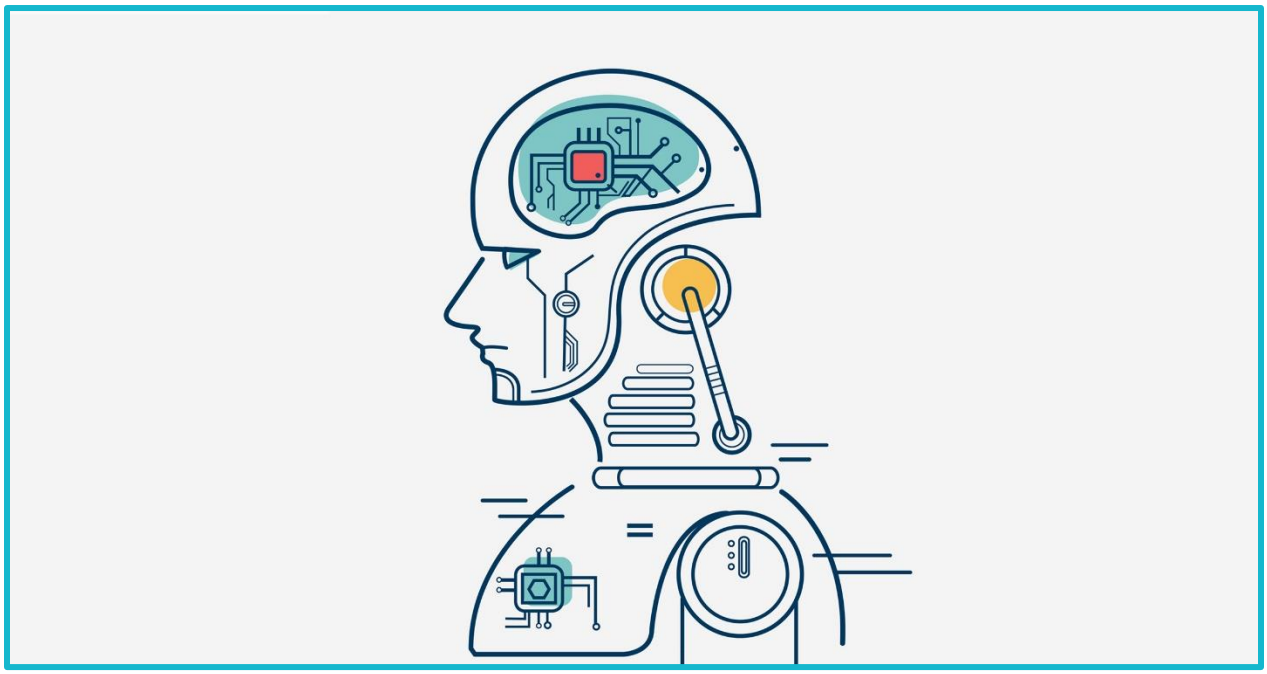


Predicția subregiunilor europene utilizând tehnici de clasificare supervizată

SOFTWARE PENTRU STATISTICĂ ȘI DATA SCIENCE

Enache Valentina
Grupa 1096

Cuprins

Prezentarea obiectivelor și a modelului statistic

Proiectul are ca obiectiv folosirea clasificării supervizate pentru a putea prezice o variabilă calitativă după mai multe variabile cantitative (variabile *predictor*).

Variabila calitativă Y este reprezentată de subregiunea geografică a Europei, aceasta fiind estimată cu ajutorul a 10 atribute de ordin economic ce măsoară performanța în diverse domenii ale țărilor din Uniunea Europeană (UE27). Așadar, tipul studiului este de clasificare.

