



**INSTITUT
POLYTECHNIQUE
DE PARIS**

SD701

EXPLORATION DE GRANDS VOLUMES DE DONNÉES

Analyse des Matchs de Football sur la demi-saison 2016 - 2017 en Ligue 1

Élèves :

LIAIGRE Clément
MECCHIA Pierre
GOERKE Alann

Enseignants :

T. VIARD



IP PARIS

30 mai 2023

Table des matières

| | |
|---|-----------|
| Introduction | 2 |
| Présentation du jeu de données | 2 |
| Objectifs du projet | 3 |
| Nettoyage des données | 4 |
| 1 Analyse des données | 5 |
| 1.1 Analyse statistique descriptive | 5 |
| 1.2 Analyse en Composantes Principales | 7 |
| 1.3 Méthode d'extraction d'ensembles d'éléments fréquents | 10 |
| 2 Mise en application d'algorithmes de Machine Learning | 13 |
| 2.1 Algorithmes utilisés | 13 |
| 3 Graphes | 16 |
| 3.1 Proportion de passes faites entre les joueurs d'une même équipe | 16 |
| 3.2 Classement des meilleurs buteurs | 17 |
| 3.3 Position d'un joueur sur le terrain | 18 |
| Conclusion | 19 |
| A Annexes | 22 |
| A.1 Fichier type XML | 22 |
| A.2 Évaluations des performances - Régression Logistique | 23 |
| A.3 Évaluations des performances - Support Vector Machines | 24 |
| A.4 Évaluations des performances - Random Forest | 25 |

Introduction

Présentation du jeu de données

Les données étudiées sont des fichiers XML correspondant à 190 matchs de la première partie de la saison de football français de 2016 / 2017, mises à disposition par la société Opta Sports, spécialisée dans la collecte et l'analyse de flux de données sportives. Chaque fichier XML, annexe A.1, est composé d'un match avec la hiérarchie suivante :

- Attributs associés au match :
 - les équipes
 - quelle équipe joue à domicile, joue à l'extérieur
 - la date et heure du match
 - ...
- Attributs associés aux événements du match :
 - le type d'évènement
 - le joueur ayant réalisé l'évènement
 - l'équipe du joueur ayant réalisé l'évènement
 - l'heure
 - ...
- Attributs spécifiques à chaque événement de match

Dans les fichiers XML décrits précédemment, les joueurs et les équipes sont désignés par un identifiant. Nous avons à disposition un fichier XML supplémentaire avec les informations de chaque joueur telles que son identifiant, son nom, son prénom, son club, sa position sur le terrain, ...Ainsi, il est possible de faire le rapprochement entre les événements de match et les informations des joueurs.

À partir des fichiers XML, nous avons créé deux DataFrames :

- un DataFrame df_event, figure 1, qui est composé des événements des 190 matchs avec 335429 lignes et 8 colonnes.

| | type_id | outcome | period_id | min | x | y | player_id | team_id |
|------------|---------|---------|-----------|-----|------|------|-----------|---------|
| Event | | | | | | | | |
| 1317941429 | 34 | 1 | 16 | 0 | 0.0 | 0.0 | None | 427 |
| 1068730640 | 34 | 1 | 16 | 0 | 0.0 | 0.0 | None | 694 |
| 1302720130 | 32 | 1 | 1 | 0 | 0.0 | 0.0 | None | 694 |
| 1455263236 | 32 | 1 | 1 | 0 | 0.0 | 0.0 | None | 427 |
| 247567393 | 1 | 1 | 1 | 0 | 49.7 | 48.7 | 229605 | 427 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 105366416 | 30 | 1 | 2 | 92 | 0.0 | 0.0 | None | 149 |
| 953695260 | 30 | 1 | 14 | 0 | 0.0 | 0.0 | None | 144 |
| 1125459369 | 37 | 1 | 14 | 0 | 0.0 | 0.0 | None | 144 |
| 1939935194 | 30 | 1 | 14 | 0 | 0.0 | 0.0 | None | 149 |
| 1127381362 | 37 | 1 | 14 | 0 | 0.0 | 0.0 | None | 149 |

335429 rows × 8 columns

FIGURE 1 – DataFrame des événements des 190 premiers matchs de Ligue 1

- un DataFrame df_player, figure 2, qui est composé des informations des joueurs de la Ligue 1 avec 762 lignes et 21 colonnes

| uID | name | team | position | first_name | last_name | birth_date | weight | height | jersey_num | real_position | ... | join_date | country | birth_place | fir |
|--------|--------------------|----------|------------|------------|-------------|------------|--------|--------|------------|----------------------|-----|------------|-----------|---------------------|-----|
| 213575 | Mehdi Tahrat | Angers | Defender | Mehdi | Tahrat | 1990-01-24 | 86 | 193 | 26 | Central Defender | ... | 2016-08-31 | Algeria | None | |
| 154011 | Flavien Tait | Angers | Midfielder | Flavien | Tait | 1993-02-02 | 70 | 175 | 20 | Attacking Midfielder | ... | 2016-07-01 | France | Longjumeau | |
| 193454 | Karl Toko Ekambi | Angers | Forward | Karl | Toko Ekambi | 1992-09-14 | 74 | 185 | 7 | Striker | ... | 2016-07-01 | Cameroon | Paris | |
| 168109 | Cheikh N'Doye | Angers | Midfielder | Cheikh | N'Doye | 1986-03-29 | 90 | 192 | 17 | Central Midfielder | ... | 2015-07-01 | Senegal | Rufisque | |
| 107613 | Romain Saïss | Angers | Midfielder | Romain | Saïss | 1990-03-26 | 80 | 190 | 28 | Defensive Midfielder | ... | 2015-07-01 | Morocco | Bourg-de-Péage | |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | |
| 194302 | Yann Bodiger | Toulouse | Midfielder | Yann | Bodiger | 1995-02-09 | 80 | 188 | 23 | Central Midfielder | ... | 2014-08-09 | France | Sète | |
| 68690 | Martin Braithwaite | Toulouse | Forward | Martin | Braithwaite | 1991-06-05 | 77 | 180 | 9 | Striker | ... | 2013-08-16 | Denmark | Esbjerg | |
| 74299 | Marc Vidal | Toulouse | Goalkeeper | Marc | Vidal | 1991-06-03 | 83 | 183 | 16 | Goalkeeper | ... | 2009-08-01 | France | Saint-Affrique | |
| 219924 | Issa Diop | Toulouse | Defender | Issa | Diop | 1997-01-09 | 92 | 194 | 5 | Central Defender | ... | 2015-07-01 | France | Toulouse | |
| 40721 | Óscar Trejo | Toulouse | Midfielder | Óscar | Trejo | 1988-04-26 | 79 | 180 | 10 | Attacking Midfielder | ... | 2013-07-18 | Argentina | Santiago del Estero | |

762 rows × 21 columns

FIGURE 2 – DataFrame des joueurs de Ligue 1

Ensuite, une jointure entre les deux DataFrames a été réalisée afin d'avoir les informations des joueurs pour chaque événement du DataFrame df_event.

Objectifs du projet

L'objectif principal du projet est de se familiariser avec l'exploration et la manipulation de grand volumes de données dans un domaine qui nous passionne : le Football.

Avant de passer aux problématiques de notre étude, une description succincte du Football s'impose afin de donner quelques clés de lecture qui permettront de mieux comprendre les éléments cités tout au long de ce rapport.

Le Football est un sport collectif qui se joue entre deux équipes de 11 joueurs. Un match de Football dure 90 minutes, divisées en deux parties de 45 minutes avec un intervalle de 15 minutes de repos entre les deux parties. L'équipe ayant marqué le plus de buts remporte le match. Les 11 joueurs d'une équipe sont répartis à des postes précis et au cours de chaque match, les 2 équipes qui s'affrontent ont une moitié du terrain qui leur est dédiée. Ainsi, en partant d'une extrémité du terrain vers l'autre, l'on retrouve :

- Le gardien : il occupe les buts pour intercepter les tirs de l'équipe adverse
- Les défenseurs : ils protègent la surface de buts pour empêcher l'équipe adverse de pénétrer celle-ci et de marquer des buts
- Les milieux : ils assurent la circulation du ballon entre les défenseurs et les attaquants
- Les attaquants : ils sont chargés de marquer des buts dans le camp adverse

Les interrogations

Les principales interrogations qui ont guidées nos recherches et nos analyses tout au long du projet sont les suivantes :

- Pourrait-on prédire le poste d'un joueur à partir des événements de matchs ?
- Quelles séries d'événements mènent à une action de but ?
- En cas de départ d'un joueur, quels joueurs potentiels pourraient le remplacer ?

Nettoyage des données

La phase de nettoyage a été relativement simple étant donné que le cœur de métier de la société Opta est de fournir aux professionnels du sport des données exploitables pour réaliser des analyses.

Notre travail sur le nettoyage des données a consisté à sélectionner les variables d'intérêt pour répondre à nos questionnements. Ainsi, cette étape nous a permis de supprimer 17 événements, sur les 50 existants, qui n'étaient pas pertinents dans le cadre de notre étude.

1 Analyse des données

Dans cette partie, nous allons effectuer dans un premier temps une analyse statistique descriptive de notre jeu de données. Puis, viendra une Analyse en Composante Principale. Enfin, la méthode d'extraction d'ensembles d'éléments fréquents sera explicitée et appliquée.

1.1 Analyse statistique descriptive

Nous avons commencé par regarder le nombre d'événements par joueur en traçant leur histogramme, visible sur la figure 3. Celui-ci nous indique que, dans notre jeu de données, il y a un grand nombre de joueurs qui ont peu d'événements sur la demi-saison. En effet, on remarque que 115 joueurs (environ 23% des joueurs) ont moins de 250 événements sur la totalité de la saison, ce qui est trop peu pour effectuer une bonne analyse par la suite. Ces joueurs s'avèrent être des remplaçants qui ne jouent pas régulièrement, ou des joueurs qui se sont gravement blessés en cours de saison. L'analyse de ces 115 joueurs n'étant pas aussi intéressante que ceux ayant joué la plupart ou la totalité des matchs, nous n'en tiendrons pas compte pour la suite de l'étude.

