Université de Lille

Master Informatique

Parcours Machine learning

De la fouille de données à l'auto-ML

Projet FIFA: Rendu Part I & II

Auteur:
Selim Lakhdar

Professeur:
Laetitia Jourdan

November 9, 2021

Contents

1 Context

Pour ce projet, nous allons exploiter le jeu de données qui a été utilisé pour générer les statistiques des joueurs dans le jeu FIFA 2019.
FIFA 19 est un jeu vidéo de football développé par EA Canada et EA Bucarest, édité par EA Sports, sorti le 28 septembre 2018 sur Nintendo Switch, PC, PlayStation 3, PlayStation 4, Xbox One et Xbox 360.
Il s'agit du vingt-sixième opus de la franchise FIFA développé par EA Sports. (Source Wikipédia).

Chaque ligne représente un joueur présent dans le jeu. Les colonnes correspondent aux informations sur les joueurs (taille, salaire, club, valeur, force, efficacité à un poste donné, ...)
Le fichier data.csv contient donc 89 attributs et 18207 joueurs (!) décrits par Age, Nationality, Overall, Potential, Club, Value, Wage, Preferred Foot, International Reputation, Weak Foot, Skill Moves, Work Rate, Position, Jersey Number, Joined, Loaned From, Contract Valid Until, Height, Weight, LS, ST, RS, LW, LF, CF, RF, RW, LAM, CAM, RAM, LM, LCM, CM, RCM, RM, LWB, LDM, CDM, RDM, RWB, LB, LCB, CB, RCB, RB, Crossing, Finishing, Heading, Accuracy, ShortPassing, Volleys, Dribbling, Curve, FKAccuracy, LongPassing, BallControl, Acceleration, SprintSpeed, Agility, Reactions, Balance, ShotPower, Jumping, Stamina, Strength, LongShots, Aggression, Interceptions, Positioning, Vision, Penalties, Composure, Marking, StandingTackle, SlidingTackle, GKDiving, GKHandling, GKKicking, GKPositioning, GKReflexes, and Release Clause.

2 Partie 1 – Analyse descriptive des données

Dans cette partie on utilisera la librairie pandas pour charger ainsi que observer notre dataset.

2.1 Préparation des données

Une première étape est de supprimer les attributs qui ne sont pas pertinents, tel que:

- Photo
- Club Logo
- Flag
- Real Face

Grâce à la fonction ***info()*** nous pouvons observer les différents attributs, leur présances ainsi que leur types. On constate, que pour certains attributs, plusieurs valeurs sont manquantes. Ainsi qu'un mauvais typage qui est du à la présence de caractère avec les chiffres que nous devrons "nettoyer" par la suite. (Fig 1)

Column	#NaN	Type	Column	#NaN	Type
Name	0	object	Age	0	int64
Nationality	0	object	Overall	0	int64
Potential	0	int64	Club	241	object
Value	0	object	Wage	0	object
Special	0	int64	Preferred Foot	48	object
International Reputation	48	float64	Weak Foot	48	float64
Skill Moves	48	float64	Work Rate	48	object
Body Type	48	object	Position	60	object
Jersey Number	60	float64	Joined	1553	object
Loaned From	16943	object	Contract Valid Until	289	object
Height	48	object	Weight	48	object
LS	2085	object	ST	2085	object
RS	2085	object	LW	2085	object
LF	2085	object	CF	2085	object
RF	2085	object	RW	2085	object
LAM	2085	object	CAM	2085	object
RAM	2085	object	LM	2085	object
LCM	2085	object	CM	2085	object
RCM	2085	object	RM	2085	object
LWB	2085	object	LDM	2085	object
CDM	2085	object	RDM	2085	object
RWB	2085	object	LB	2085	object
LCB	2085	object	CB	2085	object
RCB	2085	object	RB	2085	object
Crossing	48	float64	Finishing	48	float64
HeadingAccuracy	48	float64	ShortPassing	48	float64
Volleys	48	float64	Dribbling	48	float64
Curve	48	float64	FKAccuracy	48	float64
LongPassing	48	float64	BallControl	48	float64
Acceleration	48	float64	SprintSpeed	48	float64
Agility	48	float64	Reactions	48	float64
Balance	48	float64	ShotPower	48	float64
Jumping	48	float64	Stamina	48	float64
Strength	48	float64	LongShots	48	float64
Aggression	48	float64	Interceptions	48	float64
Positioning	48	float64	Vision	48	float64
Penalties	48	float64	Composure	48	float64
Marking	48	float64	StandingTackle	48	float64
SlidingTackle	48	float64	GKDiving	48	float64
GKHandling	48	float64	GKKicking	48	float64
GKPositioning	48	float64	GKReflexes	48	float64
Release Clause	1564	object			

Figure 1: info()

Avant de passer aux étapes suivantes nous devons gérer les valeurs manquantes. Nous allons utiliser une valeur sentinelle -1. Néanmoins, nous pouvons aussi remplacer les vlauers manquantes avec la moyenne des valeurs de cet attribut, si il est continue, sinon par Unknown pour les valeurs discrètes.

