PROJET D'ANALYSE DES DONNÉES QUALITATIVES

Author:
HASSINI Houda
LIEDRI Ibtissam

Janvier , 2020

Contents

1 Introduction

Un learning management system (LMS) ou learning support system (LSS) est un logiciel intégré qui accompagne et gère un processus d'apprentissage ou un parcours pédagogique des étudiants. Kalboard 360 est un LMS multi-agents, qui a été conçu pour faciliter l'apprentissage grâce à l'utilisation de technologies de pointe. Un tel système offre aux utilisateurs un accès synchrone aux ressources pédagogiques depuis n'importe quel appareil connecté à Internet.
Cet outil est utilisé dans différents pays par des personnes d'âge, de sexe , d'origine et de niveau différents ...Étant données l'ampleur et la variabilités des utilisateurs, il est important de comprendre les raisons des réussites mais surtout des échecs de ces derniers afin de le valoriser et de l'améliorer pour l'adapter aux besoins de chacun. Pour se faire les développeurs du système disposent d'un ensemble de données qui a été collecté à partir du système de gestion de l'apprentissage Kalboard 360 qu'ils ont mis a disposition sur Kaggle à l'adresse suivante: **Students' Academic Performance Dataset**.

2 Les données

Les données sont collectées à l'aide d'un outil de suivi d'activité de l'apprenant, appelé API d'expérience (xAPI). Le xAPI est un composant de l'architecture de formation et d'apprentissage (TLA) qui permet de suivre les progrès de l'apprentissage et les actions de l'apprenant comme lire un article ou regarder une vidéo de formation. L'API d'expérience aide les fournisseurs d'activités d'apprentissage à déterminer l'apprenant, l'activité et les objets qui décrivent une expérience d'apprentissage. L'ensemble de données se compose de 480 dossiers d'élèves et de 16 entités. Les caractéristiques sont classées en trois grandes catégories:

1. Caractéristiques démographiques telles que le sexe et la nationalité.
2. Caractéristiques des antécédents académiques tels que le niveau éducatif, le niveau scolaire et la section.
3. Caractéristiques comportementales telles que lever la main sur la classe, ouvrir les ressources, répondre au sondage des parents et satisfaction de l'école.

En plus des informations précédentes, nous disposons d'un classement qui a été fait selon les notes des élèves uniquement et qui les sépare en trois groupes (**Bon niveau** : H, **Niveau Moyen** : M et **Niveau médiocre** : L).
La liste exhaustive des 16 entités caractéristiques pour chacun des élèves:

Gender	sexe de l'élève
Nationality	nationalité de l'étudiant
Place of birth	Pays de naissance de l'étudiant
Educational Stages	Le niveau scolaire (Primaire, Collège, Lycée)
Grade Levels	La classe présice de l'étudiant
Section ID	Identifiant de la salle où l'élève a cours
Topic	La matière
Semester	Le semestre
Parent responsible for student	le parents responsable de l'étudiant
Raised hand	Le nombre de fois où l'élève a levé la main en cours
Visited resources	Le nombre de fois où l'élève a visité les ressources mises à disposition
Viewing announcements	Le nombre de fois où l'élève vérifie les nouvelles annonces
Discussion groups	Le nombre de fois où l'élève à participé aux groupes de discussion
Parent Answering Survey	Les parents ont-ils répondu à l'enquête de qualité
Parent School Satisfaction	les parents sont-ils satisfaits de l'outil
Student Absence Days	les jours d'absences de l'élève (above-7, under-7)

Un descriptif plus détaillé est disponible en annexe.

3 Objectif

Dans ce rapport nous aurons comme objectif d'identifier la façon dont les différentes entités caractéristiques contribuent à la réussite ou à l'échec d'un élève. C'est à dire que nous allons essayer d'expliquer le résultat scolaire des élèves par les différentes caractéristiques collectées.

4 Étude exploratoire

4.1 Analyse unidimensionnelle

Comme nous l'avons précisé avant , notre objectif est d'expliquer les résultats scolaire par les caractéristiques collectées, nous commençons donc d'abord par une étude exploratoire de chacune de ces caractéristiques ainsi qu'une analyse des résultats scolaires de façon générale, les résultats les plus marquants qui ont ressortent sont les suivants:

- Pour tous les individus de l'étude nous possédons l'ensemble des informations (16 variables) , il n'existe aucune valeur manquante dans le jeu de données.
- Parmi les 16 variables uniquement 4 sont des variables quantitatives, il s'agit de **Raised hand**, **Visited resources**, **Viewing announcements** et **Discussion groups**.
- Les élèves suivent des études de niveaux différents, la majorité d'entre eux sont issus du collège, les élèves des écoles primaires sont aussi assez présents contrairement aux lycéens qui eux sont moins présents sur l'ensemble des individus de l'étude.

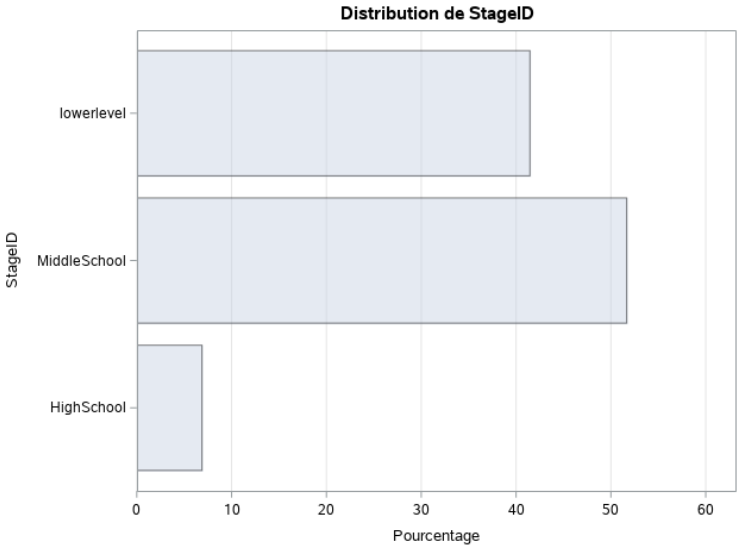


Figure 1: Distribution des niveaux scolaires dans le jeu de données

- Les élèves sont de sexe différent, mais on note que les garçon sont plus présents que les filles. L'origine varie d'un individu à l'autre, on remarque que les étudiants sont majoritairement originaire d'un pays arabe ou ayant une nationalité arabe (plus précisément les pays arabes du moyen orient où la situation vis-à-vis des droits de la femme reste délicate), ce qui pourrait expliquer la forte présence de garçon par rapport aux filles. De plus deux pays dominent à la fois la nationalité et le pays de naissance, il s'agit du Koweït et de la Jordanie.

