Analyse prédictive de données avec

Je lève la main

Introduction **2**

L'étude de cas **5**

Création de l'expérience **6**

Analyse qualitative **9**

Réduction du nombre de critères 11

Carte des individus 16

Clustering 18

Première analyse des clusters 19

Elimination des critères non pertinents 22

Deuxième analyse des clusters 24

Analyse quantitative **29**

Arbre de décision 29

Prédiction 31

Speechi

Les solutions interactives

Introduction

L'analyse prédictive a toujours été une des principales motivations de notre solution "Je Lève La Main". Dans un billet de notre blog datant de 2013, intitulé "En route vers l'analyse prédictive", nous en présentions déjà les principes :

> *Les étudiants et les professeurs se connectent à « Je lève la main » à travers un identifiant unique, tout au long de leur scolarité. Les données sont recueillies de façon fréquente et longue et l'usage qu'on peut en faire va bien au-delà des boîtiers de vote traditionnels.*

On peut utiliser ces données pour mieux personnaliser l'enseignement des élèves, mieux les aider, leur donner des schémas de travail plus adaptés, les orienter avec une plus grande fiabilité – en allant même jusqu'à prédire – peut-être – quel professeur, quelle technique pédagogique a le plus de chance de faire progresser tel ou tel élève.

Cela permet de développer de nouveaux systèmes d'évaluation prédictive (en anglais, "learning analytics").

L'utilisation systématique de ce genre de technologies, associé à des techniques de data-mining (analyse statistique des données), doit permettre d'évaluer précisément les élèves et de mieux les aider et les orienter.

En appliquant ces techniques statistiques, on peut faire progresser l'enseignement en adaptant toujours de façon plus personnalisées les matières proposées aux élèves, leur suivi, les méthodes développées par les professeurs, etc.

Il s'agit de franchir une étape supplémentaire et de permettre d'obtenir, par exemple, les résultats suivants :

- *Intervenir quand les élèves sont en difficulté (par comparaison avec les résultats moyens des autres élèves) et pouvoir leur donner un suivi spécifique*
- *Mieux personnaliser les cursus des élèves, en fonction des résultats obtenus*
- *Prédire la performance future des élèves et les orienter vers leurs matières « fortes » dans leur cursus, tout en comblant leurs manques sur les matières « faibles ».*

En résumé, l'analyse prédictive consiste à analyser l'impact de différents critères (sociologiques, technologiques, pédagogiques, …) sur les résultats des apprenants, en vue d'améliorer leur processus d'apprentissage ou de mieux les guider.

Elle s'appuie sur deux types de données :

1. la valeur des critères sur une population d'apprenants
2. les résultats de cette population sur un même ensemble de quiz

Aujourd'hui, l'application "Je Lève La Main" permet de recueillir l'ensemble de ces données :

1. la valeur des critères peut être renseignée soit par les enseignants (module "Je gère mes critères"), soit directement par les apprenants (dans leur profil).
2. les résultats à un même ensemble de quiz peuvent être obtenus en faisant en sorte que chaque enseignant récupère ces quiz sur le portail, et invitent les apprenants à y répondre dans une session (en direct ou en différé).

L'analyse des données peut être ensuite effectuée grâce à notre outil d'analyses statistiques, disponible à cette adresse :

https://statistics.jelevelamain.fr/

Un manuel d'utilisation décrit dores et déjà les manipulations de base dans cet outil (elles ne seront pas redécrites dans ce document). Il est disponible à cette adresse :

https://speechi-support.s3.amazonaws.com/JLLM-Virt/Outil-analyses-statistiques/Manuel-stats-JLLM-fr.pdf

Le présent document vient compléter ce manuel en présentant les modules avancés permettant de faire des analyses prédictives : l'Analyse en Composante Principale (ACP), le clustering et le module prédictif.

Les parties encadrées en orange sont destinées aux lecteurs non spécialistes souhaitant approfondir les concepts statistiques, elles peuvent être ignorées dans le cadre d'une lecture rapide.

A l'exception des graphiques situés dans ces parties encadrées, toutes les illustrations du manuel ont été automatiquement générées par l'outil statistique "Je Lève La Main".

Tutoriel

Les parties encadrées en vert sont destinées aux lecteurs souhaitant suivre ce manuel tout en apprenant la manipulation de notre logiciel statistique “Je Lève La Main”.

L’étude “Méthode de lecture” disponible à titre de tutoriel dans le logiciel permet de générer l’ensemble des illustrations de ce manuel, à partir de données réelles qui ont été anonymisées.