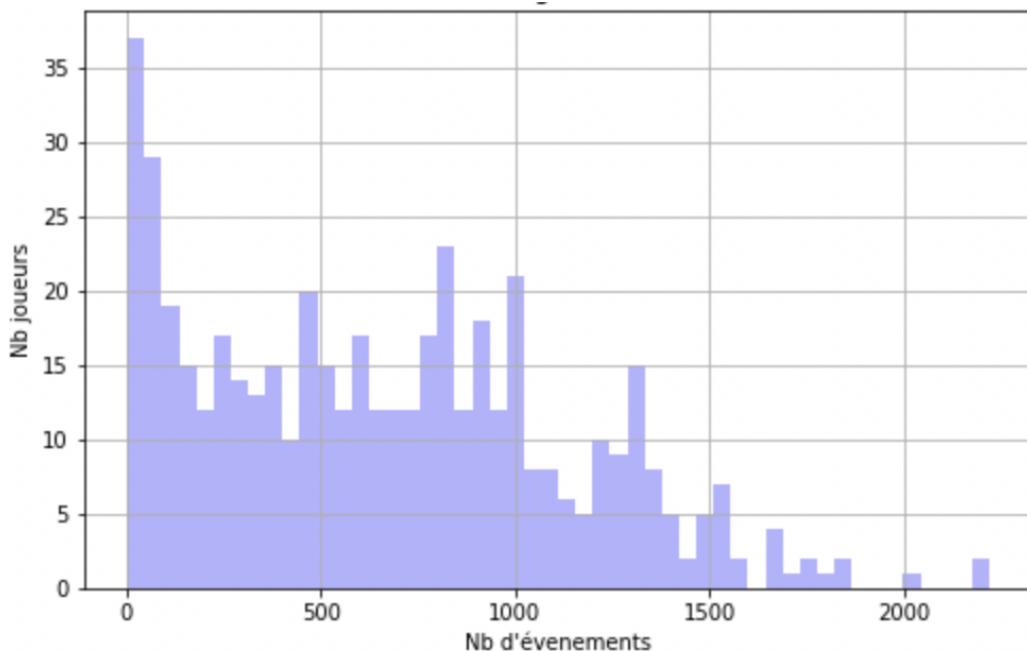


FIGURE 3 – Histogramme du nombre d'événements par joueur

La figure 4 ci-dessous, permet de mieux visualiser le nombre d'attributs propre à chacun des postes : défenseur, milieu et attaquant. Celui-ci met en évidence que les attaquants et les milieux sont ceux qui ont le plus d'événements, c'est à dire ceux qui sont le plus impliqués dans le jeu (également ceux qui touchent le plus le ballon).

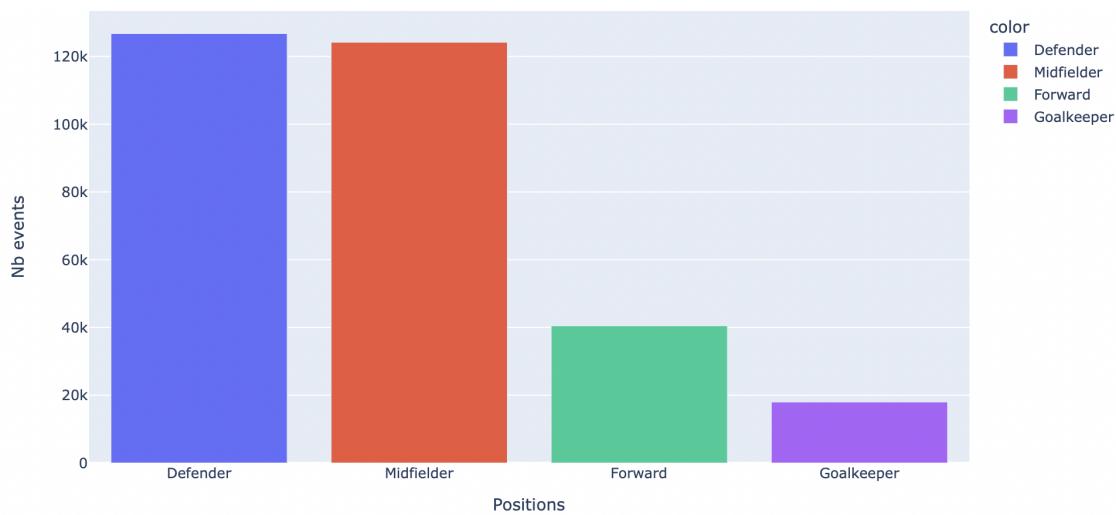


FIGURE 4 – Nombre d'évènements par poste

Il est également intéressant de comparer les fréquences d'apparitions pour chacun des types d'évènements, en fonction des postes occupés (figure 5).

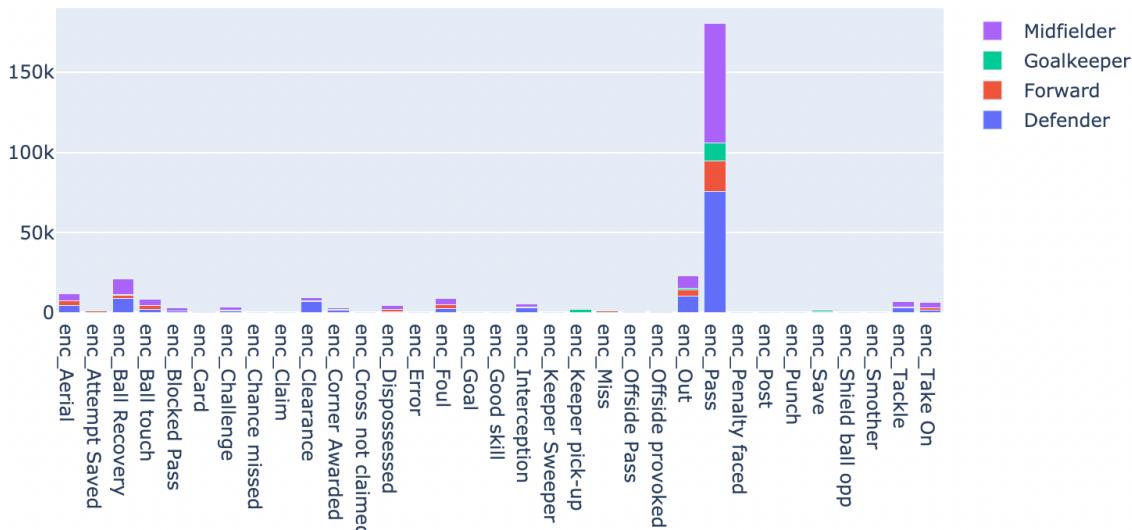


FIGURE 5 – Histogramme des événements par postes

L'évènement de type 'Passe', étant l'élément central du football, est l'évènement prédominant dans notre base de données. Ce sont les défenseurs et les milieux qui font le plus de passes, car ce sont généralement ces joueurs qui créent le jeu de leur équipe.

Du fait de l'omniprésence des passes, il est compliqué d'interpréter les autres types d'évènements. C'est pourquoi, la figure 6 représente également les fréquences d'apparitions pour chacun des types d'évènements en fonction des postes occupés, sans les passes.

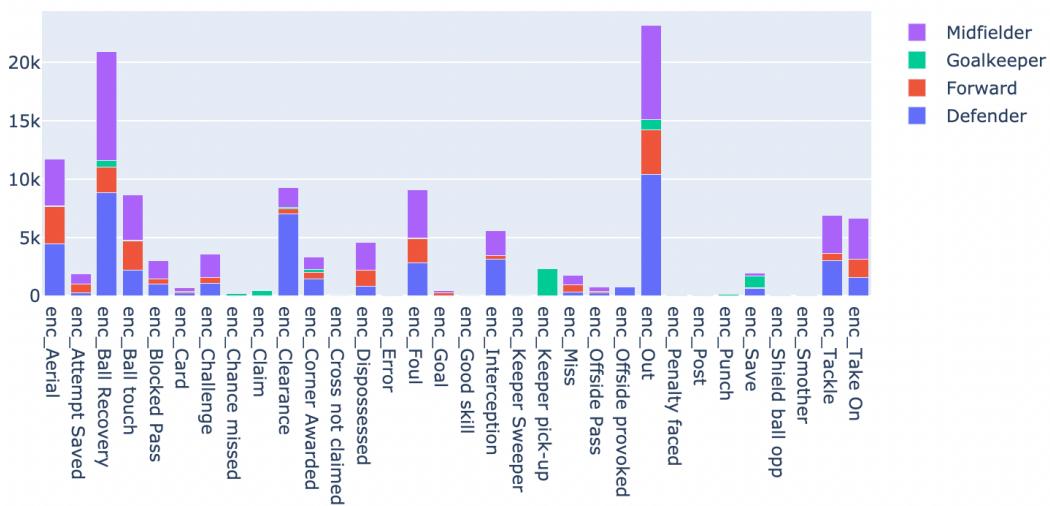


FIGURE 6 – Histogramme des évènements par postes, sans les passes

La figure 6 nous permet d'identifier des évènements caractéristiques des joueurs occupant des rôles différents. Par exemple, nous observons que les milieux et les défenseurs ont de fortes proportions à faire des récupérations de balles et à sortir le ballon en dehors du terrain comparés aux attaquants.

1.2 Analyse en Composantes Principales

Afin de réduire la dimension de notre jeu de données, nous avons mis en place une analyse en composantes principales. Le résultat de celle-ci, visible sur la figure 7, nous indique que 50% des données peuvent être représentées par uniquement trois variables.

