QUIN PROCÉS SEGUEIX LA PIPELINE ?

En aquest document estan explicades totes les pases que es duen a terme durant el preprocessament de les bases de dades en el moment que feu *click* a “Pujar dades”:

1. Inserció de dades

Les dades s’introdueixen mitjançant una interfície web o aplicació local. Aquesta estratègia permet que els fitxers d’entrada es desaquin automàticament en una estructura de carpetes local, facilitant la gestió i persistència de la informació per a processos posteriors.

2. Persistència local dels fitxers

El desat local dels fitxers d’entrada permet reutilitzar exactament la mateixa pipeline amb noves mostres del mateix dataset, sempre que es mantingui l’estructura de columnes. Això garanteix consistència en l’anàlisi i evita errors estructurals.

3. Processament independent de datasets

Cada dataset carregat s’envia a un mòdul de preprocessament específic. En aquesta fase, és crucial que l’estructura de columnes es mantingui estable. Variacions en les columnes poden comprometre la integritat del processament i generar errors crítics en les visualitzacions. Tot i així, s’han aplicat criteris de robustesa per garantir una correcta execució encara que hi hagi l’existència d’errors o de diferències amb el conjunt de dades original.
En el document de diccionaris, es troba les variables que calen per una correcta execució hi ha justificació de les variables que han estat eliminades i les que s’han conservat.

4. Generació de datasets processats

El processament genera tres sortides diferents:

- Dataset individual net de les dades del SIAD
- Dataset individual net de les dades de Beques Menjador.
- Un tercer dataset agregat, obtingut a partir de la combinació dels datasets anteriors mitjançant tècniques d’agregació (ex. mitjanes, sumatoris, etc.).

4.1. Realització del Join

A l’hora de realitzar el *join*, s’han utilitzat principalment les columnes del SIAD i algunes columnes relacionades pertanyents al dataset de les beques, a les quals s’han aplicat tècniques d’agregació. L’objectiu principal d’aquest conjunt de dades era identificar els individus que apareixen en ambdós datasets, és a dir, aquells que han patit una discriminació i que tenen un o més fills becats. D’aquesta manera, l’objectiu era obtenir una combinació d’aquestes dues fonts d’informació.

En el procés de combinació, s’han identificat algunes columnes que són constants per unitat familiar. D’aquestes, s’ha calculat la mitjana per aquelles que representen valors agregats renda familiar, volum negoci o rendiment capital mobiliari; o bé s’ha mantingut el valor constant per altres columnes com finques urbanes, família nombrosa o risc social. També s’ha registrat la quantitat d’aparicions d’un individu al dataset com a total de fills o bé s’han sumat valors com en el cas de preu de l’ajut.

A més, s'han creat columnes específiques per cada fill o familiar de l'individu que ha patit una discriminació. Així, es generarà tantes columnes com el nombre màxim de fills possibles per individu en el dataset. Si una persona té menys d'un fill, les columnes restants tindran un valor nul. Aquestes columnes són: beca garantida, discapacitat de més de 33, punts d'unitat familiar, negligència, seguiment mèdic excessiu, codi postal, sexe, data naixement, nacionalitat, centre escolar, nivell escolar i curs.

Del dataset de beques s'han agafat les columnes que semblaven més rellevants i la resta s'han omès, simplement fent una petita modificació del codi es podrien incloure quan es desitgi.