UNIVERSITÉ LIBRE DE BRUXELLES ULB

Solvay Brussels School
Economics & Management

MEMOIRE

Présenté en vue de l'obtention du Master en **Ingénieur de gestion**

Sports Analytics : Analyse des facteurs qui impactent les performances des footballeurs professionnels.

Par Mateusz Niewiarowski

Directeur : M. Germain VAN BEVER
Assesseur : M. Pierre Deville

Année académique 2023 - 2024

Je n'autorise pas la consultation de ce mémoire

Remerciements

Avant tout, je tiens à remercier le Professeur Van Bever pour son encadrement et son aide précieuse dans la structuration de ce travail. Je lui suis sincèrement reconnaissant pour sa disponibilité et ses conseils toujours pertinents, qui m'ont permis de progresser même dans les moments d'incertitude. Dès le début, le thème des sports analytics était évident, bien que la multitude de sujets possibles ait rendu l'orientation complexe. Grâce aux conseils du Professeur, j'ai pu explorer un domaine passionnant et encore peu exploré à ce jour.

Je tiens également à exprimer ma gratitude envers ma famille pour leur soutien inébranlable et leurs encouragements constants.

Résumé exécutif

Ce travail a pour objectif de prédire la performance d'un attaquant de football et comprendre les variables qui y contribuent le plus. Une approche dite de *data science* est utilisée pour mettre en évidence les variables les plus importantes et mettre en place les prédictions.

Ce travail a été motivé par une revue de littérature approfondie, qui a révélé des lacunes dans la recherche existante sur la prédiction des performances individuelles des joueurs offensifs au football. Bien que de nombreuses études aient été menées, elles se concentrent principalement sur la prédiction des scores de match ou la valorisation marchande des joueurs ((Chandre & Jennet Shinny, 2024), (Ati, Bouchet, & Ayachi Ben Jeddou, 2023), (Herold, Goes, & Nopp, 2019)). Ce projet vise donc à combler le manque d'analyse des facteurs influençant la performance des attaquants et à améliorer la prédiction de celle-ci.

La base de données utilisée après les pré-traitements contient 3033 joueurs offensifs évoluant sur cinq continents différents, afin de maintenir une grande diversité. Les analyses portent sur les données de la saison 2023. Ces données proviennent du célèbre jeu EA Sports FIFA, collectées via un projet de web scraping du site Sofifa, le site officiel en ligne contenant les données du jeu FIFA. La variable dépendante des modèles prédictifs, représentant les performances des joueurs est représenté par la variable 'overall' qui se trouve dans la base de données.

Initialement, la base de données comptait 110 variables. Dans un premier temps, les variables sans lien avec les attaquants ont été écartées. Ensuite, le filtre de corrélation a permis d'éliminer une autre série de variables pour éviter la multicolinéarité. Pour finir, plusieurs techniques de sélection de variables ont été testées pour ne retenir que les plus pertinentes.

Ensuite, cinq modèles de machine learning ont été utilisés pour prédire au mieux les performances des joueurs. Ces modèles comprenaient la régularisation linéaire de Ridge et Lasso, l'Elastic Net, la méthode des k plus proches voisins (KNN) et la forêt aléatoire. Chaque modèle a fait l'objet d'une explication détaillées afin de rendre ce travail accessible et compréhensible. Les hyperparamètres ont été optimisés selon une méthode de GridSearchCV, qui consiste à tester systématiquement toutes les combinaisons possibles des hyperparamètres spécifiés dans une grille prédéfinie. Cette méthode évalue chaque combinaison en entraînant un modèle sur un ensemble de données d'entraînement, puis en évaluant ses performances sur un ensemble de données de validation ou à l'aide de validation croisée. Ainsi, les prédictions sont assurées d'être impartiales et conformes à la configuration réelle d'utilisation des modèles.

À l'issue de cette démarche, le modèle KNN, combiné à la sélection de variables Lasso, a été choisi. Il a généré des prédictions finales satisfaisantes et cohérentes avec les attentes pour la variable 'overall'. De plus, l'analyse des coefficients de la régularisation Lasso a permis de mettre en évidence l'importance de différentes variables. Deux catégories spécifiques de variables se distinguent : tout d'abord, les qualités techniques, suivies par les aspects 'athlétiques' des joueurs offensifs. Les résultats de cette approche initiale apportent une compréhension significative au domaine en plein essor des analyses sportives.

Certaines variables reviennent fréquemment, mais leur importance varie. On remarque des tendances concernant les types de variables choisies. Premièrement, les compétences techniques des attaquants sont souvent représentées, telles que la précision du jeu de tête, les pénaltys et la qualité de tir. Les passes longues et les mouvements des attaquants sont également notables.

En ce qui concerne les aspects athlétiques, l'agilité et l'accélération des joueurs sont des facteurs clés. Les attaquants agiles peuvent changer de direction rapidement et sont plus enclins à tenter des actions spectaculaires. L'amélioration de la vitesse et de l'accélération est également importante.

Enfin, appartenir à un club en particulier influence également la note globale d'un attaquant. Les meilleurs clubs attirent les meilleurs joueurs, ce qui se reflète dans leur performance. Jouer pour un grand club offre plus d'opportunités de développement et de visibilité pour les attaquants, ce qui peut améliorer leur note globale.

Table des matières

Introduction

Pendant les 15 dernières années, nous avons vu une transition significative d'un monde où les trois quarts des données étaient analogiques à un monde où les données analogiques ne représentent qu'à peine 1 % du total (Negroponte, 1955). Les données numériques sont faciles à stocker, à chercher et à analyser, ce qui permet de convertir et d'utiliser efficacement de grandes quantités d'informations (Cao, 2017).

Le Big Data met en lumière le potentiel des données en tant que ressource précieuse pouvant être transformée en valeur économique. Ces données peuvent être utilisées de manière continue et variée, générant de la valeur. Jusqu'à récemment, les données étaient rarement réutilisées en raison de contraintes de stockage et d'analyse coûteuses. Cependant, la baisse de ces coûts transforme l'économie des données, rendant la réutilisation de plus en plus courante (Mayer-Schönberger, 2014). Cette évolution ouvre la porte à la création de nouveaux produits et services, et donc à de nouvelles sources de revenus pour les entreprises, potentiellement modifiant leurs modèles commerciaux. À mesure que plus d'entreprises prendront conscience de la valeur inexploitée que représente la réutilisation des données, le secteur privé ne manquera pas d'en recueillir plus, d'en stocker plus longtemps, de les utiliser plus fréquemment (Mayer-Schönberger, 2014).

Le domaine des sports ne fait pas exception à cette tendance, avec une croissance constante et le développement de diverses technologies d'analyse de données (Glebova & Desfontaine, 2020). L'analyse des données offre une valeur ajoutée significative aux performances des équipes professionnelles. Par exemple, elle permet d'évaluer la compatibilité entre deux joueurs pour améliorer les tactiques, de surveiller le niveau de fatigue des joueurs pour prévenir les blessures, ainsi que d'analyser les déplacements des joueurs sur le terrain (Lecuyer, 2022). De nombreuses entreprises ont émergé pour proposer des services d'analyse lors des transferts et des négociations de contrats (Oxybel, 2022).

Un définition du terme « *sport analytics* » donnée par le professeur universitaire et chercheur Benjamin Alama (Columbia University Press, 2013) est la suivante : « *Il s'agit d'utiliser des données liées au sport (des statistiques des joueurs à la météo du jour du match) pour trouver des modèles significatifs (corrélations fortes, tendances cachées, etc.) et communiquer ces modèles (à l'aide de graphiques, de tableaux, d'essais, etc.) pour aider à prendre des décisions.* (Stalbunov, 2014) »

La figure 1 ci-dessous illustre les principes du sport analytics.

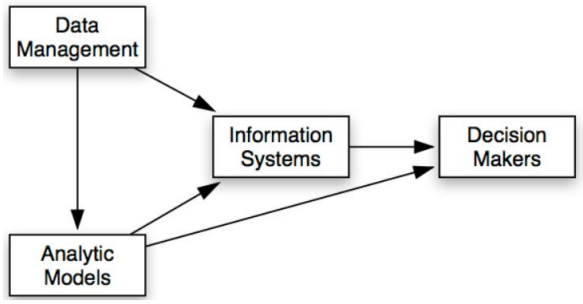


Figure 1 : Sports Analytics Framework *(Stolbunov, 2014)*

C'est seulement récemment que l'industrie du sport a connu un essor fulgurant, en adoptant des techniques de minage de données et d'apprentissage automatique de plus en plus avancées pour faciliter les opérations des clubs (Aoki, Assunção, & Vaz de Melo, 2017). Deux exemples remarquables illustrent cette tendance. Tout d'abord, l'équipe de baseball des Oakland Athletics a révolutionné l'approche du recrutement en développant une méthode novatrice basée sur l'analyse de données pour identifier les profils de joueurs idéaux, en se fondant sur des statistiques plutôt que sur des intuitions ou la réputation (Lewis, 2004). Ensuite, le cas du cycliste Christopher Froome avec l'équipe Sky, qui a remporté quatre victoires au Tour de France de 2013 à 2017 (Data Rockstars, 2021). Évoluant au sein de l'équipe avancée dans l'utilisation des données, Froome a pu optimiser ses performances et préserver sa condition physique grâce à l'analyse approfondie des données et à l'utilisation de multiples capteurs.

Actuellement, l'industrie de l'analyse sportive représente environ 1 milliard de dollars, avec un taux de croissance annuel de 30% ((Markets and Markets, 2022), (Mordor Intelligence, 2023)). Bien que cela ne représente qu'une fraction du marché global des sports, évalué entre 480 et 620 milliards de dollars, les experts prévoient qu'elle gagnera en importance dans les années à venir (QARA, 2019).

Les sports analytics offrent une multitude d'applications, allant de l'optimisation de l'entraînement des joueurs à un suivi médical renforcé, en passant par l'amélioration des stratégies pré-match et l'optimisation du processus de recrutement des joueurs (Herberger & Litke, 2021). Tout cela est rendu possible grâce à une amélioration constante de la technologie. Par exemple, les médecins peuvent maintenant surveiller de près la charge physique supportée par chaque joueur lors de chaque accélération, permettant une approche préventive pour réduire les risques de blessures et une intervention curative lorsque nécessaire (Dalgalarrondo, 2018). Les joueurs portent également des balises GPS lors des entraînements et des matchs, ce qui permet à l'équipe d'analyse de suivre chaque mouvement et de fournir des recommandations tactiques ou médicales en conséquence (Bekraoui & Leger, 2010). Ces balises GPS contribuent également à l'essor de bases de données massives et exhaustives disponibles en ligne, accélérant ainsi la recherche et facilitant la réalisation de telles analyses.

En remplaçant une partie du processus de recrutement par des approches plus axées sur les données, les clubs parviennent également à justifier en partie leurs dépenses massives. Appuyé par ses données de performance, le Real Madrid a acheté Gareth Bale pour 85 millions de livres sterling. Ont également été pris en compte ses prévisions de bénéfices de 41 millions de livres sterling sur 6 ans grâce aux ventes de maillots de football (Arrondel & Duhautois, 2022).

La plupart des clubs ont généralement déjà beaucoup de données ‘client’ qui sont naturellement collectées à travers divers canaux de consommation. Cependant, les analyses peuvent fournir des informations sur les produits et services que les supporters achètent et ceux qu'ils n'achètent pas (Arrondel & Duhautois, 2022).

De plus, pour augmenter le succès auprès des supporters actuels, l'analyse de données peut soutenir des campagnes marketing basées sur leurs habitudes (Olavsrud, 2022). Tout cela concerne simplement ce qui peut être fait avec les supporters actuels. Il existe même de plus grandes possibilités en appliquant l'analyse des données collectées sur les supporters potentiels ou tous les consommateurs du marché. Cela peut soutenir la conception de stratégies pour attirer des supporters et garantir que les campagnes marketing sont ciblées de manière appropriée (Olavsrud, 2022).

Cette introduction a offert une perspective élargie sur le domaine de l'analyse sportive. Celui-ci s'étend à une variété de sports et, combiné à l'analyse de données et à l'analytique commerciale, favorise son expansion dans le monde professionnel.

Ce projet se focalisera plus précisément sur l'analyse des variables qui influent sur les performances individuelles des joueurs de football en attaque, ainsi que sur la prédiction des performances des joueurs de football en se basant sur ces variables.

Questions de recherche

Traditionnellement, l'analyse des performances en football reposait principalement sur l'analyse des vidéos de matchs. Cependant, les avancées technologiques récentes permettent une analyse dynamique et plus contextuelle des variables. Les technologies de tracking automatique, l'analyse vidéo des mouvements, et les systèmes de positionnement global (GPS) ont considérablement amélioré la capacité à identifier des indicateurs de performance individuels et collectifs ((Rossi, 2018), (Buchheit, 2014)). Ces technologies permettent une évaluation rapide et précise des matchs, prenant en compte les paramètres contextuels rapidement changeants. Outre les approches actuelles, l'application de l'apprentissage automatique au football constitue un domaine de recherche émergent utilisé pour révéler les tendances et distinguer les équipes qui réussissent et celles qui le sont moins.

Malgré ces avancées, peu d'études se sont réellement concentrées sur l'utilisation du machine learning pour améliorer de manière significative la connaissance tactique (la compréhension des stratégies de jeu et des interactions entre les joueurs) et la performance globale des équipes de football. Les futurs travaux devront se pencher sur l'intégration de données provenant de diverses sources pour une analyse plus complète et contextualisée. Les entraîneurs et les analystes devront être formés à l'utilisation de ces nouvelles technologies pour en tirer pleinement parti.

Ainsi, de nombreuses recherches existantes se concentrent sur la prédiction du vainqueur d'un match, du score d'un match ou du positionnement d'un joueur. Par exemple, l'étude de Hucaljuk et Rakipovic démontre qu'en prenant en compte des facteurs tels que le nombre de blessures, les buts marqués, la formation de l'équipe, et d'autres variables, il est possible d'entraîner un modèle de machine learning supervisé pour prédire les scores ((Chandre & Jennet Shinny, 2024), (Ati, Bouchet, & Ayachi Ben Jeddou, 2023), (Herold, Goes, & Nopp, 2019)). Ce modèle a atteint une précision de 60 % dans ses prédictions (Ati, Bouchet, & Ayachi Ben Jeddou, 2023). De plus, ces travaux sont souvent basés sur de petites bases de données, avec un échantillon de 100 matchs, ce qui ne permet pas de tirer pleinement parti du potentiel des méthodes de machine learning (Dorfman, 2024).

Après avoir examiné la littérature existante, il apparaît clairement que la majorité des recherches se concentrent sur les résultats des matchs de football. Cependant, peu d'études se sont intéressées aux paramètres influençant les performances d'un joueur professionnel de football pour ensuite pouvoir prédire une performance générale d'un joueur de football. Ce travail se concentrera donc sur la question suivante : « Quels sont les paramètres qui influencent le plus les performances d'un attaquant professionnel de football ? » Cette question vise à vérifier, à l'aide d'une analyse de machine learning, si les paramètres considérés comme importants intuitivement seront confirmés.

Ensuite sur base de ces paramètres des modèles de machine learning supervisé seront mis en place afin pouvoir prédire les performances d'un attaquant de football.

Afin d'essayer de répondre à cette question ce travail sera basé sur une analyse des performances sportives en utilisant une base de données. Cette analyse des données nous amène à nous poser plusieurs questions :

- Quels sont les facteurs qui ont une influence sur les performances des attaquants ?
- Est-ce que ce sont forcément les paramètres physiques qui jouent le plus sur la performance ?

- Quels seront les paramètres sur lesquels les joueurs doivent travailler pour améliorer leur performance globale ?

Méthodologie

Ce mémoire se situe dans le domaine des "Sports Analytics", qui se définit comme l'analyse des performances sportives de footballeurs professionnels à l'aide d'outils d'analyse de big data et de statistiques (Pykes, 2022). Il implique initialement la recherche d'une vaste base de données. Les données 'brutes' doivent subir toute une série de pré-traitement afin de gérer les valeurs manquantes, les variables inutiles, les données redondantes, etc.

Dans le cadre de ce travail, j'ai exploité une base de données issue du site Kaggle, une plateforme proposant des bases de données sur divers sujets, acquise par Google en 2017 (Nicas, 2017). Sur Kaggle, j'ai identifié une base de données regroupant des informations sur les joueurs professionnels du populaire jeu vidéo FIFA EA Sports. Cette base de données est le fruit d'un projet de web scraping réalisé sur le site sofifa.com, qui rassemble des données précises exploitées par le jeu FIFA EA Sports. Pour les analyses de ce travail, j'ai exclusivement utilisé la base de données de l'année 2023, en me concentrant spécifiquement sur les joueurs offensifs.

La méthode CRISP[1], initialement connue sous le nom de CRISP-DM, a été développée par IBM dans les années 60 pour réaliser des projets de datamining. Aujourd'hui, elle reste la seule méthode efficace pour tous les projets de data science (Saltz, 2024). Ce modèle de processus d'exploration de données décrit une approche couramment utilisée pour résoudre les problèmes dans les domaines de l'analyse, de l'extraction et des sciences des données. Selon une enquête de la Harvard Business Review, 55% des cadres interrogés utilisent le processus CRISP pour exploiter les données de leur entreprise (O'Hara, Haylon, & Boyle, 2023).

La procédure CRISP-DM divise l'objectif global d'identification de modèles dans les données en une série de sous-tâches clairement définies, comprenant notamment la compréhension des besoins de la question de recherche, la collecte des données, l'analyse exploratoire des données, la modélisation, l'évaluation des modèles et le déploiement des résultats (Fawcett, 2013).

Une première analyse sera entreprise pour identifier les paramètres qui influent sur les performances des footballeurs lors des matchs grâce à diverses techniques telles que les 'filter methods', 'wrapper methods' et 'embbeded methods'.

Ensuite plusieurs modèles de machine learning, tels que la Ridge régression, Lasso régression, les forêts aléatoires et k-NN, seront employés pour prédire au mieux les performances des joueurs.

Une attention particulière sera accordée à la gestion de la complexité des modèles pour améliorer leur capacité à généraliser à un maximum de joueurs inconnus du modèle. L'objectif est d'obtenir des prédictions précises sans tomber dans le piège du surajustement. Il est souvent intuitif de penser que des modèles plus complexes sont meilleurs pour la prédiction, mais cette intuition erronée sera expliquée en détail plus tard dans le travail.

Pour surveiller la complexité d'un modèle, plusieurs techniques peuvent être utilisées. Une méthode courante consiste à ajuster les hyperparamètres du modèle, tels que la profondeur et la complexité des arbres dans les méthodes d'ensemble telles que les forêts aléatoires ou bien la gestion des pénalisations pour les régressions de Ridge et Lasso. En ajustant ces hyperparamètres de manière appropriée, on peut contrôler la complexité du modèle et éviter le surajustement. De plus, la validation croisée est aussi utilisée pour évaluer la capacité de

[1] **CR**oss **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining

généralisation du modèle sur des données non vues, ce qui permet de détecter tout surajustement potentiel.

1. Les thèmes de ce travail

1.1 Présentation du jeu FIFA

La série FIFA, produite par EA SPORTS (EA) depuis plus de deux décennies, est aujourd'hui la franchise de jeux vidéo de sport la plus importante au monde. Offrant une immersion dans le jeu universel du football, FIFA permet de jouer dans les plus grands clubs, championnats et avec les meilleurs joueurs de football du globe, le tout avec un niveau de réalisme et de détail impressionnant. Que ce soit pour constituer une équipe de rêve dans FIFA Ultimate Team, diriger son club favori au sommet en mode Carrière, participer à des matchs de rue avec EA SPORTS VOLTA FOOTBALL ou défier ses amis en mode Coup d'envoi, la franchise FIFA propose une expérience de jeu adaptée aux préférences.

Le célèbre jeu de football "FIFA" prend un nouveau nom, "EA Sports FC" (Mandatory, 2023). En 2022, après trois décennies de jeux portant le label "FIFA", Electronic Arts et la Fédération Internationale de Football Association (FIFA) ont pris la décision de mettre fin à leur collaboration. C'est une annonce historique, étant donné que cette collaboration existait depuis la sortie du premier jeu en 1993. La raison de cette séparation après une longue collaboration réussite est due à des négociations renouvelées concernant les droits de dénomination. La FIFA a exigé un milliard de dollars pour les quatre prochaines années, soit près du double de ce qu'EA Sports payait précédemment (L'avenir, 2023).

La franchise FIFA est l'une des plus rentables dans le monde du jeu vidéo, occupant la cinquième place des ventes historiques avec un total de 325 millions de copies écoulées depuis 1993. Bien qu'elle surpasse Grand Theft Auto (GTA) et ses 320 millions d'exemplaires vendus, elle reste devancée par des titres tels que Mario, Tetris, Call of Duty et Pokémon.

En 2023, le jeu a généré un chiffre d'affaires de 1,95 milliard de dollars, enregistrant une augmentation de 3 % par rapport à l'année précédente (Gonçalves, 2024). Le bénéfice net a également connu une croissance remarquable, s'élevant à 290 millions de dollars, soit une augmentation impressionnante de 42 % par rapport à 2022. Stuart Canfield, le directeur financier d'EA, a souligné que le changement de nom n'a apparemment pas eu d'impact négatif sur les revenus de l'entreprise (Gonçalves, 2024).

Le jeu a conquis un large public et reste toujours aussi populaire aujourd'hui, avec la récente sortie de EA Sports FC 24. Couvrant des événements majeurs comme la Coupe du Monde de football et le Championnat d'Europe, FIFA est devenu une référence incontournable dans le monde du football virtuel (Thibodeau, 2021).

Au fil des années, la série FIFA a évolué en intégrant de nouvelles fonctionnalités et améliorations techniques. Des emblèmes officiels des clubs dans FIFA 2001 à l'introduction de la barre de puissance pour les passes dans FIFA 2002, en passant par les refontes graphiques et les améliorations du gameplay dans les éditions suivantes, le jeu n'a cessé de se perfectionner pour offrir une expérience toujours plus immersive et réaliste (Trouvé, 2022).

En plus d'être un divertissement, FIFA est devenu un véritable sport à part entière, rassemblant une vaste communauté de joueurs. Des tournois officiels comme la FIFA e-world Cup attirent chaque année des compétiteurs du monde entier, offrant des récompenses attrayantes et l'opportunité de rencontrer les stars du football lors des BEST FIFA Football Awards (Thibodeau, 2021).

Avec son engagement constant envers l'amélioration de la qualité visuelle et de l'expérience utilisateur, FIFA continue de dominer l'industrie du jeu vidéo de football, établissant ainsi sa réputation de référence incontestée dans le domaine.

EA SPORTS utilise un système complexe pour garantir le réalisme de son jeu FIFA en mettant constamment à jour les données des joueurs. Actuellement, ces données sont évaluées manuellement par 8000 bénévoles qui analysent tous les matchs des différentes ligues chaque année. Ensuite, 200 éditeurs de données chez FIFA retravaillent ces informations ((Etienne, 2019), (Bohec, Le Télégramme, 2022)). Cependant, cette méthode présente des limites, notamment en ce qui concerne l'objectivité des évaluateurs. C'est là que les "data editors" d'EA Sports interviennent. Ces employés ajustent et modifient les données initiales récoltées. Ils sont plus de 200 à travailler pour l'entreprise, certains étant spécifiquement dédiés à harmoniser les notes en fonction des championnats et du niveau global du jeu (Bohec, Le Télégramme, 2022).

Les "data reviewers" constituent le premier maillon de cette chaîne. Ils sont plus de 8000 à évaluer les caractéristiques de chaque joueur dans le monde entier sur une base volontaire, recrutés par EA via les réseaux sociaux ou son site dédié (Etienne, 2019). Ils comprennent d'anciens professionnels, des coachs, des analystes et des supporters assidus, chargés d'évaluer les performances des joueurs. Leurs évaluations sont ensuite intégrées à une base de données dédiée.

À Cologne, en Allemagne, et à Guildford, au Royaume-Uni, une quarantaine de "Data Producers" supervisent le traitement des données statistiques des joueurs et des équipes((Etienne, 2019) , (Bohec, Le Télégramme, 2022)). Ils coordonnent le travail de centaines d'éditeurs de données chargés d'alimenter la base de données en ligne d'EA Sports. Cette équipe s'appuie également sur des milliers de bénévoles qui signalent les erreurs et partagent leur avis sur les notes des joueurs (Bohec, Le Télégramme, 2022).

Le processus de détermination des notes des joueurs sur une échelle de 100 par EA Sports repose sur différents attributs tels que la vitesse, l'agilité et la défense. Ces notes peuvent être ajustées tout au long de la saison pour refléter les tendances actuelles du football mondial, bien que dans certaines versions du jeu, ces mises à jour ne soient effectuées que de manière périodique. Cependant, le modèle final utilisé par EA Sports reste un secret bien gardé. L'entreprise ne divulgue ni le modèle de machine learning utilisé ni les variables spécifiques utilisées pour définir la note de chaque joueur (Vuille, 2017).