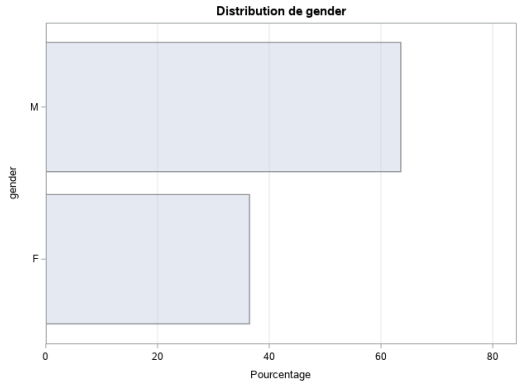


Figure 2: Distribution des élèves selon le sexe

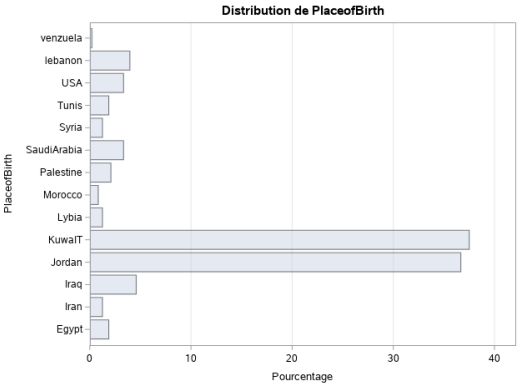


Figure 3: Pays d'origine des élèves

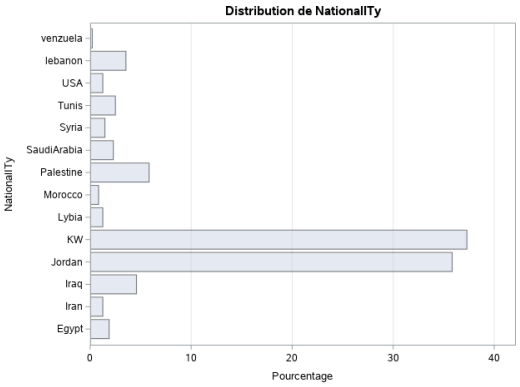


Figure 4: Nationalité des élèves

- Les élèves suivent des cours dans des niveaux différents certains niveaux regroupent de nombreux élèves comme est le cas pour **G02**,**G04**, **G06**, **G07**, **G08**, tandis que d'autre ont très peu d'élèves.

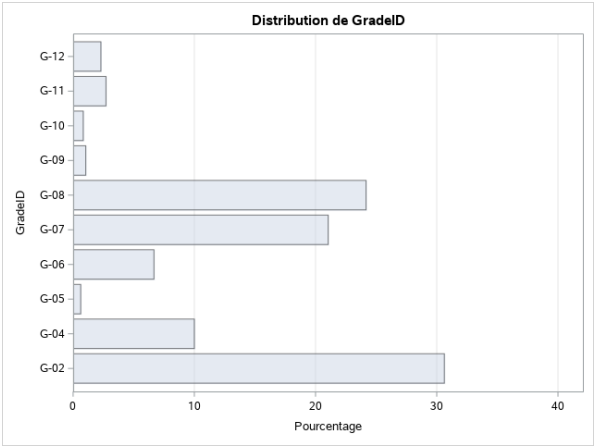


Figure 5: Les différentes classes des élèves

- Les élèves suivent des cours dans des matières différentes certaines scientifiques d'autres littéraires, il existe aussi des élèves qui suivent des cours relatifs à la théologie et d'autres qui s'intéressent au social.

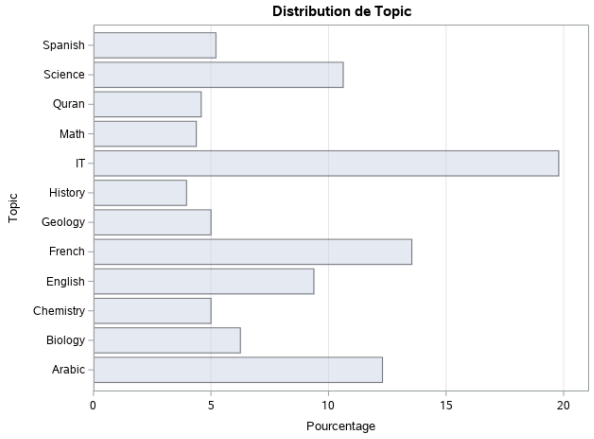


Figure 6: Les différentes matières suivis

- Les deux variables Pays de naissance et Nationalité indiquent une forte présence de deux pays (Koweït et Jordanie) les autres pays ont de très faibles proportions.
- L'étude réalisée a donnée lieu à un classement en 3 groupes, mais la visualisation ci-dessous nous montre que les élèves ne sont pas réparties de la même façon dans les différents groupes. Le groupe **M** domine, d'où l'intérêt de comprendre les raisons de réussite et d'échec pour améliorer le fonctionnement de l'outil.

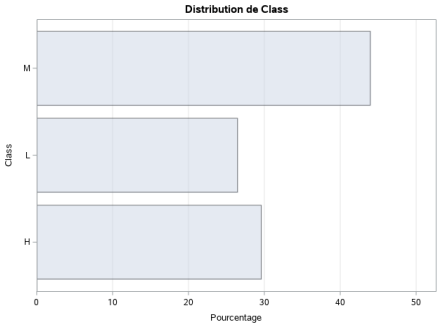


Figure 7: Proportion des trois groupes **H**,**M** et **L**

- Les élèves diffèrent en terme de taux de participation en cours durant un semestre. Certains élèves ne participent jamais(0 participation/semestre), d'autres élèves participent beaucoup(à hauteur de 100 participation par semestre) avec une moyenne de participation en cours aux alentours des 50.

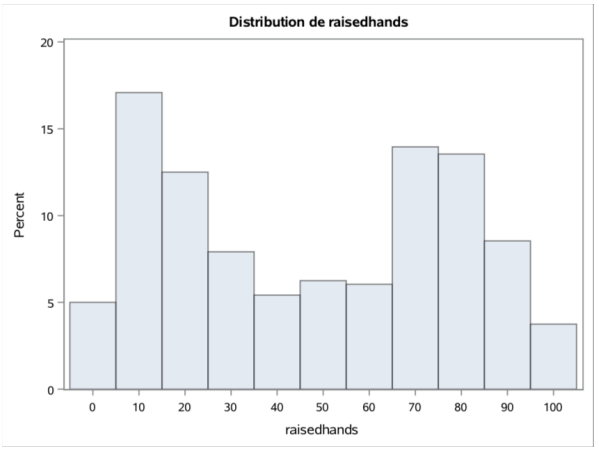


Figure 8: Histogramme du nombre de participation en classe/semestre

- Seulement 1.2% des élèves consultent souvent les annonces relatives au cours, contre plus de 7% qui ne les consultent guère. On pourra commencer à ce poser des questions concernant cette variable et sa relation avec les résultats scolaires.

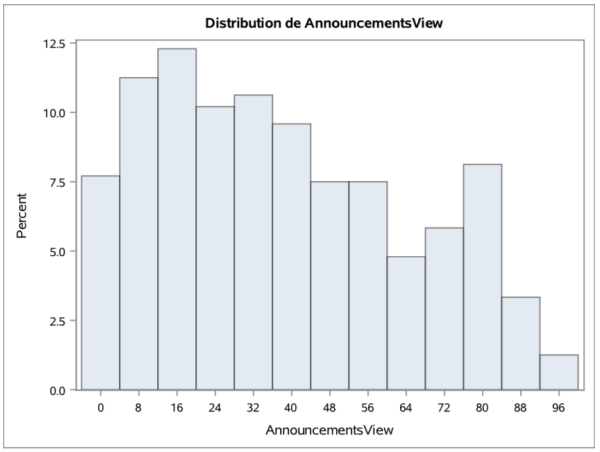


Figure 9: Histogramme du nombre de consultation des annonces

- On peut distinguer ici trois groupes d'élèves, certains qui consultent très peu les supports de cours mis à disposition, d'autres moyennement, tandis qu'on remarque certains élèves qui consultent souvent les supports de cours mis à leur disposition.

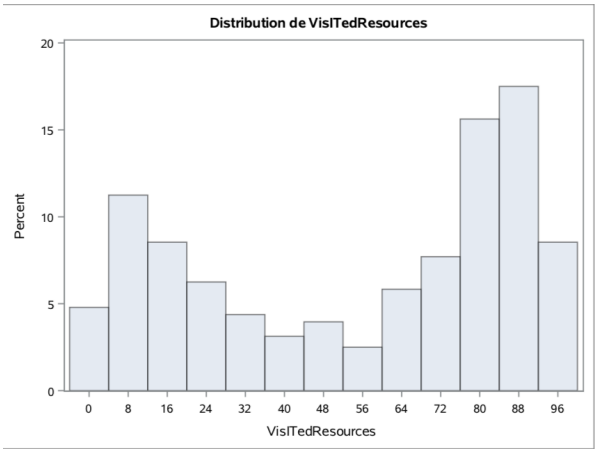


Figure 10: Histogramme du nombre de consultation des supports de cours

- On remarque que tous les élèves ont au moins participé 4 fois à une ou plusieurs discussions de groupes lors d'un semestre. On distingue là encore des élèves qui participe souvent, d'autres moyennement et d'autres peu ou très peu.

