ÉCOLE NATIONALE DES SCIENCES APPLIQUÉES D'AL-HOCEIMA
DÉPARTEMENT MATHÉMATIQUES ET INFORMATIQUE

# RAPPORT DU PROJET

## PRÉDICTION DES JOUEURS NOMMÉS AU BALLON D'OR DU FIFA


**ÉQUIPE DU PROJET**

OUMAMI SAMIRA

AMINI ALAOUI IMANE

BARHDADI MOHAMED

EL HADDAD OUSSAMA

ABDELLAOUI RACHID

LAKTIB AZIZ

**ENCADREMENT**

M. HADDOUCH KHALID

# BALLON D'OR : LES 30 NOMMÉS


NEYMAR


MODRIC


DYBALA


MARCELO


KANTE


SUAREZ


RAMOS


OBLAK


COUTINHO


MERTENS


LEWANDOWSKI


KROOS


DE BRUYNE


DE GEA


KANE


DZEKO


MESSI


GRIEZMANN


BUFFON


MANÉ


FALCAO


MBAPPÉ


RONALDO


CAVANI


BENZEMA


AUBAMEYANG


HAZARD


BONUCCI


ISCO


HUMMELS

# SOMMAIRE

# INTRODUCTION GÉNÉRALE

## CONTEXTE ET OBJECTIFS DU PROJET

La **FIFA Ballon d'or** est une récompense attribuée au meilleur joueur de football de l'année. Ce trophée individuel auquel beaucoup accordent une importance démesurée, dont les favoris sont débattus toute l'année et le vainqueur glorifié.

Créé en 1956 par le magazine France Football, ce titre était, à l'origine, attribué au meilleur joueur disposant d'une nationalité européenne évoluant dans un championnat européen. De 1995 à 2006, le Ballon d'or a été attribué au meilleur joueur évoluant dans un championnat européen sans distinction de nationalité. Depuis 2007, il récompense le meilleur joueur au monde, c'est-à-dire sans distinction de championnat ni de nationalité.

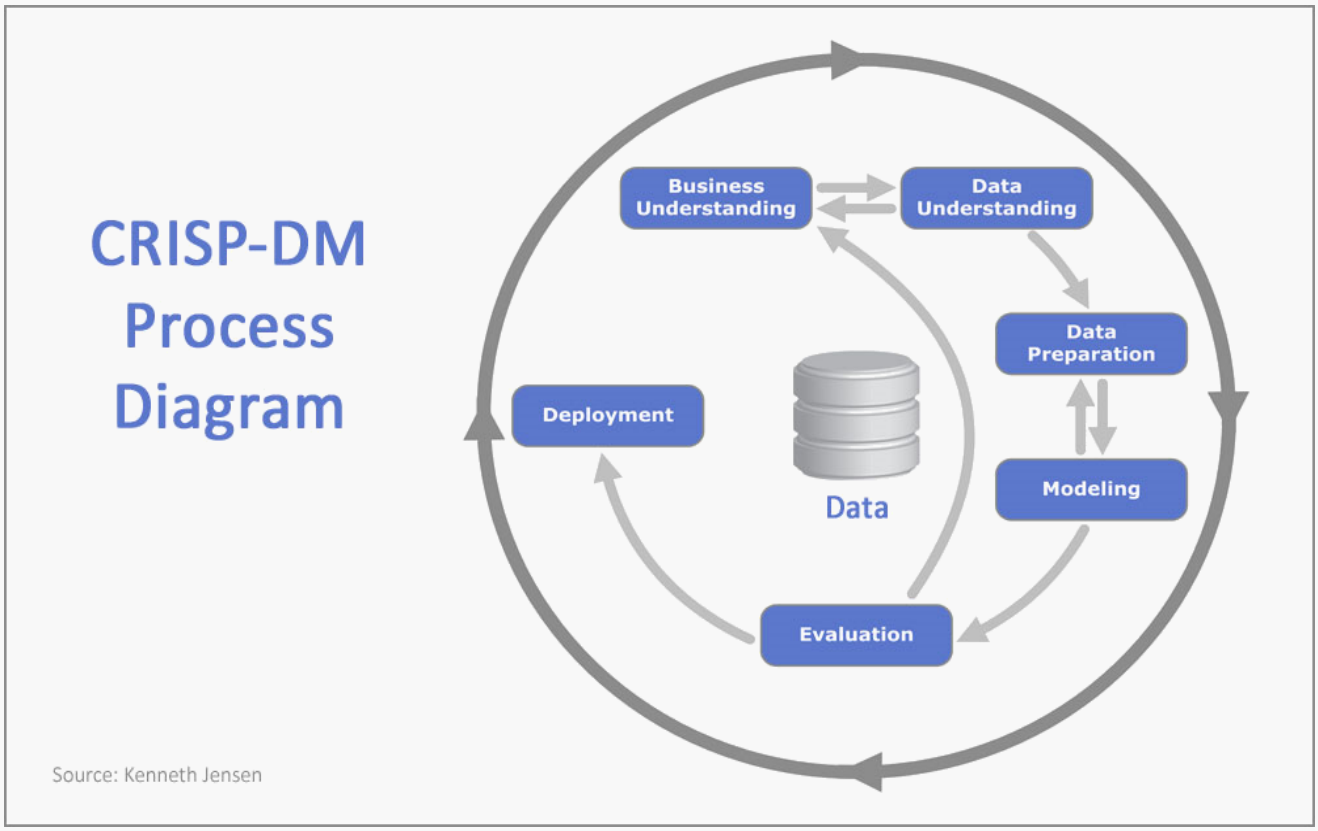
L'idée de notre intitulé « Prédiction des joueurs nommés au Ballon d'or» est d'essayer de prédire la liste des joueur sélectionnés par la FIFA, dont le gagnant du ballon d'or de l'année et sélectionné par le vote des 208 sélectionneurs des pays membres de la FIFA, en se basant sur les résultats et les statistiques de l'année courante.

Dans ce présent rapport nous verrons les différentes tâches faites pour réaliser ces objectifs tout en adoptant la méthodologie CRISP dédiée à la gestion des projets Data Mining.

Nous verrons tout d'abord la collection des données à partir des différentes sources existantes. Puis, on traitera l'implémentation de l'étude Data Mining sous R et Weka, et on finira par l'illustration des résultats obtenus.

# MÉTHODOLOGIE CRISP

## APPLICATION AU PROJET DATA MINING




## 1. Choix du sujet et identification des objectifs

Comme c'était mentionné dans l'introduction de ce rapport, le sujet concerne la prédiction des joueurs nommés pour le ballon d'or de l'année actuelle.

Notre projet était en réalité une partie complémentaire d'un autre projet intitulé « **Conception et réalisation d'un VeilleBox pour la Coupe du Monde 2018** », réalisé en module du « Business Intelligence & Veille Stratégique ».

L'objectif principal de ce projet, c'est de permettre aux passionnés à cet évènement annuel de prédire les joueurs dont le nom figura dans la liste sélectionnée par la FIFA tout en se basant sur leurs statistiques et leurs chiffres établis pour l'année de sélection.

Ce projet sera aussi un outil d'aide à la décision pour les médias afin de deviner les futures récompenses « Ballon d'or » afin de susciter la passion et l'admiration de leurs fans.

# MÉTHODOLOGIE CRISP
## APPLICATION AU PROJET DATA MINING

## 2. Compréhension des données

Les données que nous aurons besoin, sont les différentes statistiques et chiffres des 30 joueurs qui sont nommés à la FIFA ballon d'or pour la saison sportive 2016-2017. Ainsi que d'autres celles des joueurs qui ne sont pas sélectionnés pour cette récompense.