En 2023, le jeu contient une trentaine de championnats, plus de 700 équipes et 19 000 joueurs au total : la base de données du jeu Fifa est certainement l'une des plus complètes pour un jeu de simulation sportive (EA Sports FIFA 23, 2023).

1.2 Prédiction de la performance dans le football professionnel

Dans ce travail, l'approche adoptée pour prédire la note générale de performance d'un attaquant professionnel repose sur la construction de *modèles prédictifs* et l'analyse de leurs raisonnements sous-jacents, ce qui relève du domaine de la *modélisation supervisée* par *régression* (Rodrigues, Lourenço, Ribeiro, & Pereira, 2022). La régression supervisée consiste à modéliser la relation entre les caractéristiques d'entrée (variables indépendantes) et la variable cible (variable dépendante), dans le but de prédire cette variable cible pour de nouvelles données. Une condition essentielle pour cette approche est la disponibilité de données étiquetées, où la valeur de la variable cible est connue.

Par ailleurs, la base de données utilisée est assez vaste, comprenant un nombre considérable de données et de variables. Conformément aux questions de recherche, l'objectif principal était d'identifier les variables clés ayant le plus d'influence sur la variable "overall". Cela a été réalisé dans le but ultime de mettre en place un modèle de prédiction en utilisant les variables sélectionnées.

Pour répondre à ces objectifs spécifiques, quatre modèles descriptifs ont été développés, chacun comportant des variantes spécifiques. Chaque modèle adopte une approche distincte pour effectuer ses prédictions, ce qui permet de générer des informations exploitables adaptées à divers contextes.

- Régularisation de Ridge et Lasso
- La régression Elastic Net
- Méthode des k plus proches voisins (k-nearest neighbors algorithm)
  - Approche de sélection d'attributs pour la classification basée sur l'algorithme Random Forest
  - Sélection de variable via la régularisation de Lasso
- Régression Random Forest avec sélection de variable

1.3 Présentation du modèle CRISP-DM

Les quatre modèles ont été construits à l'aide de la méthodologie CRISP-DM. La procédure CRISP-DM divise l'objectif global d'identification de modèles dans les données en une série de sous-tâches clairement définies.

Les étapes suivantes sont décrites en s'appuyant sur une revue de littérature qui définit les étapes à suivre ((Wirth & Hipp, 2000), (Schröer, Kruse, & Gómez, 2021)).

Business understanding : Avant de commencer à explorer les données, il est essentiel de bien cerner les besoins de la recherche. Pour cela, il est nécessaire de définir clairement les objectifs à atteindre avec ce projet. Pour évaluer le succès de l'exploration, nous devons établir des critères précis. Ensuite, nous devons élaborer un plan de projet détaillé pour orienter notre exploration. Comme spécifié dans la question de recherche, l'objectif est de développer des modèles prédictifs en se basant sur les variables les plus significatives sélectionnées en amont.

Data understanding : Dans cette phase, il est crucial de collecter des données à partir de diverses sources et de les examiner en détail. Une étape importante consiste à décrire les données, ce qui implique d'utiliser des techniques d'analyse statistique pour comprendre les caractéristiques des données et pour classer les attributs selon leur importance. De plus, il est essentiel de vérifier la qualité des données pour s'assurer qu'elles sont fiables et pertinentes pour notre analyse.

Data preparation : Dans cette phase, il est essentiel d'établir des critères clairs pour sélectionner les données à inclure et celles à exclure. Si des problèmes de qualité sont détectés dans les données, un nettoyage s'impose pour les rendre exploitables. Ce processus de nettoyage s'effectue en appliquant séquentiellement quatre étapes distinctes :

- *Data integration* : Pendant cette phase, la manipulation concerne une base de données extraite via un projet de webscraping (les données viennent du site sofifa.com), contenant des données sur les joueurs professionnels de 2015 à 2023. La décision a été prise de focaliser l'analyse uniquement sur les attaquants pour l'année 2023.

- *Data cleaning* : Cette étape vise à résoudre le problème des valeurs manquantes dans les données.
- *Data transformation* : À cette étape, les données sont converties dans un format adéquat pour être traitées par le modèle choisi. Dans ce projet, nous nous concentrerons principalement sur la normalisation des variables et la conversion des variables catégorielles en valeurs numériques.
- *Data reduction* : Au cours de cette phase, l'objectif est de réduire le volume des données tout en préservant leurs caractéristiques essentielles. Les méthodes utilisées :
    - Lasso et Ridge
    - Random forest
    - Élimination récursive de caractéristiques

Modeling : Une fois les données préparées, les quatre modèles descriptifs sont construits à l'aide des données préparées. Les 4 modèles étudié dans ce travail sont :

- Lasso et Ridge regression
- Elastic net regression
- KNN regression
- Random forest

Evaluation : Dans cette partie nous analysons les résultats obtenus pour nos différents modèles. 6 métriques standard de mesure sont utilisées. Chaque résultat est analysé au niveau de sa pertinence :

- *Erreur absolue moyenne (MAE)* : La MAE mesure la moyenne des écarts absolus entre les valeurs estimées ou prédites et les valeurs observées. Elle est calculée en faisant la moyenne des valeurs absolues des écarts entre chaque valeur prédite et la valeur observée correspondante (Hodson, 2022).
- *Erreur quadratique moyenne (MSE)* : L'erreur quadratique moyenne permet de répondre à la question : « Quelle est l'ampleur de l'erreur de la prédiction ? ». C'est un outil d'analyse qui mesure les différences entre les valeurs réellement observées et celles prédites par un modèle. Cette erreur est toujours positive. Si les valeurs obtenues par le modèle se rapprochent des valeurs observées, les écarts sont faibles et l'erreur quadratique moyenne tend vers zéro (Kassel, 2022).
- *Coefficient de détermination (R^2)* : Le coefficient de détermination est un indice qui évalue la qualité de la prédiction d'un modèle de régression linéaire. Plus il s'approche de 1, meilleure est l'adéquation de la régression linéaire avec les données observées. Un coefficient de détermination de 1 signifie que le modèle explique parfaitement la variance des données observées, indiquant une adéquation totale entre les variables. En revanche, un coefficient proche de zéro indique que le modèle n'explique que très peu, voire pas du tout, la variance des données observées (Kassel, 2022).
- *Coefficient de détermination ajusté :* Le coefficient de détermination ajusté tient compte du nombre de variables. En effet, le principal défaut du R^2 est de croître avec le nombre de variables explicatives. Or, on sait qu'un excès de variables produit des modèles peu robustes. C'est pourquoi on s'intéresse davantage à cet indicateur qu'au R^2.
- Le critère d'information bayésien (BIC) : C'est un critère utilisé pour la sélection de modèles statistiques. Il évalue la qualité d'un modèle en prenant en compte la probabilité des données sous le modèle ainsi que la complexité du modèle. Plus le BIC

est faible, meilleur est le modèle. Le BIC pénalise les modèles comportant un grand nombre de paramètres pour éviter le surajustement (Lebarbier & Mary-Huard, 2006).

- Critère d'information d'Akaike (AIC) : Ce critère fonctionne de manière similaire au BIC mais avec une pénalisation légèrement différente. Les détails concernant ces deux critères seront donnés ultérieurement dans ce travail.

Pendant l'évaluation, nous comparons les résultats obtenus avec nos objectifs de prédiction. Ensuite, nous interprétons ces résultats pour en tirer des enseignements et envisager d'autres actions si besoin. Enfin, il est essentiel de réviser le processus dans son ensemble pour identifier les points à améliorer.

Dans la méthodologie CRISP-DM, la préparation des données et la modélisation sont étroitement liées, ce qui est essentiel pour obtenir des prédictions précises (Wirth & Hipp, 2000). Les étapes d'intégration et de transformation des données, comme la conversion des variables catégoriques en données numériques, sont communes à toutes les initiatives de modélisation. À partir de cette base commune la méthodologie consiste à rechercher la meilleure combinaison des quatre outils descriptifs pour chaque objectif défini.

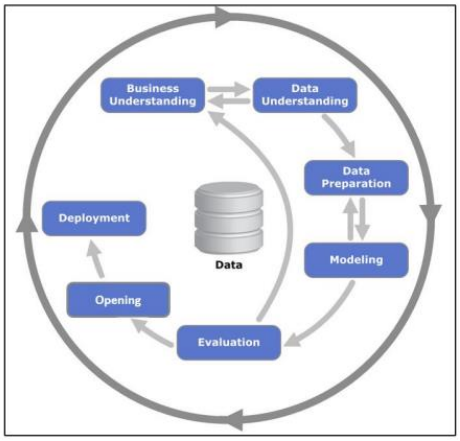


Figure 2: Modélisation de l'approche CRISP-DM (Schröer, Kruse, & Gómez, A Systematic Literature Review on Applying CRISP-DM Process Model, 2021)

1.4 Outils et logiciels utilisés

Toutes les opérations sont effectuées en utilisant le langage de programmation Python, en exploitant principalement les fonctionnalités offertes par la bibliothèque Scikit-learn. Les détails sur ces ressources, ainsi que les raisons de leur sélection, sont disponibles dans les annexes. Les outils GridSearchCV de Scikit-learn sont utilisés pour rechercher les meilleurs paramètres des modèles. Avant de faire des prédictions, les modèles doivent être entraînés sur les données. Pour garantir les meilleures performances possibles, il est crucial d'évaluer la qualité des prédictions. À cette fin, la base de données est divisée en deux ensembles : un *ensemble d'entraînement* et *un ensemble de test* (grâce à la fonction train_test_split). De plus, pour suivre l'évolution de l'apprentissage du modèle, des graphiques sont créés à l'aide de Matplotlib, une bibliothèque Python pour la visualisation des données.

Chapitre 1 : Construction des modèles

1. Présentation des variables

1.1 Collecte de données

Ce travail est basé sur une base de données trouvée sur Kaggle[2]. C'est un projet de webscraping (le fait d'extraire des données contenues sur une page web) qui reprend des données qui se trouvent dans le jeu EA Sports FIFA 2023. Le site utilisé pour l'extraction des données s'appelle sofifa[3]. C'est un site qui contient les données de tous les joueurs repris sur le jeu FIFA.

La base de données comprend à l'origine des données sur tous les joueurs pour tous les postes dans les championnats inclus dans FIFA sur les cinq continents. En raison du grand nombre de données disponibles, environ 166 674 joueurs, seules les données concernant les attaquants seront utilisées dans ce projet. L'année retenue sera celle de 2023. La base de données initiale comporte 110 attributs.

Cependant, il y a de nombreux doublons dans cette base de données en raison de la variable "fifa_update", qui reflète les mises à jour apportées aux joueurs tout au long de la saison en fonction de divers facteurs tels que les performances lors des matchs, les blessures, etc. J'ai choisi de sélectionner la dernière version de la variable "fifa_update", qui est égale à 9. En choisissant les données pour lesquelles la variable "fifa_update" est égale à 9 et en sélectionnant les postes offensifs j'obtiens une base de données de 3033 joueurs. Toutes les variables représentent les performances d'un joueur sur une saison complète.

Comme le but du travail reste de prédire la variable 'overall' pour les joueurs offensifs, j'ai décidé de ne pas sélectionner de minimum de variables.

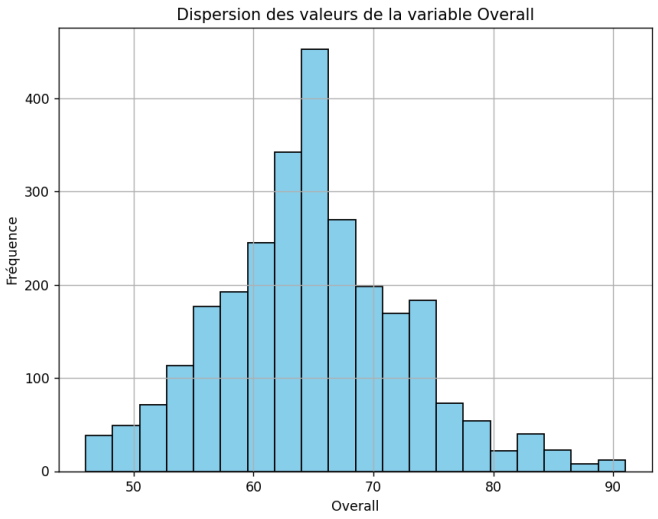


Figure 3: Dispersion de la variable dépendante 'overall'

[2] https://www.kaggle.com/datasets/stefanoleone992/fifa-23-complete-player-dataset?resource=download
[3] https://sofifa.com/?hl=fr-FR

En ce qui concerne la distribution de la variable dépendante, on observe qu'elle est centrée autour de 65. Cependant, la majorité des joueurs ont une valeur 'overall' inférieure à 65. Cela n'est pas surprenant, car en incluant tous les championnats présents dans le jeu FIFA, il y a plus de championnats de niveau faible que de championnats de haut niveau alignant les meilleurs joueurs. En effet, seuls quelques championnats européens sont classés parmi les meilleurs et alignent les meilleurs joueurs offensifs du monde. On remarque d'ailleurs qu'il y a très peu de joueurs avec une valeur 'overall' supérieure à 80, car ceux-ci font partie des meilleurs joueurs offensifs au monde.

Étant donné la sous-représentation des joueurs ayant des notes supérieures à 80, j'ai exploré deux méthodes de rééchantillonnage afin d'évaluer leur impact sur la distribution des variables.

Après avoir effectué une revue de la littérature, j'ai décidé d'explorer deux techniques pour enrichir l'échantillon des variables supérieures à 75. La première stratégie, la plus simple, consiste à générer de nouveaux échantillons par échantillonnage aléatoire pour compléter les données existantes (RandomOverSampler). La deuxième technique, appelée SMOTE (Synthetic Minority Oversampling Technique), fonctionne en sélectionnant aléatoirement un point dans la classe minoritaire, puis en calculant les k voisins les plus proches pour ce point (Torgo, Ribeiro, & Pfahringer, 2013). Des points synthétiques sont ensuite ajoutés entre le point choisi et ses voisins.

Néanmoins pour conclure j'ai décidé de ne pas appliquer ces techniques sur mes analyses finales pour 2 raisons principalement :

- Le rééchantillonnage peut introduire une variance supplémentaire dans les estimations de modèle. Par exemple, les techniques de bootstrap peuvent augmenter la variance des estimations, ce qui peut rendre les résultats moins robustes et plus sensibles aux variations des données d'entrée.
- Le rééchantillonnage peut altérer les distributions d'origine des données. Par exemple, les techniques de sur-échantillonnage comme le SMOTE (Synthetic Minority Over-sampling Technique) génèrent des échantillons synthétiques qui peuvent modifier les relations naturelles entre les variables.

1.2 Variable dépendante

Avant de mettre en place la liste des nombreuses variables indépendantes collectées, il est nécessaire de définir la cible, c'est-à-dire la variable dépendante qui représente la performance d'un joueur de football pendant la saison 2023.

La variable dépendante choisie est la variable « overall » qui est une note sur 100. Cette variable représente à quel point le joueur est bon à son poste au football. En effet chaque poste au football nécessite des qualités différentes. Ainsi les variables explicatives sélectionnées par EA Sports FIFA seront différentes en fonction du poste du footballeur professionnel. Comme le travail est focalisé sur les attaquants, la note « overall » note la qualité générale d'un attaquant de football.

1.3 Variables indépendantes

Chaque ligne correspond à un joueur offensif de football avec toutes les caractéristiques suivantes :

Initialement, sans aucun traitement particulier, la base de données contient 110 variables. Ces variables peuvent être divisées en plusieurs catégories. Pour présenter toutes ces variables, je vais utiliser un exemple précis de la base de données afin de mieux les illustrer.

1.3.1 Données relatives au jeu FIFA

- player_id: 158023
- player_url: /player/158023/lionel-messi/230009
- fifa_version: 23
- fifa_update: 9
- fifa_update_date: 2023-01-13

Ces attributs permettent de déterminer quelle version de FIFA a été utilisée pour la base de données. Comme spécifié précédemment, l'année retenue pour ce travail est 2023, avec la version ‘up_date’ égale à 9, ce qui signifie qu'il s'agit de la dernière version de FIFA pour 2023, représentant les performances sur une saison de football. De plus, les joueurs peuvent être identifiés par un numéro unique sous la variable 'player_id'.

1.3.2 Données personnelles

- short_name: L. Messi
- long_name: Lionel Andrés Messi Cuccittini
- player_positions: RW
- age: 35
- dob: 1987-06-24
- height_cm: 169
- weight_kg: 67
- value_eur: 54000000.0
- wage_eur: 195000.0
- release_clause_eur: 99900000.0

Ces variables permettent d'abord d'identifier le joueur. Ensuite, elles incluent des informations physiologiques. Enfin, elles indiquent le revenu du joueur ainsi que sa valeur marchande.

1.3.3 Données relatives au club

- league_id: 16.0
- league_name: Ligue 1
- league_level: 1.0 (1-5)
- club_team_id: 73.0
- club_name: Paris Saint Germain
- club_position: RS (right striker)
- club_jersey_number: 30.0
- club_loaned_from: nan
- club_joined_date: 2021-08-10
- club_contract_valid_until_year: 2023.0

Chaque joueur possède des données concernant le championnat dans lequel il joue. Comme pour les joueurs, le club et le championnat ont aussi leur numéro unique. On voit que les championnats sont aussi déjà classés sur une échelle de 1 à 5. Ce ne sera pas le cas pour tous les joueurs mais pour certains les dates de contrat sont connues.

1.3.4 Données relatives à la nationalité

- nationality_id: 52
- nationality_name: Argentina
- nation_team_id: 1369.0
- nation_position: RW
- nation_jersey_number: 10.0

Quelques variables sont consacrées à la nationalité de l'attaquant. Les joueurs peuvent ne pas avoir le même poste ou rôle en équipe nationale.

1.3.5 Données de performance sportive

weak_foot: 4	skill_moves: 4	international_reputation: 5
work_rate: Low/Low	body_type: Unique	real_face: Yes
preferred_foot: Left	player_tags: #Dribbler, #Distance Shooter, #FK Specialist, #Acrobat, #Clinical Finisher, #Complete Forward	player_traits: Finesse Shot, Long Shot Taker (AI), Playmaker (AI), Outside Foot Shot, Chip Shot (AI), Technical Dribbler (AI)
pace: 81.0	shooting: 89.0	passing: 90.0
dribbling: 94.0	defending: 34.0	physic: 64.0
attacking_crossing: 84	attacking_finishing: 90	attacking_heading_accuracy: 70
attacking_short_passing: 91	attacking_volleys: 88	skill_dribbling: 95
skill_curve: 93	skill_fk_accuracy: 93	skill_long_passing: 90
skill_ball_control: 93	movement_acceleration: 87	movement_sprint_speed: 76
movement_agility: 91	movement_reactions: 92	movement_balance: 95
power_shot_power: 86	power_jumping: 68	power_stamina: 70
power_strength: 68	power_long_shots: 91	mentality_aggression: 44
mentality_interceptions: 40	mentality_positioning: 93	mentality_vision: 94
mentality_penalties: 75	mentality_composure: 96	defending_marking_awareness: 20
defending_standing_tackle: 35	defending_sliding_tackle: 24	goalkeeping_diving: 6
goalkeeping_handling: 11	goalkeeping_kicking: 15	goalkeeping_positioning: 14
goalkeeping_reflexes: 8	goalkeeping_speed: nan	

Dans cette partie on voit bien que l'on reprend toutes les variables qui concernent les caractéristiques des performances sportives des joueurs. On a 7 catégories principales à savoir : l'attaque, compétences techniques, déplacement sur le terrain, force physique, mentalité, défense et enfin les habilités au poste de gardien qui elles-mêmes sont réparties en sous catégories plus détaillées. L'échelle des variables va de 1 à 100 sauf pour les variables 'weak_foot', 'skill_moves' et 'international_reputation' qui vont de 1 à 5.

Plus de détails sur les variables se trouvent dans les annexes.

2. Préparation des données

Avant d'entrer dans la phase de modélisation, il est crucial de prendre en compte divers scénarios et de réaliser certaines manipulations pour répondre aux exigences des modèles envisagés (Maharana, Mondal, & Bhushankumar, 2022). Cette étape, souvent appelée *data pre-processing*, demande beaucoup d'attention car elle reste généralement le facteur décisif pour garantir les performances des modèles d'apprentissage automatique (Kuhn & Kjell, 2013). Dans cette partie du travail, les différentes étapes de la phase du pre-processing seront revues.

Pour commencer, la base de données a été importée dans l'environnement Python afin de manipuler les données. La base de données brute pour l'année 2023, sans modifications, contient 166 674 joueurs et 110 variables. Une variable 'player_id' permet d'identifier chaque joueur. Cependant, une première analyse de la base de données révèle plusieurs incohérences, comme des doublons et des valeurs manquantes.

Dans cette partie du travail, plusieurs changements ont été apportés pour optimiser la préparation de la base de données en vue de la modélisation des prédictions.

2.1 Data integration

La data integration implique la manipulation des données dans le but de créer un ensemble de données unifié et cohérent (Luna Dong & Srivastava, 2019). Les incohérences observées dans le sous-ensemble sont causées principalement par l'absence de certaines données au vu de la confidentialité de celles-ci. Ce sont par exemple les données sur les clauses libératoires des joueurs et le salaire de certains joueurs.

Le but de ce travail est de comprendre, à partir d'une base de données comportant de nombreuses variables (110 dans ce cas), quelles sont les variables influençant le plus la note globale d'un joueur (la variable 'overall') et de prédire celle-ci. La base de données utilisée inclut les joueurs de 2015 à 2023. Le travail se concentre sur les joueurs offensifs, c'est-à-dire les attaquants, les ailiers et les milieux offensifs. Ce choix permet d'éliminer une série de variables indépendantes typiquement relatives au poste des joueurs, notamment les variables liées aux gardiens, aux défenseurs et aux milieux défensifs.

En incluant les mêmes joueurs pour chaque année, il existe un risque de redondance de données. Cela signifie que certaines observations peuvent être très similaires, voire identiques, d'une année à l'autre pour le même joueur. Cette situation peut potentiellement biaiser les modèles en surévaluant l'importance de certains joueurs ou en introduisant du bruit inutile dans les données. C'est pourquoi j'ai choisi de ne conserver que les données des joueurs de l'année 2023.

En résumé, dans la base de données utilisée, seuls les joueurs offensifs de la version 2023 du jeu ont été conservés. Comme mentionné précédemment, la distribution initiale des données a été maintenue intacte.

2.2 Data cleaning

Avant de procéder à la modélisation, il est essentiel d'évaluer le nombre de valeurs manquantes dans la base de données et de choisir comment les traiter afin de ne pas compromettre les résultats. Un traitement inadéquat de ces valeurs manquantes peut entraîner des conclusions erronées ou incomplètes (Chu, Ihab, Krishnan, & Wang, 2016).

Deux techniques de data cleaning sont envisageable : soit la suppression des valeurs manquantes, soit l'imputation d'une valeur à celles-ci sur base des données disponibles (Ridzuan & Mohd Nazmee Wan Zainon, 2019).

La première méthode, *l'exclusion*, est moins favorable, car supprimer des données peut introduire un biais dans le processus d'apprentissage et entraîner la perte d'informations essentielles (Ridzuan & Mohd Nazmee Wan Zainon, 2019).

C'est pourquoi la deuxième technique, l'imputation, est généralement privilégiée. La méthode d'imputation la plus courante consiste à remplacer les valeurs manquantes par des valeurs issues des données disponibles, telles que la moyenne ou la médiane (Ridzuan & Mohd Nazmee Wan Zainon, 2019). Cependant, pour les variables 'value_eur' et 'release_clause_eur' où des données manquaient, l'imputation par des techniques statistiques classiques peut ne pas être réaliste. En effet, ces données dépendent souvent de cas individuels et ne peuvent pas être simplement extrapolées à partir des autres observations.

Nettoyage manuel de variables :

Les variables suivantes n'ont pas de pouvoir explicatif : elles sont redondantes, telles que les noms des joueurs ou la nationalité, car ces informations sont déjà incluses sous d'autres formes, comme par exemple avec 'nationality_id'.

'player_url', 'fifa_update', 'fifa_update_date', 'long_name', 'short_name', 'league_name', 'club_name', 'nationality_name', 'player_face_url', 'player_tags', 'player_traits', 'player_positions'

Données incomplètes : Ces variables présentent des absences de données pour un nombre important de joueurs. J'ai choisi de retirer ces variables de l'analyse.

'club_joined_date', 'club_contract_valid_until_year', 'dob', 'club_loaned_from', 'nation_team_id', 'nation_position', 'release_clause_eur'

Variables relatives aux postes de gardiens ou de défenseurs : ces variables évaluent les performances à des postes auxquels il est très peu probable que les attaquants jouent.

'goalkeeping_speed', 'release_clause_eur', 'player_positions', 'real_face', 'goalkeeping_diving', 'goalkeeping_handling', 'goalkeeping_kicking', 'goalkeeping_positioning', 'goalkeeping_reflexes', 'lwb', 'ldm', 'cdm', 'rdm', 'rwb', 'lb', 'lcb', 'cb', 'rcb', 'rb', 'gk', 'ls', 'rs', 'st', 'lf', 'rf', 'lam', 'cam', 'ram', 'lm', 'lcm', 'cm', 'rcm', 'rm', 'lw', 'cf', 'rw' , 'potential'

Ces variables ont été supprimées dès le début pour plusieurs raisons expliquées ci-dessous.