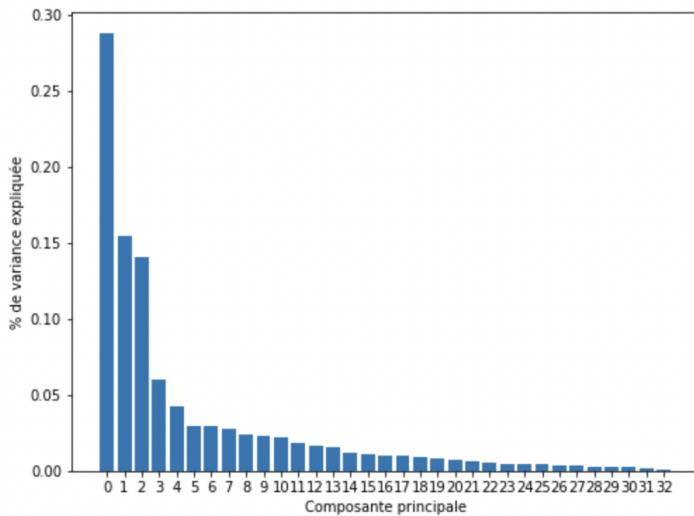


FIGURE 7 – Analyse en Composantes Principales sur les features

La difficulté ici est de réussir à interpréter ces trois variables, qui sont définies comme une combinaison linéaire des autres variables. Pour cela, il est possible de projeter nos features sur les deux premières composantes de l'ACP, cf figure 8.

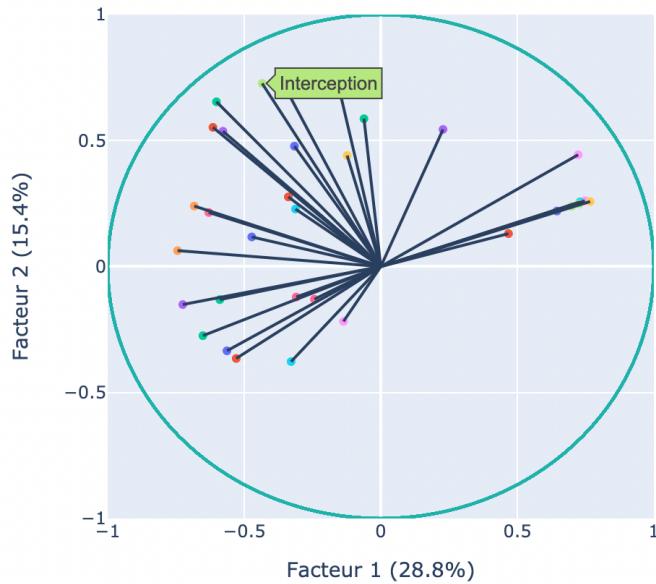


FIGURE 8 – Projection de l'ACP sur les deux première composantes

La figure 8 permet d'observer un regroupement d'évènements que l'on peut associer facilement à un poste en particulier. En effet, les valeurs positives de la première composante permettent d'isoler les gardiens de buts des autres joueurs. La deuxième composante permet de différencier les joueurs à vocation offensive de ceux à vocation défensive (ie facteur 2 négatif et positif). Par exemple, l'évènement "interception" qui est le plus souvent associé à un joueur (hors gardien) à profil défensif à pour valeur $(-0.4, 0.7)$.

La figure 9 représente quant à elle les projections entre les composantes 1 et 3 (a), puis 2 et 3 (b).

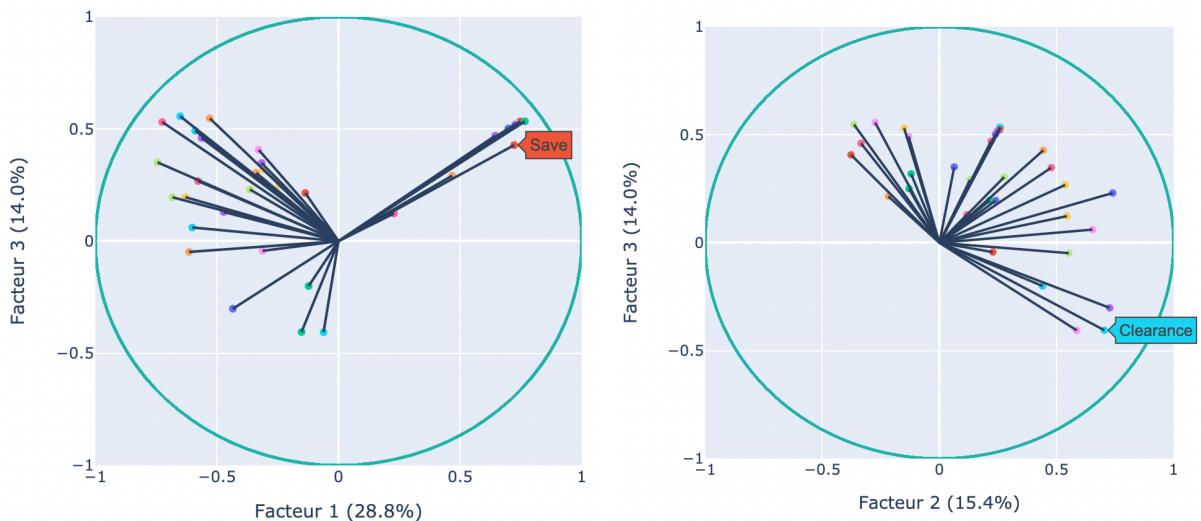


FIGURE 9 – Analyse en Composante Principale en projection sur les features

Une fois ces composantes analysées, nous avons cherché à visualiser notre jeu de données en projection selon ces composantes. Pour cela, il est possible de les projeter en deux ou trois

dimensions. Les résultats de l'ACP en trois dimensions, visible sur la figure 10, permettent de mettre en évidence que le football pratiqué par les gardiens est complètement différent de celui pratiqué par les défenseurs, les milieux et les attaquants, ce qui n'est pas surprenant. En effet, la projection en 3 dimensions donne l'impression que les gardiens sont sur un plan différent de celui des autres joueurs.

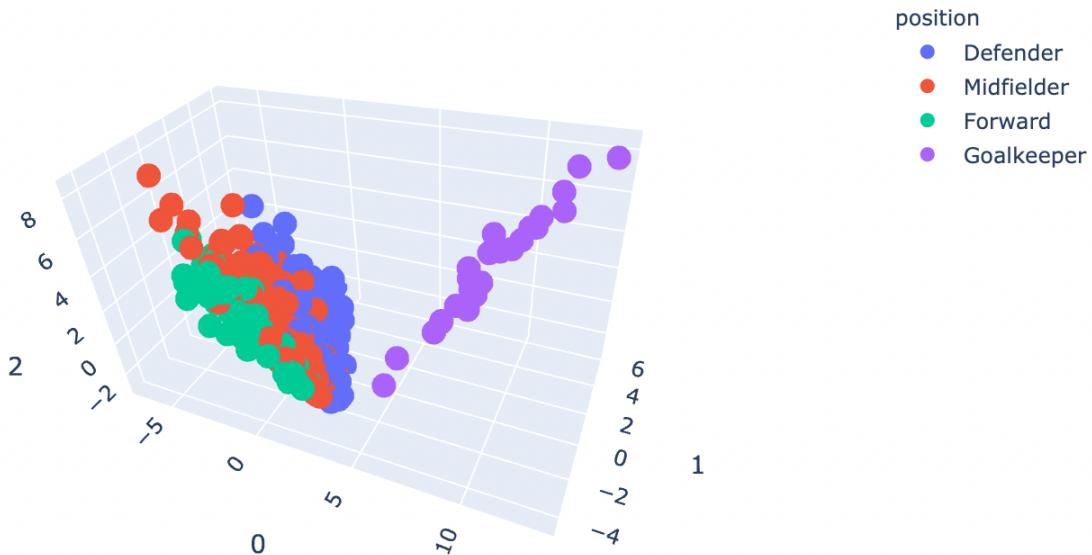


FIGURE 10 – Projection 3D - Analyse en Composante Principale

Pour valider cette remarque, il est également possible d'observer la projection de ces points sur deux composantes uniquement, par exemple les composantes 1 et 2. La figure 11 illustre la séparation très nette entre les gardiens et les autres joueurs.

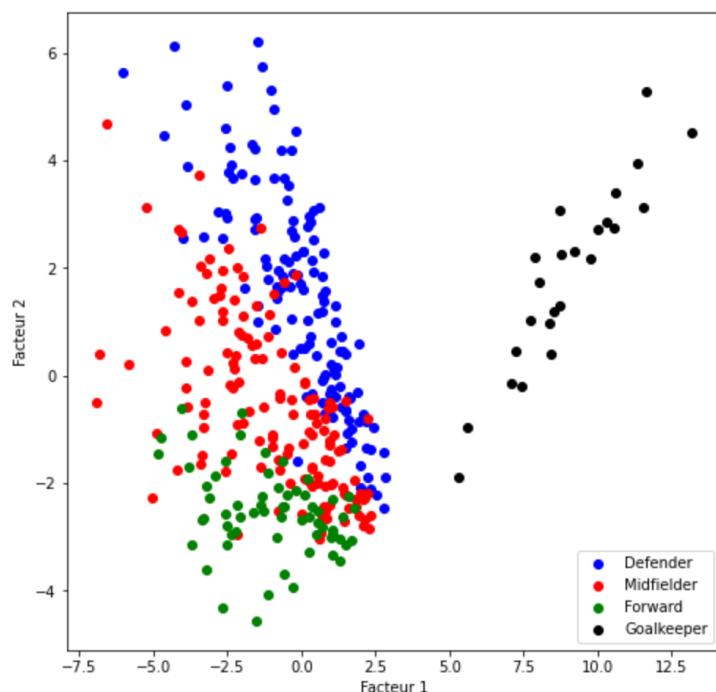


FIGURE 11 – Projection 2D - Analyse en Composante Principale

1.3 Méthode d'extraction d'ensembles d'éléments fréquents

Plusieurs approches sont possibles pour répondre à la question : "Quelles séries d'événements mènent à une action de but ?".

Les méthodes utilisées consistent à récupérer pour chaque évènement 'Goal' du DataFrame évènement, les dix événements qui précèdent l'évènement 'Goal' et d'identifier parmi ces événements, ceux qui sont les plus fréquents.

