Report final

Congo Job

June 2, 2022

Contents

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 primordia 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 and 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).

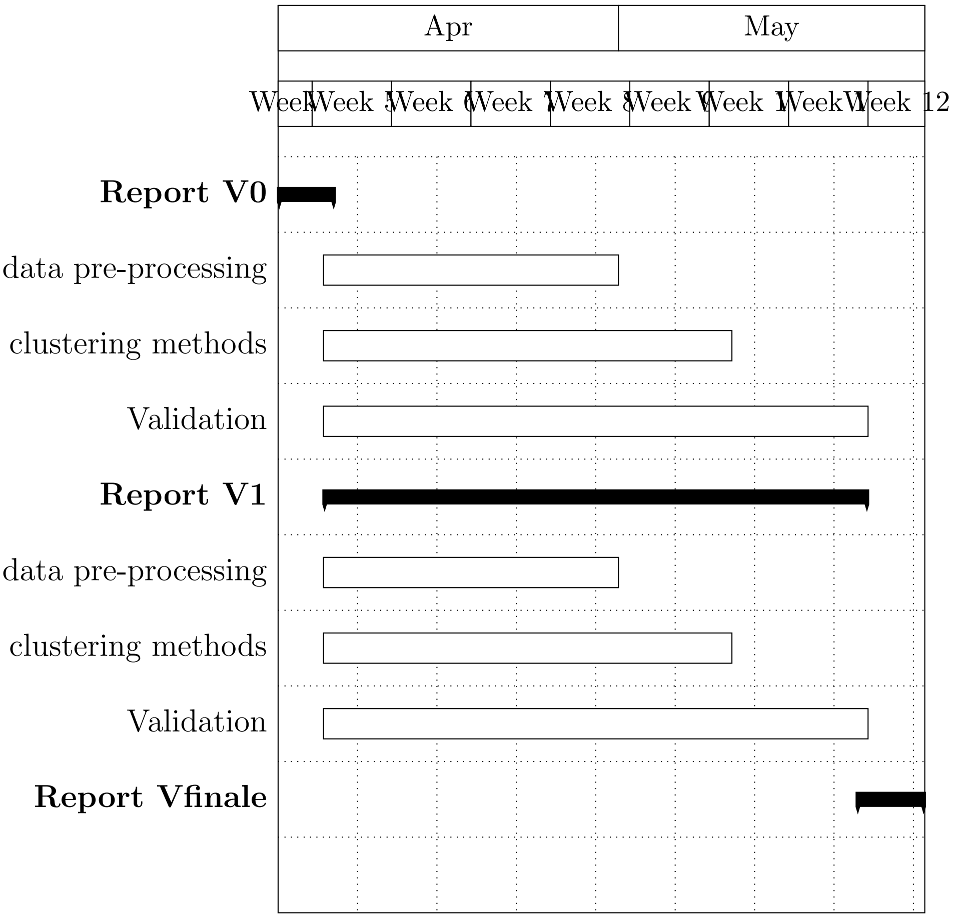
1.2 Main Objectives

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.



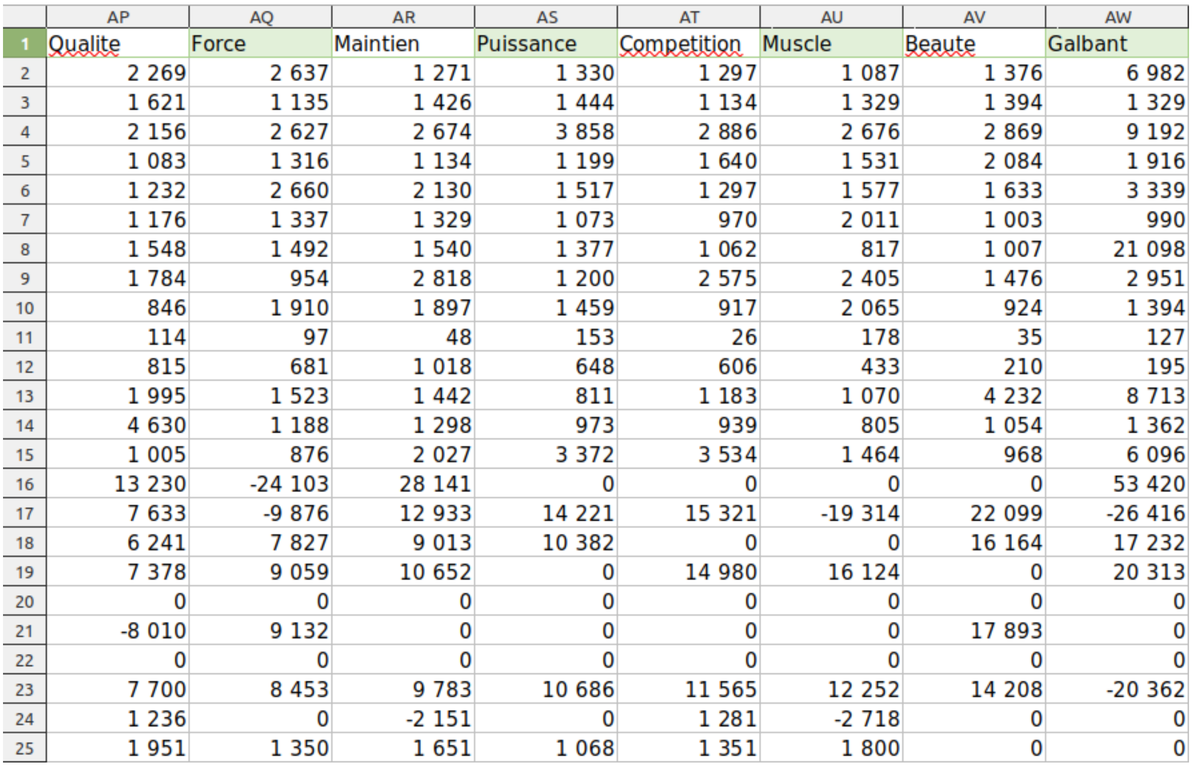
2 Description of raw data

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	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

