



UNIVERSITÉ LIBRE DE BRUXELLES

ULB

Faculté de Lettres, Traduction et Communication

La visualisation et la fouille des données dans le domaine du sport appliqué à un programme de prédiction de résultats

VACHÉ Edouard

Mémoire présenté sous la direction de Sébastien DE VALERIOLA du directeur, du co-directeur éventuel, en vue de l'obtention du diplôme de Master en Sciences et Technologies de l'Information et de la Communication.

Année académique 2020-2021

CONSULTATION DU MEMOIRE / TRAVAIL DE FIN D'ETUDES

EXEMPLAIRE DESTINÉ A LA BIBLIOTHEQUE

à insérer dans le Mémoire ou Travail de fin d'études, après la première page de couverture (page de garde)

Je soussigné

NOM (en majuscule): VACHÉ.....

PRENOM: Edoard.....

TITRE du travail : La visualisation et la fouille des données dans
le domaine du sport appliqué à un programme de
prédiction de résultats.....

AUTORISE *

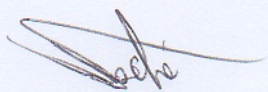
~~**REFUSE ***~~

La consultation du présent mémoire/travail de fin d'études par les utilisateurs des bibliothèques de l'Université Libre de Bruxelles.

Si la consultation est autorisée, le soussigné concède par la présente à l'Université libre de Bruxelles, pour toute la durée légale de protection de l'œuvre, une licence gratuite et non exclusive de reproduction et de communication au public de son œuvre précisée ci-dessus, sur supports graphiques ou électroniques, afin d'en permettre la consultation par les utilisateurs des bibliothèques de l'ULB et d'autres institutions dans les limites du prêt inter-bibliothèques.

Fait en deux exemplaires, Bruxelles, le 28.1.01/2021.....

Signature



* Biffer la mention inutile

1. Résumé

Ce mémoire a pour ambition de comprendre comment les méthodes de visualisation et de fouille des données, au travers de différents modèles de Machine Learning, peuvent influencer le travail d'entraîneur d'une équipe sportive. Pour illustrer l'étude la prédiction a porté sur une saison de NBA, en manipulant différents modèles. Le travail met en évidence les critères à la fois théoriques et techniques de la réalisation de ce projet. Un des points clés étant l'identification et la collecte des données exploitées, il est donc important de ne pas se précipiter durant cette phase puisque l'entièreté du projet repose sur elles. Ainsi un intérêt trop prononcé pour l'obtention de bonnes performances de prédiction peu conduire à un faible taux de résultats exploitables et éloigner le projet de l'objectif initial du travail. Un recentrage du projet de prédiction a permis de démontrer une tendance de coaching déjà observable au sein de la NBA. À la lumière des résultats obtenus, il apparaît que la fouille et la visualisation de données ont déjà eu un impact sur le jeu de la NBA et plus généralement sur tout le paysage sportif que l'on connaît aujourd'hui.

Table des matières

1	Résumé	3
2	Introduction	6
3	Motivation	15
3.1	Dessein du sujet	15
3.2	Projet de recherche	17
3.3	Position adoptée	19
3.4	Méthodologie	20
4	Données	23
4.1	Typologies des données	23
4.2	Récupération des données	32
4.3	Qualité des données	35
5	Modélisation	40
5.1	Machine Learning ou Deep Learning	40
5.2	Machine learning dans le sport	47
5.3	Modèles appliqués	52
5.3.1	k-Nearest Neighbors	55
5.3.2	Support Vector Machine	56
5.3.3	Logistic Regression	56
5.3.4	Naives Bayes	57
5.3.5	Decision Tree	57
5.3.6	Random Forest	58
5.3.7	XGBoost	58

6 Résultats	60
6.1 Résumé des résultats	60
6.2 Visualisation des résultats	65
6.3 Interprétation des résultats	79
7 Conclusion	88
8 Annexes	94
9 Glossaire	103
Bibliographie	104

2. Introduction

La révolution, l'explosion des données a transformé notre manière de voir et d'analyser notre monde. Le sport n'a pas échappé à la tendance et a été directement impacté par ce changement. Ce domaine a toujours été de près ou de loin une affaire de chiffres et de récolte statistique.

"Nous sommes actuellement à « l'Ère des mégadonnées » : la prolifération des technologies de repérages combinées avec le désir d'enregistrer et de contrôler l'activité humaine a amplifié considérablement le volume et la variété des données en circulation ainsi que la vitesse à laquelle les données se déplacent"[9, p.140].

Le Big Data n'est pas le premier mouvement statistique connu de l'histoire humaine. L'émergence des nouvelles technologies et surtout la "circulation" d'un nombre gigantesque d'informations numériques font que certains auteurs qualifient le Big Data comme un des mouvements statistiques les plus importants connus à ce jour.

Dans le monde du sport, cela a pris du temps pour arriver au niveau où nous sommes aujourd'hui. Comme bien souvent tout commença aux États-Unis. Dans le monde du Baseball, on évoque l'apparition de la "sabermétrie" ou "sabermetrics" qui est une approche statistique du Baseball et du Box Score. Les Box Score sont des dérivés des feuilles de statistiques. Terme qui de nos jours se retrouve dans la quasi totalité des sports. Celles-ci sont utilisées pour aider à déterminer la relation entre différents éléments et, dans le sport par exemple, certains pourcentages aident souvent à déterminer le succès ou non d'une équipe. Ces informations sont ensuite corrélées à un joueur, ou à une équipe puis elles sont lues, analysées pour donner une idée générale de la façon dont le jeu s'est déroulé ou de la façon dont le joueur s'est comporté pendant le match, une saison ou sa carrière.

L'apparition du Box Score au XIXème siècle est antérieur à celle des ordinateurs. Il était donc question de remplir les Box Score à la main. Malheureusement, à l'époque, il était fréquent d'avoir un manque d'information.

Le défi était donc de trouver un moyen de récupérer toutes les données désirées, voulues, sans aucune perte. Sans l'aide des technologies que nous connaissons aujourd'hui, une communauté de statisticiens du Baseball de l'époque, déploya des centaines de personnes, pour se rendre sur les terrains à travers les États-Unis dans le but de collecter les données nécessaires à la nouvelle étude scientifique de ce sport.

"Le jeu était en effet témoin de sa propre «avalanche» de numéros imprimés, une tendance qui à son tour a engendré un éventail vertigineux de nouvelles statistiques de baseball." [9, p.148].

Avant même la "révolution des données" engendrée par l'explosion des nouvelles technologies à la fin du XXème siècle, les statisticiens ont été confrontés à une problématique récurrente que connaît notre époque, à savoir comment stocker, gérer et utiliser toutes les données collectées.

Au milieu du XXème siècle, cette "avalanche" de données a eu un impact sur le monde du Baseball. Quelques équipes de la ligue national de Baseball américaine se sont attachées les services de statisticiens de l'époque pour évaluer les jeunes joueurs afin de repérer les potentielles futures stars de la ligue et créer la meilleure équipe possible. Certains statisticiens se sont focalisés, non plus uniquement sur les joueurs, mais plus particulièrement sur le jeu.

Les analyses menées par les différents statisticiens ont pu démontrer que certains faits de jeu étaient plus enclin à produire des résultats positifs que d'autres dans certaines situations précises. Dave Cameron, un écrivain et analyste de Baseball américain utilise l'exemple de l'amorti pour appuyer la phrase suivante : "Ce n'est pas toujours le mauvais mouvement, mais il est utilisé beaucoup trop souvent et dans de trop nombreuses situations où taper loin (swing) est plus susceptible de produire un résultat positif."

La dite phrase exprime bien le changement qui s'est enclenché dans le monde

du sport suite à l'ampleur qu'ont prise les statistiques. "«Moneyball» est en fait la poursuite de connaissances objectives combinées à la quête d'exploitation des inefficacités du marché." [9, p.141]. *Moneyball* est un terme provenant de l'idée que la connaissance collective des professionnels du Baseball du siècle dernier est subjective et faillible, les statistiques sont donc une des solutions pour pallier ces failles.

Dans leur article "*The Datafication of Everything' : Toward a Sociology of Sport and Big Data*", Brad et Rob Millington ont mis en avant quatre postulats concernant le sport dans l'ère du Big Data dans un but de présenter ce qui semble être des principes de base qui souligne le nouveau virage statistique du sport.

Le premier postulat, "*That Sport's Statistical Turn Exists in Reciprocity With the Wider Big Data Movement*", exprime la réciprocité qu'il y a entre les statistiques dans le sport et le mouvement du Big Data. Les auteurs considèrent le sport comme un réseau, une structure ouverte et dynamique. L'idée majeure est de montrer que cela ne s'arrête pas au simple monde du sport, il permet à travers lui de toucher d'autres domaines [9, p.150].

Le deuxième postulat qui a pour intitulé "*The Growing Presence of Advanced Analytics in High Level Sport Is Widely Deemed a Progressive Trend*", se concentre sur la tendance progressive que représente l'analyse de données dans le sport et plus particulièrement le sport de haut niveau. Lorsque Brad et Rob Millington parlent de progrès c'est au sens économique du terme, puisque l'analyse de données va permettre dans un premier temps, de prendre en compte la notion de hasard dans le sport et dans un second temps, d'éviter aux équipes, directeurs sportifs, managers, d'effectuer des choix coûteux et non rentables [9, p.151].

Dans le troisième postulat, "*The Big Data Is Increasingly Impactful Across the Sporting Landscape*", il est question de l'impact du Big Data dans le paysage sportif. Le Big Data dans le sport n'est pas qu'une question de performance, il influe par exemple sur le domaine médical du sport en prévenant d'éventuelles blessures. Les sportifs amateurs ont eux aussi été touchés au travers d'application de coaching que l'on rencontre souvent dans le fitness ou la course à pied et bien évidemment les fans par le biais des statistiques ont une nouvelle manière d'évaluer, de discuter de leurs

équipes ou joueurs favoris[9, p.153].

Enfin le quatrième et dernier postulat, "*The Big Data Has Its Discontents*", porte sur les inconvénients liés au Big Data. Un problème récurrent en lien direct avec l'utilisation de données, est la question de la privatisation des données. Avons-nous oui ou non le droit d'exploiter les données des sportifs, des équipes ou même des fans sans leur autorisation. Un autre problème soulevé concerne les intentions, qui pousse à lier le Big Data au sport, et bien souvent ce n'est pas la passion du sport mais le capitalisme et le libéralisme qui en sont la cause, qui pour beaucoup sont la source de dégradation du sport en général[9, p.155].

Ces postulats résument bien la situation actuelle du monde du sport. Le sport est devenu un domaine hyper connecté dès suite de l'arrivée des nouvelles technologies. Tout y est connecté. Et cela touche tant les joueurs, les coachs, les équipes, les supporters et les infrastructures.

Les deux exigences auxquels les nouvelles technologies permettent de répondre sont tout d'abord, la récupération des données des sportifs et ensuite, l'amélioration constante des performances. Lorsque le mot "sportif" est utilisé, il représente les athlètes professionnels et non les amateurs. L'accès à ces technologies de pointe très onéreuses n'est pas pour le moment à la portée du monde du sport amateur.

Afin d'avoir une vue d'ensemble de ce que peuvent être les nouvelles technologies dans le domaine sportif, commençons large en évoquant les équipements des stades et autres lieux sportifs pour finir par l'athlète lui même.

Quand on observe correctement ce qui nous entoure dans un stade, la première chose qui peut sauter aux yeux est le nombre de caméras. Dans un premier temps les caméras avaient pour seul et unique objectif de filmer le jeu de la meilleure des manières afin de pouvoir donner aux téléspectateurs la même impression que s'ils étaient dans les tribunes du stade. Après les matchs, il était question pour les analystes de visionner chaque action afin de décortiquer chaque partie du jeu, de récolter les données pour chaque équipe et après analyse de créer des nouvelles

stratégies de jeu. Un travail fastidieux.

Les caméras ont elles aussi évolué. Leur fonction n'est plus seulement de filmer, elles offrent aussi la capacité de récupérer les données en temps réel et de les transmettre aux analystes.

De nos jours, on peut compter en moyenne plus d'une dizaine de caméras autour des terrains. Il existe deux types de caméra désormais. Les premières sont dédiées exclusivement à la retransmission du jeu et les secondes sont chargées de collecter, d'analyser et d'afficher en temps réel les effets de diffusion télévisée, les applications grand public, l'analyse des performances ou d'autres formes encore d'évaluation.

Les joueurs sont eux aussi entourés et porteurs de capteurs, que ce soit durant les entraînements ou bien pendant les rencontres officielles. On parle de "petite" technologie de part leur taille puisque dans la majorité il s'agit de capteur que les sportifs vont avoir accroché à leurs vêtements ou à même la peau.

"Ces technologies sont relativement nouvelles dans le monde du sport. Elles sont aujourd'hui devenues un outil privilégié pour améliorer les performances des athlètes et réduire les blessures"[21, p.4].

Un des premiers sports à avoir utilisé des capteurs fixés au maillot des joueurs est le rugby. Lorsqu'on regarde un match, il est possible de voir au niveau du col, entre les omoplates, de chaque joueur ce fameux capteur qui permet tout d'abord d'évaluer les performances et la forme physique de chaque joueur en temps réel pour ensuite les transmettre directement aux différents entraîneurs, analystes afin de prendre les décisions qu'il convient pour mener à bien l'équipe vers la victoire tout en préservant la santé des joueurs.

Les équipements sportifs ne sont pas épargnés par l'arrivée des capteurs. Dans le monde du basket, toujours dans un objectif d'améliorer les performances et plus particulièrement dans ce cas précis le tir des joueurs, il existe un ballon de basket intégrant différents capteurs, logés à l'intérieur du ballon dans un petit tube. Ces capteurs donnent un tas d'information sur le tir mais aussi le nombre de rebonds, la vitesse des dribbles, la vitesse de rotation du ballon, les angles de tirs et même

l'impact des mains sur le ballon.

L'idée derrière tout cela, est de collecter un maximum de données même non structurées grâce à toutes ces caméras, ces différents capteurs et bien d'autres, même si toutes les utilisations potentielles de ces données ne sont pas encore connues.

Malheureusement, cela inclue également la possibilité que des données importantes soient perdues dans cet océan d'informations et génère les "*dark data*" qui représentent une quantité de données non négligeable que beaucoup n'arrivent pas à exploiter.

À l'ère d'une société où le capitalisme est roi et où il existe un besoin permanent, de performance, de progrès exacerbé, le sport n'a pas échappé à cette tendance. Les athlètes et les équipes veulent gagner toujours plus et battent toujours plus de records en repoussant les limites du corps humain. Les nouvelles technologies ont renforcées cette direction en apportant des outils modifiant l'appréhension du sport dans son ensemble.

Le rôle des technologies a bien changé au fur et à mesure de leur implantation dans le domaine du sport jusqu'à en devenir incontournable. Leur apport a eu des effets très positifs dans l'amélioration de performances des athlètes. Aujourd'hui impossible de s'entraîner sans que le moindre geste soit suivi par un capteur. Une toute nouvelle approche de la préparation sportive, du coaching sportif a fait alors son apparition.

Le sport demande l'apprentissage de nouveaux gestes en permanence, la répétition de ces mouvements est la meilleure manière de les maîtriser. Encore faut-il exécuter le bon mouvement. Le feedback est par conséquent indispensable et c'est à ce moment que vont intervenir les nouvelles technologies.

Deux types de feedback existent. Tout d'abord, le feedback dit naturel, humain qui n'est autre que l'ensemble des sensations ressenties durant l'exécution du mouvement. Ensuite, on parlera de feedback externe, traditionnellement donné par les entraîneurs et récemment par les équipements et appareils techniques. Nous allons

nous intéresser à l'évolution du second feedback.

Les entraîneurs sont là pour suivre, surveiller les actions des athlètes afin de corriger les imperfections, les failles observées pour obtenir de meilleurs résultats. Dans ce cas présent de coaching, les équipements technologiques servent uniquement pour la surveillance des performances de l'athlètes tandis que le feedback sera la responsabilité de l'entraîneur par le biais de conseils, de schémas ou de démonstrations.

"La méthode traditionnelle d'apprentissage moteur assisté par l'entraîneur peut être améliorée en introduisant un équipement technique capable de mesurer, calculer et présenter les propriétés de l'action exécutée." [33, p.4]. Les équipements sont le prolongement de l'entraîneur, ils permettront de récupérer des données hors de portée des capacités humaines de détection.

Le feedback donné de manière dite classique par un entraîneur, est défini comme un feedback terminal puisqu'il est partagé dans un délai plus ou moins long après que l'action, le mouvement, le geste est effectué.

Une étude sur un groupe de trois nageurs a été menée avec pour objectif d'analyser l'activité produite par ces mêmes nageurs dans différentes situations. Plusieurs équipements, technologiques ont été disposées autour du bassin, et plus particulièrement face au plot de départ dans l'alignement du couloir. "Ces données d'enregistrement ont constitué des traces audio-visuelles permettant de documenter les caractéristiques spatio-temporelles des mouvements propulsifs du nageur." [35, p.265].

Il s'agit bien ici d'un feedback classique. Une fois les nageurs hors de l'eau, après avoir réalisé les exercices demandés, l'entraîneur a pu utiliser les informations, les images prises par les équipements autour du bassin pour effectuer son retour sur la prestation de chaque nageur.

Le second feedback évoqué précédemment, considéré comme un "feedback loop" fait concurrence au feedback dit traditionnel. "Un feedback concurrent, qui est donné en temps réel pendant le déroulement de l'action, a été trouvé utile pour l'apprentissage et les entraînements dans le monde professionnel du sport." [33, p.4].

Le coach ne fait pas parti de ce feedback. Tout l'intérêt de cette nouvelle tech-

nique est de pouvoir prévenir l'athlète pendant son effort lorsqu'il effectue un geste, mouvement bon ou non pour qu'il puisse le rectifier par lui-même lors du geste suivant. "De telles solutions sont plus communément décrites comme des "biofeedback" système." [33, p.4]. En d'autres termes, des capteurs positionnés sur l'athlète sur les parties du corps à étudier vont transmettre des signaux directement à un appareil pour les traiter et renvoyer les résultats d'analyse directement à l'athlète via les différents sens du corps humain.

Revenons à l'étude faite sur les nageurs et leur activité dans l'eau. Elle se concentre aussi sur un "biofeedback" en plus du feedback traditionnel évoqué plus tôt. Il "consiste à informer le nageur en temps réel sur son efficacité gestuelle par l'intermédiaire de capteurs de pressions ou de jauges de contraintes fixés sur des plaquettes attachées aux mains du nageur." [35, p.259]

La finalité de cette étude est dans un premier temps, sportive, puisqu'elle vise à améliorer la qualité de nage de l'athlète en lui donnant en temps réel des indications permettant d'améliorer sa gestuelle au fur et à mesure qu'il nage. Dans un second temps, plus comportementale, on évalue l'attitude de l'athlète une fois que l'information est transmise, de quelle manière il va s'adapter au feedback reçu.

Les nouvelles technologies ont changé le paysage sportif. Elles ont répondu et répondent encore parfaitement au besoin de tout mesurer, quantifier. Malgré cela, la tendance laisse place aux techniques de *Machine Learning* ou de *Deep Learning*. La prédiction alliée aux nouvelles technologies est la nouvelle norme.

Le désir de performance et de gagner à tout prix est de plus en plus présent et l'informatique combinée aux mathématiques sont la clé pour aller toujours plus loin. Il est question de ne plus rien laisser au hasard, de pouvoir tout maîtriser pour être le meilleur, le plus efficace possible. Les techniques de Machine Learning et de Deep Learning permettent en partie cela.

Aujourd'hui, la priorité n'est plus la récupération des données mais la manière dont elles vont être utilisées, traitées, analysées. Les méthodes de prédiction touchent

tous les aspects du sport, tout en ayant le même objectif, d'essayer de contrôler les aspects imprévisibles du sport. On peut retrouver par exemple ces méthodes dans la prédiction de résultats ou encore en médecine du sport pour prévenir d'éventuelles blessures.

3. Motivation

3.1 Dessein du sujet

Du plus profond dont je me souviene le monde du sport en général m'a toujours attiré. Encore aujourd'hui, je passe mon temps à perfectionner mes connaissances dans ce domaine qui n'en finit pas d'évoluer.

Les sujets qui m'intéressent le plus sont, tout d'abord, comme tout bon supporter et fan qui se respecte, les résultats des différentes équipes et championnats.

Ensuite, il y a les exploits sportifs qu'ils soient collectifs ou bien individuels ils me procurent de l'admiration, de la fascination pour ce qu'accomplissent certains sportifs que jadis, je rêvais d'imiter.

Enfin, le sujet qui requiert toute mon attention, est l'évolution du jeu. Comprendre les différentes tactiques, stratégies mises en places, la préparation des équipes et des athlètes pour atteindre le plus haut niveau et y évoluer.

Un de mes objectifs de vie est de pouvoir un jour travailler dans le sport. N'étant qu'un bon sportif amateur, une carrière professionnelle est oubliée depuis bien longtemps. Cependant, ma passion me permet d'utiliser, mes connaissances en sport et de les combiner par l'intermédiaire de mes compétences en informatique acquises durant mon premier cursus d'études.

Le rôle d'analyste de données est ce vers quoi je souhaite me diriger pour mon avenir professionnel. Il mélange plusieurs aspects du monde de l'informatique qui m'ont décidées de suivre cette voie. Cela demande de la manipulation, de la gestion de données, un esprit critique, d'observation, de recherche afin de répondre à des besoins, et de trouver des solutions qui faciliteront par la suite le travail d'autrui.

Lorsque le temps de choisir un sujet de mémoire est venu, je n'avais aucun intitulé en tête. En revanche, j'étais déjà sûr de moi en ce qui concerne les domaines que je voulais traiter dans mon travail.

Ma question première, était de savoir comment lier le domaine du sport et celui de l'informatique afin que le sujet puisse être accepté. Après avoir feuilleté d'anciens mémoires des années précédentes, je me suis rendu compte de la diversité des thèmes abordés. Il ne manquait plus qu'à trouver la problématique permettant d'intégrer ces deux disciplines (sport et informatique) tout en répondant aux exigences qui m'étaient demandées.

Il était inconcevable pour moi d'effectuer mon mémoire sur un tout autre domaine. Avant même d'intégrer ce cursus je savais que j'effectuerai un mémoire qui aurait un sujet ayant rapport avec le sport.

Être capable d'étudier de près ou de loin des sportifs de haut niveau, avoir la possibilité d'aider, d'influer au travers de données sur les performances, les stratégies est quelque chose qui me fait vibrer. Il n'était donc pas forcément question de se focaliser sur la question de recherche mais savoir comment créer un sujet de recherche qui intégrerait l'analyse de données appliquée au domaine du sport.

De part mes compétences en informatique, la réalisation d'un projet concret d'analyse de données est venu tout naturellement. Ce projet, est une motivation supplémentaire et me conforte dans l'idée de conjuguer les domaines du sport et de l'informatique. Il s'agit à présent de se concentrer sur l'objet de ce projet qui ensuite permettrait d'en faire la base de ce mémoire.

La finalité de ce mémoire ne s'arrête pas simplement à la validation de mon diplôme. Il me permettra d'appuyer mes futures candidatures d'emplois et mettra ma motivation ainsi que mes connaissances en avant pour atteindre ce à quoi j'aspire pour mon futur professionnel.

3.2 Projet de recherche

La réflexion a commencé par le choix du sport sur lequel le projet est s'applique. Le Basket-Ball étant mon sport de prédilection, il a été évident de me tourner vers celui-ci.

Souhaitant faire de l'analyse de données et avec la période assez courte du mémoire, il a fallu trouver un projet qui soit réalisable en quelques mois et qui me permette d'acquérir de nouvelles compétences, tout en ayant des résultats suffisamment probant pour que le projet puisse servir de support quant à la réalisation du mémoire.

Comme évoqué dans l'introduction, tout le travail fourni dans le projet à été similaire à ce que vous avez pu lire concernant le Baseball et les box-scores. Évidemment le contexte actuel qui existe dans le monde du sport a fait que l'étude menée a permis de répondre de manière différente avec des outils plus développés aux problématiques qui ont été mises en lumière dans l'introduction.

Impossible donc d'effectuer ce projet sans un nombre acceptable de données. Contrairement aux premiers statisticiens, collectant les données à la main, les nouvelles technologies ont permis de rendre le stockage et la récupération beaucoup plus facile et rapide. Un temps de recherche a été nécessaire pour établir quel sport était le plus apte à mettre un nombre suffisant d'information.

Une chance pour le projet que la ligue nationale de basket américaine plus connue sous les initiales NBA est le championnat de Basket-Ball le plus suivi au monde et surtout il est le mieux renseigné au niveau des données.

Après mûre réflexion et avec un ensemble de données suffisamment fourni, j'ai convenu d'effectuer de la prédiction de résultats sur les matchs de la NBA.

L'objectif du projet est de prédire les résultats des matchs de la saison 2018/2019 de la ligue nationale de basket américaine de l'équipe jouant à l'extérieur. Puisqu'il

n'est pas question de prédire le score exact des équipes pour chaque rencontre. Il a été de prédire soit la victoire (WIN) soit la défaite (LOSS) de l'équipe visiteuse. Et par conséquent si l'équipe visiteuse est amenée à gagner, l'autre équipe a automatiquement perdu et inversement.

Il a été convenu pour se faire d'utiliser les données, les statistiques par match de fin de saison de chaque équipe. Les saisons qui ont été prises en compte s'étalent de 2012 à 2017 pour prédire les résultats de la saison 2018/2019.

Tout d'abord, l'ensemble des données que j'ai souhaité utiliser a dû être collecté. La récupération des données a été primordiale puisque sans elles, le projet ne pourrait aboutir. Une vérification de la qualité des données a été obligatoire pour s'assurer que les résultats qui en découleront soient corrects. Une connaissance poussée du type de données que l'on a voulu manipuler a été essentielle en vue d'y affecter des algorithmes de classification dans le cadre des prédictions des résultats des matchs de la saison NBA sélectionnée.

Ensuite, le projet a pour intention, mission de mettre en application différents modèles de Machine Learning dans une optique d'effectuer des prédictions. L'objectif premier a été de comparer les fonctionnements et les résultats de chacun. Ce travail a visé en priorité la précision des prédictions qui ont été opérées. Il n'a pas suffi simplement de prédire pour dire que l'on a prédit quelque chose, on a recherché à obtenir des résultats cohérents, probants mais surtout analysables.

Enfin, les résultats obtenus ont pour ambition d'être donnés, expliqués à un destinataire en particulier. Une partie consacrée à la visualisation des données mettra en avant, les résultats obtenus après l'application des algorithmes de Machine Learning sur l'ensemble de données ainsi les réponses quelle apportera pour répondre aux besoins du destinataire auquel la visualisation est destinée.

3.3 Position adoptée

Lorsqu'on ambitionne d'effectuer de l'analyse des données, il est important de se demander quelle position, quel point de vue, doit être adopté pour mener à bien le projet.

La position qui a été adoptée ou bien la ligne de conduite, le fil conducteur, de la réalisation du projet a été définie comme étant le destinataire du projet, la personne à qui le projet est destiné. Par conséquent tout au long du travail, il a fallu se mettre à la place du destinataire, dans chaque étapes du projet, pour en tirer les meilleurs résultats tout en répondant aux problématiques liées à cette position.

Plusieurs positions se sont présentées pour l'élaboration de cette étude.

La première position a été le point de vue d'un fan, supporter de la ligue nationale de Basket-Ball américaine. Ce point de vue, n'a fourni que peu de matière à exploiter. Un supporter classique est intéressé simplement par ce qui s'est déroulé durant les matchs sans aller plus en profondeur. Nul besoin d'appliquer des modèles de Machine Learning pour prédire quoi que ce soit puisque dans la majorité des cas il a seulement souhaité connaître rapidement et en temps réel les statistiques de son équipe pour savoir quelle équipe maîtrise le plus le jeu.

Pour répondre aux besoins d'un supporter, il suffit d'afficher le plus simplement possible les statistiques du match qu'il a regardé.

La deuxième position envisageable, a été de se placer du point de vue d'un "bookmaker". C'est à dire, "comme une personne morale (plateforme en ligne) ou physique permettant de parier sur des évènements, le plus souvent sportifs. L'intérêt d'un "bookmaker" est d'évaluer au travers des statistiques de la saison les matchs avant qu'ils aient lieu pour ensuite mettre en place les différentes paris"[44].

Cette position a offert plus de possibilité que la première. Employer des modèles de Machine Learning pour prédire les résultats peut s'avérer être très utile et avantageux pour les "bookmaker".

La troisième et dernière position à disposition, est celle destinée au coaching. Elle s'applique directement à l'encadrement d'une équipe ou bien d'un joueur. La prédiction de résultat à l'aide d'algorithmes de Machine Learning ont permis de comprendre l'usage des algorithmes mais surtout de détailler les facteurs qui ont permis d'en arriver à la réponse donnée par les différents modèles.

Cet aspect-ci, s'est focalisé précisément sur le travail de l'entraîneur. En effet, la possibilité d'évaluer la performance des joueurs de son équipe ainsi que de mettre en lumière les points forts ou faibles et aider potentiellement à façonner son coaching est un atout majeur pour un entraîneur. Contrairement aux deux autres points de vue, les performances sont ciblées et étudiées, elles sont l'objet principal de l'étude qui constituent cette position.

C'est donc cette troisième position qui a été adoptée pour la réalisation du projet, puisqu'elle concerne directement le coaching. L'étude a pour finalité via la manipulation de données, l'application de modèles de Machine Learning ainsi qu'une interprétation des réponses obtenues par l'intermédiaire de la visualisation de données, de comprendre comment de telles techniques appliquées à un sport, en particulier le Basket-Ball, ont-elles pu modifier, impacter, enrichir le coaching d'aujourd'hui.

Après plusieurs échanges et discussions avec le promoteur de ce mémoire il a été convenu, grâce aux éléments détaillés précédemment, de la question de recherche suivante :

De quelle manière la visualisation de données au travers de différents modèles de Machine Learning, peut-elle être appliquée afin d'impacter le coaching sportif ?

3.4 Méthodologie

Afin de répondre à la question de recherche exprimée dans la partie précédente, voici la méthodologie qui sera mise en place.

Ce travail sera découpé en trois chapitres bien distincts. Ces chapitres, comprendront une partie théorique du sujet, une mise en pratique de ce qui aura été évoqué dans la théorie sera faite par le biais du projet de recherche.

Ce fameux projet sera le fil conducteur au sein de chaque partie. Il permettra d'avancer dans les thèmes abordés dans ce mémoire tout en suivant son évolution.

Tout d'abord, notre premier chapitre traitera des données. Le but étant de présenter les sources utilisées, justifier le choix des données sélectionnées et comprendre en profondeur les différents aspects des données qui composent notre ensemble.