Discrete variables

In [8]:
```
data['Wage'].fillna(-1, inplace = True)
data['Preferred Foot'].fillna(-1, inplace = True)
data['Weak Foot'].fillna(-1, inplace = True)
data['International Reputation'].fillna(-1, inplace = True)
data['Work Rate'].fillna(-1, inplace = True)
data['Body Type'].fillna(-1, inplace = True)
data['Position'].fillna(-1, inplace = True)
data['Club'].fillna(-1, inplace = True)
data['Joined'].fillna(-1, inplace = True)
data['Weight'].fillna(-1, inplace = True)
data['Height'].fillna(-1, inplace = True)
data['Contract Valid Until'].fillna(-1, inplace = True)
data['Loaned From'].fillna(-1, inplace = True)
```

Figure 2: Discrete variables

Numerical variables

In [7]:
```
data['ShortPassing'].fillna(data['ShortPassing'].mean(), inplace = True)
data['Volleys'].fillna(data['Volleys'].mean(), inplace = True)
data['Dribbling'].fillna(data['Dribbling'].mean(), inplace = True)
data['Curve'].fillna(data['Curve'].mean(), inplace = True)
data['FKAccuracy'].fillna(data['FKAccuracy'].mean(), inplace = True)
data['LongPassing'].fillna(data['LongPassing'].mean(), inplace = True)
data['BallControl'].fillna(data['BallControl'].mean(), inplace = True)
data['HeadingAccuracy'].fillna(data['HeadingAccuracy'].mean(), inplace = True)
data['Finishing'].fillna(data['Finishing'].mean(), inplace = True)
data['Crossing'].fillna(data['Crossing'].mean(), inplace = True)
data['Skill Moves'].fillna(data['Skill Moves'].median(), inplace = True)
data['Jersey Number'].fillna(-1, inplace = True)
data['Release Clause'].fillna(-1, inplace = True)
```

Figure 3: Numerical variables

2.1.1 Création Groupes d'âge

Afin de faciliter l'exploitation de nos algorithmes nous allons créer des groupes d'ages. Nous allons procéder comme suit:

- -20
- 20-25
- 25-30
- 30-35
- +35

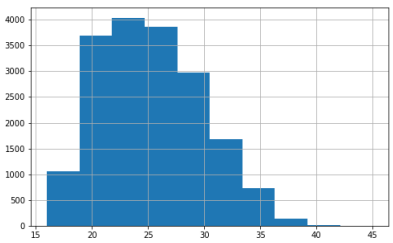

Figure 4: Initial Age distribution

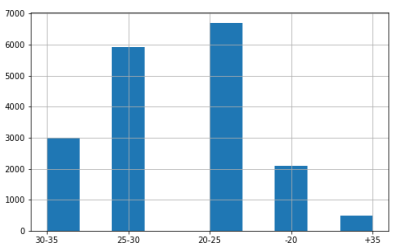

Figure 5: Age distribution groups

2.1.2 Poids/Tailles

Les attributs Poids (Weight) et Taille (Height) sont exprimés respectivement en livres (lbs) et en pouce (inch). Nous allons nettoyer l'attribut Weight en enlevant la chaine de caractère "lbs", et nous allons transformer les tailles en CM.
Nous allons changer notre valeur sentinelle à 0 pour les valeurs manquantes. Ceci sera nécéssaire pour le calcul du BMI lors de la prochaine étape.

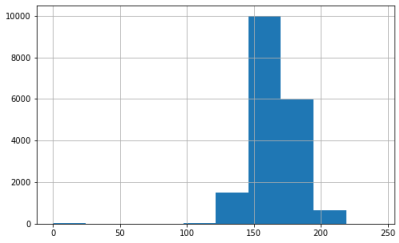

Figure 6: Weight distribution

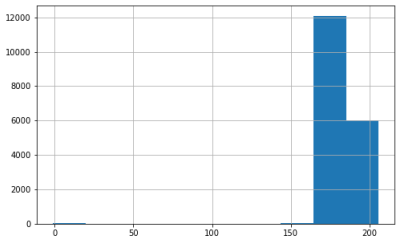

Figure 7: Height distribution

2.1.3 Création attribut BMI: Body Mass Index

Nous allons ajouté un nouvel attribut, BMI qui représente l'indice de la masse corporelle grâce à la formule:

$$poids[kg]/taille^2[m^2]$$

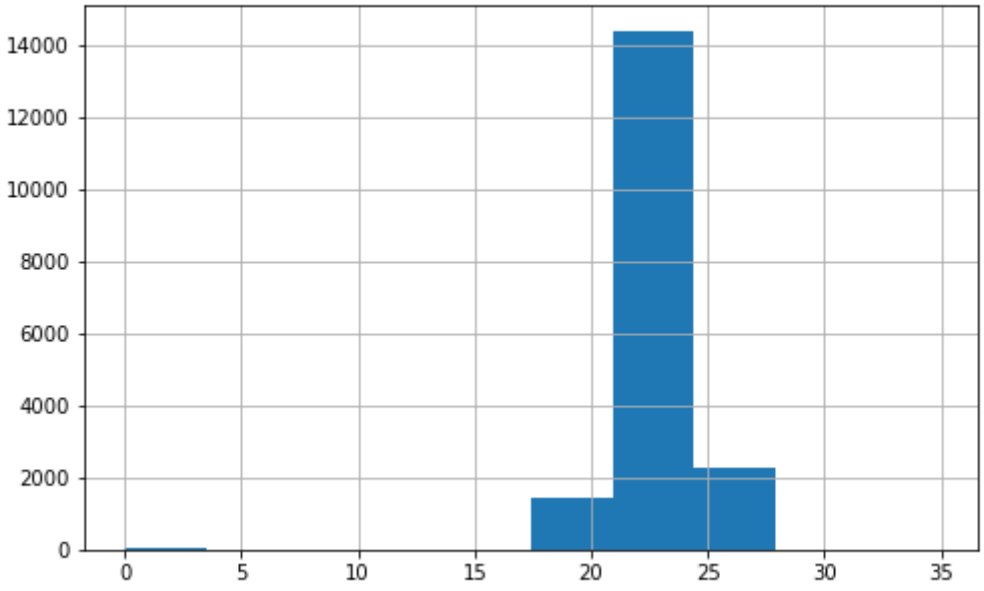

Figure 8: BMI distribution

2.1.4 Modification de la granularité des positions joueurs

Par défaut il existe plusieurs positions. Nous allons regroupper les positions en 4 groupes;

