Makhloufi Khalil 22104011

Science des données Projet

Groupe du Jeudi 14h45-16h45

Table des matières

1) Apprentissage non supervisé

1.1) Jeu de données

Le jeu de données utilisé pour le projet provient d'une étude réalisé par la FIFA (Fédération internationale de football association) aux États-Unis, en 2017. Ces données correspondent à une multitude de statistiques, qui peuvent varier sur un intervalle de 1 à 99, de plusieurs joueurs de football professionnels toute nations confondus.

On peut y constater un tableau contenant 941 lignes (940 joueurs) et 17 colonnes (17 variables). Chaque joueur de foot est décrit par 17 variables dont 15 variables quantitatives et 2 variables qualitatives.
Variables qualitatives nominales (2) :
- Nom : nom et prénom du footballeur.
- Pied fort : pied dont le joueur utilise au mieux (2 valeurs possibles : droite ou gauche).
Variables quantitatives discrètes (17) : finition, précisions de tête, passe courte, dribble, effet, passe longue, accélération, souplesse (agilité), endurance, force, interceptions, penalty, marquage, gardien position, gardien réflexes. (Tous compris entre 1 et 99)

Afin d'avoir pu obtenir ce tableau de statistiques, l'utilisation du lien suivant a vraiment été efficace pour la constitution du jeu de données. (Lien : https://data.world/raghav333/fifa-players/workspace/file?filename=fifa_cleaned.csv).
Après avoir récupéré le jeu de données, le nombre de lignes (de joueurs) est passé d'environ 14000 à presque 1000 lignes (941 joueurs exactement) et les variables qui n'avaient pas beaucoup d'importance pour différencier les données (ex : âge) ont été supprimées et 17 variables (colonnes) sont restées afin de manipuler au mieux au mieux.

1.2) Clustering

1.2.1) Chaîne de traitement

Lors d'un clustering hiérarchique, il faut importer un fichier qui contient un jeu de données que l'on veut analyser. L'outil « file » doit être implémenté en premier et l'on doit importer le jeu de données décrit précédemment. Dans un second temps, nous pouvons visualiser les données à travers un nuage de point. Le plus important dans cette partie, c'est d'établir la matrice de distance pour pouvoir réaliser le clustering. Il faut donc utiliser l'outil « Distance » qu'on peut direct relier à « distance matrice » pour explorer la matrice de distance entre les différents joueurs. Enfin grâce au distances, nous pouvons désormais réaliser le clustering hiérarchique. Pour cela, nous devons ajouter, en sortie de « distances », « Hierarchcal Clustering » qui nous fournis un dendogramme dont on peut extraire une multitude d'informations. Pas strictement néssecaires, mais nous avons la possibilité de regarder les resultats à travers un tableau. (« Data Table »)