Dans un premier temps, pour réaliser cette analyse, nous nous sommes intéressés à la méthode de fouille de textes TF-IDF. Ainsi, les dix événements précédents un évènement 'Goal' sont considérés comme une suite de chaînes de caractères (cf figure 12) et notre objectif est de déterminer quelles sont les chaînes de caractères les plus fréquentes. Dans un second temps, nous nous sommes également intéressés aux méthodes d'extraction d'éléments fréquents a-priori et FP-growth afin de trouver des règles d'associations entre l'évènement 'Goal' et les dix événements qui précèdent un but.

| all_events | |
|----------------------|---|
| 0 | Pass Pass Aerial Aerial Pass Interception Pass... |
| 1 | Save Attempt_Saved Pass Pass Punch Pass Pass C... |
| 2 | Post Pass Pass Pass Pass Out Out Intercep... |
| 3 | Pass Good_skill Take_On Take_On Aerial Save At... |
| 4 | Pass Pass Corner_Awarded Corner_Awarded Corner... |
| ... | ... |
| 477 | Foul Foul Pass Pass Pass Pass Pass Aerial... |
| 478 | Chance_missed Pass Pass Pass Pass Pass Pass Ae... |
| 479 | Pass Good_skill Take_On Take_On Pass Keeper_Sw... |
| 480 | Punch Error Punch Punch Pass Corner_Awarded Sa... |
| 481 | Save Attempt_Saved Take_On Take_On Good_skill ... |
| 482 rows × 1 columns | |

FIGURE 12 – DataFrame utilisé pour la méthode TF-IDF

Le résultat de la méthode TF-IDF nous indique que les événements 'Pass', 'Foul' et 'Aerial' sont les événements les plus fréquents avant la réalisation d'un évènement 'Goal'. La sur-représentation de l'évènement 'Pass' dans les événements des matchs ne permet pas de conclure quant à la réalisation d'un but suite à une passe, figure 13.

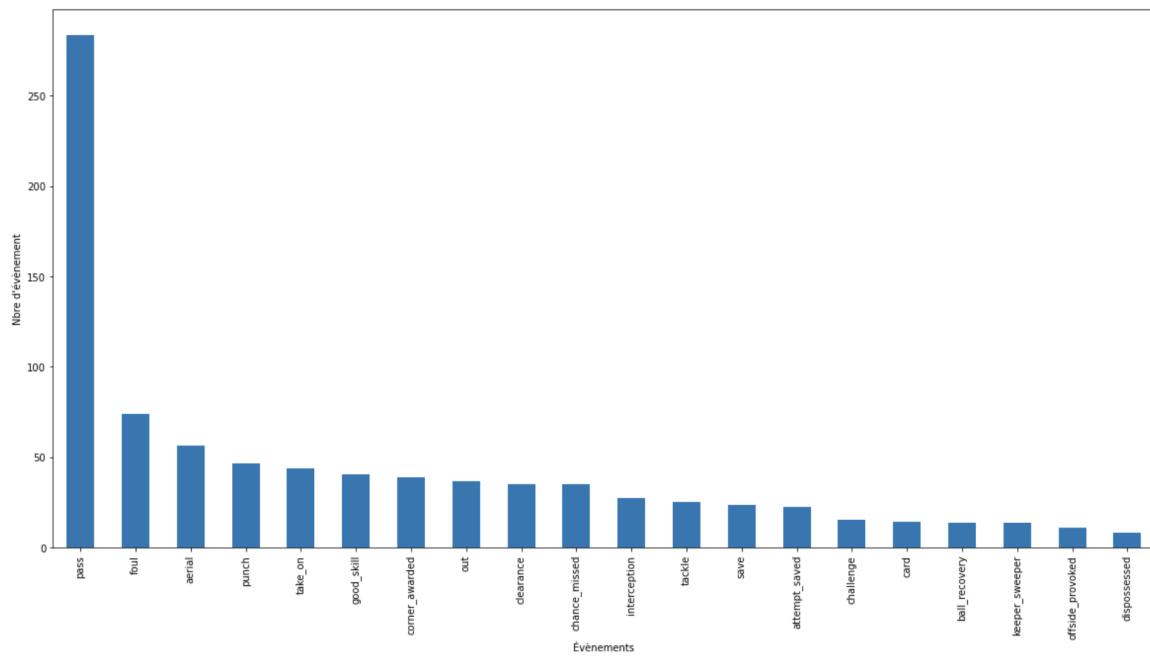


FIGURE 13 – Résultats de la méthode TF-IDF

Néanmoins, la présence des évènements 'foul' (fautes) et 'aerial' (duel aérien) peut nous montrer l'importance des corners, coups-francs ou penalty dans la réalisation d'un but, figure 14.

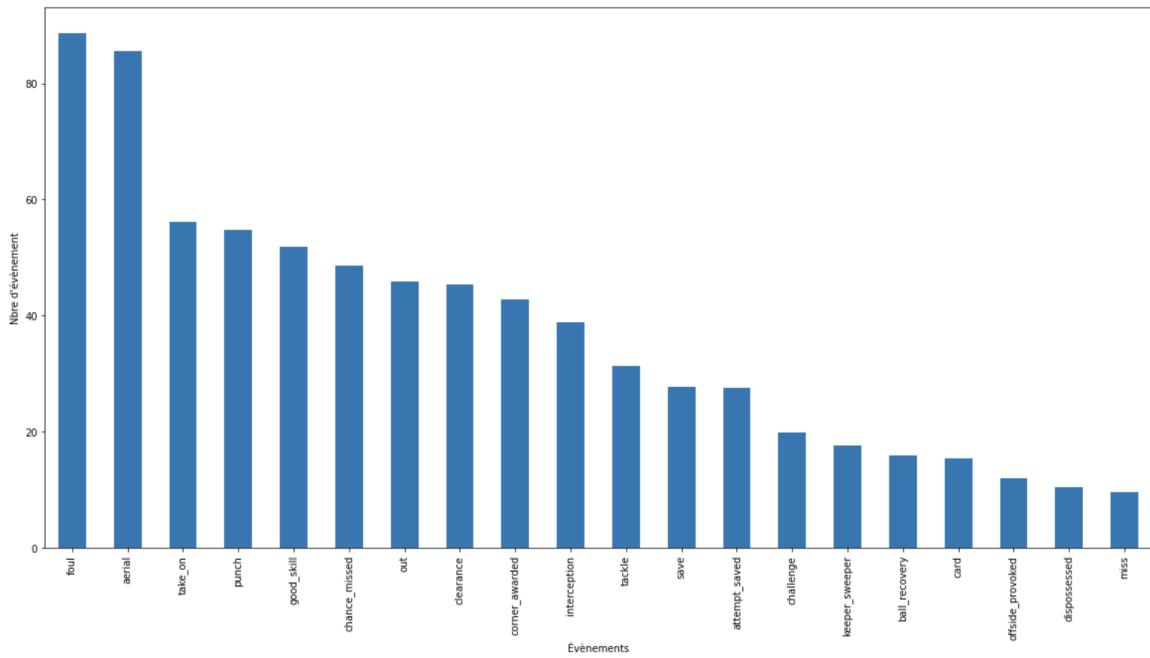


FIGURE 14 – Résultats de la méthode TF-IDF avec un seuil avec une borne max de 0.95 sur la fréquence d'apparition d'un événement

La méthode a-priori et FP-growth nous donne les mêmes résultats. Après avoir appliqué un filtre sur les nombres d'évènements antécédents supérieurs à deux et sur l'évènement résultant 'Goal'. Ainsi, les tuples d'évènements antécédents les plus fréquents engendrant l'évènement 'Goal' sont les (Pass, Aerial) et (Foul, Pass) (cf figure 15).

| | antecedents | consequents | antecedent support | consequent support |
|-----------|-----------------------------|--------------------|---------------------------|---------------------------|
| 21 | (Pass, Aerial) | (Goal) | 0.419087 | 1.0 |
| 30 | (Foul, Pass) | (Goal) | 0.273859 | 1.0 |
| 45 | (Take On, Pass) | (Goal) | 0.273859 | 1.0 |
| 27 | (Clearance, Pass) | (Goal) | 0.224066 | 1.0 |
| 33 | (Pass, Good skill) | (Goal) | 0.217842 | 1.0 |
| 24 | (Chance missed, Pass) | (Goal) | 0.209544 | 1.0 |
| 35 | (Take On, Good skill) | (Goal) | 0.192946 | 1.0 |
| 51 | (Take On, Pass, Good skill) | (Goal) | 0.192946 | 1.0 |
| 42 | (Punch, Pass) | (Goal) | 0.176349 | 1.0 |
| 39 | (Interception, Pass) | (Goal) | 0.161826 | 1.0 |

FIGURE 15 – Résultats de la méthode a-priori

La méthode a-priori confirme les résultats obtenus précédemment avec la méthode TF-IDF. La présence dans les résultats des tuples d'évènements (Take On, Pass) et (Pass, Good skill) nous indiquent que les joueurs capables de réaliser certaines prouesses techniques (Good Skill) pourraient avoir d'autant plus de facilité à marquer un but. De plus, le tuple (Clearance, Pass) pourrait indiquer qu'une contre-attaque suite aux dégagements d'un gardien a des chances d'engendrer un but.

2 Mise en application d'algorithmes de Machine Learning

Cette partie est dédiée à l'implémentation de différentes algorithmes de Machine Learning ainsi qu'à leur évaluation.

2.1 Algorithmes utilisés

Dans cette partie, nous allons chercher à répondre aux deux problématiques restantes.

Puis-je prédire la position d'un joueur à partir des événements de matchs ?

Pour évaluer nos modèles, nous avons utilisé un jeu de test équivalent à 20% de la taille totale de notre jeu de données.

