Projet datamining
Madame CARIOU

ROMAND Kyllien

DEL'CHATEAU Jean-Baptiste

2020/2021

Sommaire

Introduction

Le but de ce dossier est de montrer quelles sont les causes d'une pression artérielle anormalement élevée. Pour pouvoir expliquer cette pression artérielle, nous allons faire une analyse factorielle discriminante et une régression PLS-DA. Nous allons faire ces deux méthodes sur deux variables qui sont à expliquer, la première variable est la variable Pression et la seconde est la variables Press_arter. La variable Pression est une variable binaire à deux modalités, soit 1 si l'individu a trop de pression artérielle et 0 sinon. Pour la seconde variable Press-arter, celle-ci possède 4 modalités, lorsque l'individu à une pression artérielle inférieure à 140, un 1 lui sera affecté. SI sa pression est entre 140 et 159, le chiffre 2 lui sera affecté, si sa pression se situe entre 160 et 179, le chiffre 3 lui sera affecté et pour les individus ayant une pression artérielle supérieur à 179, le chiffre 4 leur sera affecté. La base de données que nous allons étudier se compose de 500 observations et de 10 variables dont 8 explicatives. Dans nos 8 variables explicatives, nous avons 2 variables quantitatives et 6 variables qualitatives.

Nous allons effectuer différents tests sur nos variables explicatives pour savoir lesquelles expliquent le plus justement les variables à expliquer, cette explication se fera en fonction de la qualité de notre modèle.

Dans un premier temps nous allons faire une analyse factorielle discriminante et une PLS-DA sur notre variable Pression à deux modalités. Puis dans un second temps, nous ferons une analyse factorielle discriminante et une PLS-DA sur notre variable à 4 modalités, Press_arter. Nous finirons ce dossier par une comparaison des deux méthodes et par expliquer quel est notre meilleur modèle. Vous pourrez également trouver une table des matières en fin de dossier.

La variable Pression

Dans cette première partie, nous allons réaliser une analyse factorielles discriminantes ainsi qu'une PLS-DA sur notre première variables à expliquer.

Analyse factorielle discriminantes

Une analyse factorielle discriminante permet de séparer au mieux les k groupes à partir des variables explicatives et fournit une représentation graphique du résultat (définition tirée du cours de Madame Cariou). Nous avons commencé par effectuer une discrimination sur notre base pour savoir quelles variables étaient les plus discriminantes. Sur le tableau 1, nous voyons que les variables ayant les plus petites p-values sont l'IMC, la pratique du sport et le fait de manger salé. Ce sont ces trois variables qui seront le plus discriminantes, elles seront les plus significatives, elles expliquent le mieux notre variable à expliquer. Plus la valeur de Wilks est faible, plus la variable sera discriminante.

Tableau 1 : Les valeurs de la discriminisation de nos variables explicatives

```
> round(res.desDA$power,3)
       cor_ratio wilks_lamb F_statistic p_values
Genre      0.004      0.996       1.765    0.185
Fumer      0.005      0.995       2.271    0.132
Sport      0.013      0.987       6.590    0.011
Age        0.000      1.000       0.001    0.970
Alcool     0.004      0.996       1.969    0.161
IMC        0.027      0.973      13.924    0.000
Stress     0.006      0.994       2.833    0.093
Sel        0.001      0.999       0.443    0.506
```

Grâce à la fonction discor, nous allons voir les différentes corrélations entre la composante principale et nos variables explicatives. Pour ce modèle, nous allons avoir une seule composante du fait que nous n'avons que deux modalités. Nous voyons que

les corrélations sont faibles exceptée pour l'IMC qui a une corrélation de 0,701. Cette variable sera donc importante pour l'axe.

Tableau 2 : Les corrélations entre nos variables explicatives et la composante

```
> round(res.desDA$discor,3)
           DF1
Genre   -0.253
Fumer    0.286
Sport   -0.486
Age     -0.007
Alcool   0.267
IMC      0.701
Stress   0.320
Sel     -0.127
```

La fonction discrivar nous permet de voir les coefficients de la régression linéaire, il faut donc regarder si ces coefficients sont positifs ou négatifs. Si le coefficient est positif, cela signifie que de fumer, par exemple, va faire augmenter la pression artérielle. Si le coefficient est négatif, cela signifie que la pratique du sport, par exemple, va faire baisser la pression artérielle. Nous émettons un doute par rapport à la variable sel, puisque nous savons que manger salée est mauvais pour la santé, donc celle-ci devrait être positive et dans notre modèle elle est négative, ce n'est pas logique.

