Supplier Entity Resolution si Spend Analysis

Proof of Concept

1. Contextul proiectului

Proiectul a avut ca punct de plecare o problema des intalnita in mediul de business: lipsa de coerenta a datelor despre furnizori. Baza de date continea multiple inregistrari pentru aceeasi companie, cu diferente de denumire, informatii incomplete si formate inconsistente.

Aceasta situatie afecta direct acuratetea analizelor financiare, deoarece acelasi furnizor era distribuit artificial in mai multe entitati, distorsionand rezultatele de tip Spend Analysis.

Obiectivul POC-ului a fost sa demonstrez ca acest proces poate fi automatizat intr-un mod elegant si controlat, astfel incat furnizorii sa fie identificati unitar, iar rezultatul final sa reprezinte o baza de date curata si utilizabila, fara interventii manuale costisitoare.

---

2. Abordarea folosita

Solutia a fost construita pentru a reproduce modul in care un analist uman ar lua decizii, dar cu avantajul vitezei si consistentei. Am combinat reguli de business, standardizare de date si un sistem de scor care evalueaza obiectiv fiecare varianta posibila.

2.1 Standardizare si filtrare initiala

Prima etapa a constat in curatarea datelor si uniformizarea formatelor.
Am eliminat caracterele invizibile, am normalizat denumirile si am validat informatia geografica.

Am aplicat o regula stricta: o entitate legala nu poate exista simultan in doua tari diferite. Aceasta regula a eliminat din start o mare parte din potrivirile incorecte.

2.2 Calcularea scorului de similaritate

Pentru fiecare candidat ramas valid, am calculat un scor compus din doua componente principale:

- gradul de similaritate intre denumiri
- potrivirea informatiilor geografice (tara, regiune, oras)

Scopul nu a fost doar sa potrivesc texte, ci sa inteleg contextul in care apare fiecare furnizor. Astfel, firme cu nume foarte apropiat, dar in tari diferite, sunt corect tratate ca entitati distincte.

2.3 Clasificarea rezultatelor

Pe baza diferentelor dintre scoruri, fiecare caz a fost incadrat intr-o categorie de incredere:

- Verde – potrivire clara, selectata automat
- Galben – caz ambiguu, recomandat pentru verificare
- Rosu – potrivire slaba, dar pastrata pentru continuitatea datelor financiare

---

3. Interventia umana

Pentru cazurile galbene, nu a fost necesara o analiza detaliata a tuturor atributelor. Am aplicat o regula simpla, dar extrem de eficienta: validarea pe baza codului postal.

Aceasta veriga finala de control a permis confirmarea rapida a entitatilor corecte, fara pierdere de timp in analize inutile. Intregul proces de verificare manuala a durat sub zece minute pentru intregul set de date.

---

4. Rezultate obtinute

In urma rularii solutiei:

- aproximativ 95% din inregistrari au fost rezolvate automat
- lista finala contine o singura entitate per furnizor
- doar cazurile rosii necesita decizie manuala
- datele sunt coerente si pregatite pentru analiza Spend Analysis

Au fost livrate doua tipuri de fisiere:

- un raport detaliat de audit, care contine toate variantele analizate
- fisierul final, destinat utilizarii in analiza de business

---

5. Tehnologii folosite

Proiectul a fost dezvoltat in Python, folosind Pandas pentru prelucrare de date, Difflib pentru comparari de tip fuzzy matching si OpenPyXL pentru generarea si stilizarea fisierelor Excel.

---

6. Impact pentru business

Automatizarea a redus aproximativ 90% din efortul manual asociat curatarii datelor. In acelasi timp, procesul a devenit reproductibil, transparent si mult mai fiabil decat abordarea clasica rand-cu-rand.

In loc sa consume timp cu operatiuni repetitive, analistul se poate concentra acum pe interpretarea corecta a datelor si pe analizarea cheltuielilor reale.