

Analyse Multivariée du Dataset UCI Mushroom

ACM, Clustering et Analyse Discriminante

Pierre Chambet

Maxime Adam

Janvier 2026

Télécom SudParis
Master TRIED

Abstract

Ce rapport présente une analyse du dataset UCI Mushroom (8 124 champignons, 23 variables qualitatives) pour discriminer champignons comestibles et vénéneux. Nous appliquons une démarche structurée : Analyse des Correspondances Multiples (ACM), clustering sur composantes factorielles, et analyse discriminante. L'ACM révèle que les caractéristiques d'odeur et de surface constituent les axes principaux de variation (31,3% d'inertie cumulée sur 5 axes).

Mots-clés : ACM, Classification non supervisée, Analyse discriminante, Données qualitatives

Contents

1	Introduction	3
1.1	Démarche analytique	3
2	Données et préparation	3
2.1	Description du dataset	3
2.2	Preprocessing	4
2.3	Statistiques descriptives	4
3	Analyse des Correspondances Multiples (ACM)	5
3.1	Méthodologie	5
3.2	Choix du nombre d'axes	5
3.3	Interprétation des axes factoriels	5
3.3.1	Axe 1 (7,59%) : "Surface et Odeur"	5
3.3.2	Axe 2 (6,91%) : "Modalités rares"	6
3.4	Visualisations et interprétation spatiale	6
3.5	Export	8
4	Clustering sur composantes ACM	8
4.1	Méthode	8
4.2	Résultats	8
5	Analyse discriminante	8
5.1	Modèle	8
5.2	Performance	8
6	Conclusion	8

1 Introduction

Le dataset UCI Mushroom regroupe 8 124 champignons décrits par 23 variables morphologiques qualitatives (forme du chapeau, odeur, couleur des lamelles, etc.). L'objectif est d'identifier les profils-types et les caractéristiques discriminantes pour la comestibilité.

1.1 Démarche analytique

Nous appliquons une méthodologie articulée en trois étapes :

1. **ACM** : Réduction de dimensionnalité (23 variables \rightarrow 5 axes factoriels, 31,3% d'inertie)
2. **Clustering** : Segmentation non supervisée sur composantes (CAH, K-means)
3. **Analyse discriminante** : Modélisation prédictive edible/poisonous sur facteurs ACM

Cette approche permet de combiner exploration et prédiction tout en valorisant la nature qualitative des données.

2 Données et préparation

2.1 Description du dataset

Source : UCI Machine Learning Repository (*Audubon Society Field Guide*, 1981)

Dimensions : $n = 8\,124$ champignons, $p = 23$ variables qualitatives, $K = 111$ modalités totales

Variable cible : `class` $\in \{e \text{ (edible)}, p \text{ (poisonous)}\}$, distribution équilibrée (51,8% vs. 48,2%)

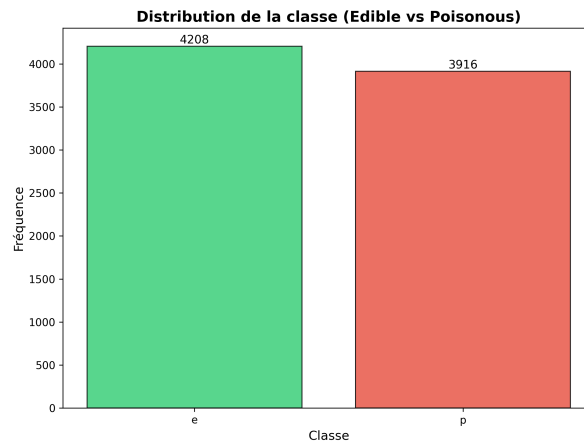


Figure 1: Distribution de la classe

2.2 Preprocessing

Valeurs manquantes : La variable `stalk-root` contient 2 480 valeurs "?" (30,5%).
Stratégie : imputation modale (modalité "b" = bulbous). Justification : préserve la distribution, évite la perte de 30% des données, compatible ACM.

2.3 Statistiques descriptives

Le Tableau 1 présente les 6 variables clés.

