Master TRIED

Rapport : Données Qualitatives

Réalisé par :

Karim ASSAAD

Année universitaire :

2016/2017

Table des matières

Introduction

L'objectif de ce projet est de comprendre le comportement des variables qualitatives, leurs liaisons et les effets des unes par rapport aux autre tout en mettant en œuvres plusieurs méthodes vues en cours comme l'analyse a correspondance multiple pour la réduction de dimension, la méthode Stepwise pour la sélection des variables, l'analyse factorielle discriminantes pour la modélisation et le K-Moyennes pour le Clustering.

Ce rapport présente les résultats les plus pertinents obtenues à l'aide de l'outil SAS et expose des réflexions bien fondées par rapports à ces résultats.

Description des données

Le jeu de données traité dans ce projet est constitué de 913 individus qui sont des professeurs et 51 variables dont 8 sont des variables socio-professionnelles. Le reste représente des variables qualitatives reliées à des questionnaires.
Il s'agit d'une enquête mise en place par Wikipédia ou cette dernière a réalisé un questionnaire dédié aux professeurs de deux universités concernant l'utilisation de Wikipédia et son intégration dans le processus d'enseignement.
Les variables socio-professionnelles sont : AGE, GENDER, DOMAIN, PhD et YEARSEXP qui réfère aux années d'expérience acquis durant l'enseignement, UNIVERSITY qui présente l'université de chaque professeur et POSITION qui, comme son nom l'indique, renseigne sur la position des professeurs au sein de leurs universités. Toutes ces variables sont qualitatives sauf la variable AGE qui est quantitative.
Les variables proposées par le questionnaire sont reparties selon le sujet qu'elles traitent, elles prennent des modalités de 1 à 5, 5 veut dire que la personne est tout à fait d'accord et 1 qu'elle n'est pas du tout d'accord. Les autres nombres désignent les cas intermédiaires ou la personne peut ne pas être totalement d'accord sans refuser le principe de la question.
Ces variables sont :

- PU1 : L'utilisation de Wikipédia permet aux étudiants de développer de nouvelles compétences
- PU2 : L'utilisation de Wikipédia améliore l'apprentissage des élèves
- PU3 : Wikipédia est utile pour l'enseignement
- PEU1 : Wikipédia est convivial
- PEU2 : Il est facile de trouver dans Wikipédia les informations que vous recherchez
- PEU3 : Il est facile d'ajouter ou de modifier des informations dans Wikipédia
- ENJ1 : L'utilisation de Wikipédia stimule la curiosité
- ENJ2 : L'utilisation de Wikipédia est amusante
- QU1 : Les articles dans Wikipédia sont fiables
- QU2 : Les articles dans Wikipédia sont actualisés
- QU3 : Les articles dans Wikipédia sont complets
- QU4 : Dans mon domaine, Wikipédia a une qualité inférieure à celle des autres ressources éducatives
- QU5 : Je fais confiance au système d'édition de Wikipédia
- VIS1 : Wikipédia améliore la visibilité du travail des élèves
- VIS2 : Il est facile d'avoir un enregistrement des contributions faites dans Wikipédia
- VIS3 : Je cite Wikipédia dans mes articles académiques
- IM1 : L'utilisation de Wikipédia est bien considérée parmi les collègues
- IM2 : Dans le milieu universitaire, le partage des ressources éducatives ouvertes est apprécié
- IM3 : Mes collègues utilisent Wikipédia
- SA1 : C'est important de partager le contenu académique dans des plates-formes ouvertes
- SA2 : C'est important de publier des résultats de recherche dans d'autres médias que des revues
- SA3 : C'est important que les étudiants se familiarisent avec les environnements collaboratifs en ligne
- USE1 : J'utilise Wikipédia pour développer mon matériel pédagogique

- USE2 : J'utilise Wikipédia comme une plate-forme pour développer des activités éducatives
- USE3 : Je recommande mes étudiants d'utiliser Wikipédia
- USE4 : Je recommande mes collègues d'utiliser Wikipédia
- USE5 : Je suis d'accord que mes étudiants utilisent Wikipédia dans mes cours
- PF1 : Je contribue aux blogs
- PF2 : Je participe activement aux réseaux sociaux
- PF3 : Je publie des contenus académiques dans des plateformes ouvertes
- JR1 : Mon université favorise l'utilisation d'environnements collaboratifs ouverts dans Internet
- JR2 : Mon université considère l'utilisation d'environnements collaboratifs ouverts sur Internet comme un mérite pédagogique
- BI1 : À l'avenir, je recommanderai l'utilisation de Wikipédia à mes collègues et étudiants
- BI2 : À l'avenir, je vais utiliser Wikipédia dans mon activité d'enseignement
- INC1 : Pour concevoir des activités éducatives utilisant Wikipédia, il serait utile : un guide des meilleures pratiques
- INC2 : Pour concevoir des activités éducatives utilisant Wikipédia, il serait utile : obtenir des instructions d'un collègue
- INC3 : Pour concevoir des activités éducatives en utilisant Wikipédia, il serait utile : obtenir une formation spécifique
- INC4 : Pour concevoir des activités éducatives utilisant Wikipédia, il serait utile : une plus grande reconnaissance institutionnelle
- EXP1 : Je consulte Wikipédia pour les questions liées à mon domaine d'expertise
- EXP2 : Je consulte Wikipédia pour d'autres questions académiques connexes
- EXP3 : Je consulte Wikipédia pour des questions personnelles
- EXP4 : Je contribue à Wikipédia (éditions, révisions, amélioration d'articles ...)
- EXP5 : J'utilise des wikis pour travailler avec mes étudiants