Dans un premier temps, on s'attardera sur l'appréhension des données qu'on pourra rencontrer dans le domaine du sport. Chaque donnée provenant de domaines bien particuliers, possèdent une typologie propre à elles.

L'étude de cette typologie de nos données permettra une meilleure compréhension de ce qu'elles signifient mais surtout de pouvoir par la suite les utiliser de la meilleure manière et d'en tirer le plus d'informations possibles.

Concernant le projet, on mettra en avant les différentes données, statistiques voulues en détaillant certaines pour bien assimiler ce à quoi elles correspondent.

Dans un deuxième temps, la récupération des données sera évoquée. Étape relativement importante et tout autant fastidieuse en fonction du nombre de données à exploiter. Dans un contexte très présent de privatisation des données en général, l'accès à des données reste très compliqué.

Tous les enjeux liés à ce contexte seront évoqués dans cette partie en expliquant les diverses méthodes utilisables pour arriver quand même à constituer un ensemble de données suffisamment conséquent pour les manipuler ensuite.

Le projet permettra d'illustrer, la méthode qui sera appliquée pour récupérer et former la collection de données qui seront exploitées, ainsi que les problèmes rencontrés dans cette quête des données de la ligue américaine nationale de Basket-Ball.

Dans un troisième et dernier temps, il sera question d'une partie essentielle du

travail de recherche lorsqu'on se sert de données : la qualité de ces données. Une mauvaise qualité peut être exprimée par des manques, des trous au sein de l'ensemble qui sera constitué et influencera directement les résultats obtenus.

Tout le projet reposera sur cette étape, il sera donc important de bien mettre en évidence dans cette partie l'importance que pourra avoir la qualité des données et les enjeux qu'il y aura autour.

Des problèmes de qualité seront expérimentés durant la réalisation du projet. Grâce à cette expérience, il sera possible de mettre en évidence les différences de résultats que l'on acquiert avec et sans un ensemble de bonne qualité.

Ensuite, ce second chapitre, traitera de la modélisation des données. Une première partie, justifiera le choix des différents modèles de Machine Learning manipulés dans le projet et détaillera le fonctionnement de chacun des modèles en mettant en lumière les avantages et inconvénients de chacun.

Comme par exemple les méthodes de classifications ou encore le besoin de standardiser les données selon les modèles afin de mieux comprendre ce qui se déroule lorsqu'on essaye de prédire les résultats des matchs.

Puis une deuxième partie, traitera de la décision de n'avoir choisi que des modèles de Machine Learning au détriment des modèles de Deep Learning.

Pour finir, cet ultime chapitre, avant la conclusion, présentera les résultats obtenus par le biais de ce projet. Ce sera à ce moment que la visualisation des données sera abordée, elle présentera sous différentes formes les réponses acquises et vulgarisera au maximum ce que l'on souhaitera démontrer afin d'être compris dans un premier temps par le destinataire des résultats, un entraîneur d'une équipe de Basket-Ballet. Dans un second temps par potentiellement d'autres personnes qui seraient susceptibles de s'y intéresser.

Le moment sera venu de faire un premier constat sur la réalisation du projet, une sorte de bilan expliquant les raisons de l'obtention de tels résultats et de quelle manière le projet pourrait être davantage performant.

4. Données

4.1 Typologies des données

Les données sont des éléments qu'on peut définir comme brutes, qui n'ont pas encore été interprétées ou mises en contexte. Elles sont extraites de ce qui est connu et qui sert comme point de départ à quelconque raisonnement ayant pour objectif de trouver une réponse en relation avec ces données.

Lorsqu'on évoque les données statistiques, une des premières choses qui vient à l'esprit d'un bon nombre de personne, est une quantité immense de chiffres et nombres en tout genre. Ces données peuvent être des informations dites numériques ou bien statistiques.

Elles sont qualifiées comme étant quantitatives. C'est à dire, qu'elles peuvent avoir soit un caractère mesurable soit un caractère repérable.

Une donnée, information est dite quantitative mesurable du moment que son caractère fait choix d'une unité de mesure appropriée. Comme par exemple avec des données traitant de la taille où la mesure peut s'exprimer en mètres ou centimètres selon le besoin de l'étude.

Tandis que pour une entité possédant un caractère repérable, ce sont des conventions qui déterminent l'échelle numérique dans laquelle la donnée a une position en fonction de sa valeur. Les températures sont des données à caractère repérable puisque l'unité pour toutes les entités recensées est la même (degrés Celsius) et que sa valeur servira de repère sur cette échelle.

Un second type de données statistiques existe. Les données dites qualitatives. On

ne peut attribuer à ces données aucune valeur ou caractéristique. C'est le cas par exemple pour, les couleurs, les textures ou encore les odeurs qui possèdent donc des propriétés physiques qualitatives.

Certains considèrent que toute donnée qui ne peut être qualifiée de quantitative est par défaut une donnée qualitative. Ce sont les deux formes de base qu'il est possible de rencontrer au moment de manipuler les données.

Il existe cependant des spécificités pour chaque donnée en fonction du domaine d'où elle est extraite.

Les informations identifiées comme appartenant au domaine du sport, sont en majorité quantitatives. Pour autant, il existe de plus en plus de données qualitatives. Ces mêmes données sont classées dans d'autres catégories qui sont propres au sport.

N'oublions pas à ce niveau qu'à chaque fois que l'on parle du monde du sport, on traite l'aspect compétitif, professionnel et non amateur du sport.

Les caractéristiques des données du sport de haut niveau vont inclure des informations spatio-temporel tout comme des informations statistiques calculées au préalable par des analystes, statisticiens.

Les entités vont être classées, catégorisées selon le sport auquel elles appartiennent. "Les informations spatio-temporelles font référence aux informations collectées dans le cadre d'une coordination spatiale et temporelle. Il comprend des informations sur la trajectoire, la trajectoire de possession, le temps de possession, la distance parcourue, etc." [34, p.51]. Les informations spatio-temporelles sont récoltées, récupérées à l'aide d'une multitude de capteurs.

Dans leur article, *A survey of competitive sports data visualization and visual analysis*, Meng Du et Xiaoru Yuan identifient deux types de données spatio-temporelles.

Pour commencer, ils distinguent des informations spatio-temporelles qu'ils qualifient d'absolues. Cette catégorie comprend les données spatio-temporelles à la fois

du sportif, de l'athlète ainsi que de la balle en fonction des objets sur le terrain. Le terrain est la référence spatiale du fait qu'il soit statique. Ces informations absolues sont une base pour l'ensemble des données spatio-temporelles et statistiques du sport quelles concernent.

Les auteurs expliquent le principe d'informations absolues de la balle, de la manière suivante : "Ce principe se réfère aux informations sur la trajectoire de la balle, y compris les coordonnées verticales et horizontales de la balle sur un court, par exemple, la hauteur de rebond d'un ballon de basket et le temps correspondant." [34, p.51].

Les mouvements de la balle dans un espace précis, en l'occurrence le terrain du sport qui est étudié, représente des données qualifiées d'absolues.

En ce qui concerne les acteurs se trouvant sur le terrain, les informations absolues qui leurs sont associées sont similaires à celles de la balle. Elles représentent leurs déplacements sur le terrain. "Informations sur les trajectoires, y compris les coordonnées verticales et horizontales des joueurs sur le terrain et le temps correspondante." [34, p.51]

Ensuite, les auteurs évoquent des données spatio-temporelles relatives. Contrairement aux entités absolues définies juste avant, les informations relatives utilisent comme système de référence le joueur ou le ballon puisqu'ils détiennent tous deux des mouvements relatifs par rapport au système de référence absolue statique qu'est le terrain.

"Au basket-ball, au football, les informations de possession utilisent un ballon ou un joueur comme référence et calculent sa position relative ; cela comprend la distance relative entre le ballon et le joueur, et le temps de possession pour chaque joueur." [34, p.51]

On peut estimer que ces données spatio-temporelles relatives représentent les faits de jeu. La possession, le tir (shooting), la tactique et bien d'autres encore correspondent aux performances d'une équipe ou des athlètes qui font le jeu. Ce type

d'informations est la priorité des analystes dans une optique d'étude et d'évaluation des performances.

Contrairement aux données spatio-temporelles, les informations statistiques n'ont pas de caractéristiques d'espace ou du temps. Les données statistiques, "se concentrent davantage sur les informations personnelles des joueurs ou des concurrents, ou sur la décision des joueurs en matière de comportement et de mouvement sur le terrain." [34, p.52].

Prenons l'exemple du basket-ball, les tirs à trois points, les lancers francs, les rebonds, les passes, les interceptions, etc., toutes ces informations sont considérées comme étant statistiques.

Cette collection d'information permet de refléter les performances d'une équipe et des joueurs pendant un match. Elles sont non seulement plus simples à récolter que les données spatio-temporelles, demandant un grand nombre de capteurs, elles permettent "d'offrir une analyse globale pour des jeux entiers et peut également offrir une analyse détaillée basée sur un événement ou un objet spécifique." [34, p.52].

Toujours en lien avec la typologie des données sportives, il existe deux types d'informations en fonction du point de vue que l'on décide d'adopter. Elles sont nommées "first person data" et "third person data".

Les informations spatio-temporelles et statistiques évoquées juste avant sont considérées comme des "third person data" puisque ces données sont produites par un point de vue extérieur sans que les acteurs principaux, les athlètes fassent référence à leurs ressentis. Ces informations "Peuvent être «captés» ou «enregistrés» comme des signaux neuro-électriques, des variables physiologiques ou des paramètres biomécaniques." [10, p.40].

Tout ce qui ne peut être verbalisé par l'athlète concerné est considéré comme une information provenant d'une position extérieure et appartient par conséquent à

la catégorie des "third person data".

En revanche, les "first person data" correspondent aux différents ressentis que peut éprouver un sportif de haut niveau durant une compétition. Certains auteurs disent de ces données qu'elles sont "non-scientifiques" à cause de la difficulté rencontrée pour les quantifier. Les difficultés rencontrées par les analystes sont compréhensibles puisque les "first person data" sont des entités qualitatives contrairement aux "third person data" qui sont quantitatives.

L'intérêt de tenir compte des "first person data" est d'avoir accès à des informations qu'aucun capteur ou autre objet n'est capable de récupérer. Elles proviennent de l'aspect psychologique du sport qui reste quelque chose de très intime, privé, singulier ou encore subjectif mais ayant un impact considérable sur les performances des athlètes. "Ici, l'intérêt est centré sur ce que le sujet vit réellement, de son propre point de vue." [10, p.40].

Les données statistiques, "third person data" ne mènent pas à une compréhension exacte du réel. Elles permettent une vision plus globale des performances produites par une équipe ou des sportifs de haut niveau. Les "first person data" sont un complément dans le but de mieux comprendre l'expérience vécue lors d'une rencontre par l'équipe ou l'athlète tout en validant les réponses mises en avant par les "third person data".

Combiner des données de différentes natures, de différents types nécessite une compréhension plus profonde de l'expérience que peuvent vivre les athlètes durant les compétitions de haut niveau. L'objectif étant de pouvoir appréhender voire maîtriser les aléas psychologiques ayant un impact énorme sur les performances sportives dans l'optique d'être le plus efficace, le plus fort peu importe dans quelle situation.

Concentrons-nous sur les informations statistiques. Elles sont l'élément central du projet qui a été réalisé. Par conséquent, il est primordial de connaître les multiples entités rencontrées, collectées et manipulées pour la suite.

La finalité étant de visualiser les résultats de l'équipe visiteuse de la meilleure des manières qu'elle que soit la position adoptée. Le sport étant un domaine qui détient des informations spécifiques et très variées, une mauvaise compréhension peut entraîner une mauvaise visualisation.

Les trois types de données qui vont suivre ne sont pas totalement distincts. "Certaines ligues sportives professionnelles utilisent des données de suivi pour calculer des données de box-score et les méta-données nécessitent généralement de mettre en perspective des données de box-score et/ou de suivi." [11, p.5].

Les informations sont très proches les unes des autres et sont souvent transformées par les statisticiens ou analystes pour les intégrer dans les box-score, ceux-ci étant la norme de retranscription des données en vue de les analyser.

Tout d'abord, les données des box-score. Comme évoqué dans l'introduction, les box-score sont des tableaux contenant différentes informations émanant simplement d'un match. Par la suite elles ont évolué pour retranscrire les données des matchs d'une saison entière voir même de la carrière d'un joueur. "On utilise ce terme pour désigner toutes les données discrètes faisant référence à un événement qui s'est déroulé pendant un match." [11, p.5]. La représentation des données sous forme de box-score est la plus répandue. Qu'on soit sportif amateur ou professionnel, fan, coach ou encore bookmaker, on a tous été en contact de près ou de loin avec des box-score.

En appliquant les critères évoqués précédemment, les données des box-score sont des entités statistiques ayant pour but de résumer un match, une saison ou même une carrière. Collectées avant à la main par un observateur extérieur et de nos jours par l'intermédiaire de capteurs, caméras et autres appareils, elles sont par conséquent catégorisées comme "third person data".

Les informations des box-score sont encore pour beaucoup assemblées à la main, malgré les avancées technologiques. On tend quand même vers une automatisation complète de la conception des entités des box-score.

Ensuite, composés en grande majorité d'informations de type statistiques, il est possible de croiser certaines données spatio-temporelles dans les box-score selon les sports. Cependant, les entités spatio-temporelles discrètes sont plus souvent catégorisés comme des "tracking data".

Contrairement aux données des box-score, les informations dites "tracking" regroupent toujours des "third person data" mais uniquement de type spatio-temporelles. Elles portent ce nom suite au développement des technologies comme les capteurs ou les caméras dans un but de suivre, de traquer tous les mouvements des athlètes par rapport à une référence spatiale représentée par le terrain où se déroule l'action.

"Ces données de suivi contrastent avec l'approche du box-score, car le volume, la variété et la précision augmentent de manière exponentielle et à des coûts moins chers." [11, p.5].

Les "tracking data" apportent une nouvelle profondeur dans l'analyse des performances du fait de leurs diversités et de leurs détails. On reviendra sur ce sujet plus tard mais les "tracking data" ont permis d'apporter une visualisation de données plus développée donnant la possibilité de combiner les données statistiques des box-score et les "tracking data".

Enfin, un dernier type de données existe, il s'agit plus d'un complément des deux premiers types expliqués précédemment. Au-delà des données très spécifiques que peuvent rassembler les "tracking data" et les données des box-score, il existe des informations que l'on peut qualifier d'additionnelles. On appelle ces données les "meta-data".

"Les meta-data peuvent concerner les règles, les capacités du stade, les caractéristiques physiques des joueurs, les couleurs des kits, les badges des équipes et les informations sur les sponsors, pour n'en citer que quelques-uns." [11, p.6].

Les "meta-data" peuvent aussi être extra sportives et provenir d'autres domaines

comme les réseaux sociaux avec les réactions en direct pendant les compétitions ou encore de la publicité pendant les pauses.

Pour la réalisation du projet, ce sont des données de type box-score de la ligue nationale de Basket-ball américaine qui ont été collectées. Voici une image représentant la source brute des informations qui ont été manipulées, exploitées.

FIGURE 4.1 – *Box-score statistiques par match des équipes, saison 2018/2019 de NBA.*

Team Per Game Stats * Playoff teams Share & Export ▼ Glossary

Rk	Team	G	MP	FG	FGA	FG%	3P	3PA	3P%	2P	2PA	2P%	FT	FTA	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS
1	Milwaukee Bucks*	82	241.2	43.4	91.1	.476	13.5	38.2	.353	29.9	52.9	.565	17.9	23.2	.773	9.3	40.4	49.7	26.0	7.5	5.9	13.9	19.6	118.1
2	Golden State Warriors*	82	241.5	44.0	89.8	.491	13.3	34.4	.385	30.8	55.3	.557	16.3	20.4	.801	9.7	36.5	46.2	29.4	7.6	6.4	14.3	21.4	117.7
3	New Orleans Pelicans	82	240.9	43.7	92.2	.473	10.3	29.9	.344	33.4	62.4	.536	17.8	23.4	.761	11.1	36.2	47.3	27.0	7.4	5.4	14.8	21.1	115.4
4	Philadelphia 76ers*	82	241.5	41.5	88.2	.471	10.8	30.2	.359	30.7	58.0	.529	21.2	27.5	.771	10.9	36.9	47.8	26.9	7.4	5.3	14.9	21.3	115.2
5	Los Angeles Clippers*	82	241.8	41.3	87.5	.471	10.0	25.8	.388	31.3	61.7	.507	22.6	28.5	.792	9.7	35.8	45.5	24.0	6.8	4.7	14.5	23.3	115.1
6	Portland Trail Blazers*	82	242.1	42.3	90.6	.467	11.0	30.7	.359	31.3	59.8	.523	19.0	23.3	.814	11.8	36.2	48.0	23.0	6.7	5.0	13.8	20.4	114.7
7	Oklahoma City Thunder*	82	242.1	42.6	94.0	.454	11.4	32.6	.348	31.3	61.3	.510	17.8	25.0	.713	12.6	35.5	48.1	23.4	9.3	5.2	14.0	22.4	114.5
8	Toronto Raptors*	82	242.4	42.2	89.1	.474	12.4	33.8	.366	29.8	55.3	.539	17.7	22.0	.804	9.6	35.6	45.2	25.4	8.3	5.3	14.0	21.0	114.4
9	Sacramento Kings	82	240.6	43.2	93.1	.464	11.3	29.9	.378	31.9	63.2	.504	16.5	22.7	.726	11.0	34.4	45.4	25.4	8.3	4.4	13.4	21.4	114.2
10	Washington Wizards	82	243.0	42.1	90.1	.468	11.3	33.3	.341	30.8	56.8	.543	18.4	23.9	.768	9.7	32.7	42.4	26.3	8.3	4.6	14.1	20.7	114.0
11	Houston Rockets*	82	241.8	39.2	87.4	.449	16.1	45.4	.356	23.1	42.0	.551	19.3	24.4	.791	10.2	31.9	42.1	21.2	8.5	4.9	13.3	22.0	113.9
12	Atlanta Hawks	82	242.1	41.4	91.8	.451	13.0	37.0	.352	28.4	54.8	.518	17.6	23.4	.752	11.6	34.5	46.1	25.8	8.2	5.1	17.0	23.6	113.3
13	Minnesota Timberwolves	82	241.8	41.6	91.3	.456	10.1	28.7	.351	31.5	62.5	.504	19.1	24.3	.787	11.3	33.5	44.8	24.6	8.3	5.0	13.1	20.3	112.5
14	Boston Celtics*	82	241.2	42.1	90.5	.465	12.6	34.5	.365	29.5	56.0	.527	15.6	19.5	.802	9.8	34.7	44.5	26.3	8.6	5.3	12.8	20.4	112.4
15	Brooklyn Nets*	82	243.7	40.3	89.7	.449	12.8	36.2	.353	27.5	53.6	.513	19.0	25.5	.745	11.0	35.6	46.6	23.8	6.6	4.1	15.1	21.5	112.2
16	Los Angeles Lakers	82	241.2	42.6	90.5	.470	10.3	31.0	.333	32.2	59.6	.541	16.3	23.3	.699	10.2	36.4	46.6	25.6	7.5	5.4	15.7	20.7	111.8
17	Utah Jazz*	82	240.9	40.4	86.4	.468	12.1	34.0	.356	28.3	52.4	.541	18.8	25.5	.736	10.0	36.4	46.4	26.0	8.1	5.9	15.1	21.1	111.7
18	San Antonio Spurs*	82	241.5	42.3	88.4	.478	9.9	25.3	.392	32.4	63.1	.513	17.2	21.0	.819	9.2	35.5	44.7	24.5	6.1	4.7	12.1	18.1	111.7
19	Charlotte Hornets	82	241.8	40.2	89.8	.448	11.9	33.9	.351	28.3	55.8	.507	18.4	23.1	.797	9.9	33.9	43.8	23.2	7.2	4.9	12.2	18.9	110.7
20	Denver Nuggets*	82	240.6	41.9	90.0	.466	11.0	31.4	.351	30.9	58.7	.527	15.8	20.9	.755	11.9	34.5	46.4	27.4	7.7	4.4	13.4	20.0	110.7
21	Dallas Mavericks	82	241.2	38.8	86.9	.447	12.5	36.6	.340	26.3	50.2	.524	18.8	25.3	.742	10.1	35.2	45.3	23.4	6.5	4.3	14.2	20.1	108.9
22	Indiana Pacers*	82	240.3	41.3	87.0	.475	9.5	25.4	.374	31.8	61.6	.517	15.8	21.1	.752	9.3	33.7	43.0	26.0	8.7	4.9	13.7	19.4	108.0
23	Phoenix Suns	82	242.4	40.1	87.4	.459	9.6	29.3	.329	30.5	58.1	.525	17.6	22.7	.779	9.1	31.3	40.4	23.9	9.0	5.1	15.6	23.6	107.5
24	Orlando Magic*	82	241.2	40.4	89.1	.454	11.4	32.1	.356	29.0	57.0	.509	15.0	19.2	.782	10.0	35.4	45.4	25.5	6.6	5.4	13.2	18.6	107.3
25	Detroit Pistons*	82	242.1	38.8	88.3	.440	12.1	34.8	.348	26.7	53.5	.500	17.3	23.1	.747	11.4	33.6	45.0	22.5	6.9	4.0	13.8	22.1	107.0
26	Miami Heat	82	240.6	39.6	88.0	.450	11.3	32.4	.349	28.3	55.6	.509	15.1	21.7	.695	11.2	35.1	46.3	24.3	7.6	5.5	14.7	20.9	105.7
27	Chicago Bulls	82	242.7	39.8	87.9	.453	9.1	25.9	.351	30.7	62.0	.496	16.2	20.7	.783	8.8	34.1	42.9	21.9	7.4	4.3	14.1	20.3	104.9
28	New York Knicks	82	241.2	38.2	88.3	.433	10.0	29.5	.340	28.2	58.8	.479	18.1	23.9	.759	10.5	34.3	44.7	20.1	6.8	5.1	14.0	20.9	104.6
29	Cleveland Cavaliers	82	240.9	38.9	87.6	.444	10.3	29.1	.355	28.6	58.5	.488	16.4	20.7	.792	10.7	31.9	42.7	20.7	6.5	2.4	13.5	20.0	104.5
30	Memphis Grizzlies	82	242.4	38.0	84.4	.450	9.9	28.9	.342	28.1	55.6	.505	17.7	23.0	.772	8.8	33.0	41.8	23.9	8.3	5.5	14.0	22.0	103.5

source by :https://www.basketball-reference.com/leagues/NBA_2019_games.html

Comme on peut le voir sur cette capture, les informations sont représentées sous la forme d'un tableau qui rappelle directement les box-score papier qui ont vu le jour pour la première fois dans le monde du Base-ball. Ces box-score sont donc disponibles de nos jours sous forme de tableaux directement en libre accès sur différents site web.

Présentés sous cette forme il est facile de distinguer qu'il est question de "third person data" et surtout d'informations de type statistiques.

Ce tableau ne montre qu'une partie des entités qui ont été utilisées. Il représente les statistiques des équipes de la NBA, de la saison 2018/2019 dont on souhaite prédire les résultats.

Une explication du tableau s'impose pour mieux comprendre les informations qu'il contient en vue de procéder à une interprétation, visualisation des données la plus adéquate possible.

Lorsqu'on prend ligne par ligne, on s'aperçoit que chacune d'entre-elle correspond aux statistiques d'une équipe en fonction de leur classement sur la saison. On va s'intéresser uniquement à la première ligne du tableau qui se réfère à l'équipe des Milwaukee Bucks.

La deuxième colonne, représenté par un "G" dans le tableau, indique le nombre de matches joués par chaque équipe. Les informations des colonnes suivantes représentent la moyenne des données récoltées sur les quatre-vingt-deux matchs disputés. En regardant les autres colonnes, on note la présence d'informations correspondant aux tirs (FG, 3PA, FT%, ...). On distingue la présence du nombre de tirs pris (FGA) ainsi que le pourcentage de tirs marqués (FG%) par match pour les deux points, trois points ou encore les lancers francs. Certaines colonnes quant à elles représentent les données liées au "playmaking", c'est à dire sur les faits de jeu qui initient une attaque ou qui évoquent une bonne défense. Comme par exemple, le nombre de rebonds offensifs et défensifs, les interceptions, les contres et les passes décisives.

Toutes ces entités représentent parfaitement les informations statistiques mise en avant dans les box-score de manière générale. Par la suite, il est de question de choisir les données qui semble être les plus aptes à traduire la réalité que l'on étudie. Parmi les informations statistiques collectés, se cache quelques entités de type "tracking" comme par exemple la vitesse de jeu. Les données dite de "tracking", après avoir été collectées par les caméras autour du terrain ont besoin d'être transformées par

les statisticiens pour être utiliser comme une information statistique.

4.2 Récupération des données

À l'heure où les données sont considérées comme le nouveau "pétrole", l'accès à ces informations se trouvant tout autour de nous, est, non seulement compliqué mais il soulève aussi de grandes questions d'un point de vue éthique.

Lorsque l'on reprend la problématique du travail de recherche que l'on mène, les données sont la pierre angulaire de ce projet. Se procurer un nombre significatif d'informations et qui plus est de bonne qualité sont la préoccupation majeure.

Pour un étudiant comme moi souhaitant effectuer de l'analyse de données sur la ligue nationale de Basket-ball américaine, comment avoir accès aux données des équipes qui composent la ligue ?

On a beau voir à la télévision un nombre incalculable de statistiques afin de donner aux supporters une expérience plus poussée des rencontres qu'ils regardent, ces informations appartiennent avant tout à la NBA qui seule autorise les diffuseurs à récolter et à exploiter ces informations.

Les entreprises spécialisées dans l'analyse de données de la NBA doivent probablement pour certaines, payer un abonnement pour avoir accès aux ressources détenues dans les bases de données de la ligue. Sans un budget suffisant impossible d'accéder aux bases de données.

Heureusement, pour pallier ce problème, des API (Application Programming Interface) existent permettant d'avoir accès aux mêmes données que celles de la NBA. Pour fournir les API en données, la plupart des développeurs d'API utilisent des méthodes de Web Scrapping.

Le Web Scrapping peut être défini comme un processus d'extraction de contenu d'intérêt du web de manière systématique, ou automatique. "Le défi consiste à traiter

ces volumes d'informations, permettant un filtrage d'informations et une intégration de données valables." [12, p.793].

Un agent logiciel ou plutôt un robot du web imite les interactions de navigation entre un humain et des serveurs web. Selon les données que l'on souhaite acquérir, le robot accède pas à pas à autant de site que nécessaire en analysant leur contenu pour trouver, extraire et structurer les informations voulues.

Au début, le Web Scrapping se concentrait sur l'extraction de ressources via le code HTML des sites web étant donné que les informations étaient directement accessibles dans leur code source.

Maintenant, à cause de l'évolution des langages de développement web, le renforcement de la sécurité des sites web ainsi que le besoin de rassembler des informations provenant d'une multitude de sources, le Web Scrapping est devenu plus complexe à mettre en place qu'auparavant. Il existe désormais des logiciels de "scrapping".

"Bien que la prolifération des services Web ait réduit le besoin de "scraper" des données Web, il existe encore des scénarios où ils sont toujours utiles." [12, p.789].

Quoi qu'il en soit, le Web Scrapping flirte avec l'illégalité. Les ressources se trouvant sur différentes plateformes web peuvent être soit brutes dans le code source et accessibles très facilement, soit elles sont protégées puisqu'elles sont stockées dans des bases de données sur des serveurs privés dont seul l'utilisateur du serveur, et/ou le propriétaire de la base de données est en mesure d'utiliser.

On peut assimiler la récupération de ces informations à du vol, du pillage, car les méthodes de Web Scrapping permettent de se servir sans aucune autorisation en contournant les sécurités rencontrées.

Les règles, les lois qui régissent Internet sont difficiles à faire respecter. On se retrouve souvent dans des situations assez floues où la frontière entre la légalité et l'illégalité est mince. La privatisation des données en est un exemple parfait. Malgré la valeur importante qu'ont les données dans la société actuelle, des questions d'éthiques se mettent en travers du chemin du Big Data.

Un accès non-autorisé aux données personnelles d'un individu est considéré comme une violation de la vie privée. De plus une fois récupérées, ces informations n'appartiennent plus à l'individu et peuvent être utilisées à des fins nuisibles contre leur propriétaire.

Un des inconvénients est le degré de sensibilité de l'information qui est récoltée. Tout d'abord, la personne qui fournit des ressources doit pouvoir contrôler la sensibilité de ce qu'elle transmet aux autres. Elle doit être en mesure de refuser l'accès aux informations les plus sensibles, les plus privées la concernant. Cependant, aujourd'hui, contrôler les données que l'on diffuse est très complexe et fastidieux. Peu de gens s'inquiètent des extraits à leur sujet.

Également, le collecteur de la donnée ne doit pas diffuser l'information si elle est considérée comme privée par l'individu. Il doit où la modifier, où la transformer afin d'enlever son caractère privé, rendant ainsi la donnée sensible utile et exploitable par de tiers personnes.

Limiter l'accès aux données peut permettre de rendre leur collecte plus saine et plus éthique qu'elle ne l'est aujourd'hui. Un individu diffuse ses informations de manière active ou passive. De façon active au moment où volontairement la personne décide de fournir ses propres données à un "data collector".

La création d'un compte sur un site web est un exemple de partage consenti de ses informations de manière active. En revanche, de façon passive, la personne transmet ses données par le biais de ses activités sur le web. Sans s'en rendre compte, un "data collector" enregistre et récupère les entités de navigation. Les cookies sur le web servent à récupérer des informations de navigation des utilisateurs qui les acceptent.

En ce qui concerne le monde du sport, les sportifs sont suivis par de nombreuses technologies qui recueillent les données de leurs performances. Il ne s'agit pas directement d'informations personnelles privées compte tenu que tout le monde peut voir leurs performances lors des compétitions en direct. Pour autant, ils n'ont pas

spécialement donné leur accord pour être surveillés comme ils peuvent l'être aujourd'hui.

On peut estimer qu'utiliser ces données relève de l'illégalité même si ce n'est pas considéré comme tel. Aussi, il faut être vigilant avec les données que l'on manipule.

Par chance, pour la réalisation du projet, un site internet du nom de *basketball-reference* (<https://www.basketball-reference.com/>) recense un très grand nombre d'informations sur la ligue nationale américaine de Basket-ball et affiche les données brutes qu'il utilise dans son code source.

Il n'a pas été question d'utiliser une API (Application Programming Interface) ou encore un logiciel de Web Scrapping, étant donné le petit volume de données nécessaires. Une partie "scrapping" des données a été exécutée à la main lors de mon travail de collecte de données.