Pour répondre à cette problématique, nous avons choisi de tester trois familles d'algorithmes de classification :

- régression : Régression Logistique (annexe A.2)
- méthode à noyaux : Support Vector Machines (annexe A.3)
- arbres de décisions : Random Forest (annexe A.4) et Gradient Boosting

Pour utiliser au mieux nos algorithmes, nous avons effectué une partie de post-processing qui se résume en l'utilisation d'un encodage One-hot permettant de passer d'une seule variable 'type_id' à 33 variables correspondant aux événements sélectionnés. Nous y avons appliqué des agrégations temporelles différentes telles que 15, 45, 90 minutes d'un match mais aussi sur la totalité de la mi-saison. Cela nous a permis de faire varier le nombre d'observations. Par exemple en considérant un intervalle de 45 minutes, nous avions 8918 observations, dont 80% (7134 observations) ont servi à l'entraînement du modèle et 20 % (1784 observations) ont servi de données de test, comme présenté dans les résultats ci-dessous.

| | date | period_id | player_id | enc_Aerial | enc_Attempt Saved | enc_Ball Recovery | enc_Ball touch | enc_Blocked Pass | enc_Card | enc_Challenge | ... | enc_Post | enc_Punch | enc_Save | € |
|------|------------|-----------|-----------|------------|-------------------|-------------------|----------------|------------------|----------|---------------|-----|----------|-----------|----------|-----|
| 0 | 2016-08-12 | 1 | 102739 | 2.0 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 1 | 2016-08-12 | 2 | 102739 | 2.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 1.0 | ... | 0.0 | 0.0 | 0.0 | 1.0 |
| 2 | 2016-08-21 | 1 | 102739 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 1.0 | ... | 0.0 | 0.0 | 0.0 | 2.0 |
| 3 | 2016-08-21 | 2 | 102739 | 1.0 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| 4 | 2016-08-27 | 1 | 102739 | 3.0 | 0.0 | 2.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 8913 | 2016-12-18 | 2 | 38265 | 0.0 | 0.0 | 1.0 | 0.0 | 1.0 | 0.0 | 1.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 |
| 8914 | 2016-12-21 | 1 | 38265 | 1.0 | 0.0 | 6.0 | 2.0 | 0.0 | 0.0 | 1.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 |
| 8915 | 2016-12-21 | 2 | 38265 | 1.0 | 0.0 | 7.0 | 1.0 | 0.0 | 0.0 | 1.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 |
| 8916 | 2017-04-05 | 1 | 38265 | 2.0 | 0.0 | 2.0 | 1.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 |
| 8917 | 2017-04-05 | 2 | 38265 | 2.0 | 0.0 | 5.0 | 1.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |

8918 rows × 39 columns

FIGURE 16 – Dataframe après encodage One-hot et agrégation sur 45min

Pour évaluer nos modèles nous avons utilisé les métriques de f1_score, validation croisée (figure 17), matrice de confusion (figure 18) et courbes ROC (figure 19). Les résultats de nos quatre algorithmes sont très proches mais nous avons choisi de sélectionner le Gradient Boosting car l'écart-type de la validation croisée est la plus faible.

| Classification Report: | | | | |
|------------------------|-----------|--------|----------|---------|
| | precision | recall | f1-score | support |
| 0 | 0.83 | 0.74 | 0.78 | 640 |
| 1 | 0.66 | 0.67 | 0.67 | 334 |
| 2 | 0.98 | 1.00 | 0.99 | 129 |
| 3 | 0.66 | 0.71 | 0.68 | 681 |
| accuracy | | | 0.74 | 1784 |
| macro avg | 0.78 | 0.78 | 0.78 | 1784 |
| weighted avg | 0.74 | 0.74 | 0.74 | 1784 |

| Cross Validation: | | | | |
|-------------------------------------|--|--|--|--|
| Score moy 7.442e-01 (+/-) 1.753e-02 | | | | |

FIGURE 17 – Gradient Boosting - Scores

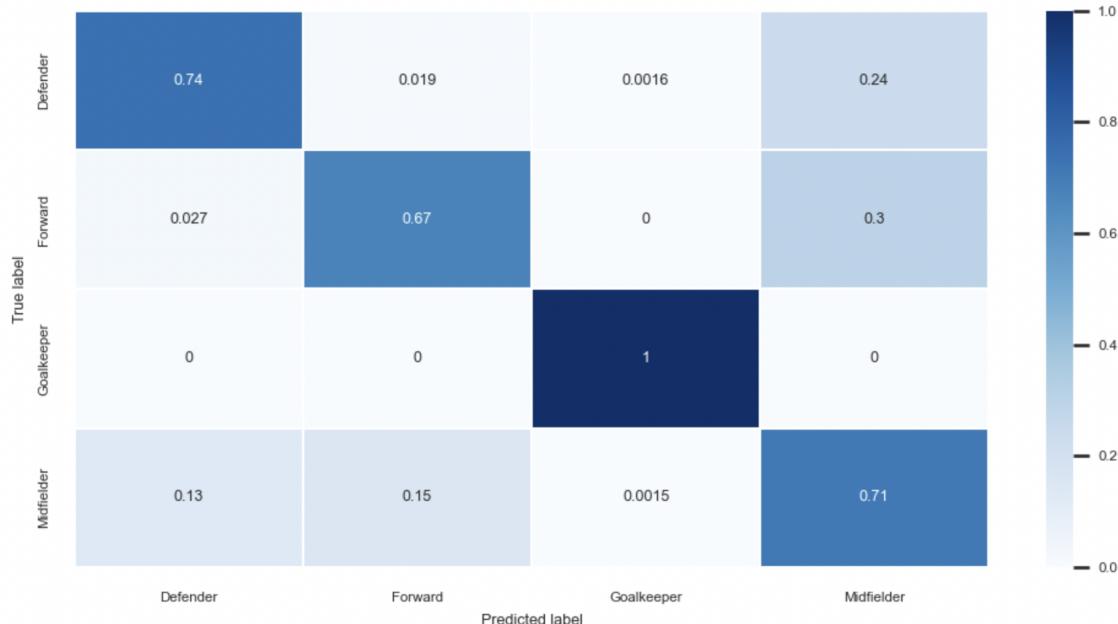


FIGURE 18 – Gradient Boosting - Matrice de Confusion

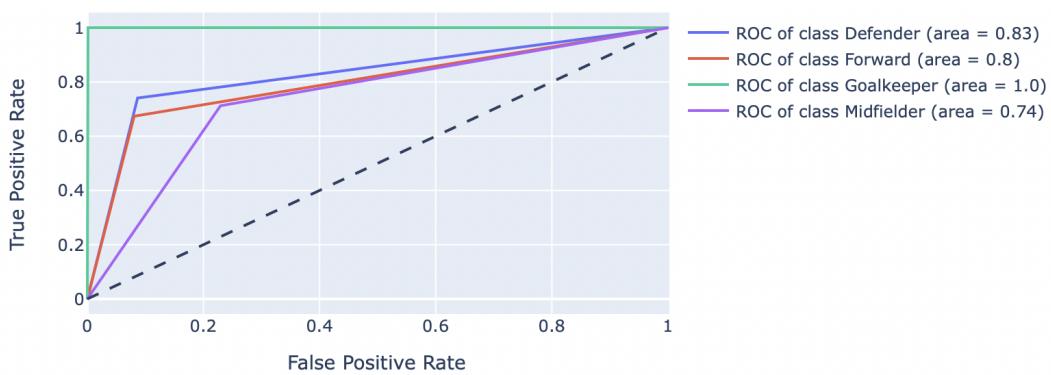


FIGURE 19 – Gradient Boosting - Courbe ROC

En cas de départ d'un joueur, quels joueurs potentiels pourraient le remplacer ?

Quant à cette problématique, nous y avons répondu avec la méthode des K-Nearest Neighbours.

L'objectif ici est de trouver le plus proche voisin d'une sélection de joueurs. Pour cela, nous avons entraîné notre modèle sur les évènements de la totalité des joueurs de Ligue 1, à l'exception de ceux du Paris Saint-Germain. Puis, nous avons testé notre modèle sur tous les évènements des joueurs de l'équipe du Paris Saint-Germain. Il n'est pas intéressant de montrer les résultats pour chacun des joueurs testés, c'est pourquoi, seuls les résultats d'une sélection de 5 joueurs possédant des caractéristiques différentes et évoluant à des postes différents vont être illustrés dans la figure 20.

Ces joueurs sont :

```
Out[62]: 21      Kevin Trapp
            3      Marquinhos
            23     Marco Verratti
            7     Hatem Ben Arfa
            1     Edinson Cavani
Name: name, dtype: object
```

FIGURE 20 – KNN - Sélection de 5 joueurs de l'échantillon test

Les joueurs, en Ligue 1, étant les plus proches avec ceux cités sur la figure précédente, sont :

```
Out[63]: array(['Benjamin Lecomte', 'Nicolas Nkoulou', 'Florent Balmont',
   'Bouna Sarr', 'Alexandre Lacazette'], dtype=object)
```

FIGURE 21 – KNN - Prédictions

Analyse

Les résultats de cet algorithme sont très intéressants pour plusieurs raisons :

- Tout d'abord, les prédictions de l'algorithme nous renvoient pour tous les joueurs, à l'exception de Hatem Ben Arfa, un plus proche voisin qui joue exactement au même poste mais dans une équipe différente.
- On pourrait alors croire que l'algorithme s'est trompé en prédisant Bouna Sarr (défenseur) pour Hatem Ben Arfa (milieu) mais ce n'est pas le cas. Ce choix est fait car sur cette saison 2016 / 2017, Bouna Sarr était identifié comme défenseur, mais avait un rôle très offensif, ce qui lui vaut des évènements similaires à ceux de Ben Arfa.
- Enfin, on peut affirmer que le style de jeu d'Edinson Cavani correspond à celui de son plus proche voisin Alexandre Lacazette : beaucoup de buts marqués, peu d'influence sur le jeu de passe de l'équipe.

Remarque

Les résultats affichés sur la figure 21 sont ceux obtenus avec la métrique de **Mahalanobis**. Ceux obtenus avec la distance Euclidienne n'étant pas cohérents. Par exemple, le plus proche voisin prédit pour le gardien du Paris Saint-Germain Kevin Trapp était un milieu : incohérent.

3 Graphes

Dans cette partie, nous allons tâcher de visualiser clairement certains aspects tirés de notre jeu de données.