Tableau 3 : Les coefficient de la régression linéaire

```
> round(res.desDA$discrivar,3)
             DF1
constant -2.396
Genre    -0.541
Fumer     0.511
Sport    -0.977
Age       0.003
Alcool    0.623
IMC       0.079
Stress    0.660
Sel      -0.144
```

Pour le test suivant, il va permettre de dire que si la composante augmente d'unité alors la pression va augmenter de 0,059. C'est donc le résultat que nous allons avoir devant notre variable à expliquer.

Tableau 4 : coefficient de la composante pour expliquer notre variable à expliquer

```
> round(res.desDA$values,3)
    value proportion accumulated
DF1 0.059        100         100
```

Nous allons finir cette discrimination par la corrélation entre notre variable à expliquer et notre composante. La corrélation est de 0,055, elle est donc très faible. Cette faible corrélation signifie que la puissance de la variable dans la composante est faible.

Tableau 5 : corrélation de notre variable à expliquer avec notre composante

```
> corRatio(res.desDA$scores[,1],base$Pression)
[1] 0.05538558
```

Nous allons effectuer une validation croisée sur notre base de données. Une validation croisée est une méthode d'estimation de fiabilité d'un modèle fondé sur une technique d'échantillonnage. L'échantillon va donc être séparé en deux groupes, un groupe servant à créer le modèle et élaborer les règles de décision ou d'affectation. Le

second groupe, le groupe teste, servant à estimer les performances du modèle. A la suite de cette cross-validation, nous allons faire une matrice de confusion pour repérer les bonnes ou mauvaises prédictions. Une matrice de confusion est une matrice qui mesure la qualité d'un système de classification, chaque ligne correspond à une classe réelle et chaque colonne correspond à une classe estimée. Grâce à la matrice de confusion, nous voyons qu'il y a 147+151 individus bien prédits soit 298 observations sur 500. Il y a 100 individus qui sont catégorisés comme ayant une forte pression artérielle, mais c'est une mauvaise prévision, de même, il y a 102 individus dans la catégorie pas de Pression alors que réellement ils en ont. Pour savoir si la qualité de notre modèle est optimale, nous allons appliquer deux taux d'erreur, un sur tout le jeu de données et un autre sur la validation croisée.

Tableau 6 : Matrice de confusion

```
> mat.confusion.app

    0   1
0 147 100
1 102 151
```

Le taux d'erreur correspond à la qualité générale du modèle. Pour trouver ce taux d'erreur, nous allons exécuter un code qui l'opération de diviser les bonnes prédictions par le nombre total de prédiction le tout soustrait de 1, pour notre étude cela donne 298/500. Pour tout le jeu de données, le taux d'erreur est donc de 40.4% et pour la validation croisée, le taux d'erreur est de 43.4%. Le taux d'erreur pour la validation croisée est supérieur au taux d'erreur pour tout le jeu de données du fait qu'en validation croisée il y a un partie d'estimation, c'est donc moins précis