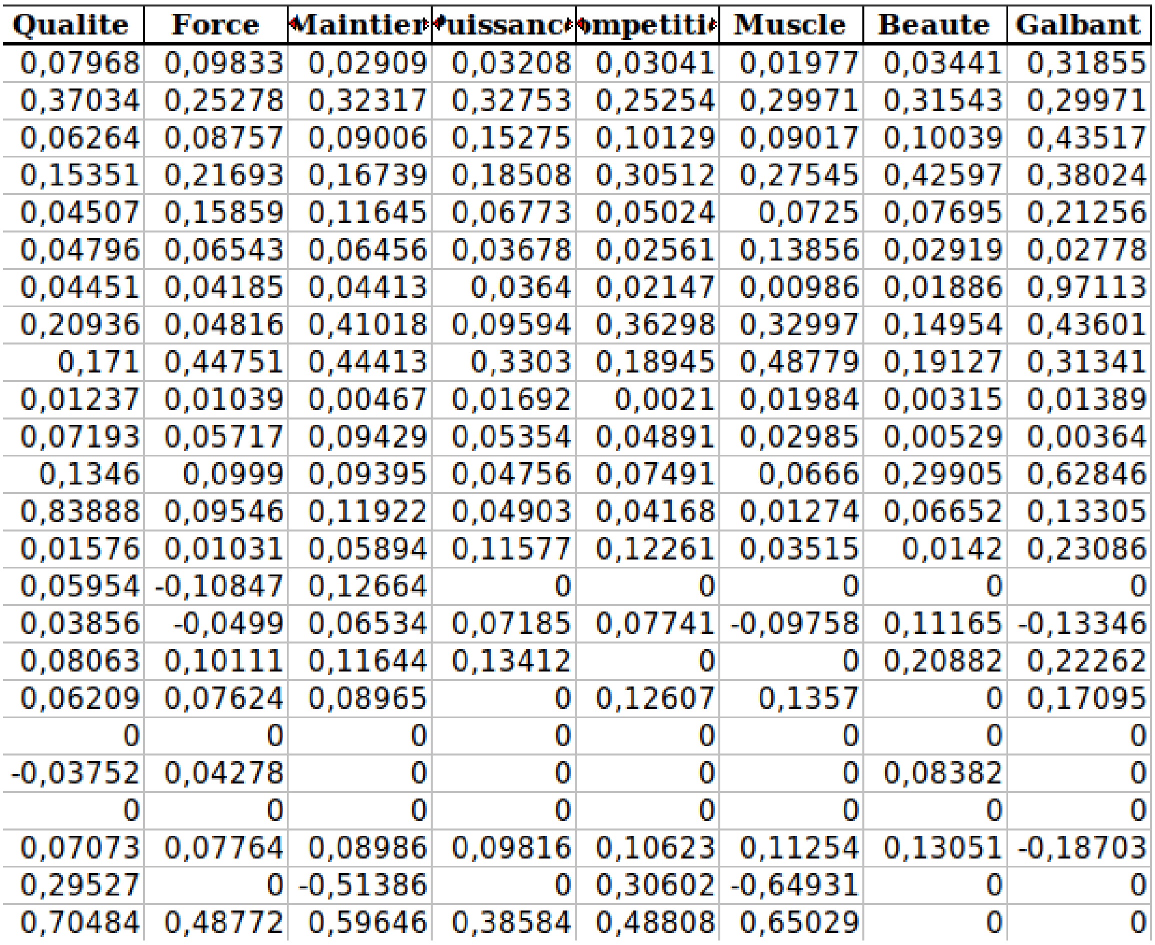
Figure 1: Raw data

3 Preprocessing

3.1 Preprocessing Method

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. Values less than 200 ms are set to 0 assuming there is a minimum response time for each question. 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



Qualite	Force	Maintier	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

Figure 2: Data processed

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 1048 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.

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

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.

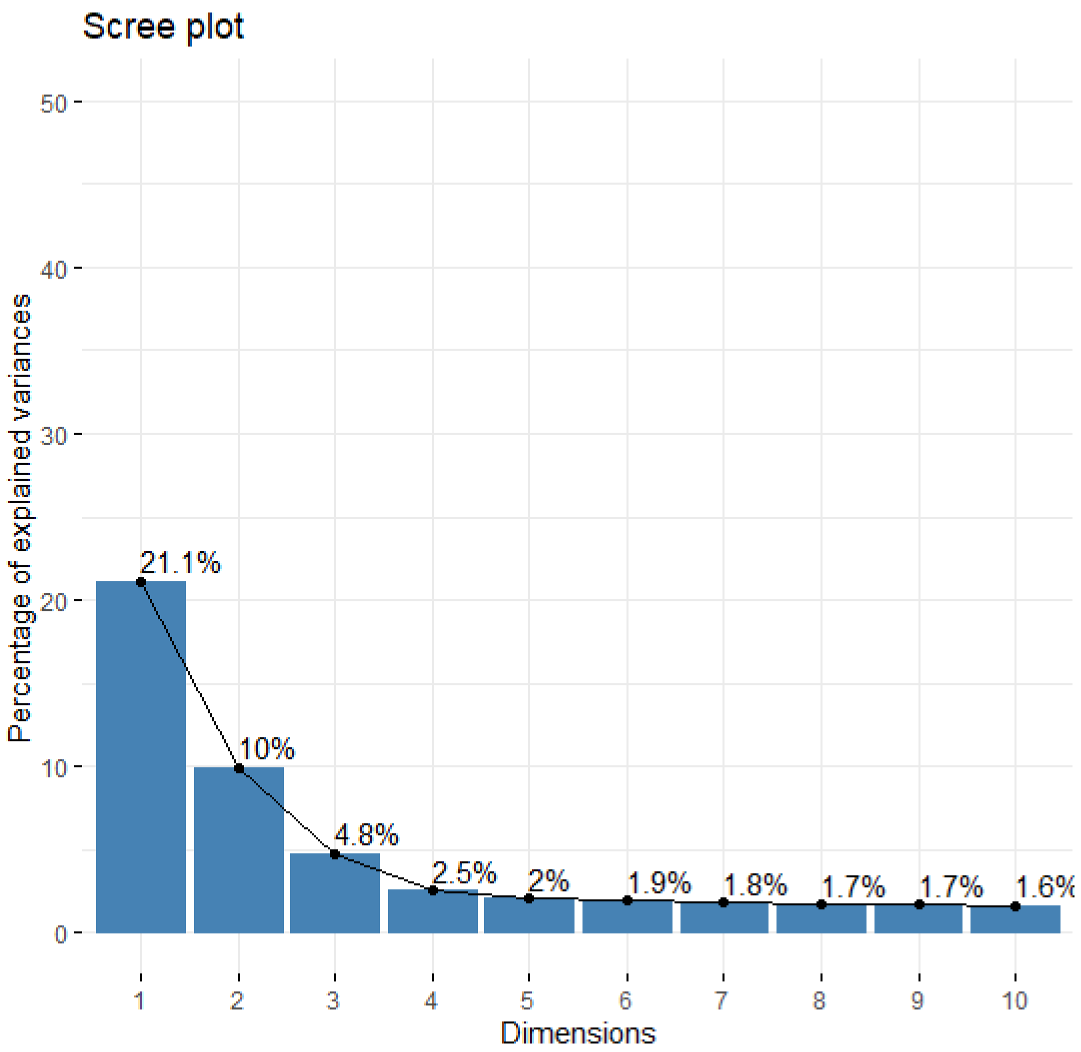


Figure 3: Percentage of variance explained by dimensions number

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 35 variables that contribute the most to the 5 principal components. The red dotted lines on the graphs indicate the average contribution value.

```
# Contributions of variables to PC1,PC2,PC3,PC4,PC5 top = 35
fviz_contrib(res.pca, choice = "var", axes = 1, top = 35)
fviz_contrib(res.pca, choice = "var", axes = 2, top = 35)
fviz_contrib(res.pca, choice = "var", axes = 3, top = 35)
fviz_contrib(res.pca, choice = "var", axes = 4, top = 35)
fviz_contrib(res.pca, choice = "var", axes = 5, top = 35)
```

