



UNIVERSITÉ  
CAEN  
NORMANDIE

UNIVERSITÉ DE CAEN NORMANDIE

PROJET ANNUEL PREMIÈRE ANNÉE MASTER INFORMATIQUE

---

**PRÉDICTION DU MEILLEUR DÉFENSEUR DE  
L'ANNÉE DE LA NBA**

---

*Encadré par :*

Mr Albrecht Zimmermann

*Réalisé par :*

Islem Messili

Bintou Ba

Ayoub Bina

# Table des matières

<b>Liste des Figures</b>	<b>IV</b>
<b>Liste des Abréviations</b>	<b>VI</b>
<b>Introduction générale</b>	<b>1</b>
<b>1 Contexte Générale</b>	<b>2</b>
1.1 Introduction . . . . .	2
1.2 Présentation du projet . . . . .	2
1.2.1 Objectif du projet . . . . .	2
1.2.2 Description détaillée du projet . . . . .	2
1.2.3 Répartition des tâches . . . . .	3
1.2.3.1 Tâches effectuées par Messili Islem . . . . .	3
1.2.3.2 Tâches effectuées par Ba Bintou . . . . .	3
1.2.3.3 Tâches effectuées par Bina Ayoub . . . . .	4
1.3 DPOY (Defensive Player Of the Year) . . . . .	4
1.4 Machine learning . . . . .	5
1.4.1 Environnement Technique . . . . .	5
1.4.2 Conclusion . . . . .	6
<b>2 Pré-traitement des données</b>	<b>7</b>
2.1 Introduction . . . . .	7
2.2 Compréhension des données . . . . .	7
2.2.1 Source de données . . . . .	7
2.2.2 Description des données . . . . .	8
2.2.2.1 Liste des joueurs ayant reçu des votes pour le DPOY . . . . .	8
2.2.2.2 Statistiques Individuelles d'un Joueur : . . . . .	9
2.2.2.3 Statistiques d'une Équipe : . . . . .	11
2.3 Préparation des données . . . . .	12
2.3.1 Extraction des données web ( Web scraping ) . . . . .	12
2.3.1.1 Listes Des Joueurs Qui Ont Reçu Des Votes Pour Le Prix DPOY . . . . .	12
2.3.1.2 Statistiques Individuelles de tous les joueurs . . . . .	12

2.3.1.3	Statistiques d'une Équipe . . . . .	13
2.3.1.4	Sources additionnelles . . . . .	14
2.3.2	Fusion des données . . . . .	14
2.4	Conclusion . . . . .	16
<b>3</b>	<b>Selection des features</b>	<b>17</b>
3.1	Introduction . . . . .	17
3.1.1	Connaissance du domaine . . . . .	17
3.1.2	Pour tout l'ensemble de données . . . . .	17
3.1.2.1	L'information mutuelle (Mutual Information) . . . . .	17
3.1.3	Pearson correlation coefficient (PCC) . . . . .	19
3.1.4	SelectFromModel Pour Random Forest Regressor/XGBoost . . . . .	20
3.1.5	Pour chaque cinq (5) années de l'ensemble de données . . . . .	21
3.1.5.1	Tau de Kendall . . . . .	21
3.1.5.2	Corrélation de Spearman . . . . .	22
3.2	Conclusion . . . . .	23
<b>4</b>	<b>Modèles de prédiction</b>	<b>24</b>
4.1	Introduction . . . . .	24
4.2	Découpage des données . . . . .	24
4.2.1	Approche basée sur des tranches de 5 à 9 ans . . . . .	24
4.2.2	Approche cumulative . . . . .	24
4.2.3	XGBoost Regressor . . . . .	25
4.2.3.1	Définition . . . . .	25
4.2.3.2	Fonctionnement . . . . .	25
4.2.4	Random Forest Regressor . . . . .	26
4.2.4.1	Définition . . . . .	26
4.2.4.2	Fonctionnement . . . . .	26
4.2.5	Entraînement . . . . .	27
4.2.6	Test . . . . .	27
4.3	Conclusion . . . . .	27
<b>5</b>	<b>Évaluation des modèles</b>	<b>28</b>
5.1	Introduction . . . . .	28
5.1.1	Coefficient de détermination R <sup>2</sup> . . . . .	28
5.1.2	Erreur quadratique moyenne MSE . . . . .	28
5.1.3	Limites de ces approches . . . . .	28
5.1.4	Autres méthodes d'évaluation . . . . .	29
5.1.5	Comparaison entre modèles avec périodes de 5 à 9 ans et modèles cumulatif . . . . .	29
5.1.6	Comparaison entre modèles avec Top 10,20,30,40,50 . . . . .	30

---

## TABLE DES MATIÈRES

5.1.7	Comparaison entre les modèles XGBoostRegressor et RandomForestRegressor . . . . .	31
5.1.8	Comparaison entre modèle par méthode de sélection d'attribut . . . . .	31
5.1.9	Le modèle le plus performant . . . . .	31
5.2	Conclusion . . . . .	31
<b>Conclusion Générale</b>		<b>32</b>
<b>Annexe</b>		<b>33</b>
1	Basket-ball . . . . .	33
2	Liste des features . . . . .	35
3	Quelques Analyses Et Visualisation Des Données . . . . .	36

# Table des figures

1.1	Pondération des votes DPOY . . . . .	5
2.1	Sources de données . . . . .	8
2.2	Page d'accueil du site Basketball Reference . . . . .	8
2.3	Liste des joueurs ayant remporté le DPOY . . . . .	9
2.4	Page d'un joueur . . . . .	10
2.5	Statistique per game d'un joueur . . . . .	10
2.6	Statistique per 100 poss d'un joueur . . . . .	10
2.7	Statistique Advanced d'un joueur . . . . .	11
2.8	Page d'une équipe . . . . .	11
2.9	Team and Opponent / Team Misc Stats . . . . .	12
2.10	Liste des joueurs ayant reçu des votes . . . . .	12
2.11	Statistique per game.csv . . . . .	13
2.12	Statistique per 100 poss.csv . . . . .	13
2.13	Statique advanced.csv . . . . .	13
2.14	Team and Opponent Stats.csv . . . . .	14
2.15	Données fusionnées . . . . .	16
3.1	Mutual Information . . . . .	18
3.2	Pearson correlation . . . . .	19
3.3	SelectFromModel . . . . .	21
3.4	MI vs MI Kendall Tau . . . . .	22
3.5	MI vs Pearson Kendall Tau . . . . .	22
3.6	MI vs MI Spearman . . . . .	23
3.7	MI vs Pearson Spearman . . . . .	23
3.8	Pearson vs Pearson Spearman . . . . .	23
4.1	XGBoost Regressor . . . . .	25
4.2	Random forest Regressor . . . . .	26
5.1	Prédictions des classements des joueurs de l'année 2017 . . . . .	29
5.2	Nombre de classement correctement prédits . . . . .	30

---

## TABLE DES FIGURES

5.3	Prédictions de classement des joueurs de toutes les années . . . . .	30
5.4	La distribution des joueurs ayant eu des votes en fonctions de leurs âge . . . . .	37
5.5	La distribution des caractéristiques d'un joueur ayant remporté le prix . . . . .	37
5.6	La distribution des caractéristiques d'un joueur n'ayant pas reçu de prix . . . . .	37
5.7	La distribution des joueurs ayant reçu des votes vs ceux qui n'ont pas reçu . . . . .	38

# Liste des Abréviations

<b>DPOY</b>	<i>Defensive Player of the Year</i>
<b>ML</b>	<i>Machine Learning</i>
<b>NBA</b>	<i>National Basketball Association</i>
<b>PCC</b>	<i>Pearson correlation coefficient</i>
<b>MI</b>	<i>Mutual Information</i>
<b>RDF</b>	<i>Random forest Regressor</i>

## Introduction générale

À la fin de chaque saison, la NBA récompense ses joueurs exceptionnels avec des prix tels que le Most Valuable Player (MVP) ou le Rookie of the Year (ROY). L'un de ces honneurs est le titre de Défenseur de l'Année (DPOY), attribué au joueur ayant eu le plus grand impact défensif au cours de la saison régulière. Historiquement, ce prix a été décerné à des ailiers forts et des pivots pour leurs performances dans le domaine du rebond et de la défense près du panier.

Cependant, le processus de sélection du DPOY a souvent été entouré d'ambiguïté. Il est difficile de distinguer l'impact individuel d'un joueur de celui de ses coéquipiers. De plus, la subjectivité des votes provenant d'un panel de plus de 120 membres des médias de la NBA à travers l'Amérique du Nord ajoute une complexité supplémentaire. Chaque membre du panel dispose de trois votes, et le joueur cumulant le plus de points remporte le prix. Bien que cette approche soit traditionnelle, elle soulève des questions sur la transparence du processus de vote et la justesse des résultats. C'est dans ce contexte que s'inscrit notre projet de fin d'études. En exploitant les données fournies par plusieurs sources, notre objectif est de développer un modèle prédictif capable de déterminer le lauréat du DPOY de manière plus objective.

La structure de notre projet comprend cinq chapitres, détaillés comme suit :

- Le premier chapitre offre une vue d'ensemble du contexte global du projet.
- Le deuxième chapitre abord les différentes phases de pré-traitement, depuis l'extraction jusqu'à la préparation de nos données.
- Le troisième chapitre est consacré à l'exploration et l'exploitation des méthodes de sélection des attributs pertinents pour la base d'apprentissage.
- Le quatrième chapitre se concentrera sur l'application de différents algorithmes pour créer les modèles de prédiction.
- Enfin, le cinquième chapitre sera dédié à l'évaluation des différents modèles de prédiction obtenus au chapitre quatre.

# **Chapitre 1**

## **Contexte Générale**

### **1.1 Introduction**

Dans ce tout premier chapitre, nous présenterons, entre autres, l'objectif du projet ainsi qu'une description détaillée de ce dernier. Par la suite, nous aborderons les fondements du Defensive Player of the Year et les concepts de base du machine learning. Enfin, nous conclurons ce premier chapitre en évoquant les outils utilisés pour la mise en œuvre de ce travail.

### **1.2 Présentation du projet**

#### **1.2.1 Objectif du projet**

L'objectif de ce projet est d'exploiter des algorithmes de machine learning pour développer un modèle de prédiction. Ce modèle permettra de prédire le joueur qui remportera le titre de meilleur défenseur de l'année de la NBA, également connu sous l'acronyme DPOY (Defensive Player of the Year).

#### **1.2.2 Description détaillée du projet**

Dans la première phase de notre projet, nous nous attacherons à rassembler un ensemble de données englobant une période d'environ trente (30) années (1993-2023), provenant de diverses sources. Une fois ces données rassemblées, nous les soumettrons à des pré traitements indispensables pour les adapter aux modèles de prédictions. Par la suite, nous procéderons à une analyse des caractéristiques (features), cherchant à identifier celles qui seront une importance capitale pour assurer la performance optimale de notre modèle. Nous mettrons en œuvre des

techniques d'apprentissage supervisé dans le but de prédire le joueur qui sera élu Joueur Défensif de l'année (DPOY) pour chaque saison.

Nous explorerons plusieurs types d'algorithmes et évaluerons ces modèles afin de déterminer leur efficacité respective.

### 1.2.3 Répartition des tâches

#### 1.2.3.1 Tâches effectuées par Messili Islem

- Automatisation de l'extraction des données depuis les ressources statistiques principales et additionnelles à l'aide de Selenium.
- Nettoyage et fusion des données pour créer un grand dataset.
- Utilisation des deux méthodes de sélection d'attributs PearsonR et SelectFromModel avec RandomForestRegressor pour déterminer l'importance des attributs.
- Implémentation de la méthode Kendall-tau.
- Entraînement des modèles XGBoost avec les attributs sélectionnés par les deux méthodes SelectFromModel et PearsonR.
- Écriture d'un script pour concaténer les résultats prédits et évaluer les modèles.
- Comparaison entre un joueur ayant remporté le titre de DPOY et un autre perdant, et réalisation de différents graphiques de distributions.
- Participation à l'interprétation des résultats des modèles de prédiction.
- Participation à l'utilisation de la méthode de sélection d'attributs Mutuelle Information et SelectFromModel avec RandomForestRegressor pour déterminer l'importance des attributs.

#### 1.2.3.2 Tâches effectuées par Ba Bintou

- Participation à l'extraction des données principales de basketball-reference.
- Utilisation de la méthode de sélection d'attributs PearsonR et SelectFromModel avec RandomForestRegressor pour déterminer l'importance des attributs.

- Entraînement des modèles Random Forest avec les attributs sélectionnés par SelectFromModel et Mutual Information.
- Participation à la création des graphiques.
- Participation à l'évaluation des différents modèles entraînés.
- Participation à l'interprétation des résultats donnés par la méthode de Kendall-tau.

#### 1.2.3.3 Tâches effectuées par Bina Ayoub

- Participation à l'extraction des ressources additionnelles.
- Utilisation de la méthode de sélection d'attributs RidgeCV pour déterminer l'importance des attributs.
- Participation à l'interprétation des résultats des modèles de prédiction.
- Participation à la création des graphiques.
- Participation à l'interprétation des résultats donnés par la méthode de Kendall-tau.
- Participation à l'évaluation des différents modèles entraînés.

### 1.3 DPOY (Defensive Player Of the Year)

Le trophée du DPOY [1] est décerné chaque année au joueur considéré comme le meilleur défenseur de la saison régulière depuis 1982-1983. La remise du trophée se fait généralement au début des séries éliminatoires.

