ESILV
ENGINEERING SCHOOL
DE VINCI PARIS

Premier League

LIVERPOOL'S EVOLUTION WITH JURGEN KLOPP

# Cahier des charges PTS DIA 8

L.F.C.

*Équipe :*
KUOCH JACKY
TRANG THOMAS
NASSAR IBRAHIM

*Encadrant :*
M. OULED DLALA IMEN

Promotion 2022

# Remerciements

Nous souhaitons exprimer notre gratitude aux personnes qui nous ont aidé et qui ont contribué à la réalisation de notre projet.
Nous remercions tout d'abord Madame OULED DLALA pour ses précieux conseils, sa bienveillance et pour nous avoir encadré tout au long du projet.
Nous remercions également les intervenants externes notamment les data scientists contactés sur les réseaux sociaux pour nous avoir conseillé et aidé pour la construction de notre projet.

# Contents

# 1 Présentation du projet

## 1.1 Problème

A l'ère du **Big Data**, les acteurs majeurs de l'économie mondiale se doivent d'être à jour avec la révolution de la donnée. La collecte massive de données devient un élément essentiel à prendre en compte pour établir une stratégie efficace, productive et adaptée aux besoins. Cette nouvelle période s'accompagne de l'apparition de nouveaux métiers qui sont aujourd'hui essentiels dans toute grande entreprise sur les différents marchés. Il est alors impossible de ne pas avoir dans ses rangs **des Data Engineers, Data Scientists ou des Data Analysts.** Cette vérité s'applique également au domaine du sport, c'est alors que naît la notion de **Sport Analytics.** Une notion encore peu démocratisée en Europe, et particulièrement en France.

## 1.2 Contexte

L'idée de réaliser ce projet est née d'un constat. Dans l'histoire du club, **Liverpool a été sacré 19 fois champion d'Angleterre.** Cependant avant le titre de la saison 2019/2020, le club n'avait plus été champion depuis 20 ans. Depuis le début de années 2010, Liverpool côtoyait rarement le haut de tableau du championnat anglais. L'arrivée de **l'entraîneur allemand Jurgen Klopp** en 2015 a marqué le début d'un changement de dimension pour le club. Cette arrivée coïncide avec **l'expansion de la Data dans le monde du football** et notamment en Premier League. De nombreuses données sont désormais exploitées et rendues publiques pour une utilisation par les médias, les équipes ou les supporters.

Aujourd'hui, les trois fans de football, mais aussi futur ingénieurs que nous sommes, voyons cette expansion de la Data comme une opportunité unique **d'allier notre projet professionnel avec notre passion commune qu'est le football.** Dans cette optique, nous avons depuis un certain temps l'objectif de réaliser diverses études de Sport Analytics en appliquant nos connaissances acquises en cours et dans nos entreprises à des domaines qui nous passionnent. Nous pensons avoir désormais des outils et des compétences pour étudier les données mises à notre disposition. L'importance de ces données s'est accentuée encore plus aujourd'hui avec la **dimension économique** que représente un sport comme le football dans lequel un joueur et/ou une équipe sont jugés sur des critères humains mais aussi statistiques.

## 1.3 Problématique

**COMMENT EXPLIQUER, À L'AIDE DE LA DATA, L'INFLUENCE DE JURGEN KLOPP SUR LE RETOUR DE LIVERPOOL AU SOMMET DU FOOTBALL EN EUROPE ?**

À l'aide de nombreuses données récoltées, étudions **l'influence de Jurgen Klopp**, entraîneur renommé pour appliquer sa propre philosophie de jeu à ses équipes, sur Liverpool. Nous appliquerons nos connaissances en matière de Data pour comparer les performances de Liverpool sous Klopp avec celles des entraîneurs précédents et comprendre les points sur lesquels Klopp a appuyé pour faire évoluer le club et sa façon de jouer.

## 1.4 Objectifs

Notre projet consisterait donc à réaliser une **analyse de données** permettant de comprendre l'évolution de Liverpool ces dernières années. Nous utiliserons nos multiples connaissances en matière de Data pour réaliser une étude permettant de mieux cerner l'évolution d'un point de vue Data.
Cette étude serait constituée de différentes phases :

- **Récupération de données, data scrapping et tri des données.**
  Source : Site webs avec data en opensource (Fbref.com, Understat.com)
  Outils: Python (Selenium, beautifulsoup, pandas)
  Extraction de données à partir de sites internet spécialisés dans les données statistiques du football.

- **Création de jeux de données et définition des paramètres pertinents.**
  Outils : Fichiers .csv
  Création de différentes tables en fonction des types de données dans une optique d'exploitation.
  Structuration des données sous forme de tables.

- **Application de modèles statistiques et implémentation d'algorithmes de Machine Learning.**
  Outils : Python, Librairie SciKitLearn
  Implémentation de modèles linéaires à plusieurs variables de Machine Learning permettant l'analyse des données.

- **Analyse des réponses, reporting de ces données, Rédaction d'un rapport et exposition des résultats trouvés.**
  Outils : Overleaf, SciKitLearn
  Interprétation des résultats et rédaction d'un article résumant l'ensemble de notre étude.

# 2 État de l'art

## 2.1 Sport Analytics

Plus que jamais, une équipe sportive est aujourd'hui une entreprise. Il est donc important de pouvoir **utiliser les données à leur disposition dans le but d'améliorer leurs performances.** Mais contrairement à une entreprise classique, on différencie deux types d'analyses de données.

Tout d'abord on distingue l'analyse de données non-sportives **(off-field analytics)** qui ont un but économique et financier. Cela se manifeste à travers des retours sur les campagnes marketing, de sponsoring ou bien les événements dédiés à promouvoir l'image du club. Ici, l'objectif est clairement d'avoir une meilleure rentabilité et de **générer plus de bénéfices d'un point de vue financier** via les différents canaux de distribution.

D'un autre côté nous avons l'analyse des données sportives **(on-field analytics)** de l'équipe. Cette analyse a pour objectif d'améliorer les performances sportives de l'équipe à différentes échelles.
Les staffs techniques, ou les équipes Data pour certains clubs, **collectent des données en permanence sur les joueurs** durant les entraînements ou les matchs pour déterminer des pistes d'améliorations pour l'effectif professionnel mais aussi les équipes B ou les centres de formations par exemple.
On utilise également les données collectées sur les adversaires dans le but de préparer les différentes confrontations et augmenter les chances de victoire. Ces données peuvent aussi servir de support pour **la prise de décision** lors de l'achat ou la vente de joueurs ou de changements au sein de la structure sportive du club.

### Exemples de métriques propres au football

**L'ELO** est une mesure qui permet de comparer le niveau de différentes équipes. Cette valeur reflète l'état de forme d'une équipe à partir de nombreux critères tels que les résultats récents de l'équipe ou les adversaires rencontrés. Cette métrique s'appuie sur des données antérieures à la date donnée pour pouvoir estimer la force d'une équipe.
On s'appuie sur **la différence de ELO entre deux équipes qui s'affrontent pour déterminer un favori.** S'appuyer sur une telle métrique pour analyser les performances d'une équipe s'avère très utile pour observer si une équipe répond aux attentes et réussit à faire face à la pression.

**Les Expected goals (xG)** [3] sont une nouvelle métrique récemment développée, qui permet **d'estimer le nombre de buts qu'un club ou un joueur va marquer au cours d'un match.** Cette métrique est calculée à partir de nombreux facteurs. Cet indicateur se base sur la capacité d'une équipe à se procurer des occasions de buts, et dans un second temps, leur qualité à être efficaces pour marquer. Les facteurs qui font varier l'xG sont la distance et l'angle du tir, la partie

du corps utilisée pour tirer, la situation de l'action, le type de passe reçue ect ...
Nous souhaitons utiliser les xG à des fins comparatives. En effet, cette métrique permet de comparer les estimations avec les faits réels. Cette démarche a également pour but d'évaluer la manière dont Liverpool aborde les matchs et comment Liverpool est désormais plus attendu dans l'approche de ses maths.
Le principe de xG est beaucoup utilisé dans les paris sportifs pour réaliser les côtes des équipes qui s'affrontent mais également en interne dans les clubs.

### Moneyball

Moneyball[1] est le nom attribué à l'approche moderne analytique, mise en place par Billy Beane, directeur général des Athletics d'Oakland, pour monter une équipe compétitive en Ligue majeure de Baseball malgré la situation financière défavorable de la franchise.

Cette approche est **une approche statistique du baseball**, soutenant l'hypothèse selon laquelle une équipe réalise de meilleures performances lorsque ses bases sont **fondées sur des analyses de données.**

À partir de données récoltées et exploitées à **l'aide d'ingénieurs et de méthodes statistiques**, Billy Beane va construire une équipe compétitive à partir de joueurs sous-évalués et pour certains, sans contrats, Cette équipe, forgée à l'aide de statistiques avancées, va finir par concurrencer les plus grandes équipes de la Ligue avec des budgets nettement supérieurs. Finalement, ces dernières finiront par adopter cette approche qui aujourd'hui, se développent dans d'autres sports et notamment dans le football.

Cette approche, ayant connu un tel succès dans divers sports, a fait l'objet de nombreux ouvrages ou de films comme le film Moneyball dont est tirée l'image 1 ci-dessous et dans lequel on suit l'histoire de Billy Beane qui va tenter de restructurer son équipe après une perte de nombreux joueurs importants et avec un budget limité.

Figure 1: Scène du film Moneyball retraçant l'histoire de Billy Beane

### The Numbers Game

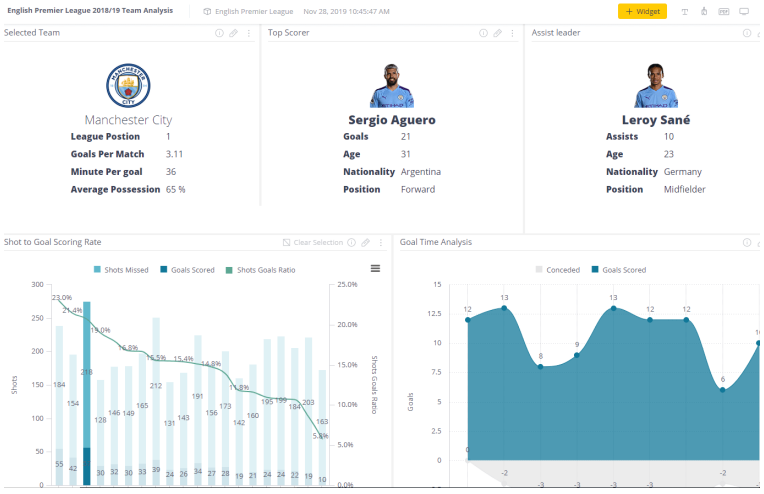
The Numbers Game [4] est un reportage retraçant l'évolution de la Data et les changements qu'elle entraîne aujourd'hui dans le monde du football. Couplées à la Data, **les équipes Data des clubs de football** entraînent des **intelligences artificielles** avec différents objectifs : préparer au mieux les entraînements, débriefer les matchs, cibler des talents à recruter et améliorer des aspects du jeu de l'effectif.

La collecte des données, associée aux IA, est comparable au rôle des émissaires ou **"scouts"** des équipes qui sont envoyés quotidiennement aux quatre coins du monde pour **observer et analyser des joueurs.** Ces données permettent d'éviter de passer à côté de talents qui, par le passé, n'auraient pas été détectés étant jugés uniquement sur des impressions et des observations subjectives. Elles permettent **d'extrapoler des données** pour étudier des performances d'un joueur dans un autre contexte.

Un exemple très parlant étant celui du joueur français N'Golo Kanté qui a explosé aux yeux du monde à un âge avancé dans le football (24 ans) suite à un transfert dans un club anglais dans lequel il a pu pleinement faire parler son talent. Lui, qui n'avait jusque là, **pas eu l'opportunité d'évoluer dans un environnement propice, n'avait pas été repéré par les grandes équipes** qui se l'arrachent aujourd'hui.

D'autres algorithmes de Machine Learning peuvent jouer **le rôle d'assistant technique** en soutien des entraîneurs en fournissant des pistes d'amélioration sur le style de jeu à adopter comme nous pouvons le voir dans la figure 2. À travers ces études et des nouveaux outils technologiques tels que des **balises portées par les joueurs** durant les entraînements ou des **caméras placées** enregistrant en continu des images des joueurs, les équipes peuvent non seulement améliorer leurs performances sportives, mais aussi extra-sportives liées à l'aspect économique du football.

Figure 2: Exemple de dashboard répertoriant les données de deux joueurs

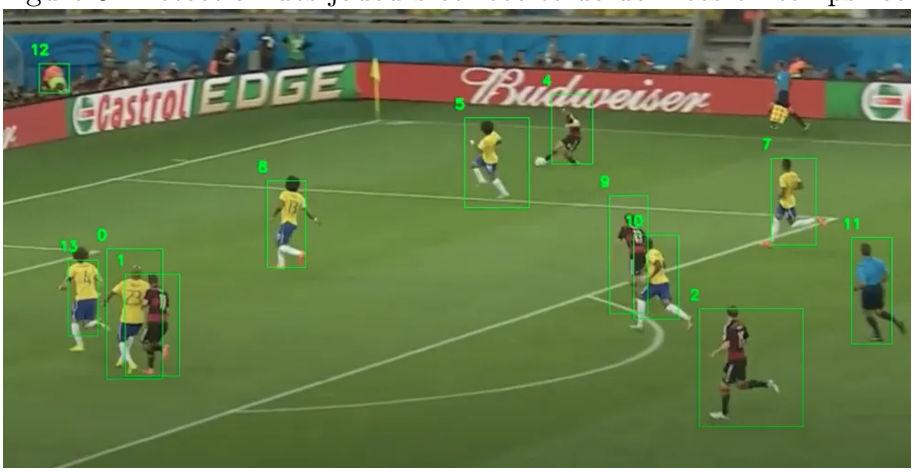


### Visionnage automatique

Le but de cette intelligence artificielle est de reproduire des fonctions du système de vision humain. Pour cela, des modèles de **Deep Learning** ont été appliqués pour détecter les positions, actions, les mouvements et les réactions de protagonistes.

