01/12/2020

TOP14 Classification des équipes et Identification des facteurs de victoires

Saison 2019-2020

CLADIERE Nathan
OPEN CLASS ROOM DATA ANALYST P8

Table des matières

1 Introduction

1.1 La data dans le sport

Les données sont utilisées dans le sport depuis plus de 30 ans. Cette utilisation de la data s'est fait connaître par le grand publique avec le film « le stratège »(2011) adapté du livre Moneyball de Michael Lewis. Cette histoire vraie illustre comment les données peuvent être utilisées pour gagner dans le baseball.

Depuis ce livre, le sport professionnel utilise de plus en plus les données. Cette augmentation s'illustre par le décuplement des systèmes de collectes :

- Depuis 2013, tous les stades de la NBA sont équipés de système de Tracking.
- Les déplacements et l'activité de chaque joueur sont suivi soit par GPS soit de système RFID.
- Les staffs des équipes professionnelles ont toutes des équipes d'analystes vidéos.
- Les clubs emploient maintenant des data analysts/scientists.

Cet essor des données apporte aujourd'hui au monde professionnel de nombreux avantages :

- Analyse de la performance :
    - Quantification et analyse des points forts et des points faibles
    - Optimisation des temps d'entraînements
- Prévention des blessures :
    - Les données biométriques et physiologiques permettent de déceler l'état de forme
- Elaboration de stratégies :
    - Adapter sa façon de jouer pour gagner, au basket le panier à 3 points est beaucoup plus utilisé depuis que les données ont relevé son importance
- Repérage de jeunes talents
- Amélioration de l'expérience des spectateurs :
    - Données en temps réel sur le match, dans le stade ou à la télévision (IBM et SAS, s'emploient pour en fournir encore plus)

L'analyse dans ce rapport se concentrera sur l'approche analyse de la performance/élaboration de stratégie dans l'esprit de la sabermétrie (quelles variables permettent de déterminer le nombre de points marqués, et donc la victoire) dans le rugby.

1.2 La data dans le rugby

Une étude recoupe toutes les études ayant été réalisées de 2007 à 2019(1) visant à déterminer les facteurs de victoires, un total de 41 articles ont été écrits à ce sujet. La plupart des articles se concentrent sur la collecte et l'analyse d'indicateurs de performances. Peu d'études se consacrent aux contextes du match (domicile, extérieur, enjeux, style d'opposition, météo).

Vingt-neuf indicateurs de performances sont utilisés à travers toutes les études et seulement quelques-uns sont communs :

- Jeu au pied
- Contre d'une touche adverse
- Essais marqués
- Points marqués au pied
- Plaquages effectués
- Turnover acquis

La dernière étude en date (2) se focalise sur la saison 2016-17 et une partie de la saison 2017-18 sur les matchs de la Premiership anglaise soit en tout 132 matchs.

La modélisation de la victoire dans cette étude s'est basée sur des données mises en forme différemment :

- Les données isolées : mesurées dans les matchs (400 m parcourus, 130 plaquages, etc.).
- Les données relatives : données par rapport à l'opposition, si l'équipe A a parcouru 300 m et l'équipe B 230 m, les données seront donc, équipe A : +70 et équipe B : -70

Ces données relatives se sont montrées plus significatives dans la précision de la modélisation (prédiction de victoire).

2 Problématiques

Trois points vont être abordés par rapport aux équipes de Top14, sur la saison 2019-20 :

1. Par rapport aux manques de contextes sur les équipes (1), peut-on établir des types de jeux ?
    a. Les équipes, sont-elles groupables par style de jeu ?
    b. Les équipes, gardent-elles les mêmes caractéristiques avec les données relatives ?
2. En reprenant le principe de la deuxième étude (2), deux modèles de prédictions seront utilisés afin de savoir :
    a. Quelle est la précision des modèles ? Quel modèle est le plus précis ?
    b. Est-ce que la précision de la modélisation change avec les données relatives ?
    c. Quels sont les facteurs de victoires pour une équipe de Top14 ?

Pour la deuxième partie, l'étude sera menée avec les points, sans les points et sans les facteurs de points (essais, transformations, et pénalités) car ces derniers prennent une importance inéluctable dans la modélisation avec les données relatives.

3 Méthode

3.1 Démarche et outils

Extraction des données

Extraction de tous les matchs de la saison 2019-2020 en format csv

Mise en forme des données

Mise en forme des données avec la librairie pandas de python

Base de données (119 matchs) :

- Données isolées
- Données relatives

Type de Jeu

Regroupement

Moyenne par équipe dans chaque match

1. Centrage et réduction des valeurs

Unité commune pour toutes les variables

2. Clustérisation par Kmeans

Regroupement des équipes ayant des caractéristiques similaires

3. Analyse en composantes principales

Analyse des caractéristiques des équipes

4. Comparaison des données

Comparaison des données isolées et relatives

plotly

Illustration

Modélisation

1. Centrage et réduction des valeurs

Unité commune pour toutes les variables

Permet d'utiliser les coefficients des modélisations comme facteur d'importance des variables

2. Première modélisation

Modélisation avec toutes les variables

Deux modèles sont comparés avec leurs précisions et leurs F1 score :

