Report V1

May 25, 2022

Contents

1 Introduction

La pratique rgulire dactivits physiques comporte de nombreux bnfices et il apparat primordial dencourager le maintien de celle-ci chez les jeunes en dveloppement (Janssen LeBlanc, 2010 ; Tremblay et al., 2011). Nanmoins, on observe un dclin de la pratique dactivits physiques au cours de ladolescence . (Garriguet Colley, 2012 ; Keating, Guan, Piero, Bridges, 2005).

Afin augmenter leur activit physique , ce projet tentera de dterminer leur niveau de motivation ou centre d'intrłt pour leur proposer une activit physique qu'il aime .

Regular physical activity has many benefits and it is essential to encourage its maintenance in It seems essential to encourage the maintenance of physical activity among developing youth (Janssen LeBlanc, 2010; Tremblay et al., 2011). Nevertheless, there is a decline in physical activity during adolescence. (Garriguet Colley, 2012; Keating, Guan, Piero, Bridges, 2005).

In order to increase their physical activity, this project will attempt to determine their level of motivation or interest to offer them a physical activity that they enjoy.

Motivation is an essential parameter for the engagement of young people in voluntary physical activity or sport. Traditionally, the level of motivation is measured by questionnaires which are restrictive to fill in. In connection with the University of Lille, an application has been developed to identify the type of motivation on a more functional mode.

2 Objective

2.1 STAPS university

2.2 Main Objectives

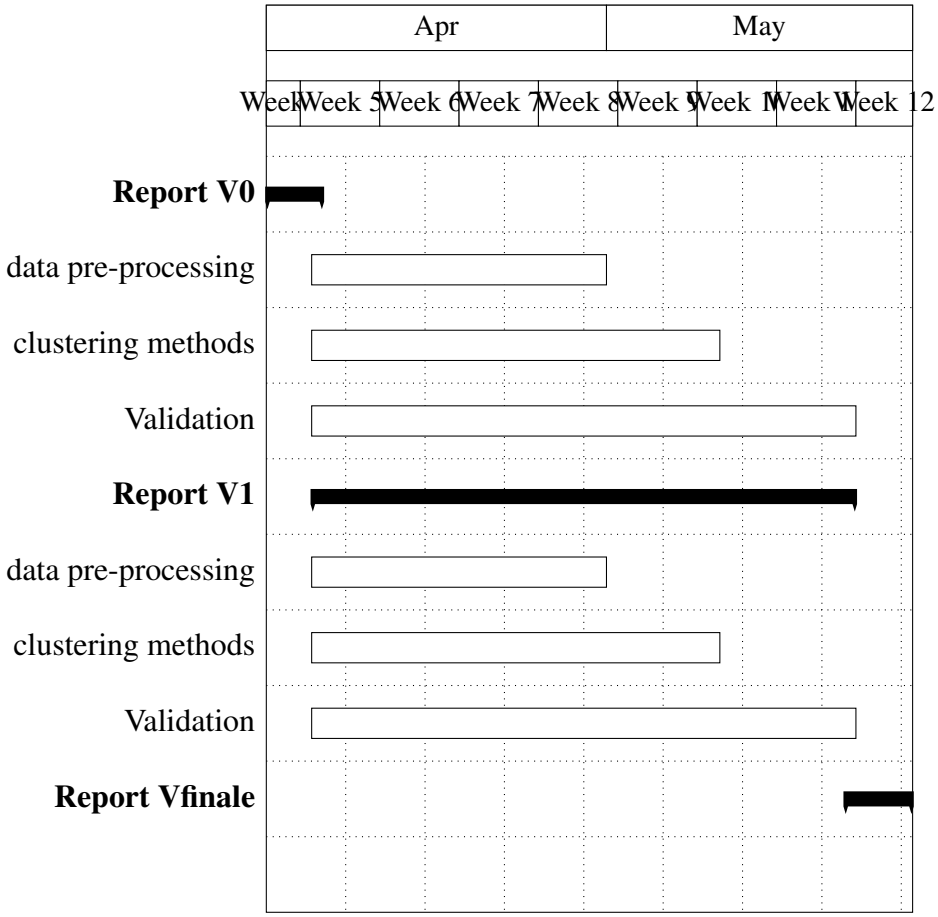
The objective of this work is to identify profiles of practitioners based on positive or negative qualifiers, i.e., to assign a profile to each cluster of the data and to estimate the strength of these profiles, i.e., the number of clusters or profiles that are most representative of the data.

2.3 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, la rgression linaire,...

We will first perform a preprocessing of the data by renormalizing the data, removing outliers, completing or removing missing values, then to analyze the data we will use different algorithms such as: K-means, principal component analysis, decision trees. Finally, we will test the robustness of our cluster by using classification algorithms such as: logistic regression, linear regression,...

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



3 Description of raw data

Le dataset contient des informations personnels sur les lycens qui sont au total 1070 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) sont obtenu 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 łtes 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".

The dataset contains personal information about the students such as their initial, high school,

gender, choice of study, parents' work and parents' support as well as date of birth, body shape (height and weight). Twenty variables measure the nature of motivation such as enjoyment, affiliation, physical condition and the degree of motivation such as SIMS intrinsic and SIMS external regulation.

