SY09 - TP1

Matthieu PETIT - Pierre-Louis LACORTE

5 avril 2017

Objectif

L'objectif de ce TP est de découvrir la statistique descriptive et l'analyse en composantes principales, via le logiciel R. Pour cela nous allons utiliser trois jeux de données, qui vont nous permettre d'appréhender les étapes d'une étude statistique dans le cadre du Data Mining. Tout d'abord par l'extraction des données et leur nettoyage. Puis par l'analyse de ces données en composantes principales, et sa mise en place grâce à R.

Table des matières

1 Statistique descriptive

1.1 Données Notes

Le jeu de donnees Notes est contenu dans un fichier sy02-p2016.csv qui contient des informations relatives aux etudiants inscrits a l'UV SY02 au semestre de printemps 2016.

1.1.1 Quelques valeurs manquantes

Ce jeu de données regroupe les résultats de 296 étudiants. Certains de ces étudiants présentent des valeurs de résultats manquantes identifiées par : NA.

Ceci peut s'expliquer par une abscence à un examen. Afin de simplifier le traitement des données on va supprimer ces étudiants de l'échantillon. Une fois ces étudiants retirés on étudie donc un échantillon de 279 étudiants.

1.1.2 Etude descriptive

Les informations sauvegardées pour chaque étudiants sont :
Nom
Spécialité : Branche de l'étudiant : GI, GSM, GM, GSU, GB, GP, HuTech, TC, ISS
Statut : origine de l'étudiant : échange, UTC
Dernier diplôme obtenu : dernier diplôme obtenu par l'étudiant
Note médian correcteur médian : Note obtenue au médian et correcteur de la copie
Note final correcteur final : Note obtenue au final et correcteur de la copie
Note total : Note obtenue au total. Le coefficient du médian est 0,4 et celui du final de 0,6
Resultat : Resultat obtenu à SY02 : F, FX, E, D, C, B, A

La première étude que nous pourrions faire est celle des résultats à l'UV sans distinction de la provenance des étudiants. Ces résulats sont présentés Figure 1, page 2. L'obtention de l'UV se fait pour un résultat compris entre E et A. Un F ou un FX correspond à un échec. On constate que 28% des étudiants ont échoué SY02 et que 7% ont eu l'UV avec mention (c.a.d avec A).

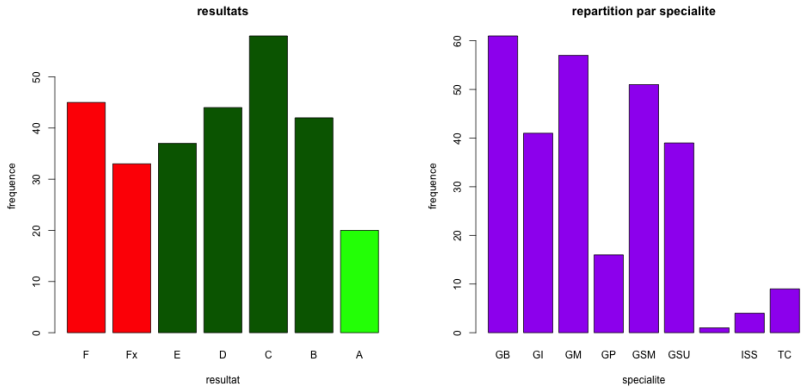

FIGURE 1 – Résultats et répartition par spécialité de l'ensemble des étudiants à SY02

La seconde étude que nous pouvons faire est celle de la répartition des spécialités. Cette répartition est représentée Figure 1. On remarque que la majoritée des spécialités représentées sont les spécialités "classiques". C'est-à-dire

les branches de GI, GP, GSM, etc.. La branche la plus présente est celle des GB avec 61 étudiants suivie des GM avec 57 étudiants. Les branches les moins représentées sont HuTech et ISS avec respectivement 1 et 4 étudiants.

1.1.3 Variables supposées liées

Le premier lien qui peut être étudié, aussi trivial soit-il, est celui de la note finale et du résultat. Cela revient à étudier la corrélation d'une variable quantitative (note) avec une variable qualitative (résultat). Pour ce faire on étudie dans un premier temps la variance des notes finales :

$$var_{notefinale} = \sum_{i=1}^{n}(x_i - \bar{x}^2) = 3291.568$$

puis la variance intergroupe (chaque résultat définit un groupe) :

$$var_{resultat} = \sum_{k=1}^{p} n_k(\bar{x_k} - \bar{x})^2 = 3100.532$$

où $\bar{x_k}$ désigne la note des étudiants du groupe k. Finalement pour étudier la relation entre une variable qualitative et une variable quantitative, on calcule le rapport de corrélation noté η^2

$$\eta^2 = \frac{var_{notefinale}}{var_{resultat}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}^2)}{\sum_{k=1}^{p} n_k(\bar{x_k} - \bar{x})^2} = 0.9419621$$

η^2 étant très proche de 1 nous pouvons en déduire que la note totale et le résultat sont extrêmement lié, ce qui était bien évidemment évident mais qui vient d'être prouvé. Ceci nous permet donc de ne travailler avec le résultat au même titre que la note totale. Nous pouvons représenter graphiquement ces liens statistiques avec la Figure 2, page 4. Les groupes sont représentés verticalement, les individus par des carrés et la moyenne de chaque groupe est représentée d'un point rouge. Finalement la moyenne de l'ensemble des étudiants est représentée par une ligne pointillée et l'écart écart de la moyenne des groupes par un trait bleu.

On peut alors se poser la question de la réussite à l'examen en fonction de la spécialité des étudiants. Ces résultats sont présentés sous forme de boîte à moustaches Figure 3, page 4 où la graduation de l'axe des ordonnées fait correspondre deux à deux respectivement les ensembles {1, 2, 3, 4, 5, 6, 7} et {F, FX, E, D, C, B, A}. *Dans ce traitement on tronque l'étudiant HuTech car une seule réalisation ne permet pas d'avoir un regard critique sur les résultats.* Les branches les plus représentées ne sont pas celle ayant le mieux réussi l'examen puisque nous constatons que les spécialités présentant les meilleurs médianes (souffrant moins des valeurs extrêmes que la moyenne) sont les TC et les GSU avec un résultat de C. A contrario les branches ayant le moins réussi sont les GSM et les ISS avec une médiane de E.