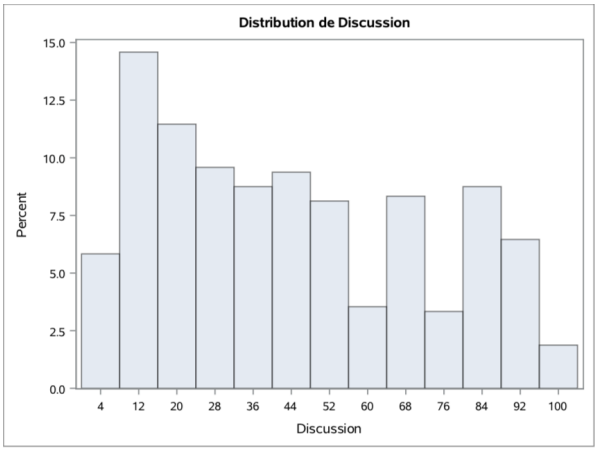


Figure 11: Histogramme du nombre de participation dans des conversation de groupes

Cette analyse unidimensionnelle de ces quatre variables quantitatives, va nous permettre, avec l'analyse des quartiles dans la transformation des ces variables en des variables qualitatives.

4.2 Analyse bi-dimensionnelle

4.2.1 Étude de la distribution des groupes selon les différentes caractéristiques

Après avoir explorer les entités caractéristiques une à une , nous nous intéressons maintenant aus relations qui les lient les unes aux autres et surtout à la relation qui existent entre ces variables et les notes obtenus par les élèves, c'est-à-dire au classement réalisé.
Ce qui pourrait expliquer naturellement les notes d'un élève est sa participation en cours, son utilisation du matériel mis à disposition, la consultation fréquentes des annonces de ces professeurs et son implication dans des discussions à propos des matières qu'ils suit.
On peut voir que comme attendu les élèves du groupe **H** ont tendances à avoir une valeur élevé pour les 4 variables quantitatives **Raised hand**, **Visited resources**, **Viewing announcements** et **Discussion groups** contrairement aux élèves qui font partie du groupe **L**.
Le groupe **H** est dominé par la gent féminine, même si elles sont moins nombreuses dans le jeu données comparé aux garçons.

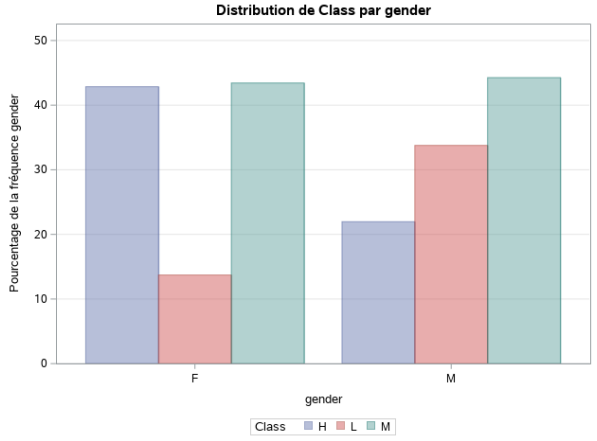


Figure 12: Proportions des deux sexe dans les trois groupes **H**,**M** et **L**

Les matières choisies par les étudiants sont aussi une variable qui détermine à quel groupe l'individu aurait plus de chance d'appartenir. Comme nous pouvons le voir sur le tableau suivant:

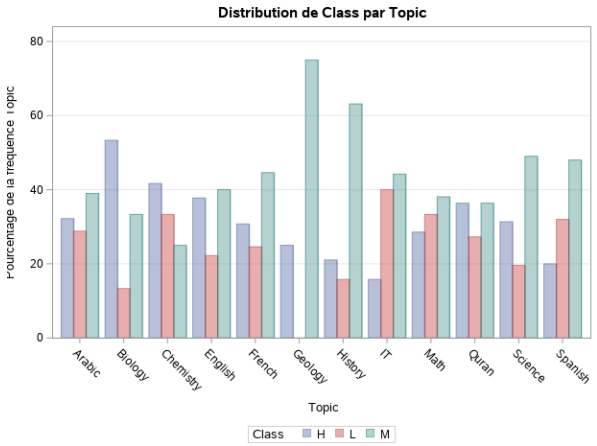


Figure 13: Proportions des matières dans les trois groupes **H**,**M** et **L**

Les matières scientiques comme **IT**, **Mathematiques** et **Chimie** correspondent aux matières avec un grand taux d'échecs contrairement à **Geologie** où aucun n'élève n'a échoué.
On note aussi, que selon le semestre, on observe une proportion différente d'individus dans chacun des groupes.

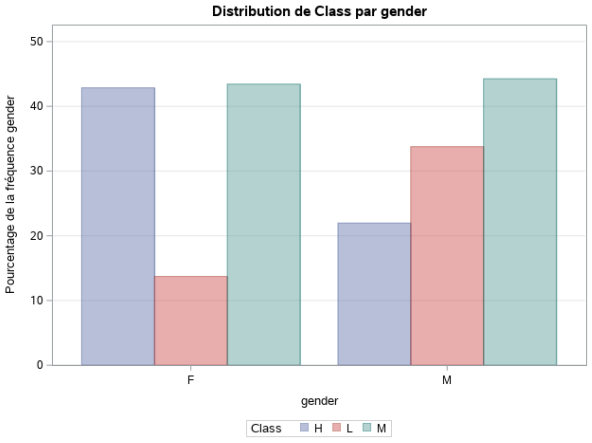


Figure 14: Les trois groupes **H**,**M** et **L** en fonction du semestre

4.2.2 Étude de l'impact des variables explicatives sur les groupes

Pour visualiser la force de la relation entre les 16 caractéristiques et le regroupement réalisé nous nous intéressant aux **V de cramer** qui est basé sur le test du χ^2 de Paerson. Ce choix est motivé par le fait que ce coefficient et borné contrairement au χ^2 et il reste stable si l'on augmente la taille de l'échantillon dans les mêmes proportions inter-modalités. Sur l'ensemble de notre jeu de données, on remarque de très fortes valeurs de **V de Cramer** sur les variables décrites dans la figure suivante:

Les trois variables : **Nombre d'absence**, **ParentansweringThestudy**, **Relation** sont ceux avec les coefficients les plus forts.