Finally the rest of the variables (71) are obtained in the following way: we ask a question: "In PE, what is the sport that you enjoyed the most?

Then one indicates to him: "We are now going to present you with words that will allow you to describe your feelings about this sport. Your job is to indicate, as quickly as possible, whether you agree or disagree with these propositions by clicking on yes or no. The response time has been taken into account in each answer. If this time is short, it means that the term seems obvious. For example, if the sport is "soccer", the student could answer "yes" quickly to the qualifier "fun", "no" quickly to the qualifier "beauty".

The possible answers to each question are "yes", "no", "I don't know". When the answer to a question is "yes", the time value is positive, negative in the case of "no" and zero in the case of "I don't know".

	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

Fig. 1: Raw data

4 Preprocessing

4.1 Preprocessing Method

Dans ce projet , Les valeurs manquantes ont t mise 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 a t calcul en utilisant les donnes non signes dans le but de rduire les valeurs extrłme 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 a t faite par ligne dans le but de conserver ce qui est "important" pour chaque personne.

In this project, the missing values have been set to zero assuming that the qualifier does not interest the students concerned and that they could have answered: "I don't know". For the management of the outliers, those which are higher than 5*standard deviation have been set to zero in order not to impact the weight given to each word. The standard deviation was calculated using the unsigned data in order to reduce the extreme values and avoid the possible compensation of the values. The students who answered "I don't know" to all these questions were not considered in the rest of the project. The normalization was done by line in order to keep what is "important" for each person.

4.2 Results of processed data

Qualite	Force	Maintien	uissanc	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

Fig. 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 nettoie contient 1050 lycens et 71 caractristiques.

The most ”important“ and least ”important“ words for each person are respectively close to either 1 or -1 . Those which are less ”important“ are close to zero.

The cleaned dataset contains 1050 students and 71 features.

5 Clustering

5.1 Principal component analysis

L'analyse en composantes principales est une mthode de la famille de l'analyse des donnes et plus gnralement de la statistique multivarie, qui consiste transformer des variables lies entre elles (dites corrles en statistique) en nouvelles variables decorrles les unes des autres. Ces nouvelles variables sont nommes composantes principales ou axes principaux. Elle permet au statisticien de rsumer l'information en rduisant le nombre de variables.

Dans notre projet, nous allons utiliser ACP pour rduire la dimension des donnes en quelques variables et garder les donnes les plus importants.

Pour dterminer le nombre de composante optimale , nous avons utiliser la fonction suivante dans Rstudio.

Principal component analysis is a method of the data analysis family and more generally of multivariate statistics, which consists in transforming multivariate statistics, which consists in transforming variables that are linked to each other (called ”correlated“ in statistics) into new variables that are decorrelated from each other. These new variables are called ”principal components“ or principal axes. It allows the statistician to summarize information by reducing the number of variables.

In our project, we will use PCA to reduce the size of the data to a few variables and keep the most important data.

To determine the number of optimal components, we use the following function in Rstudio:

```
fviz_eig(res.pca, addlabels = TRUE, ylim = c(0, 50))
```

Cette fonction permet d'avoir le graphique des valeurs propres.Les valeurs propres mesurent la quantit de variance explique par chaque axe principal.Les valeurs propres sont grandes pour les premiers axes et petits pour les axes suivants.

The eigenvalues measure the amount of variance explained by each principal axis. The eigenvalues are large for the first axes and small for the following axes.