La FIFA reposes sur plusieurs variables pour sélectionnées les joueurs (Le nombres de minutes jouées, la performance, Les buts marqués, Championnats remportés..). Dans notre projet nous avons limité les variables explicatives ont 16 :

| | |
|---|---|
| **Player** | Le nom des joueurs nommés ou non |
| **Age** | Leurs âges |
| **Confederation** | La confédération dont leurs équipes appartiennent, tous UEFA |
| **Position** | Gradient de but, Milieu de terrain ou Attaquant |
| **Team** | L'équipe dont appartient le joueur |
| **Nationality** | Leurs équipes nationales |
| **Matches** | Le nombre de matches joués lors de la saison 2016-2017 |
| **Goals** | Le nombre de buts marqués (Pour les gardiens sera contre) |
| **Selection** | Le nombre de fois de sélection à l'équipe nationale |
| **Minutes** | Le nombre de minutes jouées lors de la saison 2016-2017 |
| **Performance** | Elle se calcule à base de multiples indices (offensive, défensive...) |
| **Titres** | Le nombre de championnats remportés |
| **yellowCard** | Le nombre de carton jaunes durant les matches joués |
| **redCard** | Le nombre de rouge jaunes durant les matches joués |
| **Gender** | Le sexe des joueurs, c'est évidement masculin pour tous |
| **isSelected** | **Yes** si le joueur est nommé, Sinon **No**. |

Certaines variables seront sujet de traitement après, car soit il ne contribue pas à la nomination des joueurs, soit qu'il n'influence pas sur la prédiction qu'on souhaite réaliser. Cette tâche sera traitée en ce qui vient.

# MÉTHODOLOGIE CRISP
## APPLCATION AU PROJET DATA MINING

## 3. Préparation des données

### ▪ Collection, Nettoyage et Réduction des données :

Pour la collection des données, nous avons essayé de trouver une API qui nous permettra d'extraire les données détaillées des joueurs de la saison 2016-2017, la chose qui était impossible (à part les services payants), ce qui nous a obligé à collecter les données nous-même à partir de ces trois sites web fournissant des statistiques réelles sur les joueurs :

| | |
|---|---|
| Squawka | **Squawka.com** est le site web de la société Squawka de médias numériques axée sur l'utilisation de statistiques et de données à savoir pour alimenter le contenu pour les fans du football. |
| WhoScored.com | **WhoScored.com** vous permet d'accéder aux scores des matchs en direct, aux résultats finaux et aux notes des joueurs évoluant dans les meilleurs championnats et compétitions. |
| france football | **Francefootball.fr** est la plateforme de la revue française France Football qui permet de suivre l'actualité; sportive du football, les résultats, les classements, les transferts de foot. |

Le résultat de ces données après avoir les collecter, les organiser dans un fichier « Players Data.csv » est la base de données ci-dessous qu'on exploitera par la suite pour élaborer notre modèle de prédiction sous R & sous Weka.

# MÉTHODOLOGIE CRISP

## APPLCATION AU PROJET DATA MINING

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Player | Age | Confederation | Position | Team | Nationality | Matches | Goals | Selection | Minutes | Performance | Titres | yellewCard | redCard | Gender | IsSelecte |
| 2 | Neymar | 25 | UEFA | attaquant | PSG | Brésil | 45 | 20 | 83 | 3971 | 1570 | 2 | 15 | 1 | Male | Ye |
| 3 | Marcelo | 29 | UEFA | Défenseur | Real Madrid | Brésil | 48 | 7 | 48 | 3855 | 436 | 5 | 5 | 0 | Male | Ye |
| 4 | Philippe Coutinho | 25 | UEFA | Milieu | Liverpool | Brésil | 36 | 14 | 32 | 2532 | 1243 | 0 | 2 | 0 | Male | Ye |
| 5 | Paulo Dybala | 23 | UEFA | attaquant | Juventus Turin | Argentine | 50 | 20 | 12 | 3497 | 1022 | 1 | 3 | 0 | Male | Ye |
| 6 | Lionel Messi | 30 | UEFA | attaquant | FC Barcelone | Argentine | 57 | 58 | 114 | 4787 | 2480 | 2 | 9 | 0 | Male | Ye |
| 7 | Nabil Fekir | 24 | UEFA | attaquant | Lyon | France | 8 | 13 | 11 | 539 | 510 | 0 | 5 | 0 | Male | N |
| 8 | Mohamed Salah | 25 | UEFA | Milieu | Liverpool | Egypt | 31 | 17 | 78 | 2476 | 990 | 0 | 0 | 0 | Male | N |
| 9 | Houssem Anouar | 19 | UEFA | Milieu | Lyon | France | 2 | 3 | 2 | 1059 | -2 | 0 | 0 | 0 | Male | N |
| 10 | Eden Hazard | 26 | UEFA | Mileu | Chelsea | Belgique | 43 | 17 | 82 | 3375 | 1845 | 1 | 3 | 0 | Male | Ye |
| 11 | Robert Lewandowski | 29 | UEFA | attaquant | Bayern Munich | Pologne | 47 | 43 | 91 | 4021 | 1289 | 1 | 5 | 0 | Male | Ye |
| 12 | Harry Kane | 24 | UEFA | attaquant | Tottenham | Angleterre | 38 | 35 | 33 | 3154 | 1210 | 0 | 4 | 0 | Male | Ye |
| 13 | Toni Kroos | 27 | UEFA | Milieu | Real Madrid | Allemagne | 46 | 4 | 80 | 3978 | 1126 | 5 | 10 | 0 | Male | Ye |
| 14 | Mats Hummels | 28 | UEFA | Défenseur | Bayern Munich | Allemagne | 42 | 3 | 62 | 3270 | 1039 | 1 | 6 | 0 | Male | Ye |
| 15 | Raheem Sterling | 26 | UEFA | Milieu | Manchester City | Angleterre | 33 | 14 | 35 | 2513 | 651 | 0 | 7 | 0 | Male | N |
| 16 | Sergio Aguero | 29 | UEFA | attaquant | Manchester City | Argentine | 31 | 20 | 83 | 2409 | 908 | 0 | 4 | 1 | Male | N |
| 17 | Alexis Sanchez | 29 | UEFA | attaquant | Arsenal | Chili | 37 | 25 | 119 | 3266 | 1492 | 0 | 6 | 0 | Male | N |
| 18 | Romelu Lukaku | 24 | UEFA | attaquant | Manchester United | Belgique | 38 | 24 | 61 | 3217 | 957 | 1 | 3 | 0 | Male | N |
| 19 | Jamie Vardy | 31 | UEFA | attaquant | Leicester City | Angleterre | 35 | 12 | 19 | 2801 | 378 | 0 | 2 | 1 | Male | N |
| 20 | Radamel Falcao | 31 | UEFA | attaquant | AS Monaco | Colombie | 43 | 30 | 62 | 2937 | 723 | 1 | 6 | 0 | Male | Ye |
| 21 | Luis Suarez | 30 | UEFA | attaquant | FC Barcelone | Uruguay | 56 | 40 | 88 | 4698 | 1387 | 2 | 13 | 0 | Male | Ye |
| 22 | Edinson Cavani | 30 | UEFA | attaquant | PSG | Uruguay | 53 | 50 | 95 | 4286 | 1289 | 2 | 6 | 0 | Male | Ye |
| 23 | N'Golo Kanté | 26 | UEFA | Milieu | Chelsea | France | 41 | 2 | 20 | 3526 | 701 | 1 | 11 | 0 | Male | Ye |
| 24 | Antoine Griezmann | 26 | UEFA | attaquant | Atlético Madrid | France | 53 | 26 | 54 | 4559 | 1081 | 0 | 4 | 0 | Male | Ye |
| 25 | Karim Benzema | 29 | UEFA | attaquant | Real Madrid | France | 46 | 17 | 81 | 3038 | 679 | 5 | 0 | 0 | Male | Ye |
| 26 | Kylian Mbappé | 18 | UEFA | attaquant | PSG | France | 44 | 26 | 18 | 2633 | 613 | 2 | 2 | 0 | Male | Ye |

*La base de données collectée Players Data.csv*

Ce fichier « **Players Data.cvs** » contient des données réelles pour la saison 2016-2017sur 67 joueurs dont 30 joueurs sont nommée pour la FIFA Ballon d'or et le reste non.