J'ai choisi de conserver des joueurs de tous niveaux et continents afin d'avoir une représentation variée. Dans un premier temps, étant donné que je me concentre exclusivement sur les joueurs offensifs, j'ai décidé de retirer toutes les variables relatives au poste de gardien. En effet, les performances des attaquants ne sont pas évaluées en fonction de leurs compétences dans les

buts, par ailleurs ils ne peuvent pas utiliser leurs mains pendant le jeu. De même, j'ai éliminé les variables liées aux postes en défense, car il est rare de voir des attaquants jouer en défense, au vu du nombre de joueurs dans une équipe et de la tactique du football.

J'ai également supprimé les joueurs dont la valeur en euros (‘value_eur’) était égale à zéro ou pour lesquels il n'y avait pas de données disponibles. Suite à une analyse, il s'est avéré que ces joueurs étaient soit très âgés, soit qu'il manquait de nombreuses autres données les concernant.

En ce qui concerne la variable 'release_clause_eur', j'ai décidé de supprimer en raison de la difficulté d'accès à ces données. En effet, ces informations sont souvent peu disponibles publiquement, que ce soit pour les joueurs très connus ou moins connus.

De plus, j'ai identifié des variables répétitives qui nécessitaient d'être supprimées. Par exemple, la variable 'dob' (date de naissance) et 'age' contiennent des informations similaires. De même, les variables 'player_url', 'fifa_update' et 'fifa_update_date' ne sont pas liées aux performances des attaquants, mais plutôt aux mises à jour du jeu FIFA à partir desquelles les données ont été extraites.

Par ailleurs, les variables 'long_name' et 'short_name' sont redondantes car la base de données contient également 'player_id', que j'ai choisi de conserver pour identifier les joueurs lors du traitement des données. Pour la même raison j’ai enlevé ‘club_name’ et ‘nationality_name’.

En ce qui concerne les variables 'player_tags' et 'player_traits', elle ajoutait une description générique sur les joueurs. Ces descriptions n’apportaient pas de plus-value en termes d’information.

Après avoir effectué un premier traitement manuel des variables, j'ai effectué une analyse approfondie à l'aide d'une matrice de corrélation. Mon objectif était de comprendre les relations entre les variables indépendantes et d'identifier celles qui étaient fortement corrélées entre elles, ce qui pourrait indiquer une redondance ou un risque de multi colinéarité. J'ai également examiné les corrélations entre ces variables indépendantes et la variable explicative afin de déterminer celles qui avaient le plus de potentiel pour expliquer la variabilité de cette dernière.

La matrice de corrélation avec le coefficient de Pearson a été utilisée pour analyser les variables quantitatives. Le détail de la matrice de corrélation se trouve en annexe. La matrice de corrélation affiche uniquement les corrélations supérieures à 0.75.

Ainsi, le premier ensemble de variables fortement corrélées concerne celles faisant référence aux postes offensifs : 'ls', 'st', 'rs', 'lf', 'cf', 'rf', 'rw', 'lam', 'cam', 'ram'. La corrélation entre toutes ces variables est proche de 1. Initialement, j'ai envisagé de ne conserver qu'une seule variable, 'st', et de supprimer les autres. Cependant, même en gardant uniquement 'st', cette variable conserve une importance significative dans tous mes modèles. Cette observation n'est pas surprenante, car la note globale d'un joueur offensif est une approximation très similaire à sa note pour le poste de 'st' (attaquant). Pour conclure, j'ai éliminé toutes les variables relatives à la position.

De plus, j'ai identifié que les variables suivantes sont fortement corrélées entre elles, et j'ai donc décidé de les retirer :

"movement_reactions", "skill_ball_control", "mentality_positioning", "attacking_finishing", "potential", "attacking_short_passing", "attacking_volleys" ,"skill_dribbling", "power_shot_power" ,"power_long_shots", "pace", "mentality_composure", "passing", "shooting"," dribbling", "defending", "defending_sliding_tackle", "mentality_vision"

Pour finir, j'ai décidé de retirer la variable 'potential' en raison de sa construction. En effet, cette variable est définie comme suit :

Potential = overall + potentiel d'augmentation

où le *'potentiel d'augmentation*' varie de 1 à 4.

Il est évident que la variable 'potential', par sa définition même, est presque entièrement redondante par rapport à la variable 'overall'. Conserver la variable 'potential' introduirait une forte colinéarité, car 'potential' est essentiellement une version modifiée de 'overall' avec l'ajout d'une petite valeur fixe. Par conséquent, garder les deux variables dans l'analyse ne fournirait pas d'informations supplémentaires significatives et pourrait même compliquer l'interprétation des résultats du modèle en créant une dépendance artificielle. Pour ces raisons, j'ai décidé de supprimer la variable 'potential' de mon jeu de données afin de garantir l'intégrité et la simplicité de l'analyse.

2.3 Data transformation

2.3.1 Modification des variables

Avant de soumettre le sous-ensemble de données à un modèle prédictif, il faut le transformer afin que la forme des données satisfasse aux exigences du modèle (Qlik, 2022).

La transformation consiste en la numérisation des attributs catégoriques du sous-ensemble de données. Cette numérisation implique la transformation d'un attribut contenant n observations et x valeurs distinctes en x variables binaires, chacune comportant n observations. Chaque valeur binaire indique la présence (1) ou l'absence (0) de l'attribut correspondant. Une liste des attributs numérisés et de leur signification est disponible en annexe.

'colonnes_a_modifier' = ['ls', 'st', 'rs', 'lf', 'cf', 'rf', 'rw', 'lam', 'cam', 'ram'].

Pour toutes les valeurs comprises dans 'colonnes_a_modifier' elles sont composées de la manière suivante : note générale *'ls' = note de base + variable* allant de 1 à 4 ce qui se présente ainsi dans la base de données '65+3'. Le bonus maximal de 3 est attribué dans le jeu en prenant en compte la forme actuelle du joueur et les différentes tactiques mises en place par le coach. Dans le cadre de ce travail j'ai décidé de garder la note de base.

La variable 'work_rate' est présentée sous la forme « lazy/active », contenant en fait deux informations distinctes. La première sous-variable indique le *'work_rate' défensif* (lazy) et la deuxième contient le *'work_rate' offensif* (active). Pour capturer ces valeurs distinctes, deux colonnes ont été ajoutées à ma base de données afin de refléter correctement les efforts effectués dans le travail offensif et défensif du joueur.

Pour la variable 'club_position', j'ai choisi de regrouper les valeurs 'RES' et 'SUB', qui signifient respectivement réserviste ou remplaçant, car ces deux valeurs indiquent que le joueur n'est pas titulaire.

La standardisation des données est nécessaire lorsque les caractéristiques de l'ensemble de données présentent de grands écarts entre leurs valeurs ou sont mesurées dans des unités différentes.

Ces écarts dans les données des caractéristiques initiales posent des problèmes pour de nombreux modèles d'apprentissage automatique. Par exemple, pour les modèles basés sur le calcul de la distance, une caractéristique avec une large gamme de valeurs aura une influence disproportionnée sur la distance calculée. De nombreuses recherches démontrent qu'une caractéristique avec une échelle plus grande pourrait dominer les autres caractéristiques (Jaadi, 2022). Cela pourrait conduire à des modèles d'apprentissage automatique erroné (Peshawa Jamal & Faraj, 2014). De même, les valeurs extrêmes peuvent cacher les autres données d'une caractéristique, ce qui peut affecter la mise à l'échelle standard (Tanioka & Yadohisa , 2012).

2.3.2 Mise à l'échelle des données

Analysons plusieurs fonctions disponibles dans python :

$$\text{StandardScaler()}^4 : z = \frac{Xi - mean(x)}{stdev(x)} \qquad (2.1)$$

La standardisation est une méthode de mise à l'échelle qui rend les données indépendantes de leur échelle d'origine. Elle transforme la distribution statistique des données de telle sorte que la moyenne soit égale à 0 et l'écart type égal à 1 (scikit-learn).

$$\text{RobustScaler()}^5 : X_{robust} = \frac{Xi - Xi\ median}{IQR} \qquad (2.2)$$

Les valeurs aberrantes peuvent avoir un impact négatif sur la mise à l'échelle standard des données d'entrée. C'est pourquoi nous devrions plutôt utiliser une mise à l'échelle robuste. Contrairement à la mise à l'échelle standard qui utilise la moyenne et l'écart type, la mise à l'échelle robuste utilise la médiane et l'écart interquartile (IQR) à la place (Singh, 2022).

La mise à l'échelle robuste évalue la distance de chaque point de données par rapport à la médiane de l'ensemble d'entrée. Les valeurs mises à l'échelle auront leur médiane et leur IQR fixés respectivement à 0 et 1 (scikit-learn). Les valeurs aberrantes n'affectent pas la médiane car elle ne dépend pas de chaque valeur de la liste.

$$\text{MinMaxScaler()}^6 : X_{scaled} = \frac{(Xi - Xi\ (min))}{(Xi\ (max) - Xi\ (min))} \qquad (2.3)$$

Cette méthode ne fonctionne pas bien sur les données aberrantes. Si les données ne suivent pas une distribution uniforme ou sont fortement asymétriques, la mise à l'échelle MinMaxScaler peut ne pas être la meilleure option. Elle peut compresser les valeurs dans une plage très étroite, perdant ainsi des informations importantes (Sharma, 2022). Cette méthode est utile pour des algorithmes qui supposent que toutes les caractéristiques sont centrées autour de zéro et ont une variance de taille similaire (scikit-learn).

Puisqu'il n'existe pas de techniques spécifiques pour déterminer les meilleures méthodes de mise à l'échelle pour tous les ensembles de données, il est nécessaire d'expérimenter différentes approches à travers de multiples essais avec des algorithmes de machine learning (Sharma, 2022). De plus, pour chaque expérience particulière, la base de données sera

[4] Fonction disponible dans python
[5] Idem
[6] Idem

différente ; ainsi, la meilleure façon de développer le modèle optimal pour un ensemble de données spécifique est de tester différents algorithmes de machine learning en les associant à diverses approches de mise à l'échelle (Ahsan, Kumar Saha, & Siddique, 2021).

2.4 Data reduction

2.4.1 Fléau de la dimension

Étant donné le nombre considérable de variables initialement présentes dans ma base de données et suite au processus de transformation des données, cela peut poser des défis lors de l'application des modèles prédictifs.

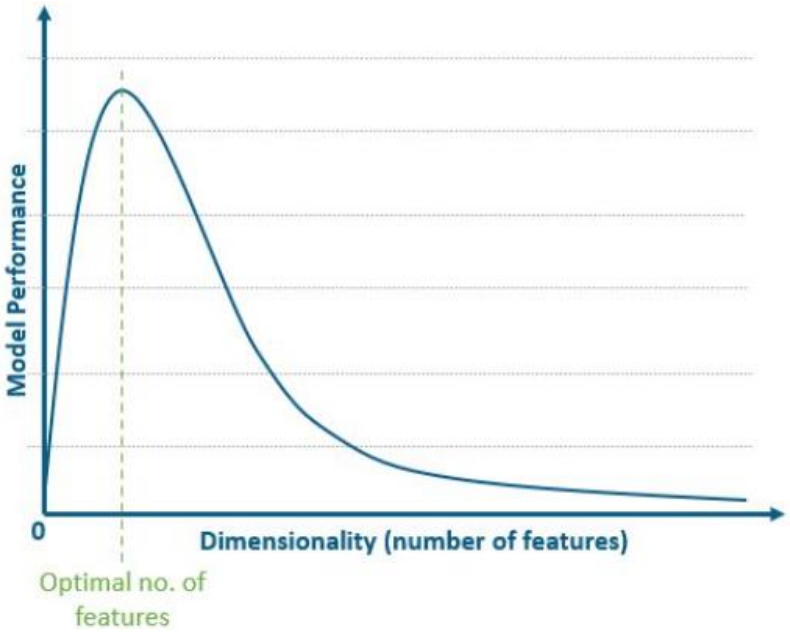


Figure 4: Limitation des performances des modèles en fonction de la dimensionalité (Shetty, 2021)

L'apprentissage automatique excelle dans l'analyse de données comportant de nombreuses dimensions, mais il devient plus difficile de créer des modèles significatifs à mesure que le nombre de dimensions augmente.

Le graphique montre que lorsque le nombre de caractéristiques augmente, la performance du classifieur augmente également jusqu'à ce que nous atteignions le nombre optimal de caractéristiques (Shetty, 2021). Ajouter davantage de caractéristiques basées sur la même taille que l'ensemble d'entraînement dégradera ensuite la performance du model.

Au vu du grand nombre de dimension, les modèles peuvent faire face au problème du surajustement. Il y a un risque de produire un modèle qui pourrait être très performant pour prédire la classe cible sur l'ensemble d'entraînement mais échouer misérablement lorsqu'il est confronté à de nouvelles données (Márquez, 2022). Ce qui veut dire que le modèle n'a pas le pouvoir de généralisation. C'est pourquoi il est si important d'évaluer les modèles sur des données jamais vues auparavant.

Pour gérer les espaces de décision, il est nécessaire d'identifier les variables qui n'apportent pas d'informations pertinentes mais qui complexifient le travail des modèles. Le processus de traitement de ces variables est appelé "dimentionality reduction", et peut être mis en place via la *sélection* des caractéristiques ou la *réduction* des caractéristiques (Jovic, Brkic, & Bogunovic).

La méthode de réduction de caractéristiques nuit à l'interprétabilité des données sélectionnées (Altman & Krzywinski , 2018). En effet le nouveau groupe de variables sélectionné est une combinaison linéaire des données de base. C'est pourquoi dans ce travail c'est la méthode de feature selection qui sera utilisée. La méthode qui a été gardée est la wrapper method avec comme modèle d'évaluation le Random Forest.

2.4.2 Recursive feature elimination

Il s'agit d'une méthode de sélection de fonctionnalités basée sur un 'wrapper'. Elle élimine les attributs redondants dont la suppression a le moins d'impact sur l'erreur d'entraînement, tout en conservant les fonctionnalités indépendantes et importantes pour améliorer les performances de généralisation du modèle. Cette méthode utilise une procédure itérative de classement des attributs, qui est un exemple d'élimination de fonctionnalités en arrière. Elle commence par construire le modèle en utilisant toutes les fonctionnalités, puis classe les attributs en fonction de leur importance (Misra & Yadav, 2020). Ensuite, elle supprime l'attribut le moins important, reconstruit le modèle et recalcule l'importance des attributs.

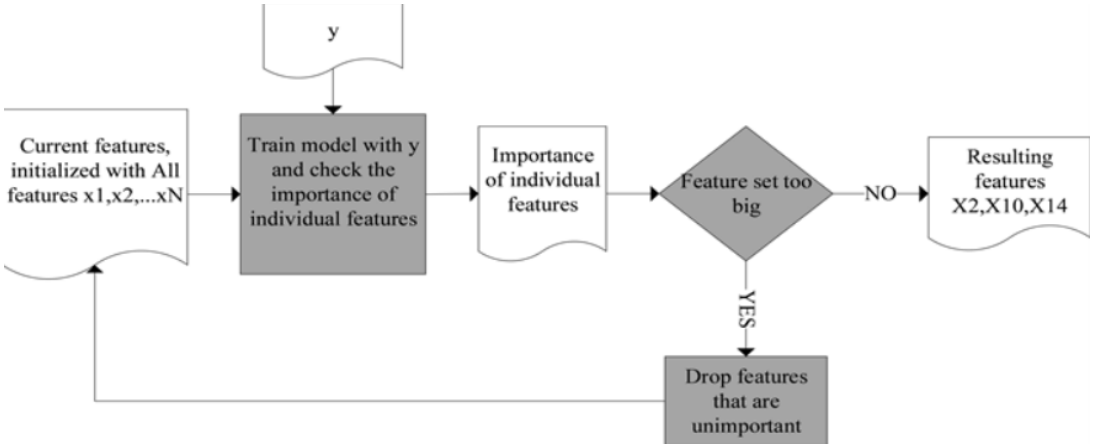


Figure 5: Illustration du fonctionnement de l'élimination récursive de caractéristiques (Hou, 2020)

Dans le cadre de ce travail, le modèle utilisé pour estimer les sous ensemble de données est appelé la forêts aléatoire (Random Forest). Ce modèle est détaillé dans la section modélisation. Dans cette partie du travail plusieurs raisons sont mises en avant afin de justifier le choix du modèle.

Les forêts aléatoires sont robustes aux valeurs aberrantes dans les données. Les valeurs aberrantes ont moins d'impact sur les performances du modèle car les arbres de décision divisent les données en fonction de sous-ensembles de caractéristiques et sont moins sensibles aux points de données individuels. De plus, dans une forêt, chaque arbre est entraîné sur un sous-ensemble aléatoire des données, ce qui réduit encore l'impact des valeurs aberrantes (Kumar & Gandhi, 2021).

De plus, les forêts aléatoires sont moins sujettes au surajustement par rapport aux arbres de décision individuels. Chaque arbre de la forêt est entraîné sur un sous-ensemble (méthode de bootstrap) des données, et à chaque division, il ne considère qu'un sous-ensemble aléatoire de caractéristiques (Kumar & Gandhi, 2021). Cette sélection aléatoire aide à décorréler les arbres et les empêche de simplement mémoriser les données d'entraînement.

Il est important de préciser qu'il n'existe pas de méthode de sélection de fonctionnalités universelle adaptée à chaque ensemble de données. Chaque ensemble de données possède des caractéristiques uniques qui influencent les mécanismes de fonctionnement de la méthode

de sélection des fonctionnalités. Par conséquent, plusieurs méthodes de sélection de fonctionnalités doivent être testées. De plus, il n'y a aucune relation entre les régressions et les méthodes de sélection de fonctionnalités : on ne sait pas quelle méthode de sélection de fonctionnalités est la meilleure pour KNN[7] par exemple. Par conséquent, il est nécessaire de vérifier les combinaisons de régressions et de méthodes de sélection de caractéristiques afin de construire un modèle de régression performant (Jeon & Oh, 2020).

2.4.3 Lasso features selection

Une autre méthode testée dans ce travail est la sélection de variables basée sur la régularisation Lasso[8]. Cette technique implique l'ajout d'un terme de pénalité à la fonction de coût de la régression linéaire, ce qui contraint la somme des valeurs absolues des coefficients des variables à rester en dessous d'un certain seuil. En conséquence, Lasso tend à réduire certains coefficients à zéro, permettant ainsi une sélection automatique des variables les plus importantes. Cette méthode est particulièrement utile pour gérer des ensembles de données avec un grand nombre de variables et pour prévenir le surapprentissage en simplifiant le modèle.

3. Modélisation

3.1 Types de modèles utilisés

Le sous-ensemble est ensuite introduit dans des modèles. Dans ce travail des modèles prédictifs sont développé, mais nous nous concentrons également sur le fait de rendre ces prédictions faciles à comprendre, ce qui fait partie de la modélisation descriptive. Les modèles prédictifs utilisés sont de l'intelligence artificielle et plus précisément du *'supervised traditional machine learning*' du type white-box dans un objectif de prédire une variable continue ((El Morr & Ali-Hassan , 2019), (Loyola-González, 2005), (Unit 21, 2024)). Ces modèles sont caractérisés par leur entraînement sur des données dites étiquetées, ce qui signifie que les valeurs de la variable dépendante sont connues. Ainsi, les modèles sont entraînés sur l'ensemble de données d'entraînement afin de découvrir des patterns qu'ils pourront ensuite appliquer à des données inconnues.

Selon les recherches (Molnar, 2023) quatre modèles peuvent s'y retrouver dont les modèles linéaires généralisés, les arbres de décision, les K-nearest Neighbors (KNN) et les Naive Bayes.

3.2 Mesure de performances

En vue d'optimiser les modèles, plusieurs mesures de performances ont été considérées : la MAE (Mean Absolute Error), la RMSE (Root Mean Square Error), le coefficient de détermination R^2, le critère d'information d'Akaike (AIC) et le Critère d'information bayésien (BIC).

3.2.1 MAE

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \qquad (3.1)$$

[7] KNN est un modèle de régression qui est détaillé dans la partie modélisation.
[8] La régression de Lasso est détaillée dans la partie modélisation

L'erreur absolue $|\boldsymbol{y} - \boldsymbol{\hat{y}}|$ représente la quantité d'erreur présente dans les mesures, mesurée par la différence entre la valeur mesurée et la valeur "vraie". Utiliser le symbole de la valeur absolue est nécessaire, car parfois la mesure peut être plus petite, donnant ainsi un nombre négatif. L'erreur moyenne absolue (MAE) est la moyenne de toutes ces erreurs absolues (J.P. Keating, 1985).

En d'autres termes, la MAE évalue l'amplitude moyenne des erreurs de prédiction, sans prendre en compte leur direction (Everitt & Skrondal, 2010). De plus le MAE est moins sensible aux valeurs aberrantes que le RMSE. Elle reflète l'échelle des données prédites, dans le cadre de ce travail, elle représentera donc la différence entre la note « overall » prédite et la réalité, sans accorder de poids supplémentaire aux erreurs plus importantes.

3.2.2 MSE et RMSE

$$MSE = \frac{1}{n}\sum_{i=1}^{n} e_i^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad (3.2)$$

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \qquad (3.3)$$

Le MSE (Mean Squared Error) est une mesure statistique qui évalue l'ampleur moyenne des différences entre les valeurs prédites et observées. Le MSE est pratique pour entraîner les modèles car il est facile à optimiser. C'est parce que sa dérivée est simple à calculer, ce qui permet de mettre à jour les paramètres du modèle plus efficacement lors de l'entraînement avec des algorithmes comme la descente de gradient. Contrairement au MAE, le MSE utilise le carré des écarts, ce qui le rend plus sensible aux valeurs aberrantes. Cependant, l'utilisation de cette mise au carré rend les résultats moins intuitifs, car les unités des données deviennent différentes et moins faciles à interpréter (Jain, 2016).

Le RMSE est particulièrement sensible aux valeurs aberrantes dans un ensemble de données. Lorsque ces valeurs extrêmes sont présentes, elles peuvent avoir un effet disproportionné sur le calcul du RMSE, ce qui peut fausser la mesure de performance du modèle de régression (Jain, 2016). Il est important d'identifier et de traiter ces valeurs aberrantes de manière appropriée avant de calculer le RMSE. En négligeant ces valeurs, un risque de fausser considérablement l'estimation de l'erreur globale du modèle se présente. C'est pourquoi au préalable une étape de nettoyage de données a eu lieu. L'avantage du RSME qui est la racine carrée du MSE est que les données deviennent à nouveau plus facilement interprétables car on regagne les mêmes unités. Tout comme le MSE, le RMSE est pratique pour entraîner les modèles car il est facile à optimiser.

3.2.3 Le coefficient de détermination

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y_i})^2}{\sum_{i=1}^{n}(y_i - \bar{y_i})^2} \qquad (3.4)$$

Le coefficient de détermination, R^2, indique quelle partie de la variation de la variable dépendante peut être expliquée par la variation de la variable indépendante dans un modèle de régression. En d'autres termes, il mesure à quel point les valeurs de la variable dépendante peuvent être prédites à partir des valeurs de la variable indépendante utilisées dans le modèle. Le coefficient de détermination R^2 prend des valeurs entre 0 (la variance par rapport à la moyenne n'est pas expliquée par le modèle) et 1 (l'entièreté de la variance est expliquée par le modèle) (Komentar, 2022).

Dans les modèles linéaires OLS, les coefficients sont déterminés en minimisant la somme des carrés des résidus (SSres). Bien que le R^2 soit un excellent moyen d'évaluer la performance d'un modèle, il est important de noter qu'il a tendance à augmenter avec le nombre de variables indépendantes (Komentar, 2022).

L'augmentation naturelle du R^2 à mesure d'ajouter davantage de variables indépendantes à un modèle peut être trompeuse (Leach & Henson , 2007). En effet, cette augmentation ne garantit pas nécessairement une amélioration de la capacité du modèle à expliquer la variance de la variable dépendante. Par conséquent, le R^2 seul peut ne pas fournir une indication précise de la qualité du modèle, surtout lorsqu'il est comparé entre des modèles avec un nombre différent de variables indépendantes.

3.2.4 Le coefficient de détermination ajusté

Pour remédier à ce problème, il existe une version ajustée du coefficient de détermination. Cette formule prend en compte le nombre d'observations et le nombre de variables indépendantes.

$$R^2_{ajusté} = 1 - \left[\frac{(n-1)}{(n-k-1)}\right] \times (1 - R^2), \qquad (3.5)$$

où $\mathrm{n} = \text{nombre d'observations dans l'échantillon}$, et $\mathrm{k} = \text{nombre de variables indépendantes}$

En considérant le calcul du R^2 et puisque n est supérieur à k, à mesure que des variables sont ajoutées au modèle, le quotient entre parenthèses devient plus grand. On comprend donc que la formule est construite pour ajuster et pénaliser l'inclusion de coefficients dans le modèle (Leach & Henson , 2007).