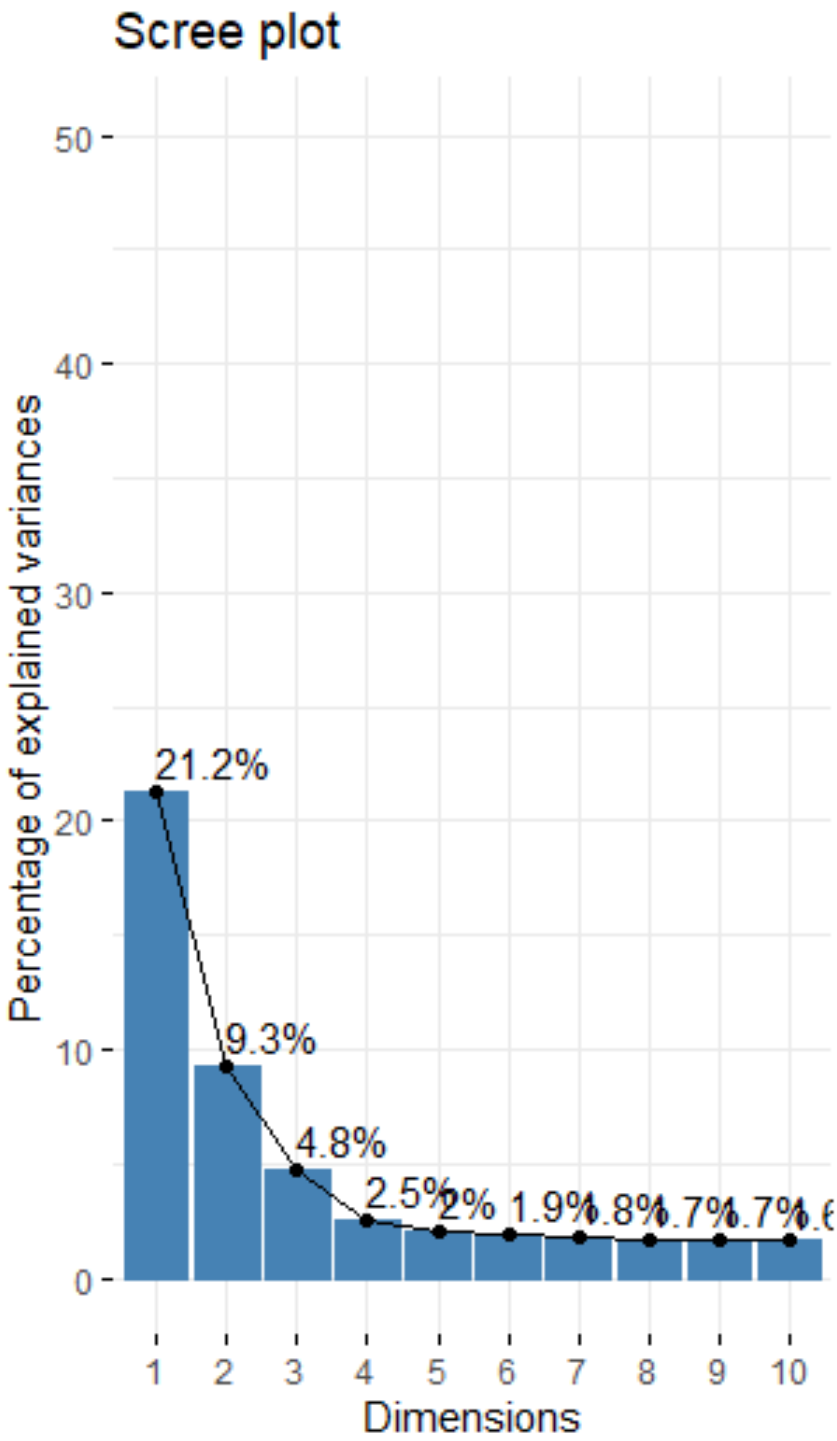


Fig. 3:

D'aprs le graphique ci-dessus, nous pourrions vouloir nous arrłter 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.

According to the graph above, we might want to stop at the fifth principal component stop at the fifth principal component because the variation is less after the fifth.

However 39.79760 % of the information (variances) contained in the data is retained by the in the data is retained by the first 5 principal components.

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.

The graphs below show the top 30 variables contributing the most to the principal components. The red dotted lines on the graphs indicate the average contribution value.