L’étude de cas

Pour illustrer notre propos, nous allons nous baser sur une étude de cas concernant le niveau de lecture des élèves de CP, dans laquelle le logiciel d’évaluation “Je Lève La Main” nous a permis d’obtenir les informations suivantes :

- les résultats d’un panel de 500 élèves de CP à un quiz de compréhension de lecture
- les valeurs de 13 critères pour chaque élève du panel :

Critères liés à l’établissement (4 critères) :

Nom	Description	Domaine	Détails
tclass	Taille de la classe	0-5	0 : < 15 élèves 1 : 15 - 19 2 : 20 - 24 3 : 25 - 29 4 : 30 - 34 5 : plus de 34
pub	Établissement public / privé	0-1	0 : Privé 1 : Public
xpprof	Nombre d’années d’expérience de l’enseignant en CP	0-25	
mlect	Méthode de lecture utilisée par l’enseignant	0-3	0 : Globale 1 : Dominante globale 2 : Dominante syllabique 3 : Syllabique

Critères liés aux élèves (9 critères) :

Nom	Description	Domaine	Détails
nlivr	Nombre de livres achetés ou empruntés par mois	0-10	
nhist	Nombre d’histoires lues par les parents par mois	0-30	
sexe	Sexe de l’élève	0-1	0 : Fille 1 : Garçon
necr	Nombre d’heures passés devant un écran par jour (tablette, télévision, ordinateur…)	0-5	

petu	Niveau d'études du père	0-5	0 : Aucun / Brevet 1 : BEP / CAP 2 : Bac 3 : Bac + 2 4 : Bac + 2 à Bac +5 5 : Bac + 8
metu	Niveau d'études de la mère	0-5	0 : Aucun / Brevet 1 : BEP / CAP 2 : Bac 3 : Bac + 2 4 : Bac + 2 à Bac +5 5 : Bac + 8
mtrav	La mère de l'élève exerce une activité	0-1	0 : Non 1 : Oui
ptrav	Le père de l'élève exerce une activité	0-1	0 : Non 1 : Oui
hdom	Heure de retour du premier parent au domicile	0-3	0 :<= 17h 1 : 17h01 - 18h00 2 : 18h01-19h00 3 : > 19h00

Création de l'expérience

Avant de commencer, nous devons définir une expérience concernant tous les élèves ayant répondu au quiz de compréhension de lecture.

Pour cela, nous créons une nouvelle expérience de type "prise de niveau" que nous nommerons "méthode de lecture".

La première fois que vous accéderez à la partie "Expérience", l'expérience "Méthode de lecture" sera automatiquement disponible. Vous pourrez donc ignorer cette étape.

Nous ouvrons cette expérience et ajoutons le quiz “Compréhension lecture” aux quiz de l’étude.

Pour ajouter le quiz “Compréhension lecture”, qui contient 20 questions, aux quiz de l’étude, il vous suffit de le chercher par son nom dans la barre de recherche puis de cliquer sur la petit flèche à droite du quiz.

Le groupe “Population interrogée”, qui comporte les personnes ayant répondu aux quiz de l’étude, est automatiquement actualisé. Il comporte 500 personnes.

Le groupe “Population interrogée” contient l’ensemble des élèves ayant répondu aux quiz sélectionnés.

L’étude pouvant être complétée par d’autres professeurs interrogeant de nouveaux élèves, le groupe “Population interrogée” peut être différent au moment où vous lisez ces lignes. Pour être certain d’utiliser les même données que celles qui ont été utilisées lors de la rédaction de ce manuel, vous pouvez utiliser le groupe “Étude méthode de lecture” en lieu et place du groupe “Population interrogée” dans la suite de ce manuel.

Groupes
Groupes disponibles
Population interrogée
Étude méthode de lecture
Population entière
Rechercher
Population interrogée
500
quizId
Étude méthode de lecture
500
Groupes de l'etude

Analyse qualitative

Dans l'onglet "ACP", partie "Configuration de l'analyse", nous sélectionnons l'étude "méthode de lecture" que nous venons de créer. Le groupe "Population interrogée" est sélectionné par défaut, nous cliquons sur "OK".

La liste des critères renseignés pour la population interrogée apparaît alors. Nous retrouvons les 13 critères de l'étude, qui sont cochés par défaut (nous ne voyons ici que les premiers critères car il est possible de les faire défiler verticalement).

Quand nous cliquons sur "Lancer l'analyse", une **Analyse en Composantes Principales (ACP)** est effectuée.