- Regression logistique
- RandomForest

3. RFECV

Sélections des variables utiles

4. Deuxième modélisation

Comparaison des deux modèles

3.2 Les variables étudiées

Tableau d'explication des variables utilisées pour définir le type d'équipe

Nom de la variable	Explication de la variable
Equipe	Nom de l'équipe
Balles portées par les trois-quarts Balles portées par les troisièmes lignes Balles portées par le cinq de devant	Utilisations des différents postes en attaque
Jeu au pied Passes Passes après contact Nombre de rucks	Style de jeu avec la balle
Mètres Plaquages Balles portées Temps de possession	Désir de posséder ou non la balle
Nombre de participants aux rucks offensifs Nombre de participants aux rucks Défensifs	Rapport aux phases de conservations

Liste des variables utilisées pour les modélisations (en plus des variables précédemment citées):

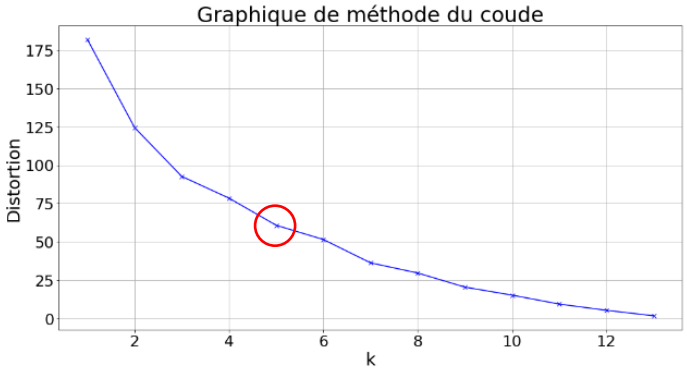
- Pour chaque type de poste :
  - Nombre de balles portées
  - Nombre de défenseurs battus
  - Nombre de mètres
  - Nombre de plaquages manqués
  - Nombre de passes après contact
  - Nombre de pénalités concédées
  - Nombre de plaquages
- Nombre de franchissements
- Nombre de transformations
- Nombre d'essais
- Nombre de défenseurs battus
- Nombre de plaquages manqués
- Nombre de pénalités transformées
- Nombre de pénalités défensives concédées
- Nombre de pénalités offensives concédées
- Nombre de pénalités concédées dans son camp
- Nombre de lancés en touche perdus
- Nombre de lancés en touche gagnés
- Ratio de touches gagnées
- Nombre de turnovers concédés
- Nombre de turnovers acquis
- Nombre de mêlées
- Ratio de gain de mêlées
- Ratio de rucks gagnés
- Temps de jeu effectifs
- Ratio des coups de pied de conversions

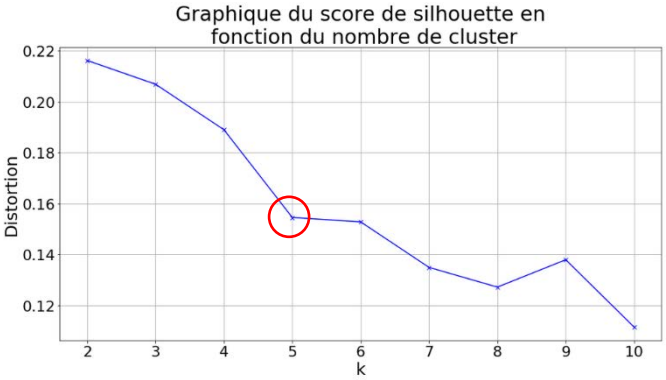
4 Quels sont les types d'équipes

4.1 Regroupement des équipes

Le choix du nombre de cluster/regroupement se fait par la méthode du coude sur un graphique de distorsion pour minimiser la distance intra classe. Cette décision est complétée par l'étude du coefficient de silhouette, devant être minimal pour obtenir le nombre de clusters (k) optimal.