Le lauréat est choisi par un groupe de plus de 120 de la presse sportive et de la télévision américaine et canadienne. Chacun de ces journalistes vote pour trois joueurs, attribuant cinq points au premier, trois points au deuxième et un point au troisième. Le joueur qui a accumulé le plus grand nombre de points se voit remettre le trophée, symbolisant ainsi son impact défensif exceptionnel tout au long de la saison. La figure 1.1 illustre comment les votes pour le DPOY sont pondérés

<b>Vote</b>	<b>Points</b>
First-place	5
Second-place	3
Third-place	1

FIGURE 1.1 – Pondération des votes DPOY

## 1.4 Machine learning

Le ML [2] ou apprentissage automatique est un domaine scientifique, et plus particulièrement une sous-catégorie de l'intelligence artificielle. Elle consiste à laisser des algorithmes découvrir des « patterns », à savoir des motifs récurrents, dans les ensembles de données. Ces données peuvent être des chiffres, des mots, des images, des statistiques... .

### 1.4.1 Environnement Technique

- **Python** [3] : est un langage de programmation de haut niveau interprété. Il est très sollicité par une large communauté de développeurs et de programmeurs. La syntaxe simple et facile à apprendre de Python met l'accent sur la lisibilité et réduit donc le coût de maintenance du programme. Les bibliothèques (packages) python encouragent la modularité et la réutilisabilité des codes.
- **Requests** : est une bibliothèque Python simplifiant les requêtes HTTP. Elle facilite l'envoi de demandes web avec des fonctionnalités comme la gestion automatique des sessions et des cookies.
- **Beautiful Soup (beautifulsoup4)** : est une bibliothèque Python, facilite l'extraction d'informations d'HTML et XML. Elle simplifie la recherche dans le code source des pages web pour une manipulation aisée des données.
- **Pandas** : bibliothèque Python, est essentielle pour la manipulation et l'analyse de données. Elle offre des structures puissantes, comme les DataFrames, simplifiant le traitement des données tabulaires.

- **lxml** : lxml, bibliothèque Python, est dédiée au traitement rapide de fichiers XML et HTML. Elle permet d'analyser, manipuler et générer des documents de manière flexible.
- **Selenium** : est suite d'outils open-source, automatise les tests logiciels sur des applications web. Elle simule le comportement d'un utilisateur réel, utilisée aussi pour le web scraping et la surveillance automatisée de sites web.
- **ChromeDriver** : est un composant de Selenium spécialement conçu pour automatiser les interactions avec le navigateur web Google Chrome. Il agit comme une liaison pour envoyer des commandes et récupérer des informations.
- **Matplotlib** : est une bibliothèque de visualisation de données pour le langage de programmation Python. Elle offre une gamme de fonctions et de méthodes pour créer des graphiques et des visualisations, telles que des lignes, des barres, des histogrammes et bien plus.
- **Scikit-learn** : est une bibliothèque Python open-source dédiée au machine learning, offrant une gamme d'outils pour l'analyse de données et la création de modèles prédictifs.

#### 1.4.2 Conclusion

Dans ce premier chapitre, nous avons défini les objectifs du projet, détaillé sa structure, introduit des concepts clés tels que le machine learning et le "Defensive Player of the Year", et présenté les outils technologiques à utiliser. Cette compréhension préalable forme le socle de la prochaine phase de pré-traitement des données.

## Chapitre 2

# Pré-traitement des données

### 2.1 Introduction

Ce chapitre sera consacré à une étape cruciale de notre projet : le pré-traitement des données. Nous utiliserons différentes sources de données pour créer un ensemble de données prêt pour les phases suivantes.

### 2.2 Compréhension des données

#### 2.2.1 Source de données

- **Source principale :** Notre ensemble de données provient principalement du site Basketball Reference [4], une ressource en ligne utilisée pour obtenir des statistiques et des informations sur le basketball. Ce site offre des données détaillées sur les joueurs, les équipes, les saisons et les matchs de la NBA, ainsi que sur d'autres ligues professionnelles de basketball. Nous avons collecté les données de la saison 1993 à 2023 pour notre projet.
- **Sources additionnelles :** Nous avons renforcé notre ensemble de données par d'autres informations provenant d'autres sources, notamment : NBA.COM [5], ESPN [6], Github [7] et BBALL-INDEX [8].



FIGURE 2.1 – Sources de données

The screenshot shows the Basketball Reference homepage with the title "Basketball Stats and History". The main content is the "2023-24 NBA Standings" table, which lists 30 NBA teams with their win-loss records. Below the table is a "Play Immaculate Grid" interactive game. The right sidebar features advertisements for "Stathead Basketball Powered By Basketball Reference" and "THE MOST POWERFUL SPORTS DATABASE ON THE INTERNET".

Team	W	L	Team	W	L
BOS (1)	54	14	DEN (2)	47	20
MIL (2)	44	24	MIN (2)	47	21
CLE (3)	43	25	DEN (3)	47	21
NOP (4)	41	27	LAC (4)	42	25
ORL (5)	40	28	NOP (5)	41	26
PHL (6)	38	30	SAC (6)	39	28
IND (7)	38	31	DAL (7)	39	29
MIA (8)	37	31	PHO (8)	39	29
CHI (9)	34	35	LAL (9)	37	32
ATL (10)	30	38	GSW (10)	35	32
BKN (11)	26	42	HOU (11)	32	35
TOR (12)	23	45	UTA (12)	29	39
CHO (13)	17	51	MEM (13)	23	46
DET (14)	12	56	POR (14)	19	49
WAS (15)	11	57	SAS (15)	19	53

FIGURE 2.2 – Page d'accueil du site Basketball Reference

## 2.2.2 Description des données

### 2.2.2.1 Liste des joueurs ayant reçu des votes pour le DPOY

Nous avons collecté la liste des joueurs ayant reçu des votes au cours des 30 années sur lesquelles notre étude se concentre. Cette liste inclut des indicateurs tels que le classement (Rang) dans le vote, les points remportés (pts won) représentant le nombre de points attribués au joueur, et le nombre de premières places reçues (first). La figure 2.3 présente un exemple de la liste des joueurs qui ont reçu des votes pour DPOY en 2023.

NBA & ABA Defensive Player of the Year (Hakeem Olajuwon Trophy) Award Winners																	
Since 2022-23 the NBA DPOY is awarded the Hakeem Olajuwon Trophy																	
NBA Winners Share & Export ▾ Glossary																	
Season	Lg	Player	Voting	Age	Tm	G	MP	PTS	TRB	AST	STL	BLK	FG%	3P%	FT%	WS WS/48	
2022-23	NBA	Jaren Jackson Jr.	(V)	23	MEM	63	28.4	18.6	6.8	1.0	3.0	.506	.355	.788	6.6	.177	
2021-22	NBA	Marcus Smart	(V)	27	BOS	71	32.3	12.1	3.8	5.9	1.7	.03	.418	.331	.793	5.6	.116
2020-21	NBA	Rudy Gobert	(V)	28	UTA	71	30.8	14.3	13.5	1.3	0.6	2.7	.675	.000	.623	11.3	.248
2019-20	NBA	Giannis Antetokounmpo	(V)	25	MIL	63	30.4	29.5	13.6	5.6	1.0	1.0	.553	.304	.633	11.1	.279
2018-19	NBA	Rudy Gobert	(V)	26	UTA	81	31.8	15.9	12.9	2.0	0.8	2.3	.669	.000	.636	14.4	.268
2017-18	NBA	Rudy Gobert	(V)	25	UTA	56	32.4	13.5	10.7	1.4	0.8	2.3	.622	.000	.682	8.1	.214
2016-17	NBA	Draymond Green	(V)	26	GSW	76	32.5	10.2	7.9	7.0	2.0	1.4	.418	.308	.709	8.2	.160
2015-16	NBA	Kawhi Leonard	(V)	24	SAS	72	33.1	21.2	6.8	2.6	1.8	1.0	.506	.443	.874	13.7	.277
2014-15	NBA	Kawhi Leonard	(V)	23	SAS	64	31.8	16.5	7.2	2.5	2.3	0.8	.479	.349	.802	8.6	.204
2013-14	NBA	Joakim Noah	(V)	28	CHI	80	35.3	12.6	11.3	5.4	1.2	1.5	.475	.000	.737	11.2	.190
2012-13	NBA	Marc Gasol	(V)	28	MEM	80	35.0	14.1	7.8	4.0	1.0	1.7	.494	.071	.848	11.5	.197
2011-12	NBA	Tyson Chandler	(V)	29	NYK	62	33.2	11.3	9.9	0.9	0.9	1.4	.679	.000	.689	9.5	.220
2010-11	NBA	Dwight Howard	(V)	25	ORL	78	37.6	22.9	14.1	1.4	1.4	2.4	.593	.000	.596	14.4	.235
2009-10	NBA	Dwight Howard	(V)	24	ORL	82	34.7	18.3	13.2	1.8	0.9	2.8	.612	.000	.592	13.2	.223
2008-09	NBA	Dwight Howard	(V)	23	ORL	79	35.7	20.6	13.8	1.4	1.0	2.9	.572	.000	.594	13.8	.234
2007-08	NBA	Kevin Garnett	(V)	31	BOS	71	32.8	18.8	9.2	3.4	1.4	1.3	.539	.000	.801	12.9	.265
2006-07	NBA	Marcus Camby	(V)	32	DEN	70	33.8	11.2	11.7	3.2	1.2	3.3	.473	.000	.729	7.6	.155
2005-06	NBA	Ben Wallace	(V)	31	DET	82	35.2	7.3	11.3	1.9	1.8	2.2	.510	.000	.416	10.1	.168
2004-05	NBA	Ben Wallace	(V)	30	DET	74	36.1	9.7	12.2	1.7	1.4	2.4	.453	.111	.428	8.5	.153
2003-04	NBA	Metta World Peace	(V)	24	IND	73	37.2	18.3	5.3	3.7	2.1	0.7	.421	.310	.733	8.0	.141
2002-03	NBA	Ben Wallace	(V)	28	DET	73	39.4	6.9	15.4	1.6	1.4	3.2	.481	.167	.450	10.6	.176
2001-02	NBA	Ben Wallace	(V)	27	DET	80	36.5	7.6	13.0	1.4	1.7	3.5	.531	.000	.423	11.6	.190

FIGURE 2.3 – Liste des joueurs ayant remporté le DPOY

#### 2.2.2.2 Statistiques Individuelles d'un Joueur :

Nous avons collecté des données sur les performances individuelles des joueurs :

- **Par Match (Per game)** : Les statistiques "per game" constituent l'évaluation de la performance individuelle des joueurs à chaque match. Les principales variables incluses sont l'âge du joueur, les points marqués par match (PTS), les minutes jouées par match (MP), les rebonds par match (TRB), le nom de l'équipe du joueur (Tm), les passes décisives par match (AST), et les interceptions par match (STL).
- **Par 100 Possessions (per 100 poss)** : Le classement "per 100 poss" offre une mesure de la performance d'un joueur pour chaque tranche de 100 possessions, indépendamment du temps de jeu réel. Les statistiques incluent, entre autres, les rebonds par 100 possessions (REB), les passes décisives par 100 possessions (AST), et les rebonds défensifs par 100 possessions de l'équipe (DRB).
- **Avancées (Advanced)** : Les statistiques avancées aident à comprendre plus en détail comment chaque joueur performe, individuellement et en équipe. Par exemple, le PER (Player Efficiency Rating) mesure l'efficacité d'un joueur en prenant en compte des choses comme les points marqués, les passes, les rebonds, les interceptions et les contres. De la même manière, les Win Shares (WS) évaluent comment un joueur contribue aux victoires de son équipe en combinant différents aspects de sa performance. Ces chiffres plus avancés

donnent une vision plus précise des compétences individuelles et de l'impact d'un joueur sur le jeu.

The screenshot shows the Basketball Reference website for LeBron James. At the top, there's a navigation bar with links like 'Players', 'Teams', 'Seasons', 'Leaders', 'Scores', 'WNBA', 'Draft', 'Stathead', 'Newsletter', and 'Full Site Menu Below'. The main content area features a large photo of LeBron James and his bio: 'LeBron Raymone James - Twitter, KingJames - Instagram, knoxjames'. It includes his position as Small Forward, Power Forward, Point Guard, Center, and Shooting Guard; shooting style as 'Shoots: Right'; height as 6 ft 2 1/2 in (200cm); weight as 113kg; and birth date as December 30, 1984. His high school is St. Vincent-St. Mary in Akron, Ohio. Below the bio is a summary table for the 2023-24 season and career statistics. A sidebar on the right is titled 'Protégez vos clients grâce à la sécurité intégrée multicouche' and includes a 'Créer un compte' button. The bottom of the page has a 'LeBron James Overview' section with links to 'Game Logs', 'Splits', 'Shooting', 'Lineups', 'On/Off', 'More', and '2023-24 Lakers'.

FIGURE 2.4 – Page d'un joueur

This screenshot displays LeBron James' game logs from basketball-reference.com. The table is titled 'LeBron James Overview' and includes columns for 'Regular Season' and 'Playoffs'. It lists various seasons from 2003-04 to 2023-24, along with career statistics. The table includes detailed stats for each game, such as Age, Tm, Lg, Pos, G, GS, MP, FG, FGA, FG%, FT, FTA, FT%, 3P, 3PA, 3P%, 2P, 2PA, 2P%, FT, FTA, FT%, ORB, DRB, TRB, AST, STL, BLK, TOV, PF, PTS, and OTRtg. A note at the top indicates that bolded values represent league leaders. The bottom of the table shows the total career statistics for LeBron James.