En plus de l'avantage du coefficient de détermination, l'ajustement utilisé dans la formule précédente nous permet également de comparer des modèles avec différents nombres de variables indépendantes. Encore une fois, la formule ajuste le nombre de variables entre un modèle et un autre et nous permet de faire une comparaison homogène.

3.2.5 Critère d'information d'Akaike

$$AIC = -2 \times \log(L) + 2\,k, \qquad (3.6)$$

où $\mathrm{L} = \text{la vraisemblance maximisée}$, et $\mathrm{k} = \text{nombre de variables indépendantes}$

L'AIC utilise l'estimation du maximum de vraisemblance (log-vraisemblance) d'un modèle comme mesure d'ajustement. La log-vraisemblance est une mesure de la probabilité d'observer les données d'un modèle. Le modèle avec la vraisemblance maximale est celui qui "ajuste" le mieux les données (Lancelot & Lesnoff, 2005).

Le maximum de vraisemblance est une méthode statistique utilisée pour estimer les paramètres d'un modèle de probabilité de manière à maximiser la probabilité d'observer les données. Autrement dit, c'est une méthode statistique permettant de trouver les paramètres d'un modèle de probabilité les plus "vraisemblables" pour expliquer des données observées. Cela permet de comparer plusieurs modèles et de déterminer lequel ajuste le mieux les données.

L'AIC est faible pour les modèles ayant des log-vraisemblances élevées. Cela signifie que le modèle ajuste mieux les données, ce qui est souhaitable (Zajic, 2022). Cependant, l'AIC ajoute un terme de pénalité pour les modèles avec une complexité de paramètres plus élevée, car plus de paramètres signifient qu'un modèle est plus susceptible de surajuster les données d'entraînement (Zajic, 2022).

En résumé, le critère d'information d'Akaike est un score numérique unique qui peut être utilisé pour déterminer lequel parmi plusieurs modèles est le plus susceptible d'être le meilleur modèle pour un ensemble de données donné. Ce critère compare les modèles de manière relative, ce qui signifie que les scores AIC ne sont utiles qu'en comparaison avec d'autres scores AIC pour le même ensemble de données. Un score AIC plus bas est préférable.

3.2.6 Critère d'information bayésien

$$BIC = -2 \times log(L) + k \log(n), \qquad (3.7)$$

où $\mathrm{L} = \text{la vraisemblance maximisée}$, $\mathrm{k} = \text{nombre de variables indépendantes}$, et

$\mathrm{n} = \text{nombre d'observations}$

On remarque que, tout comme l'AIC, le BIC se base également sur le logarithme du maximum de vraisemblance. Cependant, le terme de pénalisation diffère de celui de l'AIC. En effet, le BIC inclut une pénalité proportionnelle à $\log(\mathrm{n})$. Cela signifie que le BIC prend en compte non seulement la qualité de l'ajustement du modèle, mais aussi la complexité du modèle en termes de nombre de paramètres (Stata, 2019). En conséquence, pour de grands ensembles de données, la pénalité pour chaque paramètre supplémentaire est plus élevée, favorisant ainsi les modèles plus simples (Stata, 2019).

En d'autres termes, plus il y a de données pour entraîner les modèles, plus il est probable que le BIC identifiera correctement le modèle qui reflète fidèlement les relations dans les données.

En résumé, l'évaluation d'un modèle de régression ne devrait pas se limiter à une seule mesure de performance, il est important de prendre en compte plusieurs critères pour obtenir une compréhension approfondie de sa performance.

3.3 Validation croisée - Cross validation

Après avoir entraîné un modèle de Machine Learning sur des données étiquetées, il est important de vérifier ses performances sur de nouvelles données pour garantir sa précision. Pour évaluer ces performances, on utilise généralement la technique de la validation croisée, qui est une méthode de rééchantillonnage permettant de tester un modèle même avec des données limitées (Kofi Nti, Nyarko-Boateng, & Aning, 2021).

Le but principal reste de surveiller le surapprentissage (l'overfit) du modèle de prédiction, c'est-à-dire lorsque le modèle surperforme sur les données d'entrainement (étiquetées) mais qu'il ne

sait pas généraliser ses performances sur de nouvelles données de test (non étiquetées). Il va capturer chaque tendance individuelle, en manquant la tendance principale (Kofi Nti, Nyarko-Boateng, & Aning, 2021). La figure 6 ci-dessous illustre bien les 3 cas qui peuvent apparaitre. Le 3ème cas « d'overfit » illustre bien le sur ajustement de la régression.

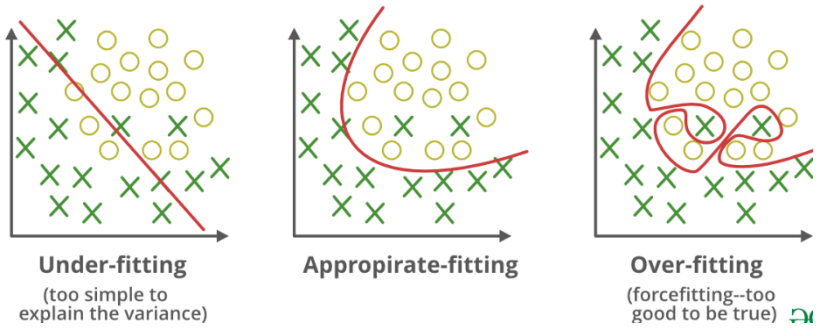


Figure 6 : Illustration des différents ajustements (MathWorks)

La validation croisée implique de diviser les données en ensembles distincts, généralement appelés 'folds', de manière aléatoire. Ensuite, le modèle est entraîné sur une partie de ces folds (K-1) et testé sur le 'fold' restant. Ce processus est répété jusqu'à ce que chaque 'fold' ait été utilisé comme ensemble de test (Naujoks, Medium, 2019). Les performances du modèle sont évaluées en calculant les scores et les erreurs pour chaque 'fold', puis en prenant la moyenne de ces scores pour obtenir une métrique de performance globale (Naujoks, Medium, 2019). Si un modèle surapprend les données d'entraînement, cela se reflétera par une grande variance entre les performances sur les différents folds. La figure 7 illustre bien ce principe.

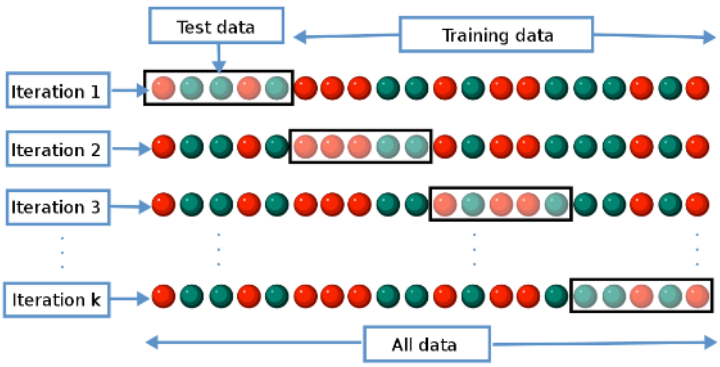


Figure 7 : Fonctionnement de la validation croisée (Naujoks, Medium, 2019)

L'approche de validation croisée est une méthode robuste pour évaluer les performances des modèles, car elle permet de réduire les biais liés à la manière dont les données sont divisées en ensembles d'entraînement et de test. En comparaison, l'approche Train-Test Split consiste à diviser les données en un seul ensemble d'entraînement et un seul ensemble de test, ce qui peut entraîner une estimation moins fiable des performances du modèle, notamment si les données sont limitées. Il est important d'utiliser les deux techniques ensemble pour obtenir une évaluation plus complète et fiable des performances du modèle (Peshawa Jamal & Faraj, 2014).

3.4 Sélection de variables

En règle générale, un modèle complexe risque de sur-ajuster lorsque son erreur sur les données d'entraînement est faible tandis que son erreur sur les données de test est élevée. Pour

comprendre le surapprentissage, on peut décomposer l'erreur de généralisation en biais et variance. Le biais représente la tendance d'un modèle à systématiquement apprendre les mêmes erreurs, tandis que la variance représente sa tendance à apprendre des fluctuations aléatoires qui ne correspondent pas au modèle réel (Domingos, 2000).

Le moyen le plus populaire de lutter contre le surapprentissage est la régularisation. En pénalisant les modèles avec une plus grande capacité et en favorisant ceux plus simples avec moins de paramètres, le surajustement peut être contrôlé.

Néanmoins, d'après plusieurs recherches ((Kotsilieris , Anagnostopoulos, & E. Livieris , 2022), (Tian & Zhang, 2022)), nous ne pouvons pas être sûrs qu'une méthode de régularisation particulière résout totalement le problème de surapprentissage. Il est facile d'éviter le surajustement en tombant dans le problème inverse, c'est-à-dire le sous-ajustement. Pour éviter simultanément les deux, il faut prendre un régresser en fonction du problème en question, ce qui, bien sûr, n'est pas connu à l'avance (Kotsilieris , Anagnostopoulos, & E. Livieris , 2022). C'est pourquoi aucune technique ne fonctionnera toujours parfaitement. Il est donc essentiel d'utiliser la régularisation avec précaution pour chaque problème et d'expérimenter.

Une généralisation correcte devient d'autant plus difficile à mesure que la dimensionnalité des exemples augmente en raison de la 'malédiction de la dimensionnalité'. De nombreux algorithmes qui fonctionnent bien dans de faibles dimensions deviennent intraitables lorsque l'entrée est de grande dimension. En grandes dimensions, les caractéristiques non pertinentes posent de nombreux problèmes dans le raisonnement basé sur la similarité, comme les régresser du plus proche voisin (KNN). À mesure que la dimensionnalité augmente, de plus en plus d'exemples deviennent les plus proches voisins d'un échantillon typique. On pourrait penser que rassembler plus de fonctionnalités ne fait jamais de mal puisqu'au pire, elles ne fournissent aucune nouvelle information sur la classe. Cependant, leurs avantages peuvent être contrebalancés par la malédiction de la dimensionnalité.

3.4.1 Filtre de corrélation

Une première étape simple qui permet de déjà avoir un aperçu de la redondance des variables est l'analyse de corrélation des variables indépendantes. Cela peut entraîner une surpondération de l'importance de ces variables dans les modèles, ce qui peut biaiser les prédictions. De plus dans certains modèles comme la régression linéaire, des variables fortement corrélées peuvent rendre les coefficients instables (Saporta, 2011). De petites variations dans les données peuvent entraîner des changements importants dans les coefficients des variables. Pour finir l'impact individuel des coefficients n'est plus facile à interpréter.

Une étape d'identification des paires de variables fortement corrélées, d'exclusion ou de combinaison en une seule variable peut déjà aider à réduire le nombre de variables.

Dans ce travail, si des variables sont corrélées à plus de 85%, seule celle qui possède la plus grande interprétabilité et de sens pour la prédiction sera conservée.

Comme le montre la matrice de corrélation en annexe, il est possible de repérer plusieurs variables corrélées. Par exemple, toutes les variables relatives à la position du joueur présentent une forte corrélation. Tous les détails se trouvent dans la section 2.4 Data Cleaning.

3.4.2 Forward features selection - Random Forest

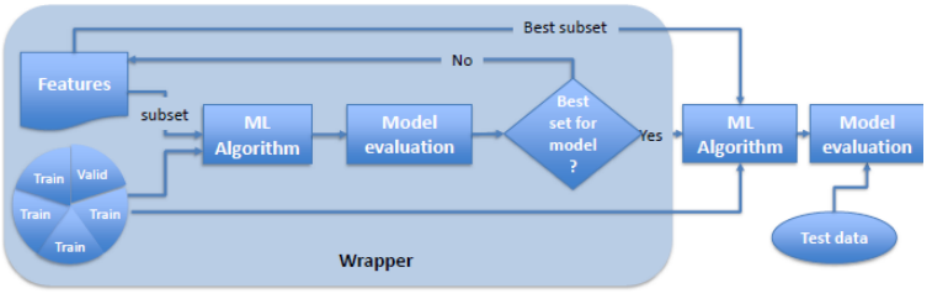


Figure 8 : Illustration de la méthode Wrapper (Hsu & Hsieh, 2011)

Dans un premier temps il est intéressant de comprendre le principe du 'wrapper selection method' (Chen & Jin, 2015).

La méthode wrapper aborde la sélection des ensembles de variables comme un problème de recherche, où différentes combinaisons sont préparées, évaluées et comparées les unes aux autres. Elle utilise un modèle prédictif pour évaluer chaque combinaison de fonctionnalités et assigner des scores de performance à ces combinaisons (Chen & Jin, 2015). Ce processus est illustré dans la *figure 8*. Nous commençons avec un ensemble de variables qui est testé par un modèle. À chaque itération, le modèle évalue un sous-ensemble de données. Si l'évaluation s'améliore à chaque test, le processus continue. Cependant, le processus est interrompu dès que l'évaluation du sous-ensemble de données commence à diminuer.

Deux techniques de sélection existent (Bee Wah & Ibrahim, 2018) :

- Forward selection : Dans ce cas on commence avec un modèle vide et ajoute progressivement une à une les variables les plus significatives. Le processus continue jusqu'à ce qu'aucune autre variable ne puisse améliorer la performance du modèle.

- Backward selection : Dans ce cas on commence avec un modèle contenant toutes les variables. À chaque étape, une variable est retirée du modèle. Le processus continue jusqu'à ce qu'aucune autre variable ne puisse améliorer la performance du modèle.

Les méthodes d'enveloppe peuvent utiliser une recherche intégrale pour évaluer toutes les combinaisons possibles de caractéristiques, mais elles peuvent aussi utiliser des méthodes empiriques pour explorer l'espace des caractéristiques de manière plus efficace. Peu importe la méthode utilisée, l'évaluation de la performance du modèle reste le critère principal pour sélectionner les caractéristiques les plus pertinentes.

Dans le cadre de ce travail, j'ai choisi d'implémenter l'algorithme RandomForestRegressor() comme modèle d'évaluation du wrapper.

Les « wrapper methods » en sélection de caractéristiques sont des outils performants, utilisés dans le domaine de l'apprentissage automatique pour identifier les caractéristiques les plus pertinentes dans un ensemble de données. Contrairement aux méthodes de filtrage qui évaluent les caractéristiques de manière indépendante, les méthodes d'enveloppe utilisent un modèle machine learning pour évaluer l'impact des caractéristiques sur la performance globale du modèle (Chen & Jin, 2015).

Les avantages des méthodes wrapper résident dans leur capacité à prendre en compte les interactions entre les caractéristiques et la tâche d'apprentissage, ce qui peut améliorer les

performances du modèle. Cependant, elles nécessitent plus de puissance de calcul, surtout avec de grands ensembles de données, et elles risquent de surapprendre si le modèle devient trop complexe ou si les données d'entraînement sont insuffisantes.

En résumé, les méthodes wrapper offrent une approche efficace et flexible pour la sélection de caractéristiques, en permettant aux modèles d'apprentissage automatique de choisir les caractéristiques les plus pertinentes pour résoudre une tâche spécifique. Elles sont largement utilisées dans de nombreux domaines de l'apprentissage automatique pour améliorer la précision des modèles et faciliter l'interprétation des données.

3.5 Modèles machine learning

3.5.1 Régression linéaire

C'est l'un des modèles les plus basiques mais aussi des plus simples pour prédire un résultat en utilisant une fonction linéaire sur des données d'entrée. Au-delà de la simple régression linéaire qui cherche à estimer la relation entre deux variables, la régression linéaire multiple calcule les poids de plusieurs variables explicatives pour prédire la variable dépendante.

Pour un modèle impliquant un ensemble de n variables explicatives, la relation est représentée par :

$$\hat{y} = \beta_0 + \sum_{i=1}^{n} \beta_i x_i \qquad (3.1)$$

où $\hat{y}$ est la prédiction de la variable dépendante, β_0 est la constante et β_i est le coefficient des variables indépendantes.

La méthode des moindres carrés ordinaires (OLS) vise à optimiser les coefficients et la constante tout en minimisant la fonction de coût, selon la formule suivante, pour un jeu de données contenant un nombre m de lignes :

$$\sum_{i=1}^{m}(y_i - \hat{y_i}) = \sum_{i=1}^{m}\left(y_i - \sum_{j=0}^{n} \beta_i x_{ij}\right)^2 \qquad (3.2)$$

Avant d'appliquer la régression linéaire aux données, elles seront normalisées pour des raisons de mise à l'échelle et d'interprétation.

3.5.2 Régularisation de Ridge et Lasso

La régularisation ajoute un terme à la fonction de régression qui pénalise l'ampleur de la valeur des coefficients de régression β_j :

$$Min\ RSS(M) = \sum_{i=1}^{n}(y_i - \hat{y_i})^2 \text{ où n = nombre d'observations,}$$

$$\text{et } \hat{y_i} = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} \text{ où k = nombre de variables indépendantes}$$

$$\rightarrow Min\ Reg(M) = \sum_{i=1}^{n}(y_i - \hat{y_i})^2 + \lambda R(\beta_i) \quad \text{terme de régularisation} \qquad (3.3)$$

λ est un terme qui contrôle le degré de pénalisation

Les deux principales fonctions de régularisation utilisées : Ridge et Lasso

- Ridge : Utilise la norme L_2 du β_i $\longrightarrow$ $R = L^2 = \|\beta_j\|_2^2 = \sum_{j=1}^{k} \beta_j^2$ (3.4)
- Lasso : Utilise la norme L_1 du β_i $\longrightarrow$ $R = L^1 = \|\beta_j\|_1 = \sum_{j=1}^{k} |\beta_j|$ (3.5)

3.5.2.1 Rigde (L2 regularisation)

La régression Ridge est une méthode d'estimation de paramètres populaire utilisée pour résoudre le problème de colinéarité qui survient fréquemment dans la régression linéaire multiple. La régression Ridge ajoute un terme de pénalité à la procédure d'estimation des moindres carrés ordinaires (OLS). Comme l'illustre clairement la formule 3.4, la pénalité est $\sum_{j=1}^{k} \beta_j^2$ (Arashi, Roozbeh, Gasparini, & Hamzah, 2021). Dans la formule ci-dessus 3.3, si λ est nul, alors nous obtenons OLS. Ce terme de pénalité pénalise les coefficients élevés, qui ont tendance à se produire en cas de multicolinéarité. En pénalisant les coefficients élevés, la régression Ridge réduit l'impact de la multicolinéarité sur les estimations des coefficients.

Dans la modélisation de régression simple, la présence de multicolinéarité conduit de manière évidente à des estimations de paramètres incohérentes. Les moindres carrés ordinaires (OLS), conduisent à un modèle inexact et instable car il n'est pas robuste au problème de multicolinéarité (Arashi, Roozbeh, Gasparini, & Hamzah, 2021). Selon OLS dans l'ensemble, un seul ensemble de bêtas est trouvé, ce qui donne la somme résiduelle des carrés (RSS) la plus basse.

Le meilleur modèle est-il celui qui a le RSS le plus bas ?

Dans le mot « non-biaisé », il faut également prendre en compte le « biais ». Le biais est la différence entre la prédiction moyenne du modèle et la valeur correcte que l'on essaie de prédire (Arif , 2019). Premièrement, le modèle sans biais trouve la relation entre les caractéristiques et la variable dépendante, tout comme le fait la méthode OLS. Ce modèle ajustera les observations de manière à minimiser entièrement le RSS. Cependant, les conséquences pourraient ne pas être favorables et des problèmes de surapprentissage pourraient survenir. Dans l'ensemble, le modèle ne fonctionnera pas correctement avec le nouvel ensemble de données (Arif , 2019).

La régression Ridge introduit un biais en réduisant les coefficients, mais elle réduit également la variance des estimations. L'ampleur du retrait est contrôlée par un paramètre de réglage, ici noté λ (lambda). L'augmentation de λ augmente le retrait et, par conséquent, augmente le biais mais diminue la variance. Ce compromis permet à la régression Ridge de trouver un équilibre entre biais et variance, conduisant finalement à des prédictions plus stables et plus précises, notamment en présence de multicolinéarité (Arif , 2019).

La régression Ridge est préférable lorsque toutes les caractéristiques sont supposées pertinentes ou lorsqu'un ensemble de données présente une multicolinéarité, car elle peut gérer plus efficacement les entrées corrélées en répartissant les coefficients entre elles.

Étant donné la nature de la pénalisation β_j^2 les coefficients non signifiants sont amenés à être proche de zero mais jamais égal à 0.

3.5.2.2 Lasso (L1 régularisation)

La régularisation Lasso est couramment utilisée en apprentissage automatique pour traiter des ensembles de données volumineux, car elle facilite la sélection automatique des caractéristiques. Pour ce faire, elle ajoute un terme de pénalité à la somme des carrés des résidus (RSS), $\sum_{j=1}^{k} |\beta_j|$, qu'elle multiplie ensuite par un paramètre de régularisation λ (formule 3.3). Ce paramètre de régularisation contrôle le degré de régularisation appliqué. Des valeurs plus élevées de λ augmentent la pénalité, réduisant davantage les coefficients vers zéro. Par

conséquent, certaines caractéristiques du modèle deviennent moins importantes voire sont éliminées, ce qui permet une sélection automatique des caractéristiques. À l'inverse, des valeurs plus faibles de lambda réduisent l'impact de la pénalité, préservant ainsi plus de caractéristiques dans le modèle (Kumar D. , 2024).

La régularisation Lasso favorise la simplicité du modèle, ce qui peut aider à prévenir les problèmes de multicolinéarité et de surajustement dans les ensembles de données (Kavlakoglu, 2024). La multicolinéarité se produit lorsque des variables indépendantes sont fortement corrélées, ce qui complique la modélisation causale. Les modèles surajustés ont du mal à généraliser sur de nouvelles données, réduisant ainsi leur utilité. En réduisant certains coefficients de régression à zéro, la régression Lasso peut efficacement éliminer les variables indépendantes du modèle, évitant ainsi ces problèmes potentiels lors de la modélisation. De plus, cette simplicité peut améliorer l'interprétabilité du modèle par rapport à d'autres techniques de régularisation comme la régression Ridge (ou régularisation L2).

Néanmoins contrairement à la régression de Ridge, le lasso a tendance à choisir une seule variable parmi un ensemble de variables lorsque les corrélations par paires sont toutes très élevées. Il est également indépendant de la variable choisie (Hebiri & Lederer, 2012).

3.5.3 Elastic Net

$$\sum_{i=1}^{M}(y_i - \hat{y}_i) + \lambda\left(\frac{1-\alpha}{2}\sum_{j=1}^{m}\widehat{\beta_j^2} + \alpha\sum_{j=1}^{m}\left|\widehat{\beta}_J\right|\right) \qquad (3.6)$$

ElasticNet combine les avantages de Ridge et de Lasso en introduisant un terme de pénalité qui est une combinaison pondérée des pénalités $L1$ et $L2$. Le paramètre α contrôle le poids attribué à chaque pénalité. Quand α est égal à 1, ElasticNet fonctionne comme la régression Lasso, tandis que quand il est égal à 0, c'est comme la régression Ridge. En ajustant α entre 0 et 1, l'ElasticNet peut trouver un compromis entre la sélection de variables et la réduction des coefficients, offrant ainsi une technique de régularisation flexible et efficace. C'est-à-dire que le modèle Elastic Net peut gérer la multicolinéarité et effectuer simultanément la sélection de caractéristiques (Zou & Hastie, 2005).

S'il existe de grandes corrélations entre les variables prédictives dans des circonstances typiques n > p, il a été démontré empiriquement que les performances de prédiction du Lasso sont dominées par la Ridge régression (Rakotomalala, 2019).

3.5.4 k-nearest neighbors regressor (KNN)

La méthode des k plus proches voisins est caractérisée par le terme d'apprentissage paresseux (lazy learning), dans le sens où l'algorithme ne construit pas de modèle prédictif à proprement parler, mais se base sur l'ensemble des données d'entraînement pour déduire les valeurs des prédictions pour l'ensemble de validation (Zhang Z. , 2016).

La classification KNN comporte au moins deux points importants (Luo & Xu Yu, 2014) : (i) la mesure de similarité entre deux points de données et (ii) la sélection de la valeur k. La valeur k représente le nombre de point similaire à sélectionner autour afin de faire la prédiction. De nombreuses méthodes ont été proposées pour résoudre le premier problème, telles que la distance euclidienne, la distance de Manhattan et la distance de Minkowski et leurs variantes. Selon plusieurs recherches, la conclusion commune du premier problème est que différentes

applications nécessitent différentes mesures de distance ((Luo & Xu Yu, 2014), (Zhang & Li, 2017)). Le deuxième problème, à savoir la sélection de la valeur k en utilisant simplement la distance euclidienne pour calculer la similarité (ou la distance) entre deux points de données. En résumé, l'algorithme des k-plus proches voisins vise à trouver les voisins les plus proches d'un point donné pour lui prédire sa valeur.

Pour déterminer quels points de données sont les plus proches d'un point de prédiction donné, il est nécessaire de calculer la distance entre le point de prédiction et les autres points de données. Ces mesures de distance sont utilisées pour définir des frontières de décision qui divisent les points de prédiction en différentes régions (Zhang & Li, 2017).