3.1 Proportion de passes faites entre les joueurs d'une même équipe

Il est intéressant de regarder, au sein d'une même équipe, quels peuvent être les schémas préférentiels en termes de passes, voire les affinités entre joueurs.

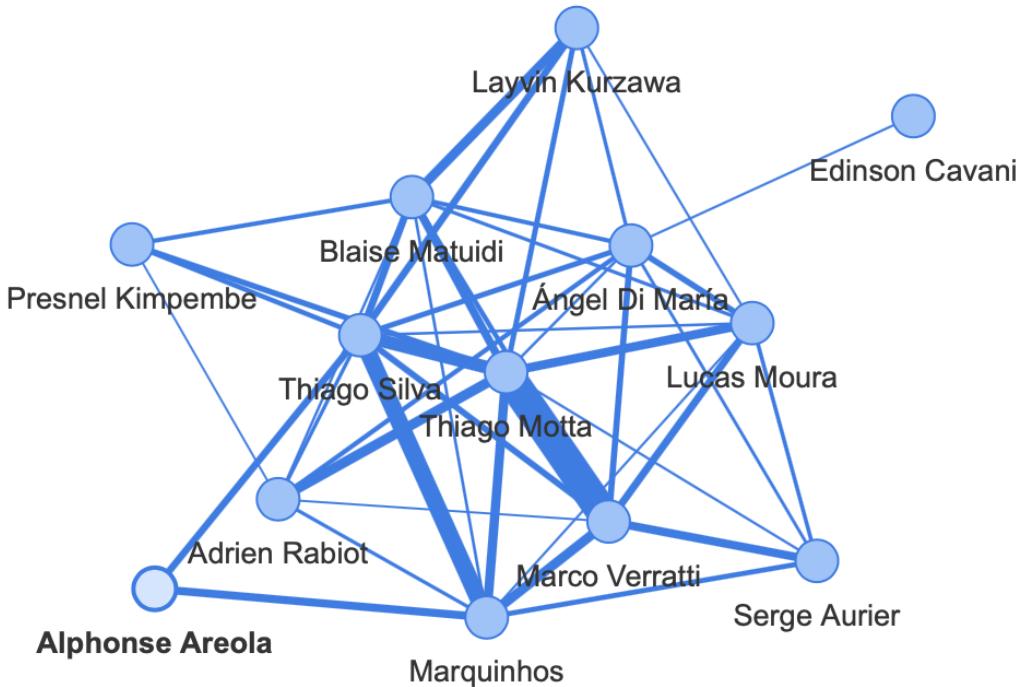


FIGURE 22 – Proportion de passes entre les joueurs du Paris Saint Germain

La figure 22 nous donne plusieurs informations sur le style de jeu du Paris Saint Germain. Tout d'abord, la relation entre Marco Verratti et Thiago Motta est très importante comparée aux autres. Ces deux joueurs sont au coeur du jeu de cette équipe, ce sont comme on dit les "plaques tournantes". On peut également remarquer qu'Edinson Cavani se trouve en dehors de ce système de jeu, il n'a qu'une seule connexion avec les autres joueurs. Cela s'explique par sa position de numéro 9 sur le terrain, et de par son profil de jeu atypique qui n'est pas celui de toucher beaucoup de ballons mais de marquer des buts.

En comparant les figures 22 et 23, on peut distinguer plus ou moins deux styles de jeu. Tout d'abord, des équipes comme le Paris Saint-Germain et Monaco ont un style de jeu axé sur la possession avec beaucoup de passes, beaucoup de connexions entre les joueurs et également un ou deux joueurs au coeur du jeu :

- Pour le Paris Saint-Germain : Verratti - Motta
- Pour Monaco : Fabinho - Sidibé

En revanche, la figure 23 montre bien qu'une équipe comme Metz à un style de jeu beaucoup plus désorganisé que Monaco et le Paris Saint-Germain. Cette équipe possède bien un joueur

au centre de schéma tactique, mais les relations entre les autres joueurs ne sont pas aussi fortes que celles des grandes équipes. Cela peut traduire, en partie, le fait que Metz soit dans les derniers du classement de Ligue 1 sur cette moitié de saison 2016 / 2017.

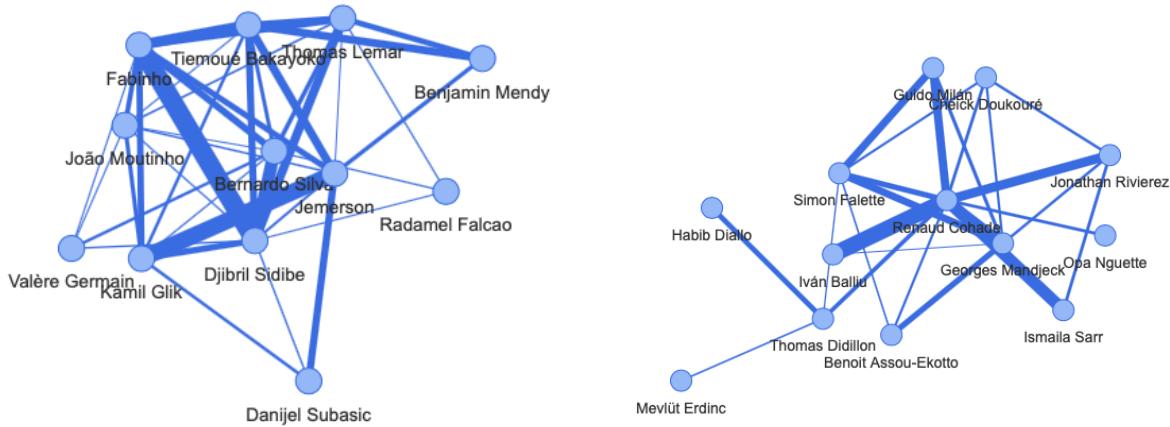


FIGURE 23 – Proportion de passes pour Monaco (a) et Metz (b)

3.2 Classement des meilleurs buteurs

Une autre possibilité qui s'offre à nous est la visualisation du nombre de buts cumulés par joueur sur ces quelques mois. Par soucis de lisibilité de la figure 24, nous avons uniquement affiché l'évolution des buts pour les 10 meilleurs buteurs à la fin de cette première moitié de saison.

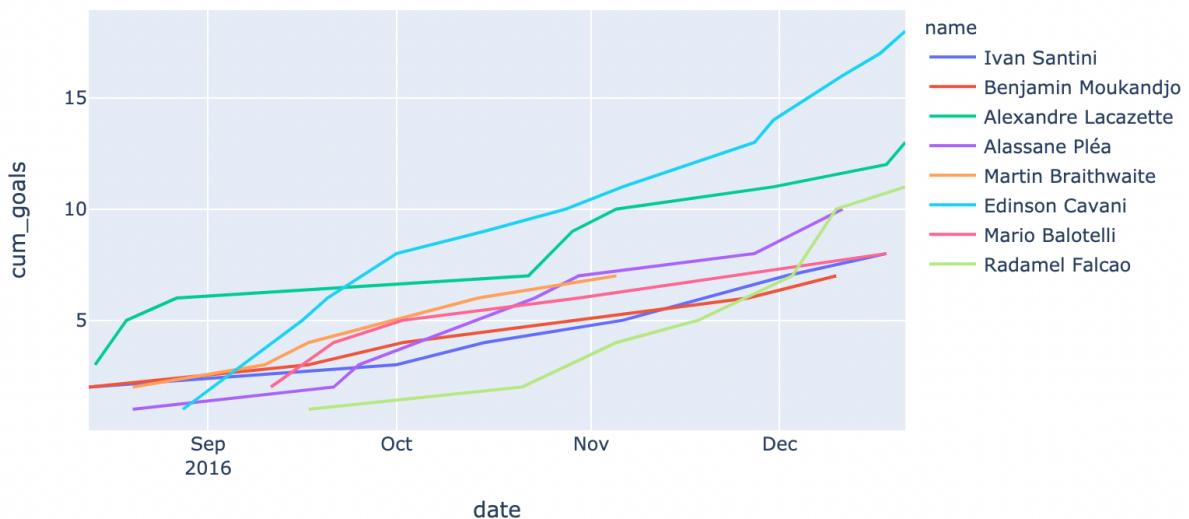


FIGURE 24 – Classement des 10 meilleurs buteurs sur la partie de saison

Les trois meilleurs buteurs, à l'issue de cette première partie de championnat de Ligue 1 2016 / 2017 sont :

1. Edinson Cavani, Paris Saint-Germain
2. Alexandre Lacazette, Lyon
3. Radamel Falcao, Monaco

3.3 Position d'un joueur sur le terrain

Dans un premier temps, il est possible de représenter le positionnement d'un joueur sur le terrain en utilisant sa densité de présence calculée à partir des coordonnées x et y fournies par notre jeu de données. La figure ci-dessous représente la densité de positionnement du joueur du Paris Saint-Germain Hatem Ben Arfa lors d'une journée tirée au hasard sur la première partie de saison.

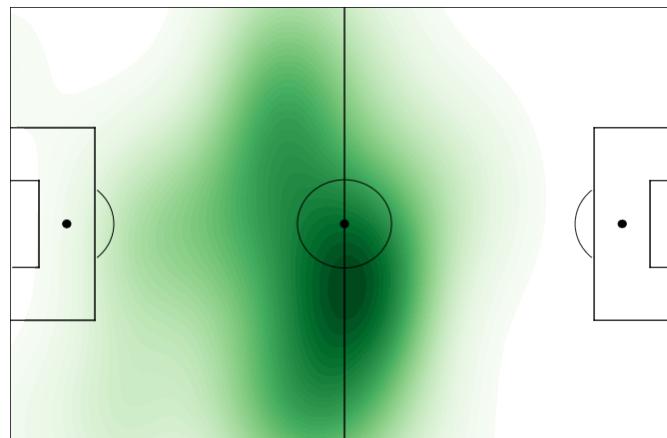


FIGURE 25 – Positionnement d'Hatem Ben Arfa sur un match

La figure 25 montre que Hatem Ben Arfa a une très forte présence dans le coeur du jeu au milieu du terrain. Il n'est en revanche peu, ou pas du tout actif dans les extrémités du terrain.

