Projet Cogmont

Azat, Lucas, Matthieu, Etienne

March 13, 2019

Contents

Introduction

Pour finaliser notre 1ère année de master nous devons réaliser un projet en lien avec les statistiques et l'analyse de données. C'est pourquoi nous avons choisi ce sujet, l'analyse de pédagogies éducatives au sein d'une école primaire. Mêlant sciences cognitives, étude statistique et analyse de données nous avons tout au long du semestre mener à bien ce projet en collaboration avec notre tutrice pédagogique et notre référante client. Nous avons reçu les résultats d'élèves d'une école primaire sous forme de tableur que nous avons ensuite trié pour ne garder que les années d'études avec suffisamment de promotion pour pouvoir faire une études transversale. Ensuite nous avons nettoyer les données et commencer l'analyse.

1. Contexte

Ce projet nous à été proposé par l'Institut des sciences cognitives - Marc Jeannerod spécialisé dans en neurosciences. L'UMR 5304 créée en 2007 est un des deux laboratoires de l'Institut des Sciences Cognitives –

Marc Jeannerod. L'UMR 5304 est un laboratoire interdisciplinaire qui intègre l'expertise de chercheur des Sciences de la Vie (psychologie cognitive, neurosciences) et de médecine (pédo-psychiatrie, neuro-pédiatrie) avec celle de chercheur des Sciences Humaines et Sociales (linguistique computationelle et théorique et philosophie) pour étudier la nature et la spécificité de l'esprit humain.

1.1 Les besoins d'un changement éducatif

Le député et mathématicien Cédric Villani a publié un rapport pour renforcer l'apprentissage des mathématiques à l'école. Les élèves français sont aujourd'hui plus que médiocres dans cette discipline. Pourtant, jusqu'en 1985, l'enseignement des maths en France était reconnu comme l'un des meilleurs. Or, l'étude internationale "Trends in International Mathematics and Science Study" (TIMSS) 2015 qui mesure les performances en mathématiques et en sciences des élèves en fin de CM1 classe la France dernière des pays de l'Union européenne et la France obtient un score en dessus de la moyenne internationnale. Pour mettre un terme à cette tendance inquiétante de la dégradation du niveau des élèves français en mathématiques, le gouvernement est à la recherche de nouvel pédagogie d'enseignement des mathématiques.

1.2 Les pedagogies

Pédagogie Montessori: La pédagogie Montessori est une méthode d'éducation créée en 1907 par Maria Montessori.
La pédagogie se base sur trois principe:
-l'autodiscipline: les enfants sont libres de choisir l'activité qu'ils souhaitent faire parmi celles qui leur sont proposées.
-L'action en périphérique: Selon Maria Montessori, il est plus profitable d'agir sur son environnement plutôt que sur l'enfant lui-même (comme des classes multi-âge).

Pédagogie "Conventionnelle": La pédagogie traditionnelle est celle du modèle transmissif. Selon le triangle pédagogique de Jean Houssaye, cette pédagogie privilégie la relation entre l'enseignant et le savoir. Autrement dit, l'enseignant expose un savoir sous forme de cours magistral, généralement suivi d'exercices ou/et de leçons à apprendre. L'élève doit intégrer et appliquer le savoir exposé par l'enseignant.

1.3 Présentation des données de départ

Notre jeu de données est composé de trois fichiers Excel (.xlsx), avec les résultats de chaque promotion au test cognitif mis en place par l'équipe de recherche l'institut des sciences cognitives.

MathsJetons_2015-2016.xlsx : Pour l'année 2015/2016.

MathsJetons_2015-2016.xlsx : Pour l'année 2016/2017.

MathsJetons_2016-2017.xlsx : Pour l'année 2017/2018.

Chaque jeu de données représente les résultats questions par questions (en comptant les sous-questions) des élèves ainsi que leurs catégories pédagogiques et des informations telles que l'encadrant, le niveau scolaire, la langue natale, l'âge, le type de classe (mélangé entre plusieurs section ou pas), l'année de passage du test et leur école. Il y à 10 questions divisées en sous questions, ce qui fait un total de 34 réponses. Chaque question est indépendante et pour répondre à la sous question suivante il faut une bonne réponse à la sous-question précédente, sauf pour la question 4, toutes ses sous-questions sont indépendantes. Une bonne réponse correspond à un 1 et une mauvaise réponse à un 0, sauf pour la réponse à la question 1 qui est la valeur de comptage maximale de l'enfant. Ici les élèves viennent tous de l'école de REP+ 5Réseau d'Education Prioritaire).

1.4 Nettoyage des données

Les données ayant déjà été travaillées l'année dernière le travail nécessaire en datamanagement n'a pas été excessif. Il nous a fallu tout de même renommer certaines variables pour les rendre plus lisibles, supprimer certaines questions car elles n'avaient été posées qu'à certaines classes... Les questions étant posées de manière à ce qu'au sein d'une même tache, il faille réussir les questions dans l'ordre pour passer à celles plus dures nous avions beaucoup de NA dès qu'une question est dure. Nous avons donc décidé de changer ces NA et les considérer comme une question que l'élève n'aurait pas réussit. Et pour ne pas passer à coté de l'information : "n'a pu aller plus loin", nous avons crée deux jeux de données : un vectoriel, un composé de scores correspondant à la somme des questions au sein d'une même tâche.

1.5 Recodage des variables

Afin de ne pas influencer notre jugement sur nos résultats, nous avons dans un premier temps décidé de rendre anonyme le pédagogie enseignée pour chaque classe. Chaque pédagogie fut donc renommée en "P1" et "P2". De ce fait nous n'avons pas pu privilégier une pédagogie plus qu'une autre subjectivement parlant. Aux 3 quarts du projet environ, nous avons reçu les données de nouveaux individus, une cinquantaine, et avons par la même occasion décidé d'enlever cet anonymat. Notre travail étant déjà réalisé, seul l'interprétation sur le jeu de données comportant les nouveaux individus en plus importe. Comme dit précédemment les questions sont divisées en sous questions, ces questions sont regroupables en groupe : un groupe qu'on appellera "variable au-dela", un groupe "variable outil" puis un groupe "variable objet". Chacun de ces groupes fait appelle à une tache pédagogique en particulier (le calcul, la mémoire...).

2. Méthodologie

Afin de répondre au mieux à notre problématique nous avons fait le choix d'utiliser plusieurs méthodes statistiques différentes pour analyser nos données. Pour cela nous avons dans un premier temps utilisé une méthode qui permet de résumer l'information globale du jeu de données : l'analyse factorielle, et la classification ascendante hiérarchique (pour faire des regroupement de variables). Puis dans un but prédictif nous avons utilisé la régression logistique et les arbres de régressions. Plusieurs tests ont été fait en parallèle, comme celui du chi2, de student..

2.1 Analyse factorielle

Les méthodes d'analyse factorielle que nous avons utilisé ici sont l'analyse des correspondances multiples (ACM) et l'analyse en composantes principales (ACP), qui sont des méthodes de synthétisation du nombre de dimensions pour les données qualitatives et quantitatives. Cela nous permet d'appréhender plus rapidement le jeu de données, et avoir une première idée de ce qui diffère les individus entre eux (ou ce qui les rapproche). L'ACM permet dans un nuage à N dimensions, en cherchant les plans orthogonaux qui maximisent la variance entre les individus, à résumer celles ci en 4 voire 5 dimensions. l'ACM a été réalisée sur les données vectorisées, celles ci ont été prises comme variables actives (celles qui définissent le placement des individus sur le graphe) et les variables portant sur la pédagogie, la question 1, et l'âge en illustratives (ajoutée après le placement des individus sur le graphe). Le principe était le même pour l'ACP qui a été faite ensuite.