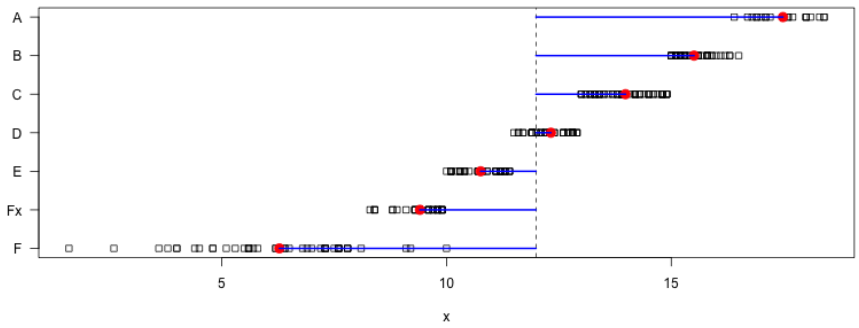

FIGURE 2 – Etude des notes pour chaque résultat

La réussite des étudiants semble donc liée au moins en partie à la spécialité de l'étudiant. On notera que les TC sont ceux qui reussissent globalement le mieux l'examen. Ceci peut cependant être expliqué : SY02 est une UV prise généralement en branche, les TC qui prennent cette UV sont donc des étudiants qui pensent être à même de réussir et qui ont assez d'avance pour prendre des UV de branche. Pour résumé, seul de bons TC prennent cette UV ce qui explique ce résultat.

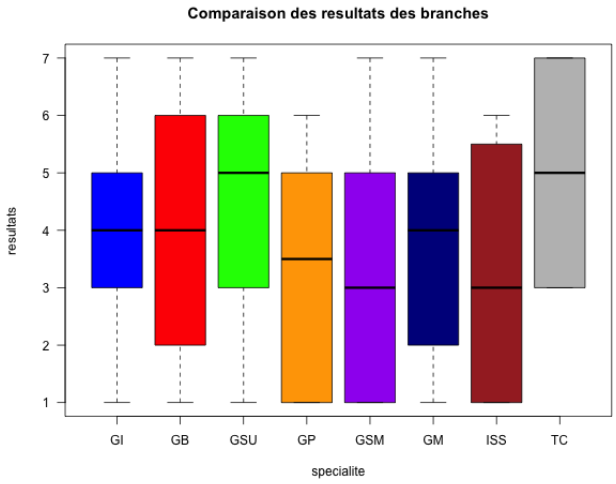

FIGURE 3 – Réussite à l'examen pour chaque spécialité de l'échantillon

Une autre corrélation que nous pourrions étudier serait celle des notes et des

correcteurs. Les correcteurs admettent-ils la même moyenne de note lors des corrections ? On peut représenter cette étude par des boxplots Figure 4 mettant en lien les notes attribuées par chaque correcteur au médian et au final.

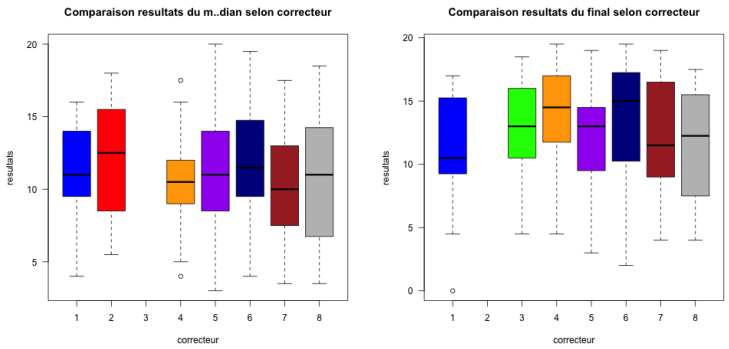

FIGURE 4 – Notes attribuées par les correcteurs

On peut constater que les correcteurs n'ont pas la même distribution de notes ce qui induit que le correcteur influence au moins en partie le résultat de l'étudiant.

1.1.4 Confirmation des observations

Nous avons pu voir précédemment qu'il semblait y avoir un lien entre la spécialité de l'étudiant et son résultat à l'examen. Pouvons-nous le confirmer statistiquement ? Pour ce faire on va étudier la corrélation de la variable quantitative "note totale" avec celle de la qualitative "spécialité". Nous constatons alors les résultats suivants :

$$var_{notefinale} = 3291.568$$

$$var_{specialité} = \sum_{i=1}^{n}(x_i - \bar{x}^2) = 256.8697$$

$$\eta^2 = \frac{var_{notefinale}}{var_{resultat}} = 0.0780387$$

Nous récupérons donc un η^2 très proche de zéro ce qui semble contredire l'observation précédente. Une appartenance à une spécialité ne permet donc pas de prédire le résultat de l'étudiant avec précision. La représentation graphique Figure 5 de cette relation vient confirmer une telle indépendance.

La dernière étude sur le jeu de notes que nous pouvons faire est celle de la relation entre le dernier diplôme obtenu et le résultat de l'étudiant. On est ici dans le cadre de l'étude de 2 variables qualitatives { *diplôme* - *résultat* }. L'outils que nous utilisons est un tableau de contingence Figure 6 pour voir le lien entre ces variables.