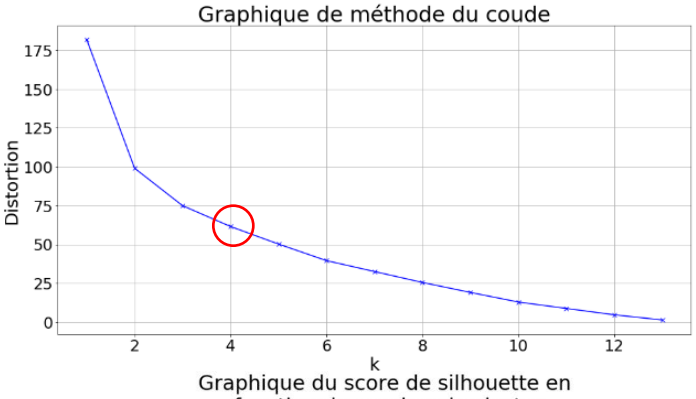
Données isolées

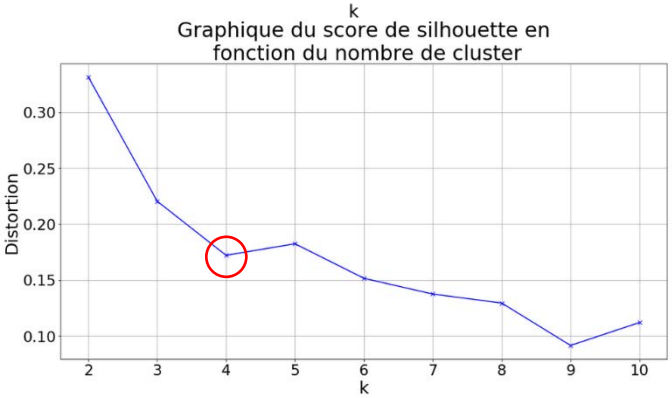




Pour les données isolées on choisira donc 5 Clusters

Données Relatives





Pour les données on choisira 4 clusters

	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster4
Données isolées	• **SUA** • **CAB** • **CA** • MHR • SF • RCT	• **AB** • SR • **LOU** • SP • **ST**	• **UBB**	• **R92**	• ASM
Données relatives	• **SUA** • **CAB** • **CA**	• **ST** • **LOU** • **AB**	• **UBB** • ASM • SR	• MHR • SP • **R92** • SF • RCT	

Tableau de classification des équipes

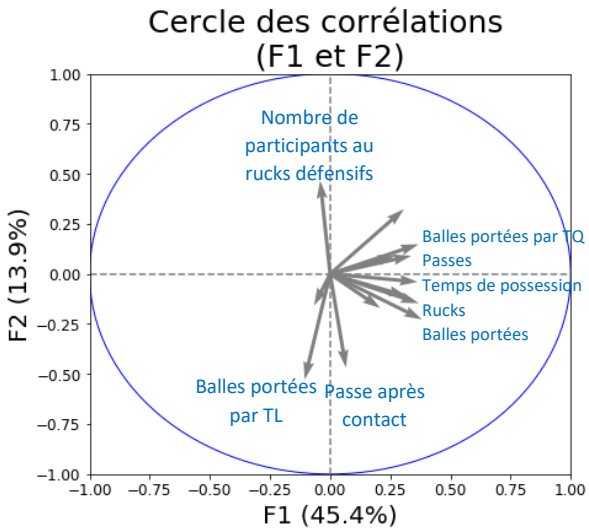
Il y a 7 clubs qui changent de regroupement avec les données relatives, 3 d'entre eux restent cependant dans un même cluster.

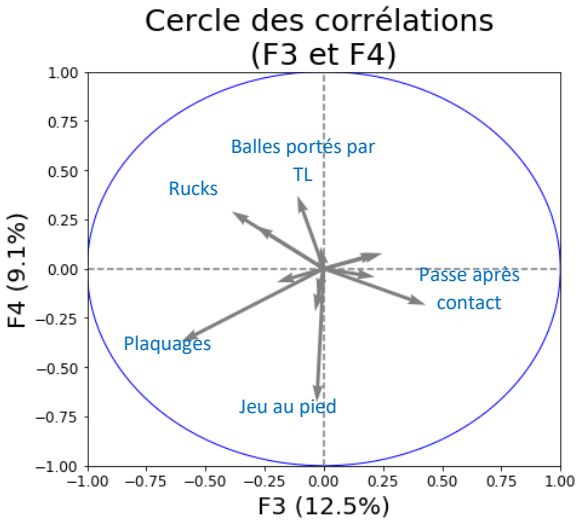
Il faut donc maintenant identifier les caractéristiques de ces clusters et identifier les différences de regroupement avec les deux types de données.

4.2 Plus de lisibilité avec l'analyse en composantes principales

L'analyse en composantes principales permet de regrouper plusieurs variables entre elles. L'étude du cercle de corrélations et l'analyse de la valeur des variables sur les composantes servent à définir les composantes.

4.2.1 Analyse des données isolées





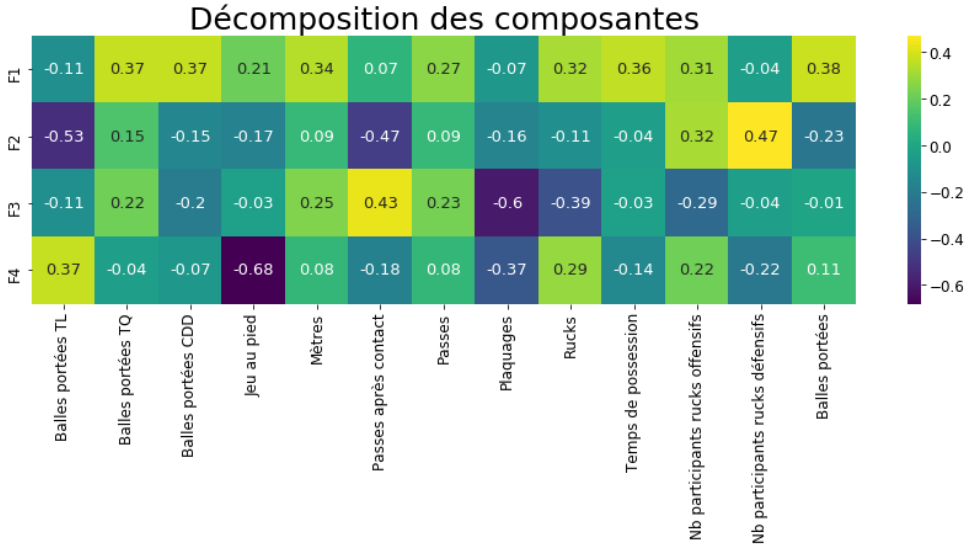
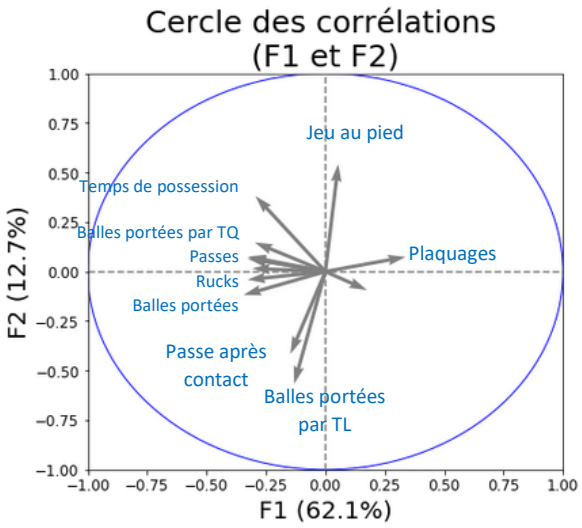