FIGURE 2.5 – Statistique par game d'un joueur

This screenshot shows LeBron James' per 100 possessions statistics from basketball-reference.com. The table is titled 'LeBron James Overview' and includes columns for 'Regular Season' and 'Playoffs'. It lists various seasons from 2003-04 to 2023-24, along with career statistics. The table includes detailed stats for each game, such as Age, Tm, Lg, Pos, G, GS, MP, FG, FGA, FG%, FT, FTA, FT%, 3P, 3PA, 3P%, 2P, 2PA, 2P%, FT, FTA, FT%, ORB, DRB, TRB, AST, STL, BLK, TOV, PF, PTS, and OTRtg. A note at the top indicates that bolded values represent league leaders. The bottom of the table shows the total career statistics for LeBron James.

FIGURE 2.6 – Statistique par 100 poss d'un joueur

CHAPITRE 2. PRÉ-TRAITEMENT DES DONNÉES

LeBron James Overview		Game Logs	Spills	Shooting	Lineups	On/Off	More	2023-24 Lakers	Back to top ▾																	
Advanced		Bold indicates league leader	Share & Export	▼ Glossary																						
Regular Season		Playoffs																								
Season	Age	Lg	Po <sup>s</sup>	G	MP	PER	TS% <sup>1</sup>	SPR%	FT%	ORB% <sup>2</sup>	DRB% <sup>2</sup>	TRB% <sup>2</sup>	AST% <sup>2</sup>	STL% <sup>2</sup>	BLK% <sup>2</sup>	TOV% <sup>2</sup>	USG% <sup>2</sup>	OWS	DWS	WS	WS/48	OBPM	DBPM	VORP		
2003–04	19	NBA	SG	79	3122	18.3	488–445	306	3.5	11.6	7.6	27.8	2.2	1.3	13.9	28.2	2.4	2.6	5.1	.078	2.3	1.6	1.7	2.9		
2004–05 <sup>3</sup>	20	NBA	SF	338	25.7	554–538	370	3.8	17.0	10.2	32.9	2.8	1.1	11.8	29.7	9.7	4.6	14.3	.203	7.0	1.7	8.6	9.1			
2005–06 <sup>4</sup>	21	CLE	SF	79	3361	28.1	568–502	208	4.6	17.1	9.8	32.8	2.0	1.5	10.7	33.1	12.0	4.3	16.3	.232	7.5	1.6	9.1	9.4		
2006–07 <sup>5</sup>	22	CLE	SF	78	3190	24.5	591–551	492	1.91	3.0	16.6	9.6	29.1	2.1	1.3	11.5	31.0	8.0	5.7	13.7	.206	5.0	2.2	8.1	8.1	
2007–08 <sup>6</sup>	23	CLE	SF	79	3027	29.1	568–219	470	4.9	17.8	11.1	37.3	2.4	2.1	11.4	33.5	10.7	4.6	15.2	.242	8.2	2.6	10.9	9.8		
2008–09 <sup>7</sup>	24	CLE	SF	71	3054	28.0	559–514	370	2.0	18.9	11.5	38.0	2.4	2.4	11.1	33.4	13.7	6.3	15.0	.318	9.3	3.7	13.2	11.6		
2009–10 <sup>8</sup>	25	CLE	SF	79	3062	28.4	594–554	253	2.0	18.9	11.5	38.0	2.4	2.4	11.1	33.4	13.7	6.3	15.0	.309	9.1	3.7	13.2	11.6		
2010–11 <sup>9</sup>	26	NBA	SF	79	3083	27.2	594–584	446	3.3	18.7	11.4	37.6	2.1	1.3	12.8	31.5	10.3	5.3	15.6	.244	6.3	1.6	8.1	7.0		
2011–12 <sup>10</sup>	27	NBA	SF	82	3236	20.7	565	117	320	5.0	19.7	12.6	33.4	2.6	1.7	12.3	32.0	10.0	4.5	14.5	.209	8.2	3.2	10.9	7.6	
2012–13 <sup>11</sup>	28	NBA	SF	76	2877	31.6	640	180	395	4.4	20.8	13.1	36.4	2.4	1.9	12.4	35.2	16.0	4.7	19.3	.322	9.3	2.4	11.7	9.9	
2013–14 <sup>12</sup>	29	NBA	SF	77	2902	29.3	649	226	412	3.6	18.9	11.5	32.0	2.2	0.8	14.4	31.0	12.3	3.7	15.9	.264	7.8	1.1	8.8	7.9	
2014–15 <sup>13</sup>	30	CLE	SF	69	2493	25.9	577	265	413	2.4	16.6	9.6	38.6	2.3	1.6	15.3	32.3	7.4	2.9	10.4	.199	6.1	1.0	7.1	5.7	
2015–16 <sup>14</sup>	31	CLE	SF	76	2709	27.5	598	199	347	4.7	18.8	11.8	36.0	2.0	1.5	13.2	31.4	9.6	4.0	13.6	.242	7.0	2.0	9.0	7.5	
2016–17 <sup>15</sup>	32	CLE	SF	74	2794	27.0	619	654	355	4.0	20.7	12.6	41.3	1.6	1.3	16.1	30.0	9.8	3.0	12.9	.221	6.4	1.2	7.6	6.7	
2017–18 <sup>16</sup>	33	CLE	SF	79	<b>FB</b> 3020	28.6	621	257	338	3.7	22.3	13.1	44.4	1.9	2.0	16.1	31.6	11.0	3.0	14.0	.221	7.3	1.4	8.7	<b>8.2</b>	
2018–19 <sup>17</sup>	34	NBA	SF	55	1937	25.6	589	288	321	3.1	12.3	12.4	39.4	1.7	1.4	13.3	31.6	4.7	2.6	7.2	.179	6.4	1.8	7.0	4.9	
2019–20 <sup>18</sup>	35	NBA	PG	76	2316	25.5	598	289	321	3.0	12.3	12.4	39.4	1.7	1.4	15.1	31.5	6.2	3.6	9.8	.204	6.6	1.8	8.4	6.1	
2020–21 <sup>19</sup>	36	NBA	PG	45	1504	24.2	662	304	310	2.2	12.6	12.6	39.1	1.5	1.5	15.2	31.9	3.0	2.6	5.6	.179	5.9	2.1	8.1	3.8	
2021–22 <sup>20</sup>	37	NBA	C	56	2086	26.2	561	367	295	3.3	20.4	11.8	30.6	1.7	2.5	12.5	32.3	5.2	2.3	7.5	.172	6.9	0.8	7.7	5.1	
2022–23 <sup>21</sup>	38	NBA	PG	55	1954	23.9	583	309	268	3.7	20.8	12.5	35.5	1.2	1.4	11.6	33.3	3.2	2.4	5.6	.138	5.5	0.6	6.1	4.0	
2022–24 <sup>22</sup>	39	NBA	PG	60	2106	23.5	561	293	302	2.8	19.2	11.4	36.5	1.7	1.4	13.7	29.2	4.5	2.3	6.8	.155	5.6	0.8	6.4	4.5	
Career	NBA			1481	56159	27.0	598	336	388	3.6	11.8	11.3	36.5	2.1	1.6	13.1	31.5	18.1	10.3	26.0	.224	7.0	1.7	8.7	15.0	
11 seasons	CLE	SF		3331	3770	27.0	578	211	408	3.6	17.0	10.8	36.1	2.2	1.6	12.9	31.7	10.7	4.6	16.3	.154	.223	6.9	1.8	8.7	8.95
6 seasons	CLE	SF		339	11901	26.4	598	324	303	5.1	21.2	12.2	38.6	1.6	1.6	13.0	31.1	26.7	15.8	42.6	.172	6.2	1.5	7.3	28.4	
4 seasons	NBA	PG		294	11108	26.6	622	184	246	4.0	19.5	12.1	34.2	2.3	1.4	13.5	31.1	47.2	18.1	65.3	.281	7.0	1.0	9.8	33.1	

FIGURE 2.7 – Statistique Advanced d'un joueur

### 2.2.2.3 Statistiques d'une Équipe :

- **Statistiques de l'équipe et de l'adversaire (Team and Opponent Stats)** : Fait référence aux données collectives de l'équipe et aux statistiques de ses adversaires. Ces statistiques fournissent une vue d'ensemble de la performance d'une équipe ainsi que de la manière dont elle se compare à ses concurrents. Parmi ces indicateurs, on trouve les performances telles que ORB (Offensive Rebounds), mesurant les rebonds pris du côté offensif, et DRB (Defensive Rebounds), qui indique le nombre de rebonds pris du côté défensif.
  - **Statistiques diverses de l'équipe (Team Misc)** : Fait référence à une catégorie de statistiques regroupant des mesures variées qui ne rentrent pas nécessairement dans les catégories traditionnelles telles que les points, les rebonds, ou les passes décisives. Ces statistiques diverses peuvent englober une gamme d'indicateurs offrant une perspective plus complète sur la performance d'une équipe.

The screenshot shows the basketball-reference.com website. At the top, there's a navigation bar with links for Home, About, Contact, Log In, and Sign Up. Below that is a search bar with placeholder text "Enter Person, Team, Section, etc." and a "Search" button. The main content area features a large banner for "BASKETBALL REFERENCE" with a magnifying glass icon. Below the banner, there's a navigation menu with links for Players, Teams, Seasons, Leaders, Score (with a red notification dot), WNBA, Draft, Stathead, Newsletter, and a "Full Site Menu Below". A "Create Account" and "Add File Logo" link are also present. The main title "CHICAGO 2023-24 Chicago Bulls Roster and Stats" is displayed in large, bold letters. Underneath the title, there's a "Previous Season" link. The page contains detailed information about the Chicago Bulls' 2023-24 season, including their record (18-34), last game (W 116-108 vs. HOU), home record (11-17), away record (7-23), coach (Tom Thibodeau), executive (Antanas Kavaliauskas), and team statistics for Points Scored (111.6), Points Allowed (110.5), and other metrics like Win %, Opp PTS/G, and Net Rating. It also lists their attendance (1,194,381), box office revenue (\$126.00M), and arena usage (United Center). A "GROUPON" advertisement for "Tacos Tacos" is visible on the right side of the page.

FIGURE 2.8 – Page d'une équipe

Team and Opponent Stats																								
Ranks are per game (except for MP, which are total) and sorted descending (except for TOV and PF); opponents ranked are flipped; year/year calculations are also per game Share & Export ▾ Glossary																								
		G	MP	FG	FGA	FG%	3P	3PA	3P%	2P	2PA	2P%	FT	FTA	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS
Team	61	14640	2516	5423	.464	798	2171	.368	1718	3252	.528	1090	1392	.783	782	2009	2791	1459	444	243	807	1086	6920	
Team/G	240.0	41.2	88.9	.464	13.1	35.6	.368	28.2	53.3	.528	17.9	.22.8	.783	12.8	.32.9	.45.8	23.9	7.3	4.0	13.2	17.8	113.4		
Lg Rank	17	23	18	23	13	13	15	24	20	23	14	14	13	2	13	5	29	20	29	11	4	19		
Year/Year	-1.4%	-1.8%	-0.5%	-0.006	3.4%	-0.4%	+0.014	-4.1%	-0.6%	-0.19	-7.8%	-10.3%	+0.022	2.0%	-3.2%	-1.8%	4.3%	13.3%	-3.9%	1.9%	-12.4%	-2.2%		
Opponent	61	14640	2492	5262	.474	793	2124	.373	1699	3138	.541	916	1210	.757	608	1861	2469	1584	411	323	791	1195	6693	
Opponent/G	240.0	40.9	86.3	.474	13.0	34.8	.373	27.9	51.4	.541	15.0	19.8	.757	10.0	30.5	40.5	26.0	6.7	5.3	13.0	19.6	109.7		
Lg Rank	17	10	2	17	13	13	20	7	4	13	3	3	2	5	3	2	10	6	19	22	10	3		
Year/Year	-1.4%	0.2%	-2.2%	+0.011	-0.2%	-4.6%	+0.016	0.4%	-0.4%	+0.004	-19.0%	-16.3%	-0.025	-1.5%	-4.4%	-3.7%	3.3%	11.4%	17.4%	1.8%	-5.2%	-3.0%		

Team Misc		Share & Export ▾		Glossary																				
		Advanced		Offense Four Factors		Defense Four Factors																		
		W	L	PW	PL	Mov	SOS	SRS	ORtg	DRtg	Pace	FTR	3PAR	eFG%	TOV%	ORB%	FT/FGA	eFG%	TOV%	DRB%	FT/FGA	Arena	Attendance	
Team	36	25	37	24	3.72	-0.04	3.68	118.1	114.3	96.0	.257	.400	.538	11.8	29.6	.201	.549	12.0	76.8	.174	Madison Square Garden (IV)	610,765		
Lg Rank	8	21	9	9	15	8	10	11	30	13	11	19	13	1	13	18	16	8	4		5			

FIGURE 2.9 – Team and Opponent / Team Misc Stats

## 2.3 Préparation des données

### 2.3.1 Extraction des données web ( Web scraping )

#### 2.3.1.1 Listes Des Joueurs Qui Ont Reçu Des Votes Pour Le Prix DPOY

Nous avons entamé le processus en téléchargeant toutes les pages HTML des joueurs ayant reçu des votes en utilisant la bibliothèque requests de Python. Pour automatiser cette tâche, nous avons fait appel à la bibliothèque Selenium, en utilisant des URL spécifiques pour chaque joueur. Une fois les pages HTML récupérées, les tableaux de votes de chaque année ont été extraits et enregistrés dans des fichiers CSV pour une analyse ultérieure.