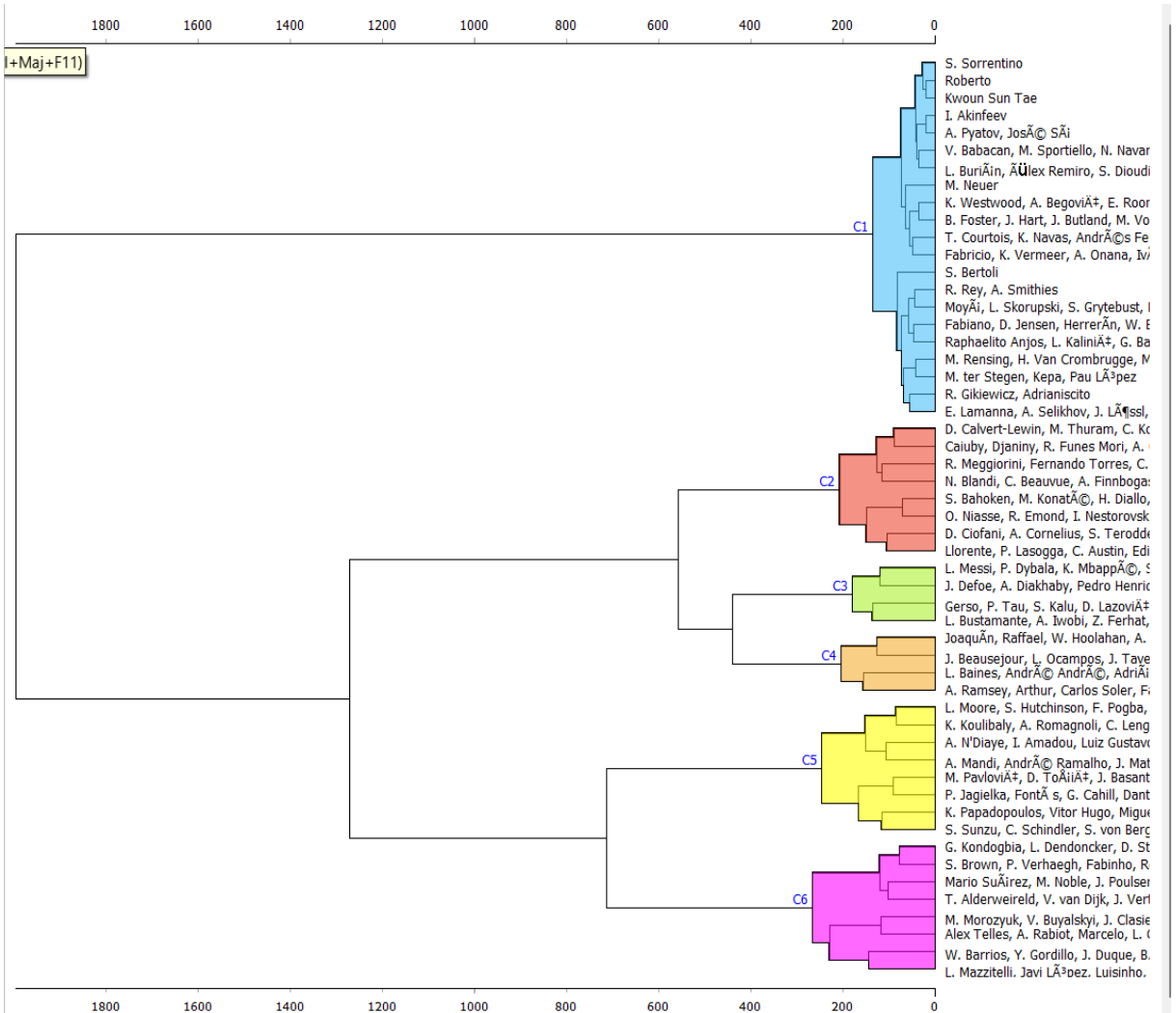
1.2.2) Explication des résultats

Tout d'abord, lorsque l'on réalise un clustering hiérarchique sur ce jeu de donnée, nous pouvons voir dans un premier temps qu'il remarque si le footballeur est un joueur de champ ou non. Voici ci-dessous une capture d'écran montrant le dendrogramme réaliser grâce au clustering. Dans ce cas-là, le top N est très efficace lorsqu'il est égal à 6 car il permet vraiment de voir les principales différences entre les 941 joueurs de football.
Cela peut faire penser à une supposition : le clustering va repartir les joueurs en 4 groupes qui pourrait correspondre à leurs poste respective. Observons de plus près le clustering hiérarchique afin d'en savoir un peu sur l'apprentissage non supervisé de ce jeu de données.

Premièrement, parmi ceux qui ne sont pas des joueurs de champs (C1), on peut observer absolument tous les gardiens de buts. De plus, il met d'un côté les gauchers et les droitiers. Parmi les gauchers, il crée deux cluster en mettant d'une part les gardiens qui ont une meilleure statistique en réflexe qu'en plongeon et à l'inverse, il met dans le second clusters ceux qui sont meilleur en plongeon. C'est exactement le même processus pour les droitiers.

Exemple 1 : Si l'on prend le joueur nommé « T.Courtois ». Nous voyons bien que ce n'est pas un joueurs de champ donc un gardien (C1). De plus nous pouvons voir que c'est un gaucher et qu'il possède un meilleur plongeon que de réflexes.

Deuxièmement, on a les joueurs de champ, le clustering observe si le joueur a un style offensive ou défensive. D'une part, parmi ceux qui sont défensive et gauchers, il repartie les joueurs afin de montrer si le joueur a plus-tôt un style d'un défenseur ou d'un milieu de terrain plus tôt défensive. Même classification pour les footballeur qui ont un style défensive mais droitiers. D'une autre part, en ce qui concerne les joueurs offensive



sdroitiers, le clustering hiérarchique met en place deux groupes, l'un comporte les joueurs qui sont plus-tôt des attaquants et le second cluster est composé de plusieurs milieu de terrain ayant un style plus tôt offensive. Même principe de clustering pour les joueurs offensives gauchers.

