

Report V1

Congo Job

May 27, 2022

Contents

1	Introduction	1
1.1	Sport et sciences sociales	1
1.2	Main Objectives	2
1.3	Specific Objectives	2
2	Description of raw data	2
3	Preprocessing	3
3.1	Preprocessing Method	3
3.2	Results of processed data	3
4	Clustering	5
4.1	Principal component analysis	5
4.2	Hierarchical Clustering on Principal Components	11
4.3	Results of HCHC	13
5	Classification	14
5.1	Choice of Classification algorithm	14
5.2	Feature selection	16
5.3	Results final model	17
6	Conclusion	18

1 Introduction

The regular practice of physical activities has many benefits such as as an improvement of mental health, prevention of cardiovascular diseases, limiting weight gain and many others.

However, we observe a decline in the practice of physical activities. It appears primordial to encourage young people to maintain a physical activity or to become more active, it is in this optics that this project is registered.

1.1 Sport et sciences sociales

Created in Strasbourg(1979) by Bernard Michon , the Sport and Social Sciences research unit remains the only STAPS research unit in the Grand Est region and is recognized as a key research structure in the social sciences of sport in France and Europe. With more than 20 researchers (full and associate) and 17 doctoral students, it produces reference works and articles (more than 174 publications).

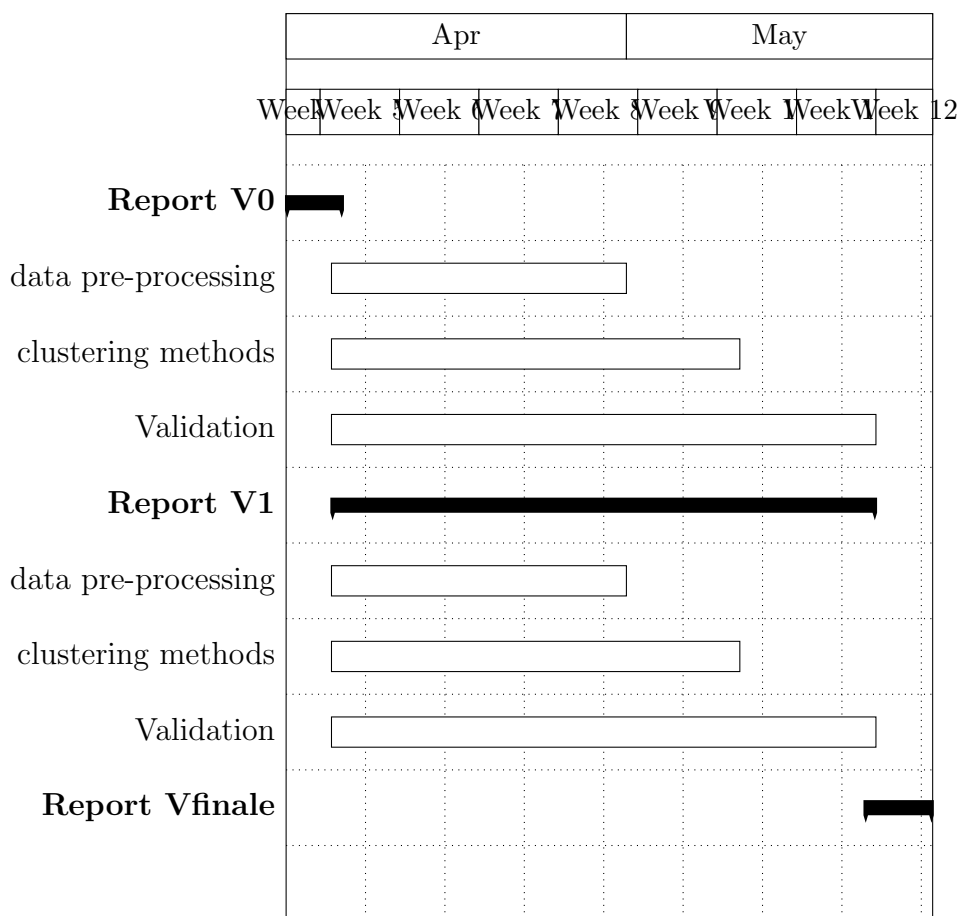
1.2 Main Objectives

The objective of this work is to identify practitioner profiles based on positive or negative qualifiers, i.e., to assign a profile to each cluster in 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.

1.3 Specific Objectives

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, k-nearest neighbors,...

The Gantt chart below gives us a quick overview of the organization of the work over time.



2 Description of raw data

The dataset contains personal information about the high school students (1070 participants) such as their initial, high school, gender, study choice, 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) were collected as follows: We ask a question: "In PE, what is the sport that you enjoyed the most?"

Then we indicate: "We are now going to present you with words that will allow you to

describe how you feel 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	Compétition	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

The missing values are set to zero assuming that the qualifier is not of interest to 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 is 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 is done by line in order to keep what is "important" for each person.

3.2 Results of processed data

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.

Qualite	Force	Maintien	Puissance	Compétitivité	Muscle	Beauté	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

The cleaned dataset contains 1050 high school students and 71 features.

4 Clustering

4.1 Principal component analysis

Principal Component Analysis (PCA) is used to reduce the size of the data to a few variables and keep the most important data. It is a method of multivariate statistics, which consists in transforming variables related to each other (called "correlated" in statistics) into new variables decorrelated from each other.

To determine the number of optimal components, the function `fviz_eig` of Rstudio was used. This function allows to have the graph of the eigenvalues.