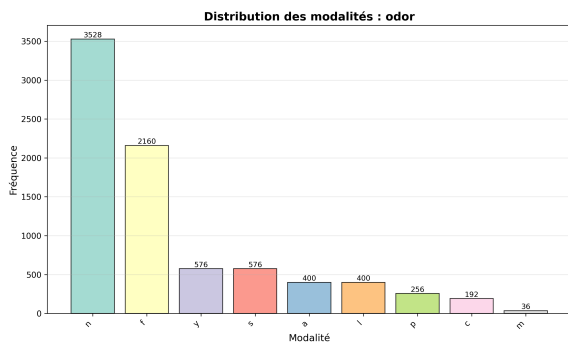
Table 1: Variables principales (top 6)

Variable	N_{mod}	Top modalité	Freq. (%)
class	2	e (edible)	51,8
odor	9	n (none)	43,4
gill-color	12	b (buff)	21,3
spore-print-color	9	w (white)	29,4
cap-color	10	n (brown)	28,1
gill-attachment	2	f (free)	97,4

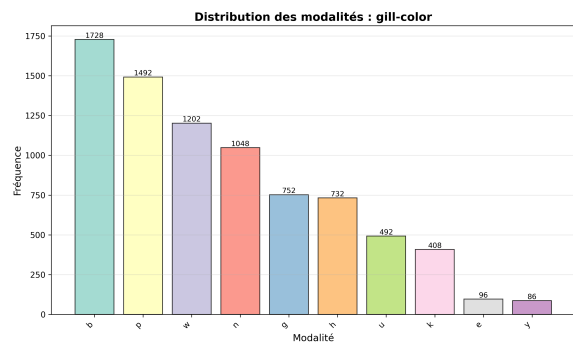
Analyse bivariée odeur \times classe (Tableau 2) : association quasi-parfaite. Les odeurs agréables (almond, anise) sont 100% comestibles ; les odeurs fétides (foul, pungent) sont 100% vénéneuses. Variable hautement discriminante.

Table 2: Tableau croisé odeur \times classe (extrait)

Odeur	Comestible	Vénéneux	Total
none (n)	3 408	120	3 528
foul (f)	0	2 160	2 160
almond (a)	400	0	400
anise (l)	400	0	400
pungent (p)	0	256	256



(a) Odeur



(b) Couleur lamelles

Figure 2: Distributions des modalités clés

3 Analyse des Correspondances Multiples (ACM)

3.1 Méthodologie

L'ACM transforme les 22 variables descriptives (111 modalités) en axes factoriels orthogonaux via le Tableau Disjonctif Complet (TDC). Inertie totale : $I_{tot} \approx 4.27$.

3.2 Choix du nombre d'axes

Nous conservons **k = 5 axes** (31,3% d'inertie cumulée). Justification : coude visible après l'axe 5 (Fig. 3), compromis interprétabilité/information.

Table 3: Valeurs propres et inerties

Axe	λ	Inertie (%)	Cumul (%)
Dim1	0,324	7,59	7,59
Dim2	0,295	6,91	14,49
Dim3	0,271	6,33	20,83
Dim4	0,243	5,68	26,51
Dim5	0,203	4,76	31,27