Exemple 2 : Si l'on prend le footballeur nommé « L. Messi ». Tout d'abord, c'est un joueur de champs donc pas un gardien. Il est gaucher et a un style offensive. Parmi les joueurs offensive, nous observons bien que c'est un attaquant. (C3)

Exemple 3 : Si l'on prend le footballeur nommé « Virgil Van Dijk». Tout d'abord, c'est un joueur de champs donc pas un gardien. Il est droitier et a un style défensive. Parmi les joueurs défensive, nous observons bien que c'est un défenseur. (C5)

Pour conclure, le clustering repère si le joueur est un goal, si c'est le cas il regarde son pied fort et le classe en fonction de son plongeon et de son réflexe. Le cas contraire, il regarde aussi le pied fort des autres joueurs de champs. Deux possibilités : style offensive ou défensive. Enfin quatre possibilité parmi les précédentes, soit le joueur est un défenseurs, milieu défensive, milieu offensive ou un attaquant.

2) Apprentissage supervisé

2.1) Jeux de données

2.1.1) Jeu de données étiquetés

Tout d'abord, pour l'apprentissage supervisé, le jeu de données étiquetés utilisé est exactement le même que celui utilisé pour l'apprentissage non supervisé avec les mêmes variables quantitatives et nominales mais avec 310 données (joueurs de football).
Sauf qu'il y une variable (classe) qui s'ajoute intitulé « poste » qui correspond à une variable qualitative nominale qui est composé de 4 issus : « Gardien, Défenseur, Milieu, Attaquant ». La méthode pour avoir obtenu ce jeu de données était seulement de reprendre le jeu de donnée précédent et de restreindre le nombre de lignes (de joueurs) de 1000 a 310 et d'ajouter la variable « poste ».

2.1.2) Jeu de données à prédire

Deuxièmement, pour le jeu de données à prédire, les données à prédire correspondent à une liste de 20 joueurs populaires de football où cette fois-ci, les données ont été crée dans un fichier CSV sur excel. Il comporte les même variables que le jeu données utilisé pour l'apprentissage non supervisé mais avec un nombre de données beaucoup moins important.

2.2) Modèles d’apprentissage

2.2) Chaîne de traitement (Donnée étiquetés)

Tout d’abord, comme pour l’apprentissage non-supervisé, nous avons besoin d’importer un ficher. Nous réutilisons l’outil « File » et nous importeront le jeu de données expliquer dans la partie 2.1.1. L’usage de « data table » en sortie de « File » permet de voir toutes les données réuni dans un tableaux pour bien vérifier qu’il s’agit du bon jeu de données. Nous voulons classer nos données, nous avons donc besoin de plusieurs modèles qui se placeront tous à la sortie de « File ». Test & score permet de voir les compétences de chaque modèles, plaçons le à la sortie des modèles mais aussi à la sortie de « File ». Ensuite, nous voulons voir les différentes classification sous forme d’une matrice, utilisons « Confusion Matrix » à la sortie de « Test & Score » et observons attentivement cette matrice. Enfin comme précédemment, nous pouvons visualiser les résultats sous forme d’un tableau puis les représenter sous forme d’un nuage de points.

2.2.2) Explication des résultats (Donnée étiquetés)

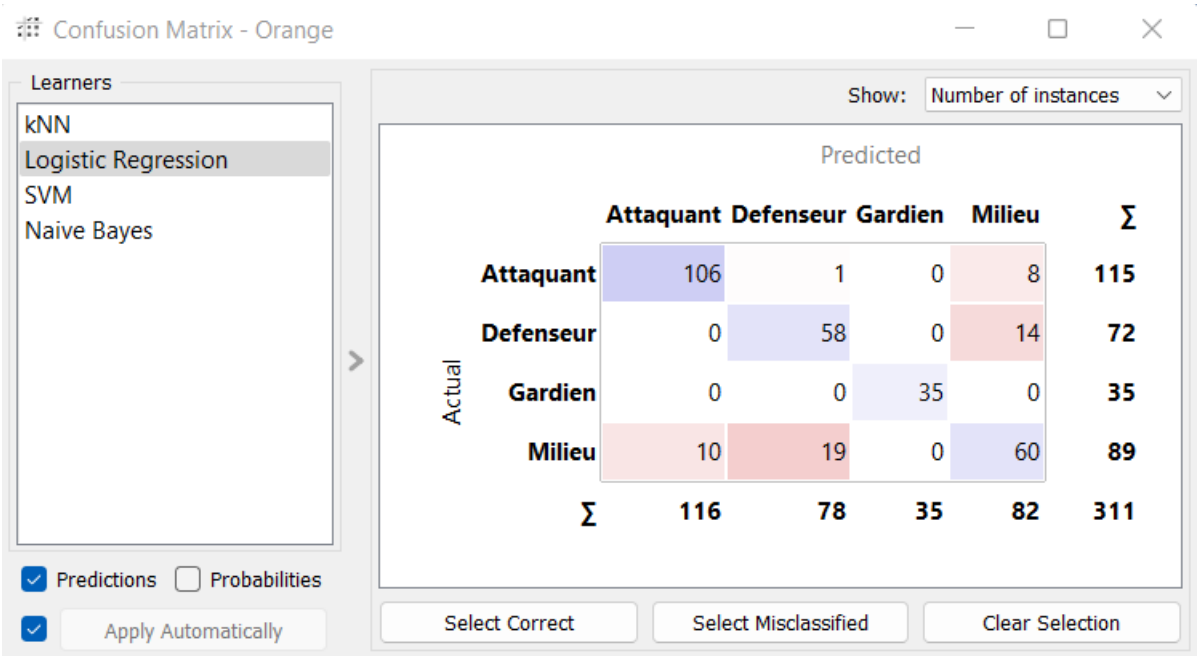
Pour pouvoir observer les performances de plusieurs modèles (nous utiliserons les modèles suivant : kNN, SVM, Naive Bayes, Logistic Régression), il faut lui donner des données (ici le jeu de donnée présenter ci-dessus). De plus, nous avons besoin qu’une variable joue le rôle de « target » (le target sera la variable poste). Lorsqu’on veut évaluer les performances d’un modèles il faut examiner son rappel, sa précision, sa f-mesure et son accuracy (mais ce ne sera pas la seule chose à vérifier). Nous avons donc besoin de l’outil test & score.

