Report V1

Congo Job

May 27, 2022

Contents

1 Introduction

La pratique rgulire dactivits physiques comporte de nombreux bnfices tel qu' une amlioration de la sant mentale ,la prvention des maladies cardiovasculaire , limiter la prise de poids et bien d'autres .

Cependant, on observe un dclin de la pratique dactivits physiques. il apparat primordia dencourager les jeunes maintenir une activit physique ou devenir plus actif , c'est dans cette optique que s'inscrit ce projet .

1.1 Sport et sciences sociales

Cre en 1979 par Bernard Michon Strasbourg ,lunit de recherche Sport et sciences sociales demeure la seule unit de recherche STAPS du Grand Est et est reconnue comme une structure de recherche incontournable en sciences sociales du sport dans le paysage franais et europen.

Regroupant plus de 20 chercheurs (titulaires et associs) et 17 doctorants, elle ralise des ouvrages et des articles de rfrence (plus de 174 publications).

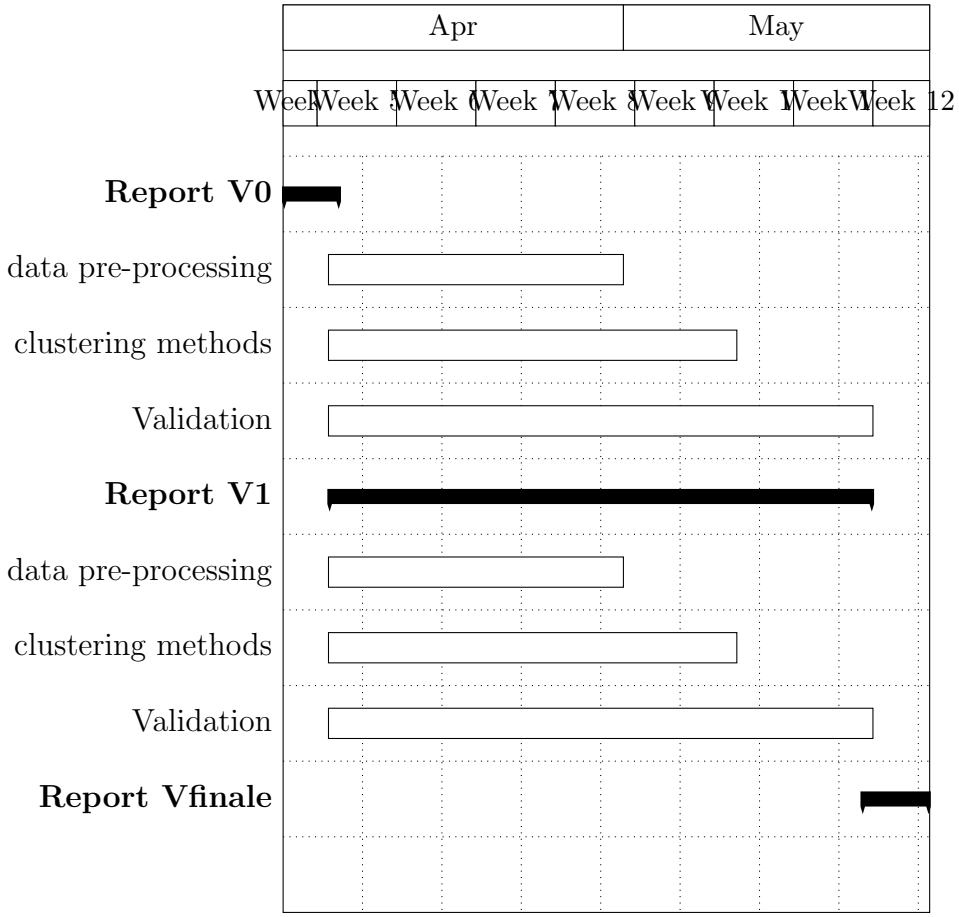
1.2 Main Objectives

L'objectif de ce travail est d'identifier des profils de praticiens bass sur des qualificatifs positifs ou ngatifs, c'est--dire d'attribuer un profil chaque cluster des donnes et d'estimer la force de ces profils, c'est--dire le nombre de clusters ou de profils qui sont les plus reprsentatifs des donnes.

1.3 Specific Objectives

Nous allons d'abord effectuer un prtraitement des donnes par la renormalisation des donnes , la suppression des valeurs aberrantes, complter ou supprimer les valeurs manquantes, puis pour analyser les donnes nous utiliserons diffrents algorithmes tels que : K-means, analyse en composantes principales, arbres de dcision. Enfin, nous testerons la solidit de notre cluster en utilisant des algorithmes de classification tels que : la rgression logistique, les k- plus proches voisins,...

Le diagramme de Gantt ci-dessous nous donne un aperu rapide de l'organisation du travail dans le temps .



2 Description of raw data

Le dataset contient des informations personnels sur les lycens (1070 participants)tel que leur initial, le Lyce, le sexe , le choix d'tude ,le travail des parents et le support des parents ainsi que la date de naissance , la morphologie de la personne (la taille et le poids) . Une vingtaine

des variables mesurent la nature de la motivation par exemple la jouissance , l'affiliation , la condition physique et le dgr de motivation tel que SIMS intrinsic et SIMS external regulation .

Enfin le reste des variables (71) recolt de la manire suivante : On pose une question : "En EPS, quel est le sport que vous avez le plus apprci ?".

Puis on lui indique : "Nous allons maintenant te prsenter des mots qui vont te permettre de dcrire ton ressenti par rapport ce sport. Votre travail consiste indiquer, le plus rapidement possible, si vous tes d'accord ou non avec ces propositions en cliquant sur oui ou non. Le temps de rponse a t pris en compte dans chaque rponse. Si ce temps est court, cela signifie que le terme semble vident. Par exemple, si le sport est "le football", l'lve pourra rpondre "oui" rapidement au qualificatif "plaisir", "non" rapidement au qualificatif "beaut".

Les rponses possibles chaque question sont "oui", "non", "je ne sais pas". Lorsque la rponse une question est "oui", la valeur du temps est positive, ngative dans le cas de "non" et zro dans le cas de "je ne sais pas".

	AP	AQ	AR	AS	AT	AU	AV	AW
1	Qualite	Force	Maintien	Puissance	Competition	Muscle	Beaute	Galbant
2	2 269	2 637	1 271	1 330	1 297	1 087	1 376	6 982
3	1 621	1 135	1 426	1 444	1 134	1 329	1 394	1 329
4	2 156	2 627	2 674	3 858	2 886	2 676	2 869	9 192
5	1 083	1 316	1 134	1 199	1 640	1 531	2 084	1 916
6	1 232	2 660	2 130	1 517	1 297	1 577	1 633	3 339
7	1 176	1 337	1 329	1 073	970	2 011	1 003	990
8	1 548	1 492	1 540	1 377	1 062	817	1 007	21 098
9	1 784	954	2 818	1 200	2 575	2 405	1 476	2 951
10	846	1 910	1 897	1 459	917	2 065	924	1 394
11	114	97	48	153	26	178	35	127
12	815	681	1 018	648	606	433	210	195
13	1 995	1 523	1 442	811	1 183	1 070	4 232	8 713
14	4 630	1 188	1 298	973	939	805	1 054	1 362
15	1 005	876	2 027	3 372	3 534	1 464	968	6 096
16	13 230	-24 103	28 141	0	0	0	0	53 420
17	7 633	-9 876	12 933	14 221	15 321	-19 314	22 099	-26 416
18	6 241	7 827	9 013	10 382	0	0	16 164	17 232
19	7 378	9 059	10 652	0	14 980	16 124	0	20 313
20	0	0	0	0	0	0	0	0
21	-8 010	9 132	0	0	0	0	17 893	0
22	0	0	0	0	0	0	0	0
23	7 700	8 453	9 783	10 686	11 565	12 252	14 208	-20 362
24	1 236	0	-2 151	0	1 281	-2 718	0	0
25	1 951	1 350	1 651	1 068	1 351	1 800	0	0

Figure 1: Raw data

3 Preprocessing

3.1 Preprocessing Method

Les valeurs manquantes sont mis zros en supposant que le qualificatif n'intresse pas les tudiants concerns et qu'ils auraient pu rpondre : "je ne sais pas". Pour la gestion des valeurs aberrantes , celles qui sont suprieur 5*cart-type ont t mise zros en fin de ne pas impacter le poids donne chaque mot. L'cart-type est calcul en utilisant les donnes non signes dans le but de rduire les valeurs extrme et viter la possible compensation des valeurs. Les tudiants ayant rpondu "je ne sais pas" toutes ces questions n'ont pas t considr dans la suite du projet . La normalisation est faite par ligne dans le but de conserver ce qui est "important" pour chaque personne.