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

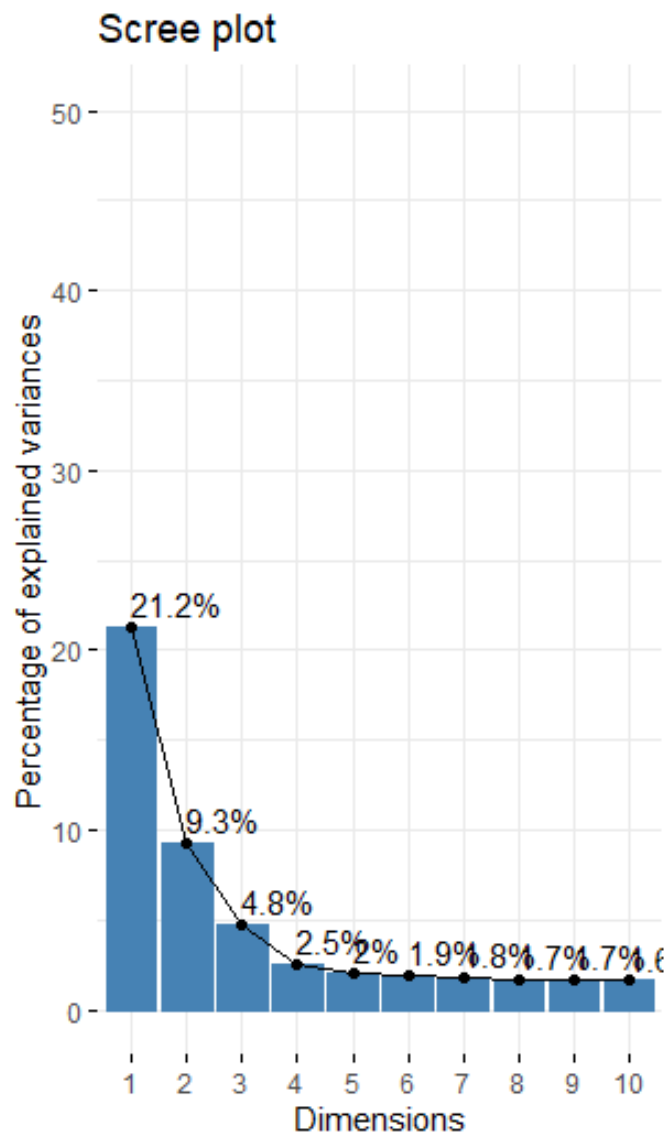


Figure 3:

From the graph above, we might want to stop at the fifth principal component because 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 first 5 principal components. in the data is retained by the first 5 principal components.

The graph below shows the top 30 variables that contribute the most to the 5 principal components. The red dotted lines on the graphs indicate the average contribution value.

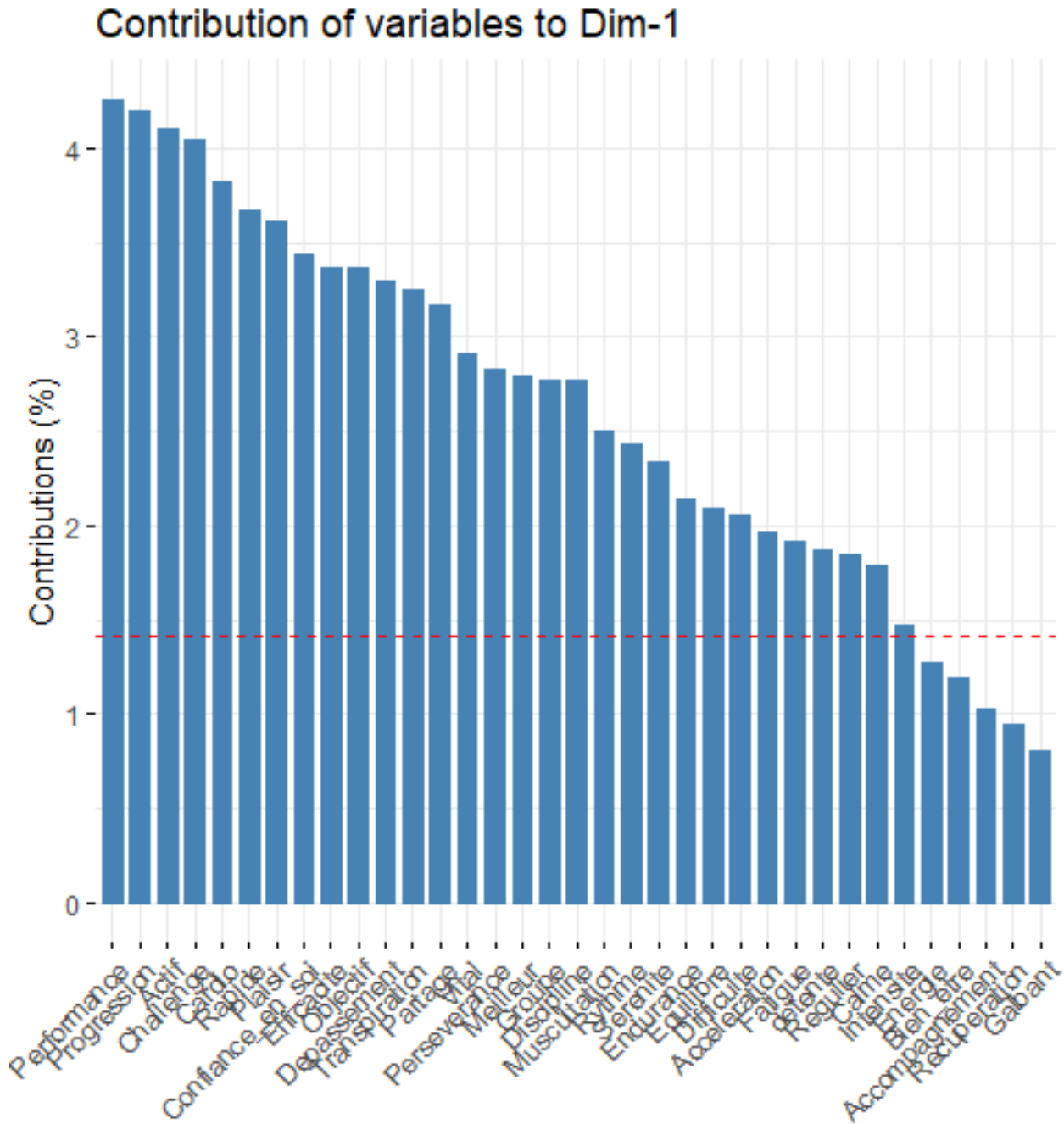


Figure 4:

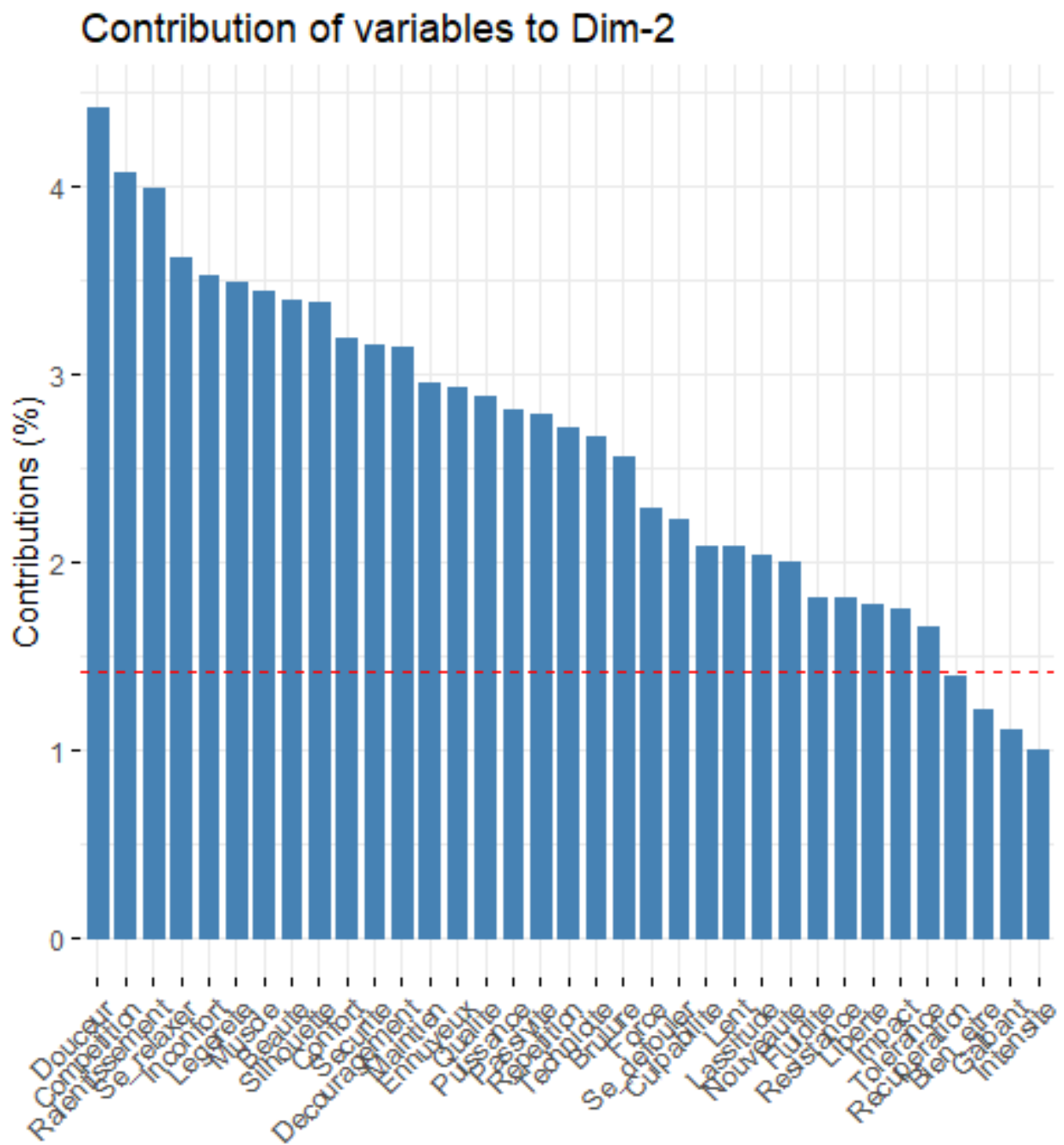


Figure 5:

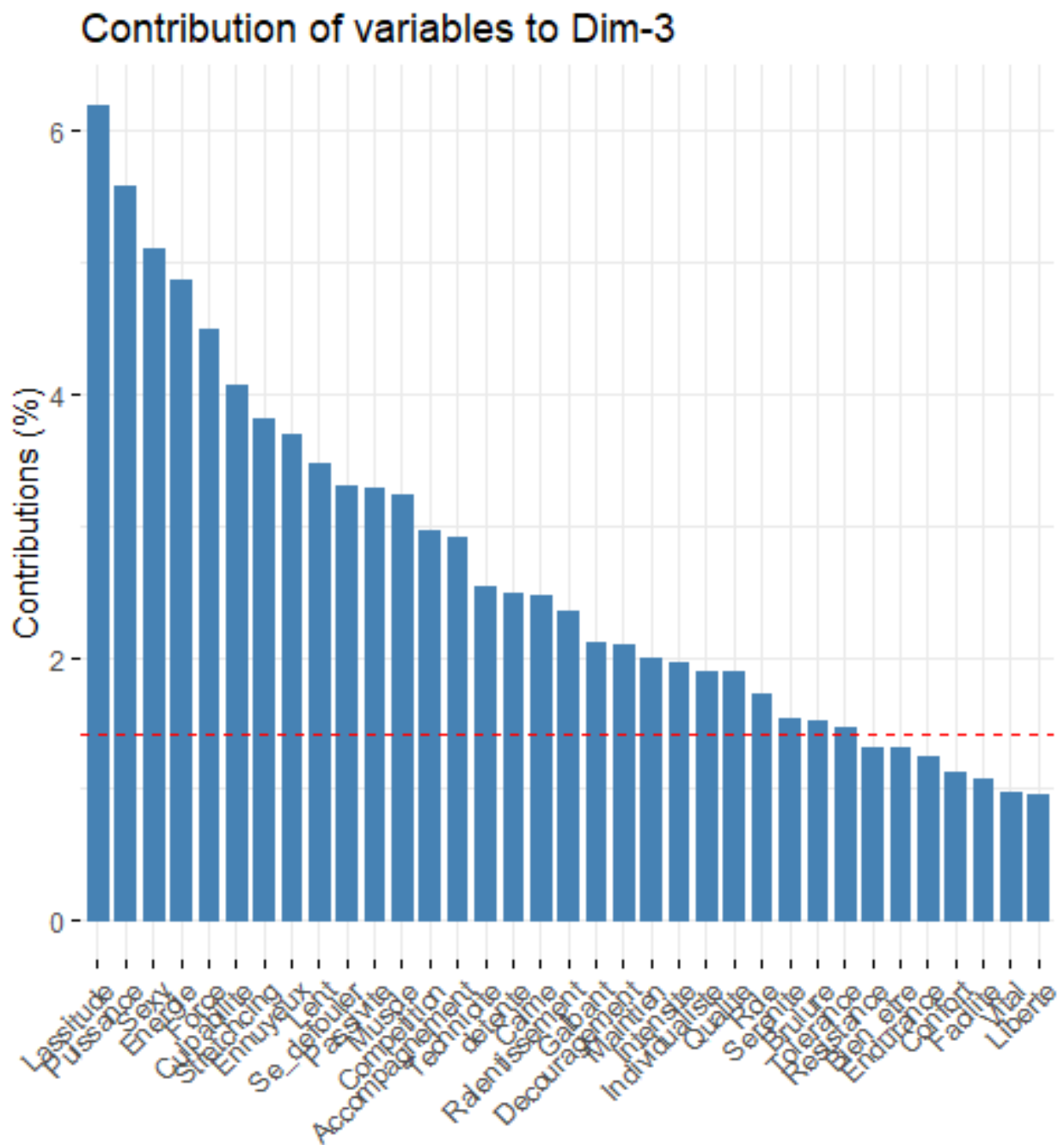


Figure 6:

Contribution of variables to Dim-4

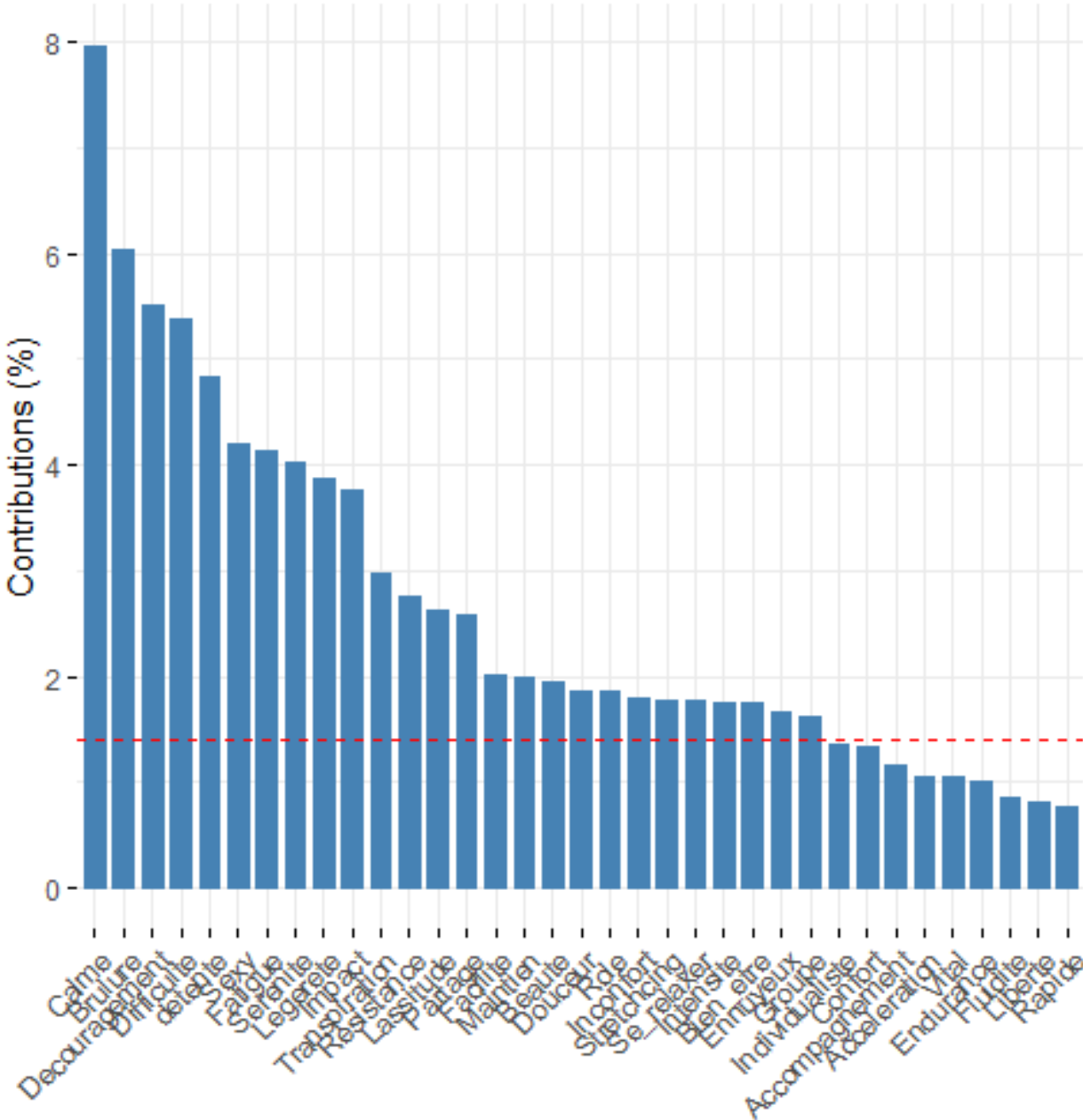


Figure 7:

