Documentație proiect PCLP3 - Partea I

Ivan Andrei-Cristian, 311CA

May 2025

1 Tipul problemei

Problema aleasă presupune plecarea de la un dataset care conține informații relevante pentru predicția magnitudinii unor cutremure. Astfel, problema aleasă este o problemă de **regresie**.

2 Structura setului de date

- Setul de date va conține 1 000 linii și va fi împărțit aleator între test și train.
- Subset de antrenare: 800 de instanțe.
- Subset de testare: 200 de instanțe.

3 Dataset

Setul de date va conține 8 coloane, fiecare coloană având tipurile de date următoare:

- zi: `int64`
- longitudine: `float64`
- latitudine: `float64`
- adancime epicentru: `float64`
- tip placa: `object`
- replici: `int64`
- magnitudine ultimul: `float64`
- magnitudine: `float64`

Setul de date va fi generat aleator. Datele vor fi generate folosind distribuția normală și uniformă, presupunând următoarele ipoteze:

- Zilele vor reprezenta zilele unei luni, deci vor fi cuprinse între 0 și 30.
- Latitudinea va fi curpinsă între 0 și 90, iar longitudinea între 0 și 180.
- Adâncimea epicentrului va fi cuprinsă între 50 și 100 km.
- Plăcile pot fi: **divergente, convergente** sau **de transformare**.
- Ultime magnitudini vor fi cuprinse între 1.5 și 9.
- Magnitudinile vor fi calculate conform formulei: 2 + un număr random între 0 și 3 în funcție de tipul plăcii $- 0.2 \cdot$ (magnitudine ultimul $- 2$) $+ 0.1 \cdot$ replici $+ 0.025 \cdot$ adâncime. Ne asigurăm la final că magnitudinea este cuprinsă între 2 și 9.

La final, vom alege un procent aleator între 0% și 10% de valori din fiecare coloană pentru a le elimina.

4 EDA

4.1 Analiza valorilor lipsă

- Pe coloana zi lipsesc 49 linii (4.9%)
- Pe coloana longitudine lipsesc 40 linii (4%)
- Pe coloana latitudine lipsesc 66 linii (6.6%)
- Pe coloana adancime epicentru lipsesc 35 linii (3.5%)
- Pe coloana tip placa lipsesc 14 linii (1.4%)
- Pe coloana replici lipsesc 27 linii (2.7%)
- Pe coloana magnitudine ultimul lipsesc 23 linii (2.3%)