- GOAL
- DEF
- MID
- FWD

GK	G	Gardien
RB	DD	Défenseur droit
RWB	DLD	Défenseur latéral droit
LB	DG	Défenseur gauche
LWB	DLG	Défenseur latéral gauche
CB	DC	Défenseur central
CDM	MDC	Milieu défensif central
CM	MC	Milieu central

CAM	MOC	Milieu offensif central
LM	MG	Milieu gauche
LW	AG	Ailier gauche
LF	AVG	Avant centre gauche
RM	MD	Milieu droit
RW	AD	Ailier droit
RF	AVD	Avant centre droit
CF	AT	2eme attaquant
ST	BU	Buteur

Figure 9: Default position

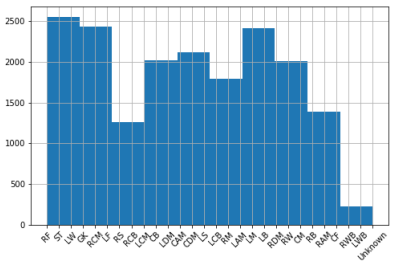

Figure 10: Old Position distribution

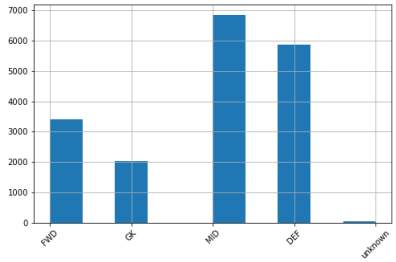

Figure 11: New Position distribution

2.1.5 Valeurs actuelles

Nous allons garder que les valeurs actuelles pour les performances. Nous garderons ici notre valeur sentinelle -1 pour représenter les valeurs manquantes.

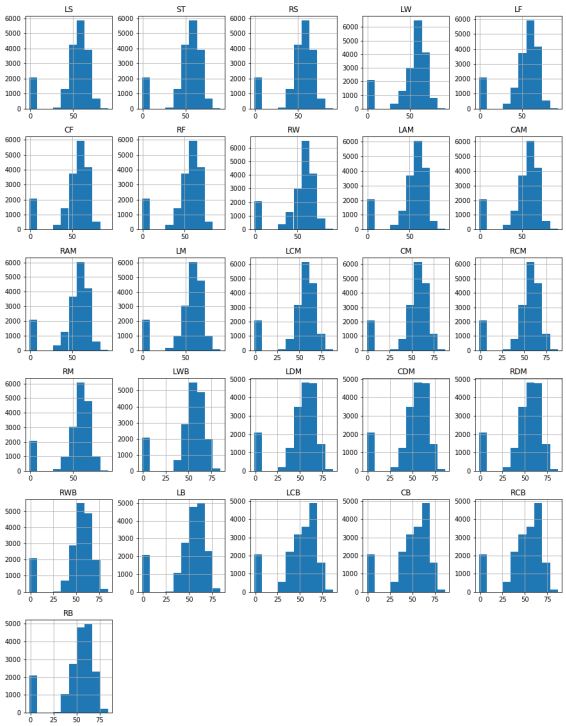

Figure 12: Stats distribution

2.1.6 Valeurs financières

Nous allons adapter les attributs financier. Enlever les symboles et normaliser les valeurs.