Statistiques pour la table de Class par StudentAbsenceDays

Statistique	DDL	Valeur	Prob
Khi-2	2	225.2048	<.0001
Test du rapport de vraisemblance	2	264.4685	<.0001
Khi-2 de Mantel-Haenszel	1	18.9847	<.0001
Coefficient Phi		0.6850	
Coefficient de contingence		0.5651	
V de Cramer		0.6850	

Statistiques pour la table de Class par ParentAnsweringSurvey

Statistique	DDL	Valeur	Prob
Khi-2	2	95.3646	<.0001
Test du rapport de vraisemblance	2	100.0803	<.0001
Khi-2 de Mantel-Haenszel	1	7.2768	0.0070
Coefficient Phi		0.4457	
Coefficient de contingence		0.4071	
V de Cramer		0.4457	

Statistiques pour la table de Class par Relation

Statistique	DDL	Valeur	Prob
Khi-2	2	81.3655	<.0001
Test du rapport de vraisemblance	2	83.9076	<.0001
Khi-2 de Mantel-Haenszel	1	35.4672	<.0001
Coefficient Phi		0.4117	
Coefficient de contingence		0.3807	
V de Cramer		0.4117	

Figure 15: Les **V de Cramer** les plus haut observés

Pour la première variable nous observons un coefficient de Cramer de 0.68 ce qui parait logique, la performance scolaire est très liée à la présence de l'individu lors des cours, un individu toujours absent a très peu de chance de réussir etc...
Les deux autres variables ont des coefficients de l'ordre de 0.40 ce qui indique que les variables ont une forte liaison avec les groupements réalisés, on peut penser que l'impact de parents est très important dans la scolarité des enfants, plus le parent s'intéresse à son enfant est participe à son apprentissage plus celui-ci a de chance de réussir.
On remarque aussi que les variables **Nationnality**, **PlaceOfBirth**, **Topics**, **Grade** ont de très faibles valeurs pour ce même coefficients mais ceci est peut-être dû au nombre de modalités très élèvé par rapport à la taille de l'échantillon étudié. Cette observation a motivé la mise en place de regroupement comme suit:

- **Nationnality** et **PlaceOfBirth** majoritairement constitué d'individu du Koweït et la Jordanie, nous avons donc garder ces deux modalités et nous avons transformé les autres en **"others"** c'est-à-dire des individus venant d'un autre pays, on plus sur notre ensemble des données c'est deux variables sont identiques pour chacun des individus donc nous avons décidé de ne garder que la variable indiquant la nationalité .
- **Grade** Cette variable décrit la classe de chaque individu. Pour cette variable nous avons testé plusieurs regroupements de classes en regroupant les classes successives (par 2, 3, 4...) et nous avons garder la configuration qui maximisait le χ^2 . Le groupement optimal est celui en 4 modalités crée en regroupant les modalités 3 à 3:
  - G01-03: Grades1
  - G04-06: Grades2
  - G07-09: Grades3
  - G09-12: Grades4
- **StageID** La variable StageID contient de l'information relative au niveaux d'études(low, middle,high). Cette information est aussi contenue dans la variable GradeID. Suite à cette redondance d'information, on a décidé de supprimer la varibale StageID.
- **Topics** Pour cette variable nous n'avons pas obtenu une amélioration significative de χ^2 en faisant des regroupement (matière scientifique , matière littéraire ..), nous avons donc décidé de garder les 7 modalités.

Nous avons choisi de transformer toutes les 4 variables quantitatives(**Raisedhands, AnnoucementsView, Discussion, VisIsTedRessources** en des variables qualitatives en se basant sur l'analyse des valeurs des quartiles et quelques statistiques descriptives comme la valeur moyenne et les valeurs extrêmes. Cette transformation est motivée par le fait que ces variables peuvent être vu comme des variables représentatives d'une fréquence étant donné que leurs valeurs sont bornés à 100. Nous avons fait recours à cette transformation afin d'homogénéiser la nature des variables dans notre jeu de données afin de n'avoir que des données qualitatives car la proportion des données quantitatives est minime par rapport au données qualitatives.
Par exemple, pour la variable **Discussion**:

	count	mean	std	min	25%	50%	75%	max
Discussion	480.0	43.283333	27.637735	1.0	20.0	39.0	70.0	99.0

Figure 16: Statistiques descriptives de la variable Discussion

On remarque que les valeurs des quartiles **Q1** et **Q2** sont relativement proches. Ces deux valeurs sont aussi proches de la moyenne de la variable(**mean = 43**), ainsi la transformation de la variable **Discussion** se fera en 3 modalités:

Intervalle	$\leq Q_1$	$> q_1$,$\geq mean$	$>$ mean
Modalités	**peu**	**moyen**	**souvent**

5 Analyse des correspondances multiples

Nous nous intéressons maintenant à l'analyse multidimensionnelle de notre jeu de données. Nous optons pour une analyse des correspondances multiples (ACM), une méthode factorielle adaptée aux tableaux dans lesquels un ensemble d'individus est décrit par un ensemble de variables qualitatives.
Nous effectuons une ACM avec l'ensemble de données obtenues après traitement en mettant la variable des groupements en supplémentaire. Afin de simplifier la visualisation, nous nous contentant d'étudier les observations dans le plan composé par les deux premiers axes.

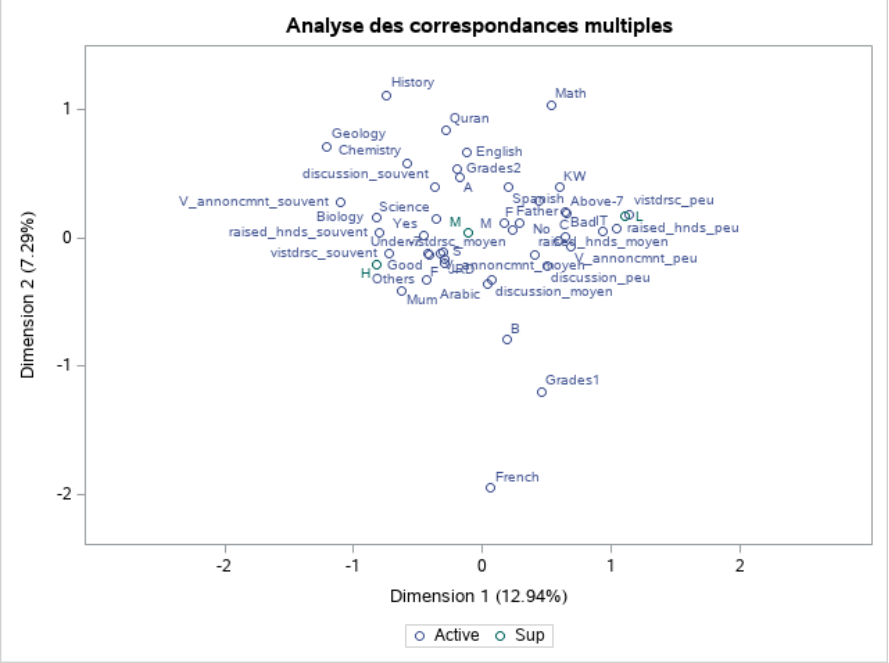


Figure 17: Nuage des modalités

Nous nous intéressons pour interpréter les résultats uniquement aux modalités dont la contribution est significativement supérieur à la masse. On peut voir dans le tableau suivant la liste des modalités retenues:

modalité	Masse	contribution axe 1	contribution axe 2
Grades1	0.0219	0.0164	**0.2016**
B	0.0249	0.0031	**0.0988**
Biology	0.0045	**0.0110**	0.0007
Chemistry	0.0036	0.0044	**0.0076**
English	0.0067	0.0003	**0.0187**
French	0.0097	0.0002	**0.2348**
Geology	0.0036	**0.0187**	**0.0112**
History	0.0028	0.0056	**0.0221**
IT	0.0141	**0.0445**	0.0002
Math	0.0031	0.0032	**0.0211**
Quran	0.0033	0.0010	**0.0145**
F	0.0365	**0.0111**	0.0028
Mum	0.0293	**0.0413**	0.0314
raised_hnds_peu	0.0179	**0.0701**	0.0006
raised_hnds_souvent	0.0336	**0.0768**	0.0003
vistdrsc_peu	0.0354	**0.0677**	0.0036
V_annoncmnt_peu	0.0363	**0.0618**	0.0012
V_annoncmnt_souvent	0.0173	**0.0756**	0.0087
Above-7	0.0284	**0.0432**	0.0063