2.2 Classification Ascendante Hiérachique

N'ayant aucune information au préalable sur le thème des questions, leur regroupement...etc Mais sachant que certaines questions faisaient appelle aux mêmes compétences. Nous avons utilisé une variante de la classification ascendante hiérarchique (CAH) afin de partitionner nos variables. Nous avons aussi utilisé la CAH classique qui consiste à regrouper les individus selon leur points communs cela en partant d'une inertie

interclasse maximale, pour arriver à une inertie interclasse de 0. Nous avons utilisé la CAH classique afin de partitionner nos individus dans un but descriptif.

2.3 Tests statistiques

Un test statistique est une procédure de décision entre deux hypothèses. Il s'agit d'une démarche consistant à rejeter ou à ne pas rejeter une hypothèse statistique, appelée hypothèse nulle, en fonction d'un jeu de données. Pour le projet nous avons utilisé le test paramétrique. Un test paramétrique est un test pour lequel on fait une hypothèse paramétrique sur la distribution des données sous H0. Les hypothèses du test concernent alors les paramètres de cette distribution. En fonction du type de données, nous avons utilisé le t-test de Student et le test de proportion de succès.

2.4 Regression Logistique

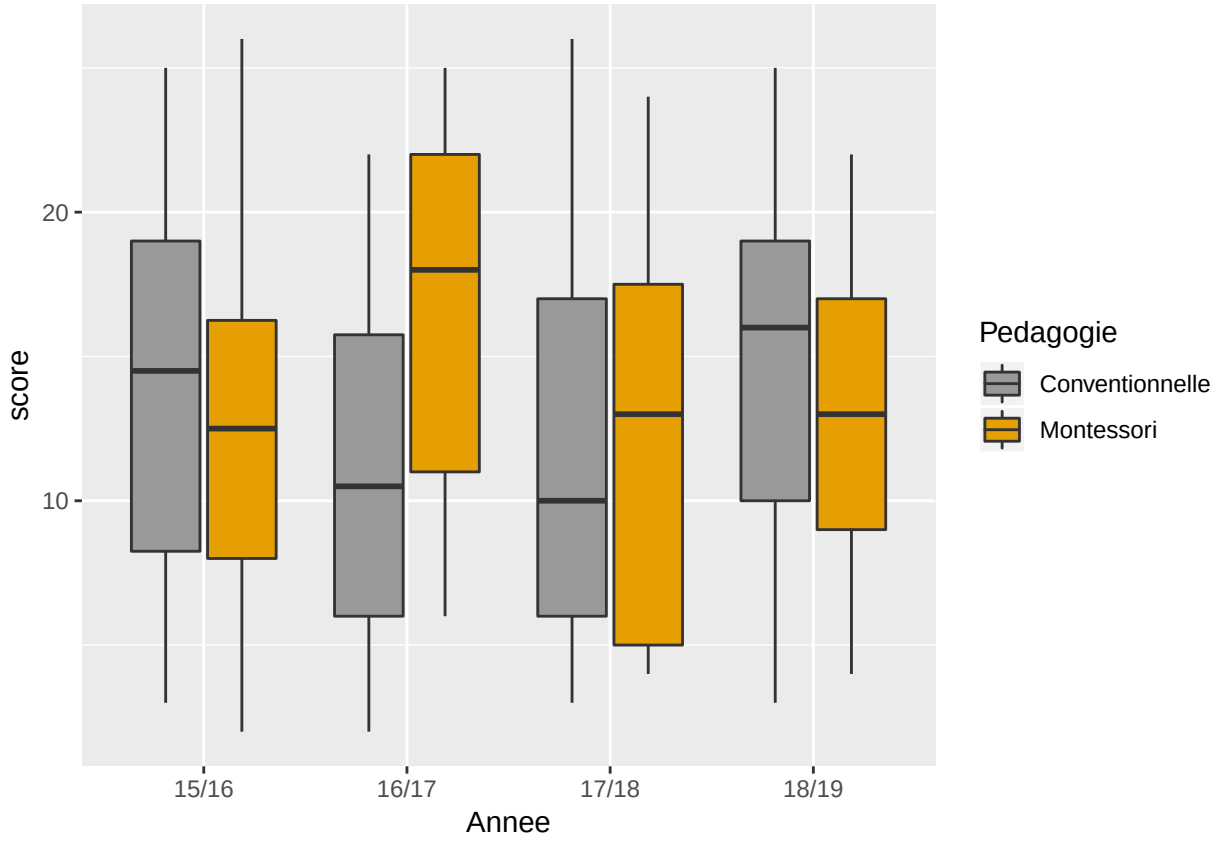
Notre problématique étant de voir s'il existe un lien entre la façon d'enseigner et les réponses au test, nous avons voulu essayer de prédire la méthode d'enseignement à l'aide des réponses des élèves avec la régression logistique. Cette méthode permet de modéliser une classification, à l'aide notamment de l'odds ratio. Cela revient à calculer la probabilité : $P(1|X) = \frac{e^{b_0+b_1x_1+\ldots+b_jx_j}}{1+e^{b_0+b_1x_1+\ldots+b_jx_j}}$. avec $b_0 = ln\frac{p(1)}{p(0)} + a_0$ et $b_j = a_j$.

2.5 Arbre de regression

L'arbre de régression est une technique d'apprentissage supervisé, qui permet en analysant un grand nombre de données, de prédire une variable à expliquer. Ils sont beaucoup utilisés dans le domaine du marketing, et plus récemment dans le domaine du machine learning (apprentissage automatique). Dans un premier temps il s'agit d'exprimer la variable à expliquer en fonction d'un maximum de variables explicatives, puis d'élaguer l'arbre afin de minimiser l'erreur, soit l'écart entre la valeur prédite et la valeur réelle. Cela revient donc à faire une régression logistique sur les données, puis d'appliquer l'algorithme de construction d'arbre à partir des résultats.

3. Analyse exploratoire

3.1 Univariée | Bivariée



En ne regardant que certaines années on peut voir clairement une différence entre la pédagogie montessorienne et la pédagogie conventionnelle vis à vis du nombre de bonnes réponses. Toutefois lorsqu'on regarde la vue d'ensemble, on peut voir que celon les années, une fois la pédagogie montessorienne est supérieur à la conventionnelle, parfois c'est l'inverse. On peut donc s'attendre à ce qu'on ne puisse pas prédire quelle pédagogie permet d'obtenir un meilleur score global.

3.2 Multivariée

Afin de traiter l'information présente dans le jeu de données de la meilleure façon, nous avons procédé à 2 analyses multivariées : 1 sur le jeu de données qualitatif sous forme de vecteurs, et 1 sur le jeu de données quantitatif sous forme de scores.

3.2.1 Analyse des Correspondances Multiples

La réalisation d'une ACM comme première approche sur le jeu de données a permis de mieux comprendre ce qui différencie les individus dans notre jeu de données et à la fois d'avoir un premier résultat sur la différence entre les deux pédagogies selon cette méthode.