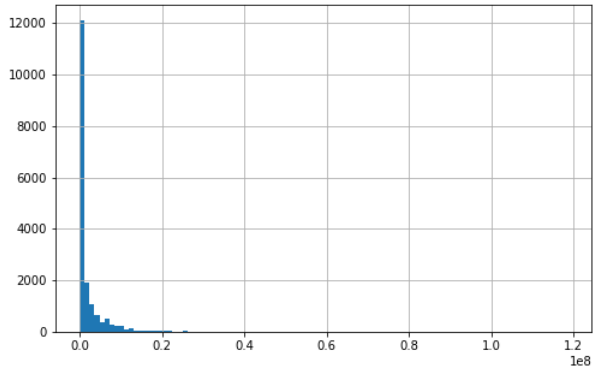

Figure 13: Value distribution

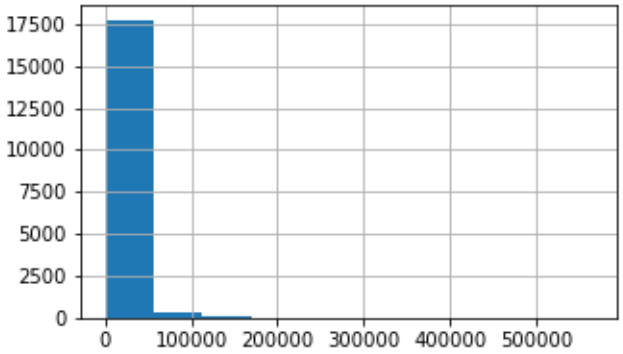

Figure 14: Wage distribution

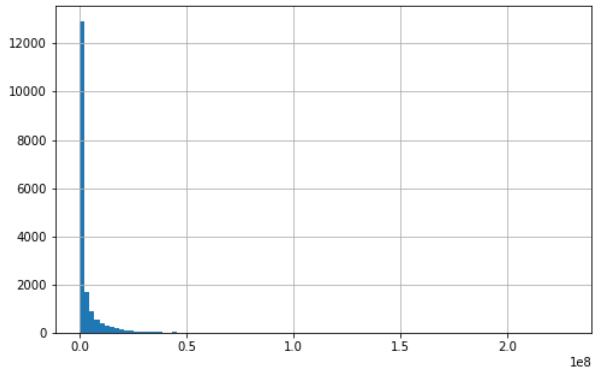

Figure 15: Release Clause distribution

2.1.7 Discrétisation de Wage en DWage

Dans cette partie nous allons discrétiser l'attribut Wage en DWage. Pour ce faire nous allons tout d'abord observer la distribution de ces valeurs grâce à un histograme avec différentes valeur de bin (ie: paquet).

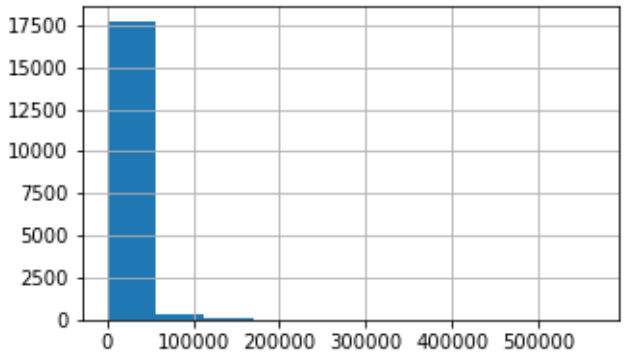

Figure 16: Wage distribution with default bin

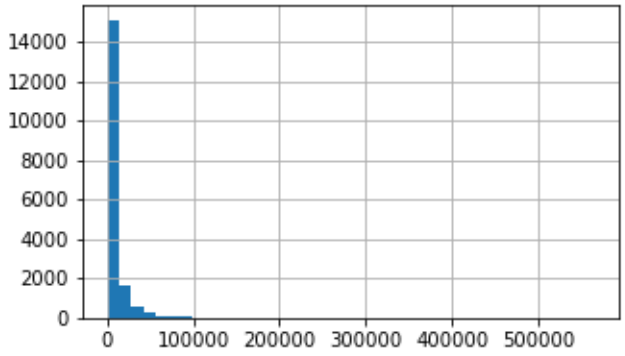

Figure 17: Wage distribution with bin=40

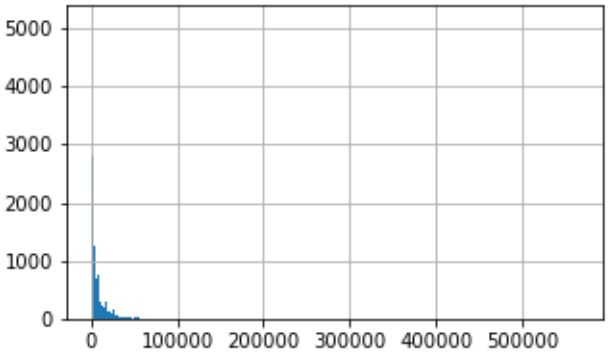

Figure 18: Wage distribution with bin=500

On remarque ici que la plus part des vlauers sont concentrées entre [0, 100.000]. Néanmoins, nous pouvons observer qu'il y a plusieurs groupes entre ces deux bornes. Nous allons séparer ces valeurs

en 10 intervalles (+1 qui représente les valeurs sentinelles). Nous aurions pu choisir de découper en moins d'intervalles pour avoir de plus gros paquets (tout dépend de l'utilisation future).

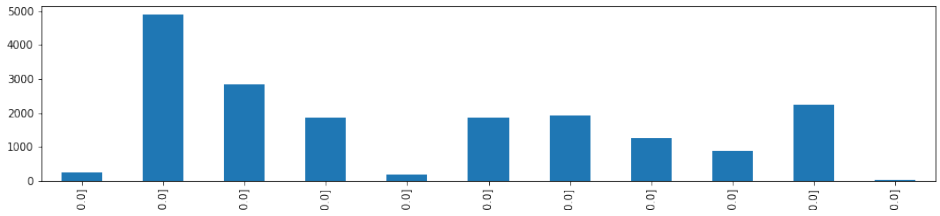

Figure 19: Wage intervalle distribution

2.1.8 Discrétisation de Value en DValue

Nous allons faire la même chose que pour DWage.