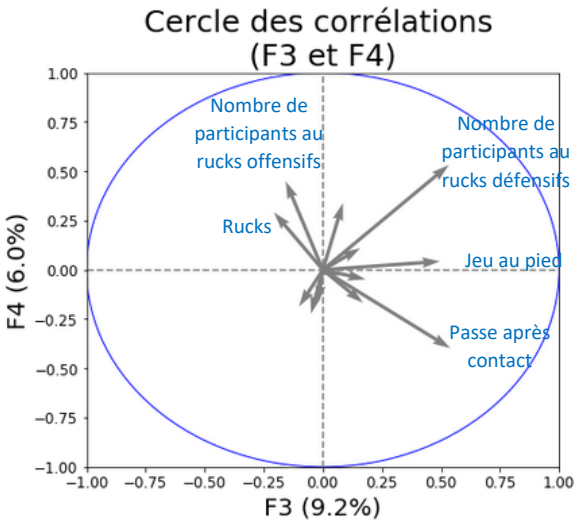
Tableau d'analyse des composantes

Composantes	Définition	Variables associées
F1	Désir de développer du volume de jeux	Balles portées par les ¾ ; Balles portées par le cinq de devant ; Mètres ; Passes ; Rucks ; Nombre de participants aux rucks offensif ; Balles ;Temps de possession
F2	Jeux dans la défense (inversement) et intensité dans les rucks défensifs	Inversement : Balles portés par les troisièmes lignes ; passe après contact ; Positivement : Nombre de participants dans les rucks défensifs
F3	Défend beaucoup	Inversement : Plaquage
F4	Jeu au pied	Inversement : Jeu au pied Positivement : Balles portées par les troisièmes lignes

4.2.2 Analyse des données relatives

Même variables dans la dimension de faire plus que l'adversaire





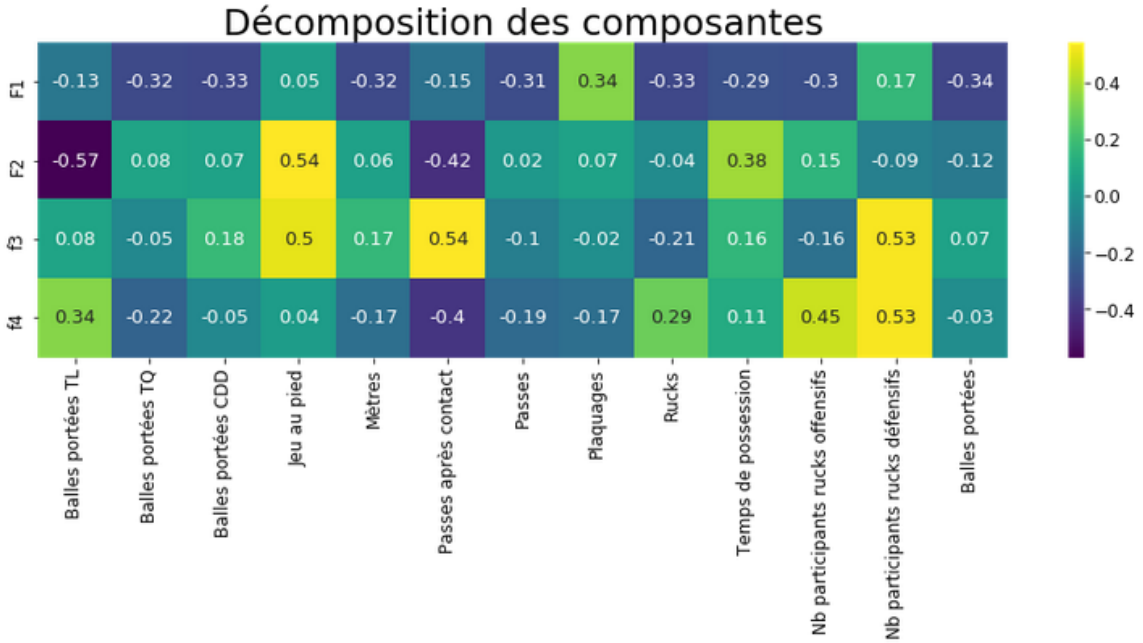
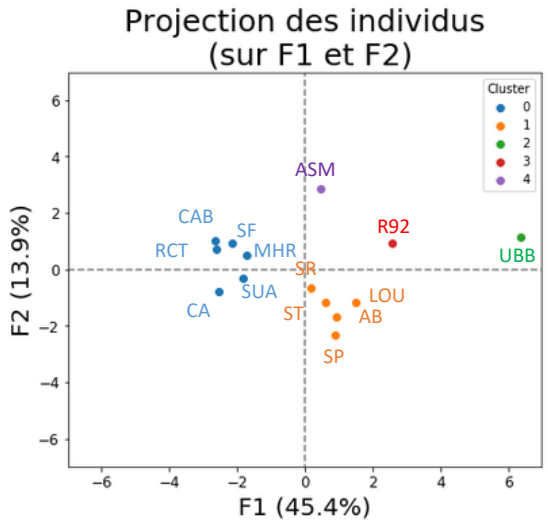
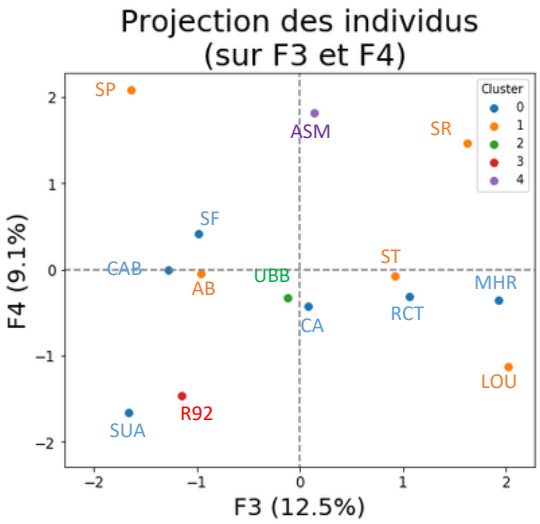