3.2 Results of processed data

Qualite	Force	Maintien	Puissanc	mpetiti	Muscle	Beaute	Galbant
0,07968	0,09833	0,02909	0,03208	0,03041	0,01977	0,03441	0,31855
0,37034	0,25278	0,32317	0,32753	0,25254	0,29971	0,31543	0,29971
0,06264	0,08757	0,09006	0,15275	0,10129	0,09017	0,10039	0,43517
0,15351	0,21693	0,16739	0,18508	0,30512	0,27545	0,42597	0,38024
0,04507	0,15859	0,11645	0,06773	0,05024	0,0725	0,07695	0,21256
0,04796	0,06543	0,06456	0,03678	0,02561	0,13856	0,02919	0,02778
0,04451	0,04185	0,04413	0,0364	0,02147	0,00986	0,01886	0,97113
0,20936	0,04816	0,41018	0,09594	0,36298	0,32997	0,14954	0,43601
0,171	0,44751	0,44413	0,3303	0,18945	0,48779	0,19127	0,31341
0,01237	0,01039	0,00467	0,01692	0,0021	0,01984	0,00315	0,01389
0,07193	0,05717	0,09429	0,05354	0,04891	0,02985	0,00529	0,00364
0,1346	0,0999	0,09395	0,04756	0,07491	0,0666	0,29905	0,62846
0,83888	0,09546	0,11922	0,04903	0,04168	0,01274	0,06652	0,13305
0,01576	0,01031	0,05894	0,11577	0,12261	0,03515	0,0142	0,23086
0,05954	-0,10847	0,12664	0	0	0	0	0
0,03856	-0,0499	0,06534	0,07185	0,07741	-0,09758	0,11165	-0,13346
0,08063	0,10111	0,11644	0,13412	0	0	0,20882	0,22262
0,06209	0,07624	0,08965	0	0,12607	0,1357	0	0,17095
0	0	0	0	0	0	0	0
-0,03752	0,04278	0	0	0	0	0,08382	0
0	0	0	0	0	0	0	0
0,07073	0,07764	0,08986	0,09816	0,10623	0,11254	0,13051	-0,18703
0,29527	0	-0,51386	0	0,30602	-0,64931	0	0
0,70484	0,48772	0,59646	0,38584	0,48808	0,65029	0	0

Figure 2: Data processed

Les mots les plus "importants" et les moins "importants" pour chaque personne sont respectivement proche de soit 1 ou de -1 . Ceux qui sont moins "importants" sont proches de zero.

Le jeu de donnes nettoiye contient 1050 lycens et 71 caractristiques.

4 Clustering

4.1 Principal component analysis

L'analyse en composantes principales (ACP) est utilis pour rduire la dimension des donnes en quelques variables et garder les donnes les plus importants. C'est une mthode de la statistique multivarie, qui consiste transformer des variables lies entre elles (dites corrles en statistique) en nouvelles variables decorrles les unes des autres.

Pour dterminer le nombre de composante optimale , la comande fviz_eig de Rstudio, a t solicit. Cette fonction permet d'avoir le graphique des valeurs propres.

Les valeurs propres mesurent la quantit de variance explique par chaque axe principal. Elles sont grandes pour les premiers axes et petites pour les axes suivants.

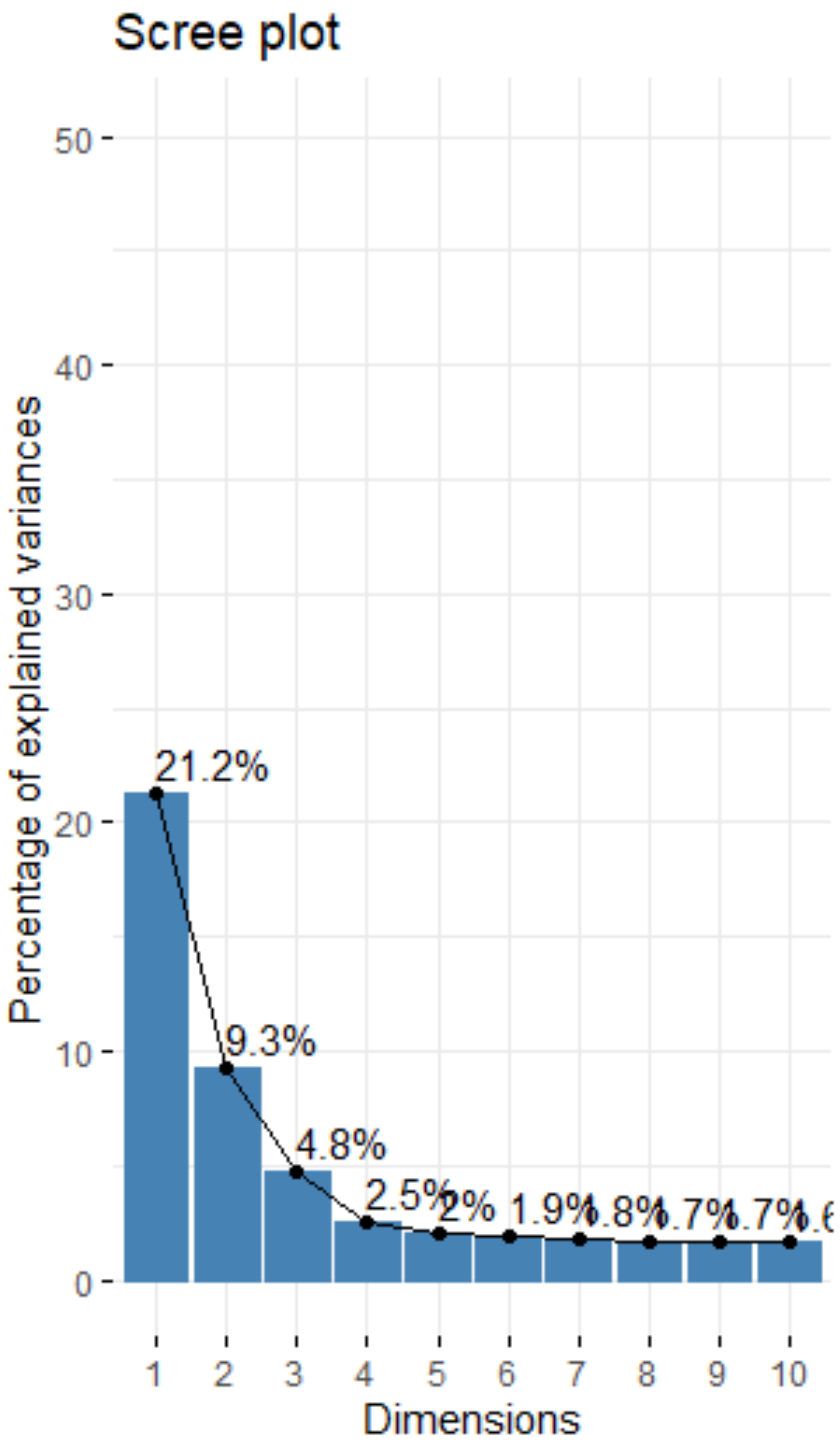


Figure 3:

D'aprs le graphique ci-dessus, nous pourrions vouloir nous arrter la cinquime composante principale car la variation est moindre aprs la cinquime. Cependant 39.79760 % des informations (variances) contenues dans les donnes sont retenues par les 5 premires composantes principales.