En examinant les variables sélectionnés précédemment et en se basant cette fois-ci sur les coordonnées, on peut voir que l'axe 1 se caractérise à sa gauche (valeurs négatives)par la présence d'individus suivant des cours de biologie, de géologie et de science dont le parent responsable est la mère, c'est des individus qui participent souvent en classe et qui regardent souvent les annonces. Ce coté de l'axe s'oppose à sa droite, du coté des valeurs positives nous retrouvons des individus qui participent peu en classe, ne visitent pas souvent les ressources et regardent rarement les annonces. Ces individus du coté là de l'axe sont aussi caractérisés par des absences supérieures à 7 jours, c'est des individus qui ont suivi des cours d'informatique lors du premier semestre.
Le bas de l'axe 2 quant à lui ce caractérise par des individus qui suivent des cours de langue française, ces individus appartient à la classe B et sont des élèves des classes (G-01 à G-04). Tandis que le haut de l'axe regroupent des individus qui suivent des cours de chimie, d'anglais, de géologie, d'histoire, de maths et de Quran.
Si on s'intéresse maintenant à la variable des groupements que nous avons mis en supplémentaire, on remarque que la modalité **L** se positionne dans le coté droit de l'axe 1 et très proche du centre de l'axe 2, cette modalités et donc décrites par la même description que le coté droit de l'axe c'est a dire que cette modalités et rencontrés fréquemment chez des individus qui participent peu en classe, ne visitent pas souvent les ressources et regardent rarement les annonces. Ces individus du coté là de l'axe sont aussi caractérisés par des absences supérieures à 7 jours, c'est des individus qui ont suivi des cours d'informatique lors du premier semestre.
La modalité **H** s'oppose à la modalité **L** en se positionnant à gauche de l'axe 1 mais très proche de la valeur 0 sur l'axe 2, elle est observée chez les individus suivant des cours de biologie, de géologie et de science dont le parent responsable est la mère, c'est des individus qui participent souvent en classe et qui regarde souvent les annonces.
La modalité **M** n'est pas bien représentée car sa qualité de représentation est de 0.001. nous ne pouvons pas l'interpréter.

6 Classification hiérarchique

L'étude unidimensionnelle et bidimensionnelle de notre jeu de données suggère fortement la présence de groupes sous-jacents, les descripteurs varient d'un individus à l'autre formant ainsi des groupes, par exemple les individus qui sont lycéens et qui suivent les mêmes matières ... Sous ce principe nous avons choisi d'explorer cette possibilité en réalisant une classification hiérarchique ascendante qui consiste à rassembler des individus selon un critère de ressemblance défini au préalable qui s'exprimera sous la forme d'une matrice de distances, exprimant la distance existant entre chaques individus pris deux à deux. Nous effectuons cette classification sur l'ensemble des nouvelles variables obtenus par ACM.
Le résultats de la classification représenté sur le plan 1-2 est le suivant :

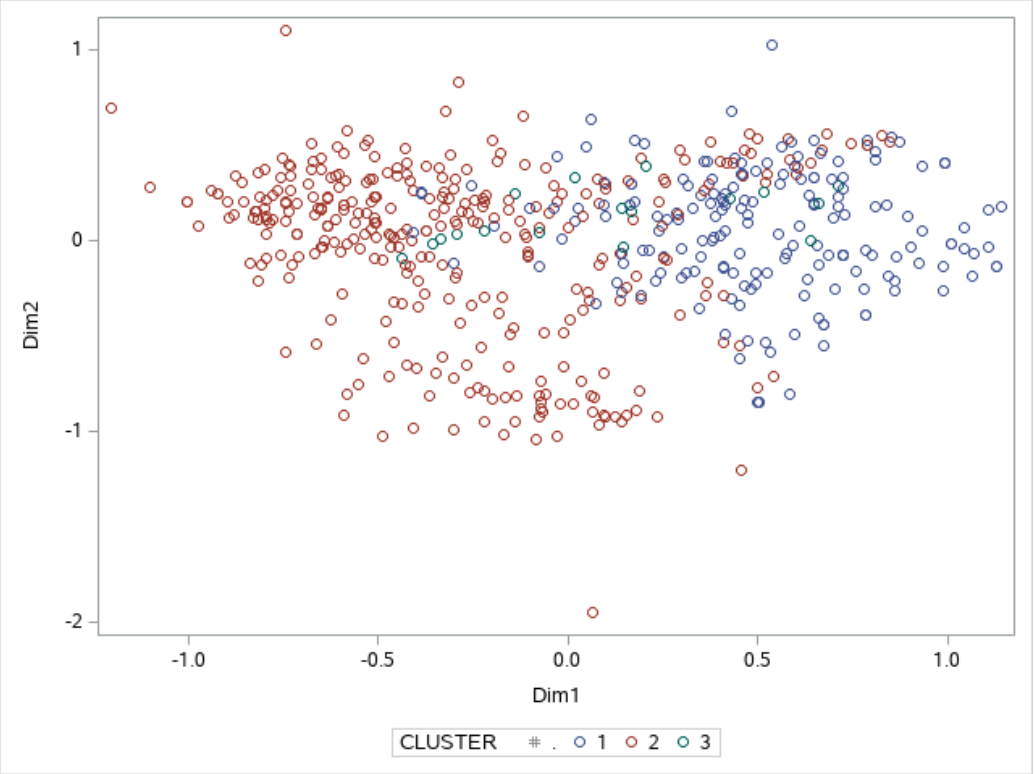


Figure 18: Nuage des modalités

Malheureusement nous n'avons pas pu estimer le nombre optimale de cluster, ni interpréter les résultats de la CAH grâce au barycentre des cluster car la version étudiante que nous utilisons de SAS bloque l'accès à l'ensemble des informations de la classification comme vous pouvez le voir dans l'annexe 2 nous pouvons pas avoir le tracer du dendrogramme. Il aurait fallu d'abord estimer le nombre optimale de cluster grâce au dendrogramme, puis effectuer une classification avec ce nombre de classe puis ensuite interpréter chaque classe en comparant son barycentre au barycentre du nuage de point en utilisant les nouvelles coordonnées obtenues grâce à la classification hiérarchique. Cependant, nous avons testé plusieurs valeurs de nombre de cluster afin de pouvoir en tirer quelques conclusions, faire une représentation, obtenir deux classes qui sont bien séparées et une classe qui est plutôt confondue, en utilisant un nombre de cluster plus élevé nous obtenons des résultats similaires quelques classes bien séparées et une classe ou deux qui viennent se confondre avec les autres.
Nous pouvons relier ces résultats a ce que nous avons observé précédemment lors de l'interprétation de l'ACM. Tout d'abord nous remarquons que les classes ne sont pas bien séparées car nous avons fait la classification sur l'ensemble des dimensions et nous représentons les résultats sur les deux premiers axes, les classes sont alors peu séparés, d'une autre part, nous retrouvons un résultat similaire à ceux de l'interprétation de l'ACM, deux classes différentes se situent d'un coté et d'un autre du plan. la classe confondus avec les deux autres pourrait correspondre à la classe **M** qui n'était pas caractérisée par les variables sur le plan de l'axe1 et de l'axe deux lors de l'interprétation de l'ACM. Afin de relier ces clusters à ceux réalisés sur la base des notes,

nous devrions comparer les distributions des trois modalités sur les cluster mais nous n'avons pas réussi à retrouver les résultats et à les récupérer comme nous l'avons montrer dans l'annexe 2.