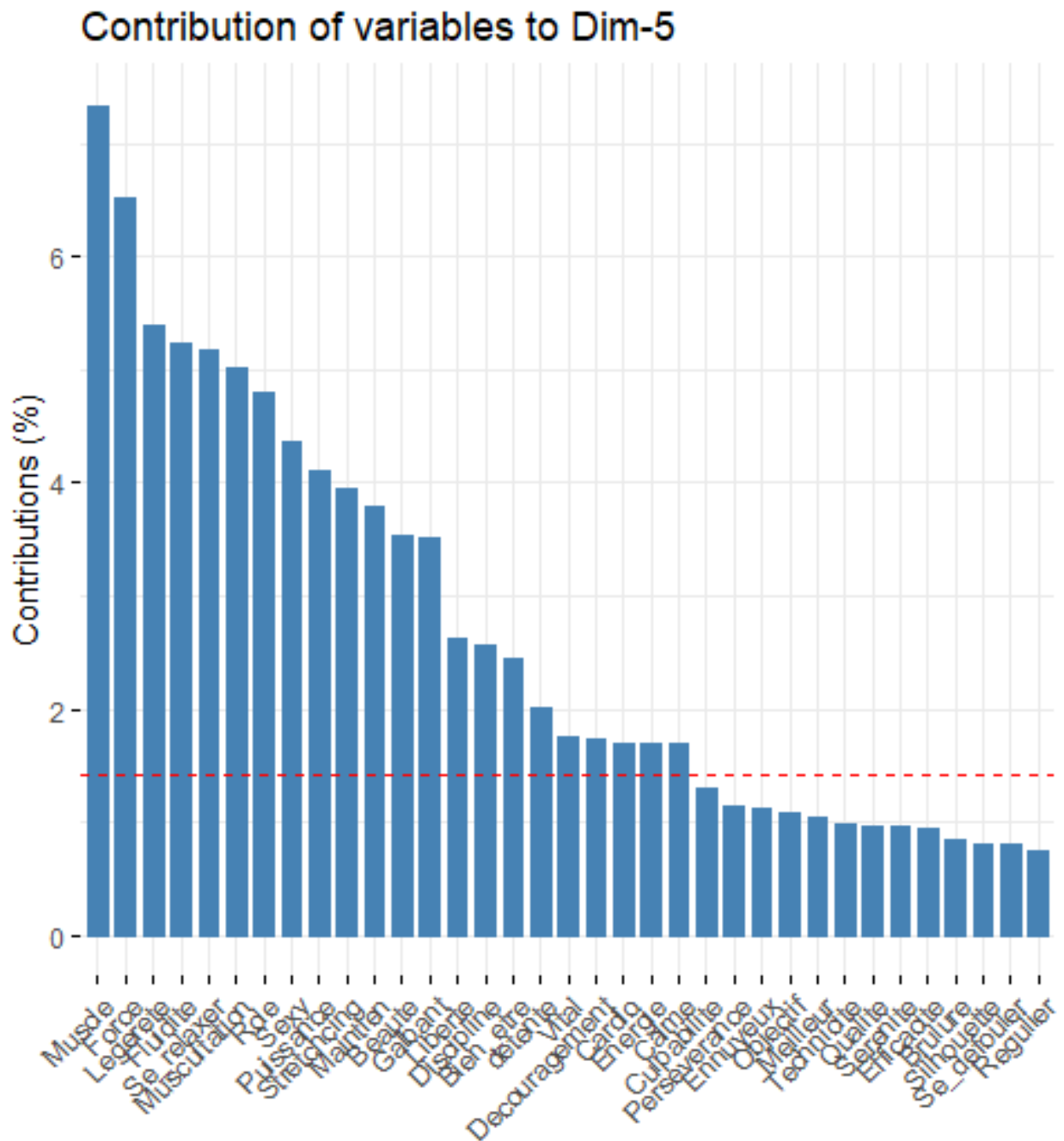


Figure 8:

The least important variables are "Recuperation" and "Facilite".

4.2 Hierarchical Clustering on Principal Components

To perform 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 regrouping and
- Partitioning clustering, in particular the k-means method.

The HCPC method algorithm has 4 main steps:

1. Perform a PCA. Choose the number of dimensions to retain by specifying the argument `ncp`. In our case, the value is 5.
2. Apply the hierarchical classification 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. Perform k-means to improve the initial partitioning obtained in step 3.