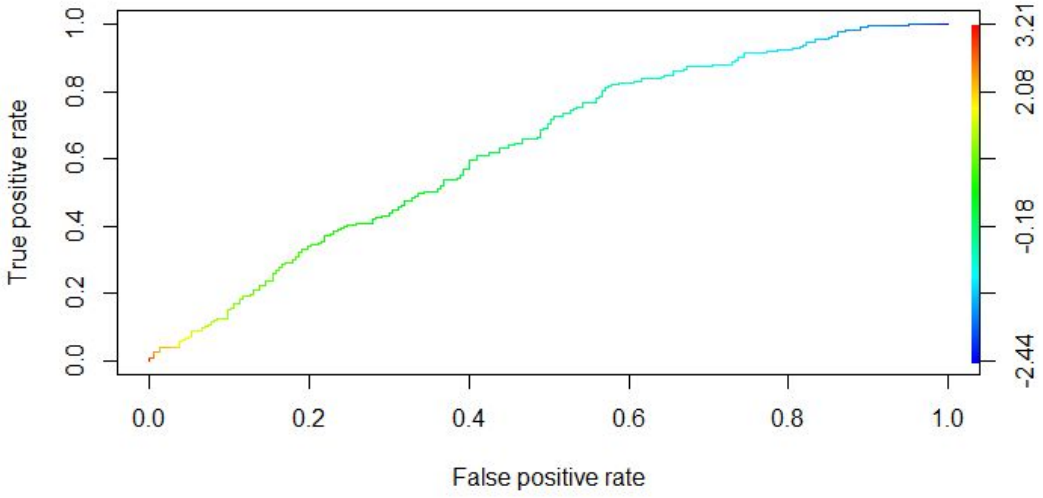
.

Tableau 7 : Les deux taux d'erreur

```
> #Taux d'erreur sur tt le jeu de données
> tx.err.app <- 1- sum(diag(mat.confusion.app)/sum(mat.confusio
n.app))
> tx.err.app
[1] 0.404
> #taux d'erreur en validation croisé
> res.geoDA$error_rate
[1] 0.434
```

Nous allons voir une dernière chose pour cette AFD, c'est la courbe de ROC.

Graphique 1 : Courbe ROC



PLS-DA

La PLS-DA est une adaptation de la régression PLS au contexte de la discrimination :

- X est le tableau des variables explicatives
- Y représente les groupes a priori, est recodée en un tableau disjonctif complet. Une régression PLS2 est alors effectuée.

Une fois les composantes PLS déterminées, l'affectation se fait selon une règle géométrique ou par analyse linéaire discriminantes sur la base des composantes de la PLS-DA (définition tirée du cours de Madame Cariou). Nous continuons de travailler sur le même échantillon à 500 observations et deux modalités pour notre variable à expliquer.

Dans un premier temps, nous avons fait la matrice de confusion du modèle PLS-DA, nous voyons qu'elle est différente, en PLS-DA, nous avons le même nombre de bonne prédiction, soit 298, mais c'est au niveau des mauvaises prédiction qu'il y a des différences avec la matrice de confusion de l'AFD. Nous avons 202 observations où les prédictions sont mauvaises. Au niveau du taux d'erreur, nous trouvons le même taux d'erreur que pour l'analyse factorielle discriminante avec un taux équivalent à 0,404.

Tableau 8 : Matrice de confusion en PLS-DA et taux d'erreur

```
> my_pls1$confusion
          predicted
original   0   1
       0 144 103
       1  99 154
```

```
> my_pls1$error_rate
[1] 0.404
```

Avant de voir le nombre de composantes optimales à choisir, nous allons regarder quelles sont nos variables qui ont le plus d'impact sur notre variable à expliquer. Pour ce faire, nous allons regarder dans le tableau du modèle VIP quelles sont les composantes dont la valeur est supérieure à 1. Dans ce tableau, nous voyons que les variables Sport et IMC sont les variables impactant le plus

notre variable Pression. Nous pouvons donc dire que le fait de faire du sport et/ou d'avoir un IMC correspond à une corpulence normale, diminue le fait d'avoir un pression artérielle à risque.

Tableau 9 : Impact de nos variables sur Y

```
> my_pls1$VIP
       Component 1 Component 2 Model VIP
Genre    0.69306119   0.6943510 0.6943510
Fumer    0.78577303   0.7856392 0.7856392
Sport    1.33286569   1.3278277 1.3278277
Age      0.01962352   0.1364383 0.1364383
Alcool   0.73184336   0.7357287 0.7357287
IMC      1.92340917   1.9160996 1.9160996
Stress   0.87718356   0.8741342 0.8741342
Sel      0.34750883   0.3769099 0.3769099
```

Nous allons maintenant nous intéresser aux nombre de composantes optimales pour notre modèle. Dans le tableau suivant, le premier taux d'erreur correspond à deux composantes, le deuxième chiffre correspond à trois composantes et ainsi de suite. Grâce au tableau 10, nous voyons les différents taux d'erreur, nous allons donc prendre le taux d'erreur qui est le plus petit, 0.404 qui correspond à un nombre de composantes égales à 2. Nous allons donc refaire la matrice de confusion avec ce nombre de deux composantes. Nous pouvons déjà dire que le taux d'erreur que nous avons trouvé en début de partie correspondait au taux le plus faible. Il serait donc logique de trouver une matrice de confusion égale à la matrice de confusion trouvée en début de partie. Le tableau 11, nous montre la matrice de confusion avec deux composantes, et nous voyons que c'est la même que celle en début de partie. L'analyse que nous avons donc faite était la bonne.