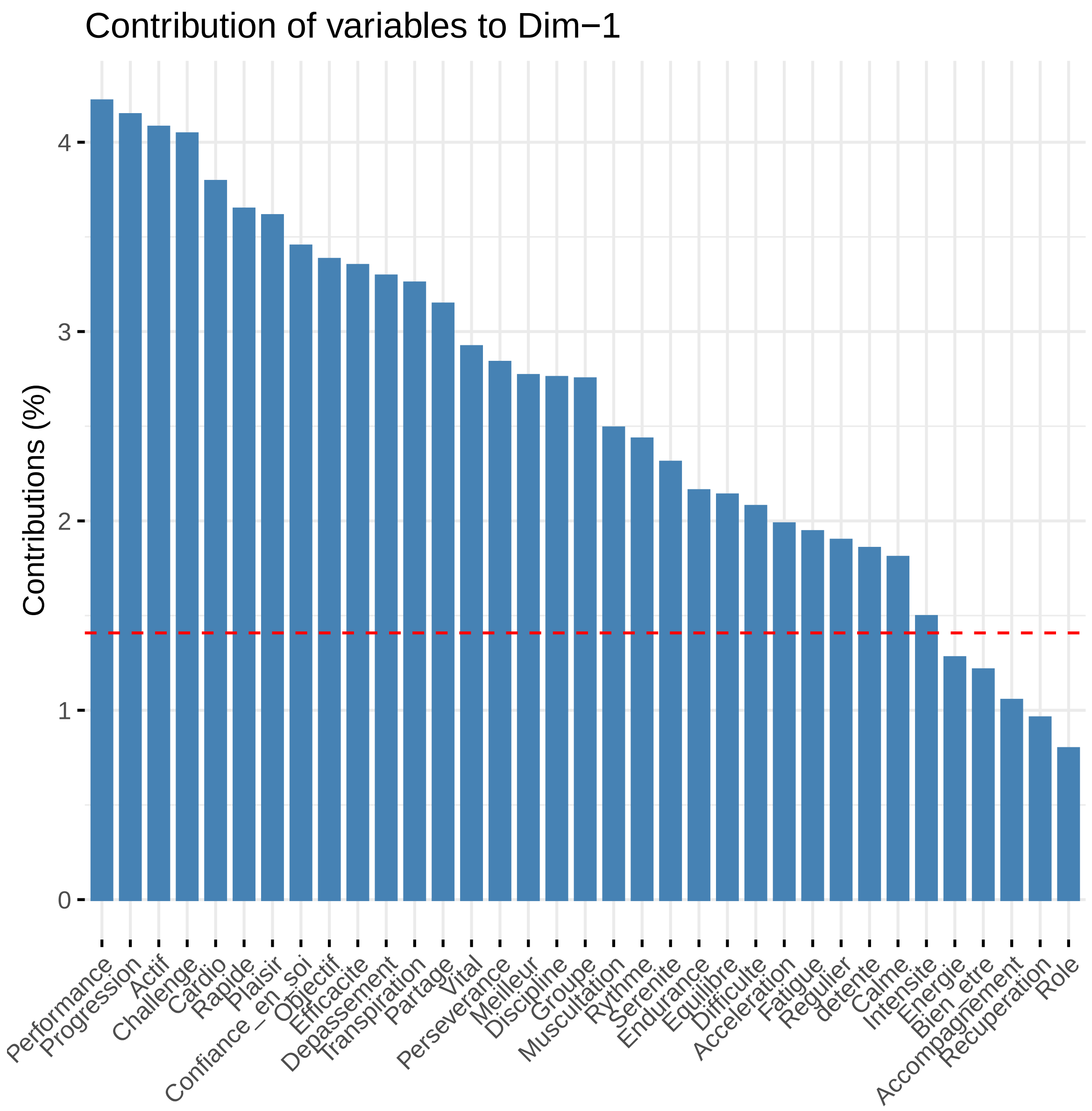


Figure 4: Variable contributions to dimension 1

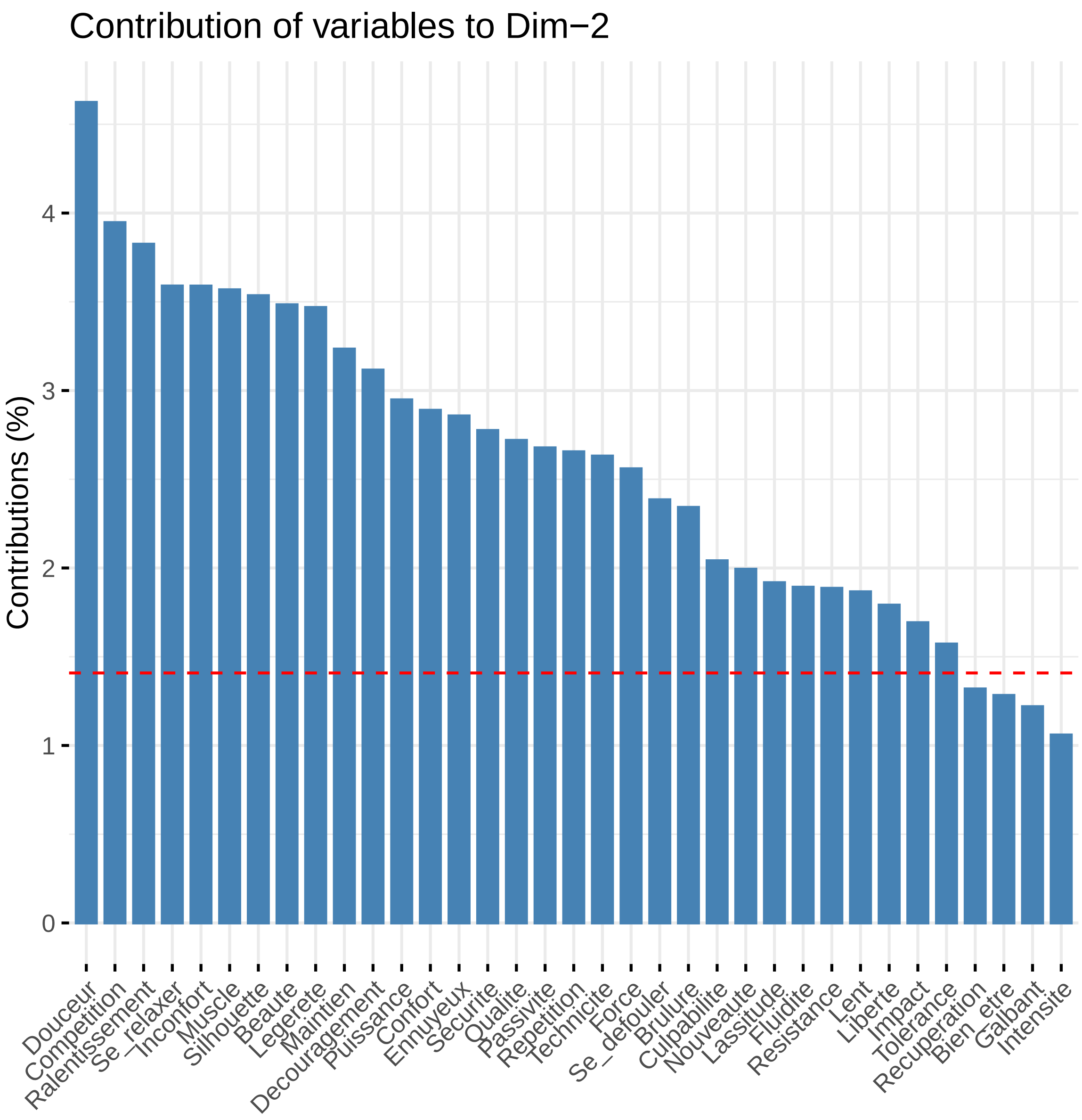


Figure 5: Variable contributions to dimension 2

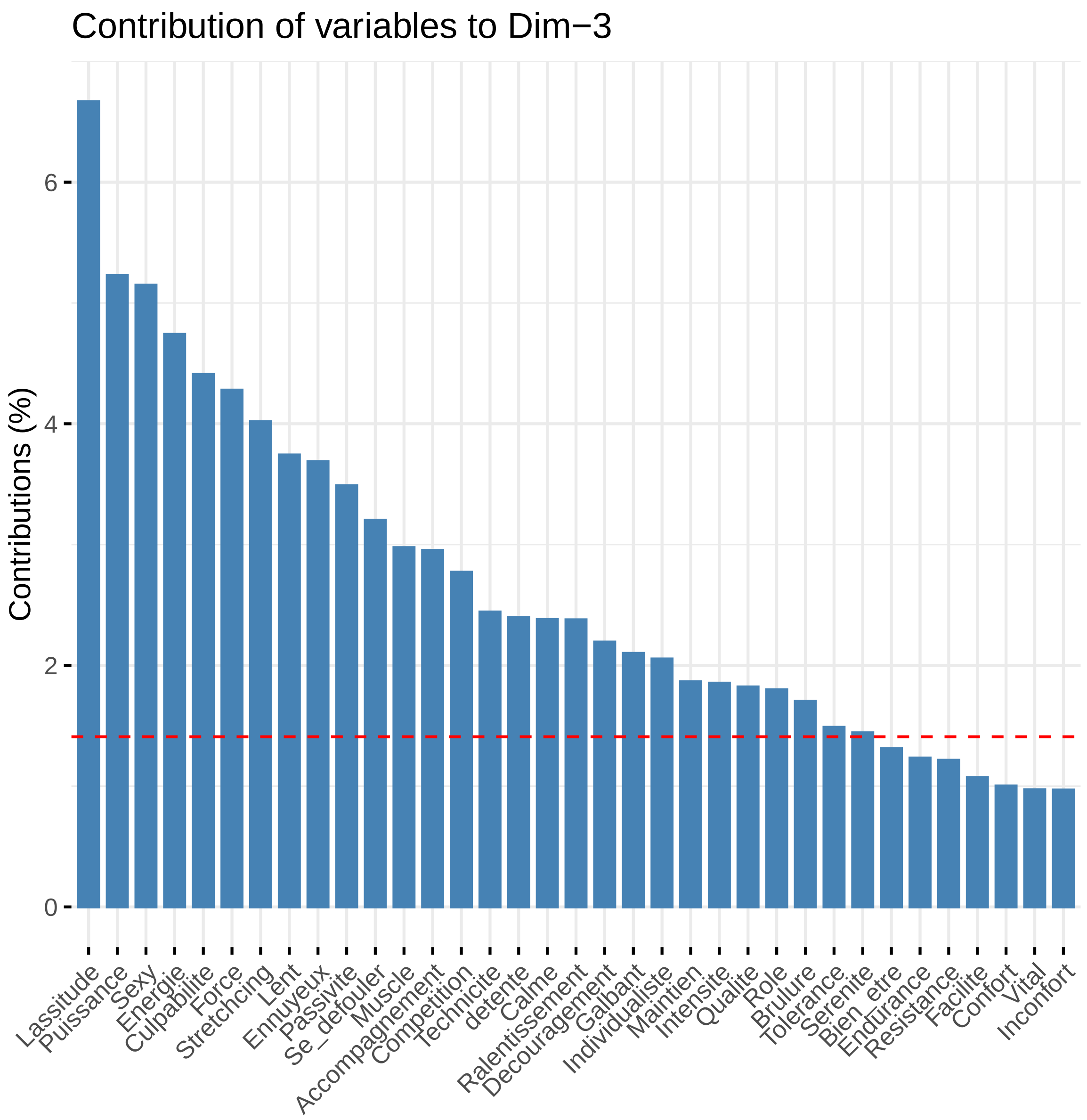


Figure 6: Variable contributions to dimension 3

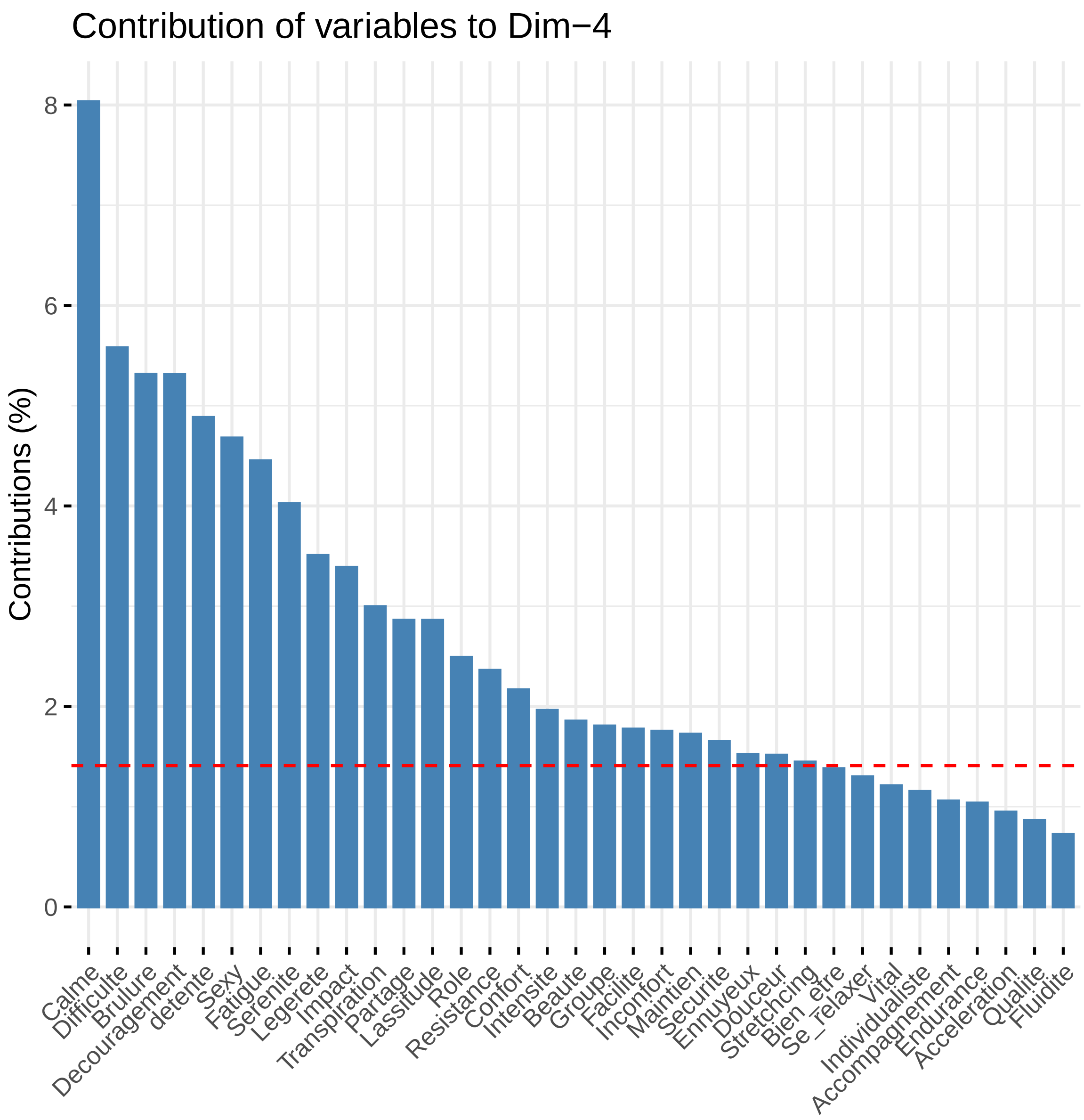


Figure 7: Variable contributions to dimension 4

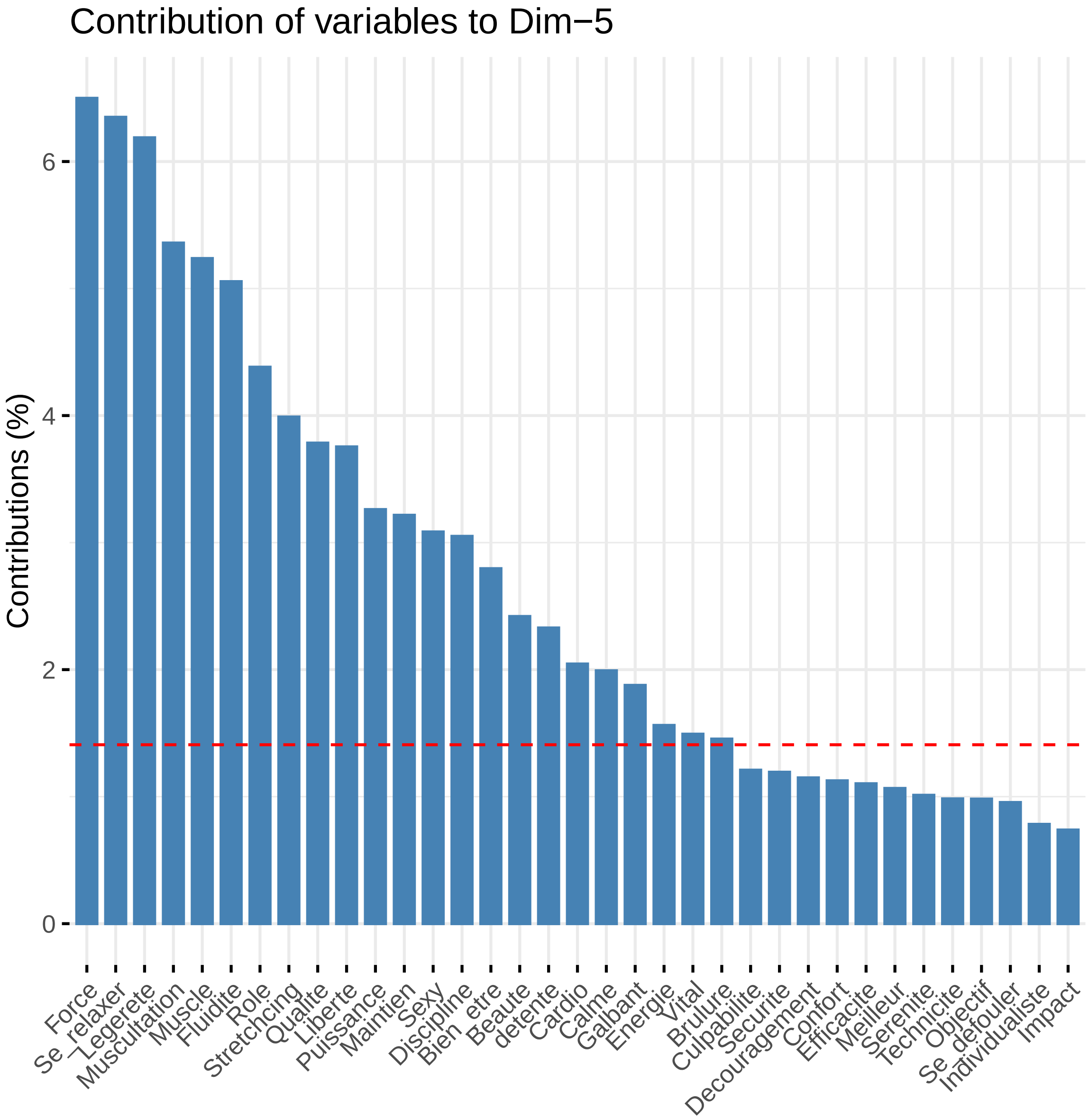


Figure 8: Variable contributions to dimension 5

By combining the most important variables in our different axes, we keep 69 variables. The least important variables are "Recuperation" and "Facility". We can say that all the qualifiers are important to describe our dataset.

4.2 Hierarchical Clustering on Principal Components (HCPC)

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.

4.3 Implementation of HCPC

We start by computing the principal component analysis (PCA). The argument ncp = 5 is used in the PCA() function to keep only the first five principal components. Then, the HCPC is applied on the PCA result.

Here are the main lines of code in Rstudio for the implementation:

```
# Compute PCA with ncp = 5
res.pca <- PCA(data_base , ncp = 5 ,graph = TRUE)
# Compute hierarchical clustering on principal components
res.hcpc <- HCPC(res.pca,nb.clust=3,consol=FALSE,graph=TRUE)
```

To visualize the dendrogram generated by the hierarchical clustering, we will use the fviz_dend() function

```
fviz_dend(res.hcpc,
          cex = 0.5,
          palette = "jco",
          rect = TRUE, rect_fill = TRUE,
          rect_border = "jco",
          labels_track_height = 0.5 )
```

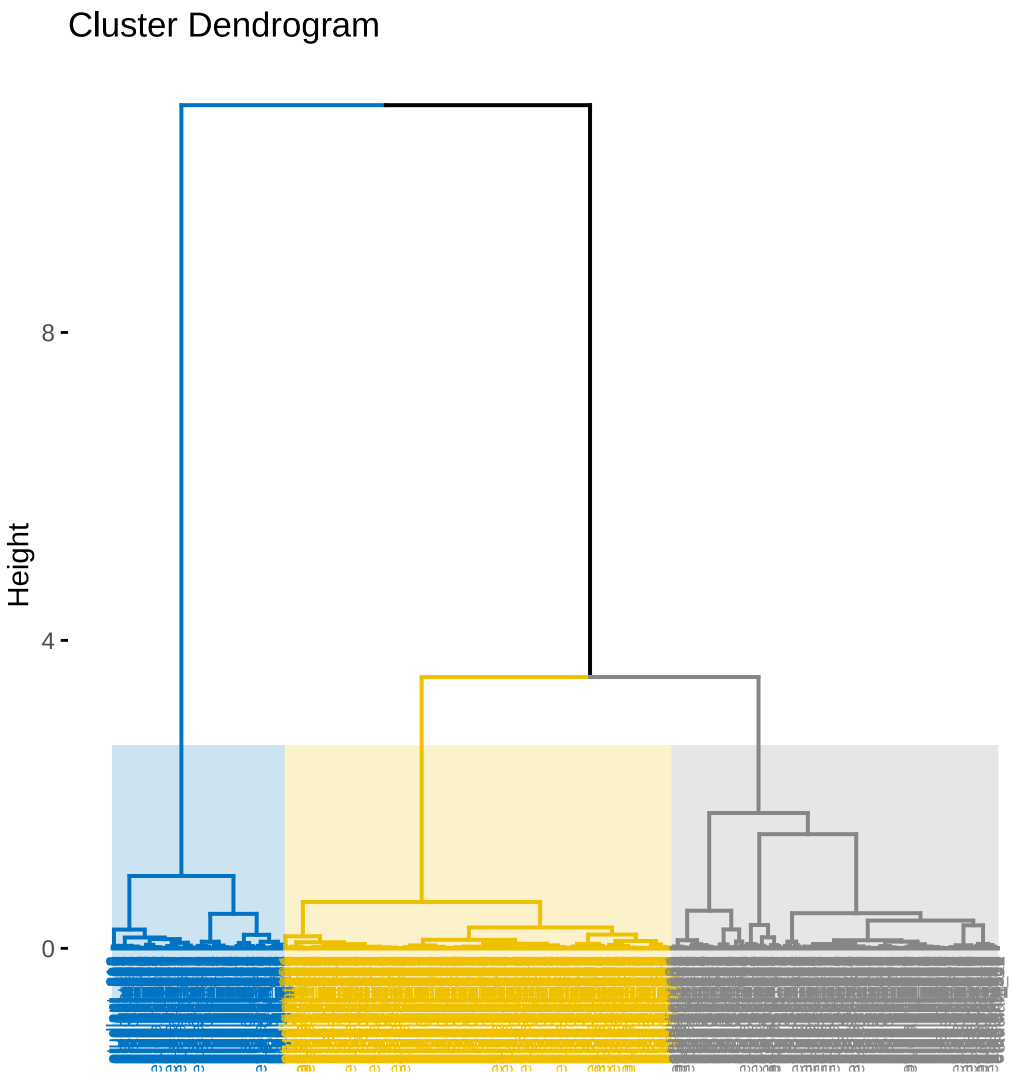


Figure 9: Dendogram

The dendrogram suggests a 3-cluster solution.However, the size ratio between each cluster disproportionate.