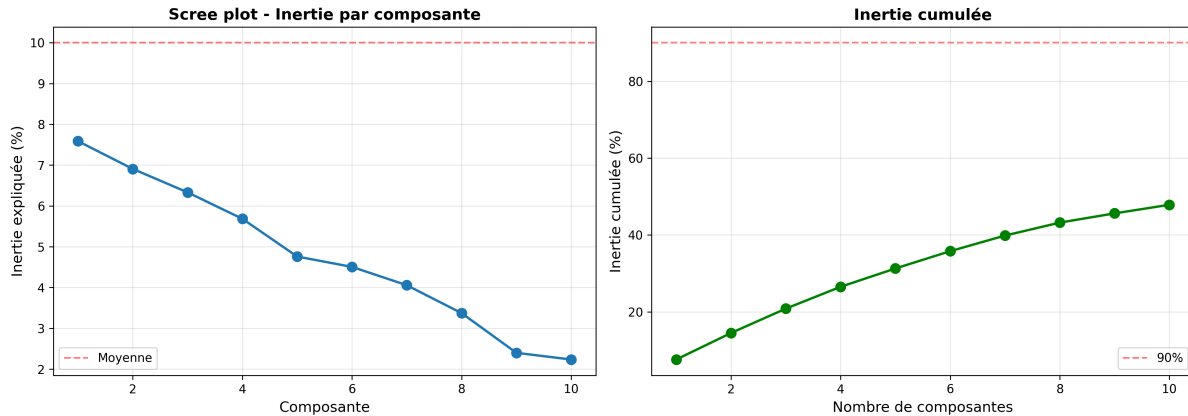


Figure 3: Scree plot et inertie cumulée

3.3 Interprétation des axes factoriels

3.3.1 Axe 1 (7,59%) : "Surface et Odeur"

Top contributions (Table 4) : `ring-type__1` (anneau large, 6,68%), `stalk-surface-*__k` (surface soyeuse, 6,4%), `odor__f` (odeur fétide, 5,49%).

Table 4: Top 5 contributions axe 1

Modalité	Coord.	Contrib. (%)
ring-type__l (large)	+1,73	6,68
stalk-surface-below-ring__k	+1,27	6,41
odor__f (foul)	+1,21	5,49
ring-type__p (pendant)	-0,67	3,05
odor__n (no odor)	-0,62	2,36

Interprétation : Axe oppose champignons à texture lisse + odeur forte (pôle +, majoritairement vénéneux) vs. champignons sans odeur + anneau pendant (pôle -, neutres). Pouvoir discriminant fort.

3.3.2 Axe 2 (6,91%) : "Modalités rares"

Top contributions : gill-attachment__a (8,7%, effectif 3%), stalk-color-*__o (7,2%, effectif <1%).

Interprétation : Effet de taille (modalités rares éloignées du barycentre). Oppose champignons atypiques vs. "moyens". Moins discriminant pour la classe, utile pour identifier sous-groupes.

3.4 Visualisations et interprétation spatiale

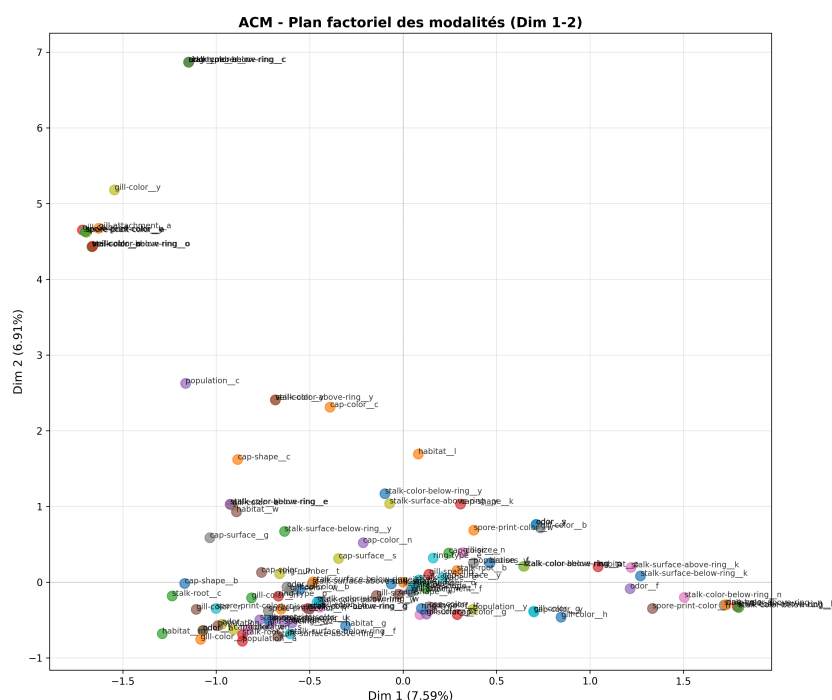


Figure 4: Plan factoriel des modalités (axes 1-2)

Analyse spatiale des modalités (Fig. 4) :

- **Dispersion axe 1** : Opposition claire entre modalités à gauche (`odor__n`, `ring-type__p`) et à droite (`odor__f`, `ring-type__l`), confirmant l'interprétation "surface + odeur"
- **Modalités excentrées axe 2** : Les modalités rares (`gill-attachment__a`, `gill-color__y`) sont très éloignées du centre, illustrant l'effet de taille
- **Centre de gravité** : Les modalités fréquentes se concentrent autour de l'origine (`cap-shape__x`, `gill-attachment__f`)