Par la suite on élaborera notre modèle aidant à la prédiction de réponse « oui ou non » pour des nouvelles valeurs entrées pour la saison 2017-2018.

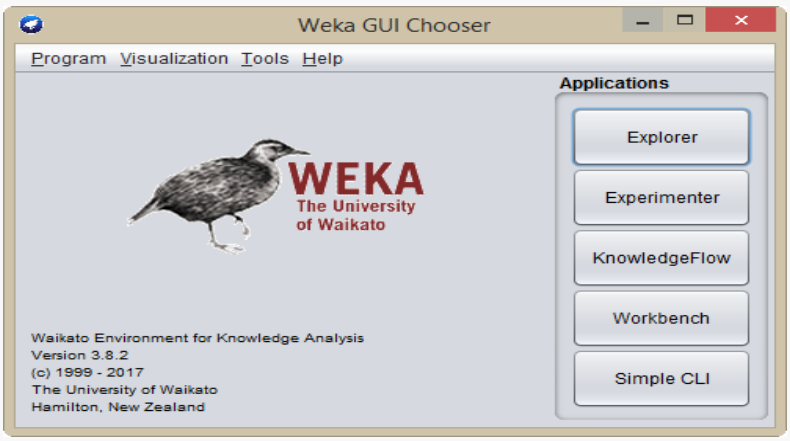
# MÉTHODOLOGIE CRISP

## APPLCATION AU PROJET DATA MINING

## 4. Modélisation et Evaluation sous Weka

### ▪ Weka : C'est quoi ?

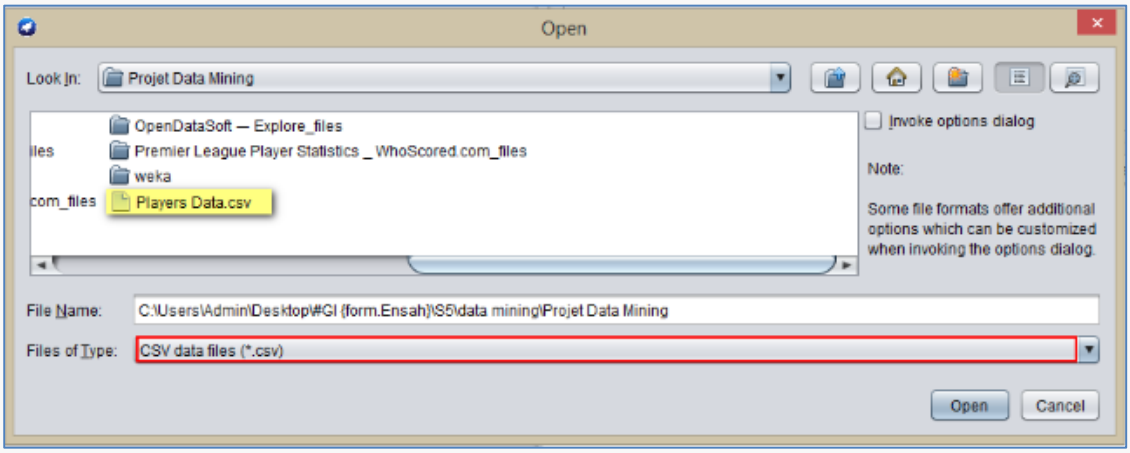
« **Environnement Waikato pour l'analyse de connaissances** » est une suite de logiciels d'apprentissage automatique. Écrite en Java, développée à l'université de Waikato en Nouvelle-Zélande. Weka est un logiciel libre disponible sous la Licence publique générale GNU, portable car il est entièrement implémenté en Java et donc fonctionne sur quasiment toutes les plateformes modernes, et en particulier sur quasiment tous les systèmes d'exploitation actuels, et contient une collection complète de préprocesseurs de données et de techniques de modélisation.



