**Pràctica 2**

1. Descripció del dataset. Perquè és important i quina pregunta/problema pretén

respondre?

El dataset serà el que vam generar a la pràctica 1. Consisteix en un conjunt de dades provinents del rànquing d’atletisme espanyol dels últims cinc anys. Així doncs conté el conjunt d’atletes i resultats d’aquests en els últims anys. Creiem que les dades tenen utilitats molt diverses, però ens hem volgut centrar en utilitzar les dades per esbrinar la progressió dels atletes, així doncs, ens agradaria respondre a la pregunta següent; Podem saber quina serà la marca d’un atleta X en una prova Y la següent temporada? O més fàcil, millorarà el atleta X la seva marca la temporada següent?

Som conscients que en les dades del rànquing hi poden faltar molts atributs que poden ser decisius en aquesta pregunta (Lesions, condicions d’entrenament, disponibilitat de material...), però ens agradaria poder fer-nos una idea general.

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

En principi, per el què fa a la integració no afegirem més dades al dataset. Tot i això, creiem que segons les necessitats de l’estudi es podrien afegir més anys de dades (en el rànking hi tenim fins als últims 15 anys) o més atletes ja que en el datset només es mostren les 1000 millors i per a proves amb molt participants en podríem trobar bastantes més.

En canvi la selecció creiem que si que és important en el nostre cas ja que en les dades hi trobem bastants atributs que no són rellevants en l’estudi que es vol dur a terme. A continuació llistem els atributs i indicarem els que ens semblen rellevants:

- **Tipus de competició o pista**: Pista Coberta, Aire Lliure o Ruta. Aquesta és una dada que no té molta rellevància ja que la majoria de proves són diferents segons el tipus de pista, per exemple les curses de 5km, 10km, 15km… només es donen en la ruta i si es fan en pista les condicions són tant diferents que no es considera la mateixa prova. En canvi els 100m per posar un altre exemple només es fan al Aire lliure, ja que en pista coberta no es poden fer perquè la pista té unes mides diferents i en ruta no té sentit. Així doncs aquest atribut creiem que no és necessari i l’eliminarem, ja que amb la prova en tenim prou.

- **Sexe**. Indica el sexe del atleta. És rellevant, però passa semblant que amb el camp anterior, ja que en principi les proves masculina i femenina són diferents. Crec que es pot prescindir d’aquest camp.

- **Prova**: Si ens fixem en el camp de prova, es pot veure que ja conté els dos camps previs (per ex: «100m MASC. AL»). Aquest camp és indispensable.

- **Posició en el rànquing**: Aquest camp indica la posició del atleta en la temporada concreta. Aquesta pot ser una dada rellevant ja que pot mostrar una relació amb els atletes que tenen més recursos, ja que els primers del rànquing acostumen a tenir algun ajut (econòmic, accés a instal·lacions d’alt rendiment, ajut mèdic...)

- **Marca**: Indica la millor marca del atleta en la temporada. Aquest és un altre dels camps clau.

- **Vent**: Indica el registre del vent en la prova que es va fer la marca. No creiem que sigui un camp rellevant ja que depèn de la competició (no es pot preedir) i no es registra en totes les proves.

- **Atleta**: indica la persona que va fer la marca, és indispensable com a identificador.

- **Data de naixement**: Ens indica l’edat del atleta, és important per saber amb quants anys va fer la marca

- **Categoria**: No és molt rellevant ja que va relacionat amb l’edat de l’atleta

- **Número de llicència**: És un identificador del atleta, el problema que pot tenir és que hi ha alguns atletes que canvien de número de llicència (perquè canvien de comunitat autònoma o perquè un any no renoven la fitxa). Creiem que és millor el nom ja que és menys habitual que es canviï

- **Federació**: Indica la comunitat autònoma del atleta. No tenim clar que sigui significatiu en la marca del atleta. Tot i que podria ser un indicador de on hi ha pistes més bones i millors entrenadors, la realitat es que els atletes competeixen en diverses comunitats al llarg del any i entrenen en comunitats diferents a la de la seva federació. Així que no és molt rellevant.

- **Club**: Indica el club de l’atleta. En principi pel que fa a la marca no hauria d’influir, ja que en l’atletisme el club no dona indicis de l’entrenador o la ubicació del atleta (sobretot en edat senior)

- **Posició** (en la prova que es va fer la marca): Aquest atribut en podem prescindir totalment ja que la posició depèn de la competició i no està relacionat amb la marca. Amb la mateixa marca pots quedar primer a una cursa popular, i últim en una carrera professional.

- **Lloc on es va fer la marca**. És un altre atribut prescindible, ja que el lloc no és molt significatiu respecte a la marca, perquè un atleta pot anar a competir on vulgui.

- **Data de la marca**. Creiem que aquest camp si que és clau ja que ens indica la temporada i l’edat de l’atleta quan va fer la marca.

Així doncs ens quedarem amb els següents atributs:

Prova, Marca, Atleta, ¿Club? I Edat (l’obtindrem ajuntant les dades de data de naixement i la data de la marca).

Apart de la selecció d’atributs pensem que podem reduir la quantitat, principalment ja que hi ha moltes proves que no tenen a veure una amb la altre, i per tant és molt probable que l’estudi s’hagi de fer prova a prova.

3. Neteja de les dades.

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

Les dades s’han hagut d’arreglar bastant, principalment perquè hem trobat incongruències o formats erronis, deguts a errors en la pujada de dades. Per exemple hi ha unes files que no tenen els atributs ben distribuïts dins el csv, perquè estaven en un format diferent.

També hem vist que en l’atribut de Marca hi ha dades que contenen una «A» al final, és degut a que en el rànquing es posa una A al final quan la marca s’ha fet en Alçada (més de 1500m), tot i que és una cosa que es senyalitza la marca es conta com a vàlida, i el que hem fet ha estat eliminar les A’s.

3.2. Identificació i tractament de valors extrems.

4. Anàlisi de les dades.

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

anàlisis a aplicar).

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.

6. Resolució del problema. A partir dels resultats obtinguts, quines són les conclusions?

Els resultats permeten respondre al problema?

7. Codi: Cal adjuntar el codi, preferiblement en R, amb el que s’ha realitzat la neteja,

anàlisi i representació de les dades. Si ho preferiu, també podeu treballar en Python.