7 Sélection de variables

Nous souhaitons maintenant prédire notre variable de groupements grâce aux variables explicatifs dans un objectif de mieux comprendre les interactions entre les deux et caractériser. Avant de procéder à ceci nous avons décidé de faire une selection de variables afin de choisir uniquement les dimensions qui caractérisent le mieux notre variable cible qui est le groupement en trois classes. Pour ceci nous avons réalisé une sélection de variable avec la méthode stepwise sur les différentes dimensions obtenues grâce à l'ACM. La méthode Stepwise sélectionne les variables en supprimant des prédicateurs du modèle existant ou en en y ajoutant sur la base des résultats d'un test F. C'est une méthode de sélection ascendante et d'élimination descendante. Les résultats obtenus sont les suivants: La décision de ne pas se contenter de choisir les deux ou trois premiers

Synthèse de la sélection Stepwise

Etape	Nombre dans	Saisi	Supprimé	R carré partiel	Valeur F	Pr > F	Lambda de Wilk	Pr < Lambda	Corrélation canonique moyenne au carré	Pr > ASCC
1	1	Dim1		0.1032	27.37	<.0001	0.89684534	<.0001	0.05157733	<.0001
2	2	Dim9		0.0152	3.65	0.0266	0.88325536	<.0001	0.05882976	<.0001
3	3	Dim13		0.0116	2.79	0.0625	0.87298267	<.0001	0.06432888	<.0001
4	4	Dim8		0.0108	2.59	0.0762	0.86353228	<.0001	0.06947392	<.0001
5	5	Dim21		0.0097	2.31	0.1001	0.85515321	<.0001	0.07411261	<.0001
6	6	Dim30		0.0096	2.27	0.1039	0.84697190	<.0001	0.07875954	<.0001

Figure 19: Résultats de la sélection de variable

axe arbitrairement pour effectuer nos apprentissages dans la suite, vient du fait que les axes sont triés dans l'ordre d'inertie ce qui ne signifie pas qu'elles sont les plus explicatives de la variable cible étudiée car ce n'est pas le critère utilisé par l'ACM. Comme nous pouvons l'observer les dimensions selectionnées sont la dimension une mais aussi la dimension 30 ce qui montre que les variables ne sont pas triées par pertinence explicative et justifie alors la nécessité de passer par un sélection de variables.
Dans la suite, nous allons nous contenter d'utiliser les dimensions qui ont été choisies ici.

8 Analyse discriminante

L'analyse discriminante est une technique statistique qui a pour but d'expliquer et prédire l'appartenance d'un individu à des groupes prédéfinis en se basant sur une série de variables prédictives.
Dans ce qui suit, nous nous intéresserons dans un premier lieu à l'application d'une analyse discriminante sur les variables sélectionnées à partir de la méthode **STEPWISE**(cf: section7), on comparera par la suite les performances pour des méthodes de discrimination paramétriques et non paramétriques.
On effectuera dans un second temps, la même démarche mais sur l'ensemble des variables et on comparera les résultats obtenus.
Le but serait de trouver la méthode permettant de discriminer au mieux les classes d'individus.

8.1 Procédure CANDISC

CANDISC est une procédure utilisée pour effectuer une discrimination canonique. La discrimination canonique est une technique de réduction de dimension dans laquelle nous trouvons des combinaisons linéaires des variables quantitatives qui fournissent une séparation maximale entre les classes ou les groupes.
Une discrimination canonique s'effectue comme suit:

1. Calcul des moyennes des variables

2. Effectuer une analyse en composantes principales sur les moyennes calculées précedemment, en pondérant chaque moyenne par le nombre d'observations dans la classe. Les valeurs propres sont égales au rapport de la variation inter-classe à la variation intra-classe dans le sens de chaque composante principale.
3. Retransformer les composantes principales dans l'espace des variables d'origine, en obtenant les variables canoniques.

On commence par appliquer une discrimination canonique en 2 variables canoniques sur les variables sélectionnées précédemment par la méthode stepwise.

La procédure CANDISC

	Corrélation canonique	Corrélation canonique ajustée	Erreur type approchée	Corrélation canonique au carré	Valeurs propres de Inv(E)*H = CanRsq/(1-CanRsq)				Test de H0 : les corrélations canoniques de la ligne en cours et toutes celles qui suivent sont égales à zéro				
					Valeur propre	Différence	Proportion	Cumulé	Rapport de vraisemblance	Valeur de F approchée	DDL num.	DDL den.	Pr > F
1	0.346608	0.328884	0.040244	0.120137	0.1365	0.0977	0.7786	0.7786	0.84697190	6.80	12	942	<.0001
2	0.193345	0.176848	0.044029	0.037382	0.0388		0.2214	1.0000	0.96261772	3.67	5	472	0.0029

Figure 20: Résultats de la procédure CANDISC

On remarque que les deux valeurs de la corrélation canonique au carré sont très faibles, ceci démontre que la variance de la variable cible **(class)** est très peu expliquée par les coefficients de l'ACM.

8.2 Discrimination bayésienne

La discrimination bayésienne se base sur la règle de décision bayésienne qui consiste à produire une estimation de la probabilité a posteriori d'affectation d'un individu à un groupe. En d'autres termes, l'objectif d'une discrimination bayésienne est de produire une règle d'affectation qui permet de prédire, pour une observation donnée, sa valeur associée de Y(la variable cible) à partir des valeurs prises par X(tableau des données). En applique une dicrimination bayésienne sur les variables séléctionnées dans la section7, en utilisant deux méthodes:

- La première consiste à utiliser le meme ensemble pour entrainer et tester le modèle. Cette méthode n'est pas fiable étant donné que les résultats seront très biaisés et les performances élevées sans pour autant l'etre pour un autre ensemble de test différent.
- La deuxième méthode vient remédier à ce problème, c'est la méthode de la validation croisée qui consiste à diviser l'échantillon original en k échantillons, puis sélectionner un des k échantillons comme ensemble de test et les k-1 autres échantillons constitueront l'ensemble d'apprentissage.

Pour les deux méthodes, on obtient les résultats suivants:

On remarque que dans les deux cas, l'erreur est presque la meme et notre modèle se trompe en moyenne 1 fois sur 2 à chaque fois dans la classification des individus selon la variable class. On remarque aussi que la modalité **M** est la plus mal-classifiée parmi les trois modalités **H, L, M**, ce qui rejoint le résultat de l'ACM dans laquelle on avait du mal à interpréter la modalité **M**.
Dans l'optique d'améliorer les performances de nos modèles, on va utiliser des méthodes décisionnelles de discrimination(KNN,boule de rayon R).

La procédure DISCRIM
Synthèse de classification pour données de calibration : WORK.FICH_TOT
Synthèse de resubstitution utilisant Fonction discriminante linéaire

Nombre d'observations et pourcentage classifiés dans Class

De Class	H	L	M	Total
	18 37.50	20 41.67	10 20.83	48 100.00
H	72 50.70	36 25.35	34 23.94	142 100.00
L	20 15.75	76 59.84	31 24.41	127 100.00
M	62 29.52	72 34.29	76 36.19	210 100.00
Total	172 32.64	204 38.71	151 28.65	527 100.00
A priori	0.33333	0.33333	0.33333	

Estimations du compte des erreurs pour Class

	H	L	M	Total
Taux	0.4930	0.4016	0.6381	0.5109
A priori	0.3333	0.3333	0.3333	

La procédure DISCRIM
Synthèse de classification pour données de calibration : WORK.FICH_TOT
Synthèse de validation croisée utilisant Fonction discriminante linéaire

Nombre d'observations et pourcentage classifiés dans Class

De Class	H	L	M	Total
	18 37.50	20 41.67	10 20.83	48 100.00
H	69 48.59	36 25.35	37 26.06	142 100.00
L	21 16.54	74 58.27	32 25.20	127 100.00
M	65 30.95	74 35.24	71 33.81	210 100.00
Total	173 32.83	204 38.71	150 28.46	527 100.00
A priori	0.33333	0.33333	0.33333	

Estimations du compte des erreurs pour Class

	H	L	M	Total
Taux	0.5141	0.4173	0.6619	0.5311
A priori	0.3333	0.3333	0.3333	

Figure 21: Résultas de la discrimination bayésienne

8.3 Discrimination décisionnelle par K-plus proches voisins

L'algorithme des KNN est utilisé comme une méthode d'affectation d'un vecteur X comme suit:

1. Choix d'un entier K
2. Calcule de la distance d(x,x_i) où M est la métrique de Mahalanobis qui est la matrice inverse de la matrice de variance.
3. Retient des K observations pour lesquelles les distances sont les plus petites.
4. Compte du nombre de fois d'apparition de chacune de ces k observations dans une classe.

