Méthodes statistiques pour Les données qualitatives

Rapport de projet

TRIED 2024 / 2025

Soumis à: Pr. N'deye Niang Keita

Soumis par :

DERRAR Achraf Nedjmeddine

JAI Ilyass

Lien: https://github.com/achrafndd05/DQ_projet

Introduction

L’analyse des données joue un rôle fondamental dans la compréhension et l’interprétation des phénomènes étudiés dans de nombreux domaines. Ce projet s’inscrit dans cette dynamique en proposant une exploration approfondie d’un jeu de données choisi, à travers des méthodes statistiques avancées.

L’objectif de cette étude est de mettre en lumière les relations entre les différentes variables grâce à une approche combinant des outils de statistique descriptive, une analyse factorielle des correspondances multiples (ACM), une classification non supervisée et une analyse discriminante. Ces techniques permettront d’identifier les structures sous-jacentes des données et d’en extraire des informations pertinentes.

Ce rapport détaille la démarche méthodologique adoptée, en exposant les résultats obtenus ainsi que leur interprétation. L’analyse réalisée vise à fournir une meilleure compréhension des données et à dégager des tendances significatives, facilitant ainsi la prise de décision et l’approfondissement des connaissances sur le sujet étudié.

Présentation du jeu de données

Le jeu de données utilisé dans ce projet, qui est intitulé “Student Performance”. Les données contiennent des informations sur **395 élèves** de deux écoles secondaires portugaises. Les notes des élèves, ainsi que les caractéristiques démographiques, sociales et scolaires ont été collectées à l'aide de rapports et de questionnaires scolaires. Les données sont réparties entre deux fichiers de données pour le cours de mathématiques et le cours de portugais respectivement ; comme ils contiennent les mêmes variables pour les mêmes élèves, nous limitons notre analyse à l'ensemble de données du cours de mathématiques uniquement. Les données peuvent être téléchargées ici : https://archive.ics.uci.edu/ml/datasets/student+performance. Les données contiennent 33 variables qualitatives et quantitatives :

Qualitative (categorical) :

1. school - student's school (binary: "GP" - Gabriel Pereira or "MS" - Mousinho da Silveira)

2. sex - student's sex (binary: "F" - female or "M" - male)

3. address - student's home address type (binary: "U" - urban or "R" - rural)

4. famsize - family size (binary: "LE3" - less or equal to 3 or "GT3" - greater than 3)

5. Pstatus - parent's cohabitation status (binary: "T" - living together or "A" - apart)

6. Mjob - mother's job (nominal: "teacher", "health" care related, civil "services" (e.g. administrative or police), "at_home" or "other")

7. Fjob - father's job (nominal: "teacher", "health" care related, civil "services" (e.g. administrative or police), "at_home" or "other")

8. reason - reason to choose this school (nominal: close to "home", school "reputation", "course" preference or "other")

9. guardian - student's guardian (nominal: "mother", "father" or "other")

10. schoolsup - extra educational support (binary: yes or no)

11. famsup - family educational support (binary:yes or no)

12. paid - extra paid classes within the course subject (binary: yes or no)

13. activities - extra-curricular activities(binary:yes or no)

14. nursery-attended nursery school(binary:yes or no)

15. higher - wants to take higher education(binary: yes or no)

16. internet - Internet access at home(binary: yes or no)

17. romantic - with a romantic relationship (binary: yes or no)

Qualitative (Numeric) :

18. Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade,

3 – secondary education or 4 – higher education)

19. Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade,

3 – secondary education or 4 – higher education)

20. traveltime-hometoschooltraveltime(numeric:1-<15min.,2-15to30min.,3-30min.to1hour,or4->1hour)

21. studytime-weekly study time (numeric:1- < 2 hours ,2- 2to5hours,3- 5 to 10 hours, or4- >10hours)

22. famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)

23. freetime-free time after school (numeric: from 1- very low to 5- very high)

24. goout- going out with friends(numeric:from 1- very low to 5- very high)

25. Dalc-work day alcohol consumption(numeric: from 1- very low to 5- very high)

26. Walc-weekend alcohol consumption(numeric:from 1- very low to 5- very high)

27. health-current health status(numeric:from 1-very bad to 5- very good)

Quantitative (numeric)

28. age - student' sage (numeric: from 15 to 22)

29. failures - number of past class failures(numeric: n if 1<=n<3 , else 4)

30. absences - number of school absences (numeric: from 0 to 93)

31. G1-first period grade for maths(numeric: from 0 to 20)

32. G2 - second period grade for maths (numeric: from 0 to 20)

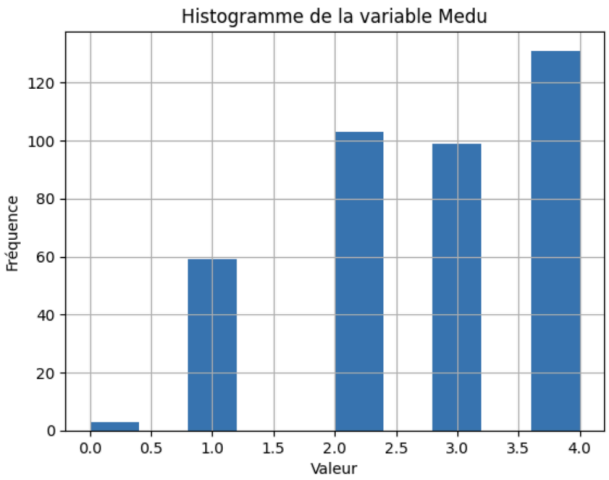
33. G3 - final grade for maths (numeric: from 0 to 20, output target)-**this is our target variable**

Analyse unidimensionnelle:

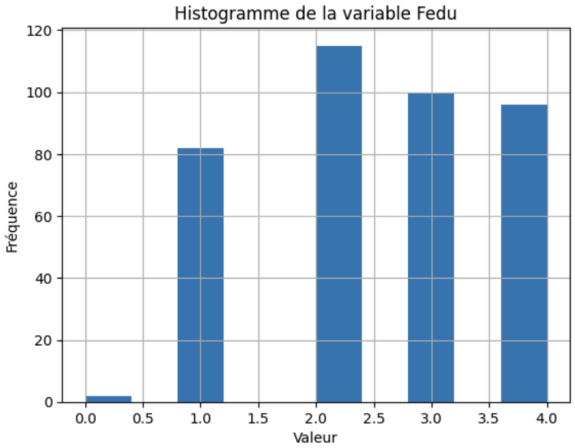
- The following categorical variables were created using K-means :

1. G3 - final grade for maths (nominal: “<10”, “>=10”, output target)
2. Age - student's age (nominal: “15 to 15”, “16 to 16”, “17 to 17”, “18 to 22”)
3. Absences - number of school absences (nominal: “0 to 4”, “5 to 13”, “14 to 75”)
4. failures - number of past class failures (nominal: zero, non-zero)

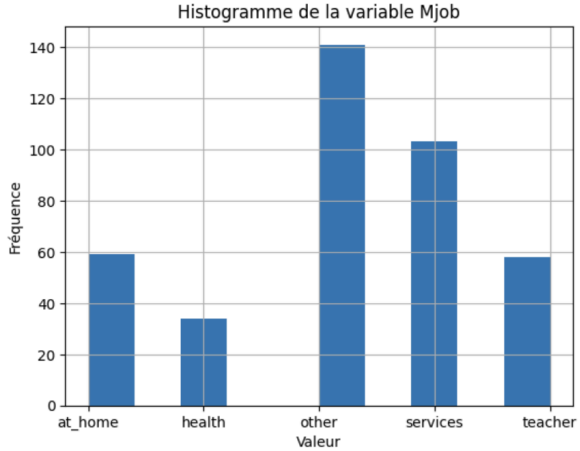
Nous allons présenter les variables que nous avons identifiées comme pertinentes. Les autres figures sont disponibles en annexe:



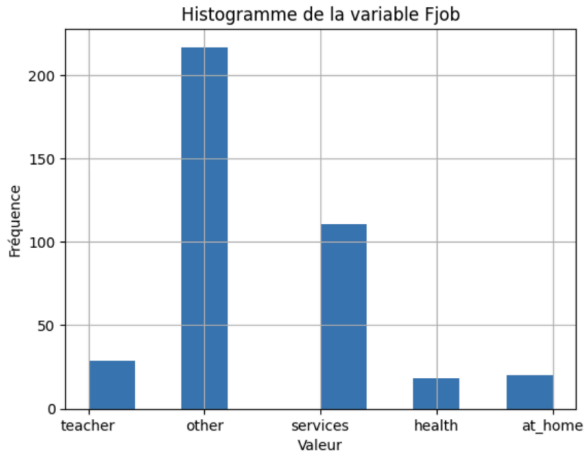
On remarque que la plupart des mères ont un niveau d'éducation élevé, avec 131 ayant fait des études supérieures (Medu = 4) et 99 ayant atteint le secondaire (Medu = 3). Le niveau intermédiaire (Medu = 2) est aussi bien représenté avec 103 cas. En revanche, seules 59 mères ont arrêté à l'école primaire (Medu = 1) et très peu, seulement 3, n'ont reçu aucune éducation (Medu = 0). Globalement, ces chiffres montrent que la majorité des mères ont un bon niveau d'instruction, ce qui peut avoir un impact positif sur la scolarité de leurs enfants



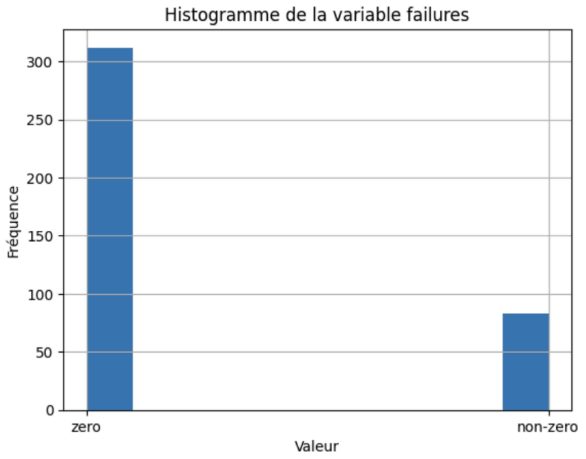
On observe une répartition assez équilibrée entre les différents niveaux. Le niveau intermédiaire (Fedu = 2) est le plus fréquent avec 115 cas, suivi de près par le secondaire (Fedu = 3) avec 100 cas et l'enseignement supérieur (Fedu = 4) avec 96 cas. En revanche, 82 pères se sont arrêtés à l'école primaire (Fedu = 1), et seulement 2 n'ont reçu aucune éducation (Fedu = 0), ce qui peut avoir un impact positif sur la scolarité de leurs enfants.



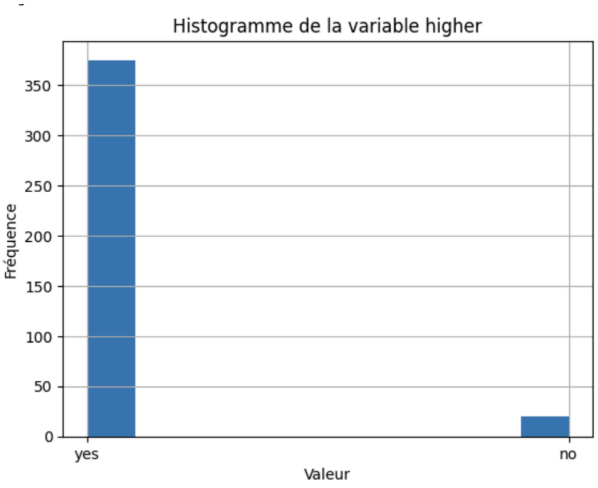
La majorité des mères occupent un emploi dans la catégorie "autre" (141), mais on ne sait pas précisément de quel type d'emploi il s'agit. Viennent ensuite celles travaillant dans les "services" (103), 59 mères sont au foyer, 58 sont enseignantes, et seulement 34 travaillent dans le secteur de la santé. En somme, la plupart des mères ont des emplois peu spécifiques.



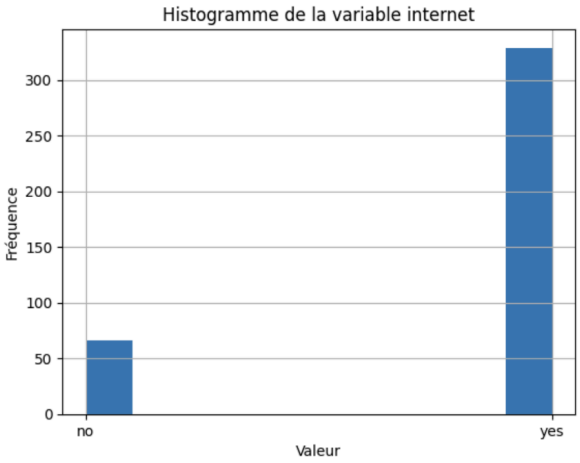
La majorité des pères occupent un emploi dans la catégorie "autre" (217), mais on ne sait pas précisément de quel type d'emploi il s'agit. Viennent ensuite ceux travaillant dans les "services" (111), 29 pères sont enseignants, 20 sont au foyer, et seulement 18 travaillent dans le secteur de la santé. En somme, la plupart des pères ont des emplois peu spécifiques.



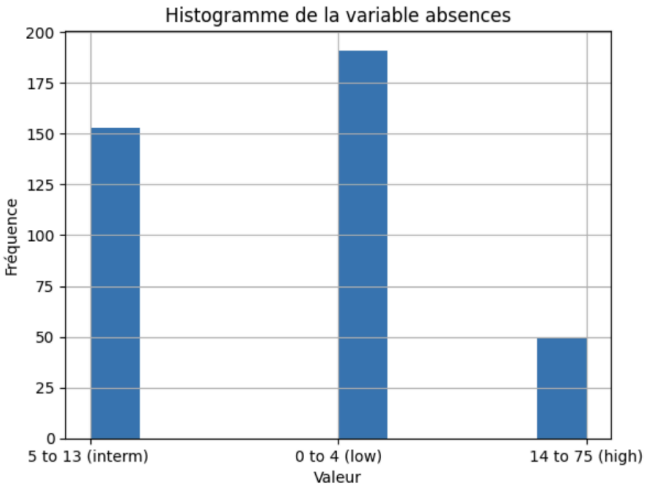
La majorité des individus n'ont eu aucune défaillance scolaire, avec 312 cas de "zero". En revanche, 83 personnes ont eu des échecs scolaires (catégorie "non-zero"). Cette dernière catégorie regroupe toutes les personnes ayant eu au moins un échec, avec un nombre de défaillances compris entre 1 et 4.



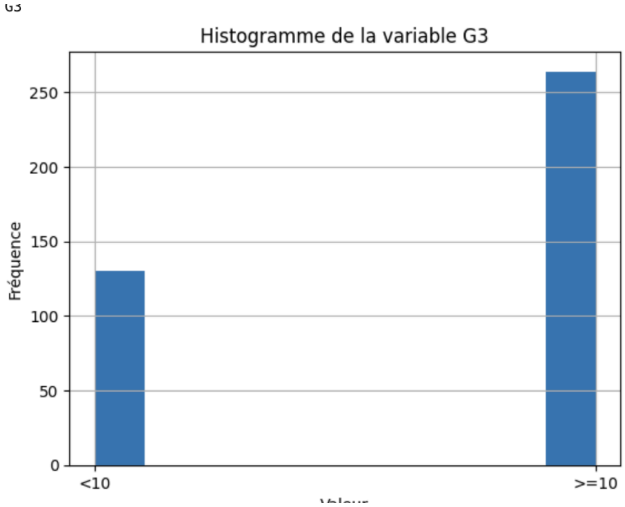
La majorité des individus (375) souhaitent poursuivre des études supérieures, tandis que seulement 20 ne souhaitent pas en faire.



La majorité des individus (329) ont accès à Internet à domicile, tandis que 66 n'en disposent pas. Cela peut influencer leur niveau d'études, car l'accès à Internet facilite l'apprentissage en ligne, la recherche d'informations et l'accès à des ressources éducatives.



La majorité des individus ont un nombre d'absences faible, avec 191 cas dans la catégorie "0 à 4 absences". Viennent ensuite ceux ayant un nombre d'absences intermédiaire, entre 5 et 13, avec 153 cas. Enfin, 50 individus ont un nombre d'absences élevé, compris entre 14 et 75. Un nombre élevé d'absences peut avoir un impact négatif sur les performances scolaires et l'engagement des élèves.



