,64%	76,07%	1,23	16,09%	0,51		0,0000
C1	marital-status	Divorced	31,11%	13,58%	2,29	29,94%	0,43		0,0000
C1	occupation	Other-service	23,64%	10,08%	2,35	30,65%	0,37		0,0000
C1	occupation	Adm-clerical	18,39%	11,49%	1,60	20,92%	0,19		0,0000
C1	workclass	Private	76,63%	69,42%	1,10	14,43%	0,16		0,0000
C1	education	HS-grad	39,74%	32,32%	1,23	16,07%	0,15		0,0000
C1	occupation	Priv-house-serv	1,57%	0,50%	3,16	41,32%	0,11		0,0000
C1	occupation	Exec-managerial	9,34%	12,46%	0,75	9,79%	-0,10		0,0000
C1	education	Prof-school	0,55%	1,71%	0,32	4,20%	-0,11		0,0000
C1	education	Masters	3,05%	5,44%	0,56	7,34%	-0,12		0,0000
C1	occupation	Prof-specialty	8,47%	12,64%	0,67	8,77%	-0,14		0,0000
C1	workclass	Self-emp-not-inc	4,61%	7,91%	0,58	7,61%	-0,14		0,0000
C1	workclass	Self-emp-inc	1,27%	3,47%	0,37	4,78%	-0,15		0,0000
C1	occupation	Craft-repair	7,93%	12,51%	0,63	8,28%	-0,15		0,0000
C1	occupation	Machine-op-inspct	2,10%	6,19%	0,34	4,43%	-0,21		0,0000
C1	education	Bachelors	9,26%	16,43%	0,56	7,36%	-0,22		0,0000
C1	marital-status	Never-married	19,14%	33,00%	0,58	7,58%	-0,32		0,0000
C1	relationship	Wife	0,36%	4,77%	0,08	0,99%	-0,32		0,0000
C1	relationship	Own-child	3,57%	15,52%	0,23	3,01%	-0,43		0,0000
C1	income	>50K	6,36%	23,93%	0,27	3,47%	-0,51		0,0000
C1	gender	Male	27,51%	66,85%	0,41	5,38%	-0,81		0,0000
C1	relationship	Husband	0,19%	40,37%	0,00	0,06%	-1,29		0,0000
C1	marital-status	Married-civ-spous	0,00%	45,82%	0,00	0,00%	-1,49		0,0000