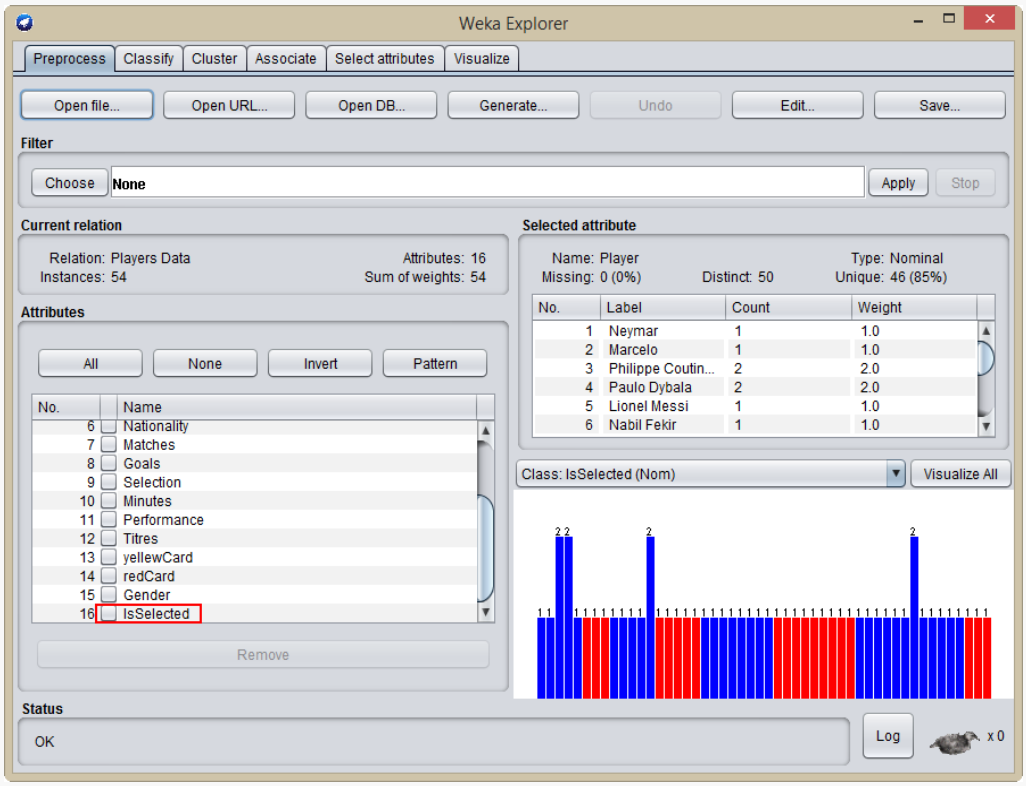

*Weka GUI Chooser*

### ▪ Chargement des données « Players Data.csv » sur Weka :

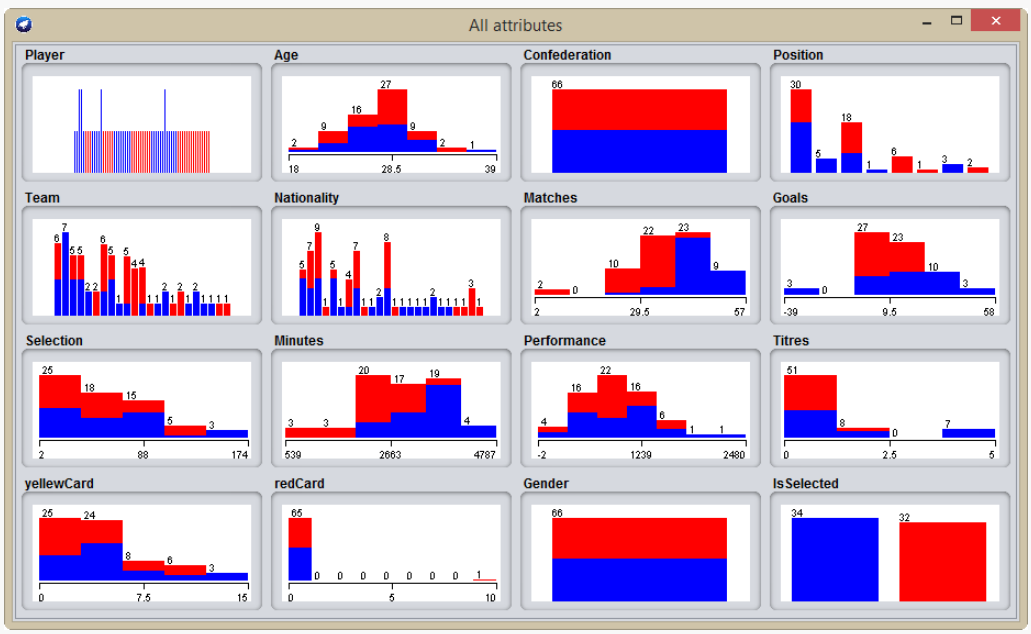
Après avoir installé Weka parfaitement, on passe au chargement des données collectées et stockées dans la base de données « **Players Data.csv** ». Dès qu'elles sont chargées sur l'environnement de travail Weka, les données sont maintenant à être manipulées, mais avant ceci on affiche une visualisation globale résumant les variables explicatives du Dataset chargé.




*Chargement des données sur Weka*




*Données sur les joueurs chargées sur Weka*




*Visualisation graphique des attributs*

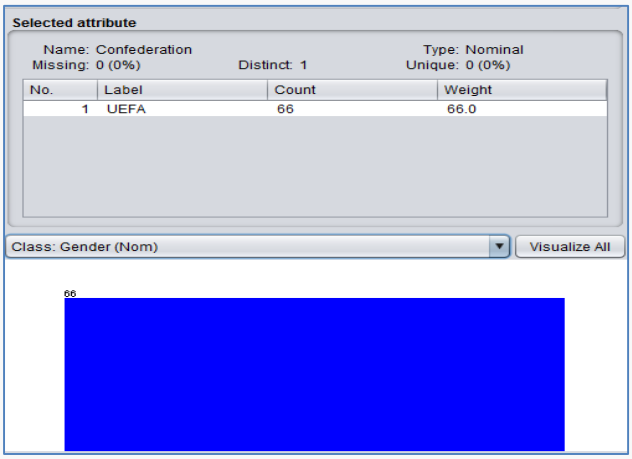
# MÉTHODOLOGIE CRISP

## APPLCATION AU PROJET DATA MINING

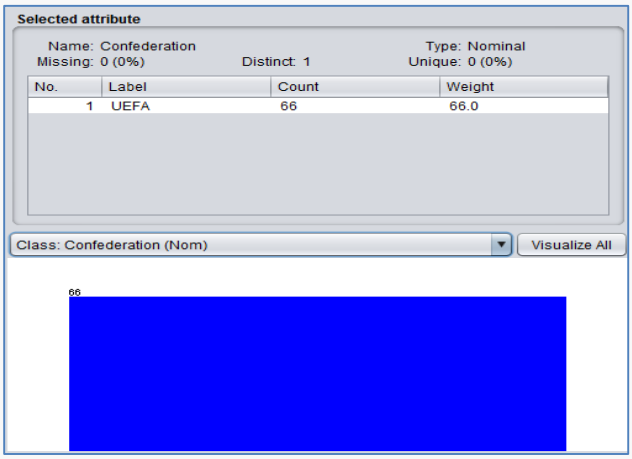
### ▪ Prétraitement des données :

C'est une étape primordiale dans chaque projet Data Mining car il permet d'augmenter la qualité de prédiction, ainsi qu'il nous garantit l'élaboration d'un modèle parfait qui peut être déployé sans risque.

▪ Pour ceci on supprime tout d'abord, les variables invariantes dans notre Dataset, qui sont « Confederation » et « Gender »

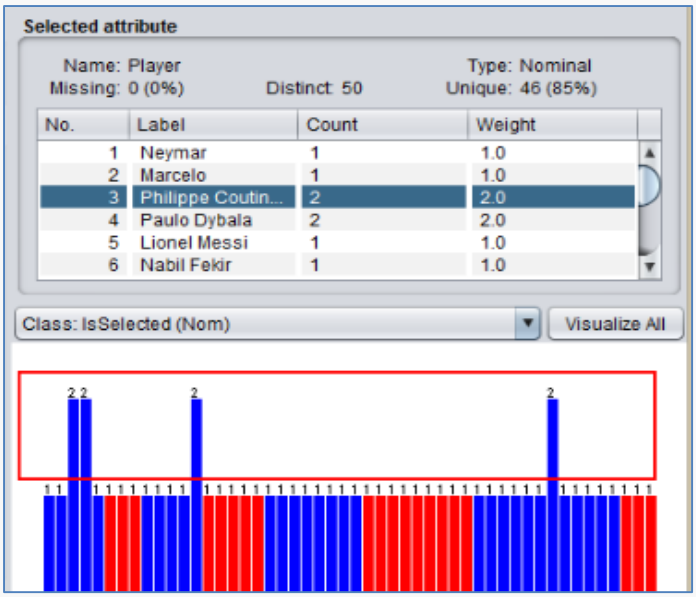



*la variable invariante « Confederation »*

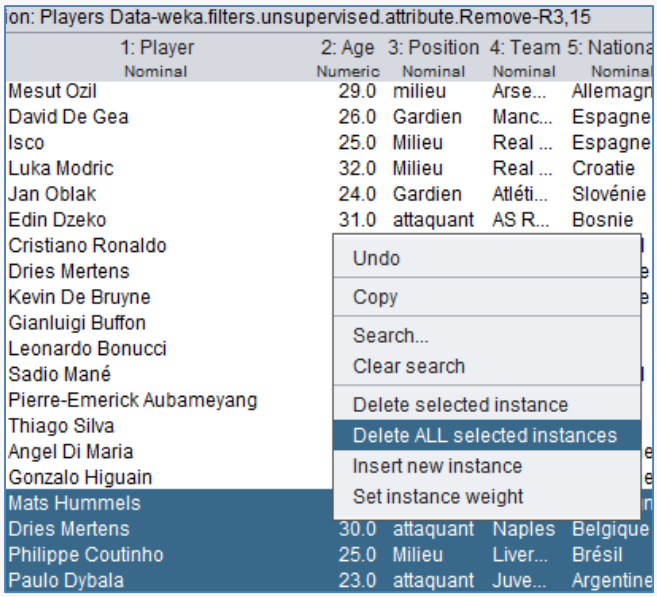



*la variable invariante « Gender »*

▪ Puis, on passe à la suppression des doublons dans l'onglet« **Edit** ». Dans notre cas et comme le montre les figures ci-dessous, 4 individus sont des doublons :



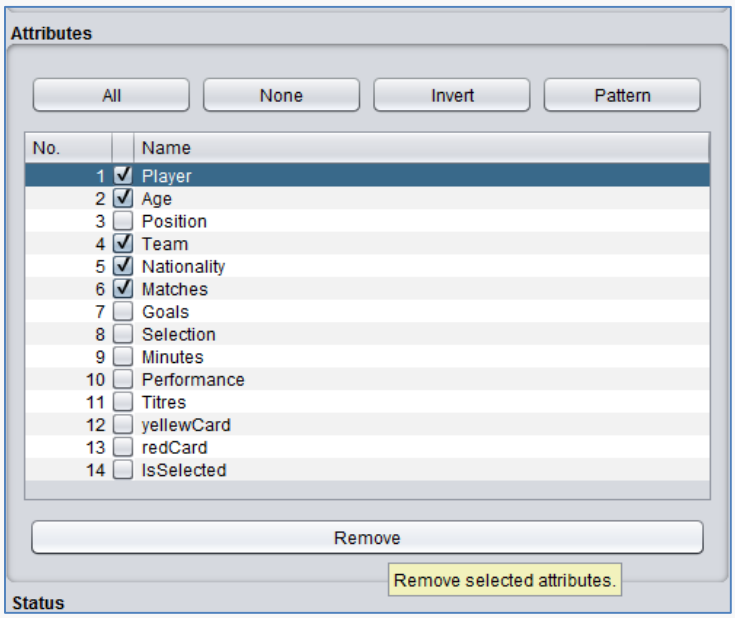

*La détection de 4 joueurs doublons*




*La suppression des doublons*

# MÉTHODOLOGIE CRISP
## APPLCATION AU PROJET DATA MINING

- Finalement, on supprime les variables qui n'influencent pas sur la prédiction, comme le nom du joueur, l'équipe et la nationalité.