Unnamed: 0	Rank	Player	Age	Tm	First	Pts Won	Pts Max	Share	G ...	STL	BLK	FG%	3P%	FT%	WS	WS/48	DWS	DBPM	DRtg	
0	0	Jaren Jackson Jr.	23	MEM	56.0	391.0	500	0.782	63	—	1.0	3.0	0.506	0.355	0.788	6.6	0.177	3.8	2.0	105
1	1	Brook Lopez	34	MIL	31.0	305.0	500	0.618	78	—	0.5	2.5	0.531	0.374	0.784	8.0	0.161	3.9	1.3	109
2	2	Evan Mobley	21	CLE	8.0	101.0	500	0.202	79	—	0.8	1.5	0.554	0.216	0.674	8.5	0.151	4.8	1.6	108
3	3	Draymond Green	32	GSW	3.0	34.0	500	0.064	73	—	1.0	0.8	0.527	0.306	0.713	4.7	0.098	3.1	2.6	112
4	4	Bam Adebayo	25	MIA	1.0	18.0	500	0.036	75	—	1.2	0.8	0.540	0.083	0.806	7.4	0.137	3.8	0.8	111
5	5	Giannis Antetokounmpo	28	MIL	0.0	14.0	500	0.028	63	—	0.8	0.8	0.553	0.275	0.645	8.6	0.204	3.7	2.7	108
6	6	OG Anunoby	25	TOR	0.0	8.0	500	0.016	67	—	1.9	0.7	0.476	0.387	0.838	4.7	0.094	2.9	0.7	113
7	7	Jrue Holiday	32	MIL	0.0	8.0	500	0.016	67	—	1.2	0.4	0.479	0.384	0.859	6.7	0.148	2.8	0.1	112
8	8	Nic Claxton	23	BRK	0.0	7.0	500	0.014	76	—	0.9	2.5	0.705	0.000	0.541	9.2	0.195	4.0	2.2	108
9	9	Joel Embiid	28	PHI	0.0	7.0	500	0.014	66	—	1.0	1.7	0.548	0.330	0.857	12.3	0.259	3.9	2.3	109
10	10	Alex Caruso	28	CHI	0.0	2.0	500	0.004	67	—	1.5	0.7	0.455	0.364	0.808	3.6	0.109	2.6	3.3	109
11	11	Jimmy Butler	33	MIA	0.0	1.0	500	0.002	64	—	1.8	0.3	0.539	0.350	0.850	12.3	0.277	2.9	2.0	112

FIGURE 2.10 – Liste des joueurs ayant reçu des votes

#### 2.3.1.2 Statistiques Individuelles de tous les joueurs

Pour automatiser ce processus, nous avons remarqué que les URLs étaient légèrement modifiées à la fin. Les listes de joueurs étaient séparées par ordre alphabétique, ce qui nous a permis d'exploiter cette astuce. Nous avons utilisé ces URLs pour extraire les pages HTML contenant des liens vers chaque page individuelle de chaque joueur. Ensuite, nous avons automatisé l'extraction des statistiques par match, par 100 possessions et avancées de chaque joueur en utilisant

Selenium et le Chrome Web Driver, et les avons enregistrées dans des fichiers CSV qui ont été fusionnés par la suite. Le processus a été long car le site avait un mécanisme de défense contre les automatisations. Nous avons surmonté cela en insérant des délais entre chaque extraction.

	Season	Age	Tm	Lg	Pos	G	GS	MP	FG	FGA	...	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS
0	1991-92	23	WSB	NBA	PF	76	43	29.3	4.0	7.8	...	.807	2.4	3.5	5.9	1.6	0.7	0.6	1.5	3.0	10.4
1	1992-93	24	WSB	NBA	SF	81	8	22.5	3.8	7.0	...	.727	1.9	2.8	4.7	1.8	0.6	0.4	1.9	2.4	9.8
2	1993-94	25	WSB	NBA	SF	3	0	11.7	1.0	2.7	...	.700	0.3	2.0	2.3	0.7	0.7	0.3	0.7	1.3	4.3
3	1994-95	26	WSB	NBA	SF	40	0	8.7	1.0	2.2	...	.667	0.7	1.0	1.7	0.5	0.4	0.2	0.4	1.3	2.6
5	1996-97	28	SEA	NBA	SF	70	21	14.0	1.6	3.6	...	.720	1.1	1.4	2.4	0.7	0.4	0.3	0.9	1.5	4.3

FIGURE 2.11 – Statistique per game.csv

	Season	Age	Tm	Lg	Pos	G	GS	MP	FG	FGA	...	TRB	AST	STL	BLK	TOV	PF	PTS	Unnamed: 29	ORtg	DRtg
0	1991-92	23	WSB	NBA	PF	76	43	2229	6.7	13.0	...	9.9	2.6	1.1	1.0	2.5	5.0	17.5	NaN	113	110
1	1992-93	24	WSB	NBA	SF	81	8	1823	8.2	15.2	...	10.3	3.9	1.3	0.8	4.1	5.1	21.5	NaN	108	111
2	1993-94	25	WSB	NBA	SF	3	0	35	4.3	11.4	...	10.0	2.9	2.9	1.4	2.9	5.7	18.6	NaN	106	107
3	1994-95	26	WSB	NBA	SF	40	0	346	6.0	13.1	...	9.8	2.6	2.3	1.3	2.3	7.6	15.0	NaN	104	110
4	1996-97	28	SEA	NBA	SF	70	21	982	6.1	13.6	...	9.3	2.8	1.7	1.2	3.4	5.8	16.2	NaN	101	105

5 rows × 32 columns

FIGURE 2.12 – Statistique per 100 poss.csv

	Unnamed: 0	Rank	Player	Age	Tm	First	Pts	Pts Won	Pts Max	Share	G	...	STL	BLK	FG%	3P%	FT%	WS	WS/48	DWS	DBPM	DRtg
0	0	1	Jaren Jackson Jr.	23	MEM	56.0	391.0	500	0.782	63	...	1.0	3.0	0.506	0.355	0.788	6.6	0.177	3.8	2.0	105	
1	1	2	Brook Lopez	34	MIL	31.0	309.0	500	0.618	78	...	0.5	2.5	0.531	0.374	0.784	8.0	0.161	3.9	1.3	109	
2	2	3	Evan Mobley	21	CLE	8.0	101.0	500	0.202	79	...	0.8	1.5	0.554	0.216	0.674	8.5	0.151	4.8	1.6	108	
3	3	4	Draymond Green	32	GSW	3.0	34.0	500	0.068	73	...	1.0	0.8	0.527	0.305	0.713	4.7	0.098	3.1	2.6	112	
4	4	5	Bam Adebayo	25	MIA	1.0	18.0	500	0.036	75	...	1.2	0.8	0.540	0.083	0.806	7.4	0.137	3.8	0.8	111	
5	5	6	Giannis Antetokounmpo	28	MIL	0.0	14.0	500	0.028	63	...	0.8	0.8	0.553	0.275	0.645	8.6	0.204	3.7	2.7	108	
6	6	7T	OG Anunoby	25	TOR	0.0	8.0	500	0.016	67	...	1.9	0.7	0.476	0.387	0.838	4.7	0.094	2.9	0.7	113	
7	7	7T	Irue Holiday	32	MIL	0.0	8.0	500	0.016	67	...	1.2	0.4	0.479	0.384	0.859	6.7	0.148	2.8	0.1	112	
8	8	9T	Nic Claxton	23	BRK	0.0	7.0	500	0.014	76	...	0.9	2.5	0.705	0.000	0.541	9.2	0.195	4.0	2.2	108	
9	9	9T	Joel Embiid	28	PHI	0.0	7.0	500	0.014	66	...	1.0	1.7	0.548	0.330	0.857	12.3	0.259	3.9	2.3	109	
10	10	11	Alex Caruso	28	CHI	0.0	2.0	500	0.004	67	...	1.5	0.7	0.455	0.364	0.808	3.6	0.109	2.6	3.3	109	
11	11	12	Jimmy Butler	33	MIA	0.0	1.0	500	0.002	64	...	1.8	0.3	0.539	0.350	0.850	12.3	0.277	2.9	2.0	112	

FIGURE 2.13 – Statique advanced.csv

### 2.3.1.3 Statistiques d'une Équipe

Tout d'abord, nous extrayons manuellement toutes les abréviations des équipes, par exemple : 'OKC', 'SEA', 'MEM', etc., et les mettons dans une liste.

Ensuite, nous formons une nouvelle liste contenant les liens de toutes les équipes en utilisant les abréviations précédemment extraites.

Nous lançons ensuite une fenêtre automatisée pour parcourir chaque lien d'équipe. Pour chaque équipe, nous extrayons les tableaux "team\_misc" et "team\_and\_opponent", et les sauve-

gardons dans des fichiers .csv

Unnamed: 0	Year	G_team	MP_team	FG_team	FGA_team	FG%_team	3P_team	3PA_team	3P%_team	...	eFG%_misc	TOV%_misc	ORB%_misc	FT/FGA_misc	eFG%
0	0	2002	82.0	19780	2851	6535	0.436	336	1096	0.307	...	0.462	15.4	27.8	0.204
1	0	2003	82.0	19930	3049	6743	0.452	467	1279	0.365	...	0.487	14.2	27.0	0.212
2	0	2004	82.0	19880	2963	6657	0.445	447	1314	0.340	...	0.479	13.9	29.5	0.234
3	0	2005	82.0	19705	2802	6271	0.447	531	1486	0.357	...	0.489	14.3	27.1	0.243
4	0	2006	82.0	19855	2746	6125	0.448	590	1578	0.374	...	0.496	13.9	25.8	0.241
5	0	2007	82.0	19980	2998	6448	0.465	500	1362	0.367	...	0.504	15.2	25.9	0.285
6	0	2008	82.0	19805	3060	6737	0.454	620	1779	0.349	...	0.500	14.0	23.8	0.225
7	0	2009	82.0	19805	2865	6311	0.454	398	1106	0.360	...	0.486	14.8	25.8	0.249
8	0	2010	82.0	19905	3223	6875	0.469	344	1020	0.337	...	0.494	13.7	31.3	0.235
9	0	2011	82.0	19880	3200	6801	0.471	309	926	0.334	...	0.493	13.0	28.9	0.218

FIGURE 2.14 – Team and Opponent Stats.csv

### 2.3.1.4 Sources additionnelles

Les données ont été extraites de manière similaire et le processus a été automatisé. Cette approche nous a permis de collecter et d'organiser un large ensemble de données pour notre projet, tout en minimisant et en optimisant le temps d'extraction.

### 2.3.2 Fusion des données

Pour préparer nos données en vue de l'entraînement de notre modèle de prédiction, nous avons fusionné les données extraites de chaque joueur avec les données de son équipe et de ses adversaires.

- Nous avons commencé par extraire les noms des joueurs à l'aide de leurs identifiants et des pages de statistiques HTML déjà téléchargées, ce qui a permis de relier chaque vecteur de statistiques à son joueur respectif.
- Ensuite, nous avons fusionné les statistiques extraites, y compris les statistiques "per game", "per 100 possessions" et "advanced", en centralisant toutes les informations pertinentes dans un seul dataframe.
- Nous avons identifié l'équipe de chaque joueur et fusionné ces données avec les statistiques individuelles en respectant l'année des statistiques, ajoutant ainsi une dimension contextuelle aux performances des joueurs.
- Nous avons également inclus des données narratives telles que le nombre de titres gagnés, les votes pour des prix, et la présence éventuelle de camarades d'équipe ayant reçu des prix, enrichissant ainsi le dataset.

- En outre, nous avons fusionné différentes données additionnelles extraites au préalable, telles que les informations sur les matchs, les classements et d'autres statistiques spécifiques, pour obtenir un ensemble de données complet et exhaustif.
- Les données manquantes ont été remplacées par des zéros pour assurer la cohérence du dataset.
- Nous avons regroupé toutes les statistiques dans un grand ensemble couvrant 30 ans, en ajoutant une colonne "année" pour différencier les statistiques par année, facilitant ainsi les comparaisons temporelles.

Lors de cette fusion, nous avons pris des mesures pour éliminer

- Redondance, notamment pour des indicateurs tels que "Age", "Tm", "Lg", "Pos", "G", "GS", "MP" etc...présents dans les ensembles "per game", "per 100 poss", et "advanced"
- L'incohérence entre les champs était compromise, car plusieurs champs des ensembles de données "per game", "advanced", "per 100 poss" partageaient les mêmes noms. Pour remédier à cette situation, des suffixes ont été ajoutés aux noms des champs afin de résoudre ce problème.
- Nous avons remplacé les données manquantes par des zéros pour garantir la cohérence du dataset et faciliter les analyses ultérieures sans erreurs causées par des valeurs absentes.
- Enfin, nous avons séparé les données en caractéristiques (features) et en cible (target) pour l'entraînement du modèle.

Pour chaque joueur, les champs obtenus incluent :

- La saison (année)
- Le nom du joueur et son rang dans le vote pour le DPOY ("-1" pour ceux n'ayant pas reçu de votes)
- Les statistiques "Per Game" pour la même saison
- Les statistiques "Per 100 Poss"
- Les statistiques "Advanced"
- Le nom de son équipe et leurs statistiques "Team and Opponent Stats"

- Les statistiques "Team Misc".