La majorité des individus ont obtenu une note finale en mathématiques supérieure ou égale à 10, avec 264 cas. En revanche, 130 individus ont obtenu une note inférieure à 10. Les résultats suggèrent que la plupart des élèves réussissent, mais une proportion significative a encore des difficultés.

Analyse bidimensionnelle

L’analyse statistique des relations entre les différentes catégories met en évidence plusieurs associations significatives. Nous nous basons sur les résultats du test du chi2, du V de Cramer et de la p-value pour interpréter ces relations.

- **G2 et G3 :** La corrélation entre G2 (deuxième période de note) et G3 (note finale) est très forte. Le chi2 est de 288.60 et le V-Cramer est de 0.86, ce qui montre que ces deux variables sont fortement liées. Comme G1 et G2 sont trop liées à G3, on les a exclues de l’analyse.

- **G1 et G3 :** G1 (première période de note) est aussi fortement liée à G3, avec un chi2 de 210.61 et un V-Cramer de 0.73. Comme pour G2, cette variable a été exclue.

- **Nombre d'échecs passés (failures) et G3 :** Le nombre d'échecs passés est fortement lié à G3, avec un chi2 de 41.83 et un V-Cramer de 0.33. Cela montre que plus un étudiant a d'échecs avant, plus ses notes finales (G3) seront probablement basses.

- **Sorties (goout) et G3 :** La variable sorties (sortir avec des amis) est moins liée à G3. Le chi2 est de 15.53 et le V-Cramer est de 0.20. Bien que cette relation soit significative (p-value = 0.0037), son impact reste faible sur les résultats.

- **Âge (age) et G3 :** L'âge des étudiants a une corrélation modérée avec G3, avec un chi2 de 10.83 et un V-Cramer de 0.17. Cela signifie que l'âge a un effet, mais pas aussi fort que d'autres facteurs.

- **Absences et G3 :** Le nombre d'absences est également fortement lié à G3, avec un chi2 de 9.90 et un V-Cramer de 0.16. Cela montre que plus un étudiant s'absente, plus ses résultats risquent de diminuer.

- **Niveau d'éducation de la mère (Medu) et G3 :** Le niveau d'éducation de la mère montre une faible corrélation avec G3, avec un chi2 de 6.18 et un V-Cramer de 0.13. L'éducation de la mère n'a donc pas un impact aussi fort que d'autres variables sur les résultats.

- **Niveau d'éducation du père (Fedu) et G3 :** De la même façon, le niveau d'éducation du père a une faible corrélation avec G3, avec un chi2 de 7.25 et un V-Cramer de 0.14.

- **higher (Souhait de poursuivre des études supérieures) et G3 :** Le fait qu'un étudiant souhaite poursuivre des études supérieures est modérément lié à G3, avec un chi2 de 9.76 et un V-Cramer de 0.16. Cette relation est statistiquement significative (p-value = 0.0018), ce qui suggère que les étudiants qui veulent poursuivre leurs études ont des résultats un peu meilleurs. Cependant, l'impact reste modéré comparé à d'autres facteurs.

Autres variables avec faible corrélation: comme **le travail de la mère (Mjob)**, **le travail du père(Fjob)**, **le choix de l'école**, **les activités extra-scolaires**, **le soutien éducatif**, **les relations familiales**, **le temps libre**, etc., montrent des corrélations faibles ou non significatives avec G3.

	Variable 1	Variable 2	Chi2	V-Cramer	p-value
31	G2	G3	288.604114	0.855861	< 0.0001
30	G1	G3	210.605213	0.731116	< 0.0001
14	failures	G3	41.828744	0.325829	< 0.0001
25	goout	G3	15.527196	0.198517	0.0037
2	age	G3	10.832018	0.166019	0.0127
29	absences	G3	9.901979	0.158732	0.0194
20	higher	G3	9.761971	0.157406	0.0018
7	Fedu	G3	7.246812	0.135621	0.1234
6	Medu	G3	6.180208	0.125243	0.1861
26	Dalc	G3	5.918365	0.122561	0.2053
11	guardian	G3	5.302843	0.116013	0.0706
8	Mjob	G3	5.002695	0.112682	0.2870
10	reason	G3	4.345124	0.105015	0.2265
24	freetime	G3	3.952767	0.100162	0.4124
15	schoolsup	G3	3.881609	0.099256	0.0488
22	romantic	G3	3.676549	0.096599	0.0552
17	paid	G3	3.515581	0.094461	0.0608
28	health	G3	3.432628	0.093339	0.4882
13	studytime	G3	3.235196	0.090615	0.3568
9	Fjob	G3	2.179279	0.074372	0.7028
23	famrel	G3	2.032893	0.071831	0.7297
1	sex	G3	1.869477	0.068883	0.1715
27	Walc	G3	1.597921	0.063684	0.8092
21	internet	G3	1.468378	0.061048	0.2256
16	famsup	G3	1.286122	0.057134	0.2568
3	address	G3	1.040176	0.051381	0.3078
12	traveltime	G3	1.036539	0.051291	0.7924
5	Pstatus	G3	0.786881	0.044690	0.3750
4	famsize	G3	0.729300	0.043023	0.3931
0	school	G3	0.369710	0.030632	0.5432
18	activities	G3	0.045004	0.010687	0.8320
19	nursery	G3	0.037038	0.009696	0.8474

Multi-dimensional Analysis

L'Analyse des Correspondances Multiples (ACM) est une méthode statistique utilisée pour analyser les relations entre variables qualitatives. Elle étudie les proximités entre individus décrits par deux ou plusieurs variables qualitatives ainsi que les proximités entre les modalités de ces variables. Le résultat de l'ACM est présenté comme la figure suivante:

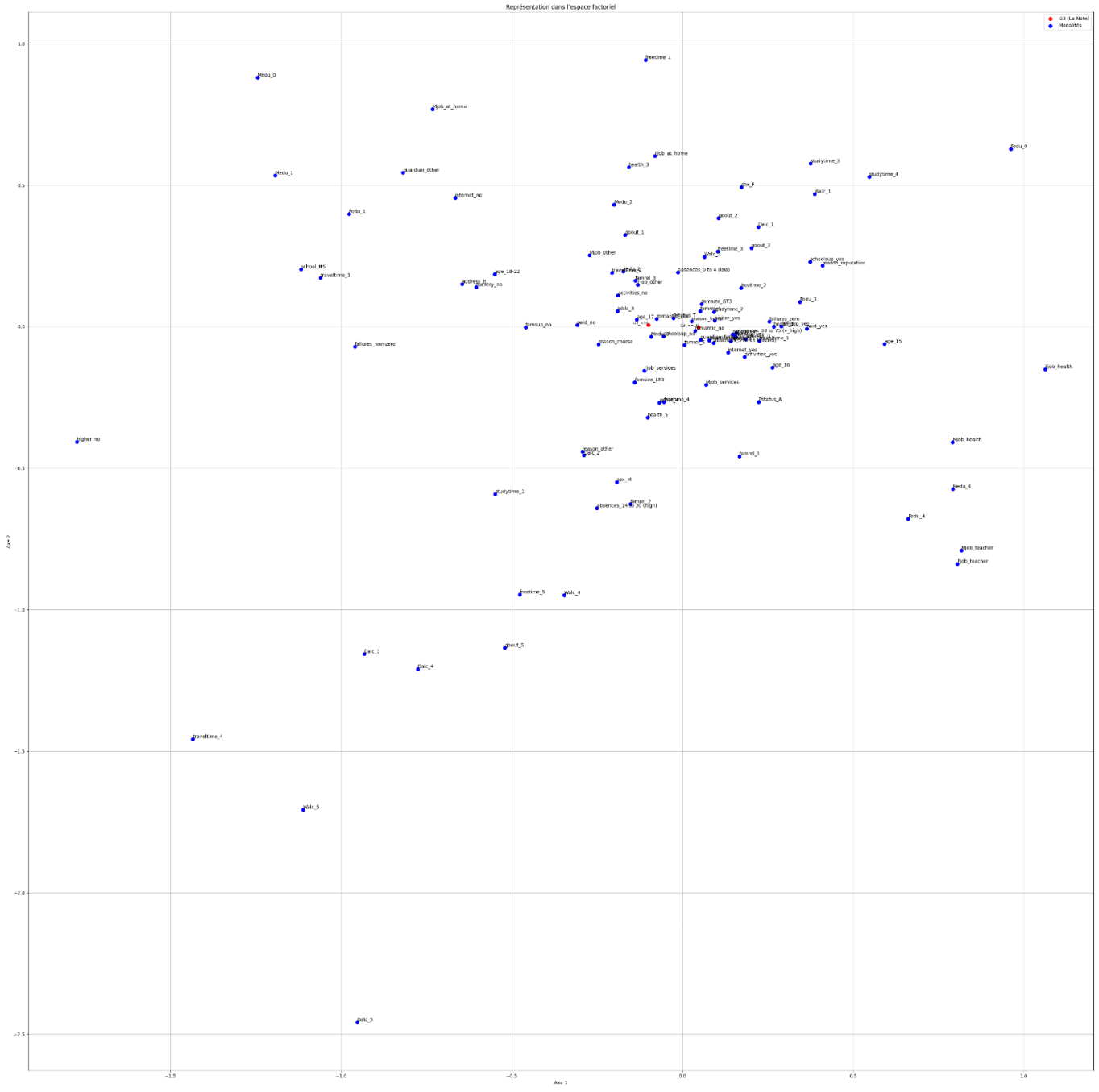


Figure - Le premier plan factoriel de l'ACM

Axe 1 : Facteur Académique et Socio Économique

Le 1er axe est principalement influencé par le niveau d'éducation des parents (Medu et Fedu), les échecs scolaires (failures), et les conditions d'apprentissage (temps de trajet et temps d'étude).

- Les élèves dont la mère et/ou le père ont un niveau d'éducation élevé (Medu_4, Fedu_4, correspondant à un enseignement supérieur) sont situés à une extrémité de l'axe.
- À l'opposé, ceux dont les parents ont un faible niveau d'éducation (Medu_1, Fedu_1, correspondant à l'école primaire, voire Medu_0 et Fedu_0, aucun niveau d'éducation) sont souvent associés à un plus grand nombre d'échecs scolaires (failures_non-zero).
- Le temps de trajet long (traveltime_3, traveltime_4) et un faible temps d'étude (studytime_1) semblent également liés aux difficultés académiques.

Le 1er Axe oppose les étudiants issus de familles avec un haut niveau d'éducation parentale et peu d'échecs scolaires à ceux dont les parents ont un faible niveau d'éducation et qui rencontrent des difficultés académiques.

Axe 2 : Facteur Social et Comportemental

Le 2eme axe distingue les étudiants en fonction de leur genre (sex_F, sex_M) et de leurs habitudes de vie, notamment la consommation d'alcool (Dalc, Walc) et les relations sociales.

- Les étudiants ayant une forte consommation d'alcool le week-end (Walc_5) et une consommation quotidienne élevée (Dalc_1) sont situés à une extrémité de l'axe.
- Le temps libre (freetime_1, freetime_5) et la qualité des relations familiales (famrel_1, famrel_2) sont également liés à cet axe, suggérant que celui-ci reflète les différences dans les comportements sociaux et familiaux des élèves.

Le 2eme Axe Distingue les étudiants en fonction de leur sexe et de leurs habitudes de vie (consommation d'alcool, temps libre, relations familiales).

Interprétation de la variable cible

Nous avons ajouté la variable cible en tant que variable supplémentaire. Ces modalités sont représentées en rouge sur le graphique. Les modalités cibles situées proches de certaines variables indiquent une forte corrélation entre cette classe cible et ces variables.

L'analyse des résultats scolaires (G3) montre que certains facteurs jouent un rôle clé dans la réussite des élèves. Ceux qui obtiennent une moyenne finale supérieure ou égale à 10 sont souvent issus de familles où les parents ont un bon niveau d'éducation (Medu, Fedu), ont connu peu d'échecs scolaires et consacrent assez de temps aux études.

À l'inverse, les élèves avec une moyenne inférieure à 10 rencontrent plus de difficultés, souvent liées à un faible niveau d'éducation des parents, plusieurs échecs scolaires et des contraintes comme un long trajet pour aller à l'école. D'autres éléments, comme le mode de vie (sorties, consommation d'alcool, relations familiales), ont une influence plus légère, mais peuvent aussi impacter les résultats.

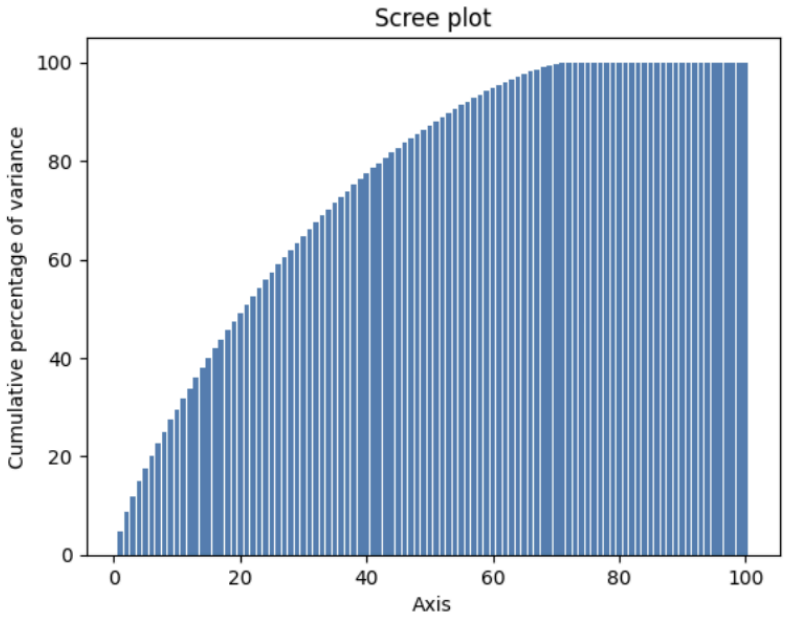


Figure - Les variances expliquées de l'ACM

Modalités dominantes: Ces modalités ont à la fois un poids élevé et une forte contribution aux axes.

1. school_GP
    - Poids : 5,9 %
    - Contribution Axe 1 : 0,56 %
    - Contribution Axe 2 : 0,02 %
2. higher_yes
    - Poids : 6,3 %
    - Contribution Axe 1 : 0,25 %
    - Contribution Axe 2 : 0,02 %
3. internet_yes
    - Poids : 5,6 %
    - Contribution Axe 1 : 0,43 %
    - Contribution Axe 2 : 0,25 %

Modalités Discriminantes mais Peu Fréquentes: Ces modalités ont un faible poids mais une forte contribution aux axes. Elles sont rares mais très influentes.

1. school_MS
    - Poids : 0,8 %
    - Contribution Axe 1 : 4,24 %
    - Contribution Axe 2 : 0,17 %
2. Medu_1
    - Poids : 1,0 %
    - Contribution Axe 1 : 6,2 %

    - Contribution Axe 2 : 1,51 %
3. higher_no
    - Poids : 0,3 %
    - Contribution Axe 1 : 4,65 %
    - Contribution Axe 2 : 0,3 %

4. Medu_4

    - Poids 2.2 %
    - Contribution à l'axe 1 6.06 %
    - Contribution à l'axe 2 3.86 %

5. Fedu_1

    - Poids 1.4 %
    - Contribution à l'axe 1 5.77 %
    - Contribution à l'axe 2 1.17 %

6. Variable : Fedu_4

    - Poids 1.6 %
    - Contribution à l'axe 1 3.1 %
    - Contribution à l'axe 2 3.97 %

On a effectué une nouvelle analyse en composantes multiples (MCA) en utilisant uniquement les 10 variables les plus significatives : 'failures', 'goout', 'age', 'absences', 'higher', 'Fedu', 'Medu', 'Dalc', 'guardian', et 'Mjob'.