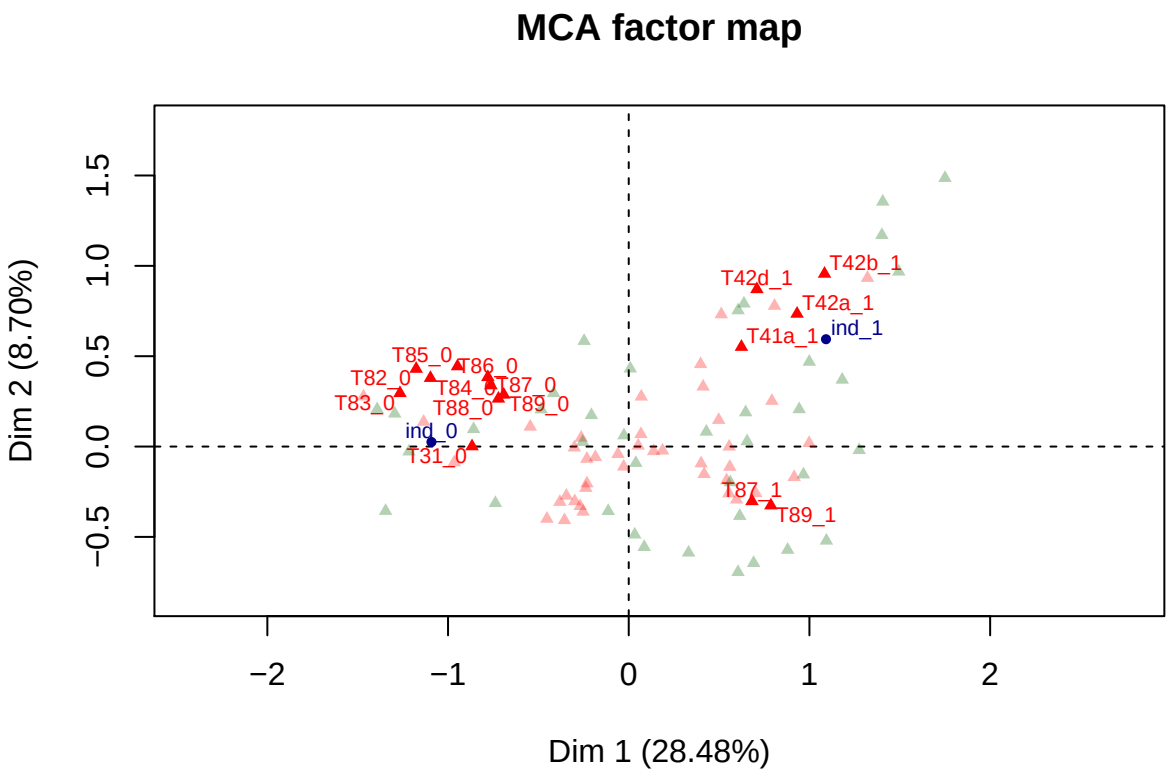


Figure 1: Graphe des modalités sur le plan principal

Le graphe précédent représente les 15 modalités qui contribuent le plus au placement des individus sur le plan principal. On peut donc voir que la dimension 1 oppose des modalités qui concernent la réussite à une question (avec un "_1" à la fin), à droite, à des modalités qui concernent l'échec d'une question (avec un "_0"), à gauche. Plus un élève réussira le questionnaire, plus il se trouvera à droite sur le graphe des individus. Cette interprétation est confirmée par l'ajout de deux individus fictifs : ind_1 et ind_0, qui comporte respectivement des succès à toutes les questions et des échecs à toutes les questions. L'individu ayant réussi en totalité le questionnaire se trouve à droite alors que l'individu ayant raté en totalité le questionnaire se trouve à gauche. De plus nous pouvons voir que les questions qui discriminent le plus la réussite ou non de l'examen sont les questions 4 et 8 (de part leur forte contribution). Toutefois cette première analyse n'aura pas permis de différencier les deux pédagogies, la variable projetée en supplémentaire sur le plan principal n'est pas significativement liée à celui ci.

Dans un second temps nous avons refait une ACM mais cette fois ci sur les données vectorielles. Cela afin de prendre en compte la succession de certaines questions qui se regroupent en "compétences".

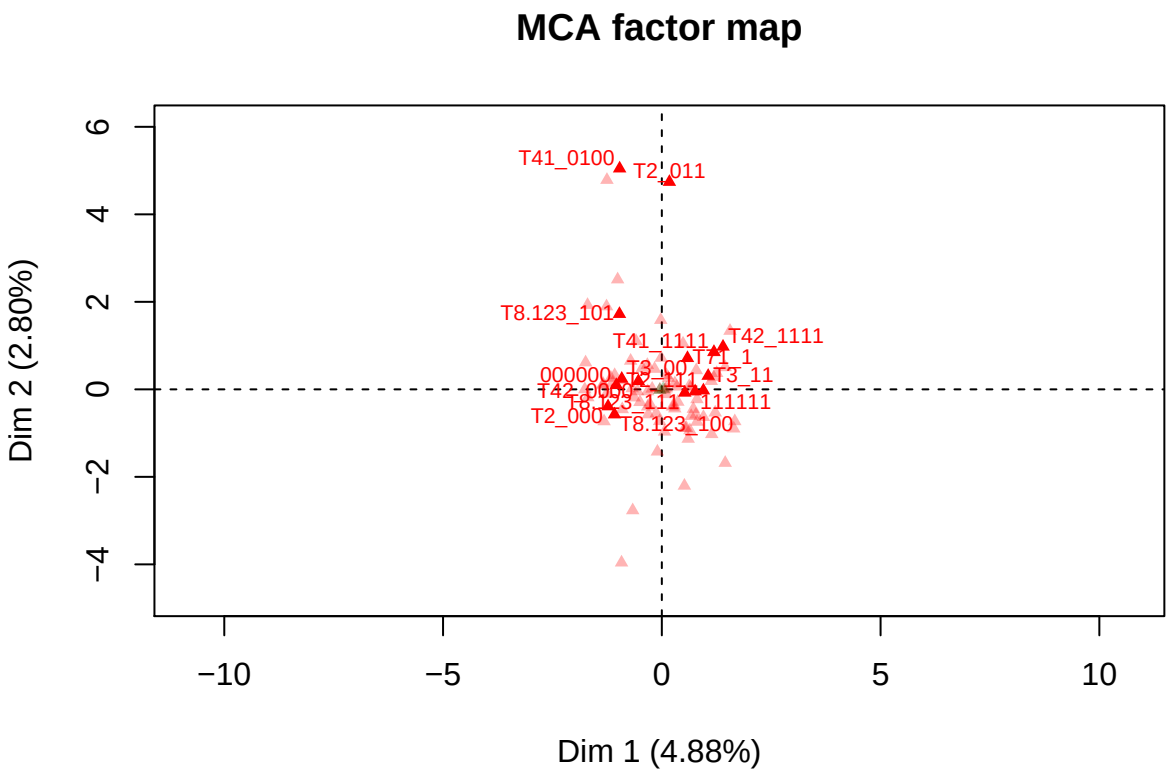


Figure 2: Graphe des modalités sur le plan principal

A nouveau la première dimension oppose les individus ayant réussit les totalités (ou la majorité pour certaines) des question à droite à ceux qui n'en ont réussi aucune à gauche. Nous ne pouvons pas non plus observer de différence significative entre les 2 pédagogies. On peut voir cette fois ci avec plus de précision les questions qui discriminent la réussite au questionnaire. Ce sont Les question 4, 3, 8 et 5.

Dans ces deux cas nous avons pu aussi observer un lien significatif entre les variables qualitatives, portant sur les réponses à la question 1 et l'age de l'élève, et le placement des individus sur la première dimension. En conséquent on peut dire qu'il existe un lien entre la réussite à l'examen et le fait qu'un enfant sache compter "loin" et dans une moindre mesure, qu'il soit âgé.

Enfin nous avons voulu voir si en classifiant les individus suite à l'ACM nous obtenions des groupes d'individus propre à une pédagogie ou non. Pour cela nous avons utilisé la classification ascendante hiérarchique (CAH).