USERWIKI (Wikipedia registered user) : est la variable à expliquer. C'est une variable qualitative qui a les modalités 0 et 1. La modalité 0 veut dire que l'individu ont un compte Wikipédia et la modalité 0 veut dire qu'il l'a pas.

Analyse des données

USERWIKI	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	784	86.25	784	86.25
1	125	13.75	909	100.00

Tableau 1 : les Fréquences de la variable USERWIKI

Ce tableau nous montre qu'il n'y a pas une équivalence entre la fréquence des professeurs qui sont inscrit sur Wikipédia et les professeurs qui ne le sont pas. On peut voir qu'il plus que 80% des professeurs qui ne sont pas inscrits, cela peut avoir des effets sur la modélisation de cette variable, néanmoins on peut considérer qu'une fréquence de 125 n'est pas très mauvaise et donc le cas reste tolérable.

Tri à plat des variables explicatives

Pour mieux comprendre les données et les visualiser d'une manière simple, j'ai procédé par analyser les tableaux de fréquences de toutes les variables explicatives.

Les résultats obtenus m'ont permis de déduire que les variables peu1, peu2, im3, sa1, sa2 et sa3 ont une valeur très faible pour leurs premières modalités par rapport aux autres modalités.

On remarque aussi que les valeurs de la dernière modalité pour les variables pf1, pf3, exp4 est négligeable par rapport aux valeurs obtenues pour le reste des modalités.

On a aussi constaté que la tranche d'âge des différents individus varie entre la trentaine et la cinquantaine donc une catégorie d'âge moyenne.

Fréquences des variables explicatives en fonction de la variable à expliquer

Puisqu'on traite un problème de modélisation, il faut bien comprendre la relation de cause à effet entre la variable à expliquer et les variables explicatives. Pour cela j'ai fait les tableaux de fréquence de variables en fonction de la variable à expliquer et dont la modalité 0 présente les professeurs non-inscrits sur Wikipédia et la modalité 1 les professeurs inscrits.

On s'intéresse ici à une partie du questionnaire qui traite le rôle de Wikipédia dans l'amélioration de l'enseignement et l'apprentissage des étudiants et qui est présentée par les variables pu1, pu2 et pu2.

Les variables pu1 et pu2 ont presque le même comportement, en effet les personnes qui ont un compte Wikipédia ont donné des avis entre favorables et moyen donc ils valorisent le rôle de Wikipédia dans le développement des techniques de l'éducation, d'un autre côté ces même personnes ont donné un avis moyen donc ne sont pas tout à fait d'accord par rapport à l'utilité de Wikipédia dans l'enseignement.

On va maintenant analyser les résultats obtenues pour les personnes qui n'ont pas de comptes Wikipédia, ces personnes sont d'accord pour le fait que Wikipédia aide à améliorer le niveau de l'enseignement en contrepartie, ils ne sont pas d'accord avec le fait que Wikipédia soit vraiment utile pour l'éducation.

On passe maintenant à analyser les variables qui traitent la convivialité de Wikipédia et la facilité de son utilisation pour l'utilisateur et qui est présenté par les questions peu1, peu2, peu3. Les réponses ont été similaires pour peu1 et peu2 qui ont donné un bon avis pour les modalité 0 et 1, quand a la variable peu3 la réponse est moyen pour les deux modalités 0 et 1.

On considère à présent Les variables enji1 et enji2 qui font référence au rôle de Wikipédia à éveiller la curiosité et au fait que son utilisation soit amusante, pour les deux modalités 0 et1 les individus sont d'accord pour le fait que Wikipédia amuse et rend plus curieux.

Concernant la qualité des articles dans Wikipédia et qui est présentée par les variables qu1, qu2, qu3 et qu4, les individus ayant un compte Wikipédia trouve que les articles sont de très bonne qualité et de source fiables, par contre les personnes qui ne possèdent pas de compte ne sont pas très d'accord pas cette affirmation.

Pour les hypothèses vis1, vis2 et vis3 les individus enregistrés sur Wikipédia sont d'accord et ceux qui ne sont pas enregistrés sont moyennement d'accord.