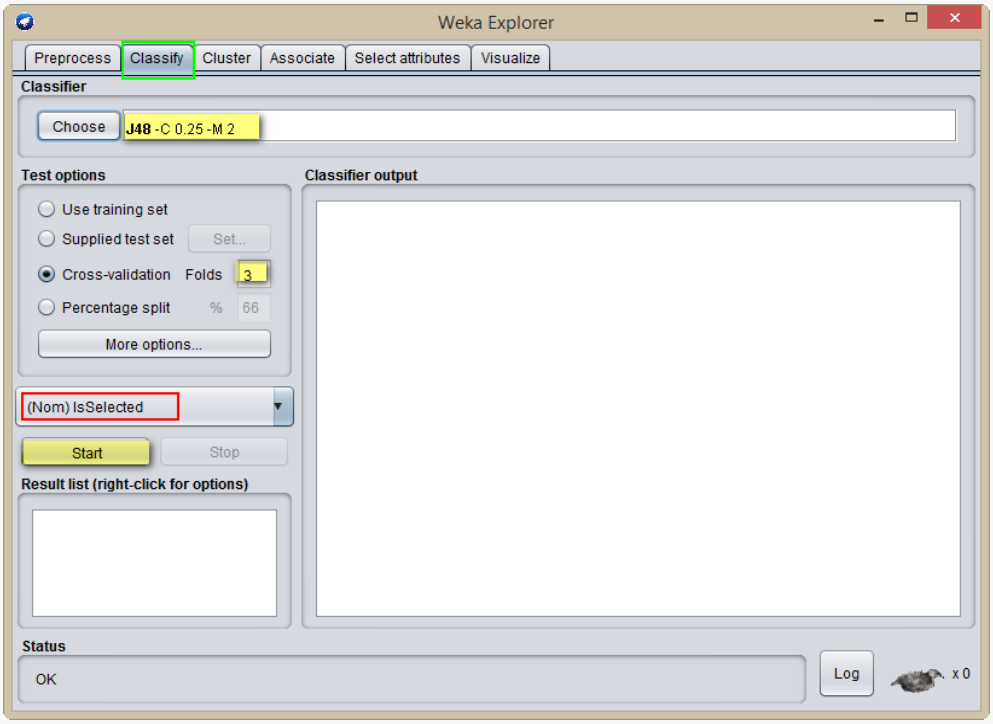

*Suppression des variables non utiles*

- Dans notre cas les valeurs des variables sont tous existantes et non atypique, et donc nous n'avons pas besoins de faire ce genre de traitement.

### Choix de méthode et algorithmes utilisés :

Une fois que les données sont bien traitées, on passera à l'élaboration de notre modèle. Nous avons choisi « **Les arbres de décision** » comme étant méthode de prédiction, qui est un outil d'aide à la décision et à l'exploration de données. Il permet de modéliser simplement, graphiquement et rapidement un phénomène mesuré plus ou moins complexe. Sa lisibilité, sa rapidité d'exécution et le peu d'hypothèses nécessaires à priori expliquent sa popularité actuelle.

Pour établir l'arbre de décision on se déplace sur l'onglet **Classify** (qui désigne prédire et non pas classifier), et on choisit l'option Use training set de Test options comme suit :

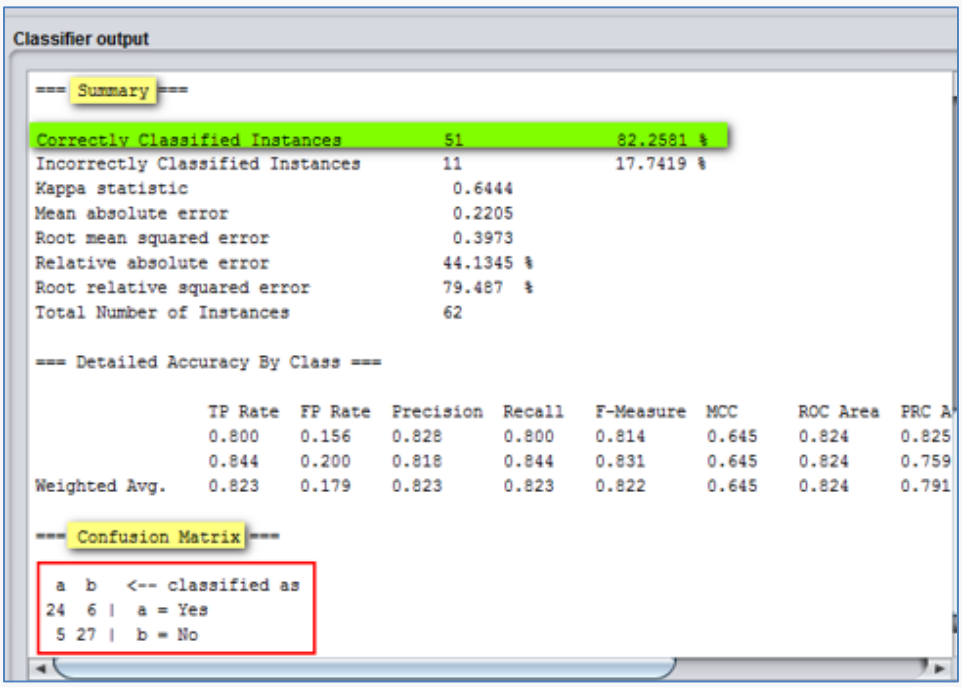



*Suppression des variables non utiles*

La zone Test options permet de choisir de quelle façon l'évaluation des performances du modèle appris se fera.

- L'option **Use training set** utilise l'ensemble d'entraînement pour cette évaluation.
- L'option **Supplied test set** va utiliser un autre fichier.
- Lorsque l'option **Cross-validation** est sélectionnée, l'ensemble d'apprentissage est coupé en 3 (si Folds vaut 3). L'algorithme va apprendre 3 fois sur 2 parties et le modèle sera évalué sur le troisième restant. Les 3 évaluations sont alors combinées.

Avec l'option **Percentage split**, c'est un pourcentage de l'ensemble d'apprentissage qui servira à l'apprentissage et l'autre à l'évaluation. Ensuite, cliquer sur le bouton **Choose** de Classifier pour choisir un algorithme parmi ceux proposés par WEKA. Dans notre cas nous allons utiliser l'algorithme J48, on clique finalement sur « **Start** ».