Après avoir appliqué le script pour récupérer les données désirées, on se retrouve avec un ensemble de données composé de 9193 lignes et 67 colonnes soit un total de 615931 données.

Les informations contenues dans l'ensemble correspondent à toutes les rencontres qui ont eu lieu de la saison 2012/2013 à la saison 2018/2019. Ensuite, pour chaque équipe, il a été question de récupérer les statistiques de la saison précédente de celle de la rencontre. Par exemple, lorsqu'on regarde une rencontre de 2016/2017, les statistiques de la saison 2015/2016 des deux équipes sont utilisées.

Toutes les ressources présentes dans l'ensemble n'ont pas été utilisées. On détaillera le choix des données exploitées lorsqu'on évoquera la partie modélisation du travail.

4.3 Qualité des données

De nos jours beaucoup d'activités humaines reposent sur les données, qu'elles soient stockées ou bien manipulées sur un ordinateur. La pression continue pour

augmenter la productivité, la compétition mondiale ou encore l'application de technologies de pointe obligent et forcent les entreprises à améliorer leur processus de production de biens et de services en éliminant le gâchis, en réduisant les pertes de temps, en augmentant la vitesse de réponse aux clients. La qualité des données est un élément essentiel dans l'amélioration des processus de conception des biens et services pour répondre au mieux aux besoins des utilisateurs.

"Les recherches sur la qualité des données ont débuté à l'étranger dans les années 1990, et de nombreux chercheurs ont proposé différentes définitions de la qualité des données et des méthodes de division des dimensions de la qualité." [28, p.2]. Certains chercheurs ont défini la qualité des données comme "fitness for use".

Le "fitness for use" est un terme désignant la capacité à interpréter et utiliser des données pour répondre à un ou plusieurs besoins dans un contexte particulier. La qualité des données quant à elle est l'écart entre le "fitness for use" optimal, c'est-à-dire les données le plus pertinentes et adéquates pour répondre à un but précis, ayant la même portée que les données du réel, et le "fitness for use" des données du réel.

Un autre terme a été créé, il s'agit de "data quality dimension" qui n'est autre qu'un ensemble d'attributs définissant un standard de bonne qualité et servant à évaluer la qualité des données.

La qualité des données doit faire face à de nouveaux challenges au fur et à mesure que le Big Data évolue. Tout d'abord, le volume de données exploité ne cesse d'augmenter et influe directement sur la qualité. Il est difficile de contrôler, d'intégrer, de nettoyer et d'obtenir un haut niveau de qualité de données. Transformer les informations qui sont catégorisées comme des ressources non-structurées en entités structurées est un processus très long.

Ensuite, le Big Data a ouvert la porte à une diversification des sources, de la provenance des données, ce qui entraîne la création de différents types de données. Trois types de données sont recensés, celles dites non-structurées, celles qualifiées de

semi-structurées et les dernières catégorisées comme structurées.

L'objectif est bien évidemment d'avoir le plus de données structurées et pour cela il faut passer par un long processus de transformation. "La quantité de données non structurées occupe plus de 80% de la quantité totale de données existantes." [28, p.3].

De plus, la rapidité avec laquelle se forment les données est sans précédent. Il faut agir vite si on souhaite utiliser certaines données. "Si les entreprises ne peuvent pas collecter les données requises en temps réel ou répondre aux besoins en données sur une très longue période, elles peuvent alors obtenir des informations obsolètes et non valides." [28, p.3].

D'un côté, il n'existe pas encore d'outils suffisamment performant pour analyser en temps réel les informations. De l'autre côté, utiliser des ressources devenues obsolètes conduit généralement à de mauvaises prises de décisions.

Enfin, la valeur des données est proportionnelle aux nombres de données. Plus le nombre augmente plus la valeur relative de ces mêmes données va diminuer, puisque avec un grand volume d'information il est très probable que certaines entités n'apportent rien, qu'elles n'aient qu'une très faible valeur.

La taille d'un ensemble de données n'est pas gage de qualité, il faut donc passer du temps à analyser les informations récoltées pour affiner l'ensemble afin qu'il augmente en qualité.

Le Big Data est un domaine relativement nouveau. C'est pour cela qu'il n'existe que très peu de standards et qu'aucune définition stricte n'a été formulée en ce qui concerne la qualité des données et les critères de qualité.

"La littérature diffère sur la définition de la qualité des données, mais une chose est certaine : la qualité des données dépend non seulement de ses propres caractéristiques, mais aussi de l'environnement commercial utilisant les données." [28, p.4]. Seules les données qui répondent correctement aux besoins des utilisateurs sont considérées comme étant de bonne qualité.

Avant le début de l'ère du Big Data, les standards de qualité étaient définis en partant du point de vue de l'utilisateur, c'est-à-dire celui qui fournissait les informations. En effet, en tant qu'utilisateurs directs ou indirects il était plus simple de s'assurer de la qualité des données de leur point de vue.

Aujourd'hui, avec la diversification des sources des données les personnes qui fournissent, produisent les données n'en sont plus forcément les utilisateurs. C'est à ce moment précis qu'interpréter, mesurer la qualité des informations récoltées devient difficile.

Afin d'avoir la meilleure qualité de données, un travail en amont de leur utilisation est nécessaire. Avant même de récolter quoi que ce soit, définir, déterminer un objectif permet de mieux cibler les données que l'on souhaite acquérir. Tout doit être déterminé à l'avance pour réduire au maximum les risques de mauvaise qualité dans l'ensemble de données.

"Les utilisateurs de Big Data choisissent rationnellement les données à utiliser en fonction de leurs objectifs stratégiques ou de leurs exigences commerciales, telles que les opérations, la prise de décision et la planification.".[28, p.7].

À l'échelle du projet, on a du faire face à des problèmes de qualité des données. Durant les discussions servant à définir les paramètres du travail, l'idée de départ était de prédire les résultats des rencontres NBA en s'appuyant sur les statistiques des saisons allant de 2010/2011 à 2018/2019.

Pendant la phase de récupération des données, au moment de vérifier le bon fonctionnement du script de Web Scrapping, un manque d'information est apparu. La saison 2011/2012 n'était pas renseignée sur le site, impossible d'utiliser l'ensemble de données contenant un vide pour une saison entière. Suite à ce problème, il a été utile de vérifier sur le site l'ensemble des saisons présentes et si toutes les informations étaient bien enregistrées. Après recherches et vérification il a été convenu de commencer depuis la saison 2012/2013 et non 2010/2011.

Une autre difficulté encore liée à la qualité des données est arrivé plus tard pendant les premiers tests de prédiction sur une petite partie de la collecte des données. Les résultats obtenus concernant ce problème seront détaillés dans la partie traitant de la modélisation du projet. Les résultats n'étaient pas satisfaisants, même après avoir vérifié à multiple reprises le programme de prédiction, il était impératif de revoir complètement l'ensemble de données.

En vérifiant l'ensemble de données, certaines équipes n'apparaissaient pas autant de fois que les autres alors qu'elles disputent toutes le même nombre de rencontres par saison. Ceci s'explique par le fait qu'il est habituel dans le monde du sport, de voir des équipes changer de nom, de logo et même de ville au fil des années.

Depuis 2012 dans la ligue nationale américaine de Basket-ball, trois équipes avaient changé de nom et de ville. L'équipe des Hornets de la ville de Charlotte était anciennement appelée les Bobcats, celle des Pélicans de la Nouvelle-Orléans était autre fois surnommée les Hornets et pour finir l'équipe des Nets du New-Jersey a déménagé pour devenir les Nets de Brooklyn.

Le simple fait de renommer les équipes pour rendre notre ensemble homogène a donné des résultats significatifs attestant de la bonne qualité des données manipulées.

Comme évoqué précédemment, lorsqu'on souhaite manipuler des données, il ne faut pas négliger l'attention qu'on leurs porte. Étudier en amont les différentes spécificités que peuvent avoir les données, en vérifier leur provenance et surtout prendre le temps de contrôler la qualité des informations que l'on a récupérées et que l'on désire analyser. Tout ce travail a un impact énorme sur la suite ainsi que sur les résultats que l'on obtient.

5. Modélisation

5.1 Machine Learning ou Deep Learning

Le Machine Learning ou apprentissage automatique (selon la traduction littérale), est une évolution des algorithmes de calculs conçu pour simuler l'intelligence humaine en apprenant du contexte environnant.

L'apprentissage automatique peut aussi être défini comme un sous-ensemble de techniques d'intelligence artificielle qui permet aux systèmes informatiques d'apprendre de l'expérience antérieure (observations de données) et d'améliorer leur comportement pour une tâche donnée.

Un algorithme dit de Machine Learning est un processus informatique utilisant des données pour réaliser une tâche sans être réellement programmé afin de produire des résultats spécifiques. On qualifie ces algorithmes d'apprentissage automatique de "soft coded" puisqu'ils ont comme particularité de modifier ou adapter automatiquement leur architecture par la répétition des actions pour devenir toujours plus performant dans la tâche qui leur est donné d'accomplir.

Un parallèle peut être effectué avec l'apprentissage d'un mouvement d'un sportif. En reprenant l'exemple des nageurs qui utilisaient des appareils pour améliorer leurs mouvements de bras durant la nage, on peut distinguer la même approche d'apprentissage avec les algorithmes de Machine Learning.

La répétition de mouvements, d'actions est la clé pour obtenir les meilleurs résultats possibles tout en prenant en compte les remarques des mouvements précédents.

Le processus d'adaptation est considéré comme un entraînement, il est commu-

nément appelé "training" et sera identifié comme tel dans la suite du travail. Le "training" représente un échantillon de données d'entrée, fourni avec les résultats souhaités.

Les algorithmes d'apprentissage automatique se configurent automatiquement de façon optimale afin de produire non seulement les résultats souhaités en appliquant les données de "training", mais peut aussi se généraliser pour obtenir les résultats désirés à partir de nouvelles données auparavant invisibles. C'est cette seconde partie que l'on définit comme la phase d'apprentissage, de "learning" du Machine Learning.

Le "learning", l'apprentissage, est la capacité avec laquelle on arrive à changer selon une stimulation externe tout en se rappelant de ses expériences précédentes. Le Machine Learning est par conséquent une approche informatique qui donne de l'importance aux techniques permettant au sujet d'augmenter ou d'améliorer ses aptitudes au changement, en fonction de la tâche qui lui a été assignée.

L'objectif principal du Machine Learning est d'étudier et d'améliorer en permanence des modèles statistiques, mathématiques pouvant être entraînés, avec des données correspondant au contexte désiré, une seule fois ou en continu afin d'en déduire le futur en prenant des décisions sans avoir une connaissance complète des éléments influents, des facteurs externes.

Pour atteindre un but spécifique, le logiciel recevant toutes ces informations, a comme rôle de choisir la meilleure action possible et d'observer les résultats qui en découlent. C'est une approche statistique d'apprentissage, essayant de trouver la meilleure probabilité, celle ayant le plus de chance de réussir dans le calcul de l'action.

Dans le prolongement, le Deep Learning (littéralement : apprentissage automatique en profondeur) peut être considéré comme l'évolution du Machine Learning, une version plus poussée de celui-ci. "L'apprentissage en profondeur est une forme d'apprentissage automatique qui permet aux ordinateurs d'apprendre de l'expérience et de comprendre le monde en termes de hiérarchie de concepts." [20, p.1].

Le Deep Learning, part du principe qu'étant donné que les ordinateurs sont capables seuls de rassembler énormément de connaissances grâce aux expériences acquises des précédentes actions effectuées, l'humain n'a donc plus besoin de fournir à l'ordinateur toutes les informations normalement requises.

Le Deep Learning tout comme le Machine Learning, est basé sur la hiérarchie de concepts qui donne la possibilité aux ordinateurs d'apprendre des concepts compliqués en les concevant à partir de concepts plus simples. Ces hiérarchies sont sous formes de couches profondes.

La hiérarchie de concepts est une des caractéristiques de base de l'aptitude humaine à résoudre des problèmes et à apprendre. Cette capacité permet à l'humain de conceptualiser le monde à différents degrés ainsi que de traduire ces concepts d'un niveau à un autre c'est-à-dire de manière hiérarchique. Contrairement aux ordinateurs qui de leur côté ne peuvent résoudre des problèmes qu'à un seul niveau. Le Deep Learning grâce à l'apport des modèles mathématiques, permet aux ordinateurs d'être dotés des mêmes capacités que l'être humain.

Cependant, l'apprentissage profond contrairement au Machine Learning repose sur des algorithmes de réseaux de neurones complexes. Ces réseaux de neurones artificiels sont calqués sur ceux du cerveau humain. La profondeur de l'apprentissage dépend du nombre de neurones présents dans l'algorithme.

Les neurones forment des couches qui analysent une à une les informations des couches précédentes. L'algorithme conservera en mémoire les réussites et supprimera les échecs des couches précédentes tout en modifiant son modèle pour en améliorer son apprentissage.

Le Deep Learning a prouvé son efficacité dans de nombreux domaines par rapport au Machine Learning. Néanmoins, sa complexité est aussi un de ses points faibles puisque pour le moment, il est compliqué de comprendre avec exactitude son mode de fonctionnement.

Alors, comment choisir entre le Machine Learning et le Deep Learning ? La position que l'on adopte pour mener à bien un travail ainsi que les objectifs liés à ce travail sont les clés pour prendre cette décision.

La position du bookmaker, souhaitant prédire les résultats des rencontres de n'importe quel sport, est de chercher à avoir, en permanence la plus grande précision de prédiction possible sans forcément connaître les détails du mécanisme. Les deux types d'intelligence artificielle peuvent être utilisables pour résoudre sa problématique.

Pourtant, le Deep Learning semble apparaître comme le plus approprié pour y répondre. En effet, la profondeur que donne le Deep Learning conduit dans la majorité des cas à de meilleurs résultats que le Machine Learning. C'est exactement ce que recherche le bookmaker.

Lorsqu'on reprend la position adoptée du projet, à savoir, un entraîneur d'une équipe de la ligue nationale américaine de Basket-Ball qui souhaite utiliser des techniques d'apprentissage automatique pour améliorer les performances de son équipe, dans sa conception, le Machine Learning est la meilleure option.

Connaître les résultats des matchs peut paraître comme un énorme avantage quand on se positionne en tant qu'entraîneur. En réalité, le plus important n'est pas de savoir quand on va gagner mais plutôt quand on va perdre. Personne n'aime la défaite. C'est d'ailleurs pour cela que les sportifs s'entraînent et travaillent autant, pour être les meilleures. C'est maintenant que le rôle de l'entraîneur prend tout son sens, il doit éviter les défaites.

Contrairement au Deep Learning, le Machine Learning donne la possibilité de comprendre le raisonnement, le travail de l'algorithme. Certes, les résultats ne sont pas aussi performant que le Deep Learning mais ce n'est pas le seul intérêt du Machine Learning. Pour un entraîneur, être capable de comprendre les facteurs responsables des défaites prédites est un atout majeur sur lequel il peut cibler et corriger les faiblesses de son équipe.

La visualisation des données est la clé, c'est elle qui met en lumière, qui vulgarise le fonctionnement des algorithmes de Machine Learning afin que les entraîneurs, tout comme les joueurs puissent être en mesure de les comprendre.

Le nombre de neurones, de couches qui composent les algorithmes de Deep Learning ainsi que tous les paramètres nécessaires pour les appliquer sont un frein et empêchent une visualisation complète de ce qui se déroule au sein de ces algorithmes.

En revanche, le Deep Learning peut être appliqué pour répondre à certains besoins d'un entraîneur. Sur un terrain de Basket-ball on compte cinq joueurs par équipe. Si un entraîneur souhaite connaître parmi tout son effectif, quel est le meilleur cinq de départ selon chaque poste, alors il peut utiliser le Deep Learning .

En prenant comme données d'entrée, le nom des cinq titulaires qu'il a l'habitude de faire jouer ainsi que les autres joueurs et les statistiques de chacun, le Deep Learning grâce aux réseaux de neurones et aux multitudes de couches est capable de définir le meilleur cinq de départ.

Seule la réponse donnée par l'algorithme est importante, l'entraîneur n'a en aucun cas besoin de se connaître les éléments précises qui ont entraînés ce résultat.

Dans le cadre du projet, pour toutes les raisons évoquées juste avant, le choix s'est logiquement porté sur l'application d'algorithmes de Machine Learning plutôt que de Deep Learning. La visualisation des données est primordiale dans le travail et le Machine Learning est le plus apte à pouvoir répondre aux besoins et objectifs du projet. Dans l'avenir, si le travail venait à se développer davantage, allier Machine Learning et Deep Learning permettrait de pousser encore plus loin l'analyse de performances et la prise de décision.

Avant de détailler les différents modèles appliqués, il faut commencer par introduire les approches les plus communes du Machine Learning. Elles vont permettre de mieux comprendre certains choix effectués pour la réalisation du travail.

Tout d'abord, il existe l'apprentissage supervisé. Un scénario décrit comme su-

pervisé est caractérisé par le concept d'un superviseur ayant pour rôle de fournir un ensemble de données dit de "training" composé des données d'entrées et de sorties attendues.

Par exemple, en prenant le cas du projet, les rencontres de chaque saison avec les statistiques des équipes qui s'affrontent représente les données d'entrées. Le résultat des rencontres exprimé sous la forme suivante, "WIN" pour la victoire et "LOSS" pour la défaite représentent les données de sorties attendues.

Une fois ces informations transmises à l'algorithme, il est en mesure d'adapter ses paramètres pour réduire autant que possible son taux d'erreur. L'algorithme effectue plusieurs itérations. Si les données qu'il traite sont suffisamment cohérentes, l'écart entre les prédictions effectuées et les résultats attendus se réduiront au fur et à mesure des récurrences jusqu'à se rapprocher autant que possible de zéro et par la même occasion augmente la précision globale de l'algorithme. Si l'écart est égal à zéro, la prédiction correspond au résultat souhaité.

Dans ce type de scénario, il est aussi possible de transmettre des échantillons que l'algorithme n'a jamais vus. On les appelle généralement des ensembles de "test" ou encore de "validation".

Pour ce faire, l'ensemble dit de "training" comme son nom l'indique, sert à entraîner le modèle afin qu'il se généralise dans l'optique de réduire au maximum les problèmes de "surapprentissage", en anglais "overfitting".

"Overfitting", traduit par "sur-apprentissage" ou "sur-ajustement", ou encore "sur-interprétation", désigne le moment où le modèle prend en compte les détails et le bruit présents dans les données de "training" ayant un impact négatif sur les performances du modèle appliqué aux données de "test".

Dans le détail, cela signifie que le bruit présent dans les informations de "training" est intégré par le modèle et impacte négativement sa généralisation avec les données de "validation".

En opposition à l'"overfitting", si un modèle n'est ni en capacité d'être entraîné, ni capable de se généraliser avec les données de "validation", c'est alors un problème qualifié d'"underfitting".

Selon le type de données de sorties attendues, il existe deux types d'obstacle dans la prédiction.

Dans un premier temps, si l'on souhaite prédire des données ayant des valeurs continues, il est question d'un problème de régression. La régression sert dans la prédiction de quantité. Cette dernière est majoritairement continue mais peut être discrète du moment qu'il s'agit d'un nombre entier. Par exemple, si l'objectif du projet avait été de prédire le score des rencontres de la saison 2018/2019, cela aurait été un problème de régression.

Dans un second temps, en prenant pour exemple le contexte du projet souhaitant prédire la victoire ou la défaite, une donnée de sortie avec un nombre discret de possibilité en l'occurrence ici deux valeurs possibles "WIN" et "LOSS", il s'agit d'un problème de classification ou bien de classification binaire puisque les données de sorties sont divisées en deux classes. Elles sont appelées labels ou catégories.

Un modèle de classification est appliqué dans la majorité des situations pour prédire des variables discrètes, le label ou la catégorie des données de sorties. Cependant, il est possible de voir des modèles de classification prédire une valeur continue telle que la probabilité d'appartenance d'une donnée à chaque classe de sortie.

Ensuite, l'approche non-supervisée à l'inverse de celle supervisée, s'applique sur des données ne possédant pas d'étiquette. Aucun superviseur, humain n'a identifié les données comme c'est le cas dans l'approche supervisée. Cette autre approche apprend des modèles en utilisant deux méthodes.

La première est la méthode probabiliste. Un des principaux procédés probabilistes employés dans l'approche non-supervisée, est le "Clustering" que l'on peut traduire par le regroupement. L'objectif de cette technique est de regrouper un ensemble de données dans des groupes plus communément appelés "cluster", afin que les données composant ces "cluster" soient les plus proches possible au niveau de leurs caractéristiques.

La deuxième méthode a déjà été évoquée au moment de la description du Machine Learning et du Deep Learning, ce sont les réseaux de neurones. Le réseau dans

son fonctionnement essaye d'imiter les caractéristiques des informations transmises en entrée et s'auto-corrige à l'aide de ses expériences précédentes.

Enfin, il existe l'approche semi-supervisée qui comme son nom le laisse entendre correspond à un mélange des approches supervisées et non-supervisées. Elle utilise des données labellisées et non-labellisées. On la retrouve lorsque l'ensemble de données à exploiter contient un grand nombre d'informations dont seulement une partie correspond à des entités labellisées.

Encore une fois, on peut souligner l'importance, d'avoir à la fois une problématique de travail précise, détaillée mais aussi de connaître en profondeur les différentes caractéristiques des données qui constituent l'ensemble à analyser. Derrière cette partie, c'est l'entièreté du travail qui en dépend. Le choix des méthodes de Machine Learning ou Deep Learning, les approches nécessaires en fonction des données à analyser ainsi que les algorithmes qui sont appliqués sur ces informations, découlent de ce qui a été identifié en amont.

5.2 Machine learning dans le sport

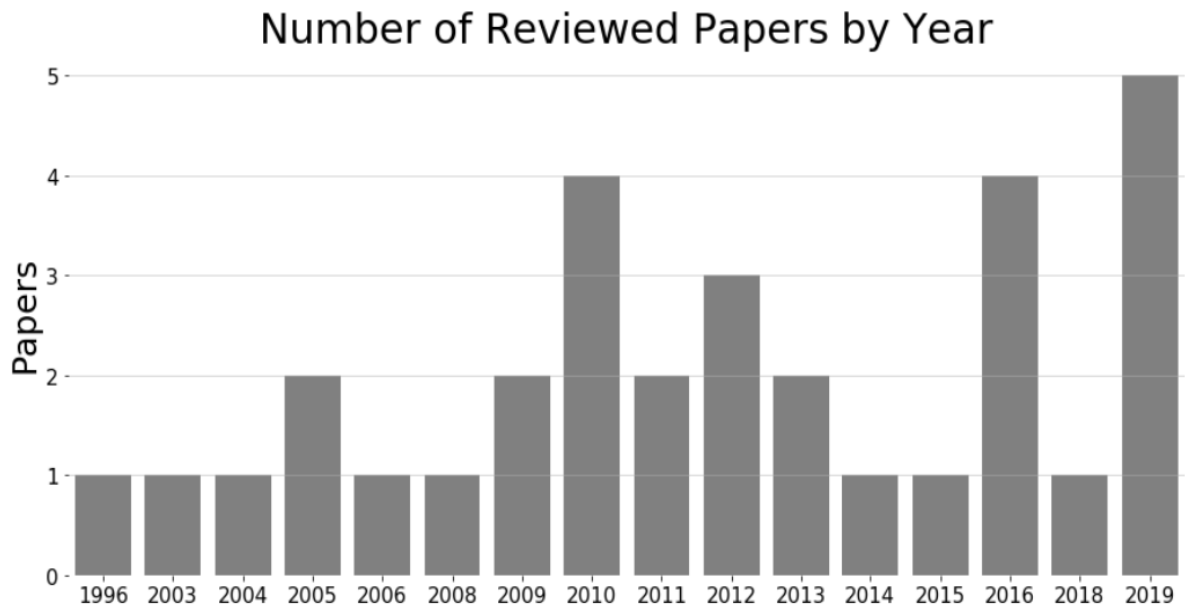
La nature imprévisible du sport rend la prédiction de performances, de résultats difficile en partie dû au nombre de facteurs potentiels pouvant influencer ces résultats et ces performances. La disponibilité sur internet des informations du monde du sport, a amplifié l'intérêt de beaucoup pour la prédiction des résultats sportifs notamment dans le cadre des paris. Elle peut être tout aussi influente pour les joueurs, les directions d'équipe ou bien les entraîneurs. L'analyse des performances peut identifier les facteurs importants qui valorisent les résultats positifs d'une équipe et sur lesquels des tactiques appropriées peuvent s'appuyer.

La prédiction dans le monde du sport est dans la majorité des cas, abordée comme un problème de classification. Vouloir prédire le score exact d'une rencontre, le nombre de points marqués ou encore les rebonds attrapés par un joueur, est jugé plus compliqué puisqu'il s'agit d'un problème de régression, et que prédire

seulement deux classes comme dans un problème reste plus facile et rapide.

Les chercheurs Rory Bunker et Teo Susnjak, au travers de leur étude intitulé *The application of machine learning techniques for predicting results in team sport*, ont cherché à déterminer quels étaient les algorithmes de Machine Learning qui avaient tendances à être utilisés le plus souvent. Voici trois graphiques issus de leur étude permettant de mieux comprendre l'importance et la progression du Machine Learning dans le monde du sport, ainsi que les modèles que la plupart des chercheurs ont appliqués dans le cadre de leur travaux.

FIGURE 5.1 – *Nombre d'articles examinés par an.*



source by : *Bunker (Rory) and Susnjak (Teo). The application of machine learning techniques for predicting results in team sport : A review (pp.4).*

La figure 5.1 ci-dessus, exprime bien l'accroissement de l'intérêt porté par les scientifiques au monde du sport en lien avec le Machine Learning. Elle représente le nombre d'études effectuées dans ce domaine entre 1996 et 2019, où au minimum, une technique de Machine Learning a été appliquée. Pendant pas mal d'années, la prédiction de résultats était considérée comme appartenant au monde des statistiques et ce n'est qu'à l'avènement de l'apprentissage automatique, fin des années

1990, que le rapport au sport a pris de l'ampleur.

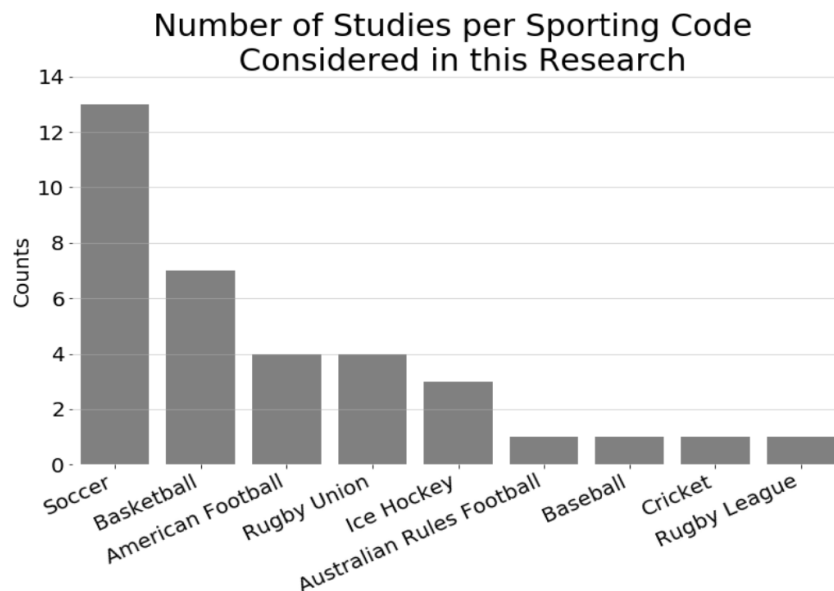
" La première étude dans ce domaine a été publiée en 1996 et le sujet a depuis lors suscité un intérêt accru pour la recherche." [39, p.3] Au sein des études recensées depuis 1996 par les auteurs de l'article *The application of machine learning techniques for predicting results in team sport*, neuf sports ont été identifiés.

En s'attardant sur la figure 5.2, le sport sur lequel le plus d'études ont été effectuées, n'est autre que le football (soccer).

Étant le sport le plus populaire sur le globe, le voir en première position n'est pas une surprise. Le nombre de matchs se jouant par an dans le monde permet d'acquérir une très grande quantité de données statistiques bien identifiées et renseignées donnant une matière solide à exploiter et à analyser.

Par exemple, dans le cadre des paris sportifs, il est la cible privilégiée puisqu'il est le sport le plus suivi, le plus télévisé au monde et touche plus de personnes que n'importe quel autre sport.

FIGURE 5.2 – *Nombre d'études par sport concernant les modèles de prédiction.*



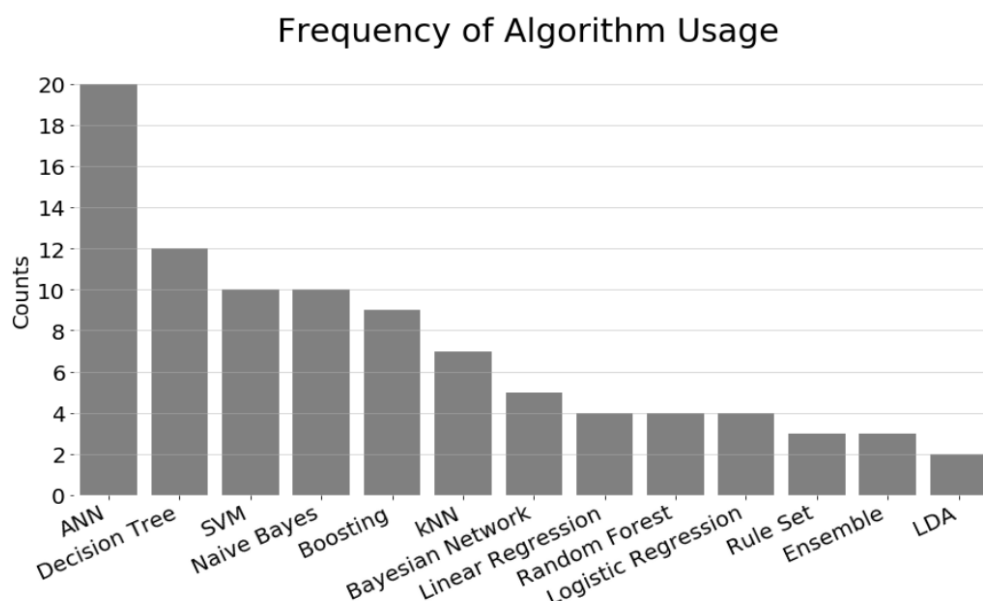
source by : *Bunker (Rory) and Susnjak (Teo). The application of machine learning techniques for predicting results in team sport : A review (pp.6).*

Le basket-ball arrive juste derrière le football (soccer). Extrêmement populaire aux États-Unis, il reste cependant un peu plus confidentiel dans certains pays. Les États-Unis, fondateur du basket-ball, est surtout le pays dans le monde qui investit le plus d'argent dans la pratique et l'étude de ce sport, dès l'école primaire et jusqu'au niveau professionnel. La culture du sport représente beaucoup pour les américains. C'est pour ces raisons que le basket-ball et le football-américain font partie du trio de tête, au niveau des sports étudiés.