La propriété de l'ACP que nous allons exploiter ici est sa capacité à produire des graphiques fournissant un point de vue privilégié sur les données, comme le **cercle des corrélations** (qui va nous permettre de mieux identifier les ensembles de critères redondants) ou **la carte des individus** (qui va nous permettre de mieux séparer les individus sur la base des critères).

L'Analyse en Composantes Principales (ACP)

L'ACP est une technique utilisée dans des problèmes mettant en jeu de nombreuses variables. Elle combine des variables liées (ou corrélées) pour créer de nouvelles variables décorrélées les unes des autres, appelées composantes principales.

La construction des composantes principales peut être expliquée de manière géométrique.

Imaginons que nous connaissons les valeurs de critères c1 et c2 sur une population. Nous pouvons représenter chaque individu dans un repère d'axe c1 et c2. Chaque individu est un point dont les coordonnées sont les valeurs pour les critères c1 et c2.

L'ACP effectue un changement de repère, de telle manière que le nuage de points formé par les individus s'étale le plus possible selon un premier axe p1 (le long de la droite décrite ci-avant), puis selon un deuxième p2, et ainsi de suite. Ces nouveaux axes sont les composantes principales.

Le nouveau repère peut être vu comme une rotation du repère initial. Dans ce nouveau repère, les données varient le plus selon p1, puis p2, et ainsi de suite.

Dans notre étude, nous allons utiliser la capacité de l'ACP à nous fournir un point de vue privilégié sur les données. Comme les individus varient le plus sur les premières composantes principales, on peut projeter le nuage de points sur ces axes en le déformant le moins possible (c'est à dire en préservant au maximum les distances : deux individus éloignés le resteront globalement, et inversement).

Si nous reprenons le nuage de points de notre exemple, on peut constater que l'on peut le projeter sur l'axe p1 en gardant la majorité de l'information sur les distances entre individus.

Ce processus de projection permet de mieux mettre en évidence des groupes d'individus ayant des valeurs similaires de critères, appelés "**clusters**". Dans la figure ci-dessus on peut par exemple identifier un petit groupe d'individus sur la gauche, détaché du reste de la population.

Il permet également d'observer la population de manière commode quand nous sommes en présence de 3 critères ou plus : comme la visualisation de la population dans un espace de dimension 3 ou plus n'est pas aisée, il est utile de la représenter dans un espace de dimension 2 (un plan) en étant le plus fidèle possible à la répartition initiale des individus.

C'est ce qui est fait en projetant les données dans un plan défini par les deux premières composantes principales. Cette projection est appelée "**carte des individus**" (voir encadré dédié pour plus d'informations).

Réduction du nombre de critères

La partie "Corrélations entre critères" va nous aider à éliminer d'éventuels critères redondants.

Le premier graphique de cette partie est appelé "Matrice des corrélations". Pour chaque couple de critères, un point de couleur indique si ces critères sont (linéairement) liés ou non. Un point bleu indique que les deux critères sont corrélés (point bleu) ou anti-corrélés (point rouge).

Deux critères corrélés (corrélation proche de 1) évoluent dans le même sens. Par exemple, le point bleu au croisement de nlivr et nhist signifie que si nlivr augmente, hist augmente. En d'autres termes, plus il y a de livres à la maison, et plus le nombre d'histoires lues par les parents augmente. Certaines corrélations sont plus subtiles, celle entre petu et metu, qui indique les parents ont très souvent le même niveau d'études.

Deux critères anti-corrélés (corrélation proche de -1) évoluent en sens opposé. Par exemple, le point rouge au croisement de nhist et hdom signifie que si nhist augmente, hdom diminue (ou inversement : si hdom augmente, nhist diminue). En d'autres termes, plus les parents rentrent tard à la maison, et moins ils lisent d'histoires à leur enfant.

Corrélation linéaire

Le coefficient de corrélation que nous utilisons dans cette étude est le coefficient de Pearson. Il mesure la force d'un lien linéaire entre deux critères.

Il existe un lien linéaire entre deux critères c1 et c2 si les valeurs de c2 peuvent être déduites de celles de c1 (et inversement) en utilisant une équation de droite.

Cela peut être vu qualitativement en représentant les individus d'une population dans un repère d'axe c1 et c2 (chaque individu est un point dont les coordonnées sont les valeurs pour les critères c1 et c2).

Si c1 et c2 sont linéairement corrélés, alors le nuage de points formé par la population décrit globalement une droite, illustrée en bleu dans l'exemple ci-dessous :

Le coefficient de corrélation donne une mesure quantitative du lien linéaire entre les variables. Il est calculé en mesurant la distance entre la valeur réelle de c2 et la valeur théorique donnée par la droite bleue (chaque distance est représentée par un petit segment noir dans la figure ci-après).