It is possible to visualize the individuals on the principal components map and to color the individuals according to the cluster they belong to. The function fviz_cluster()[in factoextra] can be used to visualize individual clusters.

```

fviz_cluster(res.hcpc,repel = TRUE,
             show.clust.cent = TRUE,
             palette = "jco",
             ggtheme = theme_minimal(),
             main = "Factor map")
```

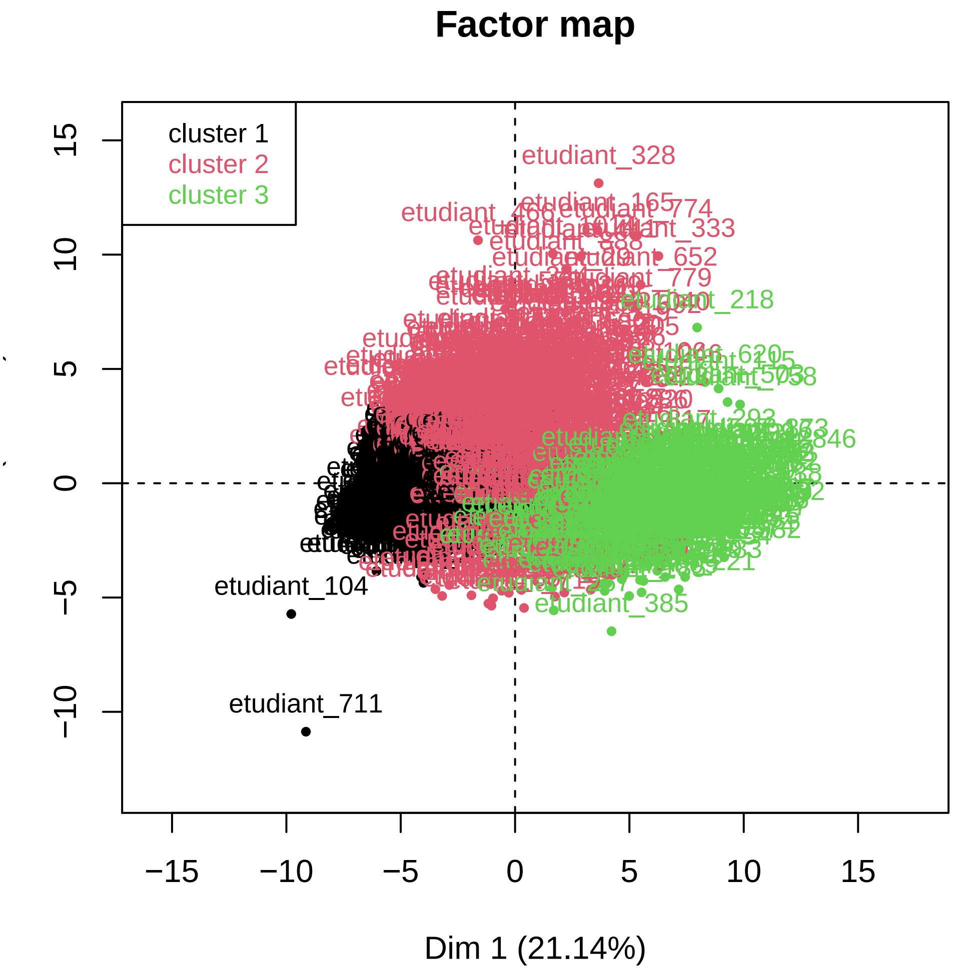


Figure 10: Visualisation of individuals cluster

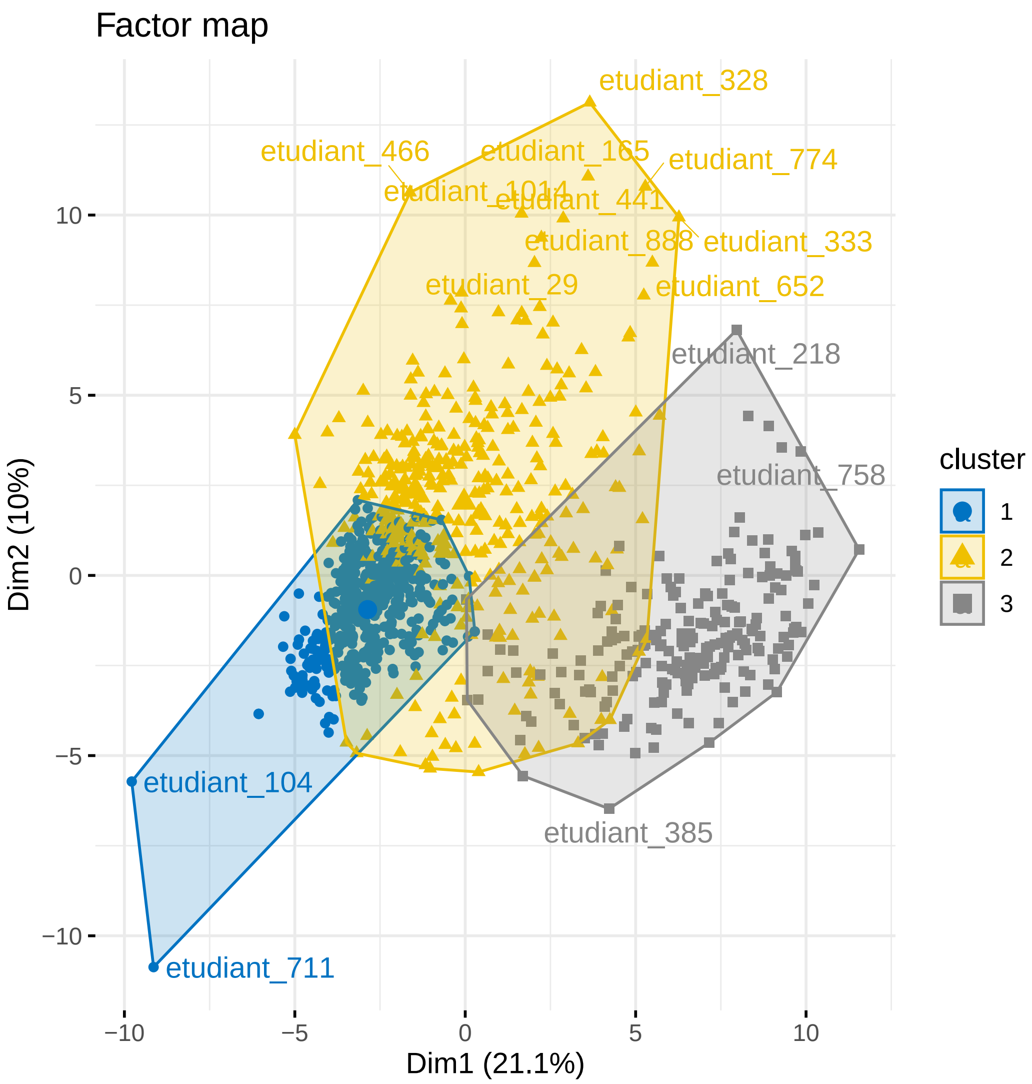


Figure 11: Visualisation of individuals cluster

4.4 Results of HCPC

Le nombre d'tudiants dans chaque cluster est 547, 387 et 204 .
The cluster 1 is made of individuals sharing :

- high values for the variables Galbant
- low values for variables like Progression, Objectif, Challenge,Transpiration, Actif, Perseverance, Confiance_en_soi, Performance, Endurance and Plaisir (variables are sorted from the weakest).

The cluster 2 is made of individuals sharing :

- high values for variables like Puissance, Competition, Se_defouler, Muscle, Se_relaxer, Technicite, Force, Legerete, Qualite and Resistance (variables are sorted from the strongest).
- low values for variables like Meilleur, Partage, Vital, Depassement, Groupe, Cardio, Serenite, Efficacite, Plaisir and Actif (variables are sorted from the weakest).

The cluster 3 is made of individuals sharing :

- high values for variables like Actif, Progression, Performance, Cardio, Partage, Plaisir, Depassement, Challenge, Rapide and Efficacite (variables are sorted from the strongest).
- low values for variables like Confort, Securite, Galbant, Douceur, Force, Ennuyeux, Inconfort, Qualite, Maintien and Ralentissement (variables are sorted from the weakest).