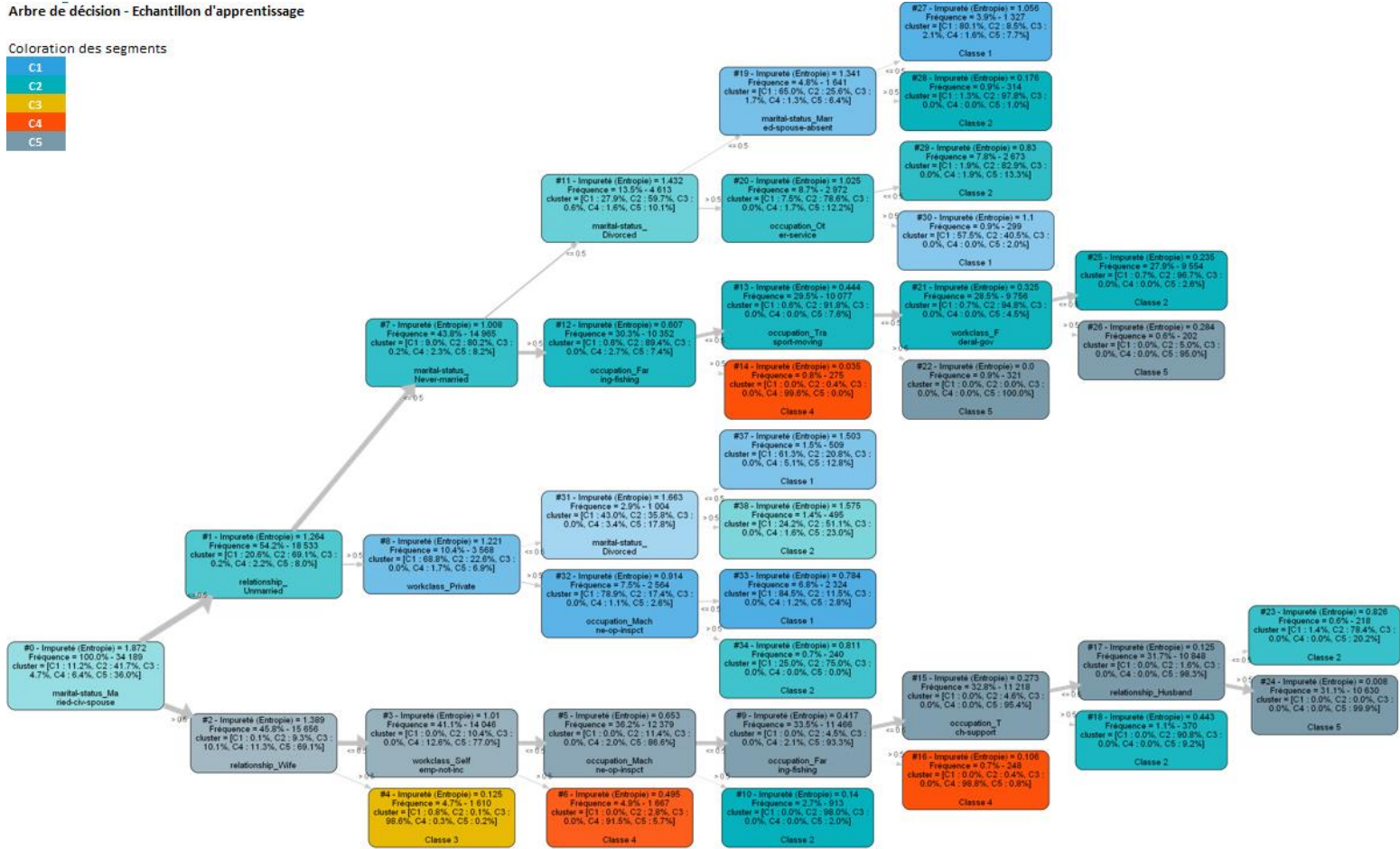
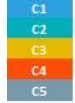
	Variable	Modalité	% de la modalité dans la Classe	% de la modalité dans l'échantillon total	Lift	% de la Classe dans la modalité	H de Cohen	Taille de l'effet (H de Cohen)	pvalue
C3	relationship	Wife	100,00%	4,77%	20,95	99,01%	2,70		0,0000
C3	gender	Female	99,87%	33,15%	3,01	14,24%	1,84		0,0000
C3	marital-status	Married-civ-spous	100,00%	45,82%	2,18	10,31%	1,65		0,0000
C3	income	>50K	46,92%	23,93%	1,96	9,27%	0,49		0,0000
C3	occupation	Adm-clerical	22,44%	11,49%	1,95	9,23%	0,30		0,0000
C3	occupation	Prof-specialty	20,23%	12,64%	1,60	7,57%	0,21		0,0000
C3	workclass	Local-gov	9,79%	6,42%	1,53	7,21%	0,12		0,0000
C3	education	11th	1,91%	3,71%	0,51	2,43%	-0,11		0,0000
C3	workclass	Private	63,69%	69,42%	0,92	4,34%	-0,12		0,0000
C3	occupation	Protective-serv	0,43%	2,01%	0,22	1,02%	-0,15		0,0000
C3	occupation	Farming-fishing	0,78%	3,05%	0,26	1,21%	-0,17		0,0000
C3	occupation	Handlers-cleaners	1,21%	4,24%	0,29	1,35%	-0,19		0,0000
C3	marital-status	Married-spouse-at	0,00%	1,29%	0,00	0,00%	-0,23		0,0000
C3	occupation	Transport-moving	1,04%	4,82%	0,22	1,02%	-0,24		0,0000
C3	relationship	Other-relative	0,00%	3,08%	0,00	0,00%	-0,35		0,0000
C3	marital-status	Widowed	0,00%	3,11%	0,00	0,00%	-0,35		0,0000
C3	marital-status	Separated	0,00%	3,13%	0,00	0,00%	-0,36		0,0000
C3	occupation	Craft-repair	2,04%	12,51%	0,16	0,77%	-0,44		0,0000
C3	income	<=50K	53,08%	76,07%	0,70	3,30%	-0,49		0,0000
C3	relationship	Unmarried	0,00%	10,49%	0,00	0,00%	-0,66		0,0000
C3	marital-status	Divorced	0,00%	13,58%	0,00	0,00%	-0,75		0,0000
C3	relationship	Own-child	0,00%	15,52%	0,00	0,00%	-0,81		0,0000
C3	relationship	Not-in-family	0,00%	25,76%	0,00	0,00%	-1,06		0,0000
C3	marital-status	Never-married	0,00%	33,00%	0,00	0,00%	-1,22		0,0000
C3	relationship	Husband	0,00%	40,37%	0,00	0,00%	-1,38		0,0000
C3	gender	Male	0,13%	66,85%	0,00	0,01%	-1,84		0,0000