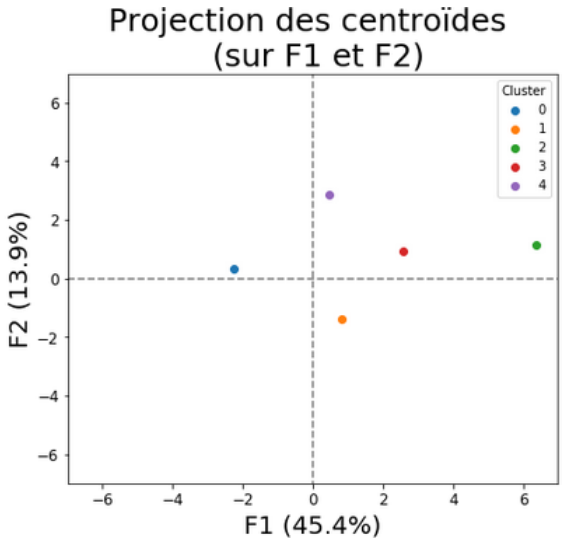
Tableau d'analyses des composantes

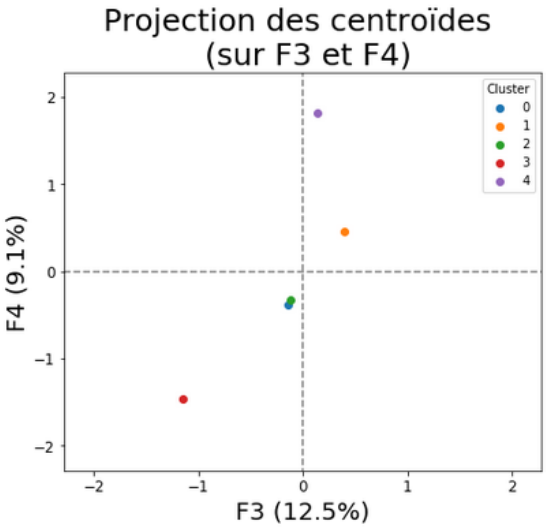
Composantes	Définition	Variables associées
F1	Désir ou capacité à développer plus de volume de jeu que l'adversaire	Inversement : Balles portées par les ¾ ; Balles portées par le cinq de devant ; Mètres ; Passes ; Nombre de rucks ; Temps de possession ; Nombre de participants aux rucks offensifs, Balles portées, Positivement : Plaquage
F2	Utilisation des troisièmes lignes ou du jeu au pied	Inversement : Balles portés par les troisièmes lignes, Passes après contact Positivement : jeu au pied
F3	Redondance	Positivement : Jeu au pied, Passes après contact, Nombre de participants aux rucks défensifs
F4	Volonté de présence dans les rucks ou de faire vivre la balle	Positivement : Nombre de participants aux rucks défensifs et offensifs, nombre de rucks Inversement : passes après contact

4.3 Description des différents clusters (isolées)









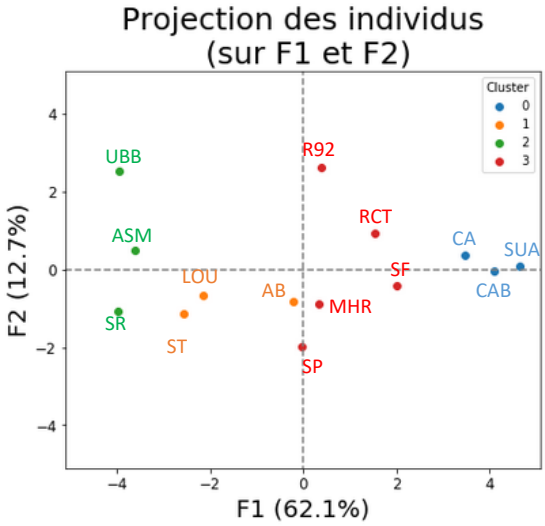
La dissociation des différents clusters se fait essentiellement sur la composante F1, qui montre 45% de l'inertie du nuage de points. Pour rappel, cette composante correspond au désir de développer du volume de jeu. L'UBB influence énormément cet axe et y est donc très bien représenté.