Voici les lignes de codes principales dans Rstudio 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))
```

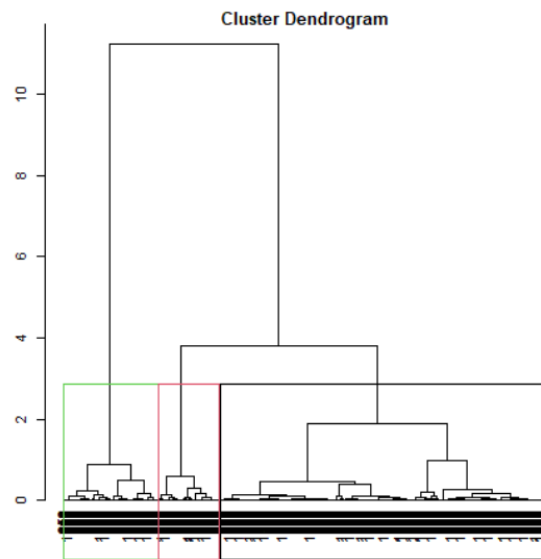


Figure 9: Hierarchical tree

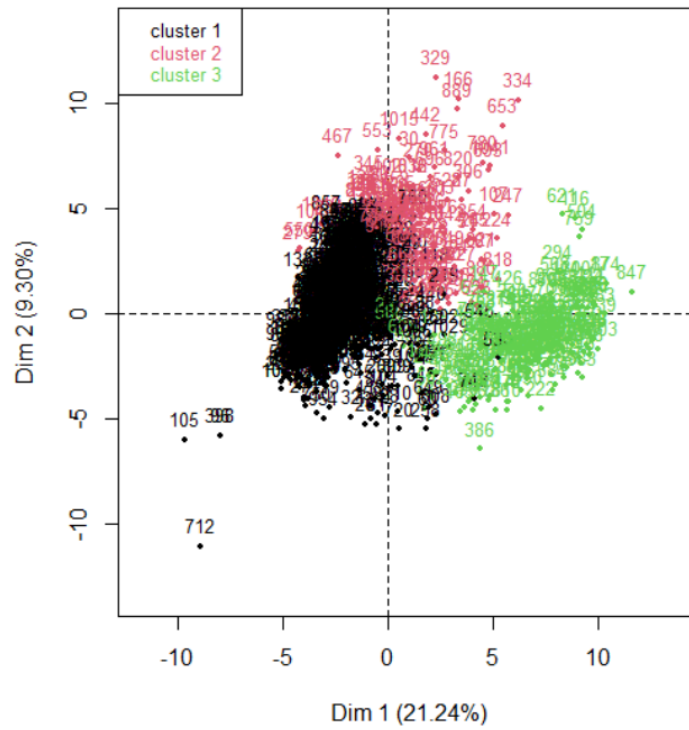


Figure 10: Ascending Hierarchical Classification of the individuals

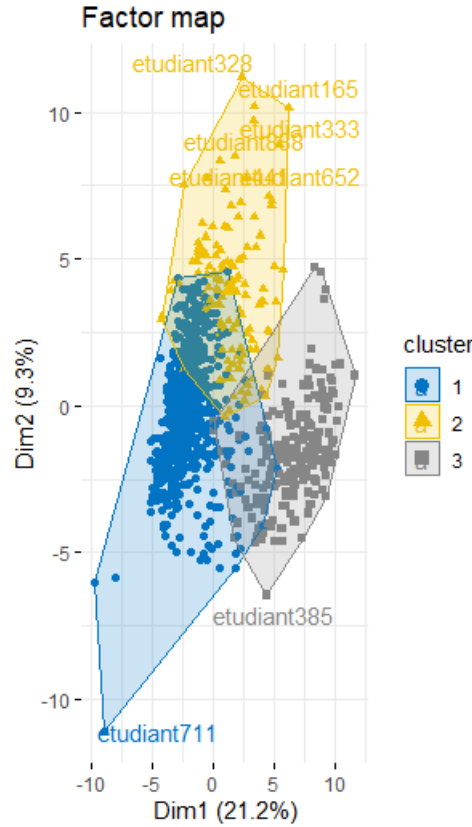


Figure 11: Ascending Hierarchical Classification of the individuals

4.3 Results of HCHC

Cluster 1 consists of individuals sharing:

- high values for the variables Galbant, Culpability, Boredom, Stretching, Security, Discouragement, Slowness, Lassitude, Discomfort, and Passivity (variables are sorted from the highest).
- low values for variables such as Progression, Perspiration, Performance, Active, Challenge, Fun, Goal, Perseverance, Self-confidence and Cardio (the variables are sorted from the lowest).

Cluster 2 consists of individuals who share:

- High values for variables such as Blast Off, Power, Competition, Technicality, Quality, Energy, Comfort, Muscle, Strength, and Intensity (variables are sorted from highest).
- Low values for the variables Sexy, Best, Calm and Vital (the variables are sorted from the lowest).

Cluster 3 consists of individuals who share:

- High values for variables such as Progression, Active, Performance, Challenge, Cardio, Sharing, Fun, Overcoming, Fast, and Efficiency (variables are sorted from highest).
- Low values for variables such as Comfort, Safety, Shaping, Softness, Boring, Strength, Holding, Quality, Beauty and Discomfort (variables are sorted from the lowest).