Ce modèle a été entraîné de manière très intensive pour pouvoir être opérationnel, à travers les **algorithmes de réseaux de neurones.** Cela a dépassé les capacités de l'oeil humain étant donné que le réseau de neurones est capable de se concentrer sur toutes les parties de l'image. Les modèles d'entraînement **décomposent l'image** en plusieurs parties notamment grâce aux formes, couleurs (à travers les pixels), repères sur le terrain qui permettent à l'IA de mieux se repérer comme sur l'image à la figure 3.

Figure 3: Détection des joueurs et récolte de données en temps réel




Depuis moins de 5 ans, le développement du **visionnage automatique** commence à s'installer dans tous les sports majeurs. Cette innovation est utilisée à de nombreuses fins, notamment à **collecter les données comme le nombre de passes, nombre de kilomètres courus par joueurs, tirs et nombreuses autres statistiques.** Ces tâches étaient réalisées à l'époque par des personnes, ce qui engendrait donc des incertitudes sur les données dues aux erreurs humaines. Pouvoir développer un tel outil **facilite la collecte de données et réduit donc le nombre d'erreurs.** Une des principales forces d'un tel outil est de pouvoir collecter des données sur des matchs où il était auparavant compliqué d'envoyer des émissaires et grâce à cela, nous obtenons une plus **grande base données** sur laquelle nous pouvons étudier les potentiels futures stars de notre sport et ne plus passer à côté de jeunes joueurs prometteurs. En effet, les données récoltées aujourd'hui sont meilleures car elles sont moins soumises aux **erreurs humaines.** L'usage du Deep Learning pour réaliser la récolte de ces données est justifiée et nécessaire, en effet aucune autre technologie n'aurait pu réaliser cette tâche qui est très complexe.

## 2.2 Machine Learning

***Using Machine Learning to understand why Real Madrid have been so poor in La Liga in the last decade*** **[2]**

Brandon Dominique a cherché à comprendre pourquoi durant la dernière décennie, le Real Madrid, pourtant si dominant sur la scène européenne, n'a pas réussi à s'imposer sur la durée à l'échelle nationale. Pour cela, il a décidé d'utiliser le **Machine Learning** en créant deux datasets, un pour le Real Madrid et un pour le FC Barcelone, regroupant chacun des **données sur les adversaires** face auxquels les deux équipes ont perdu des points entre les saisons 2009-10 et 2017-18. Pour trouver des résultats il utilise l'algorithme des **K-Means Clustering** pour déterminer le profil des équipes contre qui ces deux équipes rencontrent des difficultés.

Ses **datasets** comportent les données suivantes : la date du match, le nombre de jours de repos, la localisation des matchs, le niveau des adversaires basés sur leur **elo et leur classement en fin de saison**, les **côtes** d'une victoire du Real ou de Barcelone, les **xG** (à partir de la saison 2014-15) et finalement le **nombre de points remportés** (0 pour une défaite et 1 pour un match nul).

L'algorithme des K-Means Clustering est un algorithme **d'apprentissage non-supervisé** permettant de diviser nos observations en K partitions. Il permet d'analyser un ensemble de variables afin de **regrouper les données similaires en clusters.** Couplé à cet algorithme, Brandon a utilisé la Silhouette Analysis pour déterminer le K et ainsi avoir un résultat le plus adapté possible à ces données.

L'étude est divisée en **sous parties** pour pouvoir analyser différents aspects : les matchs avant la saison 2014-15 (sans xG), les matchs après 2014-15 (avec xG), les données regroupées selon les entraîneurs des deux clubs et les saisons regroupées selon une victoire ou non du championnat en fin de saison. Un des **inconvénients** de cette étude, qui est un inconvénient très récurrent dans les études statistiques sur le sport. En effet, lors des analyses, nous faisons **l'hypothèse** selon laquelle les équipes jouent avec leur **effectif complet** puisqu'il est difficile de connaître à travers la donnée l'état de forme ou de santé de l'effectif.

Le fait d'avoir ajouté **le classement en fin de saison des équipes** en plus du elo pour déterminer le niveau d'une équipe est un choix très pertinent qui a permis une meilleure conclusion sur le type d'équipes face auxquelles le Real Madrid perdait des points. Et enfin, le fait de ne pas avoir étudier les matchs où **le championnat était déjà mathématiquement remporté ou perdu** est un choix cohérent avec la réalité footballistique où les équipes ont tendance à ne pas jouer à fond une fois le titre remporté ou perdu.

Cette étude a permis de déceler des ressemblances dans les différentes saisons où le Real Madrid n'a pas su remporter le titre, à savoir :

- Le fait de perdre des points face à ses concurrents directs.
- Ne pas obtenir de bons résultats à l'extérieur face aux équipes de milieu de tableau.
- Le fait de se créer beaucoup d'occasions mais de ne pas les concrétiser.
- Ne pas bien débuter ces saisons.

***Predictive analysis and modelling football results using Machine Learning approach for English Premier League***

Le but de cette étude[5] est de **prédire les résultats des matchs de Premier League.** Pour cela, les deux chercheurs se sont basés sur les **données du championnat de 2005 à 2016.** Le but est donc d'établir le meilleur modèle de classification de données parmi ceux existants. Nous avons ici différents types d'algorithme **d'apprentissage supervisé,** le but étant de trouver le plus efficace pour prédire les résultats des matchs. Pour entraîner les modèles, ils ont décidé de se baser sur les résultats de 2005 à 2014. Les différents algorithmes de classification utilisés sont : **Naïve Bayes, SVM, Random Forest et Gradient Boost.**

Les données utilisées sont divisées en 2 parties : **données par équipe et données calculées entre 2 équipes.** Pour chaque équipe nous avons des données tels que le nombre de corners, but ou encore les tirs cadrés. Nous avons aussi des notes pour chaque équipe **extraites du jeu vidéo Fifa.** En plus des données propres à chaque équipe, pour chaque rencontre nous avons des champs calculés entre les données de chaque équipe. Par exemple pour chaque match, nous calculons la différence de buts marqués durant toute la saison ou encore la différence de la note de l'attaque de chaque équipe. Ces données calculées possèdent une meilleure distribution, ce qui est bénéfique pour les modèles de prédiction.

Les différents tests ont démontré que la classification des **matchs 'nuls'** sont les plus difficile car c'est le **résultat le plus rare** durant un match de foot. Cependant ce résultat a été mieux classé par les modèles ensemblistes que le modèle probabiliste. Comme on peut le voir sur la figure suivante 4 :

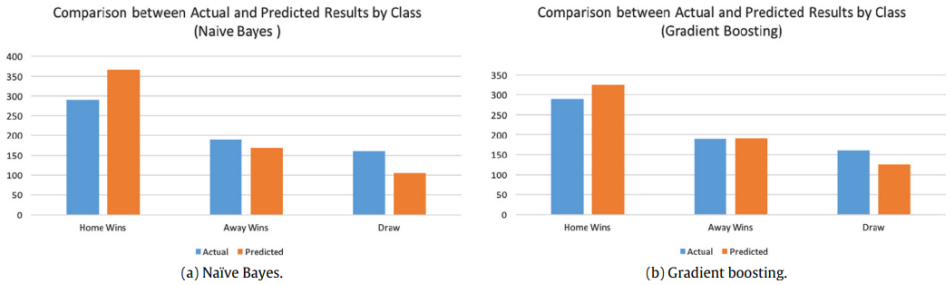



**Fig. 10.** Comparison between the true and predicted results by class for naïve Bayes and gradient boosting.

Figure 4: Comparaison des prédictions pour deux modèles différents

Nous avons donc une **moyenne de précision d'environ 60%** pour chaque classe à prédire pour les différents modèles. Nous avons donc presque une équivalence entre les différents modèles selon les hyperparamètres qui vont être choisis comme on peut le voir à la figure 5.

**Table 4**
Linear SVM results.

| (a) Confusion matrix | | | |
|---|---|---|---|
| | Predicted win | Predicted loss | Predicted draw |
| Actual win | 254 | 36 | 0 |
| Actual loss | 96 | 93 | 0 |
| Actual draw | 122 | 39 | 0 |
| (b) Precision–recall table | | | |
| | Precision | Recall | F1-score |
| Home wins | 0.54 | 0.88 | 0.67 |
| Away wins | 0.55 | 0.49 | 0.52 |
| Draws | 0.0 | 0.0 | 0.0 |

**Table 5**
RBF SVM results.

| (a) Confusion matrix | | | |
|---|---|---|---|
| | Predicted win | Predicted loss | Predicted draw |
| Actual win | 238 | 37 | 15 |
| Actual loss | 84 | 96 | 9 |
| Actual draw | 110 | 36 | 15 |
| (b) Precision–recall table | | | |
| | Precision | Recall | F1-score |
| Home wins | 0.55 | 0.80 | 0.67 |
| Away wins | 0.57 | 0.51 | 0.54 |
| Draws | 0.39 | 0.09 | 0.15 |

**Table 6**
Random forest results.

| (a) Confusion matrix | | | |
|---|---|---|---|
| | Predicted win | Predicted loss | Predicted draw |
| Actual win | 225 | 36 | 29 |
| Actual loss | 64 | 101 | 24 |
| Actual draw | 86 | 40 | 35 |
| (b) Precision–recall table | | | |
| | Precision | Recall | F1-score |
| Home wins | 0.60 | 0.78 | 0.68 |
| Away wins | 0.57 | 0.53 | 0.55 |
| Draws | 0.40 | 0.22 | 0.28 |

**Table 7**
Gradient boosting results.

| (a) Confusion matrix | | | |
|---|---|---|---|
| | Predicted win | Predicted loss | Predicted draw |
| Actual win | 222 | 37 | 31 |
| Actual loss | 58 | 99 | 32 |
| Actual draw | 88 | 31 | 42 |
| (b) Precision–recall table | | | |
| | Precision | Recall | F1-score |
| Home wins | 0.60 | 0.77 | 0.67 |
| Away wins | 0.59 | 0.52 | 0.54 |
| Draws | 0.40 | 0.26 | 0.31 |

Figure 5: Résultats des prédictions pour différents modèles

## 2.3 Data Visualization

Les études réalisées par les équipes Data des clubs de foot doivent être présentées aux entraîneurs et staff techniques. Ces derniers, parfois non initiés à l'univers de la Data, doivent pouvoir comprendre les différentes analyses fournies par les Data Scientists et Analysts du club. Pour cela, les équipes Data ont recours à la **Data Visualization** qui permet de mettre en forme, à travers des graphiques ou des dashboards, les résultats déduits des données récoltées.

### Nicolas Pepe

Jon Ollington, fan du club de football d'Arsenal, spécialiste en Data Visualization, profite de ses connaissances pour **analyser les performances** de son équipe favorite et rédige des articles pour décrire ses analyses. sur le site arseblog.news. Dans le cas présent, il cherche à comprendre les piètres performances, d'un point de vue statistique, de la nouvelle recrue du club, Nicolas Pepe, compte tenu des hautes attentes qu'avaient suscité son arrivée.

L'étude s'est focalisé sur **4 aspects statistiques** pour comparer les performances de Nicolas Pepe entre sa magnifique saison à Lille et sa première saison à Arsenal : les statistiques liées aux **Buts** (nombre de buts, nombre de tirs, nombre de tirs cadrés, le ratio de but/tir et les xG), les statistiques liées à la **Possession** (nombre d'occasions de tirs créées, nombre de touches de balles, dribbles réussis, nombre de fautes provoquées et de penaltys obtenus), les statistiques liées à la **Défense** (pressing, nombre d'interceptions, nombre de tacles réussis, et nombre de ballons récupérés) et les statistiques sur les **Passes** (nombre de passes décisives, les xA, les passes clés, les passes dans la partie de terrain adverse et le nombre de passes reçues).

Il a réalisé diverses représentations du terrain à partir de **plots programmés en Python** pour montrer des aspects du jeu qui ont évolués entre les deux saisons, à savoir **les circuits de passes**, son positionnement sur le terrain ou encore les tracés de ses actions de buts. Ces plots sont visuellement très parlants puisqu'ils permettent de visualiser un **réel changement** dans le rôle joué par Nicolas Pepe dans le dispositif d'Arsenal. Contrairement à son rôle central à Lille, il est désormais la troisième, voire quatrième option offensive dans le jeu d'Arsenal comme on peut le voir à la figure 6. Le jeu passe beaucoup moins par Pepe et son côté droit, il est beaucoup moins sollicité dans les circuits de passes et produit beaucoup moins **d'actions dangereuses.**

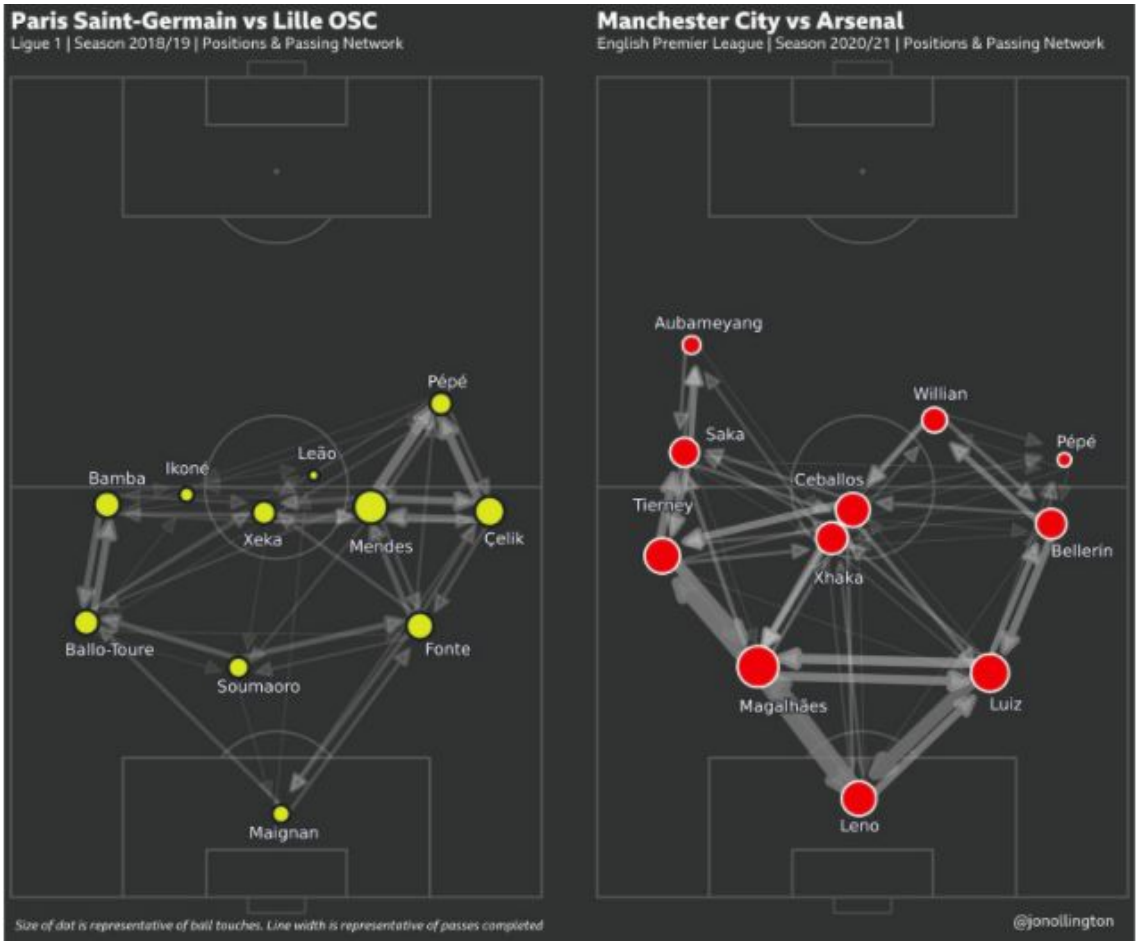



Figure 6: Comparaison des circuits de passes entre deux saisons de Nicolas Pepe

Le comparatif des statistiques sur les deux saisons permet de dresser un réel constat sur son adaptation à Arsenal. Il est beaucoup moins décisif au vue de ses nombres de buts, de passes décisive, nombres de dribbles tentés, réussis et nombres de passes reçues.