Variabilele predictor sunt:

- **PIB**: PIB (reflectă valoarea totală a tuturor bunurilor și serviciilor produse, mai puțin valoarea bunurilor și serviciilor utilizate pentru consumul intermediar în producția lor; indicator exprimat în prețuri curente, ca milioane de euro);

- **PIB per capita**: PIBPerCapita (PIB-ul pe cap de locuitor, exprimat ca euro per capita în prețuri curente);

- **PIB real per capita**: PIBRealPerCapita (indicator calculat ca raport dintre PIB real și populația medie a unui an specific și exprimat ca euro per capita);

- **Salariul mediu**: SalariuMediu (indicator exprimat în euro);

- **Procentul angajaților plătiți cu salariul minim pe economie**: ProcentSalariuMinim;

- **Rata șomajului**: RataSomaj (șomerii ca procent din forța de muncă, cuprinzând persoanele cu vârste între 15 și 74 de ani);

- **Exportul bunurilor și serviciilor ca procent din PIB**: ExportProcentPIB (valoarea exporturilor de bunuri și servicii, împărțită la PIB în prețuri curente);

- **Importul bunurilor și serviciilor ca procent din PIB**: ImportProcentPIB (valoarea importurilor de bunuri și servicii împărțită la PIB în prețuri curente);

- **Raportul dintre export și import**: ExportImportRatio;

- **Cheltuieli de cercetare și dezvoltare ca procent din PIB**: CheltuieliRsiDProcentPIB (cheltuielile interne brute pentru cercetare și dezvoltare ca procent din PIB).

Variabila țintă este:

- **Subregiunea** (conform EuroVoc, subregiunile din Europa sunt: Europa de Nord, Europa Centrală și de Est, Europa de Sud, Europa de Vest).

Pentru setul de învățare s-au extras date de pe Eurostat corespunzătoare anilor 2018-2019, iar pentru setul de testare au fost folosite date din 2014.

Implementarea modelului și prezentarea rezultatelor

Pentru setul de învățare, după înlocuirea valorilor lipsă cu media, observăm caracteristicile statisticii descriptive:

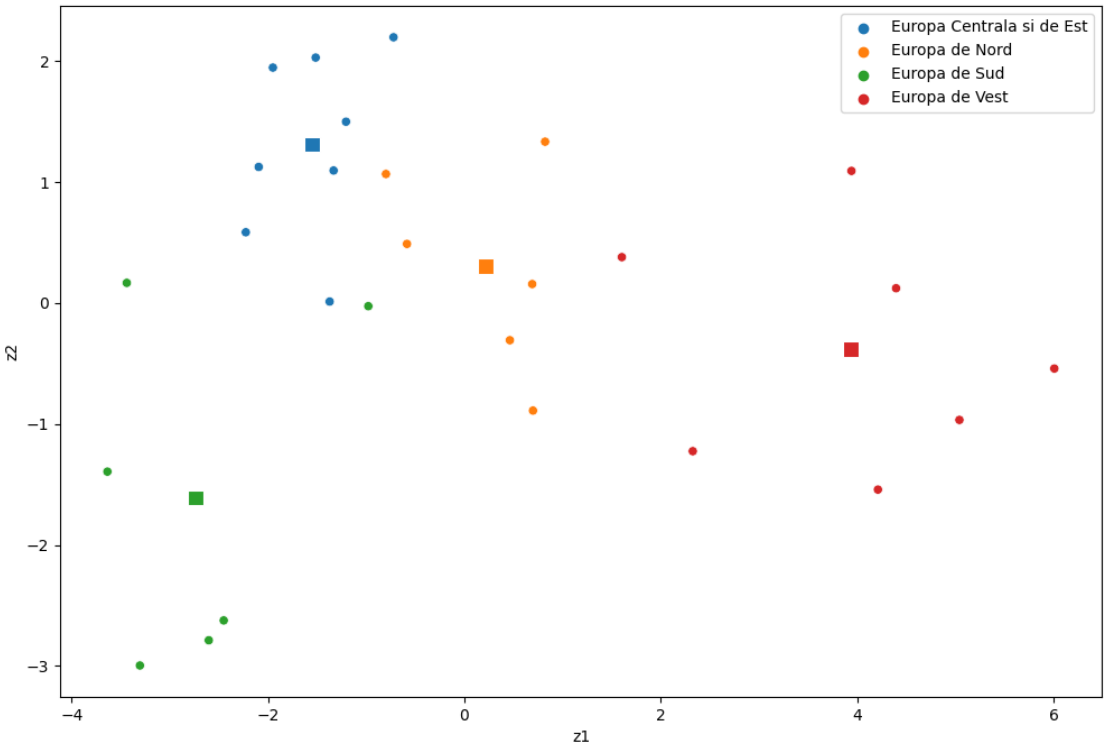
	PIB (milioane euro)	PIB per capita	PIB real per capita	Rata șomaj	Export procent PIB	Import procent PIB	Cheltuieli R&D procent PIB
count	27	27	27	27	27	27	27
mean	517,181,185	31,581	27,642	6.04	70.38	65.69	1.65
std	817,365,303	20,861	17,546	3.32	38.54	32.18	0.9
min	134,624,00	8,780	6,840	2.0	31.5	28.5	0.48
25%	577,387,00	17,155	14,885	4.05	45.2	43.4	0.91
50%	223,337,400	25,310	21,800	5.4	60.1	60.1	1.4
75%	475,335,750	42,540	36,555	6.65	82	74.1	2.17
max	3,449,050,000	102,200	83,640	17.3	208.8	172.8	3.39