Pour les variables im1, im2 et im3 qui s'intéressent aux appréciation par rapport à Wikipédia, les professeurs de modalité 0 sont moyennement d'accord avec les hypothèses énoncées par les variables et les professeurs de modalités 1 sont très d'accord avec ces hypothèses.

En regardant les variables sa1, sa2 et sa3 on voit que les professeurs des deux modalités 0 et 1 sont d'accord avec les hypothèses présentées par ces variables.

Les professeurs de modalité 0 sont moyennement d'accord avec les hypothèses énoncées par les variables bi1, bi2, jr1 et jr2, par contre les professeurs de modalités 1 sont très d'accord avec les mêmes hypothèses.

Les professeurs de modalités 0 trouvent rejettent les hypothèses présentées par les variables use1, use2, use3, use4, use5, pf1, pf2 et pf3, en contrepartie les professeurs de modalités 1 sont d'accord avec ces hypothèses.

Enfin, les variables exp1, exp2, exp3, exp4, exp5, ink1, ink2, ink3 et ink4 ont des valeurs bien distribuées entre toutes les modalités.

Comme conclusion, on peut dire qu'il est tout à fait attendus que les professeurs inscrit sur Wikipédia et qui contribue à ses articles donne des avis favorables par rapport à cet outil, cependant les personnes qui n'ont pas de compte Wikipédia ont tendance à donner un avis moyen en général, d'un autre cote les personnes qui donnent un avis complètement défavorables sont très peu et cela revient au fait que la plus part des personnes utilisent Wikipédia et peu sont les personnes à ne pas la connaitre.

A première vue, on peut dire que les hypothèses posées par Wikipédia aux professeurs ne sont pas aussi pertinentes que ça et il fallait être plus précis et faire des hypothèses plus distinctives par rapport à la variable cible.

Nettoyage des données

Après la consultation des tableaux de fréquences j'ai trouvé judicieux d'éliminer quelques variables qui ont l'air de ne pas avoir un effet sur la variable à expliquer. Les choix sont basés surtout sur les résultats des tests de Cramer et de Chi-deux qui permettent de mettre en évidence la relation entre les variables qualitatives.

Tout d'abord j'ai commencé par l'élimination de tous les variables socio-professionnel puis j'ai enlevé toutes les variables reliées à la facilité perçue d'utilisation de Wikipédia (PEU1, PEU2, PEU3), toutes les variables relier au Plaisir perçu (ENJ1, ENJ2), toutes les variables reliées aux partages (SA1, SA2, SA3), toutes les variables reliées aux incitations (, inc1, inc2, inc3, inc4) et toutes les variables reliées aux Pertinence de l'emploi (JR1, JR2). De plus j'ai éliminé les variables suivantes : PU1, PU2, QU1, Qu3, Qu4, Im2, Im3, USE5, EXP3 ce qui est fait un total de 20 variables explicatives restantes.

Analyse a Correspondance Multiple

On procède maintenant à effectuer l'analyse a correspondances multiples afin de transformer les variables qualitatives en quantitatives, mieux comprendre les liaisons entre les variables et pour réduire les dimensions.

Cette méthode sera aussi utile pour adapter les données à l'analyse factorielle discriminante (AFD).

La représentation du tableau des inerties est la suivante :

Inertia and Chi-Square Decomposition

Singular Value	Principal Inertia	Chi-Square	Percent	Cumulative Percent
0.64693	0.41852	5959.8	13.29	13.29
0.47994	0.23034	3280.1	7.31	20.60
0.36863	0.13589	1935.0	4.31	24.91
0.34464	0.11878	1691.4	3.77	28.68
0.32914	0.10833	1542.7	3.44	32.12
0.29609	0.08767	1248.4	2.78	34.91
0.28437	0.08086	1151.5	2.57	37.47
0.27046	0.07315	1041.6	2.32	39.80
0.25717	0.06614	941.8	2.10	41.89
0.25365	0.06434	916.2	2.04	43.94
0.25240	0.06370	907.1	2.02	45.96
0.25032	0.06266	892.2	1.99	47.95
0.24450	0.05978	851.3	1.90	49.85
0.24171	0.05842	832.0	1.85	51.70
0.23920	0.05722	814.8	1.82	53.52
0.23586	0.05563	792.1	1.77	55.28
0.23211	0.05387	767.1	1.71	56.99
0.22869	0.05230	744.7	1.66	58.65
0.22612	0.05113	728.1	1.62	60.28
0.22405	0.05020	714.8	1.59	61.87
0.22074	0.04873	693.9	1.55	63.42
0.21735	0.04724	672.7	1.50	64.92
0.21576	0.04655	662.9	1.48	66.40
0.21214	0.04500	640.8	1.43	67.82
0.20896	0.04366	621.8	1.39	69.21
0.20815	0.04333	617.0	1.38	70.59
0.20616	0.04250	605.2	1.35	71.93
0.20340	0.04137	589.1	1.31	73.25
0.19926	0.03970	565.4	1.26	74.51
0.19785	0.03914	557.4	1.24	75.75
0.19493	0.03800	541.1	1.21	76.96
0.19418	0.03771	537.0	1.20	78.15
0.19093	0.03645	519.1	1.16	79.31

