Fiche bilan SAE

Nom de la SAE	**Projet Classification**		semestre / Période	S4
volume horaire consacré par l'étudiant	avec enseignant	6h	en autonomie	4h
coéquipiers :	Sami Said			

Sujet spécifique	**Classification d'élément projet finale**
Objectifs	Analyse du jeu de données Classification et Interpétation Importer et Comprendre les Données Examiner les relations entre les variables Modéliser les relations à travers la régression Linéaire Multiple Optimiser et évaluer le modèle

Livrables	Chargement du fichier
	`foot <- read.csv("D:/Classification/Foot.csv",sep=";",dec= ".",header = TRUE, encoding = "latin1")`
	Tout d'abord, nous avons supprimé les doublons car cela faisait parfois apparaître un même joueur plusieurs fois. Ensuite, nous avons modifié la colonne correspondant aux noms des joueurs. Ils sont désormais assignés comme noms de lignes, afin de faciliter l'identification des lignes par le nom du joueur au lieu d'un index numérique. Ensuite, nous avons retiré la première colonne du jeu de données, qui correspondait aux postes des joueurs. Nous avons filtré le jeu de données pour qu'il ne reste que les joueurs ayant joué plus de 5 matchs. Ensuite, nous avons réduit le nombre de postes pour n'en garder que 4, à savoir Gardien (GK), Défenseur (DF), Milieu de Terrain (MF) et Attaquant (FW).
	`foot = foot %>% distinct(Player,.keep_all=T)`
	`row.names(foot) = foot$Player`
	`foot = subset(foot , select= -c(1))`
	`summary(foot)`
	`##  Pos  Squad  Age  MP`
	`##  Length:384  Length:384  Min. :16.00  Min. : 1.00`
	`##  Class :character  Class :character  1st Qu.:23.00  1st Qu.: 7.00`
	`##  Mode :character  Mode :character  Median :26.00  Median :14.00`
	`##  Mean :26.24  Mean :12.46`
	`##  3rd Qu.:29.00  3rd Qu.:18.00`
	`##  Max. :38.00  Max. :23.00`
	`##  Min  Goals  Shots  SoT`
	`##  Min. : 3.0  Min. : 0.000  Min. : 0.0000  Min. : 0.0000`
	`##  1st Qu.: 285.8  1st Qu.: 0.000  1st Qu.: 0.4475  1st Qu.: 0.0000`
	`##  Median : 765.0  Median : 0.000  Median : 1.0500  Median : 0.2750`

Fiche bilan SAE

Bilan de la SAE
(reproduire le tableau autant de fois que de compétences mobilisées dans la SAÉ)

Compétence	**Analyser statistiquement les données**
Apprentissages critiques sollicités	Comprendre l'intérêt des analyses multivariées pour synthétiser et résumer l'information portée par plusieurs variables
	Apprécier les limites de validité et les conditions d'application d'une analyse
	Prendre conscience de la différence entre modélisation statistique et analyse exploratoire
Composantes essentielles à respecter	En tenant compte du contexte de l'étude (économique, socio-démographique, commerciale, clinique...)
	En mettant en évidence les grandes tendances et les informations principales
	En tenant compte du contexte inférentiel (variabilité de l'échantillon)

Ma démarche

Savoirs / connaissances	Savoir-faire	Savoir-être
Réaliser et interpréter une ACP Déterminer et analyser les clusters	Analyse Interprétation Explication	Travail d'équipe

Evaluation du résultat

- Ce que je trouve bien réalisé, pourquoi ?