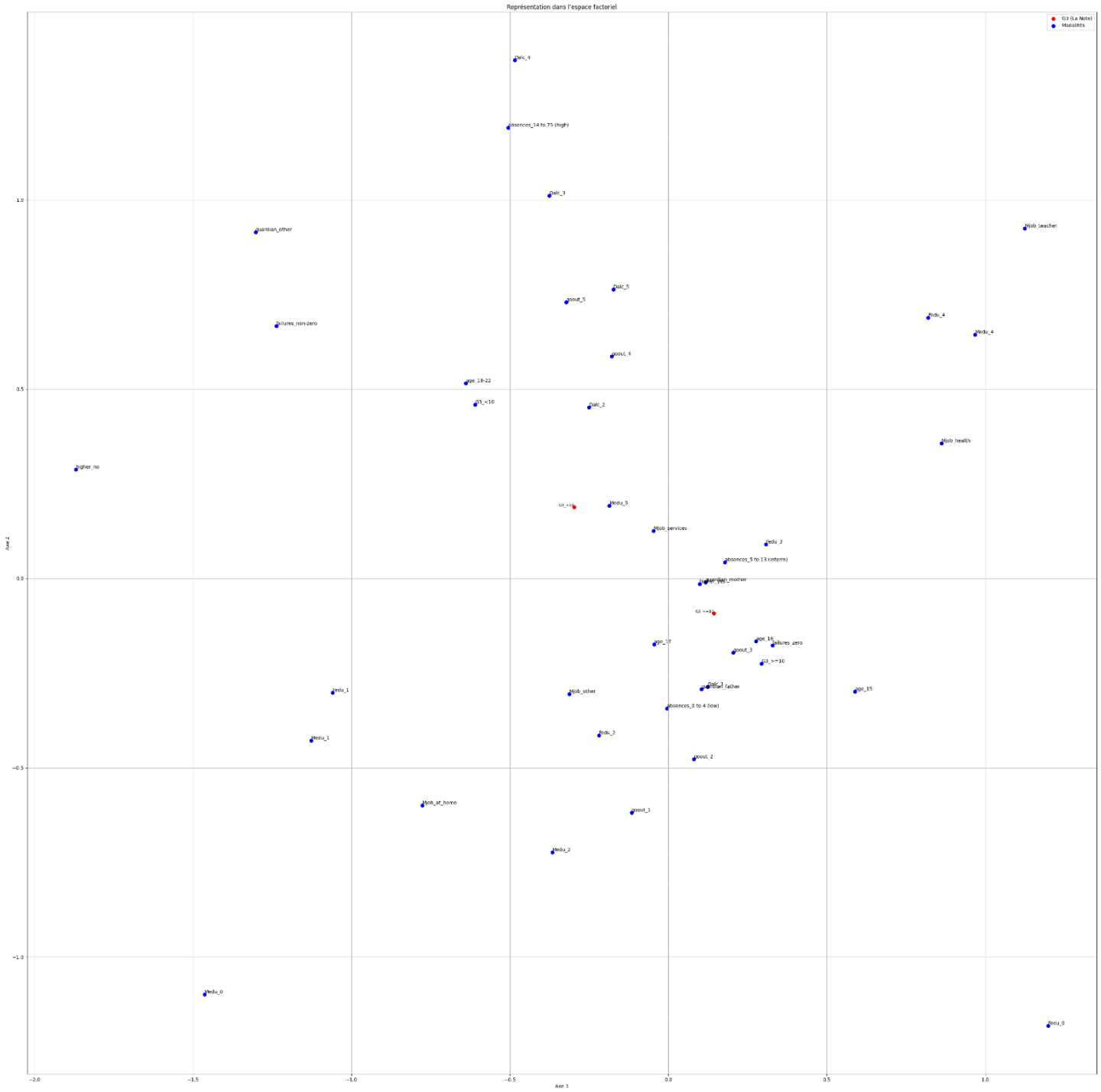


Figure - Le premier plan factoriel de l'ACM_2

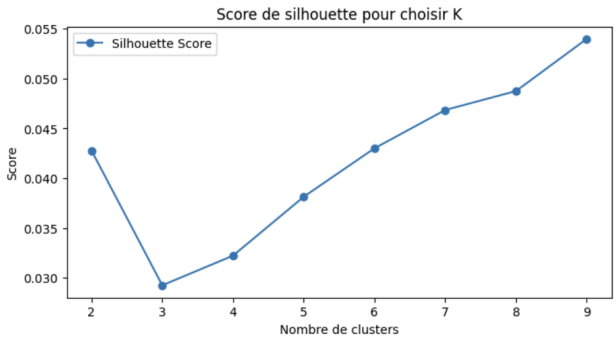
L'axe 1 semble être fortement influencé par les modalités liées aux échecs scolaires et au niveau d'éducation des parents, indiquant que cet axe pourrait capturer des variations importantes liées aux performances académiques et à l'environnement familial des étudiants.

```
Modalités qui ont les cos2 les plus élevés sur l'axe 1
                     Axe 1     Axe 2
failures_non-zero  0.406837  0.118107
failures_zero      0.405409  0.117353
Fedu_1             0.293927  0.023905
Fedu_4             0.215960  0.152325
Medu_1             0.223388  0.032258
Medu_4             0.464935  0.205500
Mjob_teacher       0.217855  0.147228
------------------------------------------------
Modalités qui ont les cos2 les plus élevés sur l'axe 2
                              Axe 1     Axe 2
absences_14 to 75 (high)  0.037028  0.205177
Medu_4                    0.464935  0.205500
```

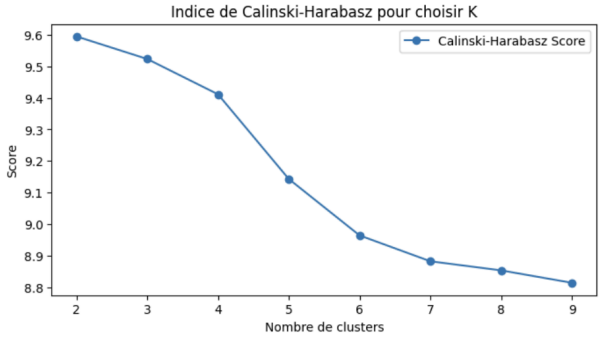
L'axe 2, quant à lui, semble capturer davantage les absences et le niveau d'éducation maternelle, suggérant que cet axe est davantage lié à des facteurs comportementaux et sociaux des étudiants, tels que les absences et l'éducation parentale.

Clustering

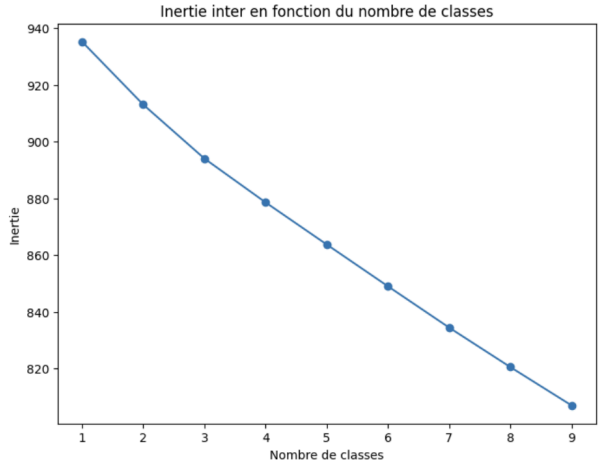
Nous passons maintenant à la classification des individus à l'aide d'une **Classification Ascendante Hiérarchique (CAH)**. Pour cela, nous utiliserons les 50 premières composantes principales de 1er ACM réalisé, qui expliquent 90% de la variance, garantissant ainsi une représentation fidèle des données tout en réduisant la dimensionnalité.



Afin de déterminer le nombre optimal de clusters, nous effectuons plusieurs itérations de la CAH en utilisant la méthode de Ward comme critère d'agrégation, qui minimise l'inertie intra-classe. L'évaluation des résultats à l'aide des indices de Silhouette, Calinski-Harabasz, et inertie indique qu'un partitionnement optimal se situe autour de 4 clusters, comme le montrent les figures suivantes.



Enfin, nous procédons à la segmentation finale des individus en 4 classes et représentons leur répartition sur le plan factoriel en colorant chaque point selon son cluster. Cette visualisation permet d' interpréter la structure des groupes en relation avec les axes principaux de l'ACM.



