Projet M2 Fouille de données
2021-2022

Contexte

Jeu de données : 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.

Comme son nom ne l'indique pas, il s'agit d'un fichier tsv (tab separated values) : les différentes colonnes sont séparées par des tabulations.
Certaines colonnes correspondent au score sur 100 du joueur à différents postes. Voici la liste des abréviations utilisées pour les postes, qui pourra vous servir pour l'interprétation des résultats.

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

Il n'est pas nécessaire de s'y connaître en foot pour faire le projet, l'objectif est de fournir des arguments scientifiques sur ce que vous aurez observé dans les données, plutôt que des arguments d'experts en foot.

Logiciels : Vous utiliserez WEKA ou Orange pour faire vos analyses, R si nécessaire.

Rendu 2 rapports sont à faire : un rapport pour les parties 1 et 2, un deuxième rapport pour la partie 3. Ces rapports peuvent être réalisés en binôme.

Partie 1 – Analyse descriptive des données

Dans un premier temps, il convient de prendre en main les données. Vous vous servirez du support de cours, des visualisations offertes par WEKA, Orange ou R ou de votre tableur préféré et du document « analyse de données » pour réaliser une première analyse descriptive de données.

Votre rapport contiendra l'analyse à destination d'un joueur de Fifa19.

Par la suite, nous appellerons cette personne le décideur. Vous l'aiderez à tirer quelques premières conclusions.

1.1 Préparation des données

Le fichier ne peut pas être utilisé tel quel. Il faut transformer les données pour les rendre compréhensibles par les logiciels d'analyse.

Importation dans Weka/Orange

L'importation dans Weka n'est pas automatique. En effet, le TSV contient des données que WEKA n'arrivera pas à transformer. Vous pouvez aller sur https://weka.wikispaces.com/ARFF+%28book+version%29 pour avoir un rappel des entêtes d'un fichier `arff`.

Par exemple, il est rappelé qu'un fichier `arff` doit avoir une entête, qu'une donnée manquante est représentée par un `?`, que du texte de type `string` est entre `''`.

De plus, pour appliquer les algorithmes en partie 2, il est nécessaire de créer ou modifier certaines colonnes :

- Création d'un **groupe d'âge** GA : -20, 20-25, 25-30, 30-35, +35
- Création d'un **attribut BMI** (body mass index) ou indice de masse corporelle (IMC) : poids[kg] / taille^2[m^2] (attention aux unités !!)
- Modification de la granularité des **positions joueurs** : vous créerez un attributs position_joueur avec 4 valeurs : GK pour les goals, DEF pour tous les défenseurs, MID pour tous les milieux et FWD pour tous les attaquants
- Certaines colonnes contiennent des « + », par exemple « 88+2 ». Il s'agit de la valeur avant la sortie du jeu, puis la valeur actualisée à la sortie du jeu. Pour l'analyse des données, vous pouvez ignorer la valeur au-delà du +.
- D'autres colonnes contiennent des **valeurs financières** sous le format €105.3M ou encore €77K : à vous de les formater pour qu'Orange/Weka les accepte ET que ce soit unifié.
- Le poids est stocké en **livres** sous la forme 159lbs : à reformater pour enlever le « lbs » final. De la même manière, vous aurez à reformater les mesures en **pouces** pour obtenir un format numérique lisible.
- Discrétisation de Value en DValue : proposez une discrétisation pertinente
- Discrétisation de Wage en DWage : proposez une discrétisation pertinente
- Suppression des attributs **photo** et **Club Logo**

1.2 Analyse

Attributs et Statistique descriptive

Présentez les différents attributs (type, signification, …), donnez quelques statistiques descriptives et quelques visualisation appropriées.

Analyse de corrélation

Réalisez une étude de la corrélation entre les variables (on réalisera directement les matrices de corrélation). Discutez cette étude.

Prise en main de données

- Créez l'équipe de 11 joueurs la plus chère
- Créez l'équipe de 11 joueurs la plus forte

Est-ce la même ? sinon qu'est-ce que cela vous permet de conclure par rapport à l'analyse de la corrélation de la question précédente ?

Partie 2 – Segmentation

Pour cette partie, nous limiterons l'étude à certains attributs et certaines instances. Vous éliminerez les goals de cette étude. Vous vous limiterez à l'études des variables numériques et vous omettrez pour le calcul des groupes les variables : Player value, wages and overall rating
Attention à la normalisation des données dans les algorithmes que vous utiliserez …

1. Faites tourner Kmeans et trouver le nombre optimal de groupe
2. Interprétez les groupes
3. Trouvez un remplaçant pour le joueur MBappé
4. Testez différents algorithmes de clustering, comparez les résultats

A RENDRE pour les parties 1 et 2 : les fichiers qui correspondent à la préparation des données, et un compte-rendu au format PDF qui répond aux questions de l'analyse (partie 1) et présente l'étude de segmentation que vous avez réalisée (partie 2).

Partie 3 – Prédiction

On cherche à savoir prédire les variables discrétisées Dwage et la variable Dvalue
Vous utiliserez l'ensemble des attributs crées en partie 1.
Vous testerez différentes méthodes de classification pour chaque variable à prédire (que vous avez discrétisée).
Vous utiliserez également quelques algorithmes de régression pour prédire les valeurs d'origine de wage et vale.
Voici quelques questions auxquelles vous devez répondre :

1. Peut on prédire le prix d'un joueur ?
2. Peut on prédire son score ?
3. Est-ce que ce qui rend un joueur cher rend un français cher ? i.e. si on se restreint aux joueurs français arrive t-on aux mêmes conclusions ?
4. Un décideur vous dit qu'il ne sélectionne les joueurs que par rapport à Overall Ratings, Wage, International Reputation, Weak Foot et Skill Moves. Que pouvez-vous lui dire par rapport à vos différents tests de prédiction ?

A RENDRE pour la partie 3 : un compte-rendu au format PDF qui présente l'étude de prédiction que vous avez réalisée et qui permet de répondre aux questions posées.