Cependant une telle comparaison entre deux saisons peut parfois conduire à de **mauvaises conclusions** étant donné les styles de jeu très différents que peuvent prôner les équipes et les entraîneurs. Il est aussi important de prendre en compte le fait que Nicolas Pepe découvre un nouveau championnat avec un **style de jeu très différent** de la Ligue 1 avec un entraîneur à la philosophie de jeu très prononcé, à savoir le tiki-taka qui ne correspond pas forcément aux qualités de Nicolas Pepe.

# 3 Récupération des données

## 3.1 Outils de Scrapping

### 3.1.1 Selenium

**Selenium** est une libraire de commandes disponible en divers langages de programmation permettant **d'automatiser des interactions** avec un navigateur internet. Selenium offre des méthodes pour effectuer les actions suivantes :

- Trouver un élément sur une page web.
- Cliquer sur élément.
- Naviguer à travers des pages web.
- Prendre des screenshots sur une page.

Selenium rend donc possible des fonctions telles que la vérification de l'existence d'un élément donné sur une page web, accéder automatiquement à d'autres pages en interagissant avec les éléments d'une page et **trouver les valeurs d'un élément donné.**

À l'aide de ces différentes fonctions, il est donc possible de récupérer des données sur des sites webs mettant à disposition des statistiques footballistiques. Cependant, récupérer toutes ces données peut parfois prendre beaucoup de temps. C'est pour cela qu'à l'aide de notebooks codés en Python, nous avons **automatisé de nombreuses tâches de scrapping** et pu récupérer un grand nombre de données.

## 3.2 Choix des source de données

### 3.2.1 Fbref.com

**Fbref** a permis de récolter des données sur des aspects de jeu ciblés. Fbref regroupe un grand nombre de statistiques avancées sur un seul même site. De nombreuses données mises à disposition sur Fbref sont offertes par StatsBomb (société récoltant des données et mises à disposition des clubs), et sont donc **uniquement disponibles sur ce site.**

### 3.2.2 Understat.com

**Understat** est le premier site à avoir répertorié les **Expected Goals**, ou plus communément appelés **xG**. Une des premières statistiques avancées sur le football permettant de déterminer le pourcentage de chances qu'une occasion se termine en but.
En sommant les xG de chaque joueur d'une équipe, on obtient les xG d'une équipe et déduisons ainsi le pourcentage de chance de gagner un match.

### 3.2.3 Goalzz.com

**Goalzz** a permis de récupérer l'ensemble des matchs de Liverpool, toutes compétitions confondues, depuis 2012 et les données principales sur les différentes rencontres. Les données récupérées ont servi de **base à la construction de notre base de données.**

### 3.2.4 Clubelo.com

**ClubElo** est le site à l'origine de la métrique **"Elo"** utilisée pour mesurer le niveau d'une équipe à une date donnée. A l'image du Elo utilisé dans le monde des échecs, le Elo permet de connaître la dynamique d'une équipe et son niveau à une échelle internationale puisque le Elo classe toutes les équipes du monde. Calculer la **différence de Elo** entre deux équipes lors d'un match permet de savoir si une équipe est favorite ou s'il s'agit d'un match dans lequel les deux équipes possèdent un niveau équivalent.

### 3.2.5 transfermarkt.com

**Transfermarkt** a permis de récupérer des données sur les arrivées et départs au sein de l'équipe de Liverpool. Transfermarkt fournit le **détail des transferts** à savoir les équipes d'arrivée et de départ et les montants de transferts. De telles données témoignent de **l'attractivité de Liverpool, la capacité à attirer des grands joueurs** mais aussi de la **confiance des dirigeants envers l'entraîneur** en place.

## 3.3 Données récupérées

| Données | Sources | Aspect de jeu recherché |
|---|---|---|
| Score mi-temps, score final, score par quart d'heure | Understat / Liverpool data all* | Temps forts récurrents durant un match |
| Date, nombre de jours de repos, résultat du match | Liverpool data all* | Temps forts récurrents durant une saison |
| Taux de possession, nombre de ballons touchés, nombre de passes, résultat | Goalzz / Fbref | Importance de la possession de balle et des passes |
| Nombre de tirs, nombre de tirs cadrés, taux de conversion, résultat | Goalzz | Capacité à se créer beaucoup d'occasions et à les concrétiser |
| Différence de Elo par match | ClubElo / Liverpool data all* | Capacité à remporter les rencontres face aux concurrents directs |
| Domicile/Extérieur, nombre d'occasions | Fbref | Influence d'Anfield sur les rencontres |
| Formations utilisées | Understat / Goalzz | Utilisation variée ou non de formations de jeu |
| Origine des buts, type de situations menant à des buts | Understat | Variété des situations de jeu menant à des buts |
| Stats défenseurs, stats gardien, nombre de buts encaissés, transferts | Fbref / Understat / TransferMarket | Importance accordée à la défense depuis l'arrivée de Jurgen Klopp |
| Fautes, tacles, cartons | Liverpool data all* | Agressivité des joueurs |

Liverpool data all* : Fichier Excel créé par Jacky à partir d'informations récupérées sur Kaggle couplées à des données récupérées sur différents sites webs.

# 4 Data Processing et Exploration des données

## 4.1 Data Processing

### 4.1.1 Nettoyage des données

Après avoir récupéré nos données, un **nettoyage des données fut nécessaire** dans les datasets récupérés. Pour ce faire, nous avons supprimé des colonnes jugées inutiles dans notre analyse. Nous avons modifié la colonne "Pourcentage de possession" à l'aide des fonctionnalités python pour retirer le "%" et uniquement conserver le pourcentage et le convertir en type de données int.

### 4.1.2 Création de variables

À partir des données récupérées, nous avons pu créer des variables supplémentaires utiles à nos analyses.

- **Matchday** : Cette variable numérique permet de connaître à quelle journée de championnat se déroule une rencontre. Une telle variable nous permet d'étudier la capacité de l'équipe à performer sur une certaine période d'une saison.
- **home0away1** : Cette variable permet d'encoder numériquement la variable at anfield (booléen) et de pouvoir étudier l'influence de la localisation du match dans les algorithmes nécessitant des variables numériques.
- **EloDiff** : À partir de l'API fournie par ClubElo, nous avons pu récupérer le Elo de Liverpool depuis l'existence de la métrique et nous avons par la suite réalisé un matching avec les dates des rencontres pour récupérer le Elo à une date donnée. Pour récupérer le Elo des adversaires de Liverpool à chaque match, nous avons du faire appel à Selenium pour automatiser la récupération du Elo du club adverse à chaque date de match de Premier League.
- **RestDays** : Nous avons cherché à connaître le nombre de jours de repos entre chaque match. Pour calculer cette donnée, nous avons eu recours à notre dataset récolté sur Goalzz.com dans lequel nous avions répertorié toutes les rencontres, toutes compétitions confondues.En faisant la différence entre les dates de chaque match, nous obtenions donc le nombre de jours de repos entre chaque match.
- **Points_won** : Dans un championnat, le classement des équipes se fait en fonction du nombre de points gagnés au fil des matchs. Cette colonne consiste à faire correspondre le nombre de points gagnés en fonction du résultat du match. L'équipe qui gagne le match remporte 3 points, s'il y a match nul les deux équipes remportent 1 point et l'équipe perdante remporte aucun point.

- **Buts cumulés encaissés et marqués** : Une saison de championnat national est composée de 38 matchs. Dans notre dataset, nous avons récupéré un total de 8 saisons. Cette colonne présente le nombre de buts cumulés qui ont été marqués et encaissés au moment où le match se joue dans la saison.

## 4.2 Exploration des données

Cette partie présente les principales explorations réalisées selon toutes les données que nous avons. Ces explorations ont été réalisées avec l'aide des librairies **Matplotlib** et **Seaborn**. Le but de ces explorations est de développer des intuitions concernant les observations qui seront retournées par l'application des modèles de Machine Learning.

Nous avons réalisé nos explorations de données selon différents axes de réflexion. Nous allons essayer de tirer le maximum d'interprétations et d'explications concernant l'évolution du jeu de Liverpool. Dans les différents graphiques qui suivent nous nous sommes basés sur le modèle suivant:

| | | |
|---|---|---|
| V | Victoire | 3 points |
| N | Nul | 1 point |
| D | Défaite | 0 point |

### 4.2.1 L'évolution des résultats de Liverpool

Pour commencer nous allons visualiser l'évolution et la répartition des résultats de Liverpool depuis la saison 2012-2013 comme nous le voyons à la figure 28.

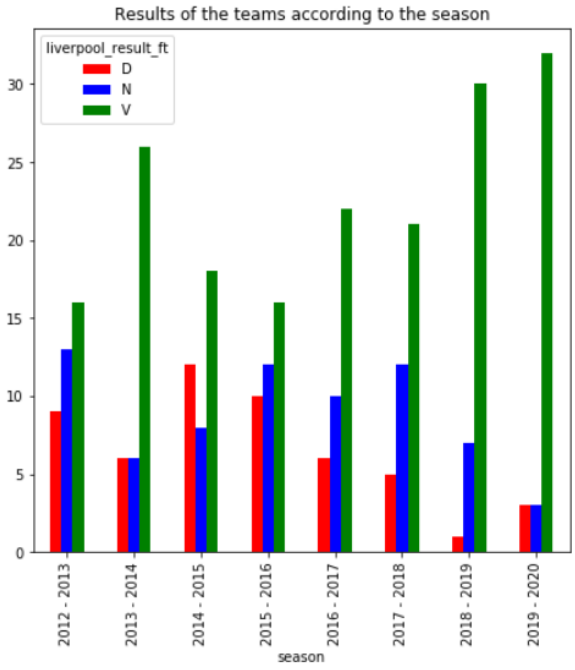



Figure 7: Résultats de Liverpool suivant la saison

Jürgen Klopp est arrivé aux commandes en Octobre 2015. Cette arrivée marque un **changement important** dans la dynamique des résultats du club. En effet, par rapport à son prédécesseur on voit **un taux de victoire en constante évolution et un taux de défaites qui diminue beaucoup depuis 2015-2016.** En général, il est assez dur pour un nouvel entraîneur de s'acclimater à son nouveau club, cependant Klopp a su en l'espace de seulement 1 an renverser la dynamique d'un club qui était déjà à l'origine un club historique, ce qui rendait sa mission d'autant plus compliquée.

Puis nous allons voir l'évolution de **l'influence de jouer à domicile pour Liverpool.** Il est important de spécifier que Liverpool est particulièrement connu pour la ferveur qui règne autour du club historique anglais. **Ayant un des meilleurs publics du monde, Liverpool puise énormément de force dans le soutien** apporté par ses supporters. Dans le prochain graphique, nous allons approfondir le graphique précédent en précisant si Liverpool joue à domicile. Pour cela, la variable que nous avons crée à savoir : at_anfield (type : bool, **TRUE** si Liverpool joue à domicile) a été utilisée pour ajouter le détail à ce graphique 8.

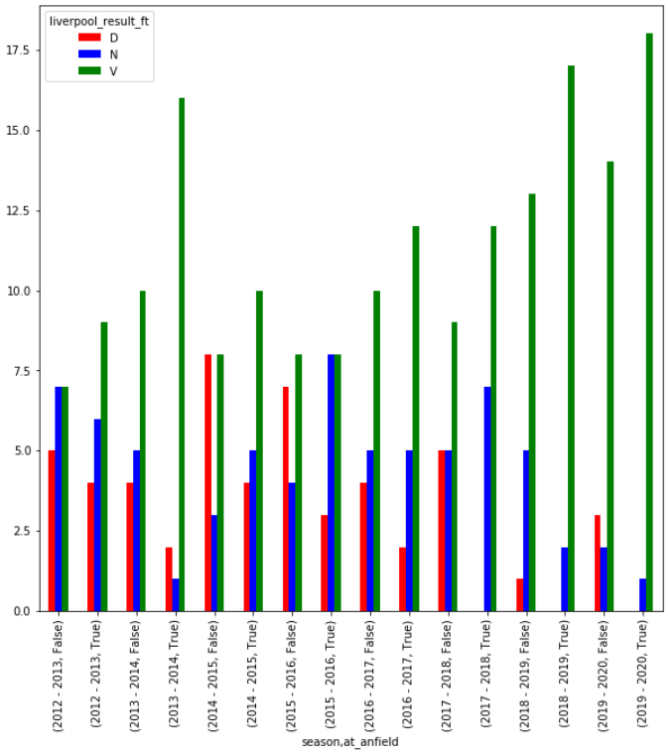



Figure 8: Résultats de Liverpool suivant la saison et le stade

Nous voyons que **la majorité des matchs gagnés sont à domicile.** Il est également très important de souligner qu'entre 2017 et 2020, Liverpool n'a connu aucune défaite à domicile ce qui est une grande prouesse, nous pouvons assimiler cela à la dynamique apportée par Klopp.