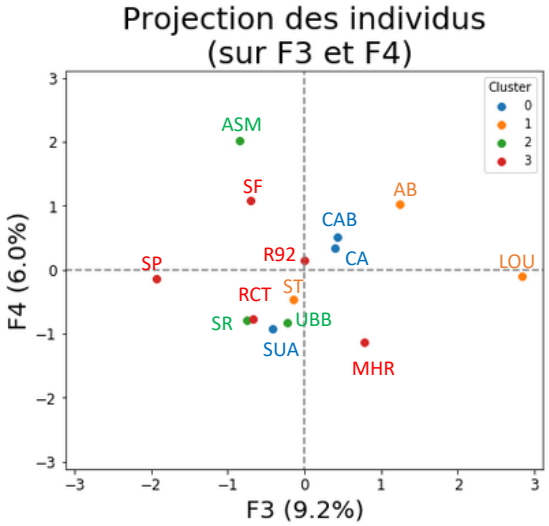
La différence entre les clusters 1,3 et 4 se fait principalement sur la composante F2 qui symbolise l'activité dans les rucks défensifs ou l'ASM met beaucoup d'intensité (57 participants en moyenne) contrairement à la SP (38 participants en moyenne).

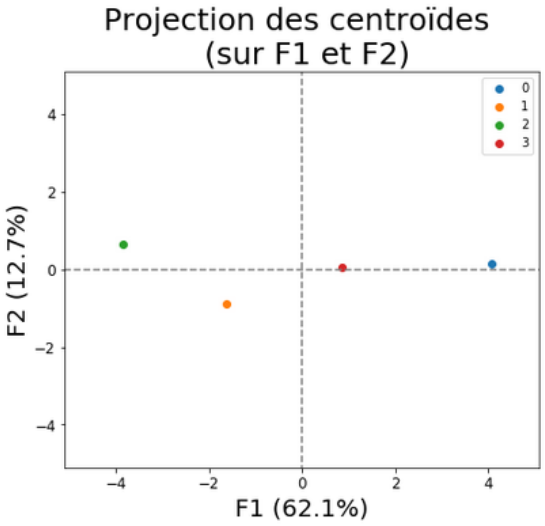
Même si toutes les équipes du cluster 0 ne sont pas très bien représentées sur F1, ce cluster se distingue en développant peu de volume de jeu, comme le RCT qui ne porte que 99 balles en moyenne, contrairement à l'UBB qui porte 121 la balles en moyenne.

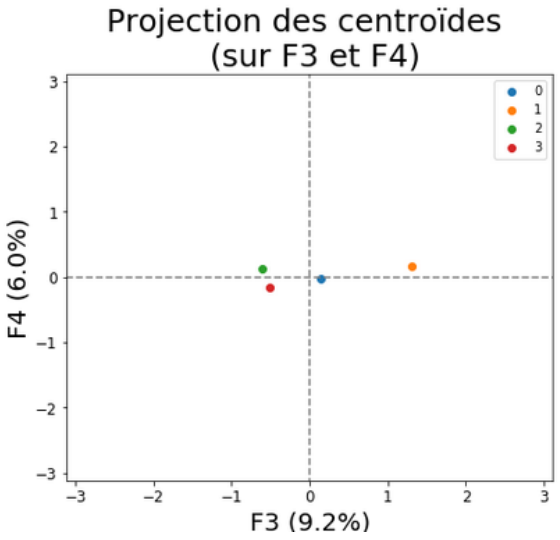
Les composantes F3 et F4, représentant 20% de l'inertie du nuage de point, ne permettent pas une distinction claire des clusters. On peut tout de même voir une différence dans l'utilisation du jeu au pied, notamment l'ASM qui l'utilise peu (18 jeu au pied en moyenne par match) alors que le Racing 92 l'utilise plus (24 jeux au pied en moyenne par match).

4.4 Description des différents clusters (relatives)









La dissociation des clusters se fait aussi essentiellement sur la composante F1 correspondant à la capacité de jouer plus que l'adversaire (62% de l'inertie du nuage de points). Le nombre de balles portées en plus pour le cluster 2 est de 13 en moyenne, tandis que le cluster 0 en porte 15 en moins.

Dans les clusters 1 et 3 la stratégie de contre-attaque face à une récupération d'un jeu au pied est différente, visible avec la composante F2. Le R92 joue au pied 4 fois en plus en moyenne par match, tandis que la SP utilise 5 jeux au pied de moins en moyenne par match.

Les composantes F3 et F4 ne correspondent qu'à 15 % de l'inertie du nuage de points et ne permettent pas de dissocier réellement les clusters.

4.5 Ce qu'il faut retenir des différents clusters

Dans les deux types de données, il est possible d'établir des types de jeux.