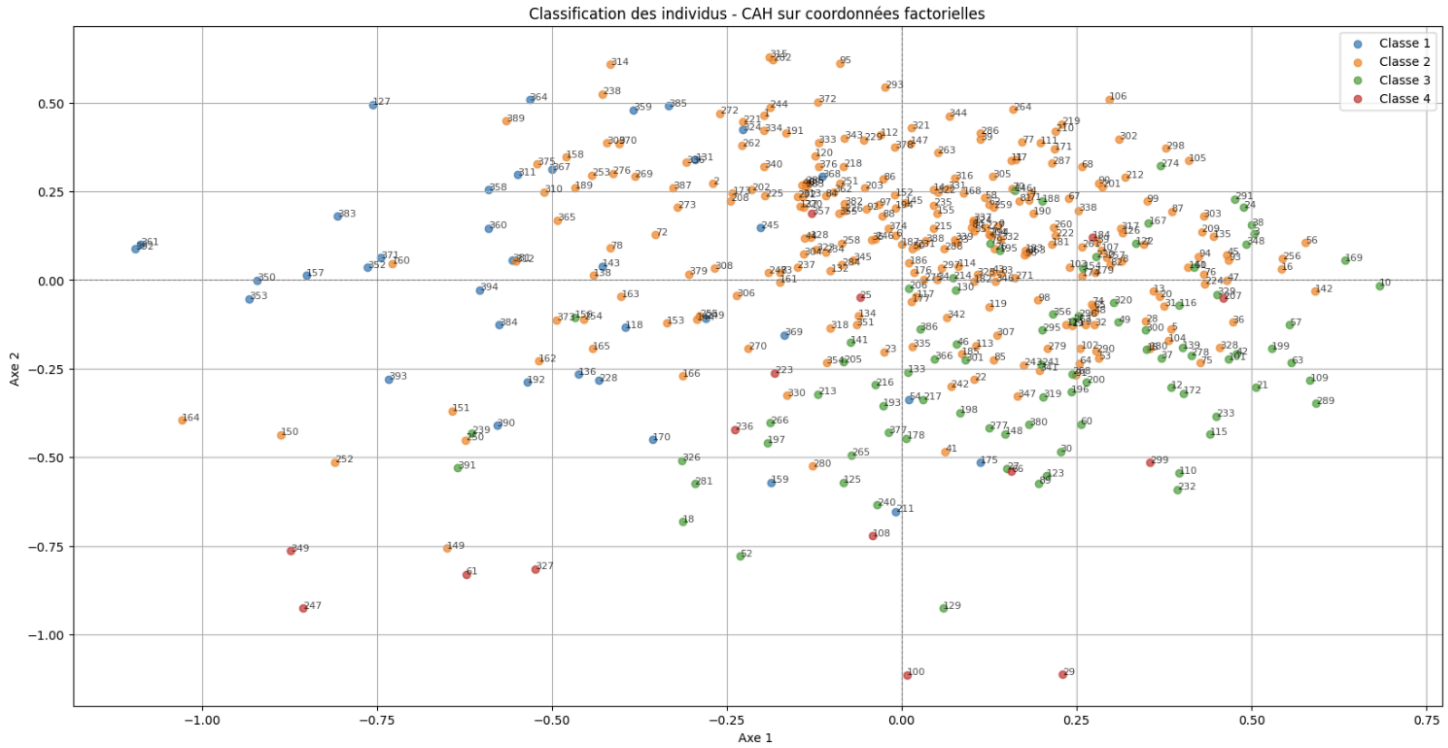


Figure - Projection du clustering (ACH) sur le premier plan factoriel de l'ACM

Le **cluster 1** est assez petit, avec seulement 17 étudiants. Parmi eux, la majorité (11) ont des notes **>=10**, ce qui montre qu'il s'agit d'un groupe d'étudiants avec des performances plutôt moyennes.

Le **cluster 2** est le plus grand, avec 276 étudiants. Ici, la majorité ont des notes **>=10** (189 étudiants), tandis qu'une minorité a des notes **<10** (87 étudiants). Ce groupe représente surtout des étudiants avec des performances bonnes à moyennes.

Le **cluster 3** regroupe seulement 15 étudiants, dont 12 ont des notes **>=10**. C'est un petit groupe d'étudiants relativement performants, mais assez isolés.

Enfin, le **cluster 4** contient 86 étudiants, avec une répartition plus équilibrée entre ceux ayant des notes **>=10** (52 étudiants) et ceux ayant des notes **<10** (34 étudiants). Cela montre un groupe plutôt diversifié en termes de performance.

Ces résultats ne sont pas vraiment utiles pour prédire la classe de **G3**, car le clustering ne semble pas offrir une séparation claire entre les groupes de performance. Bien que certains clusters montrent une répartition des notes, il n'y a pas de distinction nette qui permettrait de prédire efficacement si un étudiant obtiendra une note **>=10** ou **<10**. Les clusters sont trop mélangés, avec des groupes qui contiennent à la fois des étudiants ayant des bonnes et mauvaises performances, ce qui réduit la capacité prédictive du modèle.

	cluster	G3	count
0	1	<10	6
1	1	>=10	11
2	2	<10	87
3	2	>=10	189
4	3	<10	3
5	3	>=10	12
6	4	<10	34
7	4	>=10	52

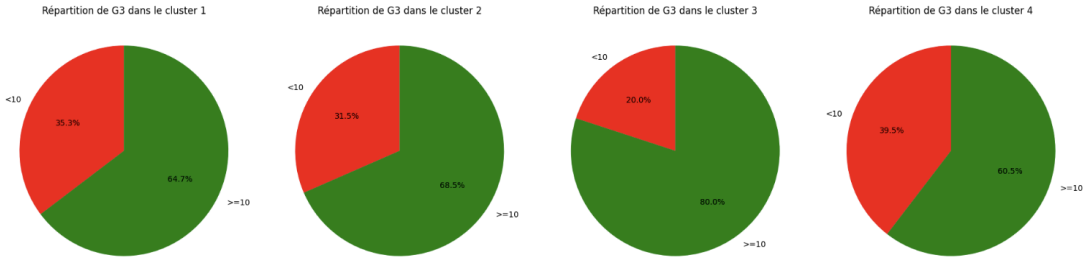


Figure - Répartition de G3 dans chaque Cluster

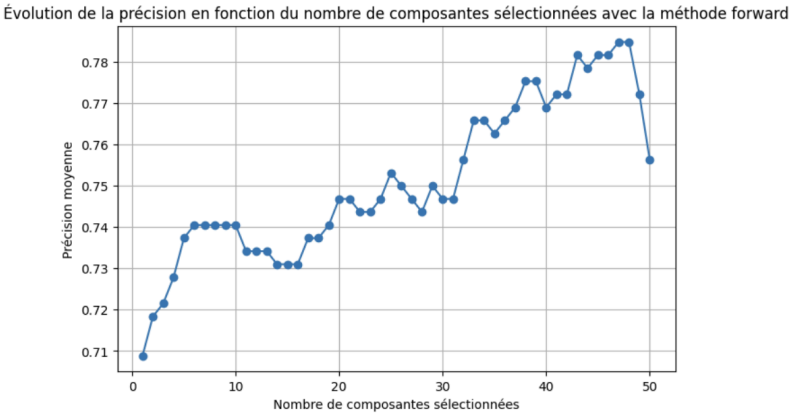
Analyse discriminante

L'Analyse Discriminante est une technique statistique qui permet de classer des données en fonction de caractéristiques spécifiques et de groupes déjà définis. Elle est souvent utilisée pour simplifier des jeux de données en réduisant leur dimension ou pour effectuer une classification supervisée. L'objectif principal de cette méthode est de rendre les groupes aussi distincts que possible, tout en réduisant la variabilité à l'intérieur de chaque groupe. Dans certains cas, elle peut aussi projeter les données dans un espace de dimension plus petite, tout en gardant la séparation entre les groupes. On distingue deux types d'Analyse Discriminante : l'Analyse Discriminante Linéaire (LDA) et l'Analyse Discriminante Quadratique (QDA). LDA est utilisée quand on suppose que les variables suivent une distribution normale dans chaque groupe et que la frontière de séparation entre les groupes est linéaire, ce qui la rend idéale pour des jeux de données plus petits. En revanche, QDA est utilisée lorsque les frontières de séparation sont plus complexes, souvent quadratiques, et est mieux adaptée aux jeux de données de plus grande taille.

Une variante de l'Analyse Discriminante Linéaire est l' **LDA avec sélection pas-à-pas (stepwise)**, qui inclut une méthode automatique de sélection des caractéristiques les plus pertinentes. Dans cette approche, un processus itératif ajoute ou supprime des variables en fonction de critères statistiques (généralement les valeurs p), afin de réduire le nombre de caractéristiques utilisées tout en conservant, voire en améliorant, la performance du modèle. Cette méthode peut être utilisée pour éliminer les variables moins significatives et améliorer la précision du modèle tout en conservant la séparation entre les groupes.

Note: Nous avons divisé notre jeu de données en deux parties : 80% pour l’entraînement et 20% pour le test.