En plus de ces champs, de nouveaux champs ont été créés pour ajouter des statistiques "narratives" :

- A-t-il remporté le DPOY la saison précédente ?
- Combien de fois a-t-il déjà remporté le DPOY ?
- Le meilleur tour que son équipe a atteint pendant les séries éliminatoires d'une saison où il a remporté le DPOY.
- Le nombre d'autres joueurs dans son équipe ayant reçu des votes pour le DPOY pendant une saison précédente (ou la même saison).

Unnamed: 0	Season	Year	Player_name	Rank	Dpoy_titles	Dpoy_votes	First_votes	won_dpoy_last_season	teamates_with_dpoy	...	eFG%_misc	TOV%_misc	ORE
0	0	2022-23	2023	Jaren Jackson Jr.	1.0	1.0	2.0	56.0	False	0.0	...	0.540	11.7
1	1	2022-23	2023	Brook Lopez	2.0	0.0	2.0	31.0	False	2.0	...	0.555	12.7
2	2	2022-23	2023	Evan Mobley	3.0	0.0	1.0	8.0	False	0.0	...	0.556	12.3
3	3	2022-23	2023	Draymond Green	4	1.0	8.0	3.0	False	0.0	...	0.571	14.1
4	4	2022-23	2023	Bam Adebayo	5.0	0.0	4.0	1.0	False	1.0	...	0.530	12.4
...	...	...	...	...	...	...	...	...	...	...	...	...	...
16279	16279	1992-93	1993	Orlando Woolridge	-1.0	0.0	0.0	0.0	False	1.0	...	0.473	12.5
16280	16280	1992-93	1993	Orlando Woolridge	-1.0	0.0	0.0	0.0	False	0.0	...	0.495	14.8
16281	16281	1992-93	1993	James Worthy	-1.0	0.0	0.0	0.0	False	0.0	...	0.486	13.7

FIGURE 2.15 – Données fusionnées

## 2.4 Conclusion

Dans ce chapitre, après avoir exploré et préparé notre ensemble de données, nous entamons la phase cruciale de sélection des caractéristiques

# Chapitre 3

## Selection des features

### 3.1 Introduction

Dans cette section, nous procéderons à la sélection des caractéristiques essentielles pour le modèle. Cette étape revêt une importance particulière dans le processus de modélisation, visant à identifier les variables ayant une influence significative sur les prédictions du modèle.

#### 3.1.1 Connaissance du domaine

Dans notre démarche de sélection d'attributs, nous avons exclu certaines caractéristiques offensives pour les joueurs. Cela signifie qu'on a décidé de ne pas inclure certaines mesures de performance offensives spécifiques qui pourraient ne pas être aussi pertinentes pour notre modèle de prédiction des performances globales des joueurs. Par contre, nous avons conservé les caractéristiques offensives des adversaires tout en supprimant les caractéristiques défensives de ces derniers.

#### 3.1.2 Pour tout l'ensemble de données

Dans un premier temps, nous avons appliqué les méthodes de sélection pour toutes les années afin d'obtenir des résultats généraux.

##### 3.1.2.1 L'information mutuelle (Mutual Information)

La mutualité d'information [9] est une mesure qui évalue le degré de dépendance ou de relation entre deux variables aléatoires. Fréquemment utilisée en théorie de l'information et en statistiques, elle permet de quantifier dans quelle mesure la connaissance de la valeur d'une

variable réduit l'incertitude associée à l'autre variable. Cette mesure repose sur le concept d'entropie, qui représente la quantité d'incertitude ou de désordre inhérente à une variable aléatoire.

- **Résultats :** Nous avons initié notre processus de sélection des attributs en évaluant les scores d'information mutuelle de toutes les caractéristiques en fonction de leur relation avec notre target qui est le nombre de points reçus par les joueurs. Sur la figure 3.1, nous constatons que les statistiques liées à la défense, comme les blocs (BLK), les interceptions (STL) et les possessions défensives gagnées (DWS), étaient parmi les plus importantes dans notre classement. De plus, d'autres données supplémentaires étaient également bien classées, ce qui suggère que d'autres sources de statistiques pourraient être plus précises. En outre, des mesures telles que le nombre de matchs commencés (GS) et les minutes jouées (MP) se sont également retrouvées en haut du classement, ce qui est logique puisqu'un joueur doit souvent participer à de nombreux matchs au cours de la saison

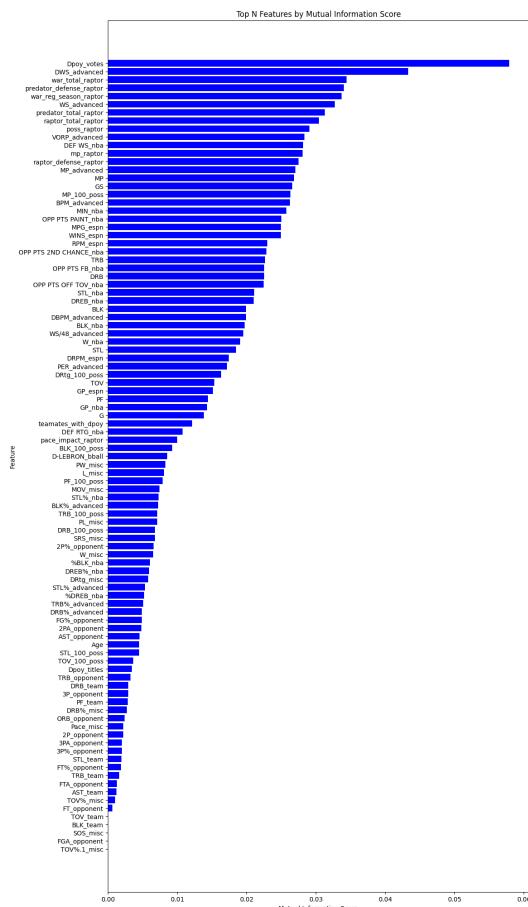


FIGURE 3.1 – Mutual Information

### 3.1.3 Pearson correlation coefficient (PCC)

PCC[10] est un coefficient de corrélation qui mesure la corrélation linéaire entre deux ensembles de données. Cette méthode fonctionne en calculant le coefficient de corrélation de Pearson entre chaque attribut et la variable cible. Les attributs avec une forte corrélation sont plus linéairement dépendants et ont donc presque le même effet sur la variable dépendante. Par conséquent, lorsqu'il y a une forte corrélation entre deux attributs, on peut éliminer l'un des deux. Cela peut aider à réduire la dimensionnalité, améliorer l'interprétabilité du modèle et potentiellement améliorer les performances du modèle en réduisant le bruit et le surapprentissage.

- **Résultats** nous avons observé que ces résultats étaient similaires à ceux obtenus précédemment. Les attributs liés à la défense tels que BLK, STL et DWS figuraient parmi les premiers dans le classement, tout comme les ressources additionnelles. De plus, le nombre de matchs commencés (GS) et les minutes jouées (MP) se sont également retrouvés en haut du classement.

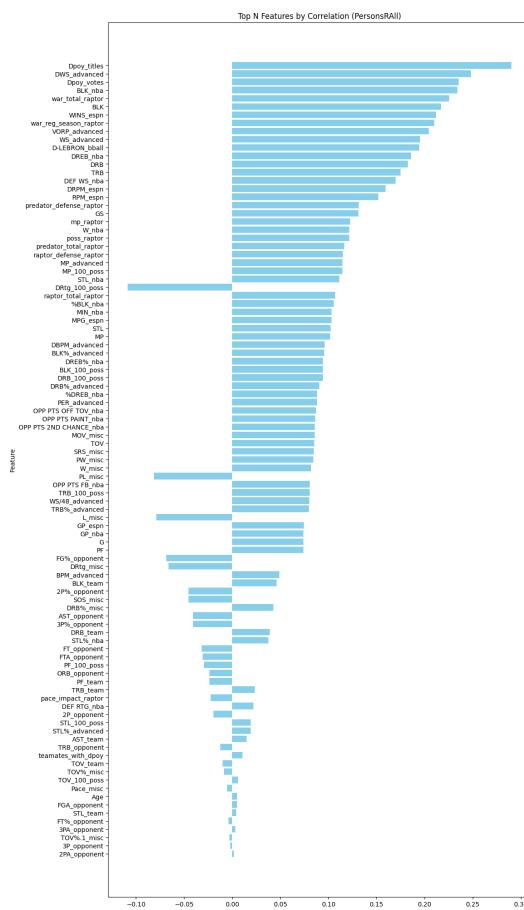


FIGURE 3.2 – Pearson correlation

### 3.1.4 SelectFromModel Pour Random Forest Regressor/XGBoost

La méthode SelectFromModel avec un modèle XGBoost est utilisée pour sélectionner des caractéristiques en fonction de leur importance dans le modèle XGBoost. Tout d'abord, un modèle XGBoost est ajusté sur les données. Ensuite, l'importance de chaque caractéristique est calculée en fonction de sa contribution à l'amélioration de la précision du modèle. Les caractéristiques dont l'importance dépasse un seuil prédéfini sont sélectionnées. Ce seuil peut être déterminé manuellement ou automatiquement en utilisant des méthodes telles que la médiane ou la moyenne des importances des caractéristiques.

La méthode SelectFromModel avec un régresseur Random Forest est utilisée pour sélectionner des caractéristiques en fonction de leur importance dans un modèle Random Forest. Tout d'abord, un modèle Random Forest est ajusté sur les données. Ensuite, l'importance de chaque caractéristique est calculée en mesurant sa contribution à la réduction de l'impureté dans les arbres de décision du modèle. Les caractéristiques dont l'importance dépasse un seuil prédéfini sont sélectionnées. Ce seuil peut être déterminé manuellement ou automatiquement en utilisant des méthodes telles que la médiane ou la moyenne des importances des caractéristiques.

- **Résultats :** Les résultats des méthodes Random Forest Regressor et XGBoost étaient très similaires. Ces méthodes permettent de capturer les relations entre les différentes variables, expliquant ainsi les variations observées d'une personne à une autre et selon les méthodes d'imputation utilisées. Parmi les caractéristiques les plus importantes, on retrouve les statistiques de STL (interceptions), BLK (contres) et DWS (part des victoires défensives), soulignant leur importance dans le modèle. Les caractéristiques associées à des ressources additionnelles se sont également classées en tête, confirmant leur pertinence dans la prédiction des résultats. De manière surprenante, certaines caractéristiques liées aux performances offensives des équipes adverses figuraient également parmi les premières, suggérant que la faiblesse offensive de l'adversaire peut favoriser les performances défensives d'un joueur. En outre, l'âge des joueurs a également été identifié comme un facteur déterminant, soulignant l'importance de l'expérience dans l'obtention de bons résultats.

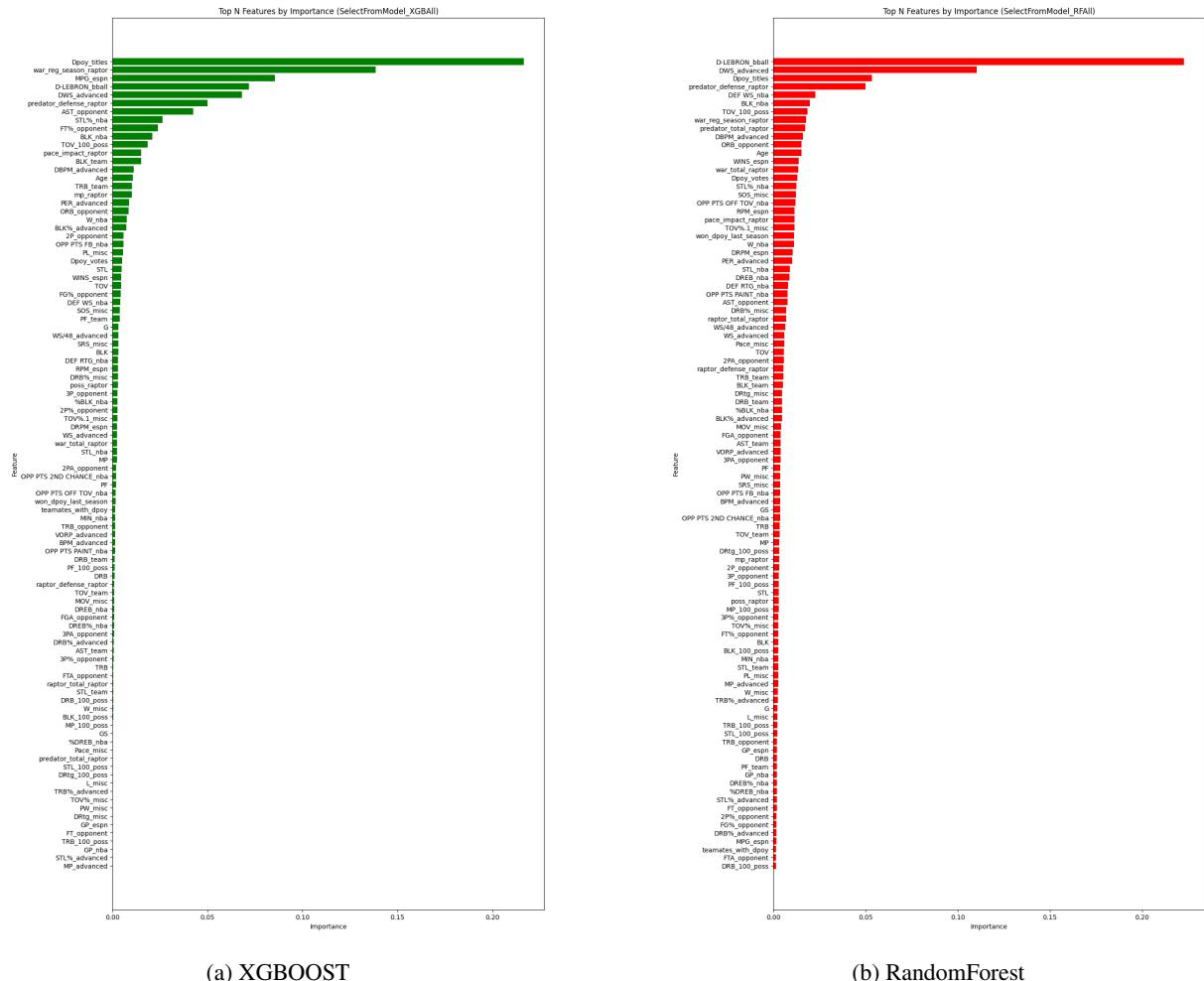


FIGURE 3.3 – SelectFromModel

### 3.1.5 Pour chaque cinq (5) années de l'ensemble de données

Nous avons réalisé une analyse en segmentant nos données sur une période de 30 ans en tranches de cinq années, pour chaque méthode de sélection. Cette approche présente plusieurs avantages, notamment l'identification des tendances et la détection de changements dans les facteurs importants au fil du temps. Cependant, interpréter manuellement les nombreux résultats obtenus était difficile. Pour surmonter cet obstacle, nous avons utilisé la méthode de Kendall Tau afin d'évaluer si les caractéristiques considérées comme importantes varient d'une période à une autre.

