Projet Cogmont

Azat, Lucas, Matthieu, Etienne

March 13, 2019

Contents

Introduction

1. Contexte

Ce projet nous à été proposé par l'Institut des sciences cognitives - Marc Jeannerod spécialisé dans la neuroscience. 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, en décembre 2016, dernière édition de classement Pisa (Programme for international student assessment) la France a fini 24ème sur 72, en recule par rapport à la dernière édition. 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 "Traditionnelle": 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 européen.

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 Ambroise Croizat à Vault-en-Velin.

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 fallut 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 dûre. 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 tache.

Recodage des variables

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 regression logistique et les arbres de regressions. Plusieurs tests ont été fait en parallèles, 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 synthétisations du nombre de dimensions pour les données qualitative et quantitatives. Cela nous permet d'appréhender plus rapidement le jeu de donnée, et avoir une première idée de ce qui diffèrent les individus entre eux (ou ce qui les rapproche). L'ACM permet dans un nuage à N dimension, en cherchant les plans orthogonaux qui maximisent la variance entre les individus, à résumer celles ci en 4 voir 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 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 regression logistique. Cette méthode permet de modéliser une classification, à l'aide notamment de l'odds ratio

2.3 Classification Ascendante Hiérachique

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.

2.4 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 explicative, 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

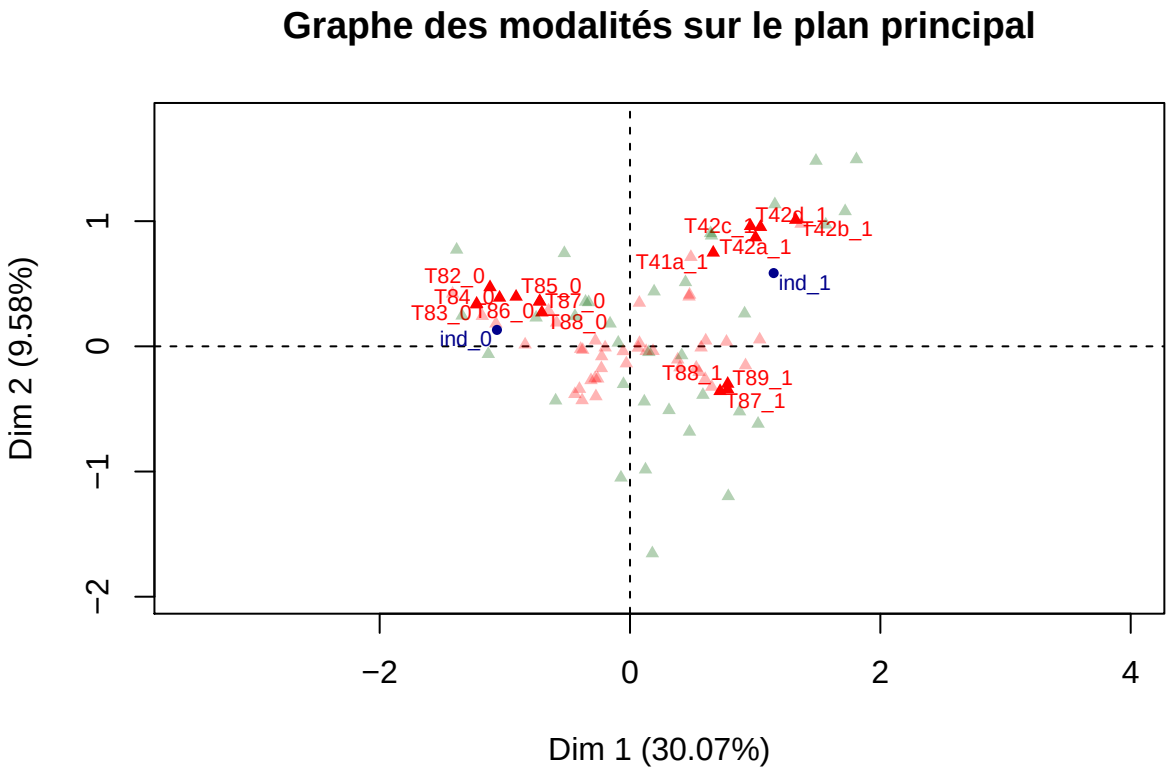