Distance euclidienne (p=2) : C'est la mesure de distance la plus couramment utilisée, et elle est limitée aux vecteurs à valeurs réelles. En utilisant la formule ci-dessous, elle mesure une ligne droite entre le point de requête et l'autre point à mesurer (Meraghni & Abdallah, 2019).

Distance Manhattan (p=1) : Il s'agit également d'une autre métrique de distance populaire, qui mesure la valeur absolue entre deux points. Elle est également appelée distance de Manhattan ou distance de bloc de ville car elle est couramment visualisée avec une grille, illustrant comment on pourrait naviguer d'une adresse à une autre via les rues de la ville (Meraghni & Abdallah, 2019).

Distance Minkowski : Cette mesure de distance est la forme généralisée des métriques de distance euclidienne et de Manhattan. Le paramètre, p, permet la création d'autres métriques de distance. La distance euclidienne est représentée par cette formule lorsque p est égal à deux, et la distance de Manhattan est désignée avec p égal à un (Meraghni & Abdallah, 2019).

Dans l'algorithme KNN, la valeur de k détermine le nombre de voisins utilisés pour classer un point donné. Par exemple, avec k=1, le point est classé selon la classe de son voisin le plus proche. Choisir k est un équilibre car des valeurs différentes peuvent conduire à un sur-ajustement ou un sous-ajustement. Des valeurs plus faibles de k peuvent donner une variance élevée mais un faible biais, tandis que des valeurs plus élevées peuvent entraîner un biais élevé mais une variance plus faible. Le choix de k dépendra des données, avec plus de valeurs aberrantes ou de bruit favorisant des valeurs de k plus élevées. Il est recommandé d'utiliser un nombre impair pour k pour éviter les égalités de classement, et la validation croisée peut aider à choisir la meilleure valeur de k pour les données du travail (Luo & Xu Yu, 2014).

Un désavantage de l'algorithme KNN est sa sensibilité au choix de la valeur de k (Zhang & Li, 2017). Malgré des recherches sur ce sujet, la sélection de la valeur k dans l'algorithme KNN reste très complexe et sujette à débat.

3.5.5 Forêts aléatoires (random forest regressor)

Les forêts aléatoires créent automatiquement des arbres de décision non corrélés. Pour ce faire, chaque arbre de décision est construit en utilisant un ensemble aléatoire de variables (bagging/boostrap aggregation) (Machová, Barčák, & Bednár , 2006). Cela en fait un excellent modèle pour travailler avec des données très variée.

Le Bagging, ou Bootstrap Aggregating, est une technique d'apprentissage automatique utilisée pour rendre les modèles plus fiables et précis. Elle consiste à créer plusieurs sous-ensembles de données d'entraînement en échantillonnant de manière aléatoire avec remplacement (Machová, Barčák, & Bednár , 2006).

Lors de la prédiction, les résultats de ces modèles de base sont combinés, souvent en faisant la moyenne pour les tâches de régression, afin d'obtenir la prédiction finale. Le bagging réduit le surapprentissage en ajoutant de la diversité parmi les modèles de base et améliore les performances globales en diminuant la variance et en augmentant la robustesse.

L'algorithme de forêt aléatoire ne nécessite pas de mise à l'échelle des données. Même après avoir fourni des données sans mise à l'échelle, il conserve une bonne précision. De plus, les valeurs aberrantes n'ont pas d'impact significatif sur les forêts aléatoires. Les variables sont regroupées pour y parvenir (George & Sumathi, 2020).

Néanmoins le modèle de Random Forest reste très sensible par rapport aux paramètres du modèle. Les paramètres présentés ci-dessous viennent directement de la documentation scikit-learn dans python (Scikit Learn). J'ai décidé d'utiliser les paramètres suivants :

'n_estimators' : Indique le nombre d'arbres dans la forêt distincts (Scikit Learn). Par conséquent, quand il y a plus d'arbres, le modèle peut mieux s'adapter à différents types de données et éviter de trop se concentrer sur les données d'entraînement. Cela veut dire qu'il sera plus fiable pour prédire avec des nouvelles données. Néanmoins il faut faire attention car s'il y a trop d'arbres le modèle peut faire sur sur-apprentissage (Ali, Khan, & Ahmad, 2012).

'max_depth' : Indique la profondeur maximale de l'arbre (Scikit Learn). Si aucun, les nœuds sont développés jusqu'à ce que tous les derniers nœuds soient pures ou jusqu'à ce que toutes les feuilles contiennent moins d'échantillons min_samples_split. Si la profondeur est trop faible, le modèle risque de ne pas saisir toutes les subtilités des données, ce qu'on appelle le sous-ajustement (Ali, Khan, & Ahmad, 2012). En revanche, si la profondeur est trop élevée, le modèle peut s'accrocher à des détails insignifiants dans les données d'entraînement, ce qu'on appelle le surajustement. Pour bien généraliser, il faut trouver un juste équilibre entre ces deux extrêmes.

'min_samples_split' : Indique le nombre minimal d'échantillons requis pour diviser un nœud (Scikit Learn). Un min_samples_split trop faible peut entraîner un surajustement en permettant à l'arbre de diviser les données de manière excessive, capturant ainsi le bruit présent dans les données d'entraînement. En revanche, un min_samples_split trop élevé peut conduire à un sous-ajustement en limitant les divisions de l'arbre, ce qui peut ne pas permettre de capturer efficacement la structure des données. C'est pourquoi en général un min_samples_split plus élevé rend le modèle plus robuste aux valeurs aberrantes et au bruit dans les données, car il limite la complexité de l'arbre, réduisant ainsi le risque de surajustement aux points de données isolés (Ali, Khan, & Ahmad, 2012).

'min_samples_leaf' : Indique le nombre minimum d'échantillons requis pour se trouver sur un nœud final (scikit-learn). Un min_samples_leaf trop faible peut entraîner un surajustement en permettant à l'arbre de se diviser trop finement, ce qui capture le bruit dans les données. À l'inverse, un min_samples_leaf trop élevé peut conduire à un sous-ajustement en limitant la croissance de l'arbre, compromettant sa capacité à capturer la structure des données (Ali, Khan, & Ahmad, 2012). Un min_samples_leaf plus élevé rend le modèle plus robuste aux valeurs aberrantes en limitant la probabilité que chaque feuille contienne des valeurs aberrantes. Cela réduit la sensibilité de l'arbre aux points de données isolés, améliorant ainsi sa capacité à généraliser sur de nouvelles données.

Pour améliorer les performances de la régression Random Forest, on utilise l'optimisation des hyperparamètres qui consiste à tester différentes combinaisons de réglages. Évaluer une

régression uniquement avec les données d'entraînement peut causer un problème de surajustement. Pour éviter cela, on utilise la validation croisée avec la méthode de recherche de grille pour trouver les meilleurs réglages.

GridSearchCV teste toutes les combinaisons de valeurs passées dans les hyperparamètres et évalue le modèle pour chaque combinaison en utilisant la méthode de validation croisée (George & Sumathi, 2020). Ainsi, après avoir utilisé cette fonction, nous obtenons l'exactitude/la perte pour chaque combinaison d'hyperparamètres et nous pouvons choisir celle qui offre les meilleures performances.

Chapitre 2 : Déploiement des modèles

1. Régression de Ridge

Dans un premier temps, j'ai standardisé mes données. Les techniques de mise à l'échelle, telles que MinMaxScaler, RobustScaler et StandardScaler, sont principalement utilisées pour normaliser les données afin d'améliorer la convergence des algorithmes et de garantir que toutes les variables sont sur la même échelle. Elles ne modifient pas fondamentalement la structure des données ou les relations entre les variables, mais elles sont essentielles pour assurer la stabilité et la fiabilité des résultats du modèle de régression de Ridge (Peshawa Jamal & Faraj, 2014).

La standardisation des données, effectuée par StandardScaler, facilite également l'interprétation des coefficients dans le modèle de régression Ridge. Puisque la régularisation est appliquée de manière égale à toutes les variables, les coefficients standardisés peuvent être comparés pour évaluer l'importance relative des variables dans le modèle (Melkumova & Shatskikh, 2017).

Au cours de ce travail 3 techniques de standardisations ont été testée pour la Ridge régression. Selon les recherches il est compliqué de prédire quelle méthode de standardisation sera la plus efficace pour notre base de données ((Singh, 2022), (Melkumova & Shatskikh, 2017)).

Ensuite, dans l'étude du modèle de régression de Ridge pour prédire les performances générales des joueurs de football, une anomalie a été observé : les mesures de performance du modèle ne varient que très peu avec différents paramètres *lambda*, malgré la régularisation introduite par cette méthode[9]. Cette stabilité peut être interprétée par plusieurs hypothèses.

Une hypothèse est que les données sont relativement stables et bien structurées. La régularisation introduite par Ridge pourrait donc ne pas avoir un impact significatif sur la variance des prédictions. Cette stabilité des données s'explique principalement par une distribution équilibrée des données et une faible présence de valeurs aberrantes. De plus lors de la section du prétraitement des données toutes les variables indépendantes corrélées entre elles et non-corrélées avec la variable dépendante ont été supprimées. Ainsi, même avec des valeurs de *lambda* différentes, les prédictions du modèle restent cohérentes et peu influencées par la régularisation.

Par ailleurs, bien que le modèle conserve 54 variables après la suppression des variables corrélées, il reste relativement complexe. En effet, la régularisation de Ridge conserve toutes les variables, mais atténue l'importance de certaines en réduisant leurs coefficients à des valeurs proches de zéro. Cependant, il est probable que ces variables contribuent de manière équilibrée à la prédiction de la variable cible, rendant le modèle moins sensible aux variations de lambda (Chiang, Liu, & Zhang, 2018). Cela peut indiquer que le modèle est capable de capturer efficacement les nuances des relations entre les variables, indépendamment du niveau de régularisation introduit par Ridge. Cette hypothèse peut s'observer au niveau des valeurs des coefficients des variables. En effet, les différents coefficients sont relativement proches sans présence de valeur excessive. Cela suggère que la régularisation introduite par Ridge maintient les coefficients dans une plage raisonnable, ce qui indique une distribution équilibrée de l'importance des différentes variables dans la prédiction de la variable cible. Ainsi,

[9] Des détails sont disponibles en annexe

même avec des valeurs de *lambda* différentes, les coefficients des variables restent relativement stables, ce qui contribue à la stabilité globale des performances du modèle.

Une autre hypothèse pourrait concerner l'interaction des variables. Si les variables interagissent de manière complexe dans le modèle, la régularisation introduite par Ridge pourrait ne pas avoir un impact significatif sur les performances (Chiang, Liu, & Zhang, 2018). Si les relations entre les variables sont subtiles et non linéaires, Ridge peut ne pas être en mesure de les capturer efficacement. Cela pourrait expliquer pourquoi les performances du modèle restent stables malgré les variations de *lambda*. En effet on remarque que très peu de variables dont les *betas* sont proches de zéro en faisant varier le *lambda* de régularisation. Le modèle prend toujours en compte les mêmes variables et en leur attribuant quasi la même importance.

1.1 Résultats

Après avoir testé plusieurs combinaisons et vérifié les résultats à l'aide de la validation croisée, j'ai obtenu les meilleures performances en utilisant la standardisation avec StandardScaler().

- λ cross validé : 0.5
- MAE: 2.174
- MSE: 7.657
- RMSE: 2.767
- R^2 score: 0.892
- R^2 ajusté : 0.88
- Akaike Information Criterion (AIC): 1918.493
- Bayesian Information Criterion (BIC): 2077.33

Vu que la régularisation de Ridge garde toutes ses variables et abaisse certaines *betas* proches de zéro il n'est pas étonnant de voir que les mesures de régression soient plutôt bonnes. Ceci explique aussi pourquoi le R^2 et le R^2 ajusté ont des valeurs très proches.

1.2 Sélection de variables

Dans le tableau ci-dessous, je présente le classement des variables qui contribuent le plus à la prédiction de la variable 'overall'. On remarque que les deux variables ayant le plus d'importance sont le 'physique' et 'l'attacking_heading_accuracy'. Ensuite, on observe que les variables suivantes ont des coefficients qui décroissent graduellement mais qui restent assez proches les uns des autres.

Overall =	
physic	3.335
attacking_heading_accuracy	2.334
skill_long_passing	1.476
mentality_penalties	1.279
skill_moves	1.060
value_eur	1.016
power_strength	-0.950
mentality_aggression	-0.895
skill_curve	0.722
movement_acceleration	0.716
movement_agility	0.713
league_level	-0.707
age	0.625

wage_eur	0.539
power_jumping	-0.523
attacking_crossing	0.433
mentality_interceptions	-0.422
nationality_id	-0.405
league_id	-0.383
defending_marking_awareness	0.365
club_team_id	-0.301
skill_fk_accuracy	0.229
power_stamina	-0.200
movement_sprint_speed	0.183
movement_balance	-0.158
weak_foot	0.131
height_cm	-0.105
Defensive_work_rate	-0.104
weight_kg	0.053
international_reputation	-0.052

Sur base des coefficients, certaines tendances sont mises en évidence :

On remarque que les variables liées aux qualités techniques des attaquants sont souvent bien notées et donc importantes pour la prédiction de leurs performances. Parmi celles-ci, on note particulièrement 'attacking_head_accuracy', 'skill_long_passing', 'skill_moves', 'skill_curve' et 'mentality_penalties'.

Augmenter la précision des passes longues permet de transformer efficacement la défense en attaque. La qualité des frappes avec effet enroulé et l'effet sur la balle sont également cruciaux pour la performance offensive. De plus, une bonne note en tirs de pénalty augmente significativement les chances de marquer.

Selon la régularisation de Ridge, on peut remarquer que pour les joueurs offensifs, ce ne sont pas les aspects d'agressivité dans le jeu qui sont importants. En effet, ces joueurs sont plutôt caractérisés par leurs qualités techniques. Cette constatation est appuyée par les *betas* négatifs des variables telles que 'power_strength' et 'mentality_aggression'.

La force, bien qu'elle puisse augmenter les chances de remporter des défis physiques avec des adversaires, et l'agressivité, qui peut aider à l'emporter dans un défi 50/50, ne sont pas des facteurs déterminants pour les performances des joueurs offensifs selon le modèle Ridge. Les qualités techniques restent les caractéristiques prédominantes pour ces joueurs.

Par ailleurs, il apparaît qu'il y a un lien assez fort entre la valeur marchande du joueur et sa note générale. En effet, ce lien est assez logique, car la valeur marchande d'un sera d'autant plus grande que le joueur est performant.

Un raisonnement similaire peut être appliqué au salaire du joueur, car logiquement au plus un joueur est performant, au plus il sera payé afin de le garder dans le club. Néanmoins on constate que l'impact est plus petit.

Ce qui peut paraitre surprenant c'est que des variables comme 'nationality_id', 'league_id', 'club_team_id' et 'international_reputation' aient des *betas* négatives. Ces observations sont contre-intuitives car spécifiquement pour les variables 'league_id' et 'club_team_id', où l'on

pourrait penser qu'un joueur évoluant dans un championnat ou une équipe de haut niveau aurait une influence positive sur sa performance. Appartenir à un certain club ou ligue peut influencer négativement ou positivement la performance, en raison des niveaux de compétition.

2. Régression de Lasso

Puisque la pénalisation L1 ajoute $\sum_{j=1}^{m}|\beta_j|$ à la fonction RSS, chaque coefficient est soumis à une contrainte de régularisation proportionnelle à sa valeur. Cela veut dire que les coefficients plus grands seront régularisés de manière plus significative que les coefficients plus petits (Kumar D. , 2024). L'algorithme Lasso réduira donc plus les coefficients des variables à plus petite échelle, même si elles sont importantes pour la prédiction, car cela réduira la pénalité globale de régularisation appliquée à ces coefficients. Par conséquent, cela peut conduire à une sous-représentation ou à une suppression de certaines variables importantes simplement parce qu'elles ont des échelles différentes par rapport aux autres variables (Kumar D. , 2024).

Au cours de ce travail les 3 techniques de standardisations ont été testée pour la régression Lasso:

- StandardScaler et RobustScaler : Dans les 2 cas lorsque le modèle est lancé les valeurs des R^2 sont respectivement -60.08 et -9.85. Un score R^2 négatif signifie que les prédictions sont moins bonnes que si l'on prédisait systématiquement la valeur moyenne. De plus pour certains lambdas dans la pénalisation de Lasso, notre modèle ne converge plus. Ces 2 techniques ne sont par principe pas les plus adaptées à la régression Lasso.
- MinMaxScaler : Dans le cas de la régression Lasso, c'est la seule méthode de standardisation qui donne des valeurs cohérentes en termes de performances du modèle. En fonction des *lambdas* de pénalité choisi entre 0.5 et 0.9 le R^2 est entre 0,67 et 0,77. Ceci est logique car au moins la pénalité est élevée au plus de variables seront présentent avec des *betas* significatives ce qui augmente potentiellement l'explication de la variable 'overall'.

2.1 Résultats

Après avoir ajusté le paramètre de régularisation pour différents lambdas et confirmé les résultats par validation croisée, les performances de ce modèle sont les suivantes :

- λ cross validé : 0.4
- MAE: 3.413
- MSE: 19.026
- RMSE: 4.361
- R^2 score: 0.733
- R^2 ajusté : 0.729
- Akaike Information Criterion (AIC): 2712.7194
- Bayesian Information Criterion (BIC): 2789.734

Le modèle présente des résultats moins bon en comparaison avec les résultats de la Ridge régression. Néanmoins ceci n'est pas étonnant car contrairement à la Ridge régression, la Lasso régression mets à zéro certaines betas. Ainsi le nombre de variables indépendantes qui expliquent le modèle est plus petit. Cela conduire à un modèle plus simple et plus interprétable,

mais aussi potentiellement moins précis, surtout si certaines des variables éliminées sont en réalité importantes pour la prédiction de la variable cible.

2.2 Variables sélectionnées

Analysons les 18 variables sélectionnées par la régression de Lasso. L'interprétation des coefficients dans une Lasso régression restent similaires à ceux d'une régression linéaire. Plus le coefficient est élevé en valeur absolue, plus la variable correspondante a un impact sur la prédiction de la variable cible. On peut voir que les 7 premières variables sont les plus significatives. Après celles-ci l'ampleur des betas est plus petite.

OVERALL =	
'height_cm'	15.718
'mentality_penalties'	13.575
'movement_balance'	12.880
'skill_long_passing'	12.641
'attacking_heading_accuracy'	9.192
'power_strength'	9.127
'movement_agility'	8.946
'movement_acceleration'	5.058
'age'	3.559
'weak_foot'	3.429
'preferred_foot'	2.946
'Attacking_work_rate'	2.565
'movement_sprint_speed'	1.579
'power_stamina':	1.053
'skill_curve'	0.525
'club_team_id'	0.464
'mentality_aggression'	0.178

Les variables sélectionnées par la régularisation de Lasso restent assez similaires à celles de la régularisation de Ridge. Parmi elles, on trouve des variables telles que 'mentality_penalties', 'skill_long_passing', 'attacking_heading_accuracy', 'movement_agility' et 'age'.

Contrairement à la régularisation de Ridge, la régularisation de Lasso met l'accent sur la puissance et l'agressivité de l'attaquant. La taille ainsi que 'power_strength' et 'movement_balance' sont des variables significatives dans ce modèle.

Par ailleurs, la variable 'club_team_id' a un impact positif ce qui veut dire qu'un attaquant qui joue dans un meilleur club, aura de meilleure possibilité de développement.

Ce modèle prend en compte de nouvelles variables, notamment 'weak_foot' et 'preferred_foot'. Ces deux variables mesurent la capacité des joueurs à utiliser leurs deux pieds. Une meilleure maîtrise du pied le plus faible permet d'améliorer les performances du joueur.

3. ElasticNet

L'Elastic Net intègre les avantages des régularisations Ridge et Lasso tout en atténuant leurs inconvénients grâce à une combinaison des pénalités $L1$ (Lasso) et $L2$ (Ridge). Cela se traduit par une fonction de pénalisation mixte où la pénalité totale est une combinaison linéaire des deux pénalités : la contrainte $L1$ contribue à la sélection des variables, tandis que la contrainte $L2$ stabilise les coefficients. Le paramètre *alpha* contrôle la proportion de chacune des pénalités dans le modèle (Zou & Hastie, 2005). Par exemple, un *alpha* de 0.5 signifie que les pénalités $L1$ et $L2$ contribuent de manière égale (scikit-learn).

En combinant ces deux formes de régularisation, l'Elastic Net offre plusieurs avantages. Comme la régression Ridge, l'Elastic Net réduit la variance des estimations des coefficients, ce qui aide à stabiliser les prédictions, en particulier lorsque les variables explicatives sont fortement corrélées. Cette approche évite les problèmes de multicolinéarité en répartissant les poids de manière plus équilibrée entre les variables corrélées.

Par ailleurs, comme la régression Lasso, l'Elastic Net peut réduire certains coefficients à zéro, permettant ainsi une sélection automatique des variables. Cela simplifie le modèle et améliore son interprétabilité en éliminant les variables non informatives. L'Elastic Net parvient donc à combiner la réduction de variance de Ridge et la capacité de sélection de variables de Lasso, ce qui est particulièrement utile dans les contextes où les prédicteurs sont nombreux et potentiellement colinéaires.'

Par ailleurs, l'Elastic Net atténue les principaux inconvénients de Ridge et Lasso. Contrairement à Lasso, qui peut être instable et choisir une seule variable parmi un groupe de variables corrélées, l'Elastic Net stabilise les coefficients grâce à la composante $L2$ (Zou & Hastie, 2005). Cela permet de conserver plus d'information utile et d'éviter une exclusion excessive de variables corrélées. Simultanément, l'Elastic Net évite de conserver toutes les variables comme Ridge, grâce à sa capacité de sélection de variables via la composante $L1$.

3.1 Résultats

Dans le cadre de ce travail, après la standardisation avec StandardScaler, le meilleur modèle a été obtenu avec un *lambda*, contrôlant le degré de pénalisation, fixé à 0.2 et un *alpha* de 0.75. Les performances du modèle sont les suivantes :

- MAE: 2.245
- MSE: 8.053
- RMSE: 2.837
- R^2: 0.887
- R^2 ajusté : 0.884
- Akaike Information Criterion (AIC): 1942.340
- Bayesian Information Criterion (BIC): 2048.236

La régularisation $L1$, qui constitue 75 % de l'impact de la régularisation du modèle, permet à Lasso de jouer un rôle majeur. Cette régularisation $L1$ aide à éliminer certaines variables non informatives, simplifiant ainsi le modèle en ne gardant que les variables les plus pertinentes. Cela signifie que le modèle profite de la capacité de Lasso à effectuer une sélection de variables, ce qui est particulièrement utile lorsque certaines des variables explicatives n'apportent pas de valeur prédictive significative.

Ensuite, la régularisation L2, qui représente les 25 % restants, joue également un rôle crucial. Cette régularisation L2 aide à gérer la multicolinéarité entre les variables explicatives. En réduisant légèrement la magnitude de tous les coefficients sans en mettre aucun à zéro, elle stabilise les coefficients, évitant qu'ils ne deviennent trop instables en présence de variables corrélées. Cela contribue à une meilleure interprétabilité et à une stabilité accrue du modèle.

Enfin, la valeur de *lambda* de 0.2 indique que le modèle applique une régularisation modérée. Cette régularisation modérée est nécessaire pour éviter le surajustement, tout en ne pénalisant pas trop les coefficients, ce qui pourrait conduire à un sous-ajustement. En d'autres termes, les données nécessitent une certaine régularisation pour améliorer la généralisation, mais une régularisation trop forte n'est pas nécessaire.

Comparé à la régression de Ridge, l'Elastic Net performe mieux en termes de BIC (2077 contre 2048), ce qui suggère une meilleure gestion de la complexité en rapport avec le nombre de paramètres qu'il utilise, malgré une pénalisation plus forte. Ainsi, l'Elastic Net favorise la simplicité et la robustesse, surtout face à la multicolinéarité.

3.2 Variables sélectionnées

Étant donné que l'objectif est de généraliser au mieux sur de nouvelles données, le BIC pourrait être un critère plus approprié dans ce contexte, car il pénalise plus fortement la complexité du modèle. Par conséquent, le modèle Elastic Net, avec son BIC plus bas, pourrait être privilégié.

Overall =	
attacking_heading_accuracy	1.878
mentality_penalties	1.343
skill_long_passing	1.104
skill_moves	1.067
value_eur	0.894
skill_curve	0.873
physic	0.671
power_strength	0.631
wage_eur	0.609
league_level	-0.587
attacking_crossing	0.532
power_stamina	0.522
movement_acceleration	0.462
age	0.437
movement_agility	0.409
league_id	-0.341
skill_fk_accuracy	0.292
nationality_id	-0.262
movement_sprint_speed	0.243
club_team_id	-0.219
club_jersey_number	-0.091
weak_foot	0.089

En comparant les variables sélectionnées par l'Elastic Net avec celles de Ridge et Lasso, on constate clairement l'effet combiné des deux régularisations sur le choix des variables. La

similitude des variables sélectionnées est confirmée par l'Elastic Net, où les variables communes aux deux régularisations sont présentes.