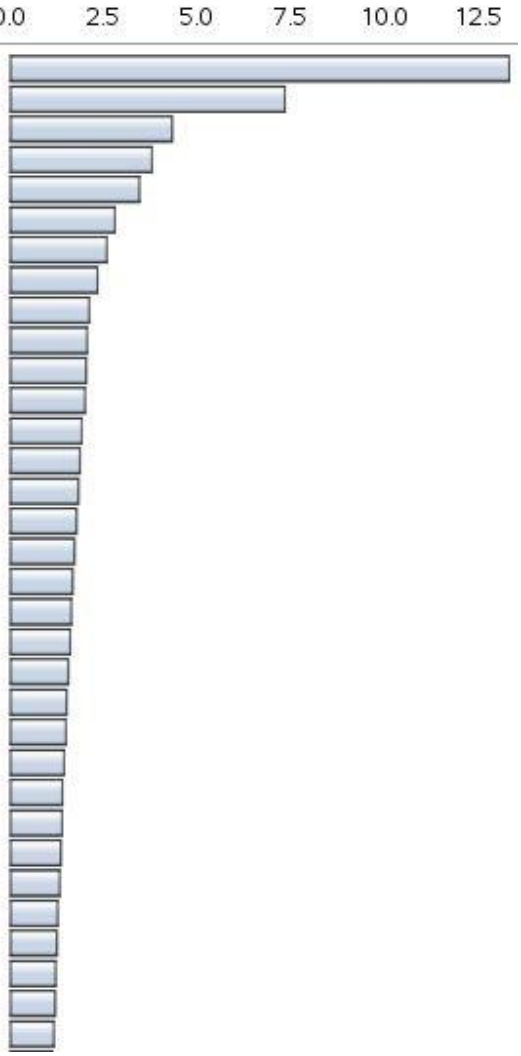


Tableau 2 : Représentation des inerties provenant de L'ACM

D'après le tableau on remarque qu'aucune valeur propre n'est plus grande que 1, pour ce fait il faut procéder par la méthode de coud pour le choix des dimensions. On peut voir clairement qu'entre la deuxième et la troisième valeur propre il y a une grande chute puis les valeurs commencent à dégrader légèrement, Cela m'amène à choisir d'afficher les résultats des 2 axes principaux, mais si on prend en considération le fait que même si ces deux axes présentent relativement plus d'informations que les autres axes, ils ne présentent en réalité que 20% de l'information ce qui signifie que si on traite seulement ces 2 axes on aura environ 80% de l'information perdue.

Ci-dessous les projections des données sur les deux premiers axes.