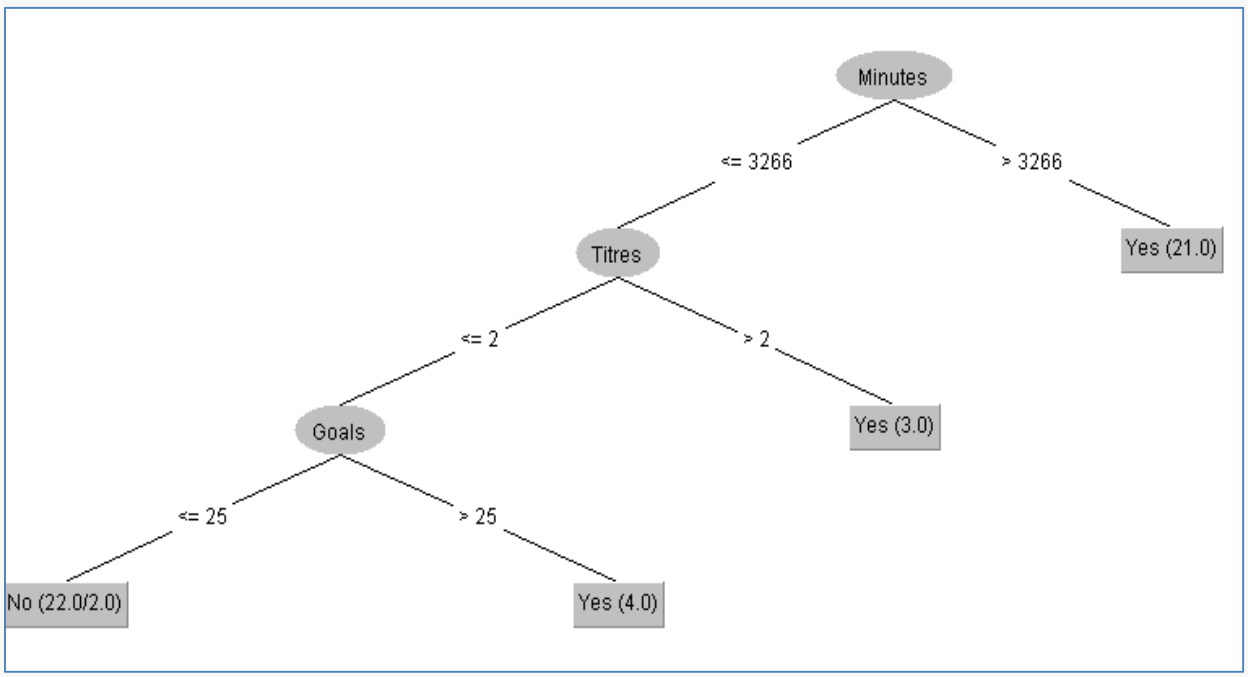

*Les informations sur l'arbre de décision construit*

# MÉTHODOLOGIE CRISP

## APPLCATION AU PROJET DATA MINING

Après la génération de l'arbre, nous mettons en œuvre le code équivalent à l'arbre (Figure 9), et nous l'appliquons après sur la base de données des joueurs de l'année courante, afin de recevoir la liste des joueurs qui pourraient être sélectionnés pour gagner le Ballon d'or.

Pour afficher l'arbre de décision, cliquer droit dans la partie **Result list** (right-click for options), et on choisit l'option **Visualize tree**. L'arbre de décision s'affiche ressemblant à ceci :




*L'arbre de décision obtenu via l'algorithme J48*

### ▪ Evaluation et Déploiement du modèle :

Une fois notre arbre est construit, on l'applique sur un échantillon d'individu avec les mêmes valeurs des 12 premiers joueurs de la source de données originales et on obtient les mêmes résultats très satisfaisants. Les résultats d'exploitation du modèle établi est illustré dans la figure ci-dessous.

# MÉTHODOLOGIE CRISP

## APPLICATION AU PROJET DATA MINING

```
=== Predictions on test set ===

inst#,actual,predicted,error,prediction
1,1:?,1:Yes,,1
2,1:?,1:Yes,,1
3,1:?,1:Yes,,1
4,1:?,1:Yes,,1
5,1:?,1:Yes,,1
6,1:?,2:No,,0.909
7,1:?,2:No,,0.909
8,1:?,2:No,,0.909
9,1:?,1:Yes,,1
10,1:?,1:Yes,,1
11,1:?,1:Yes,,1
12,1:?,1:Yes,,1
```

*L'arbre de décision appliqué sur un échantillon de joueurs*

### ▪ Conclusion :

Pour la mise en œuvre d'un tel projet sous Weka nous avons choisi d'adopter un arbre décisionnel et nous avons commencé par le nettoyage des données. Ensuite, nous avons générer notre arbre décisionnel afin d'aborder enfin la phase de la réalisation qui consiste à la mise en œuvre de l'application proposée dans le projet « **VeilleBox World Cup 2018**».

# MÉTHODOLOGIE CRISP
## APPLCATION AU PROJET DATA MINING

## 5. Modélisation et Evaluation sous R

### ▪ R : C'est quoi ?

*Logo de R*

R est un langage informatique dédié aux statistiques et à la science des données. L'implémentation la plus connue du langage R est le logiciel GNU R.

Le langage R est dérivé du langage S développé par John Chambers et ses collègues au sein des laboratoires Bell. GNU R est un logiciel libre distribué selon les termes de la licence GNU GPL et disponible sous GNU/Linux, FreeBSD, NetBSD, OpenBSD, Mac OS, X et Windows.

### ▪ Chargement des données « Players Data.csv » sur R :

Comme c'était le cas pour Weka, on commence par le chargement des mêmes données collectées, mais cette fois stockées dans la base de données de type Excel « **Players Data.xlsx** ». Ceci est fait grâce au package « xlsx » qui fournit la fonction read.xlsx (). Ces données chargées seront mémorisé dans un Data Frame qu'on nommera **Ballon.**

```
> library(xlsx)
> Ballon<-read.xlsx(file="BallonOr.xlsx" ,1, header=TRUE)
```

```
> Ballon
                Player Age Confederation   Position          Team Nationality Matches Goals Selection Minutes Performance Titres yellewCard redCard Gender
1               Neymar  25          UEFA  attaquant           PSG      BrÃ©sil      45    20        83    3971        1570      2         15       1   Male
2              Marcelo  29          UEFA DÃ©fenseur   Real Madrid      BrÃ©sil      48     7        48    3855         436      5          5       0   Male
3     Philippe Coutinho  25          UEFA     Milieu     Liverpool      BrÃ©sil      36    14        32    2532        1243      0          2       0   Male
4         Paulo Dybala  23          UEFA  attaquant  Juventus Turin   Argentine      50    20        12    3497        1022      1          3       0   Male
5         Lionel Messi  30          UEFA  attaquant    FC Barcelone   Argentine      57    58       114    4787        2480      2          9       0   Male
6          Nabil Fekir  24          UEFA  attaquant           Lyon      France       8    13        11     539         510      0          5       0   Male
7       Mohamed Salah  25          UEFA     Milieu     Liverpool       Egypt      31    17        78    2476         990      0          0       0   Male
8       Houssem Anouar  19          UEFA     Milieu           Lyon      France       2     3         2    1059          -2      0          0       0   Male
9         Eden Hazard  26          UEFA     Milieu        Chelsea     Belgique      43    17        82    3375        1845      1          3       0   Male
10  Robert Lewandowski  29          UEFA  attaquant   Bayern Munich      Pologne      47    43        91    4021        1289      1          5       0   Male
11         Harry Kane  24          UEFA  attaquant        Tottenham  Angleterre      38    35        33    3154        1210      0          4       0   Male
12          Toni Kroos  27          UEFA     Milieu   Real Madrid     Allemagne      46     4        80    3978        1126      5         10       0   Male
13        Mats Hummels  28          UEFA DÃ©fenseur   Bayern Munich     Allemagne      42     3        62    3270        1039      1          6       0   Male
14      Raheem Sterling  26          UEFA     Milieu Manchester City  Angleterre      33    14        35    2513         651      0          7       0   Male
15        Sergio Aguero  29          UEFA  attaquant Manchester City   Argentine      31    20        83    2409         908      0          4       1   Male
16      Alexis Sanchez  29          UEFA  attaquant         Arsenal       Chili      37    25       119    3266        1492      0          6       0   Male
17       Romelu Lukaku  24          UEFA  attaquant Manchester United     Belgique      38    24        61    3217         957      1          3       0   Male
18         Jamie Vardy  31          UEFA  attaquant    Leicester City  Angleterre      35    12        19    2801         378      0          2       1   Male
19      Radamel Falcao  31          UEFA  attaquant        AS Monaco      Colombie      43    30        62    2937         723      1          6       0   Male
20        Luis Suarez  30          UEFA  attaquant    FC Barcelone       Uruguay      56    40        88    4698        1387      2         13       0   Male
21      Edinson Cavani  30          UEFA  attaquant           PSG       Uruguay      53    50        95    4286        1289      2          6       0   Male
22      N'Golo KantÃ©  26          UEFA     Milieu        Chelsea      France      41     2        20    3526         701      1         11       0   Male
23    Antoine Griezmann  26          UEFA  attaquant  AtlÃ©tico Madrid      France      53    26        54    4559        1081      0          4       0   Male
24      Karim Benzema  29          UEFA  attaquant   Real Madrid      France      46    17        81    3038         679      5          0       0   Male
25      Kylian MbappÃ©  18          UEFA  attaquant           PSG      France      44    26        18    2633         613      2          2       0   Male
26       Sergio Ramos  31          UEFA DÃ©fenseur   Real Madrid     Espagne      44    10       149    3942         670      5         13       0   Male
27      Dimitri Payet  30          UEFA     Milieu     West Ham      France      33     3        36    2759        1184      0          2       0   Male
28      Julian Draxler  24          UEFA     Milieu           PSG     Allemagne      30     4        40    2063         760      2          2       0   Male
29    Zlatan Ibrahimovic  36          UEFA  attaquant Manchester United      Sweden      28    17       116    2442         875      1          7       0   Male
30       Alvaro Morata  25          UEFA  attaquant        Chelsea     Espagne      26    15        23    1341         514      1          8       0   Male
31      Nicolas Otamendi  29          UEFA  Defenseur  Manchester City   Argentine      30     1        51    2592         959      0          9       0   Male
32    Shkodran Mustafi  25          UEFA  Defenseur         Arsenal     Allemagne      26     2        20    2274         520      0         11       0   Male
33        Gareth Barry  36          UEFA     Milieu         Everton  Angleterre      33     2        53    2115         300      0         10       0   Male
```

