ML
Tema 2 — Inteligență Artificială

Alexandru Sima (332CA)

22 mai 2025

Rezumat

Hello world!

Cuprins

1 Poluarea aerului

Primul set de date conține date despre diferiți parametri măsurați ai aerului, în peste 20.000 de orașe din întreaga lume. Prin antrenarea unui model de învățare automată, se dorește clasificarea orașelor în funcție de gradul de riscuri pentru sănătate.

1.1 Analiza datelor

1.1.1 Analiza valorică

Setul de date conține 15 atribute, dintre care 7 numerice și 8 categorice (incluzând și atribul țintă "AQI_Category").

	AQI_Value	CO_Value	Ozone_Value	NO2_Value	PM25_Value	VOCs	SO2
înregistrări	23463	23463	21117	23463	23463	23463	23463
val. medie	72.01	1.36	35.23	43.08	68.51	185.05	4.44
dev. standard	56.05	1.83	28.14	196.07	54.79	140.48	5.95
val. minimă	6.00	0.00	0.00	0.00	0.00	12.41	-18.52
cuartila 1	103.26	1.00	21.00	0.00	35.00	103.26	0.73
mediana	142.97	1.00	31.00	1.00	54.00	142.97	4.28
cuartila 3	204.22	1.00	40.00	4.00	79.00	204.22	7.91
val. maximă	1280.98	133.00	222.00	1003.06	500.00	1280.98	234.69

TODO

1.1.2 Echilibrul claselor

1.1.3 Corelația între atribute

1.1.4 Atribute numerice

Aplicând testul Pearson pentru a determina corelația liniară dintre atributele numerice, se obține matricea din (1). Se poate presupune astfel că atributele "AQI_Value", "PM25_Value" și "VOCs" sunt foarte puternic corelate între ele, având coeficientul de corelație ≥ 0.98 și că atributul "NO2_Value" nu este corelat cu niciun altul. Într-adevăr, primele 3 atribute sunt puternic corelate, acest fapt observându-se trasând graficele valorilor (3). Testul Pearson oferă însă doar informații despre corelația liniară, astfel că, aplicând testul Spearman, se obține matricea din (2), care arată existența unor corelații între "NO2_Value" și alți parametri.

1.2 Preprocesarea datelor

1.2.1 Imputarea valorilor lipsă

1.2.2 Eliminarea valorilor extreme

1.2.3 Eliminarea atributelor redundante

Urmărind analiza corelației dintre atributele numerice efectuată la (1), se observă că atributele "AQI_Value", "PM25_Value" și "VOCs" sunt foarte puternic corelate, deci 2 dintre cele 3 pot fi eliminate. Am ales să păstrez "AQI_Value", neexistând diferențe semnificative între numărul de valori înregistrate ale fiecărui atribut sau distribuțiile acestora.

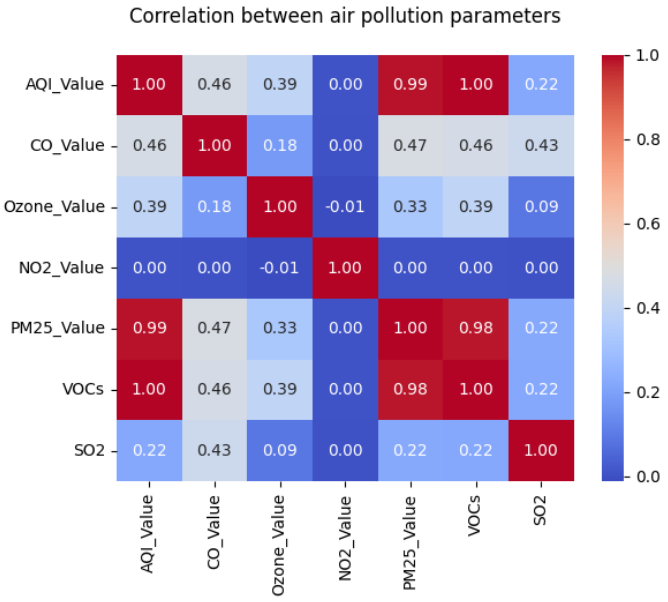

Figura 1: Corelația dintre atributele dataset-ului

1.2.4 Normalizarea datelor

1.3 Învățarea automată

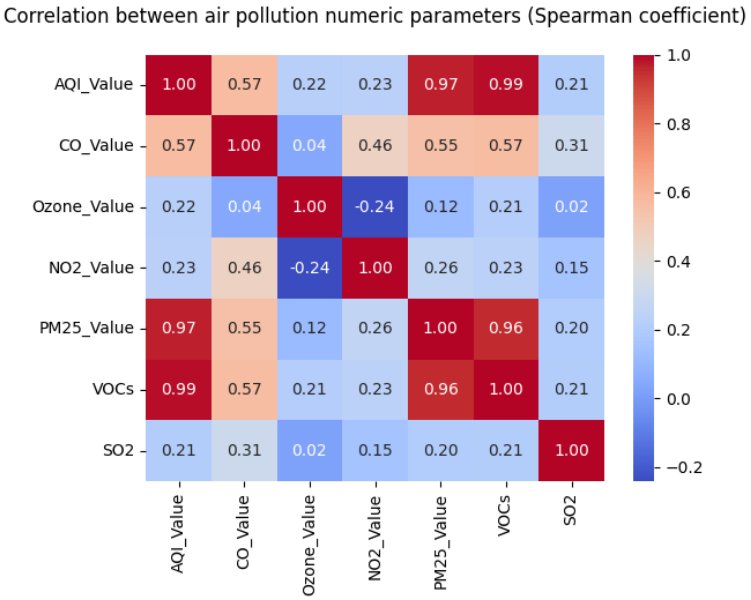

Figura 2: Corelația dintre atributele dataset-ului, folosind coeficientul Spearman

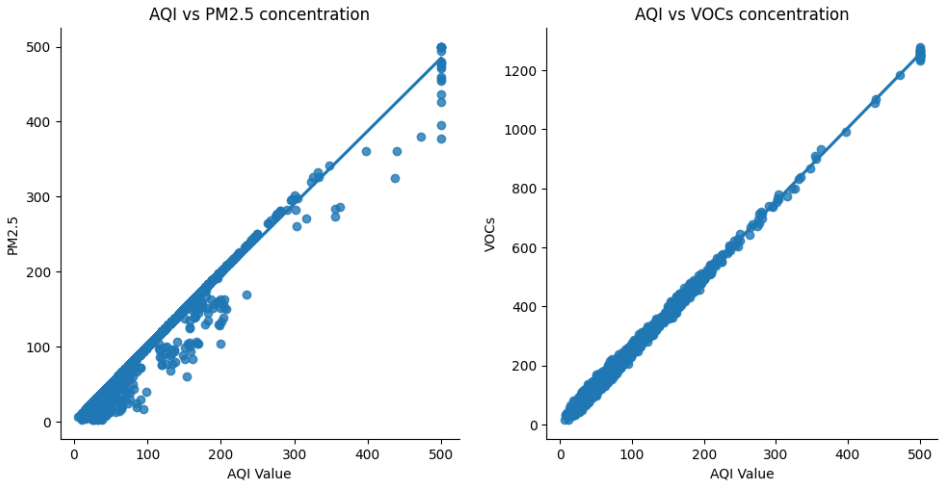

Figura 3: Corelația liniară dintre AQI_Value și PM25_Value, respectiv VOCs