Les algorithmes, les modèles de prédiction appliqués aux données sportives reviennent plusieurs fois pour la plupart des sports. Aucun de ces modèles n'appartient exclusivement au monde du sport, ils sont tous manipulés dans divers domaines. La figure 5.3 ci-dessous, met en avant la fréquence d'usage des modèles, trouvés dans les études dédiées au monde du sport.

L'algorithme le plus employé a pour acronyme "ANN" signifiant *Artificial Neural Network* traduisible par "Réseau de Neurone Artificiel". On le retrouve en haut du classement puisque qu'il est celui dont les résultats de prédiction sont les plus élevés dans quasiment tous les sports désignés dans la figure 5.2 et ce peu importe le nombre de variables et de matchs qui lui est transmis. Une précision de prédiction avoisinent les 80% ou plus est remarqué lorsqu'il est appliqué seul ou bien avec d'autres modèles.

FIGURE 5.3 – *Fréquence d'usage des algorithmes de prédiction.*



source by : *Bunker (Rory) and Susnjak (Teo). The application of machine learning techniques for predicting results in team sport : A review (pp.5).*

Cela ne veut dire pas pour autant dire que les autres modèles ne sont pas viables. Mais ils sont simplement moins performants que le modèle ANN sur des données sportives. Certains en revanche, permettent une bien meilleure visualisation des données et des réponses obtenues que lui.

Chaque modèle réussit plus ou moins bien à obtenir un haut pourcentage de bonne prédiction selon le type de données sur lesquelles il travaille.

Pour la réalisation du projet, en s'appuyant sur les travaux de Rory Bunker et Teo Susnjak, les modèles adoptés pour répondre à l'étude sont présents dans la figure 5.3 (kNN, Naïves bayes, SVM, Decision tree, ect.).

Une présentation et une description du fonctionnement de chacun d'eux est effectuée dans le paragraphe suivant pour mieux comprendre les facteurs ayant eu un impact sur les réponses acquises.

5.3 Modèles appliqués

Avant de commencer la description des différents modèles exploités pendant le projet, un descriptif détaillé des variables composant l'ensemble de données du projet s'impose pour mieux comprendre sur quelles matières les modèles de prédiction sont appliqués.

Tout d'abord, les informations récupérées représentent plusieurs instants, moments d'une rencontre de Basket-Ball.

L'aspect offensif du jeu, correspond aux points que l'on peut marquer durant un match. L'ensemble contient par conséquent, le pourcentage de tirs réussis, le nombre de tirs tentés pendant le match, le nombre de panier marqués, que ce soit un tir à 2 ou à 3 points, les passes décisives et les rebonds offensifs.

L'aspect défensif du jeu quant à lui, correspond aux rebonds défensifs, aux interceptions, et aux contres.

Des informations complémentaires concernant les faits de jeu, en l'occurrence les fautes commises, les "Turnovers" représentant les pertes de balles, la vitesse de jeu des équipes, le pourcentage de "True Shoot" représentant l'efficacité d'une équipe au tir et celui de "l'efficient Field Goal" qui est un réajustement du pourcentage de tir marqué, font partie des variables analysées. De même, certaines statistiques plus globales, telles que le "Offensive Rating", "Defensive Rating" et le "Net Rating" indiquant les performances des équipes sur le plan offensif pour marquer des points et défensif pour dissuader l'autre l'équipe de marquer. Les pourcentages de rebonds offensifs et défensifs pris après un tir ou un lancer franc manqué, se retrouvent dans l'ensemble de données.

D'autres variables ont été récoltées. Cependant, après plusieurs lectures traitant du même sujet, que le projet réalisé, certaines informations ont été estimées comme non essentielles car n'ayant que peu d'impact dans la prédiction des résultats (WIN ou LOSS).

"L'ensemble des variables prédictives comprenait : le pourcentage de tirs sur le terrain, le pour-

centage de trois points, le pourcentage de lancers francs, les rebonds offensifs, les rebonds défensifs, les passes décisives, les vols, les blocages, les revirements, les fautes personnelles et les points. Pour ce qui est de la prévision des matchs non joués, les moyennes des variables de la saison en cours se sont avérées donner de meilleures performances que la moyenne des statistiques des cinq derniers matchs."[39, p.12].

La première partie du travail consiste à prédire les résultats des rencontres de la NBA. Pour ce faire la prédiction s'effectuera sur le résultat de l'équipe visiteuse, celle qui joue à l'extérieur.

Lors de la transmission des données aux modèles, les informations sont structurées de la manière suivante :

Chaque ligne qui compose l'ensemble dit de "training" représente une rencontre de la NBA de la saison 2012/2013 à la saison 2017/2018. La ligne contient dans un premier temps, les variables appartenant à l'équipe visiteuse que l'on a nommée "Visitor". Dans un deuxième temps, celles de l'équipe jouant à domicile quant à elle définie comme "Home". Et dans un dernier temps, la dernière colonne de l'ensemble de données de "training" correspond au résultat du match, c'est-à-dire "WIN pour une victoire et "LOSS" pour une défaite, toujours de l'équipe visiteuse.

FIGURE 5.4 – Échantillon de l'ensemble des données du projet.

Visitor.Pace	Visitor.TS.	Visitor.eFG.	Visitor.TOV.	Visitor.OffRB.	Visitor.DefRB.	Visitor.Assists.Turnover	Home.Field.Goal.Attempts	Home.Field.Goal.	Home.3.Point.Attempts	Home.3.Point.	Home.Offensive.Ret.
92.5	0.508	0.472	14.2	27.3	70.9	1.248366	81.2	0.422	19.3	0.346	12.7
90.4	0.535	0.496	14.7	19.7	72.4	1.594595	79.0	0.469	15.6	0.359	10.4
91.4	0.527	0.489	13.4	23.4	74.8	1.492857	80.6	0.457	16.8	0.326	12.1
94.2	0.556	0.516	14.1	27.7	74.3	1.558442	83.6	0.448	14.6	0.362	10.7
90.7	0.526	0.474	13.1	29.2	72.3	1.328571	78.1	0.440	16.3	0.340	10.6
91.7	0.529	0.492	13.5	27.5	73.4	1.468966	79.3	0.438	13.9	0.346	11.7
94.7	0.510	0.472	13.0	29.1	70.5	1.340278	82.8	0.452	16.9	0.375	13.9
92.9	0.562	0.528	12.8	25.1	76.0	1.705882	77.3	0.451	11.8	0.333	11.0
91.4	0.527	0.489	13.4	23.4	74.8	1.492857	83.8	0.456	12.8	0.323	13.0
92.3	0.539	0.505	13.3	22.9	69.1	1.604317	82.5	0.458	19.6	0.343	10.9
90.8	0.515	0.473	13.6	29.8	72.7	1.344828	81.3	0.455	21.8	0.357	12.1
90.5	0.534	0.491	14.2	29.1	74.8	1.490066	82.1	0.443	20.9	0.346	11.1
93.0	0.567	0.516	15.3	27.8	72.1	1.134969	82.8	0.478	21.3	0.393	10.3
90.7	0.526	0.474	13.1	29.2	72.3	1.328571	80.2	0.414	13.5	0.295	10.3
94.2	0.556	0.516	14.1	27.7	74.3	1.558442	78.2	0.441	27.0	0.375	11.2
91.7	0.529	0.492	13.5	27.5	73.4	1.468966	81.0	0.454	20.2	0.370	9.9
93.7	0.521	0.481	12.9	27.7	70.9	1.666667	77.1	0.460	15.0	0.367	7.7
89.1	0.523	0.490	13.2	32.6	74.3	1.650000	81.2	0.422	19.3	0.346	12.7
94.7	0.510	0.472	13.0	29.1	70.5	1.340278	82.3	0.433	21.6	0.332	12.1
91.4	0.525	0.481	13.0	30.2	73.8	1.535211	77.3	0.451	11.8	0.333	11.0

On constate sur la figure 5.4 que les variables composant l'ensemble de données ne sont pas toutes encodées avec la même unité. Certaines statistiques sont des pourcentages, d'autres sont des nombres entiers et d'autres comme par exemple le "Net Rating" sont des entiers avec une valeur soit positive soit négative.

Ce n'est pas un problème pour quelques algorithmes et modèles de Machine Learning comme les arbres de décisions ou la forêt aléatoire. Alors que d'autres algorithmes de classification tels que le k-Nearest-Neighbors, Support Machine Vector ou encore Naïves Bayes ne sont pas en mesure de travailler correctement lorsque les données ne sont pas harmonisées.

Par conséquent, un travail de standardisation des variables doit être effectué avant de les transmettre aux différents algorithmes. Le type de standardisation qui a été appliqué dans le projet est un des plus classiques. L'objectif est de transformer la valeur des variables pour que celles-ci aient la même unité de référence.

Pour ce faire, la variable que l'on souhaite transformer est soustraite à la donnée ayant la plus petite valeur de la colonne pour ensuite être divisée par la valeur maximum de la colonne en question qui elle même est soustraite à la valeur minimum de la même colonne ($varX = (varX - \min(colonne)) / (\max(colonne) - \min(colonne))$).

En revanche, la standardisation des données peut réduire l'importance de certaines variables. Comme il est question de transformer les informations, elles peuvent ne plus exprimer la même chose une fois modifiées. Plusieurs standardisations existent, il est judicieux d'en essayer quelques-unes dans la mesure du possible et de garder celles qui dénaturent le moins la valeur des données.

Maintenant que la composition de l'ensemble de données est éclaircie, la présentation des modèles d'apprentissage appliqués va permettre d'avoir une vision plus précise de leur fonctionnement et par la même occasions être en mesure de faciliter la compréhension des réponses qu'ils apportent.

5.3.1 k-Nearest Neighbors

L'algorithme k-Nearest Neighbors, plus connu sous son acronyme kNN, se rapporte à la famille des algorithmes d'apprentissage supervisé. Souvent utilisé pour répondre à des problèmes de classification ou régression, il est considéré, par beaucoup, comme étant l'algorithme le plus simple et facile à mettre en place.

Son fonctionnement est le suivant : il prend en informations d'entrée un ensemble de données étiquetées ayant des valeurs de sorties correspondant au même label, lui servant à entraîner son modèle. Son rôle est de trouver les plus proches voisins de la nouvelle donnée.

Premièrement il s'agit de définir le nombre de voisins représenté par K et calculer par la suite la distance entre les données déjà labellisées (K voisins) et la nouvelle variable entrante.

Deuxièmement, l'algorithme a pour mission de sélectionner le nombre K de voisins en fonction de la distance calculée précédemment. Pour ensuite au sein de ces K voisins, déterminer le nombre de points se rapportant à chaque catégorie.

Enfin, en fonction de la catégorie la plus présente parmi les K voisins, l'algorithme attribuera la nouvelle variable à la catégorie la plus représentée. C'est à ce moment que le modèle généré par l'algorithme kNN est prêt à être employé.

5.3.2 Support Vector Machine

On le croise plus souvent sous la forme SVM, l'algorithme Support Vector Machine tout comme le kNN fait partie des techniques d'apprentissage supervisé. Le concept de cet algorithme est de projeter les données dans un espace de grande dimension pour être capable de les séparer.

Les variables sont disposées sur un espace de deux dimensions. Une droite dite de séparation est calculée par l'algorithme et ensuite placée sur l'espace de deux dimensions où se situe les données. Une marge représentant la distance entre la droite et la donnée la plus proche de chaque côté est définie. Comme son nom l'indique la droite a pour mission de séparer les données en autant de classes que nécessaire.

En prenant l'exemple des résultats des matchs NBA, la droite va séparer en deux l'espace de deux dimensions, les variables se trouvant au-dessus de la droite seront attribuées à la classe "WIN" tandis que si elles se trouvent en dessous elles iront dans la classe "LOSS".

5.3.3 Logistic Regression

C'est un des algorithmes d'apprentissage automatique le plus simple et le mieux interprétable qui existe. Capable de manipuler des données à la fois discrètes et continues, la Régression Logistique est un classique des algorithmes de Machine Learning. Ce qui le caractérise est sa capacité à comparer, et à étudier la relation entre deux variables quantitatives. On parle d'un modèle dit linéaire.

Contrairement aux autres modèles acceptant des variables quantitatives comme "WIN" et "LOSS" dans le projet, et afin d'appliquer cet algorithme, les deux classes à prédire ont été transformées de sorte à ce qu'elles soient numériques. Par conséquent, la classe "WIN" est devenu 1 et la classe "LOSS", 0.

La Régression Logistique, est un modèle statistique. Elle permet de prédire la

probabilité qu'un évènement arrive, une victoire de l'équipe (1) ou bien une défaite (0), à partir des coefficients de régression.

Un seuil est défini pour permettre de déterminer à quel moment on considère que l'évènement est réussi ou non. Par exemple, en ce qui concerne l'évènement de victoire d'une équipe, si le seuil est défini à 0,45, toutes les valeurs prédites supérieures au seuil de l'évènement seront donc susceptibles de se produire.

5.3.4 Naïves Bayes

La classification naïves bayésienne est une forme de classification probabiliste basée sur le théorème de Bayes avec une forte indépendance des hypothèses, d'où sa qualification de "naïf".

Il repose sur le principe selon lequel l'existence d'une caractéristique pour une classe est indépendante par rapport à l'existence des autres caractéristiques. Par conséquent, il permet de prédire une classe donnée selon un ensemble de caractéristiques à partir de probabilités. L'algorithme calcul donc la probabilité de chaque caractéristique pour ensuite les comparer à celle de la variable inconnue.

Ainsi, en prenant l'exemple d'un légume, on pourrait prédire si un légume est une tomate, une carotte ou une asperge en fonction de sa couleur, de sa forme, et d'autres caractéristiques, en considérant à part chacune d'elles.

5.3.5 Decision Tree

L'"arbre de décision" ou encore "arbre décisionnel", contribue à prendre une décision au moyen d'une série de questions, que l'on peut aussi appeler des "tests", dont la réponse conduira à la décision finale. En analyse de décision, un arbre de décision peut être utilisé pour représenter le processus qui a amené aux décisions prises.

Dans chaque arbre de décision, la question posée représente un noeud, l'endroit où une branche se sépare en deux autres branches. Selon la réponse et en fonction d'elle, il s'agira de s'orienter vers l'une des deux branches pour arriver jusqu'à la

feuille, symbolisant la fin de l'arbre, l'extrémité où se trouve la réponse finale. L'arbre se construit en séparant l'ensemble des données en sous-ensembles selon la valeur d'une caractéristique d'entrée. Ce processus est répété sur chaque sous-ensemble obtenu de manière récursive.

L'algorithme est en charge de trouver pour chaque noeud de l'arbre décisionnel, la question la plus pertinente possible. Il calcule pour toutes les caractéristiques, le gain d'information potentiel en fonction de la caractéristique sélectionnée.

5.3.6 Random Forest

La forêt aléatoire est catégorisée comme une méthode d'ensemble. En d'autres termes, elle combine et associe plusieurs résultats pour obtenir un résultat final. Composée généralement d'une dizaine voire d'une centaine d'arbres de décision et ajustable selon les besoins par validation croisée, elle est une méthode d'estimation de fiabilité d'un modèle basé sur une technique d'échantillonnage.

Chaque arbre entraîné représente un sous-ensemble de l'ensemble de données et propose un résultat, qui est ensuite ajouté aux résultats des autres arbres de la forêt pour amener à une réponse finale.

Cette méthode s'appelle le "bagging". Dans l'optique de créer un modèle robuste, l'ensemble de données est divisé en plusieurs sous-ensembles aléatoires. Puis les sous-ensembles sont tous entraînés, produisant autant de modèles que de sous-ensembles. Enfin, les réponses de chacun sont fusionnées pour n'en faire ressortir qu'une seule.

5.3.7 XGBoost

XGBoost est le nom d'une bibliothèque optimisée du modèle "Gradient Boosting" d'apprentissage automatique, qui se veut être hautement performante et flexible. Le "Gradient Boosting" est une méthode d'agrégation de modèles. Il crée un ensemble d'algorithmes d'apprentissage faibles. En d'autres termes, il combine différents modèles de Machine Learning ou de Deep Learning afin d'obtenir une meilleure précision de prédiction, que celle qu'aurait pu obtenir seul les modèles qu'il a sélectionnés.

tionné.

Par exemple, lorsqu'un modèle obtient de mauvaises performances par rapport aux autres, l'algorithme "Gradient Boosting" va alors décider de "booster", d'améliorer les performances de celui-ci, résultant parfois à des performances surpassant celles de tous les autres modèles.

Lors de la création de chaque modèle, l'algorithme calcule le gradient, la variabilité ou le poids de plusieurs variables afin d'indiquer le chemin, la direction qu'il faut prendre pour atteindre le minimum recherché (la précision la plus élevée).

Cet algorithme est un cas particulier du "boosting". Le "boosting" est une méthode permettant de transformer les modèles d'apprentissage automatique que l'on considère comme faibles, ayant de mauvaises performances, en un modèle d'apprentissage automatique fort.

Un algorithme dit de "boosting" tel que le "Gradient Boosting" ou encore le "Ada Boost", a pour principe de modifier les arbres de décisions. Il donne un poids égal, à toutes les observations présentes dans l'arbre de décisions qu'il forme. Le poids de celle-ci est augmenté lorsque l'algorithme estime qu'il est difficile de la classer et diminue le poids de la variable quand elle est facile à classer. Ce processus est effectué un certain nombre de fois.

L'objectif est d'améliorer les performances de classification en créant un modèle unique plus efficace et performant, à partir d'autres modèles individuels.

6. Résultats

Après avoir détaillé et expliqué les différents concepts et notions permettant de comprendre les différents enjeux ainsi que les démarches et les éléments composant le projet qui réalisé, il est temps à présent de passer à la présentation des résultats obtenus et de rentrer plus en profondeur dans le projet.

6.1 Résumé des résultats

Les résultats présentés précédemment correspondent à la précision de prédiction des différents modèles ayant été manipulés. Dans cette partie, on verra que les résultats évoluent en fonction des situations rencontrées pendant la réalisation du projet, et seront mis en lumière les différents facteurs déjà abordés en amont du travail ayant impacté les réponses obtenues.

Tout d'abord il a été décidé, d'exploiter uniquement la saison 2012/2013 de la ligue nationale américaine de basket-ball en vue de tester les algorithmes de prédiction et de comprendre leur fonctionnement, pour ensuite généraliser leur application à l'ensemble des données récoltées. La saison a été découpée de sorte que 70% des rencontres soient utilisées pour former l'ensemble de données de "training" et les 30% restant ont constitué la catégorie de données dite de "test" servant pour la prédiction.

Au début, seul l'algorithme k-Nearest-Neighbors a été manipulé, dans le but à la fois de tester la qualité de l'ensemble de données en attendant d'avoir des résultats satisfaisants et pour ensuite appliquer des algorithmes plus complexes.

Les premiers pourcentages de prédiction se situaient entre 49% et 51% de réussite.

Ce type de pourcentage exprime des prédictions aléatoires, sans qu'aucun modèle de prédiction fonctionnel n'est pu être créé à partir des données fournies en entrée. Un mauvais résultat qui révèle un problème se situant, soit dans l'ensemble de données soit dans l'utilisation de l'algorithme de prédiction.

Afin de vérifier la qualité des informations utilisées, il a été convenu de prédire les résultats des rencontres avec comme seule variable le score des deux équipes à la fin de la rencontre. Une prédiction infaillible qui par conséquent devait arriver 100% de réussite, avec des données de bonne qualité.

Malheureusement, le résultat fut le même, avec toujours plus ou moins 50% de précision de prédiction. C'est à ce moment que le problème de qualité des données a été identifié. Il provenait des équipes ayant changées de nom en cours de saison, point déjà évoqué plus en détails dans la partie qualité des données du travail.

Une fois le problème résolu, le test de prédiction des résultats avec comme unique information le score des rencontres, a été reconduit et a obtenu avec succès 100% de précision de prédiction.

Dès lors, maintenant que la vérification de la véracité des données est faite, place à la première prédiction utilisant toutes les données brutes contenues dans notre ensemble telles qu'elles ont été extraites.

Cette fois-ci, les résultats montre une légère amélioration dans les réponses obtenues. La précision de prédiction avec l'algorithme k-Nearest-Neighbors tourne maintenant autour des 55% et 60%. On peut donc déjà affirmer que, l'algorithme a réussi à créer un modèle de prédiction, même si celui-ci reste cependant faiblement performant.

L'échantillon de données d'entrée transmis à l'algorithme comportait toutes les informations contenues dans le box-score (voir la figure 4.1). Beaucoup de données assez proches et similaires constituent l'ensemble et certaines se sont potentiellement neutralisées au moment de générer le modèle. Cela a provoqué une situation d'"overfitting", et a conduit le modèle à trop se généraliser en capturant les fluctuations et variations aléatoires des données ainsi que le bruit qu'elles produisent.

À cet instant précis la composition de l'ensemble de données est de 67 variables,

36 données statistiques par équipe ainsi que la variable reprenant les résultats des rencontres. La lecture de différents articles spécialisés, notamment celui de Rory Burnak et Teo Susjna, *The application of machine learning techniques for predicting results in team sport* [39], complété par des passages du livre, *Mathletics* écrit par Wayne L. Winston [27], traitant de l'importance de certaines variables par rapport à d'autres entraînant une sélection plus précise et pertinente des variables transmises aux algorithmes de prédiction, l'échantillon de données a été affiné pour ne garder qu'au total 37 variables.

FIGURE 6.1 – *Tableau précision des modèles de prédiction, saison 2012/2013*

Modèle de prédiction	Précision	Nb de variables	Nb de matchs	Standardisation
kNN	63.61%	37	1314	Min/Max
kNN	61.83%	37	1314	zscore
SVM	66.16%	37	1314	Min/Max
SVM	65.14%	37	1314	zscore
Naïves Bayes	65.9%	37	1314	Min/Max
Naïves Bayes	65.9%	37	1314	zscore
Logistic Reg	67.18%	37	1314	Min/Max
Logistic Reg	65.14%	37	1314	zscore
Decision Tree	68.45%	37	1314	None
Random Forest	61.58%	37	1314	None

La figure 6.1, représente un récapitulatif des précisions obtenues pour chaque algorithme de prédiction manipulé. Tout d'abord, le même nombre de rencontres et de variables ont été transmises à tous les algorithmes. Ensuite, deux types de standardisation (transformation des données afin qu'elles soient toutes encodées de la même manière) ont été appliqués à seulement quatre des six algorithmes exploités.

La première standardisation affectée à l'ensemble des variables, identifiée par "Min/Max" dans le tableau de la figure 6.1, a déjà été détaillée dans une partie précédente. C'est la standardisation la plus classique que l'on peut retrouver de manière général. La secondes standardisation, est appelée "Zscore" ou encore "Cote

Z" en statistique. Elle correspond au nombre d'écarts types qu'il y a entre une variable et la moyenne de la colonne correspondante à la variable. La valeur des variables transformées par le "Zscore" peut être soit négative, soit positive. L'intérêt d'appliquer deux standardisations est de comparer l'impact de chacune d'elles sur les résultats affichés par les modèles.

Enfin, les modèles d'"arbre de décision" et de "forêt aléatoire" n'étant pas impacté par les différentes échelle de valeur que comportent l'ensemble de données, ils n'ont donc pas besoin de recevoir des variables transformées.

En regardant les résultats de plus près, une nette augmentation du pourcentage de prédiction est facilement perceptible. Une précision de prédiction qui oscille entre 61.58% pour la plus basse et 68.45% pour la plus haute. L'arbre de décision est le modèle le plus performant pour prédire les résultats d'une partie des rencontres de la saison 2012/2013. En revanche, la forêt aléatoire qui est en quelque sorte l'évolution des arbres de décision est celle qui donne la précision la plus faible.

Finalement sur les deux standardisations appliquées, on ne constate pas une réelle différence sur les résultats obtenus. Aucune des deux n'est plus efficace que l'autre, il s'agit de standardisations très similaires.

Après avoir des résultats plus que satisfaisants avec seulement les rencontres de la saison 2012/2013 de la NBA, la prédiction des résultats de la saison 2018/2019 avec comme ensemble de données de "training" les saisons allant de 2012/2013 à 2017/2018 est maintenant possible.

FIGURE 6.2 – *Tableau précision des modèles de prédiction, toutes saisons confondues*

Modèle de prédiction	Précision	Nb de variables	Nb de matchs	Standardisation
kNN	59.83%	37	9193	Min/Max
kNN	58.99%	37	9193	zscore
SVM	63.80%	37	9193	Min/Max
SVM	61.81%	37	9193	zscore
Naïves Bayes	61.28%	37	9193	Min/Max
Naïves Bayes	61.51%	37	9193	zscore
Logistic Reg	61.66%	37	9193	Min/Max
Logistic Reg	61.31%	37	9193	zscore
Decision Tree	62.80%	37	9193	None
Random Forest	62.73%	37	9193	None
XGBoost	62.67%	37	9193	None

Tout comme le tableau précédent, la figure 6.2 reprend l'ensemble des résultats avec les mêmes algorithmes de prédiction et les standardisations employées. Le changement porte non pas sur le nombre de variables composant l'ensemble de données mais sur le nombre de rencontres ayant servies pour entraîner le modèle. Ainsi, l'échantillon est dorénavant composé de 9193 matchs disputés de la saison 2012/2013 à la saison 2017/2018.

Cette fois, les résultats sont nettement moins significatifs que ceux précédemment révélés dans la figure 6.1. On constate une baisse de la précision pour chaque modèle de prédiction excepté la "forêt aléatoire" qui a été le modèle le plus performant pour prédire les résultats de la saison 2018/2019. Parallèlement, on constate que l'algorithme k-Nearest-Neighbors en passant sous le seuil des 60%, devient le moins performant.

Le modèle de "boosting" (XGBoost) obtient une précision de prédiction supérieure à tous les autres modèles. Cependant, il n'a pas réussi à surpasser les modèles "SVM" et "Random Forest".

Au final, aucune différence majeure est à noter en ce qui concerne les deux

standardisations utilisées, celles-ci donnant des réponses quasiment équivalentes.

6.2 Visualisation des résultats

La visualisation des données est une étape cruciale dans un travail d'analyse de données. L'enjeu de ce travail est la capacité à vulgariser au maximum les résultats obtenus afin de les rendre accessibles à des personnes n'ayant aucune connaissances techniques, et sans pour autant perdre de l'information. L'objectif est aussi de répondre aux besoins des personnes ciblées. Il n'existe pas de visualisation de données standard. Cela dépend du type d'informations que l'on souhaite vulgariser et des besoins du destinataire.

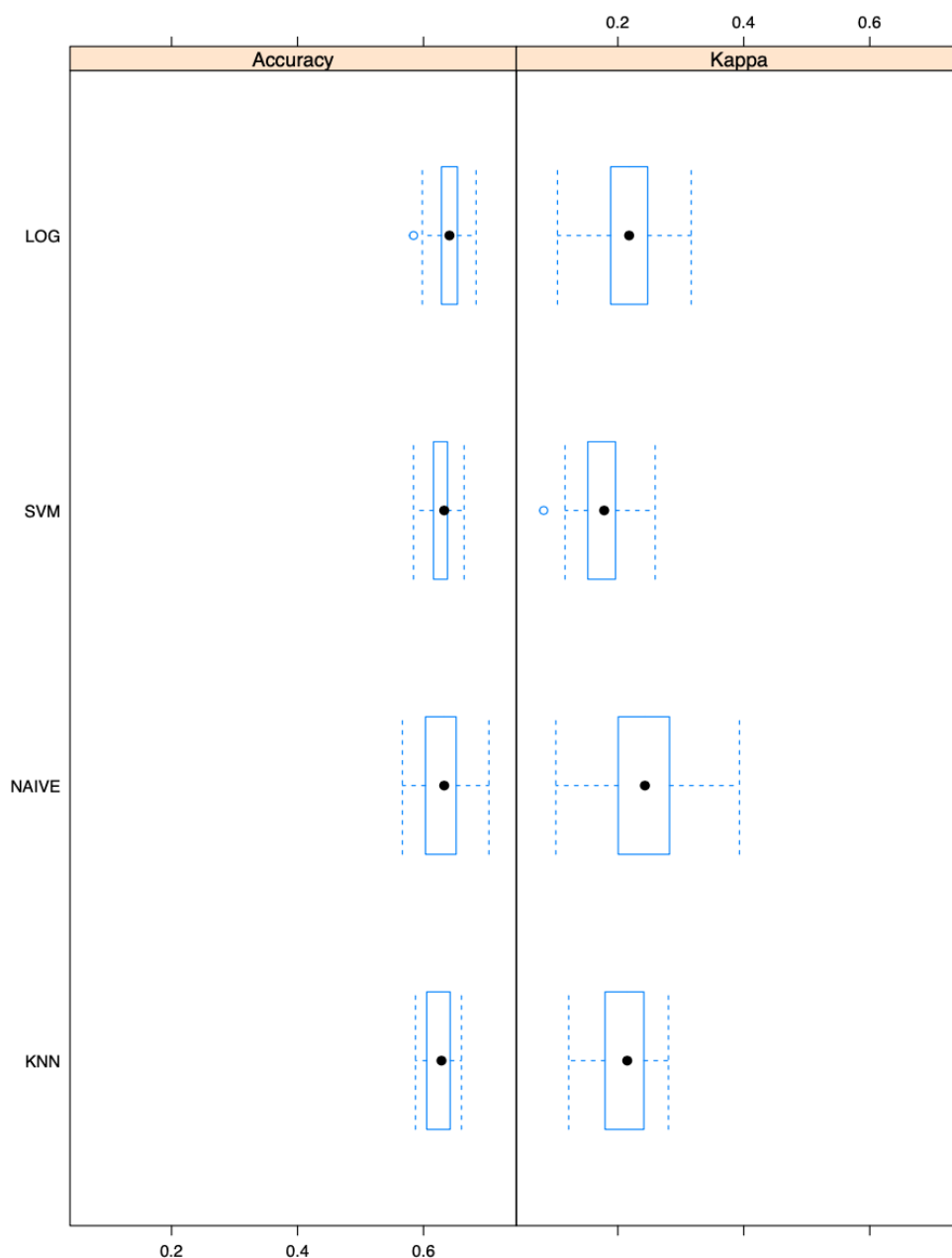
Dans le cadre du projet et en prenant en compte la position adoptée pour le réaliser, il s'agit de répondre aux potentielles questions que peut se poser un entraîneur sur les performances de son équipe en début ou même en cours de saison, afin d'adapter son coaching en fonction des résultats qui lui ont été présentés. La visualisation porte principalement sur les modèles de Machine Learning manipulés pour le travail ainsi que les résultats présentés en amont.