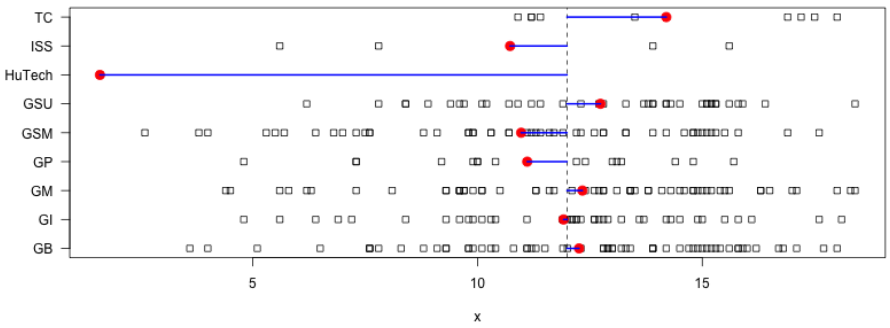

FIGURE 5 – Notes obtenues par spécialité

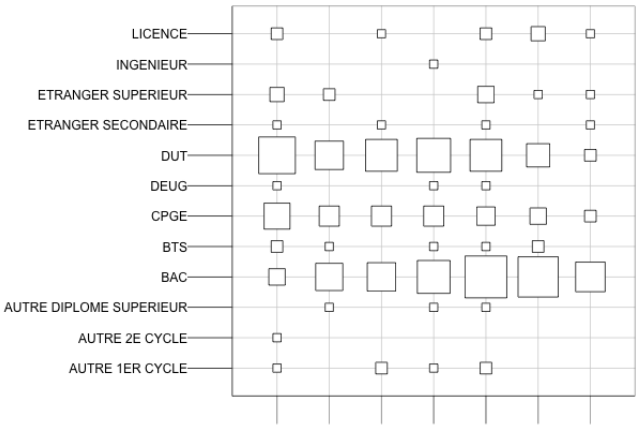

FIGURE 6 – Contingence des diplômes et des résultats

De nombreuses valeurs du tableau étant inférieures à 5 réalisations nous ne pouvons pas effectuer un test du χ^2. Les effectifs étant relatifs à une taille d'échantillon. Ils ne sont pas très intéressants si nous voulons la rapporter au niveau de la population. Il semble plus indiqué de passer aux proportions, nous pouvons les extraire de différentes manières et le plus intéressant est de ramener les effectifs par rapport aux totaux marginaux en ligne ou en colonne. Les pourcentages qui nous permettent de mieux discerner les informations importantes sont les pourcentages en colonne Table 1.

On remarque de ce tableau de contingence que les diplômés d'un BAC représentent 37% des étudiants mais participent à hauteur de 57% et 65% aux B et A. Ces étudiants sont donc des étudiants passant avec succès les examens alors qu'au contraire, si nous prennons le cas des CPGE qui représentent 13%

	F	Fx	E	D	C	B	A	TOT
1^{er} cycle	2.22	0.00	5.41	2.27	3.45	0.00	0.00	2.15
2^{e} cycle	2.22	0.00	0.00	0.00	0.00	0.00	0.00	0.36
Dip. sup.	0.00	3.03	0.00	2.27	1.72	0.00	0.00	1.08
BTS	4.44	3.03	0.00	2.27	1.72	4.76	0.00	2.51
CPGE	22.22	18.18	16.22	13.64	8.62	9.52	10.00	13.98
BAC	8.89	33.33	32.43	36.36	44.83	57.14	65.00	37.99
DEUG	2.22	0.00	0.00	2.27	1.72	0.00	0.00	1.08
DUT	44.44	36.36	40.54	38.64	25.86	19.05	10.00	31.90
Etranger 2^{nd}	2.22	0.00	2.70	0.00	1.72	0.00	5.00	1.43
Etranger sup.	6.67	6.06	0.00	0.00	6.90	2.38	5.00	3.94
Ingénieur	0.00	0.00	0.00	2.27	0.00	0.00	0.00	0.36
Licence	4.44	0.00	2.70	0.00	3.45	7.14	5.00	3.23
TOTAL	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00

TABLE 1 – Tableau de contingence avec pourcentage en colonnes

des étudiants participent eux à hauteur de 20% des étudiants qui échouent à l'examen (obtenant F ou Fx). Au même titre les étudiants sortant de DUT et qui représentent 38% des étudiants, représentent 44 des F et 38% des Fx. On semble donc voir que certains diplômes favorisent un bon résultat (BAC ou LICENCE) alors que d'autres semblent tendre vers des échecs plus nombreux (CPGE ou DUT).

1.2 Données Crabs

Ce jeu de données regroupe les caractéristiques de 200 crabes. Il sont séparés en deux espèces : "bleue" et "orange" (100 individus chacune avec 50 mâles et 50 femelles). De plus cinq caractéristiques morphologiques des crabes nous sont données :

FL : la taille du lobe frontal
RW : la largeur arrière du crabe
CL : la longueur de la carapace
CW : la largeur de la carapace
BD : la hauteur du corps

Pour résumer ce jeu de données nous observons qu'il y a deux variables qualitatives : l'espèce (sp) et le sexe (sex) ; et cinq variables quantitatives, énoncées ci-dessus.

	FL	RW	CL	CW	BD
FL	1.0000000	0.9069876	0.9788418	0.9649558	0.9876272
RW	0.9069876	1.0000000	0.8927430	0.9004021	0.8892054
CL	0.9788418	0.8927430	1.0000000	0.9950225	0.9832038
CW	0.9649558	0.9004021	0.9950225	1.0000000	0.9678117
BD	0.9876272	0.8892054	0.9832038	0.9678117	1.0000000

TABLE 2 – Corrélation entre les variables qualitatives des données crabs

1.2.1 Différences morphologiques des crabes en fonction de leur espèce ou de leur sexe

Pour établir une première analyse de ces données nous avons tracé les diagrammes en boîtes à moustaches des différentes variables quantitatives en fonction des variables qualitatives. Nous obtenons donc les figures suivantes :

Comme nous pouvons l'observer, il y a, dans tous les cas, des mensurations plus grandes pour les crabes de l'espèce "orange". En effet nous observons qu'un chevauchement des boîtes, toujours au profit de l'espèce orange. À l'inverse, il n'est pas vraiment possible de supposer un lien entre sexe et mensurations.

1.2.2 Peut-on identifier l'espèce ou le sexe en fonction des mensurations d'un crabe ?

Nous allons tenter de répondre à cette problématique, pour cela, nous pouvons tracer deux graphes matriciels, le premier comparant les mensurations selon l'espèce, et le second selon le sexe.

Nous pouvons ainsi analyser séparément ces graphes. Le premier concernant l'identification de l'espèce en fonction de ces caractéristiaues morphologiques, nous observons qu'il n'y a pas vraiment de rapport entre deux des variables quantitatives qui engendre un graphique avec une séparation nette des nuages de points orange et bleu. On en déduit que l'analyse des mensurations d'un crabe ne permet pas de déterminer facilement s'il est de l'espèce orange ou s'il est de l'espèce bleue.

A l'inverse, pour ce qui est de l'identification du sexe, nous observons quelques rapports présentant des nuages de points bleu et rose bien distincts. Nous en déduisons que si nous souhaitons identifier le sexe d'un crabe, nous pourrons donc observer les binômes RW/FL, RW/CL, RW/CW et RW/BD, pour lesquels les droites bleues et roses sont les plus distinctes.