Afin de vérifier la cohérence du graphe avec la réalité, la même figure est représentée pour autre joueur : Serge Aurier, possédant un profil plutôt défensif.

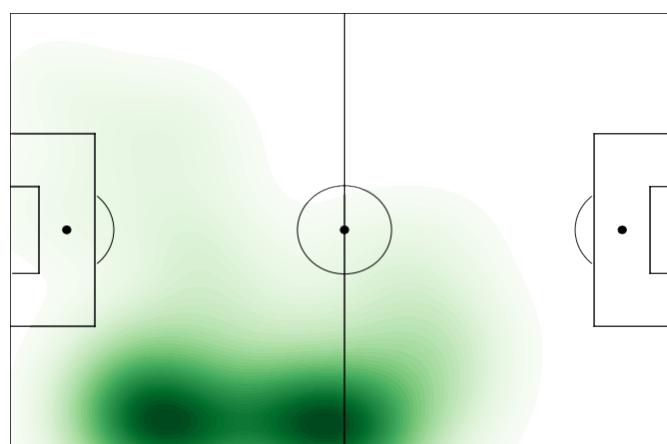


FIGURE 26 – Positionnement de Serge Aurier sur un match

En comparant les deux figures 25 et 26, on peut voir que les zones de forte densité de présence pour chacun des deux joueurs sont cohérentes avec les postes occupé en club sur cette partie de saison. En effet, Serge Aurier est défenseur droit avec le Paris Saint Germain, et la figure 26 montre bien que sa présence est presque entièrement située sur le côté droit du terrain, et plutôt en défense ainsi qu'au milieu du terrain.

Conclusion

Ce projet nous a permis de mettre en application des algorithmes et méthodes étudiées sur un sujet qui nous passionne. Nous sommes parvenus à traiter l'ensemble des questions posées initialement :

- Pourrait-on prédire le poste d'un joueur à partir des événements de matchs ?
- Quelles séries d'événements mènent à une action de but ?
- En cas de départ d'un joueur, quels joueurs potentiels pourraient le remplacer ?

La réussite de la stratégie de match reposant sur la bonne occupation des joueurs de leurs postes respectifs, la prédition des postes des joueurs à partir des évènements de match est un bon indicateur de la réussite de la stratégie de jeu. Cependant, l'on s'aperçoit que certains postes sont plus facilement prédictibles par les algorithmes que d'autres. C'est notamment le cas pour les gardiens et les défenseurs. A contrario, il y a souvent une confusion entre la prédition des attaquants et des milieux. Cela s'explique par le fait que les joueurs occupant ces postes ont tendance à avoir une certaine polyvalence et ont donc une similarité dans le style de jeu. La différence principale entre attaquants et milieux réside dans le nombre de buts. Cependant, les buts étant des évènements rares, l'importance de la variable dans la classification a pu être sous-estimée par l'algorithme.

En ce qui concerne la seconde question, il y a des séries d'événements spécifiques qui précèdent la réalisation d'un but. Ainsi, nous avons pu mettre en évidence, les évènements fautes, duels aériens, gestes techniques comme des événements caractéristiques de l'arrivée d'un but. Certains de ces événements relèvent des qualités des joueurs alors que d'autres relèvent de la circulation du ballon. Ces analyses pourraient ainsi permettre aux équipes d'adapter leur style de jeu (stratégie de jeu) afin de mettre l'accent sur la provocation de certains évènements amenant à inscrire un but.

Enfin, pour la dernière question, le choix de la métrique de distance dans l'algorithme du K-NN influence les résultats sur les potentiels joueurs qui pourraient remplacer un autre en cas de départ. Dans le cas où cette métrique est cohérente, le joueur prédit, ayant des proportions d'évènements similaires à celui que l'on souhaite remplacer, occupe généralement le même poste. Il existe cependant des cas où le joueur prédit comme remplaçant potentiel occupe actuellement un poste différent du joueur à remplacer ; même si la réalité montre que ce joueur remplaçant possède toutes les compétences requises. Ainsi, indépendamment du modèle (choix d'un algorithme autre que le K-NN), le remplacement d'un joueur ne se ferait pas uniquement sur la base des évènements de match.

Ouverture

Notre étude est une ébauche des analyses possibles de différentes problématiques importantes dans le Football. Néanmoins, il reste des pistes d'amélioration qu'il serait intéressant d'explorer :

- L'utilisation de modèles plus performants pour faire des analyses temporelles et séquentielles plus poussées. Les modèles que nous avons utilisés nécessitaient une grande agrégation des données qui a pu causer "une perte" relative d'informations.
- L'enrichissement des données d'évènements de match avec des données sur les compétences techniques et autres caractéristiques intrinsèques des joueurs.
- La prise en compte des stratégies de "positionnement" et d'occupation de jeu initiales, par exemple la composition ou le ratio défenseurs, milieux, attaquants (4, 4, 2), (5, 3, 2) etc.

Table des figures

| | | |
|----|--|----|
| 1 | DataFrame des événements des 190 premiers matchs de Ligue 1 | 2 |
| 2 | DataFrame des joueurs de Ligue 1 | 3 |
| 3 | Histogramme du nombre d'événements par joueur | 5 |
| 4 | Nombre d'évènements par poste | 6 |
| 5 | Histogramme des évènements par postes | 6 |
| 6 | Histogramme des évènements par postes, sans les passes | 7 |
| 7 | Analyse en Composantes Principales sur les features | 7 |
| 8 | Projection de l'ACP sur les deux première composantes | 8 |
| 9 | Analyse en Composante Principale en projection sur les features | 8 |
| 10 | Projection 3D - Analyse en Composante Principale | 9 |
| 11 | Projection 2D - Analyse en Composante Principale | 9 |
| 12 | DataFrame utilisé pour la méthode TF-IDF | 10 |
| 13 | Résultats de la méthode TF-IDF | 11 |
| 14 | Résultats de la méthode TF-IDF avec un seuil avec une borne max de 0.95 sur la fréquence d'apparition d'un événement | 11 |
| 15 | Résultats de la méthode a-priori | 12 |
| 16 | Dataframe après encodage One-hot et agrégation sur 45min | 13 |
| 17 | Gradient Boosting - Scores | 14 |
| 18 | Gradient Boosting - Matrice de Confusion | 14 |
| 19 | Gradient Boosting - Courbe ROC | 14 |
| 20 | KNN - Sélection de 5 joueurs de l'échantillon test | 15 |
| 21 | KNN - Prédictions | 15 |
| 22 | Proportion de passes entre les joueurs du Paris Saint Germain | 16 |
| 23 | Proportion de passes pour Monaco (a) et Metz (b) | 17 |
| 24 | Classement des 10 meilleurs buteurs sur la partie de saison | 17 |
| 25 | Positionnement d'Hatem Ben Arfa sur un match | 18 |
| 26 | Positionnement de Serge Aurier sur un match | 18 |

A Annexes

A.1 Fichier type XML

```

<!-- Copyright 2001-2016 Opta Sportsdata Ltd. All rights reserved.
--><!-- PRODUCTION HEADER
produced on:           valde-jobq-a02.nexus.opta.net
production time:      20161202T090651,965Z
production module:    Opta::Feed::XML::Soccer::F24
--><Games timestamp="">
<Game id="" away_team_id="" away_team_name="" competition_id=""
competition_name="" game_date="" home_team_id="" home_team_name=""
matchday="" period_1_start="" period_2_start="" season_id=""
season_name="">
<Event id="1273539596" event_id="149" type_id="1" period_id="1"
min="16" sec="46" player_id="9980" team_id="148" outcome="0"
x="63.9" y="41.5" timestamp="" last_modified=""
version="1471180635123">
<Q id="1072503433" qualifier_id="5"/>
<Q id="413957330" qualifier_id="155"/>
<Q id="1753987908" qualifier_id="212" value="26.5"/>
<Q id="736792328" qualifier_id="141" value="58.8"/>
<Q id="247565947" qualifier_id="140" value="86.5"/>
<Q id="1916282819" qualifier_id="213" value="0.5"/>
<Q id="1364259608" qualifier_id="152"/>
<Q id="267456351" qualifier_id="56" value="Center"/>
</Event>
<Event id="1123235964" event_id="279" type_id="44" period_id="1"
min="16" sec="48" player_id="168568" team_id="143" outcome="1"
x="12.6" y="38.4" timestamp="" last_modified=""
version="1471182567542">
<Q id="1831462194" qualifier_id="285"/>
<Q id="1205705974" qualifier_id="233" value="301"/>
<Q id="1323670578" qualifier_id="56" value="Back"/>
</Event>
<Event id="1289446" event_id="301" type_id="44" period_id="1"
min="16" sec="48" player_id="170034" team_id="148" outcome="0"
x="87.4" y="61.6" timestamp="" last_modified=""
version="1471182566166">
<Q id="247457625" qualifier_id="286"/>
<Q id="1588156035" qualifier_id="56" value="Center"/>
<Q id="618345020" qualifier_id="233" value="279"/>
</Event>
<Event id="1243377840" event_id="177" type_id="12" period_id="1"
min="16" sec="49" player_id="168568" team_id="143" outcome="1"
x="10.4" y="38.2" timestamp="" last_modified=""
version="1471181396307">
<Q id="773789077" qualifier_id="56" value="Back"/>
<Q id="1577857858" qualifier_id="213" value="5.1"/>
<Q id="759777798" qualifier_id="140" value="20.4"/>
<Q id="1786261241" qualifier_id="141" value="2.7"/>
<Q id="1211769768" qualifier_id="212" value="26.3"/>
<Q id="293681174" qualifier_id="15"/>
</Event>
<Event id="819945595" event_id="178" type_id="49" period_id="1"
min="16" sec="52" player_id="166552" team_id="143" outcome="1"
x="19.7" y="2.1" timestamp="" last_modified=""
version="1471181398373"/>