On teste cette méthode pour différentes valeurs de K est on a choisi celle qui minimise l'erreur. Le K optimal obtenu est **K_Optimal = 4** avec une erreur de 0.53.

La procédure DISCRIM
Synthèse de classification pour données de calibration : WORK.FICH_TOT
Synthèse de resubstitution utilisant 4 plus proches voisins

Nombre d'observations et pourcentage classifiés dans Class

De Class	H	L	M	Total
	14 29.17	15 31.25	19 39.58	48 100.00
H	81 57.04	24 16.90	37 26.06	142 100.00
L	10 7.87	90 70.87	27 21.26	127 100.00
M	43 20.48	47 22.38	120 57.14	210 100.00
Total	148 28.08	176 33.40	203 38.52	527 100.00
A priori	0.33333	0.33333	0.33333	

Estimations du compte des erreurs pour Class

	H	L	M	Total
Taux	0.4296	0.2913	0.4286	0.3832
A priori	0.3333	0.3333	0.3333	

La procédure DISCRIM
Synthèse de classification pour données de calibration : WORK.FICH_TOT
Synthèse de validation croisée utilisant 4 plus proches voisins

Nombre d'observations et pourcentage classifiés dans Class

De Class	H	L	M	Total
	18 37.50	15 31.25	15 31.25	48 100.00
H	63 44.37	31 21.83	48 33.80	142 100.00
L	22 17.32	62 48.82	43 33.86	127 100.00
M	59 28.10	56 26.67	95 45.24	210 100.00
Total	162 30.74	164 31.12	201 38.14	527 100.00
A priori	0.33333	0.33333	0.33333	

Estimations du compte des erreurs pour Class

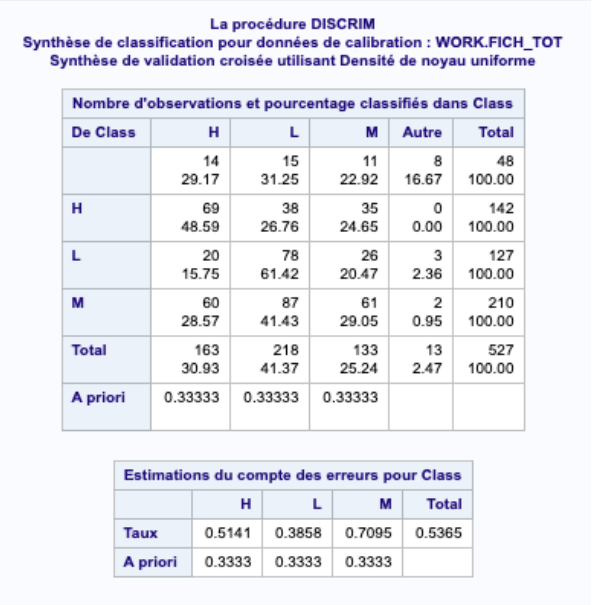
	H	L	M	Total
Taux	0.5563	0.5118	0.5476	0.5386
A priori	0.3333	0.3333	0.3333	

Figure 22: Résultas de la discrimination décisionnelle par un KNN

En examinant les résultats du KNN, on ne constate pas une amélioration dans les résulats. L'erreur reste toujours élevée et aux alentours des 50%.

8.4 Discrimination décisionnelle par boule de rayon R

Contrairement à une méthode de discrimination linéaire, le choix des paramètres de l'estimation d'une densité de noyau est plus cruciale que le choix du K pour un KNN. En effet, pour une descrimination décisionnelle par boule de rayon R, on est obligé de choisir le noyau de la boule ainsi que son rayon R. On a testé pour différentes valeurs de R. L'erreur est minimale pour une boule de rayon $\mathbf{R = 2}$ et de noyau **uniforme**.
On obtient les resultats suivants:



La procédure DISCRIM
Synthèse de classification pour données de calibration : WORK.FICH_TOT
Synthèse de validation croisée utilisant Densité de noyau uniforme

Nombre d'observations et pourcentage classifiés dans Class

De Class	H	L	M	Autre	Total
	14 29.17	15 31.25	11 22.92	8 16.67	48 100.00
H	69 48.59	38 26.76	35 24.65	0 0.00	142 100.00
L	20 15.75	78 61.42	26 20.47	3 2.36	127 100.00
M	60 28.57	87 41.43	61 29.05	2 0.95	210 100.00
Total	163 30.93	218 41.37	133 25.24	13 2.47	527 100.00
A priori	0.33333	0.33333	0.33333		

Estimations du compte des erreurs pour Class

	H	L	M	Total
Taux	0.5141	0.3858	0.7095	0.5365
A priori	0.3333	0.3333	0.3333	

Figure 23: Résultats de la discrimination decisionnelle par boule

8.5 Conclusion:

Les performances des trois méthodes de discrimination sont presque similaires avec une erreur de **50%**.
Ces résultas laissent supposer l'hypothèse que nous avons peut etre perdu une certaine variance importante dans les données en sélectionnant un sous-ensemble des dimensions. Pour vérifier celà, on refera la meme procédure de l'analyse discriminante mais sur l'ensemble complet des dimensions.
On obtient les résultats suivants:

La procédure DISCRIM
Résultats de classification pour données de calibration : WORK.FICH_TOT
Résultats de validation croisée utilisant Fonction discriminante linéaire

Nombre d'observations et probabilités a posteriori de la moyenne classifiées dans Class

De Class	H	L	M
	17 0.5702	16 0.5898	15 0.5261
H	72 0.5595	31 0.5345	39 0.4933
L	30 0.5291	58 0.5745	39 0.4929
M	75 0.5529	73 0.5716	62 0.5106
Total	194 0.5532	178 0.5677	155 0.5033
A priori	0.33333	0.33333	0.33333

Estimations du compte des erreurs de la probabilité a posteriori pour Class

Estimation	H	L	M	Total
Stratifié	0.3939	0.4222	0.5624	0.4595
Non stratifi	0.3891	0.4247	0.5559	0.4566
A priori	0.3333	0.3333	0.3333	

Figure 24: Discrimination bayesienne

La procédure DISCRIM
Synthèse de classification pour données de calibration : WORK.FICH_TOT
Synthèse de validation croisée utilisant 4 plus proches voisins

Nombre d'observations et pourcentage classifiés dans Class

De Class	H	L	M	Total
	18 37.50	23 47.92	7 14.58	48 100.00
H	58 40.85	33 23.24	51 35.92	142 100.00
L	31 24.41	55 43.31	41 32.28	127 100.00
M	74 35.24	67 31.90	69 32.86	210 100.00
Total	181 34.35	178 33.78	168 31.88	527 100.00
A priori	0.33333	0.33333	0.33333	

Estimations du compte des erreurs pour Class

	H	L	M	Total
Taux	0.5915	0.5669	0.6714	0.6100
A priori	0.3333	0.3333	0.3333	

Figure 25: Discrimination par KNN (K = 4)

La procédure DISCRIM
Synthèse de classification pour données de calibration : WORK.FICH_TOT
Synthèse de validation croisée utilisant Densité de noyau uniforme

Nombre d'observations et pourcentage classifiés dans Class

De Class	H	L	M	Autre	Total
	0 0.00	0 0.00	0 0.00	48 100.00	48 100.00
H	6 4.23	4 2.82	2 1.41	130 91.55	142 100.00
L	3 2.36	6 4.72	4 3.15	114 89.76	127 100.00
M	2 0.95	6 2.86	10 4.76	192 91.43	210 100.00
Total	11 2.09	16 3.04	16 3.04	484 91.84	527 100.00
A priori	0.33333	0.33333	0.33333		