Fiche bilan SAE

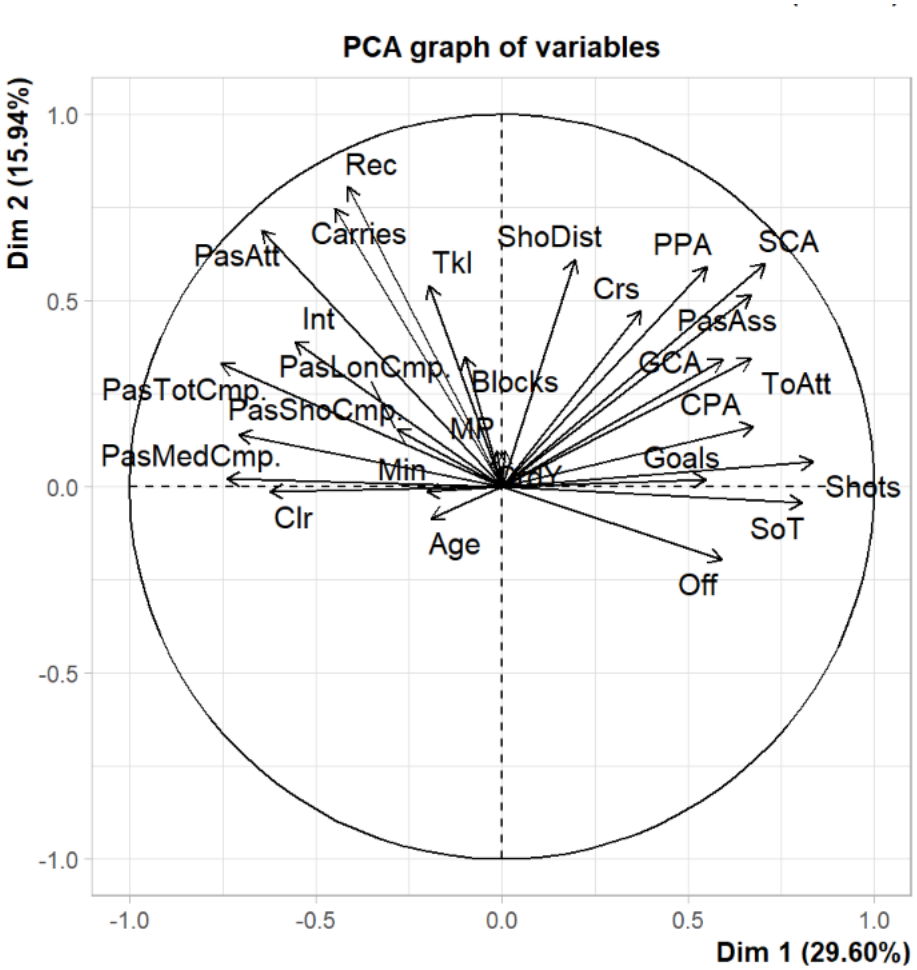
L’ACP est son analyse du a son contexte a été très bien réalisé.

- Ce que je n’ai pas bien compris ; ce qui serait à améliorer pour une prochaine fois : pourquoi ? comment ?

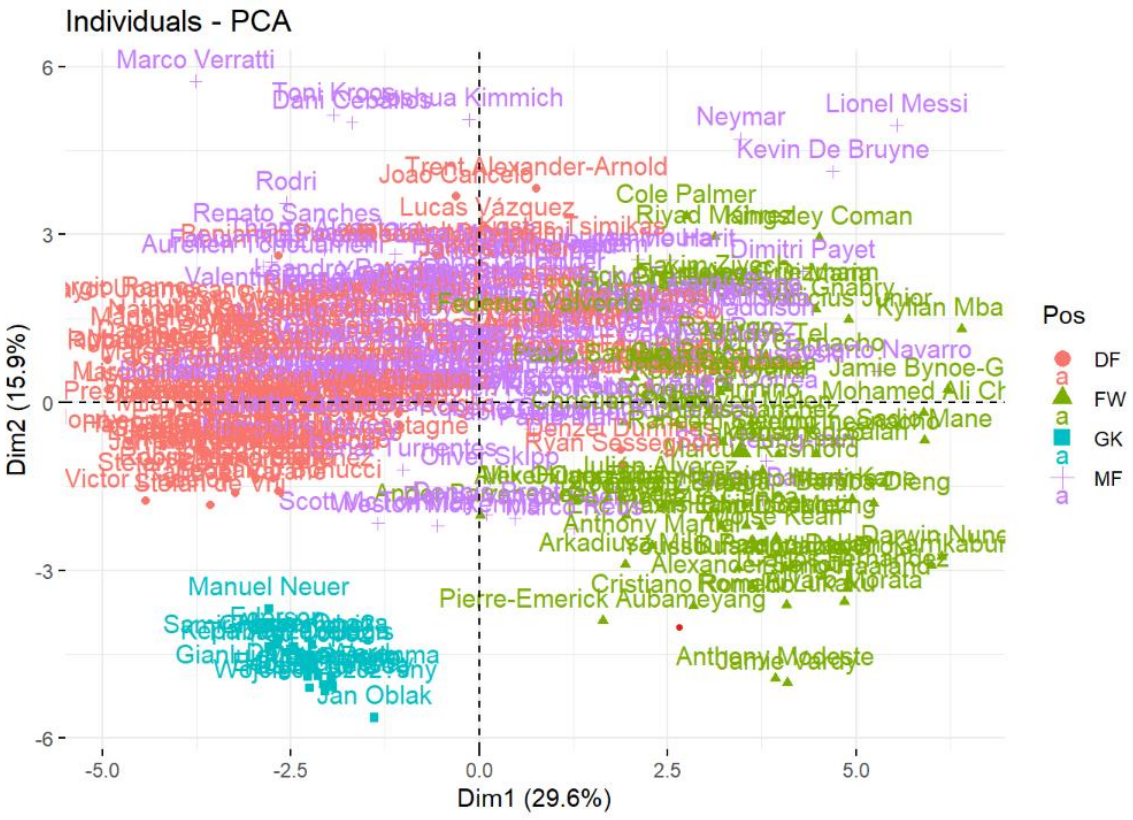
  Ce qui serait à améliorer pour la prochaine fois serait l’analyse et l’interprétation des clusters.