Le graphique 9 présente l'évolution du résultat du match selon le statut de Liverpool ( favori ou outsider).

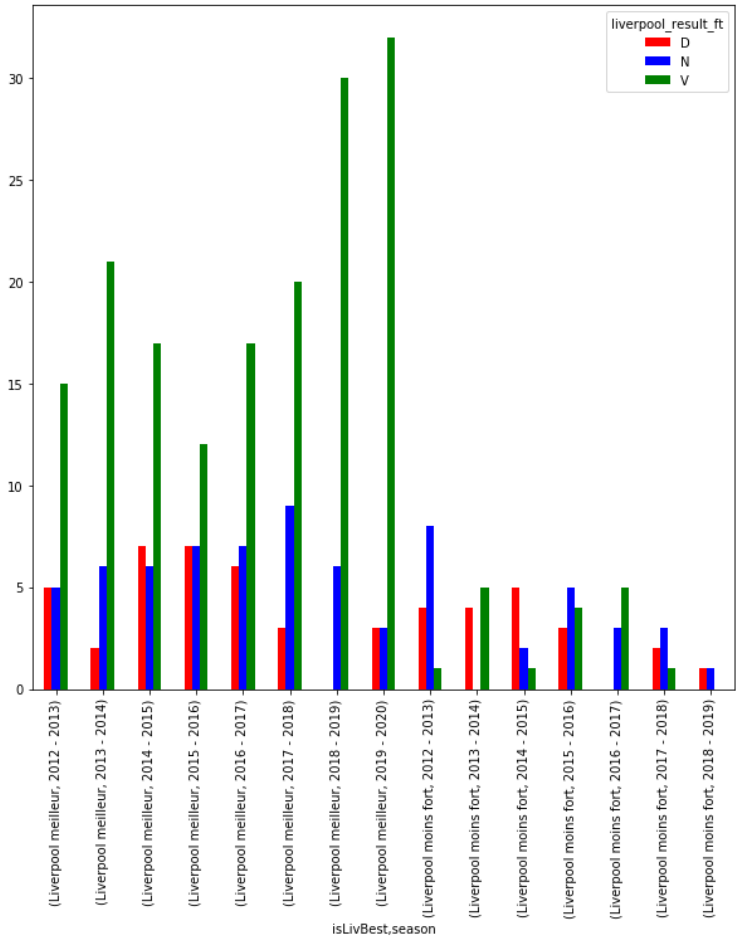



Figure 9: Résultats de Liverpool suivant la saison et la différence de Elo

Nous pouvons déduire de ce graphique que depuis l'arrivée de Klopp en 2015, l'équipe de Liverpool **s'en sort très bien quand elle est favorite** (c'est-à-dire meilleure que leur adversaire). Cependant on voit tout de même quelques défaites mais qui restent négligeables face au nombre de victoires.

NB : on constate que **durant la saison 2019-2020, Liverpool n'a joué aucun match sans être favori**, ce qui traduit leur suprématie à l'échelle nationale durant cette période.

### 4.2.2 Les caractéristiques du jeu selon le résultat

Nous avons opté pour des box plots dans cette partie pour pouvoir représenter la répartition des statistiques selon le résultat du match.

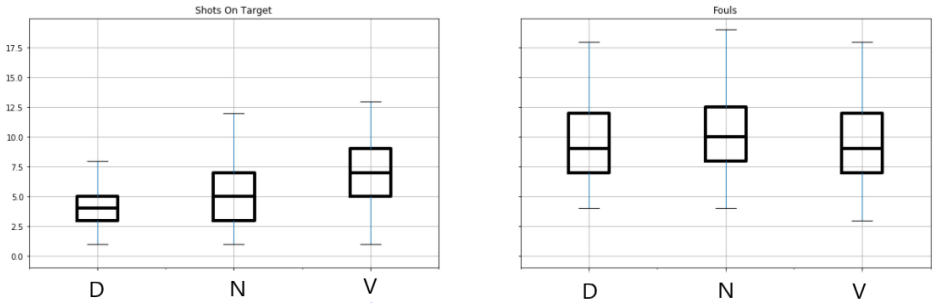



Figure 10: Nombre de tirs cadrés et fautes suivant le résultat

On voit à travers les représentations ci-dessus que **quand Liverpool gagne, la répartition de tirs cadrés est plus élevée** ce qui traduit leur domination offensive lors des victoires. En revanche quand ils perdent la répartition des tirs cadrés est très faible ce qui démontre **une incapacité totale à attaquer lors des défaites.** En ce qui concerne les fautes commises, on voit que la médiane du nombre de fautes est plus bas quand Liverpool gagne ses matchs. Ce qui signifie qu'ils jouent mieux sans faire de fautes.

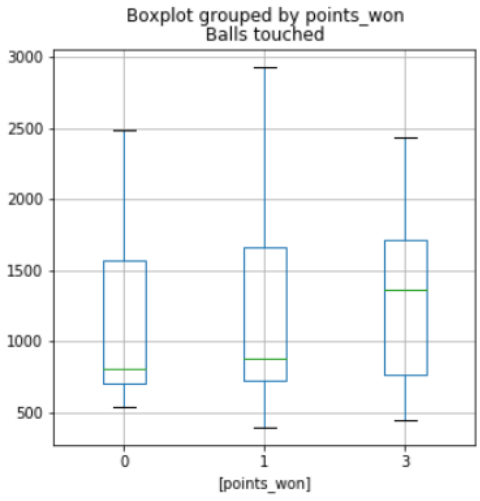



Figure 11: Nombre de ballons touchés selon le résultat

La différence de position de la médiane sur ce boxplot à la figure 11 est flagrante selon le résultat du match. On voit que quand le match est gagné la médiane est à environ 1400 ballons touchés tandis que pour une défaite, il y a environ 700 à 800 ballons touchés par match. On peut en déduire que **Liverpool a beaucoup de mal quand ils n'ont pas le contrôle du match.**

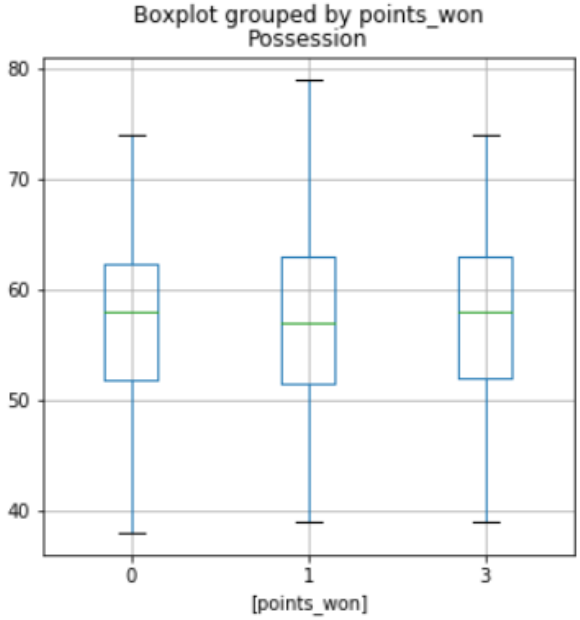



Figure 12: Possession de balle selon le résultat

Nous voyons dans le boxplot 12 ci-dessus une donnée qui nous éclaire mieux sur le gestion du ballon de Liverpool. En effet nous voyons que les écarts inter-quartiles sont approximativement égaux selon les différents résultats. **Cependant la médiane est plus haute quand Liverpool perd un match, ce qui traduit une possession stérile.** On en déduit que Liverpool est une équipe qui **aime généralement avoir le contrôle du ballon.**

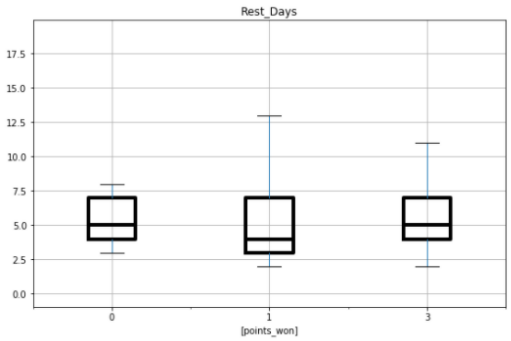



Figure 13: Nombre de jours de repos selon le résultat du match

Voilà une donnée surprenante à la figure 13, c'est le nombre de jours de repos entre chaque matchs. On voit que les médianes sont assez proches. On en déduit **qu'un bon nombre de jours de repos n'est pas forcément favorable à Liverpool et que l'enchaînement des matchs est un facteur clé** dans la réussite de Liverpool.

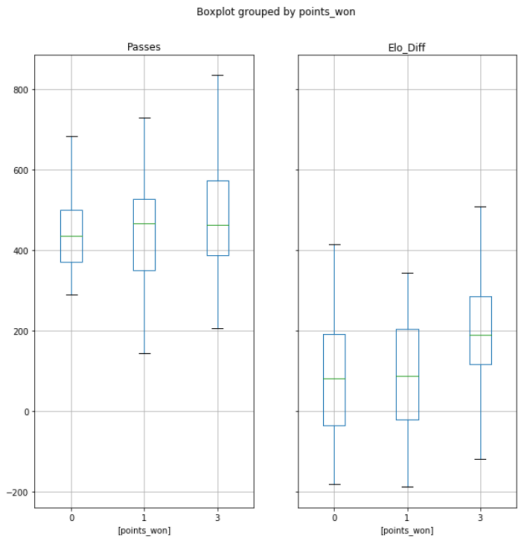



Figure 14: Nombre de passes et différence de Elo selon le résultat du match

Ci-dessus nous pouvons voir que **le nombre de passes n'influe pas forcément sur le résultat du match.** Cependant sur le boxplot à la figure 14 concernant le Elo_Diff (Elo_Liverpool - Elo_Adverse) que Liverpool a beaucoup plus tendance a gagner contre des équipes plus faibles avec une médiane a environ 200 contre 100 pour les défaites. On en déduit que Liverpool respecte généralement son rôle de favori.

### 4.2.3 L'apport de Jürgen Klopp dans le jeu

Depuis l'arrivée de Jürgen Klopp à Liverpool, le classement de Liverpool est en constante évolution. Lors de sa première saison en 2015-2016 il a fini 8ème du classement et en 3 ans il a réussi à se placer dans le podium (2ème en 2018-2019) avant d'être sacré champion en 2019-2020. La figure 15 compare les résultats de Klopp et de son prédécesseur. On voit que Klopp a :

- Un taux de défaites plus de 3 fois inférieur à celui de Rodgers.
- un taux de victoires de 82 % contre 53% pour Rodgers ce qui est plutôt significatif sur l'impact de l'entraîneur allemand.

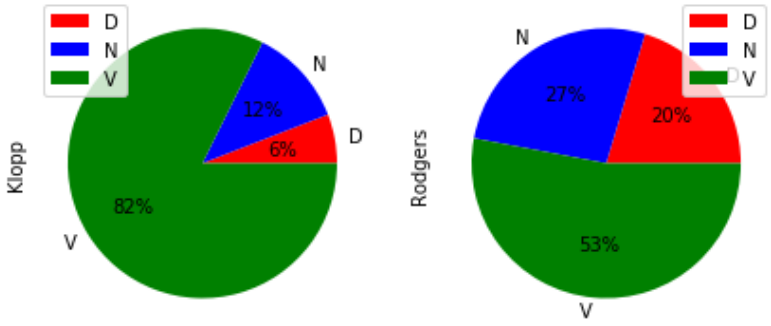



Figure 15: Répartition des résultats entre Klopp et son prédécesseur

Nous allons expliquer les changements que Klopp a instauré et comment il a transformé cette équipe. Pour cela, nous allons nous appuyer sur les statistiques dans le jeu et sur sa philosophie tactique.

Sur les boîtes à moustaches, figure 16 ci-dessous on a la répartition des tirs cadrés et des fautes commises par Liverpool sous les deux entraîneurs. La médiane des tirs cadrés sous Klopp est d'environ 7 tirs cadrés par match tandis que pour Rodgers elle est à 6, ce qui traduit **un penchant plus offensif sous le management de Klopp.**
En ce qui concerne les fautes commises, l'équipe sous Rodgers commettait environ 10 fautes par matchs contre 7.5 pour les équipes des Klopp, ce qui montre que Klopp ne privilégie pas l'agressivité mais plutôt une défense "propre" avec du sang froid.

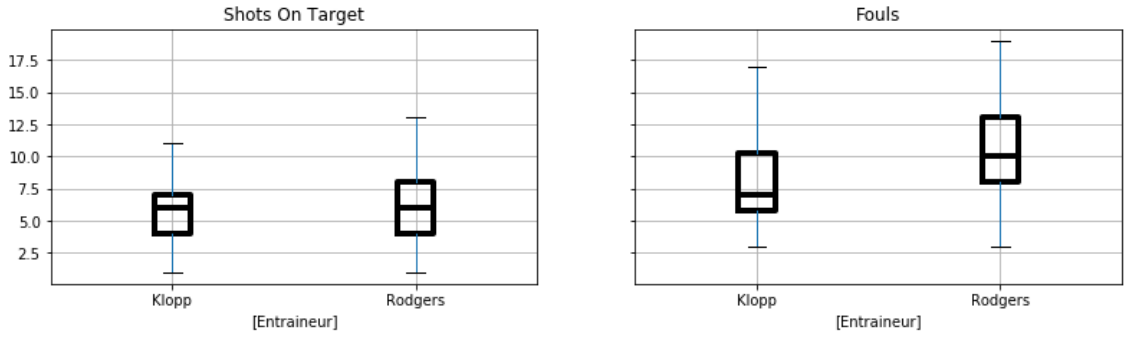



Figure 16: Taux de tirs cadrés et fautes commises par entraîneurs

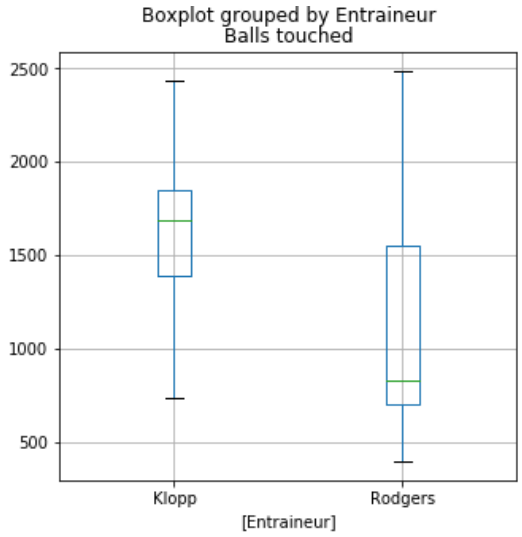



Figure 17: Nombre de ballons touchés par entraîneur

La figure 17 ci-dessus nous permets de voir l'aspect du contrôle du jeu de chaque entraîneur. La médiane de ballons touchés par matchs sous Klopp est de 1700 ballons/match contre 800 ballons/match représentant plus de la moitié, ce qui est énorme. Cela démontre que Jürgen Klopp a instauré un **jeu plus basé sur la maîtrise et la main mise sur le ballon** que Rodgers.