1.2.3 Etude de la corrélation entre les différentes mensurations morphologiques des crabes

Nous pouvons, en étudiant ce tableau de corrélation entre les différentes mensurations morphologiques des crabes, déduire que ces variables sont très corrélées. En effet, leur coefficient de corrélation est toujours supérieur à 0,88. Plus précisément, les variables les plus corrélées sont CL et CW, c'est à dire la

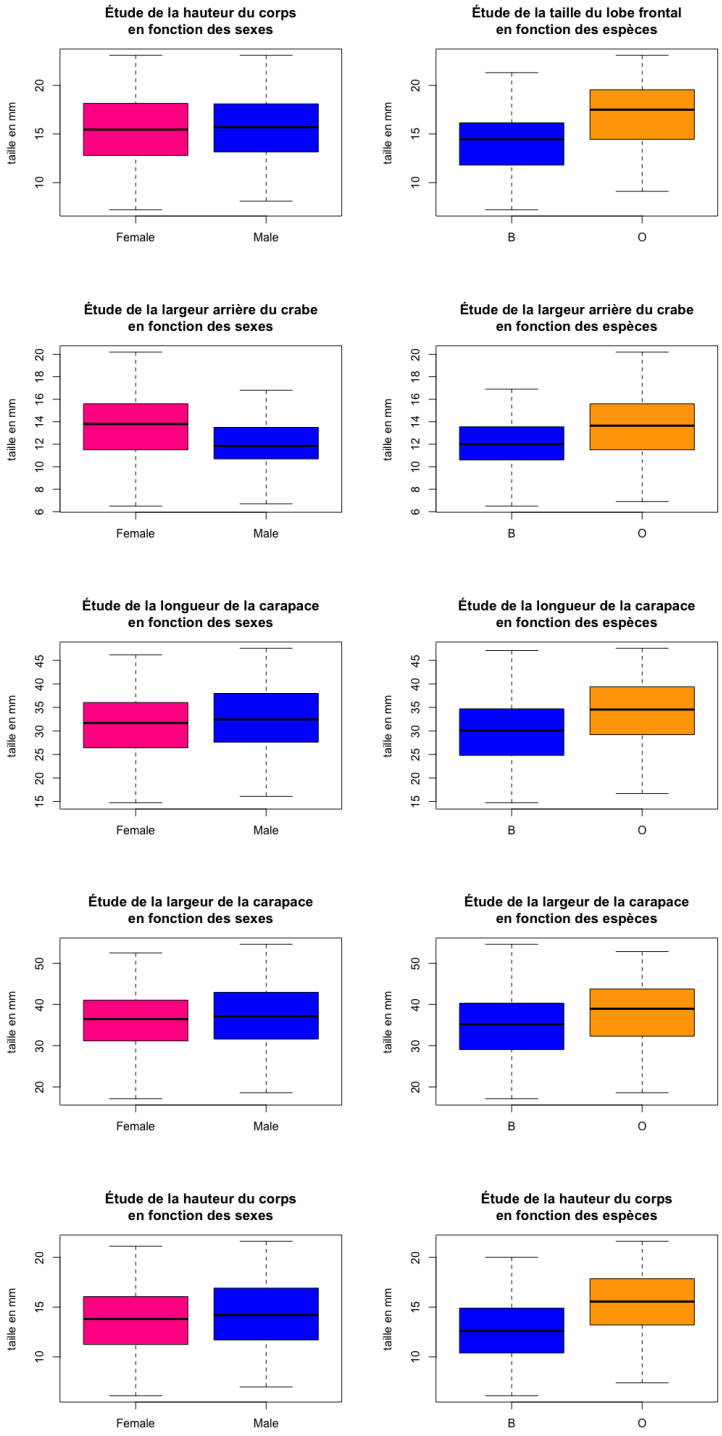

FIGURE 7 – Diagrammes en boîtes

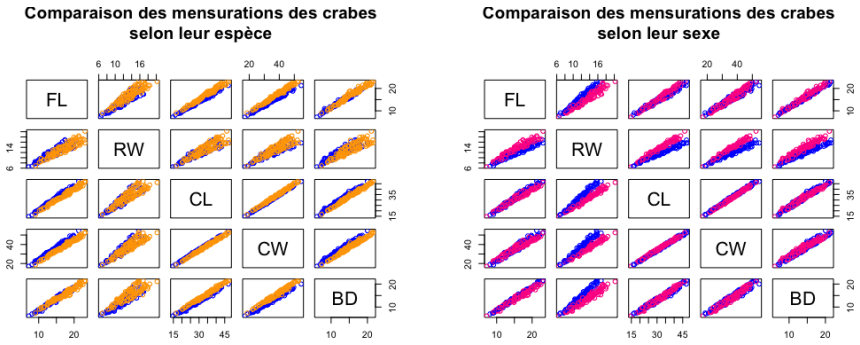

FIGURE 8 – Graphes matriciels

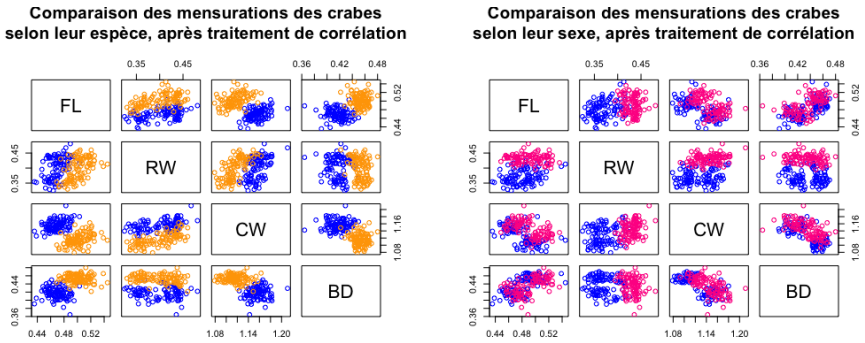

FIGURE 9 – Graphes matriciels après décorrélation

longueur et la largeur de la carapace des crabes. Cette corrélation peut sembler logique, pour que les crabes possèdent une morphologie harmonieuse, il est nécessaire que ces mensurations soient proportionnelles.

1.2.4 Etude matricielles après traitement de décorrélation

Afin de réduire l'effet de corrélation entre les mesures, nous allons traiter les données afin de les rendre analysables plus facilement. Nous choisissons donc de ne garder qu'une seule des valeurs entre CL et CW, car ce sont les deux variables les plus corrélées. On élimine CL, pour ne garder que FL, RW, CW et BC, et on les divise par CL. C'est ce traitement qui permet d'eviter l'effet de taille entre les variables.

Nous obtenons ainsi les graphes matriciels suivants, qui permettent de séparer plus distinctement les nuages de points correspondant aux espèces, ou aux sexes des crabes.