Les graphique ci-dessous montre le top 30 des variables contribuant le plus aux 5 composantes principales. Les lignes en pointill rouge, sur les graphiques, indique la valeur contribution moyenne.

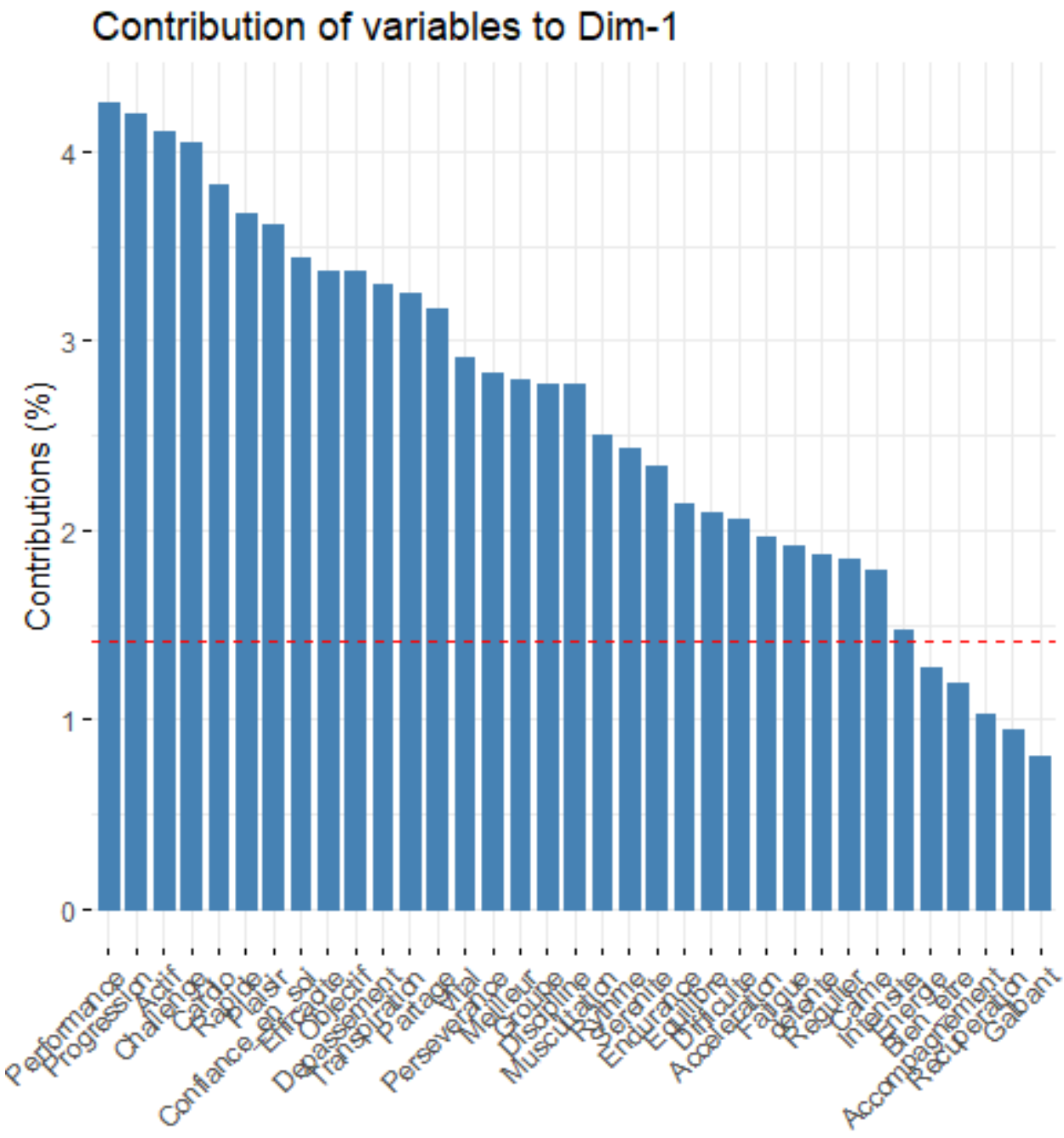

Contribution of variables to Dim-1
Contributions (%)
0
1
2
3
4
Performance
Progression
Actif
Challenge
Cardio
Rapide
Plaisir
en soi
Confiance
Efficacité
Objectif
Dépassement
Transpiration
Partage
Vital
Perseverance
Meilleur
Groupe
Discipline
Musculation
Rythme
Sérénité
Endurance
Equilibre
Difficulté
Accélération
Fatigue
détente
Réguler
Calme
Intensité
Energie
Bien être
Accompagnement
Récupération
Galbant

Figure 4:

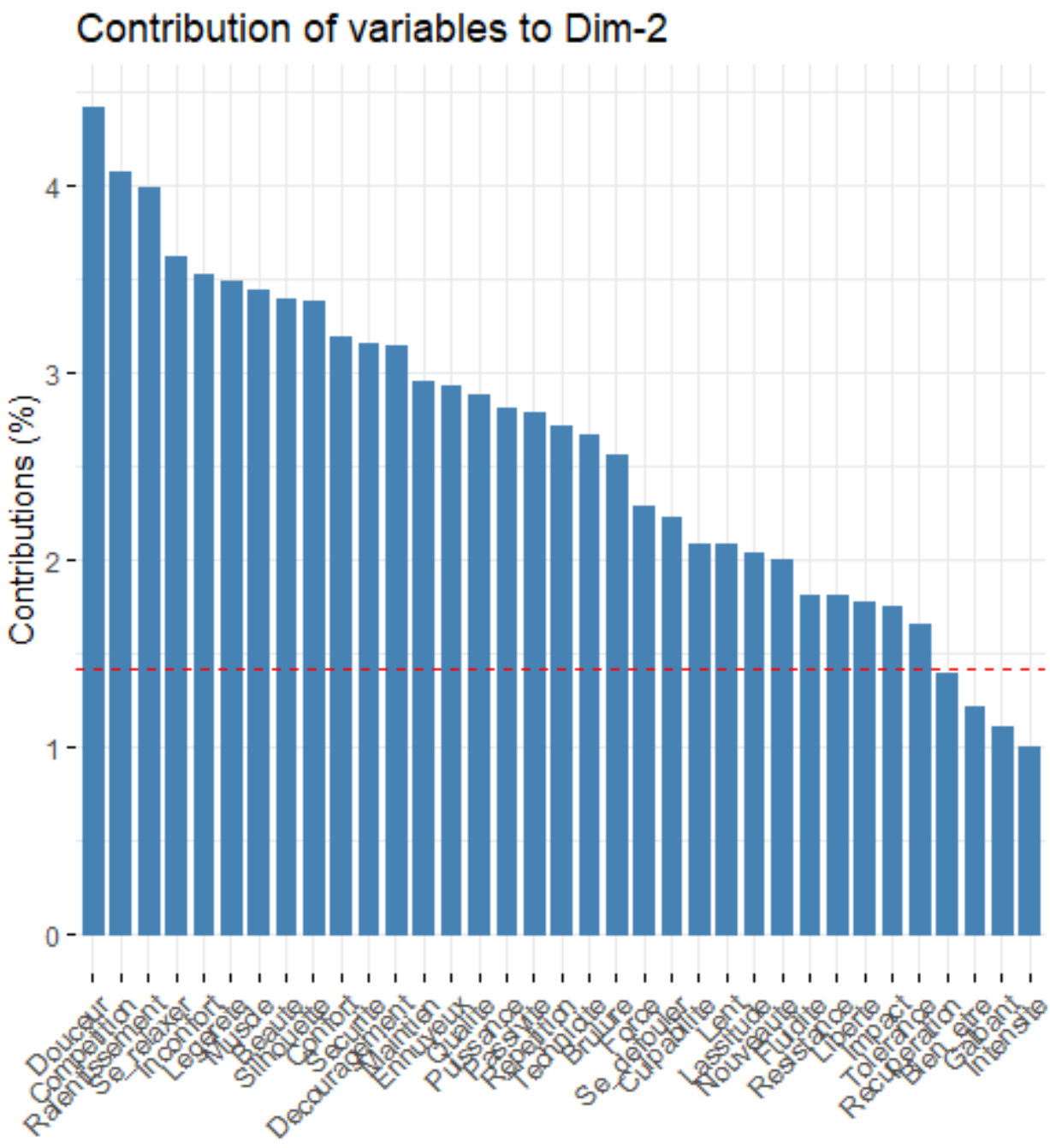

Contribution of variables to Dim-2
Contributions (%)
0
1
2
3
4
Douceur
Compétition
Raeni_ssement
Se_reaxer
Inconfort
Legerete
Musqe
Beaute
Silhouette
Confort
Securite
Decouragement
Maintien
Ennuyeux
Qualite
Puissance
Passivite
Repetition
Technicite
Brulure
Force
Se_depouiller
Culpabilite
Lent
Lassitude
Nouveaute
Fluidite
Resistance
Liberte
Impact
Tolerance
Recuperation
Bien_etre
Gabant
Intensite