La dissociation des clusters se fait essentiellement sur la composante F1, produire du jeu ou produire plus de jeu que l'adversaire.

L'UBB qui produit beaucoup de jeu se retrouve logiquement dans le cluster ou les équipes en produisent plus que l'adversaire, dans le même schéma, on retrouve le SUA, le CAB et le CA qui produisent peu de jeu, donc produisent en général moins que les adversaires.

Le ST, le LOU et l'AB restent dans une volonté de produire du jeu, et légèrement plus que l'adversaire.

Le MHR, la SP et le RCT produisent peu de jeu, mais ils leur arrivent dans produire plus que l'adversaire.

Par contre l'ASM et le SR adaptent plus leurs jeux en fonctions des circonstances (adversaires, lieux, conditions climatique ou autres).

Il y a donc des différences très claires entre certains clubs qui produisent beaucoup de jeu et d'autres moins. Cependant l'établir sur une moyenne est sans doute un peu réducteur.

Il pourrait être intéressant de faire une clusterisation sur les données de l'ensemble des matchs, puis de voir si tous les matchs d'une équipe se retrouve dans un même cluster.

Dans la même logique, on pourrait voir si les équipes jouent de la même façon face au lieu (domicile, extérieur) ou face au type d'adversaire.

5 Facteurs de victoires

5.1 Quel meilleur modèle pour prédire la victoire

Pour chaque modèle, l'ensemble des données a été divisé en deux, une partie pour entraîner les modèles, une autre pour le tester.

Les modèles ont été évalués selon deux mesures :

- La précision : proportion de résultats positifs réels identifiée correctement
- F1 score : moyenne harmonique de la précision et du rappel (proportion de résultats positifs réels identifiée correctement)

Les scores avec les points et facteurs de points ne seront pas pris en compte car trop révélateurs dans la base de données relatives. Ce constat est aussi vrai pour la base de donnée isolée.

Tableau de comparaison des modélisations

Modélisation	Variable initiale	Données utilisées	Précision	F1 score
RF+RFECV	Totalité	Rela	100,00%	1
LR+RFECV	Totalité	Rela	94,94%	0,94
RF	Totalité	Rela	94,94%	0,94
LR+RFECV	Sans points	Rela	92,40%	0,91
LR	Totalité	Rela	91,14%	0,9
RF+RFECV	Sans points	Rela	89,87%	0,89
LR	Sans points	Rela	88,61%	0,87
LR	Sans points et FP	Rela	84,81%	0,83
RF	Sans points	Rela	83,55%	0,92
LR+RFECV	Sans points et FP	Rela	83,54%	0,81
LR+RFECV	Totalité	iso	81,01%	0,79
LR	Totalité	iso	79,75%	0,77
RF	Totalité	iso	79,75%	0,75
RF	Sans points	iso	79,75%	0,74
LR+RFECV	Sans points	iso	78,48%	0,77
LR	Sans points	iso	78,48%	0,76
RF+RFECV	Totalité	iso	77,22%	0,74
RF+RFECV	Sans points	iso	75,95%	0,7
LR	Sans points et FP	iso	73,42%	0,72
RF+RFECV	Sans points et FP	Rela	72,15%	0,7
LR+RFECV	Sans points et FP	iso	69,62%	0,67
RF	Sans points et FP	Rela	67,09%	0,63
RF	Sans points et FP	iso	67,09%	0,59
RF+RFECV	Sans points et FP	iso	64,55%	0,59

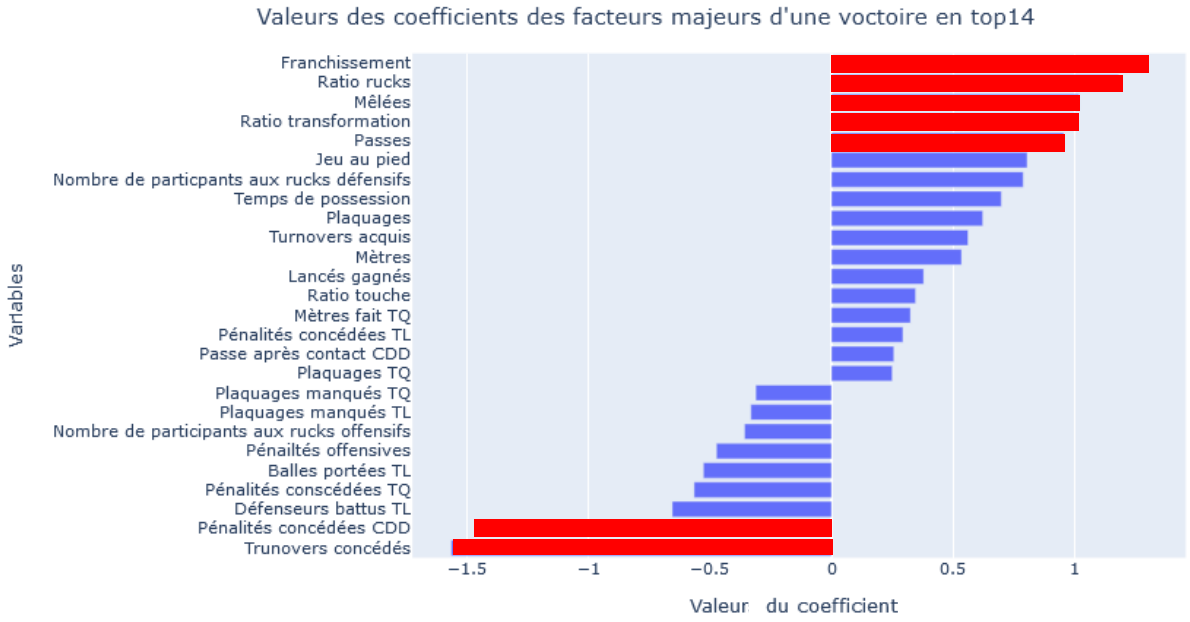
RF :RandomForest, LR : Regression logistique
RFECV : recursive feature elimination with cross validation
FP: facteurs de points