La figure 18 concernant la possession du ballon selon l'entraîneur confirme nos propos précédents quand on voit la possession médiane sous Klopp qui est de 62% contre 57% sous Rodgers. Cela vérifie bien que **Klopp a la main mise sur le jeu.**

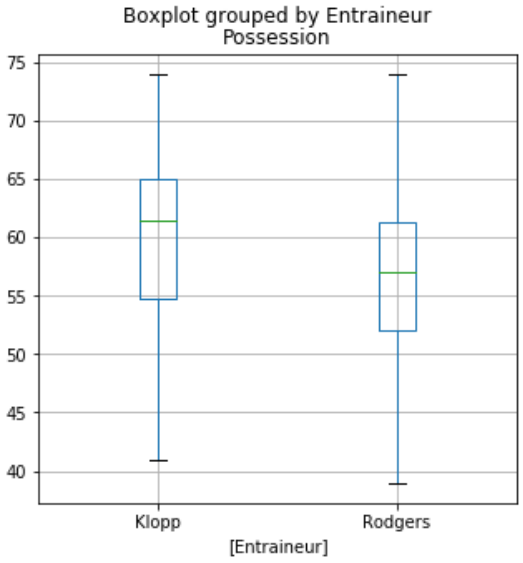



Figure 18: Possession par entraîneur

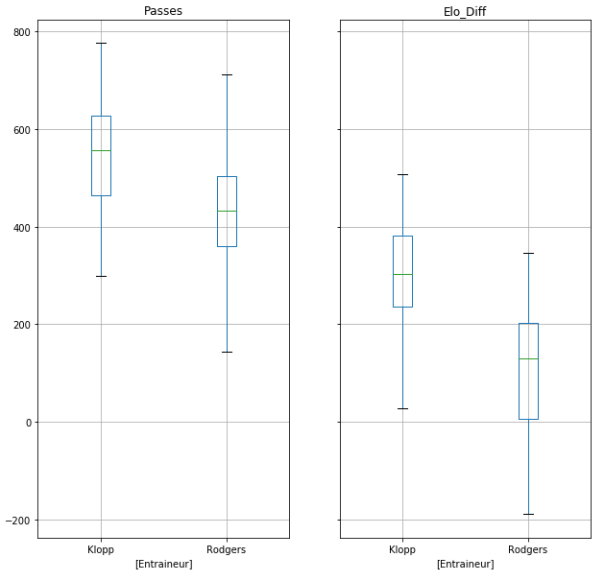



Figure 19: Passes et Elo_Diff par entraîneur

Enfin les dernières explorations que nous verrons ci-dessus concernent le nombre de passes et la différence de Elo.

- Le nombre de passes des équipes de Klopp tournent autour de 580 contre 420 pour Rogders, ce qui souligne une nouvelle fois la disparité entre les deux types de contrôles du jeu.
- La différence de Elo est la différence entre le Elo de Liverpool et de son adversaire, plus le Elo_Diff est haut, plus Liverpool est favori. Si Liverpool est favori cela signifie qu'ils sont supérieurs à leur adversaire. Le Elo_Diff a une médiane de 300 sous Klopp contre 150 pour Rodgers. On peut conclure que sous Klopp, les équipes de Liverpool sont bien meilleures que sous Rodgers d'un point de vue statistique à l'échelle européenne puisque le Elo est un classement international.

# 5 Choix des algorithmes

## 5.1 Choix des données

Dans cette section, nous allons nous concentrer sur les algorithmes de Classification que nous avons utilisé pour déterminer le résultat des matchs de Liverpool. Notre colonne cible ici est 'liverpool_result_ft'. Ce champs contient 3 classes distinctes : Victoire, Nul et Défaite.

La liste des modèles de classification que nous avons testés sont :

- Modèles linéaires
  - Modèles discriminants
    - Support Vector Machine
    - Logistic Regression
  - Modèles génératifs
    - Linear Discriminant Analysis
    - Gaussian Naïve Bayes
- Modèles ensemblistes
  - Avec Bootstrap
    - Random Forest
  - Sans Bootstrap
    - Extra Trees

Pour cette première comparaison, les hyperparamètres de chaque algorithme sont initialisés par défaut.

Les features sélectionnées pour prédire les classes sont les suivantes :

1. Tirs
2. Sauvetages du gardien
3. Centres
4. Fautes
5. Hors-jeu
6. Corners
7. Cartons jaunes
8. Cartons rouges
9. Passes
10. Interceptions
11. Contres
12. Arrêts du gardien
13. Tacles
14. Nombre de jours de repos
15. Différence de Elo
16. Localisation du match encodée

Ces colonnes nous semblaient être les statistiques essentielles et pertinentes à la compréhension et à la représentation, par les chiffres, d'une rencontre footballistique. De plus, ces features ont été sélectionnées en construisant dans un premier temps une matrice de corrélation. À l'aide de la matrice construite, nous avons procédé à une sélection de données en éliminant les variables ayant une forte corrélation entre elles pour éviter les problèmes de multicolinéarité.

## 5.2 Classification

Les algorithmes de classification font partie de la famille **des algorithmes d'apprentissage supervisé.** En fournissant un ensemble de données préparées et étiquetées du résultat attendu, l'algorithme apprend de chaque exemple en ajustant ses paramètres de façon à diminuer l'écart entre les résultats obtenus et les résultats attendus. **La marge d'erreur se réduit ainsi au fil des entraînements**, avec pour but, d'être capable de généraliser son apprentissage à de nouveaux cas.

### 5.2.1 Logistic Regression

La régression logistique est un algorithme de Machine Learning supervisé. Elle permet d'étudier les relations entre un ensemble de variables X et une variable cible Y. En supposant que Y est une variable de **Bernoulli** et X les données associés à cette variables, l'objectif est d'estimer la valeur suivante :

$$p(x) = P[Y = y|X = x] = E[Y|X = x]$$

On étudie donc la probabilité que Y soit égale à une certaine valeur y parmi les possibles classes de prédiction possibles.

### 5.2.2 Linear Discriminant Analysis (LDA)

**La Linear Discriminant Analysis** fait partie des techniques d'analyse discriminante prédictive. Il s'agit d'expliquer et de prédire l'appartenance d'un individu à une classe. La variable à prédire est forcément catégorielle (discrète), ici la victoire, la défaite ou le match nul. L'objectif est d' **établir une règle d'affectation qui permet de prédire, pour une observation donnée, sa valeur associée** à partir des valeurs prises par cette observation.

La LDA est étroitement liée à l'ACP (Analyse en composantes principales) et à l'analyse factorielle puisqu'elles recherchent les combinaisons linéaires les plus aptes à expliquer la donnée fournie.

### 5.2.3 Support Vector Machine

**Le Support Vector Machine (SVM)** est un modèle de classification ou de régression qui appartient aux **classificateurs linéaires**, c'est-à-dire qu'ils utilisent une séparation linéaire des données. Le modèle repose sur **la notion de frontière** pour délimiter le périmètre de chaque classe. Si les données ne sont pas linéairement séparables, on applique des transformations qui permettent de rendre cette séparation faisable. De plus, le but de cette manoeuvre est de maximiser la marge qui sépare les classes. Une des caractéristiques du modèle SVM est qu'il est plus performant quand il a peu de données d'entraînement.

### 5.2.4 Bagging Classifier

Le terme Bagging désigne la contraction de "Bootstrap Aggregation". C'est une méthode qui a pour but d'améliorer des classificateurs simples (comme les arbres de décisions). La bagging vise à réduire la variance du classificateur utilisé. Pour ce faire, on crée un certains nombre de nouveaux échantillons d'entraînement à partir de notre dataset, en utilisant le "bootstraping". Il s'agit de créer un nouvel échantillon aléatoire avec remise. En parallèle, un classificateur est initié à partir du nouvel échantillon aléatoire. La prédiction est définie par la classe la plus représentée parmi tous les classifieurs initiés.

- **Le Random Forest**, traduit par "forêt aléatoire" est un algorithme de classification ou de régression et se base sur le principe du **"tree bagging".** Le "tree bagging" consiste à construire de manière aléatoire n arbres de décision. Ce modèle effectue un apprentissage en parallèle sur plusieurs arbres de décisions entraînés de manière aléatoire. Pour faire sa prédiction ce modèle va mettre en commun la prédiction de chacun de ses sous arbres et prendre la classification qui aura la meilleure fréquence. Dans chaque arbre, nous utilisons qu'une partie des variables X. Nous avons décidé d'implémenter le Random Forest car :
  - L'implémentation est simple et intuitive.
  - Réduction du risque d'overfitting (sur-apprentissage).

  C'est pourquoi le Random Forest est un des modèles les plus utilisés pour traiter les problèmes de classification.

### 5.2.5 Extra Trees (Extremely Randomized Trees)

est un modèle très similaire au Random Forest. Ils se basent tous les deux sur un grand nombre d'arbres de décisions créés aléatoirement. La prédiction est donc réalisée de la même manière. Cependant on relève **deux différences majeures** entre les modèles :

- Dans les arbres d'un Extra Trees, la variable qui divise l'échantillon est choisie de manière aléatoire. Contrairement au Random Forest qui calcule la meilleure variable à utiliser.
- Dans un modèle Extra Trees, tous les classifieurs sont entraînés sur le dataset d'origine. Il n'y a donc pas de bootstraping pour le choix des échantillons.

### 5.2.6 Gaussian Naive Bayes

L'algorithme des Gaussian Naive Bayes est un type de classification bayésienne probabiliste basée sur le **théorème de Bayes** avec une forte indépendance des hypothèses.

Un classifieur bayésien naïf suppose que l'existence d'une caractéristique pour une classe, est indépendante de l'existence des autres caractéristiques. L'avantage de ce classifieur est qu'il requiert relativement peu de données d'entraînement pour estimer les paramètres nécessaires à la classification, à savoir moyennes et variances des différentes variables. En effet, **l'hypothèse d'indépendance des variables** permet de se contenter de la variance de chacune d'entre elles pour chaque classe, sans avoir à calculer de matrice de covariance.

## 5.3 K-Means

Contrairement aux algorithmes de classification, les K-Means appartiennent à la classe des algorithmes **d'apprentissage non supervisé.** Ces algorithmes doivent opérer à partir d'exemples non annotés. L'apprentissage par la machine se fait de manière entièrement indépendante sans que l'on fournisse des exemples de résultats.

Nous avons créé deux fonctions pour appliquer l'algorithme des K-means fourni par la libraire Sci-Kit Learn. Ces fonctions nous permettent d'afficher les résultats de l'algorithme et les valeurs pour chaque cluster. La seconde fonction permet de déterminer le nombre de clusters à utiliser pour chaque application de l'algorithme à l'aide de la Silhouette Analysis.

### 5.3.1 L'algorithme des K-Means

L'algorithme de K-means est une méthode de **partitionnement des données.** L'objectif de cet algorithme est de diviser un ensemble d'observations en k groupes, appelés clusters. Le clustering est une méthode **d'apprentissage non supervisé** avec l'objectif de trouver **des patterns dans des lots de données.** Ainsi les données similaires se retrouvent dans un même cluster.Deux données dites 'similaires' auront une distance réduite, alors que deux objets différents auront une distance de séparation plus grande.Dans le cas de **données quantitatives**, la distance utilisée le plus souvent en clustering est la distance euclidienne. C'est la distance géométrique classique définie telle que : Dans un espace vectoriel $E^n$, la **distance euclidienne** d entre deux observations $x_1$ et $x_2$ se calcule comme suit :

$$d(x_1, x_2) = \sqrt{\sum_{j=1}^{n} (x_1 n - x_2 n)^2}$$

### 5.3.2 La fonction K-Means

Nous avons donc créé une fonction pour appliquer l'algorithme des K-means. La fonction prend **en paramètres le nom d'un fichier et le nombre de clusters.** La fonction récupère la donnée stockée dans un fichier csv et applique l'algorithme sur le lot de données en cherchant à regrouper les données ressemblantes au sein des clusters. On récupère ensuite les moyennes pour chaque variable et chaque cluster pour les afficher ainsi que le nombre d'observations affiliées à chaque cluster.

```
def KMC (filename, n) :

    data = pd.read_csv(filename)
    kmeans = KMeans(n_clusters= n, random_state= 0)
    kmeans.fit(data)
    centers = kmeans.cluster_centers_

    for c in range (len(centers)) :
        print ("---------------------- Values for Cluster --------", c+1)
        for i in range (len(centers[c])) :
            print ((data.columns.values[i]), ": ", np.round(centers[c][i], 4))
        print (np.count_nonzero(kmeans.labels_ == c), " Games included in this
   cluster")
    return
```

### 5.3.3 Silhouette Analysis

La **Silhouette Analysis** est utilisée pour déterminer des coefficients de silhouette. Ce coefficient est la différence entre la distance moyenne avec les points du même cluster avec les points des autres clusters. Si cette différence est négative, le point est en moyenne plus proche du groupe voisin que du sien : il est donc **mal classé**. À l'inverse, si cette différence est positive, le point est en moyenne plus proche de son groupe que du groupe voisin : il est donc **bien classé**.

Ainsi, en modifiant le nombre de clusters et en calculant la **moyenne des coefficients de silhouette** pour toutes les observations on peut sélectionner le nombre de clusters qui sera le plus susceptible de décrire nos observations. C'est comme cela que nous avons sélectionné **le nombre de clusters** pour les algorithmes de K-means appliqués dans notre analyse.