Figure 5:

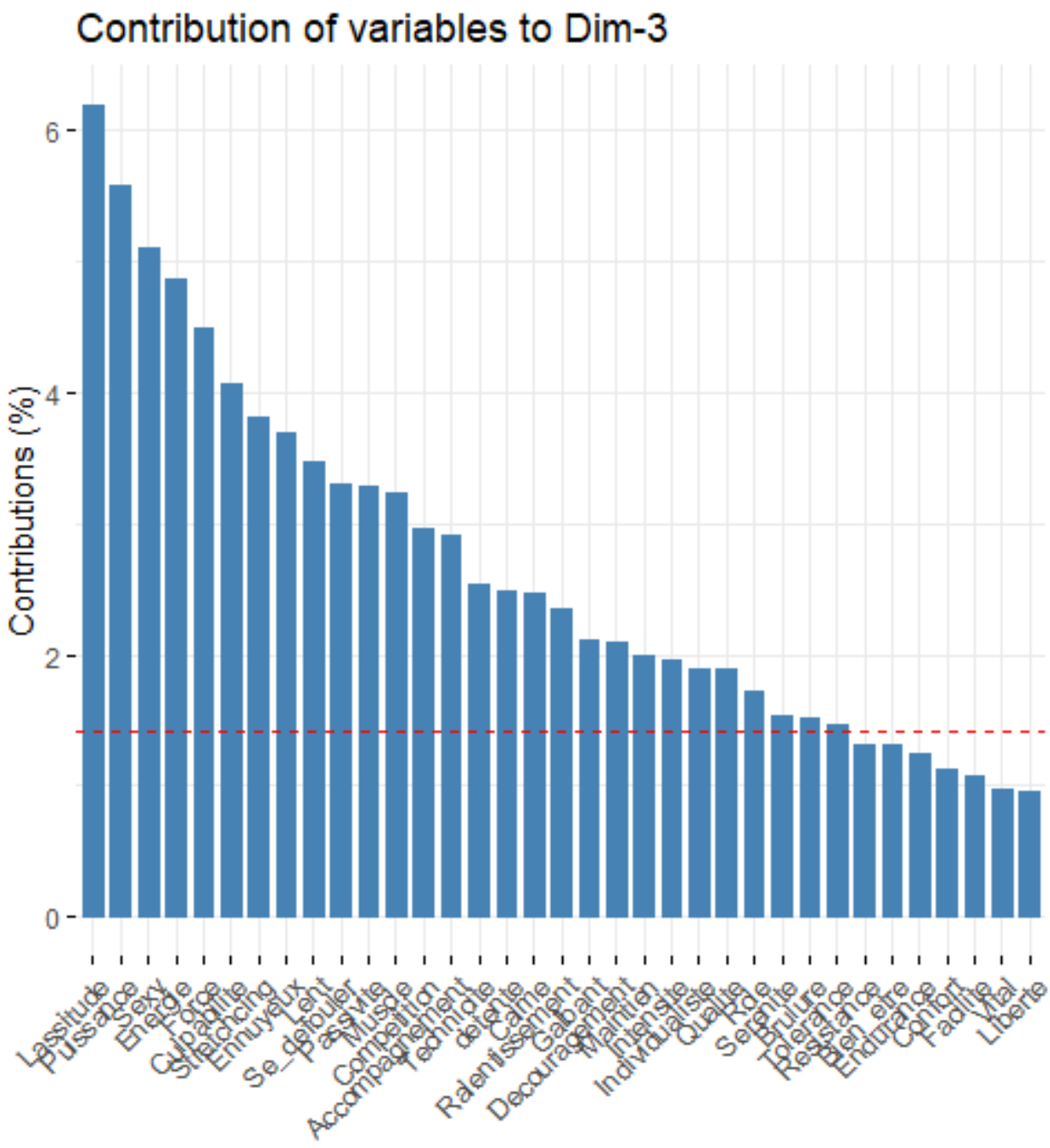

Contribution of variables to Dim-3
Contributions (%)
0
2
4
6
Lassitude
Puissance
Sexy
Energie
Force
Culpabilite
Stretching
Ennuyeux
Lent
Se_detouler
Passivite
Musee
Competition
Accompagnement
Technicite
detente
Calme
Ralentissement
Galbant
Decouragement
Maitrise
Intensite
Individualise
Qualite
Rage
Serenite
Brulure
Tolerance
Resistance
Bien etre
Endurance
Confort
Facilite
Vital
Liberte

Figure 6:

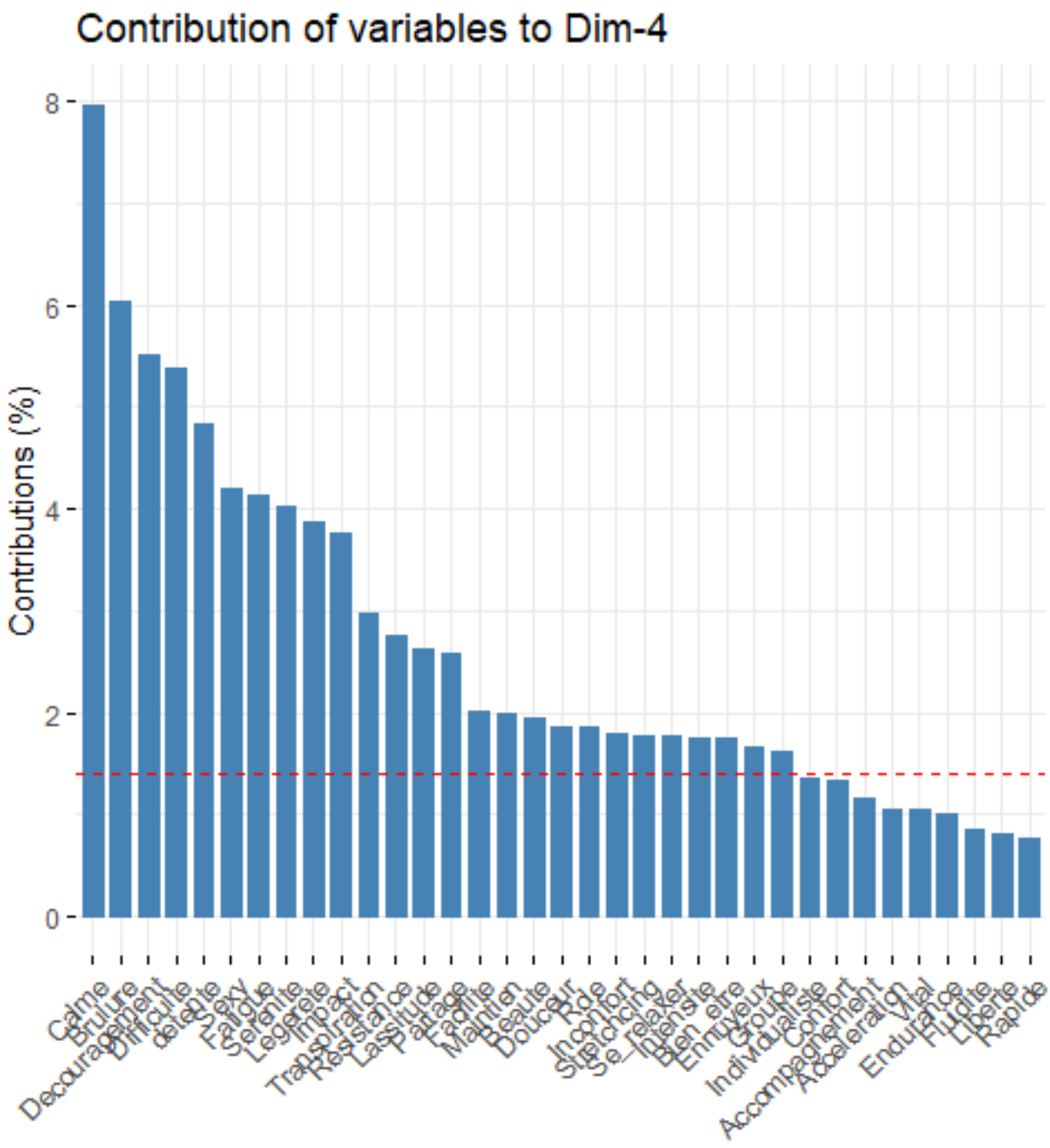

Contribution of variables to Dim-4
Contributions (%)
8
6
4
2
0
Calme
Brulure
Decouragement
Difficulte
detente
Sexy
Fatigue
Serenite
Legerete
Impact
Transpiration
Resistance
Lassitude
Paralysie
Faiblie
Maintien
Beaute
Douceur
Rode
Inconfort
Stretching
Se_relaxer
Intensite
Bien_etre
Ennuyeux
Groupe
Individualiste
Confort
Accompagnement
Acceleration
Vital
Endurance
Fluidite
Liberte
Rapide

Figure 7:

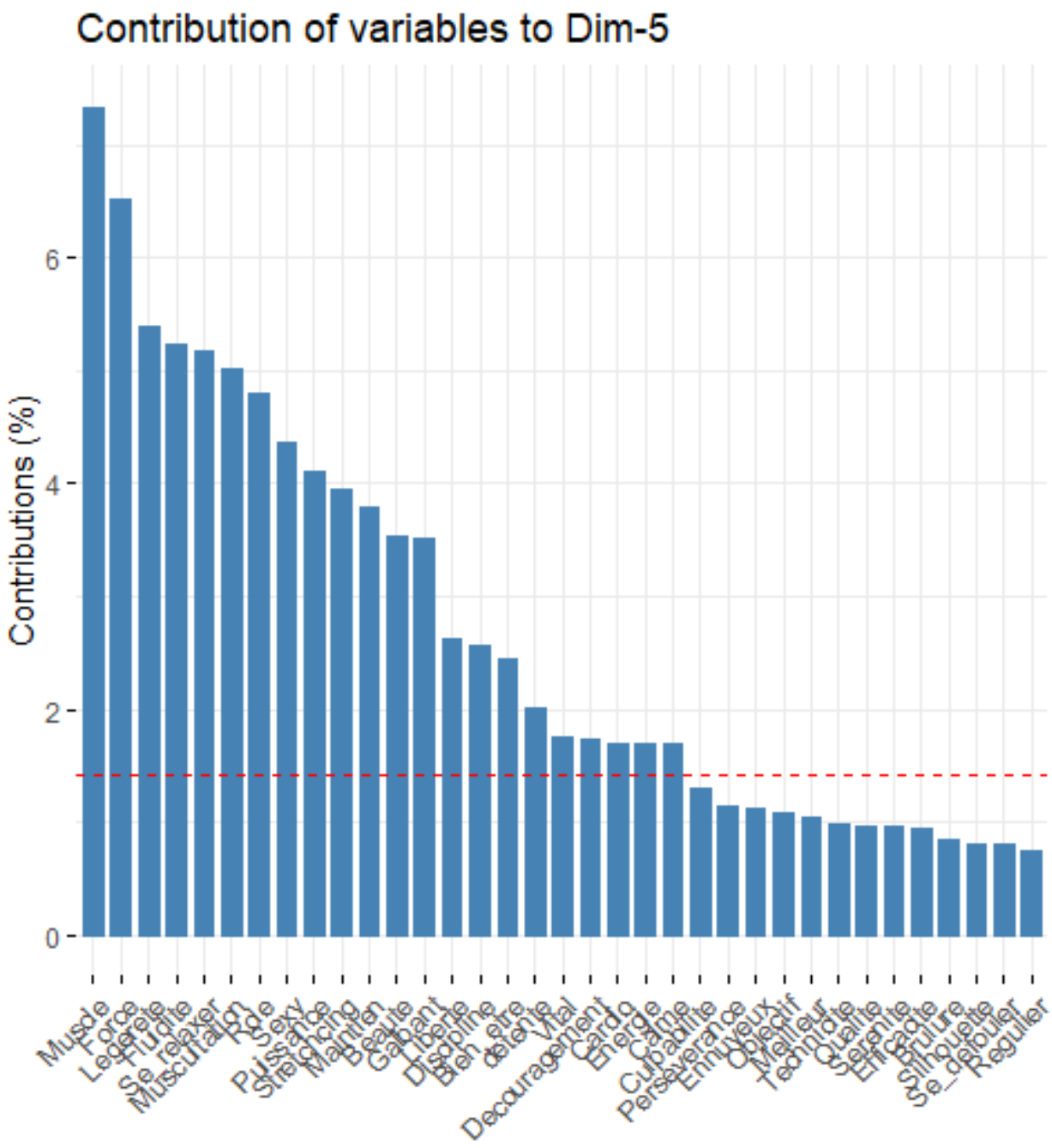


Figure 8:

Les variables les moins importants sont les suivantes : "Recuperation" et "Facilit".

4.2 Hierarchical Clustering on Principal Components

Pour raliser le clustering, nous allons utilis Hierarchical Clustering on Principal Components(HCPC). Cette mthode permet de combiner les trois mthodes standards utilises dans les analyses de donnes multivaries :

- Mthodes en composantes principales (PCA, CA, MCA, FAMD, MFA),
- Regroupement hirarchique et
- Clustering de partitionnement, en particulier la mthode des k-moyennes.

L'algorithme de la mthode HCPC a 4 principales tapes :

- 1) Effectue une ACP. Choisisse le nombre de dimensions retenir en spcifiant largument ncp. Dans notre cas ,la valeur est 5.
- 2) Applique la classification hirarchique sur le rsultat de ltape 1.
- 3) Choisisse le nombre de groupes en fonction du dendrogramme obtenu ltape 2. Un partitionnement initial est effectu.Dans notre cas ,la le nombre de groupes est 3.
- 4) Effectue le k-means pour amliorer le partitionnement initiale obtenu ltape 3.