#### 3.1.5.1 Tau de Kendall

Le coefficient de corrélation de Kendall[11], souvent appelé tau de Kendall, est une mesure de la corrélation de rang entre deux variables. Il évalue dans quelle mesure les rangs de deux

variables ordinaires sont associées.

Les observations que nous avons fait après l'application du Tau de Kendall sont : la stabilité des caractéristiques sélectionnées par les différentes méthodes s'est maintenue à travers les segments de cinq ans, bien que des changements mineurs aient été constatés, ce qui suggère des évolutions dans les critères de performance au fil du temps.

De plus, les différentes méthodes de sélection ont abouti à des résultats similaires, montrant ainsi qu'elles identifient des facteurs importants de manière cohérente pour prédire les classements des joueurs.

Cependant, des variations notables dans les caractéristiques importantes entre les périodes ont été observées. Ces changements peuvent être le reflet des évolutions dans le jeu de la NBA, des stratégies des équipes ou des critères d'évaluation des performances des joueurs.

	Period	1993_1997	1998_2002	2003_2007	2008_2012	2013_2017	2018_2023	All_years_1993_2023
1	MI vs MI Kendall's tau							
2	1993_1997	1.0	0.3224236604054341	0.34373252854386055	0.3235629225074373	0.32907694640838187	0.3607180110986937	0.34740280523015654
3	1998_2002	0.3224236604054341		1.0	0.7106483706379184	0.6652366476060485	0.6665410331895898	0.5304529685356811
4	2003_2007	0.34373252854386055	0.7106483706379185		1.0	0.669462441569379	0.6479313219300455	0.5787224471434999
5	2008_2012	0.3235629225074372	0.6652366476060484	0.669462441569379	0.9999999999999998	0.6779846659364731	0.6292858162629938	0.7378420136469379
6	2013_2017	0.32907694640838187	0.6665410331895898	0.6479313219300455	0.9999999999999998	0.6779846659364731	0.6821147489863069	0.682114748986307
7	2018_2023	0.3607180110986937	0.5304529685356812	0.5787224471434999	0.6292858162629938	0.6821147489863069	1.0	0.6310369866651028
8	All_years_1993_2023	0.3474028052301565	0.7572465182911484	0.7511716431222666	0.7378420136469379	0.745007636498718	0.6310369866651027	0.9999999999999998

FIGURE 3.4 – MI vs MI Kendall Tau

	Period	1993_1997	1998_2002	2003_2007	2008_2012	2013_2017	2018_2023	All_years_1993_2023
1	MI vs MI Kendall's tau							
2	1993_1997	1.0	0.3224236604054341	0.34373252854386055	0.3235629225074373	0.32907694640838187	0.3607180110986937	0.34740280523015654
3	1998_2002	0.3224236604054341		1.0	0.7106483706379184	0.6652366476060485	0.6665410331895898	0.5304529685356811
4	2003_2007	0.34373252854386055	0.7106483706379185		1.0	0.669462441569379	0.6479313219300455	0.5787224471434999
5	2008_2012	0.3235629225074372	0.6652366476060484	0.669462441569379	0.9999999999999998	0.6779846659364731	0.6292858162629938	0.7378420136469379
6	2013_2017	0.32907694640838187	0.6665410331895898	0.6479313219300455	0.9999999999999998	0.6779846659364731	0.6821147489863069	0.682114748986307
7	2018_2023	0.3607180110986937	0.5304529685356812	0.5787224471434999	0.6292858162629938	0.6821147489863069	1.0	0.6310369866651028
8	All_years_1993_2023	0.3474028052301565	0.7572465182911484	0.7511716431222666	0.7378420136469379	0.745007636498718	0.6310369866651027	0.9999999999999998

FIGURE 3.5 – MI vs Pearson Kendall Tau

### 3.1.5.2 Corrélation de Spearman

La corrélation de Spearman [12] est une mesure statistique non paramétrique qui évalue la force et la direction de l'association entre deux variables classées. Nous avons fait les mêmes observations que le Tau de Kendall , la stabilité des caractéristiques sélectionnées par les différentes méthodes s'est maintenue à travers les segments de cinq ans.

19	MI vs MI Spearman's								
20	1993_1997	1.0	0.439541916731821	0.46857878371082595	0.45456983013212304	0.4590716333746816	0.498020336298647	0.4781110677207274	
21	1998_2002	0.43954191673182097		1.0	0.8875306669230467	0.853231593837415	0.8443340917473808	0.7246448629637707	0.919912145234414
22	2003_2007	0.468578783710826	0.8875306669230468		1.0	0.8403049558146277	0.8335808218981476	0.7646191450357342	0.9112436173318949
23	2008_2012	0.454569830132123	0.853231593837415	0.8403049558146277		1.0	0.864769422517657	0.824944016031015	0.8931066446171023
24	2013_2017	0.45907163337468154	0.8443340917473809	0.8335808218981475	0.864769422517657		1.0	0.864037971783622	0.9093672177782929
25	2018_2023	0.498020336298647	0.7246448629637708	0.7646191450357342	0.824944016031015	0.864037971783622		1.0	0.8332469159615327
26	All_years_1993_2023	0.4781110677207274	0.919912145234415	0.9112436173318949	0.8931066446171024	0.9093672177782928	0.8332469159615327		0.9999999999999999

FIGURE 3.6 – MI vs MI Spearman

28	MI vs Pearson's Spearman's								
29	1993_1997	0.46588312967067813	0.22541591215456117	0.2552736756146741	0.2882565668986446	0.30635566922848567	0.1941267238413697	0.22369606631391986	
30	1998_2002	0.5898325586473842	0.6751932426644355	0.7088721838710994	0.6864868883155998	0.678700846513875	0.5979601191138928	0.6809761594052575	
31	2003_2007	0.6170365784550056	0.6586588175100564	0.7058406812285418	0.6724223612969102	0.6913355416441997	0.6061006726853804	0.679744453382788	
32	2008_2012	0.5924883805900631	0.546581586629576	0.6439857770112607	0.6663908509494628	0.6774004603759375	0.5501512365920737	0.649165289789306	
33	2013_2017	0.6134814415788265	0.654100487903117	0.6877052392678003	0.6967133697346992	0.719292876314582	0.6370955445269837	0.711772566360505	
34	2018_2023	0.5556737976678615	0.49195120827407957	0.5846373713619323	0.6046666544993001	0.66111542567671	0.5513634716010339	0.619535506591189	
35	All_years_1993_2023	0.6770377618106886	0.6803885272701335	0.7356829402770034	0.7470756497941617	0.7554656948494548	0.6642929952125468	0.7418285839993938	

FIGURE 3.7 – MI vs Pearson Spearman

37	Pearson's vs Pearson's Spearman's								
38	1993_1997	1.0	0.835551215119754	0.8527967245127434	0.841014972478732	0.8231977195012901	0.7586959079346787	0.8992823668016541	
39	1998_2002	0.835551215119754		1.0	0.9341847589110244	0.9059304029002606	0.8930031535565551	0.799156448548755	0.9401074692073433
40	2003_2007	0.8527867245127435	0.9341847589110245		1.0	0.9082108310052073	0.9025809204542554	0.8111521449508275	0.9497007160592766
41	2008_2012	0.841014972478732	0.9059304029002605	0.9082108310052073		1.0	0.9143877959939861	0.8532244718476811	0.9327346571966884
42	2013_2017	0.8231977195012902	0.8930031535565552	0.9025809204542553	0.9143877959939861		1.0	0.8926728881385634	0.9348233556109731
43	2018_2023	0.7586959079346787	0.799156448548755	0.8111521449508275	0.8532244718476811	0.8926728881385633		1.0	0.869682005649468
44	All_years_1993_2023	0.8992823668016541	0.9401074692073433	0.9497007160592767	0.9327346571966885	0.934823355610973	0.8696820056494681		0.9999999999999999

FIGURE 3.8 – Pearson vs Pearson Spearman

## 3.2 Conclusion

Ce chapitre a été dédié à la sélection des caractéristiques, une étape cruciale dans le processus de création de notre modèle. Nous avons appliqués différentes techniques pour identifier les caractéristiques les plus importantes, cherchant ainsi à optimiser la performance de notre modèle.

## Chapitre 4

# Modèles de prédiction

### 4.1 Introduction

Ce chapitre se concentrera sur la phase de modélisation, abordant le choix et la configuration des différents algorithmes appliqués à la base d'apprentissage pour créer les modèles de prédiction. Le défi que nous examinons concerne un problème de régression. Nous allons explorer des algorithmes tels que RandomForest, XGBoost.

### 4.2 Découpage des données

#### 4.2.1 Approche basée sur des tranches de 5 à 9 ans

Cette méthode consiste à entraîner un modèle de prédiction sur un segment de données historiques de 5 à 9 ans, puis à utiliser ce modèle pour prédire l'année suivante. Par exemple, pour prédire l'année 1998, nous entraînerions le modèle sur les données de 1993 à 1997. Ensuite, pour prédire 1999, nous utiliserions les données de 1993 à 1998, et ainsi de suite.

Cela permet au modèle d'apprendre des tendances et des motifs sur une période de temps spécifique avant de faire une prédiction.

#### 4.2.2 Approche cumulative

Dans cette approche, nous utilisons toutes les données disponibles jusqu'à l'année précédent celle que nous souhaitons prédire comme ensemble d'entraînement. L'année que nous voulons prédire est utilisée comme ensemble de test. Par exemple, pour prédire l'année 2010, nous entraînerions le modèle sur toutes les données collectées de 1993 à 2009. Cette méthode peut

être utile pour capturer toutes les informations historiques et les tendances à long terme qui pourraient influencer la prédiction.

### 4.2.3 XGBoost Regressor

#### 4.2.3.1 Définition

XGBoost [13], qui signifie Extreme Gradient Boosting, est une bibliothèque d'apprentissage automatique évolutive et distribuée à arbre de décision boosté par gradient (GBDT). Il fournit une amélioration des arbres parallèles. Il est couramment adopté en raison de ses performances remarquables et de son efficacité, que ce soit pour résoudre des problèmes de régression et de classification.

#### 4.2.3.2 Fonctionnement

Le processus de Gradient Boosting commence par une étape d'initialisation, où une prédiction initiale est faite pour chaque échantillon du jeu de données d'entraînement. Ensuite, les résidus sont calculés en soustrayant les valeurs prédictes des valeurs observées. Ces résidus sont utilisés pour construire un arbre de décision, où les données sont divisées en maximisant un critère de gain à chaque fractionnement. Les valeurs de sortie de chaque feuille de l'arbre sont déterminées en utilisant les résidus. Ensuite, les résidus sont mis à jour, où la sortie de l'arbre devient la nouvelle valeur résiduelle. Ce processus est répété avec de nouveaux arbres, chaque arbre apprenant des résidus des arbres précédents, jusqu'à ce que les résidus cessent de diminuer ou qu'un nombre spécifié d'arbres soit atteint.

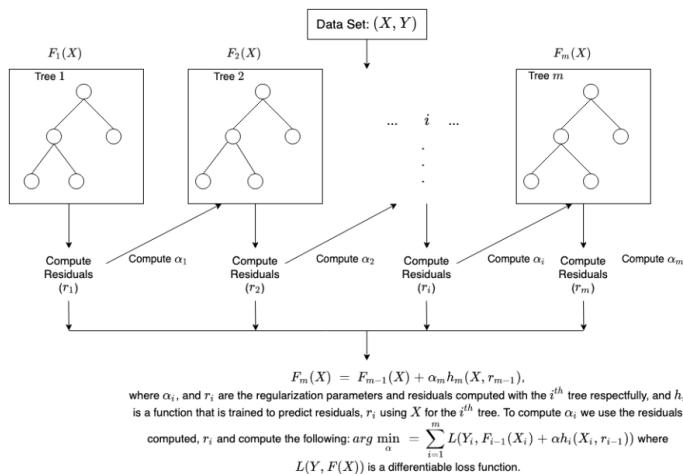


FIGURE 4.1 – XGBoost Regressor

## 4.2.4 Random Forest Regressor

### 4.2.4.1 Définition

Un modèle de régression Random Forest combine plusieurs arbres de décision pour créer un seul modèle. Chaque arbre dans la forêt est construit à partir d'un sous-ensemble différent des données et fait sa propre prédition indépendante. La prédition finale pour une entrée est basée sur la moyenne ou la moyenne pondérée de toutes les prédictions des arbres individuels.