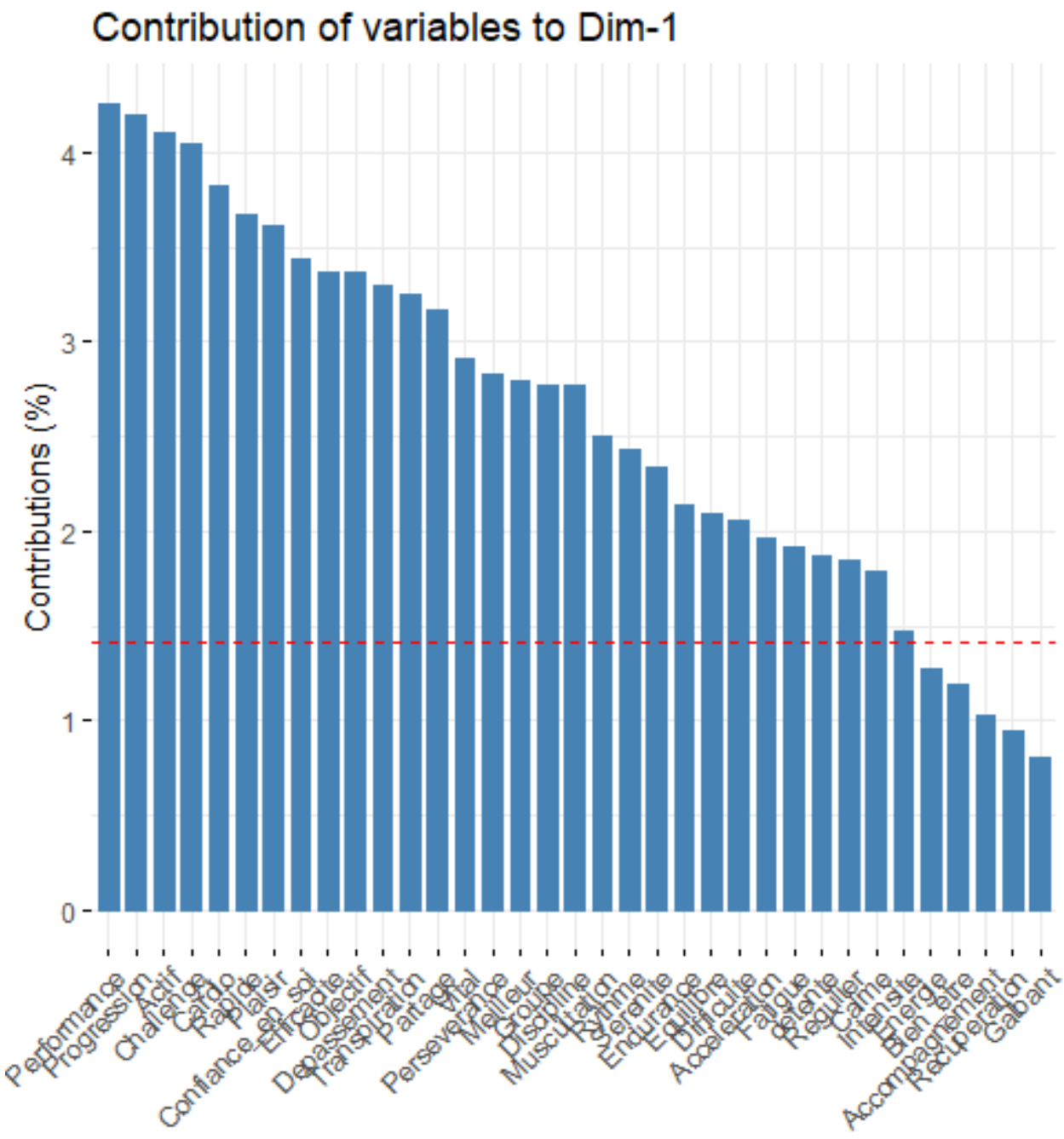

Contribution of variables to Dim-1
Contributions (%)
0
1
2
3
4
Performance
Progression
Actif
Challenge
Cardio
Rapide
Plaisir
Confiance_en soi
Efficacité
Objectif
Depassement
Transpiration
Partage
Vital
Perseverance
Meilleur
Groupe
Discipline
Musculation
Rythme
Serenité
Endurance
Equilibre
Difficulté
Acceleration
Fatigue
detente
Reguler
Calme
Intensité
Energie
Bien etre
Accompagnement
Recuperation
Gabant

Fig. 4:

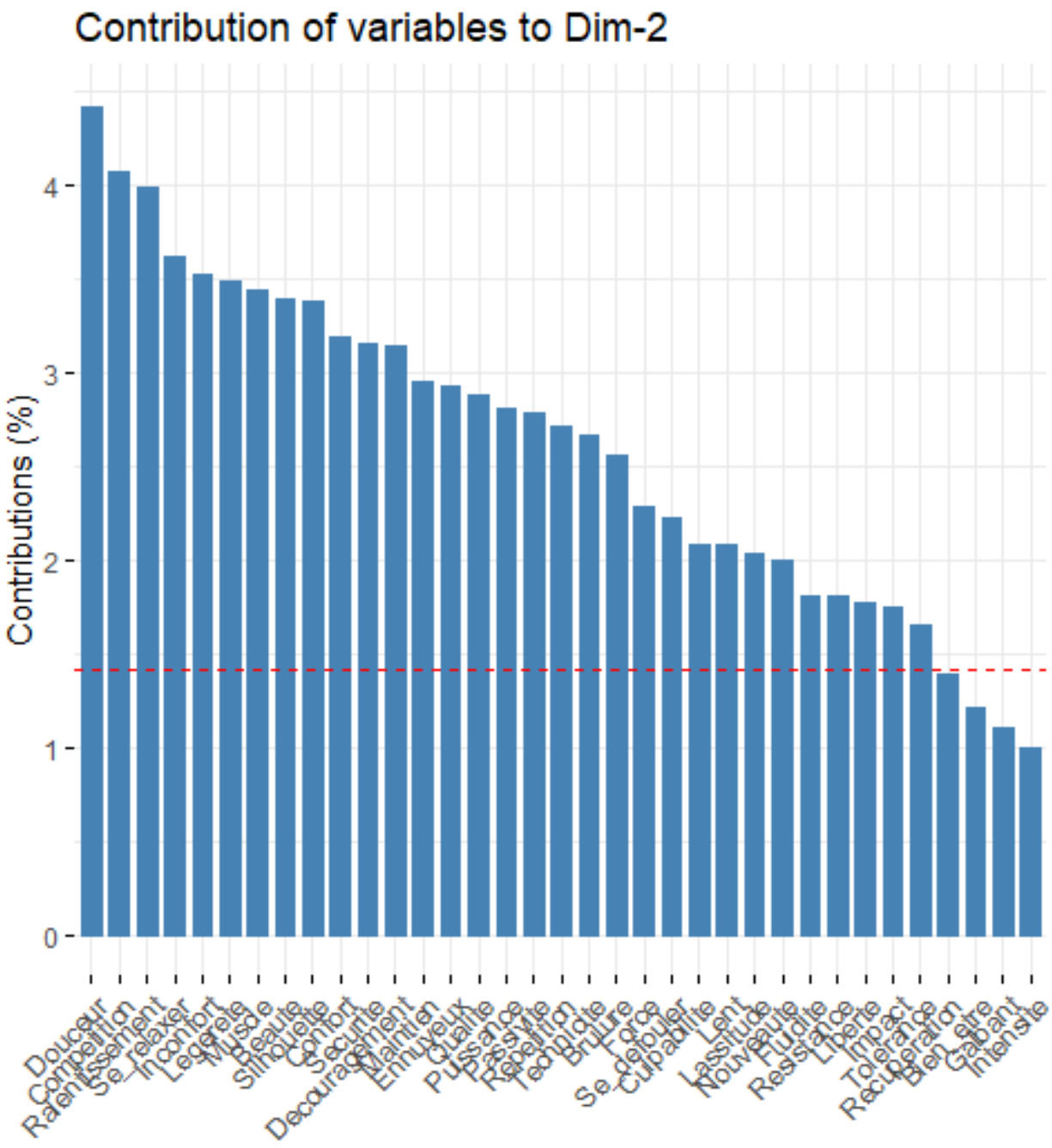

Contribution of variables to Dim-2
Contributions (%)
0
1
2
3
4
Douceur
Competition
Ralentissement
Se_relaxer
Inconfort
Legerete
Musde
Beaute
Silhouette
Confort
Securite
Decouragement
Maintien
Ennuyeux
Gaiete
Puissance
Passivite
Repetition
Rapidite
Bruture
Force
Se_defouler
Culpabilite
Lent
Lassitude
Nouveaute
Fluidite
Resistance
Liberte
Impact
Tolerance
Recuperation
Bien_etre
Gabant
Intensite

Fig. 5:

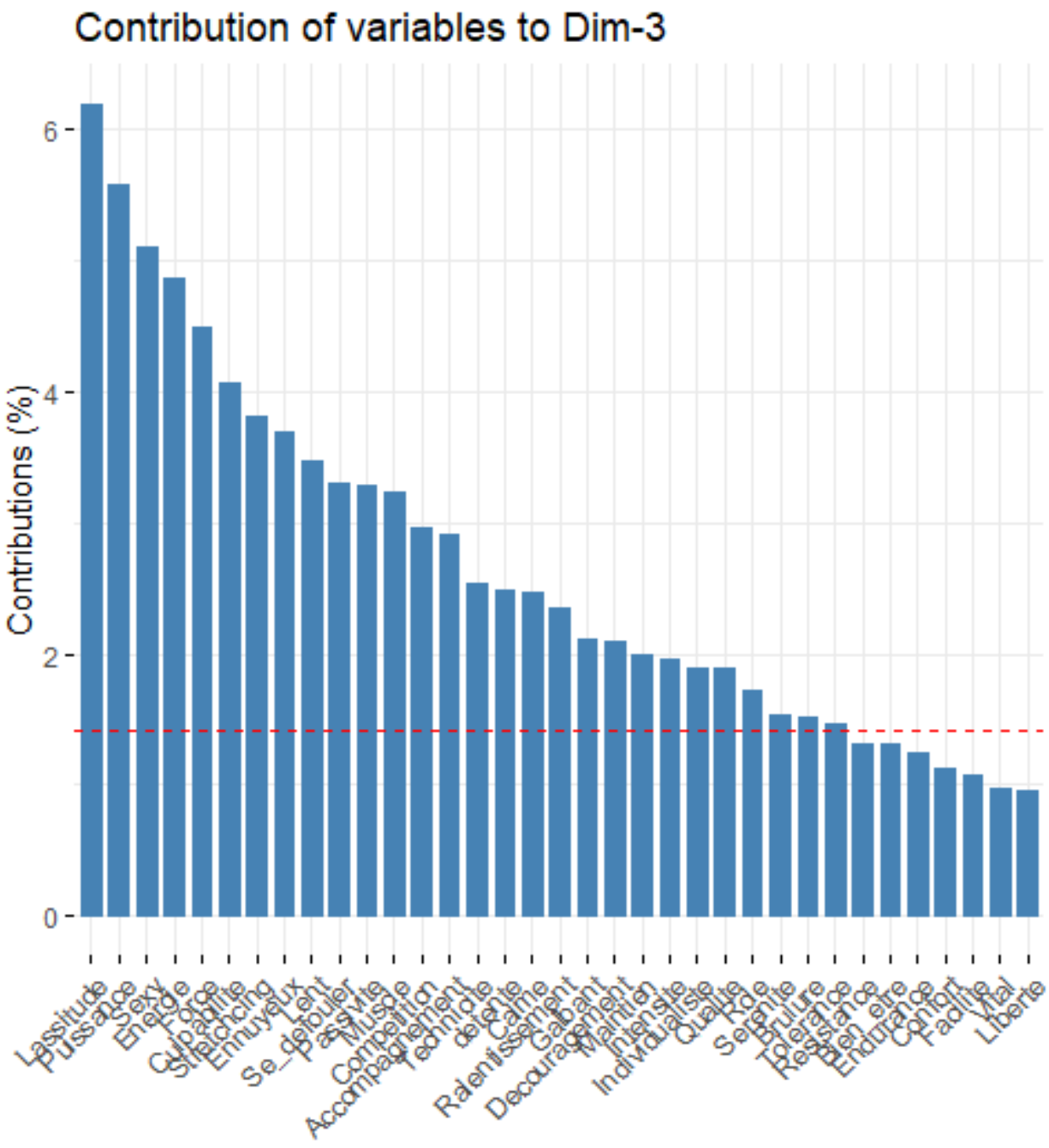

Contribution of variables to Dim-3
Contributions (%)
0
2
4
6
Lassitude
Puissance
Sexy
Energie
Force
Capable
Stretching
Ennuyeux
Lent
Se_delasser
Passif
Musclé
Competition
Accompagnement
Technique
detente
Calme
Ralentissement
Decouragement
Gabant
Maitrise
Intensite
Individualiste
Qualite
Rage
Serenite
Brulure
Tolerance
Resistance
Bien_etre
Endurance
Confort
Facilite
Vital
Liberte