L'analyse des facteurs de victoires est faite avec modèle LR+RFECV sans points et FP rela. Bien que moins précis que le modèle LR, il retire déjà 22 variables et ne retire que 1% de précision.

5.2 Quelles variables influencent la victoire

Le modèle choisit se base sur les données relatives, les variables sont donc à interpréter de la façon suivante :

- Que faut-il faire de plus que mon adversaire
- Que faut-il faire de moins que mon adversaire



Bien que toutes ces variables soient importantes dans la précision certaines semblent se dégager plus que d'autres :

Ce qu'il faut faire de moins :

- Concéder des turnovers, il apparaît comme plus significatif que d'en avoir
- Concéder des pénalités avec le 5 de devant, la nature ou le lieu de ses pénalités doivent engendrer plus de points. Ces joueurs sont aussi plus dans les zones de combat ou les fautes arrivent plus facilement.

Ce qu'il faut faire de plus

- Les franchissements, créant des déséquilibres dans la défense permettent soit de marquer (directement ou en enchaînant), soit d'obtenir une pénalité, ou au moins éloigner fortement la balle de son embut.
- Le ratio de rucks positifs, bien que les franchissements soient l'idéal, les temps jeu finissent majoritairement par des rucks. Il faut donc s'assurer de conserver la balle.
- Le nombre de mêlées, phase de lancement de jeu souvent gagnée (obtention de pénalité ou gain direct) permet de gagner beaucoup de mètres.
- Le ratio de transformation au pied, bien que les essais soient plus importants dans la modélisation de victoire (vue dans la modélisation relative avec toutes les variables) la deuxième façon de marquer des points reste les coups de pied de conversions qu'il faut donc réussir.
- Le nombre de passes, les équipes produisant un plus gros volume de jeu semblent avantagées (cette variable se cumule aussi avec le temps de possession).

6 Conclusions et perspectives

Les analyses statistiques de cette étude ont conclu plusieurs choses, dans un premier temps:

- Il est possible de regrouper les équipes en fonction de leurs types de jeu.
- Face aux données relatives, les caractéristiques des équipes suivent la même logique.
- La clusterisation se fait principalement sur la capacité ou le désir de développer du volume de jeu.
- Cependant, il serait nécessaire d'entrer plus en détails dans la clusterisation afin d'établir plus précisément les styles de jeu des équipes.

Dans un second temps :

- Bien que Bennett M(1) conclut que peu de facteur influence la victoire, on a vu à travers les différents modèles que la moindre source d'information peut rajouter de la précision.
- La régression logistique semble être un modèle mieux adapté à la prédiction de victoire.
- Les facteurs majeurs d'une victoire de Top14 sont (on retrouve aussi ses résultats dans la littérature anglo-saxonne) :
    - Les Turnovers (concédés puis acquis)
    - Les pénalités concédées par le cinq de devant
    - Les franchissements
    - Le ratio de Rucks réussis
    - Le nombre de mêlées
    - Le ratio des coups de pieds de conversions
    - Le nombre de passes
- Ces facteurs sont à considérer dans l'idée relative, il faut faire plus que l'adversaire. (Moins pour les pénalités et turnovers concédés)

Il serait intéressant de modéliser les victoires selon certains critères pour voir si les facteurs majeurs changent :

- Lieu du match (domicile/extérieur)
- Type d'équipe
- Type d'opposition

7 Références

7.1 Etudes

- (1)Carmen M. E. Colomer, David B. Pyne, Mitch Mooney, Andrew McKune & Benjamin G. Serpell. Performance Analysis in Rugby Union: a Critical Systematic Review. Sports Medicine-Open.2020
- (2)Bennett M, Bezodis N, Shearer DA, Locke D, Kilduff LP. Descriptive conversion of performance indicators in rugby union. J Sci Med Sport. 2019

7.2 Articles

- Comment les clubs sportifs capitalisent sur les données pour augmenter leurs performances [en ligne]. L'usine digitale [consulté le 01/12/2020] https://www.usine-digitale.fr/article/comment-les-clubs-sportifs-capitalisent-sur-les-donnees-des-athletes-pour-augmenter-leurs-performances.N865315
- Sport et Big Data – Quand la science des données donne l'avantage sur le terrain [en ligne].Le Big Data[consulté le 01/12/2020] https://www.lebigdata.fr/sport-et-big-data?fbclid=IwAR1Uu2vVn4j-xw0E0wHt1T1TqP2BI9-REn7vVs1G_LTSG81PpUkPSmIBvCM