*Chargement des données sur le Data Frame Ballon*

# MÉTHODOLOGIE CRISP

## APPLCATION AU PROJET DATA MINING

Pour avoir une idée globale sur les données importées, on affiche un résumé statistique sur les variables via la fonction summary ().

```
> summary(Ballon)
                Player         Age          Confederation      Position                 Team
 Alexis Sanchez     : 1   Min.   :18.00   UEFA:50       attaquant:26   Real Madrid     : 7
 Alvaro Morata      : 1   1st Qu.:25.00   NA's: 1       Defenseur: 7   PSG             : 6
 Angel Di Maria     : 1   Median :29.00                 Gardien  : 3   Chelsea         : 4
 Antoine Griezmann  : 1   Mean   :27.94                 Milieu   :14   Manchester City: 4
 Cristiano Ronaldo  : 1   3rd Qu.:30.00                 NA's     : 1   Arsenal         : 3
 (Other)            :45   Max.   :39.00                                (Other)         :26
 NA's               : 1   NA's   :1                                    NA's            : 1
    Minutes      Performance         Titres      yellewCard       redCard      Gender    IsSe
 Min.   : 539   Min.   :  -2.0   Min.   :0.0   Min.   : 0.00   Min.   :0.0   Male:50   No
 1st Qu.:2450   1st Qu.: 630.5   1st Qu.:0.0   1st Qu.: 2.00   1st Qu.:0.0   NA's: 1   Yes
 Median :3154   Median : 922.0   Median :1.0   Median : 5.00   Median :0.0             NA's
 Mean   :3094   Mean   : 932.2   Mean   :1.3   Mean   : 5.18   Mean   :0.1
 3rd Qu.:3896   3rd Qu.:1203.5   3rd Qu.:2.0   3rd Qu.: 7.00   3rd Qu.:0.0
 Max.   :4787   Max.   :2480.0   Max.   :5.0   Max.   :15.00   Max.   :1.0
 NA's   :1      NA's   :1        NA's   :1     NA's   :1       NA's   :1
```

*Résumé statistique des données chargées*

### ▪ Prétraitement des données :

Le seul prétraitement à faire dans notre cas c'est la suppression des variables invariantes (Gender, Confederation), qui n'ont pas une influence sur la performance de notre modèle génère par la suite.

```
> Ballon<-Ballon[,-3]
> Ballon<-Ballon[,-14]
```

```
> summary(Ballon)
                Player         Age          Position              Team       Nationality    Matches          Goals          Selection        Minutes
 Alexis Sanchez     : 1   Min.   :18.00   attaquant:26   Real Madrid     : 7   France    : 7   Min.   : 2.00   Min.   :-39.00   Min.   : 2.00   Min.   : 539
 Alvaro Morata      : 1   1st Qu.:25.00   Defenseur: 7   PSG             : 6   Argentine : 6   1st Qu.:33.00   1st Qu.:  4.00   1st Qu.: 32.25   1st Qu.:2450
 Angel Di Maria     : 1   Median :29.00   Gardien  : 3   Chelsea         : 4   Allemagne : 5   Median :41.50   Median : 14.00   Median : 61.50   Median :3154
 Antoine Griezmann  : 1   Mean   :27.94   Milieu   :14   Manchester City: 4   Espagne   : 5   Mean   :38.78   Mean   : 14.70   Mean   : 62.70   Mean   :3094
 Cristiano Ronaldo  : 1   3rd Qu.:30.00   NA's     : 1   Arsenal         : 3   Angleterre: 4   3rd Qu.:45.75   3rd Qu.: 24.75   3rd Qu.: 86.00   3rd Qu.:3896
 (Other)            :45   Max.   :39.00                  (Other)         :26   (Other)   :23   Max.   :57.00   Max.   : 58.00   Max.   :174.00   Max.   :4787
 NA's               : 1   NA's   :1                      NA's            : 1   NA's      : 1   NA's   :1       NA's   :1        NA's   :1        NA's   :1
  Performance       Titres      yellewCard       redCard      IsSelected    NA.
 Min.   :  -2.0   Min.   :0.0   Min.   : 0.00   Min.   :0.0   No  :20   Mode:logical
 1st Qu.: 630.5   1st Qu.:0.0   1st Qu.: 2.00   1st Qu.:0.0   Yes :30   NA's:51
 Median : 922.0   Median :1.0   Median : 5.00   Median :0.0   NA's: 1
 Mean   : 932.2   Mean   :1.3   Mean   : 5.18   Mean   :0.1
 3rd Qu.:1203.5   3rd Qu.:2.0   3rd Qu.: 7.00   3rd Qu.:0.0
 Max.   :2480.0   Max.   :5.0   Max.   :15.00   Max.   :1.0
 NA's   :1        NA's   :1     NA's   :1       NA's   :1
```

*Résumé statistique des données chargées après la suppression des invariants*

# MÉTHODOLOGIE CRISP

## APPLCATION AU PROJET DATA MINING

### ▪ Choix de méthode et algorithmes utilisés :

Le modèle de notre solution implémentée sera base sur les réseaux bayésiens qui sont des modèles graphique probabilistes représentant des variables aléatoires sous la forme d'un graphe orienté acyclique. Intuitivement, ils sont à la fois :

- **Des modèles de représentation des connaissances ;**
- **Des « machines à calculer » les probabilités conditionnelles ;**
- **Une base pour des Systèmes d'aide à la décision ;**

La démarche de conception du modèle consiste sur le découpage de notre base de données sur 3 tiers, 2/3 exploités pour l'apprentissage et le reste pour tester le du modèle établi. Enfin, pour bien valider, nous allons importer une nouvelle base de données et on applique sur elle notre modèle.