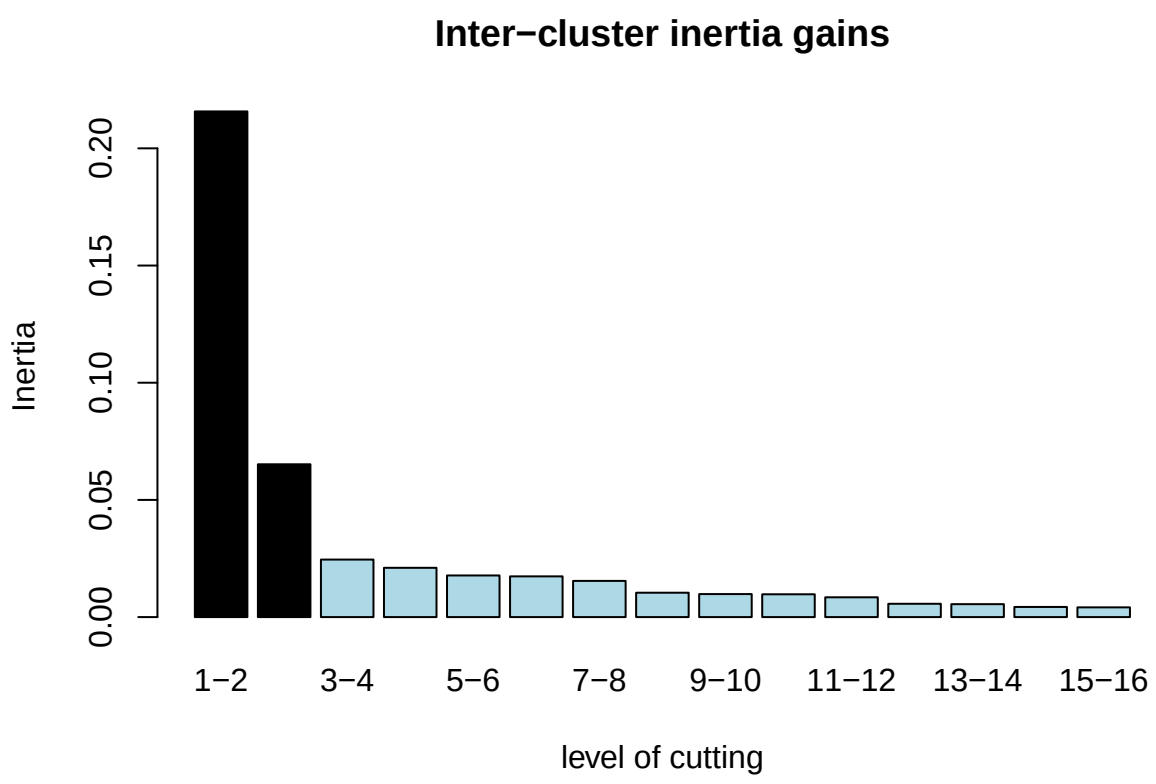


Figure 3: Diagramme des gains d'inertie

Nous pouvons observer un "saut" à la troisième classe, donc nous faisons le choix de retenir trois classes pour la CAH. Mais leur composition ne montre aucune sureprésentation d'une pédagogie plus que l'autre. Le test de chi2 entre la variable concernant la pédagogie et celle concernant la classe n'est pas significatif. Une fois de plus cela ne permet donc pas de montrer une liaison entre la pédagogie et ce qui discrimine nos classe. Au final nous obtenions une classe d'individus qui a une majorité d'échecs, une d'individus qui échouent sur les questions 4.2, et une qui d'individus qui réussissent globalement.

3.2.2 Analyse en Composantes Principales

La réalisation d'une ACP faisant suite à l'ACM a pour but d'étudier le jeu de données différemment. En effet nous avons étudié cette fois ci le jeu de données concernant les scores. Soit un jeu de données quantitatif. Afin de ne pas perdre d'informations nous avons dans un premier temps observé la matrice des corrélations entre les variables de notre jeu de données. Car plus les variables seront corrélées entre elles, plus l'ACP ne montrera que celles ci.

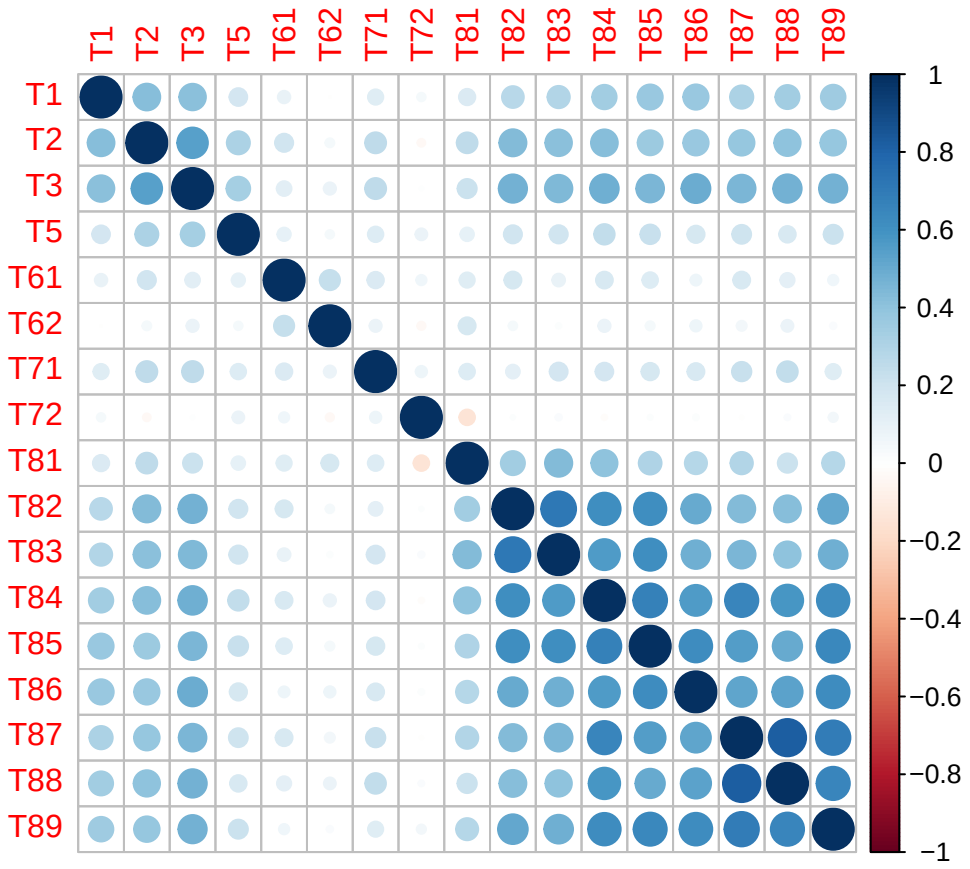


Figure 4: Matrice des corrélations

Nous avons donc fait le choix de regrouper les questions 8 en 2 groupes, aux vues des résultats. Un groupe comprenant les questions 8.1, 8.2 et 8.3, et un comportant les autres questions 8.