Tableau 10 : Les neufs taux d'erreur

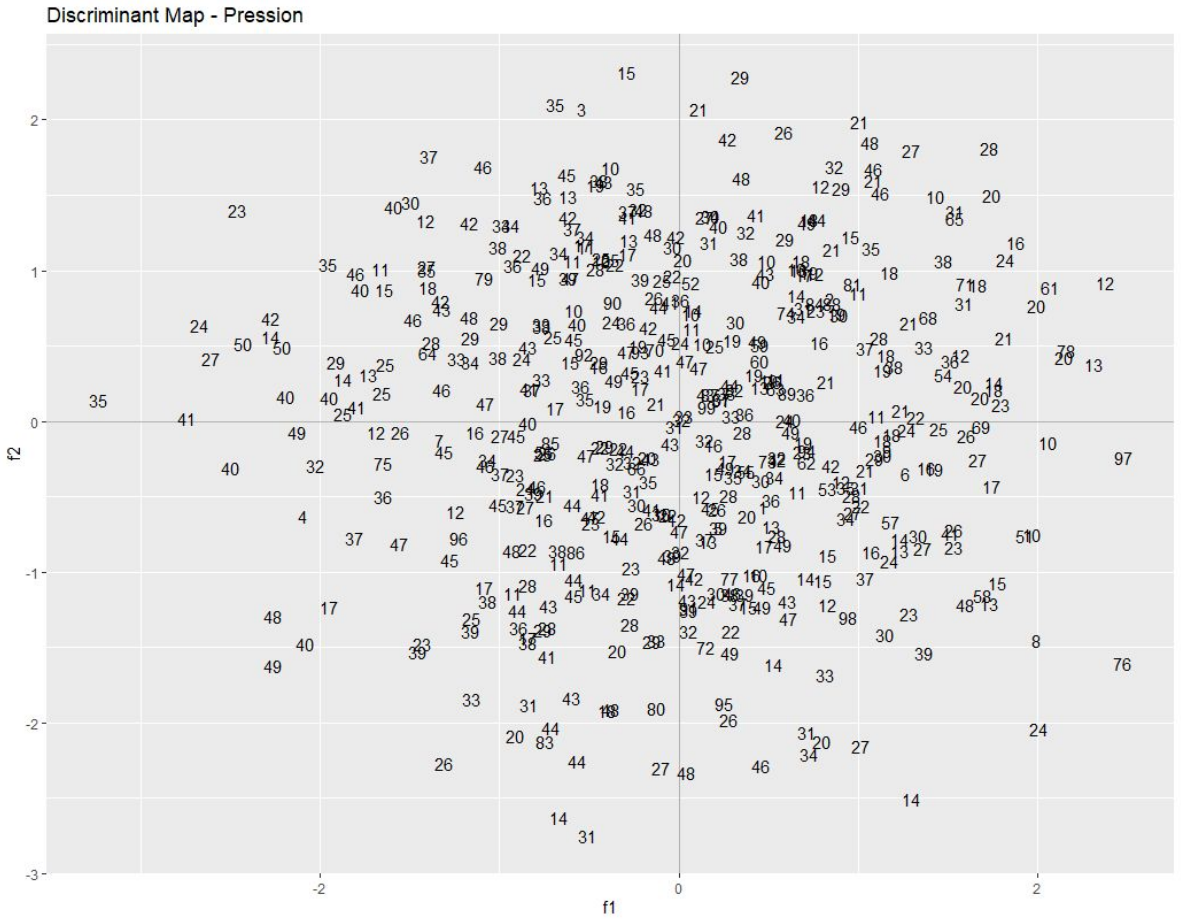
```
> err.rate
[1] 0.404 0.406 0.406 0.406 0.406 0.406 0.406 0.406 0.406
```