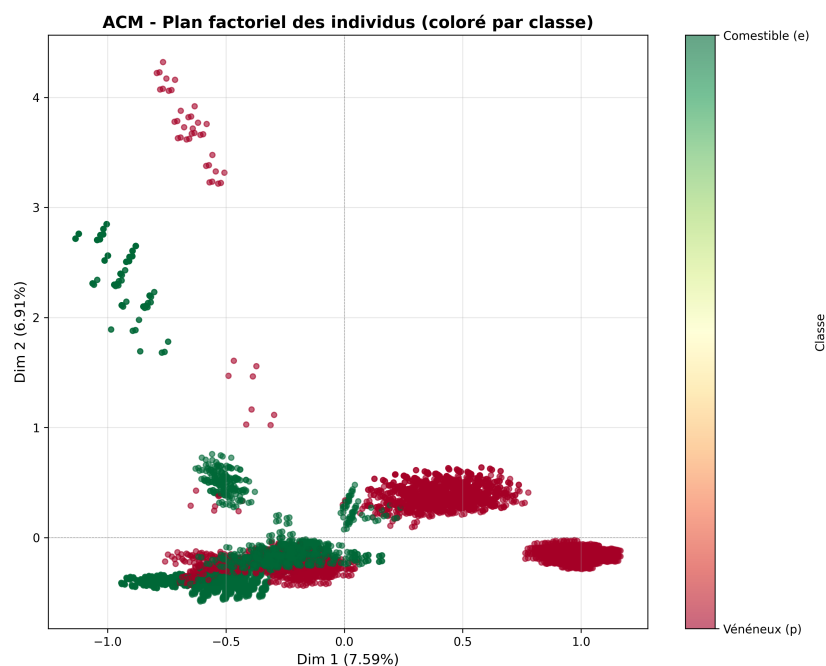


Figure 5: Plan factoriel des individus (axes 1-2), colorés par classe

Analyse de la séparation des classes (Fig. 5) :

- **Tendance de séparation** : Les champignons vénéneux (rouge) se concentrent plutôt à droite de l'axe 1, les comestibles (vert) à gauche, cohérent avec les contributions (`odor__f` vs. `odor__n`)
- **Chevauchement important** : Les nuages se superposent fortement, indiquant que les axes 1-2 seuls (14,5% inertie) ne suffisent pas pour une discrimination parfaite
- **Potentiel discriminant** : L'utilisation des 5 premiers axes (31,3% inertie cumulée) dans l'analyse discriminante (Section 5) devrait significativement améliorer la séparation

Conclusion ACM : La réduction de dimensionnalité est effective (22 variables → 5 axes interprétables), avec un axe 1 montrant un fort pouvoir discriminant. La suite de l'analyse (clustering, discriminante) exploitera ces composantes pour modéliser la comestibilité.

3.5 Export

Fichiers générés : `mca_coords.csv` ($8\,124 \times 10$ coordonnées), `mca_eigenvalues.csv`, figures. Recommandation : utiliser $k=5$ axes pour clustering et analyse discriminante.

4 Clustering sur composantes ACM

[Section rédigée par Personne B]

4.1 Méthode

CAH et K-means sur coordonnées factorielles ($k=5$ axes).

4.2 Résultats

[Choix nombre de clusters, dendrogramme, profils]

5 Analyse discriminante

[Section rédigée par Personne B]

5.1 Modèle

LDA sur composantes ACM.

5.2 Performance

[Matrice de confusion, taux de succès, validation croisée]

6 Conclusion

L'ACM a révélé que les caractéristiques de surface et d'odeur constituent les axes principaux de variation (31,3% d'inertie sur 5 axes). La variable `odor` présente une association quasi-parfaite avec la classe edible/poisonous, confirmée par l'axe 1. *[À compléter avec résultats clustering et discriminante].*

Limites : Inertie expliquée modérée (typique ACM), certaines modalités rares génèrent des effets de taille.

Perspectives : Comparer avec Random Forest, tester sur autres datasets mycologiques.