Comme avec la régularisation de Ridge, les coefficients des variables dans le modèle Elastic Net sont répartis de manière assez uniforme, diminuant progressivement. Les variables les plus significatives sélectionnées par l'Elastic Net correspondent principalement à celles identifiées par la régularisation de Lasso, ce qui est attendu étant donné une pénalité alpha de 0.75. Les variables importantes dans la régularisation de Ridge, telles que 'physic', 'skill_moves', 'value_eur' et 'wage_eur', sont également présentes dans le modèle Elastic Net.

4. KNeighborsRegressor (KNN)

Une caractéristique importante de KNN (K-Nearest Neighbors) est sa sensibilité à la dimensionnalité, c'est-à-dire au nombre de caractéristiques utilisées pour la prédiction. Dans la régression avec KNN, la prédiction est effectuée en prenant la moyenne ou la médiane des variables des plus proches observations. Cette méthode peut être fortement influencée par la dimensionnalité des données (Kouiroukidis & Evangelidis, 2011).

La principale raison de la nécessité de réduire la dimensionnalité réside dans ce qu'on appelle le "fléau de la dimension". À mesure que le nombre de dimensions augmente, le volume de l'espace augmente de manière exponentielle. Cela entraîne une dilution des données, ce qui signifie que les points de données deviennent de plus en plus éloignés les uns des autres dans un espace de haute dimension. La notion de proximité perd alors son sens, rendant difficile la recherche de voisins proches pertinents pour le modèle KNN (Kouiroukidis & Evangelidis, 2011). En conséquence, l'efficacité et la précision du modèle peuvent diminuer car il devient plus complexe de trouver des voisins fiables et pertinents pour faire des prédictions.

À mesure que le nombre de dimensions augmente, la quantité de données nécessaire pour représenter efficacement l'espace augmente également. Cela signifie que plus le nombre de dimensions croît, plus la quantité de données requise pour produire des résultats fiables peut devenir extrêmement grande (Kouiroukidis & Evangelidis, 2011). Cette augmentation des exigences en données peut rendre l'entraînement et l'utilisation du modèle KNN peu pratiques, en particulier avec des ensembles de données limités.

Avec un nombre réduit de dimensions, le modèle KNN devient plus stable et moins sensible aux variations dans les données. La réduction de la dimensionnalité aide à minimiser l'impact des caractéristiques inutiles, contribuant ainsi à un modèle plus robuste et fiable. Moins de dimensions signifient également que les caractéristiques restantes sont probablement plus pertinentes pour la prédiction, ce qui améliore la performance globale du modèle.

C'est pourquoi, dans ce travail, j'ai choisi d'utiliser deux techniques de sélection de variables pour optimiser mon modèle KNN.

4.1 KNeighborsRegressor – sélection de variables avec Lasso

J'ai tout d'abord repris la régression de Lasso. La régression de Lasso me permet de sélectionner un ensemble de variables ce qui réduit la dimensionalité. Ensuite, ces variables sont introduites dans le modèle KNeighborsRegressor[10].

[10] La fonction Python pour utiliser la régression du modèle KNN.

Le modèle KNeighborsRegressor nécessite aussi un ‘parameter tunning’ en utilisant la méthode GridSearchCV. Ainsi parmi tous les paramètres qui caractérisent le KNeighborsRegressor tels que le nombre de voisins, la mesures de distances seront testé afin de choisir la meilleure combinaison. Il en ressort que les meilleurs paramètres sont les suivants (scikit-learn, KNeighborsRegressor) :

- 'n_neighbors': 8, nombre de voisins sélectionnés
- ‘wheights’: ‘uniform’, poids de de chaque voisin
- 'algorithm': 'ball tree', algorithme utilisé pour rechercher les voisins
- 'leaf_size': 10, paramètre qui contrôle la taille des feuilles pour les algorithmes 'ball_tree'

4.1.2 Résultats

Ainsi ce modèle avec les meilleurs paramètres mène aux résultats suivants :

- MAE: 1.824
- MSE: 5.737
- RMSE: 2.395
- R^2 Score: 0.919
- R^2 ajusté : 0.917
- Akaike Information Criterion (AIC): 1629.766
- Bayesian Information Criterion (BIC): 1726.035

Ce modèle à présent est celui qui présente les meilleurs résultats. Les valeurs de MAE, MSE et RMSE montrent que le modèle a une erreur relativement faible, ce qui indique une bonne précision des prédictions. Les scores R^2 et R^2 ajusté, étant à 0.92, montrent que le modèle explique bien la variance des données, ce qui signifie qu'il est efficace. Les valeurs de l'AIC et du BIC indiquent que le modèle est bien équilibré entre complexité et qualité des prédictions. Puisque la régression Lasso a été utilisée pour la sélection des variables, il est probable que seules les variables les plus pertinents ont été retenus, évitant ainsi le surajustement.

4.2 KNeighborsRegressor – recursive feature selection

Dans cette étude, une méthode appelée "recursive feature selection» a été employée pour sélectionner les variables les plus pertinentes en évaluant un modèle. Ensuite, cet ensemble de variables a été introduit dans un modèle de KNeighborsRegressor.

Comme pour tous les modèles, le Random Forest regressor a aussi été ajusté avec les meilleurs hyperparamètres grâce à la fonction GridSearchCV.

Initialement, j’ai remarqué que les variables "value_eur" et "wage_eur" dominaient nettement par rapport aux autres variables. Après analyse, je me suis rendu compte que ce phénomène était un biais de représentativité[11]. Premièrement, il est évident de constater une disproportion dans les valeurs de ces deux variables en fonction du score "overall". En effet, l'évolution de ces deux variables suit une distribution exponentielle. Ceci ne surprend pas car, en général, les clubs sont prêts à payer des sommes astronomiques pour les joueurs les mieux notés.

Deuxièmement, les variables "value_eur" et "wage_eur" sont à une échelle bien plus grande que les autres variables. Cela induit une importance apparente plus élevée dans le modèle, car les arbres de décision ont tendance à effectuer des divisions en fonction de la variance expliquée par chaque caractéristique.

[11] Une illustration se trouve en annexe

4.2.1 Résultats

Le modèle présente les résultats suivants :

- MAE: 2.360
- MSE: 9.041
- RMSE: 3.006
- R^2 Score: 0.873
- R^2 ajusté : 0.870
- Akaike Information Criterion (AIC): 2003.646
- Bayesian Information Criterion (BIC): 2139.915

On peut constater que dans l'ensemble, le modèle s'ajuste bien aux données avec un R^2 ajusté de 0.870. De plus les erreurs des données restent faibles ce qui confirme le bon ajustement.

Néanmoins, si l'on compare les critères AIC et BIC ceux-ci sont nettement supérieur à ceux du modèle de sélection de variables avec Lasso. Une raison qui pourrait expliquer cette différence se trouve dans les variables sélectionnées. Les variables choisies par les 2 modèles diffèrent légèrement. Ainsi d'après les résultats BIC, le modèle RFE a sélectionné des variables qui ont une relation non linéaire plus complexe que la sélection de variables Lasso.

4.2.2 Variables sélectionnées

Suite à la fonction 'recursive features elimination' dans python, il en ressort que ces variables sont les plus significatives :

- age
- mentality_interceptions
- power_strength
- power_stamina
- movement_agility
- movement_sprint_speed
- movement_acceleration
- skill_long_passing
- skill_fk_accuracy
- skill_curve
- attacking_heading_accuracy
- attacking_crossing
- physic
- international_reputation
- skill_moves
- nationality_id
- club_team_id
- league_id
- mentality_penalties
- defending_standing_tackle

Les variables choisies par la méthode de l'élimination récursive des caractéristiques ont été classées avec un rang égal à un, ce qui indique leur importance prioritaire. Ces variables sont très semblables à celles sélectionnées par les modèles précédents. Cependant, pour la première fois, la variable 'international_reputation' est incluse. Cette sélection n'est pas surprenante, car les joueurs les plus réputés sont généralement ceux qui performent le mieux.

5. Random Forest regressor

Petit rappel pourquoi est-ce que le random forest regressor est utile dans le cadre de ce travail ?

Le Random Forest Regressor est une méthode robuste pour modéliser des données réelles grâce à son utilisation de multiples arbres de décision. Ces arbres sont formés sur des sous-ensembles aléatoires des données, ce qui atténue l'impact des valeurs aberrantes et permet d'ignorer les données manquantes sans nécessiter de prétraitement supplémentaire (Kumar D. , 2024). De plus, le modèle filtre le bruit en sélectionnant aléatoirement un sous-ensemble de caractéristiques à chaque division d'arbre, ce qui permet de se concentrer sur les tendances générales des données plutôt que sur les fluctuations aléatoires. En conclusion, le Random Forest Regressor est efficace pour gérer les problèmes de qualité et de propreté des données, offrant ainsi une méthode fiable pour la modélisation dans divers domaines d'application. Cette mesure est ensuite normalisée pour que la somme des scores d'importance dans la forêt aléatoire soit égale à 1.

Comme mentionné dans la partie sur Random Forest le modèle commence avec un ‘paramètre tunning’. Ainsi comme pour les autres modèles la fonction GridSearchCV a été utilisée. Les meilleurs hyperparamètres sont les suivants :

- 'max_depth': 20,
- 'min_samples_leaf': 50,
- 'min_samples_split': 60,
- 'n_estimators': 40

Ensuite comme cela a été expliqué dans la partie précédente, le modèle de Random Forest possède une méthode de sélection de variables. Dans un premier temps je m’aperçois que la variable ‘value_eur’ possède une importance écrasante par rapport aux autres variables. Après une analyse je comprends que le problème est dû à un biais de représentativité[12]. En effet, intuitivement au plus la valeur du joueur est grande au plus son overall est élevé.

5.1 Variables sélectionnées

C’est pour cela que j’essaie de relancer le modèle mais cette fois ci en supprimant la variable ‘value_eur’. Le modèle sélectionne 16 variables. La somme de toutes les importances est égale à 1. J’obtiens les importances suivantes :

Feature	Importance
skill_curve	0.364
attacking_heading_accuracy	0.214
international_reputation	0.198
physic	0.074
skill_long_passing	0.027
mentality_penalties	0.023
attacking_crossing	0.023
power_stamina	0.022
skill_moves	0.019
skill_fk_accuracy	0.007
movement_sprint_speed	0.006
club_team_id	0.006

[12] Une illustration se trouve en annexe

movement_agility	0.003
movement_acceleration	0.002
league_id	0.002

Les quatre premières variables se démarquent nettement comme les plus importantes. En comparaison avec tous les modèles précédents, il est notable que 'attacking_heading_accuracy' est une variable de grande importance, toujours présente. Cependant, dans le modèle de Random Forest, 'skill_curve' est accordée la plus grande importance, même si elle était également présente dans les modèles précédents, mais dans une moindre mesure. Enfin, une nouvelle variable qui émerge avec une importance significative est 'international_reputation'. Aucun des modèles précédents n'avait mis en évidence cette variable. Cette observation n'est pas surprenante, car en dehors des caractéristiques physiques, la réputation internationale d'un joueur est généralement liée à ses performances, ce qui peut influencer sa note globale.

5.2 Résultats

On remarque directement que les 4 premières variables ont le plus d'importance.

- MAE: 2.499
- MSE: 10.139
- RMSE: 3.184
- R^2 : 0.858
- R^2 ajusté : 0.853
- Akaike Information Criterion (AIC): 2169.976
- Bayesian Information Criterion (BIC): 2319.193

On peut constater que dans l'ensemble, le modèle s'ajuste bien aux données avec un R^2 ajusté de 0,85. De plus les erreurs des données restent faibles ce qui confirme le bon ajustement.

Néanmoins, si l'on compare les critères AIC et BIC ceux-ci sont nettement supérieur à ceux des modèles comparatifs. Une raison qui pourrait expliquer cette différence se trouve dans les variables sélectionnées. Les variables choisies par le modèle diffèrent légèrement. Ainsi d'après les résultats BIC, le modèle Random Forest a sélectionné des variables qui ont une relation non linéaire plus complexe.

En effet, Random Forest accorde beaucoup d'importances à quatre variables. Trois d'entre elles sont des variables qui reviennent dans tous les modèles : 'skill_curve', 'attacking_heading_accuracy' et 'physic'. Néanmoins une nouvelle variable avec beaucoup d'importance est mise en avant à savoir 'international_reputation'.

Cependant, les résultats montrent que la méthode de Random Forest utilisée dans ce travail n'est pas la plus efficace. Même si elle utilise une approche ensembliste basée sur des arbres de décision qui peut gérer des relations complexes entre les caractéristiques, d'autres méthodes semblent mieux convenir pour les données du travail.

6. Conclusion

6.1 Synthèse des performances des modèles

	Ridge	Lasso	Elastic Net	KNN - Lasso	KNN – RFE	Random Forest
MAE	2.174	3.413	2.245	**1.824**	2.360	2.499
R2 ajusté	0.88	0.729	0.884	**0.917**	0.870	0.853
AIC	1918.493	2712.71	1942.340	**1629.766**	2003.646	2169.976
BIC	2077.33	2789.734	2048.236	**1726.035**	2139.915	2319.193

Le modèle qui s'adapte le mieux aux données tout en gérant le mieux la complexité du modèle est le modèle KNN avec la sélection de variables de Lasso. En éliminant les variables non informatives ou redondantes, le modèle se concentre sur les caractéristiques les plus importantes, améliorant ainsi la qualité des prédictions. La réduction de la dimensionnalité permet à KNN de fonctionner plus efficacement et de se concentrer sur les caractéristiques les plus informatives, ce qui peut améliorer la précision des prédictions. De plus, le KNN a la capacité de capturer des relations non linéaires entre les caractéristiques et la variable cible. Cela lui permet de bien s'adapter aux particularités locales des données, offrant de bonnes performances lorsque les relations locales entre les observations sont fortes.

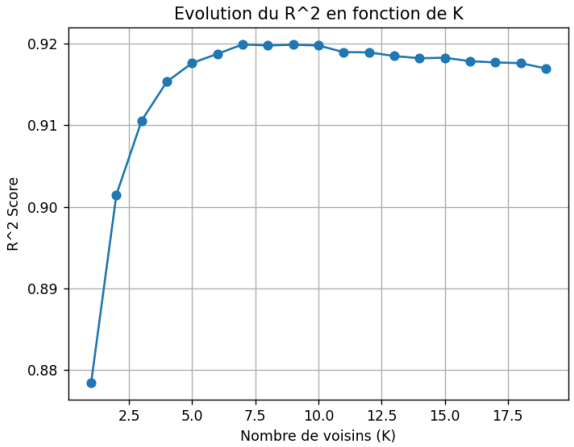


Figure 9 : Evolution du R^2 en fonction de k (via python)

Sur base des données et des variables sélectionnées, le modèle KNN, avec k optimisé à 8, prédit efficacement les performances des attaquants de football, atteignant un R^2 de 92%. Cela montre que le modèle capture bien les détails des données tout en restant fiable pour différentes situations.

Ensuite, la régression de Ridge offre une approche simple et efficace pour les relations linéaires, tandis qu'Elastic Net apporte une flexibilité supplémentaire pour gérer les caractéristiques corrélées et potentiellement non informatives. Les deux modèles montrent une bonne capacité à généraliser aux données, avec l'Elastic Net étant légèrement plus complexe mais plus robuste.

Le modèle de Random Forest regressor montre de solides performances dans l'ajustement des données, avec des métriques telles qu'un R^2 ajusté de 0,85, indiquant une bonne capacité à

expliquer la variance des données. De plus, les erreurs de prédiction, telles que la MAE, la MSE et la RMSE, sont toutes relativement faibles, ce qui confirme la précision du modèle.

Cependant, il est important de noter que les critères d'information tels que l'AIC et le BIC sont légèrement supérieurs à ceux des modèles comparatifs. Cela suggère que malgré la performance globalement solide du modèle, il est plus complexe. Ceci est dû à la sélection de variables qui diffèrent légèrement de celles des modèles comparatifs.

Pour résumé, les modèles linéaires régularisés comme Ridge et Elastic Net offrent des performances acceptables mais inférieures, suggérant une certaine linéarité, mais pas suffisante pour expliquer pleinement la variance. La performance de Random Forest indique une complexité des interactions qui nécessitent des modèles sophistiqués pour une capture précise. L'utilisation de techniques de sélection de caractéristiques comme dans KNN - Lasso est essentielle pour améliorer la performance du modèle en mettant en avant les caractéristiques les plus pertinentes.

Dans ce projet, il est important de ne pas perdre de vue l'objectif principal : prédire la note 'overall' avec la plus grande précision possible et comprendre quelles variables sont les plus importantes. En utilisant la base de données disponible, plusieurs méthodes de régression et de sélection de variables ont été évaluées. Ainsi, trois types de modèles ont été examinés : la régression linéaire et régularisée, les approches locales telles que le KNN, et les méthodes d'ensemble comme le Random Forest.

Il en ressort que la méthode locale de KNN accompagné de la régularisation de Lasso performe le mieux dans ce travail. D’une part, même si les premières variables sélectionnées par la régularisation de Lasso sont en commun à tous les modèles, le reste des variables sélectionnées sont propre à la régularisation de Lasso. Ensuite, ces variables sont introduites dans le modèle KNN qui arrive de manière efficace à prédire la variable ‘overall’. Cette combinaison permet au modèle de se concentrer sur les variables les plus importantes et de capturer les relations locales entre les variables et la performance des attaquants de football.

Les performances des modèles linéaires régularisés (Ridge et Elastic Net) suggèrent une certaine linéarité dans les données, mais pas suffisamment pour expliquer pleinement la variance observée.

La méthode d'ensemble donne également des résultats satisfaisants. En raison de la complexité du modèle, les différences de performance observées peuvent s'expliquer par le fait que la régression linéaire cherche à réduire le biais, tandis que le bagging vise à réduire la variance. Cependant, les critères d'information (AIC et BIC) sont plus élevés, ce qui suggère une complexité peut être excessive. Bien que cette méthode soit capable de capturer des interactions complexes entre les variables, elle en devient plus difficile à interpréter.

6.2 Interprétation des variables

Variables sélectionnées par la régularisation de Lasso :

OVERALL =	
'height_cm'	15.718
'mentality_penalties'	13.575
'movement_balance'	12.880
'skill_long_passing'	12.641
'attacking_heading_accuracy'	9.192
'power_strength'	9.127
'movement_agility'	8.946
'movement_acceleration'	5.058
'age'	3.559
'weak_foot'	3.429
'preferred_foot'	2.946
'Attacking_work_rate'	2.565
'movement_sprint_speed'	1.579
'power_stamina':	1.053
'skill_curve'	0.525
'club_team_id'	0.464
'mentality_aggression'	0.178

Ainsi, en analysant les coefficients des variables normalisées, on peut déterminer leur importance relative. Les variables avec des coefficients de magnitude plus élevée sont considérées comme ayant une plus grande influence sur la prédiction de la variable cible. Cela permet de classer les variables dans un ordre d'importance pour le modèle de régression linéaire.

On constate que parmi les 8 premières variables expliquant la variable 'overall', certaines concernent les aspects 'athlétiques' d'un attaquant de football. Par exemple, la taille, l'équilibre en course, la force (qui augmente les chances de gagner les duels) et l'accélération sont toutes des variables importantes lors de la sélection. D'autre part, les autres variables concernent davantage les qualités techniques des attaquants, telles que l'exécution des penalties, la qualité de passe et le jeu de tête.

Ce modèle prend également en compte de nouvelles variables, notamment 'weak_foot' et 'preferred_foot', qui mesurent la capacité des joueurs à utiliser leurs deux pieds. Une meilleure maîtrise du pied le plus faible permet d'améliorer les performances du joueur.

En outre, la variable 'club_team_id' a un impact positif, ce qui signifie qu'un attaquant jouant dans un meilleur club a de meilleures possibilités de développement. Quant aux autres variables, elles apparaissent également dans d'autres modélisations, mais avec des degrés d'importance différents.

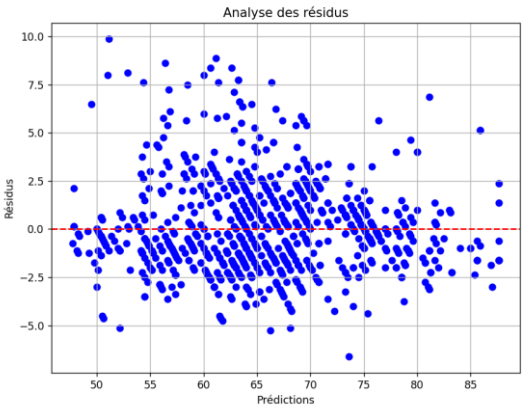


Figure 10 : Analyse des résidus (via python)

De plus, le graphique des résidus de la régression est satisfaisant, car ils semblent être aléatoirement dispersés autour de l'axe des abscisses, ne montrant pas de structure particulière.

Les variables sélectionnées par la régularisation de Ridge restent assez similaires à celles de la régularisation de Lasso. Parmi elles, on trouve des variables telles que ‘mentality_penalties’, 'skill_long_passing', 'attacking_heading_accuracy', ‘movement_agility’ et ‘age’.

Néanmoins, on remarque que les variables liées aux qualités techniques des attaquants sont souvent bien notées et donc importantes pour la prédiction de leurs performances. Parmi celles-ci, on note particulièrement ‘attacking_head_accuracy’, ‘skill_long_passing’, ‘skill_moves’, ‘skill_curve’ et ‘mentality_penalties’.

Selon la régularisation de Ridge, on peut remarquer que pour les joueurs offensifs, ce ne sont pas les aspects ‘atlétiques’ qui sont importants. En effet, ces joueurs sont plutôt caractérisés par leurs qualités techniques. Cette constatation est appuyée par les bêtas négatifs des variables telles que ‘power_strength’ et ‘mentality_aggression’.

En comparant les variables importantes pour l’Elastic Net avec celles de Ridge et Lasso, on constate clairement l'effet combiné des deux régularisations sur le choix des variables. La similitude des variables sélectionnées est confirmée par l'Elastic Net, où les variables communes aux deux régularisations sont présentes.

En ce qui concerne le modèle random forest, celui-ci accorde beaucoup d’importances à quatre variables. Trois d’entre elles sont des variables qui reviennent dans tous les modèles : ‘skill_curve’, ‘attacking_heading_accuracy’ et ‘physic’. Néanmoins une nouvelle variable avec beaucoup d’importance est mise en avant à savoir ‘international_reputation’.

On constate que les variables suivantes sont récurrentes dans tous les modèles :

- ‘attacking_heading_accuracy’
- ’mentality_penalties’
- ‘skill_long_passing’
- ’skill_moves’
- ‘skill_curve’
- ‘movement_agility’
- ‘movement_accelaration’
- ‘club_team_id’

Suite à l’analyse de tous les modèles, on constate que les variables sélectionnées sont séparées en 2 catégories distinctes. Les cinq premières variables concernent les qualités techniques des joueurs et surtout en termes de qualité de tir au but. Les variables concernées sont la précision du jeu de tête, les penalties et les effets sur le ballon lirs du tir (‘skill_curve’). De plus, la qualité de passe pour élargir le jeu ('skill_long_passing') et les déplacements de l'attaquant sont également des variables importantes.

Ensuite, les variables 'movement_agility' et 'movement_acceleration' caractérisent principalement les performances 'athlétiques' des attaquants. Les joueurs agiles peuvent pivoter plus rapidement et sont plus enclins à tenter des têtes, des volées et des coups de pied spectaculaires. Améliorer l'accélération du joueur augmentera sa vitesse et réduira le temps nécessaire pour atteindre sa vitesse de sprint maximale.

Pour finir, la variable 'club_team_id' joue également un rôle sur la note générale d'un attaquant. Il est logique que les meilleurs clubs attirent les meilleurs joueurs offensifs, et inversement. Ainsi, évoluer dans un club de premier plan offre de meilleures opportunités de développement et de visibilité pour les attaquants, ce qui se reflète dans leur note générale.

Remarques générales

La base de données utilisée dans ce travail contient des variables relativement génériques. Je pense que les data scientistes des différents clubs utilisent des données beaucoup plus précises sur les aptitudes techniques et physiques des joueurs. Néanmoins, ce travail offre déjà une première vision et une approche des variables intéressantes à prendre en compte.

Les résultats peuvent faire l'objet d'une validation sur le terrain. L'objectif initial était de recueillir l'avis de professionnels travaillant dans des clubs, par exemple. De nombreuses questions auraient pu leur être posées et l'utilisation d'une base de données réelle aurait pu produire des résultats directement applicables à une équipe.

Une demande de contact a néanmoins été envoyée à des data scientistes travaillant pour des clubs en Belgique ou en Europe. Cependant, les rares réponses reçues indiquaient soit 1) un manque de temps et des préoccupations plus importantes, soit 2) une clause de confidentialité dans leur contrat interdisant de divulguer des informations.