Tableau 11 : Matrice de confusion avec 2 composantes

```
> my_pls1$confusion
          predicted
original   0   1
       0 144 103
       1  99 154
```

Avant de conclure sur la variable pression, nous avons voulu intégrer la répartition de nos différents individus de la discrimination avec PLS-DA. Il s'agit du graphique 1, ci-dessous. Sur ce graphique, nous voyons que les individus sont répartis de façon homogène sur les axes 1 et 2, il n'y a pas de "clusters". Il y a quelques individus un peu éloignés des autres comme le 35 (le long de l'axe 1 à gauche) ou le 76 (dans le cadrant en bas à droite)

Graphique 2 : Répartition des individus sur la discrimination



Pour conclure sur cette partie, nous avons vu que nous trouvons un taux d'erreur qui est égal, de 0.404 mais une matrice de confusion différente au niveau des erreurs de prédiction puisque en faisant les additions, nous trouvons que les prédiction était les même avec 298 individus ayant une bonne prédiction. La différence entre les deux matrices, en AFD, 102 individus sont prédits avec une pression artérielle normale alors que ceux ci ont une pression anormalement élevée, en PLS-DA, ce nombre d'individus est de 99. En ce qui concerne les variables ayant un plus gros impact sur notre variable à expliquer, en PLS-DA ce sont les variables IMC et SPORT, en AFD c'est la variable IMC. La PLA-DA est donc une meilleur méthode pour nous dire les variables qui ont le plus d'impact, cette fonction sera peut être plus précise que l'AFD.

Nous allons maintenant refaire des tests sur notre deuxième variable à expliquer, Press_arter. L'analyse va être différente dans la partie suivante, car cette variable à 4 modalités au lieu de 2. Les 4 modalités ont été expliquées dans l'introduction.

La variable Press_arter

Cette partie va avoir la même structure que la variable pression, c'est-à-dire, nous allons voir l'AFD et la PLS-Da en mettant la variable étudier en variable à expliquer.

Analyse factorielle discriminante

Nous allons fonctionner de la même manière que pour la variable pression, c'est-à-dire que nous allons analyser les ratios de wilks, les corrélations entre les composantes et avec la variable à expliquer, le poids des différentes variables par rapport aux composantes et leur effet sur les composantes. Sachant que la variable Press_arter est une variable à 4 modalités, ces modalités dépendent de la valeur de la pression artérielle de l'individu, il y aura donc 3 composantes pour l'AFD.

Tableau 12 : Valeurs de la discrimination des variables explicatives

	cor_ratio	wilks_lamb	F_statistic	p_values
Genre	0.005	0.995	0.893	0.445
Fumer	0.027	0.973	4.623	0.003
Sport	0.014	0.986	2.390	0.068
Age	0.009	0.991	1.467	0.223
Alcool	0.017	0.983	2.938	0.033
IMC	0.047	0.953	8.186	0
Stress	0.010	0.990	1.718	0.162
Sel	0.004	0.996	0.726	0.537

Pour la variable pression artérielle, il y a 3 variables explicatives qui sont hautement discriminantes : l'IMC, la consommation d'alcool et le fait de fumer ; la pratique d'un sport l'est aussi mais à un seuil d'erreur plus élevé. Pour notre variable expliquée, nous allons la discriminer avec 3 composantes principales car nous avons 4 modalités.

Tableau 13 : Corrélation entre les variables explicatives et les composantes

	DF1	DF2	DF3
Genre	0.025	-0.270	0.330
Fumer	-0.462	-0.552	-0.223
Sport	0.195	-0.310	0.513
Age	-0.212	0.265	0.685
Alcool	-0.449	-0.182	0.118
IMC	-0.701	0.238	-0.056
Stress	-0.084	0.451	-0.215
Sel	0.103	-0.322	-0.183

On peut remarquer que la première composante est fortement corrélée à l'IMC, et plus légèrement avec le fait de fumer et la consommation d'alcool ; la deuxième quant à elle, est corrélé à seulement deux variables : le stress et fumer. La dernière composante est corrélée au Sport et à l'Âge.