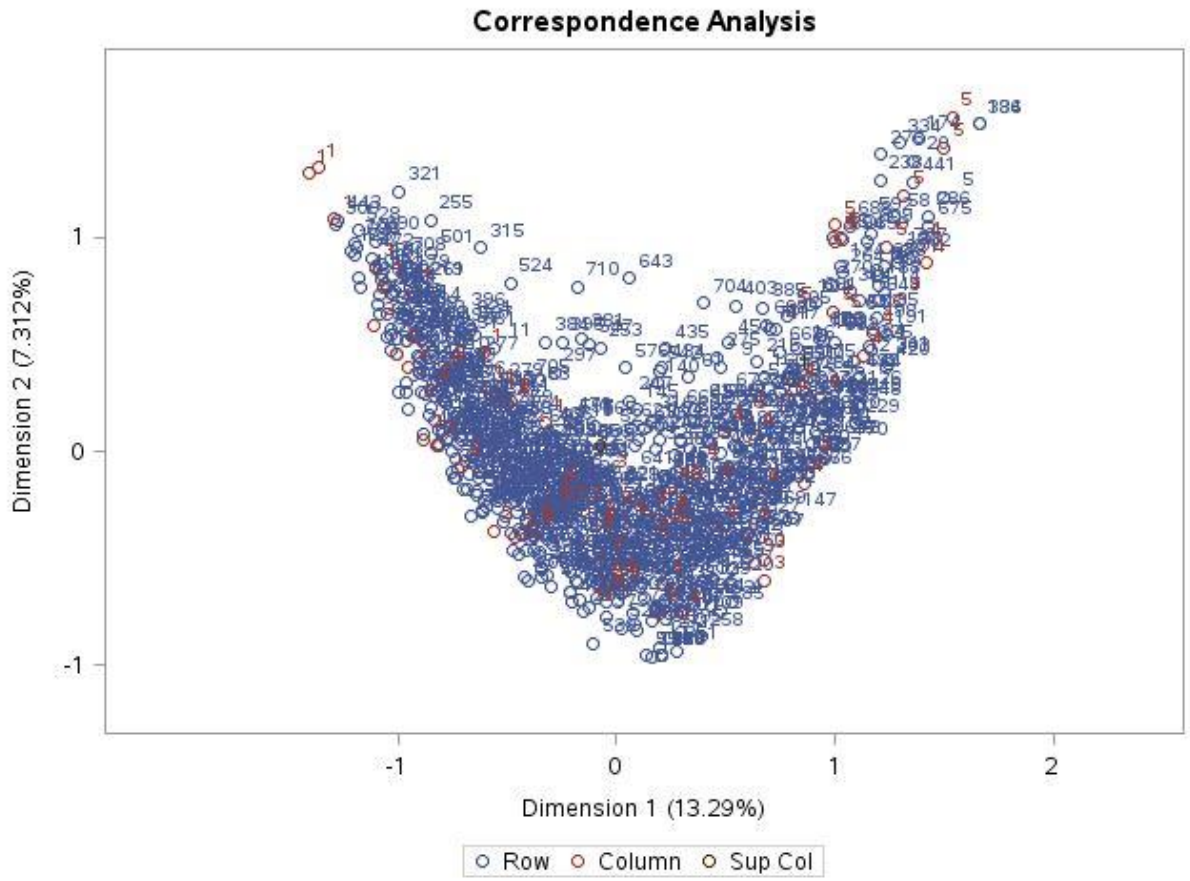


Tableau 3: Nuage des points sur le premier plan factorielle

On voit que sur le premier plan principale la projection des données a une allure polynomiale.

La variable supplémentaire n'a pas un effet évidant ici, on peut voir que les valeurs rouge et bleu sont chevauchées donc on présume qu'il est dur de distinguer entre elles.

C'est le cas aussi pour tous les plans factoriels à l'exception de quelques-uns ou on remarque que les professeurs qui sont inscrit sur Wikipédia sont condensés dans le centre du nuage. Cette observation nous amène à penser que la méthode de Discrimination Décisionnelle à Boule peut correspondre à ce problème.

Analyse Factorielle Discriminante

USERWIKI = 0					
Variable	N	Sum	Mean	Variance	Standard Deviation
Dim1	611	-3.14974	-0.00516	0.41151	0.6415
Dim2	611	-3.82679	-0.00626	0.22887	0.4784
Dim3	611	-3.63685	-0.00595	0.12963	0.3600
Dim4	611	-1.67526	-0.00274	0.12050	0.3471
Dim5	611	-1.94022	-0.00318	0.11163	0.3341
Dim6	611	-2.36291	-0.00387	0.08896	0.2983
Dim7	611	-2.86160	-0.00468	0.07853	0.2802
Dim8	611	5.76696	0.00944	0.07122	0.2669
Dim9	611	-1.42037	-0.00232	0.06664	0.2582
Dim10	611	-2.11518	-0.00346	0.06424	0.2535
Dim11	611	-3.99736	-0.00654	0.06283	0.2506
Dim12	611	0.86098	0.00141	0.06189	0.2488
Dim13	611	4.60644	0.00754	0.05675	0.2382
Dim14	611	-0.10450	-0.0001710	0.05683	0.2384
Dim15	611	-1.58768	-0.00260	0.05834	0.2415
Dim16	611	1.83429	0.00300	0.05639	0.2375
Dim17	611	-3.66888	-0.00600	0.05449	0.2334
Dim18	611	-0.55394	-0.0009066	0.05278	0.2297
Dim19	611	-1.60288	-0.00262	0.05054	0.2248
Dim20	611	-2.52925	-0.00414	0.05009	0.2238
Dim21	611	-0.88525	-0.00145	0.04763	0.2183
Dim22	611	1.02805	0.00168	0.04719	0.2172
Dim23	611	-3.14272	-0.00514	0.04665	0.2160
Dim24	611	-1.18906	-0.00195	0.04641	0.2154
Dim25	611	-0.93268	-0.00153	0.04079	0.2020