Codes

Les modèles que j'ai développés pour ce projet sont disponibles sur le lien suivant :

https://github.com/Matniew/Thesis

Annexe

Remarques concernant la régression de Ridge :

Scores de validation croisée moyens pour chaque valeur de lambda :

- Lambda = 0.3: 7.479707658760065
- Lambda = 0.4: 7.479634919198862
- Lambda = 0.5: 7.479631223047014
- Lambda = 0.6: 7.479671334442993
- Lambda= 0.7: 7.479738909379583

Les intervalles de confiance pour les coefficients de régression fournissent une plage dans laquelle nous nous attendons à ce que les véritables coefficients se trouvent avec un certain niveau de confiance à 95% dans ce cas. Si l'intervalle de confiance pour un coefficient n'inclut pas zéro, cela suggère que le coefficient est statistiquement significatif, c'est-à-dire qu'il a un effet non nul sur la variable cible.

Intervalles de confiance à 95% et significativité des coefficients :

- value_eur: 1.0165531498884741, True
- wage_eur: 0.5391385056404756, True
- age: 0.6254009602302393, True
- height_cm: -0.1051054985006985, False
- weight_kg: 0.053471551125058525, False
- league_id: -0.3838227846832385, True
- league_level: -0.7073393886687169, True
- club_team_id: -0.3019382703683249, True
- nationality_id: -0.4055634662945402, True
- preferred_foot: -0.040537247744220295, False
- weak_foot: 0.13122894796974016, False
- skill_moves: 1.0609618304440005, True
- international_reputation: -0.05244328799921005, False
- physic: 3.33553825346136, True
- attacking_crossing: 0.433271758251335, True
- attacking_heading_accuracy: 2.334955665080507, True
- skill_curve: 0.7227260142628493, True
- skill_fk_accuracy: 0.22911435252915924, True
- skill_long_passing: 1.4761021190649264, True
- movement_acceleration: 0.7169663083525162, True
- movement_sprint_speed: 0.1831700005176078, False
- movement_agility: 0.713295027977372, True
- movement_balance: -0.15879691151733955, False
- power_jumping: -0.5238635170826882, True
- power_stamina: -0.20027572465695848, False
- power_strength: -0.9504940907777293, False
- mentality_aggression: -0.8956774002370675, True

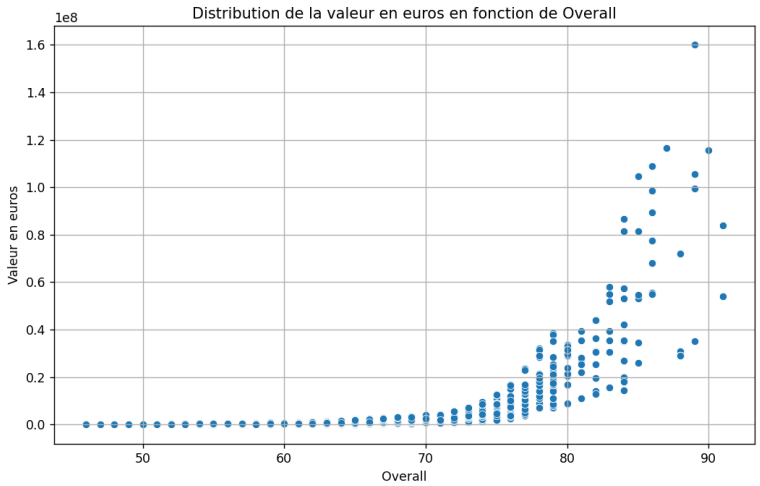
- mentality_interceptions: -0.422473806873926, True
- mentality_penalties: 1.2796927905054094, True
- defending_marking_awareness: 0.3652612235482753, True
- defending_standing_tackle: 0.03167043294381244, False
- Attacking_work_rate: -0.01657091509698816, False
- Defensive_work_rate: -0.10474089994650004, False

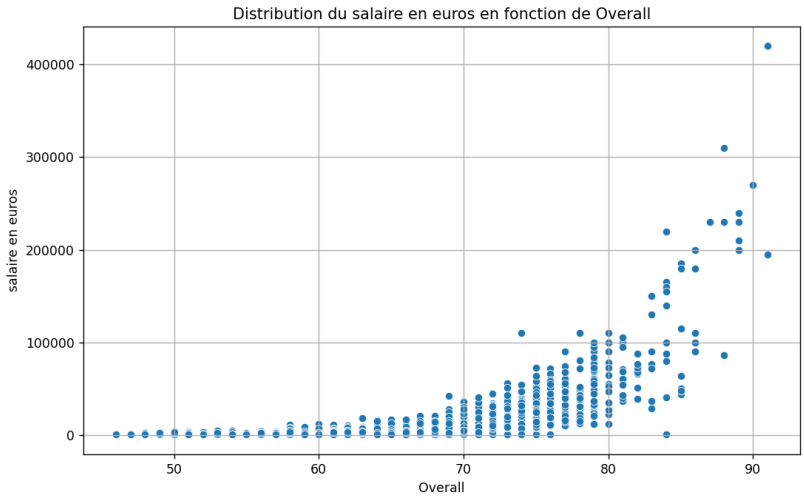
Élimination récursive de caractéristiques basée sur le Random Forest :

Un classement égal à 1 signifie que ces variables sont considérées comme les plus importantes selon l'élimination récursive de caractéristiques. Étant donné que cette méthode utilise un modèle pour évaluer l'importance des variables, nous avons également une colonne intitulée 'importance'. Cette colonne affiche l'importance attribuée aux variables par le modèle de forêt aléatoire.

Overall =	Ranking	Importance
age	1	0.006
nationality_id	1	0.009
international_reputation	1	0.165
physic	1	0.057
skill_curve	1	0.323
skill_fk_accuracy	1	0.012
skill_long_passing	1	0.044
movement_acceleration	1	0.014
movement_sprint_speed	1	0.016
movement_agility	1	0.009
power_stamina	1	0.034
power_strength	1	0.005
mentality_interceptions	1	0.006
mentality_penalties	1	0.025
defending_standing_tackle	1	0.005

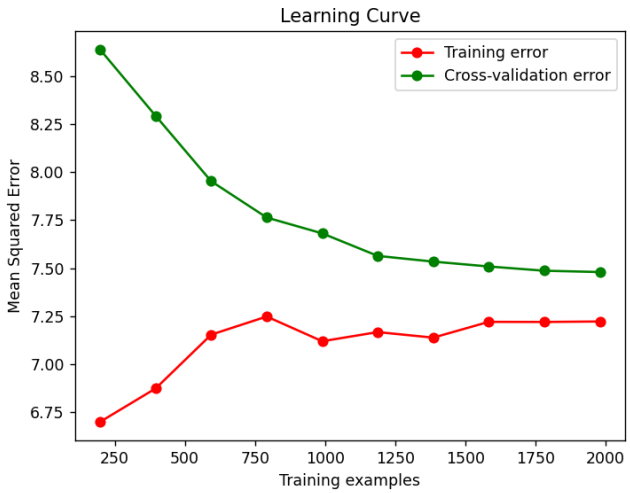
Biais de représentativité



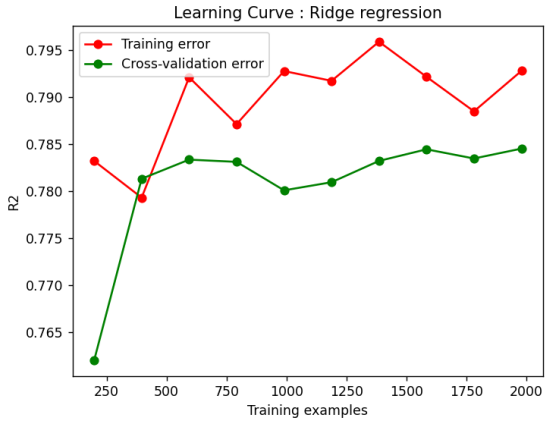


Courbes d'apprentissage :

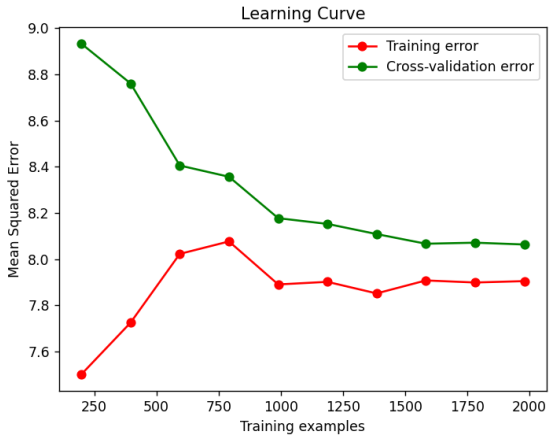
Ridge



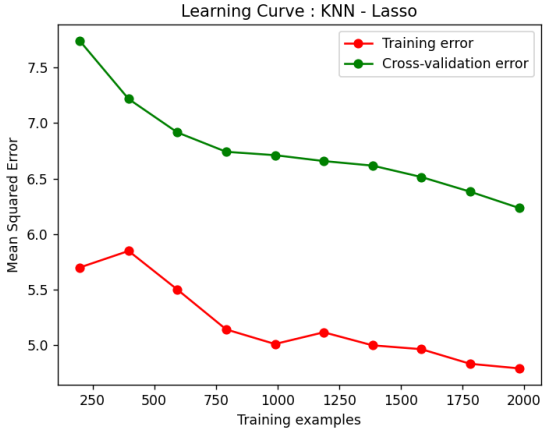
Lasso



Elastic Net

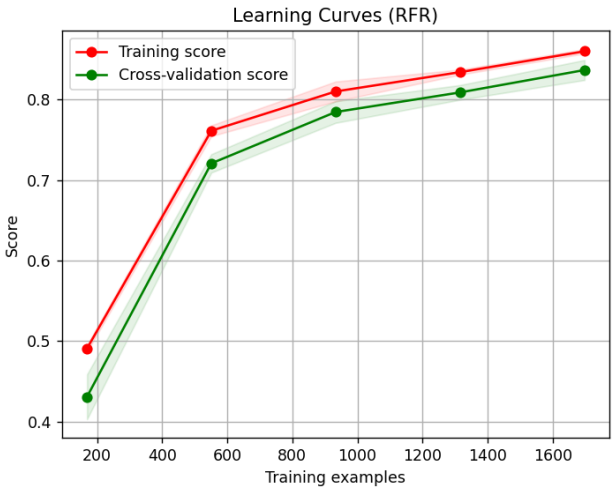


KNN - Lasso



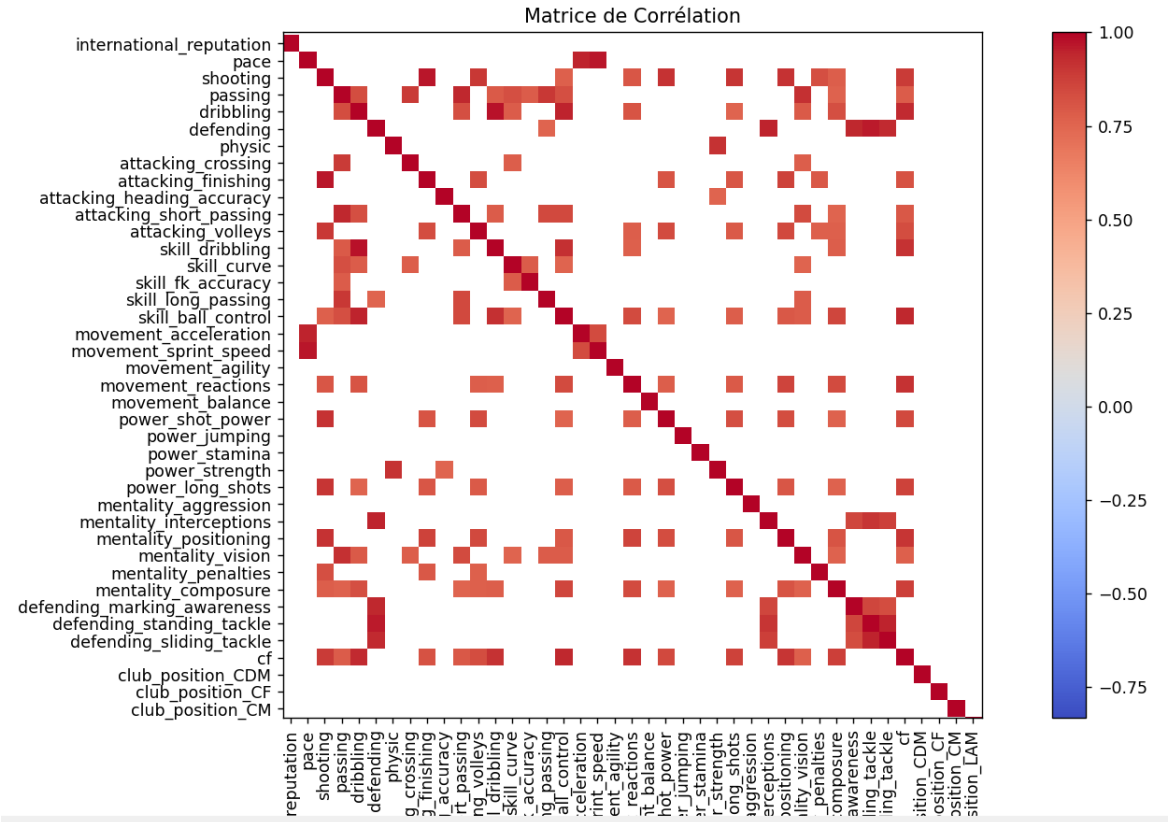
On aperçoit une tendance commune à tous les modèles. On observe à chaque fois que l'erreur d'entraînement est inférieure à l'erreur sur l'ensemble de test. De plus, au fur et à mesure que des données sont ajoutées au modèle, pour les deux sous-ensembles, le MSE diminue en convergeant et se stabilise. Ceci signifie que les modèles se stabilisent.

Random Forest :



Pour le modèle Random Forest, le principe est le même. Pour ce modèle, on mesure le R2. On observe la même tendance que pour les autres modèles, avec une convergence des deux sous-ensembles de données.

Matrice de corrélation des variables :



Le langage de programmation Python a été choisi en raison de sa popularité dans le domaine de la data science et de son large ensemble de bibliothèques open source utiles pour résoudre des tâches courantes.

Une bibliothèque représente une collection de fragments de code réutilisables. Python, en particulier, possède une bibliothèque générale appelée la 'standard library' (https://docs.python.org/3/library/index.html), qui permet d'effectuer les opérations de base courantes. En complément de celle-ci, une multitude d'autres bibliothèques ont été développées par les utilisateurs, offrant des fonctionnalités plus spécifiques. Parmi ces dernières, on trouve des bibliothèques spécialisées dans la manipulation de données et l'apprentissage automatique (machine learning), qui ont été largement exploitées pour réaliser les analyses. En utilisant ces bibliothèques, les utilisateurs peuvent bénéficier d'une gamme étendue de fonctionnalités sans avoir à tout créer de zéro.

Dans le cadre du travail en machine learning, nous avons principalement utilisé la bibliothèque open-source Scikit-learn, reconnue aujourd'hui comme l'une des plus populaires et accessibles dans ce domaine. Cette bibliothèque offre non seulement une vaste gamme d'algorithmes d'apprentissage automatique, mais également des outils de prétraitement des données et des métriques pour évaluer les performances des modèles (source : https://falksangdata.no/wp-content/uploads/2022/07/python-machine-learning-and-deep-learning-with-python-scikit-learn-and-tensorflow-2.pdf).

Toutes les techniques importées depuis la bibliothèque sklearn ont nécessité la définition de leurs paramètres, c'est-à-dire de leurs variables configurables internes, qui influencent la manière dont l'information est traitée. Ces paramètres ont des valeurs par défaut, mais peuvent être modifiés pour mieux correspondre aux données en entrée.

Explications supplémentaires sur les variables :

- ‘Weak_foot’ : sur une échelle de 1 à 5, à quel point le joueur maîtrise son pied plus faible.
- ‘Skill_moves’ : La note d'un joueur en "Skill moves" détermine l'efficacité avec laquelle il peut effectuer des mouvements techniques, ainsi que la variété des mouvements qu'il maîtrise.
- ‘International_reputation’ : Dans la série de jeux vidéo de la FIFA, l'attribut Réputation internationale représente la position et la reconnaissance d'un joueur au sein de la communauté internationale du football.
- Work_rate : Cette variables est divisée en 2. La quantité d’effort effectué en travail défensif et la quantité de travail effectué dans le travail offensif.
- Pace : Représente la vitesse d'un joueur sur le terrain.
- Shooting : Représente la capacité d'un joueur à émettre des tirs dangereux.
- Passing : Représente la qualité de passe à ses coéquipiers.
- Dribbling : Représente la capacité du joueur à contrôler le ballon et à se déplacer avec agilité, évitant les tacles adverses.
- Defending: Représente la capacité globale du joueur en défense, comprenant son marquage, ses tacles et son interception.
- Physical : Représente les attributs physiques du joueur, y compris sa force, son endurance et son saut.
- Attacking_Crossing: Représente la précision du joueur lorsqu'il centre le ballon depuis le côté du terrain pour une tête ou une reprise de volée.

- Attacking_Finishing: Représente la précision et la puissance des tirs du joueur à l'intérieur de la surface de réparation.
- Attacking_Heading_Accuracy: Représente la précision du joueur lorsqu'il tente une tête sur un centre ou un corner.
- Attacking_Short_Passing: Représente la précision du joueur lorsqu'il effectue des passes courtes et à une touche.
- Attacking_Volleys: Représente la précision et la puissance des tirs du joueur lorsqu'il reprend le ballon de volée.
- Skill_Dribbling: Représente la capacité technique du joueur à dribbler en utilisant des feintes et des changements de direction.
- Skill_Curve: Représente la capacité du joueur à effectuer des passes et des centres avec effet.
- Skill_FK_Accuracy: Représente la précision du joueur lorsqu'il tire des coups francs directs.
- Skill_Long_Passing: Représente la précision du joueur lorsqu'il effectue des passes longues et aériennes.
- Skill_Ball_Control: Représente la capacité du joueur à garder le ballon proche de ses pieds et à le protéger des adversaires.
- Movement_Acceleration: Représente la rapidité avec laquelle le joueur peut atteindre sa vitesse maximale.
- Movement_Sprint_Speed: Représente la vitesse de course maximale du joueur.
- Movement_Agility: Représente la capacité du joueur à changer de direction rapidement et à se déplacer avec fluidité.
- Movement_Reactions: Représente la rapidité de réaction du joueur face aux situations de jeu et aux actions adverses.
- Movement_Balance: Représente la capacité du joueur à rester stable et en équilibre lorsqu'il est en possession du ballon ou qu'il est bousculé par un adversaire.
- Power_Shot_Power: Représente la puissance des tirs du joueur, en dehors de la surface de réparation.
- Power_Jumping: Représente la hauteur de saut du joueur, important pour les duels aériens et les têtes.
- Power_Stamina: Représente la capacité du joueur à maintenir son endurance et son niveau d'activité physique tout au long du match.
- Power_Strength: Représente la force physique du joueur, utile pour gagner les duels à l'épaule et protéger le ballon.
- Power_Long_Shots: Représente la précision du joueur lorsqu'il tire de loin.
- Mentality_Aggression: Représente l'agressivité du joueur à la récupération du ballon et dans les duels.
- Mentality_Interceptions: Représente la capacité du joueur à anticiper les passes adverses et à les couper.

Références

Ahsan, M., Kumar Saha, P., & Siddique, Z. (2021). Effect of Data Scaling Methods on Machine Learning Algorithms and Model Performance. *School of Industrial and Systems Engineering, University of Oklahoma, Norman*. Consulté le Avril 2024, sur https://www.mdpi.com/2227-7080/9/3/52

Ali, J., Khan, R., & Ahmad, N. (2012). *Random Forests and Decision Trees.* Consulté le Avril 2024, sur https://www.uetpeshawar.edu.pk/TRP-G/Dr.Nasir-Ahmad-TRP/Journals/2012/Random%20Forests%20and%20Decision%20Trees.pdf

Altman, N., & Krzywinski , M. (2018). The curse(s) of dimensionality. *nature methods, 15*, 399-400. Consulté le Avril 2024, sur https://www.nature.com/articles/s41592-018-0019-x#citeas

Aoki, R., Assunção, R., & Vaz de Melo, P. (2017). *Luck is Hard to Beat: The Difficulty of Sports Prediction*. Applied Data Science Paper, Belo Horizonte. Consulté le Avril 2024, sur http://library.usc.edu.ph/ACM/KKD%202017/pdfs/p1367.pdf

Arashi, M., Roozbeh, M., Gasparini, M., & Hamzah, N. (2021). Ridge regression and its applications in genetic studies. *PLOS ONE*, 16. Consulté le Avril 2024, sur https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0245376

Arif , A. (2019). MASTERING RIDGE REGRESSION: COMPREHENSIVE GUIDE AND PRACTICAL APPLICATIONS. *Dataaspirant*. Consulté le Avril 2024, sur https://dataaspirant.com/ridge-regression/

Arrondel, L., & Duhautois, R. (2022). L'ARGENT DU FOOTBALL. *CENTRE POUR LA RECHERCHE ÉCONOMIQUE ET SES APPLICATIONS Paris*, 33-43. Consulté le Mai 2024, sur https://www.cepremap.fr/depot/2022/09/Opuscule60-texte.pdf

Ati, A., Bouchet, P., & Ayachi Ben Jeddou, R. (2023). Using multi-criteria decision-making and machine learning for football player selection and performance prediction: A systematic review. *HAL open science*, 1-23. Consulté le Mai 2024, sur https://hal.science/hal-04281291

Bee Wah, Y., & Ibrahim, N. (2018). *Feature Selection Methods: Case of Filter and Wrapper Approaches for Maximising Classification Accuracy.* Consulté le Avril 2024, sur https://www.researchgate.net/profile/Shuzlina-Rahman/publication/322920304_Feature_selection_methods_Case_of_filter_and_wrapper_approaches_for_maximising_classification_accuracy/links/5eb17fa545851592d6b9b50c/Feature-selection-methods-Case-of-filter-and-w

Bekraoui, N., & Leger, L. (2010). *Les systèmes d'enregistrement et d'analyse quantitatifs dans le football*. ResearchGate. Consulté le Avril 2024, sur https://www.researchgate.net/publication/246523276_Les_systemes_d'enregistrement_et_d'analyse_quantitatifs_dans_le_football

Bohec, P. (2022). Comment sont décidées les notes des joueurs sur Fifa. *Le Télégramme*. Consulté le Avril 2024, sur https://www.letelegramme.fr/sports/football/toutes-les-infos/comment-sont-decidees-les-notes-des-joueurs-sur-fifa-335876.php

Bohec, P. (2022, Septembre 29). *Comment sont décidées les notes des joueurs sur Fifa*. Consulté le Avril 2024, sur https://www.letelegramme.fr/sports/football/toutes-les-infos/comment-sont-decidees-les-notes-des-joueurs-sur-fifa-335876.php

Bohec, P. (2022, Septembre 29). *Le Télégramme.* Consulté le Avril 2024, sur Le Télégramme: https://www.letelegramme.fr/sports/football/toutes-les-infos/comment-sont-decidees-les-notes-des-joueurs-sur-fifa-335876.php

Buchheit, M. (2014). Integrating different tracking systems in football: multiple camera semi-automatic system, local position measurement and GPS technologies. *J Sports Sci.*, 1844-1857. Consulté le Mai 2024, sur https://pubmed.ncbi.nlm.nih.gov/25093242/

Cao, L. (2017). *Data science : a comprehensive overview. ACM Computing Surveys (CSUR), 50(3), 1-42*. Consulté le Avril 2024, sur https://dl.acm.org/doi/10.1145/3076253

Chandre, B., & Jennet Shinny, D. (2024). Prediction of Football Player Performance Using. *Machine Learning Algorithm*, 1-9. Consulté le Mai 2024, sur https://doi.org/10.21203/rs.3.rs-3995768/v1

Chen, G., & Jin, C. (2015). *A novel wrapper method for feature selection and its applications.* Consulté le Avril 2024, sur https://www.sciencedirect.com/science/article/pii/S0925231215001459?casa_token=UmNN1J42WaYAAAAA:g_0WQKIyiSeOhJPGMBHIQGsqrtO8iNbgi3GwctF8nrr-HUoSsvEvNZUFcjzHZmZSfgh59gA6750

Chiang, W., Liu, X., & Zhang, T. (2018). *A Study of Exact Ridge Regression for Big Data.* Seattle: IEEE Conference Publication. doi:10.1109/BigData.2018.8622274

Chu, X., Ihab, I., Krishnan, S., & Wang, J. (2016). *Data Cleaning: Overview and Emerging Challenges.* Consulté le Avril 2024, sur https://dl.acm.org/doi/abs/10.1145/2882903.2912574

Columbia University Press. (2013, Août). Sports Analytics A Guide for Coaches, Managers, and Other Decision Makers. *COLUMBIA UNIVERSITY PRESS*. Consulté le Avril 2024, sur https://cup.columbia.edu/book/sports-analytics/9780231162920

Dalgalarrondo, S. (2018). *Surveiller et guérir le corps optimal Big Data et performance sportive, pages 99 à 116* . Association pour la Recherche de Synthèse en Sciences Humaines. Consulté le Avril 2024, sur https://www.cairn.info/load_pdf.php?ID_ARTICLE=LHS_207_0099&download=1