La première visualisation possible concerne la comparaison des résultats des modèles de prédiction. Il est ainsi possible de présenter les réponses données par les modèles en les comparant, en utilisant un diagramme "en boîte" que l'on connaît plus communément appelé "boîte à moustache", comme dans la figure 6.3 ci-après.

Ce diagramme reprend les résultats des modèles k-Nearest-Neighbors (KNN), Support Vector Machine (SVM), Naïves Bayes (NB) ainsi que celui de la Régression Logistique (LOG). L'intérêt de cette visualisation est principalement de pouvoir superposer plusieurs modèles et faciliter ainsi la comparaison en faisant apparaître et comprendre la situation en un seul coup d'oeil au destinataire. La "boîte à moustache" retranscrit très simplement la répartition des observations des modèles de prédiction. La valeur correspondant à la précision (accuracy) des modèles est observable sur la partie de gauche de la figure 6.3.

Les valeurs extrêmes des observations se situent de chaque côté de la boîte représentée sur la figure par les segments verticaux en pointillés. Les bords de la boîte (traits pleins) correspondent au premier quartile (25% des observations des variables) à gauche et le troisième quartile (75% des observations des variables) à droite. Le point noir au milieu représente la précision obtenue par le modèle, quasiment équivalente à celle présentée précédemment dans la figure 6.2.

FIGURE 6.3 – *Comparaison des modèles de prédiction*



Lorsqu'on regarde l'axe des abscisses, dont la graduation représente la précision de prédiction, on constate que tous les modèles ont une précision supérieur à 0.6, c'est-à-dire plus de 60% de prédictions correctes et que parmi eux, kNN est le modèle le plus performant.

Il est également possible de représenter les performances de modèles de classification d'une autre façon. La courbe ROC, "Receiver Operating Characteristic" ou "fonction d'efficacité du récepteur" en français, mesure les performances de classificateurs binaires, ayant pour objectif de catégoriser des éléments en deux classes, en l'occurrence dans le projet en victoire (WIN) et en défaite (LOSS).

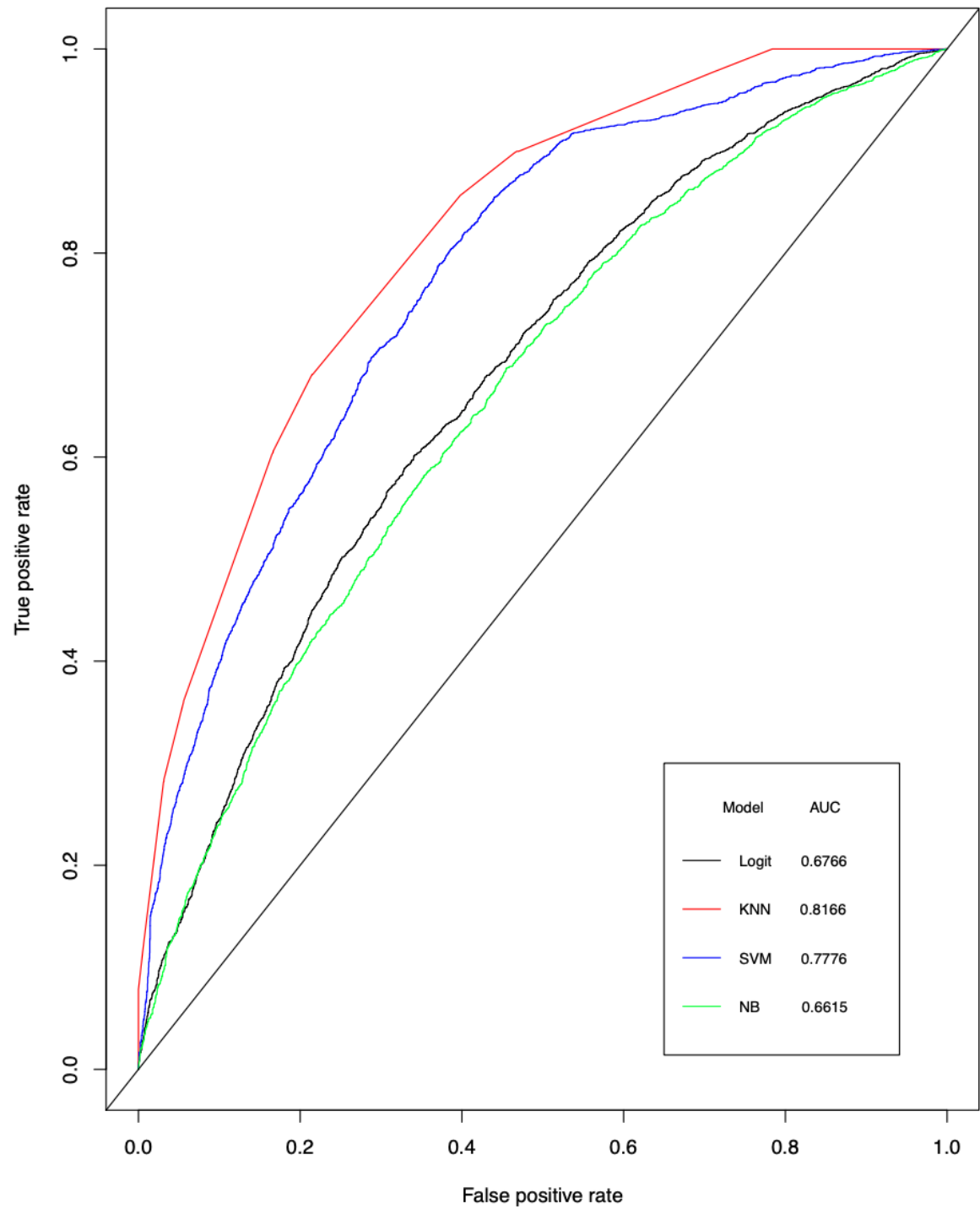
La figure 6.4, représente la courbe ROC des modèles de classification qui ont été appliqués dans le projet. Cette courbe trace le taux de vrais positifs en fonction du taux de faux négatifs. Le taux de vrais positifs est représenté sur l'axe des ordonnées de la courbe ROC. Il correspond à l'ensemble des bonnes prédictions, pour l'équipe visiteuse, que ce soit une victoire ou une défaite. L'axe des abscisses quant à lui, représente le taux de faux négatifs, et exprime l'ensemble des mauvaises prédictions, c'est-à-dire lorsque le modèle prédit une victoire de l'équipe visiteuse alors qu'en réalité celle-ci a perdu ou inversement.

Les axes ont pour minimum zéro et maximum un, la diagonale qui sépare l'espace en deux représente 0.5. Cette diagonale sert lorsque l'aire sous la courbe ROC est calculée. L'aire sous la courbe donne une mesure de la qualité d'un modèle, une aire égale à 1 représente une classification parfaite, une aire de 0.5 c'est-à-dire que la courbe qui suit la diagonale correspond à une classification sans valeur.

En observant, la courbe de chaque modèle utilisés, on peut facilement identifier qu'ils ont tous une classification positive allant au-delà de 0.5. En regardant la légende de la figure 6.4, l'aire sous la courbe y est renseigné. Sachant que plus l'aire est grande meilleur est le test. On constate que le modèle kNearest-Neighbors possède la plus grande aire sous la courbe qui est égale à 0.8166. Et que le modèle Naïves Bayes possède la plus petite avec 0.6615.

Don selon la courbe ROC évaluant et comparant les performances de classifications des modèles manipulés, kNN est désigné comme étant le plus efficace.

FIGURE 6.4 – *Courbe ROC des modèles de prédiction.*



Ensuite, un élément important de l'analyse de données est la compréhension de la corrélation qu'il peut y avoir entre les nombreuses informations qui composent l'ensemble des données exploitées. Différentes visualisations existent pour exprimer cette corrélation entre les variables.

La figure 6.5, est ce que l'on appelle une "heat map" exprimant la corrélation qu'il y a entre chaque variable que l'on retrouve dans l'ensemble de données utilisé pour le projet.

Le concept de "heat map" fait correspondre une gamme de couleurs par rapport à une taille variable, qui dans le cas présent se réfère au coefficient de corrélation entre deux variables.

La corrélation, explique une notion de liaison entre des données statistiques. Ce coefficient de corrélation est exprimé entre -1 et 1. On parle de corrélation parfaite positive lorsque le coefficient se trouve entre 0.9 et 1, et de corrélation parfaite négative quand sa valeur se situe entre -0.9 et -1. Elle est aussi considérée comme forte positive si le coefficient vaut entre 0.5 et 0.9, et forte négative si il est entre -0.5 et -0.9. À noter que toutes les corrélations ayant un coefficient supérieur à 0.3 ou inférieur à -0.3 sont considérées comme significatives.

Une corrélation entre deux variables est dite positive quand les valeurs de deux variables tendent à augmenter en parallèle. À contrario, une corrélation est qualifiée de négative quand les valeurs d'une variable diminuent pendant que celles de l'autre augmentent.

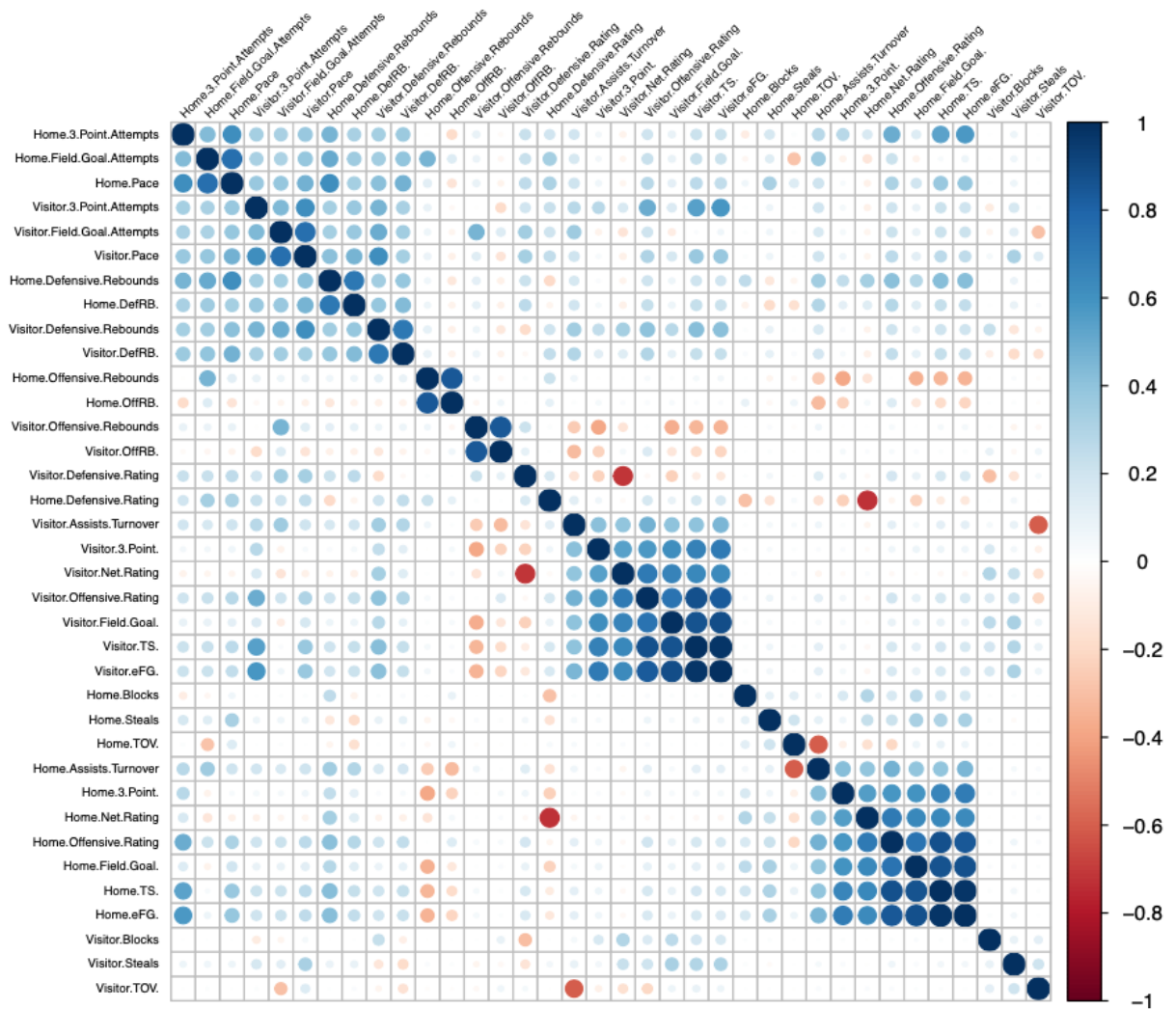
La "heat map" du projet fournit énormément d'informations. Elle est un indicateur important pour permettre la compréhension des réponses données par le biais des modèles de prédiction.

Premièrement, la diagonale de points bleu foncé, couleur exprimant une corrélation positive parfaite entre les variables, correspond à la corrélation entre ces mêmes variables ce qui donne obligatoirement un coefficient égal à 1.

Deuxièmement, on constate aisément grâce au code couleur, qu'il y a plus de corrélation positive, en bleu, que négative, en rouge.

Troisièmement, on peut voir que les variables de l'équipe visiteuse, identifiées par le préfix "Visitor" ont majoritairement tendance à avoir une forte corrélation entre elles. En revanche, la corrélation est faiblement positive ou négative voire même absente avec les variables appartenant à l'équipe jouant à domicile, variables identifiables avec le préfix "Home".

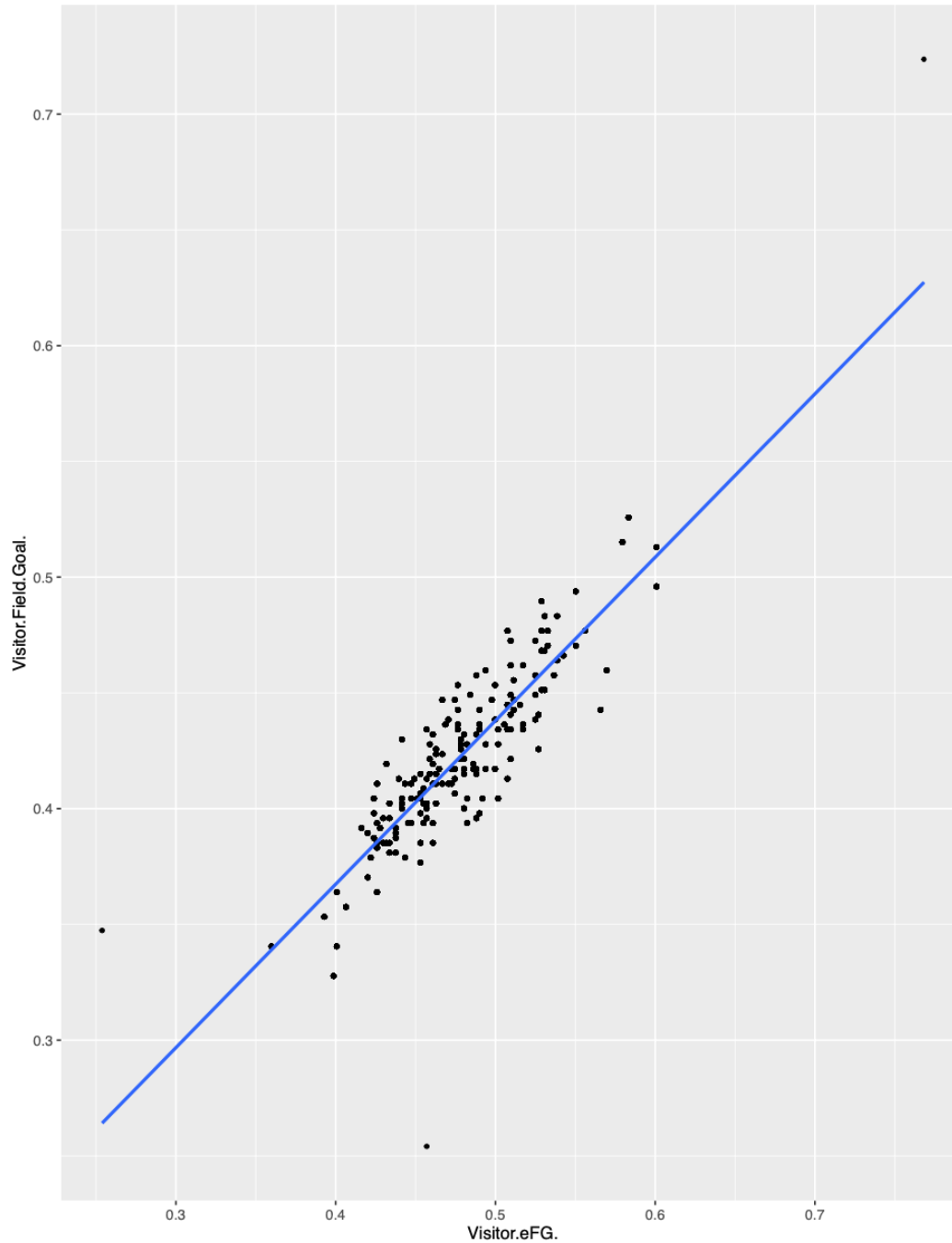
FIGURE 6.5 – *Heat Map.*



Il est aussi possible de visualiser la dite corrélation entre deux variables à l'aide d'un nuage de points permettant de mettre en évidence le degrés de corrélation qu'il y a entre elles. Par exemple, une corrélation positive existe entre les variables "Visitor.Field.Goal." et "Visitor.eFG.". Toutes deux sont déjà des données statistiques

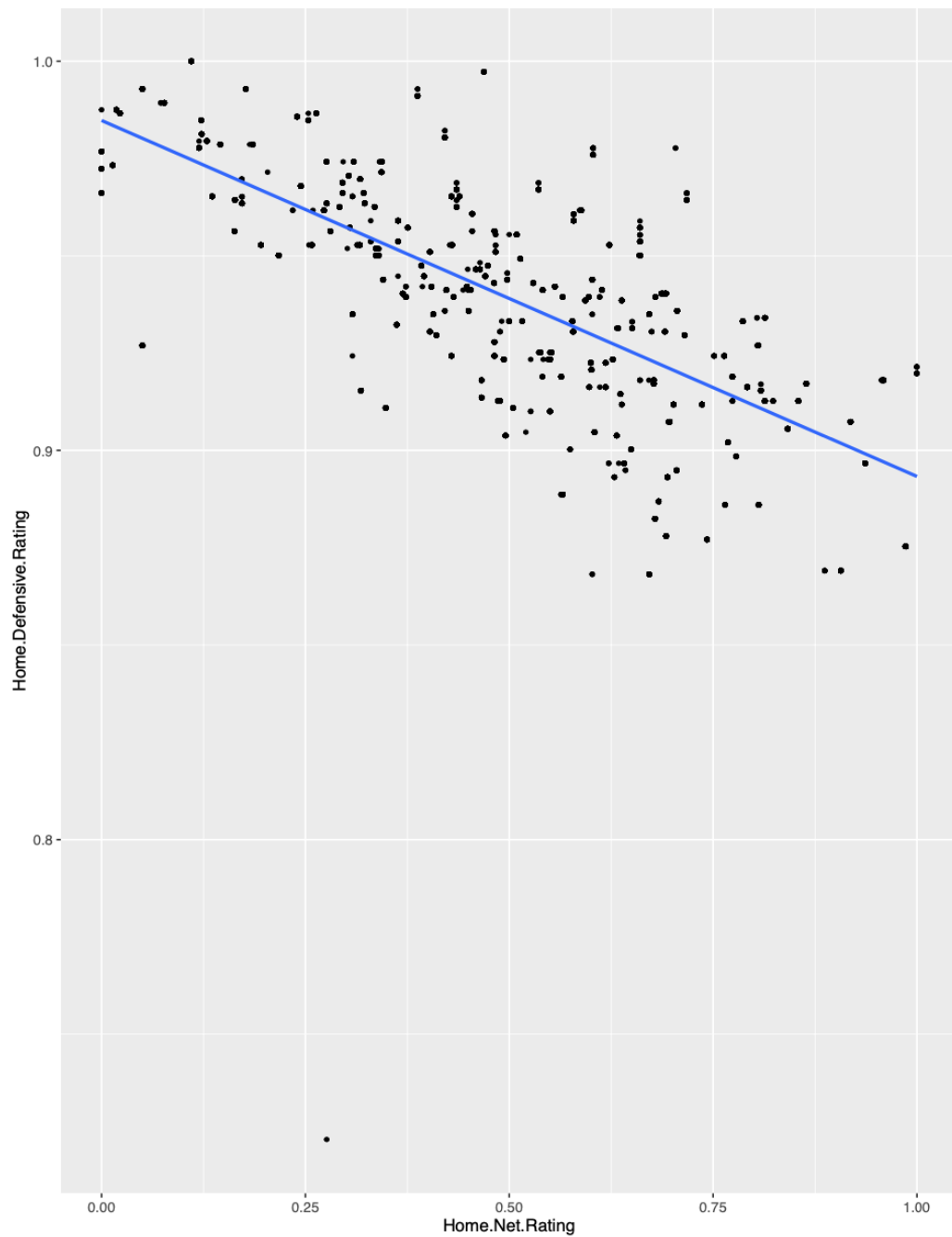
très proches. De plus, elles appartiennent à l'équipe visiteuse reconnaissable avec le préfix "Visitor". La corrélation est fortement positive, comme l'atteste la figure 6.6 ci-après. La droite de régression en bleu au vue de son sens, confirme bien cette tendance positive qu'il y a entre les deux variables. Les points quant à eux étant quasiment collés à la droite, expriment le degrés fort de positivité de cette corrélation.

FIGURE 6.6 – *Nuage de point représentant une corrélation positive.*



En considérant la figure 6.7 ci-dessous, avec les éléments évoqués plus tôt, il est facilement identifiable que ce nuage de points retranscrit une corrélation négative forte entre la variable "Home.Defensive.Rating" et la variable "Home.Net.Rating". Encore une fois, elles appartiennent à la même équipe mais elles expriment une corrélation négative entre elles comme l'indique le sens de la droite de régression en bleu.

FIGURE 6.7 – Nuage de point représentant une corrélation négative.



Enfin, la dernière de corrélation de corrélation n'a pas grand intérêt dans l'analyse qui est menée mais elle permet tout de même de voir comment se comporte deux variables qui n'ont aucune corrélation. La figure 8.4, représente la relation entre le nombre de contres effectués par l'équipe visiteuse (Visitor.Blocks) et le nombre d'interceptions faites par l'équipe évoluant à domicile (Home.Steals). La droite de régression en bleu est presque à l'horizontale signifiant une infime corrélation positive et lorsqu'on regarde la dispersion des données aucune tendance possible n'est identifiable et confirme bien la non-corrélation qui existe entre ces deux variables.

Concernant le modèle de la régression logistique, il est possible de visualiser les coefficients de régression de ce modèle. Cela permet d'identifier les variables qui ont le plus d'impact au sein du modèle de prédiction.

En regardant la figure 8.5, présente dans l'annexe, les similarités avec la "heat map" des coefficients de corrélation sont facilement identifiables. La ligne du milieu en pointillés représentant zéro est un coefficient que l'on peut qualifier de nul, donc non significatif. En bleu correspond le premier modèle de régression logistique généré avec les trente-sept variables initiales.

FIGURE 6.8 – *Importances des variables du modèle de régression logistique.*

Coefficients:				
	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-19.4610	10.3283	-1.884	0.05953 .
Visitor.Field.Goal.Attempts	-1.5007	6.7038	-0.224	0.82287
Visitor.Field.Goal.	-1.0653	12.6875	-0.084	0.93308
Visitor.3.Point.Attempts	0.7725	1.5507	0.498	0.61837
Visitor.3.Point.	9.2930	5.0702	1.833	0.06682 .
Visitor.Offensive.Rebounds	1.6162	1.4504	1.114	0.26514
Visitor.Defensive.Rebounds	0.6175	3.0370	0.203	0.83887
Visitor.Assists	-3.9380	2.1868	-1.801	0.07173 .
Visitor.Steals	2.9767	1.2620	2.359	0.01834 *
Visitor.Blocks	4.3051	0.9719	4.430	9.44e-06 ***
Visitor.Turnovers	0.8530	2.0478	0.417	0.67702
Visitor.Personal.Fouls	1.4452	1.5759	0.917	0.35914
Visitor.Points	3.0208	7.0341	0.429	0.66760
Home.Field.Goal.Attempts	17.6652	6.8412	2.582	0.00982 **
Home.Field.Goal.	28.9747	14.5334	1.994	0.04619 *
Home.3.Point.Attempts	-0.3359	1.5368	-0.219	0.82697
Home.3.Point.	-11.3464	5.8861	-1.928	0.05390 .
Home.Offensive.Rebounds	-4.0144	1.4751	-2.721	0.00650 **
Home.Defensive.Rebounds	2.8119	3.1943	0.880	0.37870
Home.Assists	0.2943	2.1563	0.136	0.89143
Home.Steals	-3.2395	1.2681	-2.555	0.01063 *
Home.Blocks	-4.1042	1.0267	-3.997	6.40e-05 ***
Home.Turnovers	4.4447	2.0343	2.185	0.02890 *
Home.Personal.Fouls	0.1379	1.6809	0.082	0.93460
Home.Points	-14.1646	7.0135	-2.020	0.04342 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

Après une première visualisation il a été convenu de ne garder que les variables les plus pertinentes, elles correspondent sur la figure 6.8 ci-dessus, aux variables ayant un point, une ou plusieurs étoiles visible sur la droite de la figure. Un second modèle a donc été créé. Les coefficients de régression des variables de ce nouveau modèle sont représentés en orange sur la figure 8.5. On constate que les variables du second modèle ont un coefficient plus important que celles du premier.

Pour autant, la précision de prédiction n'a pas augmenté bien au contraire. Le premier modèle affiche une précision de 61.66%, tandis que le second modèle qui est un réajustement du premier, donne 61.05%. Malgré un recentrage du second modèle avec seulement les données les plus significatives, on constate étrangement une moins bonne performance par rapport au premier modèle.

Il est important de s'attarder sur le modèle d'arbre de décision qui apporte un bon nombre d'informations pertinentes et qui peuvent être un atout majeur lorsqu'on se positionne en tant que coach d'une équipe ou d'un athlète.

Un élément qui peut être visualisé au sein d'un modèle d'arbre de décision, est l'importance que peuvent avoir les variables dans le modèle et l'influence qu'elles ont sur les réponses acquises. Tout comme la corrélation entre les variables qui donne un aperçu de l'homogénéité de l'ensemble des données exploitées, connaître l'importance que peuvent avoir les variables dans les prédictions des modèles permet d'avoir une vision interne de ce qu'il se passe et de comprendre comment les variables réagissent les unes vis-à-vis des autres.

La figure 8.2 met en avant l'importance de chacune des variables utilisées dans le modèle d'arbre de décision. Le graphique gauche de la figure, montre comment le modèle performerait si certaines variables étaient retirées du modèle. En se rappelant que les modèles ont pour objectif de prédire les résultats de l'équipe qui joue à l'extérieur, donc si on retire la variable la plus haute, correspondant à "Home.Net.Rating" qui est le différentiel de point entre l'attaque et la défense sur 100 possessions de l'équipe jouant à domicile, il est logique que la précision de prédiction diminue gran-

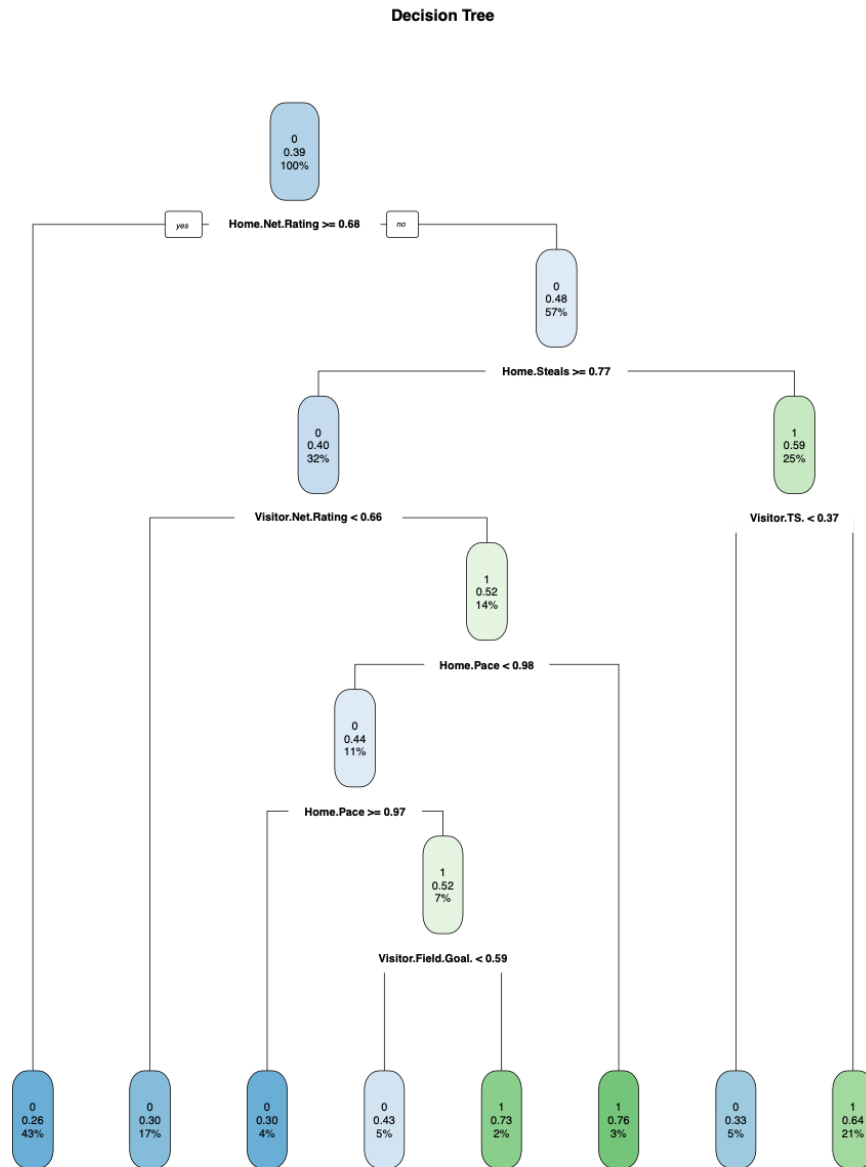
dement. C'est elle qui a le plus d'importance dans le résultat de la prédiction du modèle.