### 4.2.4.2 Fonctionnement

Le processus de construction d'une forêt d'arbres de décision commence par la construction de plusieurs arbres de décision, chacun utilisant des sous-ensembles aléatoires des données d'entraînement. Chaque arbre est formé indépendamment des autres. Ensuite, à chaque nœud de chaque arbre, une division est effectuée en fonction d'une caractéristique (ou variable) et d'une valeur seuil qui maximise la réduction de l'erreur de prédition. Une fois que tous les arbres sont construits, chaque arbre est utilisé pour prédire une valeur pour une nouvelle entrée en passant à travers chaque arbre. Ces prédictions individuelles sont ensuite combinées pour former une prédition finale. La prédition finale est souvent calculée en prenant la moyenne ou la médiane des prédictions de tous les arbres dans la forêt, bien que d'autres méthodes de combinaison puissent être utilisées.

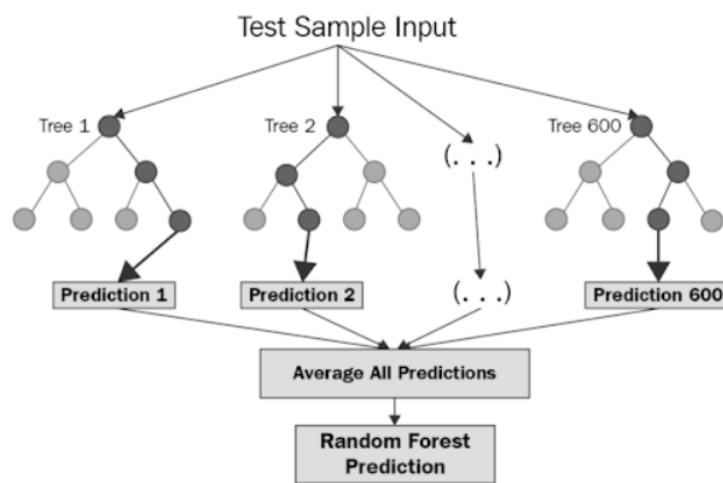


FIGURE 4.2 – Random forest Regressor

#### 4.2.5 Entraînement

- Sélection des Caractéristiques : Nous avons choisi de sélectionner les caractéristiques en utilisant différentes méthodes de sélection d'attributs, en commençant par le top 10, top 20 jusqu'au top 50 caractéristiques, selon les résultats obtenus par les quatre méthodes de sélection.

Les listes d'attributs sélectionnés par ces quatre méthodes ont été divisées en 5 segments.

- Hyperparamétrage : Pour l'entraînement des modèles, les paramètres ont été définis avec un nombre d'estimateurs compris entre 200 et 300. Une graine aléatoire (random seed) de 42 a été utilisée pour garantir la reproductibilité des résultats. Le paramètre `n_jobs = -1` a été spécifié pour utiliser tous les processeurs disponibles de la machine, maximisant ainsi la vitesse d'entraînement.

Ainsi, les entraînements ont été lancés avec les deux modèles en utilisant les top 10, top 20, jusqu'aux top 50 caractéristiques.

#### 4.2.6 Test

Pour évaluer les performances de nos modèles, nous avons classé les joueurs uniquement en fonction des points prédits. Par exemple, le premier joueur du classement est celui ayant le plus grand nombre de points prédits, ce qui en fait le gagnant selon le modèle. La position et le nom des joueurs sont affichés pour nous aider à évaluer la performance des modèles.

### 4.3 Conclusion

Dans ce chapitre, nous avons utilisé deux façons de diviser nos données pour entraîner et tester nos modèles : une méthode coupe les données en périodes de 5 à 9 ans, et l'autre utilise toutes les données jusqu'à l'année précédente pour entraîner. Ensuite, nous avons étudié deux algorithmes pour créer nos modèles : XGBoost et Random Forest Regressor.

# Chapitre 5

## Évaluation des modèles

### 5.1 Introduction

#### 5.1.1 Coefficient de détermination R<sup>2</sup>

Le coefficient de détermination mesure la proportion de la variance dans la variable dépendante qui est prévisible à partir des variables indépendantes. Une valeur de R<sup>2</sup> de 1 indique que le modèle prédit parfaitement les données, tandis qu'une valeur de 0 indique qu'il ne prédit rien de mieux que la moyenne des données

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

où  $n$  est le nombre de mesures,  $y_i$  la valeur de la mesure n°  $i$ ,  $\hat{y}_i$  la valeur prédite correspondante et  $\bar{y}$  la moyenne des mesures.

#### 5.1.2 Erreur quadratique moyenne MSE

L'Erreur quadratique moyenne (MSE) mesure la moyenne des carrés des erreurs, c'est-à-dire la moyenne des carrés des différences entre les valeurs observées et les valeurs prédites par le modèle. Plus la valeur du MSE est faible, meilleure est la performance du modèle.

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

Avec  $n$  : nombre de mesures,  $y_i$  : valeurs prédites par le modèle,  $y_i$  : valeurs observées.

#### 5.1.3 Limites de ces approches

Les méthodes d'évaluation classiques telles que le coefficient de détermination ( $R^2$ ) et l'erreur quadratique moyenne (MSE) sont souvent inefficaces dans ce contexte. La difficulté réside dans

la prédiction précise du nombre de points, étant donné que leur attribution varie d'une année à l'autre.

#### 5.1.4 Autres méthodes d'évaluation

- **Concaténation des Classements :** Nous avons concaténé les classements prédis par nos modèles pour former des DataFrames contenant les classements prédis par année. Par exemple, pour un modèle entraîné sur les top 10 caractéristiques, cela donne un tableau de dix (10) lignes par trente (30) colonnes.
- **Utilisation des Tableaux :** Une fois les tableaux sont formés, nous les utilisons pour évaluer la performance de chaque modèle.
- **Script d'Évaluation :** Nous avons écrit un script qui, en utilisant ces tableaux, compte le nombre de classements corrects afin de déterminer les modèles les plus performants.

```
Mean Squared Error (RandomForest Regressor): 305.10454914772725
R-squared (RandomForest Regressor): 0.667238004641737
```

	Player_name	Rank	Pos	Points_won	Predicted_points
4426	Kawhi Leonard	1	SF	547.0	239.870
4597	Draymond Green	2	PF	421.0	200.270
4570	Hassan Whiteside	3	C	83.0	134.795
4474	DeAndre Jordan	4.0	C	50.0	110.665
4820	Andre Drummond	10	C	3.0	61.290
4782	Tim Duncan	-1	C	0.0	43.450
4373	Rudy Gobert	7.0	C	13.0	42.140
4524	LeBron James	11T	SF	2.0	32.015
4773	Andrew Bogut	-1.0	C	0.0	27.735
4354	Kevin Garnett	-1	PF	0.0	26.940

FIGURE 5.1 – Prédiction des classements des joueurs de l'année 2017

#### 5.1.5 Comparaison entre modèles avec périodes de 5 à 9 ans et modèles cumulatif

Nous avons observé, en utilisant le script, que les modèles entraînés de manière cumulative surpassent en performances ceux entraînés avec la méthode de segmentation de 5 à 9 ans. Ceci suggère qu'il n'existe pas de tendance périodique évidente. Par conséquent, dans les prochaines comparaisons, nous ne considérerons que l'entraînement cumulatif.

Nom fichier : RDF+RDF\_All\_rank\_data\_40  
 Rank\_1994: 2 correct ranks  
 Rank\_1995: 2 correct ranks  
 Rank\_1996: 3 correct ranks  
 Rank\_1997: 0 correct ranks  
 Rank\_1998: 1 correct ranks  
 Rank\_1999: 1 correct ranks  
 Rank\_2000: 2 correct ranks  
 Rank\_2001: 3 correct ranks  
 Rank\_2002: 1 correct ranks  
 Rank\_2003: 3 correct ranks  
 Rank\_2004: 2 correct ranks  
 Rank\_2005: 1 correct ranks  
 Rank\_2006: 4 correct ranks  
 Rank\_2007: 0 correct ranks  
 Rank\_2008: 0 correct ranks  
 Rank\_2009: 0 correct ranks  
 Rank\_2010: 2 correct ranks  
 Rank\_2011: 2 correct ranks  
 Rank\_2012: 2 correct ranks  
 Rank\_2013: 0 correct ranks  
 Rank\_2014: 3 correct ranks  
 Rank\_2015: 1 correct ranks  
 Rank\_2016: 5 correct ranks  
 Rank\_2017: 0 correct ranks  
 Rank\_2018: 1 correct ranks  
 Rank\_2019: 2 correct ranks  
 Rank\_2020: 2 correct ranks  
 Rank\_2021: 2 correct ranks  
 Rank\_2022: 0 correct ranks  
 Rank\_2023: 3 correct ranks

FIGURE 5.2 – Nombre de classement correctement prédits

	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
1	1	1	3	2	2	3	1	1	1	1	6	-1	1	2	2	8	1	1	3	7.0	1.0	1	1	2.0	1.0	1.0	1.0	1.0	1.0	
2	2	3	1	1	1	1	5T	1	3T	-1	2	-1	2	6	1	1	14	2	5	1	7.0	4	2	1	-1	2.0	3.0	-1	-1.0	2.0
3	-1	4	5	-1	-1	2	3T	-1	-1	3	-1	-1	5	-1	4	-1	-1	10.0	2.0	2	13T	2	3	-1	8.0	5.0	2	9.0	4.0	5.0
4	-1.0	2	4	5	3	7T	-1	-1.0	-1	4	1	4	4	1	10	7	-1	5	4	6	4.0	3.0	4.0	3	6	3.0	11	5.0	6.0	6.0
5	-1.0	6T	6	4	5T	4	2	-1	-1	13T	3	9	4	-1	3	6	4	17	7	3.0	5	6.0	10	-1.0	15T	6T	-1.0	-1	9	7T
6	-1.0	6T	2	3	5T	-1	-1	3	-1	-1	3	5	6	5	9	2	-1	12T	-1	-1	13T	8	-1	5T	3	4	8T	6.0	-1	3.0
7	-1.0	-1	7	6T	4	7T	3T	7	-1	8	7	10	3	-1	12	11	6	9	9.0	-1	10.0	10T	7.0	-1.0	-1	10.0	3	10	9T	
8	7T	5	8T	-1	9T	6T	-1	8T	2	11	-1.0	1	13T	-1.0	7T	4	12	25	1	21	2.0	-1	11T	-1	-1.0	-1	8T	11T	-1	-1
9	4	6T	8T	-1	-1	-1	7	5T	3T	5.0	-1	11	-1	3	-1.0	10	3	6.0	12T	8.0	11	15T	-1.0	-1.0	15T	-1	-1	-1	-1.0	9T
10	7T	-1.0	-1	6T	5T	-1	8T	11T	-1	11	5	-1	-1	13T	8	6	5	10T	3	10	14	8T	-1.0	-1	-1.0	-1.0	-1.0	-1.0	1	-1.0

FIGURE 5.3 – Prédictions de classement des joueurs de toutes les années

### 5.1.6 Comparaison entre modèles avec Top 10,20,30,40,50

Les modèles entraînés avec les top 30 et 40 sont plus performants que les autres. Cela suggère que le nombre parfait d'attributs à sélectionner pourrait être entre 30 et 40.

### 5.1.7 Comparaison entre les modèles XGBoostRegressor et RandomForestRegressor

En général Randomforest est plus performant que XGboost et cela a est du au fait que moins sensible aux valeurs aberrantes et au surajustement par rapport à XGBoost.

### 5.1.8 Comparaison entre modèle par méthode de sélection d'attribut

- RandomForest avec SelectFromModel RF : Ce modèle est le plus performant, ayant correctement prédit le classement des trois premiers joueurs sur trois saisons et le gagnant du prix sur 16 saisons.
- XGBoost avec XGB SelectFromModel : Ce modèle est moins performant, mais a réussi à prédire correctement le classement des trois premiers joueurs pour une saison et le gagnant du prix sur 14 saisons.
- MI avec RDF a été moins performant que SelectFromModel RDF avec RDF, mais le modèle a réussi à prédire le gagnant pour 12 saisons, sans toutefois prédire le classement des trois premiers.
- Per avec XGB a réussi à prédire les six premiers classements en 1998 et le gagnant pour 16 saisons.

### 5.1.9 Le modèle le plus performant

Le modèle le plus performant a été obtenu en entraînant l'algorithme RandomForestRegressor avec l'utilisation des 40 meilleures attribues sélectionnées par SelectFromModel avec RDF, en employant une approche cumulative.

## 5.2 Conclusion

Dans ce dernier chapitre, nous avons comparé différents modèles de régression et techniques de sélection de caractéristiques pour prédire les performances des joueurs. Les résultats indiquent que les modèles RandomForest surpassent légèrement les modèles XGBoost et que la sélection des 30 à 40 meilleurs attributs est optimale pour obtenir les meilleures prédictions. De plus, les évaluations cumulatives des modèles ont également démontré des performances supérieures

## Conclusion Générale

Dans ce projet, nous avons abordé plusieurs étapes clés pour mener à bien notre étude. Le premier chapitre, nous avons établi le contexte global du projet, défini ses objectifs et introduit des concepts clés tels que le DPOY (Défenseur de l'Année) et le machine learning. Le deuxième chapitre s'est consacré au prétraitement des données, fournissant une description détaillée des diverses sources de données et des étapes de fusion et de préparation des données. Dans la troisième partie, orientée vers la sélection des caractéristiques, nous avons identifié les features importants. L'achèvement de ces étapes nous a préparés à aborder le chapitre 4, dans lequel nous avons exploré plusieurs méthodes en utilisant deux algorithmes, XGBoost et Random Forest. Nous avons terminé par le chapitre 5, dans lequel nous avons évalué les performances des modèles utilisés.