Tableau 14 : Coefficient des variables explicatives sur les composantes

	DF1	DF2	DF3
constant	3.859	-0.368	-2.276
Genre	0.004	-0.670	0.709
Fumer	-0.894	-1.231	-0.496
Sport	0.391	-0.598	1.040
Age	-0.021	0.021	0.051
Alcool	-1.137	-0.464	0.380
IMC	-0.082	0.030	0
Stress	-0.153	0.928	-0.472
Sel	0.061	-0.801	-0.466

Après utilisation de la fonction discrivar, on obtient l'effet que chacunes des variables explicatives exercent sur chacunes des composantes :

- Pour la première composante : on remarque une constante plutôt haute qui est diminuée par toutes les variables dites "négatives" : Fumer, Alcool, Stress, IMC et

augmenter par le sport et le genre. Cette composante est l'inverse que l'on peut imaginer.

- Pour la seconde composante : on remarque une constante faible et négative qui est augmentée par l'Age, l'IMC et le Stress ; toutes les autres variables la diminuent avec un effet beaucoup plus puissant pour le fait de fumer, d'être stressé ou de manger salé.

- Pour la dernière composante : sa constante est hautement négative, elle est diminuée par les variables Fumer, Stress et Sel, toutes les autres augmentent sauf IMC qui a un effet nul.

On remarque qu'il est difficile de voir les liens entre les composantes et les variables, et donc de nommer ou du moins attribuer un type de variable à chacune des composantes.

Tableau 15 : Coefficient des composantes dans le modèle de régression

	value	proportion	accumulated
DF1	0.078	72.543	72.543
DF2	0.022	20.235	92.777
DF3	0.008	7.223	100

Ce tableau permet de déterminer les coefficients qui s'appliquent à chacune des composantes pour expliquer la pression artérielle, on remarque que la première composante a le plus gros effet mais aussi qu'elle est plus significative (77,543 % de proportion).

Tableau 16 : Corrélation entre Pression Artérielle et les composantes

DF1	DF2	DF3
0.087	0.035	0.013

Il y a une très faible corrélation entre notre variable explicative et les différentes composantes, cela induit que la puissance de la variable dans la composante est faible.

Afin de vérifier notre modèle nous allons procéder à une validation croisée. Suite au tableau 17, on voit apparaître notre matrice de confusion, et seulement 169 des

individus sont correctement prédit ce qui correspond à seulement 33.8% de bonne prédiction ce qui est plutôt faible. Nous allons par la suite comparer les deux taux d'erreurs.

Tableau 17 : Matrice de Confusion

	"1"	"2"	"3"	"4"
"1"	84	63	43	57
"2"	21	45	23	28
"3"	17	15	22	18
"4"	5	12	11	36

Les deux taux d'erreurs, celui sur la totalité de la data et celui en validation croisée, sont respectivement de 62.6% et 70,4%. Encore une fois la méthode en validation croisée est de moins bonne qualité. Cependant les taux d'erreurs sont élevés.

PLS-DA

Suite à l'application de l'algorithme PLS, on obtient un modèle avec deux composantes.

Comme vu avec la précédente variable à expliquer nous allons dans un premier temps étudier la matrice de confusion afin de comparer les valeurs prédites et les valeurs observées ce qui permet de déterminer le taux d'erreurs et donc le taux de prédiction.

Tableau 18 : Matrice de Confusion

	"1"	"2"	"3"	"4"
"1"	229	12	0	6
"2"	108	6	0	3
"3"	68	1	0	3
"4"	52	6	0	6

On prédit correctement 238 pressions artérielles correctement, ce qui correspond à un taux d'erreurs de 51.8%. On remarque que le taux de prédiction est supérieur qu'en analyse discriminante cependant, avec la matrice de confusion on voit que notre modèle prédit surtout des 1.