Estimations du compte des erreurs pour Class

	H	L	M	Total
Taux	0.9577	0.9528	0.9524	0.9543
A priori	0.3333	0.3333	0.3333	

Figure 26: Discrimination par boule de R = 2

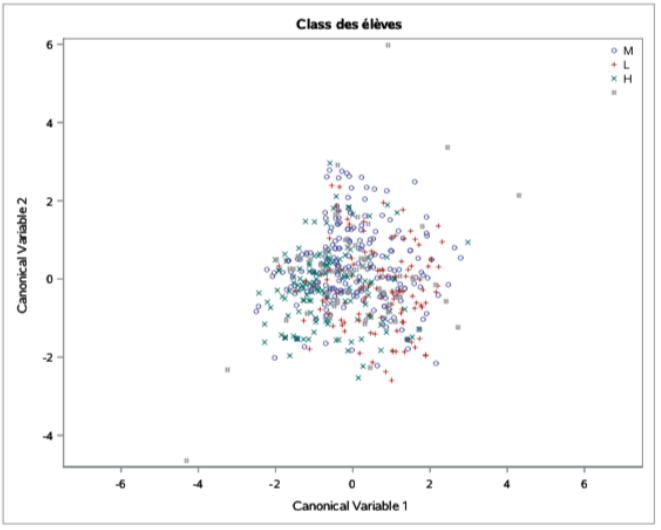


Figure 27: Graphe d'une discimination canonique

On remarque que les erreurs sont plus importantes pour une discrimination par KNN et par boule, sur l'ensemble des dimensions que sur une séléction de dimensions. Cependant une discrimination linéaire sur l'ensemble de dimensions donne de meilleures performances que sur une séléction de variables.
Pour le graphe de la disrimination canonique, on peut distinguer une séparation linéaire entre les individus de class **L** et les individus de class **H**. En effet, la composante canonique 1 permet de discriminer les individus de classe **L** des individus de la classe **H**. Cette discrimination ne ressortait pas dans le cas d'une séléction d'un sous-ensemble de variables.
Ainsi, une séléction d'un sous-ensemble de variables, a certes permis d'améliorer les performances des modèles, mais celà a supprimer la capacité de la composonte canonique 1 à discriminer entre les individus **H** et les individus **L**.
Une piste d'amélioration qui reste à explorer est la piste des classes sous-jacentes, et d'expliquer les classement non seulement par variables mais par cluster. Pour ceci, il fallait traiter chaque cluster de même que précèdent en le considérant comme un échantillon indépendant à condition que la distribution de la variables cibles soit différentes entre les cluster obtenus par la CAH.

9 Annexe

9.1 Annexe 1

Modalités des variables

Variable	Modalités
Gender	**F**: Femme
	H:Homme
Nationnality	**KW**: Kuwait
	JRD:Jordan
GradeID	**grades1**: G01-G03
	grades2: G04-G06
	grades3: G07-G09
	grades4: G10-G12
SectionID	**A**: salle de cours A
	B: salle de cours B
	C: salle de cours C
Topic	English
	Spanish, French, Arabic, IT, Math, Chemistry, Biology, Science, History, Quran, Geology
Semester	**F**: First
	S: Second
Relation	Father
	Mother
Raisedhands	**peu**
	moyen
	souvent
VisistedRessources	**peu**
	moyen
	souvent
AnnoucementsView	**peu**
	moyen
	souvent
Discussion	**peu**
	moyen
	souvent
ParentAnsweringSurvey	**Yes**: Ils ont répondu à l'enquete
	No: ils n'ont pas répondu à l'enquete
ParentSchoolSatisfaction	**Yes**: Satisfaits
	No: Insatisfaits
ParentSchoolSatisfaction	**Good**: Satisfaits
	Bad: Insatisfaits
StudentAbsenceDays	**Above-7**: plus que 7 absences par semestre
	Under-7: moins que 7 absences par semestre

9.2 Annexe 2

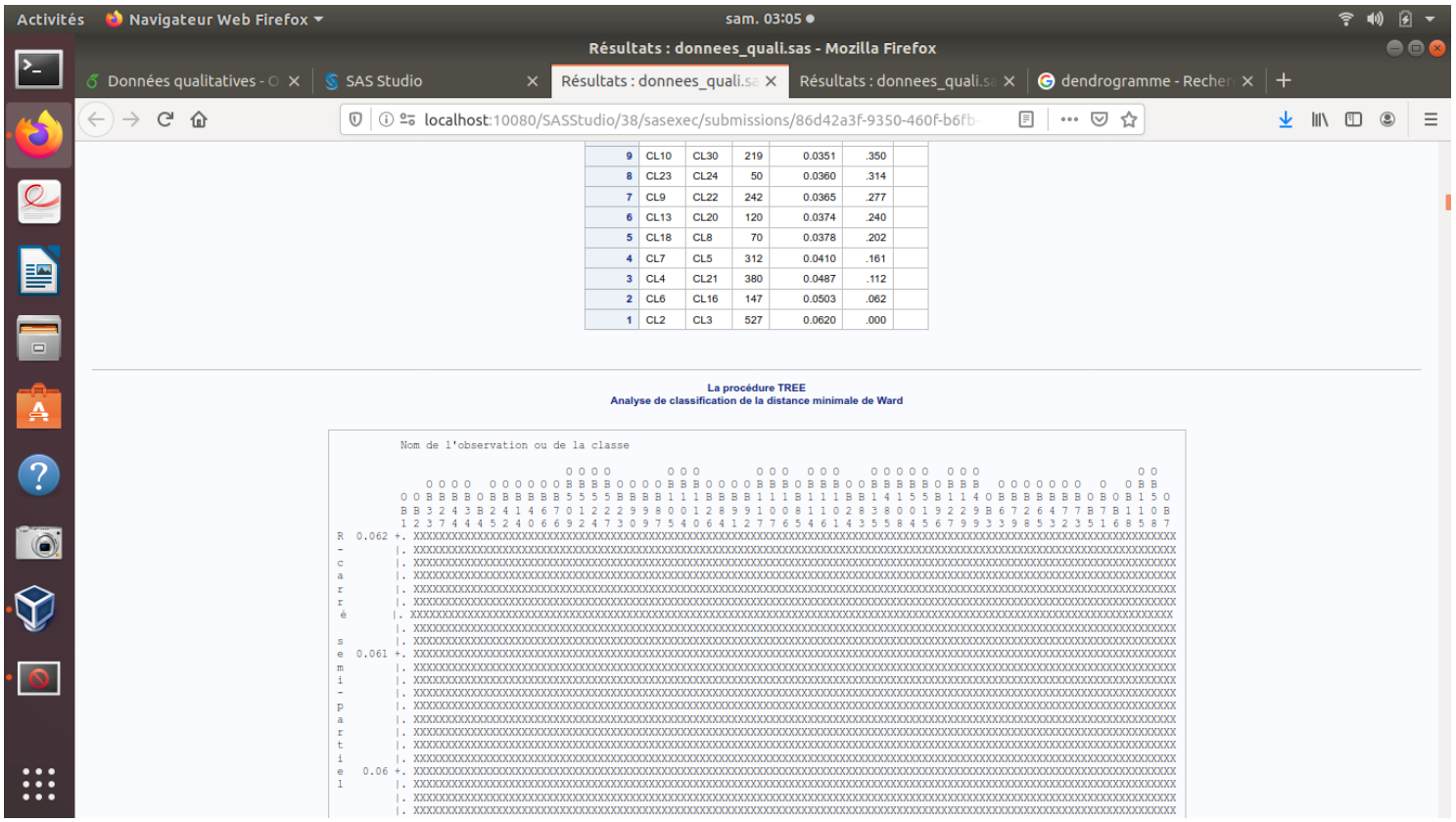


Figure 28: Affichage réduit à des croix