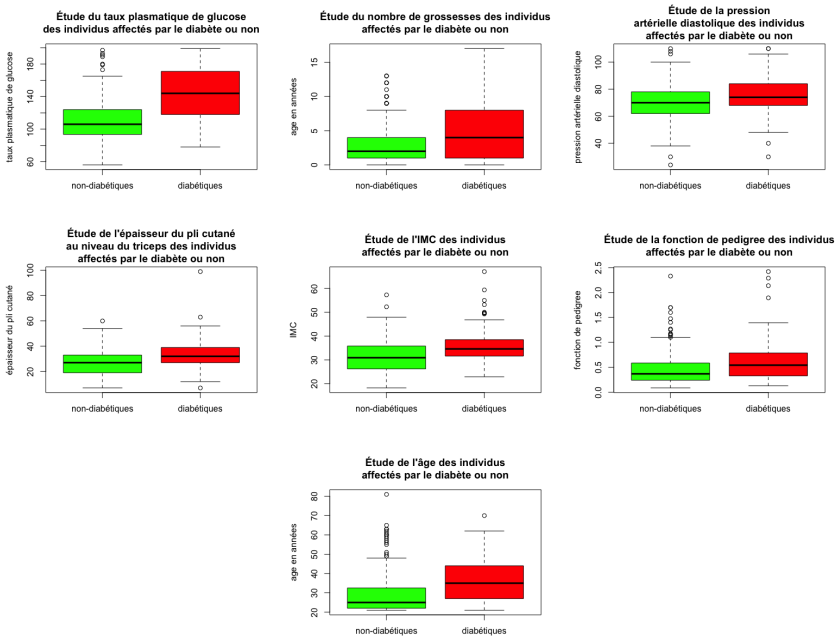

FIGURE 10 – Diagrammes en boîtes à moustaches des données PIMA

1.3 Données PIMA

Ce jeu de données regroupe 532 femmes (177 diabétiques et 355 saines), définies par 8 variables :

npreg (quantitative) : nombre de grossesses
glu (quantitative) : taux plasmatique de glucose
bp (quantitative) : pression artérielle diastolique
skin (quantitative) : épaisseur cutané du pli au niveau du triceps
bmi (quantitative) : indice de masse corporelle
ped (quantitative) : fonction de pedigree du diabète
age (quantitative) : âge
z (qualitative) : catégorie (diabétique si z=2, sinon z=1)

1.3.1 Différences entre individus diabétiques ou non

Nous allons commencer par observer les différences qu'il pourrait y avoir dans les valeurs que prennent les caractéristiques des femmes diabétiques et des femmes non-diabétiques. Nous avons choisi pour cela de représenter les différentes variables quantitatives par des diagrammes en boîtes à moustaches, un pour les femmes diabétiques, et un pour celle qui ne le sont pas.

Nous observons ainsi que les données sont assez proches dans la plupart des cas, bien qu'elles prennent toujours des valeurs plus élevées pour les femmes

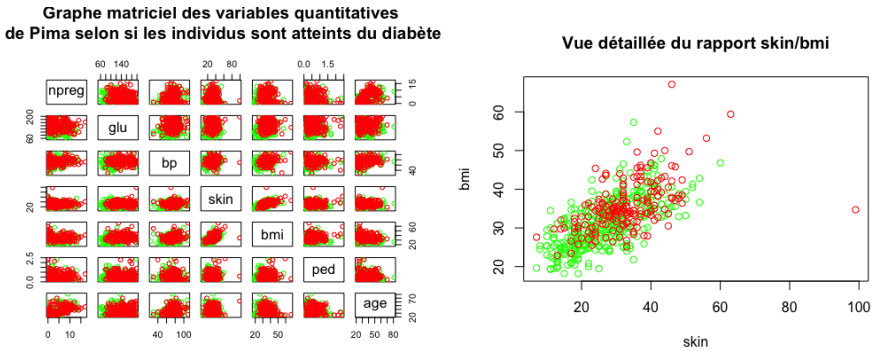

FIGURE 11 – Graphes matriciels des données PIMA

diabétiques (ces dernières sont plus âgés, on un IMC plus élevé... que les femmes non atteinte du diabète). Cependant pour quelques unes de ces variables, nous observons que les deux boites ne se chevauchent que très peu, marquant ainsi une différence plus poussée. C'est le cas de l'âge, de l'IMC, du nombre de grossesses et du taux de glucose plasmatique.

1.3.2 Peut-on déterminer si un individu est diabétique en fonction des facteurs donnés ?

Nous allons tenter de répondre à cette problématique, pour cela, nous pouvons tracer le graphe matriciel comparant les valeurs des sept variables quantitatives données, selon si un individu est diabétique ou non.

Nous observons sur ce graphe que les données semblent peu correllées, en effet, les nuages de points sont très diffus et superposés pour les individus malades ou non. Nous distinguons tout de même que deux variables qui tendent le plus à ressembler à une droite lorsqu'on représente leur ratio, ce sont "skin" et "bmi".

Nous avons donc le graphique à droite plus détaillé représentant ce ratio. Nous en déduisons que ces deux variables sont parmis les plus corrélées des données PIMA.

1.3.3 Étude de la corrélation entre les différentes variables quantitatives des données PIMA

Nous pouvons étudier le tableau de corrélation donné table 3. Il permet de voir que les couple de variables âge/npreg et bmi/skin sont les plus variables les plus corrélées (0,64 de corrélation) alors que les autres ne dépassent pas une corrélation de 0,35.

	npreg	glu	bp	skin	bmi	ped	age
npreg	1.00	0.12	0.20	0.09	0.00	0.00	0.64
glu	0.12	1.00	0.21	0.22	0.24	0.16	0.27
bp	0.20	0.21	1.00	0.22	0.30	0.00	0.34
skin	0.09	0.22	0.22	1.00	0.64	0.11	0.16
bmi	0.00	0.24	0.30	0.64	1.00	0.15	0.07
ped	0.00	0.16	0.00	0.11	0.15	1.00	0.07
age	0.64	0.27	0.34	0.16	0.07	0.07	1.00

TABLE 3 – Corrélation entre les variables qualitatives des données Pima

	npreg	glu	bp	skin	bmi	ped	age
$var_{non-diabetique}$	7.76	589.85	141.68	101.61	42.86	0.089	98.07
$var_{diabetique}$	15.35	977.50	156.84	108.05	43.71	0.159	117.44
Student (pvalue)	1.61e-07	2.2e-16	3.10e-05	4.86e-09	2.88e-12	9.26e-07	1.05e-12

TABLE 4 – Variances des données Pima, et test de Student

1.3.4 Étude de l'influence du facteur diabète sur les autres variables

Puisqu'il ne nous est pas possible de relever l'influence des variables qualitatives les unes entre elles, nous allons nous intéeresser à l'influence du facteur diabète sur ces variables. Tout d'abord nous calculons les variance intra-groupe de ces variables nous donnant la table 4. Nous avons ensuite réalisé le test de student pour chacune de ces variables. Nous pouvons observer que les p-value obtenue sont toujours largement inférieures à 0,05, donc l'hypothèse d'homogénéité entre les deux jeux de données (diabétiques et non-diabétiques) est rejetée.