Si ce coefficient est proche de 1 (en valeur absolue), alors ce lien est fort. L'exemple ci-dessus, les critères c1 et c2 ont un coefficient de corrélation de 0.91 environ.

Le signe du coefficient indique si les critères évoluent dans le même sens ou non. S'il est positif, c1 et c2 évoluent dans le même sens : si c1 augmente, c2 augmente (c'est le cas dans notre exemple). S'il est négatif, ils évoluent en sens contraire : si c1 augmente, c2 diminue (et inversement).

Coeff. corrélation proche de 1

Coeff. corrélation proche de -1

Si le coefficient est proche de 0, il n'y a pas de lien *linéaire* entre les variables. Il est important de noter que cela n'exclut pas qu'il y ait pas d'autres types de liens, plus complexes.

Par exemple, dans la figure ci-dessous, il y a un lien fort entre c1 et c2, puisqu'on a exactement c2 = c1 x c1 (les points décrivent une parabole). Mais la droite moyenne (horizontale) est une approximation médiocre, aboutissant à un coefficient de corrélation égal à 0.

On constate que de nombreux critères sont liés deux à deux, mais il n'est pas aisé d'identifier des ensembles de critères redondants. Pour ce faire, une vue plus appropriée est le cercle des corrélations.

Dans cette représentation, les critères pouvant être regroupés sont alignés (ils ont la même direction, mais pas nécessairement le même sens). Deux critères allant dans le même sens sont corrélés (par exemple petu et ptrav) et deux critères allant dans un sens opposé sont anti-corrélés (par exemple petu et necr). On peut facilement le vérifier dans la matrice des corrélations.

Le cercle des corrélations met clairement en avant deux axes d'étude :

1. Un axe "statut socio-professionel des parents" (spro), orienté à 45°, dans lequel on trouve petu, ptrav, metu, et necr;
2. un axe "lecture accompagnée à domicile" (lecta) orienté à -45°, dans lequel on trouve mtrav, hdom, nlivr et nhist.

Pour simplifier l'étude, nous n'allons conserver qu'une variable par axe (les autres pouvant être déduites avec une erreur très limitée compte-tenu des fortes corrélations) :

- la variable petu pour l'axe "spro"
- la variable nhist pour l'axe "lecta"

En renouvelant l'opération plusieurs fois (en décochant les critères déjà traités), nous finissons avec l'ensemble de critères suivant : petu, nhist, sexe, tclass et mlect. Nous pouvons vérifier que les critères conservés ne sont que très faiblement corrélés :

Carte des individus

Comme décrit dans l'encadré "Analyse en Composantes Principales", chaque individu peut être représenté par un point positionné en utilisant la valeur de ses critères. Une difficulté se pose quand il y a plus de 2 critères : les points sont alors dans un espace de dimension 3 ou plus, il faut alors les projeter dans un plan pour pouvoir les observer.

Dans la carte des individus, les individus sont projetés dans un plan qui est défini par les deux premières composantes principales (voir l'encadré "ACP" pour en avoir une définition). Ceci permet d'étaler au plus le nuage de points formé par les individus, et d'ainsi faciliter la visualisation des clusters présentés dans la prochaine partie.

Carte des individus

Considérons que l'on représente chaque individu d'une population par un point ayant pour coordonnées la valeur de ses critères. La représentation de cette population est un nuage de points dans un espace à n dimensions, où n est le nombre de critères.

Quand n > 2, il faut projeter le nuage de point dans un espace de dimension 2 (un plan) pour pouvoir le visualiser. Cette projection implique une altération des distances qui séparaient initialement les points. Pour illustrer cela, imaginons deux points A et B en dimension 3. La

Speechi
Les solutions interactives

projection de ces points sur le plan rouge altère beaucoup plus les distances que celle sur le plan vert. Le plan vert semble donc un meilleur choix que le rouge pour différencier au mieux les points.

Comme nous l'avons vu dans l'encadré concernant l'ACP, les composantes principales sont calculées de telle sorte que le nuage de point s'étale le plus selon la première composante principale p1, puis la deuxième p2, et ainsi de suite jusqu'à pn. Le plan défini par les deux premières composantes principales p1 et p2 est donc celui qui aboutira à une déformation minimale du nuage de points.

La projection de la population dans ce plan particulier est appelée "carte des individus". Elle permet de distinguer au mieux les éventuels groupements de points (appelés des clusters).

Dans l'exemple ci-dessous, nous considérons une population évaluée sur 3 critères c1, c2, c3. Les composantes principales p1 et p2 sont illustrées par des flèches de couleur, et le plan correspondant en pointillés. Après projection dans ce plan, les deux groupements de points présents en 3D restent clairement distinguables.