Aplicarea modelului LDA

Analiza discriminantă liniară este o generalizare a discriminantului liniar Fisher, ce încearcă să caute combinații liniare de variabile care să explice cât mai bine datele.

Pentru implementarea modelului de analiză discriminantă, am ales din setul de învățare (datele corespunzătoare anilor 2018-2019) variabilele predictor, adică toate variabilele cantitative, și variabila țintă, adică regiunea. În model, regiunea are 4 clase, adică Europa de Nord, Europa Centrală și de Est, Europa de Sud și Europa de Vest. S-au calculat matricea scorurilor discriminante și matricea centrelor de grupă (folosind variabilele discriminante), iar graficul rezultat arată cât de eterogene sau omogene sunt valorile.

Figura 1: Plot instanțe și centre în axele discriminante



Se poate observa că modelul ar funcționa cel mai bine pe Europa Centrală și de Est și pe Europa de Nord, în timp ce țările din Europa de Sud și Europa de Vest nu ar fi clasificate corect.

Mai departe, sunt prezentate graficele de distribuție pentru fiecare grupă, reprezentând o altă ilustrare a omogenității grupelor clasificate de model.

Pentru distribuția în axa discriminantă z1are loc o distincție clară între cele 4 subregiuni ale Europei. În graficele pentru axele discriminante z2 și z3 , distincția dintre grupe este mult mai puțin observabilă.