D’après la photo suivante, en prenant le modèle « Logistic regression », nous pouvons voir que le rappel, la précision, la f-mesure et l’accuracy sont plus-tôt bon.

Evaluation results for target (None, show average over classes)

Model	AUC	CA	F1	Precision	Recall
kNN	0.925	0.826	0.822	0.834	0.826
SVM	0.934	0.817	0.814	0.816	0.817
Naive Bayes	0.931	0.797	0.794	0.794	0.797
Logistic Regression	0.941	0.833	0.832	0.832	0.833

En comparent les modèles, celui de logistic regression semble avoir la meilleur performance. Tout de même, il faut aussi attentivement regarder la matrice de confusion que génère le modèles même si les résultats du modèle semblent être efficaces.



Observons la matrice de confusion généré par Logistic Regression. Premièrement, tous les gardiens sont correctement classés comme des gardiens. D'autre part, ce modèle a classé 10 milieu comme des attaquants, 19 milieu comme des défenseur, 14 défenseur comme des milieu, 8 attaquant comme des milieu et seulement 1 défenseur comme un attaquant.

Parmi les footballeur mal classé, cela peut être assez compréhensible car si l'on regarde attentivement ces joueurs, nous avons Luiz Gustavo qui est de base un milieu mais qui a été classé comme étant un défenseur. Luiz Gustavo a de très bonne statistiques en défense, force et interceptions (trois principales compétences d'un défenseur) mais beaucoup moins en passe. C'est pourquoi le modèle à classifier Luis Gustavo comme étant un défenseur car il a exactement les statistiques requis d'un défenseurs.

D'une autre part, Isco est un attaquant mais a été classé comme milieu. Isco n'a pas réellement une bonne statistiques de finition ni de vitesse ni de précision de tête (principales compétences d'un attaquant). En revanche, les passes (courtes longues), les dribbles (principales compétences d'un milieu de terrain) sont très élevés, ce qui laisse penser, pour le modèle Logistic Regression, qu'Isco est beaucoup plus un milieu qu'un attaquant.

Enfin, Rafael Guerreiro est de base un défenseur (gauche) qui a été classé comme un milieu. Une fois de plus, Guerreiro a de très bonne statistiques en passe et dribbles et pas beaucoup de défense ni beaucoup de physiques. Ces caractéristiques font clairement référence à un milieu de terrain d'où sa classification en milieu.

Pour conclure, les raisons pour lesquelles le modèle a mal classé quelque joueurs peuvent être compréhensible car ces joueurs ont des statistiques plus appropriés à un autre poste et peuvent être plus performant sur le terrain et profiter au maximum de leurs statistiques les plus remarquables.

On pourrait même s’amener à dire si le modèle à chercher à attribuer un second poste aux joueurs incorrectement classés et classer correctement les joueurs qui ne peuvent jouer qu’un poste seulement. Ceux-ci reste évidemment une supposition.

