**Tipologia i cicle de vida de les dades  
Pràctica 2**

**Celio Sánchez Bañuls  
José V. Grimalt Aranda**

**Màster de Ciència de Dades  
set.2021 - gen.2022**

**Taula de continguts**

[Integrants del grup 3](#_30j0zll)

1[. Descripció del *dataset*](#_2et92p0) [4](#_1fob9te)

[2. Integració i selecció de les dades d’interès a analitzar 4](#_3znysh7)

[3. Neteja de les dades. 5](#_2et92p0)

[4. Anàlisi de les dades. 5](#_tyjcwt)

[5. Representació dels resultats a partir de taules i gràfiques. 6](#_3dy6vkm)

[6.Resolució del problema. 7](#_1t3h5sf)

7[.](#_2s8eyo1) [Codi](#_17dp8vu) [11](#_2s8eyo1)

[Taula de contribucions al treball 12](#_26in1rg)

## Integrants del grup

Aquest són els integrants del grup que ha realitzat la pràctica:

|  |  |  |
| --- | --- | --- |
| **Nom** | **E-mail** | **Inicials** |
| Celio Sánchez Bañuls | [celiosanchez@uoc.edu](mailto:celiosanchez@uoc.edu) | CSB |
| José V. Grimalt Aranda | [jgrimalt@uoc.edu](mailto:jgrimalt@uoc.edu) | JGA |

## 1. Descripció del *dataset*

El conjunt de dades està compost per un únic fitxer amb format CSV, que conté informació de valoració de mercat dels 500 millors jugadors de futbol en cada posició segons els criteris del lloc web [transfermarkt.com](https://www.transfermarkt.com/). Aquest *dataset* inclou diferents tipus de dades:

* Identificació del jugador: nom i URL en [transfermarkt.com](https://www.transfermarkt.com/).
* Dades personals: edat, posició, nacionalitat, club.
* Dades sobre valoració: valor de mercat, *ranking* per posició.
* Dades estadístiques de la temporada en curs: gols marcats, targetes, substitucions...

Perquè és important i quina pregunta/problema pretén respondre?

Aquest dataset es important tan per aficionats de futbol com a com a clubs de futbol que vullguen obtindre dades rellevants sobre el valor de mercat basat en estadístiques.

En esta practica netejarem les dades i tratarem de estimar un model amb el qual a partir de les dades puguem estimar el valor de mercat a partir de cadascuna de les característiques del jugador.

## 2. Integració i selecció de les dades d’interès a analitzar

Atès que es tracta d'un conjunt de dades amb informació no lligada a cap idioma, escollim un nom en anglès que descriu el contingut. Considerem important incloure el nom de l'empresa que proporciona la informació per donar crèdit i contextualitzar l'origen de les dades.

**transfermakt-com-most-valuable-football-players-YYYYMMDD.csv**

Ací haurem d’explicar quines dades utilitzem: tot el data set o sols uns jugadors/països/clubs concrets? També si hem necessitat integrar dades d’altres datasets.

## 3. Neteja de les dades.

### 3.1. Elements buit i dades que contenen zeros.

Les dades contenen zeros o elements buits? Com gestionaries aquests casos?

Anem a revisar els valors per veure la cantitat de valors nuls que existeixen en cada atribut.

|  |  |  |
| --- | --- | --- |
| **position\_id** | *Numèric* | Identificador de la posició del jugador |
| **position\_ranking** | *Numèric* | Classificació del jugador segons la posició |
| **player** | *Text* | Nom del jugador, no es una variable númerica |
| **url** | *Text* | URL relativa del jugador en el web [transfermarkt.com](https://www.transfermarkt.com/) |
| **position** | *Text* | Nom descriptiu de la posició,no es una variable númerica |
| **age** | *Numèric* | Edat |
| **country** | *Text* | Nacionalitat |
| **club** | *Text* | Club o equip en el que juga actualment |
| **value** | *Numèric* | Valoració de mercat en € |
| **matches** | *Numèric* | Partits disputats |
| **goals** | *Numèric* | Gols marcats |
| **owngoals** | *Numèric* | Gols marcats en pròpia porteria |
| **assists** | *Numèric* | Assistències |
| **yellowcards** | *Numèric* | Targetes grogues |
| **yellow2cards** | *Numèric* | Expulsions per dues targetes grogues |
| **redcards** | *Numèric* | Expulsions per targeta roja directa |
| **subston** | *Numèric* | Nombre partits jugats sense ser titular |
| **substoﬀ** | *Numèric* | Nombre de vegades que el jugador és substituït |