Nous pouvons en concluer que le facteur diabète a un rôle non négligeable sur l'ensemble des variables quantitatives étudiées pour les données Pima.

2 Analyse en composantes principales

2.1 Exercice théorique

Nous commençons cet exercice par un rappel des données qu'il nous est demandé de traiter par l'analyse en composantes principales. Ces données se retrouvent dans la table 5.

2.1.1 Axes factoriels de l'ACP

Les moyennes des quatre variables sont respectivement 10.71, 4.05, 12.15 et 4.51. Nous allons soustraire ces valeurs à l'ensemble des données des colonnes auxquelles elles correspondent afin d'obtenir les resultats contenus dans la table 6. La somme de chaque colonne vaut maintenant 0.

correcteur	moy.median	std.median	moy.final	std.final
Cor1	10.70	3.900	10.94	4.58
Cor4	10.23	3.043	13.43	4.34
Cor5	10.97	4.413	11.82	3.97
Cor6	11.50	4.303	13.41	4.87
Cor7	10.12	4.030	11.90	4.44
Cor8	10.74	4.646	11.39	4.87

TABLE 5 – Données à traiter dans l'exercice théorique

Cor	moy.median	std.median	moy.final	std.final
Cor1	-0.00	-0.15	-1.21	0.06
Cor4	-0.47	-1.01	1.28	-0.17
Cor5	0.26	0.35	-0.32	-0.54
Cor6	0.78	0.24	1.26	0.36
Cor7	-0.59	-0.02	-0.24	-0.07
Cor8	0.02	0.58	-0.75	0.35

TABLE 6 – Valeurs avec les moyennes soustraites

Nous pouvons ensuite en déduire la matrice de variance, avec la formule :

$$V = X^T D_p X = \frac{1}{9} X^T X$$

. Cette matrice est contenue dans la table 7.

En diagonalisant la matrice de variance obtenue nous determinons les valeurs propres : qui rangées par ordre décroissant sont :

$\lambda_1 = 0.9799$, $\lambda_2 = 0.367$, $\lambda_3 = 0.0831$ et $\lambda_4 = 0.0520$

Les vecteurs propres quand à eux sont (on les appelle aussi axes principaux d'inertie) :

$$u_1 = \begin{pmatrix} -0.03682 \\ 0.29417 \\ -0.95504 \\ -0.00056 \end{pmatrix}, \; u_2 = \begin{pmatrix} -0.704 \\ -0.646 \\ -0.171 \\ -0.236 \end{pmatrix}, \; u_3 = \begin{pmatrix} -0.2332 \\ -0.0942 \\ -0.0206 \\ 0.9676 \end{pmatrix}, \; u_4 = \begin{pmatrix} 0.6694 \\ -0.6972 \\ -0.2406 \\ 0.0883 \end{pmatrix}$$

	moy.median	std.median	moy.final	std.final
moy.median	0.253			
std.median	0.161	0.317		
moy.final	0.085	-0.270	1.089	
std.final	0.054	0.054	0.015	0.118

TABLE 7 – Matrice de variance

2.1.2 Calcul des composantes principales

Maintenant que nous avons déterminé les valeurs et vecteurs propres (les axes principaux d'inertie), nous pouvons calculer la matrice des composantes principales (cf table 8) grâce à la formule suivante :

$$C = XMU = XU$$

Nous en déduiso s ensuite la représentation des six individus dans le premier plan factoriel (cf figure 12

Cor	composante 1	composante 2	composante 3	composante 4
Cor1	1.113	0.297	0.106	0.402
Cor4	-1.504	0.813	0.014	0.061
Cor5	0.404	-0.233	-0.614	-0.041
Cor6	-1.161	-1.015	0.117	0.082
Cor7	0.252	0.492	0.077	-0.324
Cor8	0.895	-0.353	0.299	-0.180

TABLE 8 – Matrice des composantes principales

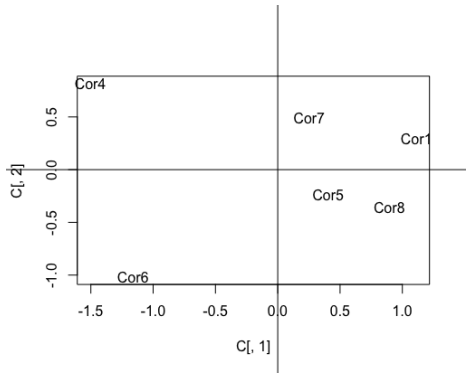

FIGURE 12 – Représentation des six individus dans le premier plan factoriel

2.1.3 Représentation graphique des quatre variables

Nous pouvons ainsi représenter graphiquement les quatre variables (moy.median, std.median, moy.final et std.final), comme indiqué sur la figure 13.

2.1.4 Calcul de somme

Nous nous intéressons maintenant au calcul de la somme suivante :
$\sum_{\alpha=1}^{k} \mathbf{c}_\alpha \mathbf{u'}_\alpha$
Les résultats pour $k = 1..4$ sont :

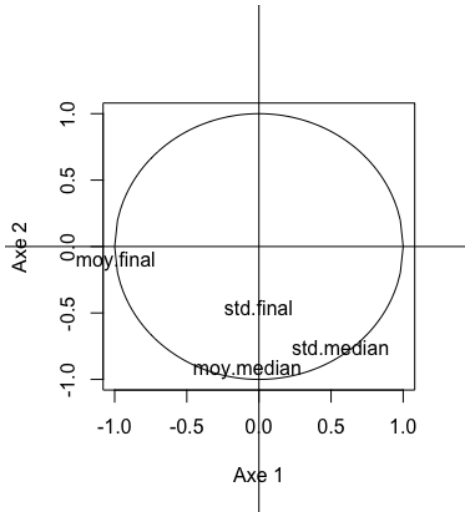

FIGURE 13 – Représentation graphique des quatre variables

$$k=1: \begin{pmatrix} -0.04099127 & 0.3274542 & -1.0630852 & -0.0006324112 \\ 0.05539082 & -0.4424834 & 1.4365293 & 0.0008545667 \\ -0.01489742 & 0.1190064 & -0.3863562 & -0.0002298367 \\ 0.04276532 & -0.3416260 & 1.1090942 & 0.0006597812 \\ -0.00928236 & 0.0741511 & -0.2407327 & -0.0001432078 \\ -0.03298509 & 0.2634977 & -0.8554494 & -0.0005088922 \end{pmatrix}$$