```

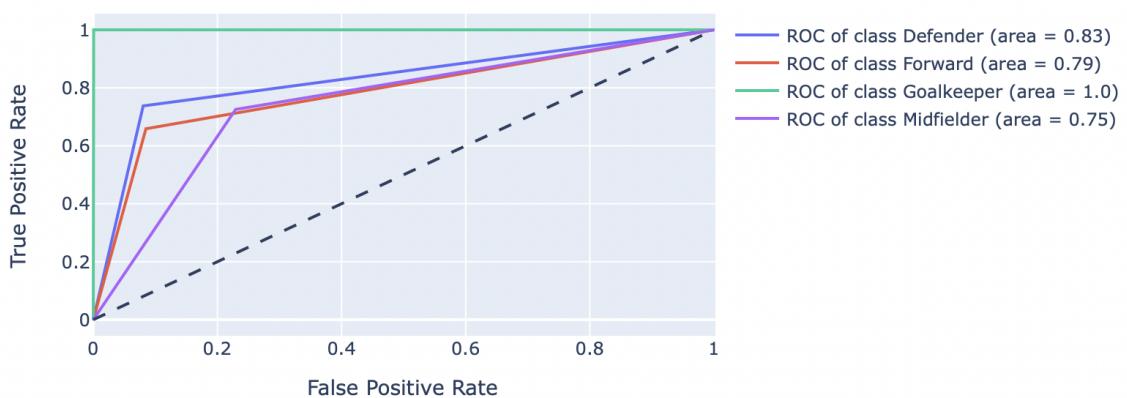
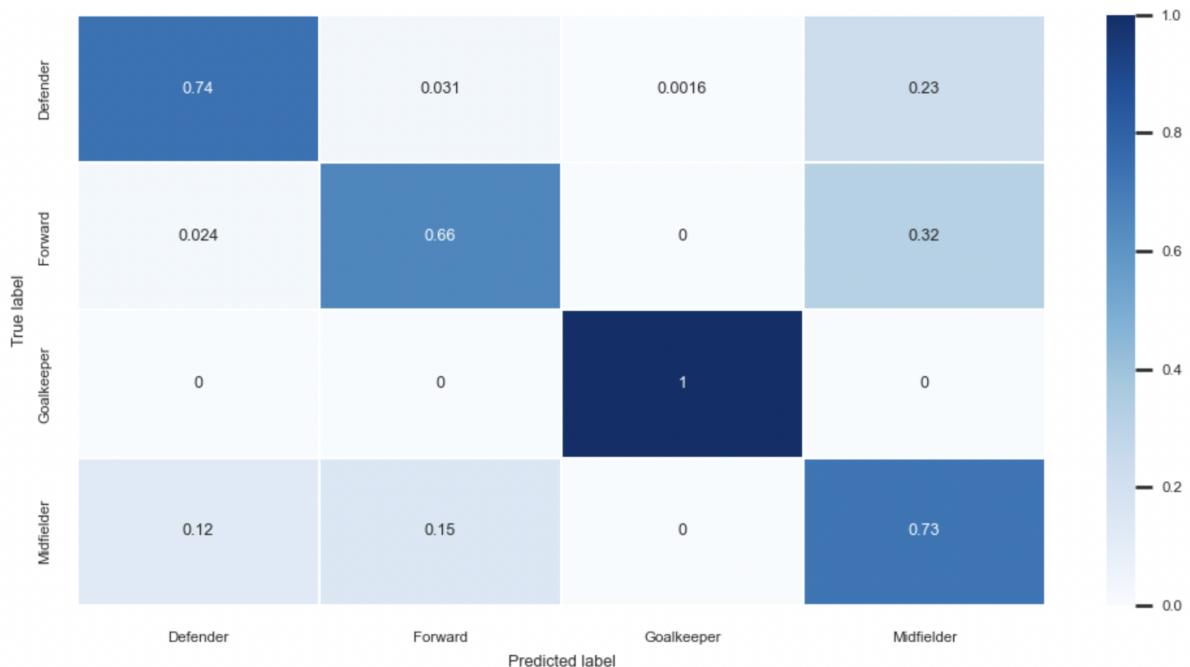
A.2 Évaluations des performances - Régression Logistique

Classification Report:

| | precision | recall | f1-score | support |
|---|-----------|--------|----------|---------|
| 0 | 0.84 | 0.74 | 0.78 | 640 |
| 1 | 0.64 | 0.66 | 0.65 | 334 |
| 2 | 0.99 | 1.00 | 1.00 | 129 |
| 3 | 0.66 | 0.73 | 0.69 | 681 |

| | | | |
|--------------|------|------|------|
| accuracy | 0.74 | 1784 | |
| macro avg | 0.78 | 0.78 | 1784 |
| weighted avg | 0.74 | 0.74 | 1784 |

Cross Validation:
Score moy $7.409e-01$ (+/-) $2.044e-02$

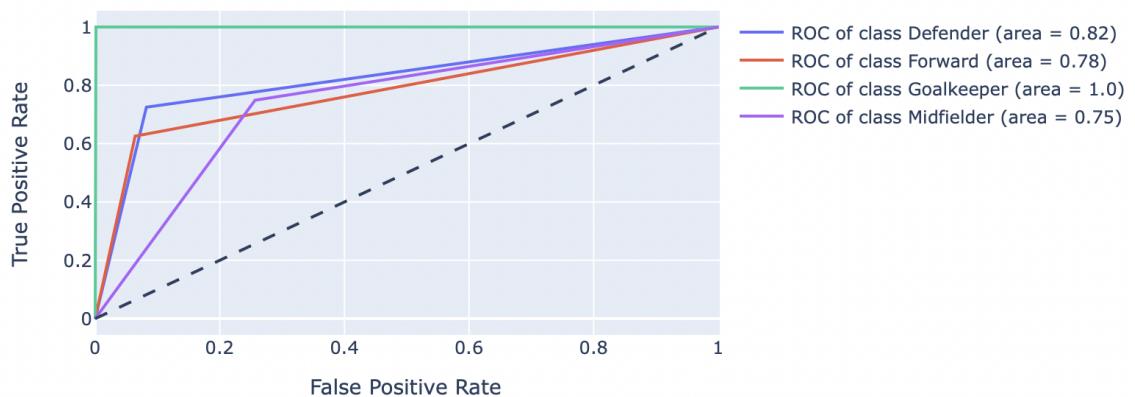
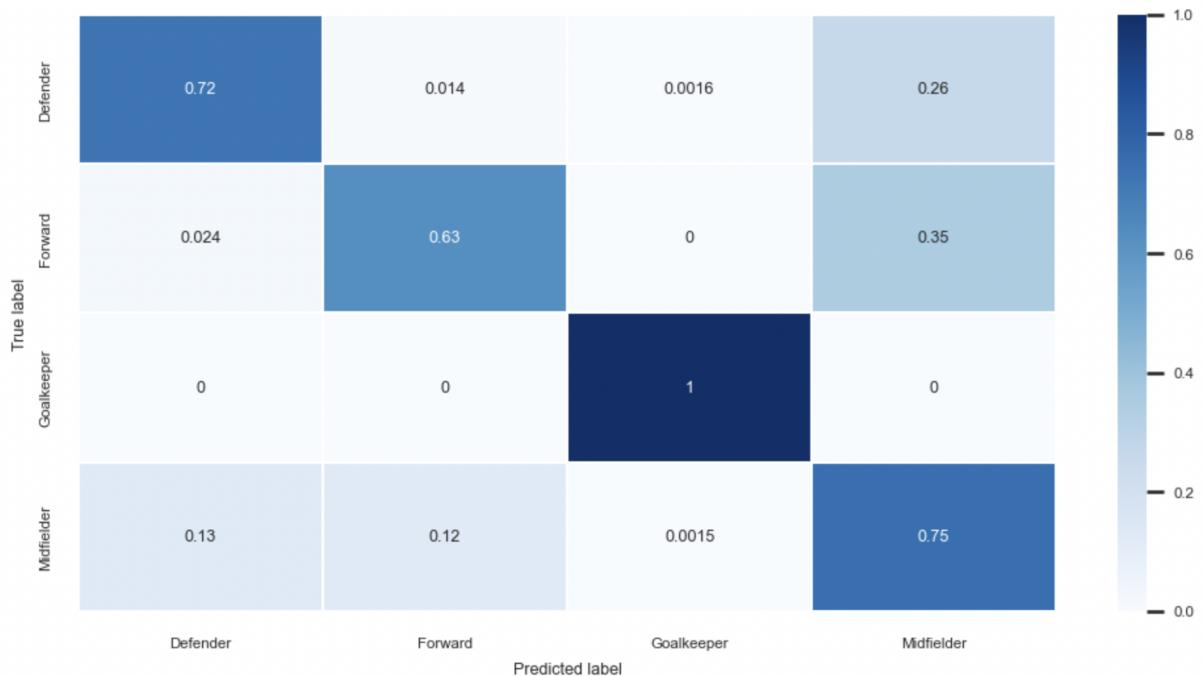


A.3 Évaluations des performances - Support Vector Machines

```
Classification Report:
precision    recall   f1-score   support
          0       0.83      0.72      0.77      640
          1       0.69      0.63      0.66      334
          2       0.98      1.00      0.99      129
          3       0.64      0.75      0.69      681

   accuracy                           0.74      1784
  macro avg       0.79      0.77      0.78      1784
weighted avg       0.74      0.74      0.74      1784

Cross Validation:
Score moy 7.440e-01 (+/-) 1.826e-02
```



A.4 Évaluations des performances - Random Forest

Classification Report:

| | precision | recall | f1-score | support |
|--------------|-----------|--------|----------|---------|
| 0 | 0.80 | 0.74 | 0.77 | 640 |
| 1 | 0.69 | 0.58 | 0.63 | 334 |
| 2 | 0.98 | 0.99 | 0.99 | 129 |
| 3 | 0.64 | 0.73 | 0.68 | 681 |
| accuracy | | | 0.73 | 1784 |
| macro avg | 0.78 | 0.76 | 0.77 | 1784 |
| weighted avg | 0.73 | 0.73 | 0.73 | 1784 |

Cross Validation:
Score moy 7.419e-01 (+/-) 1.413e-02