Voici les lignes de codes principales pour :

```
res.pca <- PCA(data_base , ncp = 5 ,graph = TRUE)
res.hcpc <- HCPC(res.pca,nb.clust=3,consol=FALSE,graph=TRUE)

plot(res.hcpc,choice = "tree")
plot(res.hcpc,choice = "map", draw.tree = FALSE)
plot(res.hcpc,choice = "3D.map")
catdes(res.hcpc$data.clust,ncol(res.hcpc$data.clust))
```

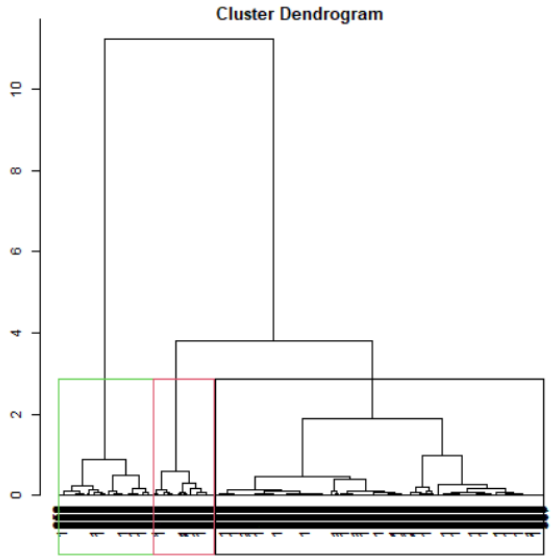


Figure 9: Hierarchical tree

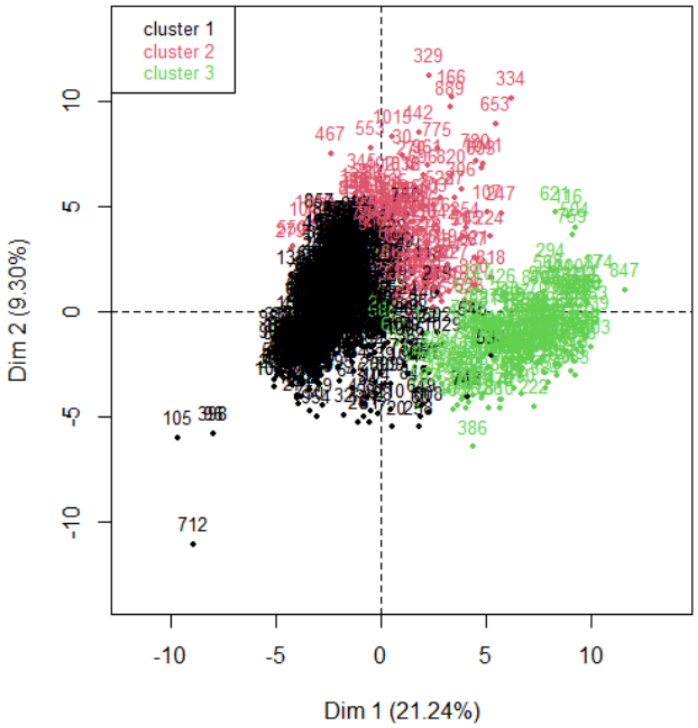


Figure 10: Ascending Hierarchical Classification of the individuals

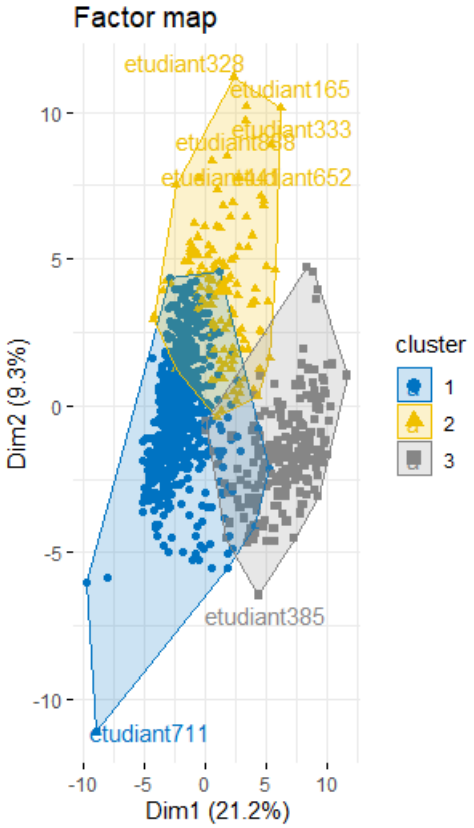


Figure 11: Ascending Hierarchical Classification of the individuals

4.3 Results of HCHC

Le cluster 1 est constitu d'individus partageant :

- des valeurs leves pour les variables Galbant, Culpabilite, Ennuyeux, Stretchcing, Securite, Decouragement, Ralentissement, Lassitude, Inconfort et Passivite (les variables sont tries partir des plus fortes).

- des valeurs faibles pour des variables comme Progression, Transpiration, Performance, Actif, Challenge, Plaisir, Objectif, Persvrance, Confiance en soi et Cardio (les variables sont tries partir des plus faibles).

Le cluster 2 est constitu d'individus partageant :

- des valeurs leves pour des variables comme Se dfouler, Puissance, Comptition, Technicit, Qualite, Energie, Confort, Muscle, Force et Intensite (les variables sont tries partir des plus fortes).

- des valeurs faibles pour les variables Sexy, Meilleur, Calme et Vital (les variables sont tries partir des plus faibles).

Le cluster 3 est constitu d'individus partageant :

- des valeurs leves pour des variables comme Progression, Actif, Performance, Challenge, Cardio, Partage, Plaisir, Depassement, Rapide et Efficacite (les variables sont tries partir des plus fortes).

- des valeurs faibles pour des variables telles que Confort, Securite, Galbant, Douceur, Ennuyeux, Force, Maintien, Qualite, Beaut et Inconfort (les variables sont tries partir des plus faibles).

5 Classification

Nous avons sparer nos donnes en 3 parties : - 80 % des donnes pour le choix de l'algorithme de slection . - 20 % des donnes pour tester le modele finale et ventuellement slectionner les colonnes les plus importantes.

5.1 Choice of Classification algorithm

4 classificateurs multi-classes sont utiliss : le classificateur vecteurs de soutien (SVC), classificateur vecteur de support linaire (LSVC), k-nearest neighbors (KNN) et rgression logistique (logreg). (KNN) et rgression logistique (logreg).

voici le graphique de la performance de chaque modle :