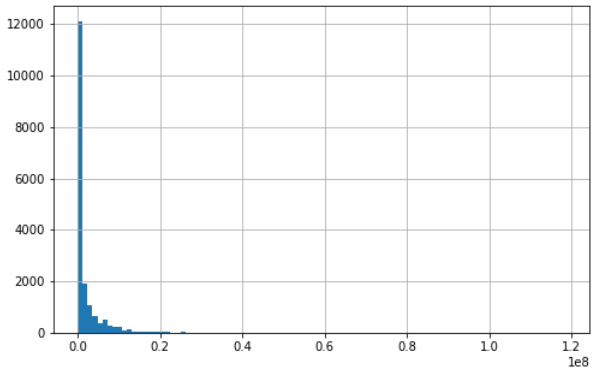

Figure 20: Wage distribution with default bin

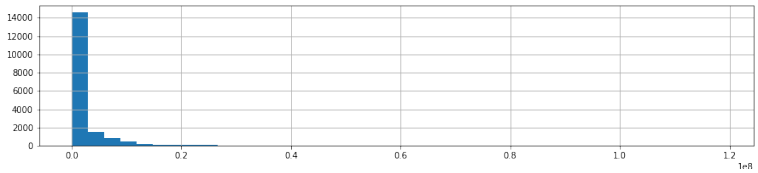

Figure 21: Wage distribution with bin=40

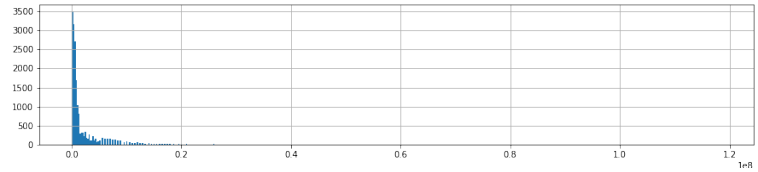

Figure 22: Wage distribution with bin=500

Apres discrétisation : On remarque ici qu'on aurait pu regroupper plusieurs groupes ensembles pour avoir moins d'intervalles. Cela, se discute selon les besoins.

