

## Resum del projecte

1. Agafem dades (les d'eficiència de cotxes de la primera classe que va trobar l'Albert) i obtenim el CSV de la pàgina web aquella (<https://www.kaggle.com/datasets/arslaan5/explore-car-performance-fuel-efficiency-data>).
2. Ns qui havia de fer l'anàlisi descriptiu previ però és el punt 2 de tasques a fer del projecte.
3. Fem el pre-processament de les dades:
  - a. Escurçar noms de variables.
  - b. Convertir els números en integers, perquè detectava els números com a strings.
  - c. Transformem el CSV a ".obj".
  - d. En la documentació mencionar que no hi havia dades missing perquè algú ja les havia tractat anteriorment (potser val la pena investigar quin tractament li han fet a les dades els senyors de la pàgina web aquella).
  - e. NOTA: la professora ha dit que necessitem la passarel·la, és un script de RStudio que fa la transformació de dades de CSV a ".obj".
  - f. A classe amb la profe hem intentat arreglar el script de R que va fer el Charlie, que fa servir la passarel·la però que no funciona correctament.
4. **IMPORTANT:** importar fitxer amb **dades SENSE regles** en el klass i seleccionar **totes les files** (no s'han de treure les dades amb cotxes elèctrics ni de dièsel, això és diferent al que vam dir a classe la setmana passada).
  - a. Anem al klass: Arxiu ☐ obre ☐ dades sense regles... i seleccionem el fitxer amb les dades de format ".obj".
  - b. NOTA: la profe ha dit que encara que siguin outliers no passa res, en teoria el klass hauria de ser capaç de classificar-los fent el clúster i creant una classe per a cada tipus de combustible.
  - c. Això **ho hem de fer tots** per a poder després crear els gràfics i tot.
5. Fer clustering:
  - a. Obtenim un dendrograma (l'arbre aquell de classes) i altres gràfics que treu el klass.
  - b. El tallem amb el klass.
6. Fer el CPG i descriptiva per classes (classificació de dades):
  - a. El "per classes" es refereix a les classes que hem obtingut de tallar el dendrograma en el pas 5.
  - b. Ha de sortir un gràfic en forma de taula. En cada cel·la de la taula tenim com un mini gràfic de barres. Les files són les classes obtingudes en el pas 5 i les columnes són els atributs/variables. Per exemple, la fila seria país i la columna seria el PIB.
7. Fer el Traffic Light Panel (TLP) segons el que ha sortit al CPG (se suposa que hem fet un anàlisi previ del CPG abans de fer el TLP):
  - a. En el Klass obtenim una sèrie de gràfics a dins d'una taula on les files són les classes i les columnes els paràmetres.
  - b. N'hi ha de dos tipus, el "normal" i l'"annotated". El normal ha estat pintat amb el nostre criteri. L'*annotated* representa la variància, la fiabilitat del resultat.

8. Fer el termòmetre i aplicar-lo al CPG, obtenim un nou TLP:
  - a. El termòmetre és semblant a PEGTI, valorar quins límits ficar per al color vermell, groc i verd.
  - b. Consisteix en ficar-li els colors segons el que hem decidit al termòmetre, de manera que podem veure visualment quins atributs (columnes) per a cada classe (files) estan en la part vermella, verda o groga.
  - c. També surten dos TLPs, un normal i un annotated.

El que hem fet fins ara és sense tenir ni idea de com estan distribuïdes les dades, sense tenir coneixement previ, però ara hauríem de ser capaços d'haver trobat alguna relació entre variables i informació d'aquestes. Amb aquesta informació podem definir ontologies (com condicions) perquè el pròxim clustering (dendograma) que fem tingui ja una idea de "com va la cosa".

1. **IMPORTANT:** importar les **dades AMB regles** al klass:
  - a. Arxiu ☐ Obre ☐ dades amb regles...
  - b. Seleccionar el ".obj".
2. Definir ontologies (com es fa el fitxer ".ont" està al vídeo penjat al drive, minut 2:06:30).
  - a. Definir base de coneixement (Knowledge-Based, surt com a KB en el pdf del projecte). Explicar la ontologia.
  - b. Fer tots els passos anteriors:
    - i. Anàlisi descriptiu univariant
    - ii. Clustering amb dendograma (per què s'ha fet aquest nou tall i per què ha canviat el dendograma)
    - iii. Class Panel Graph (quines noves conclusions es poden treure i quines diferències hi ha, per què han passat)
    - iv. Traffic Light Panel (n'hi haurà dos, el normal i l'annotated)
    - v. Termòmetre aplicat al TLP (aquest termòmetre és **el mateix** que el fet servir anteriorment. Obtenim dos TLPs un altre cop, un normal i un annotated)
  - c. Aquesta part consisteix en comparar les noves figures resultants amb les anteriors i analitzar per què pot ser que han aparegut aquestes diferències.