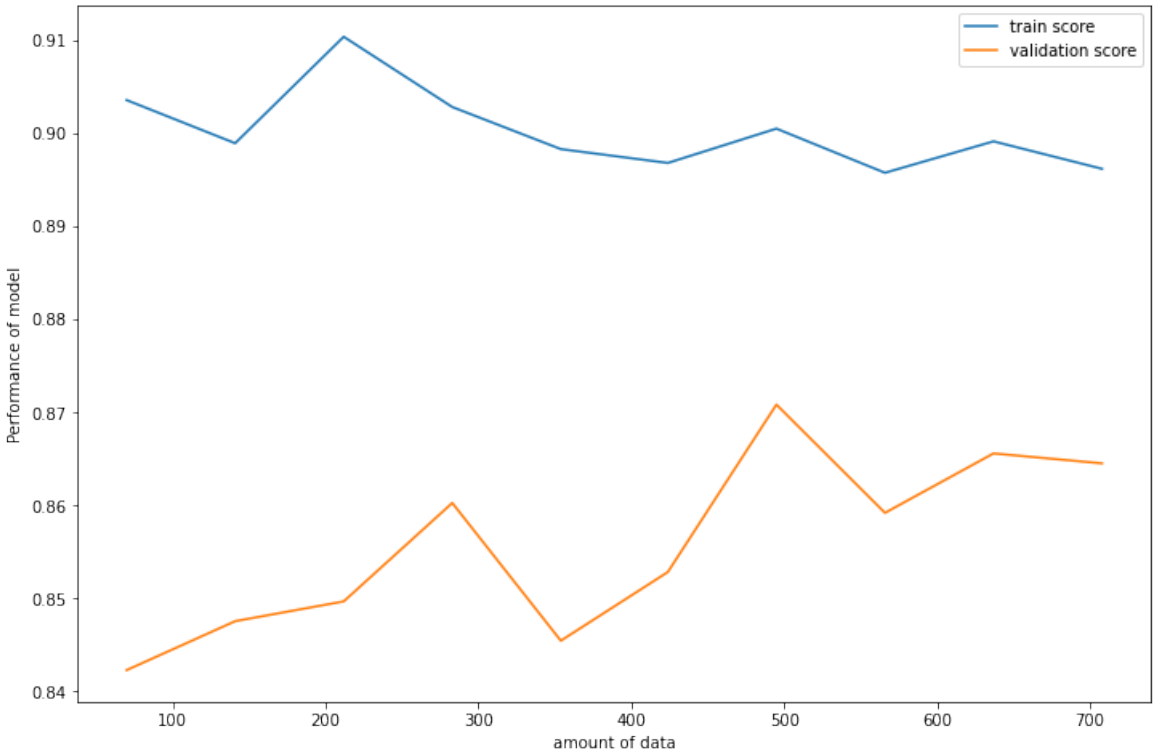


Figure 12: KNN learning curve

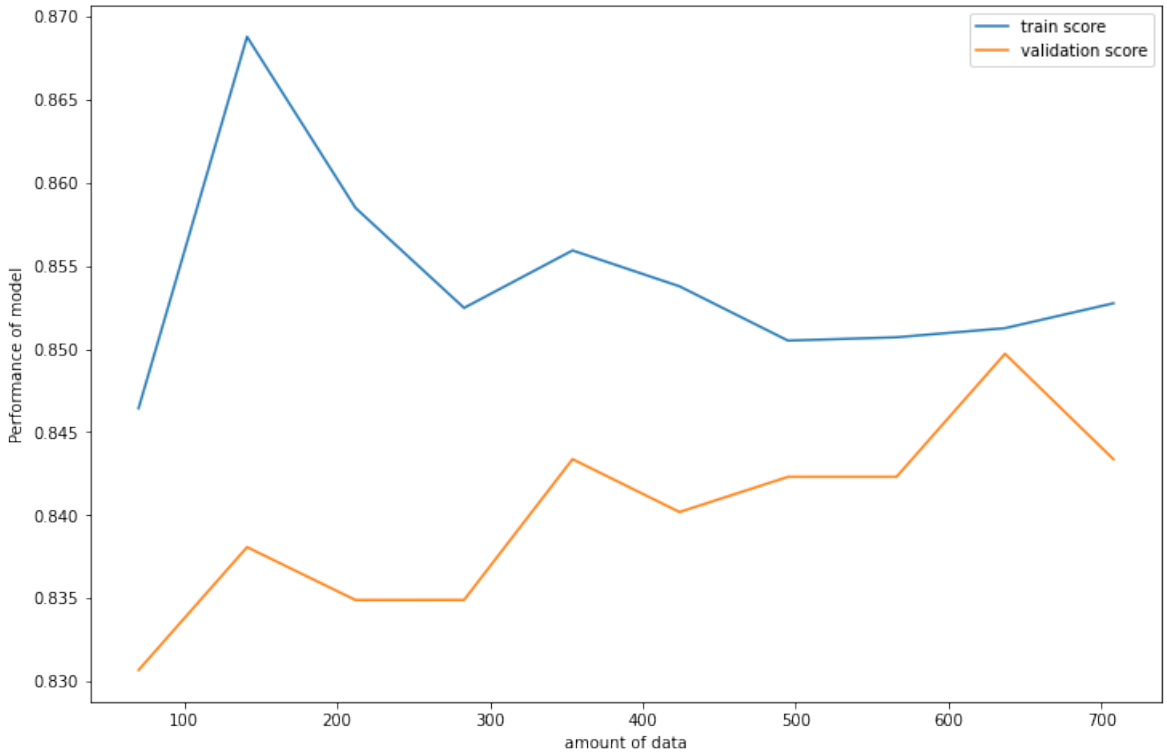


Figure 13: logreg learning curve

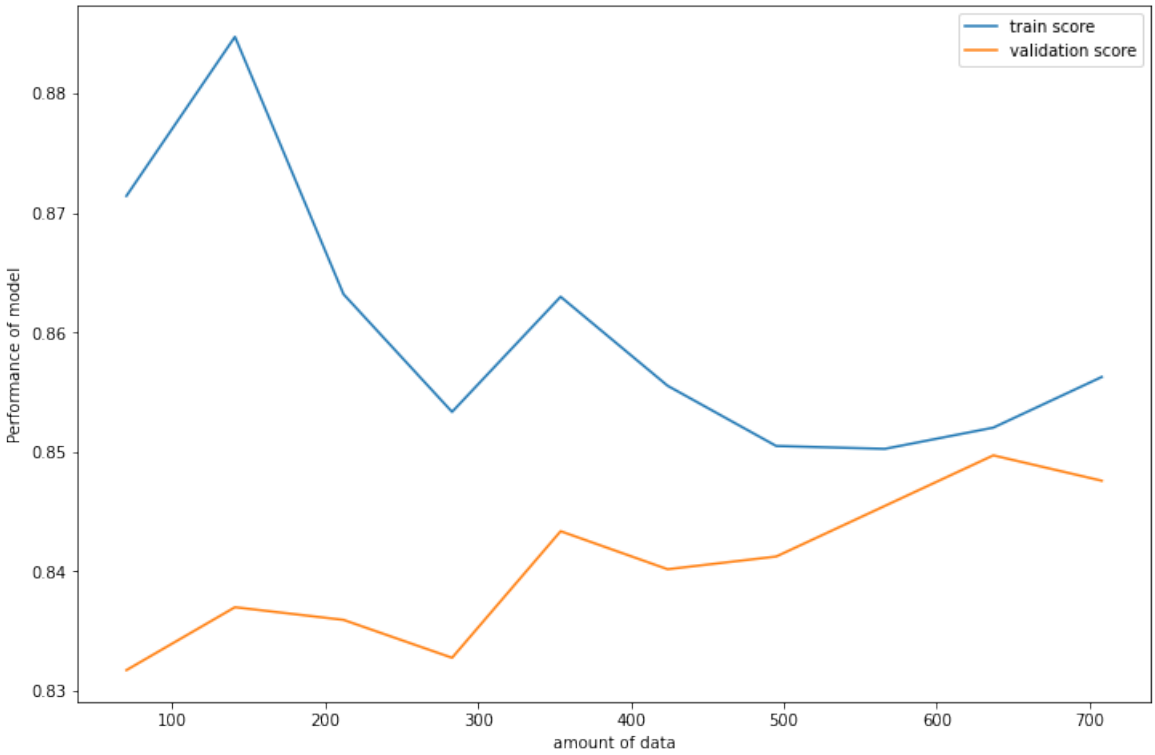


Figure 14: LSVC learning curve

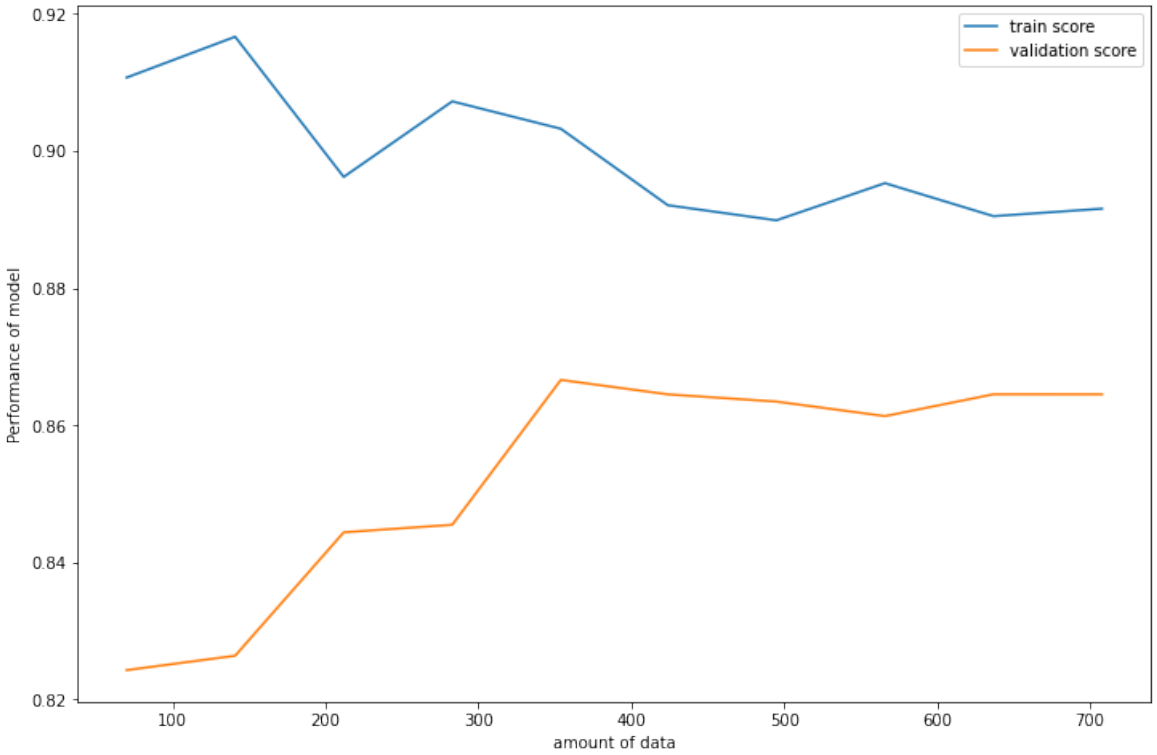


Figure 15: SVC learning curve

D'aprs les graphiques ci- dessous , la performance ou modele de lostigic Regression et SVC soit plus stable et meilleur que les autres modles. On peut dire que les deux modle ne sont pas en overfiting (score train et score val sont proches) contrairement aux 2 autres.