Pour chacun de nos datasets groupés selon les entraîneurs et le nombre de points gagnés nous avons appliqué une Silhouette Analysis dont l'output se présentait de la façon suivante :

```
For n_clusters = 2 The average silhouette_score is : 0.3952986116103981
For n_clusters = 3 The average silhouette_score is : 0.32875873532423566
For n_clusters = 4 The average silhouette_score is : 0.3456859212217401
For n_clusters = 5 The average silhouette_score is : 0.33243699073622746
For n_clusters = 6 The average silhouette_score is : 0.33981905248360367
For n_clusters = 7 The average silhouette_score is : 0.31858661393828636
For n_clusters = 8 The average silhouette_score is : 0.28902136272297996
For n_clusters = 9 The average silhouette_score is : 0.29496358629557584
For n_clusters = 10 The average silhouette_score is : 0.296109404621815
```

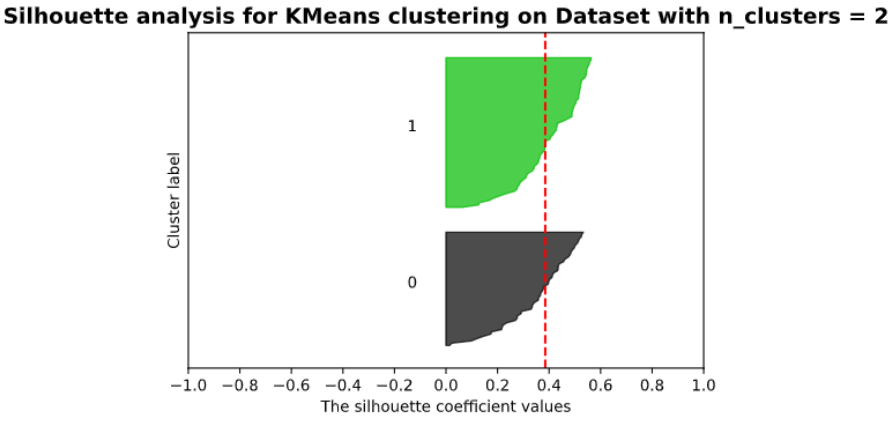



Figure 20: Exemple de Silhouette Analysis avec 2 clusters

# 6 Résultats des algorithmes

La métrique utilisée pour évaluer nos modèles est **l'accuracy.** Cette mesure indique le pourcentage de bonnes prédictions. C'est un bon indicateur qui est simple à comprendre. L'accuracy est calculée de la manière suivante :

$$Accuracy = \frac{Vrai\ positif + Vrai\ negatif}{Total}$$

## 6.1 Classification

Pour commencer cette partie d'application des modèles de classification, nous avons voulu comparer chaque modèle avec leurs paramètres par défaut pour pouvoir comparer leur score. Ainsi nous serons en mesure de pouvoir émettre une conjecture quant à l'efficacité des différents modèles à première vue. Les modèles donnant les meilleurs scores par défaut seront "tunés" dans la suite du rapport.

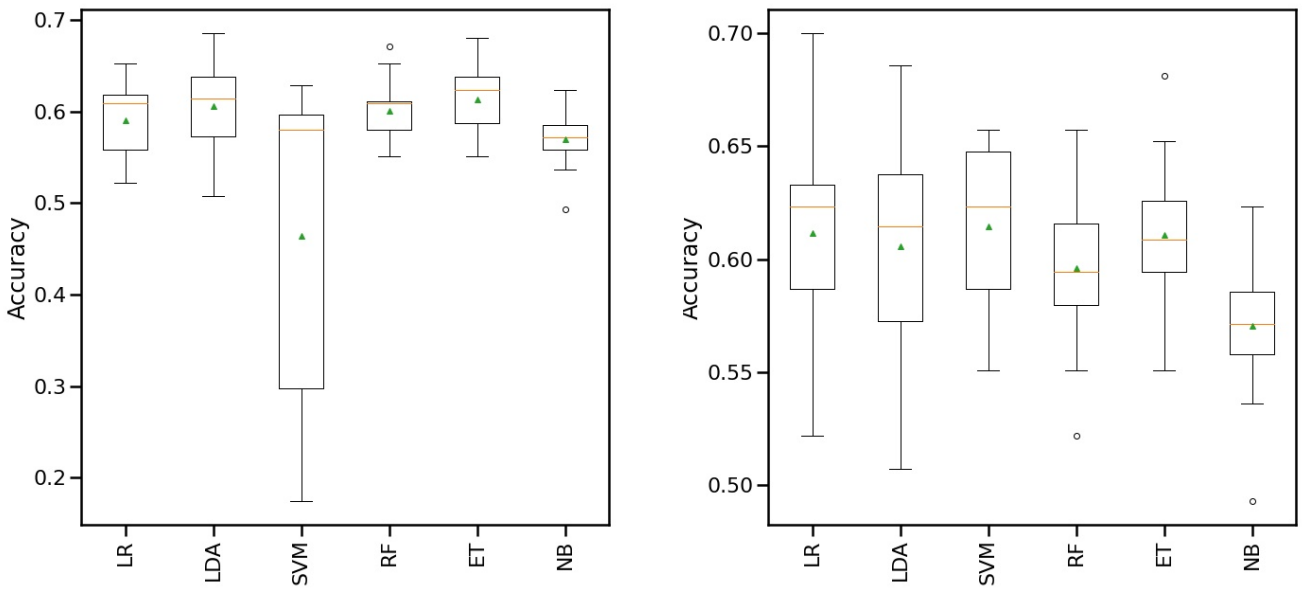



Figure 21: Comparaison des differents classificateurs avec le données non-normalisées (à gauche) les données normalisées (à droite)

Comme nous pouvons le constater ici sur la figure 21, les données normalisées permettent d'avoir une meilleure précision pour les différents modèles. Cela est dû au fait que les variables ne soient pas homogènes. En effet, même si toutes les variables sont numériques, certaines sont discrètes (comme le nombre de passes) et d'autres sont continues (comme la possession). C'est pourquoi nous allons utiliser les données normalisées pour entraîner chaque algorithme.

Pour pouvoir tester nos modèles, nous allons utiliser 30 % de nos données. Nous avons choisi ce pourcentage pour des raisons pratiques. En vue de la faible volumétrie de nos données (346 lignes), nous avons tenté de voir s'il y a une influence directe de la taille de l'échantillon d'entraînement sur la précision de nos modèles.

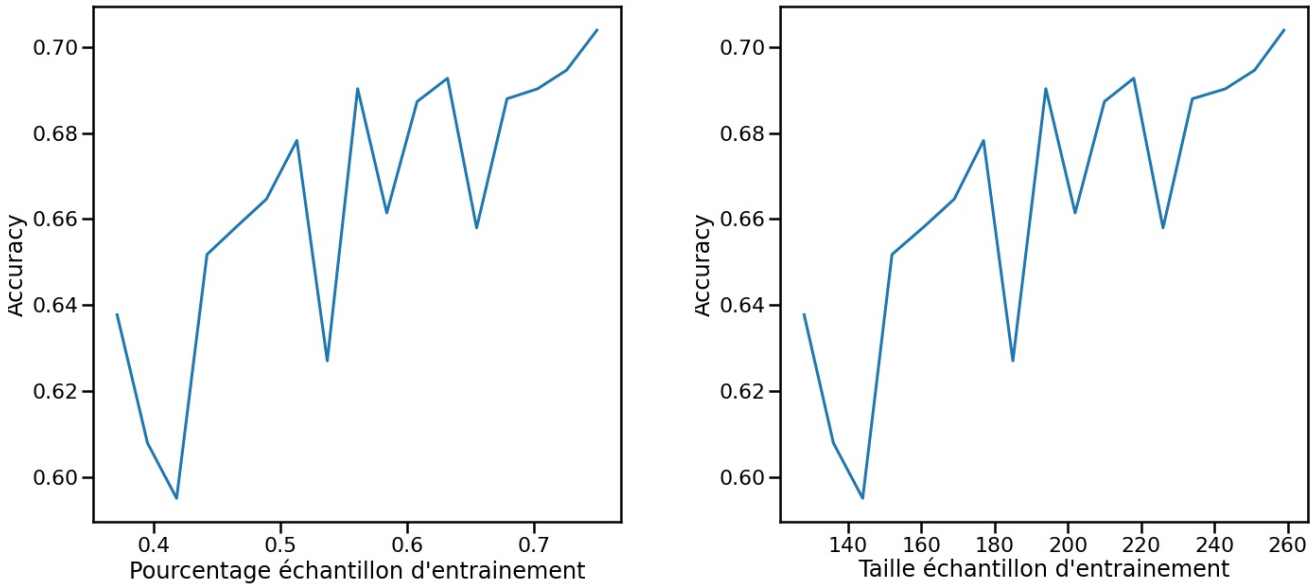



Figure 22: Accuracy en fonction de la taille d'échantillon

Nous pouvons voir sur la figure 26 qu'il n'y a pas une tendance visible et nous avons une courbe très saccadée. Nous avons donc du garder 100 lignes pour constituer notre base de test dans le but d'avoir des résultats parlant.

## 6.2 Modèles discriminant

Dans cette section nous allons nous intéresser à deux modèles lineaires : **Logistic Regressin** et **Support Vector Classifier**. Ces deux algorithmes ont pour objectifs de définir une zone par classe. Le but étant de trouver une dimension qui nous procure la meilleure séparation.

Nous avons appliqué ces modèles pour une classification multiclasses dans un premier temps. Pour éviter un sur apprentissage de la classe "Victoire", qui est la plus représentée, on utilise un régulateur L2. Cette pénalité permet de pondérer les poids affectés à chaque variable. Voici les résultats obtenus pour les deux modèles :

| | Precision | Recall | F1 Score |
|---|---|---|---|
| Défaite | 0.50 | 0.20 | 0.29 |
| Nul | 0.40 | 0.18 | 0.25 |
| Victoire | 0.63 | 0.93 | 0.75 |

| | Precision | Recall | F1 Score |
|---|---|---|---|
| Défaite | 0.53 | 0.36 | 0.43 |
| Nul | 0.25 | 0.18 | 0.21 |
| Victoire | 0.66 | 0.82 | 0.73 |

Résultat de **SVC (à gauche)** et **Logistique régression (à droite)**

Concernant les accuracies, les valeurs sont assez proches. Pour le SVM nous avons 60% de bonnes prédictions. La régression logistique est légèrement inférieure avec 58%. Cependant, pour une classification binaire la régression logistique est légèrement plus performante. Cela peut-être expliqué par la forme de la Sigmoid crée par l'algorithme.

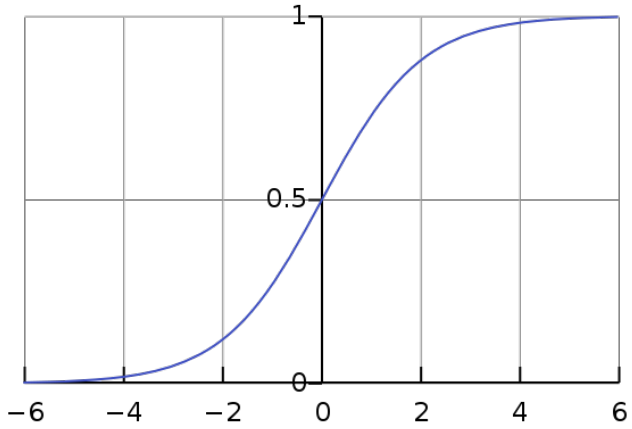



Figure 23: Fonction Sigmoïde

Voici les résultats de la classification binaire :

| | Precision | Recall | F1 Score |
|---|---|---|---|
| Nul ou Défaite | 0.55 | 0.69 | 0.61 |
| Victoire | 0.84 | 0.62 | 0.72 |

| | Precision | Recall | F1 Score |
|---|---|---|---|
| Nul ou Défaite | 0.55 | 0.69 | 0.61 |
| Victoire | 0.82 | 0.71 | 0.76 |

Résultat de **SVC (à gauche)** et **Logistique Regression (à droite)**

Pour les accuracys, nous avons un pourcentage de 70% pour la Logistique Regression contre 68% pour le SVC. Cela est confirmé par la figure du ROC de comparaison ci-dessous.

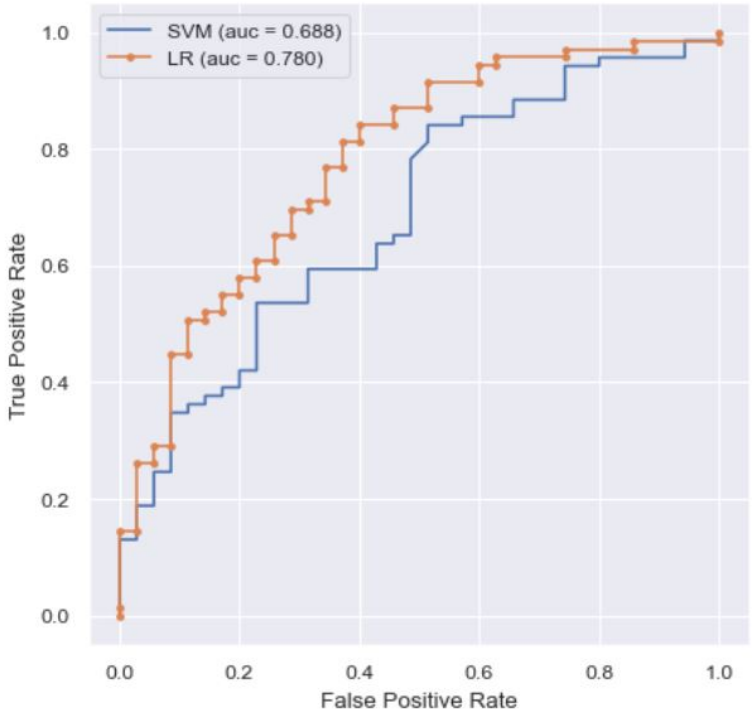



Figure 24: Courbes ROC de SVC et Logistique Regression

## 6.3 Modèles génératifs

Les modèles génératifs se basent sur les probabilités conditionnelles reposant sur le **théorème de Bayes**. Ce type de classification est intéressant pour notre étude car il s'adapte à la volumétrie, donc n'est pas affecté par les jeux de données de faible volumétrie. On calcule la probabilité d'appartenir à chaque classe en fonction de chaque variable. Pour ce faire, nous avons étudié la **Linear Discriminant Analysis** et l'algorithme des **Gaussian Naïve Bayes**. Le second algorithme suppose que toutes les variables sont indépendantes. Ces modèles nous donnent les résultats suivant :

| | Precision | Recall | F1 Score |
|---|---|---|---|
| Défaite | 0.68 | 0.62 | 0.65 |
| Nul | 0.44 | 0.13 | 0.21 |
| Victoire | 0.62 | 0.90 | 0.73 |

| | Precision | Recall | F1 Score |
|---|---|---|---|
| Défaite | 0.42 | 0.38 | 0.4 |
| Nul | 0.33 | 0.13 | 0.19 |
| Victoire | 0.56 | 0.79 | 0.66 |

Résultat de **LDA (à gauche)** et **Naïve Bayes (à droite)**

Ici on remarque que **la classe "Victoire" est la mieux prédite.** Cela est dû à la **répartition inégale des classes** dans nos données. C'est pourquoi nous nous sommes concentrés sur une classification binaire pour observer l'adaptation des modèles.