Figura 2: Distribuția în axa discriminantă z1

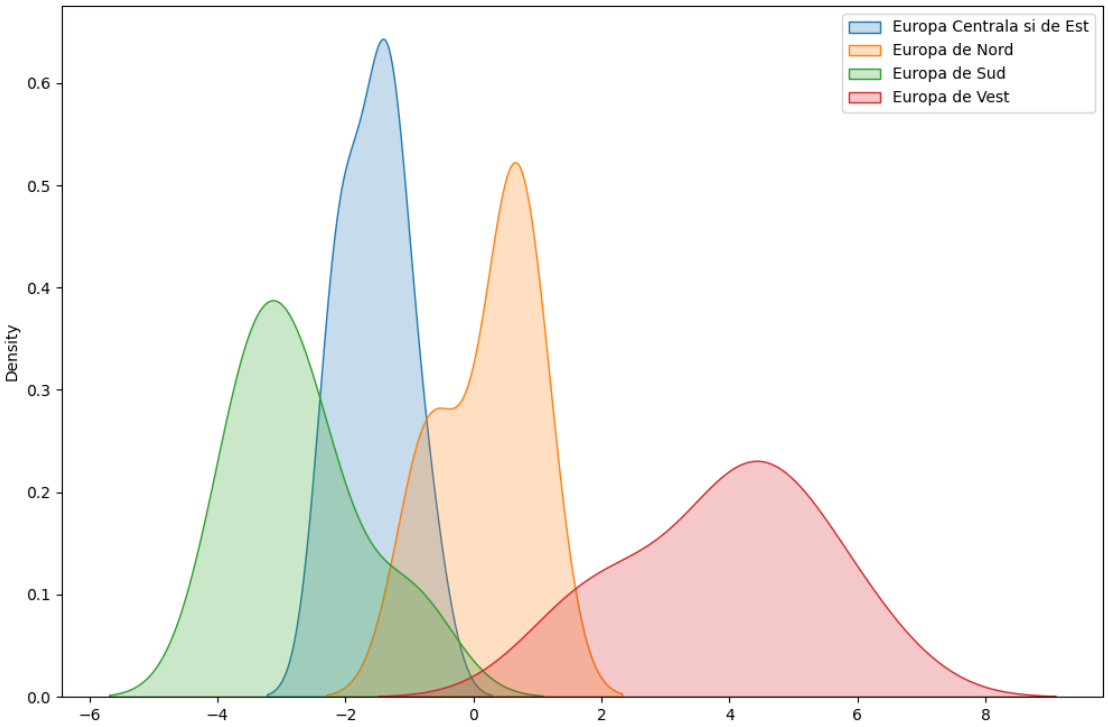


Figura 3: Distribuția în axa discriminantă z2

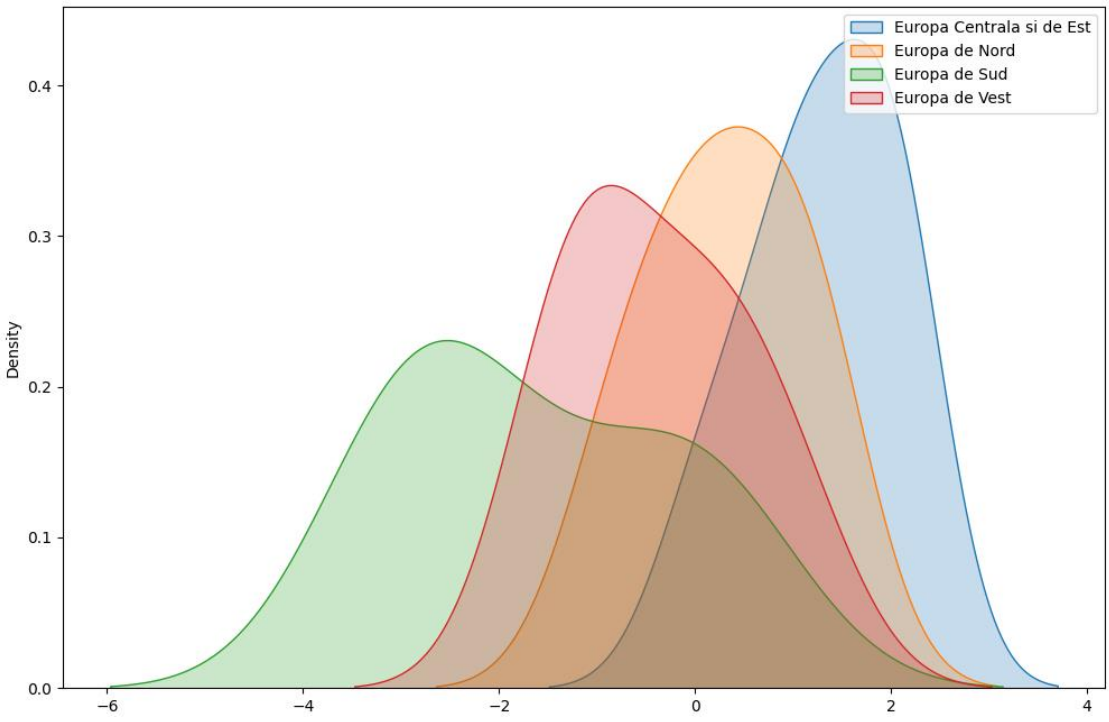
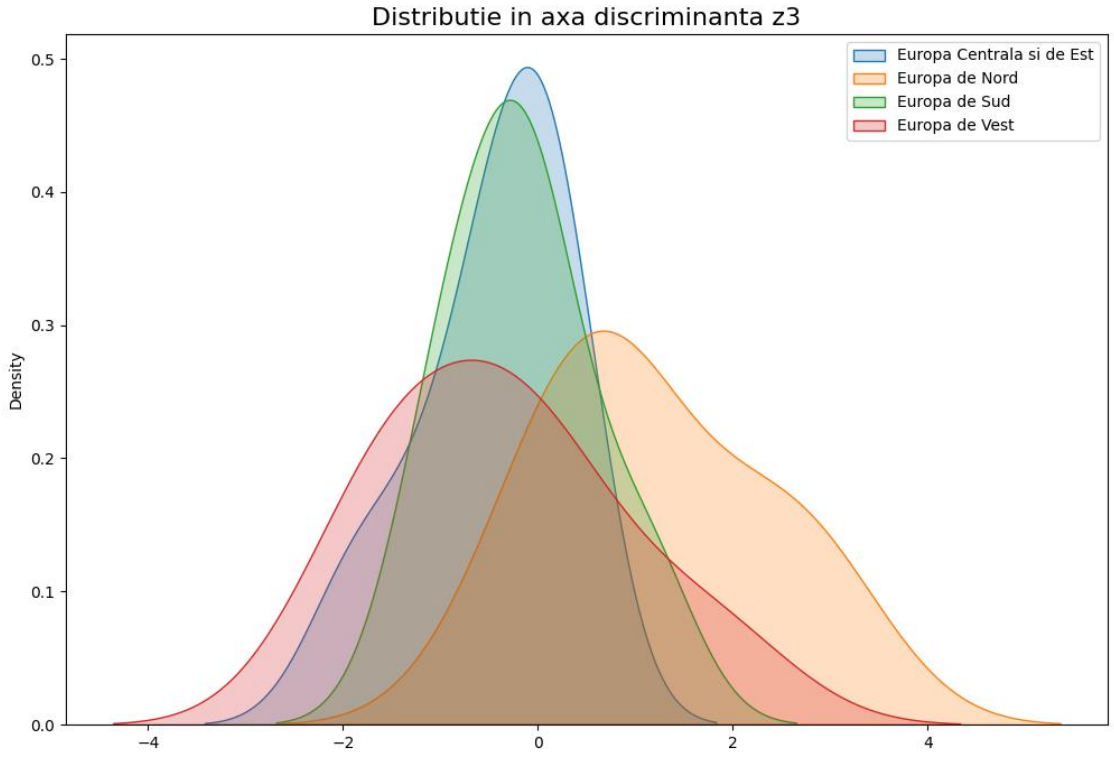


Figura 4: Distribuția în axa discriminantă z3



După clasificarea în setul de învățare (utilizând analiza discriminantă), s-a calculat matricea de clasificare eronată pentru a separa clasificările corecte de cele eronate și pentru a calcula acuratețea modelului. Aceasta a avut un scor de 88.89%.

	Europa Centrala si de Est	Europa de Nord	Europa de Sud	Europa de Vest
Europa Centrala si de Est	8	0	0	0
Europa de Nord	1	5	0	0
Europa de Sud	1	0	5	0
Europa de Vest	0	1	0	6

Aplicând algoritmul pe datele din 2014, s-a observat că, de această dată, acuratețea a fost mai mică: 66.67%.