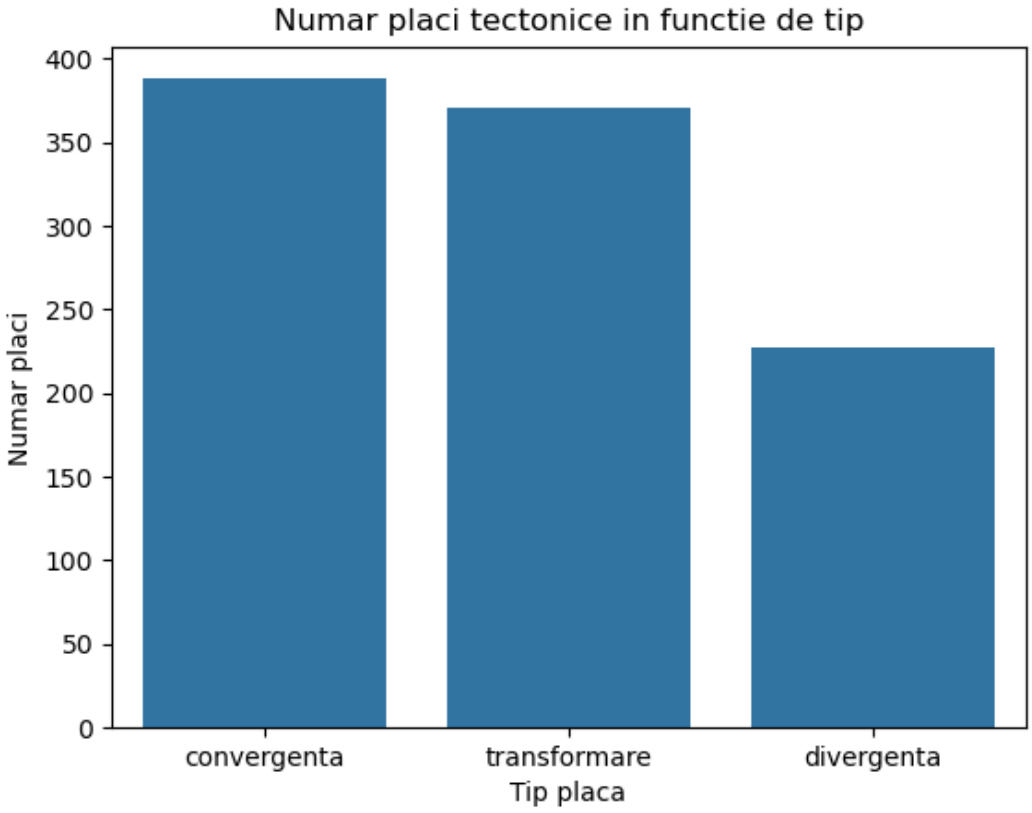
Toate datele lipsă le vom inputa. În cazul coloanelor numerice, În loc de `NaN` vom Înlocui cu media aritmetică de pe acea coloană (folosind funcția `df.mean()`, iar în cazul coloanelor de tip `object`, vom înlocui cu cel mai frecvent string de pe coloană (folosind funcția `df.mode()`).

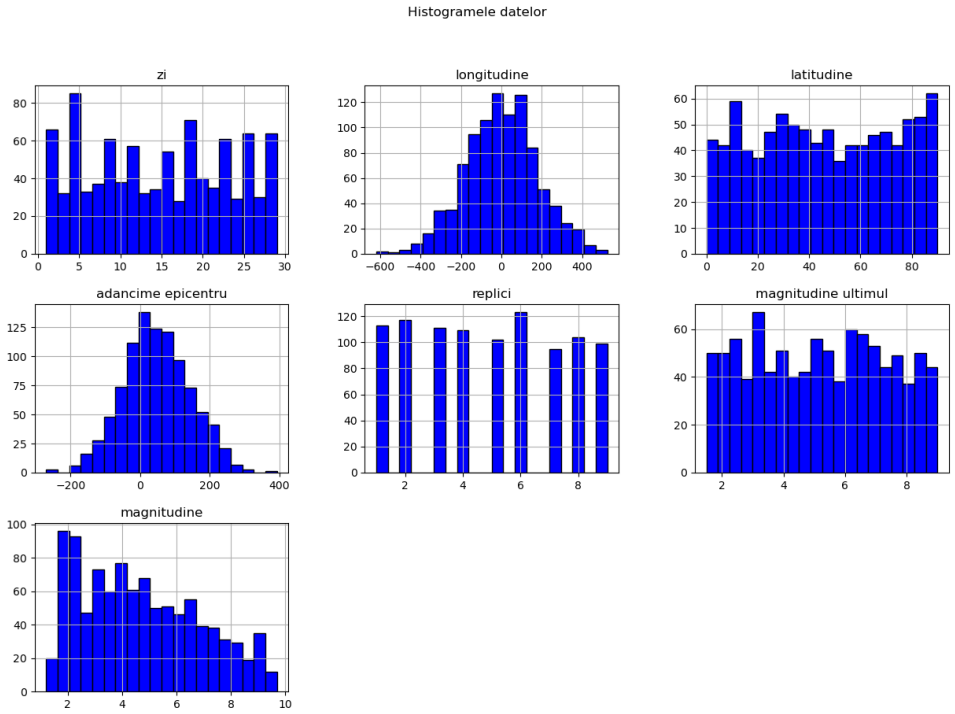
4.2 Statistici descriptive

Utilizând funcția `df.describe()`, obținem următoarele detalii:

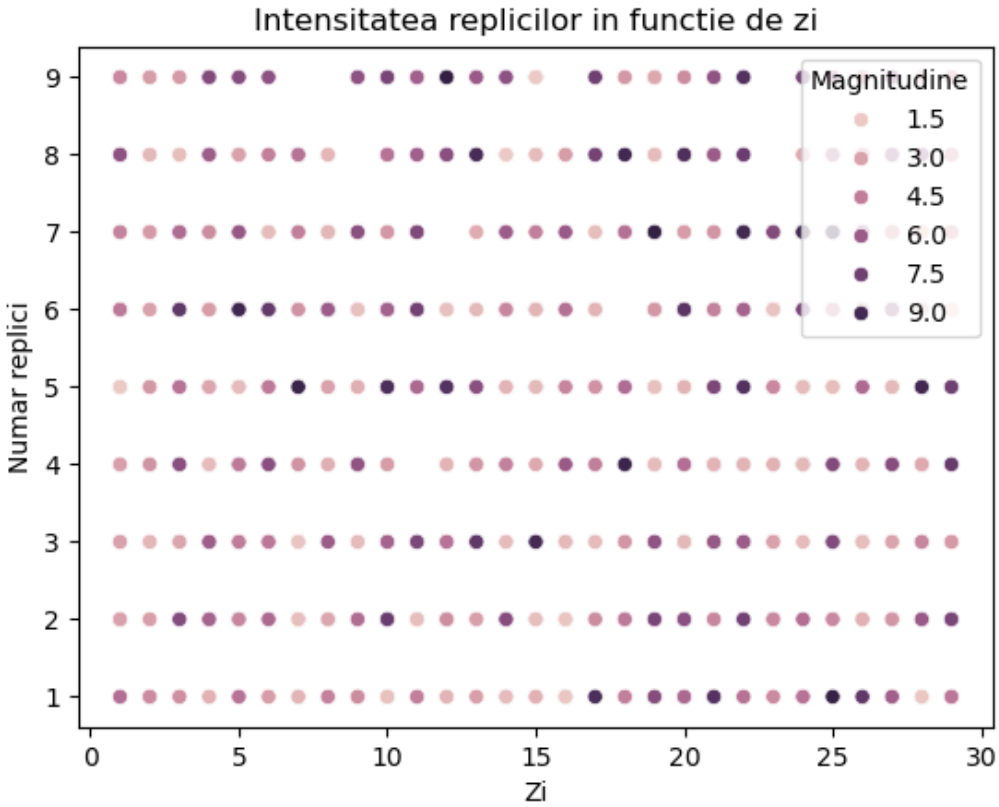
	zi	longitudine	latitudine	adancime epicentru	replici	magnitudine ultimul	magnitudine
count	951.000000	960.000000	934.000000	965.000000	973.000000	977.00000	1000.000000
mean	14.695058	1.532426	45.951523	47.743723	4.883864	5.18567	4.689800
std	8.445713	179.224720	26.577811	96.678248	2.570783	2.15649	2.176118
min	1.000000	-619.303379	0.049137	-270.427106	1.000000	1.50000	1.200000
25%	7.000000	-116.660280	23.859504	-18.065743	3.000000	3.20000	2.800000
50%	15.000000	3.931552	45.406484	42.023439	5.000000	5.20000	4.300000
75%	22.000000	114.588983	69.589967	112.990789	7.000000	7.00000	6.300000
max	29.000000	524.090951	89.982772	393.480755	9.000000	9.00000	9.700000