Pour améliorer nos modèles, nous pourrions envisager plusieurs stratégies. La normalisation des données permettrait de mettre à l'échelle les différentes caractéristiques, facilitant ainsi la convergence des algorithmes d'apprentissage. L'undersampling, qui consiste à réduire le nombre d'exemples dans la classe majoritaire, pourrait équilibrer la distribution des classes dans l'ensemble de données. Enfin, l'hyperparamétrage avec GridSearch permettrait de rechercher systématiquement les meilleures combinaisons de paramètres pour un modèle donné. Toutes ces techniques nous permettraient d'obtenir un modèle plus précis et performant.

# Annexe

## 1 Basket-ball

En basket-ball, les positions des joueurs sont généralement catégorisées en fonction de leurs rôles et responsabilités principaux sur le terrain. Les principales positions incluent :

**Meneur de jeu (Point Guard - PG) :** Souvent appelé le "général du terrain," le meneur de jeu est chargé de diriger l'offensive de l'équipe, de distribuer le ballon et de mettre en place des jeux. Compétences : Habituellement rapide, agile, et doté de bonnes compétences de passe et de maniement de balle.

**Arrière (Shooting Guard - SG) :** Souvent l'un des principaux marqueurs de l'équipe, l'arrière est connu pour ses compétences de tir, aussi bien de mi-distance que de derrière la ligne des trois points. Compétences : Polyvalent, capable de contribuer en tant que manieur de balle et marqueur.

**Ailier (Small Forward - SF) :** Les ailiers sont généralement des joueurs polyvalents qui peuvent marquer de différentes positions sur le terrain. Ils possèdent souvent une combinaison de compétences en marquage, en rebondissement et en défense. Compétences : Polyvalent, capable de jouer à l'intérieur ou à l'extérieur.

**Ailier fort (Power Forward - PF) :** Les ailiers forts jouent près du panier et sont souvent impliqués dans le rebond et le marquage près du cercle. Ils sont généralement des joueurs forts et physiques capables de marquer à l'intérieur et avec des tirs à mi-distance. Compétences : Force physique, capacité à marquer près du panier.

**Pivot (Center - C) :** Les pivots sont généralement les joueurs les plus grands de l'équipe et jouent près du panier des deux côtés du terrain. Ils sont essentiels pour le rebond, le contre et le marquage près du panier. Compétences : Grande taille, capacité à jouer en défense et

à marquer près du panier. Les pivots modernes peuvent également avoir des compétences de tir extérieur.

Le jeu de basket-ball implique des passes, des dribbles, des tirs et des mouvements tactiques pour marquer des paniers tout en défendant le panier de l'équipe adverse. Les équipes travaillent ensemble pour créer des opportunités de tir et pour empêcher l'équipe adverse de marquer. La stratégie, la coordination et les compétences individuelles des joueurs sont toutes des composantes clés du jeu.

## 2 Liste des features

- Season :** The season recorded, represented as a range (e.g., 2021-2022).
- Age :** Player's age during that season.
- Tm :** Team abbreviation where the player participated.
- Lg :** League abbreviation (e.g., NBA - National Basketball Association).
- Pos :** Player's primary playing position on the court (e.g., PG for Point Guard, SG for Shooting Guard).
- G :** Total games played.
- GS :** Games started.
- MP :** Average minutes played per game.
- FG :** Total successful field goals made.
- FGA :** Total attempts at field goals.
- FG% :** Field goal percentage (FG/FGA).
- 3P :** Total successful three-point shots made.
- 3PA :** Total attempts at three-point shots.
- 3P% :** Three-point percentage (3P/3PA).
- 2P :** Total successful two-point shots made.
- 2PA :** Total attempts at two-point shots.
- 2P% :** Two-point percentage (2P/2PA).
- eFG% :** Effective Field Goal Percentage, calculated as  $(\text{FG} + 0.5 * \text{3P}) / \text{FGA}$ .
- FT :** Total successful free throws made.
- FTA :** Total attempts at free throws.
- FT% :** Free throw percentage (FT/FTA).
- ORB :** Offensive rebounds.
- DRB :** Defensive rebounds.
- TRB :** Total rebounds (both offensive and defensive).
- AST :** Total assists.
- STL :** Total steals.
- BLK :** Total blocks.
- TOV :** Total turnovers committed.
- PF :** Total personal fouls committed.
- PTS :** Total points scored.

<b>Rank :</b>	Player's rank in the Defensive Player of the Year voting.
<b>Player :</b>	Player who received votes for the Defensive Player of the Year award.
<b>First :</b>	Number of first-place votes received.
<b>Pts Won :</b>	Total voting points received for the Defensive Player of the Year award.
<b>Pts Max :</b>	Maximum possible voting points.
<b>Share :</b>	Share of total possible points received in the voting.
<b>G :</b>	Total games played during the season.
<b>MP :</b>	Average minutes played per game.
<b>PTS :</b>	Average points scored per game.
<b>TRB :</b>	Average total rebounds per game.
<b>FG% :</b>	Field Goal Percentage.
<b>3P% :</b>	Three-Point Percentage.
<b>FT% :</b>	Free Throw Percentage.
<b>WS :</b>	Win Shares, a statistic attempting to measure a player's overall contribution to their team's wins, combining offensive and defensive contributions.
<b>WS/48 :</b>	Win Shares per 48 minutes, a rate statistic expressing Win Shares on a per-minute basis.
<b>DWS :</b>	Defensive Win Shares, a component of Win Shares specifically measuring a player's defensive contributions.
<b>DBPM :</b>	Defensive Box Plus-Minus, a box score-based metric estimating a player's overall defensive impact per 100 possessions.
<b>DRtg :</b>	Defensive Rating, a team-based statistic estimating the number of points allowed by the player's team per 100 possessions while the player is on the court. A lower DRtg is generally better as it indicates better defensive performance.

### 3 Quelques Analyses Et Visualisation Des Données

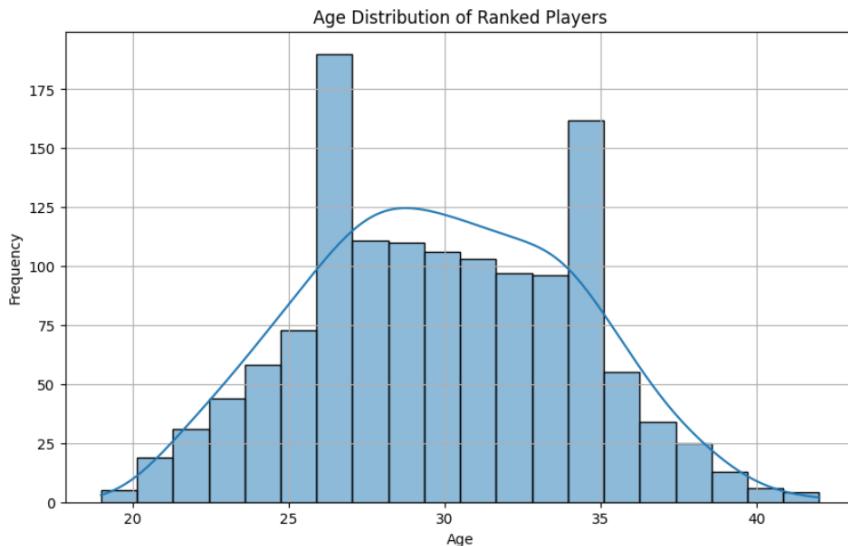


FIGURE 5.4 – La distribution des joueurs ayant eu des votes en fonctions de leurs âge

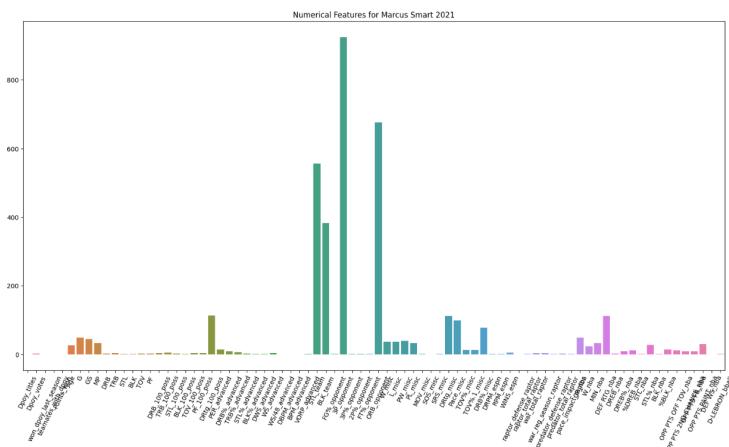


FIGURE 5.5 – La distribution des caractéristiques d'un joueur ayant remporté le prix

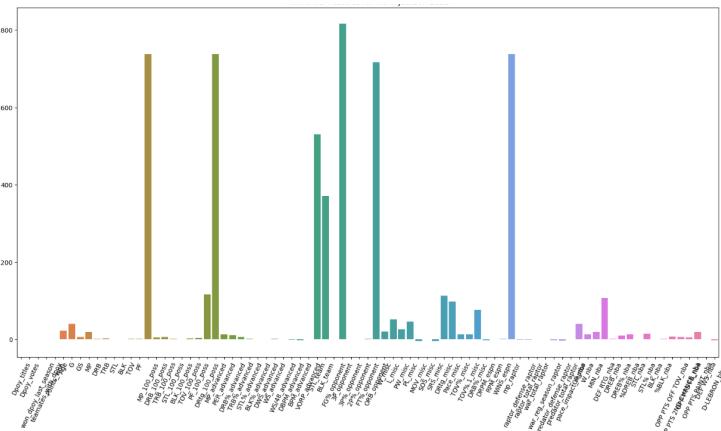


FIGURE 5.6 – La distribution des caractéristiques d'un joueur n'ayant pas reçu de prix

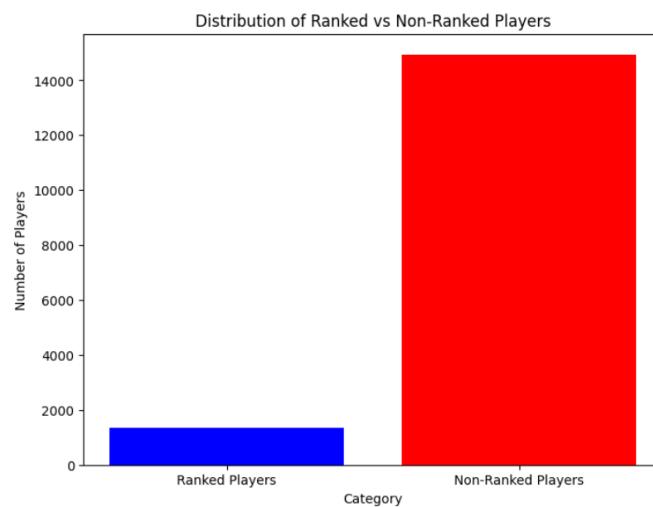


FIGURE 5.7 – La distribution des joueurs ayant reçu des votes vs ceux qui n'ont pas reçu

# Bibliographie

- [1] *NBA Defensive Player of the Year*. URL : [https://fr.wikipedia.org/wiki/NBA\\_Defensive\\_Player\\_of\\_the\\_Year](https://fr.wikipedia.org/wiki/NBA_Defensive_Player_of_the_Year).
- [2] *Machine learning*. URL : <https://datascientest.com/machine-learning-tout-savoir>.
- [3] *Python*. URL : <https://www.python.org/>.
- [4] *Basketball Reference*. URL : <https://www.basketball-reference.com/>.
- [5] : *NBA.COM*. URL : <https://www.nba.com/stats/players/defense?Season=..>.
- [6] *ESPN*. URL : [https://www.espn.com/nba/statistics/rpm/\\_/year/2023/sort/](https://www.espn.com/nba/statistics/rpm/_/year/2023/sort/).
- [7] *Github*. URL : <https://github.com/fivethirtyeight/data/tree/master/nba-raptor>.
- [8] *BBALL-INDEX*. URL : <https://www.bball-index.com/lebron-application/>.
- [9] *Mutual information*. URL : [https://fr.wikipedia.org/wiki/Information\\_mutuelle](https://fr.wikipedia.org/wiki/Information_mutuelle).
- [10] *PCC*. URL : [https://en.wikipedia.org/wiki/Pearson\\_correlation\\_coefficient..](https://en.wikipedia.org/wiki/Pearson_correlation_coefficient..)
- [11] *tau de Kendall*. URL : <https://fastercapital.com/fr/contenu/Kendall-s-tau---examen-de-Kendall-s-Tau-dans-les-statistiques-non-parametriques.html>.
- [12] *Corrélation de Spearman*. URL : [https://fr.wikipedia.org/wiki/Corr%C3%A9lation\\_de\\_Spearman](https://fr.wikipedia.org/wiki/Corr%C3%A9lation_de_Spearman).
- [13] *XGBoost*. URL : <https://pro.arcgis.com/fr/pro-app/latest/tool-reference/geoai/how-xgboost-works.htm#:~:text=XGBoost%20signifie%20extreme%20gradient%20boosting,%C3%A0%20diverses%20m%C3%A9thodes%20d'optimisation..>