Nous avons utlis GridSearchCV pour optimser les hyperparametres du modele logistic Regression. GridSearchCV nous permet de les meilleurs hyperpametre en comparant les diffrents performances de chaque combinaison grce a la technique de cross-validation.

La cross-validation consiste dcouper le jeu de donnes en k parties gales (ici k = 5). Tour tour, chacune des k parties est utilise comme jeu de test. Le reste (autrement dit, lunion des k-1 autres parties) est utilis pour l'entranement.

5.2 Feature selection

Le graphique ci- dessous nous montre la variance de chaque features. 4 candidats le seuil se demarquent : 0.8 , 0.06 , 0.04 et 0.02.

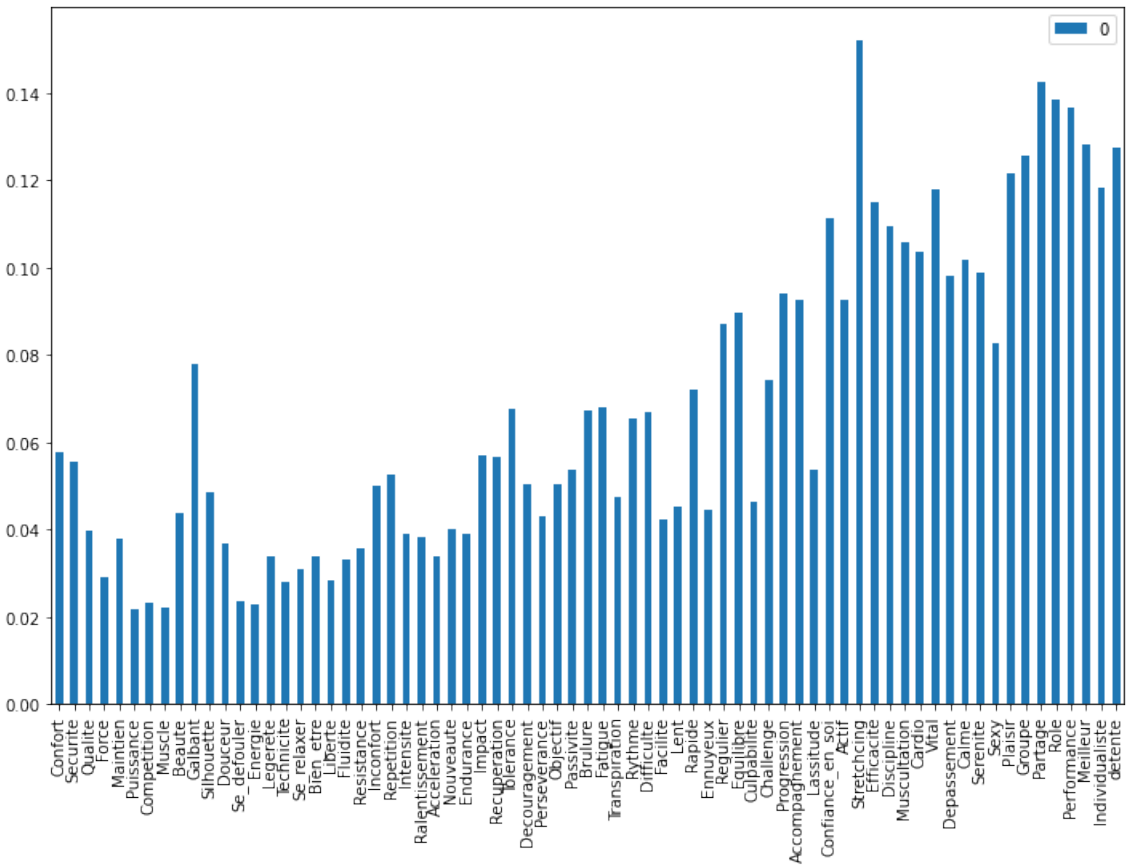


Figure 16: variance of each features

Pour liminer les valeurs infrieur ce seuil , La fonction VarianceThreshold de scikit-learn est utilis.

Au final, le seuil fix 0.02 donne de meilleurs rsultats. Aucun colonne n'a t suprime .

5.3 Results final model

	predicted: classe 0	predicted: classe 1	predicted: classe 2
Actual: classe 0	133	3	4
Actual: classe 1	1	30	0
Actual: classe 2	1	0	38

Figure 17: confusion matrix

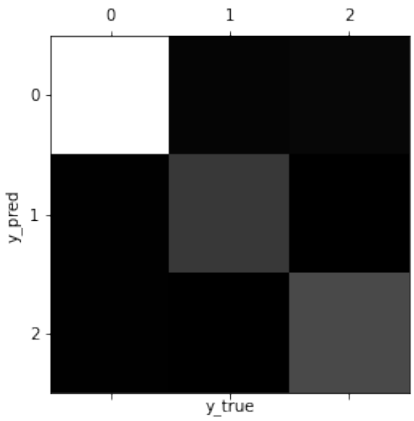


Figure 18: confusion matrix

Nous obtenons des rsultats satisfaisant : seulement 9 valeurs mal places , la valeur de 0.964 et 0.94768 pour le recall_score et le f1score.

6 Conclusion

L'objectif principal du projet tait de faire du clustering sur nos donnes. Grace au clustering HCPC , nous pouvons distinguer 3 types d'tudiants :

- les dmotivs , ceux qui recherchent le bien-tre et la simplicit dont certaines variables caractristiques du cluster sont : Culpabilite, Ennuyeux, Decouragement, Ralentissement, Inconfort
- Dans le deuxime groupe , on a ceux qui aiment les sports de combat comme la lutte, la boxe le MMA. il est caractris par ses mots : Puissance, Competition, Technicite, Qualite, Energie, Muscle, Force and Intensite
- Le dernier groupe se distingue par ses mots : Progression, Performance, Challenge, Cardio, Partage, Depassement, Rapide and Efficacite. On retrouve ceux qui apprcient la course pied et les activit de nature .

En ce qui concerne la classification ,L'un des meilleur algorithme (SVM) a donn la valeur 0.9549 de prcision, 0.9 de rappel et 0.9255 de f1-score.(Par contre, l'ensemble de donnes de ce projet n'est pas assez grand.)

References

[1] https://www.cairn.info/revue-staps-2018-2-page-99.htm

[2] https://solidarites-sante.gouv.fr/prevention-en-sante/preserver-sa-sante/article/activite-physique-et-sante

[3] https://e3s.unistra.fr/equipe/presentation/

[4] https://scikit-learn.org

[5] https://fr.wikipedia.org/wiki/Analyse_en_composantes_principales

[6] http://www.sthda.com/english/

[7] http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-
117-hcpc-hierarchical-clustering-on-principal-components-essentials/
#algorithm-of-the-hcpc-method

[8] https://husson.github.io/teaching.html

[9] https://www.youtube.com/c/MachineLearnia

[10]