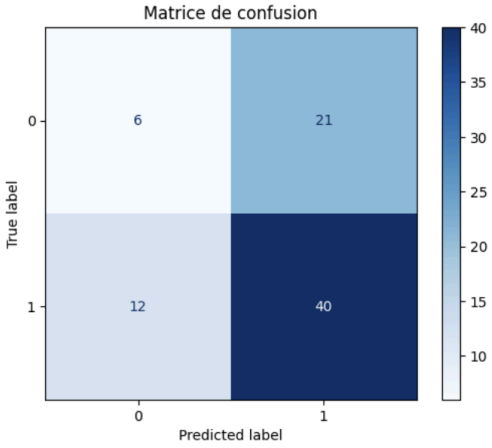
LDA:



Dans la figure, nous avons observé que la précision moyenne augmente continuellement à mesure que le nombre de composantes sélectionnées augmente, jusqu'à atteindre un pic à 47 composantes, après quoi elle diminue légèrement. Ainsi, le meilleur choix pour optimiser la précision est de sélectionner 47 composantes. De plus, la première composante (LD1) explique environ 0.71 de la précision moyenne, ce qui signifie qu'elle capture la majorité de l'information discriminante."

Les résultats obtenus montrent une **précision (Accuracy)** de 0.58, ce qui indique que le modèle classe correctement environ 58% des observations. Bien que ce ne soit pas une performance exceptionnelle, cela suggère que le modèle est mieux que le hasard, mais pourrait encore bénéficier d'améliorations.

```
Accuracy: 0.5822784810126582
AUC: 0.49572649572649574
```



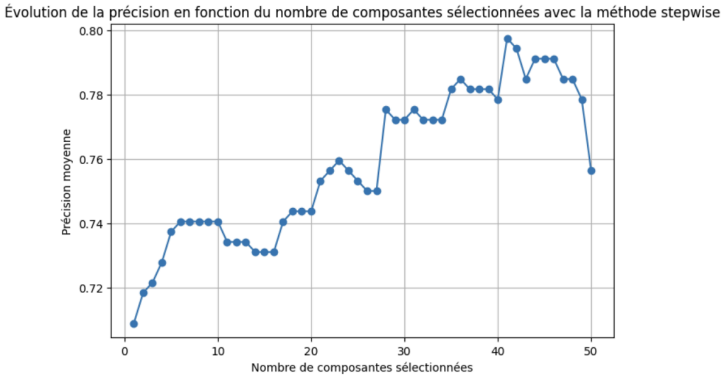
L'**AUC (Area Under the Curve)** est de 0.4957, ce qui est légèrement inférieur à 0.5, signifiant que le modèle ne distingue pas bien entre les deux classes (ici, <10 et >=10). Une AUC proche de 0.5 est souvent interprétée comme une performance similaire à celle d'un modèle aléatoire.

En examinant la **matrice de confusion**, on constate :

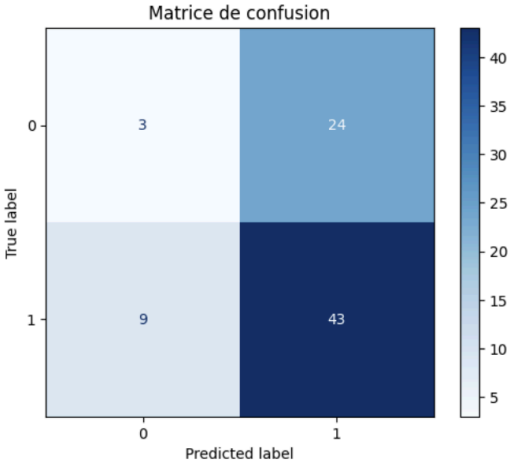
- Pour la classe "<10", le modèle a correctement classé 6 instances, mais en a mal classé 21 comme appartenant à la classe ">=10".
- Pour la classe ">=10", le modèle a correctement classé 40 instances, mais a mal classé 12 comme "<10".

Les résultats montrent qu'il faut ajouter des échantillons d'entraînement pour la classe "G < 10". En effet, le modèle semble avoir du mal à prédire correctement cette classe, avec un nombre élevé de faux positifs. En augmentant le nombre d'exemples pour la classe "G < 10", le modèle pourrait mieux apprendre à distinguer cette classe de la classe "G >= 10", ce qui pourrait améliorer à la fois la précision et l'AUC, et par conséquent la performance générale du modèle.

LDA (avec stepwise):

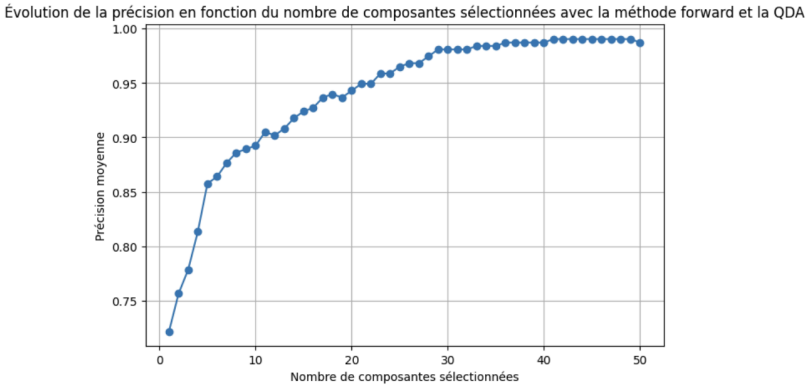


Dans cette analyse utilisant l'Analyse Discriminante Linéaire (LDA) avec sélection pas-à-pas, nous avons calculé la précision moyenne en fonction du nombre de composantes sélectionnées. La précision augmente constamment jusqu'à atteindre 41 composantes, après quoi elle diminue légèrement. Ainsi, le meilleur choix pour le nombre de composantes est de 41. De plus, la première composante (LD1) explique environ 0.71 de la précision moyenne, ce qui signifie qu'elle capture la majorité de l'information discriminante.



Les résultats montrent une précision de 0.582 et une AUC de 0.469, ce qui indique que la performance du modèle n'a pas significativement augmenté. Le modèle peine particulièrement à prédire la classe "G < 10", avec seulement 3 vrais positifs et un nombre élevé de faux positifs (24). En revanche, la classe "G >= 10" est mieux prédite. Cela suggère probablement un déséquilibre entre les classes, où la classe "G >= 10" est sous-représentée. Pour améliorer ces résultats, il serait nécessaire d'ajouter davantage d'échantillons pour la classe "G < 10".

Analyse quadratique

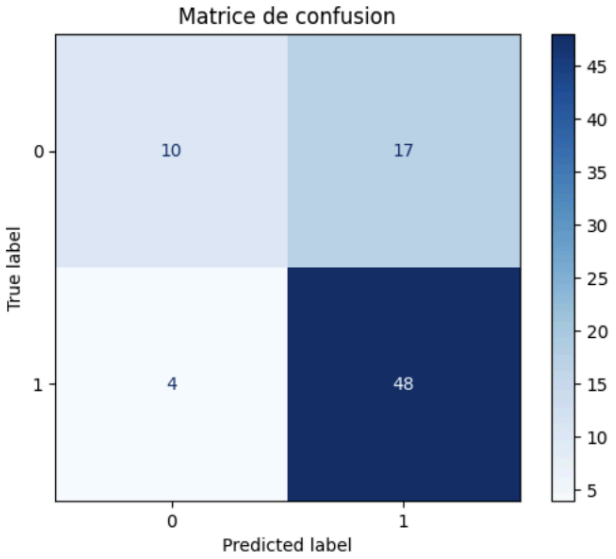


Dans la figure, on peut observer que la précision moyenne augmente continuellement à mesure que le nombre de composantes sélectionnées augmente, jusqu'à atteindre un pic à 48 composantes (précision moyenne = 0.99), après quoi elle diminue un peu. Ainsi, le meilleur choix pour optimiser la précision est de sélectionner 48 composantes. De plus, la première composante explique environ 0.73 de la précision moyenne, ce qui signifie qu'elle capture la majorité de l'information discriminante.

Le modèle de Quadratic Discriminant Analysis (QDA) semble mieux fonctionner que LDA et LDA avec sélection par étapes, surtout pour la classe "G < 10". Avec QDA, on remarque que le nombre de vrais positifs pour cette classe est de 10, ce qui est mieux que les autres modèles, où cette classe était souvent mal classée. En plus, le nombre de faux positifs est réduit à 17, ce qui montre que QDA réussit mieux à distinguer les deux classes.