On peut voir que les douze premières variables se détachent des autres au niveau de leur importance sans pour autant qu'il y ait un écart énorme entre elles. Et parmi elles, il y a plus de variables appartenant à l'équipe jouant à domicile, puisqu'elles ont un impact fort sur les résultats de l'équipe visiteuse.

Le graphique de droite de la figure 8.2 quant à lui, représente la pureté de chaque nœud à la fin de l'arbre de décision. Contrairement au graphique de gauche, aucune variable ne sort réellement du lot. On note que c'est toujours la variable "Home.Net.Rating" qui possède le plus d'importance pour la pureté des nœuds mais si on décide de la retirer cela n'affectera que très peu la pureté étant donné que les autres variables détiennent une importance relativement proche de la sienne.

Enfin, la dernière visualisation concerne toujours l'arbre de décision, cette fois-ci il s'agit de rentrer plus en détails dans le modèle et comprendre la composition de l'arbre.

FIGURE 6.9 – *Arbre de décision.*



La figure 6.9 représente le détail de l'arbre de décision final. On peut observer le cheminement que fait le modèle pour arriver au résultat de ses prédictions. Chaque noeud donne plusieurs informations. La première est la classe prédite, dans le cas présent 0 pour une défaite de l'équipe visiteuse et 1 pour une victoire. La seconde correspond à la probabilité de prédire une victoire à la fin. Et la troisième représente le pourcentage d'observation dans le noeud.

Le premier noeud donne les indications suivantes : en utilisant 100% de l'en-

semble de données, la classe prédite est celle de la défaite (0) pour l'équipe jouant à l'extérieur et que la probabilité pour une équipe de gagner est de 0.39.

La variables "Home.Net.Rating" étant été identifiée comme ayant le plus d'importance dans le modèle de l'arbre de décision, il est logique de la retrouver comme première question de l'arbre.

L'arbre affirme que si la variable "Home.Net.Rating" détient une valeur supérieur ou égale à 0.68, le modèle prédit une défaite de l'équipe visiteuse. Cependant, si sa valeur est inférieur à 0.68 alors l'arbre passe à une nouvelle question.

Dans ce nouveau noeud, l'effectif utilisé est réduit quasiment de moitié et est désormais de 57%. La probabilité de prédire une victoire a augmenté passant à 0.48 mais la classe prédite reste toujours 0. La question de ce noeud prend en compte le nombre d'interceptions de l'équipe jouant à domicile (Home.Steals). Elle dit que si la valeur de cette variable est inférieur à 0.77, les chances de victoire de l'équipe visiteuse augmentent (57%). Ainsi le noeud qui suit montre qu'avec 25% des observations de l'ensemble de données, la classe prédite est celle de la victoire (1) avec une probabilité de victoire de 0.59.

La question de ce nouveau noeud s'attarde sur l'efficacité au tir de l'équipe visiteuse (Visitor.TS%). Si la valeur de la variable est inférieure à 0.37 le modèle prédit une défaite sinon il prédit une victoire.

En revenant sur le noeud précédent, si la valeur du nombre d'interceptions (Home.Steals) est supérieur ou égale à 0.77. Le noeud suivant réduit à nouveau le nombre d'observation de l'ensemble utilisé passant à 32%, la classe prédite est encore celle de la défaite et la probabilité d'avoir une victoire est de 0.40.

Le noeud suivant traite du différentiel de point des phases offensives et défensives de l'équipe visiteuse (Visitor.Net.Rating). Si la valeur de cette variable est inférieure à 0.66, alors une défaite est prédite par le modèle, sinon les possibilités de victoire de l'équipe visiteuse s'accroissent (0.52).

À l'étape suivante la variable "Home.Pace" correspond au nombre possessions jouées par l'équipe qui est à domicile. L'ensemble de données est réduit à 14%, avec une probabilité de prédire une victoire à 0.52. Lorsque cette variable est supérieur à 0.98, le modèle donne une victoire. Autrement il basculera sur un autre noeud en réduisant à nouveau l'effectif d'observations qui passe à 11%, la classe prédite est celle de la défaite puisque la probabilité d'avoir une victoire est cette fois-ci de 0.44. En utilisant toujours cette même variable (Home.Pace), avec une valeur supérieure ou égale à 0.97, la défaite est prédite. Dans le cas contraire on se dirige vers le dernier noeud de l'arbre de décision.

Cet ultime noeud estime que la probabilité d'avoir une victoire est de 0.52 en n'utilisant que 7% de l'ensemble de départ. La dernière question de l'arbre repose sur le nombre de tirs marqués par l'équipe visiteuse (Visitor.Field.Goal). Si sa valeur est supérieure ou égale à 0.59 alors la victoire est prédite par le modèle, autrement c'est la défaite.

Les graphiques permettent d'approfondir ce qui a été observé simplement en surface. Ils facilitent la déduction des raisons, des facteurs qui ont poussés les modèles à donner de tels résultats. Ils donnent aussi beaucoup d'éléments pour interpréter les résultats.

Le modèle forêt aléatoire n'a pas réussi à obtenir une meilleure précision de prédiction que le modèle d'arbre de décision. Pourtant, la forêt aléatoire est en quelque sorte l'évolution du l'arbre de décision, étant donné qu'elle génère un grand nombre d'arbre pour ensuite récupérer et combiner les résultats de chaque arbre pour en déterminer une solution finale.

La forêt aléatoire pour arriver à une précision de prédiction de 62.73, comme indiqué sur la figure 6.2, a généré un total de 2000 arbre de décision. La figure 8.6 représente un histogramme mettant en avant la fréquence du nombre de noeud au sein des arbres de décision de la forêt aléatoire. On constate que certains arbres sont

composés de moins de dix noeuds tandis que certains détiennent un peu moins de 250 noeuds.

Malgré, une performance plus faible que celle de l'arbre de décision, la forêt aléatoire donne la possibilité de visualiser la valeur marginale d'une variable, c'est-à-dire, la variation d'une valeur qui est vraie compte tenu de contraintes particulières, en l'occurrence le fait pour l'équipe visiteuse de gagner (WIN) ou de perdre (LOSS).

L'intérêt est de comprendre comment la variation des variables "Home.Net.Rating" et "Visitor.Net.Rating" influe sur la prédiction soit de la classe "WIN" soit de la classe "LOSS". On peut voir sur la figure 8.7, présente en annexe, que lorsque la valeur de la variable "Home.Net.Rating" est supérieure à 5, il y a de grande chance pour que le modèle prédise la défaite de l'équipe visiteuse. Sur la figure 8.8, on constate que si la valeur de la variable "Visitor.Net.Rating" est supérieur à 8 le modèle aura plus tendance à prédire une victoire pour l'équipe visiteuse.

6.3 Interprétation des résultats

Finalement, en regardant les résultats obtenus par le biais du projet réalisé et présenté en amont dans le travail, et que l'on décide de les comparer à ceux qui ont été recensés par Rory Bunker et Teo Susnjak dans leur travail, *The application of machine learning techniques for predicting results in team sport : A review*, visibles juste en dessous dans la figure 6.10, on constate que les études ayant exploité les mêmes modèles de Machine Learning que ceux présent dans le projet, affichent des résultats proches mais légèrement supérieur.

FIGURE 6.10 – *Études de Machine Learning sur le Basketball.*

Paper	Competition	Models used	No. of features	No. of Matches	Accuracy of best model
Loeffelholz et al. (2009)	NBA	ANN (types: FFNN*, RBG, PNN, GRNN, fusions of these)	4	650	74.3%
Zdravevski & Kulakov (2009)	NBA	All models in WEKA (Logistic Regression*)	10	1230	72.8%
Ivanković et al. (2010)	Serbian First B	ANN trained with BP	51	890	81%
Miljković et al. (2010)	NBA	kNN, Decision Tree, SVM, Naïve Bayes*	32	778	67%
Cao (2012)	NBA	Simple Logistic Regression*, Naïve Bayes, SVM, ANN	46	4000	67.8%
Shi et al. (2013)	NCAAB	ANN*, C4.5 Decision Tree, RIPPER, Random Forest	7	32236	74%
Thabtah et al. (2019)	NBA	ANN, Naïve Bayes, LMT Decision Tree*	8	430	83%

source by : *Bunker (Rory) and Susnjak (Teo). The application of machine learning techniques for predicting results in team sport : A review (pp.5).*

Les lignes qui se rapportent le plus au projet réalisé sont les lignes 4 (Cao) et 5 (Shi et al.) de la figure. On observe que le nombre de matchs sélectionnés (778 et 4000) est bien moins important que celui de l'ensemble de données correspondant à toutes la saisons de NBA allant de la saison 2012/213 à la saison 2018/2019.

Lorsque l'on reprend le passage du projet où les modèles de Machine Learning ont été exploités uniquement pour la saison 2012/2013, le nombre de matchs (1314) se situait entre celui de la ligne 4 (700) et 5 (4000) de la figure. On constate que les résultats obtenus sont quasiment identiques à ceux indiqués dans la figure 6.1. Pour l'étude finale, le nombre de match constituant l'ensemble des données étudiées a été presque multiplié par dix. Ainsi, les données de l'échantillon de "training" sont passées de 1314 (matchs) à 9193 (matchs).

C'est est une des raisons pour lesquelles la précision de prédiction a diminué. Les modèles ont récupéré et assimilé toutes les variations, fluctuations aléatoires des données transmises, provoquant alors une situation d'"overfitting".

Parmi les différentes études présentes dans la figure 6.10, une seule utilise un

nombre plus élevé de match que dans le projet (32236). Le modèle ANN, manipulé dans cette étude est plus orienté Deep Learning et ne donc ne correspond pas au parti pris en début de travail de se concentrer sur des modèles de Machine Learning.

Si l'on observe maintenant le nombre de variables utilisées, on s'aperçoit sur la figure qu'il n'y a pas d'impact significatif en fonction du nombre de variables.

De plus, les informations fournies par la "heat map" (figure 6.5) sur la corrélation entre les variables, combinées à celles exprimées par la figure 6.10, permettent de mettre en exergue un soucis quant à la composition de l'ensemble de données. En effet, seules les variables appartenant à une même équipe possèdent des corrélations entre elles. (négatives, positives ou nulles).

On peut en déduire que les variables d'une équipe neutralisent l'influence que peuvent avoir celles de l'autre équipe.

De ce constat peut se poser aussi la question de la standardisation qui a été appliqué sur l'ensemble de données. Toutes les données ont été standardisées pour que certains modèles soient en mesure de les manipuler correctement. Malheureusement, les données n'étant initialement sur la même échelle, la standardisation a provoqué une perte de leur signification en leur donnant des valeurs quasi similaires alors qu'à la base elles reflétaient chacune réalité bien différentes.

Les variables qui sont le plus mises en avant par les graphiques, représentent une situation globale de l'équipe et non des faits de jeu en particulier pendant une rencontre. La variable "Home.Net.Rating", qui considérée par la plupart des modèles comme étant la plus importante. Elle reflète l'efficacité d'une équipe en attaque et en défense. C'est une donnée qui a une valeur primordiale pour un coach, car elle renseigne sur le niveau général d'une équipe et permet donc de déduire celle qui a le plus de chance de gagner.

Ce constat, est repris dans la figure 8.2. Il confirme que dans les six premières variables, définies comme les plus importantes dans la prédiction de résultats font encore référence à une situation globale de l'équipe et pas uniquement à des faits de

jeu précis d'une rencontre. On fait à nouveau ce même constat dans la figure 6.9, avec les mêmes variables présentent dans les noeuds du modèle d'arbre de décision.

Après avoir regardé en détail les résultats obtenus grâce à la visualisation, le constat correspond parfaitement avec ce qui a été déjà expliqué plus haut dans le texte : les résultats produits par les modèles dépendent entièrement des données transmises.

Dans l'étude, l'inconvénient premier de l'ensemble de données qui a été exploité est sa composition. Le choix de prendre les informations à la fois de l'équipe visiteuse et celles jouant à domicile peut paraître logique mais lorsqu'on regarde ce qu'expriment les différentes visualisations, les variables se neutralisent entre elles. Même problématique également lorsque le nombre de données, en l'occurrence dans le cas présent le nombre de matchs de chaque saison, est grand. Avec un nombre de matchs réduit on a remarqué que les modèles avaient plus de facilité pour prédire le résultat des matchs.

En sachant que dans le domaine sport, lorsque le moment est venu de préparer la nouvelle saison, les dirigeants, les entraîneurs se basent uniquement sur la saison précédente. Donc prédire les résultats de la saison 2018/2019 en s'appuyant sur les saisons des années précédentes en remontant jusqu'à 2012/2013, c'est avéré ne pas avoir une très grand intérêt, au contraire. Une des possibilités aurait pu être d'utiliser seulement les données des saisons 2016/2017 et 2017/2018 pour entraîner les modèles, permettant ainsi d'éviter au maximum les problèmes d'"overfitting" et de pousser davantage l'exploitation des modèles de Machine Learning.

Les informations statistiques du Basket-Ball sont aussi très proches d'une équipe à une autre, que ce soit dans le haut ou bien le bas du classement de ligue nationale américaine de Basket-Ball. Cette proximité entre chaque observation rend la tâche des modèles de Machine Learning plus compliquée pour qu'une variables se détache des autres. Ajoutée à cela une standardisation qui, en mettant toutes les variables sur la même échelle sans prendre en compte les spécificité de chaque variables, pro-

voque un flou qui alimente, renforce les possibilités d'avoir de l'"overfitting".

De même, la visualisation des résultats a permis de mieux comprendre les réponses transmises par les modèles d'apprentissage automatique. On a pu ainsi facilement comprendre que, lorsqu'une équipe jouant à domicile, détenait le meilleur "Home.Net.Rating" (estimation du différentiel de points entre l'attaque et la défense sur 100 possessions), celle-ci était susceptible de gagner la rencontre.

Néanmoins, en reprenant la position initialement adoptée, c'est-à-dire celle de l'entraîneur, les informations données par les modèles de Machine Learning sont difficilement exploitables. La seule décision de l'entraîneur susceptible d'être influencée par les résultats donnés, serait de concentrer son coaching sur la réduction des possessions de l'équipe adverses pour être en phase avec la variable "Home.Net.Rating".

En effet pour un entraîneur, compte tenu de son rôle dans une rencontre, ce sont les variables correspondant aux faits de jeu qui sont les plus pertinentes.

Son besoin est de mieux comprendre ce qui se déroule durant le match, d'en connaître les faits marquants lors d'une victoire ou bien lors d'une défaite.

C'est dans cette optique, que l'analyse en amont des statistiques a été décidée. L'ensemble de données a été reconstruit avec seulement les faits de jeu (playmaking, phases offensives et défensives), à savoir les passes décisives, les interceptions, les tirs, les rebonds, etc. .

En utilisant les mêmes techniques de visualisation, les informations mises en évidence collent parfaitement aux tâches quotidiennes du coach.

La figure 6.11 suivante, est un récapitulatif des résultats obtenus après avoir exploité le nouvel ensemble de données plus adapté aux attentes précises de l'entraîneur. Reprenant les variables des saisons 2016/2017 et 2017/2018 toujours dans l'objectif de prédire les résultats de la saison 2018/2019.

FIGURE 6.11 – *Tableau précision des modèles. Nouvel ensemble de données*

Modèle de prédiction	Précision	Nb de variables	Nb de matchs	Standardisation
kNN	55.72%	33	3933	Min/Max
SVM	63.26%	33	3933	Min/Max
Naïves Bayes	56.94%	33	3933	Min/Max
Logistic Reg	62.04%	33	3933	Min/Max
Decision Tree	62.42%	33	3933	None
Random Forest	60.98%	33	3933	None
XGBoost	61.10%	33	3933	None

On constate assez facilement que pour certains modèles, notamment les modèles kNN et Naïves Bayes ayant respectivement une précision de prédiction bien inférieure à 60%, que le nouvel ensemble de données n'est pas suffisant pour espérer obtenir de bonnes performances.

Dans la continuité de ce tableau, lorsque l'on compare les deux "Heat Map" (figures 6.5 et 8.1) de l'annexe, on détecte rapidement qu'au sein de ce nouvel ensemble de données, les variables n'ont que très peu de corrélation. Il est clair que la majorité de faible corrélation voir de corrélation nulle est un des facteurs qui engendre une précision basse chez certains modèles.

L'intérêt d'avoir recentré l'ensemble de données sur les variables consacrées au jeu du Basket-Ball, n'est pas dans l'optique d'obtenir de meilleurs résultats de prédiction puisqu'on vient de voir que ce n'est pas le cas mais d'apporter des réponses à un entraîneur en lui donnant des clés lui permettant d'orienter ses tactiques, ses stratégies de jeu.

À nouveau c'est le graphique représentant le modèle de l'arbre de décision qui apporte le plus d'informations. Le but n'est pas de rentrer dans le détail, simplement de s'intéresser aux questions que s'est posées le modèle.

Les variables les plus importantes sont désormais les variables défensives (figure 8.3). La variable "Visitor.Blocks", correspondant au nombre de contres effectués par

l'équipe visiteuse, se trouve dans le premier noeud. Si la valeur de celle ci est supérieure ou égale à 6, le modèle dit que selon lui l'équipe visiteuse fait passer ses chances de victoire de 0.41 au départ à 0.69. C'est elle qui conditionne dès le départ les possibilités de victoire de l'équipe. Sur la figure 6.12 ci-après, qui représente l'importance des variables dans la génération du modèle, on retrouve parfaitement ce qui a été identifié juste avant. Mais avec le nouvel ensemble de données ce sont les informations statistiques défensives qui ont le plus d'importance sur la décision finale prise par le modèle d'arbre de décision. Dans les statistiques défensives on retrouve en haut du tableau le nombre de contres, d'interceptions, de fautes commises, etc..

Pour autant les statistiques d'attaque gardent une importance capitale, le principe même du Basket-Ball étant de marquer plus de point que l'équipe adverse.

C'est parfaitement ce que démontre la figure 8.9 de l'annexe, avec l'application du modèle de "XGBoost" sur le nouvel ensemble de données. Ce modèle avait pour objectif principal d'améliorer le performance des modèles les plus "faibles", ceux affichant les performances les plus basses. Malheureusement, il n'a pas réussi à outre passer l'efficacité des modèles SVM, d'arbre de décision et de régression logistique.

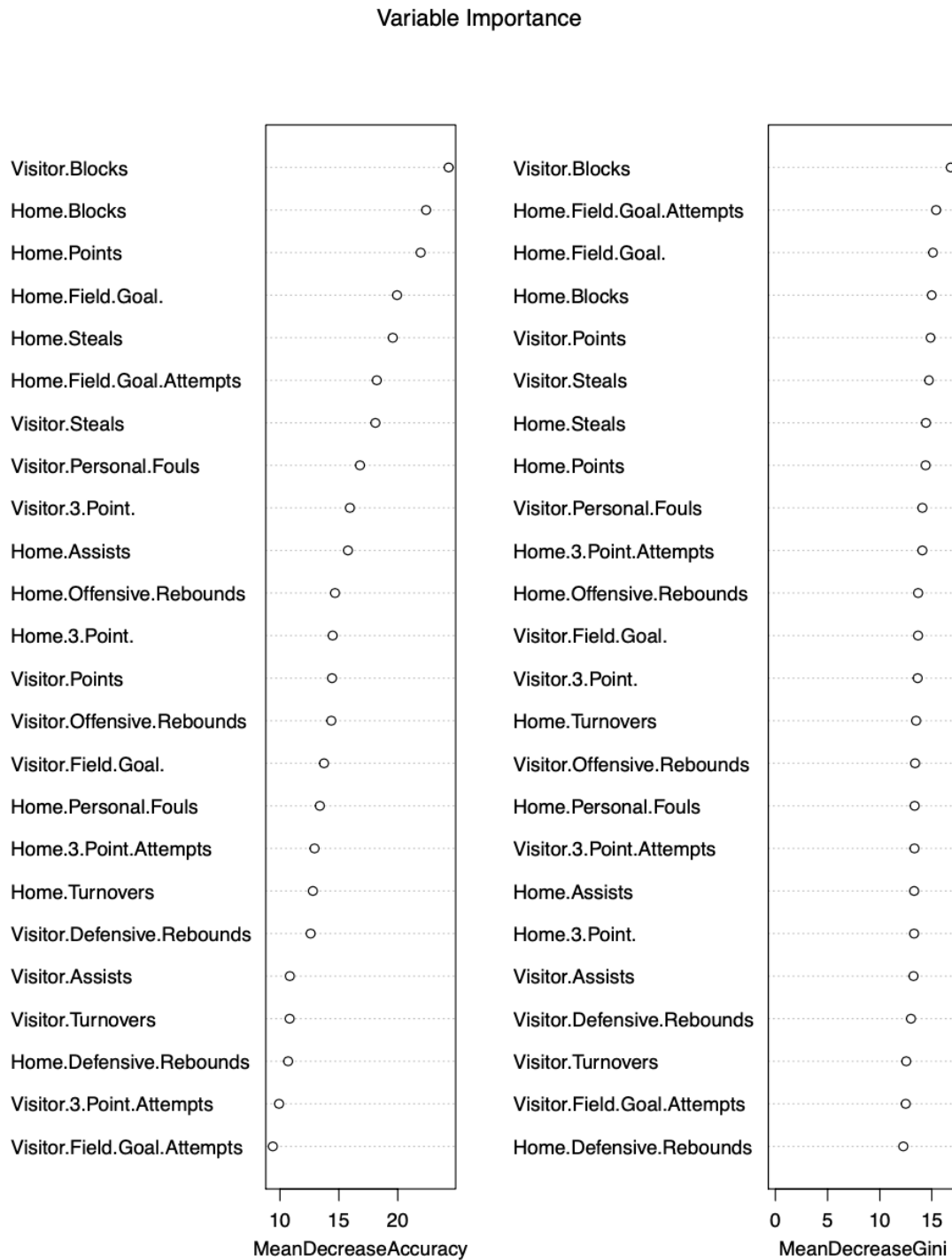
Cependant, grâce à la figure 8.9, on peut constater que l'importance des variables a changé après le travail de du modèle de "boosting". Les variables défensives se sont effacées pour laisser place aux variables offensives. Les deux variable qui se détachent des autres sont le nombre de 3-points tentés par match par l'équipe visiteuse et l'équipe jouant à domicile ("Visitor.3.Points.Attempts" et "Home.3.Points.Attempts").

Cela traduit bien le phénomène exprimé juste avant. Afin de remporter la victoire, il est plus judicieux d'éviter de jouer trop près du panier pour ne pas voir ses attaques stoppées par l'équipe et par conséquent comme le montre la figure 8.9, de se concentrer sur les tirs longue distance (3-points) qui augmentent les chances de victoire de l'équipe visiteuse.

À défaut d'avoir des prédictions performantes, la restructuration de l'ensemble

donne accès à l'entraîneur à beaucoup plus d'informations essentielles qu'auparavant. Il est désormais capable d'adapter plus facilement le jeu de son équipe avec le résultat des analyses venant d'être effectuées.

FIGURE 6.12 – *Importance des variables recentré sur les variables de jeu.*



7. Conclusion

Dans ce mémoire, l'objectif était de mettre en lumière le lien très étroit qui existe entre le monde du sport et l'analyse et la visualisation des données, à travers de techniques d'apprentissage automatique très en vogue en ce moment.

La réalisation du projet a permis de mettre en évidence concrètement quels sont les enjeux liés à ces deux domaines. Le défi principal, a porté sur l'exploration et l'identification des informations les plus appropriées et pertinentes dans l'optique de constituer un ensemble de données permettant de retranscrire le plus finement et fidèlement la conformité des observations effectuées et que l'on souhaite analyser. Un travail qui peut sembler long et fastidieux au premier abord mais qui est la pierre angulaire de tout le projet.

On a constaté pendant la phase d'exploitation de l'ensemble de données à travers les algorithmes et modèles de Machine Learning, qu'un grand nombre de données collectées et manipulées n'est pas nécessairement et forcément gage de bonne qualité, bien au contraire. Par exemple, l'ambition de vouloir prédire une saison entière, en l'occurrence celle de 2018/2019 en utilisant les statistiques des six saisons précédentes a fait apparaître des problèmes d'"overfitting", ce soucis est l'ennemi premier de tout projet portant sur de l'analyse de données. En comparant avec d'autres études traitant de près ou de loin du même sujet, dans la majorité des travaux le nombre de données n'excède jamais 4000 afin de réduire au maximum le risque d'"overfitting".

Outre le nombre, la qualité des données est primordiale. Il est donc impératif de faire en sorte que les données entrées transmises soient bien adaptées aux caractéristiques des algorithmes d'apprentissage automatique. Certains ont besoin de recevoir

des données standardisées afin d'être le plus performant possible. Ainsi les informations transmises peuvent en fonction du type d'algorithme manipulé.

Toutes ces étapes, servent à mettre les algorithmes d'apprentissage automatique dans les meilleures conditions afin qu'ils se rapprochent le plus possible des résultats espérés.

Le Machine Learning et le Deep Learning sont devenues aujourd'hui la norme lorsque l'on parle d'intelligence artificielle, d'analyse de données et de prédiction. Le Deep Learning est une évolution du Machine Learning et tend petit à petit à le suppléer. Ses performances, la profondeur de son analyse, ses facultés d'adaptation dépassent largement les capacités proposées par le Machine Learning et également celles de l'être humain à certains moments.

Pour autant, la complexité du fonctionnement interne du Deep Learning fait que dans le cadre du projet qui a été réalisé, il n'aurait pas pu apporter le même résultats que les modèles dits classiques de Machine Learning. La "simplicité" du Machine Learning par rapport au Deep Learning, a permis une meilleure compréhension et visualisation des résultats du travail effectué. Malgré ce qu'en pensent certains auteurs et chercheurs, le Machine Learning n'est pas prêt de s'effacer au profit du seul Deep Learning. Les moyens les plus complexes et les plus performants ne sont pas toujours les mieux adaptés pour à répondre à toutes les problématiques.

Dans une optique de performance des modèles de Machine Learning et afin d'obtenir la précision de prédiction la plus élevée possible, il a été formé un ensemble de données le plus efficace et le plus précis possible. Néanmoins lors de l'analyse et de l'interprétation des réponses données par les différents modèles, il c'est avéré que la précision attendue n'était pas toujours au rendez-vous mais surtout que ces résultats ne pouvaient pas être directement exploités par l'entraîneur d'une équipe pour y baser son coaching dessus.

Le projet s'est éloigné de son but premier, qui était initialement de comprendre de quelle manière la visualisation de données appliquée à des modèles de Machine Learning pouvait impacter le coaching sportif. Après les premières prédictions et visualisations des résultats, on aurait pu en conclure uniquement que cela n'apportait aucune information et plus-value utilisables pour un entraîneur d'une équipe de Basket-Ball.

En effet, on constate que l'ensemble de données manipulées est destiné uniquement aux prédictions de résultats des matchs (victoire ou défaite) et que malgré une bonne précision de prédiction, les réponses obtenues ne sont pas forcément liées entre elles.

La décision de mettre de côté les bonnes prédictions et de se concentrer sur un ensemble de données uniquement composé de variables qui permettent de retranscrire uniquement l'ensemble des faits de jeu pendant une rencontre a permis d'obtenir des résultats bien plus utiles et exploitables pour le coaching.

On a pu s'apercevoir dans l'interprétation des données que les variables les plus importantes correspondaient aux phases défensives des deux équipes. Pour un entraîneur, cette information apportée par la visualisation des données, est un avantage indéniable dans son coaching. Il sera en mesure grâce à l'analyse des différents graphiques de mettre en place de nouvelles tactiques et stratégies pour les matchs à venir.

À l'évidence les résultats obtenus traduisent parfaitement une évolution du jeu constatée depuis quelques années en NBA. Une des variables les plus probantes est celle qui exprime le nombre de contres qu'une équipe effectue par match. On peut en déduire que pour éviter que ses joueurs se fassent contrer, l'entraîneur leur demande de jouer de plus loin du cercle (panier), d'éviter de trop se rapprocher de la raquette, pour privilégier les tirs à mi-distance (2-points) et longue distance (3-points).

Ainsi, c'est précisément ce que l'entraîneur des Houston Rockets de l'époque,

Mike D'antoni, a mis en place comme tactique de 2016 à 2020. On appelle ce phénomène le "Small Ball".

Ce style de jeu, s'appuie sur le renoncement des joueurs de grande taille qui pourtant apportent à la fois de la force physique et de la présence sous le panier mais au détriment de la vitesse de jeu, au profit de joueur de plus petite taille. Par ailleurs, les joueurs de plus petite taille sont privilégiés pour leur vitesse de jeu, leur agilité et leur précision aux tirs à moyenne et longue distance.

L'idée étant d'utiliser davantage la vivacité des joueurs et la précision des tireurs longue distance (3-points) sur le terrain, plutôt que de favoriser un jeu placé au poste, ou sous le panier.

Néanmoins ce style de jeu à des avantages mais aussi des inconvénients. Si une équipe fait le choix du "Small Ball", cela laisse beaucoup plus d'espace sous le panier aux joueurs adverses de grande taille, avec une meilleure efficacité dans les rebonds au si bien défensifs qu'offensifs. L'efficacité de ce style dépend totalement de l'adresse et de la réussite aux tirs des joueurs, ce qui restreint les chances de victoire en cas de réussite médiocre.

L'adoption de ce mode de jeu n'a pas simplement impacté les équipes qui l'utilisent, elle a aussi redéfini le rôle et l'importance de certains postes, notamment pour les joueurs de grande taille évoluant au poste habituellement soit au poste de pivot, sous le panier, soit au poste d'allier fort. Historiquement, les joueurs évoluant à ces postes étaient de grande taille et athlétiques afin de résister au combat sous le panier et marquer plus facilement. Ces joueurs n'ont jamais été catégorisés comme étant ni rapides, ni agiles, ni habiles avec leurs mains, ni bons tireurs. Ces qualités étaient dévolues aux joueurs de plus petite taille.

L'arrivée du "Small Ball" dans la ligue nationale américaine de Basket-Ball, a amené les joueurs de grande tailles, à travailler et à améliorer leur mobilité, leurs tirs mi-distance et longue distance ainsi que leur qualité de dribble. Aujourd'hui, le poste de pivot a évolué vers un style plus proches du jeu des joueurs de plus petite taille.

À l'heure actuelle, les prétendants au titre de meilleur joueur de la NBA pour la saison 2020/2021, sont des joueurs de grandes tailles avec un profil atypique, qui sont capables de faire des différences au niveau offensif avec leur capacité à délivrer des passes décisives, marquer des paniers de n'importe quel endroit du terrain, mais aussi au niveau défensif, leur taille leur permettant de récupérer plus facilement des rebonds et de dissuader ou d'empêcher l'équipe adverse de marquer en réalisant des contres ou encore des interceptions.