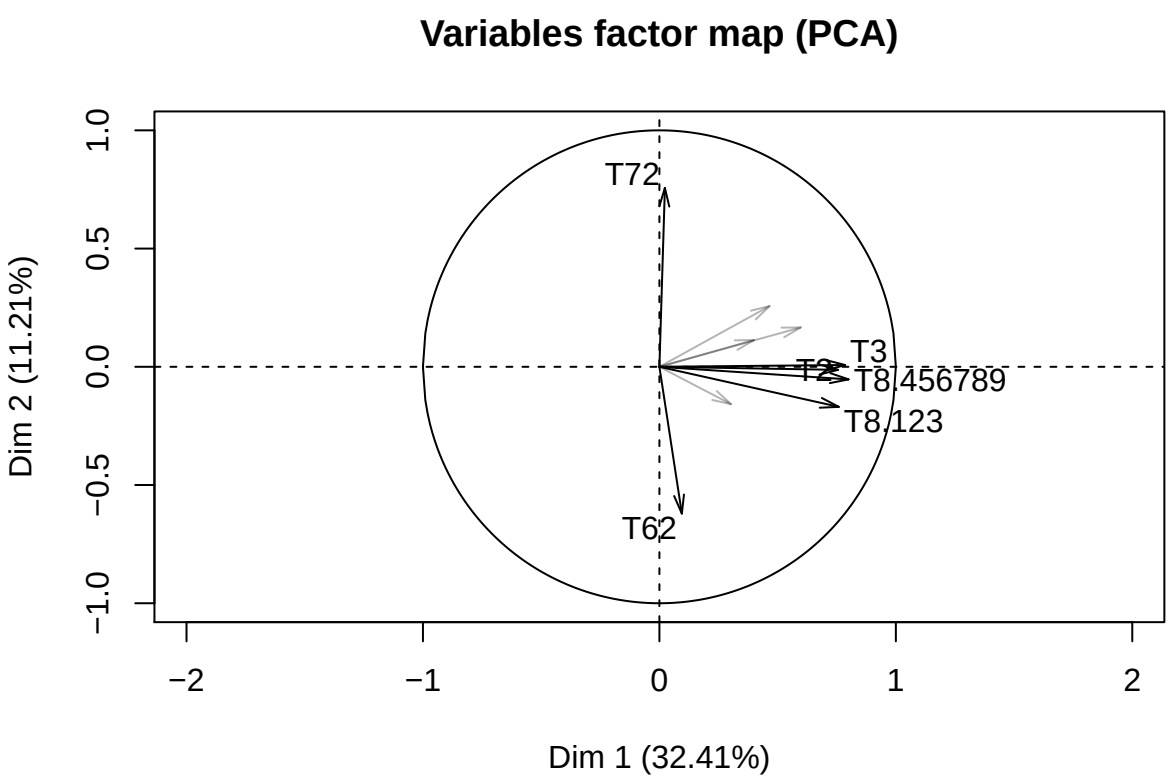


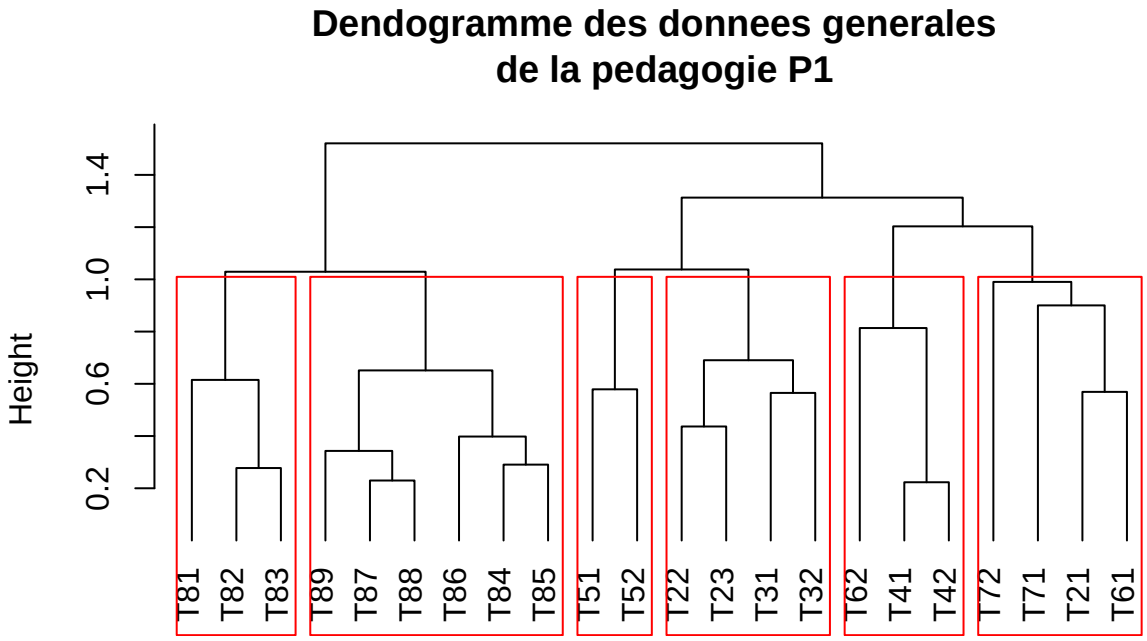
Figure 5: Cercle des corrélations du plan principal

La réalisation de l'ACP nous permet donc de voir que les individus se diffèrent sur la première dimension selon les question 2,3, et 8. Alors que la dimension 2 les différents selon les question 7.2 et 6.2. A nouveau nous n'observons pas de lien significatif entre la pédagogie enseignée et le placement des individus sur les axes factoriels.

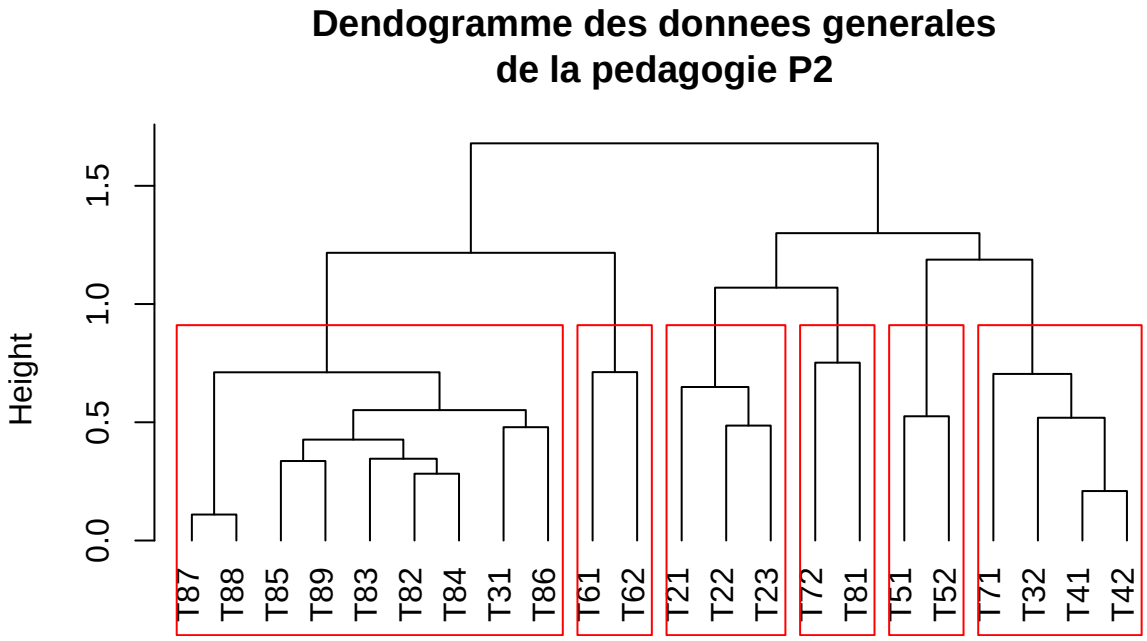
3.3 Classification Ascendante Hiérarchique des variables

N'ayant aucune information au préalable sur le thème des questions, leur regroupement...etc Mais sachant que certaines questions faisaient appellent aux mêmes compétences. Nous avons utilisé une variante de la classification ascendante hiérarchique (CAH) afin de partitionner nos variables. La CAH est une méthode de classification qui permet de regrouper des individus sein d'une même classe et qu'ils soient le plus semblables possibles tandis que les classes soient elles ,le plus dissemblables possibles.