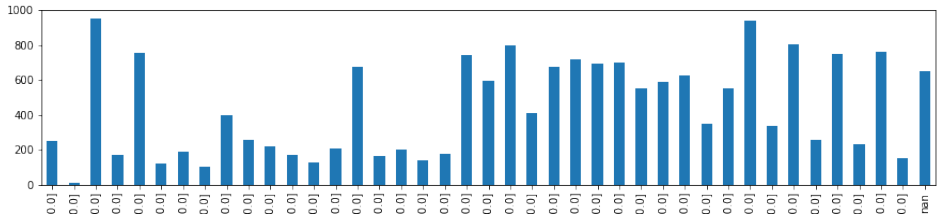

Figure 23: DValue intervalle distribution

2.2 Analyse

2.2.1 Visualisation

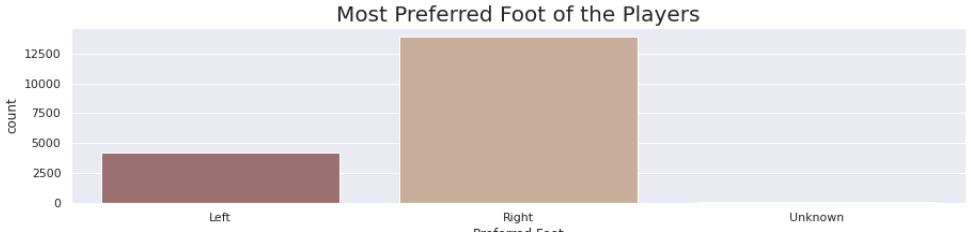

Figure 24: Prefered Foot

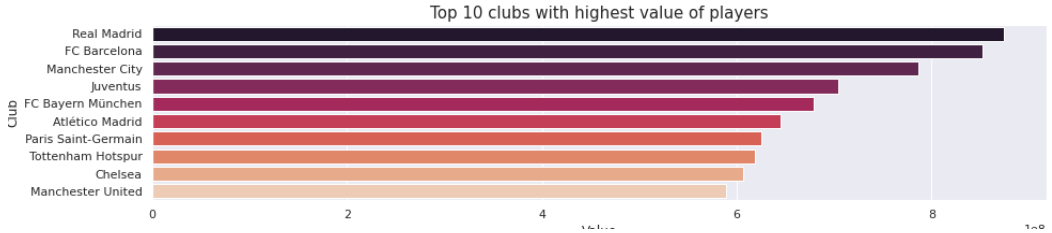

Figure 25: TOP10 Club

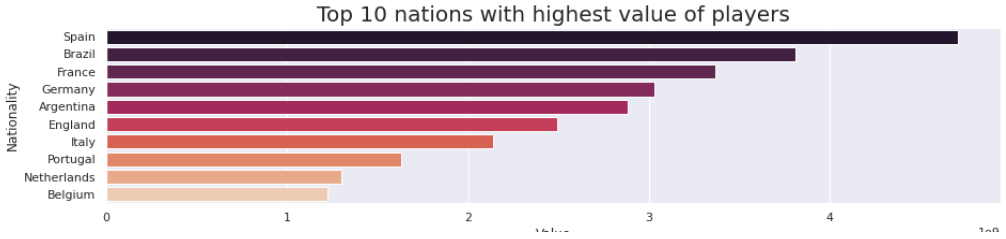

Figure 26: TOP10 Nations

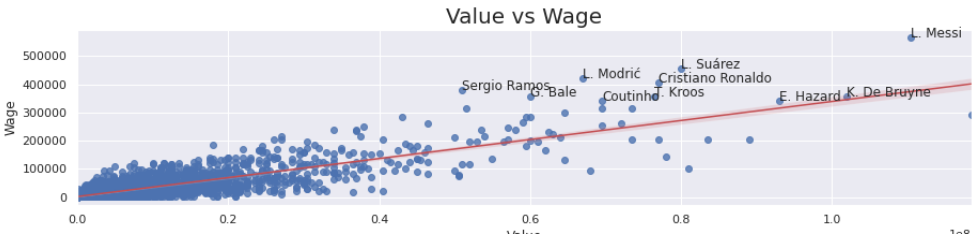

Figure 27: Value vs Worth

2.2.2 Analyse de corrélation

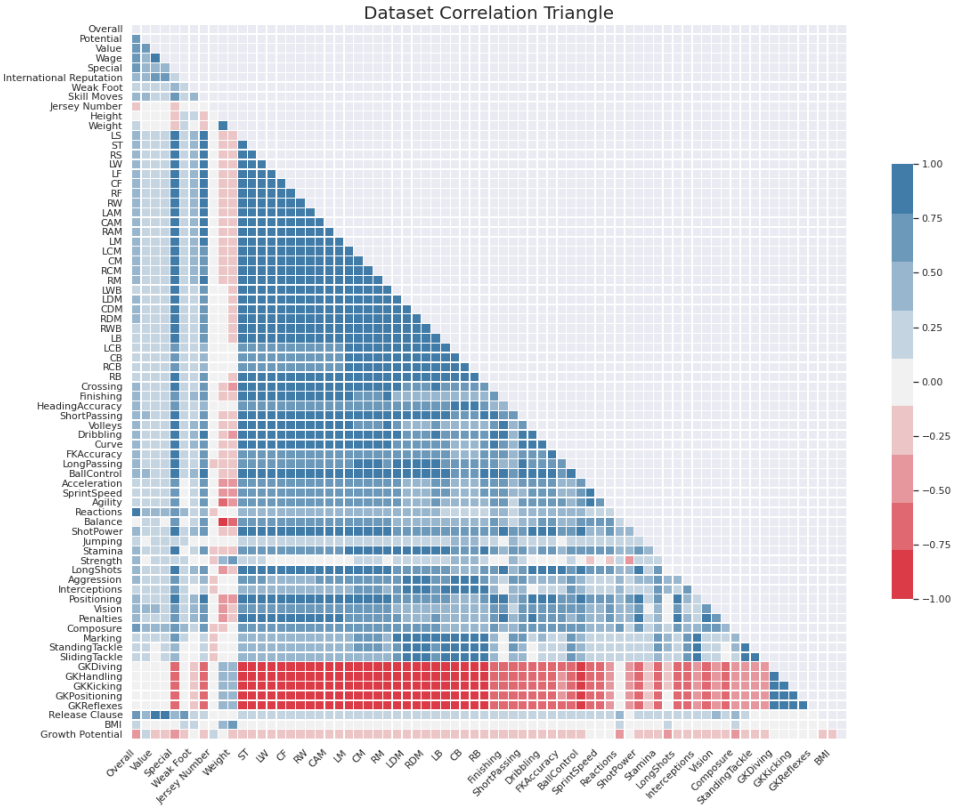

Figure 28: Correlation

On peut remarquer d'après le tableau de correlation que plusieurs variables sont fortement corréler. En particulier les variables qui représentent les stat des joueurs, telque : LS, ST, RS,

LW, LF, CF, RF, RW, LAM, CAM, RAM, LM, LCM, CM, RCM, RM.
D'autre part, on remarque les attribut GK* qui représentent les statistique du gardien, ne sont pas corrélés avec les stats précédentes.

2.2.3 Équipe la plus chère

Afin de constituer l'équipe la plus chère, nous allons sélectionner 11 joueurs qui se répartissent comme suit : 1 Gardien, 4 Defenseurs, 4 Milieus, 2 Attaquants, d'après l'attribut Value.
Gardien: Defenseur: Milieu: Attaquant:

	Name	NPosition	Overall	Nationality	Value
3	De Gea	GK	91	Spain	72000000.0

	Name	NPosition	Overall	Nationality	Value
42	S. Umtiti	DEF	87	France	57000000.0
8	Sergio Ramos	DEF	91	Spain	51000000.0
44	K. Koulibaly	DEF	87	Senegal	51000000.0
62	R. Varane	DEF	86	France	50000000.0

	Name	NPosition	Overall	Nationality	Value
4	K. De Bruyne	MID	91	Belgium	102000000.0
25	K. Mbappé	MID	88	France	81000000.0
17	A. Griezmann	MID	89	France	78000000.0
11	T. Kroos	MID	90	Germany	76500000.0

	Name	NPosition	Overall	Nationality	Value
2	Neymar Jr	FWD	92	Brazil	118500000.0
0	L. Messi	FWD	94	Argentina	110500000.0

2.2.4 Équipe la plus forte

La même chose, mais nous nous basons sur l'attribut Overall.

	Name	NPosition	Overall	Nationality	Value
3	De Gea	GK	91	Spain	72000000.0

	Name	NPosition	Overall	Nationality	Value
8	Sergio Ramos	DEF	91	Spain	51000000.0
12	D. Godín	DEF	90	Uruguay	44000000.0
24	G. Chiellini	DEF	89	Italy	27000000.0
34	M. Hummels	DEF	88	Germany	46000000.0

Les deux équipes ne sont pas identiques car les deux attributs, Overall et Value ne sont pas fortement corrélées.