### 3.2. Identificació i tractament de valors extrems.

Hi ha valors extrems? Els tenim en compte o no en l’anàlisi?

## 4. Anàlisi de les dades.

### 4.1. Selecció dels grups de dades que es volen analitzar/comparar

(planificació dels anàlisis a aplicar).

De les variables ens interesa eliminar les següents:

* Value/matches. El resultat de valor de mercat/ partits.
* Value / age. El resultat de valor de mercat/edad
* Value/goals. El resultat de valor de mercat / gols.
* Value/red cards. El resultat de valor de mercat/ tarjetes rojes.
* Value/yellow2cards. El resultat de valor de mercat /dobles tarjetes grogues,

### 4.2. Comprovació de la normalitat i homogeneïtat de la variància.

### 4.3. Aplicació de proves estadístiques per comparar els grups de dades.

En funció de les dades i de l’objectiu de l’estudi, aplicar proves de contrast d’hipòtesis,

correlacions, regressions, etc. Aplicar almenys tres mètodes d’anàlisi diferents

## 5. Representació dels resultats a partir de taules i gràfiques.

Dades recollides en la taula del conjunt de dades:

|  |  |  |
| --- | --- | --- |
| **Camp** | **Tipus** | **Descripció** |
| **position\_id** | *Numèric* | Identificador de la posició del jugador |
| **position\_ranking** | *Numèric* | Classificació del jugador segons la posició |
| **player** | *Text* | Nom del jugador |
| **url** | *Text* | URL relativa del jugador en el web [transfermarkt.com](https://www.transfermarkt.com/) |
| **position** | *Text* | Nom descriptiu de la posició |
| **age** | *Numèric* | Edat |
| **country** | *Text* | Nacionalitat |
| **club** | *Text* | Club o equip en el que juga actualment |
| **value** | *Numèric* | Valoració de mercat en € |
| **matches** | *Numèric* | Partits disputats |
| **goals** | *Numèric* | Gols marcats |
| **owngoals** | *Numèric* | Gols marcats en pròpia porteria |
| **assists** | *Numèric* | Assistències |
| **yellowcards** | *Numèric* | Targetes grogues |
| **yellow2cards** | *Numèric* | Expulsions per dues targetes grogues |
| **redcards** | *Numèric* | Expulsions per targeta roja directa |
| **subston** | *Numèric* | Nombre partits jugats sense ser titular |
| **substoﬀ** | *Numèric* | Nombre de vegades que el jugador és substituït |

Notes:

* Tots els camps de tipus numèric són de nombres enters.
* Les dades estadístiques (partits, gols, targetes, etc.) fan referència a la temporada en curs.

## 6. Resolució del problema.

A partir dels resultats obtinguts, quines són les conclusions?

Els resultats permeten respondre al problema?

## 7. Codi

Aquest és l'enllaç al repositori Git del codi amb el qual s'ha generat el *dataset*:

* URL: <https://github.com/Celiosanchez/TCVD-Practica2>

El codi s'ha desenvolupat en R, utilitzant les següents llibreries:

* **tidyverse**: és un paquet que inclou diverses llibreries útils per anàlisi estadística.

## Taula de contribucions al treball

|  |  |
| --- | --- |
| **Contribucions** | **Signatura** |
| Investigació prèvia | CSB, JGA |
| Redacció de les respostes | CSB, JGA |
| Desenvolupament del codi | CSB, JGA |