5 Classification

We separated our data in 2 parts: 80 % of the data for the choice of the selection algorithm, 20 % of the data to test the final model and eventually select the most important columns.

5.1 Choice of Classification algorithm

4 multi-class classifiers are used: the support vector classifier (SVC), linear support vector classifier (LSVC), k-nearest neighbors (KNN) and logistic regression (logreg). (KNN) and logistic regression (logreg).

Here is the graph of the performance of each model:

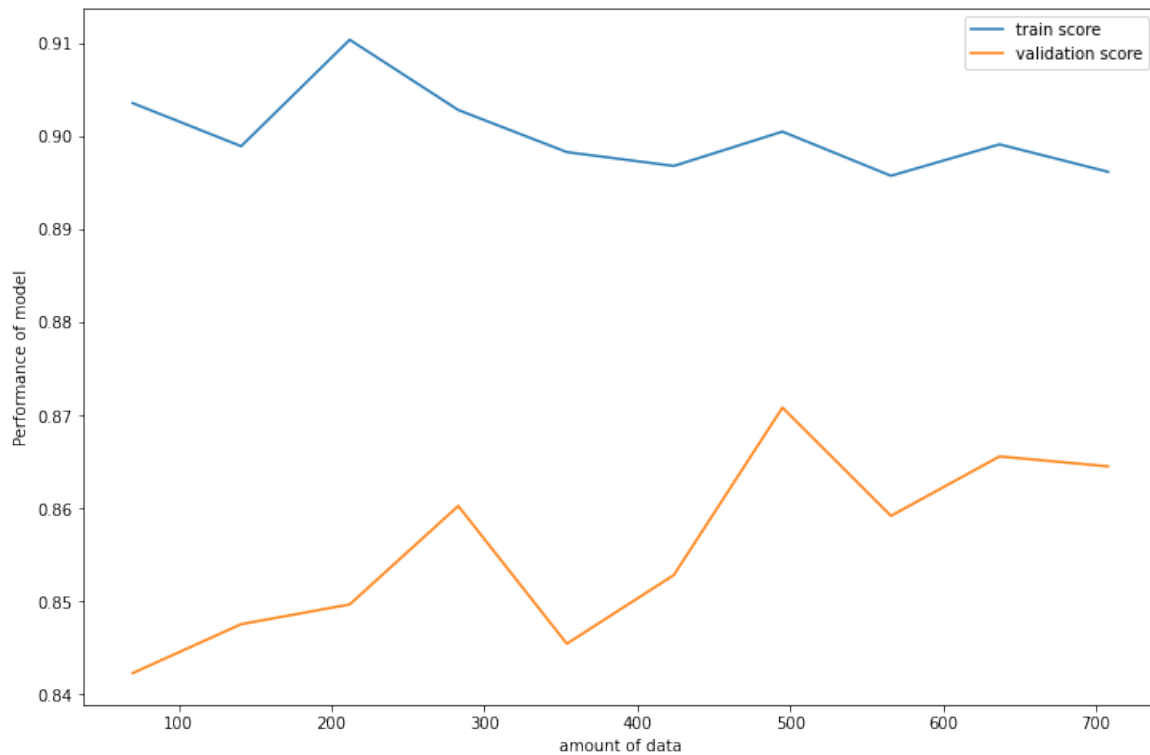


Figure 12: KNN learning curve

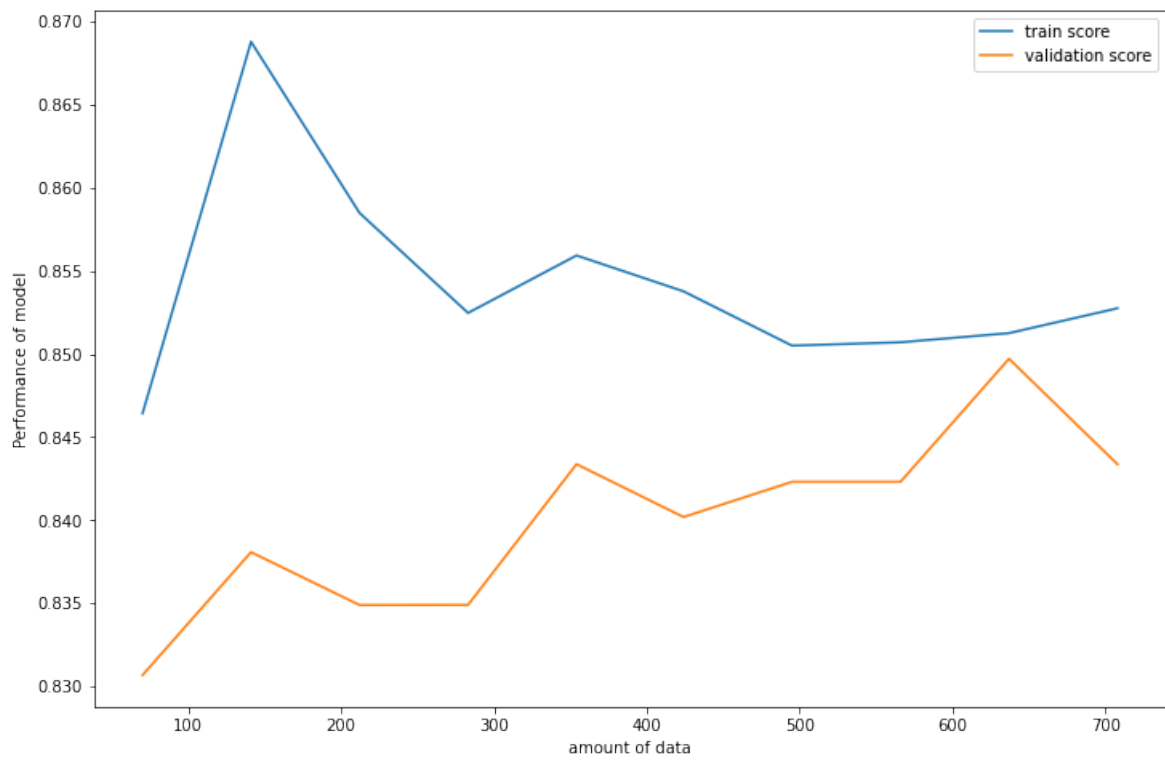


Figure 13: logreg learning curve

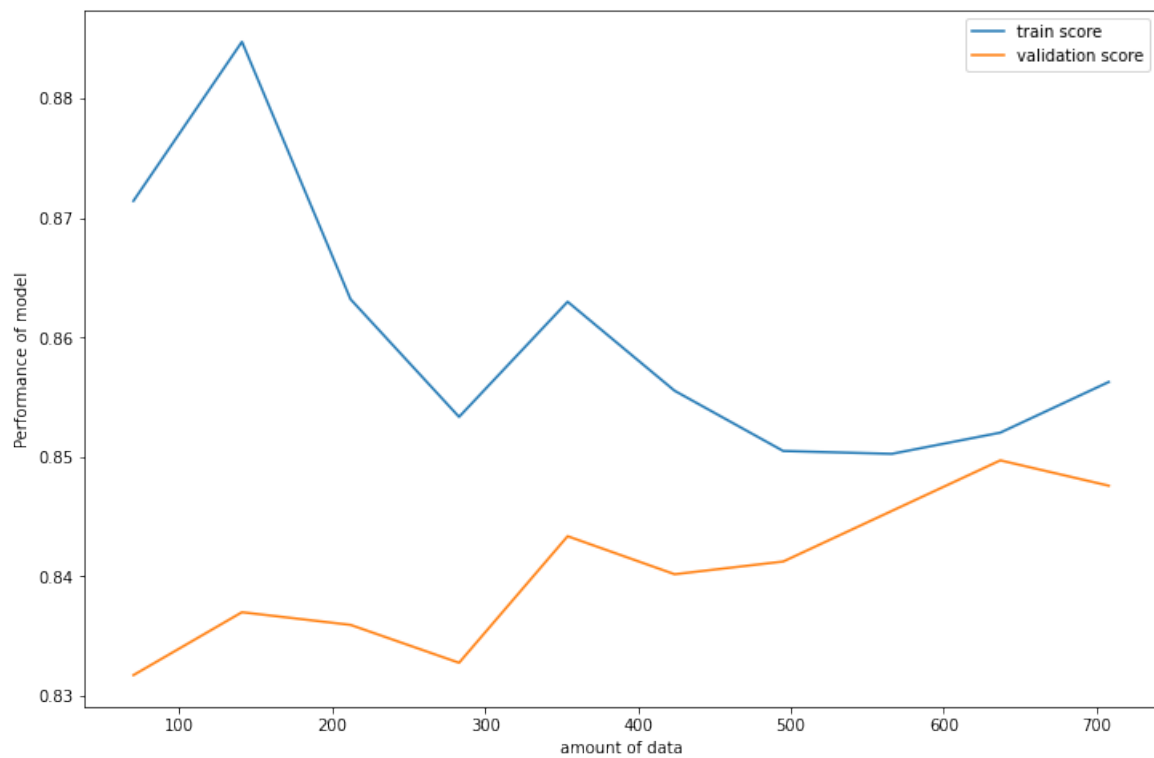


Figure 14: LSVC learning curve

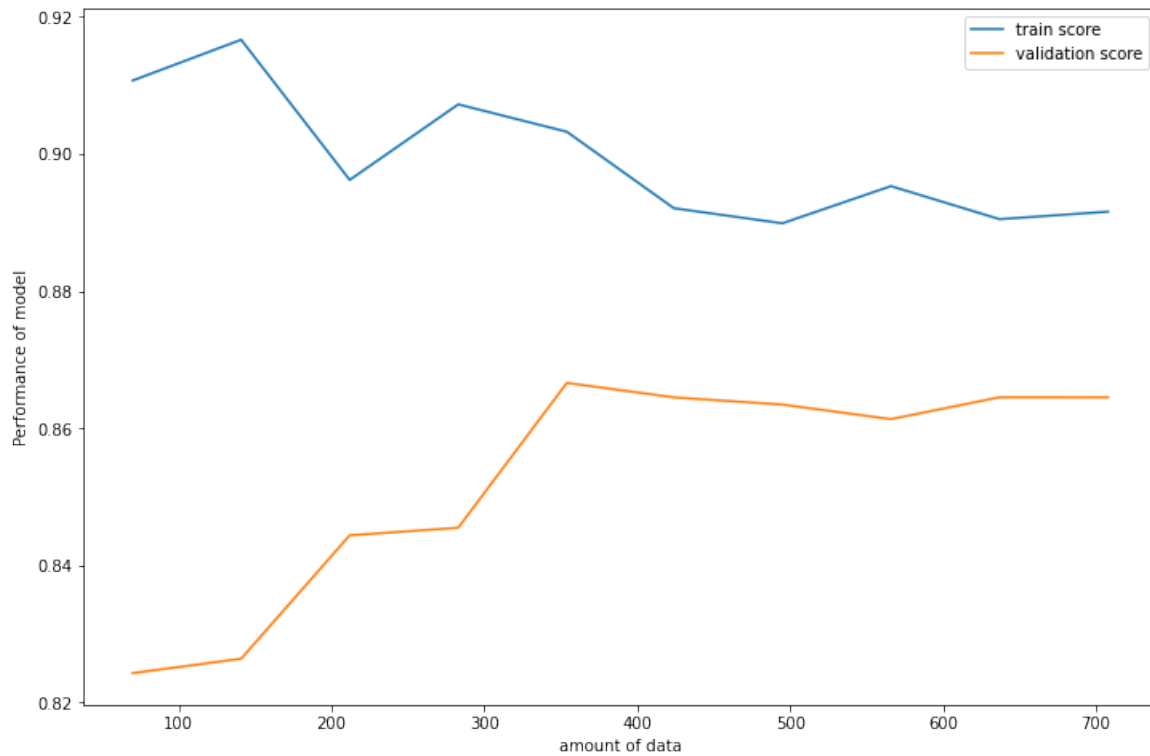


Figure 15: SVC learning curve

According to the graphs below, the performance of the logistic Regression and SVC models is more stable and better than the other models. We can say that the two models are not overfitting (train score and val score are close) contrary to the other two models.

Nous avons utilis GridSearchCV pour optimiser 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.

We used GridSearchCV to optimize the hyperparameters of the logistic Regression model. GridSearchCV allows us to select the best hyperparameters by comparing the different performances of each of each combination using the cross-validation technique.

5.2 Feature selection