Clustering

Le clustering permet de constituer des groupes les plus homogènes possibles en termes de valeurs de critères. Ces groupes sont appelés des clusters. Les individus appartenant à un même cluster auront des valeurs de critères similaires.

La figure ci-après montre le clustering de la population réalisé par l'outil en se basant sur les valeurs des 5 critères sélectionnés. Par défaut, l'outil calcule automatiquement le nombre de clusters (ici 2).

Les clusters sont totalement distincts (leur intersection est vide). Mais en 2 dimensions, les clusters semblent se recouvrir car ils sont projetés dans un espace de dimension 2 depuis un espace plus grand (5 dimensions ici). Le fait de représenter les clusters dans la carte des individus limite ces recouvrements mais ne permet pas de les éliminer.

Nous allons ajouter ce clustering en tant que groupe dans notre expérience “méthode de lecture”. Ceci va nous permettre de manipuler ce groupe dans l’onglet “Expériences”.

Notion de groupe

Un groupe est un ensemble d’utilisateurs “Je Lève La Main” satisfaisant un ensemble de contraintes sur des critères. Dans l’onglet “Expériences”, on peut par exemple créer des groupes sur des contraintes géographiques à différentes échelles : ensemble des élèves d’un établissement, d’une ville, d’un département, etc. On peut aussi utiliser des critères (personnels ou partagés par d’autres utilisateurs) pour créer des groupes, par exemple regrouper les élèves par sexe, par catégorie socio-professionnelle des parents, etc. Voir le manuel pour plus de détails.

La partie “clustering” du module ACP permet de créer automatiquement des groupes en regroupant les individus par similarité sur les critères. Un groupe est créé par cluster.

Pour ce faire, nous cliquons sur , nous conservons le nom par défaut “clustering1” et cliquons sur OK.

Première analyse des clusters

Le clustering que nous venons de générer va nous permettre de tenter d’établir un lien entre la valeur des critères qui caractérisent ces clusters et le niveau de lecture des individus appartenant à ces clusters.

Méthode d’analyse expérimentale

L’objectif d’une analyse ACP avec “Je Lève La Main” est en effet d’obtenir des clusters permettant de mettre en évidence des différences de niveau.

Dans ce cas, cela signifie que les critères sélectionnés pour obtenir ces clusters sont pertinents et qu’ils sont liés au niveau atteint par les élèves. Dans le cas contraire, l’ACP ne met en évidence aucun lien entre critères et niveau atteint - ce qui signifie que l’expérience est négative.

Pour cela, nous allons dans l’onglet “Expériences” et sélectionnons l’étude “méthode de lecture”.

Dans la partie “Groupes”, nous constatons qu’une nouvelle partie est apparue : “ACP”. Cette partie contient tous les clusterings calculés dans la partie ACP - pour le moment uniquement le clustering nommé “clustering1”.

Nous constatons que le groupe “clustering1” est bien une partition de la population interrogée. Il contient tous les individus de la population interrogée (500 personnes), et se divise en 2 sous-groupes “Cluster : 1” (255 personnes) et “Cluster 2” (245 personnes). On a bien 255 + 245 = 500.

Nous pouvons évaluer les résultats du groupe “clustering1” sur le quiz “Compréhension lecture” en ajoutant ce groupe aux groupes de l’étude :

Ensuite, nous cliquons sur l’icône de la partie “Carte de niveau” et nous obtenons le résultat suivant :

Nous constatons qu'il n'y a pas de différence significative de niveau entre les 2 clusters. Si l'on regarde les choses un peu plus en détail (en cliquant sur le carré bleu puis vert), on peut constater que la répartition des niveaux dans ces clusters est très similaire à celle de la population interrogée :

Elimination des critères non pertinents

Valeur normalisée

Les critères peuvent avoir des ordres de grandeur très hétérogènes selon leur nature. Par exemple, on peut voir que le critère “tclass” varie entre 0 et 5 alors que le critère “sexe” varie entre 0 et 1. On a donc un facteur 5 entre ces critères.

Dans ces conditions, une différence de valeur sur un critère n’a pas du tout la même importance selon le domaine sur lequel varie ce critère. Par exemple, une différence de 1 sur “tclass” est beaucoup moins importante qu’une différence de 1 sur “sexe”. En effet, quand on parle de “tclass”, “une différence de 1” sous-entend une “différence de 1 sur 5 niveaux possibles”.