Data Rockstars. (2021). Le rôle de la data science dans le sport. *Data Rockstars*. Consulté le Avril 2024, sur https://www.datarockstars.ai/le-role-de-la-data-science-dans-le-sport/

Domingos, P. (2000). A Unified Bias-Variance Decomposition. *Department of Computer Science and Engineering University of Washington*, 1-22. Consulté le Avril 2024, sur https://homes.cs.washington.edu/~pedrod/bvd.pdf

Dorfman, E. (2024). How Much Data Is Required for Machine Learning? *post industria*, 1-5. Consulté le mai 2024, sur https://postindustria.com/how-much-data-is-required-for-machine-learning/

EA Sports FIFA 23. (2023). *FIFA 23 CHAMPIONNATS ET CLUBS.* Consulté le Avril 2024, sur https://www.ea.com/fr-fr/games/fifa/fifa-23/news/fifa-23-all-leagues-clubs-teams-list

El Morr, C., & Ali-Hassan , H. (2019). *Descriptive, Predictive, and Prescriptive Analytics A practical introduction*. Springer. Consulté le Avril 2024, sur https://link.springer.com/chapter/10.1007/978-3-030-04506-7_3#Sec2

Etienne, C. (2019, Octobre 15). *Mais comment sont attribuées les notes sur FIFA ?* Consulté le Avril 2024, sur Micromania: https://www.micromania.fr/fanzone-articles-online-dossiers/mais-comment-sont-attribuees-les-notes-sur-fifa.html

Everitt, B., & Skrondal, A. (2010). *The Cambridge Dictionary of Statistics*. Cambridge University Press. Consulté le Avril 2024, sur https://www.stewartschultz.com/statistics/books/Cambridge%20Dictionary%20Statistics%204th.pdf

Fawcett, T. (2013). *Data Science for Business*. Consulté le Avril 2024, sur https://www.researchgate.net/publication/256438799_Data_Science_for_Business

George, S., & Sumathi, B. (2020). *Grid Search Tuning of Hyperparameters in Random Forest Classifier for Customer Feedback Sentiment Prediction*. Tamilnadu, India. Consulté le Avril 2024, sur https://thesai.org/Downloads/Volume11No9/Paper_20-Grid_Search_Tuning_of_Hyperparameters.pdf

Glebova, E., & Desfontaine, P. (2020). *Sport et technologies numériques : vers de nouvelles expériences* (Vol. Management du sports 3.0; Economica). EDS Desbordes. Consulté le Avril 2024, sur https://www.researchgate.net/profile/Ekaterina-Glebova-6/publication/346017927_Sport_et_technologies_numeriques_vers_de_nouvelles_experiences_spectateur/links/5fb6376792851c933f3d6ae5/Sport-et-technologies-numeriques-vers-de-nouvelles-experiences-spectate

Gonçalves, S. (2024, Février). EA Sports FC 24 bien plus rentable que FIFA, les chiffres astronomiques dévoilés. *GENTSIDE*. Consulté le Avril 2024, sur https://gaming.gentside.com/jeux-video/consoles-et-pc/ea-sports-fc-24-bien-plus-rentable-que-fifa-les-chiffres-astronomiques-devoiles_art40855.html

Hebiri, M., & Lederer, J. (2012). How Correlations Influence Lasso Prediction. *Université Paris-Est – Marne-la-Vallée*, 1-22. Consulté le Avril 2024, sur https://arxiv.org/pdf/1204.1605

Herberger, T., & Litke, C. (2021). *The Impact of Big Data and Sports Analytics on Professional Football: A Systematic Literature Review*. Conference paper, Springer Proceedings in Business and Economics. Consulté le Avril 2024, sur https://link.springer.com/chapter/10.1007/978-3-030-77340-3_12

Herold, M., Goes, F., & Nopp, S. (2019). Machine learning in men's professional football: Current applications and future directions for improving attacking play. *International Journal of Sports Science, 14*(6), 798–817. Consulté le Mai 2024, sur https://www.researchgate.net/publication/336209265_Machine_learning_in_men's_professional_football_Current_applications_and_future_directions_for_improving_attacking_play

Hodson, T. (2022). Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not. *European Geosciences Union, 15*. Consulté le Avril 2024, sur https://gmd.copernicus.org/articles/15/5481/2022/gmd-15-5481-2022-discussion.html

Hou, X. (2020). P2P Borrower Default Identification and Prediction Based on RFE-Multiple Classification Models. *Open Journal of Business and Management, 08*, 15. Consulté le Avril 2024, sur https://www.scirp.org/html/30-1531220_99070.htm

Hsu, H.-H., & Hsieh, C.-W. (2011). *Hybrid feature selection by combining filters and wrappers*. Consulté le Avril 2024, sur https://www.sciencedirect.com/science/article/abs/pii/S0957417410015198

J.P. Keating. (1985). *Encyclopedia of Statistical Sciences (pp.668-674) FirstChapter: Percentiles* . Consulté le Avril 2024, sur https://www.researchgate.net/publication/237009652_Encyclopedia_of_Statistical_Sciences

Jaadi, Z. (2022). When and Why to Standardize Your Data. *BuiltIn*. Consulté le Avril 2024, sur https://builtin.com/data-science/when-and-why-standardize-your-data

Jain, S. (2016). Regression analysis on different mitogenic pathways. *Jaypee University of Information Technology, Solan, H.P.*, 1-7. Consulté le Avril 2024, sur http://www.ir.juit.ac.in:8080/jspui/bitstream/123456789/9097/1/Regression%20Analysis%20on%20Different%20Mitogenic%20Pathways.pdf

Jeon , H., & Oh, S. (2020). Hybrid-Recursive Feature Elimination for Efficient Feature Selection. *Department of Data Science, Dankook University, Yongin*, 10. Consulté le Avril 2024, sur https://www.mdpi.com/2076-3417/10/9/3211

Jovic, A., Brkic, K., & Bogunovic, N. (s.d.). A review of feature selection methods with applications. *IEEE*. Consulté le Avril 2024, sur https://ieeexplore.ieee.org/abstract/document/7160458

Kassel, R. (2022, Juin). *Coefficient de détermination : qu'est ce que c'est et comment s'en servir ?* DataScientest. Consulté le Avril 2024, sur https://datascientest.com/coefficient-de-determination#:~:text=Concr%C3%A8tement%2C%20le%20coefficient%20de%20d%C3%A9termination,ad%C3%A9quation%20avec%20les%20donn%C3%A9es%20collect%C3%A9es.

Kassel, R. (2022). *Qu'est ce que l'erreur quadratique moyenne ?* Consulté le Avril 2024, sur https://datascientest.com/erreur-quadratique-moyenne#:~:text=Elle%20peut%20%C3%AAtre%20difficile%20%C3%A0,observ%C3%A9es%20est%20inf%C3%A9rieure%20%C3%A0%20100.

Kavlakoglu, E. (2024). Apply lasso regression to automate feature selection. *IBM*. Consulté le Avril 2024, sur https://developer.ibm.com/tutorials/awb-lasso-regression-automatic-feature-selection/

Kofi Nti, I., Nyarko-Boateng, O., & Aning, J. (2021). Performance of Machine Learning Algorithms with Different K Values in K-fold Cross-Validation. *I.J. Information Technology and Computer Science*, 61-71. Consulté le Avril 2024, sur https://www.researchgate.net/profile/Isaac-Nti-3/publication/356914937_Performance_of_Machine_Learning_Algorithms_with_Different_K_Values_in_K-fold_Cross-Validation/links/61b3101c19083169cb7f2c17/Performance-of-Machine-Learning-Algorithms-with-Different-K

Komentar, T. (2022, Mars 2). *Coefficient Of Determination Formula*. Récupéré sur keiranjk: https://keiranjk.blogspot.com/2022/03/coefficient-of-determination-formula.html

Kotsilieris , T., Anagnostopoulos, I., & E. Livieris , I. (2022). Special Issue: Regularization Techniques for Machine Learning and Their Applications. *Department of Business Administration, University of the Peloponnese*, 11. Consulté le Avril 2024, sur https://www.mdpi.com/2079-9292/11/4/521

Kouiroukidis, N., & Evangelidis, G. (2011). The Effects of Dimensionality Curse in High Dimensional kNN Search. *IEEE*. Consulté le Avril 2024, sur https://ieeexplore.ieee.org/abstract/document/6065061

Kuhn, M., & Kjell, J. (2013). *Data Pre-processing*. Springer. Consulté le Avril 2024, sur https://link.springer.com/chapter/10.1007/978-1-4614-6849-3_3

Kumar, D. (2024). A Complete understanding of LASSO Regression. *Great Learning*. Consulté le Avril 2024, sur https://www.mygreatlearning.com/blog/understanding-of-lasso-regression/

Kumar, K., & Gandhi, O. (2021). Usage of KNN, Decision Tree and Random Forest Algorithms in Machine Learning and Performance Analysis with a Comparative Measure. *Advances in Intelligent Systems and Computing*, 473-479. Consulté le Avril 2024, sur https://link.springer.com/chapter/10.1007/978-981-15-9516-5_39

Lancelot, R., & Lesnoff, M. (2005). Sélection de modèles avec l'AIC et critères d'information. *CEF-CFR*, 7. Consulté le Avril 2024, sur https://www.cef-cfr.ca/uploads/Reference/alncelotLesnoff.pdf

L'avenir. (2023, Septembre). Fini "FIFA", bienvenue chez "EA Sports FC": voici pourquoi le jeu de football mythique change de nom. *L'avenir*. Consulté le Avril 2024, sur https://www.lavenir.net/actu/conso/multimedia/2023/09/29/fini-fifa-bienvenue-chez-ea-sports-fc-voici-pourquoi-le-jeu-de-football-mythique-change-de-nom-P52YZALC5NCWXFUZPUQLNVK62Y/#:~:text=Apr%C3%A8s%2030%20ans%20de%20jeux,premier%20jeu%20sorti%20en%201993

Leach, L., & Henson , R. (2007). The Use and Impact of Adjusted R2. *Multiple Linear Regression Viewpoints, 1*, 1-12. Consulté le Avril 2024, sur https://www.glmj.org/archives/MLRV_2007_33_1.pdf#page=4

Lebarbier, E., & Mary-Huard, T. (2006). Une introduction au critère BIC : fondements. *JOURNAL DE LA SOCIÉTÉ FRANÇAISE DE STATISTIQUE*, 39-45. Consulté le Avril 2024, sur http://www.numdam.org/article/JSFS_2006__147_1_39_0.pdf

Lecuyer, C. (2022). *La révolution de la Big Data dans le sport !* Consulté le Avril 2024, sur https://blog.mbadmb.com/la-revolution-de-la-big-data-dans-le-sport/#:~:text=La%20Big%20Data%20offre%20de,aux%20sportifs%20de%20haut%20niveau.

Lewis, M. (2004). *Moneyball*. (N. &. Company, Éd.) USA. Consulté le Avril 2024

Loyola-González, O. (2005). Black-Box vs. White-Box: Understanding Their Advantages and Weaknesses From a Practical Point of View. *International Conference on Neural*

Networks and Brain. Consulté le Avril 2024, sur https://ieeexplore.ieee.org/abstract/document/8882211

Luna Dong, X., & Srivastava, D. (2019). *Big data integration*. IEEE. Consulté le Avril 2024, sur https://ieeexplore.ieee.org/abstract/document/6544914

Luo, X., & Xu Yu, J. (2014). *Advanced Data Mining and Applications*. Guilin (China): Springer. Consulté le Avril 2024

Machová, K., Barčák, F., & Bednár , P. (2006). *A Bagging Method using Decision Trees in the Role of Base Classifiers*. Acta Polytechnica Hungarica, Kosice. Consulté le Mai 2024, sur http://epa.niif.hu/02400/02461/00006/pdf/EPA02461_acta_polytechnica_hungarica_2006_02_121-132.pdf

Maharana, K., Mondal, S., & Bhushankumar, N. (2022). *A review: Data pre-processing and data augmentation techniques*. Consulté le Avril 2023, sur https://www.sciencedirect.com/science/article/pii/S2666285X22000565

Mandatory. (2023, Septembre). POURQUOI FIFA 24 CHANGE DE NOM POUR S'APPELLER EA FC 24. *Mandatory*. Consulté le Avril 2024, sur https://www.mandatory.gg/fc-24/pourquoi-fifa-24-change-de-nom-pour-sappeller-ea-fc-24/

Markets and Markets. (2022). *Sport analytic market - size, growth, report & analysis*. Consulté le Avril 2024, sur https://www.marketsandmarkets.com/Market-Reports/sports-analytics-market-35276513.html

Márquez, A. (2022). The Curse of Dimensionality. *Digital Maintenance Management*, 67-86. Consulté le Avril 2024, sur https://link.springer.com/chapter/10.1007/978-3-030-97660-6_7

MathWorks. (s.d.). *MathWorks*. Récupéré sur Surajustement: https://fr.mathworks.com/discovery/overfitting.html

Mayer-Schönberger, V. (2014). *La révolution Big Data*. Consulté le Avril 2024, sur https://www.cairn.info/revue-politique-etrangere-2014-4-page-69.htm#no6

Melkumova, L., & Shatskikh, S. (2017). Comparing Ridge and LASSO estimators for data analysis. *Procedia Engineering, 201*, 746-755. Consulté le Mai 2024, sur https://www.scribbr.fr/references/generateur/dossier/6lqdf9qY2CR4Oy1wPVL89Y/listes/4KrL09sbYnx1xywf7Gm2ZR/

Meraghni, D., & Abdallah, S. (2019). Méthodes de classifications automatiques. *UNIVERSITÉ MOHAMED KHIDER, BISKRA*, 1-50. Consulté le Avril 2024, sur http://archives.univ-biskra.dz/bitstream/123456789/13669/1/nita_manel.pdf

Misra, P., & Yadav, A. (2020). Improving the Classification Accuracy using Recursive Feature Elimination with cross validation. *Department of Computer Science, University of Lucknow, Lucknow (Uttar Pradesh), India*, 1-7. Consulté le Avril 2024, sur http://www.puneetmisra.com/admin/uploads/journals/5f136d202b8ba1.18644117.pdf

Molnar, C. (2023). *Interpretable Machine Learning A Guide for Making Black Box Models Explainable*. Consulté le Avril 2024, sur https://christophm.github.io/interpretable-ml-book/interpretability.html

Mordor Intelligence. (2023). *ports Analytics Market Size & Share Analysis - Growth Trends & Forecasts (2024 - 2029) Source: https://www.mordorintelligence.com/industry-reports/sports-analytics-market*. Mordor Intelligence. Consulté le Avril 2024, sur https://www.mordorintelligence.com/industry-reports/sports-analytics-market

Naujoks, J. (2019). Adventures in cross-validation: splitting and folding with linear regression models. *Medium*. Consulté le Avril 2024, sur https://medium.com/@johnnaujoks/adventures-in-cross-validation-techniques-with-linear-regression-models-75f4e30471

Naujoks, J. (2019, Avril 19). *Medium*. Récupéré sur Adventures in cross-validation: splitting and folding with linear regression models: https://medium.com/@johnnaujoks/adventures-in-cross-validation-techniques-with-linear-regression-models-75f4e30471

Negroponte, N. (1955). *L'Homme numérique*. Paris: Robert Laffont. Consulté le Avril 2024, sur https://www.cairn.info/revue-politique-etrangere-2014-4-page-69.htm#no6

Nicas, J. (2017, Mars 9). *Google Acquires Kaggle for Its Cloud Business*. Récupéré sur WSJ PRO: https://www.wsj.com/articles/google-acquires-kaggle-for-its-cloud-business-1489059658

O'Hara, R., Haylon, L., & Boyle, D. (2023). A Data Analytics Mindset with CRISP-DM. *Strategic Finance*. Consulté le Mai 2024, sur https://www.sfmagazine.com/articles/2023/february/a-data-analytics-mindset-with-crisp-dm

Olavsrud, T. (2022). LaLiga transforms fan experience with AI. *CIO*. Consulté le Mai 2024, sur https://www.cio.com/article/646627/laliga-transforms-fan-experience-with-ai.html

Oxybel, A. (2022). Le big data : une révolution dans le monde du football. *MCI*. Consulté le Avril 2024, sur https://mbamci.com/2022/01/big-data-revolution-football/

Peshawa Jamal , M., & Faraj, R. (2014). Data Normalization and Standardization: A Technical. *The Machine Learning Lab. at Koya University*, 1-6. Consulté le Avril 2024, sur https://www.researchgate.net/profile/Peshawa-Muhammad-Ali/publication/340579135_Data_Normalization_and_Standardization_A_Technical_Report/links/5e91b65d299bf130798fc1bd/Data-Normalization-and-Standardization-A-Technical-Report.pdf

Pykes, K. (2022, November). Sports Analytics: How Different Sports Use Data Analytics. *Data Camp*. Consulté le Septembre 19, 2023, sur https://www.datacamp.com/blog/sports-analytics-how-different-sports-use-data-analysis

QARA. (2019). *Sports Industry Insights*. QARA. Consulté le Avril 2024, sur https://medium.com/qara/sports-industry-report-3244bd253b8

Qlik. (2022). Data transformation. *Qlik*. Consulté le Avril 2024, sur https://www.qlik.com/us/data-management/data-transformation

Rakotomalala, R. (2019). *Régression régularisée*. Lyon. Consulté le Avril 2024, sur https://eric.univ-lyon2.fr/ricco/cours/slides/regularized_regression.pdf

Ridzuan, F., & Mohd Nazmee Wan Zainon, W. (2019). *Procedia Computer Science : A Review on Data Cleansing Methods for Big Data* (Vol. 161). Consulté le Avril 2024, sur https://www.sciencedirect.com/science/article/pii/S1877050919318885

Rodrigues, F., Lourenço, M., Ribeiro, B., & Pereira, F. (2022). Learning Supervised Topic Models for Classification and Regression from Crowds. *IEEE Xplore*. Consulté le Avril 2024, sur https://ieeexplore.ieee.org/abstract/document/7807338?casa_token=__Zf253_s0YAAAAA:18bkE2PfK0U8c-RhZNhrddjbI9zMTHhhkcY148mVn6SO3E4G2n5Quy-OlULFnO9mvzD-H3C1Iw

Rossi, A. (2018). Effective injury forecasting in soccer with GPS training data and machine learning. *PLOS ONE*, 1-13. Consulté le Mai 2024, sur https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0201264

Saltz, J. (2024). CRISP-DM is Still the Most Popular Framework for Executing Data Science Projects. *Data Science Process Alliance*. Consulté le 2024, sur https://www.datascience-pm.com/crisp-dm-still-most-popular/

Saporta, G. (2011). *Traitement de la multicolinéarité en régression*. Paris. Consulté le Mai 2024, sur https://cedric.cnam.fr/~saporta/multicolinearite.pdf

Schröer, C., & Kruse, F. (2021). A Systematic Literature Review on Applying CRISP-DM Process Model. *Procedia Computer Science, 181*, 526-534. Consulté le Avril 2024, sur https://www.sciencedirect.com/science/article/pii/S1877050921002416

Scikit Learn. (s.d.). *Scikit Learn*. Consulté le Mai 2024, sur RandomForestRegressor: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

scikit-learn. (s.d.). Compare the effect of different scalers on data with outliers. *scikit-learn documentation*. Consulté le Avril 2024, sur https://scikit-learn.org/stable/auto_examples/preprocessing/plot_all_scaling.html#:~:text=RobustScaler%20and%20QuantileTransformer%20are%20robust,yield%20approximately%20the%20same%20transformation.

scikit-learn. (KNeighborsRegressor). *scikit-learn*. Consulté le Mai 2024, sur https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsRegressor.html

scikit-learn. (s.d.). *scikit-learn*. Récupéré sur ElasticNet: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.ElasticNet.html

scikit-learn. (s.d.). sklearn.preprocessing.MinMaxScaler. *Documentation scikit-learn*. Consulté le Avril 2024, sur https://scikit-learn.org/0.15/modules/generated/sklearn.preprocessing.MinMaxScaler.html#:~:text=Standardizes%20features%20by%20scaling%20each,i.e.%20between%20zero%20and%20one.&text=where%20min%2C%20max%20%3D%20feature_range.

scikit-learn. (s.d.). *StandardScaler*. Consulté le Avril 2024, sur https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

Sharma, V. (2022). A Study on Data Scaling Methods for Machine Learning. *International Journal for Global Academic & Scientific Research*, 31-42. Consulté le Avril 2024, sur https://journals.icapsr.com/index.php/ijgasr/article/view/4

Shetty, B. (2021). What Is the Curse of Dimensionality? *Built-In*. Consulté le Avril 2024, sur https://builtin.com/data-science/curse-dimensionality

Singh, Y. (2022, Mars). Robust Scaling: Why and How to Use It to Handle Outliers. *Proclus Academy*. Consulté le Avril 2024, sur https://proclusacademy.com/blog/robust-scaler-outliers/#:~:text=Robust%20Scaling%20in%20Action,-We'll%20use&text=The%20standard%20scaler%20moves%20the,ranges%20of%20scaled%20values%2C%20though.&text=The%20robust%20scaler%20produces%20a,values%20than%20the

Stalbunov, V. (2014). INTRODUCTION TO SPORTS ANALYTICS. *University of Toronto Sports Analytics Student Group*. Consulté le Avril 2024, sur https://sportsanalytics.sa.utoronto.ca/2014/12/11/introduction-to-sports-analytics/

Stata. (2019). *BIC note — Calculating and interpreting BIC.* Récupéré sur Stata: https://www.stata.com/manuals/rbicnote.pdf

Stolbunov, V. (2014). INTRODUCTION TO SPORTS ANALYTICS. *University of Toronto Sports Analytics Student Group*. Consulté le Avril 2024, sur https://sportsanalytics.sa.utoronto.ca/2014/12/11/introduction-to-sports-analytics/

Tanioka, K., & Yadohisa , H. (2012). Effect of Data Standardization on the Result of k-Means Clustering. Consulté le Avril 2024, sur https://link.springer.com/chapter/10.1007/978-3-642-24466-7_7

Thibodeau, F. (2021). FIFA | Introduction Générale. *Jeux.ca*. Consulté le Avril 2024, sur https://jeux.ca/sp/fifa-introduction-generale/

Tian, Y., & Zhang, Y. (2022). A comprehensive survey on regularization strategies in machine learning. *Information Fusion, 80*, 146-166. Consulté le Avril 2024, sur https://www.sciencedirect.com/science/article/pii/S156625352100230X?casa_token=WxujokJ7Hs8AAAAA:aPNRJC3sIDqp3B3m6Q7eays3XPulmEaeUji0137tCWqLNbl1V9XS60arfFD6tcN-hT2Aw-yoaHU#sec4

Torgo, L., Ribeiro, R., & Pfahringer, B. (2013). SMOTE for Regression. *Progress in Artificial Intelligence*, 378–389. Consulté le Avril 2024, sur https://link.springer.com/chapter/10.1007/978-3-642-40669-0_33#Bib1

Trouvé, P. (2022, Mai). « EA Sports FC » : de la naissance de « FIFA » à son changement de nom, trente ans d'évolution. *Le Monde*. Consulté le Avril 2024, sur https://www.lemonde.fr/pixels/article/2022/05/10/ea-sports-fc-de-la-naissance-de-fifa-a-son-changement-de-nom-trente-ans-d-evolution_6125527_4408996.html

Unit 21. (2024). White-box Machine Learning How the Model Works & Top Benefits. *Unit 21*. Consulté le Avril 2024, sur https://www.unit21.ai/fraud-aml-dictionary/white-box-machine-learning#:~:text=White%2Dbox%20machine%20learning%20is%20a%20model%20where%20the%20algorithm,to%20reach%20it%20are%20correct.

Vuille, S. (2017). *Le mystère des notes sur «FIFA 18».* lematin. Consulté le Avril 2024, sur https://www.lematin.ch/story/le-mystere-des-notes-sur-fifa-18-164796404199

Wirth, R., & Hipp, J. (2000). *CRISP-DM: Towards a Standard Process Model for Data Mining*. Consulté le Janvier 2024, sur https://www.researchgate.net/publication/239585378_CRISP-DM_Towards_a_standard_process_model_for_data_mining

Zajic, A. (2022, Novembre 29). *What Is Akaike Information Criterion (AIC)?* Récupéré sur Built In: https://www.scribbr.fr/references/generateur/dossier/6Iqdf9qY2CR4Oy1wPVL89Y/listes/4KrL09sbYnx1xywf7Gm2ZR/

Zhang, S., & Li, X. (2017). Learning k for kNN Classification. *ACM Trans. Intell. Syst. Technol*, 8. Consulté le Avril 2024, sur https://dl.acm.org/doi/pdf/10.1145/2990508

Zhang, Z. (2016). Introduction to machine learning: k-nearest neighbors. *Ann Transl Med*, 11. Consulté le Avril 2024, sur https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4916348/

Zou, H., & Hastie, T. (2005). Regularization and Variable Selection Via the Elastic Net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 301–320. Consulté le Avril 2024, sur https://academic.oup.com/jrsssb/article/67/2/301/7109482