Documentatie

Pentru a rezolva aceasta problema am folosit limbajul de programare Python deoacere in acesta se gaseau toate facilitatile de a rezolva cu usurinta problema. Librariile folosite in aceasta problema au fost: pandas, numpy si rapidfuzzy. La aceste pachete se adauga versiunea de Python 3.9.

Pentru a asigura eficienta programului din punct de vedere a timpului de executie si a spatiului de executie, dar si pentru a eficientiza cat mai mult programul am folosit o tehnica de predictie folosita in machine learning, Fuzzy matching. Pe langa aceasta tehnica am folosit biblioteca pandas, care este specializata pe lucru cu volume mari de date.

Mai jos voi prezenta linie cu linie codul:

* Liniile 1-3: Importul pachetelor folosite in cadrul proiectului
* Liniile 6-13: Se foloseste pachetul pandas pentru a citi fisierul sursa si pentru a face conversia din “parquet” in excel. Sistemul Fuzzy solicita lucrul cu un excel pentru a putea functiona.
* Liniile 15-19: Se selecteaza dinamic coloanele in care se afla caractere numerice sau litere. Prin acest intermediu se elimina parcurgerea coloanelor in care nu se afla caractere
* Linia 20: Se defineste functia prin care vom sterge duplicatele. Parametri intrare:

1. Tabloul de date
2. Coloana pe care se afla duplicatele
3. Parametru Fuzzy: scorul de simililaritate

* Liniile 23-29: Se definesc alte functii care vor ajuta la rezolvarea problemei. “combina\_valorile\_text” combina valorile de text unice eficient, iar “combina\_valorile\_numerice” se ocupa de valorile numerice
* Liniile 32-35: Se extrag valorile duplicate, algoritmul propriu-zis de eliminare a copiilor
* Liniile 37-50. Se aplica algoritmul de fuzzy pentru a verifica duplicatele exacte si eventualele greseli. De exemplu daca se gaseste “Iphone 13” si “ iphone 13” se creeaza acelasi produs si se elimina prima varianta.
* Liniile 52-63. Se aplica functiile create si se afiseaza rezultatele