Il est donc plus pertinent de parler d’une différence *relative* de 1/5 soit 20% qu’une différence *absolue* de 1. En raisonnant ainsi, deux individus ayant une différence de 1 sur le critère “sexe” sont différents à 100%.

On passe de valeurs absolues à des valeurs relatives par un processus de normalisation qui consiste à faire le calcul suivant :

valeur normalisée = (valeur initiale - valeur minimale) / (valeur maximale - valeur minimale)

Ce calcul aboutit à un critère variant sur l’intervalle [0,1] (0 = 0%, 1 = 100% de la valeur initiale du critère) quel que soit le domaine de variation initial de ce critère.

Par exemple, si on considère un critère “année de naissance” variant sur l’intervalle [1950, 2000], on obtient les valeurs normalisées suivantes sur quelques exemples de valeurs initiales :

Valeur initiale	Valeur normalisée
1950	0
1952	0,04
1965	0,3
1974	0,48
1988	0,76
2000	1

Les valeurs obtenues sont appelées “valeurs normalisées”.

A ce stade, notre expérience a échoué puisque les clusters obtenus ne nous disent rien sur le niveau de lecture des élèves.

Analysons ce qui différencie le plus les clusters 1 et 2. Pour ceci, nous retournons dans l'onglet "ACP", partie "Valeur des critères sur les clusters", et nous demandons d'évaluer la différence de valeur de critères entre les clusters 1 et 2 :

Procédure de comparaison des clusters / Différence de valeur normalisée

Pour comparer la valeur des clusters sur les critères, l'outil procède comme suit :

1. Il calcule la valeur moyenne de chaque cluster sur chaque critère ;
2. il normalise chaque valeur pour la ramener dans l'intervalle [0,1] (voir encadré "valeur normalisée" pour plus de détails) ;
3. il calcule la différence (en valeur absolue) de ces valeurs.

En procédant de la sorte, la différence de valeur entre deux clusters sur un critère sera nécessairement entre 0 (différence de 0%) et 1 (différence de 100%), et cela, quel que soit la nature du critère.

Ceci permet de donner la même importance à chaque critère, ce qui permet de mieux identifier les critères qui varient le plus (relativement à leur ordre de grandeur). Par exemple, dans la figure suivante, le fait d'utiliser des valeurs normalisées permet de voir que c'est le critère 2 (et non le critère 1) qui varie le plus sur les clusters.

Speechi
Les solutions interactives

Utiliser des valeurs normalisées permet également d'obtenir un graphique plus lisible, car il suffit d'un seul critère variant sur un grand intervalle pour "écraser" les différences sur tous les autres (comme on peut le voir dans la partie gauche de la figure).

Nous constatons que les clusters sont totalement différents sur le critère "sexe" (la valeur 1 correspond à une différence de "100%") : un cluster ne contient que des filles, et l'autre que des garçons.

Nous avons aussi constaté que les deux clusters avaient des niveaux de lecture équivalents, nous pouvons donc en déduire que le critère "sexe" n'a pas d'impact sur le niveau de lecture.

Deuxième analyse des clusters

Le critère "sexe" n'ayant pas d'impact sur le niveau de lecture, nous pouvons l'exclure de notre étude pour tenter d'identifier les critères ayant un impact sur le niveau parmi les restants.

Cette exclusion se fait en décochant le critère "sexe" dans la liste des critères dans la configuration de l'analyse :

Nous recalculons les clusters en fixant le nombre de clusters à 4 pour bien séparer les sous-groupes tout en gardant des sous-groupes de taille raisonnable (un nombre élevé de clusters peut amener à des sous-groupes contenant très peu d'individus) .

Nous ajoutons ce nouveau clustering à l'étude sous le nom "clustering2".

Si nous regardons la composition de ce clustering, nous constatons que les sous-groupes sont de taille homogène et qu'ils possèdent au moins 100 personnes.

Si nous demandons les résultats de ce clustering sur le quiz de compréhension, nous constatons une différence significative de niveau entre les clusters 1 et 2.

Cette fois, la répartition des niveaux entre les deux clusters est nettement différente :

Nous observons également que ce qui différencie le plus ces deux clusters est le critère “petu”, puis “mlect”. Les autres critères sont quasiment identiques sur ces clusters, leur différence de valeur étant inférieure à 5%.

Nous pouvons en déduire que les critères “mect” et “petu” ont un impact sur le niveau, mais nous ne pouvons pas en dire plus. En particulier, ce n’est pas parce que ce les clusters se différencient le plus sur le critère “mlect” que ce critère a le plus d’impact sur le niveau.