3.1 Univariée

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

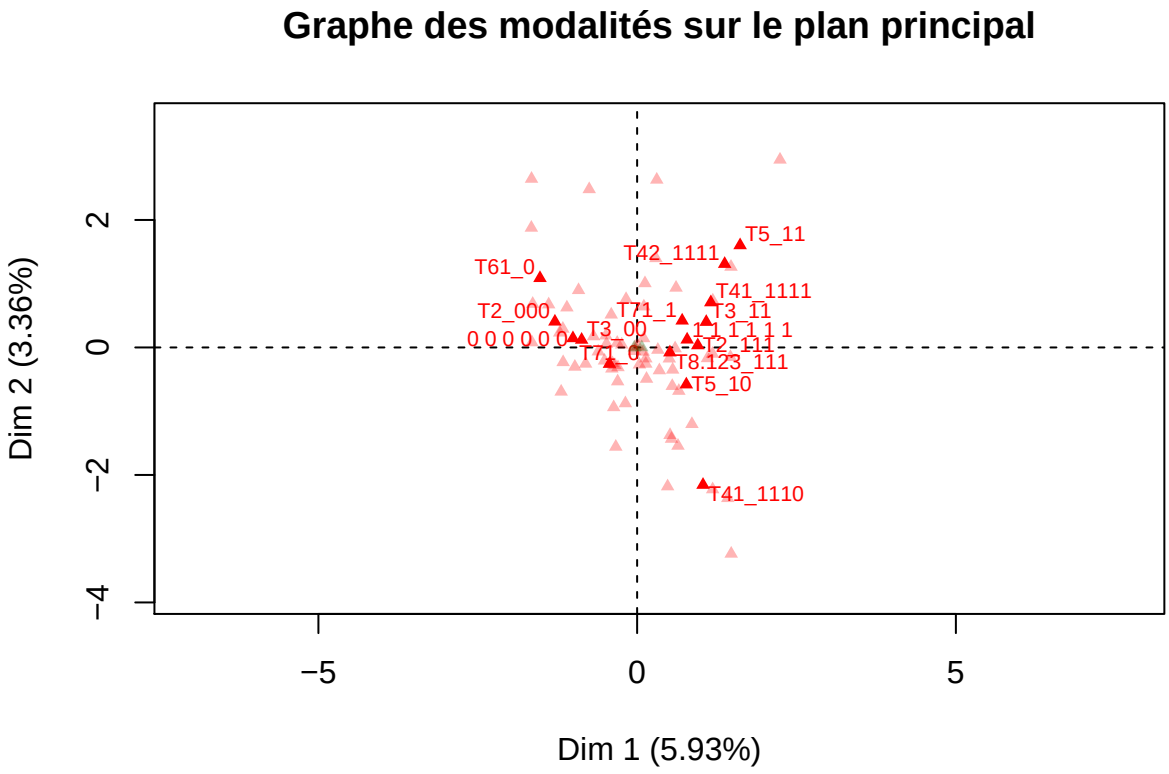
La réalisation d'une ACM comme première approche sur le jeu de données permetta de comprendre mieux ce qui différencie les individus dans notre jeu de données et à la fois d'avoir un premier résultats sur la différence entre les deux pédagogies celon cette méthode.



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 modalité 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'individus ayant réussi en totalité le questionnaire se trouve à droite alors que l'individus ayant raté en totalité le questionnaire se trouve à gauche. De plus nous pouvons voir que les questions qui descriminent 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'aurat 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 regroupe en "compétences".



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 indidivus 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 meusure, qu'il soit agé.

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.

Réalisation de 3 ACP : 1 sur le jeu de données avec les scores sans les q4 (car non numériques) 1 sur le jeu de données avec seulement deux q8 (car corrélées) 1 sur le jeu de données avec l'essentiel des q4 et des q8

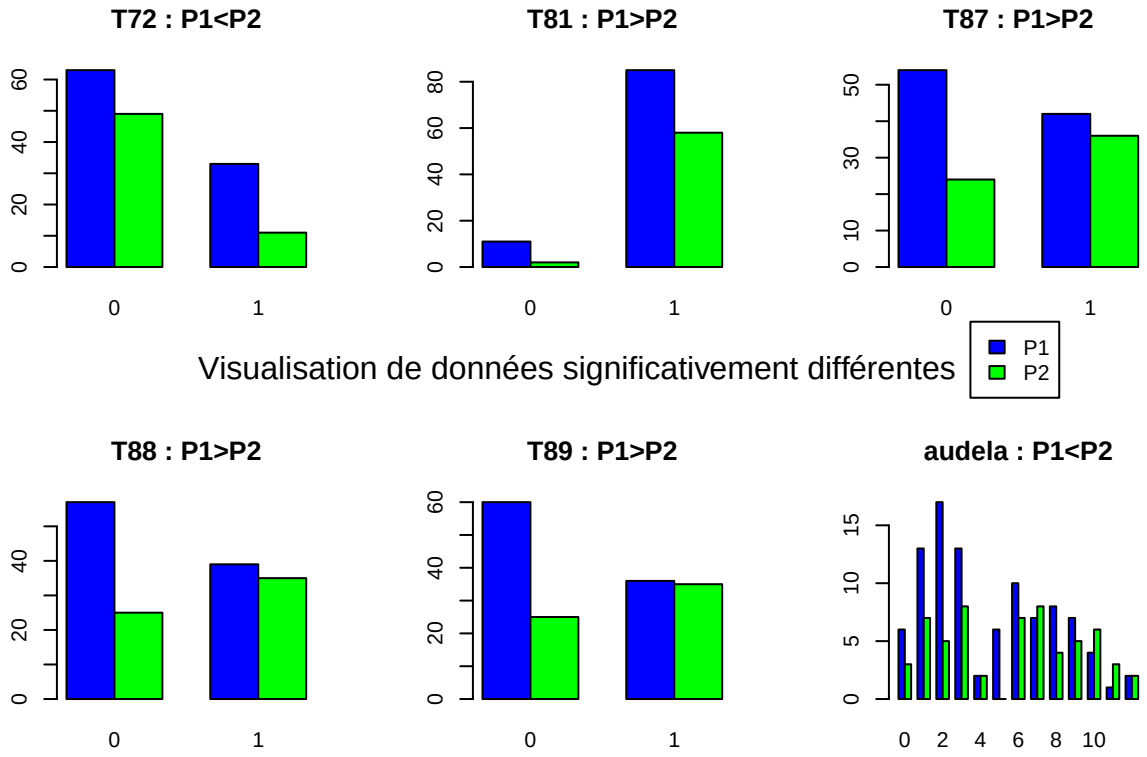
3.3 Tests

Une autre méthode permettant de comparer deux pédagogies consiste donc en des tests de signification ou des tests statistiques. Ici, en fonction du type de données, des **tests de proportion** ont été utilisés pour

comparer les succès et le **t-test de Student**. Les hypothèses nuls concernent le fait que les réponses à chaque question pour les deux pédagogies sont 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.

Après, on peut trouver la visualisation des données mentionnées ci-dessus, c'est-à-dire la visualisation de données significativement différentes. De plus, on voit ici le "gagnant".



3.4 Arbre de regression

Conclusion