Fig. 6:

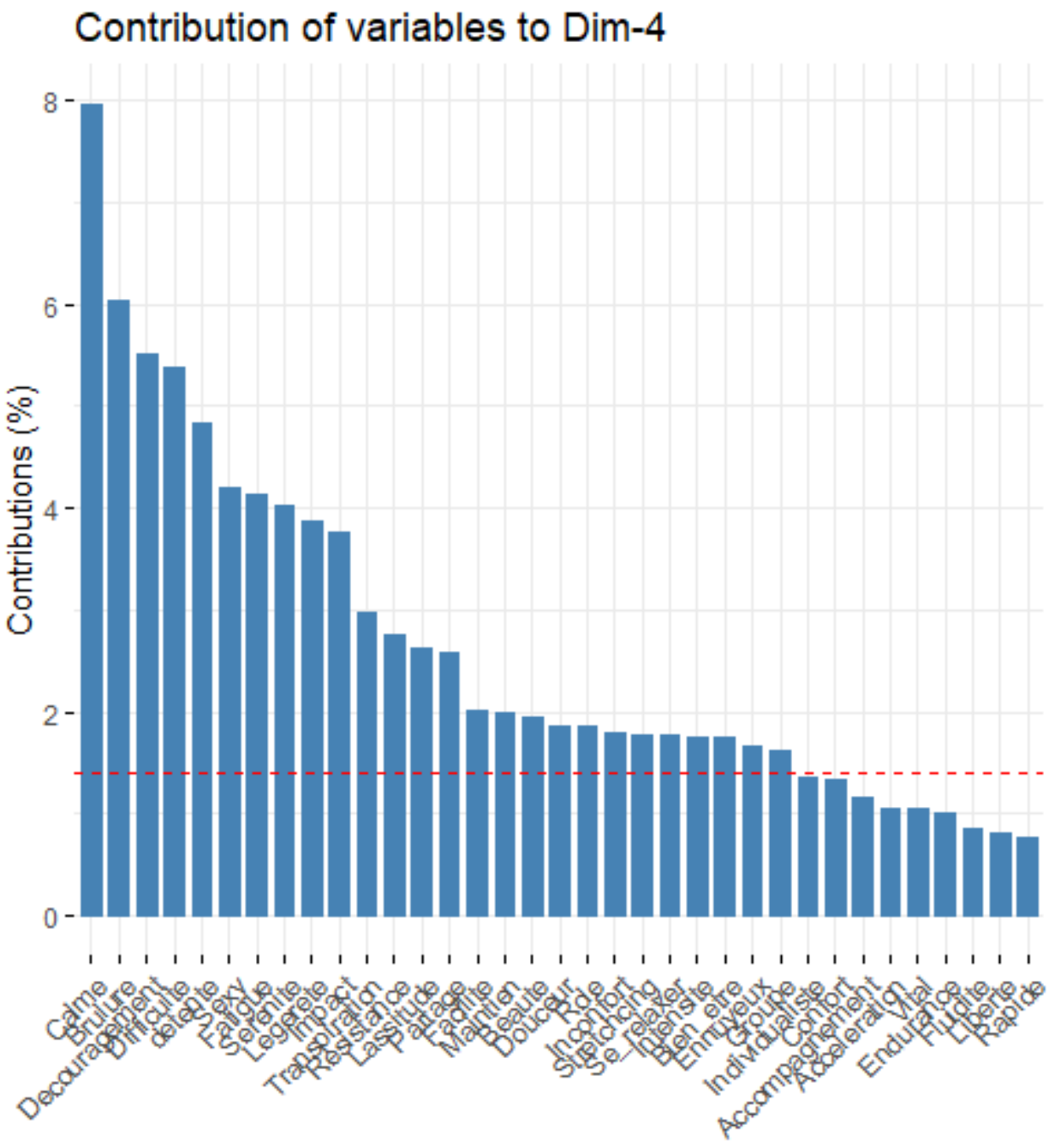

Contribution of variables to Dim-4
Contributions (%)
8
6
4
2
0
Calme
Brulure
Decouragement
Difficulte
detente
Sexy
Fatigue
Serenite
Legerete
Impact
Transpiration
Resistance
Lassitude
Partage
Facilite
Maitrise
Beaute
Douceur
Rode
Inconfort
Stretching
Se_relaxer
Intensite
Bien etre
Ennuyeux
Groupe
Individualiste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

Fig. 7:

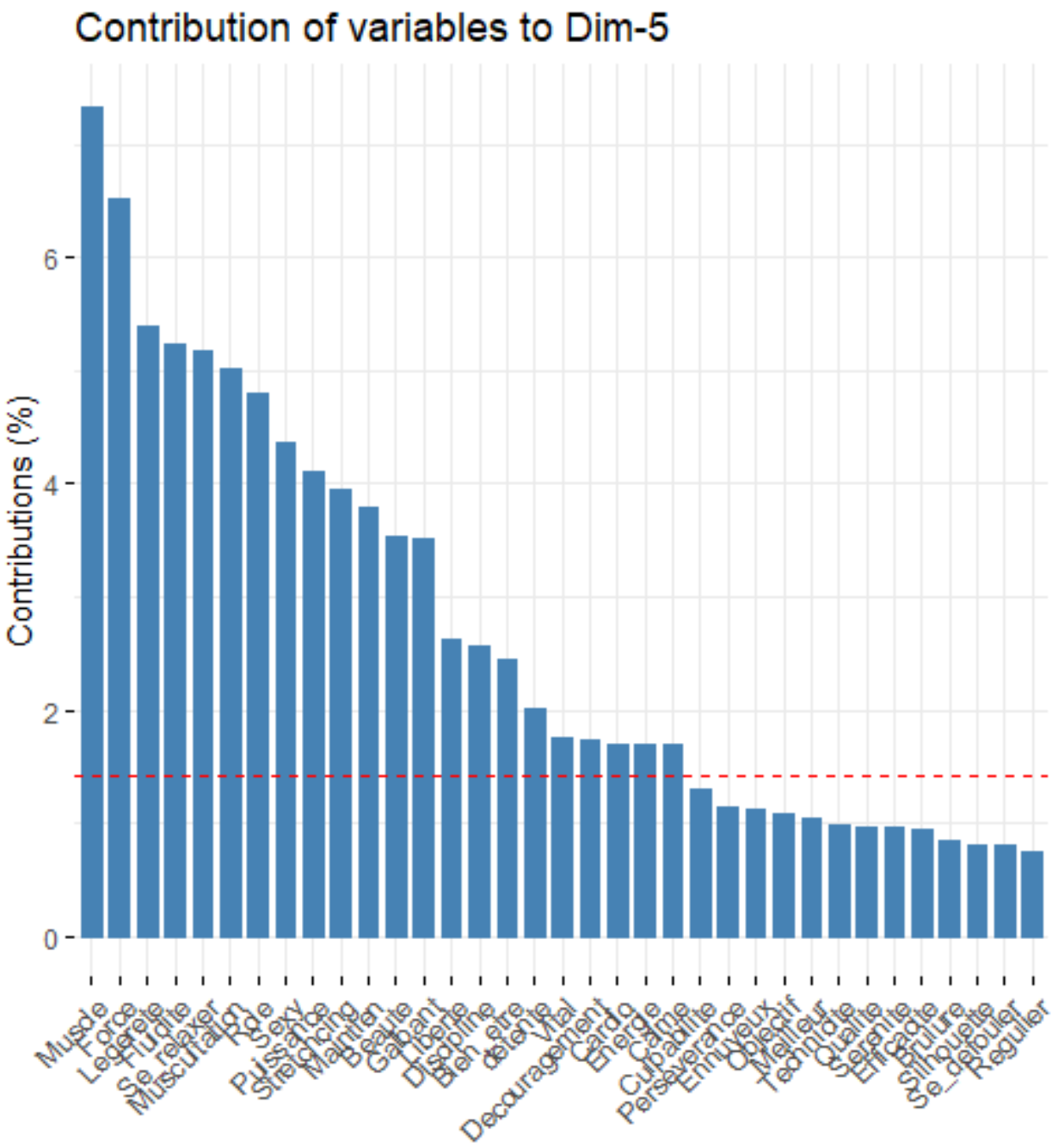


Fig. 8:

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

The least important variables are : ”Recuperation“ et ”Facilit“.

5.2 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 :

To make the clustering, we will use Hierarchical Clustering on Principal Components (HCPC). This method combines the three standard methods used in multivariate data analysis:

Principal Component Methods (PCA, CA, MCA, FAMD, MFA), Hierarchical clustering and Partitioning clustering, in particular the k-means method.

The HCPC method algorithm has 4 main steps:

1) Performs a PCA. Selects the number of dimensions to retain by specifying the argument ncp. In our case, the value is 5.

2) Apply the hierarchical clustering on the result of step 1.

3) Choose the number of groups according to the dendrogram obtained in step 2. In our case, the number of groups is 3.

4) Performs the k-means to improve the initial partitioning obtained in step 3.

Here are the main lines of code for :

```
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))
```