Nous pouvons toutefois avoir une idée de l’évolution du niveau en fonction des critères en utilisant la fonction de partitionnement de la partie “Expériences”.

Par exemple, pour observer l’impact de “mlect” sur le niveau de lecture, nous pouvons créer une partition “mlect” sur la base du critère “mlect” (voir le manuel de l’outil d’analyses statistiques pour le détail des opérations).

La partition “mlect” contient autant de sous-groupes que de méthodes de lecture. Et comme précédemment, l’ensemble des membres de ces sous-groupes donne la population interrogée.

Si nous évaluons les résultats de la partition “mlect” sur le quiz de référence, nous obtenons les niveaux suivants :

Nous voyons clairement que le niveau évolue dans le même sens que mlect, ce qui permet d’identifier une tendance mais pas d’expliquer (et encore moins de prévoir) quantitativement des résultats. C’est le but du module prédictif que nous allons utiliser dans la partie suivante.

Analyse quantitative

La partie ACP permet d'étudier une population sous l'angle des critères indépendamment de ses résultats sur des quiz.

Comme nous l'avons vu plus haut, il est possible de tenter d'établir des liens entre critères et résultats par le biais des clusters. Cependant, cette procédure nécessite de nombreux aller-retours entre les onglets "Expérience" et "ACP", et permet d'établir des relations qualitatives uniquement.

L'onglet "Expérience" contient un module dont le but est d'établir des relations quantitatives entre critères et niveau. Il s'agit de la partie "Module prédictif". Ce module est qualifié de prédictif car il permet non seulement d'expliquer le niveau en fonction des critères, mais aussi de prédire le niveau d'un élève à partir des valeurs connues sur les différents critères de l'étude.

Arbre de décision

Quand nous appuyons sur , le module commence par séparer la population interrogée en deux sous-populations : une population d'apprentissage (80% de la population interrogée) et une population de test (20%).

La population d'apprentissage est utilisée par le module pour construire un arbre de décision. Cet arbre résume les répartitions observées des niveaux sur la population d'apprentissage en fonction de seuils sur les critères. La figure ci-après montre l'arbre obtenu sur les données de notre étude :

Arbre de décision

Dans cet arbre, les noeuds contiennent dans la partie haute un niveau moyen (% de bonnes réponses) et le pourcentage de la population ayant ce niveau. Les noeuds sont colorés en utilisant un dégradé de couleur allant du vert (niveau : 100% de bonnes réponses) au rouge (0%).

La racine de l'arbre correspond au niveau moyen de la population d'apprentissage : 49% de bonnes réponses.

Nous avons un premier choix selon la valeur du critère "petu". Si sa valeur est inférieure à 1.5 (ce qui correspond aux niveaux d'études "Sans diplôme / Brevet" et "BEP / CAP"), le niveau moyen observé est environ 43% de bonnes réponses (ce qui concerne 58% de la population d'apprentissage). Si la valeur est supérieure à 1.5 (ce qui correspond à un niveau d'études "Bac ou supérieur"), le niveau moyen observé est environ 57% de bonnes réponses (42% de la population d'apprentissage).
Ensuite, d'autres choix permettent d'affiner la valeur du niveau. Il est à noter qu'un même critère peut faire l'objet de plusieurs choix successifs. Par exemple, si nous allons à droite au choix "mlect < 1.5", cela signifie que "mlect < 1.5" est faux, c'est à dire mlect >= 1.5. Ceci correspond aux valeurs 2 (dominante syllabique) et 3 (syllabique).
Il y a ensuite un autre choix, "mlect < 2.5" permettant de distinguer ces deux cas. Si nous allons à gauche, cela signifie que "mlect < 2.5" est vrai, donc nécessairement mlect = 2 = "dominante syllabique". Sinon, mlect = 3 = "syllabique".

Nous constatons que les deux premiers critères de choix sont “petu” et “mlect”, ce qui est conforme avec nos précédentes conclusions. Cependant, ce n’est pas parce que “petu” est situé à la racine de l’arbre qu’il s’agit du critère ayant le plus d’impact sur le niveau.
En effet, l’importance d’un critère est déterminée par sa contribution à la construction de chaque feuille de l’arbre (noeuds tout en bas). Par exemple, le critère “mlect” intervient deux fois dans la construction de la feuille en bas à droite (73 / 6%) et “petu” une seule fois.

L’évaluation de l’importance des critères pouvant être complexe et souvent contre-intuitive, le module présente une partie “importance relative des critères”. On parle d’importance relative car ces résultats sont associés à *cet* arbre de décision.