Description des dimensions pour la modalité 0 de USERWIKI

USERWIKI = 1					
Variable	N	Sum	Mean	Variance	Standard Deviation
Dim1	99	5.01004	0.05061	0.45762	0.6765
Dim2	99	3.19767	0.03230	0.24058	0.4905
Dim3	99	2.79090	0.02819	0.16715	0.4088
Dim4	99	1.97089	0.01991	0.11200	0.3347
Dim5	99	1.40077	0.01415	0.09029	0.3005
Dim6	99	2.15055	0.02172	0.08224	0.2868
Dim7	99	2.48436	0.02509	0.09670	0.3110
Dim8	99	-6.01773	-0.06079	0.08349	0.2890
Dim9	99	1.40222	0.01416	0.06524	0.2554
Dim10	99	1.74515	0.01763	0.06638	0.2577
Dim11	99	3.37280	0.03407	0.06818	0.2611
Dim12	99	-1.01233	-0.01023	0.06853	0.2618
Dim13	99	-4.67317	-0.04720	0.07822	0.2797
Dim14	99	0.15712	0.00159	0.07061	0.2657
Dim15	99	1.22372	0.01236	0.05167	0.2273
Dim16	99	-2.07441	-0.02095	0.05223	0.2285
Dim17	99	3.32204	0.03356	0.05008	0.2238
Dim18	99	0.42224	0.00427	0.05124	0.2264
Dim19	99	1.44148	0.01456	0.05598	0.2366
Dim20	99	2.28233	0.02305	0.05163	0.2272
Dim21	99	0.74661	0.00754	0.05734	0.2395
Dim22	99	-0.88593	-0.00895	0.04864	0.2206
Dim23	99	3.10370	0.03135	0.04634	0.2153
Dim24	99	1.28038	0.01293	0.03772	0.1942
Dim25	99	0.74227	0.00750	0.06303	0.2511

Description des dimensions pour la modalité 1 de USERWIKI

Les tableaux ci-dessus présentent les modalités 0 et 1 de la variable à expliquer en fonction des variables prévenantes de l'analyse a correspondance multiple, en effet la modalité 0 présente les professeurs qui ne sont pas inscrits dans Wikipédia et la modalité 1 fait référence aux professeurs inscrits dans Wikipédia, on s'intéresse à la comparaison des différentes valeurs obtenue pour chaque modalité. Pour le tableau à droite on remarque que les valeurs des moyennes sont négatives sauf pour quelques variables, en effet les moyennes s'approchent de 0-, pour les moyennes obtenues dans le tableau à gauche c'est tout à fait le contraire les moyennes sont de valeur positives mais aussi situes aux alentours de 0 +.

On peut voir que la variance est assez grande pour chacune des modalités, en effet la plus grande variance pour la modalité 0 est égale à 0.441 et sa plus petite variance est de l'ordre de 0.040 ainsi que pour la modalité 1 dont la plus importante valeur consiste la 1ere variable et est égale a 0.475 et la plus petite valeur est 0.063.

Toutes ces différences seront bénéfiques pour la modélisation car on pourra mieux distinguer entre les deux modalités, en effet les variations des moyennes entre les valeurs négatives pour la modalité 0 et les valeurs positives pour la modalité 1 nous permettra de visualiser chaque modalité sans que celles-ci soient chevauchées entre elles.

On va par la suite réaliser les trois modèles suivants :

Discrimination Bayésienne

Error Count Estimates for USERWIKI			
	0	1	Total
Rate	0.3044	0.2828	0.2936
Priors	0.5000	0.5000	

Méthode de Substitution

Error Count Estimates for USERWIKI			
	0	1	Total
Rate	0.3502	0.5556	0.4529
Priors	0.5000	0.5000	

Méthode de validation croisse

En utilisant la fonction bayésienne, on génère les erreurs en employant en premier temps la méthode de substitution qui nous donne une erreur de 30 % pour prédire la modalité 0 et 28% pour la modalité 1, ce qui fait en total un taux de 29% mais si on compare cette erreur à celle obtenue avec la méthode de validation croisée et qui est au alentours de 45% on peut en déduire que le modèle reste assez robuste et n'arrive pas à généraliser

Discrimination Décisionnelle (K Plus Proche voisin avec K=5)

Error Count Estimates for USERWIKI			
	0	1	Total
Rate	0.4223	0.0000	0.2111
Priors	0.5000	0.5000	

Méthode de Substitution

Error Count Estimates for USERWIKI			
	0	1	Total
Rate	0.5090	0.4848	0.4969
Priors	0.5000	0.5000	

Méthode de validation croisse

Dans ce cas aussi et en employant le modèle du k plus proche voisin en fixant à k=5 on remarque que l'erreur obtenue avec la méthode de substitution est la moitié de celle obtenue avec la méthode de validation crosse, en effet on a un taux d'erreur de 21% pour la première méthode contre un taux de 49% pour la deuxième, ce qui nous mène à conclure que ce modèle aussi est incapable de faire une bonne généralisation des donnes.

Discrimination Décisionnelle (BOULE DE RAYON R)

Error Count Estimates for USERWIKI			
	0	1	Total
Rate	0.0000	0.0000	0.0000
Priors	0.5000	0.5000	

Méthode de Substitution

Error Count Estimates for USERWIKI			
	0	1	Total
Rate	0.9885	1.0000	0.9943
Priors	0.5000	0.5000	

Méthode de validation croisse

Dans cette section on utilise le modèle de Discrimination Décisionnelle avec une boule, dans le tableau à droite pour la méthode de substitution on a un taux d'erreur nul contre un taux d'erreur très élevé au alentour de 99% pour la méthode de validation croisée, donc notre modèle a fait un sur apprentissage dans le premier cas et il est incapable de donner une erreur de généralisation donc ce modèle est à rejeter. Cela rejette l'hypothèse qu'on a proposée tout à l'heure concernant l'adéquation des données avec ce modèle, cela vient du fait que très peu de dimension suivent l'allure qu'on a décrit précédemment (les professeurs dont la modalité est 1 de USERWIKI sont condensés dans le centre du nuage).