Nous allons appliquer cette méthode sur les jeux de données en isolant les pédagogies pour comparer les regroupements.

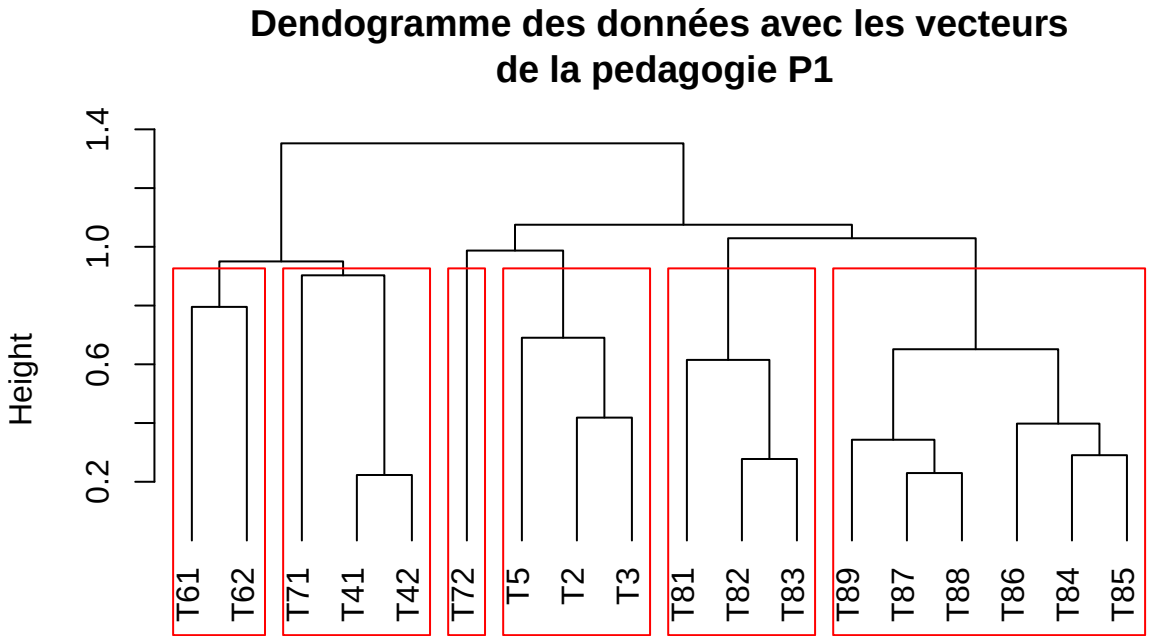


Nous avons d'abord regroupé les résultats du test venant des élèves de la pédagogie 1. Les question T2, T3, T4 et T5 sont regroupées par sous questions. Les deux sous-questions de la T6, T61 et T62 sont très éloignées, de même pour la T71 et T72. On remarque la question T81 n'est pas regroupée avec les autres questions T8 et que T71 est associée aux autres questions T8.

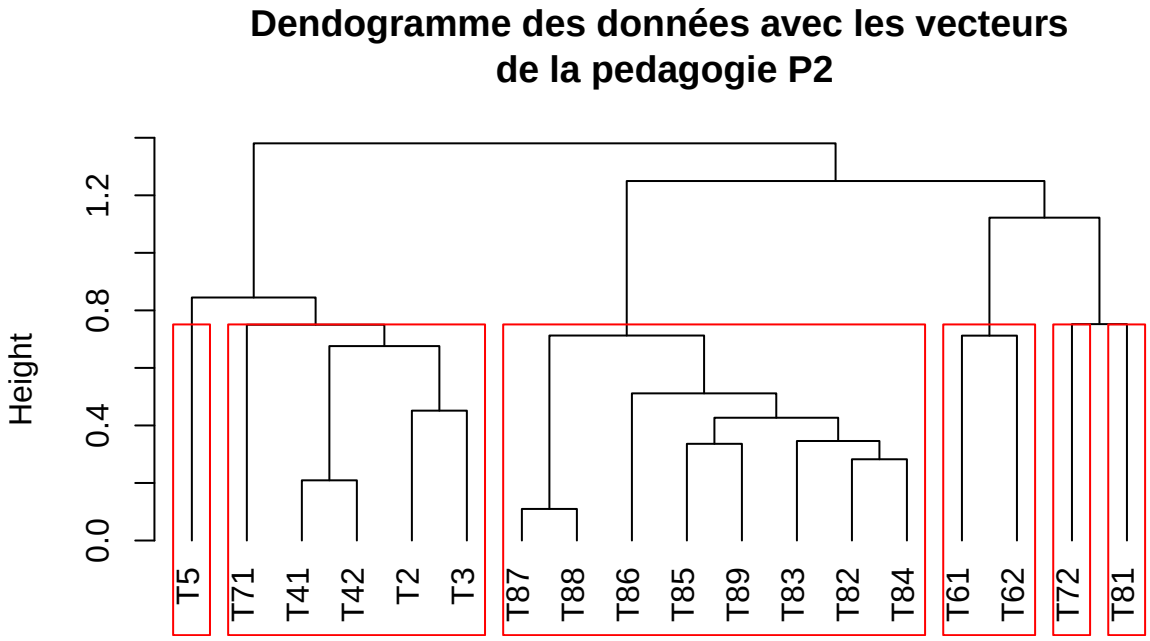


Pour les résultats du test venant des élèves de la pédagogie 2, Les questions T82 à T89 sont regroupées, la question T81 est regroupée avec T72. Les questions T31, T32 sont regroupées avec les questions T41 et T42 ainsi qu'avec la question T71 et X.quanti. T61 et T62 sont regroupées avec T21, T22 et T23 sont regroupées entre elles et T51 et T52 sont aussi regroupées entre elles. On peut d'abord remarquer les similarités avec les regroupements de la première pédagogie, nous avons toujours le couple T51, T52 et le groupe de X.quanti est constitué de de T31, T32, T41, T42 mais cette fois-ci T71 est présent dans le groupe et plus T22 et T23. Cette foi-ci seule la question T81 n'est pas regroupées avec les autres T8 et c'est la question T72 qui scinde le groupe des T8 et plus T71.

Maintenant nous allons faire le même type de classification avec les jeux de données de vecteurs et de sommes.