Le logiciel fournit un graphique mettant en évidence l’importance relative des critères. Nous pouvons constater que l’importance relative des critères “petu” et “mect” est la même, ce qui était difficile à estimer en observant l’arbre.

On peut donc en déduire que le niveau de lecture des élèves est principalement impacté, et en proportions égales, par le niveau d’études des parents et la méthode de lecture.

Prédiction

Nous pouvons enfin prédire le niveau d’un élève à partir de la connaissance que nous avons de la valeur des critères sur cet élève.

Pour cela, nous entrons le nom d’un étudiant dans la partie “Prédiction du niveau d’une personne” et appuyons sur “Prédire”.

Le module navigue alors dans l'arbre de décision en fonction des valeurs des critères connues pour l'étudiant choisi, jusqu'à atteindre une feuille (les valeurs manquantes sont complétées en utilisant les valeurs connues sur le reste de la population). Le niveau de la feuille correspond à la prédiction du niveau pour cet étudiant.

Le module prédit un niveau de 56,6% de bonnes réponses alors que le niveau réel est de 45%, soit une erreur absolue de 11,6% environ. C'est une précision remarquable quand on pense au peu d'informations utilisées pour faire cette prédiction : uniquement les valeurs des critères "petu" (utilisé deux fois) et "mlect".

Il faut cependant rester prudent car d'une part, les résultats peuvent énormément varier d'un individu à l'autre, et d'autre part, il faut prendre soin de ne pas utiliser d'individus appartenant au groupe d'apprentissage pour évaluer la qualité des prédictions.

En effet, l'arbre de décision a été construit pour donner les meilleurs résultats possibles sur le groupe d'apprentissage. Dans certaines circonstances (groupe d'apprentissage de trop petite taille ou non représentatif de la population totale par exemple), l'arbre peut trop "coller" aux données d'apprentissage. Dans ces conditions, nous obtiendrons d'excellents résultats

sur la population d'apprentissage mais des résultats médiocres sur des individus totalement nouveaux.
C'est pourquoi il est d'usage de tester la qualité des prédictions sur des individus que le module n'a jamais vu pendant la phase d'apprentissage. C'est l'objectif du groupe de test.

Pour l'expérience "Méthode de lecture" vous pouvez prédire le niveau d'un étudiant en le cherchant par son nom (écrire "Test" dans la barre de recherche, puis choisir un étudiant).

Groupe de test

Considérons qu'un critère c2 évolue de façon périodique en fonction d'un critère c1 sur une population (comme illustré par la courbe bleue sur la figure ci-dessous) et que le groupe d'apprentissage ne contienne que des exemples pour c1 appartenant à l'intervalle [-3,3].

A la fin du processus d'apprentissage, le système pourrait tout à fait aboutir à un modèle illustré par la courbe noire. Pour c1 appartenant à [-3,3], la courbe apprise par le système et la courbe réelle sont quasiment confondues. Ainsi, sur la population d'apprentissage, l'erreur de prédiction sera proche de 0.

Or, on voit clairement que pour des individus ayant une valeur de c1 hors de [-3,3], la prédiction sera de plus en plus mauvaise à mesure que c1 augmente (en valeur absolue).

Ceci peut être détecté en utilisant des individus n'appartenant pas au groupe d'apprentissage. Ce groupe est appelé "**groupe de test**". C'est uniquement sur ce groupe de test que l'erreur de prédiction sera pertinente, car la qualité des prédictions est évaluée sur des individus que le système n'a jamais vu lors de sa phase d'apprentissage. Ainsi, nous minimisons le risque que le système ait appris "par coeur" des réponses pour ces individus. On vérifie ainsi que le système a correctement généralisé le concept qui lie les critères (une généralisation correcte ici serait un

sinus par exemple).

En pratique, on utilise 80% de la population initiale pour constituer le groupe d'apprentissage et 20% pour le groupe de test.

Dans notre exemple, nous avons constitué le groupe d'apprentissage dans l'intervalle [-3,3] pour accentuer le problème de sur-apprentissage. Mais dans la pratique, il faut que le groupe d'apprentissage et le groupe de test soient tous deux des échantillons représentatifs de la population initiale. Il faudrait donc piocher des individus dans l'intervalle capturant l'ensemble de la variabilité de c1 (l'intervalle [-8, 8] ici). L'important est que les individus de ces deux groupes soient différents.

La partie "Erreur de prédiction" présente le pourcentage du groupe de test ayant eu au plus une certaine erreur de prédiction.

On peut par exemple observer que l'erreur maximale de prédiction est environ 22% sur le groupe de test, ou encore qu'une erreur d'au plus 12% a été obtenue pour 80% du groupe de test.