2.3) Prédictions

2.3.1) Chaîne de traitement (donnée à prédire)

Une fois de plus, nous devons importer le fichier qui comporte le jeu de données à prédire. Pour prédire sur le logiciel Orange, nous devons utiliser l’outil Predictions que nous mettons à la sortie de file. Pour prédire, il faut utiliser des modèles qui ont déjà été utilisé précédemment, Relions tous les modèles à prédictions et observons en détails les prédictions des données, Enfin, nous pouvons visualiser les prédictions dans Scatter Plot en ayant les deux fenêtres ouverts (prediction et nuage de points. Interprétons les résultats des prédiction.

2.3.2) Interprétation des résultats (donnée à prédire)

Nous voulons dès à présent prédire le poste des joueurs dont on a aucune informations. En effet, l’outil prédictions va clairement nous être utile pour cette partie. Voici à droite une capture d’écran montrant la prédictions des poste du modèle « Logistic Régression » des 19 joueurs concernés. En ayant un minimum de connaissances du football, tous les joueurs ont correctement été classés sauf un seul nommé Patrick Vieira (l.9). Nous savons que le français joue au milieu de terrain et non en défense. Comme dans une précédente partie analysons de près ce joueurs. Vieira a de très bonnes statistiques en passe (principale qualité d’un milieu), interceptions, force et de marquage (qualités super excellentes pour un défenseur). Mais une fois de plus, pour le modèle, ces statistique sont plus appropriés à un défenseur.

	Logistic Regression	Nom
1	Attaquant	C.Ronaldo
2	Attaquant	Van persie
3	Attaquant	Forlan
4	Gardien	Buffon
5	Gardien	Sirigu
6	Defenseur	David Luiz
7	Defenseur	Thiago Silva
8	Milieu	Iniesta
9	Defenseur	Patrick Vieira
10	Defenseur	Theo Hernandez
11	Defenseur	Dani Alves
12	Defenseur	Virgil Van dijk
13	Attaquant	Eric Cantona
14	Gardien	Van der Sar
15	Defenseur	Ferdinand
16	Milieu	Steven Gerrard
17	Milieu	Paul Pogba
18	Milieu	Lucas Paqueta
19	Attaquant	Sidney Govou

Si l’on revient à la supposition de second poste, dans ce cas la c’est exactement le cas, Patrick Vieira est à la base un milieu mais peut tout a fait jouer au poste de défenseur

	Logistic Regression	SVM	Naive Bayes	kNN	Nom
1	Attaquant	Attaqu...	Attaquant	Attaqu...	C.Ronaldo
2	Attaquant	Attaqu...	Attaquant	Attaqu...	Van persie
3	Attaquant	Attaqu...	Attaquant	Attaqu...	Forlan
4	Gardien	Gardien	Gardien	Gardien	Buffon
5	Gardien	Attaqu...	Gardien	Gardien	Sirigu
6	Defenseur	Defens...	Defenseur	Defens...	David Luiz
7	Defenseur	Defens...	Defenseur	Defens...	Thiago Silva
8	Milieu	Milieu	Milieu	Milieu	Iniesta
9	Defenseur	Milieu	Milieu	Defens...	Patrick Vieira
10	Defenseur	Milieu	Milieu	Milieu	Theo Hernandez
11	Defenseur	Milieu	Milieu	Defens...	Dani Alves
12	Defenseur	Defens...	Defenseur	Defens...	Virgil Van dijk
13	Attaquant	Attaqu...	Milieu	Attaqu...	Eric Cantona
14	Gardien	Gardien	Gardien	Gardien	Van der Sar
15	Defenseur	Defens...	Defenseur	Defens...	Ferdinand
16	Milieu	Milieu	Milieu	Milieu	Steven Gerrard
17	Milieu	Milieu	Milieu	Milieu	Paul Pogba
18	Milieu	Milieu	Milieu	Milieu	Lucas Paqueta
19	Attaquant	Attaqu...	Attaquant	Attaqu...	Sidney Govou

Si l’on compare le logistic regression à d’autres modèles et que l’on observe les différentes prédictions des quatres modèles (voir photo de gauche), on peut tout de suite remarquer que la SVM n’est pas du tout efficace car il place Sirigu, qui est de base un gardien, en attaquant ce qui est inexplicable. D’autre part, le Naive Bayes (un peu semblable au kNN) lui est pareil que logistic regression sauf qu’il considère les défenseurs droitiers et gauchers comme des milieu de terrains. Une fois de plus, cela reste tout a fait compréhensible (pas comme le SVM) car ils les statistiques parfaites pour jouer au milieu.

Pour conclure, le modèles prédit les données en fonction de leurs statistiques les plus élevés qui font référence au poste concernés. (Les statistiques de Vieira sont plus élevés dans le domaine de la défense donc il le place en défenseur).