$$k=2: \begin{pmatrix} -0.2503969 & 0.1351160 & -1.1142134 & -0.07092998 \\ -0.5174784 & -0.9686613 & 1.2966582 & -0.19145790 \\ 0.1498011 & 0.2702813 & -0.3461436 & 0.05505954 \\ 0.7582837 & 0.3155745 & 1.2837944 & 0.24085963 \\ -0.3564375 & -0.2447094 & -0.3254937 & -0.11668334 \\ 0.2162280 & 0.4923988 & -0.7946018 & 0.08315205 \end{pmatrix}$$

$$k=3: \begin{pmatrix} -0.2752941 & 0.1250546 & -1.1164140 & 0.03236433 \\ -0.5207800 & -0.9699955 & 1.2963664 & -0.17776022 \\ 0.2932238 & 0.3282414 & -0.3334668 & -0.53997649 \\ 0.7309018 & 0.3045089 & 1.2813742 & 0.35446240 \\ -0.3744752 & -0.2519988 & -0.3270880 & -0.04184795 \\ 0.1464238 & 0.4641895 & -0.8007716 & 0.37275794 \end{pmatrix}$$

$$k=4: \begin{pmatrix} -0.005844671 & -0.15555434 & -1.2132587 & 0.06791405 \\ -0.479484127 & -1.01300171 & 1.2815239 & -0.17231187 \\ 0.265413832 & 0.35720310 & -0.3234715 & -0.54364559 \\ 0.785821995 & 0.24731415 & 1.2616349 & 0.36170828 \\ -0.591729025 & -0.02574726 & -0.2490034 & -0.07051126 \\ 0.025821995 & 0.58978605 & -0.7574254 & 0.35684639 \end{pmatrix}$$

Nous pouvons observer que la somme quand $k = 4$ est égale à la matrice des composantes principales (cf. table 8).

2.1.5 Prises en compte des individus écartés

Afin de prendre en compte les données manquantes des correcteurs 2 et 3, nous remplaçons les valeurs manquantes par la moyenne de la colonne. Nous obtenons ainsi deux nouveaux graphes d'ACP (cf figure 14).

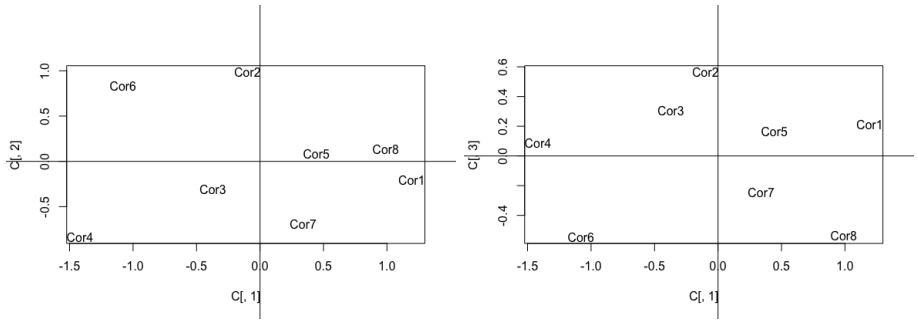

FIGURE 14 – ACP du jeu de données vu en cours, après prise en compte des données manquantes

2.2 Utilisation des outils R

Nous cherchons dans cet exercice à se familiariser avec l'utilisation des outils R liés à l'analyse en composantes principales. Pour cela nous allons nous intéresser au jeu de données notes vu en cours (et présent dans la table 9)

	math	scie	fran	lati	dm
jean	6.0	6.0	5.0	5.5	8
aline	8.0	8.0	8.0	8.0	9
annie	6.0	7.0	11.0	9.5	11
monique	14.5	14.5	15.5	15.0	8
didier	14.0	14.0	12.0	12.5	10
andré	11.0	10.0	5.5	7.0	13
pierre	5.5	7.0	14.0	11.5	10
brigitte	13.0	12.5	8.5	9.5	12
evelyne	9.0	9.5	12.5	12.0	18

TABLE 9 – Données "notes" extraites du cours

Pour effectuer l'ACP sur ces données, on va commencer par une première commande `acpnote = princomp` qui permet de traiter les données.

À partir des données récupérées à la suite de cette commande, il existe un certain nombre de commandes permettant de les représenter :

`summary(acpnote)` : retourne pour chacune des composantess son écart-type

`loadings(acpnote)` : retourne la matrice des vecteurs propres, appelée aussi matrice de rotation

`acpnote$scores` : donne pour chacune des données initiales son emplacement dans l'espace factoriel

`plot(acpnote)` : produit le diagramme en barres des composantes principales (écart-types) (cf figure 15 à gauche)

`biplot(acpnote)` : projette les données initiales dans un plan contenant aussi les axes principaux (cf figure 15 à droite). Plus précisément, cette fonction a été redéfinie pour l'étude d'une ACP, en effet, ne sont représentés ici que deux des composantes (les deux principales par défaut).

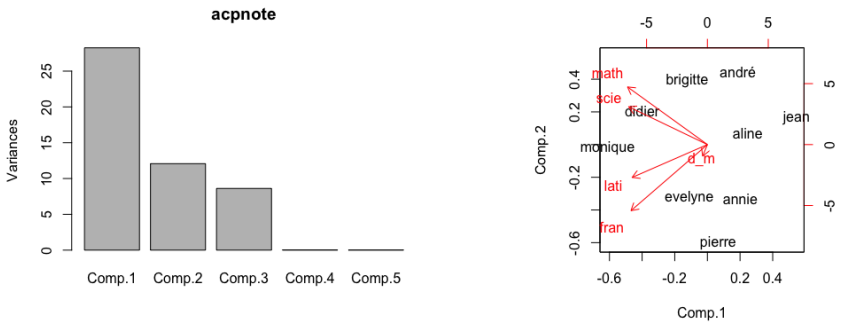

FIGURE 15 – ACP du jeu de données vu en cours

2.3 Données crabs

2.3.1 ACP sans traitement

Nous allons commencer par effectuer une analyse en composantes principales des données brutes du jeu de données crabs. Celle-ci nous permet d'obtenir les graphiques présentés dans la figure 16. Nous observons tout d'abord que la première composante est très élevée et que le graphique généré par la fonction biplot n'est pas lisible. Cela s'explique car les variables sont très corrélées.