	Name	NPosition	Overall	Nationality	Value
4	K. De Bruyne	MID	91	Belgium	102000000.0
6	L. Modrić	MID	91	Croatia	67000000.0
11	T. Kroos	MID	90	Germany	76500000.0
13	David Silva	MID	90	Spain	60000000.0

	Name	NPosition	Overall	Nationality	Value
0	L. Messi	FWD	94	Argentina	110500000.0
1	Cristiano Ronaldo	FWD	94	Portugal	77000000.0

3 Partie II - Segmentation

Avant d'appliquer les méthodes de clustering, nous allons normaliser nos données. De plus afin de "faciliter" la tâche des algorithmes, nous allons effectuer une PCA pour réduire l'espace de dimension.

3.1 Jambu Elbow

Afin de trouver le meilleur nombre de cluster, nous allons utiliser la courbe de Jambu Elbow. Nous

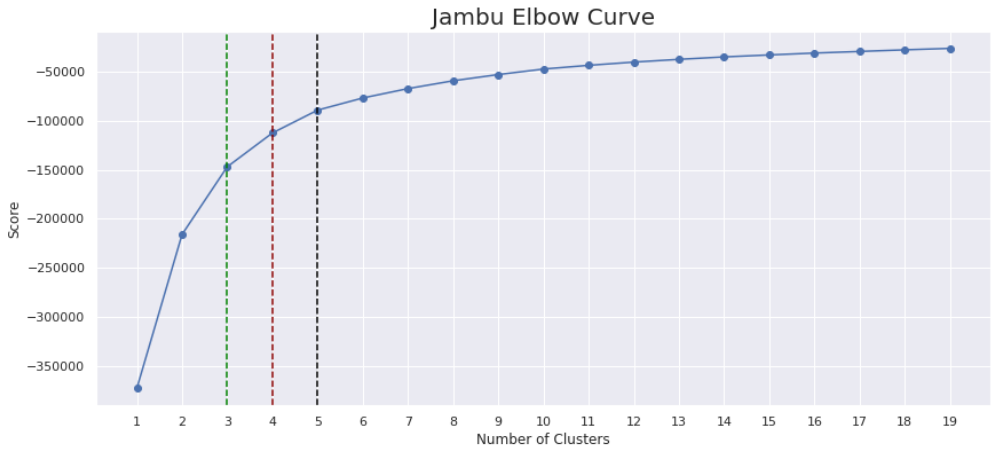

Figure 29: elbow curve

observons que la courbure se fait entre le 3 et le 5. Nous allons par la suite tester les différentes valeurs 3,4 et 5.
Nous allons observer que Kmeans arrive, plus ou moins, à bien séparer des groupes.