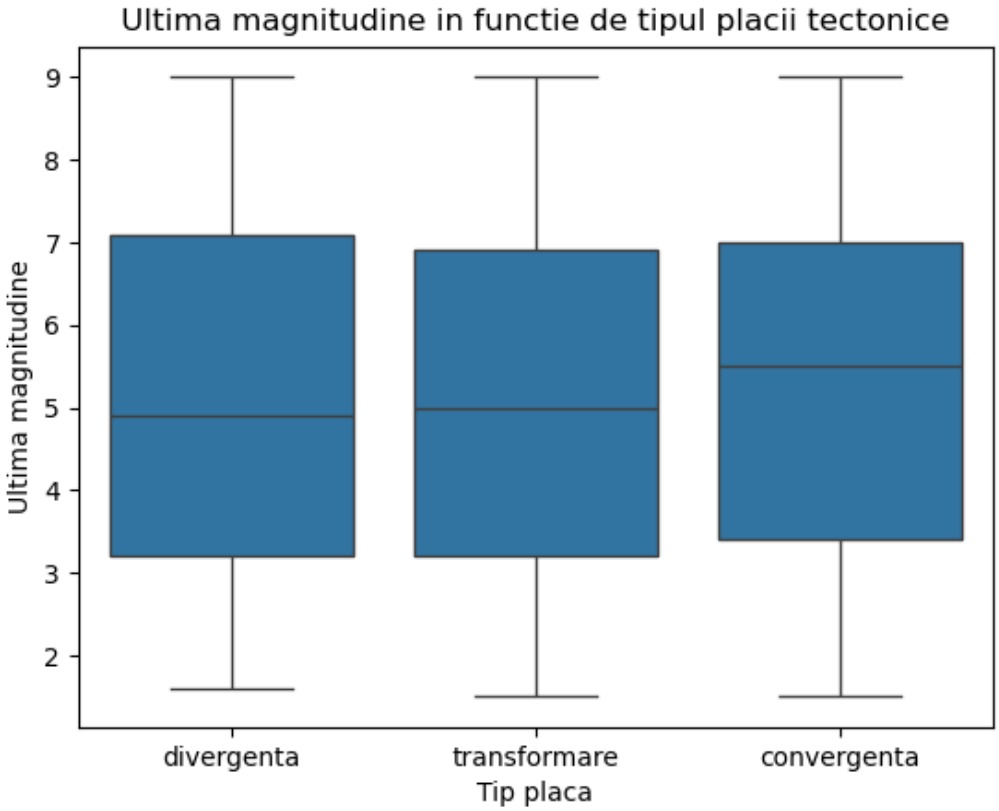
4.3 Analiza distribuției variabilelor

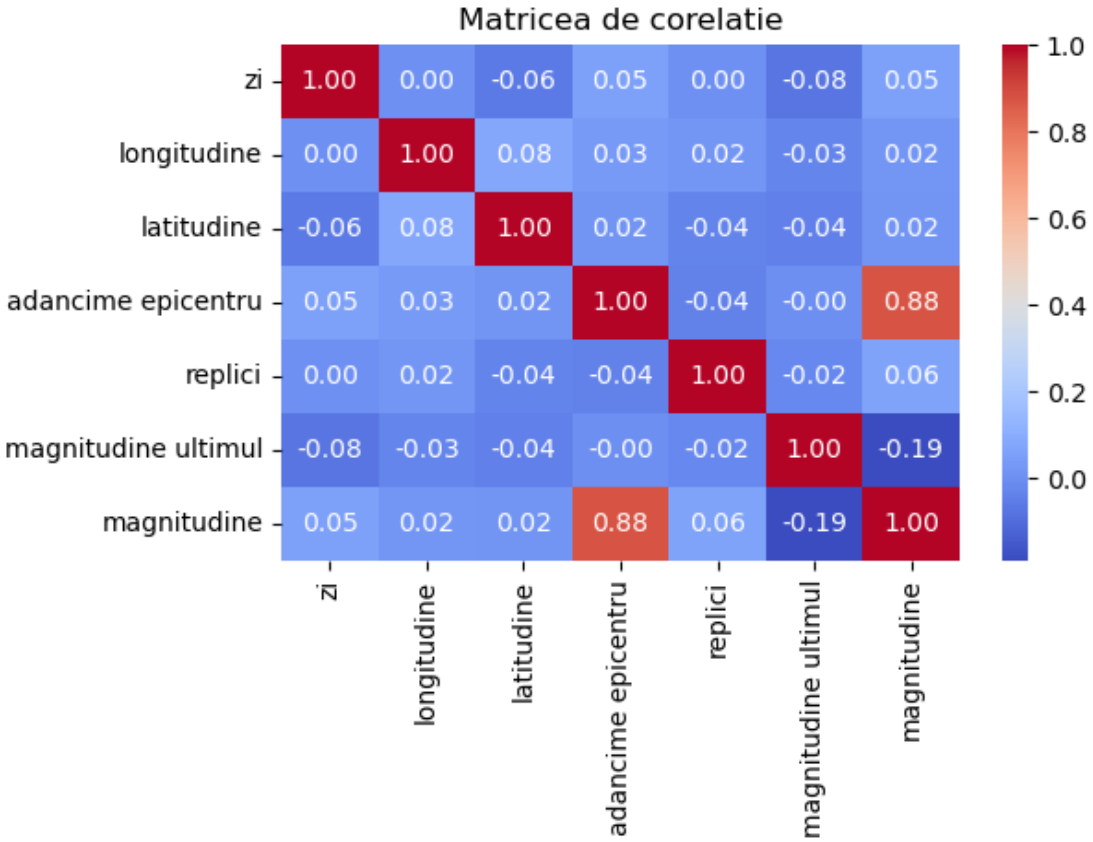




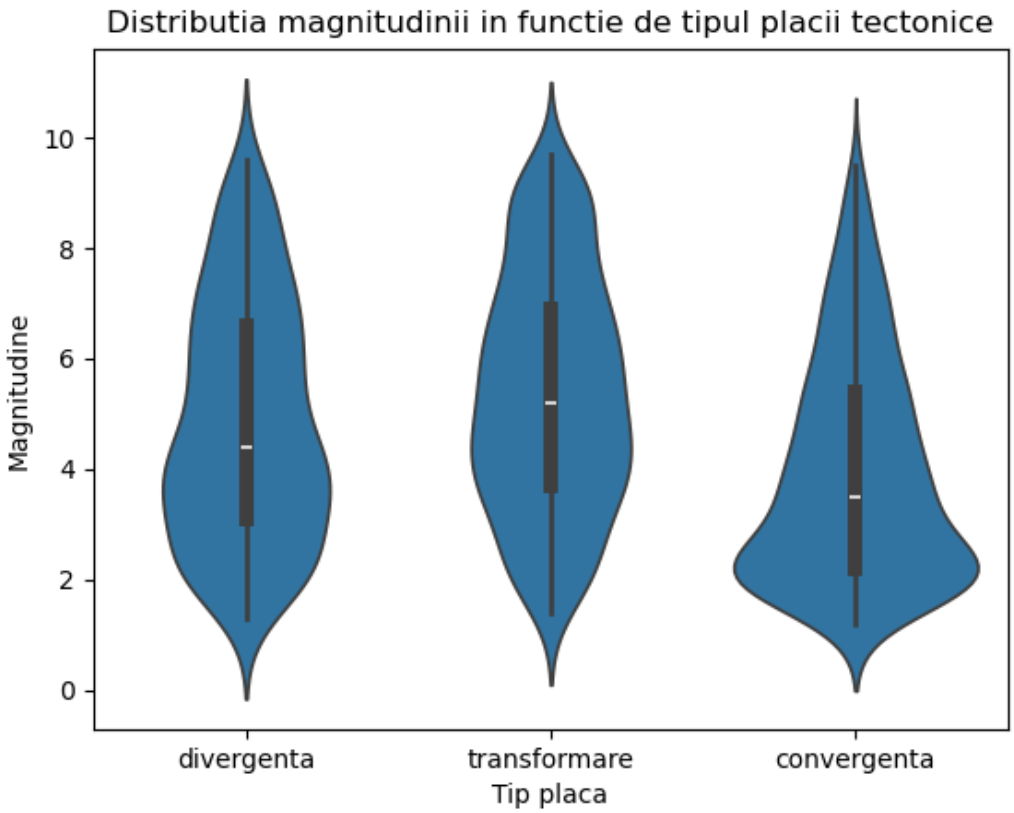
4.4 Analiza corelațiilor

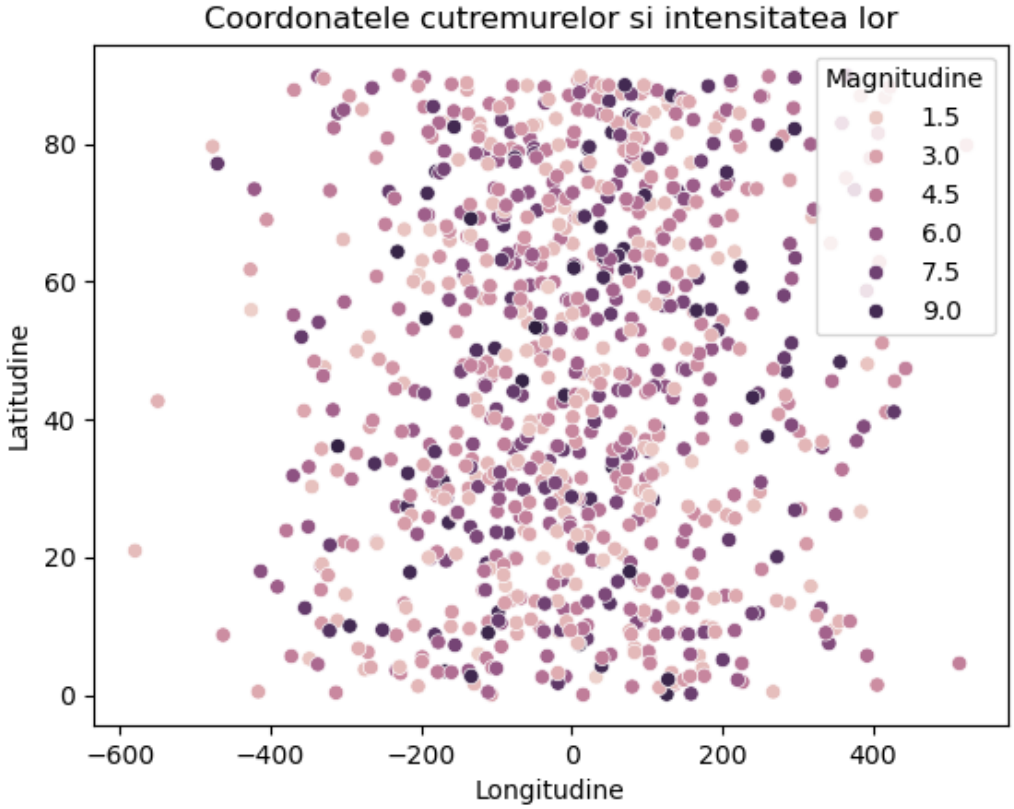



Ultima magnitudine in functie de tipul placii tectonice
Ultima magnitudine
9
8
7
6
5
4
3
2
divergenta
transformare
convergenta
Tip placa


Matricea de corelatie
zi
longitudine
latitudine
adancime epicentru
replici
magnitudine ultimul
magnitudine
1.00 0.00 -0.06 0.05 0.00 -0.08 0.05
0.00 1.00 0.08 0.03 0.02 -0.03 0.02
-0.06 0.08 1.00 0.02 -0.04 -0.04 0.02
0.05 0.03 0.02 1.00 -0.04 -0.00 0.88
0.00 0.02 -0.04 -0.04 1.00 -0.02 0.06
-0.08 -0.03 -0.04 -0.00 -0.02 1.00 -0.19
0.05 0.02 0.02 0.88 0.06 -0.19 1.00
1.0
0.8
0.6
0.4
0.2
0.0

4.5 Analiza relațiilor cu variabila țintă





5 Antrenarea și evaluarea unui model de bază

- Fiindcă abordăm o problemă de regresie, vom folosi, din biblioteca `scikit-learn`, modelul de **regresie liniară**.
- Folosind metrica **RMSE**, obținem o eroare de 0.49.