Cela peut être dû au fait que QDA est plus flexible que LDA. Alors que LDA suppose des frontières linéaires entre les classes, QDA peut gérer des frontières plus complexes, ce qui peut être utile quand les classes ne sont pas clairement séparées par une simple ligne droite. Bref, QDA réussit mieux à classifier les élèves "G < 10", ce qui améliore les résultats pour cette classe en particulier.

```
Accuracy: 0.6075949367088608
AUC: 0.49715099715099714
```



Annexe

Data link: https://archive.ics.uci.edu/dataset/320/student+performance

GitHub link: https://github.com/achrafndd05/DQ_projet

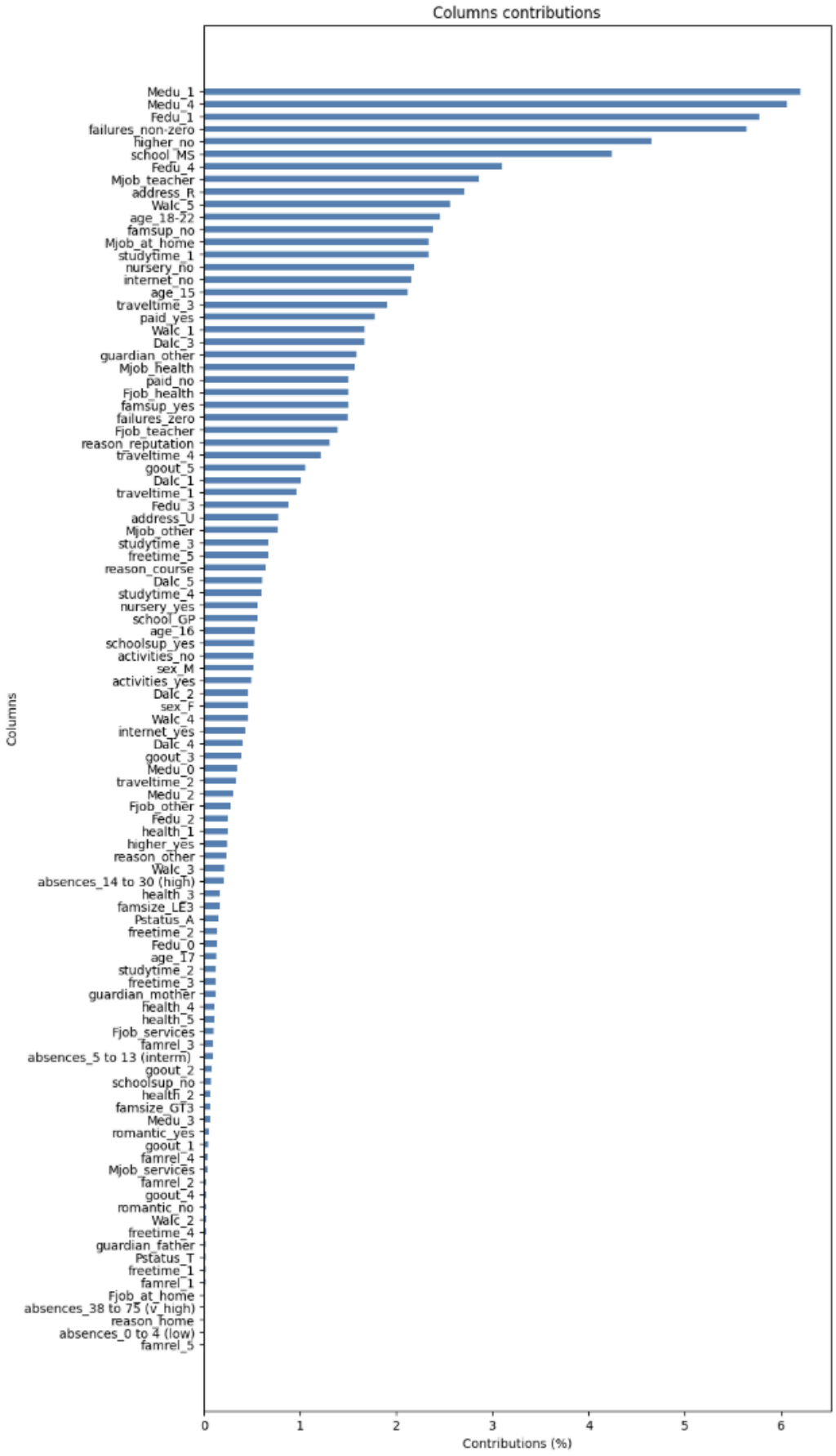


Figure - Les contributions des modalités à l'axe 1 (ACM 1)

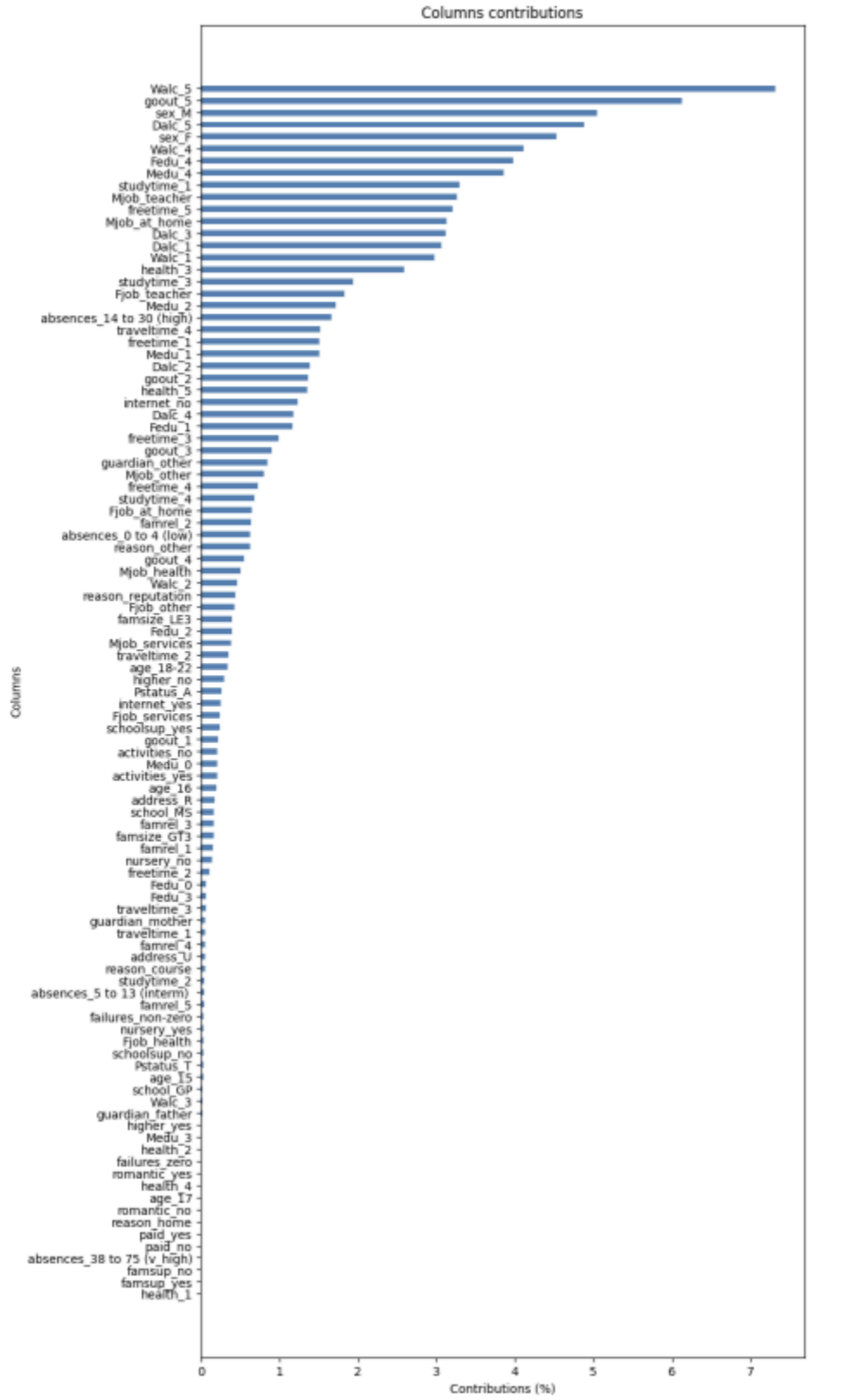


Figure - Les contributions des modalités à l'axe 2 (ACM 1)