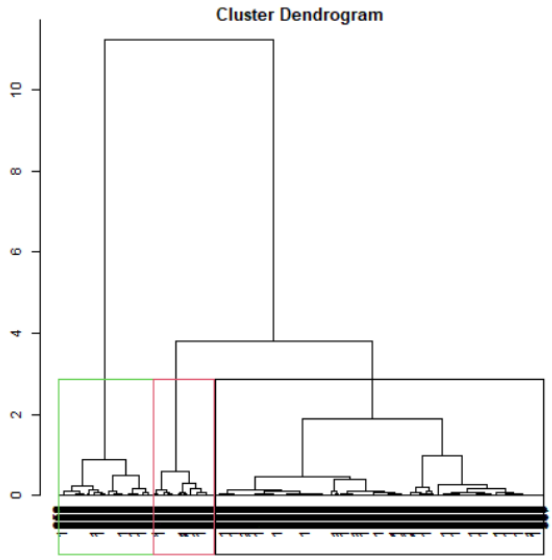


Fig. 9: Hierarchical tree

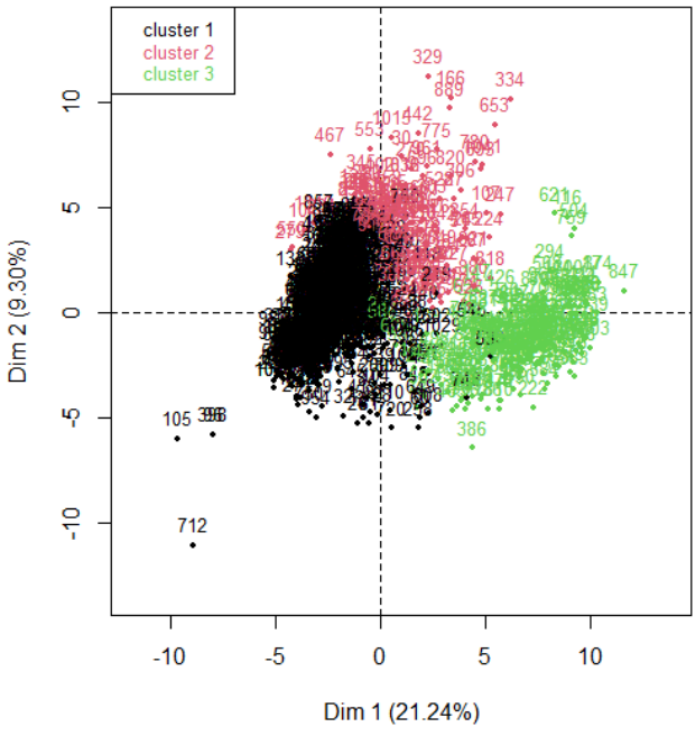


Fig. 10: Ascending Hierarchical Classification of the individuals

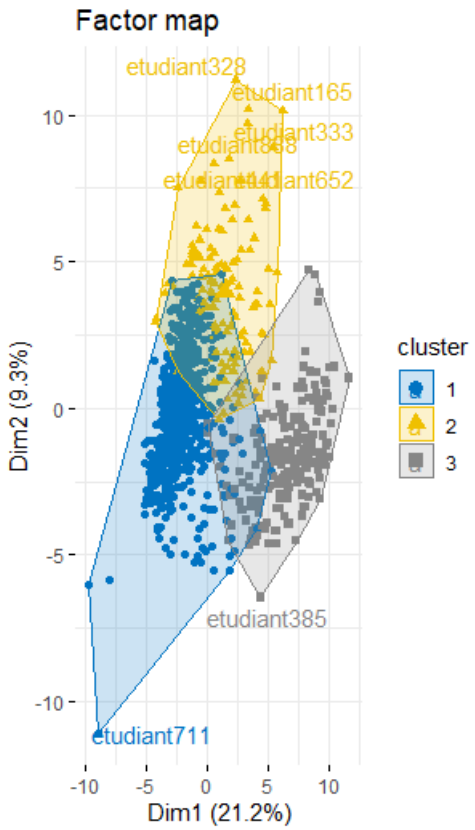


Fig. 11: Ascending Hierarchical Classification of the individuals

5.3 Results of HCHC

The cluster 1 is made of individuals sharing :

- high values for the variables Galbant, Culpabilite, Ennuyeux, Stretchcing, Securite, Decouragement, Ralentissement, Lassitude, Inconfort et Passivite (variables are sorted from the strongest).

- low values for variables like Progression, Transpiration, Performance, Actif, Challenge, Plaisir, Objectif, Perseverance, Confiance en soi and Cardio (variables are sorted from the weakest).

The cluster 2 is made of individuals sharing :

- high values for variables like Se defouler, Puissance, Competition, Technicite, Qualite, Energie, Confort, Muscle, Force and Intensite (variables are sorted from the strongest).

- low values for the variables Sexy, Meilleur, Calme and Vital (variables are sorted from the weakest).

The cluster 3 is made of individuals sharing :

- high values for variables like Progression, Actif, Performance, Challenge, Cardio, Partage, Plaisir, Depassement, Rapide and Efficacite (variables are sorted from the strongest).

- low values for variables like Confort, Securite, Galbant, Douceur, Ennuyeux, Force, Maintien, Qualite, Beaute and Inconfort (variables are sorted from the weakest).

6 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 .

6.1 Choice of Classification algorithm

6.2 Feature selection

Les donnees de validation serviront pour suprimer les valeurs a faible variance et le seuil a ete fixe 0.05 .

Voici la commande utiliser pour la raliser :

```
# elemination des colonnes a variances inferieur au seuil 0.1 ou 0.05
selector = VarianceThreshold(threshold=0.05)
selector.fit_transform(X_valid)
columns_selected = np.array(X_val.columns)[selector.get_support()]
print(columns_selected)
```

En ce qui concerne les donnes de test

6.3 Results of Classification

. presentation des resultats

7 Results and Conclusion

References

[1] https://scikit-learn.org/stable/index.html