EXEMPLE DE DÉPLOIEMENT PAR ARBRE DE DÉCISION

Modèle Tree - Matrices de confusion et métriques associées

Arbre de décision - Echantillon d'apprentissage

Coloration des segments



Matrice de confusion - Echantillon Test

		Prédit					Total
		C1	C2	C3	C4	C5	
Observé	Effectifs						
	C1	1 518	118	4	0	0	1 640
	C2	276	5 815	2	19	3	6 115
	C3	9	0	684	0	0	693
	C4	31	21	4	880	0	936
C5	91	340	4	52	4 782	5 269	
Total	1 925	6 294	698	951	4 785	14 653	

Métriques						Weighted average
accuracy	93,35%					
precision	78,86%	92,39%	97,99%	92,53%	99,94%	93,86%
recall	92,56%	95,09%	98,70%	94,02%	90,76%	93,35%
f1-score	85,16%	93,72%	98,35%	93,27%	95,13%	93,46%
support	1 640	6 115	693	936	5 269	14 653

Matrice de confusion - Echantillon d'Apprentissage

		Prédit					Total
		C1	C2	C3	C4	C5	
Observé	Effectifs						
	C1	3 511	301	13	0	2	3 827
	C2	607	13 596	1	48	15	14 267
	C3	28	0	1 588	0	0	1 616
	C4	74	62	5	2 045	0	2 186
C5	239	818	3	97	11 136	12 293	
Total	4 459	14 777	1 610	2 190	11 153	34 189	

Métriques						Weighted average
accuracy	93,23%					
precision	78,74%	92,01%	98,63%	93,38%	99,85%	93,74%
recall	91,74%	95,30%	98,27%	93,55%	90,59%	93,23%
f1-score	84,75%	93,62%	98,45%	93,46%	94,99%	93,34%
support	3 827	14 267	1 616	2 186	12 293	34 189



MICROPOL_{AI}

SEGMENTATION.AI

DEMO



**MERCI POUR
VOTRE
ATTENTION**

**→ DES
QUESTIONS ?**





MICROPOL 

Tanguy LE NOUVEL
Directeur Data Science

M +33 (0) 6 81 95 84 22
@ tlenouvel@micropole.com

WWW.MICROPOLE.COM
91-95 RUE CARNOT - 92 300 LEVALLOIS-PERRET



MICROPOL 
Data driving transformation