	Europa Centrala si de Est	Europa de Nord	Europa de Sud	Europa de Vest
Europa Centrala si de Est	5	0	3	0
Europa de Nord	3	3	0	0
Europa de Sud	0	0	6	0
Europa de Vest	1	1	1	4

Aplicarea modelului bayesian

Un alt algoritm de clasificare supervizată folosit este clasificatorul Bayesian, ce reprezintă o tehnică care atribuie etichete de clasă (dintr-o mulțime finită) instanțelor noi, reprezentate ca vectori de valori pentru diverse caracteristici.

Pentru setul de învățare, acuratețea a fost 51.85%, iar matricea de clasificări este:

	Europa Centrala si de Est	Europa de Nord	Europa de Sud	Europa de Vest
Europa Centrala si de Est	7	0	1	0
Europa de Nord	3	3	0	0
Europa de Sud	4	0	1	1
Europa de Vest	0	2	2	3

Pentru setul de testare, acuratețea modelului a fost de 59.25%, iar matricea de clasificări este:

	Europa Centrala si de Est	Europa de Nord	Europa de Sud	Europa de Vest
Europa Centrala si de Est	8	0	8	0
Europa de Nord	3	3	0	0
Europa de Sud	4	0	1	1
Europa de Vest	0	2	2	3

Aplicarea modelului SVM

Următorul model de clasificare supervizată luat în calcul este algoritmul SVM (support vector machines). Atât pentru setul de testare, cât și pentru cel de învățare, acuratețea obținută a fost de 44.44%.

Matricea pentru setul de învățare:

	Europa Centrala si de Est	Europa de Nord	Europa de Sud	Europa de Vest
Europa Centrala si de Est	8	0	0	0
Europa de Nord	6	0	0	0
Europa de Sud	4	0	1	1
Europa de Vest	4	0	0	3

Matricea pentru setul de testare:

	Europa Centrala si de Est	Europa de Nord	Europa de Sud	Europa de Vest
Europa Centrala si de Est	8	0	0	0
Europa de Nord	6	0	0	0
Europa de Sud	4	0	1	1
Europa de Vest	4	0	0	3

Aplicarea modelului arborelui de decizie

Ultimul algoritm de clasificare folosit este arborele de decizie, fiind modelul care a avut și cea mai mare acuratețe pentru ambele seturi de date.

Pentru setul de învățare (datele corespunzătoare anului 2019), algoritmul a prezis corect toate subregiunile Europei. Matricea de clasificare este:

	Europa Centrala si de Est	Europa de Nord	Europa de Sud	Europa de Vest
Europa Centrala si de Est	8	0	0	0
Europa de Nord	0	6	0	0
Europa de Sud	0	0	6	0
Europa de Vest	0	0	0	7

În ceea ce privește setul de testare (datele corespunzătoare anului 2014), acuratețea modelului este de 70.37%.

	Europa Centrala si de Est	Europa de Nord	Europa de Sud	Europa de Vest
Europa Centrala si de Est	5	3	0	0
Europa de Nord	0	6	0	0
Europa de Sud	4	0	2	0
Europa de Vest	0	1	0	6

Concluzii

Folosind numeroși algoritmi de clasificare supervizată din biblioteca *sklearn*, am încercat să prezic subregiunile Europei pentru țările Uniunii Europene, alegând ca variabile independente indicatori de ordin economic.

Pentru setul de învățare, au fost folosite date din 2018-2019, iar pentru cel de testare s-au folosit date din 2014.

Pentru ambele seturi de date, clasamentul algoritmilor de clasificare, în ordine descrescătoare, este:

- Modelul arborelui de decizie;
- Analiza liniară discriminantă;
- Clasificarea bayesiană;
- Modelul SVM.

Bibliografie

https://en.wikipedia.org/wiki/EuroVoc

https://ro.qaz.wiki/wiki/Linear_discriminant_analysis

https://ro.wikipedia.org/wiki/Clasificator_bayesian_naiv#Introducere

Sursa datelor: Eurostat