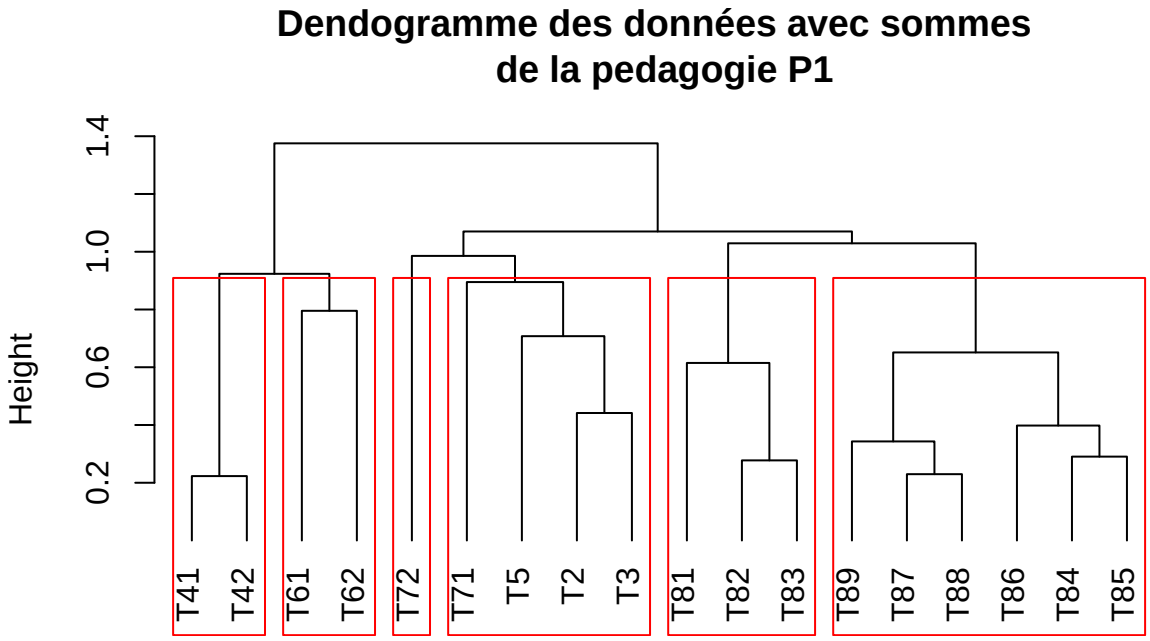


Pour les résultats du test venant des élèves de la pédagogie 1 avec les données sous forme de vecteurs nous obtenons des regroupement similaires aux précédents. T84, T85, T86, T87, T88, T89 sont regroupées avec T71 comme pour le regroupement des données générales de la pédagogie P1. T82 et T83 sont ensemble mais plus regroupées avec T81 qui est elle regroupée avec T61. T62 et T72 ne sont plus ensembles et semblent même assez éloignées. Les questions T2, T3, le couple (T41,T42) et X.quanti sont regroupées et cette fois-ci la question T5 est aussi regroupée avec.

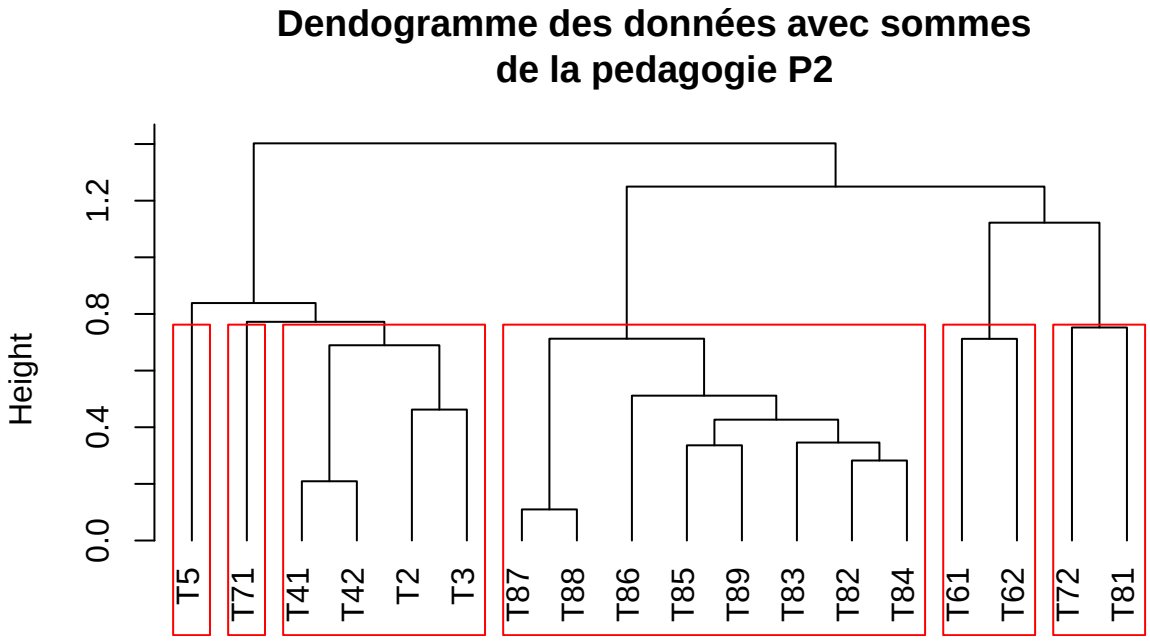


Pour les résultats du test venant des élèves de la pédagogie 2 avec les données sous forme de vecteurs nous avons aussi quelques similarités avec les regroupements des données générales de la pédagogie 2. Les questions T82 à T89 sont regroupées ensemble. T72 et T81 sont seules à définir leur groupe mais elles sont tout de même proche sur le dendrogramme. T61 et T62 sont ensemble, T2 est maintenant regroupée avec X.quanti, T5 et T3. T71 est regroupée avec T41 et T42, ces questions étaient déjà proches sur le dendrogramme des données générales.

Maintenant faisons de même avec les données avec sommes.



Pour les résultats du test venant des élèves de la pédagogie 1 avec les données sous forme de sommes nous retrouvons des regroupements déjà observés comme le groupe des question T84 à T89. T82 et T83 forment un autre groupe mais cette fois T81 est regroupée avec T61 et T62. T1, T2 T3 et T5 sont regroupées ensemble et la T72 forme une classe à elle seule.



Pour les résultats du test venant des élèves de la pédagogie 2 avec les données sous forme de sommes nous retrouvons un dendrogramme identique à celui des données avec vecteurs de la pédagogie 2. On peut en tirer les mêmes conclusions.

On remarque au final quelques différence mineures entre les dendrogrammes obtenues grâce aux données générales et ceux obtenus grâce aux données avec vecteurs mais on dénote surtout plusieurs groupe de variables qui on tendance à ressortir pour chaque type de pédagogie.

Pour la pédagogie 1 on s'aperçoit que les questions T84 à T89 sont apparemment liées, et que dans certains cas on peut aussi y lier la question T71. Les questions T82 et T83 forment aussi toujours un couple et la T81 semble liée à la T61. T1, T2, et T3 sont aussi liées.

Pour la pédagogie 2 nous observons un groupe constitué des questions T82 à T89. La question T81 semble être liée à la T72. La question T71 semble être liée aux questions T41 et T42. Les questions T1, T2, T3 et T5 semblent aussi être liées.