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 select the most important columns and test the final model.

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).

We used to evaluate our models and then optimize them using the functions below. It uses the learning_curve and GridSearchCV() functions from scikit_learn. Learning_curve() determine cross-validated training and test scores for different training set sizes. GridSearchCV allows us to select the best hyperparameters by comparing the different performances of each of each combination using the cross-validation technique.

```
# Procedure d'evalution des modeles
def evaluation(model,X_train_3 ,y_train_3 ,X_test_3 ,y_test_3):
    model.fit(X_train_3,y_train_3)
    y_pred_3 = model.predict(X_test_3)
    # print(confusion_matrix(y_test_3 , y_pred_3))
    # print(classification_report(y_test_3 , y_pred_3))

    N, train_score ,val_score = learning_curve(model, X_train_3,y_train_3,
                                  train_sizes = np.linspace(0.1,1.0,10),cv=5)

    plt.figure(figsize =(12,8))
    plt.plot(N,train_score.mean(axis = 1), label ='train score')
    plt.plot(N,val_score.mean(axis = 1), label ='validation score')
    plt.xlabel('amount of data')
    plt.ylabel('Performance of model')
    plt.legend()


# Optimisation des hyperparametres du modele
def optimiseur(model,parameters,X_train_3,y_train_3):
    grid = GridSearchCV(model,parameters)
    grid.fit(X_train_3,y_train_3)

    print("best parameters ",grid.best_params_)
    print("accuracy :",grid.best_score_)
```

Here is the graph of the performance of each model:

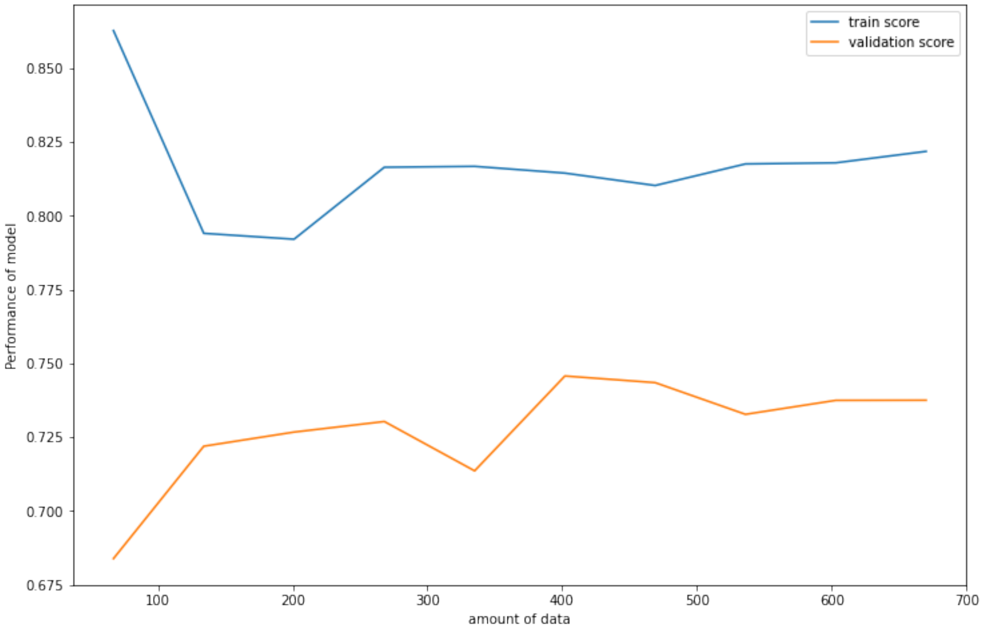


Figure 12: KNN learning curve

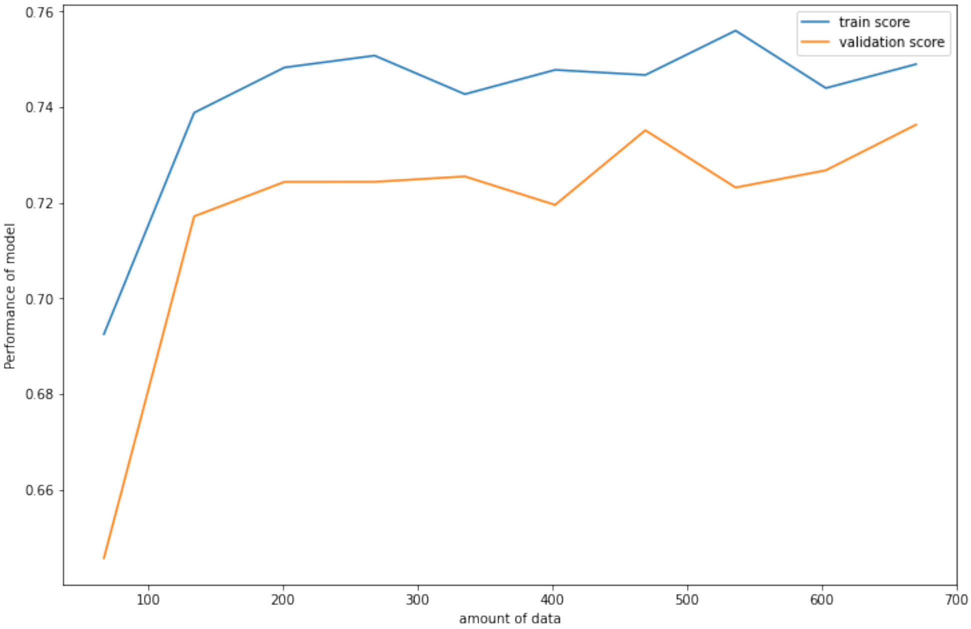


Figure 13: logreg learning curve

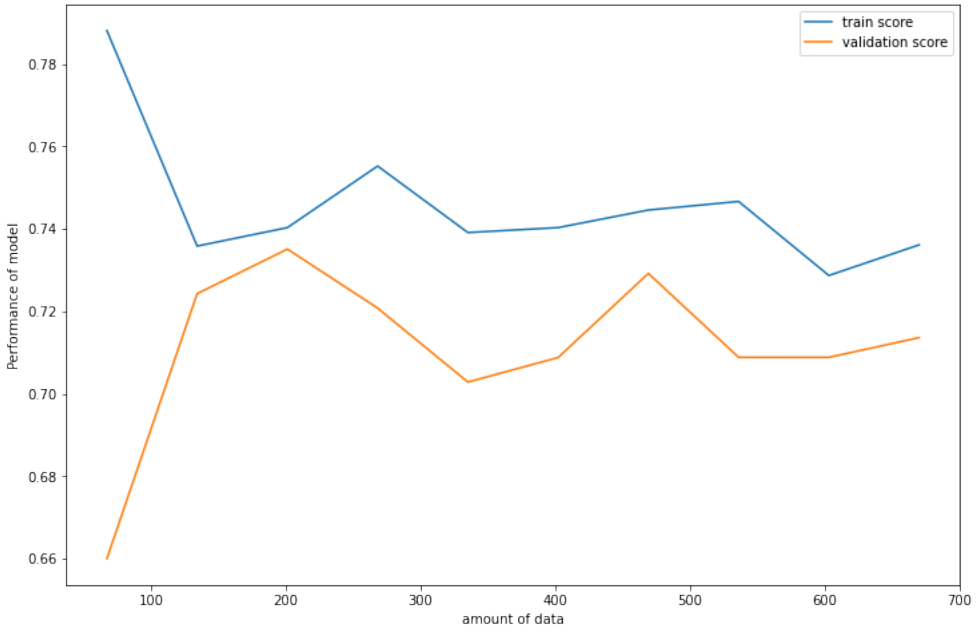


Figure 14: LSVC learning curve

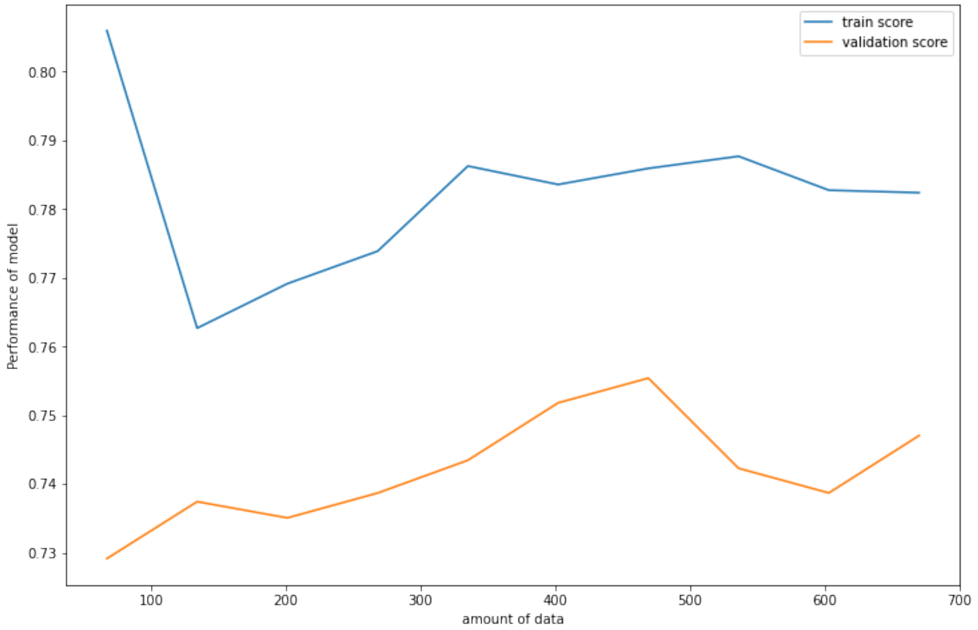
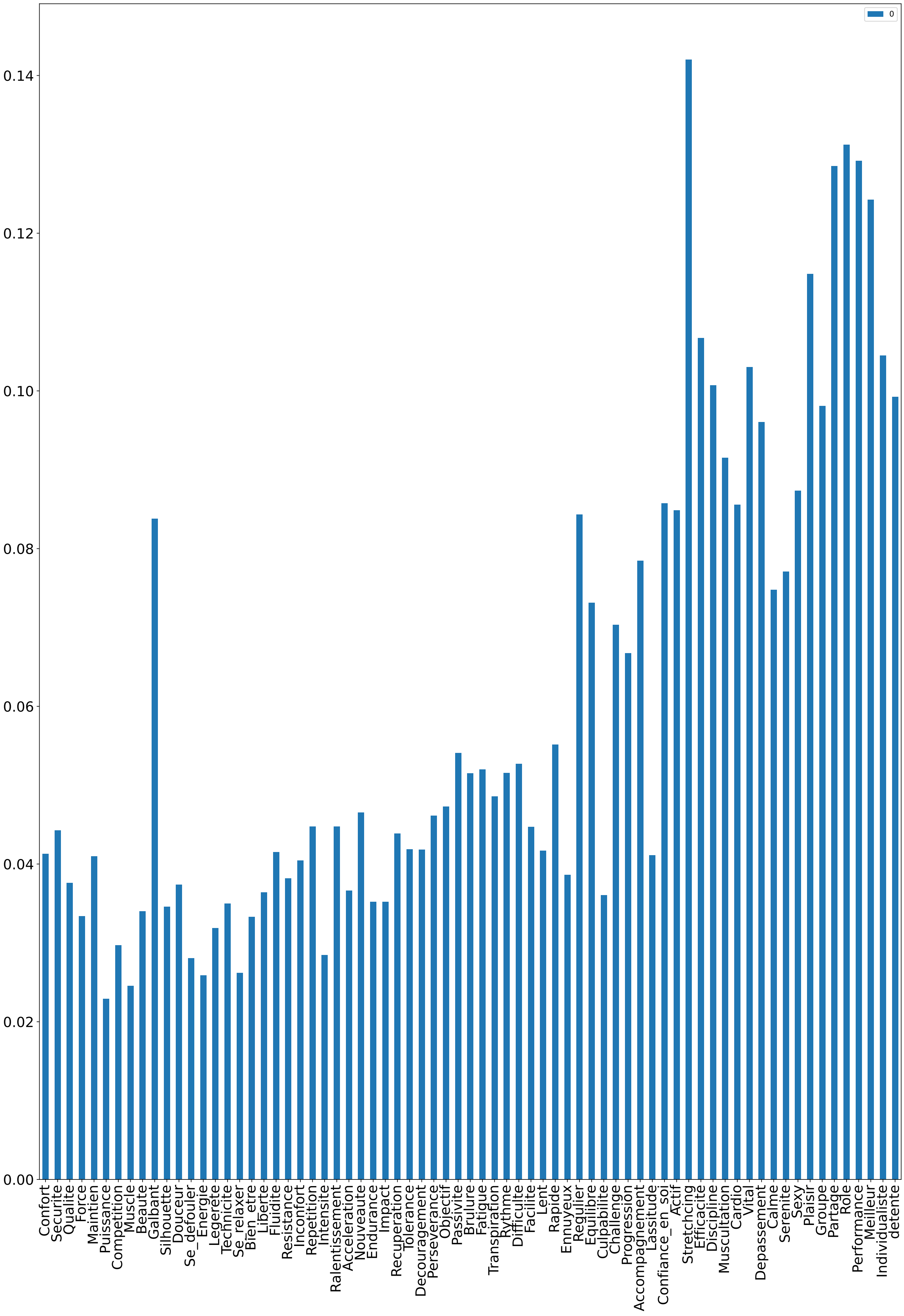


Figure 15: SVC learning curve

The performance of our 4 models is quite good, no overfiting for the 4 models. The difference between the train score and validation is quite small,it is approximately between 0.1 and 0.02 using 600 data. We selected SVC because it has the best score.

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. Stretchcing , Partage ,Role ,Performance ,Meilleur font partie des variables ayant une grande variance .


0
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
Confort
Securite
Qualite
Force
Maintien
Puissance
Competition
Muscle
Beaute
Galbant
Silhouette
Douceur
Se_defouler
Energie
Legerete
Technicite
Se_relaxer
Bien_etre
Liberte
Fluidite
Resistance
Inconfort
Repetition
Intensite
Ralentissement
Acceleration
Nouveaute
Endurance
Impact
Recuperation
Tolerance
Decouragement
Perseverance
Objectif
Passivite
Brulure
Fatigue
Transpiration
Rythme
Difficulte
Facilite
Lent
Rapide
Ennuyeux
Regulier
Equilibre
Culpabilite
Challenge
Progression
Accompagnement
Lassitude
Confiance_en_soi
Actif
Stretchcing
Efficacite
Discipline
Musculation
Cardio
Vital
Depassement
Calme
Serenite
Sexy
Plaisir
Groupe
Partage
Role
Performance
Meilleur
Individualiste
detente

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 class 1	Predicted class 2	Predicted class 3
Actual class 1	79	4	0
Actual class 2	18	67	3
Actual class 3	0	1	38

Table 1: Confusion matrix

We obtain satisfactory results: 26 misplaced values, the value of 0.890 for the precision_score and 0.895 for the recall_score and 0.888 the f1score. However, after several simulations, the components of the confusion matrix are very variable, but the performance remains the same.

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 do not like sports, whose variables that characterize them negatively are: Progression, Objective, Challenge, Perspiration, Active, Perseverance
- 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, Technicite, Qualite, Energie, Muscle, Force and Intensite
- 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 these algorithms (SVM) gave the value of 0.89 of precision .But the result of this algorithm is not satisfactory.

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