The graph below shows the variance of each feature. 4 threshold candidates stand out: 0.8 , 0.06 , 0.04 and 0.02.

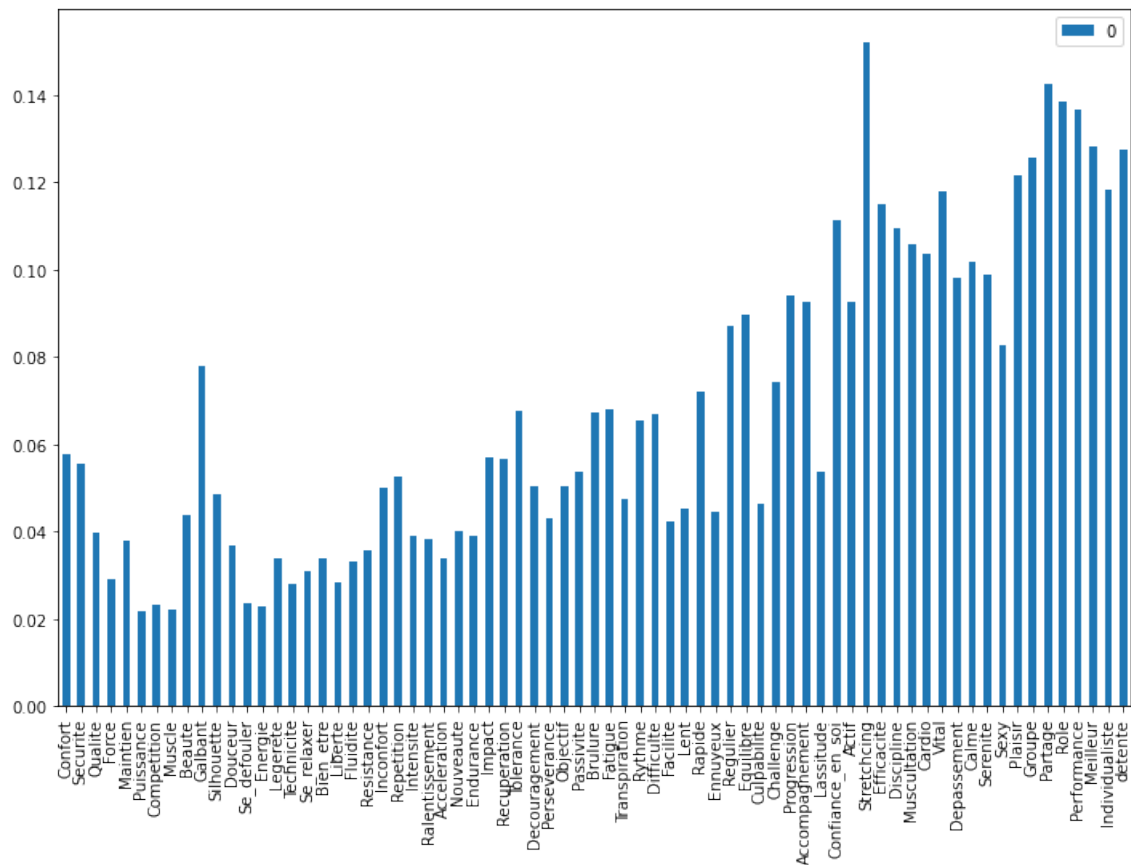


Figure 16: variance of each features

To eliminate the values below this threshold, the VarianceThreshold function of scikit-learn is used.

In the end, the threshold set at 0.02 gives better results. No column was suppressed.

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

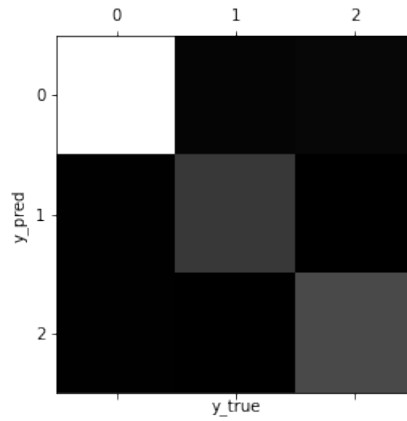


Figure 18: confusion matrix

We obtain satisfactory results: only 9 misplaced values, the value of 0.964 and 0.94768 for the recall_score and the f1score.

6 Conclusion

The main objective of the project was to do clustering on our data. Thanks to the HCPC clustering, we can distinguish 3 types of students:

- The unmotivated, those who seek well-being and simplicity, whose characteristic variables of the cluster are: Guilt, Boredom, Discouragement, Slowness, Discomfort
- In the second group, we have those who like combat sports such as wrestling, boxing and MMA. It is characterized by its words: Power, Competition, Technique, Quality, Energy, Muscle, Force and Intensity
- The last group is characterized by its words: Progression, Performance, Challenge, Cardio, Sharing, Exceeding, Fast and Efficiency. We find those who enjoy running and nature activities.

Regarding classification, one of the best algorithms (SVM) gave the value of 0.9549 of precision, 0.9 of recall and 0.9255 of f1-score.

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>