Projet Cogmont

Azat, Lucas, Matthieu, Etienne

March 13, 2019

Contents

Introduction

Contexte / Sujet

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.

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.

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.

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.

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

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..

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.

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

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.

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.

Analyse exploratoire

Univariée

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. ###ACM

ACP

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

JDD 1

Les questions 8 ressortent le plus (puis la 3), comme dans les précédentes analyses : cela nous pousse à voir la matrice des corrélations Seulement 60% de l'info sur les 4 premières dimensions

Dimension 3 expliquée par la T72.

Matrice des corrélations

Grosse corrélation entre les 8* : donc ACP biaisée (et légère sur les q4 mais suffisante pour biaiser l'analyse) Etonnemment la T1 ne ressort pas comme grosse contrib

JDD 2

Aucune démarcation entre P1 et P2 sur la première dimension.

T2 / T3 sont proches et ont un cos2 élevé : un gros score en T2 implique un gros score en T3 La dimension 1 porte sur les T2 et T3 La deuxième dimension porte sur la T62

La dimension 3 porte sur la T72 La dimension 4 porte sur la T5 En conséquent nous avons quasi 1 variable / axe l'utilité de l'acp peut être remise en question

JDD 2

A nouveau aucune démarcation entre P1 et P2

La dimension 1 porte sur les questions T2 et 3 La dimension 2 porte sur les questions T41c/d (très corrélé alors qu'on ne le voit pas dans le cor) et T61

Les dimensions 3 et 4 portent sur la question T72

L'ACP permet donc dans les 3 cas d'observer des différences au sein de la population, mais qui n'est pas significative avec la pédagogie suivit par les individus.

Recherche de regroupement de variables (cah sur les variables)

Recherche de différence significative entre p1 et p2

Regression logistique

Tests sur les différentes réponses

Tests

Donc, les tests montrent que dans les réponses à certaines questions T72, T87, T88, T89, il existe des différences significatives entre deux pédagogies. Maintenant, il faut préciser à quelle pédagogie est favorable la différence.

Ainsi, le test nous montre que pour une des questions trouvées lors du premier test - T72 les résultats de P2 sont meilleurs que ceux de P1.

Dans le test "greater"(supérieur), en plus des 3 autres questions restent, nous avons ajouté T81. Le test montre que, pour les questions sélectionnés, les résultats de P1 sont meilleurs que ceux de P2.

Tests avec nouvelles variables

En ce qui concerne les tests sur les nouvelles variables: le test de la différence générale ("two.sided") montre que les résultats de la variable audela pour P1 et P2 sont significativement différents. Donc, nous devrions aller plus loin pour trouver la meilleure pédagogie pour cela.

Les tests "less" and "greater" n'ont pas ajouté de nouvelles informations concernant la différence significative. Ils ont juste précisé que les résultats de la pédagogie P2 étaient meilleurs dans la variable audela.

Visualisation de données significativement différentes

Arbre de regression

Conclusion