2.3.2 Traitement des données pour une ACP pour explicite

Comme expliqué précedement, les données sont très corrélées. Dans l'analyse descriptive, nous avions d'ailleurs choisi de décorréler ces variables en éliminant CL (c'est à dire disant les autres variables par sa valeur). Nous avons donc choisi de traiter les données en faisant le même traitement avant d'effectuer l'analyse en composantes principales.

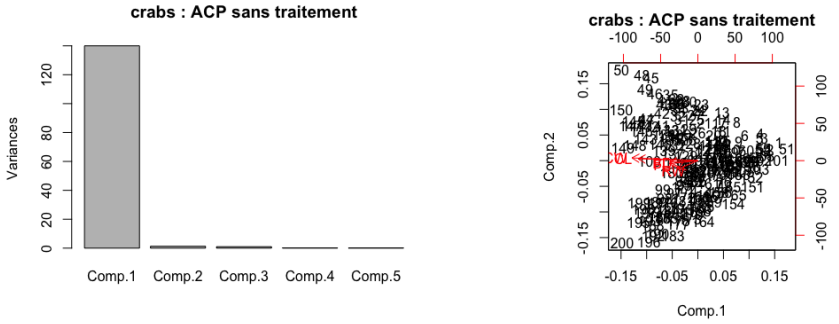

FIGURE 16 – ACP sans traitement des données crabs

Nous obtenons maintenant ces graphiques qui sont, d'une par plus lisible et a fortiori, si nous les colorons selon le sexe ou l'espèce nous permettent de distinguer quatre groupes distincts (cf. figure 18) :

Triangle bleu : femelle bleue,

Triangle orange : femelle orange,

Rond bleu : mâle bleu et

Rond orange : mâle orange

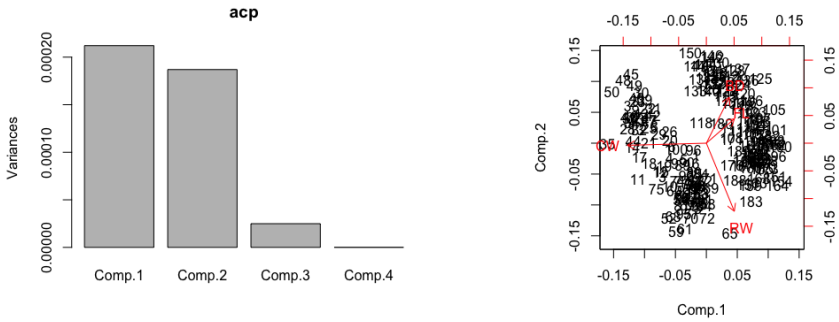

FIGURE 17 – ACP après traitement des données crabs

2.4 Données Pima

Nous allons tenter l'analyse en composantes principales des données Pima afin de distinguer de manière visuelle les groupes d'individus diabétiques et non-diabétiques. Comme nous pouvons le voir sur la figure 19 les points se regroupent en un amas au centre du graphe, et il nous est donc impossible de distinguer clairement deux groupes.

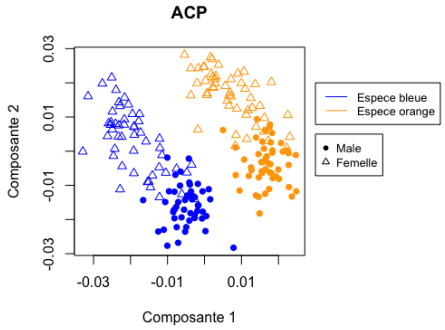

FIGURE 18 – ACP colorée des données crabs

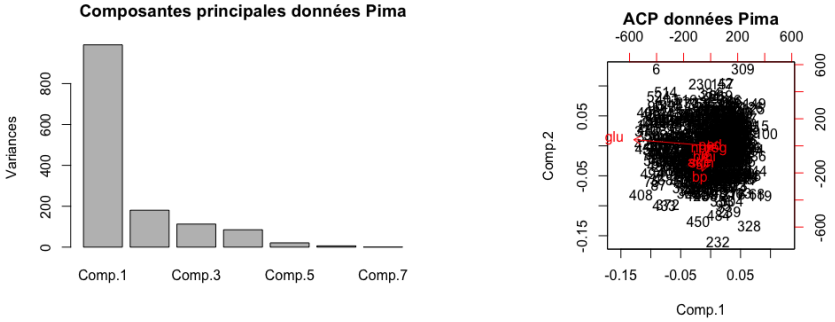

FIGURE 19 – ACP des données pima

Cependant, si nous colorons le graphe selon les individus diabétiques, on obtient la figure 20, et nous pouvons distinguer que les individus s'orientent selon l'axe de la composante glucose. En effet, les femmes atteintes du diabète sont majoritairement à gauche du graphe et les femmes saines sont à droite. Nous en déduisons, que le taux plasmatique de glucose permet d'identifier plus facilement les femmes atteintes de diabète. Plus celui-ci est élevé, plus l'individus aura une probabilité élevée d'être diabétique.

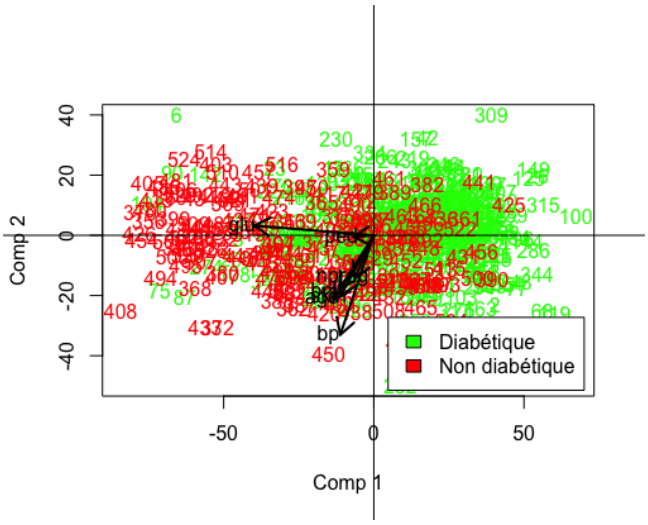

FIGURE 20 – ACP colorée des données PIMA