Lors de cette SAE, l’objectif était de classifier des éléments à l’aide d’une ACP et de clusters. On nous a fourni une base de données très vaste sur le football. Au début, j’ai rencontré certaines difficultés car la base de données était très complexe avec énormément de joueurs et de variables, donc il a fallu en supprimer certaines et ne pas prendre en compte certains joueurs (dans notre cas, ceux qui avaient joué moins de 5 matchs). Ensuite, nous avons obtenu deux graphiques que je vous joindrai en preuve, qui nous ont permis l’analyse finale. Ce sont deux cercles de corrélation qui expliquent les performances des joueurs en fonction des variables.

Eléments de preuve, ce que je peux montrer
(Choisir des éléments précis à mettre annexe)
1)


PCA graph of variables
Dim 2 (15.94%)
Dim 1 (29.60%)
Rec
Carries
PasAtt
Tkl
ShoDist
PPA
SCA
Crs
Int
PasAss
PasLonCmp.
GCA
Blocks
ToAtt
PasTotCmp.
PasShoCmp.
CPA
MP
PasMedCmp.
Min
Goals
Shots
Clr
SoT
Age
Off

2)


Individuals - PCA
Dim2 (15.9%)
Dim1 (29.6%)
Pos
DF
FW
GK
MF
Marco Verratti
Neymar
Lionel Messi
Kevin De Bruyne
Rodri
Trent Alexander-Arnold
Cole Palmer
Renato Sanches
Lucas Vázquez
Dimitri Payet
Federico Valverde
Kylian Mba
Jamie Bynoe-G
Manuel Neuer
Pierre-Emerick Aubameyang
Anthony Modeste
Jan Oblak

Fiche bilan SAE

Tester au risque 5

```
m1 <- lm(Goals~1, data)
anova(m1, mod_c2)
```

	Res.Df <dbl>	RSS <dbl>	Df <dbl>	Sum of Sq <dbl>	F <dbl>	Pr(>F) <dbl>
1	21550	251544.4	*NA*	*NA*	*NA*	*NA*
2	21389	244835.5	161	6708.914	3.640344	9.897197e-49

2 rows

Au moins un club a mis plus ou moins de buts par rapport au Milan AC avec une p valeur de 2.2e-16 on rejette H0. La pvaleur extrêmement faible de 2.2e-16 obtenue lors de la comparaison entre le modèle avec seulement l'intercept et le modèle incluant la variable 'Club' en tant que facteur suggère que l'ajout du facteur 'Club' apporte une contribution significative à l'explication de la variance dans les objectifs marqués. Ainsi, nous rejetons l'hypothèse nulle selon laquelle tous les clubs marquent plus ou moins le même nombre de buts par rapport au Milan AC, avec un risque d'erreur de 5%.