Il est fort probable que les analystes travaillant pour l'équipe des Houston Rockets, aient effectué à un autre niveau évidemment une étude similaire à celle qui a été réalisée dans ce mémoire, à la demande de l'entraîneur de l'équipe. Il en a découlé l'application du "Small Ball", visant à augmenter leurs chances de victoire. Pendant cette période l'équipe s'est classée deuxième de sa division la première année puis première de sa division durant les trois années suivantes. Elle a été classée cinquième sur l'ensemble des équipes de la NBA.

Les modèles de Machine Learning et même de Deep Learning sont des technologies qui permettent d'exploiter pleinement le potentiel de ressources des données par une analyse croisée. Selon les besoins exprimés, ils apportent une nouvelle vision du sport, plus profonde et plus technique que ce qu'une rencontre de Basket-Ball peut apporter par l'intermédiaire des informations statistiques. La prédiction de résultats avant le début d'une saison permet d'avoir un aperçu du déroulement éventuel que peut avoir la saison selon l'équipe considérée.

En revanche, prédire les résultats (victoire ou défaite) n'est pas d'une grande utilité dans le travail préparatoire de l'entraîneur. Il serait plus judicieux de prédire l'évolution des performances de l'équipe au fur et à mesure de la saison afin de mettre en évidence les points forts et les faibles de l'équipe et donner à l'entraîneur les orientations sur lesquels son équipe doit travailler pour progresser et ainsi être plus efficace.

Vulgariser les données brutes reste une valeur sûre pour le moment. Ce type de

visualisation permet de compléter celle des modèles de prédiction, puisque qu'elle donne la possibilité de suivre l'évolution dans le temps, des performances d'une équipe ou des athlètes.

Néanmoins, l'univers du Basket-Ball et plus globalement du sport est en train de changer. Les algorithmes d'apprentissage automatique sont déjà largement présents dans la partie médicale du sport pour prévenir des blessures, mais pas seulement. Ainsi aujourd'hui, la remontée des données physiques des joueurs permet à l'entraîneur de connaître presque en temps réel, l'état de forme de chacun de ses joueurs sur le terrain.

L'intérêt de cette analyse "en direct" offre la possibilité au coach d'effectuer les changements de joueurs au bon moment afin de conserver niveau constant d'efficacité et de performance de son équipe sur le terrain. La composition de l'effectif total d'une équipe en découle.

Le monde des paris sportif qui est déjà sous l'influence de la prédiction de résultats utilise des techniques de Deep Learning.

Le coaching quant à lui n'est pas entièrement dépendant de ces techniques, même si certaines tendances de jeu ("Small Ball") sont le fruit d'analyse utilisant du Machine Learning.

Au final, le sport est un domaine où certains aspects restent imprévisibles. Rien ne permet à ce jour, de connaître avec assurance le résultat d'un match ou d'une rencontre. Cette part d'incertitude fait partie de la philosophie et de la magie du sport.

8. Annexes

FIGURE 8.1 – *Heat Map recentré sur les variables de jeu.*

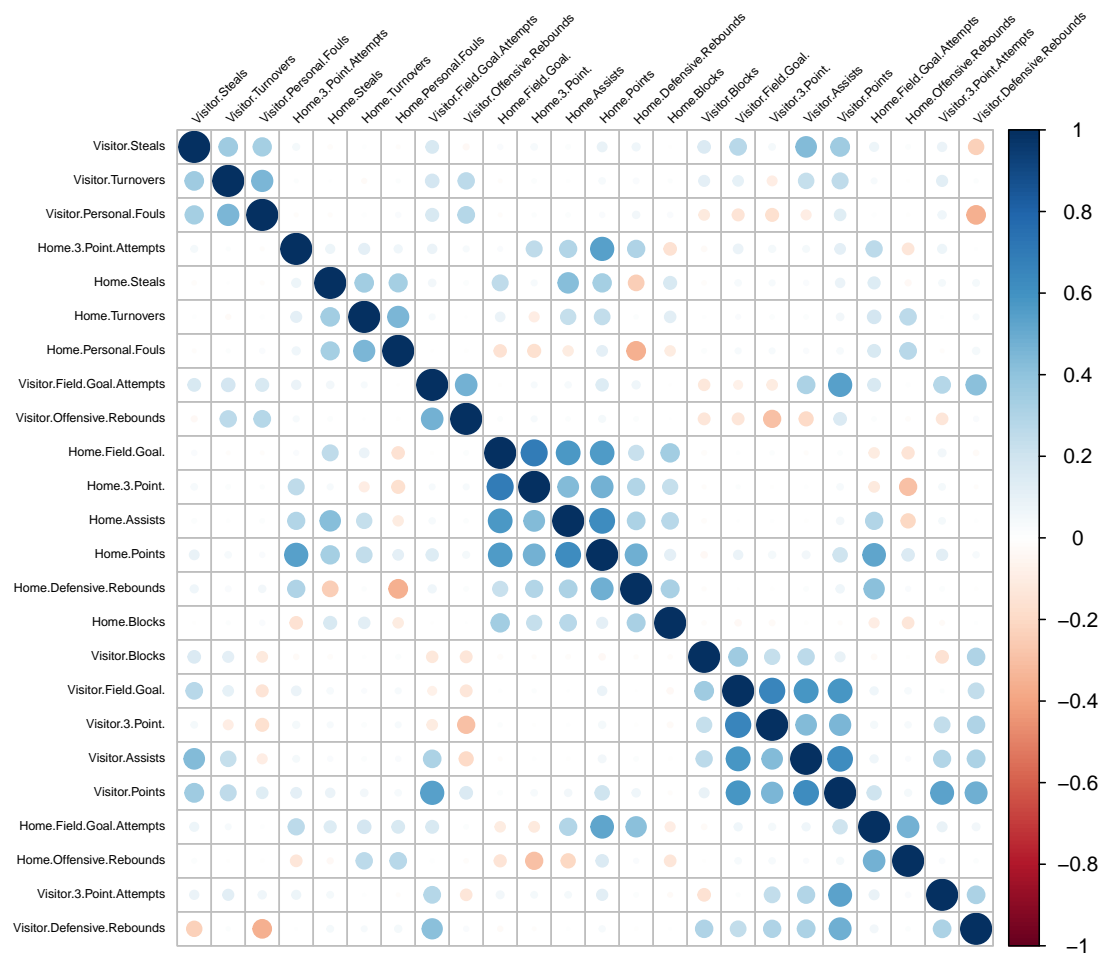


FIGURE 8.2 – *Importance des variables dans la prédiction*

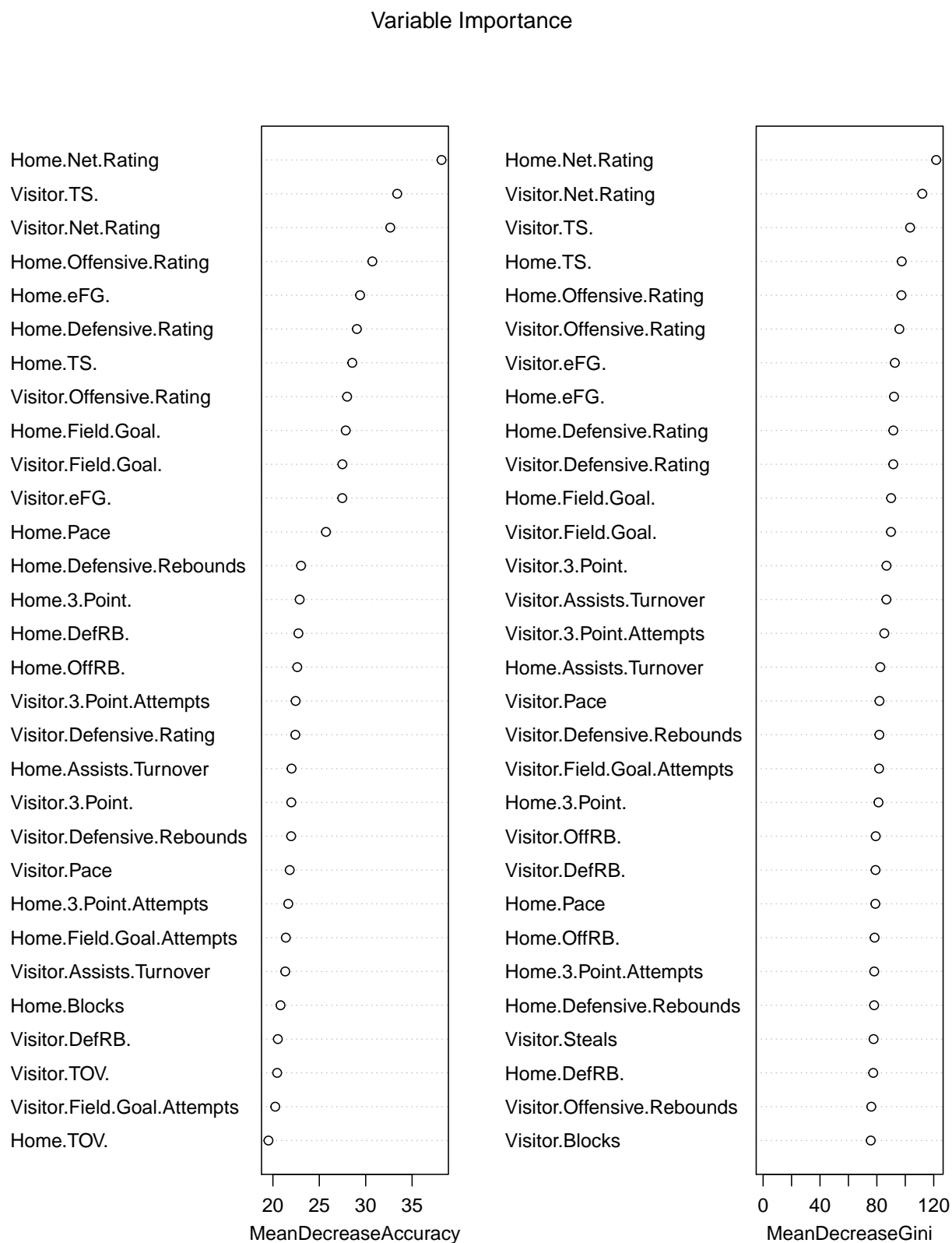


FIGURE 8.3 – *Arbre de décision recentré sur les variables de jeu*

Decision Tree

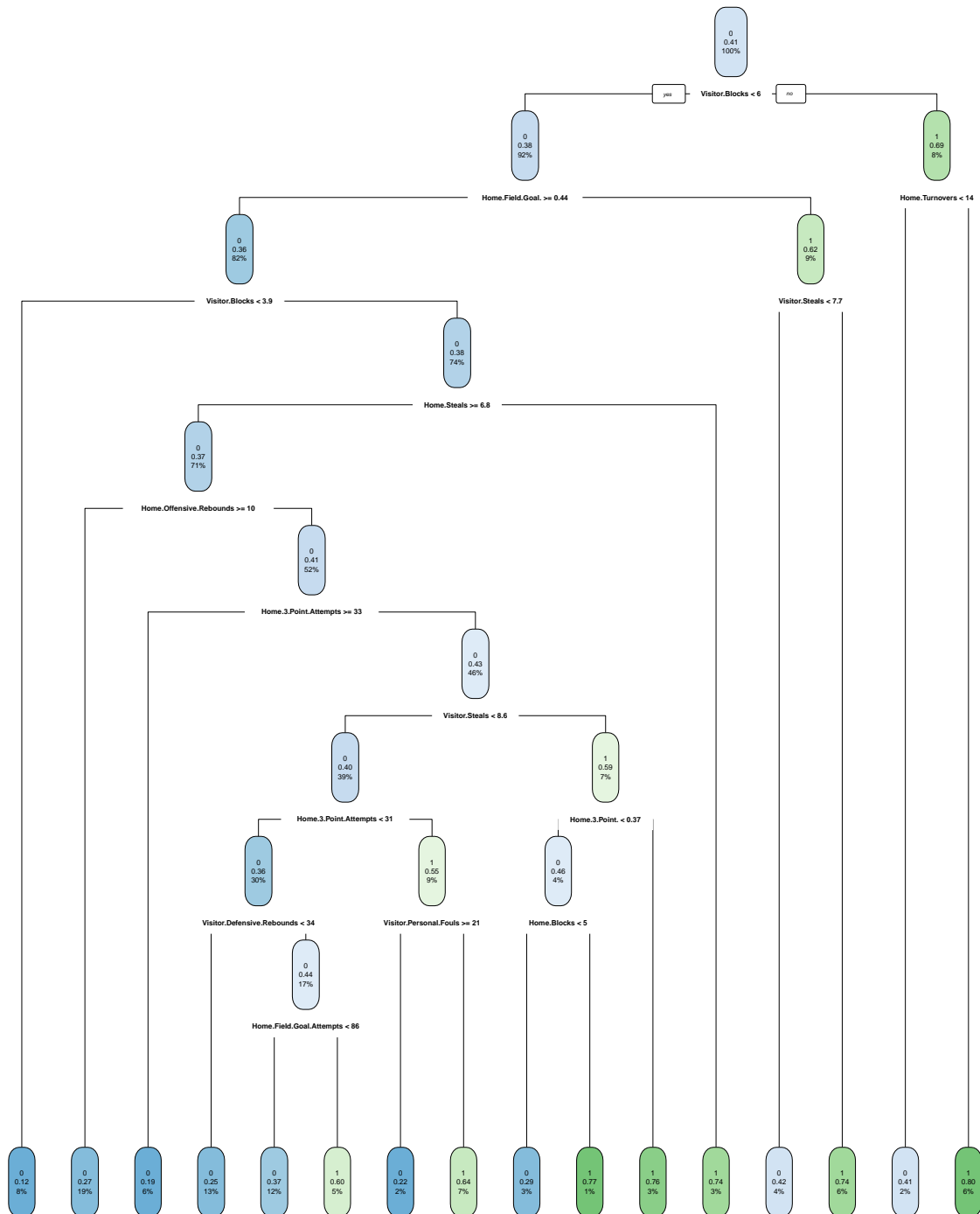


FIGURE 8.4 – *Corrélation nulle*

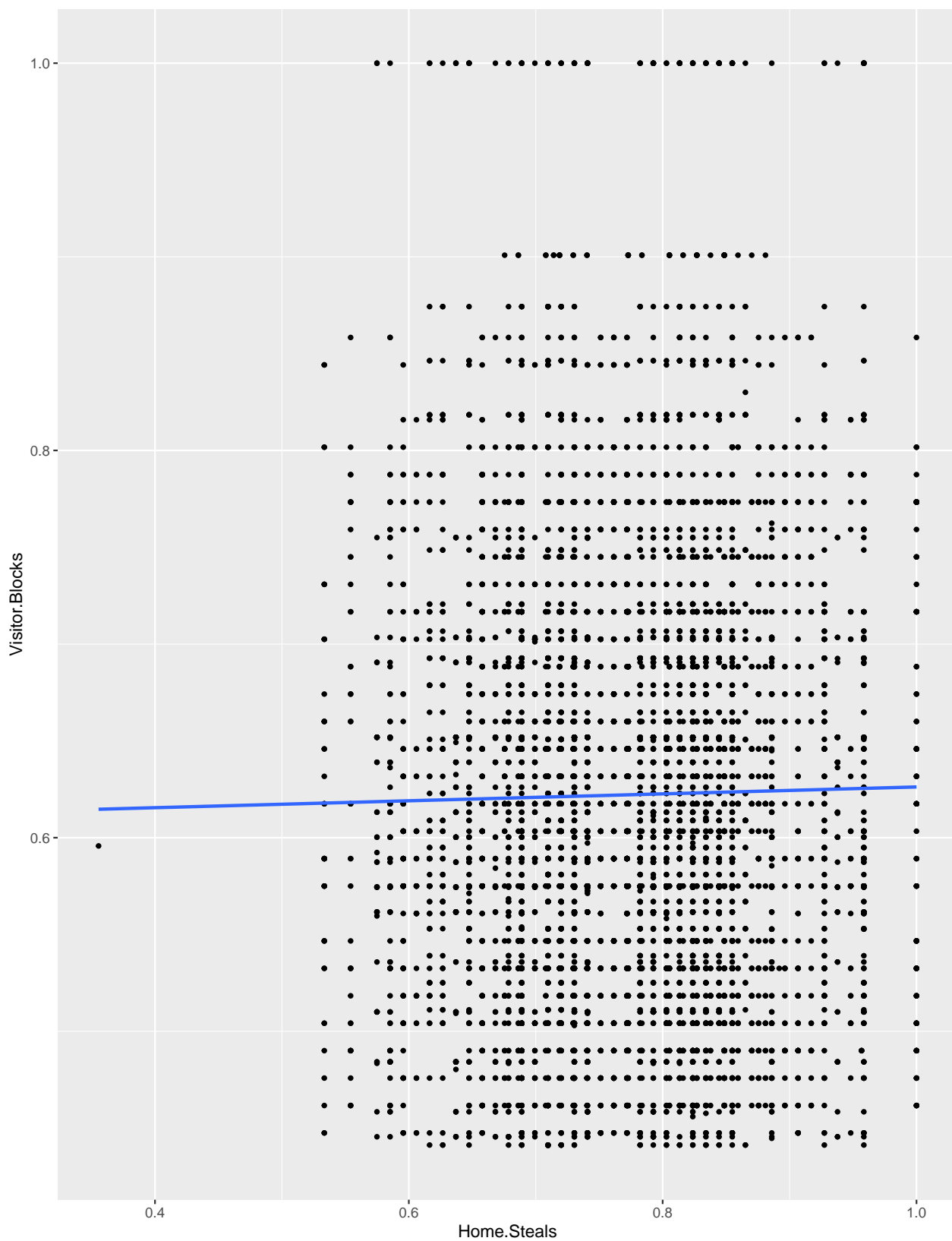


FIGURE 8.5 – Coefficients des variables de la régression logistique

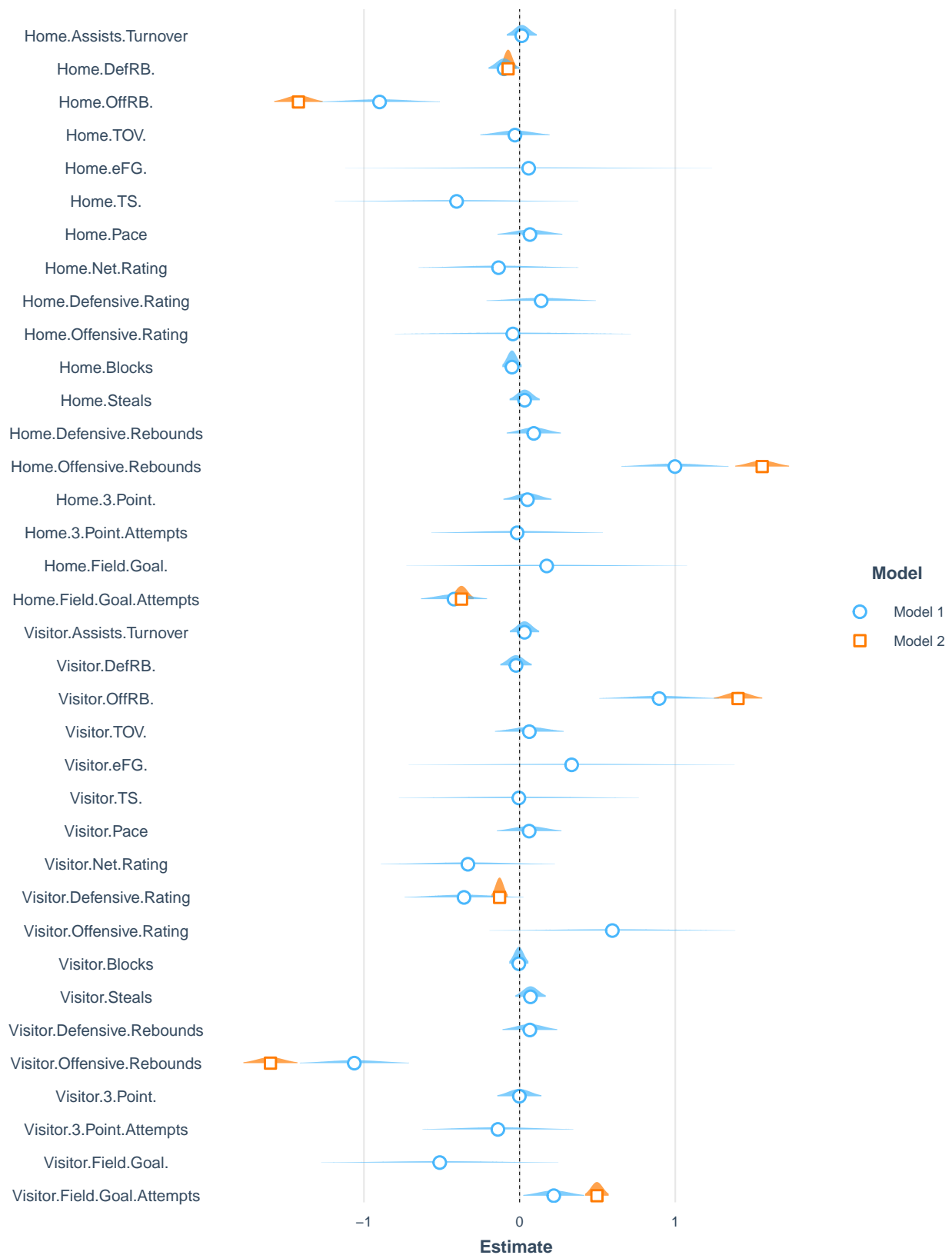


FIGURE 8.6 – *Histogramme du nombre de noeud par arbre de décision présent dans la forêt aléatoire.*