On crée alors une nouvelle variable cible qui a pour attribut : "Victoire" et "Nul ou Défaite". On observe alors les résultats suivants sur les deux modèles précédemment appliqués :

| | Precision | Recall | F1 Score |
|---|---|---|---|
| Nul ou Défaite | 0.64 | 0.50 | 0.56 |
| Victoire | 0.66 | 0.78 | 0.71 |

| | Precision | Recall | F1 Score |
|---|---|---|---|
| Nul ou Défaite | 0.71 | 0.43 | 0.5 |
| Victoire | 0.66 | 0.86 | 0.7 |

Résultat de **LDA (à gauche)** et **Naïve Bayes (à droite)**

La précision des classes est plus équilibrée avec **la classification binaire.** Nous obtenons pour la LDA une accuracy de 0.65 sur le test set et une accuracy de 0.52 pour le modèle Naïve bayes sur le test set. Ce qui prouve que l'impact de la volumétrie de chaque attribut impacte directement ces modèles probabilistes. On peut enfin confirmer cette hypothèse avec la courbe ROC :

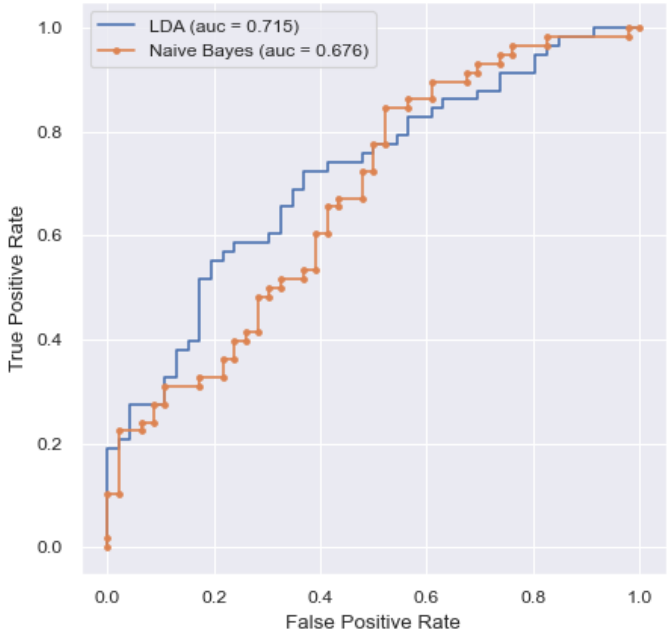



Figure 25: Courbe ROC entre LDA et Naïve Bayes

## 6.4 Modèles Ensemblistes

Dans cette section, nous allons nous focaliser sur les **Random Forests et les Extra Trees.** Ce sont des modèles robustes qui arrivent à palier le problème de variance des données des arbres de décisions classiques. Cela est rendu possible par le grand nombre d'arbres créé. Après avoir fait varier les hyperparamètres de ces deux algorithmes, nous avons constaté une similitude au niveau des différents paramètres.

Après tunning des modèles, nous avons choisi d'initialiser les deux modèles avec les mêmes valeurs pour les paramètres suivants :

- Nombre d'arbres : 94
- Profondeur maximale : 6
- Nombre minimum d'échantillons requis pour diviser un nœud interne : 3
- Nombre minimum d'échantillons requis pour une feuille : 3

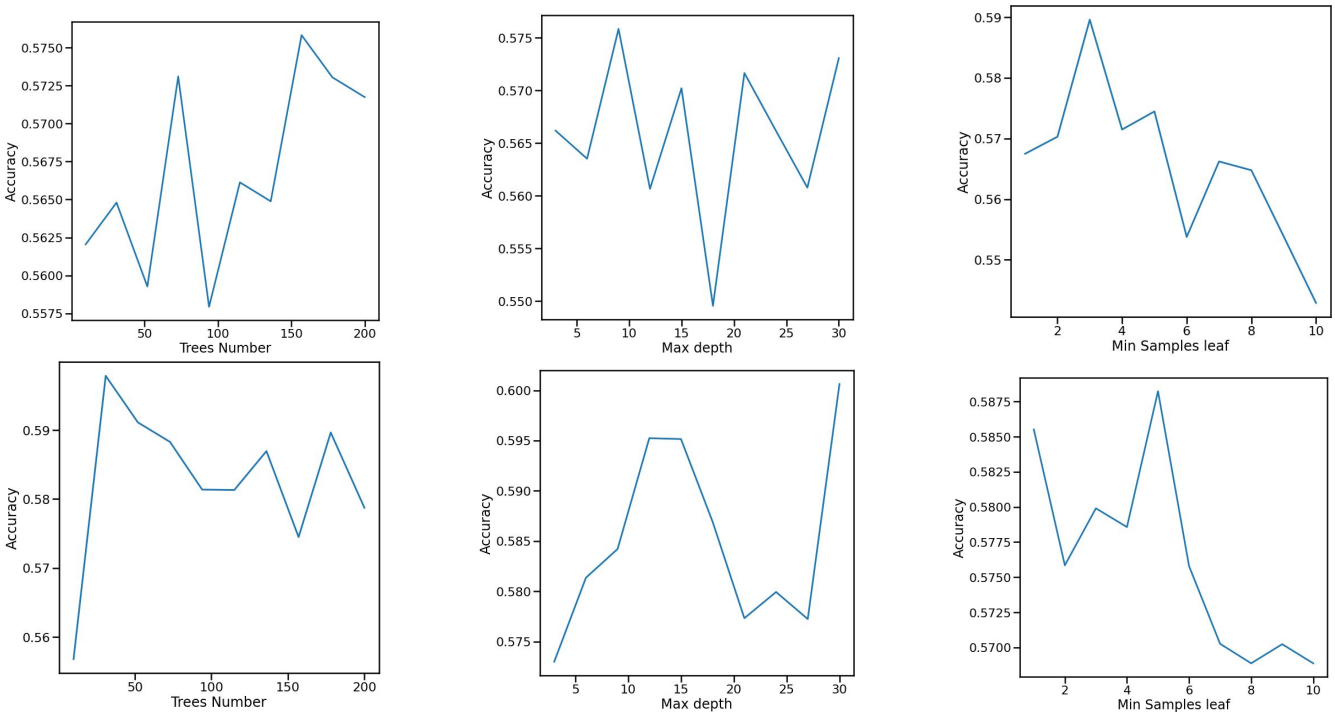



Figure 26: Variation des hyperparamètres pour le Random Forest (en haut) et l'Extra Trees (en bas)

On aurait pu penser que la méthode de **bootstraping** utilisée dans le Random Forest aiderait le modèle à mieux s'adapter que l'Extra Trees, mais les résultats de prédictions sont assez similaires. On peut expliquer cette situation par l'impact du choix aléatoire des variables de division pour l'Extra Trees. Cela se traduit aussi dans les résultats de prédictions des deux algorithmes:

| | Precision | Recall | F1 Score |
|---|---|---|---|
| Défaite | 0.30 | 0.16 | 0.21 |
| Nul | 0.25 | 0.05 | 0.08 |
| Victoire | 0.64 | 0.89 | 0.75 |

| | Precision | Recall | F1 Score |
|---|---|---|---|
| Défaite | 1.00 | 0.11 | 0.19 |
| Nul | 0.00 | 0.00 | 0.00 |
| Victoire | 0.64 | 1.00 | 0.78 |

Résultat de **Random Forest (à gauche)** et **Extra Trees (à droite)**

Pour le Random Forest tuné nous avons **une moyenne d'accuracy de 0.60** tandis que pour l'Extra Trees tuné nous avons **une moyenne d'accuracy de 0.64**.

## 6.5 K-means

### 6.5.1 K-means - Klopp - Points perdus

```
---------------------- Values for Cluster --------------- 1
matchday :  23.9412
Rest_Days :  8.1765
home_0_away_1 :  0.5882
Elo_Opponent :  1655.9412
Elo_Liverpool :  1878.0739
Elo_Diff :  222.1328
points_won :  0.6471
Shots :  15.9706
Shots On Target :  4.7353
Passes :  552.8529
Possession :  63.8529
34  Games included in this cluster
---------------------- Values for Cluster --------------- 2
matchday :  18.0667
Rest_Days :  6.2667
home_0_away_1 :  0.6
Elo_Opponent :  1823.5
Elo_Liverpool :  1831.8279
Elo_Diff :  8.3279
points_won :  0.6333
Shots :  13.9667
Shots On Target :  4.1333
Passes :  413.7667
Possession :  52.9
30  Games included in this cluster
```

Dans les saisons de Klopp, les matchs où Liverpool ont perdu des points sont souvent des matchs nuls comme le traduit la statistique de **pointswon (0.635)** et qui se regroupent en milieu de saison **(matchday = 23.94 pour le premier cluster et 18 pour le second).** Dans le premier cluster, Liverpool est favori et met la main sur le jeu et le ballon face à des adversaires supposés plus faibles d'après leur Elo. Les matchs sont principalement des matchs joués à l'extérieur.

Le second cluster regroupe les matchs face aux concurrents au titre de Liverpool. Le matchday regroupé au tour de la **journée 18** est un tournant dans les saisons en Premier League, à savoir la période du **Boxing Day,** (fête de Noël où le championnat anglais est le seul championnat européen à jouer). Liverpool, étant une grande équipe, est amenée chaque année à jouer des **grandes affiches** lors du Boxing Day. Cependant, l'existence de ce cluster témoigne de la **difficulté de Liverpool à négocier ce moment important** dans une saison. Le nombre de jours de repos n'a pas l'air d'influencer le résultat. On peut cependant émettre **l'hypothèse que les trop grosses coupures cassent le rythme de Liverpool** qui est habitué à enchaîner les matchs.

### 6.5.2 K-means - Rodgers - Points perdus

```
---------------------- Values for Cluster --------------- 1
matchday :  16.25
Rest_Days :  5.9643
home_0_away_1 :  0.6071
Elo_Opponent :  1850.1429
Elo_Liverpool :  1780.6647
Elo_Diff :  -69.4782
points_won :  0.5
Shots :  14.75
Shots On Target :  5.25
Passes :  350.3571
Possession :  50.5714
28  Games included in this cluster
---------------------- Values for Cluster --------------- 2
matchday :  17.6129
Rest_Days :  6.5484
home_0_away_1 :  0.5806
Elo_Opponent :  1634.129
Elo_Liverpool :  1793.618
Elo_Diff :  159.4889
points_won :  0.5161
Shots :  16.7419
Shots On Target :  5.5484
Passes :  433.0
Possession :  59.1613
31  Games included in this cluster
```

Quant à Brendan Rodgers, ces matchs regroupent à la fois des défaites et des matchs nuls répartis à l'équilibre selon la moyenne (0.5 pour les deux clusters) et sont majoritairement des matchs à **l'extérieur** témoignant de la difficulté de Liverpool en dehors de Anfield. Comme pour Klopp, les deux clusters séparent les observations en **deux catégories :** les matchs face aux **concurrents directs** et les matchs où **Liverpool sont favoris** et ont la possession du ballon.

La principale différence réside dans les matchs où Liverpool sont favoris. Là ou Klopp va multiplier nombre de passes (553), **Rodgers va moins faire tourner le ballon** (433 passes en moyenne). Malgré un bon taux de possession de balle compris entre 51 et 59%, les équipes de Rodgers ne parviennent pas à repartir avec les 3 points. On parle alors de possession "stérile".

Il est important de préciser que l'effectif mis à disposition influence fortement les statistiques en matière de passes de de possession. En effet, les joueurs de Liverpool à l'époque de Brendan Rodgers n'étaient pas réputés pour leur **finesse technique et leur capacité à maîtriser le ballon.** tandis que Jurgen Klopp accorde une grande importance dans ces caractéristiques lors du recrutement et du choix de ses joueurs.