D'après ces données on arrive à la conclusion que le modèle bayésien est le meilleur car il assure la plus petite marge entre l'erreur selon la méthode de validation croisse et celle de substitution donc on le retient mais il est d'une performance moyenne car même ce taux d'erreur (aux alentours de 30-45) reste assez élevée, le modèle 2 qui est base sur k plus proche voisin est aussi à retenir même si la marge est plus large entre les erreurs générées par les 2 méthodes par contre on ne peut retenir

le 3eme modèle qui repose sur la Discrimination Décisionnelle avec une boule car il surapprend les données et donne une erreur très élevé presque égale a 100% ce qui est très mauvais pour un modèle de prédiction.

Selection de Variables

Pour avoir une meilleure performance on a procédé à une sélection de variables par le biais des méthodes Stepwise. La sélection a permis de choisir les variables suivantes :

Dim36, Dim28, Dim8, Dim39, Dim40, Dim13, Dim52, Dim41, Dim17, Dim48, Dim32, Dim29, Dim17, Dim23, Dim61, Dim11, Dim30.

Ci-dessous le summary de ces variables :

Stepwise Selection Summary

Step	Number In	Entered	Removed	Partial R-Square	F Value	Pr > F	Wilks' Lambda	Pr < Lambda	Average Squared Canonical Correlation	Pr > ASCC
1	1	Dim36		0.0083	5.94	0.0151	0.99168356	0.0151	0.00831644	0.0151
2	2	Dim28		0.0082	5.84	0.0159	0.98356148	0.0029	0.01643852	0.0029
3	3	Dim8		0.0082	5.85	0.0159	0.97548253	0.0005	0.02451747	0.0005
4	4	Dim39		0.0080	5.66	0.0177	0.96771980	0.0001	0.03228020	0.0001
5	5	Dim40		0.0079	5.63	0.0180	0.96004604	<.0001	0.03995396	<.0001
6	6	Dim13		0.0063	4.42	0.0358	0.95404099	<.0001	0.04595901	<.0001
7	7	Dim52		0.0056	3.95	0.0472	0.94869844	<.0001	0.05130156	<.0001
8	8	Dim41		0.0045	3.17	0.0752	0.94442201	<.0001	0.05557799	<.0001
9	9	Dim27		0.0038	2.69	0.1013	0.94080293	<.0001	0.05919707	<.0001
10	10	Dim48		0.0038	2.68	0.1024	0.93721607	<.0001	0.06278393	<.0001
11	11	Dim32		0.0037	2.57	0.1091	0.93377262	<.0001	0.06622738	<.0001
12	12	Dim29		0.0037	2.58	0.1086	0.93032805	<.0001	0.06967195	<.0001
13	13	Dim17		0.0037	2.58	0.1085	0.92688862	<.0001	0.07311138	<.0001
14	14	Dim23		0.0037	2.56	0.1098	0.92348206	<.0001	0.07651794	<.0001
15	15	Dim61		0.0036	2.52	0.1130	0.92014425	<.0001	0.07985575	<.0001
16	16	Dim11		0.0033	2.32	0.1283	0.91707548	<.0001	0.08292452	<.0001
17	17	Dim30		0.0031	2.17	0.1410	0.91420592	<.0001	0.08579408	<.0001

Maintenant on va refaire l'AFD en utilisant les trois méthodes et en prenant en compte seulement les variables sélectionnées précédemment.

Une fois cela fait, on remarque que les erreurs de la méthode de Discrimination Décisionnelle (K Plus Proche Voisin et Boule de Rayon R) ont restées à peu près les mêmes, mais il y a eu un changement pour le cas de la méthode bayésienne.

Le résultat obtenue pour la méthode bayésienne est présenté ci-dessous.