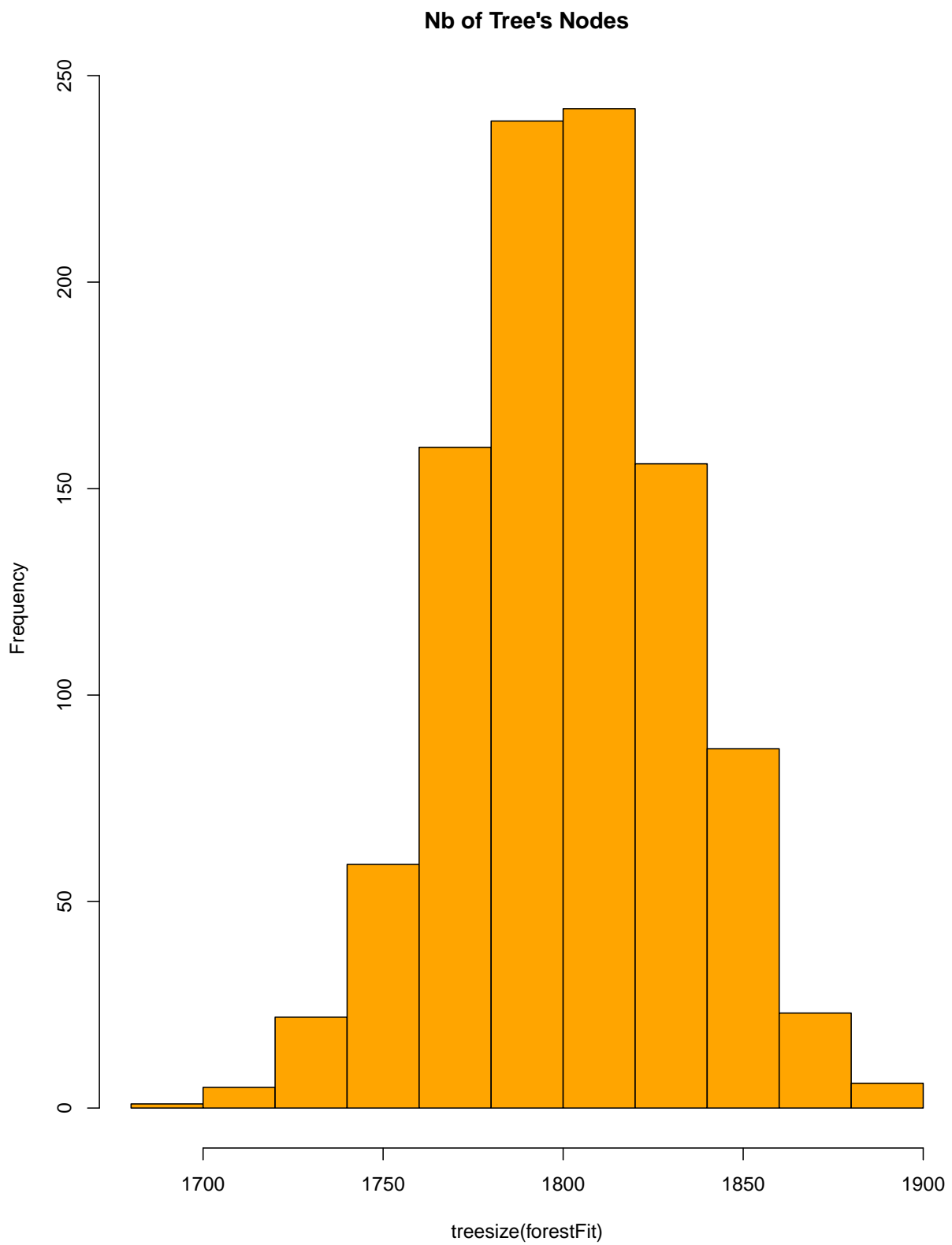


FIGURE 8.7 – "Home Net Rating" - Partial plot

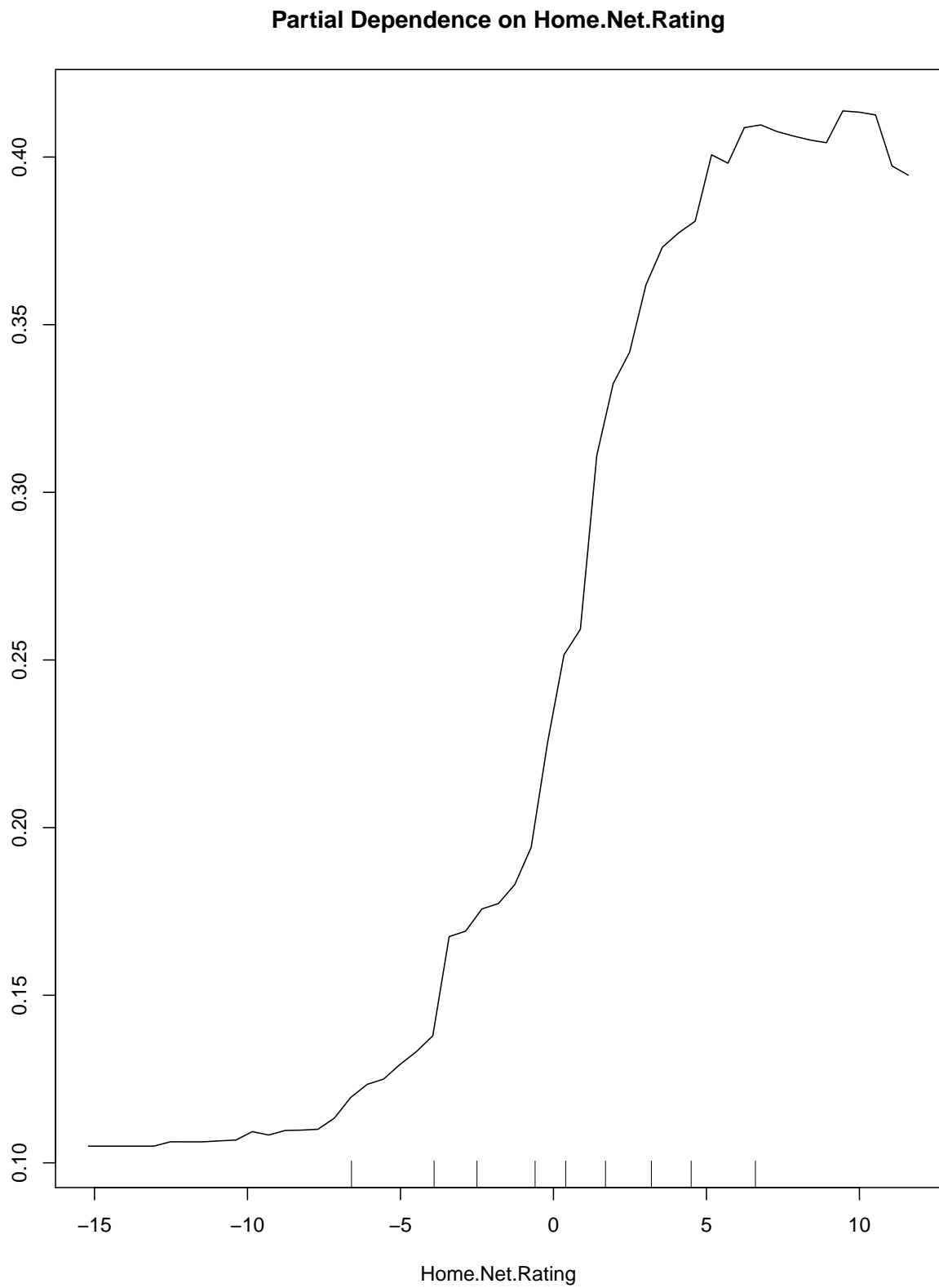


FIGURE 8.8 – "*Visitoir.Net.Rating*" - *Partial plot*

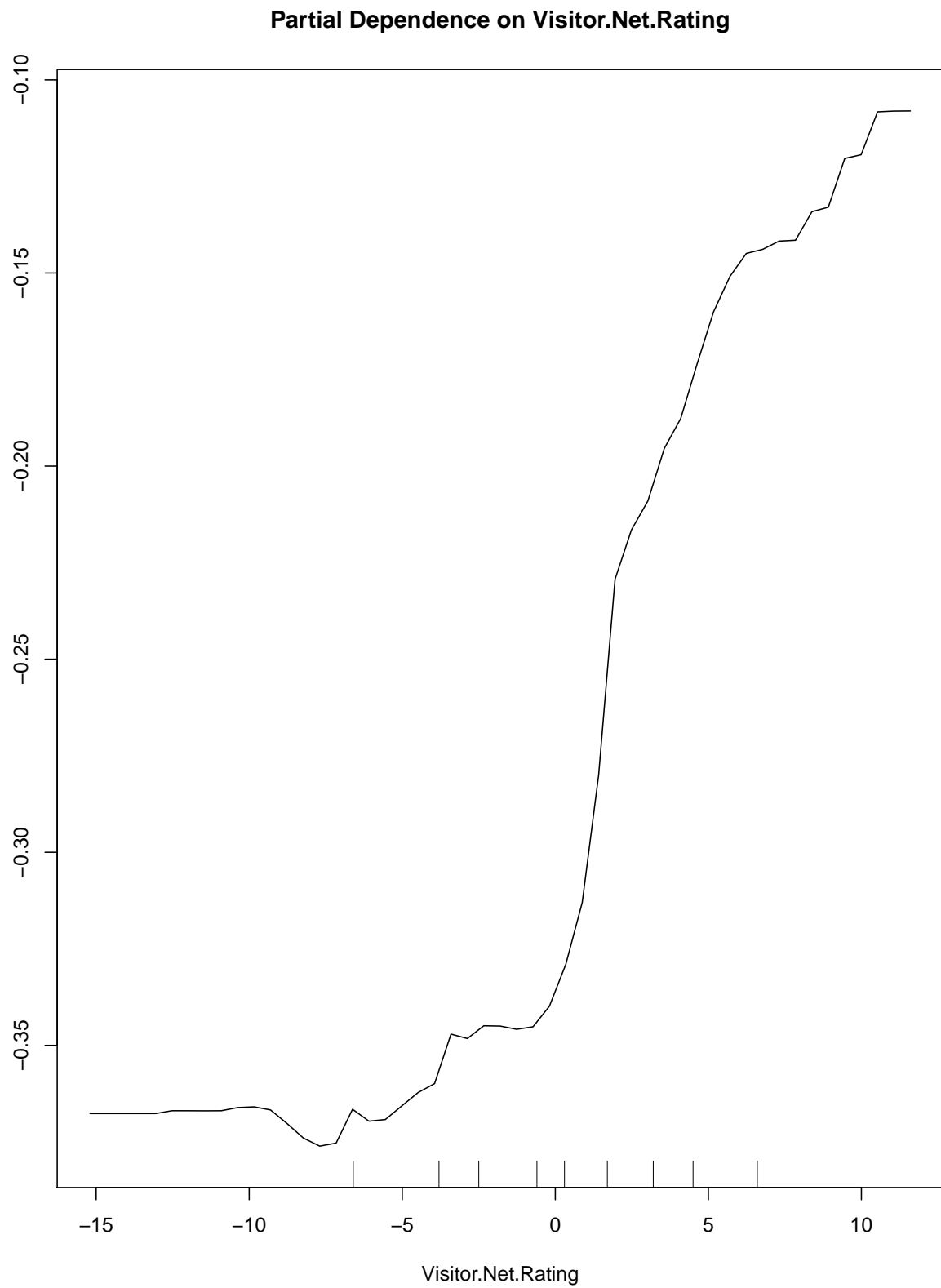
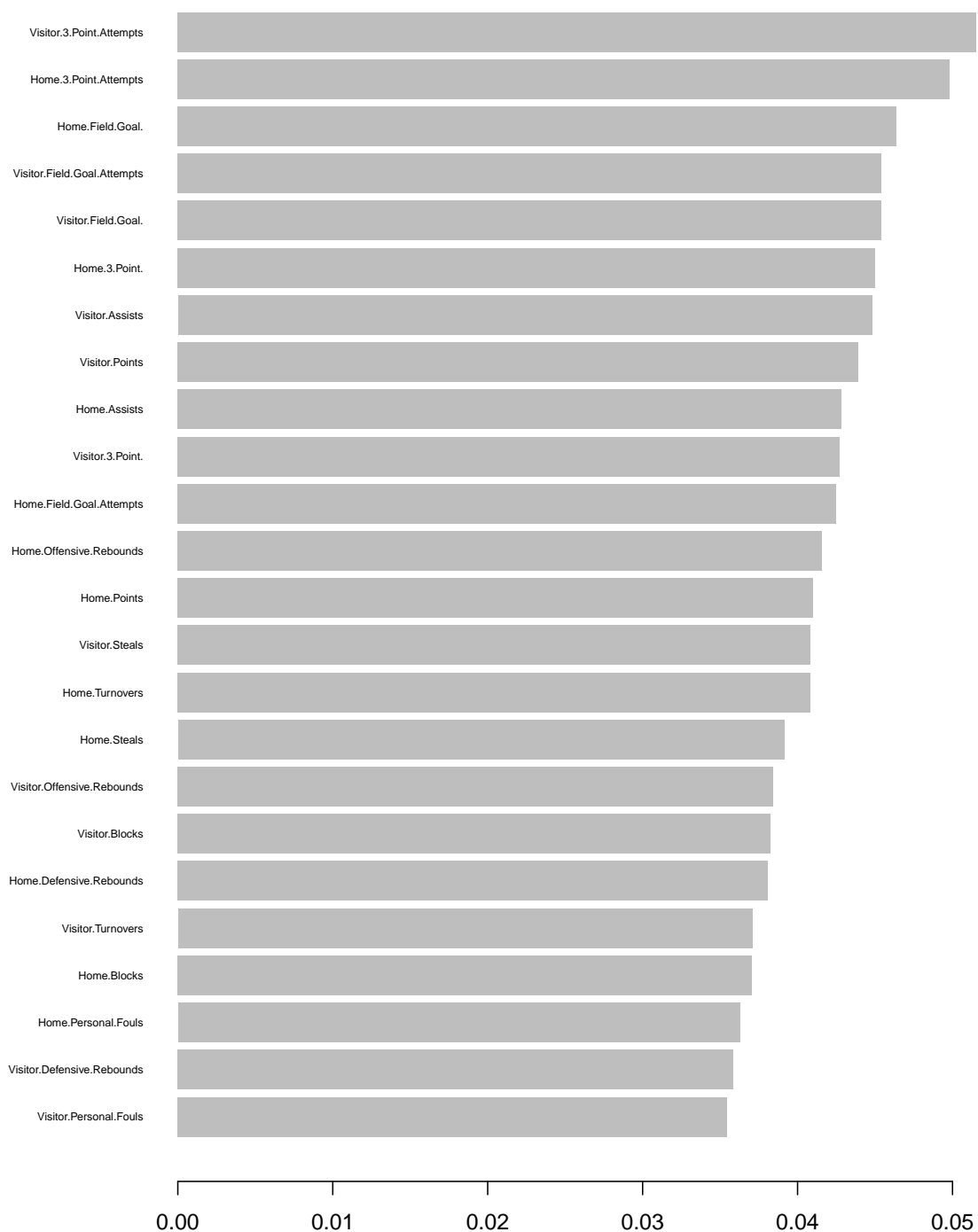


FIGURE 8.9 – *Importance des variables selon le modèle XGBoost.*



9. Glossaire

ML : Machine Learning

DL : Deep Learning

FG : Field Goals, *les tirs pris en jeu. Il prend en compte à la fois les tirs à 2 points et les tirs à 3 points.*

FGA : Field Goal Attempts, *nombre de tentatives de tirs pris en jeu.*

FG% : Field Goal Percentage, *pourcentage de tirs réussis pris en jeu.*

3P : 3-Points Field Goals, *les tirs réussis et tentés derrière l'arc comptant pour trois points.*

3PA : 3-Points Field Goal Attempts, *nombre de tentatives de tirs à 3 points pris.*

3P% : 3-Points Field Goal Percentage, *pourcentage de tirs réussis à 3 points.*

2P : 2-Points Field Goals, *les tirs réussis et tentés derrière l'arc comptant pour deux points.*

2PA : 2-Points Field Goal Attempts, *nombre de tentatives de tirs à 2 points pris.*

2P% : 2-Points Field Goal Percentage, *pourcentage de tirs réussis à 2 points.*

FT : Free Throws, *lancers-francs marqués et tentés.*

FTA : Free Throws Attempts, *nombre de lancers-francs tentés.*

FT% : Free Throws Percentage, *pourcentage de lancers-francs réussis.*

ORB : Offensive Rebounds, *rebonds offensifs.*

DRB : Defensive Rebounds, *rebonds défensifs.*

TRB : Total Rebounds, *nombre total de rebonds pris.*

AST : Assists, *passes décisives.*

STL : Steals, *interceptions.*

BLK : Blocks, *contres.*

TOV : Turnovers, *pertes de balles.*

PF : Personnel Fouls, *fautes.*

PTS : Points, *total de points marqués.*

ORtg : Offensive Rating, *estimation de points produits par les joueurs ou marqués par l'équipe sur 100 possessions.*

DRtg : Defensive Rating, *estimation du nombre de points encaissés sur 100 sessions.*

NRtg : Net Rating, *estimation du différentiel de points entre l'attaque et la défense sur 100 possessions.*

PACE : Pace factor, *estimation de la possession de la balle pendant 48 minutes.*

TS% : True Shooting Percentage, *pourcentage de l'efficacité au tir en comptant les 3-points, 2-points et les lancers francs.*

eFG% : effective Field Goal Percentage, *statistique ajustée des tirs pris en jeu en tenant compte que les tirs à 3-points valent un point de plus que les tirs à deux points.*

Bibliographie

- [1] Benito Santos (Alejandro), Theron (Roberto), Losada (Antonio), Sampaio (Jaime E.), and Lago-Peñas (Carlos). Data-driven visual performance analysis in soccer : An exploratory prototype. *Frontiers in Psychology*, vol 9 :pp.2416, 2018.

Les auteurs nous parlent du football et de la compréhension du comportement tactique collectif qui est devenue une partie intégrante de l'analyse sportive au niveau de l'élite. Toutes ces informations sont minutieusement examinées par des milliers d'analystes du monde entier à la recherche de réponses qui peuvent à long terme aider à augmenter les performances des individus ou des équipes dans leurs compétitions respectives.

- [2] Sciacchitano (Andrea), Caridi (Giuseppe Carlo Alp), and Scarano (Fulvio). A quantitative flow visualization technique for on-site sport aerodynamics optimization. *Procedia Engineering*, (no 112) :pp.412–417, 2015.

Les auteurs se sont axés sur le rôle crucial que joue l'aérodynamie dans de nombreux sports de vitesse, où les courses sont souvent gagnées par des fractions de seconde. Une compréhension approfondie du champ d'écoulement autour d'un athlète est d'une importance capitale pour optimiser la posture, la rugosité du vêtement et la forme de l'équipement des athlètes afin d'obtenir la traînée aérodynamique minimale et la vitesse maximale.

- [3] Cox (Andy) and Stasko (John). Sportvis : Discovering meaning in sports statistics through information visualization. 2002.

L'article concerne un logiciel nommé SportVis permettant d'aider les gens à découvrir le sens de l'énorme quantité de statistiques générées lors d'événements

sportifs et aussi pour aider les utilisateurs avec des tâches spécifiques à l'analyse du sport.

- [4] Phulkar (Dr. Ashish) and Kagzi (Imran I). Effect of visualization and imagery training on sports performance using sports hypnosis. *International Journal of Physical Education, Sports and Health*, vol 4(no 4) :pp.330–333, 2017.

Cet article, est basé sur une expérimentation effectuée sur 22 joueurs de cricket, afin d'examiner les effets de la technique d'imagerie utilisant l'état hypnotique sur les performances.

- [5] Guillot (Aymeric). *Visualisation en sports de combat : vaincre grace au mental*. Amphora, 2012.

ce livre fait le lien entre les méthodes de visualisation et les sports de combats. La visualisation dans les sports de combats n'est pas uniquement utilisée pour l'amélioration des performances physiques mais aussi sur le plan mental, étant un facteur important des sports de combat il est primordial d'intégrer cet aspect du sport dans la visualisation.

- [6] Ofoghi (Bahadorreza), Zeleznikow (John), MacMahon (Clare), and Raab (Markus). Data mining in elite sports : A review and a framework. *Measurement in Physical Education and Exercise Science*, vol 17(no 3) :pp.171–186, 2013.

L'objectif de cet article est de connecter les domaines du sport et de l'exploration de données à travers, la description d'un cadre pour catégoriser les sports d'élite, et la compréhension des exigences analytiques des différents problèmes d'analyse des performances.

- [7] Wang (Bin). Evaluation of sports visualization based on wearable devices. *International Journal of Emerging Technologies in Learning*, vol 12(no 12) :pp.199–126, 2017.

Afin de visualiser la classe d'éducation physique à l'école, les auteurs ont créé un système de gestion des mouvements visualisé, qui enregistre efficacement les données d'exercice de l'élève et stocke les données dans la base de données.

- [8] Moon (Bo) and Brath (Richard). Bloomberg sports visualization for pitch analysis. In *Workshop on Sports Data Visualization*, Atlanta, USA, 2013.

Cet article porte sur la visualisation des données dans le baseball au niveau par terrain, y compris les données mesurées et la vidéo. Elles peuvent être rendues efficace pour les entraîneurs et les managers grâce à une variété de techniques de visualisation des informations visant à créer des visualisations faciles à utiliser et à comprendre.

- [9] Millington (Brad) and Millington (Rob). ‘the datafication of everything’ : Toward a sociology of sport and big data. *Sociology of Sport Journal*, 32 :pp.140–160, 2015.

Cet article explore les articulations du sport et des « mégadonnées », un sujet important qui a été peu étudié jusqu’à présent.

- [10] Favier-Ambrosini (Brice) and Quidu (Matthieu). Combining first- and third-person data in sports sciences in france : Analysis of an original methodology. *Movement et Sport Sciences - Science et Motricité*, vol 2(no 100) :pp.39–52, 2018.

Cet article, évoque à l’aide de méthodologies indépendantes et de programmes de recherche compartimentés, l’étude des données à la première personne et des données à la troisième personne qui ont été tressés ensemble au cours de la dernière décennie dans le but d’accéder à une vision plus complète et plus complexe des actions.

- [11] Perin (Charles), Vuillemot (Romain), Stolper (Charles), Stasko (John), Wood (Jo), and Carpendale (Sheelagh). State of the art of sports data visualization. *Computer Graphics Forum*, vol 37(no 3) :pp.1–24, 2018.

Dans ce rapport, les auteurs ont organisé et réfléchi sur les avancées récentes et les défis dans le domaine de la visualisation des données sportives. Le but dans cet article est d’explorer la conception de nouvelles techniques de visualisation ; d’adapter les visualisations existantes à un nouveau domaine ; et de mener des études et des évaluations de conception en étroite collaboration avec des experts, y compris des praticiens, des passionnés et des journalistes.

- [12] Glez-Peña (Daniel), Lourenço (Anália), López-Fernández (Hugo), Reboiro-Jato (Miguel), and Fdez-Riverola (Florentino). Web scraping technologies in an api world. *Brefings in Bioinformatics*, 5 :pp.788–797, 2013. Le grattage de données Web, l’une des techniques les plus anciennes d’extraction de contenu Web, est toujours en mesure d’offrir un service valide et précieux à un large éventail d’applications bioinformatiques, allant des simples robots d’extraction aux méta-serveurs en ligne.
- [13] Rojas-Valverde (Daniel), Pino-Ortega (Jose), Gómez-Carmona (Carlos D), and Gutiérrez-Vargas (Randall). From big data mining to technical sport reports : the case of inertial measurement units. *BMJ Open Sport and Exercise Medicine*, vol 5 :pp.1–3.
Ce rapport vise à présenter de nouvelles méthodes de réduction des données et à proposer une nouvelle méthode d’approche pour l’analyse des résultats des unités de mesure inertielle(IMU).
- [14] Stanković (Daniel), Raković (Aleksandar), Joksimović (Aleksandar), Petković (Emilija), and Joksimović (Dina). Mental imagery and visualization in sport climbing training. *Activities in physical education and sport*, vol 1 :pp.35–38, 2011.
Le but de cet article était d’expliquer l’utilisation de l’imagerie mentale et de la visualisation dans l’entraînement d’escalade sportive. Les grimpeurs sportifs utilisent deux types de visualisation : dissociés et associés.
- [15] Messzyk (Edward) and Unold (Olgierd). Machine learning approach to model sport training. *Computers in Human Behavior*, pages pp.1499–1506, 2011.
Le but de cette étude était d’utiliser une approche d’apprentissage automatique de modéliser l’entraînement sportif, en particulier la natation, afin de surmonter le problème des définitions de classes qui se chevauchent et pour améliorer la compréhensibilité des règles.
- [16] Mężyk (Edwards) and Unold (Olgierd). Machine learning approach to model sport training. *Computers in Human Behavior*, 27 :pp.1499–1506, 2011. Le but

de cet article, de cette étude est d'utiliser une approche de Machine Learning combiné à au modèle fuzzy de afin de prédire et créer un modèle d'entraînement sportif, spécifiquement pour la natation.

- [17] Morgulev (Elia), Azar (Ofer H.), and Lidor (Ronnie). Sports analytics and the big-data era. *International Journal of Data Science and Analytics*, pages pp.213–222, 2018. L'explosion des données, avec de grands ensembles de données disponibles pour l'analyse, a affecté pratiquement tous les aspects de nos vies. L'industrie du sport n'a pas été à l'abri de ces évolutions. Nous fournissons des exemples de trois types d'analyses basées sur les données qui ont été effectuées dans le domaine du sport.

- [18] Zaïdi (H), Taïar (R), Fohanno (S), and Polidori (G). Surface flow visualization around competitive swimmers by tufts method. *Journal of Visualization*, vol 11(no 3) :pp.187–188, 2008.

Ce qui est présenté dans cette note fait partie d'une convention de partenariat avec la fédération française de natation, qui a été définie dans le cas de la préparation des JO de Pékin. Deux analyses complémentaires ont été réalisées, à savoir la mesure des dépenses énergétiques des athlètes par thermographie infrarouge et une analyse hydrodynamique spécifique.

- [19] Thornton (R Heidi), Delaney (Jace A.), Duthie (Grant M), and Dascombe (Ben). Developing athlete monitoring systems in team sports : Data analysis and visualization. *International Journal of Sports Physiology and Performance*, 2019.

Cet article évoque la collecte et l'analyse des données de surveillance des athlètes au sein des sports d'équipe dans le but d'évaluer la fatigue et les réponses d'adaptation ultérieures, d'examiner le potentiel de performance ainsi que de minimiser le risque de blessure et / ou de maladie.

- [20] Goodfellow (Ian), Bengio (Yoshua), and Courville (Aaron). *Deep Learning*. The MIT Press, Cambridge, MA, USA, 2016.

Le deep learning ou apprentissage profond est un type d'intelligence artificielle dérivé du machine learning.

- [21] Dr Daniel (James). Wearable technology in sport, a convergence of trends. *Journal of Advanced Sport Technology*, 1 :pp.1–4. La technologie portable offre la possibilité aux informations et à la vie de s'intégrer de manière transparente grâce à la mesure de la physiologie et de la biomécanique du corps. Bien que les technologies en soient encore à leurs balbutiements, elles sont aujourd'hui à la mode, d'actualité et connaissent une formidable croissance.
- [22] Aoki (K), Kinoshita (Y), Nagase (J), and Nakayama (Y). Dependence of aerodynamic characteristics and flow pattern on surface structure of a baseball. *Journal of Visualization*, vol 6(no 2) :pp.185–193, 2003.
Dans cet article, l'étude porte sur les caractéristiques aérodynamiques, telles que la portance et la traînée, d'une balle de baseball en caoutchouc officielle qui ont été étudiées expérimentalement en comparant les caractéristiques d'une sphère ayant une surface lisse et celles de sphères ayant des structures de surface différentes.
- [23] Leung (Carson K) and Joseph (Kyle W). Sports data mining : predicting results for the college football games. In *18th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems - KES 2014*, pages pp.710–719, Poland, 2014.
Au travers de ce rapport les auteurs vont utiliser leur approche d'exploration de données, qui permet de découvrir des connaissances intéressantes et de prédire les résultats de jeux de sport tels que le football universitaire pour effectuer des prédictions basées sur une combinaison de quatre mesures différentes sur les résultats historiques des matchs.
- [24] Till (Kevin), L. Jones (Ben), Cobley (Stephen), Morley (David), O'Hara (John), Chapman (Chris), Cooke (Carlton), and Beggs (Clive B). Identifying talent in youth sport : A novel methodology using higher-dimensional analysis. *PLOS ONE*, vol 11(no 5) :pp.1–18, 2016.

La prévision des performances des adultes à partir de l'identification précoce des talents dans le sport reste difficile. Le but des auteurs est d'essayer de la distinction entre les futurs joueurs amateurs et professionnels avec un haut degré de précision sans toutefois pouvoir prédire qui seront les futurs joueurs professionnels et académiques.

- [25] Nosu (Kiyoshi), Otsuka (Takashi), Okura (Taira), and Hirai (Yuriko). An immediate feedback video production system for a table tennis class. *Journal of Visualization*, vol 12(no 1) :pp.7–8, 2009.

L'article porte sur l'utilisation de la vidéo sur une classe de tennis de table afin de leur donner un feedback immédiat sur leurs compétences. L'auteur évoque la rétroaction qui peut être un élément important de la mécanique de l'acquisition des compétences, et les vidéos sont souvent utilisées en conjonction avec l'enseignement des compétences sportives pour l'éducation physique.

- [26] Capri (Harold L). *Data mining : principles, applications and emerging challenges*. Nova Publishers, New York, 2015.

Ce livre traite des principes, des applications et des nouveaux défis de l'exploration de données. L'exploration de data est un domaine de recherche où la recherche méthodologique et les moyens techniques appropriés sont expérimentés pour produire des connaissances utiles à partir de différents types de données.

- [27] Winston (Wayne L.). *Mathletics*. Princeton, Indiana, USA, 2012.

Comment utiliser des mathématiques simple pour analyser une gamme de questions statistiques et probabilistes dans le contexte du Baseball, du Basketball et du Football américain ainsi que dans les paris sportifs.

- [28] Cai (Li) and Zhu (Yangyong). The challenges of data quality and data quality assessment in the big data era. *Data Science Journal*, 14 :pp.1–10, 2015. Des données de haute qualité sont la condition préalable à l'analyse et à l'utilisation du big data et à la garantie de la valeur des données.

- [29] Haoqin (Li). Strategy and analysis of sport events based on data mining technology. *Applied Mechanics and Materials*, vol 687-691(no 3) :pp.1137–1140, 2014.

L'article porte sur l'utilisation du processus de Markov qui permet de trouver le processus focal des transitions d'action en calculant la différence de fiabilité du système afin d'appliquer la technologie d'exploration de données à l'analyse technico-tactique dans le domaine du sport.

- [30] Balafoutas (Loukas), Chowdhury (Subhasish M), and Plessner (Henning). Applications of sports data to study decision making. *Journal of Economic Psychology*, vol 75, 2019.

Le rapport traite des caractéristiques et des déterminants de la prise de décision de l'être humain dans différentes situations mais principalement lorsqu'il y a de la compétition. Les auteurs ont étudié le comportement des athlètes durant une compétition ainsi que l'aspect psychologique de la prise de décision.

- [31] Stein (Manuel), Janetzko (Halldór), Seebacher (Daniel), Jäger (Alexander), Nagel (Manuel), Hölsch (Jürgen), Kosub (Sven), Schreck (Tobias), and Keim (Daniel A.)and Grossniklaus (Michael). How to make sense of team sport data : From acquisition to data modeling and research aspects. *Data*, vol 2(no 2) :pp.1–23, 2017.

Le rapport, porte sur l'analyse des jeux de balle en équipe qui peut servir de nombreux objectifs, comme par exemple dans le coaching pour comprendre les effets des stratégies et des tactiques, ou pour obtenir des informations améliorant les performances. Cela est aussi décisif pour les entraîneurs et les analystes afin de comprendre pourquoi un certain mouvement d'un joueur ou de groupes de joueurs s'est produit, et quels sont les facteurs d'influence respectifs.

- [32] Stein (Manuel), Häußler (Johannes), Jäckle (Dominik), Janetzko (Halldór), Schreck (Tobias), and Keim (Daniel A). Visual soccer analytics : Understanding the characteristics of collective team movement based on feature-driven analysis and abstraction. *ISPRS International Journal of Geo-Information*, vol 4 :pp.2159–2184, 2015.

Les auteurs au travers de cet article ont proposé un système d’analyse visuelle pour l’identification interactive des schémas et des situations de football qui intéressent l’analyste. Ils ont analysé des matchs de football du monde réel, illustré plusieurs cas d’utilisation et recueillir des commentaires d’experts supplémentaires.

- [33] Pustisek (Matevz), Wei (Yu), Sun (Yunchuan), Umek (Anton), and Kos (Anton). The role of technology for accelerated motor learning in sport. *Personal and Ubiquitous Computing*, 2019. Dans cet article, il est question d’appareils portables mesurant une quantité physique ou physiologique d’un individu, et qui sont déjà partie de la vie quotidienne de nombreuses personnes.

- [34] Du (Meng) and Yuan (Xiaoru). A survey of competitive sports data visualization and visual analysis. *The Visualization Society of Japan*, pages pp.47–67, 2020. La visualisation de données sportives de compétition est une direction de recherche de plus en plus importante dans le domaine de la visualisation de l’information.

- [35] Gal-Petitfaux (Nathalie), Adé (David), Poizat D (Germain), and Seifert (Ludovic). L’intégration de données biomécaniques et d’expérience pour comprendre l’activité de nageurs Élite et concevoir un dispositif d’Évaluation. *Le travail humain*, vol 76(no 3) :pp.257–282, 2013.

Le but de ce rapport, était d’examiner l’activité des nageurs d’élite face à un dispositif technique sous-marin utilisé pour estimer les facteurs bio-mécaniques de la performance.

- [36] Schumaker (Robert P.), Solieman (Osama K), and Chen (Hsinchun). *Sports Data Mining*. Springer US, New York, 2010.

Ce livre nous explique la place de l’exploration de données dans le domaine du sport. Sports Data Mining rassemble en un seul endroit l’état de l’art en ce qui concerne un éventail international de sports : le baseball, le football, le basket-ball, le football et les courses de lévriers sont tous couverts, et les au-

teurs présentent les dernières recherches, développements, logiciels disponibles, et applications pour chaque sport.

- [37] Sakashita (R), Fujisawa (N), Matsuura (F), and Takizawa (K). Anaglyph stereo visualization of rhythmical movements. *Journal of Visualization*, vol 10(no 4) :pp.345–346, 2007.

Cet article vise à comparer les impressions des images planes et des images stéréo anaglyphes subissant des mouvements rythmiques en utilisant l’analyse factorielle.

- [38] Metulini (Rodolfo). Spatio-temporal movements in team sports : a visualization approach using motion charts. *Electronic Journal of Applied Statistical Analysis*, vol 10(no 3) :pp.809–831, 2017.

Cet article, apporte, une nouvelle approche de l’analyse des performances dans les sports d’équipe consiste à étudier les mouvements et les trajectoires des joueurs pendant le match. Les systèmes de suivi de pointe produisent des traces spatio-temporelles de joueurs qui ont facilité une variété de recherches visant à extraire des informations des trajectoires.

- [39] Bunker (Rory) and Susnjak (Teo). The application of machine learning techniques for predicting results in team sport : A review. *Cornell University*, pages pp.1–48, 2019.

- [40] Horvat (Tomislav) and Job (Josip). The use of machine learning in sport outcome prediction : A review. *Wiley Interdisciplinary Reviews : Data Mining and Knowledge Discovery*, pages pp.1–28, 2020. Prédire les résultats et extraire des informations précieuses est devenu attrayant non seulement pour les professionnels du sport, mais aussi pour un public plus large, en particulier dans les domaines de la gestion d’équipe et des paris sportifs.

- [41] Tani (Toshihiro), Huang (Hung-Hsuan), and Kawagoe (Kyoji). Sports play visualization system using trajectory mining method. *Procedia Technology*, (no 18) :pp.100–103, 2014.

- [42] Wentao Wang, Jiawan (Zhang), Xiaoru (Yuan), and Shixia (Liu). Matchorchestra : a generalized visual analytics for competitive team sports. *The Visualization of Japan*, 2015.

Dans cet article, les auteurs présentent un système MatchOrchestra généralisé et efficace pour analyser les sports d'équipe compétitifs en fonction de la partition musicale et de la métaphore de l'orchestre. MatchOrchestra fournit des vues sur les performances des joueurs, le statut de l'équipe, le tempo des matchs, la coopération et la confrontation des joueurs, ce qui peut aider les analystes à effectuer des tâches d'analyse spécifiques.

- [43] Chen (Wei), Lu (Junhua), Kong (Dingke), Liu (Zhiqi), Shen (Yandi), Chen (Yinyin), He (Jingxuan), Liu (Shu), Qi (Ye), and Wu (Yingcai). Gamelifevis : visual analysis of behavior evolutions in multiplayer online games. *Journal of Visualization*, vol 20, 2017.

Analyser les comportements des joueurs de jeux-vidéos multi-joueurs en ligne peut aider à comprendre la sociabilité et les caractéristiques des joueurs dans le monde virtuel. Les auteurs ont cherché à caractériser l'évolution du comportement des joueurs comme une séquence de transitions temporelles entre différents statuts.

- [44] Wikipédia. Bookmaker. <https://fr.wikipedia.org/wiki/Bookmaker>.

- [45] Yamagishi (Y) and Asai (K). The olympics and visualization. *Journal of Visualization*, vol 11 :pp.273, 2008.

L'article évoque les maillots de bain de haute technologie qui étaient devenus un sujet de conversation lors des Jeux Olympique de Pékin en 2008. Les tissus de maillots de bain étaient des matériaux scientifiquement avancés, avec des maillots de bain qui couvraient les bras et les jambes, réduisant la traînée à moins que celle de la friction de l'eau contre la peau.

- [46] Mingdong (Zhang), Li (Chen), Xiaoru (Yuan), Renpei (Huang), Shuang (Liu), and Junhai (Yong). Visualization of technical and tactical characteristics in fencing. *Journal of Visualization*, vol 22 :pp.109–124, 2018.

L’escrime est un sport qui repose fortement sur l’utilisation de tactiques. Cependant, la plupart des méthodes existantes d’analyse des données d’escrime sont basées sur des modèles statistiques dans lesquels les modèles cachés sont difficiles à découvrir. Dans cet article, les auteurs ont coopéré avec des experts pour analyser les caractéristiques techniques et tactiques des compétitions d’escrime.

Remerciements

Je tiens à remercier toutes les personnes qui ont contribué au succès de mon stage et qui m'ont aidée lors de la rédaction de ce mémoire.

Je voudrais dans un premier temps remercier, mon directeur de mémoire M.DE VALERIOLA, professeur de visualisation de données au sein de l'Université Libre de Bruxelles, de m'avoir permis de réaliser ce travail sur un sujet qui m'est cher et me passionne. Pour son enthousiasme en vers le sujet, son accompagnement, sa patience, sa disponibilité et surtout ses judicieux conseils qui ont contribué à alimenter ma réflexion.

Je remercie également toute l'équipe pédagogique de l'Université Libre de Bruxelles et les intervenants professionnels responsables de ma formation, pour avoir assuré la partie théorique de celle-ci.

Enfin, je tiens à témoigner toute ma reconnaissance aux personnes suivantes, à mes proches pour leur soutien dans la réalisation de ce mémoire. Leur aide, le partage et la confrontation des idées, leurs connaissances ainsi que leur appui et leurs encouragements dans les moments importants ont été décisif pour la réalisation de mon mémoire.

Les personnes proches autour de moi pour m'avoir apporter leurs idées, leurs savoir ainsi que leur soutien et leurs encouragements lorsque je les ai sollicité.