### ▪ Le découpage des données en 2 bases : Apprentissage et Validation

```
> ApprIndice<-sample(1:nrow(Ballon), floor(nrow(Ballon)*2/3), replace=FALSE)
> BallonAppr<-Ballon[ApprIndice,]
> BallonVal<-Ballon[-ApprIndice,]
>
```

*Découpage des données en base d'apprentissage et base de test*

### ▪ La génération du modèle en appliquant les réseaux bayésiens

La génération du modèle se fait via la fonction **naiveBayes()** du package « **naiveBayes** », et on se basant sur la base d'apprentissage **BallonAppr.**

```
> Model<-naiveBayes(IsSelected~., data=BallonAppr)
> Model

Naive Bayes Classifier for Discrete Predictors

Call:
naiveBayes.default(x = X, y = Y, laplace = laplace)

A-priori probabilities:
Y
       No       Yes
0.3939394 0.6060606

Conditional probabilities:
     Player
Y     Alexis Sanchez Alvaro Morata Angel Di Maria Antoine Griezmann  Cristiano Ronaldo David De Gea  Diego Costa Dimitr
  No      0.07692308    0.07692308    0.00000000        0.00000000       0.00000000    0.00000000 0.07692308    0.0
  Yes     0.00000000    0.00000000    0.00000000        0.00000000       0.05000000    0.05000000 0.00000000    0.0
     Player
Y     Edinson Cavani  Gareth Barry Gianluigi Buffon Gonzalo Higuain Harry Kane  Houssem Anouar       Isco Jamie Vardy
  No      0.00000000   0.07692308       0.00000000      0.07692308  0.00000000      0.00000000 0.00000000  0.00000000 0
  Yes     0.05000000   0.00000000       0.00000000      0.00000000  0.05000000      0.00000000 0.05000000  0.00000000 0
     Player
Y     Kevin De Bruyne  Kylian MbappÃ© Leonardo Bonucci  Lionel Messi  Luis Suarez Luka Modric     Marcelo Mats Hummels
  No       0.00000000     0.00000000       0.00000000    0.00000000  0.00000000   0.00000000 0.00000000   0.00000000
  Yes      0.05000000     0.00000000       0.05000000    0.00000000  0.05000000   0.05000000 0.00000000   0.00000000
     Player
```

*Génération du réseau bayésien via la fonction naiveBayese*

# MÉTHODOLOGIE CRISP

## APPLCATION AU PROJET DATA MINING

Aprés que notre modèle est bien génère, nous allons tester la fiabilité de ce dernier en l'appliquant sur la base de test (le 1/3 restant). Le test du modèle sera applique sur nos données BallonVal.

```
> ModelTest<-predict(object=Model, newdata=BallonVal)
> ModelTest
 [1] Yes No  Yes No  Yes No  No  No  Yes No  No  No  No  Yes Yes No  No
Levels: No Yes
```

*Application du modèle sur la base de test*

Pour vérifier la validité des résultats, nous allons construire notre matrice de confusion afin de tester l'erreur de prédiction. Nous remarquons que le taux d'erreur est petit, ce qui montre que le modèle est assez robuste.

```
> MC<-table(BallonTest$IsSelected, BallonTest$ModelTest)
> MC

     No Yes
  No  7   0
  Yes 4   6
> ErreurPrediction<-(MC[1,2]+MC[2,1])/sum(MC)
> ErreurPrediction
[1] 0.2352941
>
```

*La matrice de confusion et l'erreur de prédiction*

### ▪ Evaluation et Déploiement du modèle :

A la fin de chaque projet Data Mining, il fallait valider le modèle. Pour cela, nous allons importer de nouveaux données « et appliquer notre modèle RB.

```
> Validation<-read.xlsx(file="Validation.xlsx", 1, header=TRUE)
> Validation
          Player Age Confederation  Position            Team Nationality Matches Goals Selection Minutes Performance Titres yellewCard redCard Gender IsSelected
1 Cesar Azpilicueta  28         UEFA Defenseur         Chelsea     Espagne      38     1        20    3420         990      1          4       0   Male         NO
2      David Silva  32         UEFA    Milieu Manchester City     Espagne      34     4       118    2760         924     NA          6       0   Male         NO
3       Joel Matip  26         UEFA Defenseur       Liverpool   Allemagne      29     1        27    2462         882     NA          3       0   Male         NO
4       David Luiz  30         UEFA Defenseur         Chelsea      Bresil      33     1        56    2954         839     NA          6       0   Male         NO
5    Ander Herrera  28         UEFA    Milieu Manchester United     Espagne      31     1         2    2468         817     NA          6       1   Male         NO
6    Idrissa Gueye  28         UEFA    Milieu         Everton     Senegal      33     1        46    2681         694     NA         11       0   Male         NO
```

*Chargement des nouvelles données à prédire*

# MÉTHODOLOGIE CRISP

## APPLCATION AU PROJET DATA MINING

```
> ValidationTest<-predict(object=Model, newdata=Validation)
> ValidationTest
[1] No No No No No No
Levels: No Yes
> Validation<-cbind(Validation, ValidationTest)
> Validation
          Player Age Confederation  Position            Team Nationality Matches Goals Selection Minutes Performance Titres yellewCard redCard Gender IsSelected
1 Cesar Azpilicueta  28          UEFA Defenseur         Chelsea     Espagne      38     1        20    3420         990      1          4       0   Male         NO
2       David Silva  32          UEFA    Milieu Manchester City     Espagne      34     4       118    2760         924     NA          6       0   Male         NO
3        Joel Matip  26          UEFA Defenseur       Liverpool   Allemagne      29     1        27    2462         882     NA          3       0   Male         NO
4        David Luiz  30          UEFA Defenseur         Chelsea      Bresil      33     1        56    2954         839     NA          6       0   Male         NO
5     Ander Herrera  28          UEFA    Milieu Manchester United     Espagne      31     1         2    2468         817     NA          6       1   Male         NO
6     Idrissa Gueye  28          UEFA    Milieu         Everton     Senegal      33     1        46    2681         694     NA         11       0   Male         NO
  ValidationTest
1             No
2             No
3             No
4             No
5             No
6             No
>
```

*Les résultats obtenus après l'application du modèle établi*

On constate que notre modèle a bien prédit les résultats (qu'on connait auparavant). Et donc notre solution peut être déployée sans problème car il conduira à des bons résultats.

### ▪ Conclusion :

Pour étudier notre cas qui sert à prédire les joueurs nommées pour la FFIA ballon d'or, nous avons encore implémenté cet apprentissage supervisé sous R fournissant des multiples modèles de prédiction dont nous avons choisi les réseaux bayésiens connus par leur performance et par leurs bons résultats.

# CONCLUSION GÉNÉRALE

## LE PROCHAIN BALLON D'OR, ON LE CONNAIT DÉJA


Le Data Mining est un domaine pluridisciplinaire permettant, à partir d'une très importante quantité de données brutes, d'en extraire des informations cachées, pertinentes et inconnues auparavant en vue d'une utilisation industrielle ou opérationnelle de ce savoir.

Elle permet de faire la classification automatique supervisée « Prédiction » qui consiste à examiner les caractéristiques d'un objet nouvellement présenté afin de l'affecter à une classe d'un ensemble prédéfini.

On peut l'appliquer aussi pour la classification automatique non supervisée qui vise à identifier des ensembles d'éléments qui partagent certaines similarités. Notre projet intitulé « Prédiction de joueurs nommés à la FIFA ballon d'or » est un problème d'apprentissage supervisé, ce qui nous conduira à appliquer des modèles dédiés à la résolution de ce problème : Arbres de décision et Réseau Bayésiens implémentés respectivement sous Weka et sous R.

Les résultats des modèles élaborées sont très satisfaisants après avoir les appliquer sur des données de test nouvellement présentés. Nous avons bien réussi à atteindre notre objectif principal et donc nous sommes déjà capables de prédire les futurs joueurs nommées pour la FIFA Ballon d'or 2018.


FIFA
D'OR