Error Count Estimates for USERWIKI

	0	1	Total
Rate	0.3404	0.3636	0.3520
Priors	0.5000	0.5000	

Méthode de Substitution

Error Count Estimates for USERWIKI

	0	1	Total
Rate	0.3453	0.3838	0.3646
Priors	0.5000	0.5000	

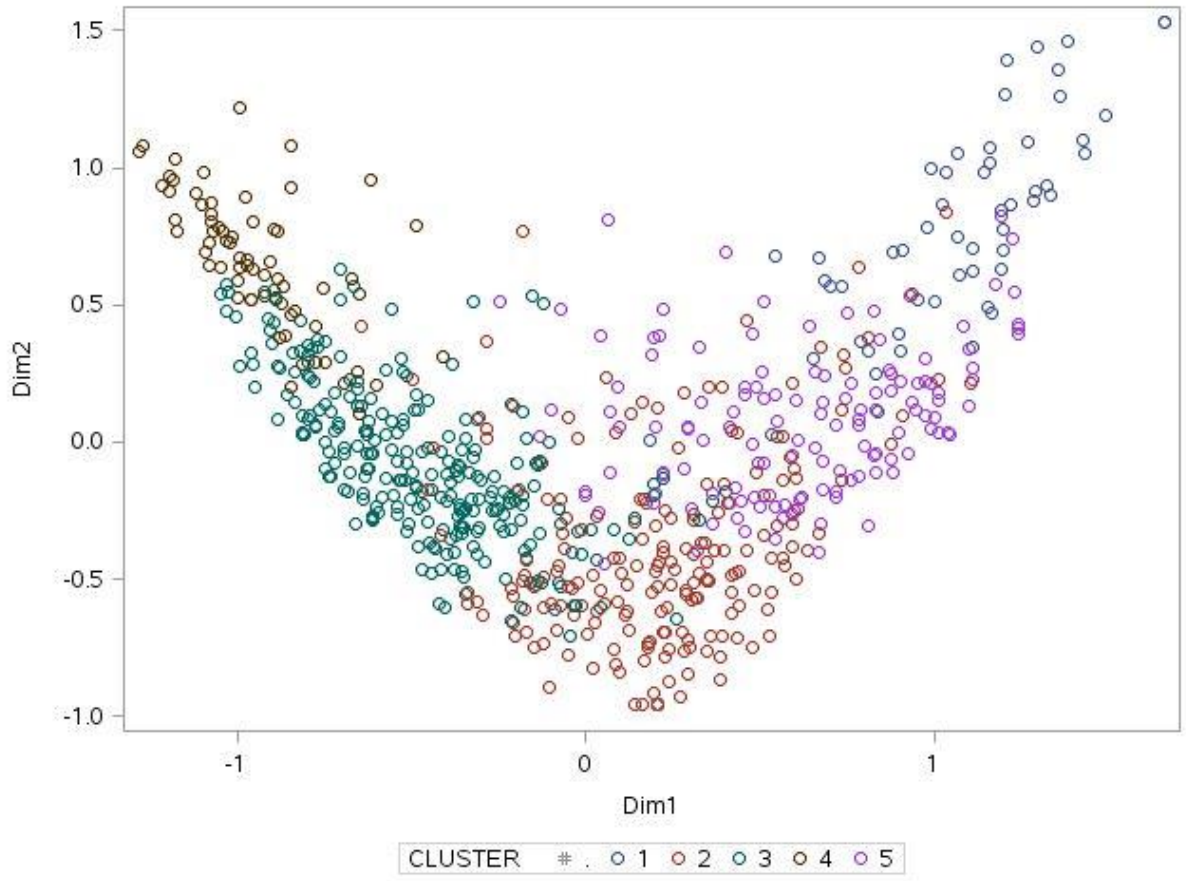
Méthode de validation croisée

On peut voir que l'erreur donnée par la méthode de substitution a augmenté par contre celle de la validation croisée a diminué. En général on peut dire que l'erreur est devenue plus stable et que ce modèle est de loin le meilleur parmi les autres.

Clustering

Les données ne sont pas linéairement séparables par rapport à la variable cible sur le premier plan factoriel donc ça ne sert à rien de découper l'échantillon en deux clusters dans l'espoir d'obtenir deux classes une modalité unique chacune, cependant il est intéressant de les deviser en sous échantillons. Ces sous échantillons regroupent des individus ayant des similarités ce qui va favoriser la séparation linéaire des données par rapport à la variable cible et par la suite va améliorer la modélisation.

Voici le résultat obtenu en découpant l'échantillons en 5 sous échantillons en utilisant la méthode des K-Moyennes :



On deduit de la figure que la classe de couleur rouge a la meilleure variance intra classe et que le decoupage semble donner une structure orodnnee sauf pour quelques points au centre qui sont distantes par rapport a toutes les classes et peuvent etres mises dans une classe a part.

Conclusion

Les méthodes employées Durant ce projet sont indispensables pour l'analyses des données qualitatives et l'apprentissage automatique, elles sont complémentaires les unes aux autres tel le cas de l'acm qui nous a fourni des nouvelles variables qui ont étés utilisées par la suite pour la modélisation avec la méthode d'analyse factorielle discriminante, de même Stepwise a permis d'améliorer la performance du modèle bayésien.

Le projet peut être développé encore plus afin d'obtenir des meilleurs résultats pour la modélisation et cela en effectuant un modèle pour chaque ensemble retournée par le Clustering.