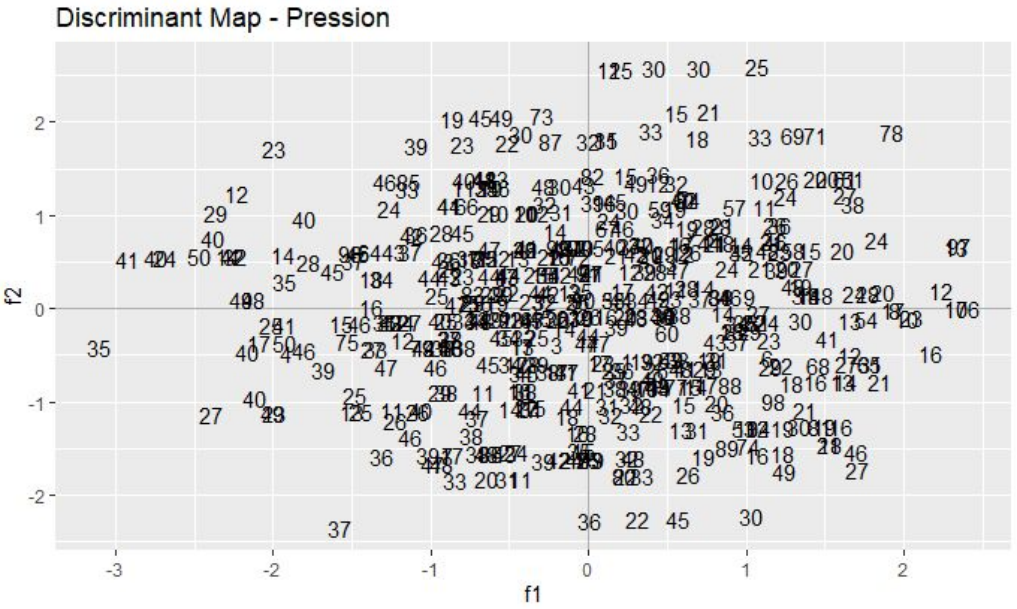
A l'aide du tableau 19, on remarque qu'elles sont les variables qui ont le plus d'impact sur nos composantes et donc sur notre variable à expliquer, c'est l'IMC qui l'impact le plus significatif suivi de l'Alcool, du Sport et de Fumer.

Tableau 19 : Impact des variables

	Component 1	Component 2	Model VIP
Genre	0.391	0.608	0.608
Fumer	1.036	1.253	1.253
Sport	0.963	0.996	0.996
Age	0.364	0.309	0.309
Alcool	1.035	1.011	1.011
IMC	2.039	1.729	1.729
Stress	0.592	0.862	0.862
Sel	0.364	0.469	0.469

Suite au graphique 3, on peut observer les différentes répartitions des individus en fonction des axes qui correspondent aux deux composantes. Aucun cluster n'apparaît cependant on peut noter des individus qui ont l'air éloignés mais il n'y pas à l'air d'avoir de valeurs atypiques.

Graphique 3 : Répartition des individus sur la discrimination



Afin de conclure concernant la variable pression artérielle, nous pouvons comparer les taux d'erreurs obtenus par les différentes méthodes utilisées, il en ressort que la méthode PLS est de meilleure qualité car son taux de prédiction est plus élevé. Cependant il en ressort dans les modèles que la variable IMC, Fumer et Sport sont les plus significatives qu'importe le modèle, on peut donc penser que ce sont les facteurs les plus déterminants pour la pression artérielle, ce qui paraît normal puisque l'IMC indique la condition physique de la personne tout comme si elle pratique ou non une condition physique, de plus le fait de fumer est une source de problème cardio-vasculaire.

Conclusion

Lorsqu'on compare les différents résultats obtenus pour les deux variables, on peut remarquer que la régression PLS est celle qui obtient les meilleurs résultats. Cependant les taux de prédiction sont plus élevés lorsqu'il faut expliquer la variable pression, on peut penser que ceci est dû à la précision supplémentaire demander à l'algorithme quant à prédire pression artérielle.

La régression logistique est dans notre cas plus intéressante à effectuer car le taux de prédiction est supérieur, cependant s'il avait fallu faire une régression à partir d'un plus grand nombre de variables explicatives et d'individus, les analyses discriminantes aurait été de meilleurs outils car le travail demandé par la régression

logistique aurait été multiplié par le nombre de variable en plus, alors que celui des analyses discriminantes n'aurait pas changé.

Table des matières