### 6.5.3 K-means - Klopp - Points gagnés

```
 ---------------------- Values for Cluster --------------- 1
matchday :  15.25
Rest_Days :  7.3333
home_0_away_1 :  0.3333
Elo_Opponent :  1822.5833
Elo_Liverpool :  1785.2694
Elo_Diff :  -37.3139
points_won :  3.0
Shots :  15.0833
Shots On Target :  6.9167
Passes :  359.75
Possession :  51.4167
12  Games included in this cluster
---------------------- Values for Cluster --------------- 2
matchday :  21.0345
Rest_Days :  5.2069
home_0_away_1 :  0.6207
Elo_Opponent :  1660.7931
Elo_Liverpool :  1821.8199
Elo_Diff :  161.0268
points_won :  3.0
Shots :  17.2759
Shots On Target :  6.4828
Passes :  439.5862
Possession :  57.3448
29  Games included in this cluster
---------------------- Values for Cluster --------------- 3
matchday :  21.125
Rest_Days :  4.625
home_0_away_1 :  0.5
Elo_Opponent :  1659.25
Elo_Liverpool :  2009.5743
Elo_Diff :  350.3243
points_won :  3.0
Shots :  17.75
Shots On Target :  7.125
Passes :  803.5
Possession :  70.5
8  Games included in this cluster
---------------------- Values for Cluster --------------- 4
matchday :  20.5
Rest_Days :  5.8333
home_0_away_1 :  0.3333
Elo_Opponent :  1803.3333
Elo_Liverpool :  2047.479
```

```
Elo_Diff :  244.1457
points_won :  3.0
Shots :  15.0833
Shots On Target :  6.3333
Passes :  514.5
Possession :  56.0
12  Games included in this cluster
---------------------- Values for Cluster --------------- 5
matchday :  20.4737
Rest_Days :  5.5263
home_0_away_1 :  0.5263
Elo_Opponent :  1630.7368
Elo_Liverpool :  2030.6659
Elo_Diff :  399.9291
points_won :  3.0
Shots :  16.1579
Shots On Target :  6.0
Passes :  514.4737
Possession :  60.7895
19  Games included in this cluster
---------------------- Values for Cluster --------------- 6
matchday :  19.3871
Rest_Days :  5.7097
home_0_away_1 :  0.3548
Elo_Opponent :  1637.3226
Elo_Liverpool :  1904.3961
Elo_Diff :  267.0735
points_won :  3.0
Shots :  18.0
Shots On Target :  7.5484
Passes :  618.7097
Possession :  62.5161
31  Games included in this cluster
---------------------- Values for Cluster --------------- 7
matchday :  16.4286
Rest_Days :  7.8571
home_0_away_1 :  0.2857
Elo_Opponent :  1908.7143
Elo_Liverpool :  1965.2826
Elo_Diff :  56.5683
points_won :  3.0
Shots :  16.5714
Shots On Target :  7.0
Passes :  367.1429
Possession :  46.7143
7  Games included in this cluster

```

Les victoires de Klopp sont séparées en **7 clusters.** Le nombre de tirs varie très peu selon les clusters avec un nombre de tirs moyens qui tournent autour d'une **quinzaine de tirs.** Dans la continuité des observations réalisées sur les défaites, on remarque ici que la majorité des clusters des victoires ont une moyenne aux alentours de **5 jours de repos.** Contrairement aux 7 et 8 jours de repos des clusters des défaites, les victoires s'enchaînent lorsque les matchs s'enchaînent à un grand rythme sans un trop grand nombre de jours de repos.

On distingue des clusters regroupant des matchs à **domicile** : 1 4 6 7, regroupant 62 matchs témoignant de la force que **Anfield** procure à Liverpool. L'existence des clusters 2 3 5 prouvent de la capacité de Liverpool de tout de même **remporter des rencontres à l'extérieur**, capacité essentielle à toute équipe espérant remporter le titre.

Deux clusters regroupent des matchs où Liverpool rencontre ses adversaires directs. Les clusters 1 et 7 avec des EloDiff respectivement égaux à -37 et 56. On remarque un schéma tactique similaire entre ces deux clusters avec un nombre de passes assez faible (359 et 367) avec un bon ratio de tirs cadrés (46%). Dans ces deux clusters, les **pourcentages de possession sont beaucoup plus faibles** que dans les autres victoires (51% et 46%), cette statistique correspond à la philosophie de jeu souvent appliquée par Jurgen Klopp face aux grosses équipes, à savoir profiter de **la vitesse de ses ailiers en procédant en contre-attaques** et en laissant donc la maîtrise du ballon aux adversaires. Inversement, dans les matchs où Liverpool est **favori**, l'équipe prend en main le ballon et met en place ses attaques avec **beaucoup de passes et des taux de possession variant de 56% à 70% !**


Figure 27: Exemple de contre-attaque de Liverpool avec 2 ailiers face à un seul défenseur.

### 6.5.4 K-means - Rodgers - Points gagnés

```
---------------------- Values for Cluster --------------- 1
matchday :  19.3333
Rest_Days :  5.8431
home_0_away_1 :  0.4314
Elo_Opponent :  1629.4118
Elo_Liverpool :  1797.0416
Elo_Diff :  167.6298
points_won :  3.0
Shots :  18.5098
Shots On Target :  7.6863
Passes :  429.8431
Possession :  57.0196
51  Games included in this cluster
---------------------- Values for Cluster --------------- 2
matchday :  22.6667
Rest_Days :  6.25
home_0_away_1 :  0.3333
Elo_Opponent :  1852.0833
Elo_Liverpool :  1837.7571
Elo_Diff :  -14.3263
points_won :  3.0
Shots :  14.1667
Shots On Target :  6.75
Passes :  291.9167
Possession :  46.1667
12  Games included in this cluster
```

Les deux clusters représentent **majoritairement des matchs à domicile**, ce qui est cohérent avec l'observation sur les défaites qui étaient majoritairement des matchs à l'extérieur. Pour Brendan Rodgers, on retrouve donc 2 clusters dont un avec 51 matchs contre 12 pour le second. Dans le premier, Liverpool est favori et réussit donc à assumer son rôle avec un **haut taux de possession (57%), un bon nombre de passes (429) et un grand nombre de tirs (18.5).**

Face aux grandes équipes, et dans un statut de **non-favori**, le style de jeu est radicalement différent avec le choix (ou non) de laisser le ballon aux adversaires comme en témoigne le pourcentage de **possession** (46%) et le nombre de **passes** (291) qui diminuent fortement par rapport au premier cluster, et qui induisent un plus petit nombre d'occasions et donc de **tirs** (14).

# 7 Discussion

La discussion est l'une des dernières étapes de l'article scientifique, elle permet de donner des éléments de réponse à la problématique posée dans ce rapport. Cette section va nous permettre d'analyser et interpréter les résultats de nos travaux pour répondre à la question de recherche retraçant la nature de l'influence de Klopp sur le club de Liverpool. À travers l'application des différents modèles, nous avons pu valider plusieurs de nos hypothèses émises lors de l'exploration des données du dataset. Ainsi cela pourra potentiellement nous permettre de déduire ce qui a permis à Klopp de changer Liverpool.

Tout d'abord, l'apprentissage supervisé nous a permis de nous rendre compte de la difficulté d'interprétation des données footballistiques. En effet, les variables utilisées sont assez hétérogènes, tant sur les types que sur l'intervalle de valeurs. C'est pourquoi les modèles qui se basent sur la probabilité conditionnelle (théorème de Bayes) performent mieux. Ces modèles linéaires ont la capacité de s'adapter aux faibles volumes de données, car elles reposent sur une grande indépendance des variables traitées. Ce type d'algorithmes nous permet d'avoir des modèles qui ont un juste équilibre entre une haute variance et un bas biais. On peut alors avoir des meilleurs prédictions sur notre test set, et ce malgré un déséquilibre des classes. Cette vision va dans le sens des analyses qui peuvent être faites dans le domaine du football, puisqu'on essaie de trouver les éléments clés qui vont nous aider à estimer le résultat d'un match. Par exemple, on peut constater que **les poids associés à chaque variable** dans le modèle de Régression Logistique reflète bien les aspects importants pour analyser un match.

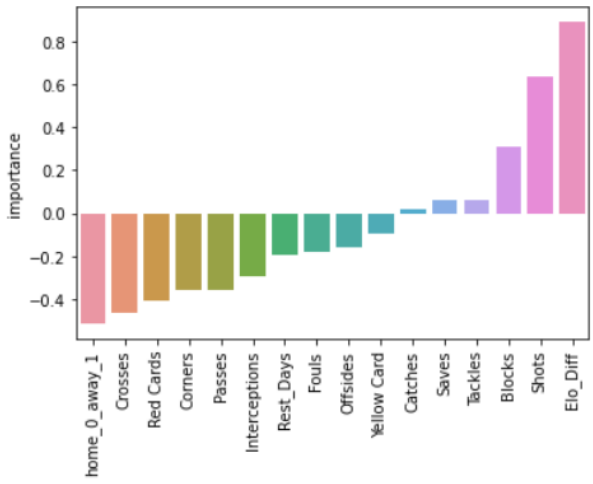



Figure 28: Importance des features dans la Régression Logistique

À partir de ces données récupérées, nous étudions le résultat du match ayant 3 classes qui sont : défaite, nul ou victoire. Parmi tous les modèles de classification implémentés, nous avons vu que c'est les modèles discriminants qui obtiennent la meilleure accuracy qui est égale à 70% sur le test set. Cette accuracy est satisfaisante étant donné que nous savons que le football ne se résume pas uniquement aux chiffres mais également à la performance et à la chance de chaque équipe. Effectivement, une équipe a beau dominer un match, elle pourra tout de même le perdre en fonction de la réussite de l'adversaire et des choix de l'arbitre.
Nous avons aussi constaté que les modèles ensemblistes ont été limité par la faible volumétrie de notre jeu de données. Malgré des accuracys assez correct ( 60% de bonnes prédictions), ces résultats restent biaisé par la proportion importante occupé par la classe "Victoire" qui est donc majoritaire dans les feuilles des arbres crées. La matrice de confusion de ci-dessous présente bien ce phénomène car nous avons ici une accuracy de 64% de bonnes prédictions. Cependant, 98% des prédiction ( 102/104 matchs) sont des "Victoires".

| | Victoire Prédite | Nul Prédit | Défaite Prédite |
|---|---|---|---|
| Vraie Victoire | 65 | 0 | 0 |
| Vrai Nul | 20 | 0 | 0 |
| Vraie Défaite | 17 | 0 | 2 |

Matrice de confusion du modèle des **Extra Trees**

D'un autre côté, par le biais de l'apprentissage non supervisé avec l'application des K-Means et la formation des différents clusters ont été très représentatifs de certaines caractéristiques de l'équipe soulevées dans l'exploration. Les différences dans le jeu amenées par Jürgen Klopp consistent à prôner un jeu plus basé sur la possession, un jeu offensif plus tranchant mais également un jeu plus épuré avec moins de fautes que son prédécesseur. Le volume de jeu de l'équipe de Klopp est également à souligner étant donné que l'abatage physique et athlétique sous Klopp requiert plus d'efforts notamment demandé par son contre pressing. Ces caractéristiques sont représentatives de la philosophie de jeu de Klopp, le "Gegenpressing", expliquant ainsi la révolution enclenchée à son arrivée en 2015.

Un aspect non visible à travers les résultats des K-Means est le laps de temps nécessaire à l'adoption des principes de jeu cités. En effet, les résultats de Jurgen Klopp lors de ses premières saisons à Liverpool n'étaient pas forcément meilleurs que ceux de Brendan Rodgers comme observé lors de l'exploration des données. Cela s'explique par le fait que Klopp a récupéré un effectif qu'il n'a pas construit et qui n'était pas adapté aux principes de jeu de Klopp. Il aurait pu être intéressant de regrouper les données par entraîneur et de les séparer en fonction des joueurs au sein de l'effectif puisque Brendan Rodgers a du s'adapter à son effectif contrairement à Jurgen Klopp qui a pu, avec l'aide et le soutien de ses dirigeants, façonner un effectif répondant à ses critères.

Nous avons également relevé l'impact de jouer à domicile. En effet, jouer dans son mythique stade d'Anfield a toujours été une grande force pour les joueurs de Liverpool que ce soit avec Rodgers ou avec Klopp. Klopp a cependant plus su profiter d'une des plus grandes armes qu'un club puisse avoir, à savoir la ferveur et la passion d'un public avec qui il a tout de suite noué une grande proximité. On le voit d'autant plus aujourd'hui, en période de pandémie sans public, les performances de Liverpool sont clairement moins tranchantes que pendant les années précédentes.

En somme, Klopp a apporté divers changements au club de Liverpool parmi lesquels: une nouvelle philosophie de jeu, des nouveaux standards pour les joueurs en termes de qualités techniques mais aussi de valeurs, une régularité dans les résultats et surtout un nouvel état d'esprit. Il a su nouer une grande proximité avec les supporters et a toujours pu compter sur la confiance presque aveugle de ses dirigeants pour construire son équipe là où Rodgers était plus remis en doute. Klopp a également su puiser dans les bases posées par Rodgers, à savoir l'importance accordée à la maîtrise du ballon et l'intensité dans le jeu. On retrouve des similarités chez les deux entraîneurs à travers la volonté de développer les joueurs à sa disposition et la rigueur demandée aux joueurs.

# 8 Conclusion

Finalement, ce projet s'est révélé très enrichissant dans la mesure où il a consisté en une approche concrète du métier d'ingénieur dans le monde des Sports Analytics.

Ce projet nous a permis d'appliquer et d'approfondir des notions étudiées en cours mais aussi d'apprendre de nouvelles notions en terme de collecte, traitement et d'interprétation de données. Les principaux problèmes, que nous avons rencontrés, concernaient le volume et l'homogénéité des données récupérées, parfois handicapant dans l'application des modèles d'apprentissage automatique. Ainsi, nous avons pu toucher du doigt la difficulté et la complexité du fonctionnement des algorithmes de Machine Learning appliqués à des données limitées en nombre.

Ce projet est avant tout un premier pas dans le monde de la Data dans le sport. À travers nos recherches et nos travaux, nous nous sommes rendus compte de l'étendue des domaines d'application de la Data dans le monde du sport comme les paris sportifs, l'analyse sportive des matchs ou encore l'analyse financière des résultats sportifs.

Ces domaines pourraient, par la suite, faire l'objet d'une étude dans la continuité de notre projet. On pensera par exemple à :

- Développement d'un modèle de prédiction des résultats des matchs.
- Analyses macro-économique et micro-économique d'une équipe.
- Développement d'un algorithme de reconnaissance vidéo destiné au visionnage automatique.

# 9 Bibliographie

## Articles

[2] Brandon Dominique. "Machine Learning Analysis : Why have Real Madrid been so poor in La Liga in the last decade ?" In: (). URL: https://docs.google.com/document/d/1n_wVq_SgYqU_9QUemZGjnySbKROs29gx8iH9gaP2ydU/edit.

[3] Cahiers du football. "Les « Expected Goals », au coeur de la révolution statistique". In: (). URL: http://www.cahiersdufootball.net/article-les-expected-goals-au-coeur-de-la-revolution-statistique-5744.

[5] Harleen Kaur Rahul Baboota. "Predictive analysis and modelling football results using Machine Learning approach for English Premier League". In: (). URL: https://www.researchgate.net/profile/Harleen-Kaur/publication/324072605_Predictive_analysis_and_modelling_football_results_using_machine_learning_approach_for_English_Premier_League/links/5cf7e9b0a6fdcc8475088f4b/Predictive-analysis-and-modelling-football-results-using-machine-learning-approach-for-English-Premier-League.pdf.

## Films et Documentaires

[1] Bennett Miller (Director). *Moneyball*. Columbia Pictures, 2011.

[4] FourFourTwo. *The Numbers Game | How Data is changing Footbal ?* URL: https://www.youtube.com/watch?v=lLcXH_4rwr4&ab_channel=FourFourTwo.