3.2 Kmeans

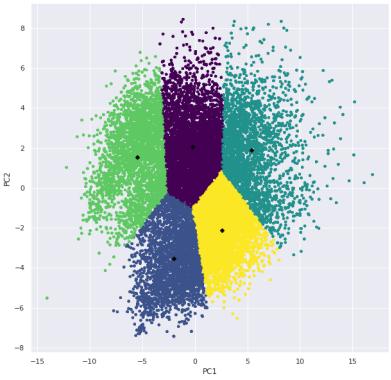

Figure 30: Kmeans K=5

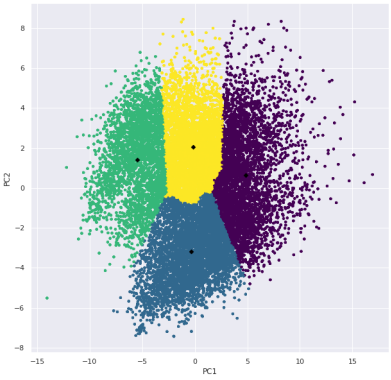

Figure 31: Kmeans K=4

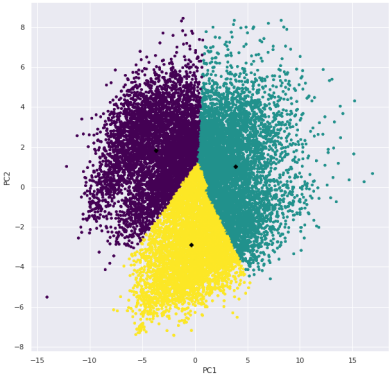

Figure 32: Kmeans K=3

3.3 Spectral Clustering

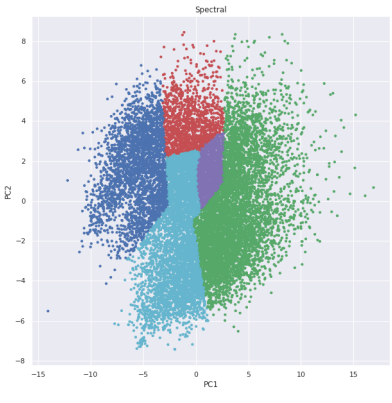

Figure 33: Spectral Clustering K=5

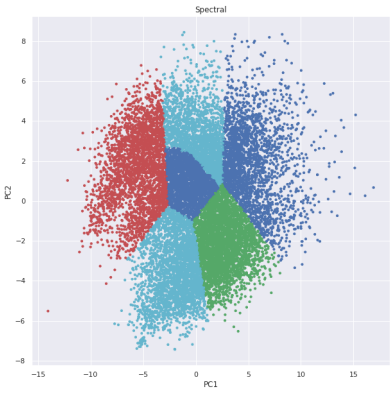

Figure 34: Spectral Clustering K=4

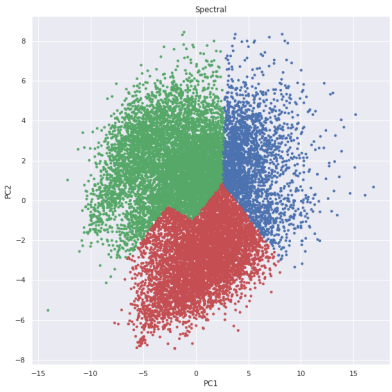

Figure 35: Spectral Clustering K=3

3.4 DBSCAN

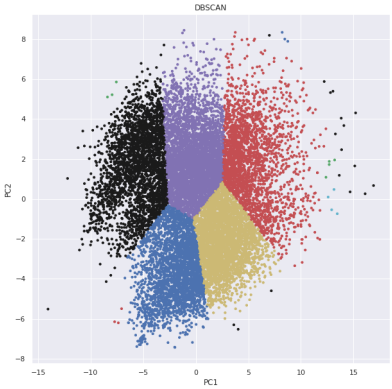

Figure 36: DBSCAN min samples 3

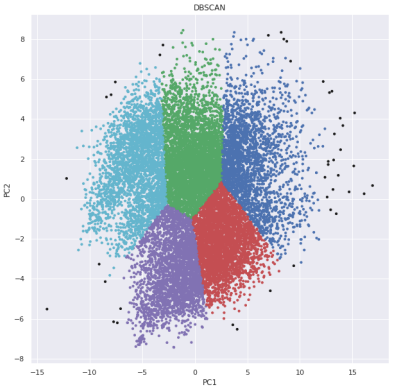

Figure 37: DBSCAN min samples 5

3.5 MiniBatchKMeans

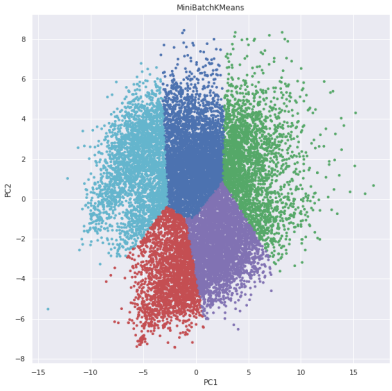

Figure 38: MiniBatchKMeans K=5

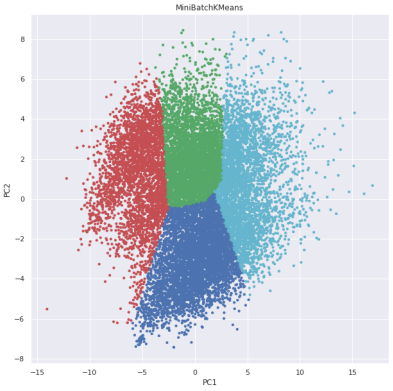

Figure 39: MiniBatchKMeans K=4

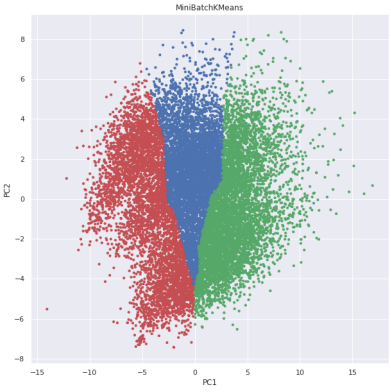

Figure 40: MiniBatchKMeans K=3

3.6 Birch

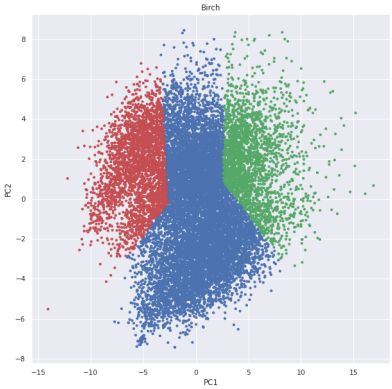

Figure 41: Birch K=5

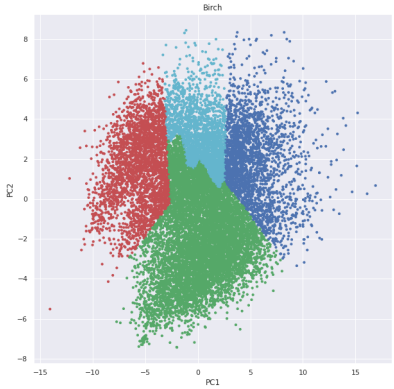

Figure 42: Birch K=4

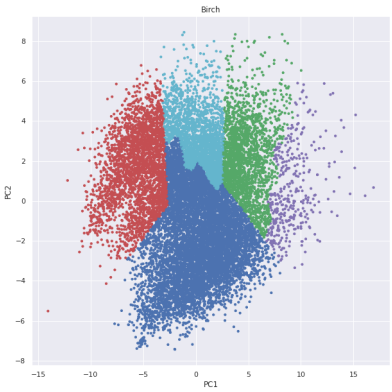

Figure 43: Birch K=3

On remarque que la pluspart des algos ne retourne pas le même clustering, mais qu'il y a quand meme une similitude. Je n'ai pas eu le temps d'effectuer une analyse complète pour cette partie ... Il faudrait créer une fonction qui plot les positions des joueurs pour voir si on arrive bien à retomber sur nos 4 groupes (GK,DEF,MID,FWD). La comparaison des algo se fera ensuite sur le pourcentage de bonne classification !

3.7 Remplacant à MBappé

En utilisant les composantes principales, et grâce à une fonction de sitance, nous allons prendre le joueur le plus proche de MBappé. Le joueur sélectionné est Cristiano Ronaldo (voir notebook)