4. Résultats

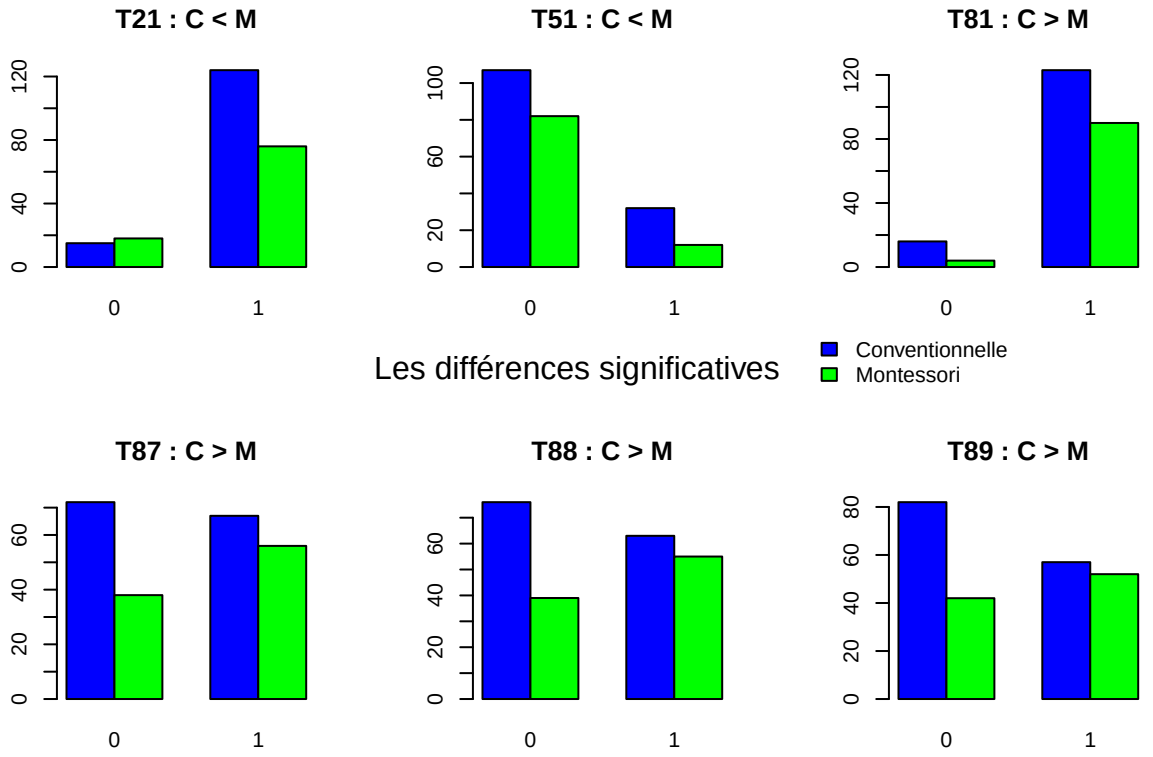
4.1 Tests statistiques

Une autre méthode permettant de comparer deux pédagogies consiste en des tests de signification ou des tests statistiques. Ici, en fonction du type de données, des **tests de proportion** ont été utilisés permettant de mettre en avant s'il existe un lien significatif entre le succès à une question et la pédagogie enseignée. Et le **t-test de Student** pour permettre de mettre en avant les liens mais cette fois ci entre les regroupements de variables et la pédagogie enseignée. Les hypothèses nuls concernent le fait que les réponses à chaque question

pour les deux pédagogies soient assez similaires. Donc, si notre hypothèse est rejetée, nous pouvons supposer que les réponses à la question “X” sont significativement différentes selon le type de pédagogie.

Les premiers tests ont été effectués sur les données originales et ont montré que seule les réponses aux questions *T72, T81, T87, T88, T89* était significativement différente pour chaque pédagogie. Ensuite, les tests ont été effectués sur de nouvelles variables, ce qui ne nous a donné que la variable *audela* comme étant significativement différente.

De plus, nous pouvons trouver la visualisation des données mentionnées ci-dessus, c'est-à-dire la visualisation de données significativement différentes. On voit donc nettement la pédagogie qui prime sur une autre ou non. Globalement il semblerait donc que la pédagogie 1 prime sur la pédagogie 2 en matière de réussite.

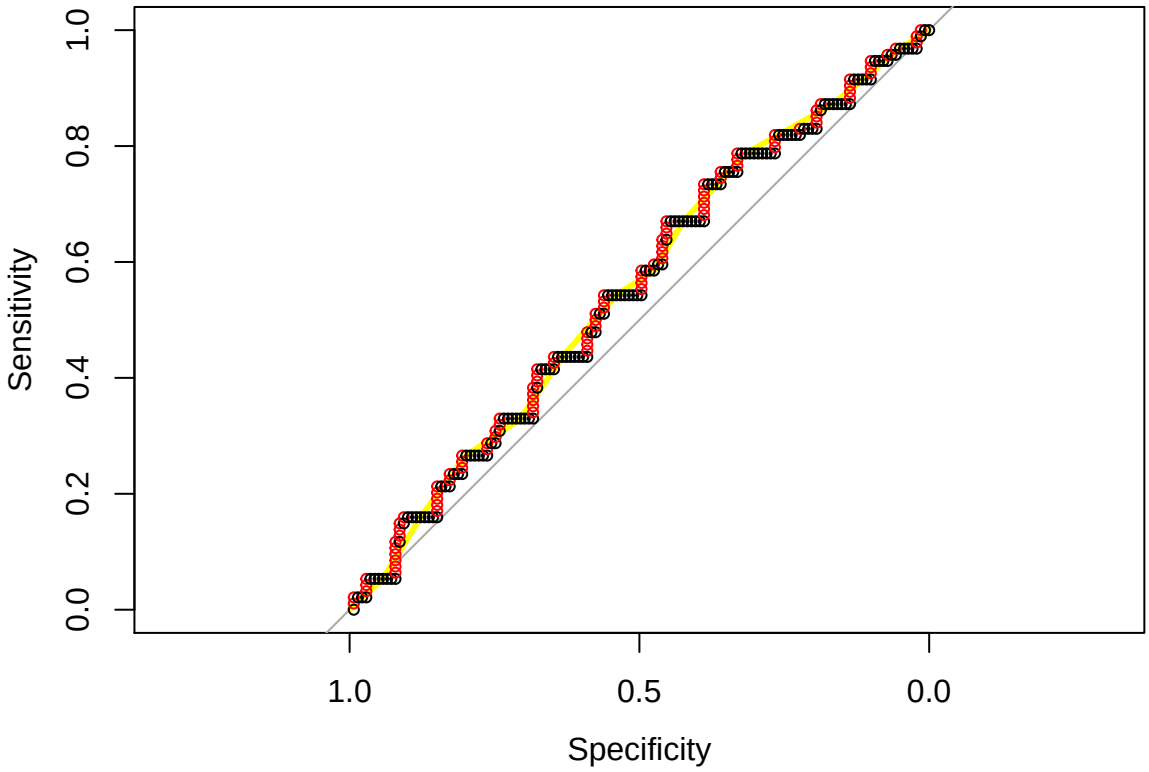


4.2 Régression logistique

La réalisation d'une régression logistique permettrait ici idéalement de prédire à quelle pédagogie appartient un élève en fonction de ses réponses au questionnaire. Nous avons donc réalisé la regression sur dans un premier temps sur toutes les variables du jeu de données, pour affiner ensuite et arriver à un modèle correct. On obtient au final un modèle composé des questions suivantes : Q21, Q22, Q31, Q41b, Q51, Q81, Q83, Q89. En soit on pourrait donc dire qu'on est capable de différencier les 2 pédagogies en fonction de leurs réponses sur les questions portants sur le dénombrement d'une collection, de la constitution d'une collection d'objets, du surcomptage (plus particulièrement de la capacité à compter 2 + 3), la création d'une collection équipotente et la reconnaissance d'une écriture chiffrée. Cela signifirait donc que chacune de ces “taches” de ce questionnaire a son importance excepté les taches 6 et 7 concernants la comparaison de deux collections et la réunion de deux collections.

Afin de valider notre modèle nous réalisons une courbe ROC. Méthode permettant de représenter le taux de bonne / fausse prédiction du modèle, plus ce taux est proche de 100% plus le modèle est bon, plus il se

rapproche des 50% plus il est aléatoire et donc inutile.



Nous obtenons finalement un AUC de 70%, ce qui équivaut à un modèle moyen.

4.3 Arbre de regression

Conclusion